Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2013)
62, *Part* 3, *pp.* 309–369

How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification

Christian Hennig

University College London, UK

and Tim F. Liao

University of Illinois, Urbana—Champaign, USA

[*Read before* The Royal Statistical Society *on Wednesday, November 14th, 2012, the President*, Professor V. S. Isham, *in the Chair*]

Summary. Data with mixed-type (metric–ordinal–nominal) variables are typical for social stratification, i.e. partitioning a population into social classes. Approaches to cluster such data are compared, namely a latent class mixture model assuming local independence and dissimilarity-based methods such as *k*-medoids. The design of an appropriate dissimilarity measure and the estimation of the number of clusters are discussed as well, comparing the Bayesian information criterion with dissimilarity-based criteria. The comparison is based on a philosophy of cluster analysis that connects the problem of a choice of a suitable clustering method closely to the application by considering direct interpretations of the implications of the methodology. The application of this philosophy to economic data from the 2007 US Survey of Consumer Finances demonstrates techniques and decisions required to obtain an interpretable clustering. The clustering is shown to be significantly more structured than a suitable null model. One result is that the data-based strata are not as strongly connected to occupation categories as is often assumed in the literature.

Keywords: Average silhouette width; Cluster philosophy; Dissimilarity measure; Interpretation of clustering; *k*-medoids clustering; Latent class clustering; Mixture model; Number of clusters; Social stratification

1. Introduction

There are various approaches for cluster analysis in the literature and a lack of clear guidance about how to choose an appropriate one for a given problem. In this paper, we explore the use of formal clustering methods for socio-economic stratification based on mixed-type data with continuous, ordinal and nominal variables.

This is guided by a general 'philosophy of clustering', which involves considerations of how to define the clustering problem of interest, how to understand and 'tune' the various available clustering approaches and how to choose between them. All this should be driven by the way that the subject matter researchers connect the aim of clustering and their interpretation of concepts like 'similarity' and 'belonging together in the same class' to the choice and formal handling of the indicators involved. The main contribution of this paper is to show in detail how this can be done and what it entails, exemplary for the socio-economic stratification application

Address for correspondence: Christian Hennig, Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK.
E-mail: c.hennig@ucl.ac.uk

 0035–9254/13/62309

but including aspects, particularly in Section 6, that are relevant for other clustering applications as well and hardly mentioned in the literature. The present approach brings statistical and sociological (or general subject matter) knowledge closer together by focusing on the interface, the 'translation task' between statistical methodology and sociological background.

Two quite different approaches are compared, namely a model-based clustering approach (Vermunt and Magidson, 2002), in which different clusters are modelled by underlying latent classes or mixture components, and a dissimilarity-based partitioning approach that is not based on probability models (k-medoids; Kaufman and Rousseeuw (1990)) with some methods to estimate the number of clusters. Such data typically arise in social stratification and generally in social science. Social stratification is about partitioning a population into several different social classes. Although the concept of social class is central to social science research, there is no agreed definition of a social class. It is of interest here whether social stratification based on formal clustering can contribute to the controversy about social class. In this paper we analyse data from the US Survey of Consumer Finances (SCF), for which the more appropriate term is 'socio-economic stratification'. Apart from computing and interpreting a clustering, we address whether the clustering captures significant structure beyond decomposing dependence structures between the indicators, which indicators are most influential for stratification and whether strata derived from the data are related to occupation categories.

Section 2 introduces the problem of social stratification including the data set and discusses the role of cluster analysis. In Section 3 latent class clustering (LCC) is introduced. Section 4 introduces k-medoids along with some indices to estimate the number of clusters. Section 5 discusses the philosophy underlying the choice of suitable cluster analysis methodology. In Section 6, this philosophy is applied to the socio-economic stratification problem by discussing in detail the decisions that researchers have to make to arrive at a clustering regarding transformation of variables, the definition of a dissimilarity measure, choice of a clustering method and finding the best number of clusters. In Section 7, the data set is analysed. Section 8 gives a concluding discussion.

The data that are analysed in the paper can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Social and socio-economic stratification

2.1. Background

The concept of social class is central to social science research, either as a subject in itself or as an explanatory basis for social, behavioural and health outcomes. The study of social class has a long history, from the social investigation by the classical social thinker Marx to today's on-going academic interest in issues of social class and stratification for both research and teaching (e.g. Grusky *et al.* (2008)). Researchers in various social sciences use social class and social stratification as explanatory variables to study a wide range of outcomes from health and mortality (Pekkanen *et al.*, 1995) to cultural consumption (Chan and Goldthorpe, 2007).

When social scientists employ social class or stratification as an explanatory variable, they follow either or both of two common practices, namely using one or more indicators of social stratification such as education and income and using some version of occupation class, which is often aggregated or grouped into a small number of occupation categories. For example, Pekkanen *et al.* (1995) compared health outcomes and mortality between white-collar and blue-collar workers; Chan and Goldthorpe (2007) analysed the effects of social stratification on cultural consumption with a variety of variables representing stratification, including education, income,

occupation classification and social status (operationalized by them). The reason why researchers routinely use some indicators of social class is that there is no agreed definition of social class, let alone a specific agreed operationalization of it.

Various concepts of social class are present in the sociological literature, including a 'classless' society (e.g. Kingston (2000)), society with a gradational structure (e.g. Lenski (1954)) and a society in which discrete classes (interpreted in various, but usually not data-based, ways) are an unmistakable social reality (e.g. Wright (1997)).

The question to be addressed by cluster analysis is to understand 'social stratification' through data, i.e. to explore what kind of unsupervised classification(s) based on the social (or socio-economic) indicators they yield. We acknowledge that data alone cannot decide the issue objectively. Some questions to address are whether the data are meaningfully clustered at all, which indicators contribute most to a clustering and to what extent clusters are aligned with features that have been connected to social stratification in the literature, here particularly occupation. Clusters may also serve as efficient reduction of the information in the data and as a tool to decompose and interpret inequality and changes over time. A latent class model was proposed for this by Grusky and Weeden (2008). A similar finite mixture model was proposed and applied to income inequality data by Liao (2006).

2.2. Indicators of socio-economic strata

A crucial question for data-based social stratification is the choice of indicators on which the classification is based, although this is restricted by what is available for observation. For the data set that is analysed here, the focus is on socio-economic class. Individuals should be allocated to classes along the dimensions of the individual's socio-economic wellbeing. There is a general consensus on how this is measured. People's current socio-economic statuses and future outlooks depend on the levels of their income and wealth (Levy and Michel, 1991). There has been a long tradition of using income to measure socio-economic wellbeing, and an influential approach to operationalize economic welfare that was proposed by Weisbrod and Hansen (1968) is a measure based on a combination of current income and current net worth (i.e. assets minus liabilities). This approach has the obvious advantage of treating someone with a great value of real estate but many liabilities as not necessarily having a high level of economic welfare.

In addition to income, occupation and education have been two important dimensions of stratification since at least Hollingshead's (1957) research. Although occupational categories may change and be updated, occupation is central to studying stratification. Blau and Duncan's (1967) landmark study of the American occupational structure started a modern tradition that was virtually unchallenged until Spilerman's (2000) criticism of the preoccupation with labour market rewards including occupation and income and with the consideration of family status as derivative of the head's occupational status, at the disregard of wealth. We intend to employ a balanced set of variables measuring rewards (occupation and income), achievement (education) and wealth (savings and housing). One of the issues of interest regarding the data set analysed is to what extent the occupation classes given there agree with classes that could be obtained from the other variables.

Although it is a standard practice in the social, economic and health sciences to represent one's socio-economic status by a minimum set of the three variables education, income and occupation, we extend this approach to including more socio-economic measures tapping one's wealth. Without available data on all the necessary components such as the values of owned properties and of current mortgages, we turn to personal savings as an indicator of net worth. An individual's savings in various accounts represent a significant part of one's net worth (see

Poterba *et al.* (1994) who studied retirement savings and net worth of American household heads aged 55–69 years). As reported by Bernheim *et al.* (2001), the correlation between self-reported saving and net worth is highly statistically significant.

Another dimension of socio-economic welfare that we examine is the diversification of the place where one keeps one's disposable income and savings. The number of checking (i.e. 'current') and savings accounts, in addition to income, is what Srivastava *et al.* (1984) examined to ascertain that their respondents did not differ from non-respondents economically. Another reason to consider the number of checking and savings accounts is also that of diversification especially in the age of globalization and internationalization when people with greater socio-economic wellbeing tend to have multiple accounts in banks in more than just their home cities, states or countries.

There are other forms of assets that an individual may own. Home ownership has often been included as an indicator of economic class, in addition to monetary items such as income and wealth (e.g. von dem Knesebeck (2002)). Life insurance is also often equated with a form of financial asset (Brennan and Schwartz, 1976). Although the majority of people earn income and have checking accounts, many may not own a home or life insurance, further differentiating socio-economic classes. We shall, however, weight some of these variables differently; see Section 6.2.

2.3. US consumer finances data

The data set that is analysed in this paper stems from the 2007 US SCF. The survey was sponsored by the Board of Governors of the Federal Reserve System in co-operation with the Statistics of Income Division of the Income Revenue Service in the USA, with the data collected by the National Opinion Research Center at the University of Chicago. The SCF employs a national area probability sample that selects households with equal probability through a multistage selection procedure and provides good coverage of widely spread characteristics. Although the area probability sampling procedure is efficient for generating a nationally representative sample, it has the two shortcomings of a potentially insufficient number of observations for certain types of financial behaviour and of non-random non-response. To deal with these potential problems, the SCF also employs a list sample that is developed from statistical records derived from tax returns, thereby guaranteeing the national representativeness of the survey. See Kennickell (2000) for further details about the methodology of the survey. The original data set available to us has 20 090 observations (individuals) and does not contain missing values. In this paper, we selected the 17 430 male observations for whom education and occupation data were available. There are obvious differences between males and females and to carry out gender comparisons, which are not the topic of this paper, it is reasonable to analyse males and females data separately.

As motivated above, we used the following eight variables (the chosen transformations are discussed in Section 6.1):

(a) *lsam*, $\log(x + 50)$ of total amount of savings as of the average of the previous month (treated as continuous),
(b) *linc*, $\log(x + 50)$ of total income of 2006 (treated as continuous),
(c) *educ*, years of education between 0 and 17 years (this is treated as ordinal (level 17 means 'graduate school and above')),
(d) *cacc*, the number of checking accounts that one has (this is ordinal with six levels (corresponding to no, 1, 2, 3, 4 or 5, or 6 or more accounts),
(e) *sacc*, the number of savings accounts, coded as above,

(f) *hous*, housing, nominal with nine levels ('neither owns nor rents', 'inapplicable', 'owns or is buying/land contract', 'pays rent', 'condo', 'co-op', 'town-house association', 'retirement lifetime tenancy' and 'own only part'),
(g) *life*, whether or not one has life insurance (binary), and
(h) *occ*, occupation class, nominal with seven levels (from 0 to 6) ('not working for pay', 'managerials and professionals', 'white-collar workers and technicians', 'lower level managerials and various professions', 'service workers and operators', 'manual workers and operators', 'farm and animal workers'). The classification of occupations follows the US Census Bureau 2006 four-digit occupation code. The detailed occupational categories were collapsed into the seven larger groups, which is a common practice, for the public version of the 2007 SCF.

3. Latent class clustering

This paper deals with the cluster analysis of data with continuous, ordinal and nominal variables. Denote the data $\mathbf{w}_1, \ldots, \mathbf{w}_n$, $\mathbf{w}_i = (\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$, $\mathbf{x}_i \in \mathbb{R}^p$, $\mathbf{y}_i \in \mathcal{O}_1 \times \ldots \times \mathcal{O}_q$ and $\mathbf{z}_i \in \mathcal{U}_1 \times \ldots \times \mathcal{U}_r$, $i = 1, \ldots, n$, where $\mathcal{O}_j$, $j = 1, \ldots, q$, are ordered finite sets and $\mathcal{U}_j$, $j = 1, \ldots, r$, are unordered finite sets.

A standard method to cluster such data sets is LCC (Vermunt and Magidson, 2002), where $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are modelled as independent and identically distributed generated by a distribution with density

$$f(\mathbf{w}) = \sum_{h=1}^{k} \left\{ \pi_h \varphi_{\mathbf{a}_h, \Sigma_h}(\mathbf{x}) \prod_{j=1}^{q} \tau_{hj}(y_j) \prod_{j=1}^{r} \tau_{h(q+j)}(z_j) \right\}, \tag{3.1}$$

where $\mathbf{w} = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ is defined as $\mathbf{w}_i$ above without subscript i. Furthermore $\pi_h > 0 \ \forall h$, $\Sigma_{h=1}^{k} \pi_h = 1$, $\varphi_{\mathbf{a},\Sigma}$ denotes the p-dimensional Gaussian density with mean vector $\mathbf{a}$ and covariance matrix Σ (which may be restricted), and $\Sigma_{y \in \mathcal{O}_j} \tau_{hj}(y) = \Sigma_{z \in \mathcal{U}_j} \tau_{h(q+j)}(z) = 1, \pi_h \geqslant 0, \tau_{hj} \geqslant 0 \ \forall h, j$.

The main assumption in density (3.1), apart from Gaussianity for the continuous variables, is that the variables are 'locally' independent within a mixture component (except for possible non-zero covariances between continuous variables).

A way to use the ordinal information in the y-variables is to restrict, for $j = 1, \ldots, q$,

$$\tau_{hj}(y) = \exp(\eta_{hjy}) / \sum_{u \in \mathcal{O}_j} \exp(\eta_{hju}), \qquad \eta_{hjy} = \beta_{j\xi(y)} + \beta_{hj}\xi(y), \tag{3.2}$$

where $\xi(y)$ is a score for the ordinal values y, the choice of which is discussed in Section 6.2. This is based on the adjacent category logit model (Agresti, 2002) as used in Vermunt and Magidson (2005a).

The parameters of density (3.1) can be estimated by the method of maximum likelihood using the EM or more sophisticated algorithms and, given estimators of the parameters (denoted by circumflexes), points can be classified into clusters (usually identified with mixture components) by maximizing the estimated posterior probability that observation $\mathbf{w}_i$ had been generated by mixture component h under a two-step model for density (3.1) in which first a mixture component $\gamma_i \in \{1, \ldots, k\}$ is generated with $P(\gamma_i = h) = \pi_h$, and then $\mathbf{w}_i$ given $\gamma_i = h$ according to the density

$$f_h(\mathbf{w}) = \varphi_{\mathbf{a}_h, \Sigma_h}(\mathbf{x}) \prod_{j=1}^{q} \tau_{hj}(y_j) \prod_{j=1}^{r} \tau_{h(q+j)}(z_j).$$

The estimated mixture component for $\mathbf{w}_i$ then is

$$\hat{\gamma}_i = \arg\max_h \hat{\pi}_h \hat{f}_h(\mathbf{w}_i), \tag{3.3}$$

where $\hat{f}_h$ denotes the density with all parameter estimators plugged in. The number of mixture components k can be selected by the Bayesian information criterion (BIC).

4. Dissimilarity-based clustering

Given a dissimilarity measure between observations (see Section 6.2), a dissimilarity-based clustering method that is suitable as an alternative to LCC is k-medoids (Kaufman and Rousseeuw, 1990). This is implemented as function `pam` in the add-on package `cluster` for the software system R (R Development Core Team, 2011). `pam` is based on the full dissimilarity matrix and therefore requires much memory. Function `clara` in `cluster` is an approximate version for Euclidean distances that can be computed for much larger data sets. It performs `pam` on several data subsets, assigns the further observations to the closest cluster medoid and selects from these the solution that is best according to the k-medoids objective function, which, for a dissimilarity measure d, is

$$g(\mathbf{w}_1^*, \ldots, \mathbf{w}_k^*) = \sum_{i=1}^{n} \min_{j \in \{1,\ldots,k\}} d(\mathbf{w}_i, \mathbf{w}_j^*) \tag{4.1}$$

for k medoids $\mathbf{w}_1^*, \ldots, \mathbf{w}_k^*$ from $\mathbf{w}$. This is similar to the popular k-means clustering, where d is the squared Euclidean distance and the 'medoids' are cluster means.

There are several possibilities to select the number of clusters for k-medoids. Some of them are listed (though treated there in connection with k-means) in Sugar and James (2003); an older but more comprehensive simulation study is Milligan and Cooper (1985). In the present paper, three such criteria are considered.

(a) *Average silhouette width ASW* (Kaufman and Rousseeuw, 1990): for a partition of $\mathbf{w}$ into clusters $C_1, \ldots, C_k$ let

$$s(i,k) = \frac{b(i,k) - a(i,k)}{\max\{a(i,k), b(i,k)\}}$$

be the so-called 'silhouette width', where

$$a(i,k) = \frac{1}{|C_h| - 1} \sum_{\mathbf{w}_j \in C_h} d(\mathbf{w}_i, \mathbf{w}_j),$$

$$b(i,k) = \min_{C_l \not\ni \mathbf{w}_i} \frac{1}{|C_l|} \sum_{\mathbf{w}_j \in C_l} d(\mathbf{w}_i, \mathbf{w}_j)$$

for $\mathbf{w}_i \in C_h$. k_{ASW} maximizes $(1/n)\Sigma_{i=1}^{n} s(i,k)$.

(b) *Calinski and Harabasz (1974) index CH*: k_{CH} maximizes the CH-index

$$\frac{\mathbf{B}(k)(n-k)}{\mathbf{W}(k)(k-1)},$$

where

$$\mathbf{W}(k) = \sum_{h=1}^{k} \frac{1}{|C_h|} \sum_{\mathbf{w}_i, \mathbf{w}_j \in C_h} d(\mathbf{w}_i, \mathbf{w}_j)^2$$

and

$$\mathbf{B}(k) = \frac{1}{n} \sum_{i,j=1}^{n} d(\mathbf{w}_i, \mathbf{w}_j)^2 - \mathbf{W}(k).$$

Originally CH was defined with the Euclidean distance as d, which connects it to k-means clustering, but k_{CH} was quite successful in Milligan and Cooper (1985) with various clustering methods.

(c) *Pearson version of Hubert's* Γ *PH*: the PH-estimator k_Γ maximizes the Pearson correlation $\rho(\mathbf{d}, \mathbf{m})$ between the vector $\mathbf{d}$ of pairwise dissimilarities and the binary vector $\mathbf{m}$ that is 0 for every pair of observations in the same cluster and 1 for every pair of observations in different clusters. PH ('normalized Γ' in Halkidi *et al.*, 2001) is a simplified version of Hubert's Γ (Baker and Hubert, 1975). The latter incurs computational problems for large data sets.

5. Clustering philosophy

The main principle of the clustering philosophy that is discussed here is as follows. There are no unique objective 'true' or 'best' clusters in a data set. Clustering requires that the researchers define what kind of clusters they are looking for. This depends on the subject matter and the aim of clustering. Selecting a suitable clustering method requires matching the data analytic features of the resulting clustering with the context-dependent requirements that are desired by the researcher.

Much of the literature in which clustering methods are introduced is based on assumptions about what the 'true clusters' are, which are rarely explicitly discussed. Typical implicit assumptions are as follows.

Assumption 1. Given a data set, there is a 'natural human intuition' of what the true clusters are.

Assumption 2. The data can be assumed to be generated from a statistical model in which different clusters are modelled by different distributions from the same family, or in which clusters are associated with density modes.

These assumptions are unsatisfactory in most cases. Neither makes any reference to the meaning of the data and to the aim of clustering in the given situation. Reference to natural human intuition is not of explanatory value and, because it is typically Euclidean, it applies neither to categorical variables nor to alternative dissimilarities.

An obvious problem with the model-based approach is that statisticians usually do not believe models to be true. If clusters are identified with components of a mixture model, they are no longer well defined if the mixture model is not precisely true (many mixture models such as the Gaussian model can approximate any distribution arbitrarily well, so there is no well-defined 'closest mixture approximating the truth').

Even in situations where a mixture model is true and most people may have a 'natural intuition' about the true clusters, these are not necessarily the clusters that a researcher is interested in.

Fig. 1(a) shows a mixture of three Gaussian distributions with $p = 2$. In some applications, it makes sense to define clusters corresponding to the mixture components and most people's intuition. However, if the application is social stratification, and the variables are, for example, income and a status indicator, this is not appropriate, because it would mean that the same social stratum (interpreted to correspond to mixture component 1) would contain the poorest people with lowest status as well as the richest people with the highest status. These observations are clustered together because flexible covariance matrices allow clusters to be defined by a certain dependence pattern between variables. The most similar observations are not necessarily brought together, which sometimes may be inappropriate even if the model is true.

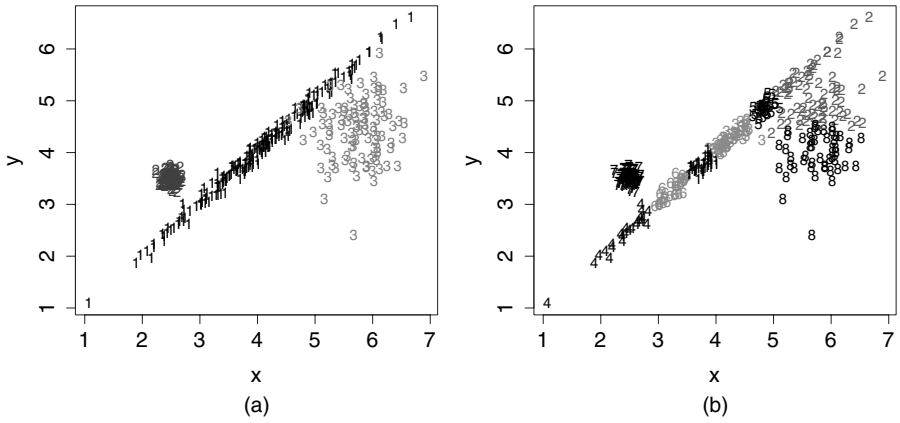

Fig. 1. Artificial data set from (a) a three-components Gaussian mixture with (b) optimal clustering from a Gaussian mixture model according to the BIC with within-component covariance matrices restricted to be diagonal

The clustering in Fig. 1(b) has been obtained by fitting a Gaussian mixture model in which all within-component covariance matrices are assumed to be diagonal, i.e. the (continuous) variables are assumed to be locally independent. These clusters are more sensible at least for social stratification even though we know that this model is wrong for the simulated data set. Assuming local independence in this sense decomposes dependence between variables into clusters. Furthermore, for most uses of social stratification (e.g. use as an explanatory variable in further studies, or informing politics) it is necessary to have descriptions of the strata in terms of the defining indicators, and an attractive feature of local independence in LCC is that the resulting strata can be interpreted in terms of the marginal distributions of the individual variables.

The role of the model assumption in cluster analysis is not the connection of the model to the truth (apart from rare applications in which the models can be directly justified), but about formalizing and exploring what kind of cluster shape the method implies. Model assumptions are helpful for making this explicit.

Supposedly 'model-free' approaches to clustering imply certain cluster shapes as well. In most applications, both model-based and supposedly model-free methods can be applied, so their characteristics need to be understood and compared with the requirements of the application.

It could be objected that clustering is often done to find latent subpopulations (Everitt *et al.*, 2011). A reviewer suspected that our philosophy applies to segmentation rather than classical cluster analysis in the sense of finding latent subpopulations, and that social stratification is a segmentation task. We do not think that this distinction is clear cut. Formally, the output of every (non-overlapping) clustering method is a segmentation. In social stratification as well as in many other applications researchers are interested in finding meaningful latent subpopulations if they exist as parts of the segmentation, but they are still interested in the segmentation if, as happens often in reality, not all the segments can be given a 'latent subpopulation' meaning. Furthermore, in unsupervised classification, knowledge about true subpopulations (such as a straightforward model for them) is not available, and using a method that segments the data in such a way that the resulting segments correspond to the researcher's concept of 'belonging together' seems to be a promising approach to find such populations.

The crucial task to choose a suitable clustering methodology according to the philosophy that is presented here is the translation of the desired 'interpretative meaning' of the clustering into data analytic characteristics of the clustering method. This encompasses various kinds of

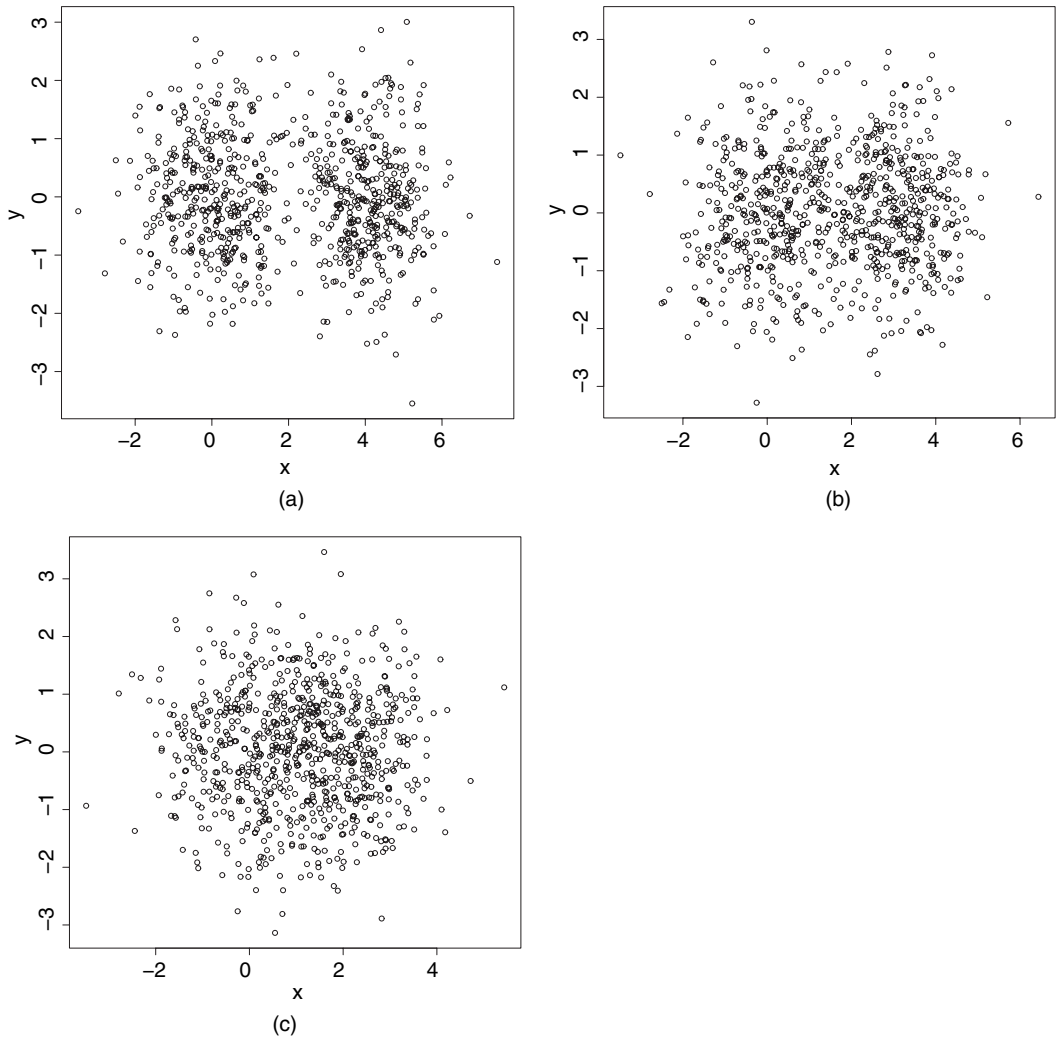

Fig. 2. Two groups of points with variance 1 along the *x*-axis and mean difference of (a) 4, (b) 3 and (c) 2

decisions that cannot be made 'objectively' from the data alone, including the choice of a clustering method, variable transformations and definition of the required concept of 'dissimilarity'.

Fig. 2 illustrates the need for such decisions regarding finding the number of clusters. It shows data from a mixture of two Gaussian distributions, the means of which have decreasing distance. There is no objective borderline difference between means distinguishing whether the underlying mixture should be treated as one or two true clusters (there is a borderline for unimodality, but in some applications interpretative clusters need to be separated much more strongly than in a borderline bimodal mixture and in some others one would want to split up even unimodal sets if for example large within-cluster dissimilarities are not tolerated).

In most applications including social stratification the researcher may not be able to justify all the required decisions in detail; sometimes different aspects of what the researcher is interested in are in conflict, and there is feedback between the desired cluster concept and the data (the researcher may want very homogeneous but not too small clusters, or may have unrealistic

expectations about existing separation between clusters). Furthermore, it is hardly possible to anticipate the implications of all required tuning decisions for the given data. As mentioned before, researchers often want to check whether the found clusters carry more meaning than an optimal partition of homogeneous random data. Therefore it is not enough to specify a cluster concept and to select a method. Clustering results need to be validated further; see Sections 7.3 and 7.6.

6. Choice of methodology

In this section, all major aspects of the choice of a suitable methodology for socio-economic stratification with the US SCF data set are discussed. Many of these aspects are relevant for other cluster analysis tasks as well.

6.1. Variable transformations

According to the philosophy that is adopted here, effective distances on the variables should reflect the 'interpretative distance' between cases, and transformations may be required to achieve this.

For the variables income and savings amount (Fig. 3) log-transformations were applied. This was not mainly done because the resulting distribution looks more healthy (particularly taming the outliers dominating the untransformed variables; see Fig. 3), but rather because in terms of social stratification it makes sense to allow proportionally higher variation within high income and/or high savings clusters; the 'interpretative difference' between incomes is rather governed by ratios than by absolute differences. The difference between two people with yearly incomes of \$2 million and \$4 million, say, should in terms of social strata be treated about equally as the difference between \$20 000 and \$40 000.

Because there are 0s in the data, the transformation $\log(x+c)$ is suitable. The choice of c is meaningful, because it governs the effective distance between zero and small non-zero savings or income. We chose $c=50$ here, which makes the effective distance between 0 and 100 equal to that between 100 and 400 (for $c=1$, this number would be about 10 000) while leaving effective distances between high values virtually unchanged. In Fig. 3(b) it can be seen that there is no longer any clear gap between zero savings and low savings, and that observations with zero and low savings can end up in the same cluster. From a purely data-intuitive point of view, the large zero-savings group is certainly a significant pattern in the data but, in terms of sociological interpretation, there is no strong reason to isolate them in their own cluster unless this is supported by other variables.

For $c=1$ (Fig. 3(c)), clusterings are too strongly dominated by the 'zero-savings' group. The precise choice of 50 is subjective, but clusterings with $c=10$ and $c=200$ had less desirable features. 50 still leaves a clear gap between people with zero and non-zero income, but the interpretative difference between zero and non-zero income is larger than regarding savings, particularly because the smallest non-zero income was considerable larger than low non-zero-savings amounts.

The ordinal coding 'no, 1, 2, 3, 4 or 5, or 6 or more accounts' of the checking and savings account variables has been chosen because numbers of accounts put together in the same class can be seen as interpretively approximately equal; precise differences between large numbers of accounts do not matter.

We treat variable standardization and weighting in the next subsection; LCC will be used with a scale equivariant model for the covariance matrices, so that standardization and weighting are only relevant for dissimilarities.

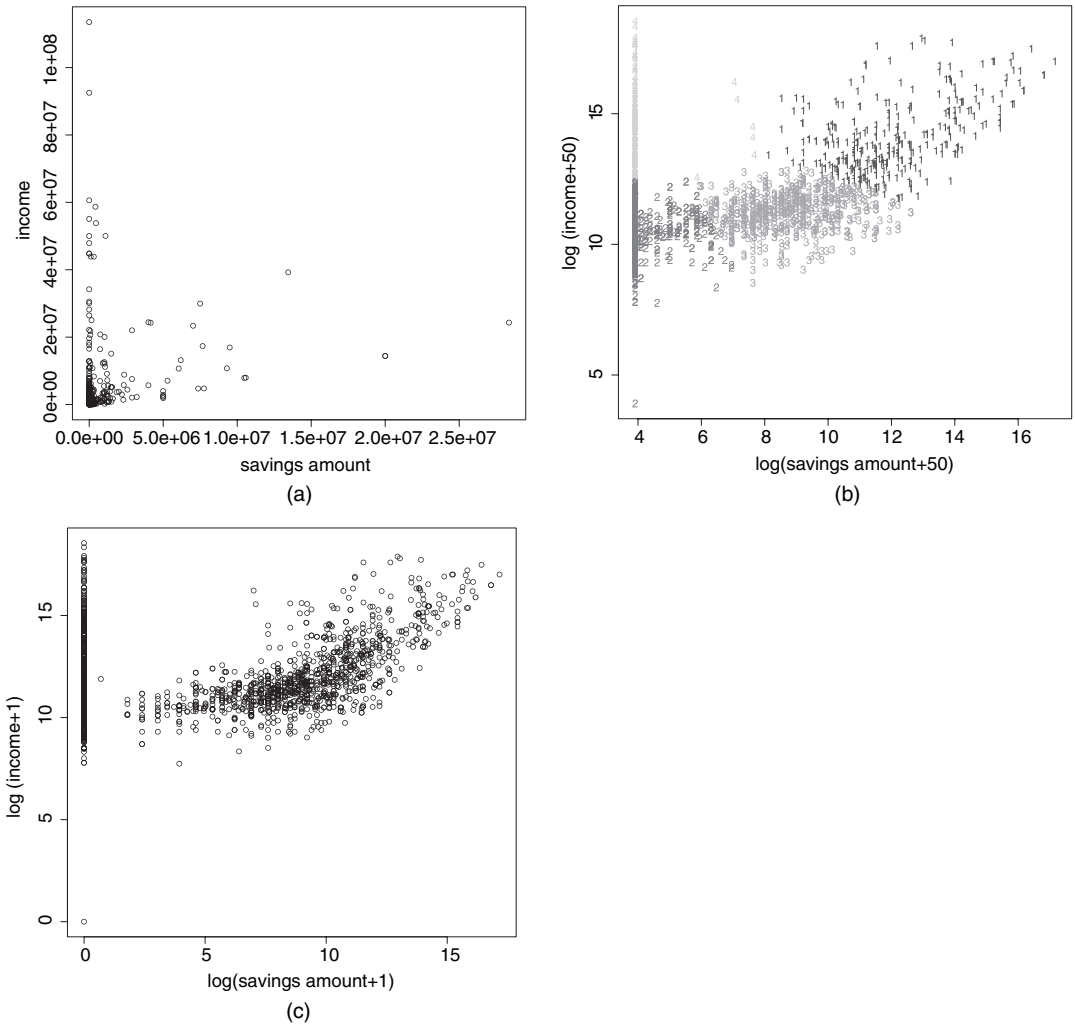

Fig. 3. Subset of income and savings data from the US SCF 2007, (a) untransformed, (b) with both variables $\log(x + 50)$ transformed (with 4-medoids clustering) and (c) with both variables $\log(x + 1)$ transformed

6.2. Definition of dissimilarity

To apply dissimilarity-based methods and to measure whether a clustering method groups similar observations together, a formal definition of 'dissimilarity' is needed. The measure should reflect what is taken as 'similar' in a given application (Hennig and Hausdorf, 2006).

The central task for mixed-type data is how to aggregate and how to weight the different variables against each other. Variablewise dissimilarities can be aggregated in a Euclidean, Gower–Manhattan or Mahalanobis style (further possibilities exist but are not discussed here).

The Euclidean distance between two objects $\mathbf{x}_i$ and $\mathbf{x}_j$ on p real-valued variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is

$$d_{\mathrm{E}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2} = \sqrt{\sum_{l=1}^{p} d_l(x_{il}, x_{jl})^2},$$

where d_l is the absolute value of the difference on variable l. The Gower distance (Gower, 1971) aggregates mixed-type variables in the same way as the Manhattan–L_1-distance aggregates continuous variables:

$$d_{\mathrm{G}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{p} d_l(x_{il}, x_{jl}).$$

Variable weights w_l can easily be incorporated in both aggregation schemes by multiplying the d_l by w_l.

The major difference between d_{G} and d_{E} is that d_{E}, by using squares, gives the variables with larger d_l more weight, i.e. two observations are treated as less similar if there is a very large dissimilarity on one variable and small dissimilarities on the others than if there is about the same medium-sized dissimilarity on all variables. Connecting this to social stratification may be controversial. In terms of interpretation it is not desirable to have extreme differences in income, savings and education within the same social class, which favours d_{E}. All other variables in the data set that is analysed here are nominal or ordinal and yield much lower maximum within-variable effective distances if treated as explained below. A further advantage of d_{E} for data sets with large n is that many computations for Euclidean distances (such as `clara` and the CH-index) do not need a full dissimilarity matrix to be handled. Therefore Euclidean aggregation is preferred here.

An alternative would be the Mahalanobis distance

$$d_{\mathrm{M}}(\mathbf{x}_i, \mathbf{x}_j)^2 = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

based on some covariance matrix estimator $\mathbf{S}$. The effect of using d_{M} is that, if variables are strongly correlated and are mainly scattered along certain 'lower dimensional directions' in data space, dissimilarities along these directions are downweighted and dissimilarities orthogonal to them are upweighted. This seems sensible if the correlation implies, in terms of interpretation, that much of the information in these variables is redundant, as, for example, in genetic set-ups involving many genes of no individual interest. However, in social stratification, variables are included in a study because they all have individual interpretative value. In such a situation, if they are statistically correlated, the fact that they share statistical information does not lower their importance for clustering, so they should not be downweighted by using d_{M} (the same argument favours 'local independence' in LCC compared with fully flexible covariance matrices or any affine equivariant covariance matrix model).

The definition of d_{E} can be extended to mixed-type variables analogously to how Gower extended d_{G}. Ordinal variables can be used with standard Likert scores $(1, 2, \ldots, |\mathcal{O}_j|)$. There are various ways to assign alternative scores, e.g. by using quantiles of an assumed underlying Gaussian distribution, assigning average ranks or based on other available variables by using techniques such as monotonic regression. 'Estimating' scores from the other variables is not sensible here, because the information in those variables is used for the clustering directly. Using Gaussian quantiles or average ranks makes the effective distances dependent on the distribution. Gaussian quantiles will normally make differences between central categories small; average ranks will assign the largest between-neighbours distances between strongly populated categories. There is no reason to expect that this reflects interpretative distance better than by using standard scores here, because treating interpretations for neighbouring categories for the variables educ, cacc and sacc as equidistant looks approximately justified. Furthermore, both schemes above fit the effective distribution to a certain homogeneous shape (the Gaussian, or, using average ranks, the uniform distribution), which runs counter to the possibility that meaningful clustering could be present in their marginal distribution.

The different values of a nominal variable should not carry numerical information. Therefore nominal variables should be replaced by binary indicator variables for all their values (let m_j denote the number of categories of variable j) to treat them symmetrically.

The variables then need to be weighted to make them comparable for aggregation. There are two aspects of this: the statistical aspect (the variables need to be standardized to have comparable distributions of values) and the substantial aspect (subject matter knowledge may suggest that some variables are more or less important for clustering).

There are various ways of standardization, e.g.

(a) range standardization (e.g. to $[0, 1]$),
(b) standardization to unit variance and
(c) standardization to unit interquartile range.

The main issue here is how the methods deal with extreme observations. Range standardization is governed by the two most extreme observations, and in the presence of outliers this can mean that pairwise differences on such a variable between a vast majority of observations could be approximately 0 and only the outliers are considerably far away from the rest. Using a robust statistic such as the interquartile range, however, means that, if there are extreme outliers on a certain variable, the distances between these outliers and the rest of the observations on this variable can still be very large and outliers on certain variables may dominate distances on other variables.

Standardization to unit variance is adopted here, which is a compromise between the two approaches that were discussed before. The gap between no income and low but non-zero income should neither dominate the contribution of the income variable too much; nor should it have an overwhelming influence on the cluster membership compared with the other variables. Another reason is that, in a situation with mixed-type variables, categorical variables must be standardized so that the dissimilarities between their levels are properly comparable with the dissimilarities between observations on continuous variables. This can be better calibrated on the basis of variances than on quantiles. Range standardization is not preferred because it makes interpretative sense that the most extreme dissimilarities in income and savings dominate dissimilarities between categories.

Nominal variables are considered next. Standardizing continuous variables to unit variance implies $E(X_1 - X_2)^2 = 2$ for independent and identically distributed X_1 and X_2. For an originally nominal variable with I categories there are I dummy variables. Let $Y_{ij}, i = 1, \ldots, I,$ be the value of observation j on dummy variable i. Dummy variables Y_i can be standardized by setting $\Sigma_{i=1}^{I} E(Y_{i1} - Y_{i2})^2 = q\, E(X_1 - X_2)^2 = 2q$ with some factor q. The rationale for this is that, in the Euclidean distance d_{E}, $\Sigma_{i=1}^{I} (Y_{i1} - Y_{i2})^2$ is aggregated with $(X_1 - X_2)^2$. $q = 1$ may seem natural, making the expected contribution from the nominal variable on average equal to the contribution from continuous variables with unit variance. This implies that $q = 1$ makes the difference between two different levels of the nominal variable larger than $E(X_1 - X_2)^2$ (because often $Y_{i1} = Y_{i2}$). This introduces wide 'gaps' between categories, which could force a cluster analysis method into identifying clusters with levels of the categorical variable. Therefore we use $q = \frac{1}{2}$. This implies that, for an originally binary variable for which the probability of both categories is about $\frac{1}{2}$, the effective distance between the two categories is about equal to $E(X_1 - X_2)^2 = 2$ (and correspondingly lower for variables with more than two categories). Larger values of q give the nominal variables higher clustering impact.

The expected value in

$$\sum_{i=1}^{I} E(Y_{i1} - Y_{i2})^2 = 2q \tag{6.1}$$

depends on the category probabilities. As a default for standardization these can be taken to be $1/I$ for each category. An obvious alternative is to estimate them from the data. This would increase the effective distance between categories for variables with higher differences between the category probabilities. In the given set-up with clear interpretative meanings of the categories we prefer not to let the effective dissimilarities between them depend on their empirical frequencies.

For ordinal variables Y with standard coding and I categories, our rationale for standardization is to enforce

$$E(Y_1 - Y_2)^2 = 2q, \qquad q = \frac{1}{1 + 1/(I-1)}, \tag{6.2}$$

with the expectation treated as above. For binary variables ($I = 2$) this is equivalent to equation (6.1), and with more levels the expected contribution converges towards that of a continuous variable. The variable educ has 18 categories with a strong and meaningful concentration on certain values (11 and above). Therefore we used the standard deviation for standardization, as for continuous variables.

Regarding substantial weighting, the housing levels are very unequally distributed. 'Owns' (72.6%) and 'pays rent' (17.9%) are by far the strongest categories. The other categories can be seen as in some sense lying between the former two in terms of interpretation. To stress the importance of owns and pays rent for dissimilarity clustering, we weighted the two dummy variables belonging to these categories up by a factor of 2, subsequently reweighting all dummy variables to keep equation (6.1) valid. This increases the effective distance between categories 3 and 4 compared with all other distances.

The ordinal variables cacc and sacc were downweighted by a factor of 0.5 because full weights make the diversification aspect overrepresented, and part of the information of these variables is shared with lsam and linc, particularly about the large number of 7343 individuals without savings (there are only five zero-income cases in the data). The life insurance indicator was downweighted by a factor of 0.5, also, because we see this as a comparably marginal indicator of the socio-economic status.

To summarize, we define the dissimilarity by

(a) Euclidean aggregation of variables,
(b) standardization of continuous variables to unit variance,
(c) using I dummy variables for each nominal variable, standardized according to equation (6.1) with $q = \frac{1}{2}$,
(d) using standard coding for each ordinal variable, standardized according to equation (6.2) and
(e) additional weighting of variables according to substantial requirements.

6.3. Comparing clustering methods

For all conceivable interpretations of social stratification it is important to find clusters with relatively small within-cluster dissimilarities. This means particularly that LCC is suitable only if it groups similar observations together. Taking the considerations in Section 5 regarding Fig. 1 and the discussion regarding the Mahalanobis distance in Section 6.2 into account, we assume that the covariance matrices are diagonal, and we do not apply existing methods that relax the local independence assumption (which *defines* a cluster shape rather than assuming that this is the objectively true shape).

An important difference between LCC and k-medoids (and also k-means) is that LCC allows strongly different within-cluster variances for different variables, whereas, for k-medoids

clustering, distances are treated in the same way regardless of their direction. This means that the variable weight in k-medoids clustering is mainly governed by dissimilarity design, whereas, in LCC, clusters may be very concentrated in one variable and very heterogeneous in one or more others. Regarding socio-economic stratification, advantages can be seen in both approaches. The LCC results will be more informative about the clustering tendency of the individual variables regardless of standardization and weighting. In contrast, the data analytic contribution of the variables may not match the interpretative importance, and LCC clusters may be too heterogeneous. It seems therefore suitable to compute an LCC clustering and to use criteria such as ASW allows us to check whether it does a good job regarding dissimilarity.

LCC as presented here assumes a clusterwise Gaussian distribution for the continuous variables. This implies that the method looks for data subsets in which deviations from a dense core are treated symmetrically, and strong deviations are penalized heavily. This requires the variables to be transformed in such a way that distances can be interpreted in the same way in all directions from the cluster centres as done in Section 6.1. Other distributional shapes (particularly with heavier tails) could be of interest but software incorporating these with mixed-type data is not available to our knowledge.

In expression (3.2) it is implicitly assumed that ordinality works by either increasing or decreasing monotonically the mixture component-specific contribution to $\tau_{hj}(y)$ through β_{hj}, which is somewhat restrictive. Without discussing the issue in detail, we rejected several alternative approaches (see, for example, Agresti (2002), Agresti and Lang (1993) and Drasgow (1986)) because of even less acceptable restrictions or the absence of straightforward generalizations to mixed-type data. A reviewer noted that model (3.2) could be tested against an unrestricted model, but using this would run counter to the presented philosophy, because it would prevent the ordinal meaning of the categories from influencing the clustering.

The methods that were introduced in Section 4 are more directly defined to make within-cluster dissimilarities small. There are many alternative dissimilarity-based clustering methods in the literature (see, for example, Kaufman and Rousseeuw (1990) and Gordon (1999)), e.g. the classical hierarchical methods, though there is no particular reason to impose a full hierarchy for social stratification. Similar data have been analysed by multiple correspondence analysis (e.g. chapter 6 of Le Roux and Rouanet (2010)). This requires continuous variables to be categorized and seems more suitable with a larger number of categorical variables.

As opposed to k-means, k-medoids clustering does not require the computation of means for nominal and ordinal categories (although this would not necessarily be detrimental for clustering). By using d instead of d^2, it is also somewhat more flexible in terms of cluster shapes (allowing within-cluster dissimilarities to vary more between clusters and variables) and less affected by outliers within clusters (Kaufman and Rousseeuw, 1990). This favours k-medoids because in social stratification a small number of moderately outlying observations can be integrated in larger clusters without affecting the medoid too much.

6.4. The number of clusters

Finding an appropriate number of clusters k is a notoriously difficult issue in cluster analysis. Following Section 5, we see finding k rather as a problem of an appropriate subject-matter-dependent definition of the required number of clusters than an estimation problem. This suggests that the problem cannot be boiled down to the choice between several apparently 'objective' criteria.

In the LCC model (3.1), finding the true k is usually interpreted as a statistical estimation problem. The BIC (the log-likelihood penalized by $\frac{1}{2}\log(n)$ times the number of free

parameters) is the most popular method. Its consistency for estimating k was proved in a simpler set-up by Keribin (2000). However, using this as a justification for the use of the BIC to estimate the number of clusters is ill posed. If the model does not hold precisely, the truth may be best approximated according to the BIC (or any consistent criterion) by a very high number of mixture components if there are only enough observations, which is of little interpretative value. The use of the BIC for estimating the number of clusters depends on n and on how strongly separated clusters are required to be (the larger n, the stronger the BIC's tendency to fit apparently homogeneous data subsets by more than one mixture component). In Section 7 it will turn out that for this reason the BIC does not do a good job here (for Gaussian mixtures there are amendments; see for example Biernacki *et al.* (2000) and Hennig (2010)).

The dissimilarity-based criteria that were introduced in Section 4 attempt to formalize a 'good number of clusters' directly. Several such criteria have been proposed in the literature, and no clear justification exists about which is best. One reason for choosing ASW, CH and PH was that their definitions allow a fairly straightforward data analytic interpretation.

ASW and CH attempt to balance a small within-cluster heterogeneity (which increases with k) and a large between-cluster heterogeneity (which increases with k as well) in a supposedly optimal way. In the definition of ASW, the dissimilarities of observations from other observations of the same cluster are compared with dissimilarities from observations of the nearest other cluster, which emphasizes separation between a cluster and their neighbouring clusters. In the definition of CH, the proportion of squared within-cluster dissimilarities is compared with all between-cluster dissimilarities, which emphasizes within-cluster homogeneity more, and is through the use of squared dissimilarities more prohibitive against large within-cluster dissimilarities. PH emphasizes good approximation of the dissimilarity structure by the clustering in the sense that whether observations are in different clusters should strongly be correlated with large dissimilarity. Because the Pearson correlation is based on squares, this again penalizes large within-cluster dissimilarities.

In real data sets, the situation is often not clear cut, and the criteria may yield vastly different solutions. For example, for the data set that is shown in Fig. 3(b), LCC–BIC yields 13 clusters to give a reasonable Gaussian mixture approximation, k-medoids with ASW yields four (shown) and with CH 16 clusters. ASW emphasizes gaps between neighbouring clusters, which hardly exist in this data set. CH penalizes large within-cluster distances more strongly and therefore ends up with many small clusters; it yields fewer clusters than ASW in some higher dimensional situations where gaps exist but the majority of within-cluster distances cannot be made small by any reasonably small number of clusters. None of these indices can be used to assess $k = 1$ and they degenerate for $k \to n$, and should therefore be interpreted with care for very small and very large k, and regarding comparisons between index values for very different values of k. Simply optimizing the indices is therefore often less suitable than looking for locally best choices of k (see Section 7.3), and by using cluster validation and subject matter criteria to make a final decision.

It is controversial in the literature whether a social stratification should rather have many small groups or a few larger groups. The main contention is whether the stratification structure is composed of a few big classes or many microclasses (Grusky and Galescu, 2005; Weeden and Grusky, 2005; Weeden *et al.*, 2007). For the reasons given above, this cannot be objectively settled by data. In the present study we concentrate on rather rough descriptions of society with numbers of clusters k up to 20 and particularly (with the aim of comparing the clusterings to the seven occupation categories) between five and 10. Very small clusters are not of much interest, because they rather refer to particularities than to the general structure of society. It is helpful if a clustering method can integrate outlying observations into the nearest larger clusters.

Therefore, ASW can be preferred to CH for similar reasons as in the k-medoids *versus* k-means discussion.

Approximating the dissimilarity structure by the clustering is a rather marginal aim, and therefore PH is of less interest than ASW and CH. We still use all three criteria in Section 7 to investigate whether certain numbers of clusters stick out in various respects. Neither ASW, nor CH nor PH can handle $k = 1$, and we shall introduce a parametric bootstrap scheme to see whether the data set is significantly clustered at all.

There are approaches for clustering that define the number of clusters implicitly and do not require criteria to compare solutions for different k. Often these attempt to find density modes, e.g. Ester *et al.* (1996). Such approaches cannot remove the requirement of user tuning but only shift the problem to the choice of bandwidths or other parameters.

6.5. Computation details

LCC and k-medoids were applied to this data set with k ranging from 2 to 20. We used `samples=100` and `sampsize=200` in R function `clara`, which is slower but more stable than with the default value. The LCC solution was computed by LatentGOLD (Vermunt and Magidson, 2005a). We computed the ASW-, CH- and PH-criterion for all k-medoids and LCC solutions, and the BIC for all LCC solutions. ASW and PH require the computation of the whole dissimilarity matrix. To prevent this, we computed them from randomly partitioning the data set into 15 subsets, computing them on every subset and averaging the results. This is much more stable than `clara`'s internal computation of ASW based on the optimal subset.

A problem with LCC is that the likelihood can degenerate or yield spurious solutions at the border of the parameter space (e.g. if a variance parameter is close to 0). LatentGOLD treats this by computing a penalized maximum likelihood estimator maximizing the sum of the log-likelihood and the logarithm of the prior density of the parameters from using Dirichlet priors for the multinomial probabilities and inverse Wishart priors for the error variance–covariance matrices. The user can tune how critical borders of the parameter space are penalized.

Given the large number of observations and a certain amount of discreteness in the income and savings variables, we decided to increase the 'error variance prior parameter' from the default value 1 to 5. Vermunt and Magidson (2005b) interpreted this parameter as follows:

> 'The number can be interpreted as the number of pseudo-cases added to the data, each pseudo-case having a squared error equal to the total variance of the indicator (dependent variable) concerned'.

The increase led to fewer clearly spurious local optima of the penalized likelihood; we tried other prior default parameters without further improvement.

7. Data analysis results

7.1. Clustering results with all variables

First, we discuss the results involving all eight variables. ASW and CH were optimal for $k = 2$ for both k-medoids and LCC. The BIC was optimal for LCC for $k = 20$, all at the border of the parameter space. PH was optimal for $k = 3$ for k-medoids.

The 2-medoids solution is fairly well interpretable, collecting 7739 (44.4%) observations with rather low savings, income, education and corresponding categories of the other variables in one cluster. In terms of the distance-based criteria it is much better than the LCC solution with $k = 2$ clusters. This holds for all k-medoids solutions compared with LCC with the same k. Selected results can be seen in Table 1.

ASW, CH and PH produce local optima for $k = 6$ and $k = 8$ for k-medoids. For LCC the

Table 1. Dissimilarity-based criteria for some k-medoids and LCC clusterings

Method	k	*Results for the following criteria:*		
		ASW	*CH*	*PH*
k-medoids	2	0.275	8216	0.443
k-medoids	6	0.211	5479	0.452
k-medoids	8	0.211	4999	0.453
LCC	2	0.212	5568	0.406
LCC	8	0.136	3138	0.387
LCC	20	0.047	1538	0.330

solutions with $k = 6$ and $k = 8$ are locally optimal according to ASW and PH, whereas CH decreases monotonically with increasing k and the BIC improves monotonically. Despite a generally decreasing trend of ASW, the value for 8-medoids is slightly better than that for 6-medoids. The number of clusters 8 enables a fairly easy interpretation with enough differentiation. We use the 8-medoid solution therefore as the preferred solution, with more evidence in its favour presented below.

Generally the dissimilarity-based criteria favour k-medoids over LCC but it is still interesting to look at an LCC solution particularly regarding unweighted variable impact. For this we chose the solution with $k = 8$ because this produces a local optimum of ASW, corresponding to k-medoids. The BIC 'optimal' solution $k = 20$ is too bad regarding ASW to consider it.

A difference between k-medoids and LCC for general k is that k-medoids produces more uniform cluster sizes. For $k = 8$, cluster proportions vary between 0.05 and 0.23, whereas those for LCC are between 0.01 and 0.24 (for some larger k, LCC even produces empty classes). Very small classes are not of much interest in social stratification.

7.2. Interpretation of clustering results

For interpreting the clustering results, we focus on the k-medoids and the LCC solutions when $k = 8$ by studying the conditional distributions for key indicator variables. We begin by lining up the eight clusters from both the k-medoids and the LCC estimations by the cluster-specific mean income values as presented in Fig. 4. The rationale for this operation is simple: income is the most obvious reward for individuals in different socio-economic classes.

Between the two clustering methods, the LCC method produced two larger higher-class clusters with relatively lower mean income whereas the k-medoids generated a greater bulge in the middle-class clusters, with much smaller-sized upper classes that earned a much higher average income. Among the k-medoids-generated classes, there appears to be a greater amount of similarity between clusters 1 and 2, between clusters 3, 4, 5 and 6, and between clusters 7 and 8. The middle class composed of clusters 3–6 comprises 61% of the cases.

Now we examine three other indicator variables of socio-economic stratification. Fig. 5 presents relative frequency distributions of estimated clusters by using both methods according to occupation, number of savings accounts and level of education. The fill patterns in Fig. 5 follow Fig. 4, i.e. the lowest class is plotted with the densest horizontal lines, the lines become sparser and the angle is raised counter-clockwise by steps for higher classes, and the highest class is plotted with the sparsest vertical lines.

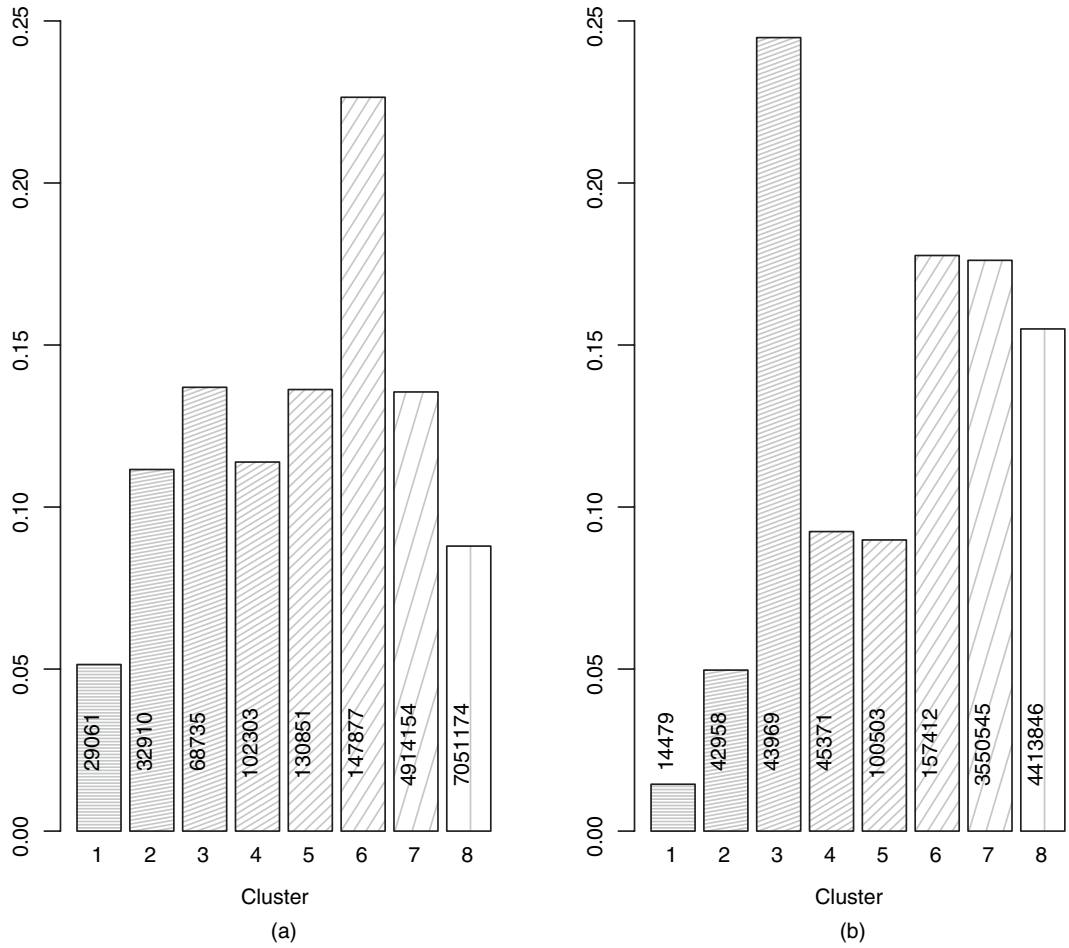

Fig. 4. Relative frequency distributions by cluster with mean income values labelled: (a) *k*-medoids solutions; (b) LCC solutions

The largest occupation group (occupation group 1) is that of managerials and professionals, dominated by the top three clusters produced by either clustering method. For the LCC solution, the top two classes have greater sizes because of their overall larger sizes (see Fig. 5). The next largest group (occupation group 0), or not working for pay, is almost equally represented by all the clusters from the *k*-medoids solution and six of the clusters from the LCC solution. Occupation group 2 consisting of white-collar workers and technicians is dominated by the second and third clusters from the top according to the *k*-medoids method but the domination is less clear according to the LCC method. Among service workers, repairers and operators (occupation group 4), the largest cluster is 5 (using either method).

The role of the number of savings accounts is less clear. According to the *k*-medoids solution, people who hold a greater number of savings accounts are in the highest two clusters but one. According to the LCC solution, people who hold no savings account are only from clusters 3 and 7, which is a rather odd pattern. The patterns of conditional distributions are intuitively appealing for education. University graduates and people with graduate education populate the top three clusters (judged by either estimation method). People with an associate or technical

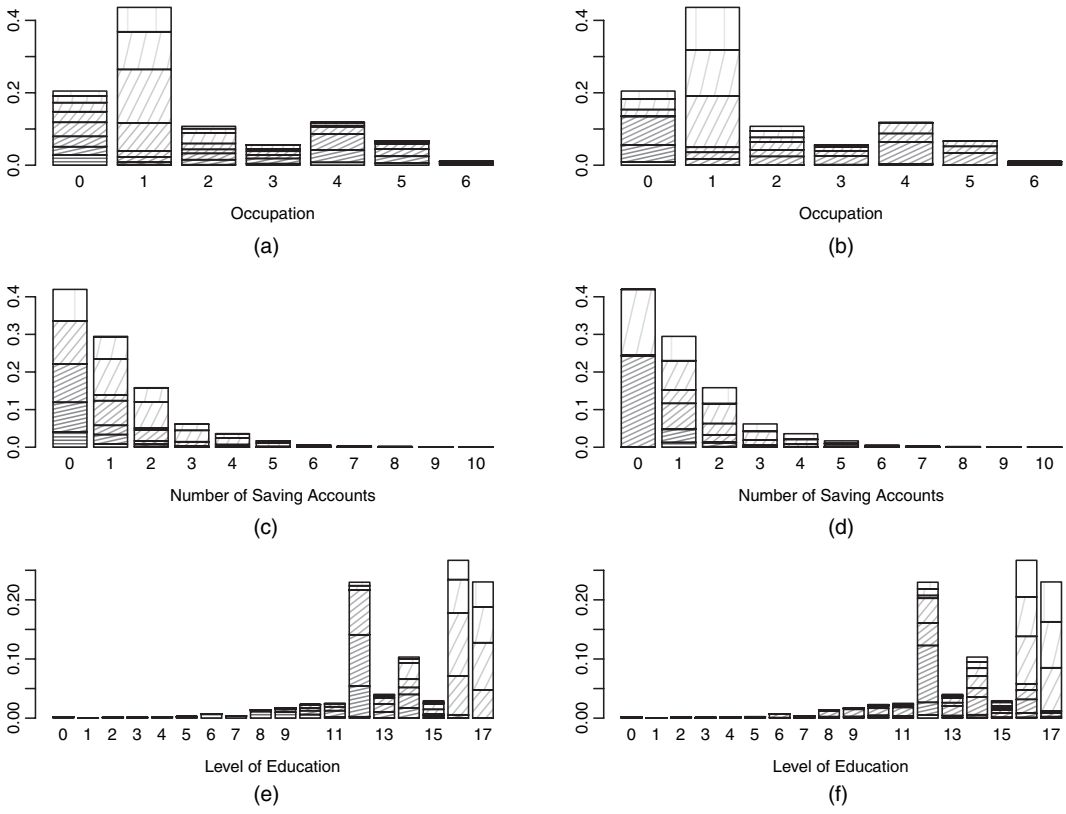

Fig. 5. Relative frequency distributions of clusters by (a), (b) occupation, (c), (d) number of savings accounts and (e), (f) level of education: (a), (c), (e) *k*-medoids solutions; (b), (d), (f) LCC solutions

degree concentrate in clusters 3–6), and people with a high school diploma only concentrate in clusters 3–5, according to the *k*-medoids solution. The distribution patterns according to the LCC solution are less clear cut, with more people in the top two clusters (7 and 8) than in the *k*-medoids results. The findings by examining cluster-specific conditional distributions of income, occupation and education collectively point to a socio-economic stratification system with eight classes with good differentiation, notably from the *k*-medoids solution.

7.3. Parametric bootstrap test for clustering

A major research question is whether one can state that the clusterings carry some meaning beyond a decomposition of the dependence between variables.

To address this question, we carried out a parametric bootstrap 'test' (because of its data-dependent nature we do not attempt to interpret the outcome in terms of formal *p*-values). The null model was designed as follows.

(a) For data generation (though not for data analysis), the nominal variables were treated as ordinal with categories ordered according to the average correlation of their dummy variables with the continuous and ordinal variables (this backed up the treatment of the owns and pays rent housing categories as the two extremes implied by the weighting above).
(b) A correlation matrix was computed by using polychoric correlations between ordinal variables (Drasgow (1986); the continuous variables were categorized into 10 categories

of approximately equal sizes for computing their correlations with the ordinal variables) and the Pearson correlation between the continuous variables.

(c) Data were generated according to a multivariate Gaussian distribution with the given correlation matrix.

(d) Ordinal categories were assigned, for the ordinal variables, by using appropriate quantiles of the corresponding marginal distributions to reproduce the original discrete marginal distributions of these variables.

The resulting distribution was interpreted as 'homogeneous', but reproducing the discrete marginal distributions and the dependence structure between the variables.

100 data sets of size 17430 were drawn from this null model and were clustered with k-medoids for k between 2 and 10 (ASW-values for the real data set were bad above $k = 10$) in the same way as for the real data set. The corresponding ASW-values can be seen in Fig. 6, along with those for the original data set.

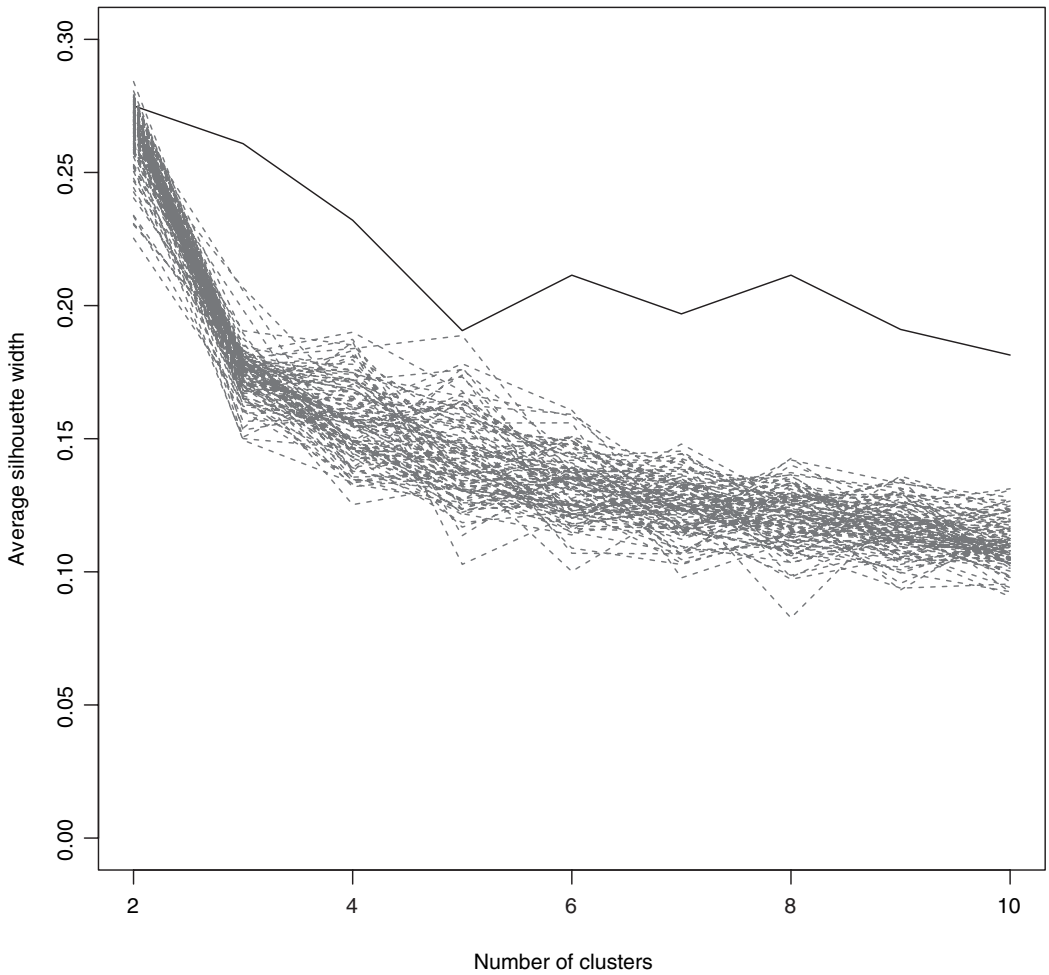

Fig. 6. Average silhouette width for US SCF data, k-medoids, k between 2 and 10 (———), and 100 parametric bootstrap samples for a null model preserving the dependence structure

Table 2. ARI between clustering on all variables (full data set) and clustering with a variable omitted†

Method	*Variables: ARI-values for the following variables:*							
	lsam	*linc*	*educ*	*cacc*	*sacc*	*hous*	*life*	*occ*
8-medoids	0.446	0.552	0.408	0.896	0.787	0.799	0.930	0.834
LCC, $k=8$	0.302	0.460	0.720	0.920	0.711	0.759	0.769	0.650

†Values near 1 mean that the variable has almost no impact.

Surprisingly, whereas the supposedly optimal $k=2$ does not look significantly better than what happens under the null model, the results for larger k clearly do. This particularly supports the solution with $k=8$, because under the null model the expected ASW goes down between 6 and 8.

Running such a test with LatentGOLD is much more difficult, and we ran only 20 bootstrapped data sets with $k=8$, producing a range of ASW-values between 0.056 and 0.079, which is much smaller than the value 0.136 from the original data set.

These results indicate that the clusterings found for $k=8$ reveal more socio-economic structure than can be explained from the marginal distributions and the dependence structure of the data alone.

7.4. Effect of variables

Variable impact was evaluated to check to what extent clusterings were dominated by certain (continuous, ordinal or nominal) variables. To do this, the clustering methods were applied to all data sets with one variable at a time left out, and the adjusted Rand index ARI (Hubert and Arabie, 1985) was computed between the resulting partition and the partition obtained on the full data set. The index compares two partitions of the same set of objects. A value of 1 indicates identical partitions; 0 is the expected value if partitions are independent. Values close to 1 here mean that omitting a variable does not change the clustering much, and therefore that the variable has a low impact on the clustering.

Table 2 shows that the biggest difference between LCC and 8-medoids in this respect is the impact of education, which is the most influential variable for 8-medoids but not particularly strongly involved for LCC. LCC is generally strongly dominated by the two continuous variables. The LCC results also show that the 'message from the data' corresponds nicely to the decision to downweight cacc, sacc and life for definition of dissimilarity, because they have a low impact despite not being downweighted in LCC.

The 8-medoids result turns out to be almost unaffected by occupation. The impact of housing is low for both methods.

7.5. Comparing clustering and occupation groups

The practice of using occupational categories compiled by the US Census Bureau or the Department of Labor as in the given data set and similar to the categories in landmark work such as Hollingshead (1957) and Blau and Duncan (1967) has been very influential for studying social stratification.

To see whether the occupation groups correspond to clusters on the other variables (which

would back up stratification based on occupation), the same procedure as before was applied to a data set in which the occ-variable was left out.

For k-medoids again a local optimum of the dissimilarity-based criteria was achieved with $k = 8$. We again used the LCC solution with $k = 8$ for comparisons.

ARI between 8-medoids and the occupation grouping was 0.091, and between LCC and the occupation grouping it was 0.058. Both of these values are close to 0, which is expected for totally unrelated clusterings, indicating that the occupation grouping has almost nothing to do with the clusterings on the other variables.

7.6. Cluster validation

Whereas in some literature the term 'cluster validation' is used almost synonymously for estimating the number of clusters, here it refers to an exploration of the quality of the achieved clusterings in the sense of how they relate to the researcher's requirements.

Many techniques can be used for this, including some methodology presented in the previous sections (particularly testing for homogeneity and the computation of indices such as CH, ASW and PH). For brevity, here we mention only that various other techniques have been applied to check interpretively relevant aspects of the clusterings, including visualization and comparison with further clusterings resulting from alternative methods or changing some tuning decisions (variable weighting, transformations and algorithmic parameters).

ARI between LCC, $k = 8$, and 8-medoids for the data set with all variables is 0.456, which implies a moderate amount of similarity between these clusterings. The value for 6-medoids and 8-medoids is 0.922, which means that these are about as similar as possible, given the different k.

Using educ as an ordinal variable is problematic in LCC because the high number of 18 categories enforces a high number of parameters. The impact of educ changed considerably when we used it as continuous, but in this case its discrete nature enforced many unstable local optima of the penalized likelihood.

Applying the bootstrap stability assessment from Hennig (2007) to the 8-medoids solution confirmed that four out of 8 clusters are very stable and the four others fairly stable, although with k treated as fixed.

8. Conclusion

We think that the application of automatic methods hoping that 'the data will enforce its true structure' is deceptive. Many decisions must be made for clustering. Section 6 is central for demonstrating the effect of the clustering philosophy that is adopted here, which connects the data analytic features of the formal methodology to the application as a rationale for making the required decisions.

The reasoning that is required for some of these decisions will often be fairly different from the typical way of thinking in the areas to which the clustering is to be applied, and therefore some work is required to connect them. This may have the helpful side effects that the researchers improve their understanding of the implications of their concepts.

In the US SCF 2007 data set, we found well interpretable clusterings with $k = 8$. The 8-medoids clustering is preferable in terms of bringing similar observations together. The most influential variables are savings amount and education. The LCC solution with unweighted variables confirmed the decision to downweight certain variables (account numbers and life insurance) for 8-medoids; it is more dominated by income and less by education. We found that socio-economic strata in these data hardly correspond to occupation classes. Both the 8-medoids and

LCC, $k = 8$, clusterings are locally optimal and established as meaningful in the sense that they contain more structural information than just the dependence structure between the data.

Acknowledgement

The second author acknowledges the sabbatical support from the University of Illinois in 2008–2009 that helped the research in this paper.

References

Agresti, A. (2002) *Categorical Data Analysis*, 2nd edn. New York: Wiley.
Agresti, A. and Lang, J. (1993) Quasi-symmetric latent class models, with application to rater agreement. *Biometrics*, **49**, 131–139.
Baker, F. B. and Hubert, L. J. (1975) Measuring the power of hierarchical cluster analysis. *J. Am. Statist. Ass.*, **70**, 31–38.
Bernheim, B. D., Garrett, D. M. and Maki, D. M. (2001) Education and saving: the long-term effects of high school financial curriculum mandates. *J. Publ. Econ.*, **80**, 435–464.
Biernacki, C., Celeux, G. and Govaert, G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattn Anal. Mach. Intell.*, **22**, 719–725.
Blau, P. M. and Duncan, O. D. (1967) *The American Occupational Structure*. New York: Wiley.
Brennan, M. J. and Schwartz, E. S. (1976) The pricing of equity-linked life insurance policies with an asset value guarantee. *J. Finan. Econ.*, **3**, 195–213.
Calinski, R. B. and Harabasz, J. (1974) A dendrite method for cluster analysis. *Communs Statist. Theor. Meth.*, **3**, 1–27.
Chan, T. W. and Goldthorpe, J. H. (2007) Social stratification and cultural consumption: the visual arts in England. *Poetics.*, **35**, 168–190.
Drasgow, F. (1986) Polychoric and polyserial correlations. In *The Encyclopedia of Statistics* (eds S. Kotz and N. Johnson), vol. 7, pp. 68–74. New York: Wiley.
Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases. In *Proc. Int. Conf. Knowledge Discovery and Data Mining*, pp. 226–231. Menlo Park: American Association for Artificial Intelligence Press.
Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*, 5th edn. New York: Wiley.
Gordon, A. D. (1999) *Classification*, 2nd edn. Boca Raton: Chapman and Hall–CRC.
Gower, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–871.
Grusky, D. B. and Galescu, G. (2005) Foundations of class analysis: a Durkheimian perspective. In *Approaches to Class Analysis* (ed. E. O. Wright), pp. 51–81. Cambridge: Cambridge University Press.
Grusky, D. B., Ku, M. C. and Szelnyi, S. (2008) *Social Stratification: Class, Race, and Gender in Sociological Perspective*. Boulder: Westview.
Grusky, D. B. and Weeden, K. A. (2008) Measuring poverty: the case for a sociological approach. In *Many Dimensions of Poverty* (eds N. Kakwani and J. Silber), pp. 20–35. New York: Palgrave–Macmillan.
Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001) On clustering validation techniques. *J. Intell. Inform. Syst.*, **17**, 107–145.
Hennig, C. (2007) Cluster-wise assessment of cluster stability. *Computnl Statist. Data Anal.*, **52**, 258–271.
Hennig, C. (2010) Methods for merging Gaussian mixture components. *Adv. Data Anal. Classifcn*, **4**, 3–34.
Hennig, C. and Hausdorf, B. (2006) Design of dissimilarity measures: a new dissimilarity measure between species distribution ranges. In *Data Science and Classification* (eds V. Batagelj, H.-H. Bock, A. Ferligoj and A. Ziberna), pp. 29–38. Berlin: Springer.
Hollingshead, A. B. (1957) Two factor index of social position. *Mimeo*. Yale University, New Haven.
Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classifcn*, **2**, 193–218.
Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data*. New York: Wiley.
Kennickell, A. B. (2000) Wealth measurement in the Survey of Consumer Finances: methodology and directions for future research. *Working Paper*. US Board of Governors of the Federal Reserve System, Washington DC. (Available from `http://www.federalreserve.gov/pubs/oss/oss2/method.html`.)
Keribin, C. (2000) Consistent estimation of the order of a mixture model. *Sankhya* A, **62**, 49–66.
Kingston, P. W. (2000) *The Classless Society*. Stanford: Standard University Press.
von dem Knesebeck, O. (2002) Social inequality and health of the elderly: classical or alternative status indicator? *Zeits. Gerontol. Geriatr.*, **35**, 224–231.
Lenski, G. E. (1954) Status crystallization: a non-vertical dimension of social status. *Am. Sociol. Rev.*, **19**, 405–413.
Le Roux, B. and Rouanet, H. (2010) *Multiple Correspondence Analysis*. Thousand Oaks: Sage.
Levy, F. and Michel, R. C. (1991) *The Economic Future of American Families: Income and Wealth Trends*. Washington DC: Urban Institute Press.

Liao, T. F. (2006) Measuring and analyzing class inequality with the Gini index informed by model-based clustering. *Sociol. Methodol.*, **36**, 201–224.
Milligan, G. W. and Cooper, M. C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.
Pekkanen, J., Tuomilehto, J., Uutela, A., Vartiainen, E. and Nissinen, A. (1995) Social class, health behaviour, and mortality among men and women in Eastern Finland. *Br. Med. J.*, **311**, 589–593.
Poterba, J. M., Venti, S. F. and Wise, D. A. (1994) Targeted retirement saving and the net worth of elderly American. *Am. Econ. Rev.*, **84**, 180–185.
R Development Core Team (2011) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Spilerman, S. (2000) Wealth and stratification process. *A. Rev. Sociol.*, **26**, 497–524.
Srivastava, R. K., Alpert, M. I. and Shocker, A. D. (1984) A customer-oriented approach for determining market structures. *J. Marktng*, **84**, 32–45.
Sugar, C. A. and James, G. M. (2003) Finding the number of clusters in a dataset: an information-theoretic approach. *J. Am. Statist. Ass.*, **98**, 750–763.
Vermunt, J. K. and Magidson, J. (2002) Latent class cluster analysis. In *Applied Latent Class Analysis* (eds J. A. Hagenaars and A. L. McCutcheon), pp. 89–106. Cambridge: Cambridge University Press.
Vermunt, J. K. and Magidson, J. (2005a) *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Belmont: Statistical Innovations.
Vermunt, J. K. and Magidson, J. (2005b) *Latent GOLD 4.0 User's Guide*. Belmont: Statistical Innovations.
Weeden, K. A. and Grusky, D. B. (2005) The case for a new class map. *Am. J. Sociol.*, **111**, 141–212.
Weeden, K. A., Kim, Y.-M., Matthew, D. C. and Grusky, D. B. (2007) Social class and earnings inequality. *Am. Behav. Scient.*, **50**, 702–736.
Weisbrod, B. A. and Hansen, W. L. (1968) An income-net worth approach to measuring economic welfare. *Am. Econ. Rev.*, **58**, 1315–1329.
Wright, E. O. (1985) *Classes*. London: Verso.
Wright, E. O. (1997) *Class Counts: Comparative Studies in Class Analysis*. Cambridge: Cambridge University Press.

Discussion on the paper by Hennig and Liao

Sabine Landau (*King's College London*)

Among many contributions Hennig and Liao's paper provides

(a) a review of methods for clustering mixed-type variables,
(b) an analysis strategy for obtaining a social stratification and
(c) a proposed new clustering philosophy.

I would like to discuss two questions.

(i) How does the analysis strategy for social stratification differ from what is typically done in applications of cluster analytic methods?
(ii) Do we need a new cluster analysis (CA) philosophy?

In the paper social stratification is defined as a categorical predictor of social, behaviour and health outcomes and it is stipulated that the relevant dimensions to be grouped are those measuring socio-economic wellbeing. The 2007 US Survey of Consumer Finances contains measures of these dimensions. Importantly the variables are of mixed type—including continuous, ordinal, nominal and binary measures.

The social stratification problem would seem an obvious cluster problem, i.e. identifying and describing distinct clusters with regard to socio-economic wellbeing variables. However, as Hennig and Liao point out, in this application there is no interest in identifying groups of people who have different covariance matrices. Only clusters that shift their centres along one of the dimensions of socio-economic wellbeing are of interest. Such a research objective differs from that typically addressed by CA. 'Clusters' are traditionally understood as latent subpopulations whose multivariate distributions vary in some feature, including their covariance matrices (Everitt *et al.*, 2011). The aim of (model-based) CA methods (finite mixtures) is to describe distributions within subpopulations. Thus one would not expect CA to be able to address the social stratification research question without modification. (Perhaps the research question would be better phrased as 'finding and describing segments'. I shall return to this point later.)

Perhaps the most important message of the paper is that subject matter understanding needs to be brought into the process of CA, to aid relevant choices. Applications of CA typically involve the following steps (Milligan, 1996):

step 1—objects to cluster;
step 2—variables to be used;
step 3—missing values;
step 4—variable standardization;
step 5—proximity measure;
step 6—clustering method;
step 7—number of clusters;
step 8—replication and testing;
step 9—interpretation.

The paper provides us with an analysis strategy. Advice is given on how to carry out these steps so that resulting partitions meet the requirements of a social stratification. Variable standardization (step 4) is not necessary if scale invariant CA methods such as finite mixtures are to be used. However, the issue needs to be addressed when the use of proximity-based CA methods is envisaged. To adapt step 5 to social stratification a new proximity measure is proposed: the new measure uses the Euclidean distance metric to aggregate variablewise dissimilarities as social strata are desired not to have extreme differences on a variable within the same class. The use of the Mahalanobis distance, which is often advocated in CA applications, is rejected. To combine continuous with categorical variables the latter are replaced by dummy variables and scaled so that the expected contribution of a single dummy variable to the aggregate proximity is half that of a (standardized) continuous variable. The paper suggests two suitable partitioning methods for social stratification (step 6):

(a) fitting a finite mixture which assumes diagonal covariance matrices for continuous variables and local independence (these assumptions are to ensure that any resulting partition does not reflect differences in the dependence structure, rather than to model the population);
(b) running a k-medoids clustering algorithm based on a suitably defined proximity matrix.

When it comes to estimating the number of groups (step 7) Hennig and Liao make the important point that the choice of the number of social strata cannot be objectively settled from the data. The optimality of any index for comparing the fit between different numbers of clusters relies on the population model being fitted. As a way forward a new parametric bootstrap procedure is suggested which allows interpreting detected groupings as extra socio-economic structure not captured by the marginal distributions and the dependence structure of the data alone.

Hennig and Liao's paper demonstrates that in a particular application such as social stratification the research aim might *not* be to identify and describe all distinct subpopulations and thus fitting wrong models can be an appropriate analysis strategy. However, in my view this does not constitute a new clustering philosophy; nor is one required. The research objective of social stratification is not the objective of traditional CA; rather it is a more general segmentation objective. I therefore would consider the very appropriate analysis strategy that is proposed here a segmentation strategy which includes the use of cluster analytic methods.

Of course this last point is largely one about terminology. I have very much enjoyed the opportunity to think about the issues that are raised in this paper. Mixed-type data have been given some long overdue attention. I trust that the proposed segmentation analysis strategy will help many applied researchers to conceptualize and implement their partitioning requirements.

It gives me great pleasure to propose the vote of thanks.

Paul S. Lambert (*University of Stirling*)

Hennig and Liao's paper makes a welcome contribution to our understanding of techniques for classification, and their effective implementation through software. The application area of the paper addresses a long-standing question about how individuals' positions within the social structure of inequality are best understood in the presence of multiple plausible indicator measures. Traditionally, writers in the social sciences find it nearly impossible to describe social inequalities without resorting to 'low k' classifications, sometimes of a relatively arbitrary character, so Hennig and Liao's analysis, which shows how a realistic mixture of variable types can be constructively combined to generate plausible low k typologies, has great potential.

The application-oriented claims of the paper are significant, and it is on those that the following comments concentrate. Some of the claims are compelling but, on cross-examination, some of them raise further questions. A convincing claim, for sure, is that more extended empirical work can be considered on the classification of individuals, taking account of multiple measures about individuals' circumstances,

and that the classification approaches that are developed in the paper make new contributions to this field. There is a wider social science literature on multiple-indicator measures such as in measuring poverty or in using predicted values from statistical models based on 'human capital' attributes (e.g. Gershuny (2002)); the latter certainly accommodates mixed-type indicator measures so it would be interesting to discuss how that compares with the recommended classificatory approach.

A more tenuous but prominent claim of the paper, however, is that occupational categories are relatively less important for understanding individuals' circumstances. Though this is a popular claim in some areas of social science, there are various reasons why the paper, thus far, might not show this conclusively. Firstly, occupations are not given much of an opportunity in the application to have an influence. Just one broad brush, low k occupation-based measure is considered, when hundreds of alternatives exist, in a range of functional forms, including low and high k categorizations and single- and multi-dimensional scales. Moreover, the occupational categories that are used include a 'not-working' classification, which by most accounts serves to conflate problematically two different forms of social inequality (i.e. inequalities between people in occupations, and inequalities between those who do and do not have occupations); a multiprocess model might be considered here, or, more simply, allocation of people to occupational categories on the basis of their last main job and/or the job of an important alter (as is common practice in social research). Second, the discussion does not really acknowledge what it is about occupation-based classifications which makes them especially appealing, which is an omission which has implications for substantive conclusions. The most relevant issue is that it is regularly argued that occupations give us stable and reliable indicators of longer-term life course circumstances of people, and that it is these that are most important to understanding an individual's position within the structure of social inequality (there are other attractions to using occupational data that are worth highlighting, such as the relative ease of collecting reliable information on them, and consistent coding schemes across societies). No social scientist would deny that there are very substantial within-occupation (or within-occupation-based category) variations in outcomes such as income, savings and lifestyle but many of these are thought to come down to 'exogenous' factors such as age, birth cohort, region and 'innate' characteristics, which we would not ordinarily want to integrate with our understanding of the social structure of inequality.

This raises a wider criticism that the terminology and theorization of 'social stratification' used by Hennig and Liao is somewhat inconsistent and varies somewhat to conventions in sociology and allied disciplines (or at least to those of European sociology—North American sociology does admittedly adapt slightly different vocabularies in this area). It is noted early on that 'social stratification is about partitioning a population into several different social classes' (page 310), yet two paragraphs later 'social class' and 'social stratification' are distinguished, and Hennig and Liao do recognize that candidate measures of position within the social stratification structure are variably continuous, high k and low k (page 311). The sociological convention is to regard 'social stratification' as a social phenomenon describing long-term and enduring social patterns to inequalities in the allocation of resources (e.g. Bottero (2005) and Kerbo (2003)). From this perspective, if the paper did want to make a strong statement about the nature of social stratification, some further points should be considered. First, the cross-sectional analysis that is presented ignores longitudinal developments in the outcomes studied, yet the measures used by Hennig and Liao have strong but differential relationships to life course stage and to cohort change—for instance, educational qualifications are much higher for younger birth cohorts, and income variations (within and between occupations) are substantially related to age. Indeed, several sociologists have tried to use classification tools in similar ways to characterize life course trajectories (e.g. Tampubolon and Savage (2012), Sturgis and Sullivan (2008) and Pollock (2007)) and these would be worth discussing in this context. In any case, without controls for age, we can suspect that empirical clusters in the patterns of response that are summarized by Hennig and Liao are linked to age and birth cohort, but if so this would not be a good reflection on the enduring character of social stratification inequalities, since an individual's position in the stratification structure is typically conceived as more about their enduring life chances than their situation in the 'here and now' (similar points could made about gender and household circumstances, though these are sensibly minimized by Hennig and Liao's focus on males only). Second, Hennig and Liao ought to recognize that, when studying the character of social stratification, many sociologists are explicitly attempting to disentangle the role of different social mechanisms on other outcomes, such as can be linked to occupational class, income, education or tenure, deliberately to achieve a meaningful account of the competing role of different social mechanisms. From this perspective, an indicator which depicts the empirical co-occurrence of quite different circumstances does not by definition illuminate our understanding of stratification.

A solution is perhaps simple—the paper might be much better expressed as being about the constellation of socio-economic circumstances in which individuals can be found at the particular moment in time (rather than enduring position in the structure of social stratification). The paper makes exciting inroads here. Others have worked on this topic, most consequentially in the commercial sector where market researchers seek to predict purchasing behaviours on the basis of a range of socio-economic measures. These approaches, however, also incorporate regional and demographic data, which are difficult to justify ignoring when the interest is in improved predictive description; it remains for the authors to demonstrate why the constellation of strictly socio-economic differences that are summarized through the paper is of particular interest (or to explore how substantially clustering patterns may vary if different combinations of input measures are applied).

Another exciting prospect for the paper could be to make it more rather than less specific. Some classificatory approaches have already been used productively to explore whether nominated economic components (such as 'employment relations' and 'employment conditions') coincide with 'social class' differences (e.g. Birkelund *et al.* (1996) and Evans and Mills (1998)). An important contribution of this paper is showing how mixed types of variables can be combined in such studies—there might be a great opportunity therefore for the authors to show through sensitivity analysis the relative benefits of using mixed functional forms in a comparable analysis of the character of social class measures.

The authors kindly made software files available which allow replication of some of their classification methods on the data set used. I reran their analysis on comparable variables from a comparably large UK survey (the British Household Panel Survey, 2005 sample). Disregarding sampling weights and several other potentially significant features of those data, I replicated the k-medoid clustering for a variety of age ranges and a variety of occupation-based social classifications, with and without a 'not-working' category included in the classification. Three things stood out in my preliminary results: there was quite some variation in the classifications identified according to different age groups and occupation-based measures; for classifications of males of all ages, age usually correlated fairly strongly with classification category, at least as strongly as did some of the other input measures; in general the metric measures of income and savings had much stronger relationships to the classification categories than the ordinal, categorical, and quasi-continuous measures (i.e. an interval scale for education), and indeed analysis using continuous occupation-based measures led to a much greater influence on the clustering by that occupational measure. My analyses are preliminary and would benefit from further replication, but the latter finding in particular might raise the question of whether the clustering mechanism might be responding disproportionately to the more finely measured continuous measures and, if so, whether this is desirable.

The vote of thanks was passed by acclamation.

Angela Montanari and Paola Monari (*University of Bologna*)

It is a great pleasure to comment on such a thought-provoking paper.

The authors' philosophy of clustering is deeply rooted in the statistical literature, whereas the strategy that they derive is definitely original.

25 years ago the *Journal of the Royal Statistical Society*, Series A, published the paper 'What is projection pursuit?' (Jones and Sibson, 1987). The method aimed at automatically finding 'interesting' low dimensional projections of a multivariate data set by optimizing a suitable measure of interestingness (the projection index). Most of the discussion following that paper and others, published around the same time, in the *Annals of Statistics* and in the *Journal of the American Statistical Association* (e.g. Huber (1985) and Friedman (1987)) equated 'interestingness' with 'showing a clustered structure' and a plethora of projection indices appeared in the literature afterwards, each of which responding to a different idea of what clusters are. Hennig and Liao cleverly remind us of this sometimes-neglected leitmotiv—clustering requires an *a priori* definition of what kind of clusters is searched—and they make many steps forward. They elegantly extend their reasoning to mixed-type data and, what is more, they provide a whole clustering strategy: practical advice on how to perform clustering according to their philosophy. Nothing is left to chance: from variable selection to dissimilarity definition, from the choice of the clustering method to the identification of the number of clusters, from bootstrap assessment of clustering to variable impact evaluation. All steps are carefully performed in a continuous dialogue between data and methods. Whereas some decisions are strongly theoretically motivated and convincing—i.e. the justification for conditional independence and the parametric bootstrap approach—others look somehow subjective—i.e. the choice of the constant in variable transformation or of the variable weights—and make us doubt about the reproducibility of the results. We agree that we can never assume that we are model free in clustering: even when we make no probabilistic assumption we certainly make assumptions on the data structure. But how

precise should these assumptions be? And how strong is their effect on the analysis? In other words, how can the authors' philosophy be extended outside the case at hand and for instance be employed in data mining, where clustering is a major tool? According to Hand *et al.* (2001) the goal of data mining is 'to find unsuspected relationships and summarize the data in novel ways'. After going through all the steps that the authors suggest, is there still room for the unsuspected and the novel?

Laura Anderlucci (*University of Bologna*)

On the comparison between clustering methods

It is always instructive to see how Christian Hennig addresses a topic by decomposing it into suitably posed subproblems. In this paper the problem is clustering and the subproblems are all the data-related major aspects of the choice of a suitable methodology. I shall concentrate my comment on the comparison between clustering methods, and in particular on the capability of latent class clustering (LCC) to group similar observations together, by providing some results on categorical data that I obtained in my doctoral dissertation (Anderlucci, 2012), for which Dr Hennig was co-supervisor. The research aimed to investigate the behaviour of an LCC and a distance-based approach in terms of quality of clustering in the case of nominal variables. Do both of these approaches lead to the same clustering? And how good are clustering methods designed to find observations that come from the same probability distribution in terms of dissimilarity-based criteria? To answer these questions a latent class model and partition around medoids were evaluated and compared in a fairly wide simulation study. The simulations were set according to the variation of several data features: the number of latent classes, the number of manifest variables, the sample size etc. For all the possible combinations of these factors (i.e. 128) we considered 2000 replications. Once both models had been run, we compared the obtained classifications according to, among others, the average silhouette width. One would expect the distance-based method to perform better than the LCC, but actually no method always outperforms the other on average, so it is not easy to make general statements. LCC, by trying to put together observations coming from the same distribution, succeeded in collecting similar observations together and in separating objects that are very different in a way that was not much worse than partition around medoids. What was surprising is that sometimes it worked even better (in 63 over 128 settings, on average). Since all the data features proved to have a significant effect on the difference in quality of clustering, it is not possible to restrict *a priori* the area of competence of the two approaches. This confirms how the application of automatic methods can be misleading and the idea that decisions (at each level) should always be made when doing clustering. In real life applications one can always perform both methods and then choose the one that gives better results in terms of some dissimilarity-based criteria.

Gero Szepannek (*Dortmund University of Technology*)

In the paper a philosophical question is raised concerning the implicit underlying assumptions of cluster analysis. In applications its aim is often settled somewhere in between purely data-driven discovery of structure and the integration of experts' knowledge and expectations on the interpretability of the results. Several preceeding methodological decisions and transformations are motivated and discussed.

If—as it is outlined in the paper—the identified clusters are supposed to be characterized by increasing values in some continous variable like income it might be an alternative to integrate this expert knowledge into the analysis, e.g. by using it as target variable for a regression tree (Breiman *et al.*, 1984) rather than performing cluster analysis. This would result in homogeneous groups ('clusters') along this prespecified variable which are in addition well interpretable in terms of similarity within other variables.

Strongly correlated variables can be interpreted to share some latent component which will be overemphasized in (non-Mahalanobis) distance calculations, if all variables are used for clustering. Nonetheless, as discussed in the paper, the use of highly correlated variables might be desirable from the expert's point of view. In any case variable correlations should be investigated in advance. This could be done by some preliminary variable clustering like it is for example implemented by the R function `corclust()` (Ligges *et al.*, 2012): using complete linkage of '$1 - |\text{corr}(X, Y)|$' results in clusters wherein all variables exceed some minimum degree of correlation. Expert knowledge helps to decide whether it is sufficient to select a single representative of each variable cluster for subsequent clustering.

Preliminary variable selection under both interpretation as well as correlation considerations strongly impacts the resulting clusters. An approach to expert-based automatic variable selection on sociodemographic data has been proposed by Roever and Szepannek (2005). Variable subsets are identified that produce the crispest clustering results while being restricted to contain representatives of several predefined well-interpretable variable subgroups.

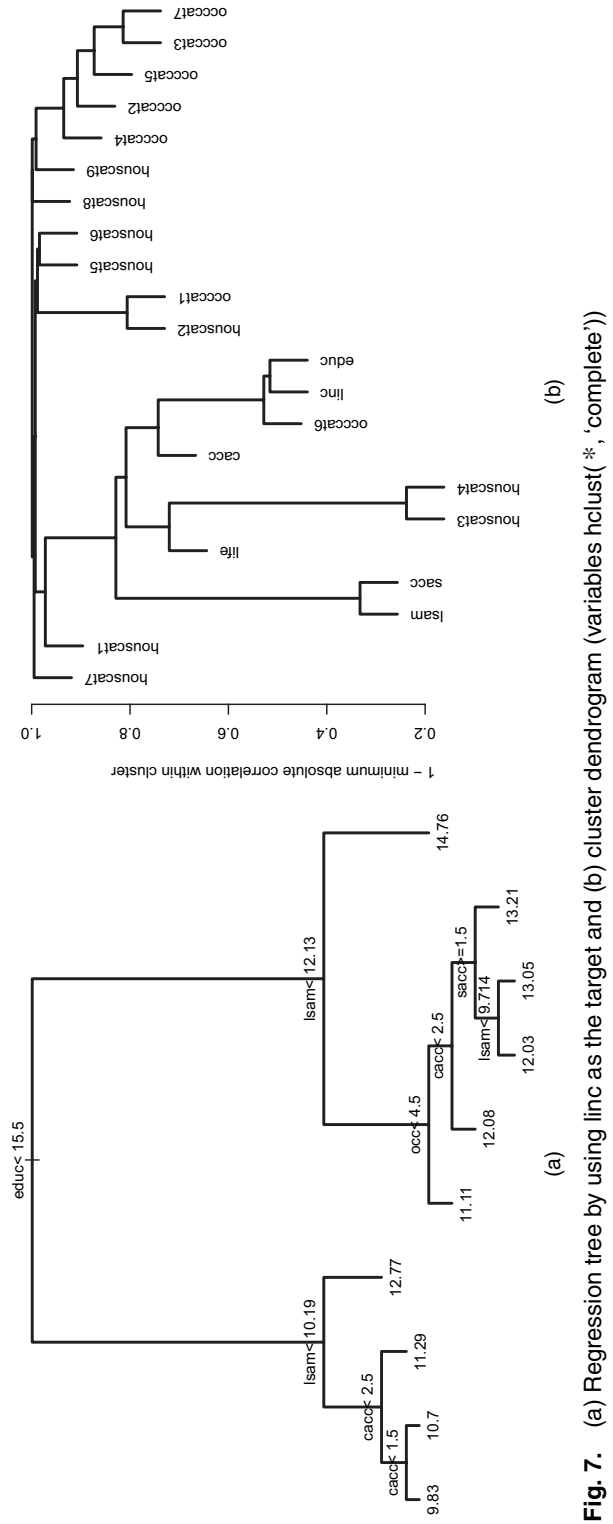

Fig. 7. (a) Regression tree by using linc as the target and (b) cluster dendrogram (variables hclust(*, 'complete'))

A bunch of preprocessing steps can be motivated although their exact parameterizations are arbitrary (e.g. the constant term of the transformed income). To investigate the sensitivity of the results to the choice of parameter values, the 'stress-tested' stability of the cluster results for slight variations in parameters should be taken into account.

Fig. 7 exemplifies results of a regression tree (using linc as target) and variable correlation clusters while the discrete nature of several transformed variables must be kept in mind.

Silvia Liverani (*Imperial College London and Medical Research Council Biostatistics Unit, Cambridge*) **and Sylvia Richardson** (*Medical Research Council Biostatistics Unit, Cambridge*)

We congratulate Hennig and Liao for a thorough clustering analysis that we hope will inspire the community at large to consider carefully the assumptions and modelling choices that they make when partitioning a date set.

They contrast their subject matter strategy with a model-based latent clustering approach, which relies on expectation–maximization for estimation and the Bayesian information criterion for choosing the number of components. Although we agree that it is important to consider specific features of each clustering problem to gain interpretability, we feel that their implementation relies on series of choices and tuning that could be considered arbitrary, and that make replication difficult. These issues can be addressed by using flexible model-based methods, such as Dirichlet process Bayesian clustering (Molitor *et al.*, 2010), where we sample from a mixture model with a variable number of clusters and can choose priors on the mixture components to reflect our beliefs. The rich posterior output that is produced can be used to learn about the partition space and its uncertainty, overcoming some of the difficulty that is faced by EM algorithms when trapped in local modes.

However, we agree with the authors that, for many applications, it is useful to show a representative partition, as an effective way to convey the output of the clustering algorithm. To characterize a partition that is most supported by the data, Dirichlet process clustering requires post-processing of its rich output. Independently of any labelling, at each iteration of the sample, we can record pairwise cluster membership and construct an $n \times n$ score matrix, with entries equal to 1 for pairs belonging to the same cluster and 0 otherwise. Averaging these matrices leads to a probability matrix S, which can be then used to identify an optimal partition. Like the authors, we propose to use the deterministic clustering procedure partitioning around medoids on the dissimilarity matrix $1-S$ and the average silhouette width to choose a representative partition C (Molitor *et al.*, 2010).

Our constructed dissimilarity matrix is based on the output of the Bayesian model, rather than Euclidean distance. It requires less tuning, avoiding choices of appropriate standardization. It allows users to specify their beliefs and context-specific knowledge in the formulation of distributional assumptions for the components and their priors. Another advantage is that the uncertainty of any parameter linked to individuals in the clusters of C can be evaluated by post-processing the Bayesian output.

Gilles Celeux (*Inria Saclay-Île-de-France*)

I congratulate Christian and Tim for their illuminating paper. I agree with the clustering philosophy that they give and from this point of view I have two remarks.

(a) I strongly agree with the opinion that the Bayesian information criterion BIC is ill posed to select the number of clusters in the model-based clustering framework and I would like to advocate the use of the integrated complete-likelihood criterion ICL of Biernacki *et al.* (2000) in place of BIC to select a useful mixture model in a clustering purpose. First, contrary to what is suggested in Section 6.4, ICL is not restricted to Gaussian mixtures. It is quite a general criterion and it could be preferred to BIC for choosing a relevant number of clusters for mixed data as well. Roughly speaking ICL = BIC + 'the entropy of the fuzzy clustering matrix'. Therefore, under the model assumptions of the mixture at hand, ICL evaluates the ability of the model to provide a meaningful partition of the (mixed) data, in addition to the goodness-of-fit aim of BIC. For this very reason, in the model-based clustering context, ICL could be expected to be more efficient than BIC under the clustering philosophy that is described in Section 5 of the paper. In particular, in many practical applications ICL does not improve monotonically as often happens with BIC (see Section 7.1). And, incidentally, it may be noted that, for latent class clustering with qualitative data, ICL does not require an asymptotic approximation and has an exact expression (Biernacki *et al.*, 2000).

(b) The undesirable occurrence of empty classes that is mentioned at the end of Section 7.1 can be avoided by imposing equal mixing proportions in latent class clustering. This constraint allows us

to avoid spurious solutions and is not too stringent in general. It is a simple way to penalize very small clusters. For this reason, it could lead to sensible clustering of social stratification data since very small classes are not of much interest in this context.

Jon R. Kettenring (*Drew University, Madison*)

The paper by Hennig and Liao presents compelling arguments for tailoring clustering methodology as tightly as possible to the application. Decisions about how to represent the data, select the appropriate algorithm, validate the results and present the conclusions should be driven by the problem at hand.

Clusters are often conceived of as point clouds that are compact (in some sense) and distinct. (Other variations allow for unusual shapes and overlapping clusters.) Fig. 1 provides an interesting example. Are there three clusters or eight, as the authors would prefer for social stratification?

The answer should be three if the goal is to find *statistical* clusters. Eight might make more sense if the goal is to find *convenience* clusters which segment the thin elliptical set of points in Fig. 1 into several cohesive contiguous groups. Hand *et al.* (2001), page 93, nicely elaborate on this distinction but then skim over it as do Hennig and Liao. With appropriate care clustering methods can be effective for both purposes.

Mahalanobis distance 'based on some covariance matrix estimator S' is described as one alternative for measuring dissimilarity involving the continuous variables. The often-considered approach of using the total covariance matrix—call it T—for this can yield very misleading results, as many researchers have noted. The authors reject this approach by using correlation arguments that only make sense if there are no clusters in the data.

This concern is closely related to the awkwardness of standardizing continuous variables to unit variance by using the square roots of the diagonal elements of T. Although widely practised, this approach can butcher cluster structure. A better approach would be to account for (as yet unknown) within-cluster variability, but how? This is a conundrum.

The authors briefly mention hierarchical methods and visualization. Whereas a full hierarchical solution may be of little interest, a partial solution that reveals major classes and subclasses could be very illuminating. Since the dimensionality of the problem is moderate, more attention to visual displays of the data, such as projection plots to reveal cluster structure, could help to validate and summarize results.

The authors make numerous thoughtful choices and compromises in the implementation of their clustering philosophy. One can quibble about the details but what really matters is whether their approach helps to resolve some of the controversy surrounding socio-economic stratification. That, of course, remains to be seen.

The following contributions were received in writing after the meeting.

Amir Ahmad and Sarosh Hashhmi (*King Abdulaziz University, Rabigh*)

We congratulate Hennig and Liao for this interesting paper. The paper provides a detailed study of the clustering results of two mixed data clustering algorithms on economic data from the 2007 US Survey of Consumer Finances. The experiments are extensive and the discussion provides a good insight into the problem.

The authors used two mixed data clustering algorithms in the paper. However, some of the recent mixed data clustering algorithms were missing in the literature survey, e.g. Huang *et al.* (2005) and Plant (2011). These algorithms could have been included in the comparative study. In the conclusion part, the authors could provide guidelines for new research work. A future work section could be included.

Grigory Alexandrovich and Hajo Holzmann (*Philipps-Universität Marburg*)

We congratulate Hennig and Liao for their elaborate application of cluster analysis in the context of socio-economic stratification. Their discussion of subject-specific issues as well as their handling of the rather complex data set are to be commended.

The concept of a cluster cannot, as the authors quite correctly emphasize, be merely defined by statistical criteria. Rather, specific information on what a cluster should look like in the application at hand is required. At this stage, however, care is certainly needed, so that the huge flexibility in performing cluster analysis is not merely used to confirm 'quantitatively' an *a priori* desired result such as a given number of social strata.

Methodologically, we feel that the latent class method is not fairly treated compared with the distance-based method. Indeed, by their very definition distance-based methods should perform better in terms of distance-based clustering evaluation criteria like the average silhouette width or the Calinski–Harabasz index than a latent class method.

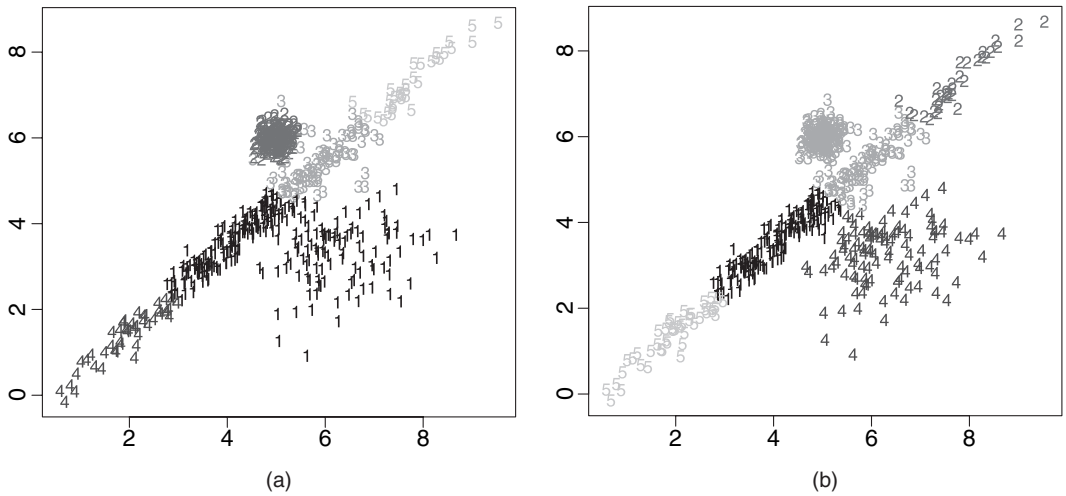

Fig. 8. (a) Latent class clustering with five components, average silhouette width 0.375, and (b) latent class fit with seven components, merged according to the average silhouette width criterion to five clusters, average silhouette width 0.519: the *k*-medoid method with five clusters (not shown) has average silhouette width 0.497

When using the Bayesian information criterion for choosing the number of components, the latent class method yields an appropriate estimate of the density of the data. However, if the underlying density cannot be well approximated by a finite mixture in the given family of densities, then the number of components can be quite large, and the resulting clustering is not meaningful. But simply—drastically—reducing the number of components is no solution either, since then the latent class method does not yield an appropriate density estimate.

Rather, we recommend merging algorithms, as proposed in Baudry *et al.* (2010), which in a top-down procedure successively merge components which together form a single cluster in an *a posteriori* analysis. Baudry *et al.* (2010) used an entropy-based criterion to select which components are to be merged in each step. Other choices are possible as well, such as merging those components so that the average silhouette width in the resulting coarser clustering is most increased. Indeed the choice of the merging criterion will (in addition to the form of the component densities) influence the final cluster shapes and can be appropriately tuned to the application at hand. One advantage compared with a distance-based method is that *a posteriori* probabilities for the final clustering are still available. Fig. 8 illustrates this procedure on a simulated data set.

Johann Bacher (*Johannes Kepler Universität, Linz*)

Hennig and Liao have discussed an important methodological and substantive question. We have analysed similar problems over the last 15 years. Our data differ with regard to the number of cases, the number of variables and, in the simulation studies, by the underlying class structure. Our experiences can be summarized as follows (Bacher, 2000; Bacher *et al.*, 2004).

(a) In accordance with Hennig and Liao, our results indicate that more than just the three classical variables (education, occupation and income) are required to detect more complex class structures.
(b) Among the software tested, LatentGold performed well. In the simulation studies, LatentGold could also detect the correct number of clusters in many situations.
(c) In contrast with Hennig and Liao, we found that it is possible to test the null model of no class structure. Information criteria and goodness-of-fit measures performed well. The question of whether a class structure exists is of theoretical and practical importance. A class structure in a society implies borders between the classes that are difficult to pass, whereas the absence of class implies that people can change their positions at any point.
(d) The absence of a class structure suggests the application of a factor analytic model. For example, LatentGold enables the use of a latent factor and a latent cluster model. The results can be compared by information measures or other fit criteria.

(e) The problems of cluster validity should be defined in a broader sense. Although stability is an important aspect, other aspects are also important. One important aspect is substantive or criteria validity that implies correlation with external criteria, e.g. in the case of how social classes correlate with voting behaviour or health behaviour.

Hennig and Liao have compared a general distance measure with a latent cluster model for different measurement levels. A third approach is to quantify the ordinal and nominal variables by using multiple correspondence analysis in a first step and to use a cluster model for quantitative variables in the second step.

Fadoua Balabdaoui (*Université Paris-Dauphine, Paris*)

I congratulate Hennig and Liao for their very thorough way of approaching the problem of clustering. Several important issues are raised by the authors, but I shall focus on only two of them.

(a) Choosing Gaussian distributions to model the mixture components is not suitable for heavier-tailed distributions. This could be rectified by replacing the Gaussian assumption by log-concavity; see for example Dharmadhikari and Joag-Dev (1988). Fitting a mixture of log-concave distributions with the objective of clustering has been already studied in Chang and Walther (2007). See also Cule *et al.* (2010) for a medical application.

(b) The problem of selecting the number of clusters is at the heart of clustering analysis. In recent work, Baudry *et al.* (2010) suggested a solution to overestimation of the Bayesian information criterion BIC and underestimation of the integrated classification likelihood (see Biernacki *et al.* (2000)) of the number of clusters: two clusters are merged if the merging yields a smaller entropy. Another criterion using entropy is the normalized entropy criterion NEC that was proposed by Celeux and Soromenho (1996). We have investigated its performance and that of the Akaike information criterion AIC and BIC for large sample sizes when the Gaussian assumption is replaced by log-concavity. Samples of sizes $n = 1000, 5000, 10000$ were generated from a mixture of $k = 4$ Gaussian distributions with means 0, 4, 8 and 11 and variances equal to 1. AIC and BIC are defined here respectively as

$$\mathrm{AIC}(k) = L_n(k) - \left\{k - 1 + 2\sum_{j=1}^{k}(N_{j,n} + 1)\right\},$$

and

$$\mathrm{BIC}(k) = L_n(k) - \frac{1}{2}\log(n)\left\{k - 1 + 2\sum_{j=1}^{k}(N_{j,n} + 1)\right\}$$

where $N_{j,n}$ is the number of knots of the log-concave maximum likelihood estimator of the jth component. Our penalization term can be explained by the fact that the logarithm of the maximum likelihood estimator is piecewise linear. On the basis of 100 replications we computed the proportion of times that the criteria considered selected the correct number of clusters. BIC had a very poor performance and hence the percentages obtained are not reported. The estimated proportions for AIC and NEC are given in Table 3. AIC seems to be doing quite well for very large sample

Table 3. Estimated proportions of correct selection of the true number of clusters in a mixture of four univariate Gaussian distributions by using AIC and NEC

n	*AIC (%)*	*NEC (%)*
1000	11	36
5000	75	51
10000	86	56

sizes. Further computations need to be done to build a clearer idea about how these criteria really perform, particularly in high dimensions.

The R programs that were used to produce the values in Table 3 are available from `http://www.ceremade.dauphine.fr/~fadoua/`.

Jean-Patrick Baudry (*Université Pierre et Marie Curie, Paris*)
I read this work with much interest because of the 'clustering philosophy' that the authors claim. This is obviously not an easy task and, from this point of view, the paper is stimulating reading.

I would be interested to know what the authors think about the following point, which concerns actually quite a general methodological choice: what question can we address from a clustering study? Indeed, the authors state that 'the occupation grouping has almost nothing to do with the clusterings on the other variables' (page 331), from which they deduce that 'the data-based strata are not as strongly connected to occupation categories as is often assumed in the literature' (in the summary). However, it is possible to derive a reasonable classification rule of the occupation with respect to the other variables (I did not have to put in much effort to obtain a 9.7% out-of-bag classification error rate with a random forest (Breiman, 2001a).) (I ran the `randomForest` R package with mtry $= 3$, which is the optimal choice according both to out-of-bag and to test set validation, and ntree $= 2000$, which is enough according to the same validation approaches. I am happy to provide the R code that I used.) Some of the occupation groups can be particularly well predicted (the group 'Manual workers and operators' receives a 2.5% out-of-bag classification error rate.) Then there is a clustering of the data which corresponds quite well to the occupation groups (the adjusted Rand index of predicted classes *versus* occupation groups is 0.88). Consequently should not the conclusion to the authors' work be that our clustering methods cannot recover this clustering? Does not the authors' analysis highlight the difficulty of the clustering task rather than answer the subject matter question ('are data-based strata connected to occupation categories?')? Of course, to obtain an unsupervised clustering which would reflect the occupation classes is much more difficult—and a completely different task—than to build a (supervised) classification rule based on the knowledge of the occupation classes and I do not claim that it is easy or even possible to obtain better results than the authors do. The point is whether or not this clustering-based approach is suitable to address the subject matter question raised.

Charles Bouveyron (*Université Paris 1 Panthéon-Sorbonne*)
First, I greatly thank the authors for this very interesting and painstaking work. I found this paper made with real care. It will certainly be very useful and instructive for economists. Such work should be in fact done for all application fields of clustering.

Regarding the clustering philosophy, which is largely discussed in the paper, I agree with the authors on the fact that clustering should be used only in conjunction with expert supervision. Indeed, it is important to keep in mind that clustering is an explanatory tool which is powerful only through the eyes of an expert in the application field.

The type of data that are used in the paper is both very interesting and representative of the actual data used in economics (and many other fields such as digital humanities or medicine). Indeed, mixed data (mixtures of qualitative and categorical data) are frequently available in applications. It should be noted that, unfortunately, the number of clustering techniques that can directly deal with this kind of data is very limited. A few recent works, such as Bouveron *et al.* (2012) and Hunt and Jorgensen (2003), consider alternatives to deal with mixed data but further effort should be made in this research direction. In particular, clustering techniques should not be limited to mixtures of quantitative and categorical data and other types of data should be explored such as networks, functions or even texts.

Another key point of the paper is of course the problem of clustering evaluation. This is an important problem in the clustering community and the limits of clustering evaluation are well highlighted in the paper. I would, however, suggest trying to evaluate the economic partitions that are obtained in the experiment with a supervised point of view. Indeed, occupation categories are often used in economics within supervised methods (regression or classification) in addition to the classical economic analysis. It may therefore be possible to evaluate the partitions obtained with the light of a supervised objective function within a cross-validation procedure.

Miguel de Carvalho (*Pontificia Universidad Católica de Chile, Santiago, and Universidade Nova de Lisboa*) **and Garritt L. Page** (*Pontificia Universidad Católica de Chile, Santiago*)
Interpretation is an important step in our investigations, and we often see it as *the* ultimate step of a data analysis (Cox and Donnelly (2011), section 1.2).

We focus on discussing a simple set-up related to the appearance of 'spurious' clusters due to (inadequate) data preprocessing, as in Fig. 3(c), illustrated by using simulated data. We suppose that there is a latent variable Z with distribution function

$$F_Z(\cdot) = \sum_{k=1}^{K} \pi_k F(\cdot; \theta_k), \tag{1}$$

whose mixture components define the 'meaningful' K clusters the researcher expects to see. The challenge is on using the data $\{X_i\}_{i=1}^n \sim F_X$ to learn about Z. Here $\pi_k \in (0, 1)$, $\Sigma_{k=1}^K \pi_k = 1$, and $\{F(\cdot; \theta) : \theta \in \Theta\}$ denotes a parametric family indexed on a parameter space Θ; more complex sampling schemes could have been used for Z (e.g. Booth *et al.* (2008), equation (2)), but equation (1) suffices for our purposes. We assume that the dependence between X and Z is described through an unknown copula function $C\{F_X(u), F_Z(v)\} = F_{X,Z}(u, v)$, for $(u, v) \in [0, 1]^2$, where $F_{X,Z}$ denotes the joint distribution function. In practice Z cannot be directly measured and therefore X (which is typically highly correlated with Z) is used as a proxy. However, we often forget that X may not be as informative about Z as one might hope, and preprocessing is used to tilt the distribution of X suitably so that it becomes more similar to that of Z.

In Section 6.1 the authors provide scientifically relevant arguments why the zero savings group of Fig. 3(c) fails to be meaningful, and thus motivating the need to employ a somewhat arbitrary $c = 50$. Additionally, a naive application of a pattern recognition technique could lead to spurious clustering—a pattern on X without any correspondent on Z. To illustrate the appearance of such spurious clusters in our set-up, consider Fig. 9 which displays 100 points simulated according to a Gumbel copula $C_\psi(p, q) = \exp(-[\{-\log(p)\}^\psi + \{-\log(q)\}^\psi]^{1/\psi})$, for $(p, q) \in [0, 1]^2$, with $\psi = 3$. The marginal for Z is a standard normal, and the marginal for X is a mixture of $N(1, 1)$ and $N(6, 1)(\pi_1 = \pi_2 = \frac{1}{2})$. This example is artificial —as in practice only $\{X_i\}_{i=1}^n$ would be observed—but it is interesting to observe that a spurious cluster on X may exist, even when Z is strongly correlated with X (Pearson correlation 0.79).

From a modelling point of view, the paper clearly puts forward the key role that subject-specific interpretations play in helping to link X to Z. Since the authors strongly advocate incorporating researcher intuition in clustering (about which we agree), should the Bayesian paradigm play a more active role in the proposed 'clustering philosophy'? In particular, product partition models have been recently devised for assessing uncertainty about the configuration of the clusters (Müller *et al.*, 2008). These methods can incorporate uncertainty associated with *a priori* 'expected' data partitions via a prior distribution assigned to the cluster configuration. The Bayesian approach is also natural for a less debatable choice of c in the preprocessing stage, or for the specification of a prior distribution on the structure of dependence between X and Z.

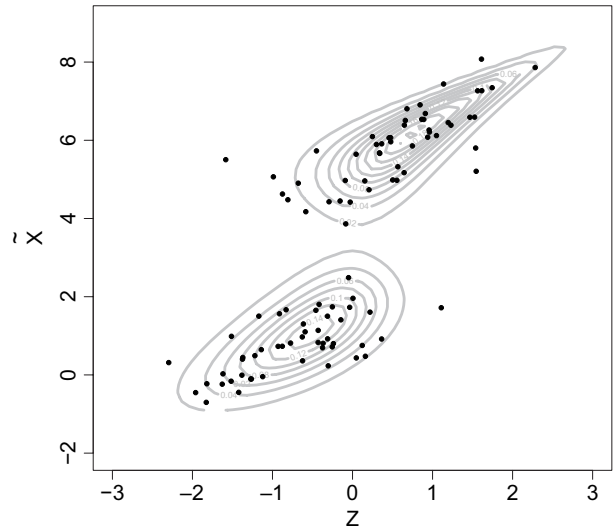

Fig. 9. Data generated from a Gumbel copula: the marginal for Z is a standard normal, and the marginal for X is a mixture of $N(1,1)$ and $N(6,1)$ ($\pi_1 = \pi_2 = \frac{1}{2}$)

Elvan Ceyhan (*Koç University, Istanbul*)

Hennig and Liao are to be congratulated for this nice contribution to the interface of statistical and sociological methodology. They employ clustering methods for mixed-type data with continuous, ordinal and nominal variables and illustrate their methodology on a real life socio-economic data set.

In clustering, data sets are divided into subsets (called *clusters*), where objects in the same cluster should be more 'similar' to each other compared with objects in other clusters (Kriegel *et al.*, 2009). Since labelling of the objects to groups is performed without prior knowledge of the class labels, clustering is also called 'unsupervised classification'. Usually, the purpose of clustering is summarization or improved understanding of the given data. However, social stratification refers to grouping a population into several different *social classes*. Hence, clustering methods may prove to be very valuable for this purpose as the authors demonstrate. Usually, in practice, clusters may not be well separated from each other. However, most clustering methods try to find an algorithm to group objects into disjoint clusters (Steinbach *et al.*, 2003). But, for socio-economic data, it is unrealistic to expect well-separated clusters, so 'fuzzy clustering' (Hoppner *et al.*, 1999), which allows partial membership of an object to several clusters, might prove to be more useful in this context.

Steinbach *et al.* (2003) provided some working definitions of a cluster, namely

(a) well-separated cluster,
(b) centre-based cluster,
(c) contiguous cluster (nearest neighbour clustering) and
(d) density-based cluster

definitions. The methods employed by the authors in the paper are latent class clustering and k-medoid clustering. The latent class clustering method yields type (d) clusters above, whereas k-medoid clustering yields type (b) clusters. For better visualization, an algorithm called WaveCluster proposed in Sheikholeslami *et al.* (1998) transforms original data into a two-dimensional grey scale image, where the clustering becomes just image segmentation.

In Section 7.2 the parametric bootstrap is employed to test the clustering method. However, the parametric bootstrap here assumes normality of the data; hence the separation of the average silhouette width of the original data in Fig. 6 might be a by-product of simulating all bootstrap samples from the same Gaussian distribution, which might not be fitting sufficiently well to the original data. Hence a non-parametric bootstrap by case resampling (Davison and Hinkley, 2007) might be more appropriate. Here B bootstrap samples from the original data would be obtained by sampling with replacement and average silhouette width values would be calculated for each sample as well as the original data and plotted together as in Fig. 6.

Charalampos Chanialidis, Peter Craigmile, Vinny Davies, Nema Dean, Ludger Evers, Maurizio Filiippone, Mayetri Gupta, Surajit Ray and Simon Rogers (*University of Glasgow*)

We thank Hennig and Liao for a thought-provoking paper. Their main argument that clustering data well depends on a large number of (possibly subjective) decisions, rather than a fixed recipe from the statistical toolbox, is well taken. However, they take this contention one step further by emphasizing the importance of incorporating substantive expert knowledge in the clustering, almost pushing towards making the analysis inferential rather than exploratory. The implications of this requirement outside the field of social studies is less clear. In many modern applications such as in genomics or in medical imaging, there may be less substantive or reliable expert knowledge that can be applied to the clustering problem.

Some of the decisions restricting the latent class clustering (LCC) model, made to allow marginal inference (i.e. forced conditional independence), have consequences that are not mentioned in this paper. For example, the restriction of the LCC model to diagonal covariance matrices means that the results will no longer be invariant under linear transformations of the variables. In this regard, we feel that the mixture of marginal Gaussian models produces a poor representation of the data in Fig. 1. For their primary analysis it would be of great interest to see whether, when run without restrictions, the unrestricted LCC model results supported the results of the authors' simplified model.

In general, perhaps a more pragmatic approach would be to emphasize the exploratory nature of cluster analysis. In such an approach, a number of different methods are run, and the differences in the results can be explained to the substantive experts through the differing underlying shape and cluster characteristics that are implied by each method. This might be a more achievable goal than eliciting non-contradictory, substantive shape, separation or size group information from the experts in each case.

A final issue is the overarching emphasis on the Gaussianity of the groups. Transformations can only go so far in terms of adjusting for departures from this assumption. Similarly, no alternative to the con-

ditional independence assumption on the categorical variables (where the intution is even more limited) is suggested for the LCC model in this work.

D. S. Coad and H. Maruri-Aguilar (*Queen Mary University of London*)
We congratulate Hennig and Liao on this thought-provoking paper, which provides a detailed treatment of how to apply cluster analysis to mixed-type data with continuous, ordinal and nominal variables. The approaches considered are latent class clustering and dissimilarity-based clustering. Their salient properties are highlighted in the context of economic data. We feel that the methodology proposed will provide a valuable template for analysing similar data sets.

In the latent class clustering model, a multivariate Gaussian density is assumed for the continuous variables. It is noted in Section 6.3 that a density with heavier tails could be of interest and that no software is currently available which incorporates such a possible extension. A natural question is whether analogous methodology has been developed for a multivariate t-distribution. Presumably, such an approach has the potential of robustness, as for continuous multivariate data (Peel and McLachlan, 2000). It would also be interesting to know whether an asymptotic test is possible for the number of clusters in the mixture. Since the standard regularity conditions do not hold under the appropriate null hypothesis, we would expect alternative methods, which obviate these difficulties, to have been considered.

Topological data analysis (Carlsson, 2009) extends techniques like agglomerative clustering, thus allowing the study of topological data features such as components and cycles. This information is usually summarized in a 'barcode', in which the horizontal axis represents the distance r at which clusters of data points are formed. The bars are grouped along the vertical axis by homology groups H_k. The first group, H_0, describes the evolution of isolated components, and so this part of the barcode is equivalent to a nearest neighbour dendrogram. The next group, H_1, describes the evolution of bidimensional cycles and higher groups correspond to higher dimensional cycles. Longer bars refer to representative features of data, whereas short bars are interpreted as short-term features or noise.

Fig. 10 depicts a barcode built with a random sample of 40 points from the US Survey of Consumer Finances data. When r is about 1.6, the data exhibit four clusters together with two or three cycles. Still in its infancy, topological data analysis results not only inform the user about data features but may also suggest suitable probabilistic models for the data.

Pietro Coretto (*Universitá degli Studi di Salerno, Fisciano*)
This paper makes an interesting important contribution because it discusses original foundational arguments in the context of cluster analysis. I find particularly inspiring the stress on model assumptions and cluster validation. The idea that model assumptions must be connected with cluster shapes is interesting, but it has its drawbacks. In fact, it is not always simple to elicit intuitions about shapes in high dimensional spaces. Since a complete elicitation of model assumptions is difficult in practice, it is crucial to find a

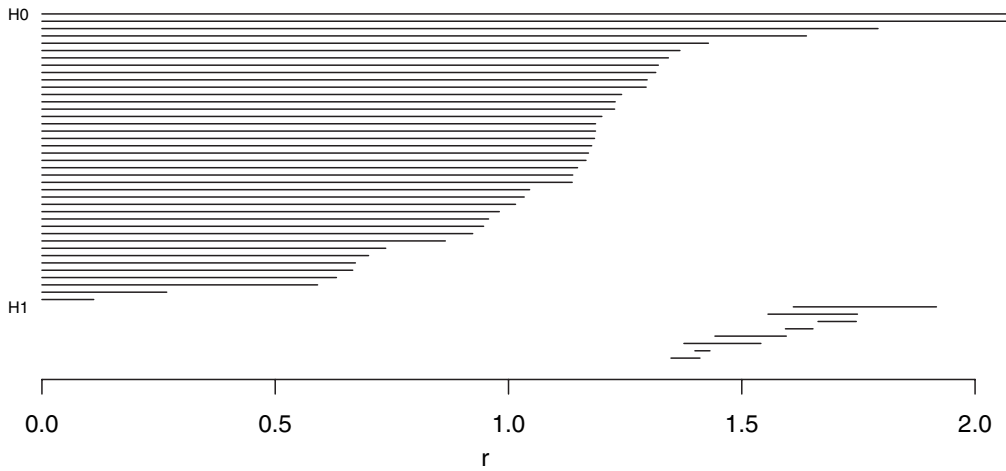

Fig. 10. Barcode for a random sample from the US Survey of Consumer Finances

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

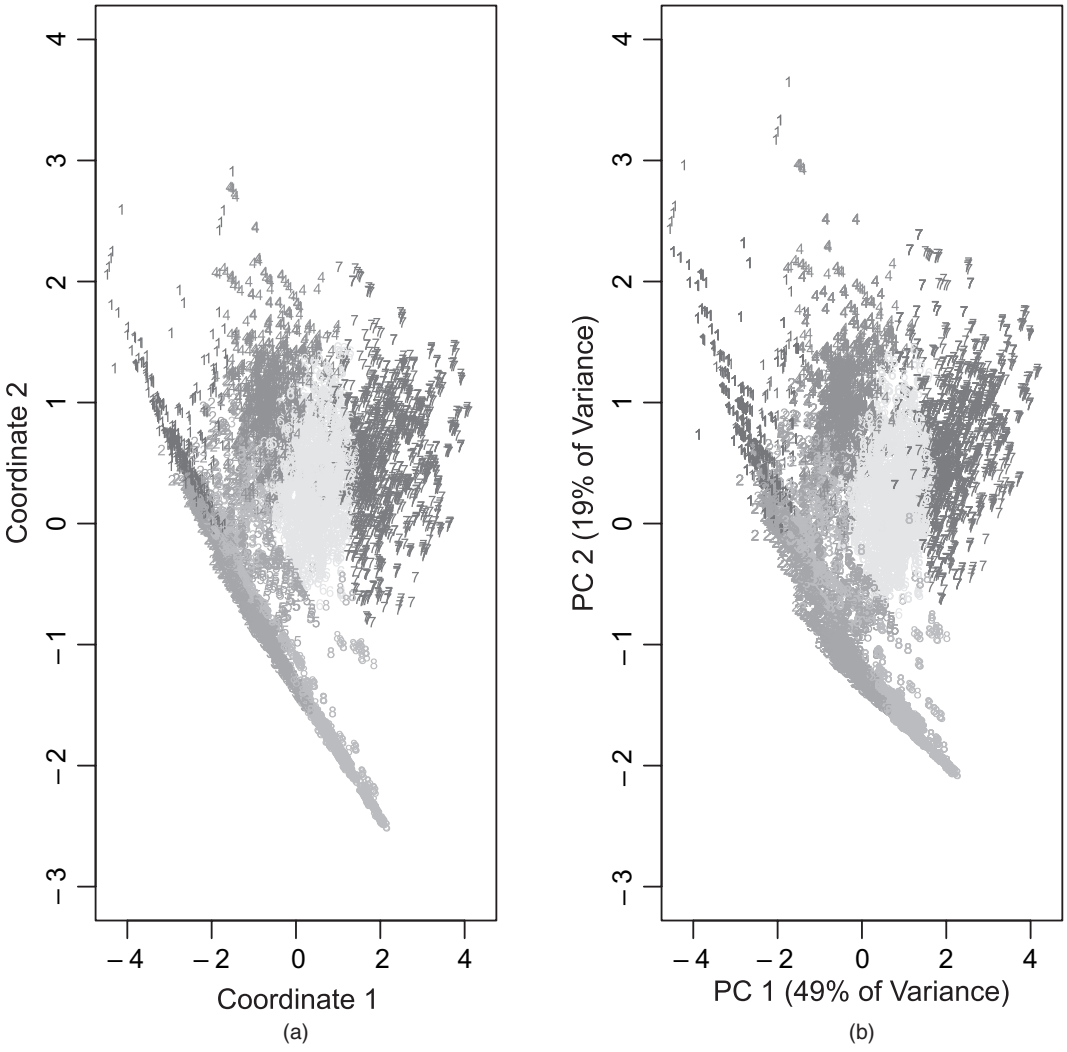

Fig. 11. (a) Multi-dimensional scaling and (b) principal components analysis

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.

To support this point, it is interesting that the two approaches can actually be very close. In particular, the dissimilarity-based clustering methods that characterize each cluster with a centre have strong links with model-based clustering methods. For instance, the k-means algorithm can be viewed as a classification EM algorithm applied to a specific Gaussian mixture model for continuous data (Celeux and Govaert, 1992) and information clustering criteria for discrete data are closely related to Bernoulli mixture models (Celeux and Govaert, 1991).

Bettina Grün and Gertraud Malsiner-Walli (*Johannes Kepler Universität, Linz*)

We thank Hennig and Liao for providing guidance on how to cluster data driven by the interpretation of subject matter researchers. In these comments we complement their work by providing insights for regularizing finite mixtures of multivariate normal distributions. We base our considerations on Vermunt and Magidson (2005), Fraley and Raftery (2007) and Frühwirth-Schnatter (2006). In essence they all use an inverse Wishart distribution as prior for the variance–covariance matrices Σ with density

$$p(\Sigma) \propto |\Sigma|^{-(\nu+p+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Psi\Sigma^{-1})\right\}.$$

The density is regular only if $\nu > p-1$. Vermunt and Magidson (2005) reparameterized this prior to have a prior sample size α and a variance–covariance matrix Λ for this prior sample, i.e.

$$\alpha = K(\nu + p + 1),$$

$$\Lambda = \frac{K}{\alpha}\Psi.$$

The default settings in Vermunt and Magidson (2005) with $\alpha = 1$ and $\Lambda = \mathrm{diag}\{\mathrm{var}(Y)\}$, the diagonal matrix of the empirical variance of the data, imply that

$$\nu = \frac{1}{K} - (p+1),$$

$$\Psi = \frac{\mathrm{diag}\{\mathrm{var}(Y)\}}{K}.$$

The default recommendation for the prior parameters in Fraley and Raftery (2007) is

$$\nu = p + 2,$$

$$\Psi = \frac{\mathrm{var}(Y)}{\sqrt[p]{K^2}}.$$

This choice of ν ensures that the prior mean exists. Taking the pth root implies that the same amount of the volume of the empirical variance of the data is covered regardless of the dimension.

Frühwirth-Schnatter (2006) suggested use of

$$\nu = p + 4,$$

$$\Psi = \phi(\nu)\,\mathrm{var}(Y),$$

where $\phi(\nu)$ is selected such that $1 - 2\phi/(\nu - p - 1)$ corresponds to the expected amount of heterogeneity explained by the differences in the component means, e.g. $\frac{1}{2}$. Her choice of ν aims at bounding the ratio of eigenvalues of the variance–covariances away from 0.

Table 4 gives the proposed default values for the various parameterizations. In addition $\tilde{\alpha}$ indicates the number of observations added to each component. In our opinion the $\tilde{\alpha}$–Λ parameterization is the easiest to interpret and to specify by

(a) fixing the prior sample size added to each component,
(b) deciding whether the empirical correlations are representative of the within-component correlations or whether these should *a priori* equal 0 and
(c) determining the fraction of the total variance assigned to the within-component variance–covariance matrix.

Table 4. Default values for the selected model in the paper with $K = 8$

Reference	ν	Ψ	α	Λ	$\tilde{\alpha}$
Vermunt and Magidson (2005)	−2.875	(1/8) diag{var(Y)}	1	diag{var(Y)}	1/8
Hennig and Liao	−2.375	(5/8) diag{var(Y)}	5	diag{var(Y)}	5/8
Fraley and Raftery (2007)	4	(1/8) var(Y)	56	1/56 var(Y)	7
Frühwirth-Schnatter (2006)	6	(3/4) var(Y)	72	1/12 var(Y)	9

Hennig and Liao add only five-eighths of an observation to each component regardless of their large sample size. Their Λ implies that *a priori* the within-component variances are assumed to be the same as the total variance. However, when intending the regularization to be effective, Frühwirth-Schnatter (2006) and Fraley and Raftery (2007) recommended to add *a priori* a considerably higher number of observations and to choose the within-cluster variance–covariance matrices to reflect the fact that the clustering aims at explaining some of the heterogeneity.

N. T. Longford (*SNTL and Universitat Pompeu Fabra, Barcelona*)

Whether a population is socially stratified or not is a matter of perspective, not of fact, because we do not have a process of arbitrating this issue by a deterministic exercise, such as the analysis of a hypothetical enumeration of the population. However, the perspective is constructive and can offer insights and understanding of the society that other less adventurous perspectives do not.

A conclusion of the paper is that in a particular setting there are eight clusters. I regard the suspension of uncertainty around this inferential statement rather unfortunate, because it reinforces the impression that in some analyses an indication of uncertainty, e.g. by estimated standard errors of the estimators, is *de rigueur*, whereas clustering is absolved of such a stricture. Suppose that there are only seven, or nine, clusters. How could we assess the magnitude of the inferential error made and what are its consequences on the validity of what is claimed to have been learned about the society from the analysis?

Within the perspective of stratification, I want to suggest that there is unlikely to be a single partition of the society that is in any meaningful way superior to some others, even within the constraints of a narrow range of aspects of the society. There are coarse partitions, to few strata, and refined partitions to many, and not all such partititions are hierarchically related. Without dismissing this hypothesis, the finding that there is a certain number of strata in a society is a possibly correct but nearly vacuous conclusion, because there may be an interpretable partition to almost any number of clusters, and what is found to fit best is not a reflection of the state of the society but an accident of the data and the balance of the variables in it.

Jorge Mateu (*University Jaume I, Castellón*), **Oscar O. Melo** (*National University of Colombia, Bogota*) **and Carlos Melo** (*District University Francisco Jose de Caldas, Bogota*)

The authors are to be congratulated on a valuable and thought-provoking contribution on the use of clustering methods for socio-economic stratification based on mixed-type data with continuous, ordinal and nominal variables.

According to Krugman (1996), the economy and society organize themselves in space. Therefore, racially communities are segregated, generating marginal and subcentres of urban areas. Then, it is of interest to consider the spatial relationship to define more appropriate clusters, and this can be pursued by using the Gower distance for a set of mixed variables (continuous, binary or multiattribute). We can then use the following expression of spatial similarity based on Gower (1968):

$$m_{ij} = \frac{\sum_{h=1}^{p_1} \left(1 - \dfrac{|x_h(\mathbf{s}_i) - x_h(\mathbf{s}_j)|}{R_h}\right) + a_{ij} + w_{ij}}{p_1 + p_2 - d_{ij} + p_3}, \qquad i, j = 1, \ldots, n,$$

where $x_h(\mathbf{s}_i)$ is the measurement of the hth continuous variable at the $\mathbf{s}_i$th localization, $\mathbf{s}_i \in \mathbb{R}^2$, R_h is the range of the hth continuous varible, p_1 is the number of continuous variables and for p_2 binary variables $a_{ij} = a(\mathbf{s}_i, \mathbf{s}_j)$ and $d_{ij} = d(\mathbf{s}_i, \mathbf{s}_j)$ are the number of positive and negative matches respectively associated with the relationship between the $\mathbf{s}_i$th and $\mathbf{s}_j$th locations. Finally, $w_{ij} = w(\mathbf{s}_i, \mathbf{s}_j)$ is the number of matches for p_3 multiattribute variables. Through the transformation

$$d_{ij} = \sqrt{(1 - m_{ij})}$$

it is possible to obtain Euclidean distances. Once this has been done, we can use the concept of geometric variability to build the clusters following the proposal of Irigoien and Arenas (2008) and Irigoien *et al.* (2008), who presented some methods that are applicable to any type of data, including mixed variables, and developed a new methodology to determine the number of clusters. In particular, let C be a set consisting of n objects, which are to be classified into k clusters C_r of size n_r respectively, $r = 1, 2, \ldots, k, k > 1$. Following their divisive criteria, for each value of $k > 1$, we can consider the ratio

$$G_k = \frac{\sum_{r=1}^{k} \sum_{r'=1}^{k} n_r n_{r'} \Delta^2(C_r, C_{r'})/2n(k-1)}{\sum_{r=1}^{k} n_r V(C_r)/(n-k)}$$

where Δ is a symmetric matrix $k \times k$, $\Delta_{rr'} = \frac{1}{2}\Delta^2(C_r, C_{r'})$, $r, r' = 1, \ldots, k$, i.e. each element of the array is half of the squared distance, which is calculated between all pairs of clusters $C_1, \ldots, C_k$, and $V(C_r)$ is the geometric variability. The value of k that maximizes this ratio provides an estimate of the number of clusters in the data. Note that this strategy of Irigoien and Arenas (2008) defining an automatic method is not Bayesian and could be a good alternative that works well with mixed variables with spatial data. It can be used in the problem presented in this paper because the different individuals do have a spatial localization, such as home, that is useful to obtain a socio-economic stratification.

G. J. McLachlan (*University of Queensland, St Lucia*)

Given the extensive use of finite mixture models in the clustering of data sets from a wide variety of fields, it is instructive to examine from time to time the basic assumptions underlying this approach to clustering. This paper serves a useful role in this respect.

It is focused on the case of mixed data where perhaps mixture models have not been applied to the same extent as with continuous data. However, various researchers have studied the use of mixture models for mixed data, concentrating on the location model, which formed the basis of the MULTIMIX procedure of Hunt and Jorgensen (1999); see, for example, chapter 5 of McLachlan and Peel (2000) where mixture models are considered for discrete and mixed data.

The authors state that 'an obvious problem with the model-based approach is that statisticians usually do not believe models to be true'. Firstly, I would contend that statisticians like working with models (see Breiman (2001b) with discussions from statisticians that included Professor Cox and Professor Efron). Secondly, it reminds me of the well-known saying that 'All models are wrong but some are useful' as attributed to the distinguished statistician Professor George Box. It is in the latter spirit that I have approached the use of mixture models for clustering. But, also, I consider that there are many situations where the component distributions in the mixture model being used to effect the clustering are relevant to describe the distribution of the data in the clusters so imposed, particularly in the biological, medical and physical sciences. A generic example concerns the situation where the data on the phenomenon under study can be modelled adequately by a single normal distribution but, when the system is perturbed, an additional normal component is needed. Although any distributional density for continuous data can be modelled with sufficiently high accuracy by a mixture of normal distributions as noted by Hennig and Liao, the key point in the aforementioned example is that it is the smallest number of normal components needed to model the distribution that is of interest.

Finally, the authors acknowledge that model assumptions are helpful for making explicit the cluster distributional assumptions being implicitly imposed with the use of so-called distribution-free methods of clustering. This is important since procedures such as k-means are often claimed to be model free whereas they are based on some particular assumption of the cluster distributions (normal distributions in equal proportions with common spherical component covariance matrices in the case of k-means).

Paul D. McNicholas and Ryan P. Browne (*University of Guelph*)

We congratulate Hennig and Liao on an interesting paper that touches on several important issues within the field of clustering. Nowadays, many clustering problems involve high dimensional data. Although we accept that finding out the 'kind of clusters' the relevant subject matter expert is after is desirable, our experience working with high dimensional data suggests that eliciting this information can be difficult. Do the authors have experience in this regard? The clustering philosophy propounded by the authors at the beginning of Section 5 'requires that the researchers define what kind of clusters they are looking for'. Are

we to take this to mean that clustering is not possible if a researcher does not know exactly what they are looking for? Must they know what they are looking for *a priori*? In our experience, subject matter experts sometimes take an 'I will know it when I see it' position, but seeing 'it' is not always possible.

The authors seem to take a pejorative view of the model-based approach to clustering. We are unsure what the authors are getting at in the fourth paragraph ('An obvious . . . ') of Section 5. What is this 'obvious problem'? What is meant by 'the mixture model is not precisely true'? McLachlan (2011) provides a cogent defence of model-based clustering, quoting, among other work, another paper read to the Society (Aitkin *et al.*, 1981). In their rejoinder, Aitkin *et al.* (1981) argued that 'when clustering samples from a population, no cluster method is *a priori* believable without a statistical model'. On the basis of their 'obvious problem' with model-based clustering, we presume that the authors disagree with the position taken by Aitkin *et al.* (1981).

Recent work on model-based clustering is increasingly based on mixture models that are not Gaussian, but this is not clear from the authors' presentation of model-based clustering. This point is perhaps most pertinent when they touch on merging mixture components (Section 6.4), citing work in this direction by Hennig (2010) as being a suitable 'amendment' when more than one Gaussian mixture component is needed to model a cluster. Although merging components can be effective, it seems remiss not to point out that merging components is not a 'get out of jail free card'; situations arise when the use of a non-Gaussian mixture model is more effective than a Gaussian mixture model followed by merging components (see Franczak *et al.* (2012) for examples with both real and simulated data).

Damien McParland and Isobel Claire Gormley (*University College Dublin*)

We congratulate Hennig and Liao on an interesting and thought-provoking paper. Many of the issues that are raised, particularly pertaining to the philosophy of clustering, should be given due consideration in any clustering analysis.

We have also been working on clustering mixed data with a socio-economic application (McParland *et al.*, 2012). We take a statistical modelling approach and so our model is more closely aligned with the latent class clustering model than the k-medoids approach. Our model considers categorical observations as nominal or ordinal manifestations of a latent continuous variable. The latent continuous data are then clustered by using a mixture of factor analysers model. Our model is estimated in the Bayesian paradigm and hence some of our clustering decisions are incorporated by using prior distributions.

In our work we found that choosing the number of clusters was a difficult problem, as it is here. We were wondering whether the authors considered any other model selection criteria for the latent class clustering model since the Bayesian information criterion does not appear to penalize sufficiently heavily for the addition of more mixture components. A regularized version of the Bayesian information criterion (Fraley and Raftery, 2007; Nyamundanda *et al.*, 2010), where a prior is applied and maximum *a posteriori* estimates are used rather than maximum likelihood estimates, may help in this regard.

We would be interested to hear whether any sensitivity analysis had been carried out to investigate the effects of reweighting variables, in particular, the choices of c, q and the arbitrary substantial reweighting of the hous, sacc and cacc variables. How would the final results change if different weights were assigned to these variables?

We found the parametric bootstrap test that was used by the authors to be a creative solution for investigating whether there is any clustering structure in the data. However, we wondered whether the authors assessed model fit. In this regard, we used a cross-validation approach to validate our model in McParland *et al.* (2012). This procedure removes entries at random from the data matrix, the model is fitted to the remaining data and the values of the missing observations are imputed. If the missing values are imputed 'correctly' a high proportion of the time, then the model is deemed to fit well. Would such an approach prove useful here?

The issues that are raised in this paper, particularly with regard to the clustering philosophy, will be considered in any further research we do in this area.

Marina Meilă (*University of Washington, Seattle*)

I congratulate the authors on carefully validating the 'model selection' via the average silhouette index. Their Fig. 6 and parametric bootstrap illustrate the unforeseen effects that such seemingly natural criteria like the silhouette suffer from. Their validation method is a variation of the idea of the *gap statistic* of Tibshirani *et al.* (2001).

Second, I offer a quantitative illustration to the phenomenon of Fig. 2, under the assumptions of latent class clustering and other simplifying assumptions. The paper suggests these separation criteria for choosing two clusters *versus* one: a large within-cluster distance (which could be quantified by the

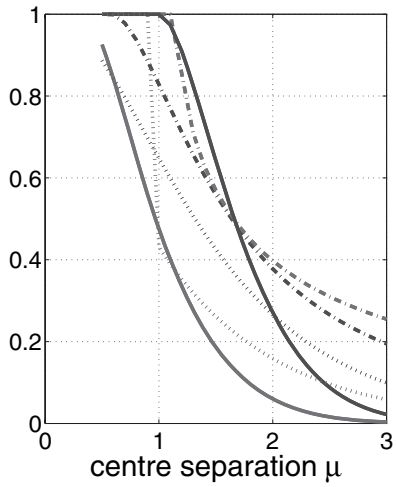

Fig. 12. Separation measures for two clusters in one dimension, $k = 2$, $\pi_1 = \pi_2 = 0.5$, $a_{1,2} = \pm\mu$ *versus* (half) the centre separation μ ($f_{\text{dip}} = f(0)/\max_x(f)$ and $p_{\text{u}} = \Pr\{0.5 - \delta \leqslant P(1|x) \leqslant 0.5 + \delta\}$ for $\delta = 0.38$): · – · –, Cauchy f_{dip}; · – · – ·, Cauchy p_{u}; ——, normal, f_{dip}; ——, normal p_{u}; · · · · · · , Laplace f_{dip}; · · · · · · , Laplace p_{u}

standard deviation of f) and bimodality. I add label confidence $P(h|x)$. Fig. 12 displays the last two criteria for three standard cluster distributions: Cauchy, normal and Laplace. Since the standard deviation grows almost linearly with the separation, it is not shown. Bimodality is measured by f_{dip} which is the height of the 'valley' relative to the peaks; p_{u} is the population proportion that is 'unsure' of its assignment. For a normal ϕ, $\mu = 1$ is the largest μ for which a mixture of two normal distributions is unimodal, and $P(h = 1|x = \mu = 1) = 0.88$, which is nicely close to a confidence of 0.9; therefore $0.5 + \delta = 0.88$ is chosen.

From Fig. 12 we can derive separation in multiple dimensions as well, with the added assumption that $\phi_{1,2}$ are identical in all dimensions except the first. Then all criteria depend on the unidimensional marginal along the first dimension.

Daniel Müllensiefen and Naoko Skiada (*Goldsmiths University of London*)

Hennig and Liao's paper undoubtedly represents a significant contribution in as much as they demonstrate how to derive guidelines for cluster analysis from *a priori* data considerations ('the cluster philosophy') as well as provide new empirical insights into the constituents of social stratification. As they point out, social stratification as a concept has been of immense practical value in the past, ranging from a predictor of health outcomes (Adler *et al.* (1994), Link and Phelan (1995) and National Center for Health Statistics (2011), page 4) to a factor influencing academic achievements (Coley, 2002; National Center for Education Statistics, 2008), intelligence (Tucker-Drob *et al.*, 2010; Turkheimer, 2003), and psychological and social behaviour in general (McLeod and Kessler, 1990), However, in most psychological contexts, social stratification has been conceptualized and implemented as an ordered variable, enabling correlation and regression applications or structural equation models. This implementation of social stratification as an ordered concept has been inherent in the idea of social class since its inception in the early 19th century (Kuper, 2004). However, a clustering treatment of variables relating to social strata results necessarily in the derivation of a categorical variable that is less practical for many psychological applications. In addition, we note that almost all variables from the 2007 US Survey of Consumer Finances data set bear a natural notion of 'low' and 'high' or 'less' and 'more' and, as such, the derivation of an aggregate variable that pertains to a sense or order seems a viable option. In terms of the approaches that have been suggested for deriving a scaled variable from a mixed-type variable data set we consider mixed-type factor analysis (Pages, 2004) that combines multiple correspondence analysis for categorical variables and principal component analysis for ordered variables in a single procedure. Similar to considerations given by Hennig and Liao, variables of different types are scaled and weighted with the aim of providing comparable contributions to the overall variance in the data. Alternatively, we also consider the LINEALS procedure and the underlying concept of bilinearization that were suggested by DeLeeuw (1988a, b) which can be used

to preprocess categorical variables with the aim of combining them with scale-type variables in structural equation modelling framework (DeLeeuw and Mair, 2009) to identify a latent common factor 'social class' eventually. Discovering agreement and disagreement between the clustering solution provided by Hennig and Liao and the scaled solutions suggested here would allow interesting insights into the relationship between the two conceptual perspectives of social stratification.

Oliver Nakoinz (*Christian-Albrechts-Universität zu Kiel*)
This paper touches on three important issues. The first is the analysis of social classes which concerns an interdisciplinary audience. The second issue is the handling of mixed scales which is momentous for many fields of analysis.

The most important part of this paper is concerned with cluster philosophy. The ideas behind the 'cluster philosophy' can be expanded to other methods. The basic idea is that the choice of methods and the exact parameters of these methods should be based on objective, theory and data. The demand to justify the choice of methods considering objective, theory and data is quite rare. This philosophy is barely found in practice but we should definitely support it. This philosophy ties together far more indigenous connections of theory and application and thus produces a considerable increase in the scientific significance and meaning of the results of quantitative analysis in research projects. Furthermore this philosophy implies the abandonment of a common view, which is the idea that different methods can validate each other. This certainly is not so. Different methods have a different function which answers different objectives. In general different methods must have different results which do not contradict but complement each other.

Coming back to cluster analysis we find a hot spot of methodological perplexity for users with moderate statistical skills from several disciplines. Hence we should welcome a decision tree for choosing the cluster method (an attempt was made for example in Nakoinz (2010)) or rather a handbook as a guide in practical research projects. This handbook would have to consider theories and objectives as well as statistical knowledge to give good guidance. Even such a guide would allow a wide range of individual decisions. For example one could discuss the choice of the k-medoid method in this paper. The authors claim that k-medoids are the right method because this method does not require mean values of nominal and ordinal variables. Instead of discussing data one could discuss theory and ask whether the cluster representative must be real elements or rather can be ideal types which need not correspond to real objects and need not have realistic values.

Finally I thank the authors for this valuable paper which can be presumed to have considerable influence on the practice of data analysis in empirical research projects in many disciplines.

Rebecca Nugent (*Carnegie Mellon University, Pittsburgh*) **and Abby Flynt** (*Bucknell University, Lewisburg*)
We enjoyed reading this paper and thank the authors for highlighting one of the common, but perhaps less discussed, philosophical problems in cluster analysis. In some ways, it is akin to 'Which came first?: the chicken or the egg?'. Should the clustering approach dictate the type of clusters? Or should the desired type of clusters dictate the approach?

We would argue that both sides have merit; however, from a practical standpoint, we would advocate a balance between the two. For this socio-economic stratification problem, Hennig and Liao do a thorough job determining the appropriate variable forms, weights, etc., all using a large amount of substantive expert knowledge. What would we do with less knowledge, or perhaps none at all? How do we define these transformations, weights and dissimilarities to ensure that we find stable, contextually meaningful clusters? This analysis strongly supports working closely on interdisciplinary teams with both statisticians and subject matter experts (which obviously has benefits). However, in practice, we wonder whether it is always feasible. Many methodologists are interested in developing clustering tools with specific statistical properties for use with a broad array of applications; many applied practitioners are simply looking for which software package to run on their data. Although we advocate conversations between the two groups, we have concerns about moving too far towards making decisions that guarantee the presence of preconceived types of clusters without leaving room for discovery of the unknown.

For this particular analysis, several transformations are very specific (e.g. $q = \frac{1}{2}$ for nominal variables) and weights are chosen 'according to substantial requirements'. There is little discussion about the sensitivity of the results to these choices. Could we reproduce essentially these same clusters by using slightly perturbed weights or parameter values? For categorical variables, what if some of these categories were empty or did not exist? How would this affect our solution? Are the clusters obtained still stable and meaningful?

With respect to fitting the latent class clustering, the authors also mention that other non-Gaussian distributional shapes might be of interest but that there are software limitations. Given the assumption that the (sets of) variables are locally independent within a mixture component, incorporating non-Gaussian or perhaps non-parametric kernel densities, at least for the continuous variables, may be a reasonable starting place.

Again, we thank the authors for their thought-provoking contribution and look forward to reading the subsequent and no doubt interesting discussion.

Akinori Okada (*Tama University, Tokyo*)

The paper describes how to deal with the data composed of mixed-type variables especially on socio-economic variables in revealing 'natural structure' of socio-economic stratification by cluster analysis of the data from the 2007 US Survey of Consumer Finances. The result that is obtained by two different clustering procedures is interesting and persuasive. In dealing with these variables, the paper examines transformation, standardization and defining dissimilarity including weighting given to the variables. The result tells us that transformation, standardization and defining dissimilarity including weighting are appropriate. But they seem to be decided *ad hoc* to deal with the data analysed in the paper. It is desirable to mention the possibility of generalizing these procedures of transformation, standardization and defining dissimilarity including weighting so that they can be used with other data for other purposes. Generalizing the present procedure will lead us to the optimal procedures of transformation, standardization and defining dissimilarity including weighting. Although the results obtained in the paper are persuasive, they do not assure that the procedures are optimal. As transformation, standardization and defining dissimilarity including weighting seem related to each other, it is desirable to mention these relationships. The present standardization pays attention to extreme observations or outliers. I think that some of the outliers can be eliminated in the present study to reveal the 'natural structure' of socio-economic stratification. Not all values or elements of the data need to be included in the analysis in the present study.

Two clustering results given by two different clustering methods are shown in the paper. One result was derived by k-medoids, and the other was derived by latent class clustering. As stated above, both of the results are persuasive. The result derived by the k-medoids method is characterized by the large clusters of middle class strata and smaller sized upper class strata, and the result derived by the latent class clustering method is characterized by large clusters of the large income strata. I would like to know why or what kind of aspects of the methods and/or the variables caused these differences between the two clustering results, which seem to be useful for transformation, standardization and defining dissimilarity as well. The comparisons of the composition of each of the eight clusters between two different clustering results, which show how the two clustering results agree, also seem interesting.

Andrea Pastore and Stefano F. Tonellato (*University Ca' Foscari, Venice*)

We congratulate Hennig and Liao for their interesting paper. One of the many important features addressed by this work is the determination of the number of clusters and we have a couple of comments about this.

(a) The indices used for the purpose (ASW, CH and PH; see Section 4) are based on the pairwise dissimilarity discussed in Section 6. Such indices are well suited to dissimilarity-based clustering but it is not clear how adequate they can be for latent class clustering. Latent class clustering, in fact, implicitly defines a dissimilarity, which determines the posterior individual membership probabilities $p_{ih} \propto \hat{\pi}_h \hat{f}_h(\mathbf{w}_i)$, which are strictly related to the model structure and parameter estimates. Dissimilarity-based criteria are in some sense biased towards partitions resulting from dissimilarity-based clustering. Perhaps this is one plausible explanation of the difference between the optimal number of groups identified by ASW, CH and PH on one side, and the number resulting from the use of the Bayesian information criterion on the other side. It might be interesting to consider a model-based dissimilarity measure, similar to that introduced by Menardi (2011) in the context of density-based clustering. This might also improve model interpretability, which is not so clear when the number of clusters is determined through model-independent dissimilarity measures.

(b) The values of ASW that were reported in Table 1 are all smaller than 0.25, except that associated with the optimal partition produced by k-medoids, which is equal to 0.275. Kaufman and Rousseeuw (1990), Table 4, page 88, suggested, admittedly 'on a rather subjective evaluation', that a value of ASW less then 0.25 means that 'no substantial structure has been found'. In the application in the paper, such low ASW-values might depend on the fact that the variables included in the

model, if considered separately, might induce rather different partitions. The resulting model-based clustering might be a compromise between such partitions, leading to relatively heterogeneous or poorly separated groups.

Gunter Ritter (*University of Passau*)

I first welcome Hennig and Liao's discussion of what clustering is about and what it can do. As a non-sociologist, I was particularly interested in their Section 5 on the 'philosophy of clustering' and their partition of the simulated data set sampled from two spherical and one elongated elliptical populations with strongly correlated normal variables shown in Fig. 1. Section 5 is about the decomposition of a data set and two related activities that come to mind are *clustering* and *quantization*. Which one is described here?

The authors reject for their purpose the idea of sampling from a parental population: assumption 2. They intend instead to unite nearby points and to separate distant points (the 'poorest' and the 'richest'). These are characteristics of *quantization*; see Graf and Luschgy (2000). Quantization is *geometric*, depends on an *a priori* given metric and is sensitive to variable rescaling. The standard algorithm is k-means, which is called Lloyd's (1982) algorithm in this context. Application of an algorithm based on a heteroscedastic, *spherical* normal mixture or classification model indeed returns the 'clusters' that are shown in Fig. 1. This phenomenon is well known in clustering and mixture analysis: the application of too narrow a model can create structure instead of revealing it; see Gordon (1999), chapter 6.

By contrast, most contributions to and applications of *clustering* are mainly concerned with detecting causes in data, for instance different genetic types of a disease in medicine or the components of data sampled from distinct normal distributions. It usually assumes independent observations coming from a mixture model, which are the basis for deeper asymptotic results on the consistent local maximum likelihood estimate. For clustering to serve its purpose, the mixture should consist of sufficiently separated components so that interiorly cohesive and exteriorly isolated clusters result; Cormack (1971). This is nicely demonstrated in Fig. 2. The assignment should be independent of units of measurement. The cluster analyst wishes to retrieve the three sources of data that were used to create Fig. 1 and this is indeed what an algorithm based on a heteroscedastic, *fully* normal mixture or classification model does. In quantization, even the normal distribution is decomposed, whereas the result should be a 'single source' in clustering.

All things considered, I agree with the referee who noted that it is quantization rather than clustering that Section 5 is about. Otherwise, the authors should have merged most of the subsets of the elongated cluster in Fig. 1 by using methods proposed in Hennig (2010).

Marc Saez (*University of Girona and Consortium for Biomedical Research Network in Epidemiology and Public Health, Girona*)

I read with interest the excellent paper of Hennig and Liao. I agree with them about the danger of applying clustering methods automatically and regardless of the direct interpretation of the implications of the methodology. However, I believe that they have not been able to exploit the opportunities of application that they present. I mean, in the description of the indicators of socio-economic strata that they do not take into account such matters as are the indicators that define social class the same for men and women and do the indicators behave identically for all age groups, in particular, for older individuals? I think not. Moreover, socio-economic indicators that they describe are all absolute. Why did they not consider indicators?; or rather why did they not relativize, considering, for example, contextual variables? It is known that individuals with similar socio-economic absolute indicators could occupy different positions on the social ladder if they lived in neighbourhoods with very discordant socio-economic characteristics. In short, I think that clustering methods are actually neutral, and what determines the final classification is the choice of the appropriate variables.

Milan Stehlík (*Johannes Kepler University, Linz*)

I congratulate Hennig and Liao for bringing up the challenging world of clustering under uncertainty! I agree completely with them that the application of automatic methods hoping that the data will enforce the true structure is deceptive. I would like to point out the existence of distributional sensitivity of dissimilarity distances. The authors mention only metric dissimilarities (e.g. Euclidean L^2, Gower–Manhattan and Mahalanobis distance).

Let us assume for simplicity that the underlying distributions form a regular exponential family (see Barndorff-Nielsen (1978)), with the sufficient statistics $t(y)$ for the canonical parameter $\gamma \in \Gamma = \{(\gamma_1, \ldots, \gamma_N), \gamma_i > 0; i = 1, \ldots, N\}$. The 'covering' property $\{t(y) : y \in Y\} \subseteq \{E_\gamma[t(y)] : \gamma \in \Gamma\}$ (see Pázman

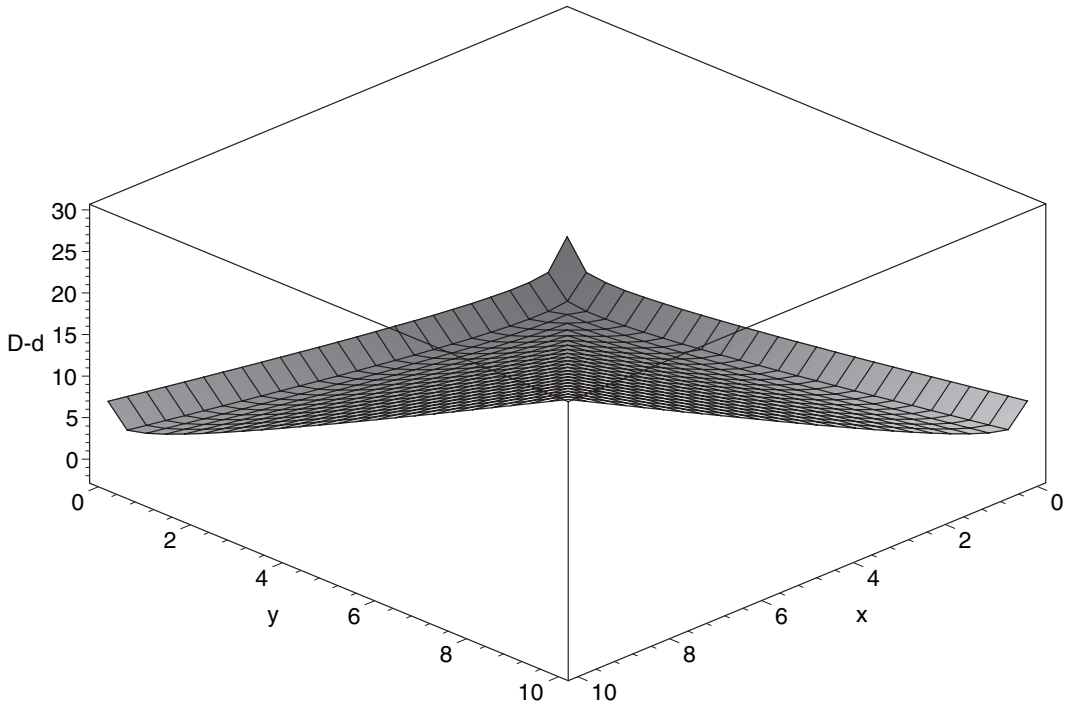

Fig. 13. $D(x,y) - d_E(x,y)$, $\gamma^* = 2$

(1993)) enables us to define the I-divergence of the observed vector y in the sense of Pázman (1993):

$$I_N(y, \gamma) := I(\hat{\gamma}_y, \gamma). \tag{2}$$

Here $I(\gamma^*, \gamma)$ is the Kullback–Leibler divergence between the parameters γ^* and γ. The distribution and other properties of expression (2) have been studied by Stehlík (2003) for the case of the gamma distribution. The I-divergence has nice geometrical properties, e.g. the Pythagorean relationship (Efron, 1978; Csiszar, 1975)

$$I(\bar{\gamma}, \gamma) = I(\bar{\gamma}, \gamma^*) + I(\gamma^*, \gamma) \tag{3}$$

for every $\gamma, \bar{\gamma}, \gamma^* \in \text{int}(\Gamma)$ such that $(E_{\bar{\gamma}}(t) - E_{\gamma^*}(t))^T(\gamma^* - \gamma) = 0$. Thus we can construct a dissimilarity measure based on equation (3) in a natural way by

$$D(y_1, y_2) = I(y_1, \gamma^*) + I(\gamma^*, y_2) \tag{4}$$

where γ^* is chosen appropriately.

Note that equation (4) is not a metric in general (because it is not symmetric for all distributions), but in the case of normal distributions it is L^2-distance. Since the data that Hennig and Liao used are not normal and also distributions of individual clusters can have a significant skewness, we should use a non-symmetric distance for informative clustering. There is an intimate relationship between ordinary Euclidean geometry and the multivariate normal distribution, where L^2-distance is fully legitimate. However, several non-normal exponential families do not in general enjoy simple Euclidean geometry.

To illustrate this, let us compare L^2-dissimilarity d_E with $D(y_1, y_2)$ for the case of two one-dimensional clusters. Cluster C1 is given by an exponential distribution and scale $\gamma_1 = 10$. Cluster C2 is characterized by an exponential distribution and scale $\gamma_1 = 2$. As we can see from Fig. 13, especially extremal values (around 0 and large values) are strongly underestimated by d_E. These extremes, however (as agreed also by the authors), can be essential for explanatory purposes in socio-economic sciences.

Douglas Steinley (*University of Missouri, Columbia*)

I congratulate Hennig and Liao on proposing a novel method for clustering data with mixed-type variables.

I can add only a couple of remarks and points of possible interest to this important advance in clustering.

(a) As alluded to by the authors, the method for choosing the number of clusters is likely to influence the final solution more than any other choice. The demonstration that latent class clustering, when combined with the Bayesian information criterion, favours up to a 20-cluster solution. How much of this is driven by the need to fit the data with the imposed constraint of local independent? For example, in Fig. 1, if each mixture was allowed to have 'free' covariance matrices, would the correct three-cluster solution be identified by mixture modelling? If the Bayesian information criterion was allowed to choose both the number of clusters and the structure of the covariance matrix, would it choose correctly? These general questions extend to the broader 'tug-of-war' between choosing an analytic approach with results that are more substantively useful *versus* an approach that could be argued to approximate the structure of the data better. As George Box said, 'All models are wrong, but some are useful'—do we concern ourselves with only the latter clause? How wrong does a model have to be before its output, even if it has the veneer of usefulness, can be considered only useful coincidentally?

(b) Regarding standardization for continuous variables, the authors indicate that there are three primary methods (e.g. range standardization, unit variance standardization and unit interquartile range standardization). Steinley and Brusco (2008) found that the most informative property was the ratio of the variance to the range, or, more simply put, a variable is more likely to exhibit clustering if, for a fixed range, the variable has a larger variance. I wonder whether such a technique could be extended to the mixed variable setting that is used in the present application. Related, it would be interesting to explore how sensitive the final clustering output is to the choice of weighting for the nominal and ordinal variables—an area of research that, for the most part, has been underexplored.

Iven Van Mechelen (*University of Leuven*)

It has been known for a long time that the outcome of several instances of statistical analysis in general and of cluster analysis in particular may strongly depend on a large number of decisions that cannot be taken in an automated way. Hennig and Liao are to be complimented for a convincing illustration of how a careful reflection on data preprocessing, aspects of the actual data analysis and characteristics of the output to be expected may contribute to a more meaningful cluster analysis. One may, however, wonder whether they went sufficiently far in this regard. The answer to this question cannot be detached from the deeper aim of the clustering. One possibility is that this aim does not pertain to the clustering *per se* but to some external goal, such as the prediction of some criterion or the supply of a suitable treatment to the elements of each of the strata resulting from the analysis. In such cases, meaningful clusterings should meet specific pragmatic criteria (e.g. capture the most critical predictive information, or a good match between strata and suitable available treatments). Alternatively, one may wish the cluster analysis to serve the Platonic principle of 'carving nature at its joints'. Such an endeavour requires a deep reflection on the nature of joints in the context under study, and on what keeps together the carved parts. In our case, this should also go with a more principled reflection on the data analysis and on the extent that this analysis touches on structures of interest underlying the (primary) data (rather than secondary derived dissimilarities) and on data-generating mechanisms. Such a reflection is largely missing in the present paper. In general, major obstacles to it are twofold: substantive researchers (especially in the social sciences) typically spend much more attention on abstract constructs than on the linkage between those constructs and empirical reality, including in particular the specifics of measurement models. From their part, statisticians hide fairly quickly behind the sayings that all statistical models are wrong and that applications in which models can be directly justified are rare, to turn subsequently to models like the latent class clustering model, with an unmistakable mathematical and generic elegance rather than a sound rooting in theories on data-generating mechanisms. Overcoming these two obstacles implies a major challenge for both substantive researchers and statisticians, which will require great efforts from both parties and an intensive iterative interaction between them.

Donatella Vicari (*Sapienza University of Rome*)

This interesting paper casts light on what I would call 'the strategy of analysis' more than clustering philosophy because it is not limited to the clustering problem, but extends to all those situations where statistical methodology meets data.

Hennig and Liao address the need to integrate the general subject matter knowledge with the statistical method, to obtain classes that are well interpretable in a context related, but not limited, to the search of socio-economic classes, which typically represent fuzzy latent concepts that are difficult to define.

It is of great value that the authors consider and discuss this problem, which is undoubtedly of interest to anyone who is involved in data analysis, either applied or methodological statisticians. Their approach has the potential to address the difficulties of all critical different choices in the *correct* perspective for the specific situation under study.

Their paper is an attempt well done to deal with a concept which is generally data specific by warning us not to search for a definition of 'cluster' that is useful in any context as an a *priori* concept, but to consider the subject matter jointly with the cluster shape that each method induces.

On one hand, such a discussion is often confined to philosophical debates on how to discover the natural clusters possibly underlying the data; on the other hand, the theoretical approach often skips such a problem by assuming a model regardless of the interest of the researcher in terms of cohesion or size of the clusters, for instance.

The authors go through all the steps that are necessary for a thorough analysis of a real case-study, starting from the selection of the variables representing different dimensions in the social stratification and the choice of the appropriate transformation for variables of mixed types, to come up to the issue of how differently to weight such variables and last, but not least, the choice of the 'best' partition in terms of the number of clusters.

I would like to focus in particular on the transformation of the variables: how to combine variables on different scales is a crucial issue especially in clustering methods, where the presence of categorical variables may dominate the solution. The choice here is to combine an appropriate 'standardization' of the variables with a downweight on a data-driven basis. I wonder whether the search for an optimal scaling of the variables from the data could be formalized to incorporate in the process the contribution of the different levels of the categorical variables better.

Lulu Wang and Jennifer A. Hoeting (*Colorado State University, Fort Collins*)

We commend Hennig and Liao for their thorough approach to clustering of mixed-type variables by using model-based and non-parametric approaches. Many practitioners will read and reference their work. We consider two areas of the paper related to the uncertainty of the results.

When Hennig and Liao discussed substantial weighting (Section 6.2), they up-weighted two of the levels of the categorical housing variable, 'owns' and 'pays rent', by a factor of 2. They should consider the effect of this weighting, since different weights for this single variable lead to new clustering results. To show this, we computed the k-medoids clustering as follows:

(a) without up-weighting owns and pays rent (i.e. no substantial weighting);
(b) without substantial weighting and combining the other seven categories of the housing variable because we think that these housing levels overlap in meaning. The resulting 'hous' variable has three categories (owns; pays rent; other).

Not surprisingly, the optimal number of clusters depends on the choice of weighting of the housing variable (Table 5 and Fig. 14). Without weighting, the local optima are $k = 7, 9$ on the basis of the average silhouette width ASW and $k = 7$ on the basis of PH. In this case, we should use $k = 7$ instead of $k = 8$ as in the paper for k-medoids. Similarly, combining other categories will lead to local optima $k = 4$ or $k = 9$. Clearly, different optimal numbers of clusters will lead to different clustering results. Moreover, on the basis of the results provided by the authors, we think that they should further defend their choice of $k = 8$ instead of $k = 6$ (another local optimum) or $k = 3$ (optimum based on PH, and which has a larger ASW than for

Table 5. Optimal number of clusters for various weighting schemes of the housing levels variable

Weighting scheme	*k for ASW*		*k for PH*	
	Overall optima	*Local optima*	*Overall optima*	*Local optima*
With weighting (as in the paper)	2	6, 8	3	6, 8
Without weighting	2	7, 9	3	7
Combine other categories	2	4, 6, 9	4	9

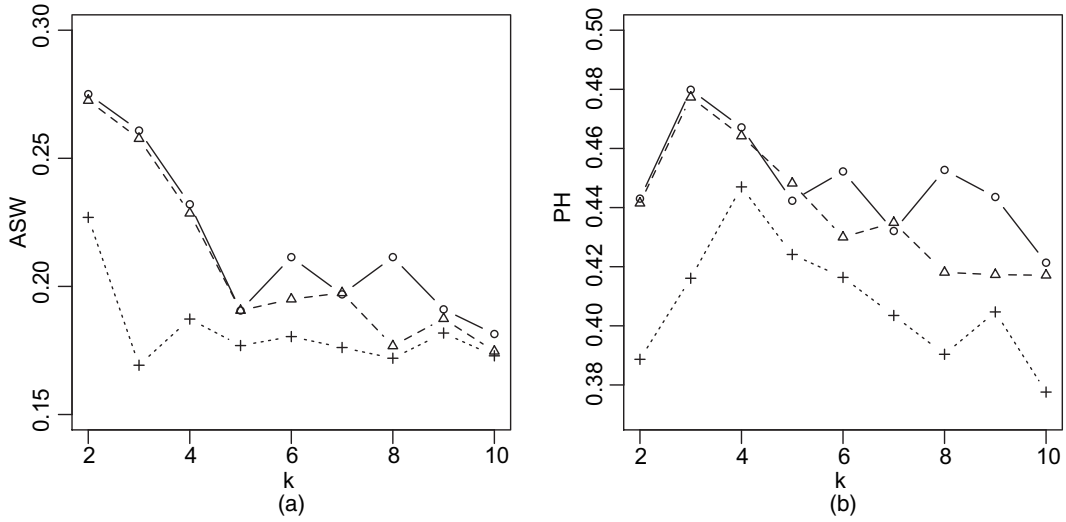

Fig. 14. (a) ASW and (b) PH by number of cluster k for various weighting schemes of the housing levels variable: ———, with weighting as in the paper; – – –, without weighting; · · · · · · ·, combine other categories

$k = 6, 8$) for k-medoids (Fig. 14). One approach to addressing the problem of uncertainty due to substantial weighting would be to adopt a Bayesian approach where the variable weights are treated as parameters with a prior distribution. However, we recognize that Bayesian estimation can be challenging because of the computational complexity.

Although we commend Hennig and Liao for the bootstrap test for k-medoids, we suggest that they consider methods to quantify the uncertainty of clustering results for latent class clustering in more detail. One advantage of model-based clustering is that the framework allows for further inference, such as computation of confidence intervals for parameter estimates or estimation of the probability that an observation belongs to one cluster. Uncertainty can be measured in several ways. For example, Leisch (2004) suggested the ratio of the size of the kth cluster to the number of observations with positive posterior probability, which is $P(X_i = k)$, the probability that X_i is in cluster k.

Ron Wehrens (*Fondazione Edmund Mach, San Michele all'Adige*) **and Mark de Rooij** (*Leiden University*)

Hennig and Liao address the notoriously difficult question of how to find an optimal solution for a vaguely defined problem, finding groups in data. Their proposed strategy is a mixture of subjective decisions based on *a priori* knowledge about the data and the aim of the investigation, and on easily computable statistical criteria.

In our opinion, the use of statistical methods should be guided by two basic principles:

(a) for one data set and one research question, different researchers should reach the same conclusions;
(b) the procedure should not be based on intentions of the researcher, client or anyone else.

Does the stategy proposed bring us closer to these two goals? This remains to be seen. No one will argue against the use of common sense and background knowledge in tackling a clustering problem, but at the same time this may lead to biased judgements.

It would probably be a sensible idea to assess, in addition, the effect of making different choices, e.g. by using computer-intensive procedures. If nothing else, it would present a band of feasible clusterings, indicating the robustness of the final conclusions with respect to the choices made. In a regression and classification context, van der Laan and Rose (2011) used ensembles of different regression techniques (linear regression, additive models, trees, etc.), governed by cross-validation procedures. The main messages in their work are that we do not know in advance which method is good for the data and that we should not make assumptions about the distributions. Something similar could be done in the case of clustering, as demonstrated by, for example, Questier *et al.* (2005). To limit the ensemble to relevant clusterings only, one could formulate prior knowledge in a set of constraints that needs to be satisfied. An example is given by the authors in their Fig. 1: the three-cluster solution is rejected because it would lead to groups that are too

heterogeneous, i.e. objects of different class are grouped together. Such knowledge, e.g. objects that should or should not be clustered together, is often partially present. Furthermore, if one thinks that variables might need a transformation the ensemble should include methodology that transforms the variables, such as the optimal scaling methods of Gifi (Gifi, 1990; Michailidis and de Leeuw, 1998).

In conclusion, we are not convinced that the promise that is implied in the title of the paper is fulfilled but at least it has provided us with food for thought.

Ruben H. Zamar and Hongyang Zhang (*University of British Columbia, Vancouver*)

Hennig and Liao present and discuss a very nice application of cluster analysis to social sciences. Perhaps the actual application should appear earlier in the paper to give even more practical meaning to the philosophical discussion of statistical methodology.

Outliers could be highly informative. Therefore one may try to isolate, highlight and give special consideration to them. For instance, there is the possibility of using impartial trimming (Cuesta-Albertos *et al.*, 1997) to robustify k-means preventing outliers from having undue influence in the clustering process and to flag them as a minority of special cases. Hence we have some reservations regarding the authors' attempt to 'accommodate' and 'tame' the outliers by using a convenient scale choice. Moreover, although application of the log-scale to positive data is commonplace in statistics we would not go as far as asserting that the difference in income between two people making 20000 and 40000 each is the same as that between two people making 2 million and 4 million each.

Reducing attention to latent class clustering and k-means simplifies the discussion but alternative approaches such as hierarchical clustering and mode climbing clustering deserve mentioning and some discussion. The authors are interested in finding not only useful partitions but also insightful interpretation for the clusters. A powerful tool for finding useful clusters and at the same time a reduced number of variables responsible for producing these clusters is the so-called 'sparse clustering approach' introduced by Witten and Tibshirani (2010). A robustified version of sparse k-means is available in R packag `RSKC` based on Kondo *et al.* (2012).

Another possible way to fulfil Hennig and Liao's goals could be to model the data as a data cube. Data cubes (Gray *et al.*, 1997) were originally designed for mining multi-dimensional data and are widely used in on-line analytical processing. They group data into cubes by using multiple cube dimension attributes (i.e. categorical and possibly some ordinal variables) and extract similarities within each group. In this application, it is possible that, for the individuals corresponding to different combinations of the categories, say 'hous' and 'occ', the clustering nature may differ. By employing the data cube technique, the clustering can be performed first within the base cube cells (i.e. cells separated by the cube dimensions) and then the relationships between different base cube cells should be further investigated to combine the cluster information of similar cells together. Instead of conducting cluster analysis based on the whole data set, such an approach will allow the cluster analysis on several similar individual groups generated by variables such as hous and occ, and hence is likely to provide more informative clusters.

The **authors** replied later, in writing, as follows.

We are extremely grateful for all the interest in our paper and all the comments, complimentary, constructive or critical. Given their large number, we apologize for being unable to respond to all the valuable comments.

Aspects of social stratification

Lambert discussed our use of the terms 'social class' and 'social stratification'. The concept of social class as discussed in the theoretical literature is indeed much more involved than a simple reference to a cluster that is social and contains a group of similar individuals, though we could only provide an extremely brief summary of it in the paper. Similarly, a serious discussion of social stratification would require an entire review paper. In the paper, we used the simple and intuitive concept of clusters in the data so that these can be related to the various theoretical definitions (Kettenring suggests, rightly, that hierarchies could also be of interest).

Our objective in the paper was to present a method for estimating socio-economic stratification by focusing on indicators that measure socio-economic wellbeing directly. We knew fully well that the generating mechanisms and the distribution of social stratification can vary according to gender, life course stages, cohorts and spatial locations. Lambert, Mateu, O. and C. Melo, and Saez all noted the possible dynamics of social stratification along the lines of the above or other contextual variables (in Saez's terms), but we do not see them as defining socio-economic stratification. These can be good topics for follow-up analyses once a

stratification system has been established with data through analysing the primary indicators of social or socio-economic stratification.

Lambert made an important point that there have indeed been efforts to combine multiple variables into a single measure of social position. Two such examples are Gershuny's (2002) Essex score, which classifies life chances and identifies the extent of access by individuals or households to the resources that determine the distribution of economic power in society, and the human development index, which is a composite measure combining life expectancy, education and income, designed by Mahbub ul Haq and Amartya Sen for the United Nations Development Programme. Both are continuous measures. No effort has been made to our knowledge to identify discernible clusters along these dimensions. In contrast, identifying such clusters has been our goal.

A complete social stratification system should include all people, whether or not they have a job. Lambert found the inclusion of the 'not-working' category problematic and suggested a multiprocess model (i.e. having *versus* not having a job and having one type of job *versus* another type). The goal in our paper is to estimate socio-economic stratification as presented by the data. Considering further what processes might be involved in getting someone a job *versus* not getting a job or getting the person one type of job *versus* another is a worthwhile topic but beyond the scope of our paper. Excluding the 'not-working' category would distort the representation of the data and of course would result in a different classification, as Lambert obtained.

A related issue, which was also raised by Lambert, is about the use of a continuous occupation-based measure. Again, it is true that such measures would produce different clustering results. However, using a continuous measure of occupation requires modelling occupation with the assistance of other variables. Continuous scales of occupation either use expert judges or population sample surveys to assign and rate occupations according to concepts such as prestige, social standing or general desirability or such a scale uses information on friendship interactions because 'incumbents of occupations that are socially similar would tend to interact more than incumbents of those that are dissimilar' (Prandy (1990), page 630). Therefore, both approaches employ some additional (latent or observed) variables which are not direct measures of occupation that are preferred as indicators of occupation.

Feller and Airoldi combined the 'less than high school' categories and redefined education as ordinal for a comparison with our clustering. They found noticeable discrepancies. The difference is mainly due to collapsing the 12 categories below high school completion. Given the importance of the education variable for the 8-medoids model (see Table 2 in the paper), this is a tremendous loss of information for which we see no good reason. Also, they worried about two cases that share similar values in education, income and occupation but belong to different clusters. But one cannot ignore the other five attributes in such comparisons.

We agree with Müllensiefen and Skiada that social stratification is most often considered as an ordinal variable. In the latent class analysis tradition, constraints can be applied to ordered latent classes. Whereas typical cluster analysis does not yield ordinality, in our post-clustering analysis, we showed how the clusters obtained line up with the distribution of a metric or ordinal variable, such as income and education. Finding out to what extent a clustering without order restrictions would give results that are consistent with the idea of ordered social classes is an informative aspect of our analysis.

Subjective decisions, substantive knowledge

Our analysis required some subjective decisions, some of which were criticized by discussants. McNicholas and Browne and Nugent and Flynt correctly noted that often available expert knowledge is not sufficiently strong to make all the required decisions. Knowledge may simply be absent, or it may be rough and qualitative where precise quantitative decisions are required. A typical example is the $\log(x+c)$ transformation for income and savings. Researchers need to realize though that such decisions are always required. Most often in the literature one would either find no transformation of such variables at all or the use of $c=1$, but certainly there is no better reason for using such 'default' choices, the implications of which are not usually discussed.

Montanari and Monari asked whether we still can find the unexpected with so much tuning. Note that making this kind of tuning choice does not predetermine the clustering result any more than a default choice. Further, addressing some other remarks, we note that most of the decisions to be made here are not of the kind that are usually addressed by Bayesian prior distributions. How to transform these variables is not a question of belief about an underlying truth, but a question of what kind of meaning the clustering should have in terms of the variables.

Wehrens and de Rooij wrote that 'the procedure should not be based on intentions of the researcher'. We think that there is an important distinction between researchers' intentions regarding the kind of in-

terpretation of the results and how the data respond to the researcher's aims. It is legitimate to cluster the data without transforming income and savings, but such a clustering would have a different meaning from ours. Researchers need to think thoroughly about what implications data preprocessing has on the meaning of the results. Here their intentions count and should be transparent (this makes the concept of formal 'optimality', as requested by Vicari and Okada, problematic). We agree only with Wehrens and de Rooij's statement regarding how the data respond to the researcher's questions.

A recurring theme of our paper was the issue of data-driven decision making. Stehlík argues that the distance should be determined by the distribution of the data; Kettenring asks for cluster-adapted schemes for standardization and distance design; several comments cite optimal scaling; some discussants mention data-driven techniques for dimension reduction. There is a potential conflict over whether the topology of the data space is based on the substantial meaning of the variables or on a purely mathematical representation of the data set (which still depends on preprocessing). Data-driven techniques as mentioned above have their merits in situations either in which the variables are of less individual interest or in which exploring the topology that is introduced by the variables is the major aim, but researchers should have in mind that the meaning of a dissimilarity-based clustering depends on the dissimilarity definition and the incorporated variables, and it is not desirable here that the data determine not only that clustering but also its meaning.

Similar points can be made regarding model-based clustering. Chanialidis and colleagues are correct about the fact that we sacrificed invariance against general linear transformations by restricting covariance matrices to be diagonal. This was intended, because the effect of using fully invariant methods is the same as of the Mahalanobis distance, as discussed in the paper, We did not want implicitly to downweight variation in the direction of correlation between our variables, which were included because of their substantial contribution to socio-economic stratification. They should have more weight than arbitrary linear combinations.

Whether non-Gaussian mixtures should be used, as suggested several times, is not only a question of the empirically observed distribution, which can be well approximated by high k mixtures of various families of distributions. It is also important that cluster shapes implied by such distributions make substantial sense. This is difficult to assess but having heavy tails within clusters is problematic if clusters should bring similar observations together.

Sensitivity, uncertainty and validation

Many discussants asked for more quantification of uncertainty and sensitivity analyses. We agree with these requests; see what is already in Section 7.6. Adding one example, we computed k-medoids clusterings for choosing c in the $\log(x+c)$ transformation for income and savings as 1, 5, 10–100 in steps of 10, 200 and 500. We chose k by maximizing the average silhouette width ASW between 5 and 9 (emulating approximately the informal decision that we made from Fig. 6 in the paper). For $c \in [50, 80]$ this yielded $k = 8$ and high adjusted Rand index values around 0.8 compared with our original clustering. However, for $c = 40$, $k = 5$ was chosen and the adjusted Rand index was only 0.59 (the lowest for all $c < 500$). For other c, k was 5, 6 or 8. It is not problematic that $c = 1$ or $c = 500$ deliver rather different clusterings, because these values are clearly different from 50 regarding interpretation, so one should not expect a very similar clustering. One would want to see stability for modifications of the method that are interpretationally 'about the same', but not necessarily if they are clearly different. Still, according to sensitivity analyses that we performed and the results that some discussants reported, some concern about stability is justified, and we concur with Longford's comment that a single partition of the society should not be declared to be universally optimal.

A general difficulty with the quantification of uncertainty is that there is a vast number of aspects to be taken into account, e.g. random variation (model-based inference explores only this), model uncertainty, tuning and choice of method, initialization of algorithms and the effect of influential observations.

We agree that visualization as suggested in some comments is helpful, as can be cross-validation and supervised methods. The figures that were provided by Feller and Airoldi do not show a strong clustering structure (apart from the zero-income and zero-savings groups, the latter of which is not very meaningful), but such projections suppress much information and we still argue that our clustering is potentially useful, and that certain (albeit weak) clustering structure exists in the data as confirmed by our Section 7.3.

Role of probability models

Landau and Ritter identified clusters with finding latent subpopulations described by homogeneous distributions. Similar points were raised by Kettenring, and McNicholas and Browne (citing Aitkin *et al.* (1981)). McLachlan also defended model-based clustering. However, one could hardly argue that this is the

'traditional' meaning of clustering, because the earliest published methods for clustering were not model based. We clearly see the benefit of using probability models for understanding the methods, for assessing uncertainty caused by random variation, and, as done in the paper, for testing a homogeneity null model. However, we reject a universalist claim for model-based clustering. Firstly, the choice of the appropriate family of distributions is far from trivial, not only because all sufficiently simple models are usually wrong in practice, but also because the identification of a cluster with a single mixture component is problematic (see Baudry *et al.* (2010) and Hennig (2010), and because identifying the number of mixture components in large real data sets is an ill-posed problem. Secondly, one could argue that the (more traditional) aim of bringing similar observations together in clusters, separated from other clusters, is more appropriately formalized in terms of dissimilarities than in terms of probability models. It is as relevant to ask how dissimilarity-based methods perform when confronted with data that are generated by latent class models, as it is to ask how good the clusters obtained from model-based clustering are in terms of dissimilarities (as done by Anderlucci in her doctoral thesis).

In the paper, we used ASW (besides other criteria) to compare the latent class clustering (LCC) with the k-medoids clustering. Alexandrovich and Holzmann as well as Pastore and Tonellato argued that this may be unfair because the LCC does not optimize such a criterion. However, for a given real data set and a clustering, it is possible to evaluate ASW, but it is not possible to observe how well an underlying probabilistic truth is estimated. Anderlucci found that in fact there are situations in which model-based clustering produces a better ASW than k-medoids. The good performance of LCC in simulations with artificial data (as mentioned by Bacher) does not mean too much for real data that are clearly different from typical simulated data with easily identifiable clustering.

Steinley asked the interesting question: 'how wrong does a model have to be before its output can be considered only useful coincidentally?'. Ideally it is possible to measure the usefulness. We used ASW here, which admittedly is a rather limited surrogate of all the uses that such a social stratification can have. Certainly it is worthwhile (as suggested by Van Mechelen) to look for external substantial criteria. The approach by Laurie Davies, cited by Coretto, starts from defining features of the data of interest to the researcher and explores whether these are in line with the model. To apply such an approach to clustering is an exciting prospect.

Test of homogeneity

Although Bacher suggested an 'absence of class structure' (and claimed that it was impossible for us to test that), Section 7.3 suggests otherwise. Some comments have been made regarding this test (actually this was a series of informal tests for each k, which could be aggregated into a single test). Geyer-Schulz asked about what is implied by the null model. Ceyhan worried about the Gaussian assumption and suggested the non-parametric bootstrap. But the non-parametric bootstrap would not provide a model for 'no clustering'; if the data are indeed clustered, non-parametric bootstrap samples are also clustered. It is true that what we rejected was a null model with a Gaussian distribution of the continuous variables. We rejected the model in favour of an alternative with stronger clustering (as measured by ASW). It would have been possible to transform latent Gaussian variables to the marginal distributions in the original data set, but we decided against this because marginals that are more clustered than a Gaussian distribution is can legitimately be highlighted as generating a significant clustering. The rationale to define the null model, to be used in other applications as well, is that the null model should capture the features of the original data that are not interpreted as indicating 'clustering' (here the correlation structure and the marginals of the categorical variables).

Further statistical aspects

Several discussants recommended trimming or removing outliers. Visual analysis (using the grand tour (Cook *et al.*, 1995)) suggested the effect of outliers on the k-medoids clustering was clearly lower than the effect of discreteness caused by categorical variables. Existing techniques for trimming have to our knowledge not yet been used with mixed-type variables, although this would certainly be of interest.

Okada wondered about what caused the differences between the 8-medoids clustering and the LCC. The k-medoids method tends to produce clusters of similar size, whereas LCC allows for more varying within-cluster variances, which enables the latter to collect a very homogeneous and small low income class.

Baudry noted that a very good supervised classification exists for occupation based on the other variables, despite our claim that occupation has a weak connection to our clustering. This illustrates how different supervised and unsupervised classification really are. Baudry called the partition that is induced by his classification rule 'clustering', but it is important to realize that the existence of a good classification rule

does not imply any clustering of the data at all (uniformly distributed data may be perfectly divided into classes by an external variable). Separation and homogeneity of clusters are not always related to supervised partitioning.

In many comments additional methodological suggestions were made. Many of them are reasonable alternatives to what we have done. Just to mention two examples, the integrated complete-likelihood criterion that was suggested by Celeux and other alternatives may indeed improve on the Bayesian information criterion and promising alternatives to LatentGold's Bayes method to prevent a degenerating likelihood were suggested by Grün and Malsiner-Walli as well as Garcia-Escudero and colleagues. We appreciate that better clusterings than ours may still be possible. Our key message in this respect is that, for choosing from a set of available methods, one needs to understand, as we attempted in the paper, how these methods relate to the meaning of the data and the aim of the analysis. We concur with Van Mechelen on the call for an even closer connection of the analysis to background knowledge and clustering aim. Modelling data-generating mechanisms would be fascinating here but was beyond the scope of our paper.

This illustrates how many choices there are and how many decisions must be made when doing cluster analysis. Developing tools for guidance as mentioned by Nakoinz is certainly a promising research direction. We hope that the structure of our paper listing the key decisions in one particular application can help in this respect.

References in the discussion

Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L. and Syme, S. (1994) Socioeconomic status and health: the challenge of the gradient. *Am. Psychol.*, **49**, 15–24.

Aitkin, M., Anderson, D. and Hinde, J. (1981) Statistical modelling of data on teaching styles (with discussion). *J. R. Statist. Soc.* A, **144**, 419–461.

Anderlucci, L. (2012) Comparing different approaches for clustering categorical data. *Tesi di Dottorato*. Alma Mater Studiorum, Università di Bologna, Bologna.

Bacher, J. (2003) A probabilistic clustering model for variables of mixed type. *Qual. Quant.*, **34**, 223–235.

Bacher, J., Wenzig, K. and Vogler, M. (2004) SPSS Two Step—a first evaluation. (Available from `http://www.statisticalinnovations.com/technicalsupport/articles.html#otherchoice`.)

Barndorff-Nielsen, O. E. (1978) *Information and Exponential Families in Statistical Theory*, pp. 111–115. New York: Wiley.

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K. and Gottardo, R. (2010) Combining mixture components for clustering. *J. Computnl Graph. Statist.*, **19**, 332–353.

Biernacki, C., Celeux, G. and Govaert, G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattn Anal. Mach. Intell.*, **22**, 719–725.

Birkelund, G. E., Goodman, L. A. and Rose, D. (1996) The latent structure of job characteristics of men and women. *Am. J. Sociol.*, **102**, 80–113.

Bischof, J. and Airoldi, E. (2012) Summarizing topical content with word frequency and exclusivity. *Int. Conf. Machine Learning, Edinburgh*. (Available from `http://arxiv.org/abs/1206.4631`.)

Booth, J. G., Casella, G. and Hobert, J. P. (2008) Clustering using objective functions and stochastic search. *J. R. Statist. Soc.* B, **70**, 119–139.

Bottero, W. (2005) *Stratification: Social Division and Inequality*. London: Routledge.

Bouveyron, C., Fauvel, M. and Girard, S. (2012) Kernel discriminant analysis and clustering with parsimonious Gaussian process models. *Preprint HAL 00687304*. Université Paris 1 Panthéon-Sorbonne, Paris.

Breiman, L. (2001a) Random forests. *Mach. Learn.*, **45**, 5–32.

Breiman, L. (2001b) Statistical Modeling: the two cultures (with discussion). *Statist. Sci.*, **16**, 199–231.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) *Classification and Regression Trees*. New York: Chapman and Hall.

Carlsson, G. (2009) Topology and data. *Bull. Am. Math. Soc.*, **46**, 255–308.

Celeux, G. and Govaert, G. (1991) Clustering criteria for discrete data and latent class models. *J. Classificn*, **8**, 157–176.

Celeux, G. and Govaert, G. (1992) A classification EM algorithm for clustering and two stochastic versions. *Computnl Statist. Data Anal.*, **14**, 315–332.

Celeux, G. and Soromenho, G. (1996) An entropy criterion for assessing the number of clusters in a mixture model. *J. Classificn*, **13**, 195–212.

Chang, G. and Walther, G. (2007) Clustering with mixtures of log-concave distributions. *Computnl Statist. Data Anal.*, **51**, 6242–6251.

Coley, R. J. (2002) *An Uneven Start: Indicators of Inequality in School Readiness*. Princeton: Educational Testing Service.

Cook, D., Buja, A., Cabrera, J. and Hurley, C. (1995) Grand tour and projection pursuit. *J. Computnl Graph. Statist.*, **4**, 155–172.

Cormack, R. M. (1971) A review of classification. *J. R. Statist. Soc.* A, **134**, 321–367.
Cox, D. R. and Donnelly, C. A. (2011) *Principles of Applied Statistics*. Cambridge: Cambridge University Press.
Csiszar, I. (1975) I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, **3**, 146–158.
Cuesta-Albertos, J. A., Gordaliza, A. and Matran, C. (1997) Trimmed *k*-means: an attempt to robustify quantizers. *Ann. Statist.*, **25**, 553–576.
Cule, M., Samworth, R. and Stewart, M. (2010) Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. R. Statist. Soc.* B, **72**, 545–607.
Davies, P. L. (1995) Data features. *Statist. Neerland.*, **49**, 185–245.
Davies, P. L. (2008) Approximating data (with discussion). *J. Kor. Statist. Soc.*, **37**, 191–240.
Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
De Leeuw, J. (1988a) Multivariate analysis with linearizable regressions. *Psychometrika*, **53**, 437–454.
De Leeuw, J. (1988b) Multivariate analysis with optimal scaling. In *Proc. Int. Conf. Advances in Multivariate Statistical Analysis* (eds S. Das Gupta and J. K. Ghosh), pp. 127–160. Calcutta: Indian Statistical Institute.
De Leeuw, J. and Mair, P. (2009) Gifi methods for optimal scaling in R: the package homals. *J. Statist. Softwr.*, **31**, 1–20.
Dharmadhikari, S. and Joag-Dev, K. (1988) *Unimodality, Convexity, and Applications*. Boston: Academic Press.
Efron, B. (1978) The geometry of exponential families. *Ann. Statist.*, **6**, 362–376.
Evans, G. and Mills, C. (1998) Identifying class structure: a latent class analysis of the criterion-related and construct validity of the Goldthorpe class schema. *Eur. Sociol. Rev.*, **14**, 87–106.
Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*, 5th edn. Chichester: Wiley.
Fraley, C. and Raftery, A. E. (2007) Bayesian regularization for normal mixture estimation and model-based clustering. *J. Classificn*, **24**, 155–181.
Franczak, B. C., Browne, R. P. and McNicholas, P. D. (2012) Mixtures of shifted asymmetric Laplace distributions. *Preprint arXiv:1207.1727v2*.
Friedman, J. H. (1987) Exploratory projection pursuit. *J. Am. Statist. Ass.*, **82**, 249–266.
Frühwirth-Schnatter, S. (2006) *Finite Mixture and Markov Switching Models*. New York: Springer.
García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008) A general trimming approach to robust cluster analysis. *Ann. Statist.*, **36**, 1324–1345.
Gershuny, J. (2002) A new measure of social position: social mobility and human captial in Britain. *Working Paper 2002-02.* Institute for Social and Economic Research, University of Essex, Colchester.
Gifi, A. (1990) *Nonlinear Multivariate Analysis*. Chichester: Wiley.
Gordon, A. D. (1999) *Classification*, 2nd edn. London: Chapman and Hall.
Gower, J. (1968) Adding a point to vector diagrams in multivariate analysis. *Biometrika*, **55**, 582–585.
Graf, S. and Luschgy, H. (2000) *Foundations of Quantization for Probability Distributions*. Berlin: Springer.
Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F. and Pirahesh, H. (1997) Data cube: a relational aggregation operator generalizing group-by, cross-tab and sub-totals. *Data Minng Knowl. Discov.*, **1**, 29–53.
Hand, D., Mannila, H. and Smyth, P. (2011) *Principles of Data Mining*. Cambridge: MIT Press.
Hennig, C. (2010) Methods for merging Gaussian mixture components. *Adv. Data Anal. Classificn*, **4**, 3–34.
Höppner, F., Klawonn, F., Kruse, R. and Runkler, T. (1999) *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition*. Chichester: Wiley.
Huang, J. Z., Ng, M. K., Rong, H. and Li, Z. Automated variable weighting in k-mean type clustering. *IEEE Trans. Pattn Anal. Mach. Intell.*, **27**, 657–668.
Huber, P. J. (1985) Projection pursuit. *Ann. Statist.*, **13**, 435–475.
Hunt, L. A. and Jorgensen, M. A. (1999) Mixture model clustering: a brief introduction to the MULTIMIX program. *Aust. New Zeal. J. Statist.*, **40**, 153–171.
Hunt, L. and Jorgensen, M. (2003) Mixture model clustering for mixed data with missing information. *Computnl Statist. Data Anal.*, **41**, 429–440.
Irigoien, I. and Arenas, C. (2008) INCA: new statistic for estimating the number of clusters and identifying atypical units. *Statist. Med.*, **27**, 2948–2973.
Irigoien, I., Fernández, E., Vives, S. and Arenas, C. (2008) Clum: a cluster program for analyzing microarray data. *Russ. J. Genet.*, **44**, 993–996.
Jones, M. C. and Sibson, R. (1987) What is projection pursuit (with discussion)? *J. R. Statist. Soc.* A, **150**, 1–36.
Kaufman, L. and Rousseuw, P. J. (1990) *Finding Groups in Data*. New York: Wiley.
Kerbo, H. R. (2003) *Social Stratification and Inequality: Class Conflict in Historical, Comparative and Global Perpective*, 5th edn. London: McGraw-Hill.
Kolenikov, S. and Angeles, G. (2009) Socioeconomic status measurement with discrete proxy variables: is Principal Components Analysis a reliable answer? *Rev. Incm. Wlth*, **55**, 128–165.
Kondo, Y., Salibian-Barrera, M. and Zamer, R. H. (2012) A robust and sparse K-means clustering algorithm. *Preprint arXiv:1201.6082v1*.
Kriegel, H.-P., Kröger, P. and Zimek, A. (2009) Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data*, **3**, 1–58.

Krugman, P. R. (1996) *The Self-organizing Economy*. Oxford: Blackwell Publishers.
Kuper, A. (ed.) (2004) Class, social. In *The Social Science Encyclopedia*, p. 111. Basingstoke: Taylor and Francis.
van de Laan, M. J. and Rose, S. (2011) *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer.
Ligges, U., Roever, C., Raabe, N., Luebke, K., Szepannek, G. and Zentgraf, M. (2012) klaR—R package for classification and visualization. (Available from `http://cran.r-project.org/package=klaR`.)
Link, B. G. and Phelan, J. (1995) Social conditions as fundamental causes of disease. *J. Hlth Socl Behav.*, **35**, extra issue, 80–94.
Linting, M., Meulman, J. J., Groenen, P. J. and van der Kooji, A. J. (2007) Nonlinear principal components analysis: introduction and application. *Psychol. Meth.*, **12**, 336–358.
Lloyd, S. P. (1982) Least squares quantization in PCM. *IEEE Trans. Inform. Theor.*, **28**, 129–137.
McLachlan, G. J. (2011) Commentary on 'Evaluating mixture modeling for clustering: recommendations and cautions' by D. Steinley and M. J. Brusco. *Psychol. Meth.*, **16**, 80–81.
McLachlan, G. J. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley.
McLeod, J. D. and Kessler, R. C. (1990) Socioeconomic status differences in vulnerability to undersirable life events. *J. Hlth Socl Behav.*, **31**, 162–172.
McParland, D., Gormley, I. C., Clark, S. J., McCormick, T. H., Kabudula, C. W. and Collinson, M. A. (2012) Clustering South African households based on their asset status using latent variable models. *Technical Report 604*. University of Washington, Seattle.
Menardi, G. (2011) Density-based Silhouette diagnostics for clustering methods. *Statist. Comput.*, **21**, 295–308.
Michailidis, G. and de Leeuw, J. (1998) The Gifi system of descriptive multivariate analysis. *Statist. Sci.*, **13**, 307–336.
Milligan, G. W. (1996) Clustering validation: results and implications for applied analysis. In *Clustering and Classification* (eds P. Arabie, L. J. Hubert and G. De Soete), pp. 341–375. Singapore: World Scientific Publishing.
Molitor, J. T., Papathomas, M., Jerrett, M. and Richardson, S. (2010) Bayesian profile regression with an application to the National Survey of Childrens Health. *Biostatistics*, **11**, 484–498.
Müller, P., Quintana, F. and Rosner, G. L. (2008) A product partition model with regression covariates. *J. Computnl Graph. Statist.*, **20**, 260–278.
Nakoinz, O. (2010) Concepts of central place research in archaeology. In *Landscapes and Human Development: the Contribution of European Archaeology: Proc. Int. Wrkshp Socio-environmental Dynamics over the Last 12,000 Years: the Creation of Landscapes, Apr. 1st–4th* (ed. Kiel Graduate School 'Human Development in Landscapes'), pp. 251–264. Bonn: Habelt.
National Center for Education Statistics (2008) Percentage of high school dropouts among persons 16 through 24 years old (status dropout rate), by income level, and percentage distribution of status dropouts, by labor force status and educational attainment: 1970 through 2007. National Center for Education Statistics, Washington DC. (Available from `http://nces.ed.gov/programs/digest/d08/tables/dt08_110.asp`.)
National Center for Health Statistics (2011) Health, United States, 2011. National Center for Health Statistics, Hyattsville. (Available from `http://www.cdc.gov/nchs/hus.htm`.)
Nyamundanda, G., Brennan, L. and Gormley, I. C. (2010) Probabilistic principal component analysis of metabolomic data. *BMC Bioinform.*, **11**, 571.
Pages, J. (2004) Analyse factorielle de données mixtes. *Rev. Statist. Appl.*, **42**, no. 4, 93–111.
Pázman, A. (1993) *Nonlinear Statistical Models*, sect. 9.1, 9.2. Dordrecht: Kluwer.
Peel, D. and McLachlan, G. J. (2000) Robust mixture modelling using the t distribution. *Statist. Comput.*, **10**, 339–348.
Pepe, M. S. and Janes, H. (2007) Insights into latent class analysis of diagnostic test performance. *Biostatistics*, **8**, 474–484.
Plant, C. and Böhm, C. (2011) INCONCO: interpretable clustering of numerical and categorical objects. In *Proc. KDD '11: 17th Association for Computing Machinery Special Interest Group in Knowledge Discovery and Data Mining Int. Conf. Knowledge Discovery and Data Mining*, pp. 1127–1135. New York: Association for Computing Machinery.
Pollock, G. (2007) Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *J. R. Statist. Soc.* A, **170**, 167–173.
Prandy, K. (1990) The revised Cambridge Scale of Occupation. *Sociology*, **24**, 629–655.
Questier, F., Put, R., Coomans, D., Walczak, B. and van der Heyden, Y. (2005) The use of CART and multivariate regression trees for supervised and unsupervised feature selection. *Chemometr. Intell. Lab. Syst.*, **76**, 45–54.
Roever, C. and Szepannek, G. (2005) Application of a genetic algorithm to variable selection in fuzzy clustering. In *Classification—the Ubiquitous Challenge* (eds C. Weihs and W. Gaul), pp. 675–681. New York: Springer.
Rubin, D. B. and Thayer, D. T. (1982) EM algorithms for ML factor analysis. *Psychometrika*, **47**, 69–76.
Sheikholeslami, G., Chatterjee, S. and Zhang, A. (1998) WaveCluster: a multi-resolution clustering approach for very large spatial databases. In *Proc. 24th VLDB Conf.*
Stehlík, M. (2003) Distributions of exact tests in the exponential family. *Metrika*, **57**, 145–164.
Steinbach, M., Ertoz, L. and Kumar, V. (2003) Challenges of clustering high dimensional data. In *New Vistas in Statistical Physics—Applications in Econophysics, Bioinformatics, and Pattern Recognition* (ed. L. T. Wile). New York: Springer.

Steinley, D. and Brusco, M. J. (2008) Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika*, **73**, 125–144.
Sturgis, P. and Sullivan, L. (2008) Exploring social mobility with latent trajectory groups. *J. R. Statist. Soc.* A, **171**, 65–88.
Tampubolon, G. and Savage, M. (2012) Intergenerational and intragenerational social mobility in Britain. In *Social Stratification: Trends and Processes* (eds P. S. Lambert, R. Connelly, R. M. Blackburn and V. Gayle), pp. 115–131. Aldershot: Ashgate.
Tibshirani, R., Walther, G. and Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc.* B, **63**, 411–423.
Tucker-Drob, E. M., Rhemtulla, M., Harden, K. P., Turkheimer, E. and Fask, D. (2010) Emergence of a gene ×socioeconomic status interaction on infant mental ability between 10 months and 2 years. *Psychol. Sci.*, **22**, 125–133.
Turkheimer, E. I. (2003) Socioeconomic status modifies heritability of IQ in young children. *Psychol. Sci.*, **14**, 623–628.
Vermunt, J. K. and Magidson, J. (2005) *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Belmont: Statistical Innovations.
Witten, D. M. and Tibshirani, R. (2010) A framework for feature selection in clustering. *J. Am. Statist. Ass.*, **105**, 713–726.