Some results for dependence in high-dimensional multiple hypothesis testing situations

Sandy Clarke
supervised by Professor Peter Hall

Mathematics and Statistics Department
The University of Melbourne

November, 2009

"discussed in more detail"

Outline

1. Problem
   - High dimensional multiple hypothesis testing
   - Stronger control
   - Procedures
   - Dependence
2. Possible solutions
   - Ignore positive dependence
   - Use conservative critical values
   - Estimate correlation structure
   - Make assumptions about the correlation structure
3. Our results
   - Weighted moving average model
   - Results for dependence
   - Impact on procedures

Hypothesis testing

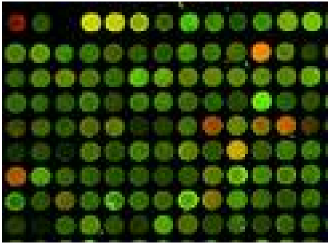
- test statistic:
  X_1
- null hypotheses:
  $H_{01} : \mu_i = 0$
- one-sided test:
  reject H_{01} if $X_1 > x$
- choose x so that:
  $P_0(X_1 > x) = \alpha$

High dimensional multiple hypothesis testing

- test statistics:
  $X_1, X_2, \ldots, X_m$ (m very large)
- null hypotheses:
  $H_{01}, H_{02}, \ldots, H_{0m} : \mu_i = 0$
- one-sided test:
  reject H_{0i} if $X_i > x$

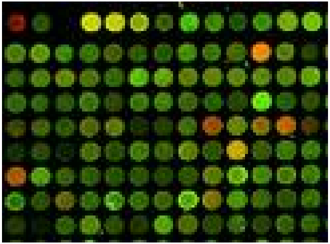
High dimensional multiple hypothesis testing

- test statistics:
  $X_1, X_2, \ldots, X_m$ (m very large)
- null hypotheses:
  $H_{01}, H_{02}, \ldots, H_{0m} : \mu_i = 0$
- one-sided test:
  reject H_{0i} if $X_i > x$
- example:
  DNA microarray expression data



High dimensional multiple hypothesis testing

- test statistics:
  $X_1, X_2, \ldots, X_m$ (m very large)
- null hypotheses:
  $H_{01}, H_{02}, \ldots, H_{0m} : \mu_i = 0$
- one-sided test:
  reject H_{0i} if $X_i > x$
- example:
  DNA microarray expression data
- curse of dimensionality: $n \ll m$



Stronger control of Type I errors

- $P_0(X_1 > x) = 0.05$ gives too many false positives
- very strict error rates

FWER: $P(\text{false rejections} \geq 1) \leq \alpha$

for example: Holm (1979)

GFWER: $P(\text{false rejections} \geq k) \leq \alpha$

for example: Lehmann & Romano (2005)

Stronger control of Type I error

- error rates which favour more rejections

FDR: $E(\frac{\text{false rejections}}{\text{rejections}}) \leq \alpha$

for example: Benjamini & Hochberg (1995)

tFDP: $P(\frac{\text{false rejections}}{\text{rejections}} \geq c) \leq \alpha$

for example: Lehmann & Romano (2005)

Types of procedures

- one-step
  compare all X_i to x which depends only on α and m
- step-down
  compare each $X_{(i)}$ to x_i from largest to smallest until one is not rejected.
- step-up
  compare $X_{(i)}$ to x_i from smallest to largest until one is rejected

Types of procedures

Example: Benjamini and Hochberg (1995)

- controls FDR at α
- step-up procedure
- algorithm:
  for x_i such that $P_0(X > x_i) = \frac{i\alpha}{m}$
  1. if $X_{(1)} > x_m$ reject $X_{(1)}, \ldots, X_{(m)}$ and exit
  2. if $X_{(2)} > x_{m-1}$ reject $X_{(2)}, \ldots, X_{(m)}$ and exit
  3. if $X_{(3)} > x_{m-2}$ reject $X_{(3)}, \ldots, X_{(m)}$ and exit
  4. etc...

The assumption of independence

The assumption of independence is rarely valid:
"It is generally assumed that genes or proteins that act together in a pathway will exhibit strong correlations among their expression values, evident as gene clusters" (p. 46)

Clarke et al (2008) in *Nature Reviews*

The assumption of independence has consequences.

Ignore positive dependence

- e.g. Benjamini and Yekutieli (2001)
- for certain kinds of positive dependence, the BH procedure controls FDR

Ignore positive dependence

- e.g. Benjamini and Yekutieli (2001)
- for certain kinds of positive dependence, the BH procedure controls FDR
- conservative control which doesn't take advantage of potential gains in power from dependence

Use conservative critical values

- e.g. Benjamini and Yekutieli (2001)
- choose x_i such that

$$P_0(X > x_i) = \frac{i\alpha}{m \sum_{i=1}^{m} i}$$

- control for general dependence

Use conservative critical values

- e.g. Benjamini and Yekutieli (2001)
- choose x_i such that

$$P_0(X > x_i) = \frac{i\alpha}{m\sum_{i=1}^{m} i}$$

- control for general dependence
- severe reduction in power

Estimate correlation structure

- e.g. Westfall and Young (1993)
- estimate correlation matrix or use resampling to 'break' the correlation
- ideally, provides the true null distribution of the test statistics

Estimate correlation structure

- e.g. Westfall and Young (1993)
- estimate correlation matrix or use resampling to ‘break’ the correlation
- ideally, provides the true null distribution of the test statistics
- practically, computationally demanding and unreliable for $n \ll m$

Make assumptions about the correlation structure

- e.g. Efron (2007)
- hierarchical Poisson structure for histogram counts of test statistics
- enables the summary of correlation by a single parameter, A, used to correct the standard FDR estimate:

$$FDR(x|A) = FDR(x)\left[1 + A\frac{x\phi(x)}{\sqrt{2}(1-\Phi(x))}\right]$$

Make assumptions about the correlation structure

- e.g. Efron (2007)
- hierarchical Poisson structure for histogram counts of test statistics
- enables the summary of correlation by a single parameter, A, used to correct the standard FDR estimate:

$$FDR(x|A) = FDR(x)\left[1 + A\frac{x\phi(x)}{\sqrt{2}(1-\Phi(x))}\right]$$

- appropriateness of the structure is questionable

Weighted moving average model

- MA_r $\qquad r : \#\{\theta_k \neq 0\}$
- $X_i = \sum_k \theta_k \epsilon_{i+k}$

  constant θ_k's, r finite, ϵ_i's iid and $-\infty < i < \infty$

Weighted moving average model

- MA_r $\qquad r : \#\{\theta_k \neq 0\}$
- $X_i = \sum_k \theta_k \epsilon_{i+k}$

  constant θ_k's, r finite, ϵ_i's iid and $-\infty < i < \infty$
- simple but not unreasonable representation
- t-statistic

Weighted moving average model

- t-statistic

$$\begin{matrix} Y_{1,1} & Y_{1,2} & \cdots & Y_{1,m} \\ Y_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \cdots & Y_{n,m} \end{matrix}$$

$Y_{ji} = \mu_i + \sum_k \theta_k \epsilon'_{j,i+k}$
for $1 \leq j \leq n$ and $1 \leq i \leq m$
with ϵ'_{ij} iid, mean 0

for n large enough, under H_{0i}:
$X_i \approx \sum_k \theta_k \epsilon_{i+k}$,
where $\epsilon_i = n^{-1} \sum_{1 \leq j \leq n} \epsilon'_{ji}$

Theoretical results

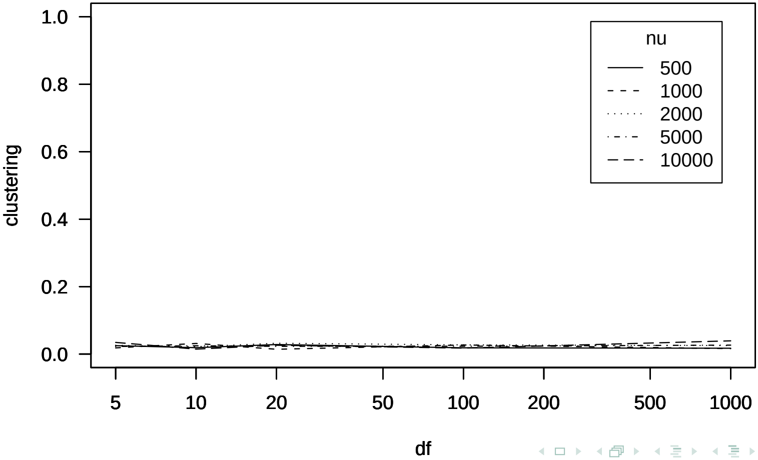
MA_r: $X_i = \sum_k \theta_k \epsilon_{i+k}$, exceedences: $\{X_i > x\}$

- no clustering of exceedences for light-tailed data
- clustering persists for heavy-tailed data:
  if $\theta_{(1)} \geq \cdots \geq \theta_{(r)}$, then

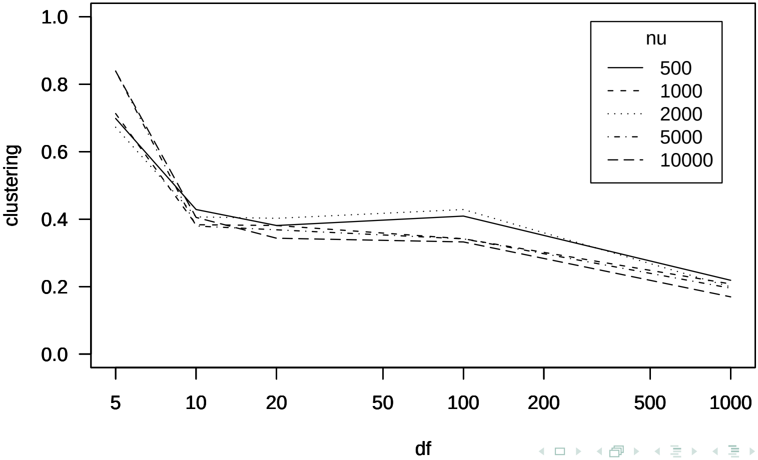
$$P(M = q | M > 0) \to \frac{\theta_{(q)}^{\rho} - \theta_{(q+1)}^{\rho}}{\theta_{(1)}^{\rho}}$$

  where M is the limiting distribution of cluster size
- intuitive explanation
- e.g. if $\theta_1 = \theta_2 = \cdots = \theta_r$, then $P(M = r | M > 0) \to 1$

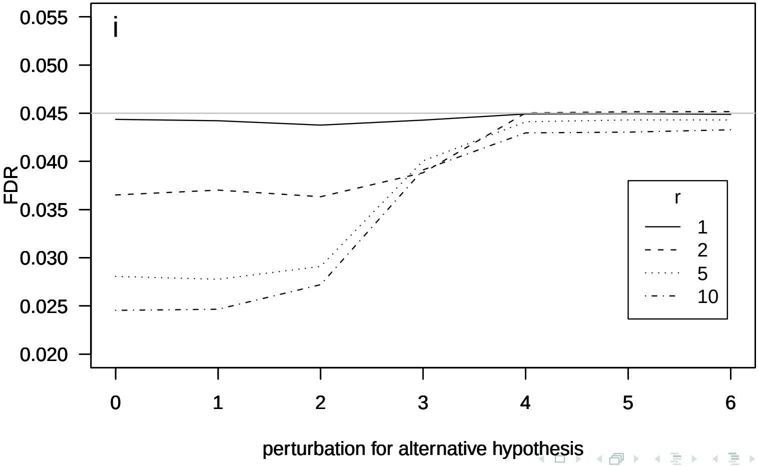
Simulation results – clustering for independent data



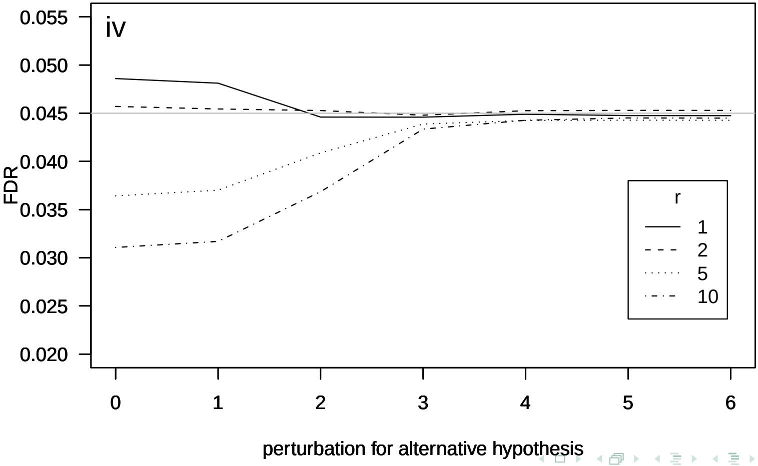
Simulation results – clustering with $r = 10$



Simulation results – FDR with $\nu = 500$ and $df = 5$

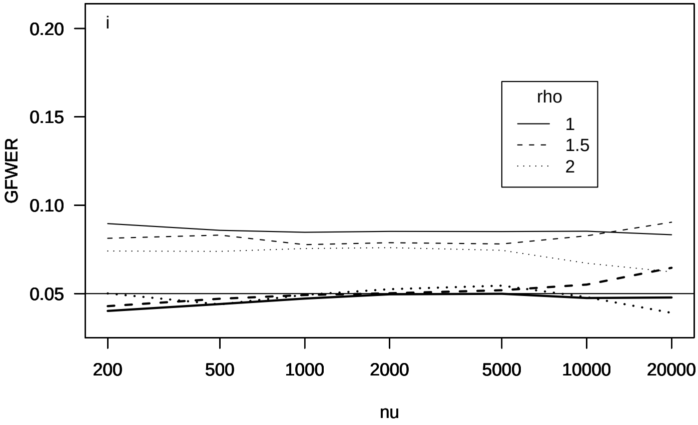


Simulation results – FDR with $\nu = 500$ and $df = \infty$



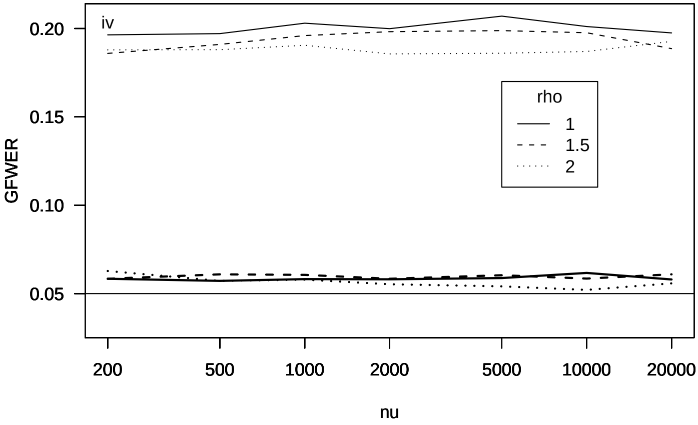
What can we do about dependence when it matters?

- estimate tail-weight and θ_k and adjust appropriately



What can we do about dependence when it matters?

- estimate tail-weight and θ_k and adjust appropriately



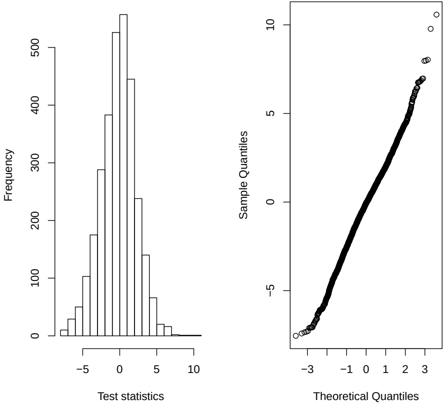
Does it ever matter?

Most data sets are normally distributed

Golub (1999)

leukemia data

- observations themselves are averages
- test statistics are averages



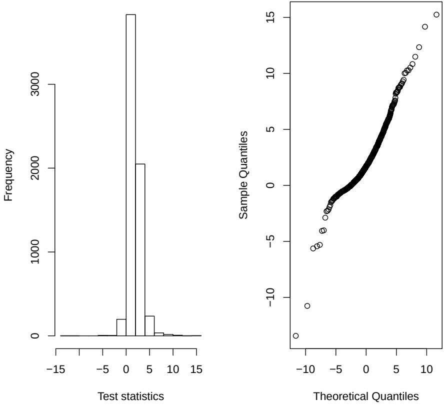
Does it ever matter?

Or at least light tailed

Callow (2000)

mouse cholesterol data

- observations themselves are averages
- test statistics are averages



Thank you

Questions?