Object oriented data analysis, power metrics, and graph Laplacians: Discussion of the paper by Tavakoli et al.

Ian L. Dryden, Simon P. Preston and Katie E. Severn,
University of Nottingham

Object Oriented Data Analysis (OODA) has seen many developments over the past decade since Wang and Marron (2007) introduced the topic, and a broad overview of the field has been given by Marron and Alonso (2014). At the heart of OODA is the need to make choices about i) what the data objects are, ii) the conceptual space in which the data objects lie, and iii) the feature space that is used for practical data analysis. The paper by Tavakoli et al. (2019) provides and an excellent exemplar of this approach to statistical analysis. Highly informed preprocessing steps are carried out, based on substantial experience from the application field. The data objects of interest are infinite dimensional covariance operators in a Hilbert space and practical analysis is carried out in the feature space of finite dimensional covariance matrices.

The main methodological contribution is the introduction of the d-covariance, which is the symmetric positive semi-definite matrix Ω that minimizes the expected squared distance of Ω to the random quantity $(X-\mu)(X-\mu)^T$, where $E[X]=\mu$. The need for the development of d-covariance is reminiscent of issues that have arisen previously in statistical shape analysis, where procedures such as Procrustes estimation produced inconsistent estimates in general for population mean shapes from Gaussian landmark models (Lele, 1993). The issue was addressed by estimating a different and more suitable population quantity, namely the Fréchet mean shape, which can be estimated consistently (e.g. see Dryden and Mardia, 2016, Chapter 13). The motivation for the new methods in Tavakoli et al. (2019) is quite similar, where the population quantity of interest has been substituted from the usual covariance matrix to the the more suitable d-covariance, which can be estimated consistently. Of course which population quantity is of most practical interest will depend heavily on the application, and the traditional covariance will always have some appeal given its traditional role in statistics.

Dryden et al. (2009) and Pigoli et al. (2014) considered the power Euclidean distances between pairs of covariance matrices and infinite dimensional covariance operators, respectively. In particular the power Euclidean metric between covariance matrices A and B is

$$d_\alpha(A,B)=|||A^\alpha-B^\alpha|||, \tag{1}$$

where $|||A|||$ is the Frobenius norm of A, and like in Tavakoli et al. (2019) the majority

of attention has been focussed on the square root scale when $\alpha = \frac{1}{2}$. If $\alpha = \frac{k}{2^m}$ where $k \geq 1, m \geq 0$ are positive integers then, using the relation $\sqrt{xx^T} = \frac{xx^T}{|x|}$ repeatedly, one can show that the d-covariance is given by

$$\mathrm{cov}_{d_\alpha} = \left\{ E\left[\frac{(X-\mu)(X-\mu)^T}{|X-\mu|^{2-2\alpha}} \right] \right\}^{\frac{1}{\alpha}}, \tag{2}$$

assuming the required moments exist. Note that any real $\alpha > 0$ can be approximated arbitrarily closely by an expression $\frac{k}{2^m}$, as the dyadic rationals are dense in $\mathbb{R}$. As commented by the authors, the effect of the regularisation can be seen in the divisor in (2). It will be interesting to develop the results of Tavakoli et al. (2019) in cases other than $\alpha = \frac{1}{2}$, for example $\alpha = \frac{3}{4}$ could provide a reasonable compromise between the popular $\alpha = \frac{1}{2}$ and the conventional $\alpha = 1$ case. Dryden et al. (2009) and Pigoli et al. (2014) also consider a variant of (1) with an additional Procrustes rotation, which has connections to the Wasserstein metric between zero mean Gaussian processes (Masarotto et al., 2018) and the Bures distance in quantum statistics. It will be interesting to explore the d-covariance for the Procrustes-type metrics.

Tavakoli et al. (2019) note that their work complements approaches for text-based analysis. In text-based corpus analysis, widely studied are word *collocations* (Gablasova et al., 2017), i.e., words that have a tendency to co-occur; and text documents represented as word-pair co-occurrence counts can be identified as networks (Severn et al., 2019). Analysis of networks is another type of OODA analysis, with wide applications in neuroscience and genetics, besides text analysis. Let $G_m = (V, E)$, comprise a set of nodes, $V = \{v_1, v_2, \ldots, v_m\}$, and a set of edge weights, $E = \{w_{ij} : w_{ij} \geq 0, 1 \leq i, j \leq m\}$, indicating nodes v_i and v_j are either connected by an edge of weight $w_{ij} > 0$, or else unconnected if $w_{ij} = 0$, and $w_{ij} = w_{ji}$ and $w_{ii} = 0$ (undirected and without loops). Any such network can be identified with its $m \times m$ graph Laplacian matrix $\mathbf{L} = (l_{ij})$, defined as

$$l_{ij} = \begin{cases} -w_{ij}, & \text{if } i \neq j \\ \sum_{k \neq i} w_{ik}, & \text{if } i = j \end{cases}$$

for $1 \leq i, j \leq m$. The space of graph Laplacians is a subset of the cone of symmetric positive semi-definite matrices (Ginestet et al., 2017). Graph Laplacians are finite dimensional matrices, but they often represent networks with large m and in that case the situation is similar to that of Tavakoli et al. (2019) except the space is restricted to the graph Laplacian subspace. Severn et al. (2019) introduce a framework for manifold value data analysis of networks that uses the Euclidean power distance (1) and introduces a unique projection from the space of covariance matrices to the subspace of graph Laplacians. The projection can be computed efficiently using quadratic programming, and further statistical analysis such as regression, principal components analysis and hypothesis testing has been developed.

One question in any OODA is "what is the best choice of metric?" The answer will very much depend on the application, the object space, and the particular criteria the user specifies for a well-performing metric. In compositional data analysis, in which the object space is the unit simplex, debate over the choice of metric has been long-running (Scealy

and Welsh, 2014), especially over whether the Euclidean metric or the log-ratio metric due to Aitchison (1983) should be preferred. Tsagris et al. (2016) introduce a one-parameter family of metrics for compositional data that involves a Box–Cox-type parameter α to be estimated from the data (Box and Cox, 1964), with the Euclidean and log-ratio metrics corresponding to special cases $\alpha = 1$ and $\alpha = 0$, respectively. For classification tasks involving various compositional data sets, the optimal value of α is frequently between 1 and 0 reflecting benefit from a compromise between Euclidean and log-ratio metrics. A data-driven choice could similarly be used to estimate the α in (1). Such as procedure is considered by Dryden et al. (2010) for covariance matrices, where α is chosen to make the power transformed data as Gaussian as possible.

References

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, 70(1):57–65.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. (With discussion). *J. Roy. Statist. Soc. Ser. B*, 26:211–252.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.*, 3(3):1102–1123.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis, with Applications in R, 2nd edition*. Wiley, Chichester.

Dryden, I. L., Pennec, X., and Peyrat, J.-M. (2010). Power Euclidean metrics for covariance matrices with application to diffusion tensor imaging. *arXiv e-prints*, page arXiv:1009.3045.

Gablasova, D., Brezina, V., and McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, 67:155–179.

Ginestet, C. E., Li, J., Balachandran, P., Rosenberg, S., and Kolaczyk, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.*, 11(2):725–750.

Lele, S. (1993). Euclidean distance matrix analysis (EDMA): estimation of mean form and mean form difference. *Math. Geol.*, 25(5):573–602.

Marron, J. S. and Alonso, A. M. (2014). Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753.

Masarotto, V., Panaretos, V. M., and Zemel, Y. (2018). Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. *Sankhya A*.

Pigoli, D., Aston, J. A. D., Dryden, I. L., and Secchi, P. (2014). Distances and inference for covariance operators. *Biometrika*, 101(2):409–422.

Scealy, J. and Welsh, A. (2014). Colours and cocktails: Compositional data analysis 2013 lancaster lecture. *Australian & New Zealand Journal of Statistics*, 56(2):145–169.

Severn, K. E., Dryden, I. L., and Preston, S. P. (2019). Manifold valued data analysis of samples of networks, with applications in corpus linguistics. *arXiv e-prints*, page arXiv:1902.08290.

Tavakoli, S., Pigoli, D., Aston, J. A. D., and Coleman, J. S. (2019). A spatial modeling approach for linguistic object data: analysing dialect sound variations across great britain. *Journal of the American Statistical Association*. To appear.

Tsagris, M., Preston, S., and Wood, A. (2016). Improved classification for compositional data using the α-transformation. *Journal of Classification*, 33(2):243–261.

Wang, H. and Marron, J. S. (2007). Object oriented data analysis: sets of trees. *Ann. Statist.*, 35(5):1849–1873.