Alternatives to OLS

When Noise Goes to Infinity and Beyond

Caleb Dame, Derek Smith, Cade Cardwell, Braden Stuart

In this paper, we use distributions with infinite variance to show that different estimation methods beyond OLS are able to provide unbiased and consistent estimators. We use Monte Carlo methods to generate data for five types of distributions with characteristics differing on variance, symmetry, and differentiability to test the limits of our models. We find that LAD and MLE are able to overcome infinite variance issues and produce consistent estimates that allow us to conduct inference, with some exceptions. These exceptions include LAD struggling with asymmetric distributions and both MLE and LAD suffering because their gradient descent solvers find non-optimal minimizers for various distributions.

I. INTRODUCTION

In this paper we explore various methods to estimate parameters of a linear relationship when the error term is distributed with infinite variance. Our goal is to show that different estimations methods, such as Least Absolute Deviations (LAD) and Maximum Likelihood Estimation (MLE), are often able to provide unbiased and consistent estimates where Ordinary Least Squares (OLS) fails to do so, permitting inference and hypothesis testing. We also seek to show the limits of LAD and MLE estimators by highlighting cases where their unique assumptions are not met and when one cannot rely on them for efficiency or unbiasedness when noise is distributed with infinite variance. This will be done via Monte Carlo approximations using various distributions to generate the noise in the data, each representing a different set of problems for estimation.

II. ESTIMATION METHODS

In evaluating distributions with infinite variances, we predict that OLS will fail to offer consistent and asymptotically normal estimates. Also, we expect LAD and MLE to perform well and give consistent and asymptotically normal estimates despite the infinite variance in the dataset when certain conditions are met.

Why would OLS fail to estimate infinite variance distributions?

Ordinary Least Squares (OLS), one of the most commonly used estimation methods, has unbiased estimates regardless of error distribution; however, implicit in OLS estimation as a method that permits inference is the assumption of finite noise variance. The coefficient of a linear model given by OLS is the vector $\beta^{OLS} = (X'X)^{-1}(X'Y)$. With the population expectation $E[X'\boldsymbol{\varepsilon}] = 0$, the Law of Large numbers guarantees that the sample mean $E[X'\boldsymbol{\varepsilon}]$ converges in probability to the population mean when the variance of the random variables are finite;

however, using OLS with noise distributed with infinite variance should still yield an unbiased estimate. Since E[X'**ε**] has infinite variance, the sample approximation does not quickly converge to the population mean, and the estimate is no longer consistent. Consistency is a key criteria for the Central Limit Theorem to operate, so the distribution of estimated parameters is no longer distributed normally; this results in standard errors that do not permit practicing inference. OLS is also particularly sensitive to outliers, which problem is exacerbated when infinite variance is introduced. As pictured in **Figure 1** (located in the appendix), the sample average of ***n*** Standard Cauchy observations produces 95% confidence intervals that are not robust to new observations since the sample mean does not converge as fast as the Central Limit Theorem would like.

Inference for M-estimators

M-estimators maximize an objective function that is a sample average of some function *m* of data and parameters. Both LAD and MLE are M-estimators and therefore require several assumptions for both identification and inference. Infinite variance distributions do not cause any of the assumptions for M-estimator identification break down. The M-estimator assumptions are included in the appendix for reference in **Note 1**. More importantly, we must rely on the following assumptions to show that $\hat{b}$is asymptotically normal in order to perform inference:

1. $\beta \in$ int (**B**)
2. $Q_n(b)$is twice continuously differentiable in a neighborhood of β
3. $\sqrt{n}\frac{\partial}{\partial b}Q_n(\beta) \rightarrow_d N(0, \Sigma)$
4. There is a H (b), continuous at β, such that
$$sup_{b\epsilon_{N(\beta)}} \left| \left| \frac{\partial^2}{\partial b \partial b'} Q_n(b) - H(b) \right| \right| \rightarrow_p 0$$
5. H (β) is nonsingular

When all of these assumptions hold, we can show:

$$\sqrt{n}(b - \beta) \rightarrow_d N(0, H^{-1}\Sigma H^{-1})$$

Assumption 2 will not hold for distributions that are not twice continuously differentiable, which will cause the M-estimators in our study to break down with distributions that will be discussed in the Data section.

Why use Least Absolute Deviation Estimators with infinite variance distributions?

While Ordinary Least Squares estimates are infamously vulnerable to outliers, Least Absolute Deviations (LAD) estimators are not nearly as sensitive to outliers and are therefore more "robust" than Least Squares Estimators. The Least Absolute Deviation Model is technically equivalent to performing a Quantile Regression for $\boldsymbol{\tau} = .5$, or in other words, the LAD estimator estimates the median effects along the linear model.

The model is much more robust than OLS because, unlike OLS, which estimates population average effects, LAD estimates population median effects, where medians are less sensitive to activity in the tails of distributions than are average. After meeting the initial assumptions for conducting inference with M-Estimators, we can easily construct Wald estimators to construct confidence intervals and do hypothesis testing after model estimation.

Why use Maximum Likelihood Estimators with infinite variance distributions?

MLE is not only consistent and asymptotically normal but, as its variance is identical to the Cramer-Rao Lower Bound, it is also asymptotically efficient. Therefore, it is the most efficient estimator possible, given that the noise distribution is known. If we can believe the strong assumptions of MLE, (and in this case we can, seeing that we generate the data ourselves,) we can assume that it will be better than OLS at predicting a distribution with an infinite variance because it is a more efficient estimator than OLS, and infinite variance noise does not violate any of the assumptions required in order to use MLE as an estimator.

General Expectations for LAD and MLE

Since the LAD estimates the population median, while the noise may have mean zero, the median of the noise may be non-zero. This would occur for any noise distribution that is asymmetric, causing the mean and the median to differ. For noise distributed identically throughout the data, as is the case for all the noise distributions in the simulations, the LAD estimate for any regression slopes should be consistent as would be equivalent for any quantile. The regression intercept would vary with each quantile estimate, shifting the line up or down the y-axis. So, in order to recover the true model, knowing the quantile regression parameter $\boldsymbol{\tau}$ that gives the same line as the population average would be necessary.

As mentioned above, MLE relies on the strong assumption of having correctly identified the true distribution of the model noise. We expect MLE to perform very well every time that the noise is correctly identified and solve for using its corresponding MLE equation that we maximize programmatically; however, when the noise is incorrectly attributed to a distribution, we will no longer have an efficient model since the solver will not have the variance associated with the Cramer-Rao Lower Bound.

We also predict that both MLE and LAD estimators will break down for distributions where $Q_n(b)$ is not twice continuously differentiable in a neighborhood of β. The Data section contains more in-depth descriptions of distributions that we believe will cause the M-estimators to break down by virtue of discontinuity or non-differentiability at the true model parameter.

Summary tables evaluating predicted model strengths and weaknesses are provided in the appendix (**Table 2** and **3**).

III. DATA

Noise Distributions

To test the performance expectations for various estimation methods under different noise distributions, we have generated homoskedastic noise from the following distributions: Standard Normal, Standard Cauchy, Q-Exponential, a mixture distribution of Standard Cauchy distributions, and a piecewise Cauchy-Normal distribution where the noise is distributed normally for positive values, and distributed Cauchy for negative values.

Since MLE methods are much simpler for functions defined everywhere and nowhere zero, we use a two-tailed Q-Exponential distribution and avoid distributions that diverge to infinity, such as the Pareto Distribution. While in reality, the Pareto distribution need never be evaluated at 0 (where it is undefined) floating point can often round small numbers to zero, resulting in a maximum immediately, and causing the MLE solution to be consistently wrong.

Data Generation

For the five noise generating distributions, we will perform Monte Carlo methods to create the following linear model for various noise:

$$\mathbf{y_i = \beta_1 x_i + \beta_0 + \varepsilon}$$

For the purposes of the paper we have chosen to estimate the equations with $\beta_1 = 1$ with $\beta_0 = 0$ throughout. The parameters will be estimated 20,000 times for each distribution of noise, using $\boldsymbol{n} = 1000$ observations. Each estimation will be computed via each estimation method for randomly generated noise values from each noise distribution, resulting in 600,000 sets of model parameters across the six estimation methods and the five noise distributions. (See **Figures 2** and **3** in the appendix for distribution pictures)

The Gaussian Distribution of noise will act as our control data. This distribution will show how OLS, LAD, and the various MLE solutions work under finite variance. Each of the others will have infinite variance and will vary in skewness, continuity, and number of fat tails in order to verify predictions of the performance of various estimators under different distributions.

The first infinite-variance distribution is the Cauchy distribution. The Cauchy distribution has a similar shape to a normal distribution and is symmetric, but has fat-tails that gives us our infinite variance to test the limits of our models. Next, like the Cauchy distribution, a Two-Tailed Q-Exponential is fat-tailed and symmetric; however, unlike the Cauchy distribution, the Two-Tailed Q-Exponential is not differentiable at zero, its mean. A mixture of Cauchy distributions will serve as the case of an asymmetric distribution, such that the mean and median are no longer identical. This distribution is a weighted sum of two Standard Cauchy distributions that have mean difference means with the entire distribution centered at their new mean. This distribution is differentiable and has infinite variance. The last distribution is a Split Cauchy-Normal noise distribution. This distribution is a piecewise distribution that is distributed Cauchy for values below zero, and Normal above zero. This distribution is not differentiable and is fat-tailed on only one side.

This data will allow us to test the limitations of MLE by changing the variance of the distribution at the mean. See **Table 1** for a summary of the properties of our distributions.

IV. METHOD

Estimation Methods and Process

For each instantiation of the linear model, the model will be estimated via each of the following ways:

- Ordinary Least Squares
- Least Absolute Deviation
- Maximum Likelihood Estimation (Cauchy, Q-Exponential, Cauchy-Mixture, Split Cauchy-Normal)

The Maximum Likelihood will be estimated each iteration for each of the Non-Gaussian noise distributions in order to compare performance of MLE when the distribution is incorrectly identified and estimated by minimizing the wrong product. Gaussian MLE Estimation is performed by maximizing the product of the probability of independent Gaussian random variables; since this mathematically simplifies to the same estimation that OLS performs, we will not estimate the Gaussian MLE.

Predictions

We predict that the LAD estimators will improperly estimate the regression constant for noise distributed as the Cauchy Mixture. While the noise of the Cauchy Mixture distribution has a mean zero, the median of the noise is non-zero. For the other distributions where the mean is the median, we predict that the LAD intercept estimate will be distributed normally and be unbiased. For all the distributions, we have homoscedasticity, so the slope estimate should be consistent and unbiased.

We expect MLE to perform very well every time that the noise is correctly identified and solved for using its corresponding MLE equation that we maximize programmatically. For

example, when we use a Cauchy noise distribution, we predict the corresponding MLE to be the most efficient. On the other hand, if one used the MLE estimate corresponding to the Q-exponential noise distribution, we expect larger variance. For mismatched MLEs that are both symmetric, we would also expect there to be no bias.

We predict that the Q-Exponential and the Split Cauchy-Normal noise distributions will cause the LAD and MLE estimators to have distributions that are not normal. In order to perform inference, both LAD and MLE estimators require the assumption that $Q_n(b)$ is twice continuously differentiable in a neighborhood of β. The LAD and MLE estimators may not have normal distributions because neither the Q-Exponential nor the Split Cauchy-Gaussian noise distributions are twice continuously differentiable at 0, the distributions' means.

V. RESULTS

Distributions of parameter estimates are shown as histograms in **Figures 4-8** (See Appendix). Summary statistics of the distributions generated by "***Noise-Method***" pairs are listed in Table 4 (See Appendix). The distributions' Mean Error, Mean Squared Error (MSE), Median Error, Median Average Deviation (MAD), and Standard Deviation are reported for each Noise-Method pair. The MSE is a traditional metric of judging an estimate's performance while the MAD is a robust measure of a distribution's spread that is most useful when variance is undefined or infinite.

In Table 4 we see that for the 20,000 trials of models ran with 1000 observations each (***n*** = 1000), OLS performed very well on Gaussian Noise, tailing LAD in only two metrics. OLS soon becomes very inaccurate as it has very large standard deviations and MSE for non-Gaussian noise. The histograms (**Figures 4-8**) for the OLS parameter distribution for non-Gaussian data

similarly have much larger variance and do not appear to be distributed normally, but more closely have a limiting distribution resembling a Cauchy distribution.

LAD performed very well on the noise of the Cauchy distribution. The slope parameter was distributed normally with variance below only that of the Cauchy-MLE solution. The distribution of the constant term beta0 for the Cauchy-LAD pair is similarly unbiased since the Cauchy distribution is symmetric, so the intercept for the median is the same as that of the mean. LAD performs poorly however when noise is distributed as a mixture of Cauchy distributions. This is potentially due to the median of the distribution falling in a low-probability area in the Probability Density Function, resulting in an estimate having a high-variance distribution that is not yet close to a normal distribution with only ($\boldsymbol{n}$ = 1000). Preliminary tests with larger sample sizes show that the distribution still limits on being normal, but may simply require a larger number of observations before inference may be performed.

The various MLE estimation methods usually performed optimally for their corresponding noise distributions, with the exception of the Split Cauchy-Normal and Two-Tailed Q-Exponential. Even when the MLE method was used to estimate a model with noise distributed differently, as can be seen in the case of the Cauchy MLE method used with Cauchy Mixture or Split Cauchy-Gaussian noise, generally, the MLE solutions we still very robust and rather unbiased.

Apparent in the histograms of all the noise distributions, the Split Cauchy-Gaussian MLE solution very often did not converge to a normal distribution, and, rather, stayed at the initialized value of 0.5. This is due to the Split Cauchy-Gaussian distribution failing one of our assumptions required to perform inference with M-Estimators. The finite sample of observations lead to the $Q_n(b)$ function that we are minimizing being a sum of discontinuous and

non-differentiable functions at the center of the distributions, the point of interest to us. Thus, $Q_n(b)$ itself becomes a non-differentiable function.

Now, when maximized via gradient descent, the Hessian evaluated at the optimal parameter $\boldsymbol{b}$, possibly will no longer be negative definite; additionally, gradient descent may also never make it into the neighborhood of the global maximizer as the sum of discontinuous functions make the optimization non-convex, causing the maximizer to settle in non-optimal maxima, resulting in biased results that do not limit on the normal distribution. Even when over 80% of the estimates never left the local maximum of $\boldsymbol{b}$=0.5, if one were to zoom in, the distribution of the other 20% of estimations centered around $\boldsymbol{b}$=1 shares the extended right tail seen in the Gaussian noise solution via Piecewise Cauchy-Gaussian MLE. So, even when gradient descent succeeded in leaving obvious local maxima, the limiting distribution still appeared biased, likely due to the Hessian being undefined in the limit at 0.

While it was expected that the Two-Tailed Q-Exponential would suffer similarly to the Split Cauchy-Gaussian distribution, LAD and MLE approaches both performed surprisingly well. Apparently, the difference between discontinuous and non-differentiable is a large one. It makes sense, however, that in the sample approximation $Q_n(b)$, a sum of function evaluations for a given set of parameters, being non-differentiable at a small fraction of the function evaluations will be less damaging to the final minimization than being discontinuous at a small fraction of the function evaluations. This results in perhaps slightly large variance in estimates and possibly slight differences in the limiting distribution preventing it from being exactly normal.

VI. CONCLUSION

Our results match the majority of our predictions made about the properties and limits of the different models. MLE and LAD are good estimation methods under the right circumstances,

and can help in situations when OLS breaks down because of a distribution with infinite variance, but are by no means perfect replacements for the ability to run an OLS model with noise distributed with finite variance. We were able to predict correctly that the split cauchy-gaussian distribution would perform terribly in our Monte Carlo simulations and that the limiting distribution of parameter estimates has larger variance or bias when noise distribution does not satisfy our model requirements. The Q-exponential was a promising distribution, as even though it didn't meet our assumptions, the data showed it was practically normal and that inference would be possible in application, if not truly normal in reality.

The next step in the research in modelling data with infinite variance noise would involve removing the element of homoskedasticity in our noise distributions or even changing the very noise distribution as the data changes. While we included this requirement in our analysis, we understand it is not an assumption that holds in all data. Future research focused on looking into the behavior of these models under the circumstance of infinite variance with heteroskedasticity would likely introduce either more tenuous assumptions or remove our ability to perform inference, but the trade-off between the two would be interesting to explore.

It is important to note that while OLS did not perform well with infinite variance, it is still mathematically the best linear estimator when noise is distributed normally, and consistent under any noise with finite variance. It is infrequent to find distributions with infinite variance, and even harder to correctly attribute characteristics such as symmetry and differentiability, let alone distribution. OLS is the most simple and well understood of all estimation models and performs well in many research publications.

VII. APPENDIX

M-estimator assumptions for identification (necessary for b to converge to the true model parameter β):
1. Qn(b),the model estimate with n observations, converges uniformly in probability to a function Q∞(b)
2. Q∞(b) is uniquely maximized at β
3. The domain of possible parameters B is compact
4. Q∞(b)is continuous

Note 1

Tables:

Distribution	**Symmetric**	**Mean = Median**	**Continuous**	**Differentiable**	**Fat-tailed**
Gaussian	Yes	Yes	Yes	Yes	No
Cauchy	Yes	N/A	Yes	Yes	Yes
Two-Tailed Q-Exponential	Yes	Yes	Yes	No	Yes
Cauchy Mixture	No	N/A	Yes	Yes	Yes
Split-Cauchy-Normal	No	N/A	No	No	One side only

Table 1

	Strength Relative to OLS	Weakness Relative to OLS	Bias Expected When
Least Absolute Deviations	Less sensitive to outliers	Much more difficult to compute and less established properties	Median ≠ Mean
Maximum Likelihood	Most efficient estimator possible	Relies on the strong assumption that the joint density function of the data is fully specified up to the finite-dimensional parameter β.	Distribution is incorrectly specified

Table 2

	Ordinary Least Squares	Least Absolute Deviations	Maximum Likelihood
Assumptions	1. Linearity 2. Full Rank 3. Exogeneity of the independent variable 4. Homoscedasticity 5. Data generation 6. Normal Distribution	1. M-Estimator Assumptions 2. $E[\varepsilon\|x]=0$ 3. $Med[y\|x]=x'\beta$	1. M-Estimator Assumptions 2. The objective Function converges uniformly. 3. The limiting function is uniquely maximized at β
Strengths	Easy to compute and well known	Less sensitive to outliers than OLS therefore more "robust" than Least Squares Estimators	Not only Consistent and Asymptotically Normal but also the asymptotically efficient estimator possible.
Breakdowns	**1n(X'X) not invertible when variance is infinite, so no model consistency or asymptotic normality**	The noise may have mean zero while the median of the noise may be non-zero. This would occur for any noise distribution that is asymmetric, causing the mean and the median to differ.	Assumption three is a very strong assumption and difficult to believe.

Table 3

		Noise				
Method		**Gaussian**	**Cauchy**	**Two-Tailed Q-Exponential**	**Cauchy Mixture**	**Split Cauchy Normal**
OLS	**Mean Error**	-0.00087	6.270396	0.006413	-0.20449	-0.93873
	Median Error	-0.0008	-0.01399	0.007409	-0.01705	**-0.00243**
	Standard Dev.	**0.108739**	749.9256	1.698639	263.3621	544.472
	MAD	**0.07301**	3.043318	0.56893	1.594668	1.462793
	MSE	**0.011825**	562427.7	2.885415	69359.65	296450.6
LAD	**Mean Error**	**-0.00027**	-0.00135	0.001858	0.011466	0.002187
	Median Error	**-0.00053**	**-0.0004**	0.004717	0.024038	0.003877
	Standard Dev.	0.136602	0.170852	**0.174121**	4.491773	0.154513
	MAD	0.092568	0.114977	**0.115674**	3.369764	0.103517
	MSE	0.01866	0.029192	**0.030322**	20.17615	0.023879
MLE	**Mean Error**	-0.00058	**-0.0009**	0.001697	**0.000109**	0.001843
Cauchy	**Median Error**	-0.00097	-0.00136	**0.002849**	**0.00022**	0.002654
	Standard Dev.	0.12461	**0.153942**	0.175606	0.100858	**0.133599**
	MAD	0.084432	**0.103711**	0.119112	0.066976	**0.090162**
	MSE	0.015528	**0.023699**	0.03084	0.010172	**0.017852**
MLE	**Mean Error**	-0.00055	-0.00107	**0.001346**	-7.60E-05	**0.001592**
Mixture	**Median Error**	-0.00082	-0.00177	0.003357	0.000329	0.002592
	Standard Dev.	0.127033	0.161366	0.189379	**0.078511**	0.135043
	MAD	0.08606	0.108665	0.127103	**0.053063**	0.090658
	MSE	0.016138	0.02604	0.035866	**0.006164**	0.018239
MLE	**Mean Error**	-0.01871	-0.01964	-0.01099	-0.00026	-0.01016
Q-exponential	**Median Error**	-0.01626	-0.01859	-0.00501	-0.00052	-0.00965
	Standard Dev.	0.238821	0.236845	0.201556	0.086754	0.197155
	MAD	0.160961	0.158256	0.128562	0.058797	0.130057
	MSE	0.057386	0.056481	0.040746	0.007526	0.038973
MLE	**Mean Error**	0.002731	-0.50082	-0.46682	-0.48944	0.02768
Split Cauchy Normal	**Median Error**	0.002728	-0.5	-0.5	-0.5	0.010293
	Standard Dev.	0.134214	0.216809	2.787769	0.095171	0.161912
	MAD	0.085114	0.5	0.5	0.5	0.097846
	MSE	0.018021	0.297829	7.989577	0.248613	0.026982

Table 4

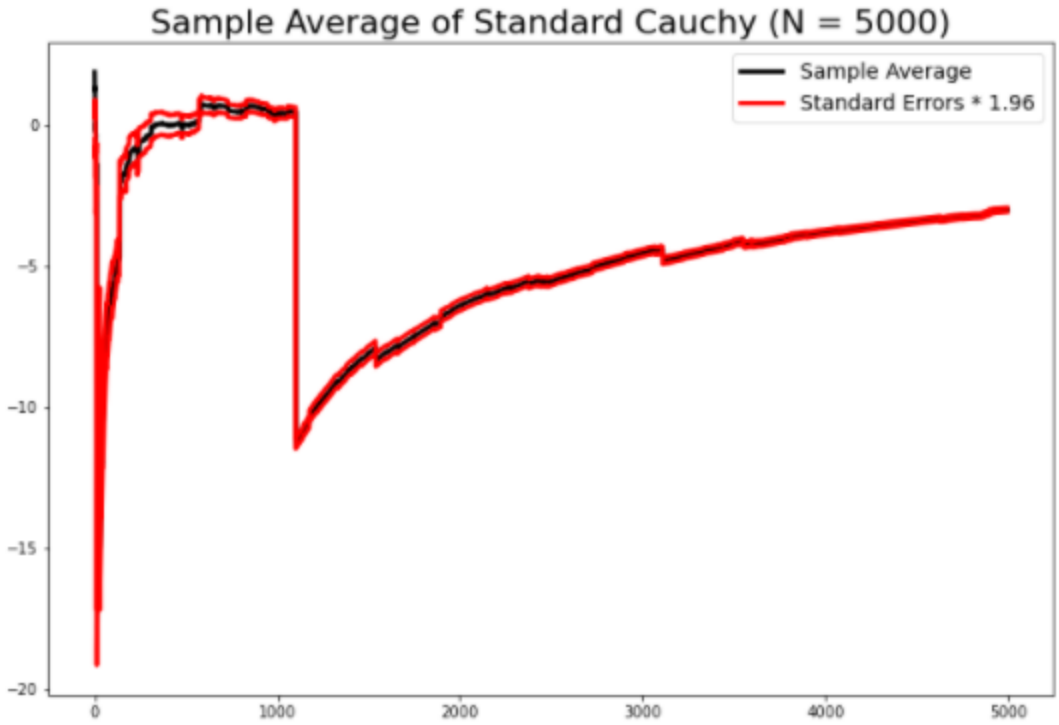

Figure 1

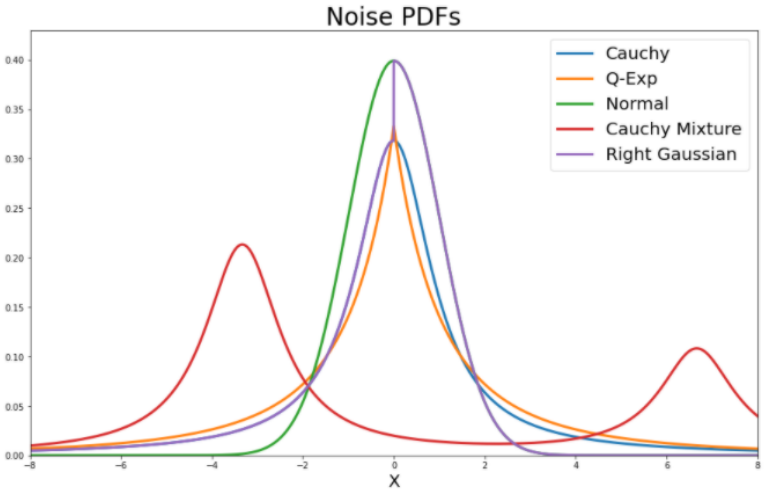

Figure 2

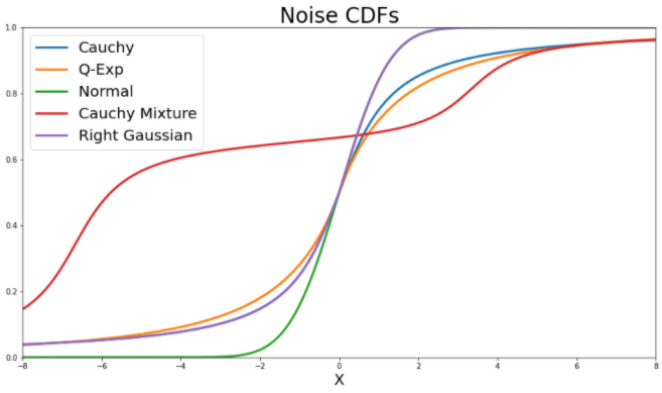

Figure 3

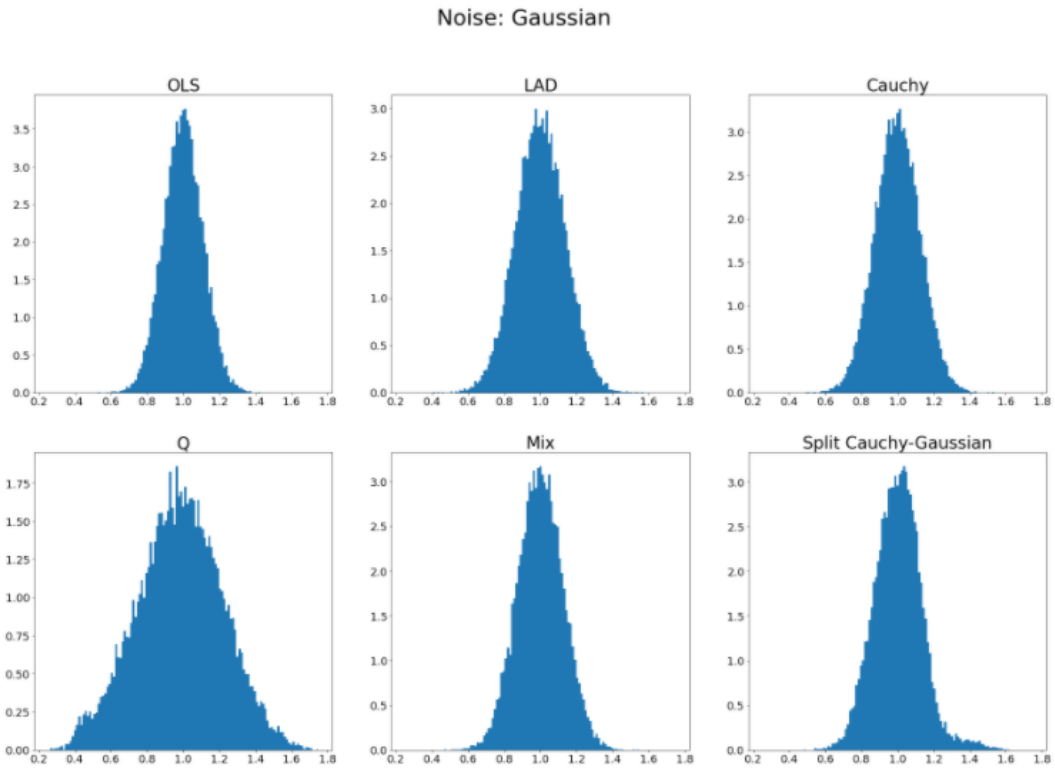

Figure 4

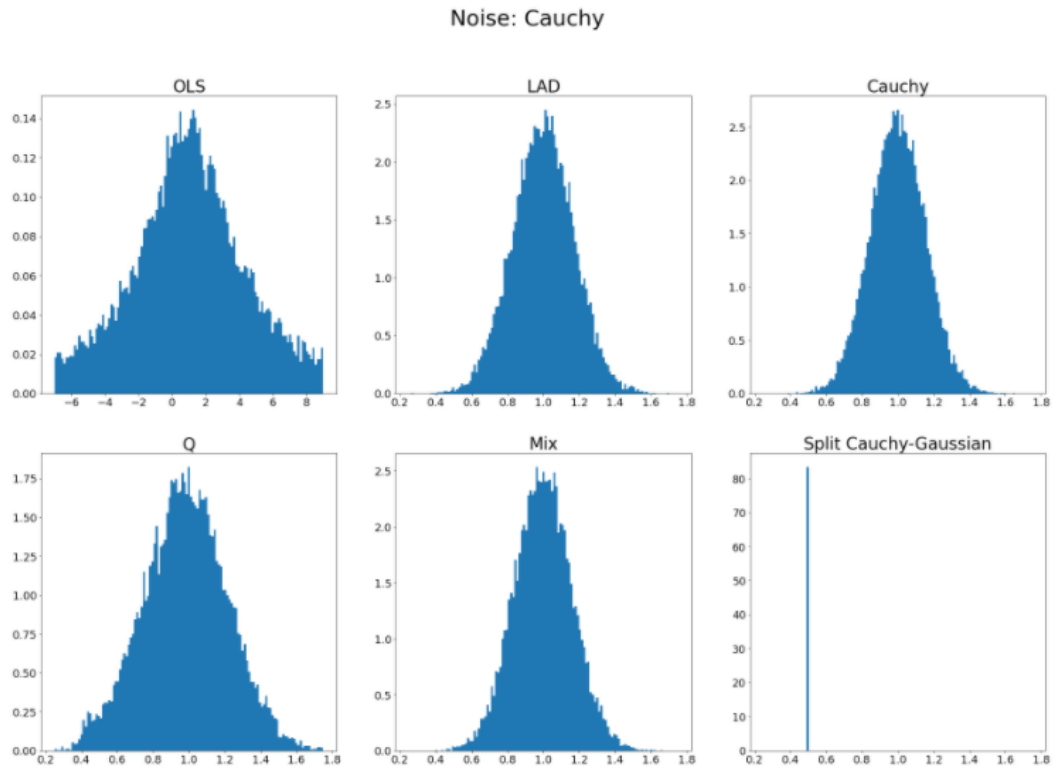

Figure 5

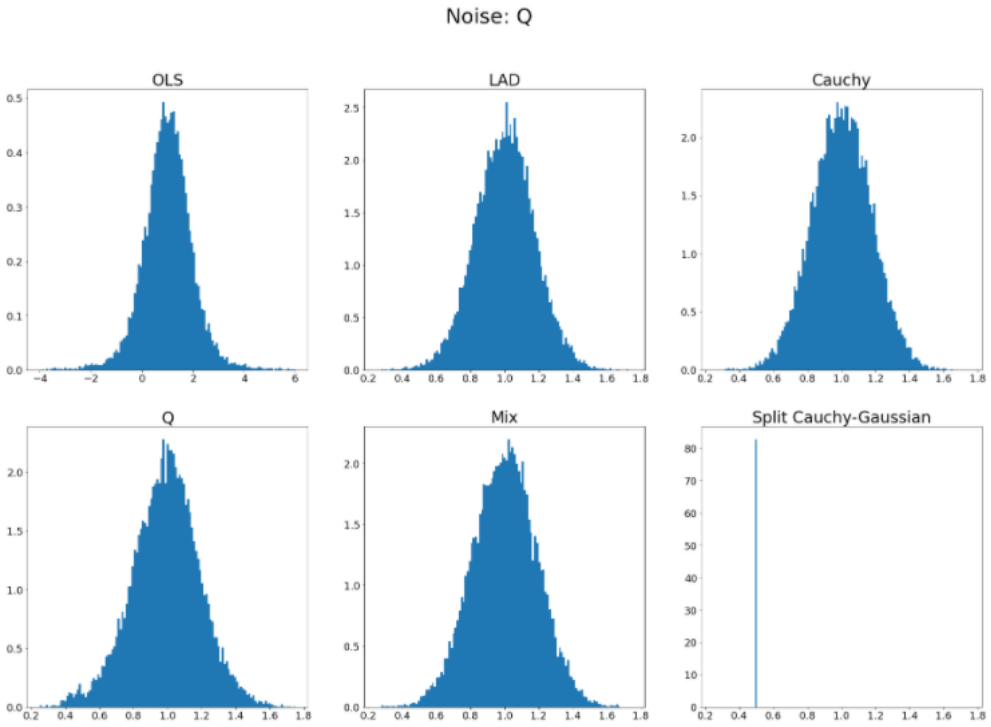

Figure 6

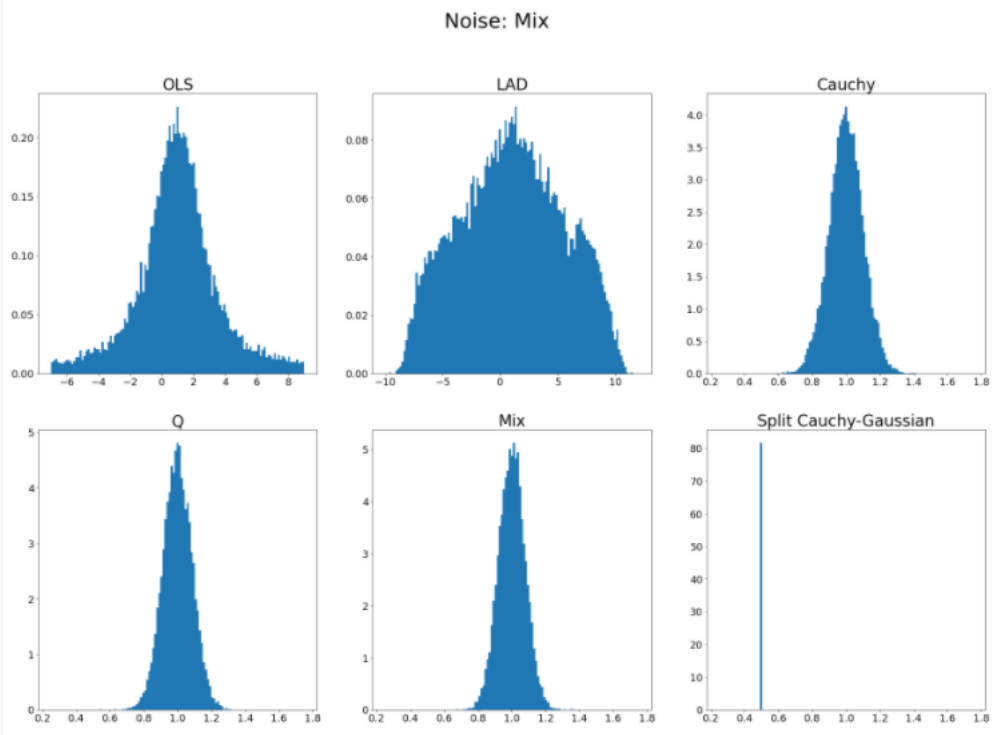

Figure 7

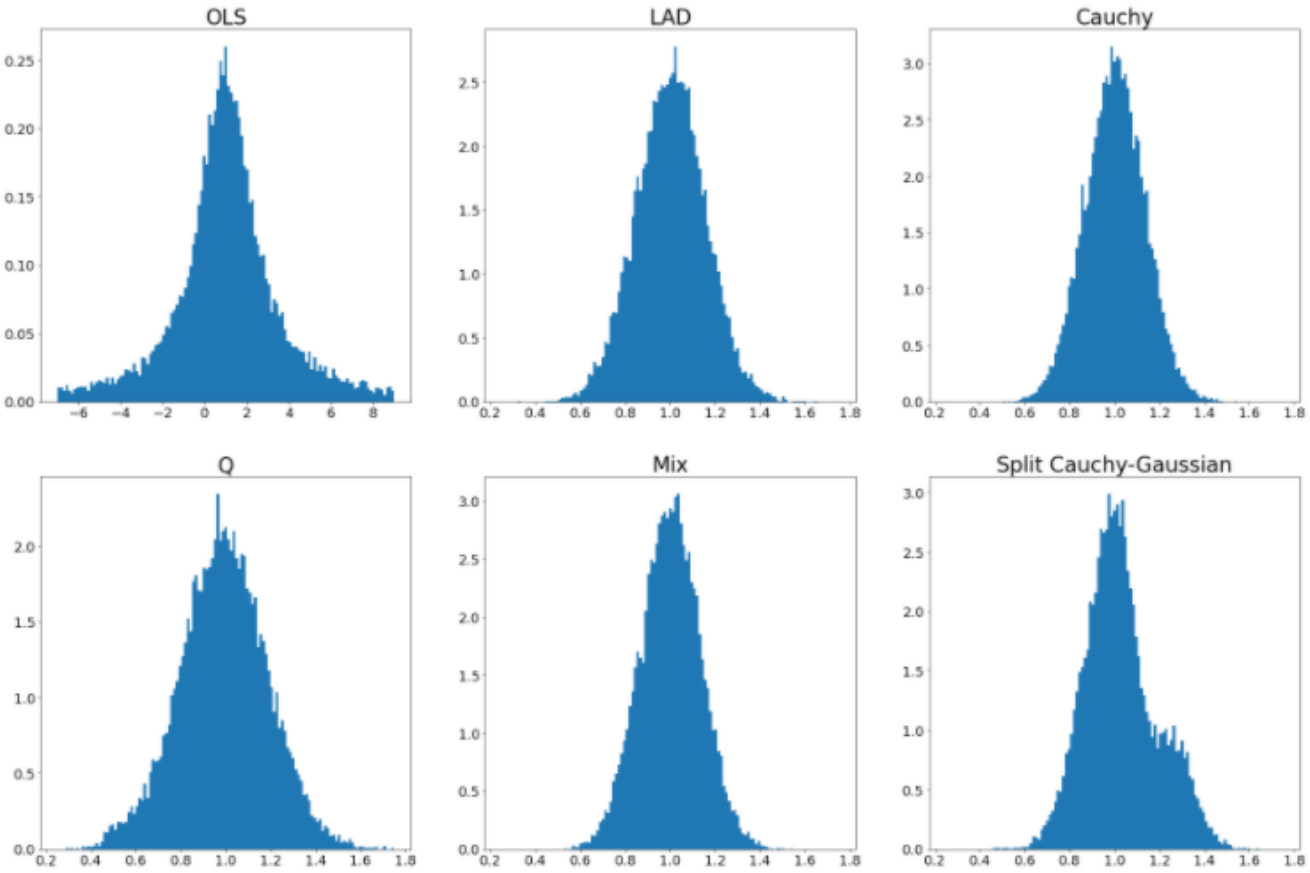
Noise: Split Cauchy-Gaussian
OLS
LAD
Cauchy
Q
Mix
Split Cauchy-Gaussian

Figure 8