"PoSI" and its Geometry

Andreas Buja

joint work with

Richard Berk, Lawrence Brown, Kai Zhang, Linda Zhao

Department of Statistics, The Wharton School
University of Pennsylvania
Philadelphia, USA

Simon Fraser 2015/04/20

Larger Problem: Non-Reproducible Empirical Findings

- Indicators of a problem
  (from: Berger, 2012, "Reproducibility of Science: P-values and Multiplicity")
  - Bayer Healthcare reviewed 67 in-house attempts at replicating findings in published research:
    $< 1/4$ were viewed as replicated.
  - Arrowsmith (2011, Nat. Rev. Drug Discovery 10):
    Increasing failure rate in Phase II drug trials
  - Ioannidis (2005, PLOS Medicine):
    "Why Most Published Research Findings Are False"
  - Simmons, Nelson, Simonsohn (2011, Psychol.Sci):
    "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant"
- Many potential causes – two major ones:
  - publication bias: "file drawer problem" (Rosenthal 1979)
  - statistical biases: "researcher degrees of freedom" (SNS 2011)

Statistical Biases – one among several

- Hypothesis: A statistical bias is due to
  an absence of accounting for model/variable selection.
- Model selection is done on several levels:
  - formal selection: AIC, BIC, Lasso, ...
  - informal selection: residual plots, influence diagnostics, ...
  - post hoc selection: "The effect size is too small in relation to the cost of data collection to warrant inclusion of this predictor."
- Suspicions:
  - All three modes of model selection may be used in much empirical research.
  - Ironically, the most thorough and competent data analysts may also be the ones who produce the most spurious findings.
  - If we develop valid post-selection inference for "adaptive Lasso", say, it won't solve the problem because few empirical researchers would commit themselves a priori to one formal selection method and nothing else.
    $\Rightarrow$ "Meta-Selection Problem"

Example: Length of Criminal Sentence

Question: What covariates predict length of a criminal sentence best?

A small empirical study:

- $N = 250$ observations.
- Response: `log-length of sentences`
- $p = 11$ covariates (predictors, explanatory variables):
  - `race`
  - `gender`
  - `initial age`
  - `marital status`
  - `employment status`
  - `seriousness of crime`
  - `psychological problems`
  - `education`
  - `drug related`
  - `alcohol usage`
  - `prior record`
- What variables should be included?

Example: Length of Criminal Sentence (contd.)

- All-subset search with BIC chooses a model $\hat{\mathbb{M}}$ with seven variables:
  - `initial age`
  - `gender`
  - `employment status`
  - `seriousness of crime`
  - `drugs related`
  - `alcohol usage`
  - `prior records`
- t-statistics of selected covariates, in descending order:
  - $|t_{\texttt{alcohol}}| = 3.95$;
  - $|t_{\texttt{prior records}}| = 3.59$;
  - $|t_{\texttt{seriousness}}| = 3.57$;
  - $|t_{\texttt{drugs}}| = 3.31$;
  - $|t_{\texttt{employment}}| = 3.04$;
  - $|t_{\texttt{initial age}}| = 2.56$;
  - $|t_{\texttt{gender}}| = 2.33$.
- Can we use the cutoff $t_{.975,250-8} = 1.96$?

Linear Model Inference and Variable Selection

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{X}$ = fixed design matrix, $N \times p$, $N > p$, full rank.
- $\epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$

In textbooks:

1. Variables selected
2. Data seen
3. Inference produced

In common practice:

1. Data seen
2. Variables selected
3. Inference produced

Linear Model Inference and Variable Selection

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{X}$ = fixed design matrix, $N \times p$, $N > p$, full rank.
- $\epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$

In textbooks:

1. Variables selected
2. Data seen
3. Inference produced

In common practice:

1. Data seen
2. Variables selected
3. Inference produced

Is this inference valid?

The PoSI Procedure — Rough Outline

- We propose to construct Post Selection Inference (PoSI) with guarantees for the coverage of CIs and Type I errors of tests.
- We widen CIs and retention intervals to achieve correct/conservative post-selection coverage probabilities. This is the price we have to pay.
- The approach is a reduction of PoSI to simultaneous inference.
- Simultaneity is across all submodels and all slopes in them.
- As a result, we obtain

  valid PoSI for all variable selection procedures!

- But first we need some preliminaries on

  Targets of Inference and Inference in Wrong Models

Submodels — Notation, Parameters, Assumptions

- Denote a submodel by the integers $\mathrm{M} = \{j_1, j_2, ..., j_m\}$ for the predictors:

$$\mathbf{X}_\mathrm{M} = (\mathbf{X}_{j_1}, \mathbf{X}_{j_2}, ..., \mathbf{X}_{j_m}) \in \mathbb{R}^{N \times m}.$$

- The LS estimators in the submodel M are

$$\hat{\beta}_\mathrm{M} = (\mathbf{X}_\mathrm{M}^T \mathbf{X}_\mathrm{M})^{-1} \mathbf{X}_\mathrm{M}^T \mathbf{Y} \in \mathbb{R}^m$$

- What does $\hat{\beta}_\mathrm{M}$ estimate?
  A: Its expectation — i.e., we ask for unbiasedness.

$$\mu := \mathbf{E}[\mathbf{Y}] \in \mathbb{R}^N \quad \textbf{arbitrary!!}$$

$$\beta_\mathrm{M} := \mathbf{E}[\hat{\beta}_\mathrm{M}] = (\mathbf{X}_\mathrm{M}^T \mathbf{X}_\mathrm{M})^{-1} \mathbf{X}_\mathrm{M}^T \mu$$

‘Slopes depend on ...’ — An Illustration

- Survey of potential purchasers of a new high-tech gizmo:
  - Response: “LoP” = Likelihood of Purchase (self-reported on a Likert scale)
  - Predictor 1: Age
  - Predictor 2: Income
- Expectation: Younger customers have higher LoP, that is, $\beta_{\texttt{Age}} < 0$.
- Outcome of the analysis:
  - Against expectations, a regression of LoP on Age alone indicates that older customers have higher LoP: $\beta_{\texttt{Age}} > 0$
  - But a regression of LoP on Age and Income indicates that, adjusted for Income, younger customers have higher LoP: $\beta_{\texttt{Age}\bullet\texttt{Income}} < 0$
  - Enabling factor: (partial) collinearity between Age and Income.
- A case of Simpson’s paradox: $\beta_{\texttt{Age}} > 0 > \beta_{\texttt{Age}\bullet\texttt{Income}}$.
  - The marginal and the Income-adjusted slope have very different values and different meanings.

Adjustment, Estimates, Parameters, t-Statistics

Notation and facts for the components of $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ and $\boldsymbol{\beta}_{\mathrm{M}}$, assuming $j \in \mathrm{M}$:

- Let $\mathbf{X}_{j \bullet \mathrm{M}}$ be the predictor $\mathbf{X}_j$ adjusted for the other predictors in M:

$$\mathbf{X}_{j \bullet \mathrm{M}} \;:=\; \left(\mathbf{I} - \mathbf{H}_{\mathrm{M} \setminus \{j\}}\right) \mathbf{X}_j \quad \perp \; \mathbf{X}_k \; \forall k \in \mathrm{M} \setminus \{j\}.$$

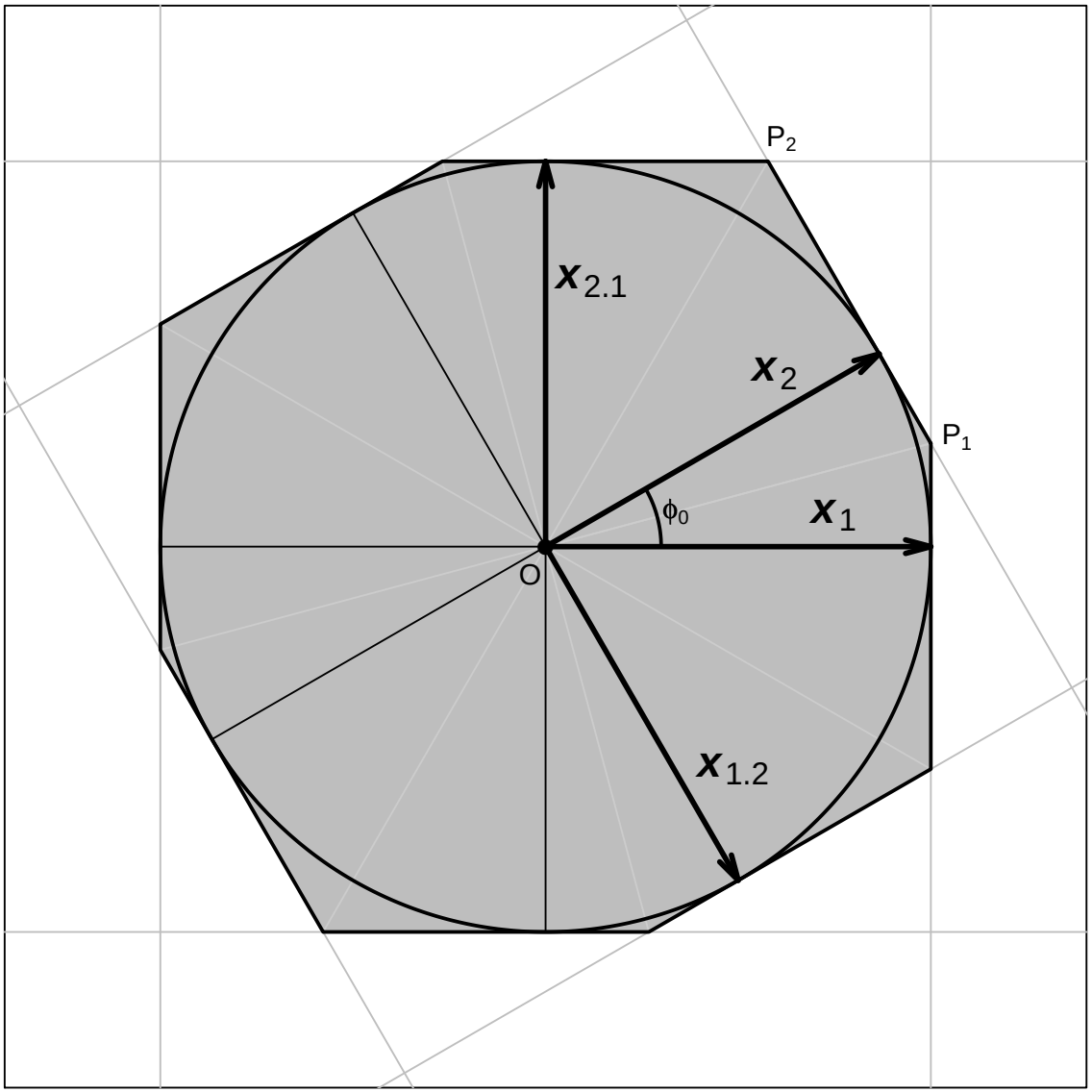
- Let $\hat{\beta}_{j \bullet \mathrm{M}}$ be the slope estimate and $\beta_{j \bullet \mathrm{M}}$ be the parameter for $\mathbf{X}_j$ in M:

$$\hat{\beta}_{j \bullet \mathrm{M}} \;:=\; \frac{\mathbf{X}_{j \bullet \mathrm{M}}^T \mathbf{Y}}{\|\mathbf{X}_{j \bullet \mathrm{M}}\|^2}\,, \qquad \beta_{j \bullet \mathrm{M}} \;:=\; \frac{\mathbf{X}_{j \bullet \mathrm{M}}^T \mathbf{E}[\mathbf{Y}]}{\|\mathbf{X}_{j \bullet \mathrm{M}}\|^2}.$$

- Let $t_{j \bullet \mathrm{M}}$ be the t-statistic for $\hat{\beta}_{j \bullet \mathrm{M}}$ and $\beta_{j \bullet \mathrm{M}}$:

$$t_{j \bullet \mathrm{M}} \;:=\; \frac{\hat{\beta}_{j \bullet \mathrm{M}} - \beta_{j \bullet \mathrm{M}}}{\hat{\sigma} / \|\mathbf{X}_{j \bullet \mathrm{M}}\|} \;=\; \frac{\mathbf{X}_{j \bullet \mathrm{M}}^T (\mathbf{Y} - \mathbf{E}[\mathbf{Y}])}{\|\mathbf{X}_{j \bullet \mathrm{M}}\| \, \hat{\sigma}} \;=\; \frac{\mathbf{X}_{j \bullet \mathrm{M}}^T}{\|\mathbf{X}_{j \bullet \mathrm{M}}\|} \, \frac{\boldsymbol{\epsilon}}{\hat{\sigma}}.$$

Geometry of Adjustment



Column space of $\mathbf{X}$ for $p = 2$ predictors, partly collinear

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathbb{M}} = \hat{\mathbb{M}}(\mathbf{Y}), \quad \mathbb{R}^N \to \mathcal{P}(\{1, \ldots p\})$

  - $\hat{\mathbb{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Facing up to post-selection inference: Confusers!
  - Target of Inference: the vector $\beta_{\hat{\mathbb{M}}(\mathbf{Y})}$, its components $\beta_{j \bullet \hat{\mathbb{M}}(\mathbf{Y})}$ for $j \in \hat{\mathbb{M}}(\mathbf{Y})$.
  - The target of inference is random.
  - The target of inference has a random dimension: $\beta_{\hat{\mathbb{M}}(\mathbf{Y})} \in \mathbb{R}^{|\hat{\mathbb{M}}(\mathbf{Y})|}$
  - Conditional on $j \in \hat{\mathbb{M}}$, the target component $\beta_{j \bullet \hat{\mathbb{M}}(\mathbf{Y})}$ has random meanings.
  - When $j \notin \hat{\mathbb{M}}$ both $\beta_{j \bullet \hat{\mathbb{M}}}$ and $\hat{\beta}_{j \bullet \hat{\mathbb{M}}}$ are undefined.
  - Hence the coverage probability $\mathbf{P}[\, \beta_{j \bullet \hat{\mathbb{M}}} \in \mathrm{CI}_{j \bullet \hat{\mathbb{M}}}(K) \,]$ is undefined.

Universal Post-Selection Inference

- Candidates for meaningful coverage probabilities:
  - $\mathbf{P}[\, j \in \hat{\mathrm{M}} \;\&\; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$ $\qquad (\leq \mathbf{P}[\, j \in \hat{\mathrm{M}}\,])$
  - $\mathbf{P}[\, \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K) \,\big|\, j \in \hat{\mathrm{M}}\,]$ $\qquad (\mathbf{P}[\, j \in \hat{\mathrm{M}}\,] = ???)$
  - $\mathbf{P}[\, \forall j \in \hat{\mathrm{M}} :\; \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\,]$

  All are meaningful; the last will be our choice.

- Overcoming the next difficulty:
  - Problem: None of the above coverage probabilities are known or can be estimated for most selection procedures $\hat{\mathrm{M}}$.
  - Solution: Ask for more!
    Universal Post-Selection Inference for all selection procedures is doable.

Reduction to Simultaneous Inference

Lemma

For any variable selection procedure $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$*, we have the following "significant triviality bound":*

$$\max_{j \in \hat{\mathrm{M}}} |t_{j \bullet \hat{\mathrm{M}}}| \;\leq\; \max_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j \bullet \mathrm{M}}| \qquad \forall\, \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^N.$$

Theorem

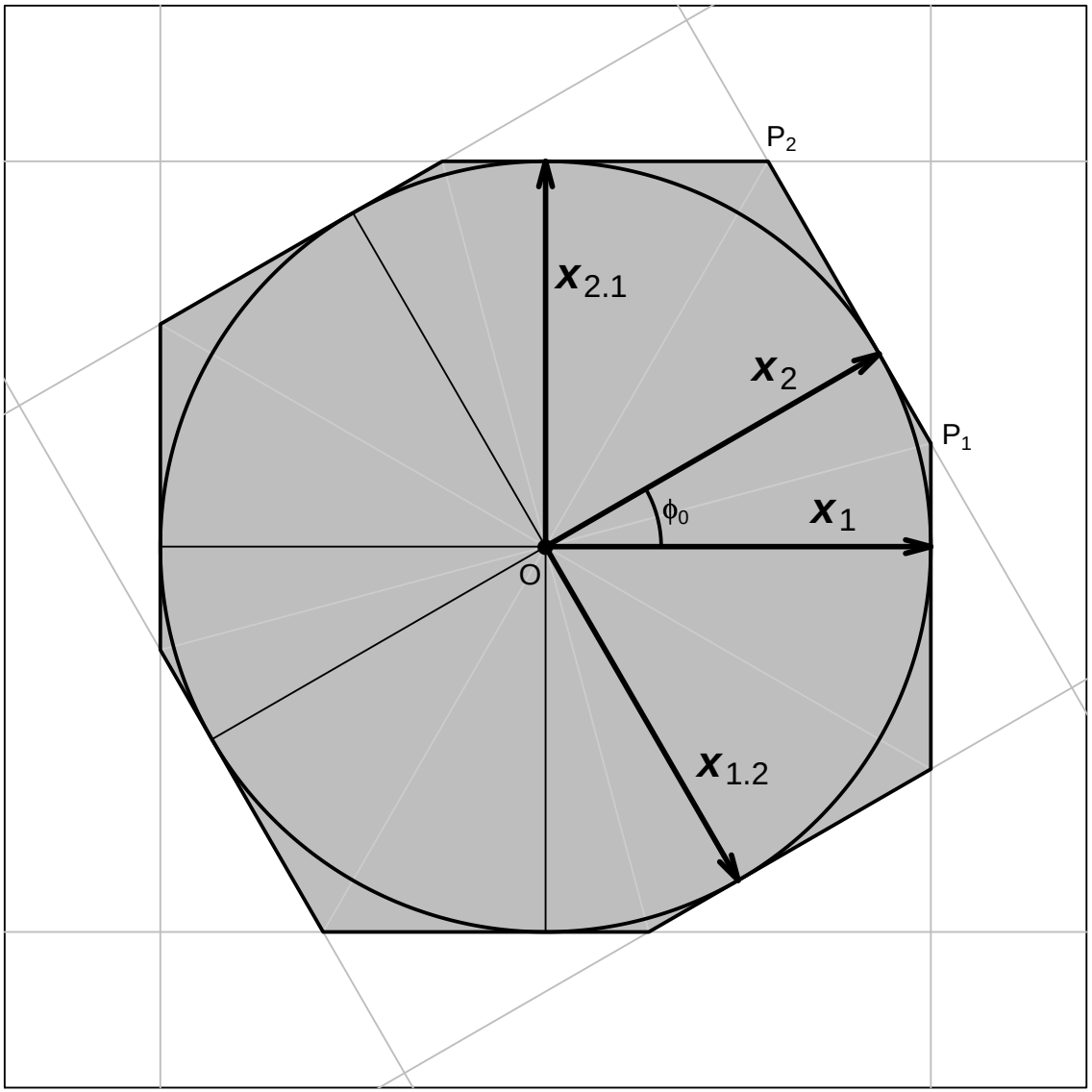
Let K *be the* $1-\alpha$ *quantile of the "max-max-*$|t|$*" statistic of the lemma:*

$$\mathbf{P}\Big[\; \max_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j \bullet \mathrm{M}}| \leq K \;\Big] \overset{(\geq)}{=} 1 - \alpha.$$

Then we have the following universal PoSI guarantee:

$$\mathbf{P}\Big[\; \beta_{j \bullet \hat{\mathrm{M}}} \in CI_{j \bullet \hat{\mathrm{M}}}(K) \;\; \forall j \in \hat{\mathrm{M}} \;\Big] \;\geq\; 1 - \alpha \qquad \forall \hat{\mathrm{M}}.$$

PoSI Geometry — Simultaneity



PoSI polytope
= intersection
of all *t*-bands.

How Conservative is PoSI?

- Is there a model selection procedure that requires full PoSI protection?
- Consider $\hat{\mathrm{M}}$ defined as follows:

$$\hat{\mathrm{M}} \;:=\; \mathrm{argmax}_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j \bullet \mathrm{M}}|$$

  A polite name: "Single Predictor Adjusted Regression" $=:$ SPAR
  A crude name: "Significance Hunting"
  a special case of "p-hacking" (Simmons, Nelson, Simonsohn 2011)

- SPAR requires the full PoSI protection — by construction!
- How realistic is SPAR in describing real data analysts behaviors?
  - It ignores the goodness of fit of the selected model.
  - It looks for the minimal achievable p-value / strongest "effect".

Computing PoSI

- The simultaneity challenge: there are $p\,2^{p-1}$ statistics $|t_{j\bullet \mathrm{M}}|$.

p	1	2	3	4	5	6	7	8	9	10
$\#\lvert t_{j\bullet \mathrm{M}}\rvert$	1	4	12	32	80	192	448	1, 024	2, 304	5, 120
p	11	12	13	14	15	16	17	18	19	20
$\#\lvert t_{j\bullet \mathrm{M}}\rvert$	11, 264	24, 576	53, 248	114, 688	245, 760	524, 288	1, 114, 112	2, 359, 296	4, 980, 736	10, 485, 760

- Monte Carlo-approximation of K_{PoSI} in R, brute force, for $p \overset{\approx}{\leq} 20$.
- Computations are specific to a design $\mathbf{X}$: $K_{\mathrm{PoSI}} = K_{\mathrm{PoSI}}(\mathbf{X}, \alpha, df)$
- Computations depend only on the inner product matrix $\mathbf{X}^T\mathbf{X}$.
  $\Rightarrow$ The limiting factor is p (N may only matter for $\hat{\sigma}^2$).
- One Monte Carlo computation is good for any α and any error df.
- Computations of universal upper bounds:

$$K_{\mathrm{univ}}(p, \alpha, df) \;\geq\; K_{\mathrm{PoSI}}(\mathbf{X}, \alpha, df) \;\; \forall \mathbf{X}_{\ldots \times p}.$$

Computing PoSI (contd.)

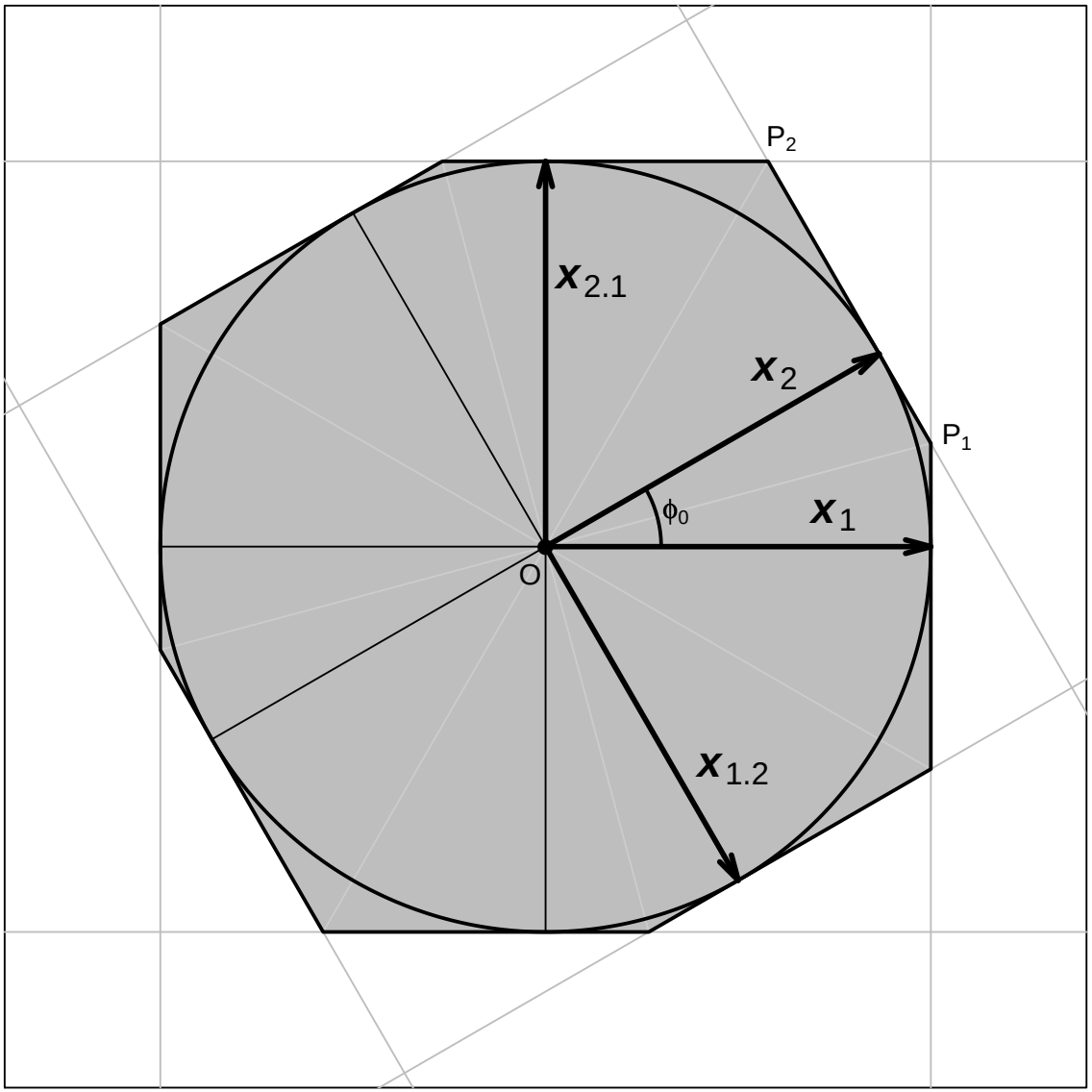
- Install code in R and play with it:

```
install.packages("http://stat.wharton.upenn.edu/~buja/PAPERS/PoSI_1.0.tar.gz",
                 type="source")

library(PoSI); help(PoSI); data(Boston, package="MASS"); summary(PoSI(Boston[,-14]))
```

- What is being computed?
  A: Lots of inner products of adjusted normalized predictors $\mathbf{X}_{j\bullet\mathrm{M}}/\|\mathbf{X}_{j\bullet\mathrm{M}}\|$ with $\epsilon \sim U(S^{d-1})$ ($d = \mathrm{rank}(\mathbf{X})$).
- We sample from $U(S^{d-1})$ and not $\mathcal{N}(\mathbf{0}, \mathbf{I})$ because the radial aspect can be integrated out exactly. Only the directional aspect needs simulating.
- Savings: Pre-projection onto column-space of $\mathbf{X}$.
  Thereafter the size of inner products is of *rank*($\mathbf{X}$), not *N*.
- For each ϵ, obtain $\max_{\mathrm{M}} \max_{j\in\mathrm{M}} \mathbf{X}_{j\bullet\mathrm{M}}^T \epsilon/\|\mathbf{X}_{j\bullet\mathrm{M}}\|$.

The Scheffé Ball and the PoSI Polytope



Circle = Scheffé Ball

The PoSI polytope is tangent to the ball.

Orthogonal Designs have the Smallest PoSI Constants

- In orthogonal designs there is no adjustment:

$$\mathbf{X}^{\text{orth}}_{j\bullet \mathrm{M}} \;=\; \mathbf{X}^{\text{orth}}_{j} \quad \forall\, \mathrm{M},\, j\,(\ni \mathrm{M})$$

- The PoSI statistic simplifies to $\max_{j=1\ldots p} |t_{j\bullet\{j\}}|$,
  hence the PoSI guarantee reduces to

  simultaneity for p orthogonal contrasts.

- The PoSI constant for orthogonal designs is uniformly smallest:

$$K_{\text{orth}}(p, \alpha, df) \;\leq\; K_{\text{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha, df) \quad \forall p, \alpha, df, \mathbf{X}_{\ldots\times p}$$

PoSI Asymptotics

- Natural asymptotics for the PoSI constant $K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha, df)$ are in terms of design sequences $p \mapsto \mathbf{X}_{\ldots\times p}$ as $p \uparrow \infty$ and $df = \infty$, i.e., σ known.
- The Scheffé constant has the following rate in p:
$$K_{\mathrm{Sch}}(p, \alpha) \;=\; \sqrt{\chi^2_{p;1-\alpha}} \;\sim\; \sqrt{p}.$$
  - This represents an upper bound on the PoSI rate.
  - We know a sharper rate bound to be $0.866\ldots\sqrt{p}$.
  - We know of design sequences that reach $0.78\ldots\sqrt{p}$.
- The lowest rate is achieved by orthogonal designs with a rate
$$K_{\mathrm{orth}}(p, \alpha) \;\sim\; \sqrt{2 \log p}.$$
- Hence there is a wide range of rates for the PoSI constants:
$$\sqrt{2 \log p} \;\lessapprox\; K_{\mathrm{PoSI}}(\mathbf{X}_{\ldots\times p}, \alpha) \;\lessapprox\; \sqrt{p}$$
- Under all circumstances, K should not be $t_{df;1-\alpha/2} = O(1)$!

Worst Case PoSI Asymptotics — Some Details

Comments on upper bound $K_{\mathrm{PoSI}} \lesssim 0.866...\sqrt{p}$:

- Ignores the PoSI structure, i.e., the many orthogonalities from adjustment.
- Based purely on the growth rate: $|\{\mathbf{X}_{j\cdot\mathbf{M}} : j \in \mathbf{M}\}| = p\,2^{p-1} \sim 2^p$
- Bound is achieved by random selection of $\sim 2^p$ random vectors:
  $\mathbf{X}_1, ..., \mathbf{X}_{p2^{p-1}} \sim U(S^{p-1})$ i.i.d. versus $\{\mathbf{X}_{j\cdot\mathbf{M}} : j \in \mathbf{M}\}$
- Reduction to radial problem: $\max_{j,\mathbf{M}:\, j\in\mathbf{M}} \langle \mathbf{U}, \mathbf{X}_{j\cdot\mathbf{M}} \rangle$ $(\mathbf{U} \sim U(S^{p-1}))$.
  $\Rightarrow$ Wyner's (1967) bounds on sphere packing apply.

Worst Case PoSI Asymptotics — Some Details

Comments on lower bound $K_{\mathrm{PoSI}} \gtrsim 0.78\sqrt{p}$:

- Best lower bound known to date is found by construction of an example.

$$\begin{vmatrix} 1 & 0 & 0 & 0 & \dots & c \\ 0 & 1 & 0 & 0 & \dots & c \\ 0 & 0 & 1 & 0 & \dots & c \\ 0 & 0 & 0 & 1 & \dots & c \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & \sqrt{1-(p-1)c^2} \end{vmatrix}$$

- This is not be the ultimate worst case yet.

Example: Length of Criminal Sentence (contd.)

- Reminder: t-statistics of selected covariates, in descending order:
  - $|t_{\texttt{alcohol}}| = 3.95$;
  - $|t_{\texttt{prior records}}| = 3.59$;
  - $|t_{\texttt{seriousness}}| = 3.57$;
  - $|t_{\texttt{drugs}}| = 3.31$;
  - $|t_{\texttt{employment}}| = 3.04$;
  - $|t_{\texttt{initial age}}| = 2.56$;
  - $|t_{\texttt{gender}}| = 2.33$.
- The PoSI constant is $K_{\text{PoSI}} \approx 3.1$, hence we would claim significance for the four variables on the left.
- For comparison, the Scheffé constant is $K_{\text{Sch}} \approx 4.5$, leaving us with no significant predictors at all.
- Similarly, Bonferroni with $\alpha/(p\,2^{p-1})$ yields $K_{\text{Bonf}} \approx 4.7$.

Conclusions

- Valid (marginal) universal post-selection inference is possible.
- PoSI is not procedure-specific, hence is conservative. However:
  - PoSI is valid even for selection that is informal and post-hoc.
  - PoSI is necessary for selection based on "significance hunting".
- Asymptotics in p suggests strong dependence of K_{PoSI} on design $\mathbf{X}$.
- Challenges:
  - Understanding the design geometry that drives $K_{\mathrm{PoSI}}(\mathbf{X})$.
  - Computations of K_{PoSI} for large p.
  - Ideally: Find a simple geometric characteristic of $\mathbf{X}$ that drives the size of K_{PoSI}.

THANK YOU!

Ways to Limit the Size of the PoSI Problem

- The full universe of models for full PoSI: all non-singular submodels
  - $\mathcal{M}_{\text{all}} = \{\text{M} : \text{M} \subset \{1, 2, ..., p\},\ 0 < |\text{M}| \leq \min(n, p),\ \text{rank}(\mathbf{X}_{\text{M}}) = |\text{M}|\}.$
- Useful sub-universes:
  - Protect one or more predictors, as in PoSI1: $\mathcal{M} = \{\text{M} : p \in \text{M}\}.$
  - Sparsity, i.e., submodels of size m' or less: $\mathcal{M} = \{\text{M} : |\text{M}| \leq m'\}.$
  - Richness, i.e., drop fewer than m' predictors from the full model: $\mathcal{M} = \{\text{M} : |\text{M}| \geq p - m'\}.$
  - Nested sets of models, as in polynomial regression, AR models, ANOVA.

PoSI Significance: Strong Error Control

For each $j \in \mathrm{M}$, consider the t-test statistic

$$t_{0,j\bullet\mathrm{M}} = \frac{\hat{\beta}_{j\bullet\mathrm{M}} - 0}{\hat{\sigma}_{\bullet}\left(\left(\mathbf{X}_{\mathrm{M}}^{T}\mathbf{X}_{\mathrm{M}}\right)^{-1}_{jj}\right)^{\frac{1}{2}}}.$$

Theorem

Let H_1 be the random set of true alternatives in $\hat{\mathrm{M}}$,
and $\hat{H}_1$ the random set of rejections in $\hat{\mathrm{M}}$:

$$\hat{H}_1 = \{(j, \hat{\mathrm{M}}) : j \in \hat{\mathrm{M}}, |t_{0,j\bullet\hat{\mathrm{M}}}| > K\} \quad \textit{and} \quad H_1 = \{(j, \hat{\mathrm{M}}) : j \in \hat{\mathrm{M}}, \beta_{j\bullet\hat{\mathrm{M}}} \neq 0\}.$$

Then

$$\mathrm{P}\left(\hat{H}_1 \subset H_1\right) \geq 1 - \alpha.$$

If we repeat the sampling process many times, the probability that all PoSI rejections are correct is at least $1 - \alpha$, no matter how the model is selected.

Back