ARTICLE TYPE

Outlier Robust Model Averaging Based on $S_p$–Criterion

Yunfei Guo$^{1,2}$ | Zhonghua Li$^*$

$^1$School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China
$^2$Mathematics Department, Yanbian University, Jilin, China

Correspondence
$^*$Zhonghua Li, School of Statistics and Data Science, Nankai University, 94 Weijin Road, Tianjin 300071, China. Email: zli@nankai.edu.cn

Summary

In this paper, we consider the problem of obtaining appropriate weights for model averaging in the face of model uncertainty and outliers. Most of the current model averaging methods are based on ordinary least squares (OLS) estimators, which are very sensitive to outliers or departures from the normality assumption on the error distribution. To overcome this problem, we propose an outlier robust $S_p$–type model averaging (SMA) procedure whenever data contain outliers. The proposed method combines the advantage of model averaging and robust model selection. Therefore, it performs better than other classical model averaging and robust model section methods, especially in the outlier data case and nearly the same as Mallow’s Model Averaging (MMA) even though in the clean data case. Moreover, the proposed criterion is quadratic in the weights, so it is operationally simple to implement. The performance of the proposed method is illustrated through some extensive simulation studies as well as real data examples.

KEYWORDS:
Linear regression, Model averaging, Model selection, Outliers, Robust

1 | INTRODUCTION

Model selection and model averaging are two main approaches to deal with model uncertainty, which commonly exists in the regression models. Model selection has been developed for many years, see Akaike Information Criterion (AIC; Akaike [1974]), Mallows’s $C_p$ (Mallows [1973]), the Bayesian Information Criterion (BIC; Schwarz [1978]), delete-one cross validation (Stone [1974]), and the Focused Information Criterion (FIC; Claeskens and Hjort [2003]), to name a few. They all intend to select the one that most approximates the true models from some potential candidate models.

As an alternative, model averaging deals with model uncertainty not by having the users select one model from among a set of candidate models according to a criterion such as $C_p$, AIC, or BIC, but instead by averaging over the set of candidate models in a particular manner. There is a long-standing literature on Bayesian model averaging; see [Hoeting, Madigan, Raftery, and Volinsky 1999] for a comprehensive review and recent works [Hu, O’Hagan, and Murphy 2018]. There is also a rapidly-growing literature on frequentist methods for model averaging, including [Buckland, Burnham, and Augustin 1997], [Hansen 2007], [Wan, Zhang, and Zou 2010], [Hansen and Racine 2012], [Liu and Okui 2013], [Lu and Su 2015], [Ando, Li et al. 2017], [Zhang, Chiou, and Ma 2018], [Sun, Hong, Lee, Wang, and Zhang 2020], [Li, Lv, Wan, and Liao 2020] and [Feng and Liu 2020], and etc.

Most of the existing estimation procedures for model selection and model averaging were based on ordinary least squares (OLS) estimators. Therefore, these methods are sensitive to outliers. Robust model selection, as an extension of model selection, has attracted more and more attention of researchers, such as a robust AIC (RAIC; Hampel 1983) and [Ronchetti 1985], a robust BIC (RBIC; Machado 1993); a robust $C_p$ (RC$C_p$; Ronchetti and Staudte 1994), a robust $T_p$ based on Wald test statistic ($RT_p$; Sommer and Huggins 1998), a robust version of cross–validation (RCV; Ronchetti, Field, and Blanchard 1997) and a weighted version of likelihood estimator (Agostinelli 2002). These proposals are mainly based on robust versions of classical selection criteria, see [Ronchetti 1997] for a comprehensive review. On the other hand, Müller and Welsh 2005 make use of stratified bootstrap to combine a measure of goodness-of-fit, a penalty term for the number of parameters and the expected prediction
error. However, the computation of their robust method is intensive because of the bootstrapping. To overcome the problem of extremely time consuming [Salibian-Barrera and Van Aelst (2008)] propose a fast and robust bootstrap method of robust model selection. Besides, [Kashid and Kulkarni (2002); Dorugade and Kashid (2010) and JadHAV, Kashid, and Kulkarni (2014)] propose a general framework of model selection in the presence of outliers and multi-collinearity.

Although robust model selection methods take outliers into account, they ignore the uncertainty in the model selection process to some extent. Model averaging, however, only consider enough model uncertainty, not for outliers. To the best of our knowledge, there are few works on outlier robust model averaging except the work by [Wang and Zou 2019], where a general Mallows-type model average (MTMA) was proposed. Although our proposed SMA and MTMA of [Wang and Zou (2019)] both aim at providing an outlier robust model averaging estimator, they are two quite different approaches. The main difference lies in that they can be seen as a model averaging version of the $S_p$ criterion (Kashid and Kulkarni 2002) and of a general Akaike-type criterion ([Burman and Nolan 1995]), respectively. A detailed explanation and comparison with the MTMA method are deferred to Section 4. Based on these considerations, we propose a frequentist model averaging method based on $S_p$ criterion (Kashid and Kulkarni 2002), which simultaneously takes model uncertainty and outliers into account.

The remainder of the article is organized as follows. Section 2 describes the model framework and model averaging estimators. In Section 3 we present the proposed weight selection criteria. Section 4 investigates the finite sample performance of our proposed method through extensive simulations and then we apply the proposed method to a real data example in Section 5. Some concluding remarks are contained in Section 6.

## 2 Model Averaging Estimation

Suppose that we observe a random sample from the following linear regression model:

$$
y_i = \mu + \varepsilon_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n,
$$

(1)

where $x_i = (x_{i1}, \ldots, x_{ip})^T$ is the $i^{th}$ $p \times 1$ independent variable vector, $y_i \in R$ is the $i^{th}$ dependent variable, $\beta = (\beta_1, \ldots, \beta_p) \in R^p$ is the corresponding coefficient vector, $\mu$ is the mean of $y_i$ and $\varepsilon_i$ is a random error with mean 0 and variance $\sigma^2$. Write $X = (x_1, \ldots, x_n)^T, y = (y_1, \ldots, y_n)^T, \mu = (\mu_1, \ldots, \mu_n)^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$, we can obtain the matrix form of model (1):

$$
y = \mu + \varepsilon = X \beta + \varepsilon.
$$

Our objective is to estimate $\mu$. To this end, we consider the set of $M$ candidate models. The $m^{th}$ model has $k_m > 0$ regressors that can be any variables in $x_i$. Without loss of generality, we assume the $M^{th}$ model is the largest model throughout this paper. For the $m^{th}$ approximating model of model (1),

$$
y_i = \sum_{j=1}^{k_m} \beta_{jm} x_{jm} + b_{im} + \varepsilon_i = \sum_{j=1}^{k_m} \beta_{jm} x_{jm} + \varepsilon_{im},
$$

(2)

for $m = 1, 2, \ldots, M$, where $x_{jm}$ for $j = 1, \ldots, k_m$ denotes the regressors in the $m^{th}$ model, $\beta_{jm}$ denotes the corresponding coefficients, $b_{im} = \mu_i - \sum_{j=1}^{k_m} \beta_{jm} x_{jm}$ is the approximation error, and $\varepsilon_{im}$ is the error term of the $m^{th}$ approximating model. The matrix form of model (2) can be written as:

$$
y = X_{(m)} \beta_{(m)} + \varepsilon_{(m)},
$$

where $X_{(m)}$ is an $n \times k_m$ regressor matrix including $k_m$ columns of $X$, $\beta_{(m)}$ is the corresponding coefficient vector, and $\varepsilon_{(m)} = (\varepsilon_{1(m)}, \ldots, \varepsilon_{n(m)})$.

The estimator of $\mu$ from the $m^{th}$ model is: $\hat{\mu}_{m} = X_{(m)} \hat{\beta}_{(m)}$, and the residual is $\hat{\varepsilon}_{(m)} = y - \hat{\mu}_{(m)}$, where $\hat{\beta}_{(m)}$ is some appropriate estimator of $\beta_{(m)}$.

Let $w = (w_1, \ldots, w_M)^T$ be a weight vector in the set $W = \{w \in [0, 1]^M: \sum_{m=1}^{M} w_m = 1\}$. The model averaging estimator of $\mu$ is then defined as:

$$
\hat{\mu}(w) = \sum_{m=1}^{M} w_m X_{(m)} \beta_{(m)}.
$$

(3)

The weight $w$ is unknown, so methods of estimating $w$ is needed. Let $k = (k_1, \ldots, k_M)$. Hansen (2007) proposed choosing weights in $\hat{\mu}(w)$ by minimizing the following Mallows criterion:

$$
C_0(w) = ||\hat{\mu}(w) - y||^2 + 2\sigma^2 w^T k.
$$

Since $\hat{\mu}(w)$ and the estimator of $\sigma^2$ in this criterion are both based on least square estimators, it is sensitive to the outliers. Therefore, the need of a more robust criterion is obvious whenever the data contain outliers. In the next section, we would propose a more robust criterion to choose weights based on $S_p$-criterion. Substituting the appropriate estimated weight, denoted by $\hat{w}$, for $w$ in (3) result in the following model averaging
CHOICE OF WEIGHTS

The key issue of most model averaging methods is how to choose the weight for each candidate model. In this section, we will provide weight selection criterion based on $S_p$-criterion. we will provide the definition and motivation of $S_p$-criterion.

### 3.1 $S_p$-criterion

First, we introduce the Mallows’s $C_p$ statistic [Mallows, 1973], which is one of most frequently used model selection criteria as follows:

$$C_p = \frac{\text{RSS}_m}{\sigma^2} - (n - 2k_m),$$

where $\text{RSS}_m$ is the residual sum of squares of the $m^{th}$ submodel with $k_m$ regressors and $\sigma^2$ is an estimate of the error variance $\sigma^2$ which is obtained from the full model. However, this method is sensitive to outliers since it is based on least square estimators which perform poorly whenever data contain influential observations. In order to overcome this problem, Kashid and Kulkarni [2002] proposed $S_p$-criterion, which is defined as follows.

**Definition 1 ($S_p$-criterion).**

$$S_p = \sum_{i=1}^{n} \frac{(\hat{\mu}_i(M) - \hat{\mu}_i(m))^2}{\sigma^2} - (p_M - 2p_m),$$

where $\hat{\mu}_i(M)$ and $\hat{\mu}_i(m)$ is the fitted value of $\mu_i$ based on the full and the $m^{th}$ candidate model, respectively. $p_M = k_M + 1$ and $p_m = k_m + 1$ are the corresponding parameters.

Now, we give a general motivation of this criterion. For more details, see Jadhav et al [2014] and Kashid and Kulkarni [2002]. Let $\hat{\mu}_i(M) = (\hat{\mu}_1(M), \ldots, \hat{\mu}_n(M))^T$ and $\hat{\mu}_i(m) = (\hat{\mu}_1(m), \ldots, \hat{\mu}_n(m))^T$ be the fitted value of $\mu_i$ from the full and subset model respectively. We consider the Euclidean (scaled) distance between these two vectors,

$$\|\hat{\mu}_i(M) - \hat{\mu}_i(m)\|^2 / \sigma^2 = \sum_{i=1}^{n} (\hat{\mu}_i(M) - \hat{\mu}_i(m))^2 / \sigma^2.$$

Formula (4) can be regarded as a measure of discrepancy or similarity between the predictive abilities of full and submodel. A submodel will be accurate enough to consider only $k_m$ regressors if the distance as given by (4) is smaller, which indicates that the vector $\hat{\mu}_i(m)$ is close to $\hat{\mu}_i(M)$. On the contrary, the prediction based on the candidate model may not be as accurate as that based on the full model if the value of (4) is large.

The estimate vectors $\hat{\mu}_i(M)$ and $\hat{\mu}_i(m)$ can be either OLS estimators or some robust estimators. The $S_p$-statistic reduces to $C_p$-statistic when OLS estimators of $\gamma$ are used (see Kashid and Kulkarni [2002]). Whenever data contain outliers or the errors follow non-normal distribution, some robust estimators such as Huber M-estimator should be used instead of OLS estimators. However, the form of $S_p$-statistic need not to be modified, which is the main advantage of our proposed method.

In the spirit of extending Mallows $C_p$ to MMA [Hansen, 2007], we will propose a method of choosing weight based on the $S_p$-statistic in the presence of outliers.

### 3.2 Choice of weights based on SMA

We propose an outlier robust $S_p$ type model averaging criterion (SMA) that is constructed as follows:

$$S(w) = \|\hat{\mu}_i(M) - \hat{\mu}_i(w)\|^2 + 2\sigma^2 w^T k = \sum_{i=1}^{n} (\hat{\mu}_i(M) - \hat{\mu}_i(w))^2 + 2\sigma^2 \sum_{m=1}^{n} w_k k_m,$$

where $\hat{\mu}_i(w)$ is the $i^{th}$ element of $\hat{\mu}(w)$, other terms are similarly defined as before.

Wang and Zou [2019] proposed a Mallows-type model average (MTMA), where the weights are chosen by a general Mallows-type criterion (MTC) that is constructed as

$$C_n(w) = \sum_{i=1}^{n} \rho (\hat{\mu}_i(w)) + C_\rho \sum_{m=1}^{M} w_k k_m,$$
where \(\delta_i(w) = y_i - \hat{\mu}(w)\), \(\rho(e)\) is some robust loss function, \(C_\rho = \sum_{i=1}^{n} \text{var} \{ \rho(y_i - x_i^T \beta) \} / \sum_{i=1}^{n} R_2 \left( \mu_i - x_i^T \beta \right)\), \(\rho\) is the derivative of \(\rho\), \(R_2\) is the second derivative of some twice differentiable function which is defined as \(R(t) := E_{\varepsilon} \rho(\varepsilon + t)\) and \(\beta\) can be estimated by \(\hat{\beta}\).

There exist essential differences and some relations between SMA and MTMA, which are summarized as follows.

- The first difference is that the residual sum of squares of the \(m\)th submodel RSS\(_m\) in the \(C_\rho\) criterion is replaced with \(\| \hat{\mu}(M) - \hat{\mu}(m) \|^2\) in our SMA, where the squared loss function has not been changed. Instead, for the MTMA estimator, the squared loss function in the \(C_\rho\) criterion is changed into a function being convex with a unique minimum, and twice differentiable in expectation.
- The other distinction is that the penalty on the average parameters \(\sum_{m=1}^{n} w_m \rho_1\) in SMA is \(2\sigma^2\), but \(C_\rho\) for the MTMA estimator.
- The weight choice criteria of SMA and MTMA degenerate into the \(S_p\)-criterion in [Kashid and Kulkarni (2002)] and a general Akaike-type criterion [Burman and Nolan (1995)], respectively, when one component of the vector \(w\) is equal to one and the remaining weights are equal to zero.
- If OLS estimators of \(\hat{\mu}(M)\) and \(\hat{\mu}(w)\) are adopted in SMA, and the loss function \(\rho(t)\) in MTMA is taken into the square loss, both of these two criteria coincide with the Mallows model average criterion proposed by [Hansen (2007)].

Our focus in this paper is mainly on the case in which the data contain outliers. In the next, we will discuss how to choose the weight in \(S(w)\). Instead of OLS estimators of \(\mu\), we should use some robust estimators such as Huber M–estimator [Huber (2004)]. Consider the model as defined in [1]. The objective is to minimize a sum of some robust loss function of the residuals

\[
\sum_{i=1}^{n} \rho(y_i - x_i^T \beta) 
\]

for some function \(\rho\). We adopt but not limited to the Huber loss function such as

\[
\rho(t) = \begin{cases} 
2\varepsilon ; & |t| \leq \varepsilon, \\
2\varepsilon |t| - \varepsilon^2 ; & |t| > \varepsilon,
\end{cases}
\]

with a tuning parameter \(\varepsilon\), which is taken to 1.345 in the simulation and real data studies. Taking derivatives of (5) with respect to (w.r.t., hereafter) \(\beta\) and setting it equal to 0, we get

\[
\sum_{i=1}^{n} \psi_i(y_i - x_i^T \beta) x_i = 0,
\]

where \(\psi\) denotes the derivative of \(\rho\) w.r.t. \(\beta\). We can obtain the Huber M–estimator \(\hat{\beta}\) of \(\beta\) via using the iteratively reweighted least squares (IRLS) method, which is available in R packages such as MASS.

Denote \(\hat{\beta}_m(M)\) the \(M\)-estimator of \(\beta_m\) based on the \(m\)th candidate model for \(m=1, \ldots, M\). Further, let \(\hat{\mu}_m = X(m) \hat{\beta}_m\) be the robust fitted values from the \(m\)th submodel. Therefore, the model average estimator of \(\mu\) combining \(M\)-estimator is

\[
\hat{\mu}(w) = \sum_{m=1}^{M} w_m \hat{\mu}_m.
\]

At the same time, the robust \(S_p\) statistic corresponding to the \(M\)-estimator \(\hat{\mu}(w)\) can be written as

\[
\hat{S}(w) = \| \hat{\mu}(M) - \hat{\mu}(w) \|^2 + 2\sigma^2 w^T k.
\]

As the variance of the error term \(\sigma^2\) is unknown in many applications, we replace it with an appropriate robust estimator from the largest model. Specifically, like [Kashid and Kulkarni (2002)], we consider the following estimator of \(\sigma^2\):

\[
\hat{\sigma}^2 = [1.4826 \text{med}\{|r(M) - \text{med}(r(M))|\}]^2,
\]

where \(r(M)\) is the residuals from the \(M\)th model and “med” denotes the median function.

Thus, we propose choosing the weights by minimizing

\[
\hat{S}(w) = \| \hat{\mu}(M) - \hat{\mu}(w) \|^2 + 2\hat{\sigma}^2 w^T k.
\]

Substituting \(\hat{w} = \arg \min_{w \in W} \hat{S}(w)\) for \(w\) in (6), we can get the following feasible robust model average estimator of \(\mu\):

\[
\hat{\mu}(\hat{w}) = \sum_{m=1}^{M} \hat{\psi}_m \hat{\mu}_m(m).
\]

Let \(\hat{\varepsilon}_m = \hat{\mu}_m(M) - \hat{\mu}_m\) and \(\hat{E} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_M)\). By some simple calculations, we can express \(\hat{S}(w)\) in the following form:

\[
\hat{S}(w) = w^T \hat{E}^T \hat{E} w + 2\hat{\sigma}^2 w^T k.
\]
It can be interestingly seen that we obtain $\hat{w}$ through solving a classic quadratic programming problem, for which numerical algorithms are readily available.

4 | SIMULATION STUDY

In this section, we conduct extensive Monte Carlo simulations to investigate the finite sample performance of the SMA versus MMA, $C_p$ and $S_p$. The simulations are based on the following model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i \sim N(0, \sigma^2),$$

where $x_{ij}$ are independent and identically distributed from standard normal distribution, $i = 1, \ldots, n$ and $j=1, 2, 3, 4$. We consider the specification of $\beta$ that $\beta = (1, 1, 0, 1, 0)$. There are $2^4 - 1 = 15$ candidate models where we assume that all models contain intercept and at least one regressor.

We consider two types of data:

- Case A: clean data which means that no outlier in the data.
- Case B: outlier data, by changing value of $y_4$ to 10, which are also introduced by Ronchetti and Staudte [1994].

We set $n \in \{25, 50, 100\}$ and $\sigma \in \{0.5, 1, 2\}$. On the base of 1000 replications, we can obtain the empirical mean squared error (MSE)

$$\text{MSE} = \frac{1}{1000} \sum_{r=1}^{1000} ||\hat{\mu}^{(r)} - \mu||^2,$$

where $\hat{\mu}^{(r)}$ represents the estimator of $\mu$ based on the $r^{th}$ replication, obtained by MMA, $C_p$, SMA or $S_p$ methods. We also consider the empirical median squared error (mSE) that is computed as follows:

$$\text{mSE} = \text{median}_{r=1, \ldots, 1000} ||\hat{\mu}^{(r)} - \mu||^2,$$

where $\hat{\mu}^{(r)}$ is defined as before, median$_{r=1, \ldots, 1000} (\cdot)$ refers to the median of a vector of length 1000.

Our simulation results are summarized in Table 1 and Figure 1. And from the presented results, we can conclude that

a. when there are outliers in the data, performance of SMA significantly outperforms other model averaging and robust model selection methods in achieving the lowest MSEs and mSEs, especially in the case where the variance becomes large and sample size is small, a meaningful and promising result.

b. in the clean data case, performance of SMA is as well as MMA and $C_p$, but clearly still better than $S_p$ method,

c. robust methods including SMA and $S_p$ are influenced by the outliers, but much smaller than classical MMA and $C_p$.

As we are not sure whether there are some outliers, which is hard to determine in practice, our proposed SMA estimator is quite a better choice whenever more uncertainty exists in the model selection process.

5 | EMPIRICAL ILLUSTRATION

To compare the finite sample performance of our proposed SMA estimators to that of MMA, $C_p$, and $S_p$ methods, we apply these estimators to analyze a well known datasets in this section. We consider the stack loss data [Brownlee 1965] obtained from 21 days of operation of a plant for the oxidation of ammonia (NH3) to nitric acid (HNO3), which is shown in Table 2. The nitric oxides produced are absorbed in a countercurrent absorption tower. Air Flow represents the rate of operation of the plant. Water Temp is the temperature of cooling water circulated through coils in the absorption tower. Acid Conc is the concentration of the acid circulating, minus 50, times 10: that is, 89 corresponds to 58.9 percent acid. Stack loss (the dependent variable) is 10 times the percentage of the ingoing ammonia to the plant that escapes from the absorption column unabsorbed; that is, an (inverse) measure of the over-all efficiency of the plant.

The model setting is similar to the simulation study. Obviously, we do not know the true data generating process (DGP). However, the data can be randomly divided into two samples, a training sample of size $n_1$ and a test sample of size $n_2$ 1000 times. We can obtain the estimation on the training set and compute the mean squared prediction error (MSPE) and median squared prediction error (mSPE) of each estimator given by $\text{MSPE} = \frac{1}{n_2} \sum_{i=1}^{n_2} (y_i - \hat{y}_i)^2$ and $\text{mSPE} = \text{median}_{i=1, \ldots, n_2} (y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the prediction for a given method. Since outliers may be present both in the training and the test sets, we will focus on the mean squared prediction error of the 95% smallest residuals for each estimator. In other
words, we will take the average of the squared prediction errors after removing the 5% largest of them (in absolute value). Then we evaluate all methods with respect to 5%-trimmed mean of squared prediction errors (5%-trimmed MSPE) and 5%-trimmed median of squared prediction errors (5%-trimmed mSPE).

Many researchers discover different outliers in this data. We provide below the list of some discoveries. See [Dodge, 1996] for more details.

- Observation 21 ([Draper and Smith, 1998]) Least Squares,
- Observations 1, 3, 4, 21 ([Cuthbert, Wood, and Gorman, 1980]) Least Squares + Plots,
- Observations 1, 2, 4, 21 ([Cook, 1979]) Cook’s Distance,
- Observations 3, 4, 21 ([Welsh et al., 1987]) L-Estimate.

Thus, it is most likely that some observations are outliers in the data set. So we can use this data to evaluate our methods. We present 5%-trimmed MSPE and 5%-trimmed mSPE results in Table 3 and Figure 2. It can be seen that our proposed SMA estimator performs best among these four methods, which is an encouraging result for users to choose.

6 CONCLUSIONS AND EXTENSIONS

Taking into consideration of the existence of both outlier observations and model uncertainty, we propose a general outlier robust model averaging method based on $S_p$ criterion, which combines the advantage of model averaging and robust model selection. In the clean data case, SMA and MMA are nearly equivalent, but whenever the data contain outliers, SMA estimators perform significantly better than other model selection and model averaging method. Besides, the proposed criterion is quadratic in the weights, so computation is a simple application of quadratic programming.

In this context, we only proposed the criterion, and evaluated the performance through simulation study and real data examples. In the future we shall prove some excellent properties of these robust average weight estimators, such as the asymptotic optimality of it. Furthermore, we can extend it to other models such as non-linear model and linear mixed models.

ACKNOWLEDGMENTS

We would like to thank the editor, the associate editor and the two anonymous referees for their constructive comments and suggestions that have considerably improved this paper.

FUNDING STATEMENT

This research was supported by the National Natural Science Foundation of China (Grant No. 12071233, 11971247) and the PhD Candidate Research Innovation Fund of Nankai University.

CONFLICT OF INTEREST DISCLOSURE

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The stack loss data ([Brownlee, 1965]) is shown in Table 2.

References


FIGURE 1 Boxplot of MSEs of different estimation methods in the two data types and under different sample size \( n \) and \( \sigma \) settings; SMA for the \( S_p \)-type Model Averaging, Sp for the Sp model selection method, MMA for the Mallows Model Averaging method and Cp for the Mallows’s Cp model selection method.
FIGURE 2 Boxplot of 5%-trimmed MSPEs and mSPEs of different estimation methods in the stack loss data; SMA for the $S_p$-type Model Averaging, Sp for the Sp model selection method, MMA for the Mallows Model Averaging method and Cp for the Mallows's Cp model selection method.

TABLE 1 Results of Simulation: the MSEs and mSEs of different estimation methods; SMA for the $S_p$-type Model Averaging, Sp for the Sp model selection method, MMA for the Mallows Model Averaging method and Cp for the Mallows’s Cp model selection method.

<table>
<thead>
<tr>
<th>n</th>
<th>$\sigma = 0.5$</th>
<th>$\sigma = 1$</th>
<th>$\sigma = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SMA</td>
<td>Sp</td>
<td>MMA</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>25</td>
<td>MSE</td>
<td>1.0521</td>
<td>1.0598</td>
</tr>
<tr>
<td></td>
<td>mSE</td>
<td>1.0383</td>
<td>1.0450</td>
</tr>
<tr>
<td>50</td>
<td>MSE</td>
<td>1.0218</td>
<td>1.0261</td>
</tr>
<tr>
<td></td>
<td>mSE</td>
<td>1.0164</td>
<td>1.0248</td>
</tr>
<tr>
<td>75</td>
<td>MSE</td>
<td>1.0103</td>
<td>1.0135</td>
</tr>
<tr>
<td></td>
<td>mSE</td>
<td>1.0102</td>
<td>1.0132</td>
</tr>
<tr>
<td>100</td>
<td>MSE</td>
<td>1.0061</td>
<td>1.0086</td>
</tr>
<tr>
<td></td>
<td>mSE</td>
<td>1.0039</td>
<td>1.0052</td>
</tr>
</tbody>
</table>

Case A: clean data

Case B: outlier data

Notes: The entries in the table are approximated with 4 decimal places. And the best estimator in each configuration is marked in bold.
### TABLE 2 Stack loss data

<table>
<thead>
<tr>
<th>Obs</th>
<th>Air Flow</th>
<th>Water Temp</th>
<th>Acid Conc.</th>
<th>stack loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80</td>
<td>27</td>
<td>58.90</td>
<td>4.20</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>27</td>
<td>58.80</td>
<td>3.70</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>25</td>
<td>59.00</td>
<td>3.70</td>
</tr>
<tr>
<td>4</td>
<td>62</td>
<td>24</td>
<td>58.70</td>
<td>2.80</td>
</tr>
<tr>
<td>5</td>
<td>62</td>
<td>22</td>
<td>58.70</td>
<td>1.80</td>
</tr>
<tr>
<td>6</td>
<td>62</td>
<td>23</td>
<td>58.70</td>
<td>1.80</td>
</tr>
<tr>
<td>7</td>
<td>62</td>
<td>24</td>
<td>59.30</td>
<td>1.90</td>
</tr>
<tr>
<td>8</td>
<td>62</td>
<td>24</td>
<td>59.30</td>
<td>2.00</td>
</tr>
<tr>
<td>9</td>
<td>58</td>
<td>23</td>
<td>58.70</td>
<td>1.50</td>
</tr>
<tr>
<td>10</td>
<td>58</td>
<td>18</td>
<td>58.00</td>
<td>1.40</td>
</tr>
<tr>
<td>11</td>
<td>58</td>
<td>18</td>
<td>58.90</td>
<td>1.40</td>
</tr>
<tr>
<td>12</td>
<td>58</td>
<td>17</td>
<td>58.80</td>
<td>1.30</td>
</tr>
<tr>
<td>13</td>
<td>58</td>
<td>18</td>
<td>58.20</td>
<td>1.10</td>
</tr>
<tr>
<td>14</td>
<td>58</td>
<td>19</td>
<td>59.30</td>
<td>1.20</td>
</tr>
<tr>
<td>15</td>
<td>50</td>
<td>18</td>
<td>58.90</td>
<td>0.80</td>
</tr>
<tr>
<td>16</td>
<td>50</td>
<td>18</td>
<td>58.60</td>
<td>0.70</td>
</tr>
<tr>
<td>17</td>
<td>50</td>
<td>19</td>
<td>57.20</td>
<td>0.80</td>
</tr>
<tr>
<td>18</td>
<td>50</td>
<td>19</td>
<td>57.90</td>
<td>0.80</td>
</tr>
<tr>
<td>19</td>
<td>50</td>
<td>20</td>
<td>58.00</td>
<td>0.90</td>
</tr>
<tr>
<td>20</td>
<td>56</td>
<td>20</td>
<td>58.20</td>
<td>1.50</td>
</tr>
<tr>
<td>21</td>
<td>70</td>
<td>20</td>
<td>59.10</td>
<td>1.50</td>
</tr>
</tbody>
</table>

### TABLE 3 5%-trimmed MSPEs and mSPEs of different estimation methods for the stack loss data: SMA for the Sp-type Model Averaging, Sp for the Sp model selection method, MMA for the Mallows Model Averaging method and Cp for the Mallows’s Cp model selection method.

<table>
<thead>
<tr>
<th></th>
<th>SMA</th>
<th>Sp</th>
<th>MMA</th>
<th>Cp</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%-trimmed MSPE</td>
<td>0.0645</td>
<td>0.0689</td>
<td>0.0710</td>
<td>0.0736</td>
</tr>
<tr>
<td>5%-trimmed mSPE</td>
<td>0.0348</td>
<td>0.0477</td>
<td>0.0455</td>
<td>0.0544</td>
</tr>
</tbody>
</table>