Inference Conditional on Model Selection with a Focus on Procedures Characterized by Quadratic Inequalities

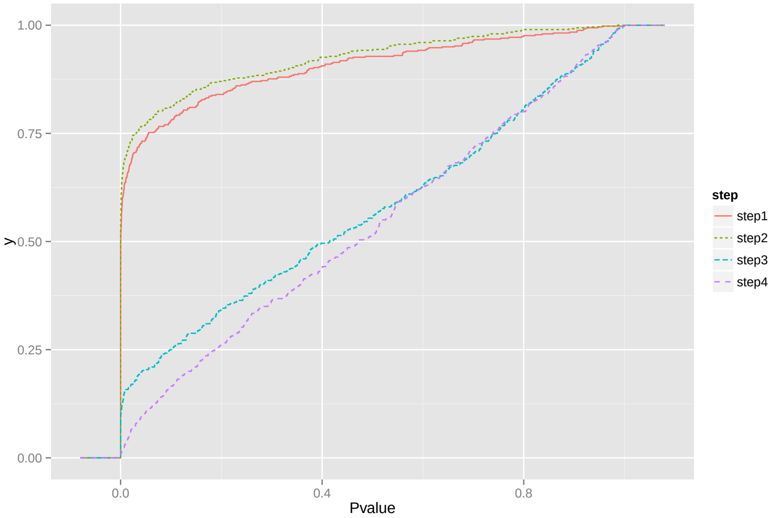
Joshua R. Loftus

Outline

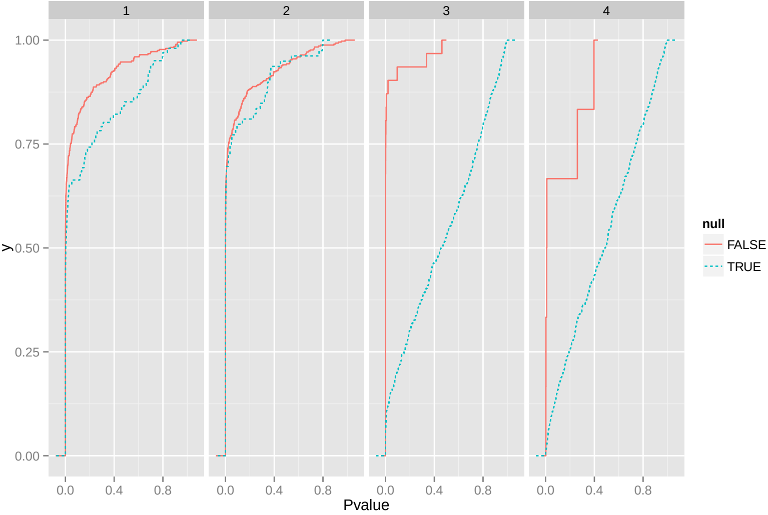
First the results (forward stepwise with groups)

- Simulation setup: $X_{20\times 200}$ with x_{ij} i.i.d. $N(0,1)$
- Predictors are grouped *a priori*. Groups of size 2, the first and second columns, the third and fourth, etc.
- $Y = X\beta + \epsilon$ with ϵ_i i.i.d. $N(0,1)$
- $\beta_1 = \cdots = \beta_4 = 2$, $\beta_i = 0$ for $i > 4$. I.e. group sparsity is 2.
- Normalize X so groups have Frobenius norm 1.
- Run forward stepwise for 4 steps. Each step adds a group–two columns–to the model.
- Compute our $T\chi$ statistic, apply appropriate (under the null) CDF transform.
- Repeat for 500 realizations.

ECDF of Pvalues in forward stepwise, by step



Null vs. non-null added variable



How did we adjust the classical χ^2 test?

- Even if the sparsity was 0 (global null), the first 4 steps would give *p*-values corresponding to the largest 4 out of 100 χ^2 statistics.
- In the same situation our $T\chi$ *p*-values would be uniform (correct).
- Let's review the backstory (briefly)

Inference with LASSO model selection

The **covariance test**

- Test with 1 degree of freedom corresponding to 1 variable entering the model
- Knots in LASSO path r.v. containing useful information
- Issue: asymptotics break down for groups of variables
- Pioneering work, new approach to combining inference with model selection

Lockhart, R.; Taylor, J.; Tibshirani, R. J.; Tibshirani, R. A significance test for the lasso. Ann. Statist. 42 (2014)

Applying geometry of random fields

Enter Geometry: the **Kac-Rice test**

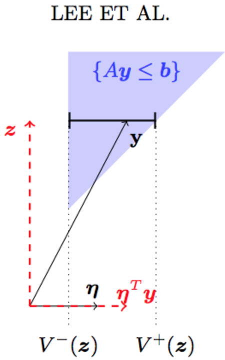
- Extend covariance test to group lasso
- Successful for first knot, lead to some new directions
- Non-asymptotic, **truncated** distributions instead of Exp(1)
- Generalized to a global null hypothesis test for a large class of penalized regression problems (including, e.g. matrix completion)
- Power not well-understood (Azaïs et al, 2015)

Taylor, J.; Loftus, J.; Tibshirani, R. J. Tests in adaptive regression via the Kac-Rice formula. arXiv Preprint. (2014)

Affine model selection events

Inference **conditional** on selection

- Focus on fixed-λ LASSO (non-sequential)
- General framework for model selection events characterized by affine inequalities
- Inference conditional on selected model. Rigorous, interpretable hypotheses



Lee, J.; Sun, D.; Sun, Y.; Taylor, J. Exact post-selection inference, with application to the lasso. arXiv Preprint. (2015)

Selective inference and optimality

The **exponential family** context

- UMPU tests by applying Lehmann-Scheffé
- Define *selective type I error*, the error criterion we want to control
$$\mathbb{P}_{M,H_0}(\text{reject } H_0 | (M, H_0) \text{ selected})$$
- Language and notation: model-hypothesis pairs

Fithian, W.; Sun, D.; Taylor, J. Optimal Inference After Model Selection. arXiv Preprint. (2014)

The goal of this work

- Our choice is to control the selective type I error
- Previous work established how to do this with affine model selection events
- Can we extend it to some non-affine examples, e.g. quadratic?

Background: the affine framework

Let $M : \mathbb{R}^n \to \mathcal{M}$ be a model selection map, with $\mathcal{M}$ denoting a space of potential models. We observe $\{M(\mathbf{y}) = m\}$ and wish to condition on this event.

For many model selection procedures (forward stepwise, LAR, LASSO, elastic net, marginal screening...) the event $\{M(\mathbf{y}) = m\}$ can be written as $\{\mathbf{A}(m)\mathbf{y} \leq \mathbf{b}(m)\}$ for some $(\mathbf{A}, \mathbf{b})$ which depend on $\mathbf{y}$ *only through* m. What does this give us?

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

MVN constrained to a polytope.

Forward stepwise example, first step

A forward stepwise model after s steps will have the form $m_s = (A_s, u_s)$ with A_s an ordered set of active variables and u_s their signs (included for computational purposes).

Denoting $\mathbf{x}_j^1 = \mathbf{X}_j / \|\mathbf{X}_j\|$, at the first step

$$u_1 (\mathbf{x}_{j_1}^1)^T \mathbf{y} \geq \pm (\mathbf{x}_j^1)^T \mathbf{y} \quad \forall j \neq j_1$$

So $-\mathbf{A}(m_1)$ has $2(p-1)$ rows of the form $u_1(\mathbf{x}_{j1}^1)^T \pm (\mathbf{x}_j^1)^T$ and $\mathbf{b}(m_1) = 0$.

TLTT. Exact Post-selection Inference for Forward Stepwise and Least Angle Regression. arXiv Preprint. (2014)

Forward stepwise example, later steps

Denoting $\mathbf{x}_j^s = \mathbf{P}_{A_{s-1}}^{\perp} \mathbf{X}_j / \|\mathbf{P}_{A_{s-1}}^{\perp} \mathbf{X}_j\|$, with $\mathbf{P}_{A_{s-1}}^{\perp}$ the projection orthogonal to $\mathbf{X}_{A_{s-1}}$, we have at the sth step

$$(u_s \mathbf{x}_{j_s}^s \pm \mathbf{x}_j^s)^T \mathbf{y} \geq 0 \quad \forall j \notin A_s$$

Form $\mathbf{A}(m_s)$ by appending the corresponding $2(p-s)$ rows to $\mathbf{A}(m_{s-1})$, so $\{M_s(\mathbf{y}) = m_s\} = \{\mathbf{z} : \mathbf{A}(m_s)\mathbf{z} \geq 0\}$.

Inference from conditional law, if we can sample from the constrained MVN.

Quadratic events: forward stepwise with groups

With predetermined groups of variables FS can add entire groups in each step. At the first step, $\mathbf{X}_{j_1}$ is a submatrix with ≥ 1 columns. Now the event that j_1 is the first group is equivalent to

$$\|(\mathbf{I} - \mathbf{X}_{j_1}\mathbf{X}_{j_1}^{\dagger})\mathbf{y}\|_2^2 \leq \|(\mathbf{I} - \mathbf{X}_{j}\mathbf{X}_{j}^{\dagger})\mathbf{y}\|_2^2 \quad \forall j \neq j_1$$

or

$$\mathbf{y}^T[\mathbf{X}_{j_1}\mathbf{X}_{j_1}^{\dagger} - \mathbf{X}_{j}\mathbf{X}_{j}^{\dagger}]\mathbf{y} \geq 0 \quad \forall j \neq j_1$$

Problem

Model selection is no longer linear in $\mathbf{y}$.
(Drop signs from definition of model).

Loftus, J.; Taylor, J. A significance test for forward stepwise model selection. arXiv preprint (2014). Acknowledgement: Léonard Blier

Is this problem important?

- Interpretation with factor models and categorical variables, e.g. inference about location on genome vs. inference about one particular variant
- Hierarchical models via overlapping groups, e.g. GLINTERNET

$$\mathbf{X} = \left[\mathbf{X}_1 \quad \middle| \quad \mathbf{X}_2 \quad \middle| \quad \mathbf{X}_1 \quad \mathbf{X}_2 \quad \mathbf{X}_{1:2}\right]$$

- Mathematically interesting: proofs can involve non-trivial applications of random field theory

Quadratic framework

Abstractly, we want to consider problems where

$$\{M(\mathbf{y}) = m\} = \{\mathbf{y}^T \mathbf{Q}(m)\mathbf{y} + \mathbf{A}(m)\mathbf{y} \leq \mathbf{b}(m)\}$$

Issues to address case-by-case

- Getting a model selection procedure into this form (if possible)
- Which things to condition on (power vs. computation)
- Form of test statistic, interpretation of hypotheses
- Complicated geometry of selection region: how to sample?

Forward stepwise with groups, step 1

Before we introduce randomness, imagine we just have some fixed ordering of the groups, with j_1 the first in this order.

Denote $\mathbf{P}^1_j = \mathbf{X}_j \mathbf{X}_j^\dagger$. Define the "event" that j_1 is the first group as

$$E_\emptyset^{j_1} = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z}^T [\mathbf{P}^1_{j_1} - \mathbf{P}^1_j] \mathbf{z} \geq k \, \mathrm{Tr}(\mathbf{P}^1_{j_1} - \mathbf{P}^1_j) \quad \forall j \neq j_1\}$$

Using AIC or BIC to compare groups of different size determines a multiplier k of the penalty on model size.

Forward stepwise with groups, step s

With $\mathbf{P}^{\perp}_{A_{s-1}}$ the projection orthogonal to $\mathbf{X}_{A_{s-1}}$, denote $\mathbf{X}^s_j = \mathbf{P}^{\perp}_{A_{s-1}}\mathbf{X}_j$ and $\mathbf{P}^s_j = \mathbf{X}^s_j(\mathbf{X}^s_j)^{\dagger}$. Now at step s, if j_s is the next group we must append the conditions

$$E^{j_s}_{A_{s-1}} = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z}^T[\mathbf{P}^s_{j_1} - \mathbf{P}^s_j]\mathbf{z} \geq k\,\mathrm{Tr}(\mathbf{P}^s_{j_1} - \mathbf{P}^s_j) \quad \forall j \notin A_s\}$$

Intersection of $p - s$ quadratic inequalities.

Model selection = intersection of quadratic inequalities

Consider the event E that a random $\mathbf{y}$ yields the active set $A_s = \{j_1, \ldots, j_s\}$. Denote $A_0 = \emptyset$, and $A_l = \{j_1, \ldots, j_l\}$ the initial segment with the same order for any $l \leq s$.

Characterization of model selection

$$E := \{M(\mathbf{y}) = A_s\} = \bigcap_{l=1}^{s} E_{A_{l-1}}^{j_l}$$

The selection event is the intersection of a list of quadratic inequalities. Depends on $\mathbf{y}$ only through A_s.

Inference conditional on selection

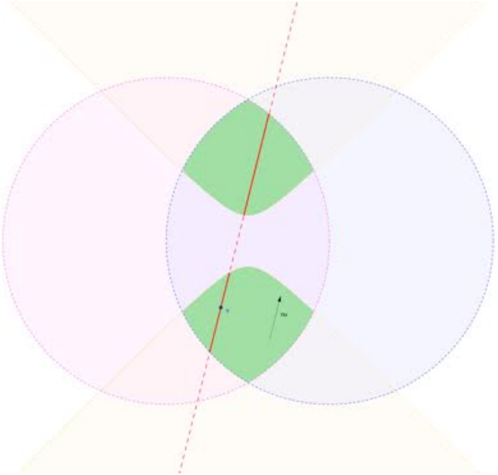
E is the support of $\mathcal{L}(\mathbf{y}|M(\mathbf{y}) = A_s)$. Challenging to sample from. Closed form results for statistics supported on one-dimensional slices through E.

Test based on $S = \|\mathbf{Py}\|_2$ for some projection $\mathbf{P}$ (think of χ statistics). Let $\mathbf{U} = \mathbf{Py}/S$ and $\mathbf{Z} = \mathbf{y} - \mathbf{Py}$.

Truncation interval $M_S = \{t \geq 0 : M(\mathbf{U}t + \mathbf{Z}) = A_s\}$. Quadratic model selection $\to$ univariate quadratics in t.

M_S is the support of $\mathcal{L}(S|A_s, \mathbf{Z})$.

Not your father's polyhedral lemma



Inference for the active groups

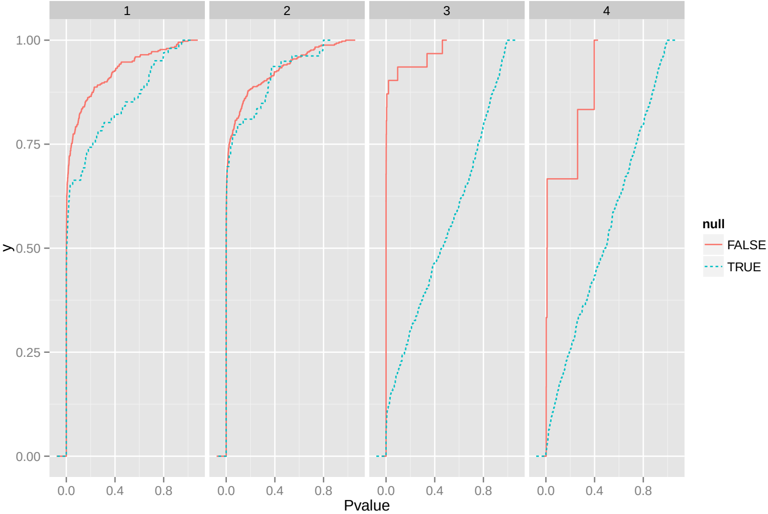
Tests for groups in A_s proceed the usual way (χ^2 test in regression). Let $\mathbf{P}_j$ be the projection for adding group j to $A_s \backslash j$, and $S = \|\mathbf{P}_j \mathbf{y}\|_2$.

Unconditionally, S is a χ-r.v., and conditionally it is just truncated to M_S. Known degrees of freedom and scale.

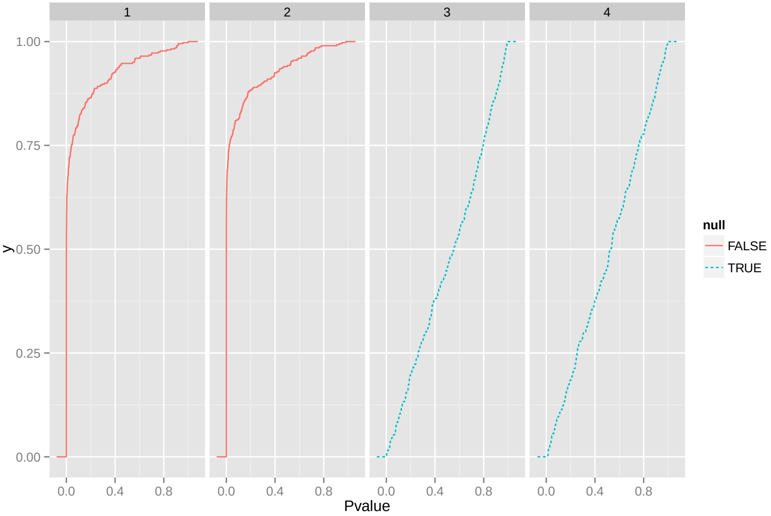
Assumptions: "saturated" model with known σ

Model assumption: $\mathbf{y} \sim N(\mu, \sigma^2 I)$ with σ^2 known.
Null hypothesis: $\mathbf{P}\mu = 0$.

Simulation results again: when does the null hold?



Conditional on correct selection



Final remarks

- **R package**: Coming soon to a CRAN near you (with Rob and Ryan Tibshirani).
- **To do**: unknown σ case, sampling from selected model (more powerful), univariate tests/intervals.
- **Drawback**: computationally expensive, after s steps we must store about $s(p-1)$ projections, each is $n \times n$. Multiply all of these by s vectors.
- Can be done in parallel.
- Another quadratic model selection procedure: k-means. Forthcoming work with Léonard Blier and Jonathan Taylor.

Recap & conclusion

- Selective inference for designs with groups of variables.
- Interpretable model selection, corrected version of classical regression tests.
- Computationally expensive.

Recap & conclusion

- Selective inference for designs with groups of variables.
- Interpretable model selection, corrected version of classical regression tests.
- Computationally expensive.

- Thanks for your attention! I owe you a coffee.
- Questions?
- (I'll be on the job market this year!)