Numerical enhancements for robust Rényi decomposable minimum distance estimators

Vaclav Kus, Jan Kucera, Domingo Morales
Department of Mathematics, FNSPE, Czech Technical University in Prague, Trojanova 13, 12000 Praha 2, Czech Republic
Departamento de Estadística, Matemáticas e Informática, Universidad Miguel Hernández de Elche, Avda. de la Universidad s/n, 03202 Elche (Alicante), España
E-mail: vaclav.kus@fjfi.cvut.cz (corresponding author)

Abstract. Different numerical aspects of Rényi pseudo-distance estimators are studied. These estimators are based on the minimization of information-theoretic divergences between empirical and hypothetical probability distributions. They are not classical distances, because the symmetry or triangle inequality does not hold. Robust properties of the minimum Rényi pseudodistance estimators are required by various applications in mathematical modeling, physics, or material science. Therefore we model the distribution of contaminated data as a mixture of the true distributions $P$ and error distribution $Q$ under different contamination level $\varepsilon$. We focus on the estimators for relatively small data samples or very sparse and scattered data with high variance, which appears mostly in high energy physics (signal and sparse background). In this case, the strict minimization leads to delta functions and it is impossible to obtain satisfactory numerical results. A way of adjusting the Rényi minimum distance estimators to these conditions is proposed. This so called ‘blurring’ is created as a convolution of Rényi distance with averaging Gaussian mask. Simultaneously, the effect of the input parameter alpha to the robustness is presented based on Monte-Carlo simulations for Gaussian model. Thus the Rényi distance is ready to be used in divergence decision trees for the signal versus background separations, e.g. in high energy physics NOvA or DUNE experiments at Fermilab.

1. Introduction

Many different fields of physical applications, such as material defectoscopy, elasticity, machine learning, or, especially high energy physics (HEP) experiments running on particle accelerators, produce a vast amount of raw data containing not only the crucial signal but also a large number of observations originating from another physical background processes. Prior to analyzing the data of the experiment, we need to separate only the desired useful observations from other physical backgrounds by using some appropriate statistical procedures based on estimators or hypotheses testing.

In this paper we deal with quite robust statistical parametric estimators and prospective tests based on minimum distance principle while using the decomposable Rényi distance recently defined in literature [1] and we study further numerical aspects of Rényi pseudo-distance estimators. These estimators are based on the minimization of information-theoretic distances (divergences, disparities) between empirical and hypothetical probability distributions. They are not classical distances because the symmetry or triangle inequality does not have to hold, however their robust properties can be excellent [2, 3].
2. Decomposable pseudo-distances and parametric estimators

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^m\}$ be a set of parametrized probability measures on a measurable space $(X, \mathcal{A})$. We consider the i.i.d. model of observations $X_1, \ldots, X_n$ governed by a distribution $Q \in \mathcal{P}^+$ with $\mathcal{P} \subset \mathcal{P}^+$. Here, $\mathcal{P}^+$ may represent a set of contaminated distributions modeled by the mixtures $Q = (1 - \varepsilon)P_\theta + \varepsilon P_{\text{err}}$ under a contamination level $\varepsilon > 0$ and contaminating distribution $P_{\text{err}}$ (e.g. physical background).

We say that the mapping $\mathcal{D} : \mathcal{P} \times \mathcal{P}^+ \rightarrow \mathbb{R}$ forms a pseudo-distance between probability measures $P_\theta \in \mathcal{P}$ and $Q \in \mathcal{P}^+$ if $\mathcal{D}(P_\theta, Q)$ is nonnegative for all $\theta \in \Theta$, all $Q \in \mathcal{P}^+$, and it holds on $\Theta$ that $\mathcal{D}(P_\theta, P_{\theta_0}) = 0 \iff \theta_1 = \theta_2,$

without requiring the symmetry or triangle inequality to be fulfilled. This pseudo-distance is called decomposable if there exist functionals $\mathcal{D}^0 : \mathcal{P} \rightarrow \mathbb{R}, \mathcal{D}^1 : \mathcal{P}^+ \rightarrow \mathbb{R}$, and a measurable mapping $\rho_0 : X \rightarrow \mathbb{R}, \theta \in \Theta$, such that for all $\theta \in \Theta$ and all $Q \in \mathcal{P}^+$ the expectation $\int \rho_0 \, dQ$ exists and the following decomposition is given by

$$\mathcal{D}(P_\theta, Q) = \mathcal{D}^0(P_\theta) + \mathcal{D}^1(Q) + \int \rho_0 \, dQ.$$  

Now, let us define $Q = \mathcal{P}^+ \cup \mathcal{P}_{\text{emp}}$, where $\mathcal{P}_{\text{emp}}$ denotes the set of all empirical distributions of the form $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$. Then we declare that a functional $T_\mathcal{D} : Q \rightarrow \Theta$ defines the minimum pseudo-distance estimator (in short min $\mathcal{D}$-estimator) if $\mathcal{D}(P_\theta, Q)$ is a decomposable pseudo-distance on $\mathcal{P} \times \mathcal{P}^+$ and the parameters $T_\mathcal{D}(Q) \in \Theta$ minimize $\mathcal{D}^0 + \int \rho_0 \, dQ$, that means

$$T_\mathcal{D}(Q) = \arg \min_{\theta \in \Theta} \left[ \mathcal{D}^0(P_\theta) + \int \rho_0 \, dQ \right], \quad \forall Q \in \mathcal{Q}.$$  

In particular, for the empirical measure $Q = P_n \in \mathcal{P}_{\text{emp}}$ we obtain

$$\hat{\theta}_{T_\mathcal{D}, n} = T_\mathcal{D}(P_n) = \arg \min_{\theta \in \Theta} \left[ \mathcal{D}^0(P_\theta) + \frac{1}{n} \sum_{i=1}^n \rho_0(X_i) \right].$$  

Note that every min $\mathcal{D}$-estimator is Fisher consistent in the sense that

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

3. Minimum Rényi distance estimator and its robustness

Assume that for some $\beta > 0$ it holds that $p^\beta, q^\beta, \ln p \in L_1(Q)$, for all $P \in \mathcal{P}$, $Q \in \mathcal{P}^+$, where $p$ and $q$ are the density functions corresponding to $P$ and $Q$. Then for all $\alpha$, $0 < \alpha \leq \beta$, with $P \in \mathcal{P}$, $Q \in \mathcal{P}^+$, the expression

$$\mathcal{R}_\alpha(P, Q) = \frac{1}{1 + \alpha} \ln \left( \int p^\alpha \, dP \right) + \frac{1}{\alpha(1 + \alpha)} \ln \left( \int q^\alpha \, dQ \right) - \frac{1}{\alpha} \ln \left( \int p^\alpha \, dQ \right)$$  

represents the family of Rényi pseudo-distances, decomposable in the sense of

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

Furthermore, for $\alpha \searrow 0$ we come to $\mathcal{R}_0(P, Q) = \lim_{\alpha \searrow 0} \mathcal{R}_\alpha(P, Q) = \int (\ln q - \ln p) \, dQ$, which completes the definition of Rényi distances for the purposes of this paper. Consequently, the parametric minimum Rényi pseudo-distance estimator with respect to $Q$ is then determined by

$$T_{\mathcal{R}_\alpha}(Q) = \begin{cases} \arg \min_{\theta} \left[ \frac{1}{1 + \alpha} \ln \left( \int p_\theta^\alpha \, dP_{\theta} \right) - \frac{1}{\alpha} \ln \left( \int p_\theta^\alpha \, dQ \right) \right] & \text{for } 0 < \alpha \leq \beta, \\ \arg \min_{\theta} \left[ - \ln \left( \int \rho_0 \, dQ \right) \right] & \text{for } \alpha = 0. \end{cases}$$
We are interested in estimators, where we replace the hypothetical distribution \( Q \) with the empirical distribution \( P_n \) based on data measured. It means that the family of resulting minimum Rényi pseudo-distance estimators, defined as \( \hat{\theta}_{\alpha,n} = T_{\alpha,n}(P_n) \in \Theta \), satisfies the condition

\[
\hat{\theta}_{\alpha,n} = \begin{cases} 
\arg \max_{\theta \in \Theta} C_{\alpha}(\theta) \left( -\frac{1}{n} \sum_{i=1}^{n} p_\theta^\alpha(X_i) \right) & \text{for } 0 < \alpha \leq \beta, \\
\arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ln p_\theta(X_i) & \text{for } \alpha = 0,
\end{cases}
\]  

(1)

with the factor \( C_{\alpha}(\theta) = \int p_\theta^{1+\alpha} d\lambda \) under the Lebesque measure \( \lambda \) on \( \mathbb{R} \).

We use the common Influence Function (IF) to measure robustness of our estimator. The IF summarizes the impact of a single date on the estimator under consideration. There are more important characteristics derivable from the influence function, but we are mainly interested in two of them. First one is the gross-error sensitivity characterized by \( \rho^* = \sup_x |IF(x; T_{\alpha,n}, \theta)| \). We require \( \rho^* \) to be finite, in other words, we want the influence function to be bounded. The second characteristic is called the rejection point

\[
\rho^* = \inf \{ r > 0 \mid IF(x; T_{\alpha,n}, \theta) = 0 \text{ for } |x| > r \}
\]

and it describes the samples treated as outliers and finally rejected. As an example, we show the IF of Rényi estimator (1), derived in [1] for parameter sigma of normal distribution \( N(\mu, \sigma) \),

\[
IF(x; T_{\alpha,n}, \sigma) = \frac{(1 + \alpha)^{5/2} \sigma}{2} \left[ \left( \frac{x}{\sigma} \right)^2 - \frac{1}{1 + \alpha} \right] \exp\left( \frac{\alpha x^2}{2\sigma^2} \right),
\]

(2)

depicted in Figure 1. We can deduce from this Figure that Rényi estimators are robust for \( \alpha > 0 \) in the sense that their influence functions are bounded. Also with higher \( \alpha \) they are steadily more robust against outliers since \( \lim_{x \to \pm \infty} IF(x; T_{\alpha,n}, \sigma) = 0 \), while the vanishing convergence is faster with increasing \( \alpha \), due to the term \( e^{-\alpha x^2} \).

4. Numerical enhancements for Rényi estimators

The numerical computations of similar variants of minimum distance estimators are not easy at all [4]. Even though the min Rényi estimator is known to be quite robust in practice, there is an
inconvenience concerning the choice of $\alpha$ for small data samples or for very sparse or scattered
data (data with high variance of the physical background). This is illustrated in Figure 2 for
$\alpha = 0.5$, where we used i.i.d. random sample of $n = 10$ values $X = (x_1, \ldots, x_{10}) \sim N(0, 2)$ for
the Rényi decomposable distance between the empirical distribution $P_{10}$ of $X$ and the normal
distribution with variable parameters $(\mu, \sigma)$. We can see that there is a large shallow minimum

![Figure 2. Rényi distance behavior.](image1)

around the point $(0.3, 2.3)$. However, a closer examination shows that all of the eight bright
points on the y axis ($\sigma = 0.001$) have steep minima with even lower values of the Rényi pseudo-
distance function. These areas specify distributions corresponding to the Dirac $\delta$-functions
$\delta_{a_i}(x)$, where $a_i = x_i$ for $i \in \{1, \ldots, 10\}$. Due to these extreme values of Rényi distance
function, the minimum Rényi estimator would prefer one of these singular estimates instead
of the large shallow minimum around the point $(0.3, 2.3)$, which is clearly much closer to the
true parameters of overall data sample distribution. Thus the estimator is so much robust and
the data are so sparse that the min Rényi estimator always takes the single datum as the only
representative of the estimated distribution and all the other data are treated as outliers. We
observed that quite similar trouble also emerges for other small sample sizes $n$.

To overcome this problematic behavior, we devised a method inspired by image processing.
The idea is to 'blur' the resulting image so that there wouldn’t be sharp and deep minima, which,
from the viewpoint of image processing, represent edges. This ‘blurring’ is created as convolution
of the Rényi distance with an averaging Gaussian mask. If we shortly denote $\mathcal{R}_{\alpha}(\mu, \sigma)$ the Rényi
distance between the empirical distribution and Gaussian $N(\mu, \sigma)$ distribution, then the Rényi
distance after averaging (blurring) is

$$\mathcal{R}_{\alpha}(\mu_i, \sigma_j) = \frac{1}{r^2} \sum_{k=i-r}^{i+r} \sum_{l=j-r}^{j+r} \mathcal{R}_{\alpha}(\mu_k, \sigma_l), \quad (3)$$

where $r$ is the radius of the averaging mask used. The Rényi distance after averaging is displayed
in Figure 3. We can see, that the minima corresponding to the Dirac $\delta$-functions are flattened
after blurring and their values are higher than the minimum at the approximate point $(0.3, 2.3)$.
The blurring mechanism not only flattens the image, but it also slightly moves the minima. To
overcome this setback and simultaneously refine the resulting Rényi estimate, we apply two-step
algorithm:

(i) First minimize the averaged distance $\mathcal{R}_{\alpha}(\mu, \sigma)$,
Then minimize the original Rényi distance $\mathcal{R}_\alpha(\mu, \sigma)$ in the neighbourhood of the local minima found by the step (i).

It means that we use the blurred (averaged) Rényi distance $\mathcal{R}_\alpha$ to find the overall approximate area of local minimum that describes the whole bunch of data and then we find the exact final estimator from the original Rényi distance $\mathcal{R}_\alpha$. Table 1 shows the differences between the three used algorithms: 1) direct minimization of the original Rényi distance, 2) direct minimization of the averaged (blurred) Rényi distance, 3) the two-step minimization described above.

Table 1. Mean minimization results from 200 repetitions of the Rényi estimators for data $(x_1, \ldots, x_{10}) \sim N(0, 2)$.

| Minimization of                 | mean $\hat{\mu}$ | mean $\hat{\sigma}$ |
|-------------------------------|-------------------|----------------------|
| Rényi distance                | 0.1742            | 0.3448               |
| Averaged Rényi distance       | 0.0058            | 2.8532               |
| Two-step minimization         | 0.0346            | 2.3832               |

The first row in Table 1 corresponds to the Rényi estimator obtained directly from (1), we can observe that the mean of the estimated parameter $\sigma$ is unacceptably small since the generated data possesses much higher value $\sigma = 2$. Conversely, from the second row of Table 1, the usage of only averaging (blurring) algorithm caused the estimator prefers parametric estimator with too high value of $\sigma$. Finally, the mean result of the proposed two-step numeric algorithm is much closer and considerably refines the previous ones.

5. Conclusion
We have studied robust properties of the minimum Rényi pseudo-distance estimators while we have focused on the estimators for small data samples or, alternatively, for very sparse and scattered data (data with high variance of potential physical background). We proposed a way of adjusting the Rényi minimum distance estimator to these conditions. Thus we have succeeded in preparation of this Rényi distance to be used for example by the divergence decision trees SDDT ([5]) developed for the signal versus background separations in high energy physics decay channels or for the robust image signal reconstructions in neutrino experiments NOvA and DUNE at FNAL.

Acknowledgments
This work was supported mainly by the grant GA16-09848S (GACR), partly by LTT18001 (MYES) and SGS18/188/OHK4/3T/14 (CTU).

References
[1] Broniatowski M, Toma A and Vajda I 2012 Decomposable pseudodistances and applications in statistical estimation J. Statist. Plann. Inference 142(9), 2574–2585
[2] Basu A, Harris IR, Hjort NL and Jones MC 1998 Robust and efficient estimation by minimising a density power divergence Biometrika 85 549–559
[3] Frydlova I, Vajda I and Kus V 2012 Modified power divergence estimators in normal model- simulation and comparative study Kybernetika 48(4) 795–808
[4] Hrabakova J, Kus V 2017 Notes on Consistency of Some Minimum Distance Estimators with Simulation Results Metrika 80 243–257
[5] Bour P, Kus V and Franc J 2016 Statistical classification techniques in high energy physics (SDDT algorithm) J. Phys.: Conf. Series 738 012034