NC 2569 Under Review in Neural Computation, Revised June 10, 2002

Selecting the Loss Function for Robust Linear Regression

Vladimir Cherkassky

Yunqian Ma

Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, U.S.A.

{cherkass, myq}@ece.umn.edu

Abstract

This paper addresses selection of the loss function for regression problems with finite data. It is well-known (under standard regression formulation) that for a *known* noise density there exist an optimal loss function under an *asymptotic* setting (large number of samples), i.e. squared loss is optimal for Gaussian noise density. However, in real-life applications the noise density is *unknown* and the number of training samples is *finite*. Robust statistics provides prescriptions for choosing the loss function using only general information about noise density; however robust statistics is based on asymptotic arguments and may not work well for finite sample problems. For such practical situations, we suggest using Vapnik's ε-insensitive loss function. We propose a practical method for setting the value of ε as a function

of known number of samples and (known or estimated) noise variance. First we consider commonly used unimodal noise densities (such as Gaussian and Laplacian). Empirical comparisons for several representative linear regression problems indicate that the proposed loss function yields more robust performance and improved prediction accuracy, in comparison with commonly used squared loss and least-modulus loss, especially for noisy high-dimensional data sets. We also performed comparisons for symmetric bimodal noise densities (where large errors occur more likely than small errors). For such (bimodal) noise, the proposed ε-insensitive loss consistently provides improved prediction accuracy (in comparison with other loss functions), for both low-dimensional and high-dimensional problems.

1 Introduction

Estimation of a real-valued function from finite set of (noisy) samples is a central problem in applied statistics. Even though the main solution approaches have been proposed centuries ago (i.e., the least-squares method by Gauss and the least-modulus method by Laplace), statistical justification of these methods is based on restrictive assumptions. That is, statistical optimality of these methods is based on two main assumptions:

- *Asymptotic setting*, i.e. good statistical properties such as asymptotic unbiasedness of the least-squares method for linear regression, holds for large number of training samples. It is not at all clear whether this property (unbiasedness) is necessary for finite-sample problems;

- *Knowledge of the noise density*, i.e., the least squares method has the smallest variance for linear regression with normal additive noise. However, in most applications, the noise density is not known.

In this paper, we consider regression problems under practical settings when the number of samples is finite, and the noise density is unknown. There are two main issues that need to be addressed by any method, that is:

- model selection (complexity control);
- choice of parameter estimation (optimization) procedure.

Model selection is outside the scope of this paper; however see (Cherkassky and Mulier, 1998; Cherkassky, 2002) for discussion of model complexity control under the framework of statistical learning theory (or VC-theory). Here we only focus on the parameter estimation procedure, or equivalently, on the choice of the loss function for regression estimation. Moreover, we only consider *linear regression* problems where the number of terms (in a linear model) is given a priori. For such problems, we consider three representative loss functions, i.e. standard squared loss, least-modulus loss (commonly used for robust regression), and ε-insensitive loss function recently proposed by Vapnik (Vapnik, 1995) for Support Vector Machine (SVM) regression. Our goal is to investigate (via empirical comparisons) appropriateness of these loss functions for *finite-sample* estimation problems with unknown noise density. Statistical characterization of different types of loss functions in terms of their robustness to different noise models is addressed in robust statistics (Huber, 1964); and robust loss functions have been extensively used for modeling real-life noisy data (i.e. least-modulus loss has been successfully used for forecasting of economic and political data in (Werbos and Titus, 1978)). However, characterization of different loss functions for finite-sample problems remains sketchy and unsatisfactory, because robust statistics provides only asymptotic characterization of unbiased estimators under maximum-likelihood approach. On the other hand, it is well known that with finite-sample estimation problems it may be preferable to use biased estimators (e.g. recall the bias-variance dilemma in statistics).

The paper is organized as follows. Section 2 reviews standard regression problem statement, describes representative loss functions, and reviews assumptions underlying each loss function. Section 3 presents empirical comparisons for finite-sample settings, for symmetric unimodal noise densities (commonly used in statistical literature) and for bimodal noise densities. Empirical results show that for unimodal noise, ε-insensitive loss (with appropriately chosen ε-value) provides superior robustness and prediction accuracy for high-dimensional sparse data sets. However, for large-sample settings (with unimodal noise) standard squared-loss provides superior prediction accuracy. For bimodal noise, the proposed ε-insensitive loss consistently yields best prediction accuracy for large-sample and small-sample settings alike. Summary and discussion are given in Section 4.

2 Loss functions for regression

We consider standard regression formulation under general setting for predictive learning (Vapnik, 1995; Cherkassky and Mulier, 1998; Hastie et al, 2001). The goal is to estimate unknown real-valued function in the relationship:

$$y = g(\mathbf{x}) + \delta \tag{1}$$

where δ is independent and identically distributed (i.i.d.) zero mean random error (noise), $\mathbf{x}$ is a multivariate input and y is a scalar output. The estimation is made based on a finite number (n) of samples (training data): $(\mathbf{x}_i, y_i), (i = 1,...,n)$. The training data are i.i.d. samples generated according to some (unknown) joint probability density function (pdf),

$$p(\mathbf{x}, y) = p(\mathbf{x})p(y \mid \mathbf{x}) \tag{2}$$

The unknown function in (1) is the mean of the output conditional probability (aka regression function)

$$g(\mathbf{x}) = \int y p(y \mid \mathbf{x}) dy \,. \tag{3}$$

A learning method (or estimation procedure) selects the 'best' model $f(\mathbf{x}, \omega_0)$ from a set of approximating functions (or possible models) $f(\mathbf{x}, \omega)$ parameterized by a set of parameters $\omega \in \Omega$. The quality of an approximation is measured by the loss or discrepancy measure $L(y, f(\mathbf{x}, \omega))$, and the goal of learning is to select the best model minimizing (unknown) prediction risk:

$$R(\omega) = \int L(y, f(\mathbf{x},\omega)) p(\mathbf{x}, y) d\mathbf{x}dy \tag{4}$$

It is known that the regression function (3) is the one minimizing prediction risk (4) with the squared loss function loss:

$$L(y, f(\mathbf{x},\omega)) = (y - f(\mathbf{x},\omega))^2 \tag{5}$$

Note that the set of functions $f(\mathbf{x},\omega)$, $\omega \in \Omega$ supported by a learning method may or may not contain the regression function (3). Thus the problem of regression estimation is the problem of finding the function $f(\mathbf{x},\omega_0)$ (regressor) that minimizes the prediction risk functional

$$R(\omega) = \int (y - f(\mathbf{x},\omega))^2 p(\mathbf{x}, y) d\mathbf{x}dy \tag{6}$$

using only the training data. This risk functional measures the accuracy of the learning method's *predictions* of unknown target function $g(\mathbf{x})$.

Since the prediction risk functional (6) is not known, we need to estimate the best model $f(\mathbf{x},\omega_0)$ using only available (training) data. For a given parametric model, its parameters are estimated by minimizing the empirical risk:

$$R_{emp}(\omega) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i,\omega)) \tag{7}$$

For example, with commonly used squared loss, the best model is found via minimization of average squared error:

$$R_{emp}(\omega) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i, \omega))^2 \tag{8}$$

In this paper we assume that in regression formulation (1) the target function is linear and its parametric form is known, i.e.

$$f(\mathbf{x}, \omega) = \omega_0 + \omega_1 x_1 + ... + \omega_d x_d \tag{9}$$

The goal is to estimate parameters of a linear regression (9) when the number of samples is finite, and the noise density in (1) is not known. The key problem for regression is the choice of a loss function used to minimize the empirical risk functional (7). For example, it is well-known that using quadratic loss function for regression problems with normal additive noise provides an asymptotically efficient (best unbiased) estimator for regression. Using standard statistical arguments (Smola and Schölkopf, 1998; Hastie et al, 2001), one can find the optimal loss function (in a maximum-likelihood sense) for a *known* noise density $p(\delta)$, that is:

$$L(y, f(\mathbf{x}, \omega)) = -\log p(y - f(\mathbf{x}, \omega)) \tag{10}$$

Two commonly used types of noise and corresponding 'optimal' loss functions are shown in Figure 1a, 1b. We note, however, that this 'optimality' is based on asymptotic arguments and it may not hold well for finite-sample settings.

In most cases the noise distribution is unknown and robust methods (Huber, 1964) provide various loss functions using only general information about the model of noise. For example, if one only knows that the noise density is a *unimodal* symmetric smooth function, then the best minimax strategy for regression estimation is provided by the least-modulus (or absolute-value) loss:

$$L(y, f(\mathbf{x},\omega)) = | y - f(\mathbf{x},\omega) | \tag{11}$$

We emphasize here that statistical optimality properties for both the least-modulus and squared loss (under respective assumptions about the noise density) should be understood under an asymptotic setting, i.e. when the number of training samples is large. Several empirical results (presented later in this paper) clearly demonstrate that statistical 'optimality' of these loss functions does not hold for finite-sample settings.

Statistical analysis of various noise models typically considers unimodal and smooth noise densities, reflecting an assumption that small errors are much more likely than large errors. In practice, however, we are often faced with symmetric bimodal distributions (that may also contain discontinuities). Figure 1c, 1d shows representative examples, i.e. uniform noise model and bimodal uniform noise density:

$$p(\delta) = \begin{cases} 0.25 & -3 < \delta < -1 \\ 0.25 & 1 < \delta < 3 \\ 0 & otherwise \end{cases} \tag{12}$$

In this paper, we shall use noise densities shown in Figure 1 as representative examples of unimodal and bimodal types of noise.

Recently, a new loss function called ε-insensitive loss has been proposed by Vapnik (1995):

$$L_{\varepsilon}(y, f(\mathbf{x},\omega)) = \begin{cases} 0 & \textit{if } |y - f(\mathbf{x},\omega)| \leq \varepsilon \\ |y - f(\mathbf{x},\omega)| - \varepsilon & \textit{otherwise} \end{cases} \tag{13}$$

This loss function had been proposed in the context of Support Vector Machine (SVM) regression. SVM approach to linear regression amounts to (simultaneous) minimization of ε-insensitive loss and minimization of the norm of linear parameters ($\|\omega\|^2$). This can be formally described by introducing (non-negative) slack variables $\xi_i, \xi_i^* \; i = 1,...n$, to measure the deviation of training samples outside ε-insensitive zone. Thus SVM regression can be formulated as minimization of the following functional:

$$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{14}$$

$$\text{Subject to} \quad \begin{cases} y_i - f(\mathbf{x}_i,\omega) \leq \varepsilon + \xi_i^* \\ f(\mathbf{x}_i,\omega) - y_i \leq \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0, i = 1,..., n \end{cases}$$

This optimization problem can be transformed into the dual problem (Vapnik, 1995), and its solution is given by

$$f(\mathbf{x}) = \sum_{i=1}^{n_{SV}}(\alpha_i - \alpha_i^*) < \mathbf{x}_i, \mathbf{x} > \tag{15}$$

with coefficient values in the range $0 \leq \alpha_i^* \leq C$, $0 \leq \alpha_i \leq C$. And $< \cdot,\cdot >$ denotes the dot product in the input space. In representation (15), typically only a fraction of training samples appear with non-zero coefficients and such training samples are called support vectors. For most applications, the number of support vectors (SVs) n_{SV} is usually much smaller than the number of training samples.

For nonlinear regression problem, SVM approach performs first a mapping from the input space onto a high-dimensional feature space, and then performs linear regression in the high-dimensional feature space using ε-insensitive loss (Vapnik, 1995; Cherkassky and Mulier, 1998; Schoelkopf et al, 1999). SVM approach enables efficient model complexity control using a special structure on a set of high-dimensional linear models in the feature space. SVM model complexity control and its attractive computational formulation have resulted in many successful applications for nonlinear regression (Schölkopf et al, 1999). However, in this paper we do not consider nonlinear regression problems, since our goal is a better understanding of different loss functions, rather than efficient complexity control.

The primary motivation (for using ε-insensitive loss) appears to be computationally efficient SVM regression formulation resulting in a sparse form of SVM solution. Namely, the value of ε controls the number of support vectors (the sparseness of SVM solutions) and thus indirectly controls model complexity and generalization. More recently, several researchers used SVM-like optimization formulation with different loss functions, i.e., squared loss, squared loss with ε-insensitive zone, and Huber loss (Vapnik, 1998; Smola et al, 1998; Suykens et al, 2001). In fact, one can use any convex loss function under SVM regression formulation; however only ε-insensitive loss yields sparse SVM solutions (Vapnik, 1998).

This paper presents a different interpretation of Vapnik's ε-insensitive loss, i.e. its advantages for finite-sample estimation problems. The main idea of ε-insensitive loss is to ignore 'small' errors during minimization of empirical risk (7); whereas 'large' errors are assigned absolute-value loss appropriate for unknown noise density. The conceptual difference between ε-insensitive loss and other types of loss functions (Huber loss, squared loss, and least modulus loss) is that the latter do not make qualitative distinction between small errors and large errors, which is crucial for finite-sample estimation problems. Unfortunately, SVM framework does not provide clear guidelines on how to select the value of ε-insensitive zone for a given data set, so in practical studies its selection depends on user experience and/or on computationally intensive data-driven techniques (Schoelkopf et al, 1999).

SVM estimates depend on both parameter C and the value of ε in formulation (14). Hence, obtaining optimal SVM solutions for linear regression requires setting optimal values for both parameters. Recently, Cherkassky and Ma (2002) proposed a practical method for selecting the value of ε and the value of regularization parameter C for SVM regression directly from the training data. Namely, the value of C is chosen as:

$$C = \max(| \bar{y} + 3\sigma_y |, | \bar{y} - 3\sigma_y |) \tag{16}$$

where $\bar{y}$ is the mean of the training responses (outputs), and σ_y is the standard deviation of the training response values. Prescription (16) can effectively handle outliers in the training data. In practice, the response values of training data are often scaled so that $\bar{y}$=0; then the optimal C is $3\sigma_y$.

The value of ε is selected as

$$\varepsilon(\sigma, n) = \tau\sigma\sqrt{\frac{\ln n}{n}} \tag{17}$$

where σ is the standard deviation of additive noise and n is the number of training samples, and τ is an empirically determined constant. The value $\tau = 3$ was suggested by Cherkassky and Ma (2002). Thus, expression (17) with $\tau = 3$ is used throughout this paper for setting the value of ε -insensitive zone. Note that using prescription (17) requires estimation of noise level σ; this can be accomplished using standard noise estimation approaches – see Cherkassky and Ma (2002) for details. In the remainder of the paper we assume that standard deviation of noise σ is known (or can be accurately estimated from data). However, we point out here that in most practical applications the noise density is unknown (and can not be accurately estimated), even when its standard deviation can be reliably estimated from the training data.

Further, it can be shown (Cherkassky and Ma, 2002) that for linear regression settings the choice of ε affects prediction accuracy much more than the choice of C parameter (provided that C parameter is larger than the value given by (16)). In other words, one can use very large values of C parameter in SVM formulation (14) without any degradation in prediction risk. For example, Fig. 2 shows SVM prediction accuracy as a function of both parameters, for a representative linear regression data set. As evident from Fig. 2, one can use any large C-value (say over 40) and then select ε -value to optimize prediction risk. Conceptually, using large C (infinite) values in SVM formulation (14) means that such formulation becomes equivalent to minimization of ε -insensitive loss for the training data, without penalization (regularization) term. This enables 'fair' comparisons between SVM and other formulations / loss functions (e.g. least squares) which do not use regularization term.

In the next section we present empirical comparisons for several linear regression estimation using three representative loss functions: squared loss, least-modulus and ε-insensitive loss with selection of ε given by (17). Our goal is to investigate the effect of a loss function on the prediction accuracy of linear regression with finite samples. Even though SVM regression has been extensively used for regression applications (Scholkopf et al, 1999), its success is mainly due to remarkable ability of SVM models to handle *nonlinear* high-dimensional problems. However, there is little consensus and understanding of the importance of ε-insensitive loss itself for standard *linear regression* estimation. The only existing study (Drucker, et al, 1997) showing empirical comparisons between SVM and ordinary least squares (OLS) for linear regression makes rather indefinite conclusions. This study applies SVM and OLS to a linear regression problem with 30 input variables, where regression estimates are obtained from 60 noisy training samples, and concludes that at high noise levels SVM is better than OLS, but at low noise levels OLS is better than SVM. This study is rather sketchy since it uses a single data set for regression comparisons, and does not describe any systematic procedure for selecting the value of ε for SVM regression.

3 Empirical Comparison

This section presents systematic comparisons of several loss functions/methods for linear regression. Specifically, we are interested in the following factors that are likely to affect generalization performance of SVM, OLS and LM methods, such as:

- low vs. high-dimensional data sets (target functions);

- sparse vs. non-sparse settings (where ‘sparseness’ can be defined as the ratio of the number of samples to the number of input variables);
- percentage of 'important’ or significant input variables for high-dimensional problems;
- the amount of additive noise (measured as SNR)
- type of noise.

First we describe experimental procedure and then present empirical results. Comparisons are organized into 2 groups, that is: *large-sample* setting for low-dimensional problems (when the number of samples is significantly larger than the number of parameters in linear regression), and high-dimensional or *sparse* problems. For high-dimensional problems (which is the case of practical interest) we also show comparisons for various settings corresponding to large/small percentage of significant input variables, high/low noise levels etc. All comparisons for different methods are shown for three representative unimodal noise densities: Gaussian, Laplacian and Uniform. In addition, we show comparisons for bimodal noise density. The goal (of comparisons) is to gain better understanding of relative advantages/limitations of different methods for linear regression: optimal least squares (OLS), least modulus (LM) and SVM regression. Note that SVM method has a tunable parameter ε selected via analytical prescription (17) for all comparisons presented in this paper. Alternatively, optimal selection of ε can be done using resampling methods. We empirically compared the resampling approach (via cross-validation) and analytical approach for selecting the value of ε, and found no significant difference in terms of prediction accuracy of SVM estimates.

Training data: we use simulated training data $(\mathbf{x}_i, y_i), (i = 1,...n)$ with random $\mathbf{x}$ -values uniformly distributed in the input space, and y -values generated according to $y = g(\mathbf{x}) + \delta$.

Target functions. Several low-dimensional and high-dimensional data sets are generated using the following target functions:

Low-dimensional $\quad g(x) = 12x_1 + 7x_2, x \in [0,1]^2 \quad$ (18)

High-dimensional function of 20 variables:

$$g(x) = 4x_1 + 4x_2 + 3x_3 + 3x_4 + 2x_5 + x_6 \cdots + x_{20}\,,\ x \in [0,1]^{20} \qquad (19)$$

We also considered variations of a high-dimensional function (19) with only a fraction of important (or significant) variables, as shown in Table 1.

Table 1

	Target function
10%	$g(x) = 4x_1 + 4x_2 + 0.01x_3 + ... + 0.01x_{20}$
25%	$g(x) = 4x_1 + 4x_2 + 3x_3 + 3x_4 + 2x_5 + 0.01x_6 ... + 0.01x_{20}$
50%	$g(x) = 4x_1 + 4x_2 + 3x_3 + 3x_4 + 2x_5 + x_6 + ... + x_{10} + 0.01x_{11} ... + 0.01x_{20}$
75%	$g(x) = 4x_1 + 4x_2 + 3x_3 + 3x_4 + 2x_5 + x_6 + ... + x_{15} + 0.01x_{16} ... + 0.01x_{20}$

Sample size. Various training sample sizes (n= 30,40,50) are used to contrast relative performance of different methods under large sample settings and sparse sample settings.

The distinction can be quantified using the ratio of the number of samples (sample size) to the number of input variables.

Additive noise. The following types of *unimodal* noise were used: Gaussian noise, uniform noise and Laplacian noise. Notice that squared loss is (asymptotically) optimal for Gaussian noise and least-modulus loss is (asymptotically) optimal for Laplacian noise density. In addition, we used *bimodal* noise (see Figure 1d) that is expected to be most appropriate for ε-insensitive loss. We also varied the noise level (as indicated by different SNR values) for high-dimensional data, in order to understand the effect of noise level on methods' performance. Signal-to-noise ratio (SNR) is defined as the ratio of the standard deviation of the true (target function) output values over the standard deviation of the additive noise.

Given a large number of experimental design factors (as defined above), we conducted many experiments with different target functions, noise levels etc. This paper presents only a small but representative subset of comparisons. However, we emphasize that all qualitative conclusions presented in this paper are consistent with a wider set of comparison results not shown here due to space limitations.

Experimental protocol. For a given training sample with specified statistical properties (sample size, noise level/type etc. as defined above) we estimate parameters of regression via minimization of the empirical risk (7) using 3 different loss functions, i.e., standard square loss, modulus loss and ε-insensitive loss (with proposed selection of ε-value). The quality of each model is evaluated as its prediction accuracy, measured as mean-squared error (MSE) between (estimated) model and the true target function. This quantity is measured using large number of independent test samples uniformly distributed in the input space. Specifically, for low-dimensional problems we used 200 test samples, and for high-dimensional problems 2,000 test samples were used to estimate the prediction risk. Since the model itself depends on a particular (random) realization of training sample (of fixed size), its (measured) prediction accuracy is also a random variable. Hence, we repeat the experimental procedure (described above) with many different realizations of training data (100 runs) and show average prediction accuracy (risk) for methods' comparison. Graphical presentation of prediction accuracy (risk) for three estimation methods uses the following labels: OLS (for ordinary least-squares method), LM (for least-modulus method) and SVM (for SVM with ε – insensitive loss using proposed optimal selection of ε). Notice that LM method is a special case of SVM with ε-insensitive loss (with ε=0).

Comparison results. Empirical comparisons for low-dimensional target function (18) are summarized in Table 2. The results show average prediction accuracy (MSE) for various methods/loss functions observed for different types of noise (with the same SNR value). As expected, OLS shows superior performance for Gaussian noise, whereas LM is the best for Laplacian noise. For unimodal noise OLS shows best performance overall, whereas SVM clearly provides superior prediction performance for bimodal noise. Good overall performance of standard OLS can be explained by noting that this data set is an example of large-sample setting, i.e. we use 30 samples to estimate 3 parameters of regression function. Hence standard statistical arguments underlying asymptotic optimality of different loss functions hold well in this case.

Table 2: Prediction accuracy for low-dimension data $g(x) = 12x_1 + 7x_2$, n=30, SNR=2 Results show the average MSE error (for 100 realizations) for different types of noise.

	Unimodal			Bimodal
	Gaussian	Uniform	Laplacian	
OLS	0.4210	0.4338	0.4002	0.4807
LM	0.6560	1.1238	0.3327	2.0734
SVM	0.5981	0.4487	0.6503	0.1993

As mentioned earlier, SVM method selects the value of ε using proposed analytic expression (17). Empirical results in Fig. 3 are shown to support our claim that expression (17) yields (near) optimal selection of ε-values. Namely, Fig. 3 shows comparisons of prediction risk obtained for the low-dimensional data set using SVM with ε-values chosen around the 'optimal' value selected by (17), that is:

- 50% smaller than selected value (denoted as '–50%' in Fig. 3);
- 50% larger than selected value (denoted as '+50%');
- Twice the selected value (denoted as '2*Sel').

Comparisons of prediction risk shown in Fig.3 (with different ε-values) show that expression (17) selects the value of ε-parameter quite well (close to optimal) for different types of noise.

Next we show comparisons for a high-dimensional target function (19). Results shown in Fig. 4 are intended to illustrate how methods' prediction performance depends on the sparseness of training data. This is accomplished by comparing prediction risk (MSE) for data sets with different sample sizes (n=30, 40 and 50) under the same SNR=2. Results in Fig.4 indicate that SVM method consistently (for all types of noise) outperforms other methods under sparse settings, i.e. for 30 samples when the ratio n/d is smaller than 2. However, for 50 samples, when this ratio is larger than 2, we approach large-sample settings, and the methods' performance becomes similar. The distinction between sparse setting and large-sample setting is not very straightforward as it also depends on the noise level. That is why comparisons in Fig. 4 are shown for a given (fixed) SNR value for all data sets. Next we show comparisons for the same high-dimensional target function (19) under sparse setting (n=30 samples) for different noise levels (SNR=1,3,5,7) in order to understand the effect of noise level on methods' performance. Results in Fig. 5 clearly show superiority of SVM method for large noise levels; however for small noise levels SVM does not provide any significant advantages over OLS. Note that MSE results in Fig. 5 are shown on a logarithmic scale, so that the difference in prediction performance (MSE) for different methods at high noise levels (SNR=1) is quite significant (i.e., of the order of 100% or more).

Additional experiments with high-dimensional data use high-dimensional target functions with a small number of important input variables (shown in Table 1). Here the goal is to understand how the methods' performance is affected by the percentage of important variables, under sparse settings. Results presented in Fig. 6 use the same sample size n=30 and noise level SNR=2 for all comparisons. Examination of results in Fig. 6 suggests that SVM provides superior performance for all target functions – due to sparseness of training data and relatively high noise level. However, relative performance of SVM is much better when the percentage of important variables is high. When the percentage of important variables is low (say 10%), SVM does not yield significant advantages vs OLS; this can be explained by noting that data sets with small number of important input variables effectively represent relatively large-sample settings.

4. Summary and Discussion

This paper presents empirical comparisons between several loss functions (methods) for linear regression with finite samples. Previous SVM regression comparison studies focused mainly on nonlinear regression settings where one needs to select several hyper-parameters (such as SVM kernel, ε-parameter and regularization parameter) in order to obtain good SVM model; and such parameter selection is usually performed by expert users and/or using resampling methods (Scholkopf, 1999). In contrast, this paper focuses on linear regression and provides an analytic prescription for selecting the value of ε-insensitive loss.

Empirical comparisons presented in this paper indicate that SVM approach clearly provides better prediction accuracy than OLS and LM for linear regression with sparse and noisy data. Our findings contradict a common opinion in the statistical literature

(Smola and Schölkopf, 1998; Hastie et al, 2001) that for a given (known) noise density there exist an optimal method (loss function).

Results presented in this paper have a number of practical and conceptual implications as discussed next. *First,* we have demonstrated practical advantages of using ε-insensitive loss (with proposed selection of ε-value) for finite-sample linear regression problems, when the training data is noisy and sparse. For low-dimensional problems (i.e. non-sparse settings) using SVM does not offer any improvement over standard OLS method, as can be expected from asymptotic statistical theory. However, advantages of SVM approach become apparent for sparse high-dimensional data sets where it outperforms ordinary least squares (OLS) even for Gaussian noise. *Second*, this superiority of ε –insensitive loss for settings where OLS method is known to be 'statistically optimal' has an important conceptual implication: statistical results developed under asymptotic settings do not hold for finite-sample high-dimensional data sets. Moreover, our results suggest that for finite-sample settings one only needs to know (or estimate) the standard deviation of noise, rather than its distribution. *Third*, our empirical results help to explain the empirical success of SVM approach for nonlinear regression, as explained next. Under standard SVM approach to nonlinear regression one first needs to perform a nonlinear mapping from an input space onto high-dimensional feature space and then perform linear regression (with special constraints) in this feature space according to SVM formulation. Minimization of regularized functional (14) for nonlinear regression should always result in *sparse settings*, where the number of parameters (in linear regression) is comparable to the number of training samples.

According to findings presented in this paper, for such sparse data sets linear SVM should use ε –insensitive loss (rather than alternative loss functions). It may be interesting to perform empirical comparisons between several nonlinear SVM regression formulations using different loss functions in (14). In contrast to a popular opinion that the loss function in (14) should be tailored to particular noise density (Smola and Sholkopf, 1998; Hastie et al, 2001) we expect an overall superiority of ε–insensitive loss for sparse high-dimensional problems. Further, our results help to understand better relative importance of ε–insensitive loss vs regularization term in SVM formulation (14). Whereas both factors are important for nonlinear SVM regression, our experience suggests that the choice of ε–insensitive loss parameter is much more important for linear regression. Even for nonlinear SVM, the quality of SVM solutions is typically much more sensitive to proper setting of ε –parameter rather than regularization parameter (Cherkassky and Ma, 2002). *Finally,* we have shown that SVM approach with ε–insensitive loss can be successfully used for symmetric *bimodal noise* distributions. This finding may be of practical importance in applications where bimodal noise models preclude application of standard OLS and LM methods.

Conceptually, Vapnik's loss function with insensitive zone can be viewed as a new approach for providing "regularization" for ill-posed estimation problems with sparse data. Our empirical comparisons indicate that such an approach can be quite useful for sparse high-dimensional data sets, provided that one can control the degree of regularization (via the choice of ε-value). This view also leads to novel interpretation of SVM formulation (14) where regularization can be performed in two ways:

(1) Standard regularization via the choice of parameter C;

(2) Regularization via the choice of ε-value.

In this paper we only considered the second approach (by intentionally setting the value of C very high). This approach works well for linear regression settings. However, for nonlinear SVM (with kernels) we may need to explore more advanced strategies for joint regularization (optimization) using both parameters C and ε. This opens new directions for SVM research.

Acknowledgements

The authors would like to thank Dr. Vladimir Vapnik for several stimulating discussions. We also acknowledge several constructive suggestions by Referee 1 that helped to improve technical quality of the paper. This work was supported, in part, by NSF grant ECS-0099906.

References

Cherkassky, V., & Ma, Y.(2002). Practical Selection of SVM Parameters and Noise Estimation for SVM Regression, *Neurocomputing* (under review).

Cherkassky, V., & Ma, Y.(2002). Selection of Meta-Parameters for Support Vector Regression, *Proc. ICANN-2002* (to appear).

Cherkassky, V.(2002). Model Complexity Control and Statistical Learning Theory, Natural Computing, 1, *Kluwer*, 109-133

Cherkassky, V., & Mulier, F.(1998). *Learning from Data: Concepts, Theory and Methods*, Wiley.

Drucker, H., Burges, C., Kaufman, L., Smola, A., & Vapnik, V.(1997). Support Vector Regression Machines, *Neural Information Processing Systems 9*, Eds. M. Moser, J. Jordan and T. Petsche, 155-161, MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J.(2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.

Huber, P., (1964). Robust estimation of a location parameter, *Ann. Math. Stat., 35,* 73-101.

Scholkopf, B., Burges, C., & Smola, A.(1999). *Advances in Kernel Methods: Support Vector Machine*, MIT Press.

Smola, A., & Schölkopf, B.(1998). A Tutorial on Support Vector Regression, NeuroCOLT Technical Report NC-TR-98-030, Royal Holloway College, University of London, UK.

Suykens, J., Vandewalle, J., & De Moor, B.(2001). Optimal control by least squares support vector machines, *Neural Networks* 14, Pergamon, 23-35.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer

Vapnik, V. (1998). *Statistical Learning Theory*, New York: Wiley.

Werbos, P., & Titus, J., (1978). An empirical test of new forecasting methods derived from a theory of intelligence: the prediction of conflict in Latin America, *IEEE Trans. Systems, Man & Cybernetics.*

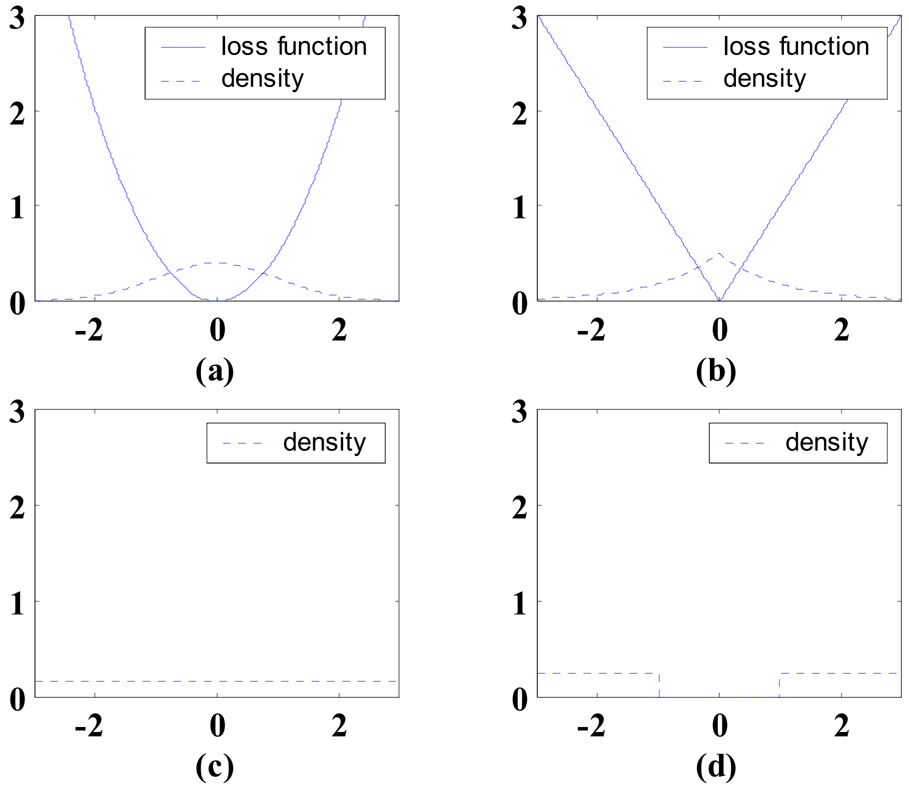


Figure 1 [Cherkassky]: (a) Gaussian noise and quadratic loss function
(b) Laplacian noise and least-modulus loss function
(c) Uniform noise
(d) Bimodal noise

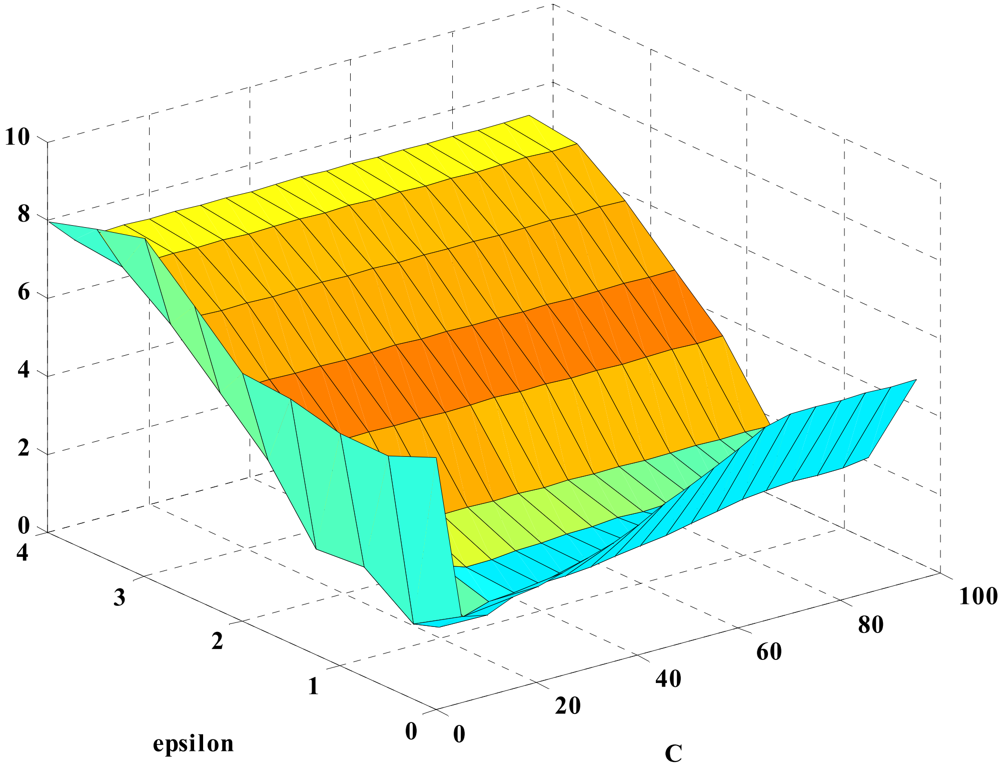


Figure 2 [Cherkassky]: Prediction Accuracy as a function of epsilon and C-values for high-dimension data set $g(x) = 4x_1 + 4x_2 + 3x_3 + 3x_4 + 2x_5 + x_6 ... + x_{20}$, n=30, SNR=2.

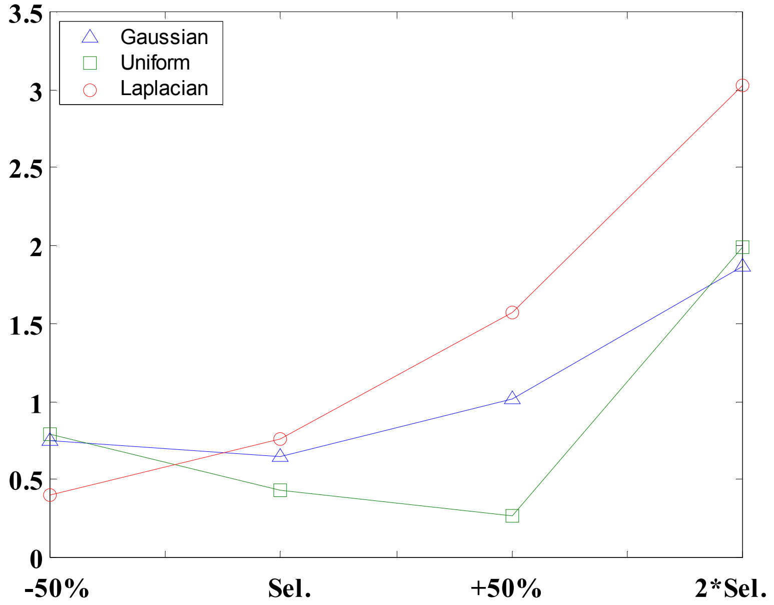
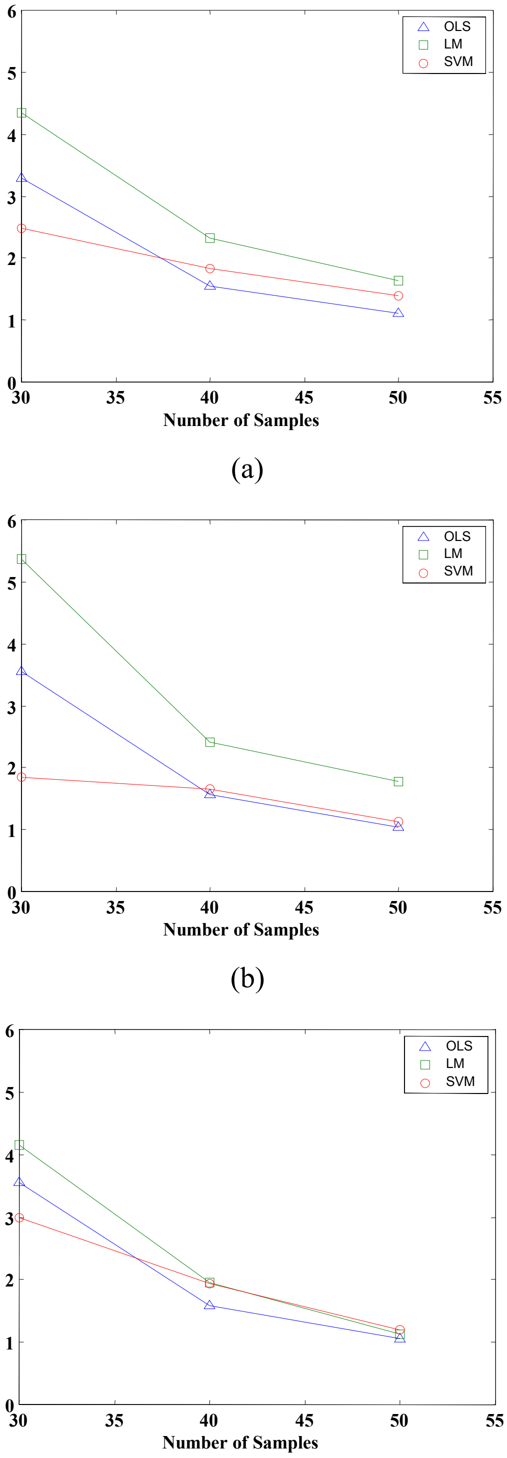


Figure 3 [Cherkassky]: Prediction Accuracy for different values of epsilon (-50%, Sel., +50%, 2*Sel.) for $g(x) = 12x_1 + 7x_2$, n=30, SNR=2 for Gaussian noise, Uniform noise and Laplacian noise

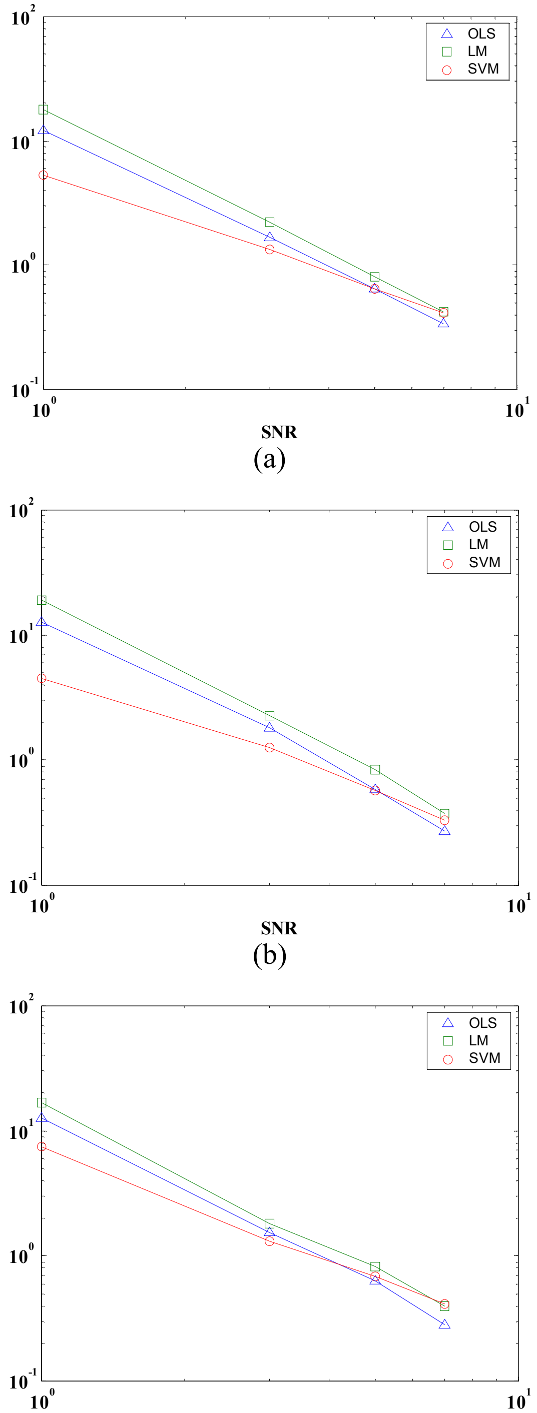


(a)

(b)

(c)

Figure 4 [Cherkassky]: Prediction Accuracy vs. sample size n=30,40,50 for high-dimensional data set $g(x) = 4x_1 + 4x_2 + 3x_3 + 3x_4 + 2x_5 + x_6 ... + x_{20}$, SNR=2 (a) Gaussian noise (b) Uniform noise (c) Laplacian noise

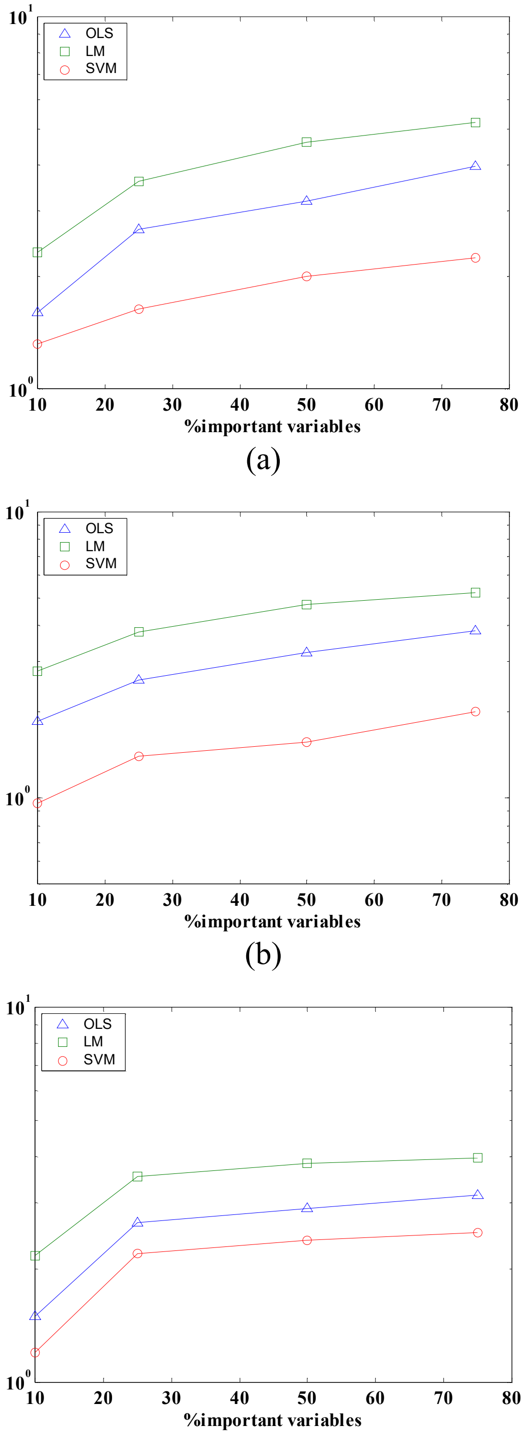


(a)

(b)

(c)

Figure 5 [Cherkassky]: Prediction Accuracy vs. SNR (1, 3, 5, and 7) for high-dimensional data $g(x) = 4x_1 + 4x_2 + 3x_3 + 3x_4 + 2x_5 + x_6 ... + x_{20}$, n=30, (a) Gaussian noise (b) Uniform noise (c) Laplacian noise



(a)

(b)

(c)

Figure 6 [Cherkassky]: Prediction Accuracy vs. %important variables for high-dimensional data, n=30, SNR=2; (a) Gaussian noise (b) Uniform noise (c) Laplacian noise