Dissimilarity Data

Grace Wahba
University of Wisconsin-Madison

June 9, 2012

Many scientific data sets have only similarity/dissimilarity information between objects or subjects of interest - including BLAST scores as a measure of pairwise similarity between proteins, genetic distance between persons in pedigrees, or between genetic marker data between study objects (plants, animals, persons), nodes in a network, and so forth. We review three papers [1] [2] [3] that utilize dissimilarity information in statistical modeling. Then we review the amazing paper [4] which investigates the the joint distribution of a $p+q$ dimensional probability distribution F_{p+q} whose support is contained in Euclidean $p+q$ space and provides a principled way of testing the null hypothesis that it can be factored into F_pF_q, that is, the first p components are independent of the last q components. The results are completely nonparametric, in that there are essentially no limitations on the specific form of the distribution(s). If there is a sample of size n from the joint distribution, the test statistic is a function of the $\binom{n}{2}$ pairwise (Euclidean) distances between all pairs of observations. Some work in progress with Jing Kong and others study adapting the method of [4] to examine correlations in certain situations when the pairwise observational information is not originally Euclidean.

References

[1] F. Lu, S. Keles, S. Wright, and G. Wahba. A framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences*, 102:12332–12337, 2005. Open Source at www.pnas.org/content/102/35/12332, PMCID: PMC118947.

[2] F. Lu, Y. Lin, and G. Wahba. Robust manifold unfolding with kernel regularization. Technical Report 1008, Department of Statistics, University of Wisconsin, Madison WI, 2005.

[3] H. Corrada Bravo, G. Wahba, K. E. Lee, B. E. K. Klein, R. Klein, and S. K. Iyengar. Examining the relative influence of familial, genetic and environmental covariate information in flexible risk models. *Proceedings of*

the National Academy of Sciences, 106:8128–8133, 2009. PMCID: PMC 2677979.

[4] G. Szekely and M. Rizzo. Brownian distance covariance. *Ann. Appl. Statist.*, 3:1236–1265, 2009.