Series of lectures on Bayesian selective inference
Lecture 4: Selection-Adjusted Bayesian Inference

Daniel Yekutieli

Statistics and OR
Tel Aviv University

Spring School of Research Unit “Structural Inference in Statistics”
17-21 March 2014, Konigs Wusterhausen

Plan

- Post-selection inference
- Predicting academic ability example
- Present selection-adjusted Bayesian approach
- Continuous parameter simulated example
- Analysis of the Dudoit and Yang ‘03 swirl microarray data

Post-selection inference

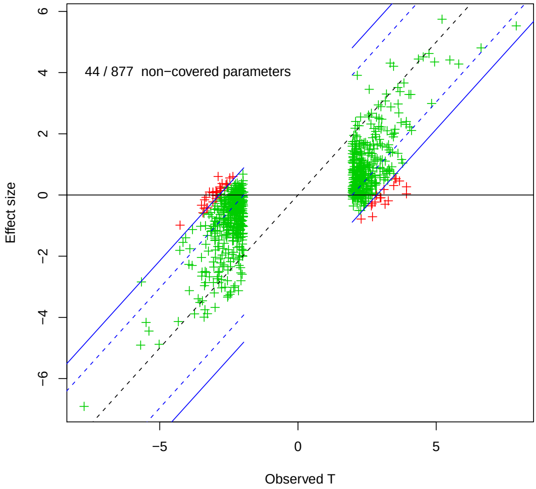
Fundamental problem in statistics: once we use the data to find interesting parameter how do we use the data again to provide valid inferences for these parameters?

- Post model selection inference of Berk et al. ’12
- Overcoming winner’s curse in Economics
- Bias reduced OR estimation in GWAS of Zhong and Prentice ’08
- Conditional frequentist CI’s of Benjamini Filthian and Weinstein ’13
- FCR approach of Benjamini and Yekutieli ’05

Why Bayesian selective inference?

1. Frequentist methods offer limited ad-hoc solutions for selection and post-selection
2. The Bayesian approach yields optimal straightforward optimal selection rules and algorithmically provides comprehensive post-selection inference
3. Sometimes we have no choice: many of the more complex statistical models are inherently Bayesian, in the sense that they include many parameters that are all assigned distributions, thus analyzing these problems means providing Bayesian selective inference

e.g. why improve the FCR approach?



Bayesian parameter selection?

1. Explicit loss minimization: Scott and Berger ‘06 – Discovery of active genes in microarray experiments

   - A gene is declared active ($\theta_i \neq 0$) if the posterior expected loss of this action is smaller than the posterior expected loss of declaring the gene inactive ($\theta_i = 0$)
   - The loss function for erroneously deciding $\theta_i = 0$ is proportional to $|\theta_i|$
   - The loss for erroneously deciding $\theta_i \neq 0$ is the fixed cost of doing a targeted experiment to verify that the gene is in fact active.

2. Bayesian FDR methods

Bayesian post-selection inference?

Not needed.

Dawid ’04 explains “ Since Bayesian posterior distributions are already fully conditioned on the data, the posterior distribution of any quantity is the same, whether it was chosen in advance or selected in the light of the data... ”

Predicting academic ability example

Assume

- True academic ability $\theta \sim N(0,1)$
- Observed academic ability in high school $Y \sim N(\theta,1)$
- Only Students with $0 < Y$ are admitted to college.

We wish to predict a student's true academic ability from his/her observed academic ability – but only if the student is admitted to college

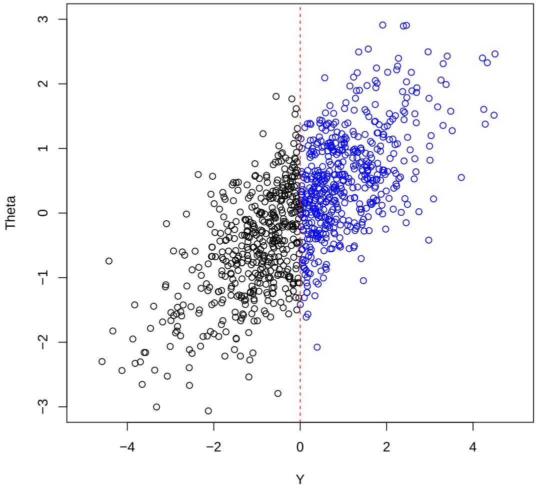
Predicting true academic ability for a college student

Consider the case of a college professor predicting θ for a student in his class

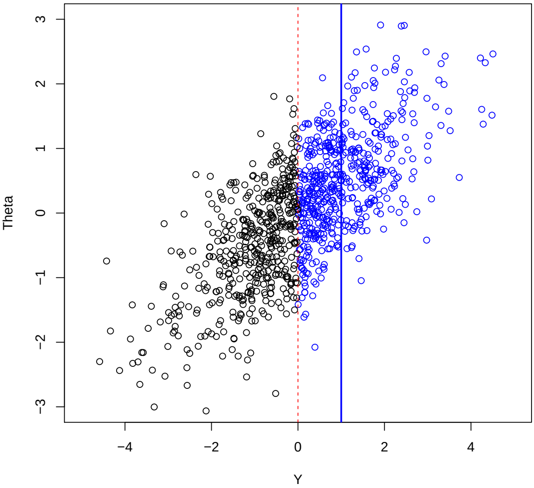
- Joint distribution of (θ, Y) is generated by drawing (θ, y) for a random high school student and keeping it only if $0 < y$.
- Joint density of (θ, y) used for *predicting* θ is

$$f_S(\theta, y) \ \propto \ e^{-\frac{\theta^2}{2}} \cdot e^{-\frac{(\theta-y)^2}{2}} / \Pr(Y > 0) \ \propto \ e^{-\frac{(\theta-y/2)^2}{2\cdot(1/2)}}$$

Distribution of (θ, Y) for a college student(s)



Dawid’s argument



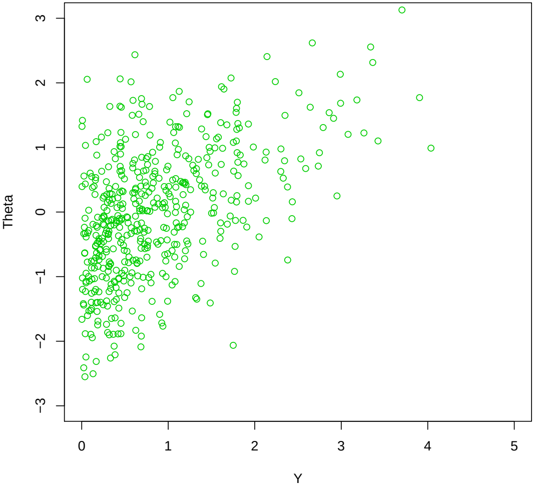
Predicting true academic ability for a high school student

Now consider the case of a guidance counselor predicting θ for a high school student coming to him for job counseling where we assume that there is a high school regulation instructing counselors to predict academic ability only for students that can be admitted to college . . .

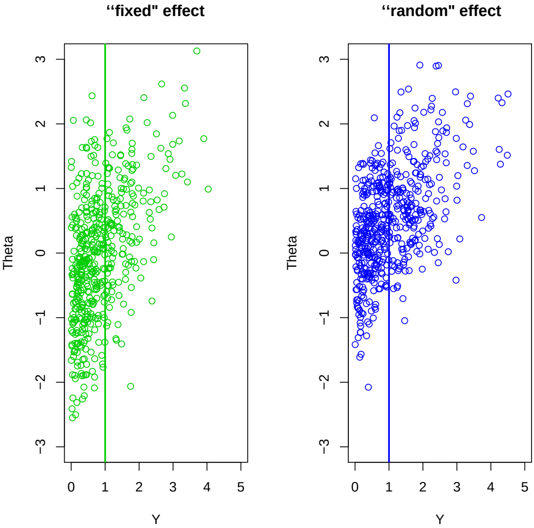
- Thus in this case $\theta \sim N(0, 1)$
- and then Y used to predict θ is drawn from the $N(\theta, 1)$ density truncated (!?) by the event $0 < Y$
- Thus in this case the joint density of (θ, y) used for predicting θ is

$$f_S(\theta, y) \;\propto\; e^{-\frac{\theta^2}{2}} \cdot e^{-\frac{(\theta-y)^2}{2}} / \Pr(0 < Y|\theta)$$

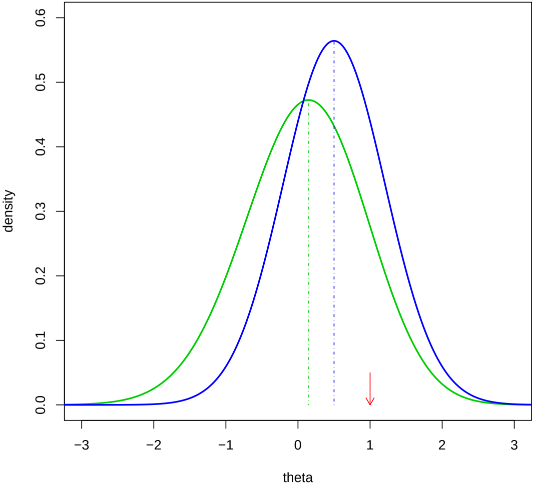
Distribution of (θ, Y) for a high school student



Joint distribution of selected (θ, Y)



Conditional density of $\theta|Y = 1$



A very strange and confusing(?) example

- The college professor and guidance counselor would predict different academic ability for the same student!
- How can that be? The predictions are Bayes rules minimizing different average risks, and a constructive way to verify this would be to explicitly write down the average risk incurred in each case

First case – “high school students admitted to college population average risk”

$$\int_\theta \phi(\theta) \cdot \int_y \phi(y-\theta) \cdot \frac{I(0<y)}{\Pr(0<Y)} \cdot (\theta - \delta(y))^2 \, dy \, d\theta$$

Second case – “particular high school student average risk”

$$\int_\theta \underline{\phi(\theta)} \cdot \int_y \phi(y-\theta) \cdot \frac{I(0<y)}{\Pr(0<Y|\theta)} \cdot (\theta - \delta(y))^2 \, dy \, d\theta$$

Problem formulation

- 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: $h_1(\theta), h_2(\theta), \ldots$.
- For each $h_i(\theta)$ there is a given subset $S^i_\Omega \subseteq \Omega$, such that inference is provided for $h_i(\theta)$ only if $y \in S^i_\Omega$ is observed.

Bayesian selective inference – a truncated data problem

- As inference is provided for $h(\theta)$ only if $y \in S_\Omega$, $Y = y$ 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 $y \in S_\Omega$.
- We define $f_S(\theta, y)$ through the average risk: if selective inference for $h(\theta)$ involves an action $\delta(Y)$ associated with a loss function $L(h(\theta), \delta)$, the $f_S(\theta, y)$ is the distribution over which the expected loss

$$r_S(\delta) = \int_\theta \int_{y \in S_\Omega} f_S(\theta, y) \cdot L(h(\theta), \delta(y))\, dyd\theta$$

  is the average risk incurred in selective inference for $h(\theta)$.
- Selection-adjusted Bayesian (saBayes) inference: Bayesian inference for $h_i(\theta)$ that is based on $f_S(\theta, y)$

Key distinction

- We call θ a “fixed” parameter if its distribution is unaffected by selection and selection is applied to the conditional distribution of Y given θ (e.g. the high school student’s θ).
- θ is a “random” parameter in cases where selection is applied to the joint distribution of (θ, Y) (e.g. the college student’s θ).
- In general the “fixed” parameters are the unknown constants and the “random” parameters are the random effects. In the swirl example we will even consider a “mixed” θ.

The components of saBayes inference

1. *The selection-adjusted prior distribution*, $\pi_S(\theta)$, is the marginal truncated distribution of θ.
2. *The selection adjusted likelihood*, $f_S(y|\theta)$, is the truncated distribution of $Y|\theta$. We will see that conditioning on θ ensures that it is the same regardless whether θ is “random” or “fixed”.
3. *The selection-adjusted posterior distribution*, $\pi_S(\theta|\ y)$, is the truncated conditional distribution of $\theta|Y = y$.

The "fixed" θ case

(Both cases we assume: $\theta \sim \pi(\theta)$, $Y|\theta \sim f(y|\theta)$)

- Marginal truncated distribution of θ is

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

- Truncated conditional distribution of $Y|\theta$

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

- Joint truncated distribution of (θ, Y) is given by

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

The “random” θ case

- The joint truncated distribution of (θ, Y)

$$f_S(\theta, y) = \frac{I_{S_\Omega}(y) \cdot \pi(\theta) \cdot f(y|\, \theta)}{\Pr(S_\Omega)} \tag{1}$$

- The marginal truncated distribution of θ

$$\pi_S(\theta) \ = \int_{S_\Omega} \frac{\pi(\theta) \cdot f(y|\, \theta)}{\Pr(S_\Omega)} dy \ = \ \frac{\pi(\theta) \cdot \Pr(S_\Omega|\, \theta)}{\Pr(S_\Omega)} \tag{2}$$

- Dividing (1) by (2) reveals that the truncated conditional dist. of $Y|\theta$ is the same regardless of whether θ is “fixed” or “random"

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

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.
- Particularly, use the same non-informative prior, i.e. treat θ as if it were a "fixed" effect $\pi_S(\theta) = \pi(\theta)$, thus

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

saBayes inference – Summary

- The selection-adjusted prior distribution is:
  1. For “random” θ: $\pi_S(\theta) \;=\; \pi(\theta) \cdot \Pr(S_\Omega|\,\theta)/\Pr(S_\Omega)$
  2. For “fixed” θ and non-informative priors: $\pi_S(\theta) \;=\; \pi(\theta)$
- The selection adjusted likelihood is

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

- The selection-adjusted posterior distribution is

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

  for $m_S(y) \;=\; \int \pi_S(\theta) \cdot f_S(y|\,\theta)d\theta$
- For “fixed” θ and non-informative priors Bayesian inference has to be corrected for selection

Return to continuous parameter-value simulation

Generate $m = 10,000$ iid (θ_i, Y_i):

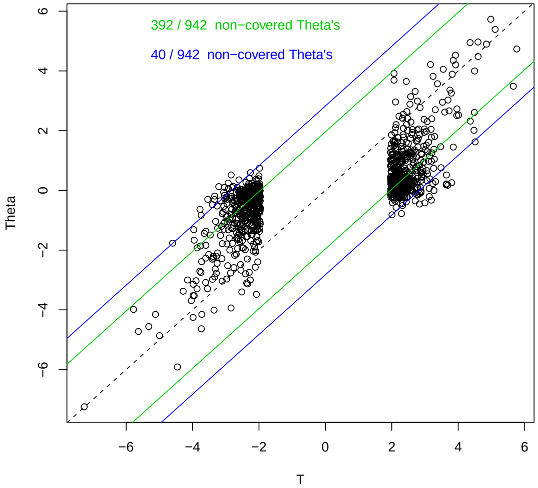
- Parameter $\theta_i \sim \pi(\theta_i)$

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

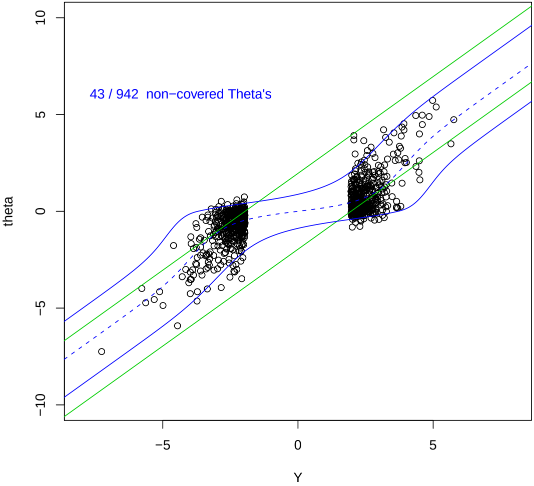
- Observations $T_i \sim N(\theta_i, 1)$

and we construct a marginal 0.95 confidence interval for θ_i only if $1.96 \leq |X_i|$

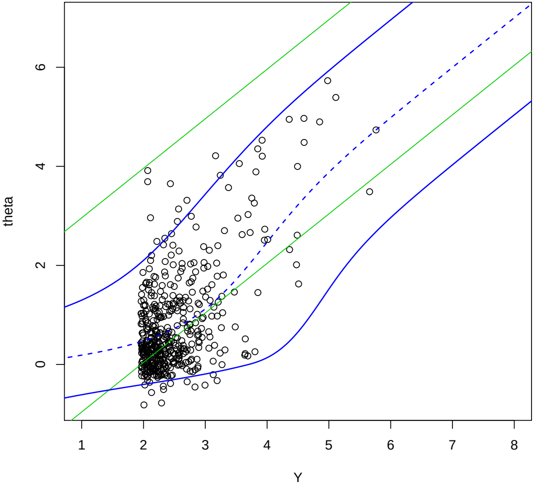
Level 0.05 FCR-adjusted CI's



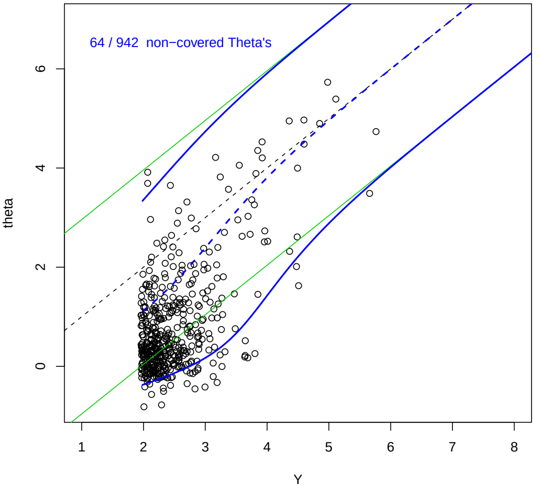
0.95 posterior credible intervals (= “random” θ saBayes)



Detail



Flat prior 0.95 saBayes posterior CI's



Swirl Zebrafish data set (Dudoit and Yang ’03)

The data includes 4, 8448 gene arrays, comparing RNA from Zebrafish with the swirl mutation to RNA from wild type fish

For Gene $g = 1 \cdots G$ ($G = 8448$):

- Parameters:
  1. μ_g expected log2-fold change in expression due to the swirl mutation
  2. σ_g^2 the variance of the log2-fold change in expression
- Statistics:
  1. $\bar{y}_g$ mean observed log2 expression ratios independent $N(\mu_g, \ \sigma_g^2/4)$
  2. s_g^2 sample variances independent $\sigma_g^2 \chi_3^2/3$.
- Likelihood:

$$f(\bar{y}_g, s_g | \, \mu_g, \sigma_g^2) \ \propto \ \sigma_g^{-4} \exp\{-\frac{1}{2\sigma_g^2}[3s_g^2 + 4(\mu_g - \bar{y}_g)^2]\}$$

Frequentists analysis: discovery of differentially expressed genes with directional FDR ≤ 0.05

1. For $g = 1 \cdots G$, compute p-value testing $H^g : \mu_g = 0$

$$t_g = \frac{\bar{y}_g}{s_g/2} \sim t_3 \;\rightarrow\; p_g = 2 \cdot \{1 - F_3(|t_g|)\}$$

2. Apply level $q = 0.10$ BH procedure to $p_1 \cdots p_G$
3. As $F_3^{-1}(1 - 0.1/(2 \cdot 8448)) = 57.10$ while $max(|t_g|) = 27.90$ the BH procedure yields no discoveries.

Hybrid analysis with the *limma* R package (Smyth '05)

Assumption: $\sigma_g^2 \sim \pi_{eB}(\sigma_g^2) = s_0^2 \cdot \nu_0/\chi^2_{\nu_0}$

1. *eBayes* function applied to $s_1^2 \cdots s_{8448}^2$ yields $s_0^2 = 0.052$ and $\nu_0 = 4.02$
2. Compute posterior mean of σ_g^2, $\tilde{s}_g^2 = (\nu_0 s_0^2 + 3s_g^2)/(\nu_0 + 3)$
3. Under $H_{0i} : \mu_g = 0$, moderated t statistic

$$\tilde{t}_g = \bar{y}_g/(\tilde{s}_g/2) \sim t_{7.02} \ \rightarrow \ \tilde{p}_g = 2 \cdot \{1 - F_{7.02}(|\tilde{t}_g|)\}$$

Analysis: Apply level $q = 0.10$ BH procedure to $\tilde{p}_1 \cdots \tilde{p}_{8448}$

Result: 245 discoveries with $|\tilde{t}_g| > 4.479$

Bayesian selective inference

1. For the population of genes in microarray estimate

$$\pi_{eB}(\sigma_g, \mu_g) = \pi_{eB}(\sigma_g) \cdot \pi_{eB}(\mu_g)$$

   and use it to derive selection rules with $Fdr = 0.05$ for classifying μ_g for each gene as positive or negative

2. Provide saBayes inference for selected genes on $h_g(\sigma_g^2, \mu_g) = \mu_g$:
   - σ_g^2 is a “random” effect with

$$\sigma_g^2 \sim \pi_{eB}(\sigma_g^2)$$

   - μ_g is a “fixed” effect for which we assign a flat prior

$$\pi_{ni}(\mu_g) \propto 1.$$

Specifying the selection rule

The eBayes prior chosen is

$$\pi_{eB}(\mu_g) = 8.5 \cdot \exp(-8.5 \cdot |\mu_g|)/2.$$

And now that for the population of genes in the microarray we have

$$f(\bar{y}_g, s_g; \mu_g, \sigma_g) = \pi_{eB}(\mu_g) \cdot \pi_{eB}(\sigma_g) \cdot f(\bar{y}_g, s_g | \, \mu_g, \sigma_g)$$

- we can use it to compute the local-fdr

$$fdr(\bar{y}_g, s_g^2) = \Pr\{sign(\mu_g) \neq sign(\bar{y}_g) | \, \bar{y}_g, s_g^2\}$$

- The directional Bayes Fdr of any selection rule S

$$Fdr_S = E_{(\bar{y}_g, s_g^2) \in S}\{fdr(\bar{y}_g, s_g^2)\} = \Pr\{sign(\mu_g) \neq sign(\bar{y}_g) | \, (\bar{y}_g, s_g^2) \in S\}$$

Selection rules

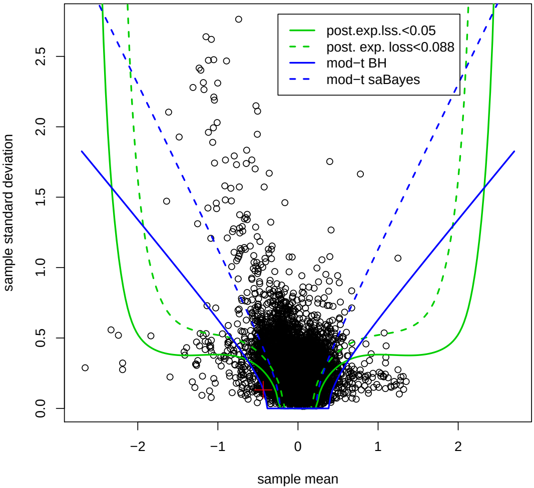
Moderated *t* statistic selection rules

1. The level $q = 0.10$ BH procedure selection rule is $|\tilde{t}_g| > 4.48$, its *Fdr* is 0.024, and it yields 245 discoveries
2. $|\tilde{t}_g| > 2.64$ is the selection rule with $Fdr = 0.05$, it yields 1124 discoveries

Selection rule based on the local FDR

1. $fdr(\bar{y}_g, s_g) < 0.05$ yields 559 discoveries.
2. $fdr(\bar{y}_g, s_g) < 0.088$ is the optimal $Fdr = 0.05$ selection rule, it yields 1271 discoveries.

Selection rules on the scatterplot of $(\bar{y}_g, s_g)$



saBayes inference for μ_{6239} for the moderated t statistic selection rules

Data for Gene 6239:

$$\bar{y}_{6239} = -0.435, s^2_{6239} = 0.0173 \Rightarrow \tilde{s}^2_{6239} = 0.037, \tilde{t}_{6239} = -4.51$$

- saBayes posterior using the eBayes prior model for μ_g

$$\pi_{eB}(\sigma^2_g) \cdot \pi_{eB}(\mu_g) \cdot f(\bar{y}_g, s^2_g | \mu_g, \sigma^2_g)$$

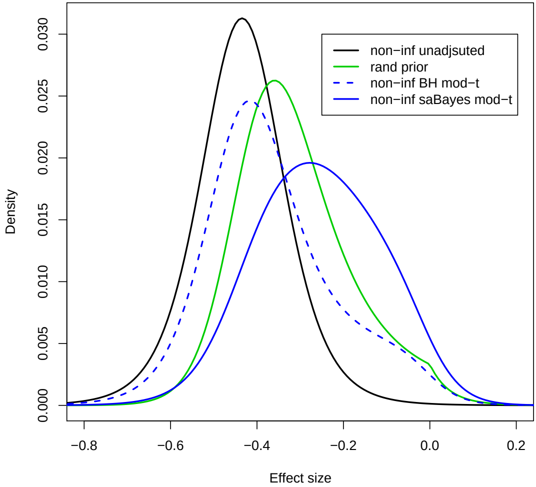
- Unadjusted posterior using the non-informative prior model for μ_g

$$\pi_{eB}(\sigma^2_g) \cdot \pi_{ni}(\mu_g) \cdot f(\bar{y}_g, s^2_g | \mu_g, \sigma^2_g)$$

- saBayes posterior using the non-informative prior model for μ_g

$$\pi_{eB}(\sigma_g) \cdot \pi_{ni}(\mu_g) \cdot f(\bar{y}_g, s_g | \mu_g, \sigma_g) / \Pr(|\tilde{t}_g| > a \mid \mu_g)$$

Marginal posterior distribution of μ_{6239}



Results

1. Unadjusted posterior using non-informative prior model for μ_g
   - $(\mu_{6239} - \bar{y}_{6239})/(\tilde{s}_{6239}/2) \sim t_{7.02}$.
   - Posterior mean and mode equal $Y_{6239} = -0.435$;
   - 0.95 cred. int. $[-0.61, -0.21)$, $\Pr(\mu_{6239} > 0) = 0.0014$.
2. eBayes prior for μ_g
   - Posterior mode is -0.36, Posterior mean is -0.31
   - 0.95 cred. int. $[-0.54, -0.01]$, $\Pr(\mu_{6239} > 0) = 0.020$.
3. saBayes posterior with non-informative prior for μ_g – $|\tilde{t}_g| > 2.64$
   - Posterior mode is -0.278, the posterior mean is -0.257
   - 0.95 cred. int. $[-0.54, 0.02]$, $\Pr(\mu_{6239} > 0) = 0.038$
4. saBayes posterior with non-informative prior for μ_g – $|\tilde{t}_g| > 4.48$
   - Posterior mode is -0.419, posterior mean is -0.367
   - 0.95 cred. int. $[-0.63, -0.02]$, $\Pr(\mu_{6239} > 0) = 0.017$

Summary

- In Bayesian analysis of large data sets, for each potential parameter, it is necessary to explicitly consider a selection rule that determines when inference is provided for the parameter and to provide inference that is based on the selection-adjusted posterior distribution of the parameter.
- Specifying a selection rule introduces an arbitrary element to Bayesian analysis. However, the selection rule is determined according to the prior distributions, and once the selection rule is determined the entire process of providing saBayes inference is fully specified and is carried out the same way as Bayesian inference.
- Specifying the selection rule involves a tradeoff between allowing too many false (or wasteful) discoveries and failing to make discoveries and reduction in the information in the selection-adjusted likelihood.