Selective Inference via the Condition on Selection Framework: Inference after Variable Selection

Jason D. Lee

Stanford University

Advisors: Trevor Hastie and Jonathan Taylor
Slides at
`http://web.stanford.edu/~jdl17/selective_inference_regression.pdf`

Selective Inference

Selective Inference is about **testing hypotheses suggested by the data.**

Selective Inference is common. In many applications there is no hypothesis specified before data collection and exploratory analysis.

- Inference after variable selection. **Confidence intervals and p-values are only reported for the selected variables.**
- Exploratory Data Analysis by Tukey emphasized using data to suggest hypotheses, and post-hoc analysis to test these.
- Screening in Genomics, only select genes with large t-statistic or correlation.
- Peak/bump hunting in neuroscience, only study process when $X_t > \tau$ or critical points of the process.

Selective Inference

Conventional Wisdom (Data Dredging, Wikipedia)

> *A key point in proper statistical analysis is to test a hypothesis with data that was not used in constructing the hypothesis. (Data splitting)*

This talk

The Condition on Selection framework allows you to specify and test hypotheses using the same dataset.

1. Selective Inference
2. Bird’s-eye View of my Research
3. Condition on Selection Framework
    - Motivation: Inference after variable selection
    - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
    - Linear Regression Review
    - Related Work
    - Selection Events in Variable Selection
    - Truncated Gaussian Pivotal Quantity
    - Testing and Confidence Intervals
5. Experiments
6. Extensions
7. End

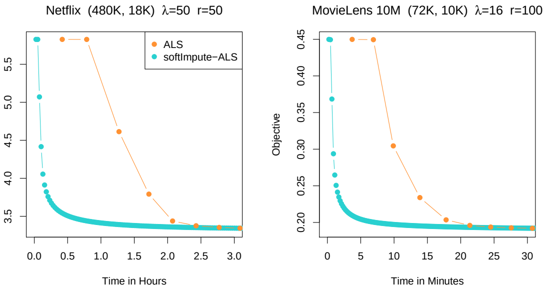
Overview of my Research

- **Machine Learning.** Large-scale matrix completion, non-negative matrix factorization, clustering, and communication-efficient sparse learning.
- **Statistical Methodology.** Selective inference, theory of regularized M-estimators, and high-dimensional statistical inference.
- **Large-scale Optimization.** Composite optimization, distributed optimization, and stochastic gradient intervals.

Fusing the three areas

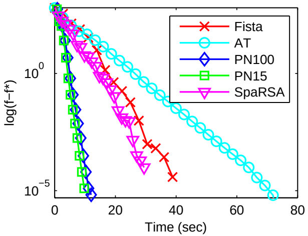
Selective Inference for Lasso = Convex Analysis/Opt + Statistics.

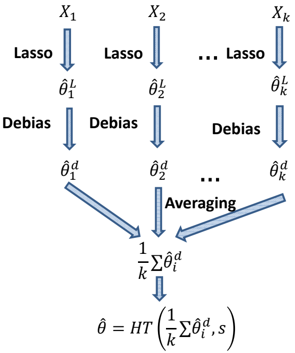
Scalable and Distributed Algorithms I



1. **Matrix Factorization.** `fast-ALS` reduces the computational cost and communication of `ALS` by $O(r)$, but with the same convergence rate. Apache Spark implementation takes 6 seconds per iteration for a $10^6 \times 10^6$ matrix with 10^9 non-zeroes, and 2 minutes per iteration for a $10^7 \times 10^7$ matrix with 10^9 non-zeroes.
2. **One-Pass Non-negative Matrix Factorization on Hadoop** Factorizes a 2TB heat transfer dataset in 50 min and factorizes a 345GB flow cytometry dataset in 20 min.

Scalable and Distributed Algorithms II





X_1 X_2 X_k

Lasso Lasso ... Lasso

$\hat{\theta}_1^L$ $\hat{\theta}_2^L$ $\hat{\theta}_k^L$

Debias Debias Debias

$\hat{\theta}_1^d$ $\hat{\theta}_2^d$ ... $\hat{\theta}_k^d$

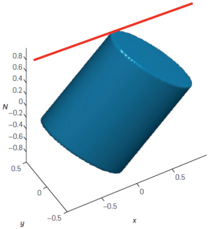
Averaging

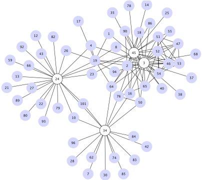
$\frac{1}{k}\Sigma\hat{\theta}_i^d$

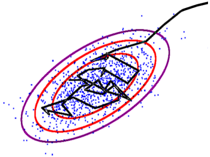
$\hat{\theta} = HT\left(\frac{1}{k}\Sigma\hat{\theta}_i^d, s\right)$

1. **Proximal Newton Method.** Designed and analyzed a second order algorithm for regularized estimators, which includes state-of-art methods such as `glmnet`,`liblinear`, and `quic`.
2. **Communication-efficient distributed sparse regression.** Proposed an algorithm for solving distributed sparse regression, which has the lowest communication cost among algorithms that achieve estimation rate $\left\|\hat{\beta} - \beta^0\right\|_2 \le \sqrt{\frac{slogp}{nk}}$.

High-dimensional model selection and estimation







- **Unified theory of model selection consistency in regularized estimators** of the form $\ell(\theta) + \rho(\theta)$. Provides structure learning consistency in graphical models, variable selection consistency in lasso, and rank consistency for nuclear norm (matrix completion).
- **Confidence intervals via stochastic gradient sampling.** SGD is one of the most common learning algorithms. We use iterates of SGD to build confidence intervals by using markov chain CLT.

Table of Contents

1. Selective Inference
2. Bird's-eye View of my Research
3. Condition on Selection Framework
    - Motivation: Inference after variable selection
    - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
    - Linear Regression Review
    - Related Work
    - Selection Events in Variable Selection
    - Truncated Gaussian Pivotal Quantity
    - Testing and Confidence Intervals
5. Experiments
6. Extensions
7. End

Motivation: Linear regression in high dimensions

1. Select relevant variables $\hat{M}$ via a variable selection procedure (k most correlated, lasso, forward stepwise ...).
2. Fit linear model using only variables in $\hat{M}$, $\hat{\beta}^{\hat{M}} = X_{\hat{M}}^{\dagger} y$.
3. Construct 90% z-intervals $(\hat{\beta}_j - 1.65\sigma_j, \hat{\beta}_j + 1.65\sigma_j)$ for selected variables $j \in \hat{M}$.

Are these confidence intervals correct?

Check by Simulation

- Generate design matrix $X \in \mathbf{R}^{n \times p}$ from a standard normal with $n = 20$ and $p = 200$.
- Let $y = \mathcal{N}(X\beta^0, 1)$.
- β^0 is 2 sparse with $\beta_1^0, \beta_2^0 = SNR$.
- Use marginal screening to select $k = 2$ variables, and then fit linear regression over the selected variables.
- Construct 90% confidence intervals for selected regression coefficients and check the coverage proportion.

Simulation

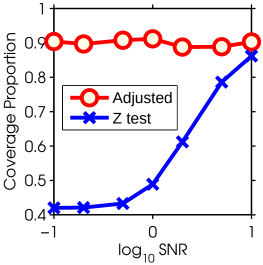


Figure: Plot of the coverage proportion across a range of SNR. The coverage proportion of the z intervals ($\hat{\beta} \pm 1.65\sigma$) is far below the nominal level of $1 - \alpha = .9$, even at SNR =5. The selective intervals (our method) always have coverage proportion $.9$.

Remember this simulation.

We will return to explain the results of the simulation.

1. Selective Inference
2. Bird's-eye View of my Research
3. Condition on Selection Framework
   - Motivation: Inference after variable selection
   - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
   - Linear Regression Review
   - Related Work
   - Selection Events in Variable Selection
   - Truncated Gaussian Pivotal Quantity
   - Testing and Confidence Intervals
5. Experiments
6. Extensions
7. End

Valid Selective Inference

Notation

- The selection function $\hat{H}$ selects the hypothesis of interest, $\hat{H}(y) : \mathcal{Y} \to \mathcal{H}$.
- $\phi(y; H)$ be a test of hypothesis H, so reject if $\phi(y; H) = 1$.
- $\{y : \hat{H}(y) = H\}$ is the selection event or "the set of y's that lead to selecting the same hypothesis H".

Definition (Selective type 1 error)

$\phi(y; \hat{H})$ is a valid selective test if

$$\mathbb{P}_0(\phi(y; \hat{H}(y)) = 1) \leq \alpha$$

Existing methods for Selective Inference

- Reduction to Simultaneous Inference: Assume that there is an apriori set of hypotheses $\mathcal{H}$ that could be tested. We can simultaneously control the type 1 error over all of $\mathcal{H}$, which implies selective type 1 error rate control for some selected $\hat{H}(y) \in \mathcal{H}$ (*e.g.* Scheffe's method).
- Data Splitting: Split the dataset $y = (y_1, y_2)$. Let $\hat{H}(y_1)$ be the selected hypothesis, and construct the test of $\hat{H}(y_1)$ using only y_2. Data splitting is "wasteful" in the sense that it is not using all the information in the first half of the data.

Condition on Selection Framework

Conditioning for Selective Type 1 Error Control

We can design a valid selective test ϕ by ensuring ϕ **is a valid test with respect to the distribution conditioned on the selection event meaning**

$$\mathbb{P}_0(\phi(y;H_i)=1|\hat{H}=H_i)\le\alpha$$

implies

$$\mathbb{P}_0(\phi(y;\hat{H}(y))=1)\le\alpha.$$

Intuition

By conditioning on the selection event, we are restricting to y's that would have led to the same hypothesis being tested. This lets us calibrate the test with respect to a distribution where the selection mechanism is deterministic and can be safely ignored.

.

More Formal argument

$$\begin{aligned}\mathbb{P}_0(\phi(y;\hat{H}(y))=1) &= \sum_i \mathbb{P}_0(\phi(y;H_i)=1|\hat{H}=H_i)\mathbb{P}_0(\hat{H}=H_i)\\ &\leq \alpha \sum_i \mathbb{P}_0(\hat{H}=H_i)\\ &\leq \alpha.\end{aligned}$$

Example: Maximum of a Normal

Let $y \sim \mathcal{N}(\mu, \Sigma)$. We would like to test $H_i : \mu_i = 0$ against the alternative $\mu_i > 0$. Let $\hat{H}(y) = H_{i^\star}$, where $y_{i^\star}$ is the maximum *i.e.* $y_{i^\star} = \max_i y_i$. We can make a test by rejecting when $y_{i^\star} > c$. Condition on selection tells us to choose the cutoff c so type 1 error rate under $y_{i^\star}|\{y_{i^\star} > \max_{i \neq i^\star} y_i\}$ is α.

Outline of the Rest of the Talk

1. Review linear regression, and cast inference after variable selection as a selective inference problem.
2. Many variable selection procedures have selection events $\{y : \hat{M}(y) = M\} = \{y : Ay \leq b\}$, where M is the subset of variables selected. Instead of selecting hypothesis, we will focus on selecting subsets of variables and testing regression coefficients of the selected variables.
3. An exact, valid selective test ϕ can be constructed for linear functions $H_0 : \eta(\hat{M}(y))^T \mu = \gamma$, using the Condition on Selection framework without sampling or numerical integration. This is done by deriving the conditional distribution of $\eta^T y$, which will be a truncated normal.
4. A selective test ϕ can be inverted to make a selective confidence interval. These confidence intervals control false coverage rate, a selective type 1 error metric (Benjamini & Yekutieli 2005).

1. Selective Inference
2. Bird's-eye View of my Research
3. Condition on Selection Framework
   - Motivation: Inference after variable selection
   - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
   - Linear Regression Review
   - Related Work
   - Selection Events in Variable Selection
   - Truncated Gaussian Pivotal Quantity
   - Testing and Confidence Intervals
5. Experiments
6. Extensions
7. End

Setup

Model

- Assume that $y_i = \mu(x_i) + \epsilon_i$
- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- $x_i \in \mathbf{R}^p$, $y \in \mathbf{R}^n$, and $\mu = \begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{bmatrix}$.
- Design matrix $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbf{R}^{n \times p}$.

Review of Linear Regression

The best linear approximation is $\beta^0 = X^\dagger \mu$. Linear regression estimates this using

$$\hat{\beta} = X^\dagger y = (X^T X)^{-1} X^T y.$$

Theorem

The least squares estimator is distributed

$$\hat{\beta} \sim \mathcal{N}(\beta^0, \sigma^2 (X^T X)^{-1})$$

and

$$\Pr\left(\beta_j^0 \in \left[\hat{\beta}_j \pm z_\alpha \sigma (X^T X)_{jj}^{-1/2}\right]\right) = 1 - \alpha.$$

Explaining the simulation

1. The confidence intervals rely on the result that $\hat{\beta}$ is Gaussian.
2. The variable selection procedure (marginal screening) returns a set of variables $\hat{M}(y)$. In particular,

$$|X_{\hat{M}}^T y| > |X_{-\hat{M}}^T y|.$$

3. For any fixed set M, $X_M^T y$ is Gaussian, but $X_{\hat{M}}^T y$ is not Gaussian!

Example

Let $y \sim \mathcal{N}(0, I)$, and $X = I$. Let $i^\star = \arg\max y_i$, then $y_{i^\star}$ is not Gaussian. The interval $y_{i^\star} \pm 2$ is not a valid 95% confidence interval!

Three possible population parameters

Population parameters

1. **Sub-model parameter.** $\beta^M = (X_M^T X_M)^{-1} X_M^T \mu = X_M^\dagger \mu$ (advocated by the POSI group).
2. **OLS parameter.** In the $n \geq p$ regime without the linear model assumption, $\beta^\star = (X^T X)^{-1} X^T \mu = X^\dagger \mu$ is the best linear approximation.
3. **The "true" parameter for $p > n$.** Assuming a sparse linear model $\mu = X\beta^0$, the parameter of interest is β^0.

For the talk, I will focus on inference for the sub-model parameter β^M, but analagous results hold for the other two.

Selective Inference in Linear Regression

Selective Inference reduces to testing $\eta(\hat{M}(y))^T\mu$.

1. Sub-model parameter. $\beta_j^{\hat{M}} = e_j^T X_{\hat{M}}^\dagger \mu = \eta(\hat{M}(y))^T\mu$, where $\eta(\hat{M}(y))^T$ is row of $X_{\hat{M}}^\dagger$.
2. OLS parameter. $e_j^T \beta_{\hat{M}}^\star = e_{j'}^T X^\dagger \mu = \eta(\hat{M}(y))^T\mu$.
3. True parameter. Under the scaling $n \gg s^2 \log^2 p$ and restricted eigenvalue assumptions, there is a parameter β^d that satisfies $\sqrt{n}\left\|\beta^d(\hat{M}) - \beta^0\right\|_\infty = o(1)$, and β^d is a linear function of μ. Valid selective inference for β^d implies asymptotically valid selective inference for β^0.

Testing regression coefficients reduce to testing an adaptive/selected linear function of μ

$$H_0 : \eta(\hat{M}(y))^T\mu = \gamma.$$

1. Selective Inference
2. Bird's-eye View of my Research
3. Condition on Selection Framework
   - Motivation: Inference after variable selection
   - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
   - Linear Regression Review
   - Related Work
   - Selection Events in Variable Selection
   - Truncated Gaussian Pivotal Quantity
   - Testing and Confidence Intervals
5. Experiments
6. Extensions
7. End

Related Work

- Significance testing for Lasso (Lockhart et al. 2013) tests for whether all signal variables are found. Our framework allows us to test the same thing with no assumptions on X and is completely non-asymptotic and exact.
- POSI (Berk et al. 2013) widen intervals to simultaneously cover all coefficients of all possible submodels. POSI is an example of reducing selective to simultaneous inference, and protects against *any* selection procedures.
- Asymptotic normality by "inverting" KKT conditions (Zhang and Zhang 2012, Van de Geer et al. 2013, Javanmard and Montanari 2013, Chernozhukov et al. 2013). Asymptotic result requires consistency of the lasso, and computational cost equivalent to solving p lasso's.
- Knockoff for FDR control in linear regression (Foygel and Candes 2014) allows for exact FDR control for $n \geq p$ without any assumptions on X.

1. Selective Inference
2. Bird's-eye View of my Research
3. Condition on Selection Framework
   - Motivation: Inference after variable selection
   - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
   - Linear Regression Review
   - Related Work
   - Selection Events in Variable Selection
   - Truncated Gaussian Pivotal Quantity
   - Testing and Confidence Intervals
5. Experiments
6. Extensions
7. End

Marginal screening or k-most correlated

Algorithm 1 Marginal screening algorithm

```
**Input:** Design matrix $X$, response $y$, and model size $k$.
Compute $|X^T y|$.
Let $\hat{M}$ be the index of the $k$ largest entries of $|X^T y|$.
Compute $\hat{\beta}_{\hat{M}} = (X_{\hat{M}}^T X_{\hat{M}})^{-1} X_{\hat{M}}^T y$
```

Warmup: Marginal screening selection event

Marginal Screening Selection Event

The marginal screening selection event is a subset of $\mathbf{R}^n$:

$$\begin{aligned}&\left\{y : \hat{s}_i x_i^T y > \pm x_j^T y, \text{ for each } i \in \hat{M} \text{ and } j \in \hat{M}^c\right\}\\ =&\left\{y : A(\hat{M}, \hat{s})y \leq b(\hat{M}, \hat{s})\right\}\end{aligned}$$

The marginal screening selection event corresponds to selecting a set of variables $\hat{M}$, and those variables having signs $\hat{s} = \text{sign}\left(X_{\hat{M}}^T y\right)$.

Lasso selection event

Lasso Selection Event

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2} \left\| y - X\beta \right\|^2 + \lambda \left\| \beta \right\|_1$$

From KKT conditions, a set of variables $\hat{M}$ is selected with sign$(\hat{\beta}_{\hat{M}}) = \hat{s}$ iff

$$\left\{ y : \mathrm{sign}(\beta(\hat{M}, \hat{s})) = \hat{s}, \left\| Z(\hat{M}, \hat{s}) \right\|_\infty < 1 \right\} = \{ y : Ay \leq b \}$$

This says that the inactive subgradients are strictly dual feasible, and the signs of the active subgradient agrees with the sign of the lasso estimate.

$$\beta(M, s) := (X_M^T X_M)^{-1}(X_M^T y - \lambda s)$$

$$Z(M, s) := X_{M^c}^T X_M (X_M^T X_M)^{-1} s + \frac{1}{\lambda} X_{M^c}^T (I - X_M (X_M^T X_M)^{-1} X_M^T) y.$$

Selection event

Selection events correspond to affine regions.

$$\{\hat{M}(y) = M\} = \{Ay \leq b\} \text{ \& } y|\{\hat{M}(y) = M\} \overset{d}{=} \mathcal{N}(\mu, \Sigma)\big|\{Ay \leq b\}$$

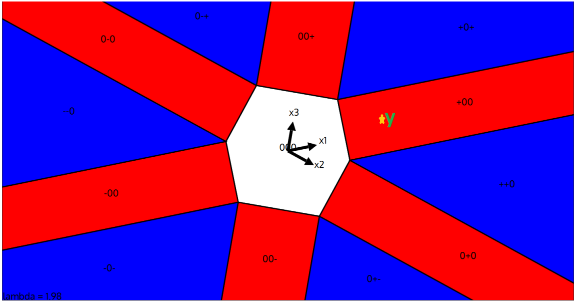


Figure: $(n, p) = (2, 3)$. White, red, and blue shaded regions correspond to different selection events. The shaded region that y falls into is where lasso selects variable 1 with positive sign.`http://naftaliharris.com/blog/lasso-polytope-geometry/`

1. Selective Inference
2. Bird's-eye View of my Research
3. Condition on Selection Framework
   - Motivation: Inference after variable selection
   - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
   - Linear Regression Review
   - Related Work
   - Selection Events in Variable Selection
   - Truncated Gaussian Pivotal Quantity
   - Testing and Confidence Intervals
5. Experiments
6. Extensions
7. End

Constrained Gaussian

Constrained Gaussians

- The distribution of $y \sim \mathcal{N}(\mu, \sigma^2 I)$ conditional on $\{y : Ay \leq b\}$ has density $\frac{1}{\mathbf{Pr}(Ay \leq b)} \phi(y; \mu, \Sigma) \mathbb{1}(Ay \leq b)$.
- Ideally, we would like to sample from the density to approximate the sampling distribution of our statistic under the null. This is computationally expensive.
- For testing regression coefficients, we only need distribution of $\eta^T y | \{Ay \leq b\}$.

Computationally Tractable Inference

It turns out

$$\eta^T y \Big| \{Ay \leq b, P_{\eta^\perp} y\} \overset{d}{=} \text{TruncatedNormal}.$$

Using this distributional result, we avoid sampling and integration.

Geometric Intuition

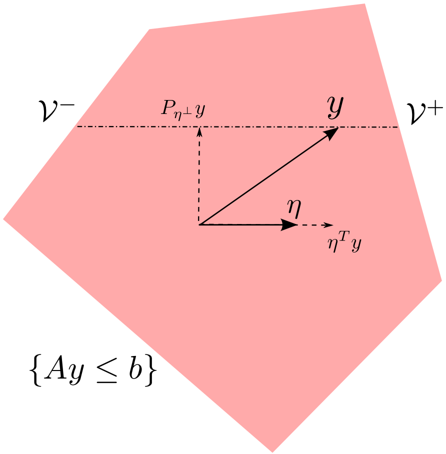


Figure: A picture demonstrating that the set $\{Ay \leq b\}$ can be characterized by $\{\mathcal{V}^- \leq \eta^T y \leq \mathcal{V}^+\}$. Assuming $\Sigma = I$ and $||\eta||_2 = 1$, $\mathcal{V}^-$ and $\mathcal{V}^+$ are functions of $P_{\eta^\perp} y$ only, which is independent of $\eta^T y$.

Truncated Normal

Corollary

The distribution of $\eta^T y$ conditioned on $\{Ay \leq b, P_{\eta^\perp} y\}$ is a (univariate) Gaussian truncated to fall between $\mathcal{V}^-(P_{\eta^\perp} y)$ and $\mathcal{V}^+(P_{\eta^\perp} y)$, *i.e.*

$$\eta^T y \mid \{Ay \leq b, P_{\eta^\perp} y\} \sim TN(\eta^T \mu, \sigma^2 \left\|\eta\right\|^2, \mathcal{V}^-, \mathcal{V}^+)$$

$TN(\mu, \sigma, a, b)$ is the normal distribution truncated to lie between a and b.

Pivotal quantity

Theorem

Let $F(x;\mu,\sigma^2,a,b)$ *denote the CDF of* $TN(\mu,\sigma,a,b)$. *Then* $F(\eta^T y;\eta^T\mu,\sigma^2\left\|\eta\right\|^2,\mathcal{V}^-,\mathcal{V}^+)$ *is a pivotal quantity*

$$F(\eta^T y;\eta^T\mu,\sigma^2\left\|\eta\right\|^2,\mathcal{V}^-,\mathcal{V}^+) \sim \textit{Unif}(0,1).$$

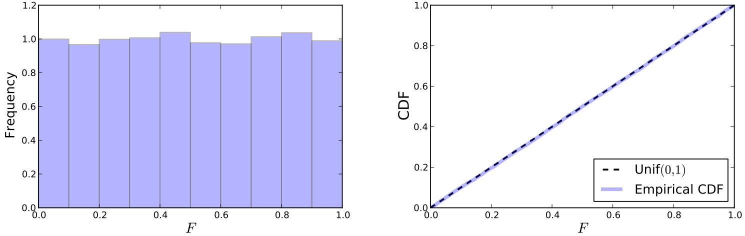


Figure: Pivot is uniform.

1. Selective Inference
2. Bird's-eye View of my Research
3. Condition on Selection Framework
   - Motivation: Inference after variable selection
   - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
   - Linear Regression Review
   - Related Work
   - Selection Events in Variable Selection
   - Truncated Gaussian Pivotal Quantity
   - Testing and Confidence Intervals
5. Experiments
6. Extensions
7. End

Testing regression coefficients

Coefficients of selected variables are adaptive linear functions

Recall, $\beta^{\hat{M}} = X^{\dagger}_{\hat{M}}\mu$, and $\hat{\beta}_{\hat{M}} = X^{\dagger}_{\hat{M}} y$. By choosing $\eta_j = X^{\dagger T}_{\hat{M}} e_j$, we have $\eta_j^T y = \hat{\beta}_j^{\hat{M}}$.

Theorem

Let $H_0 : \beta_j^{\hat{M}} = \beta_j$. The test that rejects if

$$F(\hat{\beta}_j^{\hat{M}}; \beta_j; \sigma^2 \|\eta\|^2, \mathcal{V}^-, \mathcal{V}^+) \notin \left(\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right)$$

is a α-level selective test of H_0. Choice of $(\frac{\alpha}{2}, 1 - \frac{\alpha}{2})$ is arbitrary. We can optimize endpoints to $(a, 1 - \alpha + a)$ such that the interval is UMPU, at the cost of more computation.

Other parameter targets

The same result holds exactly for the OLS parameter $\beta^\star$ and β^d.

Algorithm 2 Hypothesis test for selected variables

```
Input: Design matrix $X$, response $y$, model size $k$.
Use variable selection method (marginal screening or Lasso) to select a subset of variables $\hat{M}$.
Test $H_0 : \beta_j^{\hat{M}} = \beta_j$.
Let $A = A(\hat{M}, \hat{s})$ and $b = b(\hat{M}, \hat{s})$. Let $\eta_j = (X_{\hat{M}}^T)^\dagger e_j$.
Compute $F(\hat{\beta}_j^{\hat{M}}; \beta_j,\ \sigma^2 ||\eta_j||^2, \mathcal{V}^-, \mathcal{V}^+)$.
Output: Reject if $F(\hat{\beta}_j^{\hat{M}}; \beta_j; \sigma^2 \left\|\eta\right\|^2, \mathcal{V}^-, \mathcal{V}^+) \notin \left(\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right)$.
```

Confidence Intervals by Inverting

Confidence Intervals

Confidence interval C_j is all β_j's, where a test of $H_0 : \beta_j^{\hat{M}} = \beta_j$ fails to reject at level α.

$$C_j = \{\beta_j : \frac{\alpha}{2} \leq F(\hat{\beta}_j^{\hat{M}}; \beta_j,\ \sigma^2||\eta_j||^2, \mathcal{V}^-, \mathcal{V}^+) \leq 1 - \frac{\alpha}{2}\}$$

Interval $[L_j, U_j]$ is found by univariate root-finding on a monotone function. Solve

$$F(\hat{\beta}_j^{\hat{M}}; L_j,\ \sigma^2||\eta_j||^2, \mathcal{V}^-, \mathcal{V}^+) = \frac{\alpha}{2}$$
$$F(\hat{\beta}_j^{\hat{M}}; U_j,\ \sigma^2||\eta_j||^2, \mathcal{V}^-, \mathcal{V}^+) = 1 - \frac{\alpha}{2}$$

Similarly, the endpoints are arbitrary and can be chosen to be UMAU.

Selective Inference and FCR

Lemma (Valid selective confidence intervals)

These are valid selective intervals,

$$\mathbf{Pr}\left(\beta_j^{\hat{M}} \in C_j\right) = 1 - \alpha.$$

False coverage rate (FCR) is the confidence interval analog of FDR. FCR is also controlled.

$$\mathbb{E}\left[\frac{\textit{Number not covering and constructed}}{\textit{Number constructed}}\right] \leq \alpha$$

1. Selective Inference
2. Bird's-eye View of my Research
3. Condition on Selection Framework
   - Motivation: Inference after variable selection
   - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
   - Linear Regression Review
   - Related Work
   - Selection Events in Variable Selection
   - Truncated Gaussian Pivotal Quantity
   - Testing and Confidence Intervals
5. Experiments
6. Extensions
7. End

Comparison to data splitting and POSI

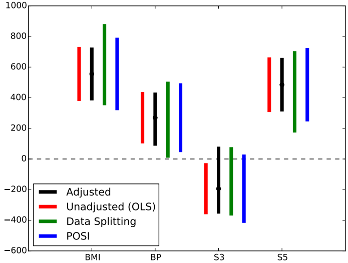


Figure: Diabetes dataset comparison of four methods: selection-adjusted (black), unadjusted (red), data splitting (green), and POSI (blue).

- Variable S3 is no longer significant after adjusting for model selection.
- Our selective intervals are approximately the same as the z-intervals for significant variables. PoSI creates narrower intervals than data splitting (1.36 vs $\sqrt{2}$ wider than nominal). Data splitting is inadmissible since a variant Condition on Selection dominates data splitting.

Non-Gaussian noise and estimated σ^2

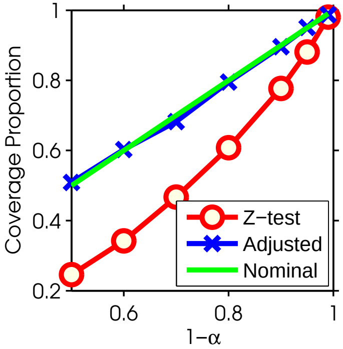


Figure: Plot of nominal coverage vs the actual coverage proportion for diabetes dataset. Simulation is done by using 2000 iterations of residual bootstrap with estimated $\hat{\sigma}^2$. The adjusted intervals always cover at the nominal level, whereas the z-test is always below.

Selective Intervals for sparse β^0 in $p > n$

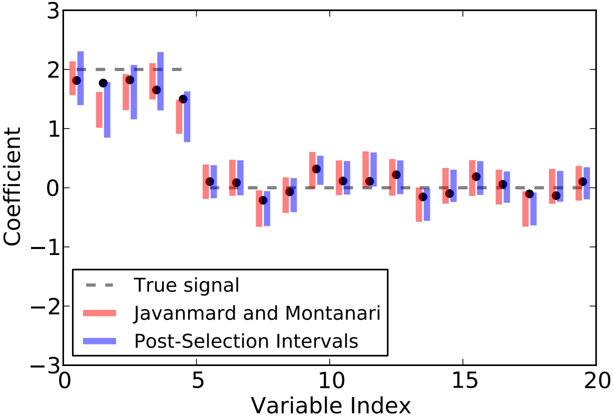


Figure: $(n, p, s) = (25, 50, 5)$ with only the first 20 coefficients being plotted. Data is generated from $y = X\beta^0 + \epsilon$ with a SNR of 2. The Javanmard-Montanari high-dimensional z-intervals do not guarantee selective coverage.

1. Selective Inference
2. Bird's-eye View of my Research
3. Condition on Selection Framework
   - Motivation: Inference after variable selection
   - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
   - Linear Regression Review
   - Related Work
   - Selection Events in Variable Selection
   - Truncated Gaussian Pivotal Quantity
   - Testing and Confidence Intervals
5. Experiments
6. **Extensions**
7. End

Follow-up work

- Testing the goodness of fit of the selected model, $H_0 : (I - P_{\hat{M}})\mu = 0$ (Lee et al. 2013)
- Selective t-tests for unknown σ^2. Non-Gaussian noise can be handled via CLT (Tian & Taylor 2015).
- Non-affine regions, only need to intersect a ray with the region to design exact conditional tests, which can be done by root-finding for "nice" sets (Lee et al. 2013, Loftus and Taylor 2014).
- Marginal screening followed by Lasso (Lv & Fan 2008), forward stepwise regression, isotonic regression, elastic net, AIC/BIC criterion with subset selection, λ chosen via hold-out set, square-root lasso (Lee & Taylor 2014, Tian et al. 2015).
- Use first half of data to select model, then do inference using the entire dataset via putting constraints only on the first half. This variant of Condition on Selection selects the same model as data splitting, but is more powerful under a strong screening assumption (Fithian, Sun, Taylor 2014).

Improving Power

Intuition: Condition on less.

- **“Selected” Model (Fithian, Sun, Taylor 2014)** If $P^{\perp}_{\hat{M}}\mu = 0$ (screening) , then we can condition on only $P_{\hat{M}-j}y$ instead of $P^{\perp}_{\eta}y$. This results in exactly the same test, since $\eta^T y$ is conditionally independent of $P^{\perp}_{\hat{M}}y$. If you run selection procedure (lasso) on only half the data $(A_1 y_1 \leq b_1)$ and use all of the data for inference, then the “selected test” benefits from conditioning on less. This test statistic can be more powerful, but requires MCMC. If screening is violated, type 1 error is not controlled, so this modification should only be used when the user is confident in the screening property.
- **Union over signs (Lee et al. 2013).** For lasso and screening, we conditioned on signs and the selected variables. We can union over all $2^{|M|}$ signs to condition on a larger set. $\eta^T y | \{P_{\eta^\perp} y, \hat{M} = M\}$ is a truncated Gaussian on a union of intervals. **Union over signs makes a huge difference for lasso.**

Beyond Selective Inference: Combinatorial Detection

Motivating example: Submatrix Detection/Localization problem (Ma and Wu 2014, Balakrishnan and Kolar 2012) with scan statistic $y^\star = \max_{C \in \mathcal{S}} \sum_{i \in C} y_i$.

- Exact tests can be designed for the intractable global maximizer statistic, and the tractable sum-test. The tests have type 1 error exactly α and detection thresholds that match the minimax analysis.
- **Heuristic greedy algorithm.** Shabalin and Nobel 2013 propose a greedy algorithm to approximate the global maximizer. By conditioning on the "path" of greedy algorithm, we obtain an exact test for the output of the greedy algorithm!

Future Work

- Non-convex regularizers (SCAD, MCP). The selection event depends on the *optimization algorithm* and the optimality conditions.
- More automatic way of defining the selection event without specific analysis for each (algorithm,hypothesis class) pair. Can it be generated in an online fashion as a primal-dual solver progresses *e.g.* using dual solutions?
- Given a single dataset and class of queries/tests, can we control validity of an adaptive sequence of queries/tests? **Implication: This would allow different research groups to share a dataset and formulate hypotheses after observing the outcome of a previous group's study.**

1. Selective Inference
2. Bird's-eye View of my Research
3. Condition on Selection Framework
   - Motivation: Inference after variable selection
   - Formalizing Selective Inference
4. Inference after Variable Selection in Regression
   - Linear Regression Review
   - Related Work
   - Selection Events in Variable Selection
   - Truncated Gaussian Pivotal Quantity
   - Testing and Confidence Intervals
5. Experiments
6. Extensions
7. End

Conclusion

Takeaways

- Standard workflow of selecting relevant variables, then reporting confidence intervals/ significance tests only of selected variables is incorrect.
- Condition on Selection allows you to properly account for the selection procedure by calibrating tests with respect to the conditional distribution. The computation time is negligible compared to the selection algorithm.
- The Condition on Selection framework applies to a large class of selection procedures: AIC, BIC, lasso, forward stepwise, marginal screening, CV for λ, combinatorial testing . . .
- Science is sequential. There is important future work in controlling the type 1 error of adaptively sequence of queries/tests.

Acknowledgments

References:

1. Jason D. Lee and Jonathan Taylor, *Exact statistical inference after marginal screening.*
2. Jason D. Lee, Dennis L Sun, Yuekai Sun, and Jonathan Taylor, *Exact post-selection inference with the Lasso.*
3. Jason D. Lee, Yuekai Sun, and Jonathan Taylor, *Evaluating the Statistical Significance of Biclusters and other Combinatorial Structures.*

Papers available at `http://stanford.edu/~jdl17/`

Thanks for Listening!

High-dimensional case

Assume that $y = X\beta^0 + \epsilon$.

What η should we use?

We can test any $\eta^T\mu = \gamma$, so how should we choose η?
Answer: Debiased Estimator.

$$\hat{\beta}^d := \hat{\beta} + \frac{1}{n}\Theta X^T(y - X\hat{\beta})$$

Observation 1: If $n \geq p$ and $\Theta = \hat{\Sigma}^{-1}$, then $\hat{\beta}^d = \hat{\beta}^{LS}$. **This suggests that we should choose an η corresponding "somehow" to the debiased estimator because this worked in the low-dimensional regime.**
Observation 2: The debiased estimator is affine in y, if the active set and signs of the active set are considered fixed.

Recall that $\hat{\beta} = \begin{bmatrix} (X_{\hat{M}}^T X_{\hat{M}})^{-1} X_{\hat{M}}^T y - \lambda(\frac{1}{n} X_{\hat{M}}^T X_{\hat{M}})^{-1} s_{\hat{M}} \\ 0 \end{bmatrix}$.

Plug this into $\hat{\beta}^d = \hat{\beta} + \frac{1}{n}\Theta X^T(y - X\hat{\beta})$ to get

$$\hat{\beta}^d = \frac{1}{n}\Theta X^T y + (I - \Theta\hat{\Sigma}) \begin{bmatrix} (X_{\hat{M}}^T X_{\hat{M}})^{-1} X_{\hat{M}}^T y - \lambda(\frac{1}{n} X_{\hat{M}}^T X_{\hat{M}})^{-1} s_{\hat{M}} \\ 0 \end{bmatrix}$$

Main Idea

Replace y with μ to make a population version.

$$\begin{aligned} \beta^d(\hat{M}, \hat{s}) :&= \frac{1}{n}\Theta X^T \mu + (I - \Theta\hat{\Sigma}) \begin{bmatrix} (X_{\hat{M}}^\dagger \mu - \lambda(\frac{1}{n} X_{\hat{M}}^T X_{\hat{M}})^{-1} s_{\hat{M}} \\ 0 \end{bmatrix} \\ &= B\mu + h \end{aligned}$$

β^d **is an affine function of** μ**.**

Selective inference for β^d.

Condition on Selection framework allows you to make a selective confidence interval for $\beta^d_{\hat{M}}$.

Selective intervals for β^d

Choose $\eta = e_j^T B$. We would like to test $\beta_j^d = \gamma$, which is equivalent to

$$\eta^T \mu = \gamma - \eta^T h = \tilde{\gamma}.$$

Thus using the framework we get,

$$\Pr(\beta^d_{j,\hat{M}} \in C_j) = 1 - \alpha.$$

Why should you care about covering β^d???

Theorem

Under $X_i \sim \mathcal{N}(0, \Sigma)$ and $n > s^2 \log^2 p$ (same assumptions as Javanmard & Montanari 2013, Zhang and Zhang 2012, and Van de Geer et al. 2014)

$$\left\| \beta^d(\hat{M}, \hat{s}) - \beta^0 \right\|_\infty \leq C \frac{s \log p}{n}.$$

Theorem

Under the same conditions as above and for any $\delta > 0$,

$$\Pr(\beta^d_{j,\hat{M}} \in C_j \pm \frac{\delta}{\sqrt{n}}) \geq 1 - \alpha$$

Valid Selective Inference

Notation

- The selection function $\hat{H}$ selects the hypothesis of interest, $\hat{H}(y) : \mathcal{Y} \to \mathcal{H}$.
- $\phi(y; H)$ be a test of hypothesis H, so reject if $\phi(y; H) = 1$.
- $\phi(y; H)$ is a valid test of H if $\mathbb{P}_0(\phi(y; H) = 1) \leq \alpha$.
- $\{y : \hat{H}(y) = H\}$ is the selection event.
- $F \in N(H)$ if F is a null distribution with respect to H.

Definition

$\phi(y; \hat{H})$ is a valid selective test if

$$\mathbb{P}_F(\phi(y; \hat{H}(y)) = 1 | F \in N(\hat{H})) \leq \alpha$$

Condition on Selection Framework

Conditioning for Selective Type 1 Error Control

We can design a valid selective test ϕ by ensuring ϕ **is a valid test with respect to the distribution conditioned on the selection event meaning**

$$\forall F \in N(H_i),\ \mathbb{P}_F(\phi(y; H_i) = 1 | \hat{H} = H_i) \leq \alpha,$$

then

$$\begin{aligned}
&\mathbb{P}_F(\phi(y; \hat{H}(y)) = 1 | F \in N(\hat{H})) \\
&= \sum_{i: F \in N(H_i)} \mathbb{P}_F(\phi(y; H_i) = 1 | \hat{H} = H_i) \mathbb{P}_F(\hat{H} = H_i | F \in N(\hat{H})) \\
&\leq \alpha \sum_{i: F \in N(H_i)} \mathbb{P}_F(\hat{H} = H_i | F \in N(\hat{H})) \\
&\leq \alpha
\end{aligned}$$

Lasso Selective Intervals

Solve Lasso at some λ, and construct confidence intervals using previous algorithm.

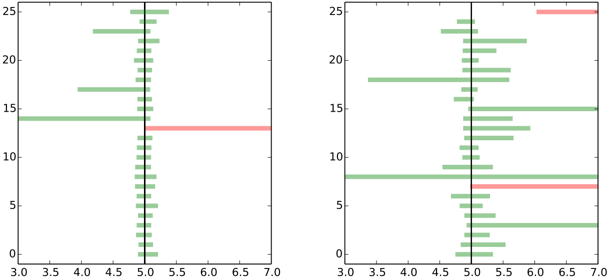


Figure: 90% confidence intervals for $\beta_1^{\hat{M}}$ for two different settings $(n, p) = (100, 50)$ and $(n, p) = (100, 200)$, over 25 simulated data sets. The truth β^0 has five non-zero coefficients, all set to 5.0, and the noise variance is 0.25. A green bar means the confidence interval covers the true value while a red bar means otherwise.

Coarsest Selection Event

Coarsest selection event

Recall that a subset/sign pair (S, s) is in bijection with a selection event. We only need to condition on the selection for the variables S, which determines η. Selection event for only variables S:

$$\begin{aligned}\left\{y : \hat{M}(y) = M\right\} &= \bigcup_{s \in \{-1,1\}^{|\hat{M}|}} \{y : (\hat{M}(y), s(y)) = (M, s)\} \\ &= \bigcup_{s \in \{-1,1\}^{|\hat{M}|}} \{y : A(M, s)y \leq b(M, s)\}\end{aligned}$$

- Condition on the coarsest partition where η is still measurable.
- The set is a union of linear constraints. Pivotal quantity, hypothesis tests, and intervals are valid for union of linear constraints.
- Strictly more powerful, and empirically performs well.

Coarsest Selection event

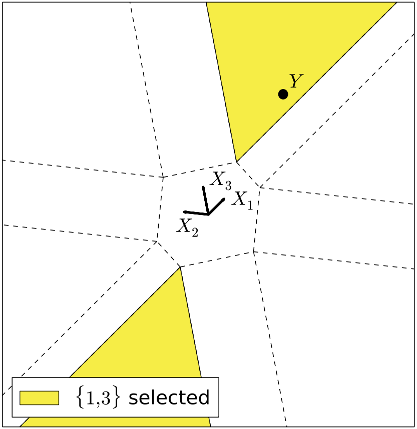


Figure: Coarsest selection event corresponds to the union of the two yellow regions. Before, we conditioned on one of the yellow wedges.

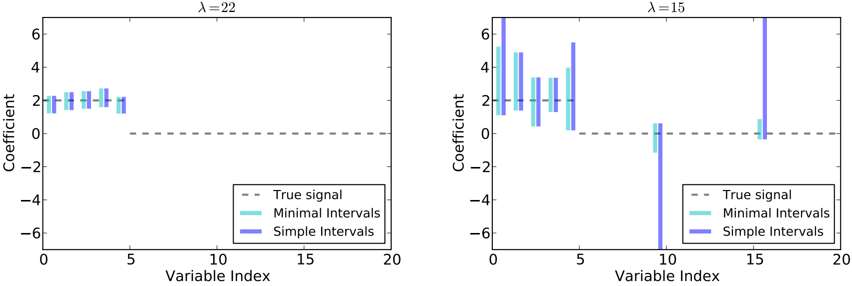


Figure: Light blue intervals are using the coarsest selection event or union of regions and dark blue are using the selection event that is one region. The simulated data featured $n = 25$, $p = 50$, and 5 true non-zero coefficients; only the first 20 coefficients are shown. The simple intervals are as good as the minimal intervals on the left plot; the advantage of the minimal intervals is realized when the estimate is unstable and the simple intervals are very long, as in the right plot.

Testing goodness-of-fit

We would like to test

$$H_0 : {\beta^0}_{-\hat{S}} = 0.$$

This means that all the true signal variables have been found, support$(\beta^0) \subset \hat{S}$.

We can test this by checking whether the unselected variables help explain the residual, or $H_0 : \left\|(I - P_{\hat{S}})\mu\right\|_\infty = 0$.

Testing goodness-of-fit

Letting $j^\star := \text{argmax}_j \; |e_j^T(I - P_{\hat{S}})y|$ and $s_j := \operatorname{sign}(e_j^T(I - P_{\hat{S}})y)$, we set

$$\eta_{j^\star} = s_{j^\star}(I - P_{\hat{S}})e_{j^\star},$$

and test $H_0 : \eta_{j^\star}^T \mu = 0$. This is a linear function of y.

Corollary

Let $H_0 : \left\|(I - P_{\hat{S}})\mu\right\|_\infty = 0$. Then, the test which rejects when

$$\left\{ F_{0,\ \sigma^2 ||\eta_j^*||^2}^{[\mathcal{V}^-, \mathcal{V}^+]}(\eta_{j^\star}^T y) > 1 - \alpha \right\}$$

is level α,

$$\mathbb{P}\left(F_{0,\ \sigma^2 ||\eta_{j^\star}||^2}^{[\mathcal{V}^-, \mathcal{V}^+]}(\eta_{j^\star}^T y) > 1 - \alpha \;\middle|\; H_0 \right) = \alpha.$$

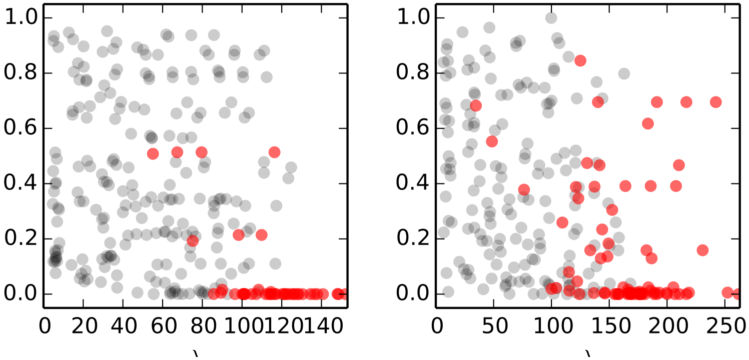


Figure: P-values for $H_{0,\lambda}$ at various λ values for a small ($n = 100$, $p = 50$) and a large ($n = 100$, $p = 200$) uncorrelated Gaussian design, computed over 50 simulated data sets. The true model has three non-zero coefficients, all set to 1.0, and the noise variance is 2.0. We see the p-values are $\mathrm{Unif}(0, 1)$ when the selected model includes the truly relevant predictors (black dots) and are stochastically smaller than $\mathrm{Unif}(0, 1)$ when the selected model omits a relevant predictor (red dots).

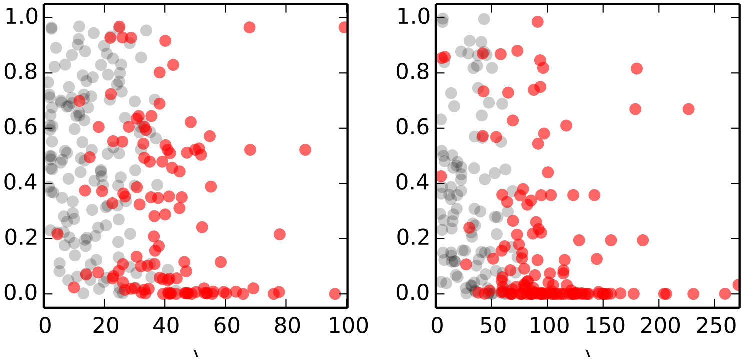


Figure: P-values for $H_{0,\lambda}$ at various λ values for a small $(n = 100,\, p = 50)$ and a large $(n = 100,\, p = 200)$ *correlated* $(\rho = 0.7)$ Gaussian design, computed over 50 simulated data sets. The true model has three non-zero coefficients, all set to 1.0, and the noise variance is 2.0. Since the predictors are correlated, the relevant predictors are not always selected first. However, the p-values remain uniformly distributed when $H_{0,\lambda}$ is true and stochastically smaller than $\mathrm{Unif}(0, 1)$ otherwise.

Other applications of Condition on Selection Framework

1. **False discovery rate control in linear regression.** Selective inference with the Benjamini-Yekutieli procedure ensures FDR control in the $n > p$ regime. By combining with recent work on debiased estimators β^d, we can ensure FDR control in sparse high-dimensional linear regression.
2. **Scan statistics and approximate scan statistics.** We work out an exact test based off the scan statistic scan statistic

$$z^\star = \max_{C \in \mathcal{S}} \sum_{i \in C} z_i .$$

Frequently scan statistics are too expensive to compute, and greedy/approximate algorithms are used to approximate the scan statistic. We work out an exact test for these algorithms too, using the Condition on Selection framework.