Post-selection inference for models characterized by quadratic inequalities

Joshua R. Loftus

Statistics Seminar
Rice University
January 25th, 2016

Outline

1. Intro on post-selection inference
2. Examples
   - Forward stepwise with groups
   - Cross-validation
3. Quadratic model selection events
4. Conclusion and future work

Section 1

Intro on post-selection inference

Regression modeling

- Data: $n \times p$ predictor matrix X, outcome of interest y
- Probability model: $y \sim N(\mu, \sigma^2 I)$
- Linear model: $\mu = \mathbb{E}[y] = X\beta$
- Sparse linear model: $\mu = X(A_0)\beta(A_0)$ for some (small) subset A_0 of the columns of X
- Given data driven methods of choosing A_0

Our goal

Inference about $\beta(A_0)$, e.g. p-values for significance testing

Regression modeling

- Data: $n \times p$ predictor matrix X, outcome of interest y
- Probability model: $y \sim N(\mu, \sigma^2 I)$
- Linear model: $\mu = \mathbb{E}[y] = X\beta$
- Sparse linear model: $\mu = X(A_0)\beta(A_0)$ for some (small) subset A_0 of the columns of X
- Given data driven methods of choosing A_0

Our goal

Inference about $\beta(A_0)$, e.g. p-values for significance testing

Regression modeling

- Data: $n \times p$ predictor matrix X, outcome of interest y
- Probability model: $y \sim N(\mu, \sigma^2 I)$
- Linear model: $\mu = \mathbb{E}[y] = X\beta$
- Sparse linear model: $\mu = X(A_0)\beta(A_0)$ for some (small) subset A_0 of the columns of X
- Given data driven methods of choosing A_0

Our goal

Inference about $\beta(A_0)$, e.g. p-values for significance testing

Regression modeling

- Data: $n \times p$ predictor matrix X, outcome of interest y
- Probability model: $y \sim N(\mu, \sigma^2 I)$
- Linear model: $\mu = \mathbb{E}[y] = X\beta$
- Sparse linear model: $\mu = X(A_0)\beta(A_0)$ for some (small) subset A_0 of the columns of X
- Given data driven methods of choosing A_0

Our goal

Inference about $\beta(A_0)$, e.g. p-values for significance testing

Motivating example using forward stepwise

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

Table: Significance test p-values for predictors chosen by forward stepwise with BIC. The outcome is log-years of potential life lost in counties of California, 8 out of 31 predictors chosen. Data: county health rankings of (University of Wisconsin Population Health Institute, 2015).

[1] In R, run `step` and then `summary`

[2] Using `selectiveInference` package

Biased p-values $\Longrightarrow$ reproducibility crisis?

This is a big problem (not *only* due to p-values, of course)

> *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).

Data splitting has issues

Most common solution, split data into training and test partitions, use the training data to select a model, and use the test data for inference

- Inefficiency: less data for model selection, less data for inference (compare to "data carving" in Fithian et al. (2014))
- Variability: results depend on partitions. Can try many partitions and find best one. What ratio to split?
- Infeasibility: correlated data, small sample, rare observations

The upshot of this talk

Our method is an alternative which does not suffer from these problems

Data splitting has issues

Most common solution, split data into training and test partitions, use the training data to select a model, and use the test data for inference

- Inefficiency: less data for model selection, less data for inference (compare to “data carving” in Fithian et al. (2014))
- Variability: results depend on partitions. Can try many partitions and find best one. What ratio to split?
- Infeasibility: correlated data, small sample, rare observations

The upshot of this talk

Our method is an alternative which does not suffer from these problems

Data splitting has issues

Most common solution, split data into training and test partitions, use the training data to select a model, and use the test data for inference

- Inefficiency: less data for model selection, less data for inference (compare to "data carving" in Fithian et al. (2014))
- Variability: results depend on partitions. Can try many partitions and find best one. What ratio to split?
- Infeasibility: correlated data, small sample, rare observations

The upshot of this talk

Our method is an alternative which does not suffer from these problems

Data splitting has issues

Most common solution, split data into training and test partitions, use the training data to select a model, and use the test data for inference

- Inefficiency: less data for model selection, less data for inference (compare to “data carving” in Fithian et al. (2014))
- Variability: results depend on partitions. Can try many partitions and find best one. What ratio to split?
- Infeasibility: correlated data, small sample, rare observations

The upshot of this talk

Our method is an alternative which does not suffer from these problems

Data splitting has issues

Most common solution, split data into training and test partitions, use the training data to select a model, and use the test data for inference

- Inefficiency: less data for model selection, less data for inference (compare to "data carving" in Fithian et al. (2014))
- Variability: results depend on partitions. Can try many partitions and find best one. What ratio to split?
- Infeasibility: correlated data, small sample, rare observations

The upshot of this talk

Our method is an alternative which does not suffer from these problems

Simulation example

Design: $X_{50 \times 100}$ generated with Gaussian columns having common correlation 0.1 (remains fixed)

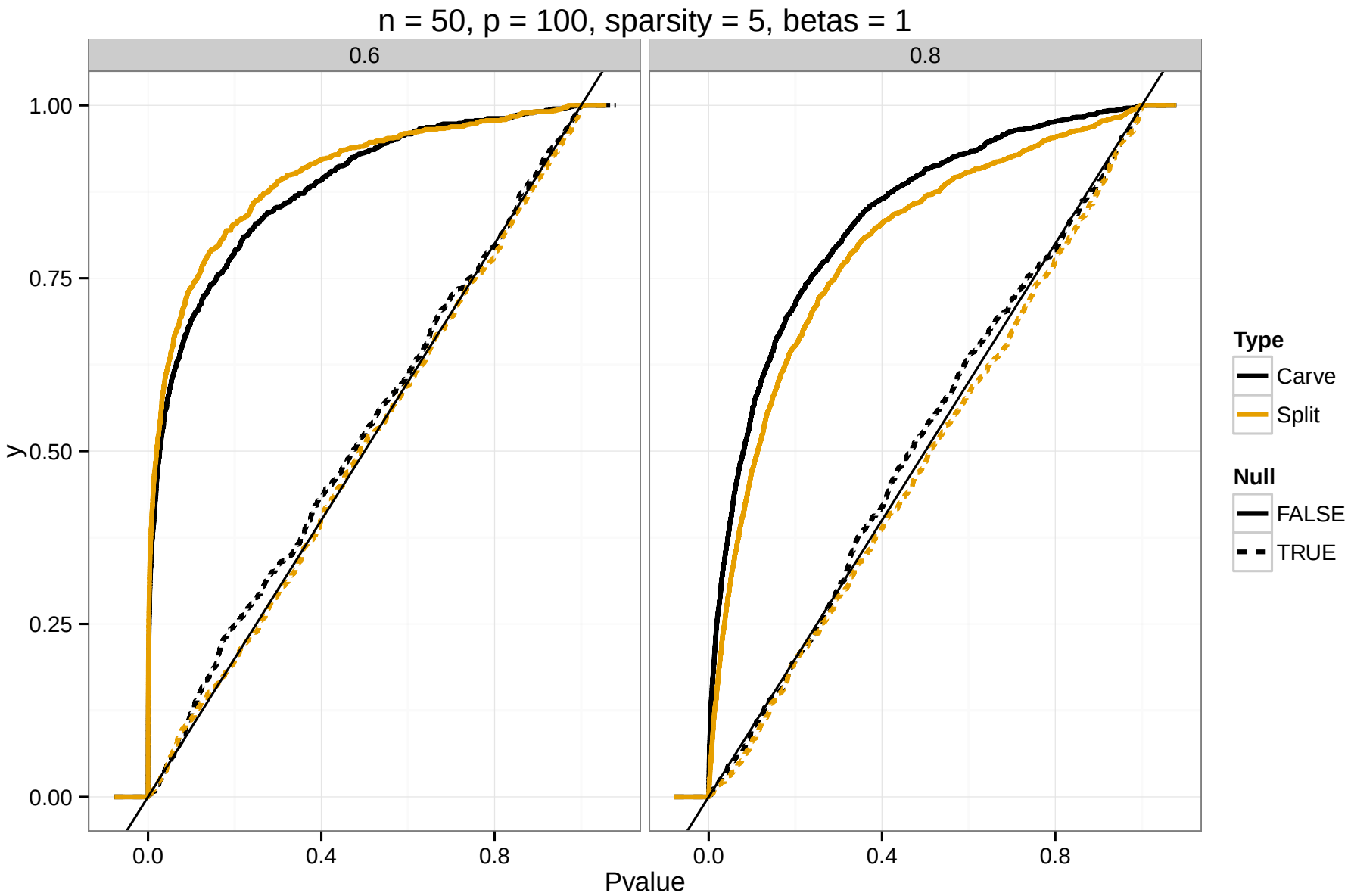
Outcome: $y = X\beta + \epsilon$ where β is 5-sparse and has nonzero entries ± 1, $\epsilon \sim N(0, 1)$ (500 instances)

Method: Forward stepwise using RIC criteria ($2 \log(p)$ instead of 2 for AIC or $\log(n)$ for BIC)

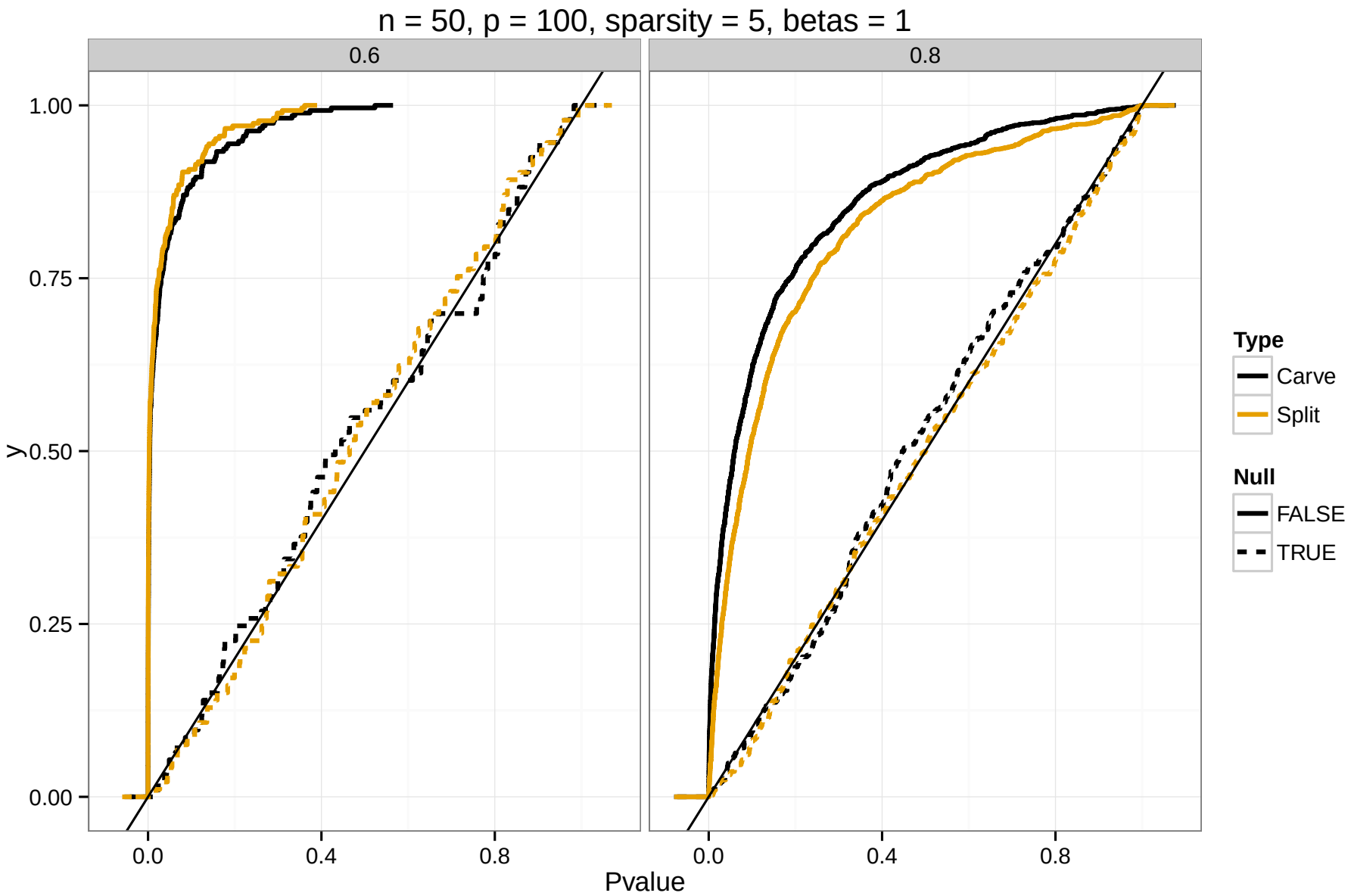
Training data: 60% for left panel, 80% for right

p-values computed using test set only (Split) or combined with selective truncated F test p-values (Carve)

ECDFs of *p*-values for selected variables



Conditional on screening: 10.8% vs 54.2%



Related work

There has been much recent work in this area. Following Berk et al. (2013), we consider two categories:

1. Full-model: inference about the parameters β_j in

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

2. Sub-model: inference about the parameters $\beta(A_0)_j$ in

$$y = X(A_0)\beta(A_0) + \epsilon$$

for some $A_0 \subset \{1, \ldots, p\}$.

Inference in the full model $\mu = X\beta$

FDR control or similar

- Screen & clean — Wasserman & Roeder (2009)
- Stability selection — Meinshausen & Bühlmann (2010)
- Empirical Bayes — Efron (2011)
- SLOPE — Bogdan et al. (2014)
- Knockoffs — Barber & Candès (2015)

Type 1 error

- Univariate treatment — Belloni et al. (2014)
- Debiasing methods — Bühlmann (2013); Javanmard & Montanari (2014); Zhang & Zhang (2014)

Inference in the sub-model $\mu = X(A_0)\beta(A_0)$

- PoSI: simultaneous inference — Berk et al. (2013)
- Selective inference, FCR — Benjamini & Yekutieli (2005)
- Answer must be valid given that the question was asked
- Conditional approach: conditions the model selection event and uses corresponding truncated probability distributions

Literature on the conditional approach

Notable early mention: Hurvich & Tsai (1990)

- Lasso, sequential — Lockhart et al. (2014)
- General penalty, global null, geometry — Taylor et al. (2015); Azaïs et al. (2015)
- Forward stepwise, sequential — Loftus & Taylor (2014)
- Fixed λ Lasso / conditional — Lee et al. (2015); Fithian et al. (2014)
- Forward stepwise and LAR — Tibshirani et al. (2014)
- Unknown σ — Tian et al. (2015); Gross et al. (2015)
- Group selection / unknown σ — Loftus & Taylor (2015)
- Cross-validation — Tian & Taylor (2015); Loftus (2015)
- Unsupervised learning — Blier et al. (2016)

Selective inference

- Null hypothesis $H_0 = H_0(A_0)$ (depends upon model)

Selective type 1 (S1) error (Fithian et al., 2014)

$$\mathbb{P}_{A_0, H_0}(\text{reject } H_0(A_0) | A_0 \text{ selected})$$

- Simultaneous error control $\implies$ S1 control
- Data splitting controls S1 error (but less power, etc)
- Methods described in this talk control S1 error
- Greater replicability
- Select $A_0 = \hat{A}(y, X)$ (e.g. by forward stepwise, Lasso)

Selective inference

- Null hypothesis $H_0 = H_0(A_0)$ (depends upon model)

Selective type 1 (S1) error (Fithian et al., 2014)

$$\mathbb{P}_{A_0, H_0}(\text{reject } H_0(A_0) | A_0 \text{ selected})$$

- Simultaneous error control $\implies$ S1 control
- Data splitting controls S1 error (but less power, etc)
- Methods described in this talk control S1 error
- Greater replicability
- Select $A_0 = \hat{A}(y, X)$ (e.g. by forward stepwise, Lasso)

Selective inference

- Null hypothesis $H_0 = H_0(A_0)$ (depends upon model)

Selective type 1 (S1) error (Fithian et al., 2014)

$$\mathbb{P}_{A_0, H_0}(\text{reject } H_0(A_0) | A_0 \text{ selected})$$

- Simultaneous error control $\implies$ S1 control
- Data splitting controls S1 error (but less power, etc)
- Methods described in this talk control S1 error
- Greater replicability
- Select $A_0 = \hat{A}(y, X)$ (e.g. by forward stepwise, Lasso)

Selective inference

- Null hypothesis $H_0 = H_0(A_0)$ (depends upon model)

Selective type 1 (S1) error (Fithian et al., 2014)

$$\mathbb{P}_{A_0,H_0}(\text{reject } H_0(A_0) | A_0 \text{ selected})$$

- Simultaneous error control $\implies$ S1 control
- Data splitting controls S1 error (but less power, etc)
- Methods described in this talk control S1 error
- Greater replicability
- Select $A_0 = \hat{A}(y, X)$ (e.g. by forward stepwise, Lasso)

Section 2

Examples

Two examples

Model selection procedures characterized by quadratic inequalities:

Forward stepwise with groups of variables (e.g. factor models)

Using **cross-validation** to choose model complexity (sparsity)

Forward stepwise with groups

At each step, add the group which results in greatest decrease of RSS.
Regress that group out of the response and out of the remaining predictors

Important examples

- Factor models
- Categorical variables
- Application: modeling flexibility e.g. GLINTERNET or GAMSEL, see Loftus & Taylor (2014)
- Application: heterogeneous subgroup treatment effects

Forward stepwise with groups

At each step, add the group which results in greatest decrease of RSS.
Regress that group out of the response and out of the remaining predictors

Important examples

- Factor models
- Categorical variables
- Application: modeling flexibility e.g. GLINTERNET or GAMSEL, see Loftus & Taylor (2014)
- Application: heterogeneous subgroup treatment effects

Heterogeneous treatment effects

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

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

Upcoming work with Mike Baiocchi: analyze power and important causal inference properties of this approach

Tracking selection event

Event characterizing step 1 of forward stepwise

$$\text{RSS}(j_1) \leq \text{RSS}(j) \quad \forall j \neq j_1$$

or

$$y^T [X_{j_1} X_{j_1}^\dagger - X_j X_j^\dagger] y \geq 0 \quad \forall j \neq j_1$$

For P groups of variables, $P - 1$ inequalities.
Step 2: $P - 2$, etc.
Computation $O(S^2 P)$ where $S \ll P$ is number of steps.

Can use AIC/BIC/RIC to choose model size

In `selectiveInference` R package, `groupfs` function with default behavior like `step`

Tracking selection event

Event characterizing step 1 of forward stepwise

$$\text{RSS}(j_1) \leq \text{RSS}(j) \quad \forall j \neq j_1$$

or

$$y^T [X_{j_1} X_{j_1}^\dagger - X_j X_j^\dagger] y \geq 0 \quad \forall j \neq j_1$$

For P groups of variables, $P - 1$ inequalities.
Step 2: $P - 2$, etc.
Computation $O(S^2 P)$ where $S \ll P$ is number of steps.

Can use AIC/BIC/RIC to choose model size

In `selectiveInference` R package, `groupfs` function with default behavior like `step`

Tracking selection event

Event characterizing step 1 of forward stepwise

$$\text{RSS}(j_1) \leq \text{RSS}(j) \quad \forall j \neq j_1$$

or

$$y^T [X_{j_1} X_{j_1}^\dagger - X_j X_j^\dagger] y \geq 0 \quad \forall j \neq j_1$$

For P groups of variables, $P - 1$ inequalities.
Step 2: $P - 2$, etc.
Computation $O(S^2 P)$ where $S \ll P$ is number of steps.

Can use AIC/BIC/RIC to choose model size

In `selectiveInference` R package, `groupfs` function with default behavior like `step`

Simulation: $n = 100$, $p = 100$, $P = 50$, sparsity $= 4$

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

Summary of forward stepwise with groups

Use classical test statistics for χ^2 or F tests of groups of variables in regression models with two additional modifications:

- Condition on the selected model by constraining the distributions to the quadratic region
- Also condition on some additional information to reduce computation to one-dimension

JL; Taylor, J. Selective inference in regression models with groups of variables. arXiv Preprint. (2015)

Summary of forward stepwise with groups

Use classical test statistics for χ^2 or F tests of groups of variables in regression models with two additional modifications:

- Condition on the selected model by constraining the distributions to the quadratic region
- Also condition on some additional information to reduce computation to one-dimension

JL; Taylor, J. Selective inference in regression models with groups of variables. arXiv Preprint. (2015)

Summary of forward stepwise with groups

Use classical test statistics for χ^2 or F tests of groups of variables in regression models with two additional modifications:

- Condition on the selected model by constraining the distributions to the quadratic region
- Also condition on some additional information to reduce computation to one-dimension

JL; Taylor, J. Selective inference in regression models with groups of variables. arXiv Preprint. (2015)

Cross-validation

One of the most commonly used techniques for choosing model complexity–e.g., number of steps in forward stepwise or value of penalty parameter λ for Lasso

Previous selective inference work, e.g. Lee et al. (2015), assumes λ fixed independently of the data

Allowing use of cross-validation to choose λ or number of steps greatly widens applicability

Notation for CV

- 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.

Examples

For each training set...

- Train LASSO models on a grid of λ values. Or fit sequentially with GLMNET. Choose λ^* minimizing cv-RSS. Finally, fit a LASSO model at λ^* on the whole data.
- 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?

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)

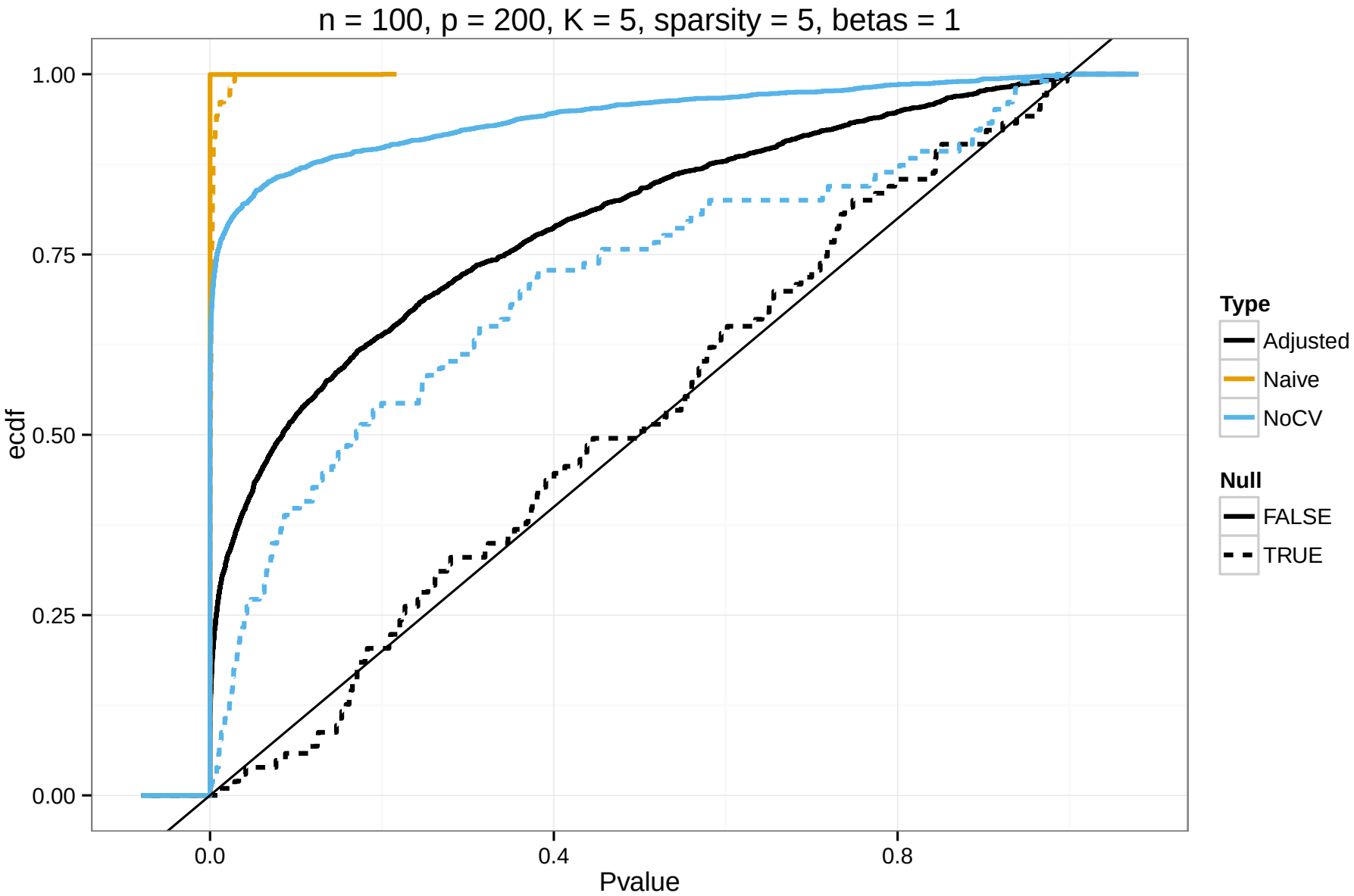
Define $Q^s_{ff} := \sum_{g \neq f} (P_{g,s})^T_f (P_{g,s})_f$ and

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

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


n = 100, p = 200, K = 5, sparsity = 5, betas = 1
ecdf
1.00
0.75
0.50
0.25
0.00
0.0
0.4
0.8
Pvalue
Type
Adjusted
Naive
NoCV
Null
FALSE
TRUE

Extension

We used forward stepwise as an example but this can also be done for the Lasso

Software implementation in a future version of `selectiveInference` R package

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

Section 3

Quadratic model selection events

Previous work: affine framework

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 *without* groups of variables

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

MVN constrained to a polytope.

Quadratic model selection framework

For many other M (e.g. forward stepwise with groups, cross-validation), event can be decomposed as

Quadratic selection event

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

Important: these Q, a, b are constant on E_m

Conditioning on $E_m \iff$ finding intersection of quadratics

Quadratic model selection framework

For many other M (e.g. forward stepwise with groups, cross-validation), event can be decomposed as

Quadratic selection event

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

Important: these Q, a, b are constant on E_m

Conditioning on $E_m \iff$ finding intersection of quadratics

Quadratic model selection framework

For many other M (e.g. forward stepwise with groups, cross-validation), event can be decomposed as

Quadratic selection event

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

Important: these Q, a, b are constant on E_m

Conditioning on $E_m \iff$ finding intersection of quadratics

Reduction to one-dimension

Test statistics have form $T^2 = \|Py\|_2^2$.

Write $y = Py + z$ where $z = (I - P)y$, $u = Py/\|Py\|_2$

Find $S_m = \{t : tu + z \in E_m\}$

One-dimensional support

Conditional on E_m **and** z, u, only remaining variation is $T = \|Py\|_2$ which has truncated support S_m.

Under $P\mu = 0$ (null hyp.), T is a χ-r.v. independent of u, z ($\sigma = 1$).

Finding intersection now reduced to solving many univariate quadratics in t, forming intersection of their positive level sets

Geometry: intersection of quadratic regions

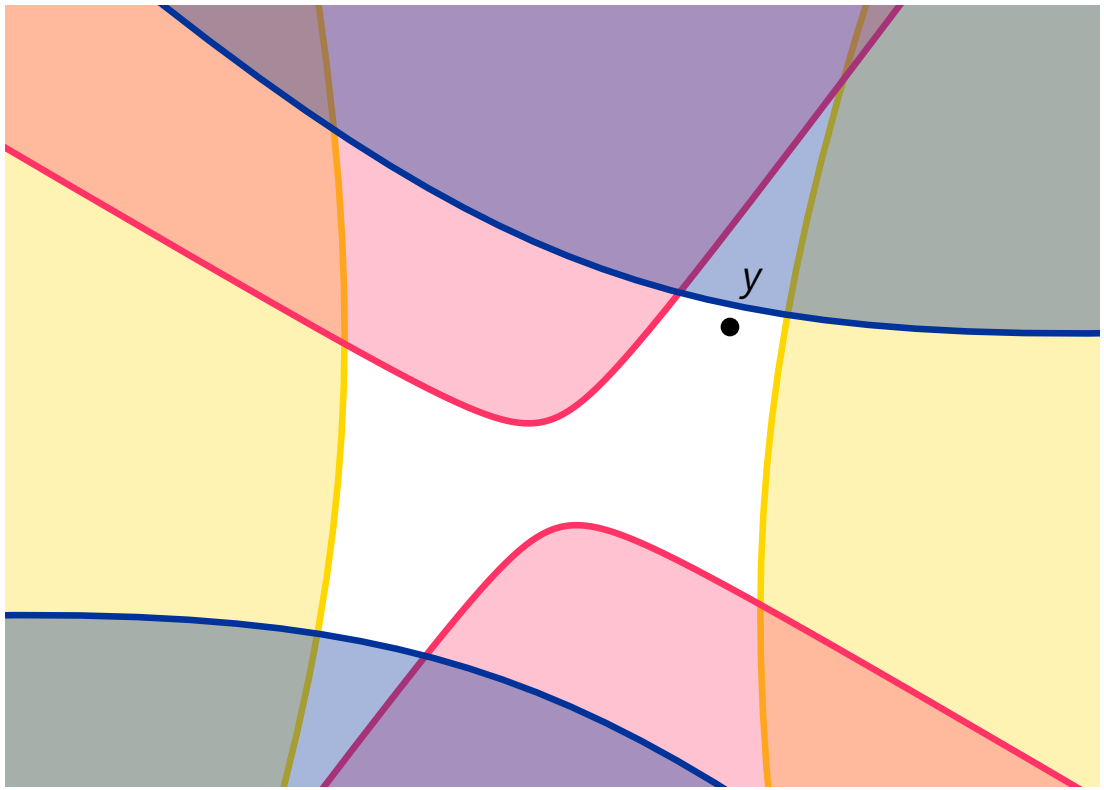


Figure: The *complement* of each quadratic is shaded with a different color. The unshaded, white region is E_m.

Geometry: intersection of quadratic regions

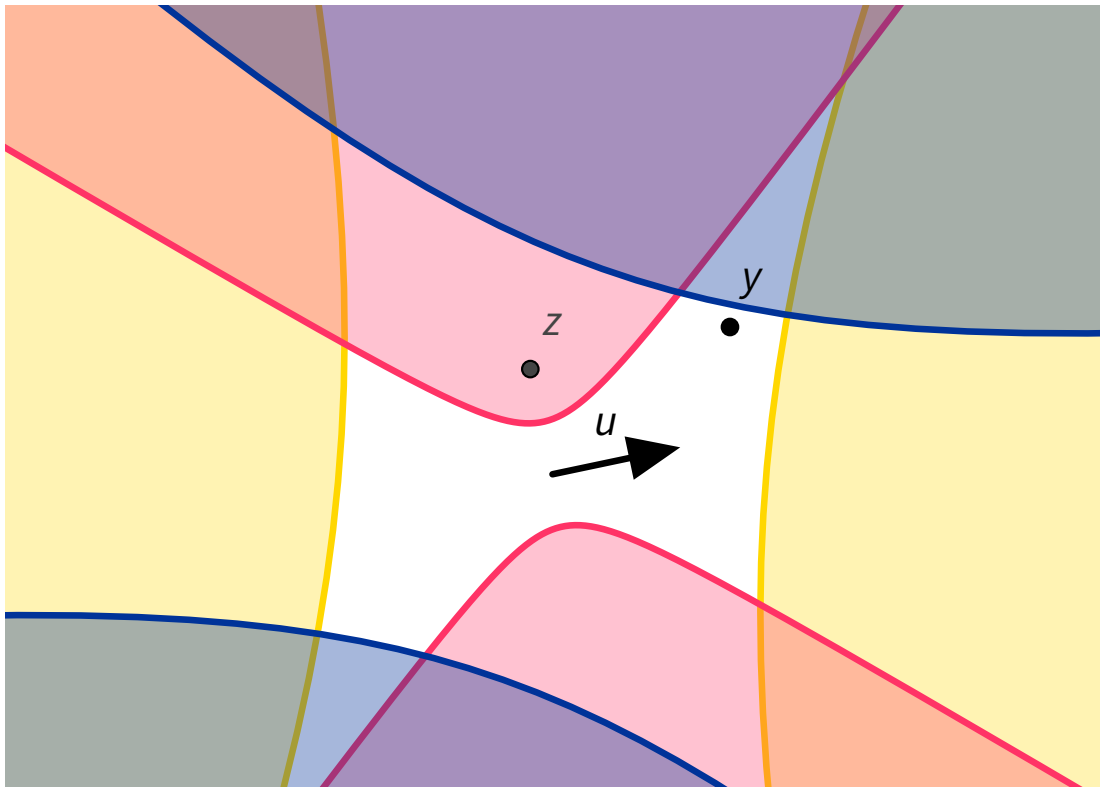


Figure: The *complement* of each quadratic is shaded with a different color. The unshaded, white region is E_m.

Geometry: intersection of quadratic regions

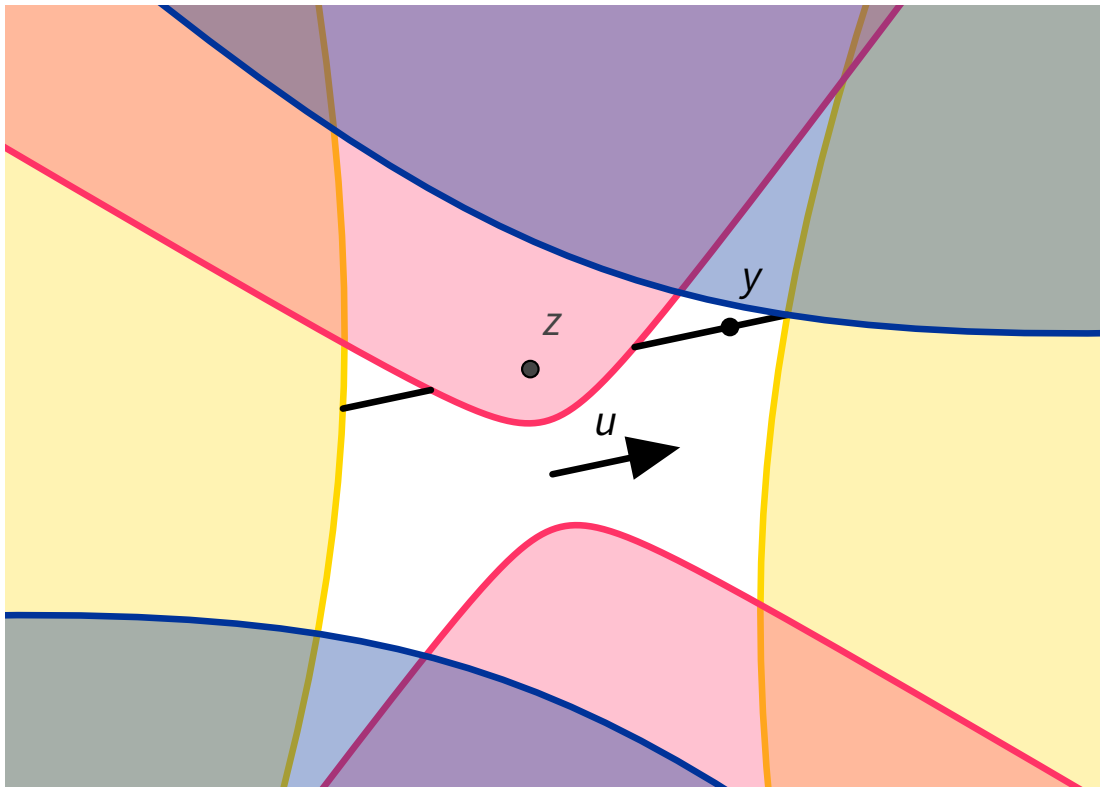


Figure: The *complement* of each quadratic is shaded with a different color. The unshaded, white region is E_m.

Geometry: intersection of quadratic regions

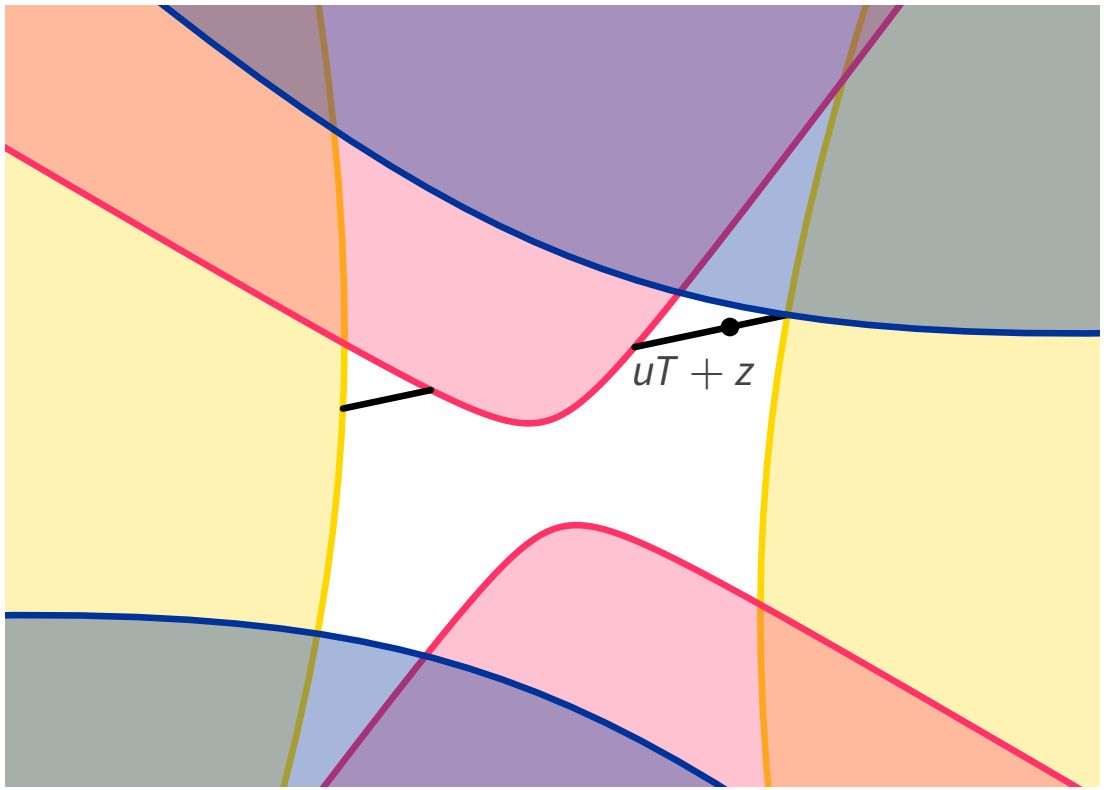


Figure: The *complement* of each quadratic is shaded with a different color. The unshaded, white region is E_m.

Section 4

Conclusion and future work

Final remarks

- R package `selectiveInference` by Rob, Ryan Tibshirani, Stephen Reid, and JL available on CRAN
- Drawback: computationally expensive
- Can be done in parallel.
- Developed custom numerical method for accurate and robust computation of truncated support for selective F tests. Details in paper
- Future work: k-means, forthcoming with Léonard Blier and Jonathan Taylor

The end

Thanks for your attention!

For more info: statweb.stanford.edu/~joftius

Section 5

References

Azaïs, J.-M., de Castro, Y., & Mourareau, S. (2015). Power of the kac-rice detection test. *arXiv preprint arXiv:1503.05093*.

Barber, R. F. & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.*, *43*(5), 2055–2085.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, *41*(2), 802–837.

Blier, L., Loftus, J. R., & Taylor, J. E. (2016). Inference on the number of clusters in *k*-means clustering. *In progress*.

Bogdan, M., Berg, E. v. d., Sabatti, C., Su, W., & Candes, E. J. (2014). Slope–adaptive variable selection via convex optimization. *arXiv preprint arXiv:1407.3824*.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, *19*(4), 1212–1242.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, *106*(496), 1602–1614.

Fithian, W., Sun, D., & Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Gross, S. M., Taylor, J., & Tibshirani, R. (2015). A Selective Approach to Internal Inference. *ArXiv e-prints*.

Ioannidis, J. P. (2005). Why most published research findings are false. *Chance*, *18*(4), 40–47.

Javanmard, A. & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, *15*(1), 2869–2909.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2015). Exact post-selection inference with the lasso. *Ann. Statist.* To appear.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *Annals of statistics*, *42*(2), 413.

Loftus, J. R. (2015). Selective inference after cross-validation. *ArXiv e-prints*.

Loftus, J. R. & Taylor, J. E. (2014). A significance test for forward stepwise model selection. *arXiv preprint arXiv:1405.3920*.

Loftus, J. R. & Taylor, J. E. (2015). Selective inference in regression models with groups of variables. *ArXiv e-prints*.

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Taylor, J. E., Loftus, J. R., & Tibshirani, R. J. (2015). Tests in adaptive regression via the kac-rice formula. *Ann. Statist.* To appear.

Tian, X., Loftus, J. R., & Taylor, J. E. (2015). Selective inference with unknown variance via the square-root lasso. *arXiv preprint arXiv:1504.08031*.

Tian, X. & Taylor, J. E. (2015). Selective inference with a randomized response. *arXiv preprint arXiv:1507.06739*.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2014). Exact Post-Selection Inference for Sequential Regression Procedures. *ArXiv e-prints*.

University of Wisconsin Population Health Institute (2015). California county health rankings. `http://www.countyhealthrankings.org`. Accessed: 2015-10-28.

Wasserman, L. & Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, *37*(5A), 2178.

Zhang, C.-H. & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 217–242.