A sequential resampling procedure for multiple testing problems with bounded risk of classification errors

Georg Hahn, Dr Axel Gandy
Department of Mathematics

Imperial College London

ABSTRACT

This project introduces a sequential algorithm which assesses the statistical significance of multiple hypotheses using a procedure which controls the False Discovery Rate.

The algorithm doesn't observe all p-values directly, but approximates them by Monte-Carlo simulation. It is designed and proven to give, with arbitrary high probability, the same classification as the one based on the exact p-values.

FALSE DISCOVERY RATE (FDR)

The FDR controls the expected proportion of incorrectly rejected null hypotheses (false positives). It is given by E(V/R), where R is the total number of hypotheses which have been declared significant and V of which come from the null.

FDR CONTROL BY BENJAMINI-HOCHBERG

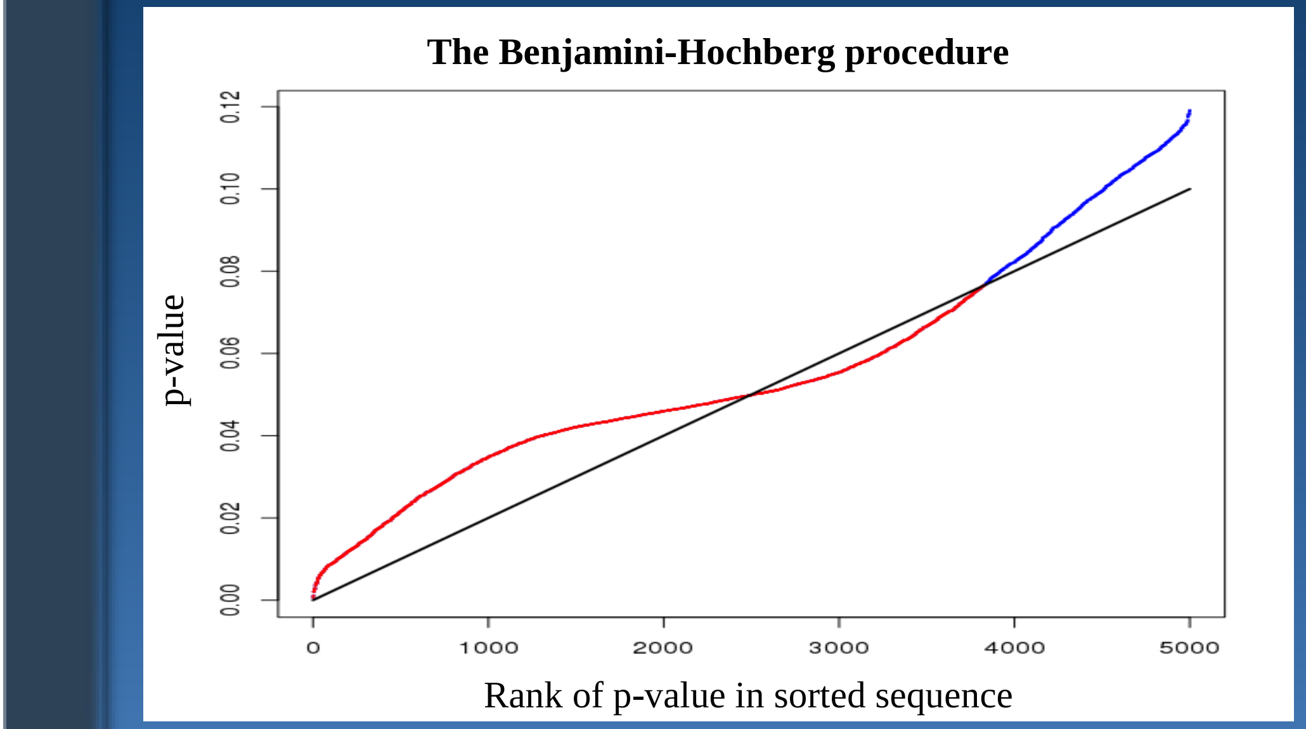

Figure: Rejecting all hypotheses up to the last one which satisfies $i*p_{(i)}/m≤α$ controls the FDR at threshold α (here α=0.1).

The method by Benjamini and Hochberg (1995) sorts all p-values first and then rejects all of them up to the last one which satisfies $i*p_{(i)}/m≤α$ (red part).

SETTING

Consider multiple hypotheses $H_{01},...,H_{0m}$ to be tested for statistical significance using a procedure which controls the False Discovery Rate at threshold α, e.g. the method by Benjamini and Hochberg (1995).

CHALLENGE

Instead of observing all p-values directly, consider the case where they can only be computed by simulation. For example this occurs when using bootstrap or permutation tests.

Naively, one could use an equal number of samples for the estimation of the p-value of each hypothesis. A more sophisticated approach is the one of Algorithm 1, where h shall denote an arbitrary function that controls the FDR at threshold α.

Algorithm 1

```
1. Set n:=0, $\underline{A}_0$:={} and $\overline{A}_0$:={1,...,m}.
2. While (|$\overline{A}_n$ \ $\underline{A}_n$| > c)
    (a) n:=n+1.
    (b) Compute confidence intervals $I_{n,i}$ ∀i=1...m.
    (c) Compute the sets $\underline{A}_n$:=h((max $I_{n,i})_{i=1,...,m}$)
        and $\overline{A}_n$:=h((min $I_{n,i})_{i=1,...,m}$).
3. Return the two sets ($\underline{A}_n$,$\overline{A}_n$).
```

ALGORITHMIC APPROACH

Algorithm 1 uses fewer samples for all those hypotheses which can already be classified with sufficient confidence and more samples for all those which are still unidentified.

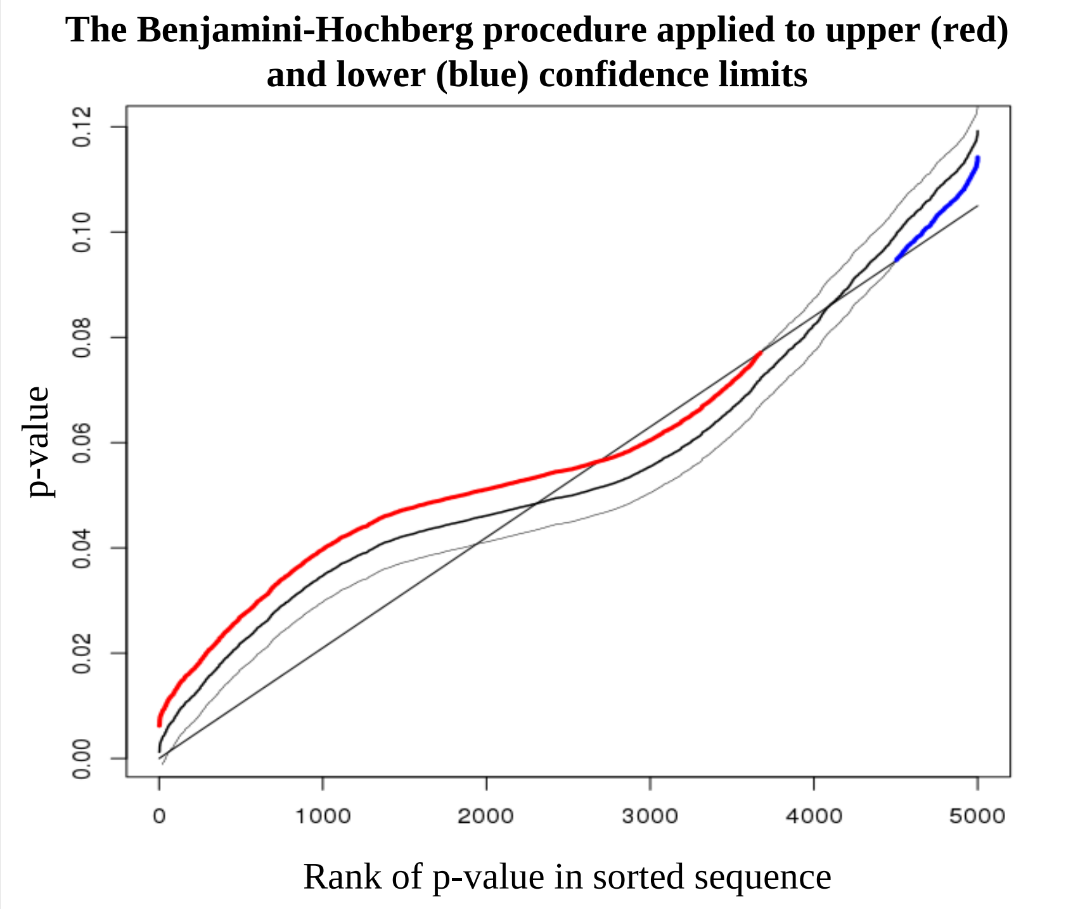

Figure: Benjamini-Hochberg with confidence limits

Assume that every p-value lies in its confidence interval. It can then be shown that if a p-value is rejected (non-rejected) according to its upper (lower) confidence limit, it will also be rejected (non-rejected) given its true p-value.

THEORETICAL RESULTS: RUNTIME

- The probability of classification errors is bounded above by ε ∀ε>0
  - The runtime needed to classify all hypotheses is infinite
    - The runtime is proven to be finite if all hypotheses are to be classified but the two closest ones to the Benjamini-Hochberg line

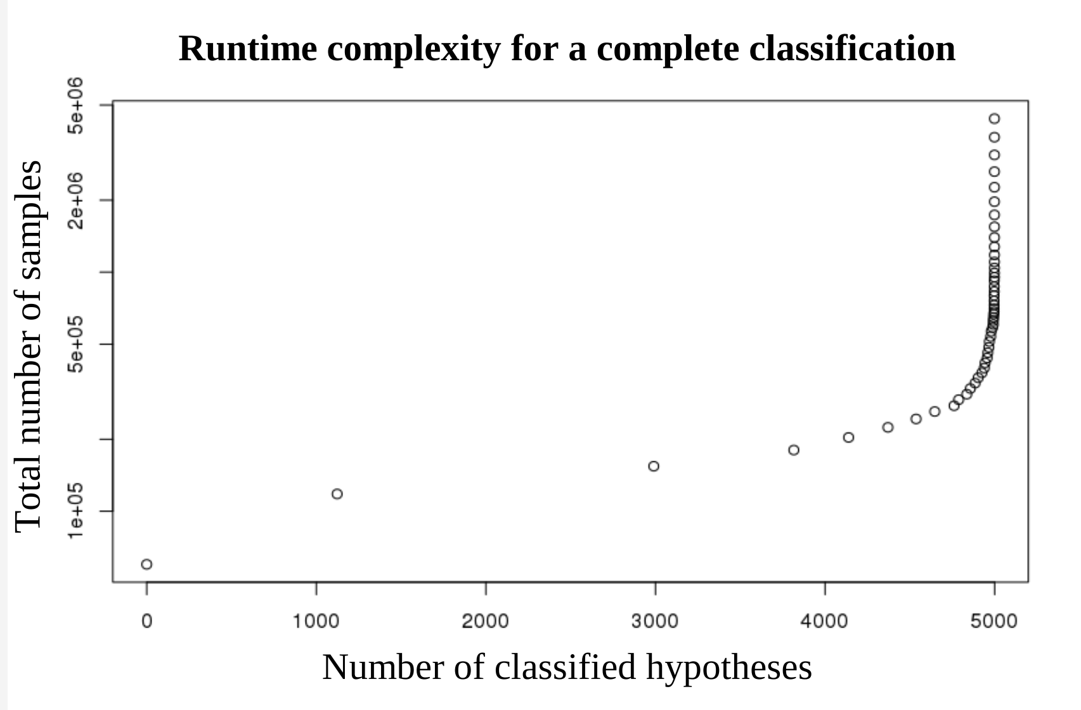

Figure: Runtime complexity to classify all m=5000 hypotheses

THEORETICAL RESULTS: CONVERGENCE

The algorithm is designed to give, with arbitrary high probability, the same classification as the one based on the exact p-values:

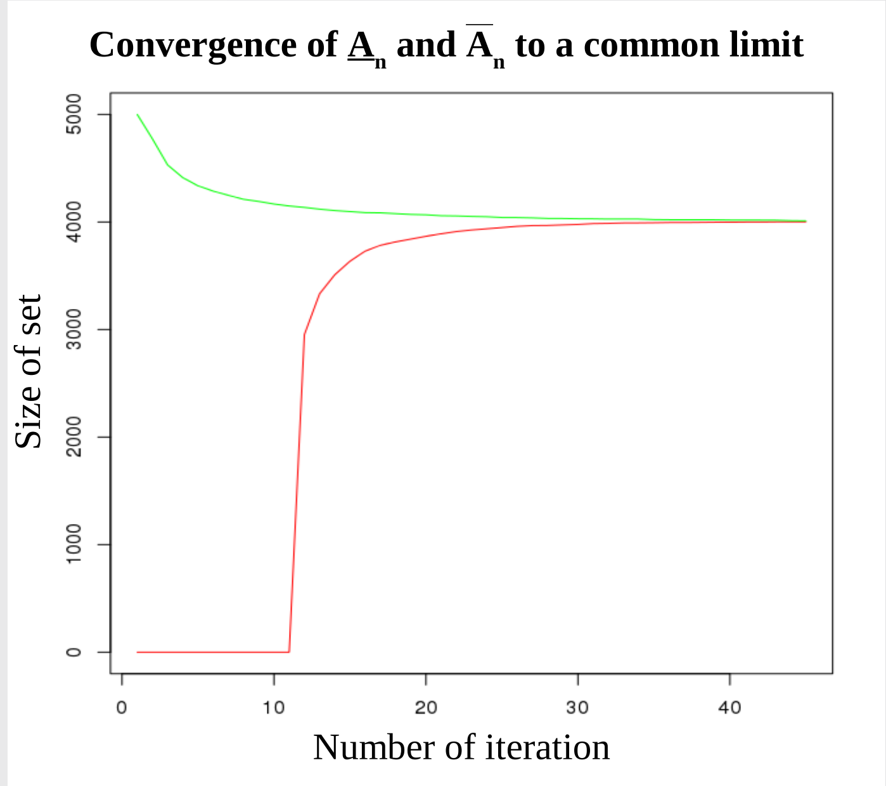

Figure: Convergence of $\underline{A}_n$ and $\overline{A}_n$ to A^{true}

It can be shown that the sequence of sets $\underline{A}_n$ ($\overline{A}_n$) is monotonically increasing (decreasing) and converges to A^{true}:=h(p_1,...,p_m), the set of rejections computed using exact p-values.

PERFORMANCE: PART I

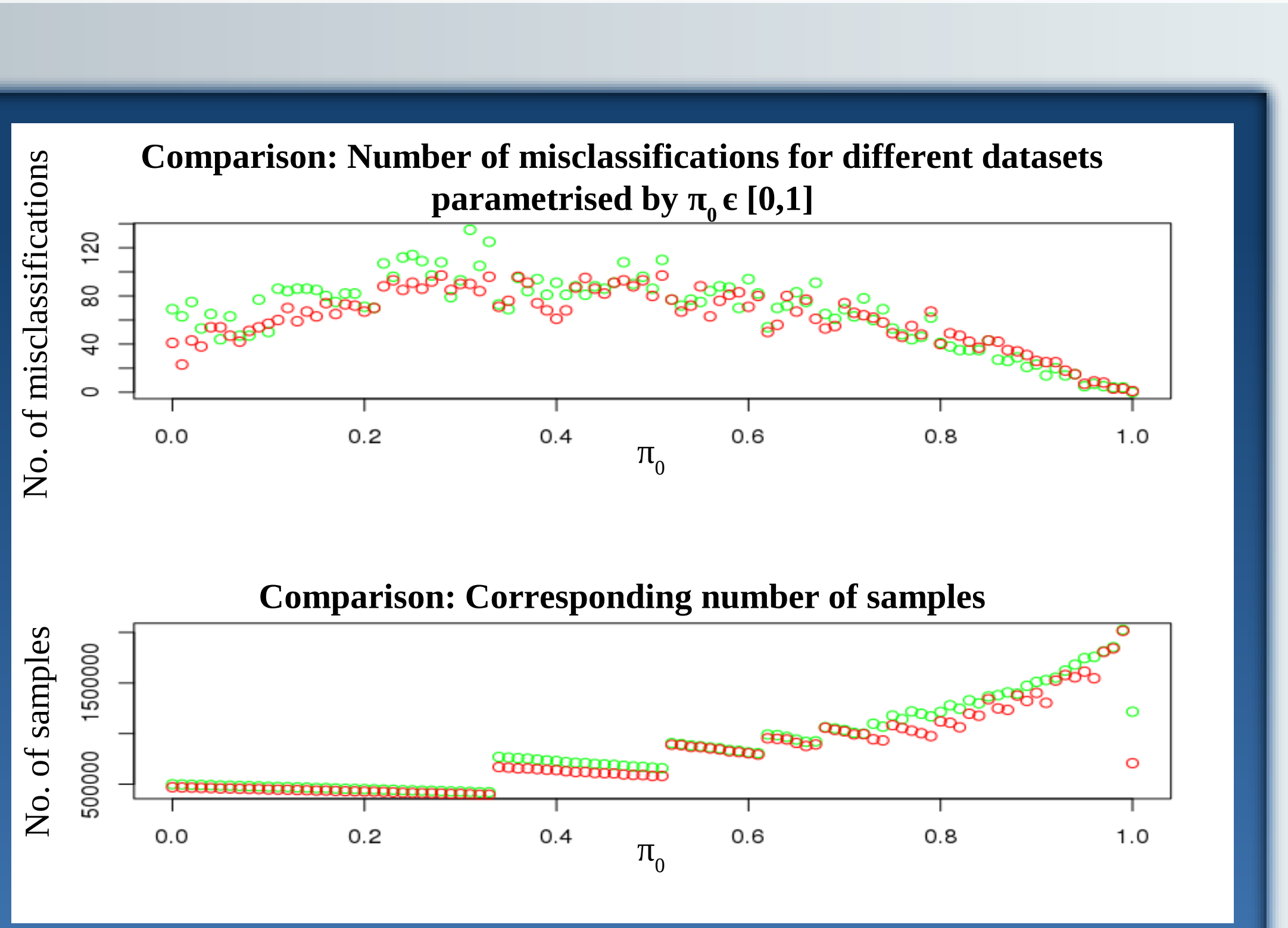

Figure: Comparison of MCFDR (in green) by Sandve et al. (2011) and Algorithm 1 (in red).

PERFORMANCE: PART II

The plot above shows a comparison of MCFDR (in green) by Sandve et al. (2011) and Algorithm 1 (in red) using test data from Sandve et al. (2011).

A total of m=5000 p-values were drawn from a probability distribution which depends on a parameter π_0 ϵ [0,1] and tested for significance using both algorithms. For each ensemble of m p-values, the total number of samples drawn and the total number of misclassifications were recorded.

The two plots show the total number of misclassifications, where the total number of samples drawn by Algorithm 1 has been adjusted to the number drawn by MCFDR.

As can be seen from the first plot, Algorithm 1 mostly outperforms MCFDR when using an equal number of samples (second plot).

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1):pp. 289-300.

Sandve, Ferkingstad and Nygard (2011). Sequential Monte Carlo multiple testing. Bioinformatics (Advance Access).

Felix Breuer. Making a Math Conference Poster with Inkscape. http://www.felixbreuer.net/ (Design of Poster)

This poster wouldn't be looking as cool as it does now without the help of Swati Chandna and Anna Fowler.