Advanced Methods in Applied Statistics

Christian Starup & Loui Wentzel

Niels Bohr Institute

March 8, 2018

Journal Article

Combining dependent *P*-values with an empirical adaptation of Brown's method

William Poole, David L. Gibbs, Ilya Shmulevich, Brady Bernard† and Theo A. Knijnenburg*,†

Institute for Systems Biology, Seattle, WA 98109-5263, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

Abstract

Motivation: Combining *P*-values from multiple statistical tests is a common exercise in bioinformatics. However, this procedure is non-trivial for dependent *P*-values. Here, we discuss an empirical adaptation of Brown's method (an extension of Fisher's method) for combining dependent *P*-values which is appropriate for the large and correlated datasets found in high-throughput biology.

Results: We show that the Empirical Brown's method (EBM) outperforms Fisher's method as well as alternative approaches for combining dependent *P*-values using both noisy simulated data and gene expression data from The Cancer Genome Atlas.

Availability and Implementation: The Empirical Brown's method is available in Python, R, and MATLAB and can be obtained from https://github.com/IlyaLab/CombiningDependentPvaluesUsingEBM. The R code is also available as a Bioconductor package from https://www.bioconductor.org/packages/devel/bioc/html/EmpiricalBrownsMethod.html.

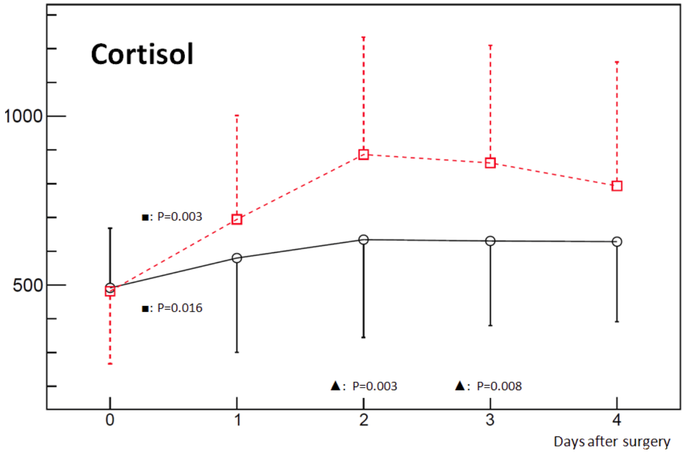
Contact: Theo.Knijnenburg@systemsbiology.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

https://doi.org/10.1093/bioinformatics/btw438

Problem

Given a dataset where one needs to calculate several or many p-values. Should one account for a possible correlation between data variables?



No Correlation solution

If the P-values are not correlated, then according to H_0 the distribution of each P-value should be uniform, and the product of P-values should then be drawn from the distribution of N products of uniform numbers:

$$P = \int_0^{\prod} \frac{(-1)^{N-1}}{(N-1)!} \cdot \ln(u)^{N-1} du \tag{1}$$

This is equivalent to a χ^2-test with $2k$ degrees of freedom called Fishers Method:

$$\Psi = \sum_{i=1}^{N} -2\log(P_i) \tag{2}$$

$$P = \phi_{2k}(\Psi) = \int_{\Psi}^{\infty} \chi^2_{2k}(x)dx \tag{3}$$

Correlation solution

However, if the data is correlated, we can't assume a uniform distribution of P-values.
Brown therefore expanded Fisher's method to include a re-scaling factor, c, such that $\Psi \sim c\chi^2_{2f}$.

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

With $W_i = -2\log(P_i)$, $E[\Psi] = 2k$ (assuming a χ^2 distribution), k is the Fisher's DoF and f the re-scaled Brown's DoF.
The combined P-value is then:

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

with $\Psi = \sum W_i$, Φ_{2k} being the cumulative distribution function of χ^2_{2f}.

Correlation solution continued

The articles contribution to Browns' method is to calculate the covariance matrix by an empirical approximation, thereby the Empirical Brown's method (EBM):

$$cov(W_i, W_j) \approx cov(w_i, w_j)$$
$$w_i = -2\log(1 - F(\overrightarrow{x_i}))$$

Kost's method uses another approach to calculate the covariance:

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

The EBM is a non-parametric approach, where $F(\overrightarrow{x_i})$ is the right-sided empirical cumulative distribution function.

Simulating data

Parameters were $\mu_i = 0$, $a = 0.8$, $n = 4$. b_j was randomly sampled from $[-0.5; 0.5]$. Each sample had 200 entries.

$$M = \begin{bmatrix} 1 & b_2 & \dots & b_j & \dots & b_n \\ b_2 & 1 & \dots & a & \dots & a \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ b_j & a & \dots & 1 & \dots & a \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ b_n & a & \dots & a & \dots & 1 \end{bmatrix} \tag{4}$$

From any sample $\vec{y}$ drawn from this distribution, n-dimensional uniform noise from $[-1; 1]$ was added:

$$\vec{x} = \vec{y} + \xi \vec{U} \tag{5}$$

They draw numbers from one axis on the multivariate normal distribution (axis 1 with correlations b_j to the others) and test the correlation to the other axes using Pearsons correlation test.

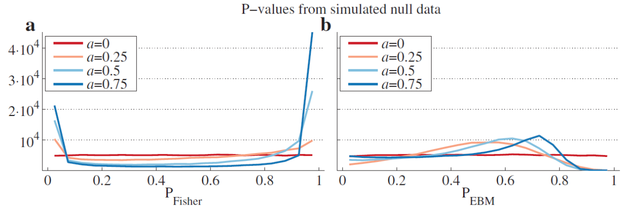


Fig. 2. *P*-values from simulated data using Fisher's method and EBM. (a) Line plot of histogram counts of *P*-values from Fisher's method applied on simulated null data with varying degrees of covariance as represented by *a*. The histogram was created by binning the *P*-values in 20 bins of size 0.05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

Ground Truth P-values

To test the different tests against correlated data, it should yield the same results as if the data was uncorrelated.

- Shuffle $\vec{y}_1$
- Calculate Ψ^* as earlier
- Repeat M times

The ground truth P-value is then

$$P_{ground} = \frac{\sum_{m=1}^{M} I(\Psi_m^* \geq \Psi)}{M} \tag{6}$$

Notice this gives a resolution in the ground truth P-value by $1/M$.

Performance results as a function of Signal to Noise ratio