On the robustness of mixture index of fit

Márton Ispány
Faculty of Informatics, University of Debrecen
Pf. 12, H–4010 Debrecen, Hungary
ispany.marton@inf.unideb.hu

Emese Verdes
Department of Measurement and Health Information Systems
World Health Organization
20 av Appia, 1211 Geneva 27, Switzerland
verdese@who.int

January 18, 2014

Abstract

The aim of this paper is to investigate the mixture index of fit in hypothesis testing problems from the point of view of robustness. The concept of contamination plot is introduced and an algorithm is proposed to determine it. Our algorithm is a remarkable application of the widely used EM algorithm by involving a two phases M-step procedure. In the parametric phase the parameters of the model in the null hypothesis are estimated using the maximum likelihood method while in the nonparametric phase the contaminating distribution is determined by a filling technique. It is proved that the objective function decreases monotonically during the iterations. Finally, the algorithm is applied and discussed when the hypothesis of independence is tested for contingency tables.

Keywords. Mixture index of fit; Kullback–Leibler distance; quantitative robustness; EM algorithm; contingency table.

1 Introduction

The problem of evaluating goodness of fit plays central role in testing statistical hypothesis. The following approach is common in parametric statistics. Firstly, the parameters involved in the null hypothesis are estimated, e.g., using the maximum likelihood method. Secondly, the agreement between the model and the data is assessed with a goodness-of-fit statistics. These statistics are usually based on different distance measures between probability distributions. Some examples for these distance measures are the Pearson χ^2, the likelihood

disparity, the Kullback–Leibler divergence, the Hellinger and the Kolmogorov distance, see Donoho and Liu [6] or Read and Cressie [15] for details. It should be remarked that, instead of the maximum likelihood method, the parameters may also be estimated by the minimum distance functional method suggested by Donoho and Liu [6].

Recently, more and more attention has been paid to the robust testing procedures, see the review of Markatou et al. [13]. In general, the above mentioned goodness-of-fit statistics are not robust in the sense that they are sensitive to the circumstances of the hypothesis testing problem. The purpose of this paper is to investigate the π^* goodness-of-fit index or mixture index of fit introduced by Rudas et al. [17] from the point of view of robustness. This goodness-of-fit measure has been widely used, e.g., in the analysis of contingency table, see Rudas et al. [17], Xi [22] and Xi and Lindsay [23], and in the item response theory, see Formann [7] and Revuelta [16].

By empirical studies it will be shown that the mixture index of fit has a kind of automatic robustness. It means that the level of π^* will remain stable under small, arbitrary departures from the null hypothesis for any hypothesis testing problem, i.e., it possesses the property of robustness of validity, see Heritier and Ronchetti [8]. The heart of our treatment is an algorithm based on the EM approach for computing the distance between the model and the observed distribution under different contamination levels and then determining the mixture index of fit. The so-called contamination plot which represents the magnitudes of distance corresponding to different contamination levels is also introduced. We apply this plot to study the robustness of goodness-of-fit measures derived from various distance measures.

The paper is organized as follows. In Section 2, the mixture index of fit and the contamination plot are introduced. Moreover, a new interpretation is also given for the mixture index of fit in the framework of robust statistics. In Section 3, a novel algorithm, the so-called EMF algorithm is described to solve the robust divergence minimization problem. We prove that this algorithm is monotone similarly to the standard EM algorithm. In Section 4, the EMF algorithm is applied to the finite discrete case. We show that the algorithm that we apply to estimate the contaminating distribution coincides with the algorithm RANK developed by Zipkin [20]. Finally, the results are demonstrated in Section 5 by the analysis of eye and hair color data which was investigated earlier by Snee [18], Diaconis and Efron [4], and Rudas et al. [17].

2 The mixture index of fit

Let us consider a statistical space $(\Omega, \mathcal{A}, \boldsymbol{P})$, where the collection of probability measures $\boldsymbol{P}$ on the sample space $(\Omega, \mathcal{A})$ is dominated by a σ–finite measure λ. It is assumed that $\boldsymbol{P}$ contains all sample distributions of interest. In the sequel, the lowercase p denotes the density of the corresponding measure $P \in \boldsymbol{P}$ 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 write shortly $P = \int p \, \mathrm{d}\lambda$. Let $\boldsymbol{M} \subset \boldsymbol{P}$ be a statistical model that we investigate. We would like to test whether it is likely that the distribution which generates the observations belongs to $\boldsymbol{M}$ or not. To evaluate the magnitude of the discrepancy between the model $\boldsymbol{M}$ and the data

various goodness-of-fit statistics can be used.

The most commonly used methods based on, e.g., the χ^2–statistics or the likelihood disparity might not be appropriate to assess goodness of fit when the sample size is very large or small, see Read and Cressie [15]. For example, the hypothesis of row-column independence is usually rejected in testing contingency tables if the sample is sufficiently large. To overcome these difficulties Rudas et al. [17] introduced a new index of fit based on mixtures. This index, called the mixture index of fit and denoted by π^*, is defined as

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

where $P \in \boldsymbol{P}$ is a probability measure. If P_n denotes the empirical measure of the sample of size n, then the $\pi^*(P_n, \boldsymbol{M})$ index measures exactly how far we are from the model $\boldsymbol{M}$ independently of the sample size. The definition of the mixture index of fit can be reformulated in the sense that the density p can be represented as a mixture of two densities of the form $p = (1-\pi)m + \pi r$, where m comes from the model and r is the density of an unrestricted R from $\boldsymbol{P}$. If π^* is small, then we can conclude that P_n is close to the model $\boldsymbol{M}$ because a great proportion of the sample can be described by the model $\boldsymbol{M}$ and only a small proportion of the sample is outside of the model $\boldsymbol{M}$. On the other hand, large value of π^* means that P_n is not close to the model $\boldsymbol{M}$ because only a small proportion of the sample can be described by the model $\boldsymbol{M}$. The mixture index of fit can be derived as the solution of a minimax optimization problem:

$$\pi^*(P, \boldsymbol{M}) = \min_{M \in \boldsymbol{M}} \max_{A \in \mathcal{A}} \left\{ 1 - \frac{P(A)}{M(A)} \right\},$$

see Xi [22] or Liu and Lindsay [12].

A new interpretation associated with the mixture index of fit can be given in the framework of robust statistics. Let d be a generalized distance measure on the space $\boldsymbol{P}$ of probability measures, i.e., we only suppose $d(P,Q) \geq 0$ for all $P, Q \in \boldsymbol{P}$ and $d(P,Q) = 0$ iff $P \equiv Q$. Thus, d is not necessarily a proper metric, neither the symmetry, nor the triangle inequality are assumed. One of the fundamental notion of robust statistics, see Huber [9, page 11], is the contamination neighbourhood defined by

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

where $\pi \in [0,1]$ is fixed and $\boldsymbol{M} \subset \boldsymbol{P}$. We call π the level of contamination. Note that $N(\boldsymbol{M}, \pi)$ is not a neighbourhood in the topological sense and it is the union of the elementary contamination neighbourhoods $N(M, \pi)$, $M \in \boldsymbol{M}$, see Huber [9, formula (4.4)]. Then, the $\pi^* = \pi^*(P, \boldsymbol{M})$ index is the least non-negative solution of the equation

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

in π. We will see in the last section that the function $C(\pi) = d(P, N(\boldsymbol{M}, \pi))$, $\pi \in [0, \pi^*]$, and its graph that we will call contamination plot, also plays an important role during the statistical decisions. Note that $C(\pi^*) = 0$.

There are several possibilities for choosing the distance measure d and this choice is strongly related to the procedure that we apply to measure the goodness of fit. In this paper, 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 $P, Q \in \boldsymbol{P}$ 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 [19]. It should be remarked that Liu and Lindsay [12] also proposed to use the information divergence to handle some undesirable features, e.g. non-differentiablity, of the mixture index of fit.

The following large deviation theory approach for stochastic comparison of tests was suggested by Bahadur ([1], [2]). Let $\{T_n : n \in \mathbb{N}\}$ be a sequence of real–valued statistics for testing the null-hypothesis $H_0 : P \in \boldsymbol{M}$, whereby H_0 is rejected for large values of T_n. In typical cases, the type I error of the test T_n defined by $\alpha_n = \sup_{P \in \boldsymbol{M}} P(T_n \geq t_n)$ tends to zero exponentially fast, where $t_n = T_n(X_1, \ldots, X_n)$ is the observed value of the test statistic based on the sample $X_1, \ldots, X_n$. It is said that the sequence $\{T_n : n \in \mathbb{N}\}$ has the (exact) Bahadur slope $c(P)$ at $P \in \boldsymbol{P}$ if

$$\lim_{n \to \infty} \frac{1}{n} \log \alpha_n = -\frac{1}{2} c(P) \quad P \ a.s.$$

Tusnády [19, Corollary 2] proved that among the tests with fixed Bahadur slope the test $T_n = D(P_n \parallel \boldsymbol{M})$ gives the best attainable exponent of type II error defined by $\beta_n = Q(T_n < t_n)$, where $Q \notin \boldsymbol{M}$ is fixed. He also proved in [19, Theorem 4] that the exact slope of this statistic is $c(Q) = 2D(Q \parallel \boldsymbol{M})$, $Q \in \boldsymbol{P}$.

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 $H_0 : P \in N(\boldsymbol{M}, \varepsilon)$, where ε is small. In this case, we have to investigate the behaviour of the applied test statistics T_n. If T_n is the information divergence, then the figure of the contamination curve in the neighbourhood of zero plays a key role because it completely determines the exact Bahadur slope. If T_n is the mixture index of fit, 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 choosen distance measure, see Donoho and Liu [6]. In the second case, we think that the question is about the behaviour of the contamination function at π^*. It will be demonstrated by numerical studies in the last section that the derivative of this function is approximately zero as the contamination level tends to π^*.

3 Robust divergence minimization by EM algorithm

In this section, an algorithm is given that we can apply for minimizing the divergence between the contaminated model and the empirical measure in that case when the model can be

parametrized by a parameter θ. To be precise, we assume that the model can be written as $\boldsymbol{M} = \{M(\theta), \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$, $d \in \mathbb{N}$. 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 $\int_\Omega f d\lambda = 1$.

Let $P \in \boldsymbol{P}$ be a given probability measure with density p. Our aim is to minimize the information divergence between P and the contaminated model $N(\boldsymbol{M}, \pi)$, 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 a modification of the EM algorithm. Denote by $(\theta^{(k)}, r^{(k)}) \in \Theta \times \Gamma$ the parameters obtained at the kth iteration. Note that $(\theta^{(0)}, r^{(0)})$ is some starting guess for the iteration.

Our EM approach is based on the standard finite mixture model, see McLachlan and Krishnan [14, p. 68]. Suppose that the sample is given as the sum of two latent layers with proportions $1-\pi$ and π. The observations at the first layer come from the model $\boldsymbol{M}$ under unknown parameter $\theta \in \Theta$ while the observations at the second layer come from an unrestricted distribution. In the E-step, at the kth iteration, the proportion of the density p that belongs to the model is calculated as follows:

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

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

$$\theta^{(k+1)} = \arg\min_{\theta \in \Theta} D(M \parallel M(\theta)). \tag{2}$$

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

$$R^{(k+1)} = \arg\min_{R \in \boldsymbol{P}} D(P \parallel (1-\pi)M(\theta^{(k+1)}) + \pi R). \tag{3}$$

While the first minimization phase is the usual maximum likelihood estimation for the parameter θ, the second one is given by a familiar water–filling technique of information theory, see Csiszár et al. [3]. The following lemma of [3], by substituting $T = (1-\pi)M$, shows how the density $r^{(k)}$ can be computed at the kth iteration.

Lemma 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*

(i) $Q(\Omega) = 1$,

(ii) $Q \geq T$, *i.e.,* $Q(A) \geq T(A)$ *for all* $A \in \mathcal{A}$,

(iii) $D(P \parallel Q)$ *is minimal*

is unique and absolute continuous with respect to 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.*

The prior inequality constraint (ii) can be replaced by more constraints such as $Q \geq T_i$, $i = 1, \ldots, r$. If $P(\operatorname{supp} \min\{f_1, \ldots, f_r\}) > 0$, where $f_i = \mathrm{d}T_i/\mathrm{d}P$, then the density of the unique solution is given by $g = \max\{\varkappa, \min\{f_1, \ldots, f_r\}\}$. One can also see that $T_i's$ may be the extremal points of a convex set of probability measures. We should also note that the minimizer will be the same under general γ-divergences, see Csiszár et al. [3] or Huber and Strassen [10, Theorem 6.1].

It is well known that the EM algorithm is an ascent algorithm, see Dempster et al. [5] or McLachlan and Krishnan [14, p. 83]. We prove that our algorithm possesses this property in the sense that the divergence $D_\pi(\theta, r)$ does not increase after an iteration. We start with a lemma which plays fundamental role in the proof and then we state our main theorem.

Lemma 2. *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space and let* f_1, f_2, *and* g *be non-negative functions on it. Then*

$$\int_\Omega \log \frac{f_1 + g}{f_2 + g} \,\mathrm{d}P \leq \int_\Omega \frac{f_1}{f_1 + g} \log \frac{f_1}{f_2} \,\mathrm{d}P.$$

Proof. By Jensen's inequality we have

$$\log \frac{f_2 + g}{f_1 + g} = \log \left(\frac{f_1}{f_1 + g} \cdot \frac{f_2}{f_1} + \frac{g}{f_1 + g} \cdot 1 \right) \geq \frac{f_1}{f_1 + g} \log \frac{f_2}{f_1}.$$

Thus, by integrating, we obtain the desired inequality.

Theorem 1. *The iterates defined by* (1), (2) *and* (3) *obey*

$$D_\pi(\theta^{(k+1)}, r^{(k+1)}) \leq D_\pi(\theta^{(k)}, r^{(k)})$$

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

Proof. It is enough to prove that

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

Here, the first term is non–positive by (3). For the second term we have

$$D_\pi(\theta^{(k+1)}, r^{(k)}) - D_\pi(\theta^{(k)}, r^{(k)}) = \int_\Omega \log \frac{(1-\pi)m(\theta^{(k)}) + \pi r^{(k)}}{(1-\pi)m(\theta^{(k+1)}) + \pi r^{(k)}} \,\mathrm{d}P. \tag{4}$$

On the other hand, by (2) we obtain

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

Let us apply Lemma 2 with the choice $f_1 = (1-\pi)m(\theta^{(k)})$, $f_2 = (1-\pi)m(\theta^{(k+1)})$, and $g = \pi r^{(k)}$. Then, one can see that the right hand side of (4) 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 minimum, see the example in Verdes [21]. In general, if D_π has several local minimum points, the sequence of iterates defined by (1), (2), and (3) depends on the choice of the 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 iterate converges to a unique D^* irrespective of its starting point.

4 The algorithm in the finite discrete case

In this section we suppose that the sample space is finite and, for simplicity, $\Omega = \{1, \dots, N\}$ and $\mathcal{A} = 2^\Omega$. Then all probability measures $P \in \boldsymbol{P}$ can be identified with their density p with respect to the counting measure. Let $X_1, \dots, X_n$ be a sample for the random variable X on the statistical space $(\Omega, \mathcal{A}, \boldsymbol{P})$. The empirical measure p_n associated with the sample is defined by

$$p_n(i) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{X_j = i\}}, \qquad i = 1, \dots, N,$$

where $\mathbb{1}_A$ denotes the indicator of an event $A \in \mathcal{A}$. Moreover, denote by $m(\theta)$, $\theta \in \Theta$, the collection of distributions belonging to the model $\boldsymbol{M}$, and let the level of contamination π be fixed. We adopt the procedure described in the previous section to minimize the objective function $D_\pi(\theta, r)$ for this discrete case.

We start the algorithm at the initial point $(\theta^{(0)}, r^{(0)})$ that we specify later. Suppose $(\theta^{(k)}, r^{(k)})$, $k \in \mathbb{N}$, is determined after the kth iteration. At the E-step of the next iteration we compute the expected model distribution divided the empirical one p_n into two parts according to the rate defined by the distributions $m(\theta^{(k)})$ and $r^{(k)}$ of the previous iteration:

$$m(i) = \frac{(1-\pi)m(\theta^{(k)}, i)}{(1-\pi)m(\theta^{(k)}, i) + \pi r^{(k)}(i)} \cdot p_n(i), \qquad i = 1, \dots, N. \tag{5}$$

In general, the normalization of m does not even belong to the model $\boldsymbol{M}$. Thus, at the M-step we first compute the likelihood projection of m to the model $\boldsymbol{M}$, i.e., $\theta^{(k+1)}$ is the maximum likelihood estimator of the parameter θ replacing the empirical distribution by m. Hence, $\theta^{(k+1)}$ is given by

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

At the second phase of the M-step, in order to determine the contaminating distribution $r^{(k+1)}$, 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

$$\sum_{i=1}^{N} p_n(i) \log((1-\pi) m(\theta^{(k+1)}, i) + \pi r(i)). \tag{6}$$

The solution of this problem is given by Lemma 1. Let $t(i) = (1-\pi) m(\theta^{(k+1)}, i)$, $i = 1, \ldots, N$, and define the numbers $\widetilde{r}(i)$, $i = 1, \ldots, N$, by the following way:

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

where the constant $\varkappa$ 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 contaminating distribution $r^{(k+1)}$ is given by $r^{(k+1)}(i) = \widetilde{r}(i)/\pi$, $i = 1, \ldots, N$. In practice, the computation of $r^{(k+1)}$ is based on the well-known filling technique similarly, for example, to the case of the separable resource allocation problem investigated by Zipkin [20], see the algorithm RANK in Ibaraki and Katoh [11, Section 2.2]. Because of this approach the following algorithm is suggested. At the first step, determine the ratios $f(i) = t(i)/p_n(i)$, $i = 1, \ldots, N$. Then, order $f(i)$'s to get $f^*(i)$'s, and denote by $t^*(i), p_n^*(i)$ the rearrengement of the sequences $t(i), p_n(i)$, $i = 1, \ldots, N$, according to this ordering. Let $\Sigma(1) = \sum_{i=1}^{N} t(i) = 1 - \pi$ and define the sequence $\Sigma(j)$ recursively by

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

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 $j^* = N + 1$. Then the constant $\varkappa$ can be derived as

$$\varkappa = \left(1 - \sum_{j=j^*}^{N} t^*(j)\right) \Bigg/ \sum_{j=1}^{j^*-1} p^*(j),$$

and the contanimating distribution can be calculated as

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

For initial values $\theta^{(0)}$ and $r^{(0)}$ we may choose the maximum likelihood estimator of the parameter in the contamination proof model and the uniform distribution, respectively.

The modified EM algorithm of this section consists of three steps: the (**E**) expectation step defined by (5), the (**M**) maximization step, where the likelihood function (6) is maximized, and the (**F**) filling step, where the contaminating distribution is computed. Hence, it is referred to this algorithm as the EMF algorithm.

5 Application to contingency tables

There were different algorithms suggested to compute the mixture index of fit for contingency tables. Rudas et al. [17] has been proposed to use the standard EM algorithm, Xi [22] has been applied the SQP (sequential quadratic programming) and Verdes [21] has been suggested to use a greedy algorithm based on the Karush-Kuhn-Tucker theorem. In this section, we adopt the EMF algorithm introduced in the previous section for computing the mixture index of fit when the row-column independence is supposed in the model.

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

The following algorithm is applied for determining the contamination curve. First, suppose that the contamination level π is fixed and the initial values $\theta^{(0)} = (\phi^{(0)}, \psi^{(0)})$, $r^{(0)}$ are obtained in some way or other. An iteration of the general EMF algorithm is reduced to the following simple form. At the E-step of the $(k+1)$th iteration, for the first latent layer, we set

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

At the M-step, the maximum likelihood estimator 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 the summation with respect to the corresponding argument. Finally, at the F-step, the filling algorithm of the previous section is used with the observed distribution $\{p_n(i,j)\}$ and the lower bounds of the prior constraints are given by $t(i,j) = (1-\pi)\phi^{(k+1)}(i)\psi^{(k+1)}(j)$, $i = 1, \dots, k$, $j = 1, \dots, \ell$. The calculations are performed with a relative error equal to 10^{-8}.

To draw the contamination plot take an enough fine grid on the unit interval $[0,1]$, i.e., divide it into K equal parts (in our case $K = 10^3$), and compute the information divergence

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

Table 1: Cross-classification of eye color and hair color

Eye color	Hair color							
	Black		Brunette		Red		Blonde	
Brown	28.33	39.67	119	0	24.09	1.91	7	0
Blue	20	0	84	0	17	0	4.94	89.6
Hazel	12.85	2.15	54	0	10.93	3.07	3.18	6.82
Green	5	0	21	8	4.25	9.75	1.24	14.76

Table 2: Decomposition of Table 1 (left: independent, right: contamination)

between the empirical distribution and the distribution given by the above iteration. We only have to find an optimal guess at each contamination level. We have proceeded as follows. At zero contamination level independently of the starting values $\phi^{(0)}$ and $\psi^{(0)}$ we have $\phi^{(k)} = \{p_n(i,+)\}$ and $\psi^{(k)} = \{p_n(+,j)\}$ for all $k = 1, 2, \ldots$. Then, at a given contamination level, let the initial values $\phi^{(0)}$ and $\psi^{(0)}$ be the results of the iteration at the previous contamination level. In this case, a faster algorithm is obtained instead of the case when we always start from the marginal distributions of p_n.

For numerical studies we consider the contingency table given by the cross-classification of eye and hair color which was analyzed earlier by Snee [18], Diaconis and Efron [4], and Rudas et al. [17], see Table 1. For the mixture index of fit 0.2961 is obtained which is smaller than the Rudas' one (0.298) given by the standard EM algorithm. We think that the Rudas' algorithm is susceptible to overestimate the π^* index. The decomposition of Table 1 into independent and contaminating parts is given by Table 2. Figure 1 represents the contamination plot for this contingency table. One can see that the contamination function is a monotone decreasing convex function. We conjecture that this fact remains true in the most of cases, but we have also found a counterexample. This plot also shows that the tangent of the contamination curve is the horizontal axis at π^* index justifying the robustness of the mixture index of fit. 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 zero. It is well known that in case of large divergence at

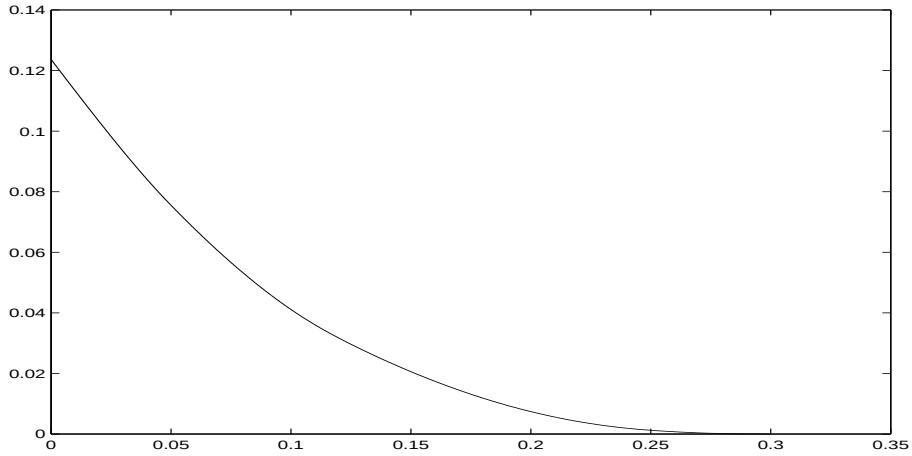

Figure 1: Contamination plot for eye-hair color table

zero the hypothesis of independence is rejected. If the contamination function dies down rapidly, then the null hypothesis is already accepted under small departure from the model, where the independence is supposed. For this contingency table the slope of the tangent of the contamination function at zero equals approximately to $-1/2$. This justifies the robustness of the test statistics based on information divergence. Another useful measure for robustness can be given by the ratio of the area under the contamination curve and the area of the triangle given by joining the points $(0, C(0))$ and $(\pi^*, 0)$. If this number is small, then the contamination function dies down rapidly. For the Table 1 this number equals 0.5563. Finally, we remark that the robustness of other goodness-of-fit measures can be investigated by similar manner.

Acknowledgement

M. Ispány has been supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund. The authors would like to express sincere appreciation to the members of the π^*–group, I. Csiszár, G. Tusnády, T. Rudas and Gy. Michaletzky for their suggestions and remarks.

References

[1] Bahadur, R. R. (1960). Stochastic comparison of tests. *Ann. Math. Stat.* **31**, 276–295.

[2] Bahadur, R. R. (1971). *Some Limit Theorems in Statistics.* SIAM, Philadelphia.

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

[4] Diaconis, P. and Efron, B. (1985). Testing for independence in a two-way contingency table: new interpretations of the chi–square statistics. *Ann. Stat.* **13**, 845–874.

[5] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B* **39**, 1–22.

[6] Donoho, D. L. and Liu, R. C. (1988). The 'automatic' robustness of minimum distance functionals. *Ann. Stat.* **16**, 552–586.

[7] Formann, A. K. (2006). Testing the Rasch model by means of the mixture fit index. *Brit. J. Mat. Stat. Psy.* **59**, 89–95.

[8] Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *J. Am. Stat. Assoc.* **89**, 897–904.

[9] Huber, P. J. (1981). *Robust Statistics.* J. Wiley & Sons, New York.

[10] Huber, P. J. and Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Stat.* **1**, 251–263.

[11] Ibaraki, T. and Katoh, N. (1988). *Resource Allocation Problems. Algorithmic Approaches.* Foundation of Computing Series. MIT Press, Cambridge.

[12] Liu, J. and Lindsay, B. G. (2009). Building and using semiparametric tolerance regions for parametric multinomial models. *Ann. Stat.* **37**, 3644–3659.

[13] Markatou, M., Stahel, W. A. and Ronchetti, E. (1991). Robust M-type testing procedures for linear models. in *Directions in Robust Statistics and Diagnostics I* (eds. Stahel, W. and Weisberg, S.), Springer, New York, pp. 201–220.

[14] McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley & Sons, New York.

[15] Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Analysis*, Springer, New York.

[16] Revuelta, J. (2008). Estimating the π^* goodness of fit index for finite mixtures of item response models. *Brit. J. Mat. Stat. Psy.* **61**, 93–113.

[17] Rudas, T., Clogg, C. C. and Lindsay, B. G. (1994). A new index of fit based on mixture method for the analysis of contingency tables. *J. Roy. Stat. Soc. B* **56**, 623–639.

[18] Snee, R. (1987). Graphical display of two-way contingency table. *Am. Stat.* **38**, 9–12.

[19] Tusnády, G. (1977). On asymptotically optimal tests. *Ann. Stat.* **5**, 385-393.

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

[21] Verdes, E. (2000). Finding and characterization of local optima in the π^* problem for two-way contingency tables. *Stud. Math. Hung.* **36**, 471–480.

[22] Xi, L. (1996). Measuring goodness-of-fit in the analysis of contingency tables with mixture based indices: Algorithms, asymptotics and inference. *PhD dissertation, Penn. State Univ.*

[23] Xi, L. and Lindsay, B. G. (1996). A note on calculating the π^* index of fit for the analysis of contingency tables. *Sociol. Method Res.* **25**, 248–259.