*LETTER* **Communicated by John Platt**

# Robust Regression with Asymmetric Heavy-Tail Noise Distributions

**Ichiro Takeuchi**
*takeuchi@pa.info.mie-u.ac.jp*
*Department of Information Engineering, Mie University, Tsu 514-8507, Japan*

**Yoshua Bengio**
*bengioy@iro.umontreal.ca*
*Université de Montréal, DIRO, Montréal, Québec, Canada*

**Takafumi Kanamori**
*kanamori@is.titech.ac.jp*
*Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8552, Japan*

**In the presence of a heavy-tail noise distribution, regression becomes much more difficult. Traditional robust regression methods assume that the noise distribution is symmetric, and they downweight the influence of so-called outliers. When the noise distribution is asymmetric, these methods yield biased regression estimators. Motivated by data-mining problems for the insurance industry, we propose a new approach to robust regression tailored to deal with asymmetric noise distribution. The main idea is to learn most of the parameters of the model using conditional quantile estimators (which are biased but robust estimators of the regression) and to learn a few remaining parameters to combine and correct these estimators, to minimize the average squared error in an unbiased way. Theoretical analysis and experiments show the clear advantages of the approach. Results are on artificial data as well as insurance data, using both linear and neural network predictors.**

## 1 Introduction

In a variety of practical applications, we often find data distributions with an asymmetric heavy tail (see the shape of the distribution in Figure 1 (ii)). Modeling data with such an asymmetric heavy-tail distribution is essentially difficult because outliers, which are sampled from the tail of the distribution, have a strong influence on parameter estimation. When the distribution is symmetric (around the mean), the problems caused by outliers can be reduced using robust estimation techniques (Huber, 1982; Hampel, Ronchetti, Rousseeuw, & Stahel, 1986), which basically intend to ignore or

 

put less weights on outliers. Note that these techniques are highly biased for asymmetric distributions; most outliers are on the same side of the mean, so downweighting them introduces a strong bias on its estimation.

We are concerned in this article with the problem of linear or nonlinear regression, that is, estimating the conditional expectation $E[Y \mid X]$, in the presence of asymmetric, heavy-tail, and unknown noise. Regression problems also suffer from the effect of outliers when the distribution of the noise (the variations of the output variable $Y$ that are not explainable by the input variable $X$) has an asymmetric heavy tail. As in the unconditional case, the robust methods that symmetrically downweight outliers (Huber, 1982; Hampel et al., 1986) are highly biased for asymmetric noise distributions.

In this article, we propose a new robust regression method that can be applied to data with an asymmetric, heavy-tail, and unknown noise distribution and does not suffer from the above bias problem. We present a theoretical analysis, numerical experiments with synthetic data, and an application to automobile insurance premium estimation. These experiments show when and how the proposed method generalizes significantly better than least squares and M estimators.

Throughout the article, we use the following notations: $X$ and $Y$ for the input and the output random variables, $F_W(\cdot)$ and $P_W(\cdot)$ for the cumulative distribution function (cdf) and the probability density function (pdf) of random variable $W$.

## 2 Traditional Robust Methods

Let us first consider the problem of estimating from a finite sample the unconditional mean $E[Y]$ of a heavy-tail distribution (i.e., the density decays slowly to zero when going toward $\infty$ or $-\infty$). The empirical average may be a poor estimator here because a few points will be sampled from the tails and may have very different values, thereby introducing a great deal of variability (from sample to sample) in the empirical average. In the case of a symmetric distribution, we can downweight or ignore the effect of these outliers in order to reduce this variability greatly. For example, the median estimator is much less sensitive to outliers than the empirical average for heavy-tail distributions.

This idea can be generalized to conditional estimators; for example, one can estimate the conditional mean $E[Y \mid X]$ from the empirical conditional median by minimizing absolute errors,

$$\hat{f}_{0.5} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_i |y_i - f(x_i)|, \tag{2.1}$$

where $\mathcal{F}$ is a set of functions (e.g., a class of neural networks), $\{(x_i, y_i), i = 1, 2, \ldots, N\}$ is the training sample, and a hat ($\wedge$) denotes estimation on data. Minimizing the above over $P(X, Y)$ and a large enough class of functions

estimates the conditional median $f_{0.5}$, that is, $P(Y < f_{0.5}(X) \mid X) = 0.5$. For regression problems with a heavy-tail noise distribution, the estimated conditional median is much less sensitive to outliers than least squares (LS) regression (which provides an estimate of the conditional average of $Y$ given $X$). However, we want to tackle problems where the goal is to estimate the conditional mean, not the conditional median, and where they differ.

In the presence of thick tails, M estimators (Huber, 1973) are among the most successful robust estimators for regression.[1] The basic idea of M estimators is to use the minimization schema,

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_i \rho(y_i - f(x_i)) \tag{2.2}$$

where $\rho$ is a loss function, with a unique minimum at zero, which is generally chosen symmetric ($\rho(-t) = \rho(t)$) when the actual noise distribution is unknown. As particular cases, $\rho(t) = \frac{1}{2}t^2$ yields the LS estimator, and $\rho(t) = |t|$ yields the median estimator, equation 2.1.

Unfortunately, these robust estimators that symmetrically downweight the influence of outliers 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.

In order to provide some quantitative evidence on this, we show in Table 1 a small Monte Carlo study on unconditional mean estimation. The values in the table are mean (variance) of 100 trials of estimation each with 10 samples, by LS, median, Cauchy (Huber, 1982), Huber (Huber, 1973), and Tukey's biweight (Beaton & Tukey, 1974) estimators. The samples are from a positive-skewed asymmetric distribution that is shifted so its mean equals zero.[2] Table 1 clearly illustrates the negative biases in all but LS estimators. Those negative biases are due to the loss functions (see equation 2.2) of the robust estimators, which in this case put less weight on data from the positive side of the distribution.

As the example suggests, traditional robust estimators are designed under an assumption of symmetric distribution. If we use them with asymmetric distribution, they yield a strong bias. In unconditional estimation problems under asymmetric distribution, it has been proved that the LS estimator (sample average) is the "best" in the sense of being a uniformly minimum variance unbiased estimator (Lehmann, 1983). For many regres-

[1] In the context of robust regression, many efforts have been devoted to deal with outliers in the input variable $X$, called leverage points, as well as those in the output variable $Y$. In this article, we do not consider outliers in $X$ and therefore do not refer to those techniques, including bounded-influence estimators (Krasker & Welsch, 1982) and S estimators (Rousseeuw & Leroy, 1987).

[2] Log-normal distributions (see section 4), shifted so their mean equals zero and their $p_\mu$, a degree of asymmetry introduced in the next section, are 0.65 (low), 0.75 (intermediate), and 0.85 (high), respectively.

Table 1: A Comparison of Estimators on Unconditional Mean Estimation Problems Under Asymmetric Distribution.

| | Degree of Asymmetry | | |
|---|---|---|---|
| Estimator | 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) |

Notes: The values are mean (variance) of 100 trials of estimation each with 10 samples by several estimators. The samples are from a positively skewed log-normal distribution (see note 2 and section 4), which is shifted as their means equal zero. Note that all but LS estimators are negatively biased, and the variances of LS estimator are much larger than the others.

sion problems as well, under asymmetric unknown noise distributions, the straightforward use of "robust" loss functions (see equation 2.2) is generally not helpful, as confirmed in section 4.

## 3 Robust Regression for Asymmetric Tails

**3.1 Motivation.** We saw 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 call its order $p_\mu$: for a distribution $P(Y)$, we define $p_\mu \triangleq F_Y(E[Y]) = P(Y < E[Y])$. Note that $p_\mu > 0.5$ ($< 0.5$) suggests $P(Y)$ is positively (negatively) skewed. When the distribution is symmetric, $p_\mu = 0.5$. Figure 1 illustrates this idea. For regression problems with an asymmetric noise distribution, we may extend median regression equation 2.1 to $p$th quantile regression (Koenker & Bassett, 1978), which estimates the conditional $p$th quantile; that is, we would like $P(Y < \hat{f}_p(X) \mid X) = p$,

$$\hat{f}_p = \underset{f \in \mathcal{F}_q}{\operatorname{argmin}} \left\{ \sum_{i:\, y_i \geq f(x_i)} p|y_i - f(x_i)| + \sum_{i:\, y_i < f(x_i)} (1-p)|y_i - f(x_i)| \right\}, \quad (3.1)$$

where $\mathcal{F}_q$ is a set of quantile regression functions. One could be tempted to obtain a robust regression with asymmetric noise distributions by estimating $f_{p_\mu}$ instead of $f_{0.5}$. But what is $p_\mu$ for regression problems? It must be defined as

$$p_\mu(x) \triangleq F_{Y|X}(E[Y \mid X = x]) = P(Y < E[Y \mid X] \mid X = x).$$

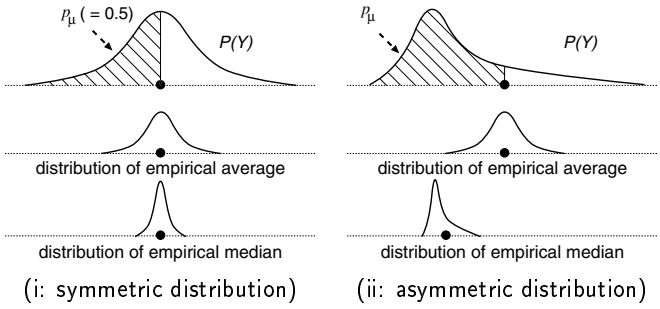


Figure 1: The schematic illustration of empirical averages and empirical medians for (i) symmetric distribution and (ii) asymmetric distribution: Distributions of a heavy-tail random variable $Y$, its empirical average and empirical median, and their expectations (black circles). In *i*, those expectations coincide, while in *ii* they do not. As indicated by the areas of slanting lines, we define $p_\mu$ as the order at which the quantile coincides with the mean. In *i*, $p_\mu = 0.5$.

So the above idea raises three problems that we address with the algorithm proposed in section 3.2: (1) $p_\mu(x)$ of $P(Y \mid X)$ may depend on $x$ in general; (2) unless the noise distribution is known, $p_\mu$ itself must be estimated, which may be as difficult as estimating $E[Y \mid X]$; and (3) if the noise density at $p_\mu$ is low (because of the heavy tail and large value of $p_\mu$), the estimator in equation 3.1 may itself be very unstable. (See Figure 2 and the discussion in the section 3.2.)

**3.2 Algorithm.** To overcome the difficulties raised above, we propose a new robust regression algorithm that can be applied to data with an unknown but asymmetric and heavy-tail noise. The main idea is to learn most of the parameters of the model using conditional quantile estimators (which are biased but robust) and to learn a few remaining parameters to combine and correct these estimators—hence the name *robust regression for asymmetric tails* (RRAT) algorithm:

**Algorithm RRAT($n$)**

**Input**: data pairs $\{(x_i \in \mathbb{R}^d, y_i \in \mathbb{R})\}$, and hyperparameters: number of quantile regressors $n$, quantile orders $(p_1, \ldots, p_n)$, function classes $\mathcal{F}_q$ and $\mathcal{F}_c$.

(1) Fit $n$ quantile regressions at orders $p_1, p_2, \ldots p_n$, each as in equation 3.1, yielding functions $\hat{f}_{p_1}, \hat{f}_{p_2}, \ldots, \hat{f}_{p_n}$, with $\hat{f}_{p_i}$: $\mathbb{R}^d \to \mathbb{R}$, with $\hat{f}_{p_i} \in \mathcal{F}_q$.

(2) Fit a least-squares regression (not necessarily linear) with inputs $q(x_i) = (\hat{f}_{p_1}(x_i), \ldots \hat{f}_{p_n}(x_i))$ and targets $y_i$, yielding a function $\hat{f}_c$: $\mathbb{R}^n \to \mathbb{R}$, with $\hat{f}_c \in \mathcal{F}_c$.

**Output**: conditional expectation estimator $\hat{f}(x) = \hat{f}_c(q(x))$.

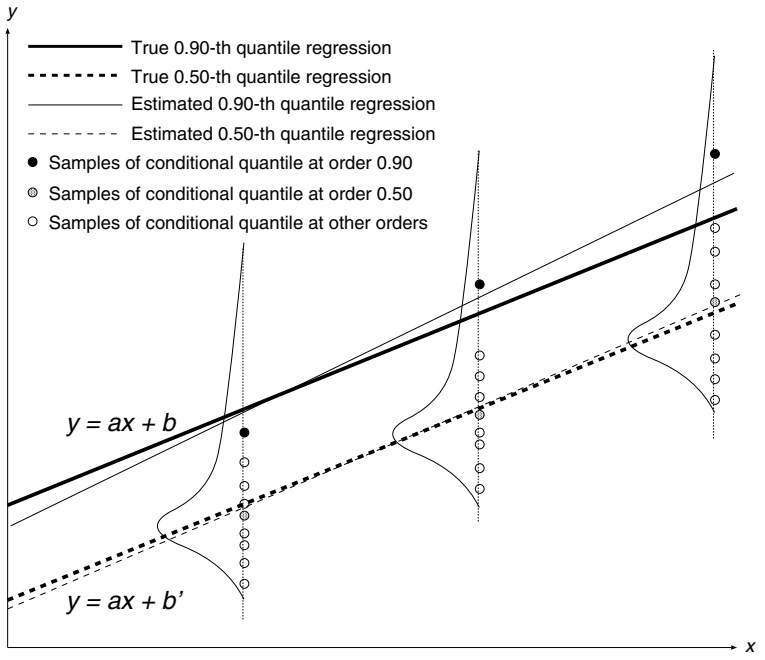


Figure 2: An introductory example of RRAT for a simple scalar linear regression problem with additive noise: $Y = aX + b + Z$, where $Z$ is a zero-mean noise random variable (independent of $X$) with an asymmetric heavy-tail distribution whose (unknown) $p_\mu$ is 0.9 and $a$ and $b$ are parameters to estimate. With additive noise, $p_\mu$ is constant with regard to $X$: $P(Y < E[Y \mid X] \mid X) = P(aX + b + Z < aX + b) = P(Z < 0) = p_\mu$. We could estimate the $p_\mu$th quantile regression (thick solid line), but it might be unreliable for large $p_\mu$ at low densities. Instead, consider a simple RRAT with $n = 1$, $p_1 = 0.5$, $\hat{f}_{p_1}$ linear in $x$, and $\hat{f}_c$ has just one additive parameter. $f_{p_1}$ (thick dotted line) has the same slope parameter $a$ as $f_{p_\mu}$ (but a different intercept $b'$ and $b$), but $\hat{f}_{p_\mu}$ (thin solid line) is more variable than $\hat{f}_{p_1}$ (thin dotted line) because of the variabilities of empirical conditional quantiles (small circles at a few values of $x$ with underlying pdf's). With RRAT, (1) estimate $a$ with $\hat{f}_{p_1}$, (2) keeping $\hat{a}$ fixed; then estimate $b$ by LS regression.

An outer model selection loop is often required in applications in order to select hyperparameters such as $n$, $(p_1, \ldots, p_n)$ and capacity control for function classes $\mathcal{F}_q$ and $\mathcal{F}_c$. For example, this can be achieved with a validation set (used in the experiments), K-fold cross-validation, or the bootstrap.

An introductory example of a very simple version of the RRAT algorithm is illustrated in Figure 2. Some of the parameters are estimated through conditional quantile estimators $f_{p_1}, \ldots, f_{p_n}$ and they are combined and corrected

by the function $f_c$ in order to estimate $E[Y \mid X]$. In general, this method yields more robust regressions when the number of parameters required to characterize $\hat{f}_c$ is small compared to the number of parameters required to characterize the quantile regressions $f_{p_i}$ (this statement is justified in more detail in section 3.5). The intuitive explanation is that the $\hat{f}_{p_i}$ parameters are estimated robustly (quantile estimators), whereas the $\hat{f}_c$ parameters are not (LS estimator). That is also the reason that RRAT can beat LS: having most of its parameters estimated robustly decreases its variance.

Problem 2 above is dealt with by doing quantile regressions of orders $p_1, \ldots, p_n$ not necessarily equal to $p_\mu$. Problem 3 is dealt with if $p_1, \ldots, p_n$ are in high-density areas (where estimation will be robust). The issue raised with problem 1 will be discussed in section 3.3. Note that RRAT with a linear $f_c$ (without constant term) is similar to some L estimators (Bickel, 1973) (linear combination of order statistics). However, the coefficients of the second level, $f_c$, are not fixed but estimated, so RRAT might be considered an adaptive L estimator. Note that RRAT also bears some similarities with ensemble methods, but relies crucially on each of the combined functions to be estimated robustly (here a quantile regression). Unlike in most ensemble methods, the training set is used to estimate the parameters of the combination function $f_c$.

**3.3 An Applicable Class of Problems.** A surprising result is that with additive or multiplicative but otherwise arbitrary noise, it is sufficient to take a single quantile function ($n = 1$) with any chosen quantile $p_1$, and it is sufficient to take an affine (two-parameter) function $f_c$ in order to map the true quantile function to the true conditional expectation. In this section and the next, we explain and generalize this result.

A large class of regression problems for which RRAT works (in the sense that the analog of problem 1 is eliminated) can be described as follows:

$$Y = g_\mu(X) + Z g_\sigma(g_\mu(X)), \tag{3.2}$$

where $Z$ is a zero-mean random variable drawn from any form of (possibly asymmetric) continuous distribution, independent of $X$. The conditional expectation is characterized by an arbitrary function $g_\mu$ and the conditional standard deviation of the noise distribution is characterized by an arbitrary positive range function $g_\sigma$. Note that the regression in equation 3.2 is a subclass of heteroscedastic regression (Greene, 1997); the standard deviation of the noise distribution is not directly conditioned on $X$, but only on $g_\mu(X)$. This specification narrows the class of applicable problems, but equation 3.2 still covers a wide variety of noise structures, as explained later.

In equation 3.2, $E[Y \mid X] = E[g_\mu(X) + Z g_\sigma(g_\mu(X)) \mid X] = g_\mu(X)$, and $p_\mu$ of the distribution $P(Z)$ coincides with $p_\mu$ of $P(Y \mid X)$ and does not depend on $x$, that is, $P(Y < E[Y \mid X] \mid X) = P(Z g_\sigma(g_\mu(X)) < 0 \mid X) = P(Z < 0 \mid X) =$

$P(Z < 0)$. Since the conditional expectation $E[Y \mid X]$ coincides with the $p_\mu$th quantile regression of $P(Y \mid X)$, we have $g_\mu(x) \equiv f_{p_\mu}(x) = E[Y \mid x]$.

The following two theorems show that RRAT($n$) "works" when large enough classes of $\mathcal{F}_q$ and $\mathcal{F}_c$ are provided by guaranteeing the existence of the function $f_c$, which transforms the outputs of $p_i$th quantile regressions $f_{p_i}(x)$ $(i = 1, \ldots, n)$ into the conditional expectation $E[Y \mid x]$ for all $x$. The cases of $n = 1$ and $n = 2$ are explained in theorem 1 and theorem 2, respectively.

The applicable class of problems with only one quantile regression $f_{p_1}$, that is, RRAT(1), is smaller than the class satisfying equation 3.2, but it is very important for practical applications because it includes additive and multiplicative noise (see section 3.4).

**Theorem 1.** *If the noise structure is as in equation 3.2, then there exists a function $f_c$ such that $E[Y \mid X] = f_c(f_{p_1}(X))$, where $f_{p_1}(X)$ is the $p_1$th quantile regression, if and only if function $h(\bar{y}) = \bar{y} + F_Z^{-1}(p_1) \cdot g_\sigma(\bar{y})$ is strictly monotonic with respect to $\bar{y}$ (taken in the range of $g_\mu$, i.e., $\{E[Y \mid x] \mid \forall x\}$). (The proof is in appendix A.)*

With the use of two quantile regressions $f_{p_1}$ and $f_{p_2}$, we show that RRAT(2), covers the whole of the class satisfying equation 3.2.

**Theorem 2.** *If the noise structure is as in equation 3.2 and*

$$p_1 \neq p_\mu, p_2 \neq p_\mu, \tag{3.3}$$

$$p_1 \neq p_2, \tag{3.4}$$

*then there exists a function $f_c$ such that $E[Y \mid X] = f_c(f_{p_1}(X), f_{p_2}(X))$, where $f_{p_i}(X)$ are the $p_i$-quantile regressions $(i = 1, 2)$. (The proof is in appendix B.)*

In comparison to theorem 1, we see that when using $n = 2$ quantile regressions, the monotonicity condition can be dropped. We conjecture that even the assumption of noise structure in equation 3.2 can be dropped when combining a sufficient number of quantile regressions. When $Y > 0$, we have that

$$E[Y \mid X] = \int_0^\infty P(Y > y \mid X)\, dy = \int_0^1 P(Y > f_q(X) \mid X) \frac{df_q(X)}{dq}\, dq$$

by substituting $y = f_q(X)$ s.t. $P(Y > f_q(X) \mid X) = q$. The integral can be unbiasedly estimated by a number of numerical integration schemes (the simplest being uniform spacing of $q$; 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 $(f_{q_{i+1}}(X) - f_{q_{i-1}}(X))/(q_{i+1} - q_{i-1})$.

However, generally increasing $n$ in RRAT may add more complexity (and parameters) to $f_c$, thereby reducing the robustness gains brought by RRAT.

**3.4 Additive and Multiplicative Noise.** Consider the case where $g_\sigma$ (in equation 3.2) is affine and $g_\mu(x) \geq 0$ for all $x$,

$$Y = g_\mu(X) + Z \times (c + d g_\mu(X)), \tag{3.5}$$

where $c$ and $d$ are constants such that $c \geq 0, d \geq 0, (c, d) \neq (0, 0)$. The conditions of theorem 1 (including monotonicity of $h(y)$) are verified for additive noise ($d = 0$, e.g., $Y = E[Y \mid X] + Z$), multiplicative noise ($c = 0$, e.g., $Y = E[Y \mid X](1 + Z)$), or combinations of both ($c > 0, d > 0$) for any form of noise distribution $P(Z)$ (continuous and independent of $X$).

**Property 1.** *If the noise structure is affine (see equation 3.5) and $g_\mu(x) \geq 0$ for all $x$, then a linear function $f_c$ is sufficient to map $f_{p_1}$ to $f$, that is, only two parameters need to be estimated by least squares. (The proof is in appendix C.)*

Additive and multiplicative noise models cover a very large variety of practical applications,[3] so this result shows that RRAT(1) already enjoys wide applicability. The fact that a linear combination function $f_c$ is sufficient in the case of additive or multiplicative noise is encouraging. The previous conjecture suggests that a linear combination could be sufficient in more general cases, albeit with more quantile regressors.

**3.5 Asymptotic Behavior.** Let us call risk of an estimator $\hat{f}(X)$ the expected squared difference between $E[Y \mid X]$ and $\hat{f}(X)$ (expectation over $X, Y$, and the training set), in the limit of large sample size. Let us write $\hat{f}_c(\hat{f}_{p_1}(X))$ for the conditional expectation obtained by RRAT(1) with finite variance.

**Theorem 3.** *Consider the class of regression problems with $Y = f(X; \alpha^*) + \beta^* + Z$, where $f$ is an arbitrary function characterized by a set of parameters $\alpha^* \in \mathbb{R}^d$ (with the assumption that the second derivative matrix of $f(X; \alpha^*)$ with respect to $\alpha^*$ is full rank) and $\beta^* \in \mathbb{R}$ is a scalar parameter, and where $Z$ is the zero-mean noise random variable drawn from a (possibly asymmetric) heavy-tail distribution (independent of $X$). Also assume that the model ($\mathcal{F}_q$) used in the first step of RRAT is well specified. Then the risk of LS regression is given by $\frac{1}{n}(d+1)Var[Z]$, while the risk of RRAT(1) is given by $\frac{1}{n}(Var[Z] + d\frac{p_1(1-p_1)}{P_Z^2(F_Z^{-1}(p_1))})$. It follows that the risk of*

[3] The specification of the conditional expectations to be nonnegative is a trivial constraint because we can shift the output data. Also, note that in many practical applications with asymmetric heavy-tail noise distributions (including ours), the output range is nonnegative.

*RRAT(1) is less than that of LS regression if and only if* $\frac{p_1(1-p_1)}{P_Z^2(F_Z^{-1}(p_1))} < Var[Z]$. *(The proof is in appendix D.)*

For instance, as we have also verified numerically, if $Z$ is log-normal (see section 4) RRAT beats LS regression on average when $p_\mu > 0.608$ (recall that for symmetric distributions $p_\mu = 0.5$).

## 4 Numerical Experiments

To investigate and demonstrate the proposed algorithm, we did three types of numerical experiments. In section 4.1, we compare RRAT, M estimators, and the LS estimator. In section 4.2, we show when RRAT works; we illustrate the performances of RRAT under several classes of data-generating processes. In section 4.3, we clear up which RRAT works; we demonstrate the performances of RRAT under several choices of hyperparameters ($\mathcal{F}_q$, $\mathcal{F}_c, n, p_1, p_2, \ldots$). In Monte Carlo simulations, $N = 1000$ training pairs[4] $(x_i, y_i)$, $(i = 1, 2, \ldots, N)$ are generated as per equation 3.2 with:

$$x_i \sim U[0,1]^d, \tag{4.1}$$

$$y_i = g_\mu(x_i) + z_i g_\sigma(g_\mu(x_i)), \tag{4.2}$$

where $g_\mu$: $\mathbb{R}^d \to \mathbb{R}$, $g_\sigma$: $\mathbb{R} \to (0, \infty)$ and $z_i$ is a random sample from log-normal distribution $\mathrm{LN}(z_0, 0, \sigma^2)$ (Antle, 1985), $Z = z_0 + e^W$ where $W \sim N(0, \sigma^2)$. The location parameter $z_0$ is chosen so that $E[Z] = 0$, and the spread parameter $\sigma$ is chosen so that $Z$ has a prescribed $p_\mu$, that is, $p_\mu = P(Z < E[Z])$. We tried only positive skews ($p_\mu > 0.5$), without loss of generality.

We measured the performances with mean squared error (MSE) on 1000 noise-free test samples (without the second term on the right-hand-side of equation 4.2), against the true conditional expectation. In each experimental setting, a number of experiments are repeated using completely independent data sets. The statistical comparisons between methods are given by the Wilcoxon signed rank test.[5] (In this section, we use the term *significant* for the 0.01 level.)

**4.1 Experiment 1: Comparison.** We compared the performance of RRAT, M estimators, and the LS estimator to see how the robust estimators behave

[4] We also tried the same experiments with different sizes of training samples. The results in experiment 1 (section 4.1) highly depend on training sample size, which we detail later. On the other hand, the training sample size did not affect the comparative results in experiment 2 (in section 4.2) and experiment 3 (in section 4.3).

[5] We used a nonparametric test rather than a Student $t$-test because the normality assumption in the $t$-test does not hold in the presence of heavy tails.

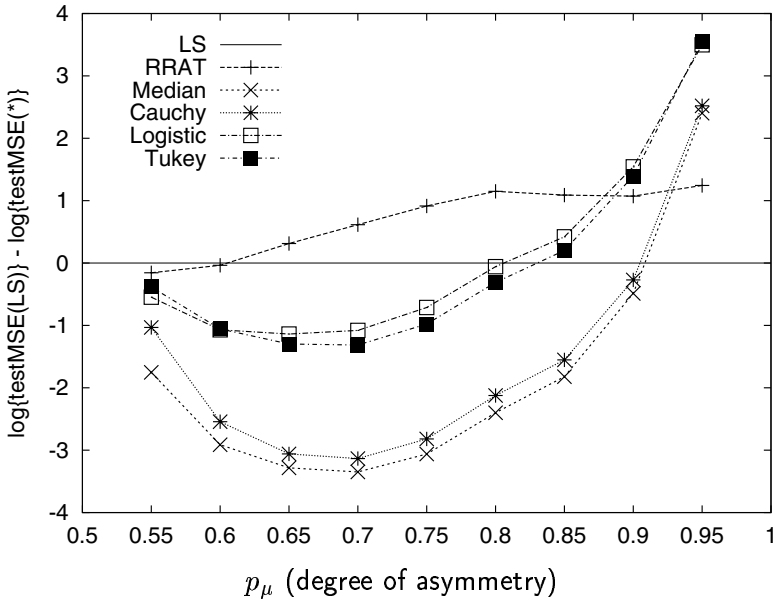


Figure 3: Performances of RRAT and M estimators compared with LS estimator. The horizontal axis ($p_\mu$) implies the degree of asymmetry (the larger $p_\mu$, the more asymmetric). The values above the zero line suggest the estimator was better than the LS estimator. Note that under intermediate asymmetric noise distributions, which are practically important, M estimators were worse even than the LS estimator, while RRAT worked better.

under asymmetric noise distributions with well-specified models. We tried several types of M estimators: median, Cauchy (Huber, 1982), Huber (Huber, 1973), and Tukey's biweight (Beaton & Tukey, 1974) estimators, each of which has its own loss function (see equation 2.2).

We employed a simple setting in this series of experiments. $g_\mu$ was affine with two inputs (with parameters sampled from $U[0, 1]$ in each experiment), $g_\sigma$ was chosen to be additive, and we tried positive skews, with $p_\mu \in \{0.55, 0.60, \ldots, 0.95\}$. We fixed $n = 1$ and $p_1 = 0.50$ in RRAT and used a well-specified $f_c$: $f_c(w) = c + w$. LS estimates and the second step in RRAT were analytically computed, and others were iteratively computed with a Polack-Ribiere conjugate gradient method. The number of independent experiments was 500. The results are summarized in Figure 3.

Figure 3 graphically shows the effect of asymmetric noise distributions on the performances of several M estimators compared with the LS estimator. To measure the degree of asymmetry in the horizontal axis, we employed $p_\mu$, where $p_\mu = 0.5$ implies symmetry and larger $p_\mu$ more asymmetry. In the

vertical axis, the comparative performances are measured by the difference in logarithm of test samples MSE where values above the zero line suggest the robust estimator performed better than LS estimator, and vice versa.

Figure 3 illustrates that when $p_\mu < 0.6$, LS was the best, RRAT was the second best, and the M estimators fared much worse. Under intermediate asymmetric noise distribution where $0.6 < p_\mu < 0.9$, RRAT was the best. For $0.6 < p_\mu < 0.8$, LS was the second best, and the M estimators were much worse. When $p_\mu$ was around 0.85, Huber and Tukey estimators beat LS regression but are slightly worse than RRAT. Finally, when the noise distribution is extremely asymmetric ($p_\mu = 0.95$), all four M estimators beat both RRAT and LS.

The results for large $p_\mu$ were not what we initially expected. We conjecture that in this case, the large biases in M estimators are not as strong as the effects of variance due to the nonrobustness of the LS estimator in the second step in RRAT.[6] Note, however, that the variances of a log-normal whose $p_\mu \in \{0.85, 0.90, 0.95\}$ are, respectively, $5 \times 10^3$, $5 \times 10^5$, and $3 \times 10^7$. 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 deterministic part of the system. For many applications with a signal-to-noise ratio that 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. In the following series of experiments, we compare RRAT only with the LS estimator.

Concerning statistical significance in Figure 3, when $0.55 \leq p_\mu \leq 0.75$, LS was significantly (0.01 level) better than any M estimator, and when $0.55 \leq p_\mu \leq 0.85$, RRAT was significantly (0.01 level) better than any M estimator. The comparison between LS and RRAT is detailed in the following sections.

**4.2 Experiment 2: When Does RRAT Work?** A series of experiments was designed to investigate the performance of RRAT with respect to $g_\mu$, $g_\sigma$, and $p_\mu$ in equation 4.2. We chose $g_\mu$ from either a class of affine functions with 2, 5, or 10 inputs (with parameters sampled from $U[0, 1]$ in each experiment), or from the class of 6-input nonlinear functions (described in Friedman, Grosse, & Suetzle, 1983).[7] $g_\sigma$ was chosen to be ei-

[6] We tried the same experiments with different sizes of training samples. The comparative performances of M estimators with respect to LS and RRAT highly depend on sample size. The overall relative performance of M estimators decreases with increasing sample size, and the $p_\mu$ threshold beyond which M estimators do better decreases with increasing sample size. On the other hand, the comparative performances between RRAT and LS did not depend on sample size. These results are consistent with our conjecture that RRAT wins over M estimators when its variance is not too large (because it is unbiased), and its variance increases when sample size is smaller or $p_\mu$ is larger. This variance is likely due to the second step of RRAT.

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

ther additive ($g_\sigma(w) = 1.0$), multiplicative ($g_\sigma(w) = 0.1w$), or a combination of both ($g_\sigma(w) = 1.0 + 0.1w$). We tried positive skews, with $p_\mu \in \{0.55, 0.60, \ldots, 0.95\}$. We fixed $n = 1$ and $p_1 = 0.50$. In the case of affine functions, we used well-specified affine models for $f_q$, and in the case of the nonlinear function, we used a neural network (NN) model for $f_q$. (NNs can be shown to be good approximators not only for ordinary LS regression but also for quantile regression; White, 1992). We used well-specified $f_c$—for additive noise, $f_c(w) = c + w$; for multiplicative noise, $f_c(w) = dw$; and in combination, $f_c(w) = c + dw$. LS estimates of affine functions and the second step of RRAT are analytically computed. LS estimates of NN functions and the first step of RRAT are implemented in an iterative manner with the stochastic gradient descent (for NNs)[8] and with Polack-Ribiere conjugate gradients (for affine classes). In this implementation, we did not use specific capacity control (no weight decay and the number of hidden units in NNs were fixed to 10 both for the first step in RRAT and for LS.). The number of independent experiments is 500 for the affine $g_\mu$ and 50 for the nonlinear $g_\mu$. The results for additive noise, multiplicative noise, and the combination of them are summarized in Figures 4, 5, and 6, respectively.

In Figures 4, 5, and 6, the values above the zero line suggest RRAT faring better than LS and vice versa. In each figure, experimental curves for 2-, 5-, and 10-dimensional affine models and for the nonlinear models are given. For more asymmetric noise (larger $p_\mu$), RRAT fares better. Note also that as the number of parameters estimated through the first step of RRAT (quantile regression $f_{p_1}$) increases, the relative improvement brought by RRAT over LS regression increases. The number of parameters of affines are 3, 6, and 11, respectively, and that of the NN is 81. The relative improvements decrease or vanish when the predictor is nonlinear and $p_\mu$ is fairly large. The possible explanations on this apparent inconsistency with theorem 3 are that the second-order approximation or full-rank assumption were not satisfied, the NN approximation violates the assumption of theorem 3 that the model is well specified, or the number of samples is not large enough to be consistent with asymptotic behavior—or all of these. For additive noise (see Figure 4), the theoretical curves derived from theorem 3 are also indicated, which almost overlap with experimental results.

Concerning statistical significance in Figures 4, 5, and 6, RRAT was significantly (0.01 level) better than LS regression when $p_\mu \geq 0.65$, as analytically expected, except when $g_\mu$ is nonlinear and $p_\mu$ is 0.90 or 0.95.

**4.3 Experiment 3: Which RRAT Works.** Another series of experiments was designed to study the performance of RRAT when varying choices of the

[8] We employed stochastic gradient, which worked better than conjugate gradient for nonlinear classes. As usual, the initial learning rate and the learning rate decrease factor were roughly selected by a few trials on the training set.

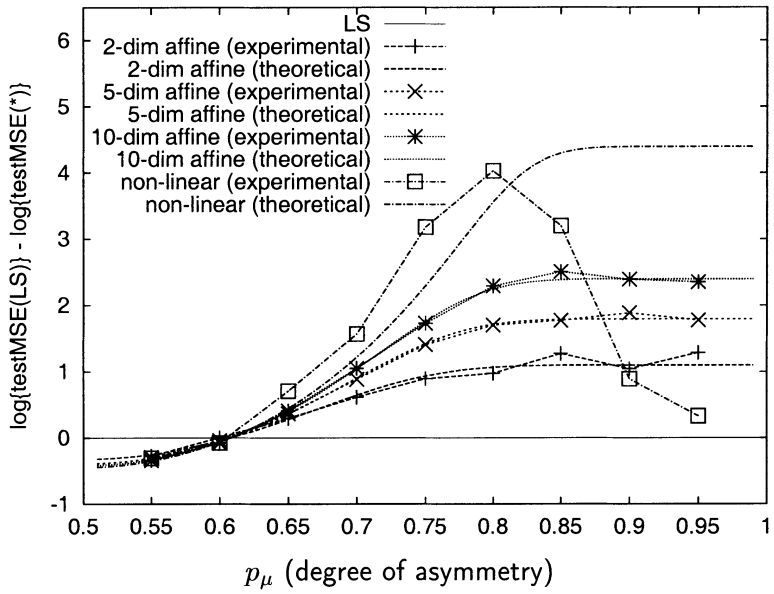


Figure 4: Performances of RRAT compared with LS estimator under additive noise. Note that the more asymmetric the noise distribution is, the better RRAT (relatively) worked in most cases. Note also that the more parameters in $f_{p_1}$, the better RRAT (relatively) worked. The theoretical curves derived from theorem 3 are also indicated, which almost overlap with empirical results.

hyperparameters $n, p_1, \ldots, p_n$. We tried $n \in \{1, 2, 3\}$ and $p_i \in \{0.2, 0.5, 0.8\}$. In the series of experiments in this section, we fixed $g_\mu$ as a two-dimensional affine function, $g_\sigma$ additive ($g_\sigma(w) = 1.0$), and $p_\mu = 0.75$. $\mathcal{F}_p$ is the class of two-dimensional affine models, and $\mathcal{F}_c$ is the class of additive constant models (i.e., they are well specified). For parameter estimation, we introduced capacity control with a weight decay parameter $w_d$ (penalty on the 2-norm of the parameters) chosen from $\{10^{-5}, 10^{-4}, \ldots, 10^{+5}\}$. The best weight decay was chosen with another 1000 validation samples that are independent of training and test samples. From 100 independent experiments, we obtained the results summarized in Table 2.

Table 2 shows the mean and its standard error over the test samples MSE for each method. The $p$-values from the Wilcoxon signed rank test (null hypothesis of no difference) are also indicated. From the given $p$-values, it is clear that all variants of RRAT worked significantly better than LS regression. Note that the choice of $p_i$ does not change the performance considerably. The true $p_\mu = 0.75$, but RRAT(1) with $p_1 = 0.20$ or 0.50 worked as well as RRAT(1) with $p_1 = 0.80$. On the other hand,

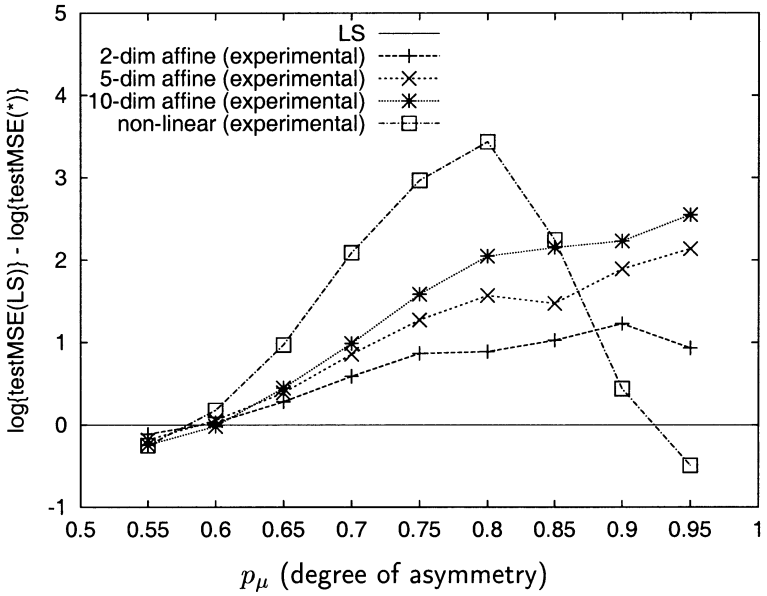


Figure 5: Performances of RRAT compared with LS estimator under multiplicative noise. Note that the more asymmetric the noise distribution is, the better RRAT (relatively) worked in most cases. Note also that the more parameters in $f_{p_1}$, the better RRAT (relatively) worked.

the choice of $n$ changes the performance considerably. (The $p$-values of the significant difference between RRAT(1) and RRAT(2) were in the range $(4.31 \times 10^{-9}, 8.65 \times 10^{-8})$, those between RRAT(1) and RRAT(3) were in the range $(7.73 \times 10^{-8}, 1.62 \times 10^{-7})$, and those between RRAT(2) and RRAT(3) were in the range $(3.50 \times 10^{-1}, 4.67 \times 10^{-1})$.) As we assumed additive noise structure here, RRAT(1) is sufficient and RRAT($n$), $n \geq 2$, are redundant, as explained with property 1. When the noise structure is more complicated (as with our insurance data), RRAT(1) might not be sufficient and RRAT($n$) with larger $n$ might be more suitable.

## 5 Application to Insurance Premium Estimation

We applied the proposed method to a problem in automobile insurance pure premium estimation: estimate the risk of a driver given his or her profile (e.g., age, type of car). In this application, we want the conditional expectation (and not another conditional moment) 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 cost). One of

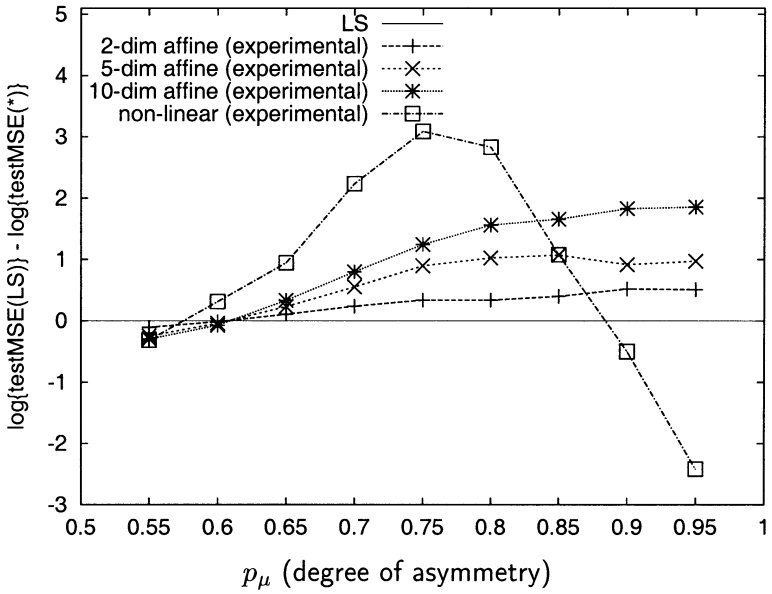


Figure 6: Performances of RRAT compared with LS estimator under the combination of additive and multiplicative noise. Note that the more asymmetric the noise distribution is, the better RRAT (relatively) worked in most cases. Note also that the more parameters in $f_{p_1}$, the better RRAT (relatively) worked except the case of nonlinear with extremely asymmetric noise.

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, the data (claim amounts) have noise (unexplainable by input, i.e., the customer's profiles) distributed with an asymmetric heavy tail extending out toward positive values. We used real-world data from an insurance company. The distribution is unknown, so the models are probably not correctly specified.

The number of input variables of the data set is 39, all discrete except one. A discrete variable with value $i$ is one-hot encoded (in a vector with all entries zero except for a 1 at position $i$), yielding input vectors with 266 elements. We repeated the experiment 10 times, each time using 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 into three equal subsets with 5000 samples, respectively, for training, validation (model selection), and testing (model comparison).

In the experiment, we compared RRAT and LS regression using linear and NN quantile predictors, that is, $\mathcal{F}_q$ is affine or a NN, fitted using conju-

Table 2: The Mean and Its Standard Error for the Average MSE of Each Method.

| | | | |
|---|---|---|---|
| LS regression | | | |
| Mean | $8.96 \times 10^{-2}$ | | |
| Standard error | $1.04 \times 10^{-2}$ | | |
| RRAT(1) | $p_1 = .2$ | $p_1 = .5$ | $p_1 = .8$ |
| Mean | $3.45 \times 10^{-2}$ | $3.41 \times 10^{-2}$ | $3.45 \times 10^{-2}$ |
| Standard error | $3.54 \times 10^{-3}$ | $3.51 \times 10^{-3}$ | $3.52 \times 10^{-3}$ |
| $p$-value | $9.76 \times 10^{-15}$ | $8.31 \times 10^{-15}$ | $9.25 \times 10^{-15}$ |
| RRAT(2) | $p_1 = .2, p_2 = .5$ | $p_1 = .2, p_2 = .8$ | $p_1 = .5, p_2 = .8$ |
| Mean | $4.56 \times 10^{-2}$ | $4.61 \times 10^{-2}$ | $4.52 \times 10^{-2}$ |
| Standard error | $4.22 \times 10^{-3}$ | $4.11 \times 10^{-3}$ | $4.20 \times 10^{-3}$ |
| $p$-value | $1.73 \times 10^{-11}$ | $4.87 \times 10^{-11}$ | $2.34 \times 10^{-11}$ |
| RRAT(3) | $p_1 = .2, p_2 = .5, p_3 = .8$ | | |
| Mean | $4.66 \times 10^{-2}$ | | |
| Standard error | $4.17 \times 10^{-3}$ | | |
| $p$-value | $3.97 \times 10^{-11}$ | | |

Notes: In the tables for RRAT($n$), $n = 1, 2, 3$, the $p$-values for the Wilcoxon signed rank test are also indicated. All variants of RRAT work significantly better than LS regression. Note that the choice of $p_i$ does not change the performance considerably, but the choice of $n$ does.

gate gradients. Capacity is controlled by weight decay $\in \{10^{-5}, 10^{-4}, \ldots 10^{5}\}$ (and in the case of the NN: early stopping, and number of hidden units $\in \{5, 10, \ldots, 25\}$), selected using the validation set. The correction-combination function $f_c$ is always affine (also with weight decay chosen using the validation set), and its parameters are estimated analytically. We tried RRAT($n$) with $n = 1, 2, 3$ and $p_i = 0.20, 0.50, 0.80$. For RRAT(1), we tried additive, affine, or quadratic correction-combination functions: $f_c \in \{c_0 + f_{p_1}, c_0 + c_1 f_{p_1}, c_0 + c_1 f_{p_1} + c_2 f_{p_1}^2\}$. For RRAT($n$), $n \geq 2$, we tried affine $f_c \in \{c_0 + c_1 f_{p_1} + c_2 f_{p_2} + \cdots\}$.

Table 3 (linear model cases) and Table 4 (NN model cases) show the $p$-values from the Wilcoxon signed rank test. In RRAT(1), the choice of $p_1 = 0.20$ does not work well, which suggests either that the underlying distribution of the data set is out of the class of equation 3.2 or the noise structure is more complicated than those tried. When $p_1 = 0.50$, the choice of $f_c$ significantly changed the performance. The worse performance of additive $f_c$ and better performance of affine $f_c$ suggest that the noise structure of the data set is more multiplicative than additive. When $p_1 = 0.80$, RRAT worked better independently of the choice of $f_c$, which suggests that the true $p_\mu$ of the data set (if it does not vary too much with $x$) stays around 0.80. Note that RRAT($n$), $n \geq 2$ always worked better than

Table 3: Experimental Comparison of RRAT and LS Regression on Linear Predictors.

| RRAT(1) | $p_1 = .2$ | $p_1 = .5$ | $p_1 = .8$ |
|---|---|---|---|
| $f_c(f_{p_1}) = c_0 + f_{p_1}$ | | | |
| $p$-value | $8.30 \times 10^{-3}$ ** | $2.97 \times 10^{-2}$ * | $2.34 \times 10^{-2}$ ○ |
| RRAT(1) | $p_1 = .2$ | $p_1 = .5$ | $p_1 = .8$ |
| $f_c(f_{p_1}) = c_0 + c_1 f_{p_1}$ | | | |
| $p$-value | $1.01 \times 10^{-1}$ — | $3.72 \times 10^{-2}$ ○ | $3.46 \times 10^{-3}$ ○○ |
| RRAT(1) | $p_1 = .2$ | $p_1 = .5$ | $p_1 = .8$ |
| $f_c(f_{p_1}) = c_0 + c_1 f_{p_1} + c_2 f_{p_1}^2$ | | | |
| $p$-value | $1.93 \times 10^{-1}$ — | $4.67 \times 10^{-3}$ ○○ | $2.53 \times 10^{-3}$ ○○ |
| RRAT(2) | $p_1 = .2, p_2 = .5$ | $p_1 = .2, p_2 = .8$ | $p_1 = .5, p_2 = .8$ |
| $f_c(f_{p_1}, f_{p_2}) = c_0 + c_1 f_{p_1} + c_2 f_{p_2}$ | | | |
| $p$-value | $1.42 \times 10^{-2}$ ○ | $3.46 \times 10^{-3}$ ○○ | $3.46 \times 10^{-3}$ ○○ |
| RRAT(3) | $p_1 = .2, p_2 = .5, p_3 = .8$ | | |
| $f_c(f_{p_1}, f_{p_2}, f_{p_3}) = c_0 + c_1 f_{p_1} + c_2 f_{p_2} + c_3 f_{p_3}$ | | | |
| $p$-value | $2.53 \times 10^{-3}$ ○○ | | |
| Best Model on Validation | | | |
| $p$-value | $2.53 \times 10^{-3}$ ○○ | | |

Notes: The $p$-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 ○ (○○) denotes RRAT being significantly better than LS regression at 0.05 (0.01) level. Note that RRAT($n$), $n \geq 2$, always beats LS regression.

LS, even though one of the components, $f_{0.20}$, was very poor by itself. It follows that the choice of $p_i$ is not critical to the success of the application; we can choose several $p_i$ and combine them. Furthermore, when we pick the best of the RRAT models (choice of $n$, $p_i$, and $f_c$ from above) based on the validation set average squared error, even better results are obtained.

## 6 Conclusion

We have introduced a new robust regression method, RRAT, that is well suited to asymmetric heavy-tail unknown noise distributions (where previous robust regression methods are not well suited). It can be applied to both linear and nonlinear regressions. A large class of generating models

Table 4: Experimental Comparison of RRAT and LS Regression on NN Predictors.

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

Notes: The $p$-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 $\circ$ ($\circ\circ$) denotes RRAT being significantly better than LS regression at 0.05 (0.01) level. Note that RRAT($n$), $n \geq 2$, always beats LS regression.

for which a universal approximation property holds has been characterized (and it includes additive and multiplicative noise with arbitrary conditional expectation). An analysis of known asymmetric heavy-tail noise distributions reveals when the proposed method beats least-square regression and M estimators. The proposed method has been tested and analyzed on both synthetic data and insurance data (the motivation for this research), showing it to outperform both least-squares regression and M estimators significantly over a wide range of asymmetric distributions.

Future work should investigate estimators of confidence intervals around the RRAT estimator (e.g., using the bootstrap or other resampling methods). Another important direction of future work is understanding and improving the behavior of nonlinear predictors for a very small signal-to-noise ratio (large $p_\mu$).

## Appendix A: Proof of Theorem 1

For all $x$, $f_{p_1}(x)$ is represented as a function of $f_{p_\mu}(x)$ as follows:

$$\begin{aligned}\forall x,\ f_{p_1}(x) &= f_{p_\mu}(x) - \left(F^{-1}_{Z\cdot g_\sigma\{g_\mu(x)\}}(p_\mu) - F^{-1}_{Z\cdot g_\sigma\{g_\mu(x)\}}(p_1)\right)\\ &= f_{p_\mu}(x) + F^{-1}_{Z\cdot g_\sigma\{g_\mu(x)\}}(p_1)\\ &= f_{p_\mu}(x) + F_Z^{-1}(p_1)\cdot g_\sigma\{g_\mu(x)\}\\ &= g_\mu(x) + F_Z^{-1}(p_1)\cdot g_\sigma\{g_\mu(x)\}, \end{aligned} \tag{A.1}$$

where $F_W^{-1}(p)$ is the $p$th quantile of random variable $W$. In equation A.1, from the first line to the second, we used $E[Z g_\sigma\{g_\mu(X)\} \mid X] = 0$. From the second to the third, we used the following property of cdf's: if $F_{aW}(av) = F_W(v)$, then $F^{-1}_{aW}(p) = a \cdot F^{-1}_W(p)$, where $a > 0$, $0 < p < 1$ are constants and $W$ is a random variable drawn from a continuous distribution. From the third to fourth, we used $f_{p_\mu}(x) \equiv g_\mu(x)$ for all $x$.

If the monotonicity of $h(\bar{y})$ in the theorem holds, there is a one-to-one mapping between $f_{p_1}(x)$ and $g_\mu(x)(= E[Y \mid x])$. It follows that there exists a function $f_c$ such that $E[Y \mid X] = f_c(f_{p_1}(X))$.

## Appendix B: Proof of Theorem 2

As in equation A.1, in the proof of theorem 1, $f_{p_1}(x)$ and $f_{p_2}(x)$ are given, for all $x$, by

$$f_{p_1}(x) = g_\mu(x) + F_Z^{-1}(p_1)\cdot g_\sigma(g_\mu(x)), \tag{B.1}$$

$$f_{p_2}(x) = g_\mu(x) + F_Z^{-1}(p_2)\cdot g_\sigma(g_\mu(x)). \tag{B.2}$$

Therefore, the proof of theorem 1 already shows the existence of the mapping from $g_\mu(x)$ to $\{f_{p_1}(x), f_{p_2}(x)\}$, with equations B.1 and B.2. Here we want to show that the mapping is one-to-one for all $x$.

Assume that it is not one-to-one; then there is at least one case where two different values $\bar{y}$ and $\bar{y}'$ are mapped to identical $\{f_{p_1}(x), f_{p_2}(x)\}$, that is,

$$\bar{y} - F_Z^{-1}(p_1)\cdot g_\sigma(\bar{y}) = \bar{y}' - F_Z^{-1}(p_1)\cdot g_\sigma(\bar{y}'), \tag{B.3}$$

$$\bar{y} - F_Z^{-1}(p_2)\cdot g_\sigma(\bar{y}) = \bar{y}' - F_Z^{-1}(p_2)\cdot g_\sigma(\bar{y}'). \tag{B.4}$$

Dividing both sides of equations B.3 and B.4 by $F_Z^{-1}(p_1)$ and $F_Z^{-1}(p_2)$, respectively (from the continuity of $Z$ and equation 3.4, $F_Z^{-1}(p_1)$ and $F_Z^{-1}(p_2)$ are

not zero), and subtracting equation B.3 from B.4 yields

$$\left(\frac{1}{F_Z^{-1}(p_1)} - \frac{1}{F_Z^{-1}(p_2)}\right)(\bar{y} - \bar{y}') = 0. \tag{B.5}$$

From the continuity of $Z$ and equation 3.4, $\frac{1}{F_Z^{-1}(p_1)} \neq \frac{1}{F_Z^{-1}(p_2)}$ and from the assumption $\bar{y} \neq \bar{y}'$, we find that the assumption that the mapping between $g_\mu(x)$ and $\{f_{p_1}(x), f_{p_2}(x)\}$ is not one-to-one is false.

It follows that there exists a function $f_c$ such that $E[Y|X] = f_c(f_{p_1}(X), f_{p_2}(X))$.

## Appendix C: Proof of Property 1

By applying $g_\sigma$ in equation 3.5 into equation A.1, we get

$$f_{p_1}(x) = g_\mu(x) - F_Z^{-1}(p_1) \cdot (c + d \cdot g_\mu(x)). \tag{C.1}$$

We can exactly obtain a linear function $f_c$,

$$\begin{aligned} g_\mu(x) &= f_c(f_{p_1}(x)) \\ &= \frac{-F_Z^{-1}(p_1) \cdot c}{1 - F_Z^{-1}(p_1) \cdot d} + \frac{1}{1 - F_Z^{-1}(p_1) \cdot d} \cdot f_{p_1}(x), \end{aligned} \tag{C.2}$$

except when the denominator appearing on the right-hand side of the second line in equation C.2 is zero.[9]

## Appendix D: Proof of Theorem 3

We consider a class of regression problems in the following form,

$$Y = f(X; \alpha^*) + \beta^* + Z, \tag{D.1}$$

where $X$ and $Y$ are the input and output random variables, respectively. $f$ is an arbitrary function characterized by a set of parameters $\alpha^* \in \mathbb{R}^d$, and $\beta^* \in \mathbb{R}$ is a scalar parameter. $Z$ is the (zero-mean) noise random variable drawn from a (possibly) asymmetric heavy-tail distribution (independent of $X$). We assume that the model is well specified, that is, the true function $f(x; \alpha^*) + \beta^*$ is a member of the class of parametric models: $M = \{f(x; \alpha) +$

[9] This problem can be solved by choosing another order $p_2 \neq p_1$. In the application of this method, we suggest trying several orders, $p_1, p_2, \ldots, p_n$, and choosing the best one or using RRAT($n$), $n \geq 2$. This is demonstrated in sections 4 and 5 for application to insurance premium estimation.

$\beta$: $\alpha \in \Theta \subset \mathbb{R}^d, \beta \in \mathbb{R}\}$, and that the training data is independently and identically distributed.

The risk of the model by LS regression is given by standard calculation of asymptotic statistics:

$$\begin{aligned}\text{risk}_{\text{LS}} &= E_{Data}\left\{E_X\{(f(x;\alpha^*)+\beta^*-f(x;\hat{\alpha}_{\text{LS}})-\hat{\beta}_{\text{LS}})^2\}\right\} \\ &= \frac{1}{n}\{\text{Var}(Z)+d\,\text{Var}(Z)\}+o\left(\frac{1}{n}\right), \end{aligned} \tag{D.2}$$

where $n$ is the number of training data and $\hat{\alpha}_{\text{LS}}$, $\hat{\beta}_{\text{LS}}$ denotes the corresponding estimated parameters by LS regression.

As explained, RRAT provides the following estimates:

$$\begin{aligned}\{\hat{\alpha}_{\text{RRAT}}, \hat{\gamma}_{\text{RRAT}}\} = \underset{\alpha,\gamma}{\text{argmin}} \Bigg\{ &\sum_{i:y_i \geq f(x_i;\alpha)+\gamma} p_1 |y_i - f(x_i;\alpha) - \gamma| \\ &+ \sum_{i:y_i < f(x_i;\alpha)+\gamma} (1-p_1) \\ &\times |y_i - f(x_i;\alpha) - \gamma| \Bigg\}, \end{aligned} \tag{D.3}$$

$$\hat{\beta}_{\text{RRAT}} = \frac{1}{n}\sum_{i=1}^{n}\{y_i - f(x_i;\hat{\alpha}_{\text{RRAT}})\}. \tag{D.4}$$

Using these estimated parameters, the risk of the model by RRAT is given by:

$$\begin{aligned}\text{risk}_{\text{RRAT}} &= E_{Data}\left\{E_X\{(f(x;\alpha^*)+\beta^*-f(x;\hat{\alpha}_{\text{RRAT}})-\hat{\beta}_{\text{RRAT}})^2\}\right\} \\ &= \frac{1}{n}\left\{\text{Var}(Z)+d\cdot\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}\right\}+o\left(\frac{1}{n}\right), \end{aligned} \tag{D.5}$$

where $P_Z$ and $F_Z$ are the pdf and cdf of the noise distribution.

From equation D.2 and D.4, RRAT yields more efficient estimates than LS regression when

$$\text{Var}(Z) > \frac{p_1(1-p_1)}{P_Z^2(F_Z^{-1}(p_1))}. \tag{D.6}$$

To prove equation D.5, let us introduce some notation. We define $|w|^+ = \max(0, w)$. We also omit the subscript $_{\text{RRAT}}$ for estimated parameters by RRAT.

First, we define the matrix $K$ as

$$K = \int \begin{pmatrix} \frac{d}{d\alpha} f(x;\alpha^*) \frac{d}{d\alpha} f(x;\alpha^*)' & \frac{d}{d\alpha} f(x;\alpha^*) \\ \frac{d}{d\alpha} f(x;\alpha^*)' & 1 \end{pmatrix} p(x)\, dx = \begin{pmatrix} M_2 & m_1 \\ m_1' & 1 \end{pmatrix},$$

where $M_2$ is a $d \times d$ matrix and $m_1$ is a $d \times 1$ vector. The risk of the RRAT is written as

$$\text{risk} = \operatorname{Tr} \operatorname{Var}(\hat{\alpha}, \hat{\beta}) K + o\left(\frac{1}{n}\right).$$

To obtain the variance of $\hat{\alpha}$, we can use the following relation between the variance and the influence function (Hampel et al., 1986):

$$\lim_{n\to\infty} n \operatorname{Var}(\hat{\alpha}, \hat{\gamma}) = \int IF(x,y) IF(x,y)' p(y \mid x) p(x)\, dy\, dx,$$

where $IF(x, y)$ is the influence function of the estimator $(\hat{\alpha}, \hat{\gamma})$. The influence function is defined as

$$IF(\tilde{x}, \tilde{y}) = \lim_{\kappa \to 0} \frac{(\alpha_\kappa, \gamma_\kappa) - (\alpha^*, \gamma^*)}{\kappa}, \tag{D.7}$$

where $(\alpha_\kappa, \gamma_\kappa)$ is given by minimizing the following with respect to $(\alpha, \gamma)$:

$$\begin{aligned} (1-\kappa) \int \Big\{ & p_1 \left| y - f(x;\alpha) - \gamma \right|^+ \\ & + (1-p_1) \left| f(x;\alpha) + \gamma - y \right|^+ \Big\} p(y \mid x) p(x)\, dy\, dx \\ + \kappa \left\{ p_1 |\tilde{y} - f(\tilde{x};\alpha) - \gamma|^+ \right. & \left. + (1-p_1) |f(\tilde{x};\alpha) + \gamma - \tilde{y}|^+ \right\}. \end{aligned}$$

To obtain the influence function we use $\frac{d}{dx}|x|^+ = \sigma(x)$[10] and $\frac{d}{dx}\sigma(x) = \delta(x)$.[11] By using these relations, we obtain $(\alpha_\kappa, \gamma_\kappa)$ as follows:

$$\begin{aligned} (\alpha_\kappa, \gamma_\kappa)' = (\alpha^*, \gamma^*)' - \kappa \frac{1}{P_Z(F_Z^{-1}(p_1))} & \left\{ 1 - p_1 - \sigma(\tilde{y} - f(\tilde{x};\alpha^*) - \gamma^*) \right\} \\ & \cdot K^{-1} \begin{pmatrix} \frac{d}{d\alpha} f(\tilde{x};\alpha^*) \\ 1 \end{pmatrix} + o(\kappa). \end{aligned} \tag{D.8}$$

Then we obtain the influence function as

$$IF(\tilde{x}, \tilde{y}) = \frac{1}{P_Z(F_Z^{-1}(p_1))} \left\{ 1 - p_1 - \sigma(\tilde{y} - f(\tilde{x};\alpha^*) - \gamma^*) \right\} K^{-1} \begin{pmatrix} \frac{d}{d\alpha} f(\tilde{x};\alpha^*) \\ 1 \end{pmatrix}.$$

[10] $\sigma(x)$ is 1 when $x \geq 0$ and 0 when $x < 0$.

[11] $\delta(x)$ is Dirac's delta function.

Then we find that the covariance matrix of the estimator $(\hat{\alpha}, \hat{\gamma})$ is written as

$$\mathrm{Var}(\hat{\alpha}, \hat{\gamma}) = \frac{1}{n} \frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2} K^{-1} + o\left(\frac{1}{n}\right).$$

Here we decompose the matrix $K^{-1}$ as

$$K^{-1} = \begin{pmatrix} H & t \\ t' & u \end{pmatrix}.$$

Then the variance matrix of $\hat{\alpha}$ is written as

$$\mathrm{Var}(\hat{\alpha}) = \frac{1}{n} \frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2} H + o\left(\frac{1}{n}\right). \tag{D.9}$$

Next we calculate the variance of $\hat{\beta}$ (written as $\mathrm{Var}(\hat{\beta})$) in equation D.4, which is decomposed as follows:

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i \mid \hat{\alpha}))\right) \\
&= \frac{1}{n^2}\sum_{i,j} E_{Data}\{z_i z_j\} && \text{(D.10)} \\
&\quad + \frac{1}{n^2}\sum_{i,j} E_{Data}\left\{\frac{d}{d\alpha} f(x_i;\alpha^*)'(\hat{\alpha}-\alpha^*)\right. \\
&\qquad\qquad \left. \times (\hat{\alpha}-\alpha^*)'\frac{d}{d\alpha} f(x_j;\alpha^*)\right\} && \text{(D.11)} \\
&\quad - \frac{2}{n^2}\sum_{i,j} E_{Data}\left\{z_i(\hat{\alpha}-\alpha^*)'\frac{d}{d\alpha} f(x_j;\alpha^*)\right\} && \text{(D.12)} \\
&\quad - \frac{2}{n^2}\sum_{i,j} E_{Data}\left\{z_i(\hat{\alpha}-\alpha^*)'\frac{\partial^2 f(x_j;\alpha^*)}{\partial\alpha^2}(\hat{\alpha}-\alpha^*)\right\} && \text{(D.13)} \\
&\quad + o\left(\frac{1}{n}\right).
\end{aligned}$$

The first term, equation D.10, is equal to $\frac{\mathrm{Var}(z)}{n}$.
The second term, equation D.11, is calculated as follows. First,

$$E_{z|X}\left\{(\hat{\alpha}-\alpha^*)(\hat{\alpha}-\alpha^*)'\right\} = \frac{1}{n}\frac{p_1(1-p_1)}{p_z(m_1)^2} H + o_p\left(\frac{1}{n}\right), \tag{D.14}$$

where $o_p(\cdot)$ is the probabilistic order with respect to $p(x_1, \ldots, x_n)$. Substituting equations D.14 to D.11, we obtain

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i,j} E_{Data}\left\{\frac{d}{d\alpha}f(x_i;\alpha^*)'(\hat{\alpha}-\alpha^*)(\hat{\alpha}-\alpha^*)'\frac{d}{d\alpha}f(x_j;\alpha^*)\right\}\\
&\quad=\frac{1}{n^2}\sum_{i,j}\mathrm{Tr}\left\{\frac{1}{n}\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}HE_X\right.\\
&\qquad\qquad\left.\times\left\{\frac{d}{d\alpha}f(x_i;\alpha^*)\frac{d}{d\alpha}f(x_j;\alpha^*)'\right\}+o\left(\frac{1}{n}\right)\right\}\\
&\quad=\frac{1}{n}\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}m_1'Hm_1+o\left(\frac{1}{n}\right). \qquad \text{(D.15)}
\end{aligned}$$

The third term, equation D.12, is calculated as follows. First, we calculate the conditional expectation $E_{z|X}\{z_i(\hat{\alpha}-\alpha^*)\}$. Now we define $(\hat{\alpha}_{(i)}, \hat{\gamma}_{(i)})$ as the estimator obtained from all the training data except $(x_i, y_i)$. By definition, $(\hat{\alpha}_{(i)}, \hat{\gamma}_{(i)})$ is independent from $z_i$. By simple calculation we find

$$\begin{aligned}
\begin{pmatrix}\hat{\alpha}\\ \hat{\gamma}\end{pmatrix}&=\begin{pmatrix}\hat{\alpha}_{(i)}\\ \hat{\gamma}_{(i)}\end{pmatrix}-\frac{1}{n}\frac{1}{P_Z(F_Z^{-1}(p_1))}\left\{1-p_1-\sigma(y_i-f(x_i\mid\hat{\alpha})-\hat{\gamma})\right\}\\
&\quad\times K^{-1}\begin{pmatrix}\frac{d}{d\alpha}f(x_i;\alpha^*)\\ 1\end{pmatrix}+o_p\left(\frac{1}{n}\right). \qquad \text{(D.16)}
\end{aligned}$$

Then the difference between $\hat{\alpha}$ and $\hat{\alpha}_{(i)}$ is calculated as

$$\begin{aligned}
\hat{\alpha}-\alpha^*&=\hat{\alpha}_{(i)}-\alpha^*-\frac{1}{n}\frac{1}{P_Z(F_Z^{-1}(p_1))}\left\{1-p_1-\sigma(y_i-f(x_i\mid\hat{\alpha})-\hat{\gamma})\right\}\\
&\quad\times\left(H\frac{d}{d\alpha}f(x_i;\alpha^*)+t\right)+o_p\left(\frac{1}{n}\right). \qquad \text{(D.17)}
\end{aligned}$$

Substituting equation D.17 in $E_{z|X}\{z_i(\hat{\alpha}-a^*)\}$, we obtain

$$\begin{aligned}
E_{z|X}\{z_i(\hat{\alpha}-a^*)\}&=\frac{1}{n}\frac{1}{P_Z(F_Z^{-1}(p_1))}\left(\int_0^\infty \tilde{z}p_z(\tilde{z})\,d\tilde{z}\right)\left(H\frac{d}{d\alpha}f(x_i;\alpha^*)+t\right)\\
&\quad+o_p\left(\frac{1}{n}\right). \qquad \text{(D.18)}
\end{aligned}$$

Then we obtain

$$
\begin{aligned}
&-\frac{2}{n^2}\sum_{i,j} E_{Data}\left\{z_i(\hat{\alpha}-\alpha^*)'\frac{d}{d\alpha}f(x;\alpha^*)\right\} \\
&\quad = -\frac{2}{n^2}\frac{1}{n}\frac{1}{P_Z(F_Z^{-1}(p_1))}\left(\int_0^\infty \tilde{z}p_z(\tilde{z})d\tilde{z}\right) \\
&\qquad \times \sum_{i,j} E_X\left\{\mathrm{Tr}\left(H\frac{d}{d\alpha}f(x_i;\alpha^*)+t\right)\frac{d}{d\alpha}f(x_j;\alpha^*)'\right\}+o\left(\frac{1}{n}\right) \\
&\quad = -\frac{2}{n}\frac{1}{P_Z(F_Z^{-1}(p_1))}\left(\int_0^\infty \tilde{z}p_z(\tilde{z})d\tilde{z}\right)\left(m_1'Hm_1+m_1't\right)+o\left(\frac{1}{n}\right) \\
&\quad = o\left(\frac{1}{n}\right). \qquad \text{(D.19)}
\end{aligned}
$$

The last equation is obtained from the definition of $H$ and $t$, that is, we can find that $Hm_1 + t = 0$.

The fourth term, equation D.13, is also $o\left(\frac{1}{n}\right)$. This can be verified by substituting equation D.17 into D.13.

Then from the previous discussion, we obtain the variance of $\hat{\beta}$ as

$$
\mathrm{Var}(\hat{\beta}) = \frac{\mathrm{Var}(Z)}{n} + \frac{1}{n}\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}m_1'Hm_1 + o\left(\frac{1}{n}\right). \qquad \text{(D.20)}
$$

Next, we calculate the covariance between $\hat{\alpha}$ and $\hat{\beta}$. $\hat{\beta}-\beta^*$ is written as

$$
\hat{\beta}-\beta^* = \frac{1}{n}\sum_{i=1}^n z_i - \frac{1}{n}\sum_{i=1}^n(\hat{\alpha}-\alpha)'\frac{d}{d\alpha}f(x_i;\alpha^*) + o_p\left(\frac{1}{\sqrt{n}}\right). \qquad \text{(D.21)}
$$

Substituting the above equation to $E_{Data}\{(\hat{\beta}-\beta^*)(\hat{\alpha}-\alpha^*)\}$, we obtain:

$$
\begin{aligned}
&E_{Data}\{(\hat{\beta}-\beta^*)(\hat{\alpha}-\alpha^*)\} \\
&\quad = \frac{1}{n}\sum_{i=1}^n E_X\left\{E_{z|X}\left\{z_i(\hat{\alpha}-\alpha^*)\right\}\right\} \\
&\qquad -\frac{1}{n}\sum_{i=1}^n E_X\left\{E_{z|X}\left\{(\hat{\alpha}-\alpha^*)(\hat{\alpha}-\alpha^*)'\right\}\frac{d}{d\alpha}f(x_i;\alpha^*)\right\}+o\left(\frac{1}{n}\right) \\
&\quad = \frac{1}{n}\sum_{i=1}^n E_X\left\{\frac{1}{n}\frac{1}{P_Z(F_Z^{-1}(p_1))}\left(\int_0^\infty \tilde{z}p_z(\tilde{z})d\tilde{z}\right)\left(H\frac{d}{d\alpha}f(x_i;\alpha^*)+t\right)\right\} \\
&\qquad -\frac{1}{n}\sum_{i=1}^n E_X\left\{\frac{1}{n}\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}H\frac{d}{d\alpha}f(x_i;\alpha^*)\right\}+o\left(\frac{1}{n}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\frac{1}{P_Z(F_Z^{-1}(p_1))}\left(\int_0^\infty \tilde{z}p_z(\tilde{z})d\tilde{z}\right)(Hm_1+t)\\
&\quad -\frac{1}{n}\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}Hm_1 + o\left(\frac{1}{n}\right)\\
&= -\frac{1}{n}\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}Hm_1 + o\left(\frac{1}{n}\right). \qquad \text{(D.22)}
\end{aligned}
$$

Now we have all the elements for calculating the risk of the estimator $(\hat{\alpha}, \hat{\beta})$. The variance of $(\hat{\alpha}, \hat{\beta})$ is given as

$$
\begin{aligned}
\mathrm{Var}(\hat{\alpha}, \hat{\beta}) &= \frac{1}{n}\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}\\
&\quad \times \begin{pmatrix} H & -Hm_1 \\ -m_1'H & \frac{P_Z(F_Z^{-1}(p_1))^2}{p_1(1-p_1)}\mathrm{Var}(Z) + m_1'Hm_1 \end{pmatrix}\\
&\quad + o\left(\frac{1}{n}\right). \qquad \text{(D.23)}
\end{aligned}
$$

On the other hand, the risk is calculated as $\mathrm{Tr}\,\mathrm{Var}(\hat{\alpha}, \hat{\beta})K + o\left(\frac{1}{n}\right)$. Then the risk is as follows:

$$
\begin{aligned}
\mathrm{Tr}\,\mathrm{Var}(\hat{\alpha}, \hat{\beta})K &= \frac{1}{n}\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}\,\mathrm{Tr}\\
&\quad \times \begin{pmatrix} H & -Hm_1 \\ -m_1'H & \frac{P_Z(F_Z^{-1}(p_1))^2}{p_1(1-p_1)}\mathrm{Var}(Z) + m_1'Hm_1 \end{pmatrix}\begin{pmatrix} M_2 & m_1 \\ m_1 & 1 \end{pmatrix}\\
&= \frac{1}{n}\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}\\
&\quad \times \left\{\mathrm{Tr}\,HM_2 - m_1'Hm_1 + \frac{P_Z(F_Z^{-1}(p_1))^2}{p_1(1-p_1)}\mathrm{Var}(Z)\right\}\\
&= \frac{1}{n}\left\{\mathrm{Var}(Z) + d\cdot\frac{p_1(1-p_1)}{P_Z(F_Z^{-1}(p_1))^2}\right\}, \qquad \text{(D.24)}
\end{aligned}
$$

where we use the relation

$$
H = M_2^{-1} + \frac{1}{1-m_1'M_2^{-1}m_1}M_2^{-1}m_1m_1'M_2^{-1}.
$$

Thus we obtain the assertion of equation D.5.

**Acknowledgments**

We thank Léon Bottou for his stimulating suggestions, as well as the following agencies for funding: NSERC, MITACS, and JSPS. Most of this work was done while I. T. and T. K. were at Université de Montréal.

**References**

Antle, C. (1985). Lognormal distribution. In *Encyclopedia of statistical sciences* (Vol. 5, pp. 134–136). New York: Wiley.

Beaton, A., & Tukey, J. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, *16*, 147–185.

Bickel, P. (1973). On some analogues to linear combinations of order statistics in the linear model. *Annuals of Statistics*, *1*, 597–616.

Friedman, J., Grosse, E., & Suetzle, W. (1983). Multidimensional additive spline approximation. *SIAM Journal of Scientific and Statistical Computing*, *4*(2), 291–301.

Greene, W. (1997). *Econometric analysis* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Hampel, F., Ronchetti, E., Rousseeuw, P., & Stahel, W. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

Huber, P. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Stat.*, *1*, 799–821.

Huber, P. (1982). *Robust statistics*. New York: John Wiley.

Koenker, R., & Bassett, G., Jr. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.

Krasker, W., & Welsch, R. (1982). Efficient bounded-influence regression estimation. *J. Am. Stat. Asso.*, *77*, 595–604.

Lehmann, E. (1983). *Theory of point estimation*. New York: Wiley.

Rousseeuw, P., & Leroy, A. (1987). *Robust regression and outlier detection*. New York: Wiley.

White, H. (1992). Nonparametric estimation of conditional quantiles using neural networks. In *Proceedings of the 23rd Symposium on the Interface, Computer Science and Statistics* (pp. 190–199). New York: Springer-Verlag.

Received August 6, 2001; accepted April 4, 2002.