SPSS TwoStep Cluster - a first evaluation

Bacher, Johann; Wenzig, Knut; Vogler, Melanie

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

LEHRSTUHL FÜR
SOZIOLOGIE

Arbeits- und Diskussionspapiere

SPSS TwoStep Cluster
– A First Evaluation

**Johann Bacher, Knut Wenzig,
Melanie Vogler**

Arbeits- und Diskussionspapiere 2004-2, 2., korr. Aufl.

**Wirtschafts- und Sozialwissenschaftliche Fakultät
Friedrich-Alexander-Universität Erlangen-Nürnberg**

LEHRSTUHL FÜR
SOZIOLOGIE

Arbeits- und Diskussionspapiere

SPSS TwoStep Cluster
– A First Evaluation

Johann Bacher, Knut Wenzig,
Melanie Vogler

Arbeits- und Diskussionspapiere 2004-2, 2., korr. Aufl.

Authors' note: The study was supported by the Staedtler Stiftung Nürnberg (Project: „Was leisten Clusteranalyseprogramme? Ein systematischer Vergleich von Programmen zur Clusteranalyse"). The paper has been presented at RC33 Sixth International Conference on Social Science Methodology: "Recent Developments and Applications in Social Research Methodology", Amsterdam, The Netherlands, August 16–20, 2004 and is submitted for publication to the "Journal of Statistical Software" (http://www.jstatsoft.org). We thank Bettina Lampmann-Ende for her help with the English version.

Friedrich-Alexander-Universität Erlangen-Nürnberg
Wirtschafts- und Sozialwissenschaftliche Fakultät
Sozialwissenschaftliches Institut
Lehrstuhl für Soziologie
2004

Arbeits- und Diskussionspapiere
des Lehrstuhls für Soziologie

Bacher, Johann, Knut Wenzig & Melanie Vogler:
SPSS TwoStep Cluster – A First Evaluation
Arbeits- und Diskussionspapiere 2004-2, 2., korr. Aufl.

Friedrich-Alexander-Universität Erlangen-Nürnberg
Lehrstuhl für Soziologie

Findelgasse 7/9
90402 Nürnberg
Postanschrift: Postfach 3931, 90020 Nürnberg

Telefon: 0911/5302-679
Telefax: 0911/5302-660

E-Mail: soziologie@wiso.uni-erlangen.de
http://www.soziologie.wiso.uni-erlangen.de

Lehrstuhlsignet: Eva Lambracht. Gesetzt mit LaTeX.

3,– €

Abstract

SPSS 11.5 and later releases offer a two step clustering method. According to the authors' knowledge the procedure has not been used in the social sciences until now. This situation is surprising: The widely used clustering algorithms, k-means clustering and agglomerative hierarchical techniques, suffer from well known problems, whereas SPSS TwoStep clustering promises to solve at least some of these problems. In particular, mixed type attributes can be handled and the number of clusters is automatically determined. These properties are promising. Therefore, SPSS TwoStep clustering is evaluated in this paper by a simulation study.

Summarizing the results of the simulations, SPSS TwoStep performs well if all variables are continuous. The results are less satisfactory, if the variables are of mixed type. One reason for this unsatisfactory finding is the fact that differences in categorical variables are given a higher weight than differences in continuous variables. Different combinations of the categorical variables can dominate the results. In addition, SPSS TwoStep clustering is not able to detect correctly models with no cluster solutions. Latent class models show a better performance. They are able to detect models with no underlying cluster structure, they result more frequently in correct decisions and in less biased estimators.

Key words:
SPSS TwoStep clustering, mixed type attributes, model based clustering, latent class models

Zusammenfassung

SPSS enthält seit Version 11.5 einen Algorithmus zur TwoStep-Clusteranalyse. Dieses Verfahren wurde in den Sozialwissenschaften unseres Wissens nach bisher nicht angewendet. Das ist eigentlich überraschend: Die weit verbreiteten Verfahren der Clusteranalyse, wie k-means und agglomerative hierarchische Verfahren, haben bekannte Schwächen, für die SPSS TwoStep Clustering wenigstens teilweise eine Lösung verspricht: Insbesondere sollen gemischt-skalierte Variablen erlaubt sein und die Anzahl der Cluster automatisch bestimmt werden. Aus diesem Grund wird der neue Algorithmus in diesem Papier mit einer Simulationsstudie evaluiert.

SPSS TwoStep ist erfolgreich, wenn die Variabeln quantitativ sind. Für gemischt-skalierte Variablen sind die Ergebnisse jedoch weniger zufriedenstellend. Ein Grund hierfür ist, dass nominalen Variabeln in der Analyse höher gewichtet werden und so verschiedene Variablen-Kombinationen die Ergebnisse dominieren können. Weiterhin findet SPSS TwoStep Cluster, selbst wenn den Daten keine Clusterstruktur zugrunde liegt. Modelle mit latenten Klassen führen hier zu besseren Ergebnissen. Sie erkennen Situationen, in denen keine Clusterstruktur vorliegt, treffen häufiger die richtige Clusterzahl und führen zu weniger verzerrten Schätzern.

Contents

List of Tables

1 Introduction

SPSS 11.5 and later releases offer a two step clustering method (SPSS 2001, 2004). According to the authors' knowledge the procedure has not been used in the social sciences until now. This situation is surprising: The widely used clustering algorithms, k-means clustering and agglomerative hierarchical techniques, suffer from well known problems (for instance, Bacher 2000: 223; Everitt et al. 2001: 94-96; Huang 1998: 288), whereas SPSS TwoStep clustering promises to solve at least some of these problems. In particular, mixed type attributes can be handled and the number of clusters is automatically determined. These properties are promising. Therefore, SPSS TwoStep clustering will be evaluated in this paper. The following questions will be analyzed:

1. How is the problem of commensurability (different scale units, different measurement levels) solved?

2. Which assumptions are made for models with mixed type attributes?

3. How well does SPSS TwoStep – especially the automatic detection of the number of clusters – perform in the case of continuous variables?

4. How well does SPSS TwoStep – especially the automatic detection of the number of clusters – perform in the case of variables with different measurement levels (mixed type attributes)?

The model of SPSS TwoStep clustering will be described in the next section. The evaluation will be done in section 3. Section 4 will give concluding remarks.

2 SPSS TwoStep Clustering

SPSS TwoStep clustering was developed by Chiu, Fang, Chen, Wang and Jeris 2001 for the analysis of large data sets. The procedure consists of two steps (Chiu et al. 2001, SPSS 2004):

Step 1: Pre-clustering of cases. A sequential approach is used to pre-cluster the cases. The aim is to compute a new data matrix with fewer cases for the next step. In order to reach this aim, the computed pre-clusters and their characteristics (cluster features) are used as new cases. The pre-clusters are defined as dense regions in the analyzed attribute-space. The number of pre-clusters depends on three parameters MXBRANCH (default value = 8), MXLEVEL (default value = 3) and INITTRESHOLD (default value = 0). The maximal number is $\text{MXBRANCH}^{\text{MXLEVEL}}$ ($8^3 = 512$). The number of pre-clusters can be smaller if a high value for INITTRESHOLD is specified. The results may depend on the input order of cases. Therefore, SPSS (2001: 2) recommends to use a random order.

Step 2: Clustering of cases. A model based hierarchical technique is applied. Similar to agglomerative hierarchical techniques, the (pre-)clusters are merged stepwise until all clusters are in one cluster. In contrast to agglomerative hierarchical techniques, an underlying statistical model is used. The model assumes that the continuous variables x_j $(j = 1, 2, \ldots, p)$ are within cluster i independent normal distributed with means μ_{ij} and variances σ^2_{ij} and the categorical variables a_j are within cluster i independent multinominal distributed with probabilities π_{ijl}, where (jl) is the index for the l-th category $(l = 1, 2, \ldots, m_l)$ of variable a_j $(j = 1, 2, \ldots, q)$.

Two distance measures are available: Euclidean distance and a log-likelihood distance. The log-likelihood distance can handle mixed type attributes. The log-likelihood distance between two clusters i and s is defined as:

$$d(i, s) = \xi_i + \xi_s - \xi_{\langle i,s\rangle} \tag{2.1}$$

where

$$\xi_i = -n_i \left(\sum_{j=1}^{p} \frac{1}{2} \log(\hat{\sigma}^2_{ij} + \hat{\sigma}^2_j) - \sum_{j=1}^{q} \sum_{l=1}^{m_j} \hat{\pi}_{ijl} \log(\hat{\pi}_{ijl}) \right) \tag{2.2}$$

$$\xi_s = -n_s \left(\sum_{j=1}^{p} \frac{1}{2} \log(\hat{\sigma}^2_{sj} + \hat{\sigma}^2_j) - \sum_{j=1}^{q} \sum_{l=1}^{m_j} \hat{\pi}_{sjl} \log(\hat{\pi}_{sjl}) \right) \tag{2.3}$$

$$\xi_{\langle i,s\rangle} = -n_{\langle i,s\rangle} \left(\sum_{j=1}^{p} \frac{1}{2} \log(\hat{\sigma}^2_{\langle i,s\rangle j} + \hat{\sigma}^2_j) - \sum_{j=1}^{q} \sum_{l=1}^{m_j} \hat{\pi}_{\langle i,s\rangle jl} \log(\hat{\pi}_{\langle i,s\rangle jl}) \right) \tag{2.4}$$

ξ_v can be interpreted as a kind of dispersion (variance) within cluster v $(v = i, s, \langle i, s\rangle)$. ξ_v consists of two parts. The first part $-n_v \sum \frac{1}{2} \log(\hat{\sigma}^2_{vj} + \hat{\sigma}^2_j)$ measures the dispersion of the continuous variables x_j within cluster v. If only $\hat{\sigma}^2_{vj}$ would be used, $d(i, s)$ would be exactly the decrease in the log-likelihood function after merging cluster i and s. The term $\hat{\sigma}^2_j$ is added to avoid the degenerating situation for $\hat{\sigma}^2_{vj} = 0$. The entropy $-n_v \sum_{j=1}^{q} \sum_{l=1}^{m_j} \hat{\pi}_{vjl} \log(\hat{\pi}_{vjl})$ is used in the sec-

ond part as a measure of dispersion for the categorical variables.

Similar to agglomerative hierarchical clustering, those clusters with the smallest distance $d(i, s)$ are merged in each step. The log-likelihood function for the step with k clusters is computed as

$$l_k = \sum_{v=1}^{k} \xi_v. \tag{2.5}$$

The function l_k is not the exact log-likelihood function (see above). The function can be interpreted as dispersion within clusters. If only categorical variables are used, l_k is the entropy within k clusters.

Number of clusters. The number of clusters can be automatically determined. A two phase estimator is used. Akaike's Information Criterion

$$AIC_k = -2l_k + 2r_k \tag{2.6}$$

where r_k is the number of independent parameters or Bayesian Information Criterion

$$BIC_k = -2l_k + r_k \log n \tag{2.7}$$

is computed in the first phase. BIC_k or AIC_k result in a good initial estimate of the maximum number of clusters (Chiu et al. 2001: 266). The maximum number of clusters is set equal to number of clusters where the ratio BIC_k/BIC_1 is smaller than c_1 (currently $c_1 = 0.04$)[1] for the first time (personal information of SPSS Technical Support). In table 2.1 this is the case for eleven clusters.

The second phase uses the ratio change $R(k)$ in distance for k clusters, defined as

$$R(k) = d_{k-1}/d_k, \tag{2.8}$$

where d_{k-1} is the distance if k clusters are merged to $k-1$ clusters. The distance d_k is defined similarly.[2] The number of clusters is obtained for the solution where a big jump of the ratio change occurs.[3]

The ratio change is computed as

$$R(k_1)/R(k_2) \tag{2.11}$$

for the two largest values of $R(k)(k = 1, 2, \ldots, k_{max}$; k_{max} obtained from the first step). If the ratio change is larger than the threshold value c_2 (currently $c_2 = 1.15$[4]) the number of clusters is set equal to k_1, otherwise the number of clusters is set equal to the solution with $\max(k_1, k_2)$. In table 2.1, the two largest values of $R(k)$ are reported for three clusters ($R(3) = 2.129$; largest value) and for eight clusters ($R(8) = 1.952$). The ratio is 1.091 and smaller than the threshold value of 1.15. Hence the maximum of 3 resp. 8 is selected as the best solution.

Cluster membership assignment. Each object is assigned deterministically to the closest cluster according to the distance measure used to find the clusters. The deterministic assignment may result in biased estimates of the cluster profiles if the clusters overlap (Bacher 1996: 311–314, Bacher 2000).

Modification. The procedure allows to define an outlier treatment. The user must specify a value for the fraction of noise, e.g. 5 (=5 %). A leaf (pre-cluster) is considered as a potential outlier cluster if the number of cases is less than the defined fraction of the maximum cluster size. Outliers are ignored in the second step.

1 The value is based on simulation studies of the authors of SPSS TwoStep Clustering. (personal information of SPSS Technical Support, 2004-05-24)

2 The distances d_k can be computed from the output in the following way:

$$d_k = l_{k-1} - l_k \tag{2.9}$$

$$l_v = (r_v \log n - BIC_v)/2 \quad \text{or} \quad l_v = (2r_v - AIC_v)/2 \quad \text{for} \quad v = k, k-1 \tag{2.10}$$

However, using BIC or AIC results in different solutions.

3 The exact decision rules are described vaguely in the relevant literature and the software documentation (SPSS 2001; Chiu et al. 2001). Therefore, we report the exact decision rule based on personal information of SPSS Technical Support. A documentation in the output, like "solution x was selected because ...", would be helpfull for the user.

4 Like c_1, c_2 is based on simulation studies of the authors of SPSS TwoStep Clustering. (personal information of SPSS Technical Support, 2004-05-24)

Table 2.1: SPSS TwoStep auto clustering results

	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]	
1	81490.274				
2	56586.953	−24903.320	1.000	1.467	
3	39624.406	−16962.547	.681	2.129	← Maximum ratio of distance
4	31686.789	−7937.617	.319	1.343	
5	25792.248	−5894.541	.237	1.010	
6	19955.794	−5836.454	.234	1.745	
7	16636.600	−3319.194	.133	1.177	
8	13825.500	−2811.100	.113	1.952	← SPSS decision (see text)
9	12414.105	−1411.396	.057	1.014	
10	11022.935	−1391.169	.056	1.036	max. number
11	9682.323	−1340.612	.054	1.755	← of clusters
12	8943.800	−738.523	.030	1.005	in phase 1

a The changes result from the previous number of clusters in the table.
b The ratio of changes is with respect to the change at the two clusters.
c The ratio of distance measures is based on the current number of clusters in relation to the previous number of clusters.

Note: SPSS TwoStep computes either BIC or AIC. If both measures are needed the procedure must be run two times. The ratio of distance measure is equal for BIC and AIC.

Output. Compared to k-means algorithm (QUICK CLUSTER) or agglomerative hierarchical techniques (CLUSTER), SPSS has improved the output significantly. An additional modul allows to statistically test the influence of variables on the classification and to compute confidence levels.

3 Evaluation

3.1 Commensurability

Clustering techniques (k-means-clustering, hierarchical techniques etc.) require commensurable variables (for instance, Fox 1982). This implies interval or ratio scaled variables with equal scale units. In the case of *different scale units,* the variables are usually standardized by the range (normalized to the range [0,1], range weighted) or z-transformed to have zero mean and unit standard deviation (autoscaling, standard scoring, standard deviation weights). If the variables have *different measurement levels,* either a general distance measure (like Gower's general similarity measure; Gower 1971) may be used or the nominal (and ordinal) variables may be transformed to dummies and treated as quantitative[1] (Bender et al. 2001; Wishart 2003).

SPSS TwoStep clustering offers the possibility to handle continuous and categorical variables. Hence, SPSS TwoStep cluster model only provides a solution for a special case of variables of mixed type. Quantitative variables with different scale units and nominal scaled variables may be simultaneously analyzed. The user must decide to handle ordinal variables either as continuous or as categorical if they are present.

Continuous variables are z-standardized by default in order to make them commensurable. This specification is not the consequence of the statistical model used in SPSS TwoStep. Hence, other procedure can be used too. Z-standardization may be appropriate in some applications, in others not (Bacher 1996: 173–198; Everitt 1981: 10; Everitt et al. 2001: 51–52). If random errors are the reasons of larger variances, z-standardization has a positive effect: Variables with large proportions of random errors are given lower weights. However, if differences between clusters result in larger variances, z-standardization has a negative effect: Variables that separate the clusters well are given lower weights. In empirical studies the reasons for high variances are unknown. Probably, both reasons apply so that both effects cancel out. Simulation studies suggest, that z-standardization is ineffective (for a summary of simulation results; see Everitt et al. 2001: 51). Better results are reported for standardization to unit range (ibidem). However, standardization to unit range is problematic for mixed type attributes (see below).

In the case of *different measurement levels*, the distance measures for the different kinds of variables must be normalized in order to make them commensurable. The log-likelihood distance uses the following normalization. If two objects i and s differ only in one categorical variable and are merged to one cluster $\langle i, s\rangle$, the log-likelihood distance is 0.602. This distance corresponds to a difference of 2.45 scale units (=standard deviation) if two objects differ in one standardized continuous variable.[2] Hence, *a difference in one categorical variable is equal to a difference of 2.45 scale units in one standardized continuous variable.* This normalization may favor categorical variables to define clusters (see below). This effect would even be stronger for standardization by range: The maximum difference between two objects that differ only in one continuous vari-

1 The term "quantitative" will be used for interval or ratio scaled variables.

2 According to formula 2.1, the log-likelihood distance $d(i, s)$ is equal to

$$d(i,s) = 0 + 0 - \left(-2 \left(\underbrace{0}_{\text{quantitative var.}} - (0.5\log(0.5) + 0.5\log(0.5)) \right) \right) = 0.6021, \tag{3.1}$$

able is 0.176[3], whereas the maximum difference between two objects that differ only in one categorical variable is 0.602. Hence, a difference in one categorical variable is given a three times higher weight than the maximum difference in continuous variables that are standardized by range. This disadvantage seems to be more serious than the above mentioned better perfomance of standardization by range. Therefore, z-standardization seems an acceptable approach.

Table 3.1 compares SPSS TwoStep clustering with other clustering procedures. K-means implementation and latent class models resp. latent probabilistic clustering models[4] are used. Compared to k-means-implementations, SPSS TwoStep allows to handle continuous and categorical variables.[5] Compared to latent class models, SPSS TwoStep performs worse. Both latent models are able to handle all kind of measurement levels. Different scale units cause no problems, they are model based transformed.

3.2 Automatic determination of the number of clusters

Chiu et al. (2001) report excellent results for the proposed algorithm to determine the number of clusters automatically. For about 98 % of the generated data sets (more than thousands), SPSS TwoStep clustering was able to find the correct number of clusters. For the rest the clusters were indistinguishable due to much overlap.

Data sets with overlapping clusters are characteristic for the social sciences. Therefore, we analyzed artificial datasets with overlap. Five different cluster models were studied (see table 3.2). One model consists of one cluster. Two models consist of two clusters with different overlap. The fourth model consists of three clusters and the fifth model consists of five clusters.

Two sets of variables were defined for each model: one data set with three independent variables and one with six independent variables. A standard normal distribution was assumed for all variables. Two (resp. four) of the variables were categorized for the analysis. A category value of 0 was assigned if the values of the variable were less than or equal to -1; if the values were greater than -1 and less than $+1$, a category value of 1 was assigned; and if the value was greater than or equal to $+1$, a category value of 2 was assigned. One of the two (resp. two of the four) variables were treated as nominal variables. The other categorical variable(s) were treated as ordinal variable(s). In order to test the effect of categorization and

if two objects i and s that differ only in one nominal variable are combinded. (ξ_i (formula 2.2) and ξ_s (formula 2.3) are zero, because "clusters" i and s contain only one object.) This corresponds to a difference of

$$+2\left(\frac{1}{2}\log\left(\hat{\sigma}^2_{\langle i,s\rangle j}+\hat{\sigma}^2_j\right)\right)=0.6021 \Leftrightarrow 10^{0.6021}=4=\hat{\sigma}^2_{\langle i,s\rangle j}+\underbrace{\hat{\sigma}^2_j}_{1} \Leftrightarrow 3=\underbrace{(x_{ij}-x_{sj})^2/2}_{\hat{\sigma}^2_{\langle i,s\rangle j}} \Leftrightarrow \pm 2.45 = x_{ij}-x_{sj}, \tag{3.2}$$

if $n-1$ is used to estimate the variances. If n is used, $x_{ij}-x_{sj}=\pm 1.73$. The different weighting of differences in categorical and continuous variables can be illustrated by the following example, too: If a standard normal distribution N(0,1) is dichotomized at point 0, the average distance between two randomly selected objects is 0.24 if the variable is treated as continuous. If the variable is treated as categorical, the average distance between two randomly selected objects is 0.30 and 1.25 times higher than the value if the variable is specified as continuous. Intuitively, one would expect that both values are equal.

3 The value of $d(i,s)$ is equal to $+2(\frac{1}{2}\log(\hat{\sigma}^2_{\langle i,s\rangle}+\hat{\sigma}^2_j))=\log(+0.5+1)=0.1761$, if we set $\hat{\sigma}^2_j=1$ in order to guarantee positive values for $d(i,s)$. If n is used to estimate the variances, $d(i,s)$ is equal 0.301.

4 Bacher (2000) proves that latent class models can be interpreted as probabilistic clustering models. In contrast to deterministic clustering techniques, cases are assigned to the cluster probabilistically.

5 Note: This comparision refers only to the solutions of the problem of commensurabilty for the different k-means implementations. Hence, this comparisions is not a benchmark test of k-means implementations or statistical clustering software.

Table 3.1: Solutions of incommensurabilty in different software packages

	k-means-clustering					latent class models	
	SPSS Quick Cluster[a]	**ALMO Prog37**[b]	**SAS**[c]	**CLUSTAN k-means**[d]	**SPSS TwoStep**[a]	**LatentGold**[e]	**ALMO Prog37**[b]
different scale units	no, syntax language can be used to standardize and weight variables, variables may be standardized with the procedure DESCRIPTIVES	z-standardization resp. different methods of weighting can be used; syntax language can be used, too	z-standardization recommended, SAS procedure STANDARD is available	z-standardization, standardization to (unit) range, weighting possible	z-standardization	not necessary, commensurability is solved model based	not necessary, commensurability is solved model based
different measurement levels	no, syntax can be used to solve the problem (see Bacher 2002; Bender et al. 2001)	nominal (including binary) and quantitative are allowed; ordinal variables are treated as interval	no, similar to SPSS syntax can be used	nominal (including binary), ordinal and quantitative	nominal and quantitative variables only, user must decide how to handle ordinal variables	nominal (including binary), ordinal and quantitative, different model based procedures	nominal (including binary), ordinal and quantitative, different model based procedures
syntax and other procedures to solve the problem of incommensurability	yes	yes	yes	Auto Script allows a script file to be run (since Version 6.03)	yes	no, except some basic transformations	yes

a SPSS Version 11.5 was used. Further information is available under http//:www.spss.com

b Version 7.1. ALMO is a standard statistical package developed by Kurt Holm (http://www.almo-statistik.de). The cluster program was developed by the author of this paper. Prog37 contains models for k-means-clustering using a general variance formula and models for probabilistic clustering. The probabilistic clustering model is described in Bacher (2000). It allows to handle variables with mixed measurement levels.

c The SAS OnlineDoc (SAS 2002) has been consulted, online available under http://v9doc.sas.com/sasdoc/

d ClustanGraphics is perhaps the best known stand alone clustering software und has been developed by David Wishart. ClustanGraphics contains a convincing dendrogramme technique and new develepments like focal point clustering, bootstrap validation and techniques for medoids (exemplars). FocalPoint uses different starting values to find the best solution. Procedures to solve the problem of incommensurability are described in (Wishart 2003) and are available in version 6.06 (personal information of David Wishart). Further information is available under http://www.clustan.com.

e Version 2.0 was used. Latent GOLD was developed by Jeroen K. Vermut and Jay Magidson. Version 3.0 is available since the beginning of 2003. The default number of start sets as well as the default number of iterations per set. Version 4.0 will come up after this summer and will contain new features. It will allow new scale types, namely "binomial count", "truncated Poisson", "truncated binomial count", "truncated normal" and "censored normal". Model validation will be improved by a bootstrap technique for the -2LL value (personal information of Jeroen Vermunt, 2004-05-24). Information is available under http://www.statisticalinnovations.com

Table 3.2: Analysed artificial datasets

Cluster models		**Variables**		
		three resp. six random N(0,1)-variables $x_1, x_2, \ldots, x_6$;	SPSS TwoStep specification:	
		three constellations of variables:		
		(1) all variables are quantitative $\Leftrightarrow$	(1) all variables are continuous	
		mixed type attributes	two of them resp. four of them were categorized	
		(2) ordinal variables are treated as quantitative $\Leftrightarrow$	(2) one resp. two of them are treated as categorical	
		(3) ordinal variables are treated as nominal $\Leftrightarrow$	(3) two resp. four of them are treated as categorical	
			substantive interpretation (for purpose of illustration)	
one cluster (M0)	cluster 1	$\mu(x_i) = 0.00; i = 1, 2, 3; n = 20000$ resp. 100 %	no class structure exists in the analyzed population	
two clusters (M2a)	cluster 1	$\mu(x_i) = -0.75; i = 1, 2, 3; n = 10000$ resp. 50 %	lower class	two classes are present,
	cluster 2	$\mu(x_i) = +0.75; i = 1, 2, 3; n = 10000$ resp. 50 %	upper class	the classes are not well seperated
two clusters (M2b)	cluster 1	$\mu(x_i) = -1.50; i = 1, 2, 3; n = 10000$ resp. 50 %	lower class	two classes are present,
	cluster 2	$\mu(x_i) = +1.50; i = 1, 2, 3; n = 10000$ resp. 50 %	upper class	the classes are well seperated
three clusters (M3)	cluster 1	$\mu(x_i) = -1.50; i = 1, 2, 3; n = 5000$ resp. 25 %	lower class	three not well seperated classes
	cluster 2	$\mu(x_i) = 0.00; i = 1, 2, 3; n = 10000$ resp. 50 %	middle class	
	cluster 3	$\mu(x_i) = +1.50; i = 1, 2, 3; n = 5000$ resp. 25 %	upper class	
five clusters (M5)	cluster 1	$\mu(x_i) = -1.50; i = 1, 2, 3; n = 3000$ resp. 15 %	lower class	three not well seperated classes plus two
	cluster 2	$\mu(x_i) = 0.00; i = 1, 2, 3; n = 10000$ resp. 50 %	middle class	inconsistent classes
	cluster 3	$\mu(x_i) = +1.50; i = 1, 2, 3; n = 3000$ resp. 15 %	upper class	
	cluster 4	$\mu(x_1) = 1.50; \mu(x_2) = \mu(x_3) = 0.00;\ n = 2000$ resp. 10 %	selfmade man	
	cluster 5	$\mu(x_1) = 0.00;\ \mu(x_2) = \mu(x_3) = 1.50;\ n = 2000$ resp. 10 %	intellectuals	

$\mu(x_i)$ = mean of a variable x_i

the performance of SPSS TwoStep for mixed type attributes the analysis was repeated with the original (uncategorized) variables.

For purpose of illustrations, the quantitative variables can be interpreted as income (e.g. personal income in the experiments with three variables, income of father and mother in the case of six variables to analyse the social status of children); the ordinal variables can be interpreted as education (e.g. personal education resp. education of father and mother) and the nominal variables can be interpreted as occupation (e.g. personal occupation resp. occupation of father and mother). The clusters represent social classes. The model with one cluster represents a society with no class structure. The model with two clusters corresponds to a two class structure (lower class and upper class are distinguishable), the model with three clusters represents a three class structure (lower class, middle class and upper class can be differentiated). The model with five clusters corresponds to a five class structure. Two of the five clusters have an inconsistent pattern. On the average, persons of cluster 4 have a high income, but a middle education (and occupation). They may be labeled the class of self-made man. In contrast, persons of cluster 5 on the average have a middle income, but a high education (and occupation). This class may be labeled as the class of intellectuals. The other clusters correspond to upper class, middle class and lower class.

For each model six artificial data sets with 20,000 cases were generated. In order to test the influence of sample size, we analyzed the total sample of 20,000 cases and subsamples with 10,000, 5,000, 2,000, 1,000 and 500 cases.

In total, 900 data sets were generated. The results are summarized in table 3.3.[6]

Continuous resp. quantitative variables only. SPSS TwoStep is able to detect the correct number of clusters for the models with two and three classes (models M2a, M2b and M3). Sample size has no effect for those configurations. Some problems occur for the model with three clusters, if only three variables are used. The procedure selects sometimes two or four clusters. However, these defects are not severe (4 of 36).

In contrast to these postive results, the model with five clusters (M5) results in wrong decisions for the number of clusters in all experiments with 20,000 cases. SPSS TwoStep is not able to find the two small inconsistent clusters that were added to the three class structure. If three variables are analyzed the results are unstable. For instance, the procedure decides for six, four, five or three clusters if the sample size is 10,000. In the case of six variables, the procedure results in three clusters.

Model M0 (no class structure) results in instable results, too. SPSS TwoStep is not able to analyze the question whether a cluster model underlies the data.

A closer look to the distance ratio reveals that a small difference between the largest distance ratio and the second largest distance ratio occurs if SPSS TwoStep has problems to detect the correct number of clusters. The situation is similar to those reported in table 2.1. Two or more cluster solutions have similar distance ratio around 1.3 to 3.8. In contrast to this ambiguous situation, a large difference is computed in those cases where SPSS TwoStep computes the correct number of clusters: There is one large distance ratio (for the different models values greater than 4.0, 8.0, 13.0, 18.0 or 50.0) whereas the second largest value is small (maximum values of 1.2 to 2.0). The results suggest that the ratio change $R(k_1)/R(k_2)$ (formula 2.11) should at least be greater than 2.5. To guarantee a higher amount of safety, values should be greater than 3.0, 3.5 or 4.0.

Different measurement levels, ordinal variables treated as categorical. The results are unsatisfactory. SPSS TwoStep only recovers the correct number of clusters for the two and three class models (M2a, M2b and M3) if six variables are used. In all other specifications, SPSS TwoStep leads to a wrong decision. For

6 The simulation study was done with the German version of SPSS 11.5. A replication with SPSS 12.0 (German version) and the English version of SPSS 11.5 reports different findings for some constellations. The differences are probably due to improvements of the algorithm (personal information of SPSS Technical Support, 2004-05-24).

Table 3.3: Number of clusters computed by SPSS TwoStep (Results of simulation studies)

Cluster Model	Sample Size n	only quantitative variables		ordinal var. treated as categorical		ordinal var. treated as continuous	
		three var.	six var.	three var.	six var.	three var.	six var.
M0	20,000	5,4,4,9,4	3,3,11,2,10	8,8,6,8,8	8,4,6,4,8	2,6,2,6,2	3,6,3,3,3
	10,000	10,2,6,9,5	6,8,7,7,5	6,3,3,8,8	5,4,4,6,4	6,2,2,2,2	3,6,3,3,3
	5,000	6,5,3,9,4	7,6,7,7,4	6,6,3,8,8	4,6,5,5,5	2,2,6,6,6	3,3,3,3,3
	2,000	4,5,7,4,11	6,5,5,4,3	3,8,3,3,8	4,4,4,8,6	2,2,6,6,6	3,3,3,3,3
	1,000	2,7,7,4,4	4,8,3,6,2	3,8,3,3,3	4,10,5,5,5	2,6,2,6,6	3,3,3,3,3
	500	3,3,6,2,2	4,4,4,5,6	3,3,6,3,8	5,4,5,5,4	2,2,2,6,6	3,3,3,3,3
M2a	20,000	2,2,2,2,2	2,2,2,2,2	10,7,7,10,7	2,2,2,2,2	3,3,3,3,3	8,8,8,8,8
	10,000	2,2,2,2,2	2,2,2,2,2	7,10,10,7,7	2,2,2,2,2	3,3,3,3,3	8,8,8,8,8
	5,000	2,2,2,2,2	2,2,2,2,2	7,7,7,10,7	2,2,2,2,2	3,3,3,3,3	8,8,8,8,8
	2,000	2,2,2,2,2	2,2,2,2,2	7,7,7,7,7	2,2,2,2,2	4,3,3,3,3	8,8,8,8,8
	1,000	2,2,2,2,2	2,2,2,2,2	7,7,7,7,7	2,2,3,2,2	3,3,3,3,3	8,8,8,8,8
	500	2,2,2,2,2	2,2,2,2,2	7,7,7,7,7	2,2,2,2,2	4,3,3,3,3	8,8,7,8,8
M2b	20,000	2,2,2,2,2	2,2,2,2,2	7,7,7,7,7	2,2,2,2,2	3,3,3,3,3	2,2,2,2,2
	10,000	2,2,2,2,2	2,2,2,2,2	7,7,7,7,7	2,2,2,2,2	3,3,3,3,3	2,2,2,2,2
	5,000	2,2,2,2,2	2,2,2,2,2	7,7,7,7,7	2,2,2,2,2	3,3,3,3,3	2,2,2,2,2
	2,000	2,2,2,2,2	2,2,2,2,2	7,7,7,7,7	2,2,2,2,2	3,3,3,3,3	2,2,2,2,2
	1,000	2,2,2,2,2	2,2,2,2,2	7,7,7,7,7	2,2,2,2,2	3,3,3,3,3	2,2,2,2,2
	500	2,2,2,2,2	2,2,2,2,2	7,7,7,7,7	2,2,2,2,2	3,3,3,3,3	2,2,2,2,2
M3	20,000	3,3,3,2,3	3,3,3,3,3	10,7,7,7,7	3,3,2,3,3	3,3,3,3,3	7,7,7,7,7
	10,000	3,3,3,3,3	3,3,3,3,3	10,7,7,7,7	3,3,3,2,2	3,3,3,3,3	7,7,7,7,7
	5,000	3,3,4,3,3	3,3,3,3,3	10,7,7,10,10	2,2,3,2,2	3,3,3,3,3	7,7,7,7,7
	2,000	3,3,3,2,3	3,3,3,3,3	7,7,7,7,7	3,3,3,2,2	3,3,3,3,3	7,8,7,7,7
	1,000	3,3,3,3,3	3,3,3,3,3	7,7,7,7,7	3,3,2,3,2	3,3,3,3,3	7,7,7,7,7
	500	3,2,3,3,3	3,3,3,3,3	7,7,7,7,7	2,3,4,2,4	3,3,3,3,3	7,7,7,7,7
M5	20,000	2,3,2,2,3	3,3,3,3,3	7,7,7,7,7	6,5,6,4,4	3,3,3,3,3	7,7,7,7,7
	10,000	6,4,4,5,3	3,3,3,3,3	7,7,7,7,7	6,5,6,3,3	3,3,3,3,3	7,7,7,7,7
	5,000	2,3,3,3,2	3,3,3,3,3	7,7,7,7,7	6,3,3,4,2	3,3,3,3,3	7,7,7,7,7
	2,000	3,3,6,2,4	3,3,3,3,3	7,7,7,7,7	6,7,2,6,6	3,3,3,3,3	7,7,7,7,7
	1,000	2,5,3,3,2	3,3,3,3,3	7,7,7,7,7	3,3,7,4,2	3,3,3,3,3	4,7,7,7,7
	500	3,2,4,2,4	3,3,3,3,3	7,7,7,7,7	6,2,3,6,3	3,3,3,3,3	4,4,7,7,7

three variables, the results are instable, resulting in different number of clusters for the same model specification. In these cases, SPSS TwoStep computes a seven or ten cluster solution that is dominated by the nominal variables: Different combination of the nominal variables, e.g. (0,0), (1,0), (1,0), (1,1), (1,2), (2,1) and (2,2), build the cluster (see below). Sample size has nearly no effect.

Different measurement levels, ordinal variables treated as continuous. The results are more disappointing than those for the model where ordinal variables are treated as categorical. SPSS TwoStep predicts for two models (M2b-six variables and M3-three variables) the correct number of clusters. However, the results for the model with three clusters are inconsistent. More variables do not lead to better results. Furthermore, the estimation for the model's parameters are biased for the model M3 with three clusters and three variables. (see table 3.5) The results for the other configurations very often depend on the categorical variables (see below). Again sample size has nearly no effect.

Reasons for poor performance in the case of mixed type attribute. One reason for the poorer performance of SPSS TwoStep clustering in the case of mixed type attribute is the loss of information if variables are categorized. Categorization limits the number of possible clusters. If one continuous variable is trichotomized the maximal number of clusters is three, etc. Moreover, categorization may destroy the underlying structure.

Another reason is the different weighting of differences in categorical and continuous variables. Categorical variables are implicitly given a higher weight (see section 3.1). Therefore, the cluster frequently corresponds simply to different combination of the categorial variables. This effect is shown in table 3.4. Table 3.4 reports the results of the analysis of model M2b for three variables. Obviously, we would have no problem to interpret the solutions substantively: "There are three classes, a lower, a middle and an upper class." However, this interpretation is wrong, because only two classes underly the structure of the data.

Qualitative evaluation. We inspected the cluster profiles of those solutions for which SPSS TwoStep predicts the correct number of clusters. The results are summarized in the tables 3.5 to 3.7.

The results for the mixed type attributes support our hypothesis (see section 3.1) that categorical variables may dominate the results because differences in nominal variables are given a higher weight than differences in continuous variables. This results in an overestimation of the differences between the clusters in the categorical variables and an underestimation of the differences between the clusters in the quantitative variables. This effect is stronger if the clusters are less seperated and overlap. For instance, for model M2a (see table 3.5) the estimated differences in the continuous variables are 1.37 resp. 1.34. The true differences are 1.53 resp. 1.50. For model M2b the differences between true and estimated parameters are smaller (estimated differences 3.00 resp. 2.98; true differences 3.03 resp. 3.00).

However, this defects are less severe. The main problem seems to be that the implicit weighting results in a wrong number of clusters. (see above) If the correct number is predicted, the bias is small.

For continuous variables (table 3.7) the results are satisfactory. True and estimated parameter do not differ too much. The findings for model M2a and three variables are one exception. The differences between the two clusters in the continuous variables are overestimated. The reason for this bias is the deterministic assignment rule that results in biased estimates if the cluster overlap. This overlap reduces if variables are added. Therefore, the bias diminishes for M2a with six variables.

Table 3.4: Results of SPSS Twostep for model M2b (three variables)

		cluster profiles[a]		frequencies[b]		
cluster	n	**quant. var**	**ord. var**	**cat. 0**	**cat. 1**	**cat. 2**
		estimated parameters				
1	5,062	1.00	1.50	0	0	100
2	9,853	0.02	1.00	0	100	0
3	5,085	−1.03	0.54	100	0	0
		true[c] *parameters*				
1	10,000	1.52	1.69	0.8	50.3	98.9
2	10,000	−1.51	0.32	99.2	49.7	1.1

a means are reported
b column percentages are reported
c The term "true" is used if the known (correct) cluster membership is used to compute cluster centers and cluster frequencies.

Table 3.5: Estimated parameters (six variables, ordinal variables are treated as categorical)

			cluster centres		frequencies of first nominal variable[a]		
model	**cluster**	n	**quant1**	**quant2**	**cat. 0**	**cat. 1**	**cat. 2**
estimated parameters for M2a	1	8582	.75	.70	0	44.4	82.3
	2	11418	−.62	−.64	100	55.6	17.7
true[b] parameters for M2a	1	10000	.77	.73	8.5	50.6	90.4
	2	10000	−.76	−.77	91.5	49.4	9.6
estimated parameters for M2b	1	9901	1.52	1.48	0.4	49.8	98.4
	2	10091	−1.48	−1.50	99.6	50.2	1.6
true[b] parameters for M2b	1	10000	1.52	1.48	0.8	50.3	98.9
	2	10000	−1.51	−1.52	99.2	49.7	1.1

a Similar results are computed for the other three categorical variables.
b The term "true" is used if the known (correct) cluster membership is used to compute cluster centers and cluster frequencies.

Table 3.6: Estimated parameters (ordinal variables are treated as quantitative/continuous variables)

			cluster centres				**frequencies of first nominal variable**[a]		
model	**cluster**	n	**quant1**	**quant2**	**ord1**	**ord2**	**cat. 0**	**cat. 1**	**cat. 2**
estimated	1	10130	1.52	1.47	1.69	1.68	0	49.6	98.4
parameters for M2a	2	9870	−1.47	−1.48	0.34	0.33	100	50.4	1.6
true[b] parameters	1	10000	1.52	1.48	1.69	1.68	8.5	50.6	90.4
for M2a	2	10000	−1.51	−1.52	0.32	0.31	91.5	49.4	9.6
estimated	1	5062	1.01	.	1.46	.	0	0	0
parameters for M3	2	9853	0.02	.	1.00	.	0	100	0
	3	5085	−1.04	.	0.54	.	100	0	100
true[b] parameters	1	5000	1.49	.	1.69	.	0.7	15.6	68.5
for M3	2	10000	0.00	.	0.99	.	31.4	69.4	31.0
	3	5000	−1.51	.	0.31	.	68.0	15.6	0.5

a Similar results are computed for the other three categorical variables.
b The term "true" is used if the known (correct) cluster membership is used to compute cluster centers and cluster frequencies.

Table 3.7: Estimated parameters (quantitative variables only)

			cluster profiles					
model	**cluster**	n	**quant1**	**quant2**	**quant3**	**quant4**	**quant5**	**quant6**
estimated	1	9829	−.84	−.81	−.82			
parameters for M2a	2	10171	.81	.78	.78			
true[a] parameters	1	10000	−.76	−.74	−.76			
for M2a	2	10000	.77	.75	.74			
estimated	1	9960	−.77	−.76	−.77	−.78	−.76	−.76
parameters for M2a	2	10040	.78	.75	.75	.73	.74	.77
true[a] parameters	1	10000	−.76	−.74	−.76	−.77	−.75	−.75
for M2a	2	10000	.77	.75	.74	.73	.74	.76
estimated	1	9987	−1.51	−1.49	−1.51			
parameters for M2b	2	10013	1.52	1.49	1.49			
true[a] parameters	1	10000	−1.51	−1.49	−1.51			
for M2b	2	10000	1.52	1.50	1.49			
estimated	1	10001	−1.51	−1.50	−1.51	−1.52	−1.50	−1.50
parameters for M2b	2	9999	1.52	1.50	1.49	1.48	1.49	1.51
true[a] parameters	1	10000	−1.51	−1.49	−1.51	−1.52	−1.50	1.50
for M2b	2	10000	1.52	1.50	1.49	1.48	1.49	1.51
estimated	1	5592	−1.54	−1.47	−1.50			
parameters for M3	2	8600	−.05	−.01	.02			
	3	5808	1.56	1.43	1.38			
true[a] parameters	1	5000	−1.51	−1.51	−1.52			
for M3	2	10000	−.00	.00	.00			
	3	5000	1.53	1.50	1.47			
estimated	1	5077	1.55	1.49	1.47	1.47	1.49	1.52
parameters for M3	2	9851	−.01	.01	.01	−.02	−.01	.00
	3	5072	−1.52	−1.51	−1.52	−1.53	−1.51	−1.51
true[a] parameters	1	5000	−1.51	−1.51	−1.52	−1.53	−1.52	−1.50
for M3	2	10000	−.00	.00	.00	−.02	.00	.00
	3	5000	1.53	1.50	1.47	1.47	1.47	1.52

a The term "true" is used if the known (correct) cluster membership is used to compute cluster centers and cluster frequencies.

3.3 Comparisons with LatentGOLD

The poor perfomance of SPSS TwoStep in some of the analyzed configurations is not a sufficient reason to reject the method. Perhaps other methods perform poorer. Therefore, we compared SPSS TwoStep with the LC-Model of LatentGOLD and the probabilistic clustering model of ALMO (see next section).

The LC-Model of LatentGOLD (Vermunt and Magidson 2000) allows to handle mixed data. The conditional probability $\pi(x_{gj/i})$ that a certain value of variable y_j is observed for person g in class (cluster) i is defined for categorical and ordinal variables as:

$$\pi(x_{gj(l)/i}) = \frac{\exp(\eta_{gj(l)/i})}{\sum_j \exp(\eta_{gj(l)/i})} \tag{3.3}$$

where $\eta_{gj(l)/i} = \beta_j + \beta_{j(l)i}$ if $x_{gj(l)} = 1$ for nominal variables and $\eta_{gj(l)/i} = \beta_j + \beta_{ji}x_{gi(l)}$ for ordinal variables.

The normal distribution is assumed for quantitative variables:

$$\pi(x_{gj/i}) = \varphi(x_{gj}/\mu_{ij}, \sigma^2_{ij}) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} e^{-0.5(x_{gj}-\mu_{ij})^2/\sigma^2_{ij}} \tag{3.4}$$

The parameters are estimated with the EM (expected maximum likelihood estimators) and the Newton-Raphson algorithm. The programme starts with EM iterations and switches to Newton-Raphson. LatentGOLD uses different starting values (default value = 10) for each model specification in order to avoid local minima. Different test statistics, including BIC and AIC are computed for model selection. The models are estimated seperately and the user must compare the models by hand. Hence, the handling of LatentGOLD is less comfortable than SPSS TwoStep.

Table 3.8 summarizes the results of our analyzed configurations. Only one analysis was done within one specification and only the sample size of $n = 20,000$ was analyzed.[7] These restrictions were necessary because LatentGOLD needs much more computing time than SPSS TwoStep clustering and model selection must be done by hand.

LatentGOLD is able to predict the correct number of clusters for all analyzed configurations. These results imply: In contrast to SPSS TwoStep, LatentGOLD is able to detect data sets without an underlying class structure. LatentGOLD is able to predict correctly the number of classes for modell M5 if all variables are quantitative. Problems occur if class dependent variances are assumed for modell M5 with three variables. The *BIC* coefficient suggests a three resp. a four class solution for modell M5 if all three variables are continuous resp. of mixed type. LatentGOLD surpasses SPSS TwoStep in the case of variables with mixed type attributes, too. Models M2a, M2b and M3 are correctly predicted in the case of three variables, whereas SPSS TwoStep needs more variables to predict correctly either model M2a, M2b or M3 (see table 3.3 results for $n = 20,000$, six variables and ordinal variables are treated as categorical).

7 Class independent variances were used (default specifications of LatentGOLD 2.0). If class dependent variances are specified problems occur to determine the correct number of clusters for model M5 - three variables (see below).

Table 3.8: BIC for the analyzed configurations computed by LatentGOLD

	three variables					six var.
cluster	**Mod. M0 BIC**	**Mod. M2a BIC**	**Mod. M2b BIC**	**Mod. M3 BIC**	**Mod. M5 BIC**	**Mod. M5 BIC**
all variables are quantitative						
1	▷170,631	197,525	241,362	215,967	208,307	416,265
2	170,663	▷189,238	▷197,963	202,962	201,428	394,539
3	170,702	189,263	197,993	▷199,412	199,099	382,600
4	170,735	189,296	198,027	199,450	199,093	381,677
5	170,772	189,333	198,065	199,481	▷198,979	▷380,605
6	170,810	189,368	198,091	199,510	199,010	380,652
7	170,846	189,405	198,126	199,543	199,042	380,712
8	170,879	189,449	198,166	199,577	199,076	380,765
9	170,912	189,479	198,206	199,613	199,113	380,818
10	170,948	189,517	198,235	199,661	199,147	380,873
11	170,997	189,559	198,280	199,684	199,183	380,937
12	171,028	189,604	198,302	199,717	199,233	380,998
variables with different measurement level (nominal, ordinal and quantitative)						
1	▷124,984	145,866	168,551	156,171	150,925	301,245
2	125,026	▷139,685	▷136,113	145,282	145,291	283,039
3	125,069	139,723	136,150	▷143,127	144,092	275,837
4	125,116	139,765	136,194	143,164	144,062	274,834
5	125,163	139,819	136,235	143,210	▷144,019	▷273,969
6	125,210	139,861	136,276	143,250	144,067	274,037
7	125,257	139,901	136,339	143,296	144,113	274,116
8	125,306	139,947	136,368	143,351	144,151	274,196
9	125,356	139,998	136,417	143,397	144,207	274,266
10	125,405	140,060	136,463	143,442	144,253	274,338
11	125,455	140,107	136,508	143,485	144,307	274,419
12	125,500	140,152	136,562	143,529	144,341	274,498

▷ correct number of clusters, _ empirical estimated number of clusters

3.4 Comparisons with ALMO

ALMO (Holm 2004) contains a probabilistic clustering model. The model is described in Bacher (2000). The conditional probability $\pi(x_{gj/i})$ that a certain value of variable y_j is observed for person g in class (cluster) i is defined as:

$$\pi(x_{gj/i}) = \pi(j(l)/i) \quad \text{für} \quad x_{gj(l)} = 1 \quad \text{for nominal variables} \tag{3.5}$$

$$p(x_{gj/i}) = \binom{m_j}{x_{gj}} p(j/i)^{x_{gj}} (1 - p(j/i))^{m_j - x_{gj}} \quad \text{for ordinal variables}^{8} \tag{3.6}$$

$$\pi(x_{gj/i}) = \varphi(x_{gj}/\mu_{ij}, \sigma^2_{ij}) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} e^{-0.5(x_{gj}-\mu_{ij})^2/\sigma^2_{ij}} \quad \text{for quantitative variables} \tag{3.7}$$

LatentGOLD and ALMO differ in the treatment of ordinal variables. LatentGOLD uses fixed category scores, ALMO a response model developed by Rost (1985). Both programmes assume a normal distribution resp. a multinominal distribution for quantitative resp. nominal variables. Additionally, LatentGOLD allows to weaken the assumption of local independence.

The parameters of the ALMO model are estimated with the EM algorithm. Different starting values must be defined by hand. Different test statistics, including BIC and AIC are computed for model selection. Different models can be estimated in one step. Therefore in contrast to LatentGOLD, summary tables of the test statistics are reported which facilitate model selection similar to SPSS TwoStep. However, ALMO needs more time – about one hour per model – for computing the parameters than LatentGOLD. Also for this reason only the same model configurations were analyzed as done with LatentGOLD.

Table 3.9 shows the results: If all variables are quantitative, ALMO succeeds in detecting the correct number of clusters except model M5 with three variables. In particular ALMO realizes the right cluster structure in model M0 and detects only one cluster in the data set and correctly predicts the number of clusters in model M5 with six variables.

In the configurations with different measurement levels, ALMO reports the right cluster structure in the models M0, M2b and M5 with six variables. The behavior of the *BIC* coefficient for model M2a is untypical. *BIC* is decreasing several times. This is an indication that local minima are found. LatentGOLD avoids these problems by using different starting values. ALMO only uses one starting point. The starting point must be varied manually by the user.

As a first summary, the ALMO results are superior to SPSS, because ALMO computes the correct number of clusters for those data sets with no underlying cluster structure and for those data sets with five clusters (when six variables are provided).

8 Version 4.0 of LatentGOLD will contain this model, too. The model will be labeled as “binomial count”.

Table 3.9: BIC for the analyzed configurations computed by ALMO

	three variables					six var.
cluster	**Mod. M0 BIC**	**Mod. M2a BIC**	**Mod. M2b BIC**	**Mod. M3 BIC**	**Mod. M5 BIC**	**Mod. M5 BIC**
all variables are quantitative						
1	▷170,631	197,525	241,362	215,967	208,307	416,265
2	170,700	▷189,251	▷197,977	202,980	201,364	394,001
3	170,757	189,317	198,043	▷199,458	199,033	382,371
4	170,820	189,382	198,104	199,523	199,035	381,277
5	170,890	189,446	198,166	199,586	▷199,099	▷380,812
6	170,953	189,507	198,227	199,643	199,163	380,925
7	171,017	189,578	198,273	199,705	199,221	381,039
8	171,085	189,638	198,344	199,770	199,273	381,157
9	171,152	189,706	198,396	199,836	199,346	381,250
10	171,221	189,763	198,466	199,888	199,407	381,355
11	171,274	189,830	198,537	199,964	199,484	381,475
12	171,350	189,901	198,580	200,017	199,530	381,597
variables with different measurement level (nominal, ordinal and quantitative)						
1	▷127,678	146,130	171,671	156,171	150,988	301,371
2	127,737	▷141,079	▷136,558	146,060	146,139	284,387
3	127,788	141,104	136,603	▷144,273	145,155	278,640
4	127,847	141,088	136,654	144,281	145,019	277,476
5	127,907	141,116	136,714	144,281	▷145,021	▷277,139
6	127,966	141,036	136,755	144,215	145,039	277,161
7	128,023	141,085	136,818	144,220	145,021	277,257
8	128,081	141,113	136,866	144,256	145,074	277,376
9	128,138	141,155	136,931	144,268	145,082	277,445
10	128,201	141,187	136,999	144,286	145,162	277,511
11	128,263	141,234	137,044	144,325	145,188	277,586
12	128,322	141,308	137,099	144,347	145,239	277,699

▷ correct number of clusters, _ empirical estimated number of clusters

4 Discussion and summary

Derived from tables 3.3, 3.8 and 3.9 the results can be summarized as shown in table 4.1.

At first glance, SPSS TwoStep algorithm realizes the right cluster structure only for 3 of the 12 tested models (25 %). The success rate of LatentGOLD and ALMO is clearly higher and reaches 100 % for LatentGOLD resp. 66 % for ALMO - in other words: LatentGOLD predicts the correct number for all analyzed constellations, ALMO for 8 of the 12 models. LatentGOLD shows the best performance. Problems only occur for the models with five clusters and three variables.

Only if all variables are quantitative the results of SPSS TwoStep algorithm are correct for those models with two or three clusters. But these models are also predicted correctly by the other two programmes.

The implicit different weighting of variables is one reason for the poor performance of SPSS TwoStep algorithm in the case of mixed data. This problem is already seen by SPSS. A solution will be offered in the coming release.[1]

ALMO obviously has problems with mixed type data, too. The assumed ordinal model is not appropriate for the data analyzed.

Those users who have to rely only on SPSS should be careful when using SPSS TwoStep for variables with different measurement levels if clusters overlap. At this point of time an application of the TwoStep algorithm to this kind of problems cannot be recommended. We recommend to use LatentGOLD for this kind of data or to avoid the problem by collecting more and mainly quantitative variables.

Finally, we want to note that SPSS TwoStep was orginally designed to cluster large data sets (e. g. several millions of cases and many variables) within an acceptable time. This aim has been realized as our simulation studies underline. SPSS TwoStep needs much less computing time than LatentGOLD and ALMO.

Table 4.1: Overview of the results for SPSS TwoStep (S), LatentGOLD (L) and ALMO (A)

number of	**three variables**					**six variables**
clusters	**Model M0**	**Model M2a**	**Model M2b**	**Model M3**	**Model M5**	**Model M5**
			all variables are quantitative			
right	L, A	S, L, A	S, L, A	(S), L, A	(L)[a]	A, L
wrong	S				S, A	S
			variables with different measurement level (nominal, ordinal and quantitative)			
right	L, A	L	L, A	L	(L)[a]	A, L
wrong	S	S, A	S	S, A	S, A	(S)

a LatentGOLD results in a wrong decision if class dependent variables are assumed.

1 SPSS plans to change the first part of ξ_v in equations (2.2) to (2.4). Instead of adding the variances of the variables $\hat{\sigma}_j^2$, a smaller positive scalar Δ_j will be defined or the scalar Δ_j may be specified by the user. (personal information of SPSS Technical Support, 2004-06-03)

References

Bacher, J. (1996). *Clusteranalyse: Anwendungsorientierte Einführung*. Oldenbourg, München [u.a.]. 2., ergänzte Auflage.

Bacher, J. (2000). A Probabilistic Clustering Model for Variables of Mixed Type. *Quality & Quantity*, 34: 223–235.

Bacher, J. (2002). Statistisches Matching: Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS. *ZA-Informationen*, 51: 38–66.

Bender, S., Brand, R., and Bacher, J. (2001). Re-identifying register data by survey data: An empirical study. *Statistical Journal of the UN Economic Commission for Europe*, 18(4): 373–381.

Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C. (2001). A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2001*, pages 263–268.

Everitt, B. S. (1981). *Cluster analysis*. Heinemann Educational Books, London, second, repr. edition.

Everitt, B. S., Landau, S., and Leese, M. (2001). *Cluster analysis*. Arnold, London, forth edition.

Fox, J. (1982). Selectiv aspects of measuring resemblance for taxonomy. In Hudson, H. C., editor, *Classifying social data. New applications of analytical methods for social science research*, pages 127–151. Jossey-Bass, San Francisco, Washington, London.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27: 857–872.

Holm, K. (2004). ALMO Statistik-System, Version 7.1. http://www.almo-statistik.de/.

Huang, Z. (1998). Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Variables. *Data Mining and Knowledge Discovery*, 2: 283–304.

Rost, J. (1985). A latent class model for rating data. *Psychometrika*, 50(1): 37–49.

SAS Institute Inc. (2002). *SAS OnlineDoc® 9*. Cary, NC. http://v9doc.sas.com/sasdoc/.

SPSS Inc. (2001). The SPSS TwoStep cluster component. A scalable component to segment your customers more effectively. White paper – technical report, Chicago. ftp://ftp.spss.com/pub/web/wp/TSCWP-0101.pdf.

SPSS Inc. (2004). TwoStep Cluster Analysis. Technical report, Chicago. http://support.spss.com/tech/stat/Algorithms/12.0/twostep_cluster.pdf.

Vermunt, J. and Magidson, J. (2000). *Latent GOLD 2.0. User's Guide*. Belmont.

Wishart, D. (2003). k-Means Clustering with Outlier Detection, Mixed Variables and Missing Values. In Schwaiger, M. and Opitz, O., editors, *Exploratory data analysis in empirical research. Proceedings of the 25th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Munich, March 14-16, 2001*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 216–226, Berlin. Springer.

Publikationen des Lehrstuhls für Soziologie

Berichte

In der Reihe „Berichte“ finden sich herausragende Forschungsergebnisse. ISSN 1437-6741 (print); ISSN 1438-4663 (online)

Wittenberg, Reinhard (Hg.): „Neues aus Wissenschaft & Praxis für Praxis & Wissenschaft“. Beiträge zum 4. Nürnberger AbsolventInnentag der Sozialwissenschaften am 4./5. Juli 2003. Bericht 2004-1 (online)

Lechner, Birgit: Freizeitverhalten von BerufsschülerInnen im Rahmen der Lebensstilforschung und Subkulturtheorie. Bericht 2001-1 (online)

Wittenberg, Reinhard: AbsolventInnen des Studiengangs Sozialwissenschaften an der Universität Erlangen-Nürnberg: Studium und Beruf. Bericht 2000-2 (online)

Wenzig, Claudia: Armutsverlaufsmuster und ihre Auswirkungen auf das Wohlbefinden bei 17- bis 24-jährigen. Eine Analyse des Sozio-ökonomischen Panels 1985-1996. Bericht 2000-1 (online)

Funk, Walter: Kriminalitätsbelastung von Deutschen und Ausländern in Nürnberg 1996. Bericht 99-2

Wittenberg, Reinhard, unter Mitarbeit von Thomas Rothe, Sandra Proske, Claudia Wenzig & Knut Wenzig: Studienabbruch sowie Studienfach- und/oder Studienortwechsel an der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität Erlangen-Nürnberg. Bericht 99-1 (online)

Arbeits- und Diskussionspapiere

In der Reihe „Arbeits- und Diskussionspapiere“ publizieren wir (Zwischen-) Ergebnisse unserer Forschungstätigkeit, Beiträge zur methodischen Diskussion und Skripten für unsere Lehrveranstaltungen.

Wenzig, Knut & Günter Buttler: Panel für Gründer in Freien Berufen – Die erste Welle im Überblick und die Bewertung der Beratungsqualität am IFB. Arbeits- und Diskussionspapiere 2004-3 (online)

Bacher, Johann, Knut Wenzig & Melanie Vogler: SPSS TwoStep Cluster – A First Evaluation. Arbeits- und Diskussionspapiere 2004-2, 2., korr. Aufl. (online)

Prosch, Bernhard & Nadine Jakob: Mobilitätsmanagement im Meinungsbild – Erste Ergebnisse einer Bevölkerungsbefragung zur Initiaitive NürnbergMOBIL. Arbeits- und Diskussionspapiere 2004-1

Dees, Werner & Claudia Wenzig: Das Nürnberger Kinderpanel - Untersuchungsdesign und Deskription der Untersuchungspopulation. Arbeits- und Diskussionspapiere 2003-5 (online)

Wittenberg, Reinhard & Manuela Schmidt: Antisemitische Einstellungen in Deutschland in den Jahren 1994 und 2002. Ein Vergleich zweier Studien des American Jewish Committee, Berlin. Arbeits- und Diskussionspapiere 2003-4 (online)

Wenzig, Knut & Johann Bacher: Determinanten des Studienverlaufs. Was beeinflusst den Studienverlauf an der WiSo-Fakultät der Friedrich-Alexander-Universität Erlangen-Nürnberg? Eine Sekundäranalyse von Daten des Prüfungsamts und der Studentenkanzlei. Arbeits- und Diskussionspapiere 2003-3 (online)

Wittenberg, Reinhard: Einführung in die sozialwissenschaftlichen Methoden und ihre Anwendung in empirischen Untersuchungen I – Skript. 3., überarb., erg. u. akt. Aufl. Arbeits- und Diskussionspapiere 2003-2 (online)

Bacher, Johann: Soziale Ungleichheit und Bildungspartizipation im weiterführenden Schulsystem Österreichs. Arbeits- und Diskussionspapiere 2003-1

Bacher, Johann & Bernhard Prosch: Lebensbedingungen und Lebensstile von Auszubildenden – Ergebnisse der Leipziger Berufsschulbefragung 2000. Arbeits- und Diskussionspapiere 2002-2 (online)

Prosch, Bernhard: Regionalmarketing auf dem Prüfstand. Ergebnisse einer Bevölkerungsbefragung zur Region Nürnberg 2001. Arbeits- und Diskussionspapiere 2002-1

Wittenberg, Reinhard: Einführung in die sozialwissenschaftlichen Methoden und ihre Anwendung in empirischen Untersuchungen I – Skript. 2., überarb., erg. u. akt. Aufl. Arbeits- und Diskussionspapiere 2001-1 (online)

Bacher, Johann: Einführung in die Grundzüge der Soziologie I – Skript. Arbeits- und Diskussionspapiere 2000-4 (online)

Wittenberg, Reinhard: Schwangerschaftskonfliktberatung. Ergebnisse einer Analyse der Nürnberger Beratungsprotokolle des Jahres 1998. Arbeits- und Diskussionspapiere 2000-3 (online)

Wittenberg, Reinhard: Techniken wissenschaftlichen Arbeitens I – Skript. Arbeits- und Diskussionspapiere 2000-2 (online)

Bacher, Johann & Reinhard Wittenberg: Trennung von Kohorten-, Alters- und Periodeneffekten. Arbeits- und Diskussionspapiere 2000-1

Prosch, Bernhard: Raum für starke Köpfe? Regionalmarketing im Meinungsbild. Arbeits- und Diskussionspapiere 99-9 (online)

Prosch, Bernhard & Sören Petermann: Zuckerbrot und Peitsche für die Hühner. Kooperation durch dezentrale Institutionen. Arbeits- und Diskussionspapiere 99-8

Wittenberg, Reinhard, Serap Asiran, Almir Krdzalic, Vanessa S. Karg & Sabine Popp: Studium, Berufswahl und Berufstätigkeit Nürnberger SozialwirtInnen zwischen 1977 und 1999. Erste Ergebnisse. Arbeits- und Diskussionspapiere 99-7

Bacher, Johann: Arbeitslosigkeit und Rechtsextremismus. Forschungsergebnisse auf der Basis des ALLBUS 1996 und der Nürnberger BerufsschülerInnenbefragung 1999. Arbeits- und Diskussionspapiere 99-6 (online)

Wittenberg, Reinhard: Einführung in die Sozialwissenschaftlichen Methoden und ihre Anwendung in empirischen Untersuchungen I - Skript. Arbeits- und Diskussionspapiere 99-5 (online)

Wittenberg, Reinhard: Antisemitische Einstellungen in Deutschland zwischen 1994 und 1998. Messprobleme und Ergebnisse. Arbeits- und Diskussionspapiere 99-4

Bacher, Johann, Christoph Gürtler, Angelika Leonhardi, Claudia Wenzig & Reinhard Wittenberg: Das Nürnberger Kinderpanel. Zielsetzungen, theoretisches Ausgangsmodell, methodische Vorgehensweise sowie wissenschaftliche und praktische Relevanz. Arbeits- und Diskussionspapiere 99-3 (online)

Wittenberg, Reinhard: Pausenverkauf, Ernährung und Gesundheit an Nürnberger Schulen. Arbeits- und Diskussionspapiere 99-2 (online)

Wittenberg, Reinhard & Dorothea Jäkel: Ernährung und Zahngesundheit an Nürnberger Hauptschulen. Arbeits- und Diskussionspapiere 99-1 (online)

Berichte sowie Arbeits- und Diskussionspapiere sind i. d. R. auch als PDF-Dokument abrufbar: http://www.soziologie.wiso.uni-erlangen.de/publikationen