ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2009 Управление, вычислительная техника и информатика № 3(8)

UDK 519.25

Sergey S. Tarima, Yuriy G. Dmitriev

STATISTICAL ESTIMATION WITH POSSIBLY INCORRECT MODEL ASSUMPTIONS

We combine a consistent (base) estimator of a population parameter with one or several other possibly inconsistent estimators. Some or all assumptions used for calculating the latter estimators may be incorrect. The suggested in the manuscript approach is not restricted to parametric families and can be easily used for improving efficiency of estimators built under nonparametric or semiparametric models. The combined estimator minimizes the mean squared error (MSE) in a family of linear combinations of considered estimators when all variances and covariances used in its structure are known. In real life problems these variances and covariances are estimated generating an empirical version of the combined estimator. The combined estimator as well as its empirical version are consistent. The asymptotic properties of these estimators are presented. The combined estimator is applicable when analysts can use several different procedures for estimating the same population parameter. Different assumptions are associated with the use of each of non-base estimators. Our estimator is consistent in the presence of wrong assumptions for non-base estimating procedures. In addition to theoretical results of this manuscript, simulation studies describe properties of the estimator combining the Kaplan-Meier estimator with the censored data exponential estimator of a survival curve. Another set of simulation examples combine semi-parametric Cox regression with exponential regression on right censored data.

Keywords: *model misspecification*, *robust estimation*, *minimum mean squared error*, *multimodel inference.*

In many applied problems researchers are challenged with statistical estimation of a population parameter θ.

In some cases θ is expressed as a functional $\theta = \int g(y)dF(y)$, where the real valued and possibly multidimensional function $g(y)$ is known but the distribution $F(y)$ may be either completely unknown (nonparametric case) or unknown with some restrictions (for example, symmetric or belongs to a parametric family). Different degrees of uncertainty about $F(y)$ are expressed by different sets of assumptions and lead to different estimating procedures. The common part is that all of these procedures attempt to estimate the same θ. The quality of estimation highly depends on how well assumptions are used in an estimating procedure and whether these assumptions are correct or not.

There are situations when θ is not easily expressed via $\int g(y)dF(y)$, for example, when θ is a regression coefficient or a distribution parameter. Then, a model dependent interpretation should be applied to θ. For example, θ may be defined as a hazard ratio between two groups in a proportional hazards regression model. Different assumptions on a baseline hazard lead to different estimating procedures. Cox model deals with a nonparametric baseline hazard, Weibull and exponential baseline hazards lead to parametric regression models. We emphasize that the interpretation of θ stays the same. Thus different estimating procedures can compete for being used for θ estimation.

Often researchers choose a single estimating procedure and proceed as if underlying assumptions are correct. These procedures start with choosing a functional form of a model and then proceed with variable selection. A detailed review of model selection procedures can be found in [1]. The major focus of recent statistical research is on variable selection methods, see Fraiman [5], Radchenko [8] and Fan [4].

Multimodel inference avoids reliance on a single model via combing several models. Bayesian model averaging is discussed by Hoeting et al. [6]. The frequentist counterpart in presented by Hjort and Claeskens [7].

Hjort and Claeskens performed averaging over a set of parametric models with the same parametric form but different number of variables.

In our work we attempt to improve properties of our base estimator by guessing on additional restrictions and thus creating grounds for using the other possibly more efficient estimators of θ. Our approach can deal with misspecification of a functional model form as well as with misspecification of the set of variables.

Section 2 derives the estimator. Its asymptotic properties are considered in Section 3. Section 4 illustrates performance of the combined estimator for various scenarios of survival function estimation.

1. Estimator

Let $Y_1,...,Y_n$ be an independent sample from an unknown distribution. If there are no additional information of any kind we can estimate θ via a base estimator $\hat{\theta}_n^{(0)}$. Hereafter n in a subscript highlights dependence on a sample of size n. We assume that the base estimator is asymptotically unbiased estimator of θ. Further we assume that there exist S sets of possibly incorrect assumptions. Each of these S assumptions can be used for building another estimator of θ, $\hat{\theta}_n^{(s)}$, $s=1,...,S$. If an assumption, say $(s')^{th}$, is correct, we may reasonably expect that $\hat{\theta}_n^{(s')}$ is a more efficient estimator of θ than $\hat{\theta}_n^{(0)}$. However, it is not known which of the S sets of assumptions are correct and which are not. It is also possible that all sets are false.

In order to avoid dealing directly with different sets of regularity conditions we assume (1) $\forall s \;\; E\hat{\theta}_n^{(s)} = \theta_n^{(s)} \underset{n\to\infty}{\to} \theta^{(s)}$, where $\theta^{(0)}=\theta$, (2) $E[\hat{\theta}_n^{(s)}]<\infty$, $\forall s$, $\forall n$, (3) under a correctly chosen model $a_{ns}\left(\hat{\theta}_n^{(s)}-\theta_n^{(s)}\right)$ has a finite variance for every n including its limiting case at $n=+\infty$, where a_{ns} is a diverging to $+\infty$ sequence of positive numbers as $n\to\infty$. Consider a family of estimators of θ

$$\hat{\theta}_n(\Lambda_n)=\hat{\theta}_n^{(0)}+\sum_{s=1}^{S}\lambda_{ns}\left(\hat{\theta}_n^{(s)}-\hat{\theta}_n^{(0)}\right),$$

where $\Lambda_n=(\lambda_{n0},...,\lambda_{nS})$. The mean squared error of $\hat{\theta}_n(\Lambda_n)$ is

$$E\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}+\sum_{s=1}^{S}\lambda_{ns}\left(\hat{\theta}_n^{(s)}-\hat{\theta}_n^{(0)}\right)\right)^2.$$

For every λ_{ni}

$$\frac{\partial}{\partial\lambda_{ni}}\mathrm{MSE}(\hat{\theta}_n(\Lambda_n))=2E\left[\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)\left(\hat{\theta}_n^{(i)}-\hat{\theta}_n^{(0)}\right)\right]+2\sum\nolimits_{s=1}^{S}\lambda_{ns}E\left[\left(\hat{\theta}_n^{(i)}-\hat{\theta}_n^{(0)}\right)\left(\hat{\theta}_n^{(s)}-\hat{\theta}_n^{(0)}\right)\right],$$

or, in a matrix form

$$\frac{\partial}{\partial\Lambda_n}\mathrm{MSE}\left(\hat{\theta}_n(\Lambda_n)\right)=2E\left[\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)\hat{\Delta}_n\right]+2\Lambda_n E\left[\hat{\Delta}_n\hat{\Delta}_n^T\right],$$

where

$$\hat{\Delta}_n=\left(\hat{\Delta}_n^{(1)},\dots,\hat{\Delta}_n^{(S)}\right)^T=\left(\hat{\theta}_n^{(1)}-\hat{\theta}_n^{(0)},\dots,\hat{\theta}_n^{(S)}-\hat{\theta}_n^{(0)}\right)^T.$$

From

$$\frac{\partial}{\partial\Lambda_n}\mathrm{MSE}\left(\hat{\theta}_n(\Lambda_n)\right)\equiv 0,$$

we find

$$\Lambda_{n0}=-E\left[\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)\hat{\Delta}_n^T\right]E^{-1}\left[\hat{\Delta}_n\hat{\Delta}_n^T\right].$$

Since $\det\left\{E\left[(\hat{\Delta}_n-a)(\hat{\Delta}_n-a)^T\right]\right\}$ is minimized at $a\equiv E\hat{\Delta}_n$ and $\det\left\{\mathrm{cov}\left(\hat{\Delta}_n,\hat{\Delta}_n^T\right)\right\}\geq 0$, the matrix of second derivatives $\frac{\partial}{\partial\Lambda_n\Lambda_n^T}\mathrm{MSE}\left(\hat{\theta}_n(\Lambda_n)\right)=E\left[\hat{\Delta}_n\hat{\Delta}_n^T\right]$ is nonnegative definite, which assures that Λ_{n0} defines the smallest MSE among $\hat{\theta}_n(\Lambda_n)$. The case when the determinant is equal to zero corresponds to multiple solutions for Λ_0, but the MSE stays at its minimum for each of them. The Moore-Penrose generalized inverse can be used for selecting one of these solutions. Then,

$$\hat{\theta}_n(\Lambda_{n0})=\hat{\theta}_n^{(0)}-E\left[\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)\hat{\Delta}_n^T\right]E^{-1}\left[\hat{\Delta}_n\hat{\Delta}_n^T\right]\hat{\Delta}_n, \tag{1}$$

provides the smallest MSE among all $\hat{\theta}_n(\Lambda_n)$ which is

$$E\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)^2-E\left[\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)\hat{\Delta}_n^T\right]E^{-1}\left[\hat{\Delta}_n\hat{\Delta}_n^T\right]E\left[\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)\hat{\Delta}_n^T\right]^T. \tag{2}$$

Due to the quadratic form at the right hand side the mean squared error of $\hat{\theta}_n(\Lambda_{n0})$ is never higher than $\mathrm{MSE}\left(\hat{\theta}_n^{(0)}\right)$. The formulas (1) and (2) cannot be used directly because $E[\cdot]$ in their expressions are not known. Applying $\hat{E}[\cdot]$ instead of $E[\cdot]$ leads to

$$\hat{\theta}_n\left(\hat{\Lambda}_{n0}\right)=\hat{\theta}_n^{(0)}-\hat{E}\left[\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)\hat{\Delta}_n^T\right]\hat{E}^{-1}\left[\hat{\Delta}_n\hat{\Delta}_n^T\right]\hat{\Delta}_n \tag{3}$$

and

$$\widehat{\mathrm{MSE}}(\hat{\theta}_n(\hat{\Lambda}_{n0}))=\hat{E}\left[\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)^2\right]-\hat{E}\left[\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)\hat{\Delta}_n^T\right]\hat{E}^{-1}\left[\hat{\Delta}_n\hat{\Delta}_n^T\right]\hat{E}\left[\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)\hat{\Delta}_n^T\right]^T. \tag{4}$$

Quantities $\hat{E}[\cdot]$ can be plug-in estimators, where the unknown distribution is substituted by its empirical estimator. Further the estimator with $\hat{E}[\cdot]$ instead of $E[\cdot]$ will be called empirical combined (EC) estimator. Following Efron [3] and Davidson and Hinkley [2] we used a nonparametric bootstrap for estimating the unknown $E[\cdot]$. See simulation studies in Section 4. Asymptotic properties of $\hat{\theta}_n(\Lambda_{n0})$ and $\hat{\theta}_n\left(\hat{\Lambda}_{n0}\right)$ are presented in Section 3.

2. Asymptotic properties

Large sample properties of $\hat{\theta}_n^{(s)}$ ($s=1,\dots,S$) are usually known, which allow describing asymptotic results for the combined estimator and its empirical version.

Theorem 1. Let $E\left|\hat{\theta}_n^{(i)}\hat{\theta}_n^{(j)}\right| < +\infty$ $(i,j=0,\dots,S)$ and as $n\to\infty$

1. sequences of positive real numbers $a_{ni}\to\infty$,

2. $\eta_{ni} \overset{d}{=} a_{ni}\left(\hat{\theta}_n^{(i)}-\theta_n^{(i)}\right)\overset{d}{\to}\eta_i$, where $E(\eta_i)=0$, $E\left(\eta_i\eta_j\right)=\sigma_{ij}<\infty$, and $\theta^{(0)}=\theta$,

3. $a_{ni}^{-1}a_{n0}=k_{ni}\to k_i\le 1$.

Then, $\xi_n \overset{d}{=} a_{n0}\left(\hat{\theta}_n(\Lambda_{n0})-\theta\right)\overset{d}{\to}\xi$ with $E(\xi_n)\to E(\xi)=0$ and

$$MSE(\xi_n)\to MSE(\xi)=\sigma_{00}-\left\|k_i\sigma_{0i}+\sigma_{00}\right\|_{i=1,\dots,S}\times$$

$$\times\left\|k_ik_j\sigma_{ij}+k_i\sigma_{0i}+k_j\sigma_{0j}+\sigma_{00}+\tau_{ij}\right\|_{i,j=1,\dots,S}^{-1}\left\|k_j\sigma_{j0}+\sigma_{00}\right\|_{j=1,\dots,S}^{T},$$

where $\tau_{ij}=\lim_{n\to\infty}a_{n0}^2\Delta_n^{(i)}\Delta_n^{(j)}$.

Proof. Since $\mathrm{MSE}\left(\hat{\theta}_n(\Lambda_{n0})\right)\le\mathrm{MSE}\left(\hat{\theta}_n^{(0)}\right)$ by construction, the condition 2 leads to $E(\xi_n)\to 0$.

$$MSE(\xi_n)=a_{n0}^2E\left[\left(\hat{\theta}_n^{(0)}-\theta\right)^2\right]\ -a_{n0}^2E\left[\left(\hat{\theta}_n^{(0)}-\theta\right)\hat{\Delta}_n^T\right]E^{-1}\left[\hat{\Delta}_n\hat{\Delta}_n^T\right]E\left[\left(\hat{\theta}_n^{(0)}-\theta\right)\hat{\Delta}_n^T\right]^T=$$

$$=E\left(\eta_{n0}^2\right)-E\left[\eta_{n0}a_{n0}\hat{\Delta}_n^T\right]E^{-1}\left[a_{n0}\hat{\Delta}_na_{n0}\hat{\Delta}_n^T\right]E\left[\eta_{n0}a_{n0}\hat{\Delta}_n^T\right]^T. \tag{5}$$

Taking into consideration

$$a_{n0}\hat{\Delta}_n^{(i)}=a_{n0}\left(\hat{\theta}_n^{(i)}-\hat{\theta}_n^{(0)}\right)=\frac{a_{n0}}{a_{ni}}a_{ni}\left(\hat{\theta}_n^{(i)}-\theta_n^{(i)}\right)+a_{n0}\left(\hat{\theta}_n^{(0)}-\theta_n^{(0)}\right)+a_{n0}\left(\theta_n^{(i)}-\theta_n^{(0)}\right)=$$

$$=k_{ni}\eta_{ni}+\eta_{n0}+a_{n0}\Delta_n^{(i)},$$

and denoting $k_n=(k_{n1},\dots,k_{nS})^T$, $K_n=diag(k_n)$, $\eta_n=(\eta_{n1},\dots,\eta_{nS})^T$ the mean squared error can be rewritten as

$$\mathrm{MSE}(\xi_n)=E\left(\eta_{n0}^2\right)-E\left[\eta_{n0}\left(K_n\eta_n+\eta_{n0}+a_{n0}\Delta_n\right)\right]\times$$

$$E^{-1}\left[\left(K_n\eta_n+\eta_{n0}+a_{n0}\Delta_n\right)\left(K_n\eta_n+\eta_{n0}+a_{n0}\Delta_n\right)^T\right],$$

$$E\left[\eta_{n0}\left(K_n\eta_n+\eta_{n0}+a_{n0}\Delta_n\right)\right]^T=$$

$$=E\left(\eta_{n0}^2\right)-E\left[\eta_{n0}\left(K_n\eta_n+\eta_{n0}+a_{n0}\Delta_n\right)\right]$$

$$E^{-1}\left[\left(K_n\eta_n+\eta_{n0}+a_{n0}\Delta_n\right)\left(K_n\eta_n+\eta_{n0}+a_{n0}\Delta_n\right)^T\right]\times$$

$$\times E\left[\eta_{n0}\left(K_n\eta_n+\eta_{n0}+a_{n0}\Delta_n\right)\right]^T=$$

$$=E\left(\eta_{n0}^2\right)-\left\|k_{ni}E\left[\eta_{n0}\eta_{ni}\right]+E\left[\eta_{n0}^2+\eta_{n0}a_{n0}\Delta_n^{(i)}\right]\right\|_{i=1,\dots,S}\times$$

$$\times\left\|k_{ni}k_{nj}E\left(\eta_{ni}\eta_{nj}\right)+k_{ni}E\left(\eta_{ni}\eta_{n0}\right)+k_{nj}E\left(\eta_{n0}\eta_{nj}\right)\ +\ E\left(\eta_{n0}^2\right)+a_{n0}^2\Delta_n^{(i)}\Delta_n^{(j)}\right\|_{i,j=1,\dots,S}^{-1}\times$$

$$\times\left\|k_{nj}E\left[\eta_{n0}\eta_{nj}\right]+E\left[\eta_{n0}^2+\eta_{n0}a_{n0}\Delta_n^{(j)}\right]\right\|_{j=1,\dots,S}^{T}.$$

Then, if the inverse exists for every n

$$\mathrm{MSE}(\xi_n) \underset{n\to\infty}{\to} \sigma_{00} - \left\| k_i\sigma_{0i} + \sigma_{00} \right\|_{i=1,\dots,S} \times$$

$$\times \left\| k_i k_j \sigma_{ij} + k_i\sigma_{0i} + k_j\sigma_{0j} + \sigma_{00} + \tau_{ij} \right\|_{i,j=1,\dots,S}^{-1} \times \left\| k_j\sigma_{j0} + \sigma_{00} \right\|_{j=1,\dots,S}^{T},$$

where $\tau_{ij} = \lim_{n\to\infty} a_{n0}^2 \Delta_n^{(i)} \Delta_n^{(j)}$. **Q.E.D.**

A note on the inverse. The τ_{ij} depends on rates of convergence of $\Delta_n^{(i)}$ and $\Delta_n^{(j)}$ and their limits.

1. If $\Delta_n^{(i)}$ converges to a $const \neq 0$ ($\hat{\theta}_n^{(i)}$ is not a consistent estimator of θ) and $a_{n0}^2 \Delta_n^{(j)}$ does not converge to zero then $\tau_{ij} = -\infty$ or $\tau_{ij} = +\infty$.

2. If $\Delta_n^{(i)}$ converges to a $\mathrm{const} \neq 0$ and $\Delta_n^{(j)} = 0$ (unbiased estimator $\forall n$) then $\tau_{ij} = 0$.

3. If convergence rates of $\Delta_n^{(i)}$ and $\Delta_n^{(j)}$ to zero are faster than a_{n0} each then $\tau_{ij} = 0$.

4. In some cases when rates of convergence are the same (for example, $a_{nj} = \sqrt{n}$, $\forall j$) τ_{ij} is a different from zero constant.

Hence, some τ_{ij} may be infinite. However, on a practical side, for every finite n the ∞ is never reached. In the expression for $MSE(\xi_n)$ the matrix with elements

$$k_{ni}k_{nj}E(\eta_{ni}\eta_{nj}) + k_{ni}E(\eta_{ni}\eta_{n0}) + k_{nj}E(\eta_{n0}\eta_{nj}) + E(\eta_{n0}^2) + a_{n0}^2\Delta_n^{(i)}\Delta_n^{(j)}$$

should be inverted. Denote this matrix as $\mathbf{B}_n^{'} = \left\| b_{ij}^{'} \right\|_{i,j=1,\dots,S}$. The inverse of $\mathbf{B}_n^{'}$ always exists but not necessarily unique, which is usually a result of linearly dependent rows (and columns) of $\mathbf{B}_n^{'}$. This comes from linear dependence among some $\hat{\theta}_n^{(i)}$, $i = 1,\dots,S$. If $\det(\mathbf{B}_n^{'}) = 0$ then the use of the Moore-Penrose generalized inverse resolves the multiplicity problem.

Another problem comes when $\mathbf{B}_n^{'}$ is of high dimensionality. Then, a large sample size is needed for estimating $\mathbf{B}_n^{'}$. A possible solution is to use only those principle components which correspond to eigenvalues above some cutoff, say 10% of the sum of eigenvalues.

Theorem 2. If conditions of Theorem 1 hold then the use of $\hat{\theta}_n(\hat{\Lambda}_{n0})$ instead of $\hat{\theta}_n(\Lambda_{n0})$ does not change its asymptotic properties.

Proof. The optimal parameter

$$\Lambda_{n0} = \left\| k_{ni}E[\eta_{n0}\eta_{ni}] + E\left[\eta_{n0}^2 + \eta_{n0}a_{n0}\Delta_n^{(i)}\right] \right\|_{i=1,\dots,S} \times$$

$$\times \left\| k_{ni}k_{nj}E(\eta_{ni}\eta_{nj}) + k_{ni}E(\eta_{ni}\eta_{n0}) + k_{nj}E(\eta_{n0}\eta_{nj}) + E(\eta_{n0}^2) + a_{n0}^2\Delta_n^{(i)}\Delta_n^{(j)} \right\|_{i,j=1,\dots,S}^{-1} \qquad (6)$$

is a continuous function of $E(\eta_{ni}\eta_{nj})$ and $E(\eta_{ni})$ $\forall i, j$, because Λ_{n0} can be represented as a ratio of two polynomials of $E(\eta_{ni}\eta_{nj})$ and $E(\eta_{n0})$. The multiplicity of inverses is resolved through the Moore-Penrose generalized inverse, which is unique. Similarly, $\hat\theta_n(\Lambda_{n0})$ is also a continuous function of $\sigma_{nij} = E\left(\eta_{ni}\eta_{nj}\right)$ and $\mu_{n0} = E(\eta_{n0})$.

From delta method

$$\hat\theta_n(\hat\Lambda_{n0}) = \hat\theta_n(\Lambda_{n0}) + \sum_{i,j=0}^{S} \frac{\partial\theta_n(\Lambda_{n0}^*)}{\partial\sigma_{nij}}(\hat\sigma_{nij} - \sigma_{nij}) +$$

$$+\frac{\partial\theta_n(\Lambda_{n0}^*)}{\partial\mu_{n0}}(\hat\mu_{n0} - \mu_{n0}) = \hat\theta_n(\Lambda_{n0}) + o_P(a_{n0}), \tag{7}$$

where Λ_{n0}^* is located between $\hat\Lambda_{n0}$ and Λ_{n0}. Thus,

$$a_{n0}\left(\hat\theta_n(\hat\Lambda_{n0}) - \theta)\right) = a_{n0}\left(\hat\theta_n(\hat\Lambda_{n0}) - \hat\theta_n(\Lambda_{n0})\right) + a_{n0}\left(\hat\theta_n(\Lambda_{n0}) - \theta)\right) =$$

$$= o_P(1) + a_{n0}\left(\hat\theta_n(\Lambda_{n0}) - \theta)\right), \tag{8}$$

which assures that asymptotic distributions of $\hat\theta_n(\hat\Lambda_{n0})$ and $\hat\theta_n(\Lambda_{n0})$ are the same. **Q.E.D.**

3. Simulation studies

3.1. Combining the Kaplan-Meier estimator with the censored exponential likelihood estimator of a survival function

Consider a right censored sample $T_1,...,T_n$, where $T_i = \min(X_i, C_i)$, $X_i \overset{d}{=} X \sim F_X$, $C_i \overset{d}{=} C \sim F_C$, X is independent of C. This sample is accompanied by $\delta_i = I(T_i < C_i)$. If the exponential family is assumed for X then it depends on a single parameter μ. The censored data likelihood is

$$L(\mu) = \prod_{i=1}^{n}[\mu\exp(-\mu T_i)]^{\delta_i}[\exp(-\mu T_i)]^{1-\delta_i} = \prod_{i=1}^{n}\exp(-\mu T_i)\mu^{\delta_i}.$$

The maximum of $L(\mu)$is reached at $\hat\mu = \left(\sum_{i=1}^{n}\delta_i\right)\left(\sum_{i=1}^{n}T_i\right)^{-1}$ leading to $\hat S_{n1}(t) = \exp(-\hat\mu t)$ an estimate of $S(t)$, which is consistent if the data actually came from an exponential distribution. Otherwise, the Kaplan-Meier (KM) estimator [9], $\hat S_{n0}(t)$, can be used. Our objective is to estimate the survival curve by combing $\hat S_{n0}(t)$ and $\hat S_{n1}(t)$.

At every time point KM estimator is asymptotically normal. For finite sample sizes some normalizing transformations may be needed, see Klein et al. [10] for details. The combined estimator can be used with the transformed estimators of the survival in a similar manner.

The combined estimator $\hat\theta_n(\hat\Lambda_{n0})$ becomes

$$\tilde S_n(t) = \hat S_{n0}(t) - \left(E\left[\hat S_{n0}(t)\hat\Delta_n\right] - S_{n0}(t)\Delta_n^T\right)\hat\Delta_n E\left[\hat\Delta_n^2\right]^{-1},$$

where $\hat{\Delta}_n = \hat{S}_{n0}(t) - \hat{S}_{n1}(t)$ and $\tilde{S}_n(t)$ is the combined estimator of $S(t)$. Estimating $\left(E\left[\hat{S}_{n0}(t)\hat{\Delta}_n\right] - S_{n0}(t)\Delta_n^T\right)$ with $\widehat{cov}\left(S_{n0}(t),\hat{\Delta}_n\right)$ and $E\left[\hat{\Delta}_n^2\right]$ via $\widehat{var}\left(\hat{\Delta}_n\right)+\hat{\Delta}_n^2$ we construct the EC estimator

$$\hat{\tilde{S}}_n(t) = \hat{S}_{n0}(t) - \widehat{cov}\left(S_{n0}(t),\hat{\Delta}_n\right)\hat{\Delta}_n\left(\widehat{var}\left(\hat{\Delta}_n\right)+\left(\hat{\Delta}_n\right)^2\right)^{-1}$$

and

$$\widehat{\text{MSE}}\left(\tilde{S}_n(t)\right) = \widehat{var}\left(\hat{S}_{n0}(t)\right) - \widehat{cov}^2\left(S_{n0}(t),\hat{\Delta}_n\right)\left(\widehat{var}\left(\hat{\Delta}_n\right)+\hat{\Delta}_n^2\right)^{-1}.$$

Simulation settings. To assess performance of $\hat{\tilde{S}}(t)$ we consider two scenarios: (1) $X_1,\ldots,X_n \sim \exp(-t)$ (standard exponential) and (2) $X_1,\ldots,X_n \sim \exp(-t^2)$ (Weibull with the scale parameter equal to 1 and its shape is set to 2). In each case, censoring follows exponential distribution with the rate of 0,75. For estimating the unknown quantities in Λ_0 we used 100 bootstrap samples. Single experiment estimators of the survival curve under different sample sizes (30 and 300) and different distributional assumptions (Exponential and Weibull) are presented on Figure 1. In order to assess MSEs of the estimators 10,000 simulations were performed in each of two Monte-Carlo experiments (exponential and Weibull) for both sample sizes. MSEs from these Monte-Carlo simulations are plotted on Figure 1.

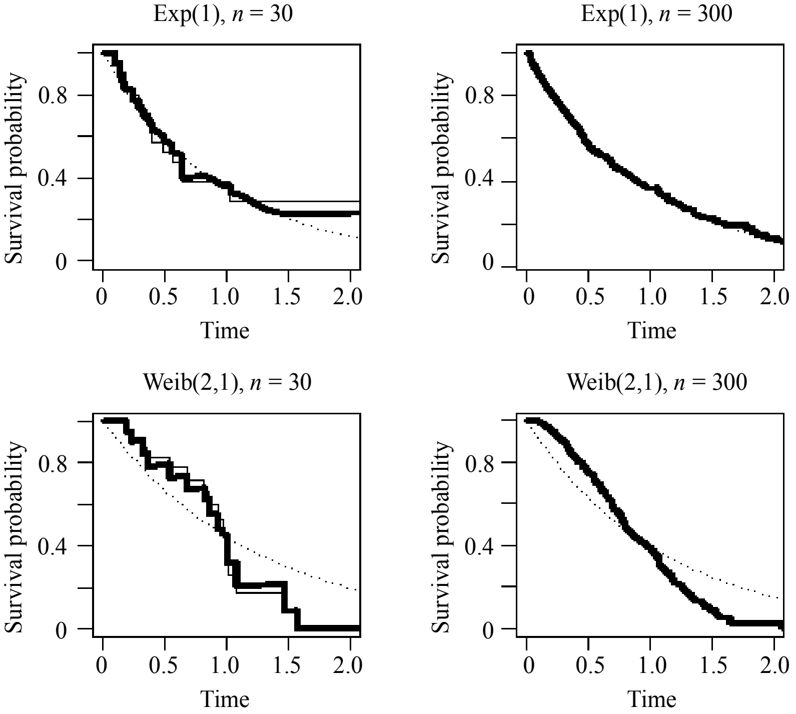

Fig. 1. Four experiments with different sample sizes and distribution. The dotted line is the parametric EXP(1) estimator. The thin solid line is the Kaplan-Meier estimator. The thick solid line is the combined estimator

The behavior of the KM estimator is bounded between 0 and 1 at a fixed sample size and time point is not necessarily normal, which means that it may take a large sample to be able to rely on normal approximation.

Figure 2 shows dynamics of the mean squared error for $\hat{S}_{n0}(t)$, $\hat{S}_{n1}(t)$, and $\hat{\hat{S}}_n(t)$. The data were drawn from the standard exponential distribution (correct model is used for building $\hat{\hat{S}}_n(t)$) and Weibull (incorrect model assumption is used for $\hat{\hat{S}}_n(t)$). Not surprisingly, the MSE of the $\hat{S}_{n1}(t)$ is always smaller then the MSEs of the other two estimators, for the first two pictures. On the other hand the Kaplan-Meier produces the highest MSE among the estimators (this is the price we pay for not using parametric assumptions). The MSE of $\hat{\hat{S}}_n(t)$ is located between the other two estimators and its advantage against the KM estimator is clearly seen till $t = 2.5$.

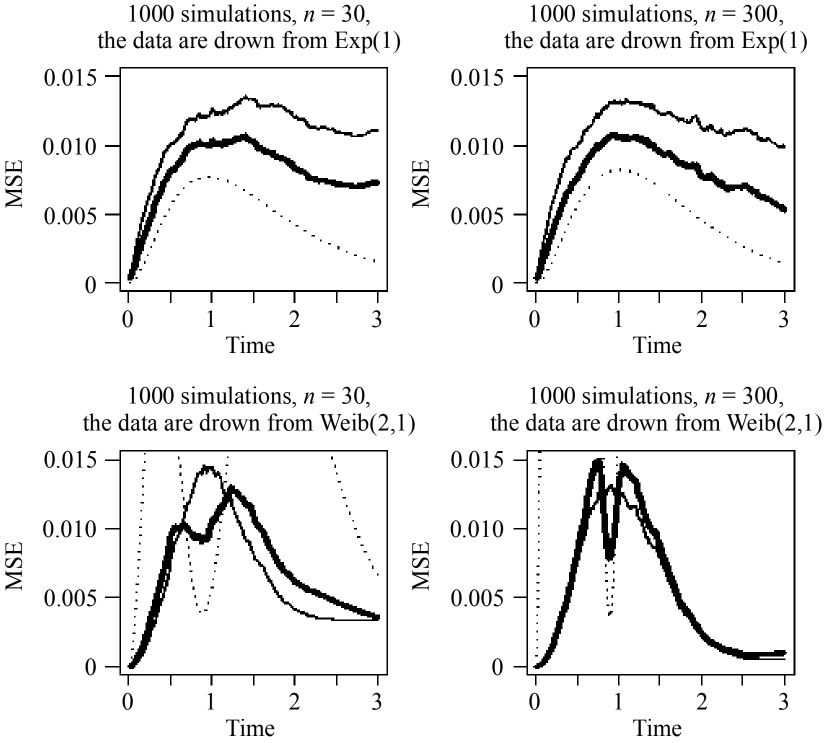

Fig. 2. Monte-Carlo MSEs. The dotted line is the Monte-Carlo MSE of the parametric EXP(1) estimator. The thin solid line is the Monte-Carlo MSE of the Kaplan-Meier estimator. The thick solid line is the Monte-Carlo MSE of the combined estimator

The last two pictures calculate $\hat{\hat{S}}_n(t)$ with a wrong parametric assumption at $n = 30$ and $n = 300$. The standard exponential assumption is violated, the data are coming from a Weibull distribution. In the ranges where the MSE of the fitted standard exponential is much higher than the MSE of the KM estimator, the MSE of $\hat{\hat{S}}_n(t)$ is close to the MSE of the KM estimator. In the place where the exponential survival crosses the KM survival curve the MSE of the $\hat{\hat{S}}_n(t)$ is slightly smaller than the KM MSE: at this time point a wrong parametric assumption leads to an unbiased estimation of the survival probability (weibull and standard exponential survival curves cross). To the left ($t \in (0.3, 0.8)$) and to the right ($t \in (1.1, 1.6)$) of the crossing area, the MSE of $\hat{\hat{S}}_n(t)$ is slightly higher than the MSE of the KM estimator. This slight increase in MSE does not violate Theorem 2, the MSE increase comes from the variability associated with Λ_{n0} estimation. At the same time, moving further away from the crossing point this MSE difference goes to zero. In all these situations (except probably the crossing point area) the use of $\hat{\hat{S}}_n(t)$ is preferable to the use of $\hat{S}_{n1}(t)$.

3.2. Combining regression parameter estimators from proportional hazards and exponential models

Cox proportional hazards model allows estimating log hazards ratios (regression coefficients of the model) under a nonparametric baseline hazard. However, if a parametric form of the baseline hazard is known then a Cox model is not a most efficient model for estimating log hazards ratios. For example, if the baseline hazard is constant then Cox proportional hazard model can be safely substituted by censored data exponential regression. Censored data Weibull regression can be used with the Weibull baseline hazard. If the proportional hazard assumption is violated for one or several covariates, the stratified on these covariates Cox proportional hazards model can be used. All these models (stratified Cox, Cox, Weibull, and exponential regressions) adjust for confounding effects of the same variables and the interpretation of regression parameters continue being the same. The difference between models is solely incorporated in the baseline hazard assumptions. The least restricted of these four is the stratified Cox model, which is built on stratum specific nonparametric baseline hazards. If we assume that the hazard in all strata is the same, the Cox model can be used. Further, assigning Weibull hazard to the baseline only two baseline hazard parameters are to be estimated (shape and scale). Setting the scale parameter equal to one the Weibull hazard becomes constant.

In this section we present a Monte-Carlo study with 10,000 repetitions. Cox model regression parameters will be improved under the constant baseline hazard guess.

Simulation settings. We generate N independent multidimentional observations (Y, A, B, C, D, E, F), where $A \sim Bern(0.5)$, $B \sim Bern(0.5)$, $C \sim Bern(0.5)$, $D \sim Bern(0.5)$, $E \sim N(0,1)$, $F \sim N(0,1)$ and $Y \sim Exp(\beta_A A + \beta_B B + \beta_C C + \beta_D D + \beta_E E + \beta_F F)$, $\beta = (\beta_A, \beta_B, \beta_C, \beta_D, \beta_E, \beta_F) = (-1, 0, 1, 0.5, 0.2, -0.5)$. Actual sample sizes ($N$) used with different simulation settings are presented in captions for Tables 1, 2, 3, and 4.

Monte-Carlo Experiment 1**:** Constant baseline hazard, six predictors, constant baseline hazard. Results of 10,000 Monte-Carlo experiments are given in Table 1. This table shows that MSEs of exponential regression parameter estimates are up to 25% smaller than MSEs of Cox model regression parameter estimates. This is not surprising since the baseline hazard is constant in our experiment and a censored data exponential regression is a valid alternative to the Cox model. MSEs of parameter estimates of the

EC estimator are up to 10% smaller than MSEs of Cox model regression parameter estimates.

Table 1

Regression parameter estimates under an exponential baseline, $\beta = (-1, 0, 1, 0.5, 0.2, -0.5)$, the sample size $N = 75$

Monte Carlo means and root MSEs	$\hat{\beta}_A$	$\hat{\beta}_B$	$\hat{\beta}_C$	$\hat{\beta}_D$	$\hat{\beta}_E$	$\hat{\beta}_F$
	(RMSE)	(RMSE)	(RMSE)	(RMSE)	(RMSE)	(RMSE)
Cox	–1.0768	0.0119	1.0680	0.5412	0.2244	–0.5407
	(0.4232)	(0.3755)	(0.4062)	(0.3824)	(0.1872)	(0.2085)
Exponential	–1.0208	0.0124	1.0159	0.5109	0.2119	–0.5101
	(0.3497)	(0.3096)	(0.2987)	(0.2988)	(0.1716)	(0.5373)
EC	–1.0618	0.0185	1.0690	0.5457	0.2222	–0.5279
	(0.3994)	(0.3503)	(0.3687)	(0.3526)	(0.1850)	(0.2009)

In order to estimate the unknown expectations we use nonparametric bootstrap. After this bootstrap based estimation the EC estimator is not optimal in terms of the smallest MSE. Moreover, the higher dimensionality of $\hat{\Delta}_n$ the more expectations should be estimated. To balance the dimensionality of $\hat{\Delta}_n$ and the amount of noise associated with its estimation we set to zero all eigenvalues contributing less than 10% from the sum of all eigenvalues. Thus, only several (less or equal than six, often two) principal components are used in the estimating procedure.

***Monte-Carlo Experiment 2*:** constant baseline hazard, six predictors, an incorrect number of parameters for the exponential model (assumed $\beta_C = \beta_D = \beta_E = \beta_F = 0$). Results of 10,000 Monte-Carlo experiments are given in Table 1. From this table we see that a wrong assumption shows a minor influence on the EC estimator. Moreover, since maximum likelihood estimators do not provide the smallest MSE and we may occasionally see a better MSE for the EC estimator.

Table 2

Regression parameter estimates under an exponential baseline, $\beta = (-1, 0, 1, 0.5, 0.2, -0.5)$, $N = 75$. We use an incorrect assumption $(\beta_C = \beta_D = \beta_E = \beta_F = 0)$ for the exponential model

Monte Carlo means and root MSEs	$\hat{\beta}_A$	$\hat{\beta}_B$	$\hat{\beta}_C$	$\hat{\beta}_D$	$\hat{\beta}_E$	$\hat{\beta}_F$
	(RMSE)	(RMSE)	(RMSE)	(RMSE)	(RMSE)	(RMSE)
Cox	–1.1056	–0.0055	1.0930	0.5521	0.2224	–0.5431
	(0.4227)	(0.3734)	(0.4118)	(0.3787)	(0.1983)	(0.2117)
Exponential	–0.5101	0.4995	0	0	0	
	(0.6020)	(0.5992)	(1.0000)	(0.5000)	(0.2000)	(0.5000)
EC	–1.0628	0.0178	1.0168	0.5148	0.2145	–0.5250
	(0.4059)	(0.3629)	(0.3889)	(0.3538)	(0.1906)	(0.2018)

Tables 1 and 2 show only a minor improvement associated with the use of the EC estimator.

***Monte-Carlo Experiments 3 and 4*:** constant baseline, 2-predictor case. Consider a simpler case with $\beta=(\beta_A,\beta_B)=(-1,1)$ and $N=40$. Table 3 presents results of Experiment 3: a correct guess, constant baseline hazard. Table 4 shows the results of the Experiment 4: an incorrect guess, we correctly assumed constant baseline hazard but we also assumed $\beta_B=0$. We observe a higher decrease of MSE comparing with the 6-predictor case (Tables 1 and 2).

In Tables 3 and 4 the MSE of the EC estimator is not larger than the MSE of Cox regression parameter estimators. Moreover, we observe an interesting effect when partially incorrect model assumptions may actually make MSE of the EC estimator smaller comparing to the Cox model.

Table 3

Regression parameter estimates under different model assumptions (Cox, exponential, or combined estimators), $\beta=(-1,1)$, $N=40$. In this model we use a correct guess that the baseline hazard is constant

Monte Carlo means and root MSEs	$\hat\beta_A$	$\hat\beta_B$
	(RMSE)	(RMSE)
Cox	−1.0933	1.0420
	(0.5543)	(0.5364)
Exponential	−1.0368	1.0168
	(0.4317)	(0.3399)
EC	−1.0353	1.0011
	(0.4830)	(0.4061)

Table 4

Regression parameter estimates under an exponential baseline and $\beta=(-1,1)$, $N=40$. We guessed that the baseline hazard is constant (correct) and $\beta_B=0$ (wrong)

Monte Carlo means and root MSEs	$\hat\beta_A$	$\hat\beta_B$
	(RMSE)	(RMSE)
Cox	−1.0700	1.0976
	(0.6093)	(0.5920)
Exponential	−0.4745	0
	(0.6842)	(1)
EC	−0.8662	0.6820
	(0.4930)	(0.5008)

***Monte-Carlo Experiment 5*:** the same as Table 1 but with $N=500$. Results of Experiment 5 are presented in Table 5, where we can see results similar to our previous experiments findings. Slight improvement of MSE are seen for the EC estimator, except for only one case: the MSE of $\hat\beta_E$ became a little bit higher. This is a result of either a simulation error or Λ_{n0} estimation.

Table 5

Regression parameter estimates under constant baseline hazard, $\boldsymbol{\beta = (-1, 0, 1, 0.5, 0.2, -0.5)}$, $\boldsymbol{N = 500}$.

Monte Carlo means and root MSEs	$\hat{\beta}_A$	$\hat{\beta}_B$	$\hat{\beta}_C$	$\hat{\beta}_D$	$\hat{\beta}_E$	$\hat{\beta}_F$
	(RMSE)	(RMSE)	(RMSE)	(RMSE)	(RMSE)	(RMSE)
Cox	–1.0180	–0.0030	1.0089	0.5057	0.2001	–0.5063
	(0.1334)	(0.1265)	(0.1293)	(0.1224)	(0.0623)	(0.0706)
Exponential	–0.0122	0.0032	1.0041	0.5035	0.1987	–0.5032
	(0.1204)	(0.1097)	(0.1089)	(0.1037)	(0.0615)	(0.0655)
EC	–1.0164	0.0037	1.0087	0.5061	0.2003	–0.5063
	(0.1291)	(0.1204)	(0.1209)	(0.1149)	(0.0624)	(0.0694)

Conclusion

In this manuscript we suggest an estimating procedure combining a consistent estimator of a population parameter with one or several others possibly inconsistent estimators. The combined estimator provides the smallest MSE among all possible linear combinations between the base and the other estimators. If assumptions used for constructing one or more of the non-base estimators are correct than we expect that the combined estimator will have a smaller MSE than the MSE of the base estimator. If there is no correlation between the base and non-base estimators, the combined estimator is equal to the base estimator. The combined estimator depends on unknown second moments. Their estimation is preformed via nonparametric bootstrap leading to the empirical combined (EC) estimator. This estimator uses the first two moments of the estimators incorporated in its structure.

The EC estimator is consistent and can be used for improving efficiency of nonparametric estimators in the presence of a possibly more efficient parametric estimator. If the parametric estimator is calculated under an incorrect model assumption, its limiting risk is not higher than the limiting risk of the original nonparametric estimator. A simulation example for improving efficiency of the Kaplan-Meier estimator with a parametric model guess illustrates the use of the EC estimator.

Our approach allows to perform multimodel inference in the presence of model misspecification. Comparing to Hjort and Claeskens [7] model averaging approach, we do not restrict our estimating procedure to variable selection in a family of parametric models. Simulation studies show how a Cox regression parameter estimates can be combined with a parametric model regression estimates.

REFERENCES

1. *Burnham P.B., Anderson D.R.* Model selection and multimodel inference. Springer, 2002.
2. *Davidson A.C. & Hinkley D.V.* Bootstrap methods and thier applications. Cambridge: Cambridge University Press, 1997.
3. *Efron B.* Censored data and the bootstrap // J. Amer. Statist. Ass. 1981. V. 76. P. 312 – 319.
4. *Fan J., Li R.* Variable selection via penalized likelihood // J. Amer. Statist. 2001. Ass. 96. P. 1348 – 1360.
5. *Fraiman R., Justel A., Svarc M.* Selection of variables for cluster analysis and classification rules // J. Amer. Statist. 2008. Ass. 103. P. 1294 – 1303.
6. *Hoeting J.A., Madigan D., Raftery A.E., Volinsky C.T.* Bayesian model averaging: a tutorial // Statistical Science. 1999. V. 19. P. 382 – 417.

7. *Hjort N.L.*, *Claeskens G.* Frequentist Model Average Estimators // J. Amer. Statist. Ass. 2003. V. 98. P. 879 – 899.
8. *Radchenko P.*, *James G.M.* Variable inclusion and shrinkage algorithm // J. Amer. Statist. Ass. 2008. V.103. P. 1304 – 1315.
9. *Kaplan E.L.*, *Meier P.* Nonparametric estimator from incomplete observations // J. Amer. Statist. Ass. 1958. V. 53. P. 457 – 481.
10. *Klein J.P.*, *Logan B.*, *Harhoff M.*, *Andersen P.K.* Analyzing survival curves at a fixed point in time // Statistics in Medicine. 2007. V. 26. P. 4505 – 4519.
11. *Klein J.P.*, *Moshenberg M.L.* Survival Analysis. Springer, 2003.

Sergey S. Tarima
Division of Biostatistics Department of Population Health
Medical College of Wisconsin Milwaukee, WI, 53226, USA
E-mail: starima@hpi.mcw.edu
Yuriy G. Dmitriev
Informatics Problems Department of Siberian Branch
of Russian Academy of Sciences (Tomsk) and Tomsk State University
E-mail: dmit@mail.tsu.ru

Поступила в редакцию 16 апреля 2009 г.