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

Jason D. Lee

Stanford University

Collaborators: Yuekai Sun, Dennis Sun, Qiang Liu, and Jonathan Taylor.

Slides at

`http://web.stanford.edu/~jdl17/selective_inference_and_debiasing.pdf`

Selective Inference

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

Selective Inference is common (Yoav Benjamini's talk). 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.

Table of Contents

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 $\hat{M}$, $\hat{\beta}^{\hat{M}} = X^{\dagger}_{\hat{M}} y$.
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

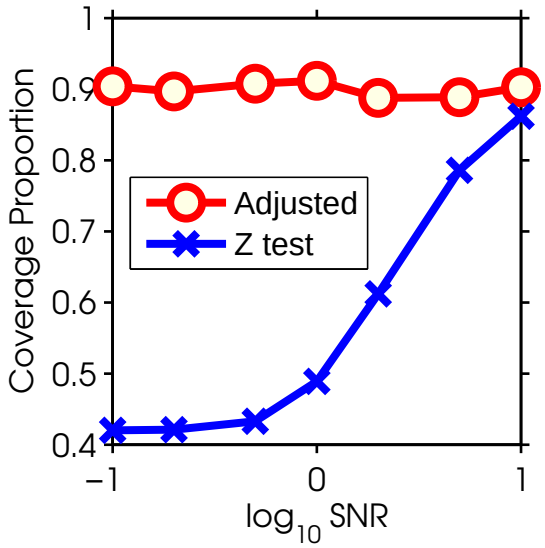


Figure: Plot of the coverage proportion across a range of SNR. The coverage proportion of the z intervals ($\hat{\beta} \pm 1.65\sigma$) is far below the nominal level of $1 - \alpha = .9$, even at SNR =5. The selective intervals (our method) always have coverage proportion $.9$.

Warning!!!!

Unadjusted confidence intervals are NOT selectively valid.

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

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

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

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

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 and PoSI).
- 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.

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 $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbf{R}^{n \times p}$.

Related Work

- 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. Taylor et al. 2014 show the significance test result can be recovered from the selective inference framework, and Taylor et al. 2014 generalize to testing global null for (almost) any regularizer.
- POSI (Berk et al. 2013) widen intervals to simultaneously cover all coefficients of all possible submodels.
- Asymptotic normality by debiasing (Zhang and Zhang 2012, Van de Geer et al. 2013, Javanmard and Montanari 2013, Chernozhukov et al. 2013).
- Oracle property and non-convex regularizers (Loh 2014). Under a beta-min condition, the solution to non-convex problem has a Gaussian distribution.
- Knockoff for FDR control in linear regression (Foygel and Candes 2014) allows for exact FDR control for $n \geq p$.

Lasso selection event

Lasso Selection Event

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

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

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

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.

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

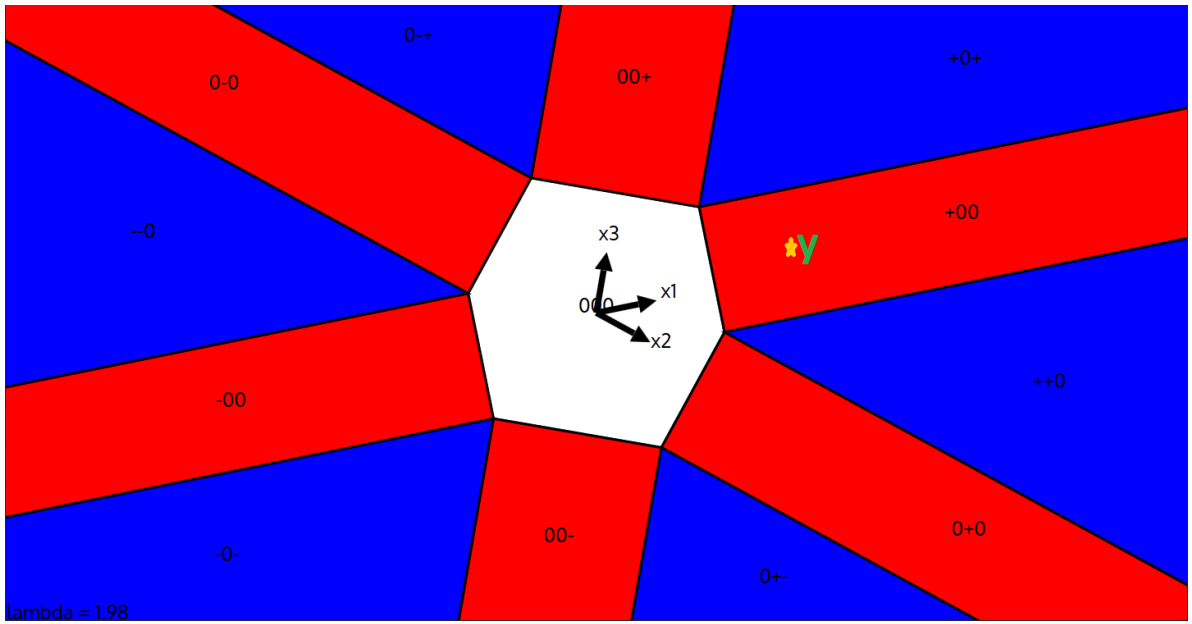


Figure: $(n, p) = (2, 3)$. White, red, and blue shaded regions correspond to different selection events. The shaded region that y falls into is where lasso selects variable 1 with positive sign.http://naftaliharris.com/blog/lasso-polytope-geometry/

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 and sensitive to value of nuisance parameters.
- For testing regression coefficients, we only need distribution of $\eta^T y | \{Ay \leq b\}$.

Computationally Tractable Inference

It turns out

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

Using this distributional result, we avoid sampling and integration.

Geometric Intuition

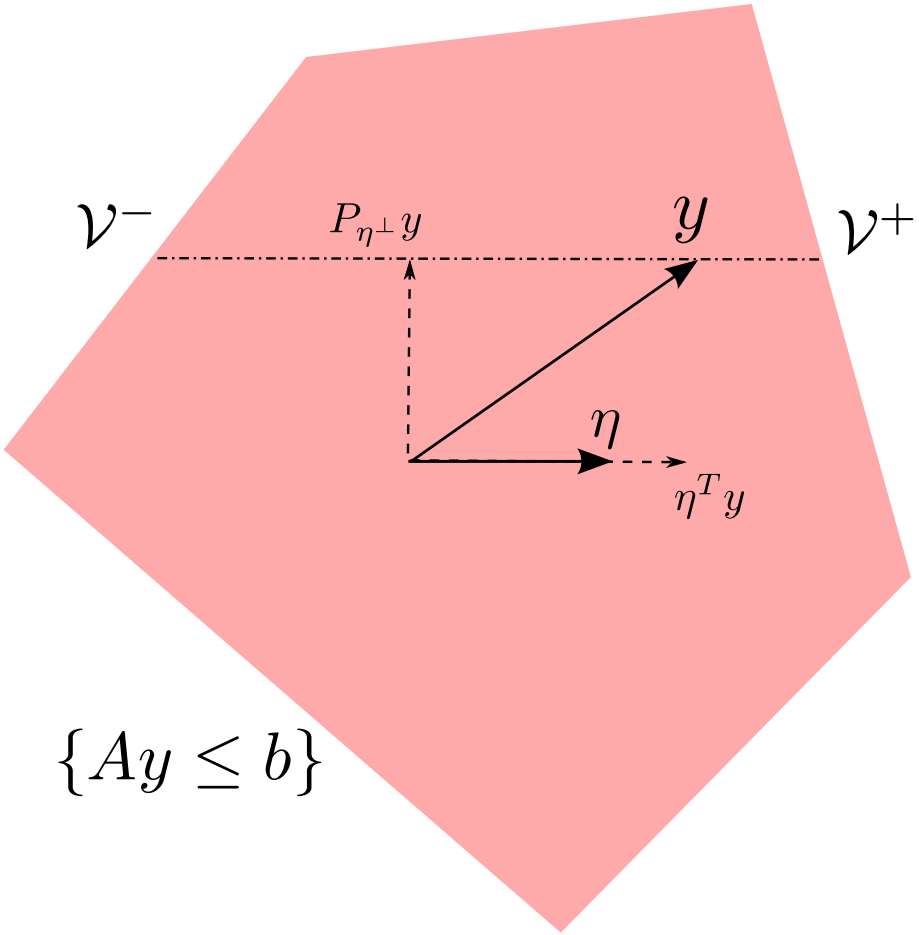


Figure: A picture demonstrating that the set $\{Ay \leq b\}$ can be characterized by $\{\mathcal{V}^- \leq \eta^T y \leq \mathcal{V}^+\}$. Assuming $\Sigma = I$ and $||\eta||_2 = 1$, $\mathcal{V}^-$ and $\mathcal{V}^+$ are functions of $P_{\eta^\perp} y$ only, which is independent of $\eta^T y$.

Testing regression coefficients

Theorem

Let $H_0 : \eta(\hat{M}(y))^T \mu = \gamma$. *The test that rejects if*

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

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.*

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 $\eta_j^T y = \hat{\beta}_j^{\hat{M}}$.

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.

Addressing some concerns.

THE Question we get

Why should I care about covering $\beta^{\hat{M}}$:

$$\Pr(\beta^{\hat{M}} \in C) = 1 - \alpha?$$

The statement that variable $j \in \hat{M}$ is significant means that variable j is significant after adjusting for the other variables in $\hat{M}$.

What some people would like

Instead, we could ask for

$$\Pr(\beta^0_{\hat{M}} \in C) = 1 - \alpha.$$

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, $\beta^0 = (X^T X)^{-1} X^T \mu = X^\dagger \mu$ 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.

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 $\sqrt{n} \left\| \beta^d(\hat{M}) - \beta^0 \right\|_\infty = o_P(1)$, and $\beta^d = B\mu + h$. 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

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

Selective confidence interval for β^0

$\beta^0_{j,\hat{M}} = \eta(\hat{M})^T \mu$ where η comes from the least squares estimator, so Condition on Selection framework allows you to construct C_j such that

$$\Pr(\beta^0_{j,\hat{M}} \in C_j) = 1 - \alpha.$$

The constructed confidence interval covers β^0 like the standard z-interval. The only difference is we make intervals for selected coordinates $\beta^0_{\hat{M}}$.

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 $\delta > 0$,

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

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

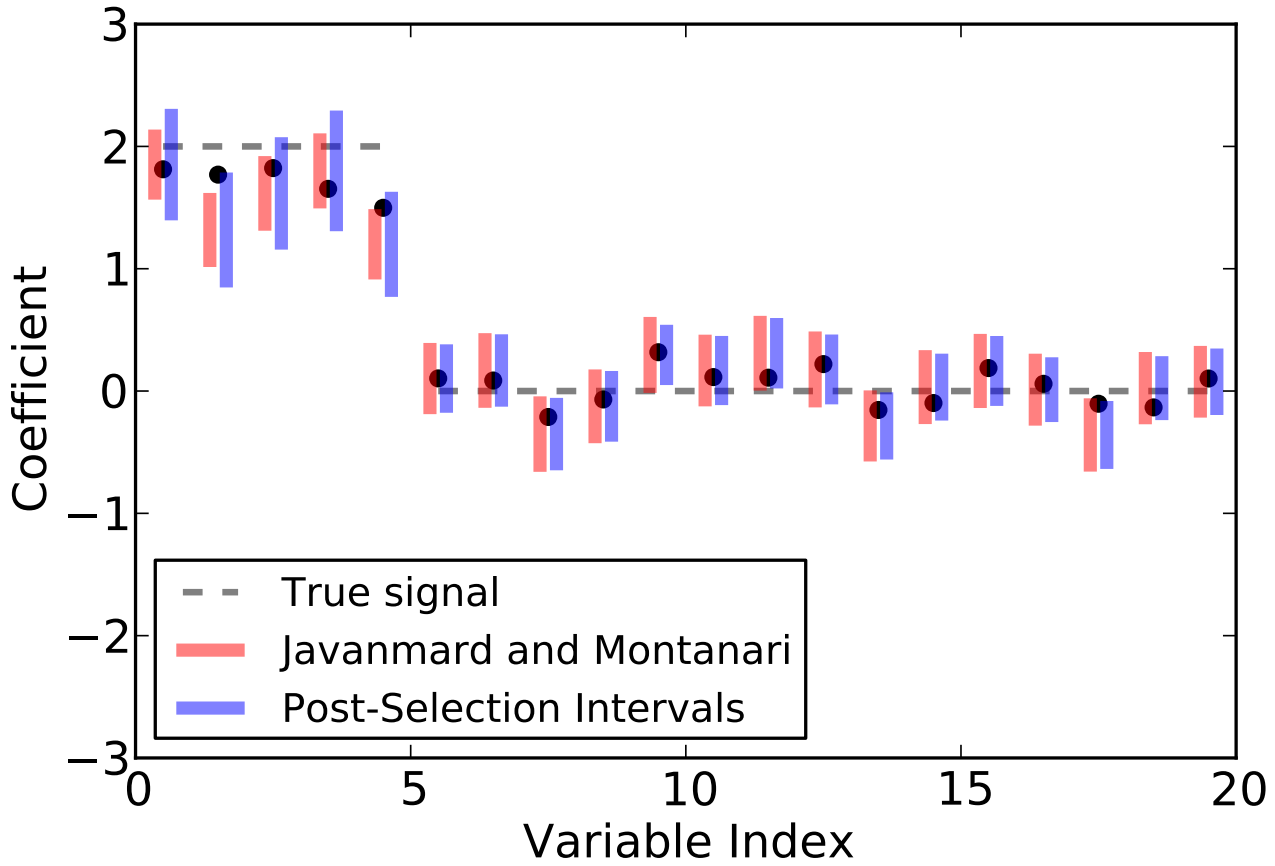


Figure: $(n, p, s) = (25, 50, 5)$ with only the first 20 coefficients being plotted. Data is generated from $y = X\beta^0 + \epsilon$ with a SNR of 2. Selective intervals (blue) control selective type 1 error, and the z-intervals (red) do not.

Follow-up work aka Roadmap to Jonathan Taylor's recent work

- Testing $H_0 : (I - P_{\hat{M}})\mu = 0$ (Lee et al. 2013)
- 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, forward stepwise regression, isotonic regression, elastic net, AIC/BIC criterion with subset selection, λ chosen via hold-out set, square-root lasso, unknown σ^2, non-Gaussian noise, and PCA (Lee & Taylor 2014, Tian and Taylor 2015, Reid et al. 2014, Choi et al. 2014, Loftus, Tian, and Taylor 2015+, Taylor et al. 2014, Loftus & Taylor 2014).
- 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 screening assumption (Fithian, Sun, Taylor 2014).

Improving Power

Intuition: Condition on less.

- **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 sampling 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.**

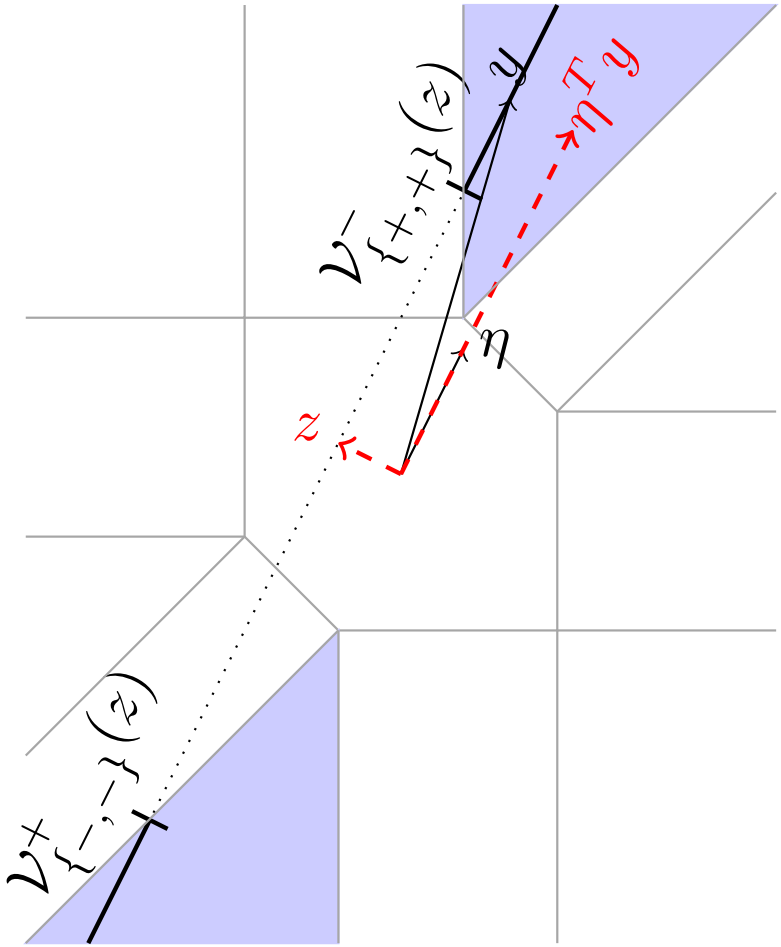


Figure: When we take the union over signs, the conditional distribution of $\eta^T y$ is truncated to a union of disjoint intervals. In this case, the Gaussian is truncated to the set $(-\infty, \mathcal{V}^+_{\{-,-\}}] \cup [\mathcal{V}^-_{\{+,+\}}, \infty)$.

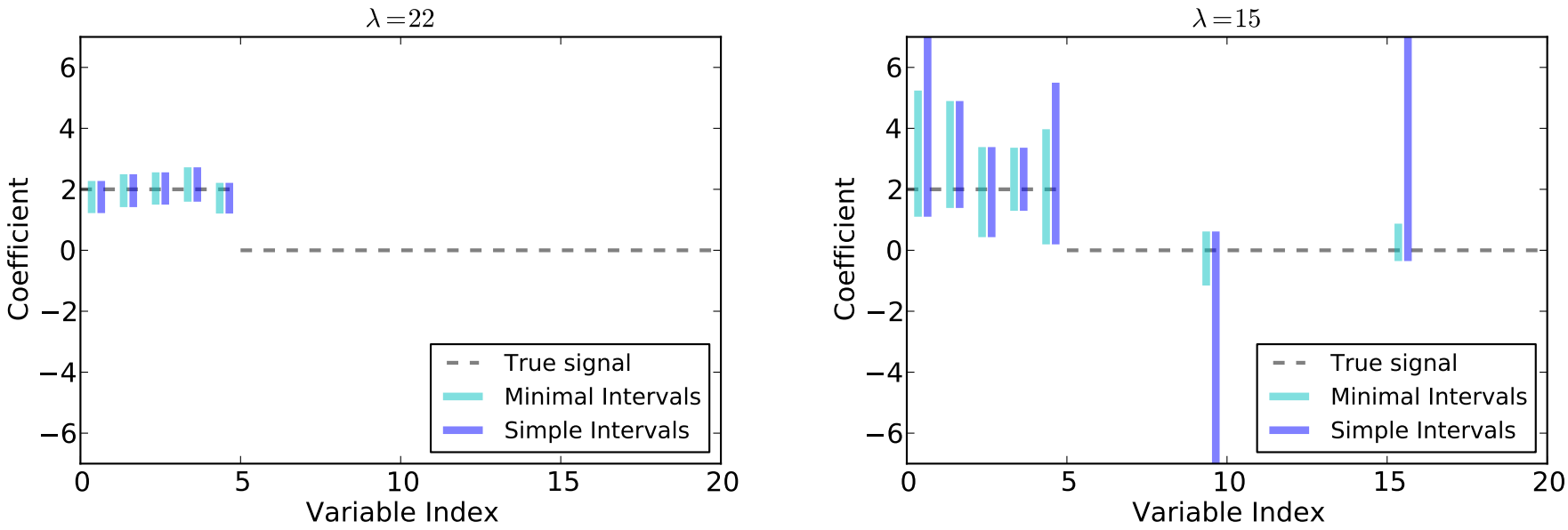
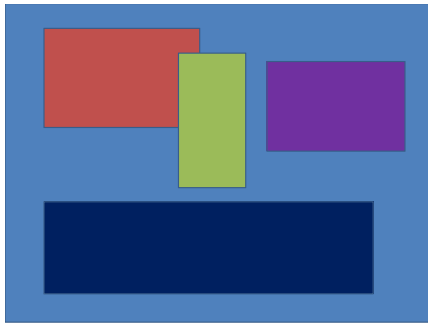


Figure: Light blue intervals are using the coarsest selection event or union of regions and dark blue are using the selection event that is one region. The simulated data featured $n = 25$, $p = 50$, and 5 true non-zero coefficients; only the first 20 coefficients are shown. The simple intervals are 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.

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.
- 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.**

The distributed setting

Given data $\{(x_i, y_i)\}_{i \in [N]}$ split (evenly) among m machines:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix}, X_k \in \mathbf{R}^{n \times p}; \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, y_k \in \mathbf{R}^n.$$

Notation:

- m machines
- subscript $k \in [m]$ indexes machines, denotes local quantities
- N total samples; $n = \frac{N}{m}$ samples per machine

The costs of computing in distributed settings

- *floating point operations*
- *bandwidth costs:* $\propto$ total bits transferred
- *latency costs:* $\propto$ rounds of communication

$$\textbf{FLOPS}^{-1} \ll \textbf{bandwidth}^{-1} \ll \textbf{latency}$$

The lasso in distributed settings

The lasso: $\hat{\beta} := \arg\min_{\beta \in \mathbf{R}^p} \frac{1}{2N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$

- *iterative:* optimization generally requires iteration
- *communication intensive:* evaluating the gradient (of the loss) at each iteration requires $O(mp)$ communication
  1. Each node forms its local gradient: $\frac{1}{n} X_k^T (y_k - X_k \beta)$.
  2. The master node averages the local gradients:

$$\frac{1}{N} X^T (y - X\beta) = \frac{1}{m} \sum_{k=1}^m X_k^T (y_k - X_k \beta).$$

Q: Distributed sparse regression with 1 round of communication?

Averaging debiased lasso's

1. Each node computes a local "debiased" lasso estimator:

$$\hat{\beta}_k \leftarrow \arg\min_{\beta \in \mathbf{R}^p} \frac{1}{2n} \|y_k - X_k \beta\|_2^2 + \lambda_k \|\beta\|_1$$
$$\hat{\beta}_k^d \leftarrow \hat{\beta}_k + \frac{1}{n} \hat{\Theta}_k X_k^T (y_k - X_k \hat{\beta}_k), \quad \hat{\Theta}_k \left(\frac{1}{n} X_k^T X_k \right) \approx I_p.$$

2. The master node averages the debiased local lasso estimators:

$$\hat{\beta}^d \leftarrow \frac{1}{m} \sum_{k=1}^m \hat{\beta}_k^d.$$

3. HT the averaged estimator to obtain a sparse estimator $\hat{\beta}^{d,ht}$.

Thresholding averaged debiased lasso's

Theorem (Lee, Liu, Sun, Taylor 2015+)

When $\|\hat{\beta}_k - \beta_0\|_1 \sim s\sqrt{\frac{\log p}{n}}, |||\hat{\Theta}_k\hat{\Sigma}_k - I_p|||_\infty \leq \sqrt{\frac{\log p}{n}}, k \in [m],$

- *For* $m \lesssim \sqrt{\frac{N}{s^2 \log p}}$: $\|\hat{\beta}^{d,ht} - \beta_0\|_\infty \sim \sqrt{\frac{\log p}{nm}}$,
- *For* $m \lesssim \sqrt{\frac{N}{s \log p}}$: $\|\hat{\beta}^{d,ht} - \beta_0\|_2 \sim \sqrt{\frac{s \log p}{nm}}$,
- *For* $m \lesssim \sqrt{\frac{N}{s \log p}}$: $\|\hat{\beta}^{d,ht} - \beta_0\|_1 \sim s\sqrt{\frac{\log p}{nm}}$.

Communication-efficiency (Tengyu Ma)

This algorithm is communication-efficient optimal. Any algorithm that achieves ℓ_2 estimation error of $\sqrt{\frac{s \log \frac{p}{s}}{nm}}$ needs $O(pm)$ communication.

Computation

The bottleneck is computation of Θ, which requires p lasso's. Can we get something similar to the debiased estimator that is of lower computational cost?
Computing Θ is very wasteful since we only need $\frac{1}{n}\Theta X^T(y - X\hat{\beta})$ to form the debiased estimator.

Averaging debiased lasso's

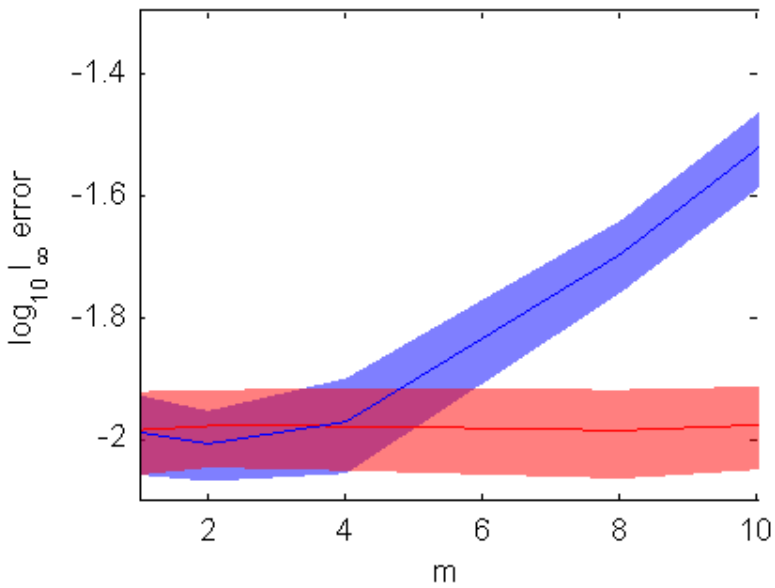


Figure: $(N, p, s) = (10000, 4000, 20)$. Red curve is lasso on all the data. Blue curve is our proposed estimator. The blue line achieves the same estimation error as the red line until $m \geq 5$.

Acknowledgments

Selective Inference:

1. Jason D. Lee, Dennis L Sun, Yuekai Sun, and Jonathan Taylor, *Exact post-selection inference with the Lasso. (Version 4 on arXiv is the most complete)*
2. Jason D. Lee and Jonathan Taylor, *Exact statistical inference after marginal screening.*
3. Jason D. Lee, Yuekai Sun, Jonathan Taylor, *Evaluating the statistical significance of submatrices.*

Communication-efficient regression

1. Jason D. Lee, Yuekai Sun, Qiang Liu, and Jonathan Taylor, Communication-efficient sparse regression, Forthcoming.

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

Thanks for Listening!