Recent Advances in Post-Selection Statistical Inference

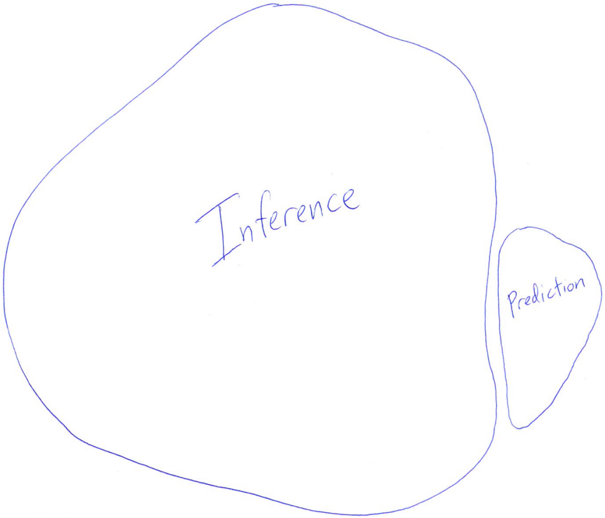
Robert Tibshirani, Stanford University

June 26, 2016

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

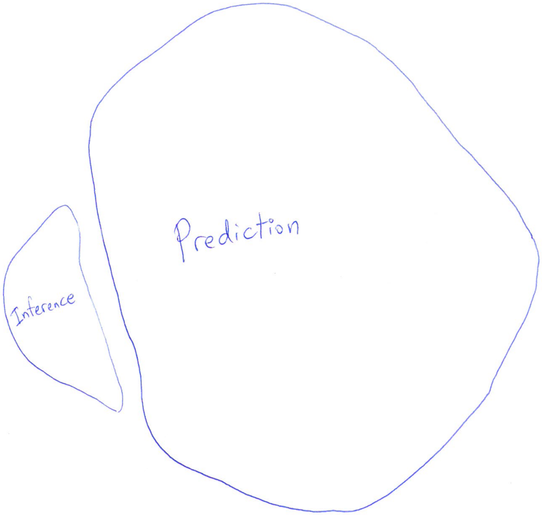
Thanks to Jon, Ryan & Will for help with the slides.

Statistics versus Machine Learning



How statisticians see the world?

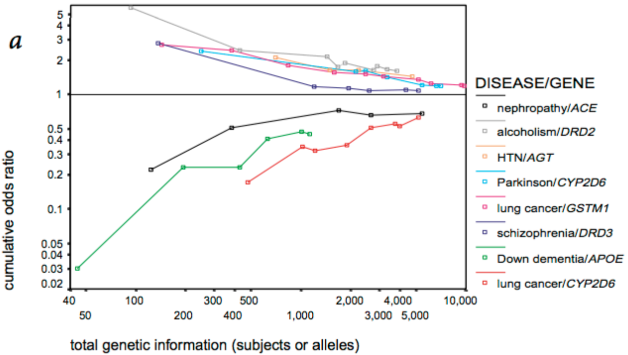
Statistics versus Machine Learning



How machine learners see the world?

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: John 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

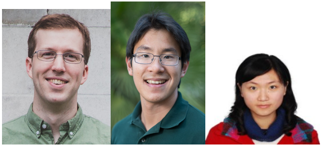
Ryan Tibshirani ,
CMU. PhD student of Taylor
2011

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

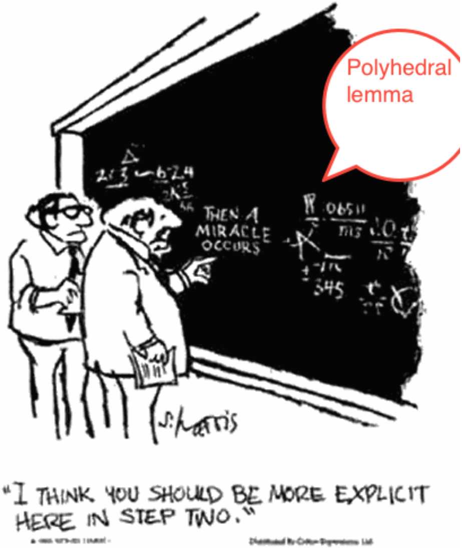
Now at CMU

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. Gaussian regression and generalized linear models— logistic regression, Cox model etc
3. When to stop Forward stepwise? FDR-controlling procedures using post-selection adjusted p-values
4. **New R package !!!!** `selectiveInference` **!!!!!**

NOT COVERED

1. Exponential family framework: more powerful procedures, requiring MCMC sampling
2. Data splitting, data carving, randomized response

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 $(x_i, y_i), i = 1, 2, \dots N$; $x_i = (x_{i1}, x_{I2}, \dots x_{ip})$.
- 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**
$$\operatorname{argmin}\Big\{\sum_i (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda \cdot \sum_j |\beta_j|\Big\}$$

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

Post selection inference

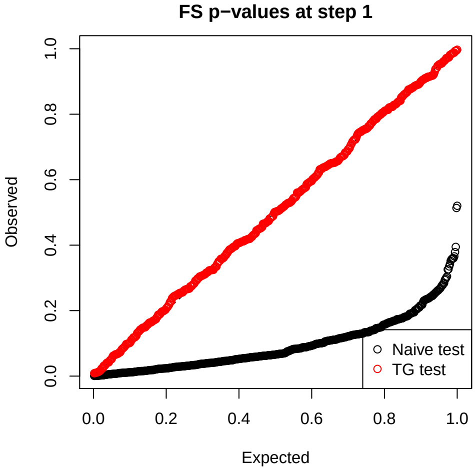
Example: Forward Stepwise regression

	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.

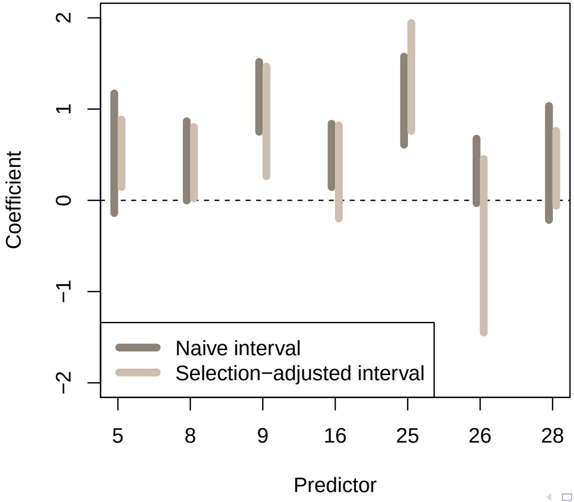
Simulation: $n = 100$, $p = 10$, and $y, X_1, \ldots X_p$ have i.i.d. $N(0,1)$ components



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* :

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

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

Some relevant literature

- Early work of Kiefer (1976, 1977), Brownie and Kiefer (1977), Brown (1978) is related in spirit, but very different focus
- False coverage-statement rate (FCR) control: Benjamini and Yekutieli (2005), Benjamini (2010), Rosenblatt and Benjamini (2014)
- Selective inference as multiple inference: Berk, Brown, Buja, Zhang, Zhao (2013) account for selection in regression over all possible procedures
- Extended by Bachoc, Leeb, Potscher (2014) to cover inference for predicted values
- Leeb and Potscher (2006, 2008) present impossibility results on estimating the conditional or unconditional distributions of post-selection estimators
- **Debiasing** approach has a different goal: Zhang, & Zhang, Van de Geer , Buhlmann, Ritov & Dezeure, Javanmard & Montanari et al, and Cai .

A key mathematical result

Polyhedral lemma: Provides a good solution for Forward Stepwise; 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)

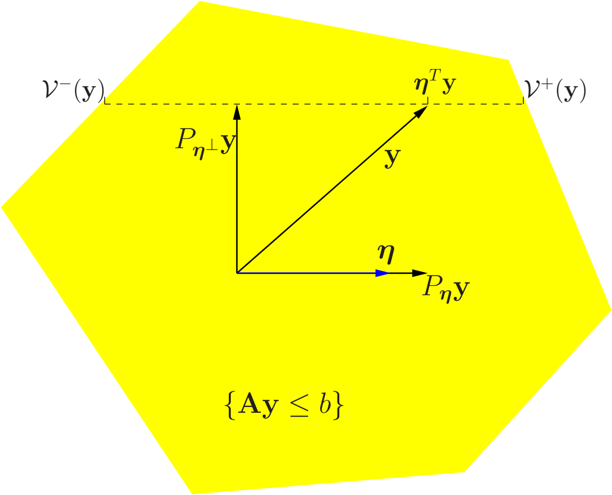
The polyhedral lemma

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

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

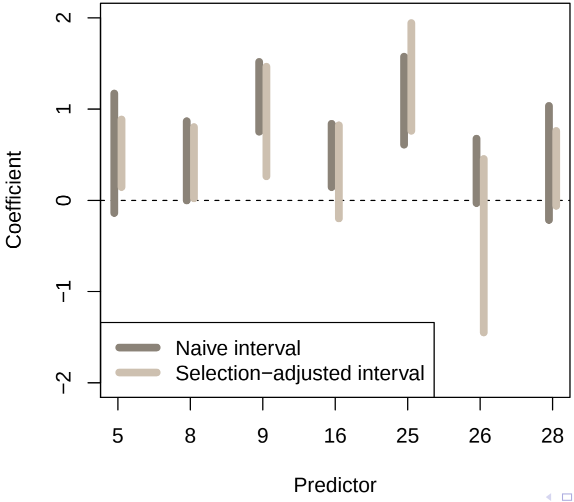
(truncated Gaussian distribution), where V^-, V^+ are (computable) values that are functions of η, A, b.

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


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

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)

Extension to Generalized Linear Models

Logistic regression, Cox Proportional hazards model, Graphical Lasso

- ℓ_1- penalized GLM, estimator $\hat{\beta}_M$ (selected model M). Define one-step estimator

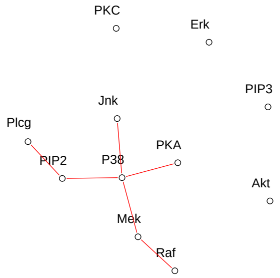
$$\bar{\beta}_M = \hat{\beta}_M + \lambda \cdot I_M(\hat{\beta}_M)^{-1} s_M \tag{1}$$

  where $I_M(\hat{\beta}_M)$ is the $|M| \times |M|$ observed Fisher information matrix of the submodel M evaluated at $\hat{\beta}_M$; s_M is sign vector of solution.
- Resulting constraints from KKT conditions:

$$\left\{ \text{diag}(s_M) \left[\bar{\beta}_M - I_M(\hat{\beta}_M)^{-1} \lambda s_M \right] \geq 0 \right\}, \tag{2}$$

- apply polyhedral lemma, to get post-selection, asymptotically valid p-values and selection intervals

Application to graphical lasso



Protein pair	P-values
Raf -Mek	0.789
Mek -P38	0.005
Plcg- PIP2	0.107
PIP2 -P38	0.070
PKA -P38	0.951
P38 -Jnk	0.557

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 JRSSB 2015

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:

G'Sell, Wager, Chouldchova, Tibs JRSSB 2015

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

Comparison to “Knockoffs” (Barber + Candes)

- In our experiments, the selective p-values yielded much higher power than knockoffs: but they control different notions of FDR.
- Knockoffs are a general procedure, applicable more broadly

R package

On CRAN: **selectiveInference**. Forward stepwise regression, Lasso, Lars, Logistic regression, Cox Model

```
gfit <-glmnet(x,y)
```
(or family="binomial" or "survival")

```
beta <- coef(gfit, s=lambda)
out <- fixedLassoInf(x,y,beta,lambda)
```

```
fsfit<- fs(x,y)
out <- fsInf(fsfit,x,y)
```

Ongoing work on selective inference

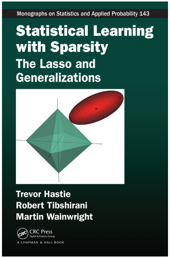
- Forward stepwise with grouped variables (Loftus and Taylor)
- Many means problem (Reid, Taylor, Tibs)
- Asymptotics (Tian and Taylor,)
- Asymptotics and Bootstrap (Ryan Tibshirani+friends)
- Internal inference— comparing internally derived biomarkers to external clinical factors– Gross, Taylor, Tibs
- data carving, randomized response

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 → Tibshirani
- New book: Hastie, Tibshirani, Wainwright



PDF free online. Has a chapter on selective inference.

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?

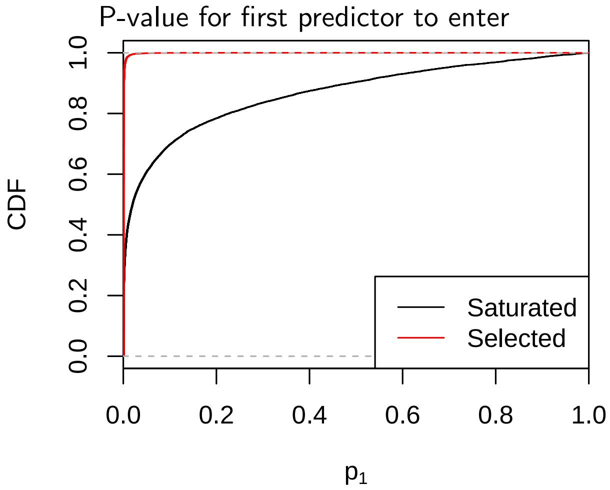
Exponential family framework

- 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
- We will return to these p-values when we discuss **stopping rules with FDR control**

Two signals of equal strength

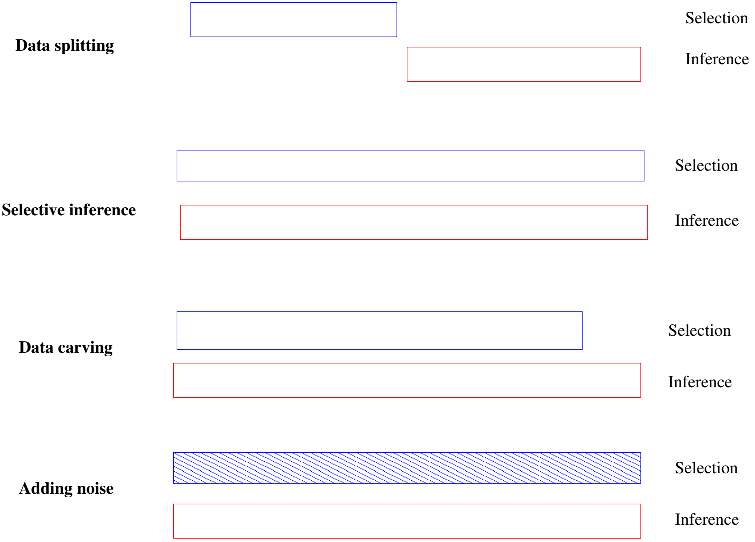


Data splitting, carving, and adding noise

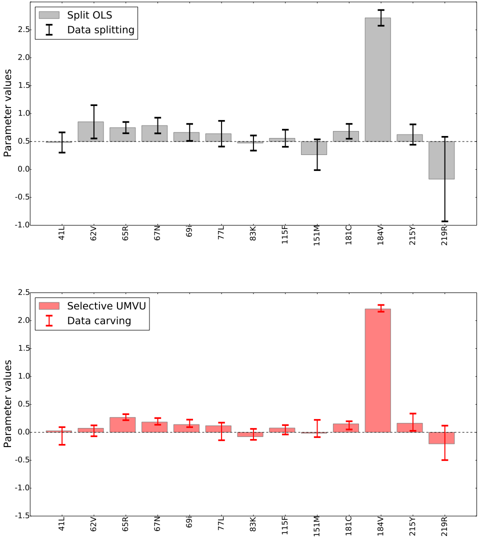
Further improvements in power

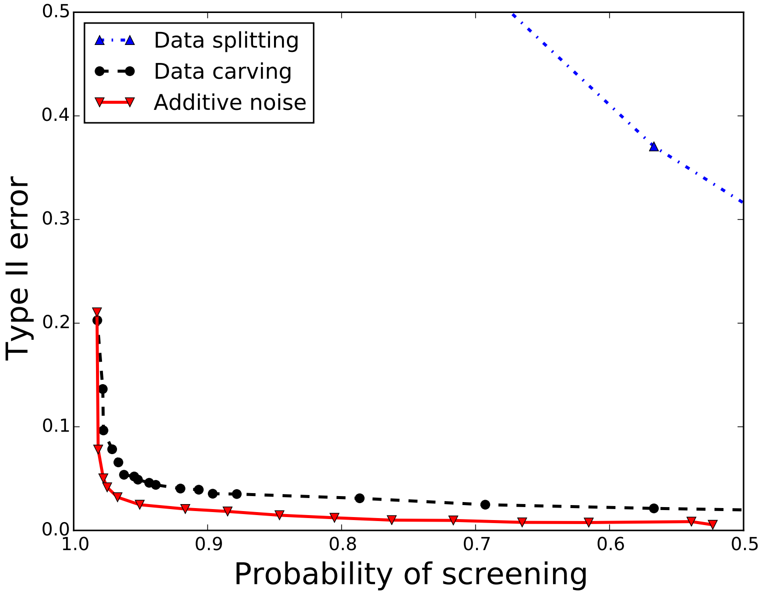
Fithian, Sun, Taylor, Tian

- 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
Probability of screening
0.5
0.4
0.3
0.2
0.1
0.0
1.0
0.9
0.8
0.7
0.6
0.5