Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays

Peter Müller
Dpt of Biostat UT, M.D. Anderson Cancer Center
Giovanni Parmigiani
Johns Hopkins U.
Christian Robert
Université Paris Dauphine & CREST, INSEE
Judith Rousseau
Université Rene Descartes, Paris & CREST, INSEE

Outline

Sample size choice for massive multiple comparisons.

I. Decision problems:

1. Terminal decision

2. Sample size

3. Simulation: preposterior MCMC

II. Prob Model:

4. A hierarchical Gamma/Gamma model

5. A mixture model extension

6. Results

Summary

- Discussion indep of prob model
- Two dec problems: sample size & multiple comparison
- *One* loss function for both.
- Terminal decision under reasonable loss functions:
  Reject if marg posterior prob $>$ threshold t.

Summary (ctd.)

- Sample size: based on same loss function
- Supplemented by power calculation and sensitivity analysis
- Evaluation by (easy) preposterior M.C.
- Prob models: (mixture of) gamma/gamma

0. Microarrays

Statistically, **massive multiple testing:** comparison of genes under two (or more) conditions to detect differentially expressed genes

	Gene A	Gene B	Gene C	...
Patient I	1	0	0	...
Patient II	1	0	1	...
Patient III	0	0	0	...
...	...	...	...	...

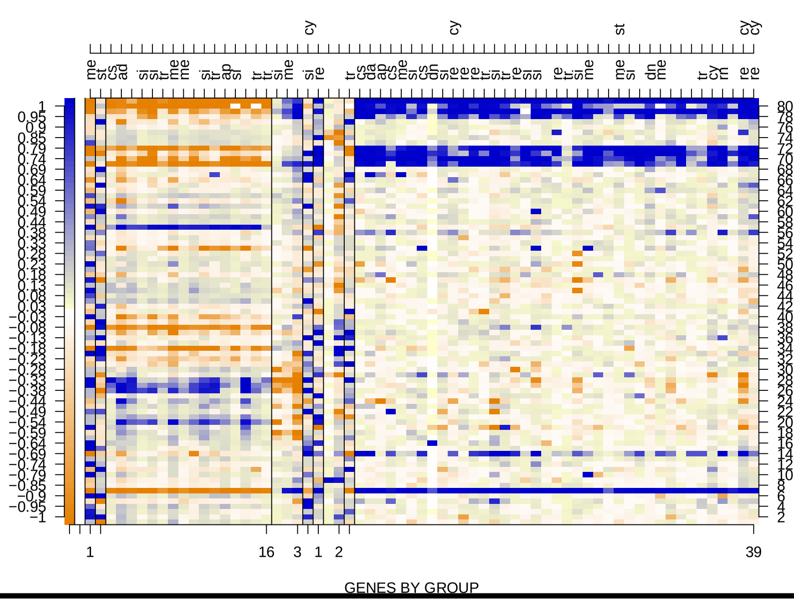
Huge number of genes and moderate to large number of cases

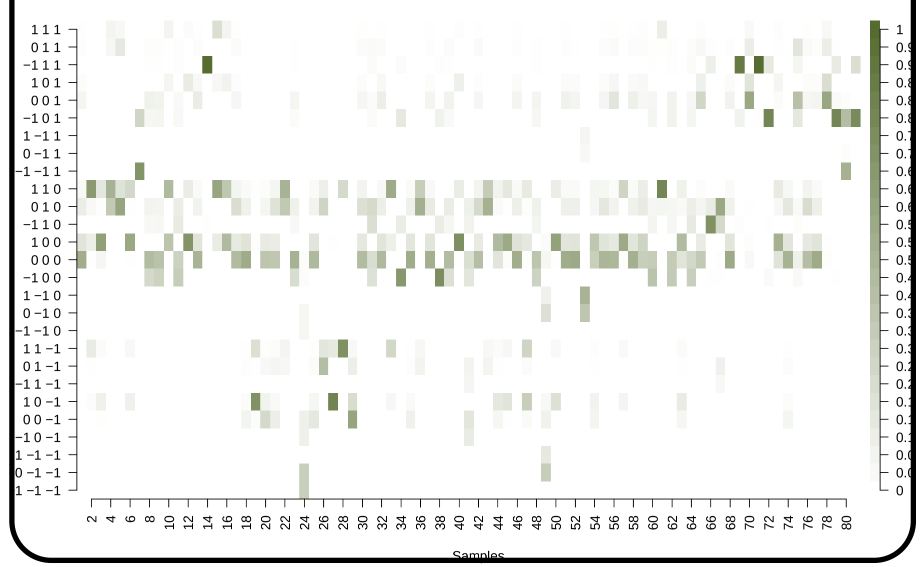
Costly preliminary step to even more costly assays (e.g., RT-PCR)

[Parmigiani et al., 2002 JRSSb]

Observations: intensities y_{ij} (dye fluorescence)

	Gene A	Gene B	Gene C	
Patient I	.67	-.12	.08	...
Patient II	-.54	.19	.01	...
Patient III	-.13	-.15	.03	...
...	...	...	...	...





I. Decision Problems

Sequential decision problem:

First sample size, then terminal decision

1. Sample size choice:

- *Before* experiment, marginalize over putative data
- Decide # arrays J

2. Multiple comparisons:

- *After* experiment, conditional on observed data
- Decide about n genes: differential vs. non-differential

1. Terminal Decision

Decision and Hypotheses: n comparisons, $i = 1, \ldots, n$:

truth $z_i \in \{0, 1\}$

decision $d_i \in \{0, 1\}$

n **huge** plus hierarchical structure:

isolated hypothesis testing one gene at a time very inefficient

Natural Bayesian framework to pull strength from all observations with help from Decision Theory (loss function)

A central notion: FDR & FNR

[Benjamini & Hochberg, 95, JRSSb]

False Discovery Rate: generalizes type I error for multiple comparisons.

Let $D = \sum d_i =$ # rejections.

$$\text{FDR}(d, z) = \frac{\sum d_i (1 - z_i)}{D} = \frac{FD}{D}$$

False negative rate: akin to type II error

$$\text{FNR}(d, z) = \frac{\sum (1 - d_i) z_i}{n - D} = \frac{FN}{n - D}$$

Posterior mean FDR & FNR:

[Genovese & Wasserman, Valencia 7 & JRSSb 02]

Posterior expected FDR

Let $v_i = \Pr(z_i = 1 \mid y)$:

$$\overline{\text{FDR}} = \mathbb{E}(\text{FDR} \mid y) = \mathbb{E}\left(\frac{\sum d_i(1-z_i)}{D}\Big| y\right) = \frac{\sum d_i(1-\boldsymbol{v_i})}{D} = \frac{\overline{\text{FD}}}{D}$$

Simplification : only need marginal post probs $\boldsymbol{v_i}$.

Same for FNR

$$\overline{\text{FNR}} = \mathbb{E}(\text{FNR} \mid y) = \frac{\sum(1-d_i)\boldsymbol{v_i}}{n-D} = \frac{\overline{\text{FN}}}{D}$$

(Expected) Loss functions

Combine FD(R) and FN(R)

1. $L_N(d,y) = c\,\overline{\text{FD}} + \overline{\text{FN}} \quad (c > 0)$
2. $L_R(d,y) = c\,\overline{\text{FDR}} + \overline{\text{FNR}} \quad (c > 0)$
3. $L_{2N}(d,y) = (\overline{\text{FD}}, \overline{\text{FN}})$
4. $L_{2R}(d,y) = (\overline{\text{FDR}}, \overline{\text{FNR}})$

Multicriteria decision problems:

L_{2N}: $\min_d \overline{\text{FN}}$ subject to $\overline{\text{FD}} \le \alpha$.

L_{2R} : $\min_d \overline{\text{FNR}}$ subject to $\overline{\text{FDR}} \le n\,\alpha$.

[Back to basics?!]

Optimal terminal decision

Can show: optimal terminal decision (arg min posterior expected loss)

$$d_i = \mathbb{I}_{(v_i > t)}$$

under **all** four loss functions.

Threshold t_L^*: Loss functions differ only in $t = t_L^*$:

$$t_N^*(y_J) = c/(1+c), \qquad \text{[constant]}$$
$$t_R^*(y_J) = \ldots \qquad \text{[implicit]}$$

$$t_{2R}^*(y_J) = \min\{t:\ \overline{\text{FDR}}(t, y_J) \le \alpha\},$$
$$t_{2N}^*(y_J) = \min\{t:\ \overline{\text{FD}}(t, y_J) \le \alpha\},$$

For

$$\begin{aligned} L_R &= c\frac{\sum d_i(1-v_i)}{D} + \frac{\sum(1-d_i)v_i}{n-D} \\ &= C_1(D) - C_2(D)\sum d_i v_i + C_3(D)\sum v_i \frac{\sum(1-d_i)v_i}{n-D} \\ &\ge C_1(D) + C_2(D)\sum_{i=n-D+1}^{n} v_{(i)} + C_3(D)\sum_{i=1}^{n-D} v_{(i)} \\ &\ge C_1(D^\star) + C_2(D^\star)\sum_{i=n-D^\star+1}^{n} v_{(i)} + C_3(D^\star)\sum_{i=1}^{n-D^\star} v_{(i)} \end{aligned}$$

we derive

$$\boldsymbol{t^\star = v_{(n-D^\star)}}$$

Properties

Decision: $d_i = \mathbb{I}_{(v_i > t)}$

L_N: t is constant and FDR $\longrightarrow 0$

L_{2R}: $\overline{\text{FDR}}$ is constant (??)

L_{2N}: $\overline{\text{FD}}$ is constant (??)

L_R: ???

2. Sample Size Determination

Preposterior mean loss $\rightarrow$ sample size J.

$$L_R^m(J) = \mathbb{E}_{y_J}[\min_d\{L_R(d, y_J)\}],$$
$$L_N^m(J) = \mathbb{E}_{y_J}[\min_d\{L_N(d, y_J)\}],$$

and

$$L_{2R}^m(J) = \mathbb{E}_{y_J}\left[\min_d\{\overline{\text{FNR}}(d, y_J) \mid \overline{\text{FDR}}(d, y_J) \le \alpha\}\right]$$
$$L_{2N}^m(J) = \mathbb{E}_{y_J}\left[\min_d\{\overline{\text{FN}}(d, y_J) \mid \overline{\text{FD}}(d, y_J) \le \alpha\}\right]$$

Expectation: $\mathbb{E}[\cdot]$ w.r.t. y_J

(model parameters already integrated in L)

Nested optimization: min w.r.t. d

Rates of Convergence (in J)

$$P(z = 1|y, \eta) = \frac{1}{1 + e^{-J\Delta}\sqrt{J}}$$

but ...Big... bummer:

$$\overline{\text{FNR}}(y_J, t^*) = O_P(\sqrt{\log J/J})$$

under L_{2R}, L_{2N}, L_N

and

$$\overline{\text{FN}}(y_J, t^*) = O_P(n\sqrt{\log J/J})$$

under L_R

Much too slow/flat!!

Power/Sensitivity

Power: prob accept as function of true effect & J

$$\beta(\rho) = \Pr\{v_i(y) > t(y) \mid \rho\} = \int \mathbb{I}_{v_i(y) > t(y)}\, \mathrm{d}p(y|\rho)$$

where ρ level of differential expression

($\rho = 0$ meaning no difference, i.e. $z = 0$)

Incorporation in the design and possible calibration of loss

3. Simulation

Compute L and L^m by **preposterior** simulation

Simulation: Loop over repeated simulations

1. *Prior:* $(\omega^o, z^o) \sim p(\omega, z)$.
2. *Data:* $y_{J_1} \sim p(y_{J_1} \mid \omega^o, z^o)$.
3. *Grid:* $J = J_0, \ldots, J_1$, let $y_J \subset y_{J_1}$

   [requires **only one** simulation run]

- *Posterior MCMC:* Compute $v_i = \text{Pr}(z_i = 1 | y_J)$.

  [**easy MCMC**, starting with **true** ω^o]
- *Thresholds:* Compute the cutoffs t^*_L
- Save $\overline{\text{FN}}(t^*, y_J)$ and $\overline{\text{FNR}}(t^*, y_J)$.
- Save (J, ρ^o_i, d_i).

Simulation (ctd.)

Curve Fitting of Monte Carlo Experiments:

1. *Expected loss* $L^m(J)$, $\overline{FN}_m$ *and* $\overline{FNR}_m$*:*

   Curve through $(J, \overline{\text{FNR}})$ and $(J, \overline{\text{FN}})$ using $\sqrt{\log J / J}$
2. *Power:* Fit curve through (J, ρ^o_i, d_i).

Optimal sample size:

Use $\widehat{L}^m(J)$ and power curves for an informed choice.

II. Probability Model

[Newton et al. (01, J Comp Bio), Newton & Kendziorski (03)]

4. Gamma/Gamma hierarchical model

Observed gene expression:

$$X_{ij} \sim \text{Ga}(a, \theta_{0i}) \text{ and } Y_{ij} \sim \text{Ga}(a, \theta_{1i}).$$

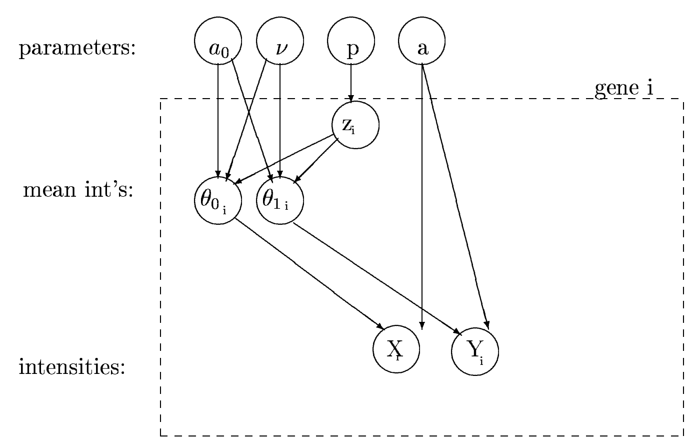
Mean expressions:

$$\theta_{0i} \sim \text{Ga}(a_0, \nu)$$

$$\theta_{1i} = \begin{cases} \theta_{0i} & \text{if } z_i = 0 \\ \sim \text{Ga}(a_0, \nu) & \text{if } z_i = 1 \end{cases}$$

Differential expression:

$$\text{Pr}(\theta_{0i} = \theta_{1i}) = \text{Pr}(z_i = 0) = p.$$



Closed Form Expressions

We find

$$p(X_i, Y_i | z_i = 0, \eta) = \left\{ \frac{\Gamma(2Ja + a_0)}{\Gamma(a)^{2J}\Gamma(a_0)} \right\} \frac{(\nu)^{a_0} \left(\prod_j X_{ij} \prod_j Y_{ij}\right)^{a-1}}{[(\sum_j X_i + \sum_j Y_i + \nu)]^{2a+a_0}}$$

and

$$p(X_i, Y_i | z_i = 1, \eta) = \left\{ \frac{\Gamma(aJ + a_0)}{\Gamma(a)^J \, \Gamma(a_0)} \right\}^2 \frac{(\nu\nu)^{a_0} \left(\prod_j X_{ij} \prod_j Y_{ij}\right)^{a-1}}{[(\sum_j X_{ij} + \nu)(\sum_j Y_{ij} + \nu)]^{a+a_0}} ,$$

5. Mixture of Ga/Ga hierarchical model

- ***Mixture of Gammas:***

$$X_{ij} \sim \int \mathsf{Ga}(a, \theta_{0i}\, r_{ij}) \; \mathsf{d}p(r_{ij}|w)$$
$$= \sum_{k=1}^{K} w_k \mathsf{Ga}(a, \theta_{0i}\, r_{ij}^k)$$

and

$$Y_{ij} \sim \int \mathsf{Ga}(a, \theta_{0i}\, s_{ij}) \; \mathsf{d}p(r_{ij}|w)$$

Addresses issues of noise in data collection and experimental conditions, based on a pilot run on control tissue

- ***Plus slide specific mixture:***

$$(X_{ij}|r_{ij}, g_j) \sim \mathsf{Ga}(a, \theta_{0i}\, g_j\, r_{ij})$$

and

$$(Y_{ij}|s_{ij}, g_j) \sim \mathsf{Ga}(a, \theta_{1i}\, g_j\, s_{ij}),$$

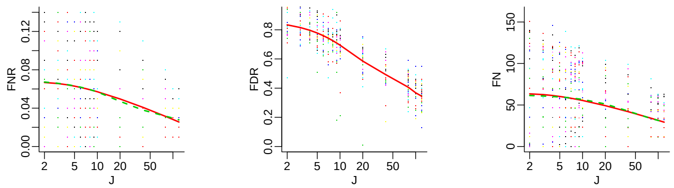
for outlier slides, again based on pilot data.

- MCMC: remains (almost) the same (add'al RJMCMC bits for # of components on both mixtures, found to be $K = 3$ and $L = 2$ on the dataset used in Newton et al., 2001)

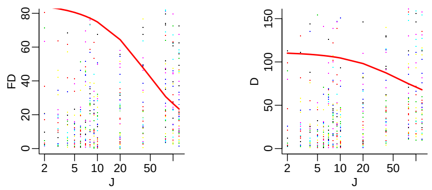
6. Results

Simulation from the estimated mixtures ($K = 3, \; L = 2$) rather than from the prior (use of pilot dataset)

Results L_N: t^*_N fixed by design

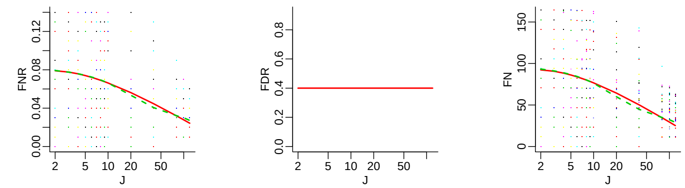


(a) $\overline{\text{FNR}}_m(J, L_N)$ (b) $\overline{\text{FDR}}_m(J, L_N)$ (c) $\overline{\text{FN}}(J, L_N)$

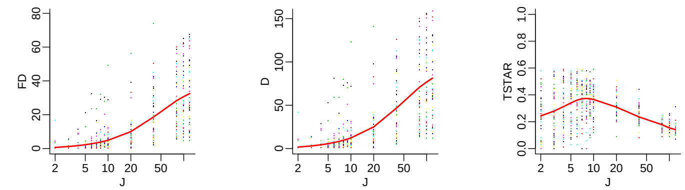


(d) $\overline{\text{FD}}_m(J, L_N)$ (e) $\overline{\text{D}}_m(J, L_N)$

Results L_{2R}: $\overline{\text{FDR}}_m$ fixed by design $\alpha = .4$

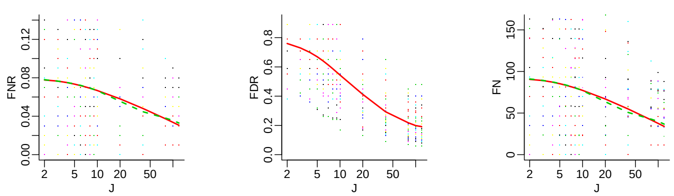


(a) $\overline{\text{FNR}}_m(J, L_{2R})$ (b) $\overline{\text{FDR}}_m(J, L_{2R})$ (c) $\overline{\text{FN}}_m(J, L_{2R})$

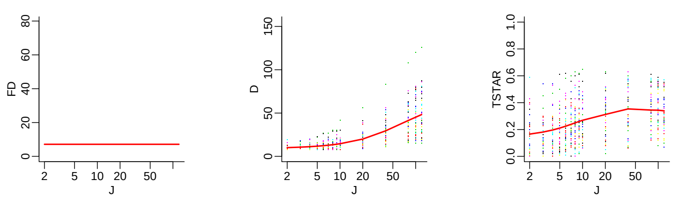


(d) $\overline{\text{FD}}_m(J, L)$ (e) $\overline{\text{D}}_m(J, L_{2R})$ (f) $\overline{t^*}_m(J, L_{2R})$

Results L_{2N}: $\overline{\text{FD}}$ fixed by design $\alpha_N = 0.1n\bar{p}\alpha/(1-\alpha)$

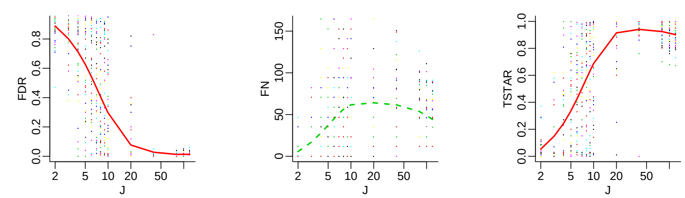


(a) $\overline{\text{FNR}}_m(J, L_{2N})$ (b) $\overline{\text{FDR}}_m(J, L_{2N})$ (c) $\overline{\text{FN}}_m(J, L_{2N})$



(d) $\overline{\text{FD}}_m(J, L_{2n})$ (e) $\overline{\text{D}}_m(J, L_{2n})$ (f) $\overline{t^*}_m(J, L_{2N})$

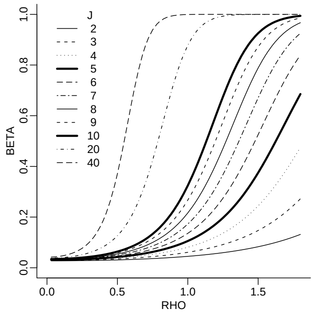
Results L_R : arrrrgh!!!



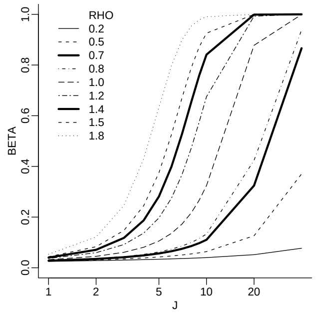
(a) $\overline{\text{FDR}}_m(J, L_R)$ (b) $\overline{\text{FN}}_m(J, L_R)$ (c) $\overline{t^*}_m(J, L_R)$

Awkward jump in $\overline{\text{FDR}}_m$ (and $\overline{t^*}_m$).

Power $\beta = \mathbb{E}_{y_J}\{\text{Pr}(d = 1|\rho, y_J)\}$ against $\rho = \log(\theta_{0i}/\theta_{1i})$ and J



(a) β against ρ (by J)



(b) β against J (by ρ)

Summary (again!)

- Discussion indep of prob model
- Two dec problems: sample size & multiple comparison
- *One* loss function for both.
- Terminal decision under reasonable loss functions:
  Reject if marg posterior prob $>$ threshold t.
- Sample size: based on same loss function
- Supplemented by power calculation
- Evaluation by (easy) preposterior M.C.
- Prob models: (mixture of) gamma/gamma