A commentary on Noyce, Küssner and Sollich: Quantifying Shapes

CHRISTIAN HENNIG
Department of Statistical Science, University College London

ABSTRACT: This review focuses on the statistical aspects of "Quantifying Shapes". A number of decisions made by the authors are discussed regarding the choice of methods and how exactly they are applied and their results interpreted. Most space is devoted to issues of cluster analysis and low dimensional embedding, with further remarks concerning GP regression and classification.

Submitted 2013 May 15; accepted 2013 June 14.

KEYWORDS: *Cluster analysis, regression, spectral clustering, principal components*

THE authors of "Quantifying Shapes: Mathematical Techniques for Analysing Visual Representations of Sound and Music" use Gaussian process (GP) regression, principal component analysis, cluster analysis and GP classifiers to analyze a data set consisting of music visualizations of 20 stimuli by 71 participants. Being a statistician with somewhat limited knowledge in music perception research, I will focus on statistical aspects of the paper.

Generally I agree with the authors that the techniques they have used can be helpful for the analysis of this kind of data and may reveal interesting structures, and I find the selection of methods reasonable (although of course there are promising alternatives in the literature). On the other hand, more attention should be paid to a number of details. Many of the authors' results depend, sometimes critically, on tuning decisions and choices to be made by the researchers, and the importance of these choices and their implications need to be acknowledged.

The first thing that struck me was that data stem from 20 stimuli, but stimuli are not distinguished in the authors' analyses. The authors give a reason for this, but it still does not go without saying that what participants do does not depend on the stimulus. This would have been worth exploring (I would have liked to see a demonstration of how plots such as in Figs. 1-3 relate to corresponding plots for single stimuli).

REGRESSION

Regarding the GP regression, I wonder why the authors confine themselves to the estimation of the covariance function k. The mean function m models the direct connection between the inputs and the output. Assuming this to be constant zero implies that any connection between input and output is unsystematic, determined by random noise and correlation with other data points only. To me it seems that the mean function would have been at least as interesting as the covariance function. The authors defend their decision by noting that Figures 1-3 actually show that the method produces fits that systematically connect the output to the input. But the fits depend heavily on the given data, and assuming the mean function to be zero means that what we see is explained, apart from the observed data points, by autocorrelation only, but not by systematic dependence of y on x (obviously estimating m(x) would imply even less prior knowledge than assuming it to be constant zero). The authors are right, though, that this approach is standard in the literature on GP regression, probably for identifiability reasons, and other methods of nonlinear regression would have been needed for modeling a nonconstant mean function.

CLUSTER ANALYSIS

The authors continue with a section on cluster analysis, on which there is much to say (cluster analysis is among my core areas of expertise). To begin with, the authors decided to use for their analyses the parameters estimated by the GP regression. This is a reasonable option, but there are many alternatives. The important issue to address here is whether similarity in the space of regression parameters is an appropriate representation of how subject-matter experts would assess the music visualisations of two

participants to be similar. I wonder whether this could be achieved in a clearer way by computing dissimilarity measures from the time series data produced by the participants directly (e.g., Douzal-Chouakria & Nagabushan, 2007), and then performing dissimilarity-based cluster analysis or embedding by multidimensional scaling.

If GP regression parameters are used, the similarity implied by this, as well as the results of most methods used by the authors, depend potentially quite heavily on whether and how the different parameters are standardized. There are various possible methods of standardization which have different implications (see, for example, Hennig and Liao (2013), who have many general thoughts on the choice of an appropriate clustering method). Figures 4 and 5 indicate that sizes and variations of the different parameters are not properly comparable without standardization and/or transformation, so it seems reasonable to standardize to unit variance, as the authors did for the PCA. The same argument would have applied to spectral clustering, but the authors did not apply it there, not because they had positive arguments for non-standardization, but rather, apparently, because they saw leaving the data alone as some kind of "safer option" given the higher complexity of the method.

I agree that data visualization is a valuable exploratory tool for clustering. However, the authors to some extent blur the distinction between clustering and lower-dimensional embedding by treating Principal Components Analysis (PCA)—which is a dimension reduction technique—together with genuine clustering techniques in the "clustering" section, and by mainly focusing on the resulting lower-dimensional embedding when discussing spectral clustering (SC). Although it occasionally works well for highlighting clusters in practice, there is no guarantee at all that clusters, if they exist, can be nicely seen in the first two PCA dimensions. There are better projection techniques, more specifically designed to find clustering (Tyler, Critchley, Duembgen, & Oja, 2009) and/or more robust against outliers (Hubert, Rousseeuw, & Vanden Branden, 2005).

The problem with the SC embedding is rather the opposite of the problem with PCA: the embedding is justified in the literature as a basis for clustering (though there is no reason to expect that two dimensions are enough; generally the authors seem too keen on throwing away clustering information that cannot be shown in a nice 2D picture), but its justification as a proper and intuitive visual representation of the data is much weaker (see von Luxburg's (2007) tutorial as cited by the authors). The SC embedding transforms distances effectively in such a way that points that are deemed candidates for being in the same cluster move closer together, producing a view that can look deceptively clustered compared to the similarity structure in the original data space. In particular, SC does not deliver a simple linear projection as the authors claim. So it may be that there is indeed little structure in the data, and PCA tells us about this more honestly than the SC embedding.

The use of the SC embedding as input for GMM seems quite inappropriate. Standard SC in fact uses the embedding as clustering input, but varying within-cluster variance/covariance structures are a major characteristic of GMM and there is no justification to expect such features in the SC embedding; it would have been more promising to use SC in the original form with k-means clustering and more than two dimensions; the literature is full of criteria that can be used to assess the number of clusters, so that this problem alone is not a good reason to use a variational Bayesian technique. On the other hand, GMM could have been applied to well-interpretable subsets of the GP parameters, a more than 2D solution of PCA or other representations of the data that preserve the original geometry better than the SC embedding (even if then clustering would have been less pronounced). Computing clusterings based on the SC embedding of estimated GP regression parameters means that two layers of complex processing are put between the raw data and the computation of the clustering, which makes the outcome difficult to interpret. Some processing may be needed, e.g., dimension reduction, but still the simpler and more intuitive the data processing, the clearer the meaning of the patterns that are finally found.

A number of further tuning decisions have been made. As explained by von Luxburg (2007), there are various options for running SC. GMM allows the specification of various models for covariance matrices (e.g., same or different within all clusters; spherical; diagonal). Bayesian prior parameters are required. The authors do not discuss these choices (though at least they list most of them in the appendix), nor the sensitivity of their conclusions against different choices. As admitted by the authors in their conclusion, some of these choices may have a large impact.

CLASSIFICATION

Regarding classification, I find part [8] of the appendix much less detailed and clear than the other parts, so I have difficulty in understanding what exactly was done and whether this is appropriate. As the authors acknowledge, there would have been straightforward alternatives, for example applying standard

classification techniques such as Linear Discriminant Analysis or Support Vector Machines to the vector of estimated GP regression parameters.

In the beginning of the section, I had expected that classification is applied not so much because it is of direct interest, but rather in order to make statements about how strongly and in which way the groups differ. While I cannot imagine a real situation in which it would be of practical importance to tell apart trained and untrained participants based on their music visualizations, I can well see the use of characterizing the difference between these groups, so I was rather surprised that later the authors comment on the success regarding the former aim, which seems quite irrelevant to me. Regarding characterizing the difference between the groups, classification-adapted projection techniques (Hennig, 2004) could be of interest, but even before that one may look at parameter-wise differences between groups and one-way Multivariate Analysis of Variance.

CONCLUSION

Despite criticizing some of the authors' decisions, I find this a very stimulating paper, applying methods the use of which is worth exploring. Regarding the results, the GP regression seems most convincing, whereas I am not sure that what was learnt from cluster analysis and classification about the data is of much interest. I think that a better job could have been done in this respect. Particularly, I do not think that the authors have demonstrated that PCA is unsuitable and SC is much better. Better results could probably have been obtained by both methods looking at more than two dimensions, also keeping in mind that structure other than clear clustering may also be of interest, and that the SC embedding is specifically designed for clustering, as opposed to visually exploring the structure of the data space.

REFERENCES

Douzal-Chouakria, A., & Nagabhushan, P.N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification,* Vol. 1, No. 1, pp. 5-21.

Hennig, C. (2004) Asymmetric linear dimension reduction for classification. *Journal of Computational and Graphical Statistics,* Vol. 13, No. 4, pp. 930-945.

Hennig, C., & Liao, T.F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 62, No. 3, pp. 309-369.

Hubert, M., Rousseeuw, P., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, Vol. 47, No. 1, pp. 64-79.

Tyler, D.E., Critchley, F., Duembgen, L., & Oja, H. (2009). Invariant coordinate selection (with discussion). *Journal of the Royal Statistical Society, Series B*, Vol. 71, No. 3, pp. 549-592.