Stat Papers (2013) 54:727–739
DOI 10.1007/s00362-012-0458-x

REGULAR ARTICLE

Maximum likelihood estimation for ordered expectations of correlated binary variables

Wojciech Gamrot

Received: 11 January 2012 / Revised: 11 June 2012 / Published online: 3 July 2012

Abstract A multivariate binary distribution that incorporates the correlation between individual variables is considered. The availability of auxiliary information taking the form of simple ordering constraints on their expected values is assumed. The problem of constructing constraint-preserving estimates for expectations is formulated as conditional maximization of convex likelihood function for corresponding multinomial distribution with suitably chosen restrictions. Starting values for convex optimization algorithms are proposed. The proposed estimator is consistent under mild assumptions.

Keywords Binomial parameter · Inequality constraints · Maximum likelihood

Mathematics Subject Classification (2000) 62F30

1 Introduction

The problem of estimating ordered probabilities has already been studied for more than fifty years. Ayer et al. (1955) considered estimators for a sequence of binomial parameters known to satisfy a set of inequalities defining a simple order. They have proven that a recursive procedure for isotonic regression, later known as pool–adjacent–violators–algorithm (PAVA) yields the maximum likelihood estimator of these probabilities satisfying the constraints. The problem was also independently studied by Brunk (1955) and van Eeden (1956, 1957, 1958) who provided important generalization including the possibility of non-simple ordering. Since then, these results were further developed in several papers involving studies of the existence and properties of

W. Gamrot (✉)
Department of Statistics, University of Economics, 1 Maja 50, 40-287 Katowice, Poland
e-mail: wojciech.gamrot@ue.katowice.pl

maximum likelihood estimates as in Katz (1963), Sackrowitz and Strawderman (1974), Parsian and Sanjari Farsipour (1997) and Charras and van Eeden (1991). Alternative estimators were proposed by Sackrowitz (1982) and Perron (2003). Extensions to several auxiliary variables were proposed by Burdakov et al. (2004). Recently, Bayesian and minimax approaches to this problem have been emphasized by a paper of Marchand and MacGibbon (2000). Computation algorithms were provided including those by Lee (1983), Best and Chakravarti (1990), Qian (1992), Block et al. (1994), Ahuja and Orlin (2001), Hansohm (2007) as well as Hansohm and Hu (2012). Implementations of PAVA and other computational methods for isotonic regression are also widely available in statistical packages as reported by de Leeuw et al. (2009). The complete discussion of the literature on estimation of ordered probabilities exceeds the scope of this paper. A good summary of the state of knowledge is presented in monographs by Barlow et al. (1972), Robertson et al. (1988) and especially by van Eeden (2006). It appears interesting that all mentioned results share a common feature: individual binary variables are conveniently assumed to be independent which greatly facilitates the computation of the likelihood function. The results for correlated binary variables with ordered expectations are not known to this author. On the other hand, in practice sometimes the independence of individual variables cannot be guaranteed. Such situations may arise in the insurance industry during the calculation of insurance premiums (see Sundt 1999, Wolny-Dominiak and Trzęsiok 2008). These may depend on probability of a claim-generating event to occur. Such probabilities for various claim generating events may be known to satisfy a simple order, while occurrences of individual claims are correlated. For example, the probability of individual persons within the household contracting a particular disease to be insured against may be known to depend monotonically on their age or some other relevant factor, while occurrences of individual claims are correlated due to possibility of contagion, genetic similarity or common environmental hazards. Hence, in this paper the problem of estimating ordered probabilities is generalized by allowing for a dependence between binary variables. Their joint distribution is found to be a special case of a multinomial one. Consequently, the original problem is re-formulated as estimation of multinomial parameters satisfying suitably chosen restrictions.

2 Binary random vectors

Consider the vector of binary random variables $\mathbf{x} = [x_1, \ldots, x_r]' \in \{0, 1\}^r$ with $P(x_i = 1) = p_i$ for $i = 1, \ldots, r$ being unknown probabilities. Hence the expectation of $\mathbf{x}$ is:

$$E(\mathbf{x}) = [E(x_1), \ldots, E(x_r)]' = [p_1, \ldots, p_r]' = \mathbf{p}$$

Let us also assume that $p_1, \ldots, p_r$ are known to satisfy a simple ordering:

$$p_1 \leq p_2 \leq \cdots \leq p_r \tag{1}$$

Moreover let us assume that each of these probabilities is also known to satisfy individual constraints in the form:

$$d_i \leq p_i \leq u_i \tag{2}$$

for $i = 1, \dots, r$. These additional constraints are not associated with correlation issue nor with ordering constraints but reflect some additional external knowledge which may be available. If such a knowledge is absent then simply $d_1 = d_2 = \dots = d_r = 0$ and $u_1 = u_2 = \dots = u_r = 1$. Now, the problem to be considered is to estimate $p_1, \dots, p_r$ from the sample in such a way that the estimates satisfy (1) and (2). The usual approach is to use maximum likelihood principle based on the assumption that $x_1, \dots, x_r$ are independent and the sample is i.i.d. This leads to a likelihood function

$$L_{ind}(\mathbf{p}) = \prod_{i=1}^{r} \binom{n}{k_i} p_i^{k_i} (1 - p_i)^{n-k_i} \tag{3}$$

where $k_1, \dots k_r$ represent counts of ones respectively for $x_1, \dots, x_r$ in the sample of size n. Such a likelihood function is maximized by

$$\widehat{\mathbf{p}} = [\widehat{p}_1, \dots, \widehat{p}_r]' \tag{4}$$

where $\widehat{p}_i = k_i/n$ for $i = 1, \dots, r$. However this statistic does not necessarily satisfy conditions (1) and (2). To find a restricted maximum likelihood estimator that minimizes (3) with respect to (1) an algorithm was proposed by Ayer et al. (1955). Known as pool–adjacent–violators–algorithm (PAVA) it relies on repeatedly averaging sample proportions for which (1) is violated until it is not. Several modifications of this procedure were later considered for more complicated situations including partial support for (2) introduced in the paper of McKeown and Jewell (2010). However, to our knowledge the assumption of independence between binary variables always played a crucial role and was not abandoned. We will now drop this assumption.

3 Restricted estimation for correlated variables

A binary vector $\mathbf{x}$ of size r may take $k = 2^r$ different values. Let vectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ each of length r represent these values, so that

$$\begin{aligned} \mathbf{a}_1 &= [1, 1, \dots, 1]' \\ \mathbf{a}_2 &= [1, \dots, 1, 0]' \\ &\vdots \\ \mathbf{a}_k &= [0, \dots, 0, 0]' \end{aligned}$$

The last vector $\mathbf{a}_k$ contains only zeros, which will turn out to be useful later. These vectors may be arranged in a matrix:

$$\mathbf{A} = [a_{ij}] = [\mathbf{a}_1, \ldots, \mathbf{a}_k]'$$

of size $k \times r$. Any realization $\mathbf{a}_z$ of the vector $\mathbf{x}$ may be represented by the index $z \in \{1, \ldots, k\}$ having a multinomial distribution characterised by a vector of respective multinomial probabilities $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]' \in \langle 0, 1\rangle^k$ satisfying

$$\mathbf{1}_k' \cdot \boldsymbol{\mu} = 1$$

while $\mathbf{1}_t = [1, \ldots, 1]'$ will in general represent a vector of ones of any length $t \in \{1, 2, \ldots\}$. Parameters $p_1, \ldots, p_r$ depend on $\mu_1, \ldots, \mu_k$ through the formula:

$$\mathbf{p} = \mathbf{A}'\boldsymbol{\mu} \tag{5}$$

Hence, one may attempt to estimate components of $\mathbf{p}$ by first estimating all individual components of $\boldsymbol{\mu}$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ represent individual realizations of $\mathbf{x}$ in a sample of size n. Then let

$$\mathbf{f} = [f_1, \ldots, f_k]'$$

be the vector of sample counts associated with respective realizations of z and corresponding to elements of $\boldsymbol{\mu}$ so that $f_i = \#\{j : \mathbf{x}_j = \mathbf{a}_i\}$ for $i = 1, \ldots, k$. As noted e.g. by Lehmann and Cassella (1998) the likelihood function for a sample drawn from a multinomial distribution takes form:

$$L(\boldsymbol{\mu}) = \frac{n!}{f_1! \cdot \ldots \cdot f_k!} \mu_1^{f_1} \cdot \ldots \cdot \mu_k^{f_k} \tag{6}$$

However, instead of directly maximizing $L(\boldsymbol{\mu})$ it is convenient to minimize the negative log-likelihood:

$$L_0(\boldsymbol{\mu}) = -\log\left(\frac{n!}{f_1! \cdot \ldots \cdot f_k!}\right) - \sum_{i=1}^{k} f_i \log(\mu_i) \tag{7}$$

When restrictions for $p_1, \ldots, p_r$ are disregarded the above function is minimized (unconditionally) at

$$\widehat{\boldsymbol{\mu}} = [\widehat{\mu}_1, \ldots, \widehat{\mu}_k]' = n^{-1}\mathbf{f}$$

which would lead to (4) via (5). However, it may still happen that estimates of $\mathbf{p}$ computed this way do not satisfy (1) or (2). Moreover, known results on restricted estimation of multinomial parameters presented in the paper of Jewell and Kalbfleisch (2004) are not applicable here, as each of individual restrictions on p_i involves more than two values of μ_i. To find a (conditional) minimum of the negative log-likelihood function (7) satisfying restrictions (1) and (2) one has to re-formulate these restrictions in terms of $\mu_1, \ldots, \mu_k$ and numerically locate the (conditional) minimum of $L_0(\boldsymbol{\mu})$.

Let $\tilde{\mathbf{a}}_j = [a_{1j}, \ldots, a_{kj}]'$ represent the j-th column of the matrix $\mathbf{A}$ for $j = 1, \ldots, r$. Let us denote:

$$\begin{aligned}\mathbf{A}_+ &= [\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_{r-1}]\\ \mathbf{A}_- &= [\tilde{\mathbf{a}}_2, \ldots, \tilde{\mathbf{a}}_r]\end{aligned}$$

and

$$\mathbf{A}_\circ = \mathbf{A}_- - \mathbf{A}_+$$

of size $k \times (r-1)$ each. The set of $r-1$ conditions (1) will be satisfied when

$$\mathbf{A}_\circ' \boldsymbol{\mu} \geq \mathbf{0}_{r-1}$$

while $\mathbf{0}_t = [0, \ldots, 0]'$ will in general represent a vector of zeros of length $t \in \{1, 2, \ldots\}$ and any inequality is assumed to hold for two matrices when it holds for all their respective elements. By denoting $\mathbf{d} = [d_1, \ldots, d_r]'$ and $\mathbf{u} = [u_1, \ldots, u_r]'$ conditions (2) may be expressed as:

$$\mathbf{d} \leq \mathbf{p} \leq \boldsymbol{u}$$

or equivalently:

$$\mathbf{d} \leq \mathbf{A}'\boldsymbol{\mu} \leq \mathbf{u}$$

Hence, the restricted maximum likelihood estimate $\widehat{\boldsymbol{\mu}}_\#$ of $\boldsymbol{\mu}$ may be calculated as a global solution to the optimization problem:

$$\begin{cases} L_0(\boldsymbol{\mu}) \to \min \\ \mathbf{A}_\circ'\boldsymbol{\mu} \geq \mathbf{0}_{r-1} \\ \mathbf{d} \leq \mathbf{A}'\boldsymbol{\mu} \leq \mathbf{u} \\ \mathbf{0}_k \leq \boldsymbol{\mu} \leq \mathbf{1}_k \\ \mathbf{1}_k'\boldsymbol{\mu} = 1 \end{cases} \tag{8}$$

To eliminate the equality constraint the last component μ_k of $\boldsymbol{\mu}$ will be expressed as a function of $\boldsymbol{\mu}_* = [\mu_1, \ldots, \mu_{k-1}]'$, according to the formula:

$$\mu_k = 1 - \sum_{i=1}^{k-1} \mu_i = 1 - \mathbf{1}_{k-1}'\boldsymbol{\mu}_*$$

This lets us represent $L_0(\boldsymbol{\mu})$ as the function of $\boldsymbol{\mu}_*$ taking the form:

$$L_0(\boldsymbol{\mu}_*) = -\log\left(\frac{n!}{f_1! \cdot \ldots \cdot f_k!}\right) - \sum_{i=1}^{k-1} f_i \log(\mu_i) - f_k \log\left(1 - \sum_{i=1}^{k-1} \mu_i\right)$$

In order to re-formulate (8) let us denote the $(k-1)\times r$ matrix obtained by dropping the last row from $\mathbf{A}$ by $\mathbf{A}_*$, and the $(k-1)\times(r-1)$ matrix obtained by dropping the last row from matrix $\mathbf{A}_\circ$ by $\mathbf{A}_{\circ *}$. Dropped rows contain only zeros. Consequently $\mathbf{A}$ and $\mathbf{A}_\circ$ may respectively be expressed in the form:

$$\mathbf{A}=\begin{bmatrix}\mathbf{A}_*\\ \mathbf{0}'_r\end{bmatrix}$$

$$\mathbf{A}_\circ=\begin{bmatrix}\mathbf{A}_{\circ *}\\ \mathbf{0}'_{r-1}\end{bmatrix}$$

As a result, the optimization problem may be replaced with an equivalent one involving only inequality constraints:

$$\begin{cases} L_0(\boldsymbol{\mu}_*)\rightarrow \min\\ \mathbf{A}'_{\circ *}\boldsymbol{\mu}_*\geq \mathbf{0}_{r-1}\\ \mathbf{d}\leq \mathbf{A}'_*\boldsymbol{\mu}_*\leq \mathbf{u}\\ \boldsymbol{\mu}_*\geq \mathbf{0}_{k-1}\\ \mathbf{1}'_{k-1}\boldsymbol{\mu}_*\leq 1\end{cases} \tag{9}$$

4 Computational issues

The matrix inequality $\mathbf{d}\leq A'_*\boldsymbol{\mu}_*\leq \boldsymbol{u}$ represents in a concise way $2r$ individual inequalities. However, if for some $i\in\{1,\ldots,r\}$ a trivial value of $d_i=0$ (or $u_i=1$) occurs then corresponding individual inequality $d_i\leq p_i$ (or $p_i\leq u_i$) is always satisfied due to requirements: $\boldsymbol{\mu}_*\geq \mathbf{0}_{k-1}$ and $\mathbf{1}'_{k-1}\boldsymbol{\mu}_*\leq 1$ in (9). Also, if for any $i\in\{1,\ldots,r\}$ there exists some $j\in\{1,\ldots,i-1\}$ such that $d_j\geq d_i$ then the inequality $d_i\leq p_i$ is always satisfied thanks to ordering constraints. Moreover, if for any $i\in\{1,\ldots,r\}$ there exists some $j\in\{i+1,\ldots,r\}$ such that $u_j\leq u_i$ then the inequality $p_i\leq u_i$ is always satisfied thanks to ordering constraints. From a computational point of view it is desirable to eliminate any redundant inequalities. Let $\mathbf{g}=[g_1,\ldots,g_r]'$ where

$$g_i=\begin{cases}1 & \text{when } d_i=0 \text{ or } d_i\leq \max_{j<i}(d_j)\\ 0 & \text{otherwise}\end{cases}$$

for $i=1,\ldots,h$ and $\mathbf{h}=[h_1,\ldots,h_r]'$ where

$$h_i=\begin{cases}1 & \text{when } u_i=1 \text{ or } u_i\geq \min_{j>i}(u_j)\\ 0 & \text{otherwise}\end{cases}$$

for $i=1,\ldots,h$. Let $m_d=\mathbf{1}'_r\mathbf{g}$ and $m_u=\mathbf{1}'_r\mathbf{h}$ respectively denote the number of nonzero elements in $\mathbf{g}$ and $\mathbf{h}$. Let $\mathbf{d}_w$ be a vector obtained by dropping from $\mathbf{d}$ all m_d elements corresponding to nonzero components in $\mathbf{g}$. Let $\mathbf{A}_*\boldsymbol{d}$ be the $(k-1)\times(r-m_d)$ matrix obtained from $\mathbf{A}_*$ by dropping all m_d columns corresponding to nonzero components in $\mathbf{g}$. Let $\mathbf{u}_w$ be a vector obtained by dropping from $\mathbf{u}$ all m_u elements corresponding to nonzero components in $\mathbf{h}$. Let $\mathbf{A}_{*}\boldsymbol{u}$ be the $(k-1)\times(r-m_u)$ matrix obtained from $\mathbf{A}_{*}$ by dropping all m_u columns corresponding to nonzero components in $\mathbf{h}$. The problem (9) may then be transformed to another equivalent form:

$$\begin{cases} L_0(\boldsymbol{\mu}_*) \to \min \\ \mathbf{B}'\boldsymbol{\mu}_* \geq \mathbf{b} \end{cases} \tag{10}$$

where

$$\mathbf{B} = [b_{ij}] = \begin{bmatrix} \mathbf{A}_{\circ *} & \mathbf{A}_{*}\mathbf{d} & -\mathbf{A}_{*}\mathbf{u} & \mathbf{I} & -\mathbf{1}_{k-1} \end{bmatrix}$$

is of size $(k-1)\times h$ and $h = 3r + k - m_d - m_u - 1$ while $\mathbf{I}$ represents an identity matrix of size $(k-1)\times(k-1)$ and

$$\mathbf{b} = [b_1, \ldots, b_h]' = \begin{bmatrix} \mathbf{0}'_{r-1} & \mathbf{d}'_w & -\mathbf{u}'_w & \mathbf{0}'_{k-1} & -1 \end{bmatrix}'$$

is of length h. Solving of the problem (10) is facilitated by the following lemma:

Lemma 1 *The function $L_0(\boldsymbol{\mu}_*)$ is convex*

Proof The first term does not depend on $\boldsymbol{\mu}_*$ so it is constant and hence it is convex. As the logarithm is a concave function, and the counts $f_1, \ldots, f_{k-1}$ are non-negative, then the second term is a linear combination of concave functions with non-positive weights and hence it is convex. Let us consider the third term. Its Hessian is given by:

$$\mathbf{H} = \frac{f_k}{\left(\sum_{i=1}^{k-1}\mu_i - 1\right)^2}\mathbf{J}$$

where $\mathbf{J} = \mathbf{1}_{k-1}\mathbf{1}'_{k-1}$ is a matrix of size $(k-1)\times(k-1)$ with all elements equal to unity. Then for any $\mathbf{x} = [x_1, \ldots, x_{k-1}]' \neq \mathbf{0}_{k-1}$ one may derive:

$$\mathbf{x}'\mathbf{H}\mathbf{x} = \frac{f_k}{\left(\sum_{i=1}^{k-1}\mu_i - 1\right)^2}\mathbf{x}'\mathbf{J}\mathbf{x} = \frac{f_k}{\left(\sum_{i=1}^{k-1}\mu_i - 1\right)^2}\left(\sum_{i=1}^{k-1} x_i\right)^2 \geq 0$$

which means that the third term in $L_0(\boldsymbol{\mu}_*)$ is convex too. Consequently, the function $L_0(\boldsymbol{\mu}_*)$ constructed as a sum of convex terms is also convex. □

The problem (10) involves minimization of a convex criterion function and a set of linear constraints that is also convex. Hence, it may be solved by using standard convex programming methods such as interior point method discussed in detail by Boyd and Vandenberghe (2004) or the adaptive barrier majorization-minimization (MM) algorithm considered by Lange (2001).

In order to use iterative methods mentioned above it is neccessary to identify some feasible starting point for the iterative process satisfying constraints in (10). This may be achieved in two steps. First, one may assume without a loss of generality that

$u_i \leq u_j$ and $d_i \leq d_j$ for any $i < j$ as the opposite situation is precluded by (1). The vector of marginal pobabilities representing an initial estimate of $\mathbf{p}$ satisfying the constraints may then be constructed according to the formula:

$$\widehat{\mathbf{p}}_0 = \mathbf{d} + \mathbf{q}(\mathbf{u} - \mathbf{d})$$

where $\mathbf{q} = [q_1, \ldots, q_r]'$ and

$$q_i = \frac{i}{r+1}$$

for i=1,. . .,r. Let $\circ$ represent the Hadamard product. Any vector of multinomial probabilities that leads to $\widehat{\mathbf{p}}_0$ via (5) also satisfies constraints of (10). In particular this is true for the vector

$$\widehat{\boldsymbol{\mu}}_0 = \exp(\log(\mathbf{1}_k\widehat{\mathbf{p}}_0' \circ \mathbf{A} + \mathbf{1}_k(\mathbf{1}_r - \widehat{\mathbf{p}}_0)' \circ (\mathbf{1}_k\mathbf{1}_r' - \mathbf{A}))\mathbf{1}_k)$$

which is obtained by assuming that $x_1, \ldots, x_r$ are independent. Consequently, first $k-1$ elements of $\widehat{\boldsymbol{\mu}}_0$ may serve as an initial estimate of $\boldsymbol{\mu}_*$. Once the global minimum for the problem (10) is located at some point, say $\widehat{\boldsymbol{\mu}}_{*\#}$, the maximum likelihood estimate $\widehat{\boldsymbol{\mu}}_\# = [\widehat{\mu}_{1\#}, \ldots, \widehat{\mu}_{k\#}]'$ of $\boldsymbol{\mu}$ is obtained as:

$$\widehat{\boldsymbol{\mu}}_\# = \left[\widehat{\boldsymbol{\mu}}_{*\#}', 1 - \mathbf{1}_{k-1}'\widehat{\boldsymbol{\mu}}_{*\#}\right]' \tag{11}$$

The concept of self-concordance introduced by Nesterov and Nemirovski (1994) will be useful to analyse the complexity of a numeric optimization problem. Let us recall that a convex function $f : \boldsymbol{R} \to \boldsymbol{R}$ is a self-concordant function if $f'''(x) \leq 2f''(x)^{3/2}$ for any x in the domain of f. A function $f : \boldsymbol{R}^n \to \boldsymbol{R}$ is self concordant functon of x if $f(x + tv)$ is self-concordant function of t for all v and all x in the domain of f. This lets us state the following result:

Lemma 2 *The function $L_0(\boldsymbol{\mu}_*)$ is self-concordant*

Proof Negative logarithm is self-concordant and the composition of a self-concordant function with affine function is also self-concordant as indicated by Boyd and Vandenberghe (2004). Hence the expression $-\log\left(1 - \sum_{i=1}^{k-1} \mu_i\right)$ as well as $-\log(\mu_i)$ for $i \in \{1, \ldots, k-1\}$ are all self-concordant. As $f_1, \ldots, f_k$ are nonnegative integers, the function $L_0(\boldsymbol{\mu}_*)$ is a linear combination of self-concordant terms with weights greater than or equal to one (as terms with no counts vanish), and hence it is also self-concordant. □

The self-concordance property may be used to establish a rigorous upper bound on the total number of Newton steps required to solve a problem using barrier method. This may be done using formulas given by Boyd and Vandenberghe (2004) but will not be elaborated here.

5 Estimator properties

If all constraints are satisfied strictly at $\boldsymbol{\mu}$ (which in other words means that $\boldsymbol{\mu}$ belongs to the interior of the feasible region) then $\widehat{\boldsymbol{\mu}}_{\#}$ may be shown to be a consistent estimator for $\boldsymbol{\mu}$. This is stated as follows:

Theorem 1 *Let $\hat{\boldsymbol{\theta}}$ be a (weakly) consistent estimator of a parameter vector $\boldsymbol{\theta}$ computed by maximizing the likelihood function $L(\cdot)$. Let $\hat{\boldsymbol{\theta}}_{\#}$ be an estimator for $\boldsymbol{\theta}$ computed by maximizing the same likelihood function subject to the constraint $\boldsymbol{\theta} \in C$ for some non-hollow set C. If $\boldsymbol{\theta} \in interior(C)$ then $\hat{\boldsymbol{\theta}}_{\#}$ is (weakly) consistent for $\boldsymbol{\theta}$.*

Proof Let $\hat{\boldsymbol{\theta}}^{(n)}$ and $\hat{\boldsymbol{\theta}}_{\#}^{(n)}$ be realizations of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\#}$ for the sample size n. Let the event ω correspond to the equality $\hat{\boldsymbol{\theta}}_{\#}^{(n)} = \hat{\boldsymbol{\theta}}^{(n)}$ being satisfied. As $\boldsymbol{\theta}$ belongs to the interior of C, there exists some ϵ_0 such that $\{\boldsymbol{\delta} : |\boldsymbol{\delta} - \boldsymbol{\theta}| \le \epsilon_0\} \subseteq C$. Since $|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| \le \epsilon_0$ implies ω we have $P(\omega) \ge P(|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| \le \epsilon_0)$ and $P(\neg\omega) \le P(|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| > \epsilon_0)$ so that:

$$\begin{aligned}
P(|\hat{\boldsymbol{\theta}}_{\#}^{(n)} - \boldsymbol{\theta}| > \epsilon) &= P(|\hat{\boldsymbol{\theta}}_{\#}^{(n)} - \boldsymbol{\theta}| > \epsilon|\omega)P(\omega) + P(|\hat{\boldsymbol{\theta}}_{\#}^{(n)} - \boldsymbol{\theta}| > \epsilon|\neg\omega)P(\neg\omega) \\
&\le P(|\hat{\boldsymbol{\theta}}_{\#}^{(n)} - \boldsymbol{\theta}| > \epsilon|\omega)P(\omega) + P(\neg\omega) = P(|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| > \epsilon|\omega)P(\omega) + P(\neg\omega) \\
&\le P(|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| > \epsilon|\omega)P(\omega) + P(|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| > \epsilon|\neg\omega)P(\neg\omega) + P(\neg\omega) \\
&= P(|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| > \epsilon) + P(\neg\omega) \le P(|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| > \epsilon) + P(|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| > \epsilon_0)
\end{aligned}$$

Hence, from consistency of $\hat{\boldsymbol{\theta}}$ we have for any $\epsilon > 0$:

$$\begin{aligned}
\lim_{n\to\infty} P(|\hat{\boldsymbol{\theta}}_{\#}^{(n)} - \boldsymbol{\theta}| > \epsilon) &\le \lim_{n\to\infty} P(|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| > \epsilon) + \lim_{n\to\infty} P(|\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}| > \epsilon_0) \\
&= 0 + 0 = 0
\end{aligned}$$

which means that $\hat{\boldsymbol{\theta}}_{\#}$ is consistent for $\boldsymbol{\theta}$. □

Corollary 1 *If $p_1 < p_2 < \ldots < p_r$ and $d_i < p_i < u_i$ for $i = 1, \ldots, r$ then $\hat{\boldsymbol{\mu}}_{\#}$ is (weakly) consistent for $\boldsymbol{\mu}$.*

It is important to note that the above proof is based on a tacit assumption that the conditional extremum of the criterion function is located exactly. In practice it may be located numerically with arbitrarily high accuracy so this condition may safely be considered satisfied. The corresponding estimate of the vector $\mathbf{p}$ is then calculated according to the formula:

$$\widehat{\mathbf{p}}_{\#} = \mathbf{A}'\widehat{\boldsymbol{\mu}}_{\#} \tag{12}$$

It is also possible to construct an estimator for the covariance matrix of $\mathbf{x}$:

$$V(\mathbf{x}) = E((\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))')$$

in the form of a statistic:

$$\widehat{V}(\mathbf{x}) = \mathbf{A}'(\mathrm{diag}(\widehat{\boldsymbol{\mu}}_{\#}) - \widehat{\boldsymbol{\mu}}_{\#}\widehat{\boldsymbol{\mu}}'_{\#})\mathbf{A}$$

If all constraint inequalities are strict so that $p_1 < p_2 < \ldots < p_r$ and $d_i < p_i < u_i$ for $i = 1, \ldots, r$, then by the invariance theorem of Goldberger (1964) the statistic $\widehat{\mathbf{p}}_{\#}$ is a consistent estimator for $\mathbf{p}$ and $\widehat{V}(\mathbf{x})$ is a consistent estimator for $V(\mathbf{x})$. By invariance theorem discussed by Bartoszyński and Niewiadomska-Bugaj (2008) both statistics are respectively maximum likelihood estimators for $\mathbf{p}$ and $V(\mathbf{x})$. Both statements are not guaranteed to be true when some of constraint inequalities are not satisfied strictly.

6 Numerical example

Let $r = 3$ so that three probabilities $p_1 \leq p_2 \leq p_3$ are to be estimated. For simplicity let $\mathbf{d} = \mathbf{0}_3$ and $\mathbf{u} = \mathbf{1}_3$ so individual bounds are trivial. This results in $\mathbf{B}$ and $\mathbf{b}$ taking the form:

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

and

$$\mathbf{b} = [\ 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad -1\ ]$$

First two columns in $\mathbf{B}$ above correspond to ordering constraints. The last column corresponds to the condition $\mathbf{1}'_{k-1}\boldsymbol{\mu}_* \leq 1$ and remaining columns correspond to the requirement $\boldsymbol{\mu}_* \geq \mathbf{0}_{k-1}$.

Assume that $n = 35$ realizations of random vector $\mathbf{x}$ were observed. Realized counts $\mathbf{f} = [f_1, \ldots, f_k]'$ associated with all $k = 2^r = 8$ possible values of $\mathbf{x}$ and corresponding sample proportions $\widehat{\boldsymbol{\mu}} = [\widehat{\mu}_1, \ldots, \widehat{\mu}_k]'$ are shown in Table 1.

This results in a vector of sample proportions:

$$\widehat{\mathbf{p}} = \mathbf{A}'\widehat{\boldsymbol{\mu}} = [0.4286 \ \ 0.5429 \ \ 0.2857]'$$

that clearly violates ordering constraints. At the same time, the solution to the problem (10) is located at the point:

$$\widehat{\boldsymbol{\mu}}_{\#} = [0.0286 \ \ 0.0571 \ \ 0.1408 \ \ 0.1878 \ \ 0.2054 \ \ 0.1232 \ \ 0.0286]'$$

Table 1 Observed realizations of the vector **x** in the numerical example

i	$\mathbf{a}_i$	f_i	$\widehat{\mu}_i$
1	1 1 1	1	0.0286
2	0 1 1	2	0.0571
3	1 0 1	3	0.0857
4	0 0 1	4	0.1143
5	1 1 0	10	0.2857
6	0 1 0	6	0.1714
7	1 0 0	1	0.0286
8	0 0 0	8	0.2286

The proposed estimator is then calculated via (11) and (12) and it takes the value:

$$\widehat{\mathbf{p}}_{\#} = [0.4033, 0.4143, 0.4143]'$$

The PAVA-based estimator of Ayer et al. (1955) is computed in this case by averaging all elements in $\widehat{\mathbf{p}}$ (firstly, the second and third element, then all three of them) and it takes the value:

$$\widehat{\mathbf{p}}_{PAVA} = [0.4190, 0.4190, 0.4190]'$$

Hence, it has been demonstrated that the proposed estimator and the PAVA-based one may take different values while both satisfy ordering constraints.

7 Conclusions

The proposed procedure clearly produces different estimates than those obtained by PAVA. If the constraints are known to be satisfied strictly, realizations of proposed variance estimator may be compared with those known for PAVA to compare accuracy of both estimation procedures. Also, still more analytical research is needed to assess the properties of proposed estimator when some or all constraints are not satisfied strictly.

From the technical point of view an obvious limitation of the proposed approach lies in rather poor scalability, as the number of optimization variables grows exponentially with r. The exact threshold of applicability depends on available hardware and software and hence it is rather hard to be correctly pinpointed. Nevertheless, the proposed estimator should be useful for binary vectors with small number of components, especially in situations, where large sample size may be expected. It seems that after minor modifications the proposed approach might also be used to deal with non-simple orderings of estimated parameters.

Acknowledgements The work was partially supported by the Grant No N N111 558540 from the Ministry of Science and Higher Education. The author is indebted to two anonymous referees for helpful comments that led to generalization of the consistency theorem and improvement of the manuscript.

References

Ahuja RK, Orlin JB (2001) A fast scaling algorithm for minimizing separable convex functions subject to chain constraints. Oper Res 49:784–789

Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E (1955) An empirical distribution function for sampling with incomplete information. Ann Math Stat 6(4):641–647

Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD (1972) Statistical inference under order restrictions. Wiley, New York

Bartoszyński R, Niewiadomska-Bugaj M (2008) Probability and statistical inference. Wiley, New York

Best MJ, Chakravarti N (1990) Active set algorithms for isotonic regression; a unifying framework. Math Program 47:425–439

Block H, Qian S, Sampson A (1994) J Comput Graph Stat 3(3):285–300

Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, New York

Brunk HB (1955) Maximum likelihood estimates of monotone parameters. Ann Math Stat 26:607–616

Burdakov O, Grimwall A, Hussian M (2004) A generalized PAV algorithm for monotonic regression in several variables. In: COMPSTAT proceedings in computational statistics. Physica-Verlag/Springer, Heidelberg, pp 761–767

Charras A, Eeden Cvan (1991) Bayes and admissibility properties of estimators in truncated parameter spaces. Can J Stat 19:121–134

de Leeuw J, Hornik K, Mair P (2009) Isotone optimization in R: pool–adjacent–violators algorithm (PAVA) and active set methods. J Stat Softw 32(5):1–24

Goldberger AS (1964) Econometric theory. Wiley, New York

Hansohm J (2007) Algorithms and error estimations for monotone regression on partially preordered sets. J Multivar Anal 98:1043–1050

Hansohm J, Hu X (2012) A convergent algorithm for a generalized multivariate isotonic regression problem. Stat Papers 53(1):107–115

Jewell NP, Kalbfleisch JD (2004) Maximum likelihood estimation of ordered multinomial probabilities. Biometrics 5(2):291–306

Katz MW (1963) Estimating ordered probabilities. Ann Math Stat 34:967–972

Lange K (2001) Numerical analysis for statisticians. Springer, New York

Lee CC (1983) The min–max algorithm and isotonic regression. Ann Stat 11:467–477

Lehmann EL, Cassella G (1998) Theory of point estimation. Springer, New York

Marchand E, MacGibbon B (2000) Minimax estimation of a constrained binomial proportion. Stat Decis 18:129–167

McKeown K, Jewell NP (2010) Miclassification of current status data. Lifetime Data Anal 16:215–230

Nesterov Y, Nemirovski A (1994) Interior point polynomial algorithms in convex programming, studies in applied mathematics 13. SIAM, Philadelphia

Parsian A, Farsipour NS (1997) Estimation of parameters of exponential distribution in the truncated space using asymmetric loss function. Stat Papers 38:423–443

Perron F (2003) Improving on the MLE of *p* for a binomial(*n*,*p*) when *p* is around 1/2. In: Moore M, Froda S, Leger C (eds) Mathematical statistics and applications: Fesschrift for Constance van Eeden. IMS lecture notes and monograph series, vol 43. IMS, Hayward, pp 45–61

Qian S (1992) Minimum lower sets algorithm for isotonic regression. Stat Probab Lett 15:31–35

Robertson T, Wright FT, Dykstra RL (1988) Order restricted statistical inference. Wiley, New York

Sackrowitz H (1982) Procedures for improving the MLE for ordered binomial parameters. J Stat Plan Inference 6:287–296

Sackrowitz H, Strawderman W (1974) On the admissibility of the M.L.E. for ordered binomial parameters. Ann Stat 2:822–828

Sundt B (1999) An introduction to non-life insurance mathematics. VVW, Karlsruhe

van Eeden C (1956) Maximum likelihood estimation of ordered probabilities. Proc Kon Nederl Akad Wetensch Ser A 60:128–136

van Eeden C (1957) Maximum likelihood estimation of partially or completely ordered probabilities. Proc Kon Nederl Akad Wetensch Ser A 59:444–455

van Eeden C (1958) Testing and estimating ordered parameters of probability distributions. Ph.D. thesis, University of Amsterdam, Amsterdam

van Eeden C (2006) Restricted parameter space estimation problems: admissibility and minimaxity results. Springer, New York

Wolny-Dominiak A, Trzęsiok M (2008) Monte Carlo simulation applied to a'priori rate making, In: Proceedings of 26th international conference mathematical methods in economics 2008. Technical University of Liberec, Liberec, 17–19 Sept 2008