Mixture Modeling of the *p*-value Distribution

3/12/2009

Mixture Modeling of the *p*-value Distribution

- First proposed by Allison, D. B. , Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A., Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data, *Computational Statistics and Data Analysis*, **39**, 1-20.
- Model *p*-value distribution as a mixture of a Uniform(0,1) distribution (corresponding to true nulls) and a Beta(α,β) distribution (corresponding to false nulls).
- Pounds and Morris (2003) propose mixture of Uniform(0,1) and Beta(α,1). (BUM model)

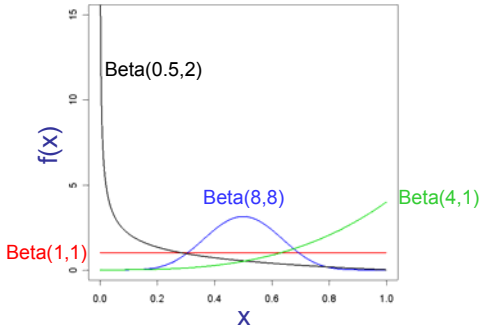
Beta Distributions

- A Beta(α,β) distribution is a probability distribution on the interval (0,1).
- The probability density function of a Beta(α,β) distribution is given by

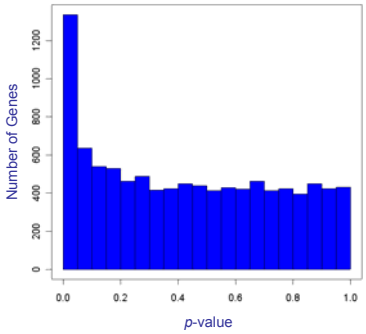
$$f(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \text{ for } 0<x<1.$$

- The mean of a Beta(α,β) distribution is $\frac{\alpha}{\alpha+\beta}$
- The variance of a Beta(α,β) distribution is $\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Various Beta Distributions



Model distribution of observed *p*-values as a mixture of uniform and beta



p-value density is assumed to be a mixture of a uniform density and a beta density.

$$g(p) = \pi_0 + \pi_1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}$$

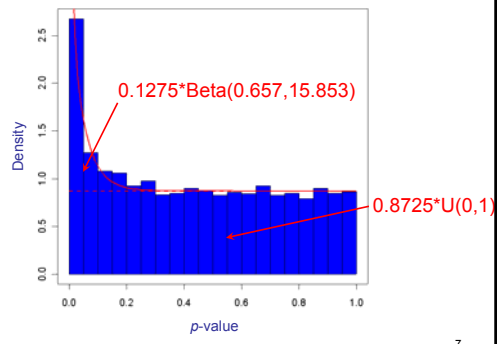
π_0 and π_1 are non-negative mixing proportions that sum to 1.

Matching up with our previous notation we have $\pi_0=m_0 / m$ and $\pi_1=m_1 / m$.

The parameters π_0, π_1, α, and β are estimated by the method of maximum likelihood assuming independence of all *p*-values.

Numerical maximization is necessary.

$\hat{\pi}_0 = 0.8725$ $\hat{\pi}_1 = 0.1275$ $\hat{\alpha} = 0.657$ $\hat{\beta} = 15.853$



7

Posterior Probability of Differential Expression

- Bayes Rule: $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$
- $P(H_{0i} \text{ is False} \mid p_i = p) = \frac{P(H_{0i} \text{ is False})P(p_i = p \mid H_{0i} \text{ is False})}{P(p_i = p)}$

$$= \frac{\pi_1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{\pi_0 + \pi_1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}$$

8

Posterior Probability of Differential Expression (continued)

- The posterior probability of differential expression is the probability that a gene is differentially expressed given its *p*-value.
- It can be estimated by replacing the unknown parameters π_0, π_1, α, and β in the previous expression by their maximum likelihood estimates.

9

	p-values	Estimated Posterior Probability of D.E.
1.	0.000001111	0.9862353
2.	0.000020858	0.9632383
3.	0.000025233	0.9618519
4.	0.000028355	0.9593173
5.	0.000032869	0.9572907
501.	0.009275782	0.7381684
502.	0.009286863	0.7380571
503.	0.009318375	0.7377411
504.	0.009332409	0.7376005
505.	0.009347553	0.7374489

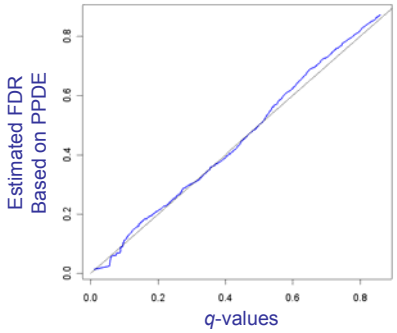
10

Relationship between Posterior Probability of Differential Expression (PPDE) and FDR

- 1 - average PPDE for a list of genes provides an estimate of the FDR for that list of genes.
- For example, the estimated FDR for the top 5 genes is 1-(0.986+0.963+0.961+0.959+0.957)/5=0.035.
- The properties of this approach to estimating FDR are unknown.

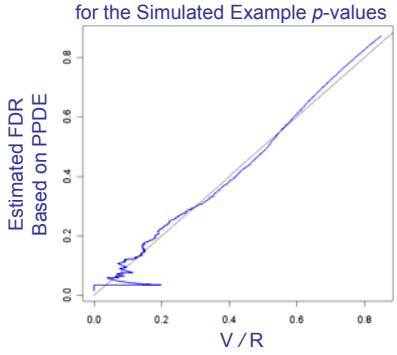
11

Plot of FDR Estimates Based on PPDE vs. *q*-value for the Simulated Example *p*-values



12

Plot of FDR Estimates Based on PPDE vs. Actual Ratio of False Positives to Number of Rejections for the Simulated Example *p*-values



13

FDR Estimates Based Directly on the Estimated Mixture Model

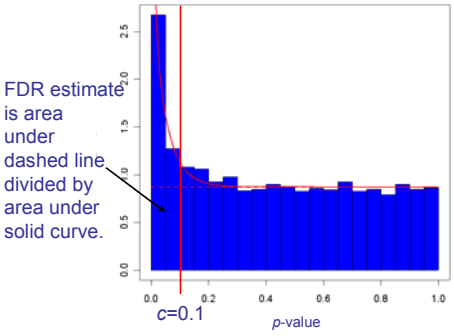
- $P(H_{0i} \text{ is True} \mid p_i \le c) = \dfrac{P(H_{0i} \text{ is True})P(p_i \le c \mid H_{0i} \text{ is True})}{P(p_i \le c)}$

$$= \frac{\pi_0 c}{\pi_0 c + \pi_1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^c p^{\alpha-1}(1-p)^{\beta-1}}$$

- Replacing the parameters in the expression above with their estimates gives an estimated "FDR" for any significance cutoff *c*.

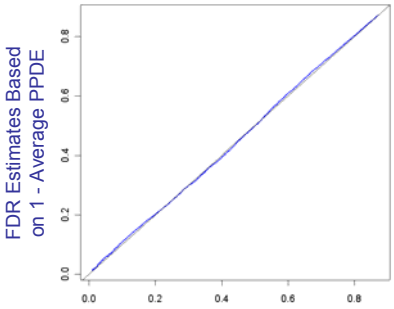
14

FDR Estimates Based Directly on the Estimated Mixture Model



15

Comparison of Mixture Model Methods for Estimating FDR



16

Comments

- The two methods will produce similar FDR estimates when there are a large number of closely spaced *p*-values.
- The method based on 1 – average PPDE may be useful for estimating the FDR in a list of genes that does not necessarily include the most significant genes.
- The method based directly on the estimated mixture model is probably preferable in the usual case where a list will consist of the most differentially expressed genes.

17