Calculation of the efficient influence function

Brice Ozenne

September 28, 2021

1 Exercise 1

Consider a random variable X for which we would like to estimate the cumulative distribution function (CDF) $F_{\mathbb{P}}(x) = \mathbb{P}[X \leq x]$. X could for instance be glucose level. We have a patient with a glucose level of x and we would like to know how often such "low" level occurs. To do so we have a sample of n iid observations $(X_i)_{i \in \{1,\ldots,n\}}$.

We will use the efficient influence function to derive a good estimator. This is done by calculating the Gateau derivative at δ_X of the CDF:

$$\Psi(\mathbb{P}) = F_{\mathbb{P}}(x) = \int_{-\infty}^{x} d\mathbb{P}(t)$$

To do so we introduce a small variation ε at a given point X_i, i.e. we perturbate the CDF of X into $\mathbb{P}_{\varepsilon(i)} = \mathbb{P} + \varepsilon\delta_{X_i}$. Here $\delta_.$ indicate the dirac mesure in the point X_i. We note the derivative of this perturbation with respect to ε is $\frac{\partial \mathbb{P}_{\varepsilon(i)}}{\partial \varepsilon}\Big|_{\varepsilon=0} = \delta_{X_i}$ which is a shorthand for $\mathbb{1}_{x=X_i}$. We therefore obtain:

$$\begin{aligned}
\frac{\partial \Psi(\mathbb{P}_{\varepsilon(i)})}{\partial \varepsilon}\Bigg|_{\varepsilon=0} &= \int_{-\infty}^{x} d\, \frac{\partial \mathbb{P}_{\varepsilon(i)}(t)}{\partial \varepsilon}\Bigg|_{\varepsilon=0} \\
&= \int_{-\infty}^{x} d\delta_{X_i} \\
&= \mathbb{1}_{X_i \leq x}
\end{aligned}$$

Averaging this derivative over n iid observations leads to the usual estimator (empirical distribution function):

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \Psi(\mathbb{P}_{\varepsilon(i)})}{\partial \varepsilon}\Bigg|_{\varepsilon=0} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} = F_{\mathbb{P}_n}(x)$$

2 Exercise 2

Consider a set of random variables $O = (Y, A, X) \in \mathbb{R} \times \{0, 1\} \times \mathbb{R}^d$. We denote by $\mathbb{P} = \mathbb{P}(dy, a, dx)$ its joint distribution and assume to observe a sample of iid observations $(O_i)_{i\in\{1,\dots,n\}} = (Y_i, A_i, X_i)_{i\in\{1,\dots,n\}}$. We could for instance measure the glucose level Y following two different diet $A = 0$ and $A = 1$ in patients with various BMI and age $X = (X_1, X_2)$.

We would like to estimate the conditional mean of Y when A is set to a certain value (say a). In our example that could be to estimate the mean glucose level after each diet. For a given a, we will denote:

- $f_{\mathbb{P}}(A, X) = \int_{y\in\mathbb{R}} y dF_{\mathbb{P}}(Y|A, X)$ the conditional expectation of Y given A and X. Here $F_{\mathbb{P}}(Y|A, X)$ denotes the CDF of Y conditional to A and X.
- $\pi_{\mathbb{P}}(X) = \mathbb{P}[A = a|X]$ the conditional probability of $A = a$ given X.
- $\mu_{\mathbb{P}}$ the marginal distribution of X (i.e. CDF of X).

We first re-express of target parameter:

$$\Psi_a(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[f_{\mathbb{P}}(a, X)] = \int_{x\in\mathbb{R}^d} f_{\mathbb{P}}(a, x) d\mu_{\mathbb{P}}(x)$$

as an (explicit) function of the joint distribution:

$$\Psi_a(\mathbb{P}) = \int_{x\in\mathbb{R}^d} \left(\int_{y\in\mathbb{R}} y \frac{\mathbb{P}(y, a, x)}{\int_{y\in\mathbb{R}} \mathbb{P}(y, a, x)} \right) d \left(\sum_{a^*\in\{0,1\}} \int_{y\in\mathbb{R}} \mathbb{P}(x, a^*, y) \right)$$

Similarly to the previous exercise we will compute the Gateau derivative at δ_O of the target parameter. This time $\mathbb{P}_{\varepsilon(i)} = \mathbb{P} + \varepsilon\delta_{O_i}$ and the derivative of this perturbation with respect to ε is $\frac{\partial \mathbb{P}_{\varepsilon(i)}}{\partial\varepsilon}\Big|_{\varepsilon=0} = \delta_{O_i}$. To ease calculation we will start by evaluating this derivative for the marignal distribution of X:

$$\begin{aligned}
\frac{\partial \mu_{\mathbb{P}_{\varepsilon(i)}}(x)}{\partial\varepsilon}\bigg|_{\varepsilon=0} &= \sum_{a^*\in\{0,1\}} \int_{y\in\mathbb{R}} \frac{\partial \mathbb{P}_{\varepsilon(i)}(x, a^*, y)}{\partial\varepsilon}\bigg|_{\varepsilon=0} \\
&= \sum_{a^*\in\{0,1\}} \int_{y\in\mathbb{R}} \delta_{O_i}(y, a^*, x) \\
&= \delta_{X_i}(x)
\end{aligned}$$

and for the conditional expectation:

$$
\begin{aligned}
\left.\frac{\partial f_{\mathbb{P}_{\varepsilon(i)}}(a,x)}{\partial\varepsilon}\right|_{\varepsilon=0} &= \int_{y\in\mathbb{R}} y\frac{\left.\frac{\partial\mathbb{P}_{\varepsilon(i)}(y,a,x)}{\partial\varepsilon}\right|_{\varepsilon=0}}{\int_{y\in\mathbb{R}}\mathbb{P}(y,a,x)} - \int_{y\in\mathbb{R}} y\frac{\mathbb{P}(y,a,x)\int_{y\in\mathbb{R}}\left.\frac{\partial\mathbb{P}_{\varepsilon(i)}(y,a,x)}{\partial\varepsilon}\right|_{\varepsilon=0}}{\left(\int_{y\in\mathbb{R}}\mathbb{P}(y,a,x)\right)^2} \\
&= \int_{y\in\mathbb{R}} y\frac{\delta_{O_i}(y,a,x)}{\mathbb{P}(a,x)} - \int_{y\in\mathbb{R}} y\frac{\mathbb{P}(y,a,x)\int_{y\in\mathbb{R}}\delta_{O_i}(y,a,x)}{\mathbb{P}(a,x)^2} \\
&= Y_i\frac{\delta_{O_i}(a,x)}{\mathbb{P}(a,x)} - \int_{y\in\mathbb{R}} y\frac{\mathbb{P}(y,a,x)\delta_{O_i}(a,x)}{\mathbb{P}(a,x)^2} \\
&= \left(\frac{Y_i\mathbb{1}_{a=A_i}}{\mathbb{P}(a,x)} - \frac{\mathbb{1}_{a=A_i}}{\mathbb{P}(a,x)}\frac{\int_{y\in\mathbb{R}} y\mathbb{P}(y,a,x)}{\mathbb{P}(a,x)}\right)\mathbb{1}_{x=X_i} \\
&= \left(\frac{Y_i\mathbb{1}_{a=A_i}}{\mathbb{P}(a,x)} - \frac{\mathbb{1}_{a=A_i}f_{\mathbb{P}}(a,x)}{\mathbb{P}(a,x)}\right)\mathbb{1}_{x=X_i} \\
&= \frac{\mathbb{1}_{a=A_i}}{\mathbb{P}(a,x)}(Y_i - f_{\mathbb{P}}(a,x))\mathbb{1}_{x=X_i}
\end{aligned}
$$

So using the chain rule (or the product rule) we obtain that:

$$
\begin{aligned}
\left.\frac{\partial\Psi_a(\mathbb{P})}{\partial\varepsilon}\right|_{\varepsilon=0} &= \int_{x\in\mathbb{R}^d}\frac{\mathbb{1}_{a=A_i}}{\mathbb{P}(a,x)}(Y_i - f_{\mathbb{P}}(a,x))\mathbb{1}_{x=X_i}d\mu_{\mathbb{P}}(x) + \int_{x\in\mathbb{R}^d} f_{\mathbb{P}}(a,x)\delta_{X_i}(x) \\
&= \frac{\mathbb{1}_{a=A_i}}{\mathbb{P}(a,X_i)}(Y_i - f_{\mathbb{P}}(a,X_i))\mu_{\mathbb{P}}(X_i) + f_{\mathbb{P}}(a,X_i) \\
&= \frac{\mathbb{1}_{a=A_i}}{\pi(a,X_i)}(Y_i - f_{\mathbb{P}}(a,X_i)) + f_{\mathbb{P}}(a,X_i),
\end{aligned}
$$

where $\pi(a,x) = \mathbb{P}(A = a \mid X = x)$.