Title: Multivariate exponential family principal component analysis (**Research**)

Advisors:

Hien Nguyen (h.nguyen5@latrobe.edu.au; +61 04 34 25 39 52)
Florence Forbes (florence.forbes@inria.fr; +33 04 76 61 52 50)

Laboratories:

Department of Mathematics and Statistics, La Trobe University, Melbourne, Australia
LJK, Mistis, Inria Grenoble Rhone-Alpes, Montbonnot

Location: the research will ideally take place in part or totality in Melbourne Australia under the co-supervision of the Inria Mistis team in Grenoble. Specific arrangements can be negociated with the applicant.

Project description:

Principal component analysis (PCA) is a ubiquitous analytic tool for the visualization and analysis of highly multivariate data in all areas of research where quantitative analysis is conducted. See [1] for a review on the standard PCA method.

The standard PCA method assumes the fact that the multivariate data are real valued and Gaussian distributed (cf. [2]). Unfortunately this assumption is unreasonable for many data types that one typically finds in applications, such as binary, positive-valued, or bounded-valued data.

In [2], it was suggested that one could extend the standard PCA method to the analysis of these alternative data types via inverse regression using generalized linear models (GLMs; [3]) based on 1-parameter families of exponential distributions and maximum likelihood estimation, with regression coefficients being constrained to the Grassmann matrix manifold (cf. [4]).

Traditionally, optimization problems that constrain the parameter space to manifolds have been difficult to solve. However, recent innovations, such as the ManifoldOptim package for R [5], have made solving such problems significantly easier.

This project proposes to use ManifoldOptim and similar tools to develop novel algorithms for conducting PCA on non-standard data types. The core aim of the project is to then implement these algorithms as an R package, that will be made available for the scientific public to use.

The ideal candidate for this project will have experience with optimization, regression analysis, matrix algebra, and programming in R. Upon completion of this project, there are numerous directions for further research, and the project can lead directly into a PhD thesis.

Bibliography:

[1] Jolliffe I. Principal component analysis. In International encyclopedia of statistical science 2011 (pp. 1094-1096). Springer, Berlin, Heidelberg.

[2] Cook RD. Fisher lecture: Dimension reduction in regression. Statistical Science. 2007;22(1):1-26.

[3] McCullagh P, Nelder JA. Generalised linear models 2nd edn. Monographs on statistics and applied probabilitity. 1989.

[4] Edelman A, Arias TA, Smith ST. The geometry of algorithms with orthogonality constraints. SIAM journal on Matrix Analysis and Applications. 1998;20(2):303-53.

[5] Martin S, Raim AM, Huang W, Adragni KP. ManifoldOptim: An R interface to the ROPTLIB library for Riemannian manifold optimization. arXiv preprint arXiv:1612.03930. 2016 Dec 12.