Stochastic and Physical Monitoring Systems
SPMS 2012
June 25 - June 30, 2012
Zlenice, Czech Republic

Testing of robustness and efficiency of Rényi divergence estimators of probability densities

Jan Kučera

Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Trojanova 13, 12000 Prague, Czech Republic

Email: kucerj28@fjfi.cvut.cz

Abstract. In this contribution we study Rényi pseudo-distance estimators which are based on minimization of information-theoretic divergences between empirical and hypothetical probability distribution. These distances are more robust (than e.g. MLE estimators) against outliers and other measurement errors potentially present in the data sets. Robustness of these estimators is described by influence function. In [1] and [4] authors found explicit formulas for enumeration of Rényi pseudodistances in normal families and for their influence functions. We focus on finding explicit formulas for other families (Weibull, Cauchy, Exponential) and finding influence functions for these estimators. We perform computer simulations for pseudorandom contaminated and uncontaminated data sets, different sample sizes and different Rényi pseudodistance parameters.

Key words: Rényi pseudodistances; $\phi-$divergences; robustness; minimum pseudodistance estimators.

1 Introduction and basic definitions

In this contribution we study Rényi pseudo-distance estimators which are based on minimization of information-theoretic divergences between empirical and hypothetical probability distribution. They are not classical distances, because the symmetry or triangle inequality does not have to hold. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^m\}$ be set of probability measures on measurable space $(\mathcal{X}, \mathcal{A})$. We will apply the estimators in the statistical model with i.i.d. observations $X_1, \ldots, X_n$ governed by distribution P_r. Because we will be interested in robustness, we allow the case $P_r \notin \mathcal{P}$ and therefore we define another set $\mathcal{P}^+ = \mathcal{P} \cup \{P_r\}$. In this paper we will take into consideration distribution P_r of contaminated data modeled as a mixture of distributions $P_r = (1-\varepsilon)P + \varepsilon Q$, where ε is the the contamination level, P is the distribution of the non-contaminated part of the data and Q is the distribution of the contamination.

Definition 1 *We say that mapping $\mathfrak{D} : \mathcal{P} \times \mathcal{P}^+ \to \mathbb{R}$ is pseudo-distance between proba-*

bility measures $P \in \mathcal{P}$ and $Q \in \mathcal{P}^+$ if it holds

$$\mathfrak{D}(P_\theta, Q) \geq 0 \quad \textit{for all } \theta \in \Theta \quad \textit{and} \quad Q \in \mathcal{P}^+ \tag{1}$$

and

$$\mathfrak{D}(P_\theta, P_{\tilde{\theta}}) = 0 \;\Leftrightarrow\; \theta = \tilde{\theta}. \tag{2}$$

This pseudo-distance is decomposable if there exist functionals so that $\mathfrak{D}^0 : \mathcal{P} \to \mathbb{R}$, $\mathfrak{D}^1 : \mathcal{P}^+ \to \mathbb{R}$ and measurable mapping $\rho_\theta : \mathcal{X} \to \mathbb{R}$, $\theta \in \Theta$, so that $\forall\theta \in \Theta$ and $\forall Q \in \mathcal{P}^+$ the expectation $\int \rho_\theta \mathrm{d}Q$ exists and

$$\mathfrak{D}(P_\theta, Q) = \mathfrak{D}^0(P_\theta) + \mathfrak{D}^1(Q) + \int \rho_\theta \mathrm{d}Q. \tag{3}$$

Definition 2 *We say that a functional $T_\mathfrak{D} : \mathcal{Q} \to \Theta$, for $\mathcal{Q} = \mathcal{P}^+ \cup \mathcal{P}_{emp}$ defines minimum pseudo-distance estimator (min $\mathfrak{D}$-estimator) if $\mathfrak{D}(P_\theta, Q)$ is a decomposable pseudo-distance on $\mathcal{P} \times \mathcal{P}^+$ and parameters $T_\mathfrak{D}(Q) \in \Theta$ minimize $\mathfrak{D}^0 + \int \rho_\theta \mathrm{d}Q$, that means*

$$T_\mathfrak{D}(Q) = \arg\min_{\theta\in\Theta} \left[\mathfrak{D}^0(P_\theta) + \int \rho_\theta \mathrm{d}Q\right] \quad \forall Q \in \mathcal{Q}. \tag{4}$$

In particular, for $Q = P_n = \frac{1}{n}\sum_{i-1}^n \delta_{X_i} \in \mathcal{P}_{emp}$

$$\hat{\theta}_{\mathfrak{D},n} = T_\mathfrak{D}(P_n) = \arg\min_{\theta\in\Theta} \left[\mathfrak{D}^0(P_\theta) + \frac{1}{n}\sum_{i-1}^n \rho_\theta(X_i)\right]. \tag{5}$$

Every min $\mathfrak{D}$-estimator is Fisher consistent in the sense that

$$T_\mathfrak{D}(P_{\theta_0}) = \arg\min_{\theta\in\Theta} \mathfrak{D}(P_\theta, P_{\theta_0}) = \theta_0, \quad \forall\theta_0 \in \Theta. \tag{6}$$

Theorem 1 *Let for some $\beta > 0$ it holds that*

$$p^\beta, q^\beta, \ln p \in \mathrm{L}_1(Q), \quad \forall P \in \mathcal{P}, Q \in \mathcal{P}^+.$$

Then for all α, $0 < \alpha \leq \beta$, and for $P \in \mathcal{P}$, $Q \in \mathcal{P}^+$ the expression

$$\mathfrak{R}_\alpha(P, Q) = \frac{1}{1+\alpha}\ln\left(\int p^\alpha \mathrm{d}P\right) + \frac{1}{\alpha(1+\alpha)}\ln\left(\int q^\alpha \mathrm{d}Q\right) - \frac{1}{\alpha}\ln\left(\int p^\alpha \mathrm{d}Q\right) \tag{7}$$

represents the family of pseudo-distances decomposable in the sense

$$\mathfrak{R}_\alpha(P, Q) = \mathfrak{R}^0_\alpha(P) + \mathfrak{R}^1_\alpha(Q) - \frac{1}{\alpha}\ln\left(\int p^\alpha \mathrm{d}Q\right),$$

where

$$\mathfrak{R}^0_\alpha(P) = \frac{1}{1+\alpha}\ln\left(\int p^\alpha \mathrm{d}P\right), \quad \mathfrak{R}^1_\alpha(Q) = \frac{1}{\alpha(1+\alpha)}\ln\left(\int q^\alpha \mathrm{d}Q\right).$$

Moreover for $\alpha \searrow 0$ it holds

$$\mathfrak{R}_0(P, Q) = \lim_{\alpha\searrow 0} \mathfrak{R}_\alpha(P, Q) = \int (\ln q - \ln p)\, \mathrm{d}Q.$$

Rényi pseudo-distance estimator is then

$$T_{\mathfrak{R}_\alpha}(Q) = \begin{cases} \arg\min_\theta \left[\frac{1}{1+\alpha}\ln(\int p_\theta^\alpha \mathrm{d}P_\theta) - \frac{1}{\alpha}\ln(\int p_\theta^\alpha \mathrm{d}Q)\right] & \text{for } 0 < \alpha \leq \beta, \\ \arg\min_\theta \left[-\ln(\int p_\theta \mathrm{d}Q)\right] & \text{for } \alpha = 0. \end{cases} \tag{8}$$

We are interested in estimators, where we replace the hypothetical distribution P_r with the empirical distribution P_n. That means family of minimum Rényi pseudo-distance estimators defined as $\theta_{n,\alpha} = T_{\mathfrak{R}_\alpha}(P_n)$ for $T_{\mathfrak{R}_\alpha}(Q) \in \Theta$ with $Q \in \mathcal{P}^+$ satisfying the condition

$$\theta_{\alpha,n} = \begin{cases} \arg\max_{\theta\in\Theta} C_\alpha(\theta)^{-1} \frac{1}{n}\sum_{i=1}^n p_\theta^\alpha(X_i) & \text{for } 0 < \alpha \leq \beta, \\ \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ln p_\theta(X_i) & \text{for } \alpha = 0. \end{cases} \tag{9}$$

In [1] author derives influence function for Rényi divergence estimators as

$$\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \theta) = -\mathbf{I}_\alpha^{-1}(\theta)\left[p_\theta^\alpha(x)(s_\theta(x) - c_\alpha(x))\right], \tag{10}$$

where

$$s_\theta = \frac{\mathrm{d}}{\mathrm{d}\theta}\ln p_\theta, \qquad \dot{s}_\theta = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^T s_\theta,$$
$$c_\alpha(\theta) = \frac{\int p_\theta^{1+\alpha} s_\theta \mathrm{d}\lambda}{\int p_\theta^{1+\alpha}\mathrm{d}\lambda}, \qquad \dot{c}_\alpha(\theta) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^T c_\alpha(\theta),$$

and

$$\mathbf{I}_\alpha(\theta) = \int \left[\dot{s}_\theta - \dot{c}_\alpha(\theta) - \alpha(s_\theta - c_\alpha(\theta))(c_\alpha^T(\theta) - s_\theta^T)\right] p_\theta^{1+\alpha}\mathrm{d}\lambda. \tag{11}$$

We will use these influence functions to demonstrate robustness of minimal Rényi pseudodistance estimator. There are more important characteristics, but we will be interested in two. First one is gross-error sensitivity characterized as $\gamma^* = \sup_x |\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \theta)|$. We want it to be finite, in other words, we want the influence function to be bounded. The other one is called rejection point and is defined as $\rho^* = \inf\{r > 0 \mid \mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \theta) = 0 \text{ for } |x| > r\}$. It describes values of samples, which will be treated as outliers and rejected.

2 Application to specific families

In [1] authors present formulas for computing and evaluating Rényi pseudodistance estimator in normal family. Our goal was to study these estimators in other families. In the next chapter we present our results together with some examples.

Laplace family

We use min Rényi estimators for estimating parameter $\theta = (\mu, \lambda)$ in the Laplace family, where the probability density is

$$p_\theta = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}} \tag{12}$$

and therefore the estimator according to (9) is

$$\theta_{\alpha,n} = \arg\max_{\theta\in\Theta} \left[(2\lambda)^{-\frac{\alpha}{1+\alpha}} \frac{1}{n} \sum_{i=1}^{n} \exp\left[-\alpha \frac{|x_i - \mu|}{\lambda} \right] \right]. \tag{13}$$

$\alpha\backslash n$	500		
	$m(\mu)$	$s(\mu)$	$eref(\mu)$
	$m(\lambda)$	$s(\lambda)$	$eref(\lambda)$
0.0	−0.005	0.071	1.000
	4.593	3.617	1.000
0.05	−0.000	0.071	1.018
	4.168	3.192	1.284
0.1	0.003	0.066	1.163
	3.719	2.741	1.742
0.2	−0.002	0.068	1.098
	2.888	1.914	3.570
0.3	−0.004	0.068	1.108
	2.215	1.241	8.490
0.5	0.001	0.069	1.056
	1.546	0.568	40.516
1.0	0.003	0.074	0.927
	1.221	0.256	199.663

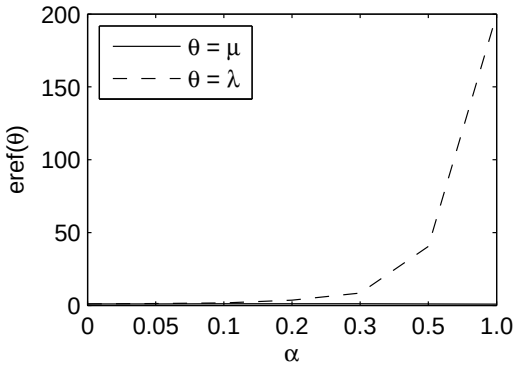


Table 1: Rényi: $p_\theta = \mathrm{L}(0,1)$, data: $(1-\varepsilon)\mathrm{L}(0,1) + \varepsilon\mathrm{L}(0,10)$, $\varepsilon = 0.4$

For creation of Table 1 we generated contaminated data (number of observations $n = 500$) as a mixture of two Laplace distributions $(1-\varepsilon)\mathrm{L}(0,1)+\varepsilon\mathrm{L}(0,10)$ with $\varepsilon = 0.4$. Then we used Rényi estimators and estimated the parameters $\mu,\ \lambda$ and repeated this process $m = 1000$ times. Numbers in the table are mean value and variance of the estimated parameters. In the last column, there is relative efficiency given by

$$eref(p) = \sqrt{\frac{\frac{1}{m}\sum_{k=1}^{m}(\hat{p}_{\mathrm{MLE},k} - p_{\mathrm{real}})^2}{\frac{1}{m}\sum_{k=1}^{m}(\hat{p}_{\alpha,k} - p_{\mathrm{real}})^2}}, \tag{14}$$

where $\hat{p}_{\mathrm{MLE},k}$ is MLE estimator, p_{real} is the real parameter of the contaminated distribution and $\hat{p}_{\alpha,k}$ is minimal Rényi estimator. As we can see, the efficiency raises with increasing α. This depends on the level of contamination, estimators with higher α are more robust against outliers, so with increasing level of contamination with more dispersed observations the relative efficiency will raise, because of the robustness against outliers.

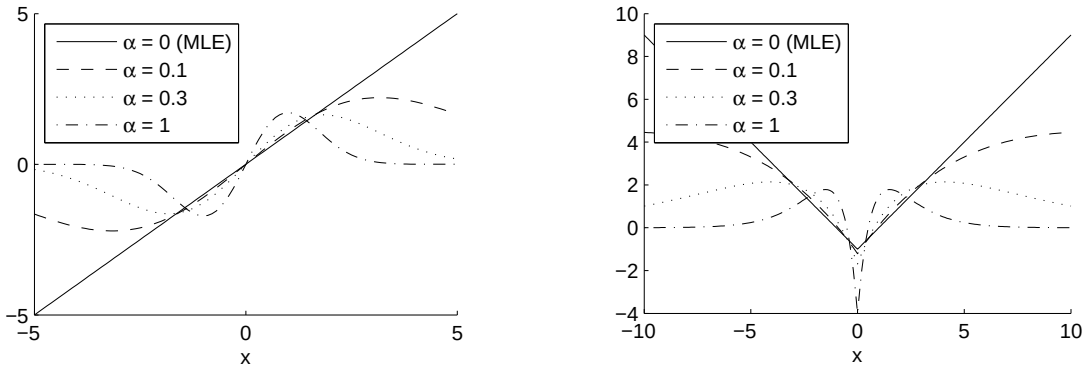
If we compute the influence functions according to (10), where we put $\theta = \mu$ (estimated parameter), $\lambda := 1$ (known parameter) respectively $\theta = \lambda$, $\mu := 0$, we get

$$\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \mu) = (1+\alpha)^{\frac{3}{2}}(x-\mu)e^{-\frac{\alpha}{2}(x-\mu)^2}, \tag{15}$$

respectively

$$\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \lambda) = (1+\alpha)^2\left(-\lambda + (1+\alpha)|x|\right)e^{-\frac{\alpha|x|}{\lambda}}. \tag{16}$$

We can see from (15), (16), that our estimators are robust for $\alpha > 0$ in the sense that



$\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \mu = 0)$, $\lambda = 1$ is known $\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \lambda = 1)$, $\mu = 0$ is known

Figure 1: Influence functions of Rényi estimator for Laplace distribution

their influence functions are bounded. Also with higher α they are robust against outliers, because $\lim_{x\to\pm\infty} \mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \cdot) = 0$ and the convergence is faster with increasing α because of the $e^{-\alpha x}$ which appears in both functions.

Exponential family

If we use min Rényi estimators in the exponential family, due to similarity with the Laplace family, we get very similar results. We have probability density

$$p_\theta = \frac{1}{\lambda}e^{-\frac{x-\mu}{\lambda}} \tag{17}$$

and therefore the estimator is

$$\theta_{\mathfrak{R}_\alpha,n} = \arg\max_{\theta\in\Theta} \lambda^{-\frac{\alpha}{1+\alpha}} \frac{1}{n}\sum_{i=1}^{n} \exp\left[-\alpha\frac{x_i-\mu}{\lambda}\right]. \tag{18}$$

Table 2 was generated in the same manner as Table 1, but in this case we used mixture of two exponential distributions with different parameters. Relative efficiency of estimators of λ are very similar to the estimators in Laplace family, but in this case efficiency of estimators of μ is decreasing with higher α. Although the efficiency is low, we can see that the estimates are really close to the actual value for all cases of α. Here we can clearly see, that there has to be some experiments regarding the choice of α for the specific problem, because in this case it would be better to use $\alpha \sim 0.3$.

$\alpha\backslash n$	500		
	$m(\mu)$	$s(\mu)$	$eref(\mu)$
	$m(\lambda)$	$s(\lambda)$	$eref(\lambda)$
0.0	0.003	0.004	1.000
	4.659	3.683	1.000
0.05	0.003	0.005	0.951
	4.184	3.207	1.319
0.1	0.003	0.004	1.070
	3.738	2.761	1.780
0.2	0.003	0.005	0.932
	2.887	1.911	3.712
0.3	0.003	0.004	1.005
	2.196	1.224	9.057
0.5	0.003	0.005	0.738
	1.545	0.569	41.931
1.0	0.007	0.012	0.145
	1.217	0.255	208.991

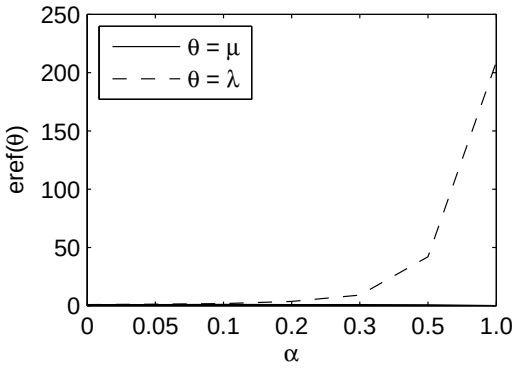
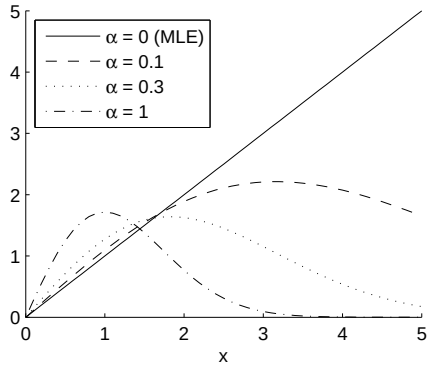


Table 2: Rényi: $p_\theta = \mathrm{E}(0,1)$, data: $(1-\varepsilon)\mathrm{E}(0,1) + \varepsilon\mathrm{E}(0,10)$, $\varepsilon = 0.4$

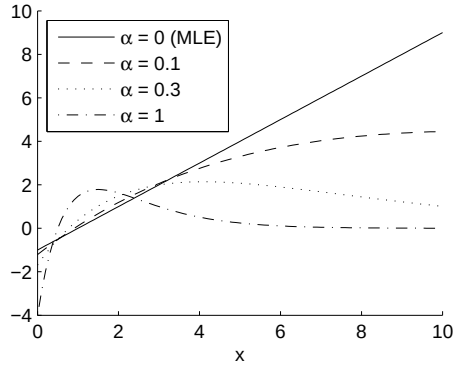
We compute the influence functions according to (10). When we put $\theta = \mu$, $\lambda := 1$, for $\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \mu)$ we get the same result as in (15). For $\theta = \lambda$, $\mu := 0$, we get

$$\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \lambda) = (1+\alpha)^2(-\lambda + (1+\alpha)x)e^{-\frac{\alpha x}{\lambda}} \tag{19}$$

which as we can see differs from (16) only in the use of x instead of $|x|$.



$\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \mu = 0)$, $\lambda = 1$ is known



$\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \lambda = 1)$, $\mu = 0$ is known

Figure 2: Influence functions of Rényi estimator for exponential distribution

From (19) and Figure 2 we can see, that influence function has the same properties with respect to robustness of the estimator as the influence function in Laplace family.

Cauchy family

For Cauchy family with parameters $\theta = (\mu, \sigma)$ and probability density

$$p_\theta = \frac{1}{\pi\sigma}\left(1 + \left(\frac{x-\mu}{\sigma}\right)^2\right)^{-1} \tag{20}$$

minimal Rényi pseudodistance estimator is given by

$$\theta_{\alpha,n} = \arg\max_{\theta\in\Theta}\left[\sigma^{-\frac{\alpha}{1+\alpha}}\frac{1}{n}\sum_{i=1}^{n}\left(1 + \left(\frac{x_i-\mu}{\sigma}\right)^2\right)^{-\alpha}\right]. \tag{21}$$

$\alpha\backslash n$	500		
	$m(\mu)$	$s(\mu)$	$eref(\mu)$
	$m(\sigma)$	$s(\sigma)$	$eref(\sigma)$
0.0	0.012	0.112	1.000
	2.359	1.389	1.000
0.05	−0.001	0.099	1.286
	2.262	1.282	1.174
0.1	−0.001	0.096	1.361
	2.134	1.156	1.445
0.2	0.001	0.096	1.374
	1.904	0.926	2.251
0.3	−0.004	0.093	1.467
	1.738	0.760	3.346
0.5	−0.002	0.088	1.639
	1.491	0.519	7.168
1.0	−0.001	0.096	1.377
	1.225	0.272	26.129

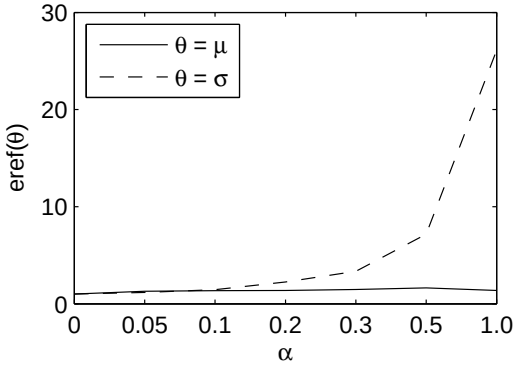


Table 3: Rényi: $p_\theta = \mathrm{C}(0,1)$, data: $(1-\varepsilon)\mathrm{C}(0,1) + \varepsilon\mathrm{C}(0,10)$, $\varepsilon = 0.4$

In Table 3 we can see, that relative efficiency of Rényi estimators increases with higher α in estimates of both parameters.

In the Cauchy family we were able to compute the influence function only for the estimated parameter $\theta = \mu$ and known parameter $\sigma := 1$. In that case we get

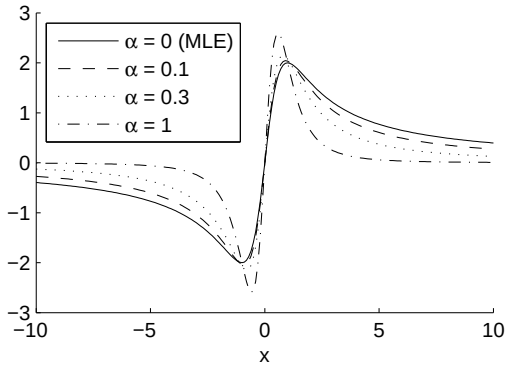
$$\mathrm{IF}(x; T_{\mathfrak{R}_\alpha}, \mu) = \sqrt{\pi}\frac{\Gamma(3+\alpha)}{\Gamma\left(\frac{3}{2}+\alpha\right)}\left(\frac{1}{1+(x-\mu)^2}\right)^{1+\alpha}(x-\mu). \tag{22}$$

We can see from (22) that for $\alpha > 0$ the influence function is bounded and converges to 0.

Weibull family

We were also able to obtain some results in the Weibull family with parameters $\theta = (\mu, \lambda, k)$ and probability density

$$p_\theta = \frac{k}{\lambda}\left(\frac{x-\mu}{\lambda}\right)^{k-1}\exp\left[-\left(\frac{x-\mu}{\lambda}\right)^k\right]. \tag{23}$$



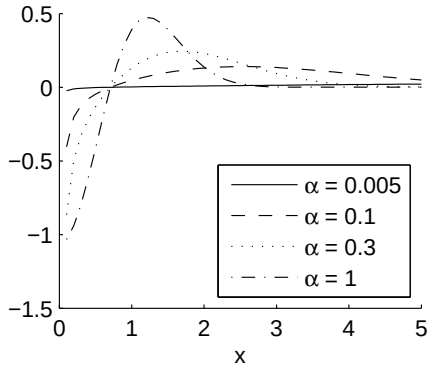
IF$(x; T_{\mathfrak{R}_\alpha}, \mu = 0)$, $\sigma = 1$ is known

Figure 3: Influence function of Rényi estimator for Cauchy distribution

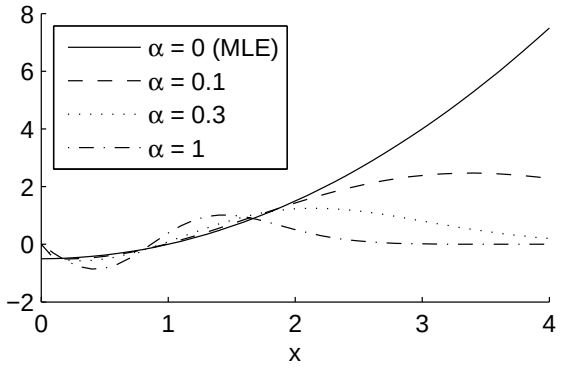
Rényi estimator is given by

$$\begin{aligned}\theta_{\alpha,n} &= \arg\max_{\theta\in\Theta}\left(\frac{k}{\lambda}\right)^{\frac{\alpha}{1+\alpha}}(1+\alpha)^{\frac{\alpha}{1+\alpha}\frac{1+\alpha+k}{k}}\Gamma\left(\frac{1+\alpha+k}{k}\right)^{-\frac{\alpha}{1+\alpha}} \\ &\quad \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\lambda}\right)^{\alpha(k-1)}\exp\left[-\alpha\left(\frac{x_i-\mu}{\lambda}\right)^{k}\right]. \qquad (24)\end{aligned}$$

We will show influence functions only in figures due to complexity and length of the formulas. The general shape of the functions is still present. We can see, that the estimator is robust against outliers and that the influence function is bounded for $\alpha \neq 0$.



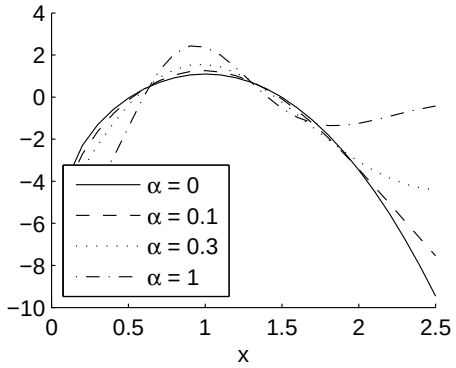
IF$(x; T_{\mathfrak{R}_\alpha}, \mu = 0)$, $\lambda = 1$, $k = 2$ are known



IF$(x; T_{\mathfrak{R}_\alpha}, \lambda = 1)$, $\mu = 0$, $k = 2$ are known

Figure 4: Influence function of Rényi estimator for Weibull distribution

Table 4 was created in different manner than previous ones. We have MLE for parameter λ, but there is none for μ. Because of this, we related the results of estimation of μ to the estimator with $\alpha = 0.001$. We can see from the table, that the estimates of μ don't differ significantly with different α. On the other hand estimates of λ are much more efficient than MLE. Although Weibull distribution has 3 parameters, we estimated



IF$(x; T_{\mathfrak{R}_\alpha}, k = 2)$, $\mu = 0$, $\lambda = 1$ are known

Figure 5: Influence function of Rényi estimator for Weibull distribution

only 2, because program we used for computing was optimized for minimization of 1- and 2-dimensional distances.

$\alpha \backslash n$	500		
	$m(\lambda)$	$s(\lambda)$	$eref(\lambda)$
	$m(\mu)$	$s(\mu)$	$eref(\mu)$
0.0	6.462	5.481	1.000
0.001	−0.006	0.289	1.000
0.05	1.015	0.289	358.924
	0.005	0.289	0.999
0.1	1.016	0.284	371.784
	0.007	0.282	1.044
0.2	1.000	0.293	351.092
	−0.005	0.288	1.004
0.3	1.010	0.283	376.025
	0.006	0.293	0.968
0.5	0.998	0.298	337.894
	−0.002	0.289	0.998
1.0	1.005	0.292	352.589
	0.001	0.287	1.012

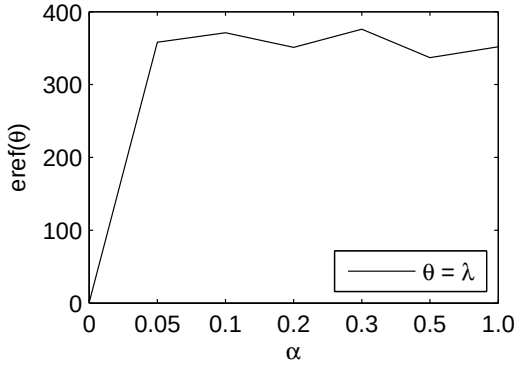


Table 4: Rényi: $p_\theta = \mathrm{W}(0, 1, 2)$, data: $(1 - \varepsilon)\mathrm{W}(0, 1, 2) + \varepsilon\mathrm{W}(0, 10, 2)$, $\varepsilon = 0.4$

3 Conclusion

We have studied robust properties of the minimum Rényi pseudodistance estimators. We have focused on finding formulas for computing these estimators and formulas of their influence functions. We also presented estimation in distributions contaminated with more dispersed observations to present the robustness of Rényi estimators against outliers.

References

[1] Michel Broniatowski, Igor Vajda. *Several Applications of Divergence Criteria in Continuous Families*. Research report No 2257 September 2009, UTIA AV CR, Prague, 2009.

[2] Igor Vajda. *Information - Theoretic Methods in Statistics*. Research report No 1834 October 1995, UTIA AV CR, Prague, 1995.

[3] Michel Broniatowski, Aida Toma, Igor Vajda. *Decomposable pseudo-distances and Applications in Statistical Estimation*. arXiv:1104.1541v1, 2011.

[4] Radim Demut. *Robust properties of minimum divergence density estimators*. Diploma Thesis, CTU, Prague 2010.

[5] By Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, M. C. Jones. *Robust and efficient estimation by minimising a density power divergence*. In Biometrika, 85, 549-559, 1998.

[6] Friedrich Liese, Igor Vajda. *On Divergences and Informations in Statistics and Information Theory*. IEEE Transactions on Information Theory, Vol. 52, No. 10,4394-4412, 2006.