Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 24

5-15-2020

Support Vector Machine-based Modified Sp Statistic for Subset Selection with Non-Normal Error Terms

Shivaji Shripati Desai
Department of Statistics, Gopal Krishna Gokhale College, Kolhapur (MS), India., ssd.stats@gmail.com

D N. Kashid
Shivaji University, Kolhapur, Maharashtra, India., dnk_stats@unishivaji.ac.in

Support Vector Machine-based Modified Sp Statistic for Subset Selection with Non-Normal Error Terms

Cover Page Footnote

Authors are thankful to Prof. Shlomo S. Sawilowsky, The Editor and the reviewers for positive comments and suggestions which have improved the standard of this article.

Journal of Modern Applied Statistical Methods
May, 2019, Vol. 18, No. 1, eP2659
doi: 10.22237/jmasm/1571545600

בס"ד

ISSN 1538 – 9472

Support Vector Machine-based Modified Sp Statistic for Subset Selection with Non-Normal Error Terms

Shivaji Shripati Desai
Gopal Krishna Gokhale College,
Kolhapur, India

D N. Kashid
Shivaji University,
Kolhapur, India

Support vector machine (SVM) is used for estimation of regression parameters to modify the sum of cross products (Sp). It works well for some nonnormal error distributions. The performance of existing robust methods and the modified Sp is evaluated through simulated and real data. The results show the performance of the modified Sp is good.

Keywords: support vector machine (SVM), support vector regression (SVR), robust subset selection, Sp- statistic, SSp- statistic

Introduction

Regression models are used in almost all fields for establishing the relationship between a variable and a set of variables. Such models are useful tools to predict future values of the response variable given the values of predictors. The multiple linear regression model is given by

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{Y}$ is $n \times 1$ vector of observations on response variable, $\mathbf{X}$ is a known $n \times k$ matrix of observations on $k \times 1$ predictors with 1's in the first column, $\hat{a} = (\beta_0, \beta_1, \beta_2, ..., \beta_{k-1})'$ is $k \times 1$ vector of unknown regression parameters and $\boldsymbol{e}$ is $n \times 1$ vector of unknown errors. The assumptions on the model in (1) are $E(\boldsymbol{e}) = 0$, $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2 I$ $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2 I$ and $\boldsymbol{e} \sim N_n(0, \sigma^2 I)$ where, I is the identity matrix of order $n \times n$ and σ^2 is error variance.

doi: 10.22237/jmasm/1571545600 Accepted May 27, 2017; Published May 15, 2020.
Correspondence: Shivaji Shripati Desai, ssd.stats@gmail.com

Important steps in regression are to obtain the estimates of regression parameters, choose an appropriate model for the given data and to predict the future values of response variable as accurate as possible. Least squares (LS) method is generally used for estimation of parameters in linear regression. The performance of least squares method is excellent if underlying assumptions are true, while it deteriorates when the data contains an outlier and (or) the distribution of error variable is non-normal. Huber's (1981) M-estimator is usually used when error distribution is nonnormal but it is close to normal, and Jaeckel's (1972) rank-based estimator performs well for almost any possible distribution of error (Birkes & Dodge, 1993, pp. 111).

In the set of possible predictors to be included in the model, some of them may be redundant and are required to be eliminated based on the observed data, which will give sufficient predictive accuracy. This is popularly known as subset selection in regression. Miller (2002) indicated fitting a model with a large number of predictors is neither economical nor practicable and in practice usually a model based on a small subset of predictors gives more accurate predictions.

There is considerable literature on subset selection methods in regression (Hocking, 1976; Thompson, 1978a, b; Rao & Wu, 1989). Standard texts such as Miller (2002), Draper and Smith (2003) and Montgomery et al. (2006) provided a good description of subset selection methodologies.

The majority of subset selection methods, including the Cross product (Cp) criterion (Mallow, 1973), are based on the LS estimator of $\boldsymbol{\beta}$. In view of the performance of LS estimator in the presence of outlier and nonnormal error distribution, the subset selection methods based on such estimates will select the wrong subset, as demonstrated in Ronchetti and Staudte (1994).

Many methods were proposed for the choice of a subset by minimizing a criterion, including those by Akaike (1973), Schwartz (1978), Shibata (1984), Rao and Wu (1989), and Kundu and Murali (1996). Some robust subset selection procedures were proposed, including the Robust AIC (Ronchetti, 1985) and Robust versions of Mallow's (1973) Cp, called RCp (Ronchetti & Staudte, 1994). Kashid and Kulkarni (2002) suggested a more general Sum of Cross products (Sp) criterion for subset selection. It uses scaled difference between robust predicted values from subset and full model to perform subset selection in linear regression. Their results showed performance of Sp is better than Cp in the presence of outliers. Baierl et al. (2007) proposed a robust version of BIC based on Huber's M-estimator and called it as Robust BIC.

These M-estimators works well under certain assumptions. They are not guaranteed to produce a good subset selection if the data do not support the underlying assumptions. An alternative is to base a subset selection procedure on a data dependent prediction method such as the Support Vector Machine (SVM), the focus of

this study. It is a growing area in machine learning introduced by Boser et al. (1992) in COLT. The basic task of SVM is to explore data (input-output pairs) and provide optimally accurate predictions for unseen data (Nalbantov, 2003). Vapnik et al. (1997) used SVM for regression and called it as Support Vector Regression (SVR). The SVR problem is formulated as a convex optimization problem which has the advantage of being free from local minima. SVMs based on the ε insensitive loss for regression are consistent and robust even for heavy-tailed distributions (Christmann et al., 2008, 2009; Messem & Christmann, 2010). SVR is used here for parameter estimation and to obtain predicted values from the full model and subset models.

Support Vector Regression

In SVR, an unknown regression function $f(x_i)$ based on data set (x_i, y_i), i = 1,2,..,n of input vectors $\boldsymbol{x}_i \in R^{k-1}$ (ith row of design matrix $\mathbf{X}$ excluding first element 1) and associated target $y_i \in R$, is estimated in the form,

$$y_i = f(\boldsymbol{x}_i) + e_i \tag{2}$$

where e_i is error term.

For linear regression, the function can be written as,

$$f(\boldsymbol{x}_i) = b + \boldsymbol{x}_i \boldsymbol{w}, \tag{3}$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_{k-1})' \in R^{k-1}, b \in R$ is a bias and $\boldsymbol{x}_i \boldsymbol{w}$ is a dot product of $\boldsymbol{x}_i$ and $\boldsymbol{w}$.

Thus, Equation (2) becomes,

$$y_i = b + \boldsymbol{x}_i \boldsymbol{w} + e_i, \quad i = 1,2,\ldots\ldots,n.$$

In matrix notations,

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\hat{\boldsymbol{a}} = (b, w_1, w_2, \ldots.., w_{k-1})'$, $\boldsymbol{Y}$, $\mathbf{X}$ and $\boldsymbol{e}$ are the same as defined in Equation (1). The above equation is equivalent to Equation (1).

Using the ε insensitive loss function (Vapnik, 2001), the regression problem can be written in the form of convex optimization problem (Smola & Schölkopf, 2004) as

$$\text{Minimize} \frac{1}{2} \left\| \boldsymbol{w}^2 \right\| \tag{4}$$

$$\text{Subject to } y_i - (\boldsymbol{x}_i \boldsymbol{w} + b) \leq \varepsilon \ , i = 1,2,\ldots,n, \tag{5}$$

$$(\boldsymbol{x}_i \boldsymbol{w} + b) - y_i \leq \varepsilon \ , i = 1,2,\ldots,n \ , \tag{6}$$

where $\varepsilon > 0$ is predetermined constant which controls the noise tolerance.

By introducing non negative slack variables ξ_i and ξ_i^* (which measures the deviations of training samples outside ε insensitive zone), the above optimization problem becomes (Vapnik, 2001):

$$\text{Min.} \frac{1}{2} \|\boldsymbol{w}\|^2 + \text{C} \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{7}$$

$$\text{Subject to } y_i - (\boldsymbol{x}_i \boldsymbol{w} + b) \leq \varepsilon + \xi_i \ , \tag{8}$$

$$(\boldsymbol{x}_i \boldsymbol{w} + b) - y_i \leq \varepsilon + \xi_i^* \ , \tag{9}$$

$$\text{and } \xi_i, \xi_i^* \geq 0, i = 1,2,\ldots,n, \tag{10}$$

where $\text{C} > 0$ is the regularization factor which determines (the cost of error) tradeoff between flatness of regression function and amount of deviations outside the ε insensitive zone which are tolerated.

Using Lagrange's multipliers method, the dual of above optimization problem can be expressed as (Gunn, 1998),

$$\text{max} . -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \boldsymbol{x}_i \boldsymbol{x}_j^{'} - \varepsilon \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) + \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) y_i \tag{11}$$

$$\text{Subject to } \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 \ , \tag{12}$$

$$\text{and } 0 \leq \alpha_i \leq \text{C} \ , 0 \leq \alpha_i^* \leq \text{C}, \tag{13}$$

where α_i and α_i^* , $i = 1,2,\ldots,n$ are Lagrange's multipliers that act as forces pushing the predictions towards the target value y_i.

Above quadratic programming problem can be solved for obtaining the values of α_i and α_i^*. Using Karush-Kuhn-Tucker conditions (Smola & Schölkopf, 2004) the weight vector is given by

$$\boldsymbol{w}^{'} = \sum_{i=1}^{n_{nsv}} (\alpha_i - \alpha_i^*) \boldsymbol{x}_i \tag{14}$$

$$\text{and } f(\boldsymbol{x}) = \sum_{i=1}^{n_{nsv}} \left(\alpha_i - \alpha_i^*\right) \boldsymbol{x}_i \boldsymbol{x}' + b, \tag{15}$$

where n_{nsv} denotes the number of support vectors.

The value of bias b is given by (Gunn, 1998),

$$b = -\frac{1}{2}\left(\boldsymbol{x}_r + \boldsymbol{x}_s\right)\boldsymbol{w} \tag{16}$$

where $\boldsymbol{x}_r$, and $\boldsymbol{x}_s$ are the support vectors (i.e. any input vector which has nonzero value of either α_i or α_i^* respectively).

The performance of SVR strongly depends on proper setting of regularization parameter (C) and the value of ε. Such parameters are called as meta parameters. The values of meta parameters C and ε are not known in advance and must be obtained from the training data.

Subset selection using SSp statistic

Consider a regression model in Equation (1) as a full model. Partition the $\mathbf{X}$ matrix and vector $\boldsymbol{\beta}$ as

$$\mathbf{X} = \left[\mathbf{X}_{(1)} : \mathbf{X}_{(2)}\right] \text{and } \boldsymbol{\beta} = \left[\boldsymbol{\beta}_{(1)} : \boldsymbol{\beta}_{(2)}\right]',$$

where $\mathbf{X}_{(1)}$ is an $n \times p$ matrix of observations on p –1 predictor variables with 1's in the first column and $\mathbf{X}_{(2)}$ is an $n \times (k-p)$ matrix of observations on remaining $(k-p)$ predictor variables. $\boldsymbol{\beta}_{(1)} = (\beta_0, \beta_1, \beta_2, .., \beta_{p-1})'$ is p × 1 vector of regression parameters corresponding to $(p-1)$ predictor variables and $\boldsymbol{\beta}_{(2)}$ is $(k-p \times 1$ vector of regression parameters corresponding to remaining $(k-p)$ predictor variables. In these notations the full model becomes

$$\boldsymbol{Y} = \boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{X}_{(2)}\boldsymbol{\beta}_{(2)} + \boldsymbol{e} \tag{17}$$

A subset model based on a $(p-1)$ predictors is given by

$$\boldsymbol{Y} = \boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{e} \tag{18}$$

Similarly, for SVR partition the weight vector $\boldsymbol{w}$ and write the sub model as

$$\boldsymbol{Y} = b_{(1)} + \mathbf{X}_{(1)}\boldsymbol{w}_{(1)} + \boldsymbol{e} \tag{19}$$

where $\mathbf{X}_{(1)}$ is an $n \times (p-1)$ matrix of the observations on $(p-1)$ predictors and $\boldsymbol{w}_{(1)}$ is a $(p-1) \times 1$ vector of the regression coefficients based on the fitted sub model.

For a subset model, the dual of optimization problem can be expressed as

$$\max . -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha_{i(1)}-\alpha_{i(1)}^{*}\right)\left(\alpha_{j(1)}-\alpha_{j(1)}^{*}\right)\boldsymbol{x}_{i(1)}\boldsymbol{x}_{j}^{'}-\varepsilon\sum_{i=1}^{n}\left(\alpha_{i(1)}-\alpha_{i(1)}^{*}\right)+$$

$$i = 1nai(1) - ai(1) * yi \tag{20}$$

$$\text{Subject to}: \sum_{i=1}^{n}\left(\alpha_{i(1)}-\alpha_{i(1)}^{*}\right)=0 \tag{21}$$

and $0 \le \alpha_{i(1)} \le C$, $0 \le \alpha_{i(1)}^{*} \le C$, where $\alpha_{i(1)}$ and $\alpha_{i(1)}^{*}$, $i=1,2,\ldots,n$ are Lagrange's multipliers. The weight vector is given by

$$\boldsymbol{w}_{(1)}^{'}=\sum_{i=1}^{n_{nsv}}\left(\alpha_{i(1)}-\alpha_{i(1)}^{*}\right)\boldsymbol{x}_{i(1)} \tag{22}$$

and

$$f(\boldsymbol{x})=\sum_{i=1}^{n_{nsv}}\left(\alpha_{i(1)}-\alpha_{i(1)}^{*}\right)\boldsymbol{x}_{i(1)}\boldsymbol{x}'+b_{(1)}. \tag{23}$$

The value of bias $b_{(1)}$ is given by (Gunn-1998),

$$b_{(1)}=-\frac{1}{2}\left(\boldsymbol{x}_{r(1)}+\boldsymbol{x}_{s(1)}\right)\boldsymbol{w}_{(1)} \tag{24}$$

By obtaining estimates $\hat{\boldsymbol{w}}$ of $\boldsymbol{w}$ and $\hat{b}$ of b using SVM, the predicted value of y based on the full model is given by

$$\hat{\mathrm{y}}_{ik}=\boldsymbol{x}_{i}\hat{\boldsymbol{w}}+\hat{b}\,,\ i=1,2,\ldots\ldots,n \tag{25}$$

The predicted value of y based on sub model is given by

$$\hat{y}_{ip}=\boldsymbol{x}_{i(1)}\hat{\boldsymbol{w}}_{(1)}+\hat{b}_{(1)},\ i=1,2,\ldots\ldots,n \tag{26}$$

where $\hat{\boldsymbol{w}}_{(1)}$ is the estimator of $\boldsymbol{w}_{(1)}$, $\hat{b}_{(1)}$ is the estimator of $b_{(1)}$ and $\boldsymbol{x}_{i(1)} \in R^{p-1}$ (i^{th} row of $\mathbf{X}_{(1)}$ excluding first element 1)

Kashid and Kulkarni (2002) defined the Sp Statistic for subset selection based on predicted values from full and subset models using M-estimator given by

$$\mathrm{Sp} = \sum_{i=1}^{n} \frac{\left(\hat{y}_{ik} - \hat{y}_{ip}\right)^2}{\sigma^2} - \left(k - 2p\right), \tag{27}$$

where, σ^2 is replaced by its estimate usually based on full model, k and p are number of parameters in full model and subset model, respectively.

The Sp statistic takes into account closeness of predictions obtained from subset model and incorporates the complexity in the form of number of predictors involved in the model. The penalty term doesn't increase with sample size. If the Sp statistic is based on the estimates obtained from SVR, its performance is not good. Hence, a complexity term is added to the criterion so as to increase its ability to identify the correct subset model. The modified Sp statistic is called the SSp statistic and is given by

$$\mathrm{SSp} = \sum_{i=1}^{n} \frac{\left(\hat{y}_{ik} - \hat{y}_{ip}\right)^2}{\sigma^2} - \left(k - p\right) + g\left(n, p\right), \tag{28}$$

where error variance σ^2 is usually unknown and is to be replaced by its suitable estimate. As discussed in Bozdogan and Haughton (1998) the term $g\left(n, p\right)$ is a non-negative penalty function which increases as the number of parameters increases. In this study, to calculate SSp statistic use SVR estimates which are robust, and thus SSp is also robust.

For $g\left(n, p\right) = p$, $\boldsymbol{\beta}$ is replaced by its LS estimate then SSp statistic is equivalent to Mallow's Cp statistic, which is given by

$$\mathrm{Cp} = \sum_{i=1}^{n} \frac{\left(y_i - \hat{y}_{ip}\right)^2}{\sigma^2} - \left(n - 2p\right). \tag{29}$$

If $g\left(n, p\right) = p$ and the M estimator of $\boldsymbol{\beta}$ is used in (28), then SSp statistic is equivalent to the Sp statistic.

Simulation results

The estimation accuracy of SVR strongly depends on selection of meta parameters C and ε. When the values of meta parameters C and ε are not known they have to be obtained from the data itself. There are several methods available in the literature for selection of C amongst which, the method proposed by Desai and Kashid (2015) uses C = MQD for better performance. In the simulations to perform SVR, this method was used. Take $\varepsilon = \mathrm{C} \times 10^{-6}$ and $\mathrm{C} = \mathrm{MQD} = \mathrm{Max.}\left(\left|\mathrm{Me} - 3\mathrm{QD}\right|, \left|\mathrm{Me} + 3\mathrm{QD}\right|\right)$, where Me and QD are median and quartile deviation of y values respectively.

A simulation design

The following simulation study is carried out for the performance evaluation of the subset selection methods. The observations on predictor variables are generated from U(0, 1) and are fixed. Observations on error term are generated from N(0,1), students t distribution with 2 d.f., Laplace(0,1), Mixture of Normal {0.2N(5,1) + 0.8N(0,1)}, standard Cauchy and Slash distributions{ratio of N(0,1) and U(0,1)}.

To obtain the observations on response variable following models are used.

Model-I : $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + e$

Model-II : $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + e$

For model-I, we consider three sets of regression parameters (2, 5, 0, 0), (2, 5, 4, 0) and (2, 5, 4, 8), and call them as Model-IA, Model-IB and Model-IC respectively. For model-II, take four sets of regression parameters (2, 0, 6, 0, 0), (2, 9, 6, 0, 0), (2, 9, 0, 4, 8) and (2, 9, 6, 4, 8), and call them as Model-IIA, Model-IIB, Model-IIC and Model-IID, respectively.

For each sample size 20, 30, 50 and 100 the experiment is repeated 1000 times. To evaluate the performance of different methods, the probabilities of selecting the optimal and correct models are calculated for various combinations of error distributions and sample sizes. SSp statistic with six different penalty functions is used. The SSp statistic using the penalty functions $g_1(n,p) = p$, $g_2(n,p) = 2p$, $g_3(n,p) = p\sqrt{n}$, $g_4(n,p) = p\left(\sqrt{n}+1\right)$, $g_5(n,p) = plog(n)$ and $g_6(n,p) = p\left(log(n)+1\right)$ is called as SSp1, SSp2, SSp3, SSp4, SSp5 and SSp6 respectively.

In regression true value of variance is often known. If it is not known, it has to be estimated from the data itself. The estimation of variance plays an important role in regression. From the available estimators of σ^2, use $\hat{\sigma}_1^2 = (1.4826\text{MAD})^2$ to calculate Sp, where $MAD = Median\left|r_i - Median(r_i)\right|$ and $r_i = y_i - \hat{y}_i$ (Birkes & Dodge, 1993). The performance of $\hat{\sigma}_1^2$ estimator works well when the M-estimator is used for obtaining regression coefficients (Kashid & Kulkarni, 2002). If SVR is used for the prediction obtain $\hat{\sigma}_1^2$, then $\hat{\sigma}_1^2$ will under-estimates the true value of σ^2. This is verified in the simulation study. To improve the performance, define $\hat{\sigma}_2^2$ using SVR estimates as,

$$\hat{\sigma}_2^2 = \left[1.4826 \times \left(0.8 + \frac{k}{n}\right) \times \mathrm{P}_{60}\left\{abs(r_i)\right\}\right]^2 \tag{30}$$

where P_{60} is 60th percentile of absolute residuals, k and p are the number of parameters of the full and subset model respectively. The performance of $\hat{\sigma}_2^2$ using SVR estimates is verified in thesimulation study (not shown here). In the simulation study for calculation of SSp statistics, use $\hat{\sigma}_2^2$. For clarity of notations, call Sp1 for Sp statistic when estimator of variance is $\hat{\sigma}_1^2$ and Sp2 for Sp statistic when estimator of variance $\hat{\sigma}_2^2$ is used.

The M-estimator is calculated using Huber loss function (Huber,1981) with bending point 1.345. Also, for comparison purposes, consider the Robust Akaike's Information Criterion (RAIC) (Ronchetti, 1985) and Robust Bayes Information Criterion (RBIC) (Baierl et al., 2007). The RAIC is defined as

$$\text{RAIC} = 2\sum_{i=1}^{n} \rho(r_{i;p}) + 2p \tag{31}$$

where $\rho(\cdot)$ is Huber function and $r_{i;p} = \left(y_i - \boldsymbol{x}^*_{i(1)}\tilde{\boldsymbol{\beta}}_{(1)}\right)/\hat{\sigma}$, $\hat{\sigma}$ is some robust estimate of σ, $\tilde{\boldsymbol{\beta}}_{(1)}$ is the M-estimator of $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{x}^*_{i(1)} \in R^p$ (i^{th} row of $\mathbf{X}_{(1)}$). RBIC is defined as

$$\text{RBIC} = n\log\sum_{i=1}^{n} \rho\,(y_i^{(s)} - \boldsymbol{x}^*_{i(1)}\tilde{\boldsymbol{\beta}}_{(1)}) + p\log\left(n\right) \tag{32}$$

where, $y_i^{(s)}$ is standardized observation obtained by subtracting the median and dividing by 1.486MAD of y values, $\tilde{\boldsymbol{\beta}}_{(1)}$ is the M-estimator of $\boldsymbol{\beta}_{(1)}$. Due to space constraint some of the simulation results are summarized in the Tables 1 to 3 and Figures 1 to 4.

From Tables 1–3 and Figures 1–4 we observed the following:

For model IA, predictors X_2 and X_3 are redundant, performance of SSp3 and SSp4 is better than RAIC, Sp1 and Sp2 and it is compatible with RBIC. For model IB, predictor X_3 is redundant, performance of SSp3 and SSp4 is better than RAIC, RBIC, Sp1 and Sp2 (except for Slash distribution). For sample size $n = 100$ and Cauchy and Slash errors, the performance of SSp3, SSp4, SSp5 and SSp6 is better than RAIC, RBIC, Sp1 and Sp2. For Model IC which is full as well as optimal model, performance of SSp1 is compatible with Sp1, Sp2, RAIC and RBIC for large sample size.

For Model IIA, performance of SSp3, SSp4, SSp5 and SSp6 is better than SSp1, SSp2, RAIC, Sp1 and Sp2. Also it is compatible with RBIC. The performance of SSp2 is better than SSp1, SSp4 is better than SSp3, and SSp6 is better than SSp5 (except for large sample). For Models IIB and IIC, the performance of SSp3, SSp4, SSp5 and SSp6 is better than SSp1, SSp2, RAIC, Sp1 and Sp2, and is compatible with RBIC. Model IID is full model and optimal model. For this model, performance of SSp1 is compatible with Sp1, Sp2, RAIC and RBIC for large sample size. For

Table 1. Probabilities of selecting Optimal and Over fitted (Correct) models for Model I B

			SSp with New S.D. and penalty									
Error	S.S.	Model	SSp1	SSp2	SSp3	SSp4	SSp5	SSp6	Sp1	Sp2	RAIC	RBIC
N(0, 1)	20	Optimal	0.633	0.706	0.766	0.759	0.739	0.766	0.757	0.820	0.745	0.873
		Correct	0.349	0.261	0.134	0.104	0.200	0.150	0.239	0.172	0.253	0.125
	30	Optimal	0.647	0.773	0.913	**0.932**	0.857	0.889	0.775	0.824	0.797	0.916
		Correct	0.353	0.226	0.079	0.056	0.141	0.108	0.225	0.176	0.203	0.084
	50	Optimal	0.662	0.772	0.947	**0.960**	0.882	0.905	0.802	0.827	0.806	0.948
		Correct	0.338	0.228	0.053	0.040	0.118	0.095	0.198	0.173	0.194	0.052
	100	Optimal	0.652	0.765	**0.976**	**0.988**	0.901	0.927	0.817	0.832	0.822	0.972
		Correct	0.348	0.235	0.024	0.012	0.099	0.073	0.183	0.168	0.178	0.028
t-2	20	Optimal	0.589	0.633	0.579	0.560	0.628	0.595	0.653	0.714	0.682	0.702
		Correct	0.306	0.208	0.080	0.058	0.137	0.095	0.279	0.184	0.261	0.064
	30	Optimal	0.670	0.753	0.835	0.825	0.813	0.833	0.722	0.784	0.721	0.810
		Correct	0.312	0.213	0.060	0.039	0.131	0.082	0.269	0.204	0.259	0.031
	50	Optimal	0.690	0.786	**0.944**	**0.949**	0.891	0.924	0.742	0.794	0.753	0.959
		Correct	0.308	0.211	0.042	0.032	0.105	0.068	0.258	0.206	0.246	0.017
	100	Optimal	0.712	0.814	**0.985**	**0.990**	0.932	0.950	0.752	0.792	0.748	0.991
		Correct	0.288	0.186	0.015	0.010	0.068	0.050	0.248	0.208	0.252	0.009
			SSp with New S.D. and penalty									
Error	S.S.	Model	SSp1	SSp2	SSp3	SSp4	SSp5	SSp6	Sp1	Sp2	RAIC	RBIC
Laplace	20	Optimal	0.621	0.682	0.702	0.692	0.712	0.710	0.669	0.740	0.704	0.878
		Correct	0.340	0.243	0.112	0.081	0.172	0.131	0.305	0.224	0.285	0.085
	30	Optimal	0.680	0.780	0.894	0.888	0.857	0.885	0.705	0.783	0.748	0.924
		Correct	0.316	0.209	0.065	0.055	0.126	0.090	0.295	0.216	0.251	0.057
	50	Optimal	0.698	0.808	**0.975**	**0.978**	0.920	0.950	0.763	0.815	0.768	0.971
		Correct	0.301	0.191	0.024	0.020	0.079	0.049	0.237	0.185	0.232	0.029
	100	Optimal	0.752	0.857	**0.992**	**0.994**	0.945	0.961	0.778	0.804	0.775	0.976
		Correct	0.248	0.143	0.008	0.006	0.055	0.039	0.222	0.196	0.225	0.024
Mixture	20	Optimal	0.618	0.703	0.780	0.773	0.740	0.778	0.725	0.807	0.769	0.879
		Correct	0.364	0.260	0.119	0.087	0.194	0.134	0.271	0.187	0.231	0.119
	30	Optimal	0.649	0.760	0.917	0.936	0.854	0.887	0.773	0.817	0.787	0.905
		Correct	0.351	0.240	0.080	0.059	0.144	0.111	0.227	0.183	0.213	0.094
	50	Optimal	0.640	0.769	0.948	0.958	0.880	0.906	0.796	0.837	0.812	0.941
		Correct	0.360	0.231	0.052	0.042	0.120	0.094	0.204	0.163	0.188	0.059
	100	Optimal	0.670	0.778	**0.978**	**0.980**	0.902	0.935	0.829	0.848	0.809	0.954
		Correct	0.330	0.222	0.022	0.020	0.098	0.065	0.171	0.152	0.191	0.046
Cauchy	20	Optimal	0.545	0.529	0.429	0.385	0.492	0.452	0.539	0.565	0.517	0.240
		Correct	0.190	0.123	0.042	0.027	0.081	0.050	0.230	0.147	0.245	0.022
	30	Optimal	0.664	0.722	0.617	0.573	0.712	0.671	0.619	0.678	0.609	0.254
		Correct	0.255	0.138	0.028	0.020	0.065	0.044	0.275	0.193	0.242	0.005
	50	Optimal	**0.729**	**0.802**	**0.862**	**0.840**	**0.881**	**0.887**	0.722	0.780	0.685	0.380
		Correct	0.260	0.171	0.020	0.013	0.072	0.046	0.260	0.192	0.287	0.004
	100	Optimal	**0.762**	**0.857**	**0.988**	**0.986**	**0.956**	**0.968**	0.713	0.774	0.730	0.581
		Correct	0.238	0.142	0.005	0.004	0.041	0.029	0.287	0.226	0.270	0.001
Slash	20	Optimal	0.374	0.311	0.192	0.159	0.252	0.210	0.399	0.395	0.398	0.114
		Correct	0.150	0.085	0.025	0.016	0.054	0.031	0.204	0.106	0.147	0.008
	30	Optimal	0.509	0.496	0.311	0.265	0.436	0.373	0.540	0.544	0.487	0.123
		Correct	0.203	0.106	0.014	0.009	0.042	0.024	0.199	0.140	0.192	0.000
	50	Optimal	0.669	0.704	0.510	0.461	0.659	0.615	0.692	0.718	0.643	0.162
		Correct	0.243	0.150	0.008	0.002	0.053	0.032	0.222	0.170	0.240	0.000
	100	Optimal	**0.723**	**0.809**	**0.841**	**0.818**	**0.902**	**0.909**	0.732	0.763	0.719	0.286
		Correct	0.275	0.177	0.003	0.002	0.058	0.037	0.263	0.231	0.271	0.001

Table 2. Probabilities of Selecting Optimal and Over fitted (Correct) models for Model II C

			SSp with New S.D. and penalty									
Error	S.S.	Model	SSp1	SSp2	SSp3	SSp4	SSp5	SSp6	Sp1	Sp2	RAIC	RBIC
N(0, 1)	20	Optimal	0.549	0.647	0.741	0.744	0.698	0.729	0.694	0.779	0.735	0.867
		Correct	0.436	0.322	0.170	0.137	0.247	0.190	0.303	0.214	0.262	0.125
	30	Optimal	0.597	0.688	0.872	0.895	0.788	0.832	0.752	0.812	0.774	0.908
		Correct	0.403	0.312	0.124	0.098	0.211	0.165	0.248	0.188	0.226	0.092
	50	Optimal	0.609	0.722	0.935	0.948	0.850	0.891	0.792	0.836	0.822	0.959
		Correct	0.391	0.278	0.064	0.050	0.149	0.108	0.208	0.164	0.178	0.041
	100	Optimal	0.612	0.734	**0.982**	**0.986**	0.882	0.916	0.820	0.848	0.852	0.977
		Correct	0.388	0.266	0.018	0.014	0.118	0.084	0.180	0.152	0.148	0.023
t-2	20	Optimal	0.581	0.638	0.607	0.576	0.640	0.622	0.650	0.725	0.666	0.692
		Correct	0.334	0.215	0.089	0.064	0.152	0.106	0.286	0.174	0.267	0.079
	30	Optimal	0.635	0.725	0.867	0.864	0.813	0.856	0.730	0.815	0.731	0.873
		Correct	0.353	0.254	0.077	0.060	0.154	0.102	0.269	0.179	0.260	0.052
	50	Optimal	0.672	0.772	0.934	0.928	0.875	0.898	0.743	0.792	0.751	0.968
		Correct	0.325	0.224	0.033	0.026	0.113	0.086	0.253	0.204	0.249	0.022
	100	Optimal	0.662	0.788	**0.988**	**0.992**	0.912	0.942	0.742	0.804	0.773	0.990
		Correct	0.338	0.212	0.012	0.008	0.088	0.058	0.258	0.196	0.227	0.008
Laplace	20	Optimal	0.607	0.670	0.714	0.698	0.711	0.719	0.669	0.756	0.671	0.824
		Correct	0.354	0.259	0.118	0.092	0.180	0.130	0.297	0.189	0.304	0.113
	30	Optimal	0.634	0.758	0.894	0.911	0.840	0.873	0.699	0.788	0.737	0.936
		Correct	0.364	0.240	0.094	0.067	0.156	0.120	0.301	0.209	0.262	0.060
	50	Optimal	0.668	0.775	0.956	0.957	0.890	0.923	0.740	0.815	0.756	0.978
		Correct	0.332	0.225	0.038	0.031	0.108	0.074	0.260	0.185	0.244	0.022
	100	Optimal	0.718	0.810	**0.982**	**0.990**	0.938	0.960	0.762	0.808	0.777	0.979
		Correct	0.282	0.190	0.018	0.010	0.062	0.040	0.238	0.192	0.223	0.021
Mixture	20	Optimal	0.561	0.657	0.749	0.760	0.713	0.745	0.701	0.798	0.734	0.879
		Correct	0.425	0.316	0.165	0.124	0.242	0.186	0.293	0.192	0.263	0.116
	30	Optimal	0.563	0.685	0.868	0.897	0.789	0.833	0.751	0.814	0.797	0.914
		Correct	0.437	0.315	0.130	0.101	0.211	0.166	0.249	0.186	0.203	0.086
	50	Optimal	0.615	0.717	0.930	0.942	0.852	0.885	0.788	0.844	0.803	0.936
		Correct	0.385	0.283	0.068	0.055	0.148	0.115	0.212	0.156	0.197	0.064
	100	Optimal	0.642	0.760	**0.974**	**0.984**	0.902	0.924	0.844	0.860	0.836	0.971
		Correct	0.358	0.240	0.026	0.016	0.098	0.076	0.156	0.140	0.164	0.029
Cauchy	20	Optimal	0.506	0.485	0.393	0.349	0.465	0.419	0.535	0.534	0.520	0.245
		Correct	0.233	0.152	0.049	0.036	0.098	0.054	0.235	0.142	0.228	0.018
	30	Optimal	0.673	0.739	0.745	0.710	0.773	0.772	0.685	0.752	0.617	0.366
		Correct	0.285	0.185	0.040	0.032	0.109	0.067	0.253	0.170	0.297	0.010
	50	Optimal	0.697	0.790	0.808	0.769	0.854	0.849	0.688	0.762	0.693	0.492
		Correct	0.281	0.169	0.018	0.013	0.071	0.046	0.289	0.205	0.295	0.004
	100	Optimal	0.756	**0.848**	**0.982**	**0.980**	**0.958**	**0.966**	0.744	0.804	0.700	0.604
		Correct	0.242	0.150	0.002	0.002	0.040	0.030	0.252	0.192	0.300	0.000
Slash	20	Optimal	0.330	0.281	0.160	0.123	0.229	0.177	0.382	0.345	0.375	0.125
		Correct	0.157	0.080	0.030	0.021	0.047	0.033	0.173	0.092	0.197	0.016
	30	Optimal	0.548	0.565	0.448	0.404	0.550	0.505	0.582	0.626	0.531	0.195
		Correct	0.278	0.183	0.040	0.025	0.093	0.062	0.244	0.155	0.240	0.007
	50	Optimal	0.610	0.673	0.524	0.485	0.657	0.620	0.647	0.691	0.688	0.294
		Correct	0.283	0.170	0.023	0.013	0.071	0.049	0.249	0.176	0.260	0.002
	100	Optimal	0.702	0.802	0.828	0.804	**0.914**	**0.916**	0.736	0.788	0.724	0.349
		Correct	0.294	0.190	0.004	0.004	0.048	0.034	0.254	0.202	0.268	0.000

Table 3. Probabilities of selecting Optimal (and Correct model) for Model II D

S. Size	Error	Sp-(SVR with New S.D.) with penalty									
		SSp6	SSp6	SSp6	SSp6	SSp6	SSp6	Sp1	Sp2	RAIC	RBIC
20	N(0, 1)	0.991	0.977	0.936	0.919	0.963	0.953	0.997	0.992	1.000	0.999
	t2	0.926	0.886	0.739	0.694	0.836	0.771	0.935	0.899	0.957	0.798
	L(0, 1)	0.966	0.944	0.871	0.839	0.914	0.888	0.966	0.944	0.986	0.960
	Mixture	0.987	0.984	0.941	0.915	0.973	0.947	0.994	0.990	1.000	0.999
	Cauchy	0.777	0.669	0.499	0.433	0.588	0.530	0.758	0.664	0.753	0.266
	Slash	0.510	0.401	0.208	0.159	0.310	0.226	0.527	0.420	0.572	0.147
30	N(0, 1)	1.000	0.997	0.984	0.972	0.992	0.987	1.000	1.000	1.000	1.000
	t2	0.971	0.957	0.857	0.805	0.924	0.896	0.975	0.962	0.991	0.910
	L(0, 1)	0.997	0.990	0.939	0.919	0.968	0.952	0.995	0.992	0.999	0.994
	Mixture	0.997	0.995	0.976	0.961	0.989	0.984	1.000	0.999	1.000	1.000
	Cauchy	0.884	0.827	0.593	0.536	0.753	0.678	0.852	0.808	0.884	0.308
	Slash	0.710	0.607	0.308	0.248	0.475	0.389	0.665	0.584	0.754	0.167
50	N(0, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	t2	0.999	0.999	0.991	0.988	0.997	0.997	0.999	0.999	0.999	0.973
	L(0, 1)	1.000	1.000	0.999	0.998	0.999	0.999	1.000	1.000	1.000	1.000
	Mixture	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Cauchy	0.984	0.978	0.898	0.874	0.960	0.946	0.985	0.978	0.967	0.365
	Slash	0.939	0.897	0.621	0.562	0.802	0.744	0.928	0.901	0.882	0.144
100	N(0, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	t2	0.999	0.999	0.999	0.999	0.999	0.999	1.000	1.000	1.000	0.999
	L(0, 1)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Mixture	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Cauchy	0.996	0.996	0.981	0.978	0.991	0.990	1.000	1.000	0.997	0.469
	Slash	0.994	0.991	0.857	0.824	0.959	0.945	0.994	0.994	0.977	0.237

large samples almost all criteria select optimal model with probability 1 except for Cauchy and Slash errors. For small samples ($n = 20$) the performance of RBIC is greater than SSp so, for a small sample size it is suggested to use RBIC.

In the simulation, for every model there are six different values of probabilities of selecting optimal model corresponding to six different error distributions. For sample size $n = 100$, obtain the average and standard deviation of these six probabilities for each model. The results are reported in Table 4**.**

From the summary statistics it is observed that when full model is an optimal model then SSp1 and SSp2 perform better than all other statistics considered in this simulation. When full model is not an optimal model then SSp3 and SSp4 works better than others.

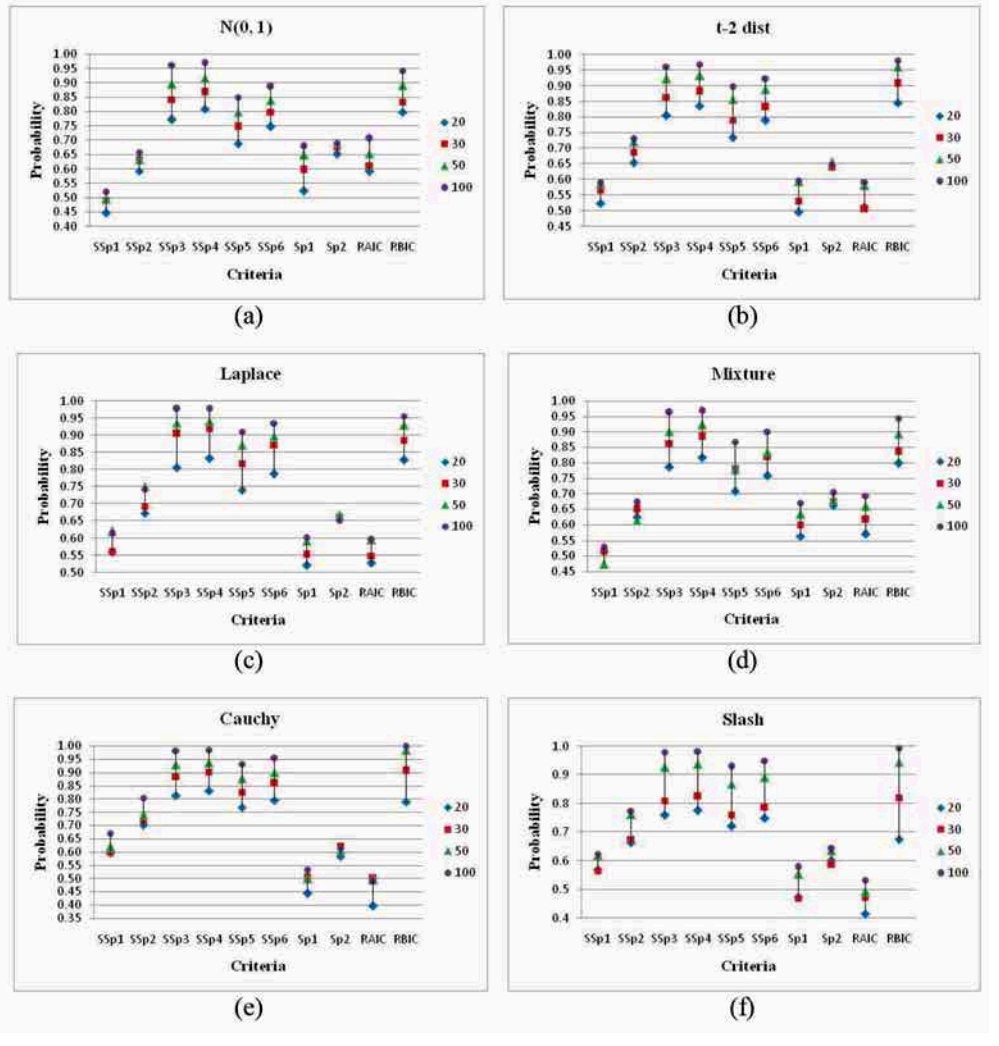

Figure 1. Probabilities of selecting optimal model : Model-I A

Real Data Application

***Example 1*:** To observe the performance of various criteria, consider Brownlee's stack loss data (Hand et al., 1994, pp 156) which contains observations from 21 days operation of a plant for the oxidation of ammonia as a stage in the production of nitric acid. The predictor variables are X_1= air flow, X_2= cooling water inlet temperature (^{o} C), X_3= acid concentration (%) and the response variable Y = stack loss. Stack loss is the percentage of the ingoing ammonia that escapes unabsorbed.

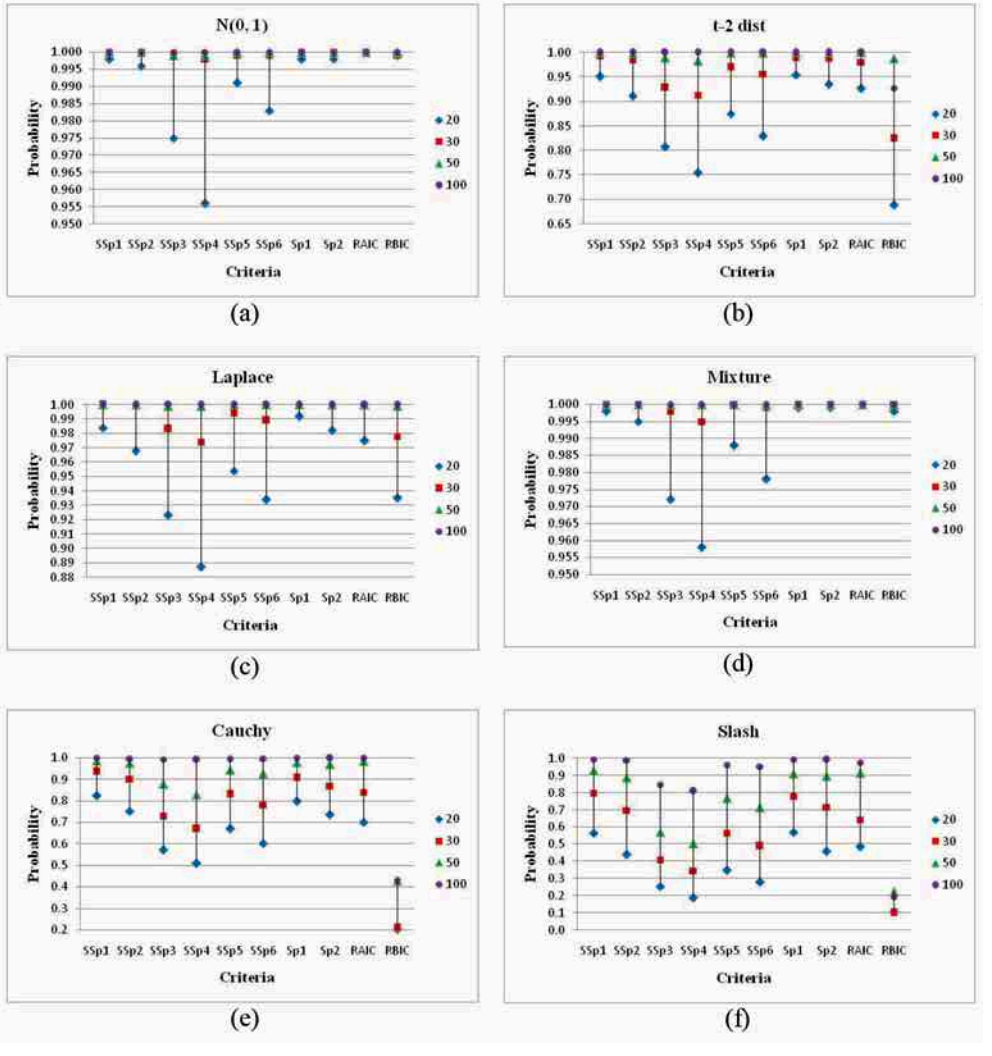

Figure 2. Probabilities of selecting optimal model: Model-I C

As mentioned in Montgomery et al. (2006, p. 396) observation no. 21 is an influential observation because it has a standardized residual of –2.64 from LS fit (Rousseuw & Leroy, 1987, p. 226/227; Rousseuw & van Zomeren, 1990). For a single outlier, replace observation 21 by 7.5 instead of original value 1.5. As a result, standardized residual corresponding to observation no. 21 becomes 4.01 from LS fit, which indicates that observation no. 21 is a potential outlier. Similarly, for two outliers, replace observation 21 by 7.5 and observation14 by 6 instead of original value 1.2, the corresponding standardized residuals from LS fit are 2.63 and 2.56 respectively, which indicates that observations 21 and 14 are outliers.

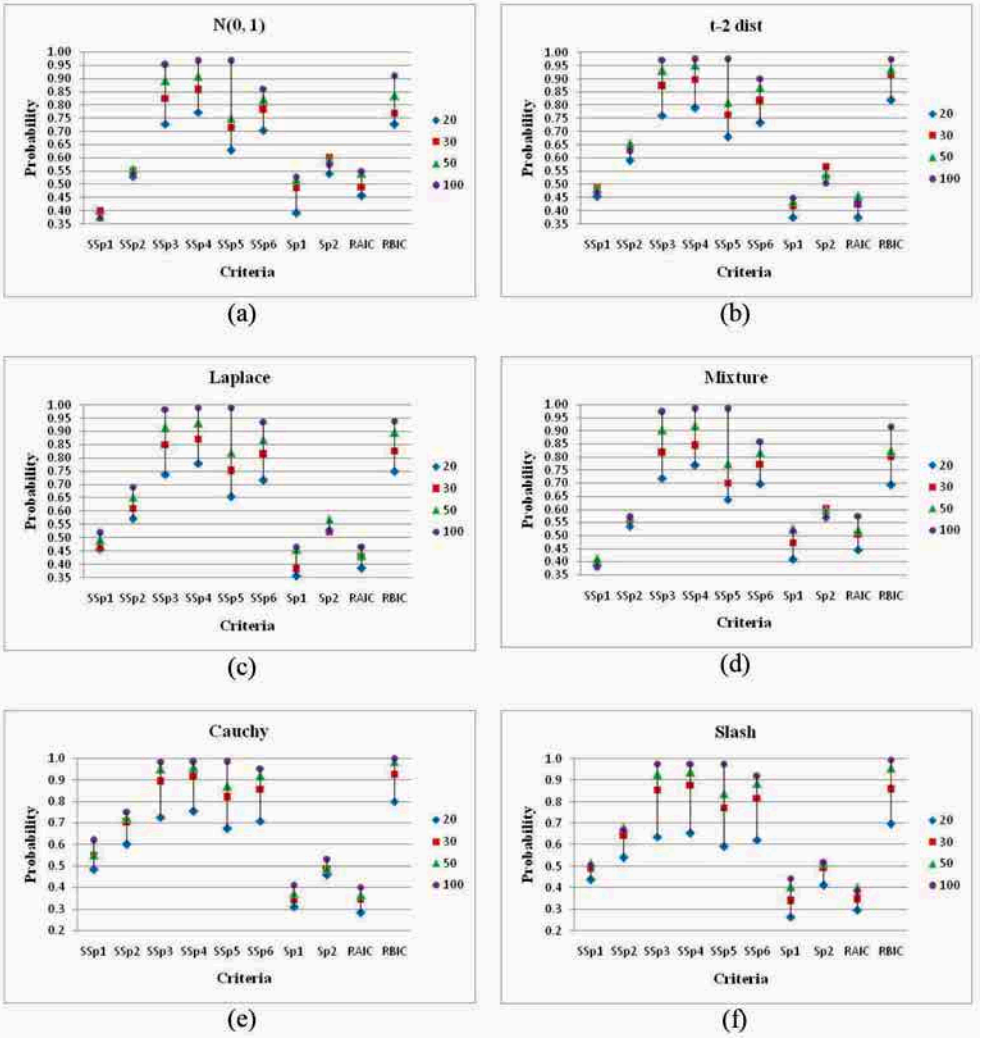

Figure 3. Probabilities of selecting optimal model : Model-II A

For original and outlier data, calculate the SSp statistic with various penalties and $\hat{\sigma}_2^2$, Sp statistic using SVR (with C = MQD and ε = C $\times 10^{-6}$) estimates and M-estimator, also Cp statistic using LS. Sp1 refers to the Sp statistic using M estimator and $\hat{\sigma}_1^2$ and Sp2 for Sp statistic using M estimator and $\hat{\sigma}_2^2$. Also, for comparison purpose, calculate RAIC and RBIC for the original and outlier data. The results are presented in the Table 5.

For the original data, all statistics Sp, Cp, SSp with all penalties, Sp1, Sp2, RAIC and RBIC select the subset $\{X_1, X_2\}$. For one outlier, statistics Sp, SSp with all

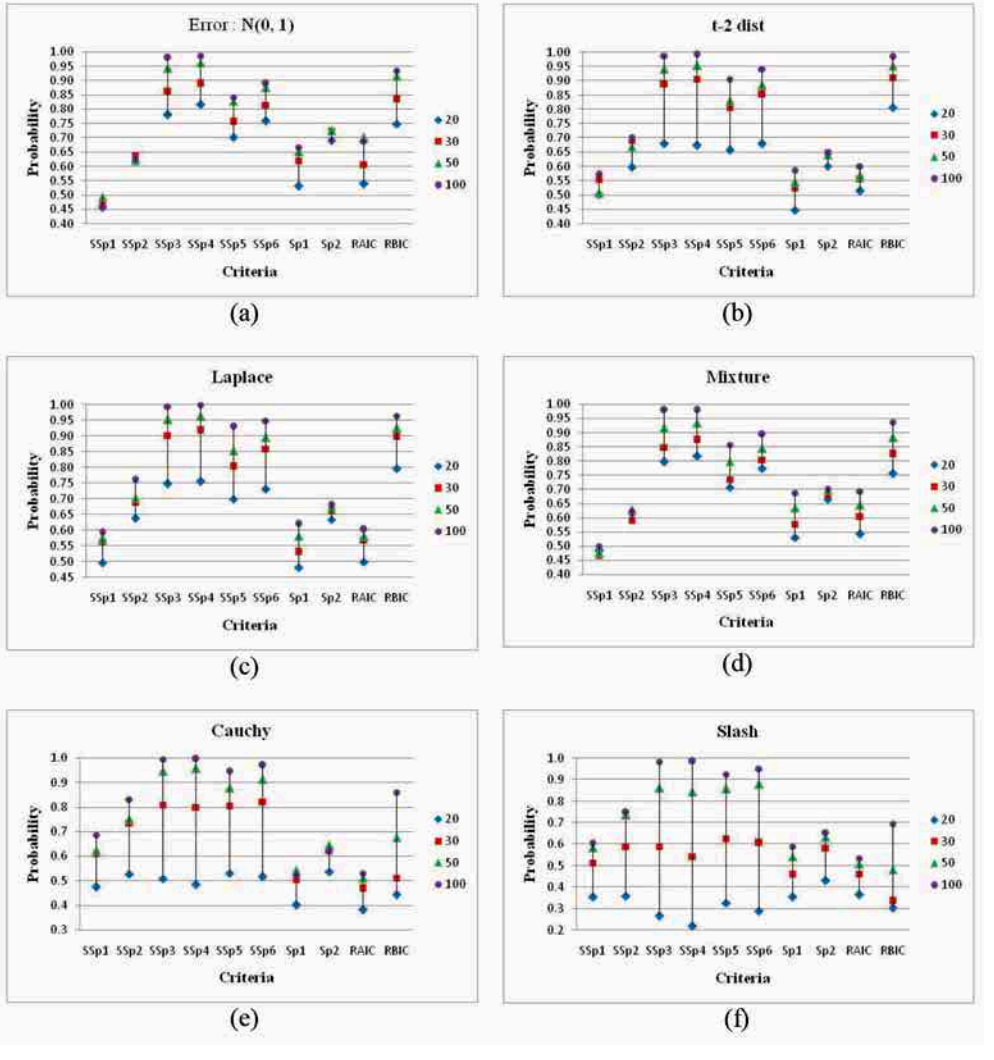

Figure 4. Probabilities of selecting optimal model : Model-II B

penalties and RAIC select the same subset $\{X_1, X_2\}$ but Cp, Sp2 and RBIC select wrong subset $\{X_1\}$. For two outliers, statistics SSp with all penalties select the same subset $\{X_1, X_2\}$, while others select wrong subsets. In particular, Sp and Cp select subset $\{X_1, X_2, X_3\}$ and Sp1, Sp2, RAIC and RBIC select subset $\{X_1\}$.

***Example 2*:** Consider the wine quality data from Montgomery et al. (2006, Table B.14, p. 578), which contains 38 observations on response variable wine quality (Y) based on five predictor variables as clarity (X_1), aroma (X_2), body (X_3), flavor (X_4)

Table 4. Summary Statistics for probability of selecting optimal model for sample size n=100

Model	Statistic	SSp1	SSp2	SSp3	SSp4	SSp5	SSp6	Sp1	Sp2	RAIC	RBIC
Model IA	Average	0.591	0.730	**0.971**	**0.975**	0.898	0.925	0.610	0.658	0.602	0.970
	S.D.	0.058	0.057	0.009	0.007	0.033	0.027	0.056	0.033	0.086	0.025
Model IB	Average	0.712	0.813	**0.960**	**0.959**	0.923	0.942	0.770	0.802	0.767	0.793
	S.D.	0.044	0.039	0.059	0.069	0.025	0.022	0.046	0.033	0.042	0.294
Model IC	Average	**0.999**	**0.998**	0.973	0.967	0.993	0.992	0.999	0.998	0.996	0.759
	S.D.	0.002	0.004	0.062	0.077	0.015	0.019	0.002	0.004	0.010	0.354
Model IIA	Average	0.476	0.641	**0.973**	**0.979**	**0.979**	0.905	0.469	0.538	0.467	0.956
	S.D.	0.093	0.079	0.011	0.009	0.009	0.038	0.045	0.028	0.079	0.039
Model IIB	Average	0.569	0.715	**0.987**	**0.990**	0.901	0.933	0.612	0.666	0.608	0.896
	S.D.	0.082	0.081	0.006	0.007	0.042	0.033	0.059	0.031	0.072	0.107
Model IIC	Average	0.682	0.790	**0.956**	**0.956**	0.918	0.937	0.775	0.819	0.777	0.812
	S.D.	0.053	0.040	0.063	0.075	0.027	0.022	0.046	0.028	0.060	0.272
Model IID	Average	**0.998**	**0.998**	0.973	0.967	0.992	0.989	0.999	0.999	0.996	0.784
	S.D.	0.003	0.004	0.057	0.071	0.016	0.022	0.002	0.002	0.009	0.342

and oakiness (X_5). Test this data for outliers and multicollinearity using Minitab. This data contains only one influential observation, which is observation 20 that has a standardized residual of 2.74. Apply the procedure described in Example 1 for subset selection to this original data. Use all the considered criteria SSp with all penalties, Cp, Sp, Sp1, Sp2, RAIC and RBIC, and select the subset $\{X_2, X_4, X_5\}$.

For a single outlier, replace observation 20 by 39.5 instead of original value 7.9, as a result its standardized residual become 5.51, which indicates that observation 20 is a potential outlier. Apply the same procedure for this outlier data. Using the criteria SSp2, SSp3, SSp4, SSp5 and SSp6, select the subset $\{X_2, X_4, X_5\}$. The criteria Sp1, Sp2 and RAIC select a different subset $\{X_1, X_3, X_4, X_5\}$ and RBIC selects the subset $\{X_4\}$.

Discussion

The modified Sp criterion was used for subset selection in regression in the presence of outliers and/or error distribution is non normal. Implementation of modified Sp criterion requires a penalty term. The choices for the penalty terms are not limited to those mentioned in this paper. A more suitable penalty can give superior performance than listed here. The proposed modification makes the criterion free from assumption of the distribution of errors and is even mitigates the effect of outliers. The simulation results confirm these findings.

Table 5. Model selection for Brownlee's stack loss data.

Subset	Sp (SVM)	Cp	SSp						Sp1	Sp2	RAIC	RBIC
			SSp1	SSp2	SSp3	SSp4	SSp5	SSp6				
Stack Loss data (Original)												
1*	6.11	13.34	9.37	11.37	16.54	18.54	18.54	15.46	11.95	8.74	42.91	68.42
2	118.77	28.93	182.28	184.28	189.44	191.44	191.44	188.37	65.20	47.69	73.78	80.69
3	290.65	148.26	446.06	448.06	453.22	455.22	455.22	452.15	265.80	194.43	148.05	95.91
12	**2.63**	**2.95**	**2.97**	**5.97**	**13.71**	**16.71**	**16.71**	**12.10**	**3.44**	**3.05**	**34.07**	**64.60**
13	7.51	14.39	10.46	13.46	21.21	24.21	24.21	19.59	12.82	9.91	43.94	70.93
23	125.65	30.16	191.77	194.77	202.52	205.52	205.52	200.90	64.15	47.46	74.41	83.31
123	4.00	4.00	4.00	8.00	18.33	22.33	22.33	16.18	4.40	4.30	34.79	66.67
One Outlier												
1	5.52	**1.72**	7.99	9.99	15.15	17.15	17.15	14.08	4.32	**2.58**	114.03	**90.25**
2	151.59	18.67	219.53	221.53	226.70	228.70	228.70	225.62	136.70	81.69	201.02	102.48
3	337.67	21.72	489.03	491.03	496.20	498.20	498.20	495.12	375.53	224.42	276.59	109.30
12	**2.35**	2.05	**2.51**	**5.51**	**13.26**	**16.26**	**16.26**	**11.64**	**2.98**	2.59	**113.86**	92.88
13	9.07	3.67	12.24	15.24	22.98	25.98	25.98	21.37	5.45	4.06	115.51	93.19
23	158.05	17.87	227.99	230.99	238.74	241.74	241.74	237.12	129.06	77.93	198.61	105.05
123	4.00	4.00	4.00	8.00	18.33	22.33	22.33	16.18	4.09	4.05	115.27	95.81
Two Outliers												
1	4.02	4.88	10.60	12.60	17.76	19.76	19.76	16.69	**0.87**	**0.72**	**132.12**	**93.45**
2	14.04	16.59	225.06	227.06	232.22	234.22	234.22	231.15	82.33	68.48	194.30	101.75
3	8.26	10.22	472.81	474.81	479.98	481.98	481.98	478.90	172.73	143.67	229.06	105.28
12	4.13	4.21	**3.16**	**6.16**	**13.91**	**16.91**	**16.91**	**12.29**	2.50	2.42	134.02	96.47
13	3.83	4.66	23.81	26.81	34.56	37.56	37.56	32.94	3.06	2.88	133.38	96.37
23	9.38	11.38	244.32	247.32	255.07	258.07	258.07	253.45	71.93	60.17	188.15	103.88
123	**4.00**	**4.00**	4.00	8.00	18.33	22.33	22.33	16.18	4.75	4.63	135.34	99.41

*Note: The symbol ij means the variables X_i and X_j are in the model.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *Proceedings of the Second International Symposium of Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Baierl, A., Futschik, A., Bogdan, M., & Bieecek, P. (2007). Locating multiple interacting quantitative trait loci using robust model selection. C*omputational Statistics and Data Analysis, 51*(12), 6423-6434. doi: 10.1016/j.csda.2007.02.010

Birkes, D., & Dodge, Y. (1993). *Alternative Methods of Regression*. New York: John Wiley and Sons, Inc. doi: 10.1002/9781118150238

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *COLT '92 Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). New York: ACM. doi: 10.1145/130385.130401

Bozdogan H., & Haughton D. M. A. (1998). Informal complexity criteria for regression models. *Computational Statistics and Data Analysis, 28*(1), 51-76. doi: 10.1016/s0167-9473(98)00025-5

Christmann, A., & Van Messem, A. (2008). Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research, 9*, 623–644.

Christmann, A., Van Messem, A., Steinwart, I. (2009). On consistency and robustness properties of support vector machines for heavy-tailed distributions. *Statistics and Its Interface, 2*(3), 311–327. doi: 10.4310/sii.2009.v2.n3.a5

Desai, S. S., & Kashid, D. N. (2015). Estimation of regression parameters using svm with new methods for meta parameter. International *Journal of Data Mining, Modeling and Management, 7*(3), 239-256. doi: 10.1504/ijdmmm.2015.071449

Draper, N. R., & Smith, H. (2003). *Applied regression analysis, Third edition*. New York: John Wiley and Sons, Inc.

Gunn, S. R. (1998). Support vector machines for classification and regression (Technical Report). School of Electronics and Computer Science, University of Southampton.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., & Ostrowski, E. (1994). *A handbook of small data sets*. London: Chapman and Hall. doi: 10.1007/978-1-4899-7266-8

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics, 32*(1), 1-49. doi: 10.2307/2529336

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley and Sons, Inc. doi: 10.1002/0471725250

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Annals of Mathematical Statistics, 43*(5), 1449-1458. doi: 10.1214/aoms/1177692377

Kashid, D. N., & Kulkarni, S. R. (2002). A more general criteria for subset selection in multiple linear regressions. *Communication in Statistics Theory and Method, 31*(5), 795-811. doi: 10.1081/sta-120003653

Kundu, D., & Murali, G. (1996). Model selection in linear regression. *Computational Statistics and Data Analysis, 22*(5), 461-469. doi: 10.1016/0167-9473(96)00008-4

Mallow, C. L. (1973). Some comments on Cp. *Technometrics, 15*(4), 661-675. doi: 10.1080/00401706.1973.10489103

Miller, A. J. (2002). *Subset Selection in Regression*. Chapman and Hall. doi: 10.1201/9781420035933

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introduction to linear regression analysis, Third edition*. New York: John Wiley and Sons Inc.

Nalbantov, G. (2003). Short horizon value growth style rotation with support vector machines (Unpublished doctoral thesis). Maastricht University, Netherlands.

Rao, C. R., and Wu, Y. (1989). A strongly consistent procedure for model selection in regression problem. *Biometrika, 71*, 43-49.

Ronchetti, E. (1985). Robust model selection in regression. *Statistics & Probability Letters, 3*(1), 21-23. doi: 10.1016/0167-7152(85)90006-9

Ronchetti, E., & Staudte, R. G. (1994). A robust version of Mallow's Cp. *Journal of The American Statistical Association, 89*(426), 550-559. doi: 10.1080/01621459.1994.10476780

Rousseeuw, P. J., & Leroy A. M. (1987). *Robust regression and outlier detection*. New York: John Wiley and Sons, Inc. doi: 10.1002/0471725382

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*(411), 633-639. doi: 10.1080/01621459.1990.10474920

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461-464. doi: 10.1214/aos/1176344136

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika, 68*(1), 45-54. doi: 10.1093/biomet/68.1.45

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199-222. doi: 10.1023/b:stco.0000035301.49549.88

Thompson, M. L. (1978a). Selection of variables in multiple regression: Part I. A review & evaluation. *International Statistical Review, 46*(1), 1-19. doi: 10.2307/1402505

Thompson, M. L. (1978b). Selection of variables in multiple regression: Part II. A review & evaluation. *International Statistical Review, 46*(2), 126-146. doi: 10.2307/1402809

Van Messem, A., & Christmann, A. (2010). A review on consistency and robustness properties of support vector machines for heavy-tailed distributions. *Advances in Data Analysis and Classification, 4*(2-3), 199–220. doi: 10.1007/s11634-010-0067-2

Vapnik, V., Golowich, S., & Smola, A. (1997). Support vector method for function approximation, regression estimation & signal processing. In Mozer, M., Jordan, M., & Petshe, T. (Eds.), *Advances in Neural Information Processing Systems 9* (pp. 281-287), Cambridge, MA: MIT Press.

Vapnik, V. (2001). *The nature of statistical learning theory, Second Edition*. New York: Springer.