Recent Advances in Post-Selection Statistical Inference

Robert Tibshirani, Stanford University

December 9, 2015

Breiman lecture, NIPS 2015

Joint work with Jonathan Taylor, Richard Lockhart, Ryan Tibshirani, Will Fithian, Jason Lee, Yuekai Sun, Dennis Sun, Yun Jun Choi, Max G'Sell, Stefan Wager, Alex Chouldechova

Thanks to Jon, Ryan & Will for help with the slides.

My history

- PhD Statistics 1984 (Stanford); went home to University of Toronto (Biostatistics & Statistics)
- Met **Geoff Hinton** 
- and **Zoubin Ghahramani, Carl Rasmussen, Rich Zemel, Radford Neal, Brendan Frey, Peter Dayan**
- They were fitting models with **more parameters than observations!!**

My history at NIPS

My history at NIPS

- Starting in the early 90's, attended pretty regularly with Leo Breiman, Jerry Friedman & Trevor Hastie.

My history at NIPS

- Starting in the early 90's, attended pretty regularly with Leo Breiman, Jerry Friedman & Trevor Hastie.
- A few hundred people; The Statisticians were minor celebrities; "easy" to get a poster accepted.
- Jerry: **"The currency is ideas!"**

My history at NIPS

- Starting in the early 90's, attended pretty regularly with Leo Breiman, Jerry Friedman & Trevor Hastie.
- A few hundred people; The Statisticians were minor celebrities; "easy" to get a poster accepted.
- Jerry: **"The currency is ideas!"**
- But I haven't been for about 15 years

My recent experience:

My recent experience:

- Submitted papers to NIPS.

My recent experience:

- Submitted papers to NIPS. **No luck**

My recent experience:

- Submitted papers to NIPS. **No luck**
- Tried again

My recent experience:

- Submitted papers to NIPS. **No luck**
- Tried again **No luck**

My recent experience:

- Submitted papers to NIPS. **No luck**
- Tried again **No luck**
- Tried yet again

My recent experience:

- Submitted papers to NIPS. **No luck**
- Tried again **No luck**
- Tried yet again **No luck**

My recent experience:

- Submitted papers to NIPS. **No luck**
- Tried again **No luck**
- Tried yet again **No luck**
- Gave up

My recent experience:

- Submitted papers to NIPS. **No luck**
- Tried again **No luck**
- Tried yet again **No luck**
- Gave up

My recent experience:

- Submitted papers to NIPS. **No luck**
- Tried again **No luck**
- Tried yet again **No luck**
- Gave up
-

My recent experience:

- Submitted papers to NIPS. **No luck**
- Tried again **No luck**
- Tried yet again **No luck**
- Gave up
-
-
- Got invited!

My recent experience:

- Submitted papers to NIPS. **No luck**
- Tried again **No luck**
- Tried yet again **No luck**
- Gave up
-
-
- Got invited! (**because I'm getting old?**)

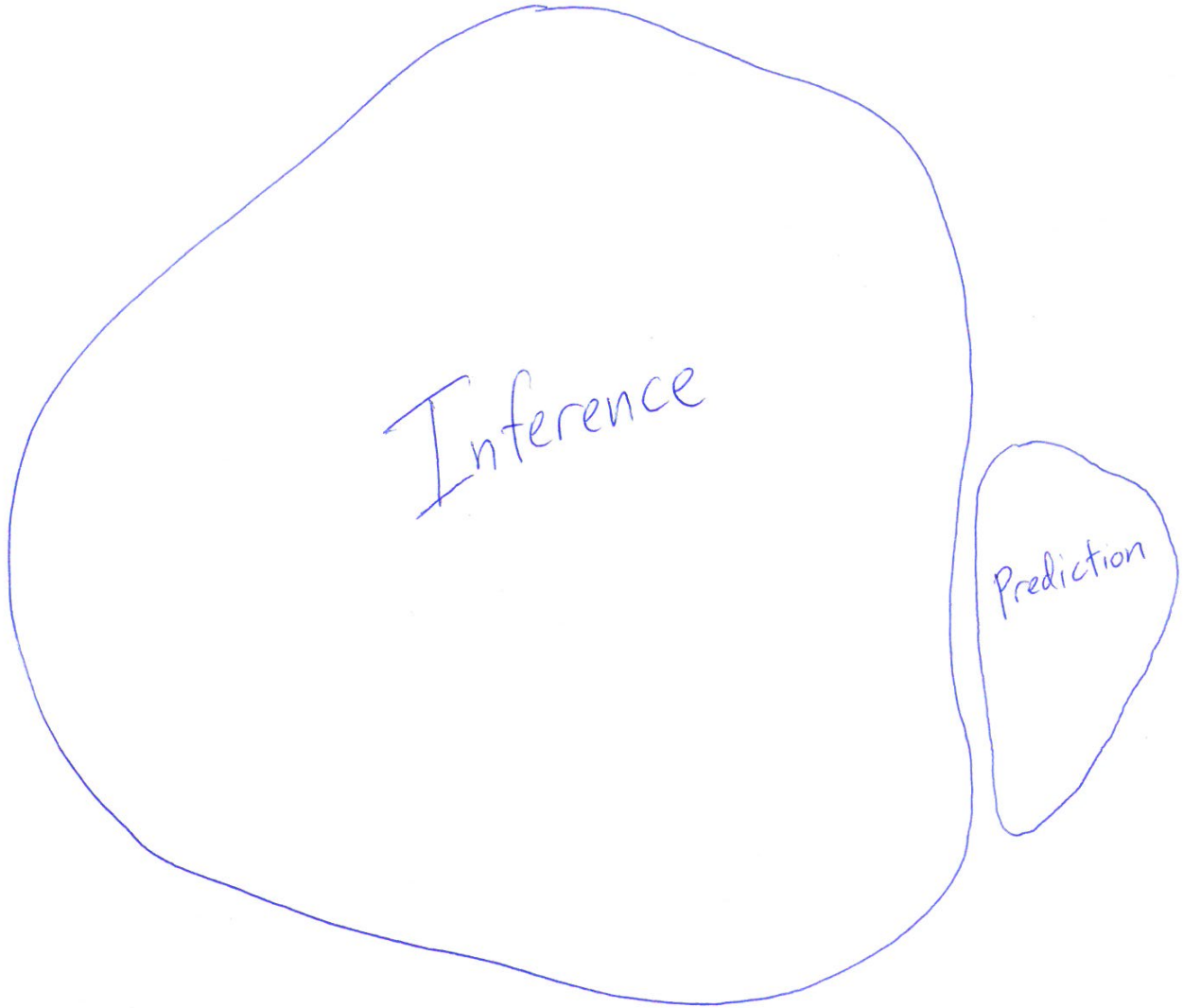
This talk is about Statistical Inference

Construction of p-values and confidence intervals for models and parameters fit to data

Frequentist (non-Bayesian) framework

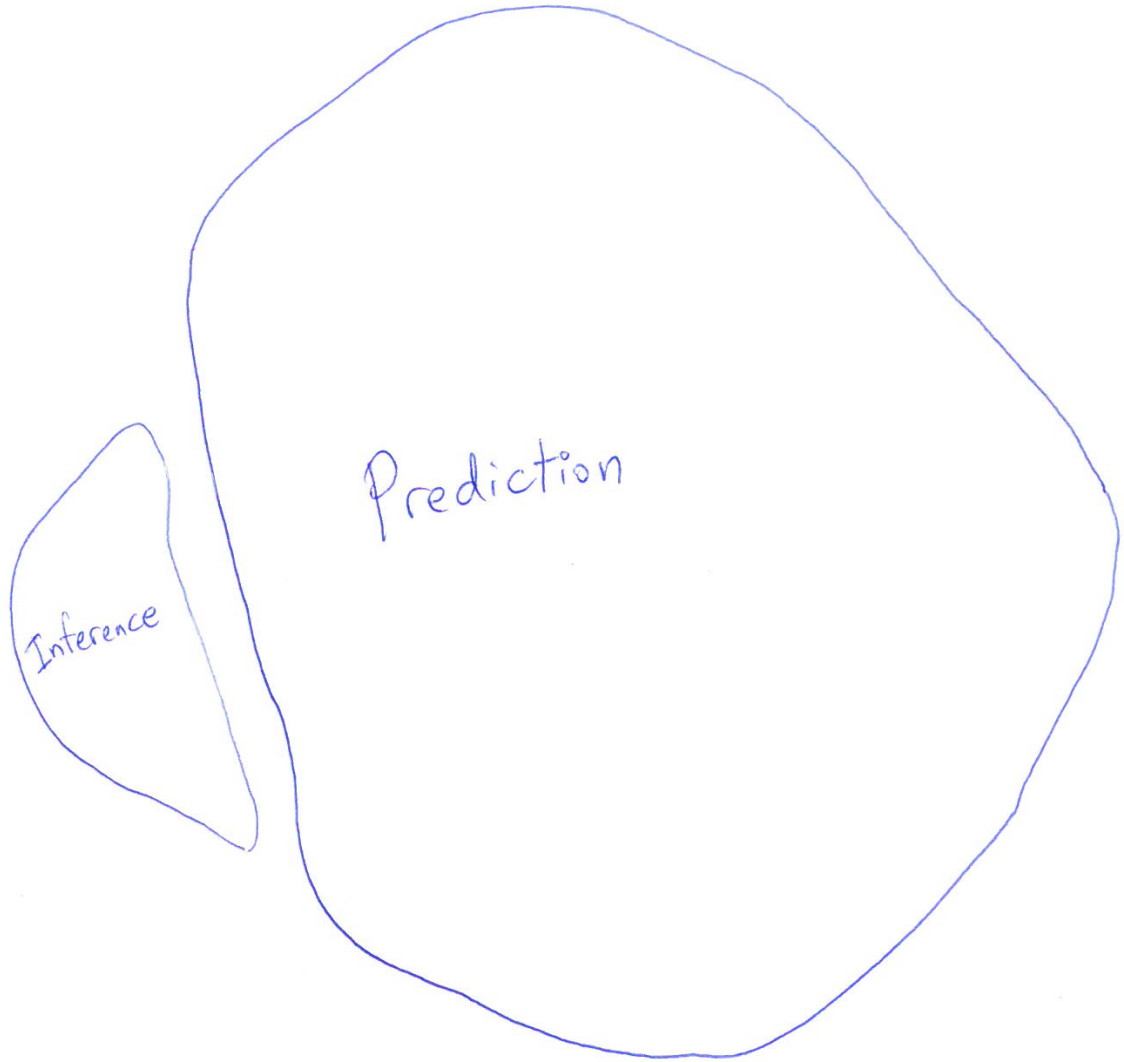
The word "**deep**" appears only once in what follows. ($5 U.S to the person who spots it first)

Statistics versus Machine Learning



How statisticians see the world?

Statistics versus Machine Learning



How machine learners see the world?

Quoting Leo

Statistical Science
2001, Vol. 16, No. 3, 199–231

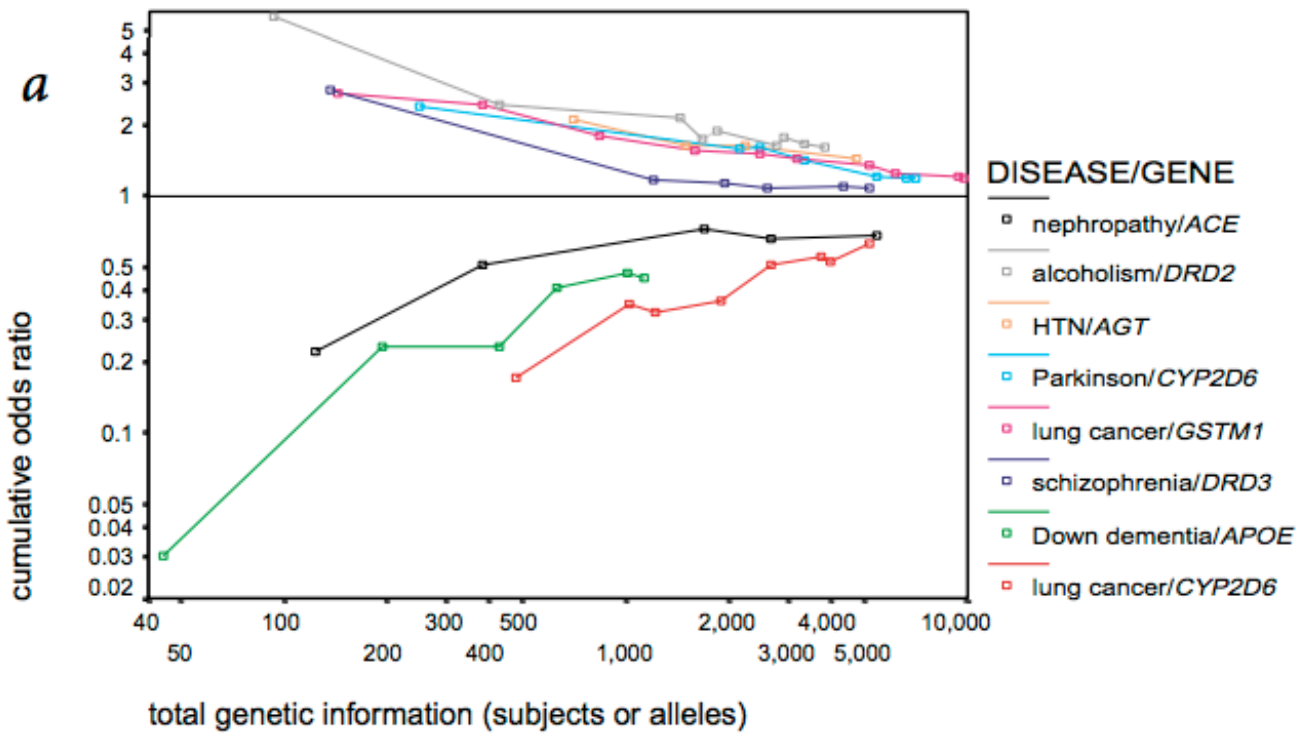
Statistical Modeling: The Two Cultures

Leo Breiman

There are **two cultures** in the use of statistical modeling to reach conclusions from data. One assumes that the **data are generated by a given stochastic data model.** The other uses **algorithmic models** and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to **irrelevant theory, questionable conclusions**, and has kept statisticians from working on a large range of interesting current problems.

Why inference is important

- In many situations we care about the identity of the features—e.g. biomarker studies: Which genes are related to cancer?
- There is a crisis in reproducibility in Science: Ioannidis (2005) "Why Most Published Research Findings Are False"



The crisis- continued

- Part of the problem is non-statistical- e.g. **incentives** for authors or journals to get things right.
- But part of the problem is **statistical**– we search through large number of models to find the "best" one; we don't have good ways of assessing the strength of the evidence
- today's talk reports some progress on the development of statistical tools for assessing the strength of evidence, after model selection

Our first paper on this topic: An all "Canadian" team

Richard Lockhart
Simon Fraser University
Vancouver
PhD . Student of David Blackwell, Berkeley, 1979

Jonathan Taylor
Stanford University
PhD Student of Keith Worsley, 2001

Ryan Tibshirani ,
CMU. PhD student of Taylor
2011

Rob Tibshirani
Stanford

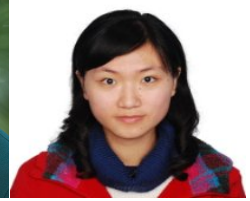
Fundamental contributions by some terrific students!



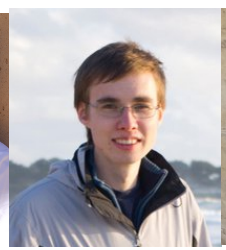
Will Fithian **Jason Lee** **Yuekai Sun**

UC BERKELEY





Max G'Sell, CMU Dennis Sun- Google Xiaoying Tian, Stanford

Yun Jin Choi Stefan Wager Alex Chouldchova Josh Loftus

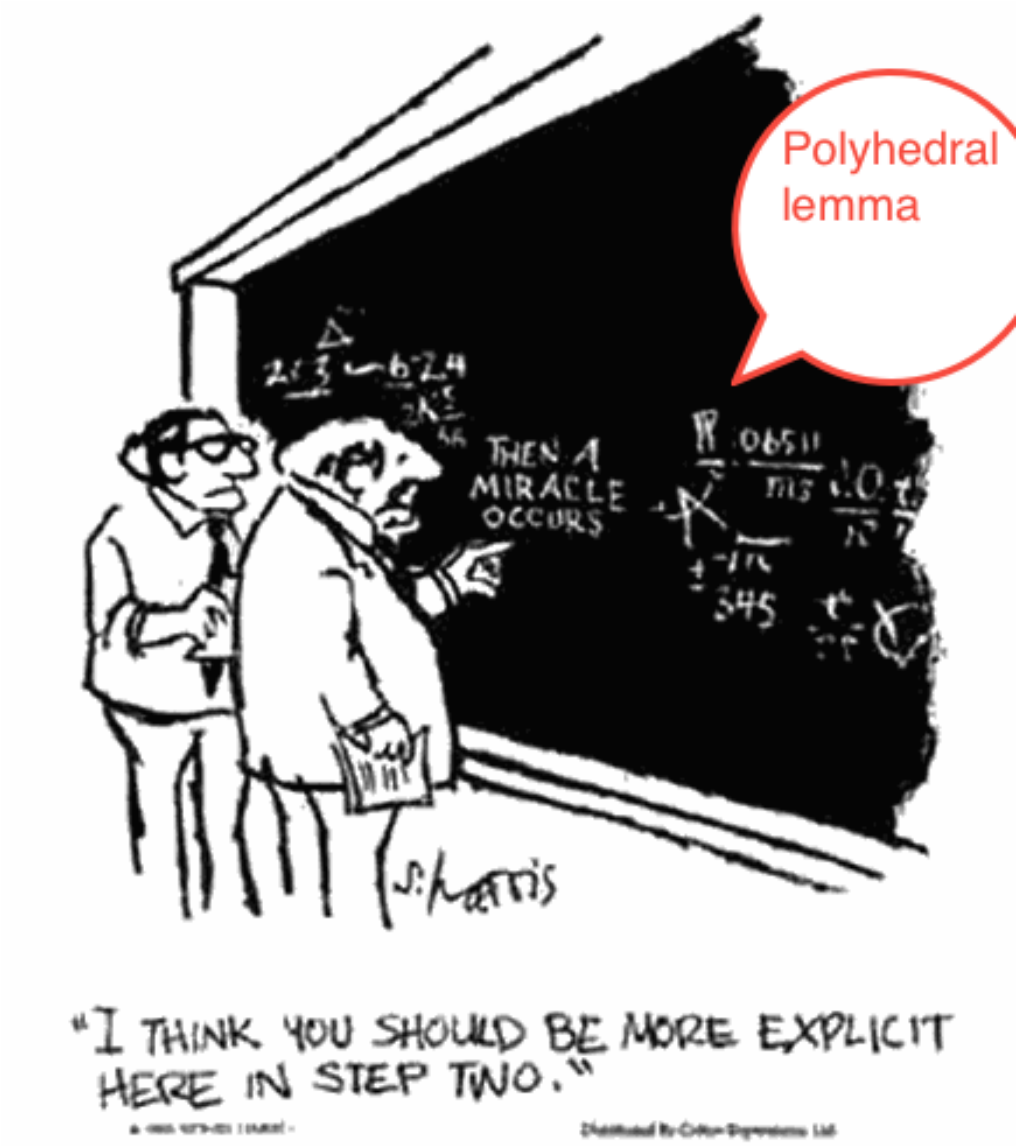
Now at CMU

STANFORD

Some key papers in this work

- Lockhart, Taylor, Tibs & Tibs. **A significance test for the lasso**. Annals of Statistics 2014
- Lee, Sun, Sun, Taylor (2013) **Exact post-selection inference with the lasso**. arXiv; To appear
- Fithian, Sun, Taylor (2015) **Optimal inference after model selection**. arXiv. Submitted
- Tibshirani, Ryan, Taylor, Lockhart, Tibs (2016) **Exact Post-selection Inference for Sequential Regression Procedures**. To appear, JASA
- Tian, X. and Taylor, J. (2015) **Selective inference with a randomized response**. arXiv
- Fithian, Taylor, Tibs, Tibs (2015) **Selective Sequential Model Selection**. arXiv Dec 2015

What it's like to work with Jon Taylor



Outline

1. The post-selection inference challenge; main examples— **Forward stepwise regression and lasso**
2. A simple procedure achieving exact post-selection type I error. No sampling required— explicit formulae.
3. Exponential family framework: more powerful procedures, requiring MCMC sampling
4. **Data splitting, data carving, randomized response**
5. When to stop Forward stepwise? FDR-controlling procedures using post-selection adjusted p-values
6. A more difficult example: **PCA**
7. Future work

What is post-selection inference?

Inference the old way (pre-1980?) :

1. Devise a model
2. Collect data
3. Test hypotheses

Classical inference

Inference the new way:

1. Collect data
2. Select a model
3. Test hypotheses

Post-selection inference

Classical tools cannot be used post-selection, because they do not yield valid inferences (generally, too optimistic)

The reason: classical inference considers a fixed hypothesis to be tested, not a random one (adaptively specified)

Leo Breiman referred to the use of classical tools for post-selection inference as a "**quiet scandal**" in the statistical community.

(It's not often Statisticians are involved in scandals)

Linear regression

- Data $(x_i, y_i), i = 1, 2, \ldots N$; $x_i = (x_{i1}, x_{I2}, \ldots x_{ip})$.
- Model
$$y_i = \beta_0 + \sum_j x_{ij}\beta_j + \epsilon_i$$
- **Forward stepwise regression**: greedy algorithm, adding predictor at each stage that most reduces the training error
- **Lasso**
$$\text{argmin}\Big\{\sum_i (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda \cdot \sum_j |\beta_j|\Big\}$$

  for some $\lambda \geq 0$.
  Either fixed λ, or over a path of λ values (Least angle regression).

Post selection inference

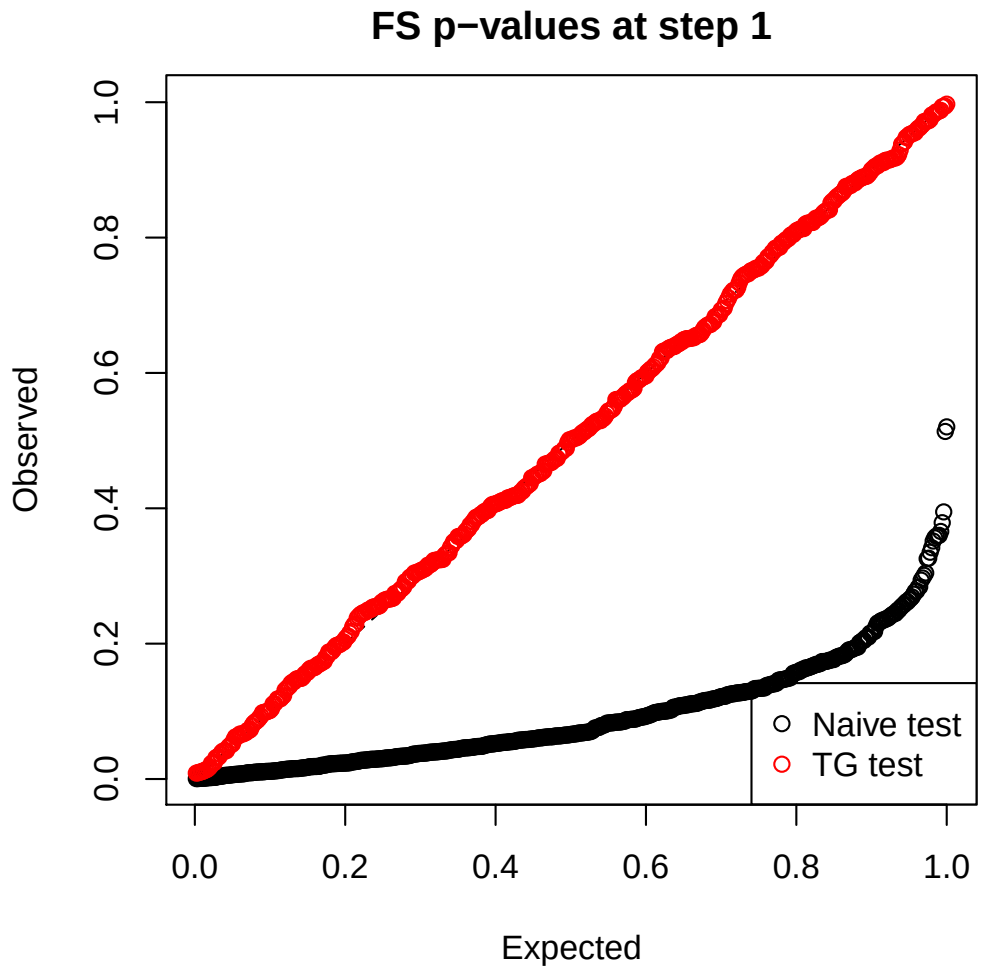
Example: Forward Stepwise regression

	FS, naive	FS, adjusted
lcavol	0.000	0.000
lweight	0.000	0.012
svi	0.047	0.849
lbph	0.047	0.337
pgg45	0.234	0.847
lcp	0.083	0.546
age	0.137	0.118
gleason	0.883	0.311

Table: *Prostate data example:* $n = 88, p = 8$. *Naive and selection-adjusted forward stepwise sequential tests*

With Gaussian errors, P-values on the right are **exact** in finite samples.

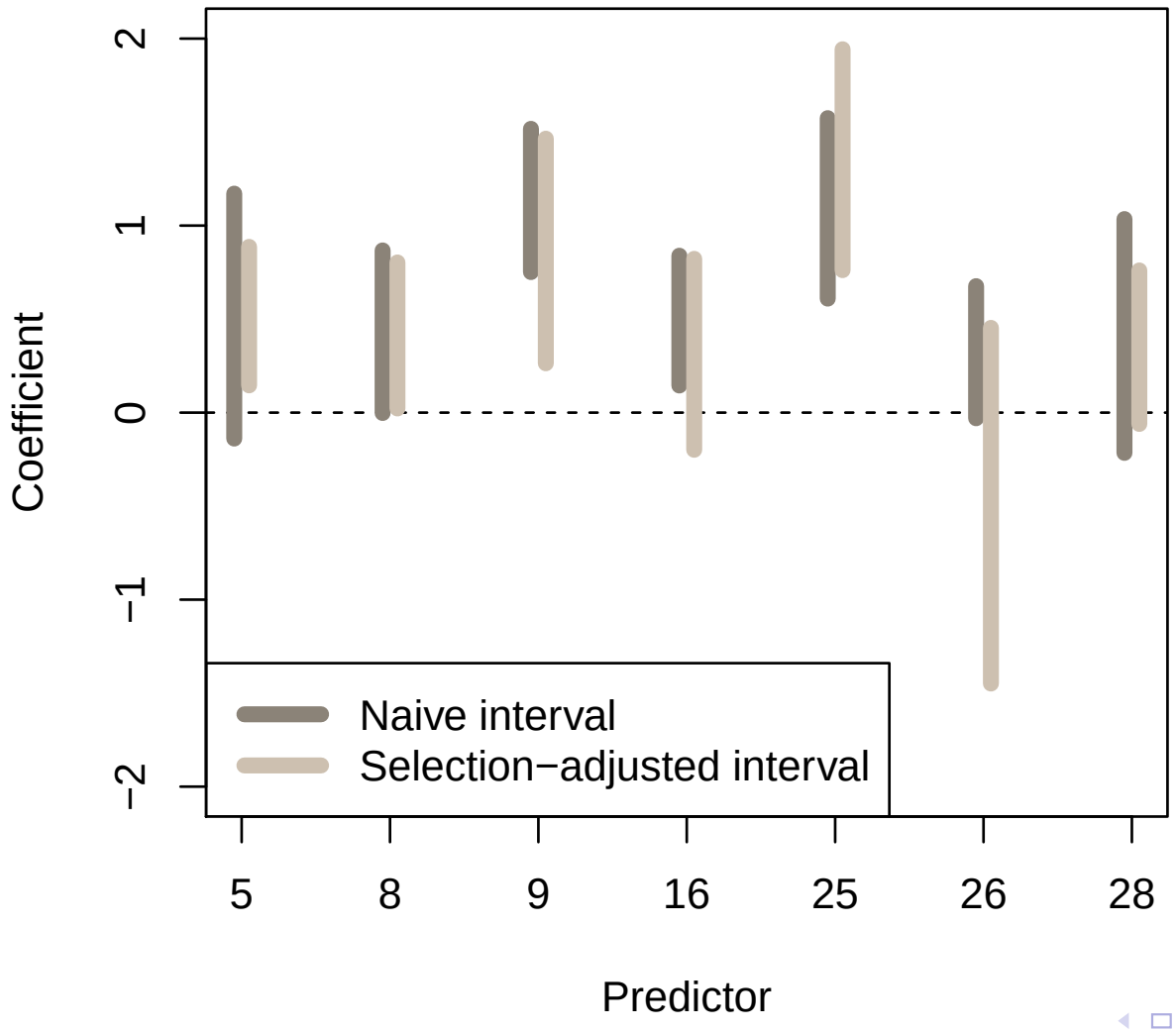
Simulation: $n = 100$, $p = 10$, and $y, X_1, \ldots X_p$ have i.i.d. $N(0, 1)$ components



Adaptive selection clearly makes χ^2_1 null distribution invalid; with nominal level of 5%, actual type I error is about 30%.

Example: Lasso with fixed-λ

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



Formal goal of Post-selective inference

[Lee et al. and Fithian, Sun, Taylor]

- Having selected a model $\hat{M}$ based on our data y, we'd like to test an hypothesis $\hat{H}_0$. Note that $\hat{H}_0$ will be **random** — a function of the selected model and hence of y
- If our rejection region is $\{T(y) \in R\}$, we want to control the *selective type I error* :

$$\text{Prob}(T(y) \in R | \hat{M}, \hat{H}_0) \leq \alpha$$

Existing approaches

- **Data splitting** - fit on one half of the data, do inferences on the other half. Problem- fitted model changes varies with random choice of "half"; loss of power. More on this later
- **Permutations and related methods**: not clear how to use these, beyond the global null

A key mathematical result

Polyhedral lemma: Provides a good solution for FS; an optimal solution for the fixed-λ lasso

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.

Some intuition for Forward stepwise regression

- Suppose that we run forward stepwise regression for k steps
- $\{y : Ay \leq b\}$ is the set of y vectors that would yield the same predictors and their signs entered at each step.
- Each step represents a competition involving inner products between each x_j and y ; Polyhedron $Ay \leq b$ summarizes the results of the competition after k steps.
- Similar result holds for Lasso (fixed-λ or LAR)

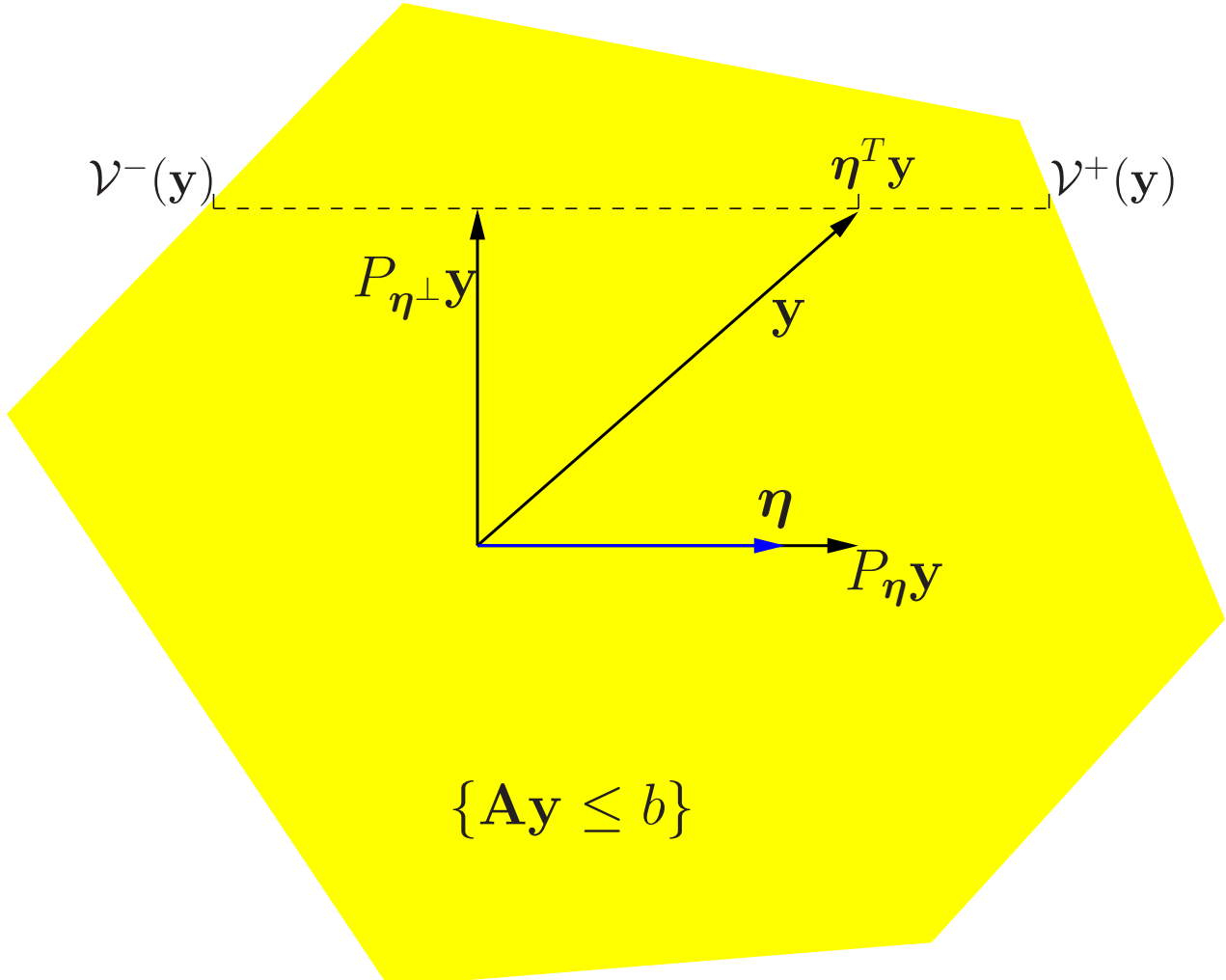
The polyhedral lemma

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

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

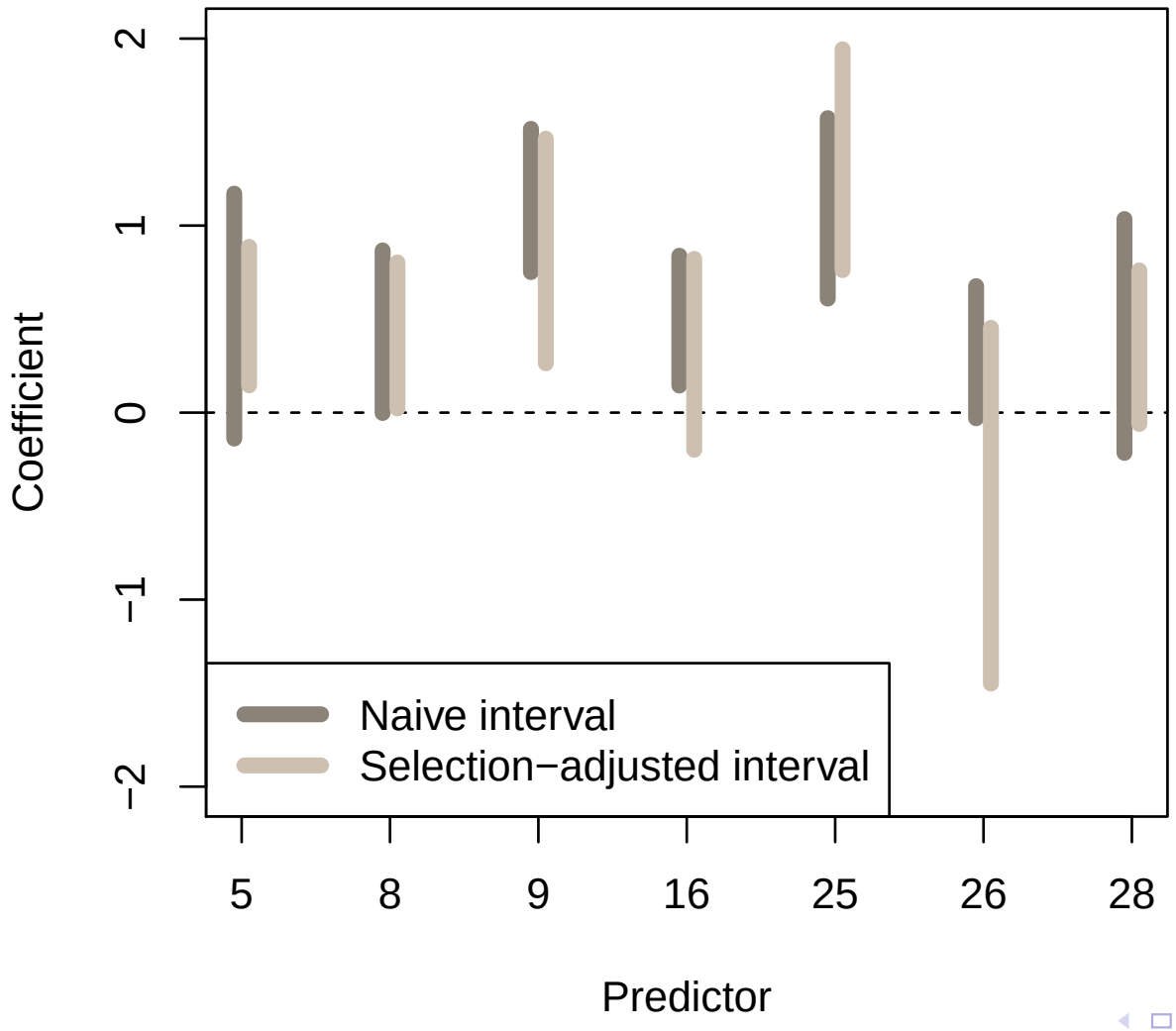
(truncated Gaussian distribution), where V^-, V^+ are (computable) values that are functions of η, A, b.

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.



Example: Lasso with λ estimated by Cross-validation

- Current work- Josh Loftus, Xiaoying Tian (Stanford)
- Can condition on the selection of λ by CV, and addition to the selection of model
- Not clear yet how much difference is makes (vs treating it as fixed)

Improving the power

- The preceding approach conditions on the part of y orthogonal to the direction of interest η. This is for computational convenience– yielding an analytic solution.
- **Conditioning on less $\rightarrow$ more power**

  Are we conditioning on too much?

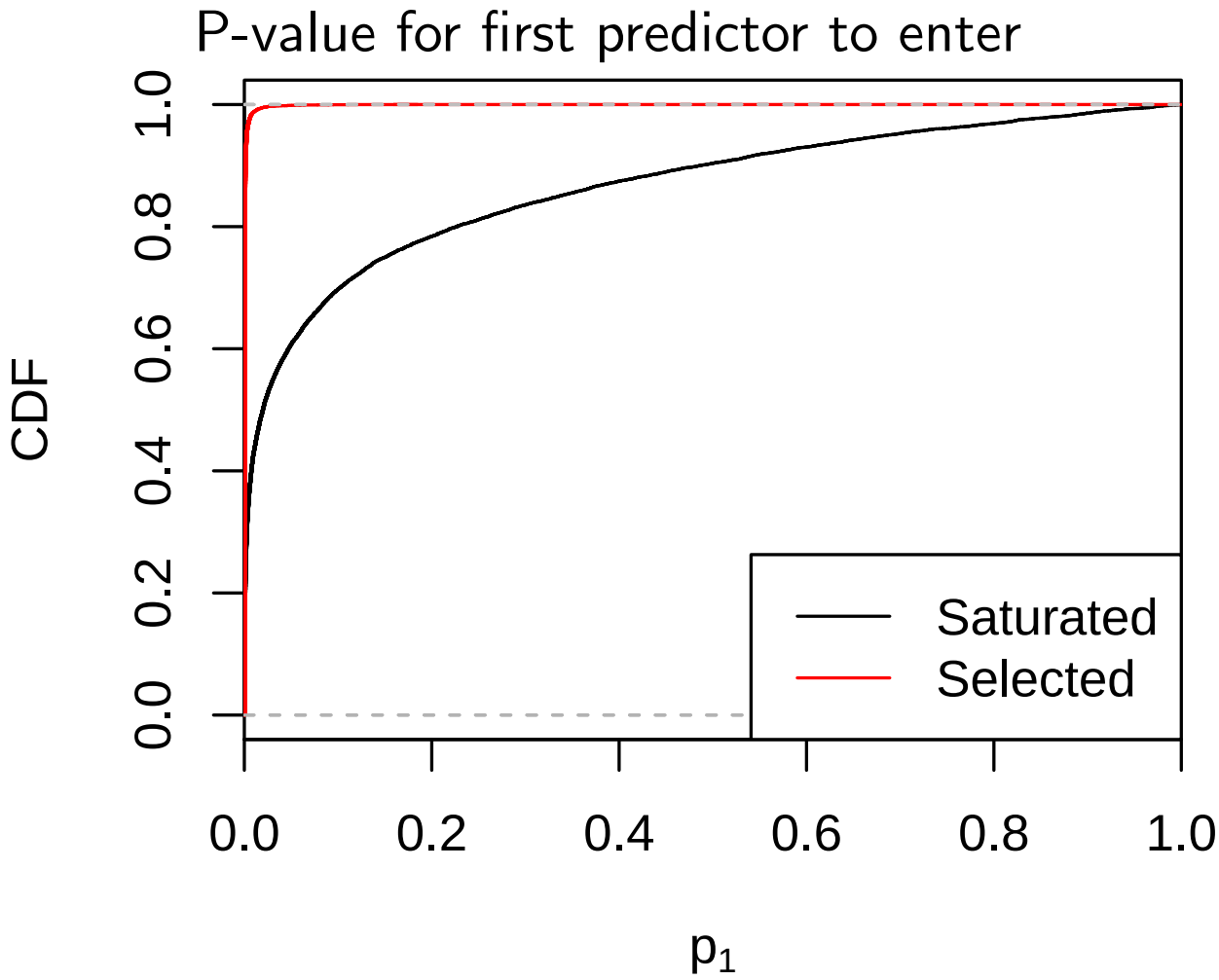
Exponential family framework

- Fithian, Sun and Taylor (2014) develop an optimal theory of post-selection inference: their **selective** model conditions on less: just the sufficient statistics for the nuisance parameters in the exponential family model.

  Saturated model $y = \mu + \epsilon \ \rightarrow$ condition on $P_{\eta^\perp} y$
  Selective model : $y = X_M \beta_M + \epsilon \ \rightarrow$ condition on $X_{M/j}^T y$

- Selective model gives the **exact** and **uniformly most unbiassed powerful test** but usually requires accept/reject or MCMC sampling.
- For the lasso, the **saturated** and **selective** models agree; sampling is not required

Two signals of equal strength



Outline

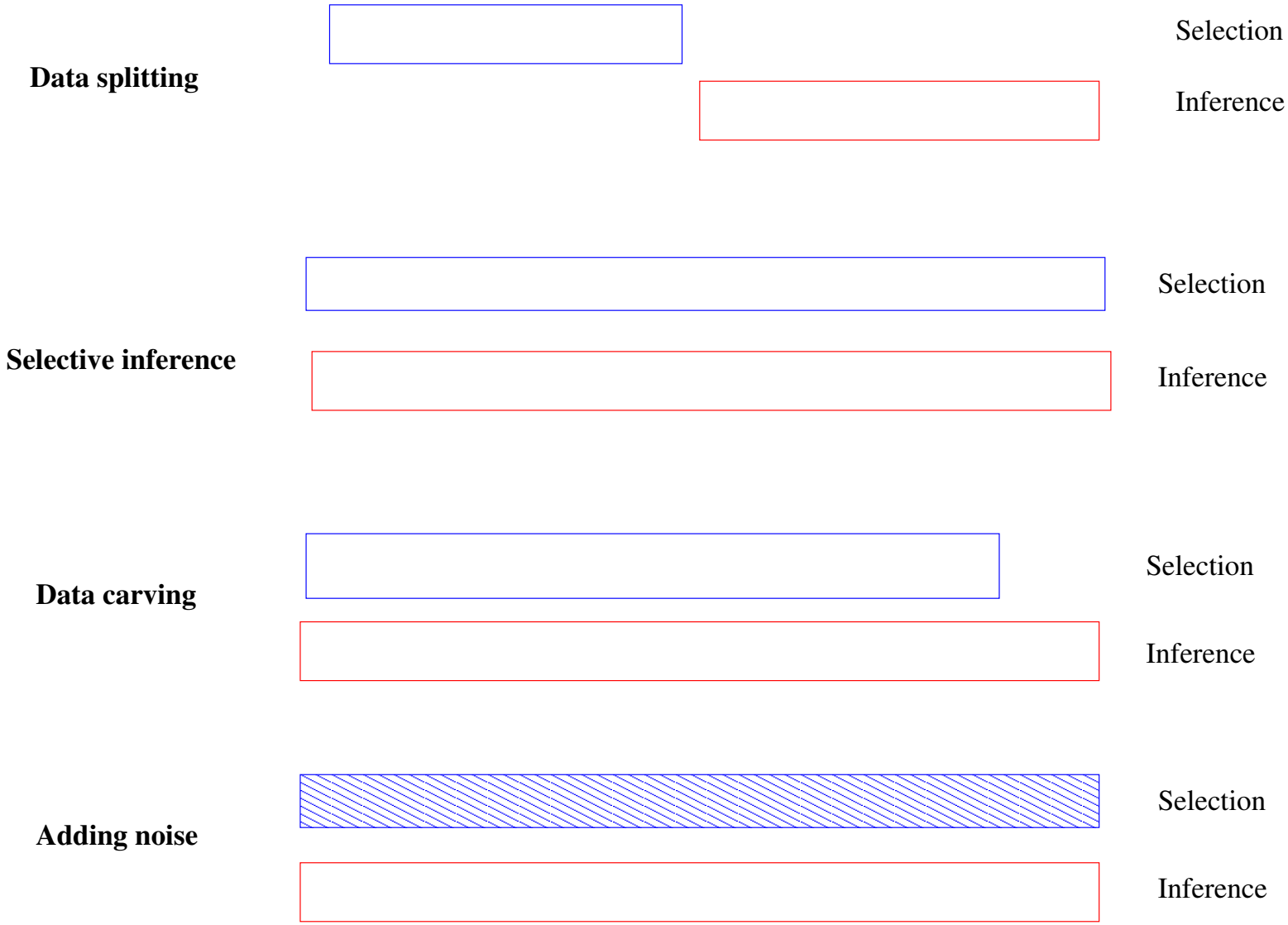
1. The post-selection inference challenge; main examples— Forward stepwise regression and lasso
2. A simple procedure achieving exact post-selection type I error. No sampling required— explicit formulae.
3. Exponential family framework: more powerful procedures, requiring MCMC sampling
4. $\Longrightarrow$ **Data splitting, data carving, randomized response**
5. When to stop Forward stepwise? FDR-controlling procedures using selection adjusted p-values
6. A more difficult example: **PCA**
7. Future work

Data splitting, carving, and adding noise

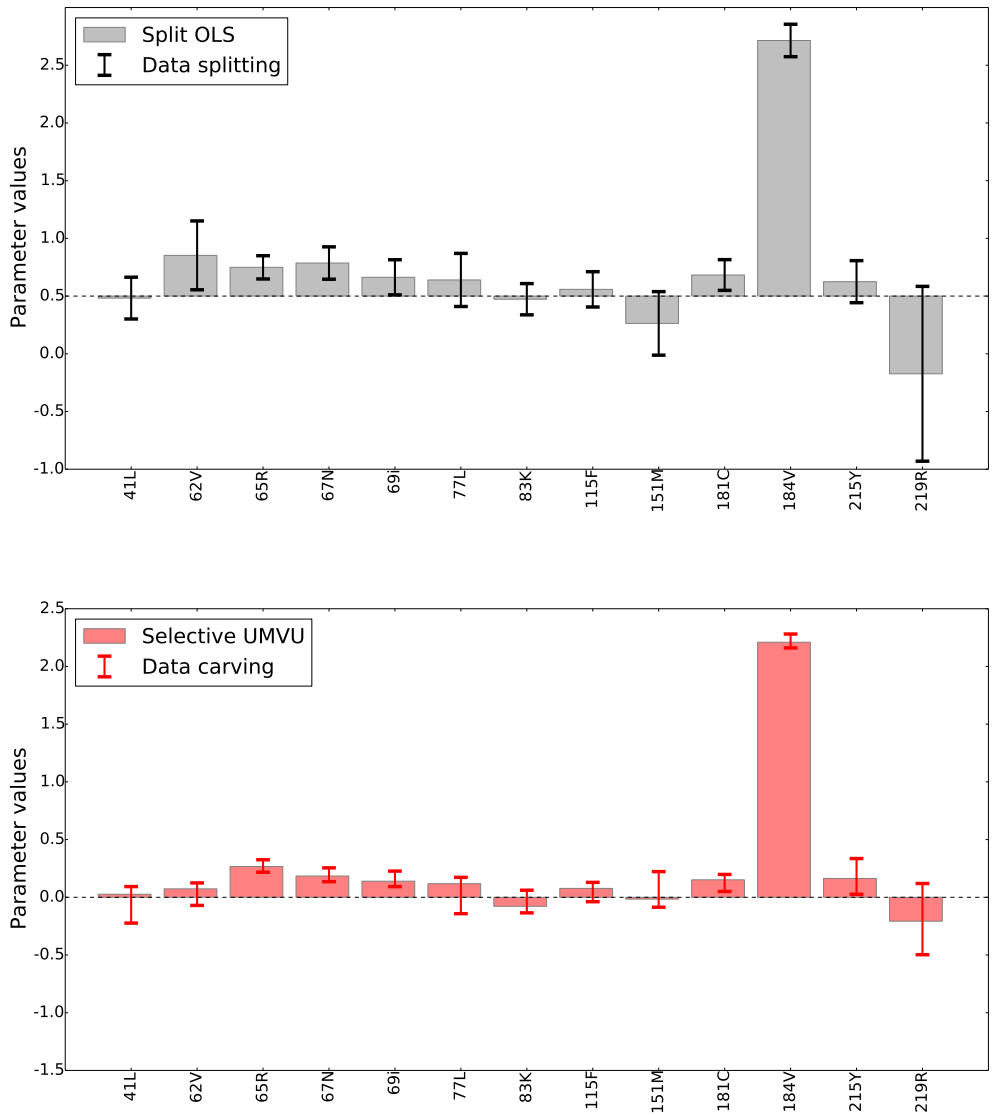
Further improvements in power

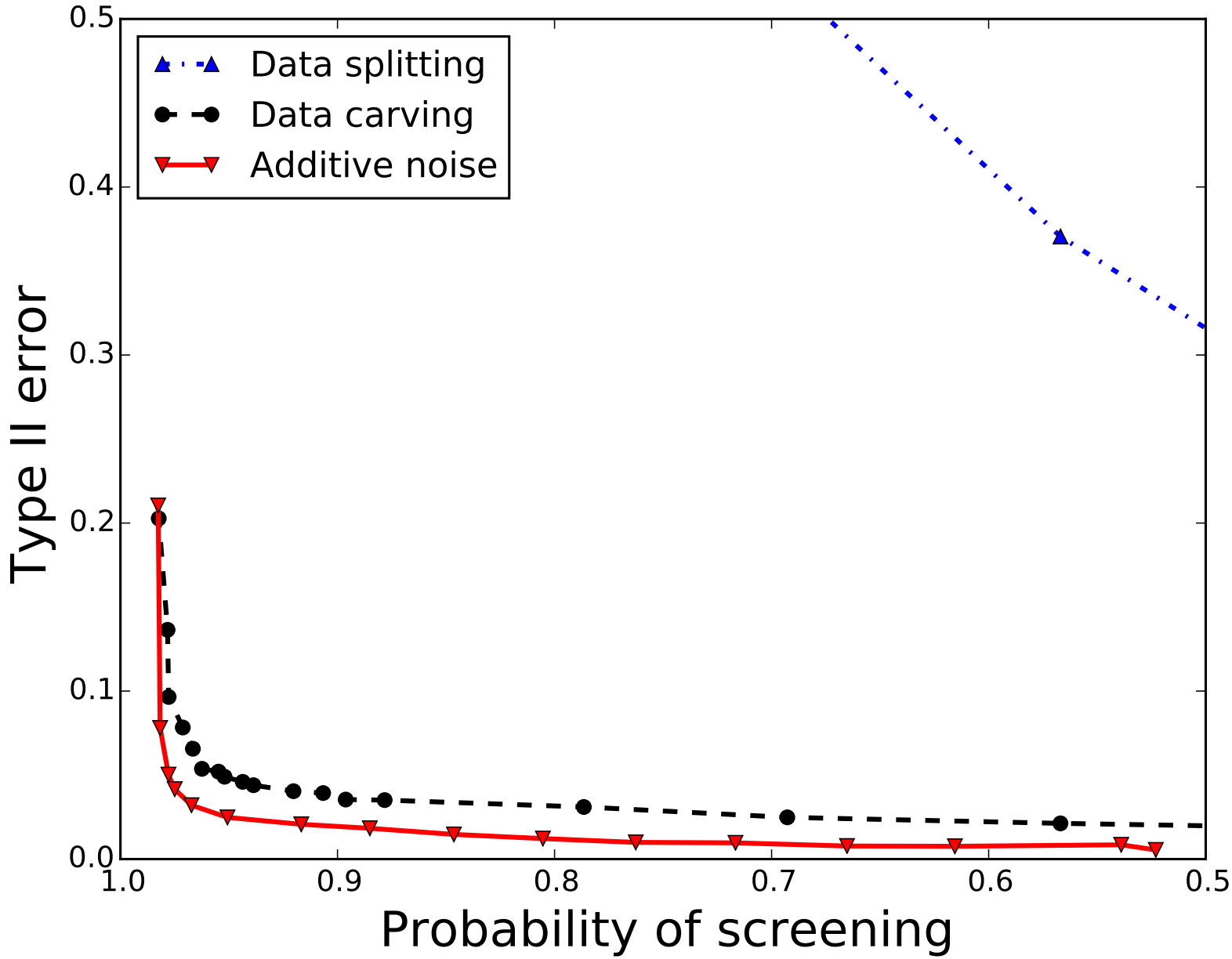
Fithian, Sun, Taylor, Tian

- Selective inference yields correct post-selection type I error. But confidence intervals are sometimes quite long. How to do better?
- **Data carving:** withholds a small proportion (say 10%) of data in selection stage, then uses all data for inference (conditioning using theory outlined above)
- **Randomized response:** add noise to y in selection stage. Like withholding data, but smoother. Then use unnoised data in inference stage. Related to **differential privacy** techniques.


Data splitting
Selection
Inference
Selective inference
Selection
Inference
Data carving
Selection
Inference
Adding noise
Selection
Inference

HIV mutation data; 250 predictors




Data splitting
Data carving
Additive noise
Type II error
0.0
0.1
0.2
0.3
0.4
0.5
1.0
0.9
0.8
0.7
0.6
0.5
Probability of screening

What is a good stopping rule for Forward Stepwise Regression?

	FS, naive	FS, adjusted
lcavol	0.000	0.000
lweight	0.000	0.012
svi	0.047	0.849
lbph	0.047	0.337
pgg45	0.234	0.847
lcp	0.083	0.546
age	0.137	0.118
gleason	0.883	0.311

- Stop when a p-value exceeds say 0.05?
- We can do better: we can obtain a more powerful test, with FDR (false discovery rate) control

False Discovery Rate control using sequential p-values

G'Sell, Wager, Chouldchova, Tibs

Hypotheses	H_1	H_2	H_3	$\ldots$	H_{m-1}	H_m
p-values	p_1	p_2	p_3	$\ldots$	p_{m-1}	p_m

- Hypotheses are considered **ordered**
- Testing procedure must reject $H_1, \ldots, H$ for some $\in \{0, 1, \ldots, m\}$
  - E.g., in sequential model selection, this is equivalent to selecting the first k variables along the path

Goal

Construct testing procedure $= (p_1, \ldots, p_m)$ that gives FDR control.
Can't use standard BH rule, because hypothesis are ordered.

A new stopping procedure:

ForwardStop

$$\hat{k}_F = \max\left\{ k \in \{1, \ldots, m\} : \frac{1}{k}\sum_{i=1}^{k}\{-\log(1-p_i)\} \leq \alpha \right\}$$

- Controls FDR even if null and non-null hypotheses are intermixed.
- Very recent work of Li and Barber (2015) on **Accumulation tests** generalizes the forwardStop rule

Requirement: p-values that are independent under the null

- The sequential p-values from the saturated model, are **not independent under the null**
- New paper: "**Selective Sequential Model Selection**"- Fithian, Taylor, Tibs & Tibs presents a general framework for deriving **exactly independent** null-p-values, for FS regression, least angle regression etc.
- These are based on the selective model, using the familiar "**max-T**" statistic
- they require accept/reject or hit-and-run MCMC sampling

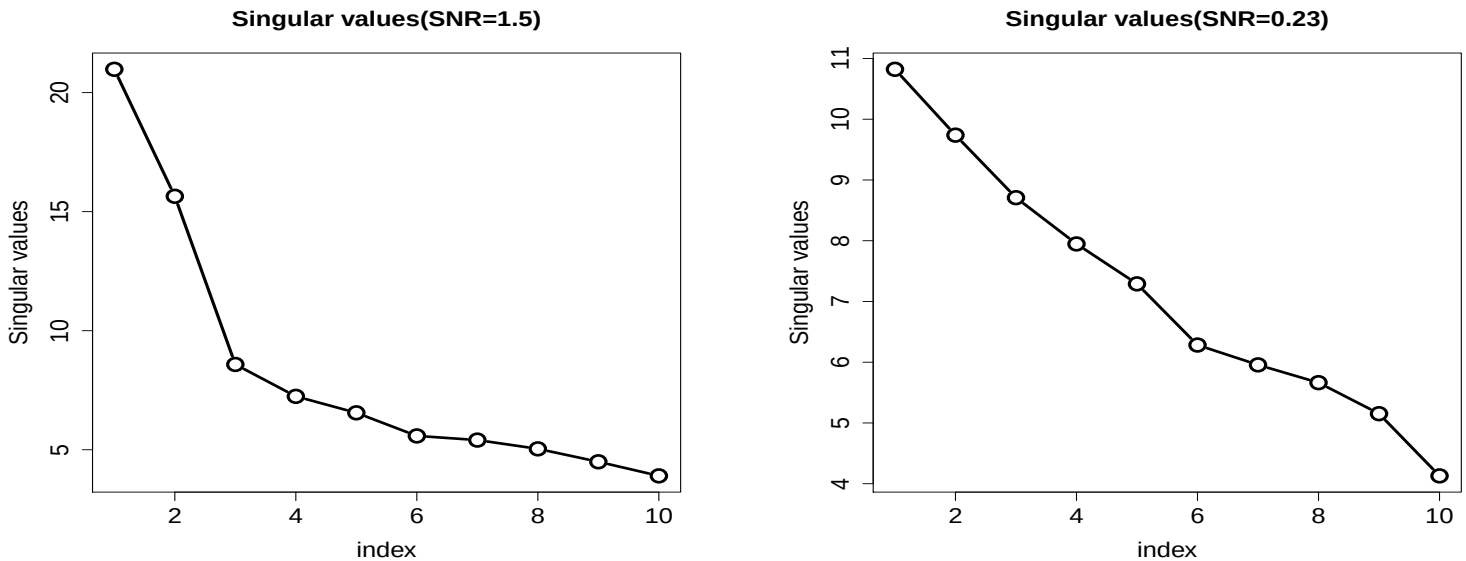
Step	Variable	Nominal p-value	Saturated p-value	Max-t p-value
1	bmi	0.00	0.00	0.00
2	ltg	0.00	0.00	0.00
3	map	0.00	**0.05**	0.00
4	age:sex	0.00	0.33	0.02
5	bmi:map	0.00	0.76	0.08
6	hdl	0.00	0.25	0.06
7	sex	0.00	0.00	0.00
8	glu^2	0.02	0.03	**0.32**
9	age^2	0.11	0.55	0.94
10	map:glu	0.17	0.91	0.91
11	tc	0.15	0.37	0.25
12	ldl	0.06	0.15	0.01
13	ltg^2	0.00	0.07	0.04
14	age:ldl	0.19	0.97	0.85
15	age:tc	0.08	0.15	0.03
16	sex:map	0.18	0.05	0.40
17	glu	0.23	0.45	0.58
18	tch	**0.31**	0.71	0.82
19	sex:tch	0.22	0.40	0.51
20	sex:bmi	0.27	0.60	0.44

A more difficult example: choosing the rank for PCA

- Choosing rank in PCA analysis is a difficult problem
- Can view PCA of increasing ranks as a stepwise regression of the columns of matrix on its singular vectors
- Condition on the selected singular vectors; continuous, not discrete (as in forward stepwise)

PCA

Scree Plot of (true) rank $= 2$



(p-values for right: 0.030, 0.064, 0.222, 0.286, 0.197, 0.831, 0.510, 0.185, 0.126.)

Ongoing work on selective inference

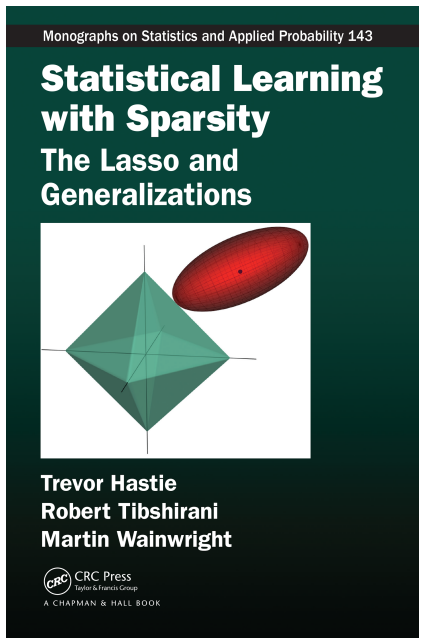
- Forward stepwise with grouped variables (Loftus and Taylor)
- Many means problem (Reid, Taylor, Tibs)
- Asymptotics (Tian and Taylor)
- Bootstrap (Ryan Tibshirani+friends)
- Internal inference— comparing internally derived biomarkers to external clinical factors– Gross, Taylor, Tibs
- **Still to come:** logistic regression, non-linear models

Conclusions

- Post-selection inference is an exciting new area. Lots of potential research problems and generalizations (grad students take note)!!
- Coming soon: Deep Selective Inference ®

Resources

- Google $\rightarrow$ Tibshirani
- R package on CRAN: **selectiveInference**. Forward stepwise regression, Lasso, Lars
- New book: Hastie, Tibshirani, Wainwright



Has a chapter on selective inference.