Grade of Membership Analysis: Newest Development with Application to National Long Term Care Survey Data*

Mikhail Kovtun Igor Akushevich Kenneth G. Manton
Center for Demographic Studies, Duke University, Durham, NC

H. Dennis Tolley
Department of Statistics, Brigham Young University, Provo, UT

March 18, 2004

Abstract

Newest developments in the theoretical foundation of the Grade of Membership (GoM) analysis performed by authors bring new insights into the value of GoM analysis, providing new ways to interpret estimates as well as new numerical methods to obtain estimates.

In this presentation we discuss the most important results obtained recently and apply the new methodology to the National Long Term Care Survey (NLTCS) data. We analyze the obtained results by providing an interpretation for the estimates and by investigating the errors arising from the limited size of the sample. This analysis shows that GoM model performs well in the case considered here.

1 Introduction

The grade of membership analysis was introduced in [10, 11], and later developed in [8] and other articles. Another possible point of view on GoM analysis was developed in [9]. Recently authors developed a new approach (described in details in [4].) This presentation summarize most important facts of [4] and demonstrates usefulness of the method by examining National Long Term Care Survey data.

The grade of membership (GoM) analysis considers a number of discrete measurements on individuals. The goal of GoM analysis is to derive some properties of a population as well as of an individual based on results of these measurements. In the sense that it tries to uncover the underlying hidden structure,

*This research was supported by grants from National Institute of Aging.

GoM relates to the latent structure analysis [1, 2, 3]. However, GoM uses different algorithms and is based on different mathematical foundation.

A good example of the applicability of GoM is in making a medical diagnosis. A physician has to make a conclusion about health state of a patient based on a number of measurements, which include both objective measurements (such as measuring blood pressure) and subjective measurements (such as asking specific questions about health.) Decision making requires knowledge about the population (what does it mean "to be healthy", or "to have this disease") derived from results of similar measurements performed on other individuals.

The ability to derive properties of a population is provided by sampling a sufficient number of individuals, whereas the ability to derive properties of an individual is provided by a sufficiently large number of measurements on each individual (in practice, several dozen.) The situation is complicated by the fact that the measurements, on the one hand, should relate to the same underlying problem (individual state,) while, on the other hand, *must* be different (to avoid bias for a particular individual.) Thus there is no obvious relation between results of different measurements.

2 Theoretical results

2.1 The formulation of the problem

GoM analysis considers J discrete measurements, represented by random variables $X_1, \ldots, X_J$, with the set of outcomes of j^{th} measurement being $\{1, \ldots, L_j\}$. The main assumption of GoM is: for some K, there exists a K-dimensional random vector G such that for every j a regression of Y_j on G is linear. Here Y_j is an L_j-dimensional random vector, $Y_j = \mathbf{1}_l$ if $X_j = l$ (where $\mathbf{1}_l$ denotes a vector which has l^{th} component equal to 1, and all other components equal to 0.)

This essentially means that one can consider an $|L| = L_1 + \cdots + L_J$-dimensional random vector of probabilities $\beta = (\beta_{jl})_{jl}$, where j ranges from 1 to J, and for every j, l ranges from 1 to L_j. Realizations of this random vector are distribution laws for individual random vectors $X^i = (X^i_1, \ldots, X^i_J)$, i.e. $\Pr(X^i_j = l) = \beta^i_{jl}$. Let μ_β be a probabilistic measure describing the distribution of β. Then the linear regression hypothesis is equivalent to an assumption that a support of μ_β is a K-dimensional linear subspace Q of $\mathbb{R}^{|L|}$. Let $\Lambda = \{\lambda^1, \ldots, \lambda^K\}$, $\lambda^k = (\lambda^k_{jl})_{jl}$, be any basis of Q, and for $\beta \in Q$, let $g = (g_k)_{k=1,\ldots,K}$ be its coordinates in basis Λ. Then the random vector G is the random vector β written in coordinates g. Let μ_g be a measure μ_β written in coordinates g.

Another assumption made in GoM analysis is that random variables $X_1, \ldots, X_J$ are conditionally (or locally) independent, i.e.

$$\Pr(X_1 = \ell_1 \wedge \cdots \wedge X_J = \ell_J \mid G = g) = \prod_j \Pr(X_j = \ell_j \mid G = g) \qquad (1)$$

This assumption is widely used in the latent structure analysis. One pos-

sible motivation for such an assumption is that all "randomness" in individual random variables $X_1^i, \ldots, X_J^i$ comes from errors in measurements, and error in one measurement does not depend on error in another one.

Let $\ell = (\ell_1, \ldots, \ell_J)$ be an integer vector with $0 \le \ell_j \le L_j$. Such a vector represents the outcome of J measurements, and $\ell_j = 0$ means that we do not take into account the outcome of the j^{th} measurement. Thus, a value of $\ell_j = 0$ in a vector ℓ means that the vector is a marginal vector across all values of the j^{th} measurement. Let $\mathcal{L}^0$ be a set of all such vectors, and for every $\mathcal{J} \subseteq \{1, \ldots, J\}$ let $\mathcal{L}^{[\mathcal{J}]}$ be a set of vectors having 0's exactly on places from $\mathcal{J}$. Let $v = (v_1, \ldots, v_K)$ be an integer vector with $v_k \ge 0$, and for every integer $J' \ge 0$ let $\mathcal{V}[J']$ be a set of such vectors satisfying the additional condition $\sum_k v_k = J'$.

In this language, the values of interest are unconditional moments of the distribution μ_β

$$M_\ell(\mu_\beta) = \int \prod_{j\,:\,\ell_j \neq 0} \beta_{j\ell_j}\, \mu_\beta(d\beta) \tag{2}$$

and conditional moments of distribution μ_g,

$$\mathcal{E}(G^v \mid X = \ell) = \int \prod_k g_k^{v_k} \frac{\prod_{j\,:\,\ell_j \neq 0} \sum_k g_k \lambda_{jl}^k}{M_\ell(\mu_\beta)}\, \mu_g(dg) \tag{3}$$

The unconditional moments $M_\ell(\mu_\beta)$ are the probabilities of obtaining the response pattern ℓ (under assumptions of the model.) Thus, frequencies of response patterns ℓ in a sample, denoted f_ℓ, are consistent and efficient estimators for unconditional moments $M_\ell(\mu_\beta)$.

The conditional moments $\mathcal{E}(G^v \mid X = \ell)$ express our knowledge of the state of the individual (represented by random vector G) based on outcomes of the measurements. These values are not directly estimable from the observations. The goal of GoM analysis is to obtain estimates for these conditional moments.

The conditional moments $\mathcal{E}(G^v \mid X = \ell)$ provide the basis for a parsimonious summary of the data under the assumption of the GoM model. If the model fits to data, the conditional moments provide predicted outcomes to use for a goodness of fit test and a model to forecast future values.

2.2 The main system of equations

We have shown in [4] that the GoM model defined above is fully described by a system of equations (with respect to variables α_{jl}^k and h_ℓ^v)

$$\begin{cases} \sum_k \alpha_{jl}^k h_\ell^{v+\mathbf{1}_k} = h_{\ell+\boldsymbol{\iota}_j}^v, & J' \in [0..J-1],\ \ v \in \mathcal{V}[J'], \\ & \mathcal{J} \subseteq [1..J] : |\mathcal{J}| > J',\ \ \ell \in \mathcal{L}^{[\mathcal{J}]}, \\ & j \in \mathcal{J},\ \ l \in [1..L_j] \\ h_\ell^{(0,\ldots,0)} = M_\ell, & \ell \in \mathcal{L}^0 \\ \sum_{v \in \mathcal{V}[J']} \frac{(\sum_k v_k)!}{\prod_k v_k!} h_{(0,\ldots,0)}^v = 1, & J' \in [0..J] \end{cases} \tag{4}$$

More precisely,

1. Any basis Λ of Q together with conditional moments $\mathcal{E}(G^v \mid X = \ell)$ calculated in this basis give a solution of (4) (λ^k_{jl} should be substituted for α^k_{jl}, and $M_\ell(\mu_\beta) \cdot \mathcal{E}(G^v \mid X = \ell)$ should be substituted for h^v_ℓ.)
2. Under mild conditions, *every* solution of (4) gives a basis of Q and conditional moments calculated in this basis.

Consequently, we can obtain conditional moments by solving this system of equations. We see that the observed moments, $M_\ell(\mu_\beta)$, play a crucial role in the actual estimation of the GoM model. Two important points in task of fitting a model are choosing the dimensionality K and estimating a basis Λ. We detail each of these in the following section.

2.3 The moment matrix

Let us write a vector of moments $(M_{\boldsymbol{\iota}_j})_{jl}$ together with incomplete vectors $(M_{\boldsymbol{\iota}_{j'}+\boldsymbol{\iota}_j})_{jl\,:\,j\neq j'}$, etc., as columns of a matrix, with places for which we do not have moments filled by question marks. We refer to this incomplete matrix as the *moment matrix*. The moment matrix contains a column for every $\ell \in \mathcal{L}^0$. Figure 1 gives an example of a portion of a moment matrix for the case $J = 3$, $L_1 = L_2 = L_3 = 2$. Columns in this matrix correspond to $\ell = (000)$, (100), (200), (010), (020), (001), (002), (110); other columns are not shown.

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

Figure 1: Example of moment matrix

Note that certain moments (which are replaced by question marks in the moment matrix) are not observable. The reason for this is that we do not have possibility to perform a measurement on an individual multiple times independently, and since individuals are heterogeneous (have different probabilities of outcomes of measurements,) we do not have multiple realizations of independent identically distributed random variables.

For a moment matrix M let its completion $\bar{M}$ be a matrix obtained from M by replacing question marks by arbitrary numbers. We have shown that

the moment matrix always has a completion, in which all columns belong to Q (recall that Q is a K-dimensional subspace containing all individual vectors of probabilities, $(\beta^i_{jl})_{jl}$.) Thus, if the moment matrix has sufficient rank (which is the case in most practical situations,) a basis of Q may be obtained from this matrix. And we have an estimator of the moment matrix in form of frequency matrix.

In particular, the (uncompleted) moment matrix gives a way to estimate lower boundary for the dimensionality of the GoM problem, K: it may not be lower then rank of nonsingular minor of the moment matrix. In practice, we will use the frequency matrix, which is an approximation of the moment matrix. In this case a minor will be considered as nonsingular only if it is nonsingular for a range of values of the moments, say all moments within a two standard deviation interval of the observed frequencies.

2.4 Maximum likelihood considerations

The second issue arising from the system of equation (4) is the method of estimating the basis and conditional moments with respect to this basis using the system (4). To date we have not implemented a stable method based on (4). However, an approximation can be derived which is based on the original formulation of a GoM estimation algorithm as described in [10]. It suggested that estimates for basis Λ and conditional expectations $\mathcal{E}(G^v \mid X = \ell)$ should be obtained by maximization of the function:

$$\prod_i \Big(\prod_j \sum_k g_{ik} \lambda^k_{jx^i_j} \Big) \tag{5}$$

where i ranges over individuals in the sample, and x^i_j is the outcome of j^{th} measurement on i^{th} individual.

This approach was motivated by maximum likelihood reasoning: if a vector $g_i = (g_{i1}, \ldots, g_{iK})$ is the hidden state (i.e. the value of random vector G) of individual i, then (5) is the probability of observing outcomes x^i_j.

As estimates for g_i may depend only on outcomes of measurements for the i^{th} individual, they are equal for individuals with equal outcomes. Thus, (5) may be rewritten as:

$$\prod_\ell \Big(\prod_j \sum_k g_{\ell k} \lambda^k_{j\ell_j} \Big)^{f_\ell} \tag{6}$$

But these likelihood functions contain incidental parameters ($g_{\ell k}$ in case of (6) and g_{ik} in case of (5),) and therefore the estimates need not be consistent; and, in general, they are not. Nevertheless, (5) and (6) may be used to obtain reliable estimates, in the following sense.

By straightforward but tedious algebraic steps one may show that solving main system of equations (4) is equivalent to maximization of the function

$$\prod_{\ell} \Big(\prod_{j} \sum_{k} g_{\ell^{[j]}k} \lambda^{k}_{j\ell_j} \Big)^{f_\ell} \tag{7}$$

where $\ell^{[j]}$ denotes vector ℓ with j^{th} component replaced by 0.

Furthermore, recently we have shown (manuscript in preparation) that when J tends to infinity, the point where (6) reaches its maximum converges to the point where (7) reaches its maximum. Thus, for sufficiently large number of measurements, maximization of (6) (or, equivalently, (5)) provides approximate estimates for the values of interest (a basis Λ and conditional expectations with respect to it.) Numerical results reported below are derived by maximization of (6) and using the estimates as solution to (7).

The advantage of maximizing (6)) is that there exists a well established and efficient numerical procedure. We are working on new numerical procedures based on (4) and (7).

3 Application to NLTCS data

We applied the above theoretical consideration to the National Long Term Care Survey (NLTCS) data.

The National Long Term Care Survey is a longitudinal survey designed to study changes in the 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 resources for caregiving. The survey began in 1982, and follow-up surveys were conducted in 1984, 1989, 1994, and 1999. A sixth follow-up survey will be conducted during 2004. A detailed description of NLTCS may be found at `http://nltcs.cds.duke.edu/`.

We considered a sample of approximately 5,000 individuals from 1999 NLTCS wave, and selected 27 questions, which characterize disability level with respect to activities of daily living, instrumental activities of daily living, and physical impairment. Details about these questions may be found in [5, 6, 7].

Every individual is subject to 27 questions, 20 of which have 2 possible answers, and 7 have 4 possible answers. Thus, we have $L_1 = \cdots = L_{20} = 2$, $L_{21} = \cdots = L_{27} = 4$, and $|L| = 68$.

3.1 Analysis of the moment matrix

The first task is to find out the dimensionality of the GoM problem, K. As it was mentioned above, K is the rank of the moment matrix, and thus it might be estimated as the rank of the frequency matrix.

We used the singular value decomposition to estimate the rank of the frequency matrix (more precisely, we used a matrix, which elements are numbers of corresponding response patterns in the sample — which is size-of-sample times bigger then frequency matrix.) As the frequency matrix is incomplete, we actually made decomposition not of the whole matrix, but of its left bottom corner of size 31×31. The singular values are given in table 1.

Table 1: Singular values of the frequency matrix

σ_1	39292.861	σ_7	90.993	σ_{13}	28.077
σ_2	4780.040	σ_8	70.721	σ_{14}	20.488
σ_3	791.574	σ_9	56.614	σ_{15}	24.379
σ_4	212.424	σ_{10}	49.728	σ_{16}	5.564
σ_5	162.809	σ_{11}	36.629	σ_{17}	0.000
σ_6	119.109	σ_{12}	30.257	σ_{18}	0.000

The computations show that the hypothesis $K = 6$ fits the data under the assumption that frequencies are in two standard deviations interval from true moments. However, there is no significant gap between the 6^{th} and 7^{th} singular numbers. This suggests that a support of distribution is an ellipsoid of full dimensionality which is thinner in higher dimensions, and choosing a particular value for K approximates this ellipsoid by lower-dimensional ellipsoid obtained from the true one by collapsing a number of smaller axis.

3.2 Comparing classic estimates with the frequency matrix

We compare estimates obtained by maximization of (6) with the frequency matrix.

In the GoM model, columns of frequency matrix belongs to the linear span of vectors $\lambda^1, \ldots, \lambda^K$. Thus, quality of approximation is characterized by closeness of columns of frequency matrix to the above linear span.

We calculate angles between columns of the frequency matrix and linear span of $\lambda^1, \ldots, \lambda^K$. The table 2 contains for the first 12 columns of frequency matrix (a) euclidean length of a column, (b) euclidean distance between a column and the linear span; and (c) angle (in degrees) between a column and linear span.

Another applicable test is how close is the moment matrix generated by the model parameters to the observed frequency matrix. We generate the moment matrix from the model parameters by formula:

$$M_\ell = \frac{1}{N} \sum_i \Big(\prod_{j\,:\,\ell_j \neq 0} \sum_k g_{ik} \lambda^k_{j\ell_j} \Big) \tag{8}$$

and compare the generated moment matrix with the frequency matrix derived from the data. The result of comparison is given in the table 3. Cells of the table correspond to the 10×6 fragment of the moment matrix; values are differences between calculated moments and frequencies expressed in standard deviations. The bottom row of the table gives averages over the columns; the average over the whole matrix is 1.840.

One can see that the model moments fit the observed data pretty well except several cases.

Table 2: Angles between linear span of $\lambda^1, \ldots, \lambda^K$ and columns of the frequency matrix

Length	Distance	Angle
4.19464	0.16533	2.25889
4.00205	0.42757	6.13302
4.26949	0.16902	2.26887
3.75365	0.33482	5.11746
4.51090	0.15884	2.01796
3.75087	0.25945	3.96634
4.59338	0.16875	2.10537
3.87431	0.40055	5.93423
4.39225	0.15977	2.08467
3.74094	0.24287	3.72234
4.64082	0.15478	1.91123
3.75590	0.33040	5.04680
4.46987	0.15564	1.99548

3.3 Distribution of individual states

The figures 2 and 3 show a 2-dimensional projections of distribution of individual's hidden state (described by the random vector G.) We give only 2 of the 15 possible projections of 6-dimensional picture into a 2-dimensional coordinate plane. From these figures we see that the values of conditional expectations vary considerably across individuals indicating a high level of heterogeneity.

3.4 Correlation with diseases

To demonstrate validity of GoM analysis, we investigated correlation between the estimated hidden state represented by random vector G and 7 diseases: diabetes, coronary diseases, renal diseases, stroke, cancer, Alzheimer disease, and Parkinson disease. We used linked Medicare data collected from 1989 to 2001 to obtain medical diagnoses for individuals in the sample.

We first clusterize hidden states using standard cluster analysis algorithm. We calculate the distance between cases as conventional Euclidean distance in 6-dimensional space, and distance between clusters as Euclidean distance between their centers of gravity. For demonstrational purposes, we chose a reduction to 6 clusters. For analysis with the number of clusters ranging from 5 to 12 there are no significant differences in the overall picture.

For every disease we calculate frequency of the disease in the sample and in the each cluster. The results are presented in figures 4 through 10. In these figures the average over the entire sample is plotted as a flat line. The means and standard deviations for each cluster are plotted as a point with associated confidence lines.

Table 3: Relative differences between calculated moments and frequencies (in standard deviations)

2.870	0.665	4.308	1.167	0.806	0.765
2.870	1.170	4.705	0.155	2.653	4.211
0.069	0.022	0.209	0.325	3.760	0.195
0.069	0.196	1.794	0.929	2.516	3.741
0.425	0.306	2.801	0.475	10.636	0.366
0.425	0.090	2.344	1.010	2.736	3.705
0.169	0.650	2.213	0.750	3.325	1.399
0.169	0.539	0.575	1.245	2.846	3.287
0.979	0.173	0.934	0.894	0.500	2.798
0.979	0.080	1.324	0.074	1.176	2.310
2.686	0.729	2.941	1.224	2.623	2.466

One can see that there exists a significant correlation between the clustered hidden state and diseases in all cases except cancer. Correlation is higher for diseases that contribute more to disability, as one would expect. Of course, there is no functional dependency here, as NLTCS questionary was not designed to diagnose these diseases.

4 Discussion

In this presentation we have discussed the implementation of the GoM model. As shown in [4], and verified here, the essence of a GoM analysis is representable in a system of equations (4) relating the observed unconditional moments with hidden conditional moments. Examination of this system of equations shows a parallel with principal component methods for continuous data. The difference lie in the fact that data in a GoM analysis are discrete with no ordinal scale required and, consequently, the "factorization" into the principal axes entails moments of order higher then two. Thus a GoM analysis will determine a coordinate system and a set of coordinates for each individual that explains the most "variation." Variation here is not determined by a least squares measure but rather by a quasilikelihood model given by (7).

In this paper we used approximation provided by the classical algorithm for maximization of the likelihood introduced in [10]. Although direct numerical methods of solving (4) or (7) are possible, such have not been worked out at this point.

The NLTCS data used to illustrate the methodology confirm our intuition. Quality of the constructed model was verified by two methods: numerical analysis of how close are columns of the frequency to the subspace predicted by the model, and how model moments deviate from the observed frequencies. Both tests demonstrate reliability and good predictive power of the model. Estima-

tion of the rank of the moment matrix (using singular value decomposition) allows us to determine the dimensionality of the GoM problem. This estimation provides strong statistical evidence in favor of dimensionality $K = 6$ used in other investigations of the NLTCS data.

The GoM analysis (the factorization into important components) split the individuals into into groups with very different disease outcomes. Recall, however, that the variables used here did not include variables commonly considered as risk factors for cancer.

From the practical point of view, the most important theoretical result is relation between the moment matrix and subspace that supports the distribution under question. It provides easy and vivid tests for hypothesis about dimensionality of distribution, and allows to justify results of other methods to obtain a basis of the support of the distribution.

The use of moment matrix also allows to handle missing data easily, as frequencies of response patterns may be calculated based only on a subset of individuals who gave all required answers.

Analysis of figures 2 and 3 (as well as more elaborated mathematical analysis) shows that vectors g_i, representing individual hidden state, occupy 5-dimensional body — rather than lower-dimensional manifold. This shows that linear regression hypothesis is suitable for the case of NLTCS data, and further reduction of dimensionality is not possible. If these hidden states occupy lower-dimensional manifold (for example, 2-dimensional sphere in 3-dimensional space), the further reduction of dimensionality is possible, but it requires to employ nonlinear regression hypothesis. The case of nonlinear regression is subject for the future work.

Another important feature of GoM analysis is that it converts discrete initial data into continuous state. It is especially useful when one investigates changes over time.

References

[1] Bartholomew, D.J., & Knott, M. (1999) *Latent Variable Models and Factor Analysis.* 2nd ed., London: Arnold; New York: Oxford University Press.

[2] 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.

[3] Heinen, T. (1996) *Latent class and discrete latent trait models: similarities and differences.* Thousand Oaks, Calif.: Sage Publications.

[4] Kovtun, M., Akushevich, I., Manton, K.G., & Tolley, H.D. (2003) *Grade of Membership Analysis: One Possible Approach to Foundations.* Submittet for publication to Annals of Statistics.

[5] Manton, K.G., Corder, L., & Stallard, E. (1993) *Changes in the Use of Personal Assistance and Special Equipment from 1982 to 1989: Results from the 1982 and 1989 NLTCS.* The Gerontologist **33**(2), 168–176.

[6] Manton, K.G., Stallard, E., & Corder, L. (1997) *Changes in the age dependence of mortality and disability: cohort and other determinants.* Demography, **34**(1), 135–157.

[7] Manton, K.G., Stallard, E., & Corder, L. (1998) *The dynamics of dimensions of age-related disability 1982 to 1994 in the U.S. elderly population.* Journal of Gerontology, **53A**(1) B59–B70.

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

[9] Wachter, K. (1999) *Grade of Membership Models in Low Dimensions: Geometry and Robustness.* Statistical Papers, **40**, 439-458.

[10] Woodbury, M., & Clive, J. (1974) *Clinical pure types as a fuzzy partition.* Journal of Cybernetics **4**, 111–121.

[11] Woodbury, M., Clive, J., & Garson, A. (1978) *Mathematical typology: a grade of membership technique for obtaining disease definition.* Computers and Biomedical Research, **11**, 277–298.

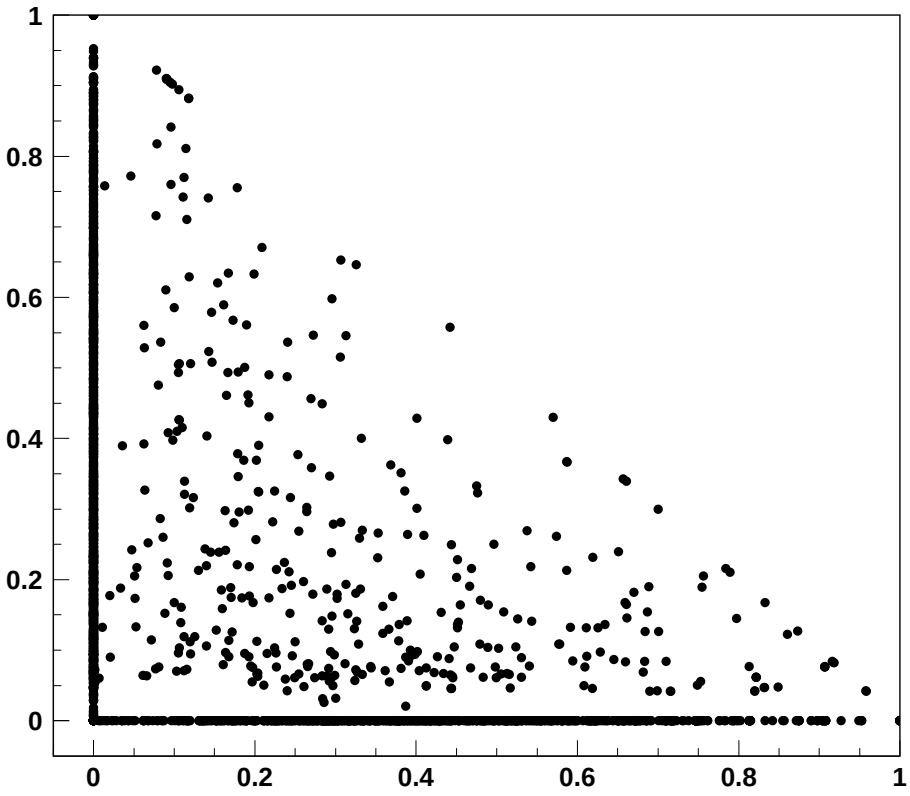

Figure 2: Distribution of hidden state: coordinates 2 and 6

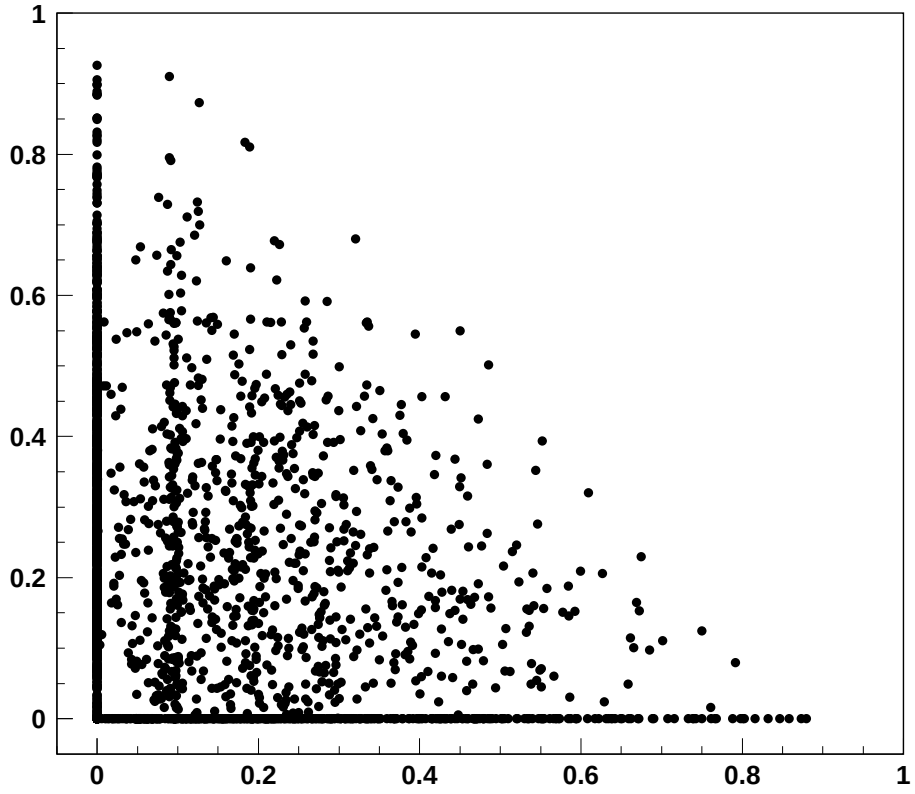

Figure 3: Distribution of hidden state: coordinates 4 and 5

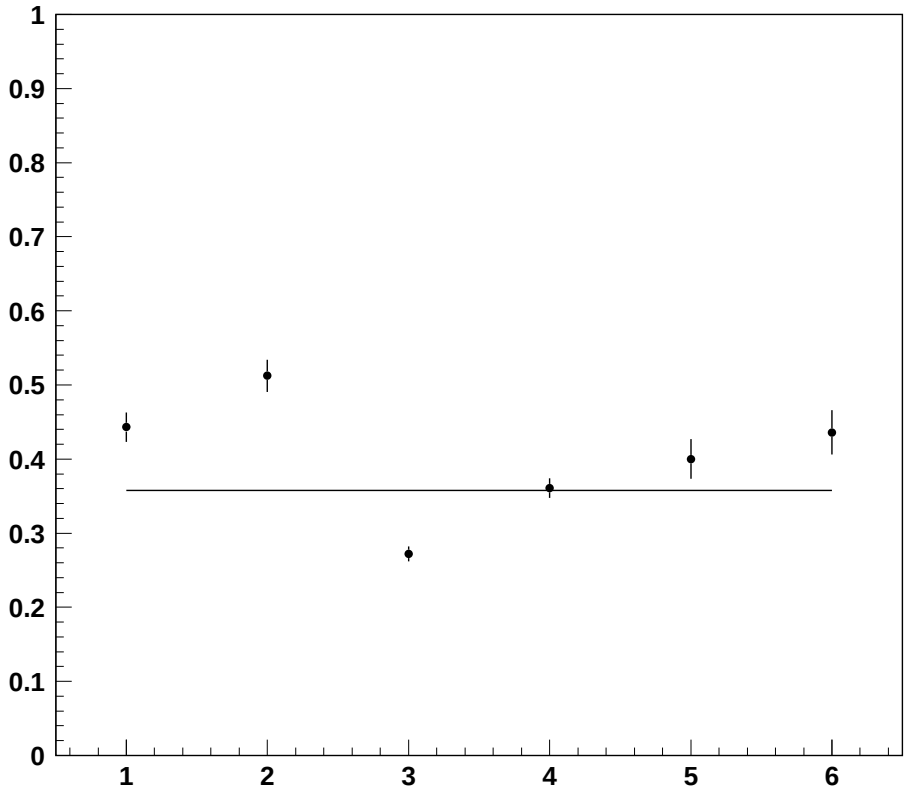

Figure 4: Frequencies of diabetes in the sample (horizontal line) and in the clusters

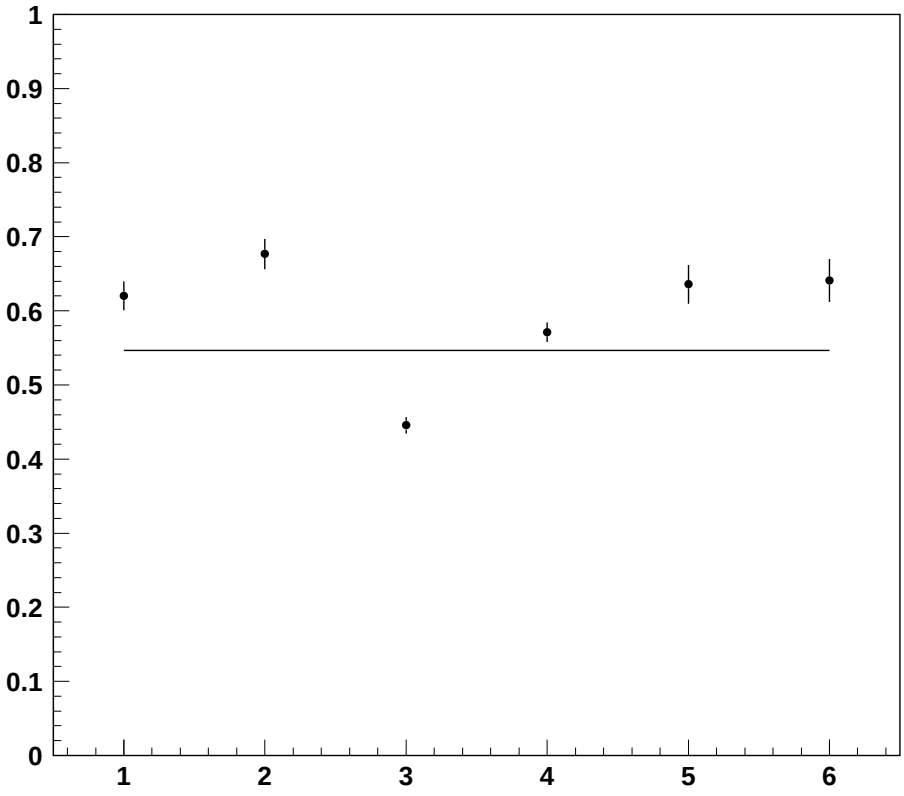

Figure 5: Frequencies of coronary diseases in the sample (horizontal line) and in the clusters

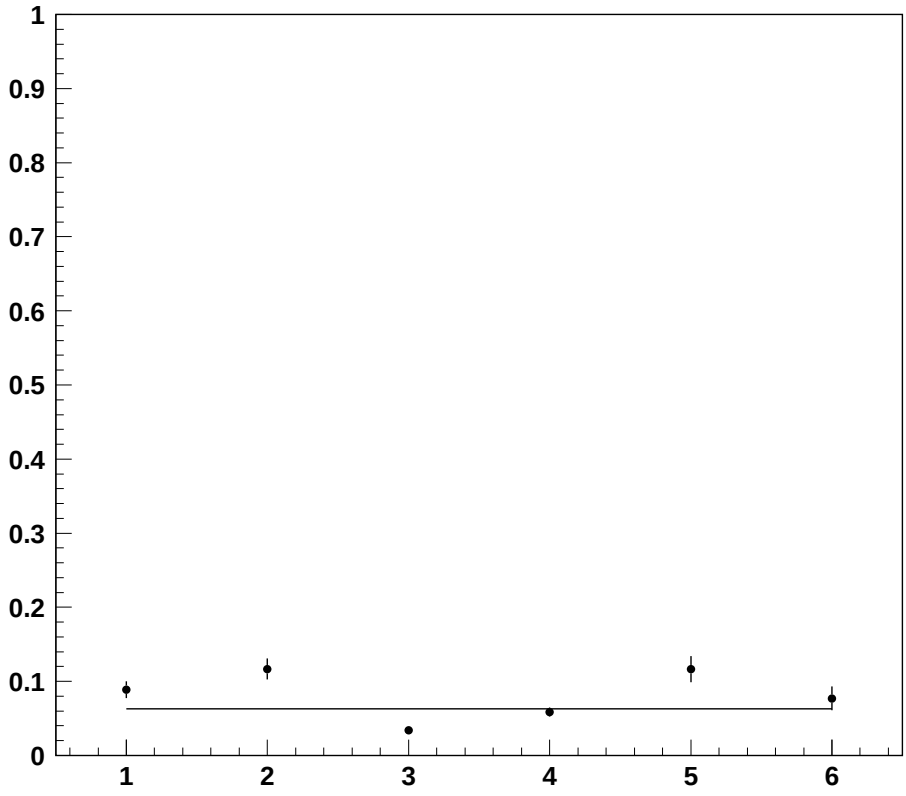

Figure 6: Frequencies of renal diseases in the sample (horizontal line) and in the clusters

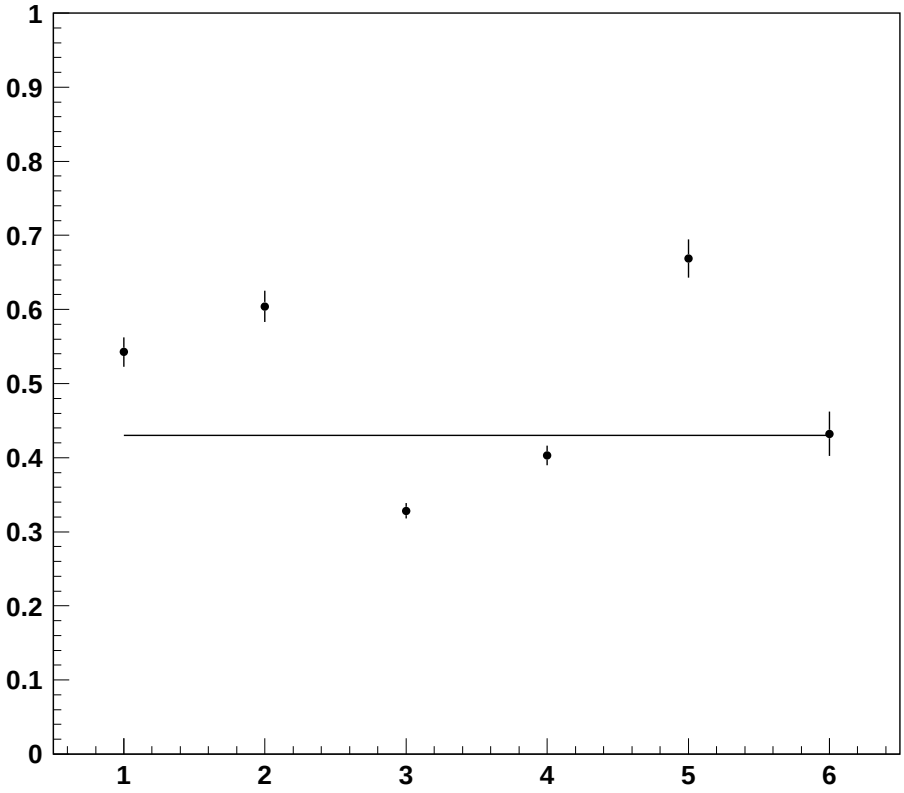

Figure 7: Frequencies of stroke in the sample (horizontal line) and in the clusters

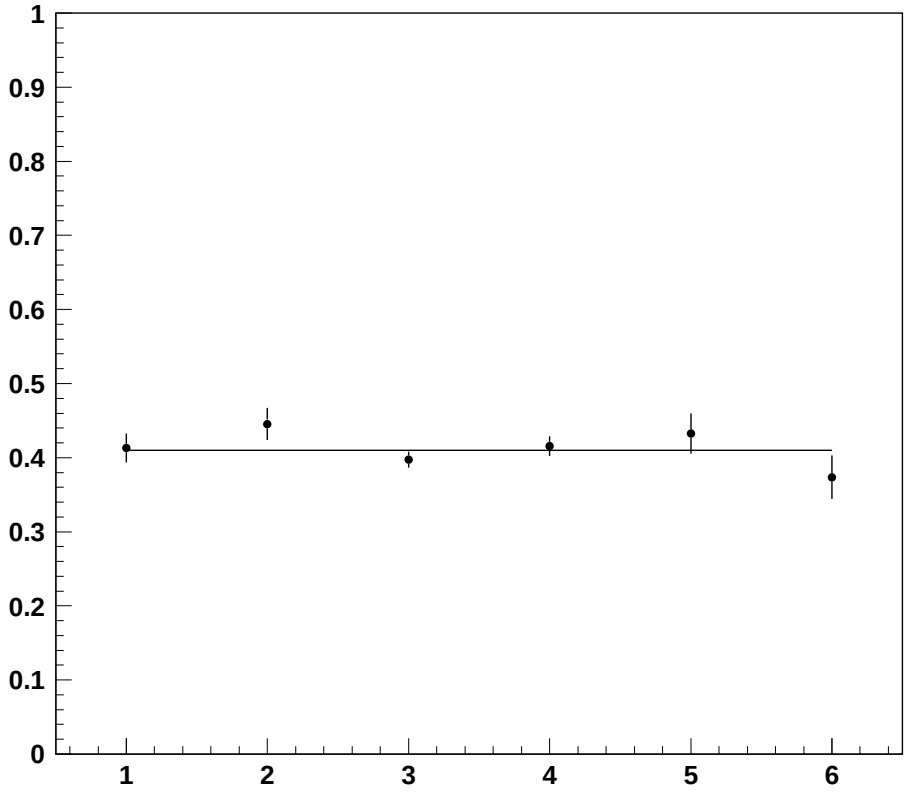

Figure 8: Frequencies of cancer in the sample (horizontal line) and in the clusters

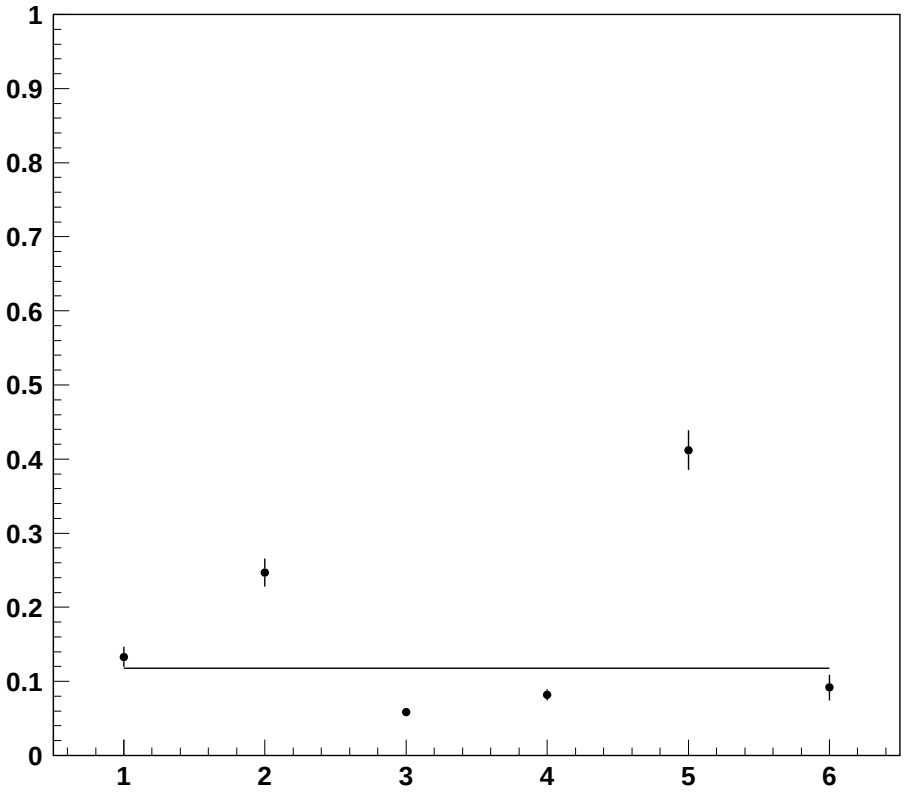

Figure 9: Frequencies of Alzheimer disease in the sample (horizontal line) and in the clusters

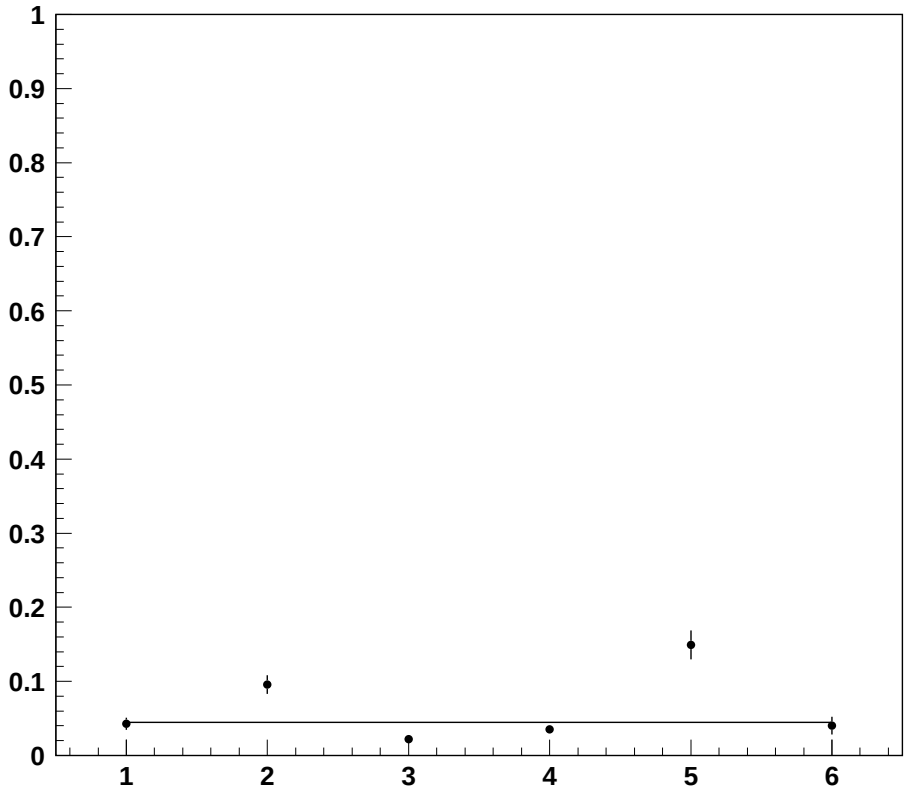

Figure 10: Frequencies of Parkinson disease in the sample (horizontal line) and in the clusters