Synthetic Interventions

Dennis Shen

Joint work with

Anish Agarwal  Devavrat Shah
Bridging causal inference & machine learning

1. Synthetic Interventions
   [A. Agarwal, D. Shah, D. Shen]
   (synthetic controls $\rightarrow$ tensor completion)

2. Causal Matrix Completion
   [A. Agarwal, M. Dahleh, D. Shah, D. Shen]
   (synthetic controls $\rightarrow$ matrix completion)
Clinical trial study w. Alzheimer’s Therapeutics company

- 2 year study
- 1000+ subjects
- 4 therapies (1 placebo)
Alzheimer's clinical trial study

**Inconclusive**

Average treatment effect for all 3 therapies was insignificant

**Costly**

Total cost of trial: $500M - $1B USD (cost of recruiting one patient: $5k – $100k USD)

Ethical concerns of testing on human subjects

**Question we set out to answer**

Maybe therapies were effective for subset of patients?

Can we estimate ADAS-COG score for each patient under each therapy?
A question:
Can we estimate ADAS-COG score for each patient under each therapy?

A framework:
Causal inference as tensor completion
Potential outcomes: a tensor viewpoint

Potential outcomes: $Y_{nt}^{(d)}$

What if patient $n$ on visit $t$ had been assigned therapy $d$ [Neyman’23, Rubin’74]

(Alice’s ADAS-COG score on visit $t$ under therapy 1)

(Alice’s ADAS-COG score on visit $t$ under therapy 2)

(Alice’s ADAS-COG score on visit $t$ under placebo)

(Alice’s ADAS-COG score on visit $t$ under therapy 3)
What data we had from

Pre-RCT baseline of ALL patients under placebo
(outcomes under therapy 1, therapy 2, therapy 3 missing)

Post-PCT of patients that received therapy 3
(outcomes under placebo, therapy 1, therapy 2 missing)

Fundamental challenge of causal inference:
only observe one outcome
want to know all potential outcomes
RCTs – randomization but no personalization

What RCT estimates
Average treatment effect

\[
\text{Avg}(\text{Therapy 1}) - \text{Avg}(\text{placebo})
\]

What RCT cannot estimate
Individual treatment effect

Why? Can only observe Alice under ONE intervention

Limitation of RCTs
What works best on average may not work best for each individual

Why are RCTs beloved?
Explicit randomization

Randomization but NO personalization
Counterfactual estimation = Tensor Completion

"causal inference is a missing data problem" vis-à-vis
"tensor completion is a missing data problem"

“What if placebo subjects received therapy 2?"

Same questions, different language

<table>
<thead>
<tr>
<th>Causal Inference</th>
<th>Tensor Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>causal estimand</td>
<td>error metric (norm)</td>
</tr>
<tr>
<td>confounded data</td>
<td>missing not at random data</td>
</tr>
<tr>
<td>observational &amp; experimental studies</td>
<td>sparsity patterns</td>
</tr>
</tbody>
</table>
Alzheimer's clinical trial study w. TauRx Therapeutics

A question:
Can we estimate ADAS-COG score for each patient under each therapy?

A framework:
Causal inference as tensor completion

An answer:
Synthetic interventions (SI)
Bob and Alice: counterfactuals of interest

Suppose Bob received therapy 1 after RCT (under placebo prior to RCT)

What if
Bob remained under placebo?
Bob received therapy 2?
Bob received therapy 3?

Suppose Alice remained under placebo after RCT (under placebo prior to RCT)

What if
Alice received therapy 1?
Alice received therapy 2?
Alice received therapy 3?
A partial answer: synthetic controls (SC) [Abadie et al ‘03, ‘10]

Estimates counterfactuals in \textit{absence} of intervention

Model learning:

$$\hat{\beta} = \text{Convex}(y_{\text{pre}}, Y_{\text{donors}})$$

Counterfactual prediction:

$$\hat{y}_{\text{post}} = Y_{\text{donors \ post}} \hat{\beta}$$

Pre-intervention data \textit{NOT} available for therapy

Pre-intervention data \textit{available} for placebo

“Results on estimation with multiple interventions are absent in the literature” [Abadie’20]
What if Alice got therapy 1, therapy 2, therapy 3?

Necessary for clinical case study

What if Bob remained under placebo?

Estimate counterfactual if policy did not occur:
- Police reform [Rydberg’18]
- Brexit [Opatrny’19]
- Tax legislation [Abadie et al’19]

“Results on estimation with multiple [in presence of] interventions are absent in the literature”
[Abadie’20]

“One of the most important innovations in the policy evaluation literature in the last 15 years”
[Athey and Imbens’16]

Significant impact of estimating counterfactuals in presence of intervention
A full answer: synthetic interventions [Agarwal, Shah, Shen ‘21]

Estimates counterfactuals in absence & presence of intervention

Model learning:
\[ \hat{\beta} = \text{PCR}(y_{\text{pre}}, Y_{\text{donors}}) \]

Counterfactual prediction:
\[ \hat{Y}_{\text{post}} = Y_{\text{post}} \hat{\beta} \]

"What if Alice received therapy 2?"

NO additional data over RCT data used!
1. Where model is applied

When can a linear model be transferred across interventions? i.e., transfer learning, distribution shift, causal transportability...

2. How model is learned

Convex regression

$$\hat{\beta} = \text{Convex}(y_{\text{pre}}, Y_{\text{pre}}^{\text{donors}})$$

Principal component regression (PCR)

$$\hat{\beta} = \text{PCR}(y_{\text{pre}}, Y_{\text{pre}}^{\text{donors}})$$

PCR is crucial to proving our formal statistical guarantees
When does synthetic interventions work?

Causal framework, statistical guarantees
Essential questions

learn under control

predict under intervention

When can a linear model be transferred between different interventions?

What type of confounding is allowed in observational data?
Why linear?—*low rank tensor*

\[ Y = \sum_{\ell=1}^{r} \mathbf{u}_{\ell} \otimes \mathbf{v}_{\ell} \otimes \mathbf{w}_{\ell} + \mathbf{\varepsilon} \]

- **Low rank**
- **Time latent factors**
- **Unit latent factors**
- **Intervention latent factors**
- **Stochasticity**
Low rank implies linear span inclusion

Produce counterfactuals for unit n under intervention d

\[ u_n = \sum_{j \in I^{(d)}} \beta_j^* \cdot u_j \]

We require:

\[ I^{(d)} = \text{units under intervention } d \]

Holds w.h.p if factors sampled independently

 Exists \textbf{invariant} linear model across time, interventions
What type of confounding?—*selection on latent factors*

\[ Y = \sum_{\ell=1}^{r} u_\ell \otimes v_\ell \otimes w_\ell + \varepsilon \]

(Kallus ‘18, Athey ‘21)

\[ \mathcal{I} \perp \varepsilon \mid \mathcal{LF} \]

(latent factors)

\[ \mathcal{I} \]

(intervention assignments)

\[ \varepsilon \]

(stochasticity)

\[ \mathcal{LF} \]

(potential outcomes tensor)
Identification

Model learning:

\[ \hat{\beta} = \text{PCR}(y_{\text{pre}}, Y_{\text{donors pre}}) \]

Counterfactual prediction:

\[ \hat{y}_{\text{post}} = Y_{\text{donors post}} \hat{\beta} \]

\[ \mathbb{E}[y_{\text{post}}] = \mathbb{E}[X_{\text{post}}] \cdot \beta^* \]
When transferrable?—*subspace inclusion*

$$\text{complexity}(X_{\text{post}}) \leq \text{complexity}(X_{\text{pre}})$$

“test” set

“train” set

$\hat{\beta}$

$\hat{\beta}$

$Y_{\text{pre}}$

$Y_{\text{post}}$

$Y_{\text{donors pre}}$

$Y_{\text{donors post}}$
When transferrable?—*subspace inclusion*

\[
\text{colspan}(X_{\text{post}}) \subseteq \text{colspan}(X_{\text{pre}})
\]

"test" set \quad \text{hypothesis test} \quad \text{"train" set}
Putting it all together

When does SI work?

Low-rank tensor factor model

$u_{nl} = \sum_{j \in I^{(d)}} \beta^*_j \cdot u_{j l}$

linear span inclusion

$Y^{(d)}_{nt} \perp \mathcal{A} \mid \mathcal{LF}$

selection on latent factors

$\text{colspan}(\mathbb{E}[Y_{\text{pre}}^{\text{donors}}]) \supset \text{colspan}(\mathbb{E}[Y_{\text{post}}^{\text{donors}}])$

subspace inclusion
Model estimation

Model learning:

\[ \hat{\beta} = \text{PCR}(y_{\text{pre}}, Y_{\text{donors, pre}}) \]

\[ \|\hat{\beta} - \beta^*\|_2 = o(1) \]

minimum norm model
Consistency

Counterfactual prediction:

\[ \hat{y}_{\text{post}} = Y_{\text{donors}} \hat{\beta} \]

\[ |\text{avg}(\hat{y}_{\text{post}}) - \text{avg}(\mathbb{E}[y_{\text{post}}])| = o(1) \]
Asymptotic normality

Counterfactual prediction:

\[ \hat{y}_{\text{post}} = Y_{\text{donors}} \hat{\beta} \]

\[ \text{avg}(\hat{y}_{\text{post}}) \sim \mathcal{N}(\text{avg}(\mathbb{E}[y_{\text{post}}]), \sigma^2(\beta^*)) \]

enables confidence intervals
Validation study setup

Training data

Test data
Validation study setup

Training data

Test data
Validation study results

Graph showing the validation study results for different matching methods: Naive, Matching, and SI.
Importance of subspace inclusion

- Train and test data obey **different distributions**
- Subspace inclusion **holds**

- Train and test data obey **same distribution**
- Subspace inclusion **fails**
Additional case studies

**Ride-sharing**

Question:
Passenger wait time if U.S. cities adopted routing policies used in Brazil?

Estimated answer:
Larger U.S. cities would see greatest increase in wait time

observational

**COVID-19 policy evaluation**

Question:
U.S. COVID-19 mortality rates if stricter social distancing adopted?

Estimated answer:
> 5x drop in mortality rates if mobility restriction increased by approx. 30%

observational

**Poverty Action Lab @ MIT**

Question:
Uptake in childhood immunization rates if personalized policies used across villages in India?

Estimated answer:
~6x uptake in immunization if policy personalized to each village

limited experimental
SI—key takeaway

provably learns all $N \times D$ causal estimands with
(i) $N \times 2$ observations (requires meas. under common intervention)
(ii) confounded data that respects selection on latent factors
Why PCR?

Intuition:
Low-rank signal is spectrally concentrated, noise is spectrally diffused

1. PCR enforces spectral sparsity in $Y_{\text{donors}}^{\text{pre}}$
2. PCR implicitly de-noises $Y_{\text{donors}}^{\text{pre}}$
Principal Component Regression

First generalization error result for error-in-variables regression (+ high-dimensional, fixed design setting)

Causal inference

Novel framework: causal inference as tensor completion

Synthetic Interventions

Addressed unresolved problem of counterfactuals in presence of interventions

[Abadie’20]

Novel results for synthetic controls literature itself
Causal tensor completion—looking forward

Causal inference

What \((\text{estimand, confounding, study})\) combinations allow identification?

Tensor completion

What \((\text{error metric, sparsity pattern})\) combinations allow completion?

modeling

If achievable, what are the \text{computational/statistical trade-offs}?

algorithmic
THANK YOU

dshen24@berkeley.edu

Acknowledgements:

Alberto Abadie, Abdullah Alomar, Romain Cosson, Esther Duflo, Anna Mikusheva, Rahul Singh