Causal Inference by Minimizing the Dual Norm of Bias

Nathan Kallus

Cornell University and Cornell Tech

www.nathankallus.com

Matching Zoo

- It's a zoo of matching estimators for causal effects:
  - PSM, NN, CM, CEM, GenMatch, Mean-Matching
- What are the inherent differences?
- What does it mean to match on covariates?

- This talk:
  - New classification: worst-case bias minimizing (WCBM)
  - Encompasses many existing methods
  - Reveals motivation as ***optimality*** for particular ***structure***
  - Gives rise to new ***kernel matching*** estimators

Set-up

- tl;dr: measure SATT under unconfoundedness using a covariate-matched control sample
- Subjects $i = 1, \ldots, n$:
  - Two treatments: "treatment" (T_i=1) & "control" (T_i=0)
    - $\mathcal{T}_0 = \{i : T_i = 0\}$, $\mathcal{T}_1 = \{i : T_i = 1\}$
  - Observe: covariates X_i, treatment T_i, and outcome Y_i=$Y_i(T_i)$
    - Unseen counterfactual potential outcomes $Y_i(0)$, $Y_i(1)$
    - X=(X_1,…, X_n), T=(T_1,…, T_n) are the whole sample
- Assume unconfounded:

$$\mathbb{E}[Y_i(t) \mid T_i, X_i] = \mathbb{E}[Y_i(t) \mid X_i], \ \mathbb{P}(T_i = t \mid X_i) > 0$$

- Want to measure SATT:

$$\text{SATT} = \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} (Y_i(1) - Y_i(0))$$

Set-up

- Estimate by making a *matched control sample*

$$\hat{\tau}_W = \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} Y_i - \sum_{i \in \mathcal{T}_0} W_i Y_i$$

$$W_i \geq 0 \qquad \sum_{i \in \mathcal{T}_0} W_i = 1$$

- "Honest" weights $W(T, X)$ that only depend on T, X
  - No causal effect mining allowed!
- Weight types:
  - General weights $\mathcal{W}_0^{\text{general}} = \{W_{\mathcal{T}_0} \in \mathbb{R}_+^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$
  - Matched subset w/o rep:
    $\mathcal{W}_0^{\text{w/o rep.}} = \{W_{\mathcal{T}_0} \in \{0, 1/n_0'\}^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$
  - Matched multi-subset w/ rep:
    $\mathcal{W}_0^{\text{w/ rep.}} = \{W_{\mathcal{T}_0} \in \{0, 1/n_0', \dots\}^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$

Decomposing Bias

- Define "bias" (misnomer – more like error)

$$\text{bias} = \mathbb{E}\left[\hat{\tau}_W - \text{SATT} \middle| X, T\right]$$

- Let $f_0(x) = \mathbb{E}[Y_i(0) \mid X_i = x],\ \epsilon_i = Y_i(0) - f_0(X_i)$

Theorem: $\hat{\tau}_W - \text{SATT} = B(W; f_0) + E(W)$

$$B(W; f) := \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} f(X_i) - \sum_{i \in \mathcal{T}_0} W_i f(X_i)$$

$$E(W) := \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} \epsilon_i - \sum_{i \in \mathcal{T}_0} W_i \epsilon_i$$

And, **under unconfoundedness,**

$$\mathbb{E}[E(W) \mid X, T] = 0$$

$$\mathbb{E}[\hat{\tau}_W - \text{SATT} \mid X, T] = B(W; f_0)$$

Worst-Case Bias

- Under unconfoundedness, our bias is $B(W; f_0)$
- Involves an *unknown* function f_0
- Consider guarding against any possible such function
  - $B(W; f)$ scales linearly in f
  - Consider worst-case relative to a magnitude $\|f\| \in [0, \infty]$
  - I.e., minimize

$$\max_f \frac{|B(W; f)|}{\|f\|}$$

Worst-Case Bias

- For worst-case bias to be well-defined, assume:
  - $\|\cdot\|$ is a semi-norm on $\mathcal{F} = \{f : \|f\| < \infty\}$
    - (implies $\mathcal{F}$ is a linear subspace)
  - $\mathcal{F}/\{f : \|f\| = 0\}$ forms a Banach space
    - (complete normed vector space)
  - Contrasts $f \mapsto B(W; f)$ are continuous maps for any W
    - (equivalently, $\exists M(W) : B(W; f) \leq M(W)\,\|f\|$ since continuous operators on B space = bounded operators)
- The dual space $\mathcal{F}^*$ = {continuous linear operators}
- Is a Banach space with norm $\|A\|_* = \sup_{\|f\| \leq 1} A(f)$

Dual Norm of Bias

- Dual norm of bias is the normalized worst-case bias:

$$\begin{aligned}\mathfrak{B}(W;\mathcal{F}) &= \max_f \frac{B(W;f)}{\|f\|} \\ &= \max_{\|f\|\leq 1} B(W;f) \\ &= \|B(W;\cdot)\|_*\end{aligned}$$

Definition 1. A matching method $W(T,X)$ is said to be *worst-case bias minimizing (WCBM)* if for some $\mathcal{W}$ and $\|\cdot\|$ satisfying assumptions we have

$$W(T,X) \in \arg\min_{W\in\mathcal{W}} \mathfrak{B}(W;\mathcal{F}) \neq \mathcal{W}.$$

Existing Methods as WCBM

- Surprising fact:
  most covariate-matching methods are WCBM!

- $\Rightarrow$ Reveals structural motivations of different matching methods
  - Choose the method that matches your structural beliefs
- $\Rightarrow$ WCBM is the *right* framework

Nearest-Neighbor Matching

- NNM: Find a control match for each treated unit and minimize the sum of pairwise distances per $\delta(x, x')$
  - Can be with or without replacement
  - Hansen, 2004 & 2006; Rubin, 1973; Cochran, 1953
  - Classically, not necessarily minimal sum of distances
- Usually, Mahalanobis

$$\delta(x, x') = \sqrt{(x - x')\hat{\Sigma}^{-1}(x - x'}$$

Nearest-Neighbor Matching

Theorem: Nearest neighbor matching wrt $\delta(x, x')$ **with** replacement is WCBM with

$$\|f\| = \sup_{x \neq x'} \frac{f(x) - f(x')}{\delta(x, x')}$$

and either $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in \mathbb{R}_+^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$

or $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in \{0, 1/n_1, \dots\}^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$

Nearest-Neighbor Matching

Theorem: Nearest neighbor matching wrt $\delta(x, x')$ **without** replacement is WCBM with

$$\|f\| = \sup_{x \neq x'} \frac{f(x) - f(x')}{\delta(x, x')}$$

and either $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in [0, 1/n_1]^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$

or $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in \{0, 1/n_1\}^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$

Caliper Matching

- CM: find smallest caliper size and pairs such that all pairwise distance can fit within the caliper
  - Raynor, 1983; Cochran & Rubin, 1973
  - Classically, not necessarily optimal caliper
  - When with replacement, (almost) same as NNM

Theorem: Caliper matching wrt $\delta(x, x')$ **without** replacement is WCBM with $\|f\| = \|f\|_{\partial(\hat{\mu}_n, \delta)}$ where

$$\|f\|_{\partial(\mu,\delta)} = \mathbb{E}_{\mu\otimes\mu}\left[\frac{(f(x)-f(x'))}{\delta(x,x')} \mid x \neq x'\right]$$ and $\hat{\mu}_n$ is the EDF

And either $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in [0, 1/n_1]^{\mathcal{T}_0} : \sum_{i\in\mathcal{T}_0} W_i = 1\}$

or $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in \{0, 1/n_1\}^{\mathcal{T}_0} : \sum_{i\in\mathcal{T}_0} W_i = 1\}$

Coarsened Exact Matching

- CEM: match exactly within each stratum, as defined by a *coarsening function* $C : \mathcal{X} \to \{1, \ldots, M\}$
  - E.g., if there are 5 treated subjects and 3 control subjects in a given stratum then each of the control subjects is given weight proportional to 5/3 (weights sum to one)
  - Iacus et al., 2011

Theorem: CEM with coarsening fn C is WCBM with

$$\|f\| = \begin{cases} \sup_{x \in \mathcal{X}} |f(x)| & \left|f^{-1}(C^{-1}(j))\right| = 1 \ \forall j, \\ \infty & \text{otherwise,} \end{cases}$$

and $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in \mathbb{R}_+^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$

Mean Matching

- MM: subsample the control population to have similar sample mean to treated population wrt

$$M_V(W) = \left\| V^{-1/2} \left(\frac{1}{n_1} \sum_{i \in \mathcal{T}_1} X_i - \sum_{i \in \mathcal{T}_0} W_i X_i \right) \right\|_2$$

  - Rubin, 2012; Rubin, 1973; Greenberg, 1953
  - Classically, not necessarily optimal

Theorem: Mean matching is WCBM with

$$\|f\|^2 = \begin{cases} \beta^T V \beta + \beta_0^2 & f(x) = \beta_0 + \beta^T x, \\ \infty & \text{otherwise.} \end{cases}$$

and $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in \{0, 1/n_1, \dots\}^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$ (w/ repl)
or $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in \{0, 1/n_1\}^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$ (w/o repl)

Kernel Matching

- Most matching methods are WCBM
- Each corresponded to particular structure / functional space
- What about other spaces?
- In ML, reproducing kernel Hilbert spaces (RKHS) are very common for *generalizing* learned function
  - E.g. kernelized SVM, kernel ridge regression, kernel PCA, …
- Via WCBM, kernels can be used for matching too!

Reproducing Kernel Hilbert Space

- HS = inner product space that is a Banach space
- RKHS = HS with continuous evaluations
- By Riesz representation theorem,
  PSD kernel $\mathcal{K}(x, x') \longleftrightarrow$ RKHS
  - Polynomial kernel $\mathcal{K}_s(x, x') = (1 + x^T x'/s)^s$
    - Spans polynomials deg ≤ s (finite-dim)
  - Exponential kernel $\mathcal{K}(x, x') = e^{x^T x'}$
    - Infinite dimensional **C_0-universal**
  - Gaussian kernel $\mathcal{K}_s(x, x') = e^{-s^2||x-x'||^2}$
    - Infinite dimensional **C_0-universal**

Kernel Matching

- Kernel Gram matrix $K_{ij} = \mathcal{K}(X_i, X_j)$

Theorem:

$$\mathfrak{B}^2(W;\mathcal{F}) = \frac{1}{n_1^2} e_{n_1}^T K_{\mathcal{T}_1,\mathcal{T}_1} e_{n_1} + W_{\mathcal{T}_0}^T K_{\mathcal{T}_0\mathcal{T}_0} W_{\mathcal{T}_0} - \frac{2}{n_1} e_{n_1}^T K_{\mathcal{T}_1,\mathcal{T}_0} W_{\mathcal{T}_0}$$

- Minimize over different domains
  → different matching methods

Kernel Matching

- General weight kernel matching

  $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in \mathbb{R}_+^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$

- Discrete kernel matching with replacement

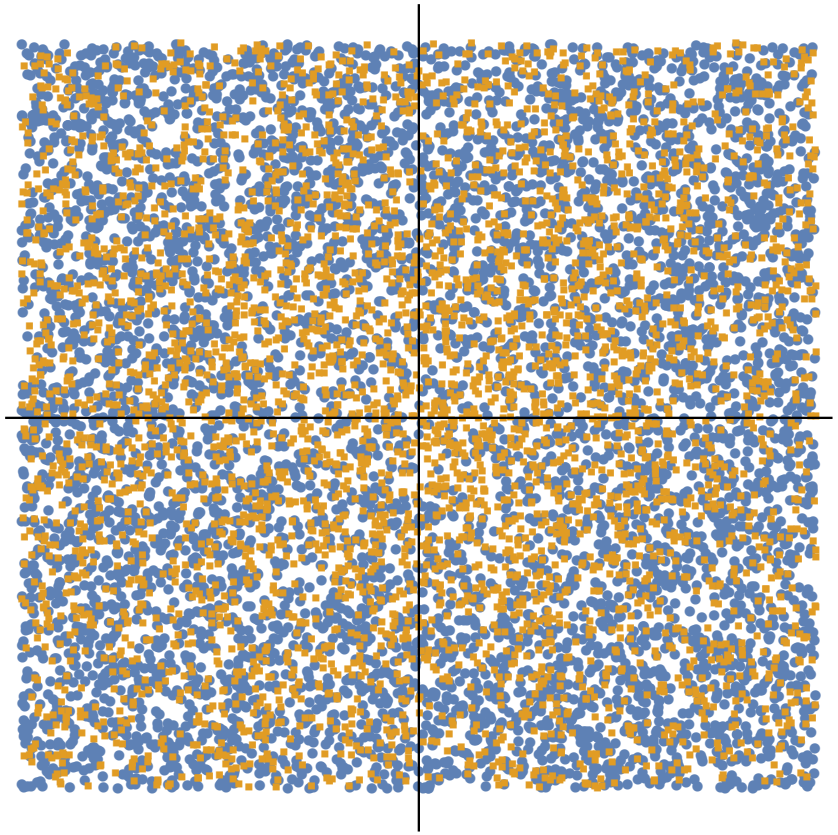
  $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in \{0, 1/n_1, \dots\}^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$

- Discrete kernel matching without replacement

  $\mathcal{W}_0 = \{W_{\mathcal{T}_0} \in \{0, 1/n_1\}^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} W_i = 1\}$

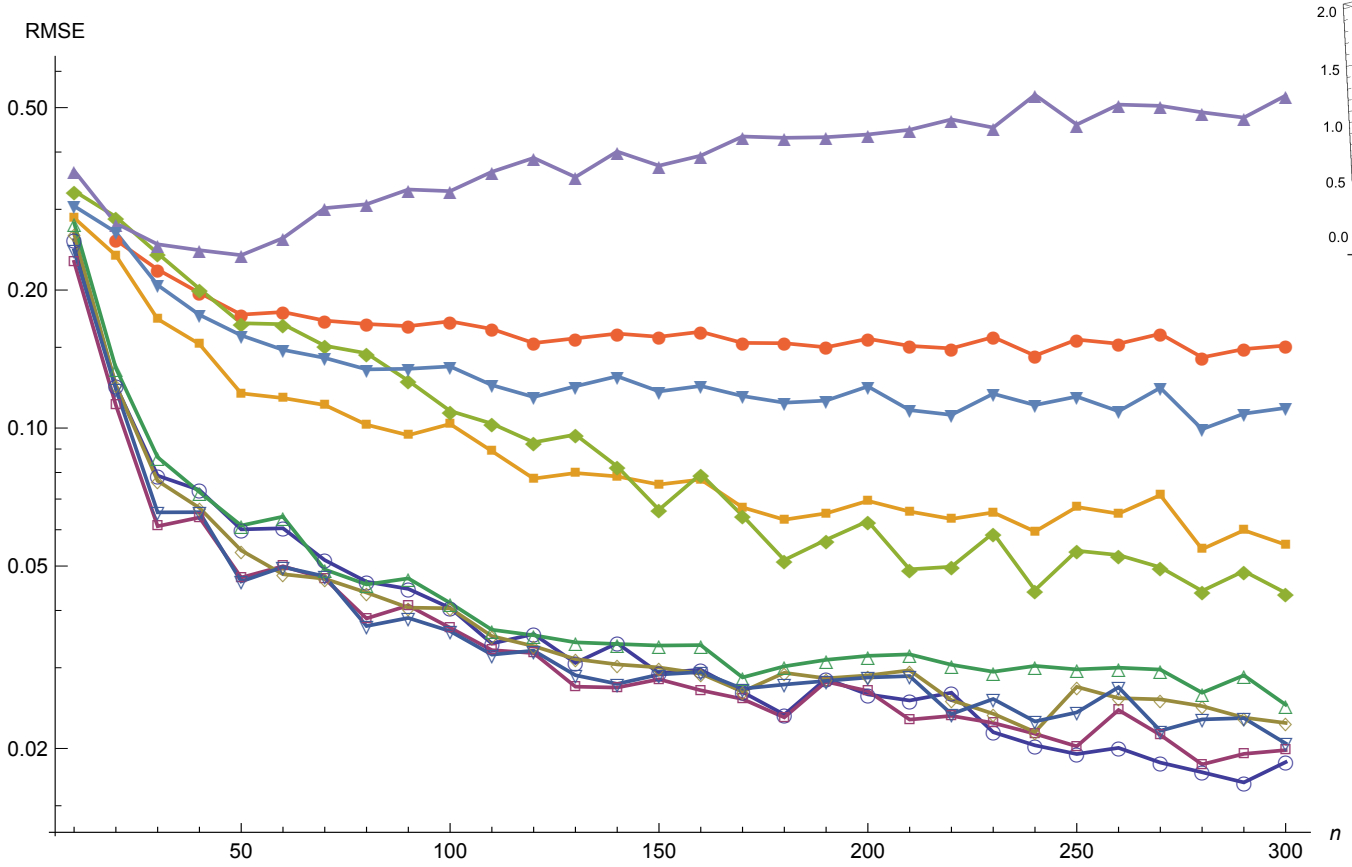
Numerics

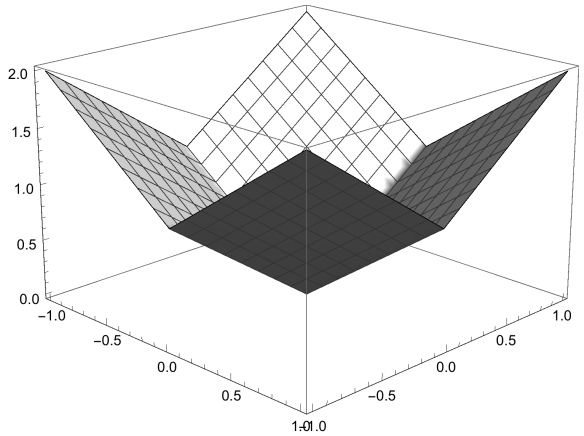
- Hypothetical observational study
- $X_i \in \mathbb{R}^2$ distributed uniformly on [-1, 1]^2
- $T_i \sim \text{Bernoulli}(0.8/(1 + \sqrt{2}\,\|X_i\|_2))$
- $Y_i(0) = f_0(X_i) + \epsilon_i$
- $\epsilon_i \sim \mathcal{N}(0, 0.1)$
- Various forms for f_0
- Measure RMSE



Numerics: L1 norm

$$f_0(x) = |x_1| + |x_2|$$

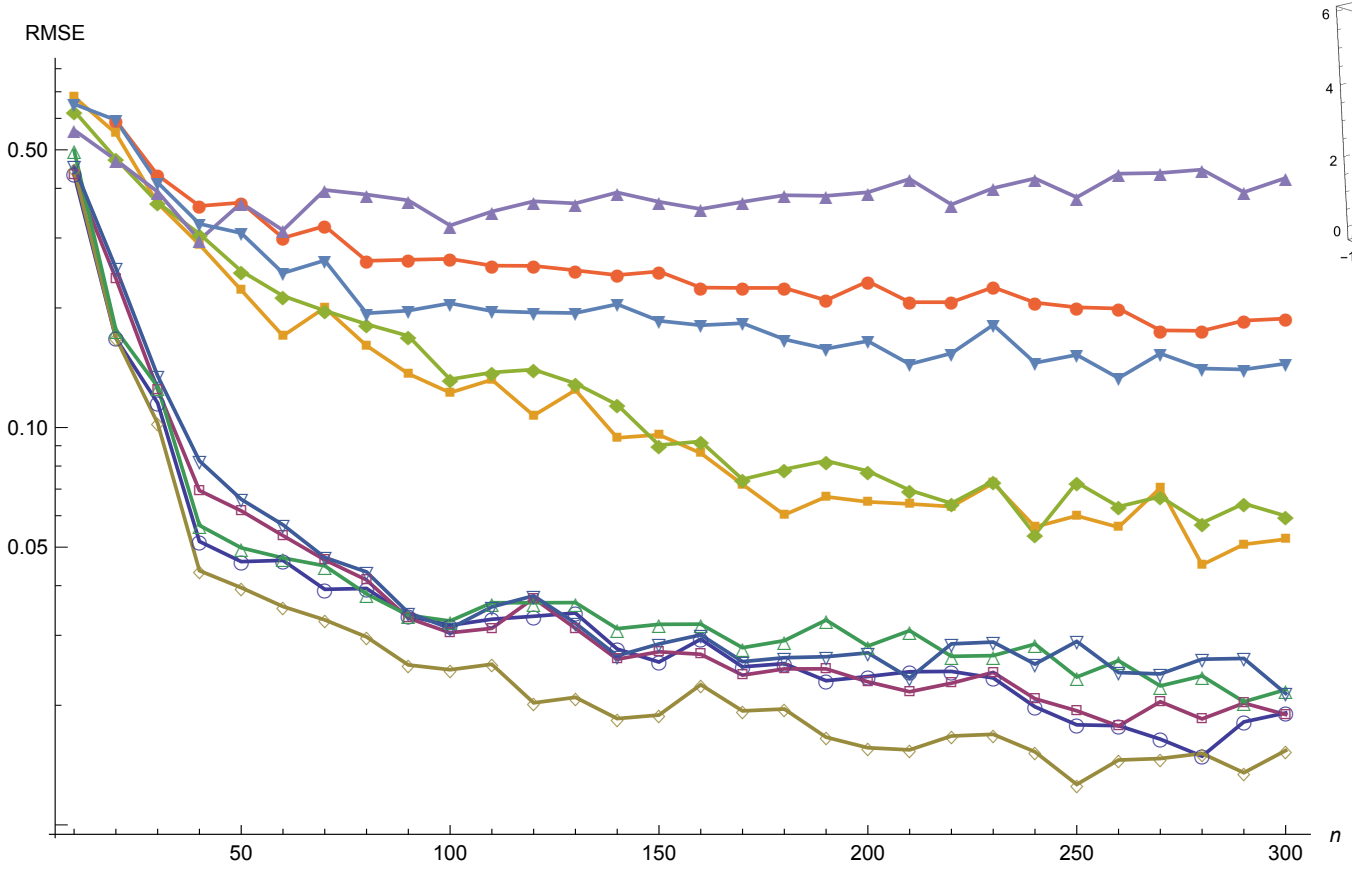


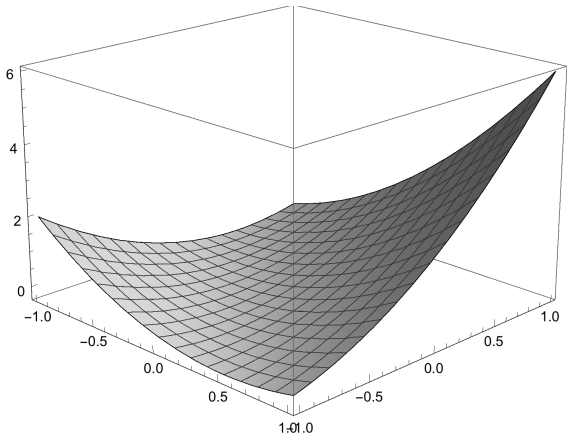


No matching
One−to−one
CEM
Mahal. means
PSM
Quad kernel weight
Exp kernel weight
Gauss kernel weight
Exp kernel match
Gauss kernel match

Numerics: quadratic

$$f_0(x) = (x_1 + x_2) + (x_1 + x_2)^2$$

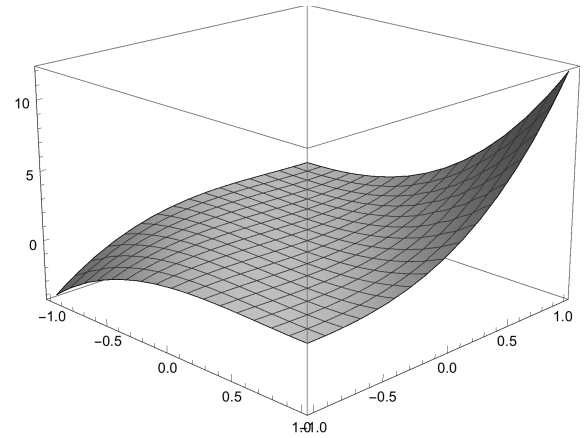


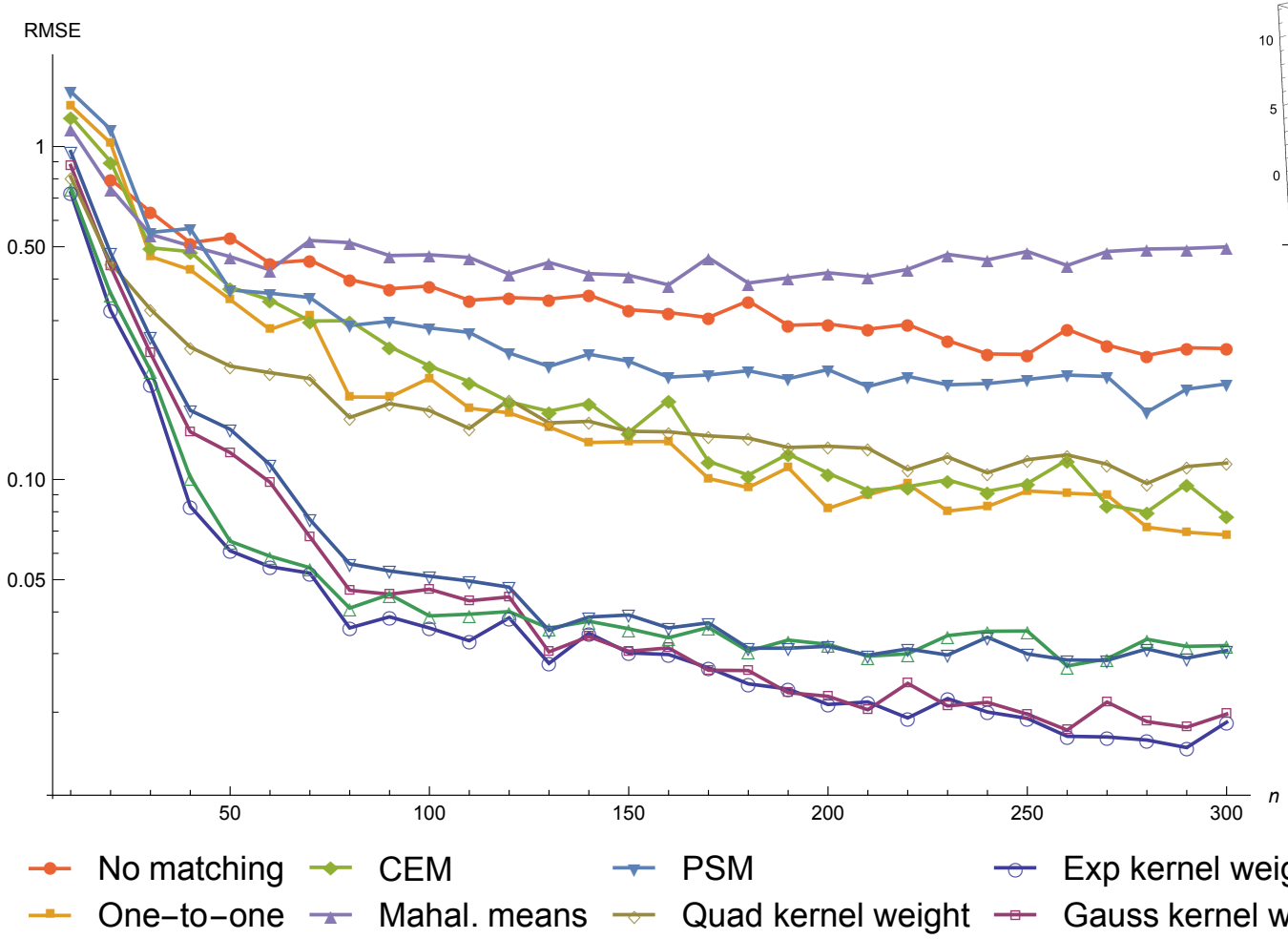


No matching
One−to−one
CEM
Mahal. means
PSM
Quad kernel weight
Exp kernel weight
Gauss kernel weight
Exp kernel match
Gauss kernel match

Numerics: cubic

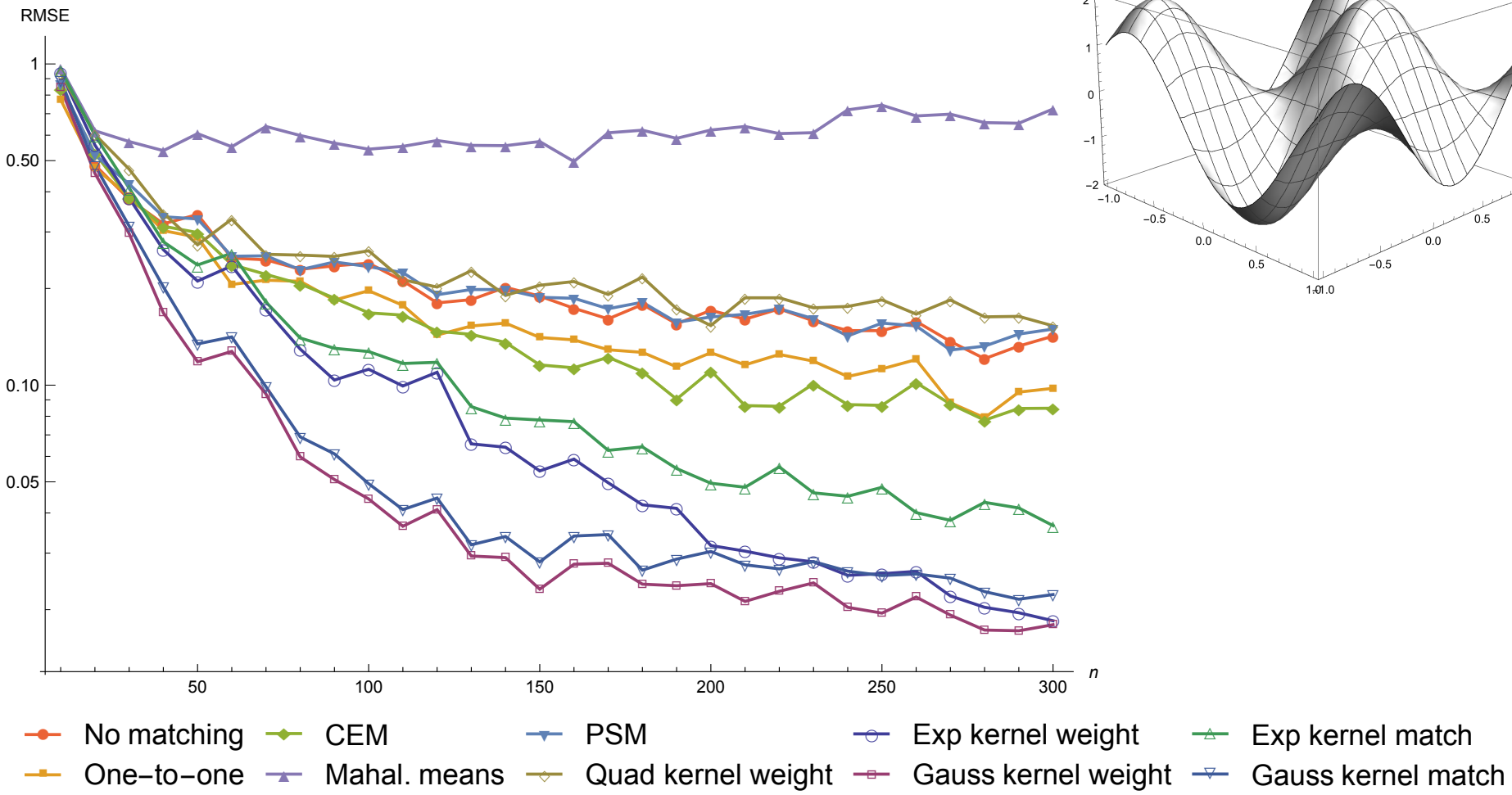
$$f_0(x) = (x_1 + x_2)^2 + (x_1 + x_2)^3$$





Numerics: sinusoidal

$$f_0(x) = \sin(\pi(x_1 + x_2)) + \cos(\pi(x_1 - x_2))$$



… huh?

- WCBM offers a general framework for matching estimators
- Structure $\longrightarrow$ imbalance metric and matching methods that minimize imbalance
- This recovers existing matching methods and uncovers structural underpinnings
- New methods: ***kernel matching***