# Exact Statistical Inference after Model Selection.

Jason D Lee

Dept of Statistics and Institute of Computational and Mathematical Engineering, Stanford University

Joint work with Jonathan Taylor, Dennis Sun, and Yuekai Sun.
February 2014

# Motivation: Linear regression in high dimensions

1. Select relevant variables $\hat{S}$ via a variable selection procedure ($k$ most correlated, lasso, OMP ...).
2. Fit a linear regression model using only the variables in $\hat{S}$.
3. Return the selected set of coefficients $\hat{S}$ and the coefficients $\hat{\beta}_{\hat{S}}$.
4. Construct confidence intervals $95\%$ confidence intervals $(\hat{\beta}_j - 1.96\sigma_j, \hat{\beta}_j + 1.96\sigma_j)$.
5. Test the hypothesis $H_0 : \beta_j = 0$ by rejecting when $\left|\frac{\beta_j}{\sigma_j}\right| \geq 1.96$.

Are these confidence intervals and hypothesis tests 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 $y = X\beta^0 + \epsilon$.
- $\epsilon \sim N(0, 1)$.
- $\beta^0$ 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 $\beta$ and check the coverage proportion.

# Simulation

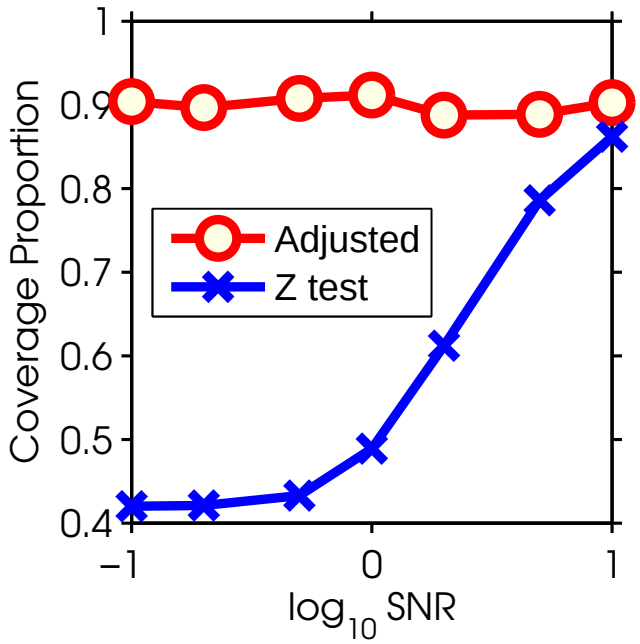



**Figure:** Plot of the coverage proportion across a range of SNR. The coverage proportion of the z intervals is far below the nominal level of $1 - \alpha = .9$, even at SNR =5. The adjusted intervals (our method) always have coverage proportion $.9$.

# 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 predictor ($f(x) = \beta^T x$) is $\beta^\star = X^\dagger \mu$. Linear regression estimates this using

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

**Theorem**

*The least squares estimator is distributed*

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

*and*

$$\Pr\left(\beta_j^\star \in \left(\hat{\beta}_j - z\sigma(X^T X)_{jj}^{-1/2}, \hat{\beta}_j + z\sigma(X^T X)_{jj}^{-1/2}\right)\right) = 1 - \alpha.$$

# Explaining the simulation

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

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

3. For any fixed set $S$, $X_S^T y$ is Gaussian, but $X_{\hat{S}}^T y$ 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.

# Condition on selection framework

This talk is about a framework for post-selection inference, i.e. the selection procedure is adaptive to the data. The main idea is

**condition on selection**

1. Represent the selection event as a set of affine constraints on $y$.
2. Derive the conditional distribution and pivotal quantity for linear contrasts $\eta^T y$.
3. Invert the pivotal quantity to obtain confidence intervals for $\eta^T \mu$.

# Related Work

- POSI (Berk et al. 2013) widen intervals to simultaneously cover all coefficients of *all* possible submodels. The method is extremely conservative and is only computationally feasible for $p \leq 30$.
- Asymptotic normality by "inverting" KKT conditions (Zhang 2012, Buhlmann 2012, Van de Geer 2013, Javanmard 2013). Asymptotic result that requires consistency of the lasso.
- 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.

# Preview of our results

- The results are exact (non-asymptotic). Only assume $X$ is in general position, and no assumptions on $n$ and $p$ (e.g. $n > s \log p$).
- We assume that $\epsilon$ is Gaussian and $\sigma^2$ is known.
- The constructed confidence intervals satisfy

$$\mathbf{Pr}\left(\beta^{\star}_{j\in\hat{S}} \in [L^j_\alpha, U^j_\alpha]\right) = 1-\alpha,$$

  where $\beta^{\star}_{j\in\hat{S}} = X^{\dagger}_{\hat{S}}\mu$.
- Test for whether the lasso/marginal screening have found all relevant variables.
- Framework is applicable to many model selection procedures including marginal screening, lasso, OMP, and non-negative least squares.

# Marginal screening

**Algorithm 1** Marginal screening algorithm

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

# Marginal screening selection event

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

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

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

# Lasso selection event

## Lasso

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

KKT conditions provide us with the selection event. A set of variables $\hat{S}$ is selected with $\text{sign}(\hat{\beta}_{\hat{S}}) = \hat{s}$ if

$$\left\{ y : \text{sign}(U(\hat{S}, \hat{s})) = z_{\mathcal{E}}, \left\| W(\hat{S}, \hat{s}) \right\|_{\infty} < 1 \right\} = \{ y : A(\hat{S}, \hat{s}) y \leq b(\hat{S}, \hat{s}) \}$$

where

$$U(S, s) := (X_S^T X_S)^{-1} (X_S^T y - \lambda z_S)$$
$$W(S, s) := X_{-S}^T (X_S^T)^{\dagger} z_S + \frac{1}{\lambda} X_{-S}^T (I - P_S) y.$$

# Partition via the selection event

## Partition decomposition

We can decompose $y$ in terms of partition, where $y$ is a different constrained Gaussian for each element of the partition.

$$y = \sum_{S,s} y \; \mathbb{1}\left(A(S,s)y \leq b(S,s)\right)$$

## Theorem

*The distribution of $y$ conditional on the selection event is a constrained Gaussian,*

$y|\{(\hat{S},\hat{s}) = (S,s)\} \stackrel{d}{=}$ *Gaussian constrained to* $\{x : A(\hat{S},\hat{s})x \leq b\}$.

# Constrained Gaussian

- The distribution of $y \sim \mathcal{N}(\mu, \sigma^2 I)$ conditional on $\{y : Ay \leq b\}$ has density $\frac{1}{\mathbf{Pr}(Ay \leq b)} \phi(y; \mu, \Sigma) \mathbb{1}(Ay \leq b)$.
- Although we understand the distribution of $y$ condition on selection is a constrained Gaussian, the normalization constant is computationally intractable.
- We would like to understand the distribution of $\eta^T y$, since regression coefficients are linear contrasts, $\hat{\beta}_{j \in \hat{S}} = e_j^T X_{\hat{S}}^{\dagger} y$.
- Instead, we show $\eta^T y$ is a (univariate) truncated normal.

## Lemma

*The conditioning set can be rewritten in terms of $\eta^T y$ as follows:*

$$\{Ay \leq b\} = \{\mathcal{V}^-(y) \leq \eta^T y \leq \mathcal{V}^+(y), \mathcal{V}^0(y) \geq 0\}$$

*where* $\alpha = \frac{A\Sigma\eta}{\eta^T\Sigma\eta}$, $\mathcal{V}^0 = \mathcal{V}^0(y) = \min_{j:\ \alpha_j=0} b_j - (Ay)_j$,

$$\mathcal{V}^- = \mathcal{V}^-(y) = \max_{j:\ \alpha_j<0} \frac{b_j - (Ay)_j + \alpha_j\eta^T y}{\alpha_j}$$

$$\mathcal{V}^+ = \mathcal{V}^+(y) = \min_{j:\ \alpha_j>0} \frac{b_j - (Ay)_j + \alpha_j\eta^T y}{\alpha_j}.$$

*Moreover, $(\mathcal{V}^+, \mathcal{V}^-, \mathcal{V}^0)$ are independent of $\eta^T y$.*

# Geometric Intuition

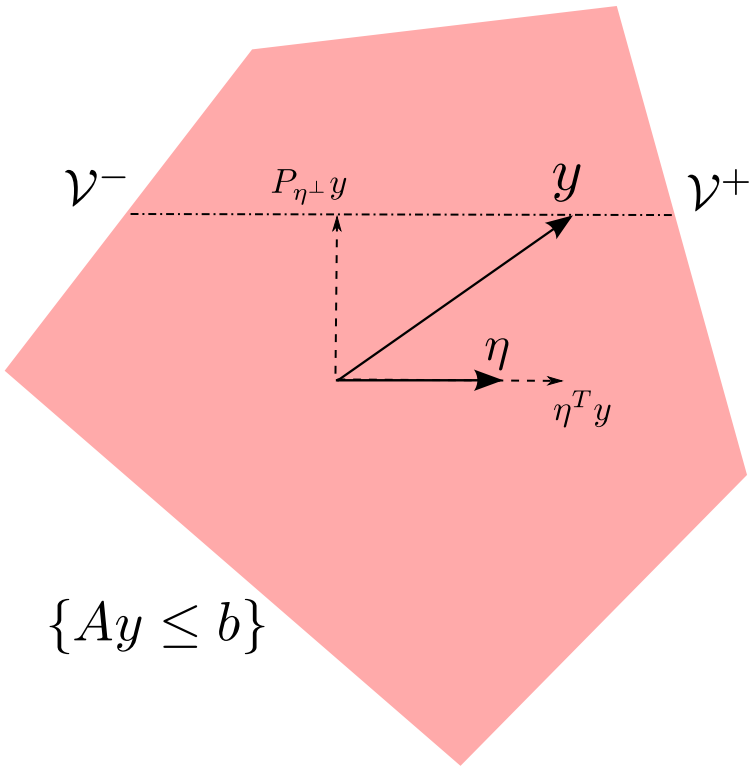



**Figure:** A picture demonstrating that the set $\{Ay \leq b\}$ can be characterized by $\{\mathcal{V}^- \leq \eta^T y \leq \mathcal{V}^+\}$. Assuming $\Sigma = I$ and $||\eta||_2 = 1$, $\mathcal{V}^-$ and $\mathcal{V}^+$ are functions of $P_{\eta^\perp} y$ only, which is independent of $\eta^T y$.

# Truncated Normal

## Corollary

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

$$\eta^T y \mid \{Ay \leq b, \mathcal{V}^+(y) = v^+, \mathcal{V}^-(y) = v^-\} \sim TN(\eta^T \mu, \eta^T \Sigma \eta, v^-, v^+)$$

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

# Pivotal quantity

**Theorem**

*Let* $\Phi(x)$ *denote the CDF of a* $N(0,1)$ *random variable, and let* $F(x;\mu,\sigma^2,a,b)$ *denote the CDF of* $TN(\mu,\sigma,a,b)$

$$F(x;\mu,\sigma^2,a,b) = \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}.$$

*Then* $F(\eta^T y;\eta^T\mu,\eta^T\Sigma\eta,\mathcal{V}^-(y),\mathcal{V}^+(y))$ *is a pivotal quantity*

$$F(\eta^T y;\eta^T\mu,\eta^T\Sigma\eta,\mathcal{V}^-(y),\mathcal{V}^+(y)) \sim \mathit{Unif}(0,1)$$

# Hypothesis testing

**Testing contrasts $\eta^T \mu$.**

The pivotal quantity allows us to test $H_0 : \eta^T \mu = \gamma_0$. Under $H_0$,

$$F(\eta^T y; \gamma_0, \eta^T \Sigma \eta, \mathcal{V}^-(y), \mathcal{V}^+(y)) \sim \text{Unif}(0,1)$$

The test that rejects if $F(\eta^T y; \gamma_0, \eta^T \Sigma \eta, \mathcal{V}^-, \mathcal{V}^+) > 1 - \alpha$ is an $\alpha$-level test of $H_0$.

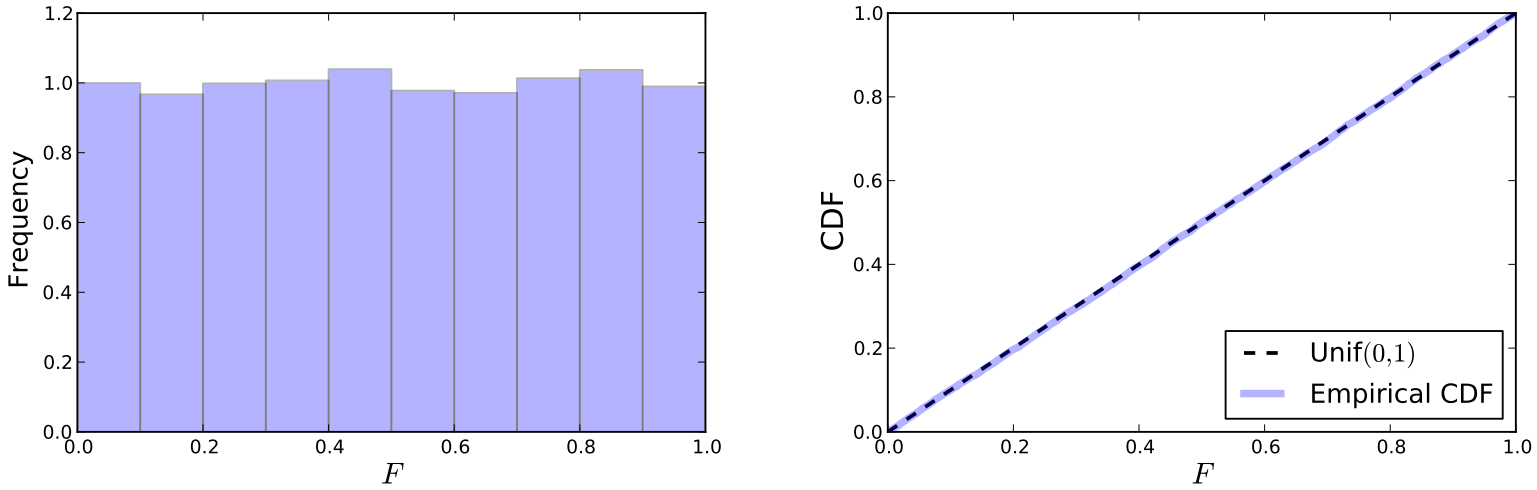



**Figure:** Histogram and empirical distribution of $F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y)$ obtained by sampling $y \sim N(\mu, \Sigma)$ constrained to $\{Ay \leq b\}$. The distribution is very close to Unif$(0, 1)$.

# Testing regression coefficients

Recall, $\beta^\star_{\hat{S}} = X^\dagger_{\hat{S}}\mu$, and $\hat{\beta}_{\hat{S}} = X^\dagger_{\hat{S}}y$.

By choosing $\eta_j = X^{\dagger T}_{\hat{S}} e_j$, we have $\eta_j^T y = \hat{\beta}_{j\in\hat{S}}$, which is the regression coefficient with respect to design $X_{\hat{S}}$.

**Theorem**

*Let* $H_0 : \beta^\star_{j\in\hat{S}} = \beta_j$. *The test that rejects if*
$F(\hat{\beta}_{j\in\hat{S}}; \beta_j, \eta_j^T\Sigma\eta_j, \mathcal{V}^-, \mathcal{V}^+) > 1 - \frac{\alpha}{2}$ *or*
$F(\hat{\beta}_{j\in\hat{S}}; \beta_j, \eta_j^T\Sigma\eta_j, \mathcal{V}^-, \mathcal{V}^+) < \frac{\alpha}{2}$ *is an* $\alpha$*-level test of* $H_0$.

**Algorithm 2** Hypothesis test for selected variables

1: **Input:** Design matrix $X$, response $y$, model size $k$.
2: Use variable selection method (marginal screening or Lasso) to select a subset of variables $\hat{S}$.
3: Specify the null hypothesis $H_0 : \beta^{\star}_{j\in\hat{S}} = \beta_j$.
4: Let $A = A(\hat{S}, \hat{s})$ and $b = b(\hat{S}, \hat{s})$. Let $\eta_j = (X^T_{\hat{S}})^{\dagger} e_j$.
5: Compute $F(\hat{\beta}_{j\in\hat{S}}; \beta_j,\ \sigma^2||\eta_j||^2, \mathcal{V}^-, \mathcal{V}^+)$, where $\mathcal{V}^-$ and $\mathcal{V}^+$ are computed via the $A$, $b$, and $\eta$ previously defined.
6: **Output:** Reject if $F(\hat{\beta}_{j\in\hat{S}}; \beta_j,\ \sigma^2||\eta_j||^2, \mathcal{V}^-, \mathcal{V}^+) < \frac{\alpha}{2}$ or $F(\hat{\beta}_{j\in\hat{S}}; \beta_j,\ \sigma^2||\eta_j||^2, \mathcal{V}^-, \mathcal{V}^+) > 1 - \frac{\alpha}{2}$.

# Confidence Intervals

## Confidence interval

Confidence interval $C(j, y)$ is all $\beta_j$'s, where a test of $H_0 : \beta^{\star}_{j \in \hat{S}} = \beta_j$ fails to reject at level $\alpha$.

$$C(j, y) = \{\beta_j : \frac{\alpha}{2} \leq F(\hat{\beta}_{j \in \hat{S}}; \beta_j,\ \sigma^2 ||\eta_j||^2, \mathcal{V}^-, \mathcal{V}^+) \leq 1 - \frac{\alpha}{2}\}$$

Interval $[L^j, U^j]$ is found by solving $F(\hat{\beta}_{j \in \hat{S}}; L^j,\ \sigma^2 ||\eta_j||^2, \mathcal{V}^-, \mathcal{V}^+) = 1 - \frac{\alpha}{2}$. and $F(\hat{\beta}_{j \in \hat{S}}; U^j,\ \sigma^2 ||\eta_j||^2, \mathcal{V}^-, \mathcal{V}^+) = \frac{\alpha}{2}$.

**Algorithm 3** Confidence intervals for selected variables

1: **Input:** Design matrix $X$, response $y$, model size $k$.
2: Use variable selection method to select a subset of variables $\hat{S}$.
3: Let $A = A(\hat{S}, \hat{s})$ and $b = b(\hat{S}, \hat{s})$. Let $\eta_j = (X_{\hat{S}}^T)^\dagger e_j$.
4: Solve for $L^j$ and $U^j$ where $\mathcal{V}^-$ and $\mathcal{V}^+$ are computed using the $A$, $b$, and $\eta_j$ previously defined.
5: **Output:** Return the intervals $[L^j, U^j]$ for $j \in \hat{S}$.

**Lemma**

*For each $j \in \hat{S}$,*

$$\mathbf{Pr}\left(\beta^\star_{j\in\hat{S}} \in [L^j, U^j]\right) = 1 - \alpha.$$

Solve Lasso at some $\lambda$, and construct confidence intervals using previous algorithm.

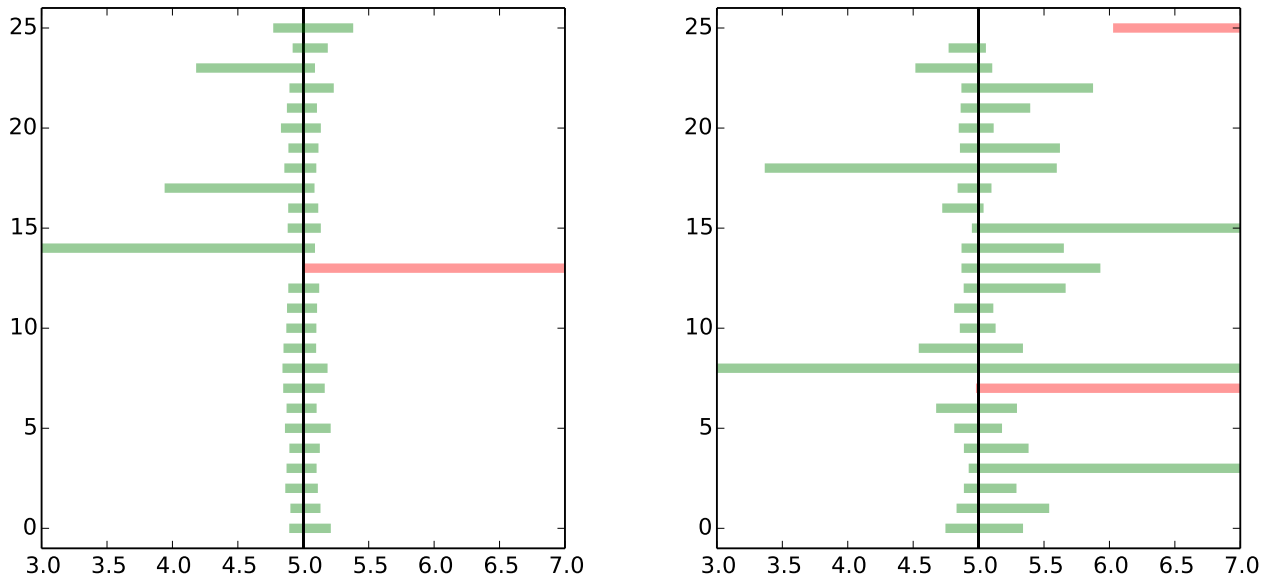
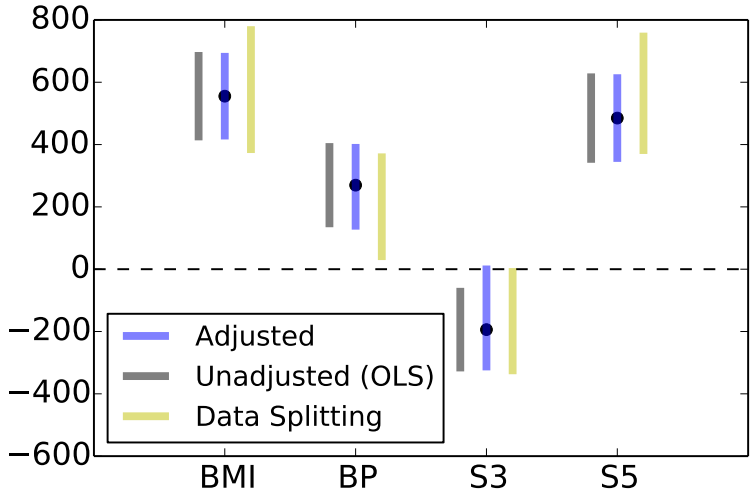



**Figure:** 90% confidence intervals for $\hat{\beta}^{\star}_1$ for two different settings $(n, p) = (100, 50)$ and $(n, p) = (100, 200)$, over 25 simulated data sets. The truth $\beta^0$ has five non-zero coefficients, all set to 5.0, and the noise variance is 0.25. A green bar means the confidence interval covers the true value while a red bar means otherwise.




- Blue line is our adjusted intervals, gray line is OLS intervals which ignore selection, and hellow line is the intervals computed using data splitting.
- Variable S3 is no longer significant after adjusting for model selection.
- Our adjusted intervals are approximately the same as the OLS intervals for significant variables. Data splitting widens the intervals by $\sqrt{2}$.

# Non-Gaussian noise and estimated $\sigma^2$

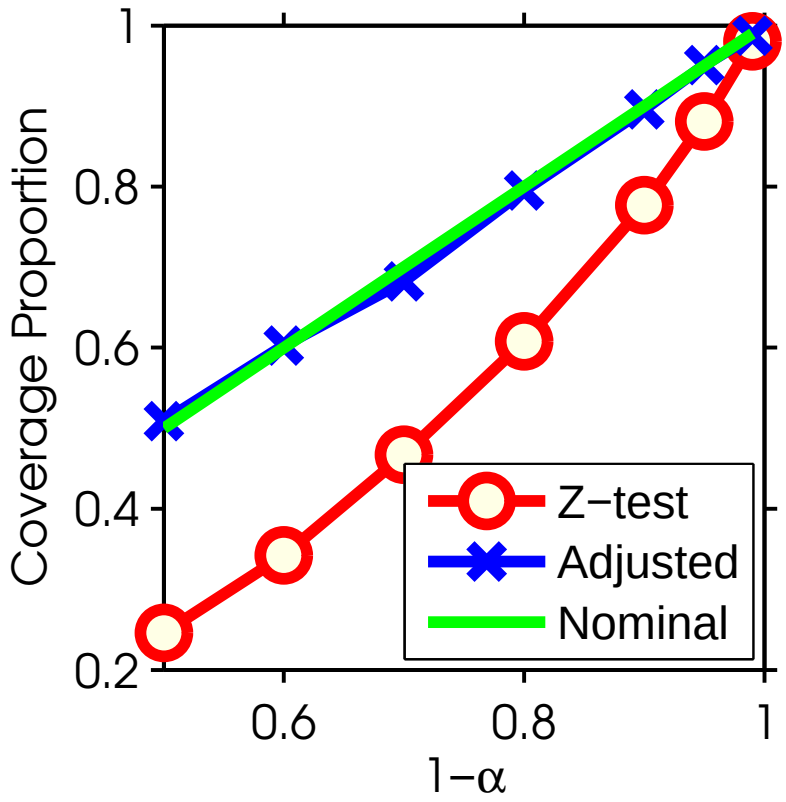



**Figure:** Plot of $1-\alpha$ vs the coverage proportion for diabetes dataset. Selection is Simulation is done by using $2000$ iterations of residual bootstrap. The adjusted intervals always cover at the nominal level, whereas the z-test is always below.

# Minimal post-selection inference

## Minimal selection event

Recall that each pair $(S, s)$ is in bijection with a selection event.
We only care about the selected variables $S$, not the signs $s$.
Selection event for only variables $S$:

$$\left\{y : \hat{S}(y) = S\right\} = \bigcup_{s \in \{-1,1\}^{|\hat{S}|}} \{y : (\hat{S}(y), s(y)) = (S, s)\}$$

$$= \bigcup_{s \in \{-1,1\}^{|\hat{S}|}} \{y : A(S, s)y \leq b(S, s)\}$$

- Condition on the coarsest partition where $\eta$ is still measurable.
- The set is a union of linear constraints. Pivotal quantity, hypothesis tests, and intervals are valid for union of linear constraints.
- Empirically results in shorter confidence intervals, at the cost more computation.

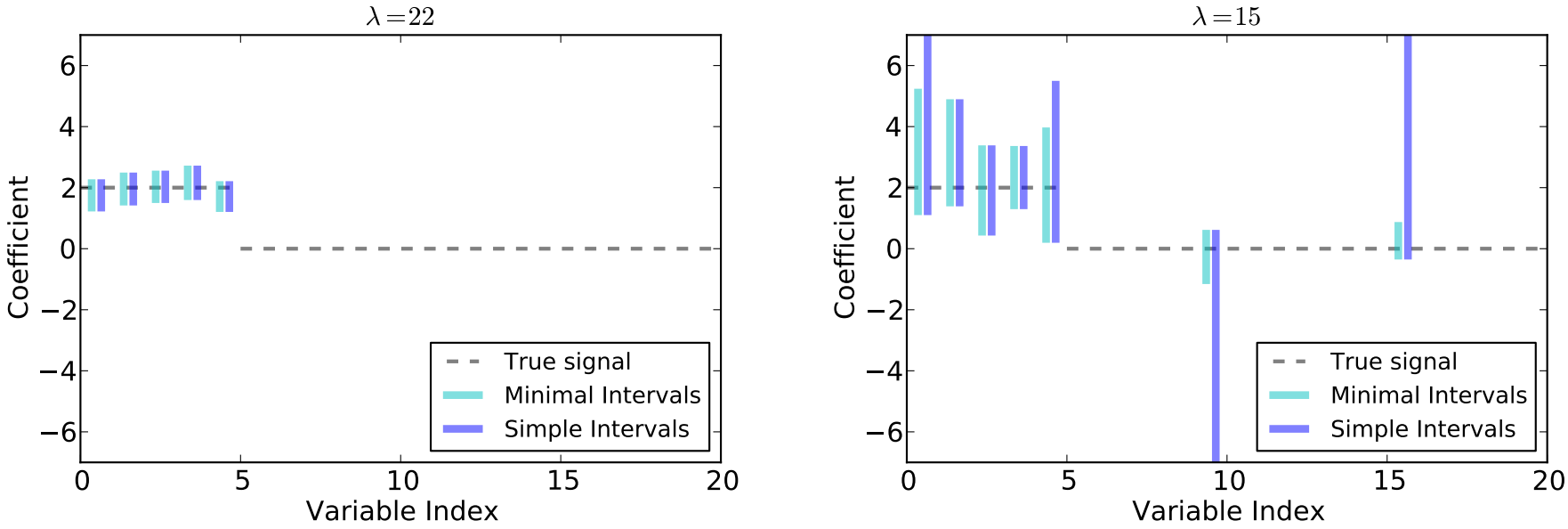



**Figure:** Comparison of the minimal and simple intervals as applied to the same simulated data set for two values of $\lambda$. The simulated data featured $n = 25$, $p = 50$, and 5 true non-zero coefficients; only the first 20 coefficients are shown. (We have included variables with no intervals to emphasize that inference is only on the selected variables.) We see that 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.

# More model selection procedures

Easily generalizes to other model selection procedures!

- Orthogonal matching pursuit/ forward stepwise regression.
- Screen+clean procedures such as marginal screening followed by Lasso.
- Constrained least squares (Non-negative least squares, isotonic regression).
- LARS (Taylor et al. 2014) and elastic net.
- Any polyhedral regularizer.

# Extensions

- Testing the goodness of fit of the selected model, $H_0 : (I - P_{\hat{S}})\mu = 0$.
- Non-Gaussian noise (Tian and Taylor 2014).
- Logistic regression, and conditional maximum likelihood.
- Pathwise algorithm for stopping Lasso that controls FWER.
- Estimating $\sigma^2$.

# Acknowledgments

Thanks to Trevor Hastie and other members of the Hastie, Tibshirani and Taylor group for feedback.

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.*

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

Thanks for Listening!

# 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 $s_j := \text{sign}(e_j^T(I - P_{\hat{S}})y)$, 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 contrast of $y$.

## Corollary

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

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

*is level* $\alpha$,

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

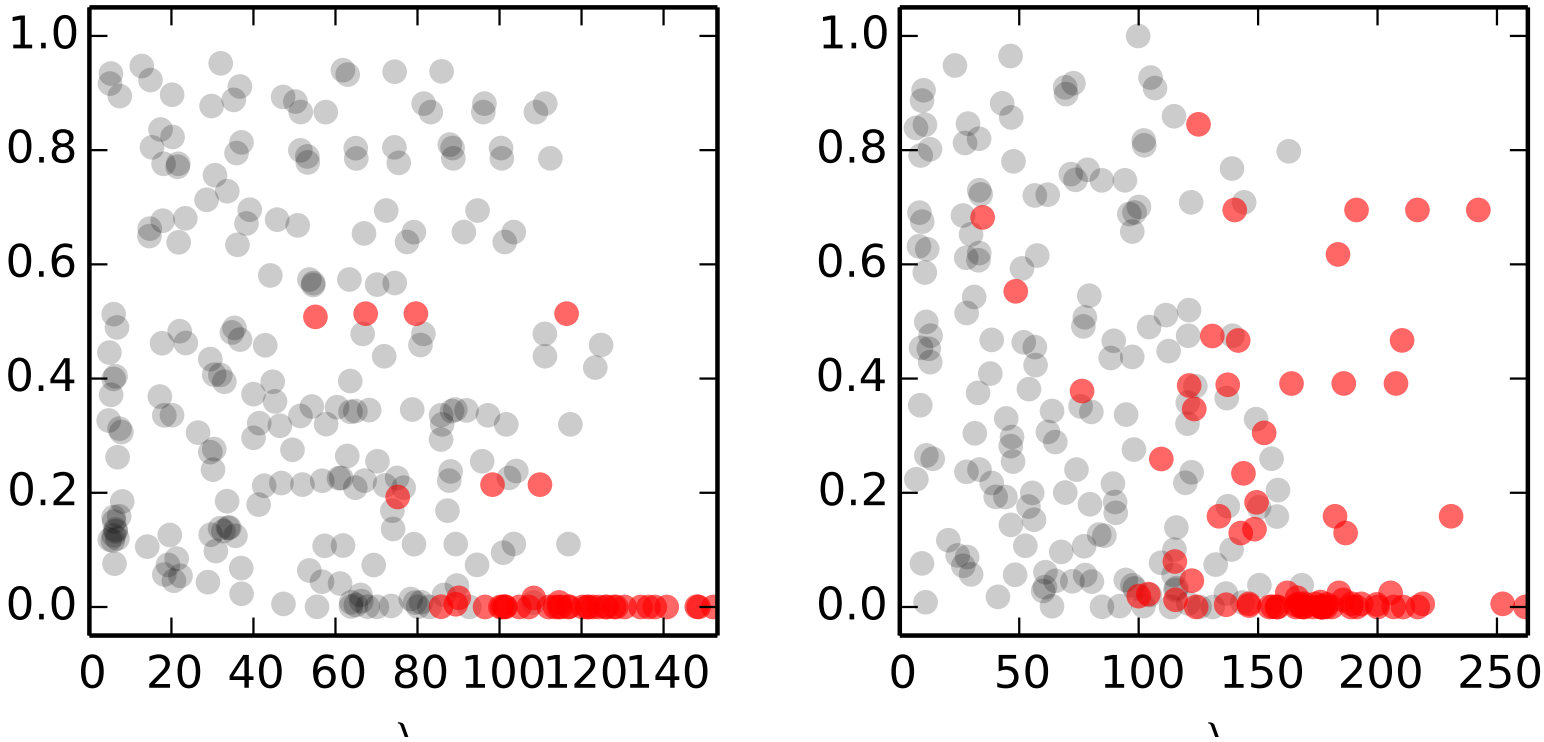



**Figure:** P-values for $H_{0,\lambda}$ at various $\lambda$ values for a small $(n = 100,\ p = 50)$ and a large $(n = 100,\ p = 200)$ uncorrelated Gaussian design, computed over 50 simulated data sets. The true model has three non-zero coefficients, all set to 1.0, and the noise variance is 2.0. We 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).

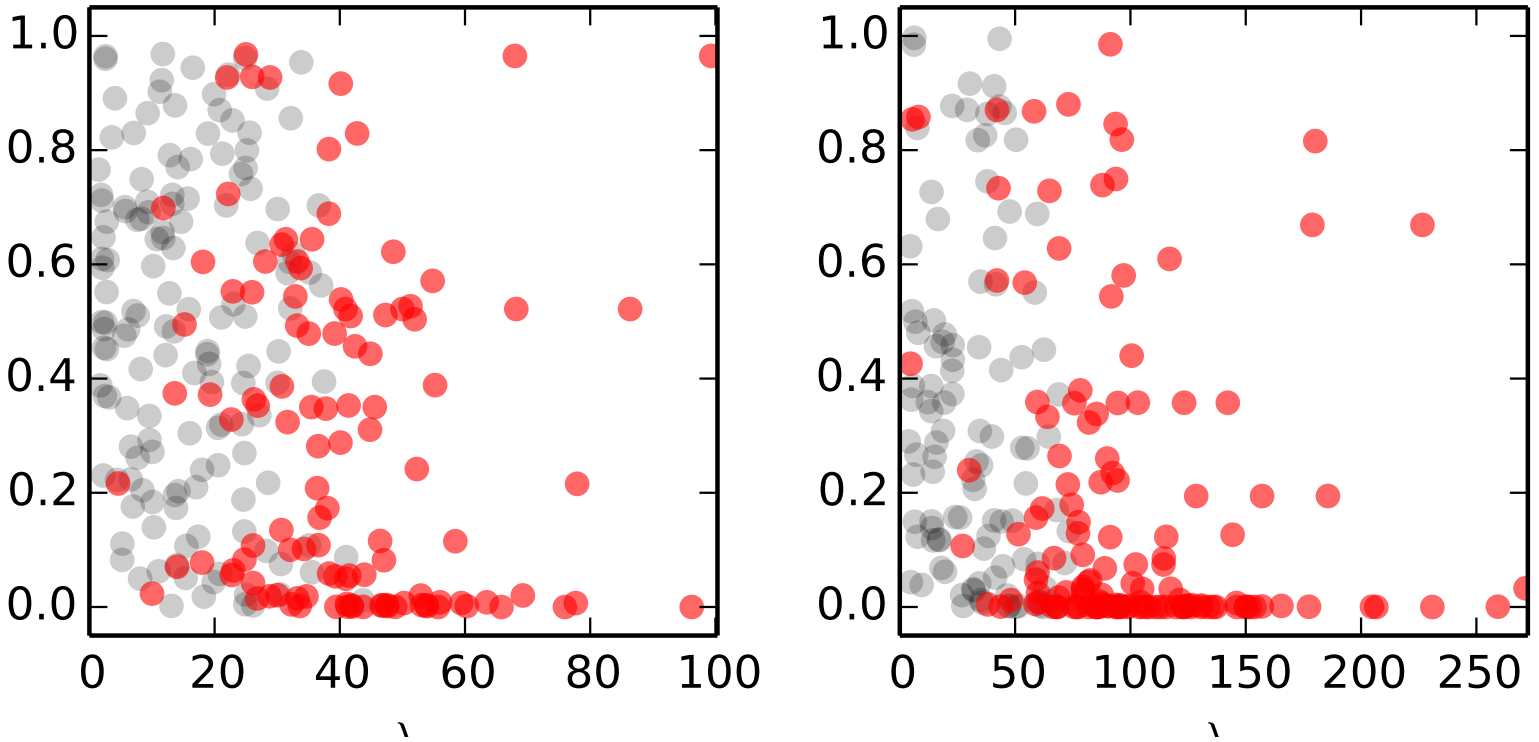



**Figure:** P-values for $H_{0,\lambda}$ at various $\lambda$ values for a small $(n = 100,\ p = 50)$ and a large $(n = 100,\ p = 200)$ *correlated* $(\rho = 0.7)$ Gaussian design, computed over 50 simulated data sets. The true model has three non-zero coefficients, all set to 1.0, and the noise variance 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.