Discussion of: A spatial modeling approach for linguistic object data: Analyzing dialect sound variations across Great Britain, by Shahin Tavakoli et al.

Alexander Petersen[a1] and Hans-Georg Müller[b2]

[a] Department of Statistics and Probability, University of California, Santa Barbara

[b] Department of Statistics, University of California, Davis

April 2019

[1]Research supported by NSF Grant DMS-1811888

[2]Research supported by NSF Grant DMS-1712864.

Congratulations to Tavakoli et al. (2019) for their timely and interesting contribution to the emerging field of object data, as exemplified by covariance objects in the context of linguistic analysis. Their paper is an outstanding example of how challenges in a specific application area can lead to new ideas and concepts in statistical methodology and data analysis, which in turn provide new insights for the applied area. It thus reflects some of the best features of interdisciplinary work. The idea of modeling relationships between languages and dialects in terms of spoken language is far from straightforward to implement. It leads essentially to functional data where the structure of the data is assumed to vary with location. As Tavakoli et al. (2019), in the following abbreviated as TPAC (Tavakoli, Pigoli, Aston, Coleman), nicely demonstrate, the spoken language perspective then requires that one deal with complex data objects and to solve the associated statistical challenges. After suitable preprocessing of the raw signals, one arrives at means and covariances. A particularly interesting feature of the data challenge tackled by TPAC is that these objects are spatially indexed.

There seems currently no universal term for the kind of highly complex and non-Euclidean data that are increasingly collected and of which the spatially indexed covariances encountered in TPAC are an example. In addition to *Object Data*, this area has been described as *Object Oriented Data Analysis* (Wang et al. 2007; Marron and Alonso 2014), in analogy to *Functional Data Analysis*, a term that was introduced by Jim Ramsay (Ramsay and Silverman 2005) for a specific type of object data, where one considers curves and functions as data objects. Key ideas of functional data analysis actually originated much much earlier (Grenander 1950; Rao 1958; Kleffe 1973), but for a long time these were not associated with a catchy name (for further details, see, e.g., Wang et al. 2016). Functional data analysis has the advantage that usually the functions are viewed as elements of a Hilbert space, and thus of a vector space with an inner product, which makes it easy to work with projections such as basis expansions, while neither an in-

ner product nor a vector space structure are usually assumed when one deals with more general types of data objects, such as covariances.

To emphasize that one aims at methodology for samples of complex data, we refer to random variables that take values in metric spaces as random objects and a simple reference to this area is therefore *Random Objects* (Müller 2016). Yet another term for the analysis of object data is *Fréchet analysis*, which includes Fréchet integration (Petersen and Müller 2016), Fréchet regression (Petersen and Müller 2019) and Fréchet analysis of variance (Dubey and Müller 2019), and primarily refers to random objects that are metric-space valued. Fréchet analysis builds upon the fundamental contributions of Fréchet in this area, specifically his extension of the notion of a mean from Euclidean to general metric spaces (Fréchet 1948). Almost any analysis of random objects involves Fréchet means, and indeed TPAC also adopt this basic concept.

Generally, two approaches have emerged for random objects. In the first approach, which encompasses a large class of practical examples, random objects correspond to data points on a manifold or, more generally, a manifold stratified space (Marron and Alonso 2014). In this case, one can make use of local tangent approximations (Bhattacharya and Patrangenaru 2003; Yuan et al. 2012) and geodesic distances to formulate methods that, to some extent, behave similarly to their Euclidean counterparts. In the more general case where random objects are viewed as elements of a metric space, one has not much more than a distance to go by (Székely et al. 2007; Lyons 2013). In the latter case, the scope can be much more general but one has fewer tools available; notably, local linear approximations cannot usually be exploited. Thus, depending on the specific task, the metric approach for random objects can be considerably more challenging; specific features of the metric space and its associated probability measure play a major role, where such features may range from the geometry of the space to certain entropy conditions.

TPAC utilize the metric approach to study the dependence of the mean vector and the covariance matrix of the MFCC vector on spatial location. The estimated mean vector at a location x is obtained by a Nadaraya-Watson kernel estimator in (3.3). Here and also in the corresponding covariance estimator in (3.5), which is also of Nadaraya-Watson type, a problem may arise when obtaining estimates near the boundaries of the domain, especially when they are complex and irregular, as in the case of Great Britain. Boundary problems associated with Nadaraya-Watson type spatial smoothing are known to be often severe and may give rise to undersmoothing bandwidth choices when using prediction error criteria, as choosing larger bandwidths near boundaries can lead to be substantial bias (Müller and Stadtmüller 1999). These effects can be greatly mitigated by instead using smoothers of local linear type. Specifically, for the case of $p-$dimensional predictors, which could include the spatial coordinates considered in TPAC, Petersen and Müller (2019) proposed an extension of Fréchet means to conditional Fréchet means for metric-space valued random objects, including an implementation of these conditional means with local least squares type smoothers.

The idea for this extension is to characterize local linear regression as sequence of weighted Fréchet means, then analyze "bias" and "stochastic deviation" separately (note that these quantities are not defined in the usual sense, due to the metric-space valued nature of the random objects, which means that one cannot form differences). The target is the conditional Fréchet mean

$$m_{\oplus}(x) := E_{\oplus}(Y|X = x) := \text{argmin}_{\omega \in \Omega} E(d^2(Y, \omega)|X = x),$$

where (Ω, d) is the metric space of interest; in TPAC this is the space of covariance matrices with the metric d_S given by d_S-covariance, or alternatively, the Frobenius metric.

Consider for a moment the simpler case where $\Omega = \mathcal{R}^1$, i.e., the responses are scalar,

so that $m_\oplus(x) = m(x) = E(Y|X = x)$ is the usual conditional mean. Then, for predictors $X \in \mathcal{R}^p$ and a distance d_g in the predictor space (where perhaps the Euclidean metric may be adequate as Great Britain is not big, so that the geodesic metric d_g might not be needed), consider the local linear estimate with bandwidth h and univariate kernel K given by

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} K_h(d_g(X_i, x))(Y_i - \beta_0 - \beta_1(X_i - x))^2, \quad K_h(\cdot) = h^{-1}K(\cdot/h).$$

The minimizers can be viewed as M-estimators

$$(\beta_0^*, {\beta_1}^*) \;=\; \underset{\beta_0,\beta_1}{\operatorname{argmin}} \int K_h(d_g(z, x)) \left[\int y \mathrm{d}F_{Y|X}(z, y) - (\beta_0 + \beta_1(z - x)) \right]^2 \mathrm{d}F_X(z),$$

where $F_{X,Y}$ is the joint distribution of (X, Y). Thus, the usual local linear estimate $\hat{m}(x) = \hat{\beta}_0$ targets the localized Fréchet mean

$$\tilde{l}(x) = \beta_0^* = \underset{y\in\mathcal{R}}{\operatorname{argmin}} E\left[(Y - y)^2 s(Y, x, h)\right]$$

for a specific weight function s that can be explicitly computed.

This suggests the generalization to general metric spaces (Ω, d) given by

$$\tilde{l}_\oplus(x) = \underset{\omega\in\Omega}{\operatorname{argmin}} \tilde{L}_n(\omega)$$

where $\tilde{L}_n(\omega) = E\left[s(Y, x, h)d^2(Y, \omega)\right]$ depends on n through the bandwidth $h = h_n$ only. Then we may plug in empirical estimators of these target quantities to obtain

$$\hat{l}_\oplus(x) = \underset{\omega\in\Omega}{\operatorname{argmin}} \hat{L}_n(\omega).$$

This can be viewed as an M-estimator, which makes it possible to harness the theory of empirical processes to obtain asymptotic results. In TPAC, Ω is the space of positive definite matrices and the metric is $d = d_S$. Conceptually, one then has a sample from a random pair (X, Y), where X is a random location and Y the corresponding covariance matrix object. The target is the function $\Sigma(x) = \text{argmin}_{\omega \in \Omega} E(d_S^2(Y, \omega)|X = x)$. Under a twice-differentiability assumption, the above local Fréchet estimator attains the usual pointwise rate of convergence $O_P(n^{-1/3})$ for two-dimensional smoothing, if optimal bandwidths are used. For further details, see Petersen and Müller (2019).

One challenge of the linguistic data considered by TPAC is that covariance matrices Y are not given but must be estimated from n_l repeated measurements of the random vector at location x_l, where it is assumed that $n_l/n \geq c$ for a constant c that does not depend on l. This specifically also requires to estimate the mean vector from the n_l repeats. It turns out that repeated measurements at the same location are actually not needed in order to obtain consistent estimation of $\Sigma(x)$, however the rate of convergence will suffer if only one vector is observed at each location. This has been worked out in detail in the context of an applied analysis of the co-myelination in the developing brain (Petersen et al. 2019), where for each child one has only one or sometimes a small number of MRI recordings at random ages, and at each age one observes the vector of myelination levels across 21 brain regions. To assess the co-myelination process, the target is the covariance matrix of the 21-dimensional myelination level vector as a function of age t, i.e., $\Sigma(t)$. The above described local Fréchet regression with local linear implementation was applied to obtain consistent estimation. We noticed in this context that boundary effects are a major problem when using Nadaraya-Watson kernel smoothing, which we do not recommend for this reason. It seems relatively straightforward to extend this method to the two-dimensional case, and it would be interesting to compare such an extension with the approach presented in TPAC.

Since in the above described method repeated measurements at a given location and the calculation of a covariance matrix from the repeated measurements as an intermediate step are actually not needed, the estimation of $\Sigma(x)$ in the linguistic application can proceed in just one step. In such a single step approach one uses the observed vectors as input rather than first obtaining covariances in an intermediate step. This then weakens somewhat the argument that the metric for constructing the intermediate covariance matrices should be aligned with the metric used for the spatial smoothing, as in the above one-step method one needs to choose only one metric.

While any metric for symmetric positive definite matrices could be used, the implementation of the one-step method in Petersen et al. (2019) proceeded with the Frobenius metric, since the resulting targets then are ordinary covariance matrices that are interpretable as such, while interpretation is tricky when one chooses a different covariance metric. Especially in contexts where the covariance function $\Sigma(x)$ has a direct interpretation, or is used in a secondary analysis such as regression or classification, this can be an important argument in favor of the Frobenius metric. However, this is clearly less of a concern in the linguistic application of TPAC, where the random vectors represent some complex features of the spoken words. Nevertheless, the choice of a good metric for symmetric positive definite matrices is a complex issue, not least as deviations from target criteria themselves depend on the chosen metric (Petersen and Müller 2016).

The solution for the intricate problem of visualizing spatially indexed means and covariances proposed by TPAC is especially nice, namely to represent mean vectors through their one-dimensional principal components and covariances by plotting their distance to the covariance at a given location of interest. Figures 5 and 6 suggest that there exist sharp changes in the covariance across certain regions in Great Britain, and an alternative approach to the one proposed in TPAC that may allow to segment the country into dialect

zones might be to fit a spatial step function. Such an approach could lead to a partition of the spatial domain into connected sets, on each of which mean and covariance would be assumed to be constant, with sharp changes across the partition sets. The problem then is to determine the boundaries and the values for each partition. While there exists a body of previous work on this, it is focused on the narrower problem of change boundary estimation for real-valued data (Müller and Song 1994; Mammen and Tsybakov 1995) and has not been developed yet for more general random objects. Assuming such a partition is available, a spatial step function for the linguisitic covariances may provide an added level of interpretability to the visualizations of TPAC, even if the true spatial covariance is not believed to be exactly constant on the partition sets. In this case, each partitioned region A is associated with an average covariance, quantified by the *Fréchet integral* (Petersen and Müller 2016)

$$\Sigma_A = \underset{\Lambda}{\operatorname{argmin}} \int_A d_S^2(\Sigma(x), \Lambda) \mathrm{d}x,$$

as Λ ranges over the space of covariance matrices of the appropriate size. The estimated covariances $\hat{\Sigma}(x)$ can be similarly summarized to obtain $\hat{\Sigma}_A$.

In conclusion, the article by TPAC makes an important contribution to the rapidly evolving methodology for random objects, is a model for interdisciplinary research, and exposes intriguing questions for future research.

REFERENCES

Bhattacharya, R. and Patrangenaru, V. (2003), "Large sample theory of intrinsic and extrinsic sample means on manifolds - I," *Annals of Statistics*, 31, 1–29.

Dubey, P. and Müller, H.-G. (2019), "Fréchet Analysis of Variance for Random Objects," *Biometrika, to appear – arXiv preprint arXiv:1710.02761.*

Fréchet, M. (1948), "Les éléments aléatoires de nature quelconque dans un espace dis-

tancié," *Annales de l'Institut Henri Poincaré*, 10, 215–310.

Grenander, U. (1950), "Stochastic processes and statistical inference," *Arkiv för Matematik*, 1, 195–277.

Kleffe, J. (1973), "Principal components of random variables with values in a separable Hilbert space," *Statistics: A Journal of Theoretical and Applied Statistics*, 4, 391–406.

Lyons, R. (2013), "Distance covariance in metric spaces," *Annals of Probability*, 41, 3284–3305.

Mammen, E. and Tsybakov, A. B. (1995), "Asymptotical minimax recovery of sets with smooth boundaries," *The Annals of Statistics*, 502–524.

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

Müller, H.-G. (2016), "Peter Hall, Functional Data Analysis and Random Objects," *Annals of Statistics*, 44, 1867–1887.

Müller, H.-G. and Song, K. S. (1994), "Maximin estimation of multivariate boundaries," *Journal of Multivariate Analysis*, 50, 265–281.

Müller, H.-G. and Stadtmüller, U. (1999), "Multivariate Boundary Kernels and a Continuous Least Squares Principle," *Journal of the Royal Statistical Society: Series B*, 61, 439–458.

Petersen, A., Deoni, S., and Müller, H.-G. (2019), "Fréchet Estimation of Time-Varying Covariance Matrices From Sparse Data, With Application to the Regional Co-Evolution of Myelination in the Developing Brain," *arXiv preprint arXiv:1806.09690 Annals of Applied Statistics, in press.*

Petersen, A. and Müller, H.-G. (2016), "Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data," *Biometrika*, 103, 103–120.

Petersen, A. and Müller, H.-G. (2019), "Fréchet regression for random objects with Euclidean predictors," *Annals of Statistics*, 47, 691–719.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer Series in Statistics, New York: Springer, 2nd ed.

Rao, C. R. (1958), "Some Statistical Methods for Comparison of Growth Curves," *Biometrics*, 14, 1–17.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and testing dependence by correlation of distances," *Annals of Statistics*, 35, 2769–2794.

Tavakoli, S., Pigoli, D., Aston, J. A., and Coleman, J. (2019), "A Spatial Modeling Approach for Linguistic Object Data: Analysing dialect sound variations across Great Britain," *Journal of the American Statistical Association (to appear)*.

Wang, H., Marron, J., et al. (2007), "Object oriented data analysis: Sets of trees," *Annals of Statistics*, 35, 1849–1873.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016), "Functional Data Analysis," *Annual Review of Statistics and its Application*, 3, 257–295.

Yuan, Y., Zhu, H., Lin, W., and Marron, J. (2012), "Local polynomial regression for symmetric positive definite matrices," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 697–719.