Bounding the k-family-wise error rate using resampling methods

Tijl De Bie and John Shawe-Taylor

University of Bristol, K.U.Leuven, University College London

May 2006

Overview

- I will discuss a methodology to bound
  - the Family-Wise Error Rate (FWER)
  - the k-FWER
  - the False Discovery Proportion (FDP)

  by means of randomisation (resampling)
- Want to achieve:
  - quantification and understanding of conclusions of multiple testing
  - take account of dependencies between tests
- (Results build on previous work by Meinshausen and others! See references)

Overview

- Single and multiple hypothesis testing (notation / formalism / terminology)
- Randomisation testing
- Bounding the FWER, k-FWER, FDP by resampling
- Practical results

Motivating example

Assumption:

- Gene A is unrelated to tissue type (e.g. 'cancerous' / 'non-cancerous')

Observation:

- Expression of Gene A in a set of tissue samples
- Label of each of these tissue samples

Question:

- Is the observation 'surprising', given our assumption?
- If so, something more interesting than the null hypothesis may be true...

A pattern function to quantify the state

- Aim: draw statistical conclusions about state of the world X based on a finite sample
- 'Particular state': quantified by means of a *pattern function*

$$\pi : X \rightarrow \mathbb{R}$$

- For example, $\pi =$ correlation of the expression with the label (-1/1)

The null hypothesis

- Which statistical conclusions can we draw?
- One computes an upper bound on the probability to see a pattern as strong as $\pi(X) \geq \sigma$, assuming a null hypothesis, the 'dull' assumption
- Null hypothesis = a set Ω_0 of distributions $\mathcal{D}$ over X
- Hence, level of surprise is quantified by p:

$$\forall \mathcal{D} \in \Omega_0 : P_{X \sim \mathcal{D}} (\pi(X) \geq \sigma) \leq p$$

- p is called the p-value

Rejecting the null hypothesis

- If p is small, the level of surprise is large $\rightarrow \Omega_0$ *rejected*
- If Ω_0 holds nevertheless: *false rejection* (aka: type I error)
- If Ω_0 does not hold: *true rejection*
- If Ω_0 does not hold but still $\pi(X) < \sigma$: type II error

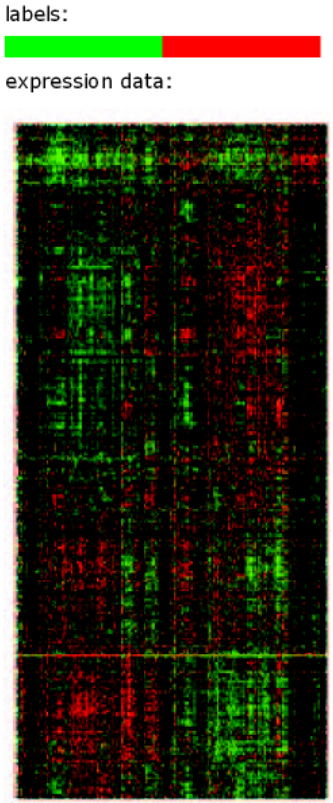
test $\setminus$ Ω_0	true	false
accept	just acceptance	type II error
reject	type I error	true rejection

- Challenge: keep type I and II errors small

Motivating example

Back to our genes...

- Microarrays gather data for all n genes in genome



Motivating example

Back to our genes...

- Microarrays gather data for all n genes in genome
- Doing n single hypothesis tests will cause many false rejections
- Assuming the genes are independent and Ω_0 holds, pn genes are expected to be rejected! (e.g. for $20,000$ genes and $p = 0.01$, this is 200)
- Should all these genes be subjected to a closer look of the biologist??
- Need for a new formalism

A set of pattern functions

- Multiple testing: a set of pattern functions, each quantifying one aspect of the state of the world:

$$\Pi = \{\pi_\alpha : \alpha \in A\}$$

  with A and index set of size $|A|$
- For example: π_α measures correlation of gene α with the labels
- Now we have $|A|$ different tests

$$\pi_\alpha(X) \geq \sigma_\alpha$$

  (note: X contains all expression data, for all genes)

A set of pattern functions

- Null hypothesis: Ω_0 contains all distributions of X where the labels and the gene expression data are independent
- Each gene may provide evidence to reject null hypothesis now! If:

$$\pi_\alpha(X) \geq \sigma_\alpha$$

- FWER: probability to reject null hypothesis assuming that it holds (i.e. the probability than any of the genes motivates rejection)
- The FWER is $\leq p$ if

$$\forall \mathcal{D} \in \Omega_0 : P_{X \sim \mathcal{D}}\left(\exists \alpha \in A : \pi_\alpha(X) \geq \sigma_\alpha\right) \leq p$$

Alternative null hypotheses

- Assuming that all genes are unrelated to the labels is probably too simplistic (hopefully!)
- Some will be truly unrelated
- How can we do more than merely reject the null hypothesis or not, based on some genes as a witness?
- In particular: can we say anything about these genes that show surprising behaviour?

Alternative null hypotheses

- Null hypothesis $\Omega_0\left(\overline{A}\right)$ = the sets of all distributions where all but the subset $\left\{\pi_\alpha : \alpha \in \overline{A}\right\} \subseteq \Pi$ of $k = \left|\overline{A}\right|$ pattern functions is jointly distributed as under a $\mathcal{D} \in \Omega_0$
- Hence: we tolerate $k = \left|\overline{A}\right|$ positive genes (i.e. genes that are *potentially* related with the labels)
- Then the probability of at least one *false* rejection is

$$\forall \mathcal{D} \in \Omega_0\left(\overline{A}\right) : P_{X\sim\mathcal{D}}\left(\exists \alpha \in A\backslash\overline{A} : \pi_\alpha\left(X\right) \geq \sigma_\alpha\right) \leq p$$

Alternative null hypotheses

- Further define:

$$\overline{\Omega}_0 = \bigcup_{\overline{A} \subseteq A} \Omega_0 \left(\overline{A}\right)$$

- Under any distribution from null hypothesis $\overline{\Omega}_0$, the probability of at least one *false* rejection is bounded by p if:

$$\forall \overline{A}, \forall \mathcal{D} \in \Omega_0 \left(\overline{A}\right) : P_{X \sim \mathcal{D}} \left(\exists \alpha \in A \backslash \overline{A} : \pi_\alpha \left(X\right) \geq \sigma_\alpha\right) \leq p$$

- Hence, if p is small and there are rejections α, then we must conclude that these are true rejections and belong to $\overline{A}$

Alternative null hypotheses

- Note: we never reject the null hypothesis $\overline{\Omega}_0$
- But: we would reject many $\Omega_0\left(\overline{A}\right)$ with too small sets $\overline{A}$
- Namely: we would reject all $\Omega_0\left(\overline{A}\right)$ for which

$$\exists \alpha \in A \backslash \overline{A} : \pi_\alpha\left(X\right) \geq \sigma_\alpha$$

Now, how to compute this upper bound p?

Bonferroni correction

Lemma (Bonferroni correction)

Suppose $A \subseteq \mathbb{N}$ is a countable index set for a class of pattern functions $\{\pi_\alpha : \alpha \in A\}$. Let $q_\alpha \in (0,1)$ such that

$$\sum_{\alpha \in A} q_\alpha \leq 1.$$

If we choose σ_α such that $\forall \mathcal{D} \in \Omega_0$:

$$P_{X \sim \mathcal{D}} \left(\pi_\alpha \left(X \right) \geq \sigma_\alpha \right) \leq p q_\alpha,$$

then $\forall \overline{A} \subseteq A, \forall \mathcal{D} \in \Omega_0 \left(\overline{A} \right)$:

$$P_{X \sim \mathcal{D}} \left(\exists \alpha \in A \backslash \overline{A} : \pi_\alpha \left(X \right) \geq \sigma_\alpha \right) \leq p.$$

Bonferroni correction

- The most well-known approach to bound the FWER
- Proof is a straightforward application of the union bound
- Correct for all Ω_0 – for arbitrary dependencies between genes
- Therefore very conservative :-(

Randomisation (permutation) testing

So far for the general introduction to multiple testing...
Now a general introduction to randomisation testing as a technique!

- Computation intensive method to perform hypothesis testing
- I Will first discuss single hypothesis testing
- Explicitly taking finiteness of random sample into account
- Then expand ideas to multiple hypothesis testing
- Everything relies on a transformation group G, such that

$$P_{X\sim\mathcal{D}}(X) = P_{X\sim\mathcal{D}}(g(X))$$

for $g \in G$ and $\mathcal{D} \in \Omega_0$

Randomisation (permutation) testing

Theorem (The p-value using randomisation testing)

- *Given a pattern function* π,
- *a null hypothesis* Ω_0 *invariant under* $g \in G$,
- *a fixed* $p \in (0, 1)$,
- *and the set* $\{g(X) : g \in G\}$.

Choose $\sigma(X)$ *such that*

$$p = \frac{\#\{g \in G : \pi(g(X)) \geq \sigma(X)\}}{|G|}$$

Then, $\forall \mathcal{D} \in \Omega_0$:

$$P_{X \sim \mathcal{D}}(\pi(X) \geq \sigma(X)) = p.$$

Subsampling the transformation group

- Usually, the number of group elements is prohibitively large
- Then a subsample G_m of size $|G_m| = m$ is used to estimate p:

$$\widehat{p} = \frac{\#\left\{g \in G_m : \pi\left(g\left(X\right)\right) \geq \sigma\left(X\right)\right\}}{m}$$

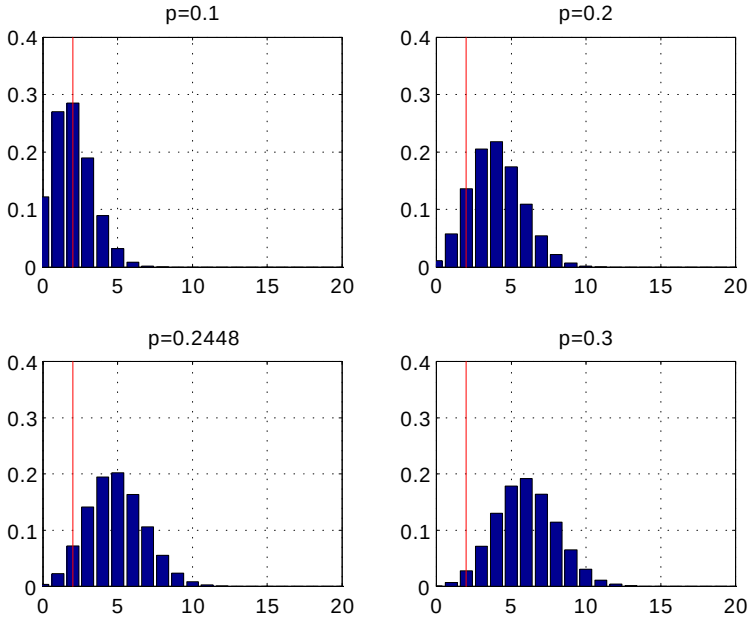
- It can be shown (Langford, see references) that with probability $\geq 1 - \delta$ over the random sample G_m,

$$p \leq \overline{\mathrm{Bin}}\left(m, \widehat{p}m, \delta\right)$$

where $\widehat{p} \leq \overline{\mathrm{Bin}}\left(m, \widehat{p}m, \delta\right) \leq \widehat{p} + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}$ approaches $\widehat{p}$ from above for increasing m

Subsampling the transformation group

- Upper bound the bias p (probabilistically) by random sampling
- E.g.: 20 coin tosses, 2 successes



- Upper bound $\overline{\mathrm{Bin}}\,(20, 2, 0.1)$: largest bias that gives probability ≥ 0.1 to see 2 or fewer successes
- Any larger value for p would imply that this few observations is unlikely

Subsampling the transformation group

Theorem (The p-value using randomisation testing)

- *Given a pattern function* π,
- *a null hypothesis* Ω_0 *invariant under* $g \in G$,
- *a fixed* $\widehat{p} \in (0,1)$,
- *and the set* $\{g(X) : g \in G_m\}$ *with* G_m *a random subset of* G.

Choose $\sigma(X)$ *such that*

$$\widehat{p} = \frac{\#\{g \in G_m : \pi(g(X)) \geq \sigma(X)\}}{m}$$

Then, with probability $\geq 1 - \delta$ *over* $G_m, \forall \mathcal{D} \in \Omega_0$:

$$P_{X \sim \mathcal{D}}(\pi(X) \geq \sigma(X)) \leq \overline{Bin}(m, \widehat{p}m, \delta).$$

The FWER

- Thus far:
  - Multiple testing framework
  - How to use transformation invariants of the null hypothesis and randomisation testing to perform single hypothesis testing
- Now: use of randomisation testing for multiple tests, bounding
  - FWER
  - k-FWER
  - k-FWER uniformly
  - FDP
- First the FWER: the probability of at least one false rejection

$$P_{X\sim\mathcal{D}}\left(\exists\alpha\in A\backslash\overline{A}:\pi_\alpha\left(X\right)\geq\sigma_\alpha\left(X\right)\right)$$

The FWER

Theorem (Bounding the FWER using randomisation testing)

- *Given a pattern class* Π,
- *a null hypothesis* Ω_0 *invariant under* $g \in G$,
- *a fixed* $\widehat{p} \in (0,1)$,
- *and the set* $\{g(X) : g \in G_m\}$ *with* G_m *a random subset of* G.

Choose $\sigma_\alpha(X)$ *such that*

$$\widehat{p} = \frac{\#\{g \in G_m : \exists \alpha : \pi_\alpha(g(X)) \geq \sigma_\alpha(X)\}}{m}$$

Then, with probability $\geq 1-\delta$ *over* $G_m, \forall \overline{A} \subseteq A, \forall \mathcal{D} \in \Omega_0(\overline{A})$:

$$P_{X\sim\mathcal{D}}\left(\exists \alpha \in A \backslash \overline{A} : \pi_\alpha(X) \geq \sigma_\alpha(X)\right) \leq \overline{Bin}(m, \widehat{p}m, \delta).$$

The k-FWER

- In practice, we are happy to tolerate more than one false rejection
- If only one is tolerated: $\widehat{p}$ is very large for any reasonable choice of $\sigma_\alpha(X)$
- Overly conservative – very few true rejections (many type II errors!)
- Remedy: upper bound probability to observe at most k false rejections (instead of at most 1): the k-FWER
- Instead of bounding

$$P_{X\sim\mathcal{D}}\left(\exists\alpha\in A\backslash\overline{A}:\pi_\alpha(X)\geq\sigma_\alpha(X)\right)$$

now bound

$$P_{X\sim\mathcal{D}}\left(\exists k \text{ different } \alpha\in A\backslash\overline{A}:\pi_\alpha(X)\geq\sigma_\alpha(X)\right).$$

The k-FWER

Theorem (Bounding the k-FWER using randomisation testing)

- *Given a pattern class* Π,
- *a null hypothesis* Ω_0 *invariant under* $g \in G$,
- *a fixed* $\widehat{p} \in (0,1)$, *and a fixed* k,
- *and the set* $\{g(X) : g \in G_m\}$ *with* G_m *a random subset of* G.

Choose $\sigma_\alpha(X)$ *such that*

$$\widehat{p} = \frac{\#\{g \in G_m : \exists k \text{ different } \alpha : \pi_\alpha(g(X)) \geq \sigma_\alpha(X)\}}{m}$$

Then, with probability $\geq 1-\delta$ *over* $G_m, \forall \overline{A} \subseteq A, \forall \mathcal{D} \in \Omega_0(\overline{A})$:

$$P_{X\sim\mathcal{D}}\left(\exists k \text{ different } \alpha \in A \backslash \overline{A} : \pi_\alpha(X) \geq \sigma_\alpha(X)\right) \leq \overline{Bin}(m, \widehat{p}m, \delta).$$

A uniform bound on the k-FWER

- In practice it may be hard to prespecify the k, the number of false rejections tolerated
- However, the previous bound holds for any fixed k, not uniformly for all k
- Can we obtain a uniform bound, for all $k \in \mathcal{K} \subseteq \{1, 2, ..., |A|\}$?
- Here the set $\mathcal{K}$ should be chosen so as to include all potentially interesting values

A uniform bound on the k-FWER

- Instead of considering

$$P_{X\sim\mathcal{D}}\left(\exists k \text{ different } \alpha \in A\backslash\overline{A} : \pi_\alpha(X) \geq \sigma_\alpha(X)\right)$$

now consider

$$P_{X\sim\mathcal{D}}\left(\exists k \in \mathcal{K} : \exists k \text{ different } \alpha \in A\backslash\overline{A} : \pi_\alpha(X) \geq \sigma_\alpha(X,k)\right)$$

- Note that we have different thresholds $\sigma_\alpha(X,k)$ for different k (we have to make this explicit now)
- For larger k, the threshold should be smaller (we tolerate more false rejections)
- Hence: the $\sigma_\alpha(X,k)$ should be non-increasing

A uniform bound on the k-FWER

Theorem (Bounding the k-FWER using randomisation testing)

- *Given a pattern class* Π,
- *a null hypothesis* Ω_0 *invariant under* $g \in G$,
- *a fixed* $\widehat{p} \in (0,1)$, *a set* $\mathcal{K} \subseteq \{1, 2, ..., |A|\}$,
- *and the set* $\{g(X) : g \in G_m\}$ *with* G_m *a random subset of* G.

Choose $\sigma_\alpha(X, k)$ *non-increasing in* k, *such that*

$$\widehat{p} = \frac{\#\left\{g \in G_m : \exists k \in \mathcal{K} : \exists k \text{ different } \alpha : \pi_\alpha(g(X)) \geq \sigma_\alpha(X, k)\right\}}{m}$$

Then, with probability $\geq 1 - \delta$ *over* $G_m, \forall \overline{A} \subseteq A, \forall \mathcal{D} \in \Omega_0(\overline{A})$:

$$\begin{aligned} & P_{X \sim \mathcal{D}}\left(\exists k \in \mathcal{K} : \exists k \text{ different } \alpha \in A \backslash \overline{A} : \pi_\alpha(X) \geq \sigma_\alpha(X, k)\right) \\ \leq \quad & \overline{Bin}(m, \widehat{p}m, \delta). \end{aligned}$$

The FDP

- We can easily check the total number of rejections for $\sigma_\alpha(X, k)$: $\#\text{rej}(k)$
- The uniform k-FWER provides an immediate upper bound on the number of false rejections: $k-1$
- $\Rightarrow$ a lower bound on the number of true rejections as: $\#\text{trej}(k) \geq \#\text{rej}(k)\text{-}(l-1)$
- Note however that $\#\text{trej}(k)$ cannot decrease with increasing k. Hence:

$$\underline{\#\text{trej}(k)} = \max_{l \leq k} \left(\#\text{rej}(l) - (l-1)\right)$$

- In particular, $\underline{\#\text{trej}(\max\{k \in \mathcal{K}\})}$ gives a lower bound on the total number of positives

The FDP

- We already have an upper bound on the false rejections of $k-1$
- But, we can get a possibly tighter one: $\overline{\#\text{frej}(k)}$=#rej($k$)-$\underline{\#\text{trej}(k)}$
- Based on these quantities, we can bound the False Discovery Proportion as:

$$\overline{\text{FDP}(k)} = \frac{\overline{\#\text{frej}(k)}}{\#\text{rej}(k)}$$

Technical details

- 2 microarray datasets: Alon and Golub
- Compare point-wise bound for the k-FWER with the uniform bound
- Do this for different choices of $\mathcal{K}$
- One technical issue remains: how to choose $\sigma_\alpha(X, \cdot)$?
- In principle, this is entirely free
- For convenience, we choose $\sigma_\alpha(X, \cdot) = \sigma(X, \cdot)$ independent of α

How to choose the thresholds?

- Compute $\{\pi_\alpha(g(X)) : g \in G_m \subseteq G, \alpha \in A\}$
- For each $g \in G_m$, sort $\{\pi_\alpha(g(X)) : \alpha \in A\}$ in decreasing ordering
- Then, pick the k'th largest numbers from $\{\pi_\alpha(g(X)) : \alpha \in A\}$ (there are m of them), and put this in a set $\mathcal{S}(X, k)$
- Interpretation: $S(X, k)$ contains an empirical estimate of the distribution of the k'th strongest correlation with the labels, under Ω_0
- Choose $\sigma^r(X, k) =$ the r'th largest value of $\mathcal{S}(X, k)$
- Result: non-increasing functions of k that vary roughly as the sorted $\{\pi_\alpha(g(X)) : \alpha \in A\}$

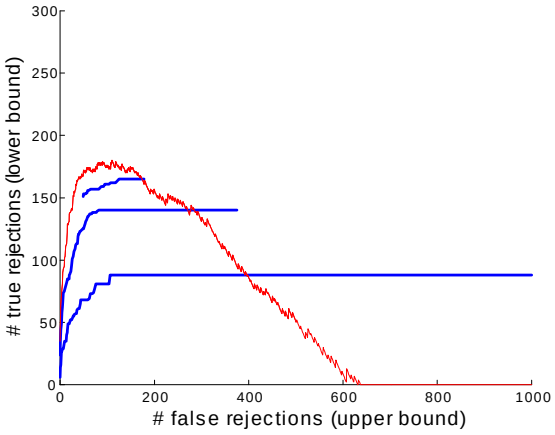
Alon and Golub datasets

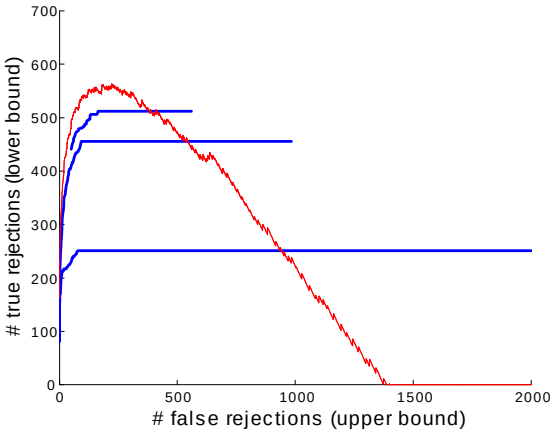
- Statistics of data sets

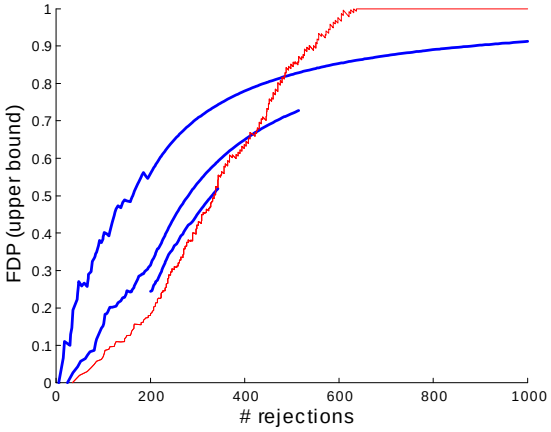
# genes	# arrays tissue 1	# arrays tissue 2
2000	40	22
7129	47	25

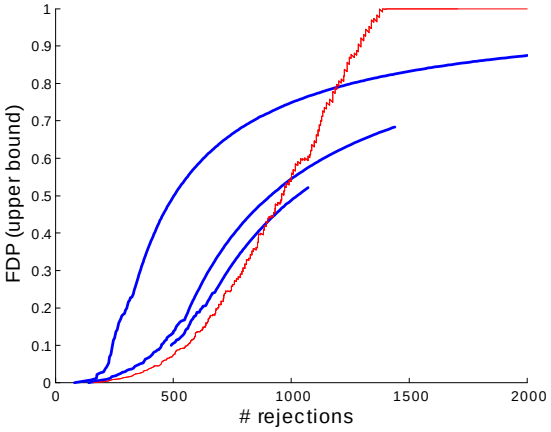
- Pattern functions used: the Wilcoxon rank sum test statistic

Alon and Golub datasets









Achievements

- Based on randomisation testing, which ensures:
  - More limited applicability (need to identify G)
  - But all dependencies are adequately taken into account
  - For that reason: practical relevance!
- Finite sample bound for randomisation testing
  - Separates randomness in g and X (δ and $\overline{\text{Bin}}\left(m, \widehat{p}m, \delta\right)$)
  - This is relevant when one sample g is used for several multiple tests
- Uniform bounds over regions of k in k-FWER bounds
  - Relevant where a priori unclear how large k should be taken
  - Allows to bound FDP