How to find an appropriate clustering for mixed type variables with application to socioeconomic stratification

Christian Hennig and Tim F. Liao

Department of Statistical Science, UCL,
Department of Sociology, University of Illinois

November 28, 2011

Abstract

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 BIC 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. According to this philosophy, model assumptions serve to understand such implications but are not taken to be true. It is emphasised that researchers implicitly define the "true" clustering and number of clusters by the choice of a particular methodology. The researcher has to take the responsibility to specify the criteria on which such a comparison can be made. The application of this philosophy to socioeconomic data from the 2007 US Survey of Consumer Finances demonstrates some techniques to obtain an interpretable clustering in an ambiguous situation.

Keywords: social stratification, cluster philosophy, latent class clustering, mixture model, k-medoids clustering, dissimilarity measure, number of clusters, average silhouette width, interpretation of clustering

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 cluster analysis methods for social stratification based on mixed type data with continuous, ordinal and nominal variables.

A main feature of the paper is that the social stratification application illustrates a more 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.

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 not based on probability models (k-medoids, Kaufman and Rouseeuw, 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 classes is central to social science research, there is no agreed-upon definition of a social class. It is of interest here whether social stratification based on formal clustering can contribute to the controversy about social classes. In this paper we analyse data from the US Survey of Consumer Finances, for which the more appropriate term is "socioeconomic stratification", though many of the considerations apply to social stratification in general.

The philosophy behind the choice of a cluster analysis method in the present paper is that it should be driven by the way concepts like "similarity" and "belonging together in the same class" are interpreted by the subject-matter researchers and by the way the clustering results are used. This requires thinking different from what is normally found in the literature. Particularly, model assumptions are no longer treated as referring to an "underlying truth", but rather serve to characterise the methods based on them.

Section 2 introduces the problem of social stratification and discusses how cluster analysis can contribute to it. In Section 3 latent class clustering 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. Aspects of this are the definition of dissimilarity, treated in Section 6, and the choice of the number of clusters, discussed in Section 7. Section 8 compares the clustering methods that were introduced in previous sections. In Section 9, the US Survey of Consumer Finances dataset is analysed. Section 10 gives a concluding discussion.

There is a vast literature on cluster analysis beyond the scope of our paper that we do not attempt to review here. Some justification of our choice of methods is given in Section 8.

2 Social stratification

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 todays ongoing academic interest in issues of social class and stratification for both research and teaching purposes (e.g., Grusky, Ku, and Szelnyi 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, 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 (a variable they operationalised themselves). The reason for researchers routinely using some indicators of social class is that there is no agreed-upon definition of social class, let alone a specific agreed-upon operationalisation of it. Neither is the use of social stratification unified.

Various different concepts of social classes 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 "let the data speak" concerning the issues discussed in the literature, acknowledging that data alone cannot decide the issue objectively. Our aim is to see if meaningful or useful social or socioeconomic classes may emerge from a cluster analysis. It is of interest whether clear clusters are apparent in data consisting of indicators of social or socioeconomic class, but also how these data can be partitioned in a potentially useful way even without assuming that the classes correspond to essentially different underlying subpopulations. They may serve as efficient reduction of the information in the data and as a tool to decompose and interpret inequality. A latent class model was proposed for this by Grusky and Weeden (2008). A similar finite model was proposed and applied to income inequality data by Liao (2006). It is also of interest how data clusterings relate to theoretical concepts of social stratification.

A crucial question for data based social stratification is the choice of variables on

which the classification is based, although this is restricted by what is available for observation. For the dataset analysed here, the focus is on socioeconomic classes. The allocation of individuals to such classes is an exercise of classification along the dimensions of the individuals socioeconomic wellbeing. There is a general consensus on how socioeconomic wellbeing is measured. People's current socioeconomic 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 socioeconomic wellbeing, and an influential approach to operationalise economic welfare 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 a lot of 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. While occupational categories may change and get updates, occupation is central to studying stratification. Blau and Duncan's (1967) landmark study of the American occupational structure started a modern tradition that was more or less unchallenged until Spilerman's (2000) criticism of the preoccupation with labour market rewards including such as occupation and income and with the consideration of family status as derivative of 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 analysed dataset 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 socioeconomic status by a minimum set of the three variables education, income, and occupation, we extend this approach to including more socioeconomic 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 individuals savings in various accounts represent a significant part of ones net worth (see Poterba, Venti, and Wise (1994) who studied retirement savings and net worth of American household heads aged 55-69). As reported by Bernheim, Garrett, and Maki (2001), the correlation between self-reported saving and net worth is highly statistically significant.

Another dimension of socioeconomic welfare we examine is the diversification of the place one keeps ones disposable income and saving. The number of checking and savings accounts, in addition to income, is what Srivastava, Alpert, and Shocker (1984) examined to ascertain that their respondents did not differ from nonrespondents economically. Another reason to consider the number of checking and savings accounts is also that of diversification especially in the age of globali-

sation and internationalisation when people with greater socioeconomic wellbeing tend to have multiple accounts in banks located 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). While the majority of people earn income and have checking accounts, many may not own a home or life insurance, further differentiating socioeconomic classes. We will however weight some of these variables differently, see Section 9.1.

According to the philosophy of clustering applied in this paper, it has to be decided how "similarity" and "belonging together in the same class" are to be interpreted in social stratification. This is a difficult task that cannot be solved uniquely because of the controversy about the meaning of "social classes" and the various different conceivable uses of a social classification. However, the clarification of the required decisions on the statistical side may help with improving the understanding of what is implied by speaking of "classes".

3 Latent class clustering

This paper deals with the cluster analysis of data with continuous, ordinal and nominal variables. Denote the data $\mathbf{w}_1, \ldots, \mathbf{w}_n,\ \mathbf{w}_i = (\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i),\ \mathbf{x}_i \in I\!R^p,\ \mathbf{y}_i \in \mathcal{O}_1 \times \ldots \times \mathcal{O}_q,\ \mathbf{z}_i \in \mathcal{U}_1 \times \ldots \times \mathcal{U}_r,\ i = 1, \ldots, n$, 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 datasets is latent class clustering (LCC; Vermunt and Magidson 2002), where $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are modelled as i.i.d., 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 $\pi_h > 0\ \forall h,\ \sum_{h=1}^{k} \pi_h = 1,\ \varphi_{\mathbf{a},\Sigma}$ denotes the p-dimensional Gaussian density with mean vector $\mathbf{a}$ and covariance matrix Σ (which may be restricted), and $\sum_{y \in \mathcal{O}_j} \tau_{hj}(y) = \sum_{z \in \mathcal{U}_j} \tau_{h(q+j)}(z) = 1,\ \pi_h \geq 0, \tau_{hj} \geq 0 \forall h, j$.

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

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

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

where $\xi(y)$ is a score for the ordinal values y. This is based on the adjacent-category logit model (Agresti, 2002) as used in Vermunt and Magidson (2005a). The score can be assigned by use of background information, certain scaling methods (e.g., Gifi, 1990), or as standard Likert scores (i.e., numbers $1, 2, \ldots, |\mathcal{O}_j|$) for the ordered values. In some sense, through ξ, ordinal information is used at interval scale level, as in most techniques for ordinal data.

The parameters of (3.1) can be fitted by the method of maximum likelihood (ML) using the EM- or more sophisticated algorithms, and given estimators of the parameters (denoted by hats), points can be classified into clusters (in terms of interpretation identified with mixture components) by maximising the estimated posterior probability that observation $\mathbf{w}_i$ had been generated by mixture component h under a two-step model for (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 estimated by the Bayesian Information Criterion (BIC). A problem with the ML-estimator 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 zero). In the software package LatentGOLD (Vermunt and Magidson, 2005a) this is treated by computing a penalised ML-estimator maximising the sum of 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 penalised.

4 Dissimilarity based clustering

Assuming that a dissimilarity measure between observations exists (discussed in Section 6), a dissimilarity based clustering method that is suitable as an alternative to LCC is k-medoids (Kaufman and Rouseeuw, 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 too much memory for large datasets (as the one analysed here). Function clara in `cluster` is an approximative version for Euclidean distances that can be computed for much larger datasets. 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 estimate the number of clusters for k-medoids. Some of them are listed (though treated there in connection to 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 (see Section 7 for more discussion of them).

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 $\mathbf{w}_i \in C_h$. The ASW estimate k_{ASW} maximises $\frac{1}{n}\sum_{i=1}^{n} s(i,k)$.

Calinski and Harabasz index (CH) (Calinski and Harabasz, 1974). The CH estimate k_{CH} maximises the CH index $\frac{\mathbf{B}(k)(n-k)}{\mathbf{W}(k)(k-1)}$, where

$$\mathbf{W}(k) = \sum_{h=1}^{k} \frac{1}{|C_h|} \sum_{\mathbf{w}_i, \mathbf{w}_j \in C_h} d(\mathbf{w}_i, \mathbf{w}_j)^2, \quad \text{and}$$
$$\mathbf{B}(k) = \frac{1}{n} \sum_{i,j=1}^{n} d(\mathbf{w}_i, \mathbf{w}_j)^2 - \mathbf{W}(k).$$

In the original definition, assuming Euclidean distances, $\mathbf{B}(k)$ is the between-cluster means sum of squares and $\mathbf{W}(k)$ is the within-clusters sum of squares.

This index was quite successful in Milligan and Cooper (1985). Using squared dissimilarities in the index makes it more directly connected to k-means than to k-medoids. However, Milligan and Cooper (1985) used the index with various clustering methods some of which were not based on squared dissimilarities. CH can be more easily computed for large datasets than ASW because a complete dissimilarity matrix is not required.

Pearson version of Hubert's Γ (PH) The PH estimator k_Γ maximises 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. Hubert's Γ

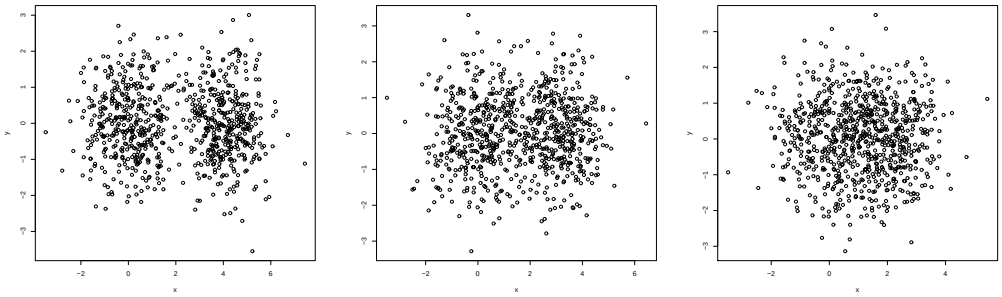

Figure 1: Two groups of points with variance 1 along the x-axis and mean difference of 4, 3 and 2.

(Baker and Hubert 1975) was originally defined in a way similar to the above definition, but with Goodman and Kruskal's rank correlation coefficient Γ instead of the Pearson correlation. The Pearson version (as proposed under the name "Normalised Γ" in Halkidi, Batistakis and Vazirgiannis, 2001) is used here for computational reasons.

5 Clustering philosophy

The main principle of the clustering philosophy discussed here is as follows. There are no unique objective "true" or "best" clusters in a dataset. 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 to match the data analytic features of the resulting clustering with the context-dependent requirements desired by the researcher. The present paper attempts to give the reader guidance on how to think about the necessary decisions.

Much of the literature in which clustering methods are introduced is based on not explicitly discussed assumptions about what the "true clusters" are (the issue is acknowledged in some books, e.g., Gordon 1999). In most cases, one or both of the following are assumed (often without realising that they are not always compatible).

1. Given a dataset, there is a "natural human intuition" of what the true clusters are.

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 (be it a mixture or fixed partition model). A less restrictive version of this,

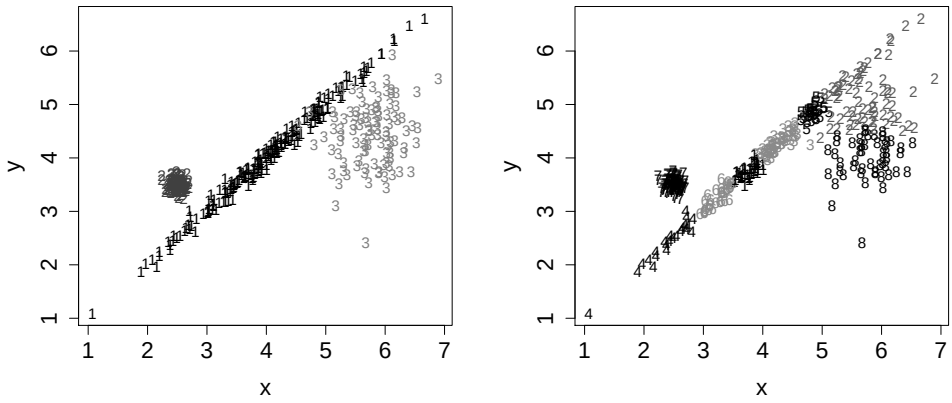

Figure 2: Artificial dataset from a 3-components Gaussian mixture (components indicated on left side) with optimal clustering (right side) from Gaussian mixture model according to the BIC with within-component covariance matrices restricted to be diagonal.

> which has not been applied yet to mixed type data and is therefore not further discussed here, is to assume that the data are generated by a distribution and clusters are identified with density modes or other features of it.

Neither of these make 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 in typical real situations with unknown truth, there is no agreement about it. Imagine two initially strongly separated populations and shift them closer together as in Figure 1. Obviously there is no clear intuition from when on precisely this is a single rather than two clusters. "Human clustering intuition" is normally Euclidean, and neither applies 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. It does not help to claim that the model may be "approximately true", because, e.g., in any reasonably defined neighbourhood of a mixture with two Gaussian components there are Gaussian mixtures (i.e., the latent class model (3.1) with $q = r = 0$) with arbitrarily larger numbers of components.

We now illustrate that 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 a researcher may be interested in.

The left side of Figure 2 shows a mixture of three Gaussian distributions with $p = 2$. In some applications, it makes sense to define the appropriate number of clusters as 3, with clusters corresponding to the mixture components. However, if the application is social stratification, and the variables are, e.g., income and a status indicator, this is not appropriate, because it would mean that the same social stratum (interpreted to correspond to mixture component no. 1) would contain the poorest people with lowest status as well as the richest people with the highest status. The cluster concept imposed by the Gaussian mixture model may correspond to the purely data based intuition of most people here, but it does not necessarily bring the most similar observations together, which, for some applications, may be inappropriate even if the Gaussian mixture model is true (more examples in which Gaussian mixture components are different from the cluster intuition of most people can be found in Lindsay and Ray, 2005 and Hennig, 2010).

One alternative to fitting a fully flexible Gaussian mixture is to assume a model in which all within-component covariance matrices are assumed to be diagonal, i.e., the (continuous) variables are assumed to be locally independent. This gives more sensible clusters for the dataset of Figure 2 (see right side) at least for social stratification, because clusters could no longer be characterised by features of dependence ("income increases with status" for component no. 1, comprising low income/low status as well as high income/high status) even though we know that this model is wrong for the simulated dataset. Assuming local independence in this sense decomposes dependence between variables into clusters. It still does not guarantee that observations are similar within clusters, because variances may differ between components.

Gaussian mixtures are very versatile in approximating almost any density, so that it is in fact almost irrelevant whether data have an obvious "Gaussian mixture shape" in order to apply clustering based on the Gaussian mixture model. 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 formalising and exploring what kind of cluster shape the method implies. Model assumptions are helpful for making this explicit. It could be objected that clustering is often (although not in social stratification) done in oder to find "true" underlying populations (Everitt et al., 2011), but even such populations can only be properly defined based on substantial interpretation and cannot be easily identified with straightforward statistical models.

Supposedly "model-free" approaches to clustering imply certain cluster shapes as well, which are often less obvious. Therefore model assumptions, despite of not being "true", are helpful because they are informative about the method's characteristics. In most applications, both model-based and supposedly model-free methods can be applied, so that their characteristics need to be understood and compared to the requirements of the application. For social stratification this will be done in Section 8.

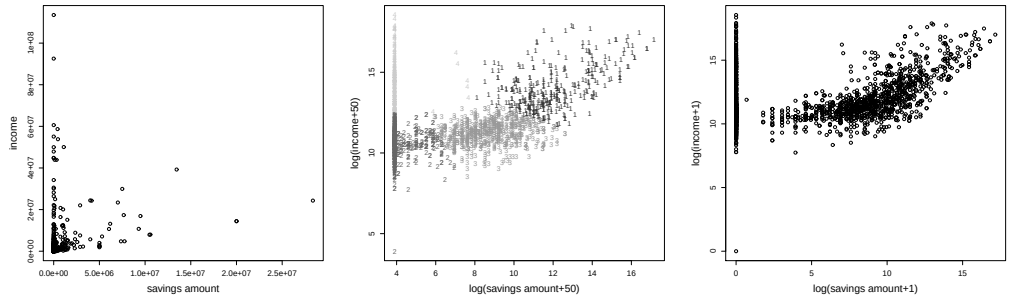

Figure 3: Subset of income and savings data from US Survey of Consumer Finances 2007, untransformed (left), with both variables log(x+50)-transformed (middle, with 4-medoids clustering), and with both variables log(x+1)-transformed.

In most applications including social stratification the researcher may not be able to justify all the required decisions in detail, and there is feedback between the desired cluster concept and the data, which, e.g., may confront the researcher with the fact that prior expectations about existing separation between clusters were unrealistic. Furthermore, it is hardly possible to anticipate the implications of all required tuning decisions for the given data. Therefore it is not enough to specify a cluster concept and to select a method. Clustering results need to be validated further, as illustrated in Section 9.

An important task related to the philoosophy presented here is the translation of the desired "interpretative meaning" of the clustering into data analytic characteristics of the clustering method. Such thinking is unfamiliar to most researchers, and even statisticians are rarely explicit about it. Therefore, nonstandard reasoning needs to be applied that cannot be backed up by literature, and significant subjective input is required.

In conclusion, the researchers should not hope that the data will give them the "true" clustering if only they choose the optimal methodology, because the choice of methodology defines implicitly what kinds of clusters will be found. Researchers need to define first how their idea of clusters can be formalised in a data analytic way. Apart from the choice of a clustering method, further thought is required for the definition of "dissimilarity" (Section 6), and the definition of a suitable number of clusters (Section 7).

6 Dissimilarity definition

To apply dissimilarity based methods and to measure whether a clustering method groups similar observations together, a formal definition of "dissimilarity" is needed.

For similar reasons as before, this is a highly nontrivial task that depends on decisions of the researcher, because the measure should reflect what is taken as "similar" in a given application (Hennig and Hausdorf, 2006).

The crucial task for mixed type data is how to aggregate and how to weight the different variables against each other. Variable-wise 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 $\mathbf{x}_i, \mathbf{x}_j$ on p continuous variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is

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

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:

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

Variable weights w_l can easily be incorporated in both aggregation schemes by multiplying the d_l (equivalent to multiplying the variables) 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 one 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 dataset 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 datasets with many observations is that many computations for Euclidean distances (such as clara and the CH index) do not need the handling of a full dissimilarity matrix. Therefore Euclidean aggregation is preferred here.

An alternative would be the Mahalanobis distance

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

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 a certain "lower dimensional direction" in data space, dissimilarities along these directions are downweighted and dissimilarities orthogonal to them are up-weighted. A key question

about the potential use of d_M is whether in a situation where several variables are correlated, the information in these variables (as far as it is consistent with the direction of correlation) should be down-weighted. This seems sensible if the correlation implies, in terms of interpretation, that much of the information in these variables is redundant, as, i.e., in genetic setups 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 that they should not be down-weighted by using d_M. In general this implies that in situations where the information in variables is "interpretatively independent", affine equivariance as enforced by d_M is not desirable, because it gives information purely statistical weight, as opposed to interpretative weight, which is increased if several substantially interesting variables point in the same direction (this also favours "local independence" in LCC compared to fully flexible covariance matrices).

The definition of d_E can be extended to mixed type variables in a similar way in which Gower extended d_G. Ordinal variables can be used with standard scores. Alternative scores can be used if available. There are various ways to assign scores, e.g., using quantiles of an assumed underlying Gaussian distribution, assigning average ranks, or based on other available variables using techniques such as monotonic regression. The effective distance between the categories of the ordinal variable should reflect their "interpretative distance", and the crucial question is whether such a data dependent scheme achieves this. "Estimating" scores from the other variables is not sensible here, because the information in those variables is used for the clustering directly already. 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. We do not see any reason why this reflects "interpretative distance" better than using standard scores in the given application. Particularly, both schemes fit the effective distribution to a certain homogeneous shape (the Gaussian, or, using average ranks, the uniform distribution), which runs counter to the idea that potential information regarding 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; technically only $(m_j - 1)$ binary variables would be needed, but this would lead to asymmetric treatment of the categories).

The variables then need to be weighted (or, equivalently, standardised) in order to make them comparable for aggregation. There are two aspects of this, the statistical aspect (the variables need 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). The rest of this section focuses on the statistical aspect, but further weighting is discussed in Section 9.1.

Before weighting, it is important to consider transformations of the continuous variables. The main rationale for transformation in the philosophy adopted here is that distances on the transformed variable should reflect the "interpretative distance" between cases.

For example, for the variables giving income and savings amount on the left side of Figure 3, log-transformations were applied. This was not mainly done because the resulting distributional pattern looks more healthy (particularly taming the outliers dominating the untransformed variables, see Figure 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" (which is our main rationale for the choice of a transformation) between incomes is rather governed by ratios than by absolute differences. The difference between two people with yearly incomes of \$2m and \$4m, say, should in terms of social strata be treated about equally than the difference between \$20,000 and \$40,000.

Because there are zeroes 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 nonzero savings/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 the middle of Figure 3 it can be seen that there is no clear gap anymore 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 an own cluster unless this is supported by other variables.

For $c = 1$ (right panel of Figure 3), clusterings are too strongly dominated by the "zero savings" group. The precise choice of 50 is subjective and cannot be motivated precisely, but clusterings with $c = 10$ and $c = 200$ had less desirable features. 50 still leaves a clear gap between people with zero and nonzero income, but the interpretative difference between zero and nonzero income is larger than regarding savings, particularly because the smallest nonzero income was considerable larger than low nonzero savings amounts.

There are various ways that standardisation can make the variation of continuous variables comparable, which comprise

- range standardisation (for example to $[0, 1]$),
- standardisation to unit variance,
- standardisation to unit interquartile range.

The main issue here is how the methods deal with extreme observations. Range standardisation is that this is governed by the two most extreme observations, and in presence of outliers this can mean that pairwise differences on such a variable between a vast majority of observations could be approximately zero and only the outliers are considerably far away from the rest. Using a robust statistic such as the interquartile range on the other hand 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.

In the present paper standardisation to unit variance is adopted, which is a compromise between the two approaches discussed before. The gap between no income and low but nonzero 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 have to be standardised so that the dissimilarities between their levels are properly comparable to the dissimilarities between observations on continuous variables. This can be better calibrated based on variances than based on quantiles. Furthermore, it makes interpretative sense that the most extreme dissimilarities in income and savings are allowed to dominate dissimilarities between categories (which would not be possible using range standardisation).

Nominal variables are considered next. Standardising continuous variables to unit variance implies $E(X_1 - X_2)^2 = 2$ for i.i.d. X_1, 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 standardised by setting $\sum_{i=1}^{I} E(Y_{i1} - Y_{i2})^2 = qE(X_1 - X_2)^2 = 2q$ with some factor q. The rationale for this is that in the Euclidean distance d_E, $\sum_{i=1}^{I}(Y_{i1} - Y_{i2})^2$ is aggregated with $(X_1 - X_2)^2$. It may seem natural to set $q = 1$, so that the expected contribution from the nominal variable equals, on average, the contribution from continuous variables with unit variance. However, if the resulting dissimilarity is used for clustering, there is a problem with $q = 1$, namely that, because of the fact that the distance between two identical categories is zero, it makes the difference between two different levels of the nominal variable larger than $E(X_1 - X_2)^2$, and therefore it introduces wide "gaps" (i.e., subsets between which there are large distances), which could force a cluster analysis method into identifying clusters with levels of the categorical variable. Therefore we recommend $q = \frac{1}{2}$, though larger values can be chosen if it is deemed, in the given application, that the nominal variables should carry higher weight (which is not the case here).

$q = \frac{1}{2}$ 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).

There is another consideration regarding the standardisation of the Y_i-variables so that

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

The expected value depends on the category probabilities. As a default, for standardisation these can be taken to be $\frac{1}{I}$ for each category. An obvious alternative is to estimate them from the data. This would increase the effective distance between categories with higher differences between the category probabilities. In the given setup 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, we suggest

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

as rationale for standardisation, which for binary variables ($I = 2$) is equivalent to (6.1), and $q \to 1$ for $I \to \infty$ means that with more levels the expected contribution converges toward that of a continuous variable (this implies that in terms of dissimilarity, it makes almost no difference whether in Section 9.1 the education variable with discrete values between 0 and 17 is treated as ordinal or continuous). The same considerations as above apply to the computation of the expected value.

To summarise, we define the dissimilarity by

- Euclidean aggregation of variables,
- suitable transformation of continuous variables,
- standardisation of continuous variables to unit variance,
- using I dummy variables for each nominal variable, standardised according to (6.1) with $q = \frac{1}{2}$,
- using standard coding for each ordinal variable, standardised according to (6.2), and
- additional weighting of variables according to substantial requirements.

7 The number of clusters

Finding an appropriate number of clusters k is a notoriously difficult issue in cluster analysis. In most of the literature the issue is treated as if it were clear that there is a "true" number of clusters that can be estimated. Following the resoning in Section 5, we do not agree. Finding k is rather a problem of an appropriate subject matter-dependent definition of the required number of clusters than an estimation

problem. Whereas choosing a cluster analysis method for given k requires rather qualitative judgements about cluster shapes of interest, choosing a method to find the best k given clusterings for given k requires rather quantitative judgements. Decisions involved are whether clusters are required to be separated by gaps and how clear a gap needs to be, whether gaps within a cluster can be tolerated, whether and how large dissimilarities can be tolerated within a cluster, how small a subset of the data is allowed to be in order to be regarded as a stand-alone cluster, whether and to what extent it is important that a clustering remains stable under small changes of the data, what would constitute a borderline situation for a data subset whether or not it should be split up into two clusters, and whether k should be in a certain range in order to be useful.

This suggests that the problem cannot be boiled down to the choice among several apparently "objective" criteria. It essentially requires user-tuning. There is no purely data-dependent way to decide from which distance downwards the two populations in Figure 1 should be interpreted as a single cluster instead of two.

In the LCC model (3.1), finding the true k is usually interpreted as a statistical estimation problem. Often the BIC, which penalises the log-likelihood for every k with $\frac{1}{2}\log(n)$ times the number of free parameters in the model is applied. It can be expected that the BIC estimates k consistently if the model really holds (proven for a simpler setup by Keribin 2000). However, the problem is in fact ill-posed. If the model does not hold precisely, the consistency result does not work as a justification of the BIC, because the true situation may be best approximated by very many mixture components if there are only enough observations, which is not of interpretative value. How useful the BIC is for deciding about the number of "interpretative clusters" depends on to what extent Gaussian mixture components (not required to be separated) used to approximate a dataset match the cluster concept of the researcher. "Penalised log-likelihood" as a general concept weighting the quality of fit against the complexity of the model makes sense, but giving the precise BIC penalty interpretative meaning is difficult (Biernacki, Celeux and Govaert, 2000, e.g., propose a criterion that penalises complexity in a stronger way; one may use Bayesian prior distributions for this aim, or try to generalise the methods for merging Gaussian mixture components in Hennig, 2010, to mixed type data).

The dissimilarity-based criteria introduced in Section 4 are attempts to formalise a "good number of clusters" directly. Several such criteria are proposed in the literature, and no clear justification exists which one is best. One reason for choosing ASW, CH and PH was that their definitions allow a more or less 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 emphasises separation between a cluster and their neighbouring clusters. In the definition of CH, the proportion of squared within-cluster dissimilarities is compared to all between-cluster dissimilarities, which emphasises within-cluster homogeneity more, and is through the use of squared dissimilarities more prohibitive against large within-cluster dissimilarities. PH emphasises 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 penalises large within-cluster dissimilarities strongly. Such data-analytic interpretations can be used to assess the methods based on certain requirements of the application (see Section 8), although it is rarely possible to connect the precise form of the index to the subject matter.

In real datasets, the situation is often not clear cut, and the criteria may yield vastly different solutions. For example, for the dataset shown in the middle of Figure 3, LCC/BIC yields 13 clusters to give a reasonable Gaussian mixture approximation, k-medoids with ASW yields 4 (shown) and with CH 16 clusters. ASW emphasises gaps between neighbouring clusters, and such gaps hardly exist. CH penalises large within-cluster distances stronger 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. All these indices cannot be used to assess $k = 1$ and 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.

The indices therefore do not make objective decisions, but rather serve as helpful indications. Due to better comparability for similar values of k, local optima are of interest as well as global ones, and looking at several criteria can be helpful to identify a solution that is good in several respects, although it should still be decided how well the indices can be related to the clustering aim.

In some situations in which formal and graphical analyses suggest that it is illusory to get a unique well justified value of k, one could fix the number of clusters at some value roughly seen to be useful. Usually the literature implies that fixing k means that its "true" value is known for some reason. This is hardly ever the case in practice, but estimating k in practice requires user tuning anyway, and therefore in ambivalent situations it may well be tuned directly.

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, however, cannot remove the requirement of user tuning, but only shift the decision problem to tuning parameters such as bandwidths.

8 Interpretative comparison of clustering methods

In this section, the previous considerations are applied to discussing the clustering methods regarding their suitability for social stratification. An obvious difficulty with this is that the requirements of social stratification are not clear enough to be directly "translated" into the formal framework in which the methods are defined. E.g., it is not possible to specify upper bounds for within-cluster dissimilarities.

It is controversial in the literature whether a stratification should rather have many small groups or a few larger ones. The main contention is whether the stratification structure is composed of a few big-classes or many micro-classes (Grusky, 2005; Weeden and Grusky, 2005; Weeden et al., 2007). For the reasons given in Section 7, 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 5 and 10. One reason is that for increasing k numerical issues such as local optima become more problematic for all clustering methods. 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. It is not a big problem that neither ASW nor CH nor PH can handle $k = 1$, but we will introduce a parametric bootstrap scheme to see whether the dataset is significantly clustered at all.

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 only suitable if it groups similar observations together. For most uses of social stratification (e.g., use as an explanatory variable in further studies, or for informing social policy) it is necessary to have descriptions of the strata that are meaningful outside the dataset that was used for deriving them. For this reason 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. Taking also the considerations in Section 5 regarding Figure 2 into account, we assume covariance matrices to be 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 one, cp. p. 4 of Bartholomew and Knott, 1999). This makes clustering scale invariant, so that the standardisation discussion of Section 6 is irrelevant for LCC.

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 distances are treated in the same way regardless of their direction. This means that the variable weight in k-medoids 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 socioeconomic 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 standardisation and weighting. On the other hand, 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 the ASW allow to check whether it does a good job regarding dissimilarity.

It would be possible to restrict covariance matrices even more in LCC, particularly to be equal and spherical, but this looks less attractive than using k-medoids or k-means, which implicitly have similar restrictions but can generalise them to the nominal amnd ordinal variables.

LCC as presented here assumes a cluster-wise 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 penalised heavily. Variables need to be transformed in such a way that distances can be interpreted symmetrically, which is done as in Section 6. Other distributional shapes (particularly with heavier tails) could be of interest but software incorporating this with mixed type data is not available to our knowledge.

In (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 this issue in detail, we acknowledge that there are several alternative approaches for modelling the effect of ordinality (see, e.g., Agresti, 2002), which incorporate different restrictions such as a certain stochastic order of mixture components (Agresti and Lang, 1993). The latter is difficult to justify in a multidimensional setting in which components may differ in a different way for different variables. The assumption of a cluster-wise underlying Gaussian distribution for the ordinal variables (Drasgow, 1986) may be worthwhile to explore. A reviewer noted that (3.2) could be tested against an unrestricted model (at least for homogeneous data), but using an unrestricted model would run counter to the presented philosophy, because it would prevent the ordinal meaning of the categories from influencing the clustering.

The methods introduced in Section 4 are more directly defined in order to make within-cluster dissimilarities small. There are many alternative dissimilarity based clustering methods in the literature (see, e.g., Kaufman and Rouseeuw, 1990, Gordon, 1999), for example the classical hierarchical ones, but most of these are unfeasible for too large datasets. Furthermore, 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, however, requires continuous variables to be categorised and seems more suitable with a larger number of categorical variables.

Compared to k-means clustering, k-medoids looks more attractive for mixed type data because the medoids, as opposed to mean vectors, are required to be members of the dataset, so no means are computed for nominal and ordinal categories

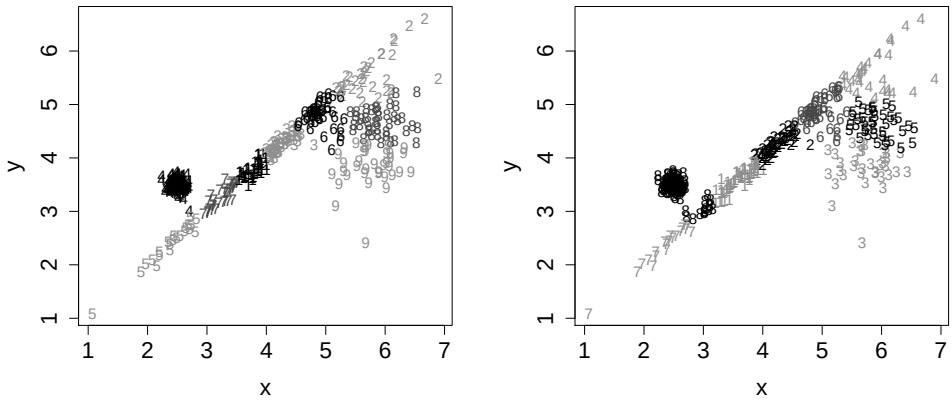

Figure 4: Artificial dataset with 9-medoids (left side) and 8-means clustering (right side), number of clusters chosen optimally according to the CH-criterion for both methods.

(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; although still distances in all "directions" are treated in the same way) and less affected by outliers within clusters (see Kaufman and Rouseeuw, 1990). The difference is illustrated in Figure 4. The k-medoids and k-means clustering (both with number of clusters optimised by the CH-criterion between 2 and 9) are similar (as often), but cluster 8 in the k-means solution seems to include too many points from the diagonal mixture component, causing a big gap within the cluster. The corresponding k-medoids cluster 4 only contains one of these points. This happens because the k-medoids criterion can tolerate larger within-cluster dissimilarities in cluster 5 (corresponding to cluster 7 in the k-means solution) without influencing the medoid too much, but the lower left point from the diagonal component influences the mean heavily. This favours k-medoids because in social stratification a small number of moderately outlying observations should be integrated in larger clusters without affecting the medoid too much.

Regarding the number of clusters, the main aim is not to approximate the dissimilarity structure by the clustering, and therefore PH is not of direct interest. ASW can be preferred to CH for similar reasons as those given in the k-medoids vs. k-means discussion. We still use all three criteria in Section 9 in order to investigate whether certain numbers of clusters stick out in various respects, and we do not only consider the global, but also local optima.

9 Data Analysis of US Consumer Finances data

9.1 Dataset, preprocessing and method tuning

In this section, we apply the methods to data from the 2007 US Survey of Consumer Finances (SCF). The survey was sponsored by the Board of Governors of the Federal Reserve System in cooperation with the Statistics of Income Division of the Income Revenue Service in the US, 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 nonrandom nonresponse. 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 dataset 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 which education and occupation data was available. There are obvious differences between males and females and in order to carry out gender comparisons, which are not the topic of this paper, it seems to be reasonable to analyse male and female data separately.

We used the following eight variables (see Section 2 for justifications):

lsam $\log(x+50)$ of total amount of savings as of the average of last month (treated as continuous),

linc $\log(x+50)$ of total income of 2006 (treated as continuous),

educ years of education between 0 and 17; this is treated as ordinal (level 17 means "graduate school and above"),

cacc number of checking accounts that one has; this is ordinal with 6 levels (corresponding to no/1/2/3/(4 or 5)/(6 or more) accounts; this is because precise differences between large number do not seem to matter here),

sacc number of savings accounts, coded as above,

hous housing, nominal with 9 levels: "neither owns nor rents", "inapplicable", "owns or is buying/land contract", "pays rent", "condo", "co-op", "townhouse association", "retirement lifetime tenancy" and "own only part",

life whether or not one has life insurance (binary),

occ occupation class, nominal with 7 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 United States Census Bureau 2006 4-digit occupation code. The detailed occupational categories were collapsed into the seven larger groups, a common practice, for the public version of the 2007 SCF.

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. In order 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 re-weighting all dummy variables in order to keep (6.1) valid. This increases the effective distance between categories 3 and 4 compared to all other distances.

The ordinal variables cacc and sacc were downweighted by a factor of 0.5 because the diversification aspect would be overrepresented with two fully weighted variables, and part of the information of these variables is shared with lsam and linc, particularly about the large number of 7,343 individuals without savings (there are only 5 "zero income" cases in the data). The life insurance indicator was downweighted by factor 0.5, too, because we see this as a comparably marginal indicator of the socioeconomic status.

As explained in Section 3, LCC depends on prior distributions. 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) interpret 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 visibly spurious local optima of the penalised likelihood; we tried to other prior default parameters further without improvement. Problems with spurious solutions could be handled by discretising the continuous variables and treating them as ordinal, but this would remove information in the continuous marginal distribution, which we considered as worthwhile.

LCC and k-medoids were applied to this dataset with k ranging from 2 to 20. We used `samples=100, sampsize=200` in R-function clara, which is slower but more stable than the default value. The LCC solution was computed by LatentGOLD. 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. In order to prevent this, we computed them from randomly partitioning the dataset into 15 subsets, computing them on every

		Criterion		
Method	k	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

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

subset, and averaging the results. This was much more stable than clara's internal computation of the ASW based on the optimal subset.

9.2 Clustering results with all variables

First we discuss the results involving all eight variables. The global optimum values of the criteria were assumed at the border of the range of k. ASW and CH were optimal for $k = 2$ for both k-medoids and LCC. The BIC was optimal for LCC for $k = 20$, potentially improving further above that value. Only PH was optimal for $k = 3$ for k-medoids.

The 2-medoids solution is fairly well interpretable, collecting 7,739 (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 indicators it is much better than the LCC solution with 2 clusters. This holds for all k-medoids solutions compared to LCC with the same k. Selected results can be seen in Table 1.

All three criteria ASW, CH and PH produce local optima for $k = 6$ and $k = 8$ for k-medoids. For LCC the 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 the ASW, the value for 8-medoids is slightly better than that for 6-medoids. The number of clusters 8 generally enables a fairly easy interpretation with enough differentiation, and 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 aim we chose the solution with $k = 8$ because this produces a local optimum of the ASW, and corresponds with the chosen k-medoids solution. The BIC-"optimal" solution $k = 20$ is too bad regarding the ASW to consider it.

One major difference between k-medoids and LCC for general k is that k-medoids

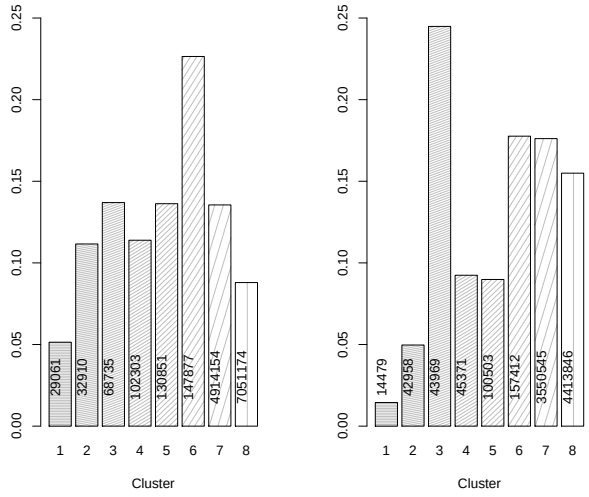

Figure 5: Relative frequency distributions by cluster with mean income values labeled (k-medoids solutions on the left and LCC solutions on the right).

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). Although social strata do not need to be of uniform size, too small classes are probably not of much interest.

9.3 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 Figure 5. The rationale for this operation is simple: income is the most obvious reward for individual in different socioeconomic classes.

Comparing the two clustering methods, the LCC method produced two larger higher-class clusters with relatively lower mean income while the k-medoids generated a greater buldge 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 to 6 comprises 61% of the cases. The k-medoids solution seems to reflect the real-world situation better.

Now we examine three other indicator variables of socioeoconomic stratification.

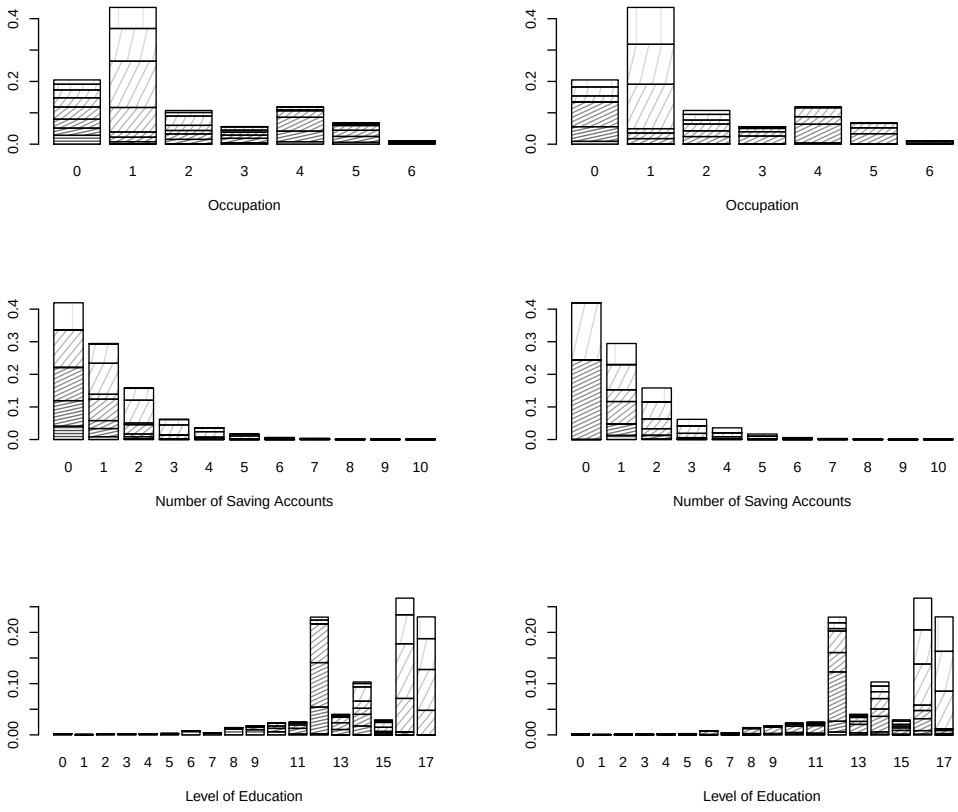

Figure 6: Relative frequency distributions of clusters by occupation (top row), number of savings accounts (middle row), and level of education (bottom row) (k-medoids solutions on the left and LCC solutions on the right).

Figure 6 presents relative frequency distributions of estimated clusters using both methods according to occupation, number of savings account, and level of education. The fill patterns in the figure follow Figure 5. That is, the lowest class is plotted with the densiest horizontal lines, the lines get sparser and the angle is raised counterclockwise by steps for higher classes, and the highest class is plotted with the sparsest vertical lines.

The largest occupation group (occupation group 1) is that of manegerials 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 (cf. Figure 5). The next largest group (occupation group 0), or not working for pay, is more or less 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 a bit less clear. According to the k-medoids solution, people who held a greater number of savings accounts are located in the highest two clusters but one. According to the LCC solution, people who held no savings account are only from clusters 3 and 7, 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 degree concentrate in clusters 3 to 6), and people with a high-school diploma only concentrate in clusters 3 to 5, according to the k-medoids solution. The distribution patterns according to the LCC solution are a bit less clearcut, with more people making into the top two clusters (7 and 8). The findings by examining cluster-specific conditional distributions of income, occupation, and education collectively point to a socioeconomic stratification system with eight classes with good differentiation, notably from the k-medoids solution.

9.4 A 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.

In order 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:

- For data generation (though not for data analysis), the nominal variables were treated as ordinal with categories ordered according to the average

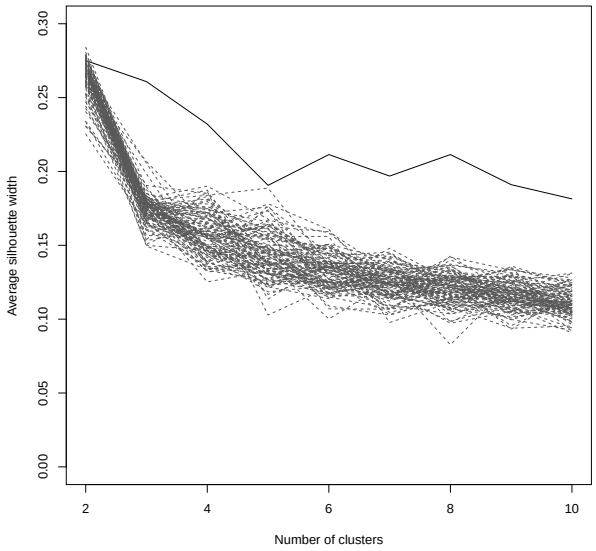

Figure 7: Average silhouette width for US consumer finances data, k-medoids, k between 2 and 10 (solid line), and 100 parametric bootstrap samples for a null model preserving the dependence structure.

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).

- A correlation matrix was computed using polychoric correlations between ordinal variables (Drasgow, 1986; the continuous variables were categorised into ten categories of approximately equal sizes for computing their correlations with the ordinal variables) and the Pearson correlation between the continuous variables.
- Data were generated according to a multivariate Gaussian distribution with the given correlation matrix.
- Ordinal categories were assigned, for the ordinal variables, using appropriate quantiles of the corresponding marginal distributions in order 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.

Hundred datasets of size 17,430 were drawn from this null model, and were clustered with k-medoids for k between 2 and 10 (ASW-values for the real dataset were bad above $k = 10$) in the same way as the real dataset. The corresponding ASW-values can be seen in Figure 7, along with those for the original data set.

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 only ran 20 bootstrapped datasets with $k = 8$, producing a range of ASW-values between 0.056 and 0.079, which are much smaller as well than the value of 0.136 from the original dataset.

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

9.5 Impact of variables

Variable impact was evaluated in order to check to what extent clusterings were dominated by certain (continuous or nominal) variables. In order to do this, the clustering methods were applied to all datasets 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 one obtained on the full dataset. The

	Variable impact (Rand)							
Method	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/8	0.302	0.460	0.720	0.920	0.711	0.759	0.769	0.650

Table 2: ARI between clustering on all variables (full dataset) and clustering with a variable omitted. Values near one mean that the variable has almost no impact.

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.

Among the categorical variables, occ has the strongest influence on LCC. The LCC results also show that the "message from the data" corresponds nicely to the decision to downweight cacc, sacc and life for dissimilarity definition, because they have a low impact despite not being downweighted in LCC.

Another feature of the 8-medoids result is that the clustering is almost unaffected by occupation. The impact of housing is low for both methods.

9.6 Comparing clustering and occupation groups

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

In order 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 (without the parametric bootstrap) was applied to a dataset in which the occ-variable was left out.

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

The 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 very low and close to the value 0 to be expected for totally unrelated clusterings, and indicate that the occupation grouping has almost nothing to do with the clusterings on the other variables.

9.7 Cluster validation

Whereas in some literature the term "cluster validation" is used almost synonymously for estimating the number of clusters by optimising some validation indices, 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.

A large amount of techniques can be used for this, including some methodology presented in the previous sections (particularly testing for homogeneity and the computation of various dissimilarity based indices). Because of space limitations, we only mention here that various other techniques have been applied in order to check interpretatively relevant aspects of the clusterings, including visualisation and comparison with further clusterings resulting from alternative methods or changing some tuning decisions (variable weighting, transformations, algorithmic parameters).

The ARI between LCC/8 and 8-medoids for the dataset 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 a lot when we used it as continuous, but in this case its discrete nature enforced many unstable local optima of the penalised 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.

10 Conclusion

We think that the application of automatic methods hoping that "the data will enforce its true structure" is deceptive. Many application-based decisions have to be made for clustering, particularly regarding weighting, transformation and standardisation of the variables. We hope that the present work can also inspire researchers applying cluster analysis in other fields to connect their choice of methodology more consciously to their cluster concept and their desired interpretation. The data analytic reasoning 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. Not all decisions can be made independently of the data. Even in situations where cluster analysis is only applied as an exploratory tool, such thinking is necessary in order to understand the meaning of whatever is found.

In the US Survey of Consumer Finances 2007 dataset, we found a well interpretable clustering and we could establish that it contains more structural information than just the dependence structure between the data, and in this sense it is a meaningful clustering. However, selecting one clustering as "best" requires subjective decisions and different clusterings could be justified by similar arguments.

More research is required concerning some of the involved decisions, a better systematic characterisation of the behaviour of the methods, and a more comprehensive treatment of ordinal variables.

References

Agresti, A. (2002) *Categorical Data Analysis.* Second Edition, Wiley, New York.

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. *Journal of the American Statistical Association* 70, 31-38.

Bartholomew, D. J. and Knott, M. (1999) *Latent Variable Models and Factor Analysis*, Kendall Library of Statistics, Vol. 7, 2nd Edition, Arnold, London.

Bernheim, B. D., Garrett, D. M. and Maki, D. M. (2001) Education and Saving: The Long-Term Effects of High School Financial Curriculum Mandates. *Journal of Public Economics* 80, 435-464.

Biernacki, C., Celeux, G. and Govaert, G. (2000) "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22, pp. 719-725.

Blau, P. M. and Duncan, O. D. (1967) *The American Occupational Structure*, Wiley, New York.

Brennan, M. J. and Schwartz, E. S. (1976) The Pricing of Equity-Linked Life Insurance Policies with an Asset Value Guarantee. *Journal of Financial Economics* 3, 195-213.

Calinski, R. B., and Harabasz, J. (1974) A Dendrite Method for Cluster Analysis, *Communications in Statistics*, 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. Inn S. Kotz and N. Johnson (eds.) *The Encyclopedia of Statistics*, Volume 7. Wiley, New York, 68-74.

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases. *Proceedings 1996 International Conference on Knowledge Discovery and Data Mining (KDD'96)*, 226231.

Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011) *Cluster Analysis* (5th ed.). Wiley, New York.

Gifi, A. (1990) *Nonlinear Multivariate Analysis*. Wiley, Chichester.

Goodman, L. A. (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61, 215-231.

Gordon, A. D. (1999) *Classification* (2nd ed.). Chapman & Hall/CRC, Boca Raton.

Gower, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrics* 27, 857-871.

Grusky, D. B. (with Galescu, G.). (2005) Foundations of class analysis: A Durkheimian perspective. In: E.O. Wright (Ed.), *Approaches to class analysis*, Cambridge University Press, Cambridge, 51-81.

Grusky, D. B., Ku, M. C. and Szelnyi, S. (2008) *Social Stratification: Class, Race, and Gender in Sociological Perspective.* Westview, Boulder, CO.

Grusky, D. B. and Weeden, K. A. (2008) Measuring Poverty: The Case for a Sociological Approach. In: Kakwani, N. and Silber, J. (eds.): *Many Dimensions of Poverty*, Palgrave-Macmillan, New York, 20-35.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001) On Clustering Validation Techniques, *Journal of Intelligent Information Systems* 17, 107-145.

Hennig, C. (2007) Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis* 52, 258-271.

Hennig, C. (2009) A Constructivist View of the Statistical Quantification of Evidence. *Constructivist Foundations* 5, 39-54.

Hennig, C. (2010) Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification* 4, 3-34.

Hennig, C. and Hausdorf, B. (2006) Design of dissimilarity measures: a new dissimilarity measure between species distribution ranges. In: Batagelj, V.; Bock, H.-H.; Ferligoj, A.; Ziberna, A. (eds.): *Data Science and Classification.* Springer, Berlin, 29-38.

Hollingshead, A. B. (1957) Two factor index of social position. Yale University Mimeograph, New Haven: CT.

Hubert, L. and Arabie, P. (1985), Comparing Partitions, *Journal of Classification* 2, pp. 193-218.

Jha, A. (2011) The mathematical law that shows why wealth flows to the 1

Kaufman, L. and Rouseeuw, P. J. (1990) *Finding Groups in Data*, Wiley, New York.

Kennickell, A. B. (2000) Wealth Measurement in the Survey of Consumer Finances: Methodology and Directions for Future Research. Working paper (May), `http://www.federalreserve.gov/pubs/oss/oss2/method.html`

Keribin, C. (2000) Consistent estimation of the order of a mixture model, *Sankhya A,* 62, pp. 49-66.

Kingston, P. W. (2000) *The Classless Society.* Stanford University Press, Stanford, CA.

Lenski, G. E. (1954) Status Crystallization: A Non-Vertical Dimension of Social Status. *American Sociological Review* 19, 405-413.

Le Roux B. and Rouanet H. (2010) *Multiple Correspondence Analysis.* SAGE, Thousand Oaks (CA).

Levy, F., and Michel, R. C. (1991) *The Economic Future of American Families: Income and Wealth Trends.* Urban Institute Press, Washington DC.

Liao, T. F. (2006) Measuring and Analyzing Class Inequality with the Gini Index Informed by Model-Based Clustering. *Sociological Methodology* 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, pp. 159-179.

Pekkanen, J., Tuomilehto, J., Uutela, A., Vartiainen, E., Nissinen, A. (1995) Social Class, Health Behaviour, and Mortality among Men and Women in Eastern Finland. *British Medical Journal* 311, 589-593.

Poterba, J. M., Venti, S. F. and Wise, D. A. (1994) Targeted Retirement Saving and the Net Worth of Elderly American. *American Economic Review* 84, 180-185.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, `http://www.R-project.org`

Ray, S. and Lindsay, B. G. (2005) The topography of multivariate normal mixtures. *Annals of Statistics* 33, 2042-2065.

Spilerman, S. (2000) Wealth and Stratification Process. *Annual Review of Sociology*, 26, 497-524.

Srivastava, R. K., Alpert, M. I. and Shocker, A. D. (1984) A Customer-Oriented Approach for Determining Market Structures. *Journal of Marketing* 84, 32-45.

Sugar, C. A. and James, G. M. (2003) Finding the Number of Clusters in a Dataset: an Information-Theoretic Approach. *Journal of the American Statistical Association* 98, 750-763.

Tukey, J. W. (1962) The Future of Data Analysis, *The Annals of Mathematical Statistics*, 33, 1-67.

Vermunt, J. K., and Magidson, J. (2002) Latent class cluster analysis. In: J.A. Hagenaars and A.L. McCutcheon (eds.), *Applied Latent Class Analysis*, Cambridge University Press, Cambridge, 89-106.

Vermunt, J. K. and Magidson, J. (2005a) *Technical Guide for Latent GOLD 4.0: Basic and Advanced.* Statistical Innovations Inc, Belmont Massachusetts.

Vermunt, J. K. and Magidson, J. (2005b) *Latent GOLD 4.0 User's Guide.* Statistical Innovations Inc, Belmont Massachusetts.

von dem Knesebeck, O. (2002) Social Inequality and Health of the ElderlyClassical or Alternative Status Indicator? *Zeitschrift für Gerontologie und Geriatrie* 35, 224-231.

Weeden, K. A., and Grusky, D. B. (2005) The Case for a New Class Map. *American Journal of Sociology* 111, 141-212.

Weeden, K. A., Kim, Y-M, Matthew, D. C., and Grusky, D. B. (2007) Social Class and Earnings Inequality. *American Behavioral Scientist* 50, 702-736.

Weisbrod, B. A., and Hansen, W. L. (1968) An Income-Net Worth Approach to Measuring Economic Welfare. *American Economic Review* 58, 1315-1329.

Wright, E. O. (1985) *Classes.* Verso, London.

Wright, E. O. (1997) *Class Counts: Comparative Studies in Class Analysis.* Cambridge University Press, Cambridge.