The π^* index: computation, characterization and application of a new goodness of fit measure

doktori (PhD) értekezés

Verdes Emese

Debreceni Egyetem

Debrecen, 2002

Ezen értekezést a Debreceni Egyetem Matematika doktori program Valószínűségelmélet és matematikai statisztika alprogramja keretében készítettem 1995–2002. között és ezúton benyújtom a Debreceni Egyetem doktori Ph.D. fokozatának elnyerése céljából.

Debrecen, 2002. január

.....
Verdes Emese
jelölt

Tanúsítom, hogy Verdes Emese doktorjelölt 1995–2002. között a fent megnevezett doktori alprogram keretében irányításommal végezte munkáját. Az értekezésben foglaltak a jelölt önálló munkáján alapulnak, az eredményekhez önálló alkotó tevékenységével meghatározóan hozzájárult. Az értekezés elfogadását javaslom.

Debrecen, 2002. január

.....
Dr. Arató Mátyás
témavezető

Acknowledgement

Here I would like to thank all the people who have contributed to my dissertation.

First of all, to the members of the π^* group, Dr. Gábor Tusnády, Dr. Imre Csiszár, Dr. György Michaletzky, Dr. Tamás Rudas and Dr. Márton Ispány for their helpful instructions and useful advice.

To Dr. Mátyás Arató and Dr. Gyula Maksa encouraging me.

Finally, to Dr. Sándor Baran helping me in editing this thesis.

Köszönetnyilvánítás

Itt szeretnék köszönetet mondani mindazoknak, akik hozzájárultak a disszertációm elkészítéséhez.

Először is a π^* csoport tagjainak, Dr. Tusnády Gábornak, Dr. Csiszár Imrének, Dr. Michaletzky Györgynek, Dr. Rudas Tamásnak és Dr. Ispány Mártonnak segítőkész útmutatásaikért és hasznos tanácsaikért.

Dr. Arató Mátyásnak és Dr. Maksa Gyulának, hogy bátorítottak.

Végül pedig Dr. Baran Sándornak, hogy segített a disszertációm szerkesztésében.

Contents

General notations

n	sample size
n_i	observed counts
m_i	expected counts
N	number of cells of a contingency table or cardinality of a finite sample space
$(\Omega, \mathcal{A}, \mathbf{P})$	statistical space
$\mathbf{P}$, $\mathbf{M}$	collection of probability measures
P, M, R	probability measures
p, m, r	probability densities
P_n	empirical measure
p_n	empirical density
$\theta \in \Theta$	model parameter
θ_n	estimated model parameter
X, Y, Z	random variables or vectors of n observations
D	design matrix
k	number of groups (when grouping the data)
k, l	number of rows and columns
ϕ, ψ	row and column marginals
$f : \mathbb{R}^d \to \mathbb{R}$	objective function
$g : \mathbb{R}^d \to \mathbb{R}^m$	constraint function
$N(\mathbf{M}, \pi)$	contamination neighbourhood

Preface

In sociology, most of the variables being involved in a statistical analysis are categorical, so contingency table analysis: modeling and model verification has received a great deal of attention in recent decades. The emphasis of this thesis is model verification, namely, a new goodness of fit approach, called the π^* index.

The main objective is to propose algorithms for the computation of this index. A second goal is to analyze theoretically the problem of computing it and answering some theoretical questions relating π^*, e.g. robustness. Finally, a wide scale of applicabilities is presented through numerical examples taken from different related disciplines.

Chapter 1

Introduction and preliminary results

1.1 Literature review

One of the first goodness of fit tests was introduced by Pearson [36], the so called Pearson χ^2 statistic. Another traditional goodness of fit statistic is the likelihood ratio χ^2 statistic. Various other goodness of fit statistics have been proposed. Other measures of fit include the Freeman-Tukey statistic, which is defined as

$$F^2 = 4\sum_i \left(\sqrt{n_i} - \sqrt{m_i}\right)^2;$$

the modified loglikelihood ratio statistic or minimum discrimination information statistic (Kullback, 1959),

$$ML = 2\sum_i m_i \log(m_i/n_i);$$

and the Neyman χ^2 statistic

$$NC = \sum_i \frac{(n_i - m_i)^2}{n_i},$$

n_i-s and m_i-s denoting the observed and expected counts, respectively. All of these statistics have asymptotic χ^2 distribution.

Cressie and Read [40] proposed a unified analysis using a power divergence family of statistics given by

$$PD(\lambda) = \frac{2}{\lambda(1+\lambda)}\sum_i n_i \left(\left(\frac{n_i}{m_i}\right)^{\lambda} - 1\right),$$

where λ is a real valued parameter. This statistic makes a link among traditional test statistics through the parameter λ. Choosing λ to be 1 we obtain the Pearson χ^2 statistic, however if λ tends to 0, the limit is the loglikehood statistic. The other goodness of fit statistics mentioned earlier are also special cases of the power divergence statistic family: choosing $\lambda = -2, -1$ and $-1/2$, NC, ML and F^2 are obtained, respectively.

One common thing of these goodness of fit statistics is that they all have an asymptotic χ^2 distribution, however we have a bad approximation of this distribution if the sample size is too small, and on the other hand, if the sample size is too big, we always tend to reject the null hypothesis. Moreover the χ^2 value is not informative in this second case due to the fact that the χ^2 value is proportional to the sample size. This second problem is illustrated by the following example.

Consider the following table of eye color and hair color (Snee, 1974; Diaconis and Efron, 1985), where the sample size is 592. The test of independence gives a Pearson χ^2 statistic $\chi^2 = 138.29$ and a liklelihood ratio statistic $L^2 = 146.44$ with degrees of freedom $df = 9$. These χ^2 statistics lead to the rejection of independence on every usual significance level. On the other hand, Diaconis and Efron [13] showed that about 10% of all possible 4×4 tables with sample size 592 have $\chi^2 \leq 138.92$.

Table 1.1: Cross–classification of eye color and hair color(n=592)

Eye color	Hair color			
	Black	Brunette	Red	Blonde
Brown	68	119	26	7
Blue	20	84	17	94
Hazel	15	54	14	10
Green	5	29	14	16

The second table is a 5×4 table cross-classifying number of children by annual income (Cramer, 1946; Diaconis and Efron, 1985). The sample size is very large: 25263. The Pearson χ^2 statistic and the likelihood ratio statistic for testing independence are $\chi^2 = 568.56$ and $L^2 = 569.42$ with degrees of freedom $df = 12$. These statistics again lead to the rejection of independence, moreover the values of these statistics indicate an even worse fit. However Diaconis and Efron [13] showed that among all possible 5×4 tables with sample size 25263 only about 2.1×10^{-7} have $\chi^2 \leq 568.56$. Their conclusion is that even though the latter statistics are more significant, the second table actually lies much 'closer' to independence than the first one.

The above 'volume test' has two drawbacks: from one side it is limited to an independence model and from the other side, it is applicable only for two-way con-

Table 1.2: Cross-classification of number of children by annual income (n=25263)

No. of	Annual income			
children	0-1	1-2	2-3	3+
0	2161	3577	2184	1636
1	2755	5081	2222	1052
2	936	1753	640	306
3	225	419	96	38
3+	39	98	31	14

tingency tables. Approximately ten years after the volume test was proposed, Rudas, Clogg and Lindsay [42] introduced a mixture model approach. The mixture model approach is more general being able to apply to any model, not just an independence model, and to tables of any dimension, not just two-way tables. We note that the mixture model approach gives a result consistent with the Diaconis-Efron volume test approach, in identifying the table closest to independence, but otherwise the numbers are clearly on a different scale.

1.2 Definition of the π^* index

The original definition was introduced by Rudas et al.[42] for contingency table analysis. If P is an observed contingency table and $\mathbf{M}$ is a model then the π^* index is defined by

$$\pi^* = \pi^*(P, \mathbf{M}) = \inf\{\pi : P = (1-\pi)M + \pi R,\ M \in \mathbf{M}, R \in \mathbf{P},\ 0 \leq \pi \leq 1\},$$

where P, M and R are contingency tables of the same size and $\mathbf{P}$ is the set of all contingency tables of this size. So π^* can be interpreted as the smallest fraction of the population outside the model $\mathbf{M}$. Hence, if π^* is small, we will conclude that we are close to the model as only a small fraction of the population cannot be described by $\mathbf{M}$, and on the contrary, if π^* is big, we will conclude that we are not so close to the model as a great fraction of the population cannot described by $\mathbf{M}$ even in the best case. Note, that P can be both table of probabilities and table of frequencies, so we can work with the whole population or with a sample. Also note, that in the second case we obtain an estimate for the true population parameter π^*.

This definition can be extended to probability measures. Let us consider a statistical space $(\Omega, \mathcal{A}, \mathbf{P})$, where the collection of probability measures $\mathbf{P}$ on the sample space $(\Omega, \mathcal{A})$ are dominated by a σ–finite measure λ. It is assumed that $\mathbf{P}$ contains

all sample distributions of interest. In the sequel, the lowercase p denotes the density of the corresponding measure $P \in \mathbf{P}$ with respect to λ. Conversely, for a density p, we denote by P the probability measure $P(A) = \int_A p \, \mathrm{d}\lambda$, $A \in \mathcal{A}$, that we shall write shortly $P = \int p \, \mathrm{d}\lambda$. Let $\mathbf{M} \subset \mathbf{P}$ be a statistical model that we investigate. Then again

$$\pi^* = \pi^*(P, \mathbf{M}) = \inf\{\pi : P = (1-\pi)M + \pi R, \ M \in \mathbf{M}, R \in \mathbf{P}, \ 0 \leq \pi \leq 1\}, \quad (1.2.1)$$

where $P \in \mathbf{P}$ is the observed probability measure. If P_n denotes the empirical measure of the sample then the $\pi^*(P_n, \mathbf{M})$ index measures exactly how far we are from the model $\mathbf{M}$ independently of the sample size. The definition of π^* can be reformulated in the sense that the density p can be represented as a mixture of two densities of the form

$$p = (1-\pi)m + \pi r, \quad (1.2.2)$$

where m comes from the model and r is the density of an unrestricted R from $\mathbf{P}$.

1.3 Results of different authors concerning the π^* index

1.3.1 Applications of the π^* index for contingency tables

Mobility tables The mixture model can be used for analyzing mobility tables. Clogg, Rudas and Matthews [43] applied this approach to the occupational mobility table taken from the famous study by Blau and Duncan [8], as condensed by Knoke and Burke [28]. This table cross-classifies American men in 1962 according to their current occupation category and their father's occupation category. First they proposed a model to be investigated, then they embedded this model in the mixture model, and finally the residuals were examined. So this method served as kind of rapprochement between modeling and graphical techniques for the analysis of categorical data. The models they chose were independence, quasi independence and quasi uniform association models. These three models are nested in the sense that for two way tables of a fixed size, all the independent distributions are contained among the quasi independent ones and these are contained among the ones where quasi uniform association holds true. The mixture index of fit is monotone in this case, see Rudas et al. [42] so this is not very surprising that the estimated π^* values decreased as 0.31, 0.147 and 0.052 indicating that independence model may account for nearly 70% of the population, the quasi independence model for nearly 85%, and the model of quasi uniform association for nearly 95% of the population. Note that the standard statistical decision based on χ^2 statistics is not monotone in the above sense. Analyzing residuals they found the pattern of misfit gaining information about the local structure of the popultion not described by the model. This suggested ways to modify the original model. Residuals in the mixture model are very different from

ordinary residuals in two very important aspects. First, these residuals are always valid, in the sense that representation from which they are derived is always valid in contrast to the usual residuals that are based on the assumption that model **M** is true for the entire population, which may or may not be correct. Second, our residuals are always nonnegative and can be given a probability distribution interpretation, and therefore any technique can be used for analyzing them that can be used to any probability distribution.

Differential Item Functioning (DIF) Differential item functioning is an item response pattern in which members of different demographic groups have different conditional probabilities of answering a test item correctly, given the same level of ability. Rudas and Zwick [44] used the mixture approach to estimate the fraction of the population for which DIF occurs. The question here is what amount of the population cannot be described by the model of conditional independence of 'test result' and 'demographic group' conditioned on 'ability'. The authors analyzed data from the 1993 Advanced Placement Physics B Exam of the Educational testing Service included several multiple choice items. The goal of the analysis was to detect male-female DIF. There were data avaible on 9104 male and 4118 female examinees. For the first 10 items being considered the π^* values turned to be between 0.02 and 0.06 indicating low level of DIF. Here again, examining residuals gave further information about the differences among ability levels being considered. It was possible to pinpoint those parts of te population where DIF occured. For most items, this was the part of the population with lower ability levels.

1.3.2 Extensions to other models

π^* regression Rudas [45] suggested to apply the mixture approach to regression models with normal and uniform error structures. He found that in both cases the minimum mixture estimates of the regression parameters are the parameter estimates of the minimax or Chebysev regression, i.e. the maximal deviance has to be minimized instead of the sum of the deviances that is the case in ordinary least squares regression. As π^* regression coincides with minimax regression, whenever minimax estimation is used, the mixture index of fit provides a natural approach for measuring model fit and for variable selection. Using mixture method is advised especially when we have short tailed error distributions, as in this case minimax regression performs better than least squares regression (see Narula, Welington, 1985). An example illustrating that the mixture index of fit has a straightforward interpretation and can be used for model selection was the analysis of a set of petrol refinery data analyzed earlier by Wood [56] and Narula and Wellington [33]. In this set of data, the dependent variable is the octane number of the product of a petrol refinery unit, and there are four independent variables describing various aspects of the production process. It was known that the designed range output of the refinery unit is $90-94$ octanes. How

good is the fit of the regression model based on all the four explanatory variables and how much worse is the fit of the regressions based on subsets of the variables only? How much is gained by including, say, a third explanatory variable in addition to two? Rudas found that even with using all four explanatory variables, only 54% of the data could be assumed to have come from a regression model, so the model fit is not very good. Improvements of fit when entering 2, 3 and 4 explanatory variables were 44%, 4% and 0.4% suggesting two explanatory variables to be included in the model.

Relationship with the correlation coefficient The mixture index of fit can be applied to any kind of data and to any statistical model, and so it can be also used to give appealing interpretations to well known statistical quantities. Rudas, Clogg and Lindsay [42] considered the relationship between the mixture index of fit and the correlation coefficient. When two variables have a joint normal distribution, their correlation coefficient can be used as a measure of the strength of their association. However the correct assesment of the amount of association when the correlation coefficient takes on other values than 0, ± 1 is difficult, because of the lack of an intuitive interpretation. The authors has found that π^* is the following function of the correlation coefficient.

$$\pi^* = 1 - \sqrt{\frac{1 - |\varrho|}{1 + |\varrho|}},$$

where $|\varrho|$ is the absolute value of the correlation coefficient. For example, when the correlation is 0.6, at most 50% of the population can be described by independence. This is an intuitively clear interpretation of the meaning of the given value of the correlation coefficient.

1.3.3 Asymptotic properties

Xi [58] examined the asymptotic properties of the π^* index and the corresponding model parameters. She found that if the model is not correct the estimate of the π^* index and the corresponding parameter estimates are asymptotically normal.

Theorem 1.3.1. *Let P be the true distribution and $\mathbf{M}$ be the class of baseline models; $P \notin \mathbf{M}$, $M_\theta \in \mathbf{M}$ and $\mathbf{M} = \{M_\theta : \quad \theta \in \Theta\}$. Under regularity conditions, the maximum likelihood estimator of π^* is asymptotically normal, with $\sqrt{n}(\widehat{\pi}^* \to \pi^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\pi^*})$, where $\sigma^2_{\pi^*}$ is the asymptotic variance for the index π^*. Additionally, the corresponding estimator of parameter θ, $\widehat{\theta}$, has an asymptotic multivariate normal distribution as well, with $\sqrt{n}(\widehat{\theta} \to \theta) \xrightarrow{d} MVN(\mathbf{0}, \mathbf{\Sigma}_{\pi^*})$, where $\mathbf{\Sigma}_{\pi^*}$ is the asymptotic variance matrix for $\widehat{\theta}$.*

When the model is true, asymptotic normality need not hold any longer. In this case Xi showed consistency of the π^* index and the model parameters.

Theorem 1.3.2. *Let P be the true distribution, $P = M_\theta$, $\theta \in \Theta$. Then the maximum likelihood estimator of π^* is consistent with respect to $\mathbf{0}$, so are the corresponding parameter estimators of θ with respect to the true value θ.*

She also showed that the rate of convergence in the above theorem is $\sqrt{n}$.

1.4 Some additional theoretical background being used in the following chapters

1.4.1 Generalized Linear Models (GLM)

The concept of GLM, which is a broad class of models was introduced by Nelder and Wedderburn [34]. Generalized linear models are specified by three components: a random component, which identifies the probability distribution of the response variable; a systematic component, which specifies a linear function of explanatory variables that is used as a predictor; and a link describing the functional relationship between the systematic component and the expected value of the random component.

The random component of a GLM consists of independent observations $Y = (Y_1, ..., Y_n)^t$ from a distribution in the natural exponential family.

The systematic component of a GLM relates a vector $\eta = (\eta_1, ..., \eta_n)^t$ to a set of explanatory variables through a linear model

$$\eta = D\theta$$

Here D is a model matrix consisting of values of explanatory variables for the n observations, and θ is a vector of model parameters. The vector η is called the linear predictor.

The third component of a GLM is a link between the random and systematic components. Let $\mu_i = E(Y_i)$, $i = 1, ..., n$. Then μ_i is linked to η_i by $\eta_i = g(\mu_i)$, where g is a monotonic differentiable function. Thus the model links expected values of observations to explanatory variables through the formula

$$g(\mu_i) = D(i, \cdot)\theta, \quad i = 1, ..., n.$$

The function $g(\mu) = \mu$ gives the identity link $\eta_i = \mu_i$, specifying a linear model for the mean response. The link function that transforms the mean to the natural parameter is called the canonical link.

Loglinear models Choosing the random component coming from the Poisson distribution, the systematic component to be determined by categorical explanatory variables and the link function being the log function we obtain the loglinear models:

$$\log m_i = D(i, \cdot)\theta, \quad i = 1, ..., N, \tag{1.4.3}$$

where m_i denotes the observed frequency in the i-th cell and N is the number of cells of the contingency table.

Logistic regression If the random component is from the Bernoulli distribution, the systematic component is determined by mixed explanatory variables and the link function is the logit function we obtain the logistic regression:

$$\log \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = D(i, \cdot)\theta, \quad i = 1, ..., n. \tag{1.4.4}$$

1.4.2 Measuring goodness of fit in logistic regression

Logistic regression is an increasingly popular statistical method used in many areas, e.g. in the social sciences. Here a binary response variable is related to one or more potential explanatory variables through the so called logistic function (1.4.4). θ is estimated by the ML method. However, evaluating goodness of fit is not so easy. There are different methods proposed. When the number of distinct covariate vectors is relatively small comparing to the sample size n, the traditional χ^2 method (Agresti, 1990) can be applied. Difficulties arise with continuous covariates where the number of distinct covariate vectors is close to n. In these cases, very often the observations are grouped using some grouping strategy. The most popular test, that is used by most of the computer packages is Hosmer and Lemeshow's test [21]. They group the observations according to the predicted probabilities of the event putting approximately the same number of subjects in each group and then compare the expected and observed frequencies using the χ^2 statistic. Problems arise when the estimated probabilities approach either zero or one which is the case in many applications due to the above grouping strategy. Another problem is that different computing packages form different groups and altough all of them apply Hosmer and Lemeshow's test, they conclude to different results [38]. Another possibility is to compute a measure in the spirit of R^2 of ordinary least squares regression. A traditional way of it to compute the proportion of cases predicted correctly. Let the predicted value of the response variable be 1 if the predicted probability of the event is greater than 0.5 and let it be 0 otherwise. This measure has several problems (Weisberg, 1978). In particular, there is no baseline or null expectation to compare the correct prediction rate with. Other measures of this type, called pseudo R^2 measures are outlined in Aldrich and Nelson [3] and McKelvey and Zavoina [27]. The backdraw of these measures is, that they are based on the assumption that a dichotomous dependent variable is only a proxy for the true interval level dependent variable that cannot be measured properly and whenever the dependent variable is truly binary, this assumption is not valid. Our π^* approach belongs to the first group of indices. First an appropriate grouping strategy will be chosen based on the theory of multivariate histograms and then the π^* index will be computed using these groups.

Table 1.3: Logistic regression results

θ	SE	Wald	Sig.
-25.89	9.32	7.71	0.005
12.12	4.33	7.81	0.005
10.79	4.19	6.63	0.001

Table 1.4: Results of the Hosmer-Lemeshow test using different computer packages

computer package	number of groups	χ^2	df	Sig.
SAS	10	24.23	8	0.002
Minitab	10	7.81	8	0.453
SPSS	10	11.10	8	0.195
BMDP	10	17.25	8	0.028
SYSTAT	10	20.92	8	0.007

As a numerical example, we consider Finney's data [19] used in many textbooks to illustrate logistic regression. The data consist of 39 observations with two covariables. The response is the occurence of restriction on the skin of the digits, and the covariables are the rate and volume of inspired air. Fitting a logistic regression model to the data we have the following results indicating that both covariables are significant.

Assessing goodness of fit the Hosmer-Lemeshow test gives different results using different statistical packages: SAS, Minitab, SPSS, BMDP and SYSTAT as shown in 1.4. Altough all five software packages are performing the same goodness of fit test, they are obviously using different algorithms to form the groups, which results in radically different conclusion about the goodness of fit.

Pigeon and Heyse [38] reanalised these data. They formed only 4 groups of the observations (instead of 10 formed by the above statistical packages) as they found the number of observations were too small for more groups. The test statistic they used was a modification of the Pearson χ^2 statistic:

$$J^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(n_{ij} - m_{ij})^2}{\varphi_j m_{ij}}$$

where φ_j is an adjustment factor handling the underdispersion in the χ^2 distribution,

Table 1.5: Results of the test proposed by Pigeon and Heyse

covariable used for grouping	number of groups	J^2	df	Sig.
X_1	4	0.49	3	0.920
X_2	4	3.28	3	0.350

n_{ij} and m_{ij} are the observed and expected frequencies for the events and the nonevents in the k groups.

Pigeon and Heyse has proved [37] that this statistic has an asymptotic χ^2 distribution with $k-1$ degrees of freedom.Their grouping strategy was also different, they grouped the data according to a chosen covariable. The authors argued that modifications were needed as in the Hosmer-Lemeshow test very often the estimated probabilities approaches either 0 or 1 for the first and the last groups according to the grouping strategy putting the low probabilities and high probabilities for events together and so the χ^2 test has failed. Pigeon and Heyse's results sorting and grouping the observations by the first or the second covariable can be found in Table 3. As no significant lack of fit could be detected under any of their two grouping strategies the authors concluded that the model provided a reasonable fit of the data.

Chapter 2

Algorithms to compute the π^* index

In this section different algorithms will be presented for the computation of the π^* index. Some of them are very general, e.g. the EM algorithm, that can be applied to find the maximum likelihood estimator of an arbitrary general model having missing data, however some are very special, tailored for the loglinear models, see the SQP algorithm. As the ideas behind these algorithms are also very different, various notations and two possible parametrizations of the π^* problem will be used. Sometimes the GLM parametrization will be applied (simulated annealing, minimax algorithm and SQP algorithm), but in other cases the row-column marginals and their product play important role so the marginals are chosen to be the model parameters (EM and EMF algorithm).

2.1 The EM algorithm

The EM algorithm is a numerical method for finding maximum likelihood estimates. At the heart of every EM algorithm is some notion of missing data. Data can be missing in the ordinary sense of a failure to record certain observations on certain cases. Data can also be missing in a theoretical sense. We can think of E, or expectation, step of the algorithm as filling in the missing data.Once the missing data are reconstructed, then the parameters are estimated in the M, or maximization, step. One of the advantages of the EM algorithm is numerical stability as it leads to a steady increase in the likelihood of the observed data. Besides this, the EM handles parameter constraints very nicely building them into the M step. In contrast, competing methods use special techniques to cope with parameter constraints. A negative feature of the EM algorithm is its slow convergence in a neighbourhood of the optimal point. This rate reflects the amount of missing data in a problem.

A sharp distinction is drawn in the EM algorithm between the observed, incomplete data Y and the unobserved, complete data X. Some function $t(X) = Y$ collapses X onto Y. For instance, if we represent X as (Y, Z), with Z as the missing data, then t is a projection onto the Y component of X. The general idea is to choose X so that maximum likelihood becomes trivial for the complete data. Assume $f(X \mid \theta)$ is the probability density belonging to the complete data X. In the E step the conditional expectation

$$E(\ln f(X \mid \theta) \mid Y, \theta^k) \tag{2.1.1}$$

is computed, where θ^k is the current estimated value of θ. In the M step, we maximize (2.1.1) with respect to θ. This yields the new parameter estimate θ^{k+1}, and we repeat these two steps until converge occurs. The essence of the EM algorithm is that maximizing (2.1.1) increases the loglikelihood $\ln g(Y \mid \theta)$ of the observed data.

2.1.1 EM algorithm for the π^* problem

Let P, M and R be the probability measures defined in (1.2.1) with densities p, m and r and let assume that the model can be written as $\mathbf{M} = \{M(\theta), \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$ $(d \in \mathbb{Z}_+)$. Suppose further that the sample is given as the sum of two latent layers with proportion $1 - \pi$ and π. The first layer comes from the model $\mathbf{M}$ under unknown parameter $\theta \in \Theta$ and the observations in the second layer come from an unrestricted distribution with density r.

In the E–step the proportions of the observed density p that belong to the model and the unrestricted part are calculated as follows:

$$m \propto \frac{(1-\pi)m(\theta^{(k)})}{(1-\pi)m(\theta^{(k)}) + \pi r^{(k)}} \cdot p$$

$$r \propto \frac{\pi r^{(k)}}{(1-\pi)m(\theta^{(k)}) + \pi r^{(k)}} \cdot p.$$

The background of these formulae is the classical Bayes formula. Note that, in general, the right hand sides are not density functions, in order to get densities they have to be normalized.

In the M-step a maximum likelihood estimation is performed for the parameter θ over the parameter space Θ in the first layer, however the second layer remains unchanged.

2.1.2 Algorithm in the finite discrete case

In this section we suppose that the sample space is finite and, for simplicity, $\Omega = \{1, \ldots, N\}$ and $\mathcal{A} = 2^\Omega$. Then all probability measures $P \in \mathbf{P}$ can be identified with its density p with respect to the counting measure. Let $X_1, \ldots, X_n$ be a sample for

the random variable X on the statistical space $(\Omega, \mathcal{A}, \mathbf{P})$. The empirical measure p_n associated with the sample is defined by

$$p_n(i) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{X_j = i\}}, \qquad i = 1, \ldots, n.$$

Moreover, denote by $m(\theta)$, $\theta \in \Theta$, the collection of distributions belonging to the model $\mathbf{M}$, and let π be fixed.

Step 1. Initialization: Let $\theta^{(0)}$ be the maximum likelihood estimate corresponding to the empirical measure p_n and let $r^{(0)}$ be the uniform distribution.

Then repeat Step 2 and Step 3 until converge occurs. After the kth iteration these steps are the following:

Step 2. The E-step:

$$m(i) = \frac{(1-\pi) m(\theta^{(k)}, i)}{(1-\pi) m(\theta^{(k)}, i) + \pi r^{(k)}(i)} \cdot p_n(i), \qquad i = 1, \ldots, N$$

$$r(i) = \frac{\pi r^{(k)}(i)}{(1-\pi) m(\theta^{(k)}, i) + \pi r^{(k)}(i)} \cdot p_n(i), \qquad i = 1, \ldots, N.$$

Normalize both functions to get probability distributions.

Step 3. The M-step: the model part is defined by the model parameter

$$\theta^{(k+1)} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{N} m(i) \log m(\theta, i),$$

and the unrestricted part is unchanged

$$r^{(k+1)}(i) = r(i), \qquad i = 1, \ldots, N.$$

2.1.3 Application to contingency tables

Here a special application will be showed, where the contingency table is a two-way table and the model is the independence model.

In order to parametrize the row–column independence let $\Theta = \mathbb{S}^k \times \mathbb{S}^l$, where k and l are the number of rows and columns, respectively, and $\mathbb{S}^k \subset \mathbb{R}^k$ denotes the k–dimensional simplex, i.e., $\mathbb{S}^k = \{(x_1, \ldots, x_k) : \sum_{i=1}^{k} x_i = 1, \text{ and } x_i \geq 0,\ i = 1, \ldots, k\}$. Then we may identify a distribution from the model $\mathbf{M}$ with $\theta = (\phi, \psi)$, where ϕ is the row marginal and ψ is the column marginal distribution, respectively. Let n_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, l$, be the observed contingency table with sample size $n = \sum_{i,j} n_{ij}$. Then the empirical measure p_n associated with the table $\{n_{ij}\}$ is given by $p_n(i, j) = n_{ij}/n$, the observed proportion in cell (i, j).

Fixing π again the steps of the algorithm are the following:

Step 1. Initialization: Let $\theta^{(0)} = (\phi^{(0)}, \psi^{(0)})$ obtained from the marginals of $p_n(i,j)$, $i = 1, \ldots, k$, $j = 1, \ldots, l$ and let $r^{(0)}(i,j) = 1/kl$, $i = 1, \ldots, k$, $j = 1, \ldots, l$.

Repeat Step 2 and Step 3 while converge occurs. After the kth iteration these steps are the following.

Step 2. The E-step: We set

$$m(i,j) = \frac{(1-\pi)\phi^{(k)}(i)\psi^{(k)}(j)}{(1-\pi)\phi^{(k)}(i)\psi^{(k)}(j) + \pi r^{(k)}(i,j)} p_n(i,j),$$

$$r(i,j) = \frac{\pi r^{(k)}(i,j)}{(1-\pi)\phi^{(k)}(i)\psi^{(k)}(j) + \pi r^{(k)}(i,j)} p_n(i,j).$$

Step 3. The M-step: Here the maximum likelihood estimate is given by taking the marginals of m:

$$\phi^{(k+1)}(i) = m(i,+)/m(+,+), \qquad \psi^{(k+1)}(j) = m(+,j)/m(+,+),$$

where $+$ denotes summation with respect to the argument. $r^{(k+1)}(i,j)$ is a normalization of $r(i,j)$, $i = 1, \ldots, k$, $j = 1, \ldots, l$.

$$r^{(k+1)}(i,j) = r(i,j)/r(+,+).$$

2.1.4 Results for the eye-hair color and income examples

To compute the π^* index we took an enough fine grid on the unit interval and for each setting of the π value (starting from 0) we computed the divergence between the observed and the 'best' mixture tables. The algorithm terminated when this divergence became less than 0.0001. The π^* estimates obtained such a way underestimated the true π^* value as no perfect fit was needed. This reflects in the following π^* values and in Tables 2.1 and 2.2 presenting table decompositions for the two examples. $\pi^* = 0.288$ (eye color), $\pi^* = 0.0914$ (income).

2.2 The SQP algorithm

This algorithm which is available in the MATLAB package finds the constrained minimum of a function starting at an initial estimate. This is generally referred to as constrained nonlinear optimization and can be expressed as

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x)$$

$$\text{subject to } g_i(x) \leq 0, \quad i = 1, \ldots, m \tag{2.2.2}$$

where $x \in \mathbb{R}^d$, f is the objective function ($f : \mathbb{R}^d \to \mathbb{R}$) and g is the vector of constraints ($g : \mathbb{R}^d \to \mathbb{R}^m$).

Table 2.1: Table decomposition of the eye–hair color table computed by the EM algorithm (left:fit, right: lack of fit)

Eye	Hair color			
color	B	B	R	B
B	28.83	119.66	24.86	7.09
B	20.17	83.70	17.39	4.96
H	13.10	54.36	11.29	3.22
G	5.24	21.77	4.52	1.29

Eye	Hair color			
color	B	B	R	B
B	38.75	0.00	0.98	0.00
B	0.00	0.06	0.00	88.47
H	1.80	0.00	2.62	6.71
G	0.00	7.05	9.39	14.61

Table 2.2: Table decomposition of the income table computed by the EM algorithm (left:fit, right: lack of fit)

No.	Annual income			
	0-1	1-2	2-3	3+
0	1992.6	3619.9	1590.5	752.1
1	2784.4	5058.3	2222.5	1050.9
2	823.4	1495.9	657.3	310.8
3	110.5	200.7	88.2	41.7
3+	38.9	70.7	31.1	14.7

No.	Annual income			
	0-1	1-2	2-3	3+
0	150.4	0.0	575.4	870.4
1	0.0	3.1	0.0	0.0
2	104.8	242.6	0.0	0.0
3	112.6	214.8	7.0	0.0
3+	0.0	26.4	0.0	0.0

In constrained optimization most of the methods are the translation of the constrained problem to a basic unconstrained problem by using a penalty function for constraints, which are near or beyond the constraint boundary. In this way the constrained problem is solved using a sequence of parametrized unconstrained optimizations, which in the limit converge to the constrained problem. These methods are now considered relatively inefficient and have been replaced by methods which have focused on the solution of the Kuhn-Tucker equations (2.2.3).

$$\begin{aligned} df(x^*) + \sum_{i=1}^{m} \lambda_i^* dg_i(x^*) = 0 \\ \lambda_i^* g_i(x^*) = 0 \quad i = 1, \ldots, m \\ \lambda_i^* \geq 0 \quad i = 1, \ldots, m \end{aligned} \tag{2.2.3}$$

The Kuhn-Tucker equations are necessary conditions for optimality for a constrained optimization problem. If the problem is a so called convex programming problem, that is $f(x)$ and $g_i(x)$ $i = 1, \ldots, m$, are convex functions, then the Kuhn-Tucker equations are both necessary and sufficient conditions for a global solution point.

The first equation describes a cancelling of the gradients between the objective function and the active constraints at the solution point. In order for the gradients to be cancelled, Lagrangian multipliers ($\lambda_i, \quad i = 1, \ldots, m$) are necessary to balance the deviations in magnitude of the objective function and constraint gradients. Since only the active constraints are included in this canceling operation, constraints which are not active must not be included in this operation and so are given Lagrangian multipliers equal to zero. This is stated implicitly in the last two equations of (2.2.3).

The solution of the Kuhn-Tucker equations form the basis to many nonlinear programming algorithms. These algorithms attempt to compute directly the Lagrangian multipliers. The SQP algorithm is one from this family. The principal idea of it is the formulation of a quadratic programming subproblem based on a quadratic approximation of the Lagrangian function (2.2.4)

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \tag{2.2.4}$$

The quadratic programming subproblem (2.2.5) is obtained by linearizing the nonlinear constraints.

$$\underset{v \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2} v^t H^{(k)} v + df(x^{(k)})^t v$$

$$dg_i(x^{(k)})^t v + g_i(x^{(k)}) \leq 0 \quad i = 1, \ldots, m \tag{2.2.5}$$

This subproblem can be solved using any quadratic programming algorithm. The solution is used to form a new iterate.

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} v^{(k)}.$$

The step length parameter $\alpha^{(k)}$ is determined by an appropriate line search procedure. The matrix $H^{(k)}$ is a positive definite approximation of the Hessian matrix of the Lagrangian function (2.2.4) that can be updated by any of the quasi-Newton methods.

So the SQP algorithm consists of the following three stages:

Step 1: Update the Hessian matrix $H^{(k)}$ of the Lagrangian function (2.2.4) .

Step 2: Compute the solution of the quadratic programming problem (2.2.5).

Step 3: Compute the step length parameter $\alpha^{(k)}$ with a line search procedure.

2.2.1 The π^* problem for contingency tables

Using the GLM form of loglinear models (1.4.3) it can be seen easily that the problem of finding the π^* index is equivalent with the following problem for loglinear models.

$$\text{maximize} \sum_i \exp\left(D\left(i,.\right)\theta\right)$$

$$\text{subject to } D\theta \leq \begin{pmatrix} \log n_1 \\ \vdots \\ \log n_N \end{pmatrix},$$

where $D\left(i,.\right)$ refers to the i-th row of the design matrix, θ is the vector of model parameters and n_i-s are the observed frequencies, $i = 1, \ldots, N$ Then choosing f to be $-\sum_i \exp\left(D\left(i,.\right)\theta\right)$, g to be

$$D\theta - \begin{pmatrix} \log n_1 \\ \vdots \\ \log n_N \end{pmatrix}$$

and substituting θ into x we obtain a constrained optimization problem of the form (2.2.2) that can be solved by the MATLAB package.

2.2.2 Results for the eye-hair color and income examples

The SQP algorithm works very nicely for the first example: it is very fast and attains the best decomposition not only close to that. This latter fact can be seen from the perfect fit of the mixture table and from the number of zeros in the lack of fit table (that will be discussed in Chapter 3), see Table 2.3. However, it fails in the second example which shows the big drawback of the SQP algorithm, namely that it has to be started from a good starting point to get the optimal solution. Since the maximum likelihood estimate (which is a common starting point for all the algorithms being used here) is not like that, the algorithm converged to another local optimum (Table 2.4). The π^* values computed are 0.2959 and 0.6239.

2.3 Simulated annealing

The simulated annealing algorithm is an algorithm for combinatorial optimization, which means that it amounts to finding the 'best' or 'optimal' solution among a finite or countably infinite number of alternative solutions. Considerable effort has been devoted to constructing and investigating methods for solving these kinds of problems from the 60's. Among the methods developed different classes can be formed. From one side one might choose between two options having very large problems. Either

Table 2.3: Table decomposition of the eye–hair color table computed by the SQP algorithm (left:fit, right: lack of fit)

Eye	Hair color			
color	B	B	R	B
B	28.33	119.00	24.08	7.00
B	20.00	84.00	17.00	4.94
H	12.85	54.00	10.92	3.17
G	5.00	21.00	4.25	1.23

Eye	Hair color			
color	B	B	R	B
B	39.66	0.00	1.91	0.00
B	0.00	0.00	0.00	89.05
H	2.14	0.00	3.07	6.82
G	0.00	8.00	9.75	14.76

Table 2.4: Table decomposition of the income table computed by the SQP algorithm (left:fit, right: lack of fit)

No.	Annual income			
	0-1	1-2	2-3	3+
0	30	3577	33	7
1	2755	1215	8	1
2	936	292	1	111
3	225	1	91	38
3+	39	98	31	9

No.	Annual income			
	0-1	1-2	2-3	3+
0	2131	0	2150	1628
1	0	3865	2214	1050
2	0	1460	639	194
3	0	417	4	0
3+	0	0	0	0

one goes for optimality at the risk of very large amounts of computational time, or one goes for quickly obtainable solutions at the risk of suboptimality. The first option constitutes the class of optimization algorithms, while the second constitutes the class of approximation algorithms. From the other side there are general algorithms that are applicable to a wide variety of problems, and tailored algorithms using problem-specific information and so having applications to a restrictive set of problems.

Simulated annealing is a high quality general algorithm. In nature it is an approximation algorithm.

The algorithm is based on the strong analogy between large combinatorial problems and the physical process annealing. The annealing process of solids contains two steps: 1. increase the temperature to a maximum value at which the solid melts. 2. decrease carefully the temperature until the particles arrange themselves in the ground state of the solid. The ground state of the solid is obtained only if the maximum tem-

perature is sufficiently high and the cooling is done sufficiently slow. Otherwise the solid will be frozen into a meta-stable state rather than into the ground state.

The Simulated annealing algorithm is based on the Metropolis algorithm which is a simulation of the evolution of a solid to a thermal equilibrium. The Metropolis algorithm generates a sequence of states of the solid in the following way. Given a current state i with energy E_i, the following state (obtained by a small displacement of the particle) is j with energy E_j. If the energy difference $E_j - E_i$ is nonpositive, the state j will be accepted as the current state. If the energy difference is positive, the state j will be accepted with probability

$$\exp\left(\frac{E_i - E_j}{k_B T}\right),$$

where T denotes the temperature and k_B is a physical constant called Boltzmann constant. The above acceptance rule is called the Metropolis criterion and the algorithm that goes with it is known as the Metropolis algorithm.

If the lowering of the temperature is done sufficiently low, the solid can reach a thermal equilibrium at each temperature. In the Metropolis algorithm this is achieved by generating a large number of tranzitions at a given temperature value.

Then the Simulated annealing algorithm can be viewed as an iteration of Metropolis algorithm with decreasing value of control parameter that plays the role of temperature.

2.3.1 Description of the algorithm

This algorithm is a slight modification of the one described in [5]. Let x be a stochastic vector and let $f(x)$ be the function to minimize. The algorithm starts from a given point $x^{(0)}$ and generates a sequence of points $(x^{(0)}, x^{(1)}, \ldots, x^{(k)}, \ldots)$. New candidate points are generated around the current point $x^{(k)}$applying random moves along each coordinate direction, in turn. The new coordinate values are uniformly distributed in intervals centered around the corresponding coordinate of $x^{(k)}$. If the point falls outside the definition domain of f a new point is randomly generated until a point belonging to the definition domain is found. A new point x is accepted as current point according to the Metropolis criterion: if $f(x) - f(x^{(k)}) < 0$ x is accepted ($x^{(k+1)} = x$) else it is accepted with probability $p = \exp(-(f(x) - f(x^{(k)})/T))$, where T is the temperature parameter. The algorithm starts with a given step vector v (v contains the step size in each coordinate direction) and with the initial temperature determined in the beginning of the algorithm. It performs a given number of cycles (a cycle contains 1 step in each coordinate direction) between two adjustments of the vector v (which makes the algorithm follow better the behaviour of the function) and a given number of adjustments of v between two temperature reductions.

The best point reached is recorded as x^*.

The algorithm stops when the values of f before the last N_ε temperature reductions and $f(x^*)$ are close enough (their distances are less than a parameter ε computed by the program). It also stops when the temperature became less than $s \cdot T_0$, where s is a parameter, T_0 is the initial temperature.

2.3.2 Algorithmic steps

Step 1. Initialization: This involves determining the starting vector $x^{(0)}$, the starting step vector v, the initial temperature T_0, the number of cycles between two adjustments of the step vector v: N_s, the number of adjustments of v between two temperature reductions N_T, the number of temperature reductions to test for termination N_ε, the distance parameter ε and a parameter s of the stopping criterion connected with the temperature.

Step 2: Starting from the point $x^{(k)}$ generate a random point x in the coordinate direction h.

$$x_h = x_h^{(k)} + rv(h), \quad h = 1, \ldots, d, \quad r \in [-1, 1].$$

Step 3: If x is not contained in the definition domain then return to Step 2.

Step 4: Compute $f(x)$. If $f(x) \leq f(x_k)$ then $x_{k+1} = x$. If $f(x) > f(x_k)$ then accept the point with probability $p = \exp(-(f(x) - f(x^{(k)})/T))$.

Step 5: $h = h + 1$. If $h \leq d$ go to Step 2, else $h = 1$.

Step 6: If the number of cycles performed is less than N_s then go to Step2, otherwise update the step vector v.

Step 7: If the number of vector adjustments is less than N_T, then go to Step 2, else reduce the temperature.

Step 8: Terminating criterion. If the distances of the last N_ε optima $f(x^*)$ found in the last N_ε temperature reductions are less than ε or $T < sT_0$ the algorithm terminates.

2.3.3 Applications to contingency tables

Using the previous formulation of the problem

$$\text{maximize} \sum_i \exp\left(D\left(i, .\right)\theta\right)$$

$$\text{subject to } D\theta \leq \begin{pmatrix} \log n_1 \\ \vdots \\ \log n_N \end{pmatrix},$$

we have a function to be minimized ($f(\theta) = -\sum_i \exp(D(i,.)\,\theta)$) on a certain domain defined by

$$D\theta \leq \begin{pmatrix} \log n_1 \\ \vdots \\ \log n_N \end{pmatrix}.$$

Apply the algorithm for them.

2.3.4 Results for the eye-hair color and income examples

Simulated annealing never attains the optimum value itself, rather it supplies some value close to that (depending on the length of the step vector). So π^* is overestimated by this method. Besides it is very slow: the two computations were performed in 199 and 641 seconds, respectively. The only advantage of this algorithm can be that it gives a good starting point where other algorithms can start from. Though in our second example there were no successfull trials, simulated annealing has failed several times (see Table 2.6). The π^* values based on some 'typical' runs turned to be 0.3217 and 0.7471, respectively, the corresponding fit and lack of fit tables can be found in Table 2.5 and 2.6.

Table 2.5: Table decomposition of the eye–hair color table computed by simulated annealing (left:fit, right: lack of fit)

Eye	Hair color			
color	B	B	R	B
B	28.25	118.98	23.95	0.87
B	19.94	83.99	16.91	0.62
H	12.73	53.63	10.80	0.39
G	4.99	21.02	4.23	0.15

Eye	Hair color			
color	B	B	R	B
B	39.74	0.01	2.04	6.12
B	0.05	0.00	0.08	93.37
H	2.26	0.36	3.19	9.60
G	0.00	7.97	9.76	15.84

2.4 The minimax algorithm

The minimax algorithm which is also available in the MATLAB package, minimizes the maximum of a set of functions starting at an initial estimate.

$$\text{minimize } \max\{f_1(x), \ldots, f_N(x)\}, \tag{2.4.6}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}, \quad i = 1 \ldots N$. Minimax also uses an SQP method, however modifications are made to line search and Hessian.

Table 2.6: Table decomposition of the income table computed by simulated annealing (left:fit, right: lack of fit)

No.	Annual income			
	0-1	1-2	2-3	3+
0	2160	32	0	1
1	2754	11	0	87
2	935	0	71	111
3	28	2	91	37
3+	25	2	31	1

No.	Annual income			
	0-1	1-2	2-3	3+
0	0	3544	2183	1635
1	0	5069	2222	964
2	0	1752	568	194
3	196	416	4	0
3+	13	95	0	12

2.4.1 Minimax algorithm for the π^* problem

The π^* application of this method is based on the following theorem of Rudas:

Theorem 2.4.1. *For the densities $m(\theta)$ and p defined in (1.2.2)*

$$1-\pi^* = \sup_{\theta\in\Theta} \inf_{supp\, m(\theta)} \frac{p}{m(\theta)},$$

where $supp\, m(\theta)$ stands for the support of $m(\theta)$, and θ is the vector of model parameters.

Below this theorem will be applied for the finite discrete case, and then especially for two models: for the loglinear models and for logistic regression.

Denoting again by $p_n(i)$ the empirical measure associated with the sample and by $m(\theta, i)$ the measure belonging to the model, $i = 1, \ldots, N$, we have

$$1-\pi^* = \sup_{\theta\in\Theta} \inf_i \left\{ \frac{p_n(i)}{m(\theta,i)}, \quad i = 1, \ldots, N \right\}.$$

As the above set is finite this expression can be rewritten as

$$\frac{1}{1-\pi^*} = \min_{\theta\in\Theta} \max_i \left\{ \frac{m(\theta,i)}{p_n(i)}, \quad i = 1, \ldots, N \right\} \tag{2.4.7}$$

which is a minimax problem that can be obtained from (2.4.6) by substituting θ into x and choosing f_i to be $m(\theta, i)/p_n(i)$, $i = 1, \ldots, N$.

2.4.2 Minimax algorithm for contingency tables

Let $n_1, \ldots, n_N$ be the observed contingency table with sample size $n = \sum_{i=1}^{N} n_i$. Then the empirical measure p_n associated with the table $\{n_i\}$ is given by $p_n(i) = n_i/n, \quad i = 1, \ldots, N$. The corresponding model table $m(\theta, i) = m_i/n, \quad i = 1, \ldots, N$ where $m_1, \ldots, m_N$ are the expected frequencies of the model table that can be expressed as a function of the model parameters using the GLM form (1.4.3)

$$m_i = \exp(D(i, \cdot)\theta),$$

where $D(i, \cdot)$ denotes the i-th row of the design matrix and θ is the vector of model parameters. Substituting the above expressions into (2.4.7) we have

$$\frac{1}{1-\pi^*} = \min_{\theta \in \Theta} \max_i \left\{ \frac{\exp(D(i, \cdot)\theta)}{n_i}, \quad i = 1, \ldots, N \right\}.$$

This problem can be solved in the MATLAB package.

2.4.3 Results for the eye-hair color and income examples

From computational point of view the minimax algorithm is very similarly to the SQP algorithm. From one side, it is also very fast, the running time was 0.29 and 0.71 seconds. From the other side, the π^* values and so the optimal table decompositions obtained are perfect, but only when starting from a 'good' starting point. So it is excellent for the first example and it fails for the second one (see Table 2.7 and 2.8). $\pi^* = 0.2985$ and 0.6251.

Table 2.7: Table decomposition of the eye–hair color table computed by the minimax algorithm (left:fit, right: lack of fit)

Eye	Hair color			
color	B	B	R	B
B	28.81	119.00	25.34	7.00
B	20.00	79.81	17.00	4.69
H	13.53	54.00	11.50	3.17
G	5.00	19.95	4.25	1.17

Eye	Hair color			
color	B	B	R	B
B	38.18	0.00	0.65	0.00
B	0.00	4.18	0.00	89.30
H	1.46	0.00	2.49	6.82
G	0.00	9.04	9.75	14.82

Table 2.8: Table decomposition of the income table computed by the minimax algorithm (left:fit, right: lack of fit)

No.	Annual income			
	0-1	1-2	2-3	3+
0	18	3577	33	7
1	2755	1215	8	0
2	936	292	0	111
3	225	0	91	38
3+	23	98	31	9

No.	Annual income			
	0-1	1-2	2-3	3+
0	2142	0	2150	1628
1	0	3865	2214	1051
2	0	1460	639	194
3	0	418	4	0
3+	15	0	0	4

2.4.4 Minimax algorithm for logistic regression

In logistic regression most of the goodness of fit tests form k groups of the observations. Doing so we will have a $2 \times k$ table and a very similar situation as in contingency tables, only the model is different. The way of forming groups will be described in the following section, now consider how to apply the minimax algorithm assuming we have k groups of observations [54]. Consider the two-way contingency table formed by the response variable and the the k groups of the explanatory variables. Assume that we have n_{ij} observations, $i = 1, 2, \quad j = 1, \ldots, k$. Then again $p_n(i, j) = n_{ij}/n$, where n is the sample size. And again express the model table $m(\theta, i, j), \quad i = 1, 2, \quad j = 1, \ldots, k$ by the model parameters. According to (1.4.4) the conditional probabilities in the s-th column are

$$m(\theta, 1, j \mid j = s) = \frac{\exp(D(s, \cdot)\theta)}{1 + \exp(D(s, \cdot)\theta)}$$

$$m(\theta, 2, j \mid j = s) = \frac{1}{1 + \exp(D(s, \cdot)\theta)}$$

where $D(s, \cdot)$ denotes the i-th row of the design matrix and θ is the vector of the model parameters. The probability of falling in the s-th group can be estimated from the sample, it is $(n_{1s} + n_{2s})/n$. Substituting these expressions into (2.4.7) the right hand side is

$$\min_{\theta \in \Theta} \max_{s} \left\{ \frac{\exp(D(s, \cdot)\theta)/(1 + \exp(D(s, \cdot)\theta))}{n_{1s}/(n_{1s} + n_{2s})}, \frac{1/(1 + \exp(D(s, \cdot)\theta))}{n_{2s}/(n_{1s} + n_{2s})}; \quad s = 1, \ldots, k \right\}$$

which can be also solved by the MATLAB package.

2.4.5 Forming groups in logistic regression

We have seen that having k groups of the observations the π^* value can be computed for the model of logistic regression. But how should we form these groups?

We will use Finney's data [19] in the following example. Using Hosmer and Lemeshow's [21] or Pigeon and Heyse's [38] grouping strategies and forming different number of groups the π^* values, like the χ^2 values turn to be very different. In general, practice shows that different grouping strategies and forming different number of groups yield different π^* values. Then the question is what is a good grouping strategy and how many shouls be formed? Or putting it in another way, what is a good estimate of the observed density? To answer this question we turned to the theory of multivariate histograms [47].

The classical histogram is formed by defining a set of nonoverlapping intervals, called bins, and counting the number of points in each bin. Usually these bins have the same width but sometimes the bin width changes according to the data. The latter histograms are called adaptive histograms and try to handle e.g. the problem of getting bad density estimates in the tails due to the paucity of data. Adaptive histograms are better than fixed bins histograms if they are optimally constructed but they are much worse if they are derived in an ad hoc fashion. So we choose histograms with fixed bins. Then the question is how many bins should be constructed? There are different ideas determining the number of bins or the length of bin width. The most traditional one, that is used by many computer packages is Sturges' number of bins rule. It says that k should be $1 + \log_2 n$. The normal bin width reference rule suggests $h = 3.5\widehat{\sigma}n^{-1/3}$ bin width if the variable is normally distributed, where n is the sample size and $\widehat{\sigma}$ is the estimated standard deviation. Some rules give an upper or lower bound for the number of bins

$$\begin{aligned} &\text{lower bound:} && k = (2n)^{1/3} \\ &\text{upper bound:} && k = \frac{b-a}{3.729\widehat{\sigma}n^{-1/3}}, \end{aligned}$$

where (a, b) is the interval the observations fall in. For multivariate histograms the normal reference rule is $h_l = 3.5\sigma_l n^{-1/(2+d)}$, where d is the dimension, $l = 1, \ldots, d$.

Following these rules at least three at most four groups should be formed by each axis in our case. Choosing 3×3 groups we obtain the following histogram of the rate and volume of inspired air for the observations $Y = 1$.

As this histogram does not seem very fine some smoothing is necessary to get a better estimate of the empirical density. The so called Average Shifted Histogram connected to the above one will be prepared which is a practical choice for computationally and statistically efficient density estimation [47]. Roughly speaking what happens is that dividing the original bins into m parts, the values of histogram will depend not only on the number of counts in the small bins, but with decreasing weight, on the counts in their near and further small neighbourhood bins.

Figure 2.1: Histogram for Finney's data ($Y = 1$)

In the univariate case ASHs are constructed in the following way. Consider a collection of m histograms, $\widehat{f}_1, \widehat{f}_2, \ldots, \widehat{f}_m$, each with bin width h, but with bin origins $0, h/m, \ldots, (m-1)h/m$, respectively. The naive or unweighted averaged shifted histogram is defined as

$$\widehat{f}(\cdot) = \frac{1}{m}\sum_{i=1}^{m}\widehat{f}_i(\cdot).$$

Multivariate ASHs are constructed by averaging shifted multivariate histograms, each with bin width $h_1 \times h_2 \times \ldots \times h_d$. Then the multivariate ASH is the average of $m_1 m_2 \cdot \ldots \cdot m_d$ shifted histograms shifted by the d coordinate axes all possible ways. In the bivariate case the ASH is given by

$$\widehat{f}(\cdot,\cdot) = \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\widehat{f_{ij}}(\cdot,\cdot).$$

For a univariate ASH let B_l denote the narrower bins and let n_l be the bin count in B_l, $l = 1, \ldots, mk$. The height of the ASH in B_l is the average of the heights of the m shifted histograms, each of width h:

$$\frac{n_{l+1-m} + \ldots + n_l}{nh}, \quad \frac{n_{l+2-m} + \ldots + n_{l+1}}{nh}, \quad \ldots, \quad \frac{n_{l+1} + \ldots + n_{l+m-1}}{nh}.$$

Hence, a general expression for the naive ASH is

$$\begin{aligned} \widehat{f}(x;m) &= \frac{1}{m}\sum_{i=1-m}^{m-1}\frac{(m-|i|)n_{l+i}}{nh} \qquad (2.4.8)\\ &= \frac{1}{nh}\sum_{i=1-m}^{m-1}\left(1-\frac{|i|}{m}\right)n_{l+i}, \qquad x\in B_l. \end{aligned}$$

The weights on the bin counts in (2.4.8) take on the shape of an isosceles triangle with base $(-1,1)$. However other weights are also possible. The general ASH uses arbitrary weights $w_m(i)$ and is defined by

$$\widehat{f}(x;m)=\frac{1}{nh}\sum_{i=1-m}^{m-1}w_m(i)n_{l+i}, \qquad x\in B_l. \qquad (2.4.9)$$

In order that $\int \widehat{f}(x;m)dx=1$, the weights must sum to m [47]. An easy way to define the general weights is

$$w_m(i)=m\times\frac{K(i/m)}{\sum_{j=1-m}^{m-1}K(j/m)}, \qquad i=1-m,\ldots,m-1, \qquad (2.4.10)$$

where K is a continuous function defined on $(-1,1)$ called kernel function. K is often chosen to be a probability density function, such as

$$K(t)=\frac{15}{16}(1-t^2)_+^2=\frac{15}{16}(1-t^2)^2I_{[-1,1]}(t),$$

which is called the biweight kernel or quartic kernel.

Then an algorithm to compute the generalized ASH is the following:

Step 1: Construct an equally spaced mesh of width $\delta=h/m$ over the interval (a,b), and compute the correponding bin counts $\{n_l, \quad l=1,\ldots,mk\}$ for the n data points. Typically, $\delta\ll h$.

Step 2: Compute the weight vector , $\{w_m(i)\}$, as in (2.4.10).

Step 3: Compute $\{f_l, \quad l=1,\ldots,mk\}$. It can be done in an efficient manner re-ordering the operations in (2.4.9). Rather than computing the ASH estimates individually in each bin, a single pass is made through the bin counts and the 'effects' of the bins on $f_l, \quad l=1,\ldots,mk$ are computed. This modification avoids repeated weighting of empty bins.

Note that the algorithm for the univariate ASH can be easily extended to the multivariate case, only the parameters in the univariate ASH become vectors.

Table 2.9: Parameter estimates and π^* values based on the ASH with $m = 1, 2, 3$ and 4

m	θ	π^*
1	$(-3, -2, -2)$	0.55
2	$(-13, 7, 4)$	0.3
3	$(-22, 11, 8)$	0.32
4	$(-24, 12, 10)$	0.36

In the above algorithm the precise choice of m is unimportant as long as it is greater than 2 and h is well chosen [47]. However many authors studied the limiting behavior of the ASH as $m \to \infty$. It can be showed that the limiting ASH can be written as

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2.4.11}$$

where x_i is the i-th data point and $K(\cdot)$ is the kernel function of the isosceles triangle density defined by

$$K(t) = (1 - |t|)I_{[-1,1]}(t).$$

(2.4.11) is called the general kernel density estimator with kernel K, corresponding to the generalized ASH defined in (2.4.9). Graphically, what happens is that a kernel scaled by h is placed around each data point and these are added vertically to get the kernel estimate. In contrast, histograms use a rectangular kernel but do not center around the data points, rather they are placed in a rigid mesh. Note, that other kernels such as the normal density can be also used. Thus the ASH provides a direct link to the better known kernel methods. As kernel estimators are usually slow to compute, the ASH is a natural candidate for computation.

2.4.6 Computation of the π^* value for Finney's data

Using the ASHs with $m = 1, .., 4$ (Figures 2.2 – 2.5) as estimates of the empirical density, we have the following estimates for the π^*value and the regression coefficients computed by the minimax algorithm.

Note that as m increases both β-s and π^*-s tend to stabilise and β is getting close to the maximum likelihood estimate which was $\beta = (-25, 12, 10)$. The π^* value is around 0.3 which is not so big. We can join Heyse and Pigeon concluding that this model fits the data reasonably well.

Figure 2.2: ASH for Finney's data ($Y = 1$, $m = 1$)

Figure 2.3: ASH for Finney's data ($Y = 1$, $m = 2$)

Figure 2.4: ASH for Finney's data ($Y = 1$, $m = 3$)

Figure 2.5: ASH for Finney's data ($Y = 1$, $m = 4$)

Chapter 3

Some theoretical questions relating the π^* index of fit

3.1 The greedy algorithm

As we have seen there are several algorithms available to compute the π^* index. They work better or worse, but all of them give some estimate for the π^* index in an actual problem. However if the problem is very extreme like in the example at the end of this section, all of these algorithms have difficulties. We suspect that it is because our problem is equivalent to that of the traveling salesman and so no general algorithm can be given to compute the global optimum rather it can be obtained by examining all candidate solutions.

We have no proof for or against this equivalence, however to have a better insight of the problem we give an algorithm called greedy. This is not a practical algorithm for the computation of the π^* index, rather it is a discussion and characterization of the problem through algorithmic steps.

3.1.1 Description of the algorithm

Our algorithm can be applied for two way contingency tables with the model of independence. Use the notations we had for two way contingency tables in the EM and EMF algorithms. First an initial estimate of the row-column marginals are computed. With this certain number of equations are obtained in (3.1.1), however if there do not exist equalities in each row and column these marginals should be updated. An analogy with bipartite graphs will show that the number of equalities should be $k + l - 1$, if it is not attained, the number of equalities can be increased in the next step. Finally, necessary and sufficient conditions are given for the candidate point we obtained being a local extremal point.

3.1.2 Theoretical justification of the algorithm

Start from the equivalence of the original form of the problem (1.2.1) and the form introduced in the SQP algorithm stated again by the following lemma.

Lemma 3.1.1. *In the case of $k \times l$ contingency tables and the model of independence finding π^* is equivalent to the following constrained maximization problem. Find $(\phi^*(1), ..., \phi^*(k), \psi^*(1), ..., \psi^*(l))$, such that*

$$\sum_{i=1}^{k} \phi(i) = \sum_{j=1}^{l} \psi(j),$$

$\phi(i) \geq 0, \quad \psi(j) \geq 0$, *for which*

$$\sum_{i=1}^{k} \sum_{j=1}^{l} \phi(i)\psi(j)$$

is maximal assuming

$$\phi(i)\psi(j) \leq p_n(i,j), \quad i = 1, ..., k, \ j = 1, ..., l, \tag{3.1.1}$$

where $(p_n(i,j))$ is a given matrix with $p_n(i,j) \geq 0$, $\sum_{i=1}^{k} \sum_{j=1}^{l} p_n(i,j) = 1$. That is, for a given probability distribution of an $k \times l$ contingency table find the maximal independent table "under" it.

Proof According to the definition of π^* index, an observed table of cell frequencies has to be decomposed into two parts, an independent and an unrestricted (residual) part so, that the π value will be minimal. That is the sum of the cell frequencies of the first part will be maximal. Denoting the elements of the observed $k \times l$ table by $p_n(i,j)$ and the row and the column marginals of the independent part by $\phi(1), ..., \phi(k)$ and $\psi(1), ..., \psi(l)$ respectively, we conclude the lemma.

Remark 3.1.2. *Assume that $\phi(i) > 0$, $\psi(j) > 0$, $i = 1, ..., k$, $j = 1, ..., l$. Then we can take the logarithms of $\phi(i)$-s and $\psi(j)$-s and can formulate an equivalent problem*

$$\max \sum_{i=1}^{k} \exp(\log \phi(i)) \sum_{j=1}^{l} \exp(\log \psi(j))$$

$$\log \phi(i) + \log \psi(j) \leq \log p_n(i,j), \quad i = 1, ..., k, \ j = 1, ..., l.$$

Sometimes we will use this logarithmized form of the problem (3.1.1). In such cases it is supposed that $p_n(i,j) > 0$, $i = 1, ..., k, j = 1, ..., l$.

Remark 3.1.3. *Performing different algorithmic steps the condition* $\sum_{i=1}^{k}\phi(i)=\sum_{j=1}^{l}\psi(j)$ *may break, but this is not a problem as multiplying the marginals with a proper constant* v *and* $1/v$ *the condition can be regained while everything remains unchanged.*

Now return to the original form of the problem and consider the following algorithm. Start from an initial estimate of the marginals $\phi(1),...,\phi(k)$. Then (a) compute $\psi(j)$-s by taking the minimum of $p_n(i,j)\ /\phi(i)$, $i=1,...,k$ and (b) update $\phi(i)$-s by taking the minimum of $p_n(i,j)\ /\psi(j)$, $j=1,...,l$.

Lemma 3.1.4. *The equalities in (3.1.1) are preserved by repeating steps (a) and (b).*

Proof Having some row marginals $\phi(1),...,\phi(k)$ and column marginals $\psi(1),...,\psi(l)$ satisfying (1) let us suppose that there exist a $\phi(i_0)$ and a $\psi(j_0)$ such that $\phi(i_0)\psi(j_0)=p_n(i_0,j_0)$. Then in step (a) if $\phi(i_0)=0$, the statement is immediate, otherwise $\phi(i)\psi(j_0)\leq p_n(i,j_0)$ implies that

$$\psi(j_0)=\min_i\frac{p_n(i,j_0)}{\phi(i)}=\frac{p_n(i_0,j_0)}{\phi(i_0)}=\psi(j_0)$$

The proof for step (b) is similar.

Lemma 3.1.5. *Performing steps (a) and (b) the cycle terminates.*

Proof As equalities have been attained in each row and column neither ϕ-s nor ψ-s can be changed. With the above steps certain number of equalities in (3.1.1) is reached. The system of equalities in (3.1.1) can be described in terms of bipartite graphs [26]. Suppose that the rows and the columns of the matrix $(p_n(i,j))$ correspond to the vertices of a graph, and the equalities in (3.1.1) correspond to the edges of that graph.

Lemma 3.1.6. *If the graph corresponding to (3.1.1) is not a connected one (it does not contain a full tree), then the set of edges can be increased so that the value of the function* $\sum_{i=1}^{k}\sum_{j=1}^{l}\phi(i)\psi(j)$ *is increased.*

Proof If the graph is not connected, there exist sets of row and column vertices I and J such that there are no edges among any vertex in I and any vertex in J and the same is true for the sets of vertices $\overline{I}$ and $\overline{J}$, where $\overline{I}$ and $\overline{J}$ are the complementers of I and J in the set of row and column indices respectively. That is, $\phi(i)\psi(j)<p_n(i,j)$ for all $i\in I, j\in J$ and for all $i\notin I, j\notin J$. Let us define $\widetilde{\phi(i)}$ and $\widetilde{\psi(j)}$ by

$$\widetilde{\phi(i)}=\begin{cases} v\phi(i), & i\in I \\ \phi(i), & i\notin I, \end{cases}$$

and

$$\widetilde{\psi(j)}=\begin{cases} \psi(j), & j\in J \\ (1/v)\psi(j), & j\notin J, \end{cases}$$

respectively, where v is a positive constant and compute the sums $\sum_{i=1}^{k}\sum_{j=1}^{l}\phi(i)\psi(j)$ and $\sum_{i=1}^{k}\sum_{j=1}^{l}\widetilde{\phi(i)}\widetilde{\psi(j)}$.

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{l}\phi(i)\psi(j) &= \left(\sum_{i\in I}\phi(i)+\sum_{i\notin I}\phi(i)\right)\left(\sum_{j\in J}\psi(j)+\sum_{j\notin J}\psi(j)\right) = \\ &= S_{in}T_{in}+S_{out}T_{in}+S_{in}T_{out}+S_{out}T_{out},\end{aligned}$$

where $S_{in}=\sum_{i\in I}\phi(i)$, $S_{out}=\sum_{i\notin I}\phi(i)$, $T_{in}=\sum_{j\in J}\psi(j)$ and $T_{out}=\sum_{j\notin J}\psi(j)$. Similarly

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{k}\widetilde{\phi(i)}\widetilde{\psi(j)} &= \widetilde{S_{in}}\widetilde{T_{in}}+\widetilde{S_{out}}\widetilde{T_{in}}+\widetilde{S_{in}}\widetilde{T_{out}}+\widetilde{S_{out}}\widetilde{T_{out}} = \\ &= vS_{in}T_{in}+S_{out}T_{in}+S_{in}T_{out}+\frac{1}{v}S_{out}T_{out}.\end{aligned}$$

The second sum differs from the first one only in those parts where inequalities hold. As

$$vS_{in}T_{in}+\frac{1}{v}S_{out}T_{out}$$

is a convex function of $v>0$, it takes its maximum value on a finite interval at the endpoints. But what are these endpoints? As $\phi(i)\psi(j)<p_n(i,j)$ and $\widetilde{\phi(i)}\widetilde{\psi(j)}=v\phi(i)\psi(j)\le p_n(i,j)$ for all $i\in I, j\in J$ the upper bound for v is

$$\min\left\{\frac{p_n(i,j)}{\phi(i)\psi(j)},\quad i\in I, j\in J, \phi(i)\neq 0, \psi(j)\neq 0\right\}.$$

Similarly

$$\min\left\{\frac{p_n(i,j)}{\phi(i)\psi(j)},\quad i\notin I, j\notin J, \phi(i)\neq 0, \psi(j)\neq 0\right\}$$

gives an upper bound for $1/v$. Any choice provides an additional equality. So we concluded the statement of the lemma.

As the number of edges of a tree is $N-1$ if the number of the vertices is N, a connected graph has at least $N-1$ edges [26]. That is, we can obtain a vector $(\phi(1),...,\phi(k),\psi(1),...,\psi(l))$ where there are at least $k+l-1$ equalities in (3.1.1).

Now we turn to the question of whether the solution we found is a local extremal point. A necessary condition for this can be given on the basis of the Kuhn-Tucker theorem [41].

Theorem 3.1.7. *(Kuhn-Tucker theorem) Suppose $f: D\longrightarrow\mathbb{R}$ is a continuously differentiable function on the set*

$$D=\{x\in\mathbb{R}^{k+l}: g_r(x)\le 0,\quad r=1,...,m\}.$$

If f has a local minimum at $x^* \in D$, then the differential df satisfies the Lagrange multiplier condition

$$df(x^*) - \sum_{r=1}^{m} w_r dg_r(x^*) = 0,$$

where each $w_r \geq 0$ and $w_r g_r(x^*) = 0,\ r = 1, ..., m.$

Theorem 3.1.8. *The necessary condition for a vector $(\phi^*(1), ..., \phi^*(k), \psi^*(1), ..., \psi^*(l))$ to be a local extremal point of the problem in Lemma 3.1.1 is that there exists a nonnegative $k \times l$ matrix W such that it is positive only for the i-s and j-s where $\phi^*(i)\psi^*(j) = p_n(i,j)$ and*

$$\sum_{j=1}^{l} w_{ij} = \phi^*(i), \quad i = 1, ..., k,$$

and

$$\sum_{i=1}^{k} w_{ij} = \psi^*(j), \quad j = 1, ..., l.$$

Proof It is obvious that $(\phi^*(1), ..., \phi^*(k), \psi^*(1), ..., \psi^*(l))$ is an extremal point of the problem if and only if $x^* = (x_1^*, ..., x_k^*, x_{k+1}^*, ..., x_{k+l}^*)$ is an extremal point of the logarithmized problem

$$\min f(x)$$

$$g_{ij}(x) \leq 0, \quad i = 1, ..., k,\ j = 1, ..., l,$$

where

$$x_i = -\log \phi(i), \quad x_{k+j} = -\log \psi(j),$$

$$f(x) = \sum_{i=1}^{k} \exp(-x_i) \sum_{j=1}^{l} \exp(-x_{k+j})$$

and

$$g_{ij}(x) = -x_i - x_{k+j} - \log p_n(i,j), \qquad i = 1, ..., k,\ j = 1, ..., l.$$

Taking the partial derivatives of the functions f and g_{ij} we get

$$\partial_s g_{ij}(x) = \begin{cases} -\delta_{si}, & s \leq k \\ -\delta_{s(k+j)}, & s > k \end{cases} \tag{3.1.2}$$

$$\partial_s f(x) = \begin{cases} -\exp(-x_s) \sum_{j=1}^{l} \exp(-x_{k+j}), & s \leq k \\ -\exp(-x_s) \sum_{i=1}^{k} \exp(-x_i), & s > k \end{cases} \tag{3.1.3}$$

According to Theorem 3.1.7 there exist weights $w_{ij}\, i = 1, \ldots, k,\ j = 1, \ldots, l$ such that

$$\sum_{i=1}^{k}\sum_{j=1}^{l} w_{ij}\partial_s g_{ij}(x^*) = \partial_s f(x^*), \tag{3.1.4}$$

$w_i j \geq 0$ and $w_i j g_i j\,(x^*) = 0$. The left hand side of (3.1.4) can be written in the form

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{l} w_{ij}\delta_{si} &= \sum_{j=1}^{l} w_{sj}, \qquad s \leq k, \\
\sum_{i=1}^{k}\sum_{j=1}^{l} w_{ij}\delta_{(k+j)s} &= \sum_{i=1}^{k} w_{is}, \qquad s > k.
\end{aligned}$$

Substituting this into (3.1.4) and using (3.1.3) we get

$$\begin{aligned}
\sum_{j=1}^{l} w_{sj} &= \exp(-x_s^*)\sum_{j=1}^{l}\exp(-x_{k+j}^*) = \phi^*(s)\sum_{j=1}^{l}\psi^*(j), \qquad s \leq k, \\
\sum_{i=1}^{k} w_{is} &= \exp(-x_s^*)\sum_{i=1}^{k}\exp(-x_i^*) = \psi^*(s)\sum_{i=1}^{k}\phi^*(i), \qquad s > k.
\end{aligned}$$

As $\sum_{i=1}^{k}\phi^*(i) = \sum_{j=1}^{l}\psi^*(j)$, let us divide both equalities by them and let the new W matrix the original divided by this sum. This new matrix satisfies the theorem: the first statement follows from $w_{ij}g_{ij}(x^*) = 0,\ \ i = 1, ..., k,\ j = 1, ..., l$, and the second coming from the above equalities.

Corollary 3.1.9. *The necessary condition for a vector $(\phi^*(1), ..., \phi^*(k), \psi^*(1), ..., \psi^*(l))$ to be a local extremal point of the problem in lemma 3.1.1 is that for all sets of indices I and J for which $\phi^*(i)\psi^*(j) < p_n\,(i, j)\,,\ i \in I,\ j \in J$ the inequality*

$$\sum_{i\in I}\phi^*(i)\sum_{j\in J}\psi^*(j) < \sum_{i\notin I}\phi^*(i)\sum_{j\notin J}\psi^*(j)$$

should hold.

Proof Use the previous notations for $S_{in} = \sum_{i\in I}\phi^*(i)$, $S_{out} = \sum_{i\notin I}\phi^*(i)$, $T_{in} = \sum_{j\in J}\psi^*(j)$ and $T_{out} = \sum_{j\notin J}\psi^*(j)$. According to Theorem 3.1.8 $w_{ij} = 0$ for all $i \in I$, $j \in J$. On the other hand, as the corresponding graph is a connected one, there exists at least one w_{ij} $(i \notin I,\ j \notin J)$ which is not zero (see Lemma 3.1.6). This and Theorem

3.1.8 imply, that

$$T_{out} = \sum_{j \notin J} \psi^*(j) = \sum_{i \in I \cup \overline{I} j \notin J} \sum w_{ij} = \sum_{i \in I} \sum_{j \notin J} w_{ij} + \sum_{i \notin I} \sum_{j \notin J} w_{ij} =$$
$$\sum_{i \in I} \sum_{j \notin J} w_{ij} + \sum_{i \in I} \sum_{j \in J} w_{ij} + \sum_{i \notin I} \sum_{j \notin J} w_{ij} = S_{in} + \sum_{i \notin I} \sum_{j \notin J} w_{ij} > S_{in}.$$

Similarly $S_{out} > T_{in}$. And so

$$\sum_{i \notin I} \phi^*(i) \sum_{j \notin J} \psi^*(j) = S_{out} T_{out} > S_{in} T_{in} = \sum_{i \in I} \phi^*(i) \sum_{j \in J} \psi^*(j).$$

Now let us see what is the sufficient condition for a $(\phi^*(1), ..., \phi^*(k), \psi^*(1), ..., \psi^*(l))$ point to be a local extremal point. Considering the logarithmized form of the problem we can see that the constraint became linear, however the function to be maximized is a convex function. Take the Taylor series of this convex function at

$$(\log \phi^*(1), ..., \log \phi^*(k), \log \psi^*(1), ..., \log \psi^*(l))$$

and keep only the first two terms of it. Then we obtain a linear approximation of the convex function to be maximized and with this linear function the problem becomes a linear programming problem [31]. We will call it the supporting linear programming problem.

Theorem 3.1.10. *A sufficient condition for a point* $(\phi^*(1), ..., \phi^*(k), \psi^*(1), ..., \psi^*(l))$ *to be a local extremal point of the problem in Lemma 3.1.1 is that the supporting linear programming problem (of the logarithmized problem) has global optimum in*

$$(\log \phi^*(1), ..., \log \phi^*(k), \log \psi^*(1), ..., \log \psi^*(l)).$$

Proof In linear programming the optimum is obtained in a vertex. Whether this

$$(\log \phi^*(1), ..., \log \phi^*(k), \log \psi^*(1), ..., \log \psi^*(l))$$

point is a vertex of the set $\log \phi(i) + \log \psi(j) \leq \log p_n(i,j)$, $i = 1, ..., k$, $j = 1, ..., l$. The vertices are determined by 'determining' (maximal independent) equalities of the $k \times l$ inequalities. Similarly to the transportation problem we can assign a graph to the matrix of the above system of inequalities. Then the maximal independent system of equalities correspond a tree in the graph. As we got a tree in Step 2 we are in a vertex. Assume that this is the global optimum point of the supporting linear programming problem. It means that any directional derivative of the linear function is negative or zero but because of the strict convexity all the directional derivatives of the convex function will be negative. It is enough to consider only the directions of the edges leading to the neighbour vertices, as any other direction is a linear combination of these.

What happens if this vertex is not a global optimum point of the supporting linear programming problem? Then we have to check a neighbour vertex. What are the neighbour vertices? The vertices are determined by equalities. Canceling one such equality one can obtain an edge running to another vertex. This second vertex will be assigned by cutting the edge by a hyperplane, that is, by another equality. So the neighbour vertices are those points that differ in one equality from the previous one. Of course, we can work also with the original set instead of the logarithmized one.

3.1.3 Algorithmic steps

Step 1. Initialization: Generate independent random values $\phi(1), ..., \phi(k)$ from the uniform distribution on $[0, 1]$ and then divide them by $\sum_{i=1}^{k} \phi(i)$.

Step 2. Row-column updating: a,

$$\psi(j) = \min\left\{\frac{p_n(i,j)}{\phi(i)}, \quad i = 1, ..., k; \phi(i) > 0\right\},$$

b,

$$\phi(i) = \min\left\{\frac{p_n(i,j)}{\psi(j)}, \quad j = 1, ..., l; \psi(j) > 0\right\}.$$

Step 3. Generating new edges: Connect two separate subgraphs choosing

$$v = \min\left\{\frac{p_n(i,j)}{\phi(i)\psi(j)}, \quad i \in I, j \in J, \phi(i) \neq 0, \psi(j) \neq 0\right\}$$

$$\text{or} \quad \frac{1}{v} = \min\left\{\frac{p_n(i,j)}{\phi(i)\psi(j)}, \quad i \notin I, j \notin J, \phi(i) \neq 0, \psi(j) \neq 0\right\}$$

depending on which gives larger increase of the function $\sum_{i=1}^{k}\sum_{j=1}^{l} \phi(i)\psi(j)$.

Repeat Step 3 until $k + l - 1$ equalities are attained in (3.1.1).

Step 4. Testing optimality: Remove one equality from the $k + l - 1$ equalities. Then the graph corresponding to the problem will be again disconnected. Perform Step 3. If the equality removed is regained, try to remove another equality. If an equality can be removed, the new $(\phi(1), ..., \phi(k), \psi(1), ..., \psi(l))$ point will be the candidate for being a local optimum. The algorithm terminates when no equality can be changed to another one.

3.1.4 An example

Consider the problem

$$\max \sum_{i=1}^{k}\sum_{j=1}^{k}\phi(i)\psi(j)$$

$$\phi(i)\psi(j) \leq p_n(i,j), \quad i=1,...,k,\ j=1,...,k,$$

$\phi(i) \geq 0$, $\psi(j) \geq 0$ with $p_n(i,j) = 1 - \delta_{ij} + \kappa\varepsilon_{ij}$, where δ_{ij} is the Kronecker δ, κ is a small constant and ε_{ij} are independent random variables, say standard normal $i = 1,...,k$, $j = 1,...,k$. If $\kappa = 0$, it can be seen that in the local optimum points both $\phi(i)$ and $\psi(j)$ can be only 0 or 1, $i = 1,...,k$, $j = 1,...,k$. A global optimum point is attained, if the number of 1-s is $k/2$ if k is even or $(k+1)/2$ if k is not even.

If $\kappa \neq 0$, we think that the global optimum point can be found only by examining all the possibilities and so our problem is equivalent to that of the traveling salesman.

3.2 Robustness of the π^* index and the EMF algorithm

In this section the π^* index will be considered from robustness point of view. By empirical studies it will be shown that the π^* index has a kind of "automatic" robustness. It means that the level of π^* is stable under small, arbitrary departures from the null hypothesis, i.e. it has the property of robustness of validity (see Heritier and Ronchetti [20]). The heart of our treatment is an algorithm based on the EM approach for computing the π^* index and the distance between the model and the observed distribution under fixed contamination level. Then the so–called contamination plot is introduced which represents the magnitudes of distance corresponding to different contamination levels. We apply it to study the robustness of the goodness–of–fit measure derived by the distance measure. Note, that this algorithm also provides the π^* value when the distance between the model and the observed distribution approaches zero.

First a new interpretation will be given for π^* in the framework of robust statistics. Then the EMF algorithm will be developed to solve the robust divergence minimization problem. It will be proved that this algorithm is monotone similarly to the standard EM algorithm. Our algorithm will be applied to the finite discrete case. It will be seen that the algorithm that we apply to estimate the contaminating distribution coincides the algorithm RANK developed by Zipkin [59]. Finally, the results will be illustrated by the analysis of eye and hair color and income data.

3.2.1 Robustness of the π^* index

A new interpretation associated with π^* can be given in the framework of robust statistics. Let d be a generalized distance measure on the space $\mathbf{P}$ of probability

measures, i.e., we only suppose that $d(P,Q) \geq 0$ for all $P, Q \in \mathbf{P}$ and $d(P,Q) = 0$ iff $P \equiv Q$. Thus d is not a proper metric, as neither symmetry nor the triangle inequality is assumed. Moreover consider one of the fundamental notions in robust statistics, see Huber [22], the contamination neighbourhood defined by

$$N(\mathbf{M}, \pi) = \{Q : Q = (1-\pi)M + \pi R,\ M \in \mathbf{M}, R \in \mathbf{P}\},$$

where $0 \leq \pi \leq 1$ is fixed and $\mathbf{M} \subset \mathbf{P}$. π is called the level of contamination here. Note that $N(\mathbf{M}, \pi)$ is not a neighbourhood in topological sense, it is the union of the elementary contamination neighbourhoods $N(M, \pi)$, $M \in \mathbf{M}$, see [22]. Then the $\pi^* = \pi^*(P, \mathbf{M})$ index is the least non–negative solution of the equation

$$d(P, N(\mathbf{M}, \pi)) := \min_{Q \in N(\mathbf{M}, \pi)} d(P, Q) = 0$$

in π. There are several possibilities choosing the distance measure d and our choice is strongly related to measuring the goodness of fit. Here the Kullback–Leibler information divergence is applied. It is defined by

$$D(P \parallel Q) = \int_\Omega \log \frac{P}{Q} \, \mathrm{d}P = \int_\Omega p \log \frac{p}{q} \, \mathrm{d}\lambda,$$

where $P, Q \in \mathbf{P}$ and, by convention, $0 \cdot \log(0/x) = 0$ if $x \geq 0$ and $x \cdot \log(x/0) = +\infty$ if $x > 0$. We choose information divergence because the test based on it is exponential rate optimal if the admissible tests are compared in Bahadur sense, see Tusnády [51].

In the following sections the divergence of P and $N(\mathbf{M}, \pi)$ will be minimized with fixed π. Then the minimum divergence as a function of π will be plotted. It will be called the contamination plot, which attains 0 when $\pi = \pi^*$. However not only this point is a question, but also the shape of the contamination plot is of primary importance considering the problem from robustness point of view.

The robustness of a hypothesis testing problem can be considered from two points of view: (a) what is the influence of a small, arbitrary departure from the null hypothesis and (b) what happens when the sample distribution is changed within a small contamination neighbourhood. In the first case we would like to test the hypothesis $H_0 : P \in N(\mathbf{M}, \varepsilon)$, where ε is near 0, and the behaviour of the applied test statistics T_n must be investigated. If T_n is the above information divergence, then the figure of the contamination curve in the neighbourhood of zero plays a key role. However if T_n is the π^* index then it is exactly linear, i.e. $\pi^*(\varepsilon) = \pi^*(0) - \varepsilon$, where $\pi^*(\varepsilon)$ denotes the π^* index under ε contamination of the null hypothesis. This shows a kind of "automatic" robustness independently of the chosen distance measure (see Donoho and Liu [16]). In the second case we think that the question is the behaviour of the contamination function at π^*. We shall see by numerical studies that the derivative of this function is approximately zero as the contamination level tends to π^*.

3.2.2 Divergence minimization under fixed contamination level

Let us assume again that the model can be written as $\mathbf{M} = \{M(\theta), \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$ ($d \in \mathbb{Z}_+$). Denote by Γ the collection of all density functions with respect to λ, i.e. $f \in \Gamma$ iff $f \geq 0$ λ almost surely and $\int_\Omega f d\lambda = 1$. Furthermore let $P \in \mathbf{P}$ be a given probability measure with density p. Our aim in this section is to minimize the information divergence between P and the contaminated model $N(\mathbf{M}, \pi)$, where the contamination level π is fixed. Define the function

$$D_\pi(\theta, r) = D(P \parallel (1-\pi)M(\theta) + \pi R),$$

where $\theta \in \Theta$, $r \in \Gamma$ and $R = \int r d\lambda$. In order to minimize D_π over $\Theta \times \Gamma$ we apply the EM algorithm. Suppose that $(\theta^{(k)}, r^{(k)}) \in \Theta \times \Gamma$ is obtained after the kth iteration.

Our EM approach is based on the EM algorithm for standard finite–mixture models. Suppose that the sample is given as the sum of two latent layers with proportion $1-\pi$ and π. The first layer comes from the model $\mathbf{M}$ under unknown parameter $\theta \in \Theta$ and the observations in the second layer come from an unrestricted distribution. In the E–step the proportion of the density p that belongs to the model is calculated as follows:

$$m \propto \frac{(1-\pi)m(\theta^{(k)})}{(1-\pi)m(\theta^{(k)}) + \pi r^{(k)}} \cdot p. \tag{3.2.5}$$

Note that the right hand side is not a density function in general. We should normalize it, but it is not of primary importance in running the algorithm.

The M–step consists of two minimization phases. The first one requires the minimization of $D(M \parallel M(\theta))$ with respect to θ over the parameter space Θ, where $M = \int m \, \mathrm{d}\lambda$, i.e., $\theta^{(k+1)}$ is defined as

$$\theta^{(k+1)} = \arg\min_{\theta \in \Theta} D(M \parallel M(\theta)). \tag{3.2.6}$$

We should remark that the measure $M(\theta^{(k+1)})$ is the likelihood projection of M onto the parametric model $\mathbf{M}$. At the second step the probability measure $R^{(k+1)}$ with density $r^{(k+1)}$ is determined as

$$R^{(k+1)} = \arg\min_{R \in \mathbf{P}} D(P \parallel (1-\pi)M(\theta^{(k+1)}) + \pi R). \tag{3.2.7}$$

While the first minimization phase is a usual maximum likelihood estimation for the parameter θ, the second one is given by a familiar water–filling technique of information theory [11]. The following lemma justifies this second step (substituting $T = (1-\pi)M$), moreover it shows how the density $r^{(k)}$ can be computed in the kth iteration.

Lemma 3.2.1. *Let P and T be two measures on the measurable space $(\Omega, \mathcal{A})$ such that $P(\Omega) = 1$ and $0 \leq T(\Omega) \leq 1$. The measure Q for which*

1. $Q(\Omega) = 1$,
2. $Q \geq T$ *(i.e.* $Q(A) \geq T(A)$ *for all* $A \in \mathcal{A}$*)*,
3. $D(P \,||\, Q)$ *is minimal*

is unique and absolute continuous w.r.t. P. *The Radon-Nikodym derivative* $g = \mathrm{d}Q/\mathrm{d}P$ *is given by* $g = \max\{\varkappa, f\}$, *where* $f = \mathrm{d}T/\mathrm{d}P$ *and* $\varkappa$ *is chosen so that* $\int_\Omega g \,\mathrm{d}P = 1$ *holds.*

Proof. It is clear that the optimal measure Q satisfying the desired properties is absolutely continuous w.r.t. P. First, suppose that $T \ll P$. Let S be an arbitrary probability measure such that $S \ll P$ and $S \geq T$. Denote the Radon–Nikodym derivative $\mathrm{d}S/\mathrm{d}P$ by h. Then $S \geq T$ implies that $h \geq f$ P almost surely. It is enough to prove that

$$\int_\Omega \log h \,\mathrm{d}P \leq \int_\Omega \log g \,\mathrm{d}P. \tag{3.2.8}$$

Since $\log x \leq x - 1$ for $x > 0$ and the function $h - g$ is non–negative on the set $\{\omega, g(\omega) > \varkappa\}$, we have that

$$\log h \leq \log g + \frac{1}{g}(h - g) \leq \log g + \frac{1}{\varkappa}(h - g).$$

Then the inequality (3.2.8) is given by integration, bearing in mind that h and g are probability densities. The uniqueness of Q follows from the fact that $\log x = x - 1$ iff $x = 1$. If T is not absolutely continuous with respect to P, then let $T = T_1 + T_2$ be the Lebesgue decomposition of T, where $T_1 \ll P$ and $T_2 \perp P$. One can see that the set of measures $\{S : S \geq T_1,\, S \ll P\}$ contains the set $\{S : S \geq T,\, S \ll P\}$, which proves our assertion in this case.

It is well known that the EM algorithm is an ascent algorithm, see Dempster et.al. [12]. We prove that the same property holds for our algorithm in such a sense that the divergence $D_\pi(\theta, r)$ is not increased after an EM iteration. We start with a lemma of Ispány [23], which plays fundamental role in the proof.

Lemma 3.2.2. *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space and let* f_1, f_2 *and* g *non–negative functions on it. If*

$$A := \int_\Omega \frac{f_1}{f_1 + g} \log \frac{f_1}{f_2} \,\mathrm{d}P \leq 0, \tag{3.2.9}$$

then

$$B := \int_\Omega \log \frac{f_1 + g}{f_2 + g} \,\mathrm{d}P \leq 0. \tag{3.2.10}$$

Theorem 3.2.3. *The iterates defined by (3.2.5) , (3.2.6) and (3.2.7) obey*

$$D_\pi(\theta^{(k+1)}, r^{(k+1)}) \le D_\pi(\theta^{(k)}, r^{(k)}) \tag{3.2.11}$$

for all $k \in \mathbb{N}$.

Proof. It is enough to prove that

$$\Big[D_\pi(\theta^{(k+1)}, r^{(k+1)}) - D_\pi(\theta^{(k+1)}, r^{(k)})\Big] + \Big[D_\pi(\theta^{(k+1)}, r^{(k)}) - D_\pi(\theta^{(k)}, r^{(k)})\Big] \le 0.$$

Here, the first term is non–positive by (3.2.7). For the second term we have

$$D_\pi(\theta^{(k+1)}, r^{(k)}) - D_\pi(\theta^{(k)}, r^{(k)}) = \int_\Omega \log \frac{(1-\pi)m(\theta^{(k)}) + \pi r^{(k)}}{(1-\pi)m(\theta^{(k+1)}) + \pi r^{(k)}} \, \mathrm{d}P. \tag{3.2.12}$$

On the other hand, by (3.2.6) we obtain

$$\int_\Omega \frac{(1-\pi)m(\theta^{(k)})}{(1-\pi)m(\theta^{(k)}) + \pi r^{(k)}} \log \frac{m(\theta^{(k)})}{m(\theta^{(k+1)})} \, \mathrm{d}P \le 0.$$

Let $f_1 = (1-\pi)m(\theta^{(k)})$, $f_2 = (1-\pi)m(\theta^{(k+1)})$ and $g = \pi r^{(k)}$, and apply Lemma 3.2.2. Then, one can see that the right hand side of (3.2.12) is non–positive, which was to be proved.

Since the information divergence is non–negative the sequence of divergences $D_\pi(\theta^{(k)}, r^{(k)})$, $k \in \mathbb{N}$, converges monotonically to some value D^*. In many practical applications D^* will be a local minimum. D^* is not necessarily global minimum, moreover, D_π can possess a lot of local minimums, see the example in Section 3.1.4. In general, if D_π has several local minimum points, the sequence of iterates defined by (3.2.5), (3.2.6) and (3.2.7) depends on the choice of initial guess $(\theta^{(0)}, r^{(0)})$. We conjecture that if the likelihood function is unimodal in Θ and the contamination level is small, then any iterates converges to the unique D^* irrespective of its starting point.

3.2.3 Algorithm in the finite discrete case

Using the notations introduced in Chapter 2, and fixing π again, the algorithm is very similar to that of the classical EM algorithm, only in the M step the second layer is not unchanged, rather it is optimized so that the divergence of the observed and the mixture measure will be minimal.

Step 1. Initialization: Let $\theta^{(0)}$ be the maximum likelihood estimate corresponding to the empirical measure p_n and let $r^{(0)}$ be the uniform distribution.

Then repeat Step 2 , Step 3 and Step 4 until convergence occurs. After the kth iteration these steps are the following:

Step 2. The E-step:

$$m(i) = \frac{(1-\pi)m(\theta^{(k)}, i)}{(1-\pi)m(\theta^{(k)}, i) + \pi r^{(k)}(i)} \cdot p_n(i), \qquad i = 1, \ldots, N.$$

$$r(i) = \frac{\pi r^{(k)}(i)}{(1-\pi)m(\theta^{(k)}, i) + \pi r^{(k)}(i)} \cdot p_n(i), \qquad i = 1, \ldots, N.$$

Normalize both functions to get density functions.

Step 3. The M-step: compute the likelihood projection by

$$\theta^{(k+1)} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{N} m(i) \log m(\theta, i).$$

Step 4. The F-step: the unrestricted part is computed as follows. We have to minimize the expression

$$\sum_{i=1}^{N} p_n(i) \log \frac{p_n(i)}{(1-\pi)m(\theta^{(k+1)}, i) + \pi r(i)}$$

in $r = \{r(1), \ldots, r(N)\}$, which is equivalent to the maximization of

$$\sum_{i=1}^{N} p_n(i) \log((1-\pi)m(\theta^{(k+1)}, i) + \pi r(i)). \tag{3.2.13}$$

The solution of this problem is given by Lemma 3.2.1. Let

$$t(i) = (1-\pi)m(\theta^{(k+1)}, i), \quad i = 1, \ldots, N,$$

and define the numbers $\widetilde{r}(i)$, $i = 1, \ldots, N$, by the following way:

$$\widetilde{r}(i) = \begin{cases} 0, & t(i)/p_n(i) \geq \varkappa \\ \varkappa p_n(i) - t(i), & t(i)/p_n(i) < \varkappa \end{cases}$$

where the $\varkappa$ constant fulfills the equation

$$\varkappa \sum_{\{i: t(i)/p_n(i) < \varkappa\}} p_n(i) + \sum_{\{i: t(i)/p_n(i) \geq \varkappa\}} t(i) = 1.$$

Note that the solution of this equation is uniquely determined, hence $\varkappa$ is well defined. Then the contanimating distribution $r^{(k+1)}$ is given by $r^{(k+1)}(i) = \widetilde{r}(i)/\pi$, $i = 1, \ldots, N$. Then the steps of this computation are

(a) Determine the ratios $f(i) = t(i)/p_n(i)$, $i = 1, \ldots, N$.

(b) Then order $f(i)$'s to get $f^*(i)$'s, and denote $t^*(i), p_n^*(i)$ the rearrengement of the sequences $t(i), p_n(i)$, $i = 1, \ldots, N$, according to this ordering.

(c) Let $\Sigma(1) = \sum_{i=1}^{N} t(i) = 1 - \pi$ and define the sequence $\Sigma(j)$ recursively by

$$\Sigma(j) = \Sigma(j-1) + (f^*(j) - f^*(j-1)) \sum_{i=1}^{j-1} p^*(i), \qquad j = 2, \ldots, N$$

until $\Sigma(j) \geq 1$. Denote $j^* = j$ the first index for which $\Sigma(j) \geq 1$, and if such index does not exist, i.e. $\Sigma(N) < 1$, then let $j^* = N + 1$.

(d) Then the constant $\varkappa$ is given by

$$\varkappa = \frac{1 - \sum_{j=j^*}^{N} t^*(j)}{\sum_{j=1}^{j^*-1} p^*(j)},$$

and the contanimating distribution can be calculated by the following manner

$$r^{(k+1)}(j) = \begin{cases} 0, & j \geq j^* \\ (\varkappa p_n(j) - t(j))/\pi, & j < j^* \end{cases}$$

3.2.4 Application to contingency tables

Like in the EM algorithm, the model of independece will be considered using the same notations. Fixing π again the steps of the algorithm are the following:

Step 1. Initialization: Let $\theta^{(0)} = (\phi^{(0)}, \psi^{(0)})$ obtained from the marginals of $p_n(i,j)$, $i = 1, \ldots, k$, $j = 1, \ldots, l$ and let $r^{(0)}(i,j) = 1/kl$, $i = 1, \ldots, k$, $j = 1, \ldots, l$.

Repeat Step 2, Step 3 and Step 4 while convergence occurs. After the kth iteration these steps are the following.

Step 2. The E-step: We set

$$m(i,j) = \frac{(1-\pi)\phi^{(k)}(i)\psi^{(k)}(j)}{(1-\pi)\phi^{(k)}(i)\psi^{(k)}(j) + \pi r^{(k)}(i,j)} p_n(i,j)$$

and

$$r(i,j) = \frac{\pi r^{(k)}(i,j)}{(1-\pi)\phi^{(k)}(i)\psi^{(k)}(j) + \pi r^{(k)}(i,j)} p_n(i,j)$$

$i = 1, \ldots, k$, $j = 1, \ldots, l$.

Step 3. The M-step: Here the maximum likelihood estimate is given by taking the marginals of m:

$$\phi^{(k+1)}(i) = m(i,+)/m(+,+), \qquad \psi^{(k+1)}(j) = m(+,j)/m(+,+),$$

where $+$ denotes summation with respect to the argument.

Step 4. The F-step: The filling algorithm of the previous section is applied with the distributions $\{p_n(i,j)\}$ and $\{t(i,j)\}$, this latter defined by

$$t(i,j) = (1-\pi)\phi^{(k+1)}(i)\psi^{(k+1)}(j),$$

$i = 1,\ldots,k$, $j = 1,\ldots,l$.

Table 3.1: Table decomposition of the eye–hair color table computed by the EMF algorithm

Eye	Hair color			
color	B	B	R	B
B	28.33	119	24.09	7
B	20	84	17	4.94
H	12.85	54	10.93	3.18
G	5	21	4.25	1.24

Eye	Hair color			
color	B	B	R	B
B	39.67	0	1.91	0
B	0	0	0	89.6
H	2.15	0	3.07	6.82
G	0	8	9.75	14.76

Table 3.2: Table decomposition of the income table computed by the EMF algorithm

No.	Annual income			
	0-1	1-2	2-3	3+
0	1940	3577	1564	740
1	2754	5079	2221	1051
2	793	1463	640	33
3	99	183	80	38
3+	36	67	29	14

No.	Annual income			
	0-1	1-2	2-3	3+
0	221	0	620	895
1	0	1	0	0
2	142	289	0	2
3	125	235	15	0
3+	2	30	1	0

3.2.5 Results for the eye-hair color and income examples

In this section contamination plots will be drawed for these two examples that reach axis x at the π^* value. To draw the contamination plot take an enough fine grid on the unit interval $[0,1]$ and compute the information divergence between the empirical distribution and the contaminated model distribution at fixed contamination level by

Figure 3.1: Contamination plot for the eye–hair color example

the above iteration. Note that the initial guesses can be chosen differently from the one defined in Step 1, as increasing the value of π according to the grid, the initial guesses can be the final values of the previous iteration, too. Practice shows that this way the algorithm becomes much faster. The π^* index computed by the EMF algorithm for the eye-hair color and income examples are 0.2958 and 0.1023, respectively. Table 3.1 and 3.2 show the fitted and contaminated tables and the contamination plots can be seen on Figure 3.1 and Figure 3.2 for these examples. Most of the cases the contamination function is a monotone decreasing convex function as in Figure 3.1 and Figure 3.2. These plots also show that the tangent of the contamination curve is the horizontal axis at $\pi = \pi^*$ justifying the robustness of the π^* index. In order to study the robustness of the distance measure which is the information divergence in our case, we have to examine the behaviour of the contamination curve at $\pi = 0$. If this function dies down rapidly, then the null hypothesis is already accepted under small departure from the model. This is not the case here, so the test statistics based on the information divergence is a robust one. Ispány and Verdes [23] suggested another measure for robustness defined by computing the ratio above and below the contamination curve in the triangle determined by the points $(0, C(0))$, $(0, 0)$ and $(\pi^*, 0)$. If this number is small, then the contamination curve dies down rapidly. For Table 1 it is 0.5563, which is quite a big value.

3.2.6 Comparison of the used algorithms based on their results for the eye-hair color and income examples

To make the above algorithms comparable, all of them are coded in the MATLAB package and all of them are started from the same initial estimate; the ML estimate

Figure 3.2: Contamination plot for the income example

(which means that the initial parameter vector is defined by fitting the independence model to the observed table). The running time is also recorded. Then consider the five algorithms from two points of view: running time and exactness of the estimate (see the best decompositions given by Xi [58] for the two examples in Tables 3.3 and 3.4).

Table 3.3: The best decomposition of the eye–hair color table

Eye	Hair color			
color	B	B	R	B
B	28.3	119	24	7
B	20	84	17	5
H	13	54	11	3
G	5	21	4.3	1.2

Eye	Hair color			
color	B	B	R	B
B	39.7	0	2	0
B	0	0	0	89
H	2	0	3	7
G	0	8	9.7	14.8

We find that the EM and EMF algorithms obtain the π^* estimate slower (EM in 161 and 136 seconds while EMF in 28 and 16 seconds for the two examples) as they have to run the algorithm several times with different settings of π, but the EMF algorithm is much faster than the traditional EM and it seems to be also superior from exactness point of view among the five methods. The SQP and minimax algorithms are equally fast (SQP needed only 0.49 and 0.17 seconds similarly to minimax using

Table 3.4: The best decomposition of the income table

No.	Annual income			
	0-1	1-2	2-3	3+
0	1939	3577	1564.3	740.6
1	2755	5081	2222	1052
2	793.5	1463.5	640	303
3	99.5	183.5	80.3	38
3+	36.7	67.6	29.6	14

No.	Annual income			
	0-1	1-2	2-3	3+
0	221.5	0	619.7	895.4
1	0	0	0	0
2	142.5	289.5	0	3
3	125.5	235.5	15.7	0
3+	2.3	30.4	1.4	0

0.43 and 0.71 seconds for the two problems) and yield about the same (good or bad) results for the above two examples. The slowest algorithm is simulated annealing (with running time 199 and 641 seconds), moreover, it does not attains the optimum value itself, it gets only close to that in the best case (when not converging to a different local optimum). The only advantage of simulated annealing can be to combine it with an algorithm (SQP or minimax) that needs a 'good' starting point. However, this attemption has failed with the income example. Note, that the EMF algorithm gives more information of the problem, here not only the π^* estimate is computed, but some information is also gained about robustness of the problem. Also note, that the above algorithms are not equally general: some are tailored for loglinear models (SQP, simulated annealing), however some are general and could be extended to any statistical model (EM, EMF, minimax). Generalization of the latter algorithms to different models is a possible way for further research. A third remark is that these comparisons are only the first steps in an extensive work just started investigating these algorithms through several examples and simulation studies.

Chapter 4

Applications

In this chapter the π^* index is applied in different fields: in sociology, endodontology, meteorology and pedagogy. Some problems arisen in these sciences could be solved by the tools we have for the computation of the π^* index, however in some applications only a simplification of the problem could be handled. So questions and possible ways of further research are also given in this chapter.

4.1 Temporal change of wind direction field over Hungary

The objective of this study [50] was to assess whether changes in the surface pressure field over Europe are reflected in the wind direction field, too. The data basis consists of hourly wind direction data from 1968-72 and 1991-95, from 10 meteorological stations of Hungary.

The rise in global surface air temperature, due to the increase of greenhouse gases, has probably induced a redistribution of the surface pressure field. In Europe, in the winter half-year the average values of surface pressure increased in the south and decreased in the north between 1961 and 1990, whereas in the summer half-year there were no significant changes. On the other hand, Metaxas et al. (1991) and Bartzokas and Metaxas [6] found that the average intensity of influx of cold air masses in summer, coming from the north and north-west to the south-east Europe, has increased. So he summer circulation system is also changing as a consequence of the redistribution of the surface pressure field in summer. Such changes may affect Middle -Europe to a less degree than the north-western and south-eastern regions, however we can ask what changes has occured in Hungary.

In continental Europe, in winter, the direction of the average pressure gradient is from south to north (Justyák, 1994). Thus, according to the mentioned results, this gradient inreases, too, which might give rise to changes in the circulation system,

e.g. the frequency or the average speed of southerlies winds may increase during this season. The spatial distribution of annual and monthly average sea level pressure fields in Hungary is determined by the so called "basin character", which means that in the middle of the Great Hungarian Plain, a pressure minimum can be found. This is caused by the strong warming in summer and the frequent passing through of the Mediterranean cyclones in winter.

In this study [50] the change in Hungarian wind direction field is analyzed.

Table 4.1: χ^2 values for testing homogeneity of wind direction field between 1968–72 and 1991–95

City	Time period				
	Winter	Spring	Summer	Autumn	Year
Békéscsaba	710	271	148	259	528
Debrecen	1177	468	249	841	1613
Szeged	1178	468	232	398	813
Miskolc	843	734	662	873	2044
Kékestető	1374	484	315	267	979
Budapest	1036	627	617	630	2036
Győr	1848	1035	1489	1602	4695
Szombathely	1691	979	576	433	2671
Keszthely	1517	1437	941	1524	4484
Pécs	1451	240	618	611	1511

The statistical test we first performed to assess differences in the hourly wind direction between 1968-72 and 1991-95 in 10 cities seasonally and also annually was the χ^2 test of homogeneity. As the observed hours turned to be very big numbers, this test showed significant difference of the distributions in each case which was not surprising taking into account the sensitivity of the χ^2 test for the sample size. The critical value of the χ^2 distribution at 0.05 significance level with 14 degrees of freedom is 23.65 which was highly exceeded in our examples, so we can conclude that wind direction has changed remarkably in the last twenty years. Note, however that there were big differences in the χ^2 values computed which means that the magnitude of differences in the hourly wind directions depended on the cities and the seasons very much as our table shows below.

To get a better impression about the magnitude of discrepancy from homogeneity, we computed the π^* indices, too (see Table 4.2). The π^* values were computed

Table 4.2: π^* values for testing homogeneity of wind direction field between 1968–72 and 1991–95

City	Time period				
	Winter	Spring	Summer	Autumn	Year
Békéscsaba	0.59	0.28	0.24	0.44	0.28
Debrecen	0.55	0.48	0.40	0.58	0.39
Szeged	0.55	0.48	0.39	0.31	0.25
Miskolc	0.46	0.59	0.56	0.55	0.53
Kékestető	0.67	0.50	0.53	0.31	0.50
Budapest	0.60	0.65	0.65	0.49	0.60
Győr	0.72	0.64	0.62	0.47	0.54
Szombathely	0.71	0.49	0.47	0.44	0.44
Keszthely	0.71	0.61	0.53	0.61	0.51
Pécs	0.73	0.41	0.47	0.45	0.47

the same way as for contingency tables with the model of independence, the first observations being considered as expected while the second observations considered as observed frequencies. In Tables 4.1 and 4.2 the π^* values and the χ^2 values changed similarly, with some exceptions, which is not very surprising as they measure different things. The χ^2 statistic is a sum of deviances of the observations in the 16 cells measured in two time periods, while the π^* index measures the maximum deviance in the above cells.

Tables 4.1 and 4.2 show that both χ^2 and π^* values are the greatest in winter confirming the facts described in the introduction. One can also notice, that the change of wind direction field is more remarkable in the western cities. In addition, some bigger values can be detected also in summer. So wind direction has changed in summer, too, in accordance with the results of Metaxas et al.(1991).

4.2 ISSP data

ISSP (International Social Survey Programme) is a continuing annual programme of cross-national collaboration on surveys covering topics important for social science research. It brings together pre-existing social science projects and coordinates research goals, thereby adding a cross-national, cross-cultural perspective to the in-

dividual national studies. Twenty-nine countries are members of the ISSP. The 1995 cross national survey was about national identity. Three variables: 'how proud of achievements in sports', 'how proud of achievements in science' and 'how proud of democracy' is chosen from this database and cross classified with the variable 'country'. The question is whether proudness of the above things depends on the country.

Table 4.3: Partial residual table for the model of conditional independence of 'science' and 'sports' conditioned on Germany

How proud	How proud of sports			
of science	very	somewhat	not very	not at all
very	61.2	0.0	0.0	0.0
somewhat	0.0	0.0	2.1	6.0
not very	0.0	10.5	17.0	7.2
not at all	0.1	1.9	0.1	5.0

Table 4.4: Partial residual table for the model of conditional independence of 'science' and 'sports' conditioned on Hungary

How proud	How proud of sports			
of science	very	somewhat	not very	not at all
very	162.3	0.0	7.1	5.4
somewhat	0.0	0.0	0.0	0.0
not very	0.0	35.3	33.3	0.0
not at all	0.0	5.5	7.0	6.8

As ISSP surveys contain very large samples, performing contingency table analysis with the traditional χ^2 based methods for the hypothesis of independence results in clear rejection of this hypothesis. How far are we from independence? The answer is given on the basis of what proportion of the observations can be described by the model of independence in the 'best case'. Analyzing residuals we can get a picture of the nature of deviance from independence. The π^* method can be applied also for loglinear models. We can ask, for example, whether the 'how proud of achievements in science' and the 'how proud of achievements in sport' variables are conditionally

Table 4.5: Partial residual table for the model of conditional independence of 'science' and 'sports' conditioned on Czech Republic

How proud	How proud of sports			
of science	very	somewhat	not very	not at all
very	46.0	0.0	1.1	0.0
somewhat	64.0	0.0	0.0	1.2
not very	0.0	0.0	44.7	4.2
not at all	0.0	31.0	34.8	11.7

Table 4.6: Partial residual table for the model of conditional independence of 'science' and 'sports' conditioned on Slovenia

How proud	How proud of sports			
of science	very	somewhat	not very	not at all
very	74.6	0.0	0.9	1.0
somewhat	0.0	0.0	0.0	0.0
not very	0.0	31.4	31.2	5.6
not at all	0.0	12.2	4.9	8.9

independent conditioned on 'country'. Again, residuals describe the way the data alter from this model.

Computing the π^* index of fit for the two way tables 'how proud of achievements in sports'-'country', 'how proud of achievements in science'-'country' and 'how proud of democracy'-'country' for the 23 countries involved in this study assuming independence, we obtain 0.29, 0.29 and 0.41 respectively. This can be interpreted as about 29% of the observations cannot be described by the model of independence in the first two cases and about 41% of the observations cannot be described by this model in the third case. So these tables are not very close to independence, especially the democracy table. One can form two groups of the countries: western and eastern. Computing again the π^* index we get 0.23, 0.21 and 0.23 for the western countries and 0.34, 0.28 and 0.21 for the eastern countries. We can see that the π^* indices decreased dramatically for the democracy tables which means, that altough people in the 23 countries think very differently about their democracy, the differences almost

disappear within the two groups. This is not the case for the sport table, where in the group of the eastern countries the differences of opinion even increase.

Fitting the model of conditional independence: independence of 'how proud of achievements in sport' and 'how proud of achievements in science' conditioned on 'country', the π^* is 0.31 which is again not too small. The tables of residuals show that deviation from conditional independence is caused by those who are proud of both their sport and science and those who are not proud of any of them. Of course, different countries show slightly different picture, see Tables 4.3- 4.6.

4.3 An endodontic study

The aim of this study [24] was to evaluate the influence of root canal forms on short term and long term sealing ability of two types of root canal sealers.

The objective of root canal treatment is the three dimensional obturation of the root canal space after meticulous cleaning and shaping (Nguyen, 1994). During root cleaning and canal shaping procedures nondesired canal deformities (e.g. zip, elbow, hourglass form, perforation) may be created. In most of the studies curved root canals showed significantly worse postinstrumentational canal shape, compared with straight root canals [14], [15]. Sealing ability of root canals having preparation deformities may be compromised. Some authors (Mann and McWalter, 1987) did not find statistically significant differences in shaping qualities between straight and curved root canals. In a recently published paper Wu [57] found statistically significant negative effect of apical transportation on seal.

The material of the sealer may also play role in sealing the root canal. Sealapex showed significantly more leakage after storing in water for one year than other sealers [57]. After storage in water for two years there were no significant difference between Sealapex and Pulp Canal Sealer, however ZOE containing sealers showed more leakage than Sealapex. The obturation quality of some root canal sealers were significantly different in studies Ostravik [35], Wu et al. [57] and Sen et al. [46]. The above results are on the long term sealing ability, however there is a shortage of the short term examination of the sealers.

In our experiment twenty straight and twenty curved root canals were prepared and then obturated by two types of root canal sealers: Pulp Canal Sealer and Sealapex (10 root canals by each sealer in each curvature group). Finally leakage was measured with a fluid transport model at 0, 1, 3, 9 and 12 month intervals.

The question was the dependence-independence structure of four discrete variables: 'leakage' (good, bad), 'time' (0, 1, 3, 9, 12), 'form' (straight, curved) and 'material' (Pulp Canal Sealer, Sealapex). Loglinear models are appropriate in this situation, however the four dimensional contingency table formed by the above variables was a sparse table, and so the π^* approach was used. As the research question was how 'material' and 'form' effects 'sealing' as time is passsing, the model (leakage*time*form, leakage*time*material) was fitted.

Table 4.7: Parameter estimates describing the effect of 'form' and 'material' on 'leakage'

leakage*material	leakage*form	leakage*time*material	leakage*time*form
-0.0346	-0.0076	0.1334	-0.2311
		0.1658	0.2230
		-0.1772	-0.0482
		0.0006	0.0031

The π^* value turned to be 0.04 which is very small and indicates good fit. The parameter estimates of the fitted model (see Table 4.7) show that in the first two time periods the effect of the root canal form is stronger. Considering the periods after the first month the influence of root canal sealer is greater on root canal filling quality. So straight and curved root canals were not very different in long time run, however Sealapex weaked more than Pulp Canal Sealer by progress of time.

4.4 A study on university and college admission in Hungary between 1967 and 1989

Fényes and Verdes examined [18] what kind of social effects determined admission into Hungarian universities and colleges in the late socialism.

The basis of data analysis was an all-round data collection on 'university and college admission' (admitted, declined), 'origin' (working class, non working class), 'test results' (eligible, non eligible) and 'sex' (man, woman) between 1967 and 1989. As the dependent variable was a dichotomous variable, the model was chosen to be logistic regression. However as no sample was drawn from the population the usual reasoning for statistical analysis could not be applied. So model fit was performed by the π^* approach which allows working with the whole population as well.

As 'origin' and 'sex' was supposed to effect through 'test results' a chain of logistic regression models, called quasi path models were fit to every year. First 'test results' was considered to be the dependent variable depending on 'origin' and 'sex'. Then the effect of 'test results' was examined on 'university and college admission'. As those who were not eligible based on their test scores were not admitted at all, logistic regression could not be fit to this model, instead the effects of 'origin' and 'sex' on 'university and college admission' were compared conditioning and not conditioning on 'test results'. Namely, the second model was testing the effect of 'origin' and 'sex' on 'university and college admission' only for those who were eligible while the third

model was testing the same effect for the whole population. The parameter estimates and the π^* values can be found in Tables 4.8 – 4.10.

As the π^* values are very small, models fit very well in each year. Based on the model parameters the following conclusions can be drawn.

The preference of working class students at entrance examinations was present in each year of the period between 1967 and 1989 (except for the last year), in spite of the continuous growth of the number of student positions avaible at colleges and universities and the abolition of the quota system in 1962. Therefore, the hypothesis that in the late 1980s admission was based only on achievement at entrance examinations has turned out to be wrong since the direct effect of 'origin' on admission has been proved in these years as well. The results also show that the preference of working class students at entrance examinations was not undermined by the corruption present at these exams. This finding is in contradiction with the theory of Szelényi and Aschaffenburg [49]. If we consider the whole period, the preference of working class students was generally increasing rather than decreasing. More detailed results show that after the reform of 1968 the preferential system stagnated, then with the standstill of the reform in 1973-74 it grew stronger, up until 1977, when the advantage of working class students at entrance examinations began to decrease. This process was going on throughout the rest of the period except for 1986.

Considering the distinction between the sexes, in the majority of the years of the period men were preferred at entrance examinations. Moreover, their examination results were a little better than those of women, therefore, it is not true that women achieved better results at entrance exams. The degree of preference based on sex, however, is usually smaller than that of discrimination based on origin, or more precisely, in the last years of the period it cannot even be shown. If temporal effects are also taken into consideration, it can be seen that the preference of men was decreasing a little throughout the 21 years examined. The reason for this can probably be found in the fact that men's results were growing better throughout the years, which counterbalances the smaller ratio of the number of male candidates to a certain extent.

4.5 The effect of social capital on the intention of higher educational studies among denominational secondary school pupils

After the 90s years the ratio of the denominational schools has increased remarkably in Hungary. This is such a new phenomenon of the society that so far there do not exist extensive studies evaluating these type of schools. The present study [39] attempts to answer what kind of factors determine the choice of higher educational institute among denominational secondary school pupils. The survey is based on data about 1463 11-12 years pupils of 53 denominational secondary schools drawn from the

Hungarian denominational secondary schools by stratified sampling method in 1999. Here again, logistic regression is chosen as statistical model. First this model will be fitted by the traditional maximum likelihood method. Then the model fit will be evaluated by the π^* approach.

The dependent variable is whether someone intents to go to a university or to a college (only those are considered who wants to enter any of the above). The factors that can effect the choice of higher educational institute can be formed into different groups [9], [4], [17]: the family capital, which has the following types: cultural, economical, social and external tie capital; and school environment, which has again two types: parents' cultural capital and social capital. There are several variables measuring the above things, see Table 4.11. Fitting logistic regression model to the data, the results can be seen in Table 4.11. One can see that the most important explanatory variables are 'the ratio of parents with higher educational degree' and 'the ratio of children having close circle of friends'. The remaining variables share their effects on the dependent variable and so their individual contributions to the explanatory force is very weak. So the model can be simplified to the above three variables model. π^* goodness of fit will be tested only for this simpler model.

The π^* value is 0.37 which is not very small, but taking into account that it is a three variable model missing a certain amount of explanatory force, this result is not so bad and it shows significant effect of these explanatory variables on the dependent variable. Note, that using this simpler model is also a consequence of technical difficulties, as presently 'too many' variables and mixed variables cannot be handled in our algorithms. And so we arrived at the problems and further research.

4.6 Questions and further research

Besides the extensive comparison of the algorithms some generalizations of them would be also necessary. For example, in π^* logistic regression only four continuous explanatory variables can be handled as in preparing multidimensional ASHs there is a dimensional limit (see Scott, 1992). Moreover, at this moment there is no way of working with mixed explanatory variables in this model. The possibility of applying these algorithms to other statistical models should be also investigated. Finally, it is also a question what properties have the π^* estimates for small samples, e.g. for sparse tables like in the above endodontic study.

Table 4.8: π^* logistic regression results for the first model of Section 4.4 ('test results' depending on 'origin' and 'sex')

year	origin exp(B)	sex exp(B)	π^*
1967	0.94	1.04	0.030
1968	0.88	0.99	0.050
1969	1.07	1.15	0.002
1970	1.01	1.02	0.006
1971	0.95	0.94	0.040
1972	1.02	1.05	0.010
1973	1.01	1.08	0.003
1974	1.01	1.05	0.003
1975	0.93	1.04	0.017
1976	0.87	1.05	0.005
1977	0.89	0.92	0.005
1978	0.96	0.99	0.003
1979	0.94	1.01	0.008
1980	0.92	1.08	0.001
1983	0.95	1.13	0.002
1984	1.07	1.17	0.011
1985	0.97	1.03	0.005
1986	1.02	1.02	0.022
1987	0.88	1.24	0.015
1988	0.90	1.20	0.009
1989	0.86	1.37	0.010

Table 4.9: π^* logistic regression results for the second model of Section 4.4 ('admission' depending on 'origin' and 'sex' conditioning on 'test results')

year	origin exp(B)	sex exp(B)	π^*
1967	1.53	1.66	0.034
1968	1.66	2.65	0.015
1969	1.67	2.09	0.005
1970	1.76	1.78	0.015
1971	1.68	1.28	0.002
1972	1.69	1.24	0.003
1973	1.80	1.31	0.006
1974	2.99	2.00	0.013
1975	3.31	2.32	0.000
1976	3.45	2.13	0.007
1977	5.56	2.00	0.018
1978	4.97	1.73	0.003
1979	4.35	1.79	0.007
1980	3.76	1.53	0.015
1983	3.26	1.50	0.011
1984	1.19	0.96	0.009
1985	2.86	1.37	0.012
1986	6.73	1.19	0.002
1987	1.74	1.26	0.003
1988	1.44	1.21	0.005
1989	1.12	0.72	0.005

Table 4.10: π^* logistic regression results for the third model of Section 4.4 ('admission' depending on 'origin' and 'sex' not conditioning on 'test results')

year	origin exp(B)	sex exp(B)	π^*
1967	1.16	1.30	0.037
1968	1.16	1.57	0.005
1969	1.30	1.51	0.005
1970	1.28	1.31	0.023
1971	1.24	1.07	0.039
1972	1.34	1.15	0.013
1973	1.36	1.22	0.009
1974	1.40	1.30	0.008
1975	1.31	1.31	0.036
1976	1.40	1.21	0.040
1977	1.36	1.11	0.041
1978	1.48	1.16	0.021
1979	1.32	1.17	0.014
1980	1.17	1.17	0.017
1983	1.17	1.21	0.081
1984	1.09	1.16	0.007
1985	1.12	1.08	0.005
1986	0.99	1.16	0.005
1987	0.92	1.26	0.016
1988	0.92	1.21	0.011
1989	0.87	1.33	0.013

Table 4.11: Logistic regression results for the model in Section 4.5

variable	B	S.E.	Wald	df	Sign.	exp(B)
pardiplo	-0.0990	0.2014	0.2415	1	0.6231	0.9058
paremplo	0.3819	0.1905	4.0188	1	0.0450	1.4651
econcapi	-0.0397	0.1873	0.0448	1	0.8323	0.9611
cultcapi	0.1808	0.1577	1.3143	1	0.2516	1.1982
urban	0.3112	0.1673	3.4618	1	0.0628	1.3651
famrelig	0.2776	0.2490	1.2435	1	0.2648	1.3200
pardenom	0.2886	0.1656	3.0373	1	0.0814	1.3346
parnetwo	0.0021	0.0372	0.0032	1	0.9552	1.0021
parfrien	0.1440	0.1717	0.7030	1	0.4018	1.1548
childfri	0.3458	0.1714	4.0710	1	0.0436	1.4131
meandipl	0.7043	0.1215	33.619	1	0.0000	2.0225
meanpafr	-1.2899	0.9165	1.9811	1	0.1593	0.2753
meanchfr	1.3957	0.4876	8.1944	1	0.0042	4.0377
eastwest	-0.4912	0.1803	7.4209	1	0.0064	0.6119
evangeli	0.7104	0.5534	1.6479	1	0.1992	2.0348
reformed	0.2611	0.4515	0.3344	1	0.5631	1.2983
catolic	-0.1731	0.4604	0.1415	1	0.7068	0.8410

Summary

There are several traditional goodness of fit measures used in contingency table analysis. All of these statistics have asymptotic χ^2 distribution, however we have a bad approximation of this distribution if the sample size is too small, and on the other hand, if the sample size is too big, we always tend to reject the null hypothesis. This was the reason of introducing the mixture index of fit for contingency tables (Rudas et al. [42]), the definition of which can be found in Chapter 1. This index has some advantages over the traditional goodness of fit measures, namely π^* measures the distance from the model independently of the sample size, it gives a nice impression about the amount of discrepancy between the model and the data and it can be extended to probability measures. This generalization can be also found in Chapter 1. Finally some previous results and applications concerning this index is detailed in this first part of the thesis.

In Chapter 2 a very practical question is made, namely computation of the π^* index. The following algorithms are described and used here: the EM algorithm, the SQP algorithm, simulated annealing and the minimax algorithm. Besides their detailed discussion the two examples introduced in [42], the so called eye–hair color data and the so called income data are analyzed through this chapter using and comparing the different algorithms and illustrating the nature of computation for contingency table analysis. Some of these algorithms, however are very general and can be used also for other statistical models. As an example, the minimax algorithm is applied to logistic regression at the end of Chapter 2. It is presented through a standard database, with Finney's data.

The theoretical part of the thesis (Chapter 3) examines two problems. The first one is a characterization and discussion of the computational problem connected to the π^* index for two way contingency tables with the model of independence, while the second is considering the π^* index from robustness point of view. The first part gives an explanation of the computational difficulties arising even for two-way tables (see the income example). We suspect that to compute this index belongs to the NP hard problems and so no general solution can be given. In the second part of the chapter the robustness of π^* is justified and additionally a fifth algorithm is introduced. This EMF algorithm seems to be superior among all according to our experiences.

In the last part of the thesis there are five applications of the π^* index. First some contingency table applications are presented, where the sample size is too small or too big and so traditional methods are not appropriate. Then two logistic regression models are fitted.

Összefoglaló

1 Bevezetés

A társadalomtudományokban a statisztikai elemzésekbe bevont legtöbb változó diszkrét, ezért a diszkrét változós adatelemzés: modellek illesztése és az illeszkedés jóságának a mérése egyre nagyobb figyelmet kapott az utóbbi évtizedekben. E disszertáció az illeszkedés jóságának a mérését állítja középpontba, egészen pontosan ennek egy új mérőszámát, a π^* indexet vizsgálja.

A két legklasszikusabb illeszkedést mérő mérőszám kontingencia táblákra a Pearson [36] féle χ^2 statisztika és a likelihood hányados χ^2 statisztika. További lehetséges mérőszámok a Freeman-Tukey statisztika, a módosított loglikelihood statisztika, a minimális diszkrimináló információs statisztika (Kullback, 1959) és a Neyman féle χ^2 statisztika.

Ezekben a statisztikákban az a közös, hogy mindegyik asszimptotikus χ^2 eloszlást követ, azonban ha a minta kicsi, az asszimptotikus tulajdonságok még nem érvényesülnek, másrészt pedig, ha a minta nagy, többnyire el kell vetnünk a nullhipotézist.

Ezért - a hagyományos mérőszámok egy alternatívájaként - vezette be Rudas, Clogg és Lindsay [42] a π^* indexet. Az eredeti definíció a következő. Ha P egy megfigyelt kontingencia tábla és $\mathbf{M}$ egy modell, akkor

$$\pi^* = \pi^*(P, \mathbf{M}) = \inf\{\pi : P = (1-\pi)M + \pi R,\ M \in \mathbf{M}, R \in \mathbf{P},\ 0 \leq \pi \leq 1\},$$

ahol P, M és R ugyanolyan méretű kontingencia táblák és $\mathbf{P}$ az ilyen méretű kontingenciatáblák halmaza. Így a π^* index úgy interpretálható, mint a populációnak azon legkisebb hányada, mely nem írható le az $\mathbf{M}$ modellel. Ezért, ha π^* kicsi, azt mondhatjuk, hogy közel vagyunk a modellhez, mert a populációnak csak egy kis része nem írható le $\mathbf{M}$-mel. És fordítva, ha π^* nagy, azt mondhatjuk, hogy nem vagyunk közel a modellhez, mert a legjobb esetben is a populáció egy jó része nem írható le $\mathbf{M}$-mel. Megjegyezzük, hogy P lehet valószínűségi és gyakorisági táblázat is. A második esetben egy becslést kapunk a valódi π^* értékre.

A π^* index több szempontból is előnyösebb a hagyományos mérőszámoknál: szemléletes képet ad az illeszkedés jóságáról, nem függ a mintanagyságtól, továbbá nagy

mértékben általánosítható: kiterjeszthető valószínűségi mértékekre is, így nem csak táblaelemzéseknél, hanem tetszőleges statisztikai modellekre is alkalmazható.

2 Eredmények

A 2. fejezet azoknak az algoritmusoknak a rendszerezése, melyekkel a π^* index kiszámolható. A 3. fejezetben két elméleti problémát vizsgálunk. Az első a π^* feladat, mint optimalizásiós probléma elemzése kétdimenziós táblázatokra a függetlenség modellje mellett. A második a π^* index robusztusságának a kérdése. A 4. fejezetben néhány alkalmazást mutatunk be, melyek különböző tudományágak adatelemzéseiben játszottak szerepet.

2.1 Algoritmusok

A 2. fejezetben négy algoritmus leírása és a π^* problémára való alkalmazása található: az EM, az SQP, a simulated annealing és a minimax algoritmusoké. Ezek MATLAB környezetbe való helyezésével lehetségessé vált tesztelésük és egymással való összehasonlításuk. A minta adatok Rudas, Clogg és Lindsay [42] szemszín-hajszín és jövedelem adatai voltak, ezek újraelemzésén keresztül jól láthatóak a különbségek a fenti algoritmusok között.

Néhány közülük nemcsak kontingencia táblákra, hanem általánosabb esetben is alkalmazható, pl. a minimax algoritmus a logisztikus regressziószámítás modelljére is (lásd Verdes és Rudas [54]). Megjegyezzük, hogy a logisztikus regressziószámításnál nem könnyű az illeszkedés jóságát mérni, lásd [3], [21], [27], ahol különböző eljárásokat vezettek be ennek a feladatnak a megoldására. Ezért érdekes megvizsgálni, hogy a π^* -os módszerrel kapott eredmények hogyan viszonyulnak a fenti irodalmi adatokhoz. A π^* index a logisztikus regressziószámítás illeszkedést mérő mérőszámainak abba az ágába sorolható, ahol a megfigyeléseket csoportosítjuk és ezután a táblaelemzésekhez hasonlóan nézzük a megfigyelt és illesztett gyakoriságok távolságát. A megoldás újdonsága abban áll, hogy 'jó' csoportokat választunk, ugyanis az eredmények szempontjából nagyon fontos, hogy hány csoportot alkotunk és milyen elv szerint soroljuk egy csoportba a megfigyeléseket (lásd [3]). A problémát egy kicsit átfogalmazva az empirikus eloszlás egy jó becslését keressük. Ez a dolgozatban az ASH (Average Shifted Histogram)- mal történik. Az ASH-k olyan simított hisztogramok, ahol egy megfelelő kiindulási hisztogram beosztását finomítjuk, és az így kapott kis intervallumokra olyan téglákat rakunk, amiknek a magassága nemcsak a kis intervallumokba eső megfigyelések számától, hanem, csökkenő súllyal, a közeli és távolabbi szomszédos intervallumokba eső megfigyelések számától is függ. Megmutatható [47], hogy ez a becslés az ASH beosztását finomítva a trianguláris magfüggvényhez tart, így az ASH közvetlenül kapcsolódik a magfüggvényes becslésekhez. Megjegyezzük, hogy az ASH úgy is súlyozható, hogy határértékben más magfüggvényhez, pl. normális magfüggvényhez jussunk.

2.2 Elméleti kérdések

Ahogy már láttuk, több algoritmus is létezik a π^* index kiszámítására. Ezek jobb vagy rosszabb becslést adnak π^*-ra. Ha azonban a megoldandó feladat 'szélsőséges', mint pl. a 3. fejezet végén található példa, mindegyiknek nehézsége támad. Az a sejtésünk, hogy ez azért van, mert a π^* probléma is egy NP teljes feladat, így véges idejű algoritmus nem adható a megoldására. Ezt a sejtést nem tudjuk sem cáfolni sem megerősíteni, de a disszertációban megvizsgáljuk, karakterizáljuk a problémát kétdimenziós kontingencia táblázatokra a függetlenség modellje mellett. Ezt algoritmikus lépéseken keresztül tesszük. A lépések a következők. Használjuk a 2. fejezet EM algoritmusról szóló részének jelöléseit, azaz jelölje $\phi(1), ..., \phi(k), \psi(1), ..., \psi(l)$ a modell tábla marginálisait és $p_n(i,j)$, $i = 1, ..., k$, $j = 1, ..., l$ az empirikus tábla értékeit. Ekkor a π^* indexhez kapcsolódó feltételes szélsőérték számítási feladat a következőképpen fogalmazható meg

$$\max \sum_{i=1}^{k} \sum_{j=1}^{l} \phi(i)\psi(j)$$

$$\phi(i)\psi(j) \leq p_n(i,j), \qquad i = 1, ..., k, \ j = 1, ..., l. \tag{1}$$

Először a $\phi(1), ..., \phi(k), \psi(1), ..., \psi(l)$ sor-oszlop marginálisok egy kezdeti becslését határozzuk meg. Így bizonyos számú egyenlőséghez jutunk (1)-ben. Ha még nincs minden sorban és oszlopban egyenlőség, a marginálisokat átírjuk. A páros gráfokkal meglévő analógiából következik, hogy az optimum helyen az egyenlőségek száma $k + l - 1$, ha ezt még nem értük el, a következő lépésben az egyenlőségek száma növelhető. Végül szükséges és elégséges feltételeket adunk arra nézve, hogy az a megoldás, amit találtunk lokális optimum hely. A szükséges feltétel a Kuhn-Tucker tétel [41] segítségével adható meg. Alkalmazva azt az (1) feladatra, az alábbi eredményhez jutunk.

1. Tétel. *Annak a szükséges feltétele, hogy a* $(\phi^*(1), ..., \phi^*(k), \psi^*(1), ..., \psi^*(l))$ *vektor lokális szélsőérték helye legyen az (1) problémának az, hogy létezzen egy olyan nemnegatív* $k \times l$*-es* W *mátrix, ami csak azokra az* i *-kre és* j*-kre pozitív, ahol* $\phi^*(i)\psi^*(j) = p_n(i,j)$, *továbbá*

$$\sum_{j=1}^{l} w_{ij} = \phi^*(i), \ i = 1, ..., k,$$

és

$$\sum_{i=1}^{k} w_{ij} = \psi^*(j), \ j = 1, ..., l.$$

Az optimalitás egy elégséges feltételét az alábbi tétel fogalmazza meg.

2. Tétel. *Annak egy elégséges feltétele, hogy a* $(\phi^*(1), ..., \phi^*(k), \psi^*(1), ..., \psi^*(l))$ *pont lokális szélsőérték helye legyen az (1) problémának az, hogy*

$$(\log \phi^*(1), ..., \log \phi^*(k), \log \psi^*(1), ..., \log \psi^*(l))$$

globális optimuma legyen az ((1) logaritmizált formája alapján felírt) támasztó lineáris programozási feladatnak.

A π^* indexet a robusztusság szempontjából vizsgálva a π^* egy újfajta interpretációjához jutunk. Tekintsük először a kontaminációs környezet fogalmát, lásd Huber [22]:

$$N(\mathbf{M}, \pi) = \{Q : Q = (1-\pi)M + \pi R,\ M \in \mathbf{M}, R \in \mathbf{P}\},$$

ahol $0 \leq \pi \leq 1$ rögzített és $\mathbf{M} \subset \mathbf{P}$. π itt a kontaminációs szint. Ekkor a $\pi^* = \pi^*(P, \mathbf{M})$ index az alábbi egyenlet legkisebb nemnegatív megoldása

$$d(P, N(\mathbf{M}, \pi)) := \min_{Q \in N(\mathbf{M}, \pi)} d(P, Q) = 0$$

π-ben, ahol d egy távolság mérték a két valószínűségi mérték között, ami ebben a dolgozatban a Kullback-Leibler féle információs divergencia. Az EMF algoritmus, amit a 3. fejezet második felében vezetünk be P és $N(\mathbf{M}, \pi)$ távolságát minimalizálja rögzített π mellett. Majd e minimum értékeket, mint π függvényét rajzoljuk ki, ez lesz az un. kontaminációs görbe, ami épp akkor éri el a 0-t, ha $\pi = \pi^*$. Így egyfelől egy újabb algoritmushoz jutunk, ami π^* értékét számolja, másfelől a kontaminációs görbe fontos információt ad a robusztussággal kapcsolatban. Az EMF algoritmus az EM algoritmus egy módosítása. Az E lépésben megegyeznek. Az M lépésben a standard EM algoritmus a második (nem megszorított) részt érintetlenül hagyja, az EMF algoritmus viszont ezt is optimalizálja (F lépés) az alábbi lemma alapján [11]:

3. Lemma. *Legyen* P *és* T *két mérték az* $(\Omega, \mathcal{A})$ *mérhető téren úgy, hogy* $P(\Omega) = 1$ *és* $0 \leq T(\Omega) \leq 1$. *Az a* Q *mérték, amelyre*

1. $Q(\Omega) = 1$,
2. $Q \geq T$ *(azaz* $Q(A) \geq T(A)$ *minden* $A \in \mathcal{A}$*),*
3. $D(P \parallel Q)$ *minimális*

egyértelmű és abszolút folytonos P*-re nézve. A* $g = \mathrm{d}Q/\mathrm{d}P$ *Radon-Nikodym derivált a* $g = \max\{\varkappa, f\}$ *kifejezéssel adható meg, ahol* $f = \mathrm{d}T/\mathrm{d}P$ *és* $\varkappa$*-t úgy választjuk, hogy* $\int_\Omega g \, \mathrm{d}P = 1$ *teljesüljön.*

Az EMF algoritmus monotonitását Ispány és Verdes [23] igazolta. Alkalmazva ezt az algoritmust is a szemszín-hajszín és a jövedelem adatokra, a 3. fejezet végén vonjuk le a tanulságot, hogy futási idő és az optimális megoldás megtalálása szempontjából hogyan értékelhetjük a meglévő eljárásokat. Jelenlegi tapasztalataink alapján azt állapíthatjuk meg, hogy az EMF algoritmus tekinthető a legjobbnak.

2.3 Alkalmazások

Az utolsó fejezetben öt alkalmazás található, különböző tudományágak (meteorológia, szociológia, pedagógia, orvostudomány) adatainak π^*-os elemzése. Ezek részben olyan táblaelemzések, ahol a minta túl nagy vagy túl kicsi, ezért a hagyományos eljárások nem voltak alkalmazhatók. A problémák egy másik része logisztikus regresszió számításhoz vezetett. Itt az első esetben azért nem a klasszikus eljárással történt a modell illesztése, mert nem mintával, hanem a teljes populációval volt dolgunk (és így a hipotézis vizsgálat általános elve nem érvényesülhetett). A második esetben az illeszkedés jóságát vizsgáltuk a π^* index-szel.

Bibliography

[1] Aarts, E. and Korst, J. (1989) *Simulated Annealing and Boltzmann Machines.* John Wiley: New-York

[2] Agresti, A. (1990) *Categorical data analysis.* John Wiley: New-York

[3] Aldrich, J. H. and Nelson, F. E. (1984) *Linear probability, logit and probit models.* Sage University Papers on Quantitative Applications in the Social Sciences, Sage:CA

[4] Angelusz, R., Tardos, R. (1991) *A "gyenge kötések" ereje és gyengesége.* Gondolat: Budapest

[5] Baran, Á. and Baran, S. (1997) *An application of simulated annealing to ML-estimation of a partially observed Markov chain.* Proceedings of the 3rd InternationalConference on Applied Statistics, 89-95

[6] Bartzokas, A. and Metaxas, D. A. (1996) *Northern hemisphere gross circulation types. Climatic change and temperature distribution.* Meteorol. Zeitschrift, 5, 99-109

[7] Bishop, Y. M. M., Fienberg, S.1E., Holland, P. W.(1975) *Discrete Multivariate Analysis: Theory and Practice.* MIT Press: Cambridge

[8] Blau, P. M. and Duncan, O. D. (1967) *The American occupational structure.* The Free Press

[9] Coleman, J. S. (1988) *Social capital in the creation of human capital* American Journal of Sociology, 94, 95–120

[10] Cramer, H. (1946) *Mathematical methods of statistics.* Princeton University Press: Princeton

[11] Csiszár, I., Ispány, M., Michaletzky, Gy., Rudas, T., Tusnády, G. and Verdes, E. (2001) *Divergence minimization under prior inequality constraints.* Proceedings of the 2001 IEEE International Symposium on Information Theory, 21

[12] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, B 39,1–22

[13] Diaconis, P. and Efron, B. (1985) *Testing for independence in a two way contingency table: new interpretations of the chi square statistics.* The Annals of Statistics, 13, 845–847

[14] Dobó-Nagy, C., Bartha, K., Bernáth, M., Verdes, E. and Szabó, J. (1997) *A comparative study of seven instruments in shaping the root canal in vitro.* International Endodontic Journal, 30, 124-132

[15] Dobó-Nagy, C., Bartha, K., Bernáth, M., Verdes, E. and Szabó, J. (1997) *The effect of root canal morhology on canal shape following instrumentation using different techniques.* International Endodontic Journal, 30, 133-140

[16] Donoho, D. L. and Liu, R. C. (1988) *The "automatic" robustness of minimum distance functionals.* The Annals of Statistics, 16, 552–586

[17] Ferge, Zs. (1980) *A társadalmi struktúra és az iskolarendszer közötti néhány összefüggés.* Gondolat: Budapest

[18] Fényes, H. and Verdes, E. (1999) *Döntés preferálással. Felvételi vizsgák a felsőoktatásban 1967 és 1989 között Magyarországon.* Szociológiai Szemle, 2, 58–76

[19] Finney, D. J. (1947) *The estimation from individual records of the relationship between dose and quantal response.* Biometrika, 34, 320-334

[20] Heritier, S. and Ronchetti, E. (1994) *Robust bounded–influence tests in general parametric models.* JASA, 89, 897–904

[21] Hosmer, D. W. and Lemeshow, S. (1980) *A goodness-of-fit test for the multiple logistic regression model.* Communications in Statistics, A10, 1043-1069

[22] Huber, P. J. (1981) *Robust Statistics.* John Wiley: New-York

[23] Ispány, M. and Verdes, E. (2001) *On robustness of π^* goodness of fit measure.* Journal of Mathematical Sciences, submitted for publication

[24] Juhász, A., Verdes, E., Tőkés, L. and Dobó-Nagy, C. (2001) *The influence of root canal shape on sealing ability of different root canal sealers.* International Endodontic Journal, submitted for publication

[25] Justyák, J. (1994) *Klimatológia.* Tankönyvkiadó: Budapest

[26] Kaufmann, A. (1967) *Graphs, Dynamic Programming and Finite Games.* Academic Press: New-York

[27] McKelvey, R. D. and Zavoina W. (1976) *A statistical model for the ordinal level dependent variables.* Journal of Mathematical Sociology, 4, 103-120

[28] Knoke, D. and Burke, P. J. (1980) *Loglinear models.* Sage: New-York

[29] Kullback, S. (1959) *Information theory and statistics.* John Wiley: New-York

[30] Lange, K. (1998) *Numerical Analysis for Statisticians.* Springer-Verlag: New-York

[31] Luenberger, D. G. (1973) *Introduction to Linear and Nonlinear Programming.* Addison-Wesley: London- New-York

[32] Menard, S. (1991) *Longitudinal research.* Sage: New-York

[33] Narula, S. C. and Wellington, J. F. (1985) *A branch and bound procedure for selection of variables in minimax regression.* SIAM Journal of Statistical Computation, 6, 573-581

[34] Nelder, J. and Wedderburn, R. W. M. (1972) *Generalized Linear Models.* Journal of the Royal Statistical Society, A 135, 370–384

[35] Ostravik, D., Kerekes, K. and Eriksen, H. M. (1993) *Clinical performance of three endodontic sealers.* Endodontics and Dental Traumatology, 3, 178-186

[36] Pearson, K. (1900) *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.* Philosophy Magazine, 50, 157-172

[37] Pigeon, J. G. and Heyse, J. F. (1999) *An improved goodness of fit statistic for probability prediction models.* Biometrical Journal, 41, 71-82

[38] Pigeon, J. G. and Heyse, J. F. (1999) *A cautionary note about assessing the fit of logistic regression models.* Journal of Applied Statistics, 26, 7, 847-853

[39] Pusztai, G. and Verdes, E. (2002) *A társadalmi tőke hatása a felekezeti gimnazisták továbbtanulási terveire.* Szociológiai Szemle, 1, in print

[40] Read, T. R.1C. and Cressie, N. A. C. (1988) *Goodness of fit statistics for discrete multivariate analysis.* Springer: New-York

[41] Rockafellar, R. T. (1970) *Convex Analysis.* Princeton University Press: Princeton, New Jersey

[42] 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 Royal Statistical Society, B 56, 623-639

[43] Rudas, T., Clogg C. C., Matthews, S. (1997) *Analysis of contingency tables using graphical displays based on the mixture index of fit in 'Visualisation of categorical data' ed. by Blasius and Greenacre.* 425–439, Academic Press

[44] Rudas, T. and Zwick, R. (1997) *Estimating the importance of differential item functioning.* Journal of Educational and Behavioral Statistics, 22, 31-45

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

[46] Sen, B. H., Piskin, B. and Baran, N. (1996) *The effect of tubular penetration of root canal sealers on dye microleakage.* International Endodontic Journal, 29, 23-28

[47] Scott, D. W. (1992) *Multivariate density estimation. Theory, practice and visualization.* John Wiley: New-York

[48] Snee, R. (1974) *Graphical display of two way contingency tables.* American Statistician, 38, 9-12

[49] Szelényi, Sz. and Aschaffenburg, K. E. (1993) *Volt-e a szocialista reformoknak eredménye? Osztálykülönbségek az iskolai végzettségben Magyarországon.* Szociológiai Szemle, 2, 71-79

[50] Tar, K., Verdes, E. and Márton, S. (2002) *Temporal change of wind direction field over Hungary.* Időjárás, submitted for publication

[51] Tusnády, G. (1977) *On asymptotically optimal tests.* The Annals of Statistics, 5, 385-393

[52] Verdes, E. (2000) *Finding and characterization of local optima in the π^* problem for two–way contingency tables.* Studia Scientiarum Mathematicarum Hungarica, 36, 471–480

[53] Verdes, E. (2001) *MATLAB Code to Compute the π^* Index for the Examples of the Chapter 12 'A latent class approach to measuring the fit of a statistical model' in Hagenaars, McCutcheon (eds.) : Applied Latent Class Analysis.* Cambridge University Press, in print

[54] Verdes, E. and Rudas, T. (2001) *A new goodness of fit measure for logistic regression.* Springer-Verlag, Ser. Contributions to Statistics, submitted for publication

[55] Weisberg, H. W. (1978) *Evaluating theories of congrassional roll-call voting.* American Journal of Political Science, 22, 554-557

[56] Wood, F. S. (1973) *The use of individual effects and residuals in fitting equations to data.* Technometrics, 15, 677-569

[57] Wu, M. K., Wesselink, P. R., Boersma, J. (1995) *A 1 -year follow-up study on leakage of four root canal sealers at different thiknesses.* International Endodontic Journal, 26, 44-52

[58] Xi, L. (1996) *Measuring goodness of fit in the analysis of contingency tables with mixture based indeces: algorithms, asymptotics and inference.* PhD Thesis, Department of Statistics, the Pennsylvania State University

[59] Zipkin, P. H. (1980) *Simple ranking methods for allocation of one resource.* Manegement Sci., 26, 34–43

Appendix A

Publications

Papers

[1] Túry, F., Wildmann, M., László, Zs., Szabó, P., Murányi, I., Verdes, E., Rathner, G., Dunkel, D. (1997) *Az alkoholizmus epidemiológiája egy magyar faluban.* Appendix, 2, 3-8

[2] Túry, F., Szabó, P., Verdes, E., László, Zs., Wildman, M., Rathner, G., Dunkel, D., Krch, D., Namyslowska, I. (1997) *Az alkoholizmus hazai epidemiológiája egy közép-európai összehasonlító vizsgálat alapján.* Szenvedélybetegségek, 3, 164-173

[3] Nagy, C. D., Bartha, K., Bernáth, M., Verdes, E., Szabó, J. (1997) *A comparative study of seven instruments in shaping the root canal in vitro.* International Endodontic Journal, 2, 124-132

[4] Nagy, C. D., Bartha, K., Bernáth, M., Verdes, E., Szabó, J. (1997) *The effect of root canal morphology on canal shape following instrumentation using different techniques.* International Endodontic Journal, 2, 133-140

[5] Fényes, H., Verdes, E. (1999) *Döntés preferálással. Felvételi vizsgák a felsőoktatásban 1967 és 1989 között Magyarországon.* Szociológiai Szemle, 2, 58-77

[6] Verdes, E. (2000) *Finding and characterization of local optima in the π^* problem for two-way contingency tables.* Studia Scientiarum Mathematicarum Hungarica, 36 , 471-480

[7] Ispány, M., Verdes, E. (2001) *On robustness of π^* goodness of fit measure for contingency tables.* Theory of Probability and its Applications, submitted for publication

[8] Ispány, M., Verdes, E. (2001) *On robustness of π^* goodness of fit measure.* Journal of Mathematical Sciences, submitted for publication

[9] Pusztai, G., Verdes, E. (2002) *A társadalmi tőke hatása a felekezeti gimnazisták továbbtanulási terveire.* Szociológiai Szemle, 1, accepted for publication

[10] Juhász, A., Verdes, E., Tőkés, L., Dobó-Nagy, C. (2001) *The influence of root canal shape on sealing ability of different root canal sealers.* International Endodontic Journal, submitted for publication

[11] Verdes, E., Rudas, T. (2001) *A new goodness of fit measure for logistic regression.* Springer-Verlag, Ser. Contributions to Statistics, submitted for publication

[12] Tar, K., Verdes, E., Márton, S. (2002) *Temporal change of wind direction field over Hungary.* Időjárás, submitted for publication

Books

[1] Verdes, E. (2001) *MATLAB code to compute the π^* index for the examples of Chapter 12 'A latent class approach to measuring the fit of a statistical model' in Hagenaars, McCutcheon (eds.): Applied Latent Class Analysis.* Cambridge University Press, in print

Proceedings

[1] Verdes, E. (1999) *π^* index of fit for the model of logistic regression.* Proceedings of the 4th International Conference on Applied Informatics, 327-330

[2] Verdes, E.(2000) *A new goodness of fit measure for logistic regression.* Proceedings of the Fifth Young Statiticians Meeting, Udine, Italy, 45-54

[3] Csiszár, I., Ispány, M., Michaletzky, Gy., Rudas, T., Tusnády, G. and Verdes, E. (2001) *Divergence minimization under prior inequality constraints.* Proceedings of the 2001 IEEE International Symposium on Information Theory, New-York, 21

Lecture notes

[1] Baran, S., Fazekas, I., Glevitzky, B., Iglói, E., Ispány, M., Kalmár, I., Nagy, M., Tar, L., Verdes, E. (1997) *Bevezetés a matematikai statisztikába.* Kossuth Egyetemi Kiadó, 217-223

Appendix B

Conference talks

[1] Verdes E.: *The π^* index of fit*, 3rd International Conference on Applied Informatics, Noszvaj, Hungary, 1997. augusztus 25-28.

[2] Verdes E.: *Application of a new index of fit for contingency tables*, Large Scale Data Analysis, Cologne, Germany, May 25-28, 1999

[3] Verdes E.: *The π^* index of fit for the model of logistic regression*, 4th International Conference on Applied Informatics, Noszvaj, Hungary, August 30-September 3, 1999

[4] Verdes E.: *The π^* index of fit for the generalized linear models*, Fourth Meeting of Austrian, Slovenian, Italian and Hungarian Young Statisticians, Pécs, Hungary, October 8-10, 1999

[5] Verdes E., Rudas T.: *The π^* index as an alternative for assessing goodness of fit of logistic regression*, Social Science Methodology in the New Millennium, Cologne, Germany, October 3-6, 2000

[6] Verdes E.: *The π^* index as a new alternative for assessing goodness of fit of logistic regression*, Fifth Young Statisticians Meeting, Udine, Italy, October 27-29,2000

[7] Verdes E., Rudas T.: *A new goodness of fit measure for logistic regression, Foundations of Statistical Inference: Applications in the Medical and in the Social Sciences and Industry and the Interface with Computer Science*, Shoresh, Izrael, December 17-23, 2000

[8] Verdes E., Ispány M.: *On robustness of π^* goodness of fit measure for contingency tables, XXI. Seminar on Stability Problems for Stochastic Models*, Eger, Hungary, January 28- February 3, 2001

Appendix C

MATLAB codes

1 MATLAB code for the EM algorithm

```
%obs=[10,20;30,40];
%obs=[3627,1781;1123,3469];
obs=[68,119,26,7;20,84,17,94;15,54,14,10;5,29,14,16];
%obs=[2161,3577,2184,1636;2755,5081,2222,1052;
936,1753,640,306;225,419,96,38;39,98,31,14];
% size of the table
[r,c]=size(obs);
% Normalization
su=0;
for i=1:r
for j=1:c
su=su+obs(i,j);
end
end;
emp=obs/su;
% Initialization
n=r*c; % number of cells
vecemp=reshape(emp,n,1);
mod=zeros(r,c);
```

```
dirt=zeros(r,c);
tic;
num=100; % grid
bound=0.3; % bound
inc=bound/num;
diver=zeros(num,1);
grid=zeros(num,1);
oldmod=emp;
for s=1:num
grid(s)=(s)*inc;
cont=grid(s);
oldmod=emp;
olddirt=(1/n)*ones(r,c);
oldrow=zeros(r,1);
oldcol=zeros(c,1);
diff=1;
count=0;
while diff>0.000001,
row=zeros(r,1);
col=zeros(c,1);
for i=1:r
for j=1:c
row(i)=row(i)+oldmod(i,j);
col(j)=col(j)+oldmod(i,j);
end;
end;
for i=1:r
for j=1:c
mod(i,j)=row(i)*col(j);
end;
end;
mod;
```

```
vecmod=reshape(mod,n,1);
%[div,vecdirt]=filling(vecemp,vecmod,cont);
div=0;
dirt=cont*olddirt;
vecdirt=reshape(dirt,n,1);
for i=1:n
div=div+2*vecemp(i)*log(vecemp(i)/...
((1-cont)*vecmod(i)+vecdirt(i)));
end;
div;
vecmod=reshape(mod,n,1);
vecobs=reshape(obs,n,1);
vecdirt=reshape(dirt,n,1);
for i=1:n
vecoldmod(i)=(vecobs(i)*(1-cont)*vecmod(i))/...
((1-cont)*vecmod(i)+vecdirt(i));
%oldmod=(emp-cont*dirt)/(1-cont);
vecolddirt(i)=(vecobs(i)*vecdirt(i))/...
((1-cont)*vecmod(i)+vecdirt(i));
end;
oldmod=reshape(vecoldmod,r,c);
olddirt=reshape(vecolddirt,r,c);
sum=0;
sumdirt=0;
for i=1:r
for j=1:c
sum=sum+oldmod(i,j);
sumdirt=sumdirt+olddirt(i,j);
end
end;
oldmod=oldmod/sum;
olddirt=olddirt/sumdirt;
```

```
diff=max(norm(oldrow-row),norm(oldcol-col));
oldrow=row;
oldcol=col;
count=count+1;
end;
if (div>0) & (div<0.0001)
rescont=cont
indpart=(1-cont)*mod;
dirtpart=dirt ;
fit=su*indpart
lof=su*dirtpart
end; s;
diver(s)=div;
end;toc;
rescont;
fit=su*indpart;
lof=su*dirtpart;
newplot;
plot(grid,diver), ...  % ,'b*-'), ...
title('Divergence versus contamination plot'), ...
xlabel('Contamination'), ...
ylabel('Divergence'), pause
```

2 MATLAB code for the SQP algorithm

```
function [parameter,picsillag]=hair
%obs=[10,20;30,40];
%obs=[3627,1781;1123,3469];
%obs=[68,119,26,7;20,84,17,94;15,54,14,10;5,29,14,16];
obs=[2161,3577,2184,1636;2755,5081,2222,1052;
936,1753,640,306;225,419,96,38;39,98,31,14];
[r,c]=size(obs);
nh=r*c;
```

```
nvekt=reshape(obs,nh,1);
n=1+r-1+c-1;
for i=1:r
row(i)=0;
end;
f or j=1:c
col(j)=0;
end;
minta=0;
for i=1:r
for j=1:c
row(i)=row(i)+obs(i,j);
col(j)=col(j)+obs(i,j);
minta=minta+obs(i,j);
end;
end;
for i=1:r
for j=1:c
ill(i,j)=row(i)*col(j)/minta;
end;
end;
illvekt=reshape(ill,nh,1);
illvekt=log(illvekt);
% design matrix
dsgn=design22(r,c);
% starting parameter vektor
x=dsgn\ illvekt
%x=zeros(n,1)
%x=unifrnd(-1,1,n,1)
phi=log(nvekt);
options(1)=1;
options(14)=1000;
```

```
tic;
parameter=constr('hairf',x,options,[],[],[],phi,dsgn,nh)
sum=0;
for i=1:nh
sum=sum+exp(dsgn(i,:)*parameter);
il(i)=exp(dsgn(i,:)*parameter);
end;
picsillag=1-sum/minta
toc;
sum;
minta;
mod=reshape(il,r,c)
dirt=reshape(nvekt-il',r,c)
mod=mod/minta;
dirt=dirt/minta;
return


function[f,g]=hairf(x,phi,dsgn,nh)
sum=0;
for i=1:nh
sum=sum-exp(dsgn(i,:)*x);
end;
f=sum;
for i=1:nh
g(i)=dsgn(i,:)*x-phi(i);
end;
return


function[dsgn]=design22(r,c)
nh=r*c;
for i=1:r
```

```
S(i)=0;
end;
for j=1:c
O(j)=0;
end;
k=0;
for i=1:r
S(i)=1;
if i==r,
for ii=1:r-1
S(ii)=-1;
end;
end;
for j=1:c
O(j)=1;
if j==c,
for jj=1:c-1
O(jj)=-1;
end;
end;
k=k+1;
sor=[];
for ii=1:r-1
sor=[sor S(ii)];
end;
oszl=[];
for jj=1:c-1
oszl=[oszl O(jj)];
end;
dsgn(k,:)=[1 sor oszl];
for jb=1:c
O(jb)=0;
```

```
end;
end;
for ib=1:r
S(ib)=0; end;
end;
return
```

3 MATLAB code for the simulated annealing algorithm

```
function [minim,xopt,x1,steps,acstep]=annm
% initialization
%obs=[10,20;30,40];
%obs=[3627,1781;1123,3469];
%obs=[68,119,26,7;20,84,17,94;15,54,14,10;5,29,14,16];
obs=[2161,3577,2184,1636;2755,5081,2222,1052;
936,1753,640,306;225,419,96,38;39,98,31,14];
% size of the table
[r,c]=size(obs);
nh=r*c;
nvekt=reshape(obs,nh,1);
n=1+r-1+c-1;
% computing the marginals
for i=1:r
row(i)=0;
end;
for j=1:c
col(j)=0;
end;
minta=0;
for i=1:r
for j=1:c
row(i)=row(i)+obs(i,j);
```

```
col(j)=col(j)+obs(i,j);
minta=minta+obs(i,j);
end; end;
% fit
for i=1:r
for j=1:c
ill(i,j)=row(i)*col(j)/minta;
end; end;
illvekt=reshape(ill,nh,1);
illvekt=log(illvekt);
% design matrix
dsgn=design22(r,c); % see design22.m above
% starting parameter vector
x=dsgn\illvekt
x0=x';
%x0=[1 1]
%x0=unifrnd(-1,1,1,n); % the starting point
contr1=0;
while contr1==0,
contr1=1;
for i=1:nh,
if dsgn(i,:)*x0'>log(nvekt(i)),
contr1=0;
end;
end;
if contr1==0,
x0(1)=x0(1)*0.8;
end;
end;
v0=0.01*ones(1,n); % the starting step vector
stoppar=3; % the number of successive temperature
% reductions to test for termination
```

```
steps=0; % the number of the performed steps
% during the algorithm
acstep=0; % the number of the accepted steps
dirsuc=zeros(1,n); % the number of succesive trials
% in each direction
xopt=x0 % the optima
phi=fuggvm(x0,dsgn,minta); % the initial value of the function
phiopt=phi; % the optimum
stepvar=10; % a test for step variation
rho=0.995; % a parameter for computing the
% init.  value of the temperature
tempvar=10;
c=2;
phistar=phi*ones(1,stoppar+1);
rat=0.85; q
L=50;
tmult=0.05;
% Determination of the initial temperature
tic;
testlist=[];
a=x0;
phi0=phi;
contr3=0;
while contr3==0,
contr3=1;
for j=1:stepvar,
for i=1:n,
contr1=0;
while contr1==0,
contr1=1;
b=a;
idb=b;
```

```
r=unifrnd(-1,1,1,1);
idb(i)=idb(i)+r*v0(i);
for k=1:nh,
if dsgn(k,:)*idb'>log(nvekt(k)),
contr1=0;
end;
end;
if contr1==1,
b=idb;
end;
end;
phi1=fuggvm(b,dsgn,minta);
testlist=[testlist max([phi1-phi0 0])];
a=b;
phi0=phi1;
end;
end;
T=-sum(testlist)/(length(testlist)*log(rho));
if T==0,
contr3=0;
end;
Tstop=tmult*T;
end;
% Determination of epsilon
phi0=phi;
testlist=[];
for i=1:L,
a=unifrnd(-1,1,1,n);
phi1=fuggvm(a,dsgn,minta);
testlist=[testlist abs(phi1-phi0)];
phi0=phi1;
end;
```

```
epsilon=0.001*(sum(testlist)/L);
% The body of the algorithm
phi0=phi;
contr2=0;
while contr2==0,
for stepadj=1:tempvar,
for cycnum=1:stepvar,
for h=1:n,
i=unidrnd(n,1,1);
contr1=0; % controls the generation of the new points
while contr1==0,
contr1=1;
x1=x0;
xpr1=x1;
r=unifrnd(-1,1,1,1);
xpr1(i)=xpr1(i)+r*v0(i);
for k=1:nh,
if dsgn(k,:)*xpr1'>log(nvekt(k)),
contr1=0;
end;
end;
if contr1==1,
x1=xpr1;
end;
end;
steps=steps+1;
phi1=fuggvm(x1,dsgn,minta);
deltaphi=(phi1-phi0);
if deltaphi<=0,
x0=x1;
phi0=phi1;
acstep=acstep+1;
```

```
dirsuc(i)=dirsuc(i)+1;
if phi1<phiopt,
phiopt=phi1;
xopt=x1;
end;
else
V=unifrnd(0,1,1,1);
p=exp(-deltaphi/T);
if V<p,
x0=x1;
phi0=phi1;
acstep=acstep+1;
dirsuc(i)=dirsuc(i)+1;
if phi1<phiopt,
phiopt=phi1;
xopt=x1;
end;
end;
end;
end;
end;
for i=1:n,
if dirsuc(i)>(0.7*stepvar),
v0(i)=v0(i)*(1+c*(dirsuc(i)/stepvar-0.7)/0.3);
else
if dirsuc(i)<(0.3*stepvar),
v0(i)=(v0(i)*0.3)/(0.3+c*(0.3-(dirsuc(i)/stepvar)));
end;
end;
if v0(i)>15,
v0(i)=v0(i)/10;
end;
```

```
end;
dirsuc=zeros(1,n);
end;
for i=stoppar+1:-1:2,
phistar(i)=phistar(i-1);
end;
phistar(1)=phi1
T=rat*T
check=[]; % checks the stopping rule
for i=2:stoppar+1,
check=[check abs(phistar(1)-phistar(i))<=epsilon];
end;
check=[check ((phistar(1)-phiopt)<=epsilon)];
if sum(check)==stoppar+1,
contr2=1;
end;
if T<Tstop,
contr2=1;
end;
end;
toc;
minim=phiopt;
minta=0;
for i=1:nh
modell(i)=exp(dsgn(i,:)*xopt');
maradek(i)=nvekt(i)-modell(i);
minta=minta+nvekt(i);
end;
[r,c]=size(obs);
mod=reshape(modell,r,c)
dirt=reshape(maradek,r,c)
modell=modell/minta;
```

```
maradek=maradek/minta;
rescont=phiopt/minta
xopt
return
```

4 MATLAB code for the minimax algorithm

```
function[x,pistar]=minihair
%obs=[10,20;30,40];
%obs=[3627,1781;1123,3469];
%obs=[68,119,26,7;20,84,17,94;15,54,14,10;5,29,14,16];
obs=[2161,3577,2184,1636;2755,5081,2222,1052;
936,1753,640,306;225,419,96,38;39,98,31,14];
[r,c]=size(obs);
nh=r*c;
nvekt=reshape(obs,nh,1);
for i=1:r
row(i)=0;
end;
for j=1:c
col(j)=0;
end;
minta=0;
for i=1:r
for j=1:c
row(i)=row(i)+obs(i,j);
col(j)=col(j)+obs(i,j);
minta=minta+obs(i,j);
end; end;
for i=1:r
for j=1:c
ill(i,j)=row(i)*col(j)/minta;
end;
```

```
end;
illvekt=reshape(ill,nh,1);
illvekt=log(illvekt);
% design matrix, see design22.m above
dsgn=design22(r,c);
% initial parameter vector
x0=dsgn\illvekt
options(15)=15;
tic;
x=minimax('minihairf',x0,options,[],[],[],dsgn,nvekt,minta,nh)
for i=1:nh
f(i)=exp(dsgn(i,:)*x)/nvekt(i);
il(i)=exp(dsgn(i,:)*x);
end;
for i=1:nh
B(i)=1/f(i);
end;
pistar=1-min(B)
toc;
mod=reshape((1-pistar)*il,r,c)
dirt=reshape(nvekt-(1-pistar)*il',r,c)
mod=mod/minta;
dirt=dirt/minta;
return


function[f,g]=minihairf(x,dsgn,nvekt,minta,nh)
for i=1:nh
mod(i)=exp(dsgn(i,:)*x);
f(i)=mod(i)/nvekt(i);
end;
sum=0;
for i=1:nh
```

```
sum=sum+exp(dsgn(i,:)*x);
end;
g=[minta-sum,sum-minta];
```

5 MATLAB code for the EMF algorithm

```
%obs=[10,20;30,40];
%obs=[3627,1781;1123,3469];
%obs=[68,119,26,7;20,84,17,94;15,54,14,10;5,29,14,16];
obs=[2161,3577,2184,1636;2755,5081,2222,1052;
936,1753,640,306;225,419,96,38;39,98,31,14];
% size of the table
[r,c]=size(obs);
% Normalization
su=0;
for i=1:r
for j=1:c
su=su+obs(i,j);
end
end;
emp=obs/su
% Initialization
n=r*c; % number of cells
vecemp=reshape(emp,n,1);
mod=zeros(r,c);
dirt=zeros(r,c);
tic;
num=1000; % grid
bound=0.3; % bound
inc=bound/num;
diver=zeros(num,1);
grid=zeros(num,1);
oldmod=emp;
```

```
for s=1:num
grid(s)=(s-1)*inc;
cont=grid(s);
%oldmod=emp;
oldrow=zeros(r,1);
oldcol=zeros(c,1);
diff=1;
while diff>0.000001,
row=zeros(r,1);
col=zeros(c,1);
for i=1:r
for j=1:c
row(i)=row(i)+oldmod(i,j);
col(j)=col(j)+oldmod(i,j);
end;
end;
for i=1:r
for j=1:c
mod(i,j)=row(i)*col(j);
end;
end;
vecmod=reshape(mod,n,1);
[div,vecdirt]=filling(vecemp,vecmod,cont);
dirt=reshape(vecdirt,r,c);
oldmod=(emp.*mod)./((1-cont)*mod+dirt);
%oldmod=(emp-cont*dirt)/(1-cont);
sum=0;
for i=1:r
for j=1:c
sum=sum+oldmod(i,j);
end
end;
```

```
oldmod=oldmod/sum;
diff=max(norm(oldrow-row),norm(oldcol-col));
oldrow=row;
oldcol=col;
end;
if (div>0) & (div<0.0001)
rescont=cont;
indpart=(1-cont)*mod;
dirtpart=dirt;
end;
diver(s)=div;
end;toc;
rescont
fit=su*indpart
lof=su*dirtpart
vecfit=reshape(fit,n,1);
vecfit=log(vecfit);
dsgn=design22(r,c); % see design22.m above
xopt=dsgn\vecfit
newplot;
plot(grid,diver), ...  % ,'b*-'), ...
title('Divergence versus contamination plot'), ...
xlabel('Contamination'), ...
ylabel('Divergence'), pause


function [div,sstar]=filling (s,m,cont)
t=(1-cont)*m;
n=length(s);
for i=1:n
f(i)=t(i)/s(i);
end;
[fstar,I]=sort(f);
```

```
for j=1:n
sstar(j)=s(I(j));
tstar(j)=t(I(j));
end;
S(1)=1-cont;
j = 1;
inc = 0;
while (S(j) < 1) & (j<n),
inc=inc+sstar(j);
S(j+1)=S(j)+(fstar(j+1)-fstar(j))*inc;
j = j + 1;
end;
if j == n
j=n+1;
end;
tossz=0;
if j == 1
sossz=1;
else
sossz=0;
for k=j:n
tossz=tossz+tstar(k);
end;
end;
for k=1:j-1
sossz=sossz+sstar(k);
end;
kappa=(1-tossz)/sossz;
for i=1:j-1
rstar(i)=kappa*sstar(i)-tstar(i);
end;
divup=0;
```

```
for i=j:n
rstar(i)=0;
divup=divup-sstar(i)*log(fstar(i));
end;
div=divup-sossz*log(1-tossz)+sossz*log(sossz);
for i=1:n
sstar(I(i))=rstar(i);
end;
return
```

The π^* index: computation, characterization and application of a new goodness of fit measure

Értekezés a doktori (PhD) fokozat megszerzése érdekében a matematika tudományában.

Írta: Verdes Emese okleveles matematikus, angol-magyar szakfordító

Készült a Debreceni Egyetem Matematika doktori programja (Valószínűségelmélet és matematikai statisztika alprogramja) keretében

Témavezető: Dr. Arató Mátyás

Elfogadásra javaslom: 2002.

A jelölt a doktori szigorlatot 2002. ....-n eredményesen letette.

A bizottság elnöke: Dr. Pap Gyula

.....

Az értekezést bí rálóként elfogadásra javaslom:

Dr. ....

Dr. ....

Dr. ....

A jelölt az értekezést 2002. ....-n sikeresen megvédte:

A bírálóbizottség elnöke: Dr. ....

A bírálóbizotség tagjai:

Dr. ....

Dr. ....

Dr. ....

Debrecen – 2002.