Modeling High-Dimensional Survey Data Using Latent Structure Analyses

Igor Akushevich,[*] Mikhail Kovtun,[†] Anatoliy I. Yashin[‡]

Abstract

The Linear Latent Structures (LLS) analysis assumes that the mutual correlations observed in survey variables reflect a hidden property of subjects that can be described by low-dimensional random vector. The statistical properties of LLS analysis, the algorithm for parameter estimates and its implementation, simulation studies, and application of LLS model to the National Long Term Care Survey (NLTCS) data are discussed. The results of analyses are compared numerically and analytically to predictions of the Latent Class and Grade of Membership analyses. Simulation studies demonstrate high quality of reconstruction of the major model components and demonstrate its potential to analyze survey datasets with 1000 or more questions. Applying the LLS model to the 1994 and 1999 NLTCS datasets (5,000+ individuals) with responses to over 200 questions on behavior factors, functional status, and comorbidities resulted in identified population structure with basis represented pure-type individuals, e.g., healthy, strongly disabled, having chronic diseases, etc. The components of the vectors of individual LLS scores are used to make predictions of individual lifespans.

Key Words: Multidimensional categorical data; demographic surveys; latent analysis; health state; Grade of Membership; Latent Class Model

1. Introduction

Survey data typically represent sample-based collections of measurements made with discrete outcomes for individuals. Common property of such datasets is high dimensionality, and that measured variables are highly correlated. Methods dealing with such tasks are known as latent analysis. Typical assumption in the methods is that the observed structure of multiple categorical variables are generated by the small number of latent (i.e., unobserved) variables. The task of latent analyses is to find these latent variables, estimate parameters of their distribution, and describe their properties using a sample of high dimensional categorical variables. Generally speaking, it is necessary to find the properties of a population, associated with latent variables, and properties of individuals, based on those multiple categorical measurements. It appears that both goals may be achieved simultaneously. To the increase precision of population and individual estimates, one has to increase both the sample size (i.e., the number of individuals) and the number of measurements (i.e., questions asked for each individual).

One of the best known of methods of latent analyses is the latent class model (LCM), which can be characterized as a statistical method for finding discrete subtypes of related cases (latent classes) from multivariate categorical data. Other models of this type (known as latent variable models), such as item response theory and Rasch models, differ by the assumptions made about the latent variable(s) (reviewed by Clogg, 1995, and Collins and Lanza, 2010). 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 introduced by Woodbury and Clive (1974). Manton et al. (1994) provided a detailed exposition of different version of this approach and reviewed its properties.

[*] Center for Population Health and Aging, Duke University, Durham, NC 27708

[†] Biology Department, Duke University, Durham, NC 27708

[‡] Center for Population Health and Aging, Duke University, Durham, NC 27708

Recently Linear Latent Structure (LLS) analysis has been proposed to model high dimensional categorical data (Kovtun et al., 2006, 2007, Akushevich et al., 2009). The LLS specific assumption is that the support for latent variable occupy a polyhedron of lower dimensionality. The LLS model was formulated using the mixing distribution theory. Similar to other latent structure analyses, the goal of LLS analysis is to derive simultaneously the properties of a population and individuals, using discrete measurements. The LLS, however, does not use maximization of a likelihood for parameter estimation. Instead, it uses an estimator, where the LLS parameter estimates are solutions of a quasilinear system of equations.

2. Linear Latent Structure Analysis

2.1 Structure of datasets and population characteristics.

The typical dataset analyzed by methods of latent structure analysis can be represented by the $I \times J$ matrix constituted by categorical outcomes X_j^i of J measurements on I individuals, where $i = 1, \dots, I$ and $j = 1, \dots, J$ run over individuals and measurements, respectively. Each row in the matrix corresponds to an individual and contains an individual response pattern, i.e., a sequence of J numbers with the jth 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 dozens. Thus, the results of a survey are represented by I measurements of random variables $X_1, \dots, X_J$, with the set of outcomes of the jth measurement being $\{1, ..., L_j\}$. The joint distribution of random variables $X_1, \dots, X_J$ can be described by the elementary probabilities,

$$p_\ell = \Pr\left(X_1 = \ell_1 \;\text{ and }\; \cdots \;\text{ and }\; X_J = \ell_J\right), \tag{1}$$

where $\ell = (\ell_1, ..., \ell_J)$ is an individual response pattern and $\ell_j \in \{1, ..., L_j\}$. To include into consideration marginal probabilities, we allow some components of ℓ to be 0's. For example, for three binary variables,

$$p_{(2,0,1)} = \Pr(X_1 = 2 \;\text{ and }\; X_3 = 1) = p_{(2,1,1)} + p_{(2,2,1)}.$$

Values of these 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 $f_\ell = I_\ell / I$.

2.2 LLS task: statistical, geometrical, and mixing distribution points of view.

The problem in LLS analysis is to evaluate dimension of a hidden space, identify its location in the space of larger dimension, and to evaluate hidden individuals' characteristics (coordinates in the latent sub-space) from the data. The LLS analysis is based on two assumptions. The first is the assumption about "local independence", which is common for all methods of latent structure analysis. The second is specific for LLS analysis. It is about existence of low-dimensional linear subspace associated with the latent structure. We present LLS in terms of the theory of mixing distributions, and then discuss its specific assumption from statistical and geometrical points of view.

Population characteristics are completely described by the joint distribution of random variables $X_1, \dots, X_J$ presented by probabilities (1). Among all possible joint distributions, one can distinguish independent distributions, i.e. distributions satisfying,

$$p_\ell = \Pr\left(X_1 = \ell_1 \text{ and } \cdots \text{and}\, X_J = \ell_J\right) = \prod\nolimits_j \Pr\left(X_j = \ell_j\right). \tag{2}$$

The description of an independent distribution law requires only knowing $\Pr(X_j = \ell_j)$ denoted below as β_{jl}. Vectors of probabilities $\beta = (\beta_{11}, \ldots, \beta_{JL_J})$ belong to vector space $R^{|L|}$, where $|L| = \sum_j L_j$. Indexes of the vector components run over all possible pairs of jl, i.e., corresponding to probabilities of the first outcome to the first question, of the second outcome to the first question, and so on. Requirements for β_{jl} to be probabilities restricts their domain in the vector space by

$$\sum\nolimits_{l=1}^{L_j} \beta_{jl} = 1 \qquad \text{and} \qquad \beta_{jl} \geq 0. \tag{3}$$

This domain represents the direct product of J unit simplex of dimensions L_j.

Since variables $X_1, \ldots, X_J$ in general case are not independent, the observed distribution $\{p_\ell\}$ cannot be described by the product of independent distributions, but it can be exactly described as a mixture of independent distributions. This means that each set of independent probabilities contributes to observed distribution with a weight function. This weight function is called mixing distribution. It is defined in the space of independent distributions, i.e. for each vector of probabilities β satisfying (3). Let $F(\beta)$ be the cumulative distribution function of the mixing distribution. Probabilities p_ℓ are represented as,

$$p_\ell = \int dF(\beta) \prod\nolimits_{j=1}^{J} \beta_{j\ell_j}. \tag{4}$$

Thus, latent structure analysis searches for a representation of the observed distribution as a mixture of independent distributions.

Any distribution $\{p_\ell\}$ can be represented as a mixture, so representation (4) does not restrict the family of distributions and further specifications are required. They are formulated as restrictions on the support of mixing distribution or, equivalently, on a set of mixed independent distributions. The LLS specific assumption is that this set is restricted to be a K-dimensional linear subspace of the space of independent distributions, i.e., the mixing distribution is supported by the linear subspace spanned by K basis vectors $\lambda^1, \ldots, \lambda^K$. Below this LLS assumption is considered from the point of view of pure statistical analysis and the geometry of the task.

Individual characteristics are described by individual probabilities $\beta^i_{jl} = \Pr(X^i_j = l)$ of specific outcomes ($i = 1, \ldots, I$ runs over individuals).

The LLS assumption about the existence of a low-dimensional linear subspace supporting the mixing distribution is essentially equivalent to the assumption that there exists a K-dimensional random vector G such that for every j a regression of Y_{jl} (random variable Y_{jl} equaling 1 if $X_j = l$ and 0 otherwise) on G is linear. The regression equation relates the expectation of Y_{jl}, which is β_{jl}, to the random vector G. If a specific value of G is associated with individual i (so-called LLS scores g_{ik}), then the regression takes the form,

$$\beta^i_{jl} = \sum\nolimits_{k=1}^{K} g_{ik} \lambda^k_{jl}. \tag{5}$$

The sense of the regression coefficients λ^k_{jl} and model restrictions is clarified by analyzing the geometry of the LLS task.

Vectors of individual probabilities $\beta^i = \{\beta^i_{jl}\}$, of individual responses $Y^i = \{Y^i_{jl}\}$ and the regression coefficients $\lambda^k = \{\lambda^k_{jl}\}$ lie in the permitted domain (3) of the space of independent distributions. From a geometric point of view, LLS searches a K-dimensional subspace (represented by λ^k_{jl}) in this space, which is the "closest" to the set of I points representing individual outcomes Y^i_{jl}. This linear subspace is defined by its basis $\lambda^1, \ldots, \lambda^K$, so to find the K-dimensional subspace means finding a basis, λ^k_{jl}, $(k = 1, \ldots, K)$. Every basis $\lambda^1, \ldots, \lambda^K$ defines a family of regression coefficients and vice versa. The complete set of restrictions in the LLS task allowing to consider β^i_{jl} and λ^k_{jl} as probabilities, is,

$$\sum_{l=1}^{L_J} \lambda^k_{jl} = 1, \quad \lambda^k_{jl} \geq 0, \quad \sum_{k=1}^{K} g_{ik} = 1 \quad \text{and} \quad \sum_{k} g_{ik}\lambda^k_{jl} \geq 0. \tag{6}$$

LLS scores g_{ik} characterizing an individual i are then estimated as the expectation of vector G, conditional on the respondent's answers. Basis vectors of the subspace can be interpreted as probabilities and can define the "pure types" (Manton et al., 1994). In this sense, the model decomposition (5) has the interpretation of a decomposition over pure types or over "ideal persons" whose individual probabilities are basis vectors of the subspace.

Summarizing, one can say that the LLS model approximates the observed distribution of $X_1, \ldots, X_J$ by a mixture of independent distributions with a mixing distribution supported by a K-dimensional subspace of the space of independent distribution. To specify such a model distribution it is sufficient to define the following LLS parameters:

1. A basis $\lambda^1, \ldots, \lambda^K$ of the space that supports the mixing distribution.

2. Conditional moments $\boldsymbol{E}(G|X = \ell)$.

This set of model parameters is not the only set possible. We chose it because of a number of useful properties listed below.

Property 1. The mixing distribution can be estimated in the style of an empirical distribution, i.e., when the estimator is a distribution concentrated in points $\boldsymbol{E}(G|X = \ell)$ with weights f_ℓ.

Property 2. The conditional expectations $\boldsymbol{E}(G|X = \ell)$ 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.

Property 3. When the number of measurement, J, tends to infinity, the individual conditional expectations $g_i = \boldsymbol{E}(G|X = \ell^{(i)})$, 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, provided some regularity conditions (Kovtun et al., 2011).

2.3 Moment matrix and the main system of equations.

Parameter estimation is based on two facts (Kovtun et al., 2005a,b) formulated in terms of the conditional and unconditional moments of the mixing distribution. The first is that columns of moment matrix belong exactly to the desired linear space. The second is that they obey the main system of equations.

2.3.1 Unconditional moments and the moment matrix

The first set of values in which we are interested consists of the unconditional moments of the mixing distribution $F(\beta)$,

$$M_\ell = \int dF(\beta) \prod\nolimits_{j:\ell_j \neq 0} \beta_{j\ell_j} p_\ell. \tag{7}$$

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

The first corollary of eq. (7) is that the unconditional moments are directly estimable from data and, therefore, the frequencies f_ℓ of response patterns ℓ observed in a sample are consistent and efficient estimators for conditional moments M_ℓ.

Recall that we allow some components of response pattern ℓ to be 0. In this case p_ℓ are marginal probabilities. In the definition of M_ℓ the multipliers, corresponding to 0 components of ℓ, are excluded from the product. Thus, the order of moment M_ℓ is equal to the number of non-zero components in ℓ.

All moments defined in (7) are estimable by frequencies; however, this definition does not cover all moments of a certain order. For example, moments of second order with β_{jl_1} and β_{jl_2}, (i.e., with the same j) are not estimable. This arises because the data do not include double answers to the same question. One can notice that i) all moments of first order are estimable, ii) the proportion of estimable moments decreases with the increase of order, and iii) no moments of order $J + 1$ and higher are estimable.

The moment matrix is constructed from moments of order up to J using the following formal rules:

1. Rows of the moment matrix are indexed by response patterns containing exactly one non-zero component or, equivalently, by pair indexes jl. Thus, the moment matrix contains $|L|$ rows, and their columns can be considered as vectors in $R^{|L|}$.

2. Columns of the moment matrix are indexed by all possible response patterns, including a response pattern containing all 0's. The first column is indexed by response pattern $(0, \ldots, 0)$; the next $|L|$ columns are indexed by response patterns containing one non-zero component, and so on.

3. The element on the intersection of row ℓ' and column ℓ'' is $M_{\ell'+\ell''}$, if ℓ'' has 0 at the position of the only non-zero component of ℓ' (in this case, $\ell' + \ell''$ is a meaningful response pattern; otherwise, the question mark is placed on the position of intersection of row ℓ' and column ℓ''). For example, the element of the moment matrix in row $(1, 0, 0)$ and column $(0, 2, 2)$ is $M_{1,2,2}$, and element in row $(1, 0, 0)$ and column $(1, 0, 2)$ is a question mark.

Equation (8) gives an example of a portion of the moment matrix for the case of $J = 3$ dichotomous variables, i.e., $L_1 = L_2 = L_3 = 2$.

$$\begin{pmatrix} M_{(100)} & ? & ? & M_{(110)} & M_{(120)} & M_{(101)} & M_{(102)} & ? & \cdots \\ M_{(200)} & ? & ? & M_{(210)} & M_{(220)} & M_{(201)} & M_{(202)} & ? & \cdots \\ M_{(010)} & M_{(110)} & M_{(210)} & ? & ? & M_{(011)} & M_{(012)} & ? & \cdots \\ M_{(020)} & M_{(120)} & M_{(220)} & ? & ? & M_{(021)} & M_{(022)} & ? & \cdots \\ M_{(001)} & M_{(101)} & M_{(201)} & M_{(011)} & M_{(021)} & ? & ? & M_{(111)} & \cdots \\ M_{(002)} & M_{(102)} & M_{(202)} & M_{(012)} & M_{(022)} & ? & ? & M_{(112)} & \cdots \end{pmatrix} \tag{8}$$

In this example, places for inestimable moments are filled by question marks. 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. Elements of this column can be denoted as components of vectors in $R^{|L|}$, i.e., as M_{jl}. The next six ($|L|$ in general) columns correspond to second-order moments. Blocks of diagonal elements are not estimable. Second-order moments can be also denoted via pair jl of indexes as $M_{jl;j'l'}$. The last shown column corresponds to third order moments. The notation M_{jl} and $M_{jl;j'l'}$ is used below for specific columns of the moment matrix.

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-know fact is that if a distribution in an n-dimensional Euclidean space is carried by a k-dimensional linear manifold, then the rank of the covariance matrix is equal to k, and the position of the manifold can be derived from the covariance matrix. This fact is the cornerstone of principal component analysis. Our method is based on similar ideas, adapted to having an incomplete set of second-order moments. 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.

For a moment matrix M let its completion $\bar{M}$ be a matrix obtained from M by replacing question marks with arbitrary numbers. The main fact with respect to the moment matrix is that the moment matrix always has a completion in which all columns belong to the supporting plane Λ. Thus, if the estimable part of the moment matrix has sufficient rank (which is the case in nondegenerate situations,) a basis in Λ may be obtained from this matrix. As we have a consistent estimator of the moment matrix in the form of a frequency matrix, the supporting plane may be consistently estimated.

2.3.2 Unconditional moments and main system of equations

Another set of the values of interest are the conditional moments $\boldsymbol{E}(G_k|X = \ell)$, which express knowledge of the state of individuals based on measurements. They are not directly estimable from observations. The goal of LLS analysis is to obtain estimates for these conditional moments. Explicit expressions for those of the lowest order are obtained using the Bayes theorem (Kovtun et al. 2007),

$$\boldsymbol{E}(G_n|X = \ell) = \int dF(g) g_n \frac{\prod_{j:\ell_j \neq 0} \sum_k g_k \lambda^k_{jl}}{M_\ell(\mu_\beta)}. \tag{9}$$

Analogously, higher conditional moments, including products of components of G, can be constructed. As can be seen explicitly from (9), the relation of conditional and unconditional moments in LLS analysis can be described as,

$$\sum\nolimits_k \lambda^k_{jl} \cdot \boldsymbol{E}(G_k|X = \ell) \frac{M_{\ell+l_j}}{=} M_\ell, \tag{10}$$

where vector ℓ contains 0 in position j, and $\ell + l_j$ contains l in this position. Similar equations can be written for conditional moments of higher orders. We refer to the system of equations relating conditional and unconditional moments as the main system of the equation. Kovtun et al, (2006) proved the following properties of solutions of the main system of equations: i) any basis λ^k_{jl} of Λ together with conditional moments $\boldsymbol{E}(G_k|X = \ell)$ calculated on this basis give a solution of the main system of equation; and ii) under regular conditions, every solution of the main system of equations gives a basis of Λ and conditional moments calculated in this basis. Note, that equation (10) is linear with respect to conditional moments.

The described properties of the moment matrix and solutions of the main system of equations suggest an efficient algorithm to obtain LLS estimates. First, a basis of the supporting plane can be obtained from the moment matrix, and second, conditional moments can be found by solving a linear system of equations.

2.3.3 Two illustrative examples.

Before going into detail for the algorithm and to realistic tasks of data analysis, we consider two simple illustrative examples. For both of them, assume $K = 2$, three dichotomous variables ($J = 3$), and the basis vectors are $\lambda^1 = (1, 0; 1, 0; 1, 0)$ and $\lambda^2 = ({}^1\!/\!_2, {}^1\!/\!_2; 0, 1; 0, 1)$. Then the independent distributions being mixed are defined by vectors:

$$\beta = g_1\lambda^1 + g_2\lambda^2 = g_1\lambda^1 + (1 - g_1)\lambda^2, \qquad 0 \leq g_1 \leq 1. \tag{11}$$

Thus, a mixing distribution can be given one dimensional p.d.f. $\rho(g_1)$. For the first task, we assume that the mixing distribution is uniform ($\rho(g_1) = 1 \cdot \theta(g_1) \cdot \theta(1 - g_1)$. In the second case we assume the mixing distribution is concentrated at two points with $g_1 = 0.1$ and $g_1 = 0.4$ ($\rho(g_1) = {}^1\!/\!_2[\delta(g_1 - {}^1\!/\!_{10}) + \delta(g_1 - {}^2\!/\!_5)]$). Unconditional moments are calculated using (7). Moment matrices for both cases are

$$\begin{pmatrix} \frac{3}{4} & \mathit{\frac{7}{12}} & \mathit{\frac{1}{6}} & \frac{1}{2} & \frac{1}{4} & \frac{5}{12} & \frac{1}{3} \\ \frac{1}{4} & \mathit{\frac{1}{6}} & \mathit{\frac{1}{12}} & \frac{1}{8} & \frac{1}{8} & \frac{1}{12} & \frac{1}{6} \\ \frac{5}{8} & \frac{1}{2} & \frac{1}{8} & \mathit{\frac{7}{16}} & \mathit{\frac{3}{16}} & \frac{3}{8} & \frac{1}{4} \\ \frac{3}{8} & \frac{1}{4} & \frac{1}{8} & \mathit{\frac{3}{16}} & \mathit{\frac{3}{16}} & \frac{1}{8} & \frac{1}{4} \\ \mathbf{\frac{1}{2}} & \mathbf{\frac{5}{12}} & \frac{1}{12} & \frac{3}{8} & \frac{1}{8} & \mathit{\frac{1}{3}} & \mathit{\frac{1}{6}} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{6} & \frac{1}{4} & \frac{1}{4} & \mathit{\frac{1}{6}} & \mathit{\frac{1}{3}} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \frac{5}{8} & \mathit{\frac{317}{800}} & \mathit{\frac{183}{800}} & \frac{451}{1600} & \frac{549}{1600} & \frac{67}{400} & \frac{183}{800} \\ \frac{3}{8} & \mathit{\frac{183}{800}} & \mathit{\frac{117}{800}} & \frac{249}{160} & \frac{351}{1600} & \frac{33}{400} & \frac{117}{400} \\ \frac{7}{16} & \frac{451}{1600} & \frac{249}{160} & \mathit{\frac{653}{3200}} & \mathit{\frac{747}{3200}} & \frac{101}{800} & \frac{249}{800} \\ \frac{9}{16} & \frac{549}{1600} & \frac{351}{1600} & \mathit{\frac{747}{3200}} & \mathit{\frac{1053}{3200}} & \frac{99}{800} & \frac{351}{800} \\ \mathbf{\frac{1}{4}} & \mathbf{\frac{67}{400}} & \frac{33}{400} & \frac{101}{800} & \frac{99}{800} & \mathit{\frac{17}{200}} & \mathit{\frac{33}{200}} \\ \frac{3}{4} & \frac{183}{400} & \frac{117}{400} & \frac{249}{800} & \frac{351}{800} & \mathit{\frac{33}{200}} & \mathit{\frac{117}{200}} \end{pmatrix} \tag{12}$$

Since these matrices were constructed from mixing distributions known a priory, diagonal blocks in the sub-matrix of the second order are calculable (marked by the italic font in (12)). As one can see, the rank of both these matrices is 2. Conditional moments are calculated for an outcome pattern. Choose $\ell = (001)$ and $\ell + l_1 = (101)$. Using (9) we have,

$$\boldsymbol{E}(G_1|X = (001)) = 2/3 \text{ and } \boldsymbol{E}(G_2|X = (001)) = 1/3 \tag{13}$$

for the first example and,

$$\boldsymbol{E}(G_1|X = (001)) = 17/50 \text{ and } \boldsymbol{E}(G_2|X = (001)) = 33/50 \tag{14}$$

for the second. Using corresponding elements of M_ℓ in (12) (marked by bold text) we can see that l.h.s. and r.h.s of eq. (10) equal to $5/6$ for first example and $67/100$ for the second:

$$1 \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{1}{3} = \frac{5/12}{1/2} \qquad \text{and} \qquad 1 \cdot \frac{17}{50} + \frac{1}{2} \cdot \frac{33}{50} = \frac{67/400}{1/4}. \tag{15}$$

External indexes in this example are $j = 1$ and $l = 1$.

3. Computational Algorithm for Estimating LLS model

Parameter estimations in LLS models are based on properties of the moment matrix and the main system of equations. These properties allow us to reduce a problem of estimating model parameters to a sequence of linear algebra problems. The algorithm based on linear algebra methods assures a low computational complexity.

Data to be analyzed are represented by a set of measurements X_j^i (See section 2.1). Finding a linear space and individual LLS scores is required. Estimation of the model includes four steps: i) estimating the rank of the frequency matrix, ii) finding the supporting plane, iii) choosing a basis in the found plane, iv) calculating individual conditional expectations and estimating mixing distribution. The second and fourth steps are the essence of LLS parameter estimation problem. The first step is defined as separate because sometimes the desired dimensionality of the LLS model may be provided by a researcher, and this step may be skipped. The third step requires using prior information about the processes studied, so it is also examined separately.

3.1 Moment matrix calculation

An important preceding step that deserves special attention is the moment matrix calculation. The elements of the moment matrix given by M_ℓ are approximated by observable frequencies defined as $f_\ell = I_\ell/I$, where I_ℓ is the number of individuals with outcome pattern ℓ, and I is the total number of individuals having certain (not missing) outcomes for nonzero elements in ℓ. Columns of a different order have different normalizations, e.g., the sum of first-order moments corresponding to question j is one (e.g., $M_{(010)} + M_{(020)} = 1$), while sums of columns for this j of the second-order sub-matrix are equal to corresponding first-order moment (e.g., $M_{(110)} + M_{(120)} = M_{(100)}$). General conditions of summations of the second order moments written in terms of notation defined after eq. (8) are,

$$\sum\nolimits_{l'=1}^{L_{j'}} M_{jl;j'l'} = M_{jl}. \tag{16}$$

Because of missing data, the property of normalization can be violated. This property, with or without the renormalization making the sums equaling to one, is required for the analysis. The renormalization could provide the property in the case of presence of missing data, however, this approximation can be true only assuming missing data are random.

In addition, a matrix containing standard errors (or confidence intervals) of estimates of frequencies is calculated for each element of the frequency matrix. Standard errors for binomial distribution, i.e. $\sigma_\ell = \sqrt{f_\ell(1-f_\ell)/I_\ell}$, require generalization for patterns with small I_ℓ as discussed in Brown et al. (2001).

3.2 Computational rank of the frequency matrix

The frequency matrix can be presented as a sum of the moment matrix with rank K and a matrix with a stochastic component. To define the dimensionality of the LLS problem, we have to estimate the rank of the frequency matrix eliminating the stochastic component. Specifically, we take the greatest minor of the frequency matrix that does not contain question marks. Then we calculate the singular value decomposition (SVD) and take K equal to the number of singular values that are greater than a maximum of the total standard deviation estimated as the quadratic sum of standard errors of frequencies involved in the minor.

The choice of a minor does not essentially influence the computational rank of the frequency matrix. Indeed, the geometrically specific choice of a minor (e.g. a n-dimensional minor of maximal size in left low corner of moment matrix) corresponds to projection of a part of vectors onto n-dimensional linear subspace. If the real rank of the moment matrix is much less than n, it is clear that the rank of the projections does not change.

3.3 Finding the supporting plane

All columns of the moment matrix belong to the supporting plane, and as the frequency matrix is an approximation of the moment matrix, a natural way to search for the supporting plane is to search for a plane that minimizes the sum of distances from it to the columns of the frequency matrix. In our case, however, this way is complicated by: (a) the frequency matrix is incomplete; (b) the statistical inaccuracy of approximation of moments M_ℓ by frequencies f_ℓ varies considerably over elements of frequency matrix; and (c) a sought basis should exactly satisfy conditions $\sum_{l=1}^{L_J} \lambda_{jl}^k = 1$ for every k and j. These obstacles are overcome by using some heuristic methods: (a) An iterative procedure for completion of the frequency matrix is used: after a basis of supporting plane is obtained, it is used to recalculate completion of the frequency matrix. A new frequency matrix is used for adjusting basis calculation etc. (b) Only the first and second order moments are examined, so statistical errors of different columns in this matrix are compatible. (c) Rotation of each simplex (corresponding to each question) to the hyperplane to eliminate one degree of freedom. Rotation, but not a simple projection, is required to provide the same distances between points in a simplex. Items (a) and (c) require explicit consideration.

3.3.1 Completion of the moment matrix

We consider the second-order moment matrix where for every $\bar{j}$ there are undefined elements corresponding to repeated answers to the same question. The intent of completion procedure is to approximate these elements, assuming that the supporting subspace Λ is found. Since only the completed frequency matrix is used for finding subspace Λ, and since the completion procedure uses a basis in the sought subspace Λ, it can be done within the iteration procedure. For one iteration step, it is required to find a symmetric matrix $B_{\bar{j}}$ of $L_{\bar{j}} \times L_{\bar{j}}$-dimension with positive elements and the required summation conditions such that the sum of elements in a column (or in a row) equals to the corresponding moment of the first order, i.e., $\sum_l B_{\bar{j},ll'} = M_{\bar{j}l'}$. Since we know first- and second-order frequencies (f_{jl} and $f_{jl,j'l'}$; $j \neq j'$), which only approximate exact moments (M_{jl} and $M_{jl,j'l'}$), special efforts are required to process the properties of $B_{\bar{j}}$. Columns of the second-order sub-matrix corresponding to question $\bar{j}$ are presented using known frequencies $f_{jl,\bar{j}\bar{l}}$; $j \neq \bar{j}$ and inestimable elements $B_{\bar{j},l\bar{l}}$,

$$\begin{pmatrix} f_{11;\bar{j}1} & \cdots & f_{11;\bar{j}L_{\bar{j}}} \\ \cdots & \cdots & \cdots \\ f_{1L_1;\bar{j}1} & \cdots & f_{1L_1;\bar{j}L_{\bar{j}}} \\ \cdots & \cdots & \cdots \\ B_{\bar{j},11} & \cdots & B_{\bar{j},1L_{\bar{j}}} \\ \cdots & \cdots & \cdots \\ B_{\bar{j};L_{\bar{j}}1} & \cdots & B_{\bar{j};L_{\bar{j}}L_{\bar{j}}} \\ \cdots & \cdots & \cdots \\ f_{11;\bar{j}1} & \cdots & f_{J1;\bar{j}L_{\bar{j}}} \\ \cdots & \cdots & \cdots \\ f_{JL_J;\bar{j}1} & \cdots & f_{JL_J;\bar{j}L_{\bar{j}}} \end{pmatrix} \tag{17}$$

The completion procedure is based on the fact that the rank of the moment matrix is K, which is much smaller than the dimension of matrix $|L|$. Therefore, only K columns are linearly independent. Each column of the moment matrix, being a vector in K-dimensional vector space, can be expanded over basis vectors $\lambda^1, \ldots, \lambda^K$ available after finding the subspace Λ. Known elements $f_{jl;\bar{j}\bar{l}}$ ($\bar{l} = 1, \ldots, L_{\bar{j}}$ and $j \neq \bar{j}$) of columns of the moment matrix corresponding to question $\bar{j}$ are expanded,

$$f_{jl;\bar{j}\bar{l}} = \sum\nolimits_k C_k^{\bar{j}\bar{l}} \lambda_{jl}^k \qquad (j \neq \bar{j}). \tag{18}$$

If coefficients $C_k^{\bar{j}\bar{l}}$ are found, matrix $B_{\bar{j}}$ can be constructed as $B_{\bar{j},\bar{l}'\bar{l}} = \sum\limits_k C_k^{\bar{j}\bar{l}} \lambda_{\bar{j}\bar{l}'}^k$, The number of known components of a vector $f_{jl;\bar{j}\bar{l}}$ is greater than the number of basis vectors, so coefficients $C_k^{\bar{j}\bar{l}}$ can be calculated by ordinary least squares with restrictions: $C_k^{\bar{j}\bar{l}} \geq 0$, $\sum_k C_k^{\bar{j}\bar{l}} = 1$ and $\sum_k \big(C_k^{\bar{j}\bar{l}} \lambda_{\bar{j}\bar{l}'}^k - C_k^{\bar{j}\bar{l}'} \lambda_{\bar{j}\bar{l}}^k\big) = 0$. The functional to be minimized is:

$$\sum_{jl:j \neq \bar{j}} \left(f_{jl;\bar{j}\bar{l}} - \sum_k C_k^{\bar{j}\bar{l}} \lambda_{jl}^k \right)^2 . \tag{19}$$

3.3.2 Removing restrictions

The restrictions $\sum_{l=1}^{L_J} \lambda_{jl}^k = 1$ are removed by reducing the number of rows by J (one for every group of indexes $j1, \ldots, jL_j$). Specifically, we use a linear map from $R^{|L|}$ to $R^{|L|-J}$ represented by a block-diagonal matrix A with J blocks of size $L_j \times (L_j - 1)$:

$$A_j = \begin{pmatrix} -\frac{\sqrt{L_j}-1}{L_j-1} & 1 & 0 & \ldots & 0 \\ \multicolumn{5}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots} \\ -\frac{\sqrt{L_j}-1}{L_j-1} & 0 & 0 & \ldots & 1 \end{pmatrix}. \tag{20}$$

Geometrically, such a map provides isometric rotation ($\bar{\lambda}^k = A\lambda^k$) to the hyperplane with zero first coordinate, i.e., (every block A_j defines a rotation of a unit simplex in L_j–dimensional space around a hypersurface opposite to the first vertex; the angle of this rotation is such that the first vertex moves to the point where the first coordinate equals 0). Explicitly, this rotation is $\bar{\lambda}_{jl-1}^k = A_j \lambda_{jl}^k$ in matrix form or $\bar{\lambda}_{jl-1}^k = \lambda_{jl}^k - \frac{\sqrt{L_j}-1}{L_j-1}\lambda_{j1}^k$ for $l = 2, \ldots, L_j$. New vectors $\bar{\lambda}^k$ do not possess any ties. It is easy to ascertain that such a transformation really conserves distances between points in a simplex. The reverse transformation is,

$$\lambda_{j1}^k = \frac{1 - \sum_{l=2}^{L_j} \bar{\lambda}_{jl-1}^k}{\sqrt{L_j}}, \quad \lambda_{jl}^k = \bar{\lambda}_{jl-1}^k + \frac{\sqrt{L_j}-1}{L_j-1}\lambda_{j1}^k. \tag{21}$$

3.3.3 Algorithm for identifying the subspace

The initial completion of the moment matrix is constructed in a arbitrary way, e. g, by the unitary diagonal matrix or completing by frequencies as $f_{ij} = f_i f_j$. The next preliminary step is the rotation of each simplex (corresponding to each question as described above) to the hyperplane to eliminate one degree of freedom. This produces n points $c^1, \ldots, c^n$ (images of columns of frequency matrix) in $m = (|L| - J)$-dimensional space. The problem is to find an affine plane that minimally deviates from these points in the space of individual probabilities. First, we find the center of gravity of this system

$$c^0 = \frac{1}{n}\sum\nolimits_i c^i, \tag{22}$$

and then consider a new set of points $\bar{c}^i = c^i - c^0$, that corresponds to shifting the point of origin. Now we need to find a K-dimensional linear subspace in R^m that minimally deviates from this set of points. The solution of this problem is well-known: one has to consider an $m \times m$ matrix X with components$X_{rs} = \sum_i \bar{c}_r^i \bar{c}_s^i$; this matrix is symmetric and positively defined, and thus its normalized eigenvectors are composed of an orthonormal basis in R^m. Let $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_m > 0$ be eigenvalues of matrix X, and let $z^1, \ldots, z^m$ be corresponding eigenvectors. The plane of dimensionality K that minimizes the sum of squared distances from points $\bar{c}^1, \ldots, \bar{c}^n$ is spanned by $z^1, \ldots, z^m$, and the sum of squared distances is tr $X - \sum_{k=1}^K \gamma_k$. Vectors $c^0, c^0 + z^1, \ldots, c^0 + z^{K-1}$ give us an affine basis of the sought affine plane. Finally, we apply inverses of transformation (21) to $c^0, c^0 + z^1, \ldots, c^0 + z^{K-1}$ to obtain the sought basis $\lambda^1, \ldots, \lambda^K$ of the subspace Λ.

3.4 Choice of a basis

The basis cannot be defined uniquely, and any convex combination of basis vectors keeping the LLS restrictions can be considered an alternative. A choice may be made using prior information about the process of interest. The appeal of prior information at this stage is reasonable because of the evident fact that the same dataset can be used for analyzing different (say, disability or CVD) substantive tasks.

The way how this information is used and how the procedure of specific choice of the basis is defined is a question of taste. We describe here two possible schemes used in our analyses.

A researcher specifies the characteristics of "ideal" individuals based on his/her experience in the research domain. Then he/she can construct vectors of probabilities $\beta_{j\ell_j}$ for such ideal individuals or take these individuals from the sample under consideration. The vectors of probabilities of these individuals are taken as basis vectors. If probability vectors are constructed by hand, they could be beyond polyhedron P_g, so they should be projected to P_g. The individual coordinates in this basis would represent "proximity" of the individual to the "ideal" ones.

In another scheme, the basis is obtained using assignment of LLS scores (calculated on some arbitrary basis) to K clusters, and then basis vectors $\lambda^1, \ldots, \lambda^K$ are calculated as means of probabilities β^i_{jl} over each cluster.

A researcher can develop his/her own scheme of basis selection. For example, he/she can simply use vectors already known from previous studies or construct a basis purely mathematically, e.g., from the condition of maximal linear independence of the vectors, or choose it from the set of the supportive polyhedron vertexes.

3.5 Calculation of individual conditional expectations.

When a basis of the supporting plane is found, the conditional expectations can be found from the main system of equations (10), which is a linear system after substituting the basis. The system, however, relates conditional expectations $\boldsymbol{E}(G_k|X = \ell)$ for a pattern ℓ with at least one $0th$ outcome. Thus exact system of equations (10) can be written for all patterns ℓ except patterns where all outcomes are known. For the complete patterns, we can calculate J conditional expectations, subsequently excluding one of J questions (i.e., obtaining patterns $\ell^{[j]}$, where $\ell^{[j]}$ denotes vector ℓ with j^{th} coordinate equal to 0), solving the exact system of equations for obtained patterns, and defining LLS score for complete pattern as mean over J solutions for conditional expectations for $\ell^{[j]}$ patterns. This approach can be formalized by considering a system of J system of equations:

$$\sum\nolimits_k \lambda^k_{jl} \cdot g_{\ell k} \approx \frac{f_\ell}{f_{\ell^{[j]}}}. \tag{23}$$

This is a sparse overdetermined system that is solved by minimizing the functional

$$\sum_j \left(\sum_k \lambda^k_{jl} \cdot g_{\ell k} - \frac{f_\ell}{f_{\ell^{[j]}}} \right)^2 \tag{24}$$

using least squares with restrictions $\sum_k g_{\ell k} = 1$ and $\sum_k \lambda^k_{jl} \cdot g_{\ell k} \geq 0$. It is implemented using SAS Proc NLP (SAS, Cary NC).

3.6 Mixing distribution

The mixing distribution for an analyzed set of data is approximated by empirical distribution, where an individual gives a unit contribution to the histogram of the distribution. A support of this distribution is 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,

$$p^*_\ell = \sum\nolimits_i \prod\nolimits_{j:\ell_j \neq 0} \beta^i_{j\ell_j} = \sum\nolimits_{\ell'} f_{\ell'} \prod\nolimits_{j:\ell_j \neq 0} \sum\nolimits_k g_{\ell' k} \lambda^k_{j\ell_j}. \tag{25}$$

3.7 Properties of LLS estimator

Kovtun et al. (2007) proved identifiability and consistency 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. 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.

4. Applications

4.1 Simulation Studies

Three types of simulation experiments were performed to test the predictive power of LLS model and its ability to reveal and to quantitatively reconstruct a hidden latent structure. Specifically they were focused on analyzing the quality of reconstruction of: i) linear subspace; ii) LLS mixing distribution; and iii) clustering properties. The results demonstrated an acceptable quality of reconstruction. Details of the design of these studies and results were described in Akushevich et al. (2009).

4.2 LLS and latent class models

The geometric approach, which considers independent distributions as points in finite-dimensional linear space and mixing distributions as measures in this space, allows us to clarify relationship between various branches of latent structure analysis. Here we consider relation between LLS models and latent class models (LCM).

In geometric language, latent classes are points in the space of independent distributions. If an LCM with classes $c_1, \ldots, c_m$ exists for a particular dataset, then an LLS model also exists, and its supporting subspace is the linear subspace spanned by vectors $c_1, \ldots, c_m$. Thus, dimensionality of LLS model never exceeds the number of classes in LCM. These numbers are equal if and only if LCM classes are points in general position (n points are said to be in general position, if they do not belong to any linear manifold of dimensionality smaller than $n - 1$).

If LCM classes are not in general position, however, the dimensionality of LLS model may be significantly smaller. For example, it is possible to construct a mixing distribution such that (a) it is supported by a line (i.e., dimensionality of LLS model is 2); (b) there exists LCM with J (number of variables) classes; (c) there is no LCM with smaller number of classes. If, however, the mixing distribution is supported by an infinite set (as in example 1 above), a latent class model does not exist at all, while LLS analysis performs well. On the other hand, LLS can be used to evaluate applicability of LCM: if the mixing distribution in LLS model has pronounced modality, then an LCM is more likely to exist (with the number of classes equal to number of modes). When both LCM and LLS models are applicable, the LLS model may still be model of choice, due to its lower computational complexity.

4.3 LLS and Grade of Membership Models

Parameters of GoM model are estimated by maximizing the likelihood function,

$$\prod_{\ell}\Big(\prod_{j}\sum_{k} g_{\ell k}\lambda^{k}_{j\ell_j}\Big)^{f_\ell}. \tag{26}$$

Proof of consistency of maximum likelihood estimates is not done for GoM model. Nevertheless, under modest conditions (which usually are satisfied in practical situations,) a solution of the classic GoM problem provides reliable estimates. Roughly speaking, a point of maximum of (26) converge to true values when *both* size of the sample, N, and number of measurements, J, tend to infinity. The idea of the proof is to show that when *both* size of the sample, N, and number of measurements, J, tend to infinity, then the point where maximum of (26) is achieved converges to: i) $\lambda^1, \ldots, \lambda^K$ converge to a basis $\tilde{\Lambda} = \{\tilde{\lambda}^k\}$ of the support of measure μ_β, and ii) g_ℓ converge to conditional expectations $\boldsymbol{E}(G \mid X = \ell)$, calculated with respect to the basis $\tilde{\Lambda}$. The most important question here is how to define properties, which an infinite system of measurements should satisfy. We shall show that reasonable assumption lead to the property: ”For sufficiently big J, at the point of maximum $g_{\ell'}$ is very close to $g_{\ell''}$ for every choice of ℓ', ℓ'' that differ only in one component.” Now rewrite (26) as

$$\prod_{\ell}\Big(\sum_{k} g_{\ell k}\lambda^{k}_{1\ell_1}\Big)^{f_\ell}\cdot\ldots\cdot\prod_{\ell}\Big(\sum_{k} g_{\ell k}\lambda^{k}_{J\ell_J}\Big)^{f_\ell} \tag{27}$$

then take j^{th} factor of (27) and rewrite it as

$$\prod_{\ell'\in\mathcal{L}^{[j]}}\bigg(\Big(\sum_{k} g_{\ell'+\mathbf{1}_j,k}\lambda^{k}_{j1}\Big)^{f_{\ell'+\mathbf{1}_j}}\cdot\ldots\cdot\Big(\sum_{k} g_{\ell'+(\boldsymbol{L}_j)_j,k}\lambda^{k}_{jL_j}\Big)^{f_{\ell'+(\boldsymbol{L}_j)_j}}\bigg). \tag{28}$$

Due to the above property, we have for every $l', l'' \in [1..L_j]$ that $g_{\ell'+\boldsymbol{l}'_j,k} = g_{\ell'+\boldsymbol{l}''_j,k}$. From this:

$$\sum_{l=1}^{L_j}\sum_{k} g_{\ell'+\boldsymbol{l}_j,k}\lambda^{k}_{jl} = \sum_{k} g_{\ell',k}\sum_{l}\lambda^{k}_{jl} = \sum_{k} g_{\ell',k}\cdot 1 = 1. \tag{29}$$

Thus, in (28) we have a product of positive factors, which sum is a constant. Such a product reaches maximum when factors are proportional to their powers:

$$\Big\{\textstyle\sum_k g_{\ell'+\boldsymbol{l}_j,k}\lambda^{k}_{jl} = \frac{f_{\ell'+\boldsymbol{l}_j}}{f_{\ell'}}, \quad l\in[1..L_j]\ \cdot \tag{30}$$

This means that $g_{\ell k}$ and λ^k_{jl} that deliver maximum to (26) satisfy the system of equations (10) and consequently, by the theorem 5.1 of Kovtun et al. (2007), we obtain required properties.

4.4 Application to the 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+). The used dataset is described in Akushevich et al. (2011).

The first 10 singular values of frequency matrix of NLTCS (σ_E=0.292):

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

When the dimensionality of the LLS-problem is fixed, we can complete the moment matrix using the algorithm described in Section 3.3. The sub-matrix corresponding to the first four dichotomous variables is,

$$\begin{pmatrix} 0.094 & \mathbf{0.513} & \mathbf{0.051} & 0.328 & 0.011 & 0.258 & 0.012 & 0.518 & 0.014 \\ 0.906 & \mathbf{0.487} & \mathbf{0.949} & 0.672 & 0.989 & 0.742 & 0.988 & 0.482 & 0.986 \\ 0.264 & 0.918 & 0.196 & \mathbf{0.633} & \mathbf{0.128} & 0.688 & 0.051 & 0.846 & 0.153 \\ 0.736 & 0.082 & 0.804 & \mathbf{0.367} & \mathbf{0.872} & 0.312 & 0.949 & 0.154 & 0.847 \\ 0.335 & 0.916 & 0.275 & 0.872 & 0.142 & \mathbf{0.664} & \mathbf{0.164} & 0.888 & 0.230 \\ 0.665 & 0.084 & 0.725 & 0.128 & 0.858 & \mathbf{0.336} & \mathbf{0.836} & 0.112 & 0.770 \\ 0.160 & 0.879 & 0.085 & 0.514 & 0.034 & 0.424 & 0.027 & \mathbf{0.640} & \mathbf{0.069} \\ 0.840 & 0.121 & 0.915 & 0.486 & 0.966 & 0.576 & 0.973 & \mathbf{0.360} & \mathbf{0.931} \end{pmatrix}$$

Completed values are marked in bold style.

On the basis of the cluster analysis we choose $K = 3$ clusters corresponding to i) individuals with minor chronic diseases without disability ($k = 1$), ii) individuals with medium to severe chronic diseases, severe disabled ($k = 2$), and iii) individuals with medium chronic diseases and minor to medium disability ($k = 3$). For $K = 4$ case, an additional cluster ($k = 4$) intermediate between ($k = 1$) and ($k = 3$) is added. An extended set of variables (J=230) allows us to identify two additional groups out of group i) with similar set and severity of chronic diseases: a) very active physically and socially individuals without disabilities, psychologically healthy, and b) moderately physically and social active individuals with minor disabilities and minor to moderate psychological disorders.

Polyhedrons defined by the LLS constrains for K=3 (a) and K=4 (c,e) and their filling by the LLS scores of NLTCS individuals (see Figure 1 for $K = 3$). The plot on the left shows 2D-polyhedron for $K = 3$. The case of $K = 4$ is considered by Akushevich et al. (2009). The polyhedron is defined by the LLS restrictions. In this case, the LLS scores are restricted by 130 inequalities ($\sum_k g_{ik}\lambda^k_{jl} \geq 0$) and one equality ($\sum_k g_{ik} = 1$). Basis vectors produced unit simplexes are labeled by numbers. Plots on the right demonstrate how the polyhedrons are filled by the population. For the filling, we assigned all individuals to 1,000 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.

An extended set of variables (J=230) allows us to identify two additional groups of individuals: i) individuals with high physical and social activities and without disabilities, and psychologically healthy and ii) individuals with moderate physical and social activities, with minor disabilities, and minor to moderate psychological disorders.

Mortality is modeled 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., $\mu_{(3)} = \mu_{0(3)} \exp(b_2g_2 + b_3g_3)$ and $\mu_{(4)} = \mu_{0(4)} \exp(b_2g_2 + b_3g_3 + b_4g_4)$. The estimates are b_2=0.36±0.06, b_3=1.71±0.06 for K=3, and b_2=0.28±0.07, b_3=1.26±0.07, and b_4=0.01±0.03 for K=4.

5. Conclusion

LLS is a model describing high-dimensional categorical data assuming existence of a latent structure represented by K-dimensional random vectors. This vector is interpreted as explanatory variables which can shed light on mutual correlations observed in measured categorical variables. This vector plays the role of a random variable mixing independent distribution such that the observed joint distribution is maximally close to the data. Mathematically, LLS analysis considers the observed joint distribution of categorical variables as

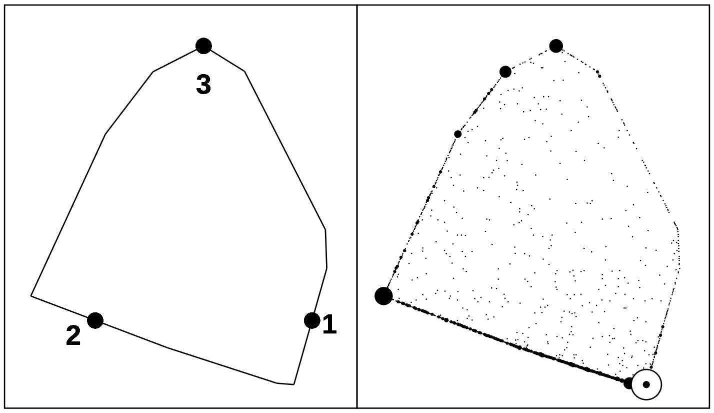

Figure 1: Polyhedrons defined by LLS constrains for K=3 and their filling by LLS scores of NLTCS individuals.

a mixture of individual joint distributions, which are assumed to be independent. Explicit consideration of the space of mixed distributions as a linear space leads to a fruitful developments, resulting in a new method as well as in a better understanding of the existing methods.

An important distinction is the existence of an algorithm capable of estimating a LLS model for large numbers of questions and individuals. The estimators of the parameters may be used for construction of second-level models (for example, when the application domain justifies assumption about parametric structure of the mixing distribution). 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.

REFERENCES

Akushevich, I., Kovtun, M., Manton, K. G., and Yashin, A. I.(2009), “Linear latent structure analysis and modeling of multiple categorical variables”, *Computational and Mathematical Methods in Medicine*, 10, 203–218.

Akushevich, I., Kravchenko, J., Akushevich, L., Ukraintseva, S., Arbeev, K., and Yashin, A., (2011), “Cancer Risk and Behavioral Factors, Comorbidities, and Functional Status in the US Elderly Population,” *ISRN Oncology* 2011, Article ID 415790, 9 pages

Brown, L.D., Cai, T.T., and DasGupta, A. (2001), “Interval estimation for a binomial proportion.” *Statistical Science* 16: 101–117.

Collins, L.M., Lanza S.T. (2010), *Latent class and latent transition analysis for the social, behavioral, and health sciences*. New York: Wiley.

Clogg, C.C. (1995), *Latent Class Models.* In “Handbook of Statistical Modeling for the Social and Behavioral Sciences”, Arminger, G., Clogg, C.C., Sobel, M.E., eds., New York: Plenum Press, 311–360.

Kovtun, M., Akushevich, I., Manton, K.G., and Tolley H.D. (2006), *Grade of Membership Analysis: One Possible Approach to Foundations.* In *Focus on Probability Theory*, Nova Science Publishers, NY, 2006, pp. 1–26.

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

Kovtun, M., Akushevich, I., Yashin A.I. (2011), “Linear Latent Structure Analysis (LLS).” In *JSM Proceedings*, Section on Nonparametric Statistics. Alexandria, VA: American Statistical Association.

Manton, K.G., Woodbury M.A., & Tolley H.D. (1994). *Statistical applications using fuzzy sets.* John Wiley and Sons, New York.

Woodbury MA and Clive J. (1974). “Clinical pure types as a fuzzy partition. Journal of Cybernetics,” 4: 111-121.