Scientific Research Publishing

American Journal of Computational Mathematics, 2019, 9, 302-316
https://www.scirp.org/journal/ajcm
ISSN Online: 2161-1211
ISSN Print: 2161-1203

Robust Estimation of the Normal-Distribution Parameters by Use of Structural Partitioning-Perobls D Method

Gligorije Perović

Department of Geodesy and Geoinformatics, Faculty of Civil Engineering, University of Belgrade, Belgrade, Serbia
Email: perg@grf.bg.ac.rs

How to cite this paper: Perović, G. (2019) Robust Estimation of the Normal-Distribution Parameters by Use of Structural Partitioning-Perobls D Method. *American Journal of Computational Mathematics*, **9**, 302-316.
https://doi.org/10.4236/ajcm.2019.94022

Received: September 3, 2019
Accepted: December 14, 2019
Published: December 17, 2019

Abstract

Quite many authors have dealt with the estimation of the parameters of normal distribution on the basis of non-homogeneous sets: Hald A. 1949 [1], Arango-Castillo L. and Takahara G. 2018 [2]. All the robust methods are based on the assumption that the results affected by gross errors can be found to the left and/or to the right of censoring, or truncated, points. However, as a rule, the (intrinsic) distribution of observations is complex (mixed) consisting of two or more distributions. Then the existing methods, such as ML, Huber's, etc., yield enlarged estimates for the normal-distribution variance. By studying better estimates the present author has invented new method, called PEROBLS D, based on the Tukeyan mixed-distribution model in which both the contamination rate (percentage) and the parameters of both distributions, forming the mixed one, are estimated, and for the parameters of the basic normal distribution better estimates are obtained than by the existing methods.

Keywords

Non-Homogeneous Sets of Observations, Tukeyan Mixed-Distributions, Robust Perobls D Method

1. Introduction and History of Robust Estimation

The history of this problem is older than 300 years. For example, *Galileo* as long ago as in 1632 used the *least absolute sum* in order to reduce the effect of observational errors to the estimate of the measured quantity [3], whereas *Rudjer Boscovich*, *is* the first who, as early as in 1757, rejected clearly *outlying observations* [4], also done by *Daniel Bernouli* 1777 [4]. The *trimmed mean* has been

used since long ago, see "*Anonymous*" 1821 [4]. The *first formal rules for rejecting of observations* were proposed by *Peirce* 1852 and *Chauvenet* 1863, and somewhat later appear the papers by *Stone* 1868, *Wright* 1884, *Irvin* 1925, *Student* 1927, *Thompson* 1935, and by many others [4].

The mixed distribution models have been also considered since long ago: *Glaisher* 1872/1873, *Stone* 1873, *Edgeworth* 1883, *Newcomb* 1886, *Jeffreys* 1932/1939 [4]. *Tukey* in 1960 [5] defined a *mixed model* as a mixture of two normal distributions of a *basic* $\Phi\left[(x-\theta)/\sigma\right]$ and of a *contaminating* $\Phi\left[(x-\theta)/3\sigma\right]$ distribution:

$$F(x)=(1-\varepsilon)\Phi\left[(x-\theta)/\sigma\right]+\varepsilon\Phi\left[(x-\theta)/3\sigma\right],$$

where $\Phi(x)=\frac{1}{2\sqrt{\pi}}\int_{-\infty}^{x}\exp\left(-t^2/2\right)\mathrm{d}t$ is the function of the standard normal distribution. Here the elements of the contaminating set appear with probability ε, (Tukey assumes ε as a small number, about 5%), and behave as *gross errors*. In this way the real distributions are represented through a normal-distribution model with weighted tails.

The term *robustness* began to be used since 1953 (introduced by *G. E. P. Box*) in order to discriminate the class of statistical procedures with little sensitivity to minor deviations from the starting assumptions. Some authors use the term *stability,* but it is less used than the term robustness. In the Foreword of his book *ROBUST STATISTICS*, *Huber* in 1981 [6] emphasizes that among the leading scientists in the late XIX and the early XX centuries there were a few statisticians —practitioners (and mentions: astronomer *Newcomb*, astrophysicist *Eddington* and geophysicist *Jeffreys*) who expressed in their studies a perfectly clear understanding of the robustness idea. They were aware of the perils caused by long tails of the functions of error distributions, so they proposed models of distribution of gross errors and derived robust variants of standard estimates. Russian geodesists, for example, in their adjustments of the first order triangulation networks allowed lower weights (about half of the original ones) to the observations of directions which do not deviate much.

The initial fundaments to the theory of robust estimation were laid by Swiss mathematician *P. J. Huber* 1964 [1] and American statistician *W. J. Tukey* 1960 [5]. Huber's article "*Robust Estimation of Location Parameter*" was the first fundament of the theory of robust estimation, which introduced an elastic class of estimates, called *M-estimates*, which have become a very useful instrument, having established which properties they have (for instance consistency and asymptotic normality). He introduced the *model of gross errors*, replacing the strict parametric model $F(x-\theta)$, with its known distribution *F*, by the *mixed model*:

$$H(x-\theta)=(1-\varepsilon)F(x-\theta)+\varepsilon G(x-\theta),$$

while a part of data ε $(0\le\varepsilon<1)$ may contain gross errors which have an arbitrary (unknown) distribution $G(x-\theta)$.

The causes of deviating from the parametric models are various and four main types of deviating from the strict parametric models can be distinguished:

1) Appearance of gross errors;
2) Rounding and grouping;
3) Using an approximate functional mode; and
4) Using an approximate stochastic model.

The fundamentals of robust methods were developed in the last century. Today there are numerous applications of robust methods, and concurrently better and more detailed solutions are sought: [2] [7]-[14].

Due to good properties of the robust methods—that it is possible to eliminate or decrease the influences of gross errors and outliers on the estimates of distribution parameters, in practice they are used more and more. Therefore, the same time, their development results. So the current development and application of the robust methods may be classified into the following groups:

- *Improvement of existing methods, such as* "A new Perspective on Robust M. Estimation" [13];
- *Solving of delicate* (*specific*) *tasks*:—for robust hybrid state estimation with unknown measurement noise statistics [14]—for optimal allocation of shares in a financial portfolio [8]—for robust estimation of 3D human poses from a single image [12]—for cubature Kalman filter for dynamic state estimation of synchronous machines under unknown measurement noise statistics [9];
- *Applications in various conditions*:—for robust estimation of the sample mean variance for Gaussian processes with long-range dependence [2]—for robust estimation of 3D human poses from a single image [12]—for robust hybrid state estimation with unknown measurement noise [14]—for estimation of mean and variance using environmental data sets with below detection limit observations [10];
- *Applications in diverse fields*:—for estimation of the sample mean variance for Gaussian processes with long-range dependence [2]—for Gaussian sum filtering with unknown noise statistics: Application to target tracking [11]—robust cubature Kalman filter for dynamic state estimation of synchronous machines under unknown measurement noise statistics [9]—for estimation of mean and variance in Fisheries [7]—for optimal allocation of shares in a financial portfolio [8]—for estimation of mean and variance using environmental data sets with below detection limit observations [10].

The proposed PEROBLS D method is aimed at eliminating the influences of gross errors and outliers on the estimates of distribution parameters, when only one contaminating distribution is present, *i.e.* in the case of Tukey's mixed distribution.

The key difference between this paper and existing studies is that the PEROBLS D method in the estimating procedure uses no distribution censoring, unlike the existing methods, but instead a structural decomposition into two distributions is used—basic and contaminating ones which have the same mean

value and then the parameters of both distributions are estimated.

Consequently, in the case that both distributions (basic and contaminating) are normal, the PEROBLS D method has the following properties:

1) Unbiased (exact) parameter estimates for the basic distribution, as the most important property;

2) Unbiased (exact) parameter estimates for the contaminating distribution;

3) Percentage estimates for fractions of basic and contaminating distributions in the mixed one.

The correctness of the method has been verified on exact (expected) values of some quantities from the mixed Tukeyan distribution, as well as on an example of simulated data for 200 measurements of one quantity.

Besides, the estimates of the mean and variance for the basic distribution have been compared with the same ones obtained by ML method, and the estimate basic distribution standard has been also compared with Tukey mad standard estimate.

As has been said, the PEROBLS D estimates are unbiased, whereas the estimates of the basic distriburion standard in both cases, according to ML and Tukey mad, are increased.

The structure of the further presentation is the following. At first definitions and notations are given, then basis of PEROBLS D method and the way of solving the formulated problem. Afterwards the existing robust estimation methods —ML and Huber's mad—are presented. Further on the PEROBLS D method is verified on examples and the solutions are compared with existing ones. Finally, there are conclusions and references.

2. Definitions and Notations

The density function for a standard normal variable Z is given as

$$f(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-z^2/2\right),\quad -\infty<z<\infty.$$

Let $F(z)$ be the notation of the distribution function for Z. The quintiles of Z will be denoted as $z_{1-\alpha}$, where α is the significance level, $z_{\alpha}=-z_{1-\alpha}$ and

$$p\left(Z\le z_{1-\alpha}\right)=F\left(z_{1-\alpha}\right)=1-\alpha.$$

A normally distributed *r. v.* X with expectation μ and variance σ^2 has the density and distribution functions, respectively:

$f(z)=f(z)/\sigma$ and $F(x)=\int_{-\infty}^{x}f(x)\mathrm{d}x=\int_{-\infty}^{x}f(z)\mathrm{d}z=F(z)$, with

$$z=(x-\mu)/\sigma. \tag{1}$$

Let as consider a random variable (natural sequence of measurements) [15]: $X_1,X_2,\cdots,X_n$, *from a normal population* and use $N\left(\mu,\sigma^2\right)$, with mean μ and standard deviation σ, (*i.e.* variance σ^2) where one assumes that the *observations* $X_1,X_2,\cdots,X_n$ are mutually *independent*. Arranging them in the *ascending order of magnitude* one obtains *order statistics* $X_{(i)},i=1,2,\cdots,n$

(Figure 1), where the points A and B defined in the following way:

$$A = X_{(n_X+1)} - 0.5d \quad \text{and} \quad B = X_{(n-n_Z)} + 0.5d\,, \tag{2}$$

where d is the width of the rounding interval for observations X.

With z from Equation (1), it will be analogously:

$$Z_{(1)} \le Z_{(2)} \le \cdots \le Z_{(n_X)} \underbrace{\le}_{z_A} Z_{(n_X+1)} \le \cdots \le Z_{(n-n_Z)} \underbrace{\le}_{z_B} Z_{(n-n_Z+1)} \le \cdots \le Z_{(n)}\,,$$

$$Z_{(n_X)} < \left(z_A = \frac{A-\mu}{\sigma}\right) \le Z_{(n_X+1)};\quad Z_{(n-n_Z)} \le \left(z_B = \frac{B-\mu}{\sigma}\right) < Z_{(n-n_Z+1)}\,,$$

where $z_{(i)} = \left(x_{(i)} - \mu\right)\Big/\sigma$.

3. Basis of PEROBLS D Method[1]

The idea of PEROBLS D method has been presented in the Least Squares book [16].

Instead of assuming the presence of gross errors in the observations within X and Z regions, used in the previous methods, in this method *the observation distribution is defined by means of Tukeyan mixed distribution* (Figure 2):

$$F(x) = (1-\varepsilon)F_1(x) + \varepsilon F_2(x)\,, \tag{3}$$

where $F_1(x)$ is *the basic*, $F_2(x)$—*the contaminating* one, whereas $0 < \varepsilon < 0.5$, noting that ε cannot exceed 0.5, because, in this case $F_2(x)$ must be taken as the basic distribution. (In geodetic applications there is mostly $0 < \varepsilon < 0.3$).

In this method the points A and B are partition points only, i. e. they are neither truncation points nor censoring ones. They are chosen so that in the domains X and Z the contaminating distribution prevails—which is one of the prerequisites to find a good (satisfactory) solution of the problem (task).

Note 1. *In geodetic measurements distributions close to the Tukeyan ones are frequent.* ▲

The designations concerning the basic and the contaminating distributions are given in Table 1.

The task is to estimate the parameters of both distributions, of basic and contaminating ones.

4. PEROBLS D Solution

The parameter estimators for both distributions will be derived from the maximal probability of the event:

$$\left\{\left(D''_{(1)} \wedge \cdots \wedge D''_{(n''_X)}\right) \wedge \left(z' \le z'_A\right) \wedge \left(z''_A \le z'' \le z''_B\right)\right.$$
$$\left. \wedge \left(D'_{(n'_X+1)} \wedge \cdots \wedge D'_{(n'-n'_Z)}\right) \wedge \left(z' > z'_B\right) \wedge \left(D''_{(n''-n'_Z+1)} \wedge \cdots \wedge D''_{(n'')}\right)\right\},$$

where $D''_{(1)} = \left(X''_{(1)} \le X'' \le X''_{(1)} + \mathrm{d}x''\right)$, $D'_{(1)} = \left(X'_{(1)} \le X' \le X'_{(1)} + \mathrm{d}x'\right)$, $\cdots$, etc.

[1]PEROBLS D is an abbreviation of the initial letters: Perović's Robust Least-Square Method; D—by distribution Decomposing.

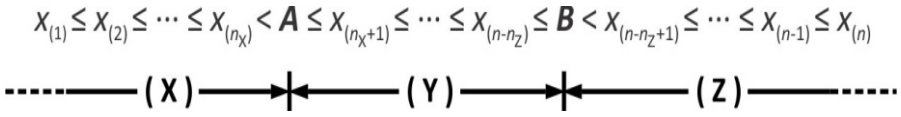

Figure 1. Partition points, A and B.

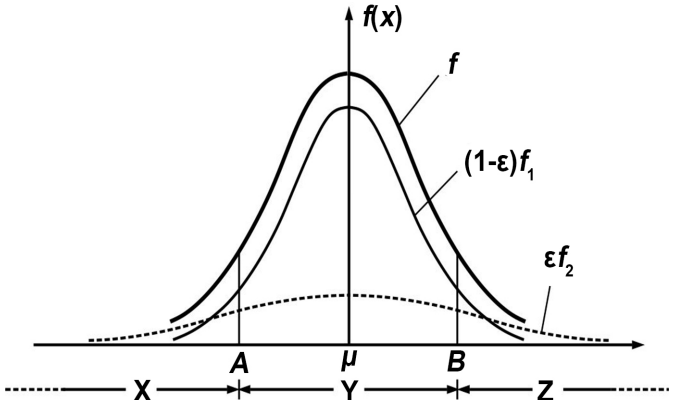

Figure 2. Tukeyan mixed distribution of normally distributed observations.

Table 1. Designations and terms of the quantities appearing in the basic and the contaminating distributions.

Designations		Terms o f Quantities
Basic Distribution	Contaminating Distribution	
X'	X''	Random variable (observation)
σ_1	σ_2	Standard deviation ($\sigma_2 > \sigma_1$)
μ	μ	Expectation
n'	n''	Number of measurements (total)
n'_X	n''_X	Number of measurements in X region
n'_Y	n''_Y	Number of measurements in Y region
n'_Z	n''_Z	Number of measurements in Z region

Let

$$\left.\begin{aligned} z' &= \frac{x'-\mu}{\sigma_1},\ z'_A = \frac{A-\mu}{\sigma_1},\ z'_B = \frac{B-\mu}{\sigma_1} \\ z'' &= \frac{x''-\mu}{\sigma_2},\ z''_A = \frac{A-\mu}{\sigma_2},\ z''_B = \frac{B-\mu}{\sigma_2} \end{aligned}\right\} \tag{4}$$

$$f(z') = \frac{1}{2\sqrt{\pi}}\exp\left(z'^2/2\right),\quad f(z'') = \frac{1}{2\sqrt{\pi}}\exp\left(z''^2/2\right).$$

Then the *likelihood function*, up to the proportionality constant k, is:

$$L = k\cdot\prod_X f\left(z''_{(i)}\right)\cdot p\left(z' \le z'_A\right)\cdot p\left(z''_A \le z'' \le z''_B\right)$$
$$\cdot\prod_Y f\left(z'_{(i)}\right)\cdot p\left(z' > z'_B\right)\cdot\prod_Z f\left(z''_{(i)}\right)\cdot\sigma_1^{-n''_Y}\cdot\sigma_2^{-n''_X-n''_Z},$$

where X, Y and Z under product sign mean: $\prod_X f\left(z''_{(i)}\right)=\prod_{i=1}^{n_X} f\left(z''_{(i)}\right)$, etc., and where:

$$n'=n'_X+n'_Y+n'_Z,\quad n''=n''_X+n''_Y+n''_Z,\quad n=n'+n''.$$

If we also introduce the notations

$$F'_A=\int_{-\infty}^{z'_A} f(z')\mathrm{d}z',\ F'_B=\int_{-\infty}^{z'_B} f(z')\mathrm{d}z',$$

$$F''_A=\int_{-\infty}^{z''_A} f(z'')\mathrm{d}z'',\ F''_B=\int_{-\infty}^{z''_B} f(z'')\mathrm{d}z'',$$

$$A'_X=\int_{-\infty}^{z'_A} z'f(z')\mathrm{d}z',\ A'_Y=\int_{z'_A}^{z'_B} z'f(z')\mathrm{d}z',\ A'_Z=\int_{z'_B}^{\infty} z'f(z')\mathrm{d}z',$$

$$A''_X=\int_{-\infty}^{z''_A} z''f(z'')\mathrm{d}z'',\ A''_Y=\int_{z''_A}^{z''_B} z''f(z'')\mathrm{d}z'',\ A''_Z=\int_{z''_B}^{\infty} z''f(z'')\mathrm{d}z'',$$

$$B'_X=\int_{-\infty}^{z'_A} z'^2 f(z')\mathrm{d}z',\ B'_Z=\int_{z'_B}^{\infty} z'^2 f(z')\mathrm{d}z',\ B'_Y=\int_{z''_A}^{z''_B} z''^2 f(z'')\mathrm{d}z'',$$

$$a'_A=f(z'_A)/F'_A,\ b'_B=f(z'_B)/(1-F'_B),$$

$$d''_{AB}=\left(f(z''_B)-f(z''_A)\right)/\left(F''_B-F''_A\right),$$

$$g''_{AB}=\left(z''_B f(z''_B)-z''_A f(z''_A)\right)/\left(F''_B-F''_A\right),$$

$$n'_X=n'F'_A,\ n'_Y=n'\left(F'_B-F'_A\right),\ n'_Z=n'\left(1-F'_B\right),$$

$$n''_X=n''F''_A,\ n''_Y=n''\left(F''_B-F''_A\right),\ n''_Z=n''\left(1-F''_B\right),$$

then the conditions $\frac{\partial \ln L}{\partial \mu}=0$, $\frac{\partial \ln L}{\partial \sigma_1}=0$ and $\frac{\partial \ln L}{\partial \sigma_2}=0$ yield the equations:

$$\left.\begin{aligned}
\frac{\partial \ln L}{\partial \mu}&=\frac{-n'_X}{\sigma_1}a'_A+\frac{n'_Z}{\sigma_1}b'_B+\frac{1}{\sigma_1}\sum\nolimits_Y z'_i-\frac{n''_Y}{\sigma_2}d''_{AB}+\frac{1}{\sigma_2}\sum\nolimits_X z''_i+\frac{1}{\sigma_2}\sum\nolimits_Z z''_i=0\\
\frac{\partial \ln L}{\partial \sigma_1}&=\frac{-n'_Y}{\sigma_1}-\frac{n'_X}{\sigma_1}z'_A a'_A+\frac{n'_Z}{\sigma_1}z'_B b'_B+\frac{1}{\sigma_1}\sum\nolimits_Y z'^2_i=0\\
\frac{\partial \ln L}{\partial \sigma_2}&=-\frac{n''_X+n''_Z}{\sigma_2}-\frac{n''_Y}{\sigma_2}g''_{AB}+\frac{1}{\sigma_2}\sum\nolimits_X z''^2_i+\frac{1}{\sigma_2}\sum\nolimits_Z z''^2_i=0
\end{aligned}\right\}\quad(5)$$

solvable only iteratively. There are many methods; here direct iterations are given.

However, within system (5), except μ, σ_1 and σ_2, the sums $\sum_Y z'_i$, $\sum_X z''_i$, $\sum_Z z''_i$, $\sum_X z''^2_i$ and $\sum_Z z''^2_i$, and the numbers n' and n'' are also unknown and they should be previously determined.

For the purpose of determining n' and n'' there are many ways. The present author has examined a few methods out of which he has adopted the least-square one. With *three relationships*:

$$\left.\begin{aligned}
F'_A n'+F''_A n''&=n_X+v_X\\
F'_B n'+F''_B n''&=n_{XY}+v_{XY}\\
n'+n''&=n+v_n
\end{aligned}\right\}\quad(6)$$

where v_X, v_{XY} and v_n are the corrections to the "observations" n_X, n_{XY} and n, $n_{XY}=n_X+n_Y$, first using LS with assuming the "observation" weights:

$P_X = 1, P_{XY} = 1$ and $P_n = 1$, we find the LS *estimates* for n' and n'':

$$n_1 = \frac{1}{D}(PU - NV), \quad n_2 = \frac{1}{D}(MV - NU), \text{ with}$$

$$\left.\begin{aligned} M &= 1 + F_A'^2 + F_B'^2, \quad D = MP - N^2 \\ N &= 1 + F_A'F_B' + F_B'F_B'', \quad U = n + n_X F_A' + (n_X + n_Y)F_B' \\ P &= 1 + F_A''^2 + F_B''^2, \quad V = n + n_X F_A'' + (n_X + n_Y)F_B'' \end{aligned}\right\}$$

and then

$$n' = qn_1 \text{ and } n'' = qn_2 \text{ with } q = \frac{n}{n_1 + n_2}.$$

In this way the condition $n' + n'' = n$ is satisfied, but the conditions: $n_X' + n_X'' = n_X$, $n_Y' + n_Y'' = n_Y$ and $n_Z' + n_Z'' = n_Z$ are not satisfied. However, since all conditions in Equations (6) cannot be satisfied simultaneously, a compromise yielding a solution close to the optimum must be accepted.

The sums $\sum_Y z_i'$, $\sum_X z_i''$, etc., can be also solved in various ways, but the present author has chosen the following one. At first we find the sums:

$$\sum_Y X_i' = \sum_Y X_i - \sum_Y X_i'',$$

$$\sum_Y (X_i' - \mu)^2 = \sum_Y (X_i - \mu)^2 - \sum_Y (X_i'' - \mu)^2,$$

$$\sum_X (X_i'' - \mu)^2 = \sum_X (X_i - \mu)^2 - \sum_X (X_i' - \mu)^2,$$

$$\sum_Z (X_i'' - \mu)^2 = \sum_Z (X_i - \mu)^2 - \sum_Z (X_i' - \mu)^2,$$

and then *by means the asymptotic theory* according to which:

$$\frac{1}{n'}\sum_{i=1}^{n'} X_i' \xrightarrow[n'\to\infty]{p} \mu = \int_{-\infty}^{\infty} x' f(x')\mathrm{d}x' = \sigma_1 \int_{-\infty}^{\infty} z' f(z')\mathrm{d}z' + \mu \int_{-\infty}^{\infty} f(z')\mathrm{d}z',$$

$$\frac{1}{n'}\sum_{i=1}^{n'} (X_i' - \mu)^2 \xrightarrow[n'\to\infty]{p} \sigma_1^2 \int_{-\infty}^{\infty} (x' - \mu)^2 f(x')\mathrm{d}x' = \sigma_1^2 \int_{-\infty}^{\infty} z'^2 f(z')\mathrm{d}z',$$

etc., it follows:

$$\sum_Y X_i' \xrightarrow[n'\to\infty]{p} n'\sigma_1 \underbrace{\int_Y z' f(z')\mathrm{d}z'}_{A_Y'} + n'\mu \underbrace{\int_Y f(z')\mathrm{d}z'}_{F_Y'},$$

$$\sum_Y (X_i' - \mu)^2 \xrightarrow[n'\to\infty]{p} n'\sigma_1^2 \underbrace{\int_Y z'^2 f(z')\mathrm{d}z'}_{B_Y'},$$

one introduces *the substitutions*:

$$\sum_X z_i'' = n''A_X'', \quad \sum_Z z_i'' = n''A_Z'', \quad \sum_Y x_i'' = n''\sigma_2 A_Y'' + n''\mu F_Y'',$$

$$\sum_Y z_i'^2 = \frac{1}{\sigma_1^2}\left[\sum_Y (X_i - \bar{X}_Y)^2 + n_Y (\bar{X}_Y - \mu)^2 - n''\sigma_2^2 B_Y''\right],$$

$$\sum_D z_i''^2 = \frac{1}{\sigma_2^2}\left[\sum_D (X_i - \bar{X}_D)^2 + n_D (\bar{X}_D - \mu)^2 - n'\sigma_1^2 B_D'\right], \quad D = X, Y.$$

where:

$$F_X'' = F_A'', \quad F_Y'' = F_B'' - F_A'', \quad F_Z'' = 1 - F_B''.$$

Using these results *the solution* of system (5) we have:

$$\mu_{k+1}=\frac{1}{n_Y}\left[\sum_Y X_i-n_k''\sigma_{2,k}A_{Y,k}''+\sigma_{1,k}\left(n_{Z,k}'b_{B,k}'-n_{X,k}'a_{A,k}'\right)+\frac{\sigma_{1,k}^2}{\sigma_{2,k}^2}\left(n_k''A_{X,k}''+n_k''A_{X,k}''-n_{Y,k}''d_{AB,k}''\right)\right], \tag{7}$$

$$\sigma_{1,k+1}^2=\frac{1}{n_{Y,k}'}\left[\sum_Y\left(X_i-\bar{X}_Y\right)^2+n_Y\left(\bar{X}_Y-\mu_{k+1}\right)^2-\sigma_{1,k}^2n_{X,k}'Z_{A,k}'a_{A,k}'+\sigma_{1,k}^2n_{Z,k}'Z_{B,k}'b_{B,k}'-\sigma_{2,k}^2n_k''B_{Y,k}''\right], \tag{8}$$

$$\sigma_{2,k+1}^2=\frac{1}{n_{X,k}''+n_{Z,k}''}\left[\sum_X\left(X_i-\bar{X}_X\right)^2+\sum_Z\left(X_i-\bar{X}_Z\right)^2+n_X\left(\bar{X}_X-\mu_{k+1}\right)^2+n_Z\left(\bar{X}_Z-\mu_{k+1}\right)^2-\sigma_{2,k}^2n_{Y,k}''g_{AB,k}''+\sigma_{1,k}^2\left(n_k'B_{X,k}'+n_k'B_{Z,k}'\right)\right], \tag{9}$$

where: $\bar{X}_X=\frac{1}{n_X}\sum_X X_i$, $\bar{X}_Y=\frac{1}{n_Y}\sum_Y X_i$, $\bar{X}_Z=\frac{1}{n_Z}\sum_Z X_i$.

Let $\boldsymbol{x}_k^{\mathrm{T}}=\left[\mu_k \quad \sigma_{1,k}^2 \quad \sigma_{2,k}^2\right]$ be the vector of parameter estimates in the k-th iteration and *d the difference vector* of these estimates from (k+1)-th and k-th iterations, then the iterations should be stopped if

$$\|\boldsymbol{d}\|<10^{-q}\|\boldsymbol{x}_k\|,\ q\in\{5,6,7,8,9\}.$$

The *points of optimal partition*, A_{opt} and B_{opt}, with $A_{opt}-\mu=-\left(B_{opt}-\mu\right)$, are found from the condition $(1-\varepsilon)f(x')=\varepsilon f(x'')$, where $1-\varepsilon=n'/n$ and $\varepsilon=n''/n$. So one obtains:

$$A_{opt}=\mu-A',\ B_{opt}=\mu+A',\ A'=\sqrt{\frac{2\ln\left(n'\sigma_2/n''\sigma_1\right)}{1/\sigma_1^2-1/\sigma_2^2}}. \tag{10}$$

The advantages of the method are:

1) *Unbiased estimators* for μ, σ_1^2 and σ_2^2, if assumptions (4) hold and A and B are close to A_{opt} and B_{opt}; and

2) *Minimal variances* for μ, σ_1^2.

The disadvantages of the method are:

1) *A high sensitivity to the choice of the points A and B*, (points A and B must be close to A_{opt} and B_{opt}), which can result in negative estimates for either of the variances σ_1^2 or σ_2^2, or for both; and

2) *Sensitivity to the choice of the initial values for the variances* σ_1^2 and σ_2^2, which, also, can result in negative estimates for one or both variances.

Therefore, the method is recommendable for applications comprising a high number of observations, (for example $n>30$).

Note 2. *If there exists the basic distribution only* (*when in Equation* (3) $\varepsilon=0$), *the method will yield either* $\sigma_1^2=\sigma_2^2$ *or negative values for one or both variances.*

5. Some Robust Estimation

Out of many robust LS methods we shall use here two of them: the method of Maximum Likelihood (ML) and Huber's mad estimation of distribution standard [6] [17].

5.1. The Maximum Likelihood (ML) Method

The ML method is based on the assumption that in the domain Y there exists only the basic distribution, unlike the domaines X and Z where in addition to the basic distribution there exist gross errors and outliers. Here the censoring points are A and B and they are defined according to Equation (2).

In the region ($X \cup Z$), due to the presence of gross errors in the observations, the distribution of the random variable X is not normal. Therefore, the estimates of the parameters μ, and σ^2 are determined on the basis of the probability of the event:

$$\left\{\left(X \le A\right) \wedge D_{(n_X+1)} \wedge \cdots \wedge D_{(n-n_Z)} \wedge \left(B \le X\right)\right\},$$

where the events $D_i = \left\{X_i \le X \le X_i + \mathrm{d}x\right\}$, $i = n_X + 1, \cdots, n - n_Z$, mean that the random variable X is within an interval $\left(X_i, X_i + \mathrm{d}x\right)$ (with differentially small $\mathrm{d}x > 0$).

Then the *likelihood function*, up to the proportionality constant, is [18]:

$$L = \frac{n!}{n_X!\, n_Z!}\sigma^{-(n-n_X-n_Z)}\left[F\left(Z_A\right)\right]^{n_X}\prod_{i=n_X+1}^{n-n_Z} f\left(z_{(i)}\right)\left[1-F\left(Z_B\right)\right]^{n_Z},$$

and the ML estimators are the solutions of the equations

$$\frac{\partial \ln L}{\partial \mu} = 0, \quad \frac{\partial \ln L}{\partial \sigma} = 0. \quad \frac{\partial \ln L}{\partial \mu} = 0, \text{ and } \quad \frac{\partial \ln L}{\partial \sigma} = 0. \tag{11}$$

Equations (11) have no analytical solution and they must be solved *iteratively.* There are several methods; here direct iterations are given:

$$\mu_{k+1} = \bar{X} - \sigma_k \frac{n_X}{n_Y}\cdot\frac{f\left(a_k\right)}{F\left(a_k\right)} + \sigma_k \frac{n_Z}{n_Y}\cdot\frac{f\left(b_k\right)}{1-F\left(b_k\right)}, \tag{12}$$

$$\sigma_{k+1}^2 = m^2 + \left(\bar{X} - \mu_{k+1}\right)^2 - \frac{n_X}{n_Y}\cdot\sigma_k^2\frac{a_k f\left(a_k\right)}{F\left(a_k\right)} + \frac{n_Z}{n_Y}\cdot\sigma_k^2\frac{b_k f\left(b_k\right)}{1-F\left(b_k\right)}, \tag{13}$$

where:

$$\bar{X} = \frac{1}{n_Y}\sum_{n_X+1}^{n-n_Z} X_i, \quad m^2 = \frac{1}{n_Y}\sum_{n_X+1}^{n-n_Z}\left(X_{(i)} - \bar{X}\right)^2,$$

$$a_k = \frac{A-\mu_k}{a_k}, \quad b_k = \frac{B-\mu_k}{a_k},$$

As initial values we can adopt $\mu_0 = \bar{X}$, and $\sigma_0^2 = m^2$.

In the present author's opinion under the preposition of existing of contaminating distribution the method yields increased estimates for σ^2.

5.2. Huber's Mad Robust Estimation of Distribution Standard

For the purpose of estimating an unknown standard σ Huber 1981 [6] and Birch and Mayers 1982 [17] proposed a median estimator for σ:

$$\sigma = \frac{mad\left(X\right)}{0.6745}, \quad \left(mad\left(X\right) = med \mid X_i - medX_j \mid\right). \tag{14}$$

6. Results and Discussion

Example 1. For the sake of verifying *the correctness of PEROBLS D method* and examining the *appropriateness of Maximum Likelihood* (*ML*) *method* in Table 2 are presented the *exact* (*expected*) *values* of some quantities from the mixed *Tukeyan distribution* (3), with $n' = 0.8n$, $n'' = 0.2n$ $(\varepsilon = 20\%)$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $\mu = 5$ and symmetrical partitioning ($n_X = n_Z$). The numbers n_X, n_Y and n_Z and the other quantities in the table are presentd for the case $n = 10^6$.

According to data in Table 2, for two methods—PEROBLS D and ML—the estimates for the corresponding quintiles are calculated and presented in Table 3. The results of estimating the variance σ_1^2 of the basic distribution by using ML methods indicate their appropriateness.

Example 2. Simulated Data. Using normally distributed $N(0, 1)$ random numbers from Tables of *Bol'shev* and *Smirnov* 1968 [19] the mixed *Tukeyan* distribution (3) is found (Table 4), with normal distributions: basic $N\left(\mu = 5, \sigma_1^2 = 1\right)$ with $n' = 160$ and contaminating $N\left(\mu = 5, \sigma_2^2 = 4\sigma_1^2 = 4\right)$ with $n'' = 40$, ($\Rightarrow n = 200$); with estimates:
basic: $\bar{X} = 4.965$, $\sigma_1^2 = 1.1070$,
contaminating: $\bar{X} = 5.000$, $\sigma_2^2 = 4.1554$.

According to Equation (10), for $\sigma_1^2 = 1$ and $\sigma_2^2 = 4$, the *optimal partition points* are:

$$A_{OPT} = 2.6452,\ B_{OPT} = 7.3548,$$

for which one obtains 6.5 *percent partition* with $n_X = 8$ and $n_Z = 5$.

Table 2. The exact (expected) values of some quantities from the mixed Tukeyan distribution (3), with $n' = 0.8n$, $n'' = 0.2n$ $(\varepsilon = 20\%)$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $\mu = 5$ and symmetrical partitioning ($n_X = n_Z$).

Designations	Partitioning $n_X + n_Z$ [%] (Censoring—for ML)			
	10	20	30	40
(A; B)	(3.0; 7.0)	(3.5; 6.5)	(3.82; 6.18)	(4.04; 5.96)
n_X	49,931	98,771	150,719	197,945
n_Y	900,138	802,458	698,562	604,110
n_Z	49,931	98,771	150,719	197,945
$\sum_X X_i$	0,109,674.4	0,269,786.7	0,460,419.4	0,646,195.1
$\sum_Y X_i$	4,500,689.1	4,012,288.6	3,492,809.5	3,020,552.9
$\sum_Z X_i$	0,389,636.5	0,717,924.7	1,046,771.0	1,333,252.0
$\sum_X \left(X_i - \bar{X}_X\right)^2$	0,032,651.4	0,062,535.6	0,092,943.8	0,120,914.5
$\sum_Y \left(X_i - \bar{X}_Y\right)^2$	0,749,827.2	0,458,297.8	0,273,552.7	0,165,779.4
$\sum_Z \left(X_i - \bar{X}_Z\right)^2$	0,032,651.4	0,062,535.6	0,092,943.8	0,120,914.5
$\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$	1600000.0	1600000.0	1600000.0	1600000.0

Table 3. The results of Estimating the Normal-Distribution Parameters by using PEROBLS D, and ML methods according to the exact (expected) values of the corresponding sums given in **Table 2**.

Method	Formula	Quantity	Partitioning [%] (Censoring—for ML)			
			10	20	30	40
PEROBLS D	(7)	μ	5.0000	5.0000	5.0000	5.0000
ML	(12)		5.0000	5.0000	5.0000	5.0000
PEROBLS D	(8)	σ_1^2	1.0000	1.0000	1.0000	1.0000
ML	(13)		1.3839	1.3247	1.2920	1.2732
PEROBLS D	(9)	σ_2^2	4.0001	4.0001	4.0001	4.0001
ML	-		-	-	-	-

Table 4. The 200 ($n = 200$) simulated observations of the *Tukeyan* mixed distribution (3), with $\mu = 5$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$ and $\varepsilon = 0.20$; ($\Rightarrow n'' = 40$, $n' = 160$).

X_i	n_i	X_i	n_i	X_i	n_i	X_i	n_i
0.3	1	3.6	4	5.0	11	6.4	4
0.9	1	3.7	3	5.1	5	6.5	2
1.9	2	3.8	5	5.2	5	6.6	4
2.2	1	3.9	5	5.3	5	6.7	2
2.5	2	4.0	3	5.4	9	6.8	2
2.6	1	4.1	3	5.5	7	6.9	1
2.7	1	4.2	7	5.6	5	7.0	1
2.8	1	4.3	7	5.7	7	7.1	2
2.9	2	4.4	8	5.8	8	7.3	1
3.0	3	4.5	7	5.9	5	7.4	1
3.2	1	4.6	7	6.0	5	7.5	1
3.3	1	4.7	4	6.1	2	8.3	1
3.4	2	4.8	3	6.2	4	9.2	1
3.5	3	4.9	6	6.3	4	9.9	1

In **Table 5**, for various choices of the partition points A and B, a survey of the parameter estimates for the basic and contaminating distributions obtained by different methods is given. The parameter estimates in the PEROBLS D method are close to the exact ones, whereas in the case of application of the ML and Huber-mad methods the variance of the basic distribution is overestimated.

The best way is to choose the partition points A and B for the PEROBLS D method from the frequency histogram (see **Figure 3**) by accepting the x values for which it, to the left and right of the distribution centre, begins to have values above the smoothing curve for the normal distribution.

Recommendation. *The estimate of standard* σ, *for the purpose of drawing the smoothing curve can be calculated according to the standard formula,* $m^2 = \sum\left(X_i - \bar{X}_X\right)^2 \Big/ (n-1)$ *where* 2% - 5% *rejected outlying observations not taken into account.* ▲

Table 5. The parameter estimates of a normal distribution based on 200 simulated observations of the Tukeyan mixed distribution with $\mu = 5$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $n' = 160$ and $n'' = 40$, $(\varepsilon = 0.20)$.

Method	Formula	Quantity	$A_{opt} = 2.65$ $B_{opt} = 7.35$ 6.5%	$A = 2.85$ $B = 6.85$ 10%	$A_1 = 3.15$ $B_1 = 6.55$ 16.5%	$A_2 = 3.15$ $B_2 = 6.15$ 23.5%
PEROBLS D	(7)	μ	**5.0130**	5.0448	5.0381	5.0188
ML	(12)		4.9670	4.9659	4.9761	4.9786
PEROBLS D	(8)	σ_1^2	**1.0382**	1.2996	0.9670	0.7963
ML	(13)		1.4721	1.4316	1.4299	1.4429
Huber (mad)	(14)			1.4067		
PEROBLS D	(9)	σ_2^2	**3.8720**	7.6081	3.3849	2.8232

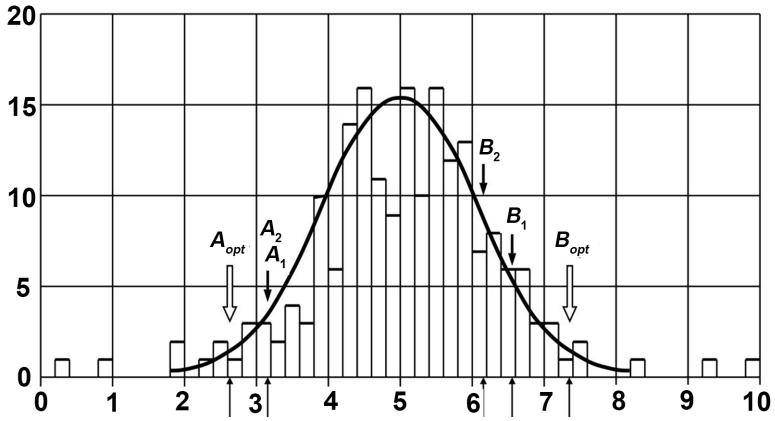

Figure 3. Frequency Histogram for 200 simulated observations of the Tukeyan mixed distribution (3) with $\mu = 5$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$ and $\varepsilon = 0.20$.

7. Conclusions

On the basis of the obtained results in Examples 1 and 2 we can conclude the following:

1) On the basis of exact (expected) values from Example 1 the validity of the PEROBLS D method in the parameter estimation (expectation and variance) for both distributions in the Tukeyian mixed distribution of observations is confirmed. Here the variance estimates for both distributions, basic and contaminating ones, are correct, *i.e.* their values are exact.

2) On the basis of simulated realistic measurements from Example 2 good (satisfactory) parameter estimates for both distributions are also confirmed.

Acknowledgements

Dr. ZORICA Cvetković from the Astronomical Observatory in Belgrade is the author of the FORTRAN programmes who performed the calculations in the examples. The ALHIDADA D.O.O. Company from Petrijevci (Croatia) has financially supported the publishing of the article. The figures were made by Dar-

ko Andjic, Dr. and graduated engineer of Geodesy from the Republic Geodetic Authority, Podgorica, Montenegro.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Huber, P.J. (1964) Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, **35**, 73-101. https://doi.org/10.1214/aoms/1177703732

[2] Arango-Castillo, L. and Takahara, G. (2018) Robust Estimation of the Sample Mean Variance for Gaussian Processes with Long-Range Dependence. 2017 *IEEE Global Conference on Signal and Information Processing* (*GlobalSIP*), Montreal, QC, 14-16 November 2017, 201-205. https://doi.org/10.1109/GlobalSIP.2017.8308632

[3] Hald, A. (1986) Galileo's Statistical Analysis of Astronomical Observations. *International Statistical Review*, **54**, 211-220. https://doi.org/10.2307/1403145

[4] Hampel, F., Roncetti, E., Rausseew, P.J. and Stael, V. (1986) Robust Statistics; The Approach Based on Influence Functions. John Wiley, New York (Russian translation, Mir, Moscow, 1989).

[5] Tukey, J.W. (1960) A Survey of Sampling from Contaminated Distributions. In: Oklin, I., Ed., *Contributions to Probability and Statistics*, Stanford University Press, Redwood City, CA.

[6] Huber, P.J. (1981) Robust Statistics. John Wiley and Sons, New York. https://doi.org/10.1002/0471725250

[7] Chen, Y. and Jackson, D. (1995) Robust Estimation of Mean and Variance in Fisheries. *Transactions of the American Fisheries Society*, **124**, 401-412. https://doi.org/10.1577/1548-8659(1995)124<0401:REOMAV>2.3.CO;2

[8] Grossi, L. and Laurini, F. (2011) Robust Estimation of Efficient Mean-Variance Frontiers. *Advances in Data Analysis and Classification*, **5**, 3-22. https://doi.org/10.1007/s11634-010-0082-3

[9] Li, Y., Li, J., Qi, J. and Chen, L. (2019) Robust Cubature Kalman Filter for Dynamic State Estimation of Synchronous Machines Under Unknown Measurement Noise Statistics. *IEEE Access*, **7**, 29139-29148. https://doi.org/10.1109/ACCESS.2019.2900228

[10] Singh, A. and Nocerino, J. (2002) Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations. *Chemometrics and Intelligent Laboratory Systems*, **60**, 69-86. https://doi.org/10.1016/S0169-7439(01)00186-1

[11] Vilà-Valls, J., Wei, Q., Closas, P. and Fernández-Prades, C. (2014) Robust Gaussian Sum Filtering with Unknown Noise statistics: Application to TargetTracking. 2014 *IEEE Workshop on Statistical Signal Processing*, Gold Coast, Australia, 29 June-2 July 2014, 416-419. https://doi.org/10.1109/SSP.2014.6884664

[12] Wang, Ch., Wang, Y., Lin, Z., Yuille, A.L. and Gao, W. (2014) Robust Estimation of 3D Human Poses from a Single Image. CBMM Memo No. 013. https://cbmm.mit.edu/sites/default/files/publications/CBMM-Memo-013.pdf https://doi.org/10.1109/CVPR.2014.303

[13] Zhou, W.-Z., Bose, K., Fan, J. and Liu, H. (2018) A New Perspective on Robust *M*-Estimation: Finite Sample Theory and Applications to Ddependence-Adjusted

Multiple Testing. *The Annals of Statistics*, **46**, 1904-1931. https://doi.org/10.1214/17-AOS1606

[14] Zhao, J. and Mili, L. (2018) A Framework for Robust Hybrid State Estimation with Unknown Measurement Noise Statistics. *IEEE Transactions on Industrial Informatics*, **14**, 1866-1875. https://doi.org/10.1109/TII.2017.2764800

[15] Perović, G. (1989) Adjustment of Calculus, Book I Theory of Measurement Errors, 2nd Revised and Enlarged Edition, Naučna knjiga, Belgrade (In Serbo-Croat).

[16] Perović, G. (2005) Least Squares. Publisher: Author, Belgrade.

[17] Birch, J.B. and Myers, R.H. (1982) Robust Analysis of Covariance. *Biometrics*, **38**, 699-713. https://doi.org/10.2307/2530050

[18] Schneider, H. (1986) Truncated and Censored Samples from Normal Populations. Marcel Dekker, Inc., New York and Basel.

[19] Bol'shev, L.N. and Smirnov, N.V. (1968) Tables of Mathematical Statistics. VTS AN SSSR, Moscow (In Russian).