STATISTICAL METHODS FOR PROBABILISTIC MODEL PARAMETER ESTIMATION FROM INCOMPLETE DATA AND THEIR APPLICATION TO THE MARGINAL PROBLEM

Jiří Vomlel

1. INTRODUCTION

This paper presents an unfinished work. Its objective is to describe properties of procedures proposed for the marginal problem solution using their analogy with some methods used in statistics. In the first part we give a brief survey of several statistical methods used for probabilistic model parameter estimation given incomplete data. In the second part of this paper two special cases of these methods are applied to the marginal problem. For simplicity the theory is given for the case of three dichotomic random variables. Generalization to more variables is straightforward.

2. PRELIMINARIES

Let A, B, C be three dichotomic random variables taking values from $\mathcal{X}_\mathcal{A}, \mathcal{X}_\mathcal{B}, \mathcal{X}_\mathcal{C}$ respectively. Further let $\mathbf{x} = (x_{ijk})_{i,j,k=1}^2$ be a vector of counts for all possible combinations of values from $\mathcal{X} = \mathcal{X}_\mathcal{A} \times \mathcal{X}_\mathcal{B} \times \mathcal{X}_\mathcal{C}$ and $\mathbf{m} = E(\mathbf{x}) = (m_{ijk})_{i,j,k=1}^2$ be its expected value. The vector $\mathbf{x}$ can be viewed as a contingency table. Suppose that the set of $N = \sum_{i,j,k=1}^{2} x_{ijk}$ observations is a random sample from an infinite population where the underlying cell probabilities are m_{ijk}/N and the distribution function is *multinomial*:

$$f(\mathbf{x}|\mathbf{m}) \quad = \quad N! \cdot N^{-N} \cdot \prod_{i,j,k=1}^{2} \frac{m_{ijk}^{x_{ijk}}}{x_{ijk}!}.$$

A *log-linear model* is specified by a generating class G of subsets of the set $\{A, B, C\}$ representing the restriction that $\log \mathbf{m} = \sum_{g \in G} u^g$. Then a *saturated log-*

linear model can be defined by a system of equations:

$$\log m_{ijk} = u^{\emptyset} + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC}, \quad i,j,k = 1,2$$

Particular models are defined by setting appropriate terms $u_g = 0$.

Estimation problem can typically be structured in terms of a "distance" measure $K(\mathbf{x}, \mathbf{m})$, where parameter estimates $\hat{\theta}$ are chosen so that the distance between $\mathbf{x}$ and $\mathbf{m} = \mathbf{m}(\hat{\theta})$ is minimized. Most often used measures are: *Minimum chi-square*, *Minimum discriminatory information*, and *Maximum likelihood.*

Minimum discriminatory information which is also called *Maximum entropy* is defined:

$$g_a^2(\mathbf{x}, \mathbf{m}) = 2 \cdot I(\mathbf{x}||\mathbf{m}) = 2 \cdot \sum_{i,j,k=1}^{2} x_{ijk} \cdot \log \frac{x_{ijk}}{m_{ijk}} = K_1 - 2 \cdot \sum_{i,j,k=1}^{2} x_{ijk} \cdot \log m_{ijk},$$

where K_1 is the constant independent of $\mathbf{m}$ and $I(\mathbf{x}||\mathbf{m})$ is usually called *Kullback-Leibler divergence.*

Within *maximum likelihood* approach an estimated value of m_{ijk} is chosen so that the given outcome is the most probable one. In the case of multinomial sampling we maximize over $\mathbf{m}$ log-likelihood function

$$L(\mathbf{m}|\mathbf{x}) = \log(\frac{N!}{N^N} \cdot \prod_{i,j,k=1}^{2} \frac{m_{ijk}^{x_{ijk}}}{x_{ijk}!}) = K_2 + \sum_{i,j,k=1}^{2} x_{ijk} \cdot \log m_{ijk},$$

where K_2 is the constant independent of $\mathbf{m}$.

We can see that in order to *maximize likelihood* function $L(\mathbf{m}|\mathbf{x})$ we need to maximize $\sum_{i,j,k=1}^{2} x_{ijk} \cdot \log m_{ijk}$, which is equivalent to *minimum discriminatory information* criteria. The duality theorem in [7] indicates that for the three basic sampling models for contingency tables (Poison , Multinomial and Product Multinomial models), *minimizing discriminatory information* is equivalent to *maximizing likelihood* provided that

$$\sum_{i,j,k=1}^{2} m_i(\hat{\theta}) = \sum_{i,j,k=1}^{2} x_i = N.$$

3. MAXIMUM LIKELIHOOD FOR COMPLETE DATA

The most widely used numerical procedure for the calculation of maximum likelihood estimates for log-linear models is the *Iterative Proportional Fitting Procedure* (IPFP) included e.g. in BMDP statistical package [6]. This method provides only estimated expected values but no estimate of its variance-covariance matrix.

Other numerical approaches are based on classical procedures for solving non-linear equations e.g. *Newton-Raphson method* or other gradient methods. These methods produce both estimated expected values and an estimate of the asymptotic covariance matrix. They are faster but tend to be unstable and require the analytical computation of likelihoods to derive expressions for the score function and Fisher information matrix.

The Iterative Proportional Fitting Procedure (IPFP) is derived from the algorithm often associated with [4], where it was used for an adjustment of cells of a contingency table to fit a set of margins.

General description of history and properties of IPFP along with iterative proportional fitting of log-linear models is given e.g. in [1] pp. 83–102.

Let us show how IPFP works on the example of hierarchical log-linear model with no second order (three-factor) interaction between variables:

$$\log m_{ijk} \quad = \quad u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}. \tag{1}$$

IPFP repeats following three steps:

1) $\hat{m}^{(0)}_{ijk} = 1 \quad i,j,k = 1,2$
We can associate any values for which the interaction structure is no more complex than the model being fitted, but values of 1 are always fine.

2) Adjust $\hat{m}^{(t+1)}_{ijk}$ so that they match each marginal table in the set of minimal sufficient statistics. In the case of our model sufficient statistics are $x_{ij\cdot}, x_{i\cdot k}, x_{\cdot jk}$ and adjustment looks like:

$$\begin{aligned}
\hat{m}^{(t+1)}_{ijk} &= \hat{m}^{(t)}_{ijk} \cdot \tfrac{x_{ij\cdot}}{\hat{m}^{(t)}_{ij\cdot}} \quad i,j,k = 1,2 \quad t = 3r \\
\hat{m}^{(t+1)}_{ijk} &= \hat{m}^{(t)}_{ijk} \cdot \tfrac{x_{i\cdot k}}{\hat{m}^{(t)}_{i\cdot k}} \quad i,j,k = 1,2 \quad t = 3r+1 \\
\hat{m}^{(t+1)}_{ijk} &= \hat{m}^{(t)}_{ijk} \cdot \tfrac{x_{\cdot jk}}{\hat{m}^{(t)}_{\cdot jk}} \quad i,j,k = 1,2 \quad t = 3r+2
\end{aligned}$$

3) Continue adjusting until the maximum difference between the sufficient statistics and their fitted values is sufficiently close to zero (measured e.g. by the Kullback-Leibler divergence):

$$\max\{\sum_{i,j=1}^{2} x_{ij\cdot} \cdot \log \frac{x_{ij\cdot}}{\hat{m}^{(t+1)}_{ij\cdot}}, \sum_{i,k=1}^{2} x_{i\cdot k} \cdot \log \frac{x_{i\cdot k}}{\hat{m}^{(t+1)}_{i\cdot k}}, \sum_{j,k=1}^{2} x_{j\cdot k} \cdot \log \frac{x_{\cdot jk}}{\hat{m}^{(t+1)}_{\cdot jk}}\} \quad \cong \quad 0.$$

A convergence proof of IPFP described in [1, 2] is based on the fact that Kullback-Leibler divergence between cell frequencies in a contingency table and their estimates in each step of this procedure $I(\mathbf{x}||\hat{\mathbf{m}}^{(t)})$ is monotone decreasing function of t.

4. MISSING DATA METHODS

Following discussions suppose that data are MCAR i.e. missing completely at random which means that missingness is not related to the variables under a study.

There are several MCAR methods:

- **Complete-Case Analysis**. In this case we take a subsample of the original cases by omitting all incomplete observations. This approach is unsatisfactory particularly if there is a lot of incomplete observations.

- **Available-Case Methods**. This method can be used e.g. in univariate analysis. It includes all cases where all variables of our interest are present (other variables need not to be present). Therefore we use all the available values. The disadvantage is that the sample base changes from variable to variable according to the pattern of missing data. In addition e.g. estimation of correlation using this method can yields contradictory results.

- **Fill-In Methods**. Fill-In methods tempt to predict missing value from the values of other variables supposing they are correlated. Then it includes imputed value in analysis.

  Imputing unconditional means. Values filled-in for a variable are means of the present values of that variable. This method is not generally satisfactory. E.g. it systematically underestimates correlations.

  Imputing conditional means. This method substitutes means that are conditioned on the variables recorded in an incomplete case. An example of this approach is Buck's method [3] used in the case of multivariate normally distributed variables. The method first estimates mean and covariance matrix from the sample mean and covariance matrix based on the complete cases, and then uses these estimates to calculate the linear regressions of the missing variables on the present variables, case by case. This method is closely related to one iteration of EM-algorithm.

- **EM-algorithm** This algorithm is described in the Section 6..

5. AN EXAMPLE OF INCOMPLETE DATA

Let there is a subsample of sample space that is fully categorized and let y_{ijk}^{ABC} be the observed frequency in the cell (i, j, k) (see Table 1).

	$j=1$		$j=2$	
	$k=1$	$k=2$	$k=1$	$k=2$
$i=1$	y_{111}^{ABC}	y_{112}^{ABC}	y_{121}^{ABC}	y_{122}^{ABC}
$i=2$	y_{211}^{ABC}	y_{212}^{ABC}	y_{221}^{ABC}	y_{222}^{ABC}

Table 1: Fully categorized subsample.

Further let there are three subsamples each observed only on two variables (A, B), (B, C) resp. (A, C). Those subsamples define supplemental tables with frequencies y_{ij}^{AB}, y_{jk}^{BC} resp. y_{ik}^{AC} (see Table 2).

	$j=1$	$j=2$
$i=1$	y_{11}^{AB}	y_{12}^{AB}
$i=2$	y_{21}^{AB}	y_{22}^{AB}

	$k=1$	$k=2$
$i=1$	y_{11}^{AC}	y_{12}^{AC}
$i=2$	y_{21}^{AC}	y_{22}^{AC}

	$k=1$	$k=2$
$j=1$	y_{11}^{BC}	y_{12}^{BC}
$j=2$	y_{21}^{BC}	y_{22}^{BC}

Table 2: Subsamples categorized only for two variables.

Let there are also three additional subsamples each observed only on one variable A, B resp. C with frequency tables y_i^A, y_j^B resp. y_k^C (see Table 3).

$i=1$	y_1^A
$i=2$	y_2^A

$j=1$	y_1^B
$j=2$	y_2^B

$k=1$	y_1^C
$k=2$	y_2^C

Table 3: Subsamples categorized only for one variable.

A vector $\mathbf{y}$ describing incomplete data

$$\mathbf{y} = ((y_{ijk}^{ABC})_{i,j,k=1}^2, (y_{ij}^{AB})_{i,j=1}^2, (y_{jk}^{BC})_{j,k=1}^2, (y_{ik}^{AC})_{i,k=1}^2, (y_i^A)_{i=1}^2, (y_j^B)_{j=1}^2, (y_k^C)_{k=1}^2)$$

contains counts of all possible combinations of values of random variables for all patterns of missing data. By the definition these counts are distributed according to the function

$$g(\mathbf{y}|\mathbf{m}) = \sum_{\mathbf{x} \in \mathbf{X}(\mathbf{y})} f(\mathbf{x}|\mathbf{m}),$$

where $\mathbf{X}(\mathbf{y})$ is the set of all vectors $\mathbf{x}$ which could be a completion of $\mathbf{y}$.

6. THE EM ALGORITHM

The EM algorithm was firstly introduced in [5] as a broadly applicable algorithm for computing maximum likelihood estimates from incomplete data. Generally every iteration of the EM Algorithm consists of two steps: *expectation step* and *maximization step*.

At the expectation step the algorithm forms the conditional expectation of the log-likelihood function for complete data $\mathbf{x}$ given the observed data $\mathbf{y}$:

$$Q(\theta|\theta^{(q)}) = E(\log f(\mathbf{x}|\theta)|\theta^{(q)}, \mathbf{y}).$$

At the maximization step a value of parameter $\theta = \theta^{(q+1)}$ maximizing $Q(\theta|\theta^{(q)})$ is found. The EM algorithm alternates between E-step and M-step using parameter θ estimate from the previous step.

The EM algorithm has a generalization called the GEM algorithm. In M-step it does not maximize $Q(\theta|\theta^{(q)})$ but only finds a value of θ such that function $Q(\theta|\theta^{(q)})$ is strictly increasing with q.

To prove that $L(\theta^{(q+1)}) \geq L(\theta^{(q)})$ Dempster *et al.* [5] use a function

$$H(\theta|\theta^{(q)}) = E(\log \frac{f(\mathbf{x}|\theta)}{g(\mathbf{y}|\theta)}|\mathbf{y}, \theta^{(q)}),$$

which according to Jensen's inequality satisfies:

$$H(\theta|\theta^{(q)}) \quad \leq \quad H(\theta^{(q)}|\theta^{(q)}). \tag{2}$$

Than we can write:

$$L(\theta^{(q+1)}) \quad = \quad Q(\theta^{(q+1)}|\theta^{(q)}) - H(\theta^{(q+1)}|\theta^{(q)}) \tag{3}$$

$$L(\theta^{(q)}) \quad = \quad Q(\theta^{(q)}|\theta^{(q)}) - H(\theta^{(q)}|\theta^{(q)}). \tag{4}$$

From the M-step of the EM algorithm (resp. the GEM algorithm) follows that:

$$Q(\theta^{(q+1)}|\theta^{(q)}) \quad \geq \quad Q(\theta^{(q)}|\theta^{(q)}). \tag{5}$$

Using inequalities (2) and (5) the result of subtraction (4) from (3) must be always nonnegative:

$$L(\theta^{(q+1)}) - L(\theta^{(q)}) \geq 0.$$

Supposing that distribution $f(\mathbf{x}|\theta)$ belongs to an exponential family the EM algorithm can be simplified.

- **Expectation step.** It adjust the values of the sufficient statistics of a log-linear model, given the incomplete data and the current value of the parameters.

$$\mathbf{t}^{(q)}(x) \quad = \quad E(\mathbf{t}(x)|\hat{\mathbf{m}}^{(q)}, \mathbf{y})$$

- **Maximization step.** This step uses the adjusted values of the sufficient statistics $\mathbf{t}^{(q)}(x)$ to find maximum likelihood estimates of $\hat{\mathbf{m}}^{(q+1)}$ for the log-linear model solving equations:

$$E(\mathbf{t}(x)|\hat{\mathbf{m}}^{(q+1)}) \quad = \quad \mathbf{t}^{(q)}(x).$$

To show how the EM Algorithm works we will again use our example with log-linear model given by formula (1). Then $x_{ij\cdot}$, $i,j = 1,2$, $x_{\cdot jk}$, $j,k = 1,2$ and $x_{i\cdot k}$, $i,k = 1,2$ are the minimal sufficient statistics for this model. Suppose there are three subsamples each observed only on two variables (A,B), (B,C) resp. (A,C). Those subsamples define supplemental tables with frequencies y_{ij}^{AB}, y_{jk}^{BC} resp. y_{ik}^{AC}.

At each step q the EM Algorithm repeats following two steps:
1)**E-step**

$$\begin{aligned}
x_{ij\cdot}^{(q)} &= y_{ij\cdot}^{ABC} + y_{ij}^{AB} + \sum_{k=1}^{2} y_{ik}^{AC} \cdot \frac{\hat{m}_{ijk}^{(q)}}{\hat{m}_{i\cdot k}^{(q)}} + \sum_{k=1}^{2} y_{jk}^{BC} \cdot \frac{\hat{m}_{ijk}^{(q)}}{\hat{m}_{\cdot jk}^{(q)}} \quad i,j = 1,2 \\
x_{\cdot jk}^{(q)} &= y_{\cdot jk}^{ABC} + y_{jk}^{BC} + \sum_{i=1}^{2} y_{ik}^{AC} \cdot \frac{\hat{m}_{ijk}^{(q)}}{\hat{m}_{i\cdot k}^{(q)}} + \sum_{i=1}^{2} y_{ij}^{AB} \cdot \frac{\hat{m}_{ijk}^{(q)}}{\hat{m}_{ij\cdot}^{(q)}} \quad j,k = 1,2 \\
x_{i\cdot k}^{(q)} &= y_{i\cdot k}^{ABC} + y_{ik}^{AC} + \sum_{j=1}^{2} y_{ij}^{AB} \cdot \frac{\hat{m}_{ijk}^{(q)}}{\hat{m}_{ij\cdot}^{(q)}} + \sum_{j=1}^{2} y_{jk}^{BC} \cdot \frac{\hat{m}_{ijk}^{(q)}}{\hat{m}_{\cdot jk}^{(q)}} \quad i,k = 1,2
\end{aligned}$$

2)**M-step**
Finding maximum likelihood estimates $\hat{m}_{ijk}^{(q+1)}$ using estimated sufficient statistics $x_{ij\cdot}^{(q)}$, $x_{\cdot jk}^{(q)}$ and $x_{i\cdot k}^{(q)}$ as if they were sufficient statistics of complete data:

$$\begin{aligned}
\hat{m}_{ij\cdot}^{(q+1)} &= x_{ij\cdot}^{(q)} \quad i,j = 1,2 \\
\hat{m}_{\cdot jk}^{(q+1)} &= x_{\cdot jk}^{(q)} \quad j,k = 1,2 \\
\hat{m}_{i\cdot k}^{(q+1)} &= x_{i\cdot k}^{(q)} \quad i,k = 1,2.
\end{aligned}$$

In this step e.g. IPFP can be employed.

7. MARGINAL PROBLEM

As the marginal problem is usually understood the problem of finding a joint probability distribution $Q((X_i)_{i\in V})$ having given distributions $P_l(X_i)_{i\in C_l}$, $l = 1,\ldots,L$ for its marginals. We are going to refer to them as to the oligodistribution, because of their few dimensional character. Sets C_l for $l = 1,\ldots,L$ are subsets of the index set V. $X_i, i \in V$ are random variables taking values from $\mathcal{X}_i$.

For any set of distributions $\{P_l((X_i)_{i\in C_l}), l = 1,\ldots,L\}$ no matter if they are consistent or not we can find corresponding pattern of incomplete data and therefore any marginal problem can be transformed into a problem of maximum likelihood estimation from incomplete data.

Continuing the example of three dichotomic variables let $X_1 = A, X_2 = B, X_3 = C$ and $\mathcal{X}_1 = a_1, a_2, \mathcal{X}_2 = b_1, b_2, \mathcal{X}_3 = c_1, c_2$. Suppose that the distributions $P_l((X_i)_{i \in C_l})$, $l = 1, \ldots, L$ are derived from incomplete data $\mathbf{y}$. For example in the case of three partially observed subsamples observed each on two variables $(A,B), (B,C)$ and (A,C) and one fully categorized subsample we can define:

$$
\begin{aligned}
P_1(A = a_i, B = b_j, C = c_k) &= \frac{y_{ijk}^{ABC}}{\sum\limits_{i,j,k=1}^{2} y_{ijk}^{ABC}} = \frac{1}{N_1} \cdot y_{ijk}^{ABC} \qquad i,j,k = 1,2 \\
P_2(A = a_i, B = b_j) &= \frac{y_{ij}^{AB}}{\sum\limits_{i,j=1}^{2} y_{ij}^{AB}} = \frac{1}{N_2} \cdot y_{ij}^{AB} \qquad i,j = 1,2 \\
P_3(B = b_j, C = c_k) &= \frac{y_{jk}^{BC}}{\sum\limits_{j,k=1}^{2} y_{jk}^{BC}} = \frac{1}{N_3} \cdot y_{jk}^{BC} \qquad j,k = 1,2 \\
P_4(A = a_i, C = c_k) &= \frac{y_{ik}^{AC}}{\sum\limits_{i,k=1}^{2} y_{ik}^{AC}} = \frac{1}{N_4} \cdot y_{ik}^{AC} \qquad i,k = 1,2
\end{aligned}
$$

8. APPLICATION OF THE EM ALGORITHM

In this section we are giong to descibe an estimation procedure whose purpose is to find a probability distribution if possible having given oligodistributions as its marginals otherwise at least being "close" to them. Suppose that a solution of our problem is the distribution

$$
\hat{Q}(A = a_i, B = b_j, C = c_k) = \frac{\hat{m_{ijk}}}{N} \qquad i,j,k = 1,2, \qquad N = \sum_{l=1}^{L} N_l,
$$

where $\hat{\mathbf{m}}$ is a maximum likelihood estimate of the parameter $\mathbf{m}$ given incomplete data $\mathbf{y}$ (see Section 6.). We can use the EM algorithm and rewrite it in the following way:

$$
\hat{Q}_{(q+1)} = \sum_{l=1}^{L} w_l \cdot \hat{Q}_{(q)} \cdot \frac{P_l}{\hat{Q}_{(q)}^{C_l}}, \tag{6}
$$

where $w_l = \frac{N_l}{N}$ are weights allocated to every distribution $P_l, l = 1, \ldots, L$ and $\hat{Q}_{(q)}^{C_l}$ denotes the marginal distribution of the distribution $\hat{Q}_{(q)}$ for the variables $(X_i)_{i \in C_l}$.

This procedure was described in [8] under the name of *method of iterative averages* together with several examples that show its behavior. A question arises, how can the maximum likelihood criteria for incomplete data be transformed into a problem of minimization of a function $\psi(Q|P_1, P_2, \ldots, P_L)$. It is shown in [9] that the sequence of iterations $Q_{(q+1)} = M(Q_{(q)})$ defined by the formula (6) converges and minimizes:

$$\psi(Q|P_1, P_2, \ldots, P_L) = \sum_{l=1}^{L} w_l \cdot I(Q \cdot \frac{P_l}{Q^{C_l}}||Q) = \sum_{l=1}^{L} w_l \cdot I(P_l||Q^{C_l}),$$

where $I(\cdot||\cdot)$ is in the Section 2. already mentioned the Kullback-Liebler divergence.

9. APPLICATION OF THE GEM ALGORITHM

Another problem whose solution we could derive from previously described methods is the problem of finding a joint probability distribution

$$Q((X_i)_{i\in V}) = \prod_{g\in G} \phi((X_i)_{i\in g})$$

which factorizes with respect to a class G of subsets of the set V having its marginals "close" to the given oligodistributions $P_l, l = 1, \ldots, L$.

In this case a version of the GEM algorithm can be employed. Suppose that we have the same oligodistributions as in the previous example and that we search for a probability distribution factorizing with respect to the class $G = \{D_1, \ldots, D_M\}$. This algorithm consists of two steps (compare it with the EM algorithm in the Section 6.):

- **E-step.** Estimation of the joint distribution marginals $R_m^{(q)}((X_i)_{i\in D_m})$, $m = 1, \ldots, M$ from the oligodistributions:

$$R_m^{(q)} \quad = \quad \sum_{l=1}^{L} w_l \cdot \left(\hat{Q}_{(q)} \cdot \frac{P_l}{\hat{Q}_{(q)}^{C_l}}\right)^{D_m} \qquad m = 1, 2, \ldots, M$$

  superscript D_m denotes marginalization to a probability distribution on $(X_i)_{i\in D_m}$

- **M-step.** One iteration (i.e. M consequent steps) of the IPFP in order to find a joint probability distribution $\hat{Q}^{(q+m)}$ factorizing with respect to G and being "closer" to the oligodistributions than $\hat{Q}^{(q)}$:

$$\hat{Q}_{(q+m)} \quad = \quad \hat{Q}_{(q+m-1)} \cdot \frac{R_m^{(q)}}{\hat{Q}_{(q+m-1)}^{D_m}} \quad m = 1, 2, \ldots, M.$$

10. CONCLUSION

This paper presents an attempt to apply the well known statistical procedures for maximum likelihood estimation from incomplete data to the marginal problem. Application of these procedures is straightforward in contrast to the problem of finding criteria which these procedures within the frame of marginal problem optimize. These criteria could be probably used to find simple convergence proofs.

REFERENCES

[1] Y. M. M. Bishop, S. E. Feinberg, and P. W. Holland. *Discrete Multivariate Analysis.* The MIT Press, Cambridge, MA, 1975.

[2] D. T. Brown. A note on approximations to discrete probability distribution. *Information and Control*, 2:386–392, 1959.

[3] S. F. Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. Roy. Statist. Soc. Ser. B*, 22:302–306, 1960.

[4] W. E. Deming and F. F. Stephan. On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.*, 11:427–444, 1940.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39:1–38, 1977.

[6] W. J. E. Dixon. *BMDP Statistical Software.* University of California Press, Berkeley, 1983 revised printing.

[7] I. J. Good. Maximum entropy for hypothesis formulation especially for multidimensional contigency tables. *Ann. Math. Statist.*, 34:911–934, 1963.

[8] R. Jiroušek and J. Vomlel. Inconsistent knowledge integration in a probabilistic model. In *Proceedings of Workshop "Mathematical Models for handling partial knowledge in A.I."*, Erice, Sicily, 1994. Plenum Publ. Corp.

[9] F. Matúš. On alternating minimization of I-divergence and averages of I-projections (unpublished manuscript). 1996.

Jiří Vomlel, Department of Decision-Making Theory,Institute of Information Theory and Automation, Pod vodárenskou věží 4, 182 08 Praha 8, Czech Republic, vomlel@utia.cas.cz