Estimating and Testing Circular Orders for Cell Cycle Genes Expression Data.

Cristina Rueda [1] Miguel Fernández [1], Sandra Barragán [1] and Shyamal D. Peddada [2]

[1] Department of Statistics and Operations Research, University of Valladolid, Valladolid, Spain

[2]Biostatistics Branch, NIEHS (NIH), Research Triangle Park, NC, USA

1 Introduction

Analysis of angular data has a long history with well-developed theory and methodology documented in several books [cf. Fisher (1993), Mardia and Jupp (2000)]. Most of the theory have been developed several decades ago. However, in recent years, and motivated by the applications, there has been a reawaked interest in drawing inferences regarding angular parameters. In particular, in the biology field to analyze periodic patterns of gene expressions. These patterns show the contribution of the genes in oscillatory biological processes, such as the cell-cycle, the circadian clock or the metabolic cycle. The time to peak expression (known as the phase angle) of such a gene can be mapped onto a unit circle and is an angular parameter.

The phase angle of a gene is often associated with its biological function. For this reason biologists are interested in estimating the phase angle of various genes involved in an oscillatory biological process. Furthermore, for a normal biological process to proceed in an orderly fashion (such as a normal cell division cycle) genes involved in the process need to express like an orchestra. Therefore, biologists are not only interested in determining the phases of various periodic genes but they are also interested in determining their relative order of expression and also in knowing about the relationships between the expressions of different experiments and species.

Suppose $\phi_1, \phi_2, \ldots, \phi_n$ are n angular parameters, then a researcher is interested in determining the order among these n parameters around the unit circle. Thus the goal is to determine whether, for example, ϕ_1 is followed by ϕ_2 which is followed by ϕ_3 and so on ϕ_{n-1}, is followed by ϕ_n which in turn is followed by ϕ_1 around the unit circle. We shall denote such a relative order among the parameters by $\phi_1 \preceq \phi_2 \preceq \cdots \preceq \phi_{n-1} \preceq \phi_n \preceq \phi_1$. Common problems of interest include (a) determining such order relationships among n angular parameters, and (b) testing the null hypothesis that the parameters satisfy a given relative order. As seen from recent literature, such problems arise naturally in a wide range of applications. The Papers: Peng et al (2005), Hughes et al (2009), Fernández et al (2012) and Slavov et al (2012), are just a few examples, among many others, dealing with this interesting issue.

Biologists routinely use ad-hoc visual methods, such as heat-maps and circle plots to make decisions regarding the relative order among angular parameters. Although the visual methods are intuitive, easy to understand and implement they ignore uncertainty associated with the estimated values of angular parameters and hence are not very satisfactory.

In this talk, the formal statistical methodology for drawing inferences regarding a relative order among a collection of angular parameters is exposed. The first steps have been taken in Rueda et al (2009) and Fernández et al. (2012. The methodology developed in these papers solve the problems of estimating and testing whether the relative order among the parameters

satisfies a pre-defined order. In this sense, this problem is analogous to "one-sample" problem in classical statistics.

In practice, however, it is not always reasonable to assume that the relative order is pre-defined with no uncertainty. Drawing inferences regarding the relative order of angular parameters in two populations while accounting for uncertainties in both samples, i.e. the "two-sample" version of the problem, is a non-trivial problem. In this talk we also formulate the desired inferential problem, addressing both estimation as well as testing problems, using a novel methodology based on rank aggregation procedures and permutation test, specifically designed to deal with circular data.

The goodness of the proposed methodology will be illustrated by analyzing cell-cycle gene expression data sets, which was the motivation for the methodology developed.

Concluding remarks and some open problems in this field will also be discussed.

References

Fernandez, M.A., Rueda, C. and Peddada. (2012). Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species, *Nucl. Acids Res.*, **40(7)**, 2823-2832.

Fisher, N.I. (1993). *Statistical Analysis of circular data*. Cambridge University Press.

Hughes, M.E. and DiTacchio, L. and Hayes, K.R. and Vollmers, C. and Pulivarthy, S. and Baggs, J.E. and Manda, S. and Hogenesch, J.B. (2009). Harmonics of Circadian Gene Transcription in Mammals, *PLoS Genetics*, **5 (4)**.

Mardia, K. and Jupp, P. (2000). *Directional Statistics*. John Wiley and Sons.

Peng, X. and Karuturi, R.K.M. and Miller, L.D. and Lin, K. and Jia, Y. and Kondu, P. and Wang, L. and Wong, L. and Liu, E.T. and Balasubramanian, M.K. and Liu, J. (2005). Identification of Cell Cycle-Regulated Genes in Fission Yeast, *The American Society for Cell Biology*, **16**, 1026-1042.

Rueda, C., Fernandez, M.A. and Peddada. (2009). Estimation of Parameters Subject to Order Restrictions on a Circle with Application to Estimation of Phase Angles of Cell-Cycle Genes. *Journal of the American Statistical Association*, **104(485)**, 338-347.

Slavov, N. and Airoldi, E.M. and van Oudenaarden, A. and Botstein, D.(2012). A conserved cell growth cycle can account for the environmental stress responses of divergent eukaryotes. *Molec. Biol. Cell*, **23**, 1986-1997.