Principal component analysis for the wrapped normal torus model

John T. Kent* and Kanti V. Mardia

Department of Statistics, University of Leeds

1 Introduction

Interest in statistical methods for directional data has had something of a renaissance in recent years through new applications in bioinformatics such as protein structure. The protein "backbone" can be largely characterized in terms of a sequence of circular angles, ϕ_j and ψ_j. Thus there is a need for new methods to handle the analysis of multiple angles.

In this paper we shall explore the development of methods for angles analogous to multivariate methods for real-valued data. In particular consider an $n \times p$ data matrix of angles θ_{jk} and consider the problem of principal component analysis to find a low-dimensional summary of the main variability in the data. Since an angle θ can be represented in $\mathbb{R}^2$ as $(\cos\theta,\ \sin\theta)$, there is a natural Euclidean representation of the data in $2p$ dimensions. However, a $2p$-dimensional principal component analyis seems excessive since there are only p degrees of freedom in the original angular dataset.

2 Moments for the wrapped normal torus distribution

Suppose that $\boldsymbol{\theta}$ is a vector of angles following a wrapped normal torus distribution; that is, $\theta_j = x_j \mod 2\pi,\ j = 1, \ldots, p$, where $\mathbf{x} \sim N_p(0, \Sigma)$. One reason for choosing this distribution is that the trigonometric moments have straightforward explicit expressions. If $\boldsymbol{\delta}$ is a $p \times 1$ vector with integer coefficients, then

$$\begin{aligned} E\{\cos(\boldsymbol{\delta}^T\boldsymbol{\theta})\} &= \exp\{-\frac{1}{2}\boldsymbol{\delta}^T\Sigma\boldsymbol{\delta}\} \\ E\{\sin(\boldsymbol{\delta}^T\boldsymbol{\theta})\} &= 0. \end{aligned}$$

In particular, for $j, k = 1, \ldots, p$,

$$\begin{aligned} E\{\cos(\theta_j)\} &= \exp\{-\frac{1}{2}\sigma_{jj}\} = c_j,\ \text{say}, \\ E\{\sin(\theta_j)\} &= 0, \\ E\{\cos(\theta_j \pm \theta_k)\} &= \exp\{-\frac{1}{2}(\sigma_{jj} \pm 2\sigma_{jk} + \sigma_{kk})\} \\ E\{\sin(\theta_j \pm \theta_k)\} &= 0. \end{aligned}$$

Combining the two versions of the last two equations yields

$$\begin{aligned} E\{\cos(\theta_j)\cos(\theta_k)\} &= c_j c_k \cosh(\sigma_{jk}) = a_{jk},\ \text{say}, \\ E\{\sin(\theta_j)\sin(\theta_k)\} &= c_j c_k \sinh(\sigma_{jk}) = b_{jk},\ \text{say}, \\ E\{\sin(\theta_j)\cos(\theta_k)\} &= 0. \end{aligned} \tag{1}$$

Store the coefficients $\{c_j\}$ as a vector $\mathbf{c}$ and the coefficients $\{a_{jk}\}$ and $\{b_{jk}\}$ as matrices A and B. Write $D = \text{diag}(\mathbf{c})$. In matrix form the covariance matrices for the cosines and sines take the form

$$\text{var}(\cos\boldsymbol{\theta}) = DAD - \mathbf{c}\mathbf{c}^T, \quad \text{var}(\sin\boldsymbol{\theta}) = DBD, \quad \text{cov}(\cos\boldsymbol{\theta}, \sin\boldsymbol{\theta}) = 0. \tag{2}$$

Thus Σ can be recovered from the trigonometric moments through the equation

$$\Sigma = \sinh^{-1}(D^{-1}\text{var}(\sin\boldsymbol{\theta})D^{-1}). \tag{3}$$

Here the notation $\sinh^{-1}(\cdot)$ applied to a matrix means that the inverse sinh function, $\sinh^{-1}(u) = \log(u + \sqrt{u^2+1})$, is applied to each element of the matrix.

These results suggest a method to estimate Σ from an $n \times p$ matrix of torus data.

(a) Calculate the sample first order trigonometric moments for the p angles, and rotate each angle so that the resultant vector points towards the positive horizontal axis.

(b) Calculate the sample second trigonometric moments corresponding to (1) and use (3) to produce an estimate of Σ.

If Σ is small (formally, write $\Sigma = \epsilon\Sigma_0$ for a fixed positive definite matrix Σ_0 and let ϵ get small), then $c_j \approx 1$ for all j and $\Sigma \approx B$. Further the three p-dimensional vectors $\sin\boldsymbol{\theta} \approx \boldsymbol{\theta} \approx \mathbf{x}$ are approximately the same as one another (treating each angle θ_j as a number in $[-\pi, \pi)$), and hence all have approximately the same covariance matrix.

3 Angular PCA

1. There are several ways in which the sine and cosine covariances might be used in a principal component analysis, including the following.

   (a) Mardia et al. (1996) suggested PCA on the p-dimensional vector $\sin\boldsymbol{\theta}$. For the wrapped normal torus model, the covariance matrix is given in (2).

   (b) Mu et al. (2005) have suggested PCA on the $2p$-dimensional vector $(\sin\boldsymbol{\theta}^T, \cos\boldsymbol{\theta}^T)^T$.

   (c) Another possibility is to pool the cosine and sine information. For the wrapped normal torus model described above, the pooled covariance matrix takes the form

   $$P = DAD - \mathbf{c}\mathbf{c}^T + DBD = D\{\exp(\Sigma) - \mathbf{1}_p\mathbf{1}_p^T\}D,$$

   where $\mathbf{1}_p$ is a p-dimensional vector of ones, and $\exp(\Sigma)$ is a $p \times p$ matrix with elements $\exp(\sigma_{jk})$. This idea is closely related to the complex PCA based on the complex vector $\exp(i\boldsymbol{\theta})$ suggested by Altis et al. (2007) and Altis et al. (2008).

   The moment results derived above will help to understand differences between these approaches more thoroughly.

2. An important use of PCA is as an exploratory method to detect clustering. A mixture of two or more wrapped normal torus distributions with a fixed number of components provides a mixture distribution with tractable moments, which in turn facilitates the use of the EM algorithm to estimate the parameters.

3. One disadvantage of the wrapped normal torus distribution is that it does not form an exponential family, unlike the sine and cosine models described e.g. in Mardia et al. (2008). In particular, there are no conjugate priors to faciliate a Bayesian analysis.

Acknowledgements

We are grateful to Mike Gilson, University of Maryland Biotechnology Institute, for drawing our attention to the importance of this problem in bioinformatics, and to Zhengzheng Zhang and Charles Taylor for helpful discussions.

References

Altis, A., Nguyen, P.H. , Hegger, R. and Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. (2007) *J. Chem. Physics*, **126**, 244111.

Altis, A., Otten, M., Nguyen, P.H., Hegger, R. and Stock, G. (2008). Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.*, **128**, 245102.

Mardia, K.V., Coombes, A., Kirkbride, J., Linney, A., and Bowie, J.L. (1996). On Statistical Problems with Face Identification from Photographs. *J. Appl. Stat.* **23**, 655–675.

Mardia, K.V., Hughes, G., Taylor, C.C. and Singh, H. (2008). Multivariate von Mises distribution with applications to bioinformatics. *Can. J. Statist.* **36**, 99–109.

Mu, Y., Nguyen, P. H. and Stock, G. (2005). Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *PROTEINS: Structure, Function, and Bioinformatics*,, **58**, 45–52.