Significance testing after stepwise model selection

Joshua Loftus

February 25, 2016

Overview

- Introductory example
- Overview of competing methods
- Explanation with a bit of mostly harmless theory
- More examples

Inference after model selection

Step	Variable	Naive[1]	Selective[2]
1	%80th Percentile Income	<0.001	0.001
2	Injury Death Rate	<0.001	0.086
3	Chlamydia Rate	0.078	0.287
4	%Obese	<0.001	0.170
5	%Receiving HbA1c	<0.001	0.335
6	%Some College	0.005	0.864
7	Teen Birth Rate	0.071	0.940
8	Violent Crime Rate	0.067	0.179

Outcome: log-years of potential life lost. Model: 8 out of 31 predictors chosen by forward stepwise with BIC.

1: In R, using `summary` on selected model

2: From `groupfs` function in `selectiveInference` package

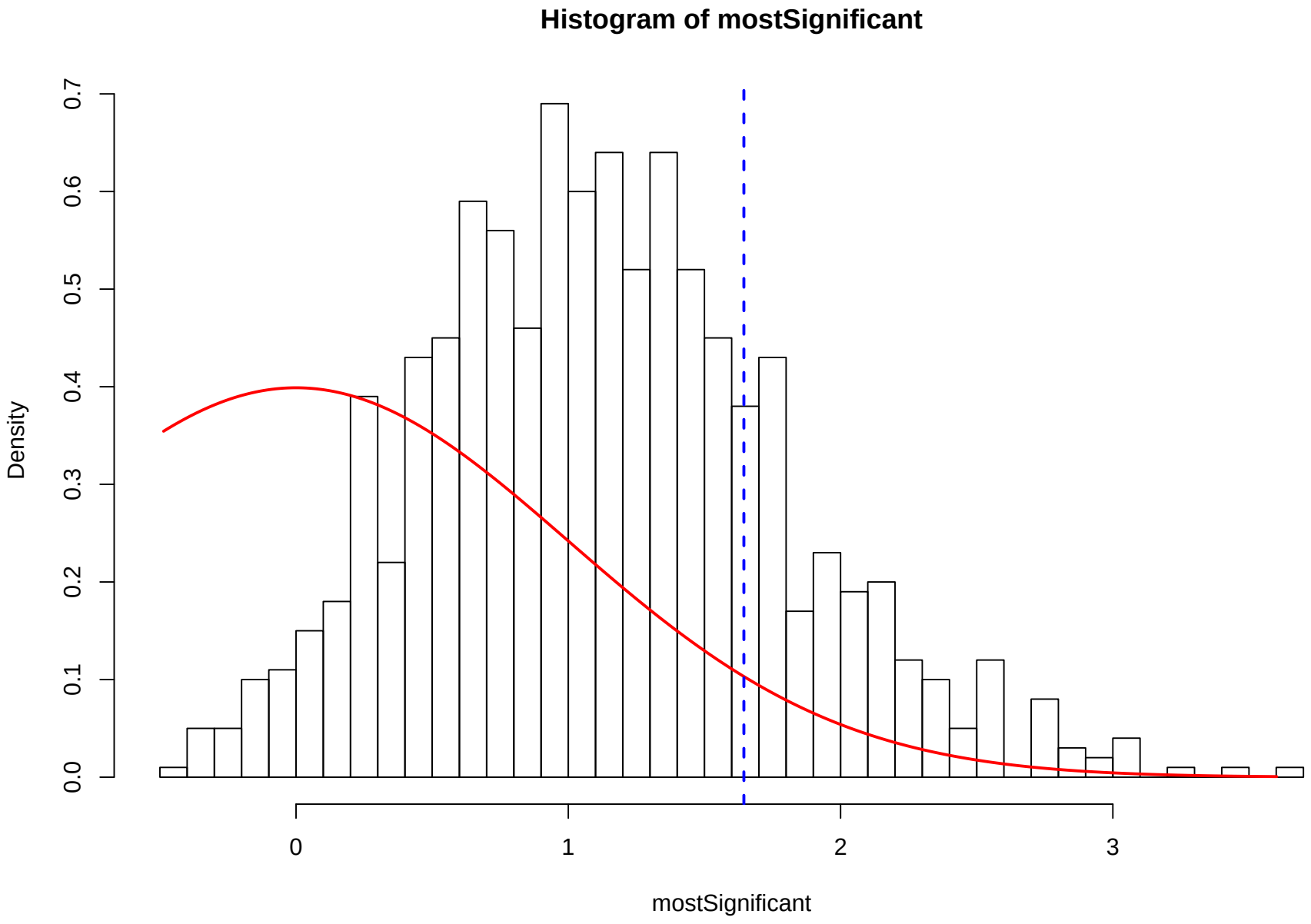
Why is the "naive" method wrong?

Simple example: choose the largest of 5 effects.

```
maxz <- function(n) return(max(rnorm(n)))
mostSignificant <- replicate(1000, maxz(5))
range <- seq(min(mostSignificant),
             max(mostSignificant),
             length.out = 1000)
```

This data is generated under a global null, all effects have 0 mean. What happens if we now compute p-values for the selected effects?

Selection bias: type 1 error 0.203 instead of 0.05



Selection can make noise look like signal

Any time we use the data to make a decision (e.g. pick one model instead of some others), we introduce selection bias

Forward stepwise, Lasso, elastic net with cross-validation, etc, all use the data in a way that would result in such bias

Significance tests, prediction error, R^2, goodness of fit tests, etc, will all suffer from selection bias

Big contributor to reproducibility crisis

> ***We conducted replications*** *of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. . . . Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size;* ***39% of effects were subjectively rated to have replicated the original result***

From Estimating the reproducibility of psychological science Open Science Collaboration (2015). See also Why most published research findings are false Ioannidis (2005).

What's the most common solution?

What's the most common solution?

Before doing any selection, set aside some **validation data**. Then, *after* the final model is chosen, use this validation set to compute prediction error, significance tests, etc

Data splitting–most commonly used acceptable practice

Survival example: Cox's PH model, regularized

- Data: 240 lymphoma patients, 7399 genes

```
train <- sample(nrow(x), 140)
x.train <- x[train,]
y.train <- Surv(y[train], status[train])
fit <- glmnet(x.train, y.train, family = "cox",
              maxit = 10000)
cv.fit <- cv.glmnet(x.train, y.train, family = "cox",
                    maxit = 10000)
coefs <- coef(fit, s = cv.fit$lambda.min)
active <- which(coefs != 0)
length(active)
```

```
## [1] 15
```

Inference for the selected model

```
test <- setdiff(1:nrow(x), train)
x.test <- x[test, active]
y.test <- Surv(y[test], status[test])
fit.test <- coxph(y.test ~ x.test)
fit.test
```

```
## Call:
## coxph(formula = y.test ~ x.test)
##
##
##              coef exp(coef) se(coef)      z     p
## x.test1  -0.2730     0.761   0.2096 -1.3027 0.190
## x.test2   0.6954     2.005   0.4394  1.5826 0.110
## x.test3   0.1218     1.130   0.3748  0.3250 0.750
## x.test4  -0.0145     0.986   0.3038 -0.0477 0.960
## x.test5   0.0755     1.078   0.1918  0.3937 0.690
## x.test6  -0.1430     0.867   0.0648 -2.2079 0.027
```

Data splitting

Pros:

- Simple: only took a few lines of code
- Robust: requires few assumptions
- Controls type 1 error, no selection bias

Data splitting

Pros:

- Simple: only took a few lines of code
- Robust: requires few assumptions
- Controls type 1 error, no selection bias

Cons:

- Reproducibility issues: different random splits, different split proportions
- Efficiency: using less data for model selection, also less power
- Feasibility: categorical variables with rare levels (e.g. rare variants)

Selective error control

New and active research area at Stanford; Jonathan Taylor, Rob Tibshirani, and many coauthors

To adjust for the selection effect, *condition* on the selected model

Mathematically, if we select M, want test a null hypothesis H_0 about M (e.g. significance test for a variable in M), we control

$$P_{M,H_0}(\text{reject } H_0 | M \text{ selected}) \leq \alpha$$

Selective error control

New and active research area at Stanford; Jonathan Taylor, Rob Tibshirani, and many coauthors

To adjust for the selection effect, *condition* on the selected model

Mathematically, if we select M, want test a null hypothesis H_0 about M (e.g. significance test for a variable in M), we control

$$P_{M,H_0}(\text{reject } H_0 | M \text{ selected}) \leq \alpha$$

If a variable "surprises" us enough to be *included in the model*, it must surprise us *again* in order to be *declared significant*

- Data splitting controls this error trivially
- Controlling this would fix reproducibility problems

A simple example: screening rule

Observe many (independent) means, select the effects which might be large, say > 1

```
N <- 10000
Zs <- rnorm(N)
bigZs <- Zs[Zs > 1]
length(bigZs)/N
```

```
## [1] 0.1637
```

A simple example: screening rule

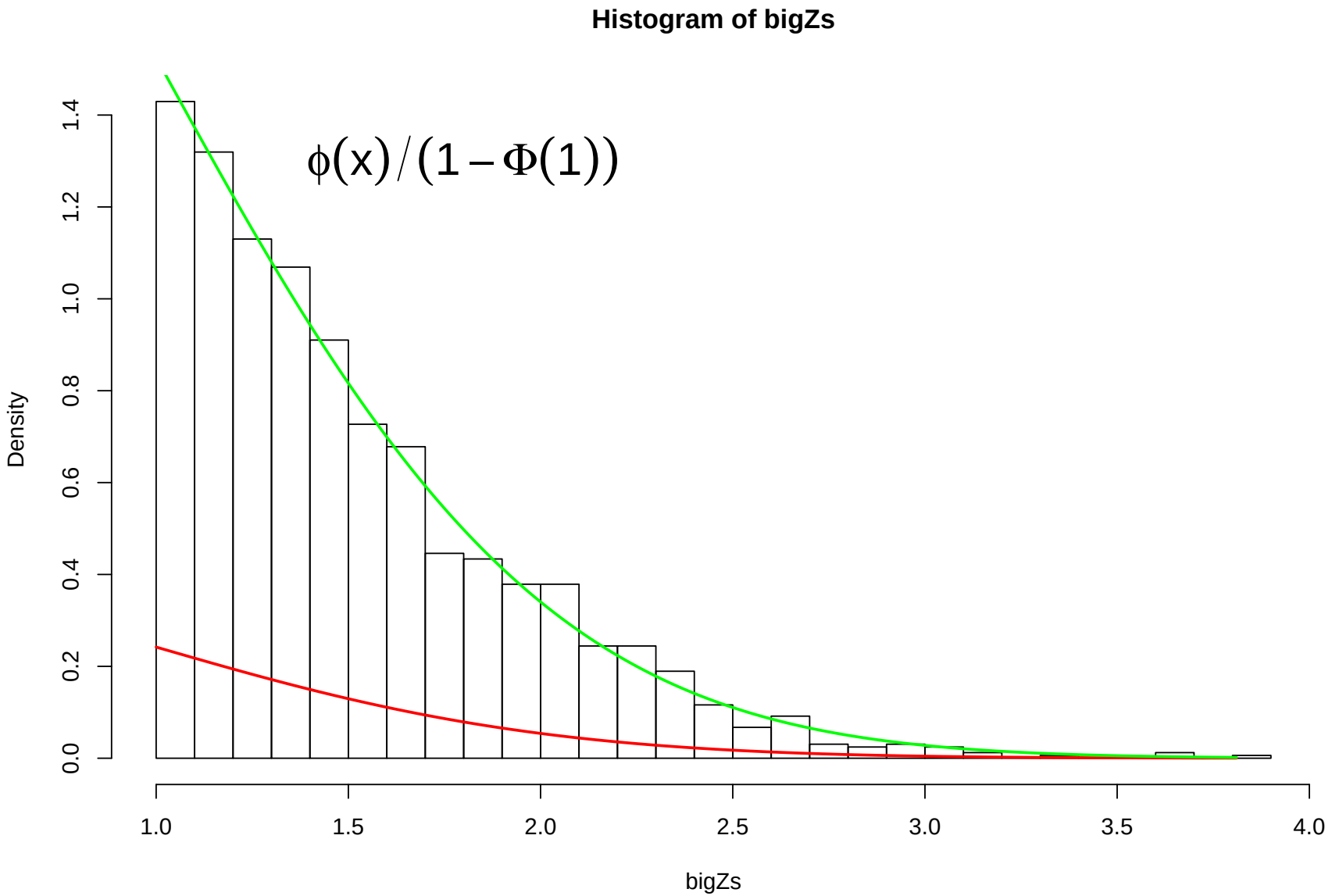
Observe many (independent) means, select the effects which might be large, say > 1

```
N <- 10000
Zs <- rnorm(N)
bigZs <- Zs[Zs > 1]
length(bigZs)/N
```

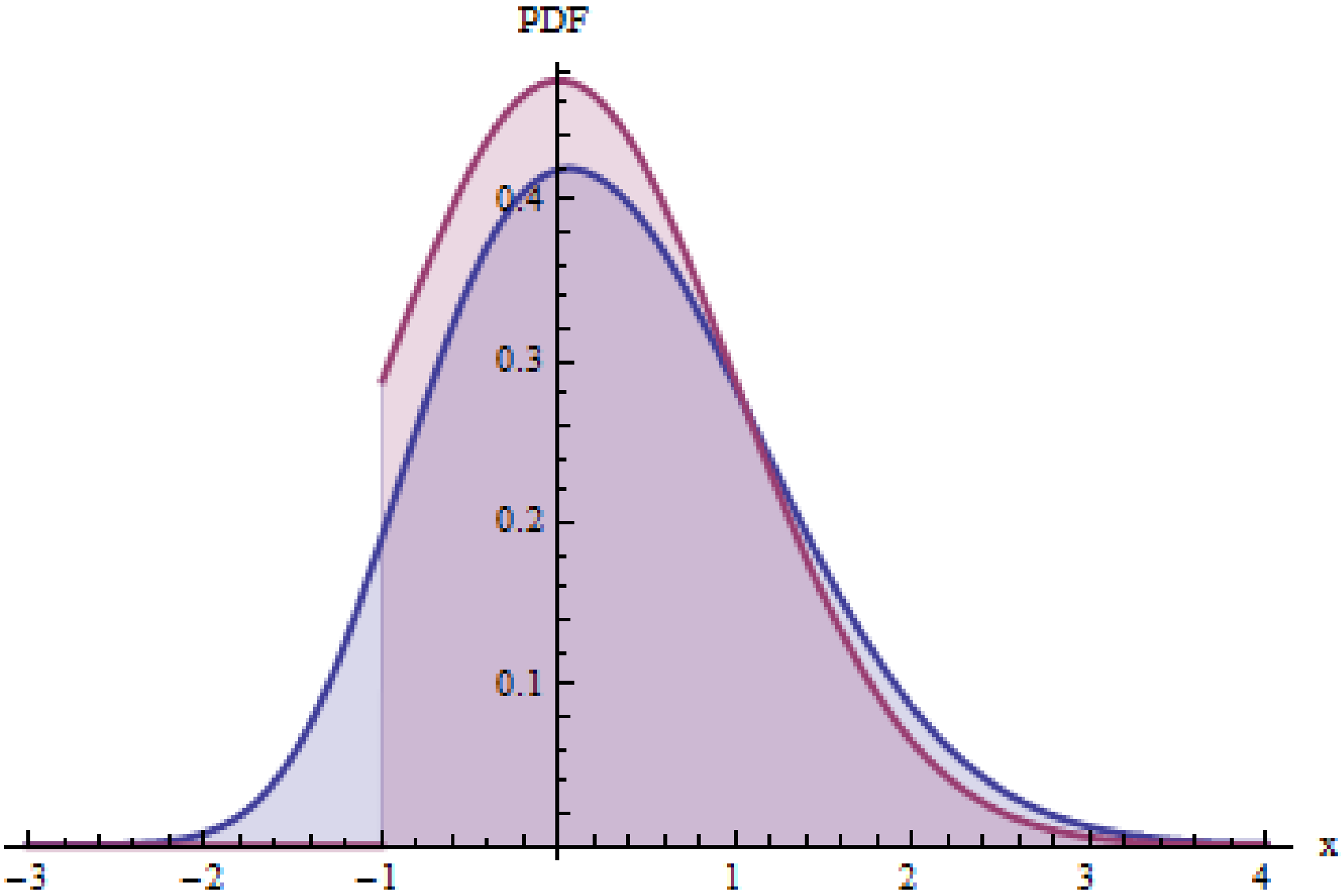
```
## [1] 0.1637
```

These are all null. What probability law can we compare them to and not mistakenly declare them all significant?

Truncated probability law



Conditional probability via truncation



Details vary depending on procedure

Most of the work in this field is in understanding the geometry of truncation for each particular model selection procedure

Exercise: truncation for the max selection rule (selecting the largest effect only)?

My dissertation work focuses on procedures with more complicated geometry (quadratic constraints), and includes a few important special cases

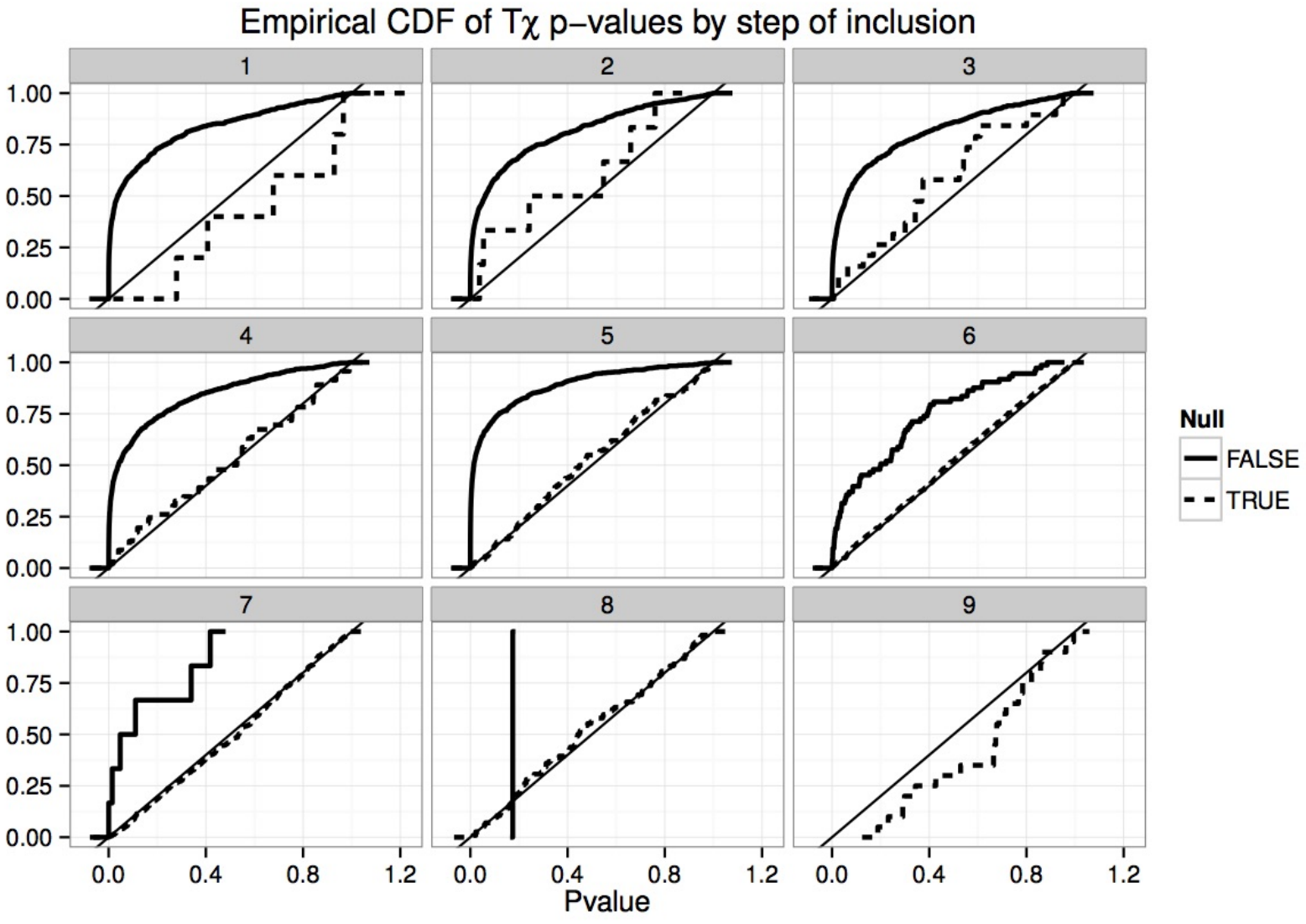
groupfs simulation: $n = 100$, $p = 100$, $P = 50$

Model size chosen with BIC. Groups 1-4 have $\|\beta\|_2 = .84$ within groups, all else 0

```
> set.seed(1)
  ...
> fit <- groupfs(x, y, ...)
> pvals <- groupfsInf(fit)
> pvals
  Group Pvalue     TF df   Size Ints    Min     Max
1     3  0.088 49.913  2 67.811    1 44.949 112.760
2     1  0.000 98.077  1 54.267    1 68.151 122.418
3     2  0.003 69.266  1 28.659    1 50.423  79.082
4     4  0.000 37.099  2 28.803    1 20.194  48.997
5    47  0.319  5.143  1  3.887    1  3.518   7.406
```

Ignoring selection, first 4 p-values are \$<\$0.001, for 47 it's 0.024

Bigger forward stepwise simulation



Heterogeneous effects

Identify subgroups of population where treatment is most effective. Interactions of treatment with various covariates. For example, biomarkers and drug efficacy (HIV example in Loftus and Taylor, (2014).

Inference for the selected treatment effects, data splitting may be unrealistic due to sample size or $p \gg n$

Ongoing work: analyze power and causal inference properties for this approach

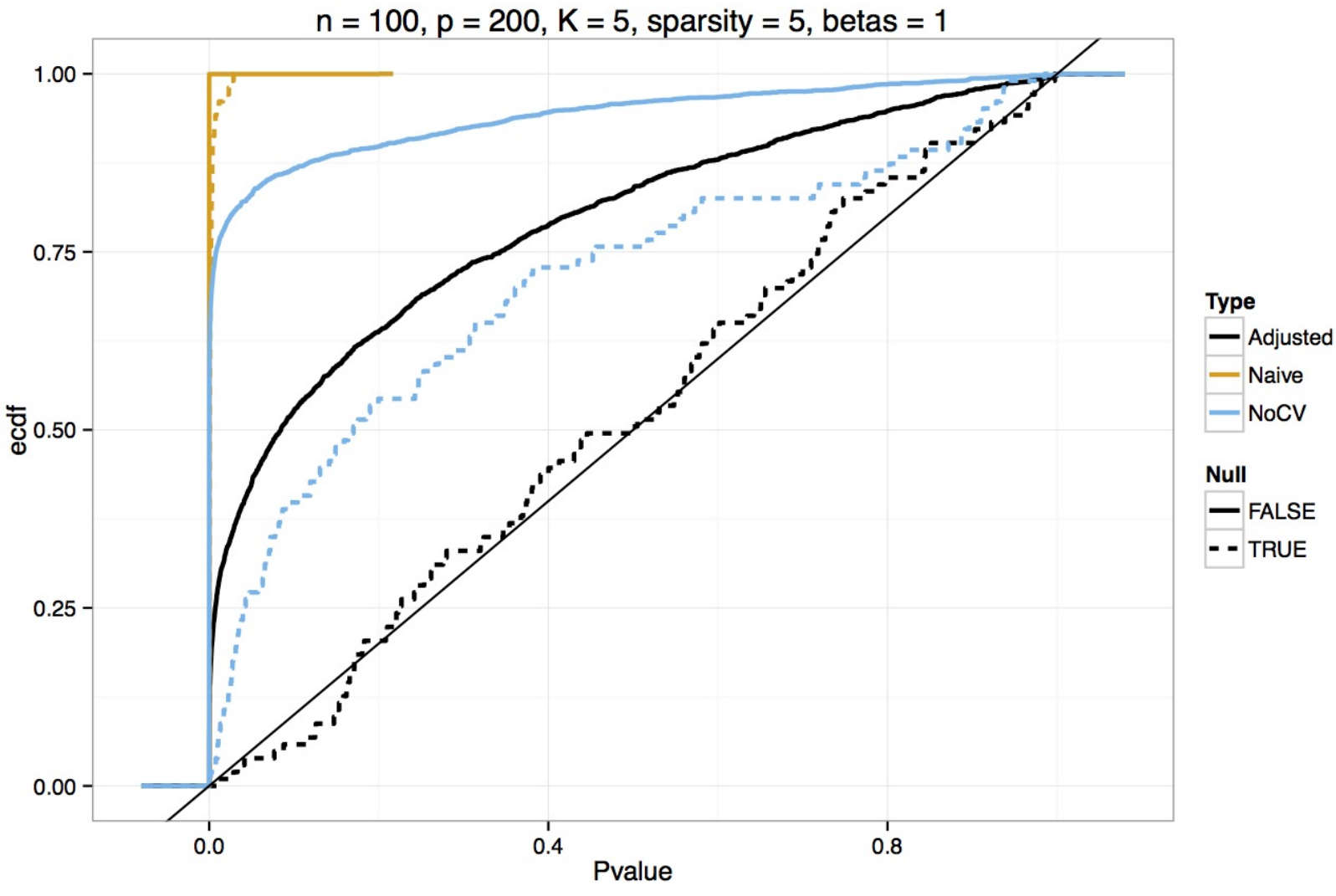
Another important example: cross-validation

Use cross-validation to choose λ in Lasso or number of steps in forward stepwise

Very complicated model selection procedure, but (recently) able to compute truncated distributions for p-values

{JL. Selective inference after cross-validation. arXiv Preprint. (2015)}

Cross-validation simulation



Remarks

Technical details in the papers, but beware:

- Tests not independent
- Computationally expensive (cross-validation especially)
- Can be low powered against some alternatives

Software implementation: `selectiveInference` R package (see also our github repo)

Which method to use for a given problem?

If n is very large, might just use data splitting (simple)

Otherwise, consider the conditional approach, especially if $p > n$ or bottlenecks like rare observations limit effective sample size

If p is small, more conservative method ("PoSI") is available, see Berk et. al. (2013).

Questions?

Thanks for your attention!

References

- Berk et al, (2010). Statistical inference after model selection. Journal of Quantitative Criminology.
- Berk et al, (2013). Valid post-selection inference. Annals of Statistics.
- Loftus, (2015). Selective inference after cross-validation. arXiv Preprint.
- Loftus and Taylor, (2015). Selective inference in regression models with groups of variables. arXiv Preprint.
- Taylor, Tibshirani (2015). Statistical learning and selective inference. **PNAS**.
- Simon et al, (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software.