This is a draft version of work in progress, content will be revisited in subsequent versions.

Robust Local Polynomial Regression with Similarity Kernels

Yaniv Shulman

yaniv@shulman.info

Abstract

Local Polynomial Regression (LPR) is a widely used nonparametric method for modeling complex relationships due to its flexibility and simplicity. It estimates a regression function by fitting low-degree polynomials to localized subsets of the data, weighted by proximity. However, traditional LPR is sensitive to outliers and high-leverage points, which can significantly affect estimation accuracy. This paper revisits the kernel function used to compute regression weights and proposes a novel framework that incorporates both predictor and response variables in the weighting mechanism. By introducing two positive definite kernels, the proposed method robustly estimates weights, mitigating the influence of outliers through localized density estimation. The method is implemented in Python and is publicly available at `https://github.com/yaniv-shulman/rsklpr`, demonstrating competitive performance in synthetic benchmark experiments. Compared to standard LPR, the proposed approach consistently improves robustness and accuracy, especially in heteroscedastic and noisy environments, without requiring multiple iterations. This advancement provides a promising extension to traditional LPR, opening new possibilities for robust regression applications.

1. Introduction

Local polynomial regression (LPR) is a powerful and flexible statistical technique that has gained increasing popularity in recent years due to its ability to model complex relationships between variables. Local polynomial regression generalizes the polynomial regression and moving average methods by fitting a low-degree polynomial to a nearest neighbors subset of the data at the location. The polynomial is fitted using weighted ordinary least squares, giving more weight to nearby points and less weight to points further away. The value of the regression function for the point is then obtained by evaluating the fitted local polynomial using the predictor variable value for that data point. LPR has good accuracy near the boundary and performs better than all other linear smoothers in a minimax sense [2]. The biggest advantage of this class of methods is not requiring a prior specification of a function i.e. a parametrized model. Instead only a small number of hyperparameters need to be specified such as the type of kernel, a smoothing parameter and the degree of the local polynomial. The method is therefore suitable for modeling complex processes such as non-linear relationships, or complex dependencies for which no theoretical models exist. These two advantages, combined with the simplicity of the method, makes it one of the most attractive of the modern regression methods for applications that fit the general framework of least squares regression but have a complex deterministic structure.

Local polynomial regression incorporates the notion of proximity in two ways. The first is that a smooth function can be reasonably approximated in a local neighborhood by a simple function such as a linear or low order polynomial. The second is the assumption that nearby points carry more importance in the calculation of a simple local approximation or alternatively that closer points are more likely to interact in simpler ways than far away points. This is achieved by a kernel which produces values that diminish as the distance between the explanatory variables increase to model stronger relationship between closer points.

Methods in the LPR family include the Nadaraya-Watson estimator [10, 18] and the estimator proposed by Gasser and Müller [7] which both perform kernel based local constant fit. These were improved on in terms of asymptotic bias by the proposal of the local linear and more general local polynomial estimators [16, 3, 9, 4, 5]. For a review of LPR methods the interested reader is referred to [2].

LPR is however susceptible to outliers, high leverage points and functions with discontinuities in their derivative which often cause an adverse impact on the regression due to its use of least squares based optimization [17]. The use of unbounded loss functions may result in anomalous observations severely affecting the local estimate. Substantial work has been done to develop algorithms to apply LPR to difficult data. To alleviate the issue [15] employs variable bandwidth to exclude observations for which residuals from the resulting estimator are large. In [3] an iterated weighted fitting procedure is proposed that assigns in each consecutive iteration smaller weights to points that are farther then the fitted values at the previous iteration. The process repeats for a number of iterations and the final values are considered the robust parameters and fitted values. An alternative common approach is to replace the squared prediction loss by one that is more robust to the presence of large residuals by increasing more slowly or a loss that has an upper bound such as the Tukey or Huber loss. These methods however require specifying a threshold parameter for the loss to indicate atypical observations or standardizing the errors using robust estimators of scale [8]. For a recent review of robust LPR and other nonparametric methods see [17, 11]

The main contribution of this paper is to revisit the kernel used to produce regression weights. The simple yet effective idea is to generalize the kernel such that both the predictor and the response are used to calculate weights. Within this framework, two positive definite kernels are proposed that assign robust weights to mitigate the adverse effect of outliers in the local neighborhood by estimating the density of the response at the local locations. Note the proposed framework does not preclude the use of robust loss functions, robust bandwidth selectors and standardization techniques. In addition the method is implemented in the Python programming language and is made publicly available. Experimental results on synthetic benchmarks demonstrate that the proposed method achieves competitive results and generally performs better than LOESS/LOWESS using only a single training iteration.

The remainder of the paper is organized as follows: In section 2, a brief overview of the mathematical formulation of local polynomial regression is provided. In section **??**, a framework for robust weights as well as specific robust positive definite kernels are proposed. Section 4 provides an analysis of the estimator and a discussion of its properties. In section 5, implementation notes and experimental results are provided. Finally, in section 6, the paper concludes with directions for future research.

2. Local Polynomial Regression

This section provides a brief overview of local polynomial regression and establishes the notation subsequently used. Let (X, Y) be a random pair and $\mathcal{D}_T = \{(X_i, Y_i)\}_{i=1}^T \subseteq \mathcal{D}$ be a training set comprising a sample of T data pairs. Suppose that $(X, Y) \sim f_{XY}$ a continuous density and $X \sim f_X$ the marginal distribution of X. Let $Y \in \mathbb{R}$ be a continuous response and assume a model of the form $Y_i = m(X_i) + \epsilon_i,\ i \in 1, \ldots, T$ where $m(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is an unknown function and ϵ_i are independently distributed error terms having zero mean representing random variability not included in X_i such that $\mathbb{E}[Y \mid X = x] = m(x)$. There are no global assumptions about the function $m(\cdot)$ other than that it is smooth and that locally it can be well approximated by a low degree polynomial as per Taylor's theorem. Local polynomial regression is a class of nonparametric regression methods that estimate the unknown regression function $m(\cdot)$ by combining the classical least squares method with the versatility of non-linear regression. The local p-th order Taylor expansion for $x \in \mathbb{R}$ near a point X_i yields:

$$m(X_i) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(x - X_i)^j := \sum_{j=0}^{p} \gamma_j(x)(x - X_i)^j \tag{1}$$

To find an estimate $\hat{m}(x)$ of $m(x)$ the low-degree polynomial (1) is fitted to the N nearest neighbors using weighted least squares such to minimize the empirical loss $\mathcal{L}_{lpr}(\cdot\,; \mathcal{D}_N, h)$:

$$\mathcal{L}_{lpr}(x; \mathcal{D}_N, h) := \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \gamma_j(x)(x - X_i)^j \right)^2 K_h(x - X_i) \tag{2}$$

$$\hat{\gamma}(x) := \min_{\gamma(x)} \mathcal{L}_{lpr}(x; \mathcal{D}_N, h) \tag{3}$$

Where $\gamma, \hat{\gamma} \in \mathbb{R}^{p+1}$; $K_h(\cdot)$ is a scaled kernel, $h \in \mathbb{R}_{>0}$ is the bandwidth parameter and $\mathcal{D}_N \subseteq \mathcal{D}_T$ is the subset of N nearest neighbors of x in the training set where the distance is measured on the predictors only. Having computed $\hat{\gamma}(x)$ the estimate of $\hat{m}(x)$ is taken as $\hat{\gamma}(x)_1$. Note the term kernel carries here the meaning typically used in the context of nonparametric regression i.e. a non-negative real-valued weighting function that is typically symmetric, unimodal at zero, integrable with a unit integral and whose value is non-increasing for the increasing distance between the X_i and x. Higher degree polynomials and smaller N generally increase the variance and decrease the bias of the estimator and vice versa [2]. For derivation of the local constant and local linear estimators for the multidimensional case see [6].

3. Robust Weights with Similarity Kernels

The main idea presented is to generalize the kernel function used in equation (2) to produce robust weights. This is achieved by using a similarity kernel function defined on the data domain $K_{\mathcal{D}} : \mathcal{D}^2 \to \mathbb{R}_+$ that enables weighting each point and incorporating information on the data in the local neighborhood in relation to the local regression target.

$$\mathcal{L}_{rsk}(x, y; \mathcal{D}_N, H) := \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x, y)(x - X_i)^j \right)^2 K_{\mathcal{D}}\left((x, y), (X_i, Y_i); H\right) \tag{4}$$

$$\hat{\beta}(x, y; \mathcal{D}_N, H) := \min_{\beta(x,y)} \mathcal{L}_{rsk}(x, y; \mathcal{D}_N, H) \tag{5}$$

Where H is the set of bandwidth parameters. There are many possible choices for such a similarity kernel to be defined within this general framework. However, used as a local weighting function, such a kernel should have the following attributes:

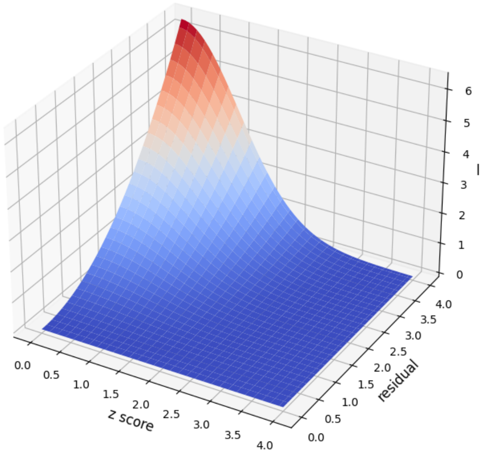
1. Non-negative, $K_{\mathcal{D}}((x, y), (x', y') \geq 0$.
2. Symmetry in the inputs, $K_{\mathcal{D}}((x, y), (x', y')) = K_{\mathcal{D}}((x', y'), (x, y))$.
3. Tending toward decreasing as the distance in the predictors increases. That is, given a similarity function on the response $s(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}_+$, if $s(y, y')$ indicates high similarity the weight should decrease as the distance between the predictors grows, $s(y, y') > \alpha \implies K_{\mathcal{D}}((x, y), (x + u, y')) \geq K_{\mathcal{D}}((x, y), (x + v, y')) \ \forall \ \|u\| \leq \|v\|$ and some $\alpha \in \mathbb{R}_+$.

In this work two such useful positive definite kernels are proposed. Similarly to the usual kernels used in (2), these tend to diminish as the distance between the explanatory variables increases to model stronger relationship between closer points. In addition, the weights produced by the kernels also model the "importance" of the pair (x, y). This is useful for example to down-weight outliers to mitigate their adverse effect on the ordinary least square based regression. Formally let $K_{\mathcal{D}}$ be defined as:

$$K_{\mathcal{D}}\Big((x, y), (x', y'); H_1, H_2\Big) = K_1(x, x'; H_1) K_2\Big((x, y), (x', y'); H_2\Big) \tag{6}$$

Where $K_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ and $K_2 : \mathcal{D}^2 \to \mathbb{R}_+$ are positive definite kernels and H_1, H_2 are the sets of bandwidth parameters. The purpose of K_1 is to account for the distance between a neighbor to the local regression target and therefore may be chosen as any of the kernel functions that are typically used in equation (2). The role of K_2 is described now in more detail as this is the main idea proposed in this work. Using K_2, the method performs robust regression by detecting local outliers in an unsupervised manner and assigns them with lower weights. There are many methods that could be employed to estimate the extent to which a data point is a local outlier however in this work it is estimated in one of the following two ways.

Figure 1: Loss function, assuming a standard quadratic function of the residual, a standard normal density for K_2 and excluding the K_1 distance kernel scaling.



Conditional Density

The first proposed method for K_2 is proportional to the estimated localized conditional marginal distribution of the response variable at the location:

$$K_2\left((x, y), (x', y'); H_2\right) = \hat{f}(y \mid x; H_2)\hat{f}(y' \mid x'; H_2) \tag{7}$$

The nonparametric conditional density estimation is performed using the Parzen–Rosenblatt window (kernel density estimator):

$$\begin{aligned} \hat{f}(y \mid x; H_2) &= \hat{f}(x, y; H_2)/\hat{f}(x; H_2) && (8)\\ &= \hat{f}(v; \mathbf{H}_v)/\hat{f}(x; \mathbf{H}_x) && (9)\\ &= \frac{|\mathbf{H}_x|^{1/2} \sum_{i=1}^{N} K_v\left(\mathbf{H}_v^{-1/2}(v - V_i)\right)}{|\mathbf{H}_v|^{1/2} \sum_{i=1}^{N} K_x\left(\mathbf{H}_x^{-1/2}(x - X_i)\right)} && (10) \end{aligned}$$

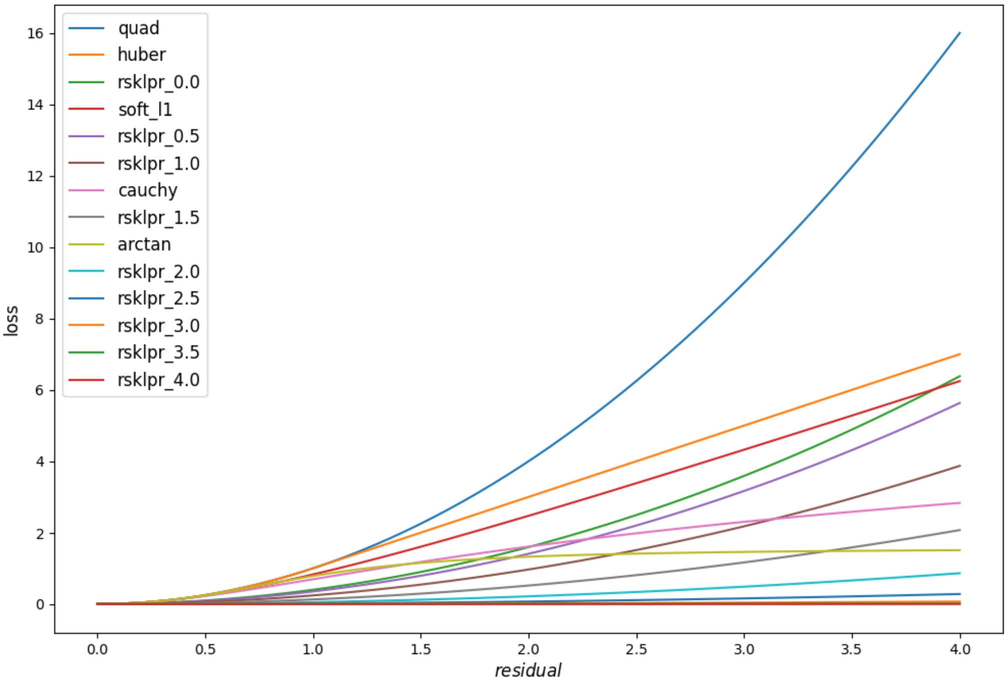
Where $v = [x, y] \in \mathbb{R}^{d+1}$ is the concatenated vector of the predictors and the response; and $\mathbf{H}_v, \mathbf{H}_x$ are bandwidth matrices.

Joint Density

The second proposed kernel is proportional to the joint distribution of the random pair, this could be useful for example to also down-weight high leverage points:

$$K_2\left((x, y), (x', y'); H_2\right) = \hat{f}(x, y; H_2)\hat{f}(x', y'; H_2) \tag{11}$$

Figure 2: The plot illustrates the proposed loss function, a number of common robust losses and the standard quadratic residual loss for comparison. It is assumed that that K_2 is equivalent to the standard normal density and the K_1 distance kernel scaling is excluded. The numbers appended to ”rsklpr” indicate how many standard deviations away from the mean the density is calculated. It is evident that the loss is heavily attenuated in regions of low density.



1

Where the joint density can be estimated using the same aforementioned approach.

Regardless of the choice of kernel, the hyperparameters of this model are similar in essence to the standard local polynomial regression and comprise the span of included points, the kernels and their associated bandwidths. Note that this estimator can be replaced with other robust density estimators and better results are anticipated by doing so however exploring this option is left for future work.

4. Properties

This section discusses some properties of the estimator. Note the notation in this section is simplified by excluding explicit mentions of D_N and H, however the analysis is conditional on the nearest neighbors in the sample, D_N.

4.1. Invariance to y at the Regression Location and Simplification of the Objective

The objective (5) is invariant to the value of y at the location (x, y) for the proposed similarity kernels.

Proof: The optimization is invariant to the scale of the objective function. Therefore:

$$\hat{\beta}(x, y) := \min_{\beta(x,y)} \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x, y)(x - X_i)^j \right)^2 K_{H_1}(x - X_i)\hat{f}(x, y)\hat{f}(X_i, Y_i) \tag{12}$$

$$= \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x, y)(x - X_i)^j \right)^2 K_{H_1}(x - X_i)\hat{f}(X_i, Y_i) \tag{13}$$

The equality holds because $\hat{f}(x, y)$ is a constant scalar that uniformly scales the weights. Since the objective is now independent of y, it follows that:

$$\hat{\beta}(x, y) := \min_{\beta(x)} \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x)(x - X_i)^j \right)^2 K_{H_1}(x - X_i)\hat{f}(X_i, Y_i) \tag{14}$$

$$:= \hat{\beta}(x) \quad \forall y \tag{15}$$

This simplification enables more efficient calculations of the estimator because the dependence on y is removed from the objective function. Note that $\hat{f}(X_i, Y_i)$ can also be replaced with $\hat{f}(Y_i \mid X_i)$ with similar results.

4.2. Weighted Arithmetic Mean of the Standard LPR

The proposed estimator is equivalent to the weighted arithmetic mean of the terms in the standard LPR loss (2), with weights $w_i = \hat{f}(X_i, Y_i)$.

Proof: Since the optimization is invariant to scaling, we have:

$$\hat{\beta}(x) := \min_{\beta(x)} \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x)(x - X_i)^j \right)^2 K_{H_1}(x - X_i)\hat{f}(X_i, Y_i) \tag{16}$$

$$= \min_{\beta(x)} \left(\sum_{i=1}^{N} \hat{f}(X_i, Y_i) \right)^{-1} \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x)(x - X_i)^j \right)^2 K_{H_1}(x - X_i)\hat{f}(X_i, Y_i) \tag{17}$$

$$= \min_{\beta(x)} \left(\sum_{i=1}^{N} w_i \right)^{-1} \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x)(x - X_i)^j \right)^2 K_{H_1}(x - X_i) w_i \tag{18}$$

The normalization by $\sum_{i=1}^{N} w_i$ shows the equivalence to the weighted arithmetic mean, with the weights $w_i = \hat{f}(X_i, Y_i)$.

4.3. Asymptotic degeneration of the estimator to the standard LPR

Asymptotically, the proposed estimator degenerates to the standard LPR when the weights w_i are uncorrelated with the standard LPR terms. Formally, as $N \to \infty$, $\hat{\beta}(x) \to \hat{\gamma}(x)$, where $\hat{\gamma}(x)$ is the standard LPR estimator, and the condition that $\left(Y - \sum_{j=0}^{p} \beta_j(x)(x - X)^j \right)^2 K_{H_1}(x - X)$

and $\hat{f}(X,Y)$ are uncorrelated holds. It is assumed that (X_i, Y_i) are independent and identically distributed (i.i.d.) random variables and that $\hat{f}(X,Y) > 0$ almost everywhere.

Proof: Define

$$g(X,Y) := \left(Y - \sum_{j=0}^{p} \beta_j(x)(x-X)^j \right)^2 K_{H_1}(x-X),$$

it follows that:

$$\hat{\beta}(x) := \min_{\beta(x)} \left(\sum_{i=1}^{N} \hat{f}(X_i, Y_i) \right)^{-1} \sum_{i=1}^{N} g(X_i, Y_i)\hat{f}(X_i, Y_i) \tag{19}$$

$$= \min_{\beta(x)} \left(\frac{1}{N} \sum_{i=1}^{N} \hat{f}(X_i, Y_i) \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} g(X_i, Y_i)\hat{f}(X_i, Y_i) \right) \tag{20}$$

As $N \to \infty$, by the law of large numbers, we obtain:

$$\left(\frac{1}{N} \sum_{i=1}^{N} \hat{f}(X_i, Y_i) \right)^{-1} \xrightarrow{a.s.} \frac{1}{\mathbb{E}\left[\hat{f}(X,Y)\right]} \tag{21}$$

$$\frac{1}{N} \sum_{i=1}^{N} g(X_i, Y_i)\hat{f}(X_i, Y_i) \xrightarrow{a.s.} \mathbb{E}\left[g(X,Y)\hat{f}(X,Y)\right] \tag{22}$$

Assuming $\mathbb{E}[\hat{f}(X,Y)] \neq 0$, it follows that:

$$\hat{\beta}(x) \xrightarrow{a.s.} \min_{\beta(x)} \frac{\mathbb{E}\left[g(X,Y)\hat{f}(X,Y)\right]}{\mathbb{E}\left[\hat{f}(X,Y)\right]} \tag{23}$$

If $g(X,Y)$ and $\hat{f}(X,Y)$ are uncorrelated, then:

$$\mathbb{E}\left[g(X,Y)\hat{f}(X,Y)\right] = \mathbb{E}\left[g(X,Y)\right]\mathbb{E}\left[\hat{f}(X,Y)\right] \tag{24}$$

$$\hat{\beta}(x) \xrightarrow{a.s.} \min_{\beta(x)} \mathbb{E}\left[g(X,Y)\right] \tag{25}$$

Therefore, as $N \to \infty$:

$$\hat{\beta}(x) \xrightarrow{a.s.} \min_{\beta(x)} \mathbb{E}\left[\left(Y - \sum_{j=0}^{p} \beta_j(x)(x-X)^j \right)^2 K_{H_1}(x-X) \right] \tag{26}$$

This is the same objective minimized by the standard LPR estimator in the asymptotic sense. Thus, the proposed estimator degenerates to the standard LPR as $N \to \infty$, provided that $g(X,Y)$ and $\hat{f}(X,Y)$ are uncorrelated. Note that one such special case is when $\hat{f}(Y \mid X)$ follows a uniform distribution.

4.4. Asymptotic Equivalence in Expected Loss Function to Standard LPR under the Normality Assumption

In this section, a detailed asymptotic analysis of the proposed estimator is provided under the assumption that the conditional distribution $f(Y \mid X)$ is normal. The goal is to demonstrate that, under certain conditions, the expected loss function simplifies to a form analogous to that of standard local polynomial regression (LPR).

Assumptions

The following assumptions are in place:

(A1) **Independent and Identically Distributed Data:** The data points $\{(X_i, Y_i)\}_{i=1}^N$ are independent and identically distributed (i.i.d.) samples from the joint distribution $f(X, Y)$.

(A2) **Kernel Functions:**

(a) The predictor X is univariate. This simplifies the analysis; extension to multivariate X would require a bandwidth matrix $\mathbf{H}$.

(b) The kernel $K_1(u)$ is a symmetric, bounded, and integrable function satisfying $\int_{-\infty}^{\infty} K_1(u)\,du = 1$.

(c) The bandwidth h for K_1 satisfies $h \to 0$ and $Nh \to \infty$ as $N \to \infty$.

(d) The similarity kernel $K_2\left((x, Y_i), (X_i, Y_i); H_2\right)$ is proportional to an estimate of the conditional density:

$$K_2\left((x, Y_i), (X_i, Y_i); H_2\right) \propto \hat{f}(Y_i \mid X_i),$$

where $\hat{f}(Y_i \mid X_i)$ is a kernel density estimator of the conditional density of Y given X, which converges to $f(Y_i \mid X_i)$ as $N \to \infty$.

(A3) **Conditional Normality:** The conditional distribution $f(Y \mid X = x)$ is normally distributed with mean $m(X)$ and variance $\sigma^2(X)$:

$$f(Y \mid X = x) \sim \mathcal{N}\left(m(X), \sigma^2(X)\right).$$

(A4) **Smoothness of Functions:** The regression function $m(X)$ and the variance function $\sigma^2(X)$ are twice continuously differentiable in a neighborhood of x.

(A5) **Finite and Positive Variance:** The variance $\sigma^2(X)$ is finite and strictly positive for all X in the domain.

Expected Loss Function

Recall that the proposed loss function is:

$$\mathcal{L}_{rsk}(x) = \sum_{i=1}^{N} \left(Y_i - \beta(X_i; x)\right)^2 K_1\left(\frac{X_i - x}{h}\right) K_2\left((x, Y_i), (X_i, Y_i); H_2\right), \tag{27}$$

where

$$\beta(X_i; x) = \sum_{j=0}^{p} \beta_j(x)(X_i - x)^j.$$

Under assumption (A2d), $K_2\left((x, Y_i), (X_i, Y_i); H_2\right) \propto f(Y_i \mid X_i)$ as $N \to \infty$. For the purpose of asymptotic analysis, it is assumed that $N \to \infty$ and that K_2 is proportional to $f(Y_i \mid X_i)$. Therefore, the expected loss function can be expressed as:

$$\mathbb{E}\left[\mathcal{L}_{rsk}(x)\right] = N \iint (Y - \beta(X;x))^2 \, K_1\left(\frac{X-x}{h}\right) K_2\left((x,Y),(X,Y);H_2\right) f(X,Y)\, dY\, dX. \tag{28}$$

Substituting $K_2 \propto f(Y \mid X)$ and $f(X,Y) = f(Y \mid X)f(X)$, we have:

$$\mathbb{E}\left[\mathcal{L}_{rsk}(x)\right] \propto N \iint (Y - \beta(X;x))^2 \, K_1\left(\frac{X-x}{h}\right) \left[f(Y \mid X)\right]^2 f(X)\, dY\, dX, \tag{29}$$

which further simplifies to:

$$\mathbb{E}\left[\mathcal{L}_{rsk}(x)\right] \propto N \int f(X) K_1\left(\frac{X-x}{h}\right) \int (Y - \beta(X;x))^2 \left[f(Y \mid X)\right]^2 \, dY\, dX, \tag{30}$$

Simplifying the Inner Integral of Y

Under assumption (A3), $f(Y \mid X)$ follows a normal distribution:

$$f(Y \mid X) = \frac{1}{\sqrt{2\pi\sigma^2(X)}} \exp\left(-\frac{(Y - m(X))^2}{2\sigma^2(X)}\right). \tag{31}$$

Therefore, the squared conditional density is:

$$\left[f(Y \mid X)\right]^2 = \left(\frac{1}{\sqrt{2\pi\sigma^2(X)}}\right)^2 \exp\left(-\frac{(Y - m(X))^2}{\sigma^2(X)}\right). \tag{32}$$

This is proportional to a normal density with mean $m(X)$ and variance $\sigma^2(X)/2$:

$$\left[f(Y \mid X)\right]^2 \propto \phi\left(Y; m(X), \frac{\sigma^2(X)}{2}\right), \tag{33}$$

where $\phi(Y;\mu,\sigma^2)$ is the normal density function.

Evaluating the Inner Integral

The inner integral is:

$$E_Y = \int (Y - \beta(X;x))^2 \left[f(Y \mid X)\right]^2 \, dY. \tag{34}$$

Since $\left[f(Y \mid X)\right]^2$ is proportional to $\phi\left(Y; m(X), \sigma^2(X)/2\right)$, we can evaluate E_Y using the expectation of square deviations under the normal distribution:

$$E_Y = (m(X) - \beta(X;x))^2 + \frac{\sigma^2(X)}{2}. \tag{35}$$

Simplifying the Expected Loss Function

Substituting E_Y back into equation (29), we have:

$$\mathbb{E}\left[\mathcal{L}_{rsk}(x)\right] \propto N \int f(X) K_1\left(\frac{X-x}{h}\right)\left((m(X)-\beta(X;x))^2 + \frac{\sigma^2(X)}{2}\right) dX. \tag{36}$$

Since $\frac{\sigma^2(X)}{2}$ does not depend on $\beta_j(x)$, minimizing $\mathbb{E}[\mathcal{L}_{rsk}(x)]$ with respect to $\beta_j(x)$ is equivalent to minimizing:

$$J(\beta(x)) = \int f(X) K_1\left(\frac{X-x}{h}\right)(m(X)-\beta(X;x))^2 \; dX. \tag{37}$$

This is the same objective function minimized by the standard LPR in expectation. Therefore, under the normality assumption and as $N \to \infty$, the expected loss functions of the proposed method and the standard LPR converge, indicating that both methods target the same regression function $m(x)$ on average.

Implications and Discussion

The convergence of the expected loss functions suggests that the robust method and standard LPR aim to estimate the same underlying regression function asymptotically. However, this equivalence in expectation does not imply identical behavior in distribution. The robust method's weighting scheme, which adjusts the influence of data points based on their estimated densities, affects the variability of the estimator.

By down-weighting outliers and points in regions of low density, the robust method can exhibit different variance properties, potentially improving robustness and efficiency in finite samples, particularly under heteroscedastic noise or the presence of outliers. This variance adjustment underscores the practical advantage of the robust method but also highlights the need for further theoretical analysis to fully characterize its asymptotic variance and distribution.

4.5. Asymptotic Bias under Non-Normal Conditional Distributions

While the proposed robust estimator aligns asymptotically with standard local polynomial regression (LPR) under the assumption of conditional normality, real-world data often deviate from this idealized condition. When the conditional distribution $f(Y \mid X)$ is not normal, particularly if it exhibits asymmetry, the asymptotic behavior of the estimator can be affected, potentially introducing bias.

To explore the implications of non-normal conditional distributions on the asymptotic properties of the proposed estimator, consider the expected loss function:

$$\mathbb{E}\left[\mathcal{L}_{\text{rsk}}(x)\right] = N \iint (Y-\beta(X;x))^2\, K_1\left(\frac{X-x}{h}\right) K_2\left((x,Y),(X,Y);H_2\right) f(X,Y)\, dY\, dX. \tag{38}$$

Assuming that $K_2 \propto f(Y \mid X)$ and recognizing that $f(X,Y) = f(Y \mid X) f(X)$, the expected loss simplifies to:

$$\mathbb{E}\left[\mathcal{L}_{\text{rsk}}(x)\right] \propto N \iint (Y - \beta(X;x))^2 K_1\left(\frac{X-x}{h}\right)[f(Y \mid X)]^2 f(X)\, dY\, dX. \tag{39}$$

When $f(Y \mid X)$ is asymmetric, the squared conditional density $[f(Y \mid X)]^2$ alters the weighting in the loss function in a way that can shift the effective mean and variance. Specifically, the expected value of Y under the squared density $[f(Y \mid X)]^2$ is generally not equal to the mean $m(X)$ of the original conditional distribution.

This shift implies that the minimization of the expected loss function may lead the estimator to converge to a value different from the true regression function $m(X)$, introducing an asymptotic bias. The magnitude and direction of this bias depend on the nature of the asymmetry in $f(Y \mid X)$.

To quantify the asymptotic bias in a general sense, consider that the mean of the squared conditional density $[f(Y \mid X)]^2$ is given by:

$$\mu'(X) = \frac{\int Y[f(Y \mid X)]^2\, dY}{\int [f(Y \mid X)]^2\, dY}. \tag{40}$$

Similarly, the variance under the squared density is:

$$\sigma'^2(X) = \frac{\int (Y - \mu'(X))^2 [f(Y \mid X)]^2\, dY}{\int [f(Y \mid X)]^2\, dY}. \tag{41}$$

The expected loss function then becomes:

$$\mathbb{E}\left[\mathcal{L}_{\text{rsk}}(x)\right] \propto N \int K_1\left(\frac{X-x}{h}\right) f(X) \left(\left(\mu'(X) - \beta(X;x)\right)^2 + \sigma'^2(X)\right) dX. \tag{42}$$

Since $\sigma'^2(X)$ does not depend on $\beta(X;x)$, minimizing $\mathbb{E}\left[\mathcal{L}_{\text{rsk}}(x)\right]$ with respect to $\beta(X;x)$ is equivalent to minimizing:

$$J(\beta(X;x)) = \int K_1\left(\frac{X-x}{h}\right) f(X) \left(\mu'(X) - \beta(X;x)\right)^2\, dX. \tag{43}$$

Therefore, the estimator $\beta(X;x)$ converges to $\mu'(X)$ rather than $m(X)$. The asymptotic bias at point x can thus be quantified as:

$$\text{Bias}(x) = \mu'(x) - m(x). \tag{44}$$

This bias arises because the mean under the squared conditional density $\mu'(X)$ differs from the mean $m(X)$ of the original conditional distribution $f(Y \mid X)$. The amount of bias depends on the degree and nature of asymmetry in $f(Y \mid X)$.

A detailed example illustrating this effect, including specific calculations of $\mu'(X)$ and $\sigma'^2(X)$ for a particular asymmetric distribution, is provided in Appendix Appendix A. This example demonstrates how the asymmetry of $f(Y \mid X)$ can lead to a shift in the estimator's asymptotic target due to the squared density weighting.

In practice, the presence of asymmetry in the conditional distribution may introduce some bias into the estimator. However, the robust weighting scheme of the proposed method can still provide advantages in terms of reducing the influence of outliers and improving estimation in the presence of heteroscedasticity or heavy-tailed errors. The trade-off between asymptotic bias and robustness to outliers should be considered in practical applications.

4.6. Relationship to Kernel Methods and RKHS

In this subsection, the relationship of the proposed method to kernel methods and Reproducing Kernel Hilbert Spaces (RKHS) is explored. The use of positive definite kernels in defining the weights $K_{\mathcal{D}}$ allows the proposed estimator to be interpreted within the RKHS framework, providing deeper insights into its properties and connections to existing kernel-based methods.

Recall that in the proposed method, the weights in the loss function (4) are defined using a compound positive definite kernel $K_{\mathcal{D}}$ on the data domain $\mathcal{D}$:

$$\mathcal{L}_{\mathrm{rsk}}(x, y; \mathcal{D}_N, H) := \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x, y)(x - X_i)^j \right)^2 K_{\mathcal{D}}\left((x, y), (X_i, Y_i); H\right). \tag{45}$$

As per equation (6), the kernel $K_{\mathcal{D}}$ is defined as a product of two positive definite kernels:

$$K_{\mathcal{D}}\Big((x, y), (x', y'); H_1, H_2\Big) = K_1(x, x'; H_1) \cdot K_2\Big((x, y), (x', y'); H_2\Big), \tag{46}$$

where K_1 is a kernel function depending only on the predictors x and x', typically chosen as the traditional distance-based kernel used in local polynomial regression, and K_2 is a kernel function that incorporates both predictors and responses.

Since $K_{\mathcal{D}}$ is a product of positive definite kernels, it is itself a positive definite kernel. Therefore, there exists a feature mapping $\phi : \mathcal{D} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space, such that:

$$K_{\mathcal{D}}\Big((x, y), (x', y')\Big) = \langle \phi(x, y), \phi(x', y') \rangle_{\mathcal{H}}. \tag{47}$$

Thus, the weights $K_{\mathcal{D}}((x, y), (X_i, Y_i))$ can be interpreted as inner products in the feature space $\mathcal{H}$. Consequently, the loss function (45) can be viewed as a weighted least squares problem where the weights are determined by the similarity between the feature representations of the data points and the point of interest.

Furthermore, consider the role of the Kernel Density Estimator (KDE) in the proposed method. The KDE at a point (x, y) is given by:

$$\hat{f}(x, y) = \frac{1}{N} \sum_{i=1}^{N} K_2\left((x, y), (X_i, Y_i); H_2\right). \tag{48}$$

Note that the KDE uses K_2, not the full kernel $K_{\mathcal{D}}$, since K_2 is the kernel used in the density estimation of the joint (or conditional) distribution involving both x and y.

Since K_2 is a positive definite kernel, there exists a feature mapping $\psi : \mathcal{D} \to \mathcal{G}$, where $\mathcal{G}$ is another Hilbert space, such that:

$$K_2\left((x,y),(x',y')\right) = \langle \psi(x,y), \psi(x',y') \rangle_{\mathcal{G}}. \tag{49}$$

Therefore, the KDE at (x,y) can be expressed in terms of inner products in the feature space $\mathcal{G}$:

$$\hat{f}(x,y) = \frac{1}{N}\sum_{i=1}^{N} \langle \psi(x,y), \psi(X_i,Y_i) \rangle_{\mathcal{G}}. \tag{50}$$

This expression shows that the KDE at (x,y) is proportional to the inner product between the feature mapping $\psi(x,y)$ and the mean of the feature mappings of the data:

$$\hat{v}_{\psi} = \frac{1}{N}\sum_{i=1}^{N} \psi(X_i,Y_i), \tag{51}$$

so that:

$$\hat{f}(x,y) = \langle \psi(x,y), \hat{v}_{\psi} \rangle_{\mathcal{G}}. \tag{52}$$

This interpretation shows that the KDE measures how closely the feature representation $\psi(x,y)$ of a point (x,y) aligns with the average feature representation $\hat{v}_{\psi}$ of the data in the space induced by K_2. In the proposed method, this alignment influences the weights in the regression, as the density estimates $\hat{f}(x,y)$ or $\hat{f}(Y_i \mid X_i)$ derived from K_2 directly affect the overall weights $K_{\mathcal{D}}((x,y),(X_i,Y_i))$. This interplay underpins the robustness and adaptability of the proposed method.

By leveraging positive definite kernels for defining $K_{\mathcal{D}}$, the method inherently operates within the RKHS framework, where weights represent similarities in feature space. This perspective highlights the connection between the kernel-based weighting and the feature mappings, offering insights into the estimator's flexibility and robustness.

5. Experiments and Implementation Notes

The proposed method was implemented in Python. Following [3], distances between pairs in each neighborhood are normalized to the range $[0,1]$. For the kernel function $K_1(x,x';H_1)$, a simple Laplacian kernel $e^{-\|x-x'\|}$ was used, as it demonstrated more efficient and consistent empirical performance than the tricube kernel suggested in [3]. For density estimation, a factorized multidimensional KDE with scaled Gaussian kernels was applied. Five bandwidth estimation methods were tested: Scott's rule [12], the Normal Reference rule, global Least Squares Cross-Validation (LSCV), local LSCV, and local Modified Least Squares Cross-Validation (MLCV). In some experiments, the bandwidth for the predictor kernel was empirically adjusted as a simple function of the window size.

For computational efficiency, certain calculations were omitted because the local regression in equation (5) is invariant to the scale of the weights. This includes excluding scaling constants fixed within a neighborhood for a specific local regression target, such as those in computing $\hat{f}(y \mid x)$ and $\hat{f}(x, y)$ in equations (7) and (11), respectively.

Experiments were conducted using a variety of synthetic benchmarks to evaluate the performance of the method's linear and quadratic variants against other local polynomial regression methods. These include LOWESS and iterative robust LOWESS [13], local linear and local constant kernel regression [13], local quadratic regression [14], and radial basis function networks [14].

The experimental setups included several non-linear synthetic curves and planes with added noise, representing dense and sparse data, homoscedastic and heteroscedastic noise characteristics, and different neighborhood sizes. The results indicate that no single method universally outperforms the others; performance varies by setting. However, the proposed method exhibited competitive performance overall, delivering the best results across numerous settings, particularly in heteroscedastic environments. It also generally outperformed its direct counterparts, LOESS/LOWESS and quadratic LPR, with just a single iteration. Moreover, the proposed method showed lower sensitivity to neighborhood size, resulting in reduced variance. This stability makes it an attractive option, especially when choosing hyperparameters without ground truth data.

The complete experimental results are available as interactive Jupyter notebooks at `https://nbviewer.org/github/yaniv-shulman/rsklpr/tree/main/src/experiments/` [1].

Simulation Studies

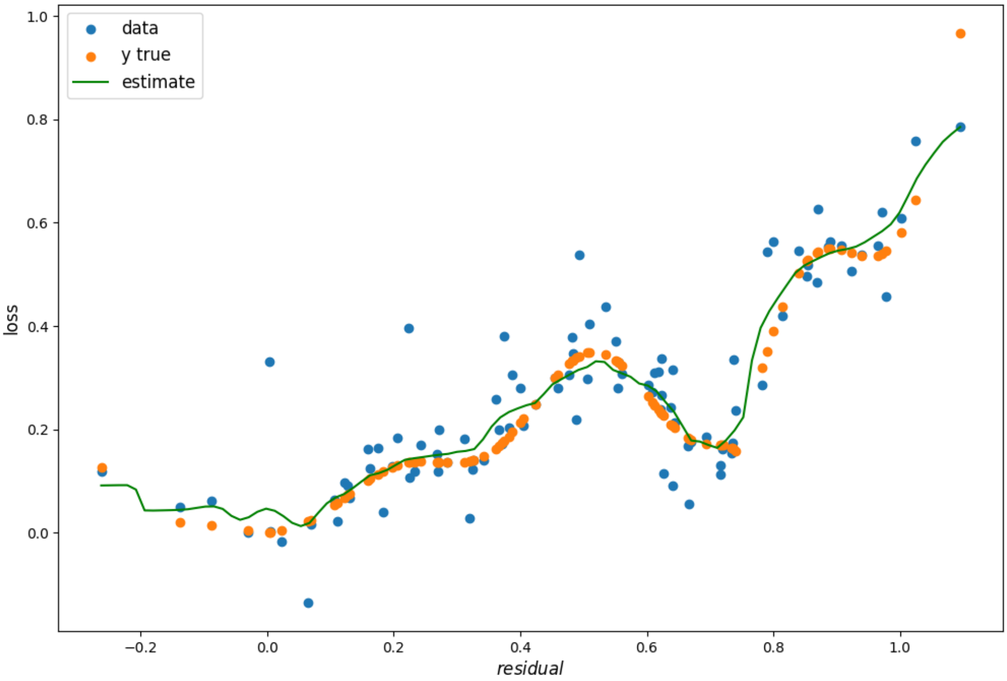
TODO: Perform empirical studies to assess the impact of asymmetry on the estimator's performance in realistic settings. Such studies may reveal that the bias introduced by asymmetry is offset by the robustness gains in finite samples.

6. Future Work and Research Directions

This work introduces a new robust variant of Local Polynomial Regression (LPR), opening several avenues for further exploration and refinement. Since the proposed method generalizes the traditional LPR, there are opportunities to replace certain standard components in equation (5) with more robust alternatives. These could include approaches such as robust methods for bandwidth selection or substituting the conventional quadratic residual function with alternatives better suited for handling outliers.

Additionally, further development of this framework may involve exploring different kernel functions K_D and assessing how robust density estimators influence overall performance. Extending the method within the RKHS framework presents another valuable direction. This could allow for the introduction of a regularization term in the loss function, enhancing control over estimator smoothness and mitigating the risk of overfitting. Through these future directions, the robustness and adaptability of the proposed method could be substantially advanced.

Figure 3: Regression example of synthetically generated 1D data with heteroscedastic noise. Additional experimental results and demonstrations including multivariate settings and bootstrap based confidence intervals are available at `https://nbviewer.org/github/yaniv-shulman/rsklpr/tree/main/src/experiments/` as interactive Jupyter notebooks [1]



1

References

[1] Project jupyter is a non-profit, open-source project, born out of the ipython project in 2014 as it evolved to support interactive data science and scientific computing across all programming languages. `https://jupyter.org/`.

[2] M. Avery. Literature review for local polynomial regression. 2010.

[3] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[4] W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[5] J. Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21, 03 1993.

[6] E. García-Portugués. *Notes for Nonparametric Statistics*. 2023. Version 6.9.0. ISBN 978-84-09-29537-1.

[7] T. Gasser and H.-G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11:171–185, 1984.

[8] R. A. Maronna, D. Martin, V. J. Yohai, and Hardcover. Robust statistics: Theory and methods. 2006.

[9] H.-G. Muller. Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, 82(397):231–238, 1987.

[10] E. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9:141–142, 1964.

[11] M. Salibian-Barrera. Robust nonparametric regression: Review and practical considerations. *Econometrics and Statistics*, 2023.

[12] D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 2015.

[13] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[14] sigvaldm. Localreg is a collection of kernel-based statistical methods. `https://github.com/sigvaldm/localreg`.

[15] V. G. Spokoiny. Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *The Annals of Statistics*, 26(4):1356 – 1378, 1998.

[16] C. J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5:595–620, 1977.

[17] P. Čížek and S. Sadıkoğlu. Robust nonparametric regression: A review. *WIREs Comput. Stat.*, 12(3), apr 2020.

[18] G. S. Watson. Smooth regression analysis. 1964.

Appendix A. Asymptotic Bias Example with Exponential Conditional Distribution

In this appendix, we provide a detailed example illustrating how asymmetry in the conditional distribution $f(Y \mid X)$ can introduce asymptotic bias in the proposed estimator due to the squared density weighting. Specifically, we consider the case where the conditional distribution of Y given X is exponential, a common asymmetric distribution.

Appendix A.1. Setup of the Example

Suppose that for each fixed X, the conditional distribution $Y \mid X$ follows an exponential distribution shifted by $m(X)$:

$$f(Y \mid X) = \lambda(X) \exp\left(-\lambda(X)(Y - m(X))\right), \quad \text{for } Y \geq m(X), \tag{A.1}$$

where $\lambda(X) > 0$ is the rate parameter, and $m(X)$ is the location parameter (shift), which represents the true regression function we aim to estimate.

The mean and variance of this distribution are:

$$\mathbb{E}[Y \mid X] = m(X) + \frac{1}{\lambda(X)}, \tag{A.2}$$

$$\mathrm{Var}[Y \mid X] = \frac{1}{\lambda^2(X)}. \tag{A.3}$$

Appendix A.2. Computing the Squared Density

The squared conditional density is:

$$[f(Y \mid X)]^2 = (\lambda(X))^2 \exp\left(-2\lambda(X)(Y - m(X))\right), \quad \text{for } Y \geq m(X). \tag{A.4}$$

This squared density is proportional to an exponential distribution with rate parameter $2\lambda(X)$:

$$g(Y \mid X) = \frac{[f(Y \mid X)]^2}{\int_{m(X)}^{\infty} [f(u \mid X)]^2 \, du} = 2\lambda(X) \exp\left(-2\lambda(X)(Y - m(X))\right), \quad \text{for } Y \geq m(X). \tag{A.5}$$

Appendix A.3. Calculating the Mean and Variance under the Squared Density

The mean and variance of Y under the squared density $g(Y \mid X)$ are:

$$\mu'(X) = \mathbb{E}_g[Y \mid X] = m(X) + \frac{1}{2\lambda(X)}, \tag{A.6}$$

$$\sigma'^2(X) = \mathrm{Var}_g[Y \mid X] = \frac{1}{(2\lambda(X))^2}. \tag{A.7}$$

Appendix A.4. Deriving the Asymptotic Bias

As per the analysis in Section 4.5, the expected loss function simplifies to:

$$\mathbb{E}\left[\mathcal{L}_{\mathrm{rsk}}(x)\right] \propto N \int K_1\left(\frac{X-x}{h}\right) f(X)\left(\left(\mu'(X)-\beta(X;x)\right)^2+\sigma'^2(X)\right) dX. \tag{A.8}$$

Minimizing with respect to $\beta(X;x)$ leads to the estimator converging to $\beta(X;x) = \mu'(X)$. Therefore, the asymptotic bias at point x is:

$$\begin{aligned}\mathrm{Bias}(x) &= \mu'(x) - m(x) &\text{(A.9)}\\ &= \left(m(x) + \frac{1}{2\lambda(x)}\right) - m(x) &\text{(A.10)}\\ &= \frac{1}{2\lambda(x)}. &\text{(A.11)}\end{aligned}$$

This expression shows that the estimator is asymptotically biased upwards by $\frac{1}{2\lambda(x)}$ compared to the true regression function $m(x)$.

Appendix A.5. Interpretation

The bias arises because the weighting induced by $[f(Y \mid X)]^2$ effectively shifts the mean of the distribution used in the expected loss function. In the case of the exponential distribution, squaring the density doubles the rate parameter from $\lambda(X)$ to $2\lambda(X)$, reducing the mean from $m(X) + \frac{1}{\lambda(X)}$ to $m(X) + \frac{1}{2\lambda(X)}$.

This shift means that the estimator targets $\mu'(X)$ instead of $m(X)$, resulting in an asymptotic bias proportional to $\frac{1}{2\lambda(X)}$.

Appendix A.6. Numerical Example

For illustrative purposes, consider $\lambda(X) = 1$ for all X. Then, the bias simplifies to:

$$\mathrm{Bias}(x) = \frac{1}{2}. \tag{A.12}$$

In this case, the estimator is asymptotically biased upwards by 0.5 units at every point x.

Appendix A.7. Implications for the Estimator

This example demonstrates that when the conditional distribution $f(Y \mid X)$ is asymmetric, the proposed estimator may not converge to the true regression function $m(X)$ as $N \to \infty$, but rather to a biased version shifted by $\mu'(X) - m(X)$.

In practice, the magnitude of the bias depends on the degree of asymmetry and the rate parameter $\lambda(X)$. For large $\lambda(X)$, the bias diminishes, and the estimator approaches $m(X)$. However, for small $\lambda(X)$, the bias becomes more significant.

Appendix A.8. Conclusion

This example illustrates how asymmetry in the conditional distribution $f(Y \mid X)$ can introduce asymptotic bias in the proposed estimator due to the squared density weighting. It underscores the importance of considering the nature of the conditional distribution when applying the robust estimator and highlights the potential trade-off between robustness to outliers and asymptotic bias.

In practical applications, one may need to assess whether the benefits of robustness outweigh the potential bias introduced, especially in cases where the conditional distribution is significantly asymmetric.