Recent Advances in Selective Inference

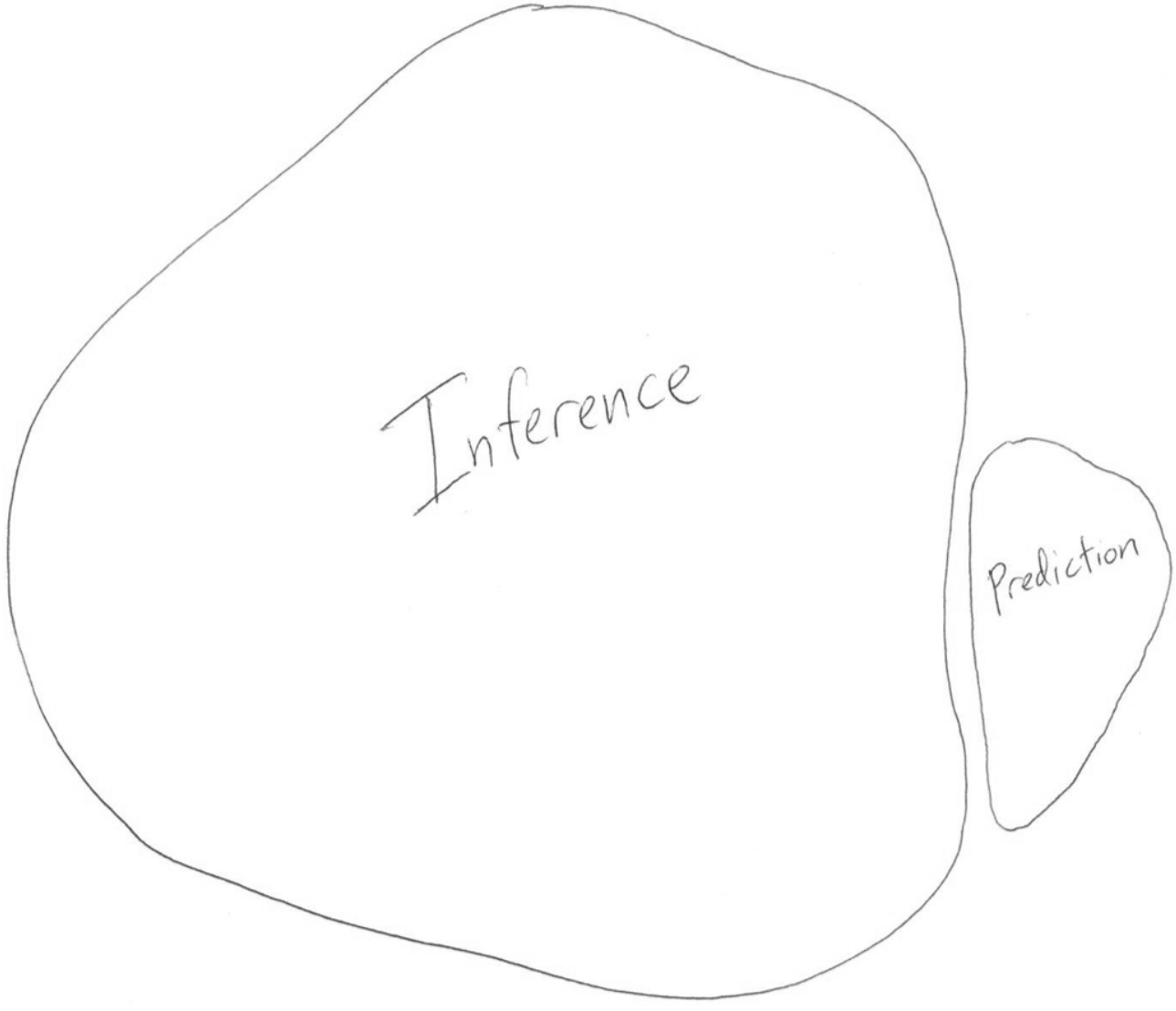
Ryan Tibshirani

Dept. of Statistics
Dept. of Machine Learning
Carnegie Mellon University

Collaborators:
Richard Lockhart, Jonathan Taylor, Rob Tibshirani,
Ale Rinaldo, Larry Wasserman

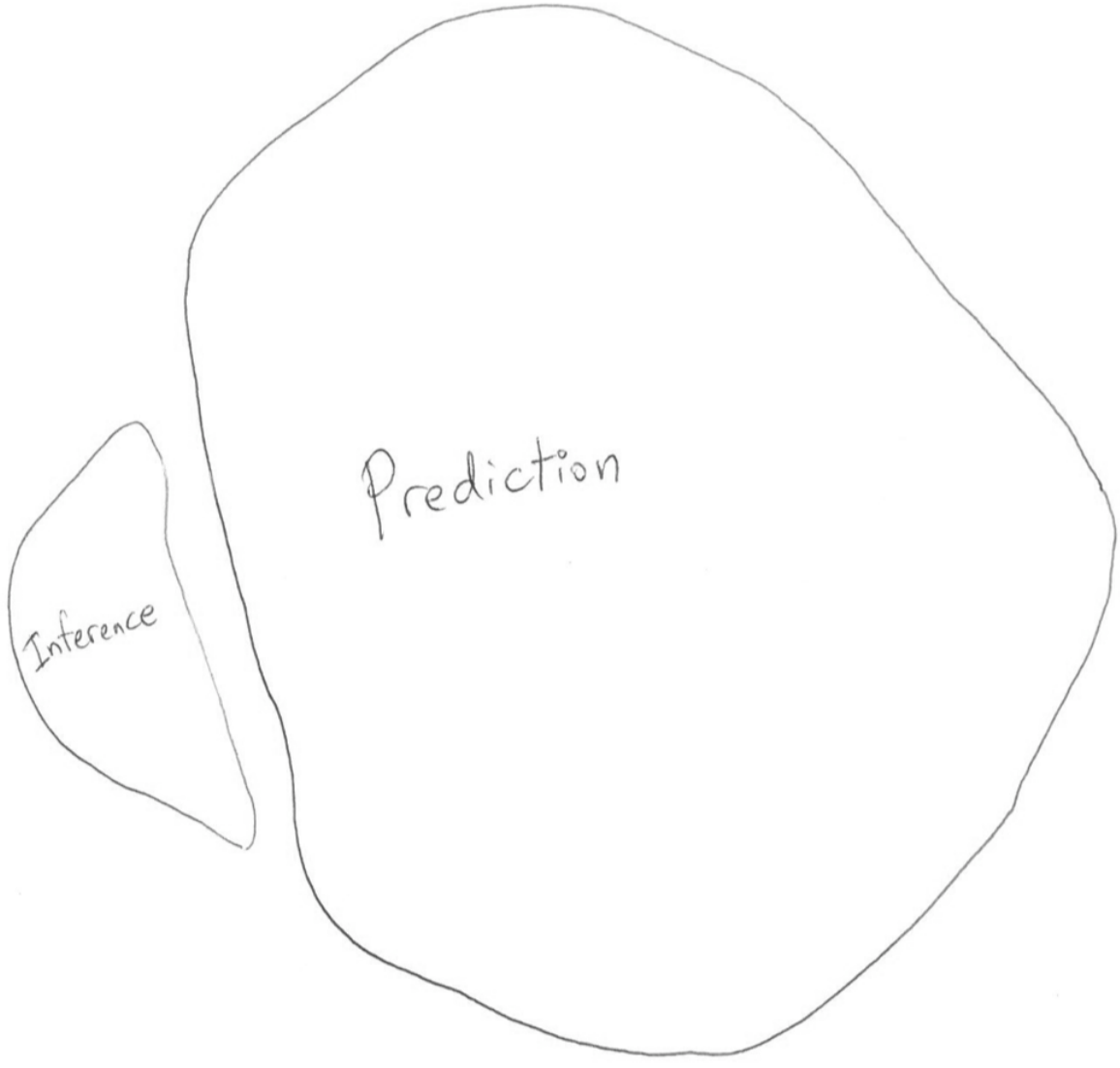
`http://www.stat.cmu.edu/~ryantibs/talks/inference-2016.pdf`

Statistics versus Machine Learning



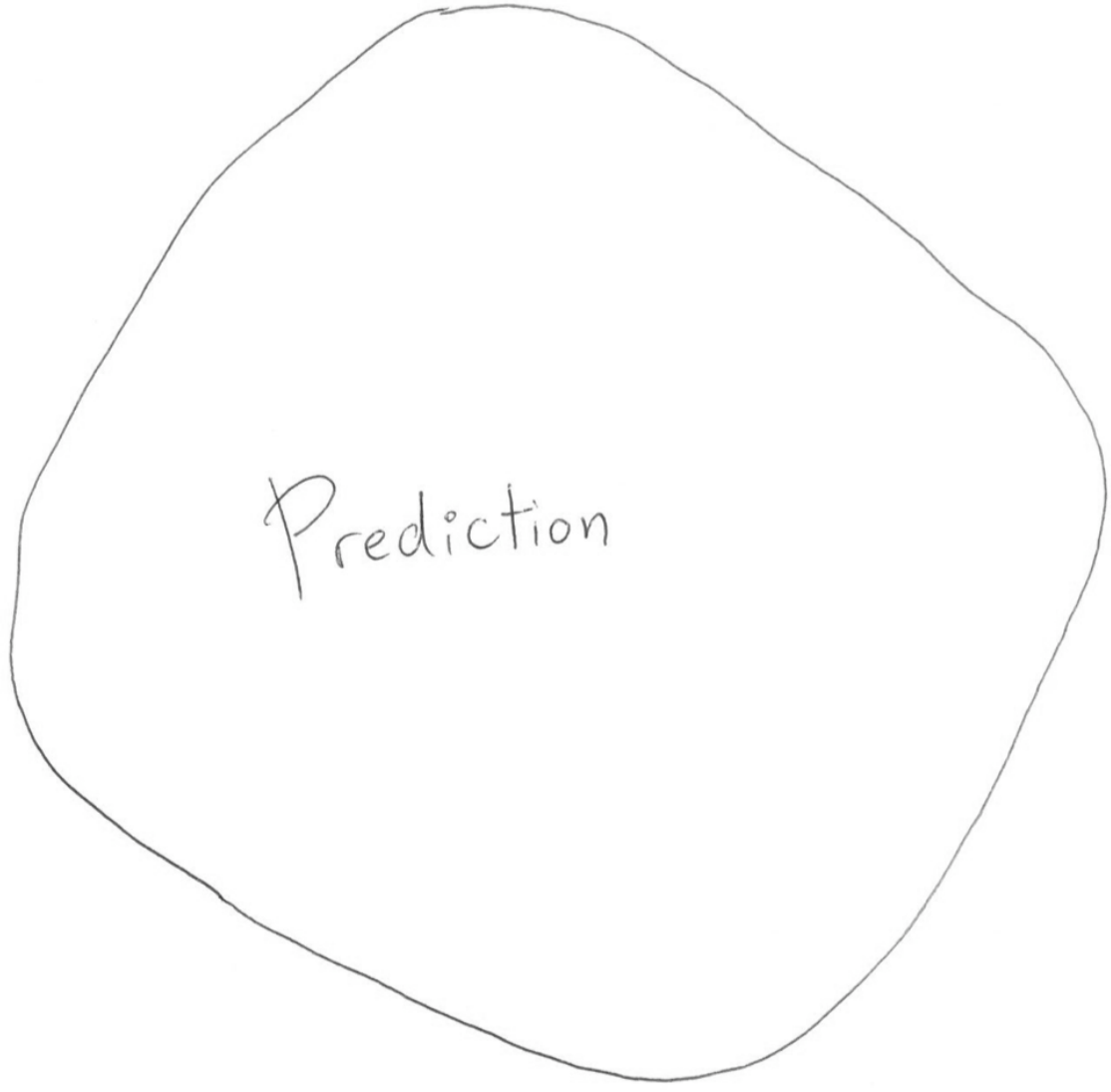
How statisticians see the world?

Statistics versus Machine Learning



How machine learners see the world?

Statistics versus Machine Learning



How machine learners see the world?

Outline

- Introduction to selective inference
- The polyhedral lemma
- Forward stepwise, LAR, and the lasso
- Sequential testing: a problem and fix
- Asymptotics: the good and the bad
- Challenges and future work

An introduction to selective inference

What is selective inference?

Statistician A:

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

Classical inference

Statistician B:

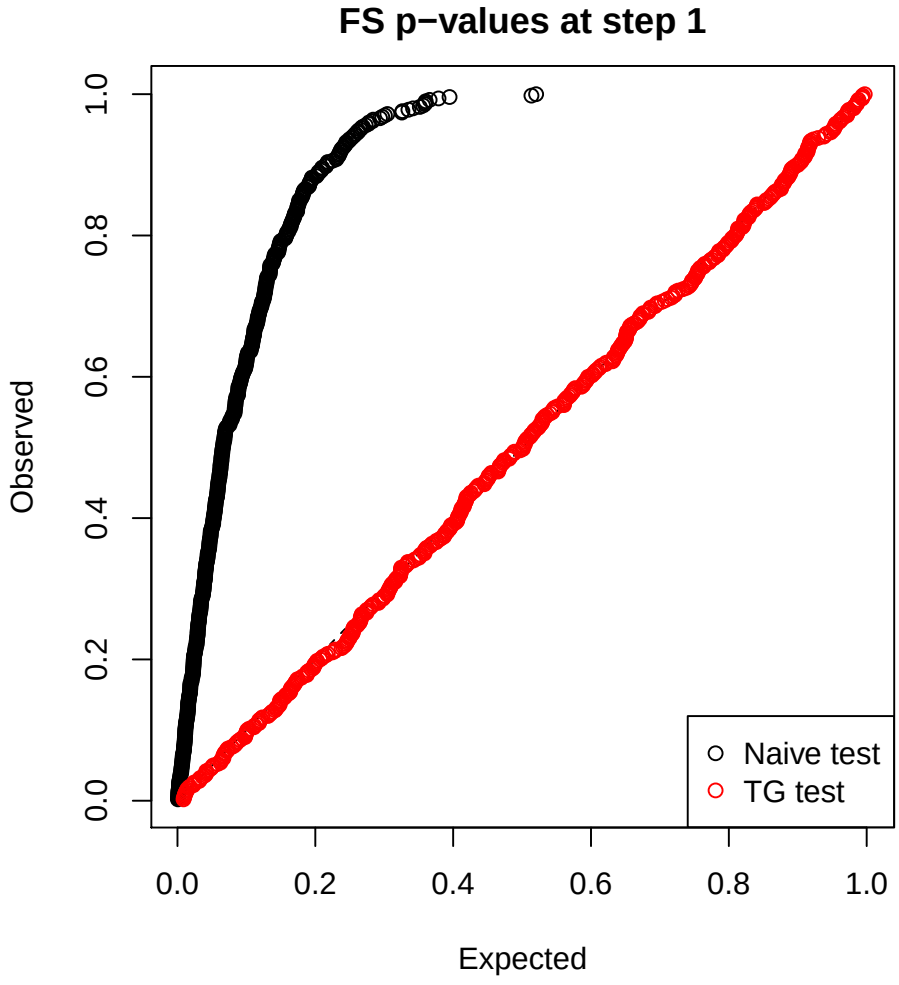
1. Collect data
2. Select a model
3. Test hypotheses

Selective 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)

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



Adaptive selection clearly makes χ^2_1 null distribution invalid ... new inference framework directly accounts for selection

Preview of results

Given $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ (columns $X_1, \ldots X_p \in \mathbb{R}^n$), with

$$y|X \sim N\big(\theta(X), \sigma^2 I\big)$$

We treat mean $\theta = \theta(X)$ as arbitrary, not assumed to be linear in X; variance σ^2 is known

Let $\hat{A}(y) = A$ denote a model selected by FS (or LAR, lasso, etc.) We will devise a statistic $T_j = T_j(y, A)$ such that

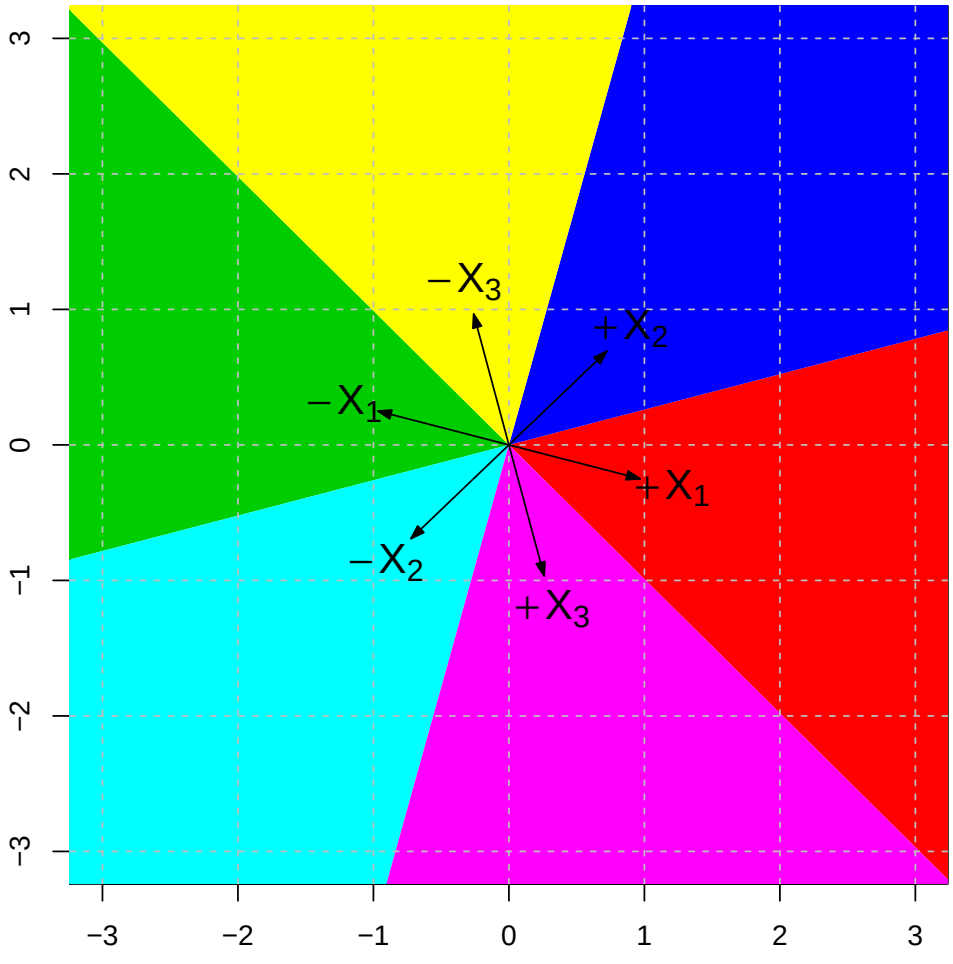
$$\mathbb{P}_{\beta_j=0}\Big(T_j \leq \alpha \;\Big|\; \hat{A}(y) = A\Big) = \alpha, \quad \text{for all } \alpha \in [0, 1]$$

Here, $j \in A$ is arbitrary, and $\beta_j = \beta_j(\theta, A) = [(X_A^T X_A)^{-1} X_A^T \theta]_j$

Inference is conditional on model selection event $\{y : \hat{A}(y) = A\}$. Together these events (collected over all A) form a partition of $\mathbb{R}^n$

Illustration: forward stepwise partition

Illustration: the model selection partition for FS, and $n = 2$, $p = 3$



Red region: $\hat{A}(y) = \{+1\}$
Blue region: $\hat{A}(y) = \{+2\}$
Pink region: $\hat{A}(y) = \{+3\}$
...

Important observation: partition elements are polyhedra

Some relevant literature

- Early work of Kiefer (1976, 1977), Brownie and Kiefer (1977), Brown (1978) is related in spirit, but very different focus
- Inference conditional on result of an F-test: Olshen (1973)
- 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

Some work from our group

- Post-selection inference for sequential regression procedures (covering FS, LAR, lasso): T., Taylor, Lockhart, Tibshirani (2014)
- Lasso at fixed λ value: Lee, Sun, Sun, Taylor (2014)
- Marginal screening: Lee and Taylor (2014)
- Many normal means: Reid, Taylor, Tibshirani (2014)
- Grouped stepwise regression: Loftus and Taylor (2014)
- Principal component analysis: Choi, Taylor, Tibshirani (2014)
- Changepoint detection, trend filtering, and graph clustering: Hyun, G'Sell, T. (2016+)

Common thread: condition on selection event, nuisance parameter; then exploit conditional distributions (exactly, or via sampling)

Even more

- Asymptotics of selective inference in low and high dimensions, under weak assumptions: T., Rinaldo, Tibshirani, Wasserman (2015)
- Asymptotics in high dimensions, under stronger assumptions: Tian and Taylor (2015)
- Theory of selective inference in exponential families: Fithian, Sun, Taylor (2015). Describes UMP selective tests
- Selective inference with a randomized response: Tian and Taylor (2015)
- Independence, and better power, in selective sequential tests: Fithian, Taylor, T., Tibshirani (2015)

This talk: polyhedral lemma, sequential regression procedures, and asymptotics (good and bad)

The polyhedral lemma

Basic setup with polyhedral constraints

Assume that $y \sim N(\theta, \Sigma)$, with mean parameter θ unknown (but covariance Σ known)

We wish to make inferences on $v^T\theta$—a linear contrast of the mean θ—conditional on $y \in \mathcal{P}$,

$$\mathcal{P} = \{y : \Gamma y \geq u\}$$

E.g., we want a test statistic (p-value), denoted by $T = T(y, v, \mathcal{P})$, with the property

$$\mathbb{P}_{v^T\theta=0}\Big(T \leq \alpha \;\Big|\; \Gamma y \geq u\Big) = \alpha, \quad \text{for all } \alpha \in [0, 1]$$

The polyhedral lemma

Lemma (Polyhedral selection as truncation)

For any v such that $v^T \Sigma v \neq 0$, and any y,

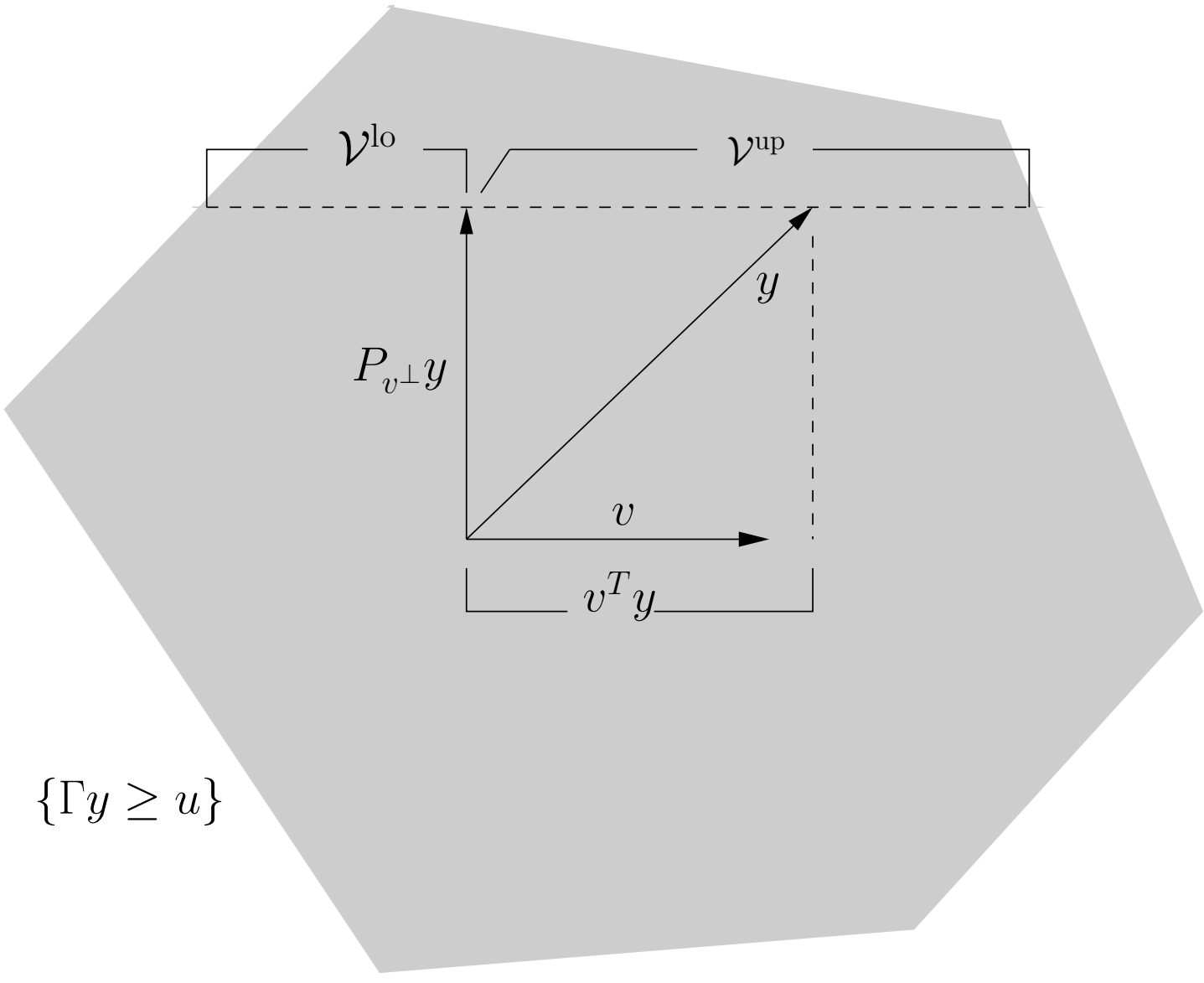
$$\Gamma y \geq u \iff \mathcal{V}^{\mathrm{lo}}(y) \leq v^T y \leq \mathcal{V}^{\mathrm{up}}(y),\ \mathcal{V}^0(y) \leq 0,$$

where $\rho = \Gamma \Sigma v / v^T \Sigma v$, and

$$\begin{aligned}
\mathcal{V}^{\mathrm{lo}}(y) &= \max_{j:\rho_j > 0} \frac{u_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j}, \\
\mathcal{V}^{\mathrm{up}}(y) &= \min_{j:\rho_j < 0} \frac{u_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j}, \\
\mathcal{V}^{0}(y) &= \max_{j:\rho_j = 0} u_j - (\Gamma y)_j.
\end{aligned}$$

If $y \sim N(\theta, \Sigma)$, then $(\mathcal{V}^-, \mathcal{V}^+, \mathcal{V}^0)(y)$ is independent of $v^T y$.

Proof


$\mathcal{V}^{\mathrm{lo}}$
$\mathcal{V}^{\mathrm{up}}$
y
$P_{v^\perp} y$
v
$v^T y$
$\{\Gamma y \geq u\}$

Inference from the polyhedral lemma

We now study $v^T y \,\big|\, \mathcal{V}^{\rm lo} \le v^T y \le \mathcal{V}^{\rm up},\ \mathcal{V}^0 \ge 0$. This is a truncated Gaussian distribution with random limits

How to use for conditional inference on $v^T\theta$? Follows two ideas:

1. For $Z \sim N^{[a,b]}(\mu, \sigma^2)$, and $F^{[a,b]}_{\mu,\sigma^2}$ its CDF, we have

$$\mathbb{P}\Big(F^{[a,b]}_{\mu,\sigma^2}(Z) \le \alpha\Big) = \alpha$$

2. Hence by conditioning on $P_{v^\perp} y$,

$$\mathbb{P}\Big(F^{[\mathcal{V}^{\rm lo},\mathcal{V}^{\rm up}]}_{v^T\theta, v^T\Sigma v}(v^T y) \le \alpha \,\Big|\, \Gamma y \ge u,\ P_{v^\perp} y\Big) = \alpha$$

and then the same if true after integrating over $P_{v^\perp y}$

Therefore to test $H_0 : v^T\theta = 0$ against $H_1 : v^T\theta > 0$, we can take as a conditional p-value $T = 1 - F_{0,v^T\Sigma v}^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}(v^T y)$, since

$$\mathbb{P}_{v^T\theta=0}\left(1 - F_{0,v^T\Sigma v}^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}(v^T y) \leq \alpha \;\middle|\; \Gamma y \geq u\right) = \alpha$$

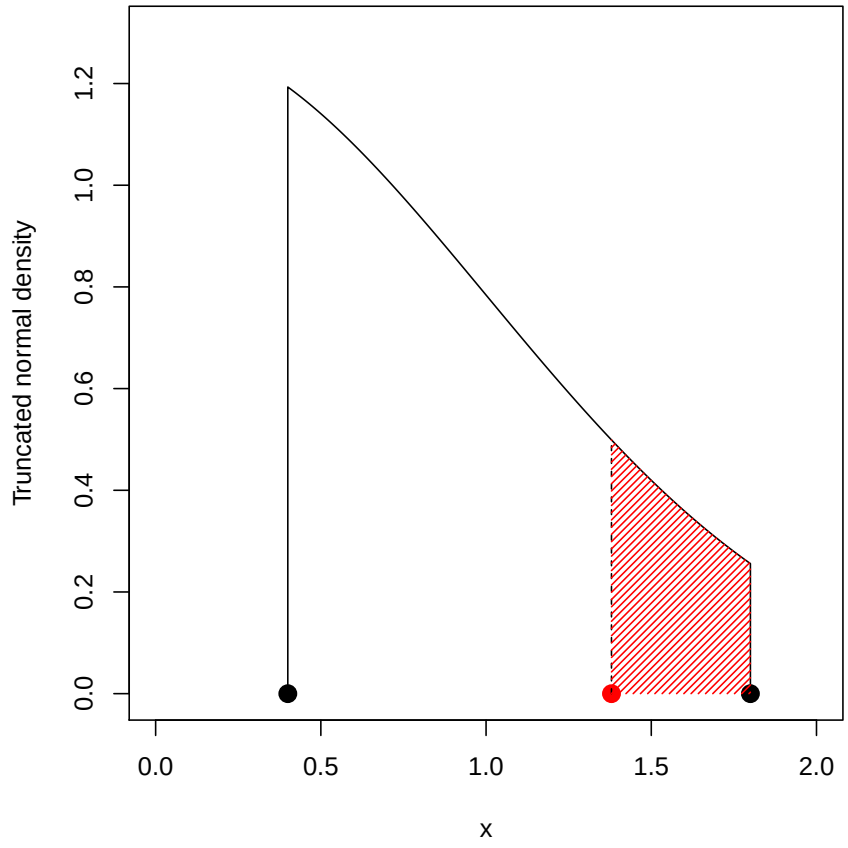
Furthermore, because the same statement holds for any fixed μ,

$$\mathbb{P}_{v^T\theta=\mu}\left(1 - F_{\mu,v^T\Sigma v}^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}(v^T y) \leq \alpha \;\middle|\; \Gamma y \geq u\right) = \alpha$$

we can compute a conditional confidence interval by inverting the pivot, yielding

$$\mathbb{P}\left(v^T\theta \in [\delta_{\alpha/2}, \delta_{1-\alpha/2}] \;\middle|\; \Gamma y \geq u\right) = 1 - \alpha$$

Simple illustration: random realizations



We observe random limits $\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}$ (black), and random point $v^T y$ (red); we form truncated normal density $f^{[\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}]}_{0, v^T \Sigma v}$, and the mass to the right of $v^T y$ is our p-value

Forward stepwise, LAR, and the lasso

How is this related to regression?

It turns out that for many sequential regression procedures:

$$\Big\{y : \text{Procedure selects model } M\Big\} = \{y : \Gamma y \geq u\}$$

for some Γ, u (depending on M). We consider the following three:

- Forward stepwise (FS): adds variables per the maximal drop in RSS (or, maximal absolute correlation with residual)
- Least angle regression (LAR): adds variables to maintain equal absolute correlation with residual among active set
- Lasso: sequence of estimates defined by solutions of

$$\min_{\beta \in \mathbb{R}^p} \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

  over tuning parameter $\lambda \geq 0$. Can be viewed as modification of LAR (deactivating variables if their coefficients hit zero)

Polyhedral sets for FS, LAR, lasso

Suppose that we run k steps of FS, LAR, or lasso, and encounter the active list A_k. Can express

$$\left\{y : \left(\hat{A}_k(y), \hat{s}_{A_k}(y)\right) = \left(A_k, s_{A_k}\right)\right\} = \{y : \Gamma y \geq 0\}$$

for a matrix Γ. Here s_{A_k} is list of active signs. Describes vectors y such that the algorithm would make the same selections (variables and signs) over k steps

E.g., for FS at first step, variable 5 is chosen with positive sign iff

$$\frac{X_5^T y}{\|X_5\|_2} \geq \pm \frac{X_j^T y}{\|X_j\|_2} \quad \text{for all } j \neq 5$$

and $2(p-1)$ rows of Γ are defined accordingly

E.g., for FS at kth step, variable j_k is chosen with a sign s_k iff

$$\frac{s_k X_{j_k}^T P_{k-1}^{\perp} y}{\|P_{k-1}^{\perp} X_{j_k}\|_2} \geq \pm \frac{X_j^T P_{k-1}^{\perp} y}{\|P_{k-1}^{\perp} X_j\|_2} \quad \text{for all } j \notin A_k$$

and $2(p-k)$ rows are appended to Γ accordingly

- Exhaustive representation for FS at step k produces Γ with about $3pk$ rows
- Similar representation for LAR, lasso at step k gives Γ with about $3pk$ rows
- LAR, special case: only $k+1$ rows for Γ!

Note: computation of our test statistic is $O(\text{number of rows of } \Gamma)$

Selective tests for FS, LAR, lasso

Assume that $y \sim N(\theta, \sigma^2 I)$, and X is fixed. No assumptions on X, except general position

Choose any contrast vector v, e.g., $v = X_{A_k}(X_{A_k}^T X_{A_k})^{-1} e_j$, so

$$v^T \theta = \left[(X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T \theta\right]_j$$

Interpretation: project θ onto variables in active set A_k, then read off partial regression coefficient of jth variable. Form Γ, compute

$$\mathcal{V}^{\text{lo}} = \max_{j:(\Gamma v)_j > 0} -\frac{(\Gamma P_{v^\perp} y)_j}{(\Gamma v)_j} \cdot \|v\|_2^2$$

$$\mathcal{V}^{\text{up}} = \min_{j:(\Gamma v)_j < 0} -\frac{(\Gamma P_{v^\perp} y)_j}{(\Gamma v)_j} \cdot \|v\|_2^2$$

The conditional p-value for $H_0 : v^T\theta = 0$ is

$$T = \frac{\Phi\left(\frac{\mathcal{V}^{\rm up}}{\sigma\|v\|_2}\right) - \Phi\left(\frac{v^T y}{\sigma\|v\|_2}\right)}{\Phi\left(\frac{\mathcal{V}^{\rm up}}{\sigma\|v\|_2}\right) - \Phi\left(\frac{\mathcal{V}^{\rm lo}}{\sigma\|v\|_2}\right)}$$

This has exact conditional size:

$$\mathbb{P}_{v^T\theta=0}\Big(T \leq \alpha \;\Big|\; \big(\hat{A}_k(y), \hat{s}_{A_k}(y)\big) = (A_k, s_{A_k})\Big) = \alpha$$

When $v = X_{A_k}(X_{A_k}^T X_{A_k})^{-1} e_k$—this tests the significance of last variable entered—and we are using LAR, test takes a very special form:

$$T = \frac{\Phi\left(\lambda_{k-1}\frac{\omega_k}{\sigma}\right) - \Phi\left(\lambda_k\frac{\omega_k}{\sigma}\right)}{\Phi\left(\lambda_{k-1}\frac{\omega_k}{\sigma}\right) - \Phi\left(\lambda_{k+1}\frac{\omega_k}{\sigma}\right)}$$

Here $\lambda_{k-1}, \lambda_k, \lambda_{k+1}$ are knots at steps $k-1, k, k+1$. Called the spacing test, measures the spacings between knots λ_k, λ_{k+1}

Example: prostate cancer data

Data from a study of $n = 67$ men with prostate cancer, measuring log PSA versus $p = 8$ clinical measures

	FS, naive	FS, TG		LAR, cov	LAR, spacing
lcavol	0.000	0.000	lcavol	0.000	0.000
lweight	0.000	0.027	lweight	0.047	0.052
svi	0.019	0.184	svi	0.170	0.137
lbph	0.021	0.172	lbph	0.930	0.918
pgg45	0.113	0.453	pgg45	0.352	0.016
lcp	0.041	0.703	age	0.653	0.586
age	0.070	0.144	lcp	0.046	0.060
gleason	0.442	0.800	gleason	0.979	0.858

Testing variables as they enter. Naive: usual t-tests; TG: truncated Gaussian tests; cov: covariance test (Lockhart, Taylor, T., Tibshirani 2014). Interesting fact: last two are asymptotically equivalent

Confidence intervals for FS, LAR, lasso

Can also compute conditional confidence interval $[\delta_{\alpha/2}, \delta_{1-\alpha/2}]$ by inverting the truncated Gaussian pivot. This has exact conditional coverage:

$$\mathbb{P}\Big(v^T\theta \in [\delta_{\alpha/2}, \delta_{1-\alpha/2}] \;\Big|\; \big(\hat{A}_k(y), \hat{s}_{A_k}(y)\big) = (A_k, s_{A_k})\Big) = 1-\alpha$$

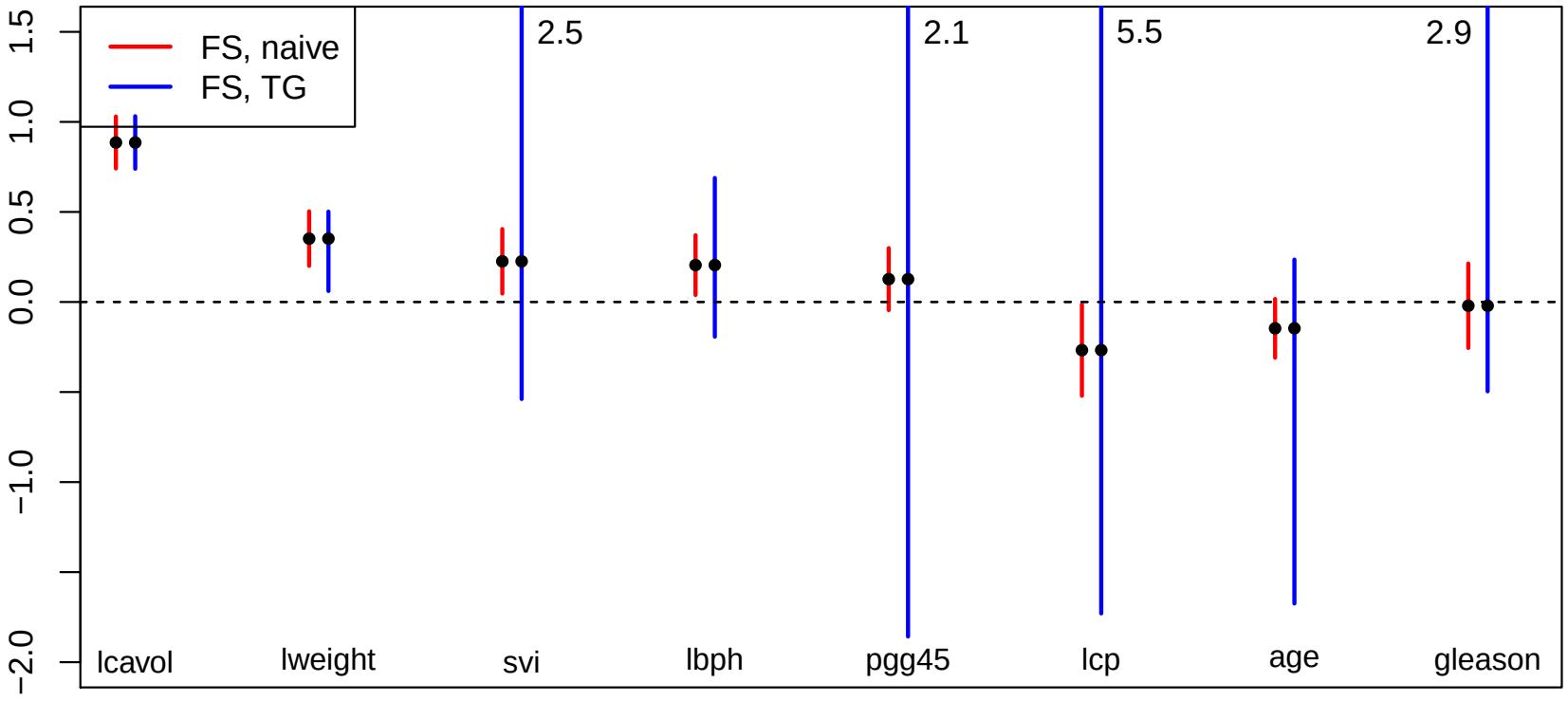
When $v = X_{A_k}(X_{A_k}^T X_{A_k})^{-1} e_k$: this traps the partial regression coefficient of last variable entered with prob $1-\alpha$, conditional on selections made so far

Can marginalize above to yield unconditional interpretation:

$$\mathbb{P}\Big(e_k^T (X_{\hat{A}_k}^T X_{\hat{A}_k})^{-1} X_{\hat{A}_k}^T \theta \in [\delta_{\alpha/2}, \delta_{1-\alpha/2}]\Big) = 1-\alpha$$

(Unlike a typical confidence interval, this tracks a moving target)

Example: back to prostate cancer data



Sequential testing: a problem and fix

A problem with testing variables as they enter

Sequentially testing variables as they enter seems like an appealing use case of proposed tools

But there is a problem with this method.[1] It helps to be concrete:

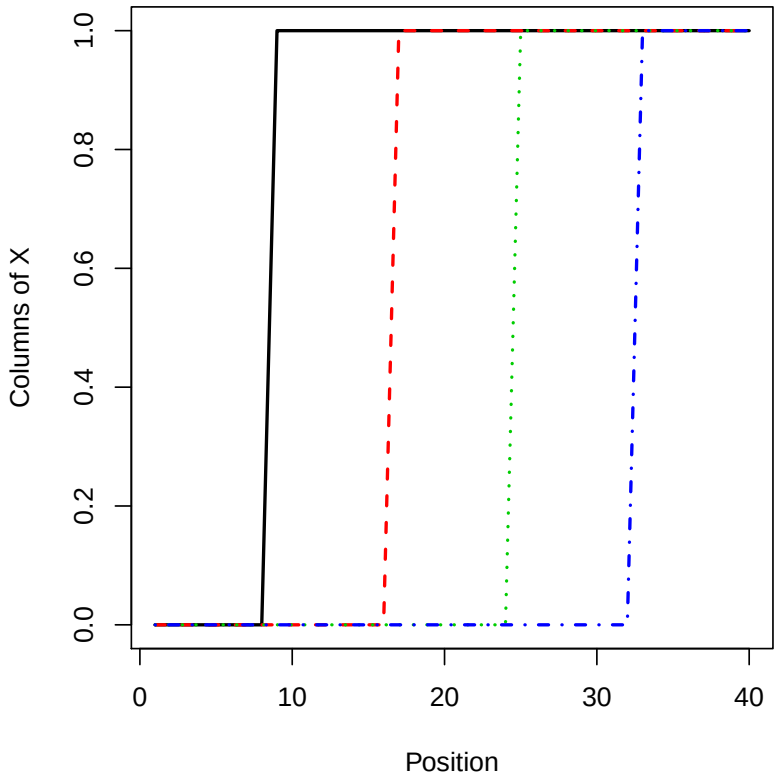
- Suppose we select variables 1,2,3 in this order
- Further suppose that θ lies in span of $X_{1:3}$
- Testing $X_1^T\theta = 0$ may not be meaningful
- Testing $[(X_{1:3}^T X_{1:3})^{-1} X_{1:3}^T \theta]_1 = 0$ is meaningful
- Thus TG test at first step may fail to reject

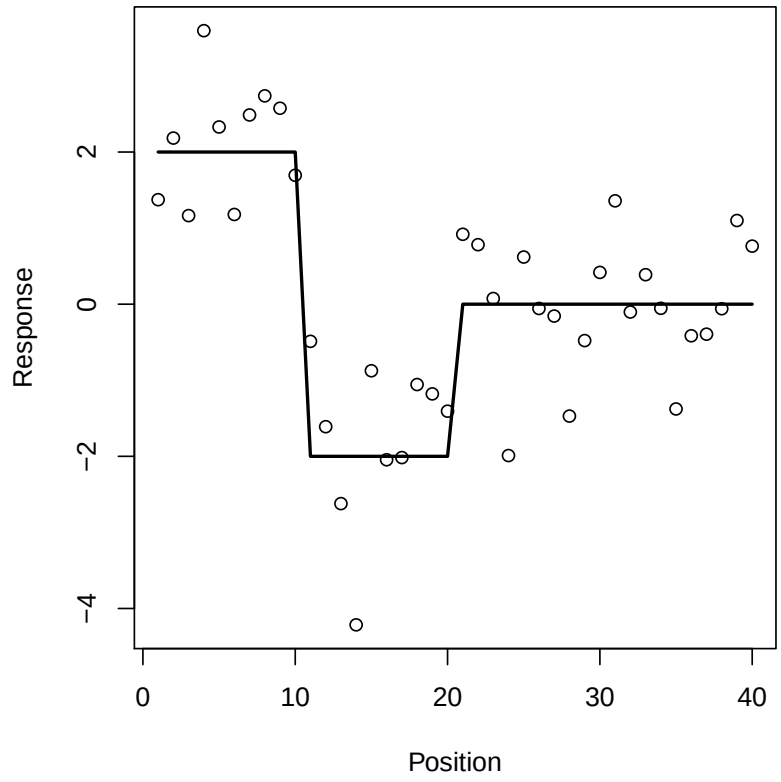
This problem can happen (not too infrequently) in practice!

[1] Many thanks to Will Fithian and Larry Brown for discussions on this.

Example: step functions

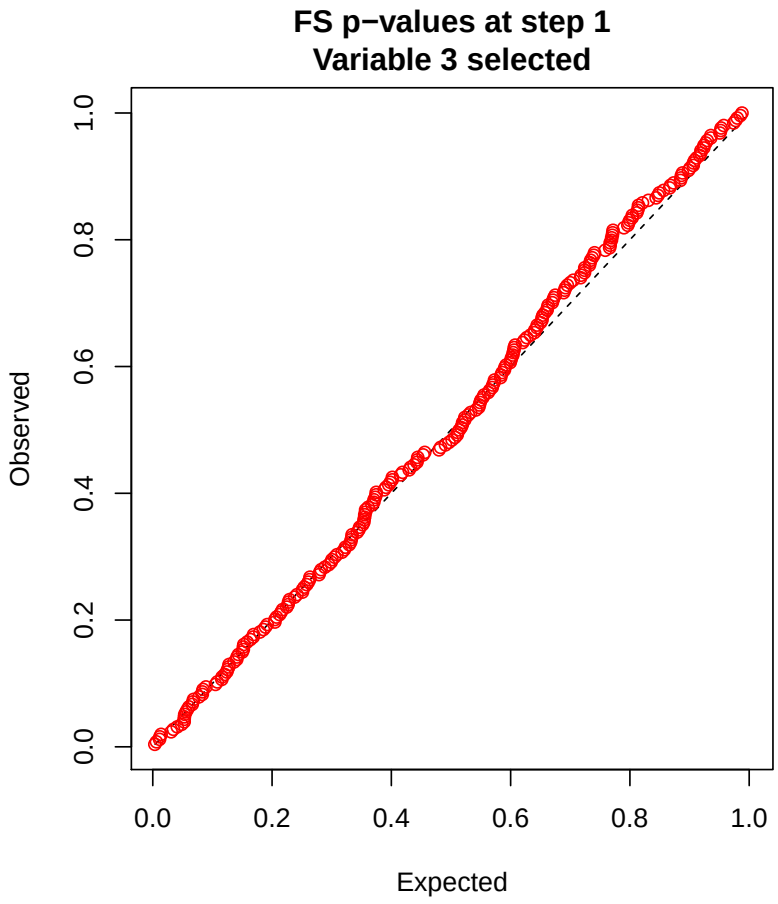
Simulation: $n = 40$, $p = 4$, and $\theta, X_1, \ldots X_p$ following structured setup. Response y generated with i.i.d. $N(0, 1)$ errors around θ

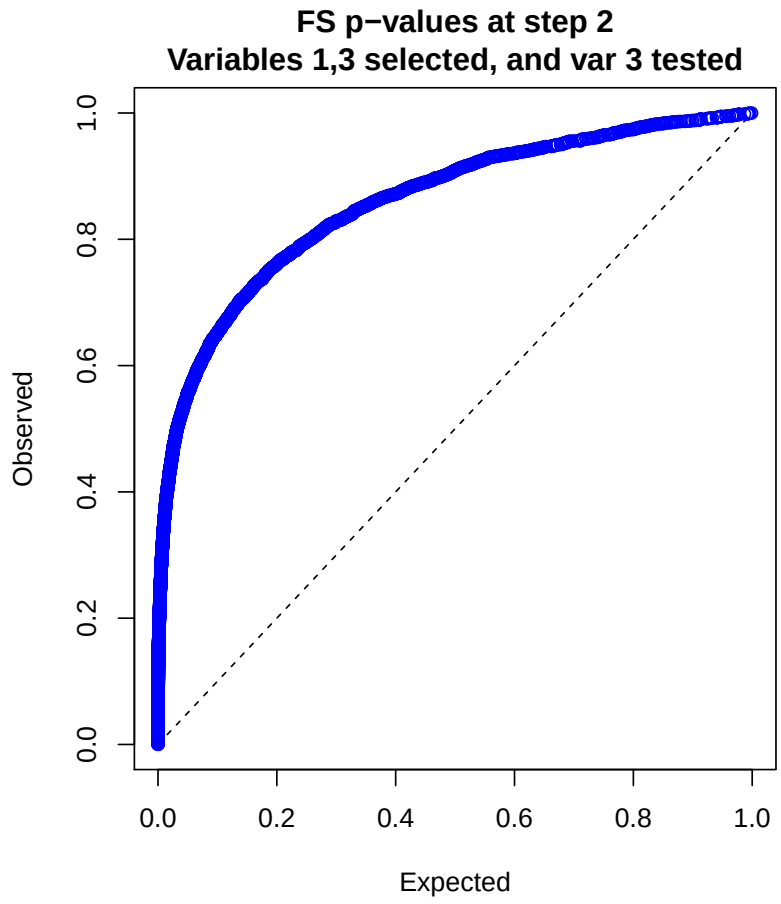




Notice that $\theta = -X_1 + X_3$, but $X_3^T \theta = 0$, exactly!

P-values for the significance of variable X_3, computed two ways:





Testing for the significance of X_3 at first step completely fails, but testing for it at second step clearly succeeds

A better way to use the proposed tools?

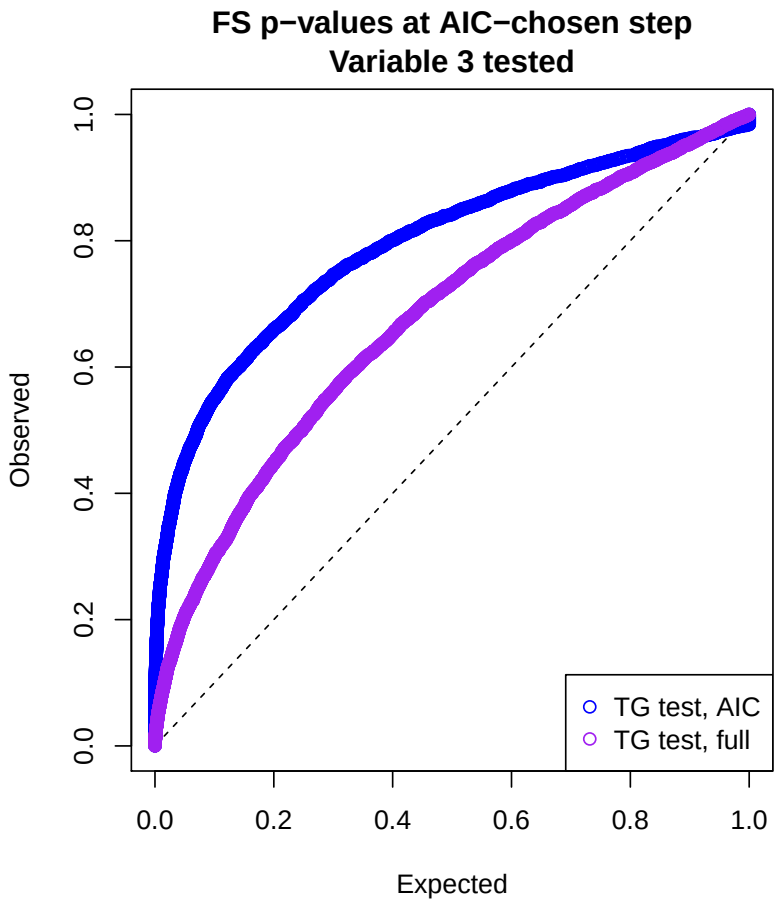
We know that the notion variable of importance hinges critically on other variables in working active model. Better (?) way to use the proposed inference tools:

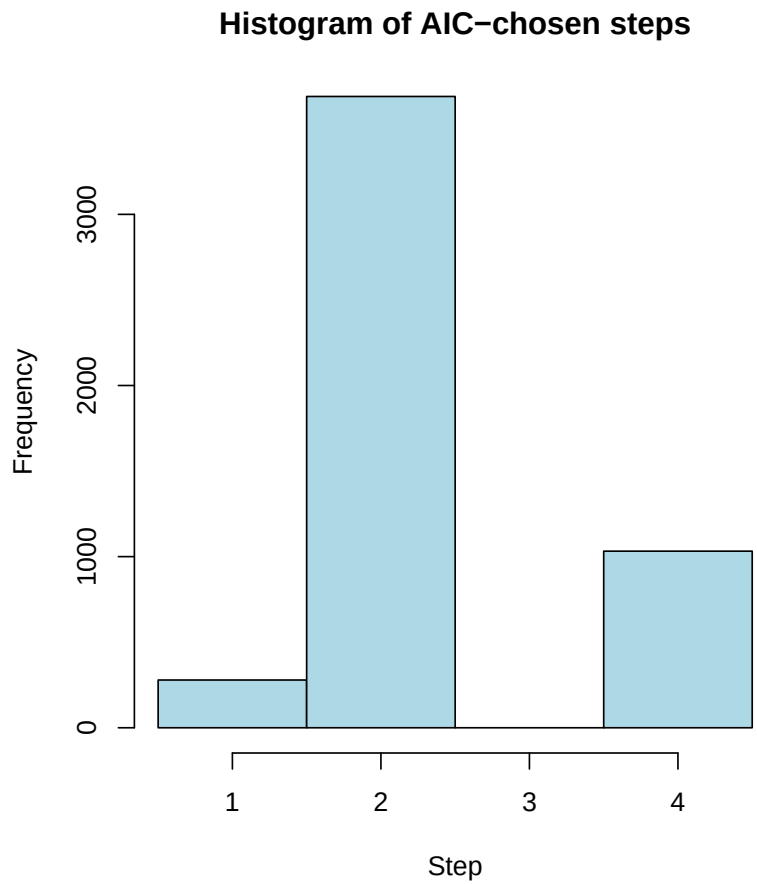
- Run the complete FS (or LAR, or lasso) path
- Select step $\hat{k}(y) = k$ based on a stopping rule like AIC or BIC
- Hope that the FS (or LAR, or lasso) active model A_k contains truly relevant variables
- Run tests on all active variables: $[(X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T \theta]_j = 0$, for all $j \in A_k$ (use Bonferonni correction for multiplicity)

To properly account for all selections that were made, the TG test statistics must be constructed conditional not only on $\hat{A}_k(y) = A_k$ (and $\hat{s}_{A_k}(y) = s_{A_k}$) but also $\hat{k}(y) = k$. Fortunately, $\{y : \hat{k}(y) = k\}$ is itself polyhedral when stopping rule like AIC or BIC is used

Example: back to step functions

With AIC-like stopping rule[2], FS path can be stopped automatically, and variable 3 tested appropriately

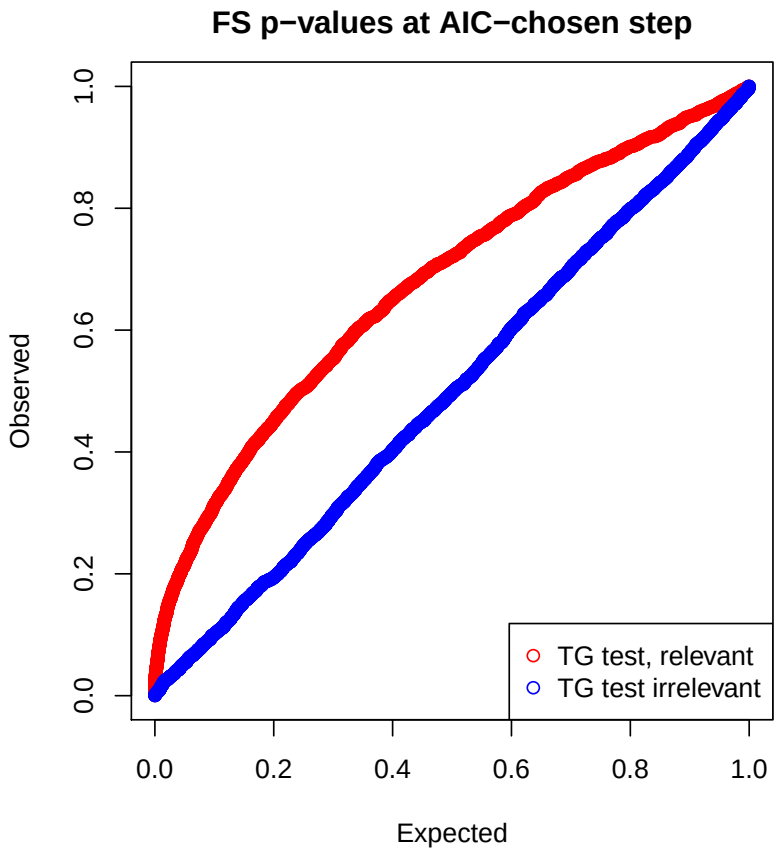


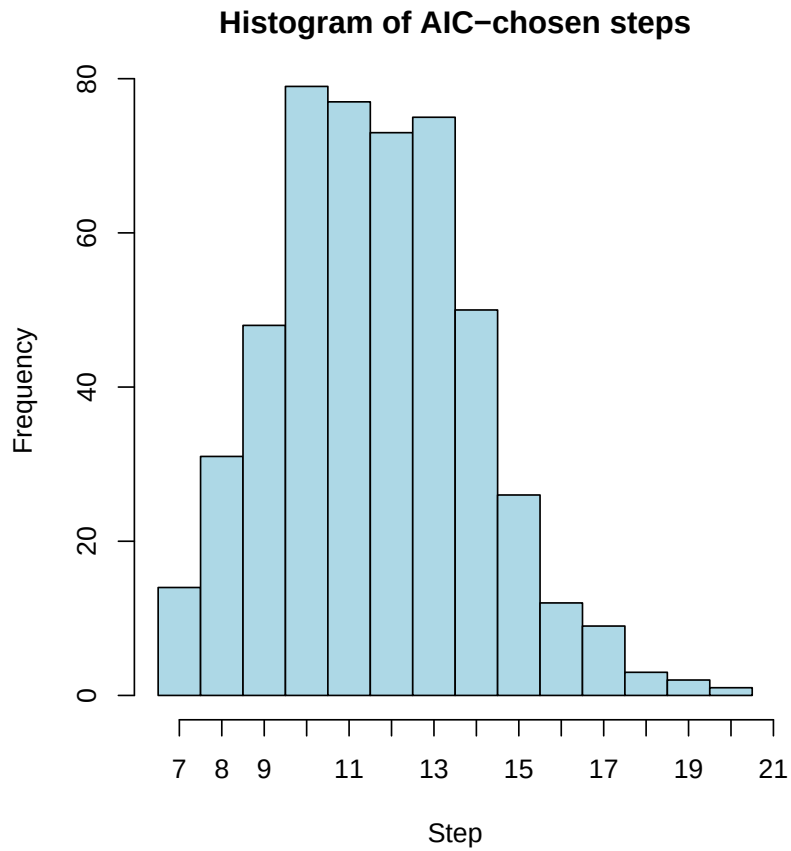


[2]We stop when AIC rises ones (rather than stopping at the AIC minimizer).

Example: AIC-stopped regression

Simulation: $n = 50$, $p = 100$, and $\epsilon, X_1, \ldots X_p$ with i.i.d. $N(0, 1)$ components. Mean $\theta = \sum_{j=1}^{5} 3X_j$, response $y = \theta + \epsilon$. Consider AIC-like stopping rule, again[3]





[3]We stop when AIC rises twice (rather than stopping at the AIC minimizer).

An alternative approach

In Fithian, Sun, Taylor (2014), and Fithian, Taylor, T., Tibshirani (2015), an important alternative perspective is described

Applies nicely to the sequential problem. At the kth step, instead of

$$H_0 : \left[(X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T \theta\right]_k = 0$$

we test

$$H_0 : \theta \in \text{span}(X_{A_k})$$

(Former called "saturated" null, latter called "selected" null.) New tests are much harder to compute: require sampling. But:

- Yields independent p-values across steps (FDR control, from G'Sell et al. 2014)
- Generally gives much better power in early steps, when relevant variables are left out of active set

Asymptotics: the good and the bad

How important is normality?

Recall that the key result of the polyhedral lemma

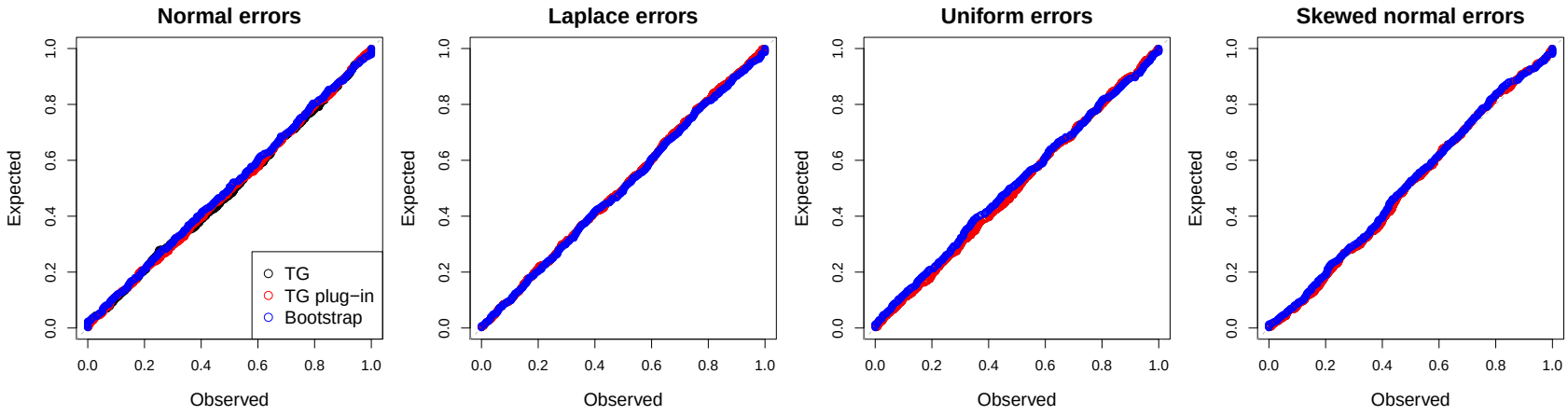
$$\Gamma y \geq u \iff \mathcal{V}^{\mathrm{lo}}(y) \leq v^T y \leq \mathcal{V}^{\mathrm{up}}(y),\ \mathcal{V}^0(y) \leq 0$$

holds deterministically; but the quantities $\mathcal{V}^{\mathrm{lo}}(y), \mathcal{V}^{\mathrm{up}}(y), \mathcal{V}^0(y)$ are functions of $P_{v^\perp} y$, and are independent of $v^T y$ only if y is Gaussian

- In practice, it is hard to break the TG tests and intervals
- As n grows, one might think $v^T y$ and $\mathcal{V}^{\mathrm{lo}}(y), \mathcal{V}^{\mathrm{up}}(y), \mathcal{V}^0(y)$ become closer to independent, because they are functions of uncorrelated quantities
- This intuition holds true when p is fixed, but not necessarily when p grows

Examples: robustness of the TG test

Simulation: $n = 50$, $p = 10$, truly null model, first step of FS



Simulation: $n = 50$, $p = 1000$, truly null model, first step of FS

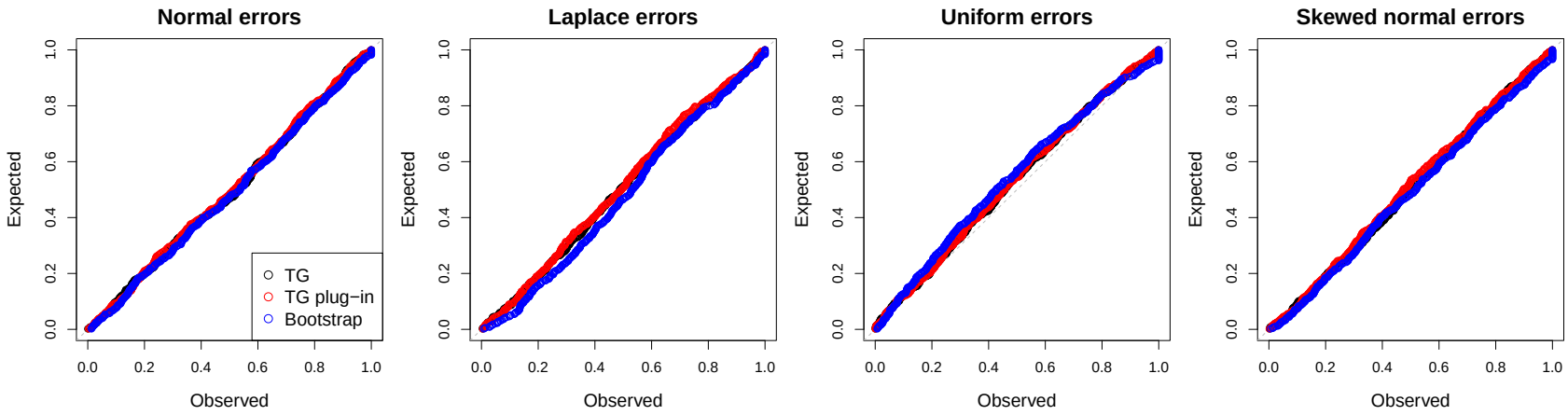
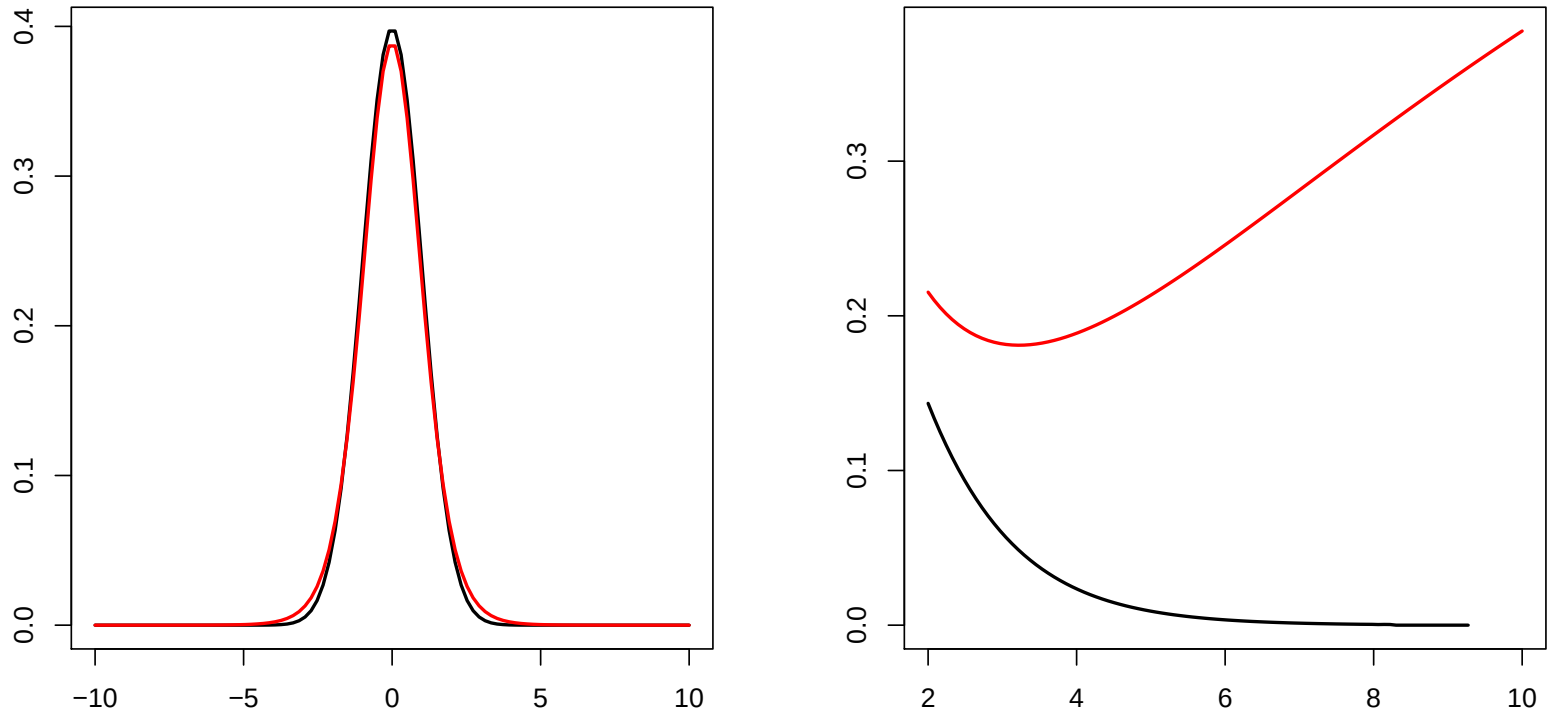


Illustration: fragility of the TG test

Illustration: two densities (left) look very similar, but conditional tail probabilities (right), $H(t) = \mathbb{P}(Z \geq t + 1 \mid Z \geq t)$, very different



TG tests rely on conditional tail probabilities; in high dimensions we can often land far in the tails

A positive asymptotic result

Theorem (Low-dimensional asymptotic setting)

Consider an asymptotic regime with $n \to \infty$ and p fixed. Assume

$$\lim_{n\to\infty} \frac{1}{n} X^T X = \Sigma \quad \textit{and} \quad \lim_{n\to\infty} \max_{i=1,\ldots n} \frac{\|x_i\|_2}{\sqrt{n}} = 0.$$

where x_i, $i = 1, \ldots n$ are the rows of X. Let T denote the TG test statistic after k steps of FS (or LAR, or lasso) and let v denote a contrast giving a partial regression coefficient. Then

$$\lim_{n\to\infty} \sup_{\theta} \sup_{F_n(\theta)} \sup_{\alpha} \left| \mathbb{P}_{v^T\theta=0}\big(T \le \alpha\big) - \alpha \right| = 0,$$

where second supremum is taken over all θ, and error distributions with mean zero and variance σ^2. Further, if $C_{1-\alpha}$ denotes the TG confidence interval, then

$$\lim_{n\to\infty} \sup_{\theta} \sup_{F_n(\theta)} \sup_{\alpha} \left| \mathbb{P}\big(v^T\theta \in C_{1-\alpha}\big) - (1-\alpha) \right| = 0.$$

A negative asymptotic result

Theorem (High-dimensional asymptotic setting)

Consider an asymptotic regime with $n, p \to \infty$*, Assume*

$$Y_{ij} = \theta_j + \epsilon_{ij}, \quad i = 1, \ldots m, \; j = 1, \ldots p$$

where ϵ_{ij}*,* $i = 1, \ldots m$*,* $j = 1, \ldots p$ *are drawn i.i.d. from*

$$\pi \cdot N(-B, 1) + (1 - 2\pi) \cdot N(0, 1) + \pi \cdot N(B, 1).$$

We perform selection based on the largest average $\bar{Y}_j$*,* $j = 1, \ldots p$*. (Equivalent regression with orthogonal design, and* $n = mp$*.) Then, with* $\log p/m \to \infty$*, and under* $\theta = 0$*, the TG statistic converges to 0 on an event with probability at least* $1/e$*. This means of course*

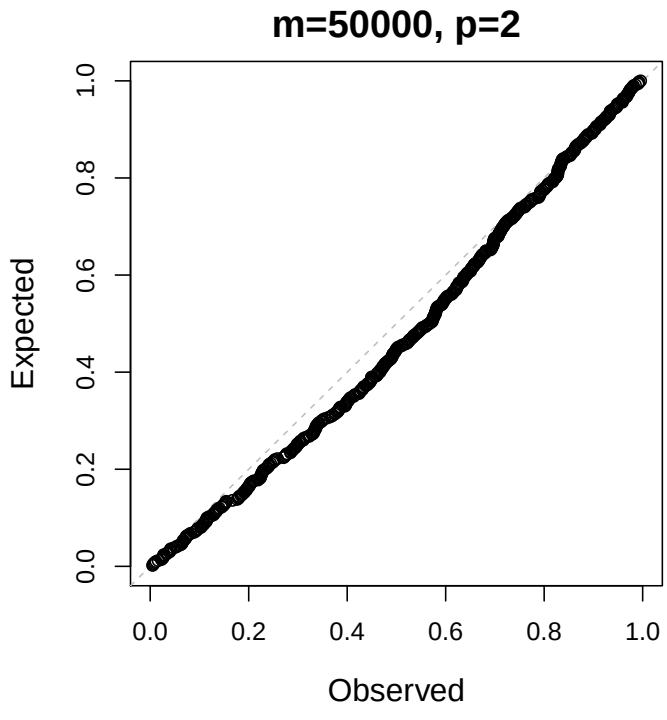
$$\liminf_{n,p \to \infty} \sup_{\theta} \sup_{F_n(\theta)} \sup_{\alpha} \left| \mathbb{P}_{v^T \theta = 0}\big(T \leq \alpha\big) - \alpha \right| \geq 1/e.$$

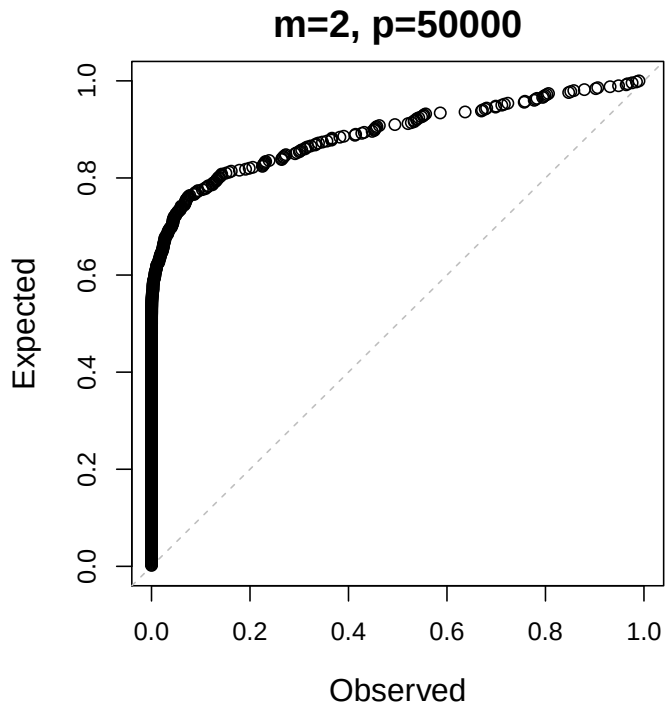
Example: failure in high dimensions

Simulation: from the many normal means setup

$$Y_{ij} = \theta_j + \epsilon_{ij}, \quad i = 1, \ldots m, \; j = 1, \ldots p$$

Recall: equivalent to regression with orthogonal design, and $n = mp$.
Set $\theta = 0$, draw errors from mixture of normals, run one step of FS





Some fixes and alternatives?

- Randomization: Tian and Taylor (2015) introduce auxiliary randomization into the response y (inspired by techniques in differential privacy). Smooths out the pivot ... helps in high dimensions?
- Bootstrap: T., Rinaldo, Tibshirani, Wasserman (2015) prove that using the dist of $v^T y^* \,\big|\, \mathcal{V}^{\mathrm{lo}}(y) \leq v^T y^* \leq \mathcal{V}^{\mathrm{up}}(y)$, over bootstrap samples y^*, yields proper (conservative) inferences in low dimensions. Helps in high dimensions?
- Sample splitting: Wasserman is all about sample splitting ... simple, transparent method with little assumptions. But can be hard to interpret, and does it struggle badly with power?

The problem (high-dim selective inference, with weak assumptions) is still pretty open

Summary, future work

Summary:

- Selective inference for FS, LAR, lasso paths: by conditioning on model selection event (polyhedron), and nuisance parameter ($n-1$ dimensional subspace orthogonal to contrast v), we can test $v^T\theta = 0$ in nearly closed-form
- Better to fit larger model, and to test significance of all active variables (rather than read off p-values sequentially)
- This assumes normal errors; method is robust to nonnormality for small p, but questionable when p is large

Challenges, future work:

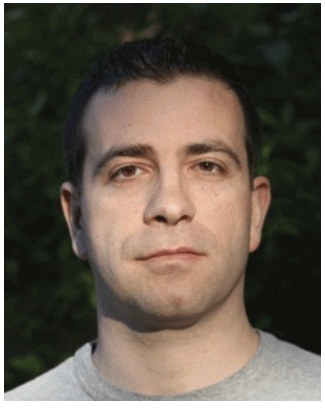
- High-dimensions: more general understanding of when selective inference works / breaks down?
- Sample splitting: loss in power made up for by robustness?
- Software: `selectiveInference` on CRAN, and on github

Acknowledgments

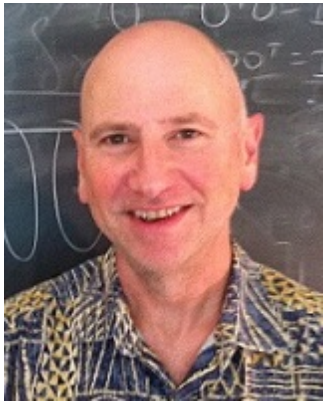
Richard Lockhart

Jonathan Taylor

Rob Tibshirani



Ale Rinaldo



Larry Wasserman

Thank you for listening!

Bonus slides

One-sided or two-sided?

The described inferential setup actually tests

$$H_0 : v^T\theta = 0 \quad \text{versus} \quad H_1 : v^T\theta > 0$$

(because it rejects when $v^T y$ is large). Hence, e.g., we should take

$$v = s_j X_{A_k}(X_{A_k}^T X_{A_k})^{-1} e_j$$

so we reject when projected population coefficient is large and has the same sign s_j as observed coefficient

For two-sided p-values, we can just use $2 \cdot \min\{T, 1-T\}$ ... does this make sense?

(Seems like not, but our intervals implicitly use a two-sided test)

Pivoting around the impossible?

Leeb and Potscher (2006, 2008) give precise impossibility results about estimating the distribution of, say, $v^T y$ after selection

To paraphrase: they show, even under normality, that there is no uniformly consistent estimate of

$$\psi(t) = \mathbb{P}\left(\frac{v^T y - v^T \theta}{\sigma \|v\|_2} \leq t \;\middle|\; y \in \mathcal{P} \right)$$

(uniform over θ). How do our results not contradict this?

We never claim (or attempt) to estimate $\psi(t)$! In our setting, this would be a complicated mixture of truncated Gaussians

Instead we form a pivot for $v^T \theta$ based on the random distribution $F(t; P_{v^\perp} y) = P(v^T y \geq t \;|\; y \in \mathcal{P},\ P_{v^\perp} y)$. There is more than one way to perform inference ...