From post-hoc analysis to post-selection inference

Daniel Yekutieli

Statistics and OR
Tel Aviv University

WHOA-PSI
St. Louis, October 2, 2016

Plan of my talk

A subjective recount

1. Connection between FDR control and post-hoc analysis
2. Post selection inference
3. Bayesian post selection inference
4. Post model selection inference

Multiple hypotheses testing framework

- m tested null hypotheses $H_1 \cdots H_m$
- m_0 null hypotheses $(P_i \sim U[0,1])$,
  $m_1 = m - m_0$ false null hypotheses $(P_i \leq U[0,1])$
- Rejecting a null hypothesis is a discovery, a false discovery is erroneously rejecting a true null hypothesis
- R is the number of discoveries and V is the number of false discoveries

$$FWE := \Pr(V > 0)$$

$$FDR := EQ, \quad Q = \begin{cases} 0 & \text{if } R = 0 \\ V/R & \text{if } R > 0 \end{cases}$$

Level α $(= 0.05)$ Bonferroni procedure

1. Reject H_i if $P_i \leq \alpha/m$

$\Rightarrow$ FWE control via Bonferroni inequality:

$$\Pr(\cup_{i \in I_0} P_i \leq \alpha/m) \leq \sum_{i \in I_0} \Pr(P_i \leq \alpha/m) = m_0 \cdot \alpha/m \leq \alpha$$

Where $I_0 \subseteq \{1 \cdots m\}$ is subset of true null hypotheses

Level q $(= 0.05)$ BH procedure

1. Sort the p-values $P_{(1)} \leq P_{2)} \leq \cdots \leq P_{(m)}$
2. Compare $P_{(i)}$ with $i \cdot q/m$
3. Let $r = \max\{i : P_{(i)} \leq i \cdot q/m\}$
4. Reject $H_{(1)} \cdots H_{(r)}$

BH '95

- Independence $FDR \leq m_0 \cdot q/m$

Benjamini and Yekutieli '01

- Independence $FDR = m_0 \cdot q/m$
- Positive dependence $FDR \leq m_0 \cdot q/m$
- Geneal dependence $FDR \leq (1 + 1/2 + \cdots + 1/m) \cdot m_0 \cdot q/m$

Interpretation of level 0.05 FDR control

All noise regime ($m_0 = m$)

- Any discovery is false – $FDR \equiv FWE$
- 0.95 probability of not making any false discovery

Signal and noise regime ($m_0 < m$)

- Many discoveries $\approx$ 0.05 false – $FWE \to 1$
- A randomly selected discovery is true with prob. 0.95

Selective vs. simultaneous inference

Benjamini and Yekutieli '05: two types of problems can arise when providing inferences in studies with multiple parameters . . .

- *Selective inference* – need to provide marginal inferences for parameters that are selected after viewing the data (e.g. microarray analysis) – solution FDR control
- *Simultaneous inference* – need to provide inferences that apply to all the parameters (e.g. subgroup analysis) – solution FWE control

Selective inference a new idea?

- Soric, JASA ‘89
  “ . . . It is mainly the discoveries that are reported and included into science . . . unless the proportion of false discoveries is kept small there is danger that a large part of science is untrue”

- Ioannidis ‘05 (Plos Medicine)
  “Why Most Published Research Findings Are False”

- Tukey ‘53; Scheffe ‘53
  post-hoc analysis

Post-hoc analysis – Scheffe's method

- $\boldsymbol{\mu} = (\mu_1 \cdots \mu_k)$ is a vector of k treatment effects, mean response in i'th treatment group is $\hat{\mu}_i \sim N(\mu_i, \sigma^2/n)$
- After viewing the data (and ANOVA) a contrast, $\boldsymbol{a\mu} = a_1\mu_1 + \cdots a_k\mu_k$ with $a_1 + \cdots + a_k = 0$, is selected

Selective inference problem: how do we use data to select $\boldsymbol{a\mu}$ and then test its significance or construct a confidence interval for it?

Solution: base inference on confidence interval

$$CI_{Scheffe}(\boldsymbol{a}, \alpha) := \boldsymbol{a}\hat{\boldsymbol{\mu}} \pm \frac{\hat{\sigma} \cdot \|\boldsymbol{a}\|}{\sqrt{n}} \cdot \sqrt{(k-1) \cdot F_{1-\alpha,k-1,N-k}}$$

that offer simultaneous coverage for all contrasts

$$\Pr_{\boldsymbol{\mu}}\{\forall \boldsymbol{a} : \ \boldsymbol{a\mu} \in CI_{Scheffe}(\boldsymbol{a}, \alpha)\} \geq 1 - \alpha.$$

Scheffe's method – FWE control

- Family of null hypotheses: $\forall \boldsymbol{a}, H_{\boldsymbol{a}}^0 : \boldsymbol{a}\boldsymbol{\mu} = 0$
- True null hypotheses: $\{\boldsymbol{a} : \boldsymbol{a}\boldsymbol{\mu} = 0\}$
- Test: reject $H_{\boldsymbol{a}}^0 : \boldsymbol{a}\boldsymbol{\mu} = 0$ if $0 \notin CI_{Scheffe}(\boldsymbol{a}, \alpha)$
- Coverage for all contrasts $\rightarrow$ FWE control for family of null hypotheses

FDR control – slightly different objective

- Post hoc analysis is concerned with valid inference for a single contrast (possibly the most significant contrast) that is specified according to the data (i.e. simultaneity a definite but costly solution for valid selective inference)

- Conditional marginal validity property of FDR more appropriate in contemporary applications (microarrays / GWAS / fMRI / nonparametric regression) that are concerned with deriving valid marginal inferences for multiple parameters that are selected after first considering m pre-specified parameters

- And indeed ... Williams, Jones and Tukey '99 suggest using the BH procedure for discovering state-to-state (pairwise) differences in educational achievement.

Selective inference framework

Benjamini and Yekutieli '05:

- m parameters $\theta_1 \cdots \theta_m$ with corresponding estimators $T_1 \cdots T_m$
- There is a selection rule $\mathcal{S}(T_1 \cdots T_m) \subseteq \{1 \cdots m\}$
- Goal: to construct valid marginal confidence interval for the selected parameters: θ_i, $i \in \mathcal{S}(T_1 \cdots T_m)$

Continuous parameter-value simulation

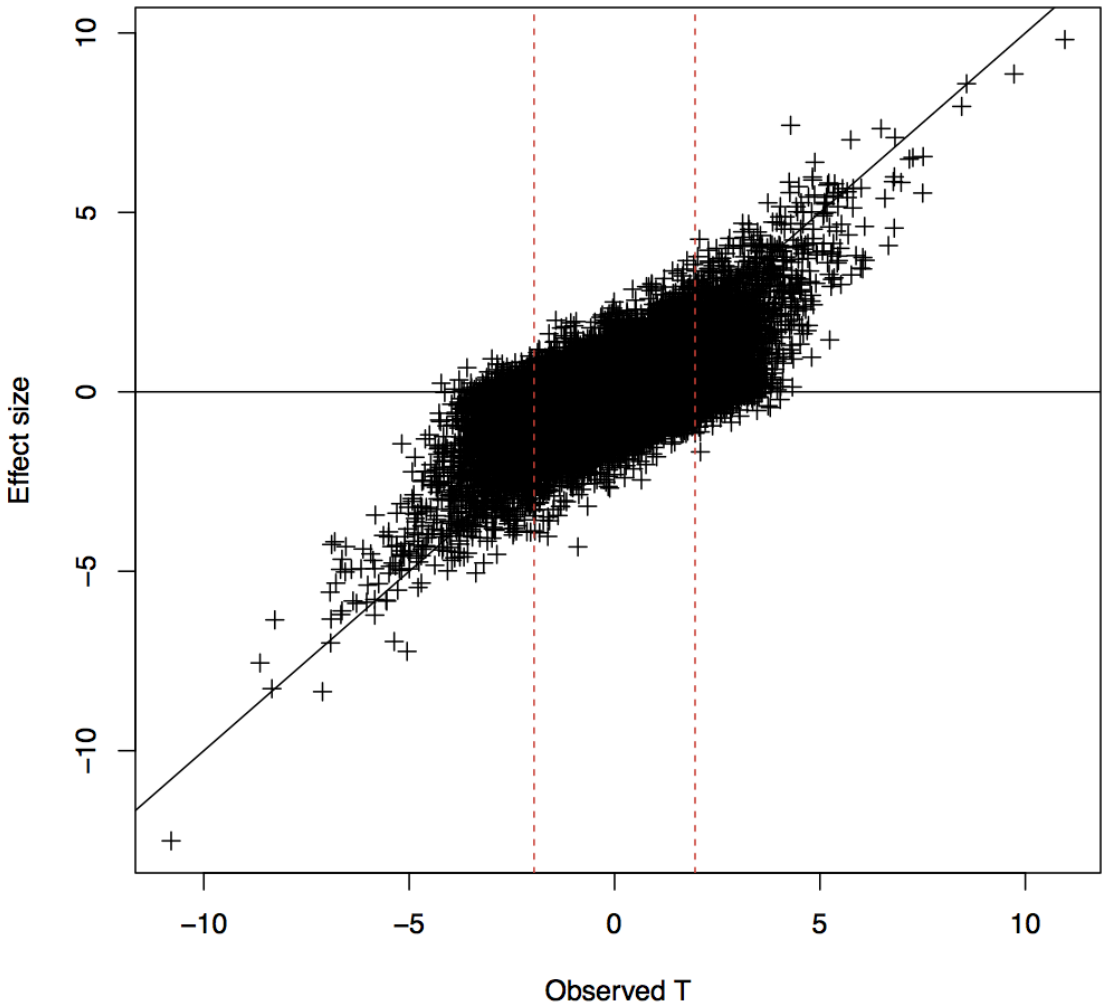
Generate $m = 10,000$ iid (θ_i, Y_i):

- Parameter $\theta_i \sim \pi(\theta_i)$

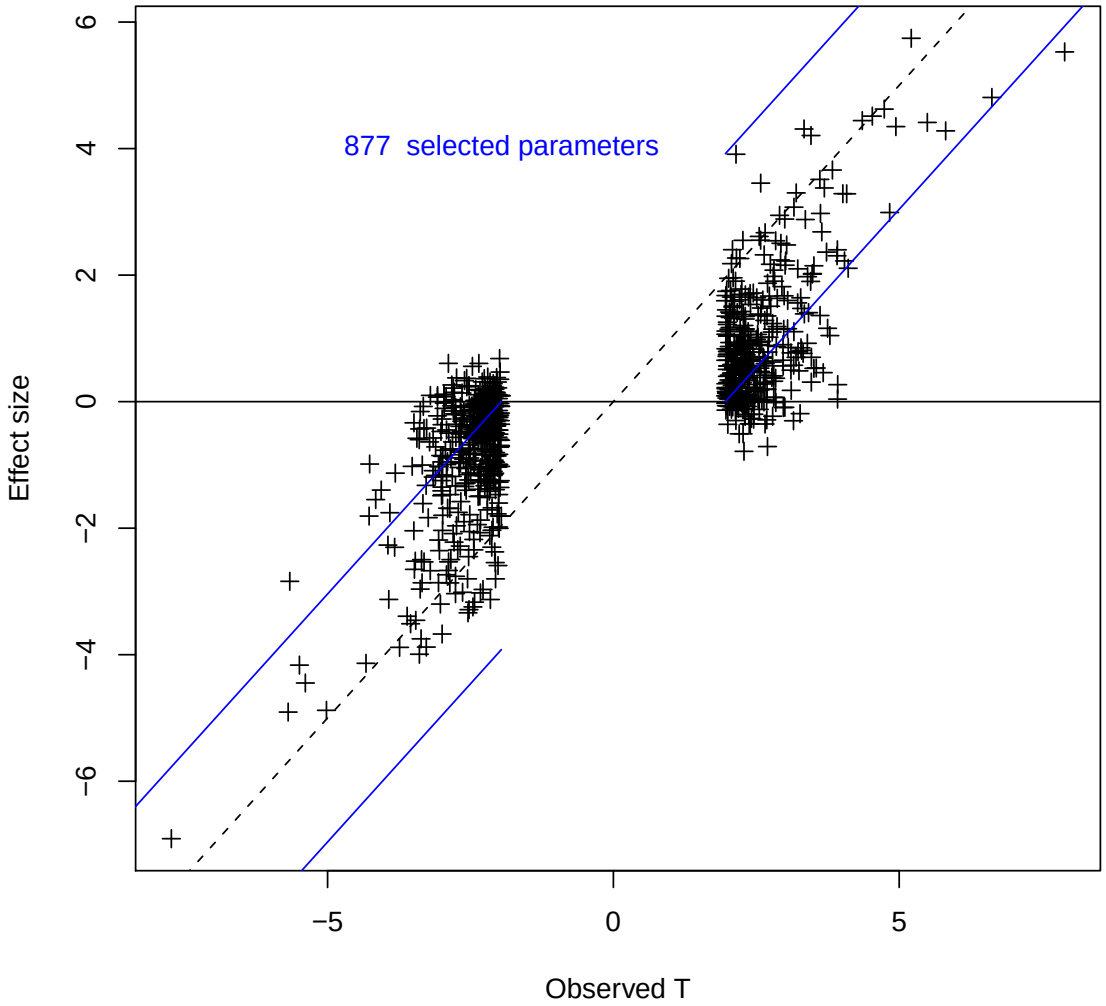
$$\pi(\theta_i) \;=\; 0.9 \cdot \frac{3 \cdot e^{-3 \cdot |\theta_i|}}{2} + 0.1 \cdot \frac{1 \cdot e^{-1 \cdot |\theta_i|}}{2}$$

- Observations $T_i \sim N(\theta_i, 1)$

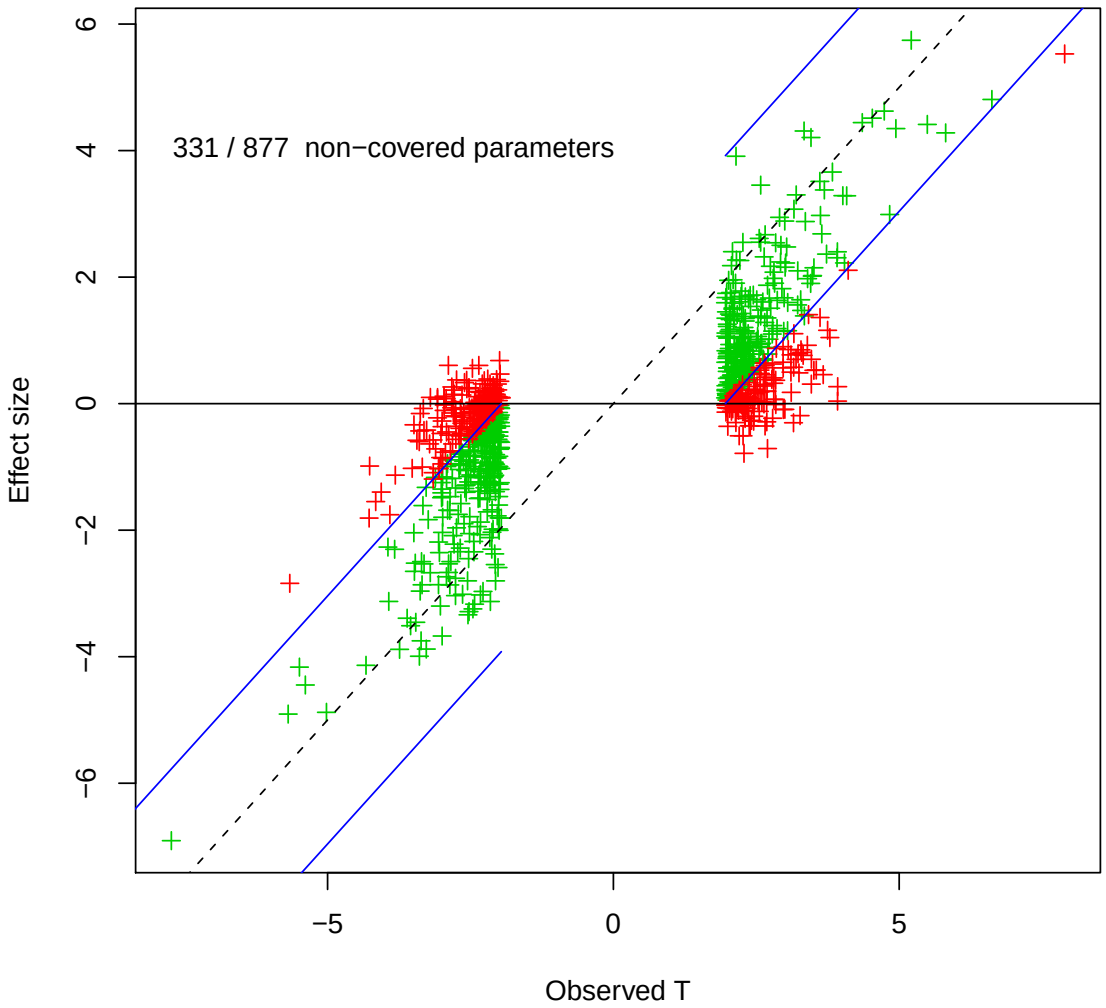
Entire data set



Marginal 0.95 CI's for θ_i with $|T_i| \geq 1.96$



CI's fail to cover 0.95 of the selected parameters



Valid marginal CI's for selected parameters

- Benjamini and Yekutieli '05 suggest the False Coverage-statement Rate as a measure for the validity of CI's constructed for the selected parameters

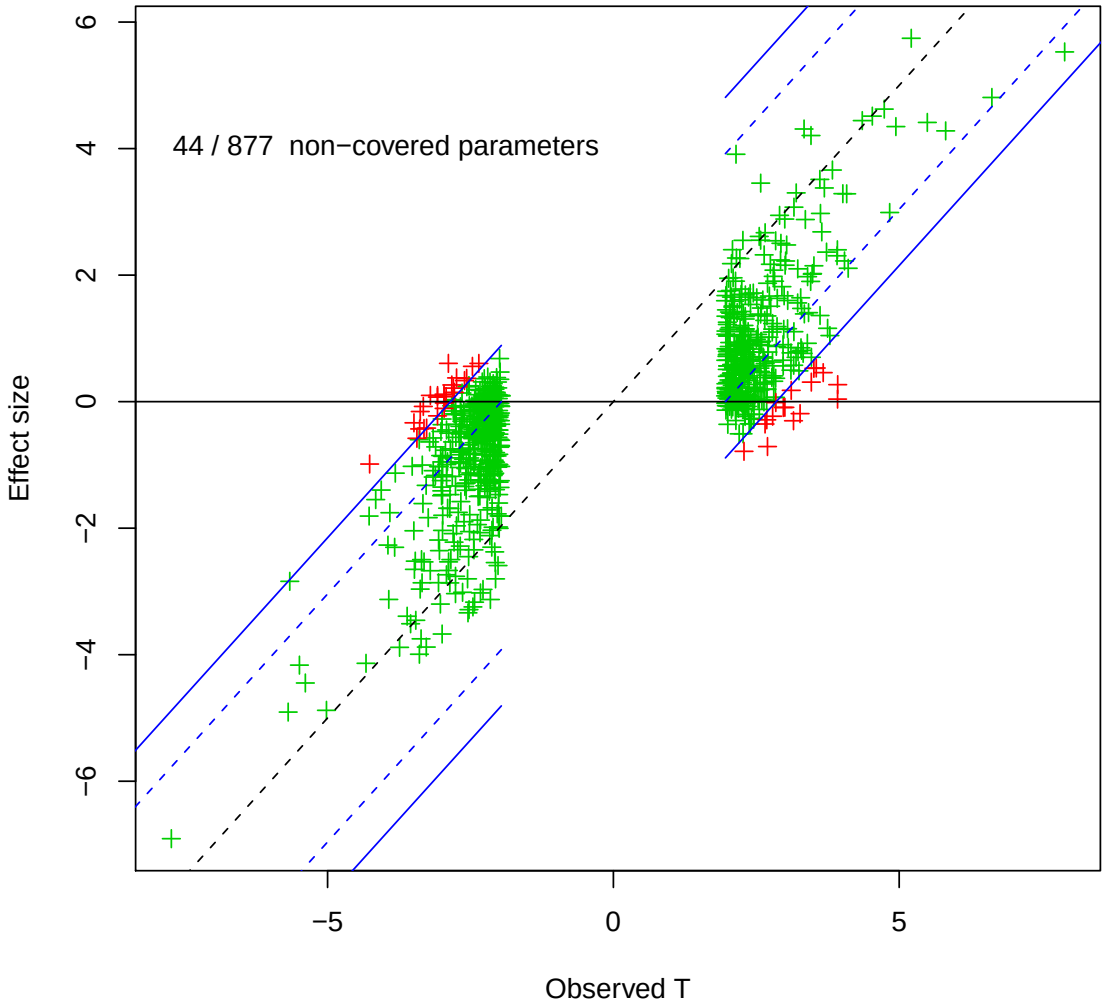
$$FCR = E\{V/\max(R,1)\}$$

where $R = |\mathcal{S}(T_1 \cdots T_m)|$ and V is the number of non-covering CI's

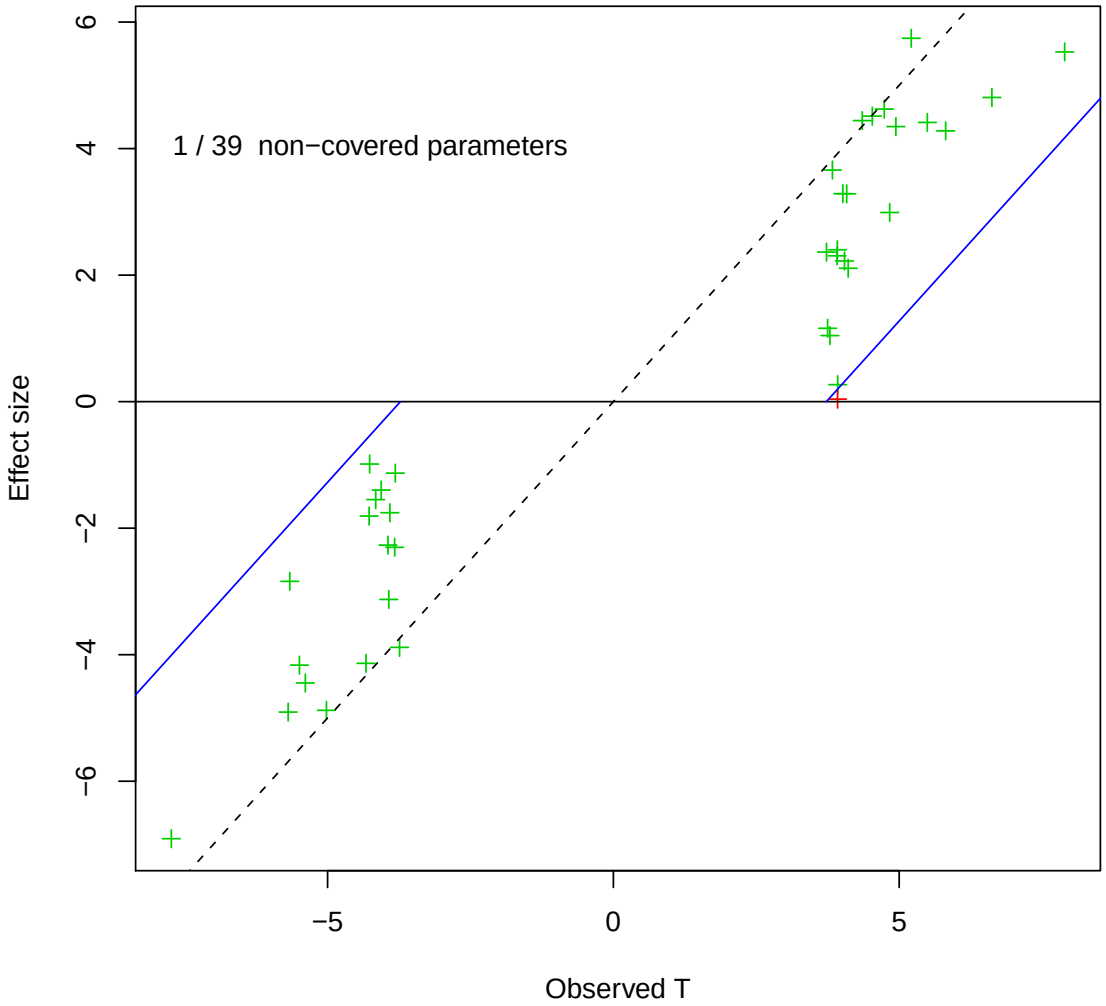
- Main result: for independent $\boldsymbol{T}$ and any selection rule $\mathcal{S}$ constructing marginal $1 - R \cdot q/m$ CI's for each selected parameter ensures $FCR \leq q$

(= *FCR adjusted CI's*)

FCR adjusted CI's for selected parameters



$\Rightarrow$ level 0.05 BH procedure



Some comments

- FDR/FCR control is a frequentist mechanism for ensuring validity statistical inference following selection.
- Alternative option is conditional frequentist CI‘s of Weinstein et al. ‘13
- BH procedure is a classifier (btwn null and non-null effects or positive or negative effects) that ensures that the discoveries are marginally true.
- Bayesian FDR for the two group model (Efron et al ‘01; Storey ‘02-3) is an explicit Bayesian classification framework. Extensions include Bayesian classifiers for dependent effects (e.g. Sun and Cai ’07) and larger (more than 2 point) parameter spaces (e.g. Heller and Yekutieli ‘14).
- Next: Bayesian post-selection inference

Bayesian analysis of continuous parameter simulation

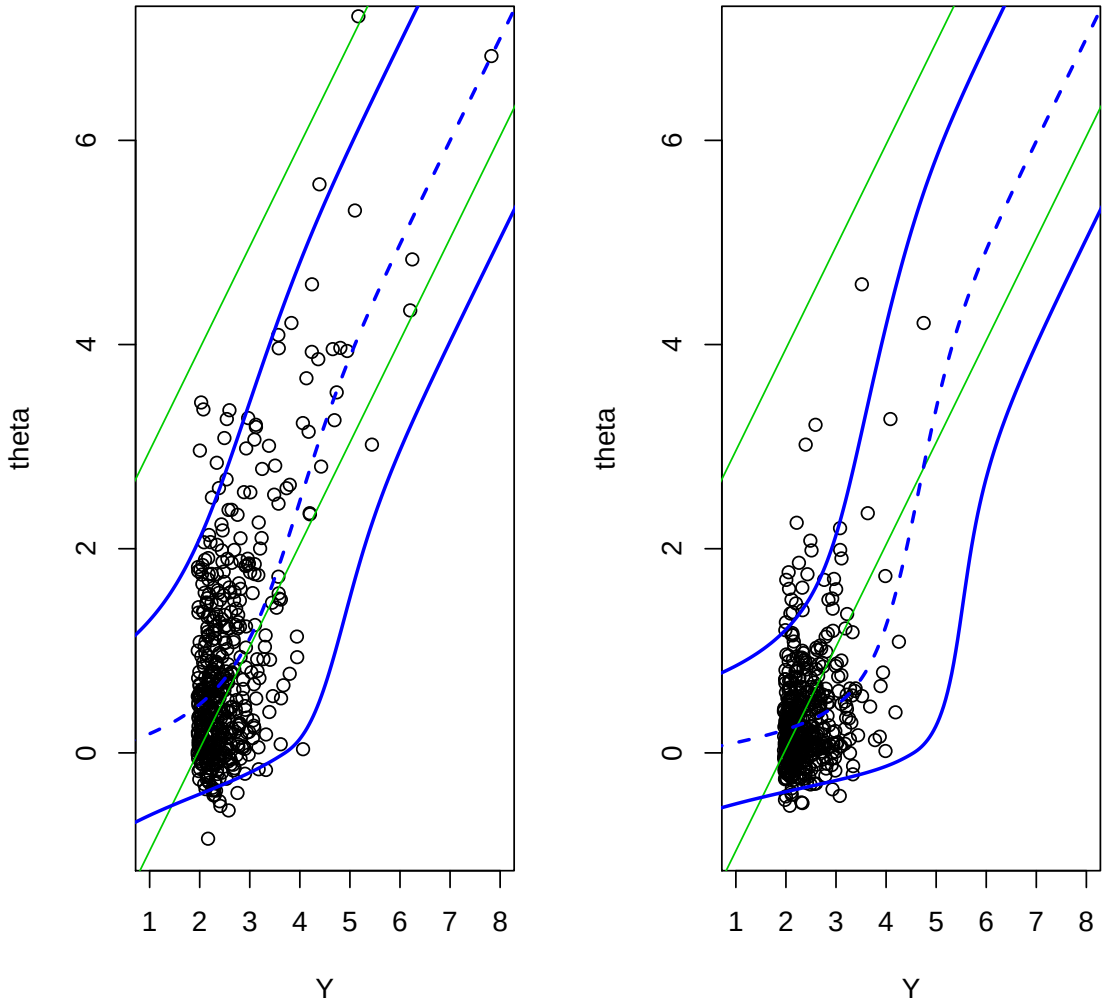
Sample of $m = 10^4$ iid (θ_i, Y_i):

- Prior $\theta_i \sim \pi(\theta_i)$

$$\pi(\theta_i) = 0.9 \cdot \frac{3 \cdot e^{-3 \cdot |\theta_i|}}{2} + 0.1 \cdot \frac{1 \cdot e^{-1 \cdot |\theta_i|}}{2}$$

- Likelihood $f(T_i; \theta_i) = \phi(T_i - \theta_i)$
- Posterior $\pi(\theta_i | T_i) \propto \pi(\theta_i) \cdot f(T_i; \theta_i)$

Joint distribution of (θ_i, T_i)



File drawer version of the simulation

- Prior $\theta_i \sim \pi(\theta_i)$

$$\pi(\theta_i) = 0.9 \cdot \frac{3 \cdot e^{-3 \cdot |\theta_i|}}{2} + 0.1 \cdot \frac{1 \cdot e^{-1 \cdot |\theta_i|}}{2}$$

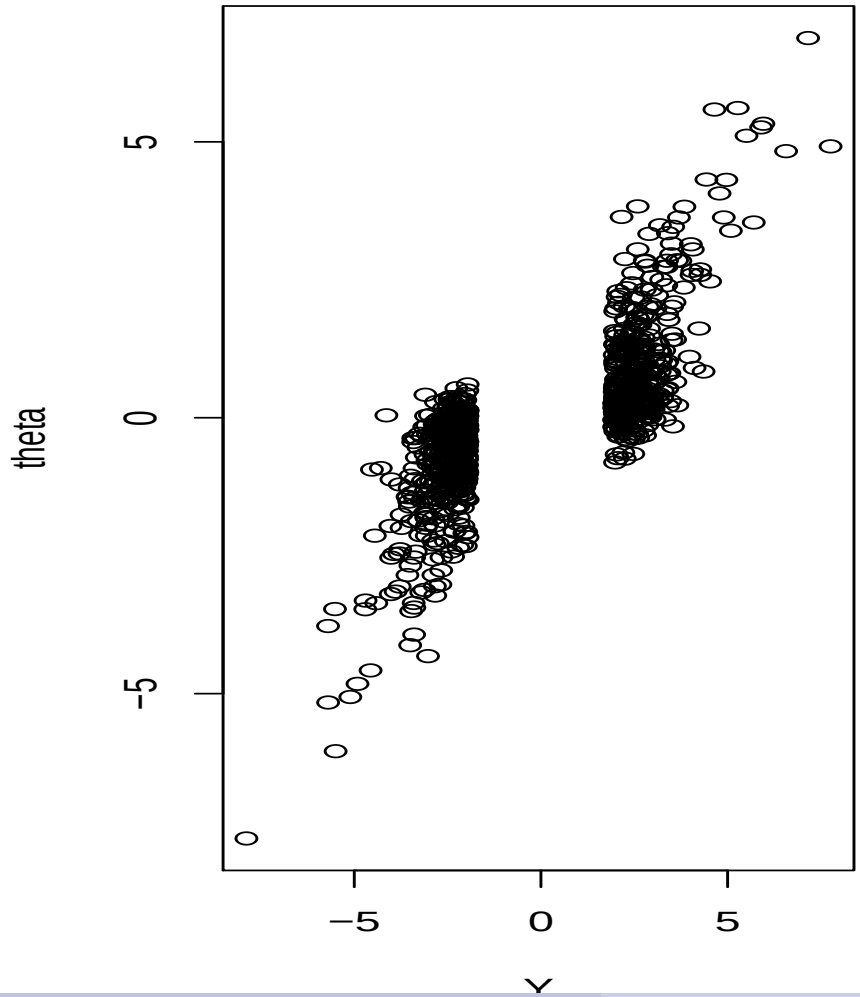
- However, only significant results are available for each parameter-value ($|T_i| \geq 1.96$). Thus T_i is drawn from conditional likelihood

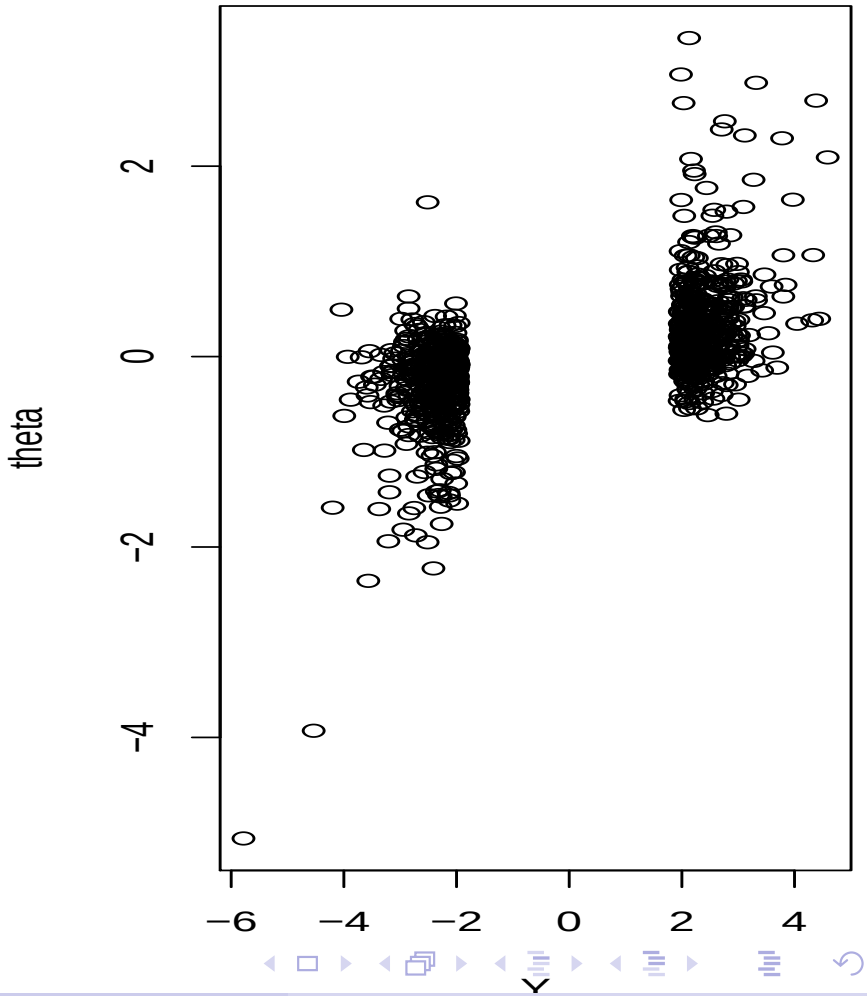
$$f_S(T_i; \theta_i) \propto \frac{\phi(T_i - \theta_i)}{\Phi(-1.96 - \theta_i) + (1 - \Phi(1.96 - \theta_i))}$$

- Posterior distribution is given by

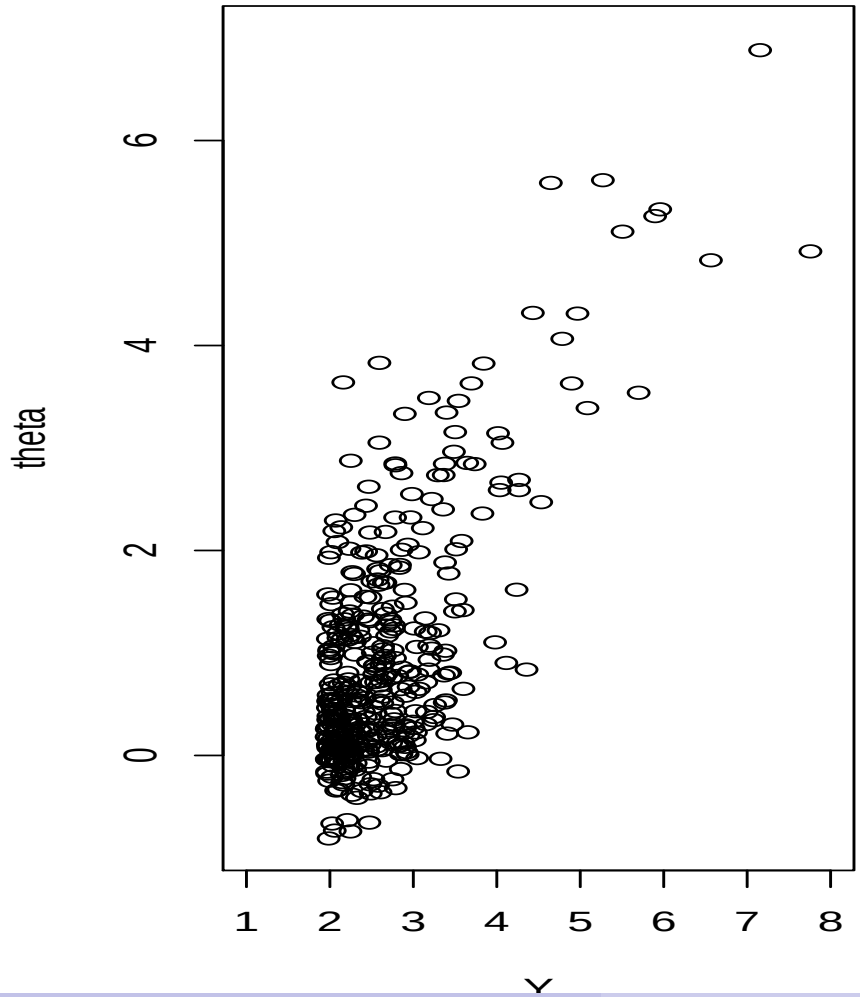
$$\pi_S(\theta_i | T_i) \propto \pi(\theta_i) \cdot f_S(T_i; \theta_i)$$

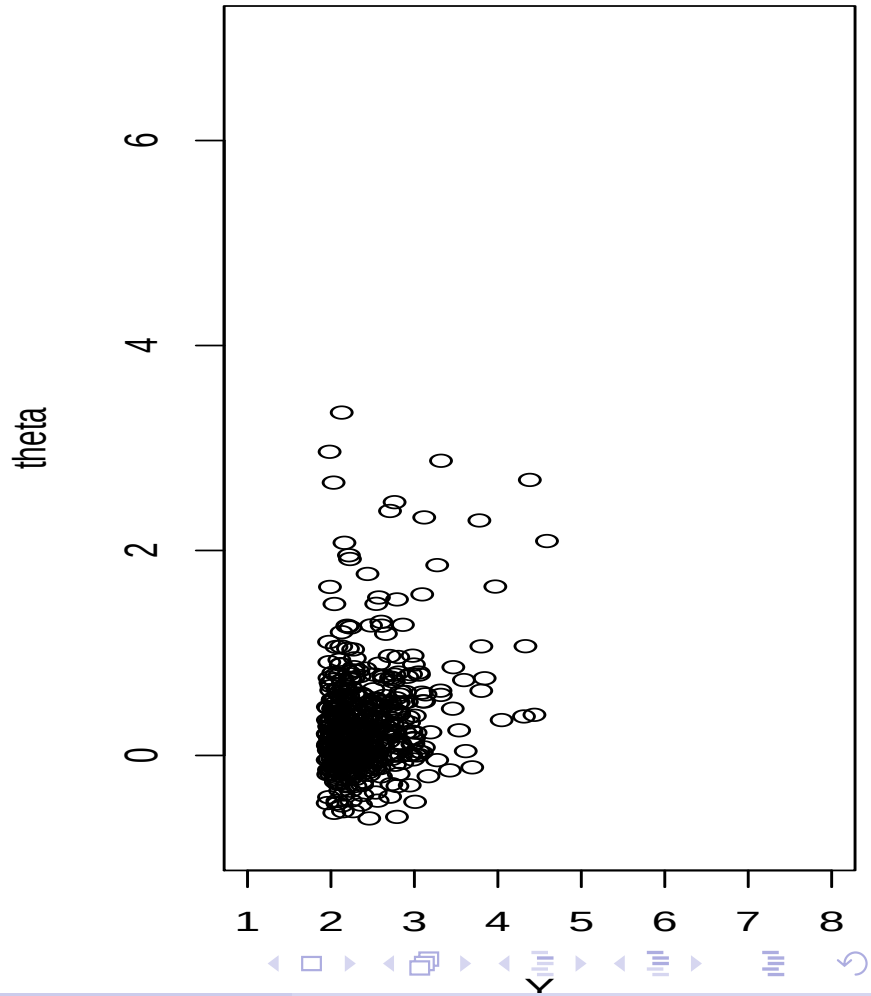
Comparison btwn truncated distributions of (θ_i, T_i)



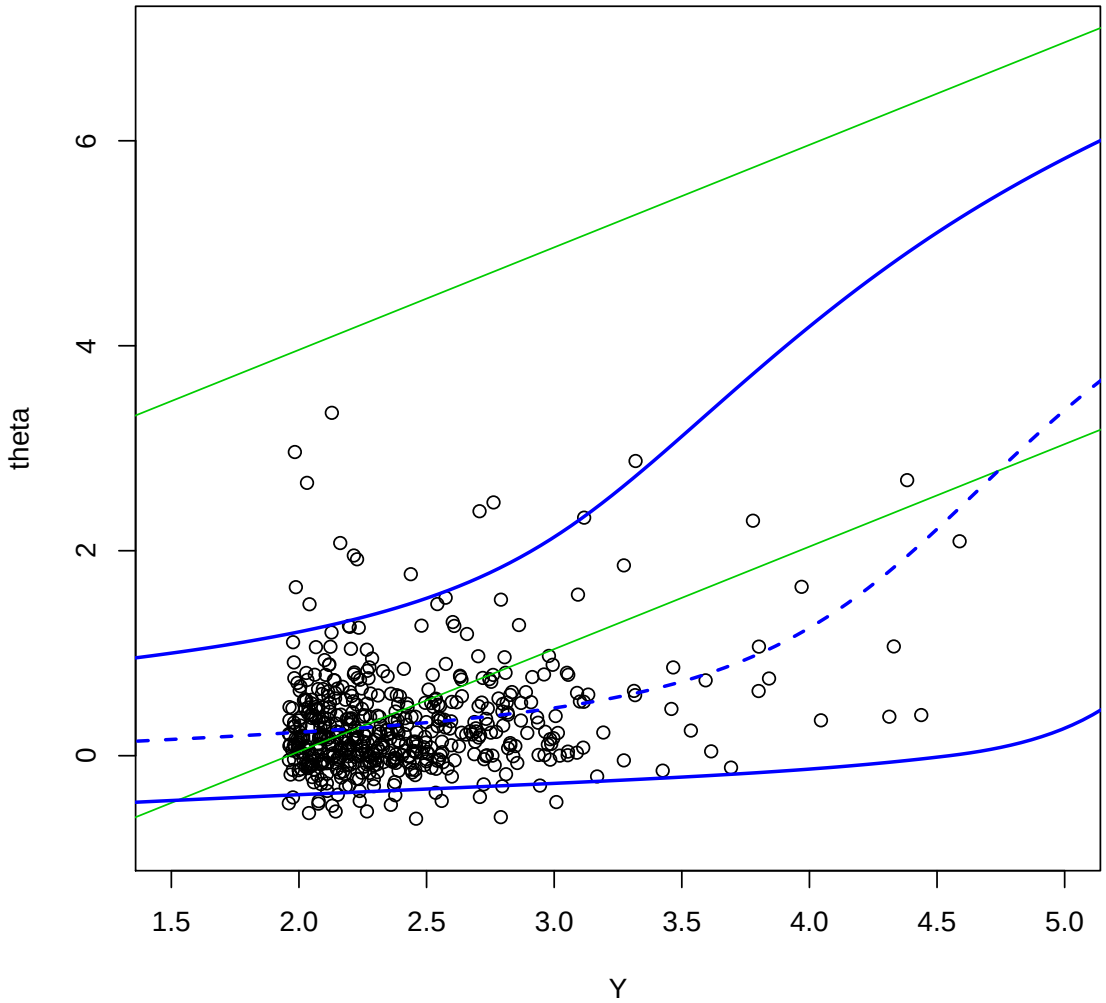


Comparison btwn the positive truncated parts





File drawer simulation – 0.95 credible intervals



Bayesian selective inference

Bayesian selective inference framework:

- θ is the parameter, Y is the data and Ω is the data sample space.
- $\pi(\theta)$ is the prior distribution and $f(y|\theta)$ is the likelihood function.
- The multiple parameters, for which inference may or may not be provided, are actually multiple functions of θ : $h_1(\theta), h_2(\theta), \ldots$
- Each $h_j(\theta)$ corresponds to a subset $S^j_\Omega \subseteq \Omega$, such that inference is provided for $h_j(\theta)$ only if $y \in S^j_\Omega$ is observed.

Bayesian selective inference – a truncated data problem

- As inference is provided for $h(\theta)$ only if $y \in S_\Omega$, then $Y = y$ used for providing selective inference for $h(\theta)$ is a realization of $f_S(\theta, y)$, the joint distribution of (θ, Y) truncated by the event that $y \in S_\Omega$.

- Where we define $f_S(\theta, y)$ through a average risk:
  if selective inference for $h(\theta)$ involves an action $\delta(Y)$ associated with a loss function $L(h(\theta), \delta(Y))$, then $f_S(\theta, y)$ is the distribution over which the expected loss

$$\int_\theta \int_{y \in S_\Omega} f_S(\theta, y) \cdot L(h(\theta), \delta(y))\, dy d\theta \tag{1}$$

  is the average risk incurred in selective inference for $h_j(\theta)$.

Selection-adjusted Bayesian inference

1. The selection-adjusted prior distribution is $\pi_S(\theta)$ the marginal truncated parameter distirubiton
2. The selection adjusted likelihood is the truncated distribution of $Y|\theta$

$$f_S(y|\theta) \;=\; I_{S_\Omega}(y) \cdot f(y|\,\theta) / \Pr(Y \in S_\Omega|\,\theta)$$

3. Bayes rules are based on the selection-adjusted posterior distribution

$$\pi_S(\theta|\,y) \;=\; \pi_S(\theta) \cdot f_S(y|\theta) / f_S(y)$$

because the average risk in (1) can be expressed

$$\int_{y \in S_\Omega^i} f_S(y) \int_\theta \pi_S(\theta|\,y) \cdot L(h_i(\theta), \delta_i(y))\; d\theta\; dy$$

saBayes for the original simulated data

- e.g. average risk for estimating θ is

$$\int_{\theta=-\infty}^{\infty} \int_{1.96 \leq |t|} \frac{\pi(\theta) \cdot \phi(t-\theta)}{\Pr(|T| \geq 1.96)} \cdot (\theta - t)^2 \, d\theta \, dt$$

- Thus

$$f_s(\theta, t) = \pi(\theta) \cdot \phi(t-\theta) / \Pr(|T| \geq 1.96)$$

- Integrating out T

$$\pi_s(\theta) = \pi(\theta) \cdot \Pr(|T| \geq 1.96 | \, \theta) / \Pr(|T| \geq 1.96)$$

- Dividing $f_s(\theta, T)$ by $\pi_s(\theta)$

$$f_s(t|\theta) = \phi(t-\theta) / \Pr(|T| \geq 1.96 | \, \theta)$$

- $\pi_S(\theta|t)$ always proportional to $f_s(\theta, t)$

saBayes for the file drawer simulated data

- e.g. average risk for estimating θ is

$$\int_{\theta=-\infty}^{\infty} \int_{1.96 \leq |t|} \frac{\pi(\theta) \cdot \phi(t-\theta)}{\Pr(|T| \geq 1.96|\theta)} \cdot (\theta - t)^2 \, d\theta \, dt$$

- Thus

$$f_s(\theta, t) = \pi(\theta) \cdot \phi(t-\theta) / \Pr(|T| \geq 1.96|\ \theta)$$

- Integrating out T

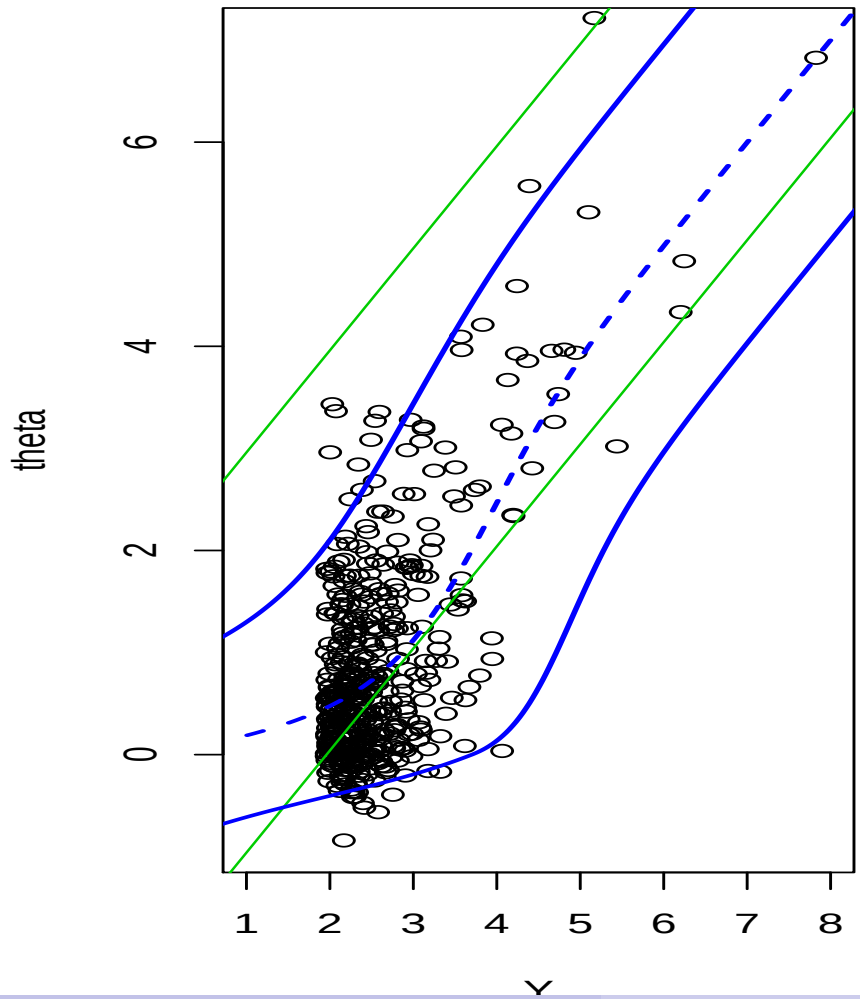
$$\pi_s(\theta) = \pi(\theta)$$

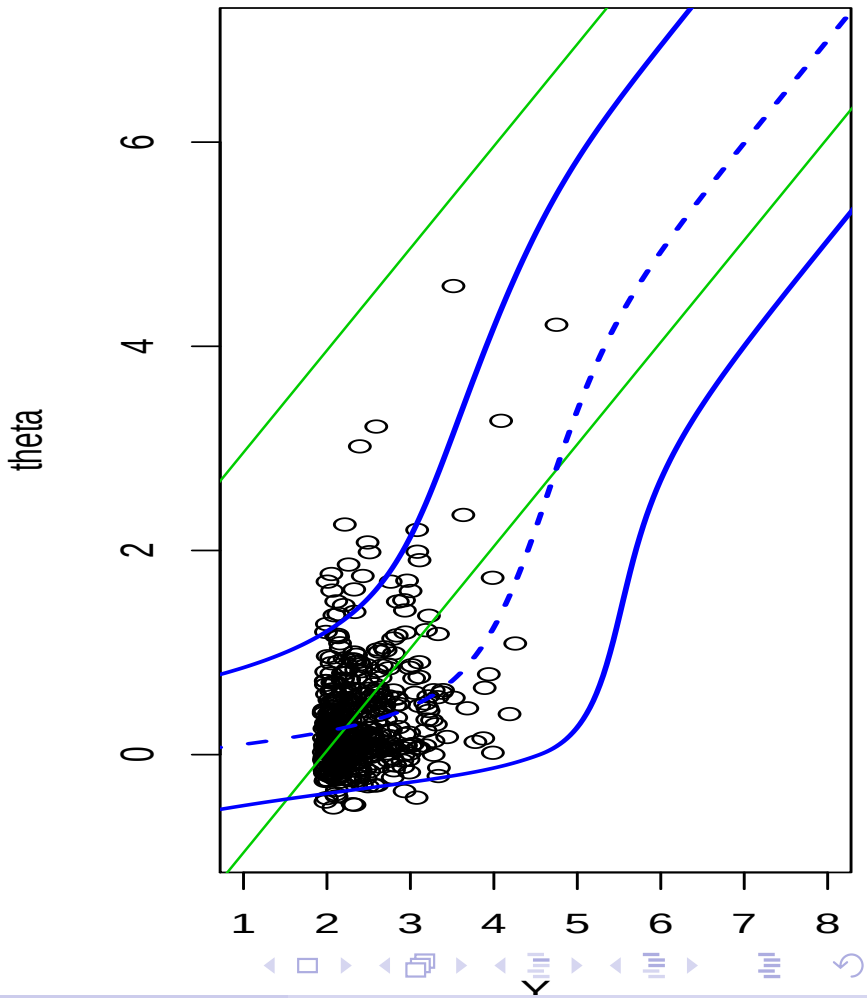
- Dividing $f_s(\theta, T)$ by $\pi_s(\theta)$

$$f_s(t|\theta) = \phi(t-\theta) / \Pr(|T| \geq 1.96|\ \theta)$$

- $\pi_S(\theta|t)$ always proportional to $f_s(\theta, t)$

Comparison of saBayes inferences





saBayes inference for non-informative priors

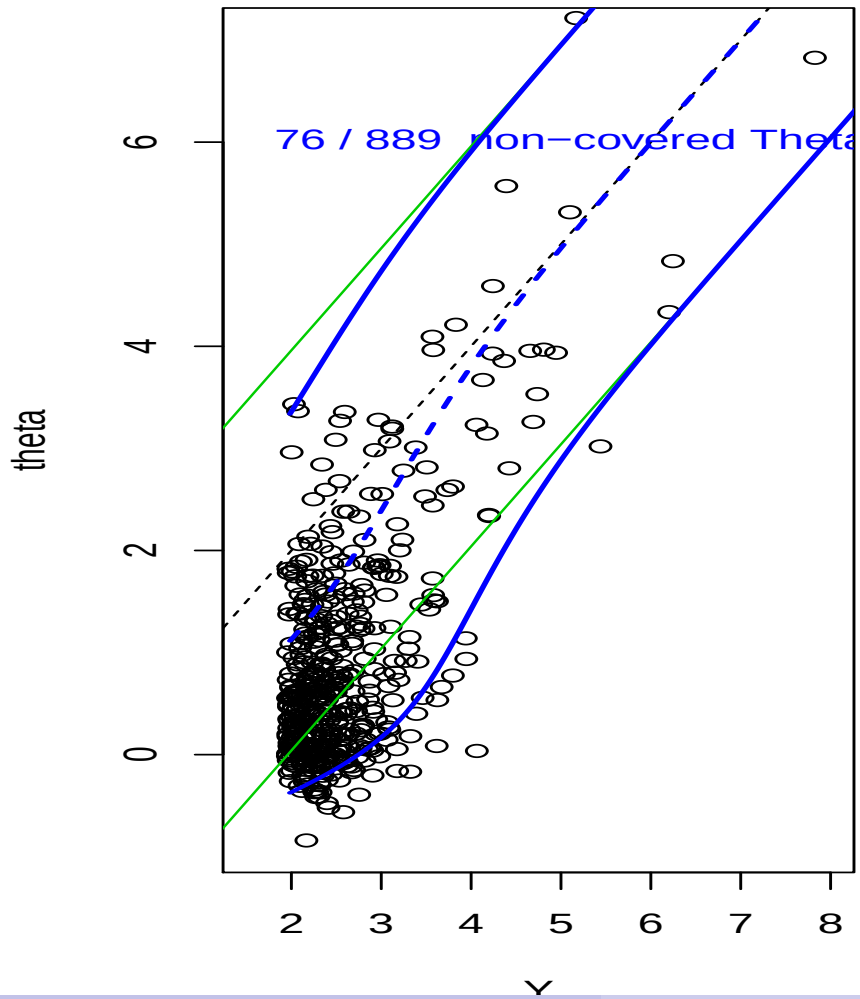
- The non-informative prior is not the marginal distribution of θ. It is used to allow conditional analysis on θ when no prior information on θ is available.
- As Y also provides all the information on θ in the truncated data problem, the prior distribution used for saBayes inference should also be non-informative.
- A simple-minded option is to use the same non-informative prior for the saBayes inference

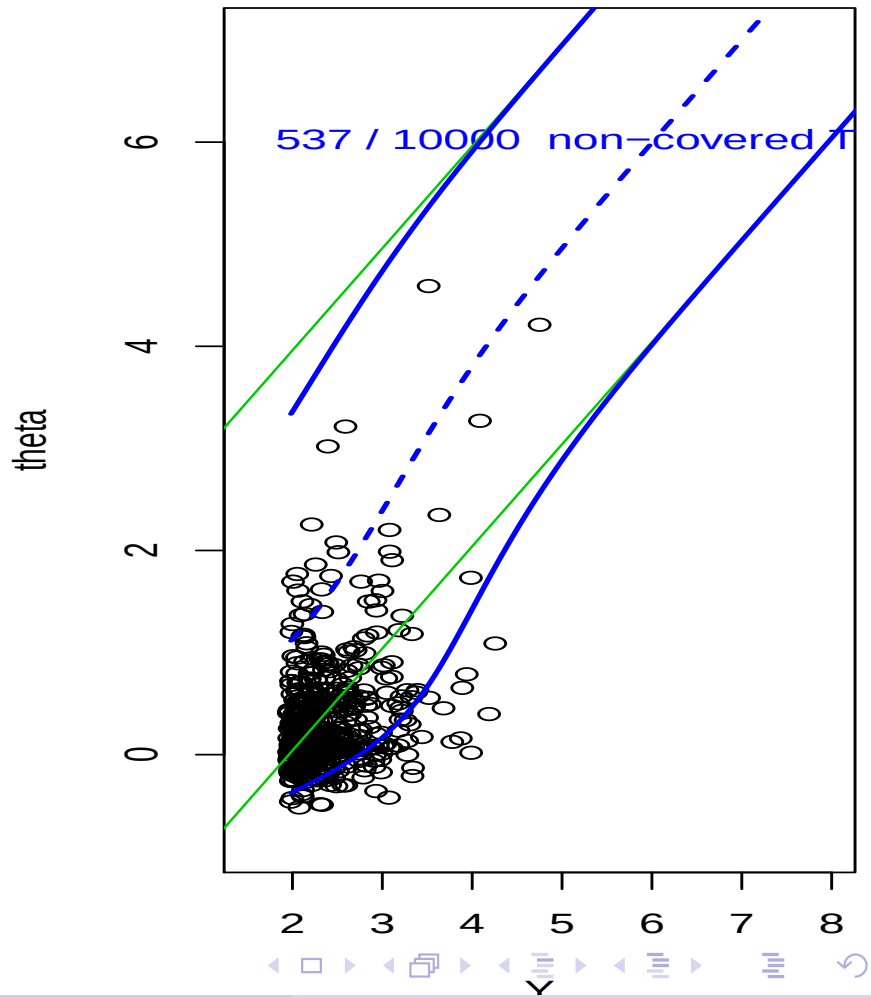
$$\pi_S(\theta) = \pi(\theta)$$

- In our example for flat priors

$$\pi_S(\theta|t) \propto f_S(t|\theta) = \phi(t-\theta)/\Pr(|T| \geq 1.96|\ \theta)$$

Non informative prior saBayes 0.95 CI's





POst Selection Inference (Berk et el. '12)

Full model interpretation of parameters

- The full model is

$$\boldsymbol{Y} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}\,, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I_{n\times n})$$

- Submodels are denoted

$$M = \{j_1 \cdots j_m) \subset \{1 \cdots p\}, \;\; \boldsymbol{X}_M = \{\boldsymbol{X}_{j_1} \cdots \boldsymbol{X}_{j_m}\}.$$

- The target of inference is $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ or some functionals thereof.
- $\boldsymbol{\beta}_M = \{\beta_i : i \in M\}$ selected after viewing the data assumed to consist of the non-zero components of $\boldsymbol{\beta}$.
- Thus, $\boldsymbol{\mu}_M = \boldsymbol{X}_M\boldsymbol{\beta}_M$ are regarded as a computational compression and a parsimonious statistical summary $\boldsymbol{\mu}$, and neither as models in their own right nor as objects of future scientific research.

Suggest an alternative interpretation for selected sub models

The submodel interpretation of parameters

- A submodel M corresponds to a subset of $\boldsymbol{\beta}$ and $\{\beta_i : i \notin M\}$, the deselected components of $\boldsymbol{\beta}$, are non-existent.
- Hence the relevant components are only those in $\boldsymbol{\beta}_M$ and these will generally differ from their siblings in $\boldsymbol{\beta}$.
- Thus selecting model M implies that the target of estimation

$$\boldsymbol{\beta}_M = (\mathbf{X}_M^T \mathbf{X}_M)^{-1} \mathbf{X}_M^T \boldsymbol{X} \boldsymbol{\beta} \quad \Leftrightarrow \quad \boldsymbol{\mu}_M = \boldsymbol{X}_M \boldsymbol{\beta}_M.$$

- With $\beta_{j \cdot M}$ denoting is the coefficient of the j'th predictor in $\boldsymbol{X}$ "adjusted" for the other predictors in M.
- $\boldsymbol{\mu}_M$ is the projection of $\boldsymbol{\mu}$ on the vector space spanned by $\mathbf{X}_M$.

Phrasing POSI as a (Bayesian) selective inference problem

Adopting the Berk et al. '12 submodel Interpretation of parameters:

- Given data generating model

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon},\ \boldsymbol{\mu} = \boldsymbol{X\beta},\ \boldsymbol{\epsilon} \sim N(0, \sigma^2 I_{n\times n}).$$

- The likelihood is $f(\boldsymbol{y}|\ \boldsymbol{\mu}, \sigma^2) = \Pi_{i=1}^{n} \phi((y_i - \mu_i)/\sigma)$.
- (The prior distribution is $\pi(\boldsymbol{\mu}, \sigma^2)$.)
- *There* is a model selection scheme: $M \to S_{\Omega}^{M}$,

$$\boldsymbol{y} \in S_{\Omega}^{M} \quad \Rightarrow \quad \text{target for inference is } \ h_M(\boldsymbol{\mu}) = \boldsymbol{\mu}_M.$$

Berk et al. '12: valid POSI via simultuaneity

- Estimator: $\hat{\boldsymbol{\beta}}_M = (\mathbf{X}_M^T \mathbf{X}_M)^{-1} \mathbf{X}_M^T \boldsymbol{Y}$ with $\hat{\beta}_{j \cdot m} \sim N(\beta_{j \cdot m}, (\mathbf{X}_M^T \mathbf{X}_M)^{-1}_{jj} \sigma^2)$
- A marginal $1 - \alpha$ CI for $\beta_{j \cdot M}$:

$$\hat{\beta}_{j \cdot m} \pm K \cdot \sqrt{(\mathbf{X}_M^T \mathbf{X}_M)^{-1}_{jj} \cdot s^2} \ , \quad K = t_{n-p, 1-\alpha/2}$$

- To ensure valid POSI (for any coeff. in any selected model) Berk et al. '12 propose using a larger $K = K(\boldsymbol{X})$ ensuring simultaneous coverage

$$\Pr\{ \ \forall M, \ \forall j \in M, \ \ \beta_{j \cdot M} \in \hat{\beta}_{j \cdot m} \pm K(\boldsymbol{X}) \cdot S.E.(\hat{\beta}_{j \cdot M}) \ \} \ \geq \ 1 - \alpha$$

Final comments

Work of Wharton group and Stanford group magnificent intellectual and technical achievements.

- Wharton group approach computationally difficult and in some cases not necessary.
- Stanford approach may be more easy to digest if we had a better understanding of the consequences of committing to the single selected model.
- Giant first steps . . .
- Happy New Year!