Computational and Mathematical Methods in Medicine
Vol. 10, No. 3, September 2009, 203–218

Linear latent structure analysis and modelling of multiple categorical variables

I. Akushevich*, M. Kovtun, K.G. Manton and A.I. Yashin

Duke University, Durham, NC, USA

(*Received 18 July 2007; final version received 29 May 2008*)

Linear latent structure analysis is a new approach for investigation of population heterogeneity using high-dimensional categorical data. In this approach, the population is represented by a distribution of latent vectors, which play the role of heterogeneity variables, and individual characteristics are represented by the expectation of this vector conditional on individual response patterns. Results of the computer experiments demonstrating a good quality of reconstruction of model parameters are described. The heterogeneity distribution estimated from 1999 National Long Term Care Survey (NLTCS) is discussed. A predictive power of the heterogeneity scores on mortality is analysed using vital statistics data linked to NLTCS.

Keywords: latent structure analysis; grade of membership; population heterogeneity; mortality; NLTCS data; disability; simulation studies

1. Introduction

Categorical data occur often in behaviour sciences, psychology, epidemiology, genetics, education, and marketing. A common feature of modern categorical data collections (e.g., mental test data, demographic surveys or gene expression data) is their high dimensionality with respect to both the sample size and the numbers of categorical variables measured for each individual.

Limited numbers of categorical variables may not precisely characterize an object due to the stochastic nature of observed objects, their heterogeneity or measurement errors. Often the phenomenon of interest cannot be observed directly, so multiple measures have to be made, each directly or indirectly related to the latent variables of actual interest. Researchers analysing such data have to: (i) extract the signal from data with a large stochastic component, (ii) extract a latent (unobservable) structure, related to the categorical data and (iii) parameterize the latent structure to interpret it substantively. The identification and description of a latent structure are not sufficient when estimations of individual parameters are additionally required. Such problems are informally solved by a physician making a diagnosis: he/she estimates an unobservable variable of interest (presence of a disease) based on categorical measures (description of symptoms and answers to physician's questions) and estimated population characteristics (physician's professional experience). Likewise, having a set of categorical measurements (e.g., a demographic survey) made on individuals, a researcher would like to discover (i) properties of the population and (ii) properties of the individuals. Methods which attempt to achieve these simultaneously are known as latent structure analysis.

*Corresponding author. Email: igor.akushevich@duke.edu

ISSN 1748-670X print/ISSN 1748-6718 online

DOI: 10.1080/17486700802259798
http://www.informaworld.com

One of the best known such methods is the latent class model (LCM), which can be characterized as a statistical method for finding subtypes of related cases (latent classes) from multivariate categorical data [7,11,19]. In LCM, individuals are assigned to one of several homogeneous classes. This requires estimation of the individual latent variable (class number). Other models of this family such as item response theory and Rasch models, discrete latent class models [14], latent distribution analysis [27,30,34], differ by the assumptions made about the latent variable(s). One method for identifying the latent structure in large categorical data sets with a simultaneous evaluation of individual scores in a state space is Grade of Membership (GoM) analysis. GoM was introduced by Woodbury and Clive [37]; see Manton *et al.* [23] for a detailed exposition and additional references. Statistical properties of GoM models were analysed by Tolley and Manton [33], Singer [31] and Wachter [36]. An example of application of the GoM approach is done in [25].

All these models belong to the family of latent structure analysis that considers a number of categorical variables measured on each individual in a sample with the intent of discovering the properties of both a population and individuals composing the population. Different approaches to this general problem, and recent developments, are described in reports collected in Hagenaars and McCutcheon [13]. The relation of specific latent structure models is discussed by Uebersax [35] and Erosheva [8]. Broad spectrum of the methods of dealing with categorical data analyses and modelling as well as place and role of the methods of LSA are described by Agresti [1] and Powers and Xie [29].

Methods of latent structure analysis can be reformulated in terms of statistical mixtures and mixing distributions [9,20,26,32]. Bartholomew [5] noted the theory of mixing distributions as a common mathematical framework of latent structure models with abilities 'to clarify, simplify and unify the disparate developments spanning over a century.' The main assumption of the latent structure analysis is the local independence assumption. Being formulated in terms of the theory of mixing distributions, it states that the observed joint distribution (of categorical random variables) can always be described as a mixture of independent distributions. The mixing distribution is considered as a distribution of the latent variable(s) that contains latent information regarding the phenomenon under study. Specific models of latent structure analysis vary by assumptions regarding properties of the mixing distribution.

LCM, GoM and many other methods of latent structure analysis use maximum likelihood for parameter estimation. Although a part of parameters describing a latent structure are regular (structural) parameters, parameters corresponding to individual states are so-called 'nuisance' parameters, the number of which increases with sample size. There are a series of non-trivial mathematical tasks, which have to be solved to estimate such individual parameters. First is the problem of identifiability, which is especially difficult for a large number of estimated parameters with complex interrelationship. Second is the problem of consistency. In the presence of nuisance parameters, consistency of estimators no longer follows from general theorems, consequently, some natural estimators need not to be consistent [28]. Thus, a proof of consistency should be provided for every estimator. Third is the availability and quality of algorithms to perform estimations. Since no general theorems are available to address these questions, all of these points have to be considered separately for each task involved in dealing with nuisance parameters.

Recently linear latent structure (LLS) analysis has been proposed to model high dimensional categorical data [16,18]. The LLS model has been formulated using mixing distribution theory. Similar to other latent structure analyses, the goal of the LLS analysis is to derive both the properties of a population and individuals using discrete measurements. The LLS, however, does not use maximization of likelihood for parameter estimation.

Instead, it uses a new estimator, where the LLS parameter estimates are solutions of a quasilinear system of equations. For this estimator, it is possible to prove consistency, to formulate conditions for identifiability, and to formulate a high-performance algorithm allowing one to handle datasets involving thousands of categorical variables [18].

Mathematically rigorous description of LLS methods as well as proofs of its basic statistical properties was provided by Kovtun *et al.* [16,18]. These papers are written at a very technical level that can complicate their use by applied researchers. The aim of this paper is to present a description of the LLS approach from a practical point of view and illustrate it by the results of simulation studies and by the analysis of a real data set. Specifically, the LLS problem formulation, its basic properties and geometrical interpretation are discussed in Section 2. Such introduction cannot completely avoid a formalism of the method, so the introduction of necessary notion is given in the Appendix. Section 3 contains the description of the results of the simulation experiments designed to check the quality of the reconstruction of model components. Section 4 discusses how the model can be applied to a specific dataset. All steps of the LLS analysis from determination of the dimensionality of the model to application of identified components of the latent structure and to prediction of longevity are discussed. The last section summarizes the results and includes a discussion of findings and perspectives.

2. Linear latent structure analysis

The typical dataset analysed by methods of latent structure analysis can be represented by the $I \times J$ matrix constituted by categorical outcomes of J measurements on I individuals. Each row in the matrix corresponds to an individual and contains an individual response pattern, i.e., a sequence of J numbers with the j^{th} number running from 1 to the number of responses L_j for that variable. In most cases L_j ranges from 2 to 5–10, and rarely exceeds several dozen.

Usually surveys are designed to study one or several major subjects of interest. The real number of investigated components is always much smaller than the number of questions. For instance, in demographic surveys directed toward disability analysis, several dozen ADL and IADL questions (27 in screener Interview Dataset of the NLTCS [22]) describe disability status. Another example is the SAT exam where hundreds of questions are asked to extract two variables (verbal and math SAT scores) describing actual knowledge and skills. Similar situations appear when the subject of investigation is determined by the processes which are only regulated by several components. For example, if a dataset is represented by DNA microarray data (i.e., measurements of expressions of thousands of genes), then such components could be expressions of regulating genes or several metagenes obtained by clustering gene expression patterns. More detailed descriptions resulting in an increase of study components is possible, but only with an increase of the precision of measurements.

Therefore, important questions to be addressed in the analysis include (i) how to extract information from the dataset and how to relate the extracted information to auxiliary phenomenological findings, (ii) how to identify quantities directly related to health outcomes or other characteristics studied in the survey or linked datasets, (iii) how to quantitatively compare individuals and (iv) how to find characteristics possessing predictive power. These can be investigated by identifying and evaluating the hidden (or latent) structure responsible for generating the data.

There exist several factors allowing us to assume that the dimension of a hidden structure is much smaller than, say, I, J or $|L| = \sum_J L_J$. The most important factor is the mutual correlations in random variables corresponding to answers to different questions. Usually such correlations are results of experimental design, where different variables are designed to be related in certain degree to studied latent variables. For example, random variables corresponding to responses to SAT questions are obviously highly correlated, and this property is essential for the test design. Other factors are related to stochastic uncertainties in these data. The existence of such uncertainties is clear due to the fact that the true individual probability of a specific response to a certain question is not exactly one nor zero. These stochastic uncertainties could prevent us from identifying certain components of the latent structure and as a result, an approximate description with decreased dimensionality has to be used. Naturally appearing research questions are: (i) what dimensionality is needed in order to give the best representation of the structure in the data, (ii) whether it is possible to find geometrical interpretations of the hidden structure and (iii) how this structure can be parameterized.

Geometrically, the sought structure can be viewed in the linear space $R^{|L|}$. Each dimension in the space corresponds to one possible outcome of a certain question. Therefore L_j components correspond to the j^{th} question and $|L|$ is the total dimension of the space. An individual response corresponds to a point in this space and is described using a vector, the components of which run over all possible pairs of jl, i.e., corresponding to the first outcome of the first question, to the second outcome of the first question and so on. A component jl of the vector of individual response equals one if its l^{th} variant was realized for j^{th} question or zero otherwise. The study population is represented by a set of I points in the linear space. If a hidden low-dimensional structure generating these data exists, then such a structure has to be reflected in the set of the points. There are three basic shapes of the structure: a set of several isolated points (or clusters), a set of points that reproduces a certain nonlinear curve or surface and a low dimensional linear subspace. These cases are investigated by methods of latent class, latent trait models and the LLS analysis, respectively. Recently, a comparison of latent structures of the GoM, Rasch (or latent trait) and latent class models was presented by Erosheva [8].

From the geometrical point of view, the LLS can be roughly understood as an approach to search a K-dimensional subspace in the linear space $R^{|L|}$, which is the closest to the set of I points, representing individual outcomes. The dimension of the space K is defined by the methods of the LLS analysis. This linear subspace is defined by its basis $\lambda^1, \ldots, \lambda^K$, where λ^k is a L-dimensional vector with components λ^k_{jl}, so to find the K-dimensional subspace means finding a basis, $\lambda^k_{jl}, (k = 1, \ldots, K)$.

In the LLS analysis the linear subspace is interpreted as the space of individual probabilities β^i_{jl}, which are expressed as

$$\beta^i_{jl} = \sum_{k=1}^{K} g_{ik}\lambda^k_{jl}. \tag{1}$$

The basis vectors of the subspace are also interpreted as probabilities and can define the so-called pure types. In this sense, the model decomposition (1) has the interpretation of decomposition over pure types or over ‘ideal persons,’ whose individual probabilities are basis vectors of the subspace. Vectors $g_i = (g_{i1}, \ldots, g_{iK})$, being coefficients of decomposition of individual probabilities over the basis vectors, represent the hidden states of individuals in which we are interested. Roughly, they can be understood as projections of the vector of individual outcomes to the linear subspace. More precisely, they are estimated as the conditional expectations of a random vector describing the latent structure conditional on an individual response pattern. To guarantee that β^i_{jl} and λ^k_{jl} have the sense of probabilities, the following restrictions have to be satisfied

$$\sum_{l=1}^{L_J} \lambda^k_{jl} = 1, \ \ \lambda^k_{jl} \geq 0, \ \sum_{k=1}^{K} g_{ik} = 1 \ \text{ and } \sum_{k} g_{ik}\lambda^k_{jl} \geq 0. \tag{2}$$

These restrictions reflect that domain of the hidden states of individuals is not a whole subspace but a convex polyhedron.

Strictly speaking, the LLS analysis is formulated in the framework of the theory of statistical mixtures and mixing distributions. Using this language, the LLS model approximates the observed distribution of J categorical measurements made for I individuals by a mixture of independent distributions (i.e., those distributions where all J variables are assumed to be independent). A mixing distribution is assumed to be supported by a K-dimensional subspace of the space of independent distributions [18]. A rigorous formulation of the task and discussion of specific properties of the model and the estimator are presented in the Appendix. Parameters to be estimated in the LLS analysis include structural constants λ_{jl}^{k} defining a basis in the supporting subspace and the mixing distribution (which can be thought of as the vector of random variable G). An important feature of the LLS analysis is that the mixing distribution is estimated non-parametrically. Properties of such kind of tasks, i.e., when an estimated 'parameter' is a whole distribution, were discussed by Kiefer and Wolfowitz [15]. Although several practical steps including analysis of underlying geometry and construction of maximum likelihood estimator were performed by Lindsay [20] and Böhning [6], the task of nonparametric estimation of continuous mixing distribution is difficult especially for large J. In the case of LLS the assumption that the support of the mixing distribution is in the linear subspace of small dimension allowed us to prove several useful properties of the model (e.g., existence of a system of equations connecting the estimated parameters), and then, using these properties to construct an estimator provided consistent and identifiable estimations of the model parameters. The basic steps of this estimation procedure include (i) estimation of the dimensionality K of the supporting subspace, (ii) estimation of structural constants λ_{jl}^{k}, which define the K-dimensional linear subspace within the supporting polyhedron, (iii) choice of the basis in the found linear subspace and (iv) estimation of mixing distribution as an empirical distribution represented by conditional LLS scores in the chosen basis (see Appendix).

The basis cannot be defined uniquely, and any linear combination keeping the LLS restrictions can be considered an alternative. The most important results of the LLS analysis, such as estimation of individual probabilities β_{jl}^{i} or relative individual position in the subspace, are independent of the basis choice, so the procedure of choice of the basis (or 'ideal persons') is arbitrary in many respects. This can be exploited to choose a basis in which the investigated phenomenon is better expressed. Here, we describe two possible schemes which are implemented in the LLS algorithm. In the first one, the researcher specifies characteristics of 'ideal' individuals based on his/her experience in the research domain. Then, he/she can empirically construct vectors of probabilities for such ideal individuals or identify these individuals from the sample. The vectors of probabilities of these individuals are taken as basis vectors. If the probability vectors are constructed empirically, they could be beyond the polyhedron, defined by (2) so they should be projected to the polyhedron. The individual coordinates calculated in this basis represent the 'proximity' of the individual to the 'ideal' ones. In another scheme, the basis is obtained using an assignment of the LLS scores (calculated on some arbitrary basis) to K clusters, and then basis vectors $\lambda^1, \dots, \lambda^K$ are calculated as means of probabilities β_{jl}^{i} over each cluster. Since the polyhedron is convex, all calculated basis vectors $\lambda^1, \dots, \lambda^K$ belong to it.

Summarizing, the goal of the LLS analysis is to define the following model parameters:

(1) A basis $\lambda^1, \dots, \lambda^K$ of the space that supports the distribution of the LLS scores.
(2) Conditional expectations (or the LLS scores) $g_{ik} = \mathsf{E}(G_k|\ell^i)$, where ℓ is J-dimensional vector of individual responses.

Conditional expectations provide knowledge about individuals. These conditional expectations can be considered coordinates in the subspace, to which all individuals belong. The population is represented by empirical distribution in points $\mathsf{E}(G|\ell)$ with weights equal to frequencies of the outcome pattern ℓ. The ability to discover such a subspace and determine individual positions in it is a valuable feature of the LLS analysis.

3. Simulation studies

Before analysing data we perform three simulation experiments which demonstrate the quality of the reconstruction of basic components of the LLS model, i.e., (i) the linear subspace, (ii) the LLS score distribution and (iii) clusters in the LLS score space.

Linear subspace. Choosing initial two-, three- and five-dimensional subspaces with 60, 120 and 240 dichotomous questions and using uniform distribution of the LLS scores, we simulated 1430 and 14,300 sets of individual responses. Then, we reconstructed a linear subspace and compared it with the initial one. The distance between linear subspaces d [10] is used as a measure of the quality of reconstruction. The distance may be interpreted as the sine of the generalized angle between subspaces. Results for d are in Table 1.

LLS score distribution. For this experiment, we simulated a sample of 10,000 individuals with 1500 dichotomous measurements, i.e., $J = 1500$, $L_j = 2$. For this simulation the position of the supporting plane was chosen randomly. The LLS scores were distributed over a union of the intervals [0.10, 0.25] and [0.50, 0.75]. The complete scheme of overall reconstruction was applied to the simulated sample. This scheme includes reconstruction of the supporting linear subspace, reconstruction of a basis on the subspace and reconstruction of the LLS score distribution in the basis. The results demonstrated an excellent quality of overall reconstruction are given in Figure 1.

Clusters in LLS score space. We also checked the ability of the LLS analysis to identify clusters of the LLS scores. We assumed the LLS score distribution was concentrated at five points with equal weights. Then, setting $K = 3$ and $I = 1000$, we simulated individual responses for $J = 100$, 200 and 500. Then, we applied a complete-linkage hierarchical clustering procedure both to the reconstructed LLS scores and to the original vector of response to combine individuals in five classes. As a measure of quality we used the fraction of individuals assigned to the correct cluster. We performed 10 simulation runs to make the calculation statistically stable. The average number of correctly classified individuals is given in Table 2.

This example demonstrates that the individual conditional expectations are much better suited for the purpose of classification than the original responses.

4. Application to NLTCS data

The National Long Term Care Survey is a longitudinal survey designed to study the changes in health and functional status of older Americans (aged 65+). It also tracks health expenditures, Medicare service use, and the availability of personal, family and community

Table 1. Distances between simulated and reconstructed subspaces.

	$N = 1{,}430$			$N = 14{,}300$		
K	$J = 60$	$J = 120$	$J = 240$	$J = 60$	$J = 120$	$J = 240$
2	0.023	0.023	0.022	0.008	0.007	0.007
3	0.075	0.072	0.070	0.023	0.023	0.023
5	0.222	0.190	0.176	0.073	0.059	0.057

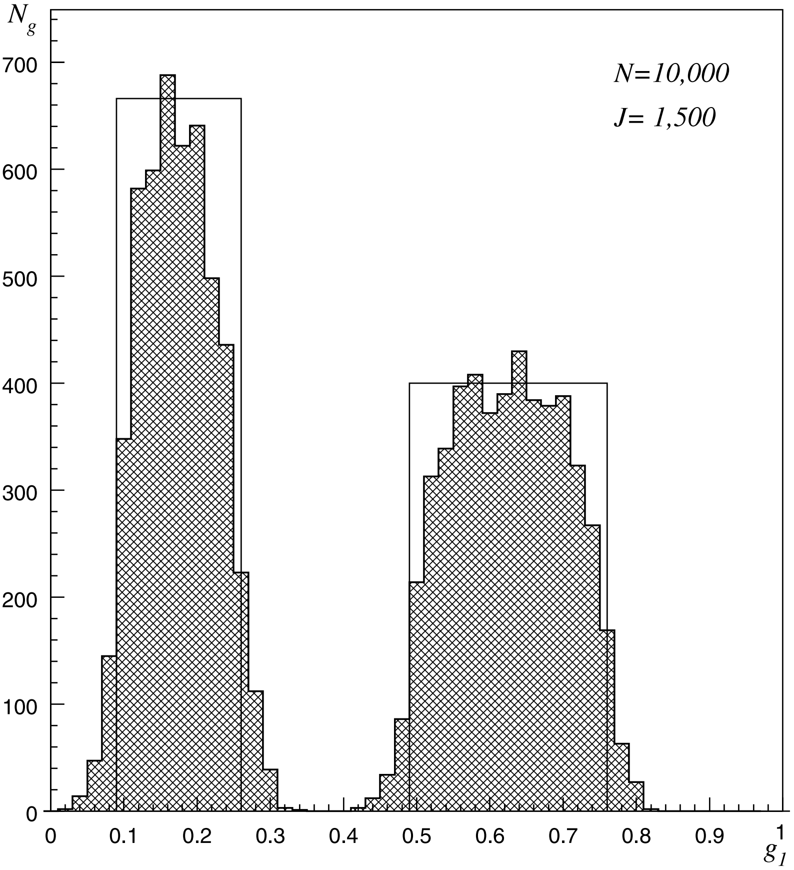

Figure 1. Reconstructed distribution of the LLS scores. Lines show the simulated the LLS score distribution.

resources for care giving. The survey began in 1982 with follow-up surveys conducted in 1984, 1989, 1994 and 1999. A sixth survey 2004 was recently finished in 2005. A detailed description of NLTCS may be found on the web at http://www.nltcs.aas.duke.edu/index.htm.

We considered a sample of 5161 individuals from the 1999 NLTCS wave, and selected 57 health related questions. 27 variables characterize a disability level with respect to activities of daily living, instrumental activities of daily living, and physical impairment. Thirty variables correspond to self-reports of chronic diseases. Details about these questions may be found in Manton *et al.* [22,24]. Then, we excluded 370 individuals with at least one missing outcome. Thus, the 4791 remaining individuals responded to 57 questions, 49 of which have two possible answers, and 8 have 4 possible answers. In total, we have $|L| = 130$.

The first task is to determine the dimensionality of the LLS problem, K, which is the rank of the moment matrix, and thus it might be estimated as the rank of the frequency matrix.

Table 2. The average number of correctly classified individuals for different J.

J	100 (%)	200 (%)	500 (%)
Original responses	77.3	82.7	85.1
LLS scores	95.2	99.8	100.0

Table 3. First 10 singular values of frequency matrix of NLTCS.

σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9	σ_{10}
39.112	3.217	1.464	0.652	0.363	0.310	0.243	0.220	0.198	0.148

We used singular value decomposition (SVD) to estimate the rank of the frequency matrix. The obtained singular values are compared with the total statistical error σ_{E} found as a quadratic sum of cell errors and equal to 0.292. The first 10 singular values are given in Table 3. Six, four and three singular numbers exceed σ_{E}, $2\sigma_{\mathrm{E}}$ and $3\sigma_{\mathrm{E}}$, respectively, so hypotheses $K = 6$, 4 and 3 can be accepted with corresponding confidence levels. For comparison and illustration we will consider the cases corresponding to the hypotheses that $K = 3$ and 4.

After the dimensionality of the LLS-problem is fixed, we apply the LLS algorithm to the estimation of the LLS parameters: basis vectors λ_{jl}^k and individual LLS scores. The basis cannot be defined uniquely, and any convex combination keeping the LLS restrictions (2) can be considered an alternative. In this analysis, we apply the following procedure to find a basis consistent with aims of the analysis. In the beginning, we estimated the LLS scores in some prior basis. Specific choice of the prior basis is not crucial, because subsequent conclusions on a final basis are made using procedures (e.g., clustering and analysis of probability vectors), which are independent of the prior basis choice. Then the LLS scores calculated in a prior basis are assigned to 7–8 clusters (i.e., for slightly larger numbers of clusters than K) using methods of cluster analysis. By analysing outcomes of typical representative respondents (with the LLS scores closest to the cluster centroid) and of probabilities β_{jl}^i calculated as means over each cluster, we can identify the cluster structure of the analysed population sample. On the basis of this analysis we choose $K = 3$ clusters corresponding to (i) healthy individuals ($k = 1$), (ii) strongly disabled individuals ($k = 2$) and (iii) individuals with chronic diseases but without evidence of disability ($k = 3$). For $K = 4$ case, an additional cluster with partly disabled individuals ($k = 4$) is added. For each such group, we created a 'typical' vector of probabilities $\bar{\beta}_{jl}^i$ by hand. Specifically, for the first group ($k = 1$), we chose unit probabilities for all answers corresponding to healthy states. For the second group ($k = 2$), unit probabilities are assigned to answers corresponding to strongly disabled states and means over samples for chronic disease questions. For the third group ($k = 3$), unit probabilities corresponding to non-disabled states and to all disease diagnoses are chosen. The final basis is constructed by a projection of all four vectors to the polyhedron in the K-dimensional subspace.

A similar analysis was performed for $K = 4$. In this case, a fourth 'typical' vector ($k = 4$) is constructed from mean frequencies over the sample. Figure 2 demonstrates the results. The plots on the left show 2D-polyhedron for $K = 3$ and two projections of 3D-polyhedron for $K = 4$. These polyhedrons are defined by the LLS restrictions (2). In this case, the LLS scores are restricted by 130 inequalities $\left(\sum_k g_{ik}\lambda_{jl}^k \geq 0\right)$ and one equality $\left(\sum_k g_{ik} = 1\right)$. The 2D-polyhedron is shown as it is in the plane normal to vector $g = (1/3, 1/3, 1/3)$. The 3D-polyhedron is in the plane normal to $g = (1/4, 1/4, 1/4, 1/4)$. Basis vectors produced unit simplexes are labelled by numbers. Plots on the right demonstrate how the polyhedrons are filled by the population. For the filling, we assigned all individuals to 1000 clusters. Each point in the plots represents one cluster. The area of each point is proportional to the number of individuals assigned to corresponding cluster. The exception is the point marked by open circles with a closed point inside. About half of the total population was assigned to this cluster.

Comparison of the LLS-score distribution for models with $K = 3$ and $K = 4$ gives us additional arguments that real dimensionality of the latent structure is three. If the

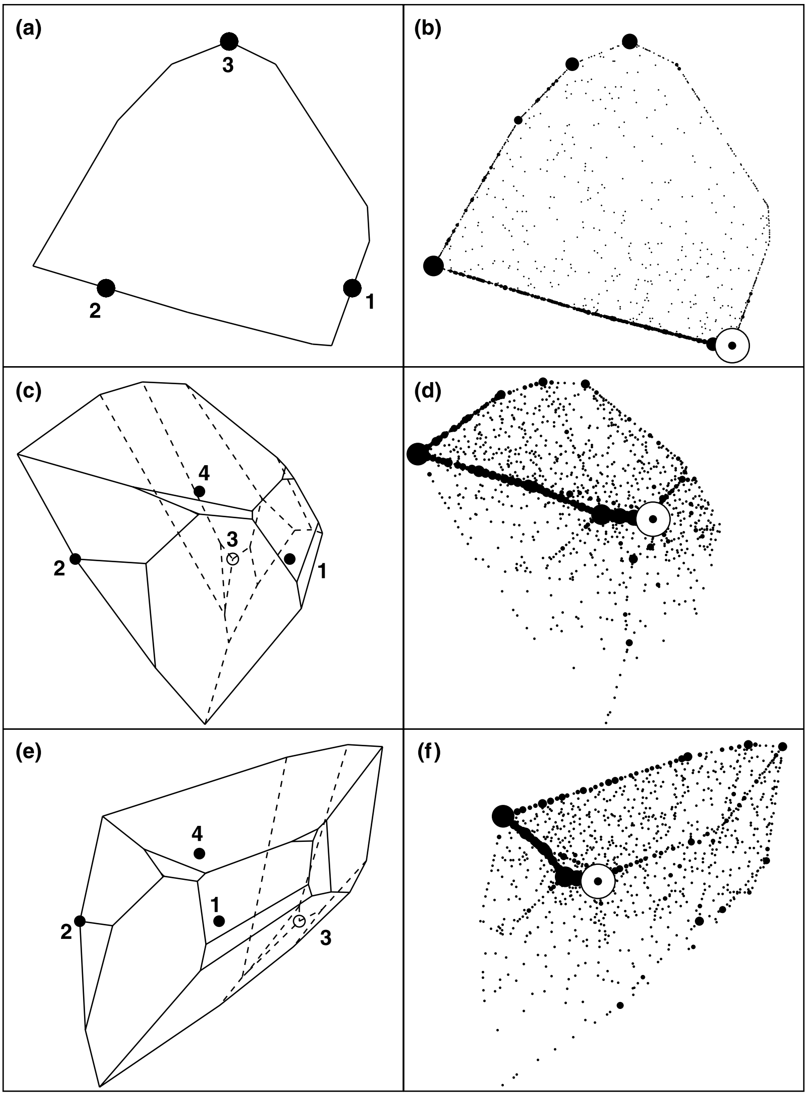

Figure 2. Polyhedrons defined by the LLS constrains for $K = 3$ (a) and $K = 4$ (c and e) and their filling by the LLS scores of NLTCS individuals (b, d, and f). See text for further explanations.

dimensionality of the model is overestimated, the distribution of the LLS scores is concentrated on the subset of the lower dimensions. Approximately, this picture is presented in Figure 2. Majority of individual LLS scores calculated in the model with $K = 4$ is concentrated in the upper plane of the 3D-polyhedron. The density of filling of the additional dimension sharply decreases with movement away from upper plane.

Different regions in polyhedrons of three- and four-dimensions correspond to individuals with different health states, so the LLS scores should be a good predictor of mortality. To check this we used Vital Statistics Data from 1999 to 2003 (August 6) linked to NLTCS. Mortality is modelled by a Cox regression, where vectors of predictors are chosen as g_2, g_3 for $K = 3$ and g_2,g_3,g_4 for $K = 4$, i.e.,

Table 4. Estimates of b_k with SE.

	b_2	b_3	b_4
$K = 3$	0.36 ± 0.06	1.71 ± 0.06	
$K = 4$	0.28 ± 0.07	1.26 ± 0.07	0.01 ± 0.03

$$\mu_{(3)} = \mu_{0(3)} \exp(b_2 g_2 + b_3 g_3) \text{ and } \mu_{(4)} = \mu_{0(4)} \exp(b_2 g_2 + b_3 g_3 + b_4 g_4) \qquad (3)$$

For both models, the healthy component (g_1) was not used as a mortality predictor. This allows us to ignore the restriction ($\sum_k g_k = 1$) and to consider the remaining components as independent variables (predictors). The estimated parameters b_i are presented in Table 4. All estimates for b_2 and b_3 are statistically significant with p-values less than 0.0001. Estimate for b_4 does not show a statistically significant effect. This is the expected result because of the original sense of the fourth group of individuals as a group (i) with intermediate state in health and disability status and (ii) corresponding to an additional fourth component the existence of which is less motivated by the above dimensionality analysis.

5. Discussion and perspectives

In this paper, we demonstrated how a quantitative analysis of survey data can be performed using methods of the LLS analysis. The LLS analysis assumes that measurements reflect a hidden property of subjects that can be described by K-dimensional random vectors. This vector is interpreted as an explanatory variable which can shed light on mutual correlations observed in measured categorical variables. The LLS analysis is intended to discover this hidden property and describe it as precisely as possible.

Basic steps of the analysis include (i) determining the dimension of the investigated database, (ii) identifying the linear subspace of the found dimensionality, which has a sense of the latent structure generating the data, (iii) choosing a basis in found subspace using methods of cluster analysis and/or prior knowledge of the phenomenon of interest, (iv) calculating empirical distribution of the so-called LLS scores which reflect individual responses in the linear subspace, (v) investigating properties of the LLS score distribution to capture population and individual effects (e.g., heterogeneity) and (vi) using components of the vectors of individual LLS scores developing a scheme of prediction of individual lifespan and future changes in health. We performed simulation studies to demonstrate the quality of reconstruction of the major components of the models: (i) low-dimensional subspace, (ii) the LLS scores distribution and (iii) clusters in the LLS score subspace. Results of the simulation studies prove the sufficient quality of reconstruction for typical sample size and demonstrate the potential of the methodology to analyse survey datasets with 1000 or more questions. This methodology was applied to the 1999 NLTCS dataset including 4791 individuals with responses to 57 questions on activities of daily living, instrumental activities of daily living, physical impairment and self-reports of chronic diseases. The estimated dimensionality is three. The components of the space correspond to healthy individuals, disabled (strongly disabled) individuals, and individuals with chronic diseases but without evidence of disability. Empirical distribution of the LLS scores in the found subspace demonstrates heterogeneity of the population with respect to these characteristics. The components of the vectors of individual LLS scores can be used as predictors of individual lifespans. To illustrate this fact we used Medicare Vital Statistics Data from 1999 to 2003 (August 6) linked to NLTCS.

Attractive feature of the methods of latent structure analysis and the LLS in particular is their capability to convert information from individual response patterns to several continuous or discrete measures. If the method of latent structure analysis is chosen properly, i.e., if an assumption on the mathematical form of the latent structure corresponds well to unobservable structure responsible for generating observable data, then such measures will absorb all available information and be largely free from stochastic uncertainties. These individual measures will constitute a useful and convenient parameterization of population heterogeneity originally hidden in a set of interrelated individual responses. Such population measures in the form of a set of individual scores have broad spectrum of applications from clinical individual testing and using them as predictors in prognostic survival models to further development of state-space models, analysis of individual trajectories in this state space, and joint analysis of data from different sources, e.g., in longitudinal analysis of surveys collected in different years. The LLS scores are obtained without parametric assumptions and reflect the underlying dimensions detectable by a used dataset. Thus, the LLS scores have to possess the properties of the best predictors, which can be extracted from used data. We investigated the properties of the LLS scores as predictors of mortality, but many more investigations are to be performed in the future.

The model estimation in the LLS analysis consists of two subsequent steps. In the first step, the dimension of the space and space itself are estimated. The dimension is estimated in the style of the principle component analysis, and all traditional statistics for model tests and significance tests for parameters can be used to estimate the quality of the estimated space in the LLS model as they were used in the principle component analysis. In the second stage, the LLS scores are estimated as expectations of the latent variable conditional on the observed outcome pattern and for the estimated linear space and appropriately chosen basis. An important feature of the LLS analysis is that no parametric assumptions are made at this stage. Estimation for standard error for the LLS score is discussed in Kovtun *et al.* [16]. Comprehensive simulation studies performed above and in Kovtun *et al.* [18] allow us to conclude that if the LLS model is correct then the quality of reconstruction of individual LLS scores is good. An inherent part in the LLS modelling is the analysis of the found latent structure, which is aimed to uncover its interpretation and therefore provides substantive sense of the results of the whole LLS analyses. Note that the underlying idea of the LLS analysis is similar to the one used in the factor analysis. A distinguishing feature is that the LLS searches a low-dimensional structure in the space of probabilities of specific responses rather than in the space of observed variables, which makes it natural for using for a set of categorical variables. This allows to construct the necessary substantive assumptions on how respective probabilities (e.g., such as (1) and (2)) are defined in terms of the LLS scores and the latent structure describing a space on which their distribution is defined.

Historically, the LLS analysis was developed as an attempt to address limitations of the Grade of Membership analysis [23] discussed by Haberman [12]. He noted that further justification of the GoM approach is required. Such justification should involve both rigorous proofs of statistical properties of GoM estimates and extensive simulation studies. Criticism of the GoM was also directed toward the availability of the so-called GoM scores representing individual information. Since GoM scores are estimated for each individual, their number increases with increasing sample size, and therefore, the estimation is non-trivial. The further development of ideas of the GoM analysis resulted in elaboration of a new method of data analysis with new remarkable properties. The first property is the existence of high performance algorithms of parameter estimation [4]. The properties of the new model allowed us to reduce the problem of estimating model parameters to a sequence of problems of linear algebra. The algorithmic approach based on linear algebra methods assures a low computational complexity and an ability to handle data involving potentially thousands of variables. Preliminary studies show that the new numerical scheme is stable and will allow the researcher to include many more variables in an analysis than was possible using the GoM analysis and other latent structure methods. The second feature of the LLS analysis is the understanding of individual scores (LLS scores) as expectations of latent variable conditional on response pattern. This leads to the extension of the set of allowed parameter values. One result of the generalization is that new scores are not restricted to $0 \leq g_i \leq 1$ and, therefore, are not interpreted as grades of membership. Because of these new features of the model we decided to use a new title for this method: Linear Latent Structure analysis.

Mathematically, the LLS approach is based on the theory of mixed distribution. This approach and the properties of the LLS models allow us to establish conditions for identifiability of the LLS models and to rigorously prove the consistency of the estimates of the LLS scores and the space supporting their distribution [18]. Currently, the work on the analytical proof that the LLS method provides non-biased estimates is in progress. However, the simulation studies allow us to state that the bias of estimates for supporting subspace and the LLS scores (if any) is small. Examples 6.1 and 6.2 considered in [18] demonstrated convergence of estimates typical for unbiased estimators.

The remainder of our discussion is devoted to two questions, which are important for future research: perspectives of longitudinal analysis and ways to handle missing data.

This analysis can be naturally generalized to the case of longitudinal data using binomial quadratic hazard model [21] or stochastic process model [38] where the LLS scores play the role of covariates. After estimating the parameters, a scheme for the projection has to be developed, which, in the general case, can be based on microsimulation procedures [2,3].

Furthermore, the LLS approach opens its own possibilities for longitudinal analyses of repeated measurements using its inherent features. Supporting planes obtained for different survey's waves may not coincide with each other. The first question to be analysed is how different supporting planes obtained from different waves are and whether a plane obtained for combined dataset is a good fit or there is a trend in the plane movement over a state space. In the first case, a plane obtained for joint analysis would be taken for longitudinal analysis. The basis becomes time independent and chosen by taking specifics of the problem into account. If a time trend of supporting plane is found, it should be explained in terms of respective changes in population characteristics measured in different waves of such experiments. The basis in this case would first be calculated for a plane for combined dataset, and then projections of basis vectors of specific plane provide a time dependence of the basis. In all cases, we can investigate the difference between results obtained for time dependent and approximate time independent bases.

Another advantage of the LLS analysis is simplicity and naturalness of handling missing data problem. Missing data are generated by the absence of responses for an individual to specific questions. The properties of the LLS approach make this kind of missing data problem relatively easy to handle. Two main sources of missing answers could be considered: first, when the failure to answer the question is random; and second, when the failure to answer the question correlates with answers to other questions. In the first case (missing data are random), the solution is provided by the fact that the input of the LLS algorithm consists of frequencies of partial response patterns (like the frequency of giving response C to the second question and response A to the fifth question). With missing data, such frequencies can be calculated simply by relating the number of individuals with a particular response pattern to the number of individuals who gave answers to the questions covered by the response pattern (as opposed to the total number of individuals). The only drawback of this method is the decreased precision of frequency estimators. As the LLS scores are expectations of latent variables conditional on the arbitrary part of the response pattern for an individual, the available part of the response pattern can be used to estimate the value of the latent variable. In the second case (missing data are not random), the absence of an answer can be considered an additional alternative for answering a multiple-choice question; in this case the standard LLS analysis can be applied.

After such investigation of missing data, the data can be filled by probabilities β^i_{jl} calculated using the LLS scores of known part of outcome pattern and the found linear subspace. This filling probability does not depend on basis choice and on prior information used for basis selection procedure. The imputation procedure being largely model independent could be useful in filling categorical data in historical cohorts or in data where data collection is difficult or costly.

LLS can be used to analyse data where a high dimensional measurement vector represents a hidden structure affecting the results of measurements. An abundance of such data recently appeared in genetic studies of biological systems where the expression of thousands of genes in cells taken from different organs and tissues of humans or laboratory animals is measured. Such measurements are performed to find appropriate regularities in biological functioning of organs and tissues of respective organisms and to detect changes in hidden biological structure due to disease, exposure to external factors, aging related changes, etc. Such an analysis will help us to better understand mechanisms of genetic regulation by suggesting genes that plays key roles in organizing a response to changes in internal or external milieu.

Acknowledgements

This research was supported by Grant Nos. P01 AG17937-05, 5P01-AG08761-16, 5UO1-AG007198-18 and R01 AG28259-01 from National Institute on Ageing.

References

[1] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, Inc, New Jersey, 2002.

[2] I. Akushevich, A. Kulminski, and K. Manton, *Life tables with covariates: Life tables with covariates: Dynamic model for nonlinear analysis of longitudinal data*, Math. Popul. Stud. 12 (2005), pp. 51–80.

[3] I. Akushevich, J.S. Kravchenko, and K.G. Manton, *Health based population forecasting: Effects of smoking on mortality and fertility*, Risk Anal. 27(2) (2007), pp. 467–482.

[4] I. Akushevich, M. Kovtun, A. Yashin, and K.G. Manton, *Linear latent structure analysis: From foundations to algorithms and applications* (2005). Available from e-Print archive arXiv.org at http://www.arxiv.org, arXiv code math.ST/0508299.

[5] D.J. Bartholomew, *Old and new approaches to latent variable modeling*, in *Latent Variable and Latent Structure Models*, G.A. Marcoulides and I. Moustaki, eds., Lawrence Erlbaum Associates, Mahwah, NJ, 2002, pp. 1–14.

[6] D. Böhning, *Computer-assisted Analysis of Mixtures and Applications*, Chapman and Hall/CRC, Boca Raton, FL, 1999.

[7] C.C. Clogg, *Latent Class Models*, in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, G. Arminger, C.C. Clogg and M.E. Sobel, eds., Plenum Press, New York, 1995, pp. 311–360.

[8] E.A. Erosheva, *Comparing latent structures of the grade of membership, Rasch, and latent class models*, Psychometrika 70 (2005), pp. 619–628.

[9] B. Everitt and D. Hand, *Finite Mixture Distributions*, Chapman & Hall, London, 1981.

[10] G.H. Golub and C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, 1996.

[11] L.A. Goodman, *Exploratory latent structure analysis using both identifiable and unidentifiable models*, Biometrika 61 (1974), pp. 215–231.

[12] S.J. Haberman, *Book review of Statistical Applications Using Fuzzy Sets*, by Kenneth G. Manton, Max A. Woodbury, and H. Dennis Tolley, J. Am. Stat. Assoc. 90 (1995), pp. 1131–1133.

[13] J. Hagenaars and A. McCutcheon (eds.), *Applied Latent Class Analysis*, Cambridge University Press, Cambridge, 2002.

[14] T. Heinen, *Latent Class and Discrete Latent Trait Models: Similarities and Differences*, Thousand Oaks, CA: Sage, 1996.

[15] J. Kiefer and J. Wolfowitz, *Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters*, Ann. Math. Stat. 27 (1956), pp. 887–906.

[16] M. Kovtun, I. Akushevich, K.G. Manton, and H.D. Tolley, *Grade of membership analysis: One possible approach to foundations*, in *Focus on Probability Theory*, Nova Science Publishers, NY, 2006, pp. 1–26.

[17] M. Kovtun, A. Yashin, and I. Akushevich, *Convergence of estimators in LLS analysis* (2005). Available from e-Print archive arXiv.org at http://www.arxiv.org, arXiv code math.PR/0508297.

[18] M. Kovtun, I. Akushevich, K.G. Manton, and H.D. Tolley, *Linear latent structure analysis: Mixture distribution models with linear constrains*, Stat. Methodology 4 (2007), pp. 90–110.

[19] P.F. Lazarsfeld and N.W. Henry, *Latent Structure Analysis*, Houghton Mifflin, Boston, 1968.

[20] B.G. Lindsay, *Mixture Models: Theory, Geometry and Applications. Vol. 5 of NSF-CBMS Regional Conference Series in Probability and Statistics*, Institute of Mathematical Statistics, Hayward, CA, 1995.

[21] K.G. Manton, E. Stallard, and B.H. Singer, *Projecting the future size and health status of the U.S. elderly population*, Int. J. Forecasting 8 (1992), pp. 433–458.

[22] K.G. Manton, L. Corder, and E. Stallard, *Changes in the use of personal assistance and special equipment from 1982 to 1989: Results from the 1982 and 1989 NLTCS*, Gerontologist 33 (1993), pp. 168–176.

[23] K.G. Manton, M.A. Woodbury, and H.D. Tolley, *Statistical applications using fuzzy sets*, John Wiley and Sons, NY, 1994.

[24] K.G. Manton, E. Stallard, and L. Corder, *Changes in the age dependence of mortality and disability: Cohort and other determinants*, Demography 34 (1997), pp. 135–157.

[25] K.G. Manton, X. Gu, H. Huang, and M. Kovtun, *Fuzzy set analyses of genetic determinants of health and disability status*, Stat. Methods Med. Res. 13 (2004), pp. 395–408.

[26] G. McLachlan and D. Peel, *Finite Mixture Distributions*, Wiley, New York, 2000.

[27] R.J. Mislevy, *Estimating latent distributions*, Psychometrika 49 (1984), pp. 359–381.

[28] J. Neyman and E.L. Scott, *Consistent estimates based on partially consistent observations*, Econometrica 16 (1948), pp. 1–32.

[29] D.A. Powers and Y. Xie, *Statistical Methods for Categorical Data Analysis*, Academic Press, New York, 2000.

[30] Y. Qu, M. Tan, and M.H. Kutner, *Random effects models in latent class analysis for evaluating accuracy of diagnostic tests*, Biometrics 52 (1996), pp. 797–810.

[31] B. Singer, *Grade of membership representations: Concepts and problems*, in *Probability, Statistics, and Mathematics: Papers in Honor of Samuel Karlin*, T.W. Andersen, K.B. Athreya and D.L. Iglehart, eds., Academic Press, Inc, New York, 1989, pp. 317–334.

[32] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York, 1985.

[33] H.D. Tolley and K.G. Manton, *Large sample properties of estimates of a discrete grade of membership model*, Ann. Inst. Stat. Math. 44 (1992), pp. 85–95.

[34] J.S. Uebersax and W.M. Grove, *A latent trait finite mixture model for the analysis of rating agreement*, Biometrics 49 (1993), pp. 823–835.

[35] J.S. Uebersax, *Analysis of student problem behaviors with latent trait, latent class, and related probit mixture models*, in *Applications of Latent Trait and Latent Class Models in the Social Sciences*, J. Rost and R. Langeheine, eds., Waxmann, New York, NY, 1997, pp. 188–195.

[36] K.W. Wachter, *Grade of membership models in low dimensions*, Stat. Pap. 40 (1999), pp. 439–458.

[37] M.A. Woodbury and J. Clive, *Clinical pure types as a fuzzy partition*, J. Cybernet. 4 (1974), pp. 111–121.

[38] A.I. Yashin and K.G. Manton, *Effects of unobserved and partially observed covariate processes on system failure: A review of models and estimation strategies*, Stat. Sci. 12 (1997), pp. 20–34.

Appendix

The results of a survey are represented by I measurements of random categorical variables $X_1, \ldots, X_J$, with the set of outcomes of the jth measurement being $\{1, \ldots, L_j\}$. The joint distribution of random variables $X_1, \ldots, X_J$ is described by the elementary probabilities

$$p_\ell = \mathsf{P}(X_1 = \ell_1 \text{ and } \ldots \text{ and } X_J = \ell_J), \tag{4}$$

where $\ell = (\ell_1, \dots, \ell_J)$ is an individual response pattern and $\ell_j \in \{1, \dots, L_j\}$. To include into consideration marginal probabilities, we allow some components of ℓ to be 0's. Values of the probabilities p_ℓ (and only these) are directly estimable from the observations. If I_ℓ is the number of individuals with pattern ℓ, consistent estimates for p_ℓ are given by frequencies $f_\ell = I_\ell / I$.

The main model assumption is that the distribution $\{p_\ell\}_\ell$ can be described as a mixture of independent distributions, i.e. distributions satisfying

$$p_\ell = \mathsf{P}(X_1 = \ell_1 \text{ and } \dots \text{ and } X_J = \ell_J) = \prod\nolimits_j \mathsf{P}(X_j = \ell_j). \tag{5}$$

In the latent structure models the probabilities $\mathsf{P}(X_j = \ell_j)$ are considered random variables (i.e., not equal for different individuals), whose distribution reflects population heterogeneity. In LLS these individual probabilities are denoted β^i_{jl} and parameterized according to (1). This is the second assumption of LLS analysis. In (1) λ^k_{jl} are constant parameters defining K-dimensional linear subspace in the space discussed in Section 2. A polyhedron in this space defined by conditions (2) represents a support for a latent variable g describing the population heterogeneity in LLS analysis. Thus, the joint distribution is represented as

$$p_\ell = \int \mathrm{d}F(g) \prod\nolimits_{j:\ell_j \neq 0} \sum\nolimits_{k=1}^{K} g_k \lambda^k_{j\ell_j} \equiv M_\ell. \tag{6}$$

where $F(\beta)$ be the cumulative distribution function of the mixing distribution. Note an important fact regarding the above equation. The value on the left-hand-side, p_ℓ, comes from the *joint distribution of* $X_1, \dots, X_J$, while the value on the right-hand-side, M_ℓ, is a moment of *mixing distribution*; the equality of these values is a direct corollary of the definition of mixture. The existence of the connection between two distinct distributions is crucial for LLS analysis.

Estimation of the model, i.e., finding k-dimensional subspace (or equivalently, estimation of parameters λ^k_{jl}) and estimation of mixing distribution, is based on two properties of the model. The first is formulated for so-called moment matrix (i.e., matrix of the moments M_ℓ) and the second is based on the existence of a system of equations relating the sought model parameters.

The construction of the moment matrix is illustrated by the following example for the toy case of $J = 3$ dichotomous variables, i.e., $L_1 = L_2 = L_3 = 2$.

$$\begin{pmatrix}
M_{(100)} & ? & ? & M_{(110)} & M_{(120)} & M_{(101)} & M_{(102)} & ? & \dots \\
M_{(200)} & ? & ? & M_{(210)} & M_{(220)} & M_{(201)} & M_{(202)} & ? & \dots \\
M_{(010)} & M_{(110)} & M_{(210)} & ? & ? & M_{(011)} & M_{(012)} & ? & \dots \\
M_{(020)} & M_{(120)} & M_{(220)} & ? & ? & M_{(021)} & M_{(022)} & ? & \dots \\
M_{(001)} & M_{(101)} & M_{(201)} & M_{(011)} & M_{(021)} & ? & ? & M_{(111)} & \dots \\
M_{(002)} & M_{(102)} & M_{(202)} & M_{(012)} & M_{(022)} & ? & ? & M_{(112)} & \dots
\end{pmatrix} \tag{7}$$

Not all moments are directly estimable from data. In this example, places for inestimable moments are filled by question marks. This arises because the data do not include double answers to the same question. The first column of the moment matrix contains moments of the first order, when only one specific outcome of one specific question is taken into account. There are no inestimable moments in the first column. The next six ($|L|$ in general) columns correspond to second-order moments. The last shown column corresponds to third order moments.

The part of the moment matrix consisting of second-order moments (which is $|L| \times |L|$ square matrix) together with the column of first-order moments is of special interest. A well-known fact is that if a distribution in an n-dimensional Euclidean space is carried by a k-dimensional linear subspace, then the rank of the covariance matrix is equal to k, and the position of the subspace can be derived from the covariance matrix. This fact is the cornerstone of principal component analysis. Our method of finding the linear subspace supporting the mixing distribution is based on similar ideas. The main property of the moment matrix is formulated as follows. When elements of the moment matrix are moments of a distribution carried by a k-dimensional subspace, every column of the moment matrix belongs to this subspace and the rank of the moment matrix is exactly K. This property allows us to apply well-established methods of computational linear algebra (specifically, methods used in principal component analysis) modified to having an incomplete set of second-order moments and exact relations among elements of the matrix like $M_{001} + M_{002} = 1$. Note, for a small J

(as in the example), there is a relatively large fraction of non-estimable components in the second-order part of the moment matrix. For increasing J, this fraction rapidly decreases.

When the supporting subspace is found and some basis in the subspace is fixed, the mixing distribution can be approximated by empirical distribution, where an individual gives a unit contribution to the histogram of the distribution. This individual contribution coincides with so-called LLS score or the conditional moments $\mathsf{E}(G_k|X=\ell)$, which express knowledge of the state of individuals based on measurements. Explicit expressions for those of the lowest order are obtained using the Bayes theorem ([16]),

$$\mathsf{E}(G_n|X=\ell)=\int \mathrm{d}F(g)g_n\frac{\prod_{j:\ell_j\neq 0}\ \sum_k\ g_k\lambda^k_{jl}}{M_\ell} \tag{8}$$

They are not directly estimable from observations, however they can be found from the main system of equations

$$\sum\nolimits_k\ \lambda^k_{jl}\cdot\mathsf{E}(G_k|X=\ell)=\frac{M_{\ell+l_j}}{M_\ell}, \tag{9}$$

where vector ℓ contains 0 in position j, and $\ell+l_j$ contains l in this position. This system becomes a linear system after substituting the basis. Thus, in the LLS analysis the mixing distribution is approximated by empirical distribution with a support of a set of I points. Probabilities of the joint distribution (4) are estimated as the sum over sample individuals or to the sum over possible outcome patterns

$$p^*_\ell=\frac{1}{I}\sum\nolimits_i\ \prod\nolimits_{j:\ell_j\neq 0}\ \sum\nolimits_k\ g_{ik}\lambda^k_{j\ell_j}=\sum\nolimits_{\ell'}\ f_{\ell'}\prod\nolimits_{j:\ell_j\neq 0}\ \sum\nolimits_k\ g_{\ell' k}\lambda^k_{j\ell_j}. \tag{10}$$

The existence of an algorithm that reduces a problem of non-parametric estimation of the model parameters to a sequence of linear algebra problems assures low computational complexity and the ability to handle problems on desktop computers data involving thousands of variables.

The described estimator is not the only possible one. We chose it because of a number of useful and attractive properties listed below (for details see [4,16–18]).

Identifiability of the LLS model

The LLS model is identifiable if and only if the moment matrix has a completion with the rank equal to the maximal rank of its completed minors. This property holds for almost all (with respect to Lebesgue measure) mixing distributions; thus, LLS models are identifiable almost surely.

Consistency of the LLS model: basis of subspace and LLS scores

The parameters of the LLS model are the exact solutions of the main system of equations, whose coefficients are true moments of the mixing distribution. The solutions of this system continuously depend on its coefficients; thus, consistency of the LLS estimates obtained by the above algorithm is a direct corollary of the known statistical fact that the frequencies are consistent and are efficient estimators of the true moments.

Possibility of nonparametric estimation of mixing distribution

The mixing distribution can be estimated in the style of an empirical distribution, i.e., when the estimator is a distribution concentrated in points $\mathsf{E}(G|X=\ell)$ with weights f_ℓ.

Extraction of individual information

The conditional expectations $\mathsf{E}(G|X=\ell)$ provide knowledge about individuals. These conditional expectations can be considered as coordinates in a phase space, to which all individuals belong. The ability to discover the phase space and determine individual positions in it is a valuable feature of LLS analysis.

Consistency of the LLS model: Mixing distribution

When the number of measurements, J, tends to infinity, the individual conditional expectations $g_i=\mathsf{E}(G|X=\ell^{(i)})$, where $\ell^{(i)}$ is the vector of responses of individual i, converge to the true value of the latent variable for this individual, and estimates of the mixing distribution converge to the true one, depending on some regularity conditions [17].