ANOVA AND FRIEDMAN DOCUMENTATION

ANDREA ZORRO-ARANDA

November 2021

We proposed a modification of the method proposed by Küffner et al. (2012). Here a new score for GRN inference derived from the ANOVA is suggested. This score is a non-linear correlation coefficient which described the likelihood of interaction between a TF and a TG. A two-way ANOVA is used to model a dependent variable (TG expression) as a response of two independent variables C and G, as well as the error. In this case, C is the effect across different experimental conditions of the differential expression and G is the similarity in the expression profiles of the TG and the TF evaluated. The null hypothesis for a two-way ANOVA is that there is no significant difference in means of C, G, and their interaction. Thus, the sum of squared deviations (SS) is divided into four components:

$$SS_T = SS_C + SS_G + SS_{CG} + SS_{error} \tag{1}$$

Where a high SS_C represents a strong difference in the expressions among conditions. A high SS_G represents a strong difference in the expression of both genes. And a high SS_{CG} indicates that the two effects are linked, which means strong differences among conditions due to strong differences between the genes. Therefore, the strength of association (η^2_+) is proportional to the fraction of SS_C of the total sum S_T, *i. e.* the fraction of the total variance that corresponds to the difference in the expression among conditions.

$$\eta^2_+ = \frac{SS_C}{SS_T} \tag{2}$$

In contrast to other correlation coefficients, η^2_+ do not identify negative correlations. Thus, reversing the signs of the TF expression profile, we can compute η^2_-. And compute the final η^2 as

$$\eta^2 = \max(\eta^2_+, \eta^2_-) \tag{3}$$

However, ANOVA has specific requirements to perform a proper application of the metric. One of them is that the distributions are assumed to be normal (Walpole et al., 2016). This might not be accurate in the case of gene expression profiles. Therefore, we proposed to compute the non-linear correlation coefficient from a Friedman Test instead of a two-way ANOVA. This is a non-parametric alternative since it does not assume normality (Hoffman, 2015). The computation of η^2 is the same as in Equations (1) to (3). The algorithm was implemented in Matlab, both the ANOVA and the Friedman method.

Hoffman, Julien I. E. (Jan. 1, 2015). "Chapter 26 - Analysis of Variance II. More Complex Forms". In: *Biostatistics for Medical and Biomedical Practitioners*. Ed. by Julien I. E. Hoffman. Academic Press, pp. 421–447.

Küffner, Robert et al. (May 15, 2012). "Inferring gene regulatory networks by ANOVA". In: *Bioinformatics* 28.10, pp. 1376–1382.

Walpole, Ronald E. et al. (Mar. 17, 2016). *Probability & Statistics for Engineers & Scientists, MyLab Statistics Update*. 9 edition. Boston: Pearson. 816 pp.