Published in final edited form as:

FEBS Lett. 2010 November 5; 584(21): 4481–4484. doi:10.1016/j.febslet.2010.10.012.

The rank product method with two samples

James A. Koziol, Ph.D.

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

Abstract

Breitling et al. (2004) 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.

Keywords

Rank product method; rank statistics; two sample comparisons

1. Introduction

In an influential paper, Breitling and colleagues [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 i^{th} gene in the j^{th} replicate of the m^{th} sample by X_{ijm} , where $1 \leq i \leq n$, $1 \leq j \leq k_m$, $1 \leq m \leq 2$. Next, rank the expression levels $X_{1jm}, X_{2jm}, \dots, X_{njm}$ within each replicate j , forming $R_{ijm} = \text{rank}(X_{ijm})$, $1 \leq R_{ijm} \leq n$, and $1 \leq m \leq 2$. Then, a suitable two sample version of Breitling's rank product statistic for the i^{th} gene is, up to a normalization constant, the product

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

RP_i is the geometric mean of the ranks of the i^{th} gene from sample 1, divided by the geometric mean of the ranks of the i^{th} gene from sample 2. Differential rankings of the i^{th} 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 of Breitling and colleagues [1], to devise a 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

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

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 $a_j(i) = \log(i)$, $1 \leq i \leq n$, $1 \leq j \leq k$. 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 $\text{Exp}(1)$. The key here is that the $\text{Exp}(1)$ 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

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

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

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

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:

$$\begin{aligned} \text{Prob}(RP_i \leq t) &= \text{Prob}(\log(RP_i) \leq \log(t)) \\ &\approx \text{Prob}(Y \leq \log(t)). \end{aligned}$$

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

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

Gil-Pelaez (1951) introduced a simple inversion formula, showing how the univariate cumulative distribution function F can be obtained by numerical inversion of its characteristic function φ :

$$F(x) = \frac{1}{2} + \frac{1}{2\pi} \int_0^\infty \frac{e^{itx}\varphi(-t) - e^{-itx}\varphi(t)}{it} dt = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{e^{-itx}\varphi(t)}{t} dt, \quad (3.2)$$

where $\text{Im}(\cdot)$ denotes the imaginary part of the argument. This formula has been exploited in similar contexts, initially by Imhof [11] and Davies [12,13], for numerical determination of the distribution of linear combinations of chi-squared random variables, which are, of course, gamma random variables. We have utilized Mathematica V.7 (Wolfram Research Inc., 2009) for numerical calculation of the integral expression for F , (3.2). As a check on accuracy, we computed in Mathematica exceedance probabilities for the quadratic forms given in Table 1 of Imhof [11], and compared them with the exact values given by Imhof. Our results are given in our Table 1, below. The probabilities determined in Mathematica agree to four places to the exact probabilities given by Imhof, with one exception, where the discrepancy is .0001. Notably, the values produced numerically from (3.2) are far more accurate than the two moment and three moment approximations to the quadratic forms investigated by Imhof.

3.2 Approximations

Calculation of the exact distribution of $\log(RP_i)$ can be tedious, hence approximations are of practical interest. The representation (2.3), together with the central limit theorem, suggest that a normal approximation to the distribution of $\log(RP_i)$ should be adequate, at least, for large k_1 and k_2 . More refined approximations are also available, as we now describe.

Recall that the moment generating function of Y in (2.4) is given by $E[e^{tY}]$, and that this expression is available from (3.1). The cumulant generating function of Y , $K_Y(t)$, is defined as the natural log of its moment generating function:

$$K_Y(t) = -k_1 \log(1+t/k_1) - k_2 \log(1-t/k_2). \quad (3.3)$$

The cumulants κ_n of Y can be determined by differentiation and evaluation at 0 of $K_Y(t)$: $\kappa_n = K_Y^{(n)}(0)$, the n^{th} derivative of $K_Y(t)$ evaluated at $t=0$. In particular, the first five cumulants of Y are:

$$\begin{aligned} \kappa_1 &= 0 \text{ (mean)} \\ \kappa_2 &= (1/k_1) + (1/k_2) \text{ (variance)} \\ \kappa_3 &= 2 \left\{ -(1/k_1)^2 + (1/k_2)^2 \right\} \\ \kappa_4 &= 6 \left\{ (1/k_1)^3 + (1/k_2)^3 \right\} \\ \kappa_5 &= 24 \left\{ -(1/k_1)^4 + (1/k_2)^4 \right\}. \end{aligned} \quad (3.4)$$

In addition, the skewness of Y is simply $\kappa_3/\kappa_2^{3/2}$, and the kurtosis of Y is κ_4/κ_2^2 .

One might expect the normal approximation to the distribution of Y to be better if both skewness and kurtosis are near 0 [the values of the normal distribution]. In this regard, we remark that, if $k_1 = k_2$, that is, equal sample sizes, then the distribution of Y is symmetric about 0 [all odd moments of Y are exactly 0.] In the presence of pronounced skewness or kurtosis, a normal approximation might be inadequate, and might be improved upon. One possibility is the Cornish-Fisher expansion [14,15,16], which we now describe.

The Cornish-Fisher expansion is a formula for approximating the quantiles of a random variable, based on quantiles of a limiting (typically normal) distribution, and correction factors involving its cumulants. This describes our situation: we take the limiting distribution of $\log(RP_i)$ as normal, and we know its cumulants to any order. The formula itself is rather cumbersome; we give it in the Appendix for reference.

In Table 2 we compare percentage points for the distribution of $\log(RP_i)$ in two instances, $n_1 = n_2 = 10$ and $n_1 = 8, n_2 = 12$, the percentage points being determined either exactly (numerical inversion of the characteristic function), a Cornish-Fisher approximation (Appendix), or a simple normal approximation (using the first two moments, (3.4)). In the symmetric case $n_1 = n_2 = 10$ (skewness = 0, kurtosis = 0.3), the normal approximation seems quite adequate, except in the tails, where the concordance of the exact values and Cornish-Fisher is quite good. In the asymmetric setting $n_1 = 8, n_2 = 12$ (skewness = 0.47, kurtosis = 0.35), discrepancies between the normal approximation and either the exact values or the Cornish-Fisher approximations are particularly pronounced with the midrange percentage points. Again, the Cornish-Fisher approximation seems quite good.

Although formidable looking, the formula is easily implemented in Excel; and, we provide an Excel workbook with embedded macro for calculation of selected quantiles of the $\log(RP_i)$ statistic, via both the normal approximation and the Cornish-Fisher expansion, for arbitrary n_1 and n_2 .

We remark that a saddlepoint approximation [17] to the exact distribution of $\log(RP_i)$ can also be developed; on the basis of prior work [18], we would expect the saddlepoint to perform comparably to the Cornish-Fisher expansion. As with Cornish-Fisher, the saddlepoint approximation relies on knowledge of the cumulant generating function (3.3), which has a rather simple form in our setting.

4. An example

For illustrative purposes, we will evaluate by means of the rank product method a dataset designated SKCC1 from [19]. SKCC1 consists of the expression levels of $n=23$ prostate-cancer related genes, evaluated on 79 patients who had undergone radical prostatectomy. Of these 79 patients, $k_1=42$ subjects had remained progression-free at 5 years post-surgery, whereas the remaining $k_2=37$ patients had experienced progression by 5 years. We calculated the $\log(RP)$ statistics from (2.2) for each of the 23 genes, the goal being to determine whether any of the genes were differentially expressed in the progressors compared to the non-progressors. The rank product method identified only one gene with a marginal p -value of less than .10, though this could be construed to be not unexpected. In Figure 1, we present a probability plot of the observed p -values: indeed, there seems to be a preponderance of p -values between .2 and .8. We would conclude that these genes fail to discriminate progressors from non-progressors.

5. Conclusions

We have introduced a two-sample rank product statistic motivated by the underlying theory of linear rank statistics. Numerical methods are available for the distribution of the log-

transformed rank product statistic in the two sample setting, and should be quite satisfactory in practice.

Our formulation of the two sample rank product statistic is analogous to parametric tests [e.g., t-tests] commonly used with microarray data. As Breitling and colleagues have previously noted, the within-replicate ranking has various advantages in this setting, obviating to some extent the normalization issues presented to researchers.

Nevertheless, it should be emphasized that the ambit of the rank product method extends beyond the microarray setting. Our example is illustrative of this: our data did not arise from a high-throughput microarray experiment as originally addressed by Breitling and colleagues, but from a more limited confirmatory experiment. Yet the rank product method is well-suited to assess differential gene expression in this limited setting. The take-home message is, regardless of its provenance, the rank product method should be viewed as a nonparametric statistical technique widely applicable to conventional hypothesis testing situations.

Acknowledgments

This research was supported in part by NIH grants PO1AI070167 and PO1CA104898. The author thanks the reviewers for perceptive comments.

Appendix - The Cornish-Fisher expansion

Let X denote a standardized random variable with mean 0 and standard deviation 1. Let X_α and Z_α denote the α quantile of X and the standard normal distribution, respectively, and let k_r denote the r^{th} cumulant of X . Then, using the first five cumulants, X_α has the Cornish-Fisher expansion [14]

$$\begin{aligned} X_\alpha \approx & Z_\alpha + (1/6)(Z_\alpha^2 - 1)\kappa_3 + (1/24)(Z_\alpha^3 - 3Z_\alpha)\kappa_4 - (1/36)(2Z_\alpha^3 - 5Z_\alpha)\kappa_3^2 \\ & + (1/120)(Z_\alpha^4 - 6Z_\alpha^2 + 3)\kappa_5 - (1/24)(Z_\alpha^4 - 5Z_\alpha^2 + 2)\kappa_3\kappa_4 \\ & + (1/324)(12Z_\alpha^4 - 53Z_\alpha^2 + 17)\kappa_3^3. \end{aligned} \quad (\text{A.1})$$

In our setting, $\log(\text{RP}_i)$ is not standardized: it has mean 0, but standard deviation $(1/k_1 + 1/k_2)^{1/2}$; see (3.4) above. Nevertheless, we can apply the Cornish-Fisher expansion (A.1) to obtain the α quantile of $\log(\text{RP}_i)$ /s.d.($\log(\text{RP}_i)$); the corresponding quantile of $\log(\text{RP}_i)$ is obtained by multiplying this quantity by s.d.($\log(\text{RP}_i)$)

References

- [1]. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*. 2004; 573:83–92. [PubMed: 15327980]
- [2]. Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the t-test for the analysis of biological microarray data. *J. Bioinf. Comp. Biol.* 2005; 3:1171–1189.
- [3]. Hong F, Breitling R, McEntree CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*. 2006; 22:2825–2827. [PubMed: 16982708]
- [4]. Jeffrey IB, Higgins DG, Culhane AE. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*. 2006; 7:359. [PubMed: 16872483]
- [5]. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*. 2008; 24:374–382. [PubMed: 18204063]

- [6]. Birmingham A, Selfors LM, Forster T, et al. Statistical methods for analysis of high-throughput RNA interference screens. *Nature Methods*. 2009; 6:569–575. [PubMed: 19644458]
- [7]. Wiederhold E, Gandhi T, Permentier HP, et al. The yeast vacuolar membrane proteome. *Mol Cell Proteomics*. 2009; 8:380–392. [PubMed: 19001347]
- [8]. Hoefsloot HC, Smit S, Smilde AK. A classification model for the Leiden proteomics competition. *Stat Appl Genet Mol Biol*. 2008; 7:8.
- [9]. Koziol JA. Comments on the rank product method for analyzing replicated experiments. *FEBS Letters*. 2010; 584:941–944. [PubMed: 20093118]
- [10]. Gil-Pelaez J. Note on the inversion theorem. *Biometrika*. 1951; 38:481–482.
- [11]. Imhof JP. Computing the distribution of quadratic forms in normal variables. *Biometrika*. 1961; 48:419–426.
- [12]. Davies RB. Numerical inversion of a characteristic function. *Biometrika*. 1973; 60:415–417.
- [13]. Davies RB. Algorithm AS 155. The distribution of a linear combination of C^2 random variables. *Applied Statistics*. 1980; 29:323–333.
- [14]. Cornish EA, Fisher RA. Moments and cumulants in the specification of distributions. Review of the International Statistical Institute. 1938; 5:307–320.
- [15]. Fisher RA, Cornish EA. The percentile points of distributions having known cumulants. *Technometrics*. 1960; 2:209–225.
- [16]. Stuart, A.; Ord, JK. Kendall's Advanced Theory of Statistics. Fifth Edition. Vol. Volume 1. Oxford University Press; New York: 1987.
- [17]. Lugannani R, Rice SO. Saddlepoint approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*. 1980; 12:475–490.
- [18]. Koziol JA. A note on Lancaster's procedure for the combination of independent events. *Biom. J.* 1996; 38:653–660.
- [19]. Koziol JA, Feng AC, Jia Z, et al. The wisdom of the commons: ensemble tree classifiers for prostate cancer diagnosis. *Bioinformatics*. 2009; 25:54–60. [PubMed: 18628288]

Figure 1.

Probability plot of the observed p-values of 23 genes comparing 42 progressors vs. 37 non-progressors, in 79 patients who had undergone radical prostatectomy. The p-values were obtained from the 23 marginal log-transformed rank product statistics (2.3), as numerically evaluated in Mathematica.

Table 1

Probability that the quadratic form Q exceeds x . Exact exceedance probabilities are from Imhof (1961); 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	.9458	.94578
	0.7	.5064	.50635
	2	.1240	.12403
$Q_2=0.6X_2^2+0.3X_2^2+0.1X_2^2$	0.2	.9936	.99354
	2	.3998	.39980
	6	.0161	.01610
$Q_3=0.6X_6^2+0.3X_4^2+0.1X_2^2$	1	.9973	.99732
	5	.4353	.43525
	12	.0088	.00877
$Q_4=0.6X_2^2+0.3X_4^2+0.1X_6^2$	1	.9666	.96664
	3	.4196	.41956
	8	.0087	.00872
$Q_5=0.7X_{6,6}^2+0.3X_{2,2}^2$	2	.9939	.99388
	10	.4087	.40866
	20	.0221	.02208
$Q_6=0.7X_{1,6}^2+0.3X_{1,2}^2$	1	.9549	.95487
	6	.4076	.40758
	15	.0223	.02232
$(1/3)Q_3 + (2/3)Q_4$	1.5	.9891	.98906
	4	.3453	.34527
	7	.0154	.01540
$(1/3)Q_3 - (2/3)Q_4$	-2	.9102	.91023
	0	.4061	.40611
	2.5	.0097	.00976
$(1/2)Q_5 + (1/2)Q_6$	3.5	.9563	.95632
	8	.4152	.41524
	13	.0462	.04623
$(1/2)Q_5 - (1/2)Q_6$	-2	.9218	.92179
	2	.4779	.47789
	7	.0396	.03963
$(1/4)(Q_3+Q_4+Q_5+Q_6)$	3	.9842	.98416
	6	.4264	.42638
	10	.0117	.01166
$(1/6)(Q_3-Q_5) + (2/6)(Q_6-Q_4)$	-3	.9861	.98614

Q	x	Exact	Approximate
	0	.5170	.51702
	4	.0152	.01520

Table 2

Percentage points of the distribution of $\log(RP_i)$ determined by numerical inversion of the characteristic function, Cornish-Fisher approximation or normal approximation.

A. $n_1 = n_2 = 10$			
statistic	char fn	Cornish-Fisher	normal approx
-2.0	0.00002483	0.00005184	0.00000387
-1.8	0.00011403	0.00016925	0.00002850
-1.6	0.00048638	0.00054721	0.00017331
-1.4	0.00161378	0.00171210	0.00087256
-1.2	0.00500671	0.00515650	0.00364518
-1.0	0.0143208	0.0143608	0.0126737
-0.8	0.0372720	0.0371532	0.0368191
-0.6	0.0869645	0.0867315	0.0898562
-0.4	0.179131	0.178998	0.185547
-0.2	0.321522	0.321558	0.327360
0.0	0.5	0.5	0.5
0.2	0.678478	0.678442	0.672640
0.4	0.820869	0.821002	0.814453
0.6	0.913036	0.913269	0.910144
0.8	0.962728	0.962847	0.963181
1.0	0.985679	0.985639	0.987326
1.2	0.994993	0.994884	0.996355
1.4	0.998386	0.998288	0.999127
1.6	0.999514	0.999453	0.999827
1.8	0.999886	0.999831	0.999972
2.0	1.000000	0.999949	0.999996
B. $n_1 = 8, n_2 = 12$			
statistic	char fn	Cornish-Fisher	normal approx
-2.0	0.00013043	0.00014618	0.00000589
-1.8	0.00038759	0.00041685	0.00004013
-1.6	0.00110905	0.00156047	0.00022794
-1.4	0.00303799	0.00309851	0.00108020
-1.2	0.0079135	0.0079631	0.0042810
-1.0	0.0194419	0.0194331	0.0142299
-0.8	0.0446093	0.0445039	0.0398257
-0.6	0.0944895	0.0943304	0.0943333
-0.4	0.182348	0.182267	0.190418
-0.2	0.316302	0.316340	0.330629
0.0	0.487775	0.487766	0.500000
0.2	0.666544	0.666472	0.669371
0.4	0.815089	0.815227	0.809582
0.6	0.913056	0.913376	0.905667

B. $n_1 = 8, n_2 = 12$			
statistic	char fn	Cornish-Fisher	normal approx
0.8	0.965048	0.965225	0.960174
1.0	0.987801	0.987743	0.985770
1.2	0.996239	0.996088	0.995719
1.4	0.998958	0.998834	0.998920
1.6	0.999737	0.999667	0.999772
1.8	0.999939	0.999907	0.999960
2.0	0.999987	0.999974	0.999994