Quantitative Social Science Methods, I, Lecture Notes: Detecting and Reducing Model Dependence in Causal Inference

Gary King
Institute for Quantitative Social Science
Harvard University

August 17, 2020

1GaryKing.org
Detecting Model Dependence

Matching to Reduce Model Dependence

Three Matching Methods

Problems with Propensity Score Matching

The Matching Frontier
Readings in Model Dependence


• Related Software: WhatIf, MatchIt, Zelig, CEM j.mp/causalinference
Readings in Model Dependence

Readings in Model Dependence


Readings in Model Dependence

- Related Software: WhatIf, MatchIt, Zelig, CEM
Readings in Model Dependence

- Related Software: WhatIf, MatchIt, Zelig, CEM

[link](https://j.mp/causalinference)
Counterfactuals

• Three types:
  1. Forecasts: What will the mortality rate be in 2025?
  2. What if Questions: What would have happened if the U.S. had not invaded Iraq?
  3. Causal Effects: What is the causal effect of the Iraq war on World GDP? (a factual minus a counterfactual)

• Counterfactuals are part of most social science research.
Counterfactuals

- Three types:

  1. Forecasts: What will the mortality rate be in 2025?
  2. What if Questions: What would have happened if the U.S. had not invaded Iraq?
  3. Causal Effects: What is the causal effect of the Iraq war on World GDP? (a factual minus a counterfactual)

- Counterfactuals are part of most social science research.
Counterfactuals

• Three types:
  1. **Forecasts** What will the mortality rate be in 2025?
  2. **What-if Questions** What would have happened if the U.S. had not invaded Iraq?
  3. **Causal Effects** What is the causal effect of the Iraq war on World GDP? (a factual minus a counterfactual)
Counterfactuals

• Three types:
  1. Forecasts What will the mortality rate be in 2025?
  2. Whatif Questions What would have happened if the U.S. had not invaded Iraq?
Counterfactuals

• Three types:
  1. Forecasts What will the mortality rate be in 2025?
  2. Whatif Questions What would have happened if the U.S. had not invaded Iraq?
  3. Causal Effects What is the causal effect of the Iraq war on World GDP? (a factual minus a counterfactual)
Counterfactuals

- Three types:
  1. **Forecasts** What will the mortality rate be in 2025?
  2. **Whatif Questions** What would have happened if the U.S. had not invaded Iraq?
  3. **Causal Effects** What is the causal effect of the Iraq war on World GDP? (a factual minus a counterfactual)

- Counterfactuals are part of most social science research
Which model would you choose? (Both fit the data well.)

• Compare prediction at $x = 1.5$ to prediction at $x = 5$

• How do you choose a model?
  
  $R^2$?
  
  Some "test"?
  
  "Theory"?

• The bottom line: answers to some questions don't exist in the data. We show how to determine which ones.

• Same for what if questions, predictions, and causal inferences
Which model would you choose? (Both fit the data well.)

- Compare prediction at $x = 1.5$ to prediction at $x = 5$
Which model would you choose? (Both fit the data well.)

• Compare prediction at $x = 1.5$ to prediction at $x = 5$
• How do you choose a model?
Which model would you choose? (Both fit the data well.)

- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- How do you choose a model? $R^2$?
Which model would you choose? (Both fit the data well.)

- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- How do you choose a model? $R^2$? Some “test”? 

![Graph showing comparisons between different models]

Detecting Model Dependence
Which model would you choose? (Both fit the data well.)

- Compare prediction at $x = 1.5$ to prediction at $x = 5$
Which model would you choose? (Both fit the data well.)

- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- The bottom line: answers to some questions don’t exist in the data. We show how to determine which ones.
Which model would you choose? (Both fit the data well.)

- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- The bottom line: answers to some questions don’t exist in the data. We show how to determine which ones.
- Same for what if questions, predictions, and causal inferences
Model Dependence Proof

To estimate $E(Y | X = x)$ at $x$, average many observed $Y$ with value $x$

Assumptions (Model-Based Inference)

1. Definition: model dependence at $x$ is the difference between predicted outcomes for any two models that fit about equally well.

2. The functional form follows strong continuity (think smoothness, although it is less restrictive)

Result

The maximum degree of model dependence: a function of the distance from the counterfactual to the data
Model Dependence Proof

Model Free Inference

To estimate $E(Y|X=x)$ at $x$, average many observed $Y$ with value $x$.

Assumptions (Model-Based Inference)

1. Definition: model dependence at $x$ is the difference between predicted outcomes for any two models that fit about equally well.

2. The functional form follows strong continuity (think smoothness, although it is less restrictive).

Result

The maximum degree of model dependence: a function of the distance from the counterfactual to the data.
Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at $x$, average many observed $Y$ with value $x$.
### Model Dependence Proof

**Model Free Inference**

To estimate $E(Y|X = x)$ at $x$, average many observed $Y$ with value $x$

**Assumptions (Model-Based Inference)**
Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at $x$, average many observed $Y$ with value $x$

Assumptions (Model-Based Inference)

1. Definition: model dependence at $x$ is the difference between predicted outcomes for any two models that fit about equally well.
Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at $x$, average many observed $Y$ with value $x$

Assumptions (Model-Based Inference)

1. Definition: model dependence at $x$ is the difference between predicted outcomes for any two models that fit about equally well.

2. The functional form follows strong continuity (think smoothness, although it is less restrictive)
Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at $x$, average many observed $Y$ with value $x$.

Assumptions (Model-Based Inference)

1. Definition: model dependence at $x$ is the difference between predicted outcomes for any two models that fit about equally well.

2. The functional form follows strong continuity (think smoothness, although it is less restrictive).

Result
Model Dependence Proof

Model Free Inference
To estimate $E(Y|X = x)$ at $x$, average many observed $Y$ with value $x$

Assumptions (Model-Based Inference)
1. Definition: model dependence at $x$ is the difference between predicted outcomes for any two models that fit about equally well.
2. The functional form follows strong continuity (think smoothness, although it is less restrictive)

Result
The maximum degree of model dependence: a function of the distance from the counterfactual to the data
A Simple Measure of Distance from The Data

Figure: The Convex Hull
A Simple Measure of Distance from The Data

Figure: The Convex Hull

- **Interpolation**: Inside the convex hull
A Simple Measure of Distance from The Data

Figure: The Convex Hull

- **Interpolation**: Inside the convex hull
- **Extrapolation**: Outside the convex hull
A Simple Measure of Distance from The Data

- **Interpolation**: Inside the convex hull
- **Extrapolation**: Outside the convex hull
- Works mathematically for any number of $X$ variables

Figure: The Convex Hull
A Simple Measure of Distance from The Data

- **Interpolation**: Inside the convex hull
- **Extrapolation**: Outside the convex hull
- Works mathematically for any number of $X$ variables
- Software to determine whether a point is in the hull (which is all we need) without calculating the hull (which would take forever), so its fast; see [GaryKing.org/what if](http://GaryKing.org/what if)

Figure: The Convex Hull
Model Dependence Example

Data:
124 Post-World War II civil wars

Dependent var:
peacebuilding success

Treatment:
multilateral UN peacekeeping intervention (0/1)

Control vars:
war type, severity, duration; development status, …

Counterfactual question:
Switch UN intervention for each war

Data analysis:
Logit model

The question:
How model dependent are the results?

Percent of counterfactuals in the convex hull:
0%

↝
without estimating any models, we know: inferences will be model dependent

For illustration: let’s find an example…. 
Model Dependence Example
Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)
Model Dependence Example

Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

- **Data:** 124 Post-World War II civil wars
Model Dependence Example
Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

- **Data**: 124 Post-World War II civil wars
- **Dependent var**: peacebuilding success
Model Dependence Example
Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

- **Data:** 124 Post-World War II civil wars
- **Dependent var:** peacebuilding success
- **Treatment:** multilateral UN peacekeeping intervention (0/1)
Model Dependence Example
Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

- **Data:** 124 Post-World War II civil wars
- **Dependent var:** peacebuilding success
- **Treatment:** multilateral UN peacekeeping intervention (0/1)
- **Control vars:** war type, severity, duration; development status,...
Model Dependence Example

Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

- **Data:** 124 Post-World War II civil wars
- **Dependent var:** peacebuilding success
- **Treatment:** multilateral UN peacekeeping intervention (0/1)
- **Control vars:** war type, severity, duration; development status,…
- **Counterfactual question:** Switch UN intervention for each war
Model Dependence Example
Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

- **Data:** 124 Post-World War II civil wars
- **Dependent var:** peacebuilding success
- **Treatment:** multilateral UN peacekeeping intervention (0/1)
- **Control vars:** war type, severity, duration; development status,…
- **Counterfactual question:** Switch UN intervention for each war
- **Data analysis:** Logit model
Model Dependence Example
Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

- **Data:** 124 Post-World War II civil wars
- **Dependent var:** peacebuilding success
- **Treatment:** multilateral UN peacekeeping intervention (0/1)
- **Control vars:** war type, severity, duration; development status,…
- **Counterfactual question:** Switch UN intervention for each war
- **Data analysis:** Logit model
- **The question:** How *model dependent* are the results?

Detecting Model Dependence
Model Dependence Example
Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

• **Data:** 124 Post-World War II civil wars
• **Dependent var:** peacebuilding success
• **Treatment:** multilateral UN peacekeeping intervention (0/1)
• **Control vars:** war type, severity, duration; development status,...
• **Counterfactual question:** Switch UN intervention for each war
• **Data analysis:** Logit model
• **The question:** How *model dependent* are the results?
• **Percent of counterfactuals in the convex hull:**
Model Dependence Example
Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

• Data: 124 Post-World War II civil wars
• Dependent var: peacebuilding success
• Treatment: multilateral UN peacekeeping intervention (0/1)
• Control vars: war type, severity, duration; development status,…
• Counterfactual question: Switch UN intervention for each war
• Data analysis: Logit model
• The question: How model dependent are the results?
• Percent of counterfactuals in the convex hull: 0%
Model Dependence Example
Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

- **Data:** 124 Post-World War II civil wars
- **Dependent var:** peacebuilding success
- **Treatment:** multilateral UN peacekeeping intervention (0/1)
- **Control vars:** war type, severity, duration; development status,…
- **Counterfactual question:** Switch UN intervention for each war
- **Data analysis:** Logit model
- **The question:** How *model dependent* are the results?
- **Percent of counterfactuals in the convex hull:** 0%
- **↝ without estimating any models, we know:** inferences will be model dependent
Model Dependence Example
Replication of Doyle and Sambanis, APSR 2000
(From: King and Zeng, 2007)

- **Data**: 124 Post-World War II civil wars
- **Dependent var**: peacebuilding success
- **Treatment**: multilateral UN peacekeeping intervention (0/1)
- **Control vars**: war type, severity, duration; development status,…
- **Counterfactual question**: Switch UN intervention for each war
- **Data analysis**: Logit model
- **The question**: How *model dependent* are the results?
- **Percent of counterfactuals in the convex hull**: 0%
- ↝ without estimating any models, we know: inferences will be model dependent
- **For illustration**: let’s find an example….
## Two Logit Models, Apparently Similar Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Original &quot;Interactive&quot; Model</th>
<th>Modified Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wartype</td>
<td>-1.742 (0.609) p=0.004</td>
<td>-1.666 (0.606) p=0.006</td>
</tr>
<tr>
<td>Logdead</td>
<td>-0.445 (0.126) p=0.000</td>
<td>-0.437 (0.125) p=0.000</td>
</tr>
<tr>
<td>Wardur</td>
<td>0.006 (0.006) p=0.258</td>
<td>0.006 (0.006) p=0.342</td>
</tr>
<tr>
<td>Factnum</td>
<td>-1.259 (0.703) p=0.073</td>
<td>-1.045 (0.899) p=0.245</td>
</tr>
<tr>
<td>Factnum2</td>
<td>0.062 (0.065) p=0.346</td>
<td>0.032 (0.104) p=0.756</td>
</tr>
<tr>
<td>Trnsfcap</td>
<td>0.004 (0.002) p=0.010</td>
<td>0.004 (0.002) p=0.017</td>
</tr>
<tr>
<td>Develop</td>
<td>0.001 (0.000) p=0.065</td>
<td>0.001 (0.000) p=0.068</td>
</tr>
<tr>
<td>Exp</td>
<td>-6.016 (3.071) p=0.050</td>
<td>-6.215 (3.065) p=0.043</td>
</tr>
<tr>
<td>Decade</td>
<td>-0.299 (0.169) p=0.077</td>
<td>-0.284 (0.169) p=0.093</td>
</tr>
<tr>
<td>Treaty</td>
<td>2.124 (0.821) p=0.010</td>
<td>2.126 (0.802) p=0.008</td>
</tr>
<tr>
<td>UNOP4</td>
<td>3.135 (1.091) p=0.004</td>
<td>0.262 (1.392) p=0.851</td>
</tr>
<tr>
<td>Wardur*UNOP4</td>
<td>— (—)</td>
<td>0.037 (0.011) p=0.001</td>
</tr>
<tr>
<td>Constant</td>
<td>8.609 (2.157) p=0.000</td>
<td>7.978 (2.350) p=0.000</td>
</tr>
</tbody>
</table>

### Log-likelihood
- Original: -45.649
- Modified: -44.902

### Pseudo $R^2$
- Original: 0.423
- Modified: 0.433
Two Logit Models, Apparently Similar Results

Effect of Multilateral UN Intervention on Peacebuilding Success

<table>
<thead>
<tr>
<th>Variable</th>
<th>Original “Interactive” Model</th>
<th>Modified Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wartype</td>
<td>−1.742 (SE 0.609, P-val 0.004)</td>
<td>−1.666 (SE 0.606, P-val 0.006)</td>
</tr>
<tr>
<td>Logdead</td>
<td>−0.445 (SE 0.126, P-val 0.000)</td>
<td>−0.437 (SE 0.125, P-val 0.000)</td>
</tr>
<tr>
<td>Wardur</td>
<td>0.006 (SE 0.006, P-val 0.258)</td>
<td>0.006 (SE 0.006, P-val 0.342)</td>
</tr>
<tr>
<td>Factnum</td>
<td>−1.259 (SE 0.703, P-val 0.073)</td>
<td>−1.045 (SE 0.899, P-val 0.245)</td>
</tr>
<tr>
<td>Factnum2</td>
<td>0.062 (SE 0.065, P-val 0.346)</td>
<td>0.032 (SE 0.104, P-val 0.756)</td>
</tr>
<tr>
<td>Trnsfcap</td>
<td>0.004 (SE 0.002, P-val 0.010)</td>
<td>0.004 (SE 0.002, P-val 0.017)</td>
</tr>
<tr>
<td>Develop</td>
<td>0.001 (SE 0.000, P-val 0.065)</td>
<td>0.001 (SE 0.000, P-val 0.068)</td>
</tr>
<tr>
<td>Exp</td>
<td>−6.016 (SE 3.071, P-val 0.050)</td>
<td>−6.215 (SE 3.065, P-val 0.043)</td>
</tr>
<tr>
<td>Decade</td>
<td>−0.299 (SE 0.169, P-val 0.077)</td>
<td>−0.284 (SE 0.169, P-val 0.093)</td>
</tr>
<tr>
<td>Treaty</td>
<td>2.124 (SE 0.821, P-val 0.010)</td>
<td>2.126 (SE 0.802, P-val 0.008)</td>
</tr>
<tr>
<td>UNOP4</td>
<td>3.135 (SE 1.091, P-val 0.004)</td>
<td>0.262 (SE 1.392, P-val 0.851)</td>
</tr>
<tr>
<td>Wardur*UNOP4</td>
<td>—</td>
<td>0.037 (SE 0.011, P-val 0.001)</td>
</tr>
<tr>
<td>Constant</td>
<td>8.609 (SE 2.157, P-val 0.000)</td>
<td>7.978 (SE 2.350, P-val 0.000)</td>
</tr>
<tr>
<td>N</td>
<td>122</td>
<td>122</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>−45.649</td>
<td>−44.902</td>
</tr>
<tr>
<td>Pseudo $R^2$</td>
<td>0.423</td>
<td>0.433</td>
</tr>
</tbody>
</table>

Detecting Model Dependence
## Two Logit Models, Apparently Similar Results

### Effect of Multilateral UN Intervention on Peacebuilding Success

<table>
<thead>
<tr>
<th>Variables</th>
<th>Original “Interactive” Model</th>
<th>Modified Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff</td>
<td>SE</td>
</tr>
<tr>
<td>Wartype</td>
<td>-1.742</td>
<td>.609</td>
</tr>
<tr>
<td>Logdead</td>
<td>-.445</td>
<td>.126</td>
</tr>
<tr>
<td>Wardur</td>
<td>.006</td>
<td>.006</td>
</tr>
<tr>
<td>Factnum</td>
<td>-1.259</td>
<td>.703</td>
</tr>
<tr>
<td>Factnum2</td>
<td>.062</td>
<td>.065</td>
</tr>
<tr>
<td>Trnsfcap</td>
<td>.004</td>
<td>.002</td>
</tr>
<tr>
<td>Develop</td>
<td>.001</td>
<td>.000</td>
</tr>
<tr>
<td>Exp</td>
<td>-6.016</td>
<td>3.071</td>
</tr>
<tr>
<td>Decade</td>
<td>-.299</td>
<td>.169</td>
</tr>
<tr>
<td>Treaty</td>
<td>2.124</td>
<td>.821</td>
</tr>
<tr>
<td>UNOP4</td>
<td>3.135</td>
<td>1.091</td>
</tr>
<tr>
<td>Wardur*UNOP4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Constant</td>
<td>8.609</td>
<td>2.157</td>
</tr>
</tbody>
</table>

| N               | 122    | 122   |
| Log-likelihood  | -45.649| -44.902|
| Pseudo $R^2$    | .423   | .433  |
Model Dependence: Same Fit, Different Predictions

In Sample Fit

Counterfactual Prediction
Detecting Model Dependence

Matching to Reduce Model Dependence

Three Matching Methods

Problems with Propensity Score Matching

The Matching Frontier
Readings, Matching

• Do powerful methods have to be complicated?
  ↝ "Causal Inference Without Balance Checking: Coarsened Exact Matching" (PA, 2011; Stefano Iacus, Gary King, and Giuseppe Porro)

• The most popular method (propensity score matching, used in 140,000 articles!) sounds magical:
  ↝ "Why Propensity Scores Should Not Be Used for Matching" (PA, 2019; Gary King and Richard Nielsen)

• Matching methods optimize either imbalance (≈ bias) or # units pruned (≈ variance); users need both simultaneously:
  ↝ "The Balance-Sample Size Frontier in Matching Methods for Causal Inference" (AJPS, 2017; Gary King, Christopher Lucas and Richard Nielsen)

• Current practice, matching as preprocessing: violates current statistical theory.
  So let's change the theory:
  ↝ "A Theory of Statistical Inference for Matching Methods in Causal Research" (Stefano Iacus, Gary King, Giuseppe Porro)
Readings, Matching

- Do powerful methods have to be complicated?
Readings, Matching

- Do powerful methods have to be complicated?
Readings, Matching

- **Do powerful methods have to be complicated?**

- **The most popular method (propensity score matching, used in 140,000 articles!) sounds magical:**
Readings, Matching

- Do powerful methods have to be complicated?
- The most popular method (propensity score matching, used in 140,000 articles!) sounds magical:
  - “Why Propensity Scores Should Not Be Used for Matching” (Gary King, Richard Nielsen) (*PA*, 2019; Gary King and Richard Nielsen)
Readings, Matching

• Do powerful methods have to be complicated?

• The most popular method (propensity score matching, used in 140,000 articles!) sounds magical:
  ~ “Why Propensity Scores Should Not Be Used for Matching” (Gary King, Richard Nielsen) (*PA*, 2019; Gary King and Richard Nielsen)

• Matching methods optimize either imbalance (≈ bias) or # units pruned (≈ variance); users need both simultaneously’:
Readings, Matching

- Do powerful methods have to be complicated?
  - “Causal Inference Without Balance Checking: Coarsened Exact Matching” (PA, 2011. Stefano Iacus, Gary King, and Giuseppe Porro)

- The most popular method (propensity score matching, used in 140,000 articles!) sounds magical:
  - “Why Propensity Scores Should Not Be Used for Matching” (Gary King, Richard Nielsen) (PA, 2019; Gary King and Richard Nielsen)

- Matching methods optimize either imbalance (∝ bias) or # units pruned (∝ variance); users need both simultaneously’:
  - “The Balance-Sample Size Frontier in Matching Methods for Causal Inference” (AJPS, 2017; Gary King, Christopher Lucas and Richard Nielsen)
Readings, Matching

• Do powerful methods have to be complicated?
  ↗ “Causal Inference Without Balance Checking: Coarsened Exact Matching” (PA, 2011. Stefano Iacus, Gary King, and Giuseppe Porro)

• The most popular method (propensity score matching, used in 140,000 articles!) sounds magical:
  ↗ “Why Propensity Scores Should Not Be Used for Matching” (Gary King, Richard Nielsen) (PA, 2019; Gary King and Richard Nielsen)

• Matching methods optimize either imbalance (∼ bias) or # units pruned (∼ variance); users need both simultaneously’:
  ↗ “The Balance-Sample Size Frontier in Matching Methods for Causal Inference” (AJPS, 2017; Gary King, Christopher Lucas and Richard Nielsen)

• Current practice, matching as preprocessing:
Readings, Matching

• Do powerful methods have to be complicated?
  ~ “Causal Inference Without Balance Checking: Coarsened Exact Matching” (PA, 2011. Stefano Iacus, Gary King, and Giuseppe Porro)

• The most popular method (propensity score matching, used in 140,000 articles!) sounds magical:
  ~ “Why Propensity Scores Should Not Be Used for Matching” (Gary King, Richard Nielsen) (PA, 2019; Gary King and Richard Nielsen)

• Matching methods optimize either imbalance (≈ bias) or # units pruned (≈ variance); users need both simultaneously’:
  ~ “The Balance-Sample Size Frontier in Matching Methods for Causal Inference” (AJPS, 2017; Gary King, Christopher Lucas and Richard Nielsen)

• Current practice, matching as preprocessing: violates current statistical theory.
Readings, Matching

- Do powerful methods have to be complicated?
  ~ “Causal Inference Without Balance Checking: Coarsened Exact Matching” (PA, 2011. Stefano Iacus, Gary King, and Giuseppe Porro)

- The most popular method (propensity score matching, used in 140,000 articles!) sounds magical:
  ~ “Why Propensity Scores Should Not Be Used for Matching” (Gary King, Richard Nielsen) (PA, 2019; Gary King and Richard Nielsen)

- Matching methods optimize either imbalance (∼ bias) or # units pruned (∼ variance); users need both simultaneously’:
  ~ “The Balance-Sample Size Frontier in Matching Methods for Causal Inference” (AJPS, 2017; Gary King, Christopher Lucas and Richard Nielsen)

- Current practice, matching as preprocessing: violates current statistical theory. So let’s change the theory:
Readings, Matching

• Do powerful methods have to be complicated?
  ~ “Causal Inference Without Balance Checking: Coarsened Exact Matching” (PA, 2011. Stefano Iacus, Gary King, and Giuseppe Porro)

• The most popular method (propensity score matching, used in 140,000 articles!) sounds magical:
  ~ “Why Propensity Scores Should Not Be Used for Matching” (Gary King, Richard Nielsen) (PA, 2019; Gary King and Richard Nielsen)

• Matching methods optimize either imbalance (∼ bias) or # units pruned (∼ variance); users need both simultaneously’:
  ~ “The Balance-Sample Size Frontier in Matching Methods for Causal Inference” (AJPS, 2017; Gary King, Christopher Lucas and Richard Nielsen)

• Current practice, matching as preprocessing: violates current statistical theory. So let’s change the theory:
  ~ “A Theory of Statistical Inference for Matching Methods in Causal Research” (Stefano Iacus, Gary King, Giuseppe Porro)
Matching to Reduce Model Dependence
Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)
Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)
Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, Political Analysis)
Matching to Reduce Model Dependence
(Ho, Imai, King, Stuart, 2007: fig.1, Political Analysis)
Matching to Reduce Model Dependence
(Ho, Imai, King, Stuart, 2007: fig.1, Political Analysis)
Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)
Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, Political Analysis)
Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)
Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)

![Graph showing the relationship between education (years) and outcome, with a trend line indicating a decline in outcome as education increases.](image-url)
The Problems Matching Solves

- Qualitative choice from unbiased estimates = biased estimator
- e.g., Choosing from results of 50 randomized experiments
- Choosing based on "plausibility" is probably worse
- Conscientious effort doesn't avoid biases (Banaji 2013)
- People do not have easy access to their own mental processes or feedback to avoid the problem (Wilson and Brekke 1994)
- Experts overestimate their ability to control personal biases more than nonexperts, and more prominent experts are the most overconfident (Tetlock 2005)
- "Teaching psychology is mostly a waste of time" (Kahneman 2011)
The Problems Matching Solves

Without Matching:

• Qualitative choice from unbiased estimates = biased estimator
  e.g., Choosing from results of 50 randomized experiments
  Choosing based on "plausibility" is probably worse

• Conscientious effort doesn't avoid biases (Banaji 2013)

• People do not have easy access to their own mental processes or feedback to avoid the problem (Wilson and Brekke 1994)

• Experts overestimate their ability to control personal biases more than nonexperts, and more prominent experts are the most overconfident (Tetlock 2005)

• “Teaching psychology is mostly a waste of time” (Kahneman 2011)
The Problems Matching Solves

Without Matching:

Imbalance
The Problems Matching Solves

Without Matching:

Imbalance $\sim$ Model Dependence
The Problems Matching Solves

Without Matching:

Imbalance $\Rightarrow$ Model Dependence $\Rightarrow$ Researcher discretion
The Problems Matching Solves

Without Matching:

Imbalance $\leadsto$ Model Dependence $\leadsto$ Researcher discretion $\leadsto$ Bias
The Problems Matching Solves

Without Matching:

Imbalance $\leadsto$ Model Dependence $\leadsto$ Researcher discretion $\leadsto$ Bias

- Qualitative choice from unbiased estimates $=$ biased estimator
The Problems Matching Solves

Without Matching:

Imbalance $\leadsto$ Model Dependence $\leadsto$ Researcher discretion $\leadsto$ Bias

- Qualitative choice from unbiased estimates = biased estimator
  - e.g., Choosing from *results* of 50 randomized experiments
The Problems Matching Solves

Without Matching:

Imbalance $\leadsto$ Model Dependence $\leadsto$ Researcher discretion $\leadsto$ Bias

- Qualitative choice from unbiased estimates = biased estimator
  - e.g., Choosing from results of 50 randomized experiments
  - Choosing based on “plausibility” is probably worse
The Problems Matching Solves

Without Matching:

Imbalance $\leadsto$ Model Dependence $\leadsto$ Researcher discretion $\leadsto$ Bias

- Qualitative choice from unbiased estimates $=\text{biased estimator}$
  - e.g., Choosing from *results* of 50 randomized experiments
  - Choosing based on “plausibility” is probably worse$^{[eff]}$
- conscientious effort doesn’t avoid biases (Banaji 2013)$^{[acc]}$
Without Matching:

Imbalance $\leadsto$ Model Dependence $\leadsto$ Researcher discretion $\leadsto$ Bias

- Qualitative choice from unbiased estimates $=$ biased estimator
  - e.g., Choosing from results of 50 randomized experiments
  - Choosing based on “plausibility” is probably worse
- Conscientious effort doesn’t avoid biases (Banaji 2013)
- People do not have easy access to their own mental processes or feedback to avoid the problem (Wilson and Brekke 1994)
The Problems Matching Solves

Without Matching:

Imbalance $\Rightarrow$ Model Dependence $\Rightarrow$ Researcher discretion $\Rightarrow$ Bias

- Qualitative choice from unbiased estimates $=$ biased estimator
  - e.g., Choosing from results of 50 randomized experiments
  - Choosing based on “plausibility” is probably worse
- Conscientious effort doesn’t avoid biases (Banaji 2013)
- People do not have easy access to their own mental processes or feedback to avoid the problem (Wilson and Brekke 1994)
- Experts overestimate their ability to control personal biases more than nonexperts, and more prominent experts are the most overconfident (Tetlock 2005)
The Problems Matching Solves

Without Matching:

Imbalance $\leadsto$ Model Dependence $\leadsto$ Researcher discretion $\leadsto$ Bias

- Qualitative choice from unbiased estimates $=\,$ biased estimator
  - e.g., Choosing from results of 50 randomized experiments
  - Choosing based on “plausibility” is probably worse
- conscientious effort doesn’t avoid biases (Banaji 2013)
- People do not have easy access to their own mental processes or feedback to avoid the problem (Wilson and Brekke 1994)
- Experts overestimate their ability to control personal biases more than nonexperts, and more prominent experts are the most overconfident (Tetlock 2005)
- “Teaching psychology is mostly a waste of time” (Kahneman 2011)
The Problems Matching Solves

**Without Matching:**

- Imbalance $\leadsto$ Model Dependence $\leadsto$ Researcher discretion $\leadsto$ Bias
The Problems Matching Solves

Without Matching:

Imbalance → Model Dependence → Researcher discretion → Bias
The Problems Matching Solves

Without Matching:

Imbalance $\leadsto$ Model Dependence $\leadsto$ Researcher discretion $\leadsto$ Bias
The Problems Matching Solves

Without Matching:

Imbalance $\sim$ Model Dependence $\sim$ Researcher discretion $\sim$ Bias

Matching to Reduce Model Dependence
The Problems Matching Solves

Without Matching:

- Imbalance $\sim$ Model Dependence $\sim$ Researcher discretion $\sim$ Bias

Matching to Reduce Model Dependence
The Problems Matching Solves

Without Matching:

Imbalance $\rightsquigarrow$ Model Dependence $\rightsquigarrow$ Researcher discretion $\rightsquigarrow$ Bias

A central project of statistics: Automating away human discretion
What’s Matching?

- **Notation:**
  - $Y_i$: dependent variable,
  - $T_i$ (1=treated, 0=control),
  - $X_i$: confounders

- **Treatment Effect for treated observation $i$:**
  \[ TE_i = Y_i - Y_i(0) = \text{observed} - \text{unobserved} \]

- **Estimate $Y_i(0)$ with $Y_j$ with a matched ($X_i \approx X_j$) control**

- **Quantities of Interest**
  1. **SATT: Sample Average Treatment effect on the Treated:**
     \[ SATT = \text{Mean}_{i \in \{ T_i = 1 \}}(TE_i) \]
  2. **FSATT: Feasible SATT**

- **Big convenience:** Follow preprocessing with whatever statistical method you’d have used without matching

- **Pruning nonmatches makes control vars matter less:** reduces imbalance, model dependence, researcher discretion, & bias
What’s Matching?

- **Notation:** $Y_i$ dep var, $T_i$ (1=treated, 0=control), $X_i$ confounders

  Treatment Effect for treated observation $i$:
  \[ \text{TE}_i = Y_i - Y_i(0) = \text{observed} - \text{unobserved} \]

  - Estimate $Y_i(0)$ with $Y_j$ with a matched ($X_i \approx X_j$) control

  **Quantities of Interest**
  1. **SATT:** Sample Average Treatment effect on the Treated:
     \[ \text{SATT} = \text{Mean}_{i \in \{ T_i = 1 \}} (\text{TE}_i) \]
  2. **FSATT:** Feasible SATT (prune badly matched treateds too)

Matching to Reduce Model Dependence

15/45
What’s Matching?

- **Notation:** $Y_i$ dep var, $T_i$ (1=treated, 0=control), $X_i$ confounders
- Treatment Effect for treated observation $i$:

$$TE_i = Y_i - Y_i(0) = \text{observed} - \text{unobserved}$$

- Estimate $Y_i(0)$ with $Y_j$ with a matched ($X_i \approx X_j$) control

**Quantities of Interest**

1. **SATT:** Sample Average Treatment effect on the Treated:
   $$SATT = \text{Mean}_{i \in \{T_i = 1\}} (TE_i)$$

2. **FSATT:** Feasible SATT (prune badly matched treateds too)

Big convenience: Follow preprocessing with whatever statistical method you'd have used without matching

Pruning nonmatches makes control vars matter less: reduces imbalance, model dependence, researcher discretion, & bias
What’s Matching?

- **Notation**: $Y_i$ dep var, $T_i$ (1=treated, 0=control), $X_i$ confounders
- Treatment Effect for treated observation $i$:

$$\text{TE}_i = Y_i(1) - Y_i(0)$$
What’s Matching?

• **Notation:** $Y_i$ dep var, $T_i$ (1=treated, 0=control), $X_i$ confounders

• Treatment Effect for treated observation $i$:

$$TE_i = Y_i(1) - Y_i(0) = \text{observed} - \text{unobserved}$$
What’s Matching?

- **Notation:** $Y_i$ dep var, $T_i$ (1=treated, 0=control), $X_i$ confounders

- Treatment Effect for treated observation $i$:

$$TE_i = Y_i - Y_i(0) = \text{observed} - \text{unobserved}$$
What’s Matching?

- **Notation:** \( Y_i \) dep var, \( T_i \) (1=treated, 0=control), \( X_i \) confounders
- **Treatment Effect for treated observation \( i \):**
  \[
  \text{TE}_i = Y_i - Y_i(0) \\
  = \text{observed} - \text{unobserved}
  \]
- **Estimate** \( Y_i(0) \) with \( Y_j \) with a matched \((X_i \approx X_j)\) control
What’s Matching?

- **Notation:** $Y_i$ dep var, $T_i$ (1=treated, 0=control), $X_i$ confounders
- **Treatment Effect for treated observation $i$:**
  
  $\text{TE}_i = Y_i - Y_i(0)$
  
  $= \text{observed} - \text{unobserved}$

- **Estimate $Y_i(0)$ with $Y_j$ with a matched ($X_i \approx X_j$) control**
- **Quantities of Interest**

  1. **SATT:** Sample Average Treatment effect on the Treated: $\text{SATT} = \text{Mean}_{i \in \{T_i = 1\}}(\text{TE}_i)$
  2. **FSATT:** Feasible SATT (prune badly matched treateds too)

Big convenience: Follow preprocessing with whatever statistical method you'd have used without matching

Pruning nonmatches makes control vars matter less: reduces imbalance, model dependence, researcher discretion, & bias
What’s Matching?

- **Notation**: $Y_i$ dep var, $T_i$ (1=treated, 0=control), $X_i$ confounders
- Treatment Effect for treated observation $i$:
  \[
  TE_i = Y_i - Y_i(0) = \text{observed} - \text{unobserved}
  \]
- Estimate $Y_i(0)$ with $Y_j$ with a matched ($X_i \approx X_j$) control
- **Quantities of Interest**
  1. SATT: Sample Average Treatment effect on the Treated:
     \[
     \text{SATT} = \text{Mean} \left( TE_i \right)_{i \in \{T_i=1\}}
     \]
What’s Matching?

- **Notation:** $Y_i$ dep var, $T_i$ (1=treated, 0=control), $X_i$ confounders
- **Treatment Effect for treated observation $i$:**

$$\text{TE}_i = Y_i - Y_i(0)$$

= observed – unobserved

- **Estimate $Y_i(0)$ with $Y_j$ with a matched ($X_i \approx X_j$) control
- **Quantities of Interest**
  1. **SATT:** Sample Average Treatment effect on the Treated:

$$\text{SATT} = \text{Mean } \left( \text{TE}_i \right)_{i \in \{T_i=1\}}$$

  2. **FSATT:** Feasible SATT (prune badly matched treateds too)
What’s Matching?

- **Notation:** $Y_i$ dep var, $T_i$ (1=treated, 0=control), $X_i$ confounders
- **Treatment Effect for treated observation $i$:**
  \[
  TE_i = Y_i - Y_i(0) = \text{observed} - \text{unobserved}
  \]
- Estimate $Y_i(0)$ with $Y_j$ with a matched ($X_i \approx X_j$) control
- **Quantities of Interest**
  1. **SATT:** Sample Average Treatment effect on the Treated:
     \[
     \text{SATT} = \text{Mean } (TE_i)_{i \in \{T_i=1\}}
     \]
  2. **FSATT:** Feasible SATT (prune badly matched treateds too)
- **Big convenience:** Follow preprocessing with whatever statistical method you’d have used without matching
What’s Matching?

- **Notation:** $Y_i$ dep var, $T_i$ (1=treated, 0=control), $X_i$ confounders
- **Treatment Effect for treated observation $i$:**

  $$TE_i = Y_i - Y_i(0) = \text{observed} - \text{unobserved}$$

- **Estimate $Y_i(0)$ with $Y_j$ with a matched ($X_i \approx X_j$) control**
- **Quantities of Interest**
  1. **SATT:** Sample Average Treatment effect on the Treated:

     $$SATT = \text{Mean} \left( TE_i \right)_{i \in \{T_i=1\}}$$

  2. **FSATT:** Feasible SATT (prune badly matched treateds too)

- **Big convenience:** Follow preprocessing with whatever statistical method you’d have used without matching
- **Pruning nonmatches makes control vars matter less:** reduces imbalance, model dependence, researcher discretion, & bias
Evaluating Reduction in Model Dependence

Hypothesis: Democratic senate majorities slow FDA drug approval time

Data: $n = 408$ new drugs (262 approved, 146 pending)

Measured confounders: 18 (clinical factors, firm characteristics, media variables, etc.)

Model: lognormal survival

QOI: Causal effect of Democratic Senate majority (identified by Carpenter as not robust)

Match: prune 49 units (2 treated, 17 control units)

Run: 262,143 possible specifications; calculate SATT for each

Evaluate: Variability in SATT across specifications

(Normally we'd only use one or a few specifications)
Evaluating Reduction in Model Dependence

Empirical Illustration: Carpenter, AJPS, 2002

Hypothesis: Democratic senate majorities slow FDA drug approval time

Data: $n = 408$ new drugs (262 approved, 146 pending)

Measured confounders: 18 (clinical factors, firm characteristics, media variables, etc.)

Model: lognormal survival

QOI: Causal effect of Democratic Senate majority (identified by Carpenter as not robust)

Match: prune 49 units (2 treated, 17 control units)

Run: 262,143 possible specifications; calculate SATT for each

Evaluate: Variability in SATT across specifications

(Normally we'd only use one or a few specifications)
Evaluating Reduction in Model Dependence
Empirical Illustration: Carpenter, AJPS, 2002

- **Hypothesis:** Democratic senate majorities slow FDA drug approval time
Evaluating Reduction in Model Dependence
Empirical Illustration: Carpenter, AJPS, 2002

- **Hypothesis:** Democratic senate majorities slow FDA drug approval time
- **Data:** $n = 408$ new drugs (262 approved, 146 pending)
Evaluating Reduction in Model Dependence
Empirical Illustration: Carpenter, AJPS, 2002

• **Hypothesis:** Democratic senate majorities slow FDA drug approval time
• **Data:** $n = 408$ new drugs (262 approved, 146 pending)
• **Measured confounders:** 18 (clinical factors, firm characteristics, media variables, etc.)
Evaluating Reduction in Model Dependence

Empirical Illustration: Carpenter, AJPS, 2002

- **Hypothesis:** Democratic senate majorities slow FDA drug approval time
- **Data:** $n = 408$ new drugs (262 approved, 146 pending)
- **Measured confounders:** 18 (clinical factors, firm characteristics, media variables, etc.)
- **Model:** lognormal survival
• **Hypothesis:** Democratic senate majorities slow FDA drug approval time

• **Data:** \( n = 408 \) new drugs (262 approved, 146 pending)

• **Measured confounders:** 18 (clinical factors, firm characteristics, media variables, etc.)

• **Model:** lognormal survival

• **QOI:** Causal effect of Democratic Senate majority (identified by Carpenter as not robust)
Evaluating Reduction in Model Dependence
Empirical Illustration: Carpenter, AJPS, 2002

- **Hypothesis**: Democratic senate majorities slow FDA drug approval time
- **Data**: $n = 408$ new drugs (262 approved, 146 pending)
- **Measured confounders**: 18 (clinical factors, firm characteristics, media variables, etc.)
- **Model**: lognormal survival
- **QOI**: Causal effect of Democratic Senate majority (identified by Carpenter as not robust)
- **Match**: prune 49 units (2 treated, 17 control units)
Evaluating Reduction in Model Dependence
Empirical Illustration: Carpenter, AJPS, 2002

- **Hypothesis**: Democratic senate majorities slow FDA drug approval time
- **Data**: \( n = 408 \) new drugs (262 approved, 146 pending)
- **Measured confounders**: 18 (clinical factors, firm characteristics, media variables, etc.)
- **Model**: lognormal survival
- **QOI**: Causal effect of Democratic Senate majority (identified by Carpenter as not robust)
- **Match**: prune 49 units (2 treated, 17 control units)
- **Run**: 262,143 possible specifications; calculate SATT for each
Evaluating Reduction in Model Dependence
Empirical Illustration: Carpenter, AJPS, 2002

• **Hypothesis:** Democratic senate majorities slow FDA drug approval time
• **Data:** $n = 408$ new drugs (262 approved, 146 pending)
• **Measured confounders:** 18 (clinical factors, firm characteristics, media variables, etc.)
• **Model:** lognormal survival
• **QOI:** Causal effect of Democratic Senate majority (identified by Carpenter as not robust)
• **Match:** prune 49 units (2 treated, 17 control units)
• **Run:** 262,143 possible specifications; calculate SATT for each
• **Evaluate:** *Variability* in SATT across specifications
Evaluating Reduction in Model Dependence
Empirical Illustration: Carpenter, AJPS, 2002

- **Hypothesis:** Democratic senate majorities slow FDA drug approval time
- **Data:** $n = 408$ new drugs (262 approved, 146 pending)
- **Measured confounders:** 18 (clinical factors, firm characteristics, media variables, etc.)
- **Model:** lognormal survival
- **QOI:** Causal effect of Democratic Senate majority (identified by Carpenter as not robust)
- **Match:** prune 49 units (2 treated, 17 control units)
- **Run:** 262,143 possible specifications; calculate SATT for each
- **Evaluate:** Variability in SATT across specifications
- (Normally we’d only use one or a few specifications)
Reducing Model Dependence

Point estimate of Carpenter's specification using raw data

Density

Matched data

Estimated in-sample average treatment effect for the treated

Raw data

Matching to Reduce Model Dependence
SATT Histogram: Effect of Democratic Senate majority on FDA drug approval time, across 262,143 specifications
Another Example: Jeffrey Koch, AJPS, 2002

Matching to Reduce Model Dependence
Another Example: Jeffrey Koch, AJPS, 2002

SATT Histogram: Effect of being a highly visible female Republican candidate across 63 possible specifications with the Koch data
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You're Doing!

- Framework: simple random sampling from a population
- Exact matching: Rarely possible; but would make estimation easy
- Assumptions:
  - Unconfoundedness: \( T \perp Y \) (0)
  - Common support: \( \text{Pr}(T=1 | X) < 1 \) (\( T=0, 1 \) are both possible)

Approximate matching (bias correction, new variance estimation): common, but all current practices would have to change

Alternative Theory of Inference: It's Gonna be OK!

- Framework: stratified random sampling from a population
- Define \( A \): a stratum in a partition of the product space of \( X \) ("continuous" variables have natural breakpoints)
- We already know and use these procedures: Group strong and weak partisans; Don’t match college dropout with 1st year grad student
- Assumptions:
  - Set-wide Unconfoundedness: \( T \perp Y \) (0) \( \mid A \)
  - Set-wide Common support: \( \text{Pr}(T=1 | A) < 1 \)
- Fits all common matching methods & practices; no asymptotics
- Easy extensions for: multi-level, continuous, & mismeasured treatments; \( A \) too wide, \( n \) too small
Assumptions to Justify Current Practice

**Existing Theory of Inference: Stop What You’re Doing!**

- **Framework:** simple random sampling from a population
- **Assumptions:**
  - **Unconfoundedness:** \( T \perp Y \mid X \)
  - **Common support:** \( \Pr(T = 1 \mid X) < 1 \) \( T = 0, 1 \) are both possible
- **Approximate matching (bias correction, new variance estimation):** common, but all current practices would have to change

**Alternative Theory of Inference: It's Gonna be OK!**

- **Framework:** stratified random sampling from a population
- **Define** \( A \): a stratum in a partition of the product space of \( X \) ("continuous" variables have natural breakpoints)
- **Assumptions:**
  - **Set-wide Unconfoundedness:** \( T \perp Y \mid A \)
  - **Set-wide Common support:** \( \Pr(T = 1 \mid A) < 1 \)
- **Fits all common matching methods & practices; no asymptotics**
- **Easy extensions for: multi-level, continuous, & mismeasured treatments;** \( A \) too wide, \( n \) too small

Matching to Reduce Model Dependence
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You’re Doing!

• Framework: simple random sampling from a population
• Exact matching: Rarely possible; but would make estimation easy
• Assumptions:
  • Unconfoundedness: \( T \perp Y \mid X \) (Healthy & unhealthy get meds)
  • Common support: \( Pr(T = 1 \mid X) < 1 \) (\( T = 0, 1 \) are both possible)

Approximate matching (bias correction, new variance estimation): common, but all current practices would have to change

Alternative Theory of Inference: It’s Gonna be OK!

• Framework: stratified random sampling from a population
• Define \( A \): a stratum in a partition of the product space of \( X \) ("continuous" variables have natural breakpoints)
• We already know and use these procedures: Group strong and weak partisans; Don’t match college dropout with 1st year grad student
• Assumptions:
  • Set-wide Unconfoundedness: \( T \perp Y \mid A \)
  • Set-wide Common support: \( Pr(T = 1 \mid A) < 1 \)

Fits all common matching methods & practices; no asymptotics
• Easy extensions for: multi-level, continuous, & mismeasured treatments; \( A \) too wide, \( n \) too small
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You’re Doing!

- Framework: simple random sampling from a population

Alternative Theory of Inference: It’s Gonna be OK!

- Framework: stratified random sampling from a population
- Define $A$: a stratum in a partition of the product space of $X$ ("continuous" variables have natural breakpoints)
- We already know and use these procedures: Group strong and weak partisans; Don’t match college dropout with 1st year grad student

- Assumptions:
  - Set-wide Unconfoundedness: $T \perp Y (0) | A$
  - Set-wide Common support: $Pr(T = 1 | A) < 1$
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You’re Doing!

• Framework: simple random sampling from a population
  • Exact matching: Rarely possible; but would make estimation easy

Alternative Theory of Inference: It’s Gonna be OK!

• Framework: stratified random sampling from a population
  • Define $A$: a stratum in a partition of the product space of $X$ ("continuous" variables have natural breakpoints)
  • We already know and use these procedures: Group strong and weak partisans; Don’t match college dropout with 1st year grad student

• Assumptions:
  • Set-wide Unconfoundedness: $T \perp Y (0) \mid A$
  • Set-wide Common support: $\Pr(T = 1 \mid A) < 1$

• Fits all common matching methods & practices; no asymptotics
  • Easy extensions for: multi-level, continuous, & mismeasured treatments; $A$ too wide, $n$ too small
Assumptions to Justify Current Practice

**Existing Theory of Inference: Stop What You’re Doing!**

- **Framework:** *simple random sampling* from a population
- **Exact matching:** Rarely possible; but would make estimation easy
- **Assumptions:**

  - **Unconfoundedness:** $T \perp Y \mid X$ (Healthy & unhealthy get meds)
  - **Common support:** $\Pr(T = 1 \mid X) < 1$ ($T = 0, 1$ are both possible)

**Alternative Theory of Inference: It’s Gonna be OK!**

- **Framework:** *stratified random sampling* from a population
- **Define $A$:** a stratum in a partition of the product space of $X$ ("continuous" variables have natural breakpoints)
- **Assumptions:**

  - **Set-wide Unconfoundedness:** $T \perp Y \mid A$
  - **Set-wide Common support:** $\Pr(T = 1 \mid A) < 1$

  Fits all common matching methods & practices; no asymptotics
  - Easy extensions for: multi-level, continuous, & mismeasured treatments; $A$ too wide, $n$ too small
## Assumptions to Justify Current Practice

### Existing Theory of Inference: Stop What You’re Doing!

- **Framework:** simple random sampling from a population
- **Exact matching:** Rarely possible; but would make estimation easy
- **Assumptions:**
  - *Unconfoundedness:* $T \perp Y(0) \mid X$ (Healthy & unhealthy get meds)

### Alternative Theory of Inference: It’s Gonna be OK!

- **Framework:** stratified random sampling from a population
- **Define** $A$: a stratum in a partition of the product space of $X$ ("continuous" variables have natural breakpoints)
- **Assumptions:**
  - *Set-wide Unconfoundedness:* $T \perp Y(0) \mid A$
  - *Set-wide Common support:* $\Pr(T = 1 \mid A) < 1$
- Fits all common matching methods & practices; no asymptotics
- Easy extensions for: multi-level, continuous, & mismeasured treatments; $A$ too wide, $n$ too small
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You’re Doing!

- Framework: simple random sampling from a population
- Exact matching: Rarely possible; but would make estimation easy
- Assumptions:
  - Unconfoundedness: \( T \perp Y(0) \mid X \) (Healthy & unhealthy get meds)
  - Common support: \( \Pr(T = 1 \mid X) < 1 \) (\( T = 0, 1 \) are both possible)

Alternative Theory of Inference: It’s Gonna be OK!

- Framework: stratified random sampling from a population
- Define \( A \): a stratum in a partition of the product space of \( X \) ("continuous" variables have natural breakpoints)
- We already know and use these procedures: Group strong and weak partisans; Don’t match college dropout with 1st year grad student
- Assumptions:
  - Set-wide Unconfoundedness: \( T \perp Y \mid A \)
  - Set-wide Common support: \( \Pr(T = 1 \mid A) < 1 \)
- Fits all common matching methods & practices; no asymptotics
- Easy extensions for: multi-level, continuous, & mismeasured treatments; \( A \) too wide, \( n \) too small
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You’re Doing!

- Framework: simple random sampling from a population
- Exact matching: Rarely possible; but would make estimation easy
- Assumptions:
  - Unconfoundedness: $T \perp Y(0) \mid X$ (Healthy & unhealthy get meds)
  - Common support: $\Pr(T = 1 \mid X) < 1$ ($T = 0, 1$ are both possible)
- Approximate matching (bias correction, new variance estimation): common, but all current practices would have to change

Alternative Theory of Inference: It’s Gonna be OK!

- Framework: stratified random sampling from a population
- Define $A$: a stratum in a partition of the product space of $X$ ("continuous" variables have natural breakpoints)
- We already know and use these procedures: Group strong and weak partisans; Don’t match college dropout with 1st year grad student
- Assumptions:
  - Set-wide Unconfoundedness: $T \perp Y(0) \mid A$
  - Set-wide Common support: $\Pr(T = 1 \mid A) < 1$ ($T = 0, 1$ are both possible)
- Fits all common matching methods & practices; no asymptotics
- Easy extensions for: multi-level, continuous, & mismeasured treatments; $A$ too wide, $n$ too small
Assumptions to Justify Current Practice

**Existing Theory of Inference: Stop What You’re Doing!**

- **Framework:** simple random sampling from a population
- **Exact matching:** Rarely possible; but would make estimation easy
- **Assumptions:**
  - Unconfoundedness: $T \perp Y(0) \mid X$ (Healthy & unhealthy get meds)
  - Common support: $\Pr(T = 1 \mid X) < 1$ ($T = 0, 1$ are both possible)
- **Approximate matching (bias correction, new variance estimation):** common, but all current practices would have to change

**Alternative Theory of Inference: It’s Gonna be OK!**

- **Framework:** stratified random sampling from a population
Assumptions to Justify Current Practice

**Existing Theory of Inference: Stop What You’re Doing!**

- **Framework:** simple random sampling from a population
- **Exact matching:** Rarely possible; but would make estimation easy
- **Assumptions:**
  - **Unconfoundedness:** $T \perp Y(0) \mid X$ (Healthy & unhealthy get meds)
  - **Common support:** $\Pr(T = 1 \mid X) < 1$ ($T = 0, 1$ are both possible)
- **Approximate matching (bias correction, new variance estimation):** common, but all current practices would have to change

**Alternative Theory of Inference: It’s Gonna be OK!**

- **Framework:** stratified random sampling from a population
- **Define $A$:** a stratum in a partition of the product space of $X$ ("continuous" variables have natural breakpoints)
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You’re Doing!

- Framework: simple random sampling from a population
- Exact matching: Rarely possible; but would make estimation easy
- Assumptions:
  - Unconfoundedness: $T \perp Y(0) \mid X$ (Healthy & unhealthy get meds)
  - Common support: $\Pr(T = 1 \mid X) < 1$ ($T = 0, 1$ are both possible)
- Approximate matching (bias correction, new variance estimation): common, but all current practices would have to change

Alternative Theory of Inference: It’s Gonna be OK!

- Framework: stratified random sampling from a population
- Define $A$: a stratum in a partition of the product space of $X$ (“continuous” variables have natural breakpoints)
- We already know and use these procedures: Group strong and weak partisans; Don’t match college dropout with 1st year grad student
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You’re Doing!

- Framework: simple random sampling from a population
- Exact matching: Rarely possible; but would make estimation easy
- Assumptions:
  - Unconfoundedness: $T \perp Y(0) \mid X$ (Healthy & unhealthy get meds)
  - Common support: $\Pr(T = 1 \mid X) < 1$ ($T = 0, 1$ are both possible)
- Approximate matching (bias correction, new variance estimation): common, but all current practices would have to change

Alternative Theory of Inference: It’s Gonna be OK!

- Framework: stratified random sampling from a population
- Define $A$: a stratum in a partition of the product space of $X$ (“continuous” variables have natural breakpoints)
- We already know and use these procedures: Group strong and weak partisans; Don’t match college dropout with 1st year grad student
- Assumptions:
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You’re Doing!

- Framework: simple random sampling from a population
- Exact matching: Rarely possible; but would make estimation easy
- Assumptions:
  - Unconfoundedness: $T \perp Y(0) \mid X$ (Healthy & unhealthy get meds)
  - Common support: $\Pr(T = 1 \mid X) < 1$ ($T = 0, 1$ are both possible)
- Approximate matching (bias correction, new variance estimation): common, but all current practices would have to change

Alternative Theory of Inference: It’s Gonna be OK!

- Framework: stratified random sampling from a population
- Define $A$: a stratum in a partition of the product space of $X$ (“continuous” variables have natural breakpoints)
- We already know and use these procedures: Group strong and weak partisans; Don’t match college dropout with 1st year grad student
- Assumptions:
  - Set-wide Unconfoundedness: $T \perp Y(0) \mid A$
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You’re Doing!

- Framework: **simple random sampling** from a population
- Exact matching: Rarely possible; but would make estimation easy
- Assumptions:
  - Unconfoundedness: $T \perp Y(0) \mid X$ (Healthy & unhealthy get meds)
  - Common support: $\Pr(T = 1 \mid X) < 1$ ($T = 0, 1$ are both possible)
- Approximate matching (bias correction, new variance estimation): common, but all current practices would have to change

Alternative Theory of Inference: It’s Gonna be OK!

- Framework: **stratified random sampling** from a population
- Define $A$: a stratum in a partition of the product space of $X$ (“continuous” variables have natural breakpoints)
- We already know and use these procedures: Group strong and weak partisans; Don’t match college dropout with 1st year grad student
- Assumptions:
  - Set-wide Unconfoundedness: $T \perp Y(0) \mid A$
  - Set-wide Common support: $\Pr(T = 1 \mid A) < 1$
Assumptions to Justify Current Practice

Existing Theory of Inference: Stop What You’re Doing!

- **Framework**: simple random sampling from a population
- **Exact matching**: Rarely possible; but would make estimation easy
- **Assumptions**:
  - *Unconfoundedness*: $T \perp Y(0) \mid X$ (Healthy & unhealthy get meds)
  - *Common support*: $\Pr(T = 1 \mid X) < 1$ ($T = 0, 1$ are both possible)
- **Approximate matching** (bias correction, new variance estimation): common, but all current practices would have to change

Alternative Theory of Inference: It’s Gonna be OK!

- **Framework**: stratified random sampling from a population
- **Define $A$**: a stratum in a partition of the product space of $X$ (“continuous” variables have natural breakpoints)
- **We already know and use these procedures**: Group strong and weak partisans; Don’t match college dropout with 1st year grad student
- **Assumptions**:
  - *Set-wide Unconfoundedness*: $T \perp Y(0) \mid A$
  - *Set-wide Common support*: $\Pr(T = 1 \mid A) < 1$
- **Fits all common matching methods & practices; no asymptotics**
## Assumptions to Justify Current Practice

### Existing Theory of Inference: Stop What You’re Doing!

- **Framework:** simple random sampling from a population
- **Exact matching:** Rarely possible; but would make estimation easy
- **Assumptions:**
  - *Unconfoundedness:* $T \perp Y(0) \mid X$ (Healthy & unhealthy get meds)
  - *Common support:* $\Pr(T = 1 \mid X) < 1$ ($T = 0, 1$ are both possible)
- **Approximate matching (bias correction, new variance estimation):** common, but all current practices would have to change

### Alternative Theory of Inference: It’s Gonna be OK!

- **Framework:** stratified random sampling from a population
- **Define $A$:** a stratum in a partition of the product space of $X$ (“continuous” variables have natural breakpoints)
- **We already know and use these procedures:** Group strong and weak partisans; Don’t match college dropout with 1st year grad student
- **Assumptions:**
  - *Set-wide Unconfoundedness:* $T \perp Y(0) \mid A$
  - *Set-wide Common support:* $\Pr(T = 1 \mid A) < 1$
- **Fits all common matching methods & practices; no asymptotics**
- **Easy extensions for:** multi-level, continuous, & mismeasured treatments; $A$ too wide, $n$ too small
Detecting Model Dependence

Matching to Reduce Model Dependence

Three Matching Methods

Problems with Propensity Score Matching

The Matching Frontier
Matching: Finding Hidden Randomized Experiments

Types of Experiments

- Balance
- Covariates:
  - Complete Randomization
  - Fully Blocked
  - Observed
  - Unobserved

On average, fully blocked dominates complete randomization for:
- Imbalance,
- Model dependence,
- Power,
- Efficiency,
- Bias,
- Research costs,
- Robustness.

E.g., Imai, King, Nall 2009: SEs 600% smaller!
Matching: Finding Hidden Randomized Experiments

Types of Experiments

- Balance: Complete Randomization vs. Fully Blocked

Observed: On average

Unobserved: On average

\[ \rightarrow \]

Fully blocked dominates complete randomization for:

- Imbalance,
- Model dependence,
- Power,
- Efficiency,
- Bias,
- Research costs,
- Robustness.

E.g., Imai, King, Nall 2009: SEs 600% smaller!
Matching: Finding Hidden Randomized Experiments

Types of Experiments

- Complete Randomization
- Fully Blocked
- Observed on average
- Unobserved on average

↝ Fully blocked dominates complete randomization for:
- imbalance,
- model dependence,
- power,
- efficiency,
- bias,
- research costs,
- robustness.

E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)

- PSM: complete randomization
- Other methods: fully blocked
- Other matching methods dominate PSM
### Types of Experiments

- **Complete Randomization**

  $\Rightarrow$ Fully blocked dominates complete randomization for:
  - imbalance,
  - model dependence,
  - power,
  - efficiency,
  - bias,
  - research costs,
  - robustness.

  E.g., Imai, King, Nall 2009: SEs 600% smaller!

### Three Matching Methods

- **PSM:** complete randomization
- **Other methods:** fully blocked
- **Other matching methods dominate PSM** (wait, it gets worse)
Types of Experiments

- **Complete Randomization**
- **Fully Blocked**

*Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!
## Types of Experiments

<table>
<thead>
<tr>
<th>Balance</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unobserved</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Types of Experiments

<table>
<thead>
<tr>
<th>Balance</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates:</td>
<td>On average</td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unobserved</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Types of Experiments

<table>
<thead>
<tr>
<th>Balance</th>
<th>Complete Randomization</th>
<th>FullyBlocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates:</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observed</strong></td>
<td>On average</td>
<td></td>
</tr>
<tr>
<td><strong>Unobserved</strong></td>
<td>On average</td>
<td></td>
</tr>
</tbody>
</table>
## Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed</strong></td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td><strong>Unobserved</strong></td>
<td>On average</td>
<td></td>
</tr>
</tbody>
</table>

- Fully blocked dominates complete randomization for: imbalance, model dependence, power, efficiency, bias, research costs, robustness.

E.g., Imai, King, Nall 2009: SEs 600% smaller!
Matching: Finding Hidden Randomized Experiments

### Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

Fully blocked dominates complete randomization for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!
### Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

~ Fully blocked dominates complete randomization
# Matching: Finding Hidden Randomized Experiments

## Types of Experiments

<table>
<thead>
<tr>
<th>Balance</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

~ Fully blocked dominates complete randomization for:
### Types of Experiments

<table>
<thead>
<tr>
<th>Balance</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates:</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td>On average</td>
</tr>
<tr>
<td>Unobserved</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

≈ *Fully blocked* dominates *complete randomization* for: imbalance,
### Types of Experiments

<table>
<thead>
<tr>
<th>Balance</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates:</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Observed</td>
<td>On average</td>
<td>On average</td>
</tr>
<tr>
<td>Unobserved</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \sim \) *Fully blocked* dominates *complete randomization* for: imbalance, model dependence,

---

Goal of Each Matching Method (in Observational Data)

- **PSM:** complete randomization
- Other methods: fully blocked
- Other matching methods dominate PSM

(wait, it gets worse)
### Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

$\sim$ Fully blocked dominates complete randomization for: imbalance, model dependence, power,
### Matching: Finding Hidden Randomized Experiments

#### Types of Experiments

<table>
<thead>
<tr>
<th>Balance</th>
<th>Complete</th>
<th>Fully</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates:</td>
<td>Randomization</td>
<td>Blocked</td>
</tr>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

〜 *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency,
Matching: Finding Hidden Randomized Experiments

### Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

~ Fully blocked dominates complete randomization for: imbalance, model dependence, power, efficiency, bias,
## Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

~ Fully blocked dominates complete randomization for: imbalance, model dependence, power, efficiency, bias, research costs,
Matching: Finding Hidden Randomized Experiments

### Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

~ Fully blocked dominates complete randomization for: imbalance, model dependence, power, efficiency, bias, research costs, robustness.
## Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed</strong></td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td><strong>Unobserved</strong></td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

Fully blocked dominates complete randomization for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!
Matching: Finding Hidden Randomized Experiments

Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

~ Fully blocked dominates complete randomization for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)
Matching: Finding Hidden Randomized Experiments

Types of Experiments

<table>
<thead>
<tr>
<th>Balance</th>
<th>Complete</th>
<th>Fully</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates:</td>
<td>Randomization</td>
<td>Blocked</td>
</tr>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

→ Fully blocked dominates complete randomization for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)

- PSM: complete randomization
### Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Observed</em></td>
<td>On average</td>
<td><em>Exact</em></td>
</tr>
<tr>
<td><em>Unobserved</em></td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

~ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

### Goal of Each Matching Method (in Observational Data)

- PSM: *complete randomization*
- Other methods: *fully blocked*
Matching: Finding Hidden Randomized Experiments

Types of Experiments

<table>
<thead>
<tr>
<th>Balance</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates:</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

\(\sim\) Fully blocked dominates complete randomization for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)

- PSM: complete randomization
- Other methods: fully blocked
- Other matching methods dominate PSM
## Types of Experiments

<table>
<thead>
<tr>
<th>Balance Covariates:</th>
<th>Complete Randomization</th>
<th>Fully Blocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>On average</td>
<td>Exact</td>
</tr>
<tr>
<td>Unobserved</td>
<td>On average</td>
<td>On average</td>
</tr>
</tbody>
</table>

> Fully blocked dominates complete randomization for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

## Goal of Each Matching Method (in Observational Data)

- PSM: complete randomization
- Other methods: fully blocked
- Other matching methods dominate PSM (wait, it gets worse)
Method 1: Mahalanobis Distance Matching

Procedure

1. Preprocess (Matching)
   - Distance \((X_c, X_t) = \sqrt{(X_c - X_t)^\top S^{-1} (X_c - X_t)}\)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > caliper

2. Estimation
   - Difference in means or a model
   - Interpretation
     - Quiz: Do you understand the distance trade offs?
     - Quiz: Does standardization help?
     - \(\Rightarrow\) Mahalanobis is for methodologists; in applications, use Euclidean!
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance \( \left( X_c, X_t \right) = \sqrt{\left( X_c - X_t \right)' S^{-1} \left( X_c - X_t \right)} \)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > caliper

2. Estimation
   - Difference in means or a model

Interpretation
- Quiz: Do you understand the distance trade offs?
- Quiz: Does standardization help?

\( \Rightarrow \) Mahalanobis is for methodologists; in applications, use Euclidean!
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess**
   - Distance: \( d(X_c, X_t) = \sqrt{(X_c - X_t)'S^{-1}(X_c - X_t)} \)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > caliper
   (Many adjustments available to this basic method)

2. **Estimation**
   - Difference in means or a model

Interpretation

- Quiz: Do you understand the distance trade offs?
- Quiz: Does standardization help?
  \( \rightarrow \) Mahalanobis is for methodologists; in applications, use Euclidean!
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. Preprocess (Matching)

• Distance \( (X_c, X_t) = \sqrt{(X_c - X_t)'S^{-1}(X_c - X_t)} \)

• Match each treated unit to the nearest control unit

• Control units: not reused; pruned if unused

• Prune matches if Distance > caliper

(Many adjustments available to this basic method)

2. Estimation
Difference in means or a model

Interpretation

• Quiz: Do you understand the distance trade offs?

• Quiz: Does standardization help?

\( \rightarrow \) Mahalanobis is for methodologists; in applications, use Euclidean!
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** *(Matching)*

2. **Estimation** Difference in means or a model
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess (Matching)**
   - Distance($X_c, X_t$) = $\sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)}$

2. **Estimation** Difference in means or a model
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** *(Matching)*
   - Distance($X_c, X_t$) = $\sqrt{(X_c - X_t)'S^{-1}(X_c - X_t)}$
   - Match each treated unit to the nearest control unit

2. **Estimation** Difference in means or a model
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** *(Matching)*
   - Distance \( (X_c, X_t) = \sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)} \)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused

2. **Estimation** Difference in means or a model

Quiz: Do you understand the distance trade offs?
Quiz: Does standardization help?
\(\rightarrow\) Mahalanobis is for methodologists; in applications, use Euclidean!
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. Preprocess (Matching)
   - Distance($X_c, X_t$) = $\sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)}$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > caliper

2. Estimation Difference in means or a model
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** (Matching)
   - Distance\( (X_c, X_t) = \sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)} \)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > *caliper*
   - (Many adjustments available to this basic method)

2. **Estimation** Difference in means or a model
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** *(Matching)*
   - Distance \( (X_c, X_t) = \sqrt{(X_c - X_t)'S^{-1}(X_c - X_t)} \)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > *caliper*
   - (Many adjustments available to this basic method)

2. **Estimation** Difference in means or a model

Interpretation
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** (Matching)
   - Distance\( (X_c, X_t) = \sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)} \)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > caliper
   - (Many adjustments available to this basic method)

2. **Estimation** Difference in means or a model

Interpretation

- **Quiz**: Do you understand the distance trade offs?
Method 1: Mahalanobis Distance Matching  
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess (Matching)**
   - Distance\((X_c, X_t) = \sqrt{(X_c - X_t)'S^{-1}(X_c - X_t)}\)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance\(>\) caliper
   - (Many adjustments available to this basic method)

2. **Estimation** Difference in means or a model

Interpretation

- **Quiz**: Do you understand the distance trade-offs?
- **Quiz**: Does standardization help?
Method 1: Mahalanobis Distance Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** (Matching)
   - Distance\( (X_c, X_t) = \sqrt{(X_c - X_t)'S^{-1}(X_c - X_t)} \)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > caliper
   - (Many adjustments available to this basic method)

2. **Estimation** Difference in means or a model

Interpretation

- **Quiz:** Do you understand the distance trade offs?
- **Quiz:** Does standardization help?

Mahalanobis is for methodologists; in applications, use Euclidean!
Mahalanobis Distance Matching

Three Matching Methods
Mahalanobis Distance Matching

Three Matching Methods
Mahalanobis Distance Matching

Age

Education (years)
Mahalanobis Distance Matching

Three Matching Methods
Mahalanobis Distance Matching

Three Matching Methods
Mahalanobis Distance Matching

Three Matching Methods
Best Case: Mahalanobis Distance Matching
Best Case: Mahalanobis Distance Matching

Three Matching Methods
Best Case: Mahalanobis Distance Matching

![Scatter plot showing data points for age and education (years). The x-axis represents education (years) ranging from 12 to 28, and the y-axis represents age ranging from 20 to 80. The data points are scattered across the plot, indicating a correlation between the two variables.](image-url)
Method 2: Coarsened Exact Matching

Procedure

1. Preprocess (Matching)
   • Temporarily coarsen $X$ as much as you're willing
     • e.g., Education (grade school, high school, college, graduate)
   • Apply exact matching to the coarsened $X$, $C(X)$
     • Sort observations into strata, each with unique values of $C(X)$
     • Prune any stratum with 0 treated or 0 control units
   • Pass on original (uncoarsened) units except those pruned

2. Estimation
   • Difference in means or a model
     • Weight controls in each stratum to equal treateds

Interpretation

• Quiz: Do you understand distance trade offs?
• Quiz: What do you do if you have too few observations?
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   • Temporarily coarsen $X$ as much as you’re willing
   • e.g., Education (grade school, high school, college, graduate)
   • Apply exact matching to the coarsened $X$, $C(X)$
   • Sort observations into strata, each with unique values of $C(X)$
   • Prune any stratum with 0 treated or 0 control units
   • Pass on original (uncoarsened) units except those pruned

2. Estimation
   • Weight controls in each stratum to equal treateds

Interpretation
• Quiz: Do you understand distance trade offs?
• Quiz: What do you do if you have too few observations?
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

Procedure

1. Preprocess (Matching)
   - Temporarily coarsen \( X \) as much as you’re willing
   - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened \( X \), \( C(X) \)
   - Sort observations into strata, each with unique values of \( C(X) \)
   - Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned

2. Estimation
   - Difference in means or a model
   - Weight controls in each stratum to equal treateds

Interpretation

- Quiz: Do you understand distance trade offs?
- Quiz: What do you do if you have too few observations?
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** *(Matching)*
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** *(Matching)*

2. **Estimation** Difference in means or a model
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

Procedure

1. Preprocess (Matching)
   - Temporarily coarsen $X$ as much as you’re willing
     - e.g., Education (grade school, high school, college, graduate)

2. Estimation Difference in means or a model
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** *(Matching)*
   - Temporarily coarsen $X$ as much as you’re willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened $X$, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units

2. **Estimation** Difference in means or a model
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** *(Matching)*
   - Temporarily coarsen $X$ as much as you’re willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened $X$, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned

2. **Estimation** Difference in means or a model
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

Procedure

1. Preprocess (Matching)
   - Temporarily coarsen $X$ as much as you’re willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened $X$, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned

2. Estimation Difference in means or a model
   - Weight controls in each stratum to equal treateds
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess (Matching)**
   - **Temporarily coarsen** $X$ as much as you’re willing
     - e.g., Education (grade school, high school, college, graduate)
   - **Apply exact matching** to the coarsened $X$, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
   - **Pass on original (uncoarsened) units except those pruned**

2. **Estimation** Difference in means or a model
   - Weight controls in each stratum to equal treateds

Interpretation
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

Procedure

1. Preprocess (Matching)
   - Temporarily coarsen $X$ as much as you’re willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened $X$, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned

2. Estimation Difference in means or a model
   - Weight controls in each stratum to equal treateds

Interpretation

- Quiz: Do you understand distance trade offs?
Method 2: Coarsened Exact Matching
(Approximates Fully Blocked Experiment)

Procedure

1. **Preprocess** *(Matching)*
   - Temporarily coarsen $X$ as much as you’re willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened $X$, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned

2. **Estimation** Difference in means or a model
   - Weight controls in each stratum to equal treateds

Interpretation

- **Quiz:** Do you understand distance trade offs?
- **Quiz:** What do you do if you have too few observations?
Coarsened Exact Matching
Coarsened Exact Matching

Education vs. Age

Three Matching Methods
Three Matching Methods
Coarsened Exact Matching

Three Matching Methods
Three Matching Methods
Three Matching Methods
Three Matching Methods
Best Case: Coarsened Exact Matching
Best Case: Coarsened Exact Matching
Best Case: Coarsened Exact Matching
Best Case: Coarsened Exact Matching
Method 3: Propensity Score Matching

Procedure

1. Preprocess (Matching)
   - Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1 \mid X) = 1 + e^{-X_i \beta}$
   - Distance ($X_c, X_t$) = $|\pi_c - \pi_t|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance $> \text{caliper}$
   - (Many adjustments available to this basic method)

2. Estimation
   - Difference in means or a model

Interpretation

• Quiz: Do you understand distance trade offs?
• Quiz: What do you do when one variable is very important?
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure
1. Preprocess (Matching)
   - Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1 | X) = 1 + e^{-X_i \beta}$
   - Distance ($X_c, X_t$) = $|\pi_c - \pi_t|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > caliper
   - (Many adjustments available to this basic method)

2. Estimation
   - Difference in means or a model

Interpretation
- Quiz: Do you understand distance trade offs?
- Quiz: What do you do when one variable is very important?
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure

1. Preprocess (Matching)
   • Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1 | X)$
   • Distance ($X_c, X_t$) = $|\pi_c - \pi_t|$
   • Match each treated unit to the nearest control unit
   • Control units: not reused; pruned if unused
   • Prune matches if Distance $> \text{caliper}$
   • (Many adjustments available to this basic method)

2. Estimation
   Difference in means or a model

Interpretation
• Quiz: Do you understand distance trade offs?
• Quiz: What do you do when one variable is very important?
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure

1. **Preprocess** *(Matching)*
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure

1. **Preprocess** *(Matching)*

2. **Estimation** Difference in means or a model
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure

1. **Preprocess** (Matching)
   - Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-x_i\beta}}$

2. **Estimation** Difference in means or a model
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure

1. **Preprocess (Matching)**
   - Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - Distance$(X_c, X_t) = |\pi_c - \pi_t|$

2. **Estimation** Difference in means or a model
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure

1. **Preprocess** *(Matching)*
   - Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - Distance($X_c, X_t$) = $|\pi_c - \pi_t|$
   - Match each treated unit to the nearest control unit

2. **Estimation** Difference in means or a model
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

**Procedure**

1. **Preprocess (Matching)**
   - Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - Distance($X_c, X_t$) = $|\pi_c - \pi_t|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused

2. **Estimation** Difference in means or a model
Method 3: Propensity Score Matching  
(Approximates Completely Randomized Experiment)

### Procedure

1. **Preprocess (Matching)**
   - Reduce $k$ elements of $X$ to scalar $\pi_i = \Pr(T_i = 1|X) = \frac{1}{1 + e^{-X_i\beta}}$
   - Distance($X_c, X_t$) = $|\pi_c - \pi_t|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > caliper

2. **Estimation** Difference in means or a model
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure

1. **Preprocess** *(Matching)*
   - Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - Distance$(X_c, X_t) = |\pi_c - \pi_t|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance$>caliper$
   - (Many adjustments available to this basic method)

2. **Estimation** Difference in means or a model
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure

1. **Preprocess** (Matching)
   - Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - Distance$(X_c, X_t) = |\pi_c - \pi_t|$  
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance$>caliper$
   - (Many adjustments available to this basic method)

2. **Estimation** Difference in means or a model

Interpretation
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure

1. **Preprocess** *(Matching)*
   - Reduce $k$ elements of $X$ to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i \beta}}$
   - Distance($X_c, X_t$) = $|\pi_c - \pi_t|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance > *caliper*
   - (Many adjustments available to this basic method)

2. **Estimation** Difference in means or a model

Interpretation

- **Quiz:** Do you understand distance trade offs?
Method 3: Propensity Score Matching
(Approximates Completely Randomized Experiment)

Procedure

1. Preprocess (Matching)
   - Reduce $k$ elements of $X$ to scalar $\pi_i = \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - Distance$(X_c, X_t) = |\pi_c - \pi_t|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance $>\text{caliper}$
   - (Many adjustments available to this basic method)

2. Estimation Difference in means or a model

Interpretation

- Quiz: Do you understand distance trade offs?
- Quiz: What do you do when one variable is very important?
Propensity Score Matching

Three Matching Methods
Propensity Score Matching

![Graph showing propensity score matching with age on the y-axis and education (years) on the x-axis. The graph displays various data points represented by 'C' and 'T' symbols, with 'C' indicating the control group and 'T' indicating the treatment group. The y-axis ranges from 0 to 80, and the x-axis ranges from 12 to 28, with 12, 16, 20, 24, and 28 clearly marked. The graph also includes a vertical line at the right side, representing the propensity score range from 0 to 1.]
Propensity Score Matching

Three Matching Methods
Propensity Score Matching

Three Matching Methods

Education (years)

Age

12 16 20 24 28

20

30

40

50

60

70

80

1

0

Propensity Score
Propensity Score Matching

Three Matching Methods
Propensity Score Matching

Three Matching Methods
Propensity Score Matching

Three Matching Methods
Best Case: Propensity Score Matching
Best Case: Propensity Score Matching
Best Case: Propensity Score Matching
Best Case: Propensity Score Matching

Three Matching Methods
Best Case: Propensity Score Matching is Suboptimal
Detecting Model Dependence

Matching to Reduce Model Dependence

Three Matching Methods

Problems with Propensity Score Matching

The Matching Frontier
Random Pruning Increases Imbalance

**Random pruning**: pruning process is independent of $X$.

- **Discrete example**
  - Sex-balanced dataset: treated $M_t$, $F_t$, controls $M_c$, $F_c$
  - Randomly prune 1 treated & 1 control
    $\rightarrow$ 4 possible datasets: 2 balanced $\{M_t, M_c\}, \{F_t, F_c\}$, 2 imbalanced $\{M_t, F_c\}, \{F_t, M_c\}$
  - $\Rightarrow$ random pruning increases imbalance

- **Continuous example**
  - Dataset: $T \in \{0, 1\}$ randomly assigned; $X$ any fixed variable; with $n$ units
  - Measure of imbalance: squared difference in means $d^2$, where $d = \bar{X}_t - \bar{X}_c$
  - $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$)
  - Random pruning $\rightarrow n$ declines $\rightarrow E(d^2)$ increases
  - $\Rightarrow$ random pruning increases imbalance

- Result is completely general
Random Pruning Increases Imbalance
Deleting data only helps if you’re careful!

- **Random pruning**: pruning process is independent of $X$.

  - **Discrete example**

    - **Sex-balanced dataset**: treated $M_t$, $F_t$, controls $M_c$, $F_c$.

    - Randomly prune 1 treated & 1 control $\Rightarrow 4$ possible datasets: 2 balanced $\{M_t, M_c\}$, $\{F_t, F_c\}$, 2 imbalanced $\{M_t, F_c\}$, $\{F_t, M_c\}$.

    - $\Rightarrow$ random pruning increases imbalance.

  - **Continuous example**

    - Dataset: $T \in \{0, 1\}$ randomly assigned; $X$ any fixed variable; with $n$ units.

    - Measure of imbalance: squared difference in means $d^2$, where $d = \bar{X}_t - \bar{X}_c$.

    - $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$).

    - Random pruning $\Rightarrow n$ declines $\Rightarrow E(d^2)$ increases.

    - $\Rightarrow$ random pruning increases imbalance.

- **Result is completely general**.
Random Pruning Increases Imbalance
Deleting data only helps if you’re careful!

• “Random pruning”: pruning process is independent of $X$

Problems with Propensity Score Matching
Random Pruning Increases Imbalance

Deleting data only helps if you’re careful!

- “Random pruning”: pruning process is independent of $X$
- Discrete example

- Randomly prune 1 treated & 1 control $\rightarrow$ 4 possible datasets: 2 balanced
  - $\{M_t, F_t\}$, $\{M_c, F_c\}$
- 2 imbalanced
  - $\{M_t, F_c\}$, $\{F_t, M_c\}$

$\implies$ random pruning increases imbalance

- Continuous example

  - Dataset: $T \in \{0, 1\}$ randomly assigned; $X$ any fixed variable; with $n$ units
  - Measure of imbalance: squared difference in means $d^2$, where $d = \bar{X}_t - \bar{X}_c$
  - $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$)

$\implies$ random pruning $\rightarrow n$ declines $\rightarrow E(d^2)$ increases

$\implies$ random pruning increases imbalance

- Result is completely general
Random Pruning Increases Imbalance
Deleting data only helps if you’re careful!

- “Random pruning”: pruning process is independent of $X$
- Discrete example
  - Sex-balanced dataset: treated $M_t, F_t$, controls $M_c, F_c$
Random Pruning Increases Imbalance
Deleting data only helps if you’re careful!

• “Random pruning”: pruning process is independent of $X$
• Discrete example
  • Sex-balanced dataset: treateds $M_t, F_t$, controls $M_c, F_c$
  • Randomly prune 1 treated & 1 control $\sim$ 4 possible datasets: 2 balanced $\{M_t, M_c\}, \{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}, \{F_t, M_c\}$
Random Pruning Increases Imbalance
Deleting data only helps if you’re careful!

- “Random pruning”: pruning process is independent of $X$
- **Discrete example**
  - Sex-balanced dataset: treateds $M_t, F_t$, controls $M_c, F_c$
  - Randomly prune 1 treated & 1 control $\sim$ 4 possible datasets: 2 balanced $\{M_t, M_c\}, \{F_t, F_c\}$
    - 2 imbalanced $\{M_t, F_c\}, \{F_t, M_c\}$
  - $\implies$ random pruning increases imbalance

Problems with Propensity Score Matching
Random Pruning Increases Imbalance
Deleting data only helps if you’re careful!

• “Random pruning”: pruning process is independent of $X$
• Discrete example
  • Sex-balanced dataset: treateds $M_t, F_t$, controls $M_c, F_c$
  • Randomly prune 1 treated & 1 control $\sim$ 4 possible datasets: 2 balanced $\{M_t, M_c\}, \{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}, \{F_t, M_c\}$
  • $\Rightarrow$ random pruning increases imbalance

• Continuous example
Random Pruning Increases Imbalance
Deleting data only helps if you’re careful!

• “Random pruning”: pruning process is independent of $X$

• Discrete example
  - Sex-balanced dataset: treateds $M_t, F_t$, controls $M_c, F_c$
  - Randomly prune 1 treated & 1 control $\sim$ 4 possible datasets: 2 balanced $\{M_t, M_c\}, \{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}, \{F_t, M_c\}$
  - $\implies$ random pruning increases imbalance

• Continuous example
  - Dataset: $T \in \{0, 1\}$ randomly assigned; $X$ any fixed variable; with $n$ units
Random Pruning Increases Imbalance
Deleting data only helps if you’re careful!

• “Random pruning”: pruning process is independent of $X$

• Discrete example
  • Sex-balanced dataset: treateds $M_t, F_t$, controls $M_c, F_c$
  • Randomly prune 1 treated & 1 control $\sim$ 4 possible datasets: 2 balanced $\{M_t, M_c\}, \{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}, \{F_t, M_c\}$
  • $\Rightarrow$ random pruning increases imbalance

• Continuous example
  • Dataset: $T \in \{0, 1\}$ randomly assigned; $X$ any fixed variable; with $n$ units
  • Measure of imbalance: squared difference in means $d^2$, where $d = \bar{X}_t - \bar{X}_c$
Random Pruning Increases Imbalance

Deleting data only helps if you’re careful!

• “Random pruning”: pruning process is independent of $X$

• **Discrete example**
  - Sex-balanced dataset: treateds $M_t, F_t$, controls $M_c, F_c$
  - Randomly prune 1 treated & 1 control $\sim 4$ possible datasets: 2 balanced $\{M_t, M_c\}, \{F_t, F_c\}$
    - 2 imbalanced $\{M_t, F_c\}, \{F_t, M_c\}$
  - $\implies$ random pruning increases imbalance

• **Continuous example**
  - Dataset: $T \in \{0, 1\}$ randomly assigned; $X$ any fixed variable; with $n$ units
  - Measure of imbalance: squared difference in means $d^2$, where $d = \bar{X}_t - \bar{X}_c$
  - $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$)
Random Pruning Increases Imbalance
Deleting data only helps if you’re careful!

• “Random pruning”: pruning process is independent of $X$
• Discrete example
  • Sex-balanced dataset: treateds $M_t, F_t$, controls $M_c, F_c$
  • Randomly prune 1 treated & 1 control $\leadsto$ 4 possible datasets: 2 balanced $\{M_t, M_c\}, \{F_t, F_c\}$
  2 imbalanced $\{M_t, F_c\}, \{F_t, M_c\}$
  • $\implies$ random pruning increases imbalance
• Continuous example
  • Dataset: $T \in \{0, 1\}$ randomly assigned; $X$ any fixed variable; with $n$ units
  • Measure of imbalance: squared difference in means $d^2$, where $d = \bar{X}_t - \bar{X}_c$
  • $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$)
  • Random pruning $\leadsto$ $n$ declines $\leadsto E(d^2)$ increases
Random Pruning Increases Imbalance

Deleting data only helps if you’re careful!

- “Random pruning”: pruning process is independent of $X$
- **Discrete example**
  - Sex-balanced dataset: treateds $M_t, F_t$, controls $M_c, F_c$
  - Randomly prune 1 treated & 1 control $\sim$ 4 possible datasets: 2 balanced $\{M_t, M_c\}, \{F_t, F_c\}$
      2 imbalanced $\{M_t, F_c\}, \{F_t, M_c\}$
  - $\implies$ random pruning increases imbalance
- **Continuous example**
  - Dataset: $T \in \{0, 1\}$ randomly assigned; $X$ any fixed variable; with $n$ units
  - Measure of imbalance: squared difference in means $d^2$, where $d = \bar{X}_t - \bar{X}_c$
  - $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$)
  - Random pruning $\sim n$ declines $\sim E(d^2)$ increases
  - $\implies$ random pruning increases imbalance
Random Pruning Increases Imbalance

Deleting data only helps if you’re careful!

• “Random pruning”: pruning process is independent of $X$

• Discrete example
  • Sex-balanced dataset: treateds $M_t, F_t$, controls $M_c, F_c$
  • Randomly prune 1 treated & 1 control $\sim$ 4 possible datasets: 2 balanced $\{M_t, M_c\}, \{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}, \{F_t, M_c\}$
  • $\implies$ random pruning increases imbalance

• Continuous example
  • Dataset: $T \in \{0, 1\}$ randomly assigned; $X$ any fixed variable; with $n$ units
  • Measure of imbalance: squared difference in means $d^2$, where $d = \bar{X}_t - \bar{X}_c$
  • $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$)
  • Random pruning $\sim n$ declines $\sim E(d^2)$ increases
  • $\implies$ random pruning increases imbalance

• Result is completely general
PSM’s Statistical Properties

1. Low Standards:
   - Sometimes helps, never optimizes
   - Efficient relative to complete randomization, but inefficient relative to (the more powerful) full blocking
   - Other methods dominate:

2. The PSM Paradox:
   - When you do “better,” you do worse
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) ↝ all \( \hat{\pi} \approx 0.5 \) (or constant within strata) ↝ pruning at random ↝ Imbalance ↝ Inefficency ↝ Model dependence ↝ Bias
   - If the data have no good matches, the paradox won’t be a problem but you’re cooked anyway.
   - Doesn’t PSM solve the curse of dimensionality problem? Nope. The PSM Paradox gets worse with more covariates.
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \implies \pi_c = \pi_t \]
PSM’s Statistical Properties

1. **Low Standards: Sometimes helps, never optimizes**
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \implies \pi_c = \pi_t \] but
     \[ \pi_c = \pi_t \not\implies X_c = X_t \]
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \implies \pi_c = \pi_t \text{ but } \pi_c = \pi_t \not\implies X_c = X_t \]

2. **The PSM Paradox:** When you do “better,” you do worse
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \implies \pi_c = \pi_t \text{ but } \]
     \[ \pi_c = \pi_t \nRightarrow X_c = X_t \]

2. **The PSM Paradox:** When you do “better,” you do worse
   - Background: Random matching increases imbalance

Problems with Propensity Score Matching
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \implies \pi_c = \pi_t \text{ but} \]
     \[ \pi_c = \pi_t \nRightarrow X_c = X_t \]

2. **The PSM Paradox:** When you do “better,” you do worse
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning)
1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \implies \pi_c = \pi_t \] but
     \[ \pi_c = \pi_t \not\implies X_c = X_t \]

2. **The PSM Paradox:** When you do “better,” you do worse
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) \( \sim \) all \( \hat{\pi} \approx 0.5 \) (or constant within strata)
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \implies \pi_c = \pi_t \text{ but} \]
     \[ \pi_c = \pi_t \notimplies X_c = X_t \]

2. **The PSM Paradox:** When you do “better,” you do worse
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) \( \sim \) all \( \hat{\pi} \approx 0.5 \) (or constant within strata) \( \sim \) pruning at random
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \Rightarrow \pi_c = \pi_t \text{ but} \]
     \[ \pi_c = \pi_t \nRightarrow X_c = X_t \]

2. **The PSM Paradox:** When you do “better,” you do worse
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) \( \sim \) all \( \hat{\pi} \approx 0.5 \) (or constant within strata) \( \sim \) pruning at random \( \sim \) Imbalance
PSM’s Statistical Properties

1. **Low Standards: Sometimes helps, never optimizes**
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     
     \[ X_c = X_t \Rightarrow \pi_c = \pi_t \text{ but} \]
     \[ \pi_c = \pi_t \not\Rightarrow X_c = X_t \]

2. **The PSM Paradox: When you do “better,” you do worse**
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) \( \sim \) all \( \hat{\pi} \approx 0.5 \) (or constant within strata) \( \sim \) pruning at random \( \sim \) Imbalance \( \sim \) Inefficiency
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \implies \pi_c = \pi_t \text{ but } \]
     \[ \pi_c = \pi_t \not\implies X_c = X_t \]

2. **The PSM Paradox:** When you do “better,” you do worse
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) \( \leadsto \) all \( \hat{\pi} \approx 0.5 \) (or constant within strata) \( \leadsto \) pruning at random \( \leadsto \) Imbalance \( \leadsto \) Inefficency \( \leadsto \) Model dependence
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     
     \[ X_c = X_t \implies \pi_c = \pi_t \text{ but} \]
     
     \[ \pi_c = \pi_t \nRightarrow X_c = X_t \]

2. **The PSM Paradox:** When you do “better,” you do worse
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) \( \sim \) all \( \hat{\pi} \approx 0.5 \) (or constant within strata) \( \sim \) pruning at random \( \sim \) Imbalance \( \sim \) Inefficiency \( \sim \) Model dependence \( \sim \) Bias
PSM’s Statistical Properties

1. **Low Standards: Sometimes helps, never optimizes**
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[
     X_c = X_t \implies \pi_c = \pi_t \text{ but } \\
     \pi_c = \pi_t \not\implies X_c = X_t
     \]

2. **The PSM Paradox: When you do “better,” you do worse**
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) \( \leadsto \) all \( \hat{\pi} \approx 0.5 \) (or constant within strata) \( \leadsto \) pruning at random \( \leadsto \) Imbalance \( \leadsto \) Inefficiency \( \leadsto \) Model dependence \( \leadsto \) Bias
   - If the data have no good matches, the paradox won’t be a problem but you’re cooked anyway.
PSM’s Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \Rightarrow \pi_c = \pi_t \text{ but} \]
     \[ \pi_c = \pi_t \not\Rightarrow X_c = X_t \]

2. The PSM Paradox: When you do “better,” you do worse
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) \( \sim \) all \( \hat{\pi} \approx 0.5 \) (or constant within strata) \( \sim \) pruning at random \( \sim \) Imbalance \( \sim \) Inefficency \( \sim \) Model dependence \( \sim \) Bias
   - If the data have no good matches, the paradox won’t be a problem but you’re cooked anyway.
   - Doesn’t PSM solve the curse of dimensionality problem?
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \Rightarrow \pi_c = \pi_t \] but
     \[ \pi_c = \pi_t \not\Rightarrow X_c = X_t \]

2. **The PSM Paradox:** When you do “better,” you do worse
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) \( \approx \) all \( \hat{\pi} \approx 0.5 \) (or constant within strata) \( \approx \) pruning at random \( \approx \) Imbalance \( \approx \) Inefficiency \( \approx \) Model dependence \( \approx \) Bias
   - If the data have no good matches, the paradox won’t be a problem but you’re cooked anyway.
   - Doesn’t PSM solve the curse of dimensionality problem? Nope.
PSM’s Statistical Properties

1. **Low Standards:** Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     \[ X_c = X_t \implies \pi_c = \pi_t \text{ but } \pi_c = \pi_t \not\implies X_c = X_t \]

2. **The PSM Paradox:** When you do “better,” you do worse
   - Background: Random matching increases imbalance
   - When PSM approximates complete randomization (to begin with or, after some pruning) \( \leadsto \) all \( \hat{\pi} \approx 0.5 \) (or constant within strata) \( \leadsto \) pruning at random \( \leadsto \) Imbalance \( \leadsto \) Inefficency \( \leadsto \) Model dependence \( \leadsto \) Bias
   - If the data have no good matches, the paradox won’t be a problem but you’re cooked anyway.
   - Doesn’t PSM solve the curse of dimensionality problem? Nope. The PSM Paradox gets worse with more covariates
PSM is Blind Where Other Methods Can See
PSM is Blind Where Other Methods Can See

Problems with Propensity Score Matching
PSM is Blind Where Other Methods Can See
What Does PSM Match?

MDM Matches

PSM Matches

Controls: $X_1, X_2 \sim \text{Uniform}(0,5)$
Treated: $X_1, X_2 \sim \text{Uniform}(1,6)$
PSM Increases Model Dependence & Bias

Model Dependence

Bias

\[ Y_i = 2T_i + X_{1i} + X_{2i} + \epsilon_i \]
\[ \epsilon_i \sim N(0, 1) \]
The Propensity Score Paradox in Real Data

Finkel et al. (JOP, 2012)

Nielsen et al. (AJPS, 2011)

Problems with Propensity Score Matching
The Propensity Score Paradox in Real Data

Finkel et al. (JOP, 2012)

Nielsen et al. (AJPS, 2011)
The Propensity Score Paradox in Real Data

Finkel et al. (JOP, 2012)

Nielsen et al. (AJPS, 2011)

Similar pattern for > 20 other real data sets we checked

Problems with Propensity Score Matching
Detecting Model Dependence

Matching to Reduce Model Dependence

Three Matching Methods

Problems with Propensity Score Matching

The Matching Frontier
Tensions in Existing Matching Methods

• Maximize one metric; judge against another: Propensity score matching, compared with var-by-var diff in means
• Choose \( n \); check imbalance after: Propensity score matching, Mahalanobis
• Choose imbalance; check \( n \) after: exact matching, CEM
Tensions in Existing Matching Methods

- Maximize one metric; judge against another: Propensity score matching, compared with var-by-var diff in means
Tensions in Existing Matching Methods

- **Maximize one metric; judge against another**: Propensity score matching, compared with var-by-var diff in means
- **Choose \( n \); check imbalance after**: Propensity score matching, Mahalanobis
• **Maximize one metric; judge against another**: Propensity score matching, compared with var-by-var diff in means

• **Choose $n$; check imbalance after**: Propensity score matching, Mahalanobis

• **Choose imbalance; check $n$ after**: exact matching, CEM
A Solution: The Matching Frontier

![Chart showing the relationship between Imbalance and Number of Units Pruned.](image)
A Solution: The Matching Frontier

The Matching Frontier

Number of Units Pruned

Imbalance

Low variance

High variance

Less biased

More biased
A Solution: The Matching Frontier

Result #1

Number of Units Pruned

Imbalance

Low variance High variance

Less biased More biased

Low variance High variance

Number of Units Pruned
A Solution: The Matching Frontier

- Number of Units Pruned
- Imbalance
- Low variance
- High variance
- Less biased
- More biased

● Result #1
● Result #2
A Solution: The Matching Frontier

![Diagram showing the matching frontier with results labeled as Result #1, Result #2, and Result #3. The x-axis represents the number of units pruned, and the y-axis represents imbalance, ranging from less biased to more biased. The results are plotted at different points on the graph, indicating varying levels of variance and bias.]
A Solution: The Matching Frontier

- Number of Units Pruned
- Imbalance
  - Low variance
  - High variance
  - Less biased
  - More biased

Result #1
Result #2
Result #3
Result #4
A Solution: The Matching Frontier

The diagram illustrates the relationship between the number of units pruned and the imbalance, with variance as a factor. The frontier curve shows the trade-off between bias and variance, where:

- **Low variance** results in less biased estimates.
- **High variance** results in more biased estimates.

The points labeled Result #1, Result #2, Result #3, and Result #4 indicate different levels of imbalance and variance, with each point representing a specific condition on the frontier.

The curve suggests that as the number of units pruned increases, the imbalance decreases, moving towards less biased results in scenarios with low variance. Conversely, in high variance scenarios, more units pruned lead to more biased results due to the increased variance.
A Solution: The Matching Frontier

The theoretical frontier (optimal) shows the trade-off between the number of units pruned and imbalance. Results #1 to #4 are plotted on the graph, with Result #4 being the least biased and Result #2 being the most biased. The graph illustrates that as variance decreases (from low to high), the bias also decreases (from more biased to less biased).
A Solution: The Matching Frontier

Number of Units Pruned

Imbalance

Low variance
High variance

Less biased
More biased

Result #1
Result #2
Result #3
Result #4

Theoretical frontier
(optimal)

IMPOSSIBLE REGION

Number of Units Pruned

Low variance
High variance
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of $N$ control units $X_0$
  - Calculate imbalance for all $(N \times n)$ subsets of rows of $X_0$
  - Choose subset with lowest imbalance

- Evaluations needed to compute the entire frontier:
  - $(N \times n)$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) “power set”
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\Rightarrow$ It's hard to calculate!

- We develop algorithms for the (optimal) frontier which:
  - Runs very fast
  - Operate as “greedy” but we prove are optimal
  - Do not require evaluating every subset
  - Work with very large data sets
  - Is the exact frontier (no approximation or estimation)
  - $\Rightarrow$ It's easy to calculate!
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:

  - [Start with matrix of \( N \) control units \( X_0 \)]
  - [Calculate imbalance for all \((N \cdot n)\) subsets of rows of \( X_0 \)]
  - [Choose subset with lowest imbalance]
  - [Evaluations needed to compute the entire frontier:]
    - \((N \cdot n)\) evaluations for each sample size \( n = N, N - 1, \ldots, 1 \)
    - The combination is the (gargantuan) “power set”
      - e.g., \( N > 300 \) requires more imbalance evaluations than elementary particles in the universe
    - \( \Rightarrow \) It's hard to calculate!

- We develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as “greedy” but we prove are optimal
  - do not require evaluating every subset
  - work with very large data sets
  - is the exact frontier (no approximation or estimation)
  - \( \Rightarrow \) It's easy to calculate!
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of $N$ control units $X_0$

  Evaluations needed to compute the entire frontier:
  - $(N \cdot n)$ evaluations for each sample size $n = N, N-1, \ldots, 1$

  The combination is the (gargantuan) "power set"
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe

  $\Rightarrow$ It's hard to calculate!

  We develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as "greedy" but we prove are optimal
  - do not require evaluating every subset
  - work with very large data sets
  - is the exact frontier (no approximation or estimation)

  $\Rightarrow$ It's easy to calculate!
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of $N$ control units $X_0$
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
How hard is the frontier to calculate?

• Consider 1 point on the SATT frontier:
  • Start with matrix of $N$ control units $X_0$
  • Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  • Choose subset with lowest imbalance

The Matching Frontier
How hard is the frontier to calculate?

• Consider 1 point on the SATT frontier:
  • Start with matrix of $N$ control units $X_0$
  • Calculate imbalance for all \( \binom{N}{n} \) subsets of rows of $X_0$
  • Choose subset with lowest imbalance

• Evaluations needed to compute the entire frontier:
  • \( \binom{N}{n} \) evaluations for each sample size $n = N, N-1, \ldots, 1$
  • The combination is the (gargantuan) "power set"
  • e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  • \( \Rightarrow \) It's hard to calculate!

• We develop algorithms for the (optimal) frontier which:
  • runs very fast
  • operate as "greedy" but we prove are optimal
  • do not require evaluating every subset
  • work with very large data sets
  • is the exact frontier (no approximation or estimation)

\( \Rightarrow \) It's easy to calculate!
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of \( N \) control units \( X_0 \)
  - Calculate imbalance for all \( \binom{N}{n} \) subsets of rows of \( X_0 \)
  - Choose subset with lowest imbalance

- Evaluations needed to compute the entire frontier:
  - \( \binom{N}{n} \) evaluations for each sample size \( n = N, N - 1, \ldots, 1 \)

The combination is the (gargantuan) “power set”
- e.g., \( N > 300 \) requires more imbalance evaluations than elementary particles in the universe

\( \hookrightarrow \) It’s hard to calculate!

- We develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as “greedy” but we prove are optimal
  - do not require evaluating every subset
  - work with very large data sets
  - is the exact frontier (no approximation or estimation)

\( \hookrightarrow \) It’s easy to calculate!
How hard is the frontier to calculate?

- **Consider 1 point on the SATT frontier:**
  - Start with matrix of $N$ control units $X_0$
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  - Choose subset with lowest imbalance

- **Evaluations needed to compute the entire frontier:**
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N - 1, \ldots, 1$
  - The combination is the (gargantuan) “power set”
How hard is the frontier to calculate?

• Consider 1 point on the SATT frontier:
  • Start with matrix of $N$ control units $X_0$
  • Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  • Choose subset with lowest imbalance

• Evaluations needed to compute the entire frontier:
  • $\binom{N}{n}$ evaluations for each sample size $n = N, N - 1, \ldots, 1$
  • The combination is the (gargantuan) “power set”
  • e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of $N$ control units $X_0$
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  - Choose subset with lowest imbalance

- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N - 1, \ldots, 1$
  - The combination is the (gargantuan) “power set”
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\nRightarrow$ It’s hard to calculate!
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of $N$ control units $X_0$
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  - Choose subset with lowest imbalance

- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N - 1, \ldots, 1$
  - The combination is the (gargantuan) “power set”
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\Rightarrow$ It’s hard to calculate!

- We develop algorithms for the (optimal) frontier which:
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of $N$ control units $X_0$
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  - Choose subset with lowest imbalance

- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) “power set”
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\implies$ It’s hard to calculate!

- We develop algorithms for the (optimal) frontier which:
  - runs very fast
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of $N$ control units $X_0$
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  - Choose subset with lowest imbalance

- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N - 1, \ldots, 1$
  - The combination is the (gargantuan) “power set”
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\leadsto$ It’s hard to calculate!

- We develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as “greedy” but we prove are optimal
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of $N$ control units $X_0$
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  - Choose subset with lowest imbalance

- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N - 1, \ldots, 1$
  - The combination is the (gargantuan) “power set”
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\approx$ It’s hard to calculate!

- We develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as “greedy” but we prove are optimal
  - do not require evaluating every subset
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of $N$ control units $X_0$
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  - Choose subset with lowest imbalance

- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N - 1, ..., 1$
  - The combination is the (gargantuan) “power set”
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\Rightarrow$ It’s hard to calculate!

- We develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as “greedy” but we prove are optimal
  - do not require evaluating every subset
  - work with very large data sets
How hard is the frontier to calculate?

• Consider 1 point on the SATT frontier:
  • Start with matrix of $N$ control units $X_0$
  • Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  • Choose subset with lowest imbalance

• Evaluations needed to compute the entire frontier:
  • $\binom{N}{n}$ evaluations for each sample size $n = N, N - 1, \ldots, 1$
  • The combination is the (gargantuan) “power set”
  • e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  • $\Rightarrow$ It’s hard to calculate!

• We develop algorithms for the (optimal) frontier which:
  • runs very fast
  • operate as “greedy” but we prove are optimal
  • do not require evaluating every subset
  • work with very large data sets
  • is the exact frontier (no approximation or estimation)
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of $N$ control units $X_0$
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of $X_0$
  - Choose subset with lowest imbalance

- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) “power set”
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\sim$ It’s hard to calculate!

- We develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as “greedy” but we prove are optimal
  - do not require evaluating every subset
  - work with very large data sets
  - is the exact frontier (no approximation or estimation)
  - $\sim$ It’s easy to calculate!
Constructing the FSATT Mahalanobis Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Covariate 1

Covariate 2

-1.0  -0.5   0.0   0.5   1.0

●

Treated

Control

Next to remove

Frontier

Number of Observations Dropped

Average Mahalanobis Discrepancy

0.0   0.1   0.2   0.3   0.4

0   5   10   15   20

The Matching Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Frontier

Covariate 1

Covariate 2

Treated
Control
Next to remove

Number of Observations Dropped
Average Mahalanobis Discrepancy

The Matching Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Frontier

Covariate 1

Covariate 2

○ Treated
○ Control
○ Next to remove

Average Mahalanobis Discrepancy

Number of Observations Dropped

The Matching Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Frontier

- Covariate 1
- Covariate 2

○ Treated
○ Control
○ Next to remove

Average Mahalanobis Discrepancy

Number of Observations Dropped

The Matching Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Frontier

- Treated
- Control
- Next to remove

Number of Observations Dropped

Average Mahalanobis Discrepancy

Covariate 1

Covariate 2

-1.0 -0.5 0.0 0.5 1.0

-1.0 -0.5 0.0 0.5 1.0

0.0 0.1 0.2 0.3 0.4

0.0 0.1 0.2 0.3 0.4

0 5 10 15 20

0 5 10 15 20

The Matching Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Frontier

The Matching Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Covariate 1

Covariate 2

-1.0 -0.5 0.0 0.5 1.0

●

Treated
Control
Next to remove

Frontier

Average Mahalanobis Discrepancy
0.0 0.1 0.2 0.3 0.4

Number of Observations Dropped
0 5 10 15 20
Constructing the FSATT Mahalanobis Frontier

Remaining Data

- Covariate 1
- Covariate 2
- Treated
- Control
- Next to remove

Frontier

- Average Mahalanobis Discrepancy
- Number of Observations Dropped
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Covariate 1
Covariate 2

-1.0 -0.5 0.0 0.5 1.0

●
●
●
●
●
●
●

Treated
Control
Next to remove

Frontier

Average Mahalanobis Discrepancy

0.0
0.1
0.2
0.3
0.4

Number of Observations Dropped

0 5 10 15 20
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Frontier

The Matching Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Covariate 1

Covariate 2

-1.0 -0.5 0.0 0.5 1.0

-1.0

-0.5

0.0

0.5

1.0

●

●

●

●

●

●

●

●

●

●

●

●

Next to remove

Frontier

Number of Observations Dropped

Average Mahalanobis Discrepancy

0.0

0.1

0.2

0.3

0.4

0 5 10 15 20
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Frontier

- Number of Observations Dropped
- Average Mahalanobis Discrepancy

- Treated
- Control
- Next to remove

The Matching Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

- Covariate 1
- Covariate 2
- Treated
- Control
- Next to remove

Frontier

- Average Mahalanobis Discrepancy
- Number of Observations Dropped
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Frontier

The Matching Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

Frontier

Covariate 1

Covariate 2

-1.0 -0.5 0.0 0.5 1.0

-1.0

-0.5

0.0

0.5

1.0

●

●

●

●

●

●

●

●

Treated

Control

Next to remove

Number of Observations Dropped

Average Mahalanobis Discrepancy

0 5 10 15 20

0.0

0.1

0.2

0.3

0.4

The Matching Frontier
Constructing the FSATT Mahalanobis Frontier

Remaining Data

○ Treated
○ Control
○ Next to remove

Frontier

Average Mahalanobis Discrepancy

Number of Observations Dropped
Constructing the FSATT Mahalanobis Frontier

Warning: figure omits details and proof!
Discrete algorithm

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

• Calculate bins
• Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:
- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

• Calculate bins
• Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

• Calculate bins
• Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:
- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units

The Matching Frontier
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Discrete algorithm

Short version:

- Calculate bins
- Until balance stops improving, greedily prune a control unit from the bin with the largest proportional difference between control and treated units
Job Training Data: Frontier and Causal Estimates

- 185 Ts; pruning most 16,252 Cs won’t increase variance much
- Huge bias-variance trade-off after pruning most Cs
- Estimates converge to experiment after removing bias
- No mysteries: basis of inference clearly revealed