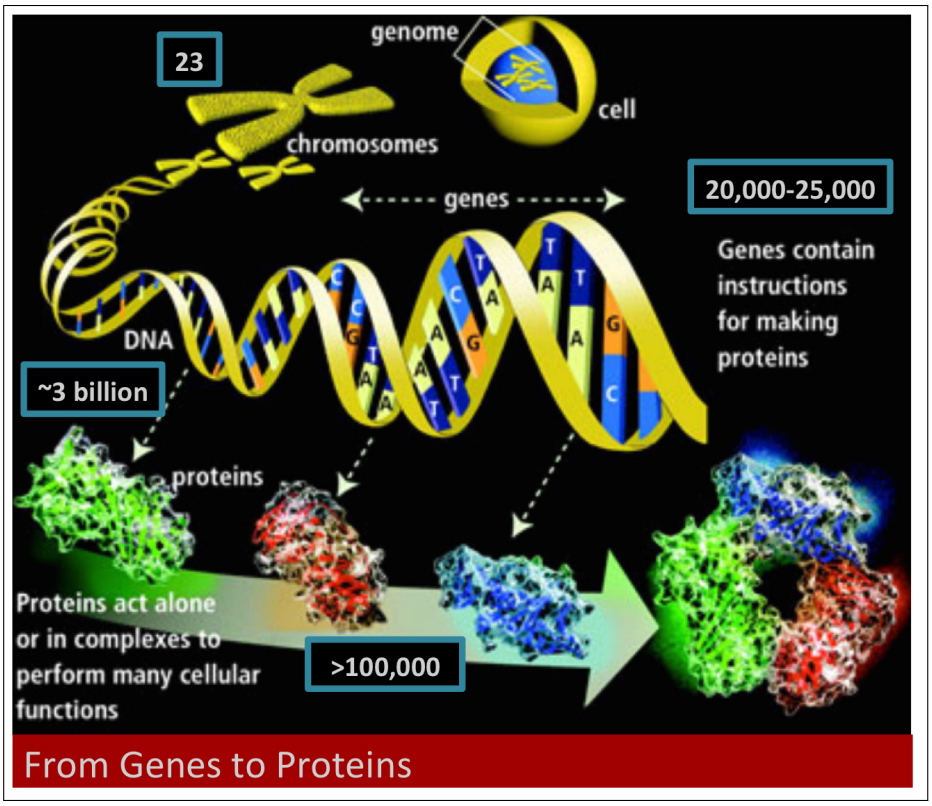
LASR 2015 — Geometry-Driven Statistics and its Cutting Edge Applications: Celebrating Four Decades of Leeds Statistics Workshops



Programme and Abstracts

Edited by K.V. Mardia, A. Gusnanto, C. Nooney & J. Voss

30th June to 2nd July 2015

WILEY

UNIVERSITY OF LEEDS

Big Data in Human Genomics/Proteomics: The amount of data generated in our world has increased exponentially so that we now live in the world of Big Data. A new field has arisen that is called Data Analytics/Data Science. In genomic science, we have 23 chromosomes leading to 20,000-25,000 genes, 3-billion-letter (ATCG) in the human genome project, millions of protein sequences, more than 100,000 protein structures, each with many large number of variables. There are many challenges from drug discovery to stem cell research. Image Courtesy of U.S. Human Genome Project and Professor Gabriela C. Freue.

Department of Statistics, University of Leeds, U.K.

Geometry-Driven Statistics and its Cutting Edge Applications:
Celebrating Four Decades of Leeds Statistics Workshops

International Conference, held in Leeds, UK, 30 June-2 July 2015,

The 33rd Leeds Annual Statistical Research (LASR) Workshop

Sponsored and organised by the Department of Statistics, University of Leeds.

Proceedings edited by

K.V. Mardia, A. Gusnanto, C. Nooney & J. Voss

Department of Statistics,

University of Leeds, Leeds, LS2 9JT.

Conference Organisers

K.V. Mardia (Chairman), A. Gusnanto, J. Voss, C. Nooney, J.M. Brennan,

R. Aykroyd, W.R. Gilks & J.T. Kent

ISBN **978 0 85316 338 1**

Circular Isotone Regression with an Application to Cell-cycle Biology.

Cristina Rueda[1], M.A. Fernández[1], S. Barragán[1], K.V. Mardia[2] and S.D. Peddada[3]

[1] Department of Statistics and OR, Universidad de Valladolid, Spain
[2] Department of Statistics, University of Oxford, Oxford, UK,
and Department of Statistics, University of Leeds, Leeds, UK
[3] Biostatistics Branch, NIEHS (NIH), Research Triangle Park, NC, USA

1 Introduction

This work is motivated by a problem encountered in Molecular Biology where researchers are interested in correlating angular data from two oscillatory systems. The observations are the time to peak expression (also known as phase angle) of periodic genes under two different conditions (dose levels, organs or even species). In particular, we deal here with expression data from genes participating in the cell-cycle. Cell-biologists are often interested in drawing inferences regarding the phase angle of cell-cycle genes since they are considered to be associated with the gene's biological function (Jensen et al 2006).

Several distinctive features should be taken into account to derive a correct model for this application. First, since the cell division cycle is a carefully orchestrated and periodic process, the peak expressions of cell-cycle genes follow an order according to their functions and the same order should apply for the tow different conditions (which will play the role of the response and the explicative variables). Then, the model should assure that the response must run exactly one cycle as the explicative runs one cycle without moving back. Second, since cells goes through 4 phases with different biological functions and even with different lengths across species, the model approach should be flexible to deal with possible different correlations from phase to phase.

While the regression model proposed in Downs and Mardia (2002) is likely to perform well when the duration of time spent by a cell in different phases of a cell-cycle is same across all species, it may be too rigid when the duration of time is not same across different species as the lengths of the four phases in which the cell-cycle is divided change from one species to another, so that the functional relationship between species may be different in each of the phases. On the other side, the non-parametric alternatives, in particular the proposal developed in Di Marzio et al (2013) do not always fulfill the two important conditions, commented above, that the regression should verify: monotonicity and synchronicity.

In this paper we develop regression models able to deal with these features. In particular, we introduce a general isotonic regression model and a flexible piecewise regression model that can be useful for drawing inferences when the duration of time spent in different phases by a cell varies across species.

2 Circular Isotonic Regression models

Let $(\psi_i, \theta_i), i = 1, ..., n$ denote a random sample from the circular response Ψ and from the circular independent variable Θ. Assume that the ψ_i values come from independent von Mises

distributions $M(\phi_i, \kappa)$.

2.1 The general Isotonic Model

The CIRE (Circular Isotonic Regression Estimator) of Ψ is defined as

$$\widetilde{\mathbf{\Phi}}^{(\boldsymbol{O})} = \arg\min_{\mathbf{\Phi} \in C_\Theta} SCE(\mathbf{\Psi}, \mathbf{\Phi}),$$

where C_Θ is the circular order induced by the independent Θ variable,

$$C_\Theta = \{\mathbf{\Phi} \in [0, 2\pi)^n : \phi_a \leq \phi_b \leq \phi_c \leq \phi_a \Leftrightarrow \theta_a \leq \theta_b \leq \theta_c \leq \theta_a\}.$$

The CIRE exists, is almost sure unique, may be obtained from circular means of adjacent angles and is the restricted MLE under the Von-Mises model (see Rueda et al 2009).
The Circular Isotonic regression Model is simply defined as $\Phi = f(\Theta)$ where f preserves the order induced by Θ.

2.2 A Piecewise Isotonic Model

Consider k different sectors (pieces) in the independent variable. Then, we have (ψ_{ij}, θ_{ij}) with $i = 1, ..., k$ and $j = 1, ..., n$ where n_i is the number of observations in sector i, and θ_i^*, $i = 1, 2, ..., k$ are the sector borders or change points with $\theta_i^* < \theta_{ij} \leq \theta_{i+1}^*$ and $\theta_{k+1}^* = \theta_1^*$.
The Piecewise Circular-Circular Model is defined as

$$\phi_{ij} = \mu + 2\arctan\left(\omega_i \tan\frac{1}{2}(\theta_{ij} - \nu_i)\right), \text{ subject to,}$$
$$\omega_i \tan\frac{1}{2}(\theta_i^* - \nu_i) = \omega_{i-1} \tan\frac{1}{2}(\theta_i^* - \nu_{i-1}) \text{ for } i = 1, \ldots, k,$$

where $\nu_0 = \nu_k$, $\omega_0 = \omega_k$, μ is a global location parameter quantifying the rotation of the response that allows a better congruence with the independent variable; ν_i is the location parameter in sector $(\theta_i^*, \theta_{i+1}^*)$ and ω_i is the slope parameter in the sector $(\theta_i^*, \theta_{i+1}^*)$.
In the particular application of cell-cycle data, the sectors are determined by four phases of the cell-cycle and the estimation problem is solved via maximum likelihood subject to the following restrictions:
Continuity Restrictions:

$$\omega_i = \frac{\tan\left(\frac{\theta_{i+1}^* - \nu_{i+1}}{2}\right)}{\tan\left(\frac{\theta_{i+1}^* - \nu_i}{2}\right)}\omega_{i+1} \text{ for } i = 1, \ldots, k-1.$$

Monotonicity Restrictions:

$$\omega_i \geq 0 \text{ for } i = 1, \ldots, k.$$

Synchronicity Restrictions:
Let $z_i = \nu_i + 2\arctan\left(\frac{1}{\omega_i}\tan\left(\frac{-\mu}{2}\right)\right)$ for $i = 1, \ldots, k$, be a possible zero of ith piece of the function. Then

$$\sharp\left\{z_i : z_i \in \left(\theta_i^*, \theta_{i+1}^*\right]\right\} = 1.$$

This model is presented in the forthcoming paper Rueda et al (2015).

3 Application

We have analyzed data from 32 periodic genes in two species of yeast, *S. cerevisiae* and *S. pombe*, considering 2 experiments from *S. cerevisiae* that we denote as **Ca** (which is the explicative in all models considered) and **Cb** (which is used as response in one of the models) and 2 experiments from *S. pombe* that we denoted as **Pa** and **Pb** (used as responses in the other two models presented here).

We have fitted the Circular-Circular regression model from Downs and Mardia (2002), the non-parametric circular model from Di Marzio et al (2012), and the two models defined in section 2.

The statistics used to select between models are, the Circular Distance Criterion $CDC(M)$, which is a sort of lack of fit criterion and is defined as $CDC(M) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}$, where $e_{ij} = 1 - (\cos(\psi_{ij} - \widehat{\psi}_{ij}))$ and a Generalized Akaike Information measure ($GAIC$) that is defined as $GAIC(M) = 2\ln(l(M)) - 2GDF$, where $l(M)$ is the model likelihood and GDF is a penalization factor obtained as the sum of the sensitivity of each fitted value to perturbation in the corresponding observed value. The main advantage of this measure is that is applicable to complex modeling procedures including nonparametric and restricted models (see Ye 1998 and Zhang et al 2012). Full details of this criterion appear in Rueda et al (2015).

The goodness of fit statistics for the different regression models are given in the following table. Full interpretation of these results together with the appropriate graphs will be given during the talk.

Experiment	Statistic	**Parametric**	**Piecewise**	**Isotone**	**Non-Parametric**
Pa/Ca	CDC	0,3938	0,3323	0,2688	0,1671
Pb/Ca	CDC	0,3682	0,1657	0,2289	0,1970
Cb/Ca	CDC	0,1570	0,1461	0,0806	0,1292
Pa/Ca	GAIC	19,4116	19,8093	22,4058	31,1877
Pb/Ca	GAIC	23,4580	37,6419	46,4377	36,4998
Cb/Ca	GAIC	55,9770	55,4740	61,9504	57,4613

4 Conclusions

Restricted regression models have proved very useful to describe relationships between cell-cycle data expressions from different species. In particular, the piecewise model has several interesting advantages for this application. It is simple and interpretable, flexible enough to describe different correlations depending on the sector and versatile, as it can handle restrictions of monotonicity and synchronicity.

Related problems arise in other fields such as in circadian biology, metabolic cycle, evolutionary psychology or motor behavior. Other application where the piecewise model will be useful is for characterizing patterns of hormones during the menstrual cycle (with three distinct phases: follicular, ovulation and luteal).

References

Di Marzio, M., Panzera, A., Taylor, C.C. (2013). Non-parametric Regression for Circular Responses. *Scandinavian Journal of Statistics*, **40**, 238-255.

Downs, T.D. and Mardia, K.V. (2002). Circular Regression, *Biometrika*, **89**, 683-697.

Jensen, J.L., Jensen T.S., Lichtenberg, U., Brunak, S. and Bork, P. (2006). Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature*, **443**, 594-597.

Rueda, C., Fernández, M.A. and Peddada, S. (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**, 338-347.

Rueda, C., Fernández, M.A., Barragán, S., Mardia, K.V. and Peddada, S. (2015). Circular Piecewise Regression with an Application to Cell-cycle Biology. *Preprint*.

Ye, J. (1998). On Measuring and Correcting the Effects of Data Mining and Model Selection. *Journal of the American Statistical Association*, **93**, 120-131.

Zhang, B., Shen, X. and Mumford, S.L. (2012). Generalized Degrees of Freedom and Adaptive Model Selection in Linear Mixed-Effects Models. *Computational Statistics and Data Analysis*, **56**, 574-586.