Leveraging Sparsity in Theoretical and Applied Machine Learning and Causal Inference

Gregory Faletto

Ph.D. Candidate, University of Southern California Marshall School of Business

Department of Data Sciences and Operations: Statistics Group

May 22nd, 2023

Zuckerberg San Francisco General Hospital and the Division of Cardiology

Outline

1 PRESTO
- A) Background and Motivation
- B) Method

2 Fused extended two-way fixed effects
- A) Causal inference background and difference-in-differences
- B) Extended two-way fixed effects
- C) FETWFE

3 Cluster stability selection

Estimating Probabilities of Rare Events is Important and Hard

Classifiers struggle in the *class imbalance* setting where one class is very rare.

Unfortunately, accurate estimates of the probabilities of rare events are often very important.

> **Health and medicine:** rare diseases can be expensive to treat, the resources to deliver effective treatment are scarce, and it can be very stressful to believe you may have a rare disease. Accurately estimating the probability of having a rare disease is crucial to treat patients effectively.

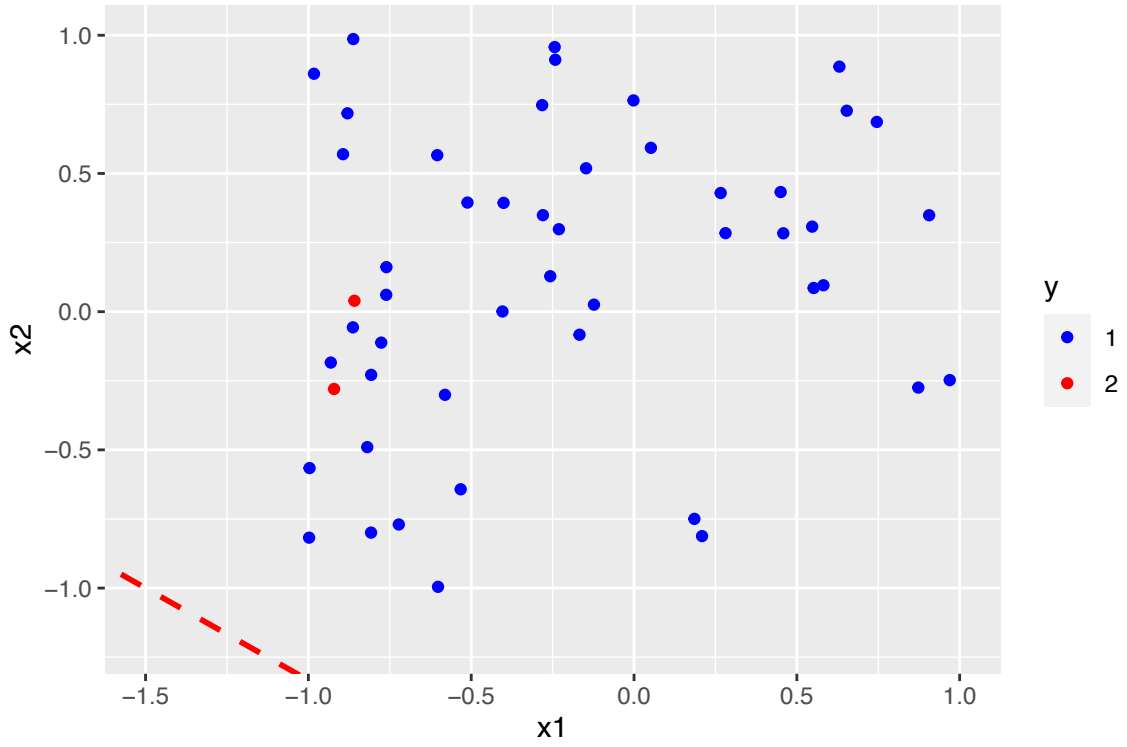
Conceptualizing the Problem

We draw 50 random points.

Binary label y generated from a logistic regression model.

Dashed line: true decision boundary (the closer to the bottom left corner you are, the more likely to be in the red class).

Because class 2 is rare, we don't observe many points near the decision boundary. Estimating the decision boundary, and $\mathbb{P}(y = 2 \mid \boldsymbol{x})$, is very difficult.



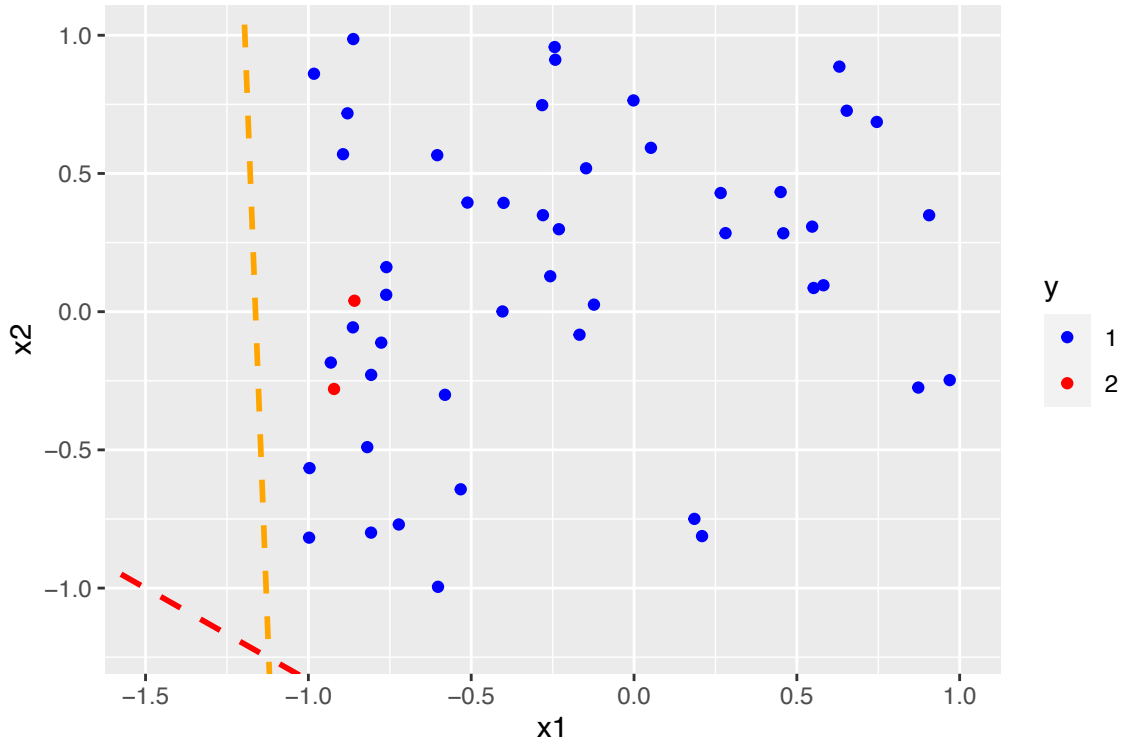
Conceptualizing the Problem

We draw 50 random points.

Binary label y generated from a logistic regression model.

Dashed line: true decision boundary (the closer to the bottom left corner you are, the more likely to be in the red class).

Because class 2 is rare, we don't observe many points near the decision boundary. Estimating the decision boundary, and $\mathbb{P}(y = 2 \mid \boldsymbol{x})$, is very difficult.



The estimated decision boundary (orange) is very bad $\Longrightarrow$ very bad rare probability estimates. (Even asymptotically!)

A Possible Solution: Leveraging Data From More Common Outcomes

Sometimes there are more common intermediate (**ordered**) outcomes.

> **Health and medicine:** sometimes there are intermediate outcomes between having a disease and being completely healthy (normal blood glucose > prediabetes > diabetes, normal blood pressure > prehypertension > hypertension, etc.)

Estimating the decision boundaries between these more common outcomes is easier (less class imbalance).

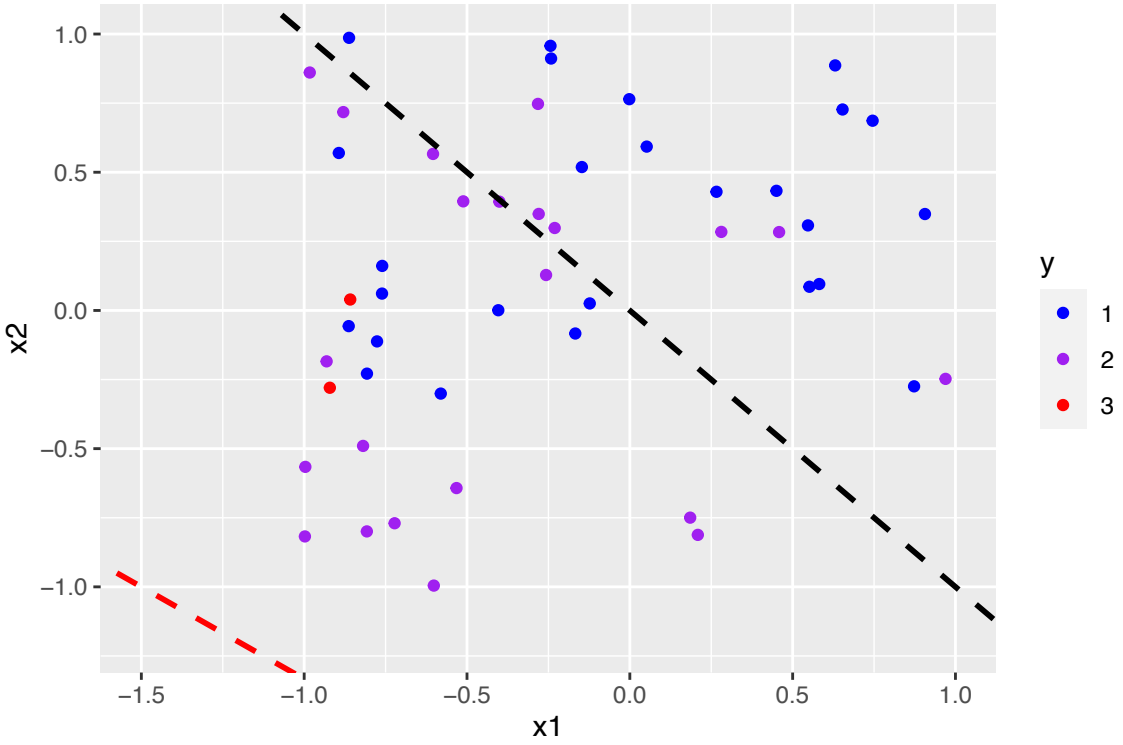
Idea: maybe the factors that are predictive of common outcomes can be used to improve predictions of rare outcomes.

A Possible Solution: Leveraging Data From More Common Outcomes

Same data as before, same red decision boundary as before. Class 2 is an intermediate outcome between class 1 and class 3. (In previous slide, classes 1 and 2 were combined.)

Black dashed line: decision boundary between classes 1 and 2.

Abundant data near decision boundary between classes 1 and 2 $\Longrightarrow$ easier to learn.



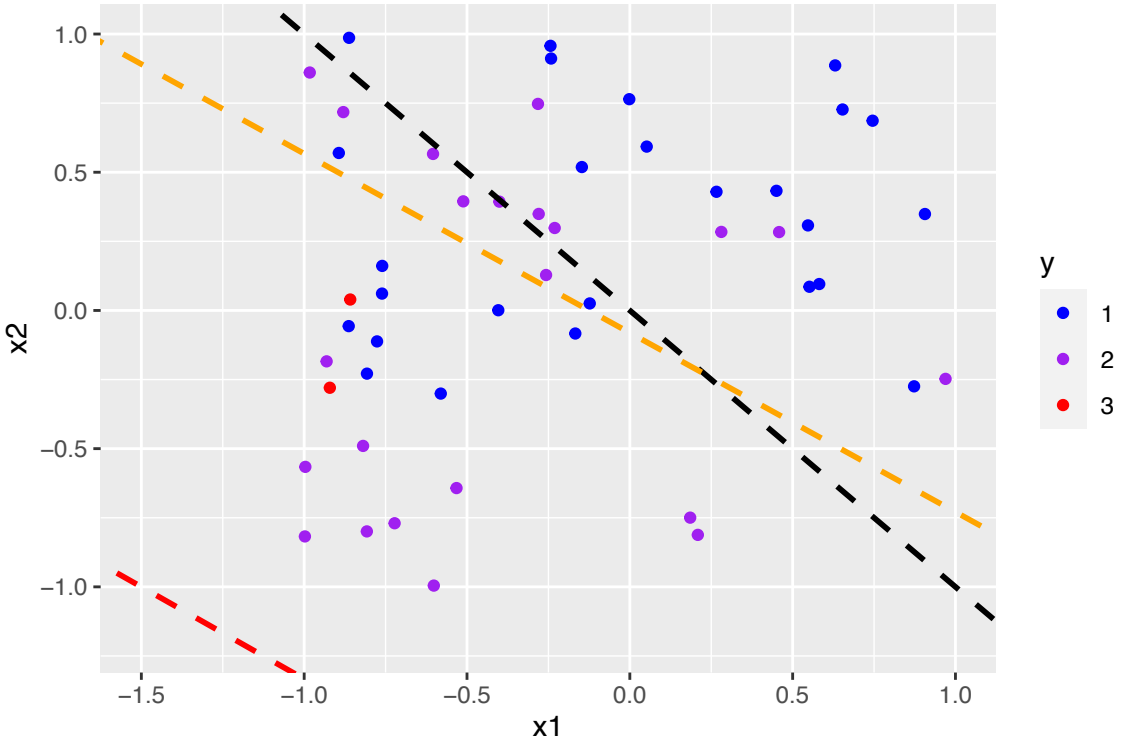
A Possible Solution: Leveraging Data From More Common Outcomes

Same data as before, same red decision boundary as before. Class 2 is an intermediate outcome between class 1 and class 3. (In previous slide, classes 1 and 2 were combined.)

Black dashed line: decision boundary between classes 1 and 2.

Abundant data near decision boundary between classes 1 and 2 $\Longrightarrow$ easier to learn.

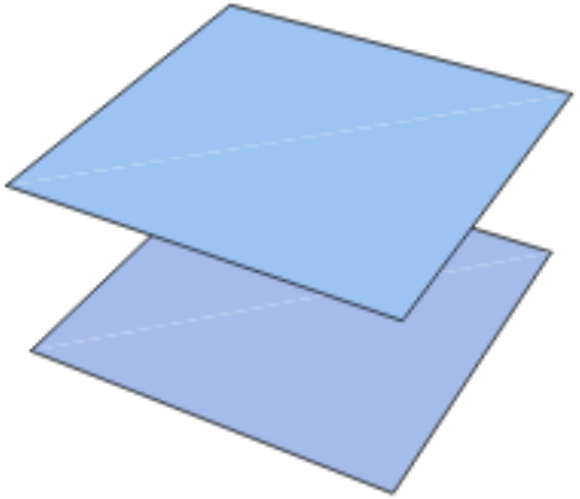
The estimated decision boundary between classes 1 and 2 (orange) does seem to fit much better.



Maybe we can leverage our more precise estimate of the decision boundary between classes 1 and 2 to better estimate the decision boundary between classes 2 and 3 $\Longrightarrow$ better estimated probabilities of lying in class 3.

A Classical Model Can Do This: The Proportional Odds Model

P*roportional odds model* or *ordered logit model* (McCullagh, 1980) for ordinal outcomes $k \in \{1,\ldots,K\}$ assumes that the decision boundaries between successive categories $y \in \{1,\ldots,K\}$ all have the same slope separated by different intercepts. (The decision boundaries are parallel.)



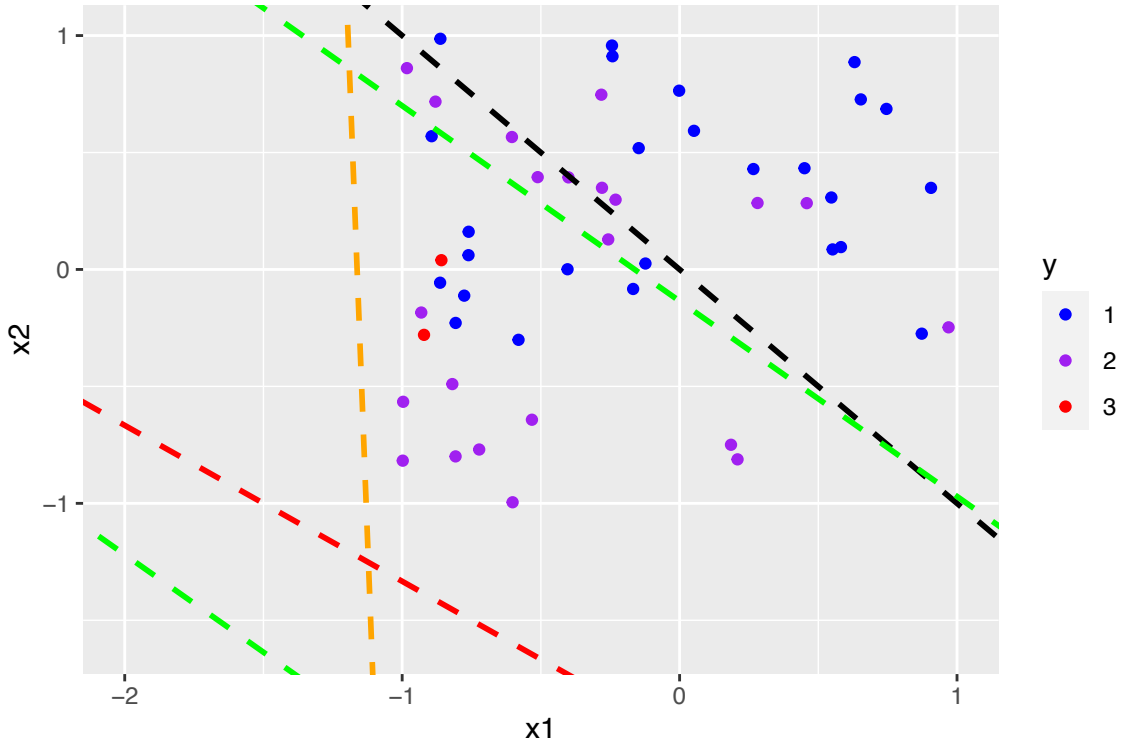
Only have to estimate one unique parameter for each of the $K-1$ decision boundaries: the intercept term. (Assumption of equal $\boldsymbol{\beta}$ means we don't have to learn much from scarce data in rare categories.)

The Proportional Odds Model Leverages Data From More Common Outcomes

Green dashed lines: estimated decision boundaries from proportional odds model. (Orange decision boundary is from logistic regression on rare class, same as before.)

Proportional odds model does seem to yield a much better estimate of the rare decision boundary!

Also yields better estimates of rare class probabilities.



Problem solved? Not quite…

So Far

So far, we've seen:

Estimating the probabilities of rare events precisely can be important.

Binary classifiers struggle in this setting (class imbalance).

Sometimes more common outcomes exist that are related to the rare outcome (on an ordinal scale.) Estimating the decision boundary associated with more common outcomes is easier.

If the common decision boundary gives us information about the rare decision boundary, then leveraging more precise estimate of the common decision boundary $\Longrightarrow$ better estimate the rare decision boundary (and, ultimately, the probabilities of rare events).

So what's wrong with just using the proportional odds model?

What's Wrong With Just Using The Proportional Odds Model? It's Unrealistically Rigid

Proportional odds model assumes that the coefficient vector associated with each decision boundary is identical. Earlier slides: when an outcome is rare, this is an improvement over separate logistic regressions to estimate each decision boundary.

But it's still unrealistically inflexible. **What if individual features have different influences on the decision boundaries at different levels?**

> Maybe a common set of characteristics predicts whether patients will suffer from a condition, but a different set of variables predicts how well their recovery will go.

Ideally, we'd like the more common decision boundaries to *inform* the estimation of the rare decision boundaries without imposing exact equality.

Outline

1 PRESTO
- A) Background and Motivation
- B) Method

2 Fused extended two-way fixed effects
- A) Causal inference background and difference-in-differences
- B) Extended two-way fixed effects
- C) FETWFE

3 Cluster stability selection

PRESTO

In brief, PRESTO estimates something like the proportional odds model, but allows the coefficient vectors for each decision boundary to be different.

However, it regularizes the differences.

> The differences in the coefficient corresponding to the same feature at adjacent decision boundaries are penalized: $|\beta_{j,k} - \beta_{j,k-1}|$
>
> If all of these differences equal 0, we have the proportional odds model. If none of them equal 0, we have completely free decision boundaries.
>
> Penalization encourages these differences to be 0 unless the data "overrules" this because allowing the coefficients to differ is "worth it" for predictive performance.

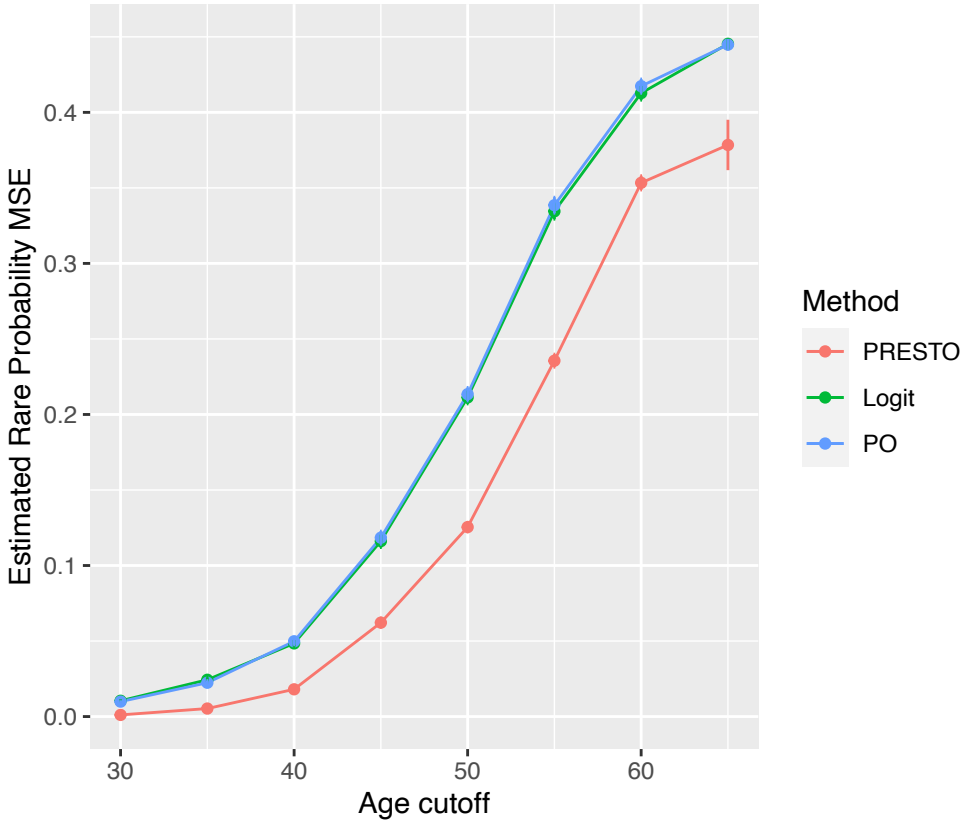
PRESTO learns from the data which parts of the common decision boundaries can be used to inform estimation of the rare decision boundaries, without the rigid proportional odds assumption.

Data Application

Data set: 3059 patients who were eventually diagnosed with diabetes. For each patient: age they were diagnosed with prediabetes, age diagnosed with diabetes, covariates.

For each age cutoff, create an ordered response variable: by this age, did the patient have prediabetes, diabetes, or neither?

Use 90% of the data to fit a model, hold out 10% for testing. Use each model to predict whether or not patient had diabetes at this age on test set. Do this 35 times for each age.



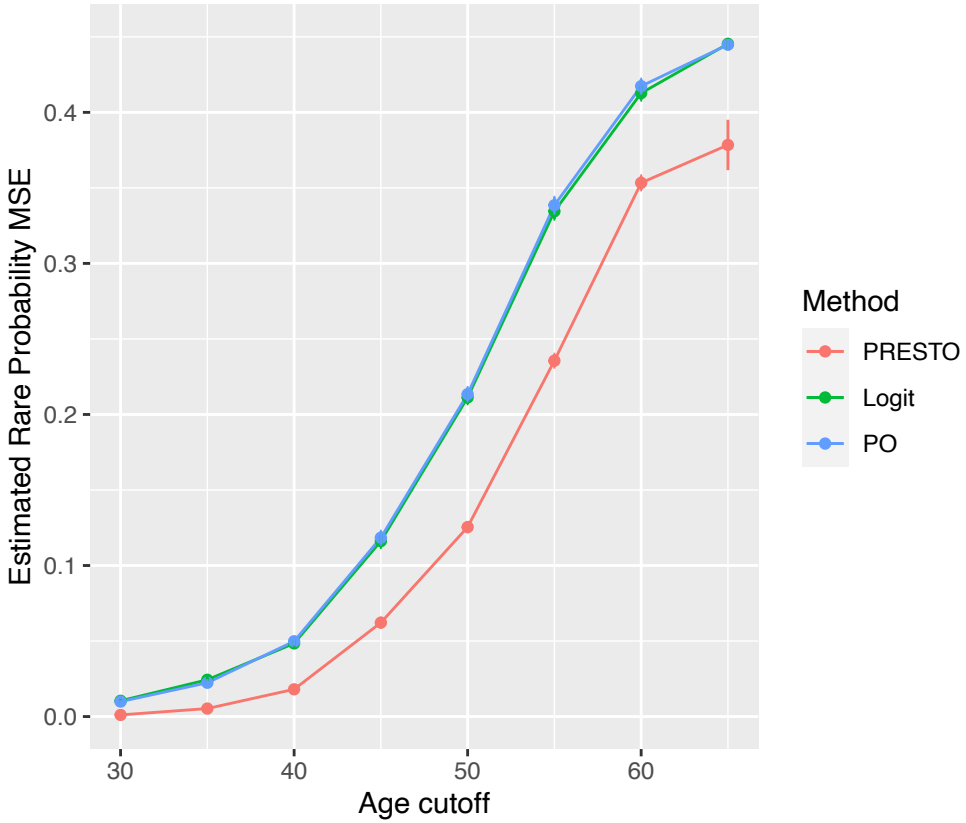
Data Application

Methods used: PRESTO, logistic regression (solely predicting whether the patient has diabetes), proportional odds model

PRESTO outperforms both other methods.

Logistic regression fails to learn from the more common cases of patients with prediabetes.

Proportional odds is too rigid, can't adapt to a different decision boundary for patients with diabetes.



Outline

1 PRESTO
- A) Background and Motivation
- B) Method

2 Fused extended two-way fixed effects
- **A) Causal inference background and difference-in-differences**
- B) Extended two-way fixed effects
- C) FETWFE

3 Cluster stability selection

Causal Inference

We observe some data. Some units receive a treatment, some don't.

For each unit, we observe some kind of numeric outcome (resting heart rate, blood pressure, etc.)

Causal effect: the difference between the unit's outcome if they do receive the treatment vs. their outcome if they don't.

Can we estimate the causal effect of the treatment?

Unit	Potential Outcomes				
	Cough Frequency (`cfa`)		Observed Variables		
	$Y_i(0)$	$Y_i(1)$	W_i	X_i (`cfp`)	Y_i^{obs} (`cfa`)
1	?	3	1	4	3
2	?	5	1	6	5
3	?	0	1	4	0
4	4	?	0	4	4
5	0	?	0	1	0
6	1	?	0	5	1

Causal Inference

Notation: W is a treatment variable (1 for treated units, 0 for untreated). $y(1)$ is the response we observe if a unit is treated, $y(0)$ is the response we observe if unit is not treated. (*Potential outcomes*)

Causal estimand (effect of treatment on the treated units):

$$\tau = \mathbb{E}[y(1) - y(0) \mid W = 1]$$

Unit	Potential Outcomes				
	Cough Frequency (`cfa`)		Observed Variables		
	$Y_i(0)$	$Y_i(1)$	W_i	X_i (`cfp`)	Y_i^{obs} (`cfa`)
1	?	3	1	4	3
2	?	5	1	6	5
3	?	0	1	4	0
4	4	?	0	4	4
5	0	?	0	1	0
6	1	?	0	5	1

Causal Inference

Fundamental problem of causal inference: we only observe one outcome for each unit.

Under randomized, double-blind assignment, the difference in average outcomes for treated and untreated units is an unbiased estimate of the treatment effect: $\hat{\tau} = \bar{y}^{(1)} - \bar{y}^{(0)}$.

$\bar{y}^{(1)}$: sample mean for treated units; $\bar{y}^{(0)}$: sample mean for control units.

Unit	Potential Outcomes				
	Cough Frequency (`cfa`)		Observed Variables		
	$Y_i(0)$	$Y_i(1)$	W_i	X_i (`cfp`)	Y_i^{obs} (`cfa`)
1	?	3	1	4	3
2	?	5	1	6	5
3	?	0	1	4	0
4	4	?	0	4	4
5	0	?	0	1	0
6	1	?	0	5	1

Causal Inference

In an observational study, things are much harder.

> Maybe patients who sought treatment are more conscious of their health, and would have seen some improvement in outcomes regardless of the treatment.
>
> Maybe patients sought treatment because of a fluky measurement, and their outcome would have reverted to the mean regardless of treatment.

In these cases and many others, the untreated units are not a reasonable baseline for comparison.

We can’t credibly estimate treatment effects unless we find a way to rule these kinds of things out.

Difference-in-Differences

Suppose we observe units at two times.

First time period: no units receive treatment.

Second time period: some do (not randomly assigned).

Let $y_t(0)$ be the untreated potential outcome for a unit at time t, and $y_t(1)$ be the treated outcome.

Goal: estimate

$$\tau = \mathbb{E}[y_2(1) - y_2(0) \mid W = 1],$$

where W denotes that the unit was treated at time 2.

Difference-in-Differences

$$\tau = \mathbb{E}[y_2(1) - y_2(0) \mid W = 1]$$

An obvious estimator: the **"before-and-after"** estimator:

> Takes the mean outcome for treated units at time 2 minus their mean outcome at time 1, when they were untreated:

$$\hat{\tau}_{BA} = \bar{y}_2^{(1)} - \bar{y}_1^{(1)},$$

where the superscript (1) conveys that in both means we include only units that were treated at time 2.

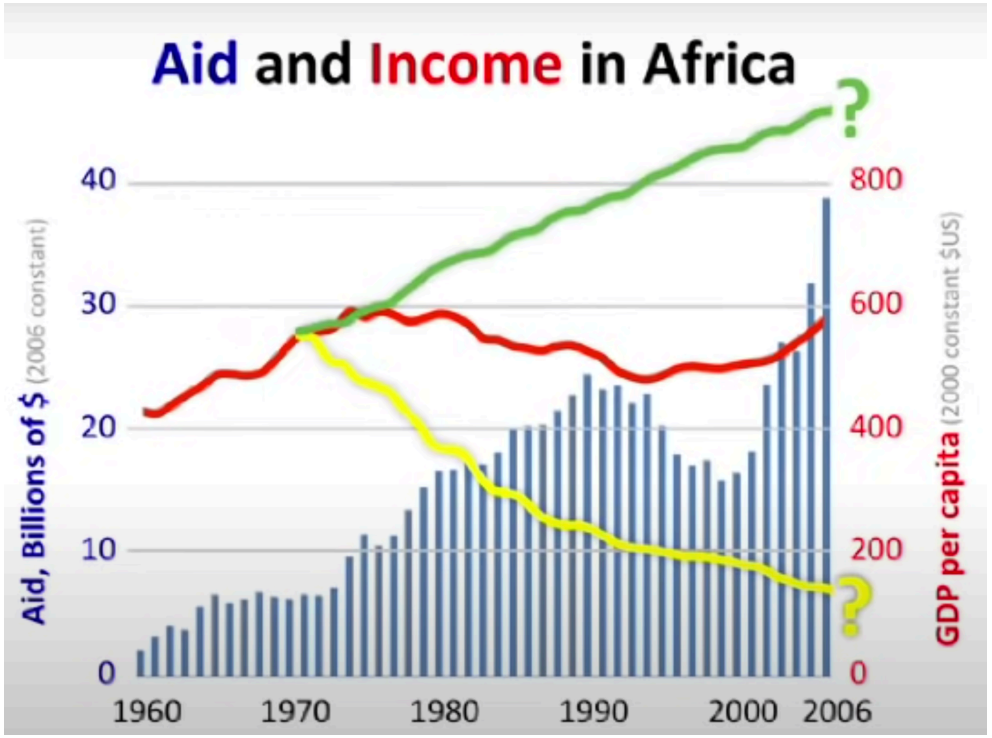
Difference-in-Differences

$$\tau = \mathbb{E}[y_{i2}(1) - y_{i2}(0) \mid W_i = 1]$$

$$\hat{\tau}_{BA} = \overline{y}_2^{(1)} - \overline{y}_1^{(1)}$$

"Before-and-after" estimator makes sense if we assume $\overline{y}_1^{(1)}$ is a good estimator for the units' (unobserved) untreated potential outcomes at time 2, $y_2^{(1)}$.

But it might not be! Can we do better?

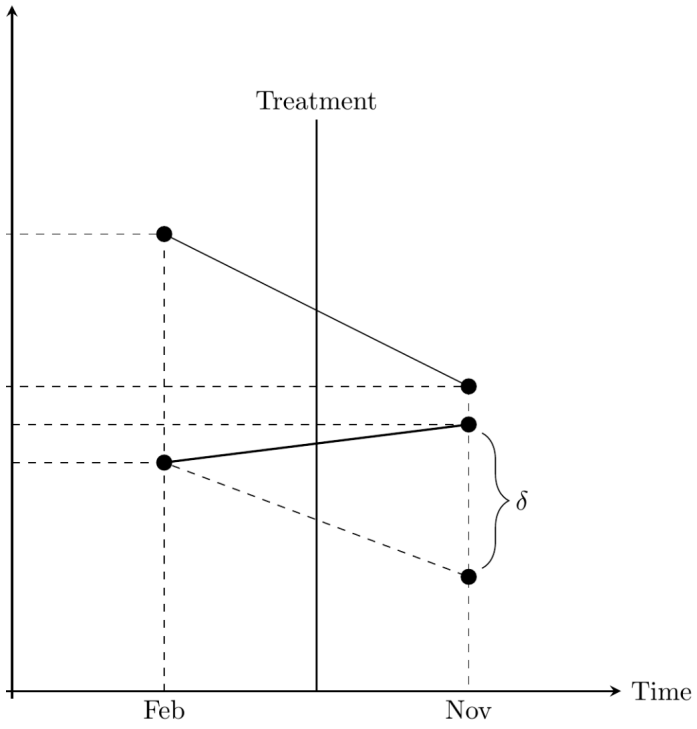


Difference-in-Differences Estimator

Fortunately, we have some information about external conditions that might have changed from time 1 to 2—**the observed outcomes from the untreated units**.

Suppose that for both units who received treatment and units who didn't, the change in their untreated potential outcome from time 1 to time 2 would have been the same (**parallel** or **common trend**):

$$\mathbb{E}[y_2(0) - y_1(0) \mid W = 1] = \mathbb{E}[y_2(0) - y_2(0) \mid W = 0]\,.$$



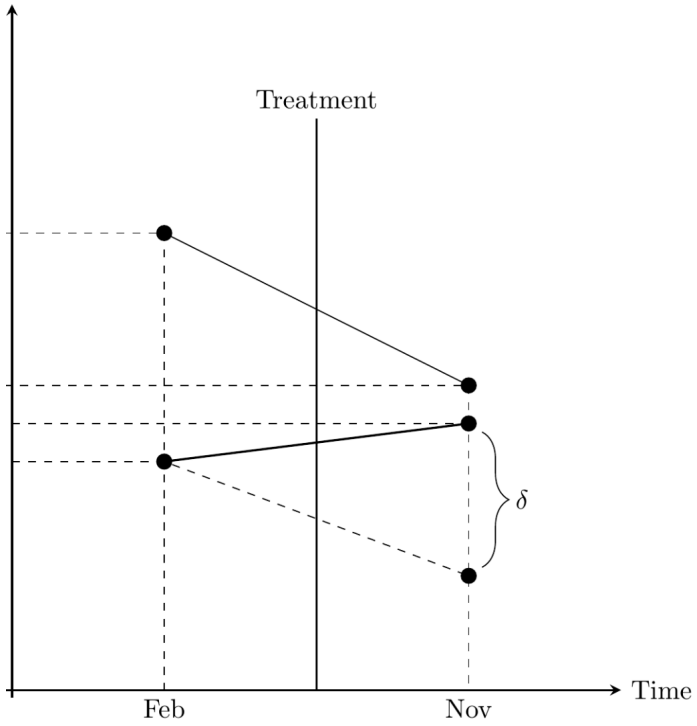
Difference-in-Differences Estimator

$$\mathbb{E}[y_2(0) - y_1(0) \mid W = 1] = \mathbb{E}[y_2(0) - y_2(0) \mid W = 0]\,.$$

Then we can improve the before-and-after estimator by adjusting for the observed change in untreated units:

$$\hat{\tau}_{\text{DID}} = \overline{y}_2^{(1)} - \overline{y}_1^{(1)} - \left[\overline{y}_2^{(0)} - \overline{y}_1^{(0)}\right],$$

This is the **difference-in-differences** estimator.



Difference-in-Differences Estimator

$$\hat{\tau}_{\text{DID}} = \bar{y}_2^{(1)} - \bar{y}_1^{(1)} - \left[\bar{y}_2^{(0)} - \bar{y}_1^{(0)}\right]$$

Difference-in-differences estimator is unbiased under :

(1) **Common trends** assumption from the last slide
(2) **"No anticipation"** assumption: the treatment to come does not affect the potential outcome of the treated units before treatment.

An equivalent way to calculate it is to estimate the coefficient $\hat{\tau}$ in the linear regression

$$y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \epsilon_{it},$$

where α_i is a separate intercept for each unit, γ_t is a separate intercept for each time, and W_{it} equals 1 if unit i was treated at time t and 0 otherwise. This is called a **two-way fixed effects** regression.

Recap

For causal inference on observational data, we need some way to make valid comparisons between the treated and control units.

The difference-in-differences method allows us to make these comparisons if we observe units at two different times, and treatment only occurs at the second time for only some units.

If the change in untreated potential outcomes for control units matches the change in untreated potential outcomes for the treated units ("common trends"), this results in unbiased treatment effect estimates.

In practice, we can calculate these estimates using a two-way fixed effects linear regression.

Can we extend this further, to more general settings?
If we observe covariates, can we use them to relax our assumptions?

Two-Way Fixed Effects

$$y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \epsilon_{it},$$

Now that we're calculating a linear regression…

Can we include controls (covariates) $\boldsymbol{X}$ in the regression to improve precision?

Can we allow an arbitrary number of time periods instead of just 2? (And can we allow units to start receiving treatment at arbitrary times after time 1?)

In general: **no!** This model will be biased.

How do we need to change the model?

Outline

1 PRESTO
- A) Background and Motivation
- B) Method

2 Fused extended two-way fixed effects
- A) Causal inference background and difference-in-differences
- **B) Extended two-way fixed effects**
- C) FETWFE

3 Cluster stability selection

Extended Two-Way Fixed Effects

Notation: suppose we observe T time periods. We have **cohorts** that receive treatment at times $r = \{2,3,\ldots,T\}$ (or some subset of those times).

Wooldridge (2021): linear regression with covariates is okay, but we need more parameters.

- Covariates must be time-invariant (pre-treatment)
- Estimate separate treatment effects for each cohort and time, τ_{rt}
- Estimate separate covariate coefficient vectors for each time and for each cohort
- Include interaction terms between covariates and treatment effects

Wooldridge calls this the **extended two-way fixed effects** model.

Extended Two-Way Fixed Effects

Intuitively, we expect extended two-way fixed effects to give use more precise estimates. **What exactly have we gained?**

Unbiased estimation of treatment effects in a much more general setting

No-anticipation and common trends assumptions can be relaxed—we can allow for anticipatory effects, or differing from common trends, as long as these can be explained by covariates.

What price do we pay?

We have a lot of parameters to estimate! If there are lots of time periods or covariates, our parameter estimates may be too noisy to be useful.

In other words, extended two-way fixed effects may be too flexible.

Intuition: we can improve this with regularization!

Extended Two-Way Fixed Effects

Wooldridge proposes an *ad hoc* remedy for the problem of too many parameters: assume some of the parameters are equal

> For example: maybe the treatment effects don't actual differ in time since treatment. So assume $\tau_{rt} = \tau_r$ for all cohorts at all post-treatment times in order to reduce the number of parameters to estimate.
>
> Or, weaken this a little: assume there is an "early treatment" and "late treatment" effect, so that we only have two treatment effects to estimate for each cohort instead of $T - r + 1$.

Problem: **is this wishful thinking?**

> Unless these assumptions are well-justified, we risk re-introducing the bias we removed by adding these parameters in the first place.

Can we use the fact that some of these restrictions probably exist in the data without putting our "thumb on the scale" by selecting the restrictions by hand?

Outline

1 PRESTO
- A) Background and Motivation
- B) Method

2 Fused extended two-way fixed effects
- A) Causal inference background and difference-in-differences
- B) Extended two-way fixed effects
- **C) FETWFE**

3 Cluster stability selection

Fused Extended Two-Way Fixed Effects

Idea: let's reduce the number of parameters to estimate using regularization.

Encode our expectations about which parameters might be equal in the way that we estimate the model, then allow the model to learn the restrictions from the data.

Estimate the extended two-way fixed effects regression, but add penalty terms that align with our belief that some of the parameters are actually equal

> Example: for each cohort, penalize the differences $|\tau_{r,t+1} - \tau_{rt}|$.
> This encodes our belief that the treatment effects in adjacent times are probably close together, so we can set them equal **unless the data gives us a good reason to estimate separate values for these parameters.**

We outsource the choice of which of these restrictions to impose to the data.

Theoretical Guarantees (High-Level Summary)

Assumptions

No anticipation of treatment of units that cannot be explained by the time-invariant covariates X.

Common trends between control units and each cohort's untreated potential outcomes, with differences that can be explained by X allowed.

Some of the restrictions we discussed exist (but we don't have to worry about knowing which ones to pick).

Treatment status is random (in the sense that we couldn't have predicted perfectly in advance which units would be treated), but allowed to be biased as long as the common trends assumption holds. Each unit must have positive probability of being assigned to each cohort, or the never-treated group.

Theoretical Guarantees (High-Level Summary)

Theorem 5.1: Fused extended two-way fixed effects (FETWFE) consistently estimates the treatment effects (as the number of units goes to infinity, the estimated treatment effects become arbitrarily precise).

Theorem 5.2: FETWFE identifies the correct restrictions with probability tending to 1 as the number of units goes to infinity.

Theorem 5.3: FETWFE is an “oracle” procedure:

> Even if the number of covariates grows asymptotically, FETWFE converges at the same rate as an ordinary least squares model estimated using all of the correct restrictions.
>
> The asymptotic variance of FETWFE is not affected by the parameters we didn’t need to separately estimate.
>
> There is a finite-sample variance estimator that allows us to construct confidence intervals for a variety of treatment effect estimators.

Simulation Studies: Setup

We generate synthetic data with 120 units, $T = 30$ time periods, 5 cohorts, and 12 covariates. This results in a total of 3600 observations and 2209 parameters to estimate.

We generate a coefficient vector where 90% of the restrictions hold (so only about 221 parameters actually need to be estimated).

We repeat the following for 350 simulations:

- Generate a random (time-invariant) covariate vector for each unit
- Generate random treatment assignments: each unit is either untreated or in one of the cohorts with equal probability
- Generate a random response using the coefficient vector and added noise
- Estimate the treatment effects using four different methods

Simulation Studies: Methods

We estimate the treatment effects using four different methods:

FETWFE

ETWFE

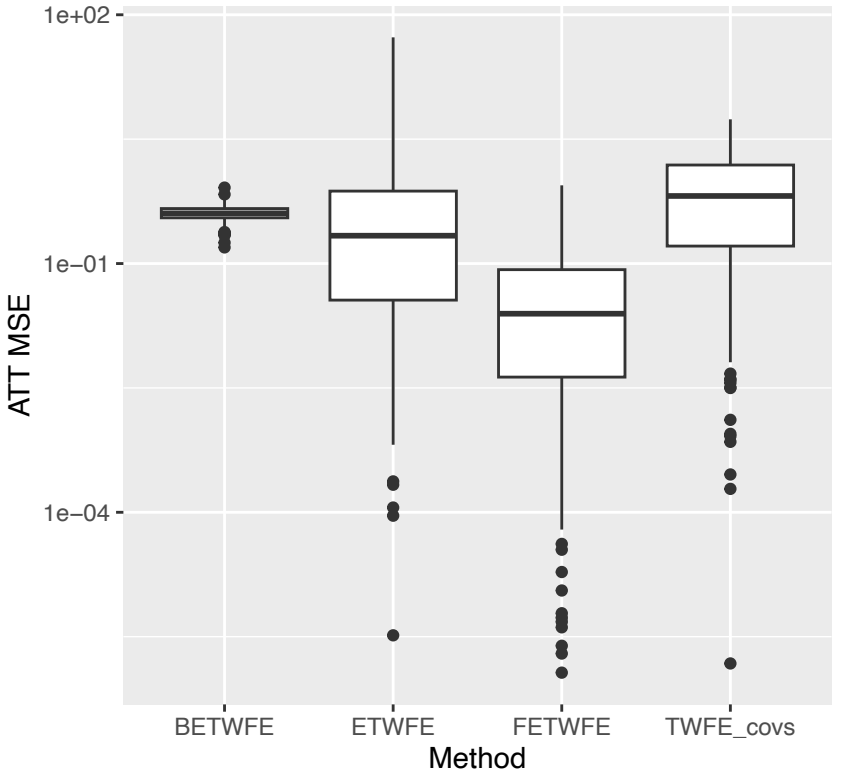
Penalized ETWFE (so adding regularization to ETWFE, but penalizing each coefficient directly rather than using fusion penalties that penalize the coefficients towards each other)

Ordinary least squares on the oversimplified model $y_{it} = \nu_r + \gamma_t + \boldsymbol{X}_i\boldsymbol{\kappa} + \tau_r W_{it,r} + \epsilon_{it}$ (separate intercept for each cohort, separate treatment effect for each cohort, added covariates)

Simulation Studies: Estimation error

We evaluate the squared error of each method in estimating the average treatment effect.

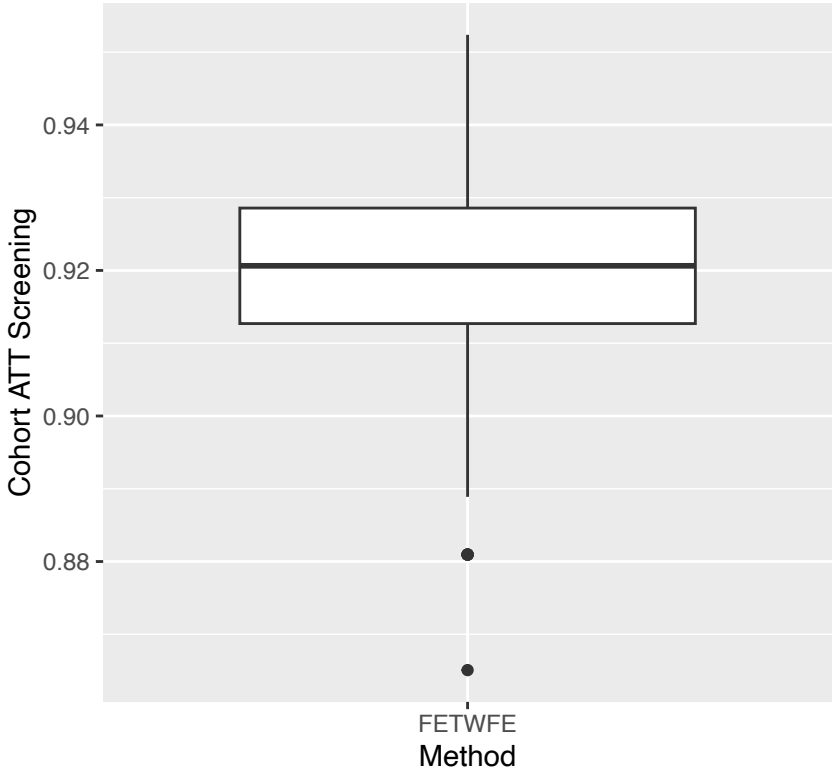
FETWFE outperforms all other methods at estimating the average treatment effect. The results are statistically significant.



Simulation Studies: Restriction Selection Consistency

On each simulation, we calculate the percentage of treatment effect restrictions that FETWFE successfully identifies.

On average, FETWFE identifies 92.1% of the restrictions in each simulation (standard error: 0.084%).

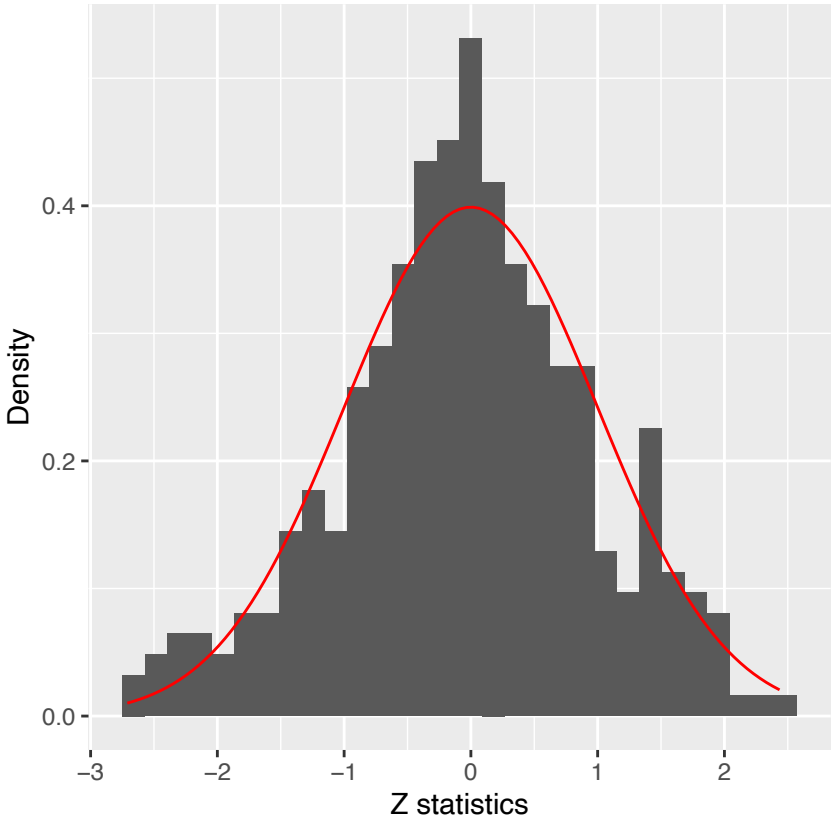


Simulation Studies: Asymptotic Distribution

We find that the distribution of Z statistics calculated from FETWFE in the previous simulation study is not approximately normal, so we conduct another simulation study to see if the asymptotics “hold up” with large enough samples.

We generate data in the same way as the previous simulation study, but with 1500 units, $T = 15$ time periods, 5 cohorts, and 2 covariates.

We construct nominal 95% confidence intervals for the treatment effect in each simulation. 94.6% of the confidence intervals contain the true treatment effect.



Outline

1 PRESTO
- A) Background and Motivation
- B) Method

2 Fused extended two-way fixed effects
- A) Causal inference background and difference-in-differences
- B) Extended two-way fixed effects
- C) FETWFE

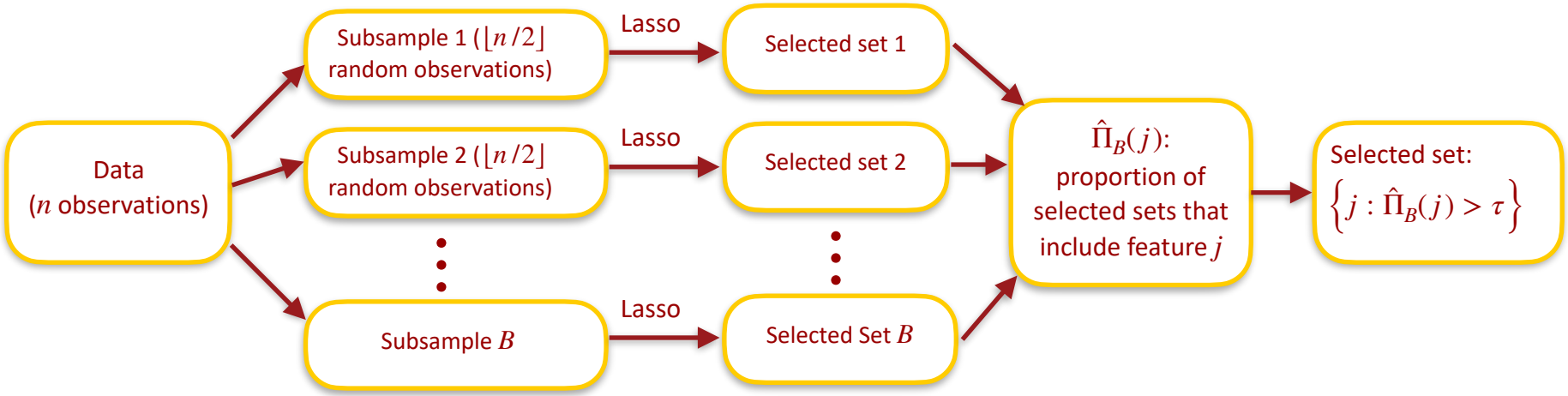
3 Cluster stability selection

Outline

Statistical Stability

"…statistical conclusions are robust or stable to appropriate perturbations to data." (Yu 2013)

Stability Selection
(Meinshausen and Buhlmann 2010)



Take repeated subsamples of size $\lfloor n/2 \rfloor$.

Run lasso on each subsample with pre-selected λ.

Return features that are chosen by lasso in a proportion of subsamples greater than pre-selected threshold τ. (Or, return top s features.)

Desirable Properties of Stability Selection

Adds stability to any feature selection method (e.g. lasso).

Guaranteed control of false discoveries under very mild assumptions.

Outline

An Observation

"Highly correlated variables… split the vote" (Shah and Samworth 2013).

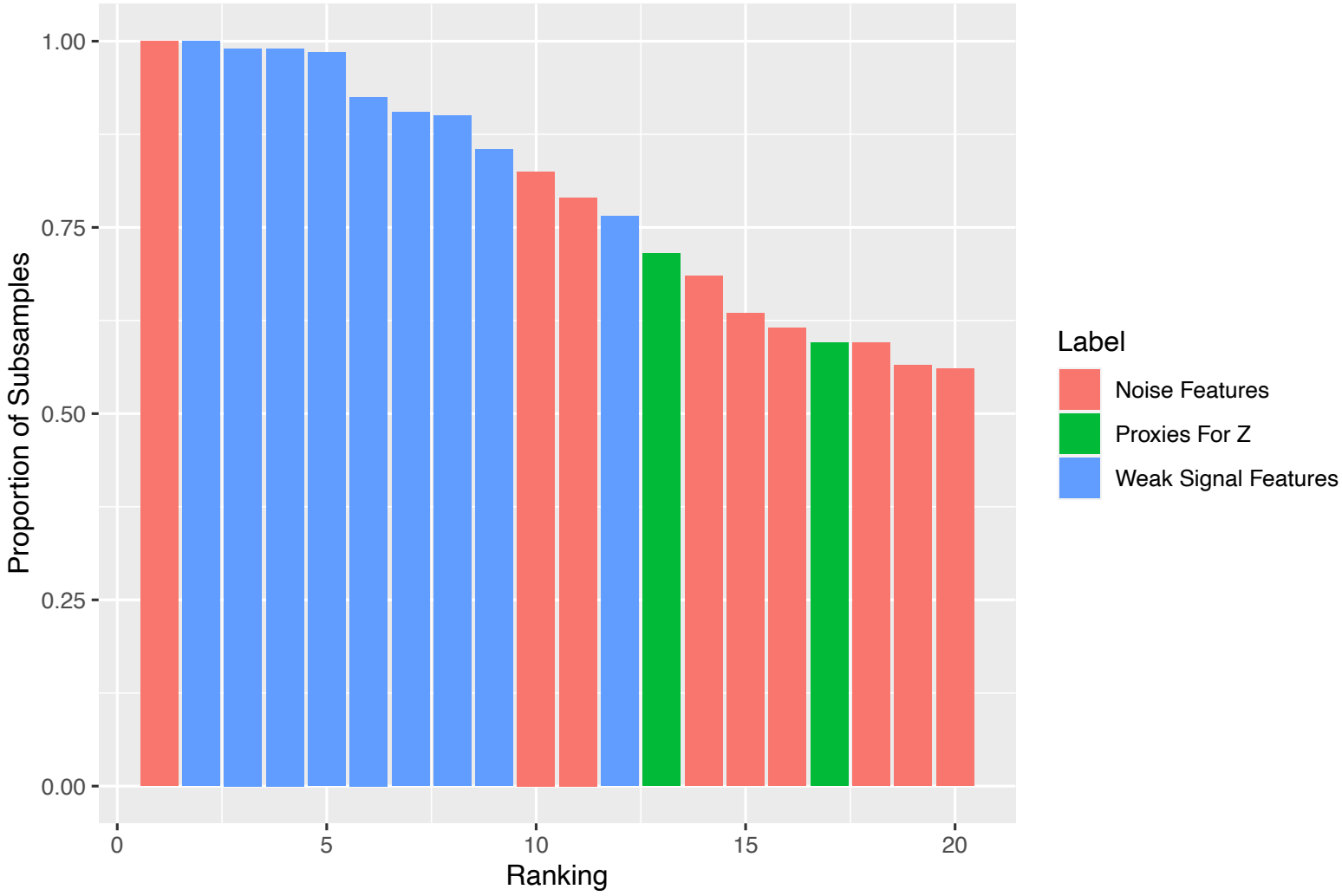
Assume we are interested in a model to make good out-of-sample predictions. We want stable feature selection, so we try stability selection with the lasso as the base procedure.

Suppose Z is in the true model. We don't observe Z, but we observe q features $X^{(\text{proxies})}$ that are highly correlated with Z (*proxies*).

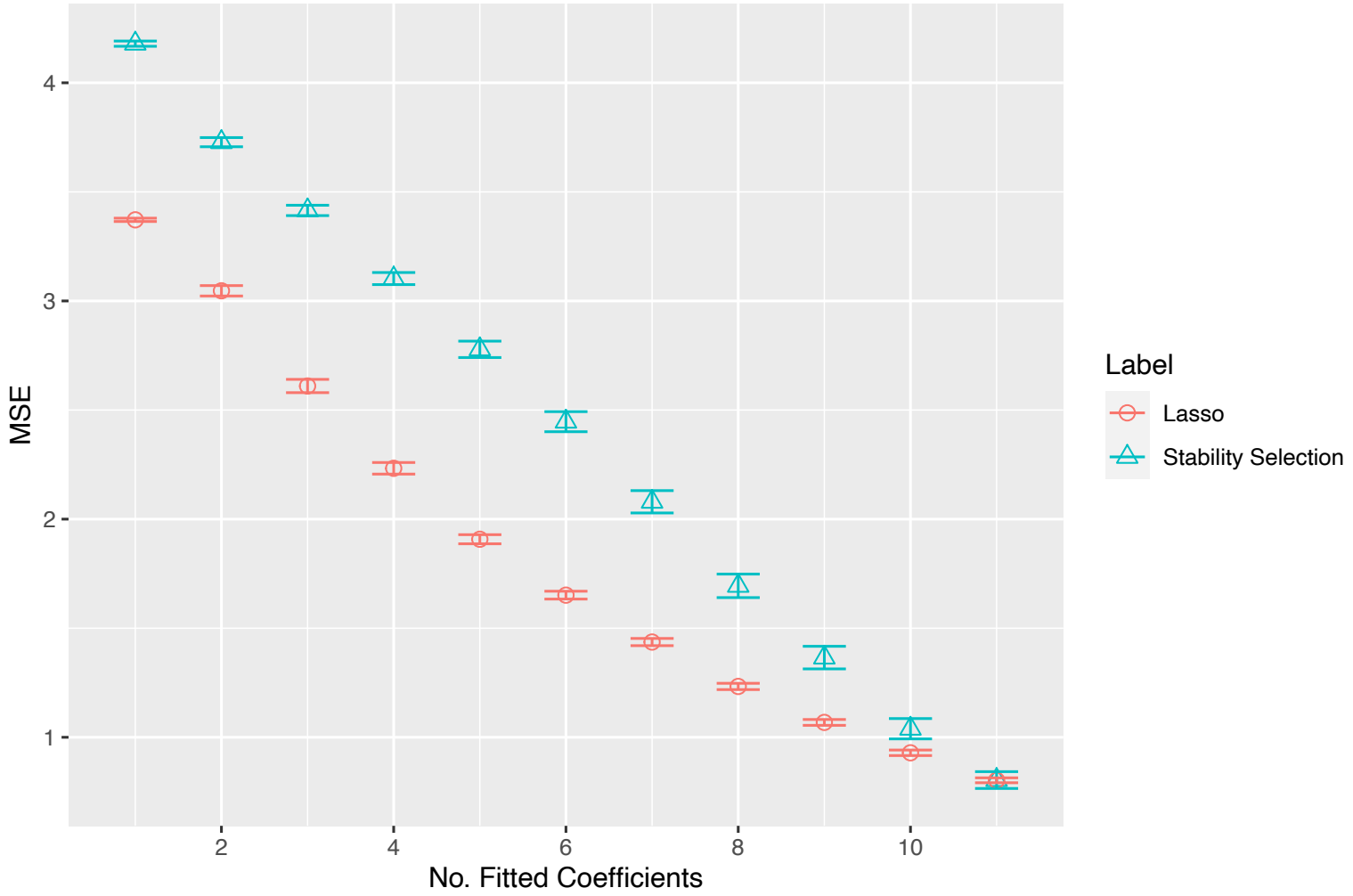
Lasso tends to choose one proxy on each fit, nearly at random.

Each proxy's selection proportion is deflated by a factor of $1/q$ (relative to what the selection proportion of Z would have been). Sorting features in order of their selection proportion results in a poor ranking of features.

Motivating Simulation



Motivating Simulation



Outline

Recall Stability Selection
(Meinshausen and Buhlmann 2010)

For each feature $j \in [p]$,

Number of subsamples taken

b^{th} subsample

Proportion of time feature j is selected

$$\hat{\Pi}_B(j) := \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left\{ j \in \hat{S}^{\lambda}\left(A_b\right) \right\}.$$

indices of the features selected by the lasso with parameter λ on sample A_b

Return features with $\hat{\Pi}_B(j) > \tau$.

(Alternatively, return s features with highest selection proportions, or return list of features ranked in descending order of $\hat{\Pi}_B(j)$.

Generalized Stability Selection

Returns a ranked list of selected clusters of features, rather than features.

Let $\mathscr{C} := \{C_1, \ldots, C_K\}$ denote the set of $K \leq p$ unique clusters.

For each cluster C_k, compute

Number of subsamples taken

b^{th} subsample

Proportion of time a feature from cluster C_k is selected

$$\hat{\Theta}_B(C_k) := \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left\{ C_k \cap \hat{S}^{\lambda}(A_b) \neq \varnothing \right\}.$$

indices of the features selected by the lasso with parameter λ on sample A_b

Return clusters with $\hat{\Theta}_B(C_k) > \tau$. (Alternatively, return s clusters with highest selection proportions, or return list of clusters ranked in descending order of $\hat{\Theta}_B(C_k)$.

Generalized Stability Selection

Also keep track of $\hat{\Pi}_B(j)$ for each individual feature.

Then for each cluster, compute weights $\boldsymbol{w}_k \in \Delta^{|C_k|-1}$ to assign to each member of the cluster. Return weights with clusters.

For regressions, construct a synthetic feature $\boldsymbol{X}_{\cdot C_k}\boldsymbol{w}_k$ for each cluster.

Regress response against s synthetic features yielded by selected clusters.

Generalized Stability Selection

Three proposals to determine weights:

1. **Sparse generalized stability selection:** Assign weight 1 to the most frequently selected cluster member, and 0 to the others.

2. **Averaged generalized stability selection:** Assign weight $1/\left|C_k\right|$ to every cluster member.

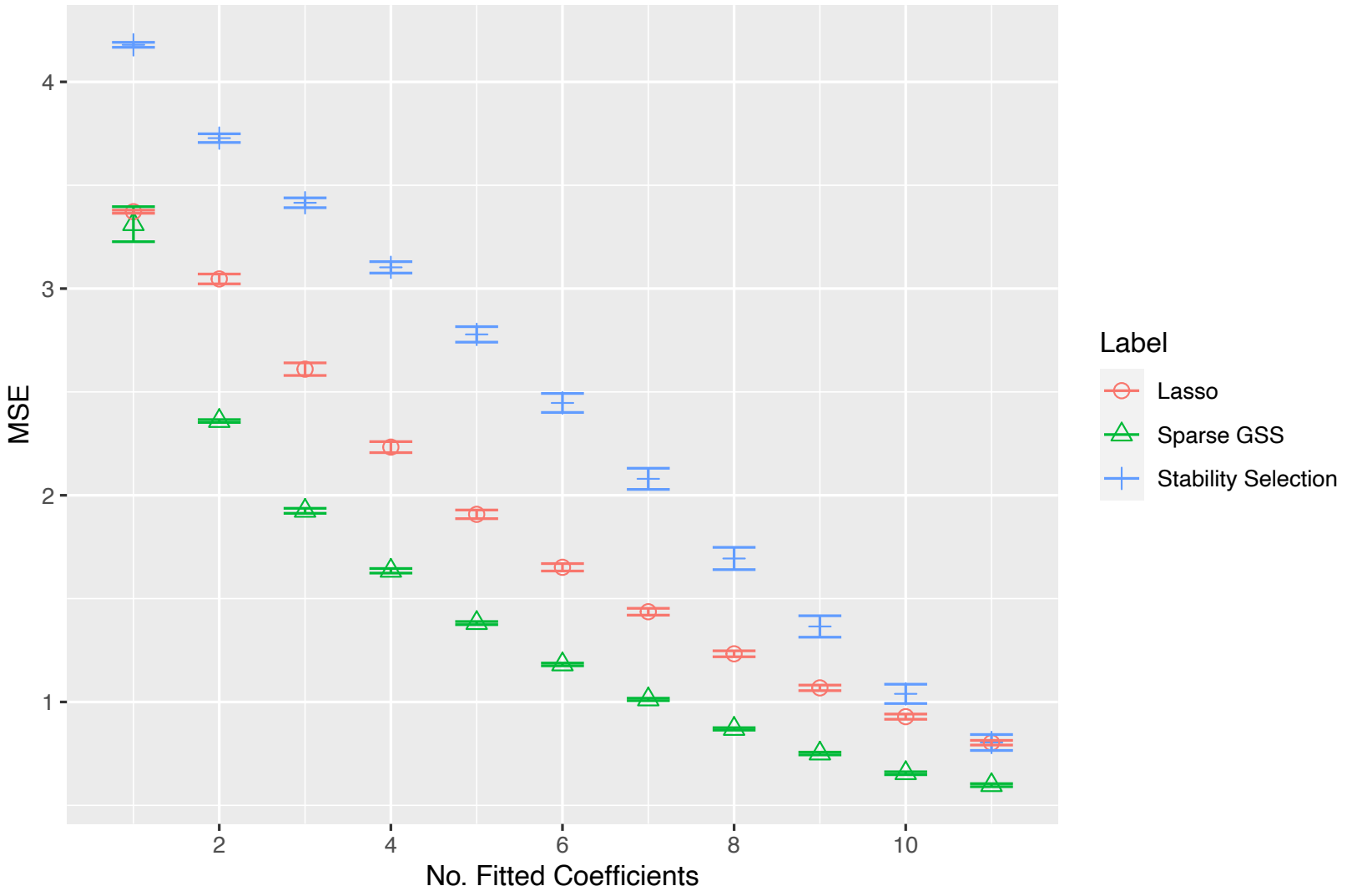
3. **Weighted averaged generalized stability selection:** Assign weight

$$\frac{\hat{\Pi}_B(j)}{\sum_{j' \in C_k} \hat{\Pi}_B(j')}$$

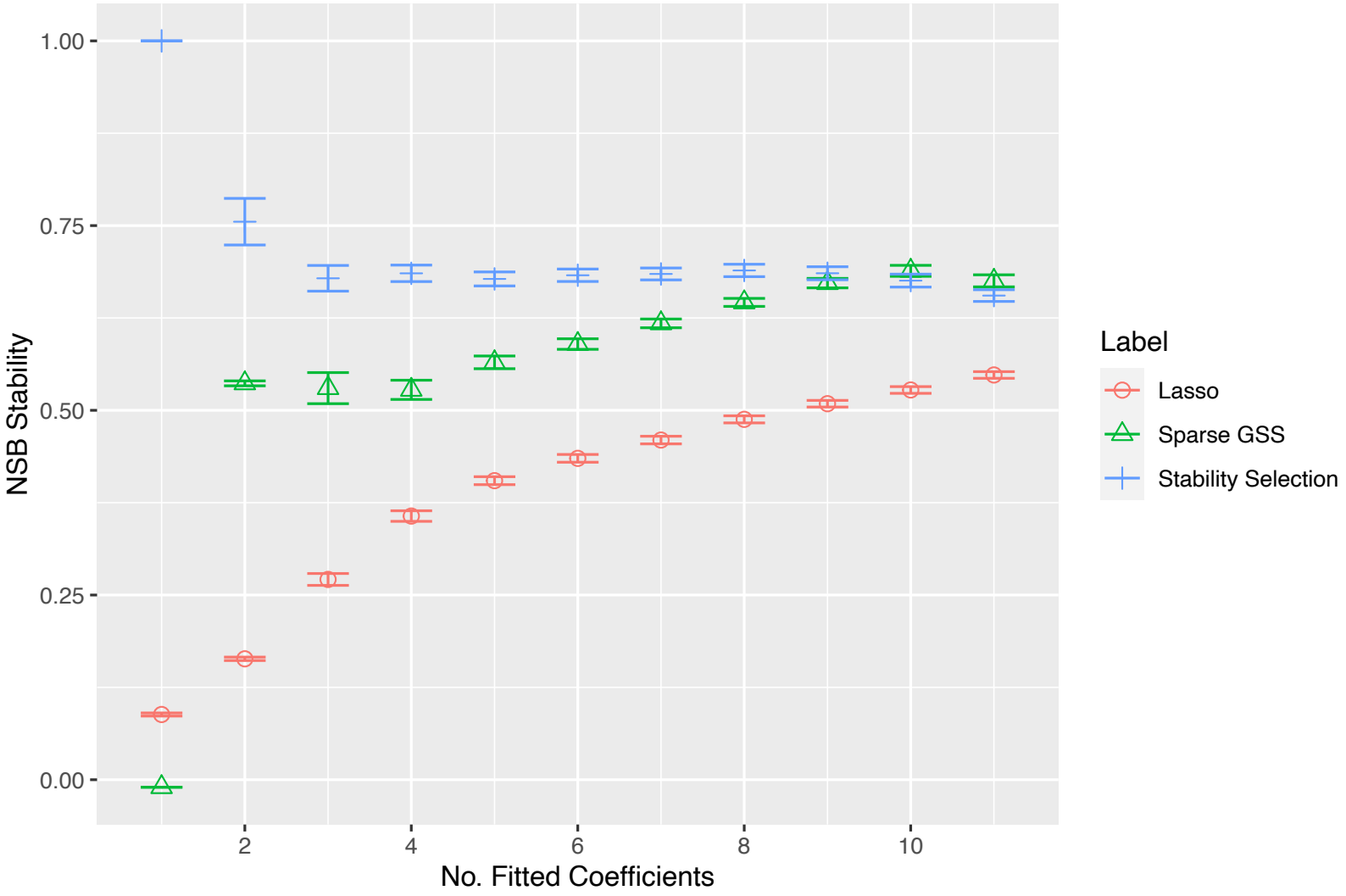
to each cluster member.

Outline

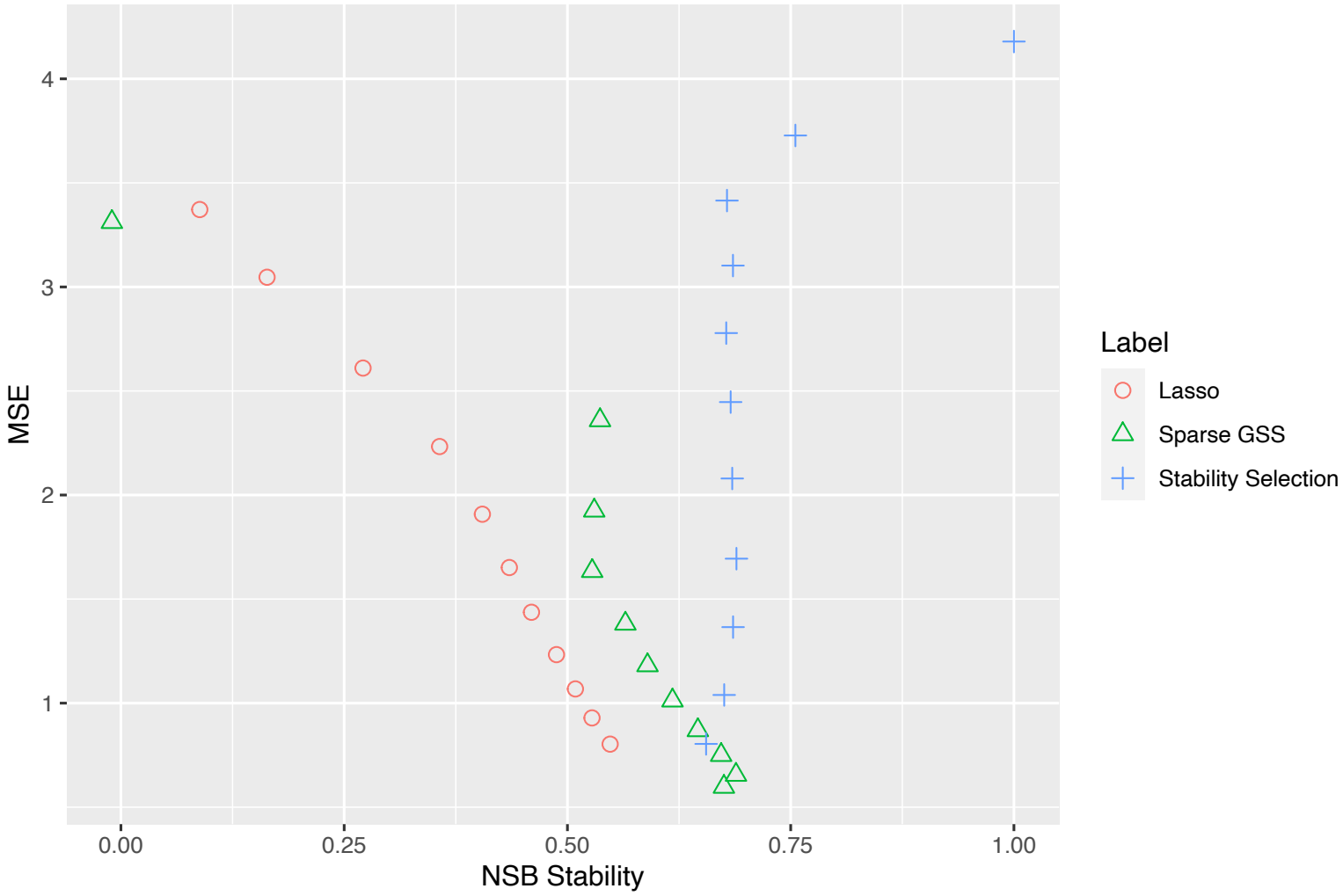
Sparse Generalized Stability Selection



Sparse Generalized Stability Selection



Sparse Generalized Stability Selection



Outline

Real Data Application—Genome-Wide Association Study

An organism's DNA typically has millions of base positions, and at each base position there is typically one more common single-nucleotide polymorphism (SNP) and one less common one.

Due to *linkage disequilibrium*, nearby SNPs tend to be highly correlated—cluster structure in data.

Goal of GWAS: use regression methods to identify SNPs associated with response (useful for e.g. diagnosing, preventing, and treating disease).

$\boldsymbol{X}$: an $(n = 1058) \times (p = 1000)$ matrix of SNPs from *Arabidopsis thaliana* plants. Each predictor takes on either the value 0 (if both SNPs at that position take on the less common value) or 1.

$\boldsymbol{y}$: logarithm of flowering time (in days) at 10° C.

Real Data Application—Genome-Wide Association Study

We repeated 100 replications of the following procedure:

Randomly divide the data into feature selection (40% of the data), training (40% of the data), and test sets.

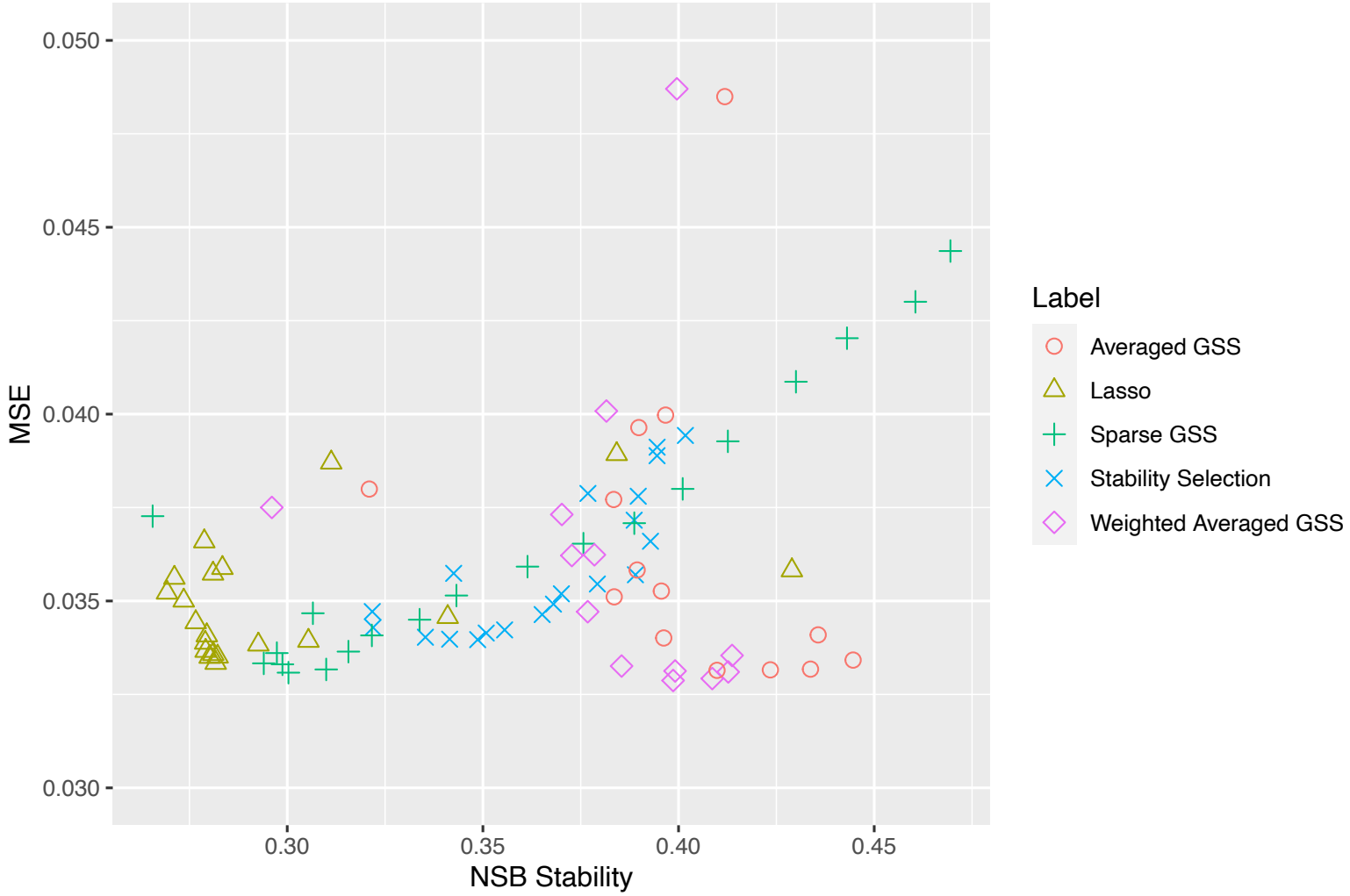
Use the feature selection and training data to estimate clusters of SNPs.

Carry out each feature selection method on the feature selection set, yielding selected sets of various sizes.

For each method and each model size, estimate a linear model via OLS on the training set using the selected features. Generate predictions from each model on the test set, and evaluate the MSE.

Also, evaluate the stability of each method at each model size across all 100 replications.

Real Data Application—Genome-Wide Association Study



Summary

Stability selection with lasso has desirable properties, but fails in case of clustered features.

Selecting one low-noise proxy (or a cluster of them) improves predictive performance when the true signal is not observed.

Generalized stability selection exploits cluster structure to correct stability selection's failure in this regime, allowing for the stable identification of important clustered features.

References

Bondell, H. D., & Reich, B. J. Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics*, 64, 115–123, 2008. https://doi.org/10.1111/j.1541-0420.2007.00843.x

D. Conn, T. Ngun, G. Li, and C. Ramirez. Fuzzy forests: Extending random forest feature selection for correlated, high-dimensional data. *Journal of Statistical Software, Articles*, 91(9):1–25, 2019. ISSN 1548-7660. doi: 10.18637/jss.v091.i09. URL https://www.jstatsoft.org/v091/i09.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2), 407–499. https://doi.org/10.1214/009053604000000067

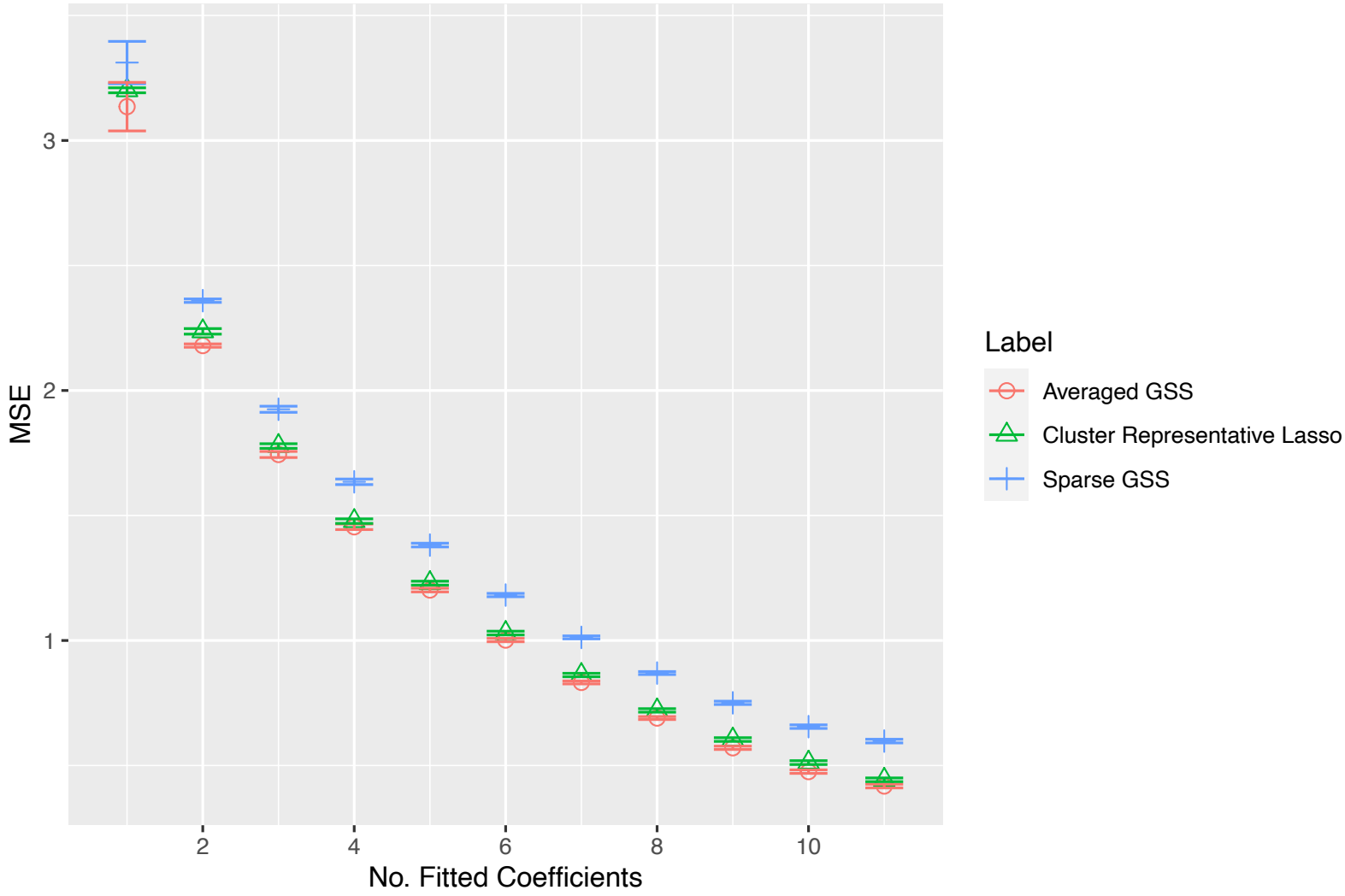
N. Meinshausen and P. Buhlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, May 2010.

R. D. Shah and R. J. Samworth. Discussion of “Correlated variables in regression: Clustering and sparse estimation” by Peter Buhlmann, Philipp Rutimann, Sara van de Geer and Cun-Hui Zhang. *Journal of Statistical Planning and Inference*, 143(11):1866–1868, 2013. doi: 10.1016/j.jspi.2013.05.022.

B. Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013. doi: 10.3150/13-BEJSP14. URL https://projecteuclid.org/download/pdfview 1/euclid.bj/1377612862.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(2):301–320, 2005. ISSN 00426989. doi: 10.1016/S0042-6989(99)00110-8.

Averaged Generalized Stability Selection



Motivating Simulation

Repeat the prior procedure 1000 times.

On each iteration:

Select feature sets of sizes $\{1,\ldots,11\}$ using:

Stability selection with the lasso as the base procedure (λ chosen in advance by cross-validation)

Lasso

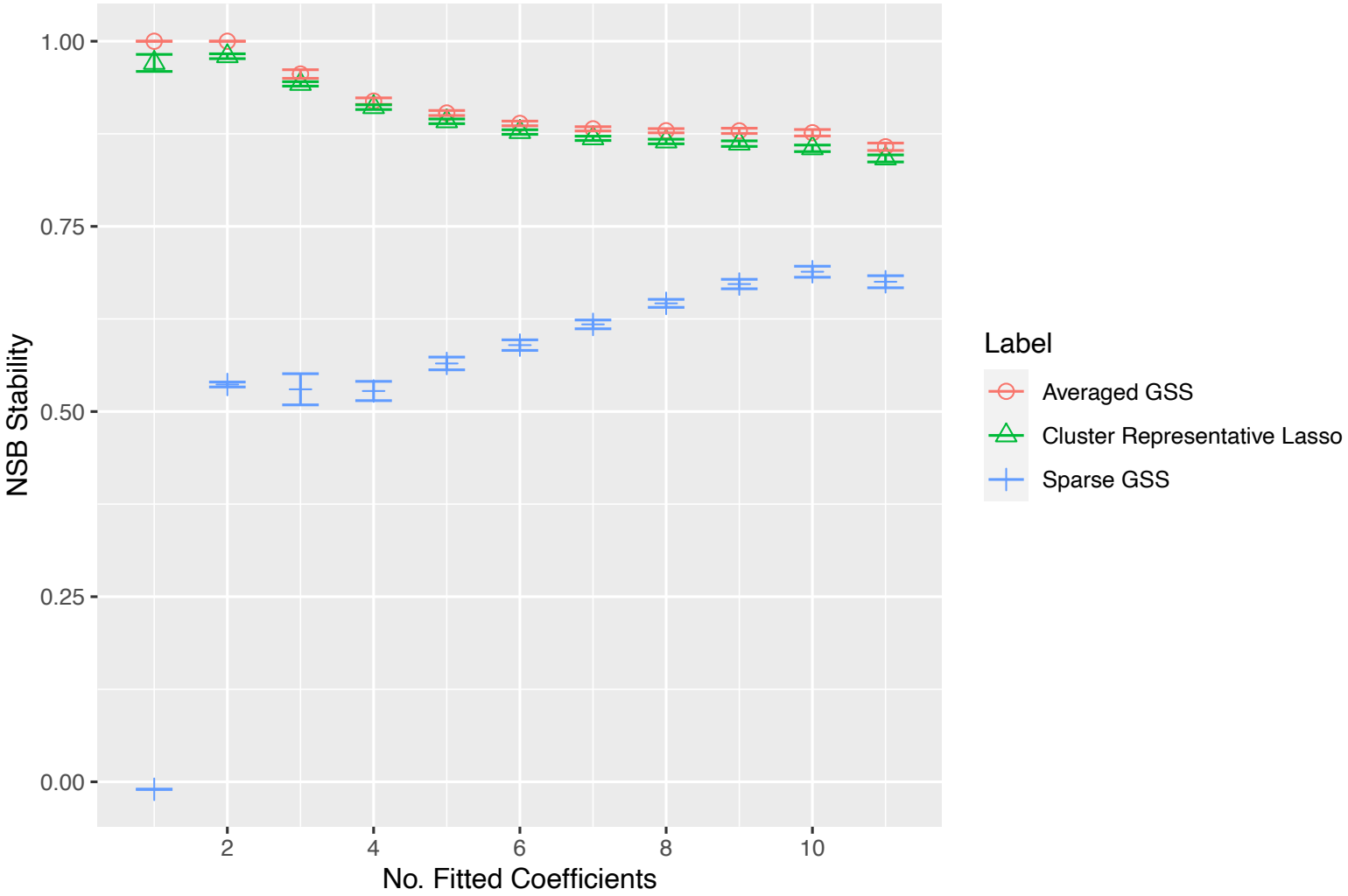
Generate a training set of size $n = 10{,}000$.

Estimate linear models on selected variables using OLS.

Calculate the mean squared error between each model's predictions and the expected value of the response:

$$\boldsymbol{\mu} = 1.5 \cdot \boldsymbol{Z} + \sum_{j=1}^{10} \frac{1}{\sqrt{j}} \cdot \boldsymbol{X}_{\cdot,(j+10)}.$$

Averaged Generalized Stability Selection



Outline

Simulation Studies: Setup

500 simulations using $n = 2500$, $p = 10$, and $K = 4$ ordered responses:

Draw $\boldsymbol{X} \in [-1,1]^{n \times p}$, where $X_{ij} \sim \text{Uniform}(-1,1)$ for all i, j.

Generate $\boldsymbol{y} \in \mathbb{R}^n$ using a relaxation of proportional odds:

$$\log\left(\frac{\mathbb{P}\left(y \leq k \mid \boldsymbol{x}\right)}{\mathbb{P}\left(y > k \mid \boldsymbol{x}\right)}\right) = \alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{x}, \qquad k \in \{1,\ldots,K-1\},$$

where $\boldsymbol{\alpha} = (0,4,6)$ (first two classes are common, last class is rare), $\boldsymbol{\beta}_1 = (1,\ldots,1)^\top$, and $\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1} + \boldsymbol{\psi}_k$ for random vectors $\boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_{K-1} \in \mathbb{R}^p$ (probability distributions specified later).

Estimate rare class probabilities by logistic regression on rare class, proportional odds, and PRESTO (penalty λ selected by cross-validation).

Calculate MSE of estimated rare class probabilities.

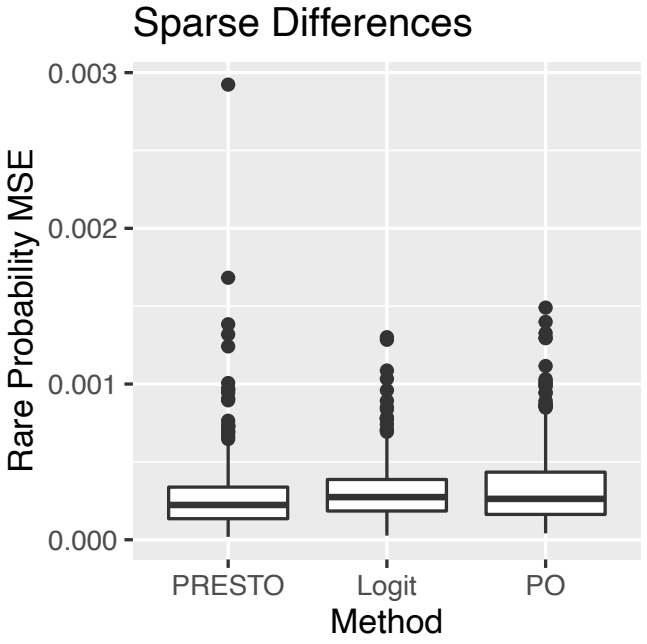
Simulation 1: Sparse Differences

$\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1} + \boldsymbol{\psi}_k$, where
$\boldsymbol{\beta}_1 = (1,\ldots,1)^\top$ and

$$\psi_{jk} = \begin{cases} 0, & \text{with probability } 2/3, \\ 0.5, & \text{with probability } 1/6, \\ -0.5, & \text{with probability } 1/6, \end{cases} \qquad j \in \{1,\ldots,p\}.$$

(Should be a favorable setting for PRESTO due to sparsity.)

The results suggest that PRESTO does in fact estimate the rare probabilities more accurately!



Simulation 2: Dense (Approximately Sparse) Differences

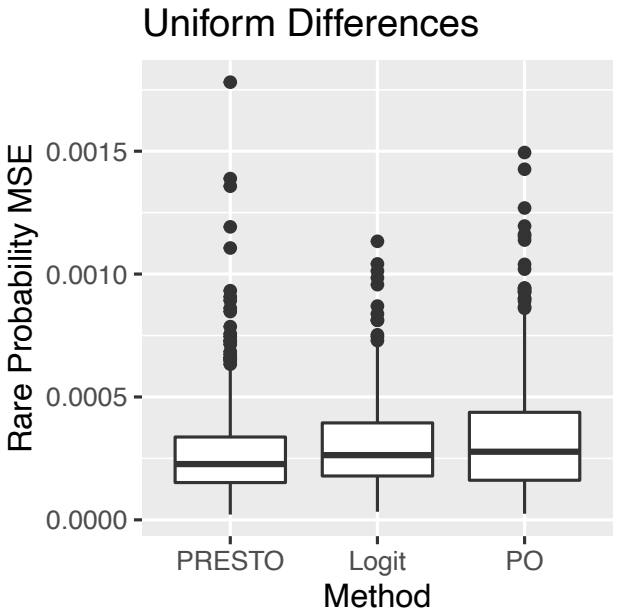
$\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1} + \boldsymbol{\psi}_k$, where
$\boldsymbol{\beta}_1 = (1,\ldots,1)^\top$ and

$$\psi_{jk} \sim \text{Uniform}(-0.5, 0.5), \quad j \in \{1,\ldots,p\}.$$

$\boldsymbol{\psi}_k$ are not sparse $\Longrightarrow$ should be harder for PRESTO.

But $\boldsymbol{\psi}_k$ can be considered "approximately" sparse (some small entries approximately equal to 0, limited number of large entries that are important to account for).

PRESTO still outperforms logistic regression and proportional odds!



Summary

Binary classifiers struggle to estimate rare probabilities (class imbalance).

If there are ordinal outcomes, a decision boundary with abundant data nearby can be leveraged to improve estimation of rare decision boundary (and rare probabilities),

Proportional odds model allows this, but imposes exactly equality of $\boldsymbol{\beta}$ vectors (unrealistically rigid).

PRESTO relaxes proportional odds, allowing $\boldsymbol{\beta}$ vectors to differ but imposing ℓ_1 penalty on differences.

This allows for best of both worlds: learn from abundant decision boundaries, but flexibly adapt for different decision boundaries between different outcomes.

References

G. James, D. Witten, T. Hastie, & R. Tibshirani. *An introduction to statistical learning with applications in R* (Vol. 112, p. 18). New York: Springer, 2021.

P. McCullagh. Regression models for ordinal data. J*ournal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.