ROBUST ISSUES IN ESTIMATING MODELS FOR MULTIVARIATE TORUS DATA

Claudio Agostinelli [1], Giovanni Saraceno [1] and Luca Greco [2]

[1] Department of Mathematics, University of Trento, (e-mail: claudio.agostinelli@unitn.it, giovanni.saraceno@unitn.it)

[2] University Giustino Fortunato, Benevento (e-mail: l.greco@unifortunato.eu)

ABSTRACT: We consider the problem of robust fitting for statistical models applied to multivariate torus data, e.g., data which are multivariate angles. We discuss two different definitions of outliers, "geometric" and "probabilistic" outliers, and the proposed robust methods to cope with them. We mainly focus on multivariate wrapped models together with some computational aspects.

KEYWORDS: circular data, multivariate torus data, outlier detection, robust estimation, wrapped models

1 Introduction

Multivariate circular data arise commonly in many different fields. Depending on the situation, observations can be thought as points on the surface of a hyper-sphere ($\mathbb{S}^{p-1}$) or as points on the surface of a torus ($\mathbb{T}^p = [0, 2\pi)^p$). While the first problem is well studied in literature, the latter received much less attention, even though it is more common. Here, we review some aspects of robust fitting of torus data according to wrapped models. The peculiarity of multivariate torus data is periodicity, that reflects in the boundedness of the sample space and often of the parametric space. Indeed, it is challenging to introduce the *geometric* concept of outliers, as points that are far from the bulk of the data. However, it is always possible to define circular outliers from a *probabilistic* point of view, as points that are unlikely to occur under the assumed model. Notice that outliers are model dependent, since they are defined with respect to the specified model. A first general attempt to develop a robust parametric technique for multivariate torus data can be found in Saraceno *et al.*, 2021 where a weighted likelihood estimator is introduced and outliers are defined using the probabilistic point of view. In contrast, Greco *et al.*, 2021 develop robust estimators based on S/M/MM-estimators as well as weighted likelihood estimators considering the geometric approach.

2 Wrapped models

Let $\mathbf{X}$ be a multivariate random variable with model density $m(\mathbf{x};\theta)$ on $\mathbb{R}^p$ parameterized by $\theta \in \Theta$. We can construct a wrapped model by $\mathbf{Y} = \mathbf{X} \bmod 2\pi$ where the mod operator is performed component-wise. The density function of $\mathbf{Y}$ takes the form of an infinite sum over $\mathbb{Z}^p$ given by

$$m^{\circ}(\mathbf{y};\theta) = \sum_{\mathbf{j}\in\mathbb{Z}^p} m(\mathbf{y}+2\pi\mathbf{j};\theta) \ .$$

A good approximation, denoted as m_J°, can be obtained, in most cases, with only few terms of the summation, so that $\mathbb{Z}^p$ is replaced by $\mathcal{C}_{\mathcal{J}} = \otimes_{s=1}^{p}\mathcal{J}$ where $\mathcal{J} = (-J,-J+1,\ldots,0,\ldots,J-1,J)$ for some fixed J. The support of $\mathbf{Y}$ is bounded and given by $[0,2\pi)^p$, for convenience, and the parametric space Θ might be restricted as well to ensure identifiability. The $p-$dimensional vector $\mathbf{j}$ represents the wrapping coefficients vector, that is, it indicates how many times each component of the $p-$toroidal data point has been wrapped. Given a sample $(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n)$, the approximated log-likelihood function is given by

$$\ell(\theta) = \sum_{i=1}^{n} \log m_J^{\circ}(\mathbf{y}_i;\theta) = \sum_{i=1}^{n} \log \sum_{\mathbf{j}\in\mathcal{C}_{\mathcal{J}}} m(\mathbf{y}_i+2\pi\mathbf{j};\theta) \ .$$

Assuming that we could observe the vectors $\mathbf{j}_i$ $(i=1,\ldots,n)$, then we would have access to the unwrapped and unobserved sample $\hat{\mathbf{x}}_i = \mathbf{y}_i + 2\pi\mathbf{j}_i$. This leads to the following log-likelihood

$$\ell_C(\theta) = \sum_{i=1}^{n} \log m(\hat{\mathbf{x}}_i;\theta) = \sum_{i=1}^{n} \log m(\mathbf{y}_i+2\pi\mathbf{j}_i;\theta) = \sum_{i=1}^{n}\sum_{\mathbf{j}\in\mathcal{C}_{\mathcal{J}}} v_{i\mathbf{j}} \log m(\mathbf{y}_i+2\pi\mathbf{j};\theta) \ ,$$

where $\mathbf{v}_{i\mathbf{j}} = 1$ or $\mathbf{v}_{i\mathbf{j}} = 0$ according to whether $\mathbf{y}_i$ has $\mathbf{j} \in \mathcal{C}_{\mathcal{J}}$ as the wrapping coefficient vector and now the $\mathbf{j}_i$s are additional unknown parameters needed to be estimated. Optimization of the above log-likelihood can be performed naturally through a Classification-Expectation-Maximization algorithm, see Nodehi *et al.*, 2021 for more details. Hereafter, we concentrate on unimodal and elliptically symmetric densities m, i.e., given a strictly decreasing and non-negative function h and set $\theta = (\mu,\Sigma)$ for a location vector parameter μ and dispersion matrix Σ, then $m(\mathbf{x};\theta) \propto |\Sigma|^{-1/2} h\left((\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right)$.

3 Outliers in multivariate torus data

Consider $0 \leq \varepsilon < 0.5$ and an arbitrary distribution $g(\mathbf{x})$ in $\mathbb{R}^p$. According to the usual gross error model, the true density $f(\mathbf{x})$ of the data is given by

$f(\mathbf{x}) = (1-\varepsilon)m(\mathbf{x};\mu,\Sigma)+\varepsilon g(\mathbf{x})$ and hence the corresponding wrapped density would have the form

$$f^{\circ}(\mathbf{y}) = (1-\varepsilon)\sum_{\mathbf{j}\in\mathbb{Z}^p} m(\mathbf{y}+2\pi\mathbf{j};\mu,\Sigma)+\varepsilon\sum_{\mathbf{j}\in\mathbb{Z}^p} g(\mathbf{y}+2\pi\mathbf{j}) \quad (1)$$

$$= (1-\varepsilon)m^{\circ}(\mathbf{y};\mu,\Sigma)+\varepsilon g^{\circ}(\mathbf{y}). \quad (2)$$

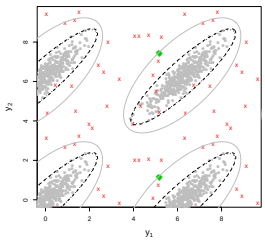
If we instead consider the approach leading to $\ell_C(\mu,\Sigma)$ and equation (1), for a given observation $\mathbf{y}_i$ we have

$$f^{\circ}(\mathbf{y}_i) \approx (1-\varepsilon)m(\mathbf{y}_i+2\pi\mathbf{j}_i;\mu,\Sigma)+\varepsilon g(\mathbf{y}_i+2\pi\mathbf{j}_i)$$

which suggests the classical geometric definition of outliers. In such cases, the degree of outlyingness of an observation is based on some "geometric" distance, e.g., the squared Mahalanobis distance. In contrast, we can define outliers directly on the torus, that is, according to equation (2), based on a "probabilistic" distance [Markatou *et al.*, 1998 and Agostinelli, 2007] where we compare the *true* density $f^{\circ}(\mathbf{y}_i)$ with the model density $m^{\circ}(\mathbf{y}_i;\mu,\Sigma)$. A measure of the agreement is provided by the finite sample Pearson residual function [Lindsay, 1994 and Markatou *et al.*, 1998], defined as $\delta_n(\mathbf{y}) = \frac{\hat{f}_n(\mathbf{y})}{\hat{m}(\mathbf{y};\theta)} - 1$ where $\hat{f}_n(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{y};\mathbf{y}_i,h)$ is a non-parametric kernel density estimate (with kernel function k and bandwidth h) of the true density $f(\mathbf{y})$ and $\hat{m}(\mathbf{y};\mu,\Sigma) = \int k(\mathbf{y};\mathbf{t},h)m(\mathbf{t};\mu,\Sigma)\,d\mathbf{t}$ is a smoothed version of the model density.

4 Example

Here, we illustrate the behavior of the robust estimators introduced in Saraceno *et al.*, 2021 and Greco *et al.*, 2021 using a simulated example. We point the reader to the cited papers for full details. The bulk of data has been drawn from a bivariate wrapped normal distribution with $\mu = 0$, $\Sigma = D^{1/2}RD^{1/2}$ where R is a random correlation matrix and $D = diag(\sigma\mathbf{1}_2)$ with $\sigma = \pi/4$. The sample size is $n = 500$ with 10% of contamination. Two types of outlying observations are considered: scattered and point-mass. It is suggested to represent circular data points after they have been unwrapped on a "flat" torus in the form $\mathbf{x} = \mathbf{y}+2\pi\mathbf{j}$ for $\mathbf{j} \in \mathcal{C}_{\mathcal{J}}$. The figure shows the unwrapped bivariate points (grey points), the scattered (red crosses) and the point-mass (green plus) outliers. The bivariate fitted models are given in the form of ellipses based on the 0.99-level quantile of a χ^2_2 distribution. We show the results obtained using maximum likelihood estimator (grey line) and the proposed robust estimators. In particular, we



	$AS(\hat{\mu})$	$\Delta(\hat{\Sigma})$
MLE	0.001478	2.096517
probabilistic	0.000491	0.002610
geometric	0.000571	0.005987

consider the robust estimators based on the weighted likelihood technique, implemented according to geometric (dotted line) and probabilistic (dashed line) outliers. Finally, the table gives some measures of fitting accuracy.

References

AGOSTINELLI, C. 2007. Robust Estimation for Circular Data. *Computational Statistics and Data Analysis*, **51**(12), 5867–5875.

GRECO, L., SARACENO, G., & AGOSTINELLI, C. 2021. Robust Fitting of a Wrapped Normal Model to Multivariate Circular Data and Outlier Detection. *Stats*, **4**(2), 454–471.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., & STAHEL, W.A. 1986. *Robust Statistics: The Approach based on Influence Functions*. Wiley.

HAWKINS, D. 1980. *Identification of Outliers*. Chapman & Hall.

HE, X. 1992. Robust Statistics of Directional Data: A Survey. *Nonparametric Statistics and Related Topics*, 87–95.

LINDSAY, B.G. 1994. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics*, **22**, 1018–1114.

MARKATOU, M., BASU, A., & LINDSAY, B. G. 1998. Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, **93**(442), 740–750.

NODEHI, A., GOLALIZADEH, M., MAADOOLIAT, M., & AGOSTINELLI, C. 2021. Estimation of parameters in multivariate wrapped models for data on a p-torus. *Computational Statistics*, **36**, 193–215.

SARACENO, G., AGOSTINELLI, C., & GRECO, L. 2021. Robust Estimation for Multivariate Wrapped Models. *Metron*. To appear.