Hindawi
Complexity
Volume 2019, Article ID 5986562, 24 pages
https://doi.org/10.1155/2019/5986562

Research Article

Geodesic Distance on Gaussian Manifolds to Reduce the Statistical Errors in the Investigation of Complex Systems

Michele Lungaroni,[1] **Andrea Murari**,[2] **Emmanuele Peluso**,[1] **Pasqualino Gaudio**,[1] **and Michela Gelfusa**[1]

[1]*Department of Industrial Engineering, University of Rome Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy*
[2]*Consorzio RFX (CNR, ENEA, INFN, Universita' di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy*

Correspondence should be addressed to Michele Lungaroni; michele.lungaroni@uniroma2.it

Received 18 March 2019; Revised 17 July 2019; Accepted 21 July 2019; Published 18 August 2019

Academic Editor: Dan Selişteanu

In the last years the reputation of medical, economic, and scientific expertise has been strongly damaged by a series of false predictions and contradictory studies. The lax application of statistical principles has certainly contributed to the uncertainty and loss of confidence in the sciences. Various assumptions, generally held as valid in statistical treatments, have proved their limits. In particular, since some time it has emerged quite clearly that even slightly departures from normality and homoscedasticity can affect significantly classic significance tests. Robust statistical methods have been developed, which can provide much more reliable estimates. On the other hand, they do not address an additional problem typical of the natural sciences, whose data are often the output of delicate measurements. The data can therefore not only be sampled from a nonnormal pdf but also be affected by significant levels of Gaussian additive noise of various amplitude. To tackle this additional source of uncertainty, in this paper it is shown how already developed robust statistical tools can be usefully complemented with the Geodesic Distance on Gaussian Manifolds. This metric is conceptually more appropriate and practically more effective, in handling noise of Gaussian distribution, than the traditional Euclidean distance. The results of a series of systematic numerical tests show the advantages of the proposed approach in all the main aspects of statistical inference, from measures of location and scale to size effects and hypothesis testing. Particularly relevant is the reduction even of 35% in Type II errors, proving the important improvement in power obtained by applying the methods proposed in the paper. It is worth emphasizing that the proposed approach provides a general framework, in which also noise of different statistical distributions can be dealt with.

1. Robust Statistics and Measurement Errors in the Science of Complex Systems

In the last decades the number of contradictory, inaccurate, and/or misleading scientific pronouncements reported in the media about complex systems has increased exponentially. From medicine to finance, very often opposing studies and findings have generated a quite high level of confusion. Moreover various expectations, predictions, and evaluations have been very often completely contradicted by reality; a very evident example is the financial crisis of 2008. This situation has considerably corroded the public confidence in expert knowledge [1] and even more generally in the sciences. The status of peer reviewed journals in various fields is not more reassuring. In a famous paper and subsequent works, it has been shown how many studies, published in the most highly respected medical journals, are refuted by other investigations in a matter of months or at the most a few years [2].

There are many causes to the worrying situation previously described, from the reduction in public funding to the corporate takeover of public functions and institutions. However, also the misguided application of inadequate statistical tools has certainly played a role and contributed to exacerbate the problem. Indeed classic significance tests are the main statistical tools in many disciplines, ranging from medicine and education to economics and psychology. In the last decades, hypothesis tests have been increasingly used also

in the so called exact sciences, such as physics, chemistry and engineering, particularly when complex systems have to be studied. These classic methods are based on certain specific assumptions, which have to be reasonably satisfied; otherwise the accuracy of the results can be seriously compromised. One fundamental assumption is that the data are randomly sampled from a Gaussian distribution. In case of hypothesis testing and comparison between independent groups, it is also typically assumed that the distributions have a common variance (even if they present different means); this is the so called homogeneity of variance or homoscedasticity hypothesis. Conversely, in practical applications the data available do not necessarily conform to these assumptions. The probability density functions sampled in experiments are not necessarily Gaussians and can present heavy tails or be skewed. Homoscedasticity is even more frequently violated.

Some of the classic parametric tests are quite robust against violation of the normality assumption, in the sense that they can preserve both the expected rates of Type I and Type II errors, even if they have quite low power compared to their nonparametric counterparts. This has been reported extensively in Sawilowsky [3] and Sawilowsky and Blair [4]. On the contrary, in the last decades, a lot of evidences have emerged showing how a blind reliance on the assumption of homoscedasticity can produce rather inaccurate results as reported in Wilcox [5, 6] and in this respect rank-based nonparametric statistics can fare even worse. In Wilcox [6] it is shown that the violation of homoscedasticity significantly affects type I errors, up to 0.5 at a significant level of 0.05. The power of the tests can be even more severely compromised by violation of the previous assumptions; indeed for distributions characterized by small departures from normality, as reported in Wilcox [5], the power of the t test can be reduced from 0.96 to 0.28. Therefore a lot of efforts have been recently devoted to developing robust tools, which seek to provide methods that compare well with popular statistical techniques, when the classic hypotheses are satisfied, but which are not unduly affected by departures from model assumptions. Developed in the framework of robust statistics, as reported in Huber and Ricchetti [7], these techniques provide quite accurate, even if slightly suboptimal results, in the case the assumptions of normality and homoscedasticity are correct, but are not compromised, if the data have been sampled from a different distribution. However, even if they are quite successful in terms of descriptive statistics, robust techniques can be affected by significant increase in Type I errors when converted into inference statistics. The developments proposed in this paper are indeed meant to improve robust statistics explicitly in this direction.

The tools developed in the framework of robust statistics in the last decades have achieved very impressive results but they do not address at all a problem typical of the experimental sciences. In the vast majority of natural science applications, the data are measurements affected by error bars. Therefore, the available data can be not only sampled from a distribution, which is quite far from a Gaussian, but they can also be affected by significant additive noise. Robust statistical methods address only the first of these two problems, the fact that the assumptions about the probability density function (pdf) generating the data are not satisfied. However, the additional uncertainty introduced by the additive noise in the measurements is not considered.

In all the formulas developed in the framework of robust statistics, the Euclidean distance is implicitly assumed to be the proper metric to adopt. The Euclidean distance, however, implicitly requires considering all data as single infinitely precise values. This assumption can be appropriate in other applications but it is obviously not so in the natural sciences, since all the measurements are typically affected by noise. As will be shown later, an inappropriate evaluation of the uncertainties in the measurements can have a major impact even on the determination of the basic statistical measures such as the mean. An alternative approach is to use a new distance between data, which would take into account the measurement uncertainties. The idea, behind the method proposed in this paper, consists of considering the measurements not as points, but as Gaussian distributions. This is a valid assumption in many scientific applications, because the measurements are affected by a wide range of noise sources, which, from a statistical point of view, can be considered random variables.

Modelling measurements not as punctual values, but as Gaussian distributions, requires defining a distance between Gaussians. This distance is the Geodesic on the Gaussian Manifold (GDGM) and can be expressed as a closed formula (see Section 3). As shown in the rest of the paper, adopting this geodesic distance increases significantly the accuracy of robust statistical tools, even when the data are affected by a limited level of noise, particularly for hypotheses testing.

With regard to the structure of the paper, next section is devoted to an introductory discussion about the importance of a proper evaluation of the errors associated with the experimental measurements in the framework of a modern theory of uncertainty. Section 3 provides the background on the main mathematical tool introduced in the paper: the Geodesic Distance on Gaussian Manifolds. The probability density functions used in the paper to test the proposed new approach are reviewed in Section 4. The main ideas behind robust statistics are introduced in Section 5. The proposed new method is applied first to the measurements of location and scale, as described in Sections 6 and 7. The impact of using the GDGM in hypothesis testing is described in Section 8. The performance of GDGM in the case the sampled pdf is asymmetric and the additive noise is heteroscedastic is exemplified in Section 9. Conclusions and lines of future work are provided in the last Section 10 of the paper.

2. The Theory of Uncertainty and the Experimental Measurements

In the science of complex systems, measurements are the basic inputs required to provide quantitative knowledge about phenomena. However, all the measurements provide limited information about the measure and since they are affected by uncertainties. At the end of last century, the theory of uncertainty was consolidated and became the dominant paradigm with the publication of the IEC-ISO “Guide to

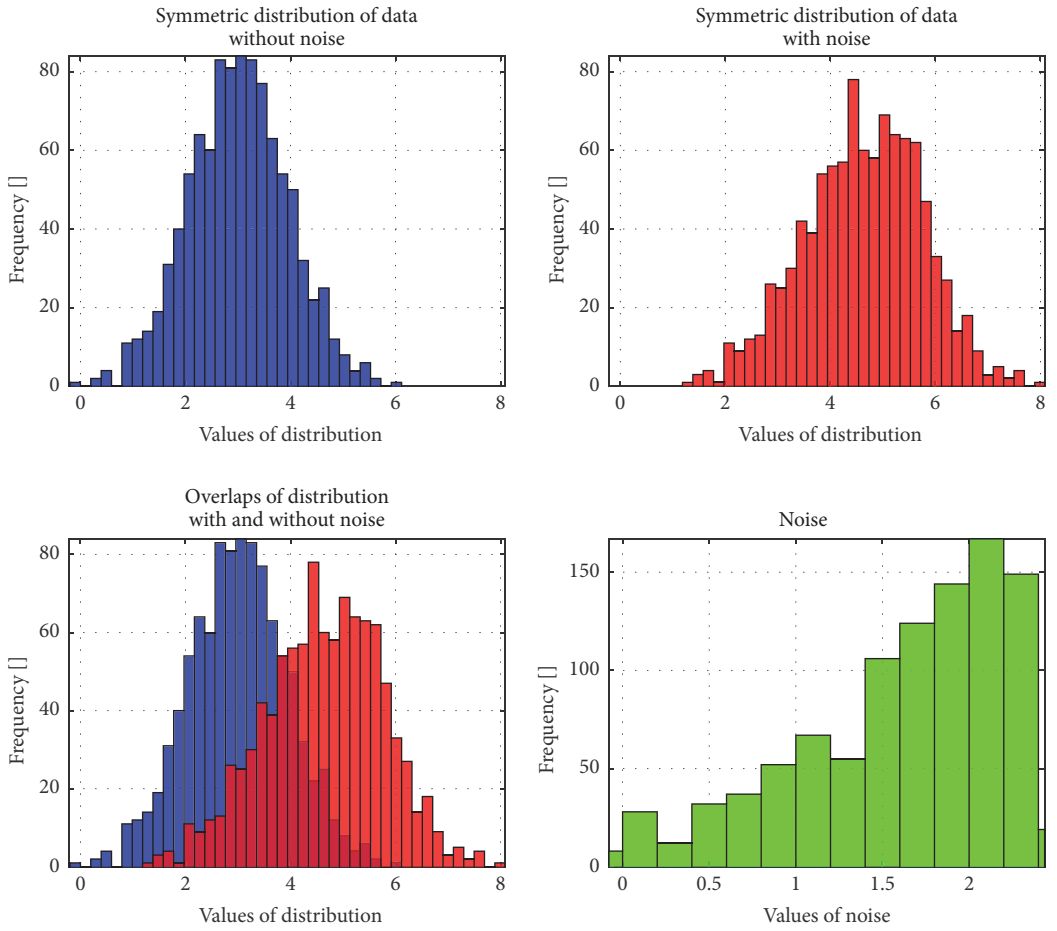


Figure 1: Distribution of data sampled from a Gaussian distribution (of $\overline{x}$=3 and s=1) affected by Gaussian noise of zero mean and standard deviation basically proportional to the absolute value of the data (see bottom right plot).

the Expression of Uncertainty in Measurement" (GUM) in 1992 [8]. Assuming that all the systematic errors have been eliminated, the uncertainty to be associated with the results of a series of measurements x_i is their standard deviation, defined as

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left[(x_i) - \overline{x}\right]^2}{n-1}} \quad (1)$$

In the previous equation, the standard deviation s is expressed in terms of the average $\overline{x}$ of the n available observations. This emphasizes how the process of calculating the average is at the very basis of the first, most elementary statistical steps, required to quantify even the most elementary quality of the available measurements. Therefore any error in the determination of the mean has the potential to invalidate the entire statistical analysis. To illustrate the potential impact of the issue let us analyse a case, for which the traditional heteroscedasticity assumption is not valid. In this simple, introductory example, we investigate the case that the measurements are generated by a Gaussian distribution of $\overline{x}$=3 and s=1. The data are considered affected by a Gaussian noise of zero mean but a varying standard deviation. In particular, the noise s increases almost linearly with the amplitude of the measurements; this is not an uncommon case in practice because in many instruments the measurement errors are a fixed percentage of the measured value. This situation is illustrated in Figure 1, which summarise a numerical experiment with 1000 data points.

The accuracy of the traditional estimates of location, mean, and mode is reported in Table 1, together with the values of the robust indicator introduced later in the paper (see Sections 3–5). The values obtained by application of our methodology, using the GDGM, are also shown. In Table 1, as for all the others reported either in the main text or in the Appendixes, the numerical values are the results obtained mediating over 1000 independent realisations. The details of the method proposed in this paper are given later; for the moment what should be retained is that the traditional statistical measures of location fare very badly for this kind of noise. Also the robust statistical indicators commit very serious overestimates. On the contrary, our approach of combining robust indicators with the GDGM reduces the inaccuracy in the estimate of location of even an order of magnitude.

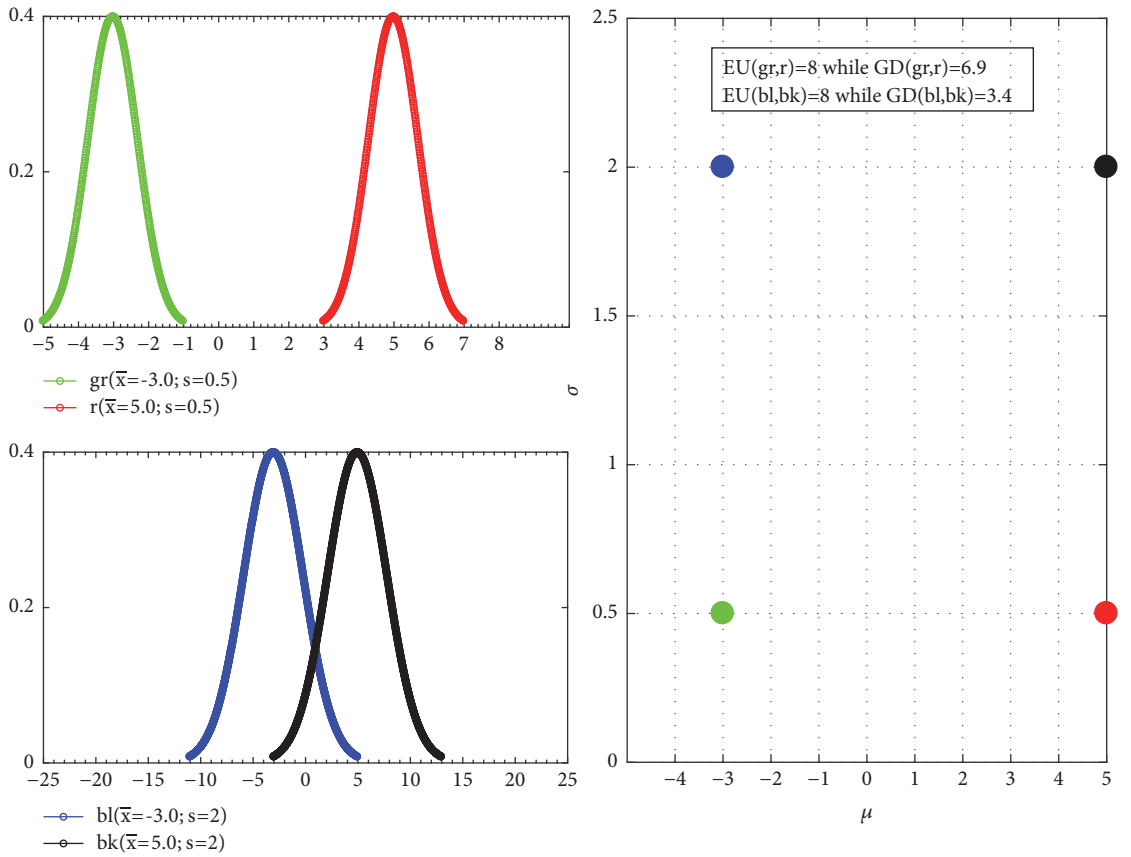


Figure 2: Examples to illustrate how the GDGM determines the distance between two Gaussians. The two couples of pdfs in the figure have the same mean but different s. The geodesic distance between the two with higher s is much smaller (see insert).

Table 1: Measures of location for the data of Figure 1. The ideal value to be obtained is 3. The acronym GD indicates the values obtained with the use of the Geodesic on the Gaussian Manifold (GDGM) proposed in the paper, which reduces the errors in the estimates even of an order of magnitude recovering almost the right value of 3.

	No Noise	Noise
Mode	2.975	4.728
Mean	3.000	4.664
GD_{Mean}	2.335	3.352
Trimmed Mean	2.999	4.687
$GD_{Trimmed}$	1.965	2.939
Wisorized Mean	3.000	4.680
GD_{Winsor}	2.122	3.088

3. Geodesic Distance on Gaussian Manifolds

As mentioned in the previous section, in the natural sciences the data available are typically the result of experimental measurements, which are affected by uncertainties referred to as noise. The sources of this uncertainty are normally several, independent and additive: as a consequence it is more than reasonable to assume that the pdf of the noise is normal. Each measurement can therefore be modelled as a probability density function of the Gaussian type, determined by its mean $\overline{x}$ and its standard deviation s:

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{\sigma^2}\right] \quad (2)$$

The set of normal distributions can be seen as a two dimensional space, or better a two dimensional manifold, parameterised by $\overline{x}$ and s. Modelling measurements not as punctual values, but as Gaussian distributions, requires defining a distance between Gaussians. The most appropriate definition of distance between Gaussian distributions is the geodesic distance, on the probabilistic manifold containing the data, which is not a Euclidean but a Riemannian space. This geodesic distance on the Gaussian manifold can be calculated using the Fisher-Rao metric as it has been applied in Cannas et al. [9] and Murari et al. [10]. For two univariate Gaussian distributions $p_1(x \mid \mu_1, \sigma_1)$ and $p_2(x \mid \mu_2, \sigma_2)$, parameterised by their mean μ_i and standard deviations

σ_i $(i = 1, 2)$, the geodesic distance GD is given by

$$GD\left(p_1 \mid |p_2\right) = \sqrt{2}\ln\frac{1+\delta}{1-\delta} = 2\sqrt{2}\tanh^{-1}\delta,$$
$$where\ \delta = \left[\frac{\left(\mu_1-\mu_2\right)^2 + 2\left(\sigma_1-\sigma_2\right)^2}{\left(\mu_2-\mu_2\right)^2 + 2\left(\sigma_1+\sigma_2\right)^2}\right]^{1/2} \quad (3)$$

A pictorial view of the potential impact of the GDGM is provided in Figure 2, which shows the distance between two couples of Gaussian distributions having the same difference in means. The standard deviations of the two Gaussians in the bottom plot are much larger than the ones of the two Gaussians, whose distance is depicted in the top plot, using the Euclidean distance results in attributing the same distance to the two couples of normal distributions. On the contrary, according to the GDGM, the distance between the two Gaussians with a larger σ is much smaller, which makes intuitive sense since they overlap much more. As will be shown in detail in the next sections, the replacement of the Euclidean distance with the GDGM improves significantly all the robust estimators, reducing the effects of noise and outliers.

4. Nonnormal Distributions

To exemplify and prove the usefulness of the method proposed in this paper, a series of numerical tests has been performed, with probability density functions of families different from the Gaussian. To simulate realistic situations in the natural sciences, Gaussian noise of various amplitudes has been added to the data sampled from the nonnormal pdfs. The tests performed in the rest of the paper have been obtained from the pdfs belonging to the families: log-normal, exponential, contaminated χ^2 and the so called g-h pdf. The analysed distributions are defined in the rest of this section (more details are provided in Wilcox 2005 [11]):

The traditional log normal distribution corresponds to the pdf:

$$f_{\log Nor}\left(x \mid \mu, \sigma\right) = \frac{e^{-(\ln x-\mu)^2/2\sigma^2}}{x\sqrt{2\pi}\sigma} \quad with\ x > 0 \quad (4)$$

The pdf of the exponential distribution can be written as

$$f_{exp}\left(x \mid \lambda\right) = \frac{1}{\lambda}e^{-x/\lambda} \quad (5)$$

The contaminated χ^2 distribution $\chi^2_k(x)$ is the sum of two χ^2 distributions, the first sampled with probability 1- Ɛ, and the second with probability Ɛ. The $\chi^2_k(x)$ is defined as

$$\chi^2_k\left(x \mid k_1, k_2, \varepsilon\right) = (1-\varepsilon)\bullet\chi^2_{k1}\left(x \mid k_1, \varepsilon\right) + \varepsilon \bullet\chi^2_{k2}\left(x \mid k_2, \varepsilon\right) \quad with\ 0 \le \varepsilon \le 1 \quad (6)$$

In (6) k indicates the number of degrees of freedom of the distribution and $\chi^2_{ki}(x)$ is

$$\chi^2_k\left(x \mid k\right) = \frac{1}{2^{k/2}\Gamma\left(k/2\right)}x^{k/2-1}e^{-x/2} \quad with\ x > 0 \quad (7)$$

The g-h distribution is a family of pdfs parameterised by the two values g and h. It consists of a normal distribution $Z = \mathcal{N}(0, 1)$ modified by the parameters g and h, which determine its asymmetry and the relevance of the tail respectively. The pdf of the g-h distribution is

$$f_{gh}\left(x \mid g, h\right) = \frac{\exp\left(gZ\right)-1}{g}\exp\left(\frac{hZ^2}{2}\right) \quad (8)$$

If g = 0 then

$$f_{gh}\left(x \mid g, h\right) = Z\exp\left(\frac{hZ^2}{2}\right) \quad (9)$$

When both parameters are zero, the standard normal $\mathcal{N}(0, 1)$ is recovered.

5. The Approach of Robust Statistics and the GDGM

In this section the needs of a special treatment for pdfs with heavy tails are reviewed. The potential of the GDGM to address the issue of noise is introduced.

5.1. The Issues Presented by Non-Gaussian Pdfs. A typical case, discussed in detail to illustrate the main ideas behind the approach proposed in this paper, is the calculation of the sample mean. When the data are not sampled from a Gaussian distribution or are affected by a significant fraction of outliers, it is well known that a trimmed mean can produce a much better estimate of location as it is reported in Wilcox [11]. An example is the case of the log normal distribution shown in Figure 3.

This is the typical example of a nonsymmetric pdf, with significant asymmetric tails. Contrary to the case of the symmetric normal distribution, in the case of the log normal the mean does not correspond to the maximum value. Indeed, as can be appreciated by Figure 3, the mean value of the distribution is quite displaced to the right of the maximum. Therefore the mean value is no more representative of the most probable value to be expected once the log normal is sampled. As a consequence, if the measures of location are meant to determine the typical value of a quantity, as it is typically the case in the science of complexity, the traditional sample mean is not really representative. The median provides a much better estimate of central tendency for distributions with tails but it can provide quite wrong estimates in the case the data are indeed sampled from a normal distribution.

Various strategies can be adopted to alleviate this problem. In the work of Wilcox [11] they are referred to as robust measures of location. A very important family of these methods is based on some form of trimming of the available data. A trimmed mean or truncated mean consists of calculating the mean of the available data after discarding the high and low end parts of the samples (typically discarding an equal amount of both). The number of samples to be discarded is usually given as a percentage of the total number

Table 2: The estimates of location for different distributions with 20 points without noise. Respectively, we have lognormal distribution $\overline{x} = 2$ and $s = 1$ parameters; exponential distribution $\lambda = 2$ parameter; contaminated χ^2 $k_1 = 4, k_2 = 40$, and $\varepsilon = 0.9$ parameters; G-h distribution $g = 0.5$ and $h = 0.5$ parameters.

	Lognormal	Exponential	Contaminated χ^2	G-h
Mode	6,56209	0,99761	3,80171	0,30349
Mean	12.28140	1.99360	7.59350	0.70637
GD_{Mean}	10.77281	1.80658	6.24093	0.44912
Trimmed Mean	8.56378	1.56370	4.19754	0.07971
$GD_{Trimmed}$	8.30072	1.52316	4.07220	0.05915
Wisorized Mean	9.13209	1.65190	4.53569	0.12788
GD_{Winsor}	8.77492	1.59667	4.34010	0.09893

Table 3: In this table the values of location for a log-normal distribution of parameters $\overline{x} = 2$ and $s = 1$ and 30 % of Gaussian noise are reported. The level of noise is expressed as percentage of the mean of the distributions. In the columns we have the location values when the distributions are composed with 20, 50, 100, and 1000 points.

	20 points Location $\overline{x}$	50 points Location $\overline{x}$	100 points Location $\overline{x}$	1000 points Location $\overline{x}$
Mode	8,5394	6,5722	5,7292	3,75215
Mean	12.24879	12.2451	12.19511	12.20451
GD_{Mean}	10.78797	10.79425	10.75084	10.80813
Trimmed Mean	8.9453	8.9077	8.78285	8.75854
$GD_{Trimmed}$	8.68679	8.63359	8.48676	8.44375
Wisorized Mean	9.40048	9.43643	9.3558	9.36122
GD_{Winsor}	9.03414	9.00968	8.89143	8.86864

Table 4: Left: case of log normal distribution sampled 20 times and 30 % of Gaussian noise. Right: case of log normal distribution sampled 100 times and 30 % of Gaussian noise. The level of noise is expressed as percentage of the mean of the distributions. The parameters of the log normal are $\overline{x} = 2$ and $s = 1$.

Methods	Error Reduction $\overline{x}$ [%]	Methods	Error Reduction $\overline{x}$ [%]
GD_{Mean}	20.09	GD_{Mean}	18.73
$GD_{Trimmed}$	0.99	$GD_{Trimmed}$	0.54
GD_{Winsor}	5.14	GD_{Winsor}	5.62

Table 5: This table reports the values of scale for a log normal distribution of parameters $\overline{x} = 2$ and $s = 1$ and 30 % of Gaussian noise. The ideal value to obtain is 15.97. The level of noise is expressed as percentage of the mean of the distributions. In the columns we have the scale values when the distributions are composed with 20, 50, 100, and 1000 points.

	20 points Scale s	50 points Scale s	100 points Scale s	1000 points Scale s
Mean	14.59655	15.23205	15.68975	16.38328
GD_{Mean}	14.67788	15.30456	15.75851	16.44366
Trimmed Mean	1.53488	0.90977	0.64213	0.20385
$GD_{Trimmed}$	2.14418	1.28424	0.88579	0.27558
Winsorized Mean	5.63982	5.73328	5.74412	5.78392
GD_{Winsor}	5.67598	5.75673	5.76635	5.80524

Table 6: Relations to calculate the confidence intervals for the hypothesis tests.

	Confidence Intervals	
	Low Limit	Upper Limit
Classic	$(\mu_{clas,1} - \mu_{clas,2}) - \Delta$	$(\mu_{clas,1} - \mu_{clas,2}) + \Delta$
Trimmed	$(\mu_{trim,1} - \mu_{trim,2}) - \Delta_t$	$(\mu_{trim,1} - \mu_{trim,2}) + \Delta_t$
Winsorized	$(\mu_{wins,1} - \mu_{wins,2}) - \Delta_w$	$(\mu_{wins,1} - \mu_{wins,2}) + \Delta_w$

Table 7: The values of Type I errors for a log normal distribution of parameters $\overline{x} = 2$ and $s = 1$ and 30 % of Gaussian noise. The level of noise is expressed as percentage of the mean of the distributions. In the columns we have the percentage of improvement of the Type I errors values through the GDGM methods compared to conventional methods. The distributions are composed with 20, 50, 100, and 1000 points. The minus sign indicates an improvement with respect to the corresponding technique calculated without the GDGM.

	20 points Type I errors [%]	50 points Type I errors [%]	100 points Type I errors [%]	1000 points Type I errors [%]
GD_{Mean}	-56.0	-71.8	-67.6	-63.2
$GD_{Trimmed}$	2.24	-10.6	-4.3	-8.7
GD_{Winsor}	-12.1	-21.5	-16.1	-17.8

Table 8: The values of power for a log normal distribution of parameters $\overline{x} = 2$ and $s = 1$ and 30 % of Gaussian noise. The level of noise is expressed as percentage of the mean of the distributions. In the columns we have reported the values of the Power obtained with the GDGM methods compared to conventional methods. The distributions are composed with 20, 50, 100, and 1000 points.

	20 points Power	50 points Power	100 points Power	1000 points Power
Mean	0.015	0.024	0.023	0.025
GD_{Mean}	0.021	0.038	0.045	0.285
Trimmed Mean	0.164	0.236	0.307	0.885
$GD_{Trimmed}$	0.194	0.288	0.335	0.957
Winsorized Mean	0.159	0.204	0.256	0.722
GD_{Winsor}	0.182	0.278	0.351	0.957

Table 9: The estimates of location or the case of a lognormal distribution with heteroscedastic noise. The results have been obtained generating 1000 points.

	No Noise	Noise
Mode	2.576	2.731
Mean	3.250	3.248
GD_{Mean}	2.765	3.257
Trimmed Mean	3.052	3.158
$GD_{Trimmed}$	2.079	2.372
Winsorized Mean	3.100	3.216
GD_{Winsor}	2.249	2.646

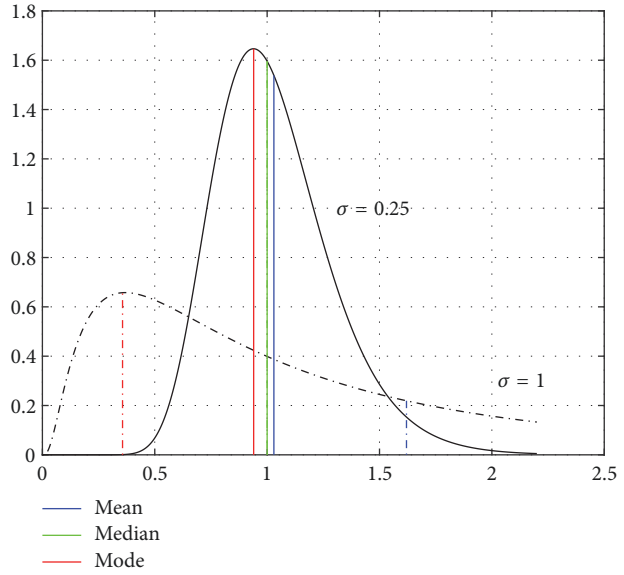


Figure 3: Log normal pdf. Comparison of mean, median, and mode of two log-normal distributions with different skewness.

Table 10: Log-normal distributions with 30% noise level.

	20 points Location $\overline{x}$	50 points Location $\overline{x}$	100 points Location $\overline{x}$	1000 points Location $\overline{x}$
Mode	8.5394	6.5722	5.7292	3.75215
Mean	12.24879	12.2451	12.19511	12.20451
GD_{Mean}	10.78797	10.79425	10.75084	10.80813
Trimmed Mean	8.9453	8.9077	8.78285	8.75854
$GD_{Trimmed}$	8.68679	8.63359	8.48676	8.44375
Winsorized Mean	9.40048	9.43643	9.3558	9.36122
GD_{Winsor}	9.03414	9.00968	8.89143	8.86864

Table 11: Log-normal distributions with 30% noise level.

20 points Scale s	50 points Scale s	100 points Scale s	1000 points Scale s
---	---	---	---
14.59655	15.23200	15.68975	16.38328
14.67788	15.30456	15.75851	16.44366
1.53488	0.90977	0.64213	0.20385
2.14418	1.28424	0.88579	0.27558
5.63982	5.73328	5.74412	5.78392
5.67598	5.75673	5.76635	5.80524

Table 12: Log-normal distributions with 30% noise level.

	20 points Type I errors	50 points Type I errors	100 points Type I errors	1000 points Type I errors
Mean	0.038	0.036	0.047	0.041
GD_{Mean}	0.009	0.018	0.016	0.014
Trimmed Mean	0.193	0.210	0.198	0.194
$GD_{Trimmed}$	0.179	0.187	0.180	0.167
Winsorized Mean	0.231	0.248	0.246	0.249
GD_{Winsor}	0.197	0.211	0.198	0.202

of samples. For most statistical applications, 5 to 25 percent of the ends are discarded; the 25% trimmed mean (when the lowest 25% and the highest 25% are discarded) is known as the interquartile mean. The trimmed mean is obviously more robust to the presence of outliers than the traditional mean. Similar considerations apply to the alternative methods to calculate robust measure of location, such as the winsorized mean discussed in the paper of Huber et al. (2009).

5.2. Application of the Geodesic Distance on Gaussian Manifolds to Measures of Location. The trimmed mean assumes that the data obtained are perfectly sampled from a distribution and they are not affected by noise. As mentioned, this is not realistic in the experimental investigations of complex systems, whose data are typically the results of complicated measurements always affected by a certain level of noise. The noise can often be modelled by a normal distribution of zero mean. The variance of the noise can often be determined experimentally but if that is not the case, the method proposed in this paper provides also a quite robust estimate of this important quantity (see Section 6.2).

To attack this second complication posed by the measurement noise, a good starting point is the observation that the mean can be considered the point having the minimum distance from the available data. This point can be calculated on the basis of the GDGM, which is an appropriate metric once the experimental values are affected by Gaussian noise. So instead of implicitly adopting the Euclidean metric to determine the trimmed and winsorized means, the more appropriate GDGM is used, which can properly take into account additive Gaussian noise.

The approach briefly described for the trimmed mean can be adopted for the other robust statistical techniques considered to determine the central tendency of pdfs. An appropriate metric, the Geodesic Distance on Gaussian Manifolds, is applied to the data already manipulated with the most appropriate robust statistical methods. So the robust statistical methods remedy the issue that the sampled distribution is not a Gaussian and the GDGM handles in a principled way the additional uncertainties due to the fact that the data is affected by noise. The details will be discussed in the next sections but, as far as the estimates of location

Table 13: Log-normal distributions with 30% noise level.

	20 points	50 points	100 points	1000 points
	Power	Power	Power	Power
Mean	0.015	0.017	0.014	0.018
GD_{Mean}	0.025	0.030	0.043	0.289
Trimmed Mean	0.183	0.210	0.300	0.888
$GD_{Trimmed}$	0.201	0.256	0.348	0.957
Winsorized Mean	0.163	0.192	0.256	0.722
GD_{Winsor}	0.184	0.261	0.367	0.968

Table 14: Log-normal with 20 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	20.09
$GD_{Trimmed}$	0.99
GD_{Winsor}	5.14

Table 15: Log-normal with 50 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	17.96
$GD_{Trimmed}$	-0.03
GD_{Winsor}	4.99

Table 16: Log-normal with 100 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	18.73
$GD_{Trimmed}$	0.54
GD_{Winsor}	5.62

Table 17: Log-normal with 1000 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	15.08
$GD_{Trimmed}$	1.38
GD_{Winsor}	6.62

are concerned, Tables 2 and 3 report a comparison between the traditional and robust estimators using the Euclidean distance and the GDGM for the case of data sampled from a log normal distribution. For the case with noise, the GDGM always improves the estimates of central tendency, by providing values closer to the mode compared to the robust statistics indicators. As already mentioned, all the results reported either in the main text or in Appendixes have been obtained by mediating over 1000 independent realisations.

6. Measures of Location

To show the potential of the proposed method, the GDGM has been applied first to the most common robust measures of location. In this section, it is assumed that the variance of the Gaussian noise has been already determined experimentally and it is known. In the next subsection an extension of the methodology is introduced to determine the variance of the noise directly from the data, without any a priori information.

6.1. Measures of Location in the Case the Variance of the Noise Is Known. The most common robust statistical measures of central tendency are the trimmed mean and the winsorized mean. The trimmed mean has already been introduced in the previous sections. The winsorized mean is a variant of the trimmed means, which involves the calculation of the mean after replacing given parts of the available samples, at the high and low end, with the most extreme remaining values. To summarise, the traditional mean, the trimmed and winsorized means are calculated according to the following formulas, in which $f(x_i)$ indicate the values sampled from the various pdfs.

The traditional mean is defined as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} f\left(x_i\right) \tag{10}$$

where N is the number of available samples. The trimmed mean is defined as

$$\mu_t = \frac{1}{N}\sum_{i=g+1}^{N-g} f\left(x_i\right) \tag{11}$$

where g corresponds to the number of trimmed points. The winsorized mean is defined as

$$\mu_w = \frac{1}{N}\sum_{i=1}^{N} f_w\left(x_i\right) \tag{12}$$

where

$$f_w\left(x_i\right) = \begin{cases} f\left(x_{g+1}\right) & if\ f\left(x_i\right) \le f\left(x_{g+1}\right) \\ f\left(x_i\right) & if\ f\left(x_{g+1}\right) < f\left(x_i\right) < f\left(x_{N-g}\right) \\ f\left(x_{N-g}\right) & if\ f\left(x_i\right) \ge f\left(x_{N-g}\right) \end{cases} \tag{13}$$

It is worth mentioning that the results reported in the rest of the paper have been obtained for symmetric versions of trimming and winsorization but the proposed alternative

Table 18: Log-normal distributions with 50% noise level.

	20 points Location $\overline{x}$	50 points Location $\overline{x}$	100 points Location $\overline{x}$	1000 points Location $\overline{x}$
Mode	8.13208	6.55561	5.54906	3.66741
Mean	12.29718	12.31377	12.16425	12.18874
GD_{Mean}	10.99257	10.92512	10.84016	10.90433
Trimmed Mean	9.65059	9.37941	9.29027	9.24034
$GD_{Trimmed}$	8.67604	8.43508	8.32643	8.30944
Winsorized Mean	9.9673	9.80449	9.71431	9.67959
GD_{Winsor}	9.07352	8.92301	8.85649	8.83833

Table 19: Log-normal distributions with 50% noise level.

20 points Scale s	50 points Scale s	100 points Scale s	1000 points Scale s
---	---	---	---
14.94729	16.17283	16.17655	16.94800
15.02021	16.2375	16.23344	16.99725
1.85818	1.09043	0.7677	0.24331
2.60785	1.54737	1.0612	0.32813
6.5626	6.68409	6.65622	6.69046
6.72474	6.77077	6.72129	6.74397

Table 20: Log-normal distributions with 50% noise level.

	20 points Type I errors	50 points Type I errors	100 points Type I errors	1000 points Type I errors
Mean	0.022	0.037	0.025	0.038
GD_{Mean}	0.009	0.009	0.007	0.011
Trimmed Mean	0.135	0.153	0.159	0.168
$GD_{Trimmed}$	0.097	0.108	0.102	0.107
Winsorized Mean	0.156	0.196	0.199	0.193
GD_{Winsor}	0.098	0.122	0.133	0.130

geodesic distance could be applied equally well to asymmetric versions of these measures.

As already briefly discussed, to apply the GDGM, which is a distance, the mean is considered the point with the minimum distance to the available samples. To this end, the data used by the robust techniques, trimmed or winsorized, are used to calculate, with an iterative process, the value with the minimum GDGM. The value of the mean is scanned over a reasonable range using the traditional mean as the first guess. The value which minimises the mean calculated with (2) is then chosen. Of course the only precaution to take is that the scan must cover a sufficiently wide range to include this minimum, but such a condition is easy to implement. To perform the iteration, the expression for the GDGM is (3), which depends on two quantities μ and σ. As mentioned, the standard deviation of the noise is assumed known or can be found with the technique described in Section 6.2. For the sample mean μ, since the noise is assumed Gaussian, the measured value is taken as the most probable value and therefore μ is identified with the experimental measured value. A series of numerical tests and theoretical considerations support this choice as can be seen in Verdolaeege [12].

In more detail, indicating with $\overline{f(x)}$ either the traditional sample mean, the trimmed mean, or the winsorized mean (depending of the context) of the available data, sampled from the pdf $f(x)$, the iterative process consists of finding the p_{mean} in the expression $(1 + p_{mean})\bullet\overline{f(x)}$ which minimises the GDGM. In this last expression p_{mean} is the parameter scanned to obtain the measure of location $\overline{x}_{model}$ by minimising the distance to the available experimental points. As mentioned earlier, for the various parameters necessary to compute the GDGM, the following values are chosen:

$$s_{data} = p_{noise}\bullet\overline{f\,(x)} \tag{14}$$

$$\overline{x}_{model} = \left(1 + p_{mean}\right) \bullet \overline{f\,(x)} \tag{15}$$

$$s_{model} = \sqrt{\frac{\sum_{i=1}^{N}\left(f\,(x_i) - \left(1 + p_{mean}\right)\bullet\overline{f\,(x)}\right)^2}{N-1}} \tag{16}$$

Table 21: Log-normal distributions with 50% noise level.

	20 points	50 points	100 points	1000 points
	Power	Power	Power	Power
Mean	0.003	0.005	0.005	0.003
GD_{Mean}	0.011	0.015	0.016	0.093
Trimmed Mean	0.106	0.109	0.144	0.512
$GD_{Trimmed}$	0.038	0.026	0.044	0.310
Winsorized Mean	0.104	0.103	0.117	0.352
GD_{Winsor}	0.053	0.058	0.070	0.351

Table 22: Log-normal with 20 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	14.28
$GD_{Trimmed}$	6.45
GD_{Winsor}	7.24

Table 23: Log-normal with 50 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	17.54
$GD_{Trimmed}$	6.99
GD_{Winsor}	7.72

Table 24: Log-normal with 100 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	15.78
$GD_{Trimmed}$	10.60
GD_{Winsor}	10.07

Table 25: Log-normal with 1000 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	11.88
$GD_{Trimmed}$	10.00
GD_{Winsor}	8.75

In these relations, p_{noise} is the percentage of Gaussian noise added to the individual samples $f(x)$ and the subscript *model* indicates the type of mean (traditional, trimmed, or winsorized).

To prove the potential of the proposed method in solving realistic problems, a series of numerical tests have been performed. The data have been sampled from the pdfs reviewed in Section 4. Various levels of Gaussian noise have been added to the individual points. A series of realisations, up to 1000, have been generated for each case and the results reported in Tables 4–9 are the averages over these realisations. The results of the systematic tests for the measures of location and scale are reported in Appendix A. In the rest of this section only the example of sampling from a log normal pdf is provided. In order to condense the information, the percentage variation between the carious estimates is reported in the main text (all the details are in Appendix A). Therefore in Tables 4–9 quantities of the form (vertical bars indicate absolute values) are reported

$$\left[\frac{\sum_{i=1}^{1000}\left(\left|\left(\overline{x}_{GD,i}-\overline{x}_{robust,i}\right)\right|/\overline{x}_{robust,i}\right)}{\#realisations}\right]100 \tag{17}$$

where the estimates using the GDGM are compared to the ones of the various robust indicators. As mentioned in the previous sections, the robust techniques are meant to provide a better estimate of central density, compared to traditional techniques; in this context this means that they tend to give results closer to the mode of the distributions from which the data are sampled (with the term mode we indicate the value where the pdf presents the main peak). The estimators using the GDGM are considered to improve the measures of location if they provide values closer to the mode of the sampled pdf compared to the robust method using the Euclidean distance.

For the case of the log normal distribution, the results for 20 and 100 samples are reported in Table 4, where the percentage improvement with respect to the traditional robust indicators is reported.

From this table, it emerges very clearly that, even at a level of noise of 30%, which is typical of many applications, the improvement provided by the GDGM is significant. Indeed the quantities calculated with the help of the GDGM are closer to the peak of the log normal pdf compared to the results obtained with the trimmed and winsorized means. In Appendix A, the results of a series of systematic tests are reported, showing how the GDGM always improves the estimates of location, over a wide range of samples and noise levels. As can be seen in the Tables in Appendix A, the GDGM allows outperforming the various robust techniques also for a quite high number of points sampled from the distribution (in the order of thousands). Appendix A also shows how this performance is not limited to the case of the log normal distribution but is equally appreciable for all the other pdfs tested.

6.2. Measures of Location in the Case the Variance of the Noise Is Not Known. In the investigation of complex systems, it is possible that the level of Gaussian noise can not be precisely quantified. Typically some experimental evidence is available but sometimes the uncertainties on the level of additive noise can be substantial. Another advantage of the GDGM is that it allows determining the level of normal noise,

Table 26: Exponential distributions with 30% noise level.

	20 points Location $\overline{x}$	50 points Location $\overline{x}$	100 points Location $\overline{x}$	1000 points Location $\overline{x}$
Mode	1.08682	0.72509	0.55488	0.24546
Mean	2.00326	2.00368	2.00076	1.99978
GD_{Mean}	1.83384	1.82025	1.81152	1.80861
Trimmed Mean	1.64069	1.61470	1.59887	1.59382
$GD_{Trimmed}$	1.60141	1.56638	1.54532	1.53794
Winsorized Mean	1.71454	1.70998	1.70169	1.70036
GD_{Winsor}	1.65804	1.63565	1.61989	1.61463

Table 27: Exponential distributions with 30% noise level.

20 points Scale s	50 points Scale s	100 points Scale s	1000 points Scale s
---	---	---	---
2.00977	2.04677	2.06789	2.08265
2.01853	2.05554	2.07682	2.09144
0.30212	0.17907	0.12788	0.04047
0.40574	0.24667	0.17285	0.05368
1.05783	1.08824	1.10303	1.10650
1.06356	1.09219	1.10667	1.10989

in particular its sigma, directly from the data. This can be achieved by scanning the sigma of the noise s_{GDGM} in the GDGM and by repeating the procedure for the identification of the mean described in the previous subsection. In the scan, the estimated mean remains constant for values of the assumed s_{GDGM} lower than the actual sigma of the noise. The mean then starts decreasing drastically when the assume s_{GDGM} is higher than the actual sigma of the noise. The inflection point is a very good estimate of the actual sigma of the additive Gaussian noise.

The approach just described is illustrated graphically in Figure 4 for the case of the log normal distribution. Data have been sampled from the pdf and then additive normal noise of zero mean and a sigma of 20% of the mean has been added. The mean of the data has then been calculated with the GDGM for a wide range of s_{GDGM}. As can be seen from the Figure, the inflection point in the mean corresponds very well to the added level of noise.

The proposed procedure has been verified for all the pdfs used in the paper and it has typically provided a reasonable estimate of the noise sigma.

The capability to derive information about the level of noise affecting the available data is of course a significant added value of the proposed technique, which can have very significant practical applications in the experimental investigations of complex systems.

7. Measures of Scale

The robust statistical methods developed in the last decades allow improving not only the estimates of location but also those of scale. The scale measures tested in this paper are reported in the following (see [5]). As a reference the classic standard deviation is defined as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} \left[f\left(x_i\right) - \mu \right]^2}{n-1}} \tag{18}$$

In the case the location is calculated with the trimmed mean, the standard deviation σ_t is defined as

$$\sigma_t = \sqrt{\frac{1}{n\left(n-1\right)\left(1-2\gamma\right)^2} \sum_{i=1}^{n} \left(f_t\left(x_i\right) - \mu_t \right)^2} \tag{19}$$

where n, γ, $f_t(x_i)$, and μ_t are the number of points sampled from the pdf, the percentage of trimming, the data of the trimmed, and the trimmed mean, respectively.

A similar definition applies to the standard deviation of the winsorized mean:

$$\sigma_w = \sqrt{\frac{1}{\left(n-1\right)} \sum_{i=1}^{n} \left(f_w\left(x_i\right) - \mu_w \right)^2} \tag{20}$$

To apply the GDGM to the standard deviations previously defined, the following formula has been applied:

$$\sigma_{GD} = \sqrt{\frac{\sum_{i=1}^{n} \left(f\left(x_i\right) - \left(1 + p_{mean}\right) \bullet \overline{f\left(x\right)} \right)^2}{n-1}} \tag{21}$$

where $\overline{f(x)}$ indicates either the traditional mean, the trimmed mean, or the winsorized mean of the available data.

The improvement in the determination of the scale using the GDGM is exemplified in Table 5 for the case of the log

Table 28: Exponential distributions with 30% noise level.

	20 points Type I errors	50 points Type I errors	100 points Type I errors	1000 points Type I errors
Mean	0.041	0.045	0.045	0.042
GD_{Mean}	0.032	0.027	0.029	0.018
Trimmed Mean	0.186	0.211	0.229	0.224
$GD_{Trimmed}$	0.183	0.215	0.219	0.221
Winsorized Mean	0.206	0.221	0.253	0.240
GD_{Winsor}	0.189	0.216	0.231	0.226

Table 29: Exponential distributions with 30% noise level.

	20 points Power	50 points Power	100 points Power	1000 points Power
Mean	0.041	0.048	0.050	0.046
GD_{Mean}	0.061	0.080	0.119	0.634
Trimmed Mean	0.292	0.434	0.589	0.999
$GD_{Trimmed}$	0.316	0.477	0.667	1.000
Winsorized Mean	0.262	0.351	0.454	0.982
GD_{Winsor}	0.287	0.420	0.597	0.998

Table 30: Exponential with 20 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	5.80
$GD_{Trimmed}$	-0.71
GD_{Winsor}	2.55

Table 31: Exponential with 50 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	5.00
$GD_{Trimmed}$	-0.26
GD_{Winsor}	2.14

Table 32: Exponential with 100 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	7.64
$GD_{Trimmed}$	-0.16
GD_{Winsor}	3.60

Table 33: Exponential with 1000 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	7.04
$GD_{Trimmed}$	-0.24
GD_{Winsor}	3.23

normal distribution. Again the impact on the other pdfs is reported in Appendixes. The results are very similar to the ones obtained by the other robust techniques, even if they are systematically slightly better. The advantages of the GDGM will become more evident in the case of hypothesis testing, described in the next section, where it will be shown how adopting the GDGM values of location and scale reduces significantly the Type I errors and improves the power.

8. Hypothesis Testing

In this section, the problems of computing confidence intervals and testing hypotheses are considered. The population variance is to be estimated from the sample variance and the standard deviation of the noise is considered known, either experimentally or by applying the method described in Section 6.2. The null hypothesis is as usual the one which assumes that the measure of location has a certain value: H_0: $\mu = \mu_0$. The alternative hypothesis is therefore $H_1 : \mu \neq \mu_0$. In the case of the classic significance tests, the null hypothesis is rejected at the confidence level $\alpha/2$ if

$$\begin{aligned} T &> t_{1-\alpha/2,n-1} \\ or\ T &< t_{\alpha/2,n-1} \\ H_1 &: \mu \neq \mu_0 \end{aligned} \tag{22}$$

where T is the student's T distribution and n the number of degrees of freedom. Robust statistical methods have been developed to test hypotheses when the usual assumptions of normal distribution of the sampled pdf and homoscedasticity are not verified. They are based on the robust estimators of location and scale introduced in the previous sections. An exhaustive treatment of these techniques can be found in Wilcox [11]. For the various robust indicators the null hypothesis can be rejected at the confidence level $\alpha/2$ if the conditions of the following inequalities are satisfied.

Table 34: Exponential distributions with 50% noise level.

	20 points Location $\bar{x}$	50 points Location $\bar{x}$	100 points Location $\bar{x}$	1000 points Location $\bar{x}$
Mode	1.06174	0.71622	0.56485	0.24778
Mean	1.99158	1.98960	2.00813	2.00242
GD_{Mean}	1.83644	1.82000	1.83880	1.82967
Trimmed Mean	1.69601	1.66348	1.67930	1.66705
$GD_{Trimmed}$	1.52821	1.49809	1.51042	1.50207
Winsorized Mean	1.75591	1.73586	1.75454	1.74685
GD_{Winsor}	1.59822	1.58391	1.60236	1.59234

Table 35: Exponential distributions with 50% noise level.

20 points Scale s	50 points Scale s	100 points Scale s	1000 points Scale s
---	---	---	---
2.12517	2.19360	2.22521	2.23517
2.13395	2.20117	2.23215	2.24189
0.33408	0.19687	0.14269	0.04497
0.46178	0.27742	0.19473	0.06013
1.15236	1.19450	1.21651	1.21824
1.18160	1.20810	1.22758	1.22814

In the case of the trimmed location estimate, according to the Yuen method described in Wilcox [5], the inequalities to evaluate are

$$T_{trim} = \frac{\mu_{trim,1} - \mu_{trim,2}}{\sigma_{wins,1-2}} > t_{1-\alpha/2,gdl}$$
$$or\ T_{trim} < t_{\alpha/2,gdl} \tag{23}$$
$$with\ gdl = \frac{(d_1 + d_2)^2}{d_1^2/(h_1 - 1) + d_2^2/(h_2 - 1)}$$

where

$$d_1 = \frac{(n_1 - 1)\,\sigma_{wins,1}^2}{h_1\,(h_1 - 1)};$$
$$d_2 = \frac{(n_2 - 1)\,\sigma_{wins,2}^2}{h_2\,(h_2 - 1)} \tag{24}$$
$$with\ h_i = n_i - 2g_i$$
$$\sigma_{wins,1-2} = \sqrt{d_1 + d_2}$$

In the case of the winsorized mean, the student test is

$$T_{wins} = \frac{\mu_{wins,1} - \mu_{wins,2}}{\sigma_{wins,1-2}} > t_{1-\alpha/2,gdl}$$
$$or\ T_{wins} < t_{\alpha/2,gdl} \tag{25}$$
$$with\ gdl = \frac{(d_1 + d_2)^2}{d_1^2/(h_1 - 1) + d_2^2/(h_2 - 1)}$$

where

$$d_1 = \frac{(n_1 - 1)\,\sigma_{wins,1}^2}{h_1\,(h_1 - 1)};$$
$$d_2 = \frac{(n_2 - 1)\,\sigma_{wins,2}^2}{h_2\,(h_2 - 1)} \tag{26}$$
$$with\ h_i = n_i - 1$$
$$\sigma_{wins,1-2} = \sqrt{d_1 + d_2}$$

The confidence intervals for these estimates can be derived from the following relations:

$$\Delta = t_{1-\alpha/2,gdl}\frac{\sigma_{clas,1-2}}{\sqrt{1/n_1 + 1/n_2}}$$
$$\Delta_{med} = t_{1-\alpha/2,gdl}\frac{\sigma_{med,1-2}}{\sqrt{1/n_1 + 1/n_2}} \tag{27}$$
$$\Delta_{trimmed} = t_{1-\alpha/2,gdl} \bullet \sigma_{wins,1-2}$$
$$\Delta_{winsor} = t_{1-\alpha/2,gdl} \bullet \sigma_{wins,1-2}$$

The extremes of the confidence intervals have been derived using the relations of Table 6. In the previous formulas and in Table 6, the subscript *clas* indicates the values calculated with the traditional methods and the Euclidean distance.

In order to verify the potential of the method proposed in this paper to help coping with Gaussian noise, all the previous tests have been calculated using also the estimates of location and scale obtained with the GDGM. Two main types of test have been performed. First, the data have been sampled by

Table 36: Exponential distributions with 50% noise level.

	20 points Type I errors	50 points Type I errors	100 points Type I errors	1000 points Type I errors
Mean	0.024	0.024	0.030	0.029
GD_{Mean}	0.018	0.017	0.013	0.016
Trimmed Mean	0.141	0.173	0.156	0.157
$GD_{Trimmed}$	0.111	0.122	0.105	0.106
Winsorized Mean	0.158	0.196	0.176	0.178
GD_{Winsor}	0.127	0.152	0.126	0.122

Table 37: Exponential distributions with 50% noise level.

	20 points Power	50 points Power	100 points Power	1000 points Power
Mean	0.026	0.029	0.038	0.026
GD_{Mean}	0.046	0.081	0.088	0.559
Trimmed Mean	0.256	0.329	0.471	0.991
$GD_{Trimmed}$	0.191	0.299	0.450	0.998
Winsorized Mean	0.237	0.292	0.371	0.931
GD_{Winsor}	0.205	0.310	0.467	0.997

Table 38: Exponential with 20 points.

Methods	Reduction $\overline{x}$ dispersion μ [%]
GD_{Mean}	4.25
$GD_{Trimmed}$	8.72
GD_{Winsor}	7.35

Table 39: Exponential with 50 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	3.80
$GD_{Trimmed}$	11.22
GD_{Winsor}	7.27

Table 40: Exponential with 100 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	5.44
$GD_{Trimmed}$	11.66
GD_{Winsor}	7.98

Table 41: Exponential with 1000 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	5.06
$GD_{Trimmed}$	13.33
GD_{Winsor}	7.53

the same pdf and noise has been added. In this case, the objective is to verify the resilience of the various tests to Type I error, i.e., to the wrong rejection of the null hypothesis (see Table 7).

In the main body of the paper this analysis has been particularised for the log-normal distribution. Further examples are reported in Appendix A. Secondly, the data have been sampled from two different distributions before adding the noise. In this case, the objective consists of determining whether the GDGM can help in reducing Type II errors, i.e., the failure to reject a false null hypothesis (see Table 8). As an example, in the following the data have been sampled from a log normal and a normal distribution. This is one of the most difficult cases since one of the two pdfs is a Gaussian. The results of similar tests for the other pdfs are again reported in Appendix A.

Some representative results of the tests for the Type I errors are reported in Table 7, where a minus sign indicates the reduction of errors achieved when using the GDGM. The use of the GDGM improves significantly the situation by reducing the Type I errors even of 20% with respect to the robust statistics techniques.

The effect of the GDGM is even more significant on the power, as can be seen in Table 8, again for some representative tests. The increase in the Power with GDGM methods can indeed reach values of the order of 35% compared to conventional methods.

9. Nonnormal Distributions and Heteroscedasticity

This section of the paper is the meant to cover the combined effects of sampling from an asymmetric distribution in presence of heteroscedastic noise. As a reference case of particular importance, the log normal distribution is analysed in detail. The data are sampled from a log normal distribution. Gaussian noise has then been added to the sampled data. The noise has zero mean and a standard

Table 42: Contaminated χ^2 distributions with 30% noise level.

	20 points Location $\overline{x}$	50 points Location $\overline{x}$	100 points Location $\overline{x}$	1000 points Location $\overline{x}$
Mode	5.78549	3.95817	2.97873	2.79684
Mean	7.58377	7.54954	7.58573	7.59473
GD_{Mean}	6.22352	6.13670	6.11370	6.09931
Trimmed Mean	4.51128	4.40502	4.32975	4.31965
$GD_{Trimmed}$	4.37176	4.30535	4.22925	4.21665
Winsorized Mean	4.78413	4.61130	4.53624	4.52361
GD_{Winsor}	4.58346	4.45736	4.37613	4.36100

Table 43: Contaminated χ^2 distributions with 30% noise level.

20 points Scale s	50 points Scale s	100 points Scale s	1000 points Scale s
---	---	---	---
10.97391	11.29413	11.57125	11.68770
11.06586	11.3861	11.66667	11.78321
0.77077	0.43540	0.30710	0.09797
1.12487	0.61454	0.42045	0.13058
3.00377	2.75198	2.73383	2.73164
3.02638	2.7614	2.74088	2.73670

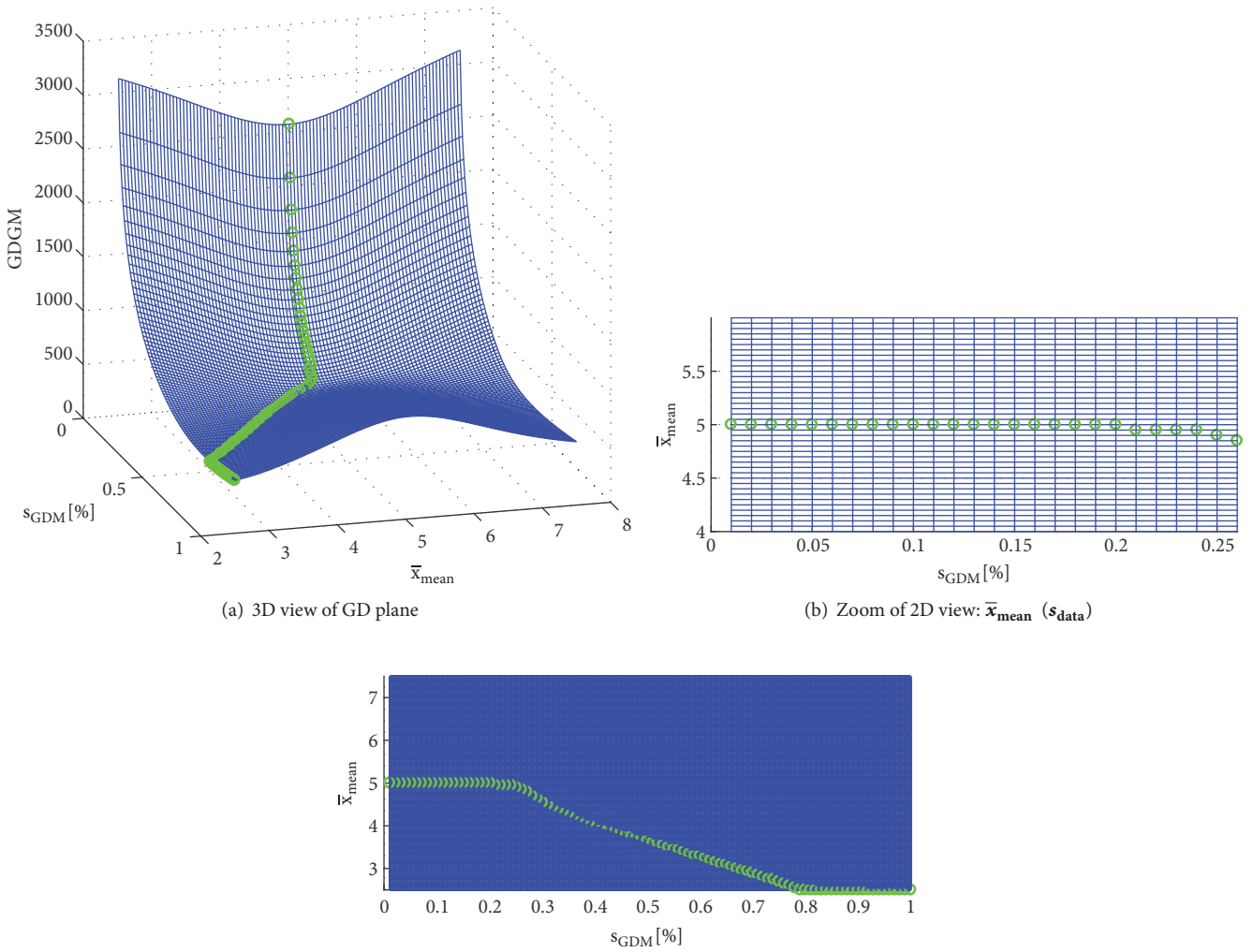
Table 44: Contaminated χ^2 distributions with 30% noise level.

	20 points Type I errors	50 points Type I errors	100 points Type I errors	1000 points Type I errors
Mean	0.035	0.040	0.048	0.038
GD_{Mean}	0.004	0.001	0.008	0.004
Trimmed Mean	0.130	0.148	0.172	0.169
$GD_{Trimmed}$	0.128	0.144	0.168	0.170
Winsorized Mean	0.202	0.214	0.230	0.236
GD_{Winsor}	0.134	0.166	0.190	0.179

Table 45: Contaminated χ^2 distributions with 30% noise level.

	20 points Power	50 points Power	100 points Power	1000 points Power
Mean	0.002	0.003	0.006	0.008
GD_{Mean}	0.008	0.044	0.087	0.930
Trimmed Mean	0.352	0.662	0.912	1.000
$GD_{Trimmed}$	0.391	0.686	0.923	1.000
Winsorized Mean	0.327	0.625	0.890	1.000
GD_{Winsor}	0.363	0.673	0.919	1.000

deviation equal to 50% of the value of the data. It is therefore a heteroscedastic noise quite common in practice, since often the uncertainties in the measurements are expressed as a percentage of their value. The effects of different types of noise are reported in Appendix B. Again the objective consists of estimating the mode, the most probable value of the data. The results of the various approaches are summarised in Table 9. From the table it is easily seen how the GDGM allows recovering values of central tendency much closer to the original distribution than the other indicators, even the robust ones. Such significant variations in the measures of location of course reverberate on the rest of the statistical quantities, from scale to hypothesis testing, again reducing significantly the errors committed by traditional robust indicators in the presence of Gaussian heteroscedastic noise.



(a) 3D view of GD plane

(b) Zoom of 2D view: $\overline{\boldsymbol{x}}_{\mathbf{mean}}$ $(\boldsymbol{s}_{\mathbf{data}})$

(c) 2D view: $\overline{\boldsymbol{x}}_{\mathbf{mean}}$ $(\boldsymbol{s}_{\mathbf{data}})$

Figure 4: Pictorial description of the procedure to find the sigma of the additive Gaussian noise in the data. The plots correspond to the case of data sampled from a log normal distribution and an additive normal noise of 20% of the log-normal mean. On the left a 3D view of the scan in the sigma used to calculate the GDGM and indicated s_{GDGM}. On the right end side, the zoomed projections on the s_{GDGM} $\overline{x}_{mean}$ plane are reported. As can be seen, the inflection point in the $\overline{x}_{mean}$ corresponds exactly to the 20% of additive Gaussian noise.

Table 46: Contaminated χ^2 with 20 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	24.94
$GD_{Trimmed}$	6.33
GD_{Winsor}	15.21

Table 47: Contaminated χ^2 with 50 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	27.91
$GD_{Trimmed}$	0.65
GD_{Winsor}	10.76

Table 48: Contaminated χ^2 with 100 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	28.79
$GD_{Trimmed}$	-0.397
GD_{Winsor}	7.094

Table 49: Contaminated χ^2 with 1000 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	29.14
$GD_{Trimmed}$	-0.34
GD_{Winsor}	6.83

Table 50: Contaminated χ^2 distributions with 50% noise level.

	20 points Location $\overline{x}$	50 points Location $\overline{x}$	100 points Location $\overline{x}$	1000 points Location $\overline{x}$
Mode	5.75158	3.96113	2.99485	2.85649
Mean	7.56302	7.61410	7.62280	7.59842
GD_{Mean}	6.22725	6.17673	6.16274	6.11750
Trimmed Mean	4.86596	4.68968	4.65874	4.60903
$GD_{Trimmed}$	4.44994	4.28575	4.26848	4.24064
Winsorized Mean	5.13555	4.90040	4.85056	4.79522
GD_{Winsor}	4.67670	4.51967	4.49452	4.45354

Table 51: Contaminated χ^2 distributions with 50% noise level.

20 points Scale s	50 points Scale s	100 points Scale s	1000 points Scale s
---	---	---	---
11.30261	11.90185	12.01272	12.08214
11.39161	11.99292	12.10357	12.17283
1.00320	0.57917	0.40993	0.12959
1.47045	0.83018	0.56349	0.17282
3.79591	3.64577	3.57781	3.55555
3.88056	3.68090	3.60212	3.57244

Table 52: Contaminated χ^2 distributions with 50% noise level.

	20 points Type I errors	50 points Type I errors	100 points Type I errors	1000 points Type I errors
Mean	0.025	0.040	0.021	0.043
GD_{Mean}	0.003	0.001	0.000	0.001
Trimmed Mean	0.118	0.123	0.104	0.111
$GD_{Trimmed}$	0.079	0.085	0.068	0.073
Winsorized Mean	0.171	0.182	0.162	0.166
GD_{Winsor}	0.088	0.118	0.097	0.094

Table 53: Contaminated χ^2 distributions with 50% noise level.

	20 points Power	50 points Power	100 points Power	1000 points Power
Mean	0.003	0.004	0.001	0.004
GD_{Mean}	0.010	0.034	0.074	0.940
Trimmed Mean	0.329	0.566	0.812	1.000
$GD_{Trimmed}$	0.339	0.659	0.874	1.000
Winsorized Mean	0.316	0.535	0.787	1.000
GD_{Winsor}	0.335	0.617	0.839	1.000

Table 54: Contaminated χ^2 with 20 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	22.14
$GD_{Trimmed}$	7.03
GD_{Winsor}	15.53

Table 55: Contaminated χ^2 with 50 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	25.48
$GD_{Trimmed}$	7.20
GD_{Winsor}	12.13

Table 56: Contaminated χ^2 with 100 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	26.55
$GD_{Trimmed}$	7.48
GD_{Winsor}	11.69

Table 57: Contaminated χ^2 with 1000 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	27.16
$GD_{Trimmed}$	6.66
GD_{Winsor}	10.30

Table 58: G-h distributions with 30% noise level.

	20 points Location $\overline{x}$	50 points Location $\overline{x}$	100 points Location $\overline{x}$	1000 points Location $\overline{x}$
Mode	0.38020	0.31542	0.28250	-0.01744
Mean	0.72107	0.74738	0.77821	0.79727
GD_{Mean}	0.45726	0.46189	0.48348	0.52147
Trimmed Mean	0.10548	0.07130	0.07681	0.06771
$GD_{Trimmed}$	0.08557	0.05222	0.05795	0.04828
Winsorized Mean	0.15617	0.12161	0.13170	0.12392
GD_{Winsor}	0.12475	0.09072	0.09795	0.08498

Table 59: G-h distributions with 30% noise level.

20 points Scale s	50 points Scale s	100 points Scale s	1000 points Scale s
---	---	---	---
3.87104	4.69402	5.32664	7.73014
3.88735	4.70586	5.33644	7.73585
0.22745	0.12645	0.08848	0.02742
0.31618	0.18211	0.12396	0.03750
0.84574	0.82853	0.82003	0.80183
0.84784	0.82991	0.82116	0.80283

Table 60: G-h distributions with 30% noise level.

	20 points Type I errors	50 points Type I errors	100 points Type I errors	1000 points Type I errors
Mean	0.037	0.041	0.031	0.049
GD_{Mean}	0.008	0.004	0.001	0.003
Trimmed Mean	0.217	0.220	0.209	0.228
$GD_{Trimmed}$	0.214	0.207	0.187	0.185
Winsorized Mean	0.276	0.296	0.273	0.307
GD_{Winsor}	0.233	0.237	0.221	0.217

Table 61: G-h distributions with 30% noise level.

	20 points Power	50 points Power	100 points Power	1000 points Power
Mean	0.006	0.007	0.001	0.006
GD_{Mean}	0.008	0.013	0.019	0.116
Trimmed Mean	0.246	0.358	0.476	0.927
$GD_{Trimmed}$	0.269	0.430	0.538	0.943
Winsorized Mean	0.246	0.358	0.476	0.927
GD_{Winsor}	0.261	0.383	0.502	0.932

Table 62: G-h with 20 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	39.60
$GD_{Trimmed}$	3.48
GD_{Winsor}	6.88

Table 63: G-h with 50 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	43.90
$GD_{Trimmed}$	3.52
GD_{Winsor}	9.95

Table 64: G-h with 100 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	42.65
$GD_{Trimmed}$	1.48
GD_{Winsor}	8.66

Table 65: G-h with 1000 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	37.42
$GD_{Trimmed}$	9.27
GD_{Winsor}	11.90

10. Conclusions

In the science of complex systems, the data are often the output of delicate measurements and therefore are typically affected by Gaussian additive noise. Since the data are also not necessarily sampled from normal pdfs, it is important to develop tools which can handle both these problems. The Geodesic Distance on Gaussian Manifolds is a principled way to address the issue of Gaussian noise. In this paper, it has been shown how the GDGM can improve the estimates of robust statistical methods, ranging from the evaluation of location and scale to hypothesis testing. The improvements obtained with GDGM for the estimates of location and scale are not negligible for all the non-Gaussian distributions tested. In the case of hypothesis testing, the advantages provided by the GDGM are quite substantial; in particular the power is significantly improved since the Type II errors can be reduced typically of more than 30% for all the pdf tested. Moreover, since the GDGM is a distance on a Gaussian manifold, it does not introduce unacceptable errors if the data are sampled from a normal pdf. A specific development of the technique allows also estimating the level of noise associated with the measurements, when this information cannot be determined experimentally.

In terms of future developments, it is planned to investigate whether the GDGM can improve also other robust techniques, in particular those belonging to the class of the so called M-estimators, which are considered better performing than the ones based on trimming [Huber et al. (2009)]. First preliminary tests indicate that there is no reason to expect that the impact of the GDGM will be less positive for this class of estimators. It is therefore considered more urgent to apply the GDGM also to the methods of robust regression, to assess whether progress can be achieved also for this class of problems. Recently new regression methods, indicated collectively as symbolic regression, have allowed relaxing the constraints of linear regression and obtain formula in nonpower law form as in the works of Murari et al. [13, 14] and in the paper of Peluso et al. [15]. The use of the GDGM seems particularly promising also in this context as it has been shown in Murari et al. [16, 17]. The proposed methodology could be therefore profitably be used in the field of Thermonuclear Fusion to help in the development of more robust scenarios [18–20].

Appendix

A.

A.1. Log-Normal Distribution. See Tables 10–25.

A.2. Exponential Distribution. See Tables 26–41.

A.3. Contaminated χ^2 Distribution. See Tables 42–57.

A.4. Contaminated G-H Distribution. See Tables 58–73.

B.

B.1. Log-Normal Distribution with Asymmetric Noise (Positive). See Table 74.

B.2. Log-Normal Distribution with Asymmetric Noise (Negative). See Table 75.

B.3. Log-Normal Distribution with Heteroscedastic Noise. See Table 76.

B.4. Log-Normal Distribution with Step Noise (Type 1). See Table 77.

B.5. Log-Normal Distribution with Step Noise (Type 2). See Table 78.

B.6. Log-Normal Distribution with Step Noise (Type 3). See Table 79.

B.7. Log-Normal Distribution with Step Noise (Type 4). See Table 80.

Data Availability

In order to obtain the numerical dataset used to carry out the analysis reported in the manuscript, please contact the corresponding author Michele Lungaroni.

Table 66: G-h distributions with 50% noise level.

	20 points Location$\overline{\boldsymbol{x}}$	50 points Location $\overline{\boldsymbol{x}}$	100 points Location $\overline{\boldsymbol{x}}$	1000 points Location $\overline{\boldsymbol{x}}$
Mode	1.03893	0.36309	0.31981	0.05504
Mean	1.01079	0.77843	0.77888	0.78485
GD_{Mean}	0.61100	0.47946	0.47803	0.51055
Trimmed Mean	0.12524	0.09354	0.07864	0.07433
$GD_{Trimmed}$	0.10583	0.07508	0.05955	0.05389
Winsorized Mean	0.16985	0.13961	0.13018	0.12525
GD_{Winsor}	0.13458	0.10507	0.09164	0.08040

Table 67: G-h distributions with 50% noise level.

20 points Scale $\boldsymbol{s}$	50 points Scale $\boldsymbol{s}$	100 points Scale $\boldsymbol{s}$	1000 points Scale $\boldsymbol{s}$
---	---	---	---
4.97806	4.86858	5.39029	7.62799
5.00121	4.88114	5.40026	7.6337
0.27053	0.13910	0.09547	0.02946
0.37237	0.19648	0.13250	0.03990
0.98774	0.88558	0.86995	0.84722
0.99131	0.88741	0.87148	0.84853

Table 68: G-h distributions with 50% noise level.

	20 points Type I errors	50 points Type I errors	100 points Type I errors	1000 points Type I errors
Mean	0.031	0.032	0.031	0.039
GD_{Mean}	0.008	0.007	0.004	0.001
Trimmed Mean	0.188	0.183	0.191	0.230
$GD_{Trimmed}$	0.183	0.171	0.174	0.166
Winsorized Mean	0.231	0.235	0.261	0.289
GD_{Winsor}	0.198	0.175	0.188	0.179

Table 69: G-h distributions with 50% noise level.

	20 points Power	50 points Power	100 points Power	1000 points Power
Mean	0.003	0.004	0.005	0.006
GD_{Mean}	0.010	0.013	0.027	0.148
Trimmed Mean	0.233	0.357	0.528	0.939
$GD_{Trimmed}$	0.255	0.385	0.556	0.949
Winsorized Mean	0.217	0.328	0.479	0.926
GD_{Winsor}	0.236	0.353	0.532	0.942

Table 70: G-h with 20 points.

Methods	Reduction $\overline{\boldsymbol{x}}$ dispersion [%]
GD_{Mean}	47.85
$GD_{Trimmed}$	-0.88
GD_{Winsor}	-2.41

Table 71: G-h with 50 points.

Methods	Reduction $\overline{\boldsymbol{x}}$ dispersion [%]
GD_{Mean}	42.90
$GD_{Trimmed}$	-0.01
GD_{Winsor}	6.39

Table 72: G-h with 100 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	44.90
$GD_{Trimmed}$	1.67
GD_{Winsor}	10.84

Table 73: G-h with 1000 points.

Methods	Reduction $\overline{x}$ dispersion [%]
GD_{Mean}	35.41
$GD_{Trimmed}$	7.96
GD_{Winsor}	15.53

Table 74: Log-normal distributions with asymmetric noise. The noise is composed of the sum of two Gaussians. The first Gaussian has zero mean and standard deviation equal to 10% of the maximum value of the data distribution $\overline{x}_1 = 0$, s_1 = 10% max(distribution); the second Gaussian has the mean equal $\overline{x}_2 = 2(s_1+s_2)$ and standard deviation s_2 = 20% max(distribution). The noise is composed of 1000 points; of these 80% of the points are sampled on the first Gaussian and the remaining on the second Gaussian. The noise has a positive mean because $\overline{x}_2 > 0$.

	No Noise	Noise
Mode	2.576	2.890
Mean	3.250	4.587
GD_{Mean}	2.765	4.101
Trimmed Mean	3.052	3.845
$GD_{Trimmed}$	2.079	2.950
Winsorized Mean	3.100	4.097
GD_{Winsor}	2.249	3.433

Table 75: Log-normal distributions with asymmetric noise. The noise is composed of the sum of two Gaussians. The first Gaussian has zero mean and standard deviation equal to 10% of the maximum value of the data distribution $\overline{x}_1 = 0$, s_1 = 10% max(distribution); the second Gaussian has the mean equal $\overline{x}_2 = -2(s_1+s_2)$ and standard deviation s_2 = 20% max(distribution). The noise is composed of 1000 points; of these 80% of the points are sampled on the first Gaussian and the remaining on the second Gaussian. The noise has a negative mean because $\overline{x}_2 < 0$.

	No Noise	Noise
Mode	2.576	2.731
Mean	3.250	3.248
GD_{Mean}	2.765	3.257
Trimmed Mean	3.052	3.158
$GD_{Trimmed}$	2.079	2.372
Winsorized Mean	3.100	3.216
GD_{Winsor}	2.249	2.646

Table 76: Log-normal distributions with heteroscedastic noise. The noise has zero mean and a standard deviation *s* equal to 50% of the value of the data. The noise is composed of 1000 points.

	No Noise	Noise
Mode	2.576	2.731
Mean	3.250	3.248
GD_{Mean}	2.765	3.257
Trimmed Mean	3.052	3.158
$GD_{Trimmed}$	2.079	2.372
Winsorized Mean	3.100	3.216
GD_{Winsor}	2.249	2.646

Table 77: Log-normal distributions with step noise (Type 1). The uniform noise is applied in the range from the minimum value to the maximum value of the distribution. The frequency of the samples decreases linearly, going to create a noise-shaped step. The noise is composed of 1000 points.

	No Noise	Noise
Mode	2.590	4.163
Mean	3.249	4.896
GD_{Mean}	2.765	3.869
Trimmed Mean	3.051	4.718
$GD_{Trimmed}$	2.078	3.112
Winsorized Mean	3.099	4.763
GD_{Winsor}	2.248	3.395

Table 78: Log-normal distributions with step noise (Type 2). The uniform noise is applied in the range from the minimum value to the maximum value of the distribution. The frequency of the samples increases linearly, going to create a noise-shaped step. The noise is composed of 1000 points.

	No Noise	Noise
Mode	2.590	6.197
Mean	3.249	6.376
GD_{Mean}	2.765	4.646
Trimmed Mean	3.051	6.307
$GD_{Trimmed}$	2.078	3.922
Winsorized Mean	3.099	6.319
GD_{Winsor}	2.248	4.126

Table 79: Log-normal distributions with step noise (Type 3). The uniform noise is applied in the range from the minus maximum value to the minimum value of the distribution. The frequency of the samples decreases linearly, going to create a noise-shaped step. The noise is composed of 1000 points.

	No Noise	Noise
Mode	2.590	-0.465
Mean	3.249	0.335
GD_{Mean}	3.120	0.245
Trimmed Mean	3.051	0.148
$GD_{Trimmed}$	2.539	0.126
Winsorized Mean	3.099	0.199
GD_{Winsor}	2.737	0.164

Table 80: Log-normal distributions with step noise (Type 4). The uniform noise is applied in the range from the minus maximum value to the minimum value of the distribution. The frequency of the samples increases linearly, going to create a noise-shaped step. The noise is composed of 1000 points.

	No Noise	Noise
Mode	2.589	2.023
Mean	3.249	2.054
GD_{Mean}	2.765	2.065
Trimmed Mean	3.051	2.020
$GD_{Trimmed}$	2.078	1.620
Winsorized Mean	3.099	2.019
GD_{Winsor}	2.248	1.799

Disclosure

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; and in the decision to publish the results.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] D. H. Freedman, *Wrong: Why Experts Keep Failing Us and How to Know When Not to Trust Them*, Little, Brown and Company, London, UK, 2010.

[2] J. P. A. Ioannidis, "Why most published research findings are false," *PLoS Medicine*, vol. 2, no. 8, article e124, 2005.

[3] S. S. Sawilowsky, "Nonparametric Tests of Interaction in Experimental Design," *Review of Educational Research*, vol. 60, no. 1, pp. 91–126, 1990.

[4] S. S. Sawilowsky and R. C. Blair, "A more realistic look at the robustness and Type II error properties of the t test to departures from population normality," *Psychological Bulletin*, vol. 111, no. 2, pp. 352–360, 1992.

[5] R. R. Wilcox, *Fundamentals of Modern Statistical Methods*, Springer, New York, NY, USA, 2001.

[6] R. R. Wilcox, *Applying Contemporary Statistical Techniques*, Academic Press, San Diego, CA, USA, 2003.

[7] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2009.

[8] IEC-ISO., *Guide to the Expression of Uncertainty in Measurement*, 1992.

[9] B. Cannas, A. Fanni, A. Murari, A. Pau, and G. Sias, "Automatic disruption classification based on manifold learning for real-time applications on JET," *Nuclear Fusion*, vol. 53, no. 9, Article ID 093023, 2014.

[10] A. Murari, P. Boutot, J. Vega et al., "Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions," *Nuclear Fusion*, vol. 53, no. 3, Article ID 033006, 2013.

[11] R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, San Diego, Calif, USA, 2nd edition, 2005.

[12] G. Verdoolaege, "A New Robust Regression Method Based on Minimization of Geodesic Distances on a Probabilistic Manifold: Application to Power Laws," *Entropy*, vol. 17, no. 12, pp. 4602–4626, 2015.

[13] A. Murari, E. Peluso, M. Gelfusa, I. Lupelli, M. Lungaroni, and P. Gaudio, "Symbolic regression via genetic programming for data driven derivation of confinement scaling laws without any assumption on their mathematical form," *Plasma Physics and Controlled Fusion*, vol. 57, no. 1, Article ID 014008, 2014.

[14] A. Murari, E. Peluso, M. Lungaroni, M. Gelfusa, and P. Gaudio, "Application of symbolic regression to the derivation of scaling laws for tokamak energy confinement time in terms of dimensionless quantities," *Nuclear Fusion*, vol. 56, no. 2, Article ID 026005, 2015.

[15] E. Peluso, A. Murari, M. Gelfusa, and P. Gaudio, "A statistical method for model extraction and model selection applied to the temperature scaling of the L-H transition," *Plasma Physics and Controlled Fusion*, vol. 56, no. 11, Article ID 114001, 2014.

[16] A. Murari, E. Peluso, M. Gelfusa, I. Lupelli, and P. Gaudio, "A new approach to the formulation and validation of scaling expressions for plasma confinement in tokamaks," *Nuclear Fusion*, vol. 55, no. 7, Article ID 073009, 2015.

[17] A. Murari, E. Peluso, M. Gelfusa, M. Lungaroni, and P. Gaudio, "How to handle error bars in symbolic regression for data mining in scientific applications," *Statistical Learning and Data Sciences*, vol. 9047, pp. 347–355, 2015.

[18] J. Ongena, P. Monier-Garbet, W. Suttrop et al., "Towards the realization on JET of an integrated H-mode scenario for ITER," *Nuclear Fusion*, vol. 44, no. 1, pp. 124–133, 2004.

[19] M. E. Puiatti, M. Mattioli, G. Telesca et al., "Radiation pattern and impurity transport in argon seeded ELMy H-mode discharges in JET," *Plasma Physics and Controlled Fusion*, vol. 44, no. 9, pp. 1863–1878, 2002.

[20] F. Romanelli and R. Kamendje, "Overview of JET results," *Nuclear Fusion*, vol. 49, no. 10, p. 104006, 2009.