Robust Cross-Validation Score Functions with Application to Weighted Least Squares Support Vector Machine Function Estimation [*]

J. De Brabanter, **K. Pelckmans**,
J.A.K. Suykens, **J. Vandewalle**, **B. De Moor**
Katholieke Universiteit Leuven
Department of Electrical Engineering, ESAT-SISTA
Kardinaal Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium
Tel: 32/16/32 11 45 Fax: 32/16/32 19 70
Email: jos.debrabanter@esat.kuleuven.ac.be

Corresponding author: Jos De Brabanter

Running title: Robust cross-validation score functions

July 2003

[*] This research work was carried out at the ESAT laboratory and the Interdisciplinary Center of Neural Networks ICNN of the Katholieke Universiteit Leuven, in the framework of the Belgian Program on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture (IUAP P4-02, IUAP P4-24, IUAP V-22), the Concerted Action Project MEFISTO of the Flemish Community and the FWO project G.0407.02. JS is a postdoctoral researcher with the National Fund for Scientific Research FWO - Flanders.

Abstract

In this paper new robust methods for tuning regularization parameters or other tuning parameters of a learning process for non-linear function estimation are proposed: repeated robust cross-validation score functions (repeated-CV $^{Robust}_{V-fold}$) and a robust generalized cross-validation score function (GCV^{Robust}). Both methods are effective for dealing with outliers and non-Gaussian noise distributions on the data. The robust procedures are based both on a robust cross-validation estimate and a robust function estimator. Simulation results for weighted Least Squares Support Vector Machine (weighted LS-SVM) function estimation are given to illustrate that the proposed robust methods outperform other cross-validation procedures and methods based on a number of other complexity criteria.

Keywords: Robust Cross-validation Score function, Robust Statistics, Influence functions, Breakdown point, M-estimators and L-estimators, Weighted LS-SVM.

1 Introduction

Most efficient learning algorithms in neural networks, support vector machines and kernel based methods (Bischop, 1995), (Cherkassky *et al.*, 1998), (Vapnik, 1999), (Hastie *et al.*, 2001) and (Suykens *et al.*, 2002b) require the tuning of some extra learning parameters, or *tuning parameters*, denoted here by θ. The tuning parameter selection methods can be divided into three broad classes:

- *Classical methods.* Cross validation, Mallows' C_p (Mallows, 1973), Akaike's information criterion (Akaike, 1973), Bayes Information Criterion (Schwartz 1979). These are more or less natural extensions of methods used in parametric modeling.

- *Plug-in methods.* The bias of an estimate of an unknown real-valued smooth function is usually approximated through Taylor series expansions. A pilot estimate of the unknown function is then "plugged in" to derive an estimate of the bias and hence an estimate of the mean integrated squared error. The optimal tuning parameters minimizes this estimated measure of fit. More complete descriptions of these approaches are given in (Härdle, 1989).

- *VC dimension.* The Vapnik-Chernovenkis theory provides a general measure of complexity, and gives associated bounds (Vapnik, 1998).

Support Vector Machines (SVM) for nonlinear function estimation, as introduced by (Vapnik, 1995, 1999, 2002) is a new methodology in the area of neural networks and nonlinear modelling. SVM is a kernel based approach which allows the use of linear, polynomial and Radial Basis Functions (RBF) and other kernels that satisfy Mercer's condition. Typically, one solves a convex quadratic programming (QP) problem in dual space in order to determine the SVM model. The formulation of the optimization problem in the primal space associated with this QP problem involves inequality constraints. Recently, least squares (LS) versions of the SVM have been investigated for function estimation and classification (Suykens *et al.*, 1999, 2002b)(Saunders *et al.*, 1998). In these LS-SVM formulations one works with equality constraints and a sum of squared errors (SSE) cost function.

For practical use, it is often preferable to have a data-driven method to select θ. For this selection process, many data-driven procedures have been discussed in the literature. Commonly used are those based on the cross-validation criterion of Stone (Stone, 1974) and the generalized cross-validation criterion of Craven and Wahba (Craven and Wahba, 1979). One advantage of cross-validation and generalized cross-validation over some other selection criteria such as Mallows' C_p, Akaike's information criterion is that they do not require estimates of the error variance. This means that Mallows' C_p, Akaike's information criterion require a roughly correct working model to obtain the estimate of the error variance. Cross-validation does not require this. The motivation behind cross-validation is easily understood, see (Allen, 1974) and (Stone, 1974). Much work has been done on the ordinary or leave-one-out cross-validation (Bowman, 1984) and (Härdle and Marron, 1985). However, the difficulty with ordinary cross-validation is that it can become computationally very expensive in practical problems. Therefore, (Burman, 1989) has introduced V-fold cross-validation. For more references on smoothing parameter selection, see (Marron, 1987, 1989) and (Härdle and Chen, 1995).

In recent years, results on L_2 and L_1 cross-validation statistical properties have become available (Yang and Zheng, 1992). However, the condition $E\left[e_k^2\right] < \infty$ (respectively, $E\left[|e_k|\right] < \infty$) is necessary for establishing weak and strong consistency for L_2 (respectively, L_1) cross-validated estimators. On the other hand, when there are outliers in the output observations (or if the distribution of the random errors has a heavy tail so that $E\left[|e_k|\right] = \infty$), then it becomes very difficult to obtain good asymptotic results for the $L_2\left(L_1\right)$ cross-validation criterion. In order to overcome such problems, a robust cross-validation score function is proposed in this paper. This is done by first treating the values of the cross-validation score function as a realization of a random variable. In a second stage, the location parameter (e.g. the mean) of this realization is estimated by a robust method. The results of this paper illustrate that the robust methods can be very effective, especially with non-Gaussian noise distributions and outliers in the data.

This paper is organized as follows. In Section 2 we describe the weighted LS-SVM for non-linear function estimation. In Section 3 we motivate the cross-validation procedure as the θ selection rule. In Section 4 the classical V-fold cross-validation score function is analysed and the repeated V-fold cross-validation score function is proposed. A robust and efficient repeated-V-fold cross-validation score function is proposed in Section 5. A robust generalized cross-validation is proposed in Section 6. In Section 7, we describe a simulation study and discuss the result. Appendix A treats statistics (location estimators) obtained as solutions of equations (M-estimators) and linear functions of order statistics (L-estimators). These classes are important in robust parametric inference. In Appendix B, some measures of robustness (e.g. breakdown point and influence function) are described. The trade-off between robustness and efficiency of the location estimators will be analysed in Appendix C. Appendix D gives a derivation of the influence function of the trimmed mean.

2 Weighted LS-SVM for robust non-linear function estimation

2.1 The unweighted case

Given a training data set of N points $\mathcal{D} = \{(x_k, y_k)\}_{k=1}^{N}$ with output data $y_k \in \mathbb{R}$ and input data $x_k \in \mathbb{R}^n$ according to

$$y_k = f(x_k) + e_k, \qquad k = 1, ..., N, \tag{1}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is an unknown real-valued smooth function that one wishes to estimate. The e_k are assumed to be i.i.d. random errors with $E[e_k] = 0$, $Var\,[e_k] = \sigma^2$ and $E[y_k|x = x_k] = f(x_k)$. One considers the following optimization problem in primal weight space:

$$\min_{w,b,e} \mathcal{J}\,(w, e) = \frac{1}{2} w^T w + \frac{1}{2}\gamma \sum_{k=1}^{N} e_k^2, \tag{2}$$

such that

$$y_k = w^T \varphi\,(x_k) + b + e_k, \qquad k = 1, \ldots, N,$$

with $\varphi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_h}$ a function which maps the input space into a so-called higher dimensional (possibly infinite dimensional) feature space, weight vector $w \in \mathbb{R}^{n_h}$ in primal weight space, random errors $e_k \in \mathbb{R}$ and bias term b. Note that the cost function $\mathcal{J}$ consists of a RSS fitting error and a regularization term, which is also a standard procedure for the training of MLP's and is related to ridge regression (Golub and Van Loan, 1989). The relative importance of these terms is determined by the positive real constant γ. In the case of noisy data one avoids overfitting by taking a smaller γ value. SVM problem formulations of this form have been investigated independently in (Saunders *et al.*, 1998) (without bias term) and (Suykens and Vandewalle, 1999).

In primal weight space one has the model

$$y(x) = w^T \varphi(x) + b. \tag{3}$$

The weight vector w can be infinite dimensional, which makes a calculation in w from (**??**) impossible in general. Therefore, one computes the model in the dual space instead of the primal space. One defines the Lagrangian

$$\mathcal{L}(w, b, e; \alpha) = \mathcal{J}(w, e) - \sum_{k=1}^{N} \alpha_k \left(w^T \varphi(x_k) + b + e_k - y_k \right), \tag{4}$$

with Lagrangian multipliers $\alpha_k \in \mathbb{R}$ (called support values). The conditions for optimality are given by

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum\limits_{k=1}^{N} \alpha_k \varphi(x_k) \\ \dfrac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum\limits_{k=1}^{N} \alpha_k = 0 \\ \dfrac{\partial \mathcal{L}}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k, \qquad k = 1, \ldots, N \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \rightarrow w^T \varphi(x_k) + b + e_k = y_k, \qquad k = 1, \ldots, N. \end{cases} \tag{5}$$

These conditions are similar to standard SVM optimality conditions, except for the condition $\alpha_k = \gamma e_k$. At this point one looses the sparseness property in LS-SVM's (Girosi, 1998), however it can be imposed in other ways as shown in (Suykens *et al.*, 2002a). After elimination of w, e one obtains the solution

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + \frac{1}{\gamma} I_N \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \tag{6}$$

with $y = (y_1, ..., y_N)^T$, $1_N = (1, ..., 1)^T$, $\alpha = (\alpha_1, ..., \alpha_N)^T$ and $\Omega_{kl} = \varphi(x_k)^T \varphi(x_l)$ for $k, l = 1, ..., N$. According to Mercer's theorem, there exists a mapping φ and an expansion

$$K(t, z) = \sum_i \varphi_i(t)\, \varphi_i(z) \qquad t, z \in \mathbb{R}^n, \tag{7}$$

if and only if, for any $g(t)$ such that $\int g(t)^2\, dt$ is finite, one has

$$\int_C \int_C K(t, z)\, g(t)\, g(z)\, dtdz \geq 0, \tag{8}$$

where C is a compact set in $\mathbb{R}^n$. As a result, one can choose a positive definite kernel $K(.,.)$ such that

$$K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l), \qquad k, l = 1, ..., N. \tag{9}$$

The resulting LS-SVM model for function estimation becomes

$$\hat{f}(x) = \sum_{k=1}^{N} \hat{\alpha}_k K(x, x_k) + \hat{b}, \tag{10}$$

where $\hat{\alpha}$, $\hat{b}$ are the solution to (**??**)

$$\hat{b} = \frac{1_N^T \left(\Omega + \frac{1}{\gamma} I_N\right)^{-1} y}{1_N^T \left(\Omega + \frac{1}{\gamma} I_N\right)^{-1} 1_N} \tag{11}$$

$$\hat{\alpha} = \left(\Omega + \frac{1}{\gamma} I_N\right)^{-1} \left(y - 1_N \hat{b}\right). \tag{12}$$

2.2 Smoother matrix

In this paper, we focus on the choice of an RBF kernel $K(x_k, x_l; h) = \exp\left\{-\|x_k - x_l\|_2^2 / h^2\right\}$. Let $\theta = (h, \gamma)^T$ and for all $\{x_k, y_k\}_{k=1}^N \in \mathcal{D}$, one has

$$\begin{aligned} \hat{f}_\theta &= \Omega\alpha + 1_N b \\ &= \left[\Omega\left(Z^{-1} - Z^{-1}\frac{J_N}{c}Z^{-1}\right) + \frac{J_N}{c}Z^{-1}\right] y \\ &= S(\theta) y, \end{aligned} \tag{13}$$

where $c = 1_N^T \left(\Omega + \frac{1}{\gamma} I_N \right)^{-1} 1_N$, $Z = (\Omega + \frac{1}{\gamma} I_N)$, J_N is a square matrix with all elements equal to 1 and $\hat{f}_\theta = (\hat{f}_\theta(x_1), \ldots, \hat{f}_\theta(x_N))^T$. The LS-SVM for regression corresponds to the case with $\hat{f}_\theta$ defined by (**??**) and

$$S(\theta) = \Omega \left(Z^{-1} - Z^{-1} \frac{J_N}{c} Z^{-1} \right) + \frac{J_N}{c} Z^{-1}. \tag{14}$$

Therefore, the LS-SVM for regression is an example of a linear smoother. This is because the estimated function in (**??**) is a linear combination of the y elements. $S(\theta)$ is known as the smoother matrix. Linear operators are familiar in linear regression (least squares fitting), where the fitted values $\hat{y}$ can be expressed as linear combinations of the output (dependent) variable y with the elements of the matrix that involves only the observations on the input (independent) variable u. Here the linear operator $H(u) = u(u^T u)^{-1} u^T$ is a projection operator also known as the hat matrix in statistics. There are some important similarities and differences between the hat matrix $H(u)$ and the smoother matrix $S(\theta)$. Both matrices are symmetric, positive semidefinite and the hat matrix is idempotent ($S^2 = S$) while the smoother matrix $S(\theta)^T S(\theta) \leqslant S(\theta)$, (meaning that $S^T S - S \leqslant 0$ is negative semidefinite). This is a consequence of the shrinking nature of $S(\theta)$. The trace of $H(u)$ gives the dimension of the projection space, which is also the number of parameters involved in the fit. By analogy one defines the effective degrees of freedom of the LS-SVM for regression (effective number of parameters) to be

$$d_{eff}(\theta) = \operatorname{tr}\left[S(\theta) \right]. \tag{15}$$

Another important property of the smoother matrix, based on an RBF kernel, is that $\operatorname{tr}\left[S(\theta) \right] < N$, except in the case ($h \rightarrow 0$, $\gamma \rightarrow \infty$) where $\operatorname{tr}[S(\theta)] \rightarrow N$.

2.3 Robust estimation

In order to obtain a robust estimate based upon the previous LS-SVM solution, in a subsequent step, one can weight the error variables $e_k = \alpha_k / \gamma$ by weighting factors v_k (Suykens *et al.*, 2002a). This leads to the optimization problem:

$$\min_{w^*, b^*, e^*} \mathcal{J}(w^*, e^*) = \frac{1}{2} w^{*T} w^* + \frac{1}{2} \gamma \sum_{k=1}^{N} v_k e_k^{*2} \tag{16}$$

such that $y_k = w^{*T}\varphi(x_k) + b^* + e_k^*, \quad k = 1, \ldots, N$. The Lagrangian is constructed in a similar way as before. The unknown variables for this weighted LS-SVM problem are denoted by the $*$ symbol. From the conditions for optimality and elimination of w^*, e^* one obtains the Karush-Kuhn-Tucker system:

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + V_\gamma \end{array}\right] \left[\begin{array}{c} b^* \\ \hline \alpha^* \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \tag{17}$$

where the diagonal matrix V_γ is given by $V_\gamma = \text{diag}\left\{\frac{1}{\gamma v_1}, ..., \frac{1}{\gamma v_N}\right\}$. The choice of the weights v_k is determined based upon the error variables $e_k = \alpha_k/\gamma$ from the (unweighted) LS-SVM case of Eq. (**??**). Robust estimates are obtained then (Rousseeuw and Leroy, 1986) e.g. by taking

$$v_k = \begin{cases} 1 & \text{if } \ |e_k/\hat{s}| \leq c_1 \\ \frac{c_2 - \ |e_k/\hat{s}|}{c_2 - c_1} & \text{if } \ \mathrm{c}_1 \leq |e_k/\hat{s}| \leq c_2 \\ 10^{-4} & \text{otherwise} \end{cases} \tag{18}$$

where $\hat{s} = 1.483 \text{ MAD}(e_k)$ is a robust estimate of the standard deviation of the LS-SVM error variables e_k and MAD stands for the median absolute deviation.

3 Classical Cross-validation Methods

3.1 Leave-one-out cross-validation score function

Next, we will motivate the cross-validation procedure as the θ selection rule. Let the distance $\Delta_{ISE}\left[f(x), \hat{f}(x;\theta)\right]$ denote the integrated squared error measure of accuracy for the estimator $\hat{f}(x;\theta)$. Write

$$\begin{aligned} \Delta_{ISE}\left[f(x), \hat{f}(x;\theta)\right] &= \int \left[f(x) - \hat{f}(x;\theta)\right]^2 g(x)\,dx \\ &= \int f^2(x)\,g(x)\,dx + \int \hat{f}^2(x;\theta)\,g(x)\,dx \\ &\quad -2\int f(x)\,\hat{f}(x;\theta)\,g(x)\,dx. \end{aligned} \tag{19}$$

Since the first term is independent of θ, minimizing this loss is equivalent to that of minimizing

$$Q = \int \hat{f}^2(x;\theta)\, g(x)\, dx - 2\int f(x)\, \hat{f}(x;\theta)\, g(x)\, dx. \tag{20}$$

Because this quantity depends on the unknown real-valued function $f(x)$ and $g(x)$ the density function over the input space. The first term of (**??**) can be computed entirely from the data, and the second term of (**??**) may be written as

$$Q_2 = \int f(x)\, \hat{f}(x;\theta)\, g(x)\, dx = E_{(x,y)}\left[y\hat{f}(x;\theta)\right]. \tag{21}$$

where the unknown $f(x)$ is replaced by the observations y at x, $E[y\,|x] = f(x)$. If we estimate (**??**) by $N^{-1}\sum_{k=1}^{N} y_k \hat{f}(x_k;\theta)$, the selection rule will be a biased estimator of $\Delta_{ISE}\left[f(x), \hat{f}(x;\theta)\right]$. The reason for the bias in the selection rule is that the observation y_k is used in $\hat{f}(x_k;\theta)$ to predict itself. This is equivalent to considering the apparent (resubstitution) estimate of the prediction error (Hastie *et al.*, 2001). There are several methods to find an unbiased estimate of $\Delta_{ISE}\left[f(x), \hat{f}(x;\theta)\right]$, for example: a plug-in method, leave-one-out technique and a modification such that bias terms cancel asymptotically. In this paper we will use the leave-one-out technique, in which one observation is left out. Therefore, a better estimator for (**??**) is

$$\hat{Q}_2 = \frac{1}{N}\sum_{k=1}^{N} y_k \hat{f}^{(-k)}(x_k;\theta), \tag{22}$$

where $\hat{f}^{(-k)}(x_k;\theta)$ denotes the leave-one-out estimator with point k left out from the training. Similarly, the first term of (**??**) may be approximated by

$$\hat{Q}_1 = \frac{1}{N}\sum_{k=1}^{N} \left(\hat{f}^{(-k)}(x_k;\theta)\right)^2. \tag{23}$$

From (**??**) and (**??**), the cross-validation function is

$$CV(\theta) = \frac{1}{N}\sum_{k=1}^{N} \left(y_k - \hat{f}^{(-k)}(x_k;\theta)\right)^2. \tag{24}$$

The above motivation is related to some ideas of (Rudemo, 1982) and (Bowman, 1984). In the context of kernel smoothing this score function for finding the bandwidth was proposed by (Clark, 1975). Wahba and Wold (1975) proposed a similar technique in the context of spline smoothing. The least squares cross-validated choice of θ for the LS-SVM estimates, based on the average squared prediction error, is the minimizer of

$$\inf_{\theta} CV(\theta) = \frac{1}{N} \sum_{k=1}^{N} (y_k - \hat{f}^{(-k)}(x_k;\theta))^2. \tag{25}$$

3.2 Generalized cross-validation score function

The GCV criterion was first proposed by (Craven and Wahba, 1979) for the use in the context of nonparametric regression with a roughness penalty. However, (Golub *et al.*, 1979) showed that GCV can be used to solve a wide variety of problems involving estimation of minimizers for (**??**).

In the leave-one-out cross-validation it is necessary to solve N separate LS-SVM's, in order to find the N models $\hat{f}^{(-k)}\left(x_k;\theta\right)$. From (**??**), the values of the LS-SVM $\hat{f}\ \left(x_k;\theta\right)$ depend linearly on the data y_k. We can write the deleted residuals $y_k - \hat{f}^{(-k)}\left(x_k;\theta\right)$ in terms of $y_k - \hat{f}\left(x_k;\theta\right)$ and the *k-th* diagonal element of the smoother matrix $S(\theta)$. The CV score function satisfies

$$CV\left(\theta\right) = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{y_k - \hat{f}\left(x_k;\theta\right)}{1 - s_{kk}(\theta)} \right)^2, \tag{26}$$

where $\hat{f}\left(x_k;\theta\right)$ is the LS-SVM calculated from the full data set $\{(x_k, y_k)\}_{k=1}^{N}$ and $s_{kk}(\theta)$ is the k-th diagonal element of the smoother matrix. The proof of (**??**) is identical to the one obtained in the development of the PRESS criterion for deciding about the complexity for parametric multivariate regression; see (Cook and Weisberg, 1982). Assuming that $\mathrm{tr}[S(\theta)] < N$ and $s_{kk} < 1$, $\forall k$, the basic idea of Generalized cross-validation is to replace the factors $1 - s_{kk}(\theta)$ by their average value, $1 - N^{-1}\mathrm{tr}[S(\theta)]$. The Generalized cross-validation score is then constructed, by analogy with ordinary cross-validation, by summing the squared residuals corrected by the square of $1 - N^{-1}\mathrm{tr}[S(\theta)]$. Since $1 - N^{-1}\mathrm{tr}[S(\theta)]$ is the same for all k, we obtain

$$GCV\left(\theta\right) = \frac{1}{N} \frac{\sum_{k=1}^{N} \left(y_k - \hat{f}\left(x_k;\theta\right)\right)^2}{\left(1 - N^{-1}\mathrm{tr}\left[S(\theta)\right]\right)^2}. \tag{27}$$

As in ordinary cross-validation, the GCV choice of the tuning parameters is then carried out by minimizing the function $GCV(\theta)$ over θ.

3.3 V-fold cross-validation score function

In general there is no reason that training sets should be of size $N-1$. There is the possibility that small perturbations, when single observations are left out, make $CV(\theta)$ too variable, if fitted values $\hat{f}(x;\theta)$ do not depend smoothly on the empirical distribution F_N or if the loss function $L\left[y,\hat{f}(x;\theta)\right]$ is not continuous. These potential problems can be avoided to a large extent by leaving out groups of observations, rather than single observations. We begin by splitting the data randomly into V disjoint sets of nearly equal size. Let the size of the vth group be m_v and assume that $\lfloor N/V \rfloor \leq m_v \leq \lfloor N/V \rfloor + 1$ for all v. For η real, $\lfloor \eta \rfloor$ denotes the greatest integer less or equal to η. For each such split we apply (**??**), and then average these estimates. The result is the V -fold cross-validation estimate of prediction error

$$CV_{V-fold}(\lambda,\gamma) = \sum_{v=1}^{V} \frac{m_v}{N} \sum_{k=1}^{m_v} \frac{1}{m_v} \left(y_k - \hat{f}^{(-m_v)}(x_k;\lambda,\gamma)\right)^2, \tag{28}$$

where $\hat{f}^{(-m_v)}$ represents the model obtained from the data outside group v. Practical experience suggests that a good strategy is to take $V = \min\left(\sqrt{N}, 10\right)$, because taking $V > 10$ may be computationally too expensive when the prediction rule is complicated, while taking groups of size at least $\sqrt{N}$ should perturb the data sufficiently to give small variance of the estimate (Davison and Hinkley, 1997). The use of groups will have the desired effect of reducing variance, but at the cost of increasing bias. According to (Beran, 1984), (Serfling, 1984) and (Burman, 1989), the bias of $CV_{V-fold}(\theta) \approx a_0 \left[(V-1)^{-1} N^{-1}\right]$, for $V = N$ (leave-one-out) the bias is of order $O(N^{-2})$, but when V is small, the bias term is not necessarily very small. The term a_0, depending on L and F_N, is of order the number of parameters being estimated. For LS-SVM, a_0 becomes a constant multiplied with the number of effective parameters. Therefore, if the number of effective parameters is not small, the $CV_{V-fold}(\theta)$ is a poor estimate of the prediction error. But the bias of $CV_{V-fold}(\theta)$ can be reduced by a simple adjustment (Burman, 1989). The adjusted V-fold cross-validation estimate of prediction error is

$$CV^{adj}_{V-fold}(\theta) = CV_{V-fold}(\theta) + \left[\frac{1}{N}\sum_{k=1}^{N}\left(y_k - \hat{f}(x_k;\theta)\right)^2 - \sum_{v=1}^{V}\frac{m_v}{N}\sum_{k=1}^{N}\frac{1}{N}\left(y_k - \hat{f}^{(-m_v)}(x_k;\theta)\right)^2\right]. \tag{29}$$

The bias of $CV^{adj}_{V-fold}(\theta) \approx a_1\left[(V-1)^{-1}N^{-2}\right]$, for some constant a_1 depending on L and F_N. The $CV^{adj}_{V-fold}(\theta)$ has a smaller bias than $CV_{V-fold}(\theta)$ and works better asymptotically as N increases. The $CV^{adj}_{V-fold}(\theta)$ is almost as simple to calculate, because it requires no additional LS-SVM fits.

4 Further Analysis of the V-fold Cross-Validation Score Function

The cross-validation procedure can basically be split up into two main parts: (a) constructing and computing the cross-validation score function, and (b) finding the tuning parameters by $\theta^* = \text{argmin}_\theta\left[CV_{V-fold}(\theta)\right]$. In this paper we focus on (a).

Let $\{z_k = (x_k, y_k)\}_{k=1}^N$ be an independent identically distributed (i.i.d.) random sample from some population with distribution function $F(z)$. Let $F_N(z)$ be the empirical estimate of $F(z)$. Our goal is to estimate a quantity of the form

$$T_N = \int L(z, F_N(z))\, dF(z), \tag{30}$$

with $L(\cdot)$ the loss function (e.g. the L_2 or L_1 norm) and where $E[T_N]$ could be estimated by cross-validation. We begin by splitting the data randomly into V disjoint sets of nearly equal size. Let the size of the v-th group be m_v and assume that $\lfloor N/V\rfloor \leq m_v \leq \lfloor N/V\rfloor + 1$ for all v. Let $F_{(N-m_v)}(z)$ be the empirical estimate of $F(z)$ based on $(N-m_v)$ observations outside group v and let $F_{m_v}(z)$ be the empirical estimate of $F(z)$ based on m_v observations inside group v. Then a general form of the V-fold cross-validated estimate of T_N is given by

$$CV_{V-fold}(\theta) = \sum_{v=1}^{V}\frac{m_v}{N}\int L\left(z, F_{(N-m_v)}(z)\right) dF_{m_v}(z). \tag{31}$$

Let $\hat{f}^{(-m_v)}(x;\theta)$ denote the regression estimate based on the $(N-m_v)$ observations outside the group v. Then the least squares V-fold cross-validated estimate of T_N is given by

$$CV_{V-fold}(\theta) = \sum_{v=1}^{V} \frac{m_v}{N} \sum_{k=1}^{m_v} \frac{1}{m_v} \left(y_k - \hat{f}^{(-m_v)}(x_k;\theta) \right)^2 . \tag{32}$$

The cross-validation score function can be written as a function of the number of $(V+1)$ means. It estimates a location parameter of the corresponding v-samples. Let $\xi = L(\vartheta)$ be a function of a random variable ϑ. In the V-fold cross-validation case, a realization of the random variable ϑ is given by $\vartheta_k = \left(y_k - \hat{f}^{(-m_v)}(x_k;\theta) \right)$, $k = 1,...,m_v$ $\forall v$ with

$$\begin{aligned} CV_{V-fold}(\theta) &= \sum_{v=1}^{V} \frac{m_v}{N} \left(\frac{1}{m_v} \sum_{k=1}^{m_v} L(\vartheta_k) \right) = \sum_{v=1}^{V} \frac{m_v}{N} \left(\frac{1}{m_v} \sum_{k=1}^{m_v} \xi_k \right) \\ &= \hat{\mu}\left(\hat{\mu}_1 \left(\xi_{11},...,\xi_{1m_1} \right),...,\hat{\mu}_V \left(\xi_{V1},...,\xi_{Vm_v} \right) \right), \end{aligned} \tag{33}$$

where ξ_{vj} denotes the j-th element of the v-th group, $\hat{\mu}_v(\xi_{v1},...,\xi_{vm_1})$ denotes the sample mean of the v-th group and $\hat{\mu}$ is the mean of all sample group means. Consider only the random sample of the v-th group and let $F_{m_v}(\xi)$ be the empirical distribution function. Then $F_{m_v}(\xi)$ depends in a complicated way on the noise distribution $F(e)$, the θ values and the loss function $L(\cdot)$. In practice $F(e)$ is unknown except for the assumption of symmetry around 0 (see Figure **??**.a). Whatever the loss function would be (L_2 or L_1), the distribution $F_{m_v}(\xi)$ is always concentrated on the positive axis with an asymmetric distribution (see Figure **??**.b). The asymmetric distribution of $\hat{\mu}_1,...,\hat{\mu}_V$, denoted by $F(\hat{\mu}_V)$ is sketched in Figure **??**.c. There is a lot of variability in the V-fold cross validated estimate, because the number of ways that N random values can be grouped into V classes with m_v in the vth class, $i=1,...,V$, and $\sum_{i=1}^{V} m_v = N$ equals $\frac{N!}{m_1!m_2!...m_V!}$.

We propose now the following procedure. Permute and split repeatedly the data - e.g. r times - into V groups as discussed. Then the V-fold cross-validation score function is calculated for each split and finally take the average of the r estimates

$$\text{Repeated\_}CV_{V-fold}(\theta) = \frac{1}{r} \sum_{j=1}^{r} CV_{V-fold,j}(\theta) . \tag{34}$$

The distribution of the Repeated_$CV_{V-fold}(\theta)$ is asymptotically normally distributed (see Figure **??**.d). The repeated V-fold cross-validation score function has about the same bias as the V-fold cross-validation score function, but the average of r estimates is less variable than for one estimate. This is illustrated in Figure **??**.

5 Repeated Robust and Efficient V-fold Cross-validation Score Function

5.1 Robustness and efficiency of location estimators

Given a random sample $\Xi = \{\xi_1, \ldots, \xi_N\}$, the expected value of the random variable Ξ is its average value and can be viewed as an indication of the central value of the density function. The expected value is therefor sometimes referred to as a location parameter. The median of a distribution is also a location parameter which does not necessarily equals the mean. Other location estimators are e.g. the β-trimmed mean and the β-Winsorized mean.

In each step of the repeated-V-fold cross-validation score function the distribution of the corresponding estimates are known (Figure **??**). In the classical cross-validation the location parameter is always estimated by the mean. In Appendix **??** two classes of location estimators, M-estimators and L-estimators, are described.

Based on the knowledge of the distributions (Figure **??**.e), we can select the candidate location estimators. First, the Huber type M-estimators have been found to be quite robust (Andrews *et al.*, 1972). For asymmetric distributions on the other hand, the computation of the Huber type M-estimators requires rather complex iterative algorithms and its convergence cannot be guaranteed in some important cases (Marazzi and Ruffieux, 1996). Secondly, the trimmed mean can be used when the distribution is symmetric (β-trimmed mean) and even when the distribution is asymmetric ((β_1, β_2)-trimmed mean). The trimmed mean is closely related to Huber's score function (Jaeckel, 1971).

In order to understand why certain location estimators behave the way they do, it is necessary to look at the various measures of robustness (Appendix **??**). There exist a large variety of approaches towards the robustness problem. The approach based on influence functions (Hampel, 1968, 1974) will be used here. The effect of one outlier on the location estimator can be described by the influence function (IF) which (roughly speaking) formalizes the bias caused by one outlier. Another measure of robustness is how much contaminated data a location estimator can tolerate before it becomes useless. This aspect is covered by the breakdown point of the location estimator.

In studies on robustness, the Tukey contamination scheme is widely used (Tukey, 1960). The generalization of this so-called super model is given by

$$F = \{F : F_\epsilon(e) = (1-\epsilon)\, F_0(e) + \epsilon G(e), \quad 0 \le \epsilon \le 1\}, \tag{35}$$

where $F_0(e)$ is some given distribution (the ideal nominal model), $G(e)$ is an arbitrary continuous distribution and ϵ is the first parameter of contamination. The contamination scheme describes the case where, with large probability $(1-\epsilon)$ the data occurs with distribution $F_0(e)$ and with small probability ϵ outliers occur according to the distribution $G(e)$. Examples of the super model are:

- *Example* 1: Super model with symmetric contamination

$$F_\epsilon(e) = (1-\epsilon)\,\mathcal{N}(0,\sigma^2) + \epsilon\,\mathcal{N}(0,\kappa^2\sigma^2), \quad 0 \le \epsilon \le 1, \quad \kappa > 1. \tag{36}$$

- *Example 2*: Super model for the mixture of the normal and Laplace or double exponential distribution

$$F_\epsilon(e) = (1-\epsilon)\,\mathcal{N}(0,\sigma^2) + \epsilon\,\mathrm{Lap}(0,\lambda), \quad 0 \le \epsilon \le 1, \tag{37}$$

where respectively κ and λ are the second parameters of contamination describing the rate of variance of $G(e)$ over the variance of $F_0(e)$ ($\kappa \gg 1$).

Given two estimates $T_1(F_N)$ and $T_2(F_N)$ of a location parameter $T(F)$, it would be sensible to choose the estimate whose sampling distribution is most highly concentrated around the true parameter value. A quantitative measure of such concentration is the mean squared error (MSE). The efficiency of $T_2(F_N)$ relative to $T_1(F_N)$ is defined to be

$$\mathrm{Eff}\,(T_2(F_N), T_1(F_N)) = \frac{Var\,(T_1(F_N)) + (E(T_1(F_N)) - T(F))^2}{Var\,(T_2(F_N)) + (E(T_2(F_N)) - T(F))^2}. \tag{38}$$

In Appendix **??**, two examples are given.

5.2 Repeated robust V-fold cross-validation score function

A classical cross-validation score function with L_2 or L_1 works well in situations where many assumptions (such as $e_k \sim N\left(0, \sigma^2\right)$, $E\left[e_k^2\right] < \infty$ and no outliers) are valid. These assumptions are commonly made, but are usually at best approximations to reality. For example, non-Gaussian noise and outliers are common in data-sets and are dangerous for many statistical procedures and also for the cross-validation score function. Given the previous derivations of robustness and efficiency, a new variant of the classical cross-validation score function is introduced based on the trimmed mean. There are several practical reasons to use this type of robust estimator, which is the least squares solution after discarding (in our case) the $g_2 = \lfloor N\beta_2 \rfloor$ largest observations:

- The trimmed mean can be applied when the sample distribution is symmetric or asymmetric.

- It is easy to compute. It is a reasonable descriptive statistic, which can be used as an estimator of the mean of the corresponding truncated distribution.

- For large N, the trimmed mean has an approximate normal distribution (Bickel and Peter, 1965). The standard deviation can be estimated based on the Winsorized sum of squares (Huber, 1970).

- It can be used as an adaptive statistic.

The general form of the V-fold cross-validation score function based on the sample mean is given in (**??**). The robust V-fold cross-validation score function based on the trimmed mean is formulated as

$$CV_{V-fold}^{Robust}\left(\theta\right) = \sum_{v=1}^{V} \frac{m_v}{N} \int_0^{F^{-}(1-\beta_2)} L\left(z, F_{(N-m_v)}\left(z\right)\right) dF_{m_v}\left(z\right). \tag{39}$$

Let $\hat{f}_{Robust}\left(x;\theta\right)$ be a regression estimate constructed via a robust method, for example the weighted LS-SVM as explained in Section 2.3. Then the least squares robust V-fold cross-validation estimate is given by

$$\begin{aligned} CV_{V-fold}^{Robust}(\theta) &= \sum_{v=1}^{V} \frac{m_v}{N} \sum_{k}^{m_v} \frac{1}{m_v - \lfloor m_v \beta_2 \rfloor} \left(y_k - f_{Robust}^{(-m_v)}(x_k;\theta) \right)_{m_v(k)}^2 \\ & \qquad \delta_{[m_v(1), m_v(m_v - \lfloor m_v \beta_2 \rfloor)]}((y_k - f_{Robust}^{(-m_v)}(x_k;\theta))^2), \end{aligned} \tag{40}$$

where $(y_k - f_{Robust}^{(-m_v)}(x_k;\theta))_{m_v(k)}^2$ is an order statistic and the indicator function $\delta_{[a,b]}(z) = 1$ if $a < z < b$ and otherwise 0.

The robust V-fold cross-validation score function can also be written as

$$CV_{V-fold}^{Robust}(\theta) = \hat{\mu}\left(\hat{\mu}_{(0,\beta_{2,1})}\left(\xi_{m_1(1)}, ..., \xi_{m_1(m_1)}\right), ..., \hat{\mu}_{\left(0,\beta_{2,V}\right)}\left(\xi_{m_V(1)}, ..., \xi_{m_V(m_V)}\right)\right). \tag{41}$$

It estimates a location parameter of the v-samples, where $\hat{\mu}_{(0,\beta_{2,v})}\left(\xi_{m_v(1)}, ..., \xi_{m_v(m_v)}\right)$ is the sample $(0,\beta_2)$-trimmed mean of the v-th group, and $\hat{\mu}$ is the mean of all the sample group $(0,\beta_2)$-trimmed mean. To use a $(0,\beta_2)$-trimmed mean, one must decide on a value of β_2. Guidelines for selection of this value can be found in (Hogg, 1974). If one is particularly concerned with good protection against outliers and if from past experience one has an idea about the frequency of occurrence of such outliers (5 to 10% is typical for many types of data) one would choose a value β_2 somewhat above the expected proportion of outliers.

Similar as presented in Section **??** for the V-fold CV score function, the data is permuted and splitted repeatedly - e.g. r times - into V groups. For each split, the robust V-fold cross-validation score function is calculated. The final result is the average of the r estimates. This procedure reduces the variance of the score function

$$\text{Repeated\_}CV_{V-fold}^{Robust}(\theta) = \frac{1}{r} \sum_{j=1}^{r} CV_{V-fold,j}^{Robust}(\theta). \tag{42}$$

Remark that the robust cross-validation score function inherents all nice properties of the trimmed mean and its IF has the same form as in Figure **??**.b.

The algorithm corresponding to the Repeated_$CV_{V-fold}^{Robust}(\theta)$ is described by following pseudo code:

Algorithm 1 - Repeated robust efficient cross-validation score function

```
for r = 1 to R,
  randomize data;
  split the data in V groups;
  for v = 1 to V,
    compute a robust regression estimator on all but the v-th group;
    compute the norm of its residuals of the v-th group;
    reject the (beta2 * n)-th largest results; % trimming
    compute the mean of the result;            % trimmed mean
  end for v
  the r-th score is the mean of the estimates;
end for r
the final result is the mean of all scores
```

Here `R`, `V` and `beta2` are the parameters corresponding to respectively R, V and β_2.

6 Robust Generalized Cross-validation Score Function

The best motivation for the generalized cross-validation (see Eq. (**??**)) is probably provided by the so-called GCV theorem first established by Craven and Wahba (1979) (see also Golub *et al.* 1979).

A natural approach to robustify the GCV is by replacing the linear procedure of averaging by the corresponding robust counterparts. Let $\xi = L(\vartheta)$ be a function of a random variable ϑ. In the GCV case, a realization of the random variable ϑ is given by

$$\vartheta_k = \left(\frac{y_k - f^*(x_k;\theta)}{1 - (1/\sum_k v_k) tr(S^*)} \right), \quad k = 1, \ldots, N \tag{43}$$

where $f^*(x_k;\theta)$ is the weighted LS-SVM as described in Section 2, the weighting of $f^*(x_k;\theta)$ corresponding with $\{x_k, y_k\}$ is denoted by v_k and the smoother matrix based on these weightings is defined as in Eq.(**??**) where Z is replaced by $Z^* = (\Omega + V_\gamma)$ with $V_\gamma =$

diag $\left\{\frac{1}{\gamma v_1},...,\frac{1}{\gamma v_N}\right\}$. The GCV can now be written as

$$GCV(\theta) = \frac{1}{N}\sum_{k=1}^{N} L(\vartheta_k) = \frac{1}{N}\sum_{k=1}^{N} \vartheta_k^2. \tag{44}$$

Using a robust analog of the sum ($(0, \beta_2)$ - trimmed mean), the robust GCV is defined by

$$GCV_{robust}(\theta) = \frac{1}{N - \lfloor N\beta_2 \rfloor} \sum_{k=1}^{N-\lfloor N\beta_2 \rfloor} \delta_{[\vartheta_{N(1)}, \vartheta_{N(N-\lfloor N\beta_2 \rfloor)}]}(\vartheta^2) \tag{45}$$

where $\delta_{[\cdot,\cdot]}(\cdot)$ is an indicator function.

7 Illustrative examples

7.1 Artificial dataset

In this example we compare eight criteria: leave-one-out CV, $CV_{V-fold}^{L_2}$, $CV_{V-fold}^{L_1}$, Akaike information criterion (AIC), Bayesian information criterion (BIC), the repeated CV_{V-fold}^{robust}, GCV and robust GCV for use in tuning parameter selection of function estimation.

First, we show three examples of estimating a sinc function where the noise model is described by: (a) contamination noise as described in (**??**) of Section **??** (Figure **??**), (b) contamination noise as described in (**??**) of Section **??**. (Figure **??**) and (c) zero mean Gaussian noise model without contamination (Table **??**). Given is a training set with $N = 150$ data points. ¿From the simulation results it is clear that in all contaminated cases the LS-SVM tuned by the classical methods are outperformed by the robust methods for tuning the weighted LS-SVM (Figure **??** and **??**). With the proposed robust procedures, the contamination has practically no influence on the tuning parameter selection. An important property of these robust procedures is that in the non-contamination case (c), it performs equally well as the classical methods (see Table **??**).

A Monte Carlo simulation (this experiment is repeated 150 times) was carried out to compare the different criteria. The LS-SVM estimates are presented with tuning parameters selected by different criteria. Figure **??** gives the boxplots of the simulations for the contamination noise as described in (**??**) of Section **??**. Figure **??** gives the boxplots of the simulations for the contamination noise as described in (**??**) of Section **??**.

7.2 Real datasets

7.2.1 Body fat data

In the body fat dataset (Penrose *et al.*, 1985) the outcome variable `'body fat'` and the 18 input variables are recorded for 252 men. The last third part of permuted observations is used as independent test set to compare the obtained results as given in Table **??**. After examination of the data, the trimming proportion of the robust cross-validation procedure was set to 5%. The results show the improved performance of the proposed robust procedures in different norms (L_1, L_2 and L_∞).

7.2.2 Boston housing data

The Boston housing dataset (Harrison *et al.*, 1978) is composed of 506 objects. There are 13 continuous variables (including the response variable `'MEDV'`) and one binary valued variable. The last third part of permuted observations is used as independent test set to compare the obtained results as given in Table **??**.

8 Conclusion

Cross-validation methods are frequently applied for selecting tuning parameters in neural network methods, usually based on L_2 or L_1 norms. However, due to the asymmetric and non-Gaussian nature of the score function, better location parameters can be used to estimate the performance. In this paper we have introduced a repeated robust cross-validation score function method by applying concepts from robust statistics (e.g. the influence function and the breakdown point) to the cross-validation methodology. We have applied a similar technique to generalized cross-validation. Simulation results illustrate that these methods can be very effective, especially with non-Gaussian noise distributions and outliers on data where the L_2 methods usually fails. The proposed methods have a good robustness/efficiency trade-off such that they perform equally well in cases where L_2 would perform optimally.

References

Akaike, H. (1973). Statistical predictor identification. *Ann. Inst. Statist. Math.* 22, 203-217.

Allen, D.M. (1974). The relationship between variables selection and data augmentation and a method of prediction. *Technometrics* 16, 125-127.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J. Rogers, W.H., Tukey, J.W. (1972). *Robust estimation of location.* Princeton, New Jersey: Princeton University Press.

Beaton, A.E., Tukey, J.W. (1974). The fitting of power series, meaning polynomials illustrated on band-spectroscopic data. *Technometrics* 16, 147-185.

Beran, R.J. (1984). Jackknife approximation to bootstrap estimates. *Ann. Statist.* 12, 101-118.

Bickel, P.J., Lehman, E.L. (1975). Descriptive statistics for non-parametric models, *Ann. Statist.* 3, 1038-1045.

Bickel, P.J. (1965). On some robust estimates of location, *Ann. Math. Statist.* 36, 847-858.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika* 71, 353-360.

Bunke, O., Droge, B. (1984). Bootstrap and cross-validation estimates of the prediction error for linear regression models, *Ann. Statist.* 12, 1400-1424.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76, 3, pp.503-514.

Cherkassky, V., Mulier, F. (1998). *Learning from Data.* Wiley, New York.

Clark, R.M. (1975). A calibration curve for radio carbon dates. *Antiquity* 49 251-266.

Cook, R.D., Weisberg, S. (1982). Criticism and Influence in Regression, San Francisco, Jossey-Bass In Leinhardt, S. (ed.) *Sociological Methodology* 1982, Chapter 8.

Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.*, 31, 377-390.

Daniel, C. (1920). Observations weighted according to order. *Amer.J. Math.* 42, 222-236.

Davison, A.C., Hinkley, D.V., (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

Fernholz, L.T. (1983). *von Mises Calculus for Statistical Functionals*, Lecture Notes in Statistics, Springer-Verlag.

Feller, W. (1966). *An Introduction to Probability theory its Applications*, Vol. I (2nd ed.) and Vol. II, Wiley, New York.

Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6) 1455-1480.

Golub, G.H., Heath, M., Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, 21, 215-223.

Hampel, F.R., Ronchetti, E.M. Rousseeuw, P.J., Stahel, W.A. (1986). *Robust Statistics, the approach based on Influence Functions*, New York, Wiley.

Hampel, F.R. (1968). *Contributions to the Theory of Robust Estimation*, Ph.D. Thesis, University of California, Berkeley.

Hampel, F.R. (1974). The influence curve and its role in robust estimation, *J. Amer. Statist. Assoc.* 69, 383-393.

Härdle, W. (1989). *Applied Nonparametric Regression, Econometric Society Monographs*, Cambridge University Press.

Härdle, W., Chen, R. (1995). Nonparametric time series analysis, a selective review with examples. *In Proceedings of the 50th ISI Session* (Beijing, 1995) vol. 1, 375-394.

Härdle, W., Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, 13, 1465-1481.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*, Springer-Verlag, Heidelberg.

Hogg, R.V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *J. Amer. Statist. Assoc.*, 348, 909-927.

Huber, P.J. (1964). Robust Estimation of a Location Parameter, *Ann. Math. Statist.*, 35, 73-101.

Jaeckel, L. (1971). Robust estimation of location: symmetry and asymmetric contamination, *Annals of Mathematical Statistics*, 42, 1020-1034.

Mallows, C.L. (1973). Some comments on C_p, *Technometrics* 15, 661-675.

Marazzi, A., Ruffieux, C. (1996). *Implementing M-estimators of the Gamma Distribution*, Lecture Notes in Statistics, vol. 109. Springer-Verlag, Heidelberg.

Marron, J.S. (1987). What does optimal bandwidth selection means for nonparametric regression estimation *In Statistical Data Analysis Based on the L_1-Norm and Related Methods* (Ed. Dodge, Y.), 379-391. North Holland, Amsterdam.

Marron, J.S. (1989). Automatic smoothing parameter selection: a survey, *Empirical Econom.*, 13, 187-208.

Penrose, K., Nelson, A., Fisher, A. (1985). Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques, *Medicine and Science in Sports and Excercise*, 17(2), 189.

Pfanzagl, J. (1969). On measurability and consistency of minimum contrast estimates, *Metrika*, 14, 248-278.

Poggio, T. Girosi, F. (1990). Networks for approximation and learning, *Proceedings of the IEEE*, 78, 1481-1497.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimation, *Scand. J. Statist.* 9 65-78.

Saunders, C., Gammerman, A., Vovk, V. (1998). Ridge regression learning algorithm in dual variables. *Proc. of the 15th Int. Conf. on Machine learning (ICML'98)*, Morgan Kaufmann, 515-521.

Schwartz, G. (1979). Estimating the dimension of a model, *Ann. of Statist.* 6, 461-464.

Serfling, R.J. (1984). Generalized L-, M-, and R-statistics, *Ann. Statist.* 12, 76-86.

Stigler, S.M. (1973). The asymptotic distribution of the trimmed mean, *Ann. Statist.* 1, 472-477.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Statist. Soc. Ser. B* 36 111-147.

Suykens, J.A.K., Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters.* Vol. 9, 293-300.

Suykens, J.A.K., De Brabanter, J., Lukas, L., Vandewalle, J. (2002a). Weighted least squares support vector machines : robustness and sparse approximation, *Neurocomputing*, Special issue on fundamental and information processing aspects of neurocomputing, 48, 1-4, Oct. 2002, 85-105.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J. (2002b). *Least Squares Support Vector Machines*, World Scientific, Singapore.

Tukey, J.W. (1960). *A Survey of Sampling of Contaminated Distributions*, In Contributions to Probability and Statistics, Stanford University Press, Stanford CA.

van der Vaart, A.W. (1998). *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Vapnik, V.N. (1998). *Statistical Learning Theory*, John Wiley & Sons, INC.

Wahba, G., Wold, S. (1975). A completely automatic french curve: fitting spline functions by cross-validation. *Comm. Statist.* 4 1-17.

Yang, Y., Zheng, Z. (1992). Asymptotic properties for cross-validated nearest neighbour median estimates in non-parametric regression: the L_1-view. In *Probability and Statistics*, 242-257. World Scientific, Singapore.

A Location Estimators

A.1 M-estimators

In this subsection we briefly review the statistics which are obtained as solutions to equations. Often the equations result in an optimization procedure, e.g. in the case of maximum likelihood estimation (MLE), least squares estimation etc.. Such statistics are called M-estimates. An important subclass of M-estimates is introduced by Huber (Huber, 1964). A related class of statistics, L-estimates is treated in the next subsection.

Let $x_1, ..., x_N$ be a random sample from a distribution F with density $p(x-\xi)$, where ξ is the location parameter. Assume that F is a symmetric unimodal distribution, then ξ is the center of symmetry to be estimated. The M-estimator $\hat{\xi}_N$ of the location parameter is defined then as some solution of the following minimization problem

$$\hat{\xi}_N = \arg\min_{\xi} \sum_{k=1}^{N} \rho\left(x_k - \xi\right), \tag{46}$$

where $\rho\left(t\right)$ is an even non-negative function called the contrast function (Phanzagl, 1969); $\rho(x_k - \xi)$ is the measure of discrepancy between the observation x_k and the estimated center. For a given density p the choice $\rho\left(t\right) = -\log p\left(t\right)$ yields the MLE. It is convenient to formulate the M-estimators in terms of the derivative of the contrast function $\Psi\left(t\right) = d\rho(t)/dt$ called the score function. In this case, the M-estimator is defined as a solution to the following implicit equation

$$\sum_{k=1}^{N} \Psi\left(x_k - \hat{\xi}_N\right) = 0. \tag{47}$$

Well-known examples of location parameter estimators are:

- *Example 1*: For $\rho\left(t\right) = t^2$, one obtains the least squares solution by minimization of $\sum_{k=1}^{N}\left(x_k - \xi\right)^2$. The corresponding score function is $\Psi\left(t\right) = t$, $-\infty < t < \infty$. For this Ψ, the M-estimate is the sample mean. The contrast function and the corresponding score function are sketched in Figure **??**.a.

- *Example 2*: For $\rho(t) = |t|$, one obtains the least absolute values by minimization of $\sum_{k=1}^{N} |x_k - \xi|$. The corresponding score function is

$$\Psi(t) = \begin{cases} -1, & t < 0 \\ 0 & t = 0 \\ 1 & t > 0. \end{cases} \tag{48}$$

  The corresponding M-estimate is the sample median. The contrast function and the respectively score function are sketched in Figure **??**.b.

- *Example 3*: Huber considers minimization of $\sum_{k=1}^{N} \rho(x_k - \xi)$, where

$$\rho(t) = \begin{cases} \frac{1}{2}t^2 & |t| \leq c \\ c\,|t| - \frac{1}{2}c^2 & |t| > c. \end{cases} \tag{49}$$

  The score function is

$$\Psi(t) = \begin{cases} -c, & t < -c \\ t & |t| \leq c \\ c & t > c. \end{cases} \tag{50}$$

  The corresponding M-estimate is a type of Winsorized mean (explained in further detail in next subsection). It turns out to be the sample mean of the modified x_k's, where x_k becomes replaced by $\hat{\xi}_N \pm c$, whichever is nearer, if $\left|x_k - \hat{\xi}_N\right| > c$. The contrast function and score function are sketched in Figure **??**.a.

- *Example 4*: Hampel (Hampel, 1968, 1974) suggested a modification to the Huber estimator:

$$\Psi(t) = \begin{cases} t & 0 \leq |t| \leq a \\ a \text{ sign}(t) & a \leq |t| \leq b \\ a\left(\frac{c-|t|}{c-b}\right) \text{sign}(t) & b \leq |t| \leq c \\ 0 & |t| > c, \end{cases} \tag{51}$$

  making $\Psi(t)$ zero for $|t|$ sufficiently large. This M-estimator has the property of completely rejecting outliers. The contrast function and score function are sketched in Figure **??**.b.

- *Example* 5: A very smooth score function, the biweight was proposed by Tukey (Tukey, 1974) and has become increasingly popular. The score function is given by

$$\Psi\left(t\right)=t\left(a^{2}-t^{2}\right)^{2}\delta_{[-a,a]}(t), \tag{52}$$

where

$$\delta_{[-a,a]}=\begin{cases}1 & \text{if } t\in[-a,a]\\ 0 & \text{otherwise.}\end{cases} \tag{53}$$

The contrast function and the respectively score function are sketched in Figure **??**.c.

A.2 *L*-estimators

L-estimators were originally proposed by Daniel (Daniel, 1920) and since then have been forgotten for many years, with a revival now in robustness studies. The description of *L*-estimators can be formalized as follows.

Let $x_1,...,x_N$ be a random sample on a distribution F, the ordered sample values $x_{N(1)}\leq ...\leq x_{N(N)}$ are called the order statistics. A linear combination of (transformed) order statistics, or L-statistic, is a statistic of the form

$$\hat{\xi}_N=\sum_{j=1}^{N}C_{N(j)}a\left(x_{N(j)}\right), \tag{54}$$

for some choice of constants $C_{N(1)},...,C_{N(N)}$ where $\sum_{j=1}^{N}C_{N(j)}=1$ and $a(\cdot)$ is some fixed function. The simplest example of an L-statistic is the sample mean. More interesting, a compromise between mean and median (trade-off between robustness and asymptotic efficiency), is the β_2-trimmed mean (Figure **??**.a) defined as

$$\hat{\mu}_{(\beta_2)}=\frac{1}{N-2g}\sum_{j=g+1}^{N-g}x_{N(j)}, \tag{55}$$

where the trimming proportion β is selected so that $g=\lfloor N\beta_2\rfloor$ and $a(x_{N(j)})=x_{N(j)}$ is the identity function. The β-trimmed mean is a linear combination of the order statistics given zero weight to a number g of extreme observations at each end. It gives equal weight $1/(N-2g)$ to the number of $(N-2g)$ central observations. When F is no longer symmetric, it may sometimes be preferable to trim asymmetrically if the tail is expected to be heavier in one direction than the other. If the trimming proportions are β_1 on the left and β_2 on the right, the (β_1, β_2)-trimmed mean is defined as

$$\hat{\mu}_{(\beta_1,\beta_2)} = \frac{1}{N-(g_1+g_2)} \sum_{j=g_1+1}^{N-g_2} x_{N(j)}, \tag{56}$$

where β_1 and β_2 are selected so that $g_1 = \lfloor N\beta_1 \rfloor$ and $g_2 = \lfloor N\beta_2 \rfloor$. The (β_1, β_2)-trimmed mean is a linear combination of the order statistics giving zero weight to g_1 and g_2 extreme observations at each end and equal weight $1/(N - g_1 - g_2)$ to the $(N - g_1 - g_2)$ central observations.

Another L-estimator is the β-Winsorized mean (Figure **??**.b). Let $0 < \beta < 0.5$, then the β-Winsorized means (in the symmetric case) is defined as

$$\hat{\mu}_{W(\beta)} = \frac{1}{N}\left(g x_{N(g+1)} + \sum_{j=g+1}^{N-g} x_{N(j)} + g x_{N(N-g)} \right). \tag{57}$$

While the β-trimmed mean censors the smallest and largest $g = \lfloor N\beta \rfloor$ observations, the β-Winsorized means replaces each of them by the values of the smallest and the largest uncensored ones.

B Measures of Robustness

B.1 Influence function

Let F be a fixed distribution and $T(F)$ a statistical functional defined on a set $\mathcal{F}$ of distributions satisfying some regularity conditions (Hampel *et al.*, 1986). Statistics which are representable as functionals $T(F_N)$ of the sample distribution F_N are called statistical functions. For example, for the variance parameter σ^2 the relevant functional is $T(F) = \int \left[x - \int x dF(x)\right]^2 dF(x)$. Let the estimator $T_N = T(F_N)$ of $T(F)$ be the functional of the sample distribution F_N. Then the influence function $IF(x; T, F)$ is defined as

$$IF(x; T, F) = \lim_{\epsilon \downarrow 0} \frac{T\left[(1-\epsilon)F + \epsilon\Delta_x\right] - T(F)}{\epsilon}. \tag{58}$$

Here Δ_x denotes the pointmass distribution at the point x. The IF reflects the bias caused by adding a few outliers at the point x, standardized by the amount ϵ of contamination. Note that this kind of differentiation of statistical functionals is a differentiation in the sense of von Mises (Fernholz, 1983). From the influence function, several robustness measures can be defined: the gross error sensitivity, the local shift sensitivity and the rejection point (Hampel, 1968, 1974).

B.2 Breakdown point

The breakdown point ϵ^* of the estimator $T_N = T(F_N)$ for the functional $T(F)$ at F is defined by

$$\epsilon^*(T,F) = \sup\left\{\epsilon : \sup_{F:F=(1-\varepsilon)F_0+\varepsilon H} |T(F) - T(F_0)| < \infty \right\}. \tag{59}$$

This notion defines the largest fraction of gross errors that still keeps the bias bounded.

B.3 Influence function of the sample mean

The classical V-fold-cross-validation is based on the sample mean. The corresponding functional $T\left(F\right) = \int x dF\left(x\right)$ of the mean is defined for all probability measures with existing first moment. ¿From (**??**), it follows that

$$\begin{aligned} IF\left(x;T,F\right) &= \lim_{\epsilon\downarrow 0} \frac{\int xd\left[(1-\epsilon)F + \epsilon\Delta_x\right](x) - \int xdF(x)}{\epsilon} \\ &= \lim_{\epsilon\downarrow 0} \frac{\epsilon\int xd\Delta_x(x) - \epsilon\int xdF(x)}{\epsilon} \\ &= x - T(F_N). \end{aligned} \tag{60}$$

The IF of the sample mean is sketched in Figure **??**.a. We see that the IF is unbounded in $\mathbb{R}$. This means that an added observation at a large distance from $T(F_N)$ gives a large value in absolute sense for the IF. The finite sample breakdown point of the sample mean has $\epsilon_N^* = 1/N$, often the limiting value $\lim_{N\to\infty} \epsilon_N^* = 0$ will be used as a measure of the global stability of the estimator. One of the more robust location estimators is the median. Although the median is much more robust (breakdown point is 0.5) than the mean, its asymptotic efficiency is low. But in the asymmetric distribution case the mean and the median does not estimate the same quantity.

B.4 Influence function of the sample (β_1, β_2)-trimmed mean

Remark that the distribution $F_{m_v}(u)$ is asymmetric (Figure **??**.b) and has only a tail at the right, so we set $\beta_1 = 0$. The corresponding statistical functional for the $(0, \beta_2)$-trimmed mean is given in terms of a quantile function and is defined as

$$\mu_{(0,\beta_2)} = T_{(0,\beta_2)}(F) = \frac{1}{1-\beta_2} \int_0^{F^-(1-\beta_2)} x dF(x) = \frac{1}{1-\beta_2} \int_0^{(1-\beta_2)} F^-(q)\, d(q). \tag{61}$$

The *quantile function* of a cumulative distribution function F is the generalized inverse $F^- : (0,1) \to \mathbb{R}$ given by

$$F^-(q) = \inf \{x : F(x) \geq q\}. \tag{62}$$

In the absence of information concerning the underlying distribution function F of the sample, the empirical distribution function F_N and the empirical quantile function F_N^{-1} are reasonable estimates for F and F^-, respectively. The empirical quantile function is related to the order statistics $x_{N(1)} \leq ... \leq x_{N(N)}$ of the sample through

$$F^-(q) = x_{N(i)}, \quad \text{for } q \in \left(\frac{i-1}{N}, \frac{i}{N} \right). \tag{63}$$

Using the influence function of the qth quantile functional $F^-(q)$, an expression is derived for the influence function $IF(x; \mu_{(0,\beta_2)}, F)$ of the $(0, \beta_2)$-trimmed mean (see Appendix **??**):

$$IF(x; \mu_{(0,\beta_2)}, F) = \begin{cases} \frac{x - \beta_2 F^-(1-\beta_2)}{1-\beta_2} - \mu_{(0,\beta_2)} & 0 \leq x \leq F^-(1-\beta_2) \\ F^-(1-\beta_2) - \mu_{(0,\beta_2)} & F^-(1-\beta_2) < x. \end{cases} \tag{64}$$

The IF of the $(0, \beta_2)$-trimmed mean is sketched in Figure **??**.b. Note that it is both continuous and bounded. The finite sample breakdown point of the $(0, \beta_2)$-trimmed mean has $\epsilon_N^* = (\lfloor N\beta_2 \rfloor + 1)/N$ and the limiting value $\lim_{N\to\infty} \epsilon_N^* = \beta_2$.

C Efficiency Robustness Trade-off

C.1 Example for the mean and median

To get an idea of efficiency-robustness trade-off, we compare the mean and the median. Given a random sample $x_1, ..., x_N$ from a symmetric distribution $F(x)$ with density $p(x)$. Let the sample mean $\bar{X}_N$ and the sample median $\tilde{X}_N$, be the estimates of a location parameter $T(F)$. The sample mean is a non-robust location estimator while the sample median is the most robust location estimator. Based on their limit distributions $\sqrt{N}\left(\bar{X}_N - T(F)\right) \rightarrow N\left(0, \sigma^2\right)$ and $\sqrt{N}\left(\tilde{X}_N - T(F)\right) \rightarrow N\left(0, \frac{1}{4p^2(0)}\right)$, the asymptotic relative efficiency (ARE) is $4p^2(0)\sigma^2$. Putting $p(x) = \phi(x)$ (the standard normal density), we have $\sigma^2 = 1$ and $p(0) = 1/\sqrt{2\pi}$ so that $ARE(\tilde{X}_N, \bar{X}_N; \Phi(x)) = 2/\pi$,where $\Phi(x)$ denotes the standard normal distribution. The efficiency loss represented by $2/\pi = 0.637$ means that in the normal case the median would require about $N = 157$ observations to achieve the accuracy that the mean achieves with $N = 100$. Substituting $p(x)$ by $p(x\,|\mu, \lambda)$ the Laplace density $Lap(x)$, we have $\sigma^2 = 2\lambda^2$ and $p(0) = 1/2\lambda$, so that $ARE(\tilde{X}_N, \bar{X}_N; Lap(x)) = 2$. The larger relative efficiency of the median in this case reflects in fact that the tail of the Laplace distribution is heavier than that of the normal distribution.

C.2 Trade-off for mean and trimmed mean

Based on large-sample behavior of estimators (Van Der Vaart, 1998) and following the approach taken by Bickel and Lehmann (Bickel and Lehmann, 1975) we compare the sample mean and the sample $(0, \beta_2)$-trimmed mean both based on N non-negative random values $x_1, ..., x_N$. To determine the asymptotic relative efficiency, we require the asymptotic distribution of the estimators.

Consider distribution functions $F_1(.), F_2(.), ...$ and $F(.)$. Let $X_1, X_2, ...$ and X denote random variables having these distributions, respectively. A probability space is defined as a triple $(\Omega_p, \mathcal{A}, \mathcal{P})$ with $\Omega_p = \{\omega_1, ..., \omega_n\}$ a set of points, $\mathcal{A}$ an algebra of subsets of Ω_p and $\mathcal{P}$ a probability distribution or a measure defined on the elements of $\mathcal{A}$. The sequence

$\{X_n\}_{n\in\mathbb{N}}$ is said to converge in distribution (or in law) to a variable X, if

$$\lim_{n\to\infty} F_n(t) = F(t), \tag{65}$$

for each continuity point t of F. This is denoted as $X_n \xrightarrow{\mathcal{D}} X$, or $\mathcal{D}\text{-}\lim_{n\to\infty} X_n = X$. The convergences in probability, with convergence probability 1 and in rth mean each represent a sence in which, for n sufficiently large, $X_n(\omega)$ and $X(\omega)$ approximate each other as a function of ω, $\omega \in \Omega_p$. This means that the distribution of X_n and X cannot be to dissimilar, whereby approximation in distribution should follow. On the other hand, the convergence $\xrightarrow{\mathcal{D}}$ depends only on the distribution functions involved and does not require that X_n and X approximate each other as functions of ω (not necessarily on a common probability space). The Central Limit Theorem (CLT) pertains to the converge in distribution of sums of random variables.

Based on the most widely known version of CLT, the Lindeberg-Levy theorem, which is a special case of the Lindeberg-Feller theorem (Feller, 1966), states that the asymptotic distribution of the sample mean is given by

$$\sqrt{N}\left[\hat{\mu} - T(F)\right] \xrightarrow{\mathcal{D}} N(0, \tau^2) \tag{66}$$

and for the $(0, \beta_2)$-trimmed mean the asymptotic distribution is given by

$$\sqrt{N}\left[\hat{\mu}_{(0,\beta_2)} - T(F)\right] \xrightarrow{\mathcal{D}} N(0, \tau^2_{(0,\beta_2)}). \tag{67}$$

with $\tau^2_{(0,\beta_2)} = \frac{1}{(1-\beta_2)}\left(\sigma^2_{(0,\beta_2)} + \beta_2\left(F^-(1-\beta_2) - \hat{\mu}_{(0,\beta_2)}\right)^2\right)$, where $\hat{\mu}_{(0,\beta_2)} = \frac{1}{(1-\beta_2)}$ $\int_0^{F^-(1-\beta_2)} x dF(x)$ is the mean and $\sigma^2_{(0,\beta_2)} = \frac{1}{(1-\beta_2)}\int_0^{F^-(1-\beta_2)}\left(x - \hat{\mu}_{(0,\beta_2)}\right) dF(x)$ is the variance of the truncated distribution. A sketch of the proof and references to other approaches to the proof are given by Bickel (Bickel, 1965) and Stigler (Stigler, 1973). On the basis of (**??**) and (**??**), it is now possible to calculate the ARE of $\hat{\mu}_{(0,\beta_2)}$ to $\hat{\mu}$ by

$$ARE(\hat{\mu}_{(0,\beta_2)}, \hat{\mu}; F) = \frac{Var\left(\hat{\mu}\right)}{Var\left(\hat{\mu}_{(0,\beta_2)}\right)} = \frac{\tau^2}{\tau^2_{(0,\beta_2)}}. \tag{68}$$

The behavior of the relative efficiency is illustrated for $N = 500$ by a Monte Carlo technique (see Figure **??**.a and b).

D The IF of the (β_1, β_2)-trimmed mean

To derive the influence function $IF(x; F^{-}(q), F)$ for the qth quantile functional $F^{-}(q)$, assume that F has a density f which is continuous and positive at $x_q = F^{-}(q)$. Let $F_\epsilon = F + \epsilon\left(\Delta_x - F\right)$ and apply (**??**)

$$\begin{aligned} T\left[F + \epsilon\left(\Delta_{x_0} - F\right)\right] &= \inf\left\{x : F(x) + \epsilon\left(\Delta_{x0}(x) - F(x)\right) \geq q\right\} \\ &= \inf\left\{x : F(x) + \epsilon\left[I\left(x \geq x_0\right) - F(x)\right] \geq q\right\} \\ &= \inf\left\{x : F(x) \geq \frac{q - \epsilon\left[I\left(x \geq x_0\right)\right]}{(1-\epsilon)}\right\}. \end{aligned} \tag{69}$$

One finds $IF(x; F^{-}(q), F) = (\partial/\partial\epsilon)\left[F_\epsilon^{-1}(q)\right]_{\epsilon=0}$ indirectly by first calculating $(d/d\epsilon)\left[F_\epsilon^{-1}(q)\right]$ for $\epsilon > 0$ and then taking $\lim_{\epsilon\downarrow 0}(d/d\epsilon)\left[F_\epsilon^{-1}(q)\right]$. From (**??**) we have $F_\epsilon^{-1}(q) = F^{-}\left(\frac{q-\epsilon[I(x \geq x_0)]}{(1-\epsilon)}\right)$. Thus

$$\begin{aligned} \frac{d}{d\epsilon}\left[F^{-}\left(\frac{q - \epsilon\left[I\left(x \geq x_0\right)\right]}{(1-\epsilon)}\right)\right] &= \frac{\frac{d}{d\epsilon}\left(\frac{q-\epsilon[I(x\geq x_0)]}{(1-\epsilon)}\right)}{f\left(F^{-}\left(\frac{q-\epsilon[I(x\geq x_0)]}{(1-\epsilon)}\right)\right)} \\ &= \frac{q - I\left(x_o \leq F^{-}(q)\right)}{f\left(F^{-}\left(\frac{q-\epsilon[I(x\geq x_0)]}{(1-\epsilon)}\right)\right)}, \end{aligned}$$

so that

$$\lim_{\epsilon\downarrow 0}\frac{d}{d\epsilon}\left[F^{-}\left(\frac{q - \epsilon\left[I\left(x \geq x_0\right)\right]}{(1-\epsilon)}\right)\right] = \frac{q - I\left(x_0 \leqslant F^{-}(q)\right)}{f\left(F^{-}(q)\right)}. \tag{70}$$

The $IF(x; F^{-}(q), F)$ is

$$IF(x; F^{-}(q), F) = \begin{cases} \frac{(q-1)}{f(F^{-}(q))} & x_0 < F^{-}(q) \\ 0 & x_0 = F^{-}(q) \\ \frac{q}{f(F^{-}(q))} & x_0 > F^{-}(q). \end{cases} \tag{71}$$

Now we can calculate the influence function of the $(0, \beta_2)$-trimmed means. Define

$$\begin{aligned} T_{(0,\beta_2)}(F_\epsilon) &= \frac{1}{1-\beta_2}\int_0^{F_\epsilon^{-1}(1-\beta_2)} y dF_\epsilon(y) \\ &= \frac{1}{1-\beta_2}\left[\int_0^{F_\epsilon^{-1}(1-\beta_2)} y dF(y) + \epsilon\int_0^{F_\epsilon^{-1}(1-\beta_2)} y d\left(\Delta_x - F\right)(y)\right]. \end{aligned} \tag{72}$$

We will find $IF(x;\mu_{(0,\beta_2)},F) = (d/d\epsilon)\left[T_{(0,\beta_2)}(F_\epsilon)\right]_{\epsilon=0}$ indirectly by first calculating $(d/d\epsilon)\left[T_{(0,\beta_2)}(F_\epsilon)\right]$ for $\epsilon > 0$ and then taking $\lim_{\epsilon\downarrow 0}(d/d\epsilon)\left[T_{(0,\beta_2)}(F_\epsilon)\right]$. From (**??**)

$$\begin{aligned}\frac{d}{d\epsilon}\left[T_{(0,\beta_2)}(F_\epsilon)\right] &= \frac{F_\epsilon^{-1}(1-\beta_2)}{(1-\beta_2)}f\left(F_\epsilon^{-1}(1-\beta_2)\right)\frac{d}{d\epsilon}\left[F_\epsilon^{-1}(1-\beta_2)\right]\\ &\quad+(\frac{1}{1-\beta_2})\left[\int_0^{F_\epsilon^{-1}(1-\beta_2)} yd\left(\Delta_x - F\right)(y)\right]\\ &\quad+\epsilon(\frac{1}{1-\beta_2})\frac{d}{d\epsilon}\left[\int_0^{F_\epsilon^{-1}(1-\beta_2)} yd\left(\Delta_x - F\right)(y)\right], \qquad (73)\end{aligned}$$

so that

$$\begin{aligned}IF(x;\mu_{(0,\beta_2)},F) &= \lim_{\epsilon\downarrow 0}\frac{d}{d\epsilon}\left[T_{(0,\beta_2)}(F_\epsilon)\right]\\ &= \frac{F^-(1-\beta_2)}{(1-\beta_2)}f\left(F^-(1-\beta_2)\right)IF(x;F^-(q),F)\\ &\quad+\frac{1}{1-\beta_2}\left[\int_0^{F^-(1-\beta_2)} yd\Delta_x(y) - \int_0^{F^-(1-\beta_2)} ydF(y)\right]\\ &= \frac{F^-(1-\beta_2)}{(1-\beta_2)}f\left(F^-(1-\beta_2)\right)IF(x;F^-(q),F) - \mu_{(0,\beta_2)}\\ &\quad+\frac{x}{(1-\beta_2)}I\left(x \le F^-(1-\beta_2)\right). \qquad (74)\end{aligned}$$

Substituting the influence function $IF(x;F^-(q),F)$, with $q=(1-\beta_2)$, given in (**??**) into (**??**) yields:

$$IF(x;\mu_{(0,\beta_2)},F) = \begin{cases} \frac{x-\beta_2 F^-(1-\beta_2)}{1-\beta_2} - \mu_{(0,\beta_2)} & 0 \le x \le F^-(1-\beta_2)\\ F^-(1-\beta_2) - \mu_{(0,\beta_2)} & F^-(1-\beta_2) < x. \end{cases} \qquad (75)$$

Captions of Tables

Table ??: Numerical performance measured on fresh test data for the results of the *sinc* function without outliers. The results compare the performances of an LS-SVM on data with a Gaussian noise model $\mathcal{N}(0, 0.1^2)$ tuned by different performance criteria as described in Section **??**. The robust procedure performs equally well as the classical methods in the non-contamination case.

Table ??: Numerical performance measured on fresh test data for the body fat dataset. The results compare the performances of an LS-SVM on this real data tuned by different performance criteria: classical L_2 V-fold CV, generalized L_2 CV, repeated robust V-fold CV and robust generalized CV. The robust procedures (robust cross-validation and weighted LS-SVM) outperform the classical methods in this case.

Table ??: Numerical performance measured on fresh test data for the Boston housing dataset. The results compare the performances of an LS-SVM on this real data tuned by different performance criteria: classical L_2 V-fold CV, L_1 V-fold CV, generalized L_2 CV, repeated robust V-fold CV and robust generalized CV. The robust procedures (robust cross-validation and weighted LS-SVM) are slightly better (2%) than the results obtained by the L_2 and L_1 V-fold cross-validation procedure. AIC, BIC and generalized cross-validation perform significantly worse than the others.

method	L_2	L_1	L_∞
L_2 Loo-CV + LS-SVM	0.000587	0.020209	0.083482
L_2 V-fold CV + LS-SVM	0.000621	0.020686	0.093063
robust V-fold CV + weighted LS-SVM	0.000586	0.020399	0.076741
L_1 V-fold CV + LS-SVM	0.000644	0.020979	0.097678
AIC + LS-SVM	0.000645	0.021227	0.091463
BIC + LS-SVM	0.000687	0.022292	0.085469
GCV + LS-SVM	0.000645	0.021227	0.091463
robust GCV + weighted LS-SVM	0.000645	0.021227	0.091463

Table 1:

method	L_2	L_1	L_∞
L_2 V-fold CV + LS-SVM	0.0000209	0.00363	0.0136
L_1 V-fold CV + LS-SVM	0.0000166	0.00306	0.0156
AIC + LS-SVM	0.0000996	0.00819	0.0256
BIC + LS-SVM	0.0000996	0.00819	0.0256
L_2 generalized CV + LS-SVM	0.0000193	0.00323	0.0138
robust V-fold CV + weighted LS-SVM	0.0000014	0.00078	0.0046
robust generalized CV + weighted LS-SVM	0.0000084	0.00226	0.0101

Table 2:

method	L_2	L_1	L_∞
L_2 V-fold CV + LS-SVM	3.9974	1.5925	4.9841
L_1 V-fold CV + LS-SVM	3.9956	1.5918	4.9804
AIC + LS-SVM	6.6044	1.4824	18.6141
BIC + LS-SVM	11.6372	1.3864	29.0393
L_2 generalized CV + LS-SVM	4.7557	1.7083	5.4784
robust V-fold CV + weighted LS-SVM	3.9158	1.5846	5.0697
robust generalized CV + weighted LS-SVM	3.9316	1.5813	5.0104

Table 3:

Captions of Figures

Figure ??: Typical distributions of random variables related to the cross-validation score function:

(a) noise distribution. In practice $F(e)$ is unknown and assumed to have zero mean;

(b) squared residual distribution, $F_{m_v}(u)$ concentrated on the positive axis with an asymmetric distribution;

(c) asymmetric distribution $F(\hat{\mu}_v)$;

(d) sampling distribution of the repeated cross-validation based on the mean;

Figure ??: Schematic representation of the sampling distribution corresponding with one point in the tuning parameter space.

Figure ??: The contrast function and the score function of (a) the L_2 norm and (b) the L_1 norm.

Figure ??: The contrast function and the score function of Huber type of M-estimators.

(a) Huber contrast and score function;

(b) Hampel's modification to the Huber estimator. This M-estimator has the property of completely rejecting outliers;

(c) Tukey's biweight contrast and score function.

Figure ??: Influence Function of (a) the mean and (b) the $(0, \beta_2)$ trimmed mean.

(a) the IF of the mean is unbounded in $\mathbb{R}$. This means that an added observation at a large distance from $T(F_N)$ gives a large value in absolute sense for the IF;

(b) The IF of the $(0, \beta_2)$ trimmed mean is both continuous and bounded in $\mathbb{R}$.

Figure ??: Schematic representation on a symmetric distribution of (a) the β-trimmed mean and (b) the β-Winsorized mean.

Figure ??: The behavior of the asymptotic relative efficiency of the trimmed mean with respect to the mean for a varying contamination ϵ studied for a sample $N = 500$ by a Monte Carlo technique:

(a) efficiency for the case that F is a symmetric distribution and for several β_2's. This plot shows that a moderate amount of trimming can provide much better protection than the mean against fairly heavy tails. Almost no efficiency is lost in the non-contamination case of $\epsilon = 0$.

(b) efficiency for the case that F is asymmetric with heavier case than a Gaussian and for several β_2.

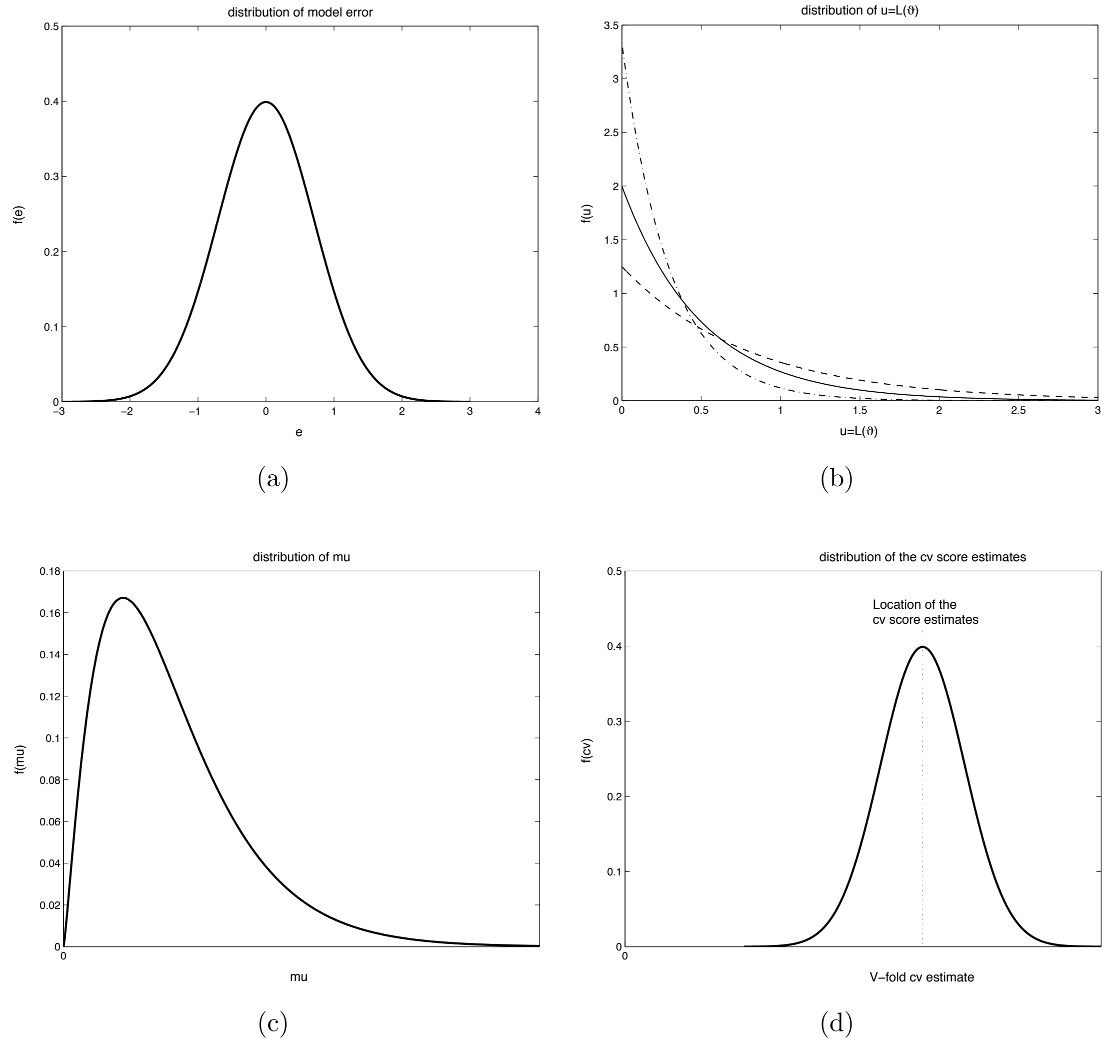
Figure ??: Experimental results comparing the performance of a LS-SVM and a weighted LS-SVM on artificial data (*sinc* function) with noise model $F_\epsilon(x) = (1-\epsilon)\mathcal{N}(0,\sigma^2) + \epsilon\mathcal{N}(0,\kappa^2\sigma^2)$ where $\epsilon = 0.15$, $\sigma^2 = 0.1^2$ and $\kappa^2 = 15$, tuned by (a) classical L_2 V-fold CV, (b) L_1 V-fold CV, (c) repeated robust CV and (d) robust generalized CV. In the L_2, L_1 and L_∞ norm, the best results are obtained by robust V-fold CV combined with the weighted LS-SVM.

Figure ??: Experimental results comparing the performance of a LS-SVM and a weighted LS-SVM on artificial data (*sinc* function) with noise model $F_\epsilon(x) = (1-\epsilon)\mathcal{N}(0,\sigma^2) +$

$\epsilon\, Lap(0, \lambda)$ where $\epsilon = 0.15$, $\sigma^2 = 0.1^2$ and $\lambda = 1$, tuned by (a) classical L_2 V-fold CV, (b) L_1 V-fold CV, (c) repeated robust CV and (d) robust generalized CV. In the L_2, L_1 and L_∞ norm, the best results are obtained by robust V-fold CV combined with the weighted LS-SVM.

Figure ??: The boxplots of the Monte Carlo simulations on artificial data (*sinc* function) for the contamination noise as described in (**??**) of Section **??** are shown. Each box in the figure gives the median and the standard deviation of the sample (a) MSE values, (b) Mean Absolute values of the errors and the (c) L_∞ norm of the errors, corresponding with the indicated method. The median of the MSE samples of the robust procedures outperform the others. In the L_2, L_1 and L_∞ norm, the best results are obtained by robust V-fold CV and robust GCV combined with the weighted LS-SVM.

Figure ??: The boxplots of the Monte Carlo simulations on artificial data (*sinc* function) for the contamination noise as described in (**??**) of Section **??** are shown. Each box in the figure gives the median and the standard deviation of the sample (a) MSE values, (b) Mean Absolute values of the errors and the (c) L_∞ norm of the errors, corresponding with the indicated method. The median of the MSE samples of the robust procedures outperform the others. Also the standard deviation of the sample performances is decreased using the robust methods. In the L_2, L_1 and L_∞ norm, the best results are obtained by robust V-fold CV and robust GCV combined with the weighted LS-SVM.



(a) (b)

(c) (d)

Figure 1:

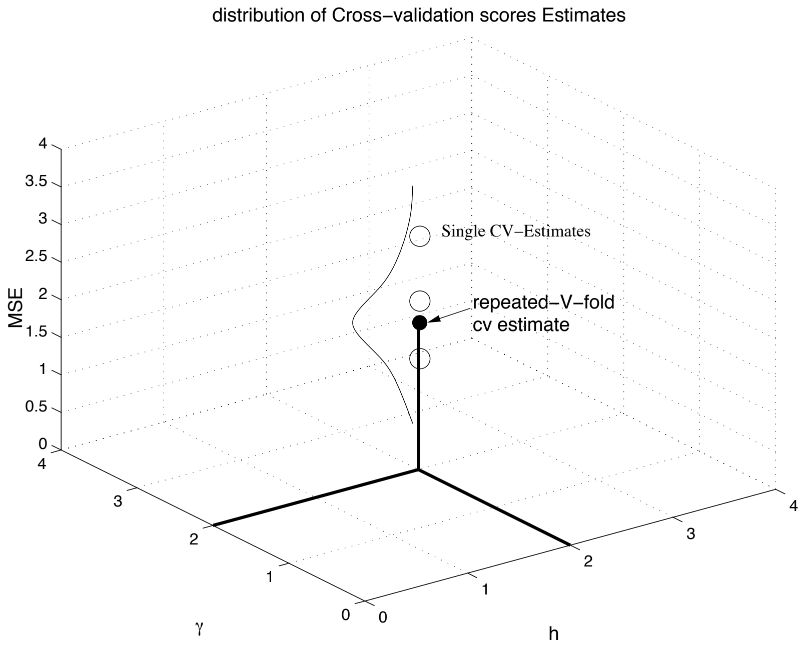
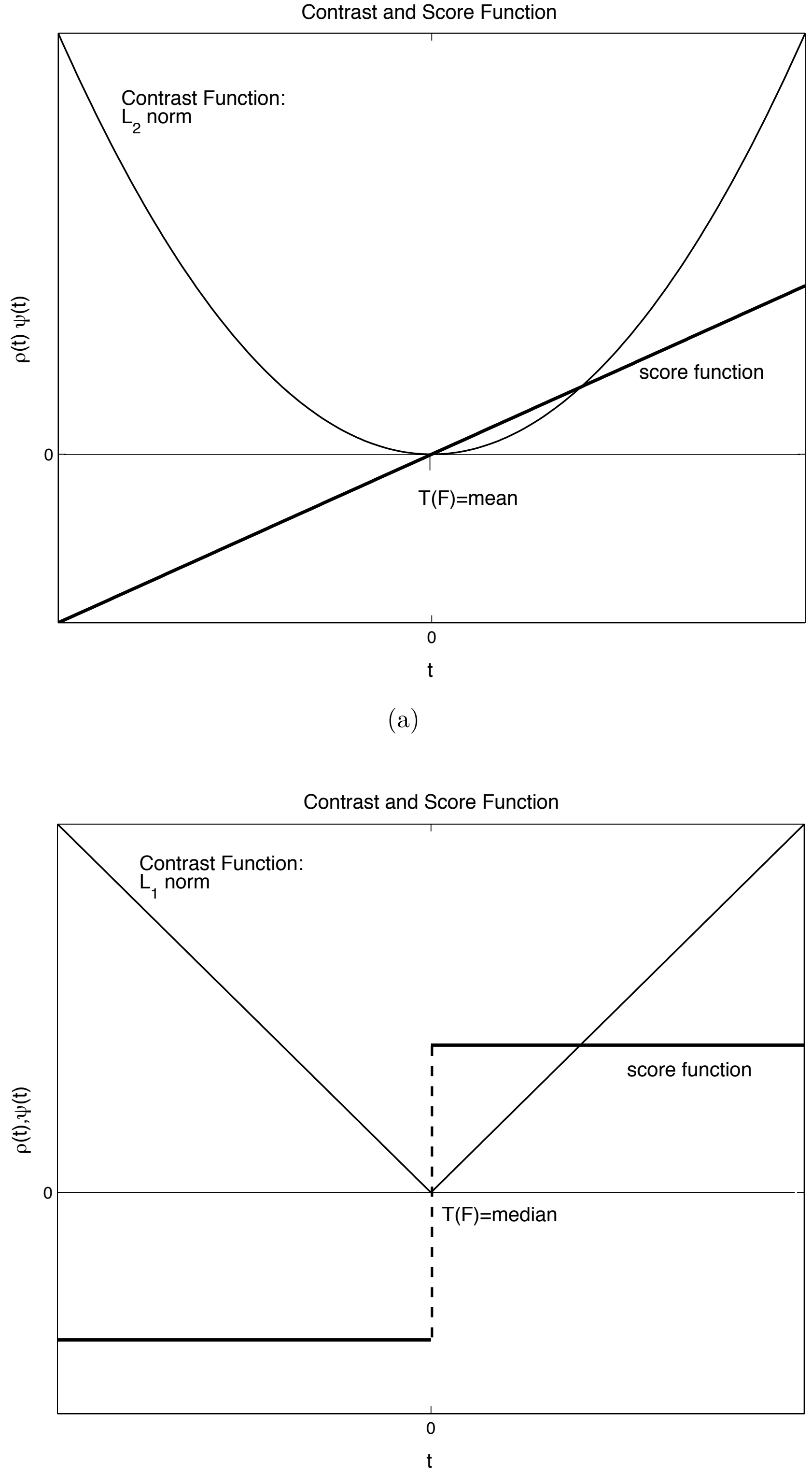

distribution of Cross−validation scores Estimates
MSE
0
0.5
1
1.5
2
2.5
3
3.5
4
γ
0
1
2
3
4
h
0
1
2
3
4
Single CV−Estimates
repeated−V−fold
cv estimate

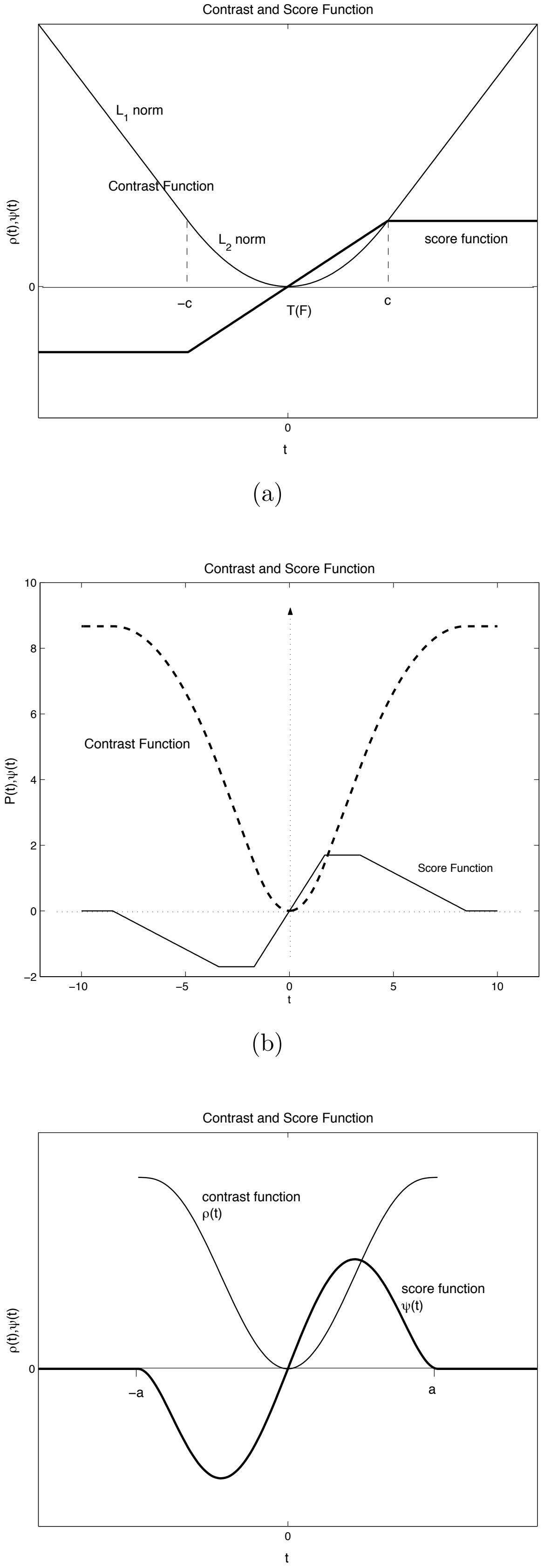
Figure 2:


Contrast and Score Function
Contrast Function:
L_2 norm
score function
T(F)=mean
$\rho(t)\ \psi(t)$
0
0
t
Contrast and Score Function
Contrast Function:
L_1 norm
score function
T(F)=median
$\rho(t),\psi(t)$
0
0
t

(a)

(b)

Figure 3:

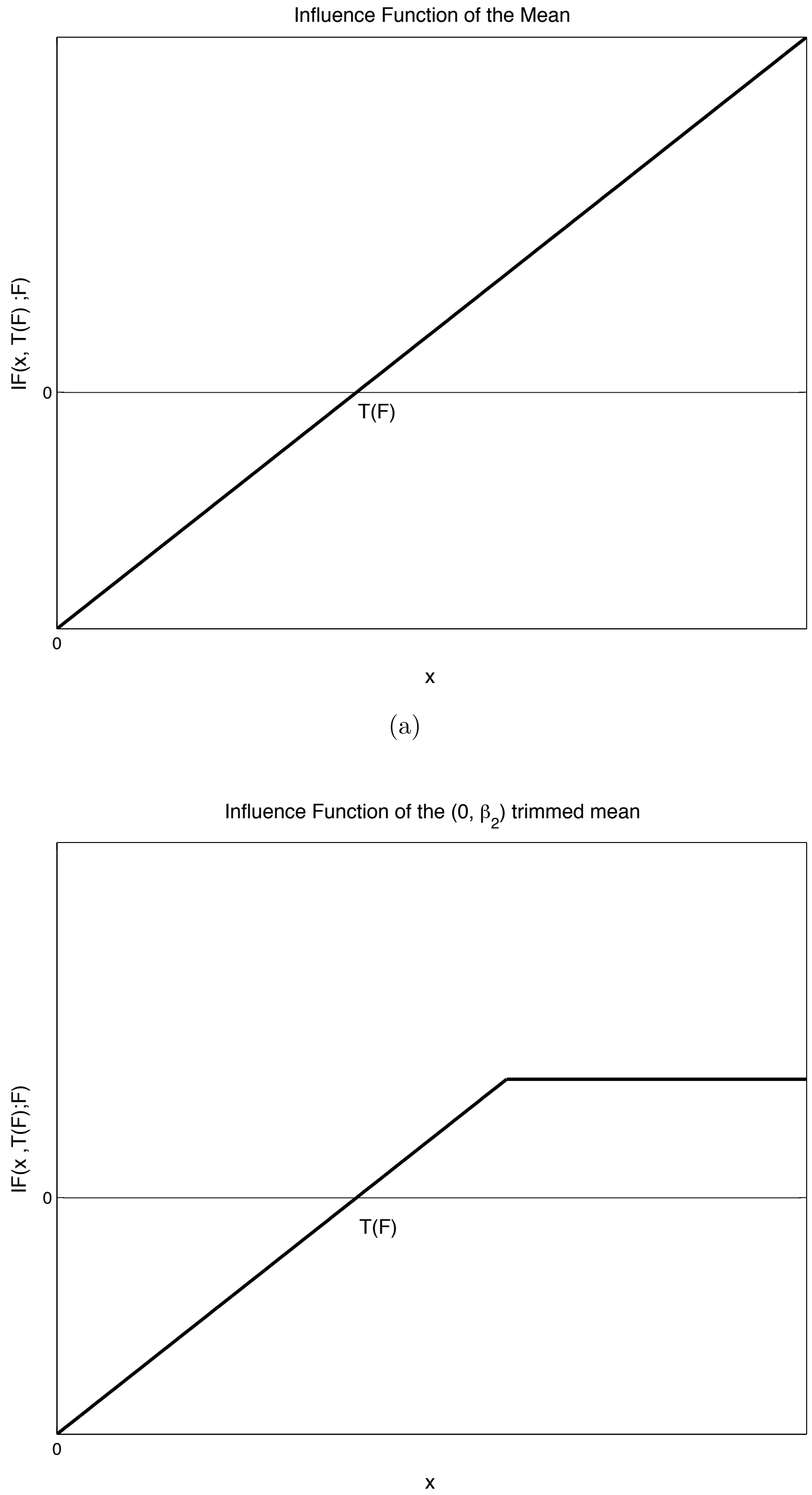

Contrast and Score Function
L1 norm
Contrast Function
L2 norm
score function
ρ(t),ψ(t)
0
−c
c
T(F)
0
t
Contrast and Score Function
10
8
6
4
2
0
−2
Contrast Function
Score Function
P(t),ψ(t)
−10
−5
0
5
10
t
Contrast and Score Function
contrast function
ρ(t)
score function
ψ(t)
ρ(t),ψ(t)
0
−a
a
0
t

(a)

(b)

(c)

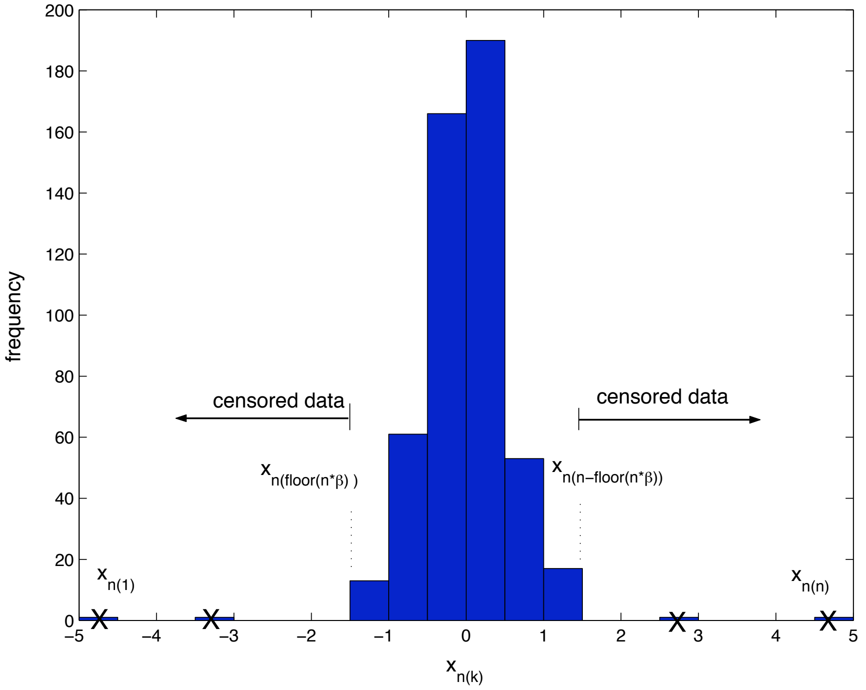
Figure 4:


Influence Function of the Mean
IF(x, T(F) ;F)
0
T(F)
0
x
Influence Function of the (0, β_2) trimmed mean
IF(x ,T(F);F)
0
T(F)
0
x

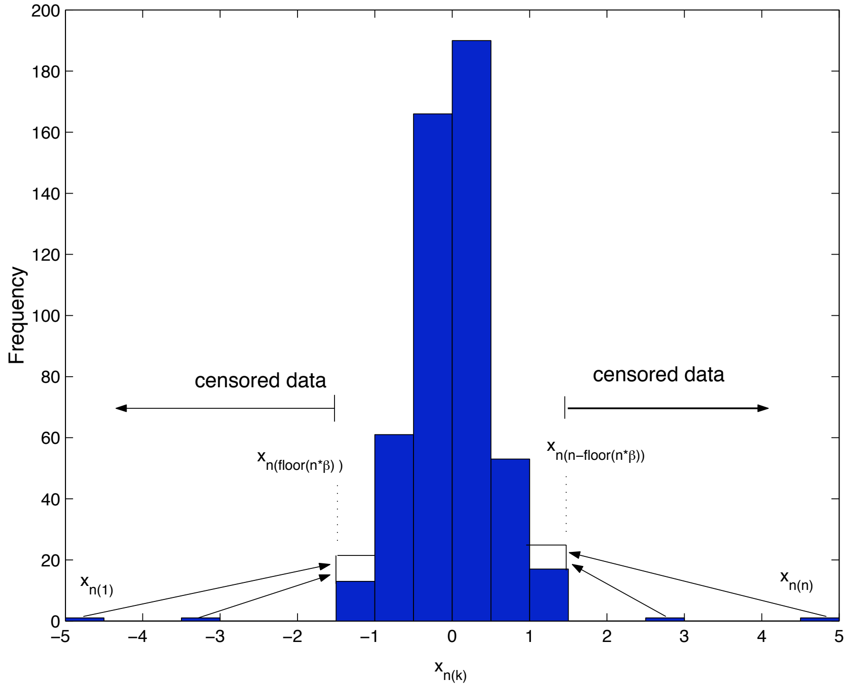
(a)

(b)

Figure 5:

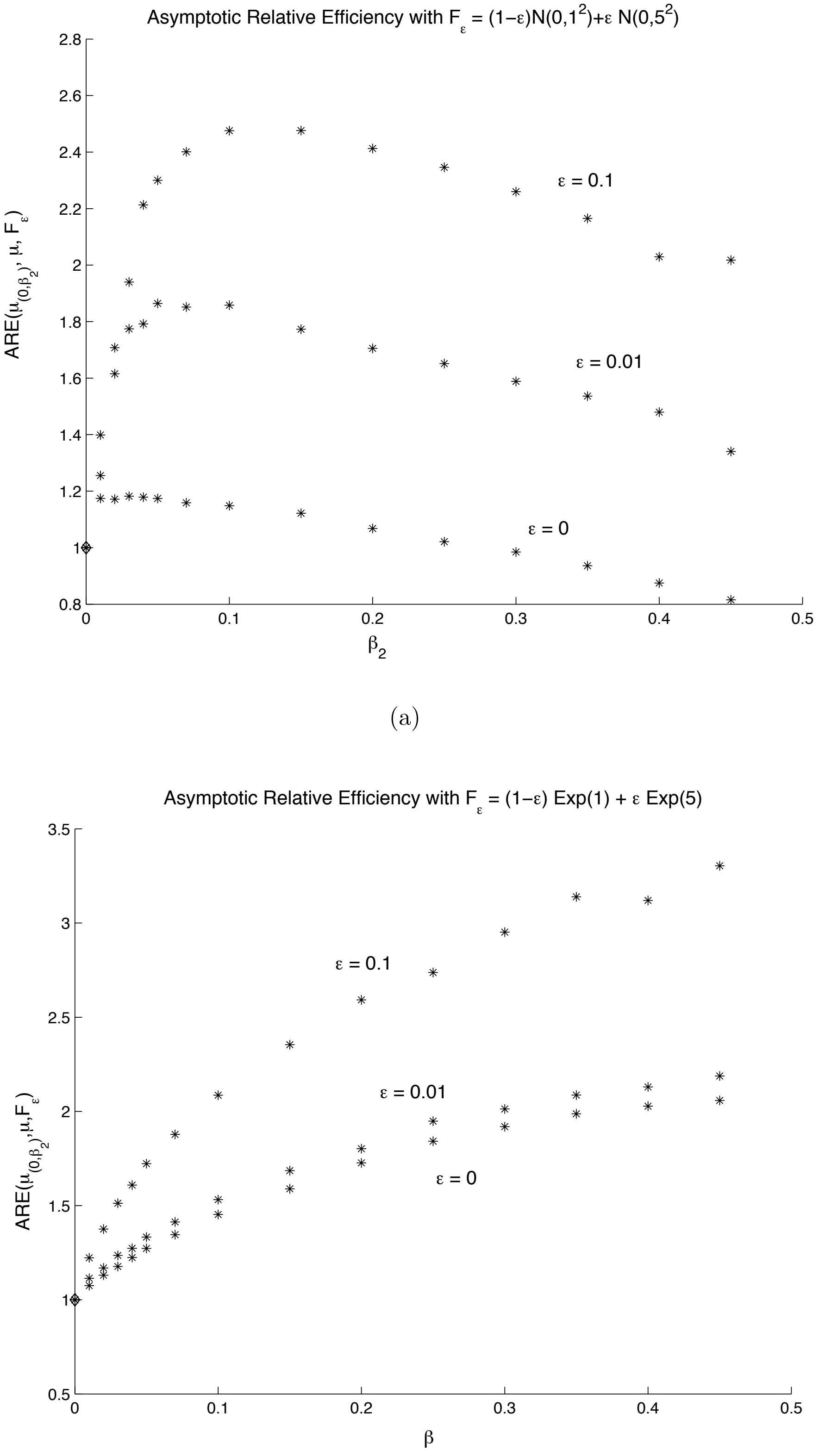

frequency
200
180
160
140
120
100
80
60
40
20
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
$x_{n(k)}$
censored data
censored data
$x_{n(floor(n*\beta))}$
$x_{n(n-floor(n*\beta))}$
$x_{n(1)}$
$x_{n(n)}$

(a)


Frequency
200
180
160
140
120
100
80
60
40
20
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
$x_{n(k)}$
censored data
censored data
$x_{n(floor(n*\beta))}$
$x_{n(n-floor(n*\beta))}$
$x_{n(1)}$
$x_{n(n)}$

(b)

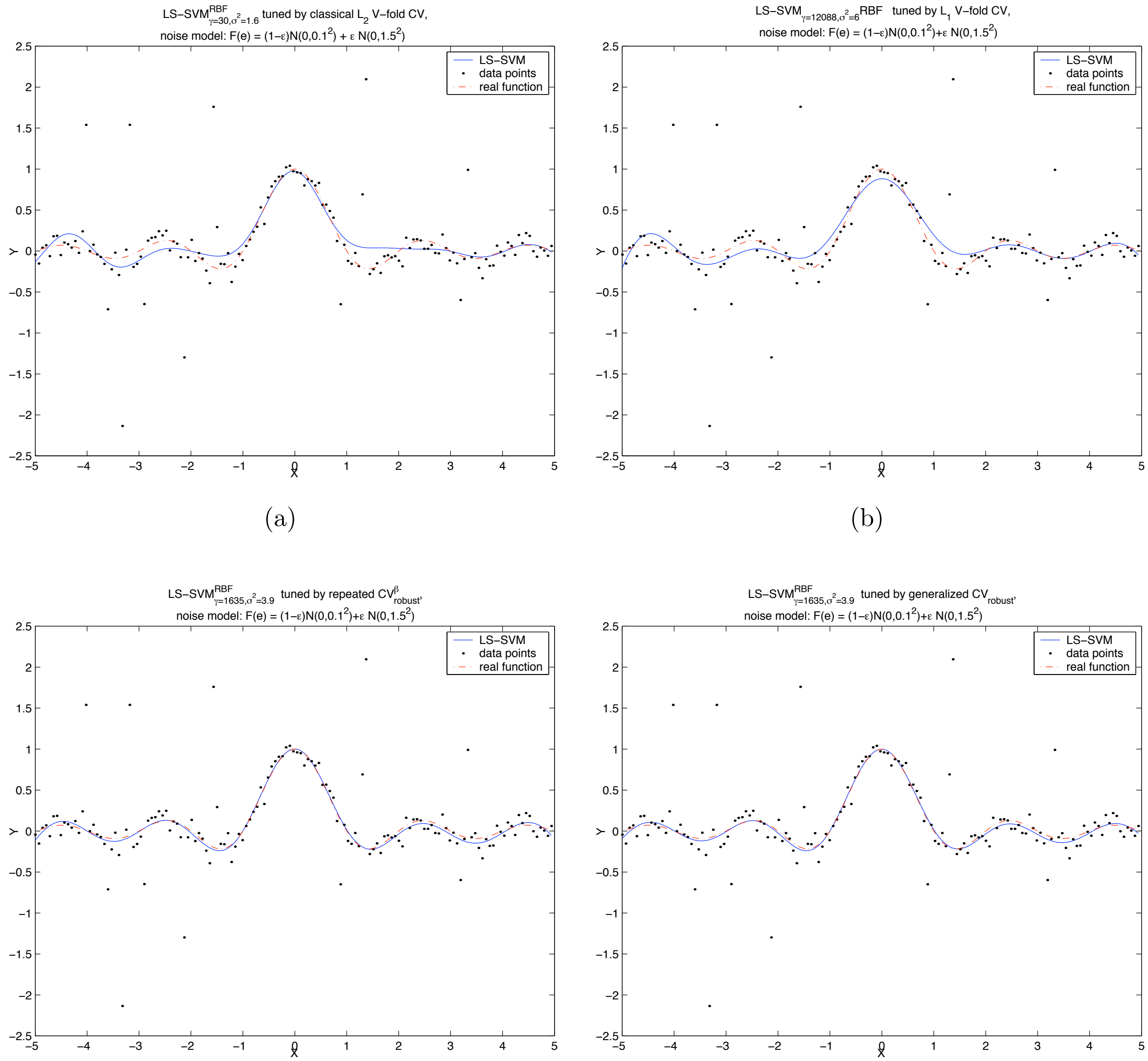
Figure 6:


Asymptotic Relative Efficiency with $F_{\varepsilon} = (1-\varepsilon)N(0,1^2)+\varepsilon\ N(0,5^2)$
$ARE(\mu_{(0,\beta_2)}, \mu, F_{\varepsilon})$
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0
0.1
0.2
0.3
0.4
0.5
β_2
$\varepsilon = 0.1$
$\varepsilon = 0.01$
$\varepsilon = 0$
Asymptotic Relative Efficiency with $F_{\varepsilon} = (1-\varepsilon)$ Exp(1) + ε Exp(5)
$ARE(\mu_{(0,\beta_2)},\mu,F_{\varepsilon})$
3.5
3
2.5
2
1.5
1
0.5
0
0.1
0.2
0.3
0.4
0.5
β
$\varepsilon = 0.1$
$\varepsilon = 0.01$
$\varepsilon = 0$

(a)

(b)

Figure 7:



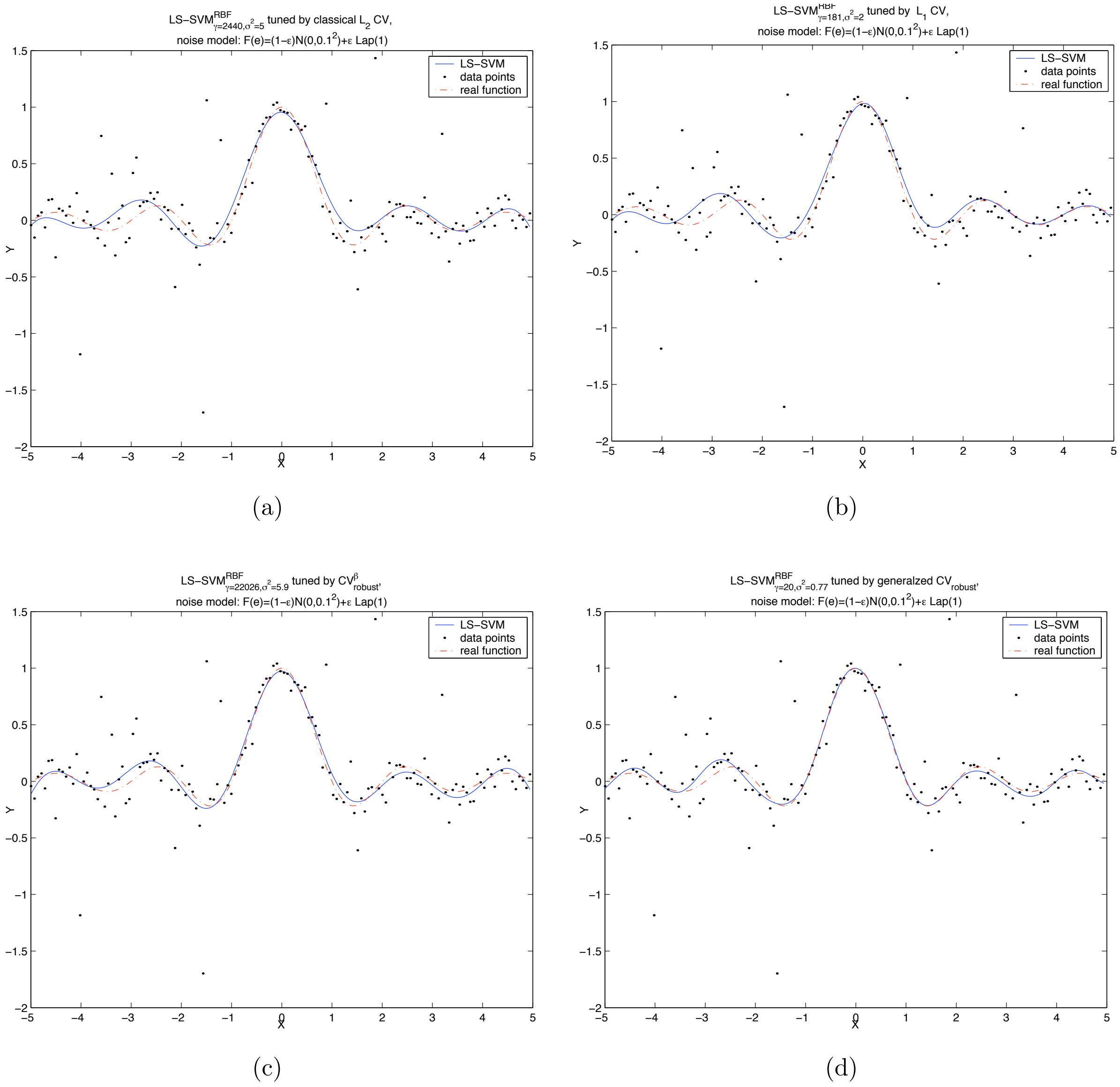
(a)

(b)

(c)

(d)

Figure 8:



LS-SVM$^{RBF}_{\gamma=2440,\sigma^2=5}$ tuned by classical L_2 CV,
noise model: $F(e)=(1-\epsilon)N(0,0.1^2)+\epsilon$ Lap(1)

(a)

LS-SVM$^{RBF}_{\gamma=181,\sigma^2=2}$ tuned by L_1 CV,
noise model: $F(e)=(1-\epsilon)N(0,0.1^2)+\epsilon$ Lap(1)

(b)

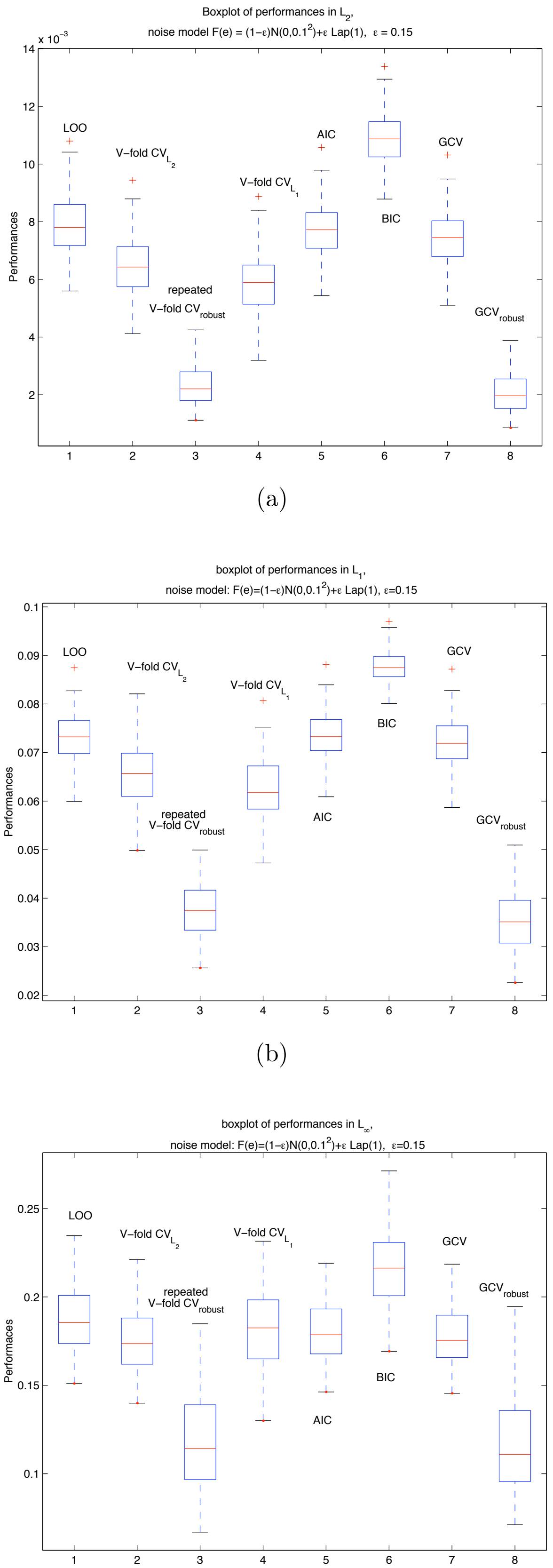
LS-SVM$^{RBF}_{\gamma=22026,\sigma^2=5.9}$ tuned by CV$^{\beta}_{robust'}$
noise model: $F(e)=(1-\epsilon)N(0,0.1^2)+\epsilon$ Lap(1)

(c)

LS-SVM$^{RBF}_{\gamma=20,\sigma^2=0.77}$ tuned by generalzed CV$_{robust'}$
noise model: $F(e)=(1-\epsilon)N(0,0.1^2)+\epsilon$ Lap(1)

(d)

Figure 9:



Boxplot of performances in L_2, noise model $F(e) = (1-\varepsilon)N(0,0.1^2)+\varepsilon$ Lap(1), $\varepsilon = 0.15$

(a)

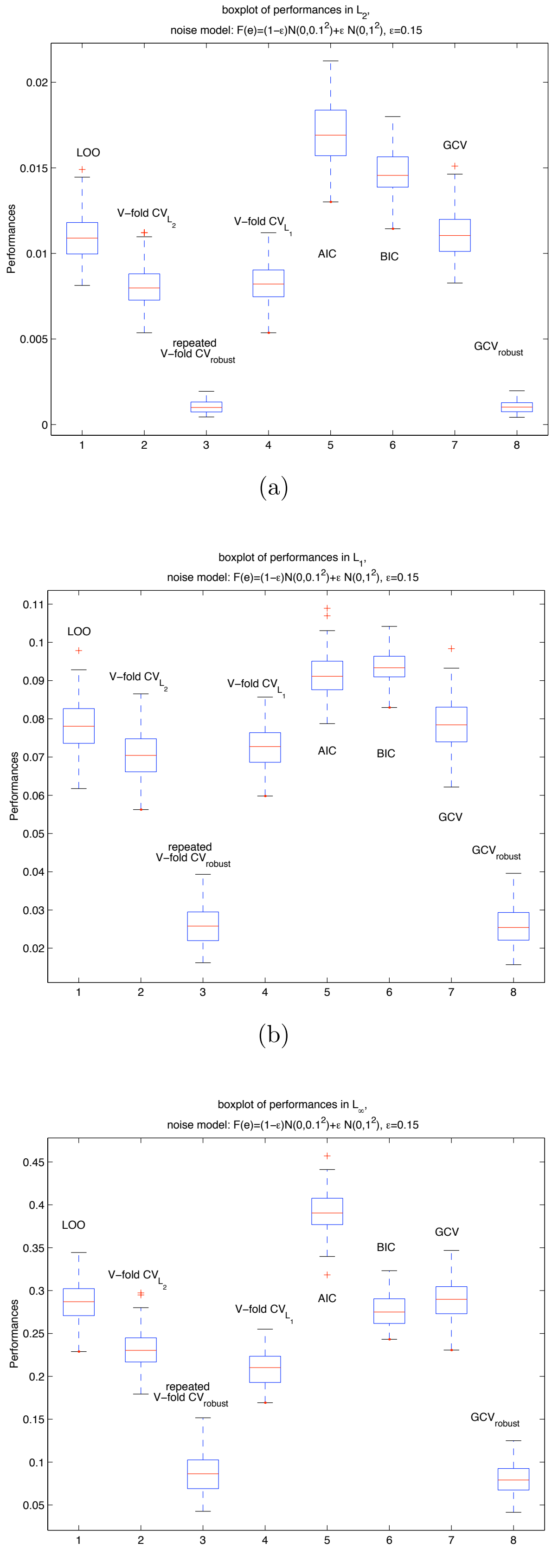
boxplot of performances in L_1, noise model: $F(e)=(1-\varepsilon)N(0,0.1^2)+\varepsilon$ Lap(1), ε=0.15

(b)

boxplot of performances in L_∞, noise model: $F(e)=(1-\varepsilon)N(0,0.1^2)+\varepsilon$ Lap(1), ε=0.15

(c)

Figure 10:



(a)

(b)

(c)

Figure 11: