Journal of Multivariate Analysis 165 (2018) 14–26

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

The joint projected normal and skew-normal: A distribution for poly-cylindrical data

Gianluca Mastrantonio

Department of Mathematics, Polytechnic of Turin, Turin, Italy

ARTICLE INFO

Article history:
Received 20 February 2017
Available online 2 December 2017

AMS subject classifications:
62H11
13P25
62F15

Keywords:
Circular data
Circular–linear distribution
Multivariate distribution
Projected normal
Skew-normal

ABSTRACT

This paper introduces a multivariate circular–linear (or poly-cylindrical) distribution obtained by combining the projected and the skew-normal. We show the flexibility of our proposal, its closure under marginalization, and how to quantify multivariate dependence. Due to a non-identifiability issue that our proposal inherits from the projected normal, a computational problem arises. We overcome it in a Bayesian framework, adding suitable latent variables and showing that posterior samples can be obtained with a post-processing of the estimation algorithm output. Under specific prior choices, this approach enables us to implement a Markov chain Monte Carlo algorithm relying only on Gibbs steps, where the updates of the parameters are done as if we were working with a multivariate normal likelihood. The proposed approach can also be used with the projected normal. As a proof of concept, on simulated examples we show the ability of our algorithm in recovering the parameter values and to solve the identification problem. Then the proposal is used in a real data example, where the turning-angles (circular variables) and the logarithm of the step-lengths (linear variables) of four zebras are modeled jointly.

1. Introduction

The analysis of circular data, i.e., observations on the unit circle, requires specific statistical tools since the circular domain is intrinsically different from the real line, and this inhibits the use of standard statistics that, if not properly modified, lead to uninterpretable results; for a general review, see [16,22,34]. A similar type of problem holds for circular densities which, beyond being non-negative and integrating to 1, should possess the property of "invariance" [27], i.e., they must be a location model under the group of rotations and reflections of the circle. This property, which expresses the need of densities that can represent equivalently the same phenomena under different reference systems, is specific to circular densities and it is sometimes overlooked [27].

Circular data are often observed along with "linear" variables, i.e., variables taking values on the real line; they are then called cylindrical in the bivariate case and poly-cylindrical in higher dimension. For example in marine research, wind and wave directions are modeled with wind speed and wave height [8,41,25] and, in ecology, animal behavior is described using measures of speed and direction, e.g., step-length and turning-angle [9,19,33,31]. In most applications, cylindrical data are modeled assuming independence between the circular and linear components; see, e.g., [8,20,30]. Ignoring dependence can lead to misleading inference since we are not considering a component of the data that can help in understanding the phenomenon under study; see, e.g., [28]. In the literature, to date, no poly-cylindrical distributions have been proposed and there are only few distributions for cylindrical data; the best known examples are those in [2,17,23], and the new density of [1]. The aim of this work is to introduce what is, to the best of our knowledge, the first poly-cylindrical distribution.

E-mail address: gianluca.mastrantonio@polito.it.

https://doi.org/10.1016/j.jmva.2017.11.006
0047-259X/

Circular and linear variables live in very different spaces and the definition of a mixed-domain distribution is not easy. The issue is even more complicated if we require flexibility, interpretable parameters and the possibility to define an efficient estimation algorithm that is easy to implement. We choose to adopt a Bayesian framework because, as we show in Section 3.1 that using standard Markov chain Monte Carlo (MCMC) methods, we are able to propose an algorithm with the required characteristics while Monte Carlo (MC) procedures [7,36] allow us to obtain posterior distributions for all the statistics we may need to describe the results.

Since circular observations often exhibit bimodality [38,39], our aim is to propose a distribution with circular marginals that can model such data. In the literature, the best known bimodal circular distributions are the projected normal ($\mathcal{PN}$) [39] and the generalized von Mises [10]. The former can be easily generalized to the multivariate setting and it has an interesting augmented density representation, based on a normal probability density function (pdf), that can be used to define circular–linear dependence. The $\mathcal{PN}$ is very flexible [40,26] with shapes that range from unimodal and symmetric to bimodal and antipodal; it is closed under marginalization and, as we show in Appendix, it possesses the invariance property. However, multivariate extensions of the generalized von Mises are not easy to handle and, in our opinion, it is not straightforward to use it as a component of a poly-cylindrical distribution.

We define our proposal constructively, starting from the $\mathcal{PN}$ and choosing a distribution for the linear component which, taking advantage of the $\mathcal{PN}$ augmented density representation, allows us to define a poly-cylindrical distribution that is flexible enough to model real data and whose parameters can be easily estimated with MCMC algorithms.

For the linear component we use a skew-normal, i.e., a generalization of the Gaussian distribution which allows more flexibility introducing asymmetry in the normal density. Its first univariate version was proposed by Azzalini [4]. Subsequent works introduced multivariate extensions and different formalizations; see, e.g., [5,13,18,37]. Among these, we found that of Sahu et al. [37] (hereafter $\mathcal{SSN}$) interesting: it can be closed under marginalization and it has an augmented density representation that, as the $\mathcal{PN}$, is based on a normal pdf.

Using this particular form of the skew-normal distribution, due to the properties listed above, we are able to define the *joint projected normal and skew-normal* ($\mathcal{JPNSN}$) poly-cylindrical distribution by introducing dependence in the normal pdfs of the augmented representations. The distribution retains the $\mathcal{PN}$ and $\mathcal{SSN}$ as marginal distributions and is closed under marginalization, i.e., any subset of circular and linear variables is $\mathcal{JPNSN}$ distributed. The MCMC algorithm we propose can be based only on Gibbs steps, updating parameters as if we were working with a multivariate normal likelihood. The density cannot be expressed in closed form but we do not consider this an issue from the point of view of model fitting, since we are able to estimate its parameters easily.

The $\mathcal{JPNSN}$ has the same identification problem as the $\mathcal{PN}$ [39], but we show that posterior values can be obtained by a post-processing of the MCMC algorithm based on the non-identifiable likelihood. The proposed algorithm can also be used with the univariate and multivariate $\mathcal{PN}$ and the spherical $\mathcal{PN}$ distribution of [14], solving their identification problem in a new way.

The algorithm, tested on simulated datasets, shows its ability to retrieve the parameters used to simulate the data, and posterior samples do not suffer from an identification issue. We used the $\mathcal{JPNSN}$ to jointly model the logarithm of step-lengths and turning-angles of four zebras observed in Botswana. Through a comparison based on the continuous rank probability scores (CRPSs) [11,12] of our proposal, the cylindrical distribution of [1] and a cylindrical version of the $\mathcal{JPNSN}$, i.e., assuming independence between zebras, we show that ignoring multivariate dependence can lead to loss of predictive ability.

The paper is organized as follows. Section 2 is devoted to the constructive definition of the distribution. In Section 3 we introduce the identification problem and how to estimate the $\mathcal{JPNSN}$ parameters. The proposal is applied to simulated examples in Section 4.1 and the real data application is discussed in Section 4.2. The paper ends with concluding remarks in Section 5. In Appendix we prove the invariance property of the $\mathcal{PN}$ and we show MCMC implementation details.

2. The joint projected normal and skew-normal distribution

In this section we build the poly-cylindrical density by first introducing the circular and linear marginals and then showing how to induce dependence.

2.1. The projected normal distribution

The $\mathcal{PN}$ is a distribution for a p-dimensional vector $\mathbf{\Theta} = (\Theta_1, \ldots, \Theta_p)$ of circular variables, i.e., $\Theta_i \in [0, 2\pi)$ is an angle expressed in radiant, obtained starting from a $2p$-dimensional vector $\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_p)$, where $\mathbf{W}_i = (W_{i1}, W_{i2})^\top \in \mathbb{R}^2$, distributed as a $2p$-variate normal with mean vector $\boldsymbol{\mu}_w$ and covariance matrix $\mathbf{\Sigma}_w$. For each $i \in \{1, \ldots, p\}$, $\mathbf{W}_i$ is normally distributed with parameters $\boldsymbol{\mu}_{w_i}$, and $\mathbf{\Sigma}_{w_i}$ is a point in the 2-dimensional space expressed using the Cartesian system. The same point can be also represented in polar coordinates with the angle Θ_i and the distance vector $R_i \in \mathbb{R}^+$. Between $\mathbf{W}_i$, Θ_i and R_i the following relations exist:

$$\Theta_i = \operatorname{atan}^* (W_{i2}/W_{i1}) \tag{1}$$

and

$$\mathbf{W}_i = R_i \begin{pmatrix} \cos \Theta_i \\ \sin \Theta_i \end{pmatrix}, \quad R_i = \|\mathbf{W}_i\|,$$