# Model based analysis of incomplete data

Tamás Rudas
Center for Social Sciences, Hungarian Academy of Sciences and
Department of Statistics, Eötvös Loránd University, Budapest
rudas@tarki.hu

Emese Verdes
World Health Organization, Geneva
verdese@who.int

Juraj Medzihorsky
Department of International Relations, Central European University, Budapest
medzihorskyj@ceu.edu

December 16, 2016

**Abstract**

The mixture index of fit may be used to measure model fit in the presence of unit nonresponse by assuming the equality of two mixture representations of the true distribution of the population. One represents the observational process as a mixture of the distributions in the observable (actual data) and unobservable (missing data) parts of the population, and the other one represents the model assumption as a mixture of a distribution from the model (fit) and of an unspecified (no fit) component. In this framework, model fit may be measured by minimizing the no-fit fraction conditioned on an assumed or observed fraction of those unobserved. The paper generalizes this framework to item nonresponse by introducing additional components in the representation of the observational process for the various observed data patterns. This approach provides the researcher with a tool to diagnose the fit of statistical models using incomplete observations, without making untestable assumptions, like data being missing at random.

Keywords: missing data, mixture index of fit, model fit, model based analysis

# 1 Introduction

roblem in the methodology of survey research. While in the data collection phase, one tries to avoid having missing data, and while during the analysis of the data, various techniques have been

proposed and are routinely applied to try to make up for the missing information, in the phase of reporting the results of the survey, the incompleteness of the original data is often suppressed. In theory, one distinguishes several reasons of having incomplete data, including coverage error (the sampling procedure is unable to select from certain groups of the population), no-contact (the selected sampling unit is not found), unit nonresponse (the sampling unit is selected and contacted but does not want to participate in the survey), and item nonresponse (a participant of the survey does not answer a certain question).

Whatever the sources and reasons for having missing data might be, the fundamental question is whether or not conclusions based on the available data are as valid as those that could be obtained if all the desired data had been collected. Potential answers to this question include the following:

(a) The results based on the complete data would be different from those based on the incomplete data actually collected and it is best not to draw any conclusions from the results of the survey. This position is unlikely to be taken by anyone in the survey business.

(b) The results based on the complete data would be different from those that could be obtained from the incomplete data actually collected, but this does not seriously limit the validity of the latter results. For example, in a survey to predict the outcome of presidential elections, one may believe that those who decide not to participate in the survey or participate but do not tell their voting intentions, are also unlikely to cast their votes, and predictions based on those who did tell their preferences may well approximate the final results.

(c) The conclusions that could be obtained if complete data were available, may also be determined using the observed incomplete data. Under this assumption, various analyses of the observed data are performed to obtain the conclusions that could be obtained using the complete data. Most importantly, various analyses are performed to guess the missing data in case of item nonresponse, often taking the form of imputation, and then using the "completed" data set for the analysis.

(d) The conclusions that could be obtained if the complete data were available are the same as those based on the observed, incomplete data. Under this assumption, the incompleteness of the data may be disregarded.

It is not the goal of this paper to give a full review of the methods proposed under the above assumptions, nor is our goal to discuss the appropriateness of these methods or of the assumptions in various missing data situations above. We only note that assumptions (c) and (d) are often formulated as certain missing at random conditions (see Little and Rubin, 1989, but also Seaman, Galati, Jackson and Carlin, 2013 about the various interpretations of this concept). In spite of the great popularity of these assumptions, the fact that they may not be verified or tested using the incomplete data only, remains their fundamental weakness and, it appears, these assumptions are adopted by researchers not so much because they trust their validity but rather because of the convenience of the analysis that follows.

This paper offers an alternative approach to analyzing incomplete data from the perspective of the fit of a statistical model. The methodology which we outline here, provides the researcher with the opportunity to diagnose model fit or misfit with reference to the data collection procedure, including various kinds of missing data, even if the observed and unobserved data are not related in any known way. Our approach is based on a generalization of the mixture index of fit which was proposed by Rudas, Clogg, and Lindsay (1994) and subsequently investigated and used by a number of authors (see, e.g., Dayton, 2003; Deben and Garcia, 2009; Formann, 2000; Knott, 2005; Liu and Lindsay, 2009; Medzihorsky, 2015, Rudas, 1999, 2002). Rudas (2005) extended the index to handle missing data in the forms of coverage error, no contact, and unit nonresponse. This extension is summarized in section 2, and section 3 introduces the generalization of the mixture index of fit to incorporate item nonresponse.

The fundamental idea behind our approach is that an estimate is produced for the true pop-

ulation distribution through the equality of two mixture representations of it. One represents the fit of the model of interest, with one component for the part of the population which is described by a distribution belonging to the model, and another one describing the part where the model is not true. The other mixture represents the observational process, with a component for those whose data were observed and further components for those whose data are missing either entirely (coverage error, no contact, unit nonresponse) or partially (various patterns of item nonresponse).

This approach has, at least, three main advantages. First, it is always correct that is, the analysis is not based on the assumption that the true distribution belongs to the hypothesis of interest. Second, it is also always correct in the sense that there is no untestable assumption made about the missing data mechanism. Third, the findings have an easy and intuitive interpretation, much simpler than those relying on achieved p-values. The method is illustrated using data from the International Social Survey Project (see GESIS, 2009).

Diagnostics of the statistical model of interest is obtained by assessing the no-fit proportions under various assumptions regarding the relative size of the unobserved component, and the estimated distributions in the not entirely observed components under these assumptions. From a substantive point of view, the estimated distributions in the various parts of the population may be interpreted similarly to the procedures described by Clogg, Rudas, and Matthews (1997), and judging the acceptability of these estimated distriburtons from a substantive point of view is highy recommended as part of the diagnostic procedure.

As the number of variables increases, there may be very many components with differing observed data patterns and the analysis may be simplified by assuming that the true distributions in some of them are identical.

Finally, Section 4 describes the computational procedures applied.

# 2 The mixture index of fit

The fundamental idea behind the mixture index of fit is that a statistical model $M$ may also give a relevant description of the population of interest if it only applies to a fraction of it, provided this fraction is large enough. In this framework, the true distribution on the population is a mixture of two components, one describing the part of the population where a distribution from the model is true and another one describing the part of the population where the model is not true. The mixture may be written in the form of

$$P = (1 - \pi)F + \pi E, \tag{1}$$

where $F$ is a distribution belonging to the model $M$ (the fit distribution), $E$ is an unrestricted distribution (the error distribution), and $(1 - \pi)$ is the fit rate. This representation is always true for some value of $\pi$. The smaller is $\pi$, the more relevant is $M$ for the population. The mixture index of fit is $1-\pi$ for the smallest possible value of $\pi$, such that representation (1) is possible.The mixture index of fit was first proposed by Rudas et al. (1994) and they gave methods for the estimation of the smallest value of $\pi$ and for the determination of confidence intervals for its true value. Xi and Lindsay (1996), Verdes (2000), and Dayton (2003) proposed improvements of the estimation techniques, and Formann (2006), Medzihorsky (2015), Hernandez, Rubio, Revuelta, and Santacreu (2006), Revuelta (2008), and Verdes and Rudas (2003) discussed applications to various choices of $M$. The application of the mixture index of fit is particularly attractive in the case of small sample sizes, when chi-squared based asymptotic testing is inappropriate, and the related properties were investigated by Formann (2000) and by Pan and Dayton (2006). Further, the minimal $\pi$, denoted as $\pi^*$, is a population quantity estimated form data and it avoids the problems associated with

using the value of the chi-squared statistics as a measure of model misfit. The mixture index of fit is also applicable when population data are available.

The framework on which the mixture index of fit relies, also provides a new way of looking at residuals. In the classical setup, if the hypothesis that the model of interest describes the population is rejected, then the residuals refer to a model which is deemed to be not relevant. If the hypothesis is not rejected, the residuals are deemed to have been due to sampling variation, and their analysis is even more questionable. In the case of the mixture index of fit, the residuals given by $E$ describe the part of the population where the model doe snot hold. For ways to analyze these residuals, see Clogg, et al. (1997).

The mixture index of fit was also applied to continuous data, including the multivariate normal distribution (Knott, 2005) and minimax regression (Rudas, 1999)

Rudas (2005) proposed an extension of the mixture index of fit to certain incomplete data situations. Coverage error, no contact, and unit nonresponse may all be modeled by assuming that a certain part of the population of interest is not available to participate in the survey. Data may only be collected from those not in this part, that is, those who are available to participate in the actual survey, and the estimates based on the observed data refer to the distribution in the part of the population which is observable. From the perspective of the observational process, the true distribution which characterizes the population may be written as a mixture, with an observable ($O$) and an unobservable ($U$) component:

$$P = (1 - \rho)O + \rho U, \tag{2}$$

where $\rho$ is the relative size of the unobservable fraction of the population. Equating (1) and (2), that is, model fit with the observational process, yields

$$(1 - \pi)F + \pi E = (1 - \rho)O + \rho U. \tag{3}$$

In (3), a small $\pi$ means that the model of interest may describe a large fraction of the population, while a small $\rho$ means that this may be said based on the observation of a large fraction of the population. Xi (1996) considered (3), although with an interpretation not related to missing data, from the perspective of simultaneously minimizing $\pi$ and $\rho$. In the missing data framework, the smallest value of $\pi$ with which a representation of the form (3) is possible for any given $\rho$, denoted as $\pi(\rho)$, is the quantity of interest. A choice of $\rho$ represents the researchers knowledge or beliefs with respect to the observational process and $\pi(\rho)$ gives the smallest possible misfit rate of the model of interest which is implied. Diagnostic uses of $\pi(\rho)$ were described in Rudas (2005), together with computational tools to estimate its value. Even if very little is known about the likely size of $\rho$, the need to explicitly consider a realistic range for it, emphasizes the inherent weakness of the analysis which would be present, even if the researcher was not forced to think about the severity of the missing data issue.

The generalization to item nonresponse to be discussed in this paper will be illustrated using the same set of data which was used to illustrate the computation and use of $\pi(\rho)$, data from the International Social Survey Programme (see GESIS, 2009).

# 3 Generalization to item nonresponse

As discussed earlier, a crucial question in the analysis of incomplete data is whether or not the missing data, or at least certain features of it, may be revealed by an analysis of the observed data. When the answer to this question is positive, certain analyses based on the incomplete data may

lead to results similar or identical to the analyses that could be based on the complete data, if they were available. When the answer is negative, then the analyses based on the incomplete data may yield results very different from those that could be obtained from the complete data. The model based analyses described in this paper may be used in those very common situations when the answer is not known or is known to be negative. The analysis based on the mixture (3) allows the distribution of the unobserved component to be similar to that of the observed component but also allows it to be quite different. The key idea in modifying the framework to include item nonresponse is to allow groups of respondents answering to different subsets of the questions to have different true (and only partially observed) distributions. For example, if only two questions, $A$ and $B$, are asked, the observed data patterns may be $AB$, $A$ only, $B$ only, and neither $A$ nor $B$. Note that the patterns here depend on which questions were answered and not on what the actual answers were or would have been. Accordingly, the population is assumed to consist of the following components:

– those available to answer and answering both $A$ and $B$;
– those available to answer but answering only $A$;
– those available to answer but answering only $B$; and
– those not available to answer or available but not answering either question.

The reason for combining those not available to participate in the survey and those available but not answering any of the questions is that no observed data are available for either group and, therefore, these groups cannot be distinguished, if there are only 2 variables. In cases, when the questionnaire contains more than 2 variables, these groups may need to be distinguished. The methods described here, do extend to those cases. Accordingly, instead of (2), the following mixture representation is considered.

$$P = \rho_1 O_{AB} + \rho_2 O_A + \rho_3 O_B + \rho U, \tag{4}$$

where $O_{AB}$ is the distribution in the component where both $A$ and $B$ may be observed, $O_A$ and $O_B$ are the distributions in the partially observable components, and $U$ is the distribution in the component that may not be observed. $O_{AB}$ may be estimated from the observed data, and so are the $A$-marginal distribution of $O_A$ and the $B$-marginal distribution of $O_B$. The relative fraction sizes $\rho_1, \rho_2, \rho_3$ and $\rho$ sum to 1. Further,

$$\rho_i/(\rho_1 + \rho_2 + \rho_3) = \rho_i/(1 - \rho), i = 1, 2, 3$$

may also be estimated from the data, as the relative fractions of the different response patterns. Also, $\rho$ is the fraction of those available but not answering any of the questions or not available, so the observed fraction of those not answering any of the questions is a lower bound for $\rho$.

Next, (4) and (1) may be equated,

$$\rho_1 O_{AB} + \rho_2 O_A + \rho_3 O_B + \rho U = (1 - \pi)F + \pi E, \tag{5}$$

and the fit (in the model) $F$, the no fit (not in the model) $E$, the entirely unobserved $U$ distributions and the unobserved components of the $O_A$ and of $O_B$ distributions may be estimated to simultaneously minimize $\pi$ and $\rho$, that is, the no-fit and the unobserved fractions. For $O_A$ and $O_B$, this involves estimating the conditional distributions $B|A$ and $A|B$, respectively, where the $A$ marginal of $O_A$ and the $B$ marginal of $O_B$ are taken (estimated) from the observed data. Depending on the structure of the missing data patterns and the model of interest, the simultaneous minimization of $\pi$ and $\rho$ may not be feasible, and a procedure where the analyst assumes a no-observation fraction $\rho$ and sees how small the no-fit fraction $\pi$ may be, can be applied.

The latter procedure is illustrated now using two variables from the 1995 ISSP survey data (N=1367). Question $A$ is 'How proud are you with the way democracy works in your country?'

and question $B$ is 'How proud are you with achievements in sport of your country?' (briefly referred to as Democracy and Sport, respectively). In our analysis, the 'can't chose', 'don't know' and 'refused to answer' categories are combined into the missing category.

The model of interest in our analysis is the independence of the two variables. First, we give the estimated representation according to the left hand side of (5) in Table 1, and then the estimated representation according to the right hand side of (5) in Table 2, with the assumption of $\rho = 0.1$.

Table 1
Mixture representation of the ISSP data with item nonresponse,
as on the left hand side of (5) with $\rho = 0.1$ (probabilities multiplied by 100)

$0.830 \times$

| Both Democracy and Sport observed | | | | | |
|---|---|---|---|---|---|
| | Sport | very proud | somewhat proud | not very proud | not proud at all |
| | Democracy | | | | |
| | very proud | 17.23 | 9.43 | 1.89 | 0.57 |
| | somewhat proud | 16.49 | 31.91 | 4.35 | 1.72 |
| | not very proud | 3.53 | 7.71 | 1.72 | 0.90 |
| | not proud at all | 0.57 | 1.31 | 0.33 | 0.33 |

$+0.039 \times$

| Only Democracy observed | | | | | |
|---|---|---|---|---|---|
| | Sport | very proud | somewhat proud | not very proud | not proud at all |
| | Democracy | | | | |
| | very proud | 0.00 | 0.14 | 8.94 | 15.05 |
| | somewhat proud | 45.45 | 0.01 | 0.15 | 6.11 |
| | not very proud | 12.85 | 5.99 | 0.00 | 0.12 |
| | not proud at all | 0.47 | 4.55 | 0.06 | 0.10 |

$+0.031 \times$

| Only Sport observed | | | | | |
|---|---|---|---|---|---|
| | Sport | very proud | somewhat proud | not very proud | not proud at all |
| | Democracy | | | | |
| | very proud | 0.00 | 51.07 | 3.56 | 0.35 |
| | somewhat proud | 7.28 | 0.00 | 0.45 | 1.80 |
| | not very proud | 18.68 | 5.73 | 0.14 | 0.02 |
| | not proud at all | 7.37 | 3.19 | 0.29 | 0.05 |

| | Unobserved | | | | | |
|---|---|---|---|---|---|---|
| | | Sport | very proud | somewhat proud | not very proud | not proud at all |
| $+0.100 \times$ | | Democracy | | | | |
| | | very proud | 0.01 | 74.83 | 3.06 | 0.20 |
| | | somewhat proud | 11.01 | 0.13 | 0.30 | 0.00 |
| | | not very proud | 5.00 | 3.07 | 0.22 | 0.00 |
| | | not proud at all | 1.70 | 0.42 | 0.04 | 0.02 |

Table 2
Mixture representation of the ISSP data with independent and unrestricted components, as on the right hand side of (5) (probabilities multiplied by 100)

| | Fit | | | | | |
|---|---|---|---|---|---|---|
| | | Sport | very proud | somewhat proud | not very proud | not proud at all |
| $+0.954 \times$ | | Democracy | | | | |
| | | very proud | 11.24 | 17.71 | 2.45 | 1.15 |
| | | somewhat proud | 17.62 | 27.77 | 3.84 | 1.81 |
| | | not very proud | 4.73 | 7.45 | 1.03 | 0.49 |
| | | not proud at all | 0.94 | 1.47 | 0.20 | 0.10 |

| | No Fit | | | | | |
|---|---|---|---|---|---|---|
| | | Sport | very proud | somewhat proud | not very proud | not proud at all |
| $+0.046 \times$ | | Democracy | | | | |
| | | very proud | 77.11 | 0.00 | 0.00 | 0.00 |
| | | somewhat proud | 0.00 | 0.15 | 0.01 | 0.00 |
| | | not very proud | 0.00 | 0.16 | 10.21 | 6.26 |
| | | not proud at all | 0.00 | 0.05 | 2.00 | 4.06 |

The operations with the tables are to be interpreted for every cell. For example in Table 1, in cell $(1,1)$, we estimate a conditional probability of 0.1723 in the part where both variables are observable, and the relative size (or probability) of this part is 0.830; further, in the component where only Democracy may be observed, the conditional probability of this cell is 0.00 and the relative size of this part is 0.039, and so forth. In total, the probability in cell $(1,1)$ is

$$0.830 \times 0.1723 + 0.039 \times 0.00 + 0.031 \times 0.00 + 0.100 \times 0.0001 = 0.143$$

and from Table 2 it is

$$0.954 \times 0.1124 + 0.046 \times 0.7711 = 0.143.$$

The interpretation of the estimates given is, that assuming the total unobserved fraction is 10%, independence may account for over 95% of the population. The estimates of the various distributions given are model-based in the sense that they are generated to reduce no-fit (i.e., to maximize fit) of independence of the two variables considered.

While among those who responded to both questions, the combination of the somewhat proud and the very proud categories were the most populous, accounting for about 75% of the respective population, in the case of those who only told their opinion about democracy, the estimated opinions about sport are even more positive, with nearly 60% being in the very proud category. Among those who told their opinion only about sport, nearly 60% of the estimated opinions about democracy appear in the very proud or somewhat proud categories. We estimate that those who were not observed at all, are typically very proud of democracy and somewhat proud of sport. On the other hand, where independence holds, the estimated distribution is quite similar to the observed distribution, and the distribution of those for whom independence does not hold is strongly concentrated on the main diagonal, with being very proud in both aspects accounting for over 77% of this population.

The decision about the appropriateness of the model of independence in this approach depends on a number substantive considerations. Is over 95% a large enough fraction? Are the estimates of the partially observed distributions reasonable? Is the assumption of 10% unobserved fraction realistic? There are, usually, no clear responses to these questions and, therefore, the assessment of the situation requires the judgement of the researcher. The traditional analysis, on the other hand, based on the untestable and surely oversimplifying assumptions of missing at random or missing completely at random, and asking whether or not independence may describe the entire population, appears to be a well automated procedure and gives the fake feeling of certainty to the researcher. We believe, that based on a sample, and, even more, based on an incompletely observed sample, such an automated response should be avoided and researchers need to be confronted with the inherent uncertainty of the situation. The final decision in the analysis, which, seen as part of the process of knowledge acquisition, is always tentative, should incorporate substantive considerations and researchers should not be encouraged to delegate the responsibility to an automated mechanism based on overly simple assumptions.

When there are $k$ variables, there may be $2^k$ different observed data patterns, and reliable estimation under the assumption that in all of these the true distributions may be different, may not be feasible for realistic sample sizes. In such cases, one may assume that in some of these components, the true distributions are identical.

For example, let there be 3 variables, $A$, $B$ and $C$. One interesting assumption, which largely reduces the computational burden is that the population distribution is the same in all components, where $A$ is observable and is also the same (though possibly different from the previous one), where $A$ cannot be observed. The distributions on the components where $A$ was observed are

$$O_A, O_{AB}, O_{AC}, O_{ABC}$$

and the distributions in all these components are assumed to be the same. Similarly, the distributions in the components where $A$ was not observed

$$O_B, O_C, O_{BC}$$

are assumed to be the same. Under this model, the approach utilized so far, namely preserving the observed marginal distributions is not feasible, as, for instance, the observed $B$ marginal distributions in $O_B$ and $O_{BC}$ may be different.

In this case, one may proceed by estimating the common distributions in the components where $A$ was observed, say $O_{A^+BC}$, and the common distribution in the components where $A$ was not observed, say $O_{A^-BC}$, to obtain a mixture representations as

$$\rho_{A^+} O_{A^+BC} + \rho_{A^-} O_{A^-BC} + \rho U = (1-\pi)F + \pi E, \tag{6}$$

where $\rho_{A^+}$ is the fraction, where $A$ was observed, and $\rho_{A^-}$ is the fraction where $A$ was not observed. Then, the value of $\pi^*$ may be estimated.

The diagnostic use of this approach is similar to the one described above. If the value of $\pi^*$ is deemed satisfactory, the estimated distributions $O_{A^+BC}$ and $O_{A^-BC}$ may be inspected from a substantive point of view, and if they do not appear realistic, the assumption made with respect to the equality of distributions in the various components may be changed. For example, a less restrictive set of assumptions is that whether or not $A$ was observed is irrelevant for the joint distribution, implying that

$$O_B = O_{AB},\ O_C = O_{AC},\ O_{BC} = O_{ABC}.$$

# 4 Computational procedures

The example in Section 3 was calculated in two different ways, the second one giving slightly better results.

The problem may be formulated as a constrained optimization problem and may be solved using commercially available statistical software. We used the Optimization Toolbox in MATLAB 6.5 (MathWorks, 2003), namely the 'fmincon' function. The parameter values to be estimated were the 12 free parameters of the only Democracy was observed table, the 12 free parameters of the only Sport was observed table, the 15 parameters of the unobserved table and the 6 parameters of the independent table. The constraints ensured that the no fit table, which was derived from 5, had nonnegative entries. The quantity to be maximized was the total of the $(1 - \pi)F$ table. This procedure resulted in a $\pi^*$ value of 0.057.

A more involved algorithm yielded the solution given in Tables 1 and 2, with a slight improvement of $\pi^*$= 0.046. This procedure was based on a function, which had the 39 parameters on the left hand side of (5) as variables and yielded $\pi^*$ for independence. This function was based on the implementation of the estimation of $\pi^*$ for log-linear models in the pistar package (Medzihorsky, 2013). The package uses the EM algorithm to fit the decomposition for a fixed $\pi$ and the Brent algorithm to find $\pi^*$. Finally, the DEoptim optimizer (Mullen, 2011) was used to find such values of the 39 parameters that yield the lowest $\pi^*$ value. The DEoptim optimizer implements the Differential Evolution algorithm for global optimization, which does not require the function to be optimized to be either differentiable or continuous. The optimizer declares that it found the optimum, if after a user-defined number of steps the value to be optimized does not reduce by a factor equal to a tolerance value multiplied by the sum of the value to be optimized plus the tolerance. The tolerance was set to $10^{-6}$ and the number of minimum steps to 100. A parallelized version of the algorithm was used, leaving the remaining options at their defaults. The algorithm declared convergence after slightly over 3000 iterations.

The approach leading to (6) requires the estimation of joint distributions based on incomplete data. For $O_{A^+BC}$, one may combine the data for the distribution of $A$ from $O_A$, $O_{AB}$, $O_{AC}$, $O_{ABC}$. The distribution of $B$ is obtained by combining data from $O_{AB}$ and $O_{ABC}$. For $C$, one can combine data from $O_{AC}$ and $O_{ABC}$. Once the univariate distribution are estimated, one proceeds to estimate the bivariate distributions. For example, to obtain the $AB$ marginal distribution, one combines the $O_{AB}$ observed distribution with the $AB$ observed marginal of the $O_{ABC}$ distribution to estimate the odds ratio between $A$ and $B$, and uses this, together with the already estimated univariate marginal distribution to determine the $AB$ marginal, using the Iterative Proportional Fitting Procedure. For details, see Rudas (1998). By the variation independence of the univariate marginal distributions and the odds ratio, this procedure always yields a bivariate distribution. Higher order distributions may be constructed in a similar way, but, in general however, there is no

guarantee, that for any system of marginal distributions there will exist a joint distribution. The conditions, under which this will happen for any data, are described in Bergsma and Rudas (2002). When a joint distribution cannot be constructed, the assumed equalities among the distributions in the various components are not tenable and need to be changed.

# 5 Acknowledgements

The first author was supported in part by Grant No TAMOP 4.2.1./B-09/1/KMR-2010-0003 from the European Union and the European Social Fund and by Grant K-106154 from the Hungarian National Scientic Research Fund (OTKA). He is also a recurrent visiting professor at the Central European University, Budapest, and the moral support received is acknowledged.

# 6 References

Bergsma, W., Rudas, T. (2002) Marginal models for categorical data. Annals of Statistics, 30, 140-159

Clogg, C. C., Rudas, T., Matthews, S. (1997). Analysis of model misfit, structure, and local structure in contingency tables using graphical displays based on the mixture index of fit. In J. Blasius M. Greenacre (Eds.), Visualization of categorical data (pp. 425-439). New York: Academic Press.

Dayton, C. M. (2003). Applications and computational strategies for the two-point mixture index of fit. British Journal of Mathematical and Statistical Psychology, 56, 1-13.

Deben, A. G., Garcia, J. E. S. (2009) The study of cells in the analysis of contingency tables from the viewpoint of Rudas, Clogg and Lindsay mixture index of fit. Revista Investigacion Operational 30, 244-249

Formann, A. K. (2000). Rater agreement and the generalized Rudas-Clogg-Lindsay index of fit. Statistics in Medicine, 19, 1881-1888. Formann, A. K. (2006). Testing the Rasch model by means of the mixture fit index. British Journal of Mathematical and Statistical Psychology, 59, 86-95.

GESIS (2009). ISSP 1995 [Data file and codebook]. Retrieved from http://www.issp.org/data.shtml

Hernandez, J. M., Rubio, V. J., Revuelta, J., Santacreu, J. (2006). A procedure for estimating intrasubject behavior consistency. Educational and Psychological Measurement, 66, 417-434.

Knott, M. (2005). A measure of independence for a multivariate normal distribution and some connections with factor analysis. Journal of Multivariate Analysis, 96, 374-383.

Little, R., Rubin, D. (1989). Statistical analysis with missing data. New York: Wiley.

MATLAB (Version 6.5, 2003) [Computer software]. Natick, MA: MathWorks, Inc.

Medzihorsky, J. (2013) An R package for the Rudas-Clogg-Lindsay mixture index of fit https://github.com/jmedzihorsky/pistar.

Medzihorsky, J. (2015) Election fraud: A latent class framework for digit based tests. Political Analysis, 23, 506-517.

Mullen, K., Ardia, D., Gil, D., Windover, D., Cline, J. (2011). 'DEoptim': An R Package for Global Optimization by Differential Evolution. Journal of Statistical Software, 40, 1-26.

Liu, J., Lindsay, B. G. (2009) Building and using semiparametric tolerance regions for parametric multinomial models. Annals of Statistics, 37, 3644-3659.

Pan, X., Dayton, C. M. (2006). Factors influencing the mixture index of model fit in contingency tables showing independence. Unpublished manuscript. Department of Measurement, Statistics and Evaluation, University of Maryland, Retrieved from http://www.education.umd.edu/EDMS/fac/Dayton/PiStar_Pan_Dayton.pdf

Rudas, T. (1998) Odds Ratios in the Analysis of Contingency Tables. Sage.

Rudas, T. (1999) The mixture index of fit and minimax regression. Metrika, 50, 163-172.

Rudas, T. (2002). A latent class approach to measuring the fit of a statistical model. In J. Hagenaars, A. McCutcheon (Eds.), Applied latent class analysis (pp. 425-439). Cambridge, UK: Cambridge University Press.

Rudas, T. (2005). Mixture models of missing data. Quality & Quantity, 39, 19-36.

Rudas, T., Clogg, C. C., Lindsay, B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. Journal of the Royal Statistical Society, Ser. B, 56, 623-639.

Revuelta, J. (2008). Estimating the $\pi^*$ goodness of fit index for finite mixtures of item response models. British Journal of Mathematical and Statistical Psychology, 61, 93-113.

Seaman, S., Galati, J., Jackson, D., Carlin, J. (2013) What is meant by "missing at random?". Statistical Science, 28, 257-268.

Verdes, E. (2000). Finding and characterization of local optima in the $\pi^*$ for two-way contingency tables. Studia Scientiarium Mathematicarum Hungarica, 36, 471-480.

Verdes, E., Rudas, T. (2003). The $\pi^*$ index as a new alternative for assessing goodness of fit of logistic regression. In Y. Haitovsky, H. R. Lerche, Y. Ritov (Eds.), Foundations of statistical inference (pp. 167-189). New York: Springer.

Xi, L. (1996) Measuring goodness-of-fit in the analysis of contingency tables with mixture-based indices: Algorithms, asymptotics and inference. Ph.D. Dissertation, Pennsylvania State Univ.

Xi, L., Lindsay, B. G. (1996). A note on calculating the $\pi^*$ index of fit for the analysis of contingency tables. Sociological Methods & Research, 25, 248-259.