Outlier Detection and Robust Estimation in Nonparametric Regression

Dehan Kong
University of Toronto

Howard Bondell
University of Melbourne

Weining Shen
University of California, Irvine

Abstract

This paper studies outlier detection and robust estimation for nonparametric regression problems. We propose to include a subject-specific mean shift parameter for each data point such that a nonzero parameter will identify its corresponding data point as an outlier. We adopt a regularization approach by imposing a roughness penalty on the regression function and a shrinkage penalty on the mean shift parameter. An efficient algorithm has been proposed to solve the double penalized regression problem. We discuss a data-driven simultaneous choice of two regularization parameters based on a combination of generalized cross validation and modified Bayesian information criterion. We show that the proposed method can consistently detect the outliers. In addition, we obtain minimax-optimal convergence rates for both the regression function and the mean shift parameter under regularity conditions. The estimation procedure is shown to enjoy the oracle property in the sense that the convergence rates agree with the minimax-optimal rates when the outliers (or regression function) are known in advance. Numerical results demonstrate that the proposed method has desired performance in identifying outliers under different scenarios.

1 Introduction

Outliers are observations that deviate markedly from the majority of data. They are commonly encountered in real data applications such as genomics, biomedical imaging and signal processing. In the presence of outliers, likelihood-based inference can be unreliable. For example, ordinary least squares estimates for regression problems are highly sensitive to outliers. To facilitate valid statistical inference, an active area of research has been devoted to outlier detection and robust statistical estimation. Popular methods include M-estimators (Huber, 1981), Generalized M-estimators (Mallows, 1975), least median of squares (Hampel, 1975), least trimmed squares (Rousseeuw, 1984), S-estimators (Rousseeuw and Yohai, 1984), MM-estimators (Yohai, 1987), weighted least squares (Gervini and Yohai, 2002) and empirical likelihood (Bondell and Stefanski, 2013). Although many of the existing robust regression approaches enjoy nice theoretical properties and satisfactory numerical performances, they usually focus on linear regression models. While nonparametric regression models have been widely used in modern statistics, there is a considerable gap in the literature on the extension of aforementioned methods to nonparametric regression problems, in which identifying outliers may be more challenging because outliers can be more easily associated with the majority of data via a nonparametric function than a linear curve. There are a few robust nonparametric estimation methods such as Cleveland (1979), Brown, Cai, and Zhou (2008) and Cai and Zhou (2009). However, these methods can only estimate the nonparametric function, and none of them can be applied to outlier detection.

In this paper, we fill in this gap by considering outlier detection and robust estimation simultaneously for nonparametric regression problems. We use univariate regression $y_i = f(x_i) + \epsilon_i$ as an illustrative example and propose to include a subject-specific mean shift parameter in the model. In particular, we add an additional subject-specific term into the nonparametric regression model, i.e. $y_i = f(x_i) + \gamma_i + \epsilon_i$, where a nonzero mean shift parameter γ_i indicates that the ith observation is an outlier. Then the problem becomes estimation of the regression function, f and mean shift parameters, γ_i's. This idea originates from Gannaz (2006); McCann and Welsch (2007); She and Owen

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

(2011) in the context of linear models, however, the extension from linear model to nonparametric models requires nontrivial effort and the results are much more flexible and useful in practice. The proposed method is not restricted to particular domains, but in general applicable for a wide range of domains including univariate and multivariate data. The extension from univariate data to the multi-dimensional and high-dimensional data is discussed in Section 6. Mateos and Giannakis (2012) proposed a robust estimation procedure based on a similar model as ours, however, there are several key differences between our paper and this reference. First, our algorithm is different and much faster. We only need to update γ_i's iteratively, while Mateos and Giannakis (2012) need to iteratively update γ_i's and f. Second, outlier detection is an important goal of our paper, and we study the performance of the method in terms of outlier detection, while the reference only focused on robust function estimation. Third, we have a better tuning method. The tuning method in Mateos and Giannakis (2012) depends on an initial function fit, which is computationally much slower. More severely, their initial function fit may not be robust and may result in a bad estimate of the error variance. As their tuning criterion is based on the estimate of the error variance, the tuning parameters selected could be completely misleading. Finally, we have investigated the asymptotic theory of our method, which is not included in that reference.

Our theoretical studies are concerned with the consistency of outlier detection and the so-called "oracle property" of the estimators. Specifically, we define the "oracle estimate" of f as the one obtained given all the outliers are known. Then an estimator of f is said to satisfy the oracle property if it possesses the same minimax-optimal convergence rate as the oracle estimate. The oracle property for mean shift parameter estimators can be defined in a similar way. A major contribution of our paper is that we derive sufficient conditions on the tuning parameters such that the estimators of f and $\boldsymbol{\gamma}$ satisfy the oracle property. In other words, our estimation procedure is not affected by the additional step of identifying the outliers. The main technique we use here is based on Müller and van de Geer (2015) with modifications to accommodate the mean shift parameter component in our case.

For mean shift regression models, regularization methods are commonly used to detect the outliers. For example, McCann and Welsch (2007) considered an L_1 regularization. She and Owen (2011) imposed a nonconvex penalty function on γ_i's to obtain a sparse solution. Kong et al. (2018) imposed an adaptive penalty function on γ_i's to obtain fully efficient robust estimation and outlier detection. In this paper, we adopt a general penalized regression framework by considering popular penalty functions such as LASSO (Tibshirani, 1996) and smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001). The proposed method also applies to multivariate nonparametric regression and semi-parametric models (e.g., partial linear models). A major challenge in extending the previous work in linear models lies in accurate and efficient estimation of both the nonparametric function and the mean shift parameters at the same time. In the literature, nonparametric estimation is usually achieved via a smoothing procedure, such as local polynomial smoothing (Fan and Gijbels, 1996), polynomial splines (Hastie and Tibshirani, 1990), regression splines (Agarwal and Studden, 1980) and smoothing splines (Wahba, 1990; Gu, 2013). In this paper, we adopt smoothing splines for the nonparametric function estimates, and propose an efficient algorithm to solve an optimization problem that involves selecting two different tuning parameters simultaneously.

The rest of the paper is organized as follows. Section 2 describes our methodology including the problem formulation, computational algorithm and tuning parameter selection. Section 3 discusses the convergence rate for our nonparametric estimates. We present some simulation results to evaluate the finite-sample performance of the proposed method in Section 4. In Section 5, we apply our method to the baseball data. We discuss some extensions to multi-dimensional and high-dimensional models in Section 6 and conclude with some remarks in Section 7. A proof sketch of the theorems is given in Section 8.

2 Methodology

We consider a univariate nonparametric mean shift model as follows,

$$y_i = f(x_i) + \gamma_i + \epsilon_i, \tag{1}$$

where the covariate x_i, lies in a bounded closed interval on the real line, and ϵ_i's are i.i.d random errors with mean 0 and finite second moment. We are interested in using mean shift parameters γ_i's as indicators of the outliers in the nonparametric regression of y_i given x_i. More precisely, if $\gamma_i \neq 0$, then its corresponding subject i is an outlier. Similarly, if $\gamma_i = 0$, subject i is a normal data point. Suppose we have n samples (x_i, y_i), and we assume the number of outliers is less than $\lceil n/2 \rceil$. In other words, less than half of the γ_i's are nonzero, and this assumption guarantees identifiability of our model.

We consider a shrinkage approach for outlier detection by pushing most of the γ_i's toward zero. Clearly,

those nonzero estimates will represent the outliers we detect under the regression model. In addition to outlier detection, we are also interested in robust estimation of the nonparametric function $f(\cdot)$. This is achieved by adopting a smoothing spline technique with a roughness penalty, which is based on the second order derivative of f.

Throughout the paper, we assume $x_i \in [0,1]$ without loss of generality. For general situations, the proposed method will still work by first scaling x_i's onto the unit interval. Denote $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^T$. We propose to solve the following minimization problem

$$\begin{aligned} &\text{argmin}_{f,\boldsymbol{\gamma}}\{\textstyle\sum_{i=1}^n (y_i - f(x_i) - \gamma_i)^2 \\ &+ \textstyle\sum_{i=1}^n P(\gamma_i, \kappa) + \lambda \int_0^1 [f''(x)]^2 dx\}, \end{aligned} \tag{2}$$

where $P(\gamma_i, \kappa)$ is a general penalty function (e.g., LASSO and SCAD), and κ and λ serve as tuning parameters, which control the outlier detection and nonparametric smoothing respectively.

It has been shown that problem (2) can be solved efficiently by reformulating the problem itself as a penalized regression problem and constructing a reproducing kernel corresponding to the penalty term $\int_0^1 [f''(x)]^2 dx$, see Wahba (1990) and Gu (2013) for more details.

In particular, denote $k_1(s) = s - 0.5$, $k_2(s) = \frac{1}{2}(k_1^2(s) - \frac{1}{12})$ and $k_4(s) = \frac{1}{24}(k_1^4(s) - \frac{k_1^2(s)}{2} + \frac{7}{240})$. Define

$$K(s,t) = k_2(s)k_2(t) - k_4(s-t). \tag{3}$$

By the representer theorem (Kimeldorf and Wahba, 1971), for $f(\cdot)$, the solution of (2) can be represented as

$$f(x) \quad = \quad d_1 + d_2 x + \sum_{i=1}^n \alpha_i K(x_i, x),$$

where K is the kernel function defined in (3), and d_1, d_2 are unknown parameters.

Let $\mathbf{1}$ be a column vector of 1, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ and $\mathbf{X} = (x_1, \ldots, x_n)^T$. Denote $\mathbf{K}$ be a $n \times n$ matrix with its (i,j)-th element taking value of $K(x_i, x_j)$. We can reformulate (2) as

$$\begin{aligned} &\text{argmin}_{d_1, d_2, \boldsymbol{\alpha}, \boldsymbol{\gamma}}\{||\mathbf{y} - d_1\mathbf{1} - d_2\mathbf{X} - \mathbf{K}\boldsymbol{\alpha} - \boldsymbol{\gamma}||_2^2 \\ &+ \textstyle\sum_{i=1}^n P(\gamma_i, \kappa) + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}\}. \end{aligned} \tag{4}$$

Denote $N = (\mathbf{1}, \mathbf{X}, \mathbf{K})$ and $\boldsymbol{\theta} = (d_1, d_2, \boldsymbol{\alpha}^T)^T$. We also define

$$\mathbf{L} = \begin{pmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times n} \\ \mathbf{0}_{n\times 2} & \mathbf{K} \end{pmatrix}.$$

Notice that for a fixed $\boldsymbol{\gamma}$, we have

$$\hat{\boldsymbol{\theta}} \quad = \quad (N^T N + \lambda \mathbf{L})^{-1} N^T (\mathbf{y} - \boldsymbol{\gamma}). \tag{5}$$

Thus, we can use the profiling idea to express the optimization problem strictly as a function of $\boldsymbol{\gamma}$. In particular, denote $H_\lambda = N(N^T N + \lambda L)^{-1} N^T$. By plugging (5) into (4), the optimization problem becomes

$$\text{argmin}_{\boldsymbol{\gamma}}\{(\mathbf{y} - \boldsymbol{\gamma})^T (I - H_\lambda)(\mathbf{y} - \boldsymbol{\gamma}) + \sum_{i=1}^n P(\gamma_i, \kappa)\}. \tag{6}$$

For large sample sizes, solving (6) can be expensive even for some popular choice of penalty functions. For example, if we use the LASSO penalty, the problem can be reformulated as a quadratic programming problem or a full sequence of solutions for κ can be found via the Least Angle Regression (LARS) algorithm (Efron et al., 2004) for each fixed λ. Since the number of parameters equals the sample size n, when n becomes large, the quadratic programming or LARS would be computationally slow. If we use the SCAD penalty, we may consider using a concave convex procedure (Kim, Choi, and Oh, 2008). However, the computation will still be quite expensive.

To overcome computational burdens, we adopt the Thresholding-based Iterative Selection Procedure (TISP) proposed in She (2009). TISP provides a feasible way to tackle the optimization problem in (4). It is a simple procedure that does not involve complicated operations such as matrix inversion. The general idea is to update $\boldsymbol{\gamma}$ by an iterative thresholding rule. More specifically, we write $(\mathbf{y} - \boldsymbol{\gamma})^T (I - H_\lambda)(\mathbf{y} - \boldsymbol{\gamma}) = ||(I - H_\lambda)^{1/2}\mathbf{y} + \{I - (I - H_\lambda)^{1/2}\}\boldsymbol{\gamma} - \boldsymbol{\gamma}||_2^2$. In this way, we can update $\boldsymbol{\gamma}$ via

$$\begin{aligned} \boldsymbol{\gamma}^{(j+1)} \quad &= \Theta(\{I - (I - H_\lambda)^{1/2}\}\boldsymbol{\gamma}^{(j)} \\ &\quad + (I - H_\lambda)^{1/2}\mathbf{y}, \kappa). \end{aligned} \tag{7}$$

where $\Theta(\cdot, \kappa)$ is a threshold function corresponding to the penalty function $P(\cdot, \kappa)$. For example, if we use the LASSO penalty, the threshold function would be

$$\Theta(x, \kappa) = \begin{cases} 0, & |x| \le \kappa; \\ \text{sgn}(x)(|x| - \kappa), & |x| > \kappa, \end{cases}$$

where $\text{sgn}(\cdot)$ is the sign function. If the SCAD penalty is used, the threshold function is defined as

$$\Theta(x, \kappa) = \begin{cases} 0, & |x| \le \kappa; \\ \text{sgn}(x)(|x| - \kappa), & \kappa < |x| \le 2\kappa; \\ \{(a-1)x - \text{sgn}(x)a\kappa\}/(a-2), & 2\kappa < |x| \le a\kappa; \\ x, & |x| > a\kappa, \end{cases}$$

where $a = 3.7$ suggested by Fan and Li (2001).

To this end, our algorithm of solving (2) can be summarized as follows. First step, we solve for $\boldsymbol{\gamma}^{(j)}$ using

(7) until convergence, and the limit would be $\hat{\boldsymbol{\gamma}}$. The starting point $\boldsymbol{\gamma}^{(0)}$ is chosen as $(I-H_\lambda)^{1/2}\mathbf{y}$. The second step is to plug in the final solution $\hat{\boldsymbol{\gamma}}$ into (5) and get $\hat{\boldsymbol{\theta}}$. The outliers we identify are those observations with $\hat{\gamma}_i \neq 0$, and the nonparametric function estimate is $\hat{f}(x) = \hat{d}_1 + \hat{d}_2 x + \sum_{i=1}^n \hat{\alpha}_i K(x_i, x)$.

The problem (2) involves two tuning parameters κ and λ, which control the outlier detection and nonparametric estimation respectively. It is important to choose these two parameters appropriately. We use grid search to choose κ and λ simultaneously through a combination of generalized cross validation (GCV) criterion and modified Bayesian information criterion (BIC).

In particular, we use the following procedure:

1. For a fixed λ, we first tune κ on a set of grid points. We calculate the smoothing matrix H_λ and set $\boldsymbol{\gamma}^{(0)} = (I - H_\lambda)\mathbf{y}$.

2. For each κ chosen from the grid points, we use (7) to solve for $\hat{\boldsymbol{\gamma}}(\lambda, \kappa)$. We define

$$\mathrm{RSS}(\lambda, \kappa) \quad = \quad ||(I - H_\lambda)(\mathbf{y} - \hat{\boldsymbol{\gamma}}(\lambda, \kappa))||_2^2$$

   and choose κ which minimizes the following modified BIC.

$$\mathrm{BIC}(\lambda, \kappa) \quad = \quad m \log(\mathrm{RSS}(\lambda, \kappa)/m) + k \log(m),$$

   where k denotes the number of nonzero in $\hat{\boldsymbol{\gamma}}(\lambda, \kappa)$ and m is the effective sample size that is defined by $m = \mathrm{tr}\{(I - H_\lambda)\}$, where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. Denote $\kappa_{opt}(\lambda)$ be the optimal parameter we choose for this fixed λ.

3. We use the GCV criterion to select the smoothing parameter λ. Denote $\hat{\boldsymbol{\gamma}}(\lambda, \kappa_{opt}(\lambda))$ as the solution of $\hat{\boldsymbol{\gamma}}$ for a fixed λ and $\kappa = \kappa_{opt}(\lambda)$. We choose the λ which minimizes the GCV

$$\mathrm{GCV}(\lambda) \quad = \quad \frac{\mathrm{RSS}\,(\lambda, \kappa_{opt}(\lambda))}{(n - \mathrm{tr}(H_\lambda))^2}$$

Note that for Step 2, we restrict $k < n/2$ because we assume that the true number of outliers is less than half of the sample size. Additionally, when the number of parameters exceeds the sample size, we will get perfect fit if the degree of freedom is large enough, which would make the BIC go to negative infinity as the residual sum of squares goes to zero. This would result in the wrong selection of the tuning parameter because it tends to select the parameter that leads to the perfect fit. In particular, when we tune κ, we only consider those κ whose corresponding solution $\hat{\boldsymbol{\gamma}}(\lambda, \kappa)$ has less than $n/2$ nonzero components.

3 Asymptotic Theories

In this section, we discuss the asymptotic properties for our method. We focus on the LASSO penalty, i.e. we consider solving the following problem

$$\begin{aligned}&\mathrm{argmin}_{f,\boldsymbol{\gamma}}\{n^{-1}\textstyle\sum_{i=1}^n (y_i - f(x_i) - \gamma_i)^2 \\ &\quad + \kappa \textstyle\sum_{i=1}^n |\gamma_i| + \lambda \int_0^1 [f''(x)]^2 dx\}.\end{aligned}$$

Other penalty functions can be treated in a similar way. Let f_0 be the true nonparametric regression function. We make the following assumptions.

(A1) The residuals $\epsilon_1, \ldots, \epsilon_n$ are generated independent and identically distributed from $N(0, 1)$.

(A2) The true nonparametric function f_0 is twice differentiable.

(A3) Denote the true mean shift parameter by $\boldsymbol{\gamma}_0$, and the number of outliers by s_0. We assume that $s_0 \log n/n = o_p(n^{-2/5})$.

Condition (A1) assumes Gaussian errors with unit variance for simplicity. It is possible to extend the results for sub-Gaussian errors and unknown variance. Condition (A2) is a standard assumption in nonparametric statistics literature. Condition (A3) essentially requires the number of outliers be of order $o_p(n^{3/5})$ without the logarithmic factor.

The following theorem gives the convergence rate for our nonparametric function estimate $\hat{f}$ and the mean shift parameter $\hat{\gamma}$.

Theorem 1. *Assume that Conditions (A1) – (A3) hold. We choose* $\kappa = c_1\sqrt{\log n}/n$ *and* $\lambda = c_2 n^{-9/10}$ *with some positive constants* c_1 *and* c_2. *Then with probability of at least* $1 - 2n^{-1} + 6\exp(-n)$,

$$\|\hat{f} - f_0\|_2 \leq C_1 n^{-2/5}, \quad \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2 \leq C_2 s_0 \log n/n \tag{8}$$

for some positive constants C_1 *and* C_2.

Theorem 1 implies that the outliers can be estimated consistently. Moreover, it states that the estimator $\hat{f}$ achieves the same minimax-optimal convergence rate for the univariate nonparametric regression function when there are no outliers. In other words, $\hat{f}$ enjoys the oracle property as if the outliers are known in advance. This result provides theoretical justification for our estimation method because it essentially suggests that the outlier detection procedure will not affect the convergence rate of the regression function estimation. Similar conclusions can be made on $\hat{\gamma}$. If we let the true number of outliers s_0 to be a constant, the rate of convergence is essentially $O_p(\log n/n)$, which agrees

with the optimal rate of $\log p/n$ for a typical regression problem with p covariates and n observations, where $p = n$ in our case.

The proof proceeds by using the same techniques in Theorems 1 and 2 of Müller and van de Geer (2015), where the authors obtained the nonparametric convergence rate for the regression function in a partial linear model with a diverging number of parametric covariates. The main difference is that we consider a mean-shift parameter for each observation without replications. Therefore the conditions (2.1–2.6) in Müller and van de Geer (2015) cannot be directly verified and it requires a careful model re-parameterization, which leads to an additional factor of $n^{-1/2}$ on the original optimal choice of $\lambda = O_p(n^{-1/5})$. More details are given in Section 8.

4 Simulation

In this section, we evaluate the proposed method through simulations. We set different sample sizes $n = 50, 100, 200$. For the true nonparametric function, we set $f(x) = 10\sin(2\pi x)$, where $x \in [0, 1]$. For x_i's, they are randomly generated from uniform$(0, 1)$. For ϵ_i's, they are generated independent and identically distributed from $N(0, 1)$. The true response is generated from $y_i^* = f(x_i)+\epsilon_i$. Among all the observations, we randomly set cn of them to be outliers, where we consider different values $c = 0.1, 0.2$. In particular, we contaminate cn of the observations by adding each of the corresponding y_i^* by b, i.e. $y_i = y_i^* + b_i$, where we consider different levels $b = 5, 10$. Denote the final observations as $\{(x_i, y_i), 1 \le i \le n\}$.

For each setting, we conduct 100 Monte Carlo replications and report the mean of the following four quantities:

1. **M** the mean masking probability (fraction of undetected true outliers)
2. **S** the mean swamping probability (fraction of good points labeled as outliers)
3. **JD** the joint outlier detection rate (fraction of simulations with 0 masking)
4. **MSE** the mean square error (MSE), which is defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}(x_i) - f(x_i)\right\}^2$$

Masking can occur when we specify too few outliers. For an outlier detection procedure, it is desirable to have masking probability as small as possible. This is the most important criterion to evaluate a robust procedure. Swamping occurs when we specify too many outliers. Although it is not as critical as masking, swamping can not be too large. The joint outlier detection rate is to report the proportion of simulations with 0 masking, i.e. we can identify all the outliers correctly, and the desired outlier detection rate is 1. The MSE is used to characterize how well the nonparametric function is estimated in the presence of the outliers. The smaller the MSE is, the better the robust procedure is.

We compare the performance of the proposed method with other robust regression approaches. As there is no other existing approach that achieves outlier detection and robust regression for univariate nonparametric regression to the best of our knowledge, we adopt a two-step approach. First, we apply regression spline to approximate the nonparametric function $f(\cdot)$ such that the problem transforms to a linear regression problem. Then we can use some of the existing outlier detection and robust regression techniques under the linear regression framework. In particular, we choose a set of B-spline basis with the degree of freedom (df= 3) and consider knots that are quantiles of the range of x depending on the number of the knots (denoted by k) we set. We choose k that minimizes the GCV, which is defined as

$$\text{GCV}(k) = \frac{\text{RSS}(k)}{(n - df - k)^2},$$

where $\text{RSS}(k)$ is the residual sum of squares when the knots number is k.

After selecting the number of knots k, we use regression splines to fit the nonparametric function with the B-spline basis. We consider a set of robust regression methods in linear regression, including MM-estimators (Yohai, 1987), Gervini and Yohai's (GY) fully efficient one-step procedure (Gervini and Yohai, 2002) with least square estimates as an initial value. For GY's method, we need to estimate the variance of ϵ_i. For simplicity, we plug in the true value $\sigma^2 = 1$, which makes the performance of GY's method better than when plugging in the estimates of σ^2 since an oracle value is used. We summarize the simulation results in Table 1 based on 100 Monte Carlo replicates.

From the results, we find that the proposed method performs reasonably well under different scenarios. It achieves much smaller masking rate and much higher joint outlier detection rate than those of using MM and GY methods for majority of the cases. The MM and GY methods work reasonably well when $b = 10$ although our method still outperforms these two methods for most of the cases. However, when b decreases to 5, these two methods work poorly. This indicates

Table 1: Simulation Results (along with the associated standard errors in brackets) for the proposed method (P), compared with MM and GY methods.

(n, c, b)	Method	M	S	JD	MSE
$(50, 0.1, 5)$	P	0.012(0.006)	0.209(0.009)	0.95	0.235(0.022)
	MM	0.446(0.031)	0.003(0.001)	0.17	1.828(1.153)
	GY	0.182(0.021)	0.04(0.003)	0.42	0.622(0.049)
$(50, 0.1, 10)$	P	0.008(0.005)	0.103(0.011)	0.97	0.61(0.293)
	MM	0.016(0.008)	0.003(0.002)	0.95	1.808(1.227)
	GY	0.026(0.008)	0.121(0.005)	0.89	1.108(0.133)
$(50, 0.2, 5)$	P	0.019(0.006)	0.251(0.007)	0.87	0.38(0.048)
	MM	0.779(0.022)	0.002(0.001)	0	83.938(78.204)
	GY	0.242(0.018)	0.096(0.005)	0.12	1.195(0.088)
$(50, 0.2, 10)$	P	0.012(0.005)	0.158(0.012)	0.93	0.667(0.175)
	MM	0.047(0.014)	0.005(0.002)	0.8	271.046(266.292)
	GY	0.037(0.009)	0.302(0.007)	0.8	2.642(0.27)
$(100, 0.1, 5)$	P	0.001(0.001)	0.155(0.007)	0.99	0.1(0.006)
	MM	0.296(0.023)	0.002(0.001)	0.12	6209.61(6209.23)
	GY	0.08(0.009)	0.028(0.002)	0.46	0.223(0.02)
$(100, 0.1, 10)$	P	0.001(0.001)	0.032(0.005)	0.99	0.177(0.086)
	MM	0(0)	0.001(0)	1	0.306(0.022)
	GY	0.003(0.002)	0.103(0.003)	0.97	0.524(0.04)
$(100, 0.2, 5)$	P	0.003(0.002)	0.245(0.007)	0.97	0.133(0.018)
	MM	0.779(0.02)	0.001(0.001)	0	1.32(0.281)
	GY	0.192(0.012)	0.082(0.003)	0.05	0.732(0.057)
$(100, 0.2, 10)$	P	0(0)	0.046(0.005)	1	0.097(0.006)
	MM	0.072(0.018)	0.011(0.003)	0.73	4×10^9 (4×10^9)
	GY	0.007(0.002)	0.307(0.004)	0.86	1.664(0.122)
$(200, 0.1, 5)$	P	0.001(0.001)	0.116(0.004)	0.98	0.053(0.003)
	MM	0.247(0.016)	0.002(0.002)	0.03	1.4×10^6 (1.4×10^6)
	GY	0.055(0.006)	0.022(0.001)	0.35	0.086(0.01)
$(200, 0.1, 10)$	P	0(0)	0.02(0.001)	1	0.041(0.002)
	MM	0(0)	0(0)	1	0.148(0.021)
	GY	0(0)	0.085(0.003)	1	0.29(0.028)
$(200, 0.2, 5)$	P	0.021(0.014)	0.206(0.006)	0.95	0.087(0.017)
	MM	0.776(0.016)	0.003(0.001)	0	2×10^8 (2×10^8)
	GY	0.139(0.006)	0.073(0.002)	0.01	0.348(0.027)
$(200, 0.1, 10)$	P	0(0)	0.028(0.002)	1	0.046(0.003)
	MM	0.026(0.007)	0.008(0.002)	0.8	2×10^{17} (2×10^{17})
	GY	0.001(0.001)	0.312(0.003)	0.98	1.039(0.055)

that our method is more sensitive in detecting outliers than the comparing methods. As sample size increases, both P and GY have better performances (e.g., smaller MSE, higher S). In terms of MSE, our method beats the other two significantly. For MM, it happens sometimes that the estimates break down and the MSE becomes unrealistically large (e.g., 10^9) due to outliers for a few Monte Carlo Studies. For GY method, although it does not break down, its MSE still doubles the MSE of our method in most cases.

5 Real Data Application

In this section, we consider the baseball data obtained from He, Ng, and Portnoy (1998), which can be downloaded from *http://www.blackwellpublishers.co.uk/rss*. The data consist of 263 North American Major League players during the 1986 season. We are interested in studying how the performance of baseball players affect their annual salaries (in thousands of dollars). We use the number of home runs in the latest year to measure the performance. We treat the annual salary as the response y_i and the number of home runs x_i as the covariate. Figure 5 shows a scatterplot of (x_i, y_i). We see that the signal-to-noise ratio is quite low. We have applied the proposed method to this data set, and found 15 outliers. We have colored the outliers in red and plotted the estimated regression curve in the same figure.

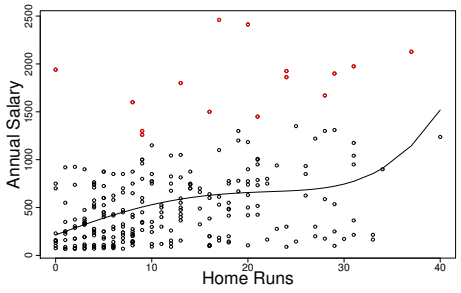

Figure 1: Data Scatter Plot: identified outliers colored in red and estimated function in solid curve.

Interestingly, the mean shift parameter estimates of these outliers are all positive, which indicates that there exists some other factors which cause these players to have a higher salary than their peers. The dataset also contains an additional variable seniority, which is the number of years the player has played in the league. The average number of years played among these outliers is 11.2, and is 7.12 years among these 'normal' players. We have performed a t-test to check whether the outliers group have a longer playing history than that of the normal group. The resulting p-value is 0.0003, which indicates that the average years played among the outliers are significant longer than the average years among normal players. Thus, it confirms our findings of the outliers by associating their high salaries with their longer career histories after adjusting for the performance effect.

6 Extension

Although we restrict our discussion to univariate regressions, the proposed method can be easily extended to nonparametric multidimensional regression and semiparametric multidimensional regression models. For example, we may consider the following partially linear model

$$y_i = f(\mathbf{x}_i) + \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{9}$$

where $f(\cdot)$ is an unknown nonparametric function, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are p-dimensional covariates, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iq})^T$ are q-dimensional covariates and ϵ_i is the random error. This model is often used in genetics to jointly model the genetic pathway effect nonparametrically and the clinical effect parametrically, see Liu et al. (2007); Kong et al. (2016) for example. A

special case of our model is the nonparametric multidimensional regression problem, where we do not have the term $\mathbf{z}_i^T\boldsymbol{\beta}$.

The mean shift model can also be used under this scenario to detect the outliers in the data. Similarly, we consider

$$y_i = f(\mathbf{x}_i) + \mathbf{z}_i^T\boldsymbol{\beta} + \gamma_i + \epsilon_i, \quad (10)$$

and we are interested in identifying the nonzero γ_i estimates.

In genetics, it is often the case that the number of genes p is larger than the sample size, which makes it challenging to do statistical inference using traditional nonparametric regression techniques. To solve this problem, Liu et al. (2007) used the kernel machine smoothing method for the nonparametric function $f(\cdot)$. Specifically, they assume the function $f(\cdot)$ resides in a functional space H_K generated by a positive definite kernel function $K(\cdot,\cdot)$. Under the kernel machine framework, estimation would be more accurate for multi-dimensional data. According to Mercer's Theorem (Cristianini and Shawe-Taylor, 2000), there is a one-to-one correspondence between a positive definite kernel function and a functional space H_K under some regularity conditions. We call H_K the Reproducing Kernel Hilbert Space generated by the kernel K. We can expand the function $f(\cdot)$ on the basis functions in H_K, where the basis functions can be represented using the kernel function. By representer theorem (Kimeldorf and Wahba, 1971), the solution of the nonparametric function $f(\cdot)$ can be written as

$$f(\mathbf{x}) = \sum_{i=1}^{n}\theta_i K(\mathbf{x},\mathbf{x}_i), \quad (11)$$

where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)^T$ are unknown parameters. There are several popular kernels such as the dth degree polynomial kernel $K(\mathbf{x}_1,\mathbf{x}_2) = (\mathbf{x}_1^T\mathbf{x}_2+1)^d$, the gaussian kernel $K(\mathbf{x}_1,\mathbf{x}_2) = \exp(-||\mathbf{x}_1-\mathbf{x}_2||^2/\rho)$ and the identity-by-state kernel, where d and ρ are the tuning parameters. Under the kernel machine framework, one often starts with certain kernel function, which implicitly determines the functional space H_K.

We would borrow the similar idea in Section 2 and consider solving the following optimization problem

$$\begin{aligned}&\text{argmin}_{f,\boldsymbol{\gamma}}\textstyle\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i) - \mathbf{z}_i^T\boldsymbol{\beta} - \gamma_i)^2\\ &\quad + \textstyle\sum_{i=1}^{n}P(\gamma_i,\kappa) + \lambda||f||^2_{H_K}.\end{aligned}$$

Let $\mathbf{y} = (y_1,\ldots,y_n)^T$, $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)^T$, $\mathbf{z} = \mathbf{z}_1,\ldots,\mathbf{z}_n{}^T$, $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)^T$ and $\mathbf{K}$ be a $n\times n$ matrix with ijth element $K(\mathbf{x}_i,\mathbf{x}_j)$. Plugging (11) into the objective function, we obtain

$$\text{argmin}_{\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}}\{||\mathbf{y}-\mathbf{z}\boldsymbol{\beta}-\mathbf{K}\boldsymbol{\theta}-\boldsymbol{\gamma}||_2^2+\sum_{i=1}^{n}P(\gamma_i,\kappa)+\lambda\boldsymbol{\theta}^T\mathbf{K}\boldsymbol{\theta}\}.$$

We would use the similar profiling and TISP method to obtain the estimates of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. However, it is unclear how to tune parameters under this scenario for a few reasons. First, the dimension of $\mathbf{x}_i$ can be much higher than the sample size, and the tuning of parameters may be affected by the curse of dimensionality. Second, for some kernels such as polynomial and gaussian kernels, we need to also tune the parameters involved in the kernel function (e.g., d and ρ), which makes the computation more complicated. Another interesting question is how to choose the kernel function that is adaptive to the data. We leave these topics for future research.

7 Discussion

In this paper, we consider the outlier detection and robust estimation for the univariate nonparametric regression problem. A mean shift parameter is introduced for outlier detection purpose. We consider a penalized approach and propose an efficient algorithm to solve the "two-component" optimization problem. We also discuss the simultaneous selection of different tuning parameters. We obtain a minimax-optimal convergence rate for nonparametric regression function estimation and show that it enjoys the oracle property in the sense that it agrees with the optimal nonparametric rate when the outliers are known in advance. Our method can be extended to nonparametric and semiparametric multidimensional regression models. It remains open to develop efficient algorithms as well as appropriate tuning procedures for these cases.

8 Proof sketch

The main idea of the proof is to re-parameterize the model (1) into a partial linear model such that the results in Müller and van de Geer (2015) are applicable. This can be done by considering

$$n^{-1/2}y = n^{-1/2}f(x) + \boldsymbol{I}\boldsymbol{\gamma} + n^{-1/2}\epsilon, \quad (12)$$

where $\boldsymbol{I}$ is the n-dimensional identity matrix and $\boldsymbol{\gamma} = n^{-1/2}(\gamma_1,\ldots,\gamma_n)^T$ is the collection of re-scaled mean-shift parameters. Then the noise level is reduced to as if there are n replications. To verify Conditions (2.1)–(2.6) in Müller and van de Geer (2015), first we note that the design matrix is the identity matrix, hence Condition (2.1)–(2.3) are easily satisfied. Condition (2.4) holds because we assume f_0 is twice differentiable. Condition (2.5) holds since we use the smoothness penalty function (second-order derivative). Condition (2.6) also holds because of the independence assumption. Then we can obtain the convergence rates as a consequence of the assumption (A3) we make on

s_0 and the results in Theorem 2 of Müller and van de Geer (2015) with the choice of $p = n$ there. The additional $n^{-1/2}$ can be absorbed by the tuning parameters κ and λ.

Achknowledgements

We thank three reviewers and the area chair for their helpful comments and suggestions. Kong's research was partially supported by the Natural Science and Engineering Research Council of Canada. Shen's research was partially supported by the Simons Foundation Award 512620.

References

Agarwal, G. G. and Studden, W. J. (1980), "Asymptotic integrated mean square error using least squares and bias minimizing splines," *The Annals of Statistics*, 8, 1307–1325.

Bondell, H. and Stefanski, L. (2013), "Efficient robust regression via two-stage generalized empirical likelihood," *Journal of the American Statistical Association*, 108, 644–655.

Brown, L. D., Cai, T. T., and Zhou, H. H. (2008), "Robust nonparametric estimation via wavelet median regression," *The Annals of Statistics*, 36, 2055–2084.

Cai, T. T. and Zhou, H. H. (2009), "Asymptotic equivalence and adaptive estimation for robust nonparametric regression," *The Annals of Statistics*, 37, 3204–3235.

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

Cristianini, N. and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 1st ed.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *The Annals of Statistics*, 32, 407–499.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66 (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*, Chapman & Hall, 1st ed.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Gannaz, I. (2006), "Robust Estimation and Wavelet Thresholding in Partial Linear Models," Tech. Rep. math.ST/0612066.

Gervini, D. and Yohai, V. J. (2002), "A class of robust and fully efficient regression estimators," *The Annals of Statistics*, 30, 583–616.

Gu, C. (2013), *Smoothing spline ANOVA models*, vol. 297 of *Springer Series in Statistics*, Springer, New York, 2nd ed.

Hampel, F. R. (1975), "Beyond location parameters: robust concepts and methods," in *Proceedings of the 40th Session of the International Statistical Institute (Warsaw, 1975), Vol. 1. Invited papers*, vol. 46, pp. 375–382, 383–391 (1976), with discussion.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized additive models*, vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, Ltd., London.

He, X., Ng, P., and Portnoy, S. (1998), "Bivariate quantile smoothing splines," *Journal of the Royal Statistical Society. Series B.*, 60, 537–550.

Huber, P. J. (1981), *Robust statistics*, New York: John Wiley & Sons Inc., wiley Series in Probability and Mathematical Statistics.

Kim, Y., Choi, H., and Oh, H.-S. (2008), "Smoothly clipped absolute deviation on high dimensions," *Journal of the American Statistical Association*, 103, 1665–1673.

Kimeldorf, G. and Wahba, G. (1971), "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, 33, 82–95.

Kong, D., Bondell, H., and Wu, Y. (2018), "Fully efficient robust estimation, outlier detection, and variable selection via penalized regression." *Statistica Sinica*, to appear.

Kong, D., Maity, A., Hsu, F., and Tzeng, J. (2016), "Testing and estimation in marker-set association study using semiparametric quantile regression kernel machine." *Biometrics*, 72, 364–371.

Liu, D., Lin, X., and Ghosh, D. (2007), "Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models," *Biometrics*, 63, 1079–1088, 1311.

Mallows, C. (1975), "On Some Topics in Robustness," *unpublished memorandum, Bell Tel. Laboratories, Murray Hill.*

Mateos, G. and Giannakis, G. B. (2012), "Robust nonparametric regression via sparsity control with application to load curve data cleansing," *IEEE Transactions on Signal Processing*, 60, 1571–1584.

McCann, L. and Welsch, R. E. (2007), "Robust variable selection using least angle regression and elemental set sampling," *Computational Statistics & Data Analysis*, 52, 249–257.

Müller, P. and van de Geer, S. (2015), "The partial linear model in high dimensions," *Scand. J. Stat.*, 42, 580–608.

Rousseeuw, P. and Yohai, V. (1984), "Robust regression by means of S-estimators," in *Robust and nonlinear time series analysis (Heidelberg, 1983)*, New York: Springer, vol. 26 of *Lecture Notes in Statist.*, pp. 256–272.

Rousseeuw, P. J. (1984), "Least median of squares regression," *Journal of the American Statistical Association*, 79, 871–880.

She, Y. (2009), "Thresholding-based iterative selection procedures for model selection and shrinkage," *Electronic Journal of Statistics*, 3, 384–415.

She, Y. and Owen, A. B. (2011), "Outlier detection using nonconvex penalized regression," *Journal of the American Statistical Association*, 106, 626–639.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B.*, 58, 267–288.

Wahba, G. (1990), *Spline models for observational data*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Yohai, V. J. (1987), "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics*, 15, 642–656.