False discovery rate and model selection

Elisabeth Gnatowski

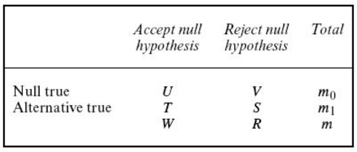
23.06.2006

Problem

- find differentially expressed genes using DNA microarrays
- number of genes much larger than number of independent samples in study $(p >> n)$
- problem of testing multiple hypotheses simultaneously
- analysing microarray data requires control of type 1 errors including balance between finding too many false-positive results and too little significant results ⇒ FDR

1. Definition of the FDR
   - Multiple Testing
   - FDR and pFDR
   - Controlling the FDR
2. Estimation of the FDR
   - Gene - specific FDR
3. Variable Selection
4. A decision theoretic framework
5. Simulation studies
   - $p < n$
   - $p > n$

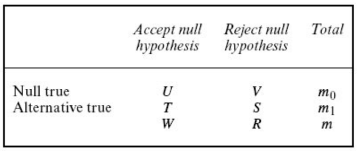
Multiple Testing



	Accept null hypothesis	*Reject null hypothesis*	*Total*
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

- Testing m Hypothesis, for m_0 of them, the null is true
- H_0 : gene is not differentially expressed
- V is equivalent to type 1 error, false-positive results
- T is equivalent to type 2 error, false-negative results
- W number of not rejected hypothesis,
  R number of rejected hypothesis

FDR and pFDR (positive false discovery rate)



	Accept null hypothesis	*Reject null hypothesis*	*Total*
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

- expected rate of false-positive results of all positive results

$$FDR = \begin{cases} E\left(\frac{V}{R}\right) & \text{falls } R > 0 \\ 0 & \text{falls } R = 0 \end{cases} = E\left[\frac{V}{R}|R > 0\right] P(R > 0)$$

- if $P(R = 0) > 0 \rightarrow$ Definition of FDR is useless $\rightarrow$ pFDR

$$pFDR = E\left(\frac{V}{R}|R > 0\right)$$

rate at which discoveries are false

Controlling the FDR

Benjamini and Hochberg (1995) propose a algorithm for selecting the hypotheses that are significant that controls the FDR:

- let $H_1, \ldots, H_G$ denote the null hypotheses to be tested, and $p_1 \leq p_2 \leq \ldots \leq p_G$ denote the corresponding, ordered, independent p-values
- let α denote the rate at which it is desired to control the FDR
- for selecting significant hypotheses first define level α and find $\hat{k} = max\left\{1 \leq k \leq G : p_k \leq \frac{\alpha k}{G}\right\}$
- reject all null hypotheses with indizes $1, \ldots, k$

strong control of the FDR at level α when the p-values are independent and uniformly distributed

1. Definition of the FDR
   - Multiple Testing
   - FDR and pFDR
   - Controlling the FDR
2. Estimation of the FDR
   - Gene - specific FDR
3. Variable Selection
4. A decision theoretic framework
5. Simulation studies
   - $p < n$
   - $p > n$

Basics

- Estimating the FDR by estimating π_0 (which is the rate of the true null hypothesis) and the joint distribution of the p - values
- the p - values of the true null hypothesis are uniformly distributed on the interval $[0, 1]$
- Theorem from Bayes:

$$\pi\left(\theta|x\right) = \frac{f\left(x|\theta\right) g\left(\theta\right)}{\int f\left(x|\theta\right) g\left(\theta\right) d\theta}$$

$\pi\left(\theta|x\right)$ posteriori distribution
$g\left(\theta\right)$ priori distribution
$f\left(x|\theta\right)$ joint distribution

sampling from posteriori distribution by MCMC

Assumptions

suppose

- we have independent test statistics $T = (T_1, \dots, T_m)$ for testing m hypotheses
- we have corresponding indicator variables $H_1, \dots, H_m$ where
  $$H_i = \begin{cases} 0 & \text{if the null hypotheses is true} \\ 1 & \text{if the alternative hypotheses is true} \end{cases}$$
- $H_1, \dots, H_m$ are a random sample from a Bernoulli distribution where $P(H_i = 0) = \pi_0;\ i = 1, \dots, m$
- $T_i | H_i = 0 \sim f_0$ and $T_i | H_i = 1 \sim f_1$ for densities f_0 and f_1
- we have the same rejection region R for each of the m hypotheses

Estimation of pFDR

by a Theorem from Storey (2002):

$$pFDR = P(H = 0 | T \in R)$$

$$= \frac{\pi_0 P(T \in R | H = 0)}{P(T \in R)}$$

Treating $H_1, \dots, H_m$ as parameters, we see that the definition of pFDR are posterior probabilities.

π_0 is the priori probability for a hypothesis to be a null hypothesis

Estimation a Gene - specific FDR

Application to a general linear model

- model
  $$E[Y_i] = \beta_{0j} + \beta_{1j} X_{ij}$$
- scientific focus: making inference about β_{ig}; fitting the model using OLS $\Rightarrow$ set of statistics $T_{11}, \dots, T_{1p}$, where T_{1j} is the least squares estimator of β_{1j} divided by its estimated standard error $(j = 1, \dots, p)$
- Using normal distribution with mean 0 and variance 1 as the null distribution for testing $H_{0g} : \beta_{1g} = 0$ we get G p - values $p_1, \dots ; p_G$

apply Algorithm of Storey (2002) to estimate the gene-specific FDR:

- fit
  $$E[Y_i] = \beta_{0j} + \beta_{1j} X_{ij}$$
  for each gene g, $g = 1, \dots, G$
- calculate a p - value using $\frac{\hat{\beta}_{1g}}{SE(\hat{\beta}_{1g})}$,
  let $p_1, \dots, p_G$ denote the G p - values
- Estimate π_0, the proportion of differentially expressed genes and $F_P(x)$, the cdf of the p - values by

  $\hat{\pi}_0(\lambda) = \frac{W(\lambda)}{(1-\lambda)G}$ and $\hat{F}_P(x) = \frac{min\{R(\gamma), 1\}}{G}$

  where $R(\gamma) = \#\{p_i \le \gamma\}$ and $W(\lambda) = \#\{p_i > \lambda\}$
- all rejection regions are of the form $[0, \gamma]$, $\gamma \ge 0$

- for any rejection region of interest $[0, \gamma]$, estimate pFDR as
  $$p\widehat{FDR}(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\hat{F}_P(\gamma)\{1 - (1-\gamma)^m\}}$$
- Estimate FDR as
  $$\widehat{FDR}_\gamma = \frac{\hat{\pi}_0(\lambda)\gamma}{\hat{F}_P(\gamma)}$$

Controlling procedure by Storey (2004)

to make sure, that the number of false-positive results does not exceed a previously defined number, it is necessary that $FDR \le \alpha$

- define a threshold function
  $$t_\alpha(F) = sup\{0 \le t \le 1 : F(t) \le \alpha\}$$
  where F is a function

  $\Longrightarrow$
- thresholding rule
  $$t_\alpha\left(\widehat{FDR}\right) = sup\left\{0 \le t \le 1 : \widehat{FDR}(t) \le \alpha\right\}$$
- reject null hypotheses $p_i \le t_\alpha(FDR_\gamma)$

- when the p - values are independent, the thresholding rule provides strong control of the false discovery rate at level α
- when $\lambda = 0$ one obtains the Benjamini and Hochberg (1995) procedure

1 Definition of the FDR
- Multiple Testing
- FDR and pFDR
- Controlling the FDR

2 Estimation of the FDR
- Gene - specific FDR

3 Variable Selection

4 A decision theoretic framework

5 Simulation studies
- $p < n$
- $p > n$

Joint hierarchical model for $(Y, \mathbf{X})$

- An alternative to fitting G models of the form $E[Y_i] = \beta_{0j} + \beta_{1j}X_{ij}$, is to treat $\mathbf{X}_i$ as independent variables and Y_i as the response variable for the ith subject. $i = 1, \dots, n$
  $\Rightarrow$ hierarchical normal regression model
- At the first stage of the model:

$$Y_i \overset{ind}{\sim} N\left(\mathbf{X}_i^T \beta, \sigma^2\right)$$

- For the second stage of the model, we introduce binary - valued latent variables $\gamma_1, \dots, \gamma_p$; conditional on them

$$\beta_i | \gamma_i \sim (1 - \gamma_i)\, N\left(0, \tau_i^2\right) + \gamma_i N\left(0, c_i^2 \tau_i^2\right)$$

where $c_1^2, \dots, c_p^2$ and $\tau_1^2, \dots, \tau_p^2$ are variance components.

- If $\gamma_j = 1$, then this indicates that the jth covariate should be included in the model,
  while $\gamma_j = 0$ implies that it should be excluded
- assume an inverse gamma (IG) conjugate prior for σ^2 and that γ_i is distributed as Bernoulli with probability $p_i; i = 1, \dots, p$

  $\Rightarrow$ multilevel model:

$$Y_i \overset{ind}{\sim} N\left(\mathbf{X}_i^T \beta, \sigma^2\right) \qquad (1)$$

$$\beta_i | \gamma_i \sim (1 - \gamma_i)\, N\left(0, \tau_i^2\right) + \gamma_i N\left(0, c_i^2 \tau_i^2\right) \qquad (2)$$

$$\gamma_i \overset{ind}{\sim} Be\left(p_i\right) \qquad (3)$$

$$\sigma \sim IG\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \qquad (4)$$

Gibbs sampling

for calculating the posterior distribution: instead of sampling from the joint posteriori distribution, sampling from the fully conditional distributions

- posterior distribution of β given Y, σ, γ is

$$N(A_\gamma \left(\sigma\right)^{-2} X^T X \hat{\beta}_{LS},\ A_\gamma)$$

  where $A = (\sigma^{-2} X^T X + D^{-1} R^{-1} D^{-1})$
- variance σ^2 is sampled from its posterior given γ and β, which is

$$IG(n + \tfrac{\nu}{2}, \left(Y - X^T \beta\right)^T \left(Y - X^T \beta\right) + \tfrac{\nu\lambda}{2})$$

- vector γ is sampled componentwise from the posterior distribution, the ith component $(i = 1, \dots, G)$ being Bernoulli with probability

$$P\left(\gamma_i = 1 | \gamma_{(i)}, \beta, \sigma\right) = \frac{P\left(\beta_i | \gamma_i = 1\right) p_i}{P\left(\beta_i | \gamma_i = 1\right) p_i + P\left(\beta_i | \gamma_i = 0\right)\left(1 - p_i\right)}$$

- from the point of view of selecting variables, we wish to consider the posterior distribution of $\gamma_1, \dots, \gamma_p$
- conditional distribution of
  $\hat{\beta}_l$ given $\sigma_l, \gamma_l = 0$ is $N\left(0, \sigma_l^2 + \tau_l^2\right)$, while that of
  β_l given $\sigma_l, \gamma_l = 1$ is $N\left(0, \sigma_l^2 + c_l^2 \tau_l^2\right)$
- the relative heights of these two densities at zero is

$$u_l = \left\{\frac{\sigma_l^2/\tau_l^2 + c_l^2}{\sigma_l^2/\tau_l^2 + 1}\right\}^{1/2}$$

  $\Rightarrow u_l = P\left(\gamma_l = 1 | \hat{\beta}_l = 0\right)$, which is $1 - locFDR$ of the lth variable at zero.
- the FDR based on $\hat{\beta}_l$ being in a critical region R is

$$FDR\left(R\right) = \frac{\int_{x \in R} \left\{2\pi \left(\sigma_l^2 + c_l^2 \tau_l^2\right)\right\}^{-1/2} exp\left\{\frac{-x^2}{\sigma_l^2 + c_l^2 \tau_l^2}\right\} dx}{\int_{x \in R} \left\{2\pi \left(\sigma_l^2 + \tau_l^2\right)\right\}^{-1/2} exp\left\{\frac{-x^2}{\sigma_l^2 + \tau_l^2}\right\} dx}$$

Some points to note

- characterization of the FDR based on a Bayesian framework → Bayesian framework provides a natural method of regularization
- we have utilized a variable selection framework to derive the FDR → procedures that select variables based on controlling the FDR will have certain risk optimality properties in the hierarchical model described above
- we have formulated a joint model and have derived FDR as a univariate quantity within this joint framework → no need to extend FDR to situations that are higher-dimensional if we use a univariate model
- in the framework presented here, dependence between the predictor variables is naturally incorporated into the definition of FDR

Bayesian variable selection procedure

Because we are using a Gibbs sampling algorithm in order do derive the posterior distribution in the model, the FDR can be derived easily:

- fixing an rejection region R, we simply count the proportion of MCMC samples in which the $\gamma = 0$ and $\beta \in R$
- based on the posterior distribution, we can develop a univariate variable selection procedure
- we can rank $P\left(\gamma_i = 0 | Y_1, \dots, Y_n\right), i = 1, \dots, G$ and select the variables with small posterior probabilities

Algorithm:

1. set level to be α and fix a rejection region R
2. fit model (1)-(4) using MCMC methods
3. based on the MCMC output, calculate $pp_i = P(\gamma_i = 0 | \hat{\beta}_i \in R)$
4. let $pp_{(1)} \leq \dots \leq pp_{(G)}$ denote the sorted values of $pp_1, \dots, pp_n$ in increasing order
5. find $\hat{k} = max\left\{1 \leq k \leq G : pp_k \leq \frac{\alpha k}{G}\right\}$, select variables $1, \dots, G$

if the predictor variables are orthogonal or whenever $P(\gamma_i = 0 | \hat{\beta}_i \in R)$ is an monotonic function of the univariate p - values the algorithm is equivalent to the Benjamini and Hochberg (1995) procedure.

1 Definition of the FDR
- Multiple Testing
- FDR and pFDR
- Controlling the FDR

2 Estimation of the FDR
- Gene - specific FDR

3 Variable Selection

4 A decision theoretic framework

5 Simulation studies
- $p < n$
- $p > n$

Risk inflation

Here we consider the hierarchical regression model from section 3 and study the properties of the variable selection procedure from a decision theoretic perspective.

- Define $R(\beta,\hat{\beta})$ to be the predictive risk of the estimator $\hat{\beta}$,

$$R(\beta,\hat{\beta}) = E_\beta \left| X\hat{\beta} - X\beta \right|^2$$

- the vector γ of latent variables can take 2^p possible values. Let $\zeta = (\zeta_1,\ldots,\zeta_G)$ denote the true model, so $\zeta_i = I\left(\beta_i \neq 0\right);\ i = 1,\ldots;G$
- The risk inflation is given by

$$RI\left(\gamma\right) = \sup_\beta \frac{R(\beta,\hat{\beta}_\gamma)}{R(\beta,\hat{\beta}_\zeta)}$$

$$RI\left(\gamma\right) = \sup_\beta \frac{R(\beta,\hat{\beta}_\gamma)}{R(\beta,\hat{\beta}_\zeta)} \qquad (5)$$

- the denominator $R(\beta,\hat{\beta}_\zeta)$ is the lowest possible risk, since it represents the risk for the ideal model
- the risk inflation reflects the worst-possible increase in risk with using a combination selection/estimation procedure

  → we wish to find procedures that minimize (5) over a large class of procedures

- Foster and George (1994): for the case of diagonal X^TX the optimal rule that minimizes (5) is a threshold rule that selects the top $(2\log G)$ variables based on the absolute magnitude of the univariate statistics

  → equivalently, the optimal threshold rule selects the $2\log G$ variables with the smallest univariate p-values
- the Benjamini-Hochberg (1995) procedure is a data-dependent threshold rule that is a special case of the class of FDR-controlling procedures proposed by Storey et al (2004)

  → thus, when $\hat{k} \approx (2\log G)$, then the Benjamini-Hochberg (1995) procedure will be the optimal from a risk inflation framework
- in general case where X^TX is nonorthogonal: the RI is bounded from below by $2\log G - o(\log G)$

1 Definition of the FDR
- Multiple Testing
- FDR and pFDR
- Controlling the FDR

2 Estimation of the FDR
- Gene - specific FDR

3 Variable Selection

4 A decision theoretic framework

5 Simulation studies
- $p < n$
- $p > n$

First situation: $p < n$

we consider the model $E[Y_i] = \beta_{0j} + \beta_{1j}X_{ij}$

- n=50 and p=10
- the true model is $E[Y] = X_1 + 1.5X_2 + 3X_3$
- the variance of the error term in all simulation studies is one, 250 simulations
- the predictors were generated with correlation $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$
- a ROC curve was constructed based on taking the top k variables (k=1,2,3,4,5 and 10) based on the estimated posterior probability

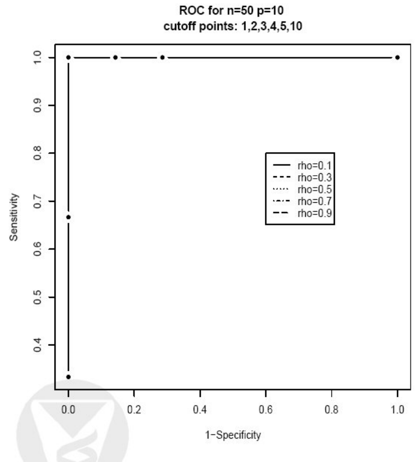


Figure 1: Plot of ROC curve for simulation setting when $n = 50$ and $p = 10$. Variables ranked univariately based on marginal posterior probability. ROC averaged across 250 simulations.

Risk behavior

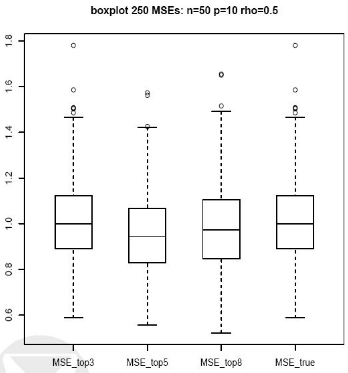
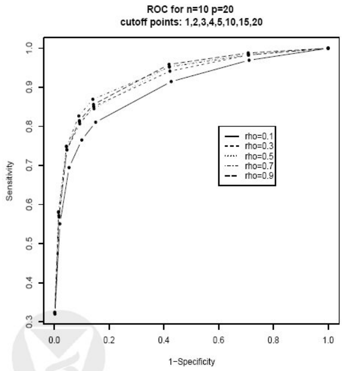
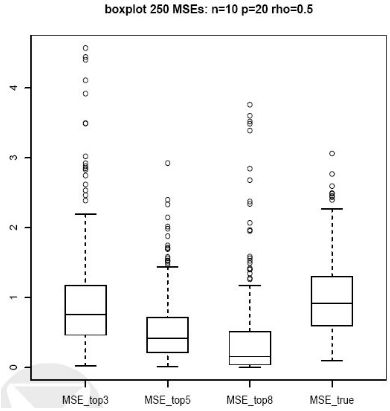


Figure 2: Mean squared errors based on taking top k variables ($k = 3, 5, 8$) and true MSE averaged across 250 simulations.

Second situation: $p > n$





Literature

- Gosh, D., W. Chen, and T. Raguhathan. 2004. The false discovery rate: a variable selection procedure. Preprint.
- Rottenkolber, M. 2005. Untersuchung von False Discovery Rate Kontrollprozeduren zur Identifikation differentiell exprimierter Gene. Diploma Thesis, Department of Statistics, University of Munich.
- Storey, J.D. 2002. A direct approach to false discovery rates. J. Roy. Statist. Soc. B 64, 479-498
- Storey JD, Taylor JE, and Siegmund D. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. Journal of the Royal Statistical Society, Series B, 66: 187-205.