NONPARAMETRIC MULTIPLICATIVE DECONVOLUTION IN SURVIVAL ANALYSIS

Sergio Brenner Miguel [1] & Nathawut Phandoidaen [2]

[1] *Institut für angewandte Mathematik, Im Neuenheimer Feld 205, Heidelberg University, Germany; brennermiguel@math.uni-heidelberg.de*
[2] *Institut für angewandte Mathematik, Im Neuenheimer Feld 205, Heidelberg University, Germany; phandoidaen@math.uni-heidelberg.de*

Abstract. We study the non-parametric estimation of an unknown survival function S with support on $\mathbb{R}_+$ based on a sample with multiplicative measurement errors. The proposed fully-data driven procedure is based on estimation of the Mellin transform of the survival function and a regularisation of the inverse of the Mellin transform by a spectral cut-off. The upcoming bias-variance trade-off is handled by a data-driven choice of the cut-off parameter. For the analysis of the variance term, we consider the i.i.d. case and incorporate dependent observations in form of Bernoulli shift processes and β-mixing sequences.

Keywords. MSC2010 Primary 62G05; secondary 62N02, 62C20;
Survival analysis, Non-parametric statistics, Multiplicative measurement errors, Functional dependence measure, Adaptivity

1 Data-driven survival function estimator under mulitplicative measurement errors

1.1 The model

In this work we are interested in estimating the unknown survival function $S : \mathbb{R}_+ \to \mathbb{R}_+$ of a positive random variable X, defined as

$$S : \mathbb{R}_+ \to [0, 1], x \mapsto \mathbb{P}(X > x),$$

given identically distributed copies of $Y = XU$ where X and U are independent of each other and U has a known density $g : \mathbb{R}_+ \to \mathbb{R}_+$. In this setting the survival function $S_Y : \mathbb{R}_+ \to \mathbb{R}_+$ of Y is given by

$$S_Y(y) := \int_0^\infty S(x)g(y/x)dx, \quad y \in \mathbb{R}_+.$$

The estimation of S using a sample $Y_1, \dots, Y_n$ from Y is thus an inverse problem called multiplicative deconvolution. We will allow for certain dependency structures on the sample $Y_1, \dots, Y_n$. More precisely, we assume that $X_1, \dots, X_n$ is a stationary process while the error terms $U_1, \dots, U_n$ will be independent and identically distributed (i.i.d.).

1.2 Estimation strategy

To solve this particular multiplicative deconvolution problem, we use the rich theory of Mellin transforms, in analogy to Brenner Miguel et al (2021). In fact, for a positive random variable Z and $c \in \mathbb{R}$ with $\mathbb{E}(Z^{c-1}) < \infty$ we can define the Mellin transform of the distribution $\mathbb{P}^Z$ as the function

$$\mathcal{M}_c[\mathbb{P}^Z] : \mathbb{R} \to \mathbb{C}, \quad t \mapsto \mathbb{E}(Z^{c-1+it}).$$

As a direct consequence of this definition, we get the convolution theorem, which states for two positive, independent random variable Z_1, Z_2 with $\mathbb{E}(Z_1^{c-1}), \mathbb{E}(Z_2^{c-1}) < \infty$, we have $\mathcal{M}_c[\mathbb{P}^{Z_1 Z_2}] = \mathcal{M}_c[\mathbb{P}^{Z_1}]\mathcal{M}_c[\mathbb{P}^{Z_2}]$. More general, we define for any function $h \in \mathbb{L}^1(\mathbb{R}_+, x^{c-1})$, the space of all measure function with $\int_0^\infty |h(x)| x^{c-1} dx < \infty$, the Mellin transform $\mathcal{M}_c[h] : \mathbb{R} \to \mathbb{C}$ by

$$\mathcal{M}_c[h](t) := \int_0^\infty h(x) x^{c-1+it} dx, \quad t \in \mathbb{R}.$$

Using this definition, we can state that, under the assumption $\mathbb{E}(X^{1/2}) < \infty$, the following calculation rule of the Mellin transform of a survival function holds true

$$\mathcal{M}_{1/2}[S](t) = (1/2 + it)^{-1} \mathcal{M}_{3/2}[\mathbb{P}^X](t), \quad t \in \mathbb{R}.$$

Addtionally, $\mathbb{E}(X^{1/2})$ implies that $S \in \mathbb{L}^2(\mathbb{R}_+)$, that is $\|S\|^2 := \int_0^\infty S^2(x) dx < \infty$, and

$$S(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} x^{-1/2-it} \frac{\mathcal{M}_{3/2}[\mathbb{P}^X](t)}{(1/2 + it)} dt, \quad x \in \mathbb{R}_+,$$

using the inverse Mellin transform, compare Brenner Miguel et Phandoidaen (2021). Thus, we propose the spectral-cut off estimator $\widehat{S}_k$ for $k \in \mathbb{R}_+$ by

$$\widehat{S}_k(x) := \frac{1}{2\pi} \int_{-k}^{k} x^{-1/2-it} \frac{\widehat{\mathcal{M}}(t)}{(1/2 + it)\mathcal{M}_{3/2}[g](t)} dt, \quad \text{with } \widehat{\mathcal{M}}(t) := n^{-1} \sum_{j=1}^{n} Y_j^{1/2+it}. \qquad (1)$$

1.3 Upper bounds and data-driven method

Setting $S_k := \mathbb{E}(\widehat{S}_k)$ and assuming that $\mathbb{E}(Y) < \infty$ we deduce the usual squared bias variance decomposition

$$\mathbb{E}(\|\widehat{S}_k - S\|^2) = \|S - S_k\|^2 + \mathbb{E}(\|\widehat{S}_k - S_k\|^2),$$

where we can decompose the variance term in a term dependent on the underlying inverse problem and a term driven by the dependence structure, that is

$$\mathbb{E}(\|\widehat{S}_k - S_k\|^2) \leq \frac{\mathbb{E}(Y_1)}{2\pi n}\int_{-k}^{k} |(1/2+it)\mathcal{M}_{3/2}[g](t)|^{-2}dt + \frac{1}{2\pi}\int_{-k}^{k} \mathbb{V}\mathrm{ar}(\mathcal{M}_{1/2}[\widehat{S}_X](t))dt$$

where $\widehat{S}_X(x) := n^{-1}\sum_{j=1}^n \mathbb{1}_{(0,X_i)}(x)$ is the empirical survival function. In total, we can show for any $k \in \mathbb{R}_+$,

$$\mathbb{E}(\|\widehat{S}_k - S\|^2) \leq \|S - S_k\|^2 + \mathbb{E}(Y_1)\frac{\Delta_g(k)}{n} + \frac{1}{2\pi}\int_{-k}^{k} \mathbb{V}\mathrm{ar}(\mathcal{M}_{1/2}[\widehat{S}_X](t))dt,$$

where $\Delta_g(k) := (2\pi)^{-1}\int_{-k}^{k}|(1/2+it)\mathcal{M}_{3/2}[g](t)|^{-2}dt$. In other words, we are able to decompose the risk of our estimator into a squared bias term, an variance term which is driven by the inverse problem and a variance term which is dependent on the dependence structure of the sample $X_1, \ldots, X_n$.

For the case of independent observations $X_1, \ldots, X_n$ we deduce that

$$\frac{1}{2\pi}\int_{-k}^{k} \mathbb{V}\mathrm{ar}(\mathcal{M}_{1/2}[\widehat{S}_X](t))dt \leq \frac{\mathbb{E}(X)}{n}$$

making it neglectable. For different dependency structures, for instance beta mixing or functional dependency measures, upper bounds for the second variance term are derived in Brenner Miguel et Phandoidaen (2021).

1.4 Data-driven choice of $\widehat{k}$

While the squared bias term $\|S - S_k\|^2$ is decreasing for k increasing, the variance term $\mathbb{E}(\|\widehat{S}_k - S_k\|^2)$ is increasing. This contrary behaviour of these terms implies that the choice of a suitable cut-off parameter $k \in \mathbb{R}_+$ is non-trivial. To handle this bias-variance dilemma, we suggest a data-driven choice of the parameter $k \in \mathbb{R}_+$ based on a penalized contrast approach, that is, for a $\chi > 0$ we set

$$\widehat{k} := \arg\min_{k \in \mathcal{K}_n} -\|\widehat{S}_k\|^2 + 2\chi\widehat{\sigma}_Y\Delta_g(k)n^{-1},$$

where $\mathcal{K}_n := \{k \in \{1, \ldots, n\} : \Delta_g(k) \leq n\}$ and $\widehat{\sigma}_Y := n^{-1}\sum_{j=1}^n Y_j^{1/2}$. Then under regularity assumptions on g, and the moment assumption $\mathbb{E}(Y_1^{5/2}) < \infty$, we can state that for all $\chi > 96$,

$$\mathbb{E}(\|S - \widehat{S}_{\widehat{k}}\|^2) \leq 6 \inf_{k \in \mathcal{K}_n} \big(\|S - S_k\|^2 + 2\chi\mathbb{E}(Y_1)\frac{\Delta_g(k)}{n}\big) + C(g,f)\left(\frac{1}{n} + \mathbb{V}\mathrm{ar}(\widehat{\sigma}_X) + \int_{-n}^{n} \mathbb{V}\mathrm{ar}(\mathcal{M}_{1/2}[\widehat{S}_X](t))dt\right)$$

where $C(g,f)>0$ is a constant depending on χ, the error density g, $\mathbb{E}(X_1^{5/2})$, $\sigma_X := \mathbb{E}(X_1)$ and $\widehat{\sigma}_X := n^{-1}\sum_{j=1}^n X_j$. Thus, we can say that the data-driven estimator $\widehat{S}_{\widehat{k}}$ realises the optimal choice of $k \in \mathbb{R}_+$ among the set $\mathcal{K}_n$ which minimise the sum of the squared bias and variance term up to an access risk consisting of an neglectable n^{-1} term and a term driven by the dependency structure of the sample $X_1, \dots, X_n$. Again, in the case of independent observations this term is neglectable, too.

1.5 Graphiques

To visualise the finite sample behaviour of our estimator, we present an extract of the simulation study of Brenner Miguel et Phandoidaen (2021) where we visualise the estimator $\widetilde{S}_{\widehat{k}} := \max(\min(\widehat{S}_{\widehat{k}}, 1), 0)$ where obviously $\|\widetilde{S}_{\widehat{k}} - S\|^2 \leq \|\widehat{S}_{\widehat{k}} - S\|^2$ holds.

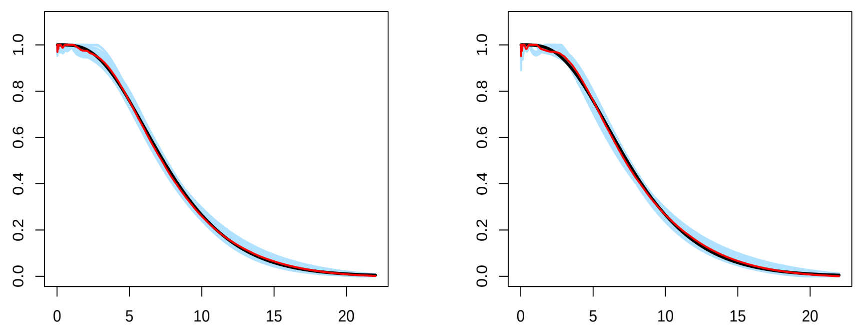

Figure 1: Considering the estimators $\widetilde{S}_{\widehat{k}}$, we depict 50 Monte-Carlo simulations with error densities $g(x) = \mathbb{1}_{(0,1)}(x)$ (left) and $g(x) = \mathbb{1}_{(0.5,1.5)}(x)$ (right) with $n = 1000$. The true survival function S of an $\Gamma_{4,0.5}$ distribution is given by the black curve while the red curve is the point-wise empirical median of the 50 estimates.

Bibliographie

Brenner Miguel, S et Comte, F. et Johannes J. (2021), *Spectral cut-off regularisation for density estimation under multiplicative measurement errors*, Electronic Journal of Statistics

Brenner Miguel, S et Phandoidaen, N. (2021), *Multiplicative deconvolution in survival analysis under dependency*, arXiv preprint arXiv:2107.05267