Significance testing after cross-validation

Joshua Loftus (jloftus@turing.ac.uk)
(building from joint work with Jonathan Taylor)
9 December, 2016
Slides and markdown source at
https://joftius.github.io/turing

Setting: regression model selection

Linear model

$$y = X\beta + \epsilon$$

- y vector of outcomes
- X predictor/feature matrix
- β parameters/weights to be estimated, assume most are "null," i.e. equal 0 (sparsity)
- ϵ random errors, assume probability distribution $N(0, \sigma^2 I)$
- Pick subset of predictors we think are non-null
- How good is the model using this subset?
- Are chosen predictors actually non-null, i.e. significant?

Type 1 error: declaring a predictor significant when it is actually null.

Motivating example: forward stepwise

Data: California county health data...
Outcome: log-years of potential life lost.
Model: 5 out of 30 predictors chosen by FS with AIC.

```
model <- step(lm(y ~ .-1, df), k = 2, trace = 0)
print(summary(model)$coefficients[,c(1,4)], digits = 2)
```

```
##                              Estimate Pr(>|t|)
## Food.Environment.Index          0.342   0.0296
## `%.With.Access`                -0.036   0.0017
## `%.Excessive.Drinking`          0.090   0.0182
## Teen.Birth.Rate                 0.026   0.0045
## Average.Daily.PM2.5            -0.225   0.0211
```

5 interesting effects, all significant. Time to publish!

What's wrong with this?

The outcome was actually just noise, independent of the predictors

```
set.seed(1)
df = read.csv("CaliforniaCountyHealth.csv")
df$y <- rnorm(nrow(df)) #!!!
```

(With apologies for deceiving you, I hope this makes the point...)

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 may introduce a selection effect (bias).

This happens with forward stepwise, Lasso, elastic net with cross-validation, etc.

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

Most common solution: data splitting

Pros:

- Simple: only takes a few lines of code
- Robust: requires few assumptions
- Controls (selective) 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)

Literature on (conditional) post-selection inference

- Frequentist interpretation — Hurvich & Tsai (1990)
- Lasso, sequential — Lockhart et al. (2014)
- General penalty, global null, geometry — Taylor, Loftus, and Tibshirani (2015), Azaïs, Castro, and Mourareau (2015)
- Forward stepwise, sequential — Loftus and Taylor (2014)
- Fixed λ Lasso / conditional — Lee et al. (2015), Fithian, Sun, and Taylor (2014)
- Forward stepwise and LAR — Tibshirani et al. (2014)
- Asymptotics — Tian and Taylor (2015a)
- Unknown σ — Tian, Loftus, and Taylor (2015), Gross, Taylor, and Tibshirani (2015)
- Group selection / unknown σ — Loftus and Taylor (2015)
- Cross-validation — Tian and Taylor (2015b), Loftus (2015)
- Unsupervised learning — Blier, Loftus, and Taylor (2016)

(Incomplete list, growing fast)

Previous work: affine model selection

- Model selection map $M : \mathbb{R}^n \to \mathcal{M}$, with $\mathcal{M}$ space of potential models.
- Observe $E_m = \{M(y) = m\}$, want to condition on this event.
- For many model selection procedures (e.g. Lasso at fixed λ)

$$\underbrace{\mathcal{L}(y|M(y) = m)}_{\text{what we want}} = \mathcal{L}(y|\underbrace{A(m)y \leq b(m)}_{\text{simple geometry}}) \quad \text{on } \{M(y) = m\}$$

MVN constrained to a polytope.

Quadratic model selection framework

For some model selection procedures (e.g. forward stepwise with groups, cross-validation), model selection event can be decomposed as

Quadratic selection event

$$E_m := \{M(y) = m\} = \bigcap_{j \in J_m} \{y : y^T Q_j y + a_j^T y + b_j \geq 0\}$$

- These Q, a, b are constant on E_m, so conditionally they are constants
- For conditional inference, need to compute this intersection of quadratics

Truncated χ significance test

Suppose $y \sim N(\mu, \sigma^2 I)$ with σ^2 known, $H_0(m) : P_m \mu = 0$, P_m is constant on $\{M(y) = m\}$, $r := \mathsf{Tr}(P_m)$, $R := P_m y$, $u := R/\|R\|_2$, $z := y - R$, $D_m := \{t \geq 0 : M(ut\sigma + z) = m\}$, and the observed statistic $T = \|R\|_2/\sigma$

Post-selection $T\chi$ distribution

$$T|(m, z, u) \sim \chi_r|_{D_m} \tag{1}$$

where the vertical bar denotes truncation. Hence, with f_r the pdf of a central χ_r random variable

$$T\chi := \frac{\int_{D_m \cap [T,\infty]} f_r(t)dt}{\int_{D_m} f_r(t)dt} \sim U[0, 1] \tag{2}$$

is a p-value controlling selective type 1 error.

Geometry problem: intersection of quadratic regions

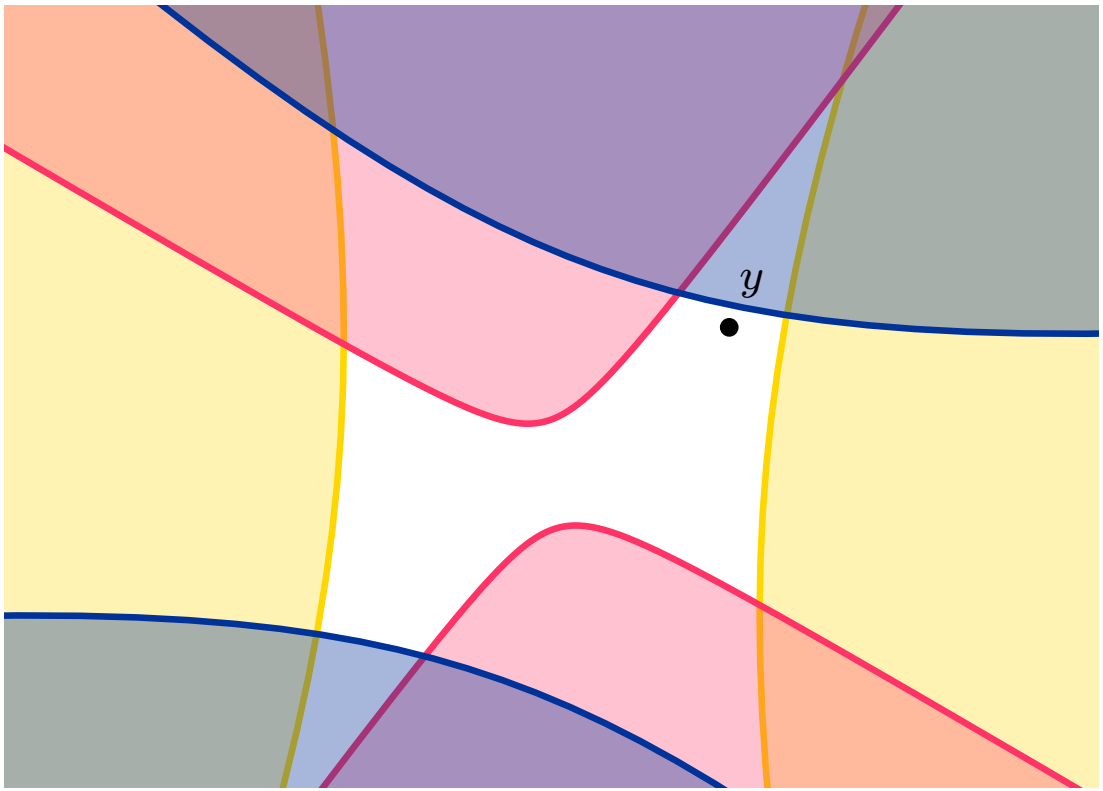


Figure 1: The *complement* of each quadratic is shaded with a different color. The unshaded, white region is E_m.

Geometry problem: intersection of quadratic regions

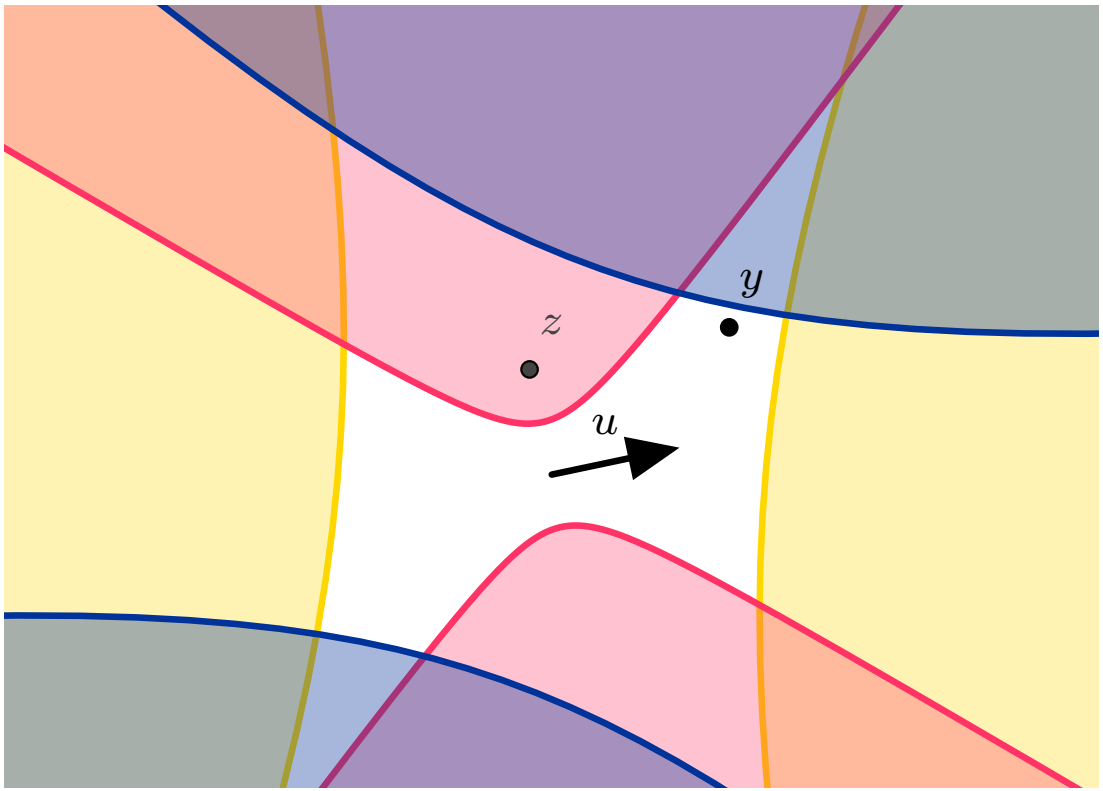


Figure 1: The *complement* of each quadratic is shaded with a different color. The unshaded, white region is E_m.

Geometry problem: intersection of quadratic regions

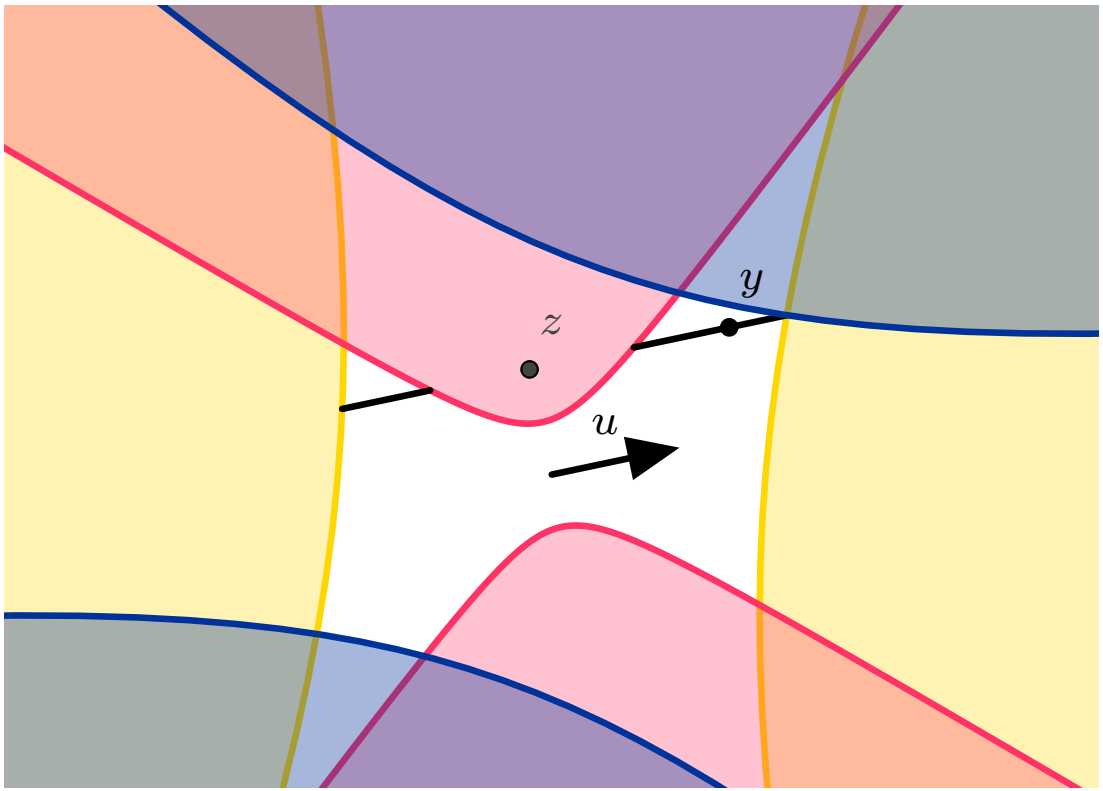


Figure 1: The *complement* of each quadratic is shaded with a different color. The unshaded, white region is E_m.

Geometry problem: intersection of quadratic regions

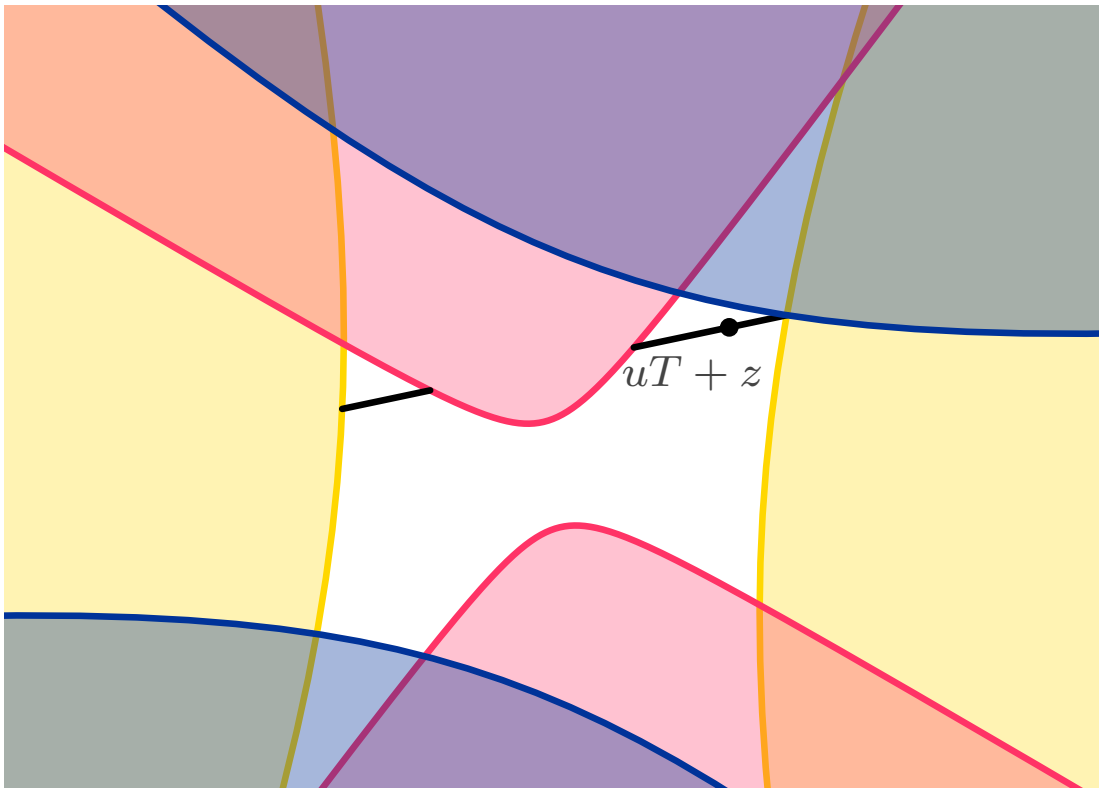


Figure 1: The *complement* of each quadratic is shaded with a different color. The unshaded, white region is E_m.

Adaptive model selection with cross-validation

- For K-fold cv, data partitioned (randomly) into $D_1, \ldots, D_K$. For each $k = 1, \ldots, K$, hold out D_k as a test set while training a model on the other $K - 1$ folds. Form estimate RSS_k of out-of-sample prediction error. Average these estimates over test folds.
- Use to choose model complexity: evaluate $RSS_{k,s}$ for various sparsity choices s. Pick s minimizing the cv-RSS estimate.
- Run forward stepwise with maxsteps S. For $s = 1, \ldots, S$ evaluate the test error $RSS_{k,s}$. Average to get RSS_s. Pick s^* minimizing this. Run forward stepwise on the whole data for s^* steps.

Can we do selective inference for the final models chosen this way?

Notation for cross-validation

- Let f, g index CV test folds.
- On fold f, model m_f at step s, and $-f$ denoting the training set for test fold f (complement of f).
- Define $P_{f,s} := X^f_{m_f,s}(X^{-f}_{m_f,s})^\dagger$ (not a projection)
- $s = \text{argmin}_s \sum_{f=1}^K \|y^f - P_{f,s} y^{-f}\|_2^2$
- Sums of squares... maybe it's a quadratic form?

Blockwise quadratic form of cv-RSS

Key result of Loftus (2015).

Define $Q^s_{ff} := \sum_{g \neq f} (P_{g,s})^T_f (P_{g,s})_f$ and

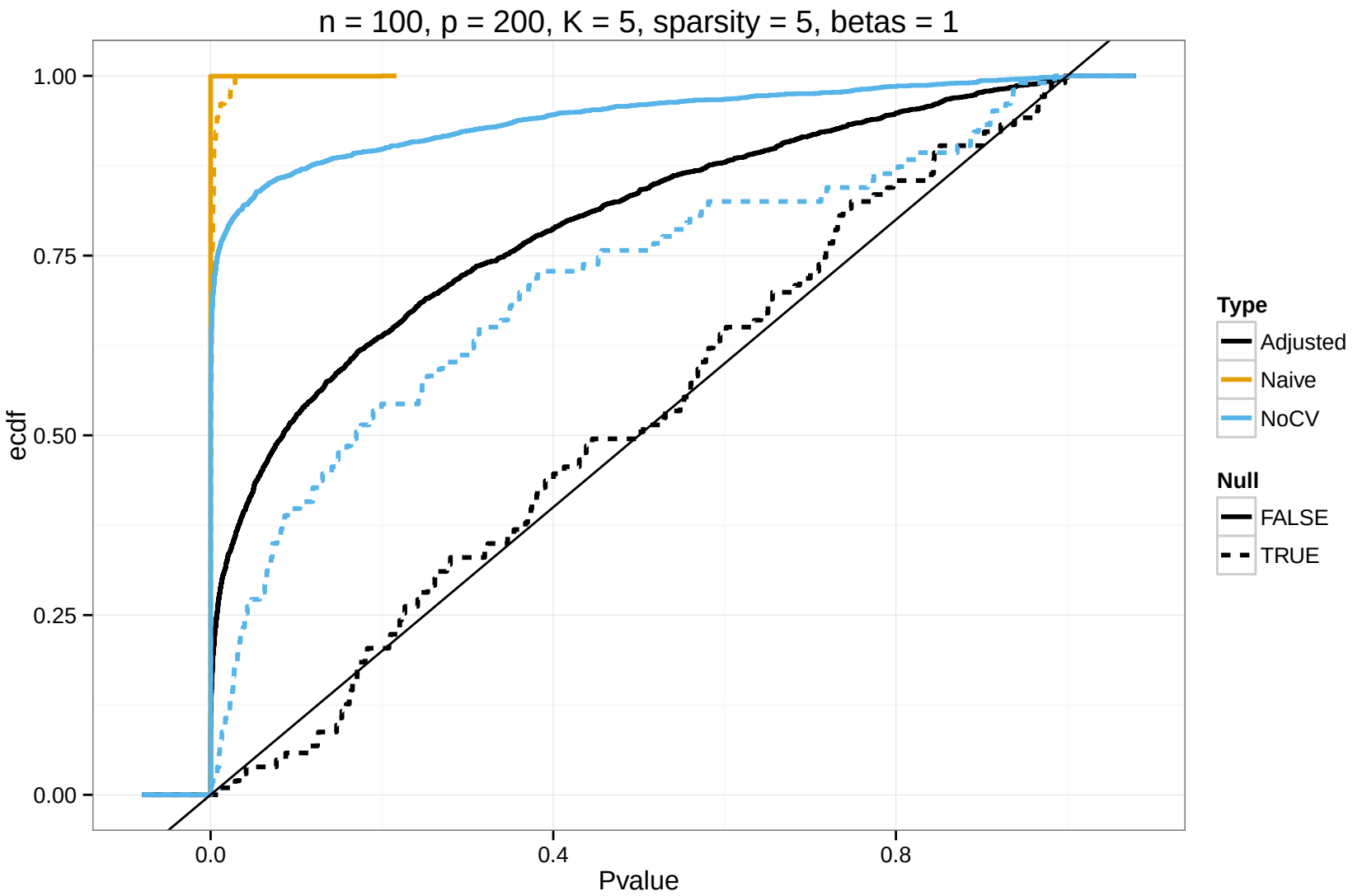
$$Q^s_{fg} := -(P_{f,s})_g - (P_{g,s})^T_f + \sum_{\substack{h=1 \\ h \notin \{f,g\}}}^{K} (P_{h,s})^T_f (P_{h,s})^T_g$$

Then with y_K denoting the observations ordered by CV-folds,

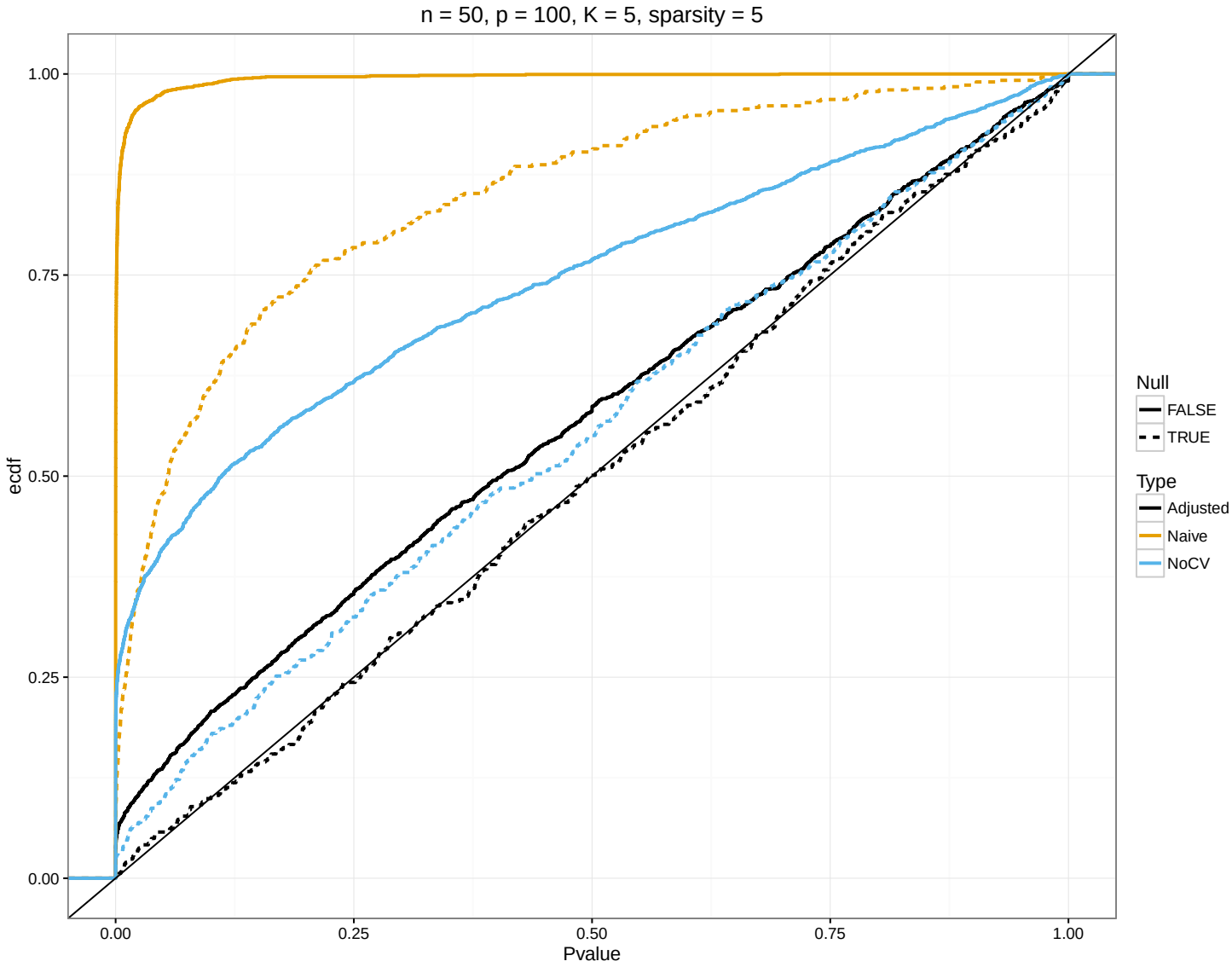
$$\text{cv-RSS}(s) = y_K^T Q^s y_K$$

This quadratic form allows us to conduct inference conditional on models selected by cross-validation

Empirical CDF: forward stepwise simulation



Empirical CDF: LAR simulation



Remarks

Technical details in the papers, a few notes:

- Tests not independent
- Computationally expensive
- May be low powered against some alternatives
- Can also do σ^2 unknown case
- Most usual limitations of model selection still apply

Software implementation: `selectiveInference` R package on CRAN
Github repo: `https://github.com/selective-inference/`

References

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

Thanks for your attention!

Questions?

jloftus@turing.ac.uk

More references

Azaïs, Jean-Marc, Yohann de Castro, and Stéphane Mourareau. 2015. "Power of the Kac-Rice Detection Test." *ArXiv Preprint ArXiv:1503.05093*.

Blier, Léonard, Joshua R. Loftus, and Jonathan E. Taylor. 2016. "Inference on the Number of Clusters in k-Means Clustering." *In Progress*.

Fithian, William, Dennis Sun, and Jonathan Taylor. 2014. "Optimal Inference After Model Selection." *ArXiv Preprint ArXiv:1410.2597*.

Gross, S. M., J. Taylor, and R. Tibshirani. 2015. "A Selective Approach to Internal Inference." *ArXiv E-Prints*, October.

Lee, Jason D, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. 2015. "Exact Post-Selection Inference with the Lasso." *Ann. Statist.*

Lockhart, Richard, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. 2014. "A Significance Test for the Lasso." *Annals of Statistics* 42 (2). NIH Public Access: 413.

Loftus, J. R., and J. E. Taylor. 2015. "Selective inference in regression