C U R S O S E C O N G R E S O S

ADISTA22
Advances in Directional Statistics

Santiago de Compostela, Spain
15-17th June, 2022
Programme and Book of Abstracts

EDITION BY
Jose Ameijeiras Alonso
Rosa M. Crujeiras Casais
M. José Ginzo Villamayor
Alberto Rodríguez Casal
Paula Saavedra Nieves

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ADISTA22: Advances in Directional Statistics

Santiago de Compostela, Spain
15-17th June, 2022
Programme and Book of Abstracts

CURSOS E CONGRESOS DA
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
N.º 268

ADISTA22: Advances in Directional Statistics

Santiago de Compostela, Spain
15-17th June, 2022
Programme and Book of Abstracts

EDITION BY

Jose Ameijeiras Alonso
Rosa M. Crujeiras Casais
M. José Ginzo Villamayor
Alberto Rodríguez Casal
Paula Saavedra Nieves

2022
Universidade de Santiago de Compostela

Maqueta
Jose Ameijeiras Alonso
Rosa M. Crujeiras Casais
M. José Ginzo Villamayor
Alberto Rodríguez Casal
Paula Saavedra Nieves

Edita
Servizo de Publicacións e Intercambio Científico
da Universidade de Santiago de Compostela
Campus Vida
15782 Santiago de Compostela
usc.gal/publicacions
DOI: https://dx.doi.org/10.15304/9788419155924
ISBN 978-84-19155-92-4

Venue

Department of Statistics, Mathematical Analysis and Optimization.

Faculty of Mathematics. Address: Lope Gómez de Marzoa, s/n, Campus Vida.

15782, Santiago de Compostela, Spain.

Organizing committee

Jose Ameijeiras Alonso, University of Santiago de Compostela, Spain.

Rosa M. Crujeiras Casais, University of Santiago de Compostela, Spain.

M. José Ginzo Villamayor, University of Santiago de Compostela, Spain.

Giovanna Jona-Lasinio, Sapienza University, Italy.

Francesco Lagona, Rome 3 University, Italy.

Christophe Ley, Ghent University, Belgium.

Arthur Pewsey, University of Extremadura, Spain.

Alberto Rodríguez Casal, University of Santiago de Compostela, Spain.

Paula Saavedra Nieves, University of Santiago de Compostela, Spain.

Presentation

This book of proceedings collects the abstracts of the talks and posters presented at the third International Workshop on Advances in Directional Statistics (ADISTA22). It was a great pleasure to be able to meet during these trying times of the continuing COVID-19 pandemic, and we are grateful to all those who have collaborated in the ADISTA22 organization. In order to allay potential fears of contracting COVID on long journeys to Santiago de Compostela, the workshop has been organized to have a hybrid format with talks and posters presented both by those participating in person in Santiago and by others participating online from their home countries. We consider this to be the optimal format for the workshop in the current circumstances.

Despite the challenges posed by the pandemic, we were delighted to be able to welcome all 57 participants who took up our invitation to participate in the workshop: 33 in person and 24 online, from no less than 18 countries (Australia, Belgium, Canada, France, Germany, Italy, Japan, Luxembourg, Mexico, Norway, Portugal, South Africa, South Korea, Spain, Taiwan, Turkey, UK, USA). One of the central themes of the postponed ADISTA20 workshop was to be a celebration of Kanti Mardia's 85th Birthday. However, its postponement meant that we were unable to mark that event collectively. Instead, ADISTA22 provided a venue to celebrate another landmark event in the development of the field: namely, the publication 50 years ago, in 1972, of Kanti's seminal monograph *Statistics of Directional Data*. Since its publication, the field of Directional Statistics has had periods in which its development has been more vigorous than in others, with interesting new applications often stimulating renewed interest.

We sincerely hope that the workshop has provided a vehicle for stimulating dialogue and collaboration, leading to important future developments in the field.

We would like to thank most sincerely the University of Santiago de Compostela (USC) for hosting the workshop and the following institutions for their financial and institutional support of the workshop: The Department of Statistics, Mathematical Analysis and Optimization (USC); The Faculty of Mathematics (USC); The Galician Centre for Mathematical Research and Technology (CITMAga); The Galician Society for the Promotion of Statistics and Operations Research (SGAPEIO); The Galician Statistics Institute (IGE); The Spanish Society of Statistics and Operations Research (SEIO); The Italian Society of Statistics (SIS); The Luxembourg Statistical Society (LSS).

Finally, we very much hope that this book of abstracts will be enjoyable and stimulating for all of those participating in the ADISTA22, and for all the scientific community interested in Directional Statistics.

The Organizing Committee

Contents

Scientific program

Day 1 (15th June)

9:00-9:30	Registration
9:30-10:00	Opening ceremony
10:00-10:20	Statistics of Discrete Distributions on Manifolds: A Journey from the Karl Pearson Roulette Wheel Data to Some Smart Health Science Data **K. V. Mardia**
10:20-10:40	A Copula Model for Trivariate Circular Data **Shogo Kato**
10:40-11:00	Hidden Markov Random Field with Copula-Based Emission Distributions for the Analysis of Spatial Cylindrical Data **Francesco Lagona**
11:00-11:30	Coffee-break
11:30-11:50	PosterView 1
11:50-12:10	The FMM Approach to Model Oscillatory Signals. The Case of the Electrocardiogram **C. Rueda**
12:10-12:30	The Circular Structure of Oscillatory Signals: Applications with CPCA **Y. Larriba**
12:30-12:50	PosterView 2
12:50-13:10	Score Matching for Microbiome Compositional Data **J. L. Scealy**
13:10-13:30	Analogues on the Sphere of the Affine-equivariant Spatial Median **A. T. A. Wood**
13:30-15:30	Lunch
15:30-15:50	Chordal-Based Tests of Uniformity on the Hypersphere **E. García-Portugués**
15:50-16:10	Asymptotic Power of Sobolev Tests for Uniformity on Hyperspheres **T. Verdebout**
16:10-16:30	A Two-Sample Statistical Test for Directional Data **E. Gutiérrez-Peña**
19:30	Reception in Pazo de Fonseca

Day 2 (16th June)

9:30-09:50	Geodesic Projection of the von Mises-Fisher Distribution for Projection Pursuit of Directional Data **S. Jung**
9:50-10:10	Directional Distributions and the Half-Angle Principle **J. Kent**
10:10-10:30	PosterView3
10:30-10:50	Adaptive Warped Kernel Estimation for Nonparametric Regression with Circular Responses **T. M. Pham Ngoc**
10:50-11:10	Bivariate Linear-Circular Panel Data Modelling with Flexible Random Effects **A. Maruotti**
11:10-11:40	Coffee-break
11:40-12:20	Poster Session
12:20-12:40	Robust Estimation with Wrapped Models for Torus Data **C. Agostinelli**
12:40-13:00	Regime Switching Models for Directional and Linear Observations **A. C. Harvey**
13:00-13:20	A Cauchy-Type Model for Data on the Cylinder **A. Pewsey**
13:20-15:30	Lunch
15:30-16:30	Old town of Santiago de Compostela walking visit
16:30	Social Activities

Day 3 (17th June)

9:30-9:50	From Topology of the Data Space to a Compatible Probabilistic Model **K. Sargsyan**
9:50-10:10	On a Nonparametric Version of the Circulas **J. Ameijeiras-Alonso**
10:10-10:30	Clustering Using Principal Nested Stratified Spheres with Application to SARS-CoV-2 RNA Structure **S. Huckemann**
10:30-10:50	Bayesian Inference for the Skew Fisher-von Mises-Langevin Model on the Sphere **C. Ley**
10:50-11:10	A Resampling-Based Assessment of Rotational Symmetry in Orientation Data **U. Genschel**
11:10-11:40	Coffee-break
11:40-12:00	Graphical Models for Circular Variables **A. Panzera**
12:00-12:20	Statistical Models and the Benford Hypothesis: A Unifying Framework **M. Di Marzio**
12:20-12:40	An Application of Square Root Transformation for Optimal Prior Selection **A. Kume**
12:40-13:00	Some Directional Models for Tree Branching Patterns **L.-P. Rivest**
13:00-13:20	On the Circular Median Absolute Deviation **G. Porzio**
13:20-15:30	Lunch
15:30-15:50	Geological Features, Parallel Lines and Circular Statistics **P. E. Jupp**
15:50-16:10	Spatial Quantiles on the Hypersphere **D. Paindaveine**
16:10-16:30	Closing ceremony
21:00	Conference dinner

Abstracts: Invited Speakers

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Robust Estimation with Wrapped Models for Torus Data

C. Agostinelli[1,*], L. Greco[2] and G. Saraceno[1]

[1]*Department of Mathematics, University of Trento, Via Sommarive, 14, Trento, Italy; claudio.agostinelli@unitn.it, giovanni.saraceno@unitn.it*
[2]*University Giustino Fortunato, Viale Raffaele Delcogliano, 12, Benevento, Italy; l.greco@unifortunato.eu*
[*]*Corresponding author*

Abstract. Multivariate circular observations, i.e. points on a torus arise frequently in fields where instruments such as compass, protractor, weather vane, sextant or theodolite are used. Multivariate wrapped models are often appropriate to describe data points scattered on p-dimensional torus. While the concept of outliers for data in an Euclidean space is extensively discussed in the literature, the corresponding idea in p-dimensional torus data is presented only in few articles for the case of 1-dimensional torus data, aka, circular data, see e.g. [1] and the references therein. After showing the effect of outliers in estimating parameters using Maximum Likelihood in data on a p-dimensional torus, we introduce a general methods to construct robust estimators using the Weighted Likelihood approach in Wrapped Models. We compare their performance with robust procedures obtained in the more classical framework of MM-estimators and we provide computational details. Finally, we provide examples based on real data sets.

Keywords: Classification EM; Expectation-Maximization; Torus Data; Weighted Likelihood Estimators; Wrapped Models.

References

[1] Agostinelli, C. (2007). Robust estimation for circular data. *Computational Statistics and Data Analysis*, **51(12)**, 5867–5875.

[2] Greco, L., Saraceno, G. and Agostinelli, C. (2021). Robust Fitting of a Wrapped Normal Model to Multivariate Circular Data and Outlier Detection. *Stats*, **4(2)**, 454–471.

[3] Nodehi, A., Golalizadeh, M., Maadooliat, M. and Agostinelli, C. (2021). Estimation of parameters in multivariate wrapped models for data on a p-torus. *Computational Statistics*, **36(1)**, 193–215.

[4] Saraceno, G., Agostinelli, C. and Greco, L. (2021). Robust Estimation for Multivariate Wrapped Models. *Metron*, **79(2)**, 225–240.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

On a Nonparametric Version of the Circulas

J. Ameijeiras-Alonso[1,*] and I. Gijbels[2]

[1]*Department of Statistics, Mathematical Analysis and Optimization, CITMAga, Universidade de Santiago de Compostela, Spain; jose.ameijeiras@usc.gal*
[2]*Department of Mathematics and Leuven Statistics Research Center (LStat), KU Leuven, Belgium; irene.gijbels@kuleuven.be*
[*]*Corresponding author*

Abstract. In this talk, we will discuss how to estimate the circulas using kernel methods. Copulas are an important tool to study dependencies for data on the real line. The analogous version of copulas for data on the torus is the circulas. We will also discuss how to derive the asymptotic bias and variance of these nonparametric estimators. The derivation of different strategies to obtain the optimal smoothing parameter will be also discussed. We will provide a plug-in version of this smoothing parameter. Finally, we will discuss some interesting applications of this circula estimator.

Keywords: Copula; Kernel Estimation; Mean Squared Error; Optimal Smoothing Parameter; Toroidal Data.

Acknowledgments. Supported by the FWO research project G.0826.15N (Flemish Science Foundation), GOA/12/014 project (Research Fund KU Leuven). The first author was supported by Grant PID2020-116587GB-I00 funded by MCIN/AEI/10.13039/501100011033 and the Competitive Reference Groups 2021-2024 (ED431C 2021/24) from the Xunta de Galicia.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Statistical Models and the Benford Hypothesis: A Unifying Framework

L. Barabesi[1], A. Cerioli[2] and M. Di Marzio[3,*]

[1]*Department of Economics and Statistics, University of Siena, Italy; lucio.barabesi@unisi.it*
[2] *Department of Economics and Management and University Centre "Robust Statistics Academy" (Ro.S.A.), University of Parma, Italy; andrea.cerioli@unipr.it*
[3]*Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, "G. D'Annunzio" University, Pescara, Italy; marco.dimarzio@unich.it*
[*]*Corresponding author*

Abstract. The Benford hypothesis is the statement that a random sample is made of realizations of an absolutely-continuous random variable distributed according to Benford's law [1]. We establish the closeness of a particular model to the Benford hypothesis through a suitable Kolmogorov distance. We then adopt a basically semiparametric approach to compare two asymptotically equivalent and optimal test statistics. We show that the proposed test statistics are invariant under scale transformation of the data, a crucial requirement when compliance to the Benford hypothesis is used to corroborate scientific theories. The empirical advantage of the proposed tests is shown through an extensive simulation study. Application to astrophysical data also motivates the methodology.

Keywords: Characteristic Function; Digit Distribution; Kolmogorov Distance; Likelihood-ratio Test; Significand.

References

[1] Berger, A. and Hill, T. P. (2015). *An Introduction to Benford's Law*. Princeton University Press, Princeton.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Chordal-Based Tests of Uniformity on the Hypersphere

E. García-Portugués[1]

[1]*Department of Statistics, Carlos III University of Madrid; edgarcia@est-econ.uc3m.es.*

Abstract. We provide a general and tractable family of tests of uniformity on the hypersphere of arbitrary dimension. The family is constructed from powers of the chordal distances between pairs of observations. It connects and extends three particular tests: Rayleigh [4], Pycke [2, 3], and Bakshaev [1]. The asymptotic null distributions of the new tests are obtained and, despite involving infinite sums of weighted chi-squared random variables, are shown to be tractable in practice. Additionally, powers of the tests against generic local alternatives are provided. In particular, explicit powers against Cauchy-like distributions on the hypersphere are derived. Numerical experiments corroborate the obtained theoretical results. Two real data applications of astronomical and biological nature illustrate the practical use of the tests for assessing uniformity on the two-dimensional sphere.

Keywords: Hypersphere; Tests; Uniformity.

References

[1] Bakshaev, A. (2010). *N*-distance tests of uniformity on the hypersphere. *Nonlinear Analysis: Modelling and Control*, **15(1)**, 15–28.

[2] Pycke, J. R. (2007). A decomposition for invariant tests of uniformity on the sphere. *Proceedings of the American Mathematical Society*, **135(9)**, 2983–2993.

[3] Pycke, J. R. (2010). Some tests for uniformity of circular distributions powerful against multimodal alternatives. *The Canadian Journal of Statistics*, **38(1)**, 80–96.

[4] Lord Rayleigh (1919). On the problem of random vibrations, and of random flights in one, two, or three dimensions. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **37(220)**, 321–347.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

A Resampling-Based Assessment of Rotational Symmetry in Orientation Data

E. Biswas[1], U. Genschel[1,*] and D. J. Nordman[1]

[1]*Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A;*
ulrike@iastate.edu, ebiswas@iastate.edu, dnordman@iastate.edu
[*]*Corresponding author*

Abstract.

This talk describes a resampling method for assessing rotational symmetry in three-dimensional orientation data. Orientation data are of interest in a wide variety of fields, including human kinematics and materials science, where each observation can be represented by a 3×3 rotation matrix $\boldsymbol{O}$ in $SO(3)$ (i.e., the set of orthogonal matrices with determinant 1), denoting the orientation of some object in a three-dimensional coordinate system (cf. [1], [3], [4], [6], [11], [14], [16]); see [9] for an introduction.

In many applications with orientation data, rotationally symmetric or *isotropic* distributions are commonly used for basic modeling purposes, which serve to conceptualize the variability in an orientation $\boldsymbol{O} = \boldsymbol{SR}$ as due to a directionally symmetric random perturbation $\boldsymbol{R}$ of an underlying location parameter $\boldsymbol{S} \in SO(3)$. In this sense, symmetric distributions for orientations $\boldsymbol{O}$ resemble a standard location model $Y = \mu + \varepsilon$ for real-valued data using symmetrically distributed errors ε. The most common distribution on $SO(3)$ of this form is the isotropic version of the Matrix Fisher distribution (cf. [4], [7]). Other such models include the isotropic Cayley distribution ([8], sec. 5.2), Bunge's Gaussian distribution [2], the Lorentzian distribution [10], the de la Vallée Poussin distribution [15], the isotropic Gaussian distribution [13], the uniform distribution [12], and the wrapped trivariate normal distribution [17].

Because rotational symmetry serves as an important, though simplifying, property for model-based inference about orientation data, formal assessments of this assumption become important. While much research as focused on testing rotational symmetry in directional data (cf. [5]), the problem of assessing symmetry has received less consideration for orientations. This talk describes a general bootstrap-based procedure for formally testing the property of rotational symmetry in orientation data. Such distributional symmetry in a data point $\boldsymbol{O} \in SO(3)$ can be translated to, or characterized by, the mutual independence of three random variables $\boldsymbol{O} \leftrightarrow (Z_1, Z_2, Z_3) \in (0, \pi] \times [0, 2\pi] \times [-1, 1]$, where two random variables have known marginal distributions and the third variable has an unknown marginal distribution. The bootstrap procedure combines resampling, of both parametric and non-parametric forms, in order to "re-create" data with rotational symmetry (i.e., the null hypothesis). Estimation steps may also be used to account for an unknown location parameter $\boldsymbol{S} \in SO(3)$ (i.e., a composite null hypothesis) and incorporated into the boot-

strap as well. Empirical processes induced by the orientation data have complex limits, which are not distribution-free and additionally include further random components when the location parameter is estimated. The resampling-based testing approach is shown to capture the true (but unknown) sampling distribution of test statistics under the null hypothesis of rotational symmetry. The performance of the bootstrap-based testing method is evaluated through numerical studies, and the testing approach is illustrated with orientation data collected in texture analysis from materials science.

Keywords: Axis-angle Representation; Bootstrap; Projected Arithmetic Mean; Random Rotation; UARS Model.

References

[1] Bingham, M. A., Nordman D. J. and Vardeman, S. B. (2009). Modeling and Inference for Measured Crystal Orientations and a Tractable Class of Symmetric Distributions for Rotations in Three Dimensions. *J. Amer. Statist. Assoc.*, **104**, 1385-1397.

[2] Bunge, H. J. (1982). *Texture Analysis in Material Science*. Butterworth, London.

[3] Chang, T. (1998). Estimating the Relative Rotation of Two Tectonic Plates from Boundary Crossings. *J. Amer. Statist. Assoc.*, **83**, 1178-1183.

[4] Downs, T. D. (1972). Orientation Statistics. *Biometrika*, **59**, 665-676.

[5] García-Portugués, E., Paindaveine, D. and Verdeboutb, T. (2020). On Optimal Tests for Rotational Symmetry Against New Classes of Hyperspherical Distributions. *J. Amer. Statist. Assoc.*, **115**, 1873-1887.

[6] Hielscher, R., Schaeben, H. and Siemes, H. (2010). Orientation Distribution Within a Single Hematite Crystal. *Math. Geosci.*, **42**, 359-375.

[7] Khatri, C. G. and Mardia, K. V. (1977). The Von Mises-Fisher Matrix Distribution in Orientation Statistics. *J. Roy. Statist. Soc. Ser. B*, **39**, 95-106.

[8] León, C. A., Massé, J. C. and Rivest, L.-P. (2006). A Statistical Model for Random Rotations. *J. Multivariate Anal.*, **97**, 412-430.

[9] Mardia, K. V. and Jupp, P. E. (2009). *Directional Statistics*. Wiley, New York.

[10] Matthies, S. (1982). Form Effects in the Description of the Orientation Distribution Function (ODF) of Texturized Materials by Model Components. *Physica Status Solidi (b)*, **112**, 705-716.

[11] Matthies, S., Muller, J. and Vinel, G. W. (1988). On the Normal Distribution in the Orientation Space. *Textures Microstruct.*, **10**, 77-96.

[12] Miles, R. E. (1965). On Random Rotations in $\mathbb{R}^3$. *Biometrika*, **52**, 636-639.

[13] Nikolayev D. I. and Savyolova, T. I. (1997). Normal Distribution on the Rotation Group $SO(3)$. *Textures Microstruct.*, **29**, 201-233.

[14] Rancourt, D., Rivest, L.-P. and Asselin, J. (2000). Using Orientation Statistics to Investigate Variations in Human Kinematics. *J. Roy. Statist. Soc. Ser. C*, **49**, 81-94.

[15] Schaeben, H. (1997). The de la Vallée Poussin Standard Orientation Density Function *Textures and Microstructures* **33**, 365-373.

[16] Stanfill, B., Genschel, U., Hofmann, H. and Nordman, D. J. (2013). Point Estimation of the Central Orientation of Random Rotations. *Technometrics*, **55**, 524-535.

[17] Yu, Q., Nordman, D. J. and Vardeman, S. B. (2014). A Wrapped Trivariate Normal Distribution and Bayes Inference for 3-D Rotations. *Statist. Sinica*, **24**, 897-917.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

A Two-Sample Statistical Test for Directional Data

E. Gutiérrez-Peña[1,*]

[1]*Department of Probability and Statistics, IIMAS, UNAM. Apartado Postal 20-126, CP 01000, Ciudad de México, Mexico; eduardo@sigma.iimas.unam.mx*
[*]*Corresponding author*

Abstract. Testing whether two random samples are drawn from the same population is an imporant problem in statistics. For directional data this problem can be particularly challenging, with tests often relying on unrealistic parametric distributional assumptions. In this talk we will discuss a nonparametric two-sample statistical test for directional data. The test is based on a method proposed in [1] and makes use of pairwise distances between observations. The resulting test statistic can be efficiently computed even in high-dimensional problems.

Keywords: Distance Matrix; Equality of Distributions.

References

[1] Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society* B, **67(4)**, 515–530.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Regime Switching Models for Directional and Linear Observations

A. Harvey[1,*] and D. Palumbo[2]

[1]*Faculty of Economics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DD, UK; ach34@cam.ac.uk*
[2]*Ca' Foscari University of Venice, Department of Economics, Cannareggio 873, 30121 Venice, Italy. Homerton College, University of Cambridge, Hills Road, Cambridge CB2 8PH, UK; dp470@cam.ac.uk*
[*]*Corresponding author*

Abstract. The score-driven approach to time series modeling is described in [1] and [2]. It provides a solution to the problem of modeling circular data and it can also be used to model switching regimes with intra-regime dynamics. Furthermore it enables a dynamic model to be fitted to a linear and a circular variable when the joint distribution is a cylinder. The viability of the new method is illustrated by estimating a model with dynamic switching and dynamic location and/or scale in each regime to hourly data on wind direction and speed in Galicia.

Keywords: Circular Data; Conditional Score; Hidden Markov Model; Cylindrical Distribution; Wind.

References

[1] Creal, D., Koopman, S. J. and Lucas, A. (2013), Generalized autoregressive score models with applications. *Journal of Applied Econometrics,* **28**, 777-95.

[2] Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: with Applications to Financial and Economic Time Series*. Econometric Society Monograph, Cambridge University Press.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Clustering Using Principal Nested Stratified Spheres with Application to SARS-CoV-2 RNA Structure

B. Eltzner[1], S. F. Huckemann[2,*], K. V. Mardia[3] and H. Wiechers[2]

[1]*Max Planck Institute for Biophysical Chemistry, Göttingen, 37077, Germany; beltzne@uni-goettingen.de*
[2]*Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, Georgia-Augusta-University Göttingen, Goldschmidstr. 7, 37077 Göttigen, Germany; huckeman@math.uni-goettingen.de, henrik.wiechers@uni-goettingen.de*
[3]*Department of Statistics, School of Mathematics, University of Leeds, LS2 9JT, UK and Department of Statistics, University of Oxford, OX1 3LB, UK; K.V.Mardia@leeds.ac.uk*
[*]*Corresponding author*

Abstract. There is an abundance of clustering methods for Euclidean data based on principal component analysis (PCA). A while ago Jung, Dryden and Marron (2010) introduced principal nested spheres (PNS) as a PCA analog on spheres. Allowing for constant curvature components, rather than only for zero curvature components, PNS is more flexible than PCA. This is in particular useful for clustering methods involving mode hunting in projections to the first PCA component: We use instead circular mode hunting on the more flexible first PNS component. To make this method applicable to data on general manifolds, one can approximate them with, map them to, or transform them to spheres, possibly stratified spheres in order to preserve topological features. We illustrate the latter for RNA data, represented by dihedral angles on a microscopic scale, yielding torus-valued data, and by landmarks on a mesoscopic scale, yielding data on a size-and-shape space. As modern methods reconstructing spatial RNA structure occasionally produce unrealistically close atom positions, called clashes, learning classes of clash-free RNA both on microscopic and mesoscopic scale, we propose consistent corrections of the RNA backbone of SARS-CoV-2 at two sites of the frameshift stimulation element where previous methods could not resolve inconsistencies reconstructed by cryo-EM.

Keywords: Adaptive Linkage Clustering; Circular Mode Hunting; Dimension Reduction; Multiscale Learning; Torus PCA.

References

[1] Jung, S., Dryden, I. L. and Marron, J. S. (2010). *Analysis of principal nested spheres*. *Biometrika*, **99(3)**, 551–568.

[2] Mardia, K. V., Wiechers, H., Eltzner, B. and Huckemann, S. F. (2021). Principal component analysis and clustering on manifolds. *Journal of Multivariate Analysis*, **57(3)**, 104862

[3] Wiechers, H., Eltzner, B., Mardia, K. V. and Huckemann, S. F. (2021). Learning torus PCA based classification for multiscale RNA backbone structure correction with application to SARS-CoV-2. *bioRxiv*, 10.1101/2021.08.06.455406.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Geodesic Projection of the von Mises–Fisher Distribution for Projection Pursuit of Directional Data

S. Jung[1,*]

[1]*Seoul National University, Seoul 08826, Korea; sungkyu@snu.ac.kr*
[*]*Corresponding author*

Abstract. We investigate geodesic projections of von Mises–Fisher (vMF) distributed directional data. The vMF distribution for random directions on the $(p-1)$-dimensional unit hypersphere $\mathbb{S}^{p-1} \subset \mathbb{R}^p$ plays the role of multivariate normal distribution in directional statistics. For one-dimensional circle $\mathbb{S}^1$, the vMF distribution is called von Mises (vM) distribution. Projections onto geodesics are one of main ingredients of modeling and exploring directional data. We show that the projection of vMF distributed random directions onto any geodesic is approximately vM-distributed, albeit not exactly the same. In particular, the distribution of the geodesic-projected score is an infinite scale mixture of vM distributions. Approximations by vM distributions are given along various asymptotic scenarios including large and small concentrations ($\kappa \to \infty, \kappa \to 0$), high-dimensions ($p \to \infty$), and two important cases of double-asymptotics ($p, \kappa \to \infty$, $\kappa/p \to c$ or $\kappa/\sqrt{p} \to \lambda$), to support our claim: geodesic projections of the vMF are approximately vM. As one of potential applications of the result, we contemplate a projection pursuit exploration of high-dimensional directional data. We show that in a high dimensional model almost all geodesic-projections of directional data are nearly vM, thus measures of non-vM-ness are a viable candidate for projection index.

Keywords: Projection Index; von Mises Distribution.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Geological Features, Parallel Lines and Circular Statistics

P. E. Jupp[1,*], I. B. J. Goudie[1], R. A. Batchelor[2,†] and R. J. B. Goudie[3]

[1]*School of Mathematics and Statistics, University of St Andrews KY16 9SS, U.K.; pej@st-andrews.ac.uk, ig@st-andrews.ac.uk*
[2]*School of Earth and Environmental Sciences, University of St Andrews KY16 9AL, U.K.*
[3]*MRC Biostatistics Unit, University of Cambridge, School of Clinical Medicine, Cambridge CB2 0SR, U.K.; robert.goudie@mrc-bsu.cam.ac.uk*
[*]*Corresponding author*
[†]*Died 15 February 2022*

Abstract. Some planes in sedimentary rocks contain features that appear to lie near equally spaced parallel lines. Determining whether or not they do so can provide information on possible mechanisms for their formation. The problem is recast here in terms of circular statistics, enabling closeness of candidate sets of lines to the points to be measured by a mean resultant length. This leads to models that are higher-dimensional versions of the quantal models introduced in [1] to hunt for *quanta*, standard units of length, of which measured lengths were thought to be multiples. For our geological problem, we develop tests of goodness of fit and estimates of the direction of the lines and of the spacing between them.

Keywords: Quantal Model; Mean Resultant Length.

References

[1] Kendall, D. G. (1974). Hunting quanta. *Phil. Trans. Roy. Soc. A*, **76**, 231–266.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

A Copula Model for Trivariate Circular Data

S. Kato[1,*] and C. Ley[2]

[1]*Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan; skato@ism.ac.jp*
[2]*Department of Mathematics, University of Luxembourg, 6, rue de la Fonte, 4365 Esch-sur-Alzette, Luxembourg; christophe.ley@uni.lu*
[*]*Corresponding author*

Abstract. We propose a new family of distributions for trivariate circular data. Its density can be expressed in simple form without involving infinite sums or integrals. The univariate marginals of the proposed distributions are the uniform distributions on the circle, and therefore the presented family is considered a copula for trivariate circular data. The bivariate marginals of the proposed distributions are members of the family of [2]. The univariate and bivariate conditional distributions are the wrapped Cauchy distributions and the distributions of [1], respectively. An efficient algorithm is presented to generate random variates from our model. Maximum likelihood estimation for the presented distributions is considered. An extension of the proposed family for multivariate circular data is briefly discussed.

Keywords: Circula; Multivariate Circular Data; Random Variate Generation; Torus; Wrapped Cauchy Distribution.

References

[1] Kato, S. and Pewsey, A. (2015). A Möbius transformation-induced distribution on the torus. *Biometrika*, **102(2)**, 359–370.

[2] Wehrly, T. E. and Johnson, R. A. (1980). Bivariate models for dependence of angular observations and a related Markov process. *Biometrika*, **67(1)**, 255–256.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Directional Distributions and the Half-Angle Principle

J. T. Kent[1,*]

[1] *University of Leeds, Leeds LS2 9JT, UK; j.t.kent@leeds.ac.uk*
[*] *Corresponding author*

Abstract. Two closely-related distributions on the circle are the wrapped Cauchy (WC) and the angular central Gaussian (ACG) distributions. The former can be obtained from the latter by angle doubling; equivalently, the latter can be obtained from the former by angle-halving. This close relationship means that statistical properties for one distribution carry over with little change to the other distribution. Sometimes the connections are obvious, but sometimes they are more subtle. More details can be found in [3]. Here are some of the key results.

1. Angle doubling and halving. The mapping $\phi \to \theta = 2\phi$ is a two-to-one mapping of the circle to itself. The reverse mapping $\theta \to \phi = \{\theta/2, \theta/2 + \pi\}$ is a one-to-two mapping of the circle to itself. Note that $\phi \sim$ ACG corresponds to $\theta \sim$ WC.

2. Transformation groups. A natural group of transformations on the circle is given by the re-scaled linear transformations. Given a nonsingular 2×2 matrix A, a unit vector $\boldsymbol{u} = (\cos\phi,\ \sin\phi)^T$ is taken to $A\boldsymbol{u}/||A\boldsymbol{u}||$. The family of ACG distributions is closed under the group of re-scaled linear transformations. Another natural group of transformations on the circle is given by the Möbius transformations. The family of WC distributions is closed under the group of Möbius transformations. Further, the two groups can be identified with one another through an unexpected Möbius identity.

3. Projections from the circle to the line. Two related projections are gnomonic and stereographic projection. The gnomonic projection of ϕ is the same as the stereographic projection of θ. Hence gnomonic projection of the ACG distribution is the same as stereographic projection of the WC distribution, and in both cases the projected distribution turns out to be the Cauchy distribution on the line. An estimation method for the parameters of one distribution (ACG, WC, or Cauchy) can be adapted immediately to estimate the parameters of the other two distributions. Several algorithms to compute maximum likelihood estimates based on EM algorithms have been explored in [4, 5, 1].

4. Extensions to the unit sphere in $\mathbb{R}^q$, $q > 2$. The results on the circle carry over to only a limited extent to higher-dimensional spheres. The ACG distribution has an immediate extension, which under gnomonic projection corresponds to a multivariate Cauchy distribution. This is different from a recent extension of the WC distribution [2], which under stereographic projection

corresponds to a multivariate t-distribution with $q-1$ degrees of freedom.

5. Practical implications. The WC distributions and ACG are often used in robustness studies in an analogous manner to the Cauchy distribution on the line [6, 7]. The Möbius transformation is often used to construct a link function for regression models involving angular variables as explanatory and response variables.

Keywords: Angular Central Gaussian Distribution; Gnomonic Projection; Möbius Transformation; Stereographic Projection; Wrapped Cauchy Distribution.

References

[1] Arslan, O., Constable, P. D. L. and Kent, J. T. (1995). Convergence behaviour of the EM algorithm for the multivariate t-distribution. *Commun. Stat.: Theor. Methods*, **24**, 2981–3000.

[2] Kato, S. and McCullagh, P. (2020). Some properties of a Cauchy family on the sphere derived from Möbius transformation. *Bernoulli*, **26**, 3224–3248.

[3] Kent, J. T. (2022) Directional distributions and the half-angle principle. arXiv:2202.06611 [math.ST].

[4] Kent, J. T. and Tyler, D. E. (1988). Maximum likelihood estimation for the wrapped Cauchy distribution. *J. Appl. Stat.*, **15**, 247–254.

[5] Kent, J. T., Tyler, D. E. and Vardi, Y. (1994). A curious likelihood identity for the multivariate t-distribution. *Commun. Stat.: Simul. Comput.*, **23**, 441–453.

[6] Tyler, D. E. (1987a). A distribution-free M-estimator of multivariate scatter. *Ann. Stat.*, **15**, 234–251.

[7] Tyler, D. E. (1987b). Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika*, **74**, 579–589.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

An Application of Square Root Transformation for Optimal Prior Selection

A. Kume[1,*], C. Villa[2] and S. G. Walker[3]

[1]*School of Mathematics, Statistics and Actuarial Science; University of Kent; a.kume@kent.ac.uk*
[2]*School of Mathematics, Statistics and Physics Newcastle University; cristiano.villa@ncl.ac.uk*
[3]*Department of Mathematics, Department of Statistics and Data Sciences, University of Texas at Austin; s.g.walker@math.utexas.edu*
[*]*Corresponding author*

Abstract. The pooling of opinions is a big area of research and has been for a number of decades [1, 2]. The idea is to obtain a single belief probability distribution from a set of expert opinion belief distributions. The paper proposes a new way to provide a resultant prior opinion based on the optimal information among all possible linear combinations of the prior densities, including negative components. This is done in the square-root density space which is identified with the positive orthant of Hilbert unit sphere of differentiable functions [3]. It can be shown that the optimal prior is easily identified as an extrinsic mean in the sphere. For distributions belonging to the exponential family the resulting calculations do not require numerical integration and can be immediately implemented in the Bayesian analysis. The idea can also be adopted for any neighbourhood of a chosen base prior and spanned by a finite set of "contaminating" directions.

Keywords: Fisher Information; Expert Opinion; Hilbert Sphere; Maximum Entropy Prior Distributions.

References

[1] Albert, D., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, A. Mengersen, K. and Rousseau, J. (2012). Combining Expert Opinions in Prior Elicitation. *Bayesian Analysis*, **7(3)**, 503-532.

[2] Genest, C. and Zidek, J. V. (1986). Combining probability distributions. A critique and annoted bibliography. *Statistical Science*, **1**, 114-148.

[3] Kurtek, S. and Bharath, K. (2015). Bayesian sensitivity analysis with the Fisher-Rao metric. *Biometrika*, **102(3)**, 601-616.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

A Hidden Markov Random Field with Copula-Based Emission Distributions for the Analysis of Spatial Cylindrical Data

F. Lagona[1,*]

[1] *University of Roma Tre - via G. Chiabrera 199 00145 Rome Italy; francesco.lagona@uniroma3.it*
[*] *Corresponding author*

Abstract. Cylindrical spatial series are bivariate vectors of angles and intensities that are simultaneously observed at a number of sites in an area of interest. These data arise frequently in environmental and ecological studies. Examples include hurrican wind satellite data, wave directions and heights that are generated by deterministic wave models, speeds and directions of marine currents recorded by a network of high-frequency radars, as well as telemetry data of animal movement.

The analysis of cylindrical spatial series is complicated by the special topology of the support on which the measurements are taken (the cylinder), and by the difficulties in modeling the cross-correlations between angular and linear measurements across space. Additional complications arise from the multimodality of the marginal distribution of the data, which are often observed under heterogeneous, space-varying conditions.

I describe a cylindrical hidden Markov random field (MRF) model that parsimoniously accounts for the specific features of cylindrical spatial series. More precisely, the data distribution is approximated by a mixture of copula-based cylindrical densities, whose parameters vary across space according to a latent Potts model. The Potts model [6] is a categorical MRF, i.e. a multinomial process in discrete space, which fulfills a spatial Markovian property: the conditional distribution at each site given the rest of the field is independent of the field values outside a neighborhood of the site. It segments an area of interest according to an interaction parameter that captures the correlation between adjacent observations and controls the smoothness of the segmentation.

Cylindrical hidden MRFs have been already proposed in the literature [5], by exploiting the Abe-Ley density [1] as emission distribution. The Abe-Ley density is a five-parameter bivariate density on the cylinder. A mixture of Abe-Ley densities therefore provides a distributional extension to allow for multimodal cylindrical data. Assuming that the mixture parameters vary according to the segmentation provided by a Potts MRF is a further extension to capture unobserved spatial heterogeneity and to allow for spatial correlation.

These proposals are here extended by considering copula-based cylindrical densities [3]. Copulas allow the marginal densities and the joint dependence structure to be modeled separately. As a result, they provide a general method for binding any pair of univariate marginal distributions together to form a bivariate distribution. This is particularly advantageous in the cylindrical setting,

because a copula can be exploited to bind two marginal densities that do not necessarily have the same support. In this work, we take this approach by binding a Weibull and a circular wrapped Cauchy together to form a cylindrical density. However, this proposal can be promptly adapted with different marginal densities, if desired.

Hidden MRFs are popular models in spatial statistics, since the seminal paper by Besag [2]. They can be seen as an extension of hidden Markov models, exploited in time series analysis, to the spatial setting. Hidden Markov models have been recently proposed for the analysis of cylindrical time series [4]. This paper extends this approach to the analysis of cylindrical spatial series.

Special computational issues arise in the estimation of the parameters of the proposed cylindrical hidden MRF model. When the spatial interaction parameter of the Potts model is equal to zero, the cylindrical hidden MRF reduces to a latent class model for independent cylindrical data and a standard Expectation-Maximization (EM) algorithm can be exploited for likelihood maximization. EM algorithms are based on the definition of a complete-data likelihood function and, under regularity conditions, provide a sequence of estimates that converges to a local maximum of the likelihood function by iteratively updating and maximizing the expected value of the complete-data log-likelihood function. When, however, the interaction parameter of the Potts model is not equal to zero, the computation of the expected complete-data log-likelihood is unfeasible and special approximation strategies are needed. Extending the composite likelihood methods proposed in [5], I describe a numerically efficient EM algorithm for estimating the parameters of the cylindrical hidden MRF.

These methods are finally illustrated on a vector field of sea currents in the Adriatic sea.

Keywords: Composite Likelihood; Copula; Cylindrical Data; Hidden Markov Random Field.

References

[1] Abe, T. and Ley, C. (2017). A tractable, parsimonious and flexible model for cylindrical data, with applications. *Econometrics and Statistics*, **4**, 91-104.

[2] Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, **48**, 259-302.

[3] Jones, M. C., Pewsey, A. and Kato, S. (2015). On a class of circulas: copulas for circular distributions. *Annals of the Institute of Statistical Mathematics*, **67**, 843-862.

[4] Lagona, F., Picone, M. and Maruotti, A. (2015). A hidden Markov model for the analysis of cylindrical time series. *Environmetrics*, **26**, 534-544.

[5] Ranalli, M., Lagona, F., Picone, M. and Zambianchi, E. (2018). Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. *Journal of the Royal Statistical Society C*, **67**, 575-598.

[6] Strauss, D. J. (1977). Clustering on coloured lattices. *Journal of Applied Probability*, **14**, 135-143.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

The Circular Structure of Oscillatory Signals: Applications with CPCA

Y. Larriba[1,*], A. Rodríguez-Collado[1] and C. Rueda[1]

[1]*Department of Statistics and Operations Research, University of Valladolid; yolanda.larriba@uva.es*
[*]*Corresponding author*

Abstract. Oscillatory systems govern many biological processes as for example, the molecular clock networks that drive tissue-specific circadian gene expressions [1]; or the neuronal dynamics orchestrated by action potential (AP) curves, that measure the fluctuation of the potential of a neuron [2]. Both circadian gene expressions and APs display up-down-up patterns periodically. Data analysis from oscillatory systems is challenging as it generally involves several noise sources and very heterogeneous patterns. The inherent circular nature of oscillatory signals enables a suitable mathematical formulation to efficiently deal with these problems. Specifically, this work addresses the circular order identification problem to solve two different problems associated with genomic data analysis involving oscillatory signals: the development of a human atlas of circadian gene expressions and a neuronal cell-type (Cre Lines) taxonomy in the mouse brain. For the circadian atlas, gene expression sample collection times are unknown, and the circular order provides timing estimates. While, for the neurons, the circular order identification provides an order among the cells that allows a visualization of the neuronal taxonomy.

The intrinsic link between oscillatory signals and circular geometry has been deeply analyzed by our research group. It is demonstrated in [3] that a circular signal $\mu(t)$, $t \in [0,2\pi)$, describing an up-down-up pattern, within the Euclidean space can be equivalently formulated, within the Circular space, as a circular signal $\phi(t)$ where $\mu(t) = \cos(\phi(t))$ and $\phi(t) \preceq \phi(t')$, $0 \leq t \leq t' < 2\pi$. A circular signal $\phi(t)$ is characterized as it preserves the order among $t \in [0,2\pi)$. Moreover, it is shown in [4] that the Frequency Modulated Möbius (FMM) model is well suited for the analysis of oscillatory signals. In particular, the FMM model is the simplest parametric model for circular signals. It is based on the Möbius link function [5] and describes a circular signal as follows:

$$\phi(t) = \beta + 2\arctan\left(\omega \tan\left(\frac{t-\alpha}{2}\right)\right),\ t \in [0,2\pi);$$

where $\omega \in [0,1]$ is a kurtosis parameter, while α and β are angular parameters of location and skewness, respectively.

In practice, the order among the times ($t \in [0,2\pi)$) may be unknown, as is the problem the circadian atlas below. If so, the first is to solve the circular order identification problem to provide timing estimates. Once the order among the samples is known, as usually happens in other applications, the angular parameters derived from the FMM fitting can be used also to derive circular orders,

which is very useful to derive the order among the components of the oscillatory system, as is shown in the applications below for the genes and or the cells.

There exist algorithms in the literature, mainly focused on genomics, to solve ordering identification. As discussed in [1], they are usually complex, *ad-hoc* or work as a black-box. Together with these procedures, there is the Circular Principal Component Analysis (CPCA), a nonlinear dimensionality reduction method, that also describes the potential circular structure of the data by its projection onto curves that are constrained to lie on the unit circle [6]. CPCA is a simple and powerful tool to solve circular order identification problem, as is shown in the applications below.

Application 1: The human circadian atlas is derived from the GTEx database, a collection of $17,382$ unordered post-mortem RNA-seq gene expression samples from 948 donors across 54 human tissues. For each tissue, CPCA is applied to estimate the temporal order among samples. Then, the FMM model is fitted to gene expression data to identify molecular rhythms, to estimate gene activation times (peak times), which also generates an order among the genes, and to describe molecular clock networks. Results are tissue-specific, and an in-deep analysis of them draws the first circadian gene expression atlas across human organs including tissue-specific top rhythmic genes, activation times, and peak phase relationships.

Application 2: The taxonomy of mouse neurons is obtained from mice APs including in the Allen Cell Type database. Specifically, a total of $1,892$ experiments from mouse cells of 24 different Cre lines have been analyzed. At the class level, Cre lines can be either glutamatergic or GABAergic. First, at the neuronal level, a multicomponent FMM model is applied to extract from the experiments electrophysiological features: amplitude, width, shape, or peak times, that characterize Cre lines types. Next, such electrophysiological and additional genetic markers are joined, and CPCA is conducted on them to define the neuronal taxonomy that provides the order and circular distances among the Cre lines. For the first time, the taxonomy has an intrinsic circular topology and allows to locate Cre lines types.

Keywords: FMM Model; CPCA; Oscillatory Signal; Circadian Gene Expressions; AP Curves.

Acknowledgments. The authors gratefully acknowledge the financial support received by the Spanish Ministry of Science, Innovation and Universities [PID2019-106363RB-I00 to C.R., Y.L. and A.R-C].

References

[1] Larriba, Y. and Rueda, C. (2022). CIRCUST: a novel methodology for reconstruction of temporal order of molecular rhythms; validation and application towards a human circadian gene expression atlas. *Preprint*.

[2] Rodríguez-Collado, A. and Rueda, C. (2021). Electrophysiological and Transcriptomic Features Reveal a Circular Taxonomy of Cortical Neurons. *Frontiers in Human Neuroscience*, **15**, 684950.

[3] Larriba, Y. and Rueda, C. and Fernández, M. A. and Peddada, S.D. (2019). Order restricted inference

in chronobiology. *Statistics in medicine*, **39(3)**, 265-278.

[4] Rueda, C. and Larriba, Y. and Peddada, S.D. (2019). Frequency Modulated Möbius Model Accurately Predicts Rhythmic Signals in Biological and Physical Sciences. *Scientific Reports*, **9**, 18701.

[5] Downs, T. D. and Mardia, K. V. (2002). Circular regression. *Biometrika*, **89(3)**, 683-697.

[6] Scholz, M. (2007). Analysing Periodic Phenomena by Circular PCA. *Proceedings of the Conference on Bioinformatics Research and Development*, **4414**, 38-47.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Bayesian Inference for the Skew Fisher-von Mises-Langevin Model on the Sphere

N. Nakhaei Rad [1,2,3], A. Bekker [3], M. Arashi [4,3] and C. Ley [5,*]

[1] *Department of Mathematics and Statistics, Mashhad Branch, Islamic Azad University, Mashhad, Iran.*
[2] *DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS), South Africa.*
[3] *Department of Statistics, University of Pretoria, Pretoria 0002, South Africa.*
[4] *Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Iran.*
[5] *Department of Mathematics, Université du Luxembourg, Esch-sur-Alzette, Luxembourg, L-4364.*
[*] *Corresponding author*

Abstract. [2] have introduced a skewed version of the famous Fisher-von Mises-Langevin (FvML) distribution on the unit sphere. After briefly discussing frequentist inferential aspects associated with this model, we will delve into its Bayesian inference. In particular, we shall consider various priors for the parameters, and we will show how the similarities/differences between priors can be quantified using the Wasserstein Impact Measure of [1]. For the computation of the posterior distributions, modifications of Gibbs and slice samplings are applied for generating samples. We will illustrate our approach on various real data sets. Finally, if time permits, we shall give an overview on Bayesian inference for directional models, starting from [3].

Keywords: Fisher-von Mises-Langevin Distribution; Prior Impact Measure; Skew-Rotationally-symmetric Distributions; Spherical Data; Wasserstein Impact Measure.

References

[1] Ghaderinezhad, F., Ley, C., and Serrien, B. (2022). The Wasserstein Impact Measure (WIM): a practical tool for quantifying prior impact in Bayesian statistics. *Computational Statistics & Data Analysis*, in press.

[2] Ley, C. and Verdebout, T. (2017). Skew-rotationally-symmetric distributions and related efficient inferential procedures. *Journal of Multivariate Analysis*, **159**, 67–81.

[3] Mardia, K. V. and El-Atoum, S. (1976). Bayesian inference for the von Mises-Fisher distribution. *Biometrika*, **63**, 203–206.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Statistics of Discrete Distributions on Manifolds: A Journey from the Karl Pearson Roulette Wheel Data to Some Smart Health Science Data

K. V. Mardia[1,2,*] and K. Sriram[3]

[1] *University of Leeds, Leeds LS2 9JT, UK; k.v.mardia@leeds.ac.uk*
[2] *University of Oxford, Oxford OX1 3LB, UK*
[3] *Indian Institute of Management Ahmedabad, India; karthiks@iima.ac.in*
[*] *Corresponding author*

Abstract. Currently, certain novel problems related to discrete distributions on manifolds are arising. Their statistical roots go back to Karl Pearson [2, 3] who, over one hundred years ago, analysed roulette wheel data from the famous Monte Carlo casino which he later used as an illustration in his seminal chi-square paper of 1900. However, at that time there were no methodologies available for analysing directional data and so, not surprisingly, he linearised the problem and constructed a test for unbiasedness of the roulette wheel data transformed to the line. The trend to linearize the problem has continued, as can be seen from the subsequent related literature.

Recently, new discrete circular data have emerged in the Smart Health Sciences, such as acrophase data for monitoring blood pressure. Motivated by these and other practical applications, we construct, in a unified way, four rich families of discrete circular distributions, based on: maximum entropy, centered wrapping, and marginalizing and conditioning circular distributions. We also consider extensions to some other manifolds. We examine in detail two families: conditional discrete (CD) and marginal discrete (MD) distributions. Some of these families are deduced from established distributions such as the von Mises and wrapped Cauchy. Others are derived directly, such as a flexible family based on trigonometric sums and the circular location family. Results relating these families to one another are discussed. Such distributions have already been studied on the line, but not in any unified way, and the talk will highlight a far-reaching characterization for when CD and MD distributions are identical on the line. This challenging characterization has not yet appeared in the literature, and we give its circular counterpart. In particular, we examine the key properties of the CD and MD distributions for the von Mises and the wrapped Cauchy distributions, including the maximum likelihood estimators of their parameters. We show how to test the hypothesis of uniformity using these distributions and examine how to determine a change-point when the data arise as a sequence: for example, streaming data from roulette wheel spins. We also consider how to fit mixtures of such distributions.

The problem of model misspecification is examined when using a continuous distribution to model circular data that are discrete. In addition to simulation studies, we illustrate the problem using acrophase data which shows how using continuous circular models for discrete circular data can lead to misleading results. Our overall paradigm is "if one has discrete circular data then one

should use a discrete circular model". On the other hand, if the discrete data arise from grouped continuous data, Sheppard correction can work in some cases, but the "loss" due to improvised inference can only be assessed after appropriate discrete modelling, which serves as a benchmark.

The talk will focus on the case of discrete circular distributions but the results are far-reaching as they apply to any manifold, with extensions including irregular lattice support, families with skewness and kurtosis, and families on the torus. It is expected that this new statistical methodology will pave the way for many further developments and this talk gives a very brief glimpse of our research manuscript [1] on this topic.

An early version of this work was presented at the open access Leeds Annual Statistics Research (LASR) Workshop 2019 and, very recently, at the IMS meeting "Interactions of Statistics and Geometry 2022" in Singapore.

Keywords: Conditioned Families; Marginalized Families; Roulette Wheel Data; Acrophase Data; von Mises Distribution; Wrapped Cauchy Distribution.

References

[1] Mardia, K. V. and Sriram, K. (2022) Families of discrete circular distributions with some novel applications. Arxiv: https://arxiv.org/abs/2009.05437v2.

[2] Pearson, K. (1894). Science and Monte Carlo. *The Fortnightly Review*, new series, **55**, 183–193.

[3] Pearson, K. (1897). *The Scientific Aspect of Monte Carlo Roulette. The Chances of Death and Other Studies in Evolution*. Vol **1**, chapter II, pages 42–62. London: Edward Arnold.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Bivariate Linear-Circular Panel Data Modelling with Flexible Random Effects

A. Maruotti[1,*] and P. Alaimo Di Loro[1]

[1]*Dipartimento di Giurisprudenza, Economia, Politica e Lingue Moderne – Libera Università Maria Ss Assunta; a.maruotti@lumsa.it, p.alaimodiloro@lumsa.it*
[*]*Corresponding author*

Abstract. Due to the substantial progress in tracking technology, recent years have seen an explosion in the amount of movement data being collected. This has led to a huge demand for statistical tools that allow ecologists to draw meaningful inference from large tracking data sets.

Here, we introduce a bivariate bidimensional mixed-effects regression model for linear-circular outcomes, motivated by the analysis of animal movements. The model is able to capture heterogeneity across animals and allows for a full association structure among outcomes, assuming a discrete distribution for the random terms, with a possibly different number of support points in each univariate profile.

Formally, let the analyzed sample count n statistical units, i.e. the animals. One linear and one circular outcomes, y_{it1} and y_{it2} respectively, along with two vectors of covariates $\mathbf{x}'_{itj} = (1, x_{itj1}, \ldots, x_{itjP_j})$ and $\mathbf{z}'_{itj} = (1, z_{itj1}, \ldots, z_{itjQ_j})$, $j = 1,2$, which can vary over outcomes, are recorded for each unit i ($i = 1,2,\ldots,n$) at each time t ($t = 1,2,\ldots,T$). We assume that y_{itj} are realizations of conditionally independent random variables, with parameters $\boldsymbol{\theta}_{itj} = (\theta_{itj1}, \theta_{itj2}, \ldots, \theta_{itjM})$.

To overcome some of the limitations associated with the popular generalized linear mixed models, we relax the exponential family distribution assumption for the linear outcome, and replace it by a general distribution family including highly skew and/or kurtotic continuous distributions. Therefore, we are able to properly model excess of positive or negative kurtosis and/or skewness. The circular outcome is instead modelled via a proper, e.g. von Mises and/or Projected Normal, distribution.

A standard way to induce dependence among responses is to assume that they share some common latent structure. Thus, the model specification is completed by connecting the J univariate submodels through a set of correlated random effects $\mathbf{u}_i = (\mathbf{u}_{i1}, \mathbf{u}_{i2})$ which account for potential heterogeneity among statistical units and correlation between outcomes. In a regression setting, the interest is usually focused upon the mean which is modeled through a linear mixed model, providing a very broad framework for modeling dependence in the data [3]. In the following, we leave the distribution of the random effect completely unspecified and invoke the non-parametric maximum likelihood approach. Moreover, we consider a bivariate bidimensional latent structure such

that the independence model is nested in the multivariate one, and different levels of heteorgeneity in the two univariate profiles can be identified.

According to model assumptions, the likelihood function in the bivariate ($J = 2$) case is given by

$$L(\cdot) = \prod_{i=1}^{n} \left\{ \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \pi_{k_1 k_2} \prod_{j=1}^{2} \prod_{t=1}^{T} f(y_{it\,j} | \mathbf{x}_{it\,j}, \mathbf{z}_{it\,j}, \mathbf{u}_{i1} = \mathbf{u}_{k_1}, \mathbf{u}_{i2} = \mathbf{u}_{k_2}) \right\} \tag{1}$$

where $\pi_{k_1 k_2} = Pr(\mathbf{u}_{i1} = \mathbf{u}_{k_1}, \mathbf{u}_{i2} = \mathbf{u}_{k_2})$ is the joint probability associated to each couple of locations $(\mathbf{u}_{k_1}, \mathbf{u}_{k_2})$. The following constraints hold $\sum_{k_1=1}^{K_1} \pi_{k_1} = \sum_{k_2}^{K_2} \pi_{k_2} = \sum_{k_1 k_2} \pi_{k_1 k_2} = 1$
with

$$\pi_{k_1} = Pr(\mathbf{u}_{i1} = \mathbf{u}_{k_1}) = \sum_{k_2=1}^{K_2} \pi_{k_1 k_2}$$

and

$$\pi_{k_2} = Pr(\mathbf{u}_{i2} = \mathbf{u}_{k_2}) = \sum_{k_1=1}^{K_1} \pi_{k_1 k_2}.$$

We would remark that the number of locations (i.e. mixture components) may vary between outcomes. Thus, we control for heterogeneity in the univariate profiles and for the association between latent effects in the two profiles. This approach results in a finite mixture with $K_1 \times K_2$ components, in which each of the K_1 locations are coupled with each of the K_2 locations of the second outcome. If $J = 1$, our proposal reduces to a univariate finite mixture model. We would further remark that dependence between the random effects, as measured for example by the correlation coefficient, could be easily computed, though it may be quite biased due to the reduced number of locations. Moreover, our proposal adopts a parameterization for the mixture probabilities, where the dependence model properly nests the independence one.

The proposal extends the works of [1] and [2].

Keywords: Longitudinal Data; Finite Mixtures; Linear-Circular Data; Mixed Effects Regression.

Acknowledgments. We would like to thank the Organizing Committee for the support for celebrating the ADISTA 22.

References

[1] Maruotti, A. (2016). Analyzing longitudinal circular data by projected normal models: a semiparametric approach based on finite mixture models. *Environmental and Ecological Statistics*, **23(2)**, 257-277.

[2] Rivest, L.-P. and Kato, S. (2019). A random effects model for clustered circular data. *Canadian Journal of Statistics*, **47(4)**, 712–728.

[3] Verbeke, G., Fieuws, S., Molenberghs, G. and Davidian, M. (2014). The analysis of multivariate longitudinal data: a review. *Statistical methods in medical research*, **23(1)**, 42–59.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Spatial Quantiles on the Hypersphere

D. Konen[1] and D. Paindaveine[1,*]

[1]*ECARES and Department of Mathematics, Université libre de Bruxelles, Avenue F.D. Roosevelt, 50, ECARES, CP114/04, B-1050, Brussels, Belgium; Dimitri.Konen@ulb.be, Davy.Paindaveine@ulb.be*
[*]*Corresponding author*

Abstract. A concept of quantiles for distributions on the unit hypersphere $\mathbb{R}^d$ is proposed. The innermost quantiles are Fréchet medians, i.e., the L_1-analog of Fréchet means. Since these medians may be non-unique, we define a quantile field around each such median m. The corresponding quantiles are directional in nature: they are indexed by a scalar order between 0 and 1 and by a unit vector in the tangent space to the hypersphere at the median. To ensure computability in any dimension, our quantiles are essentially obtained by considering the Euclidean Chaudhuri [1] spatial quantiles in a suitable stereographic projection of the hypersphere onto its tangent space at the median. Despite this link with their Euclidean antecedent, studying our quantiles requires understanding the nature of the Chaudhuri quantile in a version of the projective space where all points at infinity are identified. Parallel to what is done in [2] in the Euclidean case, we thoroughly investigate the properties of the proposed quantiles, and study in particular the asymptotic behaviour of their sample versions, which requires controlling the impact of estimating the median. Our spherical quantile concept also allows for companion concepts of ranks and depth on the hypersphere.

Keywords: Centrality Regions; Directional Statistics; Multivariate Quantiles; Spatial Quantiles; Statistical Depth.

References

[1] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, **57**, 862–872.

[2] Konen, D., and Paindaveine, D. (2022). Multivariate ρ-quantiles: a spatial approach. *Bernoulli*, **28**, 1912–1934.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Graphical Models for Circular Variables

A. Gottard[1] and A. Panzera[1,*]

[1]*Department of Statistics, Computer Science, Applications. University of Florence; anna.gottard@unifi.it, agnese.panzera@unifi.it*
[*]*Corresponding author*

Abstract. Graphical models are a key class of probabilistic models for studying the conditional independence structure of a set of random variables. Despite their potential, graphical models for angular variables seem to be under-studied. Following ([1]), we explore three probability distributions defined according to the main approaches used for specifying distributions on the $p-$dimensional torus, in terms of conditional independence and graphical models. Regarding the *intrinsic approach*, we discuss the multivariate von Mises distribution. In particular, after pointing out some aspects of this distribution as an undirected graph model, we propose a related class of directed acyclic graph models useful when a natural ordering of the angles is known *a priori*. Regarding both the *wrapped* and the *embedded approaches*, we introduce two classes of undirected graphical models, which are related, in different manners, to the *classical* Gaussian graphical model. Specifically, the first class is based on the Wrapped Normal distribution, while the second is based on the Inverse Stereographic projected Normal distribution. For these classes of undirected graphical models, we discuss issues and possible solutions, including also some more flexible models where the distributional assumptions are relaxed. The potential usefulness of the proposed classes of models is shown by modelling the conditional independence among dihedral angles characterizing the three-dimensional structure of some proteins.

Keywords: Conditional Independence; Multivariate Circular Distributions; Protein Folding Problem; Toroidal Data.

References

[1] Gottard, A. and Panzera, A. (2021). Graphical models for circular variables. *arXiv preprint*. arXiv: 2104.03194.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

A Cauchy-Type Model for Data on the Cylinder

S. Kato[1] and A. Pewsey[2,*]

[1] *Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan; skato@ism.ac.jp*
[2] *Department of Mathematics, Escuela Politécnica, University of Extremadura, 10003 Cáceres, Spain; apewsey@unex.es*
[*] *Corresponding author*

Abstract. We propose a five-parameter distribution as a unimodal model for cylindrical data. Its density can be expressed in simple closed form involving no integrals, infinite sums or special functions, its parameters have clear interpretations, and its marginal and conditional distributions are all either Cauchy or wrapped Cauchy. Method of moments and numerically-based maximum likelihood estimation are fast, and tests for both independence and goodness-of-fit are available. An analysis of data on ambient temperature and wind direction from the seminal paper of [1] illustrates the model's application.

Keywords: Cauchy Distribution; Goodness-of-fit; Independence; Parameter Estimation; Wrapped Cauchy Distribution.

References

[1] Johnson, R. A. and Wehrly, T. E. (1978). Some angular-linear distributions and related regression models. *Journal of the American Statistical Association*, **73(363)**, 602–606.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Adaptive Warped Kernel Estimation for Nonparametric Regression with Circular Responses

T. D. Nguyen[1], T. M. Pham Ngoc [2,*] and V. Rivoirard[2]

[1,2]*Laboratoire de Mathématiques d'Orsay, UMR 8628, Université Paris Saclay, 91405 Orsay Cedex France; tiendat.nguyen.mat@gmail.com, thanh.pham_ngoc@math.u-psud.fr*
[3]*Ceremade, CNRS, UMR 7534, Université Paris-Dauphine, PSL Research University, 75016 Paris, France; rivoirard@ceremade.dauphine.fr*
[*]*Corresponding author*

Abstract. In this work, we deal with nonparametric regression for circular data. We propose a kernel estimation procedure with data-driven selection of the bandwidth parameter. For this purpose, we use a warping strategy combined with a Goldenshluger-Lepski type estimator. To study optimality of our methodology, we consider the minimax setting and prove, by establishing upper and lower bounds, that our procedure is nearly optimal on anisotropic Hölder classes of functions for pointwise estimation. The obtained rates also reveal the specific nature of regression for circular responses. Finally, a numerical study is conducted, illustrating the good performances of our approach.

Keywords: Circular Data; Nonparametric Regression; Warping Method; Kernel Rule; Adaptive Minimax Estimation.

Statistical model and estimation procedrure

Assume that we have an i.i.d. sample $\{(X_j,\Theta_j)\}_{j=1}^n$ distributed as (X,Θ), where Θ is a circular random variable, and X is a random variable with density f_X supported on $\mathbb{R}$. We are interested in estimating a regression function m which contains the dependence structure between the predictors X_j and the observations Θ_j. For $x\in\mathbb{R}$, let $m_1(x)=\mathbb{E}\big(\sin(\Theta)|X=x\big)$ and $m_2(x)=\mathbb{E}\big(\cos(\Theta)|X=x\big)$. We investigate the adaptive (meaning that the estimation procedure does not need the specification of smoothness parameters) nonparametric estimation m at a given point in $\mathbb{R}$. Using the angular distance, the function m should minimize the risk $L(m(X)):=\mathbb{E}\big[1-\cos(\Theta-m(X))\big|X\big]$. It is known that the minimizer of the latter risk L is given by $m(x)=\text{atan2}(m_1(x),m_2(x))$.

Let $K:\mathbb{R}\to\mathbb{R}$ be a compactly supported kernel. Assume that F_X the cumulative distribution function of X is known and invertible, the *warping strategy* consists in introducing auxiliary functions $g_1,g_2:(0,1)\mapsto\mathbb{R}$ defined by $g_1:=m_1\circ F_X^{<-1>}$ and $g_2:=m_2\circ F_X^{<-1>}$ in such a way that

$m_1 = g_1 \circ F_X$ and $m_2 = g_2 \circ F_X$. For $u \in F_X(\mathbb{R})$, we set the estimators :

$$\widehat{g}_{1,h_1}(u) := \frac{1}{n}\sum_{j=1}^{n} \sin(\Theta_j).K_{h_1}(u - F_X(X_j)), \quad \text{and} \quad \widehat{g}_{2,h_2}(u) := \frac{1}{n}\sum_{j=1}^{n} \cos(\Theta_j).K_{h_2}(u - F_X(X_j)),$$

to estimate g_1 and g_2 respectively, and $h_1, h_2 > 0$ are bandwidths of kernel $K_{h_1}(\cdot)$ and $K_{h_2}(\cdot)$ respectively. As a consequence, for $x \in \mathbb{R}$, the estimators for m_1 and m_2 are

$$\widehat{m}_{1,h_1}(x) = \big(\widehat{g}_{1,h_1} \circ F_X\big)(x) \quad \widehat{m}_{2,h_2}(x) = \big(\widehat{g}_{2,h_2} \circ F_X\big)(x).$$

Denote $h := (h_1, h_2)$, we then finally set the estimator for $m(x)$ at $x \in \mathbb{R}$ by

$$\widehat{m}_h(x) := \text{atan2}\big(\widehat{m}_{1,h_1}(x), \widehat{m}_{2,h_2}(x)\big).$$

To get an automatic bandwidth selection, we propose a data-driven selection rule inspired from Goldenshluger and Lepski [1]. Let $\widehat{g}_{1,\widehat{h}_1}$ and $\widehat{g}_{2,\widehat{h}_2}$ be the adaptive estimators of g_1 and g_2 respectively where $\widehat{h}_1$ and $\widehat{h}_2$ are bandwidths selected by our data-driven selection rule. We first establish an oracle-type inequality for $\widehat{g}_{1,\widehat{h}_1}$ and $\widehat{g}_{2,\widehat{h}_2}$ which highlights the bias-variance decomposition of the pointwise squared risk. This key result permits us to derive convergence rates for estimating m as shown in the theorem below. Denote $\widehat{h} := (\widehat{h}_1, \widehat{h}_2)$, we define the adaptive estimator of m at $x \in \mathbb{R}$ by $\widehat{m}_{\widehat{h}}(x) := \text{atan2}\big(\widehat{m}_{1,\widehat{h}_1}(x), \widehat{m}_{2,\widehat{h}_2}(x)\big)$.

Let $\beta > 0$ and $L > 0$. The Hölder class $\mathcal{H}(\beta, L)$ is the set of functions $f : (0,1) \longmapsto \mathbb{R}$, such that f admits derivatives up to the order $\lfloor \beta \rfloor$, and for any $(y, \widetilde{y}) \in (0,1)^2$,

$$\left| \frac{d^{\lfloor \beta \rfloor} f}{(dy)^{\lfloor \beta \rfloor}}(\widetilde{y}) - \frac{d^{\lfloor \beta \rfloor} f}{(dy)^{\lfloor \beta \rfloor}}(y) \right| \leq L.\big|\widetilde{y} - y\big|^{\beta - \lfloor \beta \rfloor}.$$

Since the function $\text{atan2}(w_1, w_2)$ is undefined when $w_1 = w_2 = 0$, it is reasonable to consider the following assumption: *For given $x \in \mathbb{R}$, assume that*

$$|m_1(x)| = \big|g_1(F_X(x))\big| \geq \delta_1 > 0, \quad \textit{and} \quad |m_2(x)| = \big|g_2(F_X(x))\big| \geq \delta_2 > 0.$$

The following theorem gives an upper-bound for the mean squared error of the estimator $\widehat{m}_{\widehat{h}}$.

Theorem. Let $\beta_1, \beta_2, L_1, L_2 > 0$. Suppose that g_1 belongs to a Hölder class $\mathcal{H}(\beta_1, L_1)$, g_2 belongs to $\mathcal{H}(\beta_2, L_2)$, Let the kernel K be of order $\mathcal{L} \in (\mathbb{R}_+)$ such that $\mathcal{L} \geq \max(\beta_1, \beta_2)$. For n sufficiently large,

$$\mathbb{E}\Big[\big|\widehat{m}_{\widehat{h}}(x) - m(x)\big|^2\Big] \leq 32.\pi^2.\Big(\frac{1}{\delta_2^2} + \frac{1}{\delta_1^2}\Big).(C_1^2 + C_2^2).\max\big\{\psi_n(\beta_1)^2, \psi_n(\beta_2)^2\big\},$$

where C_1 is a constant (depending on β_1, L_1, K) and C_2 is a constant (depending on β_2, L_2, K), along with $\psi_n(\beta_1) = \big(\log(n)/n\big)^{\frac{\beta_1}{2\beta_1+1}}$ and $\psi_n(\beta_2) = \big(\log(n)/n\big)^{\frac{\beta_2}{2\beta_2+1}}$.

We also derive a lower bound which shows that the attained convergence rate of $\widehat{m}_{\widehat{h}}(x)$ is nearly optimal up to a logarithmic factor.

References

[1] Goldenshluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist*, **39(3)**, 1608-1632.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

On the Circular Median Absolute Deviation

G. C. Porzio[1,*] and H. Demni[1]

[1]*Department of Economics and Laws, University of Cassino and Southern Lazio, Italy; porzio@unicas.it, houyem.demni@unicas.it*
[*]*Corresponding author*

Abstract. The concentration of a circular distribution is typically measured through its mean resultant length, which also gives rise to the so-called circular variance. However, such a measure is not robust and a contamination can affect its value, at least in the standardized bias sense. Alternative robust measures of spread are not really available, especially if we distinguish between robust measures and robust versions of non-robust measures. For this reason, this work aims at evaluating a robust measure of spread: the Circular Median Absolute Deviation (Circular MAD). That is, the (linear) median of the distribution of the shortest arc distances from the circular median. Although its linear analogue is one of the best available measure of scale in terms of robustness (the linear MAD achieving the highest possible breakdown point), we observe that the Circular MAD has been quite neglected within the literature so far. Neither its properties and behaviour under different scenarios, nor its link with any distribution are found in the literature. This work will partially fill this gap.

Keywords: Circular Mean Absolute Deviation; Directional Data; Mean Resultant Length; von Mises Distribution.

Acknowledgments. This work has been partially funded by the BiBiNet project (grant H35F2100 0430002) within the POR-Lazio FESR 2014-2020.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Some Directional Models for Tree Branching Patterns

L.-P. Rivest[1,*]

[1]*Department of Mathematics and Statistics, Université Laval, Québec, Canada; Louis-Paul.Rivest@mat.ulaval.ca*
[*]*Corresponding author*

Abstract. Computer tomography (CT) scanning provides new opportunities for understanding tree branching patterns. It allows the construction of an analytical representation of a tree crown where each branch is a 3D line segment, classified in a level of a hierarchy, starting with the trunk (level 1), the branches attached to the trunk (level 2) and so on. The ancestors of a branch in this hierarchy are its parent, that is the branch from which it emanates, and its parent's ancestors, see [2]. This presentation considers statistical models to investigate how a branch orientation can be explained by characteristics of its ancestors. The S^2 unit vectors of level 2 branches, attached to the trunk, can be modeled with the small circle distribution of Bingham and Mardia, see [1]. The modeling of the orientations of level 3 branches branch is more complex. A small circle model could account for their scatter around their parents. As trees tend to fill up the space available, level 3 branches could also spread in directions opposite to that of the trunk. This presentation investigates a directional model that accounts for these two features. It uses a combination method first proposed in [3]. The new model belongs the Fisher-Bingham family presented in Chapter 9 of [3]. Its properties are briefly reviewed and it is then used to analyze data on the crown structure of a miniature Pixie tree that was CT scanned at the Eastern Canadian Plant Phenotyping Platform of McGill University.

Keywords: Generalized Linear Model; Generalized Rsquare; Spherical Data.

References

[1] Bingham, C. and Mardia, K. V. (1978). A small circle distribution on the sphere. *Biometrika*, **65(2)**, 379–389.

[2] Dutilleul, P., Mudalige, N. and Rivest, L.-P. (2022). Learning how a tree branches out: A statistical modeling approach. *PlosOne*. Under revision.

[3] Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*,**1**, 96.

[4] Mardia, K. V. and Jupp, P. E. (2009). *Directional statistics*. John Wiley & Sons.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

The FMM Approach to Model Oscillatory Signals. The Case of the Electrocardiogram

C. Rueda[1]

[1]*Department of Statistics and Operations Research, University of Valladolid; cristina.rueda@uva.es*

Abstract.

Introduction: Oscillatory systems arise in the different biological and medical fields. By instances, variables following the heart rhythm as those measured in the electrocardiogram (ECG) are oscillatory signals. Questions such as how to detect heart rhythm failures by automatically interpreting the ECG are only an example of relevant advances in biomedical signal analysis, a field that is in continuous advance, mainly due to the basic statistics and mathematics research. Oscillations are encountered in all areas of science, physics, and biology, and in human society as the business cycle indicators. Therefore, their processing and analysis is carried out from many different perspectives. On the one hand, the focus of the signal analyst emphasizes the time-frequency approach. On the other hand, a dynamic system is described primarily by a set of differential equations for a physicist or mathematician. Finally, there is the statistical approach, which is suitable when real signals are observed and is the focus of this talk. Specifically, we exploit the circular nature of oscillatory processes to derive efficient parametric models. Moreover, depending on the application, the purpose of the analysis differs. Some of the popular purposes are: the detection of periodicity, the extraction of features, locating fiducial marks, generating synthetic data, or denoising signals. Among those, extracting features from an observed signal is perhaps the most widely studied data analysis problem. Consequently, to define a reduced set of interpretable features and an efficient algorithm to accurately extract these features from the recorded signal are the top requirements of an efficient signal analysis method. The number of oscillatory components and the amplitude and peak time of each oscillation are among the main features to be extracted, which in the case of physiological signals, contain interesting information about a person's health condition. In this talk, a review of the main theoretical and computational properties of the FMM (Frequency Modulated Mobiüs) approach to model oscillatory signals is given, also the most significant advances that the method has yielded in the analysis of electrocardiogram signals. The importance of the ECG signal in the diagnosis and prediction of cardiovascular diseases is worth noting. The process recorded in the ECG is the periodic electrical activity of the heart. The ECG signals are mostly used as diagnostic tools, since an irregularity in any of those measurements could indicate a heart condition. However, interpreting the signals is not easy, even for trained physicians.

Methods: A single oscillation is mathematically represented in the circular space as a circular signal. The FMM approach, is a novel approach to study these signals that models the phase in the circular space using Mobiüs transforms [3, 2, 1]. It solves a variety of exciting questions with real

data; some of them, such as the decomposition of the signal into components and their multiple uses, are of general application others are specific. The underlying statistical model is a signal plus error model where the signal is described parametrically and is decomposed into several additive components. The parametric formulation facilitates the interpretability and the derivation of essential elements. Some of the main FMM parameters are circular, representing the cyclical nature of the underlying biological processes. The main virtues that make the FMM approach preferable to other approaches, are on the one hand, that the components describe specific physiological processes and the parameters can characterize, reproduce, and identify the variety of morphologies observed in each wave that compose the signal. On the other hand, the fitting algorithm provides accurate and robust model parameter estimates, discarding overfitting problems. Furthermore, by assuming simple restrictions, the parameters are identifiable.

Results: The FMM_{ecg} model, separately characterize the five fundamental waves of a heartbeat. It does so by generating parameters describing the wave shape of a heartbeat, in a similar way a physician would do manually. Diagnostic results are then calculated automatically from that data. A maximization- Identification (MI) algorithm is designed to estimate the parameters of the model. This algorithm alternates,iteratively, between a maximization, M-step, and a wave-identification, I-step. The methodology's potential to describe a variety of noisy and pathological ECG patterns it is shown.

Conclusions: The FMM is a multi-purpose approach that solves questions, such as the extraction of interpretable features, the generation of synthetic data or signal compression. The method is statistically and mathematically sound where the cyclic structure of the underlying process is represented by the circular parameters. The method outperforms data driven and model-based methods for the ECG analysis being the greatest benefit from this new discovery its potential as automatic interpretation method.

Keywords: FMM Models; Oscillatory Signal; Electrocardiogram.

Acknowledgments. The author gratefully acknowledge the financial support received by the Spanish Ministry of Science, Innovation and Universities [PID2019-106363RB-I00].

References

[1] Rueda, C. and Fernández, I. and Larriba, Y. and Rodríguez-Collado, A. (2021). The FMM Approach to Analyze Biomedical Signals: Theory, Software, Applications and Future. *Mathematics*, **9(10)**, 1145.

[2] Rueda, C. and Rodríguez-Collado, A. and Larriba, Y. (2021). A Novel Wave Decomposition for Oscillatory Signals. *IEEE Transactions on Signal Processing*, **69**, 960-972.

[3] Rueda, C. and Larriba, Y. and Peddada, S. D. (2019). Frequency Modulated Möbius Model Accurately Predicts Rhythmic Signals in Biological and Physical Sciences. *Scientific Reports*, **9**, 18701.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

From Topology of the Data Space to a Compatible Probabilistic Model

K. Sargsyan[1,*]

[1] *Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan; karsar@ibms.sinica.edu.tw*
[*] *Corresponding author*

Abstract. Recently most of the data analysis done for practical applications deals with big data. For big data, deep learning has gained particular popularity and has performed successfully in several important areas such as image recognition, machine translation, and many others. Deep probabilistic modeling (DPM) is a deep learning approach trying to account for both model/data uncertainty and using probability distributions as building blocks. As statisticians working with directional statistics know well, one cannot naively apply probability distributions used for Euclidean data to circular data. In fact, the topology of the data space restricts our choices for probability distributions (compare cases of torus and sphere). In recent years an effort has been made to equip existing DPM software frameworks with such probability distributions as building blocks for particular cases of data space topology. Often, we have a pretty good understanding of what the topology of data space should look like (an example of dihedral angles for proteins). In some cases, computational methods of topological data analysis may present a starting point. However, for big data, one may easily encounter non-trivial topology, especially after applying topological data analysis. Therefore, we present an approach to model topology of the data space via cell complexes and automatically generate corresponding probability distributions for the Pyro probabilistic programming language.

Keywords: Topology; Deep Probabilistic Model; Probability Distribution.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Score Matching for Microbiome Compositional Data

J. L. Scealy[1,*] and A. T. A. Wood[1]

[1]*Research School of Finance, Actuarial Studies and Statistics, Australian National University, 26C Kingsley Street, Canberra, ACT 2601; janice.scealy@anu.edu.au, andrew.wood@anu.edu.au*
[*]*Corresponding author*

Abstract. Compositional data are challenging to analyse due to the non-negativity and sum-to-one constraints on the sample space. It is often the case with microbiome compositional data that many of the components are highly right-skewed, with large numbers of zeros. A major limitation of currently available estimators for compositional models is that they either cannot handle many zeros in the data or are not computationally feasible in moderate to high dimensions. We derive a new set of novel score matching estimators applicable to distributions on a Riemannian manifold with boundary, of which the standard simplex is a special case. The score matching method is applied to estimate the parameters in a new flexible model for compositional data and we show that the estimators are scalable and available in closed form. We apply the new model and estimators to real microbiome compositional data and show that the model provides a good fit to the data. Some very recent extensions of this work will also be described.

Keywords: Dirichlet Distribution; Latent Variables; Multinomial Distribution; Parameter Estimation; Zeros.

References

[1] Hyvarinen, A. (2005). Estimation of non-normalised statistical models by score matching. *Journal of Machine Learning Research*, **6**, 695–709.

[2] Mardia, K. V., Kent, J. T. and Laha, A. K (2016). Score matching estimators for directional distributions. https://arxiv.org/abs/1604.08470

[3] Martin, I., Uh, H-W., Supali, T., Mitreva, M. and Houwing-Duistermaat, J. J. (2018). The mixed model for the analysis of a repeated-measurement multivariate count data. *Statistics in Medicine*, **38**, 2248–2268.

[4] Scealy, J. L. and Wood, A. T. A. (2021). Score matching for compositional distributions. *Journal of the American Statistical Association (accepted article)*.

[5] Yu, S., Drton, M. and Shojaie, A. (2019). Generalised score matching for non-negative data. *Journal of Machine Learning Research*, **20**, 1–70.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Asymptotic Power of Sobolev Tests for Uniformity on Hyperspheres

E. García-Portugués[1], D. Paindaveine[2] and T. Verdebout[2,*]

[1]*Department of Statistics, University Carlos III, Madrid, Spain; edgarcia@est-econ.uc3m.es*
[2,*]*Department of Mathematics and ECARES, ULB, Belgium; Davy.Paindaveine@ulb.be, thomas.verdebout@ulb.be*
[*]*Corresponding author*

Abstract. One of the most classical problems in multivariate statistics is considered, namely, the problem of testing isotropy, or equivalently, the problem of testing uniformity on the unit hypersphere. Rather than restricting to tests that can detect specific types of alternatives only, we consider the broad class of Sobolev tests. While these tests are known to allow for omnibus testing of uniformity, their non-null behavior and consistency rates, unexpectedly, remain largely unexplored. To improve on this, we thoroughly study the local asymptotic powers of Sobolev tests under the most classical alternatives to uniformity, namely, under rotationally symmetric alternatives. We show in particular that the consistency rate of Sobolev tests does not only depend on the coefficients defining these tests but also on the derivatives of the underlying angular function at zero.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Analogues on the Sphere of the Affine-equivariant Spatial Median

J. L. Scealy[1] and A. T. A. Wood[1,*]

[1]*Research School of Finance, Actuarial Studies and Statistics, Australian National University, 26C Kingsley Street, Canberra, ACT 2601; janice.scealy@anu.edu.au, andrew.wood@anu.edu.au*
[*]*Corresponding author*

Abstract. This talk will discuss the robust estimation of location on the sphere. Even though the influence functions for the mean direction and other location estimators are bounded, it will be argued that lack of robustness in location estimation can still be a problem on the sphere. A more relevant criterion for robustness on the sphere is standard-bias robustness, or SB -robustness, which will be defined in the talk, as opposed to just requiring that the influence function is bounded. Most of the previous literature on SB-robustness has focused on the rotationally symmetric case (see Scealy and Wood (2021) for a list of references). The talk will focus on two estimators of location that are analogues on the sphere of the affine-invariant spatial median in $\mathbb{R}^d$; see Hettmansperger and Randles (2002). The new location estimators are particularly well-suited for unimodal distributions with ellipse-like symmetry (see Rivest (1984) for discussion of this type of symmetry). Robustness properties, asymptotic behaviour and practical performance of these estimators will be described. We also prove general semi-parametric results on conditions for SB-robustness which apply to unimodal cases of parametric families such as the Kent distribution (Kent, 1982), the ESAG distribution (Paine et al, 2018) and the scaled von Mises-Fisher distribution (Scealy and Wood, 2019). An example from the geophysics literature will be described briefly. This talk is based on Scealy and Wood (2021).

Keywords: Affine Transformation; Influence Function; Robustness; SB-robustnes; Spatial Median.

References

[1] Hettmansperger, T. P. and Randles, R. H. (2002). A practical affine equivariant multivariate median. *Biometrika*, **89**, 851-860.

[2] Kent, J. T. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society, Series B*, **44**, 71-80.

[3] Paine, P. J., Preston, S. P., Tsagris, M. and Wood, A. T. A. (2018). An elliptically symmetric angular Gaussian distribution. *Statistics and Computing*, **28**, 689-697.

[4] Rivest, L.-P. (1984). On the information matrix for symmetric distributions on the hypersphere. *Annals of Statistics*, **12**, 1085-1089.

[5] Scealy, J. L. and Wood, A. T. A. (2019). Scaled von Mises-Fisher distributions and regression models for paleomagnetic directional data. *Journal of the American Statistical Association*, **114**, 1547-1560.

[6] Scealy, J. L. and Wood, A. T. A. (2021). Analogues on the sphere of the affine-equivariant spatial median. *Journal of the American Statistical Association*, **116**, 1457-1471.

Abstracts: Poster Presenters

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Circular Modal Regression with Applications to Prey Escaping Strategies

M. Alonso-Pena[1,*] and R. M. Crujeiras[1]

[1] *CITMAga, Universidade de Santiago de Compostela; mariaalonso.pena@usc.gal, rosa.crujeiras@usc.gal*
[*] *Corresponding author*

Abstract. We model the relationship between a circular response variable and a general covariate by estimating the conditional local modes. This approach allows the estimation of the most likely values of the response given the value of the predictor. The estimation procedure is based on the local maximization of the conditional kernel density estimator. Convergence rates are derived for the circular multimodal estimator and its finite sample performance is explored via simulations. The new methodology is employed to study the escape behavior of larval zebrafish from a potential predator.

Keywords: Circular Regression; Kernel Estimators; Mean Shift; Multimodal Regression.

The study of dependence between a circular response variable and a general covariate (real-valued or circular) in a nonparametric regression setting was carried out by [2]. In this context, the regression function is regarded, as usual, as the conditional mean direction of the response variable given the covariate. However, when the conditional distribution is highly asymmetric or multimodal, the estimation of the conditional expectation is not the most suitable approach. In the case where a large asymmetry is encountered, the use of quantile regression models can be a useful approach (see [3]).

A much less explored field is the consideration of the (multi)modal regression function, which allows to estimate the conditional local modes of a response variable given a covariate. In the real-valued setting, [4] proposed the use of the conditional mean shift algorithm (see [1]) in order to compute the local modes for each value of the covariate, after obtaining a kernel estimator of the conditional density. This approach results useful when the underlying conditional distribution presents more than one mode, and it can detect the most likely values of the response, as a function of the covariate.

In this work, we explore the nonparametric estimation of the multimodal regression function for a circular response variable. The motivation behind this research is the study of escape responses in larval zebrafish. In the behavioral experiment carried out by [5], the directions in which a group of zebrafish larvae escape from a robot predator were measured, as well as the angle in which

the imitating predator approached its prey. On of the aims of the experiment was to analyze the relationship between the angle of approach (predictor variable) and the escape direction (response variable), but classical regression estimators for a circular response lie in regions without data, given that for some values of the covariate there seem to exist two preferred directions. Thus, the estimation of the multimodal regression function (or, more precisely, multifunction) seems a much more appropriate approach to analyze this dataset.

Let Φ be a circular response variable with support on $[-\pi,\pi)$ and Δ a general predictor, which can be either circular or real-valued. For each value δ in the support of Δ, the circular regression multifunction is defined as the local maxima of the conditional density function, i.e.,

$$M(\delta) = \left\{ \phi \in [-\pi,\pi) : \frac{\partial}{\partial \phi} f(\phi|\delta) = 0,\ \frac{\partial^2}{\partial \phi^2} f(\phi|\delta) < 0 \right\},$$

where $f(\phi|\delta)$ represents the conditional density of $\Phi|\Delta$. In order to estimate the regression multifunction, the conditional density is estimated via a nonparametric method involving circular kernels and, then, the local modes are computed by using a conditional version of the directional mean shift algorithm, with suitable circular weights. The performance of the estimator was studied both theoretically, by obtaining the asymptotic convergence rates, and empirically by means of a simulation study.

The new methodology was applied to the zebrafish dataset, and it was obtained that only one conditional mode is estimated when the fish visualize the threat laterally (where their eyes are located), and in this case their escape behavior is contralateral, i.e., the escape response is toward the opposite side to where the predator stands. However, when the fish visualize the threat with their peripherical vision (rostral and caudal sides), two conditional modes are estimated, indicating ipsilateral escapes, that is, in these cases the larvae escape towards the same side where its predator lies.

Acknowledgments. This work was supported by Project PID2020-116587GB-I00 funded by MCIN/AE I/10.13039/501100011033 and the Competitive Reference Groups 2021-2024 (ED431C 2021/24) from the Xunta de Galicia. Research of M. Alonso-Pena was supported by the Xunta de Galicia through the grant ED481A-2019/139 from the Consellería de Educación, Universidade e Formación Profesional. The authors also acknowledge the Supercomputing Center of Galicia (CESGA) for the computational resources.

References

[1] Comaniciu, D. and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 603–619.

[2] Di Marzio, M., Panzera, A. and Taylor, C. C. (2013). Non-parametric regression for circular responses. *Scandinavian Journal of Statistics*, **40**, 238–255.

[3] Di Marzio, M., Panzera, A. and Taylor, C. C. (2016). Nonparametric circular quantile regression, *Journal of Statistical Planning and Inference*, **170**, 1–14.

[4] Einbeck, J. and Tutz, G. (2006). Modelling beyond regression functions: An application of multimodal regression to speedflow data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **55**, 461–475.

[5] Nair, A., Changsing, K., Stewart, W. J. and McHenry, M .J. (2017). Fish prey change strategy with the direction of a threat. *Proceedings of the Royal Society B: Biological Sciences*, **284**.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Permutation Tests for Three-Dimensional Rotation Data

M. Bingham[1,*]

[1] *1725 State St., University of Wisconsin-La Crosse, La Crosse, WI, USA, 54601; mbingham@uwlax.edu*
[*] *Corresponding author*

Abstract. Motivated both by collaboration with colleagues in Physical Therapy who have limited mathematics background, and teaching at a primarily undergraduate institution where student involvement in research is valued, my recent scholarship has involved developing nonparametric methods for three-dimensional rotation data. In this poster presentation, I will discuss the development of permutation tests for comparing locations of three-dimensional rotations. Results of simulation studies for exploring power and an application comparing movement around joints in the foot for three different species will also be included.

Keywords: Permutations Tests; Three-dimensional Rotation Data.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Circular Regression for Errors-in-Variables

M. Di Marzio[1], S. Fensore[1,*], A. Panzera[2], and C. C. Taylor[3]

[1]*Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University of Chieti-Pescara, Italy; marco.dimarzio@unich.it, stefania.fensore@unich.it*
[2]*DiSIA, University of Florence, Italy; agnese.panzera@unifi.it*
[3]*Department of Statistics, University of Leeds, UK; charles@maths.leeds.ac.uk*
[*]*Corresponding author*

Abstract. We study the problem of estimating a regression function when the predictor and/or the response are circular random variables in the presence of measurement errors [1]. We propose estimators whose weight functions are deconvolution kernels defined according to the nature of the involved variables. We derive the asymptotic properties of the proposed estimators and consider possible generalizations and extensions. We provide some simulation results and a real data case study to illustrate and compare the proposed methods.

Keywords: Deconvolution Kernels; Fourier Coefficients; Measurement errors; Wind Direction; CO Pollution.

References

[1] Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, New York.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Data-Driven Stabilizations of Goodness-of-Fit Tests

A. Fernández-de-Marcos[1,*] and E. García-Portugués[1]

[1]*Department of Statistics, Carlos III University of Madrid (Spain); albertfe@est-econ.uc3m.es, edgarcia@est-econ.uc3m.es*
[*]*Corresponding author*

Abstract. Exact null distributions of goodness-of-fit test statistics are generally challenging to obtain in tractable forms. Practitioners are therefore usually obliged to rely on asymptotic null distributions or Monte Carlo methods, either in the form of a lookup table or carried out on demand, to apply a goodness-of-fit test. Stephens [1] provided remarkable simple and useful transformations of several classic goodness-of-fit test statistics that stabilized their exact-n critical values for varying sample sizes n. However, detail on the accuracy of these and subsequent transformations in yielding exact p-values, or even deep understanding on the derivation of several transformations, is still scarce nowadays. We illuminate and automatize, using modern tools, the latter stabilization approach to (*i*) expand its scope of applicability and (*ii*) yield semi-continuous exact p-values, as opposed to exact critical values for fixed significance levels. We show improvements on the stabilization accuracy of the exact null distributions of the Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling, Kuiper, and Watson test statistics. In addition, we provide a parameter-dependent exact-n stabilization for several novel statistics for testing uniformity on the hypersphere of arbitrary dimension. A data application in astronomy illustrates the benefits of the advocated stabilization for quickly analyzing small-to-moderate sequentially-measured samples.

This contribution is based on the work [2].

Keywords: Exact Distribution; Goodness-of-fit; p-value; Stabilization; Uniformity.

References

[1] Stephens, M. A. (1970). Use of the Kolmogorov–Smirnov, Cramér-von Mises and related statistics without extensive tables. *J. R. Stat. Soc. Ser. B Methodol.*, **32(1)**, 115–122.

[2] Fernández-de-Marcos, A. and García-Portugués, E. (2021). Data-driven stabilizations of goodness-of-fit tests. *arXiv:2112.01808*.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Analyzing Compositional Data Using a Directional Distribution

A. Figueiredo[1,*]

[1]*Faculdade de Economia da Universidade do Porto and LIAAD - INESC TEC Porto, Rua Dr. Roberto Frias, 4200-464 Porto-Portugal; adelaide@fep.up.pt*
[*]*Corresponding author*

Abstract. Compositional data are vectors whose components are non-negative values and constrained to a constant sum, for example vectors of proportions that sum one. This type of data arise in many areas, including Agriculture, Archaeology, Biology, Economics, Environment, Geography, Geology, Medicine and Psychology. These data need to be transformed, before applying the standard statistical techniques designed for the Euclidean space. The most usual transformations of these data are based on the log-ratios of the components of the compositional vectors ([1]). Alternatively, the square-root transformation can be used to transform compositional data into directional data, as suggested in [2], and then modeled using distributions defined on the hypersphere.
In this study we use this transformation for analyzing compositional data and we model the obtained data using the Watson distribution defined on the hypersphere (for details about this distribution, see for example [3]). More precisely, we apply the square-root transformation and then for clustering compositional data we identify a mixture of Watson distributions and for the classification of compositional data into predefined groups, we use the Bayes classification rules for the Watson distribution. We apply these methods to several compositional data sets already analyzed in the literature using other transformations of the compositional data.

Keywords: Bayes Rule; EM Algorithm; Hypersphere; Watson Distribution.

Acknowledgments. This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

References

[1] Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, **44 (2)**, 139–177.

[2] Stephens, M. A. (1982). Use of the von Mises distribution to analyse continuous proportions. *Biometrika*, **69 (1)**, 197–203.

[3] Mardia, K. V. and Jupp, P. E. (2009). *Directional statistics*. John Wiley & Sons.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Nonparametric Plug-in Estimation of Spherical Highest Density Regions for Galician Surnames

M. J. Ginzo-Villamayor[1,*] and P. Saavedra-Nieves[1]

[1]*Faculty of Mathematics Rúa Lope Gómez de Marzoa, s/n. Campus vida 15782 Santiago de Compostela; mariajose.ginzo@usc.gal, paula.saavedra@usc.gal*
[*]*Corresponding author*

Abstract. Surnames (family names) can be used as a source of information for population characteristics, given that the analysis of surname patterns provides information about long–term and short–term dynamics of population movements. Linguistics considers different classifications of surnames depending on motivation, morphology or semantic. Specifically, the semantic classification in [1] establishes the following groups for Galician surnames: patronymic, toponymic and apelative.

Under this approach, Web-Scraping tools[1] described in [2] have been used to construct a random sample of 315 points on the 3-dimensional unit sphere. Specifically, the preliminary database contains a total of 1711 Galician surnames that represent 8.15% of the total number of surnames in Galicia and 86.15% of the population. After classifying them according to the Galician council they belong, each of the 315 geographical districts has been assigned a point on the 3-dimensional unit sphere where coordinates correspond to the council normalised percentage of surnames in the semantic categories apelative, patronymic and toponymic, respectively.

The main goal of this work is to estimate the Highest Density Regions (HDRs) from the previous directional sample in order to get a deeper geographical understanding of the semantic surnames distribution. Following [3], given $\tau \in (0,1)$, the $100(1-\tau)\%$ directional HDR is the subset

$$L(f_\tau) = \{x \in S^{d-1} : f(x) \geq f_\tau\} \tag{2}$$

where f denotes the density of a random vector X taking values on a $d-$dimensional unit sphere S^{d-1} and f_τ is the largest constant such that

$$P(X \in L(f_\tau)) \geq 1-\tau$$

with respect to the distribution induced by f. Remark that $L(f_\tau)$ is equal to the greatest modes of the distribution when large values of τ are considered.

Although there are other alternative routes for estimating HDRs, the plug-in approach has received considerable attention in the literature. Given a random sample on S^{d-1} of the unknown directional density f, plug-in methods reconstruct the $100(1-\tau)\%$ HDR namely $L(f_\tau)$ in (2) as

$$\hat{L}(\hat{f}_\tau) = \{x \in S^{d-1} : f_n(x) \geq \hat{f}_\tau\}$$

[1]Web Scraping converts data in unstructured format (HTML tags) on the web to an easy-to-access structured format.

where $\hat{f}_\tau$ is an estimator of the threshold f_τ and f_n denotes a nonparametric directional density estimator.
In this work, plug-in HDRs estimation methodology is applied on the constructed sample using the R package `HDiR`[2] in order to detect the existence of geographic and semantic patterns for Galician surnames.

Keywords: Directional Highest Density Regions; Galician Surnames; Plug-in Estimation; Semantic Classification.

Acknowledgments. This work was supported by the Government of Galicia (Grupos de Referencia Competitiva) ED431C 2021/24 and the Grants PID2020-116587GB-I00 and PID2020-118101GB-I00 funded by MCIN/AEI/ 10.13039/501100011033.

References

[1] Boullón-Agrelo, A. I. (2008). *I nomi nel tempo e nello spazio - V. Atti del XXII Congresso Internazionale di Scienze Onomastiche Pisa*, vol. II, chap. The surnames in Galicia today: a characterization and description, 299–310. Edizioni ETS.

[2] Ginzo-Villamayor, M. J. (2022). Statistical Techniques in Geolinguistics. Onomastic modeling. *PhD Thesis*.

[3] Saavedra-Nieves, P. and Crujeiras, R. M. (2021). Nonparametric estimation of directional highest density regions. *Advances in Data Analysis and Classification*, 1–36.

[2] `https://CRAN.R-project.org/package=HDiR`

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

New Construction of a Cylindrical Distribution from Independent Linear and Circular Distributions

T. Imoto[1,*]

[1]*School of Management and Information, University of Shizuoka, 52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan; imoto@u-shizuoka-ken.ac.jp*
[*]*Corresponding author*

Abstract. In diverse scientific fields, circular data is often obtained with linear data. Typical examples are wind direction and its speed at some point, and ball-pass direction and its position in a ball game. Such data should be modeled by a cylindrical distribution, or bivariate distribution with circular and linear marginals. In this study, we focus on a simple construction as a method of specifying marginal distributions, where the support of the linear variable is $\mathbb{R}$. Let $G(\cdot)$ be a distribution function whose probability density function (pdf) is symmetric about zero, $f_L(\cdot)$ be a linear pdf whose mean and variance are 0 and 1, respectively, and $f_C(\cdot)$ be a circular pdf whose mean direction is 0. Then the proposed pdf is defined by

$$f(x,\theta) = 2G\left(\frac{2\lambda(x-\mu_L)\sin(\theta-\mu_C)/\sigma}{1+(x-\mu_L)^2/\sigma^2}\right)\frac{1}{\sigma}f_L\left(\frac{x-\mu_L}{\sigma}\right)f_C(\theta-\mu_C),$$

$$x\in(-\infty,\infty), \quad \theta\in[-\pi,\pi),$$

where $\mu_L\in(-\infty,\infty)$ and $\sigma\in(0,\infty)$ are location and scale parameters of linear variable, respectively, $\mu_C\in[-\pi,\pi)$ is a location parameter of circular variable, and λ is a dependence parameter between two variables. The marginal pdfs of this distribution are given by $\frac{1}{\sigma}f_L\left(\frac{x-\mu_L}{\sigma}\right)$ for linear variable and $f_C(\theta-\mu_C)$ for circular variable. If $G(\cdot)$ is expressed in a closed-form like uniform, logistic, Cauchy distribution function, the proposed construction has a merit of the nonnecessity of additional normalizing constant. The other desirable properties of this construction and illustrative fittings to ball-passing data in soccer game are given in the talk.

Keywords: Bivariate Distribution about Linear and Circular Variables; Construction with Specifying Marginal Distribution; Soccer Game Data.

Acknowledgments. The author would like to thank the invitation to honorable international workshop ADISTA 22.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

CircSpaceTime: an R Package for Spatial and Spatio-Temporal Modelling of Circular Data

G. Mastrantonio[1,*], G. Jona Lasinio[2] and M. Santoro[3]

[1]*Polytechnic of Turin, Turin, Italy; gianluca.mastrantonio@gmail.com*
[2]*Sapienza University of Rome, Rome, Italy; giovanna.jonalasinio@uniroma1.it*
[3]*IAC CNR, Rome, Italy; m.santoro@iac.cnr.it*
[*]*Corresponding author*

Abstract. CircSpaceTime is the only R package, currently available, that implements Bayesian models for spatial and spatio-temporal interpolation of circular data. Such data are often found in applications where, among the many, wind directions, animal movement directions, and wave directions are involved. To analyse such data, we need models for observations at locations $\mathbf{s}$ and times t, as the so-called geostatistical models, providing structured dependence assumed to decay in distance and time. The approach we take begins with Gaussian processes defined for linear variables over space and time. Then, we use either wrapping or projection to obtain processes for circular data. The models are cast as hierarchical, with fitting and inference within a Bayesian framework. Altogether, this package implements work developed by a series of papers, by Jona Lasinio, Mastrantonio, Wang, and Gelfand [1, 2]. All procedures are written using *Rcpp*. Estimates are obtained by MCMC, allowing parallelized multiple chains run. The implementation of the proposed models is considerably improved on the simple routines adopted in the research papers. As original running examples, for the spatial and spatio-temporal settings, we use wind directions datasets over central Italy.

Keywords: *R*-package; *Rcpp*; Spatial Models.

Acknowledgments. Gianluca Mastrantonio research has been partially supported by MIUR grant Dipartimenti di Eccellenza 2018 - 2022 (E11G18000350001), conferred to the Dipartimento di Scienze Matematiche - DISMA, Politecnico di Torino.

References

[1] Mastrantonio, G., Jona Lasinio, G., and Gelfand, A. E. (2016). Spatio-temporal circular models with non-separable covariance structure. *TEST*, **25**, 331-350.

[2] Fangpo, W. and Alan, E. G. (2014). Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*, **109**, 1565-1580.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

A Flexible Functional-Circular Regression Model for Analyzing Temperature Curves

A. Meilán-Vila[1,*], R. M. Crujeiras[2] and M. Francisco-Fernández[3]

[1]*Department of Statistics, Carlos III University of Madrid, Av. de la Universidad 30, Leganés, 28911, Spain; ameilan@est-econ.uc3m.es*
[2]*CITMAga, Centre for Mathematical Research and Technology, Universidade de Santiago de Compostela, San Xerome s/n, Santiago de Compostela, 15782, Spain; rosa.crujeiras@usc.gal*
[3]*Department of Mathematics, Universidade da Coruña. CITIC, Faculty of Computer Science, Campus de Elviña s/n, A Coruña, 15071, Spain; mariofr@udc.gal*
[*]*Corresponding author*

Abstract. The analysis of a variable of interest which depends on other variable(s) is a typical issue appearing in many practical problems. Regression analysis provides the statistical tools to address this type of problems. This topic has been deeply studied, especially when the variables in study are of Euclidean type. However, there are situations where the data present certain kind of complexities, for example, the involved variables are of circular [2] or functional [6] type, and the classical regression procedures designed for Euclidean data may not be appropriate. In these scenarios, these techniques would have to be conveniently modified to provide useful results.

This work aims to design and study a nonparametric regression estimator for a model with a circular response and a functional covariate. When the response variable is of circular nature, the regression function can be defined as the minimizer of a circular risk function. It can be proved that the minimizer of this function is given by the inverse tangent function of the ratio between the conditional expectation of the sine and the cosine of the response variable. The estimator proposed in this work implicitly considers two regression models, one for the sine and another for the cosine of the response variable. Then, a Nadaraya–Watson-type estimator of the circular regression function is obtained by computing the inverse tangent function of the ratio of Nadaraya–Watson estimators for the two regression functions of the sine and cosine models. This way of proceeding has been previously used. For example, considering a regression model with an $\mathbb{R}^d$-valued covariate and a circular response, similar proposals (adapted to that context) to estimate the circular regression function were provided and studied, under the assumption of independence in [4] and also for spatially correlated errors in [3]. The estimator proposed in this work depends on a univariate kernel function and on a bandwidth parameter. The selection of the bandwidth parameter is crucial in the estimation procedure, since it controls the smoothness of the estimator. In this research, a cross-validation approach is used to select the bandwidth parameter in practice.

In this work, under certain assumptions, the asymptotic bias and variance of the proposed estimator, as well as, its asymptotic distribution, are calculated. A comprehensive simulation study is carried out to check the performance of the estimator in practice. In addition, in order to illustrate the procedure with a real dataset, we consider daily temperature records in Santiago de Compostela

(NW-Spain) for the period 2002-2019. Modeling the relation between the temperature curve (as a functional covariate) and the day of the year (as a circular response) for a certain period of time allows to compare observations and predictions given by such a model in a different period. This procedure enables to illustrate how temperature patterns change on local scale, probably due to global warming.
The numerical analysis carried out in this research was performed with the statistical environment `R` [5], using the functions supplied with the `fda.usc` package [1].

Keywords: Circular Data, Flexible Regression, Functional Data, Temperature Curves.

Acknowledgments. Research of A. Meilán-Vila and M. Francisco-Fernández has been supported by MINECO (Grant MTM2017-82724-R), MICINN (Grant PID2020-113578RB-I00), and by Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01), all of them through the ERDF. Research of R. M. Crujeiras has been supported by MICINN (Grant PID2020- 116587GB-I00), and by Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2021-24), all of them through the ERDF.

References

[1] Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). *Statistical computing in functional data analysis: The R package fda.usc*. Journal of Statistical Software, **51(4)**, 1–28.

[2] Mardia, K. V. and Jupp, P. E. (2009). *Directional statistics*. John Wiley & Sons.

[3] Meilán-Vila, A., Crujeiras, R. M. and Francisco-Fernández, M. (2021). Nonparametric estimation of circular trend surfaces with application to wave directions. *Stochastic Environmental Research and Risk Assessment*, **35**, 923–939.

[4] Meilán-Vila, A., Francisco-Fernández, M., Crujeiras, R. M. and Panzera, A. (2021) Nonparametric multiple regression estimation for circular response. *TEST*, **30**, 650–672.

[5] R Development Core Team (2022). R: a language and environment for statistical computing. *R Foundation for Statistical Computing*. `http://www.R-project.org`.

[6] Ramsay, J. O. and Silverman, B. (2005). *Functional data analysis*. Springer.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Biomechanical Data Modeling through a Multivariate Circular-Linear Model based on Vine Copulas

P. Nagar[1,*], A. Bekker[1], M. Arashi[2,1], C. Kat[3] and A. C. Barnard[4]

[1]*Department of Statistics, University of Pretoria, Pretoria 0002, South Africa; priyanka.nagar@up.ac.za, andriette.bekker@up.ac.za*
[2]*Department of Statistics, Ferdowsi University of Mashhad, Mashhad 9177948974, Iran; arashi@um.ac.ir*
[3]*Department of Mechanical and Aeronautical Engineering , University of Pretoria, Pretoria 0002, South Africa; cor-jacques.kat@up.ac.za*
[4]*Walk-A-Mile Centre for Advanced Orthopaedics, Centurion 0157, South Africa; annettechristi@gmail.com*
[*]*Corresponding author*

Abstract. High-dimensional data containing circular and linear variables is common in biomechanical and orthopedic data. In most cases the circular and linear variables are considered in isolation. The joint distribution modelling based on high-dimensional data containing circular and linear data is vital given the large amounts of directional data and the vast applications thereof. In this study, we propose a modelling framework applicable to the 6D joint distribution of circular-linear data based on vine copulas. The pair-copula decomposition concept of vine copulas represents the dependence structure as a combination of circular-linear, circular-circular and linear-linear pairs modelled by their respective copulas. This allows us to assess the dependencies in the joint distribution. This study is motivated by the modelling of biomechanical data, i.e. the fracture displacements, that is used as measure in external fixator comparisons. A case study based on the rotational and translational variables from a external fixator experiment illustrates the distribution's application.

Keywords: Circular-linear Data; Fracture Displacement; Multivariate Statistics; Vine Copulas.

Acknowledgments. This work was based upon research supported in part by the National Research Foundation (NRF) of South Africa, Reference: SRUG190308422768 grant No. 120839 and SARChI Research Chair UID: 71199, Re:IFR170227223754 grant No. 109214. Opinions expressed and conclusions arrived at in this study are those of the authors and are not necessarily to be attributed to the NRF.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Enhancing Wind Direction Prediction of South Africa Wind Energy Hotspots with Bayesian Mixture Modeling

N. Nakhaei Rad[1,2,*], A. Bekker[1] and M. Arashi[1,3]

[1] *Department of Statistics, University of Pretoria, Pretoria 0002, South Africa; najmeh.nakhaeirad@up.ac.za, andriette.bekker@up.ac.za*
[2] *DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS), South Africa.*
[3] *Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Iran; arashi@um.ac.ir*
[*] *Corresponding author*

Abstract. Wind energy production depends not only on wind speed but also on wind direction. Thus, predicting and estimating the wind direction for sites accurately will enhance measuring the wind energy potential. The uncertain nature of wind direction can be presented through probability distributions and Bayesian analysis can improve the modeling of the wind direction using the contribution of the prior knowledge to update the empirical shreds of evidence. This must align with the nature of the empirical evidence as to whether the data are skew or multimodal or not. So far mixtures of von Mises within the directional statistics domain, are used for modeling wind direction to capture the multimodality nature present in the data. In this paper, due to the skewed and multimodal patterns of wind direction on different sites of the locations understudy, a mixture of multimodal skewed von Mises is proposed for wind direction. Furthermore, a Bayesian analysis is presented to take into account the uncertainty inherent in the proposed wind direction model. A simulation study is conducted to evaluate the performance of the proposed Bayesian model. This proposed model is fitted to datasets of wind direction of Marion island and two wind farms in South Africa and show the superiority of the approach. The posterior predictive distribution is applied to forecast the wind direction on a wind farm. It is concluded that the proposed model offers an accurate prediction by means of credible intervals. The mean wind direction of Marion island in 2017 obtained from 1079 observations was 5.0242 (in radian) while using our proposed method the predicted mean wind direction and its corresponding 95% credible interval based on 100 generated samples from the posterior predictive distribution are obtained 5.0171 and (4.7442,5.2900). Therefore, our results open a new approach for accurate prediction of wind direction implementing a Bayesian approach via mixture of skew circular distributions.

Keywords: Bayes Estimation; Mixture Distribution; Modified Gibbs Sampling; Sine-skewed von Mises Distribution; Wind Direction.

Acknowledgments. This work was based upon the National Research Foundation (NRF) of South Africa, SARChI Research Chair UID: 71199; Ref.: IFR170227223754 grant No. 109214;
Ref.: SRUG190308422768 grant No. 120839, the South African DST-NRF-MRC SARChI Research Chair in Biostatistics (Grant No. 114613), STATOMET at the Department of Statistics at the University of Pretoria and DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS), South Africa. The research of the third author (M. Arashi) is supported by a grant from Ferdowsi University of Mashhad (N.2/56073). The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the CoE-MaSS or the NRF.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Copula Bounds for Circular Data

H. Ogata[1,*]

[1] *1-1 Minami-Osawa, Hachioji, Tokyo, Japan 192-0397; hiroakiogata@tmu.ac.jp*
[*] *Corresponding author*

Abstract. The copula is a powerful tool for describing the dependency of random variables. In two dimensions, the Fréchet-Hoeffding upper [lower] bound indicates the perfect positive [negative] dependence between two random variables. However, for circular random variables, the usual concept of dependency is not accepted because of their periodicity.
In this work, we give the equivalence class of circular copula function in the sense of arbitrariness of origins. Then, the copula bounds for circular data are shown. When the circular copula reaches the bounds, we say the circular random variables have perfect dependence, and it is considered as a generalization of complete dependence in [1].
We also consider modified Fréchet and Mardia families of copulas for modelling the dependency of two circular random variables. Finally, simulation studies are given to demonstrate the behavior of the model.

Keywords: Fréchet-Hoeffding Copula Bounds; Equivalence Class.

References

[1] Fisher, N. I. and Lee, A. J. (1983) A correlation coefficient for circular data. *Biometrika*, **70(2)**, 327–332.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Modeling Circular Time Series

D. Palumbo[1,*] et al.

[1] *Ca' Foscari University of Venice, Department of Economics, Cannaregio 873, 30121 Venice, Italy.*
Homerton College, University of Cambridge, Hills Road, Cambridge CB2 8PH, UK; dp470@cam.ac.uk
[*] *Corresponding author*

Abstract. Circular observations pose special problems for time series modelling, such as the frequent display of rapid movements all around the circle due to data circularity. The aim of the present paper is to develop a new comprehensive class of time series models that addresses these issues and yields a coherent model specification methodology staightforward to implement. The novel approach is based on a circular conditional distribution, such as the von Mises, while the dynamics are driven by the score of the conditional distribution following the approach of [2] and [1]. The flexibility of this approach allows its application on various directional distributions also with heavy tails, such as the Cardioid or the Skewed Cauchy. The paper shows also how the new score-driven model can be applied on non-stationary circular time series introducing a general class of non-stationary specifications. Moreover, the basic model is extended with an additional time varying parameter which captures heteroscedasticity. The asymptotic distribution of the maximum likelihood estimator is derived for a first-order model and specification tests are proposed based on the Lagrange multiplier principle. The small sample properties of the estimator are then investigated by Monte Carlo experiments and shown to be consistent with the asymptotics. The empirical performance of the model is then assessed on various datasets of wind direction.

Keywords: Directional Statistics; Dynamic Conditional Score Models; Heteroscedasticity; von Mises Distribution; Wind Direction.

References

[1] Creal, D., Koopman, S. J. and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics,* **28**, 777-95.

[2] Harvey, A. C. (2013). Dynamic Models for Volatility and Heavy Tails: with Applications to Financial and Economic Time Series. *Econometric Society Monograph*, Cambridge University Press.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Complex Valued Time Series Modeling with Relations to Directional Statistics

T. Shiohama[1,*]

[1] *18 Yamazatocho, Showa, Nagoya, 466-0673, JAPAN; shiohama@nanzan-u.ac.jp*
[*] *Corresponding author*

Abstract. Stationary time series fluctuation often shows periodic behavior and these patterns are usually summarised via a spectral density. Since the spectral density is a periodic function, it can be modeled by using a circular distribution function. In this study, several time series models are studied in relation to a circular or a cylindrical distribution. First, as an introduction, we illustrate how to model bivariate time series data using complex-valued time series in the context of circular distribution functions. Next, some other time series modeling by incorporating cylindrical distributions is illustrated. The maximum likelihood estimation procedures are introduced to estimate unknown model parameters. Some real data analyses are also performed to illustrate the proposed models' applicability.

Keywords: Circular Statistics; Cylindrical Distribution; Maximum Likelihood Estimation; Spectral Density; Time Series Analysis.

Introduction

Since the spectral density of a time series is a periodic function, this is closely related to the circular density function used in the field of directional statistics. The clear relationship between time series spectra and circular density is investigated by Taniguchi et al. [9], where the well known circular distributions including a wrapped Cauchy, von-Mises, and, Kato and Jones models [5, 6] are related to the spectral density of AR(1), MA(1), AR(2), and ARMA(1,1) models, respectively. Some circular time series models are considered in the literature, we refer to the circular Markov models of Abe et al. [2] and the higher-order circular Markov model of Ogata and Shiohama [8]. These studies also revealed the relationship between a circular data analysis and the time series spectra.

The difference between this study and that studied by Taniguchi et al. [9] is that this study focuses on the time series modeling while the latter considers circular distribution constructed from time series spectra. If we consider a stationary time series process with complex-valued autoregressive and/or moving average coefficients, the observed process is on the complex plane. There are several approaches for modeling complex-valued time series such as Jumarie [3] and Hochberg

and Orsingher [4]. Several bivariate models using complex-valued coefficients are illustrated, for example, bivariate random walks whose increments are complex-valued autoregressive models. The other example is the complex-valued process whose spectral density is given by a circular density function of the sine-skewed wrapped Cauchy distribution [1].

Related to the time series modeling whose spectral density function is the circular density, an extension of modeling time series with cylindrical distribution is also considered. Remind that the autoregressive process with order p has the roots of the characteristic polynomials outside the unit circle. Considering the roots of the characteristic polynomials which take values on a cylinder, a different aspect of autoregressive models can be constructed. In addition, modeling autocorrelation function (ACF) for the real-valued time series uses the inverse of the roots of the characteristic polynomials, and this fact motivates us to consider the modeling ACF of the autoregressive process using disc distribution whose support is inside the unit circle.

All the examples introduced in this study are applied to real data analysis. For this, the maximum likelihood estimation (MLE) is considered and its asymptotic properties are provided. Owing to the results obtained by Miyata et al. [7], some time series models do not necessarily impose an identifiability assumption, which makes clear the asymptotic properties of MLEs. To cope with non-stationarity, we also introduce a state space modeling of proposed time series models where we apply time-varying parameters models.

References

[1] Abe, T., and Pewsey, A. (2011). Sine-skewed circular distributions. *Statistical Paper*, **52**, 683–707.

[2] Abe, T., Ogata, H., Shiohama, T. and Taniai, H. (2017). A circular autocorrelation of stationary circular Markov processes. *Statistical Inference for Stochastic Processes*, **20**, 275–290.

[3] Hochberg, K. J. and Orsingher, E. (1996). Composition of stochastic process governed by higher-order parabolic and hyperbolic equations. *Journal of Theoretical Probability*, **9**(2), 511–532.

[4] Jumarie, G. (1999). Complex-valued Wiener measure: An approach via random walk in the complex plane. *Statistics & Probability Letters*, *42*(1), 61–67.

[5] Kato, S. and Jones, M. C. (2013). An extended family of circular distributions related to wrapped Cauchy distributions via Brownian motion. *Bernoulli*, **19**, 154–171.

[6] Kato, S. and Jones, M. C. (2015). A tractable and interpretable four-parameter family of unimodal distributions on the circle. *Biometrika* **102**, 181–190.

[7] Miyata, Y., Shiohama, T., and Abe, T. (2022). Identifiability of asymmetric circular and cylindrical distributions. To appear in *Sankyha A*.

[8] Ogata, H., and Shiohama. T. (2022). A mixture transition modeling for higher-order circular Markov processes. *Preparation*

[9] Taniguchi, M., Kato, S., Ogata, H., and, Pewsey, A. (2020). Models for circular data form time series spectra. *Journal of Time Series Analysis*, **41**, 808–829.

ADISTA22: Advances in Directional Statistics
Universidade de Santiago de Compostela, 2022, ISBN 978-84-19155-92-4

Learning Torus PCA Based Classification for Multiscale RNA Correction with Application to SARS-CoV-2

H. Wiechers[1,*], B. Eltzner[2], K. V. Mardia[3] and S. F. Huckemann [1]

[1] *Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, Georgia-Augusta-University, Göttingen, 37077, Germany,*
[2] *Max Planck Institute for Biophysical Chemistry, Göttingen, 37077, Germany, and*
[3] *Department of Statistics, School of Mathematics, University of Leeds, LS2 9JT, England.*
[*] *Corresponding author*

Abstract. Reconstructions of structure of biomolecules, for instance via X-ray crystallography or cryo-EM frequently contain clashes of atomic centers which are not chemically permissible. Methods to correct these clashes are usually based on simulations approximating biophysical chemistry, making them computationally expensive and often not correcting all clashes. Using RNA data, we propose fast, data-driven multiscale learning reconstructions from clash free RNA benchmark data. Multiscale here is based on two levels of shape analysis for RNA geometry. The shape of RNA at microscopic scale (suites) can be described by dihedral angles of the backbone leading to appropriate landmarks at another scale, the mesoscopic scale which is described by landmarks obtained as centers of so-called sugar rings. Based on our analysis that concentrated neighborhoods at mesoscopic scale closely relate to clusters at microscopic scale, we correct within-suite-backbone-to-backbone clashes exploiting Fréchet means at the two scales; one uses angular shape (microscopic), the other uses size-and-shape (mesoscopic). We validate our reconstructions by showing that the classes learned are in high correspondence with clusters from existing literature and that the reconstructions proposed are well within physical resolution limits. While this method is general for RNA we illustrate its power by the cutting-edge RNA example of SARS-CoV-2.

Keywords: Angular Shape Analysis; Frameshift Stimulation Element; Fréchet and Procrustes Means; Geodesic Projection; Size-and-shape Space.

References

[1] Wiechers, H., Eltzner, B., Mardia, K. V. and Huckemann, S. F. (2021). MLearning torus PCA based classification for multiscale RNA backbone structure correction with application to SARS-CoV-2. *bioRxiv*.

[2] Mardia, K. V., Wiechers, H., Eltzner, B. and Huckemann, S. F. (2021). Principal component analysis and clustering on manifolds. *Journal of Multivariate Analysis*.

List of participants

C U R S O S E C O N G R E S O S

Nº 268

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA