ESAIM: PS
January 2006, Vol. 10, p. 11–23
DOI: 10.1051/ps:2005019

ESAIM: Probability and Statistics

USING AUXILIARY INFORMATION IN STATISTICAL FUNCTION ESTIMATION

Sergey Tarima[1] and Dmitri Pavlov[2]

Abstract. In many practical situations sample sizes are not sufficiently large and estimators based on such samples may not be satisfactory in terms of their variances. At the same time it is not unusual that some auxiliary information about the parameters of interest is available. This paper considers a method of using auxiliary information for improving properties of the estimators based on a current sample only. In particular, it is assumed that the information is available as a number of estimates based on samples obtained from some other mutually independent data sources. This method uses the fact that there is a correlation effect between estimators based on the current sample and auxiliary information from other sources. If variance covariance matrices of vectors of estimators used in the estimating procedure are known, this method produces more efficient estimates in terms of their variances compared to the estimates based on the current sample only. If these variance-covariance matrices are not known, their consistent estimates can be used as well such that the large sample properties of the method remain unchangeable. This approach allows to improve statistical properties of many standard estimators such as an empirical cumulative distribution function, empirical characteristic function, and Nelson-Aalen cumulative hazard estimator.

Mathematics Subject Classification. 62G05, 62G20.

Received February 28, 2004. Accepted July 22, 2005.

Introduction

In many practical situations sample sizes are not sufficiently large and estimators based on such samples may not be satisfactory in terms of their variances. At the same time it is not unusual that some auxiliary information about the parameters of interest is available. Such additional information, if available, can be incorporated in an estimating procedure which can result in improved properties of standard methods.

Auxiliary information

Auxiliary information can be obtained from different data sources and in different forms such as census data, population based survey reports, results of previous experiments, and expert opinions or assumptions on population parameters. Information from these sources can be presented and used in a number of different

Keywords and phrases. Auxiliary information, multiple data sources, partially grouped samples, convergence rates.

[1] Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin, 53226, USA; starima@hpi.mcw.edu

[2] Clinical Biostatistics, Pfizer Inc., 50 Pequot Avenue, New London, Connecticut, 06320, USA; dmitri.pavlov@pfizer.com

ways. Census data can be used to obtain probability distributions for such parameters of interest as age, gender, household income, etc. Surveys usually report more targeted information, such as a proportion of likely voters favoring democratic (or republican) political platform, or an average income of customers shopping at a supermarket. While using census data, a researcher typically assumes zero variance, but it is not usually a case to ignore sample variability for survey based estimation. Expert opinion can be expressed in a form of a well grounded guess on one or several population parameters. Some experts can impose a set of restrictions on population parameters or an underlying distribution.

In general, auxiliary information can be of the following two types. The first type is auxiliary information of exact nature, such as census data, expert assumptions, or a set of linear restrictions. The second type is information known with some degree of uncertainty, such as survey results, estimates from previous experiments, etc.

Exact auxiliary information

One of the first attempts to use auxiliary information goes back to 1973, when Pugachev [11] suggested to use a correlation effect for incorporating auxiliary information in an estimating procedure. He considered a linear regression of a variable of interest on a variable known from auxiliary information. The latter one is usually called *auxiliary*.

Haberman [6] expressed auxiliary information in a set of linear constraints on probability measures. His idea was to find a probability measure satisfying these constrains and bringing minimum to Kulbak-Leibler divergence with an empirical measure. This approach was also brought up by Dmitriev and Ustinov [2]. They considered projections of probability measures in a class of probability measures defined by additional information. In addition to focusing on many theoretical issues of their method, they gave detailed analysis of projections onto quantile and symmetric classes of distributions. They proved that asymptotic properties of Pugachev's estimators and the estimators obtained by Kulbak-Leibler projections are identical.

In 1986, Chambers and Dunstan [1] presented quantile estimation in the presence of auxiliary variable. They proposed a model-based method incorporating auxiliary information on estimation stage. Rao *et al.* [12] extended their approach to design-based estimators.

The same "ultimate" knowledge on auxiliary variable was used by Holt [5] to modify estimators derived on data inflicted by non-responses and by Zhang [13] who minimized profile likelihood in the presence of constrains imposed by auxiliary information.

Bayesian theory provides an easy-to-use methodology for incorporating auxiliary information. Bayesian inference is based on a posterior risk minimization at an assumed prior distribution. If this prior distribution is known, the parameters of this distribution are also know. If this prior distribution is known in a form of a parametric model with a set of unknown parameters, these parameters are substituted by sample estimates and the prior is called *empirical*.

Non-parametric Bayesian approach provides statistical inference comparable to classical nonparametric inference. A probability model $p(F)$ can be assumed on an underlying distribution F. A standard assumption is a Dirichlet distribution, $p(F) = \mathcal{D}(F_0, M)$, where F_0 is an assumed distribution, and M is the parameter defining the variability around F_0.

The described methods deal mostly with auxiliary information of exact nature. The results of the research described above are not applicable directly if the exact auxiliary information cannot be obtained. Literature on using uncertain auxiliary information is not as extensive. In fact the authors found only few papers that provide methods on the use of auxiliary information presented in a form of statistical estimates.

Even though non-parametrical Bayesian approaches can efficiently emulate some cases of uncertain additional information, these methods also rely on some assumptions of parametric families or random probability measures, which makes them members of a group of exactly known auxiliary information. Empirical Bayesian approaches are usually used as a convenient practical tool for implementing classical Bayesian techniques, and hence, we are not considering them separately from the other Bayesian approaches.

Uncertain auxiliary information

Kuk and Mak [7] used a median derived from an independent sample to improve the standard median estimator.

Kulldorff [8] considered a parameter estimation problem on a partially grouped sample. If we consider frequencies of grouped observations as auxiliary information, the problem of statistical estimation on a partially grouped sample can be thought as a problem of uncertain auxiliary information. Kulldorff used a likelihood based approach for estimating on partially grouped data. He focused specifically on exponential and normal distributions.

In 1991, Gal'chenko and Gurevich [4] extended Pugachev's approach to auxiliary information obtained from a single previous experiment.

In contrast to Gal'chenko and Gurevich's method, this paper describes a more general approach for a situation when auxiliary information is available from several independent data sources. The suggested extension allows to implement auxiliary information obtained from any finite number of previous experiments. Also, there are no strict constrains on rates of convergences for the estimates used in the procedure.

Layout

Section 1 defines notation, develops extensions to Gal'chenko and Gurevich's approach, and assess asymptotic properties of the suggested estimators.

In Section 2 a cumulative distribution function (CDF) estimator with incorporated auxiliary information given by a probability estimate is presented. Empirical cumulative distribution function (ECDF) is modified by this probability estimate. Mathematics of this special case is given in details for illustrative purposes. A numerical example concludes this section.

Section 3 considers mutually uncorrelated auxiliary information. Cumulative hazard function (CHF) is estimated on partially grouped samples. In CHF case, Nelson-Aalen cumulative hazard function estimator is improved by incorporating additional data into the estimating procedure.

A short discussion is situated in Section 4.

1. Methodology

1.1. Notation

Current data set. Let $(\mathcal{X}, \mathcal{F})$ be a measurable space and $\mathcal{P}$ be a set of all probability measures on $(\mathcal{X}, \mathcal{F})$. Suppose $\mathbf{X} = (X_1, ..., X_N)$ is a vector of independent and identically distributed random variables with a common probability distribution $P \in \mathcal{P}$.

Objective. The objective is to estimate a vector of parameters $\Theta = (\theta_1, ..., \theta_S)^T$, where $\theta_s = \int_{\mathcal{X}} \varphi_s(x) \mathrm{d}P(x)$ with a known function $\varphi_s(x)$ defined on $\mathcal{X}$. The quality of this estimation is defined by a risk function $R_\Gamma(\hat{\Theta}) = \Gamma^T \mathrm{cov}(\hat{\Theta}, \hat{\Theta})\Gamma$, where $\Gamma = (\gamma_1, ..., \gamma_S)^T$ is a vector of pre-specified constants and $\hat{\Theta}$ is an unbiased estimator of Θ.

Auxiliary data sources. Auxiliary information is presented in a form of vectors of unbiased estimates $\tilde{\mathcal{B}}_i = (\tilde{\beta}_{i1}, ..., \tilde{\beta}_{iJ_i})^T$ from I independent data sources, $i = 1, ..., I$, J_i denotes the number of estimates from ith source of auxiliary information. The key assumption on auxiliary information is $\tilde{\mathcal{B}}_i$ estimates $\mathcal{B}_i = (\beta_{i1}, ..., \beta_{iJ_i})^T$, which are *shared* parameters of ith auxiliary data source and current data. In other words, the distribution of the current sample (P) and the distribution of the ith auxiliary data (denote this distribution as Q_i) can be different, but $\beta_{ij} = \int \phi_{ij}(y) \mathrm{d}Q_i(y) = \int \varphi_{ij}(x) \mathrm{d}P(x)$, $i = 1, ..., I$ and $j = 1, ..., J_i$. The distributions $Q_i(\cdot)$ and functions $\phi_{ij}(\cdot)$ may be defined on a domain different from $\mathcal{X}$. This key assumption let us think that every $\tilde{\mathcal{B}}_i$ is an estimate of $\mathcal{B}_i$ a set of parameters of P.

Estimators on current data set. Given $\mathbf{X}$ we obtain $\hat{\Theta}$ a vector of unbiased estimators of Θ and $\hat{\mathcal{B}}_i$ vectors of unbiased estimators of $\mathcal{B}_i$.

Some more notation. For further delivery we denote

- $\mathcal{B} = (\mathcal{B}_1^T, ..., \mathcal{B}_I^T)^T$ a $J \times 1$ vector column, where $J = \sum_{i=1}^{I} J_i$;
- $\hat{\mathcal{B}} = (\hat{\mathcal{B}}_1^T, ..., \hat{\mathcal{B}}_I^T)^T$ a $J \times 1$ vector column of estimates on current sample;
- $\tilde{\mathcal{B}} = (\tilde{\mathcal{B}}_1^T, ..., \tilde{\mathcal{B}}_I^T)^T$ a $J \times 1$ vector column of estimates from additional data sources;
- $\mathbf{K}_{22i}^{'} = \text{cov}(\hat{\mathcal{B}}_i, \hat{\mathcal{B}}_i)$ and $\mathbf{K}_{22i}^{''} = \text{cov}(\tilde{\mathcal{B}}_i, \tilde{\mathcal{B}}_i)$ are $J_i \times J_i$ variance covariance matrices;
- $\mathbf{K}_{22}^{'} = \text{cov}(\hat{\mathcal{B}}, \hat{\mathcal{B}})$, $\mathbf{K}_{11} = \text{cov}(\hat{\Theta}, \hat{\Theta})$, and $\mathbf{K}_{22}^{''} = \text{cov}(\tilde{\mathcal{B}}, \tilde{\mathcal{B}})$ are $J \times J$ variance covariance matrices;
- $\mathbf{K}_{12} = \text{cov}(\hat{\Theta}, \hat{\mathcal{B}})$ is a $J \times S$ covariance matrix;
- $\mathbf{K}_{22} = \text{cov}(\hat{\mathcal{B}}, \hat{\mathcal{B}}) + \text{cov}(\tilde{\mathcal{B}}, \tilde{\mathcal{B}})$ is a $J \times J$ variance covariance matrix.

From mutual independence of the sources of auxiliary information, the matrix $\mathbf{K}_{22}^{''} = \text{diag}(\mathbf{K}_{22i}^{''})$ (block diagonal matrix and $\mathbf{K}_{22i}^{''}$ are its diagonal elements, $i = 1, ..., I$).

1.2. Method

A family of unbiased estimators. Let

$$\hat{\Theta}^{\Lambda} = \hat{\Theta} + \Lambda \left(\hat{\mathcal{B}} - \tilde{\mathcal{B}} \right) \tag{1}$$

be a family of unbiased estimators, where

$$\Lambda = \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1J} \\ \dots & \dots & \dots \\ \lambda_{S1} & \dots & \lambda_{SJ} \end{pmatrix}$$

defines all possible estimators in (1).

The smallest dispersion ellipsoid. In a class of positive definite variance covariance matrices $\text{cov}(\hat{\Theta}^{\Lambda}, \hat{\Theta}^{\Lambda})$, a dispersion ellipsoid based on $\text{cov}(\hat{\Theta}^{\Lambda_0}, \hat{\Theta}^{\Lambda_0})$ is called the *smallest* if

$$\Gamma^T \text{cov} \left(\hat{\Theta}^{\Lambda_0}, \hat{\Theta}^{\Lambda_0} \right) \Gamma \leq \Gamma^T \text{cov} \left(\hat{\Theta}^{\Lambda}, \hat{\Theta}^{\Lambda} \right) \Gamma$$

for any $\Gamma = (\gamma_1, ..., \gamma_S)^T$.

Optimal estimator. An estimator is called *optimal* if it defines the smallest dispersion ellipsoid in (1).

Dispersion ellipsoid is invariant to orthogonal transformations and its shape can be uniquely identified through its eigenvalues. After eigenvectors' based orthogonal transformation T the vector $T\hat{\Theta}$ consists of uncorrelated components. Hence, without loss of generality, a vector of estimates $\hat{\Theta} = (\hat{\theta}_1, ..., \hat{\theta}_S)^T$ with mutually uncorrelated components can be assumed, and the class of unbiased estimators (1) can be represented in a form of S equalities

$$\hat{\theta}_s^{\Lambda_s} = \hat{\Theta}_s + \Lambda_s \left(\hat{\mathcal{B}} - \tilde{\mathcal{B}} \right), \tag{2}$$

where $\Lambda_s = (\lambda_{s1}, ..., \lambda_{sJ})$, $s = 1, ..., S$.

The variance of every $\hat{\theta}_s^{\Lambda_s}$ can be minimized independently from the variances of $\hat{\theta}_t^{\Lambda_t}$, $t \neq s$. Moreover, the dependence from Γ disappears, because one-dimensional variance minimization does not depend on a multiplicative constant.

Setting a gradient vector with respect to Λ identically equal to zero, that is

$$\bigtriangledown_{\Lambda} \text{Var} \left(\hat{\theta}^{\Lambda} \right) = 2\mathbf{K}_{12} + 2\Lambda \mathbf{K}_{22} = \mathbf{0}, \tag{3}$$

we find that if $\mathbf{K}_{22}$ is invertible, the solution $\Lambda_0 = \mathbf{K}_{12}\mathbf{K}_{22}^{-1}$ is invariant to any choice of Γ. This Λ_0 brings a minimum to $\Gamma^T \text{cov}(\hat{\Theta}^{\Lambda}, \hat{\Theta}^{\Lambda})\Gamma$ because second derivatives give $\Gamma^T \mathbf{K}_{22} \Gamma > 0$.

Hence, the optimal estimator is

$$\hat{\Theta}^0 = \hat{\Theta} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\left(\hat{\mathcal{B}} - \tilde{\mathcal{B}}\right) \tag{4}$$

and its smallest dispersion ellipsoid is defined by a variance covariance matrix

$$\mathbf{K}^0 = \mathbf{K}_{11} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\mathbf{K}_{12}^T. \tag{5}$$

The optimal estimator (4) is a multivariate multiple linear regression of $\hat{\Theta}$ on $\hat{\mathcal{B}} - \tilde{\mathcal{B}}$. If additional information provides exact values of $\mathcal{B}$, the optimal estimator becomes a multivariate multiple linear regression of $\hat{\Theta}$ on $\hat{\mathcal{B}}$.

The difference of the regressions $\hat{\Theta}$ on $\hat{\mathcal{B}} - \tilde{\mathcal{B}}$ and $\hat{\Theta}$ on $\hat{\mathcal{B}}$ is incorporated in the matrix $\mathbf{K}_{22} = \mathbf{K}_{22}^{'} + \mathbf{K}_{22}^{''}$. If $\mathcal{B}$ is known exactly from an additional data source, $\mathbf{K}_{22} = \mathbf{K}_{22}^{'}$. If $\mathcal{B}$ is estimated by $\tilde{\mathcal{B}}$ from an additional data source, $\mathbf{K}_{22} = \mathbf{K}_{22}^{'} + \mathbf{K}_{22}^{''}$.

Adaptive estimator. The optimal estimator (4) can be obtained only when the matrices $\mathbf{K}_{12}$ and $\mathbf{K}_{22}$ are known. However, in the majority of real life problems they are not available. The simplest solution is to substitute $\mathbf{K}_{12}$ and $\mathbf{K}_{22}$ with their consistent estimates $\hat{\mathbf{K}}_{12}$ and $\hat{\mathbf{K}}_{22}$.

After this substitution

$$\hat{\Theta}^* = \hat{\Theta} - \hat{\mathbf{K}}_{12}\hat{\mathbf{K}}_{22}^{-1}\left(\hat{\mathcal{B}} - \tilde{\mathcal{B}}\right). \tag{6}$$

By analogy with Pugachev's terminology the estimator (6) is called *adaptive*. The variance covariance matrix (5) can be estimated by

$$\hat{\mathbf{K}}^0 = \hat{\mathbf{K}}_{11} - \hat{\mathbf{K}}_{12}\hat{\mathbf{K}}_{22}^{-1}\hat{\mathbf{K}}_{12}^T. \tag{7}$$

The estimators (6) and (7) can be obtained only when $\hat{\mathbf{K}}_{22}$ is invertible. If $\mathbf{K}_{22}$ is invertible, there exists a real valued $\epsilon > 0$ such that the smallest eigenvalue of $\mathbf{K}_{22}$ is greater then this ϵ. As sample size increases, the probability that the smallest eigenvalue of $\hat{\mathbf{K}}_{22}$ is strictly greater than zero goes to one and chances of inability to invert $\hat{\mathbf{K}}_{22}$ go to zero.

1.3. **Large sample properties**

The method of Section 1.2 uses estimators and their dispersions only. The analysis of large sample properties also uses actual sample sizes of the current sample and additional data sources. Let n be a size of a current sample, m_i be a sample size of the ith additional data source, $i = 1, ..., I$.

Set $m_i = f_i(n)$, where $f_i(n)$ is an increasing function, as n going to infinity. Thus, the asymptotical properties of (4) and (6) are tied to n only.

Assume

- $\xi_n = a_n\left(\hat{\Theta} - \Theta\right) \xrightarrow{d} \xi$, where a_n is a sequence of positive real numbers such that $a_n \to +\infty$, $\xi \overset{d}{=} N(\mathbf{0}, \Sigma_{11})$, and $a_n^2\mathbf{K}_{11} \to \Sigma_{11}$;
- $\tau_n = a_n\left(\hat{\mathcal{B}} - \mathcal{B}\right) \xrightarrow{d} \tau$, where $\tau \overset{d}{=} N\left(\mathbf{0}, \Sigma_{22}^{'}\right)$ and $a_n^2\mathbf{K}_{22}^{'} \to \Sigma_{22}^{'}$;
- $\zeta_{in} = b_{in}\left(\tilde{\mathcal{B}}_i - \mathcal{B}_i\right) \xrightarrow{d} \zeta_i$, where b_{in} is a sequence of positive real numbers such that $b_{in} \to +\infty$, $\zeta_i \overset{d}{=} N\left(\mathbf{0}, \Sigma_{22i}^{''}\right)$, $i = 1, ..., I$, and $b_{in}^2\mathbf{K}_{22i}^{''} \to \Sigma_{22i}^{''}$;
- $\zeta_n = \left(\zeta_{1n}^T, ..., \zeta_{In}^T\right)^T$, where $\zeta \overset{d}{=} N\left(\mathbf{0}, \Sigma_{22}^{''}\right)$ and $\Sigma_{22}^{''} = \operatorname{diag}\left(\Sigma_{22i}^{''}\right)$.

Proposition 1. *If* $b_{in}a_n^{-1} \to w_i \in [0, +\infty)$ *and* $\Sigma_{22} = \Sigma_{22}^{'} + \operatorname{diag}(w_i^2\Sigma_{22i}^{''})$ *is positive definite then* $a_n(\hat{\Theta}^0 - \Theta) \xrightarrow{d} N(\mathbf{0}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)$, *where* $\Sigma_{12} = \operatorname{cov}(\xi, \tau)$.

Proof. Let $\eta_n = a_n(\hat{\Theta}^0 - \Theta)$ then $\eta_n = \xi_n - \mathbf{K_{12}}\mathbf{K}_{22}^{-1}\tau_n$. Since the random variable η_n is a linear combination of random variables converging to a normal distribution, η_n converges to a normal random variable, denote η. From $E(\hat{\Theta}^0) = \Theta$ find $E(\eta) = \mathbf{0}$, from (5) $\operatorname{cov}(\eta, \eta) = a_n^2\mathbf{K}^0 = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$. □

The term $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$ is always non-negative and depends on sample sizes only through w_i. If w_i are much larger than 1, the risk function decrease gained by incorporating auxiliary information is very small and the asymptotic properties of the optimal estimator are almost the same as the estimator based on current sample has. If w_i is close to 0 then the estimators obtained from additional data sources provide small variability and the asymptotic properties of the optimal estimator are close to the case of incorporating additional information of exact nature.

Denote $\mathcal{N}_0$ a class of all univariate and multivariate normal random variables with zero means and variance covariance matrices of finite elements.

Proposition 1 provides asymptotic properties of the optimal estimator (4). The asymptotic properties of (6) are presented in

Proposition 2. *Under the assumptions of Proposition 1, if the elements of $a_n^2(\hat{\mathbf{K}}_{12}-\mathbf{K}_{12})$ and $a_n^2(\hat{\mathbf{K}}_{22}-\mathbf{K}_{22})$ converge to random variables from $\mathcal{N}_0$, then $a_n(\hat{\Theta}^*-\Theta)\xrightarrow{d}\eta$ and $a_n^2(\hat{\Theta}^*-\hat{\Theta}^0)$ converges to random variables from $\mathcal{N}_0$. In other words, the asymptotic properties of the optimal and adaptive estimators are the same, and $\hat{\Theta}^*$ converges to $\hat{\Theta}^0$ with a_n^2 rate of convergence.*

Proof. The representation

$$a_n\left(\hat{\Theta}^*-\Theta\right)=a_n\left(\hat{\Theta}^0-\Theta\right)+a_n\left(\hat{\Theta}^*-\hat{\Theta}^0\right)$$

can be rewritten as

$$a_n\left(\hat{\Theta}^*-\Theta\right)=\eta_n-a_n\left(\hat{\mathbf{K}}_{12}\hat{\mathbf{K}}_{22}^{-1}-\mathbf{K}_{12}\mathbf{K}_{22}^{-1}\right)\left(\hat{\mathcal{B}}-\tilde{\mathcal{B}}\right). \tag{8}$$

From (8) it follows that it is enough to show that $(\hat{\mathbf{K}}_{12}\hat{\mathbf{K}}_{22}^{-1}-\mathbf{K}_{12}\mathbf{K}_{22}^{-1})(\hat{\mathcal{B}}-\tilde{\mathcal{B}})$ does not affect the asymptotic properties of $a_n(\hat{\Theta}^*-\Theta)$. The convergence of the elements of $a_n^2(\hat{\mathbf{K}}_{12}-\mathbf{K}_{12})$, $a_n^2(\hat{\mathbf{K}}_{22}-\mathbf{K}_{22})$, and $\hat{\mathcal{B}}-\tilde{\mathcal{B}}$ to random variables from $\mathcal{N}_0$ makes any continuous and differentiable at zero function of these elements converge to a normal random variable at the same rate. So, the elements of $a_n^2(\hat{\mathbf{K}}_{12}\hat{\mathbf{K}}_{22}^{-1}-\mathbf{K}_{12}\mathbf{K}_{22}^{-1})(\hat{\mathcal{B}}-\tilde{\mathcal{B}})$ also converge to random variables from $\mathcal{N}_0$.

Hence, 1) the elements of $a_n(\hat{\mathbf{K}}_{12}\hat{\mathbf{K}}_{22}^{-1}-\mathbf{K}_{12}\mathbf{K}_{22}^{-1})(\hat{\mathcal{B}}-\tilde{\mathcal{B}})$ converge to random variables from $\mathcal{N}_0$ with decreasing variance (as $n\to\infty$), which makes this term uninfluential to asymptotic properties of $a_n(\hat{\Theta}^*-\Theta)$, and 2) the elements of $\hat{\Theta}^*$ converge to $\hat{\Theta}^0$ with a_n^2 rate of convergence. □

The advantage of using a_n instead of $\sqrt{n}$ and b_{in} instead of $\sqrt{m_i}$ is that by this generalization it becomes possible to incorporate in the estimating procedure not only the well-spread estimators converging to their means at $\sqrt{n}$ rate of convergence but also a variety of estimators converging to normal random variables with arbitrary convergence rates. For example, kernel estimators depending on bandwidth provide certain degree of robustness, which importance cannot be underestimated at small sample sizes, however, these estimators suffer from a convergence rate slower than $\sqrt{n}$.

Remark. The estimator (7) estimates a variance covariance matrix of the optimal estimator (4). Since the elements of $a_n^2(\hat{\Theta}^*-\hat{\Theta}^0)$ converge to random variables from $\mathcal{N}_0$, the elements of $a_n^4(\mathrm{cov}(\hat{\Theta}^*,\hat{\Theta}^*)-\mathbf{K}^0)$ also converge to random variables from $\mathcal{N}_0$.

To illustrate merits of the proposed estimators, an example of characteristic function (CF) estimation with auxiliary information is considered in the following section.

1.4. **Characteristic function with auxiliary information**

A CF is defined by

$$\Phi(t)=\int_{-\infty}^{+\infty}\exp(itx)\mathrm{d}F(x),$$

where $t\in(-\infty,+\infty)$ and $\sqrt{i}=-1$.

Since $\Phi(t)$ is a function of t, the estimating procedure incorporating auxiliary information can be presented for an arbitrary time point t.

Let $X_1, ..., X_n$ be a simple random sample with a common distribution function F, then $\theta = \Phi(t)$ can be estimated by a plug in estimator

$$\hat{\theta} = \hat{\Phi}(t) = \int_{-\infty}^{+\infty} \exp(itx) \mathrm{d}F_n(x),$$

where $F_n(\cdot)$ is a ECDF. The estimator $\hat{\Phi}(t)$ is called *empirical characteristic function.*

Additional information comes from two independent data sources.

The first data source provided two estimates $F_{m_1}(a)$ and $F_{m_1}(b)$, where m_1 is the number of observations used for obtaining these estimates, a and b are constants on real line.

From the second data source a mean estimate $\tilde{X}_{m_2}$ is available. The estimator $\tilde{X}_{m_2}$ was obtained on m_2 observations with kernel smoothing.

The types of the additional estimators imply

$$\sqrt{m_1}\left(F_{m_1}(a) - F(a)\right) \xrightarrow{d} N\left(0, F(a) - F^2(a)\right),$$

$$\sqrt{m_1}\left(F_{m_1}(b) - F(b)\right) \xrightarrow{d} N\left(0, F(b) - F^2(b)\right),$$

$$m_2^{-1/3}\left(\tilde{X}_{m_2} - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right),$$

where $\mu = \int_{-\infty}^{+\infty} x \mathrm{d}F(x)$ and $\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 \mathrm{d}F(x)$.

The optimal estimator is

$$\hat{\theta}^0 = \hat{\theta} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\left(\tilde{\mathcal{B}} - \hat{\mathcal{B}}\right), \tag{9}$$

where $\hat{\mathcal{B}} = \left(F_n(a), F_n(b), \bar{X}_n\right)$, $\tilde{\mathcal{B}} = \left(F_{m_1}(a), F_{m_1}(b), \tilde{X}_{m_2}\right)$.

The matrices used in (9) are

- $\mathbf{K}_{12} = \left\| K_{12}^{(i)} \right\|_{i=1,2,3}$, where
  $K_{12}^{(1)} = n^{-1} \int_{-\infty}^{+\infty} \left(\exp(itx) - \Phi(t)\right)\left(I_{[x\le a]} - F(a)\right) \mathrm{d}F(x)$,
  $K_{12}^{(2)} = n^{-1} \int_{-\infty}^{+\infty} \left(\exp(itx) - \Phi(t)\right)\left(I_{[x\le b]} - F(b)\right) \mathrm{d}F(x)$, and
  $K_{12}^{(3)} = n^{-1} \int_{-\infty}^{+\infty} \left(\exp(itx) - \Phi(t)\right)(x-\mu) \mathrm{d}F(x)$;
- $\mathbf{K}_{22} = \left\| K_{22}^{(ij)} \right\|_{i,j=1,2,3}$, where
  $K_{22}^{(11)} = \left(n^{-1} + m_1^{-1}\right) \int_{-\infty}^{+\infty} \left(I_{[x\le a]} - F(a)\right)^2 \mathrm{d}F(x)$,
  $K_{22}^{(22)} = \left(n^{-1} + m_1^{-1}\right) \int_{-\infty}^{+\infty} \left(I_{[x\le b]} - F(b)\right)^2 \mathrm{d}F(x)$,
  $K_{22}^{(33)} = \left(n^{-1} + m_2^{-2/3}\right) \int_{-\infty}^{+\infty} (x-\mu)^2 \mathrm{d}F(x)$,
  $K_{22}^{(12)} = K_{22}^{(21)} = \left(n^{-1} + m_1^{-1}\right) \int_{-\infty}^{+\infty} \left(I_{[x\le a]} - F(a)\right)\left(I_{[x\le b]} - F(b)\right) \mathrm{d}F(x)$,
  $K_{22}^{(13)} = K_{22}^{(31)} = n^{-1} \int_{-\infty}^{+\infty} \left(I_{[x\le a]} - F(a)\right)(x-\mu) \mathrm{d}F(x)$,
  $K_{22}^{(23)} = K_{22}^{(32)} = n^{-1} \int_{-\infty}^{+\infty} \left(I_{[x\le b]} - F(b)\right)(x-\mu) \mathrm{d}F(x)$.

However, the optimal estimator depends on F, which is not available. The adaptive estimator is obtained from the optimal estimator by plugging in F_n instead of an unknown F. Since μ depends on F, a sample mean $\bar{\mu}$ can be used instead, $\bar{\mu} = \bar{X}_n = \sum_{i=1}^n X_i$.

The CF estimator incorporates two correlated probability estimators from the first data source and one smoothed kernel based mean estimator from the second data source, which is independent from the first.

Section 2 presents another illustrative special case, CDF estimation with auxiliary information, and three numerical examples.

2. Cumulative distribution function estimator

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables (i.i.d.r.v.) with an unknown CDF $F(t)$, $t \in (-\infty, \infty)$. In addition to $X_1, \ldots, X_n$, an additional data source provided an estimate $\tilde{F}_m(s)$ based on a simple random sample of m observations, s is a constant on real line. Note, actual observations from this additional data source are not available. Moreover, they can come from a different from F distribution, denote it G. The underlying assumption for incorporating $\tilde{F}_m(s)$ in statistical estimation is the distributions managing random variables in current and additional data are equal at s, that is $G(s) = F(s)$. Let $\tilde{F}_m(s)$ be a sample probability estimate, which means $\sqrt{m}(\tilde{F}_m(s) - F(s)) \xrightarrow{d} N(0, F(s) - F^2(s))$.

To modify an ECDF $F_n(t)$ we consider a class of unbiased estimators

$$F_n^{\lambda}(t) = F_n(t) + \lambda(F_n(s) - \tilde{F}_m(s)) \tag{10}$$

which is a special case of (1).

The optimal parameter λ_0 bringing the smallest in (10) variance is

$$\lambda_0 = -\frac{m}{n+m} \frac{F(\min(s,t)) - F(s)F(t)}{F(s)(1-F(s))}. \tag{11}$$

Then, applying (11) to (10) the optimal estimator becomes

$$F_n^0(t) = F_n(t) - \frac{m}{n+m} \frac{F(\min(s,t)) - F(s)F(t)}{F(s)(1-F(s))}(F_n(s) - \tilde{F}_m(s)). \tag{12}$$

The variance of (12) is

$$\operatorname{var}\left(F_n^0(t)\right) = \frac{F(t)(1-F(t))}{n} - \frac{m}{n(n+m)} \frac{\left(F(\min(s,t)) - F(s)F(t)\right)^2}{F(s)(1-F(s))}. \tag{13}$$

The actual value of λ_0 usually is not available in (12) and (13). In this case, the adaptive estimator (6) can be used, which is simplified to

$$F_n^*(t) = F_n(t) - \frac{m}{n+m} \frac{F_n(\min(s,t)) - F_n(s)F_n(t)}{F_n(s)(1-F_n(s))}(F_n(s) - \tilde{F}_m(s)). \tag{14}$$

According to Proposition 2 the estimator (14) provides the same asymptotic properties as (12).

Remark. The adaptive estimator $F_n^*(t)$ cannot be used when $F_n(s) = 0$ or $F_n(s) = 1$. To avoid it, many different amendments to $F_n^*(t)$ can be used. For example, assigning $F_n^*(t) = F_n(t)$ in these cases resolves this problem.

Denote

$$F_{n+m}(s) = \frac{n}{n+m} F_n(s) + \frac{m}{n+m} \tilde{F}_m(s).$$

If $t \leq s$ then (14) is simplified to

$$F_n^*(t) = F_{n+m}(s) \frac{F_n(t)}{F_n(s)} \tag{15}$$

and if $t > s$ (14) becomes

$$F_n^*(t) = F_{n+m}(s) + [1 - F_{n+m}(s)] \frac{F_n(t) - F_n(s)}{1 - F_n(s)}. \tag{16}$$

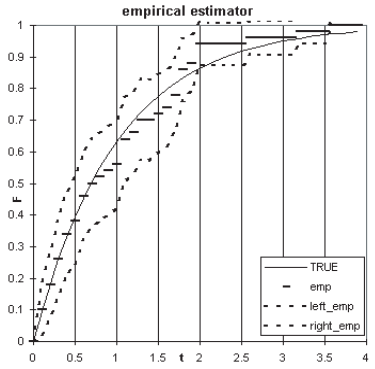

FIGURE 1. Experiment a: empirical estimator, $n = 50$.

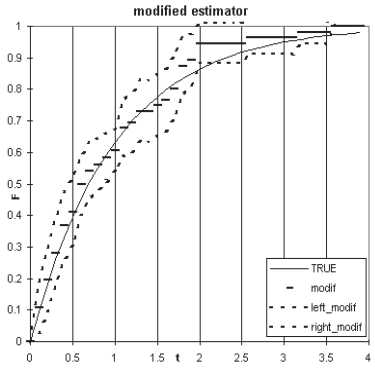

FIGURE 2. Experiment a: modified estimator, $n = 50$, $m = 150$.

From (15) and (16)

$$F_n^*(t) = F_{n+m}(s)\frac{F_n(\min(t,s))}{F_n(s)} + [1 - F_{n+m}(s)]\,\frac{F_n(\max(t,s)) - F_n(s)}{1 - F_n(s)}. \tag{17}$$

The estimator (17) was also obtained by Little and Rubin [9] as a maximum likelihood estimator on two dimensional data with ignorable missingness in the second component.

When the $m(m+n)^{-1}$ goes to 0 the estimator (14) converges to empirical cumulative distribution function estimator. As the $m(m+n)^{-1}$ goes to 1, the estimator (14) goes to the case when auxiliary information is known exactly. Applying the ultimate case with $m = +\infty$ and finite n, the ratio $m(m+n)^{-1} = 1$ and $F_n^*(t)$ becomes

$$F_n^*(t) = F_n(t) - \frac{F_n(\min(s,t)) - F_n(s)F_n(t)}{F_n(s)(1 - F_n(s))}(F_n(s) - F(s)). \tag{18}$$

The (18) can be represented as

$$F_n^*(t) = F(s)\frac{F_n(\min(t,s))}{F_n(s)} + (1 - F(s))\,\frac{F_n(\max(t,s)) - F_n(s)}{1 - F_n(s)}. \tag{19}$$

The estimator (19) is a non-parametric maximum likelihood estimator constructed with a prior knowledge of $F(s)$ [10].

In order to illustrate how $F_n^*(t)$ differs from the empirical cumulative estimator we consider the following

Example. Suppose $X_1, ..., X_n$ are obtained from $F(t) = 1 - \exp(-t)$. The estimate $\tilde{F}_m(1)$ represents auxiliary information. Three experiments were conducted: a) $n = 50$, $m = 150$; b) $n = 50$, $m = 10$; c) $n = 50$, $m = 10\,000$.

In Figures 1, 3, 5 empirical estimators with their 95% confidence intervals are shown. Figures 2, 4, 6 picture adaptive estimators and their 95% confidence intervals.

Figure 2 shows the closer t to 1 (and the stronger the correlation between $F_n(t)$ and $\tilde{F}_m(1)$), the smaller confidence interval becomes. The strongest improvement from incorporating auxiliary information comes at $t = 1$.

Figure 6 describes a situation, where m is much larger than n.

Figures 1 to 6 illustrate that the best improvement of the confidence interval corresponds to $t = 1$ because at this time point the strongest correlation between $\tilde{F}_m(t)$ and $F_n(t)$ is attained.

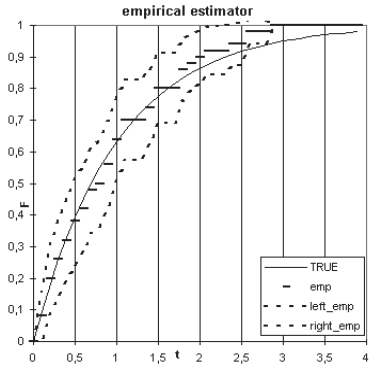

FIGURE 3. Experiment b: empirical estimator, $n = 50$.

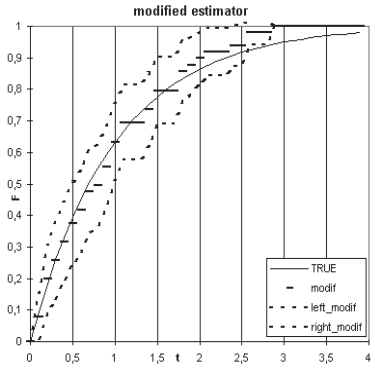

FIGURE 4. Experiment b: modified estimator, $n = 50$, $m = 10$.

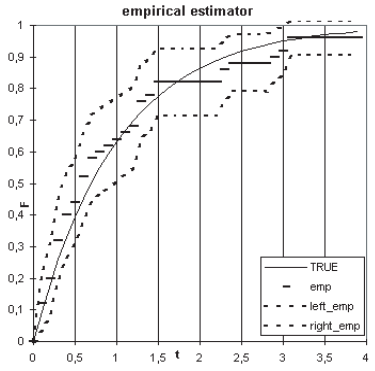

FIGURE 5. Experiment c: empirical estimator, $n = 50$.

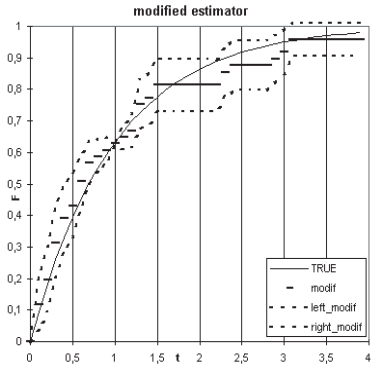

FIGURE 6. Experiment c: modified estimator, $n = 50$, $m = 10\,000$.

3. Mutually uncorrelated auxiliary information

Section 1 considers auxiliary information presented by vectors of estimates obtained from mutually independent data sources.

In this section, we consider a more specific situation, we assume that every additional data source provides one estimator only. Then, in the notation of Section 1 $J_i = 1$, $i = 1, ..., I$. The components of $\tilde{\mathcal{B}}$ are independent and, hence, uncorrelated. The matrix $\mathbf{K}_{22}^{''}$ becomes a diagonal matrix. Another reasonable assumption is the components of $\hat{\mathcal{B}}$ are also uncorrelated. This assumption sets all non-diagonal elements of $\mathbf{K}_{22}^{'}$ to zero. Summarizing these two assumptions the matrix $\mathbf{K}_{22}$ is diagonal with elements $\mathrm{Var}(\hat{\beta}_{i1}) + \mathrm{Var}(\tilde{\beta}_{i1})$. Then, the matrix $\mathbf{K}_{22}^{-1}$ is also diagonal with elements $(\mathrm{Var}(\hat{\beta}_{i1}) + \mathrm{Var}(\tilde{\beta}_{i1}))^{-1}$.

In these assumption, the optimal estimator is

$$\hat{\Theta}^0 = \hat{\Theta} - \sum_{i=1}^{I} \frac{\mathbf{K}_{12}\left(\hat{\beta_{i1}} - \tilde{\beta_{i1}}\right)}{\mathrm{Var}\left(\hat{\beta}_{i1}\right) + \mathrm{Var}\left(\tilde{\beta}_{i1}\right)}, \tag{20}$$

the variance (5) is simplified to

$$\mathbf{K}^0 = \mathbf{K}_{11} - \sum_{i=1}^{I} \frac{\mathbf{K}_{12}\mathbf{K}_{12}^T}{\mathrm{Var}\left(\hat{\beta}_{i1}\right) + \mathrm{Var}\left(\tilde{\beta}_{i1}\right)}, \tag{21}$$

where $\mathbf{K}_{12}$ is a $S \times I$ matrix with elements $\mathrm{cov}\left(\hat{\theta}_s, \hat{\beta}_{i1}\right)$, $s = 1, ..., S$, $i = 1, ..., I$.

3.1. **Cumulative hazard estimator**

Let $Y_1, \ldots, Y_n$ be i.i.d.r.v. with an unknown $F_Y(t)$, $C_1, \ldots, C_n$ be i.i.d.r.v. from $F_C(t)$, $t \in [0, \infty)$, Y is independent from C. If $T_j = \min(Y_j, C_j)$ and $\delta_j = I(Y_j \leq C_j)$, for $j = 1, \ldots, n$, paired observations $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ represent right censored data.

In survival analysis CHF is defined by

$$H(t) = \int_0^t (1 - F_Y(x-))^{-1} \mathrm{d}F_Y(x), \tag{22}$$

where $F(x-) = \lim_{\varepsilon \to 0} F(x - \varepsilon)$, $\varepsilon > 0$.

The standard approach to CHF estimation is to use Nelson-Aalen cumulative hazard estimator

$$\hat{H}_n(t) = \sum_{x \leq t} \frac{d_n(x)}{R_n(x)}, \tag{23}$$

where $d_n(x) = \sum_{i=1}^n \delta_i I(Y_i = x)$ represents the number of uncensored events Y_i registered at x, $R_n(x) = \sum_{i=1}^n I(Y_i \geq x)$ stands for the number of events Y_i registered at or after x.

The variance of (23) can be estimated by Aalen's variance estimator

$$\widehat{\mathrm{var}}(\hat{H}_n(t)) = \sum_{x \leq t} \frac{d_n(x)}{R_n^2(x)}. \tag{24}$$

Fleming [3] is a good reference for further reading on (23) and (24).

Assume that 1)

$$\tilde{H}_m(s_i) = \sum_{x \leq s_i} \frac{d_m(x)}{R_m(x)} \tag{25}$$

were obtained from an additional data source, $i = 1, ..., k$, and 2) $Y_i I\left(s_{j'-1} < Y_i \leq s_{j'}\right)$ is independent from $Y_i I\left(s_{j-1} < Y_i \leq s_j\right)$, where I is an indicator function and $j \neq j'$. The second assumption states that the number of events in periods $(s_{j-1}, s_j]$, $j = 1, ..., k$, are mutually independent.

The elements of variance covariance matrices can be consistently estimated on the basis of given right censored data by

$$\widehat{\mathrm{cov}}\left(\tilde{H}_m(s_i), \tilde{H}_m(s_j)\right) = \widehat{\mathrm{var}}(\tilde{H}_m(\min(s_i, s_j)))$$

and

$$\widehat{\mathrm{cov}}\left(\hat{H}_n(s_i), \hat{H}_n(s_j)\right) = \widehat{\mathrm{var}}(\hat{H}_n(\min(s_i, s_j))).$$

To be able to use uncorrelated auxiliary information we represent auxiliary information as

$$\Delta\tilde{H}\left(s_i\right) = \tilde{H}\left(s_i\right) - \tilde{H}\left(s_{i-1}\right) = \sum_{s_{i-1} < x \leq s_i} \frac{d_m(x)}{R_m(x)}, \tag{26}$$

where $i = 1, ..., k$, $s_0 = 0$, $\tilde{H}(0) = 0$.

By analogy with (26),

$$\Delta\hat{H}(s_i) = \sum_{s_{i-1}<x\leq s_i} \frac{d_n(x)}{R_n(x)}.$$

Since $Y_i I(s_{j-1} < Y_i \leq s_j)$ is independent from $Y_i I(s_{j'-1} < Y_i \leq s_{j'})$, where $j \neq j'$, $\Delta\tilde{H}(s_i)$ and $\Delta\hat{H}(s_i)$ are mutually uncorrelated for any i.

Then, the adaptive cumulative hazard estimator with auxiliary information is

$$\hat{H}_n^*(t) = \sum_{x\leq t} \frac{d_n(x)}{R_n(x)} - \frac{m}{n+m} \sum_{j=1}^{k} \left(\sum_{\min(t,s_{j-1})<x\leq\min(t,s_j)} \frac{d_n(x)}{R_n^2(x)} \left(\sum_{s_{j-1}<x\leq s_j} \frac{d_n(x)}{R_n^2(x)} \right)^{-1} \right) \times \left(\sum_{s_{j-1}<x\leq s_j} \frac{d_n(x)}{R_n(x)} - \sum_{s_{j-1}<x\leq s_j} \frac{d_m(x)}{R_m(x)} \right). \quad (27)$$

The variances of (27) can be estimated by

$$\widehat{\text{var}}\left(\hat{H}_n^*(t)\right) = \sum_{x\leq t} \frac{d_n(x)}{R_n^2(x)} - \frac{m}{n+m} \sum_{j=1}^{k} \left(\sum_{\min(t,s_{j-1})<x\leq\min(t,s_j)} \frac{d_n(x)}{R_n^2(x)} \right)^2 \left(\sum_{s_{j-1}<x\leq s_j} \frac{d_n(x)}{R_n^2(x)} \right)^{-1}. \quad (28)$$

The formulaes (27) and (28) represent estimators of an optimal estimator and its variance. The large sample properties of these estimators are presented in Proposition 2.

4. Conclusion

A problem of using auxiliary information is considered. Auxiliary information is presented in a form of a set of statistical estimates obtained from mutually independent additional data sources. Moment based methodology is developed for incorporating auxiliary information.

The methodology developed in this paper provides extensions for previously known procedures of incorporating auxiliary information. In particular, these extensions are 1) arbitrary convergence rates and 2) multiple sources of auxiliary information. The extension to arbitrary convergence rates provides an opportunity to incorporate in statistical estimation the estimators with a different from $\sqrt{n}$ rates of convergence. For example, many kernel estimators along with advantageous robustness acquire slow convergence rates. Another extension presents the use of additional information from several mutually independent data sources.

Large sample properties are shown for the optimal and adaptive estimators. The first proposition shows that the optimal estimator (the estimator based on known variance covariance matrices) is unbiased and provides the same or smaller variance than the estimator without using auxiliary information. The second proposition proves that the adaptive estimator developed for unknown variance covariance matrices provides the same asymptotic properties as the optimal estimator.

Detailed mathematics for incorporating one estimate from one additional data source in a cumulative distribution function is presented in a separate section. In addition, a numerical example is also provided.

As a special case, the use of mutually uncorrelated additional estimators is considered. A cumulative hazard estimator with auxiliary information is shown as an application of this special case.

Overall, the decision on the use of these estimators depends on the answers to the following questions: 1) how strong is the correlation between the estimator based on current sample and the estimator from an additional data source? 2) do we have a large enough sample size to insure that the adaptive estimator provides properties close to the optimal estimator?

The number of estimates used in an estimating procedure is another issue a researcher should be aware of. As with linear models, the more estimates we involve in a procedure, the larger sample size we need for attaining sufficient quality of estimation.

Acknowledgements. The authors would like to thank the associate editor and referees for their careful review and valuable comments that helped to improve the manuscript. This work was inspired by professor Yu.G. Dmitriev (Tomsk State University, Russia) whose help and guidance is much appreciated.

References

[1] R.L. Chambers and R. Dunstan, Estimating distribution functions from survey data. *Biometrika* **73** (1986) 597–604.
[2] Y.G. Dmitriev and Y.C. Ustinov, *Statistical estimation of probability distribution with auxiliary information* [in Russian]. Tomsk State University, Tomsk (1988).
[3] T.R. Fleming and D.P. Harrington, *Counting processes and survival analysis*. Wiley (1991).
[4] M.V. Gal'chenko and V.A. Gurevich, Minimum-contrast estimation taking into account additional information. *J. Soviet Math.* **53** (1991) 547–551.
[5] D. Holt and D. Elliot, Methods of weighting for unit non-response. The Statistician, Special Issue: *Survey Design, Methodology and Analysis* **40** (1991) 333–342.
[6] S.J. Haberman, Adjustment by minimum discriminant information. *Ann. Statist.* **12** (1984) 121–140.
[7] A.Y.C. Kuk and T.K. Mak, Median estimation in the presence of auxiliary information. *J. R. Statist. Soc. B* **51** (1989) 261–269.
[8] G. Kulldorff, *Contribution to the theory of estimation from grouped and partially grouped samples*. Almqvist & Wiksell, Stockholm (1961).
[9] R.J.A. Little and D.B. Rubin, *Statistical analysis with missing data*. Wiley (2002).
[10] A.B. Owen, *Empirical likelihood*. Chapman and Hall (2001).
[11] V.N. Pugachev, *Mixed methods of determining probabilistic characteristics* [in Russian]. Soviet Radio, Moscow (1973).
[12] J.N.K. Rao, J.G. Kovar and H.J. Mantel, On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **77** (1990) 365–375.
[13] B. Zhang, Confidence intervals for a distribution function in the presence of auxiliary information. *Comput. Statist. Data Anal.* **21** (1996) 327–342.