Pairwise Comparisons for Gene Expression Array Data (*)

Confronti a coppie per dati di espressione genica

Aldo Solari¹, Silvio Bicciato²

¹ Dipartimento di Processi Chimici dell’Ingegneria, Università di Padova
   e-mail: solari@stat.unipd.it
² Dipartimento di Scienze Biomediche, Università di Modena e Reggio Emilia

Keywords: gene expression, multiple comparisons, multiple testing, permutation test

1. Introduction

Researchers using RNA expression microarrays in experimental designs with more than two treatment groups often identify statistically significant genes with ANOVA approaches. However, the ANOVA test does not discriminate which of the multiple treatment groups differ from one another. For instance, one of the most highly successful methods (at least in terms of popularity among bench researchers), has been the resampling-based multiple test procedure (MTP, Dudoit and van der Laan, 2008), implemented in the R package multtest. For tests of equality of means in a one-way design, multiple-sample F-statistics are available, but in case of rejection of the null hypothesis one will frequently want to know more about the means than just that they are unequal. The number of genes significant by each pairwise comparison alone is useful, helping researchers determine which comparisons show the largest number of significant results (suggesting treatments with the most powerful effects on the transcriptome).

Consider the one-way layout setting. Let $X_{j,i}$ with mean $\mu_j = (\mu_{1,j}, \ldots, \mu_{m,j})^t$ be the $i$th independent m-dimensional observation on the $j$th treatment, $j = 1, \ldots, k$, $i = 1, \ldots, n_j$. The following MANOVA location-shift model is considered

$$X_{j,i} = \mu_j + \epsilon_{j,i}$$

where $\epsilon_{j,i}$ is an m-dimensional error term with $E(\epsilon_{j,i}) = 0$, such as $\epsilon_{j,i} \sim N_m(0, \Sigma)$.

The problem of all pairwise comparisons, that is, to test the $\binom{k}{2}$ null hypotheses

$$H_{j,l} = \mu_j = \mu_l, \quad j < l$$

by using permutation tests, also in the simplest case of $m = 1$, has received little attention. To obtain valid permutation tests, one must consider for each hypothesis $H_{j,l}$ only the set of permutations that become equally likely under that hypothesis. As a consequence, one has to permute observations only within the $j$th and the $l$th treatment, and not among all the $k$ treatments, obtaining $\binom{k}{2}$ separate tests and not a joint testing family (Hochberg and Tamhane, 1987). To improve on the Bonferroni-Holm stepdown procedure, Finos and Solari (2008) made effective use of the dependence structure of test statistics in order to increase the ability to detect false null hypotheses and providing strong control of the familywise error rate.

(*) Work supported by CARIPARO grant
Here we extend to the multivariate case \((m > 1)\) the work of Finos and Solari (2008) to categorize significant genes by their expression patterns (as determined by the results of pairwise comparisons, that is, to test the \(m \cdot \binom{k}{2}\) null hypotheses

\[
H_{h,j,l} = \mu_{h,j} = \mu_{h,l}, \quad h = 1, \ldots, m, \quad j < l.
\]

The algorithm was applied to a well-known study comparing the expression of breast cancer tumor tissues among individuals who are “BRCA1-mutation-positive”, “BRCA2-mutation-positive”, and “Sporadic” (Hedenfalk et al., 2001). The expression measurements used in the study consist of 3226 genes on 22 arrays; seven arrays were obtained from the BRCA1 group, eight from the BRCA2 group, and six from the Sporadic group. As in Hulshizer and Blalock (2007), we combine results from all of the pairwise comparisons for each probe set, creating a pattern ID. Pattern IDs are constructed using logic gates that use “increase”, “no significant change”, or “decrease” results from each pairwise comparison. The first pairwise comparison is assigned 1, 0, or -1; the second is assigned 10, 0, or -10; the third is assigned 100, 0, -100 and so on. In this way, the sum of each combination of pairwise comparisons for a given probe set creates a pattern ID encoding that pattern’s statistically defined shape, and allowing researchers to easily group different genes that belong to the same pattern. Further, two patterns of opposite sign and the same absolute value will be mirror reflections of one another, which may have value for assessing opposing actions in single pathways.

References


