FDR and Bayesian Multiple Comparisons Rules

Y. Xingchen G. Diana

Department of Statistical Sciences
University of California, Santa Cruz

AMS 221 Presentation, Winter 2019

Outline

Example

Gene#	Tumor1	Tumor2	$\cdots$	Tumor18	Normal1	Normal2	$\cdots$	Normal18
1	0.39	6.19	$\cdots$	0.03	6.53	-0.22	$\cdots$	-1.53
2	-10.82	0.117	$\cdots$	-1.11	24.72	1.22	$\cdots$	14.06
3	-0.78	-6.47	$\cdots$	0.03	15.69	-1.39	$\cdots$	-1.53
4	26	11.40	$\cdots$	27.14	57.25	44.71	$\cdots$	60.50
5	4.17	1.33	$\cdots$	-1.11	-3.78	-0.22	$\cdots$	-8.86
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
7457	-7.17	-1.06	$\cdots$	-2.28	0.14	-4.25	$\cdots$	-3.97

- $H_0 : r_i = 0$ versus $H_1 : r_i = 1$
- For each gene we are interested in the comparison of the two competeing hypotheses that gene i is differentially expressed versus not differentially expressed
- Many recently proposed approaches to address these type of problems are based on the controlling the FDR

Notation

- r_i denotes the unknown truth in the i-th comparison
- δ_i is an indicator for the decision to report i as differentially expressed
  - $\delta_i = 1$: decision to report a gene as differentially expressed as a discovery (rejection)
  - $\delta_i = 0$: not differentially expressed as a negative (fail to reject)

Notation

- $D = \sum \delta_i$
- false discovery, $FD = \sum (1 - r_i)\, \delta_i$. false discovery rate, $FDR = \sum (1 - r_i)\, \delta_i / D$
- false negative, $FN = \sum r_i\, (1 - \delta_i)$. false negative rate, $FNR = \sum r_i\, (1 - \delta_i)\, /\, (n - D)$

- How we proceed to control the FDR depends on the chosen paradigm
- Frequentist considers taking expectation of the FDR over repeated sampling
- We focus on the Bayesian method of controlling the FDR
  - The only unknown quantity in $FDR = \sum \delta_i \left(1 - r_i\right) D$ is the r_i
  - Let $\nu_i = P\left(r_i = 1 | Y\right)$ denote the marginal posterior probability of gene i being differentially expressed and define

$$\overline{FDR} = E\left(FDR | Y\right) = \frac{\sum \left(1 - \nu_i\right) \delta_i}{D}$$

Posterior Probabilites Adjust for Multiplicities

An example of the need for multiplicity control:

In a recent talk about the drug discovery process, the following numbers were given in illustration.

- 10,000 relevant compounds were screened for biological activity.
- 500 passed the initial screen and were studied in vitro.
- 25 passed this screening and were studied in Phase I animal trials.
- 1 passed this screening and was studied in a Phase II human trial.

This could be nothing but noise, if screening was done based on 'significance at the 0.05 level.'

If no compound had any effect,

- about $10,000 \times 0.05 = 500$ would initially be significant at the 0.05 level;
- about $500 \times 0.05 = 25$ of those would next be significant at the 0.05 level;
- about $25 \times 0.05 = 1.25$ of those would next be significant at the 0.05 level
- the 1 that went to Phase II would fail with probability 0.95.

Posterior Probabilites Adjust for Multiplicities

- In Bayesian paradigm we use posterior probabilitites to account for multiplicities.
- Posterior inference adjusts for multiplicities, and no further adjustment is required
  1. The probability model needs to include a positive prior probability of non-differential expression for each gene i
  2. The model needs to include hyperparameter that defines the prior probability mass for non-differential expression
- The marginal posterior probability of differential expression adjusts for the muliplicites

Posterior Probabilites Adjust for Multiplicities

	Observed z scores								
p_0	-5.0	-4.0	-3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
0.4	**1.0**	**1.0**	**1.0**	**0.9**	**0.5**	0.2	**0.4**	**0.9**	**1.0**
0.8	**0.9**	**0.9**	**0.8**	**0.4**	0.1	0.1	0.1	**0.4**	**0.8**
0.9	**0.5**	0.4	0.3	0.1	0.1	0.0	0.0	0.1	0.3

Table 1: Posterior probabilities of differential expression, as a function of the observed difference score z_i, under three different simulation truths, using $p_0 = 0.4$ (first row), 0.8 (second row) and 0.9 (third row) for the proportion of false comparisons. Probabilities $v_i > 0.4$ are marked in bold face.

Loss Function: L_N

- Use $\overline{FD}$ and $\overline{FN}$ for the posterior expectations

$$\overline{FD} = \sum \delta_i (1 - \nu_i)$$
$$\overline{FN} = \sum \nu_i (1 - \delta_i)$$

- Loss Function:

$$L_N = c\overline{FD} + \overline{FN}$$

- $L_R(\delta, z) = c\overline{FDR} + \overline{FNR}$
- The loss function L_N is a natural extension of $(0, 1, c)$ loss function for traditional hypothesis testing problems
- The loss for a false discovery and false negative depends on the total number of discoveries or negatives, respectively

Loss Function: L_N

- Genovese and Wasserman (2002) interpret c as the langrange multiplier in the problem of minimizing FNR subject to a bound on FDR
- Dicovering as many as possible of the components that have $\delta_i = 1$, while at the same time controlling the number of false discoveries

Optimal Rule for L_N

- The optimal decision is to declare all genes with marginal probability beyond a threshold as differentially expressed:

$$\delta_i^* = I\left(\nu_i > t\right)$$

where the optimal choice (cutoff) for t is $c/\left(c+1\right)$

Issue

Assumptions underlying the loss function

- all false negatives are equally undesirable, and
- all false positives are equally undesirable

This is inappropriate in most applications

- A false negative for a gene that is truly differentially expressed, but with a small difference across the two biologic conditions, is surely less of a concern than a false negative for a gene that is differentially expressed with a large difference. The large difference might make it more likely that follow up experiments will lead to significant results

Loss Function: L_m

Assume now the probability model includes for each gene i a parameter m_i that can be interpreted as the level of differential expression, with $m_i = 0$ if $r_i = 0$, and $m_i > 0$ if $r_i = 1$

$$L_m(m, \delta, z) = -\sum \delta_i m_i + k \sum (1 - \delta_i) m_i + cD$$

- The loss function includes a reward proportional to m_i for a correct discovery, and a penalty proportional to m_i for each false negative
- The last term cD encourages parsimony, without which the optimal decision would be to trivially flag all genes
  - Similar to the penalty term in AIC and BIC that encouage parsimony

Optimal Rule for L_m

Let $\overline{m_i} = E\left(m_i | Y\right)$ denote the posterior expected level of differential expression for gene i. The optimal rule is

$$\delta_i^* = I\left\{\overline{m_i} \geq \frac{c}{1+k}\right\}$$

Flag all genes with $\overline{m_i}$ greater than the fixed cutoff.

Generalized Loss Function: L_f

- Replace m_i by some function of m . Thus the loss will be a non-linear function

$$L_f(m, \delta, z) = -\sum \delta_i f_D(m_i) + k \sum (1 - \delta_i) f_N(m_i) + cD$$

where $f_D(m)$ and $f_N(m)$ would naturally be S-shaped, monotone functions with a minimum at $m = 0$, and perhaps level off for large levels of m

Optimal Rule for Generalized L_f

The optimal decision is

$$\delta_f^* = I\left\{\overline{f_{D_i}} + k\overline{f_{N_i}} > c\right\}$$

where $\overline{f_{N_i}} = E\left(f_N\left(m\right)|Y\right)$, the posterior expectation for $f_{N_i}\left(m_i\right)$, and similarly for f_D.

- Flag all genes with sufficiently large expected reward for discovery and/or penalty for a false negative. The rule δ_f^* follows from the fact that the choice of m_i in L_m was arbitrary

FDR and Dependence: L_D

- Dependence structure of expression across genes might be interesting to study
- If the goal is to develop a panel of biomarkers (features) to classify future samples, then we want to have low correlation of the expression levels for the differentially expressed genes
- The dependent subsets are genes with a common functionality or gens corresponding to the nodes on a pathway of interest

FDR and Dependence: L_D

- Dependence is introduced not on the observed gene expressions, but on imputed trinary indicators {-1, 0, 1}

$$e_{it} = \begin{cases} -1 & \text{if } z_{it} < -1 \\ 0 & \text{if } -1 < z_{it} < 1 \\ 1 & \text{if } z_{it} > 1 \end{cases}$$

where z_{it} is latent normally distributed random variables that introduce the desired dependence on related genes as well as regression on biologic condition

FDR and Dependence: L_D

- Let x_t denote a sample specific covariate vector including an indicator x_{t1} for the biologic condition of the sample t and other covariates
  - For example, in the case of two group comparison between tumor and normal tissue, x_{t1} could be a binary indicator of tumor
- Let $\{e_{jt}; j \in N_i\}$ denote the trinary indicator for other genes that we may include as possible parent nodes in the dependent prior model

$$z_{it} = g\left(x_t, e_{jt}, j \in N_i\right) + \epsilon_i$$

With mean function $g\left(\cdot\right)$ and standard normal residuals ϵ_i

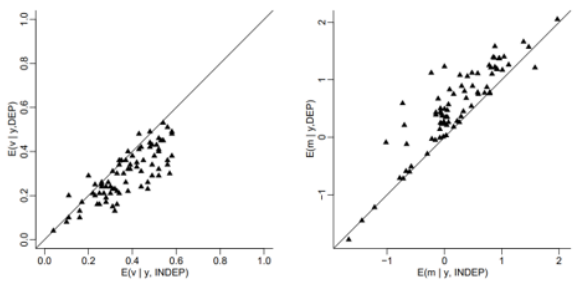
FDR and Dependence: L_D

- The regression on e_{jt} introduces the desired dependence, and the regression on x_t includes the regression on the biologic condition x_{t1}
- Let m_i denote the regression coefficient for x_{t1}
- Define Σ_1 as the correlation matrix of $\{z_{it}; \delta_i = 1\}$, the latent scores corresponding to the reported genes
- Under these construction, the loss function is modified as follows to penalize highly correlated genes while encouraging the inclusion of few highly differentially expressed genes with low correlation

$$L_D(m, \delta, z) = -k_1 \log(|\Sigma_1|) - k_2 \sum \delta_i f_D(m_i) + k_3 \sum (1 - \delta_i) f_N(m_i) + k_4 \, c \, D.$$

FDR and Dependence: L_D

- Muller et al (2006) used the aforementioned loss functions and models to study the EOC data where the goal is to compare 10 benign samples vs. 14 malignant samples
- The inference summaries ν_i and $\overline{m_i}$ change slightly when adjusting for dependence, but the impact in the final decision is visible



(a) v_i (b) $\overline{m}_i$

Figure 1: *Inference with dependent prior (y-axis) vs. indep prior (x-axis). The changes are large enough to change the decisions.*

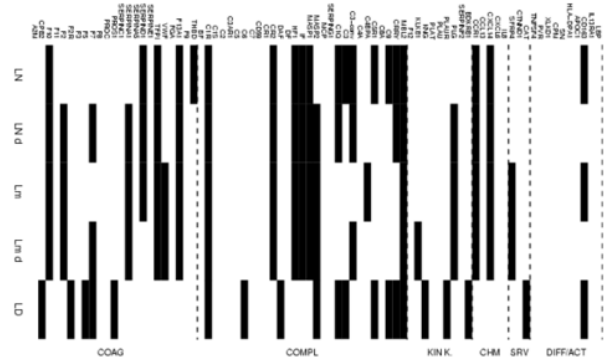


Figure 2: *Genes with $\delta_i = 1$, using L_N (top), L_N and dep model, L_m, L_m and dep model, and L_D (bottom).*

FDR and Predictive: L_F

- Microarray experiments are often carried out as hypothesis generating experiments, to screen for promising genes to be followed up in later experiments
  - For example: They consider inference for nine samples from patients with CLL by first using the microarray experiments for the initial screening and then using the real-time RT-PCR to validate the identified genes
- A loss function that based on the success of a future follow-up study should be used
- Let z_{it} be a suitably normalized measurement for gene i in sample t with $var\left(z_{it}\right) \approx 1$, similar to previous latent probit score variable
- Let y_{it}, denote the recorded outcome of the RT-PCR for gene i in sample t, which are copy numbers , interpreted as base two logarithm of the relative abundance of RNA in the sample

FDR and Predictive: L_F

- Abruzzo et al. (2005) conclude there is a bimodal distribution of correlations between the microarray and RT-PCR measurements, about half the genes show cross platform correlation $\rho_i \approx 0.8$ and half show $\rho_i \approx 0$
- Also the Standard deviation in the RT-PCR outcomes for each gene across samples is in around 0.5 to 1.5
- The cross-platform dependence can be captured by building on top the dependent prior model described earlier

$$p(y_{it} \mid z_{it}, \rho_i) = \begin{cases} z_{it} & \text{with prob. } \rho_i \\ N(0,1) & \text{with prob. } (1-\rho_i) \end{cases}$$

with $Pr(\rho_i = 0.8) = 0.5$ and $Pr(\rho_i = 0) = 0.5$

FDR and Predictive: L_F

- Muller et al (2006) build a loss function to construct a rule to identify genes that are most likely to achieve a significant result in the follow up experiment

1. For each identified gene i, select an alternative hypothesis
2. Find the sample size needed to achieve a desired power of test
3. Find the posterior predictive probability of statistically significant outcome for the future experiment
4. Define a loss function terms related to the posterior predictive probability and the sampling cost for the future experiment

- Author comments that this procedure is not a perfect reflection of the actual experimental process but gets the job done

FDR and Predictive: L_F

1. For each identified gene i, select an alternative hypothesis
   - Let $x_t \in \{-0.5, 0.5\}$ denote a centered indicator for the biological condition
     - Let $(\overline{m_i}, s_i)$ denote the posterior mean and standard deviation of m_i
     - Let $\bar{\rho}$ denote the assumed average cross-platform correlation
     - Let $\mu_{i1} = E(y_{it}|x_t > 0)$ and $\mu_{i0} = E(y_{it}|x_t < 0)$ denote the mean expression under the two biological condition in the follow up experiment The null and alternative hypothesis are as follows

$$H_0 : \mu_{i1} - \mu_{i0} = 0,\ H_1 : \mu_{i1} - \mu_{i0} = m^*_{yi}$$

where $m^*_{yi} = \bar{\rho}(\overline{m_i} - s_i)$

FDR and Predictive: L_F

2. Find the sample size needed to achieve a desired power of test
   - Let q_α denote $1-\alpha$ quantile of the standard normal distribution
   - Assuming a normal z-test, the required the sample size is as follows
   $$n_i(z) \geq 2\left[\left(q_\alpha + q_\beta\right)/m_{yi}^*\right]^2$$
   for a given significance level α and power $1-\beta$
   - the sample size is a function of data z and the choice of the alternative m_{yi}^*

FDR and Predictive: L_F

3. Find the posterior predictive probability of statistically significant outcome for the future experiment

- Let $\overline{y_{i0}}$ and $\overline{y_{i1}}$ be denote the sample average of expression level in the follow up experiment
- Let $R_i = \left\{ (\overline{y_{i1}} - \overline{y_{i0}}) \sqrt{n/2} \geq q_\alpha \right\}$ denote the statistically significant different region in the follow up experiment
- Let $\pi_i = Pr\left(R_i | Y\right)$ denote the posterior predictive probability of R_i

$$\pi_i(z) = (1 - p_\rho)\alpha + p_\rho \Phi \left[\frac{\rho^\star \overline{m}_{i1} \sqrt{n_i/2} - q_\alpha}{\sqrt{1 + \frac{n}{2}\rho^{*2} s_i^2}} \right]$$

FDR and Predictive: L_F

4. Define a loss function terms related to the posterior predictive probability and the sampling cost for the future experiment

- The loss function is defined by combining the tradeoff of small sampling cost and high success probability

$$L_F(\delta, z) = \sum_{\delta_i=1} \left[-c_1 \pi_i(z) + n_i(z)\right] + c_2 D$$

- The optimal rule is $\delta_i^* = I(n_i + c_2 \leq c_1 \pi_i)$

FDR and Predictive: L_F

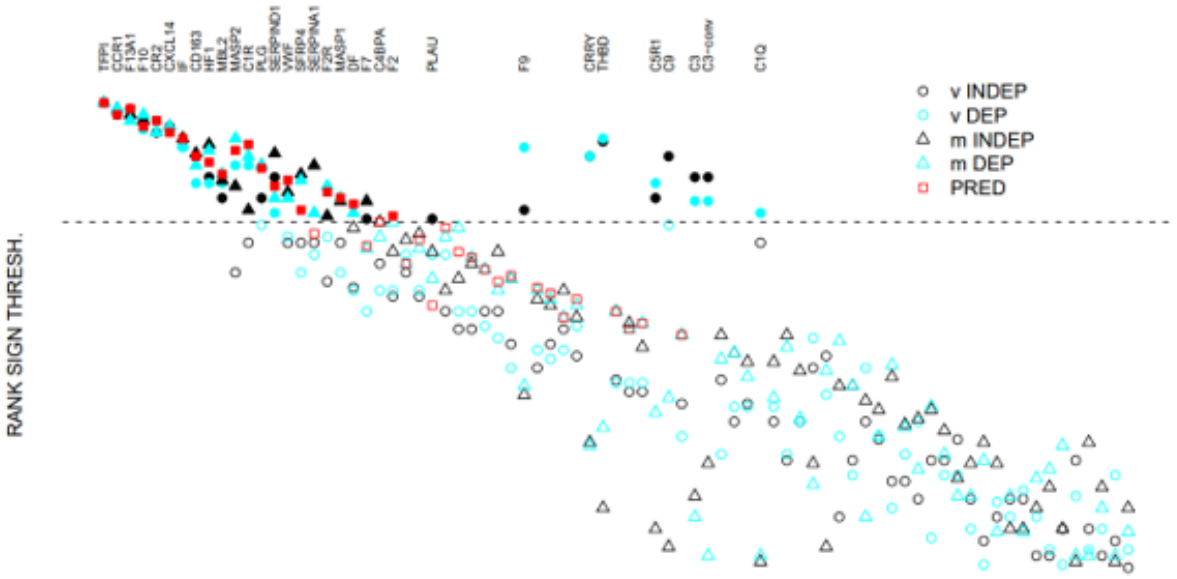


Figure 3: *Comparison of optimal rules for L_N (circles) under the independent model (black) and dependent model (light gray), L_m (triangle) under the independent (black) and dependent model (gray), and L_F (square). For each gene (horizontal axis) symbols are plotted against the rank of the corresponding significance threshold statistic (vertical axis). The reported genes under each rule are the top 20 ranked genes (above the dashed horizontal line). The symbols corresponding to selected genes are filled. The names of selected genes are shown on the vertical axis. Genes are sorted by average rank under the five criteria.*

Loss function	Rule
$L_N = c\overline{\text{FD}} + \overline{\text{FN}}$	$\delta_i = I(v_i > t)$
$L_m = -\sum \delta_i m_i + k \sum (1-\delta_i) m_i + cD$	$\delta_i = I(\overline{m}_i > t)$
$L_D = -k_1 \log(\lvert\Sigma_1\rvert) - k_2 \sum \delta_i f_D(m_i) + +k_3 \sum (1-\delta_i) f_N(m_i) + k_4 D$	no closed form
$L_F = \sum_{\delta_i=1} (-c_1 \pi_i + n_i) + c_2 D$	$\delta_i = I(n_i + c_2 \leq c_1 \pi_i)$

Parmigiani, G. and Inoue, Lr.
Decision Theory: Principles and Approaches.
Wiley, 2009.

Muller P. , Parmigiani G. and Rice K.
FDR and Bayesian Multiple Comparison Rules.
John Hopkins Univ., 2006.