Optimal Inference After Model Selection

Will Fithian
Joint work with Dennis Sun & Jonathan Taylor

December 11, 2015

Outline

Two Stages

Two stages of a statistical investigation:

1. Selection: Choose a probabilistic model for the data, formulate an inference problem.
   Ask a question
2. Inference: Attempt the problem using data & selected model.
   Answer the question

Two Stages

Two stages of a statistical investigation:

1. Selection: Choose a probabilistic model for the data, formulate an inference problem.
   Ask a question
2. Inference: Attempt the problem using data & selected model.
   Answer the question

Classical admonishment: no looking at data until stage 2

Actual practice: choose variables, check for interactions, overdispersion, ...

Two Stages

Two stages of a statistical investigation:

1. Selection: Choose a probabilistic model for the data, formulate an inference problem.
   Ask a question
2. Inference: Attempt the problem using data & selected model.
   Answer the question

Classical admonishment: no looking at data until stage 2

Actual practice: choose variables, check for interactions, overdispersion, ...

How should we relax the classical view?

Naive Inference After Selection

What is wrong with naive inference after selection?

Example (File Drawer Effect): Observe independent
$Y_i \sim N(\mu_i, 1), \;\; i = 1, \ldots, n.$

1. Restrict attention to apparently large effects

$$\hat{I} = \{i : |Y_i| > 1\}.$$

2. Nominal level-α test of $H_{0,i} : \mu_i = 0$, for $i \in \hat{I}$
(e.g., $\alpha = 0.05$: reject if $|Y_i| > 1.96$)

Naive Inference After Selection

What is wrong with naive inference after selection?

Example (File Drawer Effect): Observe independent $Y_i \sim N(\mu_i, 1), \;\; i = 1, \ldots, n$.

1. Restrict attention to apparently large effects

$$\hat{I} = \{i : |Y_i| > 1\}.$$

2. Nominal level-α test of $H_{0,i} : \mu_i = 0$, for $i \in \hat{I}$ (e.g., $\alpha = 0.05$: reject if $|Y_i| > 1.96$)

"Everyone knows" this is invalid. Why?

Naive Inference After Selection

Problem: frequency properties among selected nulls

$$\frac{\text{\# false rejections}}{\text{\# true nulls tested}} \to \frac{\mathbb{P}_{H_{0,i}}(i \in \hat{I},\ \text{reject } H_{0,i})}{\mathbb{P}(i \in \hat{I})}$$

$$= \mathbb{P}_{H_{0,i}}(\text{reject } H_{0,i} \mid i \in \hat{I})$$

Naive Inference After Selection

Problem: frequency properties among selected nulls

$$\frac{\text{\# false rejections}}{\text{\# true nulls tested}} \to \frac{\mathbb{P}_{H_{0,i}}(i \in \hat{I},\ \text{reject } H_{0,i})}{\mathbb{P}(i \in \hat{I})}$$
$$= \mathbb{P}_{H_{0,i}}(\text{reject } H_{0,i} \mid i \in \hat{I})$$

Solution: directly control selective type I error rate

$$\mathbb{P}_{H_{0,i}}(\text{reject } H_{0,i} \mid i \in \hat{I})$$

Example:

$$\mathbb{P}_{H_{0,i}}(|Y_i| > 2.41 \mid \ |Y_i| > 1) = 0.05$$

Naive Inference After Selection

Problem: frequency properties among selected nulls

$$\frac{\text{\# false rejections}}{\text{\# true nulls tested}} \to \frac{\mathbb{P}_{H_{0,i}}(i \in \hat{I},\ \text{reject } H_{0,i})}{\mathbb{P}(i \in \hat{I})}$$
$$= \mathbb{P}_{H_{0,i}}(\text{reject } H_{0,i} \mid i \in \hat{I})$$

Solution: directly control selective type I error rate

$$\mathbb{P}_{H_{0,i}}(\text{reject } H_{0,i} \mid i \in \hat{I})$$

Example:

$$\mathbb{P}_{H_{0,i}}(|Y_i| > 2.41 \mid\ |Y_i| > 1) = 0.05$$

Guiding principle when asking random questions:

The answer must be valid, given that the question was asked

False Coverage-Statement Rate

Benjamini & Yekutieli (2005): CIs for selected parameters, e.g.

- selected genes in GWAS
- selected treatment in clinical trials

Analog of FDR:

$$\mathbb{E}\left[\frac{\#\text{ non-covering CIs}}{1 \ \vee \ \#\text{ CIs constructed}}\right] \leq \alpha$$

Conditional inference used as device for FCR control (Weinstein, F, & Benjamini 2013)

Also used to correct bias (e.g. Sampson & Sill, 2005; Zöllner & Pritchard, 2007; Zhong & Prentice 2008)

Difference in perspective: should we average over questions?

Motivating Example 1: Verifying the Winner

Setup: Quinnipiac poll of 667 Iowa Republicans, May 2014:

Rank	Candidate	Result
1.	Scott Walker	21%
2.	Rand Paul	13%
3.	Marco Rubio	13%
4.	Ted Cruz	12%
⋮	⋮	
14.	Bobby Jindal	1%
15.	Lindsey Graham	0%

Question: Is Scott Walker really winning? By how much?

Problem: Winner's curse

"Question selection," not really "model selection"

Related to subset selection (Gupta & Nagel 1967, others)

Motivating Example 2: Inference After Model Checking

Two-sample problem:

$$X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F_1, \qquad Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_2$$

Motivating Example 2: Inference After Model Checking

Two-sample problem:

$$X_1, \dots, X_m \overset{\text{i.i.d.}}{\sim} F_1, \qquad Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} F_2$$

Test Gaussian model based on normalized residuals

$$R = \left(\frac{X_1 - \overline{X}}{S_X}, \; \dots, \; \frac{X_m - \overline{X}}{S_X}, \; \frac{Y_1 - \overline{Y}}{S_Y}, \; \dots, \; \frac{Y_n - \overline{Y}}{S_Y} \right)$$

If test rejects, use permutation test (e.g., Wilcoxon):

$$F_1 = ?, \qquad F_2 = ?, \qquad H_0 : \; F_1 = F_2$$

Otherwise, use two-sample t-test:

$$F_1 = N(\mu, \sigma^2), \qquad F_2 = N(\nu, \tau^2), \qquad H_0 : \; \mu = \nu$$

Model selection, strong sense

Motivating Example 3: Regression After Variable Selection

E.g., solve lasso at fixed $\lambda > 0$ (Tibshirani, 1996):

$$\hat{\gamma} = \arg\min_{\gamma} \|Y - X\gamma\|_2^2 + \lambda \|\gamma\|_1$$

"Active set" $E = \{j : \hat{\gamma}_j \neq 0\}$ induces selected model $M(E)$:

$$Y \sim N\left(X_E \beta^E, \sigma^2 I_n\right)$$

Motivating Example 3: Regression After Variable Selection

E.g., solve lasso at fixed $\lambda > 0$ (Tibshirani, 1996):

$$\hat{\gamma} = \arg\min_{\gamma} \|Y - X\gamma\|_2^2 + \lambda\|\gamma\|_1$$

"Active set" $E = \{j : \hat{\gamma}_j \neq 0\}$ induces selected model $M(E)$:

$$Y \sim N\left(X_E \beta^E, \sigma^2 I_n\right)$$

Can we get valid tests / intervals for $\beta_j^E, \quad j \in E$?

Lee, Sun, Sun, & Taylor (2013) studied slightly different problem (inference w.r.t. different model)

Random Model, Random Null

Testing null hypothesis H_0 in model M

Selective error rate: $\mathbb{P}_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected})$

Nominal error rate: $\mathbb{P}_{M,H_0}(\text{reject } H_0)$

Random Model, Random Null

Testing null hypothesis H_0 in model M

Selective error rate: $\mathbb{P}_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected})$

Nominal error rate: $\mathbb{P}_{M,H_0}(\text{reject } H_0)$

"Kosher" adaptive selection: two independent experiments

- Select M, H_0 based on exploratory experiment 1
- Test using confirmatory experiment 2

Random Model, Random Null

Testing null hypothesis H_0 in model M

Selective error rate: $\mathbb{P}_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected})$

Nominal error rate: $\mathbb{P}_{M,H_0}(\text{reject } H_0)$

"Kosher" adaptive selection: two independent experiments

- Select M, H_0 based on exploratory experiment 1
- Test using confirmatory experiment 2

M, H_0 random, but no adjustment necessary:

$$\mathbb{P}_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected}) = \mathbb{P}_{M,H_0}(\text{reject } H_0).$$

Data Splitting

Assume $Y = (Y_1, Y_2)$ with $Y_1 \perp\!\!\!\perp Y_2$

Data splitting mimics exploratory / confirmatory split:

- Select model based on Y_1
- Analyze Y_2 as though model chosen "ahead of time."

Again, no adjustment necessary:

$$\mathbb{P}_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected}) = \mathbb{P}_{M,H_0}(\text{reject } H_0).$$

Data Splitting

Assume $Y = (Y_1, Y_2)$ with $Y_1 \perp\!\!\!\perp Y_2$

Data splitting mimics exploratory / confirmatory split:

- Select model based on Y_1
- Analyze Y_2 as though model chosen "ahead of time."

Again, no adjustment necessary:

$$\mathbb{P}_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected}) = \mathbb{P}_{M,H_0}(\text{reject } H_0).$$

Objections to data splitting:

- less data for selection
- less data for inference
- not always possible (e.g., autocorrelated data)

Data Carving

Think of data as "revealed in stages:"

Let $A = \{(M, H_0) \text{ selected}\}$.

$$\underbrace{\mathscr{F}_0 \quad \subseteq \quad \mathscr{F}(\mathbf{1}_A(Y))}_{\text{used for selection}} \quad \underbrace{\subseteq \quad \mathscr{F}(Y)}_{\text{used for inference}}$$

Data Carving

Think of data as "revealed in stages:"

Let $A = \{(M, H_0) \text{ selected}\}$.

$$\mathscr{F}_0 \quad \underbrace{\subseteq}_{\text{used for selection}} \quad \mathscr{F}(\mathbf{1}_A(Y)) \quad \underbrace{\subseteq}_{\text{used for inference}} \quad \mathscr{F}(Y)$$

Conditioning on A in stage two
$\iff Y \in A$ excluded as evidence against H_0

Data Carving

Think of data as "revealed in stages:"

Let $A = \{(M, H_0) \text{ selected}\}$.

$$\mathscr{F}_0 \quad \underbrace{\subseteq}_{\text{used for selection}} \quad \mathscr{F}(\mathbf{1}_A(Y)) \quad \underbrace{\subseteq}_{\text{used for inference}} \quad \mathscr{F}(Y)$$

Conditioning on A in stage two

$\iff$ $Y \in A$ excluded as evidence against H_0

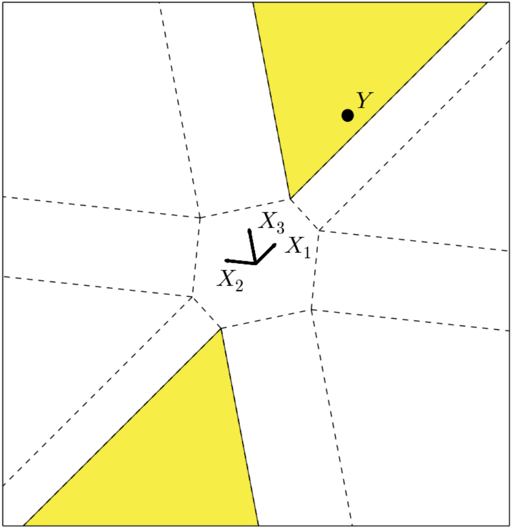
Data splitting conditions on Y_1 instead of $\mathbf{1}_A(Y_1)$

$$\mathscr{F}_0 \quad \underbrace{\subseteq}_{\text{used for selection}} \quad \mathscr{F}(\mathbf{1}_A(Y_1)) \quad \underbrace{\subseteq}_{\text{wasted}} \quad \mathscr{F}(Y_1) \quad \underbrace{\subseteq}_{\text{used for inference}} \quad \mathscr{F}(Y_1, Y_2).$$

Data Carving: Use all leftover information for inference

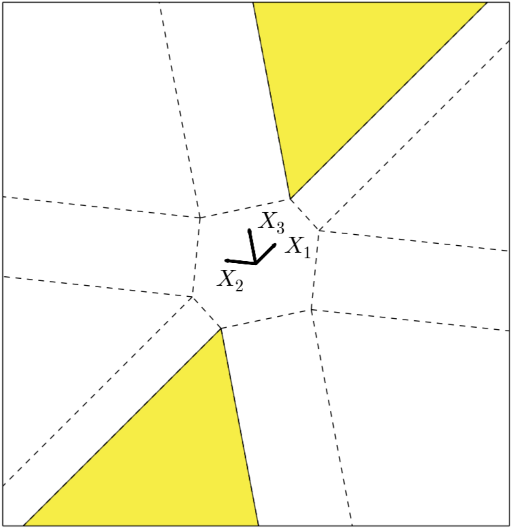
Lasso Partition

Yellow region: $\{y : \text{Variables 1, 3 selected}\}$



Lasso Partition

```
M.hat = which(coef(glmnet(X, Y), lambda) != 0)
```



Goals

Prior work on linear regression after selection with σ^2 known

Lockhart et al. (2014), Tibshirani et al. (2014), Lee et al. (2013), Loftus and Taylor (2014), Lee and Taylor (2014), ...

Our goals:

1. Formalize inference after selection
2. Understand power — can it be improved?
3. Generalize to unknown σ^2
4. Generalize to other exponential families

Outline

Selective Hypothesis Tests

Setup: Observe $Y \sim F$ on space $(\mathcal{Y}, \mathscr{F})$, F unknown

Question space: collection $\mathcal{Q}$ of all candidate testing problems q

Testing problem is a pair $q = (M, H_0)$ of

- model $M(q)$ (family of distributions)
- null hypothesis $H_0(q) \subseteq M(q)$. (wlog $H_1 = M \setminus H_0$)

Selective Hypothesis Tests

Setup: Observe $Y \sim F$ on space $(\mathcal{Y}, \mathscr{F})$, F unknown

Question space: collection $\mathcal{Q}$ of all candidate testing problems q

Testing problem is a pair $q = (M, H_0)$ of

- model $M(q)$ (family of distributions)
- null hypothesis $H_0(q) \subseteq M(q)$. (wlog $H_1 = M \setminus H_0$)

Two stages:

1. Selection: Select subset $\widehat{\mathcal{Q}}(Y) \subseteq \mathcal{Q}$ to test
2. Inference: Test H_0 vs. $M \setminus H_0$ for each $q = (M, H_0) \in \widehat{\mathcal{Q}}$

Selective Hypothesis Tests

Design hypothesis test $\phi_q(y) : \mathcal{Y} \to [0, 1]$ for question q

We only care about behavior on selection event:

$$A_q = \{q \in \widehat{\mathcal{Q}}(Y)\}$$

A_q: event that q was asked

Selective Hypothesis Tests

Design hypothesis test $\phi_q(y) : \mathcal{Y} \to [0, 1]$ for question q

We only care about behavior on selection event:

$$A_q = \{q \in \widehat{\mathcal{Q}}(Y)\}$$

A_q: event that q was asked

Test $\phi_q(y)$ is a selective level-α test if

$$\mathbb{E}_F\left[\phi_q(Y) \mid A_q\right] \le \alpha, \quad \forall F \in H_0$$

Selective power function:

$$\mathsf{Pow}_{\phi_q}(F \mid A_q) = \mathbb{E}_F\left[\phi_q(Y) \mid A_q\right]$$

Selective Hypothesis Tests

Design hypothesis test $\phi_q(y) : \mathcal{Y} \to [0, 1]$ for question q

We only care about behavior on selection event:

$$A_q = \{q \in \widehat{\mathcal{Q}}(Y)\}$$

A_q: event that q was asked

Test $\phi_q(y)$ is a selective level-α test if

$$\mathbb{E}_F\left[\phi_q(Y) \mid A_q\right] \leq \alpha, \quad \forall F \in H_0$$

Selective power function:

$$\text{Pow}_{\phi_q}(F \mid A_q) = \mathbb{E}_F\left[\phi_q(Y) \mid A_q\right]$$

NB: Selective level defined w.r.t. $F \in M(q)$
$\implies$ can design tests "one (M, H_0) at a time"

What If the Model Is Wrong?

Some (all?) M are probably misspecified ($F \notin M$).

We don't know which.

Non-adaptive inference:

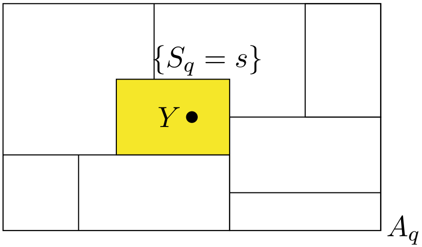
- Size of ϕ defined w.r.t. selected model M
- Guarantees vacuous when $F \notin M$
- Try to select correct or "close enough" M

Adaptive inference:

- Same situation: selective size of ϕ_q defined w.r.t. $M(q)$
- Benefit: allowed to check model

Conditioning on Selection Variables'

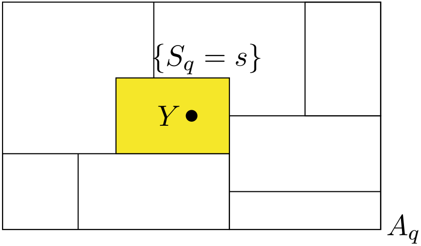
Sometimes want to condition on more than A_q:



More generally, can condition on finer selection variable $S_q(Y)$, with $A_q \in \mathscr{F}(S_q)$

Conditioning on Selection Variables'

Sometimes want to condition on more than A_q:



More generally, can condition on finer selection variable $S_q(Y)$, with $A_q \in \mathscr{F}(S_q)$, e.g.

- $S_q(Y) = Y_1$ (data splitting)
- $S_q(Y) =$ active variables and signs (inference after lasso)
  Reason: tractable computation
- can control FCR with $S_q(Y) = (\mathbf{1}_{A_q}(Y), |\widehat{\mathcal{Q}}(Y)|)$
  Reason: stronger inferential guarantee

Conditioning Discards Information

ϕ_q has selective level α w.r.t S_q if

$$\mathbb{E}_F\left[\phi_q(Y) \mid S_q(Y)\right] \overset{\text{a.s.}}{\leq} \alpha, \quad \text{on } A_q, \quad \forall F \in H_0$$

More stringent when S_q is finer

Finest: $S_q(Y) = Y$, Coarsest: $S_q(Y) = \mathbf{1}_{A_q}(Y)$

Cost: conditioning on $S_q \iff$ ignoring evidence in S_q

Leftover Information

After conditioning on $S(Y) = s$, the leftover information is

$$\mathcal{I}_{Y \mid S}(\theta; s) = \mathrm{Var}\left[\nabla \ell(\theta;\ Y \mid S = s) \mid S = s\right]$$

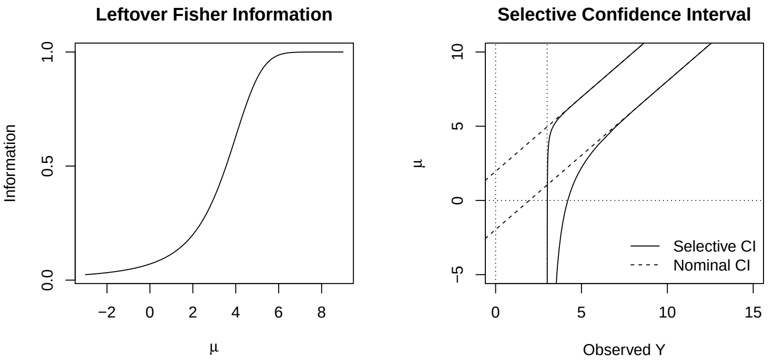
Can characterize:

$$\mathbb{E}\left[\mathcal{I}_{Y \mid S}(\theta; S)\right] \ = \ \mathcal{I}_Y(\theta) - \mathcal{I}_S(\theta) \ \preceq \ \mathcal{I}_Y(\theta).$$

$\mathcal{I}_S(\theta)$: the (average) price of selection

Leftover Information

$$Y \sim N(\mu, 1), \qquad A = \{Y > 3\}$$



Selective Tests for Exponential Families

Goal: Test $H_0 : \theta = \theta_0$, nuisance parameter ζ where

$$Y \sim \exp\left\{\theta\, T(y) + \zeta' U(y) - \psi(\theta, \zeta)\right\}\; f_0(y)$$

Selective Tests for Exponential Families

Goal: Test $H_0 : \theta = \theta_0$, nuisance parameter ζ where

$$Y \sim \exp\left\{\theta\, T(y) + \zeta' U(y) - \psi(\theta, \zeta)\right\}\; f_0(y)$$

Selection event A:

$$Y \mid A \sim \exp\left\{\theta\, T(y) + \zeta' U(y) - \psi_A(\theta, \zeta)\right\}\; f_0(y)\; \mathbf{1}_A(y)$$

Selective Tests for Exponential Families

Goal: Test $H_0 : \theta = \theta_0$, nuisance parameter ζ where

$$Y \sim \exp\left\{\theta\, T(y) + \zeta' U(y) - \psi(\theta, \zeta)\right\}\; f_0(y)$$

Selection event A:

$$Y \mid A \sim \exp\left\{\theta\, T(y) + \zeta' U(y) - \psi_A(\theta, \zeta)\right\}\; f_0(y)\, \mathbf{1}_A(y)$$

Conditioning on U eliminates ζ, base test on one-parameter family

$$\mathcal{L}_\theta(T \mid U,\; Y \in A)$$

Side constraint: selective unbiasedness

$$\mathbb{E}_\theta\; [\phi(Y) \mid A] \geq \alpha, \quad \forall \theta \neq \theta_0$$

Selective Tests for Exponential Families

$$Y \mid Y \in A \sim \exp\left\{\theta\, T(y) + \zeta' U(y) - \psi_A(\theta, \zeta)\right\}\; f_0(y)\, \mathbf{1}_A(y)$$

Proposal (F, Sun & Taylor 2014)

The UMPU selective level-α test ϕ of $H_0 : \theta = \theta_0$ rejects for $\{T < C_1(U)\} \cup \{T > C_2(U)\}$, with C_i chosen so that

$$\mathbb{E}_{\theta_0}\left[\phi(T, U) \mid U, A\right] = \alpha \qquad \text{(Selective Level } \alpha\text{)}$$

$$\mathbb{E}_{\theta_0}\left[T\,\phi(T, U) \mid U, A\right] = \alpha\, \mathbb{E}_{\theta_0}\left[T \mid U, A\right] \qquad \text{(Selectively Unbiased)}$$

Follows from Lehmann & Scheffé (1955)

Solve for cutoffs using Monte Carlo (sampling can be hard)

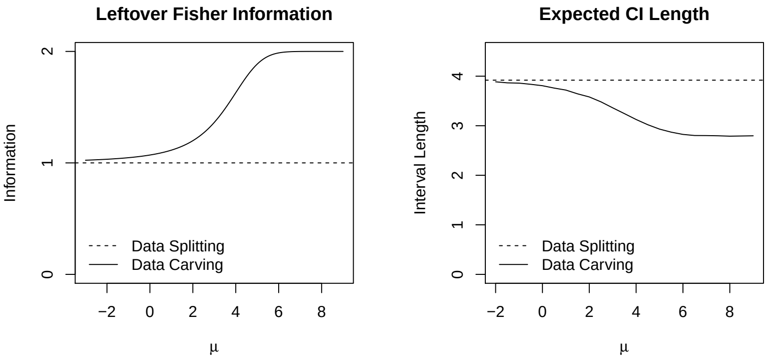
Also show: data splitting typically inadmissible

Data Splitting is Inadmissible

Compare optimal test to data splitting for

$$Y_1, Y_2 \overset{\text{i.i.d.}}{\sim} N(\mu, 1), \qquad A = \{Y_1 > 3\}$$

Optimal test based on $\mathcal{L}(Y_1 + Y_2 \mid Y_1 > 3)$, data splitting based on $\mathcal{L}(Y_2)$.



Outline

Linear Regression

Gaussian response $Y \in \mathbb{R}^n$, regressors $X \in \mathbb{R}^{n \times p}$

Select active set $E \subseteq \{1, \ldots, p\}$ based on lasso, LARS, forward stepwise, ...

Inference w.r.t. selected linear model

$$Y \sim N(X_E \beta^E,\ \sigma^2 I_n)$$

Exponential family in $\beta^E, \sigma^2 \implies$
$\exists$ UMPU selective test for $H_0 :\ \beta_j^E = 0$

Linear Regression: Selected Model

$$Y \sim \exp\left\{-\frac{1}{2\sigma^2}(y - X_E\beta)'(y - X_E\beta)\right\} \frac{1}{\sqrt{2\pi\sigma^2}}$$

Linear Regression: Selected Model

$$Y \sim \exp\left\{ \; \frac{1}{\sigma^2} \sum_{k \in E} \beta_k \; {X_k}' y - \frac{1}{2\sigma^2} \|y\|^2 - \psi(\beta, \sigma^2) \; \right\} f_0(y)$$

Linear Regression: Selected Model

$$Y \sim \exp\left\{ \frac{1}{\sigma^2} \sum_{k \in E} \beta_k \, {X_k}' y - \frac{1}{2\sigma^2} \|y\|^2 - \psi(\beta, \sigma^2) \right\} f_0(y)$$

σ^2 known:

$$T(y) = {X_j}' y, \quad U(y) = {X_{E \setminus j}}' y$$

Selective z-test for β_j on event A is based on

$$\mathcal{L}_{\beta_j} \left(X_j' Y \;\middle|\; {X_{E \setminus j}}' Y, \; A \right)$$

Condition on $(n - |E|)$-dim. hyperplane $\bigcap A$

Hit-and-run MCMC (typically $A =$ polytope)
Exact level-α tests possible w/o mixing (Besag & Clifford, 1989)

Linear Regression: Selected Model

$$Y \sim \exp\left\{ \frac{1}{\sigma^2} \sum_{k \in M} \beta_k \, X_k{}'y - \frac{1}{2\sigma^2} \|y\|^2 - \psi(\beta, \sigma^2) \right\} f_0(y)$$

σ^2 unknown:

$$T(y) = X_j{}'y, \quad U(y) = (X_{E\setminus j}{}'y, \; \|y\|^2)$$

Selective t-test for β_j on event A is based on

$$\mathcal{L}_{\beta_j/\sigma^2}\left(X_j{}'Y \;\middle|\; X_{E\setminus j}{}'Y, \; \|Y\|^2, \; A\right)$$

Condition on $(n - |E|)$-dim. hyperplane $\bigcap$ sphere $\bigcap$ A

Sample using ball $\{\|y\| \leq \|Y\|\}$ instead of sphere, then adjust

Saturated Model

What if we don't believe linear model?

Saturated Model

What if we don't believe linear model?

Idea: $Y \sim N(\mu, \sigma^2 I_n)$ (saturated model),
define least-squares parameters for "model" $E \subseteq \{1, \dots, p\}$:

$$\theta^E \triangleq \arg\min_\theta \mathbb{E}_\mu \left[\|Y - X_E \theta\|^2 \right] \\ = (X_E' X_E)^{-1} X_E' \mu$$

Used by Berk et al. (2012), Taylor et al. (2014), Lee et al. (2013), Loftus and Taylor (2014), Lee and Taylor (2014), others

Saturated Model

What if we don't believe linear model?

Idea: $Y \sim N(\mu, \sigma^2 I_n)$ (saturated model),
define least-squares parameters for "model" $E \subseteq \{1, \ldots, p\}$:

$$\theta^E \triangleq \arg\min_\theta \mathbb{E}_\mu \left[\|Y - X_E\theta\|^2\right]$$
$$= (X_E' X_E)^{-1} X_E' \mu$$

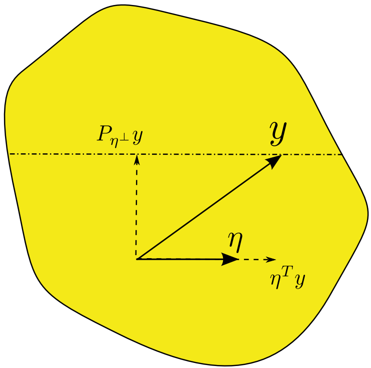
Used by Berk et al. (2012), Taylor et al. (2014), Lee et al. (2013), Loftus and Taylor (2014), Lee and Taylor (2014), others

Parameters are linear contrasts: $\theta_j^E = \eta' \mu$

σ^2 known: test of $H_0 : \theta_j^E = 0$ based on $\mathcal{L}_{\theta_j^E}\left(\eta' Y \;\middle|\; \mathcal{P}_\eta^\perp Y,\ A\right)$

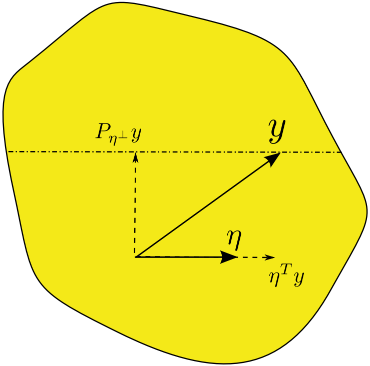
Linear Regression: Saturated Model

$\mathcal{L}_{\theta_j^E}\left(\eta' Y \;\middle|\; \mathcal{P}_\eta^\perp Y,\ A\right)$: Gaussian truncated to a "slice"



Linear Regression: Saturated Model

$\mathcal{L}_{\theta_j^E}\left(\eta' Y \;\middle|\; \mathcal{P}_\eta^\perp Y,\ A\right)$: Gaussian truncated to a "slice"



σ^2 unknown: also need to condition on $\|Y\|$

line $\bigcap$ sphere: leaves only 2 points in support

Saturated vs. Selected z-Test

Usual z-statistic $Z = \frac{\eta' y}{\sigma \|\eta\|}$

Selected-model z-test based on

$$\mathcal{L}_{\beta_j^E}\left(Z \;\middle|\; X_{M\setminus j}{}'Y,\ A\right)$$

Saturated-model z-test based on

$$\mathcal{L}_{\theta_j^E}\left(Z \;\middle|\; \mathcal{P}_\eta^\perp Y,\ A\right)$$

Selected-model test more powerful (conditions on less)

Saturated-model test more robust (valid under weaker assumptions)

Hybrid approaches exist

Simulation

Setup: regression with $n = 100, p = 200, Y \sim N(X\beta, I_n)$

True $\beta_j = \begin{cases} 7 & j = 1, \ldots, 7 \\ 0 & j > 7 \end{cases}$

X Gaussian, pairwise correlation 0.3 between variables (normalized)

Simulation

Setup: regression with $n = 100, p = 200, Y \sim N(X\beta, I_n)$

True $\beta_j = \begin{cases} 7 & j = 1, \ldots, 7 \\ 0 & j > 7 \end{cases}$

X Gaussian, pairwise correlation 0.3 between variables (normalized)

Split data into $Y^{(1)} = (Y_1, \ldots, Y_{n_1})$, $Y^{(2)} = (Y_{n_1+1}, \ldots, Y_{100})$

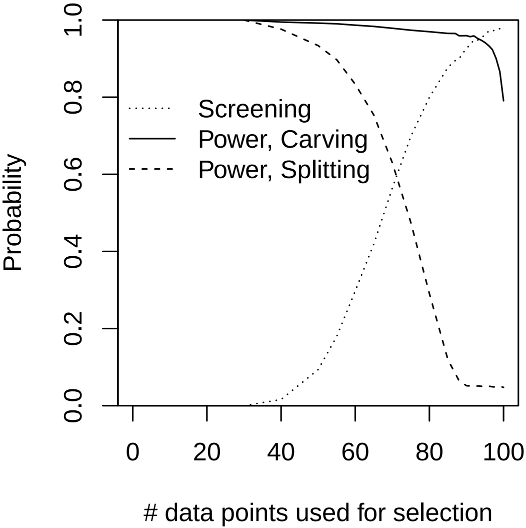
Selection: lasso on $Y^{(1)}$ using $\lambda = 2\mathbb{E}(\|X'\epsilon\|_\infty)$, $\epsilon \sim N(0, I)$
Suggested by Negahban et al. (2012)

Inference: two procedures

- Data Splitting (Split$_{n_1}$): Use $Y^{(2)}$ for inference
- Data Carving (Carve$_{n_1}$): Selected model z-test

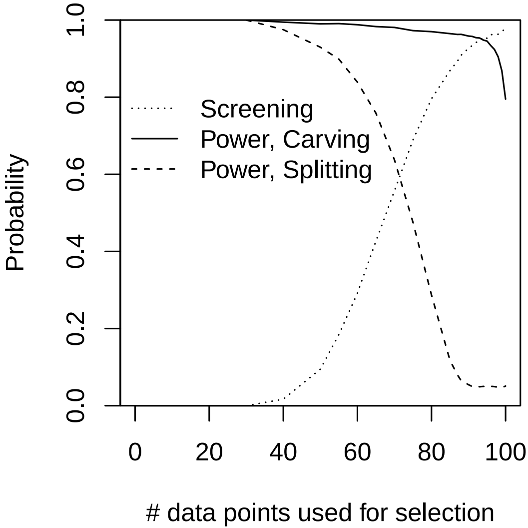
Selection–Inference Tradeoff

As n_1 varies, tradeoff between model selection quality and power



Selection–Inference Tradeoff

Robustness: same plot for t_5 errors



Outline

Motivation: Iowa Caucus

Setup: Quinnipiac poll of $n = 667$ Iowa Republicans:

Rank	Candidate	Result	Votes*
1.	Scott Walker	21%	140
2.	Rand Paul	13%	87
3.	Marco Rubio	13%	87
4.	Ted Cruz	12%	80
⋮	⋮		
14.	Bobby Jindal	1%	7
15.	Lindsey Graham	0%	0

Question: Is Scott Walker really winning?

Answer: Yes (p=0.00053), by at least 22%

p=0.022 for Gupta & Nagel method

Winner vs. Runner-Up Test

Theorem (F 2015):

Let $[d]$ denote the index of the largest count, and conclude that $\pi_{[d]} > \max_{j<d} \pi_{[j]}$ if exact, two-sided binomial level-α test of $H_0 : \pi_{[d]} \leq \pi_{[d-1]}$ rejects.
This is a valid level-α procedure.

Analogous result known for Gaussians (Gutmann & Maymin, 1987)

Winner vs. Runner-Up Test

Theorem (F 2015):

Let $[d]$ denote the index of the largest count, and conclude that $\pi_{[d]} > \max_{j<d} \pi_{[j]}$ if exact, two-sided binomial level-α test of $H_0 : \pi_{[d]} \leq \pi_{[d-1]}$ rejects.
This is a valid level-α procedure.

Analogous result known for Gaussians (Gutmann & Maymin, 1987)

Conditional approach leads to:

- Lower confidence bound for $\pi_{SW} - \max_{j \neq SW} \pi_j$
- Subset selection rule
- Stepdown procedure yielding confident ranks

Stepdown Procedure

Stepdown Procedure: Start with #1, reject until $p > .05$

Quinnipiac poll of $n = 692$ Iowa Democrats:

Rank	Candidate	Result	Votes
1.*	Hillary Clinton	60%	415
2.*	Bernie Sanders	15%	104
3.*	Joe Biden	11%	76
4.*	Don't Know	7%	48
5.	Jim Webb	3%	21
6.	Mark O'Malley	3%	21
7.	Lincoln Chafee	0%	0

FWER controlled at $\alpha = 0.05$

Sequential Model Selection

New work (F, Taylor, Tibshirani, Tibshirani):

Generate nested model sequence in algorithmic fashion

$$M_0(Y) \subseteq M_1(Y) \subseteq \cdots \subseteq M_d(Y) \subseteq M_\infty$$

e.g.

- Forward stepwise, lasso
- Graphical lasso
- "Best first" decision tree

Goal: select least complex model consistent with data
control FDR, FWER (type I error = # of extra steps)

Need to condition on subpath $M_0, \ldots, M_k$
null p-values are iid uniform (use ForwardStop, Accum. Tests)

Forward stepwise, lasso: $2p$ linear constraints afer k steps.

Diabetes Example

Step	Variable	Nominal p-value	Saturated p-value	Max-t p-value
1	bmi	0.00	0.00	0.00
2	ltg	0.00	0.00	0.00
3	map	0.00	**0.05**	0.00
4	age:sex	0.00	0.33	0.02
5	bmi:map	0.00	0.76	0.08
6	hdl	0.00	0.25	0.06
7	sex	0.00	0.00	0.00
8	glu^2	0.02	0.03	**0.32**
9	age^2	0.11	0.55	0.94
10	map:glu	0.17	0.91	0.91
11	tc	0.15	0.37	0.25
12	ldl	0.06	0.15	0.01
13	ltg^2	0.00	0.07	0.04
14	age:ldl	0.19	0.97	0.85
15	age:tc	0.08	0.15	0.03
16	sex:map	0.18	0.05	0.40
17	glu	0.23	0.45	0.58
18	tch	**0.31**	0.71	0.82
19	sex:tch	0.22	0.40	0.51
20	sex:bmi	0.27	0.60	0.44

Conclusions

Conditioning on selection generalizes data splitting

Doable in interesting problems

Conditioning $\iff$ discarding information

Knowledge of selection protocol allows us not to “overcorrect”

The End

Thanks!