Counterbalance Matching for Causal Inference

JP & Dylan Small

June 23, 2025

Mathematical Justification

Problem Setup

Let

- Y_i: a scalar outcome of interest
- $X_i \in \mathbb{R}^p$: observed baseline covariates, which we balance by matching
- $U_i \in \mathbb{R}$: unmeasured confounder we hope to get information about
- $Z_i \in \{0, 1\}$: treatment indicator

and consider the following data-generating model:

$$q_i = \mathbb{P}(Z_i = 1 \mid X_i, U_i) = g(\alpha + X_i^T \beta + \gamma U_i),\ \beta \in \mathbb{R}^p$$

where g is the logistic function defined as:

$$g(x) = \frac{\exp(x)}{1 + \exp(x)}$$

Then define a linear outcome surface as:

$$\mathbb{E}[Y_i \mid X_i, U_i, Z_i] = \tau Z_i + X_i^T \psi + \theta U_i$$

As usual, we are interested in τ, the causal effect of Z on Y, which we can write as:

$$\tau = \mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]$$

However, whenever $\gamma \neq 0$ and $\theta \neq 0$, then U both predicts treatment Z and affects outcome Y, and we have our usual problem of unmeasured confounding. From here on, we will assume mean/fine balance in X, and define our fitted propensity score compactly as:

$$\pi_i = \mathbb{P}(Z_i = 1 \mid X_i) = g(\alpha + X_i^T \beta)$$

Defining Surprise Scores

Consider the propensity-residual score, defined as:

$$S_i = Z_i - \pi_i$$

We call this quantity a **surprise score** because it measures how much the treatment assignment deviates from the propensity score. Note that for control units $i \in \mathcal{C}$, S_i is always negative, and for treatment units $i \in \mathcal{T}$, S_i is always positive. Note that in the absence of unmeasured confounding, we have the following:

$$\begin{aligned}
\mathbb{E}[S_i \mid X_i, U_i = 0] &= \mathbb{E}[Z_i - \pi_i \mid X_i, U_i = 0] \\
&= \mathbb{E}[Z_i \mid X_i, U_i = 0] - \mathbb{E}[\pi_i \mid X_i, U_i = 0] \\
&= (q_i \mid U_i = 0) - \pi_i \\
&= g(\alpha + X_i^T \beta) - g(\alpha + X_i^T \beta) \\
&= 0
\end{aligned}$$

We express the absolute surprise as:

$$s_i = |S_i| = |Z_i - \pi_i|$$

which is always non-negative and measures the absolute deviation of the treatment assignment from the propensity score. Next we will prove that surprise score is a proxy for the unmeasured confounder U_i.

Surprise Scores as Proxies for Unmeasured Confounding

Consider the expected value of the surprise score, which we can write as:

$$\begin{aligned}
\mathbb{E}[S_i \mid X_i, U_i] &= \mathbb{E}[Z_i - \pi_i \mid X_i, U_i] \\
&= \mathbb{E}[Z_i \mid X_i, U_i] - \mathbb{E}[\pi_i \mid X_i, U_i]
\end{aligned}$$

Then since we know that Z_i is Bernoulli(q_i), we can write:

$$\begin{aligned}
\mathbb{E}[S_i \mid X_i, U_i] &= q_i - \pi_i \\
&= g(\alpha + X_i^T \beta + \gamma U_i) - g(\alpha + X_i^T \beta) \\
&= g(\eta_i) - g(t_i)
\end{aligned}$$

Now consider the Taylor expansion of $g(\eta_i)$ around t_i:

$$g(\eta_i) = g(t_i) + g'(t_i)(\eta_i - t_i) + \frac{1}{2} g''(\xi)(\eta_i - t_i)^2, \ \xi \in [t_i, \eta_i]$$

Then it follows that:

$$\begin{aligned}
\mathbb{E}[S_i \mid X_i, U_i] &= g(t_i) + g'(t_i)(\eta_i - t_i) + \frac{1}{2} g''(\xi)(\eta_i - t_i)^2 - g(t_i) \\
&= g'(t_i)(\eta_i - t_i) + \frac{1}{2} g''(\xi)(\eta_i - t_i)^2
\end{aligned}$$

Note that $\eta_i - t_i = \gamma U_i$, so we can write:

$$\mathbb{E}[S_i \mid X_i, U_i] = g'(t_i)\gamma U_i + \frac{1}{2}g''(\xi)(\gamma U_i)^2$$

Define the quadratic term in the expansion above as:

$$R_i = \frac{1}{2}g''(\xi)(\gamma U_i)^2$$

From here, we note the following for the logistic function:

$$0 \leq g'(x) \leq \frac{1}{4}, \quad |g''(x)| \leq \frac{1}{4}$$

Then it follows that

$$0 \leq R_i \leq \frac{1}{8}|\gamma U_i|^2$$

And when $|\gamma U_i| \leq 1$, we have:

$$\frac{R_i}{g'(t_i)|\gamma U_i|} \leq \frac{\frac{1}{8}|\gamma U_i|}{\frac{1}{4}|\gamma U_i|} = \frac{1}{2}$$

So the quadratic error term is at most 50% of the linear term in the practical range of γU_i, and we can write:

$$\mathbb{E}[S_i \mid X_i, U_i] = q_i - \pi_i \approx g'(t_i)\gamma U_i$$

Rearranging for U_i and taking absolute values, we have:

$$|U_i| \approx \frac{|q_i - \pi_i|}{|\gamma| g'(t_i)}$$

Since we know that $0 < g'(x) \leq \frac{1}{4}$ for the logistic function, it follows that $\frac{1}{g'(t_i)} \geq 4$. Therefore, we have a lower bound on the magnitude of the unmeasured confounder:

$$|U_i| \gtrsim \frac{4}{|\gamma|}|q_i - \pi_i|$$

This shows that the magnitude of U_i is proportional to the magnitude of the expected surprise. Hoeffding's inequality tells us that the observed surprise, $S_i = Z_i - \pi_i$, will be close to its expectation, $q_i - \pi_i$. Using S_i as a proxy for its expectation, we arrive at the main result:

$$\begin{aligned} |U_i| &\gtrsim \frac{4}{|\gamma|}|S_i| \\ &\gtrsim \frac{4}{|\gamma|}s_i \end{aligned}$$

Thus, the magnitude of the observable surprise score $s_i = |S_i|$ is a proxy for the magnitude of the unmeasured confounder $|U_i|$.

Then when we select a subset of units i with $s_i = |S_i| \geq \Lambda$, we have that:

$$|U_i| \gtrsim \frac{4\Lambda}{|\gamma|}$$

So selecting units with high surprise is approximately equivalent (up to a constant factor) to selecting units with large unmeasured confounder.

Effect on the Covariance that Drives Hidden Bias

Rubin's asymptotic linear bias formula for a matched-pair estimator (1973) is:

$$\text{Bias}(\widehat{\tau}_0) = \theta \text{Cov}(U, Z \mid X)$$

He proved that after balancing the covariates X in the matched-pair estimator, the only remaining covariance is between the unmeasured confounder U and the treatment indicator Z. We re-express this variance in pair language. Denote the within-pair distance in U for pair i as:

$$\Delta U_i = U_{i,\mathcal{T}} - U_{i,\mathcal{C}}$$

Then since we know that $Z_{i,\mathcal{T}} = 1$ and $Z_{i,\mathcal{C}} = 0$, we have that:

$$\text{Cov}(U, Z \mid X) = \frac{1}{2}\text{Cov}(\Delta U_i \mid X_i)$$

By proof that JP needs to add here, we have that

$$\text{Cov}(\Delta U_i \mid |S_i| \geq \Lambda)$$

is increasing in Λ. This means counter-balancing inflates the covariance which betrays hidden bias while preserving unbiasedness for observed X. So hidden bias is most detectable when we select focal pairs with high surprise.

Information-Theoretic Approach

Define the per-pair mutual information between treatment and the latent factor U given X as:

$$I(U; Z \mid X) = \mathbb{E}\left[\log \frac{\mathbb{P}(Z \mid U, X)}{\mathbb{P}(Z \mid X)}\right]$$

Then we have that since g is monotone and S is sufficient for treatment assignment given X, we have that:

$$I(U; Z \mid X, |S| \geq \Lambda)$$

is increasing in Λ. This means that when we counter-balance, we maximize the information about U we can get from studying the F focal pairs.

Our Algorithm

Inputs.

- Treated index set $\mathcal{T}$ and donor-pool controls $\mathcal{C}$.
- Covariate matrix X, fitted propensities π, surprise scores $S_i = Z_i - \pi_i$.
- Distance metric $d_{ij} = \|X_i - X_j\|_\Sigma^2$ (Mahalanobis or user-chosen).
- Balance constraints $\{B_k\}_{k=1}^K$ (mean balance, fine balance, calipers, ...).
- Hyper-parameters: $\mathbf{F}$ (number of focal pairs), Λ (surprise threshold).

Step A: Pre-screen the candidate pool. Retain only units whose absolute surprise exceeds a safety margin

$$|S_i| \;\geq\; \Lambda + \varepsilon_\delta, \quad \varepsilon_\delta = \sqrt{\tfrac{1}{2}\log\frac{2}{\delta}}$$

so that $|S_i|$ exceeds its expectation with probability at most δ (Hoeffding bound).

Step B: Find the tightest admissible similarity radius $\kappa^\star$. Repeat until convergence (*bi-section search*):

1. Guess κ; build the graph that links each $i \in \mathcal{T}$ to controls $j \in \mathcal{C}$ with $d_{ij} \leq \kappa$.
2. Ask a max-flow or assignment solver whether *(i)* all balance constraints B_k are feasible and *(ii)* at least F treated units in that graph satisfy $|S_i| \geq \Lambda$.
3. Shrink or enlarge κ accordingly.

The result is the smallest $\kappa^\star$ that still admits a feasible match with $\mathbf{F}$ high-surprise pairs. (This is Rosenbaum–Garfinkel *threshold search*.)

Step C: Penalised assignment for the global match. Create the augmented cost matrix

$$\tilde{d}_{ij} \;=\; \begin{cases} d_{ij}, & d_{ij} \leq \kappa^\star \text{ and } |S_i| \geq \Lambda; \\ d_{ij} + M, & \text{otherwise}, \end{cases} \qquad M \gg \max_{ij} d_{ij},$$

then solve the cardinality/assignment problem

$$\min_{\mu:\mathcal{T}\to\mathcal{C}} \sum_{i\in\mathcal{T}} \tilde{d}_{i\,\mu(i)} \quad \text{s.t. } B_k \;\; (k = 1, \ldots, K).$$

The penalty M forces the optimiser to use exactly the F admissible high-surprise pairs found in Step B while choosing the cheapest controls for all other treated units.

Step D: Outputs.

- *Full matched sample*: all treated units plus their chosen controls.
- *F focal pairs*: the F pairs that triggered the $|S| \geq \Lambda$ rule (highlighted for qualitative follow-up).
- Diagnostics: covariate balance table, distribution of d_{ij} inside and outside the focal subset, and Γ-sensitivity value.

Computation notes. The threshold search (Step B) runs in $O(\log R)$ iterations where $R = \max_{ij} d_{ij} - \min_{ij} d_{ij}$; each iteration calls a standard Hungarian or max-flow routine. In practice, with $n \lesssim 10^4$, the full pipeline completes in seconds on a laptop.

Tuning hints.

- Start with $F \approx 20$ and increase if qualitative capacity allows.
- Pick Λ so that $\kappa^\star$ falls near the 10th percentile of all distances— that keeps focal pairs both "close" and "surprising".
- Sensitivity-analysis software: `sensitivitymw` or `rbounds`.