FDR doesn’t tell the whole story: Joint influence of effect size and covariance structure on the distribution of the false discovery proportions.

Alan H. Feiveson, Ph.D., James Feidler, Ph.D., Robert J. Ploutz-Snyder, Ph.D., PStat

NASA Johnson Space Center, Houston, TX

*Universities Space Research Association, Houston, TX

Abstract

As part of a 2005 Annals of Statistics paper, Gavrilov, Benjamini, and Sarkar report results of simulations that estimated the false discovery rate (FDR) for equally correlated test statistics using a well-known multiple test procedure. In our study we estimate the distribution of the false discovery proportion (FDP) for the same procedure under a variety of correlation structures among multiple dependent variables, using both simulated and real experiments. Specifically, we study the mean (the FDR), skewness, kurtosis, and percentiles of the distribution of the FDP in the context of multiple comparisons that give rise to correlated non-central t-statistics when several time periods are being compared to baseline. Even if the FDR achieves its nominal value, other aspects of the distribution of the FDP depend on the correlation between test statistics and on the variance and covariance structure, proportion of true nulls, and number of dependent variables. We show examples where the mean FDP (the FDR) is 10% as designed, yet there is a surprising probability of having 30% or more false discoveries. Thus, in a real experiment, the proportion of false discoveries could be quite different from the stipulated FDR.

Background and Significance

Gavrilov, Benjamini, and Sarkar (GBS) [1] discuss the pros and cons of multiple and family-wise error rates for controlling the FDR in a multiple-testing situation with a large number of variables which may be correlated. In particular, they prove that a simplified version of a family of multiple procedures suggested by Benjamini, Kingsley, and Yeakel (BKY) [2] does indeed control the FDR to a desired level α, when the test statistics are mutually independent. GBS then propose results of some simulations with equi-correlated and normally distributed test statistics to show that the FDR of this simplified BKY procedure is fairly robust under this dependence model.

False Discovery Proportion and Rate

As defined in [3], the adaptive step-down procedure based on m tests is as follows:

1. Let \( p_1, \ldots, p_m \) be the ordered p-values.
2. Define critical values as follows:
   \[ \alpha = \min \left\{ \frac{\xi}{1 + \xi} \mid \xi \geq 1 \right\}, \]

\[ \xi = \min \{ 1, \frac{n(1 - \alpha)}{\alpha} \}, \]

3. If \( \xi = 1 \), stop, no rejections with p-values \( p_i \leq \alpha \), otherwise repeat hypothesis rejection.

Simulated BKY Procedure

As defined in [3], the adaptive step-down procedure based on m tests is as follows:

1. Let \( p_1, \ldots, p_m \) be the ordered p-values.
2. Define critical values as follows:
   \[ \alpha = \min \left\{ \frac{\xi}{1 + \xi} \mid \xi \geq 1 \right\}, \]

\[ \xi = \min \{ 1, \frac{n(1 - \alpha)}{\alpha} \}, \]

3. If \( \xi = 1 \), stop, no rejections with p-values \( p_i \leq \alpha \), otherwise repeat hypothesis rejection.

Simulated Data Model

The FDR is 0.05 for each value of m, and the test statistics are mutually independent. The empirical distribution of the FDP is fairly robust under this dependence model.

Study Summary

- 200, 400, 800, 1600, 3200 variables
- 1000 simulations per study per simulation run
- dependence structures
  - variables and tests are completely independent
  - P1 - P5: variables are independent, but dependent multiple comparisons as a result of repeated measures design
  - D: general covariance structure between variables (weighted sum of AR1, constant correlation, and two-stage Whishart) arising from multiple comparisons
- FDR attributes studied: mean FDR, median, IQR, skewness, kurtosis

As defined in [4], the adaptive step-down procedure based on m tests is as follows:

1. Let \( p_1, \ldots, p_m \) be the ordered p-values.
2. Define critical values as follows:
   \[ \alpha = \min \left\{ \frac{\xi}{1 + \xi} \mid \xi \geq 1 \right\}, \]

\[ \xi = \min \{ 1, \frac{n(1 - \alpha)}{\alpha} \}, \]

3. If \( \xi = 1 \), stop, no rejections with p-values \( p_i \leq \alpha \), otherwise repeat hypothesis rejection.

Simplified BKY Procedure

As defined in [3], the adaptive step-down procedure based on m tests is as follows:

1. Let \( p_1, \ldots, p_m \) be the ordered p-values.
2. Define critical values as follows:
   \[ \alpha = \min \left\{ \frac{\xi}{1 + \xi} \mid \xi \geq 1 \right\}, \]

\[ \xi = \min \{ 1, \frac{n(1 - \alpha)}{\alpha} \}, \]

3. If \( \xi = 1 \), stop, no rejections with p-values \( p_i \leq \alpha \), otherwise repeat hypothesis rejection.

Simulated Data Model

The FDR is 0.05 for each value of m, and the test statistics are mutually independent. The empirical distribution of the FDP is fairly robust under this dependence model.

Example of FDP Distributions by Dependence

The best FDP distributions tend to arise when \( m \) is small (little or no dependence) and especially when there is a small proportion of nulls. Here, the distribution is more symmetrical about the nominal FDR of 0.05, and also has relatively little spread (small FDR).

Conclusions

1. When effect sizes and the proportion of nulls are both small, the actual FDR can be a lot smaller than its nominal value (0).
2. Even when the FDR is close to its nominal value, the FDP distribution can have extreme skewness in the right tail (at low and high FDR).
3. For fixed q, depending on \( m \) and \( \alpha \), skewness and other characteristics of the FDP distribution can be strongly associated with the degree of dependence between test statistics.
4. Examining the effect of correlated dependent variables on functions of test statistics, one can estimate that the correlation structure of the FDP can be strongly influenced by the covariance matrices of the correlated and dependent variables.
5. Consider controlling the FWE error rate (probability of 2 or more rejections when \( H_0 \) is true) in the presence of moderate to extreme dependence, especially when one expects a large proportion of non-null cases, but with relatively small effect sizes.

References