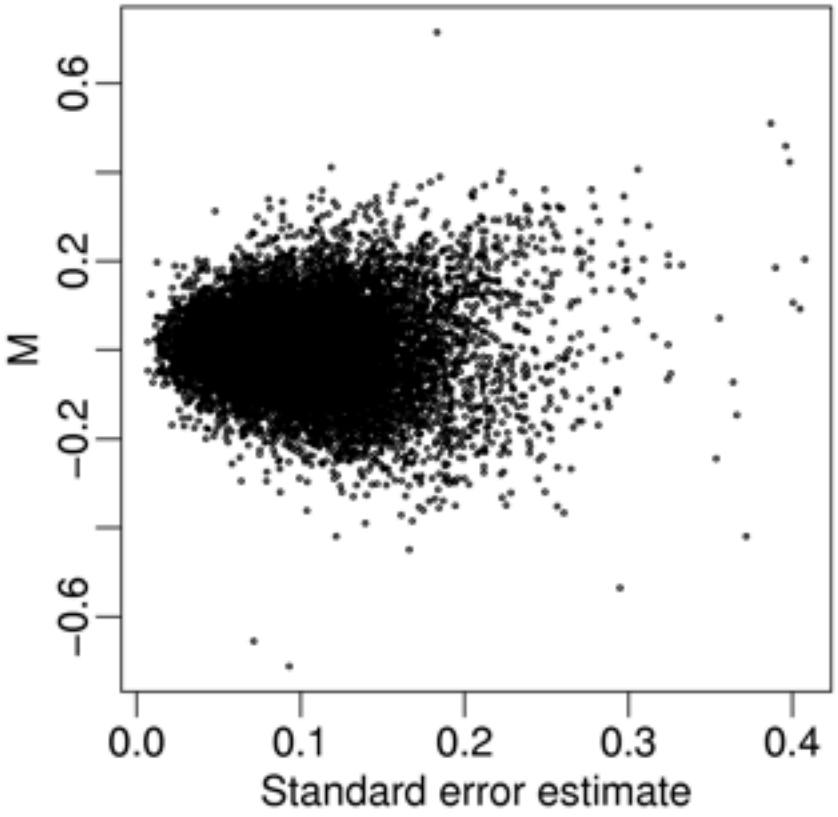
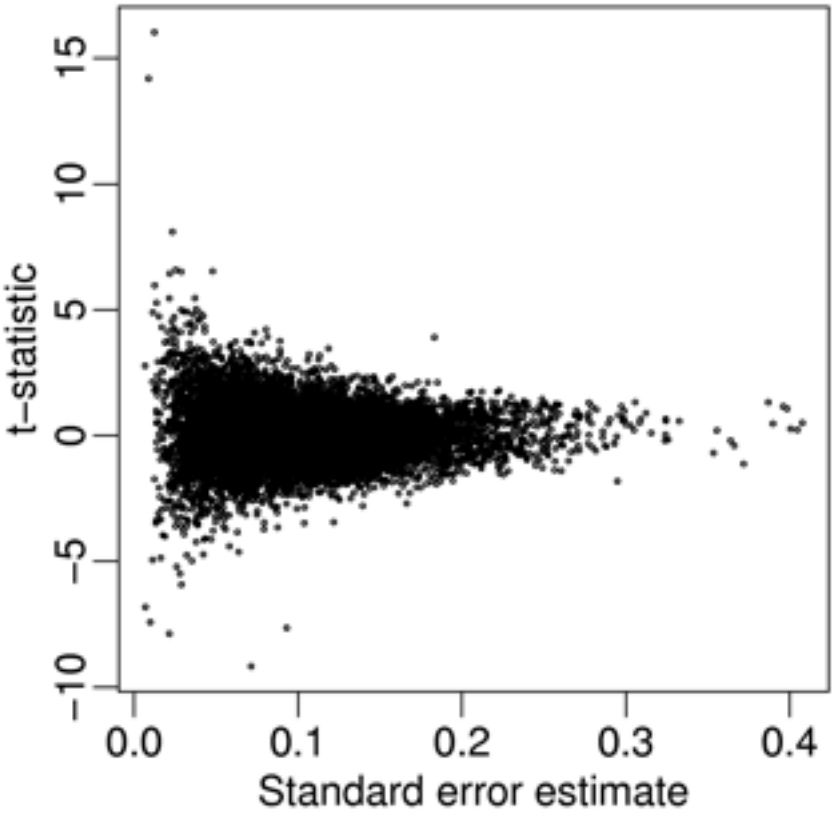
CMSC 702
Gene Expression Analysis

Differential Expression II

Data Show Problems





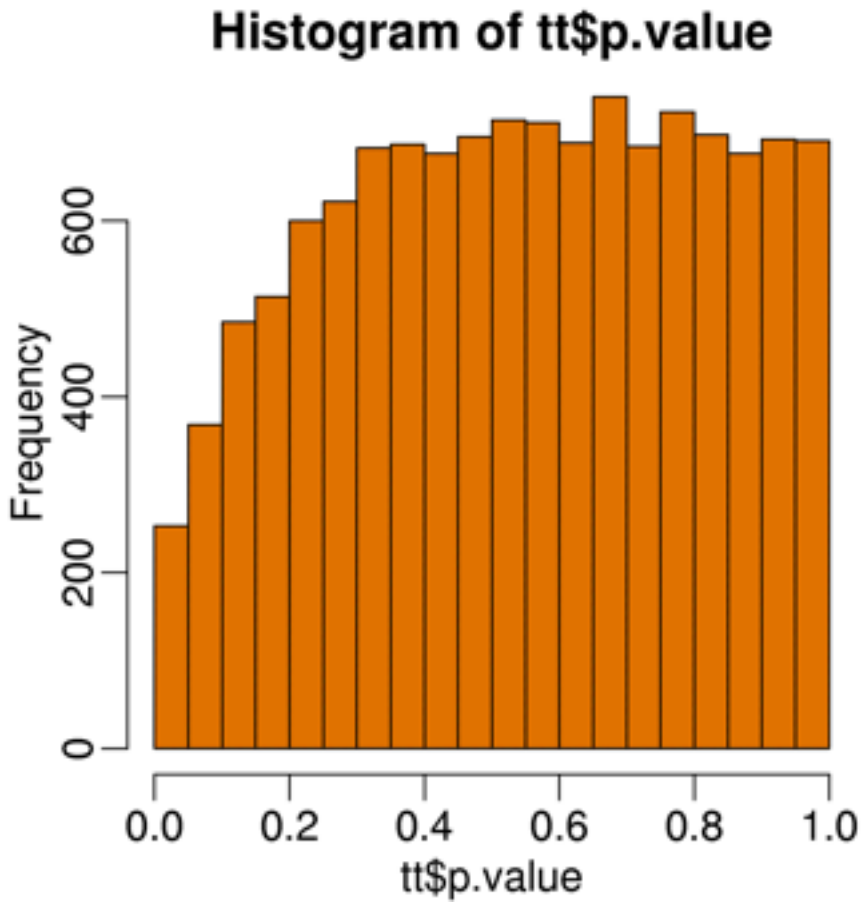
Problems

- **Problem 1: T-statistic bigger for genes with smaller standard errors estimates**
- **Implication: Ranking might not be optimal**

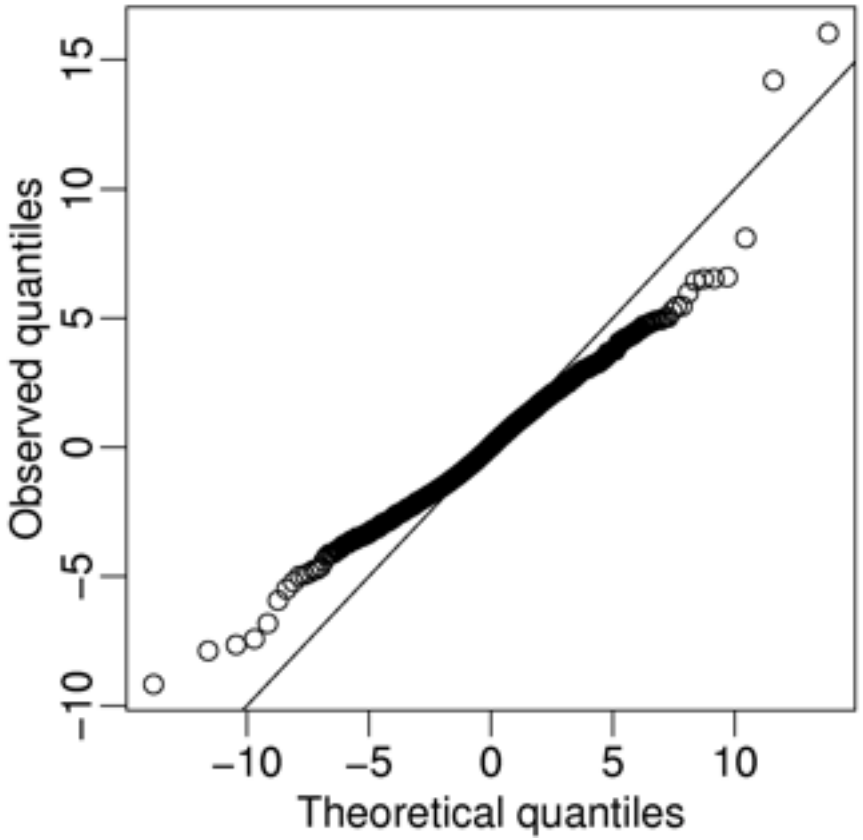
Problem 1

- **With few replicates SE estimates are unstable**
  - **The t-test divides by SE**
  - **This explains why t-test is not powerful**
- **Empirical Bayes methodology provides a statistically rigorous way of improving this estimate**
  - ***Borrowing strength across genes to estimate variance***
- **SAM, a more ad-hoc procedure, works well in practice**

Data Show Problems



Data Show Problems



Problems

- **Problem 1: T-statistic bigger for genes with smaller standard errors estimates**
- **Implication: Ranking might not be optimal**

- **Problem 2: T-statistic not t-distributed.**
- **Implication: p-values/inference incorrect**

Problem 2

- **Even if we use a parametric model to improve standard error estimates, the assumptions might not be good enough to provide trust-worthy p-values**

- **We will describe non-parametric approaches for obtaining p-values**

One final problem

- **Say we are interested in statistical inference, we need to define statistical significance. If we are ranking we may need to define a cut-off that defines *interesting enough***
- **The naïve answer to determining a cut-off is the p-values. Are they appropriate?**
- **Test for each gene null hypothesis: no differential expression.**
- **Notice that if you look at 10,000 genes for which the null is true you expect to see 500 attain p-values of 0.05**
- **This is called the multiple comparison problem. Statisticians fight about how to solve it, but not about the problem above.**
- **Main message: p-values can't be interpreted in the usual way**

So what's next?

1. **Hierarchical/Bayesian methods to estimate variance (and *Empirical Bayesian* ways to produce estimates)**
2. **Non-parametric ways of obtaining *p-values***
3. **Corrections for multiple testing**

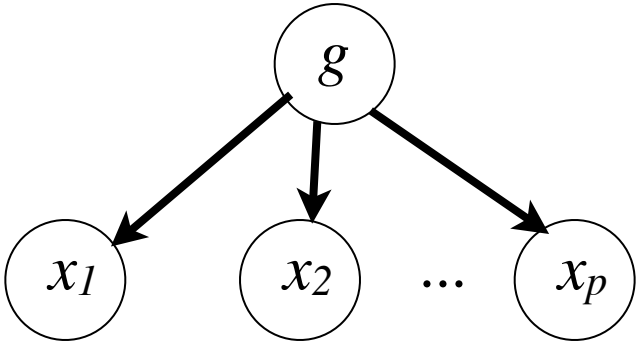
A bag of tricks for the big data analyst!

Introduction to Empirical Bayes

Naive Bayes Classifiers

Consider the "hector-recognition" computer vision problem: *is this a picture of hector?*

- Input: pixel intensities (continuous) x_i
- Output: yes/no (categorical) g

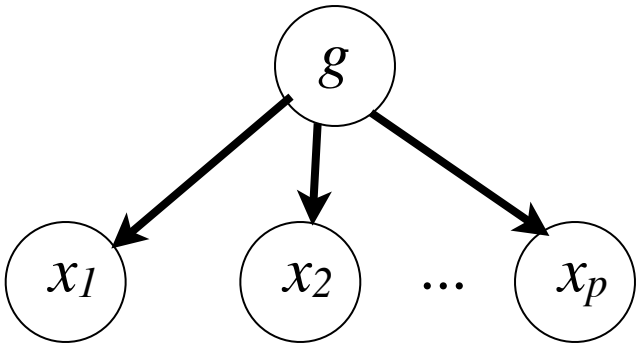


Decision: $P(g=yes \mid x_1, \ldots, x_p) > P(g=no \mid x_1, \ldots, x_p)$?

Naive Bayes Classifiers

Consider the "hector-recognition" computer vision problem: *is this a picture of hector?*

Decision: $P(g=yes \mid x_1, \ldots, x_p) > P(g=no \mid x_1, \ldots, x_p)$?



- Defining $P(g=yes \mid x_1, \ldots, x_p)$ is hard
- Defining $P(x_1, \ldots, x_p \mid g=yes)$ is easier

$$P(x_1, \ldots, x_p | g = yes) = \prod_{j=1}^{p} f(x_j | g = yes)$$

$$X_j | G \sim N(\mu_j, \sigma_j^2)$$

But if we have few pictures of Hector, we may not have enough data *per pixel* to estimate means. Then what?

Bayes Rule

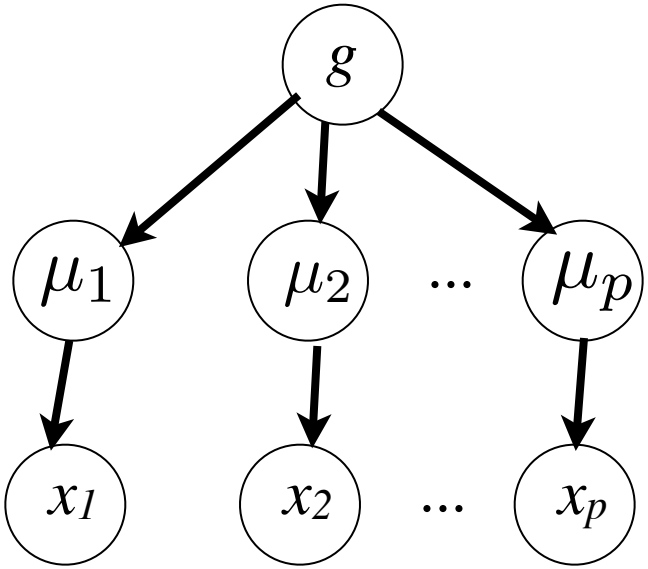
$$P(G|X) \propto P(X|G)P(G)$$

Posterior Likelihood Prior

- Statisticians use it as a way of *bringing history* into their analyses
- Priors may be fully specified (*full Bayesian*), or its parameters estimated from data *(empirical Bayesian*)
- The subject of much debate in statistics (Frequentist/ Bayesian argument)
- Useful way of thinking about *hierarchical models*

Robust Naive Bayes Classifier

Consider the "hector-recognition" computer vision problem: *is this a picture of hector?*



Many parameters to estimate (as many means as pixels), not that much data (there aren't as many images)

Idea: Pixels behave similarly, can we *pool* data to estimate pixel means

$$\mu_j | G \sim N(\mu_0, \tau^2)$$

$$X_j | \mu_j \sim N(\mu_j, \sigma_j^2)$$

BASIC TWO-STAGE SAMPLING

$$\theta \sim G$$
$$Y \mid \theta \sim f(y \mid \theta)$$

- G is the prior
- f is the sampling distribution
- Use the "rules of probability" to get the:

Posterior Distribution

$$g(\theta \mid Y) = \frac{f(y|\theta)g(\theta)}{f_G(Y)}$$

Marginal Distribution

$$f_G(Y) = \int f(y \mid u)g(u)du$$

THE BASIC GAUSSIAN/GAUSSIAN MODEL

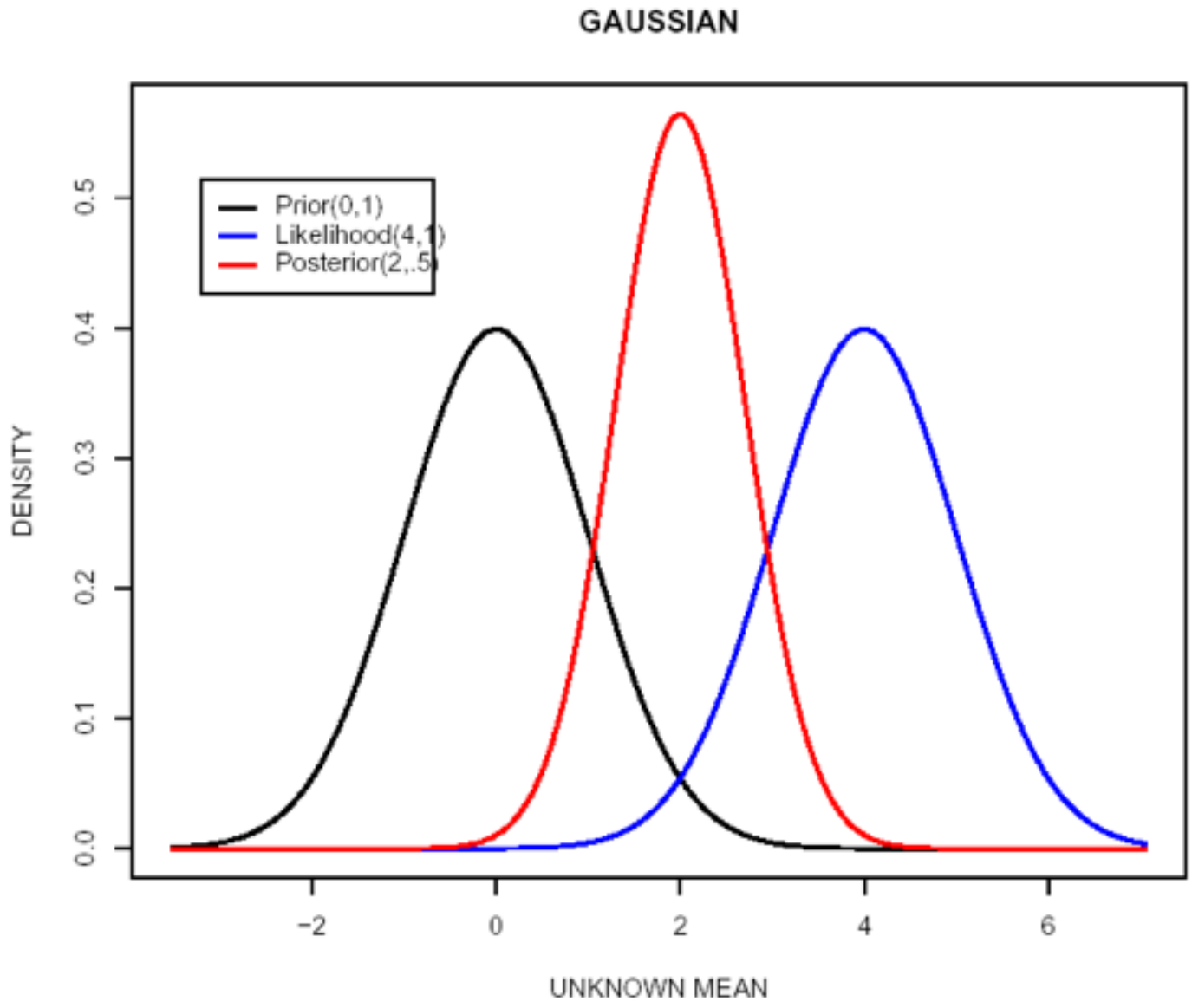
$$\begin{aligned} \text{Prior:} \quad \mathsf{G} &= N(\mu, \tau^2) \\ \text{Sampling distn.:} \quad \mathsf{f} &= N(\theta, \sigma^2) \\ \text{Marginal distn.:} \quad \mathsf{f}_G &= N(\mu, \sigma^2 + \tau^2) \end{aligned}$$

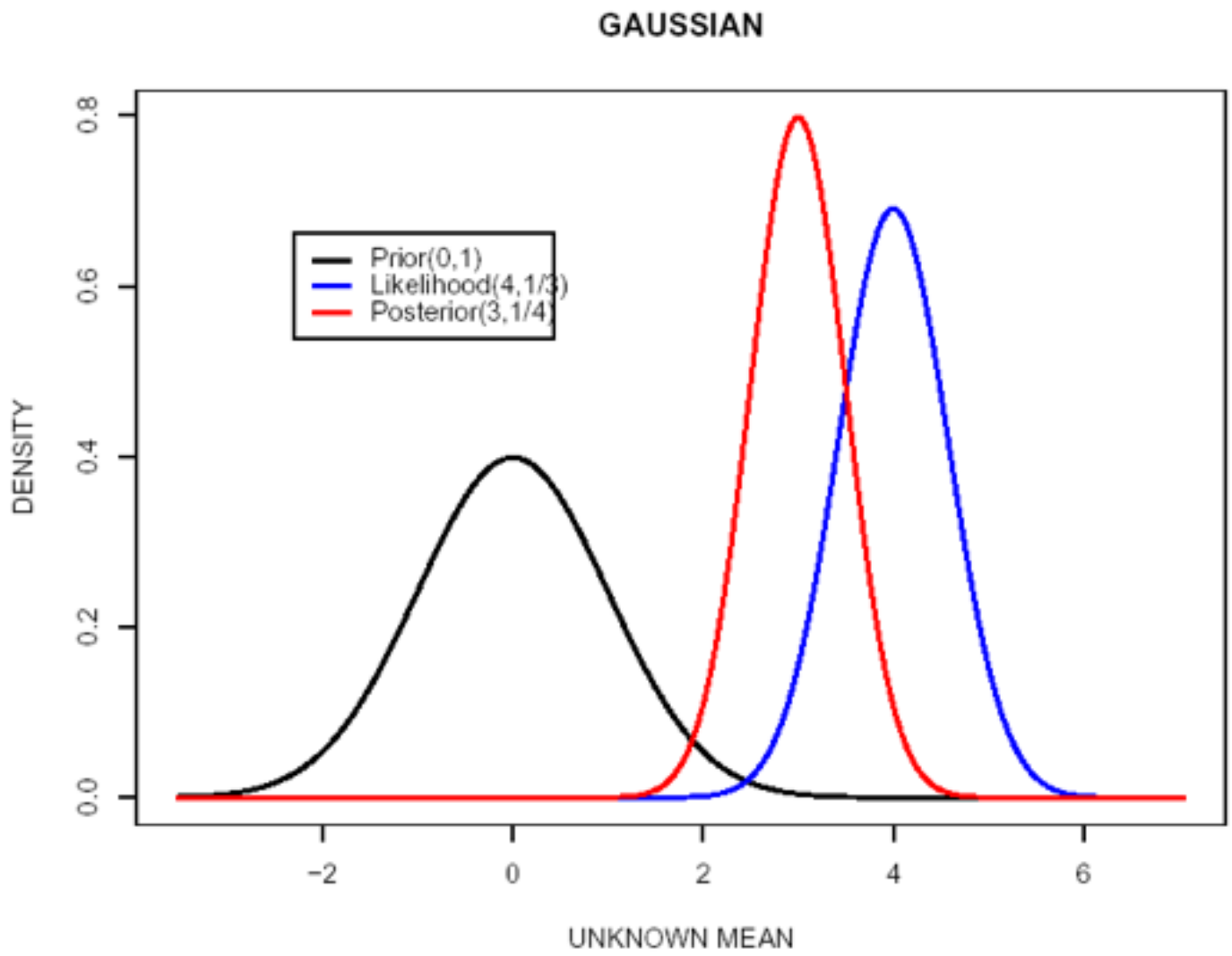
Overdispersion

- If (μ, τ^2, σ^2) are known, the posterior is Gaussian:

$$\begin{aligned} E(\theta|Y) &= B\mu + (1-B)Y \\ &= \mu + (1-B)(Y-\mu) \\ V(\theta|Y) &= (1-B)\sigma^2 \\ B &= \frac{\sigma^2}{\sigma^2 + \tau^2} \end{aligned}$$

- The Gaussian prior is conjugate
- Shrinkage and variance reduction
- Increasing σ^2 or decreasing τ^2 produces greater shrinkage


GAUSSIAN
Prior(0,1)
Likelihood(4,1)
Posterior(2,.5)
DENSITY
0.0
0.1
0.2
0.3
0.4
0.5
−2
0
2
4
6
UNKNOWN MEAN


GAUSSIAN
Prior(0,1)
Likelihood(4,1/3)
Posterior(3,1/4)
DENSITY
0.0
0.2
0.4
0.6
0.8
−2
0
2
4
6
UNKNOWN MEAN

Modeling Relative Expression

Courtesy of Gordon Smyth

Hierarchical Model

Normal Model

$$\hat{\beta}_{gj} \sim N(\beta_{gj}, c_{gj}\sigma_g^2)$$

$$s_g^2 \sim \sigma_g^2 \chi_{d_g}^2$$

Prior

$$P(\beta_{gj} \neq 0) = p$$

$$\beta_{gj} \mid \beta_{gj} \neq 0 \sim N(0, c_{0j}\sigma_g^2)$$

$$\sigma_g^2 \sim s_0^2 \left(\chi_{d_0}^2 / d_0\right)^{-1}$$

Reparametrization of Lönnstedt and Speed 2002

Normality, independence assumptions are wrong but convenient, resulting methods are useful

Posterior Statistics

Posterior variance estimators

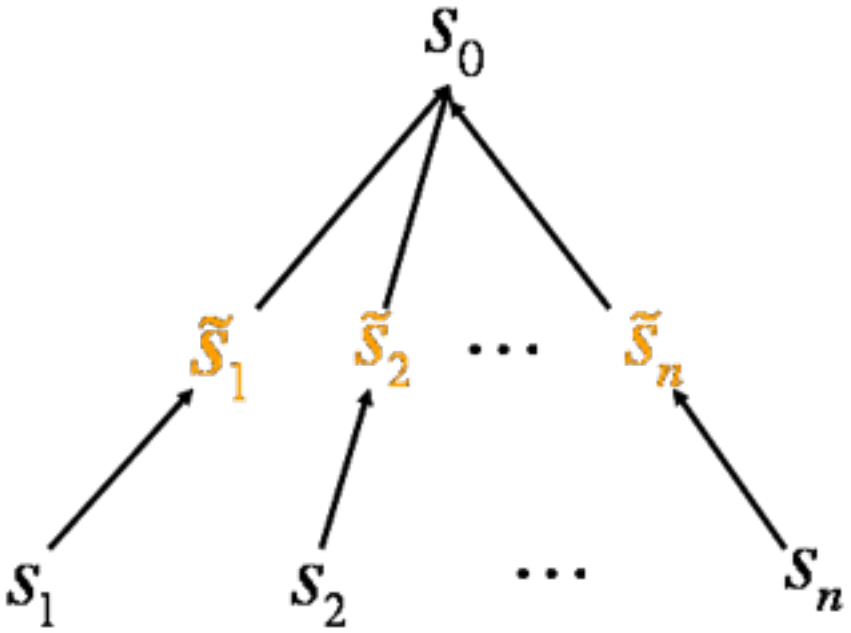
$$\tilde{s}_g^2 = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0}$$

Moderated t-statistics

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{c_{gj}}}$$

Eliminates large t-statistics merely from very small *s*

Shrinkage of Standard Deviations



The data decides whether $\tilde{t}_g$

should be closer to $t_{g,pooled}$ or to t_g