RESEARCH ARTICLE

Latent Class Cluster Analysis

Ali Ünlü

Author's affiliation:

Technical University of Munich (TUM), Munich, Germany

Corresponding author: Prof. Dr. Ali Ünlü
TUM School of Education, Technical University of Munich (TUM)
Arcisstrasse 21, 80333 Munich, Germany
Email: aligalibuenlue@gmail.com
ORCID 0000-0002-6941-5074

ABSTRACT

This paper describes the technique of exploratory latent class cluster analysis (finite mixture modeling). The classical analysis is a model-based statistical approach for identifying unobserved subgroups from observed categorical data (clustering) and for classifying cases into the identified subgroups based on membership probabilities estimated directly from the statistical model (classification).

In the first part on mathematical modeling of the paper, we introduce the data and the sampling distribution for the data as required in the analysis of latent classes (multinomial distribution), the fundamental model assumptions are reviewed, and the general unrestricted latent class model is presented. Classification of cases into the clusters using modal assignment is discussed. In the second part on inferential statistics of the paper, we briefly review the classical maximum likelihood methodology related to parameter estimation and model testing, and the information criteria AIC and SIC for model selection. In the third part on case study of the paper, the General Social Survey data are analyzed using the software Latent GOLD®. We present the Latent GOLD® profile plot and tri plot options for the graphical representation of the results. The Latent GOLD® classification output illustrating the assignment of respondents to the latent survey respondent types is also shown.

Keywords: Latent class analysis, finite mixture, model-based clustering, model-based classification, multinomial distribution, maximum likelihood, General Social Survey data, Latent GOLD® software, profile plot, tri plot

1. Introduction:

Latent class analysis (LCA) is a statistical approach for examining latent categorical variables (Andersen, 1982; Clogg, 1995; Dayton, 1998; Formann, 1984; Goodman, 1974a, b, 1978; Langeheine and Rost, 1988; Lazarsfeld and Henry, 1968; McCutcheon, 1987; Vermunt and Magidson, 2004). Applications of LCA are numerous (Hagenaars and McCutcheon, 2002; Rost and Langeheine, 1997); for example in medical research, when the accuracy (sensitivity and specificity) of a diagnostic test for screening patients for a disease and disease prevalence have to be estimated (for reviews, see Enøe et al., 2000; Hui and Zhou, 1998; Walter and Irwig, 1988).

The basic idea underlying the exploratory LCA approach to cluster analysis (Vermunt and Magidson, 2002) can be summarized as follows. The conditional probabilities of the item responses given the unobserved group memberships of a postulated latent class (LC) model (LCM) differ across unobserved subgroups, called the latent classes, clusters, or types. These subgroups form the categories of a discrete latent variable representing concepts such as attitude toward abortion, Alzheimer disease status, or knowledge state in Euclidean geometry. Having estimated these parameters and assessed the fit of the LCM (including the identification of the number of latent classes), the parameters can be compared across the subgroups to show how they differ from each other. This enables to characterize the unobserved subgroups or the latent variable, based on what is only observable, the data. In a next step, as the classification part of LCA, cases can be classified into the identified clusters based upon membership probabilities estimated directly from the fitted LCM.

These steps will be worked out in detail for this paper. In doing so, we want to embed our elaborations into a general modeling scheme, shown in Figure 1. Figure 1 presents a schematic representation of the basic modeling process in the quantitative behavioral or psychological sciences and the essentials of what is common to the majority of modeling endeavors.

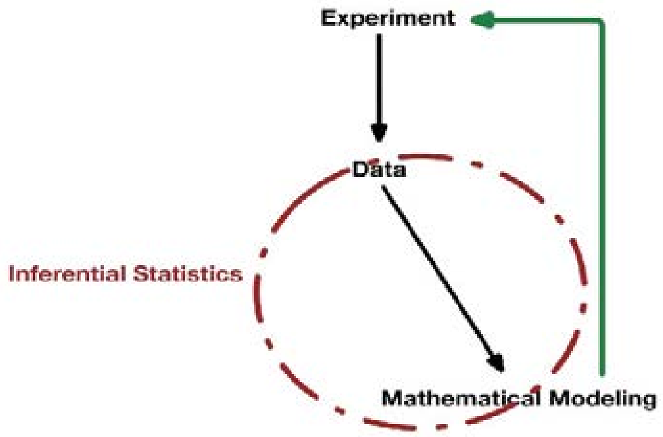


Figure 1. Schematic representation of the simplified modeling process in the quantitative behavioral or psychological sciences. The contents of the paper are structured according to this scheme.

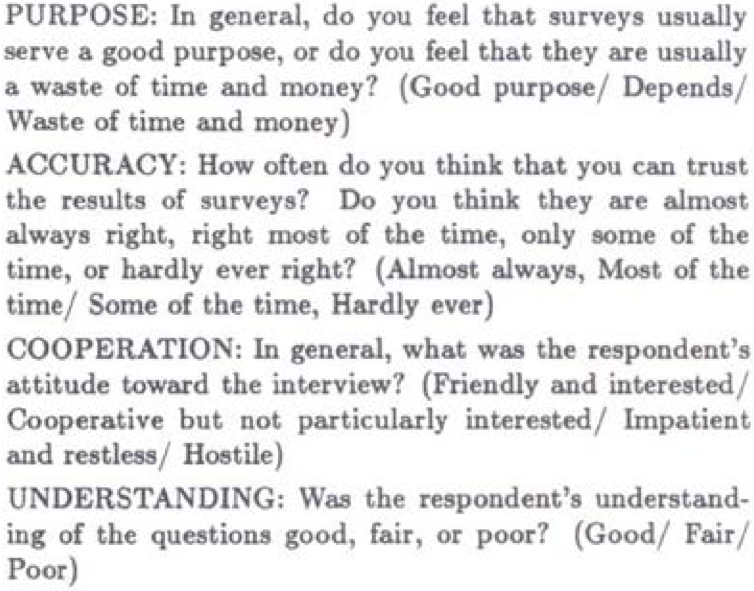
Addressed are the aspects of data, mathematical modeling, and inferential statistics in the overall modeling process. An experiment yields the encoded data (top black arrow), which are first modeled mathematically, whereas hypothetically known population parameters are presupposed (bottom black arrow). After the model has been fully developed mathematically, the, in fact, empirically unknown parameters of the derived mathematical model and hence the model itself are estimated and fully specified, respectively, from the real data using inferential statistics (full-dotted red circle). The ultimate goal then is to manufacture a theoretically validated and empirically fitted mathematical model that can be utilized to eventually describe and draw conclusions about the experiment or empirical phenomenon under consideration, from which the whole process initiated (green arrow).

This paper is structured as follows. In Section 2, we motivate the General Social Survey example test items and data, and the general problem of exploratory latent class cluster analysis. In Section 3, the mathematical modeling is discussed, including the data and sampling distribution for the data, and leading to the formulation of the general latent class model and classification problem. In Section 4, we outline the inferential statistics required for parameter estimation and model testing, and model selection criteria. The LCA analysis of the General Social Survey data based on the software Latent GOLD® is presented in Section 5. Finally, we conclude with summarizing the basic idea and a schematic overview of the overall technical procedure of exploratory latent class cluster analysis in Section 6.

2. 1982 General Social Survey and the Problem: To demonstrate a real LCA application, we take as starting point part of the empirical data from the 1982 General Social Survey (GSS). For illustration purposes, we restrict ourselves to example GSS test items and data, but for more details, refer to McCutcheon (1987).

Two indicators ascertain the respondents' opinions regarding the purpose of surveys (the PURPOSE item) and how accurate the results of surveys are (ACCURACY). Two additional indicators cover the evaluations made by the interviewers of the respondents' levels of understanding of the interview questions (UNDERSTANDING) and cooperation shown in answering the questions (COOPERATION). The original item formulations and corresponding response categories are shown in Table 1.

Table 1. The four 1982 GSS questions (McCutcheon, 1987, p. 29) considered in the example application. The PURPOSE and ACCURACY items are directly responded to by the interviewees. The items COOPERATION and UNDERSTANDING are responded to by the interviewers and cover assessments of the interviewers about the interviewees.



PURPOSE: In general, do you feel that surveys usually serve a good purpose, or do you feel that they are usually a waste of time and money? (Good purpose/ Depends/ Waste of time and money)

ACCURACY: How often do you think that you can trust the results of surveys? Do you think they are almost always right, right most of the time, only some of the time, or hardly ever right? (Almost always, Most of the time/ Some of the time, Hardly ever)

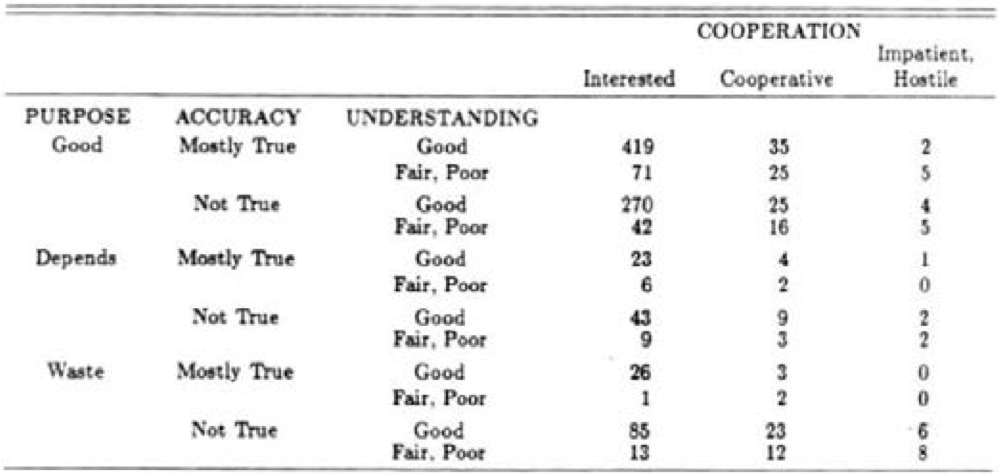
COOPERATION: In general, what was the respondent's attitude toward the interview? (Friendly and interested/ Cooperative but not particularly interested/ Impatient and restless/ Hostile)

UNDERSTANDING: Was the respondent's understanding of the questions good, fair, or poor? (Good/ Fair/ Poor)

In the example application, the first and second response categories of the item ACCURACY correspond to "Mostly True" and "Not True", respectively. For the COOPERATION item, the categories "Impatient and restless" and "Hostile" are merged into one category in the example. For the UNDERSTANDING item, in the application the categories "Fair" and "Poor" are joined.

The corresponding data set, for the subsample of sample size $N = 1,202$ white respondents, analyzed in the paper is shown in Table 2. Table 2 gives the contingency table representation of the GSS data (as compared to the frequency table representation reported later in Table 3).

Table 2. The 1982 GSS data in contingency table representation (McCutcheon, 1987, p. 30). The sample size is $N = 1,202$.



			COOPERATION		
			Interested	Cooperative	Impatient, Hostile
PURPOSE	ACCURACY	UNDERSTANDING			
Good	Mostly True	Good	419	35	2
		Fair, Poor	71	25	5
	Not True	Good	270	25	4
		Fair, Poor	42	16	5
Depends	Mostly True	Good	23	4	1
		Fair, Poor	6	2	0
	Not True	Good	43	9	2
		Fair, Poor	9	3	2
Waste	Mostly True	Good	26	3	0
		Fair, Poor	1	2	0
	Not True	Good	85	23	6
		Fair, Poor	13	12	8

A remark is in order. Latent GOLD® (Section 5) is commercial software distributed by Statistical Innovations, www.statisticalinnovations.com. The demo version of Latent GOLD® is freely available at www.statisticalinnovations.com/latent-gold-5-1/, after registration. The demo version only works with the sample data files downloaded with the software. The data set that we analyze in this paper is also contained in the SPSS system file gss82white.sav accompanying the Latent GOLD® demo version, so the contents of this paper can be rerun by the reader, at no cost.

We consider the above four items of the GSS study. In our example, the PURPOSE variable has the response categories "Good", "Depends", and "Waste". The ACCURACY variable has the categories "Mostly True" and "Not True". The UNDERSTANDING item has two categories "Good" and "Fair, Poor". The item COOPERATION has the categories "Interested", "Cooperative", and "Impatient, Hostile". Thus, we have two dichotomous and two trichotomous example items.

Given the four items, what types of respondent attitudes toward surveys may we expect in this example? As an initial and intuitive guess, for instance there may be the type of "Ideal" respondents, who endorse surveys, and who are assessed to have understood the interview questions, in contrast to the type of "Skeptic" respondents, with a high understanding level as well, but who do not endorse surveys. We see that these types are hidden, not observable, since they group together people with similar latent characteristics. Are these the only two respondent types that are possible, or stated differently, how many clusters can prevail? For example, there may be a third type of "Believers", who positively respond toward surveys in the interview, without having really understood the interview questions.

We can ask further pertinent questions. What is the share or percentage each of those types may take in the population of interest? According to Table 2, we may expect the "Ideal" attitude type toward surveys to be the dominant one in the target population. How can we detect and characterize such latent classes? How can we assign individuals of our sample to those groups, meaning predict their attitudes toward surveys?

These questions can be addressed by and answered through the use of the LCA technique. The goals of the exploratory LCA approach to cluster analysis are:

1. Identify the unobservable (e.g., survey attitude) types, clusters, or latent classes, which group together subjects or cases that share similar latent characteristics (e.g., different attitudes toward surveys). In other words, the goal of the analysis is to construct a latent typology (e.g., of survey attitude types).

   More specifically, this goal of the classical LCA approach can be subdivided into the following three subtasks:

   a. How many clusters are underlying the data?

   b. What sizes or proportions do the latent classes occupy in the population of reference?

   c. What are the characteristics of the identified types; that is, what do they mean, or how do we interpret them?

2. Classify subjects or cases of a sample into the (e.g., survey attitude) types, clusters, or latent classes. That is, practically identify for each sample unit the unobserved subgroup it may belong or be assigned to.

3. Mathematical Modeling: Formalization of the Data, General Latent Class Model, and Classification by Modal Assignment: In this section, we describe a formal representation of the data and the naturally resulting multinomial sampling distribution for the data. The basic model assumptions underlying LCA and the general unrestricted LC model are introduced. The classification of cases into the clusters using the modal assignment rule is recapitulated.

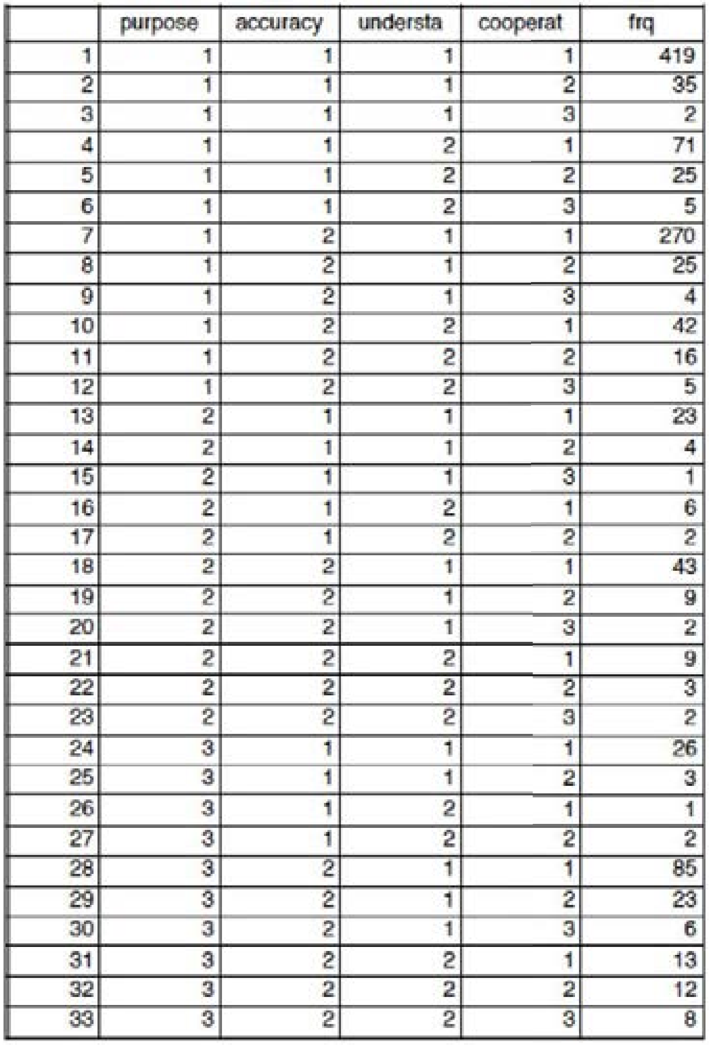
To simplify presentation, we will consider the LC model for three manifest variables or indicators, I_1, I_2, and I_3, which are assumed to be dichotomous. All of the discussion can easily be extended to more than three or polytomous indicators, with two or more than two answer categories.

3.1. Data and Sampling Distribution for the Data: Specifying, and mathematically formalizing, what the data are is not a trivial task sometimes. We do this in our context, where the data arise as follows. We consider $N \in \mathbb{N}$ (set of natural numbers with 0) subjects randomly sampled from a population of reference, who respond to any of the three indicators I_1, I_2, and I_3. The indicators are assumed to be dichotomously scored; score 1 for a positive response (e.g., solve or agree), and score 2 for a negative response (e.g., fail or disagree). In our data set (Table 3), for example, the dichotomous variable ACCURACY is scored 1 for “Mostly True” and 2 for “Not True”.

Let $\mathbf{I} = (I_1, I_2, I_3)$ denote the random vector obtained by viewing I_j for $j = 1, 2, 3$ as random variables that assume the scores 1 or 2. A realization $\mathbf{i} = (i_1, i_2, i_3)$ of $\mathbf{I}$ is called a response pattern. The set of all response patterns is the Cartesian product $\{1, 2\}^3 = \{1, 2\} \times \{1, 2\} \times \{1, 2\}$, which we write as $\mathcal{R}$.

The data are represented by the observed absolute counts $N(\mathbf{i}) \in \mathbb{N}$ of the response patterns $\mathbf{i} \in \mathcal{R}$. This can be summarized in a frequency table shown in Table 3, which lists all observed response patterns along with their absolute frequencies for our example data set.

Table 3. Frequency table of the coded 1982 GSS data.



	purpose	accuracy	understa	cooperat	frq
1	1	1	1	1	419
2	1	1	1	2	35
3	1	1	1	3	2
4	1	1	2	1	71
5	1	1	2	2	25
6	1	1	2	3	5
7	1	2	1	1	270
8	1	2	1	2	25
9	1	2	1	3	4
10	1	2	2	1	42
11	1	2	2	2	16
12	1	2	2	3	5
13	2	1	1	1	23
14	2	1	1	2	4
15	2	1	1	3	1
16	2	1	2	1	6
17	2	1	2	2	2
18	2	2	1	1	43
19	2	2	1	2	9
20	2	2	1	3	2
21	2	2	2	1	9
22	2	2	2	2	3
23	2	2	2	3	2
24	3	1	1	1	26
25	3	1	1	2	3
26	3	1	2	1	1
27	3	1	2	2	2
28	3	2	1	1	85
29	3	2	1	2	23
30	3	2	1	3	6
31	3	2	2	1	13
32	3	2	2	2	12
33	3	2	2	3	8

Having mathematically defined what the data are – the pair

$$(\mathbf{i} = (i_1, i_2, i_3) \in \mathcal{R}, N(\mathbf{i}) \in \mathbb{N})$$

of, both, a combination of the indicators' answer categories and the absolute count of the combination –, the sampling distribution for the data can be derived.

We assume that the subjects give their response patterns independent of each other. The true probability of occurrence $\rho(\mathbf{i}) > 0$ of any response pattern $\mathbf{i} \in \mathcal{R}$ is assumed to stay constant across the subjects. This is an "iid" (independent and identically distributed) replication of the same random experiment, with at least two outcomes, repeated sample size N many times, and as a result, it leads to the binomial or multinomial distribution, for two or more possible outcomes, respectively.

More precisely, the data $\mathbf{x} = \{N(\mathbf{i})\}_{\mathbf{i} \in \mathcal{R}}$ are the realization of a random vector $\mathbf{X} = \{X_{\mathbf{i}}\}_{\mathbf{i} \in \mathcal{R}}$, which is multinomially distributed over $\mathcal{R}$, and has the following formula (for the probability of the data):

$$P(\mathbf{X}=\mathbf{x}) = P\left(X_{(2,2,2)} = N\big((2,2,2)\big), \dots, X_{(1,1,1)} = N\big((1,1,1)\big)\right) = \underbrace{\frac{N!}{\prod_{\mathbf{i}\in\mathcal{R}} N(\mathbf{i})!}}_{\text{multinomial coefficient}} \prod_{\mathbf{i}\in\mathcal{R}} \rho(\mathbf{i})^{N(\mathbf{i})}.$$

Here, $\rho(\mathbf{i}) > 0$ for any $\mathbf{i} \in \mathcal{R}$ with $\sum_{\mathbf{i}\in\mathcal{R}} \rho(\mathbf{i}) = 1$, and $0 \leq N(\mathbf{i}) \leq N$ for any $\mathbf{i} \in \mathcal{R}$ with $\sum_{\mathbf{i}\in\mathcal{R}} N(\mathbf{i}) = N$.

It is interesting to note that the multinomial distribution of the data arises naturally, without any reference to some psychometric or latent variable model. That is, this data distribution holds "of its own volition". It is universally presumed in latent variable modeling.

3.2. Model Assumptions: Three Assumptions A1, A2, and A3 are pertinent to the classical LCA approach, which we summarize next.

A1 We assume that the population of reference can be partitioned into $T \in \mathbb{N}$ mutually exclusive and exhaustive subpopulations $K_1, K_2, \dots, K_T$, also called types, clusters, or classes, with unknown proportions $p(K_t) > 0$ for $1 \leq t \leq T$. Then, in particular, any element of the population belongs to exactly one of these subpopulations, and we have the natural restriction

$$\sum_{t=1}^{T} p(K_t) = 1.$$

Since these types, clusters, or classes generally group together cases (e.g., persons) who share similar unobservable characteristics (e.g., values), they are called latent. In other words, we assume a latent typology underlying the population of reference. Figure 2 gives a pictorial representation of this assumption for four latent classes.

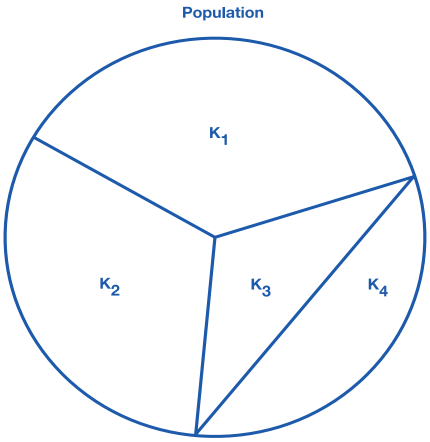


Figure 2. Population partitioned into four subpopulations representing the latent types, clusters, or classes K_1, K_2, K_3, and K_4. Each class is not empty, any two classes have no element in common (mutual exclusiveness), and the union of all classes is the entire population (exhaustiveness). In other words, every element of the population belongs to one, and exactly one, of the four classes.

Technically, the latent classes can be viewed as the realizations of a latent random variable X, and denoted by $X = K_t, t = 1, \ldots, T$.

In the GSS data example, a latent typology may consist of the survey attitude types of, for instance, "Ideal" and "Skeptic" respondents ($T = 2$), and we have $X = K_1$ and $X = K_2$, respectively.

A2 Conditional probabilities are the key quantities that provide the link between the observable level and latent level in latent variable models. Consider the following comparison of observable versus latent components:

Observable	Latent
$\mathbf{i}$	K_t
$\rho(\mathbf{i})$	$p(K_t)$
$\mathbf{I}$	X

Going from the latent level to the observable level involves use of the conditional probabilities $r(\mathbf{I} = \mathbf{i} | X = K_t)$ (r stands for "response"), whereas the transition from the observable to the latent level is accomplished by the conditional probabilities $P(X = K_t | \mathbf{I} = \mathbf{i})$. As we will see later, these two descriptions are interrelated through Bayes' rule.

For the LCM, a special sort of latent variable model, we also do need conditional probabilities. We assume that, for any of the T latent classes, there is a set of conditional probabilities for each of the observed indicators. For example, for the latent class K_1, we have the set of conditional probabilities (for notation, see below) $\{r(I_1 = 1 | X = K_1), r(I_1 = 2 | X = K_1)\}$ for I_1, and $\{r(I_2 = 1 | X = K_1), r(I_2 = 2 | X = K_1)\}$ for I_2.

We also assume that the latent classes are uniquely characterized by these conditional probabilities, both intra and inter classes. Here, "intra classes" means that any class is fully specified by its conditional probabilities. That is, there are no characteristics other than the conditional probabilities of the class that could be of relevance. "Inter classes" means that the latent classes can be uniquely distinguished from one another, solely based on their corresponding sets of conditional probabilities. That is, any two classes cannot have exactly the same conditional probabilities, across the items; there must be one item such that the conditional probabilities for one category of this item differ for the two latent classes.

Thus, we have conditional probabilities

$$0 \leq r(I_l = i_l | X = K_t) \leq 1$$

for any $1 \leq l \leq 3, i_l \in \{1,2\}$, and $1 \leq t \leq T$, and it holds the following second natural restriction (beside $\sum_{t=1}^{T} p(K_t) = 1$)

$$\sum_{i_l=1}^{2} r(I_l = i_l | X = K_t) = 1$$

for any $1 \leq l \leq 3$ and $1 \leq t \leq T$.

In the GSS data example, we can have the conditional probability for an "Ideal" respondent (type "Ideal") to express a belief that the results of surveys are "Mostly True" (indicator ACCURACY).

A3 How can we compute such joint conditional probabilities as $r(I_1 = 1, I_2 = 2, I_3 = 1 | X = K_1)$ or $r(I_1 = 2, I_2 = 1, I_3 = 2 | X = K_2)$? The following third assumption allows to express these joint probabilities by means of the item-wise probabilities introduced in Assumption A2. The assumption states that if the latent variable is held fixed, the relationships between the item responses are random and not anymore systematic.

Thus, we assume that, within any of the T latent classes, the observed scores to the indicators are independent. That is,

$$r(I_1 = i_1, I_2 = i_2, I_3 = i_3 | X = K_t)$$

$$= \prod_{l=1}^{3} r(I_l = i_l | X = K_t)$$

for any $i_l \in \{1, 2\}, 1 \leq l \leq 3,$ and $1 \leq t \leq T.$

This is the assumption of local independence, which is fundamental in classical LCA.

Two remarks are in order. First, the assumption of local independence can also be viewed as a parameter reduction. In the example with the three dichotomous indicators, there are seven independent joint conditional probabilities $r(I_1 = i_1, I_2 = i_2, I_3 = i_3 | X = K_t)$ for any latent class K_t, $t = 1, \ldots, T$. Under the assumption of local independence, this is reduced to three independent individual conditional probabilities $r(I_l = i_l | X = K_t)$ for any latent class K_t, $t = 1, \ldots, T$. Second, the assumption of local independence is restrictive, from an empirical viewpoint. This assumption requires that the latent class does not change, that is, remains the same, when answering all items of the test. This precludes transfer (e.g., learning) effects that may empirically occur (and may alter the latent class) during testing.

3.3. The General Unrestricted Latent Class Model: Assumptions A1, A2, and A3 suffice to define the general unrestricted LC model. The general unrestricted (T-class) latent class model is a multinomial probability model (cf. Bishop et al., 2007), parametrizing the multinomial cell probabilities $\rho(\mathbf{i})$ for $\mathbf{i} = (i_1, i_2, i_3) \in \mathcal{R}$ (see Section 3.1) as follows:

$$\rho(\mathbf{i}) =^{A1,W} \sum_{t=1}^{T} \{p(K_t) r(I_1 = i_1, I_2 = i_2, I_3 = i_3 | X = K_t)\}$$

$$=^{A2,A3} \sum_{t=1}^{T} \left\{ p(K_t) \prod_{l=1}^{3} r(I_l = i_l | X = K_t) \right\}.$$

W denotes the Law of Total Probability from probability theory (Gut, 2013), which states that if $\{B_i, i = 1, \dots, n\}$ is a partition of the sample space, with $P(B_i) > 0$ for $i = 1, \dots, n$, then for any event A,

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i).$$

This can be recapped in terms of the partition $\{K_t, 1 \leq t \leq T\}$ according to Assumption A1. In our context, $n = T$, B_i corresponds to $X = K_t$ (or K_t), and A is $\mathbf{I} = \mathbf{i}$ (or $(I_1 = i_1, I_2 = i_2, I_3 = i_3)$).

The LCM with T latent classes for three dichotomous items contains

$$\underbrace{(T-1)}_{\text{independent class parameters (natural restriction)}} + \underbrace{T}_{\text{per class}} \cdot \underbrace{3}_{\text{per item}} \cdot \underbrace{(2-1)}_{\text{independent item parameter (natural restriction)}}$$

independent model parameters that need to be estimated from the data. These parameters are:

$$\begin{gathered} p(K_1), p(K_2), \dots, p(K_{T-1}), \\ r(I_1 = 1|X = K_1), r(I_2 = 1|X = K_1), r(I_3 = 1|X = K_1), \\ r(I_1 = 1|X = K_2), r(I_2 = 1|X = K_2), r(I_3 = 1|X = K_2), \\ \vdots \\ r(I_1 = 1|X = K_T), r(I_2 = 1|X = K_T), r(I_3 = 1|X = K_T). \end{gathered}$$

In the sequel, we assume that all model parameters are summarized in a parameter vector

$$\theta = \big(p(K_1), \dots, p(K_T), r(I_1 = 1|X = K_1), \dots, r(I_3 = 2|X = K_T)\big),$$

which ranges over the parameter space $\Theta = (0,1]^T \times [0,1]^{6T}$.

In particular,

$$\rho(\mathbf{i}) =^{\text{LCM}} \sum_{t=1}^{T} \left\{ p(K_t) \prod_{l=1}^{3} r(I_l = i_l|X = K_t) \right\} \equiv f(\theta)$$

depends on θ, and we therefore also write $\rho_\theta(\mathbf{i})$. Mathematically, we treat (known) θ to be fixed, but arbitrary. Statistically, we need to estimate (unknown) θ based on the data (see Section 4). According to the equation $\rho_\theta(\mathbf{i}) = f(\theta)$, each θ specifies its own LCM, and we can identify both an LCM and its parameter vector θ, and simply speak of the "LCM θ". So, what is an LCM? An LCM is a parameter vector $\theta \in \Theta$, which is linked to the sampling distribution of the data via $\rho_\theta(\mathbf{i}) = f(\theta)$. This procedure of parametrizing the multinomial cell probabilities through a function of the model parameters is not specific to LCA. In latent variable models in general, $\rho(\mathbf{i})$ is assumed to be a function $\Psi(\eta)$ of model parameters η, and it is the choices of Ψ and η that define the respective psychometric model.

3.4. Classification of Cases by Modal Assignment: How can we assign individuals to the unobservable classes? What parameters can we use for this purpose? Note that, first and foremost, $r(I = i|X = K_t)$ are theoretical parameters. Their direct practical value is limited because the conditioning is on the latent variable $X = K_t$, which we cannot observe. In contrast, directly more useful especially for classification are the parameters $P(X = K_t|I = i)$, in which the roles of the manifest and latent variables are interchanged, because in this case we condition on the observable variable $I = i$. Both sorts of parameters are interrelated. The formula from probability theory that allows to interchange the roles of events in conditional probabilities is Bayes' rule (see below).

Beside the characterization of the latent classes and latent variable using the conditional probabilities postulated in Assumption A2, LC cluster analysis also allows for the assignment of cases to the latent classes based upon posterior membership probabilities. For $1 \leq t \leq T$ and $\mathbf{i} = (i_1, i_2, i_3) \in \mathcal{R}$, the posterior membership probabilities (as opposed to the prior probabilities $p(K_t)$) are

$$P(X = K_t|\mathbf{I} = \mathbf{i}) =^{\text{Bayes}'\text{rule}} \frac{p(K_t)r(\mathbf{I} = \mathbf{i}|X = K_t)}{\rho_\theta(\mathbf{i})}$$

$$=^{\text{A1,A2,A3,W}} \frac{p(K_t)\prod_{l=1}^{3} r(I_l = i_l|X = K_t)}{\sum_{t'=1}^{T}\{p(K_{t'})\prod_{l=1}^{3} r(I_l = i_l|X = K_{t'})\}}.$$

Bayes' rule (Gut, 2013) states that for two events A and B with $P(A) > 0$ and $P(B) > 0$ it holds

$$P(B|A) = P(A|B) \cdot \underbrace{\frac{P(B)}{P(A)}}_{\text{correction term}} .$$

That is, the order of conditioning the events can be changed, from "$A|B$" to "$B|A$", as long as the correction term is invoked. In our context, Bayes' rule is applied for the choice of events "$A\colon \mathbf{I} = \mathbf{i}$" and "$B\colon X = K_t$".

Note that the posterior membership probabilities $P(X = K_t|\mathbf{I} = \mathbf{i})$ solely depend on the model parameters $p(K_t)$ and $r(I_l = i_l|X = K_t)$ of an LCM. In particular, these posterior probabilities can be directly estimated from the fitted statistical model. This is why we speak of model-based classification, as opposed to distance-based methods (e.g., Kaufman and Rousseeuw, 2005).

Having introduced the relevant probabilities, the most common classification rule is modal assignment. Assign all cases with a given response pattern $\mathbf{i} \in \mathcal{R}$ to a latent class K_{t_0} $(1 \leq t_0 \leq T)$ for which the posterior membership probability $P\left(X = K_{t_0}\middle|\mathbf{I} = \mathbf{i}\right)$ is largest. That is,

$$P\left(X = K_{t_0}\middle|\mathbf{I} = \mathbf{i}\right) = \max_{1 \leq t \leq T} P\,(X = K_t|\mathbf{I} = \mathbf{i}).$$

An example may help to illustrate the modal assignment. Suppose we have the following posterior membership probabilities $P(K_1|\mathbf{i}) = 0.18$, $P(K_2|\mathbf{i}) = 0.27$, and $P(K_3|\mathbf{i}) = 0.55$ (these numbers must sum up to 1), for $T = 3$ latent classes. According to modal assignment, all subjects of the sample with the response pattern $\mathbf{i}$ are put into cluster K_3, because this cluster has the largest value (0.55).

We could also classify cases into clusters proportionally, with the sampling weights for the clusters given by their posterior membership probabilities. In the example, we would draw K_1 with probability 0.18, K_2

with probability 0.27, and K_3 with probability 0.55. That is, there is some fractional chance specified by the posterior membership probability of landing in each of the clusters.

In both cases, modal assignment and proportional assignment, we have instances of model-based classification.

4. Inferential Statistics: Parameter Estimation and Model Testing:

Reconsider Figure 1. So far, we have treated the mathematics with presupposed fixed, but arbitrary parameters of the LCM. Now, we deal with the statistics, the problem of specifying empirically plausible values for the parameters based on the data. The goals are threefold. We discuss parameter estimation by maximum likelihood, choosing plausible values for the parameters; goodness-of-fit testing based on the deviance, whether a particular choice of estimated parameters provides an adequate fit; and model selection using the criteria AIC and SIC, among competing sets of estimated parameters that do fit the data. Three classic references for the statistical techniques and results discussed in this section are Bishop et al. (2007), Cramér (1999), and Lehmann and Romano (2005), and for model selection by Burnham and Anderson (2002).

4.1. Maximum Likelihood Estimation: In Section 3.1, we have derived the multinomial formula for the probability $P(\mathbf{X}=\mathbf{x})$ of the data $\mathbf{x}=\{N(\mathbf{i})\}_{\mathbf{i}\in\mathcal{R}}$. According to Section 3.3, we have modeled the multinomial cell probabilities $\rho(\mathbf{i})=\rho_\theta(\mathbf{i})$. Thus, the probability of the data can be viewed as a function of the model parameters θ, with the data held fixed. Then, in maximum likelihood estimation, the parameter values are selected in such a way that with this choice of values the probability or likelihood of the given data is maximized. That is, we choose those values for the parameters that render the available data most probable.

More precisely, the kernel of the likelihood function (omitting the data dependent, constant multinomial coefficient), as a function of θ, for fixed data $\mathbf{x}=\{N(\mathbf{i})\}_{\mathbf{i}\in\mathcal{R}}$, is

$$\mathcal{L}(\theta;\mathbf{x}) \coloneqq \prod_{\mathbf{i}\in\mathcal{R}} \rho_\theta(\mathbf{i})^{N(\mathbf{i})}$$

$$= \prod_{\mathbf{i}\in\mathcal{R}} \left\{ \sum_{t=1}^{T} \left\{ p(K_t) \prod_{l=1}^{3} r(I_l = i_l | X = K_t) \right\} \right\}^{N(\mathbf{i})} .$$

This function $\mathcal{L}(.;\mathbf{x})$ of θ, for constant data $\mathbf{x}$, must be globally maximized over the parameter space Θ. This is called the maximum likelihood (ML) estimation problem.

The maximum of the kernel of the likelihood function cannot, in general, be obtained by analytical methods. For instance, setting equal to zero the first-order partial derivatives of this function with respect to the parameters results in a non-linear system of equations. Numerical optimization methods are required. Popular iterative methods for solving the ML estimation problem are Expectation-Maximization (EM) and Newton-Raphson (NR) type algorithms (e.g., McLachlan and Krishnan, 2008). In this paper, we use the program Latent GOLD® to do ML estimation. This program implements a hybrid EM-NR algorithm (for details, see Vermunt and Magidson, 2016).

We want to note that one application of a local optimization routine does not ensure the achievement of a global maximum. There may be several local maxima. On the contrary, the local optimization routine must be repeated multiple times, with different start values for the parameters, until a global maximum has been achieved.

In the sequel, let $\hat{\theta} \in \Theta$ be the ML estimate of the parameter vector θ, that is,

$$\mathcal{L}(\hat{\theta};\mathbf{x}) = \max_{\theta\in\Theta} \mathcal{L}(\theta;\mathbf{x})\,.$$

Viewed as a random (variable) vector, $\hat{\theta}$ is also called the ML estimator for θ.

4.2. Goodness-of-Fit Testing: The goodness-of-fit of an estimated LCM can be tested based on the log-likelihood ratio statistic (deviance)

$$L^2(\theta;\mathbf{x}) \coloneqq 2\sum_{\mathbf{i}\in\mathcal{R}} \left\{N(\mathbf{i})\ln\left(\frac{N(\mathbf{i})}{N\rho_\theta(\mathbf{i})}\right)\right\}.$$

We test the null hypothesis that some LCM θ can be used to explain the data, that is, $\rho(\mathbf{i}) = \rho_\theta(\mathbf{i})$ for all $\mathbf{i} \in \mathcal{R}$, against the alternative hypothesis that no LCM fits the data. For this test problem, we hope for non-significance, that is, we do not want to reject the null hypothesis (albeit the type II error can be high and is not controlled). If the model is correct (i.e., under the null hypothesis), given certain regularity conditions are satisfied (Birch, 1964; Bishop et al., 2007; Read and Cressie, 1988), the random variable $L^2(\hat{\theta};\mathbf{x})$, with $\hat{\theta}$ the ML estimator, is asymptotically (as $N \to \infty$) chi-squared with the number df of degrees of freedom

$$df = (|\mathcal{R}|-1) - \{(T-1) + T\cdot[3\cdot(2-1)]\}.$$

That is, the number of degrees of freedom of the asymptotic chi-square distribution equals the number of possible response patterns minus one (i.e., number of degrees of freedom in the data), $|\mathcal{R}|-1$, minus the number of independent model parameters that were estimated, $(T-1) + T\cdot[3\cdot(2-1)]$.

The name of the statistic $L^2(\hat{\theta};\mathbf{x})$ stems from the fact that it can also be written in the form of, for large N ($\approx^d$ means “approximately distributed as” and χ^2 is the chi-square distribution),

$$L^2(\hat{\theta};\mathbf{x}) \\ = -2\ln\left(\frac{\prod_{\mathbf{i}\in\mathcal{R}}\rho_{\hat{\theta}}(\mathbf{i})^{N(\mathbf{i})}}{\prod_{\mathbf{i}\in\mathcal{R}}\left(\frac{N(\mathbf{i})}{N}\right)^{N(\mathbf{i})}}\right) \approx^d \chi^2_{df}.$$

The log of the ratio of two (globally) maximized likelihoods is computed, in the numerator for the LCM with the ML estimate $\hat{\theta}$, and in the denominator for the saturated general multinomial distribution with the ML estimates $\widehat{\rho(\mathbf{i})} = N(\mathbf{i})/N$ for $\mathbf{i} \in \mathcal{R}$.

4.3. Model Selection: It is not valid to compare LC models with different numbers of latent classes by general likelihood ratio tests, which would mean restricting parameters (class proportions) to boundary values (to be zero). This would violate a regularity condition, for the true parameter vector must be an interior point of the parameter space (e.g., Birch, 1964; Bishop et al., 2007).

Therefore, we select among competing LC models using the Akaike information criterion (AIC; Akaike, 1973), with the ML estimate $\hat{\theta}$,

$$\begin{aligned} \mathrm{AIC} &\coloneqq -2\ln\left(\mathcal{L}(\hat{\theta};\mathbf{x})\right) \\ &+2\cdot \text{number of independent model parameters}, \end{aligned}$$

and the Schwarz information criterion (SIC; Schwarz, 1978)

$$\begin{aligned} \mathrm{SIC} &\coloneqq -2\ln\left(\mathcal{L}(\hat{\theta};\mathbf{x})\right) \\ &+\ln(N)\cdot \text{number of independent model parameters}. \end{aligned}$$

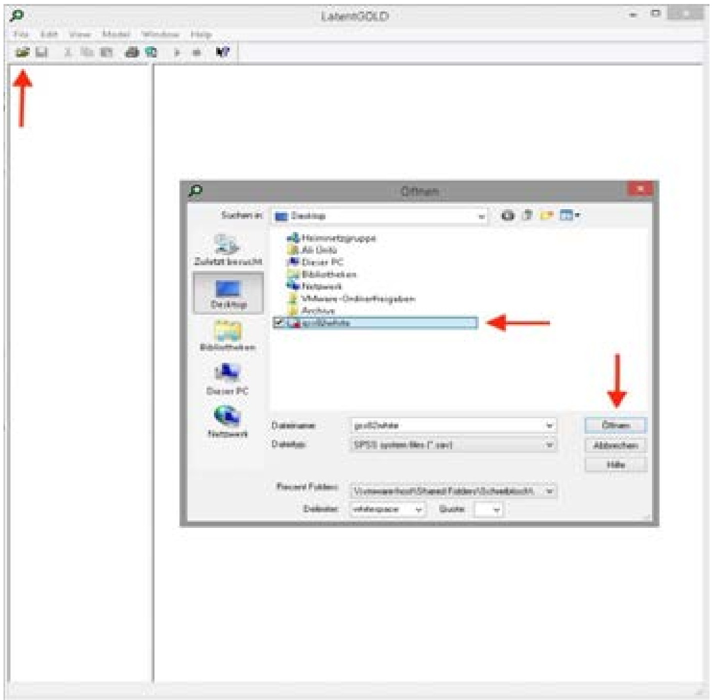
We select an LC model with the smallest value of AIC or SIC.

For further reading, a comprehensive analysis and comparison of model selection criteria for LCA are discussed in Lin and Dayton (1997), Lin (2006), and Lin (2015).

5. Application of Latent GOLD® to GSS Data:

We analyze the 1982 GSS example data set described in Section 2 using the Windows-based software Latent GOLD® for LC cluster analysis. Latent GOLD® is commercial software and can be obtained from Statistical Innovations at www.statisticalinnovations.com. They are the makers of Latent GOLD®. There is a freely available demo version of this software. The GSS data set used in the analyses of this section is also distributed as the SPSS system file gss82white.sav with the demo version.

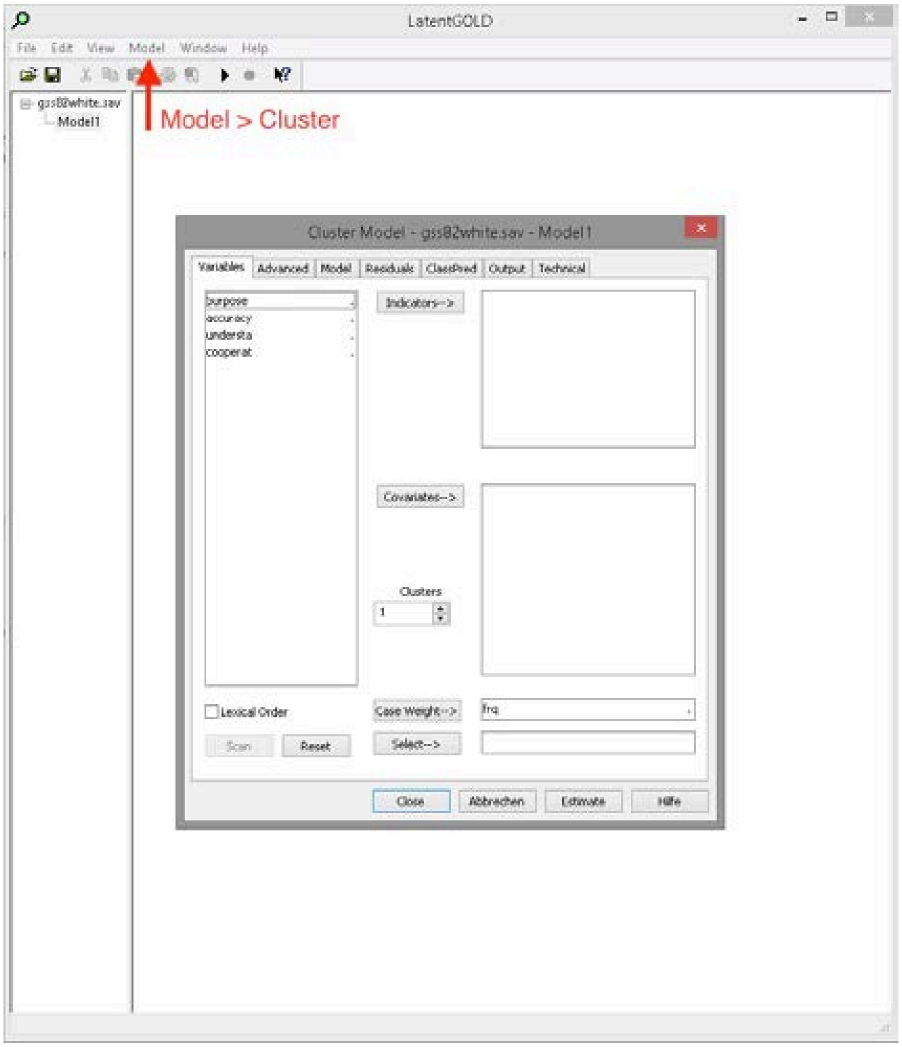
The first thing to know is how you can get your data into Latent GOLD®. Via the menu, choose "File > Open...", which brings up the dialog box shown in Figure 3 and where we select our data set, the file gss82white.sav. If necessary, change to the appropriate directory of the data file that you want to open.



Dateiname: gss82white
Dateityp: SPSS system files (*.sav)
Recent Folder: \\vmware-host\Shared Folders\Schreibtisch\...
Delimiter: whitespace Quote:

Figure 3. "File > Open..." dialog box. The red arrows show the File Open shortcut, the selected data file, and the Open button, respectively.

After having loaded the data into Latent GOLD®, via the menu, choose "Model > Cluster". This brings up the analysis dialog box for LC cluster analysis displayed in Figure 4. You can also right or double click "Model1" and select "Cluster" to open up the analysis dialog box.



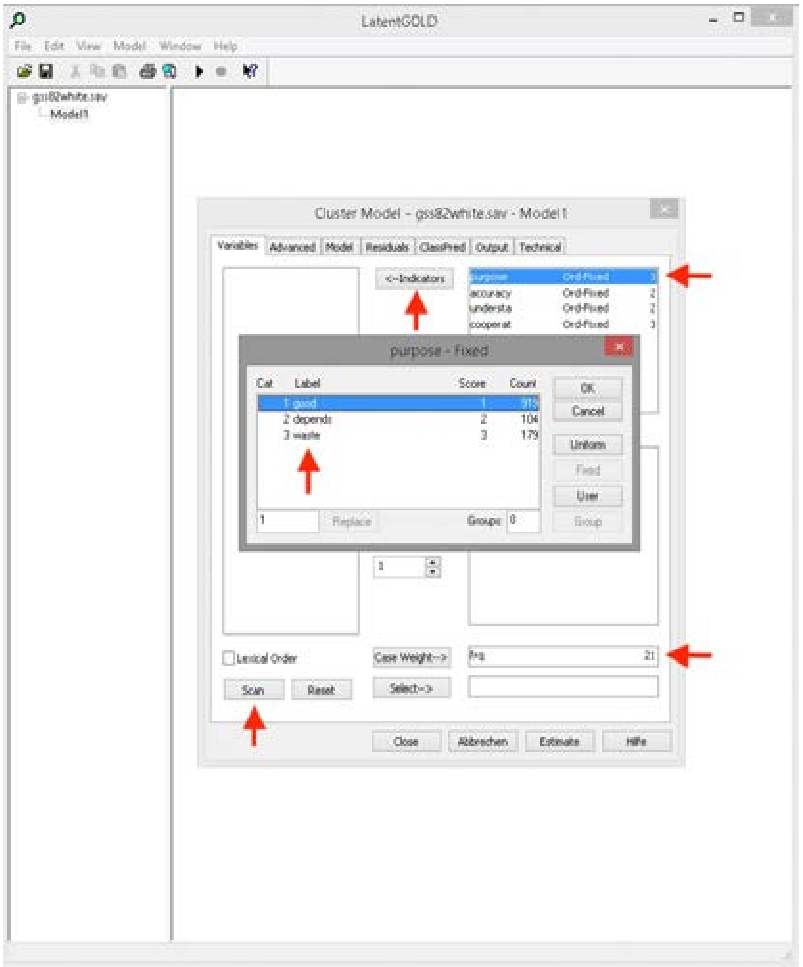
Clusters: 1

Case Weight-->: frq

Figure 4. "Model > Cluster" dialog box for LC cluster analysis.

In Figure 5, we select the variables for the analysis. For this, we use the mouse to highlight the four variables in the left-hand side variable box, and then press the "Indicators" button to place the variables in the right-hand side indicator box. To scan the data file, click on the "Scan" button.

Now, the variables or data can be inspected. For example, double clicking the PURPOSE variable allows to view its categories, labels, scores, and counts. We have a case weight variable "frq" in our SPSS data file, which automatically appears in the "Case Weight" box in Latent GOLD®.



```
purpose     Ord-Fixed   3
accuracy    Ord-Fixed   2
understa    Ord-Fixed   2
cooperat    Ord-Fixed   3
```

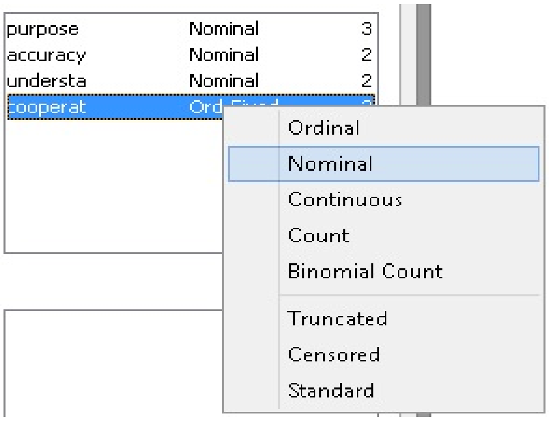
Cat	Label	Score	Count
	1 good	1	915
	2 depends	2	104
	3 waste	3	179

Groups: 0

Case Weight--> frq 21

Figure 5. Select the indicators and automatically set the case weight for LC cluster analysis, then scan the data, and explore information about a variable (e.g., the PURPOSE variable).

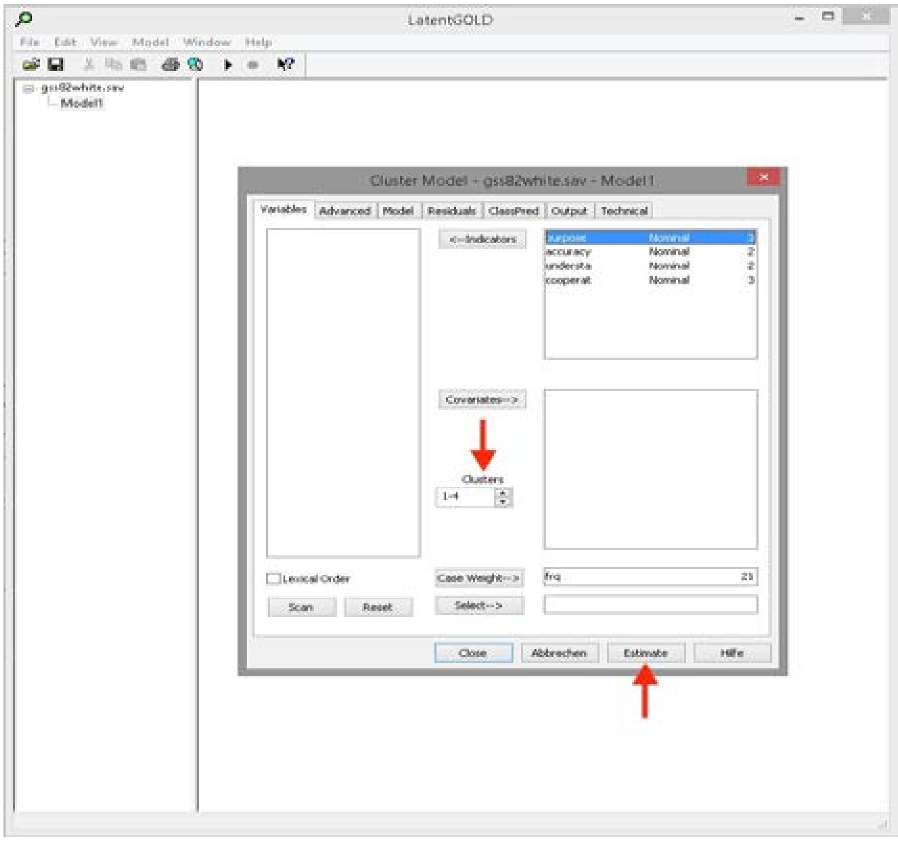
For the general LCM, we change the scale type of each variable from default "Ord-Fixed" (ordinal) to "Nominal", by right click on each variable (or all variables simultaneously if jointly highlighted) in the indicator box. This is shown in Figure 6.



```
purpose     Nominal     3
accuracy    Nominal     2
understa    Nominal     2
cooperat    Ord-Fixed   3
```

Figure 6. Pop-up menu for variable scale type. Right click a variable in the indicator box to change its scale type. For the general LC cluster analysis model, we set the scale type of all indicators to "Nominal".

Latent GOLD® allows to estimate LC models with different numbers of clusters in one run. Simply specify the number of latent classes to be estimated in the "Clusters" box, shown in Figure 7. In our example, "1-4" clusters are specified, that is, we estimate the 1-class LCM, 2-class LCM, 3-class LCM, and 4-class LCM, all in one go. To start the estimation process, we have to click the "Estimate" button at the bottom of the dialog box.



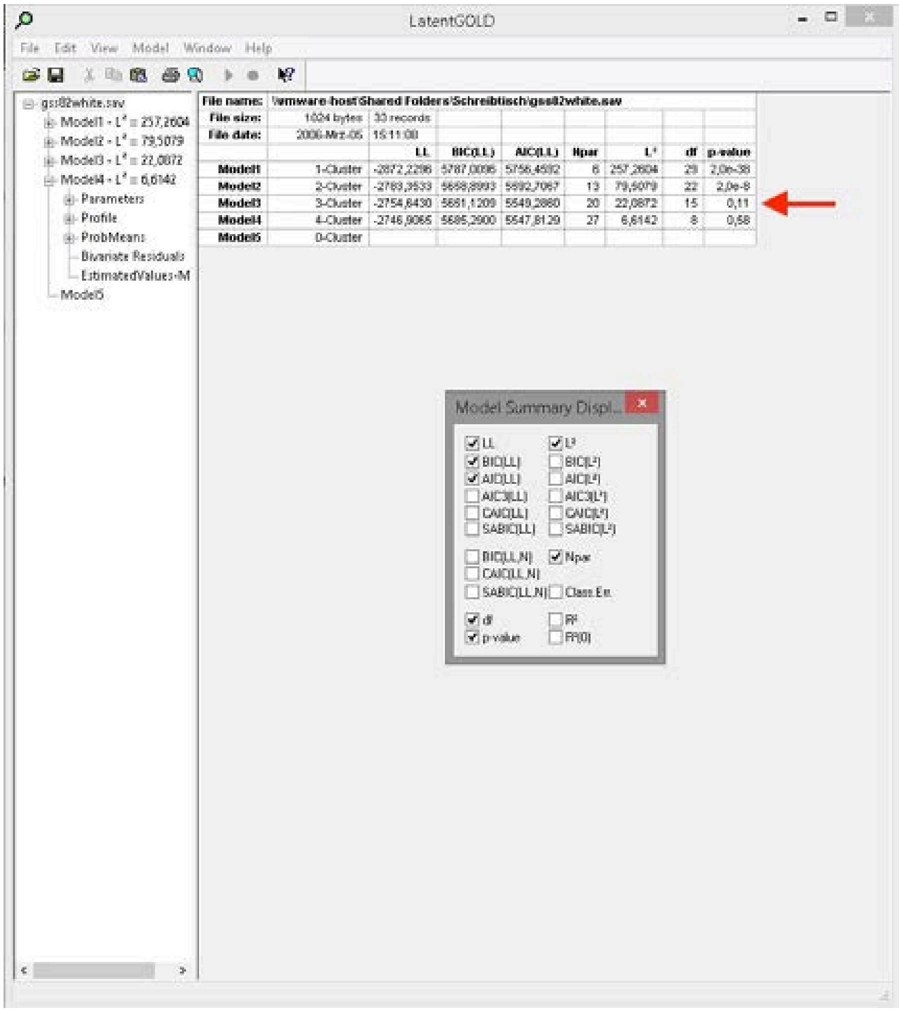
purpose	Nominal	3
accuracy	Nominal	2
understa	Nominal	2
cooperat	Nominal	3

Clusters 1-4

Case Weight--> frq 21

Figure 7. LC cluster analysis dialog box, with specified numbers of "Clusters" $T = 1, 2, 3, 4$. To estimate these four LC models, the "Estimate" button must be clicked.

After the models have been estimated, click on the data file name gss82white.sav on the left-hand side of the display (Outline Pane) to obtain a summary of all models fitted on that data on the right-hand side of the display (Contents Pane). Right click in the Contents Pane to access the Model Summary Display, containing further statistics that can be displayed if selected. Your screen should then look like in Figure 8.



- gss82white.sav
  - Model1 - L² = 257,2604
  - Model2 - L² = 79,5079
  - Model3 - L² = 22,0872
  - Model4 - L² = 6,6142
    - Parameters
    - Profile
    - ProbMeans
    - Bivariate Residuals
    - EstimatedValues-M
  - Model5

File name:	\\vmware-host\Shared Folders\Schreibtisch\gss82white.sav							
File size:	1024 bytes	33 records						
File date:	2006-Mrz-05	15:11:08						
		LL	BIC(LL)	AIC(LL)	Npar	L²	df	p-value
Model1	1-Cluster	-2872,2298	5787,0086	5756,4592	6	257,2604	29	2,0e-38
Model2	2-Cluster	-2783,3533	5668,8993	5592,7067	13	79,5079	22	2,0e-8
Model3	3-Cluster	-2754,6430	5661,1209	5549,2860	20	22,0872	15	0,11
Model4	4-Cluster	-2746,9065	5685,2900	5547,8129	27	6,6142	8	0,58
Model5	0-Cluster							

Model Summary Displ...

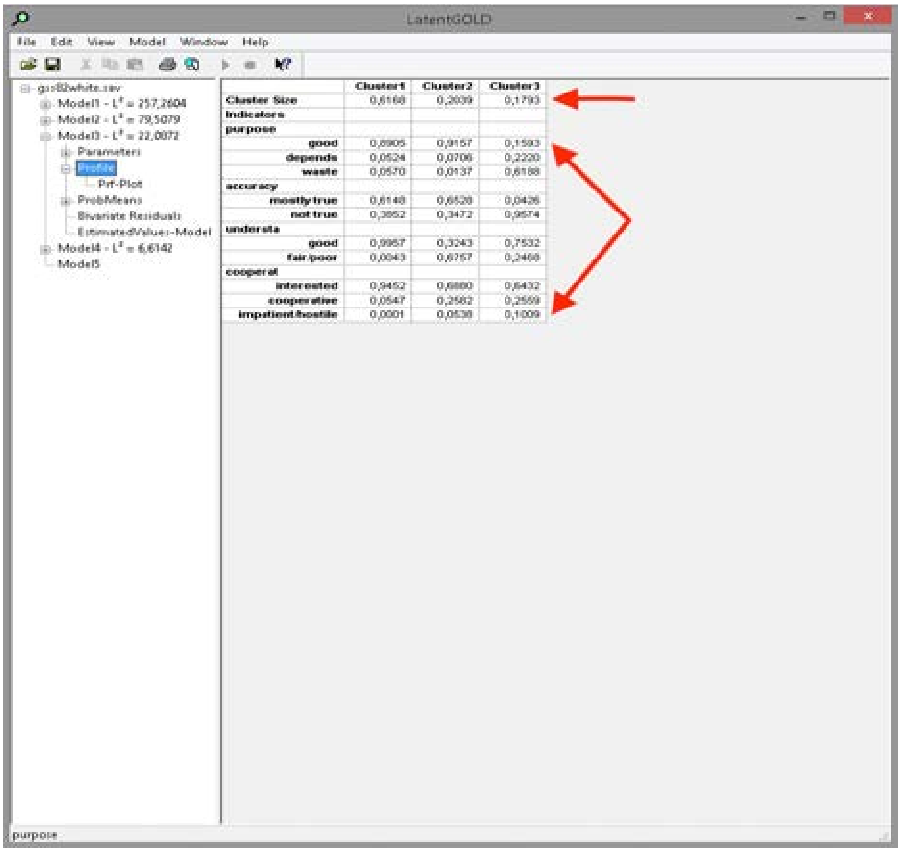
☑ LL	☑ L²
☑ BIC(LL)	☐ BIC(L²)
☑ AIC(LL)	☐ AIC(L²)
☐ AIC3(LL)	☐ AIC3(L²)
☐ CAIC(LL)	☐ CAIC(L²)
☐ SABIC(LL)	☐ SABIC(L²)
☐ BIC(LL,N)	☑ Npar
☐ CAIC(LL,N)	
☐ SABIC(LL,N)	☐ Class.Err.
☑ df	☐ R²
☑ p-value	☐ R²(0)

Figure 8. Model summary output for the fitted 1-class, 2-class, 3-class, and 4-class LC models.

In Figure 8, the "Model Summary Display" is also shown, where additional statistics can be selected. "LL" stands for the log-likelihood $\ln\left(\mathcal{L}(\hat{\theta};\mathbf{x})\right)$; "BIC(LL)" and "AIC(LL)" are the Schwarz and Akaike information criteria SIC and AIC, respectively; "Npar" stands for the number of independent model parameters; "L^2" is the deviance $L^2(\hat{\theta};\mathbf{x})$; and "p-value" is the significance probability $p = P\left(\chi^2_{df} > L^2(\hat{\theta};\mathbf{x})\right)$. "Model3" is the best model according to the heuristic model selection rule (cf. Figure 16).

We can use the output shown in Figure 8 to determine the number of clusters. The complete independence model ("Model1") and the 2-class model ("Model2") are rejected right away. They yield larger BIC (SIC) and AIC values or have highly significant p-values (cf. Figure 16). "Model3" (three clusters) and "Model4" (four clusters) are the competing LC models. We apply the heuristic rule to select among the two (cf. Figure 16). "Model3" and "Model4" have non-significant p-values of 0.11 and 0.58, respectively. That is, we cannot reject these models, for example, at an α level of 0.05. "Model3" has the minimum number of LCs ($T = 3$ versus 4) such that non-significance holds and is the one that is more parsimonious (Npar = 20 versus 27). Using this heuristic decision rule, “Model3”, the 3-class LC model, is the favored model. In the sequel, we will use this model to exemplify the software. Nonetheless, “Model4” could also be considered, but for illustration purposes we will restrict our attention to “Model3”.

We can view the estimated model parameters of the 3-class LC model, see Figure 9. In the Outline Pane, click the expand icon “+” next to “Model3” and highlight the list element “Profile”. Now, the estimated model parameters are displayed in the Contents Pane. You can also click the expand icon next to “Profile” to see the subordinate entry “Prf-Plot” in the Outline Pane in Figure 9.

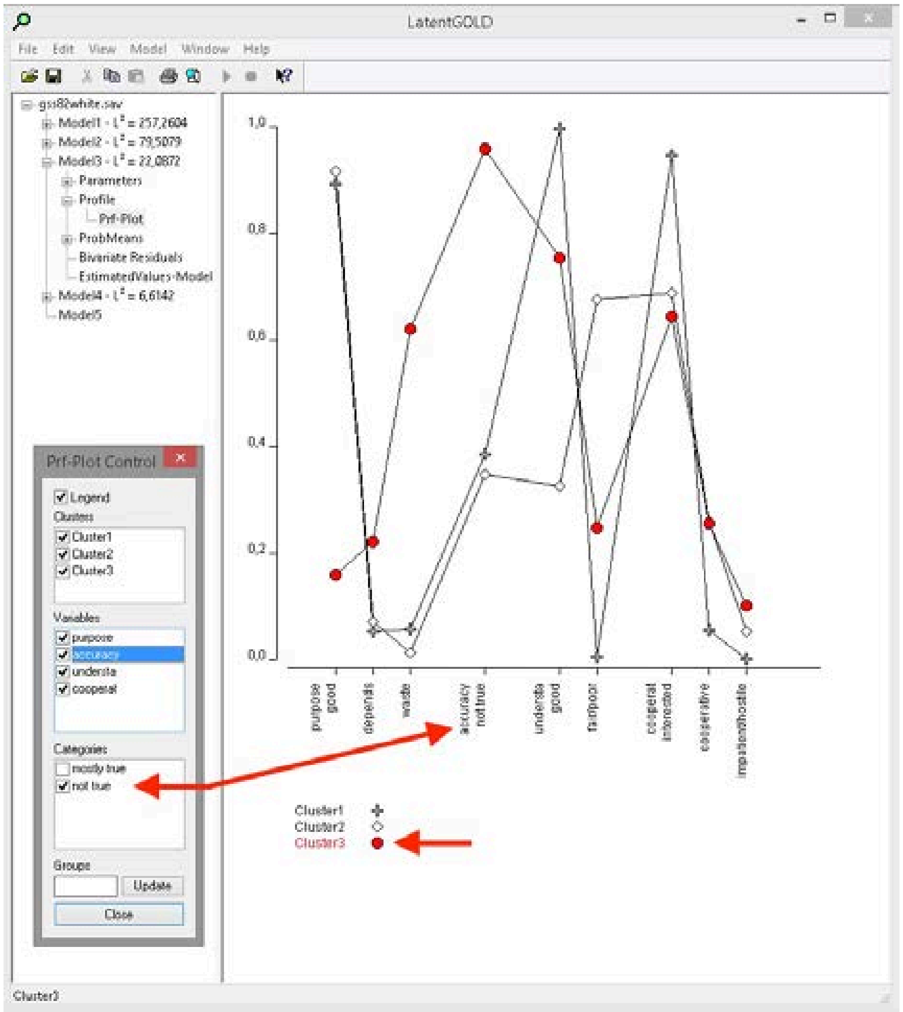


	Cluster1	Cluster2	Cluster3
Cluster Size	0,6168	0,2039	0,1793
Indicators			
purpose			
good	0,8905	0,9157	0,1593
depends	0,0524	0,0706	0,2220
waste	0,0570	0,0137	0,6188
accuracy			
mostly true	0,6148	0,6528	0,0426
not true	0,3852	0,3472	0,9574
understa			
good	0,9957	0,3243	0,7532
fair/poor	0,0043	0,6757	0,2468
cooperat			
interested	0,9452	0,6880	0,6432
cooperative	0,0547	0,2582	0,2559
impatient/hostile	0,0001	0,0538	0,1009

Figure 9. Profile output for the 3-class LC model showing the cluster sizes $p(K_t)$, for $t = 1, 2, 3$, and conditional item probabilities $r(I_l = i_l | X = K_t)$, for $t = 1, 2, 3$, $l = 1, 2, 3, 4$, $i_1, i_4 \in \{1, 2, 3\}$, and $i_2, i_3 \in \{1, 2\}$. According to the latter parameters, we can interpret “Cluster1” as the “Ideal” respondent type, “Cluster2” are the “Believers”, and “Cluster3” is the “Skeptic” type.

We observe that the estimated cluster sizes are $p(K_1) = 0.6168$, $p(K_2) = 0.2039$, and $p(K_3) = 0.1793$ for “Cluster1”, “Cluster2”, and “Cluster3”, respectively. The conditional item probabilities $r(I_l = i_l | X = K_t)$, that is, the numbers between the two red arrows, can be used to characterize or interpret, and distinguish the unobserved clusters. We can see that latent class K_1 is the type of “Ideal” respondents, who endorse surveys (PURPOSE = “Good”, 0.8905; ACCURACY = “Mostly True”, 0.6148; COOPERATION = “Interested”, 0.9452), and who are assessed to have understood the interview questions (UNDERSTANDING = “Good”, 0.9957). In contrast, latent class K_3 constitutes the type of “Skeptic” respondents, with a high understanding level (UNDERSTANDING = “Good”, 0.7532), but who do not endorse surveys (PURPOSE = “Waste”, 0.6188; ACCURACY = “Not True”, 0.9574; COOPERATION = “Impatient, Hostile”, 0.1009). Latent class K_2 is the type of “Believers”, who positively respond toward surveys in the interview (PURPOSE = “Good”, 0.9157; ACCURACY = “Mostly True”, 0.6528), without having really understood the interview questions (UNDERSTANDING = “Fair, Poor”, 0.6757).

The conditional item probabilities $r(I_l = i_l|X = K_t)$ can be represented graphically using a profile plot. This is shown in Figure 10.



☑ Legend

Clusters
☑ Cluster1
☑ Cluster2
☑ Cluster3

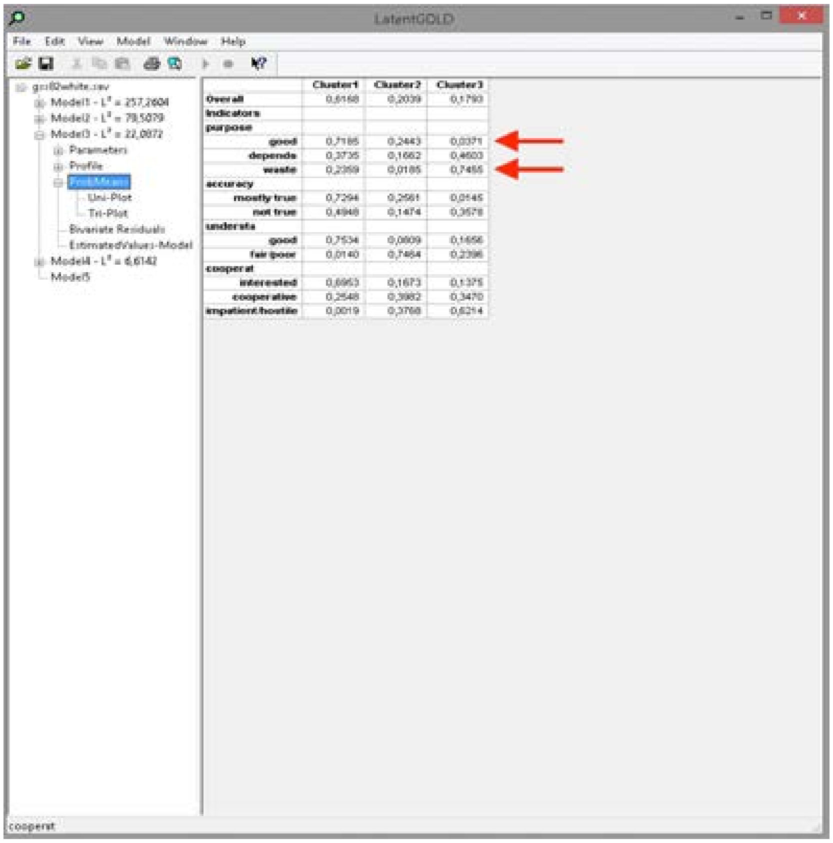
Variables
☑ purpose
☑ accuracy
☑ understa
☑ cooperat

Categories
☐ mostly true
☑ not true

Figure 10. Profile plot for the 3-class LC model. On the x-axis are the indicators with their response categories, $I_l = i_l$. On the y-axis the conditional item probabilities $r(I_l = i_l|X = K_t)$ are put, with each of the three latent classes being represented by a zigzag line. “Cluster3”, the “Skeptic” respondent type, is selected and highlighted in red. The profile plot Control Panel is displayed, for customizing the legend, clusters, variables, and categories of the plot.

In Figure 10, highlight "Prf-Plot" in the Outline Pane to produce the profile plot in the Contents Pane. To invoke the profile plot Control Panel "Prf-Plot Control" right click on the plot or select "View > Plot Control" in the menu. The Control Panel allows to customize the profile plot. For example, select the variable ACCURACY in the "Variables" box to see the checked categories of this variable. By default, for dichotomous variables only the last category is displayed (see the ACCURACY variable with its category "Not True"). This can be changed in the Control Panel (see the dichotomous variable UNDERSTANDING). Each latent class is represented by a zigzag line (we have three lines). The conditional item probabilities are placed on the y-axis. The variables with their categories are vertically put on the x-axis. Any of the clusters (e.g., "Cluster3") can be selected by clicking on the symbol next to the cluster name at the bottom of the plot, and the corresponding profile in the plot is then highlighted in red.

In Figure 11, under "ProbMeans", we have the reversed conditional probabilities $P(X = K_t|I_l = i_l)$ for landing in cluster K_t given the indicator category $I_l = i_l$. Click the expand icon "+" next to "ProbMeans" to see the available plot options.

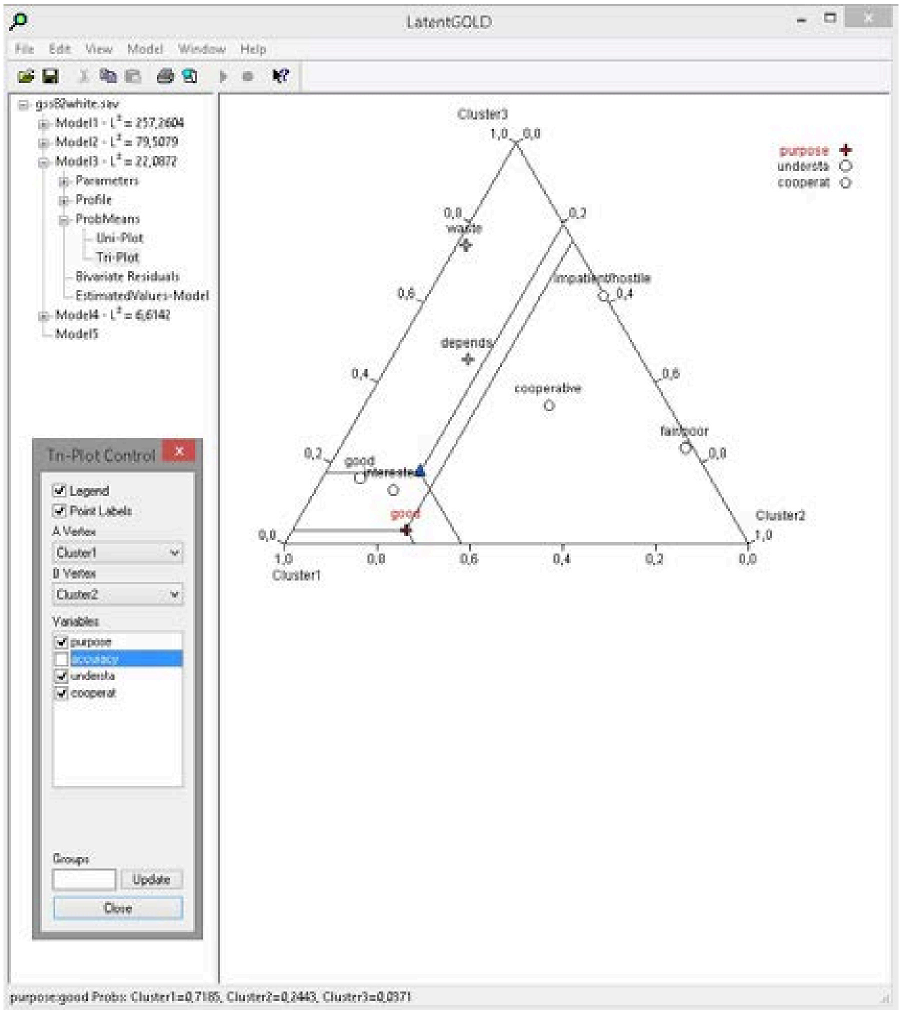


	Cluster1	Cluster2	Cluster3
Overall	0,6168	0,2039	0,1793
Indicators			
purpose			
good	0,7185	0,2443	0,0371
depends	0,3735	0,1662	0,4603
waste	0,2359	0,0185	0,7455
accuracy			
mostly true	0,7294	0,2561	0,0145
not true	0,4946	0,1474	0,3578
understa			
good	0,7534	0,0809	0,1656
fair/poor	0,0140	0,7464	0,2396
cooperat			
interested	0,6953	0,1673	0,1375
cooperative	0,2548	0,3982	0,3470
impatient/hostile	0,0019	0,3768	0,6214

Figure 11. Output of indicator-score specific posterior membership probabilities $P(X = K_t|I_l = i_l)$. The rows sum up to one. The categories "Good" and "Waste" of the PURPOSE variable are good indicators for the "Ideal" (0.7185) and "Skeptic" (0.7455) respondent types, respectively.

For example, given the category “Good” of the PURPOSE indicator, the probabilities for being in the clusters “Ideal”, “Believers”, and “Skeptic” are 0.7185, 0.2443, and 0.0371, respectively. That is, there is a high probability of landing in the cluster “Ideal”, if “Good” is scored on the PURPOSE indicator. Or, given PURPOSE = “waste”, the probabilities for the three clusters are 0.2359, 0.0185, and 0.7455, respectively, with the highest probability obtained for the cluster “Skeptic”.

These reversed conditional class probabilities $P(X = K_t | I_l = i_l)$ can be visualized using a tri plot, a barycentric coordinate display. The tri plot for the 3-class LC model is shown in Figure 12.



gss82white.sav
Model1 - L^2 = 257,2604
Model2 - L^2 = 79,5079
Model3 - L^2 = 22,0872
Parameters
Profile
ProbMeans
Uni-Plot
Tri-Plot
Bivariate Residuals
EstimatedValues-Model
Model4 - L^2 = 6,6142
Model5

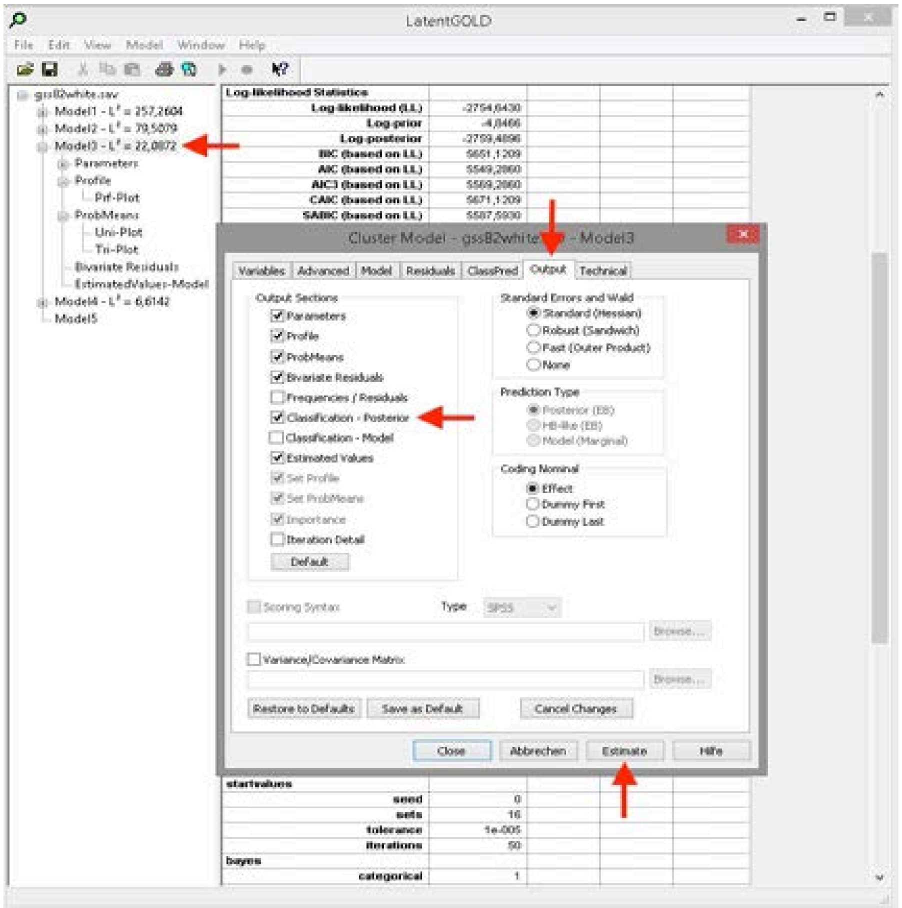
☑ Legend
☑ Point Labels
A Vertex: Cluster1
B Vertex: Cluster2
Variables
☑ purpose
☐ accuracy
☑ understa
☑ cooperat
Groups

purpose:good Probs: Cluster1=0,7185, Cluster2=0,2443, Cluster3=0,0371

Figure 12. Tri plot for the 3-class LC model. The axes represent the three clusters. The base axis is “Cluster1”, the right axis is “Cluster2”, and the left axis is “Cluster3”. Each indicator category $I_l = i_l$ is shown as a point in the triangle, with the corresponding coordinates $C = \left(P(X = K_1 | I_l = i_l), P(X = K_2 | I_l = i_l), P(X = K_3 | I_l = i_l)\right)$. The indicator-score PURPOSE = “Good” is selected and the corresponding coordinates are delineated. The filled blue triangle has the cluster sizes as the coordinates. In the Control Panel “Tri-Plot Control”, the variable ACCURACY has been disabled and is not shown in the tri plot.

Select “Tri-Plot” in the Outline Pane to obtain the tri plot in the Contents Pane. To view “Tri-Plot Control”, the Control Panel for customizing the tri plot, right click on the Contents Pane or tri plot, or select “View > Plot Control” from the menu. For example, we have disabled the variable ACCURACY in the Control Panel, with the effect that this variable is not shown anymore in the tri plot. The edges of the triangle stand for the three clusters. The filled blue triangle delineates the cluster sizes, $p(K_1) = 0.6168$, $p(K_2) = 0.2039$, and $p(K_3) = 0.1793$. We have highlighted the category “Good” of the variable PURPOSE. In the status bar at the bottom of the window, we see that the highest conditional probability $P(X = K_t|\text{PURPOSE} = \text{“Good”})$ is with “Cluster1” (0.7185), followed by “Cluster2” (0.2443), and the smallest probability is obtained for “Cluster3” (0.0371). These numbers can also be seen in barycentric coordinates of the tri plot.

We can use Latent GOLD® to classify cases into the clusters by modal assignment. For this, we have to re-estimate our 3-class LC model. However, this time, we do request the classification output in the software settings, see Figure 13.



Log-likelihood Statistics	
Log-likelihood (LL)	-2754,6430
Log-prior	-4,8466
Log-posterior	-2759,4896
BIC (based on LL)	5651,1209
AIC (based on LL)	5549,2860
AIC3 (based on LL)	5569,2860
CAIC (based on LL)	5671,1209
SABIC (based on LL)	5587,5930

Output Sections
☑ Parameters
☑ Profile
☑ ProbMeans
☑ Bivariate Residuals
□ Frequencies / Residuals
☑ Classification - Posterior
□ Classification - Model
☑ Estimated Values
☑ Set Profile
☑ Set ProbMeans
☑ Importance
□ Iteration Detail

Standard Errors and Wald
● Standard (Hessian)
○ Robust (Sandwich)
○ Fast (Outer Product)
○ None

Prediction Type
● Posterior (EB)
○ HB-like (EB)
○ Model (Marginal)

Coding Nominal
● Effect
○ Dummy First
○ Dummy Last

□ Scoring Syntax Type SPSS

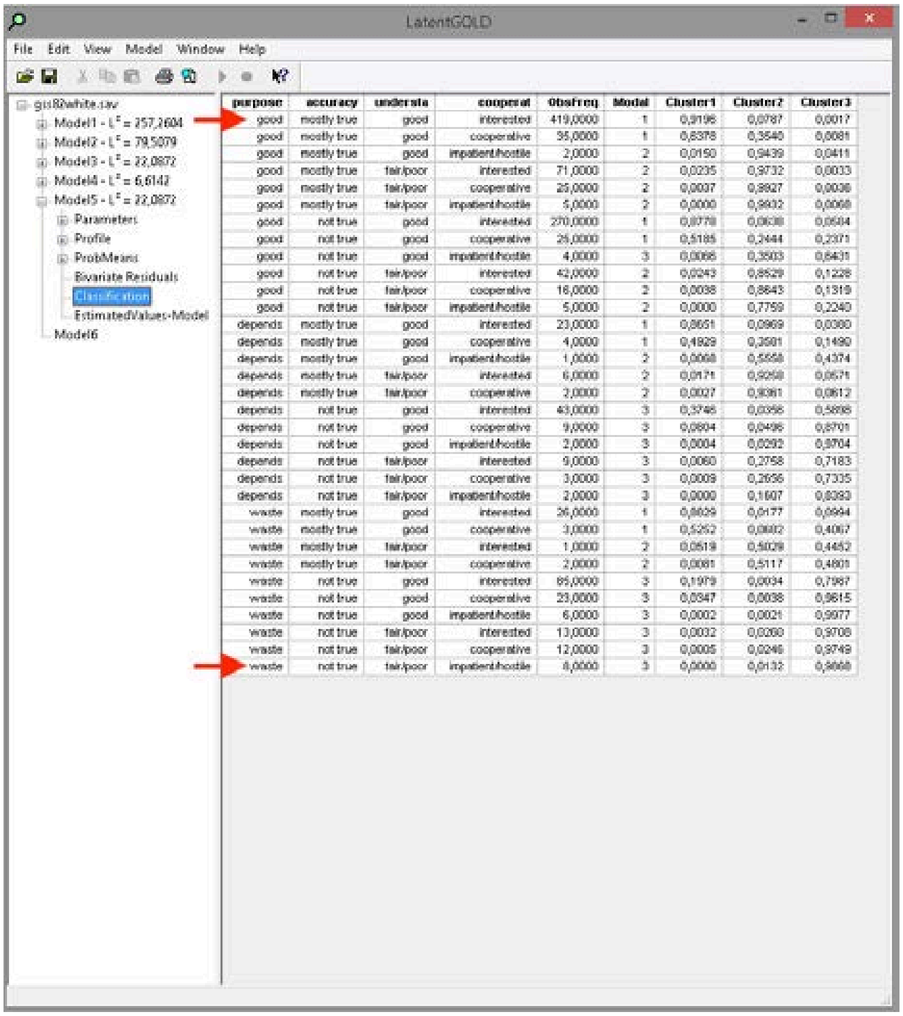
□ Variance/Covariance Matrix

startvalues	
seed	0
sets	16
tolerance	1e-005
iterations	50
bayes	
categorical	1

Figure 13. Output tab of the analysis dialog box. The analysis dialog box appears after double clicking “Model3”. Choose the tab “Output”, activate “Classification - Posterior”, and click on “Estimate”.

Call the analysis dialog box by double clicking “Model3” in the Outline Pane. Then click on the sixth tab “Output”. To request the classification by modal assignment results, mark the entry “Classification - Posterior” of the box “Output Sections” and click on “Estimate”.

The same 3-class LC model is fitted and appears as the new “Model5” in the Outline Pane. Under “Model5”, click on “Classification” on the left-hand side to obtain the classification output on the right-hand side. The Latent GOLD® screen should now look like in Figure 14.



gss82white.sav
Model1 - L² = 257,2604
Model2 - L² = 79,5079
Model3 - L² = 22,0872
Model4 - L² = 6,6142
Model5 - L² = 22,0872
Parameters
Profile
ProbMeans
Bivariate Residuals
Classification
EstimatedValues-Model
Model6

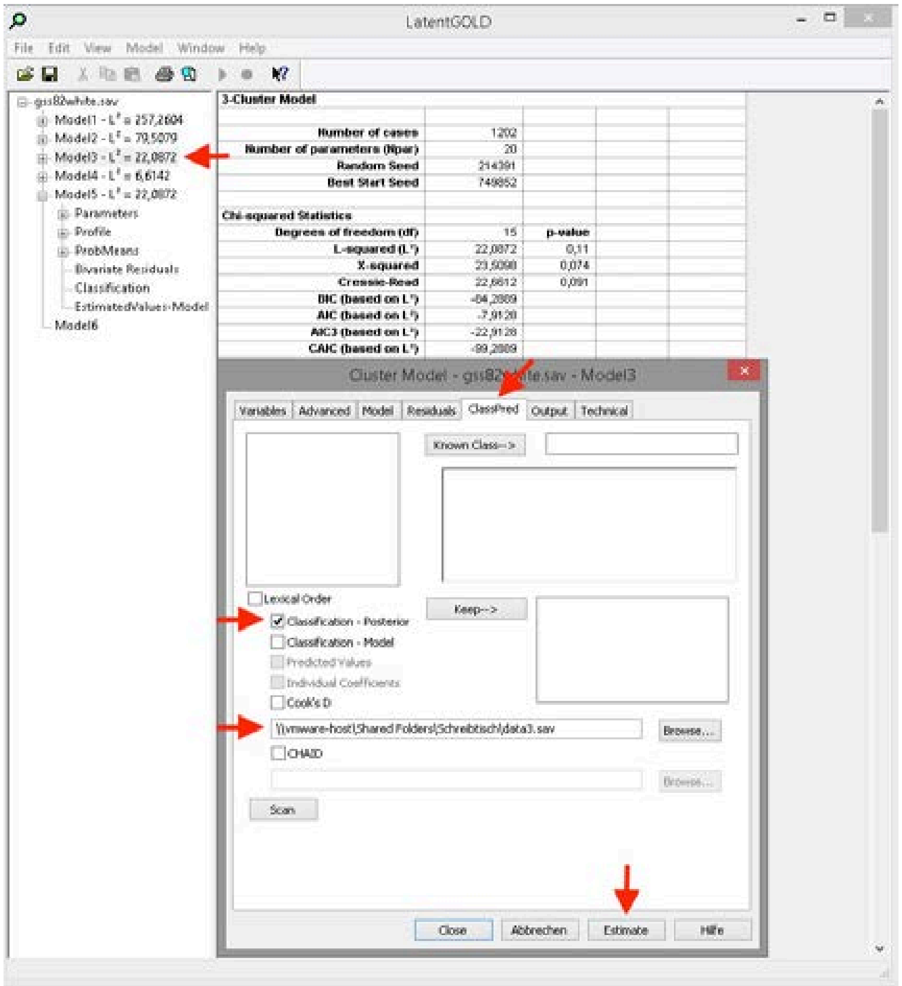
purpose	accuracy	understa	cooperat	ObsFreq	Modal	Cluster1	Cluster2	Cluster3
good	mostly true	good	interested	419,0000	1	0,9196	0,0787	0,0017
good	mostly true	good	cooperative	35,0000	1	0,6378	0,3540	0,0081
good	mostly true	good	impatient/hostile	2,0000	2	0,0150	0,9439	0,0411
good	mostly true	fair/poor	interested	71,0000	2	0,0235	0,9732	0,0033
good	mostly true	fair/poor	cooperative	25,0000	2	0,0037	0,9927	0,0036
good	mostly true	fair/poor	impatient/hostile	5,0000	2	0,0000	0,9932	0,0068
good	not true	good	interested	270,0000	1	0,8778	0,0638	0,0584
good	not true	good	cooperative	25,0000	1	0,5185	0,2444	0,2371
good	not true	good	impatient/hostile	4,0000	3	0,0066	0,3503	0,6431
good	not true	fair/poor	interested	42,0000	2	0,0243	0,8529	0,1228
good	not true	fair/poor	cooperative	16,0000	2	0,0038	0,8643	0,1319
good	not true	fair/poor	impatient/hostile	5,0000	2	0,0000	0,7759	0,2240
depends	mostly true	good	interested	23,0000	1	0,8651	0,0969	0,0380
depends	mostly true	good	cooperative	4,0000	1	0,4929	0,3581	0,1490
depends	mostly true	good	impatient/hostile	1,0000	2	0,0068	0,5558	0,4374
depends	mostly true	fair/poor	interested	6,0000	2	0,0171	0,9258	0,0571
depends	mostly true	fair/poor	cooperative	2,0000	2	0,0027	0,9361	0,0612
depends	not true	good	interested	43,0000	3	0,3746	0,0356	0,5898
depends	not true	good	cooperative	9,0000	3	0,0804	0,0496	0,8701
depends	not true	good	impatient/hostile	2,0000	3	0,0004	0,0292	0,9704
depends	not true	fair/poor	interested	9,0000	3	0,0060	0,2758	0,7183
depends	not true	fair/poor	cooperative	3,0000	3	0,0009	0,2656	0,7335
depends	not true	fair/poor	impatient/hostile	2,0000	3	0,0000	0,1607	0,8393
waste	mostly true	good	interested	26,0000	1	0,8829	0,0177	0,0994
waste	mostly true	good	cooperative	3,0000	1	0,5252	0,0682	0,4067
waste	mostly true	fair/poor	interested	1,0000	2	0,0519	0,5029	0,4452
waste	mostly true	fair/poor	cooperative	2,0000	2	0,0081	0,5117	0,4801
waste	not true	good	interested	85,0000	3	0,1979	0,0034	0,7987
waste	not true	good	cooperative	23,0000	3	0,0347	0,0038	0,9615
waste	not true	good	impatient/hostile	6,0000	3	0,0002	0,0021	0,9977
waste	not true	fair/poor	interested	13,0000	3	0,0032	0,0260	0,9708
waste	not true	fair/poor	cooperative	12,0000	3	0,0005	0,0246	0,9749
waste	not true	fair/poor	impatient/hostile	8,0000	3	0,0000	0,0132	0,9868

Figure 14. Classification by modal assignment output for the 3-class LC model. The cases in the first and last rows of the output, with completely positive and negative categories, are put into the first (“Modal” = 1) and third (“Modal” = 3) clusters, respectively. For, “Cluster1” and “Cluster3” are most likely, with highest probabilities 0.9196 and 0.9868, respectively.

Two examples are emphasized, the first and last rows of the output. The first row shows that the 419 interviewees with the response pattern (PURPOSE = “Good”, ACCURACY = “Mostly True”, UNDERSTANDING = “Good”, COOPERATION = “Interested”) are classified into the first cluster “Cluster1” / “Ideal”, because the posterior probability for this cluster is the largest among the three

(0.9196). Under “Modal”, there is the value 1 to denote this result. In the last row, we have 8 interviewees who have the response pattern (PURPOSE = “Waste”, ACCURACY = “Not True”, UNDERSTANDING = “Fair, Poor”, COOPERATION = “Impatient, Hostile”) and are assigned to the third cluster “Cluster3” / “Skeptic”. This cluster has the highest probability (0.9868). Thus, we see the value 3 under “Modal”.

The classification output can be appended to the data file, see Figure 15.



```
gss82white.sav
  Model1 - L² = 257,2604
  Model2 - L² = 79,5079
  Model3 - L² = 22,0872
  Model4 - L² = 6,6142
  Model5 - L² = 22,0872
    Parameters
    Profile
    ProbMeans
    Bivariate Residuals
    Classification
    EstimatedValues-Model
  Model6
```

3-Cluster Model		
Number of cases	1202	
Number of parameters (Npar)	20	
Random Seed	214391	
Best Start Seed	749852	
Chi-squared Statistics		
Degrees of freedom (df)	15	p-value
L-squared (L²)	22,0872	0,11
X-squared	23,5098	0,074
Cressie-Read	22,6612	0,091
BIC (based on L²)	-84,2889	
AIC (based on L²)	-7,9128	
AIC3 (based on L²)	-22,9128	
CAIC (based on L²)	-99,2889	

□ Lexical Order
☑ Classification - Posterior
□ Classification - Model
□ Predicted Values
□ Individual Coefficients
□ Cook's D
\\vmware-host\Shared Folders\Schreibtisch\data3.sav
□ CHAID

Figure 15. Append the classification output to the data file by choosing “Classification - Posterior” on the “ClassPred” tab of the analysis dialog box, which is retrieved by double clicking “Model3” in the Outline Pane. You can specify the name for the file and where to store it. The file is generated when the “Estimate” button is pressed.

Double click on “Model3” to open up the analysis dialog box. Select “Classification - Posterior” on the fifth tab “ClassPred”, specify under what name and where to store your file, and click on “Estimate”. The new file is saved and can be found at the specified location.

Via the menu “File > Save Results...” we can save the output or results for a highlighted model. The settings of a selected model can also be saved and then opened for later use. For this, select the model that you want to save in the Outline Pane, and go to “File > Save Definition...”. The save dialog box opens. Name the file ("*.lgf", for Latent GOLD® file) and specify its location, and click on "Save". Now you can re-open your saved "*.lgf" definition file, without having to re-open the data file separately. Simply go to "File > Open...". The open dialog box appears. Select the desired file and click on "Open".

6. Conclusion: In this paper, we have discussed the technique of exploratory latent class cluster analysis. This technique pursues two major goals, clustering followed by classification, both in model-based variants.

As the end point of the paper, we recap the basic idea underlying exploratory latent class cluster analysis, and we give a schematic overview of the complete procedure.

The basic idea can be summarized as follows. The conditional probabilities of a postulated LC model differ across unobserved subgroups (latent classes, clusters, or types). These subgroups form the categories of a categorical latent variable. Having estimated these parameters, assessed the fit of the LC model, and selected the proper number of latent classes, those parameters can be compared between the subgroups to show how the clusters differ from each other. This enables characterizing the unobserved subgroups, and hence, the nature of the categorical latent variable (model-based clustering). Eventually, cases of the sample can be classified into the latent types based upon posterior membership probabilities estimated directly from the LC model (model-based classification).

To sum up the entire presentation, an overall flow diagram of the analysis process of exploratory latent class cluster analysis is depicted in Figure 16.

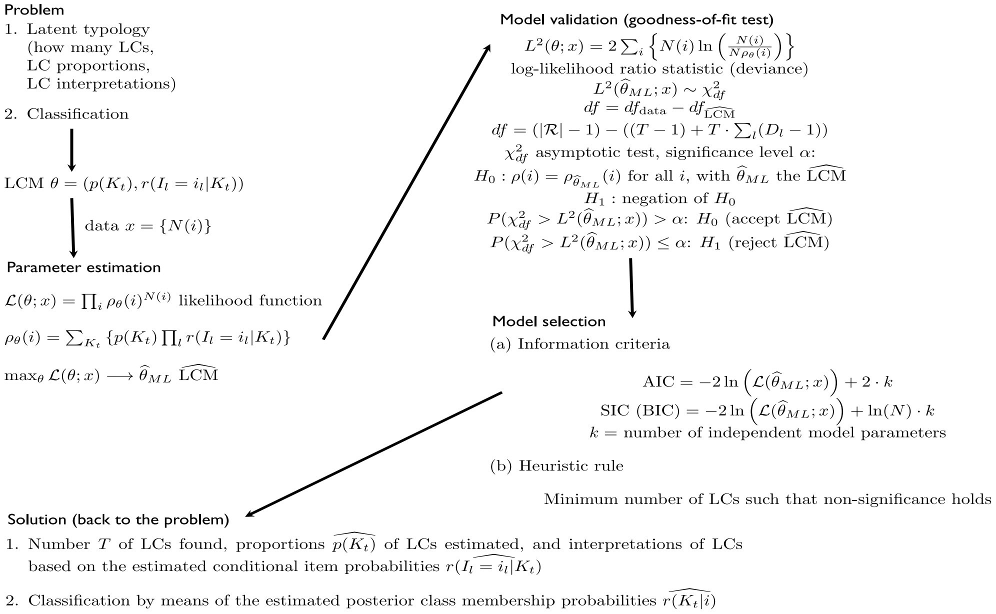

Problem
1. Latent typology (how many LCs, LC proportions, LC interpretations)
2. Classification
LCM $\theta = (p(K_t), r(I_l = i_l|K_t))$
data $x = \{N(i)\}$
Parameter estimation
$\mathcal{L}(\theta; x) = \prod_i \rho_\theta(i)^{N(i)}$ likelihood function
$\rho_\theta(i) = \sum_{K_t} \{p(K_t) \prod_l r(I_l = i_l|K_t)\}$
$\max_\theta \mathcal{L}(\theta; x) \longrightarrow \widehat{\theta}_{ML}$ $\widehat{\text{LCM}}$
Model validation (goodness-of-fit test)
$L^2(\theta; x) = 2 \sum_i \left\{N(i) \ln\left(\frac{N(i)}{N\rho_\theta(i)}\right)\right\}$
log-likelihood ratio statistic (deviance)
$L^2(\widehat{\theta}_{ML}; x) \sim \chi^2_{df}$
$df = df_{\text{data}} - df_{\widehat{\text{LCM}}}$
$df = (|\mathcal{R}| - 1) - ((T-1) + T \cdot \sum_l (D_l - 1))$
χ^2_{df} asymptotic test, significance level α:
$H_0 : \rho(i) = \rho_{\widehat{\theta}_{ML}}(i)$ for all i, with $\widehat{\theta}_{ML}$ the $\widehat{\text{LCM}}$
H_1 : negation of H_0
$P(\chi^2_{df} > L^2(\widehat{\theta}_{ML}; x)) > \alpha$: H_0 (accept $\widehat{\text{LCM}}$)
$P(\chi^2_{df} > L^2(\widehat{\theta}_{ML}; x)) \leq \alpha$: H_1 (reject $\widehat{\text{LCM}}$)
Model selection
(a) Information criteria
$\text{AIC} = -2 \ln\left(\mathcal{L}(\widehat{\theta}_{ML}; x)\right) + 2 \cdot k$
$\text{SIC (BIC)} = -2 \ln\left(\mathcal{L}(\widehat{\theta}_{ML}; x)\right) + \ln(N) \cdot k$
k = number of independent model parameters
(b) Heuristic rule
Minimum number of LCs such that non-significance holds
Solution (back to the problem)
1. Number T of LCs found, proportions $\widehat{p(K_t)}$ of LCs estimated, and interpretations of LCs based on the estimated conditional item probabilities $r(I_l \widehat{= i_l|} K_t)$
2. Classification by means of the estimated posterior class membership probabilities $\widehat{r(K_t|i)}$

Figure 16. Flow chart representation of the analysis process of exploratory latent class cluster analysis. Abbreviations: LC(s) = latent class(es), LCM = latent class model, ML = maximum likelihood, df = degrees of freedom, AIC = Akaike information criterion, and SIC (BIC) = Schwarz (Bayesian) information criterion.

References:

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Academiai Kiado.

2. Andersen, E. B. (1982). Latent structure analysis: A survey. *Scandinavian Journal of Statistics, 9*, 1–12.

3. Birch, M. W. (1964). A new proof of the Pearson-Fisher theorem. *Annals of Mathematical Statistics, 35*, 818–824. doi:10.1214/aoms/1177703581.

4. Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis.* New York: Springer. doi:10.1007/978-0-387-72806-3.

5. Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* New York: Springer. doi:10.1007/b97636.

6. Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York: Plenum Press. doi:10.1007/978-1-4899-1292-3_6.

7. Cramér, H. (1999). *Mathematical methods of statistics.* Princeton, NJ: Princeton University Press.

8. Dayton, C. M. (1998). *Latent class scaling analysis*. Thousand Oaks: Sage Publications. doi:10.4135/9781412984720.

9. Enøe, C., Georgiadis, M. P., & Johnson, W. O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine, 45*, 61–81. doi:10.1016/S0167-5877(00)00117-3.

10. Formann, A. K. (1984). *Die Latent-Class-Analyse* [*Latent class analysis*]. Weinheim: Beltz.

11. Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology, 79*, 1179–1259. doi:10.1086/225676.

12. Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231. doi:10.1093/biomet/61.2.215.

13. Goodman, L. A. (1978). *Analysing qualitative/categorial variables: Loglinear models and latent structure analysis.* Cambridge: Cambridge University Press.

14. Gut, A. (2013). *Probability: A graduate course.* New York: Springer. doi:10.1007/978-1-4614-4708-5.

15. Hagenaars, J. A., & McCutcheon, A. L. (Eds.) (2002). *Applied latent class analysis.* Cambridge: Cambridge University Press. doi:10.1017/CBO9780511499531.

16. Hui, S. L., & Zhou, X. H. (1998). Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research, 7*, 354–370. doi:10.1177/096228029800700404.

17. Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis.* Hoboken, NJ: Wiley. doi:10.1002/9780470316801.

18. Langeheine, R., & Rost, J. (Eds.) (1988). *Latent trait and latent class models.* Boston: Springer. doi:10.1007/978-1-4757-5644-9.

19. Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

20. Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses.* New York: Springer. doi:10.1007/0-387-27605-X.

21. Lin, T. H. (2006). A comparison of model selection indices for nested latent class models. *Monte Carlo Methods and Applications, 12*, 239–259. doi:10.1515/156939606778705164.

22. Lin, T. H. (2015). Model selection information criteria in latent class models with missing data and contingency question. *Communications in Statistics - Simulation and Computation, 44*, 319–331. doi:10.1080/03610918.2013.777454.

23. Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*, 249–264. doi:10.2307/1165284.

24. McCutcheon, A. L. (1987). *Latent class analysis.* Newbury Park: Sage Publications. doi:10.4135/9781412984713.

25. McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions.* New York: Wiley. doi:10.1002/9780470191613.

26. Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer. doi:10.1007/978-1-4612-4578-0.

27. Rost, J., & Langeheine, R. (Eds.) (1997). *Applications of latent trait and latent class models in the social sciences.* New York: Waxmann.

28. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. doi:10.1214/aos/1176344136.

29. Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511499531.004.

30. Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The Sage encyclopedia of social science research methods* (pp. 549–553). Thousand Oaks: Sage Publications. doi:10.4135/9781412950589.n472.

31. Vermunt, J. K., & Magidson, J. (2016). *Technical guide for Latent GOLD 5.1: Basic, advanced, and syntax.* Belmont, MA: Statistical Innovations Inc. https://www.statisticalinnovations.com/user-guides/

32. Walter, S. D., & Irwig, M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *Journal of Clinical Epidemiology, 41*, 923–937. doi:10.1016/0895-4356(88)90110-2.