Doubly robust machine learning

Michal Kolesár

ECO539B, Spring 2025

April 22, 2025

Standard semiparametrics

Debiased GMM

- A plethora of causal/structural parameters in economics can be expressed via semiparametric moment conditions $E[g(Z, \gamma_0, \theta_0)] = 0$, where γ_0 is a nuisance function (conditional mean, quantile, density etc), and θ_0 parameter of interest.
- Make two simplifications:
  1. Data is $Z = (Y, X)$, and $\gamma_0(X) = E[Y \mid X]$
  2. $g(Z, \gamma, \theta) = m(Z, \gamma) - \theta$, with $m(Z, \gamma) - m(Z, 0)$ linear in γ.

  m linear in γ allows for estimation based on doubly robust moments, rather than just locally robust/orthogonal moments (Chernozhukov et al. 2022), and simpler regularity conditions, but main ideas go through without these simplifications

$\theta_0 = E[m(Z, \gamma_0)]$ where $\gamma_0(X) = E[Y \mid X]$

- Running example: $X = (D, W)$, and $\theta_0 = E[\gamma_0(1, X)]$.
    - Under unconfoundedness, $\theta_0 = E[Y(1)]$.
    - Inference on the ATE just slightly more complicated, here $\theta_0 = E[\gamma_0(1, W) - \gamma_0(0, W)]$.
- Other examples: Average effect of shifting distribution of covariates

$\theta_0 = \int \gamma_0(x) dF_1(x) - E[\gamma_0(X)]$ (Stock 1989), average derivative $\theta_0 = E[\partial \gamma_0(D, X)/\partial D]$

…hundreds of papers applying general theory to different settings

Plug-in approach

- Simplest approach: estimate γ_0 nonparametrically, and then set $\hat{\theta}_{PI} = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}(1, W_i)$.
- Then
$$\sqrt{n}(\hat{\theta}_{PI} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\gamma}(1, W_i) - \gamma_0(1, W_i)) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\gamma_0(1, W_i) - \theta_0).$$
Second term well-behaved and asymptotically normal (corresponds to an oracle), but first typically non-zero mean, since $\hat{\gamma}$ biased for γ
- Need bias to be $o_p(n^{-1/2})$, can be hard to ensure (typically check on case-by-case basis), and doesn't hold if e.g. $\hat{\gamma}$ is estimated via the lasso $\implies$ $\hat{\theta}_{PI}$ asymptotically biased.

Standard semiparametrics

Debiased GMM

Debiasing: Key insight!

- Notice that for any γ,

$$E[\gamma(1,W)] = E\left[\frac{D}{p(W)}\gamma(X)\right] \implies E[\gamma(1,W) - \gamma_0(1,W)] = E\left[\frac{D}{p(W)}(\gamma(X) - Y)\right].$$

  where $\pi(W) = E[D \mid W]$ is propensity score.

- Suggests estimating bias as $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i}{\pi(W_i)}(\hat{\gamma}(X_i) - Y_i)$
- If $\hat{\gamma}$ estimated in separate sample and $\pi(W_i)$ known, then subtracting off $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i}{\pi(W_i)}(\hat{\gamma}(X_i) - Y_i)$ would yield estimator of θ_0 that is unbiased conditional on $\hat{\gamma}$.

Debiased GMM

- Debiased generalized method of moments (GMM) subtracts off feasible bias estimate,

$$\hat{\theta}_{DB} = \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_{\ell(i)}(1, W_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{\alpha}_{\ell(i)}(X_i)(\hat{\gamma}_{\ell(i)}(1, W_i) - Y_i)$$

  where $\hat{\gamma}_{\ell(i)}$ estimated on sample excluding i and $\hat{\alpha}$ is some estimator of $\alpha(X) := D/\pi(W)$.
- To avoid estimating γ_0 n times, usually split sample into $L = 5$ or $L = 10$ folds, and $\ell(i)$ excludes fold that i belongs to.
- This cross-fitting avoids own observation bias analogous to the jackknife instrumental variables estimator (JIVE) in instrumental variables (IV) settings.

General theory

- Since $\theta(\gamma) = E[m(Z, \gamma)]$ is a linear functional, if the functional is continuous (i.e. $E[m(W, \gamma)] \leq C\|\gamma\| =: CE[\gamma(W)^2]$ for all $\gamma \in L^2$), by the Riesz representation theorem, there exists $\alpha_0 \in L^2$ such that for all $\gamma \in L^2$

$$E[m(Z, \gamma)] = E[\alpha_0(X)\gamma(X)] \tag{1}$$

  Called Riesz representer (RR). Existence of RR equivalent to semiparametric variance bound being finite (Hirshberg and Wager 2021; Newey 1994), i.e. $\sqrt{n}$-consistent estimation possible
- In our case, $\alpha(X) = D/\pi(W)$, general recipes for construction available (e.g. Chernozhukov et al. 2022; Chernozhukov et al. 2018)

- Debiased GMM equivalent to GMM based on moment condition

$$\theta_0 = E[\psi(Z, \gamma_0, \alpha_0)], \quad \psi(Z, \gamma, \alpha) = m(Z, \gamma) + \alpha(X)(Y - \gamma(X))$$

In our case $\psi(Z) = \gamma(1, W) + (Y - \gamma(X))D/\pi(W)$.

- Moment condition doubly robust

$$\begin{aligned} E[\psi(Z, \gamma, \alpha)] - \theta_0 &= E[m(Z, \gamma) - m(Z, \gamma_0) + \alpha(X)(Y - \gamma(X))] \\ &=_{(1)} E[m(Z, \gamma) - m(Z, \gamma_0) + \alpha(X)(\gamma_0(X) - \gamma(X))] \\ &=_{(2)} E[(\alpha(X) - \alpha_0(X))(\gamma_0(X) - \gamma(X))] \end{aligned}$$

where (1) uses definition of γ_0 and (2) uses definition of RR, $E[m(Z, \gamma)] = E[\alpha_0(X)\gamma(X)]$.

Doubly robust moments

- Idea of using doubly robust moments old (Robins, Rotnitzky, and Zhao 1994, 1995). Innovation of debiased GMM is to combine it with cross-fitting, which allows for weaker regularity conditions on estimator of γ that accommodate regularized estimators (lasso, random forests, neural nets etc).
- Price: need to estimate α as well as γ, and it could be that α is non-smooth and hard to estimate
- Existence of RR allows for multiple approaches to estimating θ_0: regression (plug-in), propensity score weighting, and doubly-robust moments:

$$\theta_0 = E[m(Z, \gamma_0)] = E[\alpha_0(X)Y] = E[\psi(Z, \gamma_0, \alpha_0)].$$

Large-sample properties

- Simple conditions for asymptotic normality of $\hat{\theta}_{DB} = \frac{1}{n}\sum_{i=1}^{n}[m(Z_i, \hat{\gamma}_{\ell(i)}) + \hat{\alpha}_{\ell(i)}(X_i)(Y_i - \hat{\gamma}_{\ell(i)}(W_i))]$, since

$$\sqrt{n}(\hat{\theta}_{DB} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(R_{1i} + R_{2i} - R_{3i}) + \frac{1}{\sqrt{n}}(\psi(Z_i) - \theta_0),$$

with

$$\begin{aligned} R_{1i} &= m(Z_i, \hat{\gamma}_{\ell(i)}) - m(Z_i, \gamma_0) - \alpha_0(X_i)(\hat{\gamma}_{\ell(i)}(X_i) - \gamma(X_i)) \\ R_{2i} &= (\hat{\alpha}_{\ell(i)}(X_i) - \alpha_0(X_i))(Y_i - \gamma_0(X_i)) \\ R_{3i} &= (\hat{\alpha}_{\ell(i)}(X_i) - \alpha_0(X_i))(\hat{\gamma}_{\ell(i)}(X_i) - \gamma(X_i)) \end{aligned}$$

Key advantage of orthogonality

- $\sqrt{n}(\hat{\theta}_{DB} - \theta_0) \xrightarrow{d} \mathcal{N}(0, E[\psi(Z_i)^2])$, provided remainder terms negligible.
- Now, R_{1i}, R_{2i} mean zero, with variances going to zero if:
    1. First-step estimators consistent: $\|\hat{\alpha} - \alpha\| + \|\hat{\gamma} - \gamma\| + \|m(Z, \hat{\gamma}) - m(Z, \gamma)\| = o_p(1)$
    2. $\max_i \operatorname{var}(Y_i \mid X_i) < \infty$ and $\max_i \alpha(X_i) < \infty$ (strong overlap!)
- Bias given by $\frac{1}{\sqrt{n}} \sum_i R_{3i} = \frac{1}{\sqrt{n}} \sum_i (\hat{\alpha}_{\ell(i)}(X_i) - \alpha_0(X_i))(\hat{\gamma}_{\ell(i)}(X_i) - \gamma(X_i)) \leq \sqrt{n}\|\hat{\alpha} - \alpha_0\|\|\hat{\gamma} - \gamma_0\|$, so need
    3. $\|\hat{\alpha} - \alpha_0\|\|\hat{\gamma} - \gamma_0\| = o_p(1)$: Only product of biases needs to be small! In contrast, with plug-in approach needed bias of $\hat{\gamma}$ small
- Last condition allows for "machine learning" estimators of γ_0 (and γ).

- Debiasing plug-in estimator can be done in many ways, see e.g., Laan and Rose (2018, 2018), Hirshberg and Wager (2021), and Athey, Imbens, and Wager (2018). I'm not clear on relative merits of different approaches.
- Cross-fitting not needed for estimating β in partially linear model $Y = D\beta + g(W) + U$ (Donald and Newey 1994)
    - This is why post-double lasso (Belloni, Chernozhukov, and Hansen 2014) doesn't need to cross-fit

References i

Athey, Susan, Guido W. Imbens, and Stefan Wager. 2018. "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions." *Journal of the Royal Statistical Society,* Series B (Statistical Methodology), 80, no. 4 (September): 597–623. https://doi.org/10.1111/rssb.12268.

Belloni, Alexandre, Victor Chernozhukov, and Christian B. Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *The Review of Economic Studies* 81, no. 2 (April): 608–650. https://doi.org/10.1093/restud/rdt044.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James M. Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21, no. 1 (February): C1–C68. https://doi.org/10.1111/ectj.12097.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. 2022. "Locally Robust Semiparametric Estimation." *Econometrica* 90, no. 4 (July): 1501–1535. https://doi.org/10.3982/ECTA16294.

Donald, Stephen G., and Whitney K. Newey. 1994. "Series Estimation of Semilinear Models." *Journal of Multivariate Analysis* 50, no. 1 (July): 3–40. https://doi.org/10.1006/jmva.1994.1032.

Hirshberg, David A., and Stefan Wager. 2021. "Augmented Minimax Linear Estimation." *The Annals of Statistics* 49, no. 6 (December): 3206–3227. https://doi.org/10.1214/21-AOS2080.

Laan, Mark J. van der, and Sherri Rose. 2018. *Targeted Learning in Data Science.* Springer Series in Statistics. Cham: Springer. https://doi.org/10.1007/978-3-319-65304-4.

Newey, Whitney K. 1994. "The Asymptotic Variance of Semiparametric Estimators." *Econometrica* 62, no. 6 (November): 1349–1382. https://doi.org/10.2307/2951752.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89, no. 427 (September): 846–866. https://doi.org/10.1080/01621459.1994.10476818.

———. 1995. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association* 90, no. 429 (March): 106–121. https://doi.org/10.1080/01621459.1995.10476493.

Stock, James H. 1989. "Nonparametric Policy Analysis." *Journal of the American Statistical Association* 84, no. 406 (June): 567–575. https://doi.org/10.1080/01621459.1989.10478805.