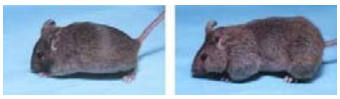
Mixture Modeling of the Distribution of *p*-values from *t*-tests

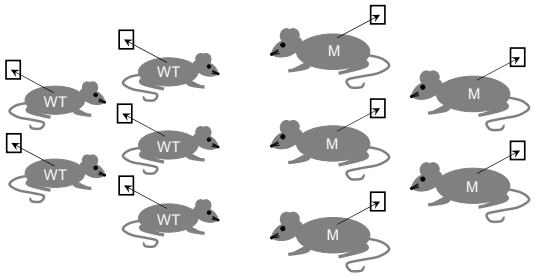
2/17/2011

Wild-type vs. Myostatin Knockout Mice



Belgian Blue cattle have a mutation in the myostatin gene.

Affymetrix GeneChips on 5 Mice per Genotype



A Typical Microarray Data Set

Gene ID	Treatment 1					Treatment 2					p-value
1	4835.8	4578.2	4856.3	4483.7	4275.3	4170.7	3836.9	3901.8	4218.4	4094.0	p_1
2	153.9	161.0	139.7	173.0	160.1	180.1	265.1	201.2	130.8	130.7	p_2
3	3546.5	3622.7	3364.3	3433.6	2757.2	3346.9	2723.8	2892.0	3021.3	2452.7	p_3
4	711.3	717.3	776.6	787.5	750.3	910.2	813.3	687.9	811.1	695.6	p_4
5	126.3	178.2	114.5	158.7	157.3	231.7	147.0	102.8	157.6	146.8	p_5
6	4161.8	4622.9	3795.7	4501.2	4265.8	3931.3	3327.6	3726.7	4003.0	3906.8	p_6
7	419.3	555.3	509.6	515.5	488.9	426.6	425.8	500.8	347.8	580.3	p_7
8	2420.7	2616.1	2768.7	2663.7	2264.6	2379.7	2196.2	2491.3	2710.0	2759.1	p_8
9	321.5	540.6	471.9	348.2	356.6	382.5	375.9	481.5	260.6	515.7	p_9
10	1061.4	949.4	1236.8	1034.7	976.8	1059.8	903.6	1060.3	960.1	1134.5	p_{10}
11	1293.3	1147.7	1173.8	1173.9	1274.2	1062.8	1172.1	1113.0	1432.1	1012.4	p_{11}
12	336.1	413.5	425.2	462.8	412.2	391.7	388.1	363.7	310.8	404.6	p_{12}
13	5718.1	4105.5	5620.9	6786.8	7823.0	1297.8	1303.8	1318.8	1189.2	1171.5	p_{13}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22690	249.6	283.6	271.0	246.9	252.7	214.2	217.9	266.6	193.7	413.2	p_{22690}

We want to test $H_{i0} : \mu_{i1} = \mu_{i2}$ for gene i=1,...,m

Test statistic for gene i: $t_i = \dfrac{\bar{X}_{i1} - \bar{X}_{i2}}{\sqrt{s_i^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

$|t_i| \sim |t|$ where

$$t \sim t(n_1 + n_2 - 2, \text{ncp} = \delta_i)$$

$$\delta_i = \frac{|\mu_{i1} - \mu_{i2}|}{\sqrt{\sigma_i^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Equivalently Expressed (EE) and Differentially Expressed (DE) Genes

- A certain proportion, say π_0, of the tested genes have expression distributions that are the same for both treatments. (EE genes)
- For other genes, the mean expression level differs between treatments. (DE genes)
- For DE genes, the degree of differential expression, summarized by the non-centrality parameter

$$\delta_i = \frac{|\mu_{i1} - \mu_{i2}|}{\sqrt{\sigma_i^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

varies from gene to gene.

Objectives

- Estimate π_0= proportion of non-centrality parameters that are zero (i.e., proportion of genes that are EE)
- Estimate $g(\delta)$= density that approximates the true distribution of nonzero non-centrality parameters.
- Estimate false discovery rates (FDR)
- Estimate falsely interesting discovery rates (FIDR)
- Perform power and sample size calculations for future experiments

7

Conditional Distribution Function of the *p*-value Given the Non-Centrality Parameter

$$F_{p|\delta}(p;\delta) = P(p\text{-value} \le p|\text{NCP} = \delta)$$

$$\begin{aligned} F_{p|\delta}(p;\delta) &= P(|t| \ge F_{|t|}^{-1}(1-p;0)|\text{NCP} = \delta) \\ &= 1 - P(|t| \le F_{|t|}^{-1}(1-p;0)|\text{NCP} = \delta) \\ &= 1 - F_{|t|}(F_{|t|}^{-1}(1-p;0);\delta) \end{aligned}$$

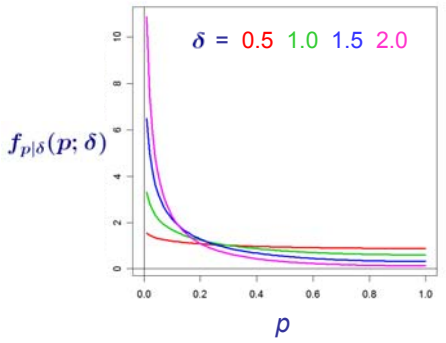
8

Conditional Density of the *p*-value Given the Non-Centrality Parameter

$$\begin{aligned} f_{p|\delta}(p;\delta) &= \frac{\partial}{\partial p} F_{p|\delta}(p;\delta) \\ &= \frac{\partial}{\partial p}\{1 - F_{|t|}(F_{|t|}^{-1}(1-p;0);\delta)\} \\ &= -f_{|t|}(F_{|t|}^{-1}(1-p;0);\delta)\frac{\partial}{\partial p}F_{|t|}^{-1}(1-p;0) \\ &= \frac{f_{|t|}(F_{|t|}^{-1}(1-p;0);\delta)}{f_{|t|}(F_{|t|}^{-1}(1-p;0);0)} \end{aligned}$$

9

Conditional Densities of *p*-values Given δ



10

The Marginal Distribution of the *t*-test *p*-value

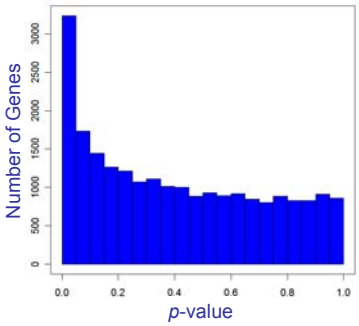
Suppose that each non-centrality parameter δ is 0 with probability π_0 and a draw from a continuous distribution $g(\delta)$ with probability $(1-\pi_0)$.

Then the marginal density of the *t*-test *p*-value is given by

$$f_p(p) = \pi_0 + (1-\pi_0)\int_0^\infty f_{p|\delta}(p;\delta)g(\delta)d\delta$$

11

Histogram of *p*-values from Two-Sample *t*-Tests



12

Approximate *g* with a Linear Spline Function

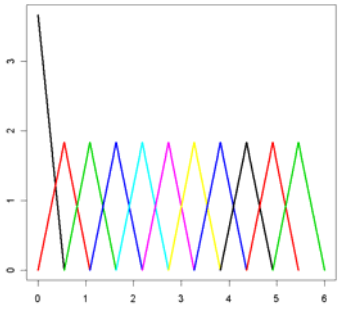
$$g(\delta) \approx g(\delta, \beta) = \sum_{k=1}^{K-1} \beta_k B_k(\delta)$$

where $B_1(\delta), \ldots, B_{K-1}(\delta)$ are B-splines normalized to be densities

and $\beta_1, \ldots, \beta_{K-1} \geq 0 \quad \sum_{k=1}^{K-1} \beta_k = 1$.

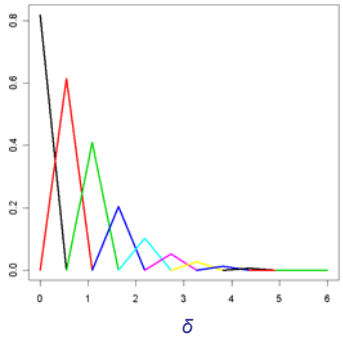
13

The B-Splines



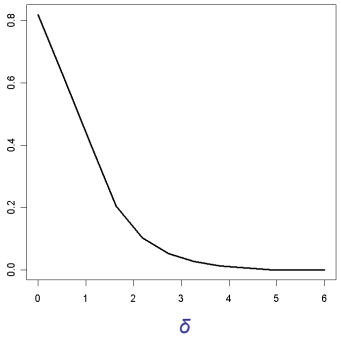
14

Weighted B-Splines



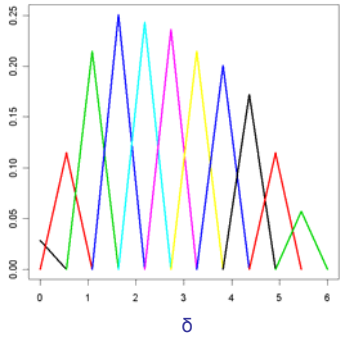
15

Linear Spline Estimate of *g*



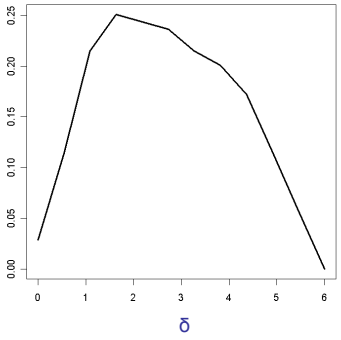
16

Weighted B-Splines



17

Linear Spline Estimate of g



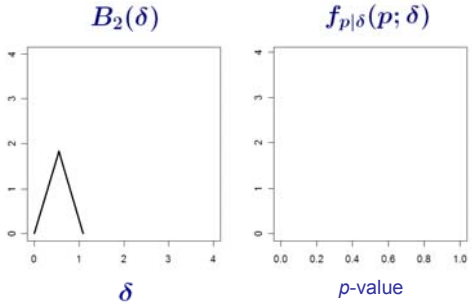
18

Approximating the Marginal Density of *p*-values

$$
\begin{aligned}
f_p(p) &= \pi_0 + (1-\pi_0)\int_0^\infty f_{p|\delta}(p;\delta)g(\delta)d\delta \\
&\approx \pi_0 + (1-\pi_0)\int_0^\infty f_{p|\delta}(p;\delta)g(\delta,\beta)d\delta \\
&= \pi_0 + (1-\pi_0)\int_0^\infty f_{p|\delta}(p;\delta)\sum_{k=1}^{K-1}\beta_k B_k(\delta)d\delta \\
&= \pi_0 \cdot 1 + \sum_{k=1}^{K-1}(1-\pi_0)\beta_k \int_{\mathcal{S}_k} f_{p|\delta}(p;\delta)B_k(\delta)d\delta \\
&\equiv \theta_1 z_1(p) + \sum_{k=1}^{K-1}\theta_{k+1}z_{k+1}(p) \\
&= \sum_{k=1}^{K}\theta_k z_k(p) \equiv f_p(p;\theta)
\end{aligned}
$$

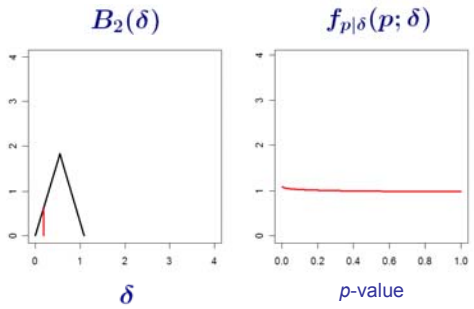
19

$$z_3(p) = \int_{\mathcal{S}_2} f_{p|\delta}(p;\delta)B_2(\delta)d\delta$$



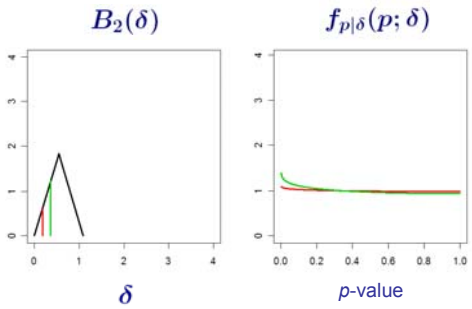
20

$$z_3(p) = \int_{\mathcal{S}_2} f_{p|\delta}(p;\delta)B_2(\delta)d\delta$$



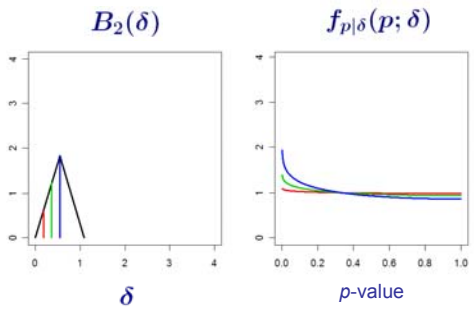
21

$$z_3(p) = \int_{\mathcal{S}_2} f_{p|\delta}(p;\delta)B_2(\delta)d\delta$$



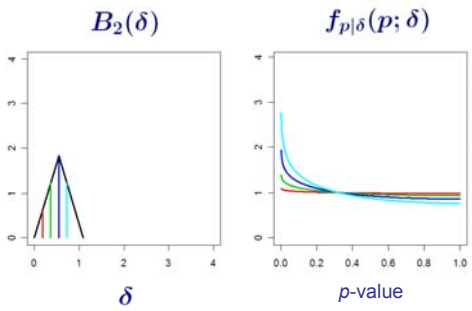
22

$$z_3(p) = \int_{\mathcal{S}_2} f_{p|\delta}(p;\delta)B_2(\delta)d\delta$$



23

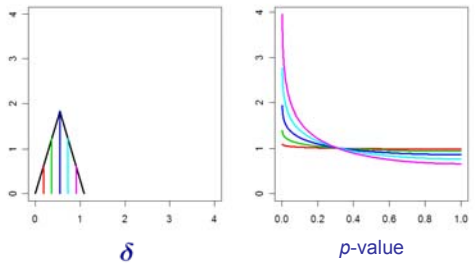
$$z_3(p) = \int_{\mathcal{S}_2} f_{p|\delta}(p;\delta)B_2(\delta)d\delta$$



24

$$z_3(p) = \int_{\mathcal{S}_2} f_{p|\delta}(p;\delta) B_2(\delta) d\delta$$

$B_2(\delta)$ $\qquad$ $f_{p|\delta}(p;\delta)$

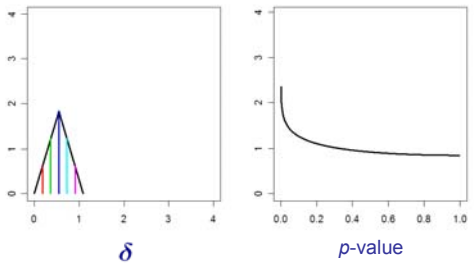


δ $\qquad$ p-value

25

$$z_3(p) = \int_{\mathcal{S}_2} f_{p|\delta}(p;\delta) B_2(\delta) d\delta$$

$B_2(\delta)$ $\qquad$ $z_3(p)$

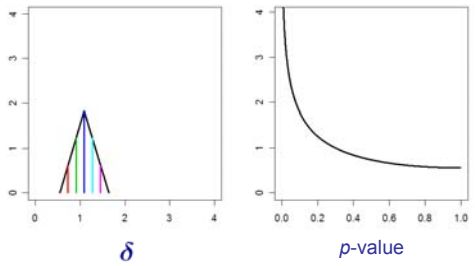


δ $\qquad$ p-value

26

$$z_4(p) = \int_{\mathcal{S}_3} f_{p|\delta}(p;\delta) B_3(\delta) d\delta$$

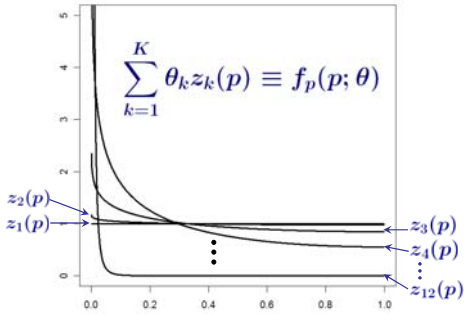
$B_3(\delta)$ $\qquad$ $z_4(p)$



δ $\qquad$ p-value

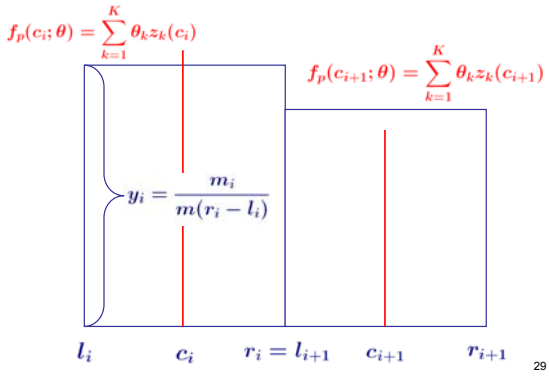
27

We seek a convex combination of the z functions that fits our empirical p-value distribution well.



28

Bin p-values into 2000 Bins



29

Find $\boldsymbol{\theta}$ that minimizes

Number of Histogram Bins (e.g., 2000)

Observed Density Height for ith Bin

Smoothing Parameter

$$SS(\theta,\lambda) = \sum_{i=1}^{N_{bin}} w_i\{y_i - f_p(c_i;\theta)\}^2 + \lambda Q(\theta)$$

Weight Assigned to ith Bin

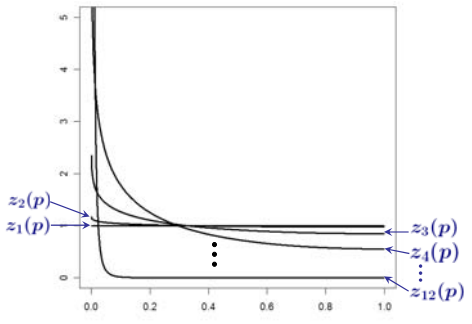
Semiparametric Approximation of Density Height for the ith Bin

Penalty for lack of smoothness in g

Solved via quadratic programming.

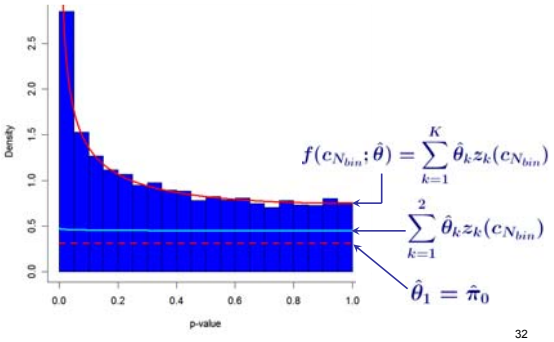
30

Solution to the penalized least squares problem yields a convex combination of z functions as a density estimate.



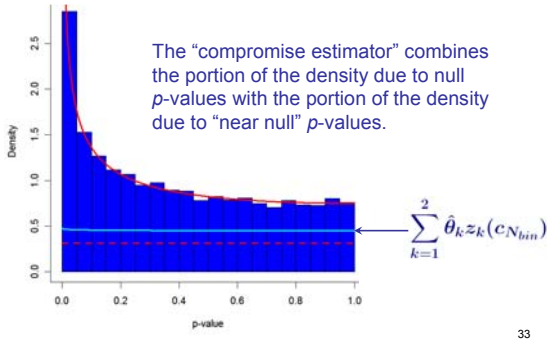
31

Two-Sample *t*-test *p*-values with Estimated Density



32

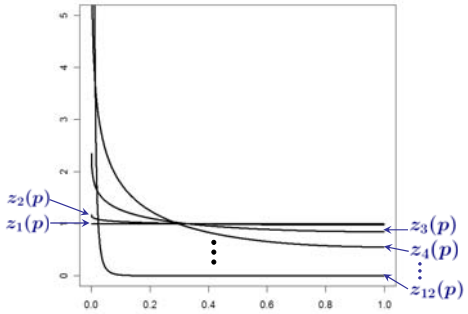
Two-Sample *t*-test *p*-values with Estimated Density



The "compromise estimator" combines the portion of the density due to null *p*-values with the portion of the density due to "near null" *p*-values.

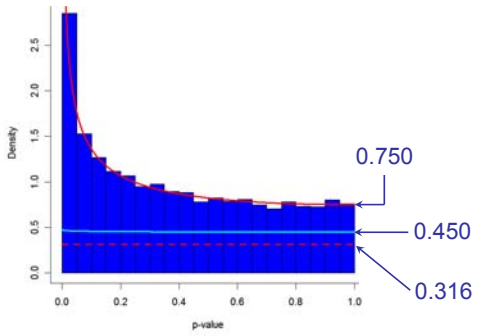
33

Note that the similarity of z_1 and z_2.



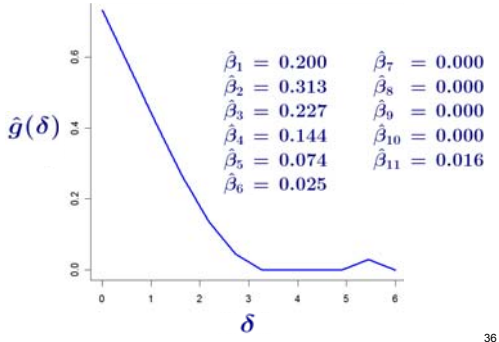
34

Two-Sample *t*-test *p*-values with Estimated Density



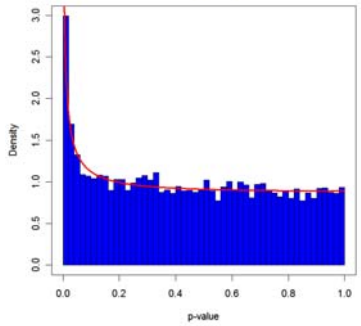
35

Estimated Density of Nonzero Non-Centrality Parameters



36

Histogram of 10000 *p*-values with Estimated Distribution



m = 10000

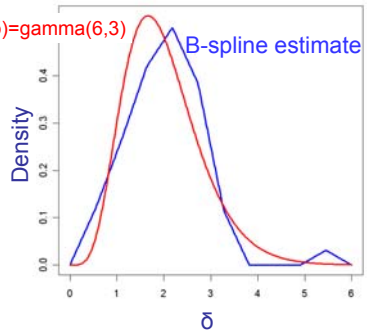
m_0 = 8500

π_0 = 0.85

g(δ)=gamma(6,3)

37

True and Estimated NCP Densities



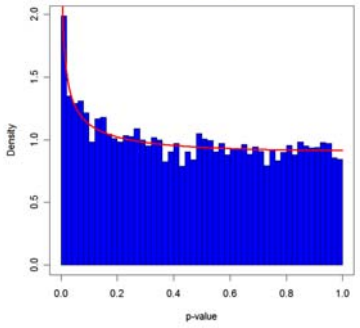
38

Parameter Estimates

$$\begin{aligned}
\hat{\pi}_0 &= 0.8583 & \hat{\beta}_6 &= 0.209 \\
\hat{\beta}_1 &= 0.000 & \hat{\beta}_7 &= 0.063 \\
\hat{\beta}_2 &= 0.065 & \hat{\beta}_8 &= 0.000 \\
\hat{\beta}_3 &= 0.143 & \hat{\beta}_9 &= 0.000 \\
\hat{\beta}_4 &= 0.229 & \hat{\beta}_{10} &= 0.000 \\
\hat{\beta}_5 &= 0.274 & \hat{\beta}_{11} &= 0.017
\end{aligned}$$

39

Histogram of 10000 *p*-values with Estimated Distribution



m = 10000

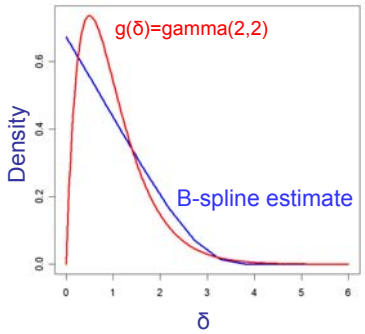
m_0 = 7500

π_0 = 0.75

g(δ)=gamma(2,2)

40

True and Estimated NCP Densities



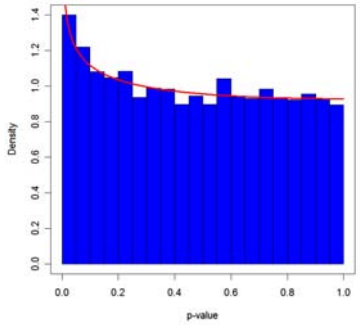
41

Parameter Estimates

$$\begin{aligned}
\hat{\pi}_0 &= 0.7848 & \hat{\beta}_6 &= 0.039 \\
\hat{\beta}_1 &= 0.184 & \hat{\beta}_7 &= 0.008 \\
\hat{\beta}_2 &= 0.297 & \hat{\beta}_8 &= 0.000 \\
\hat{\beta}_3 &= 0.226 & \hat{\beta}_9 &= 0.000 \\
\hat{\beta}_4 &= 0.156 & \hat{\beta}_{10} &= 0.000 \\
\hat{\beta}_5 &= 0.091 & \hat{\beta}_{11} &= 0.000
\end{aligned}$$

42

Histogram of 10000 *p*-values with Estimated Distribution



m = 10000

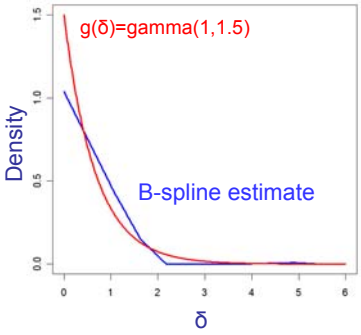
m_0 = 7000

π_0 = 0.70

g(δ)=gamma(1,1.5)

43

True and Estimated NCP Densities



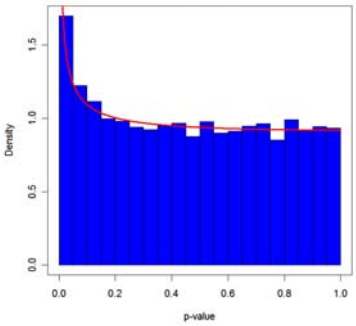
44

Parameter Estimates

$$\begin{aligned}
\hat{\pi}_0 &= 0.6975 & \hat{\beta}_6 &= 0.000 \\
\hat{\beta}_1 &= 0.284 & \hat{\beta}_7 &= 0.000 \\
\hat{\beta}_2 &= 0.398 & \hat{\beta}_8 &= 0.000 \\
\hat{\beta}_3 &= 0.229 & \hat{\beta}_9 &= 0.003 \\
\hat{\beta}_4 &= 0.081 & \hat{\beta}_{10} &= 0.004 \\
\hat{\beta}_5 &= 0.000 & \hat{\beta}_{11} &= 0.000
\end{aligned}$$

45

Histogram of 10000 *p*-values with Estimated Distribution



m = 10000

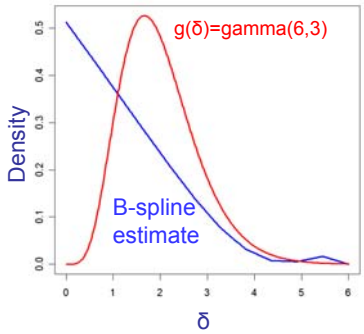
m_0 = 9000

π_0 = 0.90

g(δ)=gamma(6,3)

46

True and Estimated NCP Densities



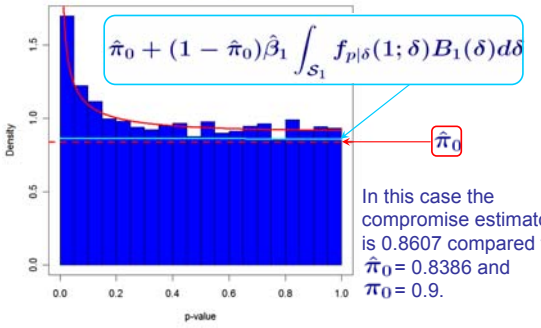
47

Parameter Estimates

$$\begin{aligned}
\hat{\pi}_0 &= 0.8386 & \hat{\beta}_6 &= 0.076 \\
\hat{\beta}_1 &= 0.140 & \hat{\beta}_7 &= 0.042 \\
\hat{\beta}_2 &= 0.239 & \hat{\beta}_8 &= 0.017 \\
\hat{\beta}_3 &= 0.197 & \hat{\beta}_9 &= 0.004 \\
\hat{\beta}_4 &= 0.156 & \hat{\beta}_{10} &= 0.003 \\
\hat{\beta}_5 &= 0.115 & \hat{\beta}_{11} &= 0.009
\end{aligned}$$

48

The Compromise Estimator



In this case the compromise estimator is 0.8607 compared to $\hat{\pi}_0$ = 0.8386 and π_0 = 0.9.

Other Estimators of π_0

There are many! Some examples include...

Lowest Slope: Schweder and Spjotvoll (1982), Hochberg and Benjamini (1990), Benjamini and Hochberg (2000).

Mixture of Uniform and Beta: Allison et al. (2002), Pounds and Morris (2003).

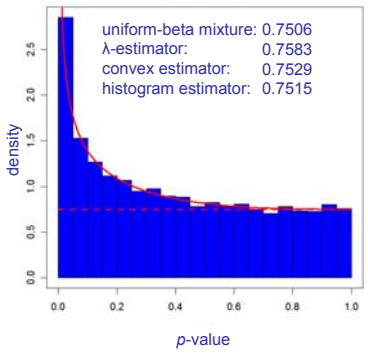
λ Threshold: Storey (2002), Storey and Tibshirani (2003)

Convex Density Estimation: Langaas, Ferkingstad, Lindqvist (2005)

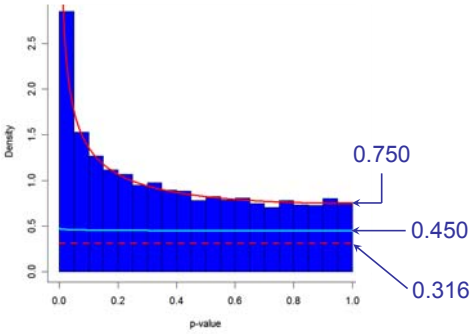
Histogram Based Estimation: Mosig et al. (2001), Nettleton et al. (2006)

Moment Based Estimator: Lai (2007)

Other Estimators of π_0 for the Myostatin Data



Myostatin *p*-values with Estimated Density



Some Simulation Results

π_0	0.95	0.95	0.95	0.70	0.70	0.70	
$g(\delta)$	near	medium	far	near	medium	far	
$\hat{\pi}_0$	0.0251	0.0348	0.0346	0.0503	0.1035	0.0732	RMSE
$f(c_{N_{bin}};\hat{\theta})$	0.0276	0.0145	0.0066	0.1519	0.0748	0.0187	RMSE
compromise	0.0206	0.0124	0.0269	0.0492	0.0268	0.0458	RMSE
convex	0.0238	0.0148	0.0121	0.1485	0.0767	0.0214	RMSE
$\hat{\pi}_0$	0.0001	-0.0228	-0.0286	-0.0063	-0.0703	-0.0233	BIAS
$f(c_{N_{bin}};\hat{\theta})$	0.0266	0.0123	-0.0014	0.1517	0.0743	0.0166	BIAS
compromise	0.0076	0.0007	-0.0221	0.0367	0.0158	-0.0169	BIAS
convex	0.0208	0.0089	-0.0020	0.1476	0.0749	0.0167	BIAS

"convex" is the estimator of Langaas et al. (2005) JRSSB **67,** 555–572

Other Quantities of Interest

Posterior Probability of Differential Expression
PPDE(p) = P(DE|p-value=p)

False Discovery Rate
FDR(c) = P(EE|p≤c)

True Positive Rate
TPR(c) = P(DE|p≤c)
= 1 – FDR(c)

True Negative Rate
TNR(c) = P(EE|p>c)

Expected Discovery Rate
EDR(c) = P(p≤c|DE)

Gadbury et al. (2004). *Stat. Meth. in Med. Res.* **13**, 325-338 discuss last three quantities in power and sample size context.

Approximation for FDR

$$\begin{aligned}\mathrm{FDR}(c) &= P(\mathrm{EE}|p \le c) = \frac{P(p \le c|\mathrm{EE})P(\mathrm{EE})}{P(p \le c)} \\ &= \frac{c\pi_0}{\int_0^c [\pi_0 + (1-\pi_0)\int_0^\infty f_{p|\delta}(p;\delta)g(\delta)d\delta]dp} \\ &\approx \frac{c\hat{\pi}_0}{c\hat{\pi}_0 + (1-\hat{\pi}_0)\int_0^\infty F_{p|\delta}(c;\delta)\hat{g}(\delta)d\delta} \\ &= \frac{c\hat{\pi}_0}{c\hat{\pi}_0 + (1-\hat{\pi}_0)\sum_{k=1}^{K-1}\hat{\beta}_k\int_{\mathcal{S}_k} F_{p|\delta}(c;\delta)B_k(\delta)d\delta}\end{aligned}$$

Approximation for EDR

$$\begin{aligned}\mathrm{EDR}(c) &= P(p \le c|\mathrm{DE}) \\ &= \int_0^c \int_0^\infty f_{p|\delta}(p;\delta)g(\delta)d\delta dp \\ &= \int_0^\infty F_{p|\delta}(c;\delta)g(\delta)d\delta \\ &\approx \int_0^\infty F_{p|\delta}(c;\delta)\hat{g}(\delta)d\delta \\ &= \sum_{k=1}^{K-1}\hat{\beta}_k\int_{\mathcal{S}_k} F_{p|\delta}(c;\delta)B_k(\delta)d\delta\end{aligned}$$

Power-Sample Size Calculations

- If we have estimates of π_0 and g(δ) from a previous experiment, we can examine how our ability to discover differentially expressed genes will vary with sample size.
- Suppose the within-treatment sample sizes for a new experiment differ from the previous experiment by a factor of η.
- If δ denotes the NCP for a gene in the previous experiment, then the NCP for the same gene in the new experiment will be $\sqrt{\eta}\delta$.
- We can see how quantities of interest very with η to guide samples size selection in the new experiment.

Relationship between New NCP and Old NCP

$$\begin{aligned}\delta_{\mathrm{new}} &= \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2\left(\frac{1}{\eta n_1} + \frac{1}{\eta n_2}\right)}} \\ &= \sqrt{\eta}\frac{\mu_1 - \mu_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \sqrt{\eta}\delta\end{aligned}$$

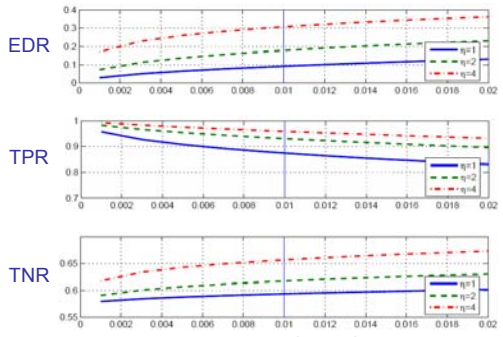
Approximate EDR for the New Experiment

$$\mathrm{EDR}^*(c) \approx \sum_{k=1}^{K-1}\hat{\beta}_k\int_{\mathcal{S}_k} F_{p|\delta}(c;\sqrt{\eta}\delta)B_k(\delta)d\delta$$

degrees of freedom change to reflect the larger sample size

replaces δ in the calculation for the previous experiment

Power-Sample Size Calculations



p-value threshold for significance

"Interesting Discovery" Rates

$$\mathrm{FIDR}(c) = P(\delta < \boxed{\delta^*} | p \leq c)$$

researcher-determined threshold
that defines "interesting discovery"

$$\mathrm{EIDR}(c) = P(p \leq c | \delta \geq \boxed{\delta^*})$$

61

Main Reference

Ruppert, D., Nettleton, D., Hwang, J.T.G. (2007). Exploring the information in *p*-values for the analysis and planning of multiple-test experiments. *Biometrics.* **63** 483-495.

62