pensynth

Easier Causal Inference through Fast Penalized Synthetic Control Estimation

Erik-Jan van Kesteren

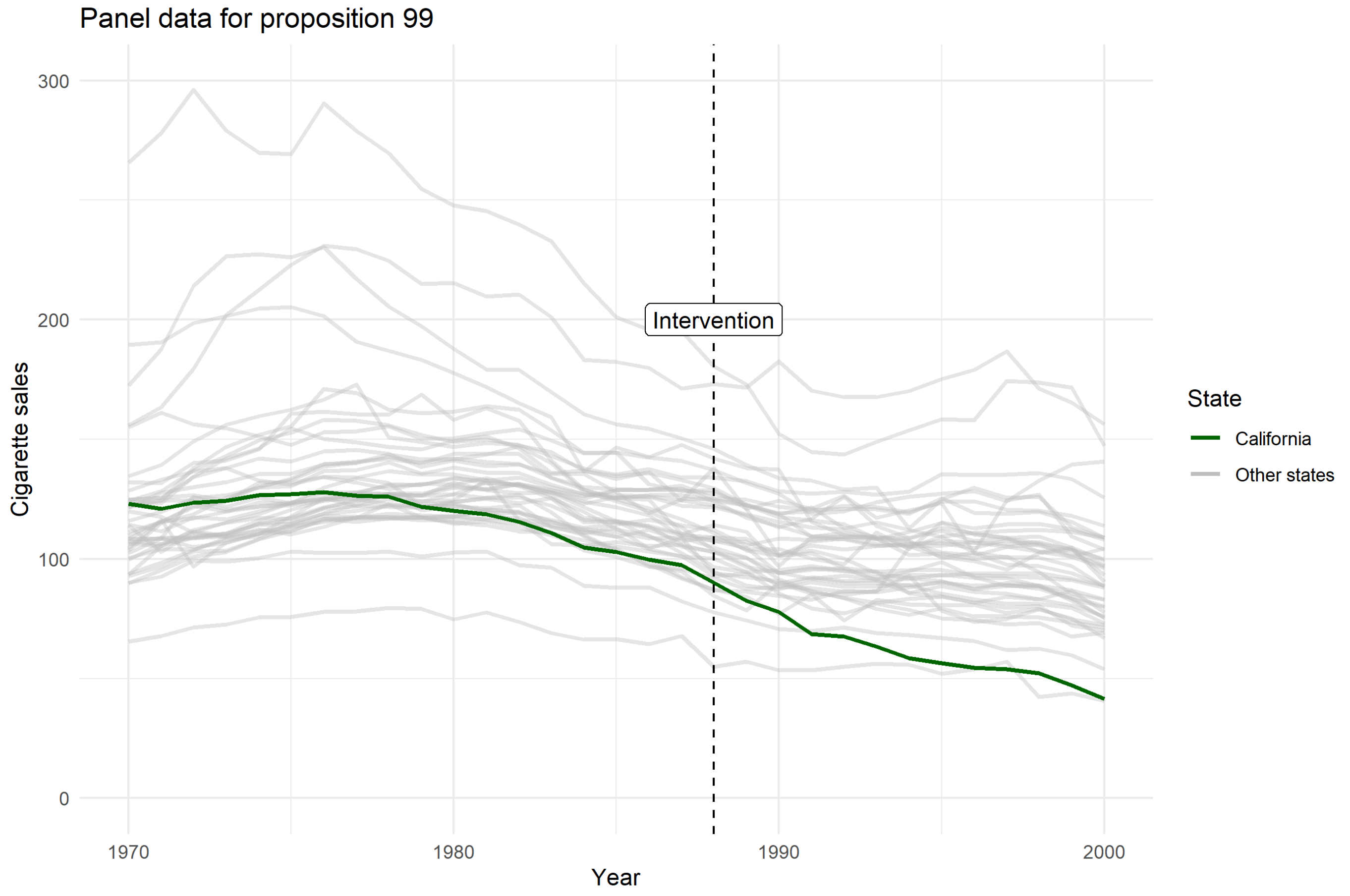
Outline

- Context: policy evaluation
- The synthetic control method at light-speed
- Synthetic control weights ill-defined
- Penalized synthetic controls
- Showcase of the package
- Conclusion

Setting the stage

Policy evaluation setting

- Follow a unit over time $1 \leq t \leq T$
- There is a policy intervention at T_0
- What is the causal effect θ_t of the intervention (at time $t > T_0$)?



Policy evaluation setting

- We measure an outcome Y
- Causal estimand: treatment effect at time $t > T_0$

$$\theta_t = Y_t^1 - Y_t^0$$

- Fundamental problem of causal inference:
  - for $t \leq T_0$ observe only Y_t^0
  - for $t > T_0$ observe only Y_t^1 (!!)

t	Y_t	Y_t^0	Y_t^1
1	7	7	NA
2	9	9	NA
3	6	6	NA
4	5	5	NA
T_0	6	6	NA
6	2	NA	2
7	3	NA	3
8	1	NA	1
…	…	…	…
T	2	NA	2

The problem of estimating the effect of a policy intervention is equivalent to the problem of estimating Y_t^0

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. Journal of Economic Literature, 59(2), 391-425.

Counterfactual estimators

- Now we're doing counterfactual estimation
- There are about a million of these methods

Matrix completion, fixed effects models, matching, diff-in-diff, standard imputation methods, (Bayesian structural) time-series models, …

The Synthetic Control Method

Synthetic controls

- Synthetic control: use J “donor units” to estimate Y_t^0
- different states, different schools, different persons which **did not receive the intervention**
- Let’s call their outcomes C_{jt} for donor unit j at time t, then we only need to compute the following weighted sum:

$$\hat{\theta}_t = Y_t^1 - \hat{Y}_t^0 = Y_t^1 - \sum_{j \in J} C_{jt} w_j$$

The synthetic control!

t	Y_t	Y_t^0	Y_t^1	C_{1t}	C_{2t}	…	C_{jt}
1	7	7	NA	2	9	…	6
2	9	9	NA	6	9	…	8
3	6	6	NA	4	3	…	5
4	5	5	NA	2	1	…	4
T_0	6	6	NA	1	2	…	7
6	2	NA	2	3	6	…	7
7	3	NA	3	2	5	…	6
8	1	NA	1	4	6	…	5
…	…	…	…	…	…	…	4
T	2	NA	2	3	4	…	6

Estimating the weights

- Which combination of donor units is the best approximation of the true Y_t^0?

- Original synthetic control method says:
  - Synth. control should “look like” intervened unit at $t \leq T_0$
  - Avoid extrapolation

Estimating the weights

- Collect P variables about intervened unit in a column vector ($X1$), and the same variables about the donor units in a $P \times J$ matrix ($X0$)
  - E.g., state size, average income, demographics, a selection or summary of pre-intervention outcomes

(there are loads of discussions what to include here)

- Then, estimate weights such that $||X1 - X0w||_2^2$ is as small as possible (THIS IS ORDINARY LEAST SQUARES REGRESSION)

Estimating the weights

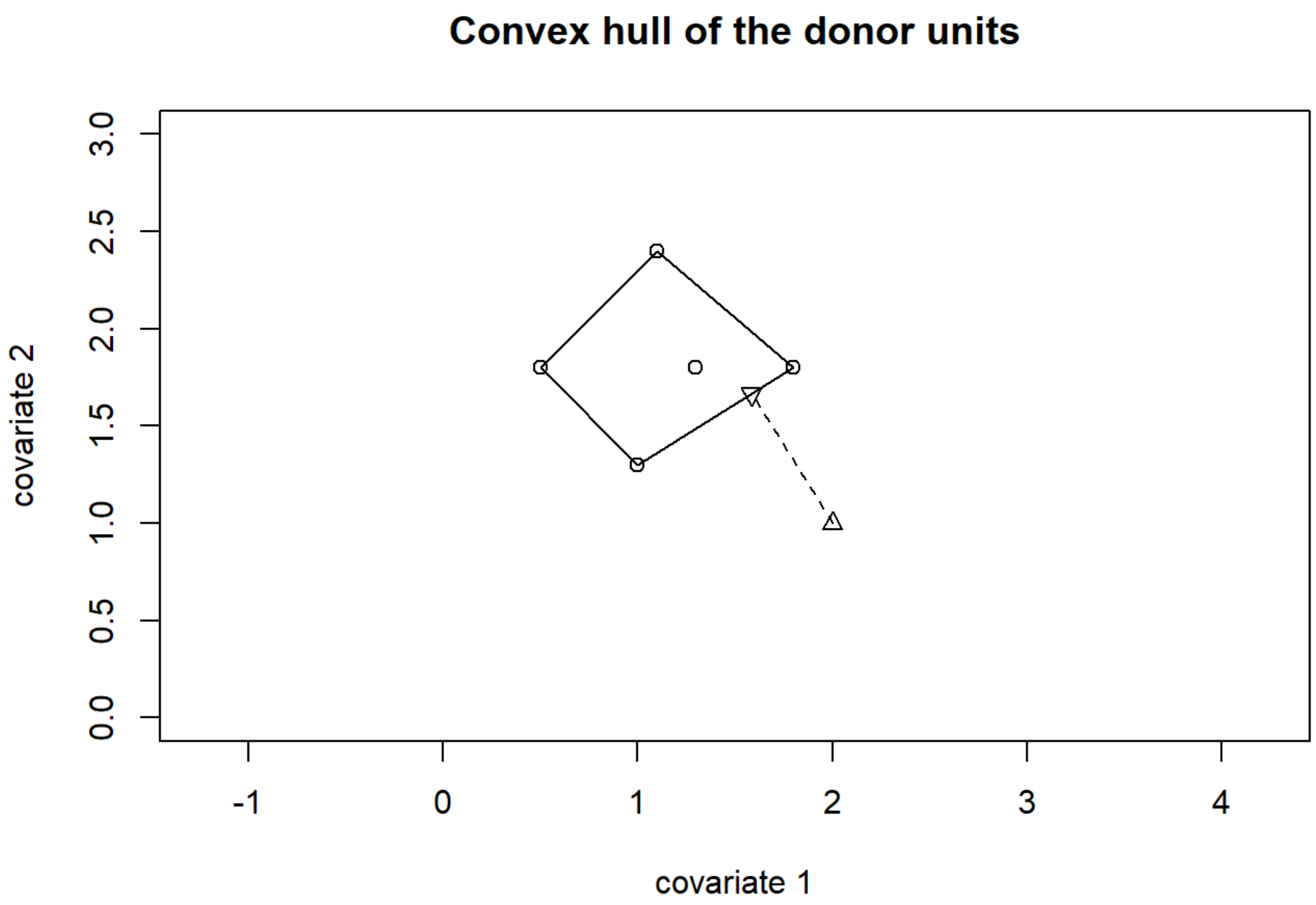
- But this is high-dimensional regression: weights not unique
- If $J \geq P$ there are infinitely many solutions where
$$||X1 - X0w||_2^2 = 0$$
- We need additional information to determine which combination of donors is best

(you can do sparse linear regression like LASSO, regularized horseshoe priors, adaptive LASSO, or other interesting things)

Convex hull constraint

- SCM additional constraint: no extrapolation
- Ensures that synthetic control unit “could plausibly exist”
- Convex hull condition, ensuring:
  - $w_j \geq 0$
  - $\sum_{j \in J} w_j = 1$

- Constrained OLS, solved using a quadratic program


Convex hull of the donor units
covariate 2
covariate 1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-1
0
1
2
3
4

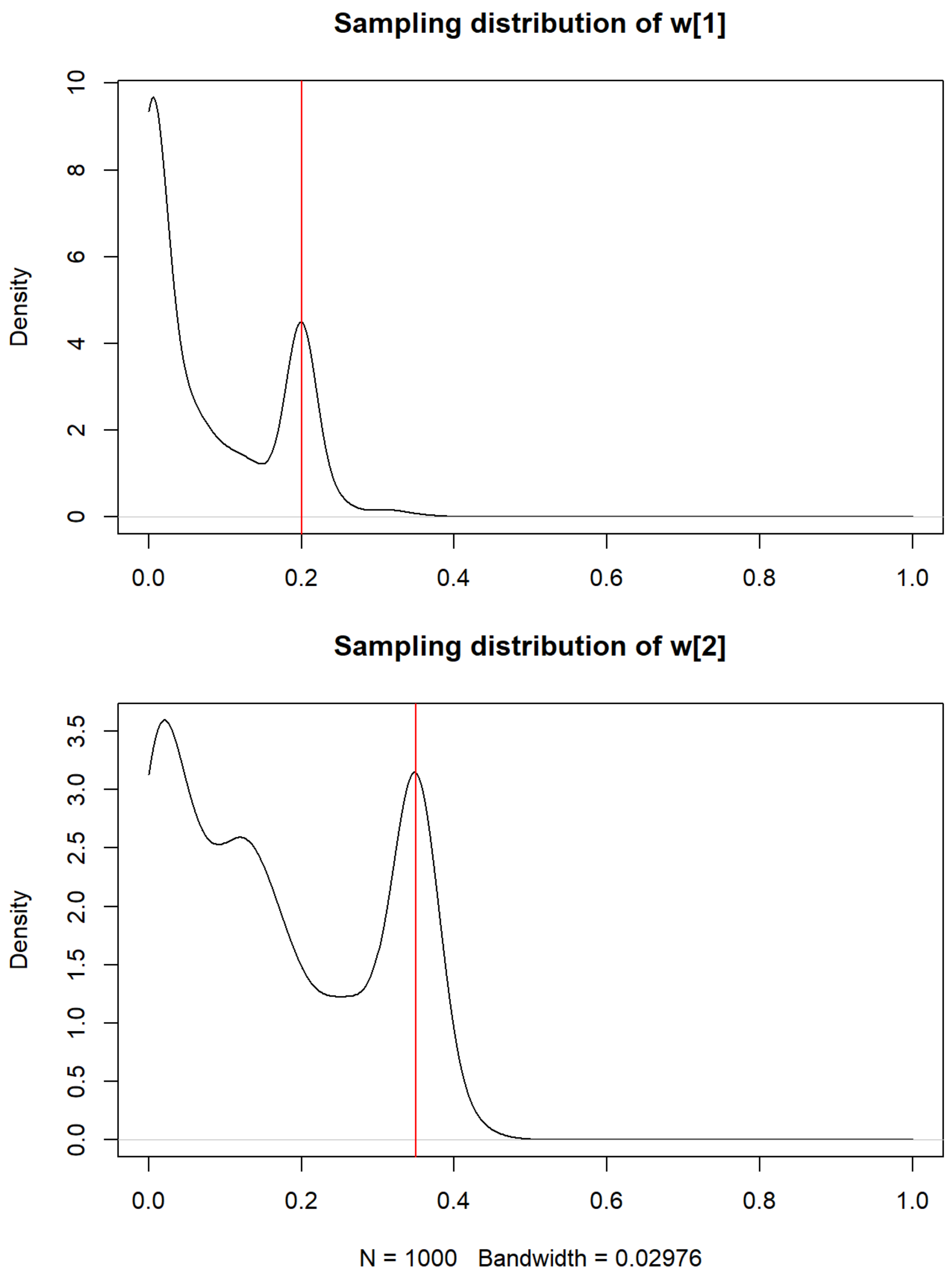
Estimating the weights

- This constraint creates sparsity as well!
- If treated unit is outside convex hull, we will get
  - $\|X1 - X0w\|_2^2 \geq 0$
  - $\sum I(w_j > 0) = P$

- Wonderful! Interpretable synthetic control!

Simple simulation

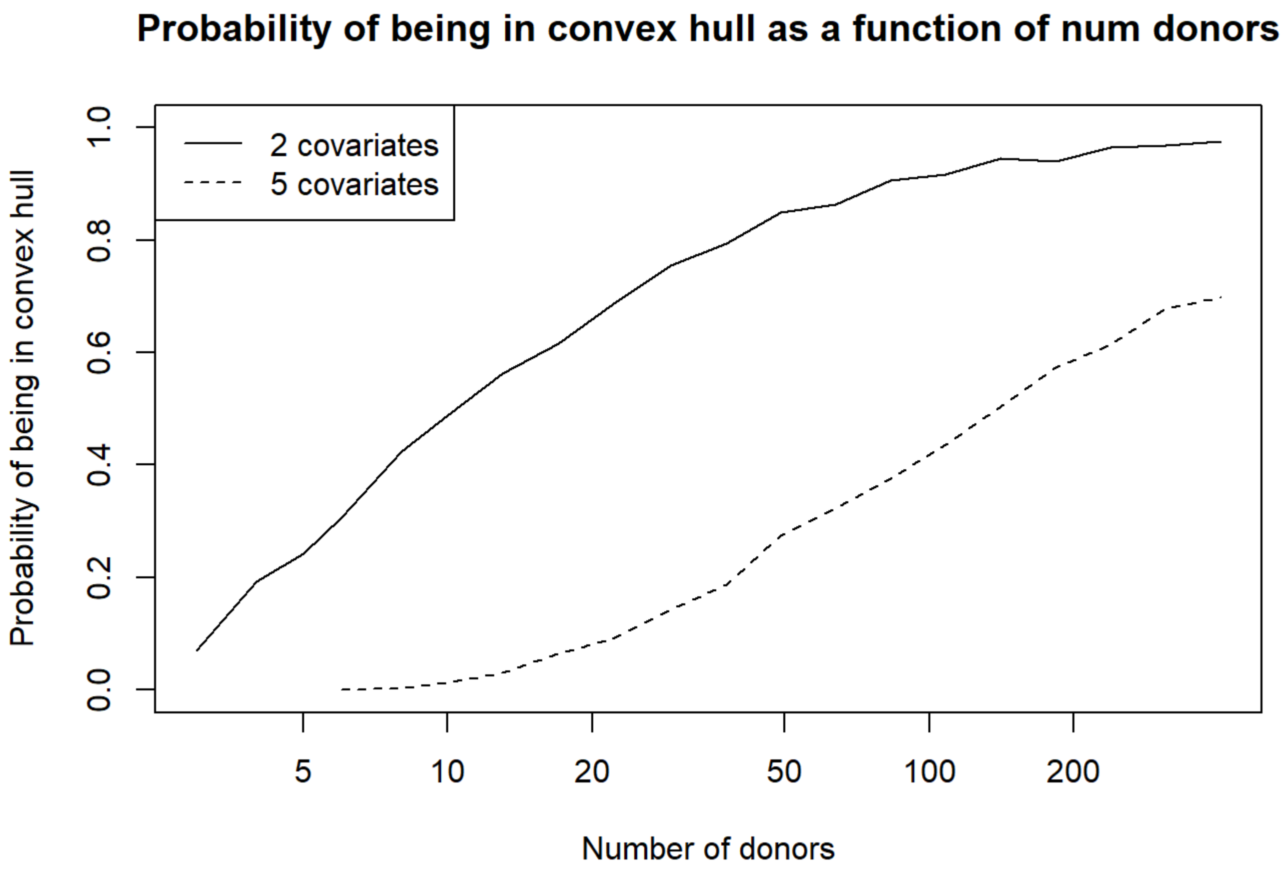
- Random normal data for $X1$ and $X0$
- $w = [.20, .35, .45, 0, 0, \ldots, 0]$
- $J = 50, P = 7$
- Use synthetic control to estimate weights
- What's happening to my weights!?!?



We are in the convex hull

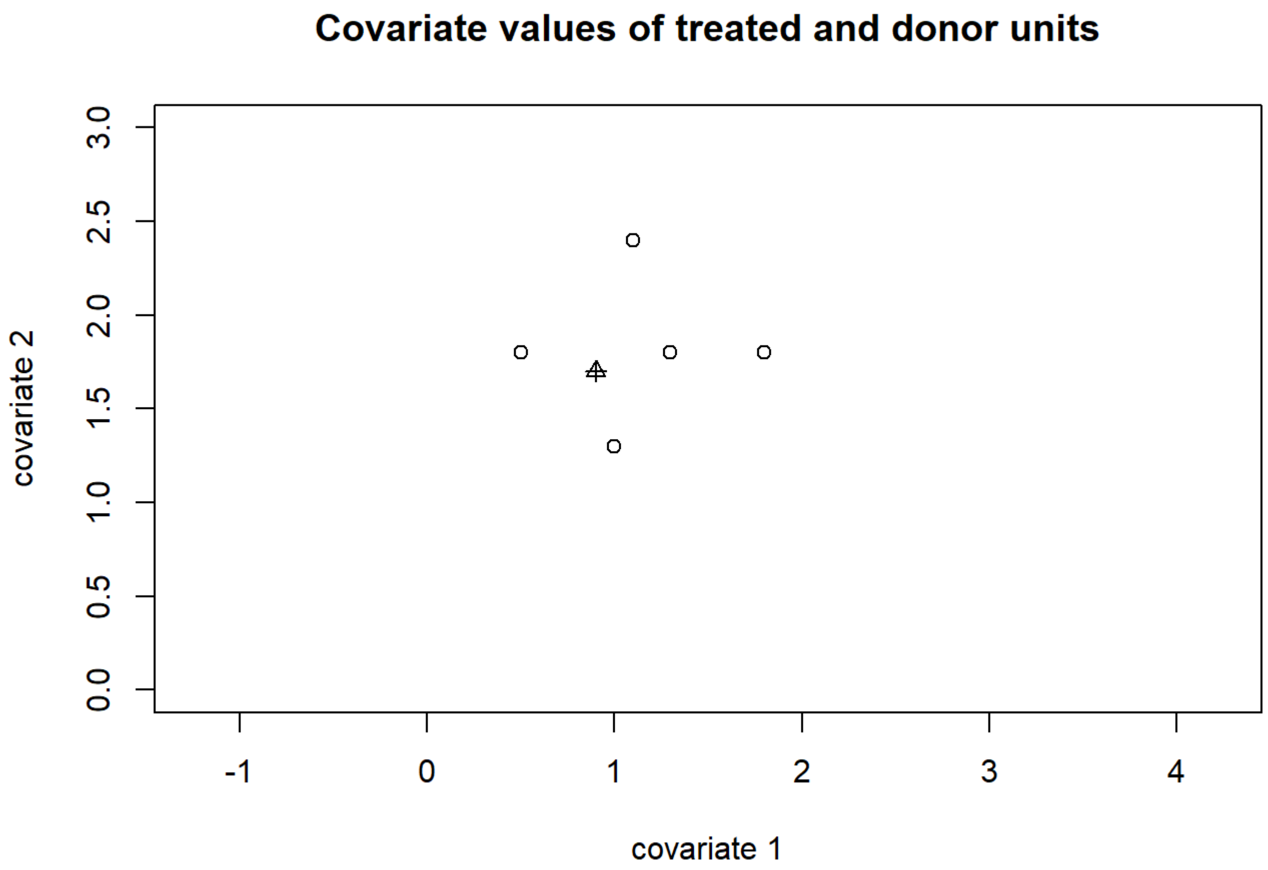
Probability of being in convex hull

- With more donors, P(in convex hull) increases
- In one application, I had >3000 donors (Dutch schools)
- This is common in studies with register data



Problem in convex hull

- Again, infinitely many solutions where $\|X1 - X0w\|_2^2 = 0$
- Even with the additional SCM constraint
- So we need something else



Penalized synthetic control

Prefer nearest neighbours (Abadie & L'Hour, 2021)

Minimize

$$\|X1 - X0w\|_2^2 + \lambda \cdot \sum_{j \in J} w_j \|X1 - X0_j\|_2^2$$

subject to

$$w_j \geq 0 \; \forall \, j \in J, \qquad \sum_{j \in J} w_j = 1$$

Penalized SCM: advantages

- Weights are well-defined always (!)
- Smoothly interpolate between synthetic control and NN matching
  - When $\lambda = 0$ pensynth equals synthetic control
  - When $\lambda = \infty$ pensynth equals nearest neighbour matching
- Can deal with multiple treated units

pensynth

pensynth Public

Pin | Unwatch 2 | Fork 1 | Starred 4

main | 2 Branches | 5 Tags | Go to file | Add file | Code

vankesteren Merge pull request #8 from vankesteren/ar1sim 331af2b · 5 months ago 60 Commits

.github	initial commit	26 days ago
R	Slightly reparameterize rarnorm checks	5 months ago
img	Update readme for new data simulation function	5 months ago
man	Update rarnorm.Rd	5 months ago
tests	set default AR parameter in simulation to 0.8	5 months ago
.Rbuildignore	Update description for CRAN submission	9 months ago
.gitignore	initial commit	26 days ago
DESCRIPTION	Bump version, test, check, document	5 months ago
LICENSE	initial commit	26 days ago
LICENSE.md	initial commit	26 days ago
NAMESPACE	Fix R CMD CHECK	9 months ago
README.md	Update readme for new data simulation function	5 months ago
pensynth.Rproj	Bump version, test, check, document	5 months ago

README | License | MIT license

Penalized synthetic control estimation

R-CMD-check passing | r-universe 0.6.0 | CRAN 0.5.1 | repo status Active

The goal of `pensynth` is to make it easier to perform penalized synthetic control in the spirit of Abadie & L'Hour (2021).

About

Penalized synthetic control estimation

Readme
Unknown, MIT licenses found
Activity
4 stars
2 watching
1 fork

Releases 5

CRAN release 0.5.1 Latest
on Mar 29

+ 4 releases

Packages

No packages published
Publish your first package

Contributors 2

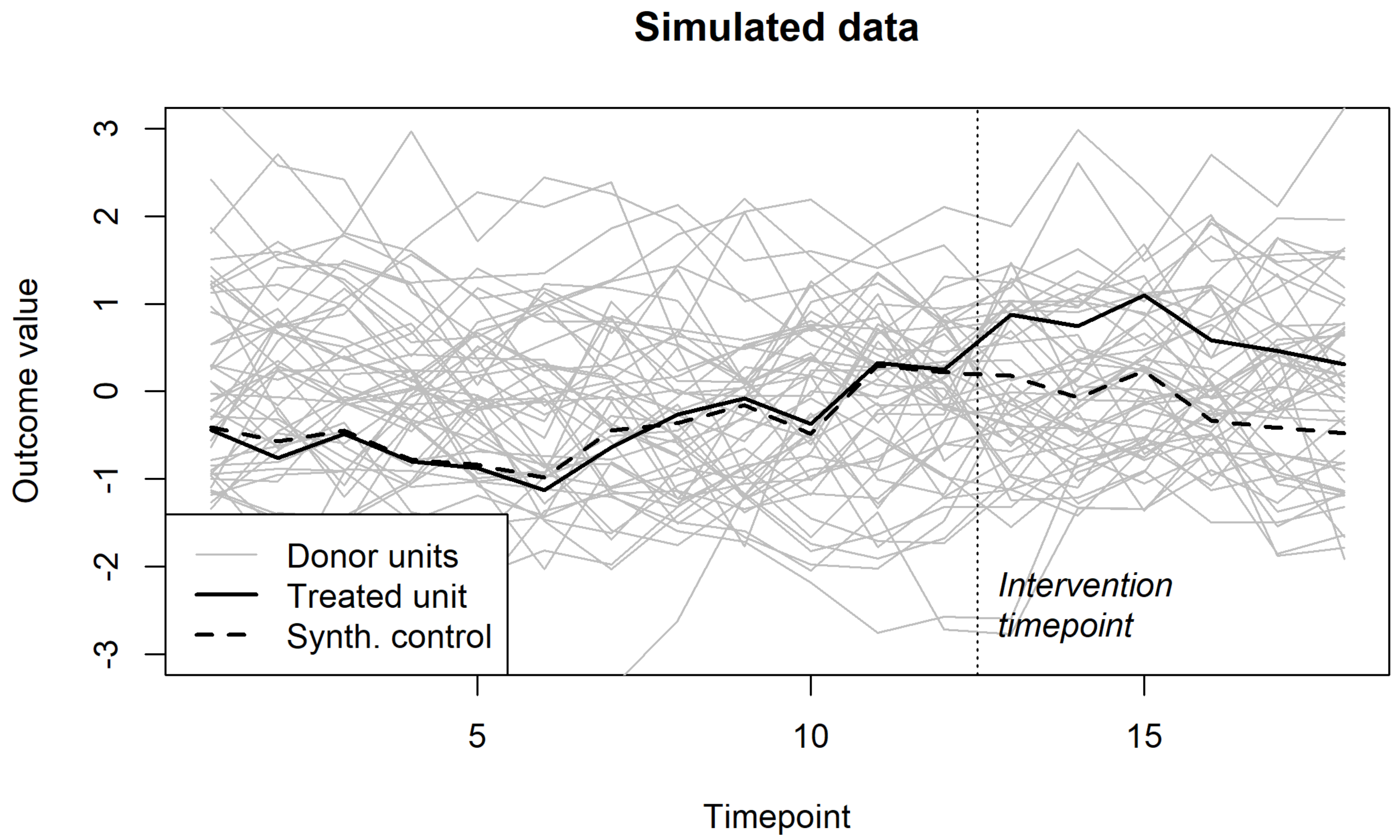
vankesteren Erik-Jan van Kesteren
isaacOnline Isaac Slaughter

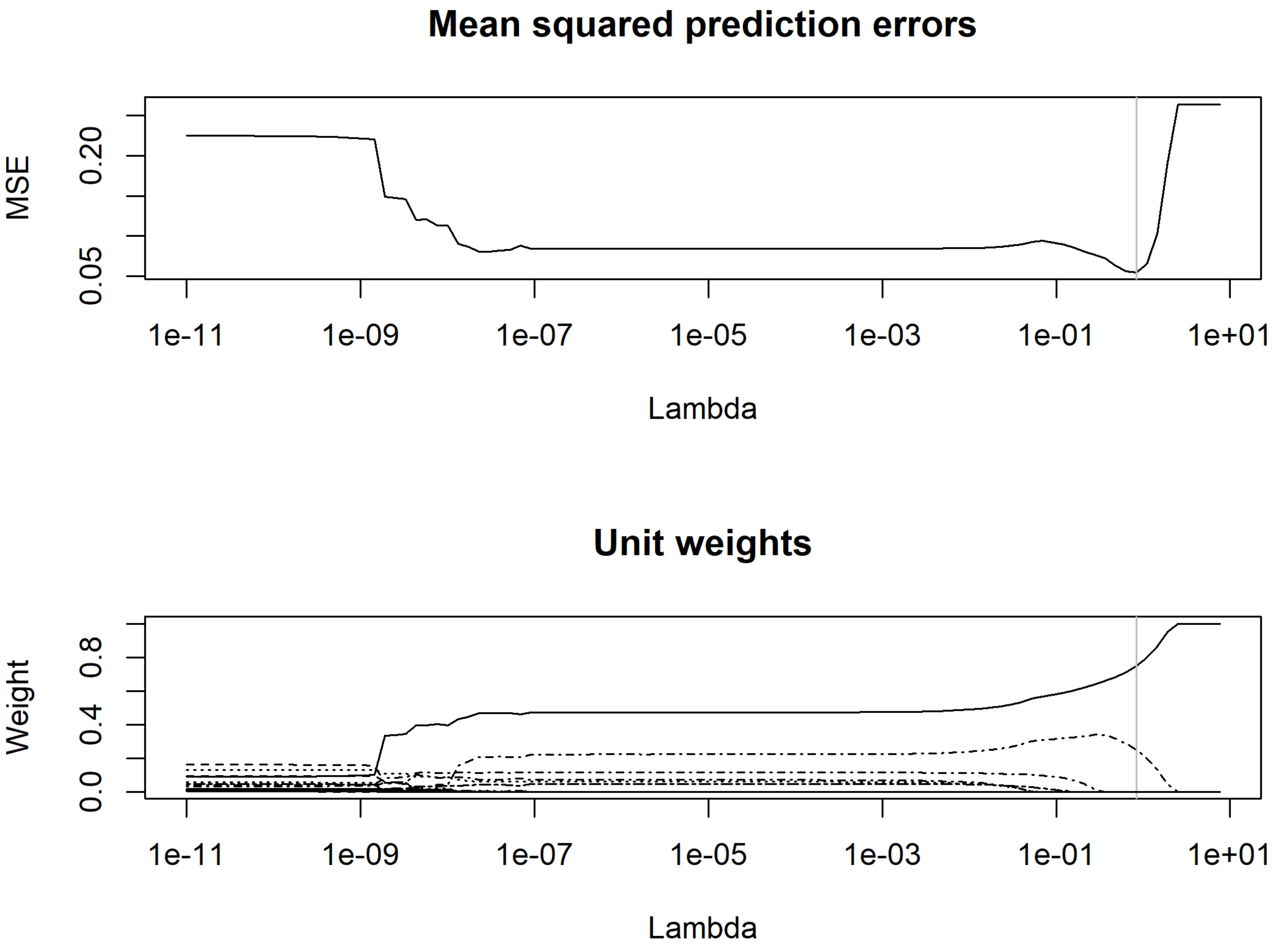
Languages

R 100.0%

pensynth R package

- Efficient and fast implementation
  - Using state-of-the art QP solver *clarabel*, written in Rust
  - Using sparse matrices for the constraints
  - Handles hundreds-thousands of donor units with ease
- Easy-to-use and convenient, understandable
  - Plotting, summarization, nice methods
  - Simulated data built-in
  - Hold-out validation on pre-intervention outcome for tuning λ
- On CRAN: https://doi.org/10.32614/CRAN.package.pensynth


Simulated data
Outcome value
3
2
1
0
-1
-2
-3
5
10
15
Timepoint
Donor units
Treated unit
Synth. control
Intervention
timepoint



Mean squared prediction errors
MSE
0.05
0.20
1e-11
1e-09
1e-07
1e-05
1e-03
1e-01
1e+01
Lambda
Unit weights
Weight
0.0
0.4
0.8
1e-11
1e-09
1e-07
1e-05
1e-03
1e-01
1e+01
Lambda

Conclusion

- Synthetic control is popular for policy evaluation
- Counterfactual estimation method
- It is ill-defined when the treated unit is in the convex hull
- Penalized synthetic control helps with this

Future work:

- Enable multiple simultaneous treated units
- Temporal cross-validation for hyperparameter tuning
- Inference through conformal prediction intervals
  (some bootstrap parts in there too?)
- Formalize when exactly this is “better” and why

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of economic literature*, *59*(2), 391-425.

Abadie, A., & L'hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, *116*(536), 1817-1834.

Ben-Michael, E., Feller, A., & Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, *116*(536), 1789-1803.

Custers, G., van Kesteren, E.-J., & Ryan, O. (2024). The limited impact of extending the school week on educational outcomes in Dutch primary education: a quasi-experimental study using penalized synthetic control analysis. https://doi.org/10.31235/osf.io/h4m7p

Doudchenko, N., & Imbens, G. W. (2016). *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis* (No. w22791). National Bureau of Economic Research.

Kellogg, M., Mogstad, M., Pouliot, G. A., & Torgovitsky, A. (2021). Combining matching and synthetic control to tradeoff biases from extrapolation and interpolation. *Journal of the American statistical association*, *116*(536), 1804-1816.

Van Kesteren, E.-J., and Slaughter, I. (2024). *pensynth: Penalized Synthetic Control Estimation.* R package version 0.5.1 https://doi.org/10.32614/CRAN.package.pensynth

Van Kesteren, E.-J. (2024) The infeasibility of synthetic controls with many donors. https://erikjanvankesteren.nl/blog/synth_optimization

Van Kesteren, E.-J. (2024) On the sparsity of the SCM convex hull constraint. https://erikjanvankesteren.nl/blog/synth_sparsity

https://odissei-soda.nl