Procrustes statistics for unlabelled points and applications

K.V. Mardia & V. Nyirongo*

University of Leeds

We consider the Procrustes statistics for a form arising from matching problems in Bioinformatics. The Procrustes statistic is related to RMSD except for a divisor. In practice, the unlabelled (atom) points are matched and one is interested in understanding how good is the resulting match.

We can consider this final match as drawn from $x_j \sim \mathrm{N}(\mu_i, \sigma^2 I_3)$ for two configurations. Thus, we can derive the p-values, giving some indication of the goodness of fit. For concentrated data, the statistic has a χ^2 distribution. The harder problem is finding the distribution of the minimum Procrustes statistic when the points are unlabelled. We will discuss this problem and the inherent difficulty. For illustrative purposes, we will use Gaussian configurations on a line.

References

Dryden, I.L. and Mardia, K.V. (1998). *Statistical Shape Analysis*. Chichester, John Wiley.

Eidhammer, I. and Jonassen, I. and Taylor, W.R. (2004). *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis* . New Jersey, John Wiley.

Gold, N.D. (2003). Computational approaches to similarity searching in a functional site database for protein function prediction. *Ph.D Thesis*. Leeds University Press.

Goodall, C.R. and Mardia, K.V. (1992). The noncentral Bartlett decompositions and shape densities. *Journal of Multivariate Analysis*, **40**, 94-108.

Goodall, C.R. and Mardia, K.V. (1993). Multivariate aspects of shape theory. *Annals of Statistics*, **21**, 848-866.

Taylor, C.C., Mardia, K.V. and Kent, J.T. (2003). Matching Unlabelled Configurations Using the EM Algorithm. *Proceedings in Stochastic Geometry, Biological Structure and Images*, 19-21. Edited by R.G. Aykroyd, K.V. Mardia and M.J. Langdon. Leeds University Press.