

The Challenge of Non-Linear Regression on Large Datasets with Asymmetric Heavy Tails

Y. Bengio, Universite de Montreal,
I. Takeuchi, Mie University,

T. Kanamori, Tokyo Institute of Technology

2920 Chemin de la Tour, Montreal, Quebec, Canada H3T 1J8, 2194 (bengio@iro.umontreal.ca)

Key Words: robust estimation, asymmetric error, heavy-tail distribution, quantile regression, neural networks.

Abstract:

Regression becomes unstable under a heavy-tail error distribution due to dominant effects of *outliers*. Traditional robust estimators are helpful under *symmetric* error, by reducing the effect of outliers equally from both sides of the distribution. Under *asymmetric* error, however, those estimators are *biased* because the outliers appear only one side of the distribution. Motivated by data-mining problems for the insurance industry, we propose in this paper a new approach to robust regression that is tailored to deal with asymmetric error. The main idea is to estimate a majority of the parameters using *quantile regression* (which may be biased but robust), and to estimate the few remaining using least squares estimator to correct the biases. Theoretical analysis shows conditions when the conditional expectations can be recovered from the conditional quantiles and what can be gained asymptotically with the proposed algorithm. Experiments confirm the clear advantages of the approach. Results are on synthetic data as well as real insurance data, using both linear and neural-network predictors.

1. Introduction

We often find distributions with an *asymmetric heavy tail* like in Fig. 1 (ii). Estimating the mean of such a distribution is essentially difficult because *outliers*, samples from the tail, have a strong influence on it. For *symmetric* distribution, the problems can be reduced using *robust* estimators [1] which intend to ignore or put less weights on outliers. Note, however, that **these robust estimators are biased under asymmetric distributions**: most outliers are on the same side of the mean, thus down-weighting them introduces a strong bias on its estimation.

We are concerned in this paper with the problem of linear or non-linear regression under asymmetric heavy-tail error distribution¹. As in the unconditional case, regres-

sion problems also suffer from outliers, and straightforward robust regression estimators (such as M-estimators) are biased under asymmetric error.

In this paper, we propose a new robust regression estimator which can be applied under (possibly) asymmetric error. We present a theoretical analysis, numerical experiments with synthetic data and an application to automobile insurance premium estimation.

Throughout the paper, we use the following notations: X and Y for the input and the output random variables, $F_W(\cdot)$ for the cumulative distribution function (cdf) density function (pdf) of random variable W . And when appropriate, the conditional distribution $Y|X = x$ is written $Y|x$.

2. Related Work

2.1 Traditional Robust Estimators

When we estimate the mean of heavy-tail distribution, the empirical median might be a better choice than the empirical average because the former reduces the effects of outliers. Similarly, in regression problems with a heavy-tail error, L_1 regression can be much more efficient than Least Square (LS) regression. Note that in general L_1 regression estimates the conditional median $F_{Y|x}^{-1}(0.5)$ and LS regression estimates the conditional mean $E[Y|x]$, but both of these conditional moments coincide in the case of symmetric noise. These ideas have been studied in the context of robust estimators. *M-estimators* [2] are among the most successfully used for regression problems². The basic idea of M-estimators is to use the minimization schema:

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \sum_i \rho(y_i - f(x_i)) \quad (1)$$

where \hat{f} denotes estimation on data, \mathcal{F} is a set of functions, (x_i, y_i) is the i -th training sample, and ρ is a loss

of the system and ϵ is the (zero-mean) error term following asymmetric heavy-tail distribution, we consider the robust estimation of f .

²Many efforts have been devoted to deal with outliers in the independent variable X , called *leverage points*, as well as those in the dependent variable Y . In this paper, we do not consider outliers in X and related techniques [3].

¹For a class of systems: $Y = f(X) + \epsilon$, where X and Y are independent and dependent variables, respectively, f is a deterministic part

Table 1: A Comparison of Estimators of Unconditional Mean under Asymmetric Distributions. The figures are mean (variance) of the estimator over 100 trials, each with 10 samples, for several methods². The samples are from a positively-skewed log-Normal distribution³ which is shifted so their mean equals zero. Note that all but LS estimators are negatively biased and the variances of LS estimator are larger than the others.

Estimator	Degree of Asymmetry		
	low	intermediate	high
LS	0.03 (0.18)	0.08 (3.02)	0.22 (134.83)
Median	-0.26 (0.08)	-1.35 (0.33)	-7.13 (1.72)
Cauchy	-0.25 (0.07)	-1.25 (0.20)	-6.36 (1.65)
Huber	-0.13 (0.07)	-0.41 (0.62)	-2.41 (35.82)
Tukey	-0.47 (0.13)	-1.39 (0.15)	-4.76 (4.57)

function, which is *symmetric*, i.e., $\rho(-t) = \rho(t)$, with a unique minimum at zero. As particular cases, $\rho(t) = \frac{1}{2}t^2$ yields the LS estimator and $\rho(t) = |t|$ yields the median estimator.

Unfortunately, **these robust estimators do not work well for asymmetric distributions**. Whereas removing outliers symmetrically on both sides of the mean does not change the average, removing them on one side changes the average considerably. (See Table 1: a small Monte-Carlo study for quantitative evidence on this.) For many regression problems as well, under asymmetric error, the straightforward use of "robust" loss functions (1) is generally not helpful, as confirmed in section 4.

2.2 Transformations

One commonly suggested solution is a transformation to symmetry. For example, the Box-Cox transformation [5] is given by

$$Y^{(\lambda)} = h(Y; \lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log Y & \lambda = 0 \end{cases}, \quad (2)$$

where λ is a parameter that characterizes the transformation.

A naive approach using transformation techniques for regression problems is as follows:

1. transform dependent variable Y with a parameter λ so that transformed variable $Y^{(\lambda)}$ be symmetrically distributed,
2. estimate a regression function $f^{(\lambda)}$, s.t. $E[Y^{(\lambda)}|x] = f^{(\lambda)}(x) + \epsilon^{(\lambda)}$, where $\epsilon^{(\lambda)}$ is a residual.
3. estimate a regression function f , s.t. $E[Y|x] = f(x) + \epsilon$ by inverse transformation $h^{-1}(f^{(\lambda)}(x); \lambda)$, where ϵ is a residual.

²Those are Cauchy [1], Huber [2] and Tukey [4].

³Log-Normal distributions (see sec.4.), whose p_μ , a degree of asymmetry introduced in the next section, are 0.65(low), 0.75(intermediate) and 0.85(high), respectively.

The advantage of transformation techniques is found in step 2, where the estimation of regression function would be more stable because the error term is symmetrically distributed.

On the other hand, we face difficulties in step 1 and 3. In step 1, we have to estimate λ from data, that is usually unstable because it depends on the estimations of higher order moments. In step 3, we have to note that the inverse-transformation of the conditional expectation of the transformed variable does not equal the conditional expectation of the original variable, i.e.

$$h^{-1}(E[h(Y; \lambda)|x]; \lambda) \neq E[Y|x]. \quad (3)$$

Thus, the simple usage of transformation techniques does not provide an unbiased estimation of $E[Y|x]$.

To estimate $E[Y|x]$ from $E[Y^{(\lambda)}|x]$ one need to compute the following

$$E[Y|x] = \int h^{-1}(E[Y^{(\lambda)}|x] + \epsilon^{(\lambda)}; \lambda) P(\epsilon^{(\lambda)}) d\epsilon^{(\lambda)} \quad (4)$$

where $\epsilon^{(\lambda)}$ is the residual of the regression for transformed data, i.e. $\epsilon^{(\lambda)} = Y^{(\lambda)}|x - E[Y^{(\lambda)}|x]$. Duan [6] proposed to use empirical estimates of $\epsilon^{(\lambda)}$, which he named *smearing estimate*:

$$E[Y|x] = \frac{1}{n} \sum_{i=1}^n h^{-1}(\hat{f}^{(\lambda)}(x) + \hat{\epsilon}^{(\lambda)}; \lambda). \quad (5)$$

We will refer to the smearing estimate in section 4.

2.3 Parametric Approaches

Some parametric approaches have been proposed for regression under asymmetric error. For example, Williams [7] proposed to use Johnson's S_u distribution, that is characterized by four parameters: each for *location*, *scale*, *skewness* and *thickness*. Here again, elaborate parameterizations of heavy-tail asymmetric distributions depend on the estimation of high-order moments, which is sensitive to outliers and much less stable than the estimations of low-order moments.

One might consider, then, approximate parametric models inspired by prior knowledge of the data. The *Log-normal* distribution [8] is a typical example of such a parametric choice. If we knew that the error term was log-normal, we could apply a maximum likelihood principle to estimate the model. Alternatively, we could use the strategy explained in the previous subsection, where we set $\lambda = 0$ and $\epsilon^{(\lambda)}$ in eq. (4) is Normal. However, unless we have explicit knowledge of the error distribution, making such assumptions can yield to poor results (note in particular that the error made on the distribution's tail can have a drastic effect on the conditional expectation when this tail is heavy).

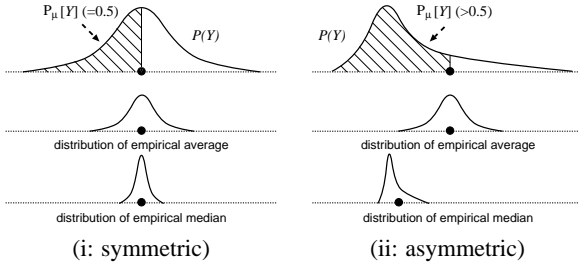


Figure 1: The schematic illustration of empirical averages and empirical medians for (i) symmetric distribution and for (ii) asymmetric distribution: Distributions of a heavy-tail random variable Y , its empirical average and empirical median, and their expectations (black circles) are shown. In (i) those expectations coincide, while in (ii) they do not. As indicated by the areas of slanting lines, we define $P_\mu[Y]$ as the order at which the quantile coincides with the mean.

3. Proposed Method

3.1 Motivation

We have seen in the previous section that the median estimator is robust but biased under asymmetric distributions. We first consider an extension of median estimator as a motivation. Under asymmetric distributions, **the median does not coincide with the mean but another quantile does**. We consider its order to characterize asymmetric distributions. For a distribution $\text{Prob}(Y)$, we define $P_\mu[Y] \triangleq F_Y(E[Y]) = \text{Prob}(Y < E[Y])$. Note that $P_\mu[Y] > 0.5$ (< 0.5) suggests $\text{Prob}(Y)$ is positively (negatively) skewed. When the distribution is symmetric, $P_\mu[Y] = 0.5$. Fig. 1 illustrates this idea.

For regression problems under asymmetric error, we may extend median regression to p -th quantile regression [9] that estimates the conditional p -th quantile $F_{Y|x}^{-1}(p)$, i.e. we want \hat{f}_p s.t. $P(Y < \hat{f}_p(x)) = p$:

$$\hat{f}_p = \underset{f \in \mathcal{F}_q}{\text{argmin}} \left\{ \sum_{i: y_i \geq f(x_i)} p |y_i - f(x_i)| + \sum_{i: y_i < f(x_i)} (1-p) |y_i - f(x_i)| \right\}, \quad (6)$$

where \mathcal{F}_q is a set of quantile regression functions. A straightforward idea for our regression task is to estimate $f_{P_\mu[Y|x]}$, instead of $f_{0.5}$. But what is $P_\mu[Y|x]$? It must be defined as $P_\mu[Y|x] \triangleq F_{Y|x}(E[Y|x]) = \text{Prob}(Y < E[Y|x])$. Accordingly the above idea raises 3 problems, which we will address with the algorithm proposed in the next subsection: (i) $P_\mu[Y|x]$ may depend on x in general, (ii) unless the error distribution is known, $P_\mu[Y|x]$ itself must be estimated, which is maybe as difficult as estimating $E[Y|x]$, (iii) if the density of error distribution at $P_\mu[Y|x]$ is low (because of the heavy tail and large value

of $P_\mu[Y|x]$), the estimator in eq. (6) may itself be very unstable.

3.2 Algorithm

We propose a new algorithm for regression under (possibly) asymmetric error. The main idea is

1. estimate a majority of the parameters with quantile regressions, each of which is biased but robust,
2. estimate the few remaining parameters to correct the bias in step 1, and combine the above estimators,

hence the name of **Robust Regression for Asymmetric Tails** algorithm (RRAT):

Algorithm RRAT(n)

Input: training data: $(x_1, y_1), \dots, (x_N, y_N)$, hyper parameters: $n \in \{1, 2, \dots\}$, $p_1, \dots, p_n \in (0, 1)$, function classes: $\mathcal{F}_q, \mathcal{F}_c$

Step 1: Using eq.(6), estimate n quantile regressions: $\hat{f}_{p_1}, \dots, \hat{f}_{p_n} \in \mathcal{F}_q$ at orders p_1, \dots, p_n respectively.

Step 2: Compute n dimensional vector $q(x_i) = (\hat{f}_{p_1}(x_i), \dots, \hat{f}_{p_n}(x_i))$ for each i . Consider a functional relationship: $y = f_c(q(x)) + \epsilon$ and estimate $f_c \in \mathcal{F}$ by LS.

Output: conditional expectation estimator $\hat{f}(x) = \hat{f}_c(\hat{f}_{p_1}(x), \dots, \hat{f}_{p_n}(x))$.

An outer model selection loop is often required in order to select hyper-parameters such as n, p_1, \dots, p_n , or capacity control parameters for \mathcal{F}_q and \mathcal{F}_c .

In general, we believe that this method yields more robust regressions when the number of parameters required to characterize \hat{f}_c is small (because they are estimated with “non-robust” LS) compared to the number of parameters required to characterize the quantile regressions $\hat{f}_{p_i}, i = 1, \dots, n$ as confirmed in subsection 4.2.

Problem (ii) above is dealt with by doing quantile regressions of orders $p_1 \dots p_n$ not necessarily equal to $P_\mu[Y|x]$. Problem (iii) is dealt with if $p_1 \dots p_n$ are in high density areas (where estimation will be stable). Unfortunately, the solution to problem (i) is less obvious. However, we have specified a class of noise structure for which the problem is avoided (see next two subsections), and we conjecture that a general solution exists when we choose a large n .

3.3 Why RRAT works (Simple Examples)

In this subsection, we consider simple examples to illustrate the underlying idea of RRAT⁵. In particular, we consider **additive noise structures** (see eq. (8)) and **multiplicative noise structures** (see eq. (10)).

⁵We consider only RRAT(1) in this subsection.

As explained in the previous subsection, we want to estimate as many of the parameters as possible in the 1st step because they are estimated through quantile regression and expected to be robust. On the other hand, in the 2nd step, we want to estimate as few parameters as possible because they are estimated through LS and not robust. In the 2nd step, we estimate a function f_c , that estimates the conditional expectation $E[Y|x]$ from the p -th quantile regression $F_{Y|x}^{-1}(p)$, i.e. with $n = 1$ we would like to have

$$E[Y|x] = f_c(F_{Y|x}^{-1}(p)). \quad (7)$$

We will see below how many parameters are required to characterize f_c in each noise structure.

[additive noise structure]

The additive noise structure is given by:

$$Y = E[Y|x] + \epsilon, \quad E[\epsilon] = 0, \quad \epsilon : \text{i.i.d.}, \epsilon \perp X, \quad (8)$$

illustrated in Fig. 2. The solid line is the conditional expectation $E[Y|x]$ and the dotted line is the p -th quantile regression $F_{Y|x}^{-1}(p)$ (e.g. $p = 0.5$). It is clear that the difference between the conditional expectation $E[Y|x]$ and the p -th quantile regression $F_{Y|x}^{-1}(p)$ is **constant** (does not depend on x), then we can write

$$E[Y|x] = F_{Y|x}^{-1}(p) + \underbrace{\{E[\epsilon] - F_{\epsilon}^{-1}(p)\}}_{\text{constant}},$$

$$E[Y|x] = c_0 + F_{Y|x}^{-1}(p). \quad (9)$$

Thus, f_c is characterized by **only a single additive parameter**.

[multiplicative noise structure]

The multiplicative noise structure is given by

$$Y = E[Y|x] \cdot \epsilon, \quad E[\epsilon] = 1, \quad \epsilon : \text{i.i.d.}, \epsilon \perp X, \quad (10)$$

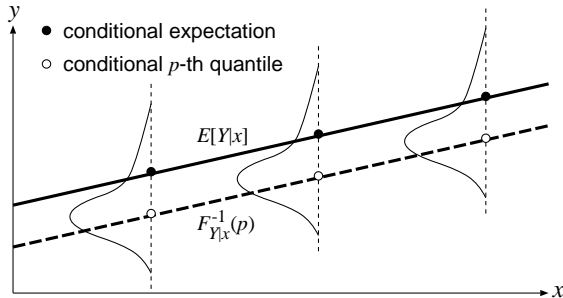


Figure 2: Additive Noise Structure: The difference between the conditional expectation $E[Y|x]$ (solid line) and the p -th quantile regression $F_{Y|x}^{-1}(p)$ (dotted line) is constant (does not depend on x).

illustrated in Fig. 3. The solid line is the conditional expectation $E[Y|x]$ and the dotted line is the p -th quantile regression $F_{Y|x}^{-1}(p)$ (e.g. $p = 0.5$). It is clear that the difference between the conditional expectation $E[Y|x]$ and the p -th quantile regression $F_{Y|x}^{-1}(p)$ is **linear in $E[Y|x]$** , then we can write

$$E[Y|x] = F_{Y|x}^{-1}(p) + \underbrace{E[Y|x] \cdot \{E[\epsilon] - F_{\epsilon}^{-1}(p)\}}_{\text{linear in } E[Y|x]},$$

$$E[Y|x] = c_1 \cdot F_{Y|x}^{-1}(p). \quad (11)$$

Thus, f_c is characterized by **only a single multiplicative parameter**.

3.4 Applicable Class of Problems

In the previous subsection, we saw two simple examples of systems in which RRAT may work well. There were two points we considered:

- **(representability)** in the 2nd step, a function f_c must exist, i.e. a small set of quantile regression $F_{Y|x}^{-1}(p_1), F_{Y|x}^{-1}(p_2) \dots F_{Y|x}^{-1}(p_n)$ must be able to represent the conditional expectation $E[Y|x]$,
- **(robustness)** the number of parameters of f_c affects the robustness of RRAT, i.e. the less parameter f_c has, the more robustness would be brought by RRAT.

In this subsection, we will clarify the class of systems for which RRAT may be applied from these two points of views (representability and robustness).

The following three theorems come directly from the consideration in the previous subsection.

Theorem 1.1. If the error structure is additive, a single additive parameter is sufficient in the 2nd step of RRAT.

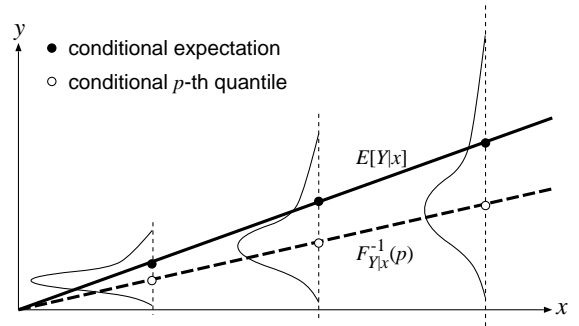


Figure 3: Multiplicative Noise Structure: The difference between the conditional expectation $E[Y|x]$ (solid line) and the p -th quantile regression $F_{Y|x}^{-1}(p)$ (dotted line) is linear in $E[Y|x]$.

proof: see eq. (9).

Theorem 1.2. If the error structure is multiplicative, a single multiplicative parameter is sufficient in the 2nd step of RRAT.

proof: see eq. (11).

Theorem 1.3. If the error structure is the combination of additive and multiplicative, two parameters are sufficient to be estimated in the 2nd step of RRAT.

proof: As a natural extension of eqs. (8) and (10), the combination of additive and multiplicative noise structures is given by

$$Y = E[Y|x] + (\alpha_0 + \alpha_1 E[Y|x]) \cdot \epsilon, \quad E[\epsilon] = 0, \quad \epsilon : \text{i.i.d.}, \epsilon \perp X. \quad (12)$$

In the same way as the derivations of eqs. (9) and (11), the difference between the conditional expectation $E[Y|x]$ and the p -th quantile regression $F_{Y|x}^{-1}(p)$ has affine relation w.r.t $E[Y|x]$, i.e.

$$E[Y|x] = F_{Y|x}^{-1}(p) + \underbrace{(\alpha_0 + \alpha_1 E[Y|x]) \cdot \{E[\epsilon] - F_{\epsilon}^{-1}(p)\}}_{\text{affine w.r.t. } E[Y|x]}, \quad (13)$$

$$E[Y|x] = c_0 + c_1 F_{Y|x}^{-1}(p).$$

Thus, f_c is characterized by two parameters **Q.E.D.**

These three theorems satisfy the both requirements of *representability* and *robustness*. From our experiences through several applications, a number of data sets seem to have both of additive and multiplicative noise. Thus, theorem 1.3 is practically important and useful.

The next theorem concerns only *representability*. With the use of two quantile regressions (RRAT(2)), we can guarantee the representability of wider class of noise structure.

Theorem 2. If the noise structure is given by

$$Y = E[Y|x] + g\{E[Y|x]\} \cdot \epsilon, \quad E[\epsilon] = 0, \quad \epsilon : \text{i.i.d.}, \epsilon \perp X, \quad (14)$$

where g is an arbitrary positive range function, and

$$p_1 \neq P_{\mu}[\epsilon], \quad p_2 \neq P_{\mu}[\epsilon], \quad (15)$$

$$p_1 \neq p_2, \quad (16)$$

then there exists a function f_c such that $E[Y|x] = f_c(F_{Y|x}^{-1}(p_1), F_{Y|x}^{-1}(p_2))$.

The proof is given in [10]. Eq. (14) includes additive or/and multiplicative noise structures as special cases,

and covers a wide variety of heteroscedastic regression systems. However, the theorem does not tell us about the number of parameters in f_c . When f_c is complicated, *robustness* might not be brought by RRAT.

We also **conjecture** that the noise structure (14) can be more general when combining a sufficient number of quantile regressions. When $Y > 0$, we have that

$$E[Y|x] = \int_0^{\infty} P(Y > y|x) dy = \int_0^1 P(Y > f_q(x)) \frac{df_q(x)}{dq} dq \quad (17)$$

by substituting $y = f_q(x)$ s.t. $P(Y > f_q(x)) = q$. The integral can be unbiasedly estimated by a number of numerical integration schemes, (the simplest being uniform spacing of q , and better results might be obtained using, for example, a Gaussian quadrature). Such approximations correspond to a **linear combination of the quantile regressors**. The derivative $\frac{df}{dq}$ can be unbiasedly (and efficiently) estimated with $(f_{q_{i+1}}(x) - f_{q_{i-1}}(x)) / (q_{i+1} - q_{i-1})$.

3.5 Asymptotic Behavior

We derived an asymptotic property of RRAT for the additive noise structure. Consider

$$Y = f(x; \beta^*) + \epsilon, \quad x \in \mathbb{R}^d, Y \in \mathbb{R} \quad (18)$$

with a function f , a set of parameters β^* and zero-mean continuous random variable ϵ . Under some regularity conditions, the risk (the expected square difference between true conditional expectation function $E[Y|x]$ and estimated function \hat{f}) are obtained:

$$\text{risk of LS} \quad : \quad \frac{1}{n} (d+1) \text{Var}[\epsilon],$$

$$\text{risk of RRAT(1)} \quad : \quad \frac{1}{n} (\text{Var}[\epsilon] + d \frac{p_1(1-p_1)}{P_{\epsilon}^2(F_{\epsilon}^{-1}(p_1))}).$$

The theorem and the proof are given in [10].

From the above result, RRAT(1)'s risk is less than LS's if and only if $\frac{p_1(1-p_1)}{P_{\epsilon}^2(F_{\epsilon}^{-1}(p_1))} < \text{Var}[\epsilon]$. For instance, as we have also verified numerically, if ϵ is *log-normal* (see sec. 4.) RRAT beats LS regression in average when $P_{\mu}[\epsilon] > 0.608$ (recall that for symmetric distributions $P_{\mu}[\epsilon] = 0.5$).

4. Numerical Experiments

To investigate and demonstrate the performances of RRAT, we did several types of numerical experiments. Due to limited space, we describe here only a brief summary (more details in [10]). In subsection 4.1 we compare RRAT with M-estimators and LS. In subsection 4.2

we illustrate the performances of RRAT for several data generating classes through comparisons with LS. In subsection 4.3 we compare RRAT with transformation techniques summarized in subsection 2.2.

4.1 Comparison with M-estimators

We compared RRAT with several M-estimators and LS, in order to see how they behave under asymmetric error.

[data generating process]

In each experiment, $N = 1000$ training data pairs (x_i, y_i) were generated by

$$x_i \sim U[0, 1]^d, \quad d = 2, \quad (19)$$

$$y_i = a_0 + a_1x_{i1} + a_2x_{i2} + \epsilon_i, \quad (20)$$

where $a_j \sim U[0, 1], j = 0, 1, 2$ and ϵ_i is from a *shifted log-normal distribution*, given by

$$p(\epsilon) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left\{-\frac{(\log \epsilon - \theta_1)^2}{2\theta_2}\right\}. \quad (21)$$

Two parameters θ_1, θ_2 were set so that $E[\epsilon] = 0$ and $P_\mu[\epsilon] \in \{0.55, 0.60, \dots, 0.95\}$, which allowed us to try several degrees of asymmetric errors (we tried only positive skews without loss of generality.).

[models and their parameter estimations]

We tried several M-estimators: median, Cauchy [1], Huber [2] and Tukey [4] estimators. In the implementations of the latter three M-estimators, we used the true scale parameter (standard deviation) for simplicity ⁶.

In each of M, LS and RRAT estimators, we used a well-specified two dimensional affine model. Parameter estimations were implemented by the conjugate gradient method for M-estimators and the 1st step of RRAT. LS and the 2nd step of RRAT were computed analytically.

RRAT(1) with $p_1 = 0.50$ was tried. The function f_c in the 2nd step of RRAT are assumed to be well specified, i.e. f_c has only a single additive parameter.

[evaluation and results]

We measured the performances with mean squared error (MSE) on 1000 *noise-free* test samples (without the error term of eq. (20)), i.e. against the true conditional expectation. In each experimental setting, 500 experiments are repeated using completely independent data sets. The statistical comparisons between methods are given by a Wilcoxon signed rank test (In this section, we use the term “significant” for 0.01 level.) The significant results are summarized in Fig. 4 and a box below.

⁶In reality, we also have to estimate scale parameter because we do not know it.

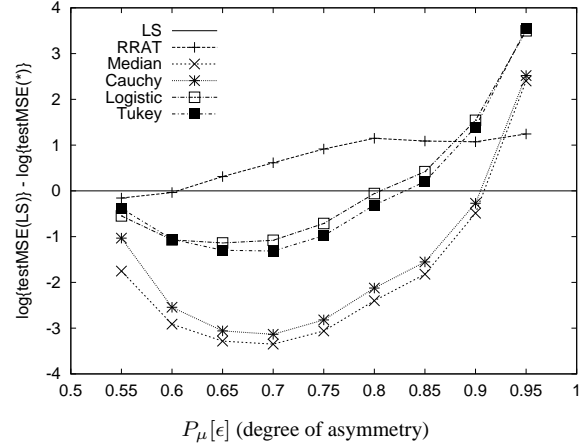


Figure 4: Performances of RRAT and M-estimators compared with LS: The horizontal axis ($P_\mu[\epsilon]$) implies the degree of asymmetry (the larger $P_\mu[\epsilon]$, the more asymmetric). The values above zero line suggest the estimator was better than LS. Note, under intermediate asymmetric error distributions, which are practically important, M-estimators were worse even than the LS, while RRAT worked better. When $0.55 \leq P_\mu[\epsilon] \leq 0.75$, LS was significantly (0.01 level) better than any M-estimators and when $0.55 \leq P_\mu[\epsilon] \leq 0.85$, RRAT was significantly (0.01 level) better than any M-estimators.

Main Results

- When error distribution is almost Normal,

$$\text{Error(LS)} < \text{Error(RRAT)} < \text{Error(M)}$$
- When error distribution has intermediate degree of asymmetry,

$$\text{Error(RRAT)} < \text{Error(LS)} \ll \text{Error(M)}$$
- When error distribution has high degree of asymmetry,

$$\text{Error(M)} < \text{Error(RRAT)} \ll \text{Error(LS)}$$

[discussion]

The results for large $P_\mu[\epsilon]$ were not what we initially expected. We **conjecture** that in this case the large *biases* in M-estimators were not as strong as the effects of variance due to *non-robustness* of the LS estimator in the 2nd step in RRAT. Note, however, that the variances of a log-Normal whose $P_\mu[\epsilon] \in \{0.85, 0.90, 0.95\}$ are respectively 5×10^3 , 5×10^5 and 3×10^7 . In this series of experiments, the expectation ranges in $[0.0, 3.0]$, which means that the noise scale is $10^3 \sim 10^7$ times larger than the deterministic part of the system. For many applications with a signal to noise ratio which is not *extremely*

small (including such noisy cases as the insurance data studied here), the M-estimators are worse than the LS estimator, because of bias.

4.2 Performances on Several Conditions

A series of experiments were designed to investigate the performance of RRAT compared with LS on several data generating processes.

[data generating process]

We used the generalized data generating process given by

$$x_i \sim U[0, 1]^d, \quad d \in \{2, 5, 10\} \quad (22)$$

$$y_i = f(x_i) + g\{f(x_i)\} \cdot \epsilon_i, \quad (23)$$

where f is among 2,5 and 10 dimensional affine functions and 6 inputs non-linear function⁷, and g is among $\{g(z) = 1, g(z) = 0.1z, g(z) = 0.1z + 1\}$ each corresponding to additive, multiplicative and combination of them in their noise structures.

[models and their parameter estimations]

When f is affine, we used well-specified models and when f is non-linear, we used a neural network (NN) for both LS and quantile regressions in the 1st step of RRAT. Only RRAT(1) with $p_1 = 0.5$ was tried and the f_c in the 2nd step of RRAT was correctly specified as described in theorems 1.1 ~ 1.3. Parameter optimization was performed by conjugate gradients for NN and quantile regressions with affine models. Others were analytically computed.

[evaluation and results]

The evaluation is same as the previous series of experiments. The number of independent experiments is 500 for the affine f and 50 for the non-linear f . The results are summarized in Figs 5, 6 and 7 and a box below.

Main Results

- In most cases, as the degree of asymmetry increases, RRAT worked relatively better.
- In most cases, as the number of parameters in f increases, RRAT worked relatively better.
- When f was affine, the experimental results almost coincide with the theoretical asymptotic analysis.

[discussion]

⁷ $f(x) = 10 \sin(\pi x_1 x_2) + 20 \sin(x_3 - 0.5)^2 + 10x_4 + 5x_5$ (does not depend on x_6).

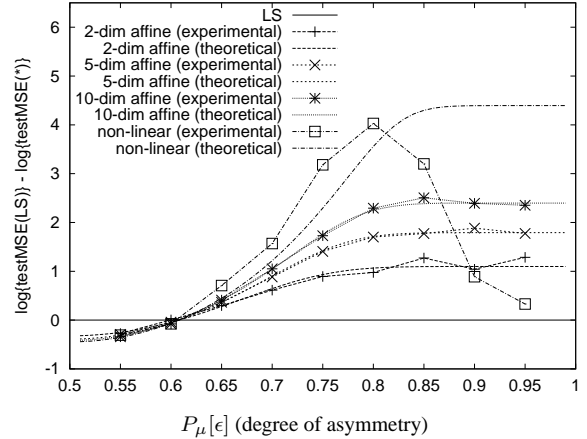


Figure 5: Performances of RRAT compared with LS with additive noise (the theoretical curves derived from asymptotic analysis in subsection 3.5 are also indicated). Note that as the degree of asymmetry increases, RRAT worked better. Note also that as the number of parameters increases, RRAT worked better.

The relative improvements decrease or vanish when the predictor is non-linear and $P_\mu[\epsilon]$ is fairly large. The possible explanation on this apparent inconsistency with result in subsec. 3.5 are: the second-order approximation or full-rank assumption were not satisfied, the NN approximation violates the assumption of result in subsec. 3.5 that the model is well-specified, and/or the number of samples is not large enough to be consistent with asymptotic behavior.

4.3 Comparison with Transformations

In this series of experiments, we compared RRAT with transformation techniques described in subsec. 2.2. In particular, the *naive transformation approach* (see below) and smearing estimate [6] were considered.

[data generating process]

In each experiment, $N = 100$ training data pairs (x_i, y_i) were generated by

$$x_i \sim U[0, 1]^d, \quad d \in \{1, 2, 5\} \quad (24)$$

$$y_i = h^{-1}\{f(x_i) + \epsilon_i, \lambda\}, \quad \lambda = 0, \quad (25)$$

where h^{-1} is the inverse function of Box-Cox transformation given by eq. (2), and ϵ_i is from standard Normal. f was either of

$$f(x) = 1 + \sum_{j=1}^d x_j \quad \text{or} \quad f(x) = \sin(2\pi \sum_{j=1}^d x_j). \quad (26)$$

[models and their parameter estimations]

The transformation technique was implemented in line with the three steps described in subsec. 2.2. In the 1st

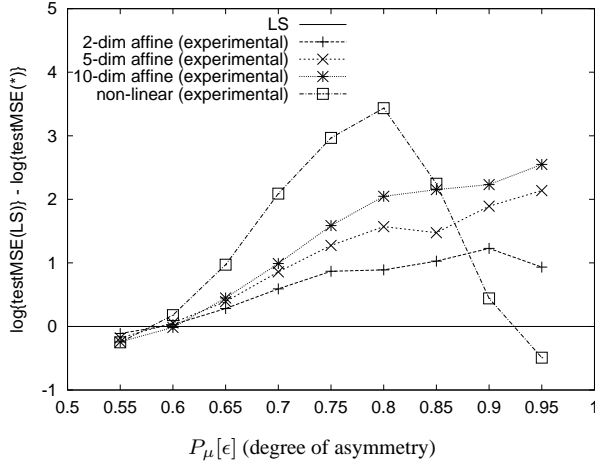


Figure 6: Performances of RRAT compared with LS with multiplicative noise

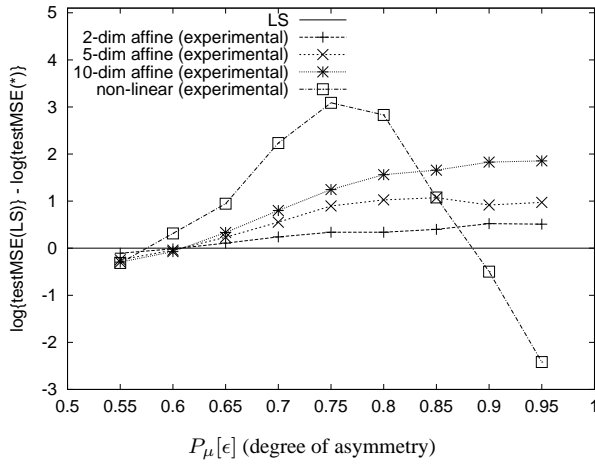


Figure 7: Performances of RRAT compared with LS with combination of additive and multiplicative noise structures

step, the λ was selected based on residuals [11], [12]⁸. In the 2nd step, regression function \hat{f}_λ was estimated by LS for transformed data. In the 3rd step, the naive transformation approach only computes the inverse transformation of the response of \hat{f}_λ . The smearing estimate was computed by eq. (5).

In the naive transformation approach, the Smearing Estimate, RRAT and LS, a neural network (NN) was used for the regression models. NN's generalization property was controlled by cross validation w.r.t the number of hidden units and weight decay penalty terms. Parameter optimization was performed with the conjugate

⁸At first, several candidates of λ were chosen. For each candidate of λ , data were transformed and the regression function \hat{f}_λ was estimated by that transformed data. Best λ , which was expected to yield symmetric error distribution, was selected according to the empirical third moment of the residuals on \hat{f}_λ .

gradients method.

The hyper-parameters of RRAT was also determined by cross-validation. In particular three quantile regressions with $p_1 = 0.25$, $p_2 = 0.50$ and $p_3 = 0.75$ were estimated and the best subset of them was selected by cross-validation.

[evaluation and results]

We evaluated estimators by MSE on 10,000 test data. We repeated 100 independent experiments. The significance of the differences was examined by pair-wise t -tests. The results are summarized in Tab. 2 and a box below.

Table 2: Experimental Comparison of RRAT, Smearing Estimate and Naive Transformation Approach with LS: \circ ($*$) denotes that the estimator was significantly better (worse) than LS in 0.05 level

$d = 1$				
estimator	affine		sin	
	MSE	p -val	MSE	p -val
LS	1.00	-	1.00	-
RRAT	0.98	.048 \circ	0.98	.003 \circ
Sme. Est.	0.99	.197	0.99	.259
Naive	1.34	.000 $*$	1.41	.000 $*$

$d = 2$				
estimator	affine		sin	
	MSE	p -val	MSE	p -val
LS	1.00	-	1.00	-
RRAT	0.94	.000 \circ	0.97	.000 \circ
Sme. Est.	.98	.187	0.96	.000 \circ
Naive	1.25	.000 $*$	1.21	.000 $*$

$d = 5$				
estimator	affine		sin	
	MSE	p -val	MSE	p -val
LS	1.00	-	1.00	-
RRAT	0.98	.048 \circ	0.95	.006 \circ
Sme. Est.	0.99	.197	0.95	.019 \circ
Naive	1.34	.000 $*$	1.29	.000 $*$

Main Results

- RRAT was **always significantly better** than LS.
- The Smearing Estimate was **sometimes significantly better** than LS.
- The naive transformation approach was **always significantly worse** than LS

[discussion]

The experimental setting employed here was not fair in the sense that the data generating process given by eqs. (24) and (25) was *well specified* for transformation approaches. In spite of such an *unfair* condition for RRAT, the latter worked better than the Smearing Estimate. It became clear from further investigation that the Smearing Estimate performs poorly when the transformation parameter λ is not correctly estimated. This is probably because the Smearing Estimate uses empirical residuals under the assumption that they are independent of x (homoscedastic). This assumption would not be satisfied when transformation parameter was not correctly estimated.

5. Insurance Premium Estimation

We have applied the proposed method to a problem in automobile insurance pure premium estimation: estimate the risk of a driver given his/her profile (age, type of car, etc.). In this application we want the conditional expectation because it corresponds to the average claim amount that customers with a certain profile will make in the future and this expectation is the main determinant of the cost of insurance.

One of the challenges in this problem is that the premium must take into account the tiny fraction of drivers who cause serious accidents and file large claim amounts. That is, claim amounts has random element that is unexplainable by customer’s profiles, which distributes with an asymmetric heavy tail extending out toward positive values. We used real data from a North American insurance company.

The number of input variables is 39, all discrete except one. The discrete variables are one-hot encoded, yielding input vectors with 266 elements. We repeated the experiment 10 times using each time an independent data set, by randomly splitting a large data set with 150,000 samples into 10 independent subsets with 15,000 samples. Each subset is then randomly split in 3 equal subsets with 5000 samples respectively for training, validation (model selection), and testing (model comparison). We tried several versions of RRAT (w.r.t $n, p_1, \dots, p_n, \mathcal{F}_q$). The results with affine models for \mathcal{F}_q are summarized in Table 3 and those with neural network models for \mathcal{F}_q are summarized in Table 4.

Table 3: Experimental comparison of RRAT vs LS (p-values) on linear predictors: The figures in the table are p -values from the Wilcoxon signed rank test, where ‘*’ (‘**’) denotes LS regression being significantly better than RRAT at 0.05 (0.01) level, ‘-’ denotes no significant difference between them and ‘o’ (‘oo’) denotes RRAT being significantly better than LS regression at 0.05 (0.01) level. **Note RRAT(n), $n \geq 2$, always beating LS regression.**

RRAT(1), $f_c(f_{p_1}) = c_0 + f_{p_1}$		
$p_1 = .2$ 8.30×10^{-3} **	$p_1 = .5$ 2.97×10^{-2} *	$p_1 = .8$ 2.34×10^{-2} o
RRAT(1), $f_c(f_{p_1}) = c_0 + c_1 f_{p_1}$		
$p_1 = .2$ 1.01×10^{-1} -	$p_1 = .5$ 3.72×10^{-2} o	$p_1 = .8$ 3.46×10^{-3} oo
RRAT(1), $f_c(f_{p_1}) = c_0 + c_1 f_{p_1} + c_2 f_{p_1}^2$		
$p_1 = .2$ 1.93×10^{-1} -	$p_1 = .5$ 4.67×10^{-3} oo	$p_1 = .8$ 2.53×10^{-3} oo
RRAT(2), $f_c(f_{p_1}, f_{p_2}) = c_0 + c_1 f_{p_1} + c_2 f_{p_2}$		
$p_1 = .2, p_2 = .5$ 1.42×10^{-2} o	$p_1 = .2, p_2 = .8$ 3.46×10^{-3} oo	$p_1 = .5, p_2 = .8$ 3.46×10^{-3} oo
RRAT(3), $f_c(f_{p_1}, f_{p_2}, f_{p_3}) = c_0 + c_1 f_{p_1} + c_2 f_{p_2} + c_3 f_{p_3}$		
$p_1 = .2, p_2 = .5, p_3 = .8$ 2.53×10^{-3} oo		
Best model on validation set		
2.53×10^{-3} oo		

Main Results (on both linear and NN)

- The performance of RRAT(1) depends on the choice of p_1 , and when the chosen p_1 were appropriate, they worked better than LS.
- The performance of RRAT(n), $n \geq 2$ always yielded the better performances than LS.
- The best RRAT based on validation set was significantly (0.01 level) better than LS.

6. Conclusion and Future Work

In this paper, we proposed a regression algorithm, RRAT, which is tailored to regression problems under asymmetric heavy-tail errors. Experiments on synthetic data as well as real insurance data confirmed the clear advantages of the approach.

Our future work includes a study of how RRAT would perform when the regression problems in hand violates the condition in eq. (14). In these conditions it would be interesting to use a linear combination of the quantile regressions as suggested in eq. (17).

Table 4: Experimental Comparison of RRAT vs LS (p-values) on NN predictors: see the caption of Table 3

RRAT(1), $f_c(f_{p_1}) = c_0 + f_{p_1}$		
$p_1 = .2$	$p_1 = .5$	$p_1 = .8$
$1.83 \times 10^{-2} *$	$1.66 \times 10^{-1} -$	$1.09 \times 10^{-2} \circ$
RRAT(1), $f_c(f_{p_1}) = c_0 + c_1 f_{p_1}$		
$p_1 = .2$	$p_1 = .5$	$p_1 = .8$
$3.99 \times 10^{-1} -$	$4.67 \times 10^{-3} \circ\circ$	$1.83 \times 10^{-2} \circ$
RRAT(1), $f_c(f_{p_1}) = c_0 + c_1 f_{p_1} + c_2 f_{p_1}^2$		
$p_1 = .2$	$p_1 = .5$	$p_1 = .8$
$4.39 \times 10^{-1} -$	$4.67 \times 10^{-3} \circ\circ$	$1.42 \times 10^{-2} \circ$
RRAT(2), $f_c(f_{p_1}, f_{p_2}) = c_0 + c_1 f_{p_1} + c_2 f_{p_2}$		
$p_1 = .2, p_2 = .5$	$p_1 = .2, p_2 = .8$	$p_1 = .5, p_2 = .8$
$3.46 \times 10^{-3} \circ\circ$	$1.42 \times 10^{-2} \circ$	$6.26 \times 10^{-3} \circ\circ$
RRAT(3), $f_c(f_{p_1}, f_{p_2}, f_{p_3}) = c_0 + c_1 f_{p_1} + c_2 f_{p_2} + c_3 f_{p_3}$		
$p_1 = .2, p_2 = .5, p_3 = .8$		
$8.30 \times 10^{-3} \circ\circ$		
Best model on validation		
$4.67 \times 10^{-3} \circ\circ$		

References

- [1] P.J. Huber. *Robust Statistics*. John Wiley & Sons Inc., 1982.
- [2] P.J. Huber. Robust regression: Asymptotics, conjectures and monte carlo. *Ann. Stat.*, 1:799–821, 1973.
- [3] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons Inc., 1987.
- [4] A.E. Beaton and J.W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16:147–185, 1974.
- [5] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Ser. B*, 26:211–246, 1964.
- [6] N. Duan. Smearing estimate: A nonparametric re-transformation method. *Journal of the American Statistical Association*, 78(383):605–610, 1983.
- [7] M. S. Williams. A regression technique accounting for heteroscedastic and asymmetric errors. *Journal of Agricultural, Biological, and Environmental Statistics*, 2(1):108–129, 1997.
- [8] K. Shimizu and K. Iwase. Uniformly minimum variance unbiased estimation in lognormal and related distributions. *Communications in Statistics - Theory and Methods*, 10(11):1127–1147, 1981.
- [9] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.
- [10] I. Takeuchi, Y. Bengio, and T. Kanamori. Robust regression with asymmetric heavy-tail noise distributions. *Neural Computation*, 14(10):in press, 2002.
- [11] J. M. G. Taylor. Measures of location of skew distributions obtained through box-cox transformations. *Journal of American Statistical Association*, 80(390):427–432, 1985.
- [12] D. V. Hinkley. On power transformations to symmetry. *Biometrika*, 62:101–111, 1975.