FEBS Letters 584 (2010) 4481–4484

FEBS Letters

journal homepage: www.FEBSLetters.org

The rank product method with two samples

James A. Koziol *

Department of Molecular and Experimental Medicine, The Scripps Research Institute, MEM216, 10550 N Torrey Pines Rd, La Jolla, CA 92037, USA

ARTICLE INFO

Article history:
Received 2 August 2010
Revised 4 October 2010
Accepted 8 October 2010
Available online 14 October 2010

Edited by Takashi Gojobori

Keywords:
Rank product method
Rank statistic
Two sample comparison

ABSTRACT

Breitling et al. (2004) [1] introduced a statistical technique, the rank product method, for detecting differentially regulated genes in replicated microarray experiments. The technique has achieved widespread acceptance and is now used more broadly, in such diverse fields as RNAi analysis, proteomics, and machine learning. In this note, we extend the rank product method to the two sample setting, provide distribution theory attending the rank product method in this setting, and give numerical details for implementing the method.

1. Introduction

In an influential paper, Breitling et al. [1] introduced a statistical technique for detecting differentially regulated genes in replicated microarray experiments. Their rank product method entails ranking expression levels within each replicate, then computing the product of the ranks across the replicates. The rank product is then compared to its sampling distribution under a permutation model for subsequent inference. The rank product method appears to be robust, with higher sensitivity and specificity than *t*-test methodology and desirable operating characteristics, as demonstrated in extensive numerical studies [2–5]. Although developed originally for microarrays, the rank product method has found widespread acceptance in diverse settings, e.g., RNAi analysis [6], proteomics [7], and machine learning model selection [8].

Koziol [9] recently described a simple method for establishing distribution theory for the original rank product statistic: first, by invoking a log transformation, a linear rank statistic is shown to be equivalent to the rank product statistic; then, distribution theory for the linear rank statistic is available through a remarkably simple and accurate approximation involving the gamma distribution.

The purpose of this note is to consider extensions of the rank product method to two sample settings. As with the original one sample version of the rank product statistic, log transformation may be invoked to an equivalent linear rank statistic formulation. Subsequent distribution theory involves weighted linear combinations of gamma random variables; numerical procedures for determining these distributions are readily available for practical use.

We outline the two sample problem with the rank product method as well as the linear rank statistic formulation in Section 2, and consider numerical issues in Section 3. We provide an example in Section 4, and conclude with some remarks in Section 5.

2. The rank product statistic

2.1. The two sample version

We briefly describe the rank product statistic in the two sample setting. We start with expression levels for n genes from k_1 independent replicates in sample 1, and k_2 independent replicates in sample 2. Denote the expression level for the ith gene in the jth replicate of the mth sample by X_{ijm}, where $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant k_m$, $1 \leqslant m \leqslant 2$. Next, rank the expression levels $X_{1jm}, X_{2jm}, \ldots, X_{njm}$ within each replicate j, forming $R_{ijm} = \text{rank}(X_{ijm})$, $1 \leqslant R_{ijm} \leqslant n$, and $1 \leqslant m \leqslant 2$. Then, a suitable two sample version of Breitling's rank product statistic for the ith gene is, up to a normalization constant, the product

$$RP_i = \left(\prod_{j=1}^{k_1} R_{ij1}\right)^{1/k_1} \div \left(\prod_{j=1}^{k_1} R_{ij2}\right)^{1/k_2}. \tag{2.1}$$

RP_i is the geometric mean of the ranks of the ith gene from sample 1, divided by the geometric mean of the ranks of the ith gene from sample 2. Differential rankings of the ith gene in the two samples would lead to excessively large or small values of RP_i, hence genes associated with sufficiently small or large RP values would be marked for further consideration. One could modify the approach

* Fax: +1 858 784 2664.
E-mail address: koziol@scripps.edu

0014-5793/$36.00 © 2010 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2010.10.012

of Breitling et al. [1], to devise a permutation approach for the distribution of the RP_i under the null hypothesis that the X_{ijm} are identically distributed (exchangeable) within each of the k_i independent replicates of each sample. A more direct method of obtaining the null distribution is available, as we now describe.

2.2. An alternative formulation

An equivalent statistic to RP_i is the monotone transformation

$$\log(RP_i) = (1/k_1)\sum_{j=1}^{k_1} \log(R_{ij1}) - (1/k_2)\sum_{j=1}^{k_2} \log(R_{ij2}). \tag{2.2}$$

Monotonicity ensures that achieved significance levels of RP_i and $\log(RP_i)$ are identical. There are two key notions reflected in this transformation. First, since the k_m replicates are independent, each component of $\log(RP_i)$ is the average of k independent, identically distributed random variables under the null hypothesis. More fundamentally, the log transformation demonstrates that the rank product method engenders replacement of the ranks R_{ij} within each replicate by rank scores $a_j(R_{ij})$, where the score function here is given simply by $a_j(i) = \log(i)$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant k$. Although we note that much of the richness and diversity of linear rank statistics arises from adoption of different score functions into the underlying construct, we will not pursue this notion here, but will restrict attention to the log score function.

2.3. A simple approximation

Following Koziol [9], there exists a simple and remarkably accurate approximation to the null distribution of $\log(RP_i)$. First, note that $R_{ijm}/(n+1)$ is approximately uniformly distributed on the unit interval (0,1), the approximation improving as n (the number of genes) increases. Next, let U_j denote a uniform random variable [that is, U_j is uniformly distributed on (0,1)]. Then $-\log(U_j)$ has an exponential distribution on the positive real line with scale parameter 1, commonly denoted Exp(1). The key here is that the Exp(1) distribution is a particular case of a gamma distribution, namely, a $\Gamma(1,1)$ distribution. [The gamma distribution is a two-parameter continuous probability distribution. The two parameters are commonly referred to as the shape parameter k and the scale parameter θ, and the distribution is denoted $\Gamma(k,\theta)$.] The sum of independent, identically distributed exponentials is also gamma distributed, with the same scale parameter, but an altered shape parameter: in our setting, $-\sum_{j=1}^{k}\log(U_j)$ has a $\Gamma(k,1)$ distribution.

Next, note that an equivalent representation of $\log(RP_i)$ is given by

$$\log(RP_i) = (1/k_1)\sum_{j=1}^{k_1} \log\left(\frac{R_{ij1}}{n+1}\right) - (1/k_2)\sum_{j=1}^{k_2} \log\left(\frac{R_{ij2}}{n+1}\right). \tag{2.3}$$

It follows that the null distribution of $\log(RP_i)$is approximately distributed as

$$Y = -(1/k_1) * Y_1 + (1/k_2) * Y_2, \tag{2.4}$$

where the Y_m are independent random variables with $\Gamma(k_m, 1)$ distributions, $m = 1, 2$.

How does all this relate to the rank product? We have the following steps:

$$\text{Prob}(RP_i \leqslant t) = \text{Prob}(\log(RP_i) \leqslant \log(t)) \approx \text{Prob}(Y \leqslant \log(t)).$$

Hence we may determine approximate critical values for RP_i by back transformation from the associated critical values for Y. We turn to numerical calculation of the distribution of Y in the next section.

3. Distribution of the rank product statistic

3.1. Exact distribution

We will use inversion of the characteristic function for determination of critical values of the log rank product statistic. Recall that, if F is a one-dimensional distribution function, its characteristic function φ is the complex function of the real variable t:

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} dF(x).$$

In particular, the characteristic function of a $\Gamma(k,\theta)$ distribution is simply

$$\varphi(t) = (1 - it\theta)^{-k}.$$

It follows that the characteristic function of $Y = -(1/k_1)*Y_1 + (1/k_2)*Y_2$ from (2.4) is given by

$$\varphi_Y(t) = E[e^{itY}] = \left(1 + \frac{it}{k_1}\right)^{-k_1}\left(1 + \frac{it}{k_2}\right)^{-k_2}. \tag{3.1}$$

Gil-Pelaez [10] introduced a simple inversion formula, showing how the univariate cumulative distribution function F can be obtained by numerical inversion of its characteristic function φ:

$$\begin{aligned} F(x) &= \frac{1}{2} + \frac{1}{2\pi}\int_0^{\infty} \frac{e^{itx}\varphi(-t) - e^{-itx}\varphi(t)}{it} dt \\ &= \frac{1}{2} - \frac{1}{\pi}\int_0^{\infty} \text{Im}\left(\frac{e^{-itx}\varphi(t)}{t}\right) dt, \end{aligned} \tag{3.2}$$

Table 1
Probability that the quadratic form Q exceeds x. Exact exceedance probabilities are from Imhof (1961) [11]; approximate exceedance probabilities are from inversion of the characteristic function, and numerical integration, in Mathematica v.7.

Q	x	Exact	Approximate
$Q_1 = 0.6X_1^2 + 0.3X_1^2 + 0.1X_1^2$	0.1	0.9458	0.94578
	0.7	0.5064	0.50635
	2	0.1240	0.12403
$Q_2 = 0.6X_2^2 + 0.3X_2^2 + 0.1X_2^2$	0.2	0.9936	0.99354
	2	0.3998	0.39980
	6	0.0161	0.01610
$Q_3 = 0.6X_6^2 + 0.3X_4^2 + 0.1X_2^2$	1	0.9973	0.99732
	5	0.4353	0.43525
	12	0.0088	0.00877
$Q_4 = 0.6X_2^2 + 0.3X_4^2 + 0.1X_6^2$	1	0.9666	0.96664
	3	0.4196	0.41956
	8	0.0087	0.00872
$Q_5 = 0.7X_{6;6}^2 + 0.3X_{2;2}^2$	2	0.9939	0.99388
	10	0.4087	0.40866
	20	0.0221	0.02208
$Q_6 = 0.7X_{1;6}^2 + 0.3X_{1;2}^2$	1	0.9549	0.95487
	6	0.4076	0.40758
	15	0.0223	0.02232
$(1/3)Q_3 + (2/3)Q_4$	1.5	0.9891	0.98906
	4	0.3453	0.34527
	7	0.0154	0.01540
$(1/3)Q_3 - (2/3)Q_4$	−2	0.9102	0.91023
	0	0.4061	0.40611
	2.5	0.0097	0.00976
$(1/2)Q_5 + (1/2)Q_6$	3.5	0.9563	0.95632
	8	0.4152	0.41524
	13	0.0462	0.04623
$(1/2)Q_5 - (1/2)Q_6$	−2	0.9218	0.92179
	2	0.4779	0.47789
	7	0.0396	0.03963
$(1/4)(Q_3 + Q_4 + Q_5 + Q_6)$	3	0.9842	0.98416
	6	0.4264	0.42638
	10	0.0117	0.01166
$(1/6)(Q_3 - Q_5) + (2/6)(Q_6 - Q_4)$	−3	0.9861	0.98614
	0	0.5170	0.51702
	4	0.0152	0.01520