direct path that bridges model assumptions to validation. Davies (1995, 2008) developed an interesting 'data approximation' approach that can be beneficial to cluster analysis. The true model, as understood in the frequentist set-up, completely disappears, and data are treated as deterministic. The statistician looks for good approximating models rather than true models. The key idea is that a distribution F is a good approximation for a set of observations $X_n := \{x_1, x_2, \ldots, x_n\}$ if F can generate samples of size n that look like X_n. In some sense the validation becomes an integral part of the model-building strategy. Here validation is intended as a step to check adequacy of the approximating model (this is different from Section 7.6). Although I hardly believe that such a concept of approximation can support both dissimilarity-based methods and model-based methods at the same time, I believe that such a concept is somehow consistent with the idea that is proposed in this paper.

Another important topic touched on by the paper is the choice of the number of clusters. The problem is old, unsolved, difficult and fascinating. In model-based clustering the choice of the number of clusters is often presented as an estimation problem because it is treated as the problem of recovering the number of mixture components. These methods are often trusted because of consistency theorems. However, under regularity assumptions plus model truth, consistency holds for the number of mixture components which do not overlap with clusters necessarily. In contrast, as highlighted in the paper, dissimilarity-based methods embody a concept of 'good number of clusters' directly, and without claiming a status of rigorous objectivity. In my view the choice of the number of clusters is not an estimation problem, but rather a choice of complexity. Within the 'data approximation' approach discussed previously, a good number of clusters is such that the corresponding approximating model is just sufficiently complex to be able to reproduce observed data.

Catherine M. Crespi and Weng Kee Wong (*University of California, Los Angeles*)

Hennig and Liao provide an excellent and thoughtful demonstration of the application of cluster analysis to socio-economic stratification. Much of their paper is concerned with incorporating nominal and ordinal variables in a cluster analysis in a manner that achieves a clustering that satisfies the researcher as grouping together observations that one would judge to 'be similar' or 'to belong in the same class' on the basis of subject matter knowledge. Essentially we are looking for a transformation from a nominal or ordinal scale to an interval or ratio scale that allows the variable to be used in the computation of dissimilarity metrics such that the 'good grouping' objective is achieved.

Nominal and ordinal variables are also an issue in principal components analysis (PCA) and, in response, a collection of methods termed 'non-linear' or 'categorical' PCA have been developed; for an introduction, see Linting *et al.* (2007); for a historical overview, see Gifi (1990). In non-linear PCA, the categories of such variables are assigned numeric values through a process of optimal scaling. The objective of optimal scaling is to optimize the properties of the correlation matrix of the quantified variables. The method maximizes the first p eigenvalues of the correlation matrix of the quantified variables, where p is the number of components that are chosen in the analysis. This maximizes the variance that is accounted for in the quantified variables. Optimal scaling and PCA model estimation are performed simultaneously.

Do the authors see utility in applying the approach of optimal scaling to cluster analysis? Could a paradigm of simultaneous estimation of optimal scaling of variables and clustering of observations be considered for cluster analysis?

Avi Feller and Edoardo M. Airoldi (*Harvard University, Cambridge*)

We commend Dr Hennig and Professor Liao on their thoughtful exposition of clustering with mixed-type variables. The methodology proposed, however, goes only part way towards addressing the issues of how to connect analysis to substantive claims.

From a substantive perspective, characterizing socio-economic stratification in terms of discrete variation, although appealing, can ultimately prove artificial. For example, why is a service worker with a high school degree earning \$50 000 per year in cluster 2 considered 'lower class' whereas a service worker with a high school degree earning \$50 000 per year in cluster 3 is considered 'middle class'? From a statistical perspective, modelling the distribution of the covariates conditionally on the cluster labels by using a mixture model creates a difficult inference problem. Tackling these hard problems can be justified in applications where labels arise naturally, such as in modelling text topics (Bischof and Airoldi, 2012) or unobserved disease states (Pepe and Janes, 2007). Nonetheless, this approach runs the risk of identifying discrete clusters in the data even in the absence of clear clustering patterns—which the authors aim to interpret as discrete socio-economic strata in their application. Additional concerns arise because the clustering method proposed is sensitive to data choices about preprocessing and variable definitions. For

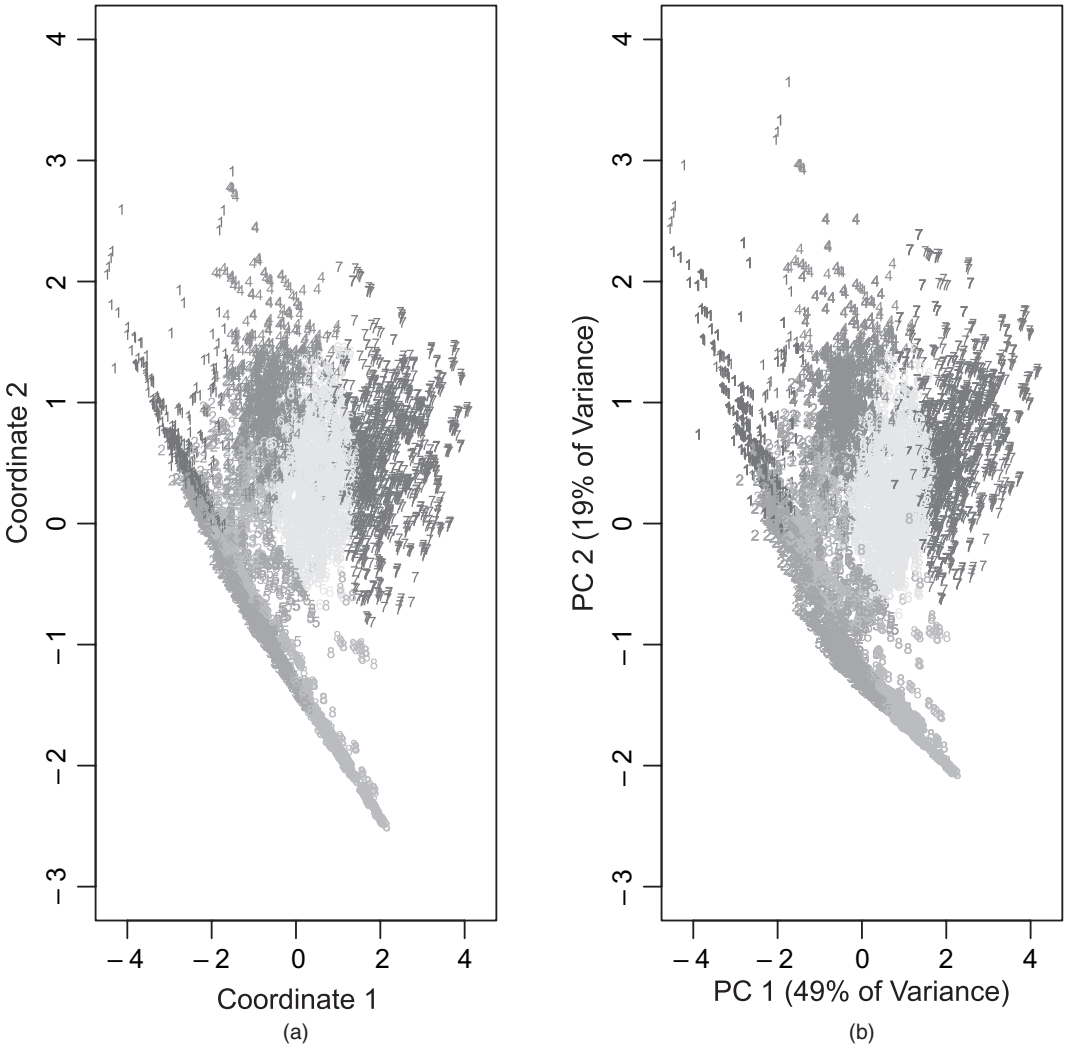

Fig. 11. (a) Multi-dimensional scaling and (b) principal components analysis

example, we recreated the authors' k-medoids procedure exactly but redefined *education* as an ordinal variable (treating 'less than high school' as a single category) rather than a continuous variable. This is arguably more appealing. However, on the basis of only this minor change, roughly a fifth of the individuals were assigned to different clusters.

To tackle some of these concerns, we would suggest a simpler approach that jointly models socio-economic status and covariates in terms of continuous—rather than discrete—variation (e.g. Rubin and Thayer (1982)). This avoids the problem of imposing social structure that may not exist. If, on the basis of such an analysis, the inferred socio-economic variables show some level of segregation, we would be comfortable separating individuals into discrete strata. To illustrate this line of reasoning, we performed simple principal components analysis and multi-dimensional scaling on the dissimilarity matrix that was used in the paper and projected the authors' preferred clusters onto these latent spaces. As shown in Fig. 11, there is no clear socio-economic stratification by using either multi-dimensional scaling (Fig. 11(a)) or principal components analysis (Fig. 11(b)). Rather, the socio-economic clusters proposed are poor descriptors of the overall variation in socio-economic status in this context. Of course, this analysis has its own shortcomings, but it clearly demonstrates how assuming discrete variation *a priori* can lead to ques-

tionable results. This continuous approach is also consistent with other research on principal components analysis with mixed-type data in socio-economic applications, such as Kolenikov and Angeles (2009).

Luis A. García-Escudero, Alfonso Gordaliza and Agustín Mayo-Iscar (*Universidad de Valladolid*)
We congratulate Hennig and Liao on an important, stimulating and practical paper. First, we stress that we completely agree about the lack of a unique objective 'truth' in clustering. Moreover, we appreciate their emphasis on explaining the importance of translating the interpretative meaning of the data and the specific aim of the clustering into properties of the methodology to apply. In fact, clustering methods should not be seen as 'black boxes' where the researcher does not play any active role.

The authors also illustrate why the equivariance paradigm, which is very important in many applications, leads to the failure of cluster methods in particular settings. Therefore, it would be interesting if the researcher could control the allowed discrepancies with respect to the Euclidean way of looking at data. This can be done by posing constraints on the group scatter matrices as in TCLUST (García-Escudero *et al.*, 1998). The stronger these constraints, the closer we are to applying Euclidean distances (and the further away from affine equivariance). The strength of the constraints must be decided by taking into account the aim of clustering. Constraints also avoid detecting spurious clusters like those reported for the latent class clustering method. TCLUST also allows fitting clusters with different weights.

It is also shown how a proper preprocessing stage of the data (again by taking into account the interpretative meaning) allows a better representation of the data in which (Euclidean) distances between points approximately reflect the dissimilarity between individuals. Then, the authors propose the use of the k-medoids method. After this transformation, we believe that any sensible clustering method would provide statistically meaningful and comparable results. For instance, application of the 8-means method gives clusters with an 80% adjusted Rand index with respect to the `clara` partition. When choosing a trimming level between 2% and 5% the trimmed 8-means method (which is available in the add-on package `tclust`) yields 77% and 80% adjusted Rand index values.

Finally, the authors comment on the potential interest of developing clustering methods for mixed-type data with heavy tails. It is important to detect anomalous observations in socio-economic stratification not only to control for their undesired effects on the clustering results, but also to explore the significance of the anomalies detected themselves. The use of trimming approaches, starting from the transformed data or considering trimmed versions of likelihoods based on expression (3.1), is surely worthwhile in this case.

Andreas Geyer-Schulz (*Karlsruhe Institute of Technology*)
I consider this paper to be a beautiful example of modelling with tender loving care. However, with regard to the overall clustering philosophy and the increasing availability of data on the Internet, I would like to know how far (and with which constraints) the modelling process can be automated or at least made adaptive.

Since the number of clusters was identified by a parametric bootstrap test for clustering which depends on the null model specified, my questions are

(a) what the assumptions underlying the null model are,
(b) how much choice we have in designing the null model and, last but not least,
(c) what the authors would consider as a natural null model.

Gérard Govaert (*Université de Technologie de Compiègne*)
Hennig and Liao must be congratulated for this interesting and valuable paper. I agree that the application of automatic methods is disappointing and that many choices must be made for clustering but their comparison between the latent class and dissimilarity-based approaches may be questionable.

On one hand, dissimilarity-based approaches require the choices of a dissimilarity and a numerical criterion. Therefore, these choices require some expertise from the user. On the other hand, the latent class approaches proposed use much less expertise. As a matter of fact, it is also possible to integrate the whole expertise of the user in these approaches. And, to obtain a balanced comparison, it would be preferable to select the mixture model taking into account this expertise. For example, the choice of the distance could be related to the choice of the mixture distributions: spherical Gaussian distributions are related to the standard Euclidean distance, general Gaussian distributions are related to the Mahalanobis or Batthacharya distances and Laplace distributions are related to Manhattan L_1-distances. Then, standardization and weighting developed in this paper are not only relevant to Euclidean distance but also to the spherical Gaussian mixture.