Some new ideas for post selection inference and model assessment

Robert Tibshirani, Stanford

WHOA!! 2018

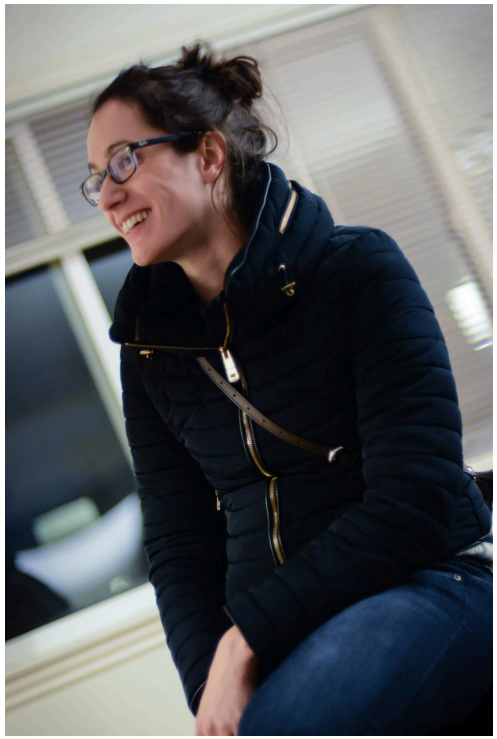
Thanks to Jon Taylor and Ryan Tibshirani for helpful feedback

Two topics

1. **How to improve post-selection inference for the lasso**:
   Keli Liu, Jelena Markovic & RT (with further generalizations by Jon Taylor)
2. **Maybe we're answering the wrong question in #1:**
   Post model-fitting exploration via "Next-Door" analysis– Leying Guan & RT

Keli Liu

Jelena Markovic



Leying Guan

Post-selection inference for the lasso

- Data $(x_i, y_i), i = 1, 2, \ldots N$; $x_i = (x_{i1}, x_{i2}, \ldots x_{ip})$. $\boldsymbol{X}$ fixed.
- Model

$$y_i = \beta_0 + \sum_j x_{ij}\beta_j + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2).$$

Post-selection inference for the lasso

- Data $(x_i, y_i), i = 1, 2, \ldots N$; $x_i = (x_{i1}, x_{i2}, \ldots x_{ip})$. $\boldsymbol{X}$ fixed.
- Model

$$y_i = \beta_0 + \sum_j x_{ij}\beta_j + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2).$$

- The **Lasso**

$$\arg \min_{\beta_0, \beta_1, \ldots, \beta_p} \left\{ \sum_i (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda \cdot \sum_j |\beta_j| \right\}$$

for some $\lambda \geq 0$.

Review of truncated Gaussian approach

Polyhedral selection events

- Response vector $y \sim N(\mu, \Sigma)$. Suppose we make a selection that can be written as

$$\{y : Ay \leq b\}$$

with A, b not depending on y. This is true for **forward stepwise regression, lasso with fixed λ, least angle regression** and other procedures.

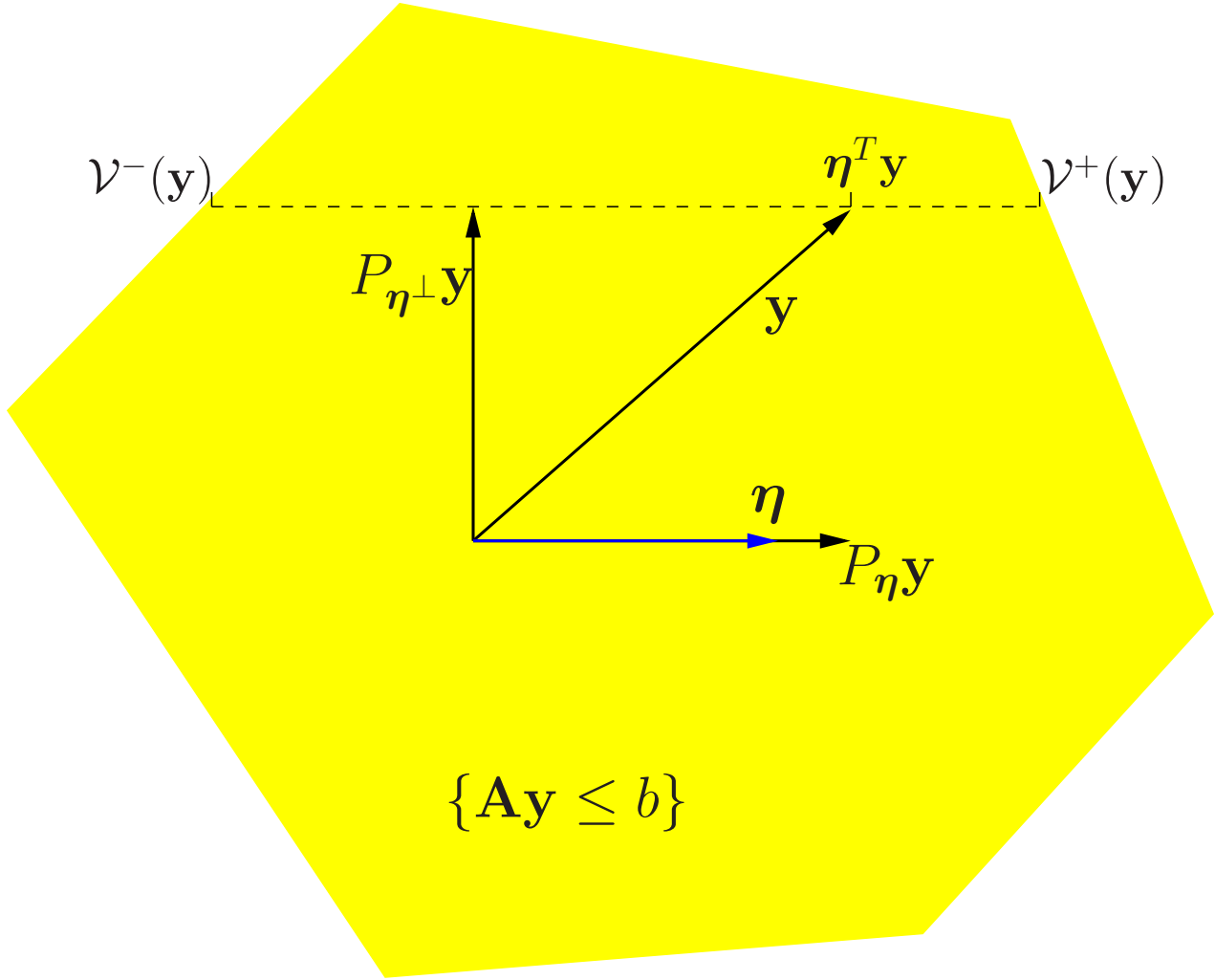
The polyhedral lemma

[Lee et al, Ryan Tibshirani et al.]
For any vector η

$$F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^\top\mu,\sigma^2\eta^\top\eta}(\eta^\top y)|\{Ay \leq b\} \sim \mathrm{Unif}(0,1)$$

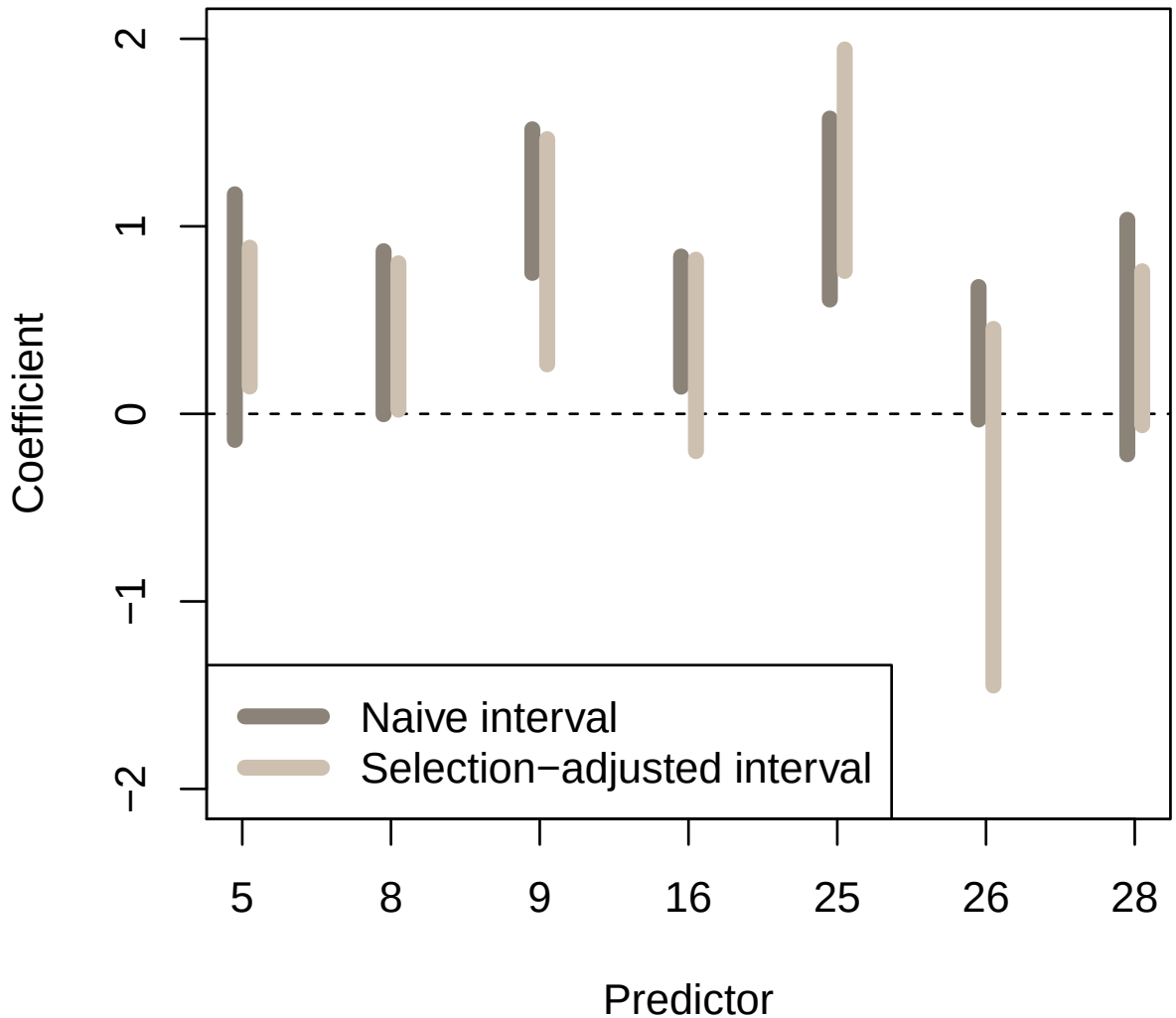
(truncated Gaussian distribution), where V^-, V^+ are (computable) values that are functions of η, A, b.

Typically choose η so that $\eta^T y$ is the partial least squares estimate for a selected variable


$\mathcal{V}^-(\mathbf{y})$
$\eta^T\mathbf{y}$
$\mathcal{V}^+(\mathbf{y})$
$P_{\eta^\perp}\mathbf{y}$
$\mathbf{y}$
η
$P_\eta\mathbf{y}$
$\{\mathbf{A}\mathbf{y} \leq b\}$

Example: Lasso with fixed-λ

HIV data: mutations that predict response to a drug. Selection intervals for lasso with fixed tuning parameter λ.



A big shortcoming of this approach

- Intervals are often very wide, can even be infinite.

A big shortcoming of this approach

- Intervals are often very wide, can even be infinite.
- Why? We have **conditioned on too much**, leaving not enough variation for inference [Fithian, Taylor- "data carving"].

A big shortcoming of this approach

- Intervals are often very wide, can even be infinite.
- Why? We have **conditioned on too much**, leaving not enough variation for inference [Fithian, Taylor- "data carving"].
- Jonathan Taylor & co-authors have worked to solve this problem by **adding noise** to the data before model fitting. This is clever and produces shorter intervals and more powerful tests.

A big shortcoming of this approach

- Intervals are often very wide, can even be infinite.
- Why? We have **conditioned on too much**, leaving not enough variation for inference [Fithian, Taylor- "data carving"].
- Jonathan Taylor & co-authors have worked to solve this problem by **adding noise** to the data before model fitting. This is clever and produces shorter intervals and more powerful tests.
- Here we show how the problem can be largely solved without randomization to provide shorter intervals.

Forming a Data Driven Query: Two Costs

1. **Variable Selection**: The data is used to decide which variables are worthy of attention, e.g., running the lasso and focusing on the active set.

Forming a Data Driven Query: Two Costs

1. **Variable Selection**: The data is used to decide which variables are worthy of attention, e.g., running the lasso and focusing on the active set.
2. **Target Formation**: Having settled on a subset $M \subset \{1, \ldots, p\}$ of variables for careful study, what should be the target of our estimation? Two choices:

Forming a Data Driven Query: Two Costs

1. **Variable Selection**: The data is used to decide which variables are worthy of attention, e.g., running the lasso and focusing on the active set.
2. **Target Formation**: Having settled on a subset $M \subset \{1, \ldots, p\}$ of variables for careful study, what should be the target of our estimation? Two choices:

   full target β_j^F, $j \in M$, where

   $$\beta^F = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mu,$$

   or **partial target**

   $$\beta^{(M)} = \left(\mathbf{X}_M^\top \mathbf{X}_M\right)^{-1} \mathbf{X}_M^\top \mu.$$

Consequences

- With the full target, our only cost is in #1.

  Our proposal: instead of conditioning on the entire active set and signs, we can condition **just on the event that a given variable X_j was chosen.**

  [minimal conditioning: it's the event that leads us to ask a question about X_j]

  This leads to a truncated Gaussian distribution on the union of two disjoint intervals, with exact coverage under Gaussian errors.

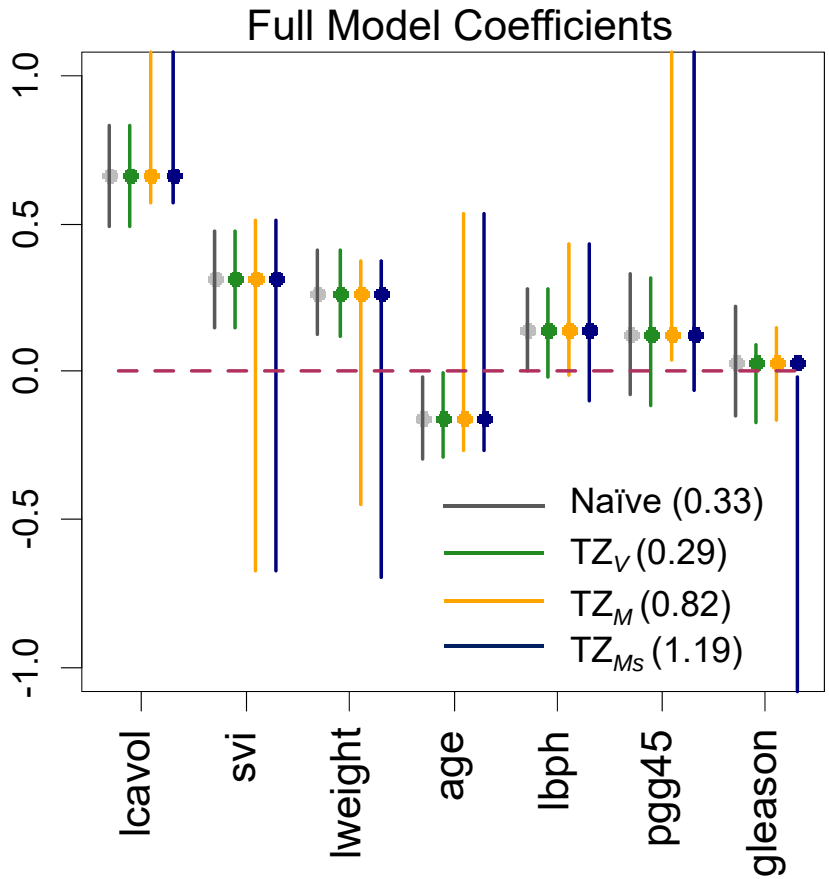
Consequences

- With the full target, our only cost is in #1.

  Our proposal: instead of conditioning on the entire active set and signs, we can condition **just on the event that a given variable X_j was chosen.**

  [minimal conditioning: it's the event that leads us to ask a question about X_j]

  This leads to a truncated Gaussian distribution on the union of two disjoint intervals, with exact coverage under Gaussian errors.
- With the partial target, we have to deal with both #1 and #2. Details in a few slides.



Prostate cancer data. Naive – ignore selection; TZ_V *– condition just on selected variable;* TZ_M *– condition on active set;* TZ_{Ms} *– condition on active set and signs (Lee et al.).*

Partial targets

Idea: we choose a subset $\hat{H} \subset \hat{M}$ of high value targets (details below). How we choose to summarize the effect of a variable $j \in \hat{M}$ depends on whether j is a high value target:

Partial targets

Idea: we choose a subset $\hat{H} \subset \hat{M}$ of high value targets (details below). How we choose to summarize the effect of a variable $j \in \hat{M}$ depends on whether j is a high value target:

- **High Value**: We summarize the effect of j using $\beta_j^{\hat{H}}$ where

$$\beta^{\hat{H}} = \left(\mathbf{X}_{\hat{H}}^\top \mathbf{X}_{\hat{H}}\right)^{-1} \mathbf{X}_{\hat{H}}^\top \mu.$$

  So our choice of target is fully adaptive for high value targets.

- **Low Value**: If variable j is selected by the lasso but is not deemed a high value target, we summarize its effect via $\beta_j^{\hat{H}\cup\{j\}}$ where

$$\beta^{\hat{H}\cup\{j\}} = \left(\mathbf{X}_{\hat{H}\cup\{j\}}^\top \mathbf{X}_{\hat{H}\cup\{j\}}\right)^{-1} \mathbf{X}_{\hat{H}\cup\{j\}}^\top \mu$$

  and $\mathbf{X}_{\hat{H}\cup\{j\}}$ is the matrix containing the high value targets as well as variable j. The coefficient $\beta_j^{\hat{H}\cup\{j\}}$ is the effect of variable j after partialing out the effect of the high value targets, i.e., it allows us to ask the question whether variable j contributes any explanatory power beyond the variables in $\hat{H}$.

Defining high and low-value targets

Stable-t: Take $\hat{H}$ to be those variables in $\hat{M}$ with t-statistics surpassing a Bonferroni corrected threshold. We first fit a OLS model using all the variables in $\hat{M}$, i.e.,

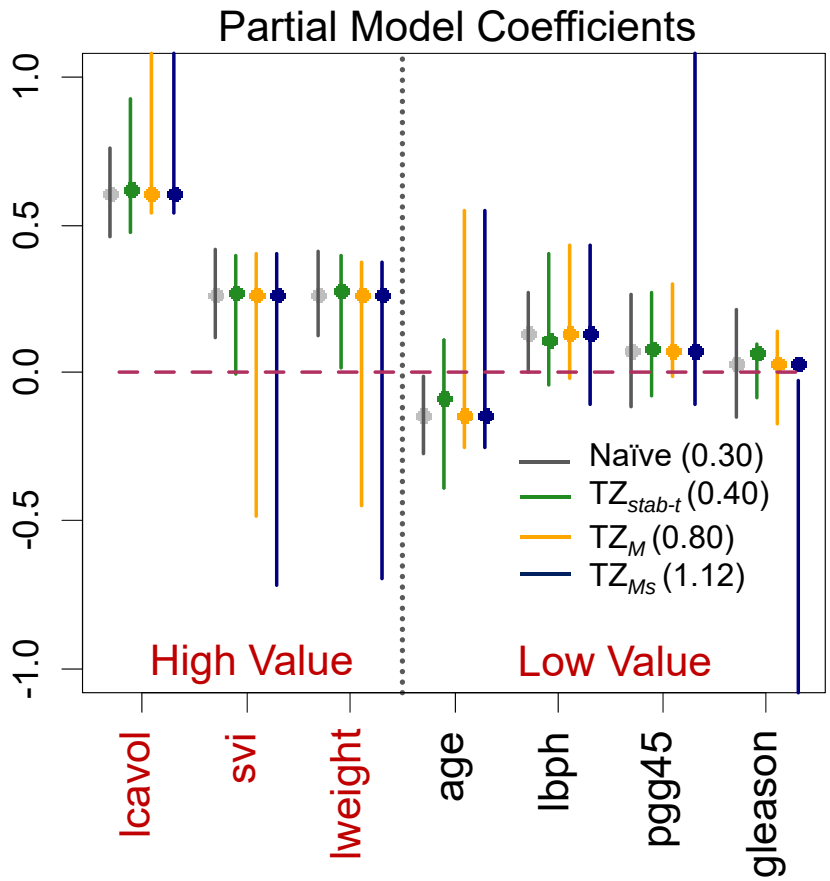
$$\hat{\beta}^{\hat{M}} = \left(\mathbf{X}_{\hat{M}}^{\top}\mathbf{X}_{\hat{M}}\right)^{-1}\mathbf{X}_{\hat{M}}^{\top}y$$

and allow j to be a high value target if the t-statistic for $\hat{\beta}_j^{\hat{M}}$ is large, i.e., if

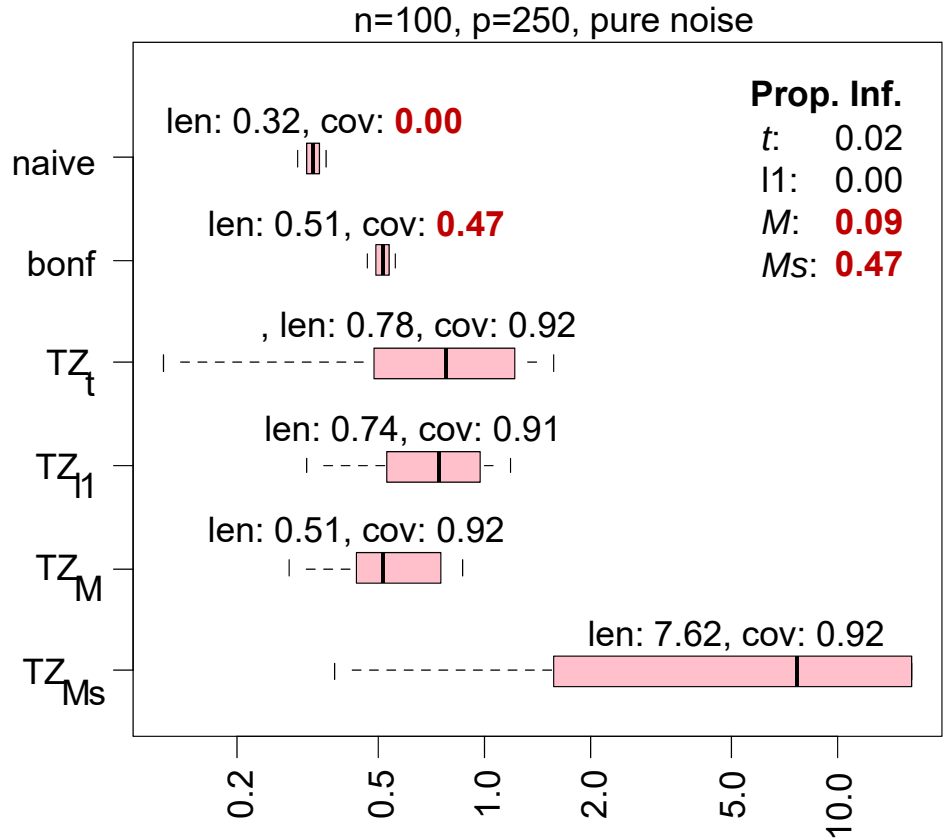
$$\left|\frac{\hat{\beta}_j^{\hat{M}}}{\sigma\left(\mathbf{X}_{\hat{M}}^{\top}\mathbf{X}_{\hat{M}}\right)_{jj}^{-1}}\right| > c$$

for some cutoff c. If we choose c by Bonferroni, it has the form $\left|\Phi^{-1}\left(\frac{\alpha}{2p}\right)\right| \approx \sqrt{2\log p}$ for large p;

We again get a truncated Gaussian over a union of intervals, and exact coverage with finite samples.



Prostate cancer data. Naive – ignore selection; TZ_V *– condition just on selected variable;* TZ_M *– condition on active set;* TZ_{Ms} *– condition on active set and signs (Lee et al.);* TZ_{stab-t} *– stable-t for high value target selection.*



Boxplot of lengths of 90% confidence intervals for "partial" regression coefficients.

Naive – ignore selection; Bonf – Bonferroni; TZ_t *– stable-t for high value target selection;* TZ_{l1} *– stable-*ℓ_1 *for high value target selection;* TZ_M *– condition on active set;* TZ_{Ms} *– condition on active set and signs (Lee et al.);*

Wrapup

- All of this is for $N > p$;

Wrapup

- All of this is for $N > p$;
- The ideas extended for the **high-dimensional full target** case via "ROSI:" in preparation with Kevin Fry, Keli Liu, Jonathan Taylor and Rob Tibshirani. Gets good power as well! Application to large GWAS problems.

Wrapup

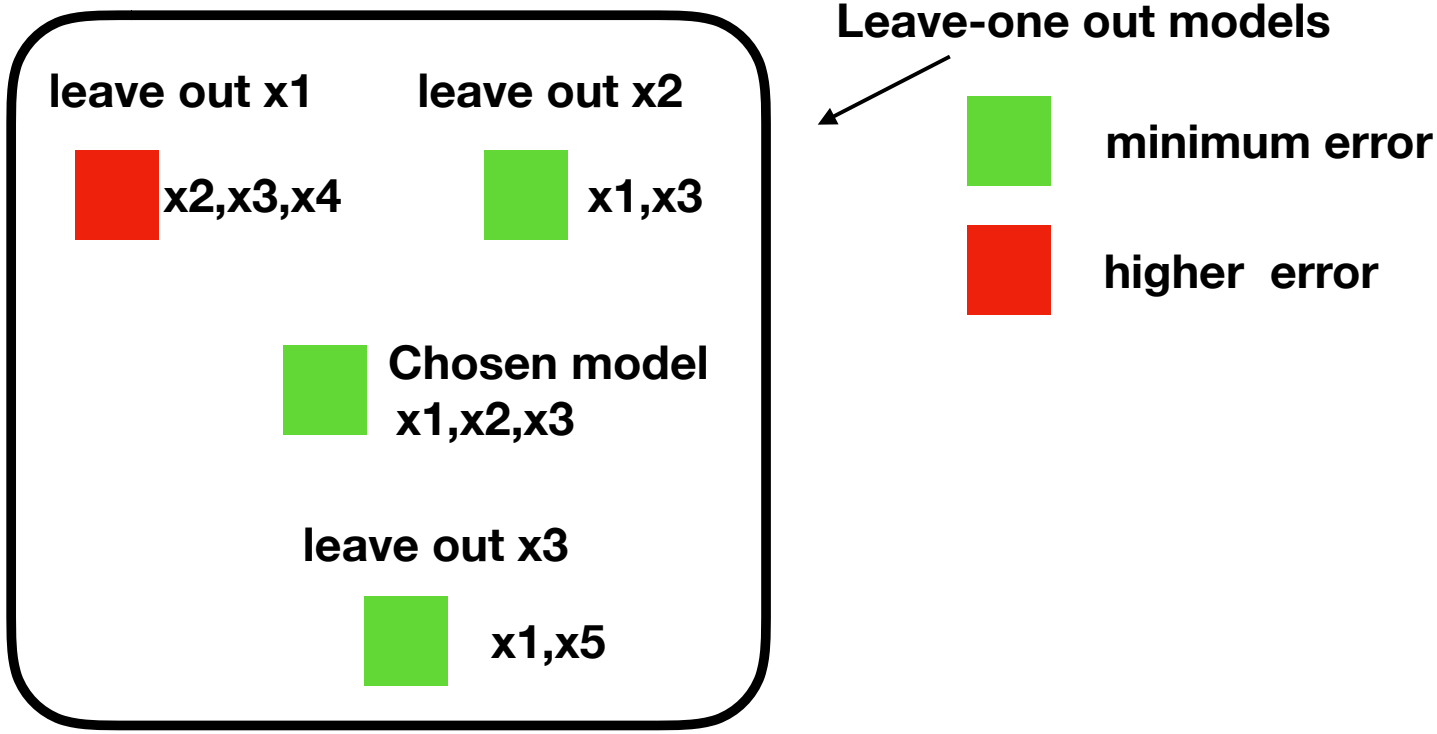
- All of this is for $N > p$;
- The ideas extended for the **high-dimensional full target** case via "ROSI:" in preparation with Kevin Fry, Keli Liu, Jonathan Taylor and Rob Tibshirani. Gets good power as well! Application to large GWAS problems.
- Will be added to our **selectiveInference** R and Python packages.

Next-door analysis

Motivation

- Having fit a model by e.g. lasso, post-selection inference (as above) focusses on significance and confidence intervals for each chosen feature
- But scientists will often have different questions:
  - Is the chosen model the uniquely best one?
  - Are there other models with similar prediction performance?
  - Is a given predictor indispensible or can it be swapped out for one or more other predictors?
- These are **model-centric** as opposed to **feature-centric** questions
- Our proposed solution is an application of the LOCO (leave-one-covariate-out) method of Lei et al (the CMU group)

  [no data splitting; focus on models, not variables]


leave out x1
x2,x3,x4
leave out x2
x1,x3
Chosen model
x1,x2,x3
leave out x3
x1,x5
Leave-one out models
minimum error
higher error

Algorithm: Next-Door analysis for the lasso

1. Fit the lasso with parameter λ chosen by cross-validation. Let the solutions be $\hat{\beta}(\hat{\lambda})$. Let S be the active set where the coefficient in $\hat{\beta}(\hat{\lambda})$ is non-zero.

2. For each $j \in S$, solve the lasso problem with the coefficient for the j^{th} predictor being fixed at 0:

$$\{\hat{\beta}_0, \hat{\beta}; \hat{\lambda}, j\} = \text{argmin}_{\beta_j=0} \frac{1}{2} \sum_i (y_i - \beta_0 - \sum_{\ell \neq j} x_{i\ell}\beta_\ell)^2 + \hat{\lambda} \sum_\ell |\beta_\ell| \quad (1)$$

   Let $\hat{\beta}(\hat{\lambda}; j)$ be the coefficients and d_j be the increase in validation error for this model relative to the base model.

3. Form an approximately unbiased estimate of d_j and test if predictor j is **indispensable**: that is, test whether the increase in estimated prediction error d_j is significantly larger than zero.

Details

- Need to condition on selection events: (1) chosen model has minimum CV error, (2) predictor j is in chosen model
- We use tricks of Markovic and Taylor (adding noise in CV) and Xiaoying Tian (adding $\pm$ noise for Cp) to obtain **approximately** debiased prediction error estimates and the bootstrap to get **approximate** type I error control

Table: *Prostate cancer results. The leftmost column shows the fitted model from the lasso, and the remaining columns show the nearby models corresponding to the removal of each predictor.*

	base	lcavol	lwt	svi	lcp	lbph	pgg45	age
lcavol	0.64		0.69	0.70	0.59	0.65	0.63	0.62
lwt	0.27	0.37		0.30	0.27	0.35	0.27	0.26
svi	0.25	0.46	0.29		0.22	0.21	0.27	0.25
lcp	-0.12	0.07	-0.11	-0.01		-0.14	-0.04	-0.11
lbph	0.18	0.21	0.29	0.14	0.19		0.18	0.17
pgg45	0.17	0.18	0.13	0.19	0.13	0.18		0.15
age	-0.08	-0.02	-0.03	-0.09	-0.07	-0.05	-0.07	
cv_error	0.61	0.90	0.65	0.64	0.62	0.61	0.63	0.60
debiased_error	0.62	0.94	0.66	0.66	0.63	0.62	0.62	0.62
test_error	0.51	0.87	0.49	0.56	0.50	0.50	0.47	0.53
selection freq		1.00	1.00	0.96	0.78	1.00	0.88	0.74
NextDoor pvalue		0.01	0.21	0.20	0.29	0.48	0.26	0.34
Feature (Post-Sel) pval		0.00	0.01	0.02	0.23	0.05	0.07	0.28

Post selection p-value, Frequency of selection $\neq$ Feature indispensability!!

Final comments

- Paper on arxiv by Guan & Tibshirani
- "NextDoor" R package will soon be on CRAN. Idea: run `glmnet` to fit model, then run `NextDoor` on the output to get post-fitting summary report