The Linkage between the Score Function and the Efficient Influence Curve

Scrambled notes: some key concepts in Targeted Maximum Likelihood Estimation

Sylvia Cheng

December 9, 2025

Disclaimer: This note is to help myself reorganize my understanding for this topic. It is not TMLE 101. Mathematical rigor is not guaranteed due to my own limitations. Use with caution.

Quick overview and why this note

In Targeted Maximum Likelihood Estimation (TMLE), a critical step in obtaining the correct final estimate is updating the initial estimate of the empirical probability distribution P_n with respect to a target parameter $\Psi(P)$ to remove bias. In the theories behind TMLE, this is done by going through a parametric path from the initial estimate and solving score equations. This path is usually referred to as the least favorable path, orleast favorable submodel.

As shown in Mark's book [1], the tangent space is the span of all score functions of all possible submodels P_ϵ, and this tangent space should contain the the Efficient Influence Curve (EIC) when it is at $\epsilon = 0$. Naturally, we can see that the EIC can be expressed as a linear combination of those score functions in general.

Excitingly, in the least favorable submodel, the score is exactly equal to the EIC. How so? The writing below exists to clarify this.

Targeting and submodels

Scores and EIC

Following notations used in [1], let P_n^0 denote the initial estimate. Denote a parametric submodel as $\{P(\epsilon) : \epsilon\}$ and $P(\epsilon = 0) = P_n^0$ (so that it is centered at P_n^0).

If $P(\epsilon)$ targets $\Psi(P)$, then its scores at $\epsilon = 0$ span the direction of the EIC, denoted as $D^*(P)$. This means that if $\epsilon \in \mathbb{R}^k$, there exists a vector $\mathbf{a} \in \mathbb{R}^k$ such that

$$\mathbf{a}^\top \left(\frac{d}{d\epsilon} \log P(\epsilon) \Big|_{\epsilon=0} \right) = D^*(P)$$

EIC and the least favorable path

I was confused about the term "least favorable" for a while, until I connected it to the direction of the EIC. This is the direction in which the estimator has the highest variance, and the parameter is most sensitive to fluctuations. It sounds pretty bad, but to reduce bias in our estimation, this is exactly the direction to go. Mathematically, the name "least favorable" comes from the fact that this direction has the minimum Fisher Information relative to the parameter of interest.

Among all submodels, there exists a submodel s.t. its score equals the EIC:

$$S(O) \equiv \frac{d}{d\epsilon} \log P(\epsilon) \Big|_{\epsilon=0} = D^*(P)(O)$$

This is the least favorable model. Furthermore, a submodel is least favorable if and only if its score spans the same linear subspace as the EIC. By having this model, we can focus on using MLE to solve the EIC until $\sum D^*(O_i) = 0$ for TMLE.

Unbiased + Lowest Variance = Efficiency!

Here we should first revisit what the Cramér-Rao Lower Bound (CRLB) is and how it is related to the Fisher Information. See lecture notes from STAT 210A for more details. Some quick notes if skipping the upcoming subsubsection:

- Fisher Information for a submodel with score S: $\mathcal{I}(\epsilon) = E[S^2]$
- The semi-parametric efficiency bound: $\text{Var}(\Psi)_{\text{eff}} = E[(D^*)^2]$

When $S(O) = D^*(O)$, since the FI is $\mathcal{I} = E[(D^*)^2]$, then estimating the parameter in this path is as hard as estimating it in the full semi-parametric model (set of all possible semi-parametric submodels). This means that maximizing likelihood along this path can remove bias the most for Ψ.

Can be skipped: CRLB and the Least favorable submodel

For our parameter of interest $\Psi(P)$ to be estimable, it must be pathwise differentiable, meaning that there exists a continuous linear functional such that for any submodel (Riesz Representation Thm.), we have

$$\frac{d}{d\epsilon}\Psi(P_\epsilon)\bigg|_{\epsilon=0} = E_P[D^*(O)S_\epsilon(O)] = \langle D^*, S_\epsilon\rangle_{L_2(P)}$$

From CRLB, we know that the variance of any unbiased estimator $\hat{\Psi}$ restricted to a specific submodel is lower-bounded by the inverse Fisher information of the submodel.

Furthermore, when estimating a smooth function of a parameter ϵ (in this case $\theta = \Psi(P_\epsilon)$), the lower bound is scaled by the squared derivative of the function (note: compare this to the Delta Method variance version):

$$\text{Var}(\hat{\Psi}) \geq \left(\frac{d\Psi}{d\epsilon}\right)^2 \cdot \text{Var}(\hat{\epsilon})_{\min}$$

and we can further write

$$\text{Var}(\hat{\Psi}) \geq \left(\frac{d\Psi}{d\epsilon}\right)^2 \cdot \frac{1}{E[S_\epsilon^2]}$$

$$\text{Var}(\hat{\Psi}(P_\epsilon)) \geq \frac{\langle D^*, S_\epsilon\rangle^2}{||S_\epsilon||^2}$$

Naturally, we can now derive the semi-parametric efficiency bound, which is defined as the supremum of the parametric bounds over all possible submodels, which is

$$\sup_{\{S\in\mathcal{T}:S\neq 0\}} \frac{\langle D^*, S\rangle^2}{||S||^2}$$

Finally, by the Cauchy-Schwarz inequality ($\langle u, v\rangle^2 \leq ||u||^2||v||^2$), we know that $\frac{\langle D^*,S\rangle^2}{||S||^2}$ is maximized when S is collinear with the EIC D^*, therefore through the least favorable submodel, whose score is equal to the EIC, we can have

$$\frac{\langle D^*, S\rangle^2}{||S||^2} = \frac{\langle D^*, D^*\rangle^2}{||D^*||^2} = E[(D^*)^2] = \text{Var}(D^*)$$

This shows how TMLE is able to achieve the efficiency bound by taking the least favorable path.

Score and EIC of a classic example: ATE for a binary outcome variable

Define the clever covariate as $H(A, W)$ and update the outcome regression $\bar{Q}^0$ as follows:

$$\text{logit}(\bar{Q}(\epsilon)) = \text{logit}(\bar{Q}^0) + \epsilon H(A, W)$$

We have the log-likelihood for a binary outcome Y as

$$L(\epsilon) = Y \ln(\bar{Q}(\epsilon)) + (1 - Y) \ln(1 - \bar{Q}(\epsilon))$$

Let $\eta(\epsilon) = \text{logit}(\bar{Q}^0) + \epsilon H$ be the linear regressor.

The score can be rewritten through the chain rule as

$$\frac{dL}{d\epsilon} = \frac{\partial L}{\partial \bar{Q}} \cdot \frac{\partial \bar{Q}}{\partial \eta} \cdot \frac{\partial \eta}{\partial \epsilon}$$

Take the derivative of all 3 terms, then we have

$$\frac{\partial L}{\partial \bar{Q}} = \frac{Y}{\bar{Q}} - \frac{1 - Y}{1 - \bar{Q}} = \frac{Y - \bar{Q}}{\bar{Q}(1 - \bar{Q})}$$

$$\frac{\partial \bar{Q}}{\partial \eta} = \bar{Q}(1 - \bar{Q})$$

$$\frac{\partial \eta}{\partial \epsilon} = H(A, W)$$

And finally, we can write

$$\begin{aligned} \frac{dL}{d\epsilon} &= \left(\frac{Y - \bar{Q}}{\bar{Q}(1 - \bar{Q})} \right) \cdot \left(\bar{Q}(1 - \bar{Q}) \right) \cdot H(A, W) \\ &= (Y - \bar{Q}(\epsilon)) \cdot H(A, W) \end{aligned}$$

Recall that the EIC here is

$$\frac{1}{n} \sum_{i=1}^{n} D^*(O_i) = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\frac{I(A_i = 1)}{g(1|W_i)} - \frac{I(A_i = 0)}{g(0|W_i)} \right) (Y_i - \bar{Q}(A_i, W_i)) + \bar{Q}(1, W_i) - \bar{Q}(0, W_i) - \Psi_n \right]$$

The last 3 terms together after summing is 0 given the design of the estimator, then we can see that by letting the score equation $\frac{dL}{d\epsilon} = 0$, we also solve the efficient influence curve equation.

References

[1] van der Laan, M. J., & Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. Springer New York.