International Journal of Statistics and Probability; Vol. 4, No. 3; 2015
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Regularized Single Index Quantile Regression Model

Chinthaka Kuruwita[1]

[1] Department of Mathematics, Hamilton College, Clinton, NY 13323, USA

Correspondence: Chinthaka Kuruwita, Department of Mathematics, Hamilton College, Clinton, NY 13323, USA. Tel: 1-315-859-4365. E-mail: ckuruwit@hamilton.edu

Received: May 17, 2015 Accepted: June 18, 2015 Online Published: July 1, 2015

doi:10.5539/ijsp.v4n3p74 URL: http://dx.doi.org/10.5539/ijsp.v4n3p74

Abstract

This article proposes a new approach for variable selection in the single index quantile regression model. Compared to existing methods, the new approach produce sparse solutions for the index vector. Performance of the new method is enhanced by a fully adaptive penalty function. Finite sample performance is studied through a simulation study that compares the proposed method with existing work under several criteria. A data analysis is given which highlights the usefulness of the proposed methodology.

Keywords: robust regression, variable selection, lasso

1. Introduction

With the advances in data generation, collection, and storage, modeling data with large number of covariates has become the rule rather than the exception. However, when the number of covariates is large, classical modeling approaches often suffer the curse of dimensionality. In the context of nonparametric regression, handling this issue becomes an extremely challenging task, because data sparseness in local neighborhoods makes it virtually impossible to perform a fully nonparametric estimation of a regression function with multiple covariates. To overcome this issue Ichimura (1990) proposed the single-index model. It is a semiparametric model that combines the strengths of parametric and nonparamntric paradigms to alleviate the issues surrounding high dimensionality. However, single-index models suffers from the same set of drawbacks of classical regression models: 1. the typical single-index model assumes that the errors have finite variance. In practice this assumption may not hold, especially with heavy tailed error distributions; 2. the single-index model attempts to model the conditional mean of the response variable. As a result the estimates becomes highly sensitive to extreme values. To overcome these difficulties and allow the estimation of various conditional quantiles of the response variable, Koenker & Bassett (1978) in their pioneering work introduced the quantile regression framework. In this article we examine the estimation and variable selection of the single-index quantile regression model.

Consider the single-index quantile regression model

$$y_i = g_\tau(\theta_\tau^\top \mathbf{x}_i) + \epsilon_i, i = 1, \ldots, n \tag{1}$$

where $\{y_i\}$'s are univariate response variables, $\{\mathbf{x}_i\}$'s are $p \times 1$ covariates and g_τ is a univariate "link" function. The $\{\epsilon_i\}$'s are independent random error terms with the τ^{th} conditional quantile to be zero, i.e. $P(\epsilon_i \leq 0|\mathbf{x}_i) = \tau$.

Estimation of model (1) has been studied by Wu et al. (2010). They proposed an iterative method to estimate the index vector and the link function. Their algorithm uses local linear estimation and regular quantile regression estimation iterativelly until convergence. The drawback of their procedure is that it does not perform variable selection. Therefore, when applied to data with large number of covariates, it performs poorly in identifying relevant predictors associated with the given response. To remedy this issue, Alkenani and Yu (2013) introduced regularization into the method proposed by Wu et al. (2010). They propose to use Least Absolute Shrinkage and Selection Operator (LASSO) and Adaptive LASSO as a regularizing mechanism to select variables in model (1). Although, adaptive LASSO has been shown to perform quite well under various settings (Zou (2006), Wu & Liu (2009), Zheng et al. (2013)) their estimation procedure is a computationally expensive one that involves minimizing a double weighted summation of n^2 items, where n is the sample size.

To overcome this difficulty Lv et al. (2014) proposed an improved version of the estimation algorithm by extending

Carroll et al. (1997)'s approach. In effect, they cut down the computational burden from n^2 to n for the estimation of the index vector and link function of model (1). In addition, to select important variables, they propose to use LASSO type regularization as a follow-up step to their estimation. The use of regularization as a secondary step has made their approach to perform poorly in estimating the link function when the number of covariates, p, is considerably large (see Figures 1 and 2). Furthermore, the resulting estimate of the index vector is not sparse and there is a large bias and variance in the estimated coeffcients and the link function. In addition, they rely on methods like average derivative estimation (ADE) of Chaudhuri et al. (1997) to obtain initial values for their algorithm. Unfortunately, ADE method require multidimensional kernel smoothing which is computationally expensive and sometimes infeasible with high dimensional covariates. Therefore, it is counterproductive to use ADE in a single-index modeling framework. In order to overcome these drawbacks, we propose the following two improvements to Lv's algorithm:

1. To use the standard unpenalized linear quantile regression estimator as the initial value for the algorithm thereby reduce the overall computational burden in estimation.
2. To penalize the estimated coefficients iteratively so that the final estimate of the index vector is sparse.

This proposal is motivated by the following observations: 1. Zhu et al. (2012) showed that the linear quantile regression estimator is a $\sqrt{n}$-consistent estimator of the index parameter even under link violations. Therefore, we can exploit this property in the single index framework as a "good" starting point for the estimation; 2. Although the algorithm proposed by Lv et al. (2014) performs well with small number of non-significant covariates, when a large number of extra (non-significant) covariates are introduced into the estimation, which is more likely to happen in practice with large datasets, we see a dramatic drop in performance in Lv's method. We demonstrate this effect in the following example:

1.1 Motivational Example

Following the standard notation in the literature (Wang et al. (2007)), decompose the index vector as $\theta_\tau = (\theta_a, \theta_b)$, where $\theta_a = (\theta_1, \ldots, \theta_{p_0})$ and $\theta_b = (\theta_{p_0+1}, \ldots, \theta_p)$. Here, p_0 is the number of non-zero covariates and p is the total number of covariates with $p_0 \ll p$. Note that θ_a is actually $\theta_{\tau a}$, but we suppress the dependence on τ for notational convenience.

Consider the following two cases in the context of model (1):

1. Set $p = 10$, $\theta_\tau = (1, 1, 1, 0, \ldots, 0)/\sqrt{3}$. That is, $p_0 = 3$ and the rest are zeros. Set $\epsilon \sim N(0, \sigma^2)$, with $\sigma = 0.1$ and $\mathbf{X} = (x_1, \ldots, x_p)$ *i.i.d.* $\sim U[0, 1]$.
2. Same as above, except $p = 50$, with $p_0 = 3$ and the rest of the 47 coefficients are zeros.

We simulated 1000 random samples of size n = 200 for each case and with three link functions:

$$g(t) = 2t + 3; \qquad g(t) = t^2; \qquad g(t) = sin(4t).$$

Remark 1

It is well known that $g(\cdot)$ and θ are unidentifiable since $g(\theta^\top \mathbf{x}) = g_c[(c\theta)^\top \mathbf{x}]$ for $c \neq 0$, where $g_c(\cdot) = g(\cdot/c)$. In order to achieve identifiability in the model the index vector θ_τ is constrained to be $\|\theta_\tau\| = 1$ with $\theta_{\tau,1} > 0$. (Ichimura, 1990; Lin & Kulasekera, 2007).

Figure 1 shows the estimated coefficients and the link function estimates of the 1000 simulated samples from case 1. As evident in Figure 1, both methods perform really well in recovering the link functions. However, when we introduce a large number of non-significant covariates into the estimation problem (case 2), the method proposed by Lv et al. (2014) was unsuccessful in recovering the link function. This is shown in Figure 2.

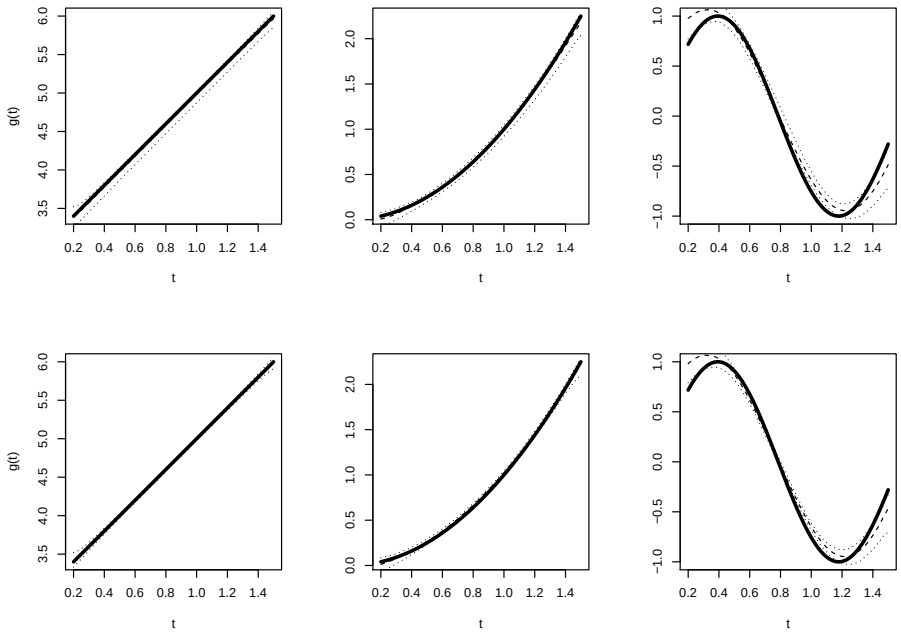

Figure 1. Comparison of estimated link functions at $\tau = 0.5$ with p=10: solid=true; dashed=50^{th}, dotted=5^{th} and 95^{th} percentiles of the quantile function estimates of 1000 simulated samples. Top row: Lv et al. (2014). Bottom row: Proposed method.

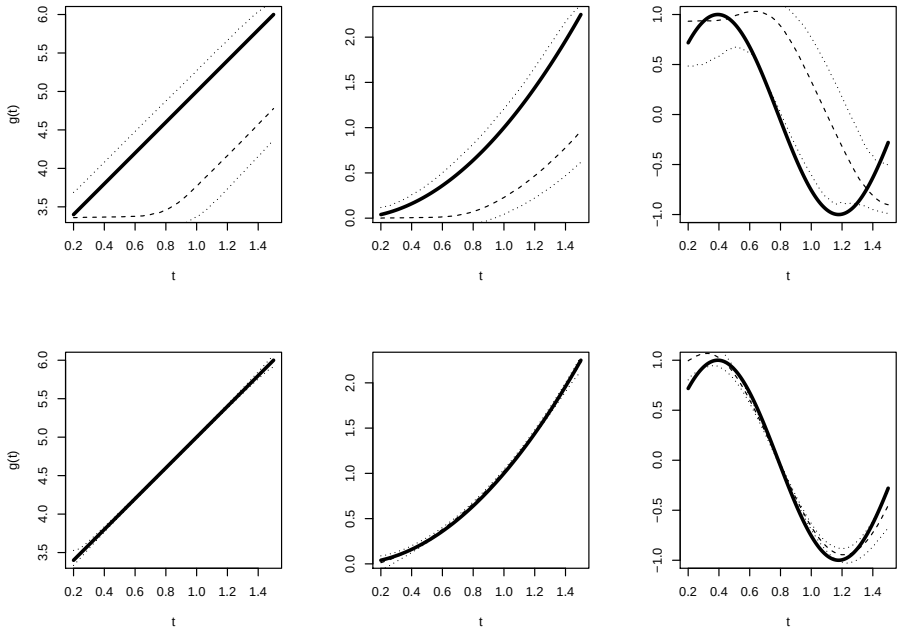

Figure 2. Comparison of estimated link functions at $\tau = 0.5$ with p=50: solid=true; dashed=50^{th}, dotted=5^{th} and 95^{th} percentiles of the quantile function estimates of 1000 simulated samples. Top row: Lv et al. (2014). Bottom row: proposed method.

This drop in effectiveness of the method proposed by Lv et al. (2014) is a direct consequence of how they chose to combine the estimation and variable selection methods. In essence, their approach follows a two-step structure.

First, they estimate the index vector using an unpenalized minimization algorithm. Then, they use the resulting index vector and the link function in a penalized estimation procedure to arrive at their final estimate of the index vector. Although, this method work well with relatively few number of covariates being non-significant (as in Figure 1), it fails to perform adequately when a large number of (non-significant) covariates are introduced (see Figure 2). In addition to unsatisfactory performance in link function estimation it fails to yield a sparse index vector estimate when there are many non-significant covariates are present in the data.

For example, Table 1 shows a brief summary of the estimated coefficients for cases 1 and 2 with the linear link $g(t) = 2t + 3$. When p=10 with only 7 non-significant covariates, both methods perform quite well. However, when the number of non-significant covariates increase to 47, Lv et al.'s method did not recover most of the zero coefficients. Furthermore, their mean squared errors of the three estimated significant variables is much larger (about five times) compared to our method.

Table 1. Summary of results: Mean Squared Error (MSE) of the non-zero coefficients and the identification of zero coefficients in Case 1 and 2 with linear link at $\tau = 0.5$.

	Lv et al. (2014)					Proposed Method				
	MSE (in 10^{-3})			no. of zeros		MSE (in 10^{-3})			no. of zeros	
p	θ_1	θ_2	θ_3	Correct	Wrong	θ_1	θ_2	θ_3	Correct	Wrong
10	0.295	0.278	0.245	5.87	0	0.159	0.168	0.156	6.8	0
50	0.875	0.860	0.872	5.28	0	0.163	0.170	0.171	46.3	0

As evident in this motivational example (Figures 1, 2 and Table 1), the overall performance of the approach proposed by Lv et al. (2014) is not satisfactory . To remedy this issue, we now present the details of our proposal and discuss how we improve the estimation of the index vector and link function in model (1).

2. Estimation Method

Step 1: Obtain an initial $\hat{\theta}^{(0)}$ from the classical unpenalized linear quantile regression method and standardize it so that $\|\hat{\theta}^{(0)}\| = 1$ with $\hat{\theta}^{(0)}_{\tau,1} > 0$.

Step 2: Given $\hat{\theta}^{(0)}$, use local linear estimation to obtain initial estimates of the link function g (and its derivative) at $\hat{\theta}^{(0)^\top}\mathbf{x}_j,\ j = 1, \ldots, n$, by solving the minimization problem

$$\min_{a_j, b_j} \sum_{i=1}^{n} \rho_\tau \left\{ y_i - a_j - b_j[\hat{\theta}^{(0)^\top}(\mathbf{x}_i - \mathbf{x}_j)] \right\} K \left\{ \frac{\hat{\theta}^{(0)^\top}(\mathbf{x}_i - \mathbf{x}_j)}{h} \right\}$$

Here $\rho_\tau(s) = |s| + (2\tau - 1)s$ is the loss function (commonly known as the "check" function), $K(\cdot)$ is a univariate kernel function and h is the bandwidth that governs the smoothness in the local linear estimation.

Step 3: Using the estimate $\hat{a}_j = \hat{g}(\hat{\theta}^{(0)^\top}\mathbf{x}_j)$ and $\hat{b}_j = \hat{g}'(\hat{\theta}^{(0)^\top}\mathbf{x}_j)$ from Step 2, obtain the final estimator of the index parameter θ by solving following penalized minimization problem:

$$\min_{\theta} \sum_{j=1}^{n} \rho_\tau \left\{ y_j - \hat{g}(\hat{\theta}^{(0)^\top}\mathbf{x}_j) - \hat{g}'(\hat{\theta}^{(0)^\top}\mathbf{x}_j)[\theta - \hat{\theta}^0]^\top \mathbf{x}_j \right\} + \lambda_n \sum_{k=1}^{p} \omega_k |\theta_k|,$$

where λ_n is the penalty parameter and $\omega \in \mathbb{R}^p$ is a data-driven weight vector that controls the penalty for each index coefficient.

Note: This step is where we differ from Lv et al. (2014). We introduce the penalty function as a main step in the estimation rather than a post estimation step. Furthermore, the penalty function that we propose is fully adaptive unlike the one proposed by Lv et al. (2014).

Step 4: Repeat Steps 2 and 3 until convergence. Denote the final estimate of θ as $\hat{\theta}$.

Once the index parameter θ is estimated, an improved estimate of the link function is obtained by smoothing $\{\hat{\eta}_i, y_i\}_{i=1}^n$ with local linear estimation. Here $\hat{\eta}_i = \hat{\theta}^\top \mathbf{x_i}$ and the minimization is identical to the one in *Step 2*. Denote the final estimate of g_τ as $\hat{g}$.

2.1 Tuning Parameter Selection

It is well known that any estimation procedure that depends on non parametric smoothing requires careful attention to bandwidth selection. Most bandwidth selection procedures relies on minimizing the mean squared error in estimation. For local linear quantile regression function estimation, Yu & Jones (1998) have developed a useful rule-of-thumb bandwidth by minimizing the asymptotic mean squared error (AMSE). This bandwidth, h_τ, is calculated as follows:

$$h_\tau = h_{mean} \left\{ \frac{\tau(1-\tau)}{\phi(\Phi^{-1}(\tau))^2} \right\}^{1/5} \tag{2}$$

where h_{mean} is the AMSE optimal bandwidth for local linear nonparametric mean regression. The functions $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and the distribution function of the standard normal distribution respectively. This is a commonly used (and computationally efficient) way to calculate AMSE optimal bandwidths for nonparametric quantile regression. We use this optimal bandwidth selector in step 2 of our algorithm.

The penalty parameters λ_n and ω can be chosen using leave-one-out cross-validation. We propose a two dimensional grid search with cross validation in terms of check loss optimality to select the optimal parameters. Let $\hat{\omega} = 1/|\tilde{\theta}|^\delta$, with $\delta > 0$, and $\tilde{\theta}$ is any $\sqrt{n}$-consistent estimator of the index vector. Define the leave-one-out cross validation score for the τ^{th} conditional quantile at (λ, δ) as

$$CV(\lambda, \delta)_\tau = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - \hat{Y}^{(-i)})$$

where $\hat{Y}^{(-i)}$ is the fitted value of the τ^{th} conditional quantile with the i^{th} observation removed. Although this approach yield check loss optimal tuning parameters, their statistical properties are not well understood especially under heavy tail error distributions where quantile regression is more applicable. Furthermore, this is a computationally expensive method due to the two dimensional nature of the minimization. Therefore, we illustrate this method only in our data analysis in section 4. To save computational time in our simulations, we used a coarse grid search to find the check loss optimal tunning parameters. In particular, we used a two-dimensional grid $\{(\lambda, \delta) : \lambda = 2, 4, \ldots, 10;\ \delta = 0.5, 1, 2, 3\}$.

3. Simulation Results

We conducted a simulation study to compare the performance of our method with that of Lv et al. (2104). It covers several aspects of estimation: mean squared error in estimation of the non-zero indices; recovery of the link function; sparseness in estimation; robustness to outliers. We present 3 examples. Examples 1 focuses on estimation of a highly non-linear link function and variable selection. Example 2 highlights the robustness to outliers. Example 3 demonstrates the behavior under a hetroscedastic error structure.

To illustrate how the presence of large number of non-significant variables affect the estimation, we use the following two cases:

1. Set $p = 10$, $\theta_\tau = (1, 1, 1, 0, \ldots, 0)/\sqrt{3}$. That is, $p_0 = 3$ coefficients are non-zero and the rest are zeros.
2. Same as above, except $p = 50$, with $p_0 = 3$ coefficients are non-zero and the rest of the 47 coefficients are zeros.

For each example we used 1000 random samples each with size $n = 200$ and compared the performance of the our method with Lv et al. (2014). To assess the performance, we choose to compare: 1. Mean squared error of the estimated non-zero components; 2. the average number of correctly classified zeros (denoted by 'Correct'); 3. the average number of incorrectly classified zeros (denoted by 'Wrong'). We used a threshold of 10^{-4} to label any coefficient estimate as zero. To assess the effectiveness in estimation of the link function, we plotted the link function estimates in summary form, i.e. the 5^{th}, 50^{th} and the 95^{th} percentiles of the estimated link functions in the 1000 simulated samples. All simulations and analyses were carried out using R statistical software (R Core Team (2015)).

3.1 Example 1

In this first example, we illustrate the behavior of the estimators under a highly non-linear link function and a homoscedastic error distribution. In particular, we generated data from

$$y = sin(4t) + 0.1\epsilon$$

where $t = \theta^T \mathbf{x}$ with θ is chosen according to the two scenarios described in cases 1 and 2. Also, x_i i.i.d Unif[0,1]. The error term ϵ follows a normal distribution with mean 0 and standard deviation 1. This is the same type of link function used by Lv et al. (2014) as their first example. The following figure shows the link function estimates for Lv et al. (2014) and our method respectively. As shown in Figure 3, the link function estimates are quite good (Figure 3 top row) for Lv's approach but the performance goes down considerably when large number of non-significant covariates are introduced (Figure 3 bottom row). In contrast, our approach performs quite well under both settings ($p = 10$ and $p = 50$) as evident in Figure 4.

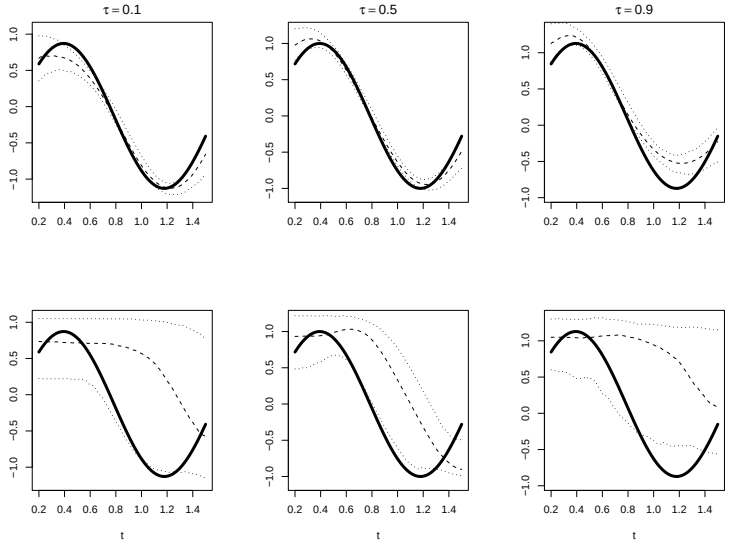

Figure 3. Comparison of estimated link functions using the method in Lv et al. (2014) with n=200: solid=true; dashed=50^{th}, dotted=5^{th} and 95^{th} percentiles of the quantile function estimates of 1000 simulated samples. Top row: $p = 10$ with 3 non-zero coefficients and 7 zero coefficients. Bottom row: $p = 50$ with 3 non-zero coefficients and 47 zero coefficients.

Next, we present the performance measures on the variable selection aspect of the two methods. The results are given for five quantiles, $\tau = 0.1, 0.25, 0.5, 0.75$, and 0.9. Table 2 summarizes the following performance metrics: the mean squared error of the three estimated non-zero indices; the average number of correctly estimated zeros; the average number of incorrectly estimated zeros.

In Case 1 (p=10), both methods were successful in selecting the three significant variables and shrinking the other 7 coefficients to zero. However, in Case 2 ($p = 50$), Lv's method was not effective in identifying the 47 non-significant variables and setting them to zero. Also, the mean squared error in Case 2 is much higher in Lv's approach compared to our method. This is a direct consequence of the lack of sparsity in their index vector estimates.

3.2 Example 2

In this example we examine how the two approaches fair when there is some contamination in the data due to outliers. We generated data from

$$y = t^2 + 0.1\epsilon\ .$$

As before, $t = \theta^T \mathbf{x}$ with θ is chosen according to the two cases described above with x_i i.i.d Unif[0,1]. The error term ϵ follows a mixture distribution of $0.9N(0, 1) + 0.1Cauchy(0, 1)$. This error distribution ensures that about 10% of the data are contaminated with very large outliers. The following figures show the link function estimates

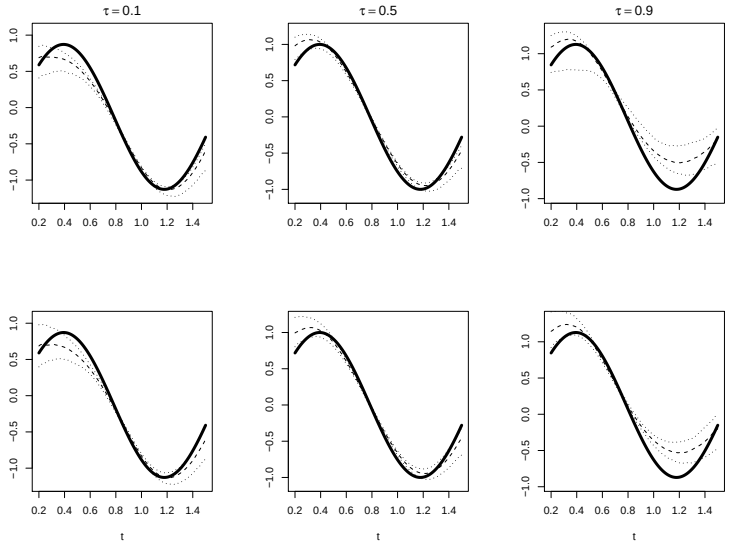

Figure 4. Comparison of estimated link functions using the proposed method with n=200: solid=true; dashed=50^{th}, dotted=5^{th} and 95^{th} percentiles of the quantile function estimates of 1000 simulated samples. Top row: $p = 10$ with 3 non-zero coefficients and 7 zero coefficients. Bottom row: $p = 50$ with 3 non-zero coefficients and 47 zero coefficients.

Table 2. Summary of results: Mean Squared Error (MSE) of the non-zero coefficients and the identification of zero coefficients in Case 1 and 2 with sin link and normal errors.

		Lv et al. (2014)					Proposed Method				
		MSE (in 10^{-3})			no. of zeros		MSE (in 10^{-3})			no. of zeros	
	τ	θ_1	θ_2	θ_3	Correct	Wrong	θ_1	θ_2	θ_3	Correct	Wrong
	0.10	0.294	0.272	0.261	5.04	0	0.267	0.288	0.266	6.97	0
	0.25	0.149	0.162	0.171	5.74	0	0.156	0.159	0.152	6.51	0
$p = 10$	0.50	0.142	0.152	0.188	5.66	0	0.127	0.122	0.121	5.70	0
	0.75	0.254	0.250	0.267	5.67	0	0.167	0.163	0.166	5.45	0
	0.90	0.340	0.275	0.323	5.78	0	0.216	0.263	0.238	6.32	0.46
	0.10	3.651	3.864	4.394	5.42	0	0.250	0.259	0.269	45.8	0
	0.25	0.349	0.33	0.359	5.01	0	0.153	0.156	0.167	46.1	0
$p = 50$	0.50	0.214	0.213	0.209	5.15	0	0.134	0.127	0.125	46.8	0
	0.75	0.270	0.277	0.263	5.26	0	0.162	0.15	0.144	46.8	0
	0.90	11.91	23.10	16.87	5.11	0.001	5.360	5.712	5.688	46.1	0.04

for Lv et al. (2014) and our method. As shown in Figure 5, the link function estimates are quite good (Figure 5 top row) for Lv's approach even when there are outliers in the data. However, as in example 1, their performance goes down considerably when large number of non-significant covariates are introduced (Figure 5 bottom row). In contrast, as evident in Figure 6, our approach performs quite well under both settings ($p = 10$ and $p = 50$).

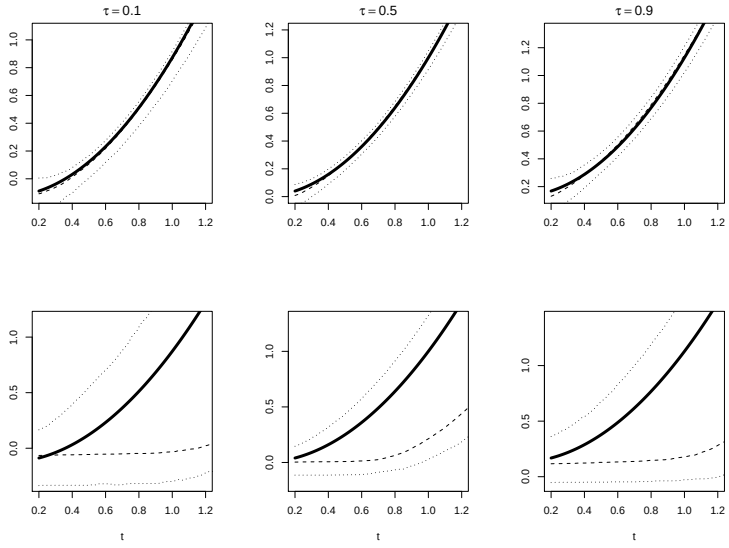

Figure 5. Comparison of estimated link functions using the method in Lv et al. (2014) with n=200: solid=true; dashed=50^{th}, dotted=5^{th} and 95^{th} percentiles of the quantile function estimates of 1000 simulated samples. Top row: $p = 10$ with 3 non-zero coefficients and 7 zero coefficients. Bottom row: $p = 50$ with 3 non-zero coefficients and 47 zero coefficients.

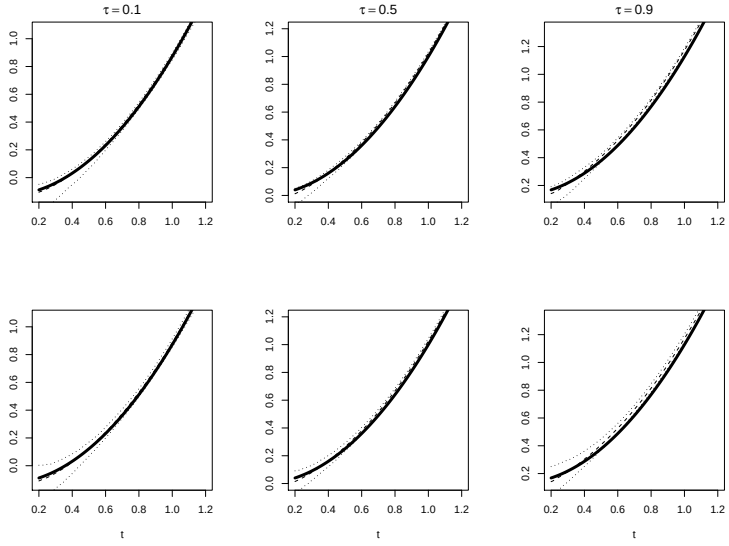

Figure 6. Comparison of estimated link functions using the proposed method with n=200: solid=true; dashed=50^{th}, dotted=5^{th} and 95^{th} percentiles of the quantile function estimates of 1000 simulated samples. Top row: $p = 10$ with 3 non-zero coefficients and 7 zero coefficients. Bottom row: $p = 50$ with 3 non-zero coefficients and 47 zero coefficients.

As before, the performance of the variable selection aspect of the estimation is studied at five quantiles, $\tau = 0.1, 0.25, 0.5, 0.75$, and 0.9. Table 2 summarizes, for cases 1 and 2, the mean squared error of the three estimated non-zero components in the index vector and the average number of correctly and incorrectly estimated zeros for both methods. We see a distinct pattern as we observed in Example 1: both methods are successful in selecting the correct number of covariates when there are few non-significant variables are present ($p = 10$), but the proposed method outperformed Lv's approach when there are many non-significant variables are involved ($p = 50$).

Table 3. Summary of results: Mean Squared Error (MSE) of the non-zero coefficients and the identification of zero coefficients in Case 1 and 2 with quadratic link and a mixture error distribution.

		Lv et al. (2014)					Proposed Method				
		MSE (in 10^{-3})			no. of zeros		MSE (in 10^{-3})			no. of zeros	
	τ	θ_1	θ_2	θ_3	Correct	Wrong	θ_1	θ_2	θ_3	Correct	Wrong
	0.10	0.761	0.822	0.851	5.88	0	0.630	0.632	0.611	6.82	0
	0.25	0.419	0.418	0.428	5.77	0	0.290	0.303	0.291	6.87	0
$p = 10$	0.50	0.298	0.293	0.281	5.92	0	0.226	0.256	0.247	6.98	0
	0.75	0.330	0.423	0.452	5.73	0	0.301	0.316	0.317	6.91	0
	0.90	0.634	0.692	0.691	5.76	0	0.743	0.770	0.793	6.70	0
	0.10	4.650	6.500	7.180	5.42	0	0.599	0.571	0.640	46.84	0
	0.25	1.150	1.290	1.320	5.06	0	0.302	0.296	0.296	46.89	0
$p = 50$	0.50	0.705	0.734	0.716	5.11	0	0.228	0.252	0.242	46.92	0
	0.75	1.230	1.260	1.090	5.14	0	0.300	0.307	0.322	46.83	0
	0.90	4.260	4.440	4.210	5.50	0	0.749	0.705	0.808	46.68	0

3.3 Example 3

In this final example, we explore how the two approaches behave under a heteroscedastic error structure. We generated data from

$$y = 2t + 3 + \sigma(t)\epsilon$$

where $\sigma(t) = 2 + cos(2\pi t)$ and $t = \theta^T \mathbf{x}$. As before θ corresponds to the two cases mentioned above with x_i i.i.d Unif[0,1]. The error term ϵ follows a distribution of $0.25N(0, 1)$. The following two figures shows the estimated link functions for Lv et al. (2014) and our method respectively.

As shown in Figure 7 top row, when the number of covariates is small ($p = 10$), the link function estimates are quite good for Lv's approach. However, their performance goes down considerably when large number of non-significant covariates are introduced (Figure 7 bottom row). In contrast, our approach performs quite well under both settings ($p = 10$ and $p = 50$) as evident in Figure 8.

Finally, the performance of the variable selection aspect for both methods is summarized in Table 4. As before, it is quite evident that the proposed method outperforms Lv's method when large number of insignificant variables are involved.

Table 4. Summary of results: Mean Squared Error (MSE) of the non-zero coefficients and the identification of zero coefficients in Case 1 and 2 with linear link and a heteroscedastic error distribution.

		Lv et al. (2014)					Proposed Method				
		MSE (in 10^{-3})			no. of zeros		MSE (in 10^{-3})			no. of zeros	
	τ	θ_1	θ_2	θ_3	Correct	Wrong	θ_1	θ_2	θ_3	Correct	Wrong
	0.10	38.2	56.7	51.9	5.46	0.083	36.9	34.4	32.1	6.74	0.077
	0.25	9.76	9.36	9.09	5.52	0.001	9.22	9.58	9.03	6.58	0.001
$p = 10$	0.50	4.94	5.05	4.78	5.54	0	4.78	5.10	4.84	6.44	0
	0.75	3.85	4.20	3.86	5.42	0	4.20	4.36	4.27	6.56	0
	0.90	6.02	5.66	6.12	5.46	0.001	6.30	6.57	7.14	6.65	0
	0.10	65.4	134	115	5.49	0.007	63.2	61.7	63.3	46.8	0.270
	0.25	27.4	35.5	34.6	5.56	0	11.1	11.1	12.1	46.8	0.005
$p = 50$	0.50	15.3	15.1	16.1	5.60	0	5.23	5.7	5.85	46.8	0.001
	0.75	15.4	15.1	17.4	5.75	0	5.19	5.26	5.34	46.7	0.001
	0.90	28.4	34.4	33.2	5.27	0	9.99	9.77	11.7	46.7	0.007

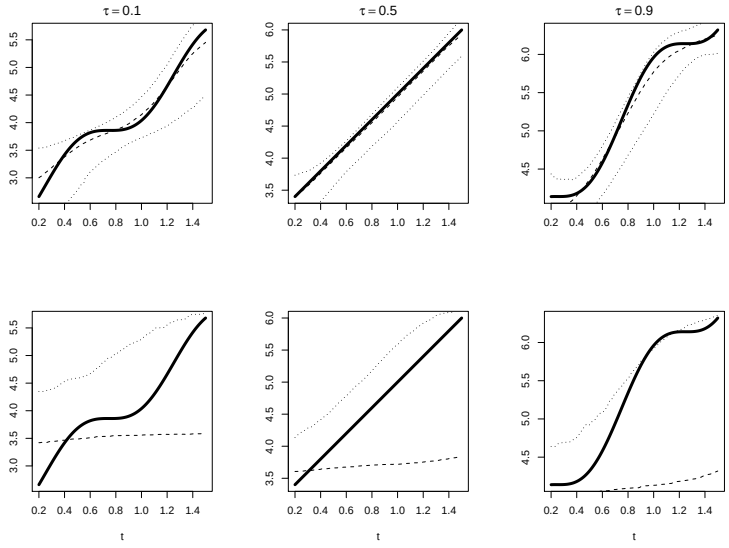

Figure 7. Comparison of estimated link functions using the proposed method with n=200: solid=true; dashed=50^{th}, dotted=5^{th} and 95^{th} percentiles of the quantile function estimates of 1000 simulated samples. Top row: $p = 10$ with 3 non-zero coefficients and 7 zero coefficients. Bottom row: $p = 50$ with 3 non-zero coefficients and 47 zero coefficients.

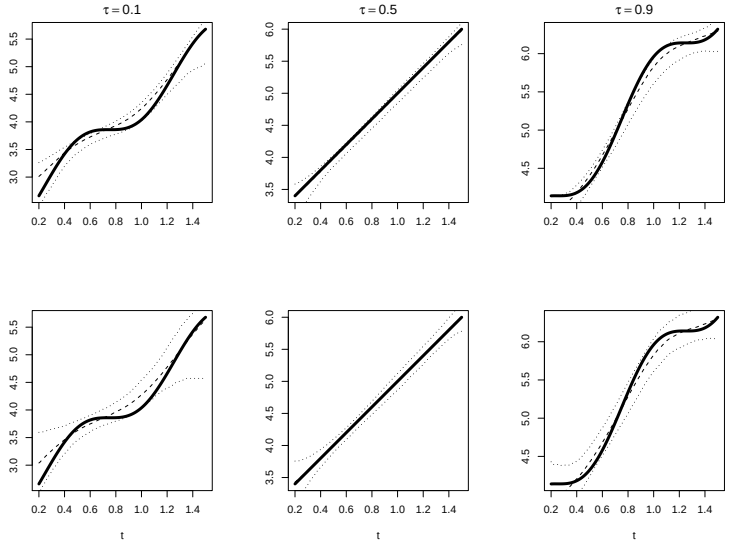

Figure 8. Comparison of estimated link functions using the proposed method with n=200: solid=true; dashed=50^{th}, dotted=5^{th} and 95^{th} percentiles of the quantile function estimates of 1000 simulated samples. Top row: $p = 10$ with 3 non-zero coefficients and 7 zero coefficients. Bottom row: $p = 50$ with 3 non-zero coefficients and 47 zero coefficients.

4. Data Analysis

We now present an application of our estimation procedure to real data. The data was originally used by Hirst et al. (1994) to understand the behavior of dihydrofolate reductase (DHFR) inhibitors. A DHFR inhibitor is a molecule that inhibits the function of the enzyme dihydrofolate reductase. Furthermore, a DHFR inhibitor is a type of antifolate. Since folate is needed by rapidly dividing cells, DHFR inhibition may be used to therapeutic advantage. For example, drugs that are based on DHFR inhibitors are used in cancer chemotherapy because it can prevent cells from dividing. Bacteria also need DHFR to grow and multiply and hence DHFR inhibitors are used

in antibacterial agents.

Table 5. Coefficient estimates of the DHFR inhibitor data. Three methods are compared: 1. linear quantile regression (LQR); 2. method by Lv et al. (2014) - denoted by (LV); 3. Proposed method - denoted by (New). A blank space indicates that the coefficient estimate is zero.

	$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.75$		
Variable	LQR	LV	New	LQR	LV	New	LQR	LV	New
1	-1.736	0.268		-2.509	0.264	0.010	0.568	0.129	
2	1.518		0.030	1.553	0.114	0.080	0.947	0.053	
3	-1.966	0.053	0.081	-2.008	0.117		-1.437	-0.008	
4	-0.600		0.046	-0.828	0.072		-1.413		
5	0.402	0.018		0.614			0.797		
6	0.006			-0.098	0.035		-1.385		0.051
7	0.518	0.079		0.977	0.115		-2.984	-0.004	
8	1.490	0.114	0.361	1.596	0.109	0.281	1.037	0.082	0.220
9	1.311	0.205		1.970	0.234		3.417	0.074	0.523
10	-2.820	0.198		-2.820	0.140		-3.904	0.151	0.210
11	2.410		0.273	1.516	0.001	0.180	1.262	0.005	0.024
12	-3.155	0.081	0.272	-2.762	0.181		-2.139		
13	0.583			0.225	0.010		0.162		
14	0.282			0.847	0.002		0.360		
15	-0.780		0.019	-0.423			-0.255		
16	0.191			0.610			0.776		
17	0.443			0.761		0.026	1.090		0.251
18	1.893	0.149		1.789	0.117		2.684		
19	-4.529	0.086	0.439	-5.469		0.481	-1.632	-0.002	
20	2.793	0.438		3.372	0.388	0.403	3.815	-0.020	0.444
21	-0.394			-0.164			-2.525	-0.008	
22	2.952	0.419	0.192	4.912	0.379	0.412	0.139		
23	-2.752	0.549		-4.290	0.455	0.563	-0.943	0.500	
24	1.973	0.249		5.004	0.457		-3.501	0.495	0.162
25									
26	-0.210	0.222		-1.640	0.202		1.594	0.365	0.110
27	4.510	0.091	0.694	4.106			3.922	0.562	0.576

The main goal of Hirst et al. (1994) is to use neural networks and logic programming to describe the quantitative structure-activity relationships (QSAR) in the inhibition of DHFR. QSAR is a model that correlates measurable or calculable physical or molecular properties to some specific biological response. Once a valid QSAR has been determined, it should be possible to predict the biological activity of related drug candidates before they are put through expensive and time-consuming biological testing (Hansch & Fujita (1964)). The QSAR method examined by Hirst et al. (1994) is for application to drug design problems where there is no receptor structure. The goal of the study is to identify the important attributes (variables) that relate to the inhibition level which is measured by $log(1/K)$, where K is the inhibition level as experimentally assayed.

The actual data set contains structural information on 74 drug variants of a particular DHFR inhibitor in *E. coli*. For each drug variant, there are 3 positions where chemical activity occur and 9 attributes per position leading to 27 total variables. These variables represent the attributes: polarity, size, flexibility, number of hydrogen-bond donors and acceptors, presence and strength of rc-acceptors and re-donors, polarisability of the molecular orbitals, and o-effect and branching. The attributes were assigned to reproduce general trends rather than precise values. For example, the size of a substituent was based on the number of carbon, nitrogen and oxygen atoms it contains, with size = 0 for hydrogen, size = 1 for single atoms and size = 2 for substituents containing two to four carbon, nitrogen or oxygen atoms. One of the variables (variable 25) had no variability and was removed from the data set. A complete description of the variables is given in Appendix A.

We analyzed this data under the single index qunatile regression framework to identify important variables that are strongly associated with the inhibition level $log(1/K)$. The following table summarizes the variables (attributes) selected by three methods: 1. linear quantile regression (LQR); 2. method by Lv et al. (2014) - denoted by (LV);

3. Proposed method - denoted by (New). The LQR is presented to use as a baseline to compare and contrast the variable selection aspect of the other two methods. We examined the impact of the 27 variables on three quantiles ($\tau = 0.25, 0.5$ and 0.75) of the conditional distribution of the response. For the ease of reading the values from the table, any blank space indicate a variable that is not selected. The tuning parameters for both methods were selected through leave-one-out cross-validation as described in section 2.1.

The main observation in Table 5 is that compared to Lv's method, the proposed approach consistently produce sparse index vectors at all three quantiles. This behavior is consistent with what we observed in our simulation study. In particular, the average number of covariates selected by Lv's method is about 16, whereas the new method selected about 10 variables on average. Both methods seems to indicate that the number of carbon, nitrogen and oxygen atoms at all three positions (variables 2,11, and 20) have an impact on the inhibition level. This is consistent what Hirst et al. (1994) reported using their classical mean regression and neural network approaches.

We are unable to comment on the actual chemical implication of these findings but it highlights the importance of looking at specific quantiles of a response variables, a task that was not possible under classical regression models. In terms of sparseness of the estimated coefficient vector, our method tends to produce more sparse estimates compared to Lv et al. (2014). This is consistent with what we observed in the simulation study.

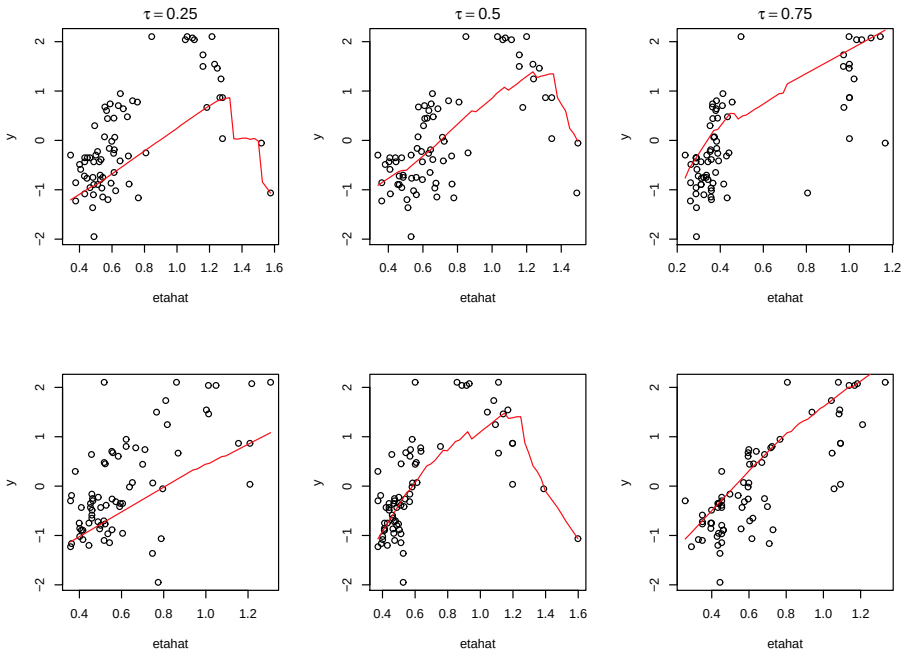

Figure 9. Comparison of estimated link functions for the DHFR inhibitor data. Top row: Method by Lv et al. (2014). Bottom row: Proposed Method

The link function estimates shown in Figure 9, points to a very interesting observation: for the conditional median ($\tau = 0.5$), the link function estimate resembles a clear quadratic pattern. This is consistent with what Hirst et al. (1994) reported in their original work using classical mean regression. They claimed that the quadratic term of the variable $X8$ (3-position polarizability) is statistically significant. This highlights the importance and the the flexibility of the single index approach in determining the appropriate link in exploratory data analysis.

5. Conclusion

The iterative estimation method that we propose in this article outperforms Lv et al. (2014) both in estimating the link function and selecting relevant variables especially when there are large number of non-significant variables are involved. The key to our success is the iterative penalizing technique and the adaptive penalty weights that we used in the estimation algorithm. Asymptotic properties of the proposed estimators are of great interest. Especially the consistency and the oracle property of the variable selection method. These aspects are currently being investigated.

References

Alkenani, A., & Yu, K. (2013). Penalized Single-Index Quantile Regression, *International Journal of Statistics and Probability*, *2*, 12-30. http://dx.doi.org/10.5539/ijsp.v2n3p12

Carroll, R.J., Fan, J., Gijbels, I., & Wand, M.P. (1997). Generalized patially linear single-index models, *Journal of the American Statistical Association*, *92*, 477-489. http://dx.doi.org/10.1080/01621459.1997.10474001

Chaudhuri, P., Doksum, K., & Samarov, A. (1997). On average derivative quantile regression, *Annals of Statistics*, *25*, 715-744. http://dx.doi.org/10.1214/aos/1031833670

Hansch, C., & Fujita, T. (1964). $p - \sigma - \pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure, *Journal of the American Chemical Society*, *86*(8), 1616-1626. http://dx.doi.org/10.1021/ja01062a035

Hansch C, Li, R., Blaney, J.M., & Langridge, R. (1982). Comparison of the inhibition of Escherichia coli and Lactobacillus casei dihydrofolate reductase by 2,4-diamino-5-(substituted-benzyl)pyrimidines: quantitative structure-activity relationships, X-ray crystallography, and computer graphics in structure-activity analysis, *Journal of Medicinal Chemistry*, *25*(7), 777-784.

Hirst, J.D., King, R.D., & Sternberg, M.J.E. (1994). Quantitative structure-activity relationships by neural networks and inductive logic programming. *Journal of Computer-Aided Molecular Design*, *8*, 405-420. http://dx.doi.org/10.1007/BF00125375

Ichimura, H. (1990). Semiparametric least squares (SLS) and weighted sls estimation of single- index models. *Journal of Econometrics*, 58, 71-120. http://dx.doi.org/10.1016/0304-4076(93)90114-K

Koenker, R., & Bassett, G. (1978). Regression quantiles *Econometrica*, *46*, 33-50. http://dx.doi.org/10.2307/1913643

Lin, W, & Kulasekera, K.B. (2007). Identifiability in single-index and additive-index models, *Biometrika*, *94*, 496-501. http://dx.doi.org/10.1093/biomet/asm029

Lv, Y., Zhang, R., Zhao, W., & Liu, J. (2014). Quantile regression and variable selection for the single-index model, *Journal of Applied Statistics*, *41*, 1565-1577. http://dx.doi.org/10.1080/02664763.2014.881786

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, *90*, 1257-1270. http://dx.doi.org/10.1080/01621459.1995.10476630

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society-B*, *58*, 267-288.

Wang, L., Li, G., & Jiang, G. (2014). Robust regression shrinkage and consistent variable selection through the LAD-Lasso, *Journal of Business and Economic Statistics*, *41*, 347-355.

Wu, Y., & Liu, Y. (2009). Variable selection in quantile regression, *Statistica Sinica*, *19*, 801-817.

Wu, T.Z., Yu, K., & Yu, Y., (2010). Single-index quantile regression, *Journal of Multivariate Analysis*, *101*, 1607-1621. http://dx.doi.org/10.1016/j.jmva.2010.02.003

Yu, K., & Jones, M. C. (2003). "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228-289. http://dx.doi.org/10.1080/01621459.1998.10474104

Zhu, L., Huang, M., & Li, R. (2012). Semiparametric quantile regression with high-dimensional covariates, *Statistica Sinica*, *22*, 1379-1401. http://dx.doi.org/10.5705/ss.2010.199

Zou, H. (2006), The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, *101*, 1418-1429. http://dx.doi.org/10.1198/016214506000000735

Appendix A

The description of the variables used in the data analysis. See Hirst et. al (1994) for more details.

Table 9. Description of variables

Variable Name	Description
X1	3-position polarity
X2	3-position number of carbon, nitrogen, and oxygen atoms
X3	3-position flexibility
X4	3-position number of hydrogen-bond donors
X5	3-position number of hydrogen-bond acceptors
X6	3-position presence and strength of π-donors
X7	3-position presence and strength of π-acceptors
X8	3-position polarizability
X9	3-position σ-effect
X10	4-position polarity
X11	4-position number of carbon, nitrogen, and oxygen atoms
X12	4-position flexibility
X13	4-position number of hydrogen-bond donors
X14	4-position number of hydrogen-bond acceptors
X15	4-position presence and strength of π-donors
X16	4-position presence and strength of π-acceptors
X17	4-position polarizability
X18	4-position σ-effect
X19	5-position polarity
X20	5-position number of carbon, nitrogen, and oxygen atoms
X21	5-position flexibility
X22	5-position number of hydrogen-bond donors
X23	5-position number of hydrogen-bond acceptors
X24	5-position presence and strength of π-donors
X25	5-position presence and strength of π-acceptors
X26	5-position polarizability
X27	5-position σ-effect
Y- Response	log 1/Ki, where Ki is the inhibition constant as experimentally assayed