Regularization of Case Specific Parameters: A New Approach for Improving Robustness and/or Efficiency of Statistical Methods

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Yoonsuh Jung

Graduate Program in Statistics

The Ohio State University

2010

Dissertation Committee:

Steven N. MacEachern, Co-Adviser

Yoonkyung Lee, Co-Adviser

Tao Shi

ABSTRACT

Regularization methods allow one to handle a variety of inferential problems where there are more covariates than cases. This allows one to consider a potentially enormous number of covariates for a problem. We exploit the power of these techniques, supersaturating models by augmenting the "natural" covariates in the problem with an additional indicator for each case in the data set. We attach a penalty term for these case-specific indicators which is designed to produce a desired effect. For regression methods with squared error loss, an ℓ_1 type penalty for case-specific parameters produces a regression which is robust to outliers and high leverage cases. Through this modification we have devised a robust LASSO which retains desirable property of the LASSO and performs better when outlying observations exist. For quantile regression methods, an ℓ_2 type penalty decreases the variance of the fit enough to overcome an increase in bias.

The paradigm thus allows us to robustify procedures which lack robustness and to increase the efficiency of procedures which are robust. Including the case-specific parameters can be viewed as a modification of the current loss function to produce better estimator. For the LASSO with the squared error loss, the modification yields Huber's loss. The check loss function in quantile regression is adjusted to be quadratic near its minimum. This modification produces an averaging effect near the target quantile thus more efficient quantile estimation in various settings. Applications

to classification procedures such as logistic regression and support vector machines are also considered. Finally, a modification to cross validation through use of a new validation function in quantile regression is investigated. The new validation function makes use of the same adjusted check loss which is used for estimation.

To my parents, my wife Hyoju An, and my daughter Ha-im.

ACKNOWLEDGMENTS

I would like to express my sincerest appreciation to my advisors, Prof. Steven MacEachern and Prof. Yoonkyung Lee for many useful research ideas, for their inspiring advice, encouragement, and guidance over the years. I thank Prof. Tao Shi for his academic suggestions and willingness to serve as my committee member. My special thanks go to my wife, Hyoju An, for her love and support while I finished the degree. I also thank my parents. Their enduring love and support helped me overcome all the difficulties I faced in the past.

VITA

December 26, 1977	Born - Seoul, Korea
February 2003	B.S. Statistics, Korea University, Seoul, Korea
June 2006	M.S. Statistics, The Ohio State University, Columbus, OH
September 2006-June 2007	Teaching Assistant, The Ohio State University, Columbus, OH
January 2008-December 2009	Statistical Consultant, The Ohio State University, Columbus, OH

FIELDS OF STUDY

Major Field: Statistics

Studies in:

Topic 1	LASSO
Topic 2	Quantile Regression
Topic 3	Classification

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Overview

As complex and high dimensional data become crucial in statistics, regularization (or penalization) methods have been actively studied and applied in many contexts. With regression often serving as a motivating theme, a host of methods for regularized model selection and estimation problems have been developed. Among these regularization methods are penalized linear regression methods (e.g. ridge regression (Hoerl and Kennard; 1970) and the LASSO (Tibshirani; 1996)), nonparametric regression methods, (e.g. smoothing splines (Wahba; 1990) and generalized additive models (Hastie and Tibshirani; 1990)), and extension to regression with a non-normal error distribution, namely, the generalized linear model (McCullagh and Nelder; 1989). Bickel and Li (2006) provide a wide and thoughtful review of regularization methods in statistics. Most of the current literature on these regularized regression techniques focuses on regularizing estimates of the regression parameters by introducing various forms of penalty for the regression coefficients. Interestingly, these existing regularization methods can be improved by introducing case-specific parameters which enable

us to modify the effect of each observation on the inference. This is the primary theme of this thesis.

To regularize each case, a case-specific indicator and a corresponding parameter is added to the existing regression problem. In the linear model, as a quick example, this is accomplished by adding an identity matrix to produce, say, $Y = X\beta + I\gamma + \epsilon$. The additional case-specific parameters must be penalized. We consider another penalty of possibly different form for the new coefficient vector, γ. Including the additional penalty for case-specific parameters does not require a regularization structure for β, although it is natural to regularize both γ and β.

A general framework for the inclusion of case-specific parameters in regularization problems can be built under the minimization of a given penalized empirical risk function. In principle, a second penalty for case-specific parameters is added to the genuine risk function. With little extra computational cost, and under mild conditions such as convexity of a risk function, it is feasible to find estimators for both the regression parameters and the case-specific parameters. Depending on the purpose of improvement, the exact form of the extra penalty is determined. Incorporation of a new penalty essentially brings a modification of the loss function which measures discrepancy between a model and data. As a result, a modified modeling procedure is produced which is more robust and/or more efficient, based on the type of additional penalty.

When an ℓ_1 type penalty is used for case-specific parameters, the squared error loss function for LASSO is adjusted to be linear for large errors. Resembling Huber's loss function in shape, this modification brings robustness to the modeling procedure. We examine the robustness of the modified LASSO in comparison with LASSO through a

simulation study where data are partially contaminated. We also study performance of the procedure on a set of language data (Baayen; 2007). The comparisons show superior performance of the robust LASSO while maintaining the desirable properties of LASSO.

On the contrary, quantile regression itself is known to be robust. Thus we consider an ℓ_2 type penalty for case-specific parameters, expecting the modified procedure to be more efficient. The additional penalty changes the V shape of the original check loss function to a form which is quadratic inside and linear outside. The averaging effect near the true quantile greatly reduces the variance while allowing a small bias. As a by-product of the averaging effect, it is observed that the undesirable crossings of the estimated quantiles, are reduced. Although modified quantile regression yields a biased estimator in a finite sample, we design the modification to produce a consistent estimator by decreasing the size of the adjustment to zero as the sample size grows. Through extensive simulations, a rule to guide the scope of adjustment is also developed. The rule is designed to be scale invariant, and it is shown to perform very well for both symmetric and asymmetric error distributions. Applications to real data also support the improved efficiency of modified quantile regression. We extend the modification to heteroscedastic models, and the modified method shows better performance than the standard quantile regression in simulation studies.

The modified check loss function provides an interesting extension to cross validation. It can serve as an alternative validation function. Employing the modified check loss function, instead of the check loss, allows us to assess the fitted quantile regression more accurately, due to the averaging effect on the prediction error.

Case-specific parameters also can be applied to classification problems. We consider several classification procedures, such as logistic regression and the support vector machine. The strategy of choosing the extra penalty for case-specific parameters is similar to the above regression problems. When a procedure lacks robustness, an ℓ_1 type penalty for γ is desirable. On the contrary, considering an ℓ_2 type penalty may provide more efficiency for support vector machines. As in regression, the additional parameters effectively modify the original loss function.

1.2 Outline of the Thesis

We provide a general description of inclusion of case-specific parameters and their regularization in Chapter 2, under the frame of risk minimization. Application of case-specific parameters to LASSO is in Chapter 3, which results in the robust LASSO. Detailed description of quantile regression, modified by case-specific parameters is discussed in Chapter 4. We consider modification of a procedure first, then develop a heuristic rule through extensive simulations to guide the size of modification. Asymptotic properties and finite sample performance of modified quantile regression are also addressed. In Chapter 5, a new cross validation function for quantile regression is investigated. The modified check loss function developed in Chapter 4 is considered as an alternative validation function. In Chapter 6, some classification procedures adjusted by case-specific parameters are examined. Concluding remarks are given in the last chapter.

CHAPTER 2

INCLUSION OF CASE-SPECIFIC PARAMETERS

2.1 General Modeling Procedure

Suppose that we have n pairs of observations denoted by (x_i, y_i), $i = 1, \ldots, n$, for statistical modeling and prediction. Here $x_i = (x_{i1}, \ldots, x_{ip})^\top$ with p covariates and the y_i's are responses. As in the standard setting of regression and classification, the y_i's are assumed to be conditionally independent given the x_i's. In this chapter, we take modeling of the data as a procedure of finding a functional relationship between x_i and y_i, $f(x; \beta)$ with unknown parameters $\beta \in \mathbb{R}^p$ that is consistent with the data. The discrepancy, or lack of fit, of f is measured by a loss function $\mathcal{L}(y, f(x; \beta))$. Consider a modeling procedure, say, $\mathcal{M}$, of finding f which minimizes the empirical risk

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i; \beta))$$

or its penalized version

$$R_n(f) + \lambda J(f) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i; \beta)) + \lambda J(f),$$

where λ is a positive penalty parameter for balancing the data fit and the model complexity of f as measured by $J(f)$. A variety of common modeling procedures are

subsumed under this formulation, including ordinary linear regression, generalized linear models, nonparametric regression, and supervised learning techniques. For brevity of exposition, we identify f with β and view $J(f)$ as a functional depending on β.

Motivated by the LASSO type ℓ_1 norm regularization, we propose a general scheme to modify the modeling procedure $\mathcal{M}$. First, we introduce case-specific parameters, $\gamma = (\gamma_1, \ldots, \gamma_n)^\top$, for the n observations and modify $\mathcal{M}$ to be the procedure of finding the original model parameters, β, together with the case-specific parameters, γ, that minimize

$$L(\beta, \gamma) = \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i; \beta) + \gamma_i) + \lambda_\beta J(f) + \lambda_\gamma J_2(\gamma). \tag{2.1}$$

If λ_β is zero, $\mathcal{M}$ is empirical risk minimization, otherwise it is penalized risk minimization. In general, $J_2(\gamma)$ measures the size of γ. When concerned with robustness, we often take $J_2(\gamma) = \|\gamma\|_1 = \sum_{i=1}^{n} |\gamma_i|$. A rationale for this choice is that with added flexibility, the case-specific parameters can curb the undesirable influence of individual cases on the fitted model. Such a case specific adjustment of the model would be necessary only for a small number of potential outliers, and the ℓ_1 norm which yields sparsity works to that effect. When concerned with efficiency, we often take $J_2(\gamma) = \|\gamma\|_2^2 = \sum_{i=1}^{n} \gamma_i^2$. This choice has the effect of increasing the impact of selected, non-outlying cases on the analysis.

2.2 Algorithm for Finding Solutions

Although the computational details for obtaining the solution to (2.1) are specific to each modeling procedure $\mathcal{M}$, it is feasible to describe a common computational strategy which is effective for a wide range of procedures. For fixed λ_β and λ_γ, the

solution pair of $\hat{\beta}$ and $\hat{\gamma}$ to the modified $\mathcal{M}$ can be found with little extra computational cost. A generic algorithm below alternates estimation of β and γ. Given $\hat{\gamma}$, minimization of $L(\beta, \hat{\gamma})$ is done via the original modeling procedure $\mathcal{M}$. Take $J_2(\gamma) = \|\gamma\|_1$ as an example here. Fixing $\hat{\beta}$, we seek to minimize $L(\hat{\beta}, \gamma)$, which decouples to a minimization of $\mathcal{L}(y_i, f(x_i; \hat{\beta}) + \gamma_i) + \lambda_\gamma|\gamma_i|$ for each γ_i. In most cases, an explicit form of the minimizer $\hat{\gamma}$ of $L(\hat{\beta}, \gamma)$ can be obtained. This adjustment is equivalent to changing the loss from $\mathcal{L}(y, f(x; \beta))$ to $\mathcal{L}(y, f(x; \beta) + \hat{\gamma})$, which we call the γ-adjusted loss of $\mathcal{L}$. Alternatively, one may view

$$\mathcal{L}_{\lambda_\gamma}(y, f(x; \beta)) := \min_{\gamma \in \mathbb{R}} \{\mathcal{L}(y, f(x; \beta) + \gamma) + \lambda_\gamma|\gamma|\} = \mathcal{L}(y, f(x; \beta) + \hat{\gamma}) + \lambda_\gamma|\hat{\gamma}|$$

as an "effective loss". Concrete examples of the adjustments will be given in the next chapters. These considerations lead to the following iterative algorithm for finding $\hat{\beta}$ and $\hat{\gamma}$.

1. Initialize $\hat{\gamma}^{(0)} = 0$ and $\hat{\beta}^{(0)} = arg\min_\beta L(\beta, 0)$ (the ordinary $\mathcal{M}$ solution).

2. Iteratively alternate the following two steps, $m = 0, 1, \ldots$

   - $\hat{\gamma}^{(m+1)} = arg\min_\gamma L(\hat{\beta}^{(m)}, \gamma)$ modifies "residuals".
   - $\hat{\beta}^{(m+1)} = arg\min_\beta L(\beta, \hat{\gamma}^{(m+1)})$. This step amounts to reapplying the $\mathcal{M}$ procedure to $\hat{\gamma}^{(m+1)}$-adjusted data although the nature of the data adjustment would largely depend on L.

3. Terminate the iteration when $\|\hat{\beta}^{(m+1)} - \hat{\beta}^{(m)}\|^2 < \epsilon$, where ϵ is a prespecified convergence tolerance.

In a nutshell, the algorithm attempts to find the joint minimizer (β, γ) by combining the minimizers β and γ resulting from the projected subspaces. Convergence of the iterative updates can be established under appropriate conditions.

2.3 Convergence

Convexity of the loss and penalty terms plays a primary role in characterizing the solutions of the iterative algorithm. For a general reference to properties of convex functions and convex optimization, see Rockafellar (1997). First, we ensure that the minimizer pair (β, γ) in each step is properly defined.

Lemma 1 *Suppose that $L(\beta, \gamma)$ in (2.1) is continuous and strictly convex in β and γ for fixed λ_β and λ_γ. Given γ, there exists a unique minimizer $\beta(\gamma) = arg\min_\beta L(\beta, \gamma)$, and vice versa.*

Proof. Given γ, choose an arbitrary $\beta^0 \in \mathbb{R}^p$ and consider $A_\beta := \{\beta \in \mathbb{R}^p \mid L(\beta, \gamma) \leq L(\beta^0, \gamma)\}$. By the continuity of L, A_β is closed. For a fixed $\lambda_\beta > 0$, it is bounded because it is contained in the ℓ_p ball of $\{\beta \in \mathbb{R}^p \mid \|\beta\|_p^p \leq L(\beta^0, \gamma)/\lambda_\beta\}$. Thus, there exists β in the compact set A_β attaining the minimum. Uniqueness follows from the strict convexity of L. Similarly, given β, there exists a unique minimizer $\gamma(\beta) = arg\min_\gamma L(\beta, \gamma)$. □

The assumption that $L(\beta, \gamma)$ is strictly convex holds if the loss $\mathcal{L}(y, f(x; \beta))$ itself is strictly convex. Also, it is satisfied when a convex $\mathcal{L}(y, f(x; \beta))$ is combined with $J(f)$ and $J_2(\gamma)$ strictly convex in β and γ, respectively.

Lemma 2 *Under the same condition as Lemma 1, let* $h : \mathbb{R}^{p+n} \to \mathbb{R}^{p+n}$ *be the mapping of* (β, γ) *to its one-step update* (β^1, γ^1) *by the iterative algorithm. Then,* h *is continuous.*

Proof. By the iterative strategy, (β^1, γ^1) depends on β only, and β^1 is a composite mapping of β to β^1. So, it is sufficient to show that the mappings of β to γ^1 and γ^1 to β^1 are continuous. Although the former, finding γ^1 given β, is a much simpler optimization than the latter in general for an array of $L(\beta, \gamma)$'s of practical interest, by the symmetry in the problem, we will show only that the mapping of γ^1 to β^1 is continuous. Dropping the superscript for notational simplicity, consider $g(\gamma) = \inf_\beta L(\beta, \gamma)$. Since $L(\beta, \gamma)$ is convex, g is convex in γ (see, for example, Rockafellar (1997), p.38) and thus continuous. For any sequence, $\{\gamma_n\}_{n=1}^\infty$ converging to γ, we want to show that the corresponding sequence of β minimizers, $\{\beta(\gamma_n)\}_{n=1}^\infty$, converges to $\bar{\beta} := \beta(\gamma)$.

As g is continuous, for $\epsilon > 0$, there is N such that $n \geq N$ implies $g(\gamma_n) \leq g(\gamma) + \epsilon$, that is, $L(\beta(\gamma_n), \gamma_n) \leq L(\bar{\beta}, \gamma) + \epsilon$. Let $M := \max\{\max_{n=1,\ldots,N} L(\beta(\gamma_n), \gamma_n), L(\bar{\beta}, \gamma) + \epsilon\}$ and $A := \{(\beta, \gamma) \mid L(\beta, \gamma) \leq M\}$. Note that A is closed and bounded for fixed λ_β and λ_γ, and it contains the sequence of $\{(\beta(\gamma_n), \gamma_n)\}_{n=1}^\infty$. Therefore $\{(\beta(\gamma_n)\}_{n=1}^\infty$ has a convergent subsequence $\{(\beta(\gamma_{n_k})\}_{k=1}^\infty$ with a limit, say, β^*. Then by the continuity of g and L,

$$L(\bar{\beta}, \gamma) = g(\gamma) = \lim_{k\to\infty} g(\gamma_{n_k}) = \lim_{k\to\infty} L(\beta(\gamma_{n_k}), \gamma_{n_k}) = L(\beta^*, \gamma). \tag{2.2}$$

The uniqueness of the minimizer $\bar{\beta}$ at γ and (2.2) imply that $\beta^* = \bar{\beta}$. Consequently, this proves that every convergent subsequence of the bounded sequence $\{(\beta(\gamma_n)\}_{n=1}^\infty$ has the same limit $\bar{\beta}$. Hence the limit of $\{(\beta(\gamma_n)\}_{n=1}^\infty$ is $\bar{\beta}$, which completes the proof.

□

Proposition 3 *Suppose that $L(\beta, \gamma)$ is strictly convex in β and γ with a unique minimizer (β^*, γ^*) for fixed λ_β and λ_γ. Then, the iterative algorithm gives a sequence of $(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})$ with strictly decreasing $L(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})$. Moreover, $(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})$ converges to (β^*, γ^*).*

Proof. For fixed λ_β and λ_γ, by the definition of $\hat{\gamma}^{(m+1)}$ and $\hat{\beta}^{(m+1)}$ in the algorithm, we have

$$L(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}) \geq L(\hat{\beta}^{(m)}, \hat{\gamma}^{(m+1)}) \geq L(\hat{\beta}^{(m+1)}, \hat{\gamma}^{(m+1)}).$$

Unless $(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}) = (\beta^*, \gamma^*)$, at least one of the inequalities must be strict by the strict convexity of L. Thus, $\{L(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})\}_{m=0}^{\infty}$ is a monotonically decreasing sequence. Since it is bounded below by zero, the sequence has a limit, say, L^*. Now consider $A := \{(\beta, \gamma) \mid L(\beta, \gamma) \leq L(\hat{\beta}^{(0)}, \hat{\gamma}^{(0)})\}$, which is closed and bounded. The sequence of the minimizers, $(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})$, is contained in A, and therefore it has a convergent subsequence $\{(\hat{\beta}^{(m_k)}, \hat{\gamma}^{(m_k)})\}_{k=1}^{\infty}$. Let $(\bar{\beta}, \bar{\gamma})$ denote the limit of the subsequence. By the continuity of L, $\lim_{k\to\infty} L(\hat{\beta}^{(m_k)}, \hat{\gamma}^{(m_k)}) = L(\bar{\beta}, \bar{\gamma})$.

Suppose that $(\bar{\beta}, \bar{\gamma}) \neq (\beta^*, \gamma^*)$, i.e., $L(\bar{\beta}, \bar{\gamma}) > L(\beta^*, \gamma^*)$. Then we can obtain the one-step update of $(\bar{\beta}, \bar{\gamma})$ denoted by $(\bar{\beta}^1, \bar{\gamma}^1)$ and further reduce the objective value by $\epsilon := L(\bar{\beta}, \bar{\gamma}) - L(\bar{\beta}^1, \bar{\gamma}^1) > 0$. By Lemma 2, the mapping of $(\bar{\beta}, \bar{\gamma})$ to $(\bar{\beta}^1, \bar{\gamma}^1)$ is continuous. So, there exists a $\delta > 0$ such that for any (β, γ) in an open ball centered at $(\bar{\beta}, \bar{\gamma})$ with a radius ϵ, $|L(\beta^1, \gamma^1) - L(\bar{\beta}^1, \bar{\gamma}^1)| < \epsilon/2$. This implies that for sufficiently large k, $L(\hat{\beta}^{(m_k+1)}, \hat{\gamma}^{(m_k+1)}) \leq L(\bar{\beta}^1, \bar{\gamma}^1) + \epsilon/2$. However, this leads to a

contradiction that $L(\hat{\beta}^{(m_k+1)}, \hat{\gamma}^{(m_k+1)}) \leq \{L(\bar{\beta}, \bar{\gamma}) - \epsilon\} + \epsilon/2 = L(\bar{\beta}, \bar{\gamma}) - \epsilon/2$. Therefore, $(\bar{\beta}, \bar{\gamma}) = (\beta^*, \gamma^*)$. Furthermore, since the limit of any convergent subsequence is the same, we conclude that $(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})$ converges to (β^*, γ^*). □

CHAPTER 3

ROBUST LASSO

3.1 Literature Review of LASSO

The LASSO (Tibshirani; 1996), an acronym for Least Absolute Shrinkage and Selection Operator, makes use of an ℓ_1 penalty to regularize a least squares model. The ℓ_1 penalization induces bias in estimates of the regression parameters, and it tends to produce "sparse" solutions in which many of the fitted regression coefficients are zero. For high dimensional problems with many extraneous covariates, the reduction in variance (through shrinkage and selection) that stems from the ℓ_1 penalty outweighs the increase in bias and leads to a more accurate inference. The LASSO is closely connected to many other methods that exploit a bias-variance tradeoff, including ridge regression (Hoerl and Kennard; 1970) and bridge regression (Frank and Friedman; 1993), with a more general form of penalty. The asymptotic properties of the LASSO type estimators are given by Knight and Fu (2000). Many other penalties have been developed with an eye toward different model selection or estimation properties. Examples include the elastic net (Zou and Hastie; 2005), the grouped LASSO (Yuan and Lin; 2006), and the grouped LASSO for logistic regression (Meier et al.; 2008). The idea of using an ℓ_1 constraint has been applied to various modeling procedures

such as kernel machines (Gunn and Kandola (2002); Roth (2004)), quantile regression (Li and Zhu (2008); Belloni and Chernozhukov (2009)), and boosting (Zhao and Yu; 2004). Along with rapid expansion of techniques, computational algorithms to fit LASSO type estimators have been actively investigated under the linear model, (Efron et al. (2004); Osborne and Turlach (2000)); (Perkins et al. (2003); Kim et al. (2008)).

3.2 Case Specific Parameters in LASSO

The LASSO estimate of $\beta = (\beta_1, \ldots, \beta_p)^\top$ is defined as the solution $\hat{\beta} \in \mathbb{R}^p$ that minimizes

$$L_\lambda(\beta) = \frac{1}{2}(Y - X\beta)^\top (Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{3.1}$$

where λ is a regularization parameter for balancing the data fit and the amount of shrinkage of β. We take this as a baseline model fitting procedure for illustration.

To reduce the sensitivity of the LASSO solution to influential observations, the given p covariates are augmented by n case indicators. Let z_i be the indicator variable taking 1 for the ith observation and 0 otherwise, and $\gamma = (\gamma_1, \ldots, \gamma_n)^\top$ be the coefficients of the case indicators. The proposed modification of the LASSO with $J_2(\gamma) = \|\gamma\|_1$ leads to the robust LASSO. For the robust LASSO, we find $\hat{\beta} \in \mathbb{R}^p$ and $\hat{\gamma} \in \mathbb{R}^n$ that minimize

$$L(\beta, \gamma) = \frac{1}{2}\{Y - (X\beta + \gamma)\}^\top \{Y - (X\beta + \gamma)\} + \lambda_\beta \sum_{j=1}^{p} |\beta_j| + \lambda_\gamma \sum_{i=1}^{n} |\gamma_i|, \tag{3.2}$$

where λ_β and λ_γ are fixed regularization parameters constraining β and γ. Just as the ordinary LASSO in (3.1) stabilizes the solution by shrinking and selecting β, the additional penalty in the robust LASSO in (3.2) has the same effect on γ, whose components gauge the extent of case influences.

The minimizer $\hat{\gamma}$ of $L(\hat{\beta}, \gamma)$ for a fixed $\hat{\beta}$ can be found by soft-thresholding the residual vector $r = Y - X\hat{\beta}$. That is, $\hat{\gamma} = sgn(r)(|r| - \lambda_\gamma)_+$. For observations with small residuals, $|r_i| \leq \lambda_\gamma$, $\hat{\gamma}_i$ is set equal to zero with no effect on the current fit and for those with large residuals, $|r_i| > \lambda_\gamma$, $\hat{\gamma}_i$ is set equal to the residual $r_i = y_i - x_i^\top \hat{\beta}$ offset by λ_γ towards zero. Combining $\hat{\gamma}$ with $\hat{\beta}$, we define the adjusted residuals to be $r_i^* = y_i - x_i^\top \hat{\beta} - \hat{\gamma}_i$. That is, $r_i^* = r_i$ if $|r_i| \leq \lambda_\gamma$, and $r_i^* = sgn(r_i)\lambda_\gamma$, otherwise. Thus, introduction of the case-specific parameters along with the the ℓ_1 penalty on γ amounts to winsorizing the ordinary residuals. The γ-adjusted loss is equivalent to truncated squared error loss which is $(y - x^\top \beta)^2$ if $|y - x^\top \beta| \leq \lambda_\gamma$, and is λ_γ^2 otherwise. Figure 3.1 shows (a) the relationship between the ordinary residual r and the corresponding γ, (b) the residual and the adjusted residual r^*, (c) the γ- adjusted loss as a function of r, and (d) the effective loss.

The effective loss is $\mathcal{L}_{\lambda_\gamma}(y, x^\top \beta) = (y - x^\top \beta)^2/2$ if $|y - x^\top \beta| \leq \lambda_\gamma$, and $\lambda_\gamma^2/2 + \lambda_\gamma(|y - x^\top \beta| - \lambda_\gamma)$ otherwise. This effective loss matches Huber's loss function for robust regression (Huber; 1981), and yields the Huberized LASSO described by Rosset and Zhu (2004). As in robust regression, we choose a sufficiently large λ_γ so that only a modest fraction of the residuals are adjusted.

3.3 Bayesian Interpretation of Robust LASSO

Tibshirani (1996) suggested that LASSO solution has a dual interpretation as a posterior mode of β when the prior distribution of β is independent double exponential. Similar Bayesian versions of LASSO related methods include Yuan and Lin (2005), Park and Casella (2008), and Hans (2009). An analogue can be drawn for the robust LASSO. Suppose that the β_j's and γ_i's have independent double exponential

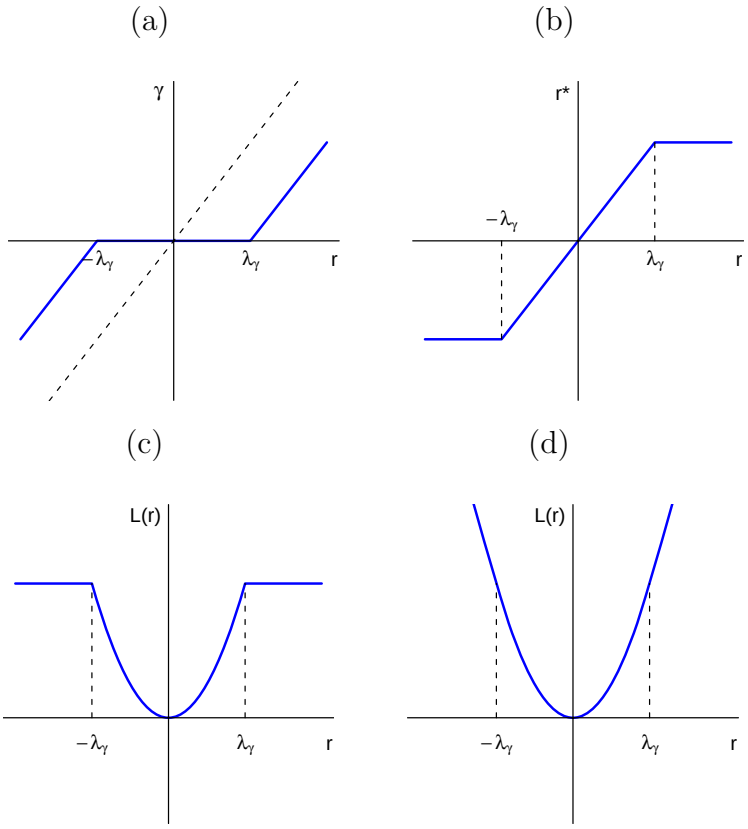

Figure 3.1: Modification of the squared error loss with a case-specific parameter. (a) γ versus the residual r. (b) the adjusted residual r^* versus the ordinary residual r. (c) a truncated squared error loss as the γ-adjusted loss. (d) the effective loss.

priors with mean 0 and the scale parameters σ_β and σ_γ, respectively. Consider a normal distribution with mean 0 and standard deviation σ for the errors ϵ_i. Treating the hyperparameters σ_β and σ_γ, and the error variance σ^2 as known constants, we have the posterior density function of β and γ

$$p(\beta, \gamma | X, Y) \propto$$
$$\exp\left(-\tfrac{1}{2\sigma^2}\{Y - (X\beta + \gamma)\}^\top \{Y - (X\beta + \gamma)\} - \tfrac{1}{\sigma_\beta}\sum_{j=1}^{p} |\beta_j| - \tfrac{1}{\sigma_\gamma}\sum_{i=1}^{n} |\gamma_i| \right).$$

Thus, the posterior mode of (β, γ) is the maximizer of $\log\ p(\beta, \gamma | X, Y)$, or equivalently the minimizer of

$$\frac{1}{2}\{Y - (X\beta + \gamma)\}^\top \{Y - (X\beta + \gamma)\} + \frac{\sigma^2}{\sigma_\beta}\sum_{j=1}^{p} |\beta_j| + \frac{\sigma^2}{\sigma_\gamma}\sum_{i=1}^{n} |\gamma_i|.$$

Reparametrization of σ^2/σ_β and σ^2/σ_γ as λ_β and λ_γ yields the objective function of the robust LASSO in (3.2). With this reformulation, we have another interpretation of λ_β and λ_γ as the so-called noise-to-signal ratios.

We note that a Bayesian might not choose the posterior mode as his or her estimator. Rather, a formal inference problem would be created, and an appropriate estimator would be chosen, as discussed in Hans (2009).

3.4 Computation

Taking the LASSO as a primary example, we describe details of computational implementation and illustrate how existing software can be easily altered to fit the robust LASSO.

3.4.1 Robust LASSO Algorithm

Recall the problem of finding the solution $\hat{\beta}$ and $\hat{\gamma}$ to the robust LASSO in (3.2) for fixed λ_β and λ_γ. The iterative algorithm in Chapter 2 can be restated for the LASSO as follows:

1. Initialize $\hat{\gamma}^{(0)} = 0$ and get the ordinary LASSO solution $\hat{\beta}^{(0)} = arg\min L(\beta, 0)$.

2. Iteratively alternate the following two steps, $m = 0, 1, \ldots$

   - $\hat{\gamma}^{(m+1)} = arg\min L(\hat{\beta}^{(m)}, \gamma) = sgn(r^{(m)})(|r^{(m)}| - \lambda_\gamma)_+$, where $r^{(m)} = Y - X\hat{\beta}^{(m)}$.
   - $\hat{\beta}^{(m+1)} = arg\min L(\beta, \hat{\gamma}^{(m+1)})$. This step is LASSO optimization with a winsorized response $Y^* = Y - \hat{\gamma}^{(m+1)} = X\hat{\beta}^{(m)} + r^{*(m)}$.

3. Terminate when $\sum_{j=1}^{p}(\hat{\beta}_j^{(m+1)} - \hat{\beta}_j^{(m)})^2/p < \epsilon$ for a prespecified small value ϵ.

For fitting β, existing algorithms for LASSO such as LARS (Efron et al.; 2004) or algorithm based on Osborne et al. (2000) can be used. Especially, Osborne et al. (2000)'s algorithm can fit the data where number of parameters exceeds the number of observations. Since the LARS algorithm can generate the entire solution path of $\hat{\beta}$ as a function of λ_β, it may be numerically more efficient to incorporate tuning of λ_β with the iterative algorithm than to find $\hat{\beta}$ for a fixed λ_β. So, we may consider joint updating of λ_β and $\hat{\beta}$ as well as λ_γ and $\hat{\gamma}$ as suggested in Gu (1992) for the similar issue of combining an iterative algorithm with tuning. Namely, one can choose the best λ_β at each iteration for fitting β and the best λ_γ for fitting γ with respect to some selection criteria.

3.4.2 Selection of the Penalty Parameters

As with methods of regularization in general, the choice of the penalty parameters is important for the effectiveness of the robust LASSO. Appropriate selection of λ_β and λ_γ needs to be combined with the iterative algorithm.

The role of λ_γ as a bending constant for winsorizing the residuals makes it sensible to set $\lambda_\gamma = k \cdot \sigma$ with a proper choice of k. The standard robust statistics literature (Huber (1981)) suggests that good choices of k lie in the range from 1 to 2. The effect of the k on estimation error is illustrated in the simulation studies in Section 3.6. Setting λ_γ in this fashion requires an estimate of σ. One may use a robust estimate by fitting a full robust regression model with an M estimator as implemented in `MASS` package. Alternatively, one can refit σ in each iteration and dynamically update λ_γ. Empirically, little difference has been observed between the two methods.

For the choice of λ_β, C_p-type risk estimates can be used. Note that the C_p criterion requires an accurate estimate of σ, and its implementation in the LARS algorithm uses $\hat{\sigma}$ from the full OLS fit. As potential outliers may significantly influence this assessment of fit through C_p, a robust estimate of σ is recommended for C_p as well. Ronchetti and Staudte (1994) obtain a robust version of C_p by replacing the squared error loss with its truncated version and p with $c \cdot p$ in C_p, where c is a constant, slightly smaller than 1, that depends on the the bending constant k. For instance, $c \approx 0.9817$ for $k = 1.345$ and 0.9985 for $k = 2$. The ordinary sum of squared residuals with the updated Y^* in the C_p evaluation is effectively the same as the empirical risk with respect to truncated squared error loss. With k as large as 2, the C_p in the LARS iterations is close to the robust C_p. This justifies selection of λ_β using C_p in each iteration.

An alternative method for selection of λ_β is generalized cross validation (GCV). Unlike C_p, its evaluation does not depend on $\hat{\sigma}^2$ although the leave-one-out cross validation identity, the theoretical basis for the GCV has not been established for the robust LASSO.

With C_p modified by the robust estimate of σ, numerical experiments indicate that the robust LASSO algorithm typically converges in several iterations, with a noticeable change in coefficients, if any, occurring in the first update only. For large k, fewer iterations are needed.

3.5 Simulations

We conducted a simulation study to investigate the sensitivity of the LASSO (or LARS) and the robust LASSO (robust LARS, $k = 2$) to contamination of the data. For brevity, we report only that portion of the results pertaining to accuracy of the fitted regression surface and inclusion of covariates in the model. Similar results were obtained for k near 2. The results differ for extreme values of k. Throughout the simulation, the standard linear model $y = x^\top \beta + \epsilon$ was assumed. Following the simulation setting in Tibshirani (1996), we generated $x = (x_1, \ldots, x_8)^\top$ from a multivariate normal distribution with mean zero and standard deviation 1. The correlation between x_i and x_j was set to $\rho^{|i-j|}$ with $\rho = 0.5$. Three scenarios were considered with a varying degree of sparsity in terms of the number of non-zero true coefficients: i) sparse: $\beta = (5, 0, 0, 0, 0, 0, 0, 0)$, ii) intermediate: $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, and iii) dense: $\beta_j = 0.85$ for all $j = 1, \ldots, 8$. In all cases, the sample size was 100. For the base case, ϵ_i was assumed to follow $N(0, \sigma^2)$ with $\sigma = 3$.

For potential outliers in ϵ, the first 5% of the ϵ_i's were tripled, yielding a data set with more outliers. We also investigated sensitivity to high leverage cases. For this setting, we tripled the first 5% of the values of x_1. Thus the replicates were blocked across the three settings. There were 100 replicates in the simulation. The C_p criterion was used to select the model.

Figure 3.2 shows mean square error (MSE) between the fitted and true regression surfaces, omitting intercepts. MSE is integrated across the distribution of a future X, taken to be that for the base case of the simulation. Over the n replicates in the simulation, $\widehat{MSE} = n^{-1}\sum_{i=1}^{n}(\hat{\beta}^i - \beta)^\top \Sigma(\hat{\beta}^i - \beta)$, where $\hat{\beta}^i$ is the estimate of β for the i^{th} replicate. LARS and robust LARS perform comparably in the base case, with the $\widehat{MSE}$ for robust LARS being greater by 1 to 6 percent. For both LARS and robust LARS, $\widehat{MSE}$ in the base case increases as one moves from the sparse to the dense scenario. $\widehat{MSE}$ increases noticeably when ϵ is contaminated, by a factor of 1.31 to 1.41 for LARS. For robust LARS, the factor for increase over the base case with LARS is 1.12 to 1.22. For contamination in X, results under LARS and robust LARS are similar in the intermediate and dense cases, with increases in $\widehat{MSE}$ over the base case. For the sparse case, the coefficient of the contaminated covariate, x_1, is large relative to the other covariates. Here, robust LARS performs noticeably better than LARS, with a smaller increase in $\widehat{MSE}$.

Table 3.1 presents results on the difference in number of selected variables for pairs of models. In each pair, a contaminated model is contrasted with the corresponding uncontaminated model. The top half of the table presents results for contamination of ϵ. The distribution of the differences in the number of selected variables for the pairs of fitted models has a mode at 0 in each scenario for both LARS and robust

LARS. There is, however, substantial spread around 0. The fitted models for the data with contaminated errors tend to have fewer variables than those for the original data, especially in the dense scenario. This may well be attributed to inflated estimates of σ^2 used in C_p for the contaminated data, favoring relatively smaller models. The effect is stronger for LARS than for robust LARS, in keeping with the lessened impact of outliers on the robust estimate of σ^2.

The bottom half of Table 3.1 presents results for contamination of X. Again, the distributions of differences in model size have modes at 0 in all scenarios. The distributions have substantial spread around 0. Under the sparse scenario in which the contamination has a substantial impact on MSE, the distribution under robust LARS is more concentrated than under LARS.

The simulation demonstrates that the proposed robustification is successful in dealing with both contaminated errors and contaminated covariates. As expected, in contrast to LARS, robust LARS is effective in identifying observations with large measurement errors and lessening their influence. It is also effective at reducing the impact of high leverage cases, especially when the high leverage arises from a covariate with a large regression coefficient. The combined benefits of robustness to outliers and high leverage cases render robust LARS effective at dealing with influential cases in an automated fashion.

3.6 Analysis of Language Data

Balota et al. (2004) conducted an extensive lexical decision experiment in which subjects were asked to identify whether a string of letters was an English word or a non-word. The words were monosyllabic, and the non-words were constructed

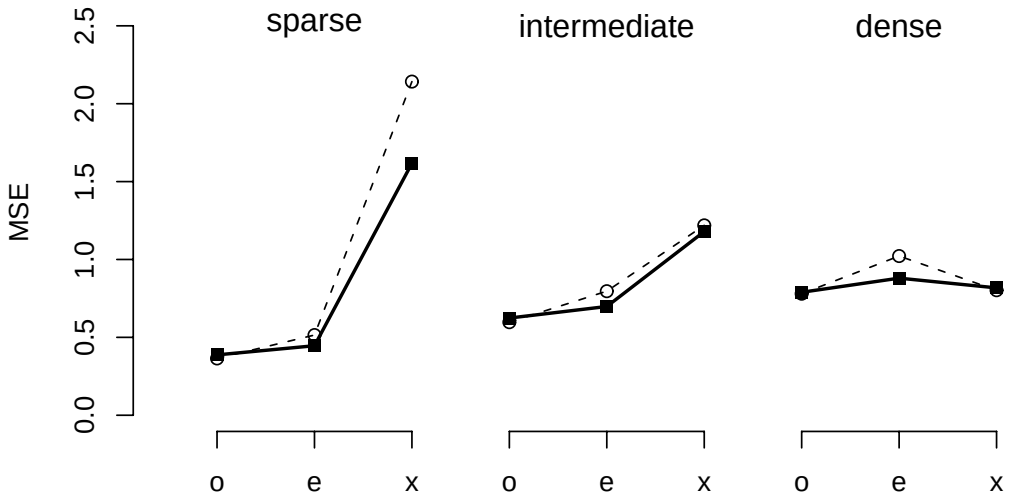

Figure 3.2: Estimate of Mean squared error ($\widehat{MSE}$) of $\hat{\beta}$ for LARS and its robust version under three different scenarios in the simulation study. In each scenario, o, e, and x indicate clean data, data with contaminated measurement errors, and data with mismeasured first covariate. The dotted lines are for LARS while the solid lines are for robust LARS. The points are the average $\widehat{MSE}$ for 100 replicates.

to closely resemble words on a number of linguistic dimensions. Two groups were studied – college students and older adults. The data consist of response times by word, averaged over the thirty subjects in each group. For each of word, a number of covariates was recorded. Goals of the experiment include determining which features of a word (i.e., covariates) affect response time, and whether the active features affect response time in the same fashion for college students and older adults. The authors make a case for the need to conduct and analyze studies with regression techniques in mind, rather than simpler ANOVA techniques.

Table 3.1: Difference in the number of selected variables for the fitted model to contaminated data from that to clean data

	LARS							robust LARS						
Scenario	-3	-2	-1	0	1	2	3	-3	-2	-1	0	1	2	3
ϵ contamination														
Sparse	5*	6	21	48	13	5	2*	1*	4	12	71	7	5	0
Intermediate	5	10	14	46	21	3	1	1	3	14	64	14	4	0
Dense	2	1	16	80	1	0	0	0	0	8	89	3	0	0
X contamination														
Sparse	7*	5	15	34	20	7	12*	5*	3	16	36	22	12	6
Intermediate	1*	5	13	55	21	3	2	1	3	18	50	23	4	1
Dense	0	0	5	93	2	0	0	0	0	4	94	2	0	0

NOTE: The entries with * are the cumulative counts of the specified case and more extreme cases.

Baayen (2007) conducts an extensive analysis of a slightly modified data set which is available in his `languageR` package. In his analysis, he creates and selects variables to include in a regression model, addresses issues of nonlinearity, collinearity and interaction, and removes selected cases as being influential and/or outlying. He trims a total of 87 of the 4568 cases. The resulting model, based on "typical" words, is used to address issues of linguistic importance. It includes seventeen basic covariates which enter the model as linear terms, a non-linear term for the written frequency of a word (fit as a restricted cubic spline with five knots), and an interaction term between the age group and the (nonlinear) written frequency of the word. We take his final model as an expert fit, and use it as a target to compare the performance of the robust LASSO to the LASSO.

We consider two sets of potential covariates for the model. The small set consists of Baayen's 17 basic covariates and three additional covariates representing a squared

term for written frequency and the interaction between age group and the linear and squared terms for written frequency. Age group has been coded as ± 1 for the interactions. The large set augments these covariates with nine additional covariates that were not included in Baayen's final model. Baayen excluded some of these covariates for a lack of significance, others because of collinearity.

The LASSO and robust LASSO were fit to all 4568 cases with the small and large sets of covariates. For the robust LASSO, the iterative algorithm in Section 2 was implemented by using LARS (Efron et al.; 2004) as the baseline modeling procedure and winsorizing residuals with λ_γ as a bending constant. The bending constant was taken to be scale invariant, so that $\lambda_\gamma = k \cdot \hat{\sigma}$, where k is a constant and $\hat{\sigma}$ is a robust scale estimate.

In all cases, the model was selected via the minimum C_p criterion. The estimated error standard deviations were approximately 0.0775 and 0.0789 for both the LASSO and the robust LASSO models with the small and large sets of covariates, respectively. A comparison of the models in terms of sum of squared deviations (SSD) from Baayen's fitted values, with the sum excluding those cases that he removed as outliers and/or influential cases, shows mixed results, with the robust LASSO having a smaller value of SSD for some values of k and larger SSD for other values. The comparison highlights the discontinuity of the fitted surface in k whenever a different model is selected by minimum C_p. In order to assess the relative performance of the two methods, we must average over a number of data sets.

To obtain the needed replicates and to examine the performance of the robust LASSO with a smaller sample size, we conducted a simulation study. For one replicate of the simulation, we sampled 400 cases from the data set and fit the LASSO

and robust LASSO. Models were selected with the minimum C_p criterion, and SSD computed on the full data set, removing the cases that Baayen did. A summary of SSD is presented in Figure 3.3. The figure also presents confidence intervals, based on $5,000$ replicates, for the robust LASSO. We see that the robust LASSO outperforms the LASSO for both the small and large covariate sets over a wide range of k.

In the simulation study, we also tracked the covariates in the selected model, with a summary displayed for $k = 1.6$ in Figure 3.4. Two key features of the coefficients are whether or not they are zero and how large they are. Several interesting features appear. First, there is little apparent difference between the 20 covariates in the small set (covariates 1 through 20) and the additional nine covariates in the large set (covariates 21 through 29). Both sets of covariates often have coefficients near 0, as indicated by the left-hand panel. The vertical bars, extending from the 10th to the 90th percentile for a given coefficient, show many effects that are, on the whole, small. The right-hand panel presents the fraction of replicates in which a given coefficient is non-zero in the model selected by the minimum C_p criterion. Again, we see little difference between the two sets of covariates. Overall, covariates in the small set had non-zero coefficients 0.73 of the time under the LASSO and 0.75 of the time under the robust LASSO. The remaining covariates had non-zero coefficients 0.68 and 0.70 of the time, respectively. The models selected by the robust LASSO averaged 21.4 covariates; those selected by the LASSO averaged 20.7 covariates. Second, in spite of the relatively high frequency with which covariates were "included" in the model, they often had very small coefficients, suggesting that there is substantial uncertainty about the "correct" model. Third, the LASSO and the robust LASSO provide coefficients of similar magnitude and with similar non-zero frequency. For

this data set and simulation, we expect this behavior, as the fraction of outlying data is small and the magnitude of the outliers is modest. Even in this difficult situation for a robust procedure, as Figure 3.3 shows, the robust LASSO can better match the expert fit. Fourth, examination of particular covariates demonstrates the appeal of regularization methods. Covariates 20 (`WrittenFrequency`) and 29 (`Familiarity`) address the same issue. See Balota et al. (2004) for a more complete description of the covariates. Both covariates appear in nearly all of the models for both the LASSO and the robust LASSO. Subjects are able to decide that a familiar word is a word more quickly (and more accurately) than an unfamiliar word, and so we see negative coefficients for both covariates. Although there seems to be no debate on whether this conceptual effect of similarity exists, there are a variety of viewpoints on how to best capture this effect. Regularization methods facilitate inclusion of a suite of covariates that address single conceptual effect. The robust LASSO retains this ability.

3.7 Conclusion

Case-specific parameters are incorporated in the LASSO with an ℓ_1 type penalty for the additional parameters. The role of the case-specific parameters is to down-weight outliers. Thus, the original LASSO is modified to produce a more robust estimator. Technically, the modification has a dual formulation. Under the dual formulation, the original squared error loss is changed to Huber's loss function. Since the magnitude of the modification can be depicted as the bending constant of Huber's loss function, the amount of regularization for case-specific parameters can be easily chosen. We demonstrate better performance of the modified LASSO than the LASSO

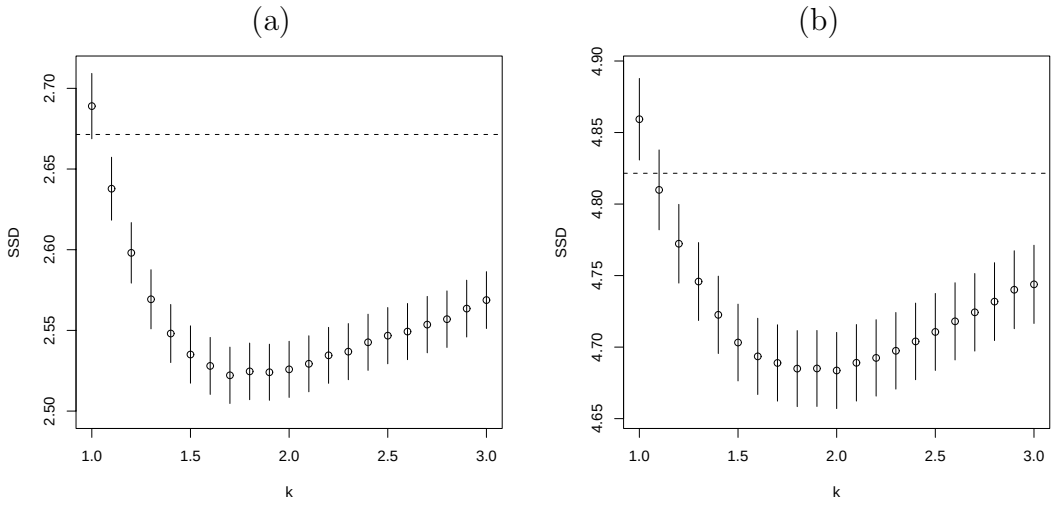

Figure 3.3: Sum of squared deviations (SSD) from Baayen's fits in the simulation study. The horizontal line is the mean SSD for the LASSO while the points represent the mean of SSDs for the robust LASSO. The vertical lines give approximate 95% confidence intervals for the mean SSDs. Panel (a) presents results for the small set of covariates and panel (b) presents results for the large set of covariates.

in an analysis of a real data set, and we also show the benefits of the modification through simulation studies.

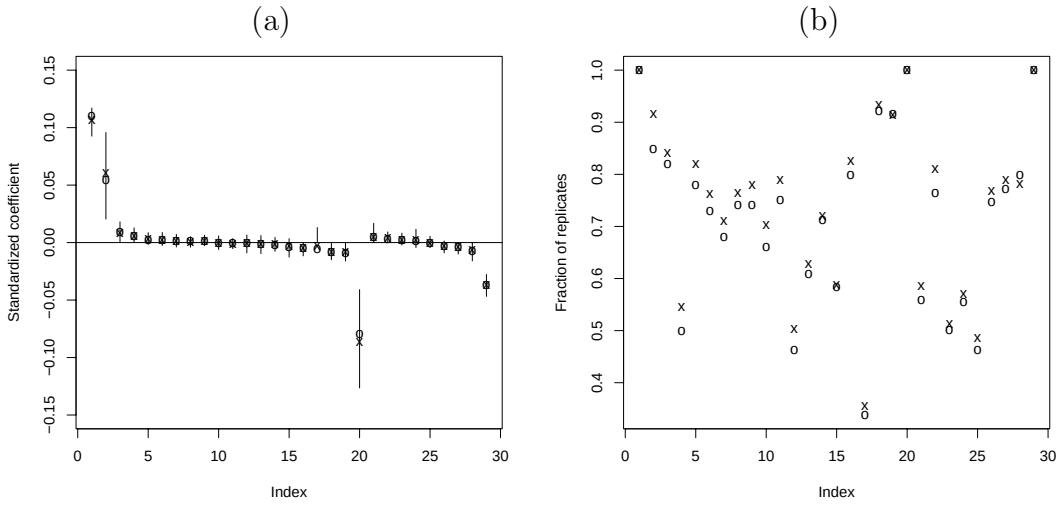

Figure 3.4: Coefficients in the simulation study. The first 20 variates are the small set of covariates, while the remaining 9 variates are in the large set of covariates. x is for the robust LASSO and o is for the LASSO. Panel (a) contains a plot of mean coefficients for standardized variates in the simulation. The vertical lines extend from the 10th to 90th percentiles of the coefficients under the robust LASSO. Panel (b) shows the fraction of replicates in which coefficients in the selected model are non-zero.

CHAPTER 4

EFFICIENT QUANTILE REGRESSION

4.1 Literature Review of Quantile Regression

Quantile regression has emerged as a useful tool for providing conditional quantiles of a response variable Y given values of a predictor X. This allows us to estimate not only the center, but also the upper or lower tail of the conditional distribution of interest. Due to the ability of quantile regression to capture the full distributional aspects, rather than only the conditional mean, quantile regression has been widely applied. Koenker and Bassett (1978) pioneered quantile regression and proved its consistency. Bassett and Koenker (1978) showed efficiency of median regression when the median is more efficient than the mean in a location model. To overcome the restriction of the *iid* linear model, He (1997) and Koenker and Zhao (1994) discuss heterogeneous error models. Nonparametric quantile regression also has been studied with a variety of kernel methods (Yu and Jones (1998), Takeuchi et al. (2006)), and with varying coefficients (Kim; 2007). Penalized quantile regression is considered in Shim et al. (2009) and Belloni and Chernozhukov (2009). Wu and Liu (2009) compared SCAD and LASSO type penalties and their asymptotic properties under quantile regression. Koenker et al. (1994), He et al. (1998) and Nychka et al. (1995)

considered quantile smoothing splines regression. In the following sections, a new form of quantile regression is proposed by employing case-specific parameters and their regularization.

4.2 Case Specific Parameters in Quantile Regression

To estimate qth regression quantile, $0 < q < 1$, the check loss function in Koenker and Bassett (1978) is employed. Quantile regression seeks a minimizer of the sum of asymmetrically weighted absolute errors with weight q on the positive errors and weight $(q-1)$ on the negative errors;

$$\rho(u) = \begin{cases} qu & \text{for } u \geq 0 \\ (q-1)u & \text{for } u < 0. \end{cases} \tag{4.1}$$

We first consider the following linear model of $y_i = x_i^\top \beta + u_i$, where u_i's are *iid* from some distribution with qth quantile equal to zero. Quantile regression finds the minimizer β of

$$L(\beta) = \sum_{i=1}^{n} \rho(y_i - x_i^\top \beta). \tag{4.2}$$

By introducing case-specific parameters γ_i, the linear model is changed to $y_i = x_i^\top \beta + \gamma_i + \epsilon_i$, and (4.2) is modified to be

$$L(\beta, \gamma) = \sum_{i=1}^{n} \rho(y_i - x_i^\top \beta - \gamma_i) + \frac{\lambda_\gamma}{2} J(\gamma), \tag{4.3}$$

where $J(\gamma)$ is a penalty for case-specific parameters. It can be verified easily that the ℓ_1 type penalty for $J(\gamma)$ is not appealing due to the piecewise linearity of the loss. For improving efficiency, an ℓ_2 type penalty for the γ is considered. Because of the asymmetry in the loss, except for $q = 1/2$, the size of reduction in the loss by case-specific parameter γ would depend on its sign. Given $\hat{\beta}$ and residual $r = y - x^\top \hat{\beta}$, if $r \geq 0$, then the positive γ would lower $\rho(r)$ by $q\gamma$, while if $r < 0$, the negative γ with

the same absolute value would lower the loss by $(q-1)\gamma$. This asymmetric impact on the loss is undesirable. Instead, a new penalty is introduced which leads to the same reduction in loss for positive and negative γ of the same magnitude. In other words, the desired ℓ_2 norm penalty needs to put $q\gamma_+$ and $(1-q)\gamma_-$ on an equal footing. This leads to the following penalty which is proportional to $q^2\gamma_+^2$ and $(1-q)^2\gamma_-^2$:

$$J_2(\gamma) := \{q/(1-q)\}\gamma_+^2 + \{(1-q)/q\}\gamma_-^2.$$

When $q = 1/2$, $J_2(\gamma)$ becomes the symmetric ℓ_2 norm. Together with case-specific parameters, additional penalty $J_2(\gamma)$, and given $\hat{\beta}$, the minimizer of (4.3), $\hat{\gamma}$ is explicitly given by

$$-\frac{q}{\lambda_\gamma} I\big(r < -\frac{q}{\lambda_\gamma}\big) + rI\big(-\frac{q}{\lambda_\gamma} \le r < \frac{1-q}{\lambda_\gamma}\big) + \frac{1-q}{\lambda_\gamma} I\big(r \ge \frac{1-q}{\lambda_\gamma}\big).$$

Then the effective loss has the form of

$$\rho^M(u) = \begin{cases} qu - \frac{q(1-q)}{2\lambda_\gamma} & \text{for} \quad \frac{1-q}{\lambda_\gamma} \le u \\ \frac{\lambda_\gamma}{2}\frac{q}{1-q}u^2 & \text{for} \quad 0 \le u < \frac{1-q}{\lambda_\gamma} \\ \frac{\lambda_\gamma}{2}\frac{1-q}{q}u^2 & \text{for} \quad -\frac{q}{\lambda_\gamma} \le u < 0 \\ (q-1)u - \frac{q(1-q)}{2\lambda_\gamma} & \text{for} \quad u < -\frac{q}{\lambda_\gamma} \end{cases} \tag{4.4}$$

Thus the modified quantile regression estimator is defined as

$$\hat{\beta}^M = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho^M(y_i - x_i^\top \beta) \tag{4.5}$$

Due to convexity of the loss function, finding minimizer of $\rho(\cdot)$ and $\rho^M(\cdot)$ is equivalent to finding zero of their derivatives; $\psi(\cdot)$ and $\psi^M(\cdot)$ where,

$$\psi^M(u) = \begin{cases} q & \text{for} \quad \frac{1-q}{\lambda_\gamma} \le u \\ \lambda_\gamma \frac{q}{1-q} u & \text{for} \quad 0 \le u < \frac{1-q}{\lambda_\gamma} \\ \lambda_\gamma \frac{1-q}{q} u & \text{for} \quad -\frac{q}{\lambda_\gamma} \le u < 0 \\ q-1 & \text{for} \quad u < -\frac{q}{\lambda_\gamma} \end{cases} \tag{4.6}$$

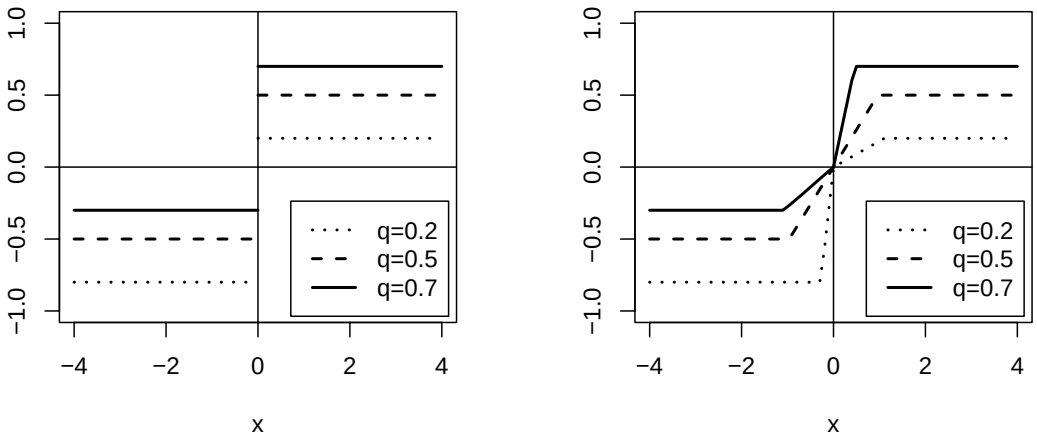

Figure 4.1: left: $\psi(x)$, right: $\psi^M(x)$ at $q = 0.2$, 0.5 and 0.7

Figure 4.1 illustrates shapes of $\psi(\cdot)$ and $\psi^M(\cdot)$ when estimating 0.2, 0.5 and 0.7 regression quantiles. Contrary to the $\psi(\cdot)$, $\psi^M(\cdot)$ is continuous at zero and has different slopes around zero. Thus modified loss function is now differentiable near the true quantile, which may produce unique solution. When estimating median regression coefficient, the two slopes become equivalent which reminds us of Huber's loss function (Huber; 1981). Newey and Powell (1987) and Efron (1991)'s asymmetric squared error loss ($|q-1(u<0)|\cdot u^2$, $q \in (0,1)$) provide another interesting connection, since our proposed method can be regarded as a hybrid of check loss and asymmetric squared error loss.

The interval of quadratic adjustment is $(-q/\lambda_\gamma, (1-q)/\lambda_\gamma)$, and we call the length of the interval of adjustment the "window width". The role of α is to determine the shrinking speed of the window width. When the λ_γ, and α are properly chosen,

the modified procedure will enjoy its advantage to the full. The following sections addresses how to set a good rule for selecting λ_γ and α.

4.3 Simulations for Finding a Rule

To develop a rule and obtain a consistent estimator, we first consider λ_γ of the form $\lambda_\gamma := c_q n^\alpha / \hat{\sigma}$ where c_q is a constant depending on q, n is the sample size, α is a positive constant, and $\hat{\sigma}$ is a robust scale estimate of the error distribution. For optimal finite sample performance, we will consider a range of α values. We use 1.4826·MAD (Median Absolute Deviation) as a robust scale estimator $\hat{\sigma}$. The form of the rule suggests that c_q should be scale invariant and depend only on the targeted quantile q.

In this section, choice of the window width will be investigated by simulation. Throughout the simulation, the linear model $y_i = \beta_0 + x_i^\top \beta + \epsilon_i$ is assumed. Following the simulation setting in Tibshirani (1996), $x^\top = (x_1, ..., x_8)$ is generated from a multivariate normal distribution with mean (0, ..., 0) and variance Σ, where $\sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.5$. The true coefficient vector β is taken to be (3, 1.5, 0, 0, 2, 0, 0, 0). Various distributions are considered for ϵ_i, including normal, t, shifted log-normal, shifted gamma, and shifted exponential error distribution. In each distribution, ϵ_i is assumed to be *iid* with median zero and variance 9 (except when the ϵ_i follows the standard normal distribution). For the t distributions, 2.25, 5, and 10 degrees of freedom are used, maintaining a variance of 9.

Several values of α were tried. After examining the results, a decision was made to set α equal to 0.3. This makes α to be independent of sample size. Thus we search only for c_q. Sample sizes range from 10^2 to 10^4, and various quantiles from 0.1 to 0.9

are considered. To gauge the performance of ℓ_2 adjusted quantile regression with λ_γ, define mean squared error (MSE) of the estimated quantile $X^\top\hat{\beta}+\hat{\beta}_0$ at a new X as

$$\begin{aligned} MSE &= E^{\hat{\beta},X}||(X^\top\hat{\beta}+\hat{\beta}_0)-(X^\top\beta+\beta_0)||^2 \\ &= E^{\hat{\beta},X}\{(\hat{\beta}-\beta)^\top X^\top X(\hat{\beta}-\beta)+(\hat{\beta}_0-\beta_0)^2\} \\ &= E^{\hat{\beta}}\{(\hat{\beta}-\beta)^\top\Sigma(\hat{\beta}-\beta)+(\hat{\beta}_0-\beta_0)^2\}. \end{aligned} \tag{4.7}$$

MSE is integrated across the distribution of a future X. The distribution of the future X is normal with mean (0, ..., 0) and variance Σ. In the simulation, MSE is approximated by a Monte Carlo estimate over 500 replicates, $\widehat{MSE} = 500^{-1}\sum_{i=1}^{500}((\hat{\beta}^i-\beta)^\top\Sigma(\hat{\beta}^i-\beta)+(\hat{\beta}_0^i-\beta_0)^2)$, where $\hat{\beta}^i$ and $\hat{\beta}_0^i$ are the estimates of β and the intercept β_0 for the i^{th} replicate, respectively. With fixed α, the window width $(\hat{\sigma}/(c_q n^\alpha))$ is a function of the constant c_q only. Thus by varying c_q, an 'optimal' window width which provides the smallest MSE can be obtained. The optimal window widths, found by a grid search, are shown in Figure 4.2 for various error distributions.

Each panel of Figure 4.3 shows a typical shape of the MSE curve as a function of window width. In general, MSE values begin to decrease as we increase the window width from zero until it hits its minimum, and increase thereafter due to an increase in bias. However, when estimating the median with normally distributed errors, MSE decreases as the window width increases. This is not surprising, given the optimality properties of least squares regression for normal theory regression. The comparisons between sample mean and sample median can be explicitly found under the t error distributions with different degrees of freedom. The benefit of the median relative to the mean is greater for thicker tailed distributions. We observe that this qualitative behavior carries over to the optimal window width. Thicker tails lead to shorter optimal windows, as shown in Figure 4.2.

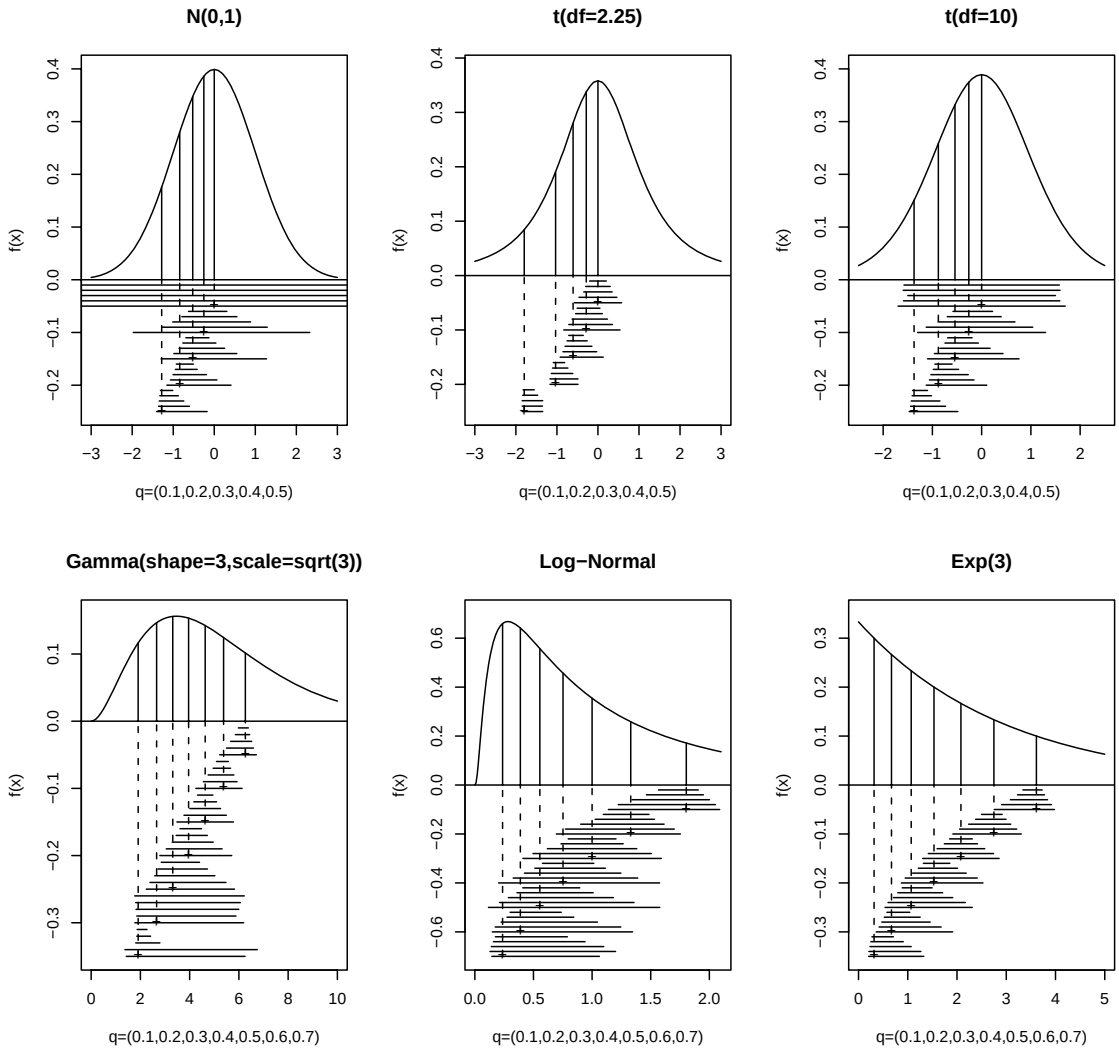

Figure 4.2: 'Optimal' intervals of adjustment for different quantiles (q), sample sizes (n) and error distributions. The vertical lines in each distribution indicate the true quantiles. The stacked horizontal lines for each quantile are corresponding optimal intervals. The five intervals at each quantile are for n= 10^2, $10^{2.5}$, 10^3, $10^{3.5}$ and 10^4.

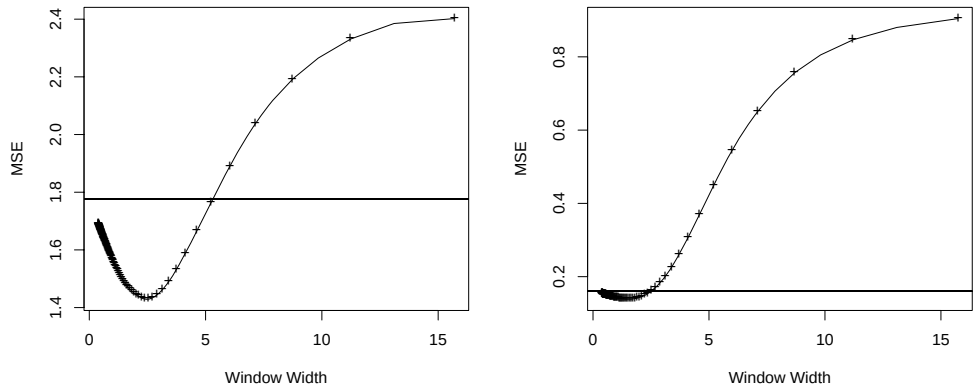

Figure 4.3: $\widehat{MSE}$ values evaluated at one hundred points marked with '+' and connected by a smoothing spline. The smallest and largest window widths in each plot correspond to the window width approximately 5% and 98% of data in it, respectively. The residual distribution is the t (df=10) distribution, sample sizes are 10^2 (left panel) and 10^3 (right panel), and the 0.2 quantile is estimated. The horizontal lines represent the $\widehat{MSE}$ values from the standard quantile regression.

4.4 Development of a Rule

Under each error distribution mentioned above, the 'optimal' constants which yield smallest $\widehat{MSE}$ are found at the quantiles 0.1, 0.2, ..., 0.9. First, omitting the median, log of the optimal constant $log(c_q)$ from the standard normal error is regressed on q to suggest a possible relationship. A significant linear relationship exists. The fitted values from this regression were used to produce values for c_q. These values were then applied to the other error distributions. However, the rule obtained from the normal distribution led to poor $\widehat{MSE}$ values when applied to skewed error distributions. This is due to the overestimation of the window width or equivalently, underestimation of c_q near the median. As we can see in Figure 4.3, too large a window may lead to a huge MSE.

As an alternative, another rule expressing the relationship between the optimal $\log(c_q)$ and q was developed from the exponential error distribution. The top left plot in Figure 4.4 shows the relationship between optimal $log(c_q)$ and q. Before fitting a linear model of $log(c_q) = \beta_0 + \beta_1 q + \epsilon$, q greater than 0.5 were converted to $1-q$, since it was judged desirable to have a rule which will work well for symmetric distributions. The solid line in the top right plot of Figure 4.4 is the fitted line using all observations, whereas the dashed line is from only observations with $q \geq 0.5$, excluding observations with + mark. The dashed line is accepted as a final rule.

The final rule is compared to the other rules from normal, t, log-normal, and gamma distributions. In Figure 4.4, the solid lines in the second and third rows represent 'optimal' rules from each distribution mentioned above (developed on quantiles ≥ 0.5) whereas the dashed line is the final rule. Numerical expression of the final rule is given by

$$c_q \approx \begin{cases} 0.5e^{-2.118-1.097q} & \text{for } q < 0.5 \\ 0.5e^{-2.118-1.097(1-q)} & \text{for } q \geq 0.5, \end{cases} \tag{4.8}$$

where q stands for the qth quantile. Under various error distributions, the estimated c_q from the rule (4.8) is employed to gauge its prediction performance. Specifically, $\widehat{MSE}$ values for quantile regression (QR), modified quantile regression with optimal c_q (OPT), and modified quantile regression with c_q chosen by the final rule (QR.M) are compared. Figures 4.5 through 4.11 show the behavior of QR, OPT, and QR.M in terms of $\widehat{MSE}$. Overall, QR.M handily outperforms standard quantile regression. Surprisingly enough, the version of finite sample performance for this modified quantile regression is often nearly optimal. This near-optimality extends across a range of residual distributions.

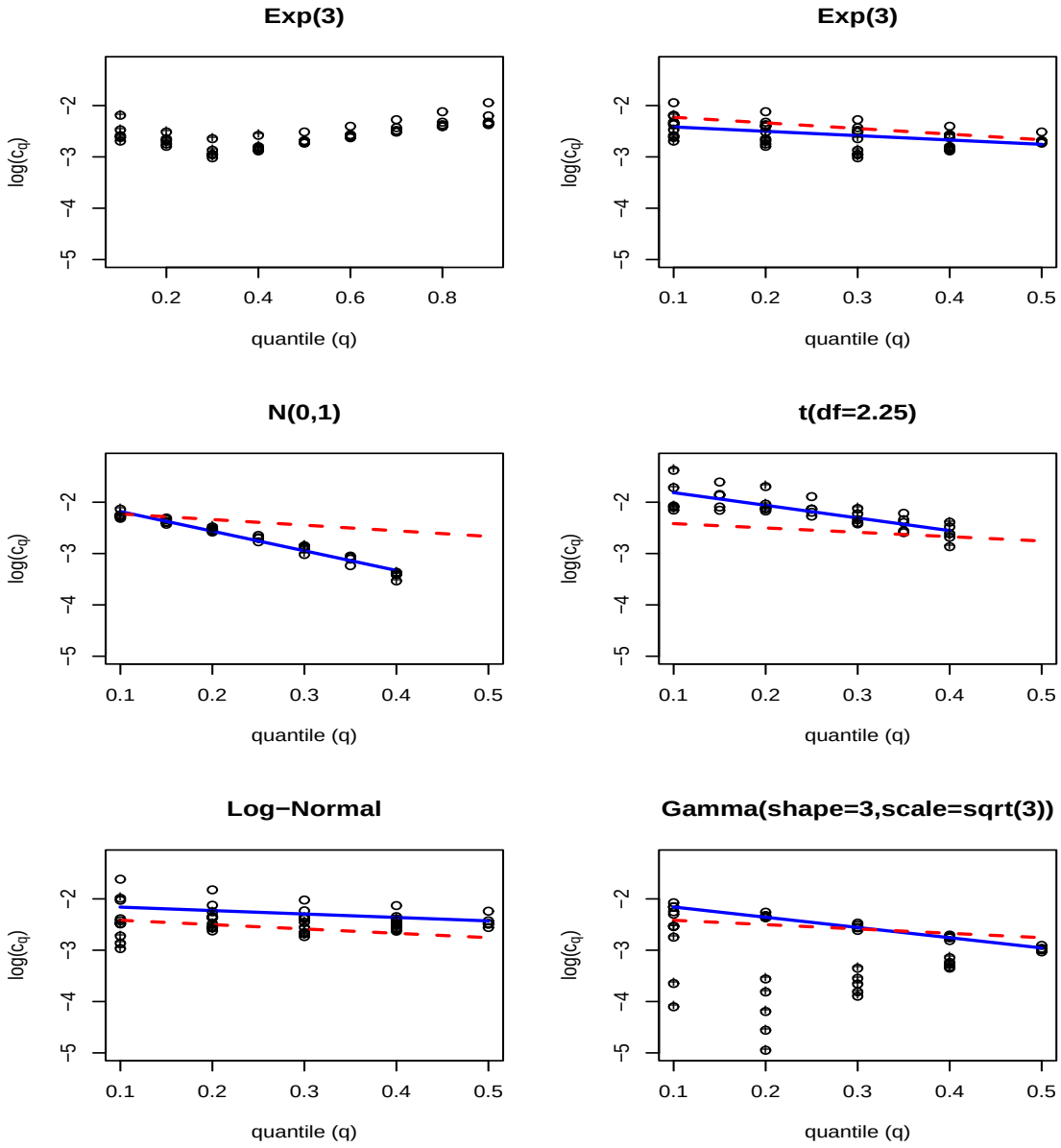

Figure 4.4: Top left: Relationship between optimal $\log(c_q)$ and quantile from the exponential distribution. Top right: Left plot is folded in half at $q = 0.5$. Circles with a + mark are from the left fold (quantile < 0.5) and the others are from the right fold (quantiles ≥ 0.5). The solid line is the fitted line using all observations whereas the dashed line excludes observations with a + mark (final rule). Solid lines in the middle and bottom plots are the rules corresponding to normal, t, log-normal, and gamma distributions compared to the final rule (dashed line).

In practice, the robust linear modeling procedure, `rlm(MASS)` in R package is ready to be utilized. Equipped with (4.6), the modified estimators can be obtained from the `rlm` function by specifying q and the corresponding rule c_q. Since the `rlm` function internally uses re-scaled MAD for the method of scale estimation, the estimate of the scale parameter in λ_γ is automatically obtained.

4.5 Asymptotic Properties

The modified quantile regression can be a consistent estimator when the window width $(1/\lambda_\gamma)$ decreases as sample size increases. In this section, we provide asymptotic properties of the modified procedures and conditions of λ_γ.

4.5.1 Asymptotic Properties of Quantile Estimator

The modified ψ function for the qth quantile is given by

$$\psi^M(u) = \begin{cases} q & \text{for} \quad \frac{1-q}{\lambda_\gamma} \le u \\ \lambda_\gamma \frac{q}{1-q} u & \text{for} \quad 0 \le u < \frac{1-q}{\lambda_\gamma} \\ \lambda_\gamma \frac{1-q}{q} u & \text{for} \quad -\frac{q}{\lambda_\gamma} \le u < 0 \\ q-1 & \text{for} \quad u < -\frac{q}{\lambda_\gamma} \end{cases}$$

Define $\hat{\theta}_n^\gamma$ to be a zero of $\sum_{i=1}^n \psi_q^\gamma(x_i - \theta)$. The following proposition shows that the modified quantile estimator, $\hat{\theta}_n^\gamma$, is asymptotically equivalent to the ordinary estimator, $\hat{\theta}_n$, of the qth quantile of a distribution under some mild conditions. Let 0 be the unique qth quantile of the distribution of X, that is, $F_X(0) = q$, $F_X(-\epsilon) < q$, and $F_X(\epsilon) > q$ for any $\epsilon > 0$.

Proposition 4 *Assume that $X_1, X_2, \ldots$ form an i.i.d. sequence from F_X. Under the following assumptions,*

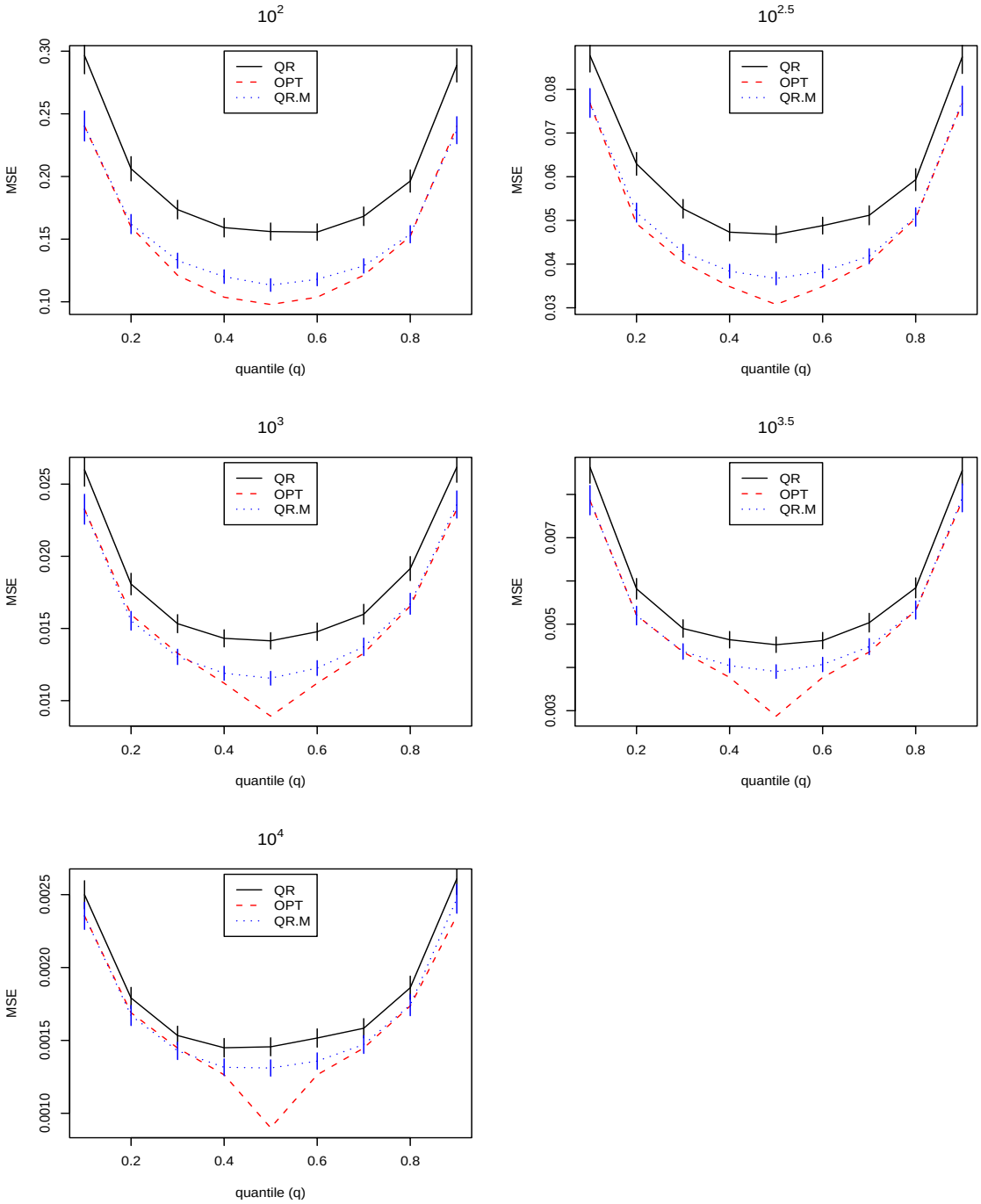

Figure 4.5: $\widehat{MSE}$ values from quantile regression (QR), modified quantile regression with optimal window width (OPT), and modified quantile regression using the rule (QR.M) under a standard normal error distribution.

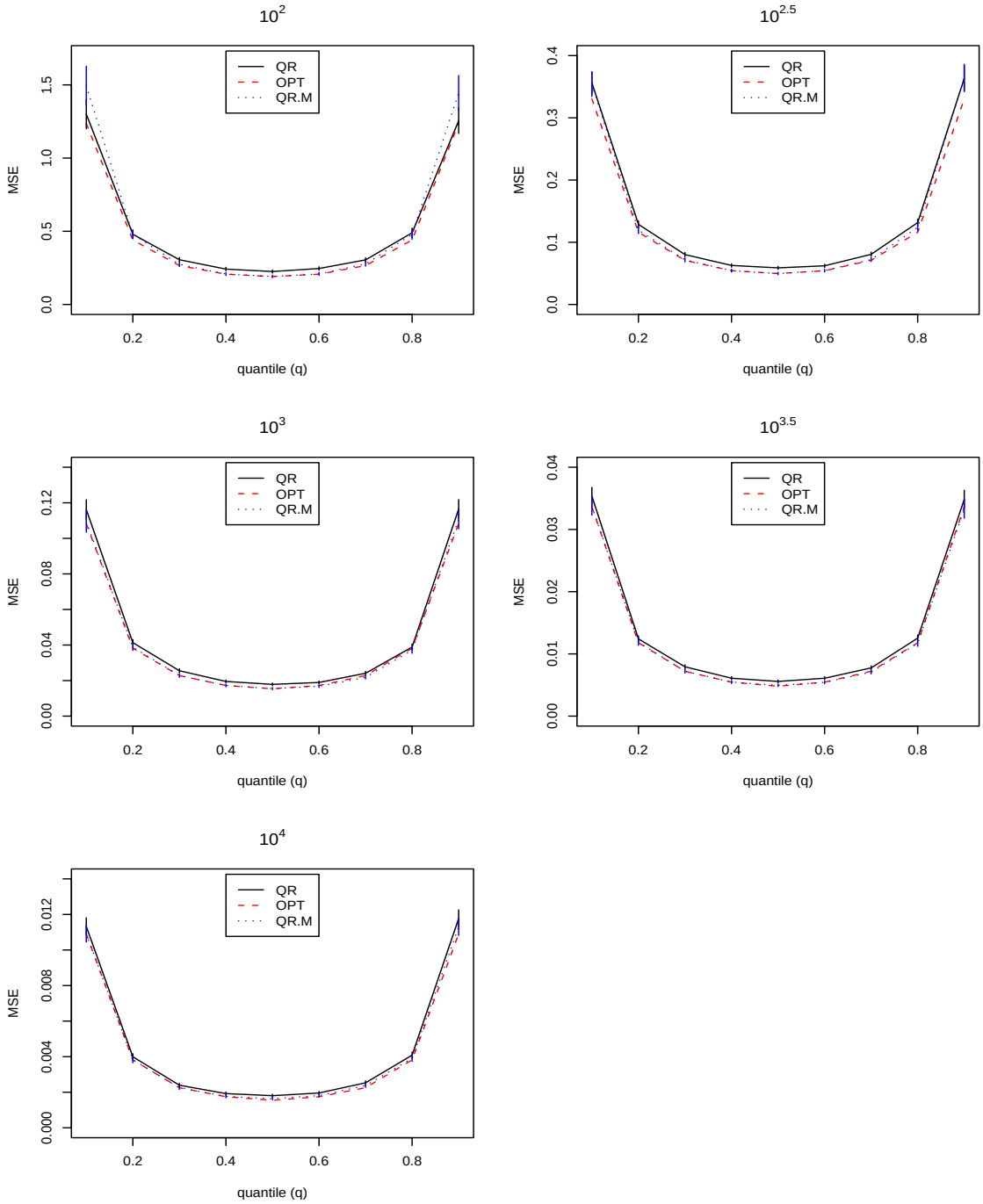

Figure 4.6: $\widehat{MSE}$ values from quantile regression (QR), modified quantile regression with optimal window width (OPT), and modified quantile regression using the rule (QR.M) under t (df=2.25) error distribution.

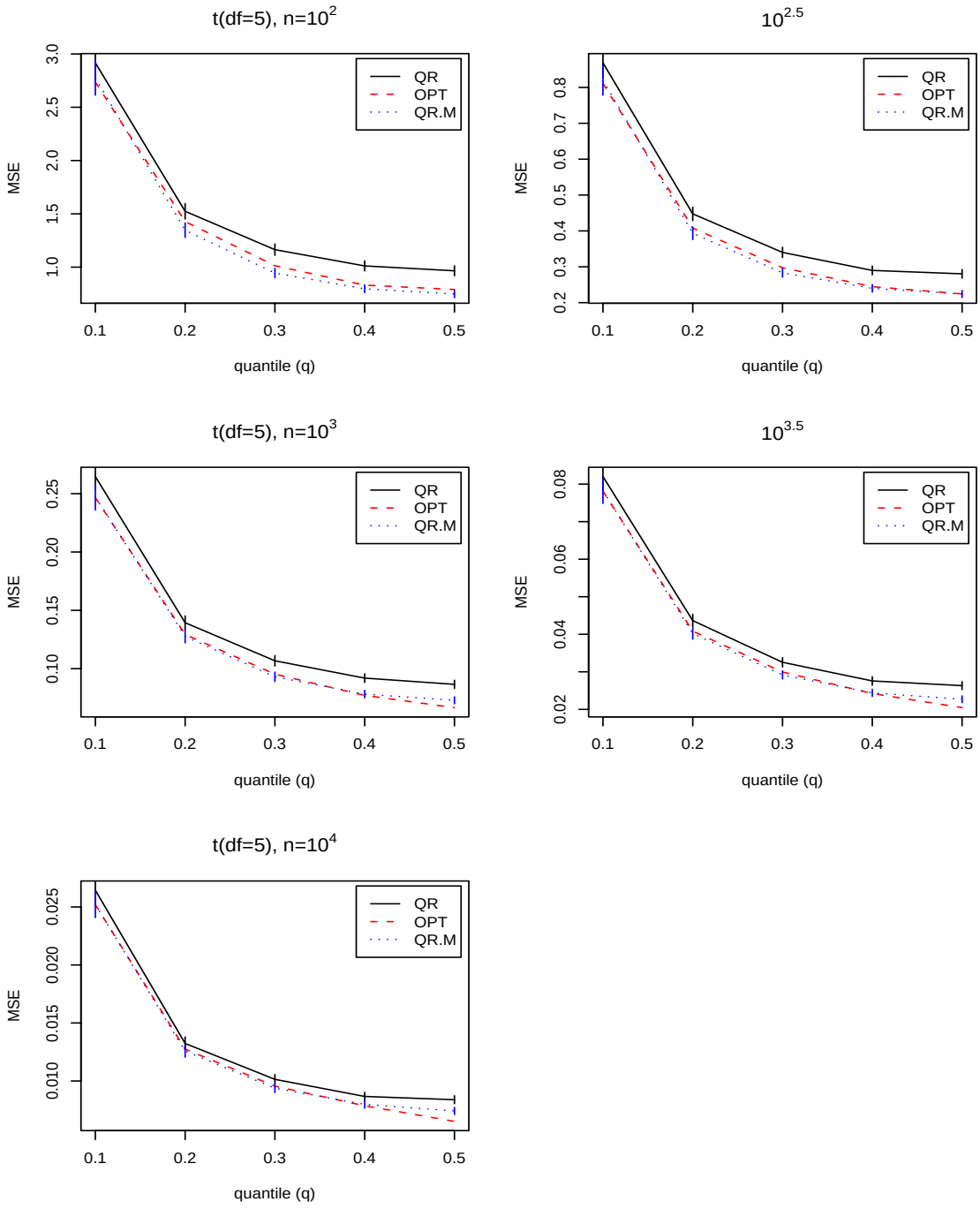

Figure 4.7: $\widehat{MSE}$ values from quantile regression (QR), modified quantile regression with optimal window width (OPT), and modified quantile regression using the rule (QR.M) under t (df=5) error distribution.

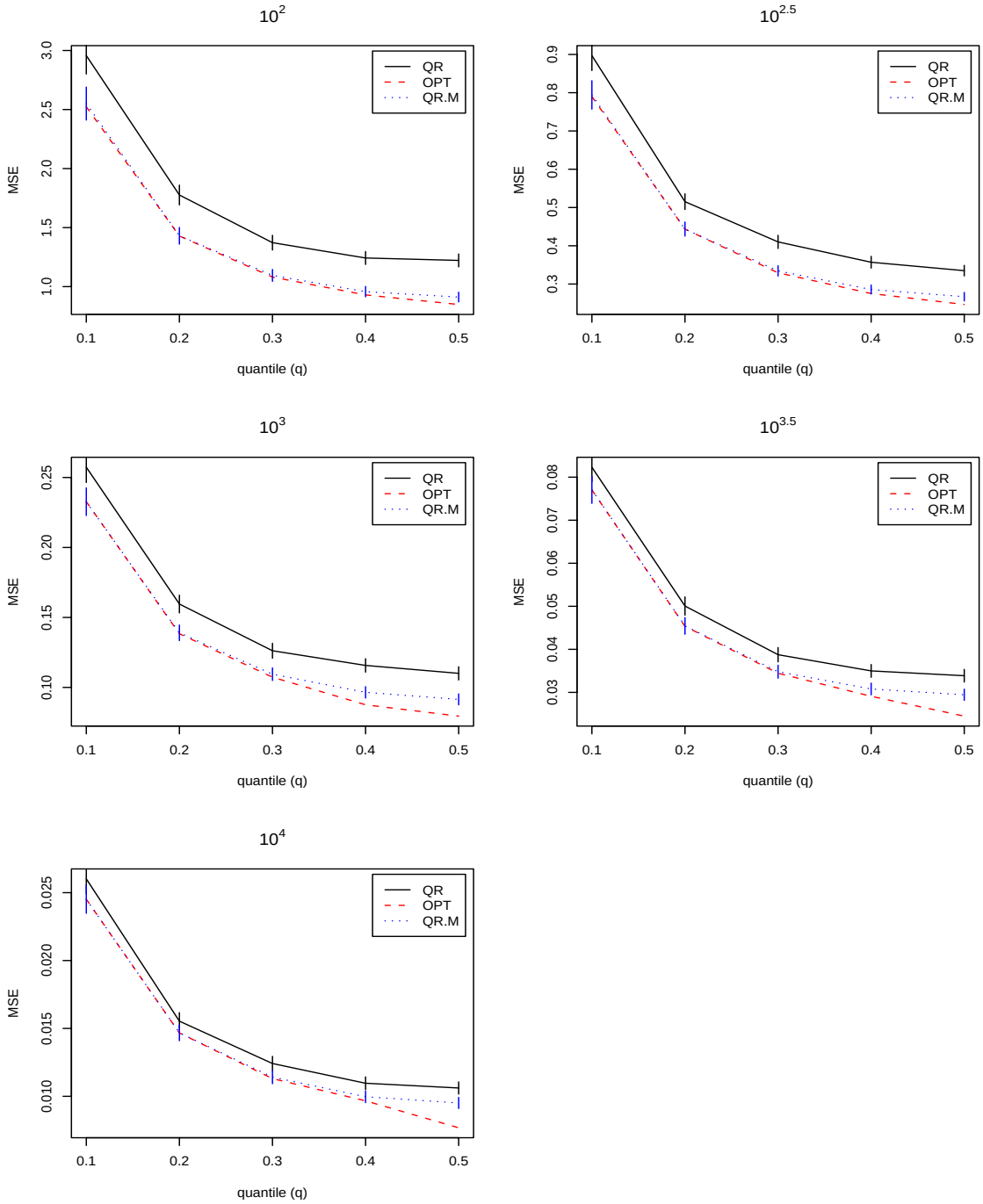

Figure 4.8: $\widehat{MSE}$ values from quantile regression (QR), modified quantile regression with optimal window width (OPT), and modified quantile regression using the rule (QR.M) under a t (df=10) error distribution.

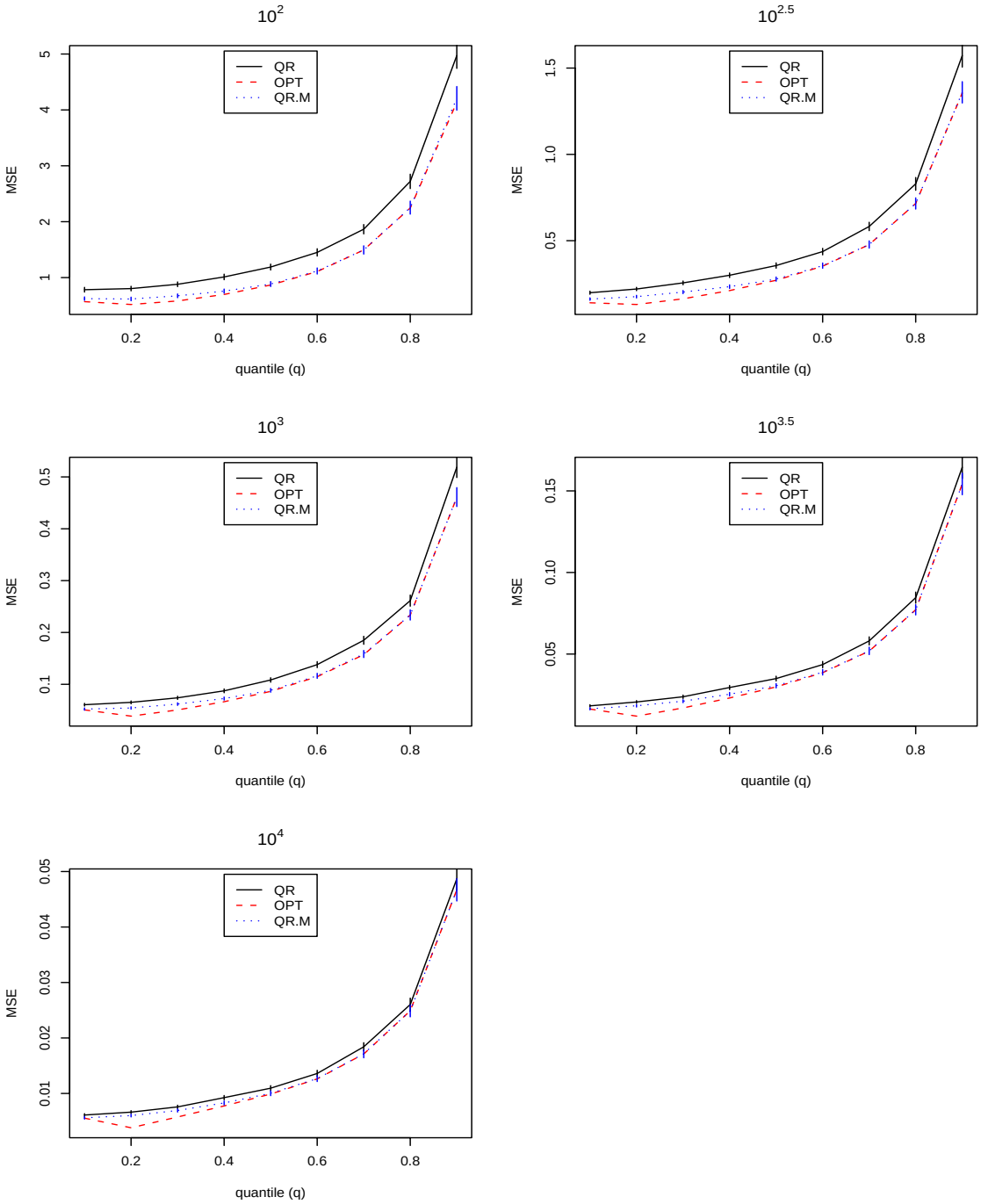

Figure 4.9: $\widehat{MSE}$ values from quantile regression (QR), modified quantile regression with optimal window width (OPT), and modified quantile regression using the rule (QR.M) under a gamma $(3,\sqrt{3})$ error distribution.

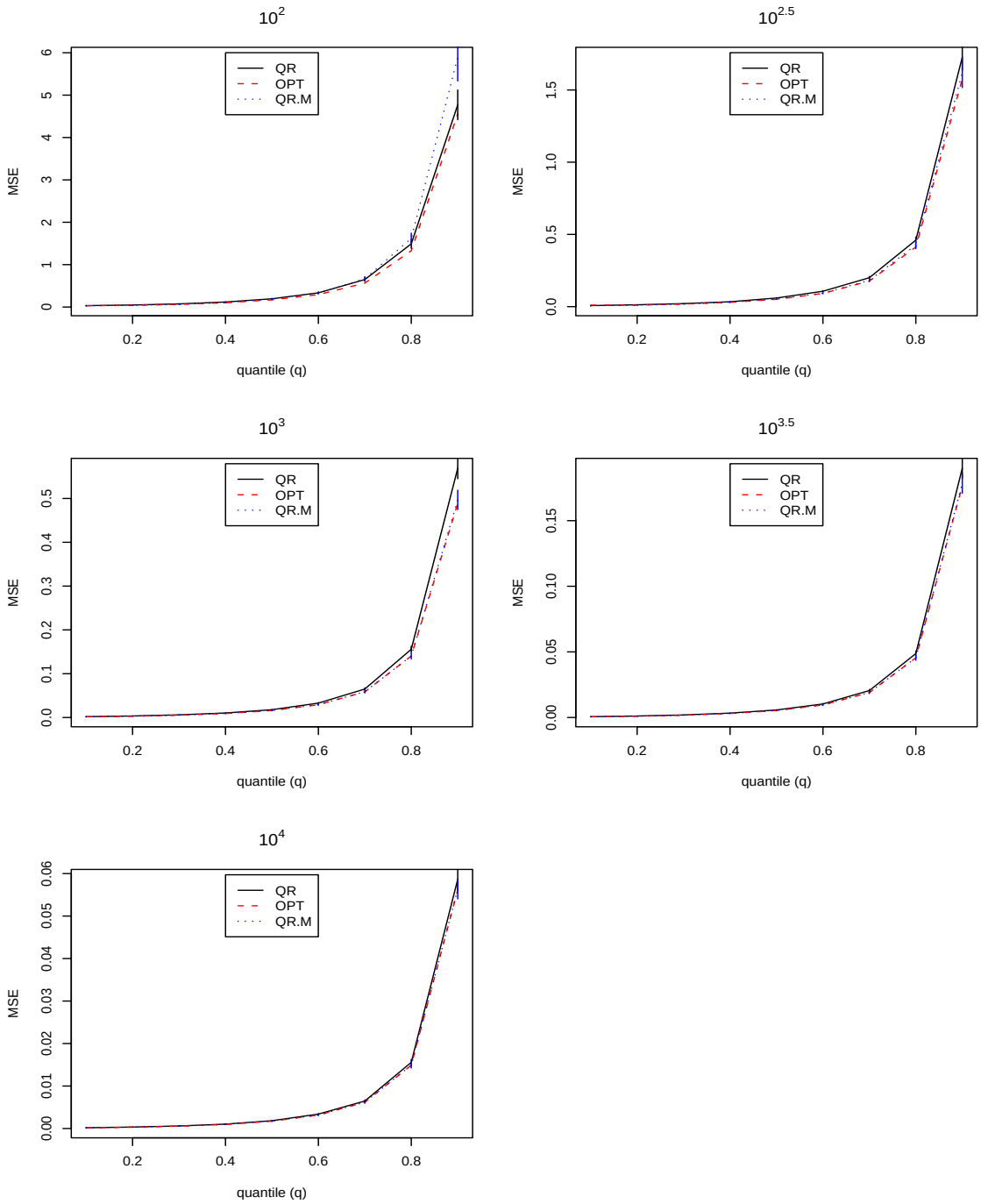

Figure 4.10: $\widehat{MSE}$ values from quantile regression (QR), modified quantile regression with optimal window width (OPT), and modified quantile regression using the rule (QR.M) under a log-normal error distribution.

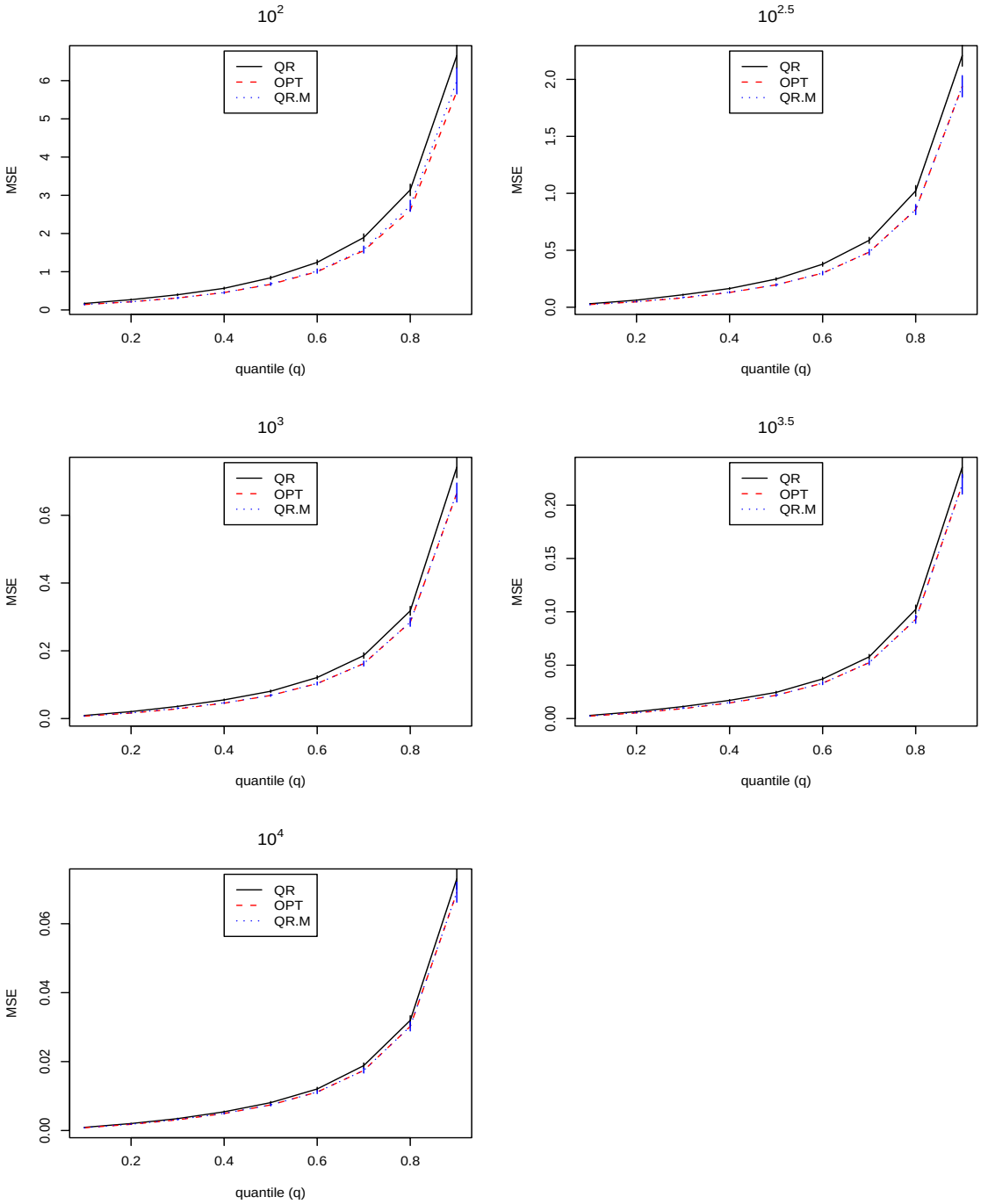

Figure 4.11: $\widehat{MSE}$ values from quantile regression (QR), modified quantile regression with optimal window width (OPT), and modified quantile regression using the rule (QR.M) under an exponential (3) error distribution.

A1) the probability density function of X, $f(x)$, has positive density at 0 and $f(x) = f(0) + f'(0)x + o(x)$ *around 0, and*

A2) $\lambda_\gamma = n^\alpha$ *for* $1/4 < \alpha \leq 1/2$,

$$\sqrt{n}(\hat{\theta}_n^\gamma - \hat{\theta}_n) \xrightarrow{p} 0.$$

Proof. Let $\hat{\theta}_n$ be an ordinary quantile estimator, i.e.,

$$\hat{\theta}_n \in [\inf\{\theta | \sum_{i=1}^n \psi_q(x_i - \theta) \leq 0\}, \sup\{\theta | \sum_{i=1}^n \psi_q(x_i - \theta) \geq 0\}].$$

Typically, this interval will be degenerate, and the sample quantile will be uniquely defined as $\hat{\theta} = \{X_{(k)} | \frac{k-1}{n} < q < \frac{k}{n}\}$. If $q = k/n$ exactly, for some integer k, $\hat{\theta}_n \in [X_{(k)}, X_{(k+1)}]$. In any event, $\hat{\theta}_n$ is a near zero of $\sum_{i=1}^n \psi_q(x_i - \theta)$, with $|\sum_{i=1}^n \psi_q(x_i - \theta)| < 1$. Under condition A1), all definitions of $\hat{\theta}_n$ are asymptotically equivalent (Wretman; 1978).

First note that $\psi_q^\gamma(x-\theta)$ is a weakly decreasing functions of θ for fixed x. So, if $\hat{\theta}_n^\gamma > \hat{\theta}_n + \delta_n$ for $\delta_n > 0$, then $\sum_{i=1}^n \psi_q^\gamma(x_i - (\hat{\theta}_n + \delta_n)) \geq \sum_{i=1}^n \psi_q^\gamma(x_i - (\hat{\theta}_n^\gamma)) \geq 0$. Likewise, if $\hat{\theta}_n^\gamma < \hat{\theta}_n - \delta_n$ for $\delta_n > 0$, then $\sum_{i=1}^n \psi_q^\gamma(x_i - (\hat{\theta}_n - \delta_n)) \leq 0$. It will be shown that the closely related $P(\sum_{i=1}^n \psi_q^\gamma(X_i - (\hat{\theta}_n + \delta_n)) > 0)$ and $P(\sum_{i=1}^n \psi_q^\gamma(X_i - (\hat{\theta}_n - \delta_n)) < 0)$ converge to zero for δ_n such that $n^{1/2}\delta_n \to 0$ and $n^{2\alpha}\delta_n \to \infty$. This leads to the conclusion that $P(|\hat{\theta}_n^\gamma - \hat{\theta}_n| \leq \delta_n) \to 1$ as $n \to \infty$, and thus $\sqrt{n}(\hat{\theta}_n^\gamma - \hat{\theta}_n) \xrightarrow{p} 0$.

To show the convergence for $\hat{\theta} + \delta_n$, we work with $\hat{\theta}_n = X_{(k)}$, and we define $S_n(\psi_q^\gamma) := \sum_{i=1}^n \psi_q^\gamma(X_i - (\hat{\theta}_n + \delta_n))$. The argument hinges on the following conditional probability, given $\hat{\theta}_n$.

$$\begin{aligned} P(S_n(\psi_q^\gamma) > 0 \mid \hat{\theta}_n) &= P\big(S_n(\psi_q^\gamma) - E(S_n(\psi_q^\gamma)|\hat{\theta}_n) > -E(S_n(\psi_q^\gamma)|\hat{\theta}_n) \;\big|\; \hat{\theta}_n\big) \\ &\leq Var(S_n(\psi_q^\gamma)|\hat{\theta}_n)/\{E(S_n(\psi_q^\gamma)|\hat{\theta}_n)\}^2. \end{aligned}$$

The last inequality is obtained by application of Chebyshev's inequality, assuming that $E(S_n(\psi_q^\gamma)|\hat{\theta}_n) < 0$. We will show that $P(E(S_n(\psi_q^\gamma)|\hat{\theta}_n) < 0)$ tends to one.

Decompose the sum, $S_n(\psi_q^\gamma)$, into three parts

$$\begin{aligned} S_n(\psi_q^\gamma) &= \sum_{i=1}^{n} \psi_q^\gamma(X_{(i)} - (\hat{\theta}_n + \delta_n)) \\ &= \sum_{i=1}^{k-1} \psi_q^\gamma(X_{(i)} - (\hat{\theta}_n + \delta_n)) + \psi_q^\gamma(-\delta_n) + \sum_{i=k+1}^{n} \psi_q^\gamma(X_{(i)} - (\hat{\theta}_n + \delta_n)). \end{aligned}$$

For computation of $E(S_n(\psi_q^\gamma)|\hat{\theta}_n)$, observe that the first term, given $\hat{\theta}_n$, can be taken as $\sum_{i=1}^{k-1} \psi_q^\gamma(Z_i - (\hat{\theta}_n + \delta_n))$, with i.i.d. Z_i $(i = 1, \ldots, k-1)$, where Z_i follows the distribution of X restricted to $(-\infty, \hat{\theta}_n)$. Similarly, the last term, given $\hat{\theta}_n$, can be re-expressed as $\sum_{i=k+1}^{n} \psi_q^\gamma(Z_i - (\hat{\theta}_n + \delta_n))$, with Z_i $(i = k+1, \ldots, n)$ having the same distribution as X restricted to $(\hat{\theta}_n, \infty)$, and independent of $Z_1, \ldots, Z_{k-1}$. Then,

$$E(S_n(\psi_q^\gamma)|\hat{\theta}_n) = (k-1)E\{\psi_q^\gamma(Z_1-(\hat{\theta}_n+\delta_n))|\hat{\theta}_n\}+\psi_q^\gamma(-\delta_n)+(n-k)E\{\psi_q^\gamma(Z_n-(\hat{\theta}_n+\delta_n))|\hat{\theta}_n\},$$

Under the assumptions A1) and A2), and making use of the fact that $\hat{\theta}_n \overset{a.s.}{\to} 0$, it can be shown that, for almost all sequences of $\hat{\theta}_n$,

$$(1/n)E(S_n(\psi_q^\gamma)|\hat{\theta}_n) = -f(0)\delta_n - f'(0)q(1-q)/(24\lambda_\gamma^2) + o_p(\delta_n).$$

The same decomposition of $S_n(\psi_q^\gamma)$ can be used to derive an expression for the conditional variance:

$$\begin{aligned} (1/n^2)Var(S_n(\psi_q^\gamma)|\hat{\theta}_n) &\leq (k-1)/n^2 E[\{\psi_q^\gamma(Z_1 - (\hat{\theta}_n + \delta_n)) - (q-1)\}^2|\hat{\theta}_n] \\ &\quad +(n-k)/n^2 E[\{\psi_q^\gamma(Z_n - (\hat{\theta}_n + \delta_n)) - q\}^2|\hat{\theta}_n]. \end{aligned}$$

Considering sequences for which $\hat{\theta}_n \to 0$,

$$(1/n^2)Var(S_n(\psi_q^\gamma)|\hat{\theta}_n) = q(1-q)f(0)/(6n^{1+\alpha}) + o_p(n^{-1-\alpha}).$$

Combining the expressions for the conditional mean and variance, we have

$$\begin{aligned} & E(S_n(\psi_q^\gamma)|\hat{\theta}_n)/\{Var(S_n(\psi_q^\gamma)|\hat{\theta}_n)\}^{1/2} \\ \leq & \frac{-f(0)\delta_n - f'(0)q(1-q)/(24\lambda_\gamma^2) + o_p(\delta_n)}{[q(1-q)f(0)/(6n^{1+\alpha}) + o_p(n^{-1-\alpha})]^{1/2}}. \end{aligned}$$

For $\delta_n = n^{-2\alpha+\epsilon}$, with $\epsilon > 0$ satisfying $(3\alpha-1)/2 < \epsilon < (2\alpha - 1/2)$, the upper bound diverges to $-\infty$. Thus, we have shown that $P(S_n(\psi_q^\gamma) > 0|\hat{\theta}_n) \stackrel{a.s.}{\to} 0$, and hence that $P(S_n(\psi_q^\gamma) > 0) \to 0$ by the dominated convergence theorem.

Similarly, with the changes that $\hat{\theta}_n = X_{(k+1)}$ and $(1/n)E\{\sum_{i=1}^n \psi_q^\gamma(X_i - (\hat{\theta}_n - \delta_n))|\hat{\theta}_n\} = f(0)\delta_n - f'(0)q(1-q)/(24\lambda_\gamma^2) + o(\delta_n)$, we can show that $P(\sum_{i=1}^n \psi_q^\gamma(X_i - (\hat{\theta}_n - \delta_n)) < 0)$ tends to zero.

□

Remark For $\lambda_\gamma = n^\alpha$ with $\alpha > 1/2$, convergence of $\sqrt{n}(\hat{\theta}_n^\gamma - \hat{\theta}_n)$ to zero in probability can be established easily.

4.5.2 Asymptotic Properties Under *Independent* Errors

Allowing a potentially different error distribution for each observation, let $Y_1, Y_2, \ldots$ be independent random variables with cdf $F_1, F_2, \ldots$ and suppose that each F_i has continuous pdf f_i. Assume that the qth conditional quantile function of Y given x is linear in x and given by $x^\top\beta(q)$, and let $\xi_i(q) := x_i^\top\beta(q)$. First consider the following regularity conditions:

(C-1) $f_i(\xi)$, $i = 1, 2, \ldots$, are uniformly bounded away from 0 and ∞ at ξ_i.

(C-2) $f_i(\xi)$, $i = 1, 2, \ldots$, admit a first-order Taylor expansion at ξ_i, and $f_i'(\xi)$ are uniformly bounded at ξ_i.

(C-3) There exists a positive definite matrix D_0 such that $\lim_{n\to\infty} n^{-1}\sum x_i x_i^\top = D_0$.

(C-4) There exists a positive definite matrix D_1 such that $\lim_{n\to\infty} n^{-1}\sum f_i(\xi_i)x_i x_i^\top = D_1$.

(C-5) $\max_{i=1,\ldots,n} \|x_i\|/\sqrt{n} \to 0$ in probability.

(C-1) and (C-3) through (C-5) are the conditions considered for the limiting distribution of the standard regression quantile estimator $\hat{\beta}_n$ in Koenker (2005), while (C-2) is an additional assumption that we make.

Koenker (2005) showed that the limiting behavior of $\sum_{i=1}^n(\rho(u_i - x_i^\top\delta/\sqrt{n}) - \rho(u_i)) \equiv Z_n(\delta)$ determines the limiting distribution of $\hat{\delta}_n = \sqrt{n}(\hat{\beta}_n - \beta)$, where $\hat{\delta}_n$ minimizes $Z_n(\delta)$ and $u_i = y_i - x_i^\top\beta(q)$. Thus we will consider the limiting behavior of

$$Z_n^M(\delta) = \sum_{i=1}^n\{\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - \rho(u_i)\} \tag{4.9}$$

Theorem 5 *Under the (C-1), (C-2), (C-3), (C-4), and (C-5), if $\alpha > 1/3$, then*

$$\sqrt{n}(\hat{\beta}^M - \beta) \xrightarrow{d} N(0, q(1-q)D_1^{-1}D_0D_1^{-1}) \tag{4.10}$$

Proof. $Z_n^M(\delta)$ can be decomposed into

$$\begin{aligned} Z_n^M(\delta) &= \sum_{i=1}^n\{\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - \rho(u_i - x_i^\top\delta/\sqrt{n})\} + \sum_{i=1}^n\{\rho(u_i - x_i^\top\delta/\sqrt{n}) - \rho(u_i)\} \\ &= \sum_{i=1}^n\{\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - \rho(u_i - x_i^\top\delta/\sqrt{n})\} + Z_n(\delta) \end{aligned}$$

To get a consistent estimator, we set the λ_γ in $\rho^M(\cdot)$ to be $\lambda_\gamma{=}c\cdot n^\alpha$ where c is a constant, n is the sample size, α is a positive constant. Making use of a Taylor series

expansion of the density, and using software (MAPLE)

$$
\begin{aligned}
& E\Big(\sum_{i=1}^{n}\{\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - \rho(u_i - x_i^\top\delta/\sqrt{n})\}\Big) + \frac{nq(1-q)}{2\lambda_\gamma} \\
= & \sum_{i=1}^{n}\int_{x_i^\top\delta/\sqrt{n}}^{\frac{1-q}{\lambda_\gamma}+x_i^\top\delta/\sqrt{n}} (\frac{\lambda_\gamma}{2}\frac{q}{1-q}(u - \frac{x_i^\top\delta}{\sqrt{n}})^2 - q(u - \frac{x_i^\top\delta}{\sqrt{n}}) + \frac{q(1-q)}{2\lambda_\gamma}) f_i(\xi_i + u)du \\
+ & \sum_{i=1}^{n}\int_{-\frac{q}{\lambda_\gamma}+x_i^\top\delta/\sqrt{n}}^{x_i^\top\delta/\sqrt{n}} (\frac{\lambda_\gamma}{2}\frac{1-q}{q}(u - \frac{x_i^\top\delta}{\sqrt{n}})^2 - (q-1)(u - \frac{x_i^\top\delta}{\sqrt{n}}) + \frac{q(1-q)}{2\lambda_\gamma}) f_i(\xi_i + u)du \\
= & \frac{q(1-q)}{6c^2n^{2\alpha}}\sum_{i=1}^{n} f_i(\xi_i) + \frac{q(1-q)}{6c^2n^{2\alpha}}\sum_{i=1}^{n}\frac{f_i'(\xi_i)x_i^\top\delta}{\sqrt{n}} + o(n^{-2\alpha+1/2})
\end{aligned}
$$

Note that $\sum_{i=1}^{n} f_i'(\xi_i)x_i^\top\delta/\sqrt{n} = O(\sqrt{n})$, as $f_i'(\xi_i)$, $i = 1, \ldots, n$, are uniformly bounded from the condition (C-2), and $|x_i^\top\delta| \le \|x_i\|_2\|\delta\|_2 \le (\|x_i\|_2^2 + \|\delta\|_2^2)/2$, while $\sum_{i=1}^{n}\|x_i\|_2^2 = O(n)$ from the condition (C-3).

Taking $C_n := -q(1-q)/(2cn^{\alpha-1}) + q(1-q)/(6c^2n^{2\alpha})\sum_{i=1}^{n} f_i(\xi_i)$, we have that

$$
E\sum_{i=1}^{n}\{\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - \rho(u_i - x_i^\top\delta/\sqrt{n})\} - C_n \to 0 \quad \text{if } \alpha > 1/4.
$$

And similarly,

$$
\begin{aligned}
& Var\Big(\sum_{i=1}^{n}\{\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - \rho(u_i - x_i^\top\delta/\sqrt{n})\}\Big) \\
= & \sum_{i=1}^{n}\frac{q^2(1-q)^2 f_i(\xi_i)}{20c^3n^{3\alpha}} + o(n^{-3\alpha+1}) \to 0 \quad for\ \alpha > 1/3
\end{aligned}
$$

Thus under the condition that $\alpha > 1/3$,

$$
\sum_{i=1}^{n}\{\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - \rho(u_i - x_i^\top\delta/\sqrt{n})\} - C_n \xrightarrow{p} 0
$$

This implies that the limiting behavior of $Z_n^M(\delta)$ is same the as that of $Z_n(\delta)$. From the proof of Theorem 4.1 in Koenker (2005), $Z_n(\delta) \xrightarrow{d} -\delta^\top W + \frac{1}{2}\delta^\top D_1\delta$, where W $\sim N(0, q(1-q)D_0)$. By the convexity argument in Koenker (2005) (see also Pollard (1991), Hjort and Pollard (1993), and Knight (1998)), $\hat{\delta}_n$, the minimizer of $Z_n^M(\delta)$,

converges to $\hat{\delta}_0 := D_1^{-1}W$, the unique minimizer of $-\delta^\top W + \frac{1}{2}\delta^\top D_1 \delta$ in distribution. This completes the proof. □

Although the modified quantile procedure is asymptotically equivalent to the standard quantile, for finite sample, there exist averaging effect within the interval of adjustment due to ℓ_2 type modification. It turns out that the adjustment prevents somewhat estimated quantiles from crossing. Notice that the lower bound of α which assures the consistency of modified quantile regression is increased to 1/3, from 1/4 for the modified quantile estimator.

The asymptotic equivalence of the regular quantile estimator and the modified one in the previous section can be proved similarly as the proof of the asymptotic equivalence of the regular quantile regression estimator and its modification.

4.5.3 Asymptotic Properties Under *Heterogeneous* Errors

We follow the same conditions in section 4.5.2. The methods described in previous sections rely on conditional independence of y_i given x_i, and on correct specification of the quantile function. They do not assume that the u_i are identically distributed. As a consequence, the results on consistency and asymptotic normality apply quite broadly. However, one might expect that an estimator that makes use of differences in the distributions f_i would be more efficient. As we show in this subsection, this is indeed the case.

There are two main ways to introduce differences amongst the densities f_i into the quantile regression problem. The first is to create a set of weights (as in one description of weighted least squares), and to define the conditional quantile function

as the minimizer of a weighted criterion. Thus, one might use a weighted version of (4.5):

$$\tilde{\beta}^M = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} f_i(\xi_i)\rho^M(y_i - x_i^\top\beta). \tag{4.11}$$

The second approach involves scaling the loss function (as in a second description of weighted least squares) to account for differing local dispersions. Thus, one might replace $\rho(y_i - x_i^\top\beta)$ with $\rho((y_i - x_i^\top\beta)/f_i(\xi_i))$.

The two approaches are identical in the content of the check loss. However, under a modified loss function, they differ. We begin with the first formulation.

CONDITION

(C-6). $\lim_{n\to\infty} n^{-1}\sum f_i^2(\xi_i)x_i x_i^\top = D_2$, where D_2 is positive definite

Under conditions (C-1) through (C-6), Theorem 5.1 in Koenker (2005) shows that $\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, q(1-q)D_2^{-1})$, where $\tilde{\beta} = \arg\min_{\beta\in\mathbb{R}^p}\sum_{i=1}^n f_i(\xi_i)\rho(y_i - x_i^\top\beta)$.

Theorem 6 *Under the (C-1), (C-2), (C-3), (C-4), (C-5), and (C-6), if $\alpha > 1/3$, then*

$$\sqrt{n}(\tilde{\beta}^M - \beta) \xrightarrow{d} N(0, q(1-q)D_2^{-1}) \tag{4.12}$$

Proof. First, consider a slight modification of Theorem 5, with a similar decomposition.

$$\begin{aligned} Z_n^{M*}(\delta) &= \sum_{i=1}^{n}(f_i(\xi_i)\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - f_i(\xi_i)\rho(u_i)) \\ &= \sum_{i=1}^{n}(f_i(\xi_i)\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - f_i(\xi_i)\rho(u_i - x_i^\top\delta/\sqrt{n})) + Z_n^*(\delta). \end{aligned}$$

Following the argument in Theorem 4.1 of Koenker (2005), $Z_n^*(\delta) \xrightarrow{d} -\delta\check{W} + \frac{1}{2}\delta^\top D_2\delta$ where $\check{W} \sim N(0, q(1-q)D_2)$. we have that

$$E\sum_{i=1}^{n}\{(f_i(\xi_i)\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - (f_i(\xi_i)\rho(u_i - x_i^\top\delta/\sqrt{n})\} - C_n^* \to 0 \quad \text{if } \alpha > 1/4,$$

where $C_n^* := -q(1-q)/(2cn^\alpha)\sum_{i=1}^n f_i(\xi_i) + q(1-q)/(6c^2n^{2\alpha})\sum_{i=1}^n f_i(\xi_i)^2$. And similarly,

$$\begin{aligned} & Var\Big(\sum_{i=1}^n \{f_i(\xi_i)\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - f_i(\xi_i)\rho(u_i - x_i^\top\delta/\sqrt{n})\}\Big) \\ = & \sum_{i=1}^n \frac{q^2(1-q)^2 f_i(\xi_i)^2}{20c^3n^{3\alpha}} + o(n^{-3\alpha+1}) \to 0 \quad for\ \alpha > 1/3 \end{aligned}$$

Thus under the condition that α >1/3,

$$\sum_{i=1}^n \{f_i(\xi_i)\rho^M(u_i - x_i^\top\delta/\sqrt{n}) - f_i(\xi_i)\rho(u_i - x_i^\top\delta/\sqrt{n})\} - C_n^* \xrightarrow{p} 0$$

and this completes the proof. □

Now let

$$D = \begin{pmatrix} D_2 & D_1 \\ D_1 & D_0 \end{pmatrix} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \begin{pmatrix} f_i^2 & f_i \\ f_i & 1 \end{pmatrix} \otimes x_i x_i^\top, \tag{4.13}$$

where $\otimes$ represents the Kronecker product. Since all D_i in (4.13) are nonnegative definite by (C-3), (C-4) and (C-6), D is also a nonnegative definite matrix. There exists an orthogonal matrix P such that

$$P^\top D P = \begin{pmatrix} D_2 & 0 \\ 0 & D_0 - D_1 D_2^{-1} D_1 \end{pmatrix}.$$

The fact that $D_0 - D_1 D_2^{-1} D_1$ is nonnegative definite and that D_1 is nonsingular imply that $D_1^{-1} D_0 D_1^{-1} - D_2^{-1}$ is also nonnegative definite, and consequently that $\tilde{\beta}^M$ is more efficient than $\hat{\beta}^M$. In practice, $f_i(\xi_i)$ in the (4.11) needs to be estimated. We do not pursue this here.

A simple simulation with a heterogeneous model given below is considered, without the hindrance of the density estimation problem. The heterogeneous error model is given by

$$y_i = \beta_0 + \beta_1 x_i + x_i \cdot u_i,$$

where the u_i's are *iid* standard normal, $(\beta_0, \beta_1)^\top = (1, 2)^\top$, and x takes one of three values, $\in \{1, 2, 3\}$. 200 data sets are generated for each of two sample sizes n=300 and n=900. In each case, the covariate, x, is distributed across the three design points.

Four different models are fit to the data; standard quantile regression (QR), modified quantile regression (QR.M), weighted QR (WQR), and weighted QR.M (WQR.M). The true q^{th} quantile regression line passes through the three points $(1, 3 + \Phi^{-1}(q))$, $(2, 5 + 2\Phi^{-1}(q))$, and $(3, 7 + 3\Phi^{-1}(q))$ for each $0 < q < 1$, where $\Phi(\cdot)$ is cumulative distribution function of standard normal distribution.

The mean squared error (MSE) between the true quantile line and each of the four fitted lines is computed for the 200 data sets for each quantile. Since the density of the error distribution at x=1 is twice as large as that at x=2, and three times larger than that at x=3, weights of (6/11,3/11,2/11) are given to the cases at x=1, 2, and 3, respectively. Table 4.1 reveals that the weighted versions (WQR and WQR.M) perform better than the unweighted counterparts in nearly every case. We suspect that the lone reveral of this pattern for QR and WQR is due to simulation variation. Surprisingly, QR.M outperforms WQR when $q > 0.3$. In these instances, WQR.M performs even better than QR.M.

The second method to introduce different weight is to consider the following location-scale model,

$$y_i = x_i^\top \beta + (x_i^\top \tau) u_i, \tag{4.14}$$

where u_i's are *iid* from distribution F with finite density f. Koenker and Zhao (1994) define the weighted quantile regression estimator $\check{\beta}_\tau$ as $\arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho((y_i - x_i^\top \beta)/x_i^\top \tau)$. With any $\sqrt{n}$-consistent estimator of τ up to a constant, they showed that $\sqrt{n}(\check{\beta}_\tau - \beta) \xrightarrow{d} N(0, q(1-q)D_2^{-1})$. Now, with $\rho^M(\cdot)$, the modified quantile regression estimator

Table 4.1: Point estimates and approximate 95% confidence intervals for MSE (multiplied by 1000), based on 200 replicates with n=300, and n=900, at selected quantiles.

	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5
			n=300		
QR	92.31(8.85)	66.68(5.47)	56.40(4.80)	44.77(3.57)	42.53(3.33)
QR.M	92.91(8.41)	62.37(5.19)	48.38(4.06)	37.48(2.94)	33.47(2.64)
WQR	85.49(8.32)	62.10(5.53)	54.93(4.97)	42.82(3.46)	42.32(3.48)
WQR.M	84.44(8.03)	57.93(5.19)	44.84(3.92)	35.55(2.89)	31.21(2.48)
			n=900		
QR	28.93(2.44)	22.10(2.04)	18.06(1.56)	15.50(1.43)	14.41(1.25)
QR.M	28.89(2.40)	21.09(1.95)	15.90(1.35)	13.13(1.14)	12.22(1.07)
WQR	27.77(2.31)	21.44(1.91)	17.73(1.55)	15.20(1.44)	14.42(1.21)
WQR.M	28.28(2.59)	20.09(1.82)	15.30(1.28)	12.58(1.10)	11.54(1.00)

$\check{\beta}_\tau^M$ is defined as

$$\check{\beta}_\tau^M = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^n \rho^M((y_i - x_i^\top\beta)/x_i^\top\tau). \tag{4.15}$$

Theorem 7 *Under (C-1)through (C-6),and* $\hat{\tau} = \kappa\tau + O_p(n^{-1/2})$ *for some scalar* κ*, if* $\alpha > 1/3$*, then*

$$\sqrt{n}(\check{\beta}_\tau^M - \beta) \xrightarrow{d} N(0, q(1-q)D_2^{-1}). \tag{4.16}$$

Proof. The proof of convergence is similar to the proof of Theorem 5, except that u_i is now expressed as $(y_i - x_i^\top\beta)/(x_i^\top\tau)$. The behavior of $\sqrt{n}(\check{\beta}_\tau^M - \beta)$ follows from consideration of $\sum_{i=1}^n \rho^M(u_i - \frac{x_i^\top\delta}{x_i^\top\tau}/\sqrt{n}) - \rho(u_i)$. First, we decompose the above expression.

$$\begin{aligned} Z_n^{M**}(\delta) &= \sum_{i=1}^n (\rho^M(u_i - \frac{x_i^\top\delta}{x_i^\top\tau}/\sqrt{n}) - \rho(u_i)) \\ &= \sum_{i=1}^n (\rho^M(u_i - \frac{x_i^\top\delta}{x_i^\top\tau}/\sqrt{n}) - \rho(u_i - \frac{x_i^\top\delta}{x_i^\top\tau}/\sqrt{n})) + Z_n^{**}(\delta). \end{aligned}$$

Observe that $Z_n^{M**}(\delta) - Z_n^{**}(\delta) - C_n \xrightarrow{p} 0$, where C_n is defined in Theorem 5. Since Theorem 2.1 of Koenker and Zhao (1994) shows $\sqrt{n}(\check{\beta}_\tau - \beta) \xrightarrow{d} N(0, q(1-q)D_2^{-1})$, we complete the proof. □

Again, the above estimator is more efficient than the unscaled version, and it achieves the same efficiency as the weighted quantile regression estimates.

The analogy that weighted least squares is to least squares as weighted quantile regression is to quantile regression, is telling. In addition to motivating two forms of modified quantile regression (we prefer the second form, due to an invariance argument), our simulation reveals the type of improvement we expect when we incorporate more features of the model into the analysis. The analogy is farther reaching. If we have approximately correct weights, we expect to see improvement over the unweighted analysis. This suggests the use of a relatively simple model for the weights, perhaps coupled with a more complex model for the quantile function.

4.6 Analysis of Engel's Data

Engel's data consists of the household food expenditure and household income from 235 European working-class households in the 19th century. Taking the log of food expenditure as a response variable, we investigate the relation between log of food expenditure and log of household income. In Figure 4.12, Engel's data is plotted after transformation of both variables. Superimposed on the scatter plot are the fitted lines from quantile regression (QR), and modified quantile regression (QR.M) using the rule developed in Section 3. Although the two methods display quite similar fitted lines, Figure 4.13 reveals the difference between QR and QR.M. We note that

these fitted lines from modified quantile regression do not across over the range of log(Household income) in the data. This is partly due to the averaging effect of the ℓ_2 adjustment to the check loss function.

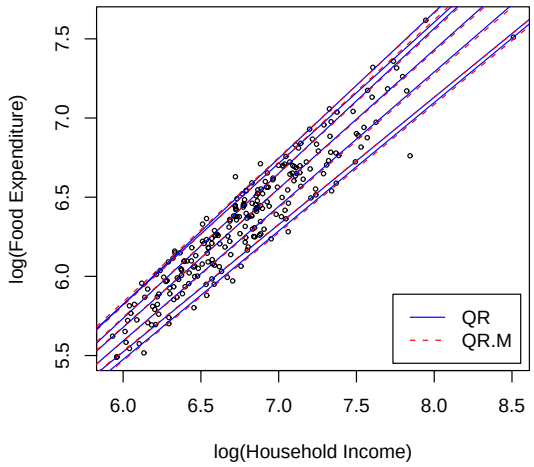

Figure 4.12: Superimposed on the scatter plot are the 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95 standard quantile regression (solid, blue) lines, and modified quantile regression (dashed, red) lines for Engel's data after log transformation of both response and predictor variables.

4.7 Analysis of NHANES Data

In this section, nonparametric versions of QR and QR.M are numerically compared for analysis of real data. The Centers for Disease Control and Prevention conduct the National Health and Nutrition Examination Survey (NHANES), a large scale survey designed to monitor the health and nutrition of residents of the United States. Many are concerned about the record levels of obesity in the population, and the survey

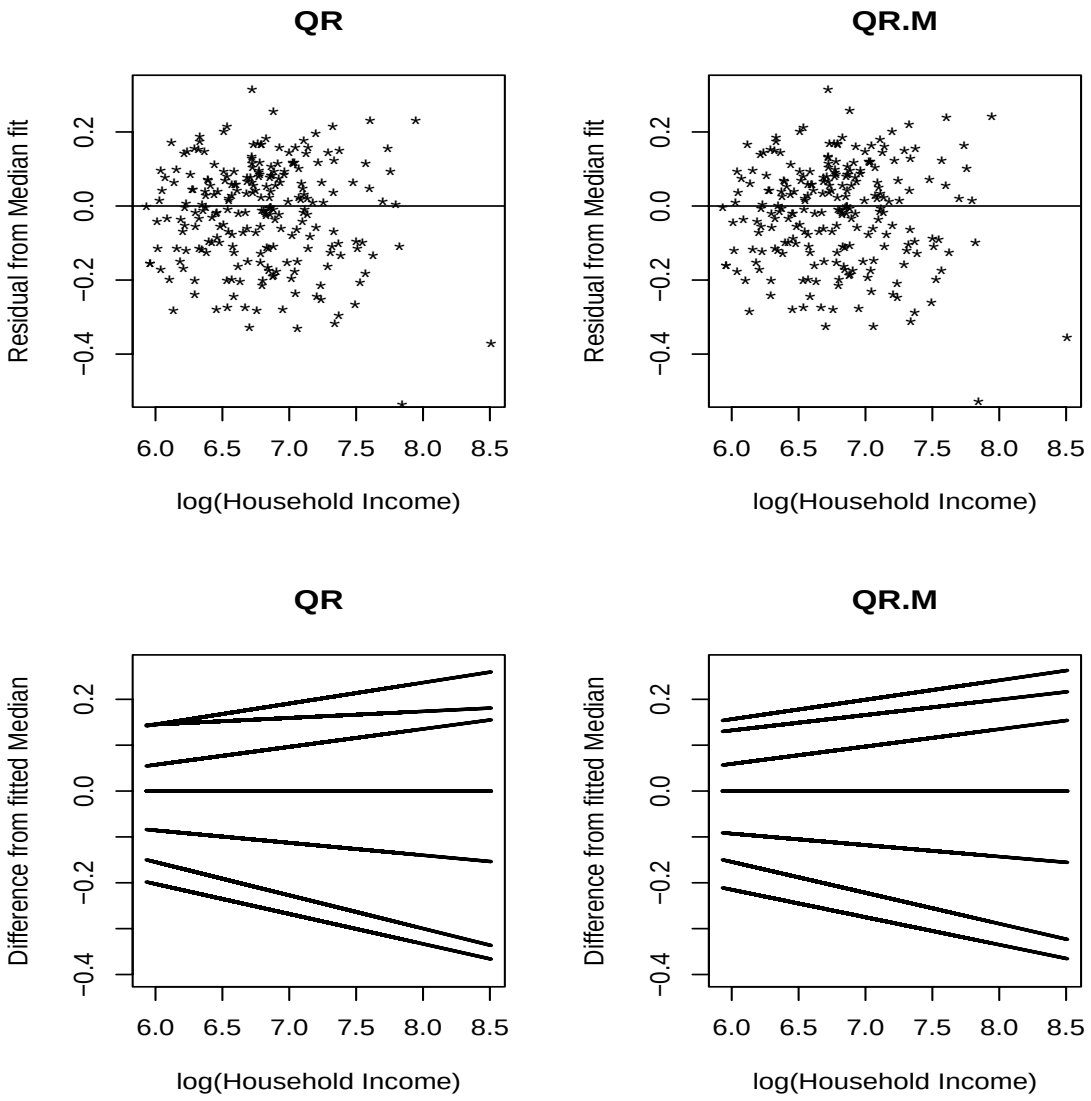

Figure 4.13: Top: Residuals from a median fit via QR and QR.M. Bottom: Differences between fitted median line and the fitted quantiles at q=0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95.

contains information on height and weight of individuals, in addition to a variety of dietary and health-related questions. Obesity is defined through body mass index (BMI) in kg/m^2, a measure which adjusts weight for height. In this analysis, we describe the relationship between height and BMI among the 5938 males over the age of 18 in the aggregated NHANES data sets from 1999, 2001, 2003 and 2005. Since BMI is weight adjusted for height, the null expectation is that BMI and height are unrelated.

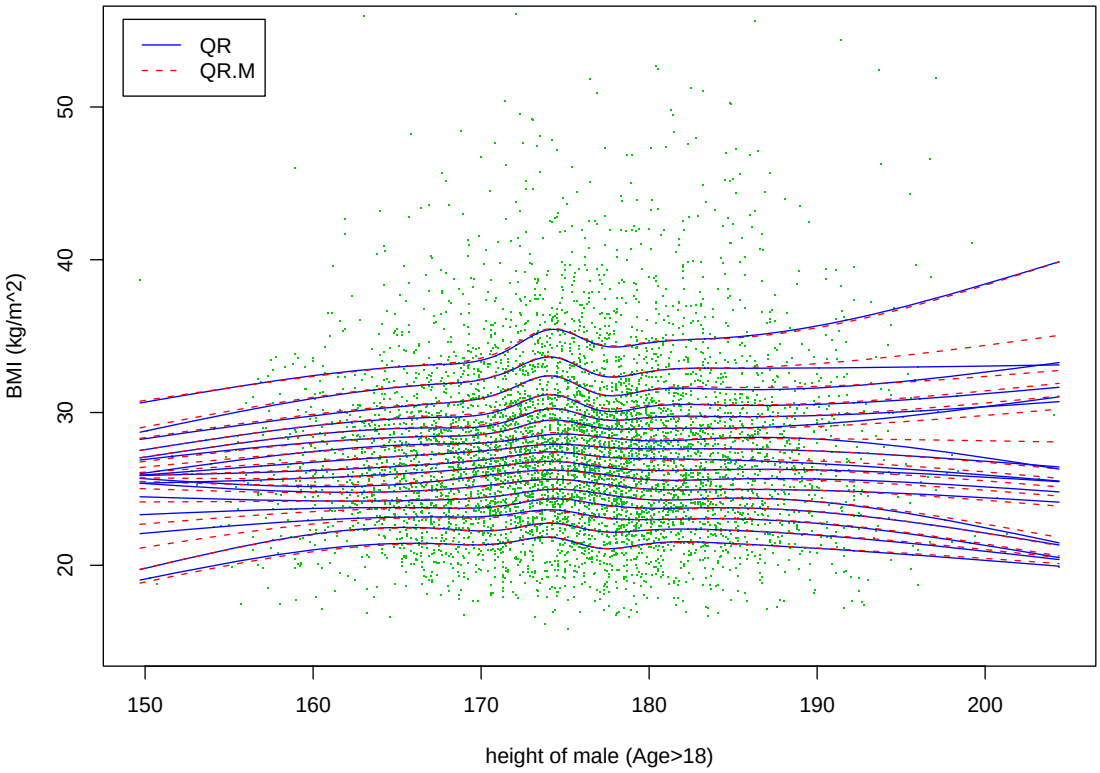

Figure 4.14: Regression spline estimates of conditional BMI quantiles in steps of 0.05, from 0.1 to 0.9 for the NHANES data. Natural spline bases and 6 knots are used in each fitted curve.

We fit a nonparametric quantile regression model to the data. The model is a six knot regression spline using the natural basis expansion. The knots (held constant across quantiles) were chosen by eye. The modified quantile regression method was directly implemented by specifying (4.6) in the `rlm` function in the R package.

Figure 4.14 displays the fits from standard (QR) and modified (QR.M) quantile regressions for the quantiles between 0.1 and 0.9 in steps of 0.05. The fitted curves show a slight upward trend, some curvature overall, and mildly increasing spread as height increases. There is a noticeable bump upward in the distribution of BMI for heights near 1.73 meters. The differences between the two methods of fitting the quantile regressions are most apparent in the tails, for example the 0.6th and 0.85th quantiles for large heights.

The predictive performance of the standard and modified quantile regressions are compared in Figure 4.15. To compare the methods, 10-fold cross validation was repeated 500 times for different splits of the data. Each time, a cross validated score was computed by

$$CV = \frac{1}{n}\sum_{i=1}^{n}\rho(y_i - \hat{y}_i), \tag{4.17}$$

where y_i is the observed BMI for an individual in the hold-out sample, $\hat{y}_i$ is the fitted value under QR or QR.M, and the sum runs over the hold-out sample. The figure contains plots of the 500 CV scores. The great majority of CV scores are to the lower right side of the 45 degree line, indicating that the modified quantile regression outperforms the standard method–even when the QR empirical risk function is used to evaluate performance. The pattern shown in these panels is consistent across the target quantiles. The pattern becomes a bit stronger when the QR.M empirical risk function is used to evaluate performance.

To compare the small sample behavior of the two methods, we split the NHANES data set into a small part for fitting and the rest for validation. CV score with the entire data can be regarded as a base CV score. Thus, the base CV score is subtracted from the CV score with small sample size to compare the *excess* CV scores only. Figure 4.16 shows this *excess* CV score as we increase the sample size from 100 to 800. In most quantiles, the *excess* CV score shows considerable reduction when the sample size increases from 100 to 200, and thereafter the difference diminishes as expected by the asymptotic equivalence of QR.M.

Modified quantile regression has an additional advantage which is apparent for small and large heights. The standard quantile regression fits show several crossings of estimated quantiles, while crossing behavior is reduced considerably with modified quantile regression. Crossed quantiles correspond to a claim that a lower quantile lies above a higher quantile, contradicting the laws of probability. Figure 4.17 shows this behavior. Fixes for this behavior have been proposed (e.g., He (1997)), but we consider it desirable to lessen crossing without any explicit fix. The reduction in crossing holds up across other data sets that we have examined and with regression models that differ in their details.

4.8 Conclusion

We have shown that case-specific indicators can be employed in quantile regression through ℓ_2 regularization of their parameters. This modification increases the efficiency of quantile regression due to an averaging effect near the target quantile. The simulation studies suggest a simple heuristic rule to choose the degree of regularization. The behavior of the heuristic rule is excellent under both symmetric and

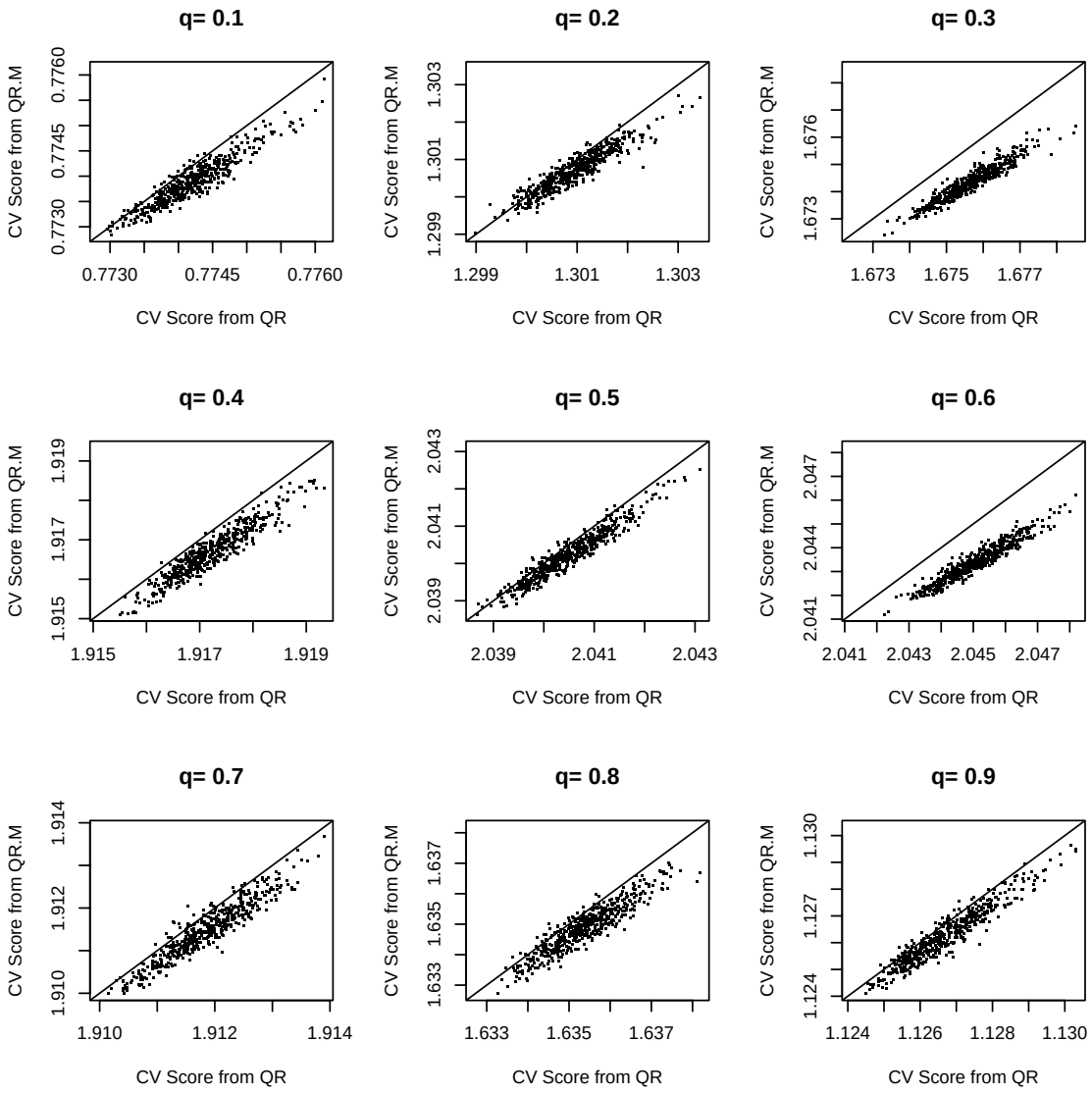

Figure 4.15: Scatter plots of 10-fold CV scores from standard quantile regression (QR) and modified method (QR.M) at various quantiles with 45 degree lines. Regression splines with natural spline bases and 6 knots are fitted to the NHANES data.

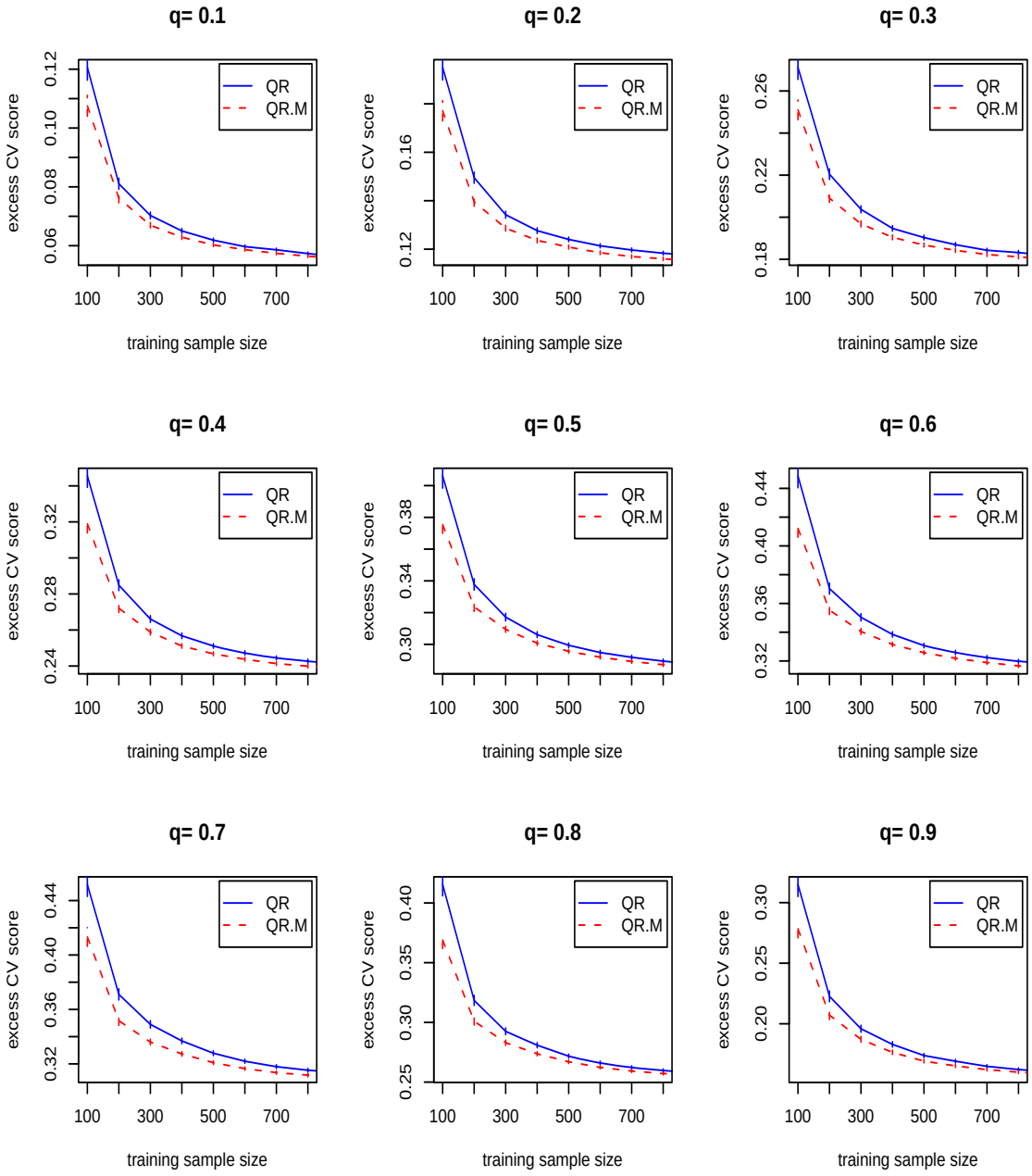

Figure 4.16: *Excess* CV scores from two methods (QR and QR.M) at various quantiles with approximate 95% confidence intervals from 500 replicates

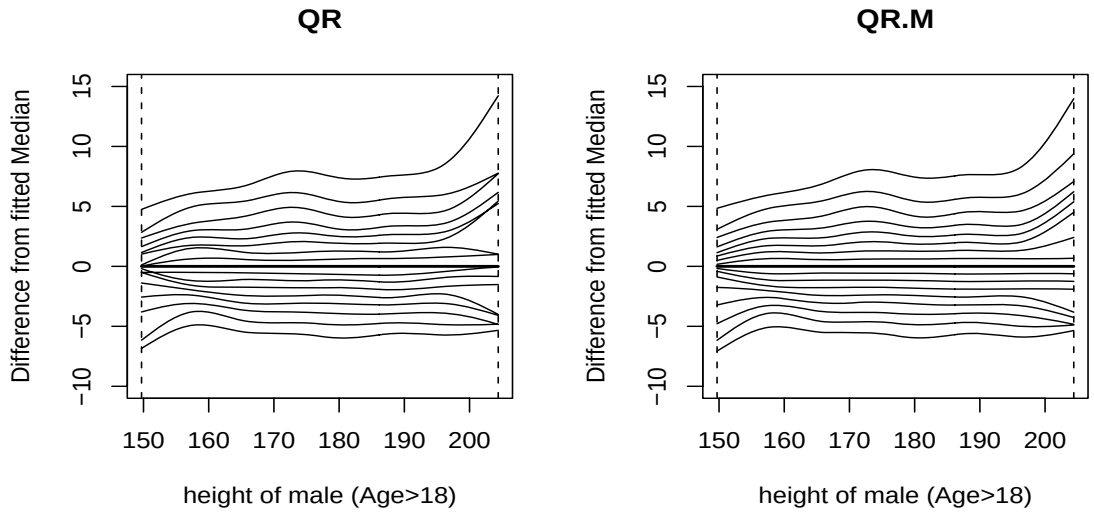

Figure 4.17: Differences between fitted median line and the other fitted quantiles for standard quantile regression (QR) and modified quantile regression (QR.M). The dashed lines are the minimum and maximum of the observed heights.

asymmetric error distributions at any conditional quantile. The analysis of Engel's data (linear fit) and the NHANES data (nonlinear fit) reveal desirable properties of modified quantile regression. For large sample behavior, theoretical results establish consistency and asymptotic equivalence to traditional quantile regression. In terms of computation, modified quantile regression requires only slight adjustment to existing software. Implementation of the method is easy, and computation is quick.

CHAPTER 5

CROSS VALIDATION IN QUANTILE REGRESSION

5.1 Literature Review of Cross Validation

Model selection is a crucial problem in statistical analysis. As a popular data driven method for model selection, cross validation is described formally in Stone (1974) and Geisser (1975). Wahba and Wold (1975) proposed cross-validation for selection of the smoothing parameter in smoothing spline regression. Among the many authors who have considered cross validation, Shao (1993) provides a rigorous treatment of model selection under the linear model. Efron and Tibshirani (1997) argued that cross validation is nearly unbiased for the future error rate, but often highly variable, and suggested a new bootstrapping method. Kohavi (1995) provides a thorough review and performs experiments comparing cross validation and bootstrap.

Typically, the loss criterion used for model fitting is employed for model validation. Just as certain modification of a loss criterion could bring robustness or better efficiency in modeling, validation with a modified criterion could lead to selection of a better model. For example, to reduce the influence of outliers in model selection, Ronchetti et al. (1997) proposed cross validation with a robust loss function when the

squared error loss is used for fitting a linear model. Leung (2005) discusses a similar issue in nonparametric regression setting.

In k-fold cross validation, the choice of k is an interesting theme (Kohavi; 1995). When k is equal to the sample size, k-fold cross validation is also called leave-one-out cross validation. In the linear model, leave-one-out cross validation is shown to be asymptotically equivalent to AIC (Stone; 1977) and inconsistent (Shao; 1997). Like AIC, this method is too conservative and suffers from asymptotic difficulties (Shao; 1993), which may lead to a larger and unrealistic model (Efron; 1983).

In this thesis, we mainly focus on cross validation in quantile regression settings. In quantile regression, the check loss function is commonly used for validation as well as fitting. In Chapter 4, a modified check loss function is employed for fitting a quantile regression model. Here, we consider use of the modified check loss as a validation function for cross validation (CV.M), and compare its performance to that of cross validation using the check loss (CV). Since the modified check loss is designed to gain the efficiency at the expense of allowing greater bias, the suggested method can be regarded as applying a new version of bias-variance tradeoff to the problem of estimating future error. Finally, the purpose of this chapter is not to discuss the choice of k but to examine the difference of two validation functions. For the remainder of the chapter, k is fixed at 10.

5.2 Modified Cross Validation Function

Recall that the modified check loss has the form of

$$\rho^M(u) = \begin{cases} qu - \frac{q(1-q)}{2\lambda_\gamma} & \text{for} \quad \frac{1-q}{\lambda_\gamma} \le u \\ \frac{\lambda_\gamma}{2}\frac{q}{1-q}u^2 & \text{for} \quad 0 \le u < \frac{1-q}{\lambda_\gamma} \\ \frac{\lambda_\gamma}{2}\frac{1-q}{q}u^2 & \text{for} \quad -\frac{q}{\lambda_\gamma} \le u < 0 \\ (q-1)u - \frac{q(1-q)}{2\lambda_\gamma} & \text{for} \quad u < -\frac{q}{\lambda_\gamma}. \end{cases} \tag{5.1}$$

The above function will be used for model validation. The length of ℓ_2 adjustment (a.k.a., the window width) is guided by the same rule we developed in Section 4.4. However, the length of adjustment for validation differs from that for fitting. For k-fold cross validation, $(k-1)$ folds are used for model fitting, and only one fold is used for validation. Thus the window width, which is based on the sample size in the validation set, is larger than the window width used to fit a modified quantile regression. Notice that the modified check loss can be employed in both fitting and validation. To separate the effect of the loss function on fitting from that on validation, we restrict our attention to the effect of modification of the loss for validation only in this chapter.

First, observe that the modified check loss is a consistent validation function. When α >0, we immediately see that $\rho^M(u) \to \rho(u)$ for every u. Furthermore, the absolute difference between the check loss and modified check loss is uniformly bounded. That is, $sup_u|\rho^M(u) - \rho(u)| = \frac{q(1-q)}{\lambda_\gamma}$. When q=1/2, the validation function in (5.1) becomes Huber's function. The performances of absolute deviation, squared loss and Huber's loss function as a validation function are compared in Leung (2005) for nonparametric kernel regression. A simulation study in the next section investigates the finite sample performance of cross validation with the modified criterion.

5.3 Simulations

The behavior of the modified check loss function in (5.1) as a validation criterion is considered under various statistical procedures. The modified check loss is compared to the check loss for validation under the three settings of 1) the linear model, 2) regression splines, and 3) smoothing splines.

5.3.1 Validation under the Linear Model

Simulation settings similar to the linear model setting in Chapter 3 are employed. The dense case is changed to β=(0,0.85,0.85,0.85,0.85,0.85,0.85,0) since the previous dense case shows no difference in selected models. The sparse case (β=(5,0,0,0,0,0,0,0)) and intermediate case (β=(3,1.5,0,0,2,0,0,0)) remain unchanged. Assuming $y = x^\top\beta + \epsilon$, we generated $x = (x_1, \ldots, x_8)^\top$ from a multivariate normal distribution with mean zero and standard deviation 1. The correlation between x_i and x_j was set to $\rho^{|i-j|}$ with $\rho = 0.5$. We consider selection of a linear model among all possible 256 ($= 2^8$) subset models.

The standard quantile regression models with smallest prediction error are selected by 10-fold cross validation based on two loss functions; check loss (CV) and modified check loss (CV.M). The scale parameter, σ, is estimated from the full least squares model. As in Section 4.3,

$$MSE = E^{\hat{\beta}}\{(\hat{\beta} - \beta)^\top\Sigma(\hat{\beta} - \beta) + (\hat{\beta}_0 - \beta_0)^2\} \tag{5.2}$$

is approximated by a Monte Carlo estimate from 1000 data sets:

$$\widehat{MSE} = \frac{1}{1000}\sum_{i=1}^{1000}((\hat{\beta}^i - \beta)^\top\Sigma(\hat{\beta}^i - \beta) + (\hat{\beta}_0^i - \beta_0)^2).$$

Table 5.1: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE, based on 1000 replicates with n=500, at selected quantiles. Sparse, intermediate and dense cases are considered.

	q=.1	q=.25	q=.5	q=.8
Sparse	3.794	9.328	16.451	7.640
	(2.602,4.986)	(7.750,10.906)	(14.811,18.091)	(6.098,9.182)
Intermediate	1.333	3.854	8.246	3.417
	(0.411,2.256)	(2.716,4.992)	(6.822,9.570)	(2.332,4.503)
Dense	0.556	1.036	2.298	0.530
	(0.190,0.921)	(0.365,1.708)	(1.739,2.858)	(0.072,0.987)

Note that $\widehat{MSE}$ values from CV and CV.M are equal only when the two methods choose the same model over every replicate. $\widehat{MSE}$ from CV ($\widehat{MSE}_{CV}$) and $\widehat{MSE}$ from CV.M ($\widehat{MSE}_{CV.M}$) gauge the relative accuracy of the two validation functions in assessing the fitted models. Table 5.1 shows the percentage reduction under three linear models, along with approximate confidence intervals for the corresponding theoretical percentage reductions. The variance of the asymptotic distribution of the above quantity is approximated through the delta method, which suggests the use of

$$\widehat{Var}(\frac{\bar{Y}}{\bar{X}}) \approx \frac{1}{n}\left[\frac{\bar{y}^2}{\bar{x}^4}s_x^2 + \frac{1}{\bar{x}^2}s_y^2 - \frac{2\bar{y}}{\bar{x}^3}\hat{\rho}s_x s_y\right].$$

The various simulation settings in Table 5.1 reveal that the modified validation function outperforms the traditional check loss function for all three linear models. As we move from the dense case to the sparse case, the advantage of CV.M over CV grows. When the true regression coefficients are all substantial and non-zero, the two methods are equivalent up to simulation variation (results are omitted). There is certainly a lower bound under which $\widehat{MSE}_{CV}$ and $\widehat{MSE}_{CV.M}$ cannot go. Given

Table 5.2: Point estimates and approximate 95% confidence intervals for percentage reduction in mean *excess* MSE, based on 1000 replicates with n=500, at selected quantiles. Sparse, intermediate and dense cases are considered.

	q=.1	q=.25	q=.5	q=.8
Sparse	5.078	12.590	22.349	10.355
	(3.494,6.662)	(10.500,14.681)	(20.189,24.508)	(8.283,12.427)
Intermediate	2.608	7.755	16.362	6.795
	(0.810,4.406)	(5.503,10.006)	(13.834,18.890)	(4.666,8.924)
Dense	2.870	6.044	13.154	2.709
	(0.995,4.744)	(2.232,9.856)	(10.108,16.200)	(0.394,5.024)

a data set, $\widehat{MSE}$ is defined as the minimum MSE of all possible subset models. Treating $\widehat{MSE}_{min}$ as an achievable base, we define the excess (*excess* MSE) to be $\widehat{MSE}_{CV} - \widehat{MSE}_{min}$ and $\widehat{MSE}_{CV.M} - \widehat{MSE}_{min}$ for the two validation functions. We then compute the percentage reduction in *excess* MSE. Table 5.2 presents the results. Again, we see that CV.M provides a more accurate assessment of the linear quantile regression model than CV does.

From Tables 5.1 and 5.2, the modified check loss function presents its potential as a validation function for linear quantile regression. In subsequent sections, we turn our attention to the performance of CV.M for nonlinear models.

In addition to the performance in terms of MSE, the extent of disagreement between the models selected by CV and CV.M is also of interest. Table 5.3 summarizes the number of agreements and disagreements among 1000 replicates. Disagreement tends to increase as the probability density at the target quantile increases, and as the scenario moves from 'dense' to 'sparse'.

Table 5.3: Number of agreements and disagreements of models selected by CV and CV.M among 1000 replicates in each scenario of sparse, intermediate and dense at several quantiles. '=' , '+', and '–' represent selection of the same model, selection of a better model (in terms of MSE by CV.M, and selection of a worse model by CV.M respectively.

	Sparse	Intermediate	Dense
	(=,+,–)	(=,+,–)	(=,+,–)
q=0.1	(694,193,113)	(795,130,75)	(914,56,30)
q=0.25	(535,309,156)	(644,222,134)	(841,89,70)
q=0.5	(471,396,133)	(584,284,132)	(810,132,58)
q=0.8	(546,308,146)	(679,202,119)	(866,77,57)

5.3.2 Validation under Regression Splines

Two simple simulation settings (adapted and modified from Yu and Jones (1998)) are considered to observe the performance of CV and CV.M under nonparametric quantile regression. Regression spline models with a natural spline basis expansion are fitted to the simulated data. A reasonable range of the number of interior knots is considered, and the number of knots is selected by 10-fold cross validation using either check loss or modified check loss. Data sets were simulated from

1. Simple quantiles; $y_i = 2 + 2\cos(x_i) + \exp(-4x_i^2) + \epsilon_i$,

2. Smooth "curvy" quantile: $y_i = 2.5 + \sin(2x_i) + 2\exp(-16x_i^2) + 0.5\epsilon_i$.

The covariate X follows a standard normal distribution, independent of ϵ_i. Two error distributions, standard normal and Exp(1), are employed for ϵ_i. True curves at some quantiles are shown in Figure 5.1 when the errors follow normal distribution. In each case, 2000 replicates of sample size n=200, 500, and 1000 were generated.

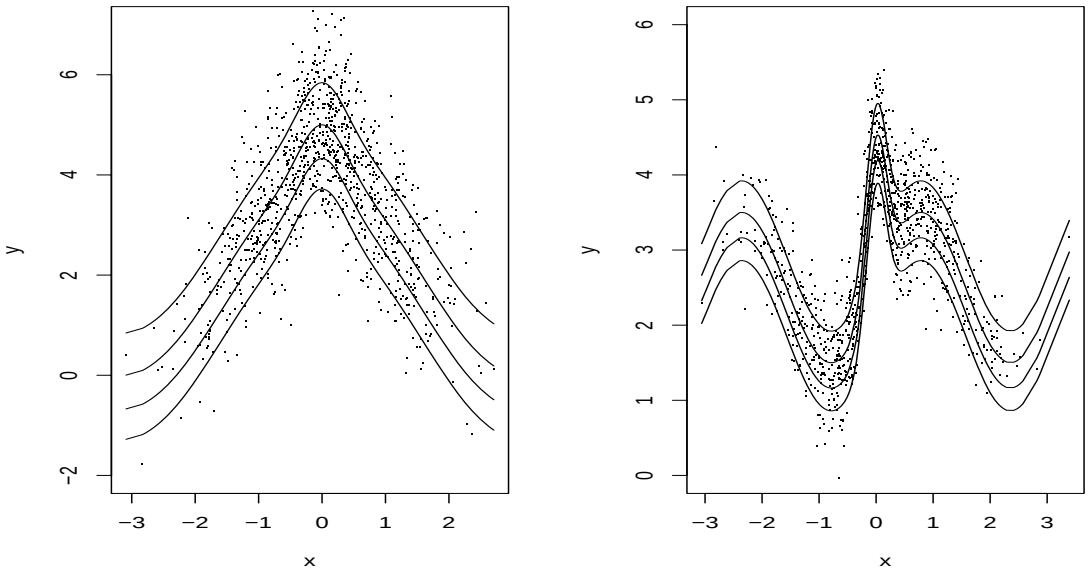

Figure 5.1: True curves for simulation 1 (left) and simulation 2 (right), at q= 0.1, 0.25, 0.5, and 0.8 quantiles, under a standard normal error distribution.

MSE is approximated by a Monte Carlo estimate given by

$$\widehat{MSE} = \frac{1}{2000}\sum_{j=1}^{2000}\frac{1}{n}\sum_{i=1}^{n}(f_s(x_i) - \hat{f}_s^j(x_i))^2,$$

where s=1 for the true underlying function in simulation 1, and s=2 for simulation 2. j indicates the jth replicate. Since there is no "true set of knots" in this simulation study, the set of knots which produces smallest MSE for a given data set is regarded as the base model being pursued. Thus, the base model depends on the generated data. The check loss and the modified check loss do sometimes select the base model, although they more typically show large deviations from the base model. CV.M selects the base model more often than CV does across all simulations carried out. For example, under the normal error distribution in the first simulation, the check loss

Table 5.4: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE under simulation 1, based on 2000 replicates with n=200, at selected quantiles. Standard normal error and Exp(1) are considered.

Error	N(0,1)	N(0,1)	Exp(1)	Exp(1)
	MSE	*excess*MSE	MSE	*excess*MSE
q=0.1	3.579	7.244	4.859	10.515
	(2.267,4.891)	(4.637,9.852)	(2.805,6.914)	(6.219,14.811)
q=0.25	5.469	11.038	4.809	10.717
	(3.863,7.076)	(7.870,14.207)	(2.912,6.705)	(6.608,14.826)
q=0.5	12.934	27.66	3.205	6.955
	(11.074,14.794)	(24.03,31.30)	(0.950,5.460)	(2.135,11.774)
q=0.8	5.605	11.89	8.463	16.728
	(4.156,7.056)	(8.89,14.88)	(6.744,10.181)	(13.463,19.991)
q=0.9	4.828	10.034	5.681	10.698
	(3.370,6.285)	(7.099,12.969)	(4.22,7.14)	(8.024,13.371)

picks the base models about 20% of the time whereas the modified check loss selects them correctly around 25% of the time. In the first simulation, both loss functions show a higher rate of selecting the base model than in the second simulation, probably due to the simpler form of the first curve, as shown in Figure 5.1. The percentage reduction in MSE and *excess* MSE when switching from check loss to modified check loss is calculated. The scale parameter, σ, is estimated by the standard deviation of the residuals from a mean smoothing spline regression fit.

Table 5.4 shows these quantities for the first simulation, when n is 200. Again, the *excess* MSE is obtained by subtracting the MSE of a base model. In every situation considered, modified check loss outperforms check loss. This tendency remains when n is 500 and 1000 as shown in Table 5.5 and Table 5.6, respectively. In the second simulation, the same quantities are investigated, and the result shows that CV.M

Table 5.5: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE under simulation 1, based on 2000 replicates with n=500, at selected quantiles. Standard normal error and Exp(1) are considered.

Error	N(0,1)	N(0,1)	Exp(1)	Exp(1)
	MSE	*excess*MSE	MSE	*excess*MSE
q=0.1	3.802	8.156	3.556	8.851
	(2.453,5.150)	(5.318,10.993)	(2.216,4.895)	(5.560,12.106)
q=0.25	5.800	13.679	8.242	21.048
	(4.337,7.263)	(10.368,16.990)	(6.841,9.644)	(17.725,24.371)
q=0.5	11.910	28.879	10.85	25.867
	(10.226,13.593)	(25.206,32.551)	(9.065,12.644)	(21.952,29.782)
q=0.8	8.232	18.75	11.308	24.396
	(6.773,9.692)	(15.610,21.908)	(9.609,13.007)	(20.974,27.818)
q=0.9	4.817	11.190	4.266	8.779
	(3.560,6.074)	(8.357,14.023)	(3.100,5.432)	(6.437,11.120)

Table 5.6: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE under simulation 1, based on 2000 replicates with n=1000, at selected quantiles. Standard normal error and Exp(1) are considered.

Error	N(0,1)	N(0,1)	Exp(1)	Exp(1)
	MSE	*excess*MSE	MSE	*excess*MSE
q=0.1	1.577	3.622	2.702	8.019
	(0.589,2.565)	(1.374,5.870)	(1.728,3.676)	(5.205,10.833)
q=0.25	4.126	10.560	5.939	18.534
	(2.928,5.324)	(7.589,13.529)	(5.019,6.868)	(15.802, 21.267)
q=0.5	8.839	23.069	10.178	27.854
	(7.492,10.186)	(19.866,26.272)	(8.756,11.600)	(24.375,31.333)
q=0.8	5.506	13.888	8.628	20.043
	(4.407,6.606)	(11.209,16.566)	(7.176,10.081)	(16.862,23.224)
q=0.9	2.723	6.951	3.967	8.502
	(1.738,3.708)	(4.492,9.409)	(2.853,5.082)	(6.168,10.836)

Table 5.7: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE under simulation 2, based on 2000 replicates with n=200, at selected quantiles. Standard normal error and Exp(1) are considered.

Error	N(0,1)	N(0,1)	Exp(1)	Exp(1)
	MSE	*excess*MSE	MSE	*excess*MSE
q=0.1	0.226	0.6324	-3.236	-6.928
	(-0.817,1.269)	(-2.283,3.547)	(-5.037,-1.435)	(-10.836,-3.020)
q=0.25	1.511	4.757	1.096	3.209
	(0.520,2.503)	(1.682,7.831)	(-0.164,2.355)	(-0.447,6.865)
q=0.5	3.222	10.420	-0.322	-1.001
	(2.257,4.186)	(7.412,13.428)	(-1.450,0.806)	(-4.522,2.520)
q=0.8	1.456	4.468	0.883	2.550
	(0.441,2.471)	(1.408,7.527)	(-0.040,1.807)	(-0.094,5.194)
q=0.9	0.918	2.631	2.289	5.659
	(0.009,1.828)	(0.045,5.217)	(1.388,3.190)	(3.454,7.865)

assesses the fitted quantile curve more accurately. There is little advantage of using the modified check loss when the sample size is small (n=200), as shown in Table 5.7. This appears to be due to the relatively complex shape of the quantile curve and the paucity of observations (20) in the hold-out sample. However, as the sample size grows, and more information is available to fit and validate the model, we can see in Table 5.8 and Table 5.9 that CV.M clearly performs better than CV.

On average, CV and CV.M select the same model about 75% of the time in the two simulations. Thus the reductions in MSE and *excess* MSE are based on only about 25% of the replicates, which makes the CV.M stand out: Given selection of different models, the reductions in MSE and *excess* MSE are substantial.

Table 5.8: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE under simulation 2, based on 2000 replicates with n=500, at selected quantiles. Standard normal error and Exp(1) are considered.

Error	N(0,1)	N(0,1)	Exp(1)	Exp(1)
	MSE	$excess MSE$	MSE	$excess MSE$
q=0.1	0.334	1.056	0.745	2.242
	(-0.388,1.055)	(-1.215,3.326)	(-0.795,2.284)	(-2.353,6.838)
q=0.25	1.348	4.850	3.652	13.981
	(0.663,2.034)	(2.425,7.274)	(2.673,4.631)	(10.472,17.491)
q=0.5	2.203	8.195	1.545	5.512
	(1.457,2.948)	(5.507,10.882)	(0.674,2.417)	(2.451,8.563)
q=0.8	1.607	5.337	1.378	4.419
	(0.905,2.309)	(3.057,7.616)	(0.746,2.009)	(2.425,6.414)
q=0.9	0.372	1.146	1.748	5.007
	(-0.302,1.047)	(-0.921,3.212)	(0.776,2.720)	(2.293,7.721)

Table 5.9: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE under simulation 2, based on 2000 replicates with n=1000, at selected quantiles. Standard normal error and Exp(1) are considered.

Error	N(0,1)	N(0,1)	Exp(1)	Exp(1)
	MSE	$excess MSE$	MSE	$excess MSE$
q=0.1	0.323	1.131	2.556	12.508
	(-0.149,0.794)	(-0.514,2.776)	(1.656,3.456)	(8.372,16.644)
q=0.25	1.493	5.992	3.087	17.114
	(0.911,2.074)	(3.718,8.266)	(2.308,3.865)	(13.126,21.101)
q=0.5	2.916	12.695	2.282	10.507
	(2.328,3.504)	(10.299,15.090)	(1.562,3.000)	(7.340,13.673)
q=0.8	1.193	4.652	0.914	3.269
	(0.484,1.903)	(1.945,2.633)	(0.353,1.475)	(1.290,5.248)
q=0.9	0.759	2.633	0.700	2.227
	(0.184,1.333)	(0.661,4.605)	(0.251,1.149)	(0.810,3.644)

5.3.3 Validation under Quantile Smoothing Splines

In this section, the performance of the check loss and the modified check loss are compared under quantile smoothing splines. Several versions of quantile smoothing splines are suggested, with slightly different forms. Bloomfield and Steiger (1983) and Jones in the discussion of Cole (1988) have proposed estimating a quantile smoothing spline model that minimizes

$$\sum_{i=1}^{n} \rho(y_i - g(x_i)) + \lambda \int (g''(x))^2 dx.$$

Koenker et al. (1994) suggested use of an ℓ_1 roughness penalty, where the above $(g''(x))^2$ is replaced by $|g''(x)|$. Nychka et al. (1995) proposed to round out the corner of the check loss in a small interval $(-\epsilon, \epsilon)$ around zero in order to make the loss function differentiable and thus improve computation. This modification is similar in form to our modified check loss. Two essential differences lie in the fact that the adjustment in Nychka et al. (1995) is chosen to be effectively zero relative to the residuals, and that our modification has an asymmetric window about zero for quantiles other than the median. By adopting Nychka et al. (1995)'s quantile smoothing splines via the function `qsreg(fields)` in the R package, we compare the fitted models selected by the check loss and the modified check loss. Again, σ needed for the window width of the modified check loss is estimated by the standard deviation of the residuals from the mean smoothing spline regression fit.

Data from sinusoid curve with period 1, along with several error distributions are simulated. That is,

$$y_i = sin(2\pi x_i) + \epsilon_i, i = 1, ..., n,$$

Table 5.10: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE based on 1000 replicates with n=500, at several quantiles. Normal error with mean zero, and standard deviations 0.2 is considered.

Reduction in	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5
MSE	4.826	6.071	9.767	9.386	10.113
	(3.274,6.378)	(4.174,7.967)	(7.356,12.178)	(7.094,11.678)	(8.121,12.105)
excessMSE	13.155	17.863	28.414	29.688	30.139
	(9.214,17.096)	(12.724,23.002)	(22.545,34.284)	(23.493,35.883)	(25.056,35.222)

where x_i's are *iid* from the standard uniform distribution, and ϵ_i's are *iid* from some error distribution. Error distributions considered are normal, t, shifted gamma, and shifted exponential distribution with median 0, and standard deviation 0.2. The t distributions, with 5 and 10 degrees of freedom, are scaled to have standard deviation 0.2. A fine grid search of the smoothing parameter is conducted to select the 'best' value by CV and CV.M. As in the previous section, the performance of the two validation functions is judged by the percentage reduction in MSE and *excess* MSE. To calculate the *excess* MSE, a base model is defined by the minimum MSE across the fine grid of smoothing parameter value given a data set. The simulation consists of 1000 replicates with sample size 500, 1000 and 2000.

The simulation results from a normal error distribution at several quantiles are shown in Tables 5.10, 5.11, and 5.12. The results for $q > 0.5$ are similar to those for $q < 0.5$, due to the symmetry of the normal distribution. There are considerable reductions when CV.M is employed, indicating that the averaging effect of CV.M when evaluating the residuals near the target quantile induces better assessment of

Table 5.11: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE based on 1000 replicates with n=1000, at several quantiles. Normal error with mean zero, and standard deviations 0.2 is considered.

Reduction in	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5
MSE	3.433	7.163	9.646	10.093	9.784
	(1.931,4.935)	(5.232,9.095)	(7.359,11.933)	(8.034,12.152)	(7.514,12.054)
$excessMSE$	10.594	21.932	31.035	32.839	34.744
	(6.174,15.014)	(16.722,27.142)	(24.958,37.112)	(27.317,38.361)	(28.658,40.830)

Table 5.12: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE based on 1000 replicates with n=2000, at several quantiles. Normal error with mean zero, and standard deviations 0.2 is considered.

Reduction in	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5
MSE	3.385	5.480	5.559	7.580	7.725
	(2.010,4.760)	(3.927,7.033)	(3.803,7.316)	(5.884,9.276)	(6.138,9.313)
$excessMSE$	11.465	20.607	23.009	31.588	32.421
	(7.129,15.800)	(15.457,25.757)	(16.824,29.194)	(25.972,37.204)	(27.257,37.584)

a fitted model. Detailed results for the other error distributions are shown in the appendix.

Overall, quantiles at high density lead to greater reduction in MSE. This tendency holds not only within an error distribution, but also between the distributions. For example, MSE reductions grow as we move from a sharp tail (normal) to thicker tails (t with df=10 and 5) at the tail quantiles. For the same reason, the extent of reductions near the median becomes smaller as we change from the normal distribution to a t with 10 degrees of freedom and further to a t with 5 degrees of freedom.

The same smoothing parameters are selected by CV and CV.M for roughly 50% to 75 % of the replicates. The remaining 25% to 50% of the replicates account for the discrepancy in the assessment of the fitted curve under CV and CV.M. Examination of the cases with a difference shows that the improvement of CV.M is mainly due to not overfitting. In contrast, CV sometimes severely undersmooths.

The undersmoothing with CV is partly demonstrated by the plots of the fitted functions which produce maximum or minimum values of $MSE_{CV.M}/MSE_{CV}$. The Figure 5.2 shows the selected models by CV and CV.M which yield the minimum value of $MSE_{CV.M}/MSE_{CV}$. In this figure, CV is poor compared to CV.M, and we can clearly observe that CV overfits the data. To the contrary, Figure 5.3 shows the data sets where CV.M is the poorest, relative to CV. Here, it seems that CV.M sometimes slightly oversmooths data. To our eyes, the mild oversmoothing in Figure 5.3 is preferable to the severe undersmoothing in the Figure 5.2. The earlier percentage reduction in MSE calculations confirm this impression. Figure 5.4 presents the distributions of the estimated smoothing parameter $\hat{\lambda}$ (in log scale) that are centered around $\hat{\lambda}$ from the base model. In Figure 5.4, only those $\hat{\lambda}$ pairs with different

Table 5.13: MSE values ($\times$ 1000) are decomposed to variance and squared bias, based on 1000 replicates with n=500, at selected quantiles. Standard normal error is considered.

Method	CV	CV.M	CV	CV.M	CV	CV.M
	variance	variance	squared bias	squared bias	MSE	MSE
q=0.1	2.080	1.949	0.136	0.142	2.216	2.090
q=0.2	1.556	1.427	0.078	0.086	1.633	1.513
q=0.3	1.297	1.173	0.079	0.080	1.375	1.254
q=0.4	1.229	1.108	0.062	0.066	1.290	1.173
q=0.5	1.205	1.032	0.065	0.076	1.270	1.108
q=0.6	1.218	1.093	0.069	0.073	1.287	1.166
q=0.7	1.292	1.133	0.077	0.085	1.368	1.218
q=0.8	1.497	1.350	0.090	0.101	1.588	1.451
q=0.9	2.028	1.906	0.125	0.126	2.152	2.032

values are included. The long lower tail of the $\hat{\lambda}$ from CV is another indication of undersmoothing.

As stated earlier, the benefits of CV.M resulted from the bias-variance tradeoff of estimating future error. This fact is readily verified when MSE value is decomposed to variance and squared bias. Table 5.13 shows the decomposition of variance and squared bias under standard normal error distribution. We can see that the reduction in variance is much larger than the increase in squared bias, thus leading to smaller MSE values. (Results for the other error distributions are in the appendix.)

5.4 Conclusion

This chapter shows that cross validation with the check loss in quantile regression is reasonable, but that the technique can be improved by use of the modified check loss function developed in Chapter 4. The proposed method has shown its superiority

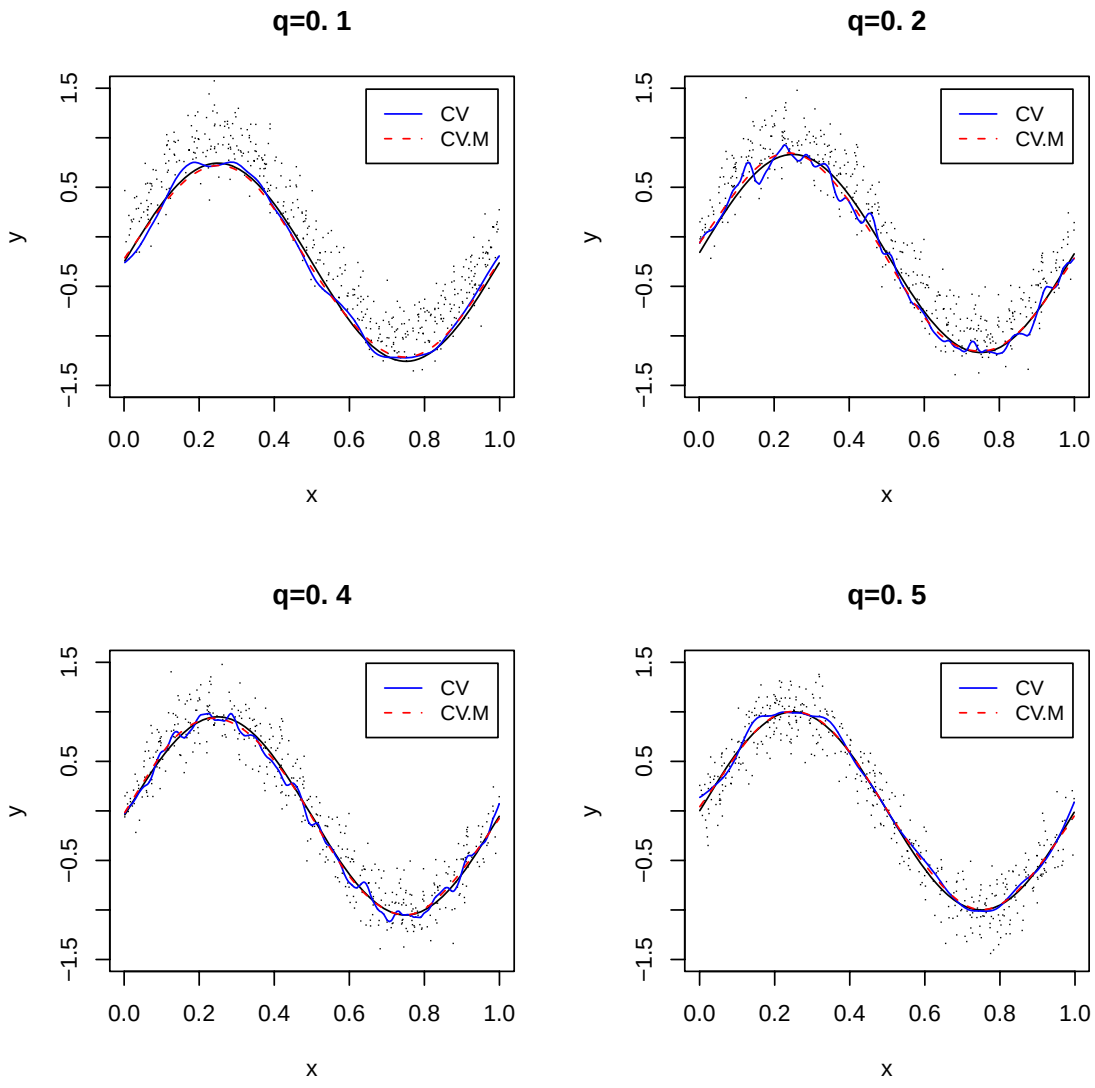

Figure 5.2: 'Best' cases of the fitted models selected by CV.M when compared with CV under a normal error distribution, with n=500.

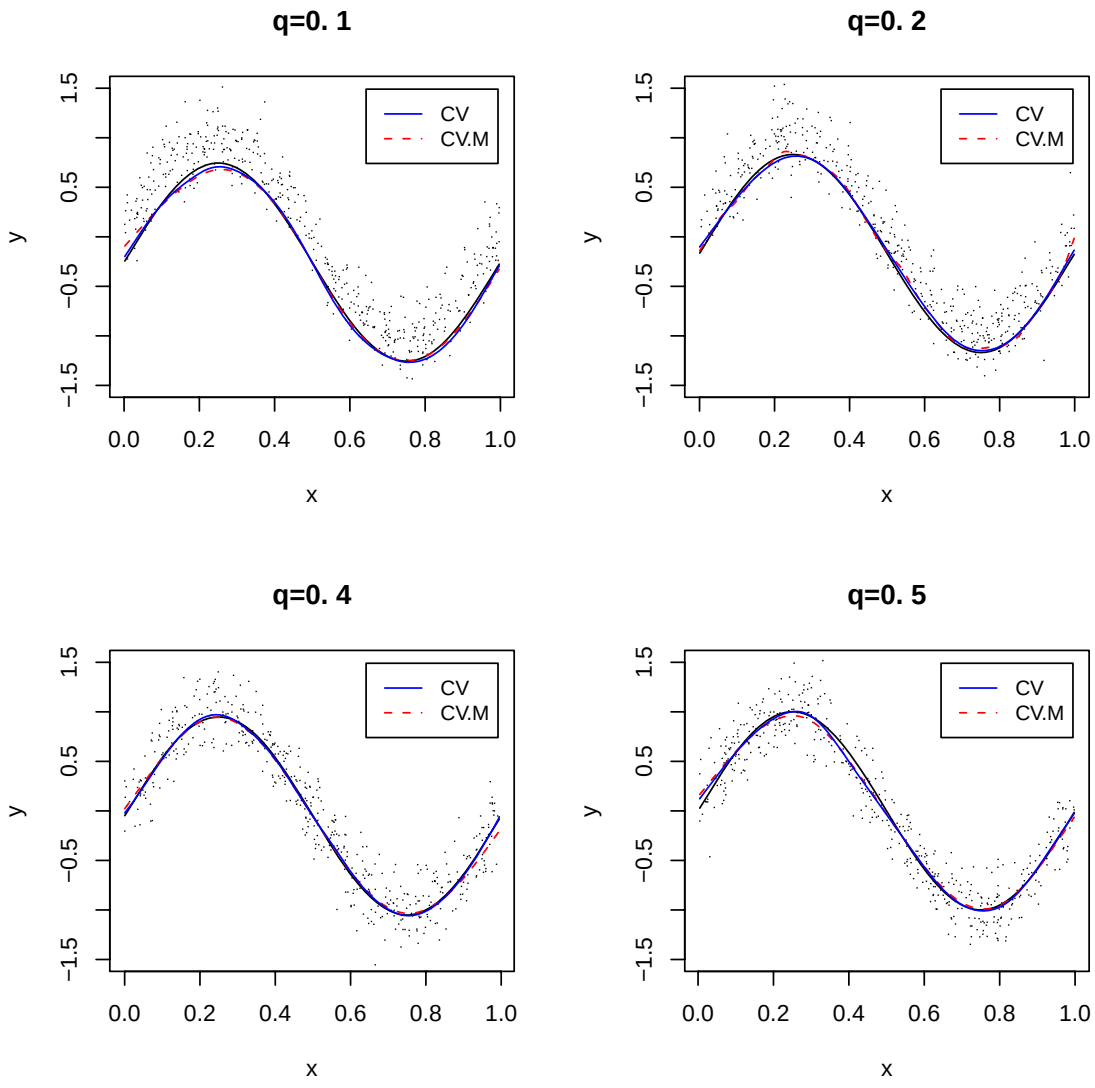

Figure 5.3: 'Worst' cases of the fitted models selected by CV.M when compared with CV under a normal error distribution, with n=500.

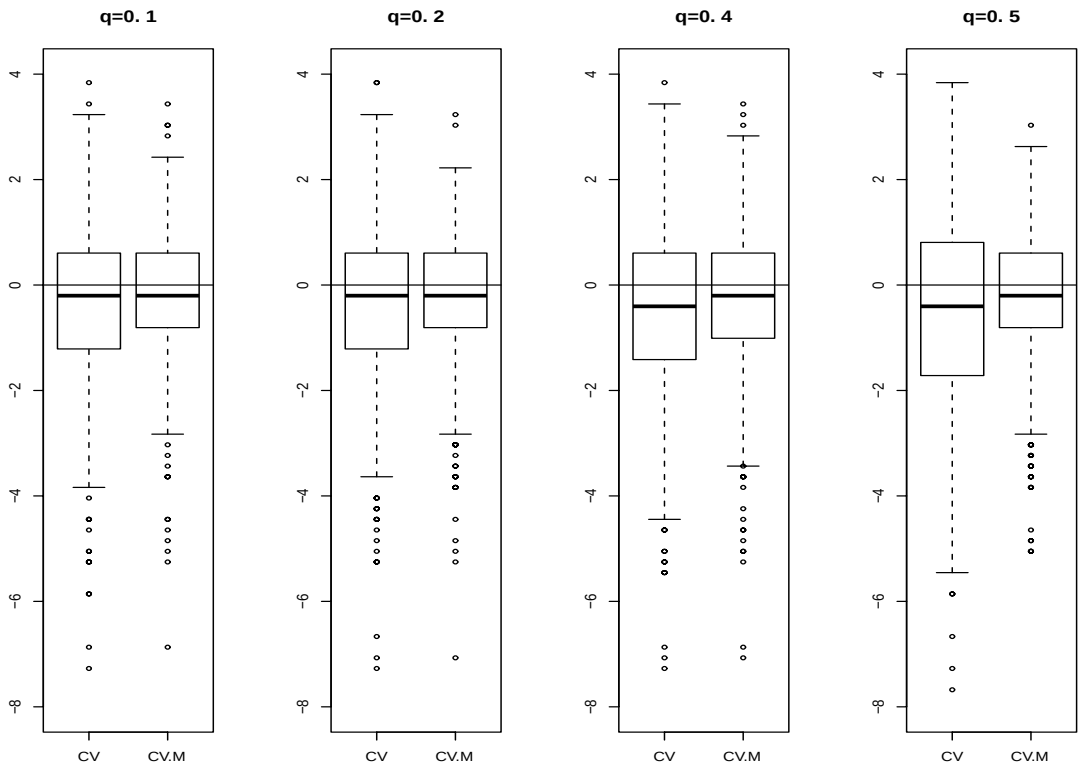

Figure 5.4: Distribution of $\hat{\lambda}$ from CV and CV.M after subtracting $\hat{\lambda}$ from the base model (in the log scale) under a normal error distribution, with n=500. Only those $\hat{\lambda}$ pairs with different values are included.

in estimating future error through diverse simulations. The advantage is mainly due to the bias-variance tradeoff in estimating future error. This is reflected in the model selection, and a range of experiments reveals that the check loss may suffer from overfitting the data, while the modified check loss reduces overfitting thus leading to better model selection.

CHAPTER 6

CASE-SPECIFIC PARAMETERS IN CLASSIFICATION

6.1 Application to Classification Problems

Now suppose that y_i's indicate binary outcomes. For modeling and prediction of the binary responses, we mainly consider margin-based procedures such as logistic regression, support vector machines (Vapnik; 1998), and boosting (Freund and Schapire; 1997). These procedures can be modified by the addition of case indicators.

6.1.1 Logistic Regression

Although it is customary to label a binary outcome as 0 or 1 in logistic regression, we instead adopt the symmetric labels of $\{-1, 1\}$ for y_i's. The symmetry facilitates comparison of different classification procedures. Logistic regression takes the negative log likelihood as a loss for estimation of logit $f(x) = \log[p(x)/(1-p(x))]$. The loss, $\mathcal{L}(y, f(x)) = \log[1 + \exp(-yf(x))]$, can be viewed as a function of the so-called margin, $yf(x)$. This functional margin of $yf(x)$ is a pivotal quantity for defining a family of loss functions in classification similar to the residual in regression.

Logistic regression can be modified with case indicators:

$$L(\beta_0, \beta, \gamma) = \sum_{i=1}^{n} \log\big(1 + \exp(-y_i\{f(x_i; \beta_0, \beta) + \gamma_i\})\big) + \lambda_\gamma \|\gamma\|_1, \qquad (6.1)$$

where $f(x;\beta_0,\beta) = \beta_0 + x^\top\beta$. When it is clear in the context, $f(x)$ will be used as a short notation for $f(x;\beta_0,\beta)$, a discriminant function in general. For fixed $\hat{\beta}_0$ and $\hat{\beta}$, γ_i is determined by minimizing

$$\log\big(1+\exp(-y_i\{f(x_i;\hat{\beta}_0,\hat{\beta})+\gamma_i\})\big)+\lambda_\gamma|\gamma_i|.$$

First note that the minimizer γ_i must have the same sign as y_i. Letting $\tau = yf$ and assuming that $0<\lambda_\gamma<1$, we have $\arg\min_{\gamma\geq 0}\log\big(1+\exp(-\tau-\gamma)\big)+\lambda_\gamma|\gamma| = \log\{(1-\lambda_\gamma)/\lambda_\gamma\}-\tau$ if $\tau\leq\log\{(1-\lambda_\gamma)/\lambda_\gamma\}$, and 0 otherwise. This yields a truncated negative log likelihood given by

$$\mathcal{L}(y,f(x)) = \begin{cases} \log\big(1+\lambda_\gamma/(1-\lambda_\gamma)\big) & \text{if } yf(x)\leq\log\{(1-\lambda_\gamma)/\lambda_\gamma\},\\ \log\big(1+\exp(-yf(x))\big) & \text{otherwise}\end{cases}$$

as the γ-adjusted loss. This adjustment is reminiscent of Pregibon (1982)'s proposal for robust logistic regression by tapering the deviance function so as to downweight extreme observations. See Figure 6.1 (b) for the γ-adjusted loss (the dashed line), where $\eta_\lambda := \log\{(1-\lambda_\gamma)/\lambda_\gamma\}$, and it is a decreasing function of λ_γ. λ_γ determines the level of truncation of the loss. As λ_γ tends to 1, there is no truncation. Figure 6.1 (b) also shows the effective loss (the solid line) for the ℓ_1 adjustment, which linearizes the negative log likelihood below η_λ.

6.1.2 Large Margin Classifiers

With the symmetric class labels, the foregoing characterization of the case-specific parameter γ in logistic regression can be easily generalized to various margin-based classification procedures. In classification, potential outliers are those cases with large negative margins. Let $g(\tau)$ be a loss function of the margin $\tau = yf(x)$. The following proposition holds for a general family of loss functions.

Proposition 8 *Suppose that g is convex and monotonically decreasing in τ, and g' is continuous. Then, for $\lambda_\gamma < -\lim_{\tau\to-\infty} g'(\tau)$,*

$$\hat{\gamma} = \arg\min_{\gamma} \; g(\tau+\gamma) + \lambda_\gamma|\gamma| = \begin{cases} g'^{-1}(-\lambda_\gamma) - \tau & \text{for } \tau \le g'^{-1}(-\lambda_\gamma) \\ 0 & \text{for } \tau > g'^{-1}(-\lambda_\gamma). \end{cases}$$

The proof is straightforward. Examples of the margin-based loss g satisfying the assumption include the exponential loss $g(\tau) = \exp(-\tau)$ in boosting, the squared hinge loss $g(\tau) = \{(1-\tau)_+\}^2$ in the support vector machine (Lee and Mangasarian; 2001), and the negative log likelihood $g(\tau) = \log(1+\exp(-\tau))$ in logistic regression. Although their theoretical targets are different, all the loss functions are truncated above for large negative margins when adjusted by γ. Thus, the effective loss $\mathcal{L}_{\lambda_\gamma}(y, f(x)) = g(yf(x)+\hat{\gamma})+\lambda_\gamma|\hat{\gamma}|$ is obtained by linearizing g for $yf(x) < g'^{-1}(-\lambda_\gamma)$.

However, the effect of $\hat{\gamma}$-adjustment depends on the form of g, and hence on the classification method. For boosting, $\hat{\gamma} = -\log\lambda_\gamma - yf(x)$ if $yf(x) \le -\log\lambda_\gamma$, and 0 otherwise. This gives

$$L(\beta_0, \beta, \hat{\gamma}) = \sum_{i=1}^{n} \exp(-y_i f(x_i; \beta_0, \beta) - \hat{\gamma}_i) = \sum_{i=1}^{n} \exp(-\hat{\gamma}_i)\exp(-y_i f(x_i; \beta_0, \beta)).$$

So, finding β_0 and β given $\hat{\gamma}$ amounts to weighted boosting, where the positive case-specific parameters $\hat{\gamma}_i$ downweight the corresponding cases by $\exp(-\hat{\gamma}_i)$. For the squared hinge loss in the SVM, $\hat{\gamma} = 1 - yf(x) - \lambda_\gamma/2$ if $yf(x) \le 1 - \lambda_\gamma/2$, and is 0 otherwise. A positive case-specific parameter $\hat{\gamma}_i$ has the effect of relaxing the margin requirement, that is, lowering the joint of the hinge individually. It allows the associated slack variable to be smaller in the primal formulation. Accordingly, the adjustment affects the coefficient of the linear term in the dual formulation of the quadratic programming problem.

As a related approach to robust classification, Wu and Liu (2007) propose truncation of margin-based loss functions and study some theoretical properties to ensure classification consistency. Similarity exists between our proposed adjustment of a loss function with γ and truncation of the loss at some point. However, it is the linearization of a margin-based loss function on the negative side that gives its effective loss, and the process essentially defines the modification in our approach. Linearization is more conducive to computation for modified procedures than is truncation. Furthermore, application of the result in Bartlett et al. (2006) shows that the linearized loss functions satisfy sufficient conditions for classification consistency, namely Fisher consistency, which is the main property investigated by Wu and Liu (2007) for truncated loss functions.

6.1.3 Support Vector Machines

As a special case of a large margin classifier, the linear support vector machine (SVM) looks for the optimal hyperplane $f(x; \beta_0, \beta) = \beta_0 + x^\top \beta = 0$ minimizing

$$L_\lambda(\beta_0, \beta) = \sum_{i=1}^{n} \Big[1 - y_i f(x_i; \beta_0, \beta)\Big]_+ + \frac{\lambda}{2}\|\beta\|_2^2, \tag{6.2}$$

where $[t]_+ = \max(t, 0)$ and $\lambda > 0$ is a regularization parameter. Since the hinge loss for the SVM, $g(\tau) = (1 - \tau)_+$, is piecewise linear, its linearization with $\|\gamma\|_1$ is void, indicating that it has little need of further robustification. Instead, we consider modification of the hinge loss with $\|\gamma\|_2^2$. This modification is expected to have the same effect of improving efficiency as in quantile regression.

Using the case indicators z_i and their coefficients γ_i, we modify (6.2), arriving at the problem of minimizing

$$L(\beta_0, \beta, \gamma) = \sum_{i=1}^{n} \Big[1 - y_i\{f(x_i; \beta_0, \beta) + \gamma_i\}\Big]_+ + \frac{\lambda_\beta}{2}\|\beta\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma\|_2^2. \qquad (6.3)$$

For fixed $\hat{\beta}_0$ and $\hat{\beta}$, the minimizer $\hat{\gamma}$ of $L(\hat{\beta}_0, \hat{\beta}, \gamma)$ is obtained by solving the decoupled optimization problem of $\min_\gamma [1 - y_i f(x_i; \hat{\beta}_0, \hat{\beta}) - y_i\gamma_i]_+ + \frac{\lambda_\gamma}{2}\gamma_i^2$ for each γ_i. With an argument similar to that for logistic regression, the minimizer $\hat{\gamma}_i$ should have the same sign as y_i. Let $\xi = 1 - yf$. A simple calculation shows that

$$\arg\min_{\gamma \geq 0} \; [\xi - \gamma]_+ + \frac{\lambda_\gamma}{2}\gamma^2 = \begin{cases} 0 & \text{if } \xi \leq 0 \\ \xi & \text{if } 0 < \xi < 1/\lambda_\gamma \\ 1/\lambda_\gamma & \text{if } \xi \geq 1/\lambda_\gamma. \end{cases}$$

Hence, the increase in margin $y_i\hat{\gamma}_i$ due to inclusion of γ is given by

$$\{1 - y_i f(x_i)\} I(0 < 1 - y_i f(x_i) < \frac{1}{\lambda_\gamma}) + \frac{1}{\lambda_\gamma} I(1 - y_i f(x_i) \geq \frac{1}{\lambda_\gamma}).$$

The γ-adjusted hinge loss is $\mathcal{L}(y, f(x)) = [1 - 1/\lambda_\gamma - yf(x)]_+$ with the hinge lowered by $1/\lambda_\gamma$ as shown in Figure 6.1 (c) (the dashed line). The effective loss (the solid line in the figure) is then given by a smooth function with the joint replaced with a quadratic piece between $1 - 1/\lambda_\gamma$ and 1 and linear beyond the interval.

6.2 Conclusion

In this Chapter, we developed theoretical basics of employing case-specific parameters in some of classification methods. As in regression problems, the form of penalty for case-specific parameters should be tailored to the needs of a problem. In logistic regression, ℓ_1 type penalty may modify the procedure to be more robust, whereas ℓ_2 type penalty in SVM will adjust the SVM to be more efficient. The modifications in this chapter have not yet fully investigated except their mathematical formula, thus are left as future research.

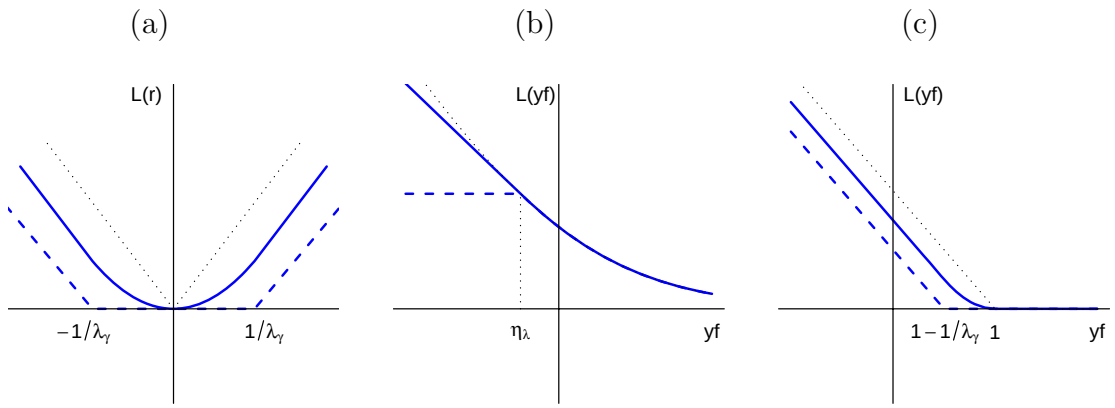

Figure 6.1: Modification of (a) absolute deviation loss for median regression with ℓ_2 penalty, (b) negative log likelihood for logistic regression with ℓ_1 penalty, and (c) hinge loss for the support vector machine with ℓ_2 penalty. The solid lines are for the effective loss, the dashed lines are for the γ-adjusted loss, and the dotted lines are for the original loss in each panel.

CHAPTER 7

CONCLUSION

This thesis has focused on theoretical developments and applications of case-specific parameters in various statistical modeling procedures. The developed methods and applications with case-specific parameters illustrate a new approach to statistical modeling. The additional case-specific parameters create a supersaturated model, and so the regularization or penalization method is adopted. This requires an extra penalty for the case-specific parameters. Depending on the characteristics of a modeling procedure, the form of the penalty for the case-specific parameters is chosen to improve inference. Modeling procedures that lack robustness can be modified to be robust, while robust models can be made more efficient. For example, an ℓ_1 type penalty in the LASSO problem results in the robust LASSO, while an ℓ_2 type penalty in quantile regression ends with efficient quantile regression. In the classification problem, an ℓ_1 type penalty in logistic regression leads to robust logistic regression, and employing an ℓ_2 type penalty in the support vector machine is expected to produce a more efficient classifier (although this has not been fully investigated). Since the additional parameters cause a modification of an existing procedure, the extent of modification is an important practical question. We provided some heuristic rules to guide the amount of modification.

The strength of case-specific parameters and their penalization lies in its broad application as a tool for improving current methods. In this thesis, several regression and classification methods are considered to illustrate the usage of case-specific parameters. Moreover, modified quantile regression naturally brings us to modified cross validation, as illustrated in Chapter 5. These examples illustrate a variety of problems to which case-specific regularization can be applied, and they show the ease with which novel techniques can be developed. We believe there are many more situations where case-specific parameters can play an important role in statistics, and we look forward to continued development of the methods.

APPENDIX A

APPENDIX

The details of the following tables are described in Chapter 5.3.3.

Table A.1: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE based on 1000 replicates with n=500, at several quantiles. Scaled t distributions (to maintain 0.2 standard deviations) with 10 degrees of freedom (2 columns in the left) and 5 degrees of freedom (2 columns in the right) are considered.

Error	t(df=10)	t(df=10)	t(df=5)	t(df=5)
	MSE	$excess MSE$	MSE	$excess MSE$
q=0.1	4.812	13.548	3.255	9.195
	(3.170,6.455)	(9.197,17.899)	(1.912,4.599)	(5.524,12.866)
q=0.2	5.726	16.684	7.034	20.325
	(3.728,7.725)	(11.321,22.048)	(4.918,9.150)	(14.837,25.813)
q=0.3	8.513	25.091	7.140	22.137
	(6.330,10.695)	(19.525,30.656)	(5.324,8.955)	(17.138,27.136)
q=0.4	9.236	28.433	7.749	25.108
	(6.918,11.553)	(22.521,34.346)	(5.613,9.886)	(19.061,31.154)
q=0.5	13.114	37.778	9.345	29.105
	(10.246,15.983)	(31.415,44.140)	(7.505,11.186)	(24.137,34.073)
q=0.6	9.843	30.402	8.240	25.090
	(7.430,12.257)	(24.243,36.562)	(6.233,10.247)	(19.647,30.532)
q=0.7	9.109	26.441	7.341	23.286
	(6.675,11.544)	(20.341,32.540)	(5.451,9.230)	(17.955, 28.618)
q=0.8	5.326	15.276	6.195	19.506
	(3.419,7.237)	(10.226,20.326)	(4.453,7.937)	(14.503,24.508)
q=0.9	4.680	12.629	3.846	10.517
	(2.917,6.443)	(8.102,17.157)	(2.028,5.663)	(5.746,15.288)

Table A.2: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE based on 1000 replicates with n=1000, at several quantiles. Scaled t distributions (to maintain 0.2 standard deviations) with 10 degrees of freedom (2 columns in the left) and 5 degrees of freedom (2 columns in the right) are considered.

Error	t(df=10)	t(df=10)	t(df=5)	t(df=5)
	MSE	$excess MSE$	MSE	$excess MSE$
q=0.1	3.429	10.702	5.334	15.682
	(1.893,4.964)	(6.196,15.207)	(3.603,7.064)	(10.983,20.381)
q=0.2	5.320	16.189	4.404	15.525
	(3.761,6.879)	(11.735,20.643)	(2.872,5.935)	(10.524,20.526)
q=0.3	6.660	23.246	7.133	24.370
	(4.784,8.536)	(17.497,28.995)	(5.189,9.077)	(18.676,30.063)
q=0.4	7.380	25.829	6.802	25.894
	(5.776,8.984)	(20.925,30.732)	(5.021,8.583)	(20.128,31.661)
q=0.5	10.016	33.115	8.457	30.922
	(7.502,12.531)	(26.538,39.692)	(6.221,10.693)	(24.407,37.437)
q=0.6	11.019	35.712	8.913	31.659
	(8.298,13.741)	(29.065,42.359)	(6.705,11.121)	(25.307,38.011)
q=0.7	7.396	25.392	9.245	31.733
	(5.459,9.334)	(19.634,31.150)	(7.132,11.357)	(25.798,37.668)
q=0.8	7.124	23.124	7.324	23.502
	(5.414,8.834)	(18.278,27.969)	(5.082,9.566)	(17.246,29.757)
q=0.9	4.733	14.529	3.733	11.117
	(3.068,6.399)	(9.788,19.270)	(2.352,5.114)	(7.224,15.010)

Table A.3: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE based on 1000 replicates with n=2000, at several quantiles. Scaled t distributions (to maintain 0.2 standard deviation) with 10 degrees of freedom (2 columns in the left) and 5 degrees of freedom (2 columns in the right) are considered.

Error	t(df=10)	t(df=10)	t(df=5)	t(df=5)
	MSE	$excess MSE$	MSE	$excess MSE$
q=0.1	3.897	12.677	5.094	16.060
	(2.302,5.492)	(7.892,17.462)	(3.112,7.077)	(10.468,21.651)
q=0.2	6.697	24.591	6.194	22.125
	(4.830,8.565)	(18.677,30.506)	(4.430,7.959)	(16.639,27.611)
q=0.3	7.831	27.561	7.007	26.660
	(5.922,9.741)	(21.960,33.163)	(5.383,8.631)	(21.359,31.961)
q=0.4	7.246	28.492	6.255	26.826
	(5.650,8.842)	(23.124,33.859)	(4.614,7.896)	(20.991,32.661)
q=0.5	8.911	34.958	6.799	28.403
	(6.834,10.987)	(28.727,41.188)	(5.093,8.504)	(22.736,34.070)
q=0.6	6.875	28.379	6.787	26.948
	(5.271,8.479)	(23.013,33.746)	(5.188,8.387)	(21.615,32.281)
q=0.7	5.402	21.895	6.275	24.185
	(3.876,6.927)	(16.450,27.340)	(4.638,7.912)	(18.607,29.762)
q=0.8	7.035	23.913	6.895	25.186
	(4.883,9.187)	(17.611,30.215)	(4.877,8.912)	(18.969,31.403)
q=0.9	3.616	11.857	4.332	14.884
	(2.359,4.872)	(7.934,15.780)	(2.636,6.027)	(9.476,20.292)

Table A.4: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE based on 1000 replicates with n=500, at several quantiles. Shifted gamma(2, scale=2) (2 columns in the left) and shifted Exp(0.2) distribution (2 columns in the right) are considered.

Error	Gamma	Gamma	Exp	Exp
	MSE	$excess MSE$	MSE	$excess MSE$
q=0.1	5.442	17.342	7.210	23.040
	(3.847,7.037)	(17.889,27.356)	(4.872,9.548)	(16.431,29.649)
q=0.2	8.633	27.466	9.444	30.248
	(6.594,10.672)	(25.220,36.987)	(7.366,11.522)	(24.499,35.998)
q=0.3	9.455	30.930	6.560	21.413
	(7.568,11.342)	(29.276,40.179)	(4.390,8.731)	(14.888,27.939)
q=0.4	11.253	38.038	8.888	26.838
	(9.324,13.182)	(35.729,47.087)	(6.080,11.695)	(19.368,34.308)
q=0.5	11.544	35.066	9.792	28.188
	(9.156,13.932)	(34.793,48.159)	(7.390,12.195)	(22.124,34.252)
q=0.6	11.720	34.181	9.140	26.749
	(9.173,14.266)	(30.680,41.593)	(6.735,11.544)	(20.364,33.135)
q=0.7	10.672	28.828	14.648	37.073
	(8.277,13.068)	(24.564,35.592)	(11.815,17.482)	(31.102,43.044)
q=0.8	7.881	21.891	10.611	27.558
	(5.906,9.856)	(15.505,25.842)	(7.970,13.253)	(21.665,33.450)
q=0.9	3.093	7.811	3.879	9.639
	(1.833,4.353)	(4.543,12.730)	(2.452,5.306)	(6.217,13.062)

Table A.5: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE based on 1000 replicates with n=1000, at several quantiles. Shifted gamma(2, scale=2) (2 columns in the left) and shifted Exp(0.2) distribution (2 columns in the right) are considered.

Error	Gamma	Gamma	Exp	Exp
	MSE	$excess MSE$	MSE	$excess MSE$
q=0.1	4.690	16.668	8.748	31.715
	(3.159,6.222)	(11.614,21.721)	(6.937,10.560)	(26.313,37.117)
q=0.2	9.329	34.174	10.591	38.718
	(7.498,11.161)	(28.881,39.467)	(8.633,12.550)	(32.809,44.627)
q=0.3	10.762	38.133	11.895	40.268
	(8.252,13.272)	(31.598,44.667)	(9.314,14.476)	(33.820,46.717)
q=0.4	9.712	35.292	8.487	29.434
	(7.784,11.640)	(29.780,40.804)	(6.478,10.496)	(23.578,35.291)
q=0.5	10.363	38.317	9.733	34.347
	(8.400,12.326)	(32.955,43.679)	(7.700,11.766)	(28.637,40.056)
q=0.6	12.717	40.278	11.675	39.170
	(10.124,15.311)	(34.216,46.341)	(9.779,13.570)	(34.276,44.064)
q=0.7	10.617	32.521	12.309	40.158
	(8.298,12.941)	(26.724,38.318)	(10.041,14.577)	(34.720,45.595)
q=0.8	7.329	21.585	6.308	19.967
	(5.263,9.396)	(16.093,27.078)	(4.585,8.031)	(15.043,24.889)
q=0.9	4.249	11.620	2.361	6.809
	(2.631,5.866)	(7.430,15.810)	(1.480,3.241)	(4.333,9.286)

Table A.6: Point estimates and approximate 95% confidence intervals for percentage reduction in mean MSE and mean *excess* MSE based on 1000 replicates with n=2000, at several quantiles. Shifted gamma(2, scale=2) (2 columns in the left) and shifted Exp(0.2) distribution (2 columns in the right) are considered.

Error	Gamma	Gamma	Exp	Exp
	MSE	$excess MSE$	MSE	$excess MSE$
q=0.1	5.872	22.622	9.051	37.441
	(4.461,7.284)	(17.889,27.356)	(7.211,10.892)	(31.763,43.119)
q=0.2	7.927	31.103	9.099	41.324
	(6.056,9.798)	(25.220,36.987)	(7.334,10.863)	(35.832,46.816)
q=0.3	8.162	34.727	10.675	45.391
	(6.507,9.818)	(29.276,40.179)	(8.861,12.488)	(40.188,50.594)
q=0.4	10.002	41.408	10.205	43.416
	(8.069,11.935)	(35.729,47.087)	(8.512,11.899)	(38.408,48.425)
q=0.5	10.391	41.476	8.899	37.997
	(8.000,12.786)	(34.793,48.159)	(6.875,10.924)	(31.762,44.232)
q=0.6	9.168	36.137	8.230	32.835
	(7.358,10.979)	(30.680,41.593)	(6.231,10.228)	(26.489,39.182)
q=0.7	7.863	30.078	7.737	29.282
	(6.092,9.634)	(24.564,35.592)	(6.077,9.396)	(23.995,34.570)
q=0.8	5.737	20.674	6.300	20.941
	(4.107,7.368)	(15.505,25.842)	(4.715,7.884)	(16.186,25.695)
q=0.9	2.739	8.637	1.855	5.573
	(1.383,4.094)	(4.543,12.730)	(0.765,2.945)	(2.362,8.785)

Table A.7: MSE values are decomposed to variance and squared bias, which multiplied by 1000 based on 1000 replicates with n=500, at selected quantiles. Scaled t distribution (to maintain 0.2 standard deviation) with 10 degrees of freedom is considered.

Method	CV	CV.M	CV	CV.M
	variance	variance	squared bias	squared bias
q=0.1	2.232	2.102	0.133	0.142
q=0.2	1.436	1.334	0.067	0.078
q=0.3	1.200	1.051	0.056	0.061
q=0.4	1.070	0.957	0.061	0.060
q=0.5	1.021	0.892	0.049	0.056
q=0.6	1.072	0.954	0.052	0.059
q=0.7	1.170	1.058	0.061	0.070
q=0.8	1.446	1.298	0.078	0.088
q=0.9	2.084	1.961	0.141	0.152

Table A.8: MSE values are decomposed to variance and squared bias, which multiplied by 1000 based on 1000 replicates with n=500, at selected quantiles. Scaled t distribution (to maintain 0.2 standard deviation) with 5 degrees of freedom is considered.

Method	CV	CV.M	CV	CV.M
	variance	variance	squared bias	squared bias
q=0.1	2.175	2.061	0.156	0.174
q=0.2	1.201	1.112	0.072	0.087
q=0.3	0.921	0.867	0.057	0.061
q=0.4	0.806	0.744	0.048	0.055
q=0.5	0.758	0.697	0.046	0.052
q=0.6	0.795	0.712	0.045	0.053
q=0.7	0.922	0.854	0.059	0.063
q=0.8	1.158	1.083	0.082	0.091
q=0.9	2.140	2.037	0.146	0.168

Table A.9: MSE values are decomposed to variance and squared bias, which multiplied by 1000 based on 1000 replicates with n=500, at selected quantiles. Shifted gamma error distribution is considered.

Method	CV	CV.M	CV	CV.M
	variance	variance	squared bias	squared bias
q=0.1	0.353	0.333	0.039	0.039
q=0.2	0.470	0.430	0.043	0.045
q=0.3	0.601	0.547	0.048	0.049
q=0.4	0.723	0.666	0.055	0.057
q=0.5	0.904	0.823	0.056	0.065
q=0.6	1.196	1.058	0.062	0.074
q=0.7	1.620	1.454	0.092	0.113
q=0.8	2.370	2.181	0.130	0.147
q=0.9	4.398	4.214	0.305	0.340

Table A.10: MSE values are decomposed to variance and squared bias, which multiplied by 1000 based on 1000 replicates with n=500, at selected quantiles. Shifted Exp(0.2) error distribution is considered.

Method	CV	CV.M	CV	CV.M
	variance	variance	squared bias	squared bias
q=0.1	0.107	0.103	0.018	0.017
q=0.2	0.222	0.210	0.024	0.023
q=0.3	0.370	0.349	0.030	0.029
q=0.4	0.541	0.505	0.037	0.038
q=0.5	0.773	0.723	0.045	0.051
q=0.6	1.125	1.011	0.057	0.074
q=0.7	1.682	1.475	0.071	0.104
q=0.8	2.671	2.402	0.123	0.158
q=0.9	5.527	5.248	0.282	0.307

BIBLIOGRAPHY

Baayen, R. (2007). *Analyzing Linguistic Data: a practical introduction to statistics*, Cambridge University Press, Cambridge, England.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D. and Yap, M. (2004). Visual word recognition of single-syllable words, *Journal of Experimental Psychology* **133**: 283–316.

Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds, *Journal of the American Statistical Association* **101**(473): 138–156.

Bassett, G. and Koenker, R. (1978). Asymptotic theory of least absolute error regression, *Journal of the American Statistical Association* **73**(363): 618–622.

Belloni, A. and Chernozhukov, V. (2009). ℓ_1-penalized quantile regression in high dimensional sparse models, *MIT Department of Economics Working Paper No. 09-11* .

Bickel, P. J. and Li, b. (2006). Regularization in statistics, *Test* **15**(2).

Bloomfield, P. and Steiger, W. (1983). *Least Ablolute Deviations: Theory, Applications, and Algorithms*, Boston, Birkhäuser.

Cole, T. (1988). Fitting smoothed centile curves to reference data, *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **151**(3): 385–418.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation, *Journal of the American Statistical Association* **78**(382): 316–331.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss, *Statistica Sinica* **1**(1): 93–125.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics* **32**(2): 407–499.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method, *Journal of the American Statistical Association* **92**(438): 548–560.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools, *Technometrics* **35**(2): 109–135.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* **55**(1): 119–139.

Geisser, S. (1975). The predictive sample reuse method with applications, *Journal of the American Statistical Association* **70**(350): 320–328.

Gu, C. (1992). Cross-validating non-Gaussian data, *Journal of Computational and Graphical Statistics* **1**: 169–179.

Gunn, S. R. and Kandola, J. S. (2002). Structural modelling with sparse kernels, *Machine Learning* **48**(115–136).

Hans, C. (2009). Bayesian lasso regression, *Biometrika* **96**(4): 835–845.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall/CRC.

He, X. (1997). Quantile curves without crossing, *The American Statistician* **51**(2): 186–192.

He, X., Ng, P. and Portnoy, S. (1998). Bivariate quantile smoothing splines, *Journal of Royal Statistical Society, Series B* **60**(part 3): 537–550.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes, *Technical report*.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**(3): 55–67.

Huber, P. J. (1981). *Robust statistics*, John Wiley & Sons, New York.

Kim, J., Kim, Y. and Kim, Y. (2008). A gradient-based optimization algorithm for lasso, *Journal of Computational and Graphical Statistics* **17**(4): 994–1009.

Kim, M.-O. (2007). Quantile regression with varying coefficients, *The Annals of Statistics* **35**(1): 92–108.

Knight, K. (1998). Limiting distributions for l_1 regression estimators under general conditions, *Annals of Statistics* **26**(2): 755–770.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators, *The Annals of Statistics* **28**(5): 1356–1378.

Koenker, R. (2005). *Quantile Regression*, Cambridge U. Press.

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

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines, *Biometrika* **81**(4): 673–680.

Koenker, R. and Zhao, Q. (1994). L-estimation for linear heteroscedastic models, *Journal of Nonparametric Statistics* **3**(3): 223–235.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, Morgan Kaufmann, pp. 1137–1143.

Lee, Y.-J. and Mangasarian, O. L. (2001). SSVM: A smooth support vector machine, *Computational Optimization and Applications* **20**: 5–22.

Leung, D. H.-Y. (2005). Cross-validatoin in nonparametric regression with outliers, *The Annals of Statistics* **33**(5): 2291–2310.

Li, Y. and Zhu, J. (2008). L1-norm quantile regression, *Journal of Computational and Graphical Statistics* **17**(1): 1–23.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*, Chapman & Hall/CRC.

Meier, L., van de Geer, S. and Bhlmann, P. (2008). The group lasso for logistic regression, *Journal of Royal Statistical Society, Series B* **70**(Part 1): 53–71.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing, *Econometrica* **55**(4): 819–847.

Nychka, D., Gray, G., Haaland, P., Martin, D. and O'Connell, M. (1995). A nonparametric regression approach to syringe grading for quality improvement, *Journal of the American Statistical Association* **90**(432): 1171–1178.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). On the lasso and its dual, *Journal of Computational and Graphical Statistics* **9**(2): 319–337.

Osborne, M. R.and Presnell, B. and Turlach, B. A. (2000). A new approach to variable selection in least squares problems, *IMA Journal of Numerical Analysis* **20**(3): 389–404.

Park, T. and Casella, G. (2008). The bayesian lasso, *Journal of the American Statistical Association* **103**(482).

Perkins, S., Lacker, K. and Theiler, J. (2003). Grafting: Fast, incremetal feature selection by gradient descent in function space, *Journal of Machine Learning Research* **3**: 1333–1356.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators, *Econometric Theory* **7**(2): 186–199.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications, *Biometrics* **38**(2): 485–498.

Rockafellar, R. T. (1997). *Convex Analysis (Princeton Landmarks in Mathematics and Physics)*, Princeton University Press.

Ronchetti, E., Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation, *Journal of the American Statistical Association* **92**(439): 1017–1023.

Ronchetti, E. and Staudte, R. G. (1994). A robust version of Mallows's C_p, *Journal of the American Statistical Association* **89**(426): 550–559.

Rosset, S. and Zhu, J. (2004). Discussion of "Least angle regression" by Efron, Hastie, Johnstone and Tibshirani, *Annals of Statistics* **32**(2): 469–475.

Roth, V. (2004). The generalized lasso, *IEEE Transactions on Neural Networks* **15**(1): 16–28.

Shao, J. (1993). Linear model selection by cross-validation, *Journal of the American Statistical Association* **88**(422): 486–494.

Shao, J. (1997). An asymptotic theory for linear model selection, *Statistica Sinica* **7**: 221–264.

Shim, J., Hwang, C. and Seok, K. H. (2009). Non-crossing quantile regression via doubly penalized kernel machine, *Computational Statistics* **24**(1): 83–94.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society, Series B* **36**(2): 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion, *Journal of the Royal Statistical Society, Series B* **39**(1): 44–47.

Takeuchi, I., Le, Q. V., Sears, T. D. and Smola, A. J. (2006). Nonparametric quantile estimation, *Journal of Machine Learning Research* **7**.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* **58**(1): 267–288.

Vapnik, V. N. (1998). *Statistical Learning Theory*, Wiley-Interscience.

Wahba, G. (1990). *Spline Models for Observational Data*, Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia.

Wahba, G. and Wold, S. (1975). A completely automatic french curve: fitting spline functions by cross validation, *Communications in Statistics* **4**: 1–17.

Wretman, J. (1978). A simple derivation of the asymptotic distribution of a sample quantile, *Scandinavian Journal of Statistics* **5**(2): 123–124.

Wu, Y. and Liu, Y. (2007). Robust truncated-hinge-loss support vector machines, *Journal of the American Statistical Association* **102**: 974–983.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression, *Statistica Sinica* **19**(2): 801–817.

Yu, K. and Jones, M. (1998). Local linear quantile regression, *Journal of the American Statistical Association* **93**(441): 228–237.

Yuan, M. and Lin, Y. (2005). Efficient empirical bayes variable selection and estimation in linear models, *Journal of the American Statistical Association* **100**(472): 1215–1225.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **68**(1): 49–67.

Zhao, P. and Yu, B. (2004). Boosted lasso, *Technical report*, Journal of Machine Learning Research.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2): 301–320.