Loui Wentzel (vln585) and Christian Starup (plf981)

Introduction:

The article "Combining dependent P-values with an empirical adaptation of Brown's method" by William Poole, David L. Gibbs, Ilya Shmulevich, Brady Bernard and Theo A. Knijnenburg published in Bioinformatics, Volume 32, Issue 17, 1 September 2016, Pages i430–i436' was chosen due to our resent experience with the method and the potential relevance it could have for other students.

Problem/Motivation:

Given a dataset where one needs to calculate several or many p-values, should one account for a possible correlation between data, and in case, how is this done?

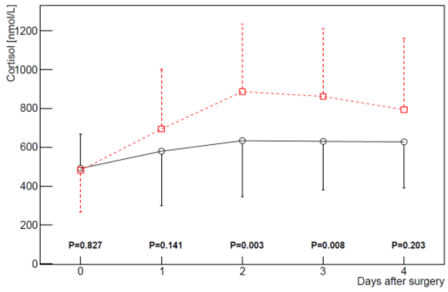


Figure 1 shows a figure from our Bachelor thesis, describing the hormone Cortisol in patients undergoing colorectal surgery as a function of days. The patients were divided into those developing cardiopulmonary complications and those who did not. We wanted to answer the question 'What is joint probability that there is a difference among the groups given the data is correlated across days?'

Correlation solution:

If the P-values are not correlated, then each should follow a uniform distribution; However if the data is correlated, we can't assume a uniform distribution of p-values; thus, a new method is needed. Brown realised this and altered Fisher's method to include a rescaling factor, c, such that $\Psi \sim c\chi^2_{2f}$. Here f is the DoF of the Brown's method, k is the DoF from Fisher's method and c is calculated as $c = \frac{k}{f}$. Brown showed that these can be calculated by:

$$f = \frac{E[\Psi]^2}{\text{var}[\Psi]}, \qquad c = \frac{\text{var}[\Psi]}{2E[\Psi]}, \qquad E[\Psi] = 2k, \qquad \text{var}[\Psi] = 4k + 2\sum_{i<j} \text{cov}(W_i, W_j)$$

Here, $W_i = -2\log P_i$. The articles contribution to this result is to evaulte the covariance matrix emperically (hence, Empirical Brown's Method, EBM). An additional method to Fisher's and Brown's is also covered in the article, created by Kost and McDermott, known as Kost method. This method approximates the covariance to a third order polynomial:

$$\text{cov}(W_i, W_j) \approx 3.263\rho_{ij} + 0.710\rho_{ij}^2 + 0.027\rho_{ij}^3$$

Where ρ_{ij} is the correlation between the data variables X_i and X_j.

The paper takes a non-parametric approch to approximate the covariance matrix in Browns' method:

$$\text{cov}(W_i, W_j) \approx E[(\vec{w_i} - E[\vec{w_i}])(\vec{w_j} - E[\vec{w_j}])], \qquad w_i = -2\log(1 - F(\vec{x_i}))$$

Where $\vec{x_i}$ is a vector of i'th data points across all correlated measurements X_i. In the example given, it would be the i'th patients level of Cortisol across all 5 days, and be a 5 by 1 vector. $F(\vec{x_i})$ is then the empirical cumulative distribution function to be evaluated individually for each day, that is acting elementwise onto the vector $\vec{x_i}$. The combined P-value is then given by:

$$P_{combined} = 1 - \Phi_{2f}(\Psi/c)$$

With Φ_{2f} being the CDF of χ^2_{2f}, $\Psi = \sum W_i = \sum -2\log P_i$, and c and f as defined before.

Simulating data:

To investigate the appropiateness of the three different tests, the writers simulated data from a multivariate normal distribution with correlation matrix

$$\Sigma = \begin{bmatrix} 1 & b_2 & \cdots & b_j & \cdots & b_n \\ b_2 & 1 & \cdots & a & \cdots & a \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ b_j & a & \cdots & 1 & \cdots & a \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ b_n & a & \cdots & a & \cdots & 1 \end{bmatrix}$$

And all means 0. Setting $a = 0.8$ and then sampling 200 points yielded highly correlated data among axes 2 through n, and when the b_j's were randomly drawn from a uniform distribution in $[-0.5; 0.5]$ it allowed to test axis 1 for correlations with the other axes. This data would not be in accordance with the null hypothesis, as the goal was to produce a test that could correctly account for the correlation to give the same results as data in accordance with the null. Uniform noise could be added to each axis, and the size is controlled by a parameter ξ:

$$\bar{x} = \bar{y} + \xi \bar{U}$$

Setting $\xi = 0$ amounted to noiseless data and a signal to noise ratio (SNR) of ∞, while setting $\xi = 2.5$ yielded SNR= 0.5.
To generate the P-values, the first axis was tested for correlations against the other axes using Pearsons correlation test.
Null-hypothesis data could be simulated by setting $b_j = 0$, and doing so should yield a uniform distribution of P-values for a perfect test, but instead yielded as shown in the graph below:

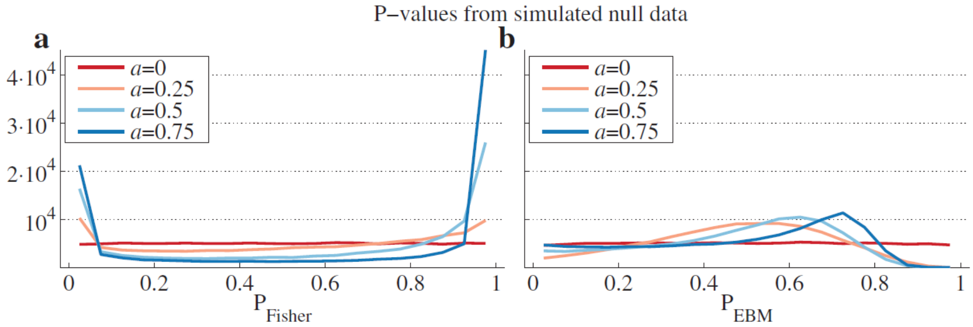


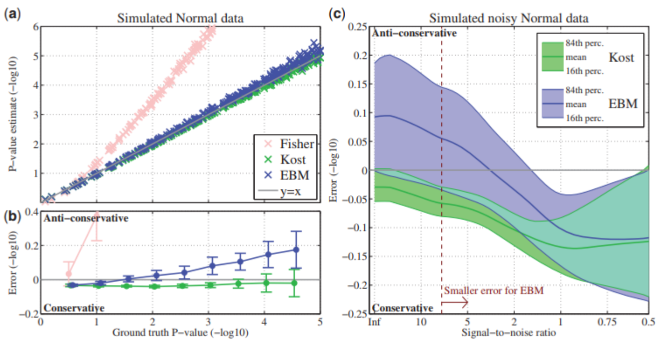
Fig. 2. *P*-values from simulated data using Fisher's method and EBM. (a) Line plot of histogram counts of *P*-values from Fisher's method applied on simulated null data with varying degrees of covariance as represented by *a*. The histogram was created by binning the *P*-values in 20 bins of size 0.05 from 0 to 1. (b) Similar to (a) but for *P*-values derived with the Empirical Brown's method

It is clear that for correlated data Fishers method is highly unqualified to produce a uniform distribution. The Empirical Browns Method yields better results, but they are not perfect.
Ground truth P-values:

To simulate the null hypothesis and test it against the already performed tests the first axis was shuffled (randomly permuted), giving no correlation with the other axes while retaining the internal correlations of the other axes. Then the test statistics were again calculated, and the suffling repeated to ensure no "unlucky" shuffling. This was repeated M times, and the resulting P-value is

$$P_{perm} = \frac{\sum_{m=1}^{M} \Theta(\Psi_m^* \geq \Psi)}{M}$$

If one has 10 or more exceedances where $\Psi_m^* \geq \Psi$, then one can use the central limit theorem to show that P_{perm} is an accurate P-value estimate.
For a correlation test able to correctly account for the correlations among the data, it was expected that the correlated tests gave the same result for the permuted P-values as for the non-permuted P-values. The permuted P-values were called the Ground Truth P-values, and plotted against the non-permuted P-values, it was expected for a correct test to follow the line $y = x$.
Performance results:



Fishers, Kost's and the EMB p-values can be seen plotted as a function of the ground truth value in fig a. where we can see that the Kost and EMB nicely follows the line y=x as just described. Fisher's however, is way more anti-conservative and estimates far lower p-values. Looking closer, one can conclude that Kost outperforms EMB for noiseless data, as it follows the line y=x more precisely and has a smaller error. However, when the SNR is reduced from infinity to lower values, the mean error from $y = x$ is smaller for EBM from a SNR of about 8 and below.

The message is that correlated P-values should be considered in the data analysis, and if one expects a clean result with SNR above 8, one should use Kost's method for combining dependent P-values with a polynomial description of the covariance matrix. But if the signal is noisy so that SNR<8, one should use the Empirical Browns Method and draw the sample w_i, from which correlations can be found.