Strict overlap with high-dimensional covariates

Yue Wu

1 Introduction

Causal inference with observational studies relies on two key assumptions: unconfoundedness and overlap. And consider the case when more covariates are included in the analysis: unconfoundedness is usually more plausible because it is more likely that unmeasured confounding variables become measured, while overlap is more difficult to satisfy because the treatment assignment becomes more predictable given more covariates.

So this paper[1] mainly focuses on the implications of overlap assumption, as well as how the restriction becomes more stronger with high-dimensional covariates.

2 Overlap Assumptions

2.1 Overlap

Assumption 1 (Overlap)

$$0 < e(X) < 1$$

Overlap assumption is important both conceptually and operationally. Conceptually, if some units deterministically receive one treatment, then we have philosophic difficulty in thinking about the counterfactual potential outcomes of another treatment group. Operationally, if overlap is bad, the estimated propensity score can be very close to 0 or 1, which makes the estimators based on IPW numerically unstable.

The normal overlap assumption, together with unconfoundedness, and of course SUTVA, is sufficient for the non-parametric identification of the Average Treatment Effect.

2.2 Strict Overlap

But for efficient semiparametric estimation, a more strict version of overlap is needed, which assumes that the probability of a unit being assigned to either group should be bounded away from 0 and 1 with some constant η.

- **Assumption 2 (Strict Overlap)** *For some constant $\eta \in (0, 0.5)$,*

$$\eta \le e(X_{1:p}) \le 1 - \eta$$

- **Assumption 3 (Strict Overlap, Likelihood Ratio Form)**

$$P_0(X_{1:p} \in A) := P(X_{1:p} \in A \mid T = 0),\ P_1(X_{1:p} \in A) := P(X_{1:p} \in A \mid T = 1)$$

$$\pi := P(T = 1)$$

By Bayes' Theorem, strict overlap is equivalent to the bound on the density ratio between P_1 and P_0

$$b_{min} =: \frac{1-\pi}{\pi}\frac{\eta}{1-\eta} \leq \frac{dP_1(X_{1:p})}{dP_0(X_{1:p})} \leq \frac{1-\pi}{\pi}\frac{1-\eta}{\eta} := b_{max}$$

3 Implications of strict overlap

3.1 Bounded general discrepancies

Implications of bounded likelihood ratios are well-studied in information theory, in which the so-called f-divergence is an important discriminating information that describe the "dissimilarity" of two probability distributions[3]. It is defined in terms of a convex function f, such as KL-divergence and χ^2-divergence.

- f-divergence:

$$D_f(Q_1||Q_0) := E_{Q_0}[f(\frac{dQ_1}{dQ_0})]$$

  - KL-divergence: $f(x) = xlog(x)$
  - χ^2-divergence: $f(x) = (x-1)^2$

The result of information theory[3] shows that an upper bound on it can be derived when the likelihood ratio is bounded. So adapting this in the context of covariate distributions between treatment and control, we can derive explicit bound of f-divergence.

Theorem 1 *Strict overlap implies*

$$D_f(P_1(X_{1:p})||P_0(X_{1:p})) \leq \frac{b_{max}-1}{b_{max}-b_{min}}f(b_{min}) + \frac{1-b_{min}}{b_{max}-b_{min}}f(b_{max}) := B_{f(1||0)}$$

$$D_f(P_0(X_{1:p})||P_1(X_{1:p})) \leq \frac{b_{min}^{-1}-1}{b_{min}^{-1}-b_{max}^{-1}}f(b_{max}^{-1}) + \frac{1-b_{max}^{-1}}{b_{min}^{-1}-b_{max}^{-1}}f(b_{min}^{-1}) := B_{f(0||1)}$$

Theorem 1 means that, strict overlap restricts general discrepancy between the covariate distribution under treatment and control.

Note here that the upper bound B_f is free of p, while the left side may not, so that actually hints us why these bounds become more restrictive for larger values of p. We can take KL-divergence as an example, KL-divergence can be expanded into a summation of p terms[4], each of which corresponds to the discriminating information added by the new covariate X_k, beyond the information contained in the already considered covariates $X_{1:k-1}$.

$$KL(P_1(X_{1:p})||P_0(X_{1:p})) = \sum_{k=1}^{p} E_{p_1}\{KL(P_1(X_{(k)} \mid X_{1:k-1})||P_0(X_{(k)} \mid X_{1:k-1}))\}$$

Therefore, in the large p limit, strict overlap implies 1.1, the average unique discriminating information contained in each covariate converges to zero.

Corollary 1.1

$$p^{-1}\sum_{k=1}^{p} E_{p_1}\{KL(P_1(X_{(k)} \mid X_{1:k-1})||P_0(X_{(k)} \mid X_{1:k-1}))\} = O(p^{-1})$$

3.2 Bounded mean discrepancy

A concrete implication of strict overlap can be derived as Theorem 2, by adapting the χ^2-function as the choice of f.

$$\mu_{0,1:p} := (\mu_{0,(1)}, ..., \mu_{0,(p)}) := E_{P_0}(X_{1:p}),\ \Sigma_{0,1:p} := var_{P_0}(X_{1:p})$$
$$\mu_{1,1:p} := (\mu_{1,(1)}, ..., \mu_{1,(p)}) := E_{P_1}(X_{1:p}),\ \Sigma_{1,1:p} := var_{P_1}(X_{1:p})$$

Theorem 2 *Strict overlap implies*

$$p^{-1}\sum_{k=1}^{p}|\mu_{0,(k)} - \mu_{1,(k)}| \leq p^{-\frac{1}{2}}min\{||\Sigma_{0,1:p}||_{op}^{1/2} \cdot B_{\chi^2(1||0)}^{1/2},\ ||\Sigma_{1,1:p}||_{op}^{1/2} \cdot B_{\chi^2(0||1)}^{1/2}\}$$

$B_{\chi^2(1||0)}$ and $B_{\chi^2(0||1)}$ are free of p, and $||\Sigma_{1:p}||_{op}$ equals the variance of the variance-maximizing one-dimensional linear projection of $X_{1:p}$, which can be seen as a proxy for the degree to which the covariates are correlated.

If the covariates are not too correlated, so that operator norm grows more slowly than p, the bound converges to zero, then strict overlap implies that the covariate means are, on average, arbitrarily close to balance, which is a strong requirement in observational studies with many covariates.

4 Strict overlap and trimming

Theorem 3 *Let* $\tilde{\phi}(X_{1:p}) = \mathbf{I}\{e(X_{1:p}) \geq 0.5\}$ *be the Bayes optimal classifier, for an overlap bound* $\tilde{\eta} \in (0, 1/2)$*, we have*

$$P(\tilde{\eta} \leq e(X_{1:p}) \leq 1 - \tilde{\eta}) \leq P(\tilde{\phi}(X_{1:p}) \neq T)/\tilde{\eta}$$

When large covariate sets enable more accurately classification, the probability that a unit has an acceptable propensity score becomes small. In this case, a trimming procedure must throw away a large proportion of the sample.

5 Summary

- When we assume strict overlap, we actually make restriction on the global discrepancies between the covariate distributions in the treated and control populations('overlap' and 'balance'), and these bounds become more restrictive as the dimension grows large, which reminds us that the overlap assumptions should be carefully considered when adjusting for rich covariates.
- f-divergence is an information that describe the "dissimilarity" of two probability distributions, and an upper bound of it can be derived when the likelihood ratio is bounded.
- By regarding the covariates $X_{1:p}$ as a subset of a stochastic process, we can conduct the population-level analysis which is independent of sample size.
- The operator norm of a covariance matrix can be seen as a proxy for the degree to which the covariates are correlated.

References

[1] D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J.S. (2017). Overlap in observational studies with high-dimensional covariates. Journal of Econometrics.

[2] Ding, P. (2024) A first course in causal inference. Chapman & Hall

[3] Rukhin, A.L., 1997. Information-type divergence when the likelihood ratios are bounded. Appl. Math. 24 (4), 415–423.

[4] Cover, T.M., Thomas, J.A., 2005. Entropy, relative entropy, and mutual information. In: Elements of Information Theory, no. x. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 13–55, chap. 2.