Robust Linear Model Selection for High-Dimensional Datasets

by

MD JAFAR AHMED KHAN

M.Sc., The University of British Columbia, 2002

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES

(Statistics)

The University of British Columbia

December 2006

Abstract

This study considers the problem of building a linear prediction model when the number of candidate covariates is large and the dataset contains a fraction of outliers and other contaminations that are difficult to visualize and clean. We aim at predicting the future non-outlying cases. Therefore, we need methods that are robust and scalable at the same time.

We consider two different strategies for model selection: (a) one-step model building and (b) two-step model building. For one-step model building, we robustify the step-by-step algorithms forward selection (FS) and stepwise (SW), with robust partial F-tests as stopping rules.

Our two-step model building procedure consists of *sequencing* and *segmentation*. In *sequencing*, the input variables are sequenced to form a list such that the good predictors are likely to appear in the beginning, and the first m variables of the list form a reduced set for further consideration. For this step we robustify Least Angle Regression (LARS) proposed by Efron, Hastie, Johnstone and Tibshirani (2004). We use bootstrap to stabilize the results obtained by robust LARS, and use "learning curves" to determine the size of the reduced set.

The second step (of the two-step model building procedure) - which we call *segmentation* - carefully examines subsets of the covariates in the reduced set in order to select the final prediction model. For this we propose a computationally suitable robust cross-validation procedure. We also propose a robust bootstrap procedure for segmentation, which is similar to the method proposed by Salibián-Barrera and Zamar (2002) to conduct robust inferences in linear regression.

We introduce the idea of "multivariate-Winsorization" which we use for robust data cleaning (for the robustification of LARS). We also propose a new correlation estimate which we call the "adjusted-Winsorized correlation estimate". This estimate is consistent and has bounded influence, and has some advantages over univariate-Winsorized correlation estimate (Huber 1981 and Alqallaf 2003).

Contents

List of Tables

List of Figures

List of Notation

n	number of observations (rows) in the dataset
d	total number of covariates
m	number of covariates in the reduced set
r_{jY}	correlation between covariate X_j and response Y
$r_{jY.1}$	partial correlation between covariate X_j $(j \neq 1)$ and response Y adjusted for X_1
$\tilde{r}_{jY.1}$	quantity proportional to $r_{jY.1}$
w.o.l.g.	without loss of generality
FS	forward selection procedure
SW	stepwise procedure
LARS	Least Angle Regression
SD	standard deviation
med	median
mad	median absolute deviation from median
MSPE	mean squared prediction error

Acknowledgements

I would like to express my most sincere and deepest gratitude to my reverend supervisor, Dr. Ruben H. Zamar. This thesis would not have been completed without his excellent guidance, generous support and continuous motivation. In the beginning, when I wanted to quit my study due to my son's illness, he inspired me to stay the course. This thesis, therefore, belongs to him. I am also very grateful to his wife for allowing me to call their residence anytime I wanted.

I offer my very special thanks to Dr. Stefan Van Aelst (University of Ghent, Belgium), my co-supervisor, for his invaluable suggestions and guidance throughout this study, and his sponsorship for my stay in Belgium. I am very grateful for the hospitality I was offered by Stefan and his wife during that time.

My sincere thanks go to Dr. Matias Salibián-Barrera, member of my supervisory committee, for his effective suggestions that helped me save many days of work. I am very grateful to him for allowing me to use some of his user-friendly computer codes.

Respectful thanks are to Dr. Harry Joe, Dr. Raymond Ng (computer science, UBC), and Dr. Roy E. Welsch (MIT) for taking their time to read my thesis and provide useful suggestions and comments.

I sincerely thank Christine Graham, Rhoda Morgan, Elaine Salameh and Peggy Ng for their help with administrative matters. Christine and Rhoda are no longer with this department, but they will be remembered forever.

My heartfelt thanks are for my dear friends Mike ("The Man from St. Petersburg"), Jean-François (we all know him as Jeff), Lawrence, Marin, Justin and other graduate students for their invaluable support during my journey to PhD. I benefitted from helpful discussions with Mike and Jean-François, and from their brilliant suggestions. (Mike also suggested how to acknowledge his suggestions!) They also made my life in Vancouver such a pleasant experience!

Finally, I would like to pay tribute to robustness with a poem that I wrote during my graduate studies:

Ashes to ashes, dust to dust –
The statistics I use must be robust!

JAFAR AHMED KHAN

The University of British Columbia

December 2006

To Reba, Roja and Aumi

Chapter 1

Introduction

1.1 Motivation

We consider the problem of building a linear prediction model when there is a large number d of candidate covariates. Large datasets usually contain a fraction of outliers and other contaminations, which are difficult to visualize and clean. The classical algorithms are much affected by these outliers and, therefore, these algorithms often fail to select the 'correct' linear prediction model that would have been chosen if there were no outliers.

We argue that it is not reasonable to attempt to predict future outliers without knowledge of the underlying mechanism that produces them. Therefore, we aim at predicting the future non-outlying cases by fitting well the *majority* of the data. For this, we need a robust method that is capable of selecting the important variables in the presence of outliers in high-dimensional datasets.

Robust model selection has not received much attention in the robustness literature. Seminal papers that address this issue include Ronchetti (1985) and Ronchetti and Staudte (1994) which introduced robust versions of the selection criteria AIC and C_p, respectively. Yohai (1997) proposed a robust Final Prediction Error (FPE) criterion (for Splus documentation). Ronchetti, Field and Blanchard (1997) proposed robust model selection by cross-validation. Morgenthaler, Welsch and Zenide (2003) constructed a selection technique to simultaneously identify the correct model structure as well as unusual observations. All these robust methods require the fitting of all submodels. One exception is the model selection based on the Wald test (Sommer and Huggins 1996) which requires the computation of estimates from the full model only.

A major drawback of the existing robust model selection methods is that they do not scale up to large dimensions, because fitting a robust model is a nonlinear optimization problem. As the number d of possible predictors increases, the number of submodels (which is $2^d - 1$) increases dramatically, making the computational burden enormous. Also, the methods that require the fitting of only the full model are not suitable, because only a few of the d covariates are typically included in the final model, and the fitting of the full model increases the numerical complexity of the methods unnecessarily.

In this study, we attempt to achieve robustness and computational suitability at the same time. That is, we attempt to develop linear prediction model building strategies that are simultaneously (i) capable of selecting the important covariates in the presence of contaminations, and (ii) scalable to high dimensions. The term "scalable" is used to indicate that the numerical complexity of the statistical methods proposed is "reasonable" (e.g., not exponential).

1.2 Model selection strategy

We consider two different strategies for the selection of a linear prediction model for high-dimensional datasets: (a) one-step model building and (b) two-step model building, which are described below.

1.2.1 One-step model building

Since for large values of d the computational burden of all possible subsets regression is enormous, we turn our focus on step-by-step algorithms like forward selection (FS) and stepwise (SW) procedures (see, for example, Weisberg 1985, Chapter 8) that can stop when certain goals are achieved.

Classical FS or SW procedures yield poor results when the data contain outliers and other contaminations, since they attempt to select the covariates that will fit well all the cases (including the outliers). Therefore, our goal is to develop robust step-by-step algorithms that will select important variables in the presence of outliers, and predict well the future non-outlying cases.

We express the classical FS and SW algorithms in terms of sample means, variances and correlations, and replace these sample quantities by their robust counterparts to obtain robust step-by-step algorithms. Similar ideas have been used for building robust estimators of regression parameters (see, for example, Croux, Van Aelst and Dehon 2003, and the references therein). We also incorporate robust partial F-tests as stopping rules during the implementation of these robust algorithms.

1.2.2 Two-step model building

Our two-step model building procedure is a blend of all possible subsets regression and step-by-step algorithms. All possible subsets regression is expected to select a better model (with respect to predictive power) than any step-by-step algorithm, but its computational burden is extremely high for large values of d. We, therefore, consider applying this procedure on a "reduced set" of covariates. Thus, we consider proceeding in two steps. The first step - which we call *sequencing* - quickly screens out unimportant variables to form a "reduced set" for further consideration. The second step - which we call *segmentation* - carefully examines different subsets of the variables in the reduced set for possible inclusion in the prediction model. These two steps are described below.

Sequencing

The goal of the first step is a drastic reduction of the number of candidate covariates. The input variables are sequenced to form a list such that the good predictors are likely to appear at the beginning of the list. The first m covariates of the list then form the reduced set from which the final prediction model will be obtained.

One strategy for sequencing the candidate covariates is to use one of the several available step-by-step or stagewise algorithms, e.g., Forward Selection (FS), or Forward Stagewise procedure (Stagewise) (see, for example, Hastie, Tibshirani and Friedman 2001, Chapter 10). We focus on the powerful algorithm recently proposed by Efron, Hastie, Johnstone and Tibshirani (2004) called Least Angle Regression (LARS), which is a mathematical solution to the Stagewise problem. LARS is computationally efficient and

has been shown to have clear statistical advantages over other step-by-step and stagewise algorithms.

Since LARS is very sensitive to contamination, our goal is to robustify LARS. We show that LARS can be expressed in terms of the mean vector and covariance matrix of the data, and we replace these classical ingredients of LARS by their robust counterparts to obtain robust LARS. We combine robust LARS algorithm with bootstrap to obtain a more stable and reliable list of covariates.

One important issue in the sequencing step is to determine the appropriate value of the number of covariates, m, for the reduced set. The probability that the reduced set contains all the important variables increases with m. Unfortunately, also the computational cost of the second step, segmentation, increases with m. Therefore, we aim to determine a "reasonable" value of m which is large enough to include most of the important variables but not so large as to make the second step impractical or unfeasible. For this purpose, we introduce a "learning curve" that plots robust R^2 values versus dimension. An appropriate value of m is the dimension corresponding to the point where the curve starts to level off.

Segmentation

When we have a reduced set of m covariates for further consideration, one reasonable approach to reach the final model is to perform all possible subsets regression on this reduced set using an appropriate selection criterion. Again, the classical selection criteria, e.g., Final Prediction Error (FPE), Akaike Information Criterion (AIC), Mallows' C_p, cross-validation (CV) and bootstrap procedures are not resistant to outliers. The robust

AIC procedure (Ronchetti 1985) has certain limitations, which are discussed in this thesis. The robust CV method (Ronchetti, Field and Blanchard 1997) is computationally expensive.

In this study, we propose computationally suitable robust CV and robust bootstrap procedures to evaluate the predictive powers of different subsets of the reduced set of covariates. Our robust bootstrap procedure is similar to the methods proposed by Salibián-Barrera (2000), and Salibián-Barrera and Zamar (2002) to conduct robust statistical inferences in linear regression. Since the performance of robust FPE procedure (Yohai 1997) has not been studied so far, we also evaluate this method in our study.

1.3 Computation of robust correlation matrices

As mentioned earlier, our approach to robustification of FS, SW and LARS consists of expressing these algorithms in terms of the mean vector and the correlation matrix of the data, and then replacing these classical ingredients by their robust counterparts. Therefore, robust estimation of the correlation matrix is a major component of the robust methods we propose in this study.

The computation of robust correlation estimates from a d-dimensional data set is very time-consuming, particularly for large values of d. Even the fast MCD algorithm by Rousseeuw and Van Driessen (1999) is not fast enough for the type of applications we have in mind. Therefore, we use robust correlations derived from a pairwise affine equivariant covariance estimator.

Interestingly, the pairwise approach for robust correlation matrix estimation is not only computationally suitable, it is also more relevant than the d-dimensional approach for robust step-by-step algorithms. Pairwise approach allows us to compute only the required correlations at each step of the algorithm. Since we intend to stop as soon as certain goals are achieved, a pairwise approach saves the computation of the correlations that are not required.

We consider robust correlations derived from a simplified version of the bivariate M-estimator proposed by Maronna (1976). This estimate is computationally efficient, affine equivariant and has a breakdown point of 1/3 in two dimensions.

For very large high-dimensional data, however, we need an even faster robust correlation estimator. Therefore, as a part of our robust LARS procedure, we propose a new correlation estimate called the "adjusted-Windsorized correlation estimate". Unlike two separate univariate Windsorizations for X and Y (see Huber 1981 and Alqallaf 2003), we propose a joint Windsorization with a larger tuning constant c_1 for the points falling in the two quadrants that contain the majority of the data, and a smaller constant c_2 for the points in the two minor quadrants. Our estimate has some advantages over the univariate-Windsorized correlation estimate.

1.4 Organization of subsequent chapters

The following chapters are organized as follows. In Chapter 2, we present the one-step model selection procedure, where we develop robust versions of FS and SW algorithms incorporated with robust partial F-tests used as stopping rules.

In Chapter 3, we present the first step (robust sequencing) of the two-step model selection procedure. Here, we robustify LARS to sequence the covariates, use bootstrap to stabilize the results obtained by robust LARS, and use "learning curves" to decide about the size of the reduced set. For the development of robust LARS, we introduce the idea of "multivariate-windsorization" which we use for robust data cleaning. We also propose a new correlation estimate which we call the "adjusted-Windsorized correlation estimate".

Chapter 4 deals with the second step (robust segmentation) of the two-step model selection procedure. Here, we review the existing classical and robust selection criteria, discuss their limitations, and propose computationally suitable robust CV and robust bootstrap procedures to evaluate the predictive powers of different subsets of the reduced set of covariates. We also evaluate the performance of robust FPE (Yohai 1997).

Chapter 5 studies the properties of the adjusted-Windsorized correlation estimate (the new correlation estimate proposed in Chapter 3). We show that the proposed estimate is consistent and has bounded influence. We obtain its asymptotic variance and intrinsic bias. We show that the tuning constants of this estimate can be chosen such that it is approximately Fisher-consistent. We also show that a smoothed version of this estimate is asymptotically normal.

In Chapter 6 we conclude by summarizing the main ideas proposed in this thesis, and the main results obtained.

Though the major chapters (Chapters 2-5) are connected conceptually, each of them is independent of the others. That is, they may be considered as individual research papers, to a certain extent.

Chapter 2

One-step Model Building: Robust Forward Selection and Stepwise Procedures

2.1 Introduction

When the number d of candidate covariates is small, one can choose a linear prediction model by computing a reasonable criterion (e.g., C_p, AIC, cross-validation error or bootstrap error) for all possible subsets of the predictors. However, as d increases, the computational burden of this approach (sometimes referred to as *all possible subsets regression*) increases very quickly. This is one of the main reasons why step-by-step algorithms like forward selection (FS) and stepwise (SW) (see, for example, Weisberg 1985, Chapter 8) are popular.

Unfortunately, classical FS or SW procedures yield poor results when the data contain outliers and other contaminations. These algorithms attempt to select the covariates that will fit well all the cases (including the outliers), and often fail to select the model that would have been chosen if those outliers were not present in the data. Moreover, aggressive deletion of outliers is not desirable, because we may end up deleting a lot of observations which are outliers only with respect to the predictors that will not be in the model.

We argued earlier that it is not reasonable to attempt to predict the future outliers without knowledge of the underlying mechanism that produces them. Therefore, our goal is to develop robust step-by-step algorithms that will select important variables in the presence of outliers, and predict well the future non-outlying cases.

We show that the list of variables selected by classical FS and SW procedures are functions of sample means, variances and correlations. We express the two classical algorithms in terms of these sample quantities, and replace them by robust counterparts to obtain the corresponding robust versions of the algorithms. Once the covariates are selected (by using these simple robust selection algorithms), we can use a robust regression estimator on the final model.

Robust correlation matrix estimators for d-dimensional data sets are usually derived from affine-equivariant, robust estimators of scatter. This is very time-consuming, particularly for large values of d. Moreover, the computation of such robust correlation matrices becomes unstable when the dimension d is large compared to the sample size n. On the other hand, only a few of the d covariates are typically included in the final model, and the computation of the whole d-dimensional correlation matrix at once

will unnecessarily increase the numerical complexity of the otherwise computationally suitable step-by-step algorithms.

To avoid this complexity, we use an affine-equivariant bivariate M-estimator of scatter to obtain robust correlation estimates for all pairs of variables, and combine these to construct a robust correlation matrix. We call this the pairwise robust correlation approach. Interestingly, this pairwise approach for robust correlation matrix estimation is not only computationally suitable, but is also more convenient (compared to the full d-dimensional approach) for robust step-by-step algorithms. The reason is as follows. The sample correlation matrix (R, say) has the property that the correlation matrix of a subset of variables can be obtained by simply taking the appropriate submatrix of R. This property allows us to compute only the required correlations at each step of the algorithm. With the pairwise robust correlation approach we keep this property.

Affine equivariance and regression equivariance are considered to be important properties for robust regression estimators (see, e.g., Rousseeuw and Leroy 1987). However, these properties are not required in the context of variable selection, because we do not consider linear combinations of the existing covariates. The only transformations that should not affect the selection result are linear transformations of individual variables, i.e., shifts and scale changes. Variable selection methods are often based on correlations among the variables. Therefore, robust variable selection procedures need to be robust against correlation outliers, that is, outliers that affect the classical correlation estimates but can not be detected by looking at the individual variables separately. Our approach based on pairwise correlations is robust against correlation outliers and thus suitable for robust variable selection. It should be emphasized that with our approach we consider the problem of "selecting" a list of important predictors, but we do not yet "fit" the

selected model. The final model resulting from the selection procedure usually contains only a small number of predictors compared to the initial dimension d, when d is large. Therefore, to robstly fit the final model we propose to use a highly robust regression estimator such as an MM-estimator (Yohai 1987) that is resistant to all types of outliers. Note that we always use models with intercept.

Croux, Van Aelst and Dehon (2003) estimated the parameters of a regression model using S-estimators of multivariate location and scatter. They also obtained the corresponding standard errors. Their estimation method can be adapted for model-building purposes. However, for the step-by-step algorithms like FS and SW, our pairwise approach has computational advantages.

The rest of this chapter is organized as follows. In Section 2.2 we review some classical step-by-step algorithms. In Section 2.3 we decompose the FS and SW procedures in terms of the correlation matrix of the data. In Section 2.4, we present robust versions of these algorihms, along with their numerical complexities. Section 2.5 presents a Monte Carlo study that compares our robust methods with the classical ones by their predicting powers. Section 2.6 contains two real-data applications. Section 2.7 is the conclusion.

2.2 Review: classical step-by-step algorithms

In this section, we review the three most important step-by-step algorithms: Forward Selection (FS), stepwise (SW) and Backward Elimination (BE). We show a serious drawback of the BE procedure, which is why we did not consider this algorithm for robustification.

2.2.1 Forward Selection (FS)

Let us have d predictors $X_1, \ldots, X_d$, and a response Y. Let each variable be standardized using its mean and standard deviation. The FS procedure selects the predictor (X_1, say) that has the largest absolute correlation $|r_{1Y}|$ with Y, and obtains the residual vector $Y - r_{1Y}X_1$. All the other covariates are then 'adjusted for X_1' and entered into competition. That is, each X_j is regressed on X_1, and the corresponding residual vector $Z_{j.1}$ (which is orthogonal to X_1) is obtained. The correlations of these $Z_{j.1}$ with the residual vector $Y - r_{1Y}X_1$, which are also called "the partial correlations between X_j and Y adjusted for X_1", decide the next variable (X_2, say) to enter the regression model. All the other covariates are then 'adjusted for X_1 and X_2' and entered into further competition, and so on. We continue adding one covariate at each step, until a stopping criterion is met.

The reason behind the 'orthogonalization', that is, the construction of $Z_{j.1}$ from X_j, is that the algorithm measures what 'additional' contribution X_j makes in explaining the variability of Y, when X_j joins X_1 in the regression model. The R^2 produced by (X_1, Z_2) is the same as the R^2 produced by (X_1, X_2), and the orthogonalization ensures maximum R^2 at each FS step.

We stop when the partial F-value for each covariate that has not yet entered the model is less than a pre-selected number, say "F-IN" (Weisberg 1985).

2.2.2 Stepwise (SW)

The SW algorithm is the same as the FS procedure up to the second step. When there are at least two covariates in the model, at each subsequent SW step we either (a) remove a covariate, or (b) exchange two covariates, or (c) add a covariate, or (d) stop. Note that, the "exchange" of two covariates is not the same as the addition (removal) of a covariate in one step followed by the removal (addition) of another covariate in the next step. Sometimes, a new covariate cannot enter the model because of an existing covariate, and the existing covariate cannot be removed according to the criterion used.

The options at each SW step are considered in the order in which they are mentioned above. A selected covariate is removed if its partial F-value is less than a pre-selected number, say "F-OUT" (Weisberg 1985). A selected covariate is exchanged with a new one if the exchange increases R^2. A covariate is added if it has the highest partial F-value among the remaining covariates, and the value is more than F-IN (as in FS). And, we stop when none of the above (removal, exchange or addition) occurs in a certain step.

2.2.3 Backward Elimination (BE)

The BE procedure (see, for example, Weisberg 1985, Chapter 8) is the opposite of FS. BE starts with the full model, and removes one covariate at each step. The covariate to remove is the one that has the smallest partial F-value among all the covariates that are currently in the model. We stop when the partial F-value for each covariate currently in the model is greater than F-OUT.

Limitation of BE

When the number d of candidate covariates is large, only a few of these covariates are typically included in the final model. However, to apply the BE algorithm, we have to compute the pairwise correlation estimates for all the d covariates (since BE starts with the full model). Therefore, BE has higher numerical complexity than that of FS (or SW). This problem will be more serious with the computation of robust correlation estimates (for robust BE). Therefore, we will not consider the BE procedure for robustification.

2.3 FS and SW Expressed in Correlations

In order to robustify the FS and SW procedures, we will now express these algorithm in terms of the original correlations of the variables.

2.3.1 FS expressed in terms of correlations

Let the d covariates $X_1, \ldots, X_d$ and the response Y be standardized using their mean and standard deviation. Let r_{jY} denote the correlation between X_j and Y, and R_X be the correlation matrix of the covariates. Suppose w.l.o.g. that X_1 has the maximum absolute correlation with Y. Then, X_1 is the first variable that enters the regression model. We call the predictors that are in the current regression model "active" predictors. The remaining candidate predictors are called "inactive" predictors. We now need the partial correlations between X_j $(j \neq 1)$ and Y adjusted for X_1, denoted by $r_{jY.1}$, to determine the second covariate X_2 (say) that enters the model.

The partial correlation $r_{jY.1}$ expressed in terms of original correlations

Each inactive covariate X_j should be regressed on X_1 to obtain the residual vector $Z_{j.1}$ as follows

$$Z_{j.1} = X_j - \beta_{j1} X_1, \tag{2.1}$$

where

$$\beta_{j1} = \frac{1}{n} X_1^t X_j = r_{j1}. \tag{2.2}$$

We have

$$\frac{1}{n} Z_{j.1}^t Y = (X_j - \beta_{j1} X_1)^t Y = r_{jY} - r_{j1} r_{1Y}, \tag{2.3}$$

and

$$\frac{1}{n} Z_{j.1}^t Z_{j.1} = (X_j - \beta_{j1} X_1)^t (X_j - \beta_{j1} X_1) = 1 - r_{j1}^2. \tag{2.4}$$

Therefore, the partial correlation $r_{jY.1}$ is given by

$$r_{jY.1} = \frac{Z_{j.1}^t (Y - \beta_{Y1} X_1)/n}{\sqrt{Z_{j.1}^t Z_{j.1}/n}\ \mathrm{SD}(Y - \beta_{Y1} X_1)} \tag{2.5}$$

Note that the factor $\mathrm{SD}(Y - \beta_{Y1} X_1)$ in the denominator of (2.5) is independent of the covariate X_j; $(j = 2, \ldots, d)$ being considered. Hence, when selecting the covariate X_j that maximizes the partial correlation $r_{jY.1}$, this constant factor can be ignored. This reduces computations and therefore is more time efficient. It thus suffices to calculate

$$\tilde{r}_{jY.1} = \frac{Z_{j.1}^t (Y - \beta_{Y1} X_1)/n}{\sqrt{Z_{j.1}^t Z_{j.1}/n}} \tag{2.6}$$

which is proportional to the actual partial correlation. $\tilde{r}_{jY.1}$ can be rewritten as follows

$$\begin{aligned} \tilde{r}_{jY.1} &= \frac{Z_{j.1}^t Y/n}{\sqrt{Z_{j.1}^t Z_{j.1}/n}} \quad \text{[since } Z_{j.1} \text{ and } X_1 \text{ are orthogonal]} \\ &= \frac{r_{jY} - r_{j1} r_{1Y}}{\sqrt{1 - r_{j1}^2}} \quad \text{[using (2.3) and (2.4)].} \end{aligned} \tag{2.7}$$

Now, suppose w.l.o.g. that X_2 (or, equivalently, $Z_{2.1}$) is the new active covariate, because it minimizes $\tilde{r}_{jY.1}$ (and thus also the partial correlation $r_{jY.1}$). All the inactive covariates should now be orthogonalized with respect to $Z_{2.1}$.

Orthogonalization of $Z_{j.1}$ wrt $Z_{2.1}$

Each inactive variable $Z_{j.1}$ should be regressed on $Z_{2.1}$ to obtain the residual vector $Z_{j.12}$ as follows

$$Z_{j.12} = Z_{j.1} - \beta_{j2.1} Z_{2.1}.$$

Here,

$$\begin{aligned}
\beta_{j2.1} &= \frac{Z_{2.1}^t Z_{j.1}/n}{Z_{2.1}^t Z_{2.1}/n} \\
&= \frac{X_2^t Z_{j.1}/n}{Z_{2.1}^t Z_{2.1}/n} \quad \text{[because of orthogonality]} \qquad (2.8) \\
&= \frac{X_2^t (X_j - r_{j1} X_1)/n}{Z_{2.1}^t Z_{2.1}/n} \quad \text{[Using (2.1) and (2.2)]} \\
&= \frac{r_{2j} - r_{21} r_{j1}}{1 - r_{21}^2} \quad \text{[using (squared) denominator of (2.7) for } j = 2\text{]}.
\end{aligned}$$

Thus, $\tilde{r}_{jY.1}$ and $\beta_{j2.1}$ are expressed in terms of original correlations.

Lemma 2.1. *Given that the numerators and denominators of the following equations*

$$\tilde{r}_{jY.12\cdots(k-1)} = \frac{Z_{j.12\cdots(k-1)}^t Y/n}{\sqrt{Z_{j.12\cdots(k-1)}^t Z_{j.12\cdots(k-1)}/n}}, \quad \textit{for all inactive } j, \qquad (2.9)$$

and

$$\beta_{jh.12\cdots(h-1)} = \frac{Z_{h.12\cdots(h-1)}^t Z_{j.12\cdots(h-1)}/n}{Z_{h.12\cdots(h-1)}^t Z_{h.12\cdots(h-1)}/n}, \quad \textit{for } h = 2, \ldots, k;\ j \textit{ inactive}, \qquad (2.10)$$

are functions of original correlations, the numerators and denominators of the following quantities can be expressed as functions of original correlations: (a) $\tilde{r}_{jY.12\cdots k}$ and (b) $\beta_{j(k+1).12\cdots k}$.

Proof. Here, $\tilde{r}_{jY.12\cdots(k-1)}$ determines the next active covariate X_k (or, equivalently, $Z_{k.12\cdots(k-1)}$). The given $\beta_{jk.12\cdots(k-1)}$ can be used to obtain the residual vector

$$Z_{j.12\cdots k} = Z_{j.12\cdots(k-1)} - \beta_{jk.12\cdots(k-1)} Z_{k.12\cdots(k-1)}. \tag{2.11}$$

Now,

$$\tilde{r}_{jY.12\cdots k} = \frac{Z^t_{j.12\cdots k} Y/n}{\sqrt{Z^t_{j.12\cdots k} Z_{j.12\cdots k}/n}}. \tag{2.12}$$

Using (2.11), the numerator of (2.12) can be written as

$$\frac{1}{n} Z^t_{j.12\cdots(k-1)} Y - \beta_{jk.12\cdots(k-1)} \frac{1}{n} Z^t_{k.12\cdots(k-1)} Y,$$

where the first part is the numerator of (2.9), $\beta_{jk.12\cdots(k-1)}$ comes from (2.10) for $h = k$, and the rest is the numerator of (2.9) for $j = k$ (because X_k was inactive at that point). Thus, the numerator of (2.12) is a function of the original correlations.

Using (2.11), the squared denominator of (2.12) can be written as

$$\frac{1}{n} Z^t_{j.12\cdots(k-1)} Z_{j.12\cdots(k-1)} - 2\beta_{jk.12\cdots(k-1)} \frac{1}{n} Z^t_{k.12\cdots(k-1)} Z_{j.12\cdots(k-1)} + \beta^2_{jk.12\cdots(k-1)} Z^t_{k.12\cdots(k-1)} Z_{k.12\cdots(k-1)}, \tag{2.13}$$

where, the first term is the (squared) denominator of (2.9), the second term is the numerator of (2.10) for $h = k$, and the last term is the denominator of (2.10) for $h = k$. This proves Part (a) of the lemma.

The quantities $\tilde{r}_{jY.12\cdots k}$ determine the next active covriate $X_{(k+1)}$. Now,

$$\beta_{j(k+1).12\cdots k} = \frac{Z^t_{(k+1).12\cdots k} Z_{j.12\cdots k}/n}{Z^t_{(k+1).12\cdots k} Z_{(k+1).12\cdots k}/n}. \tag{2.14}$$

Because of orthogonality, the numerator of (2.14) can be written as:

$$\begin{aligned}\frac{1}{n}X^t_{(k+1)}Z_{j.12\cdots k} &= \frac{1}{n}X^t_{(k+1)}(X_j - \beta_{j1}X_1 - \beta_{j2.1}Z_{2.1} - \ldots \beta_{jk.12\cdots(k-1)}Z_{k.12\cdots(k-1)}) \\ &= r_{j(k+1)} - \beta_{j1}\frac{1}{n}X^t_{(k+1)}X_1 - \beta_{j2.1}\frac{1}{n}X^t_{(k+1)}Z_{2.1} \ldots - \beta_{jk.12\cdots(k-1)}\frac{1}{n}X^t_{(k+1)}Z_{k.12\cdots(k-1)} \\ &= r_{j(k+1)} - r_{j1}r_{k1} - \beta_{j2.1}\frac{1}{n}Z^t_{(k+1).1}Z_{2.1} \cdots - \beta_{jk.12\cdots(k-1)}\frac{1}{n}Z^t_{(k+1).12\cdots(k-1)}Z_{k.12\cdots(k-1)},\end{aligned}$$

where the β's come from (2.10) for $h = 2, \ldots, k$, and the other quantities are the numerators of (2.10) for $j = k+1$, and $h = 2, \ldots, k$.

For the denominator of (2.14), we can use the relation

$$Z_{(k+1).12\cdots k} = Z_{(k+1).12\cdots(k-1)} - \beta_{(k+1)k.12\cdots(k-1)}Z_{k.12\cdots(k-1)},$$

which follows from (2.11) by replacing $j = (k+1)$. So, the denominator can be written as

$$\begin{aligned}Z^t_{(k+1).12\cdots(k-1)}Z_{(k+1).12\cdots(k-1)}/n - 2\beta_{(k+1)k.12\cdots(k-1)}Z^t_{(k+1).12\cdots(k-1)}Z_{k.12\cdots(k-1)}/n \\ +\beta^2_{(k+1)k.12\cdots(k-1)}Z^t_{k.12\cdots(k-1)}Z_{k.12\cdots(k-1)}/n,\end{aligned}$$

where the first part is the (squared) denominator of (2.9) for $j = k+1$, the second part is the numerator of (2.10) for $j = k+1$ and $h = k$, and the last part is the denominator of (2.10) for $h = k$. Therefore, $\beta_{j(k+1).12\ldots k}$ is a function of original correlations. This completes the proof. ■

FS steps in correlations

We can now summarize the FS algorithm in terms of correlations among the original variables as follows:

1. To select the first covariate X_{m_1}, determine $m_1 = \text{argmax}\,|r_j|$.

2. To select the kth covariate X_{m_k} ($k = 2, 3, \ldots$), calculate $\tilde{r}_{jY.m_1 \cdots m_{(k-1)}}$, which is proportional to the partial correlation between X_j and Y adjusted for $X_{m_1}, \cdots, X_{m_{(k-1)}}$, and then determine $m_k = \text{argmax}\,|\tilde{r}_{jY.m_1 \cdots m_{(k-1)}}|$.

Partial F-tests for stopping

At each FS step, once the "best" covariate (among the remaining covariates) is identified, we can perform a partial F-test to decide whether to include this covariate in the model (and continue the process) or to stop. The new "best" covariate enters the model only if the partial F-value, denoted by F_{partial}, is greater than $F(0.95, 1, n-k-1)$ (say), where k is the current size of the model including the new covariate. Here again, the required quantities can be expressed in terms of correlations among the original variables, as we show below.

Suppose that X_1 is already included in the model, and X_2 has the largest absolute partial correlation with Y after adjusting for X_1. To decide whether X_2 should be included in the model we perform a partial F-test using the statistic F_{partial} given by

$$
\begin{aligned}
F_{\text{partial}} &= \frac{(Y-\beta_{Y1}X_1)^t(Y-\beta_{Y1}X_1)-(Y-\beta_{Y1}X_1-\beta_{Y2.1}Z_{2.1})^t(Y-\beta_{Y1}X_1-\beta_{Y2.1}Z_{2.1})}{(Y-\beta_{Y1}X_1-\beta_{Y2.1}Z_{2.1})^t(Y-\beta_{Y1}X_1-\beta_{Y2.1}Z_{2.1})/(n-3)} \\
&= \frac{(n-3)\,(2\,\beta_{Y2.1}Z_{2.1}^tY/n-\beta_{Y2.1}^2Z_{2.1}^tZ_{2.1}/n)}{1-r_{1Y}^2-(2\,\beta_{Y2.1}Z_{2.1}^tY/n-\beta_{Y2.1}^2Z_{2.1}^tZ_{2.1}/n)} \\
&= \frac{(n-3)\,(\beta_{Y2.1}Z_{2.1}^tY/n)}{1-r_{1Y}^2-\beta_{Y2.1}Z_{2.1}^tY/n} \\
&= \frac{(n-3)\,\tilde{r}_{2Y.1}^2}{1-r_{1Y}^2-\tilde{r}_{2Y.1}^2}, \qquad (2.15)
\end{aligned}
$$

where $\tilde{r}_{2Y.1}$ is expressed in correlations in (2.7).

Similarly, when $(k-1)$ covariates $X_1,\ldots,X_{k-1}$ are already in the model, and w.l.o.g. X_k has the largest absolute partial correlation with Y after adjusting for $X_1,\ldots,X_{k-1}$, the partial F-statistic for X_k can be expressed as:

$$
F_{\text{partial}} = \frac{(n-k-1)\,\tilde{r}_{kY.12\cdots(k-1)}^2}{1-r_{1Y}^2-\tilde{r}_{2Y.1}^2-\cdots-\tilde{r}_{kY.12\cdots(k-1)}^2}, \qquad (2.16)
$$

where the partial correlations can be expressed in terms of the original correlations using Lemma 2.1.

2.3.2 SW expressed in terms of correlations

The SW algorithm starts as the FS procedure. When there are at least two covariates in the model, at each subsequent SW step we either add a covariate, or drop a covariate, or exchange two covariates, or stop.

To decide whether to add a covariate, the partial correlations of each inactive covariate X_j with Y can be computed as in the case of FS (see (2.12)) to perform a partial F-test (see (2.15) and (2.16)). To decide whether to drop an "active" covariate, we can pretend that the active covariate under consideration entered the model last, and calculate its partial correlations with Y (see (2.12), subscripts modified) to perform a partial F-test (see (2.15) and (2.16), subscripts modified).

Once an "active" covariate is dropped, the "orthogonalizations" of the other covariates (active or inactive) with this covariate that were used before to derive the partial correlations become irrelevant, and the order of the other active covariates in the model cannot be determined. Fortunately, this does not create a problem to decide the next covariate, because, for example, $\tilde{r}_{jY.346} = \tilde{r}_{jY.643}$. Therefore, we can update all relevant calculations considering the currently active covariates in any order.

Stopping criteria for SW. Unlike the FS algorithm where a stopping criterion is "optional" (we may choose to sequence all the covariates), SW has to have a built-in stopping rule, because at each step we have to decide whether to add one covariate and/or drop another. We may choose two different theoretical F percentiles as the inclusion and deletion criteria, e.g., $F(0.95, 1, n-k_1-1)$ and $F(0.90, 1, n-k_2-1)$, respectively, where k_1 and k_2 are the model sizes after inclusion and before deletion.

2.4 Robustification of FS and SW algorithms

In the last section we expressed the FS and SW algorithms in terms of sample means, variances and correlations. Because of these non-robust building blocks, these algorithms

are sensitive to contamination in the data, as shown by our simulation and real-data examples later on. A simple robustification of these algorithms can be achieved by replacing the non-robust ingredients of the algorithms by their robust counterparts. For the initial standardization, the choices of fast computable robust center and scale measures are straightforward: median (med) and median absolute deviation (mad). As mentioned earlier, most available robust correlation estimators are computed from the d-dimensional data and therefore are very time consuming (see, for example, Rousseeuw and Leroy 1987). Robust pairwise approaches (Huber 1981) are not affine equivariant and, therefore, are sensitive to two-dimensional outliers.

One solution is to use robust correlations derived from a pairwise affine equivariant covariance estimate. We consider an estimate which is inspired by the computationally suitable multivariate M-estimate proposed by Maronna (1976). We first present Maronna's estimate below.

Definition 2.1. (Maronna's M-estimate of multivariate location and scatter) *Let us have n multivariate observations $\boldsymbol{z}_i$, $i = 1, \ldots, n$. Maronna's M-estimate of the location vector $\boldsymbol{t}$ and scatter matrix $\boldsymbol{V}$ is defined as the solution of the system of equations:*

$$\frac{1}{n}\sum_i u_1(d_i)(\boldsymbol{z}_i - \boldsymbol{t}) = \boldsymbol{0}, \tag{2.17}$$

$$\frac{1}{n}\sum_i u_2(d_i^2)(\boldsymbol{z}_i - \boldsymbol{t})(\boldsymbol{z}_i - \boldsymbol{t})' = V, \tag{2.18}$$

where $d_i^2 = (\boldsymbol{z}_i - \boldsymbol{t})'V^{-1}(\boldsymbol{z}_i - \boldsymbol{t})$, and u_1 and u_2 satisfy a set of general assumptions.

For further computational ease, we considered the following simplified version of the bivariate M-estimate. We used the coordinatewise median as the bivariate location estimate and only solved (2.18) to estimate the scatter matrix and hence the correlation.

We used the function $u_2(t) = u(t) = \min(c/t, 1)$ with $c = 9.21$, the 99% quantile of a χ^2_2 distribution. For the bivariate observations $\boldsymbol{z}_i = (x_i, y_i)$, $i = 1, \ldots, n$, the steps for calculating this correlation estimate are presented below:

1. Calculate the medians m_X and m_Y, and obtain $\tilde{\boldsymbol{z}}_i = (\tilde{x}_i, \tilde{y}_i)$, $i = 1, \ldots, n$, where $\tilde{x}_i = x_i - m_X$, and $\tilde{y}_i = y_i - m_Y$.
2. Calculate the mads s_X and s_Y, and set $V_0 = \begin{pmatrix} s_X & 0 \\ 0 & s_Y \end{pmatrix}$.
3. Calculate $d_i^2 = \tilde{\boldsymbol{z}}_i' V_0^{-1} \tilde{\boldsymbol{z}}_i$, $i = 1, \ldots, n$. Then obtain $V_1 = \frac{1}{n} \sum_{i=1}^{n} u(d_i^2)\, \tilde{\boldsymbol{z}}_i\, \tilde{\boldsymbol{z}}_i'$.
4. Set $V_0 \leftarrow V_1$.
5. Repeat steps 3 and 4.

We stop when $|r(V_1) - r(V_0)| < \delta$, where $\delta > 0$ is a pre-selected small number, and $r(.)$ is the correlation coefficient calculated from the bivariate scatter matrix.

Finally, FS and SW algorithms are implemented using these robust pairwise correlations.

Robust partial F-tests. We replace the classical correlations in the partial F statistic by their robust counterparts to form a robust partial F statistic. We conjecture that the robust pairwise correlations appearing in the numerator of the F statistic are jointly normal. Therefore, under the null hypothesis, the robust F statistic is asymptotically distributed as χ^2_1. To assess our conjecture numerically, we conducted the following simulation. We generated X_1, ϵ_1 and ϵ_2 from a standard normal distribution. We then generated $Y = \beta_0 + \beta_1 X_1 + \sigma_1 \epsilon_1$, and $X_2 = \gamma_0 + \gamma_1 X_1 + \sigma_2 \epsilon_2$, where β_0, β_1, γ_0 and γ_1

are generated from a uniform distribution on $(-10, 10)$, σ_1 is chosen so that the signal-to-noise ratio equals 2, and σ_2 is chosen so that X_1 and X_2 have a particular correlation randomly chosen from a uniform distribution on $(0, 1)$. We generated 2000 datasets of size 100, and calculated the robust partial F statistic for covariate X_2 in each case. Figure 2.1 shows the qqplot of the robust partial F values against the theoretical χ^2_1 quantiles. Moreover, the average and variance of the robust F values are 0.99 ($\simeq 1$) and 2.02 ($\simeq 2$), respectively. All of these support our conjecture. Therefore, we consider it to be reasonable to use theretical F quantiles as our stopping criteria for robust FS and SW algorithms.

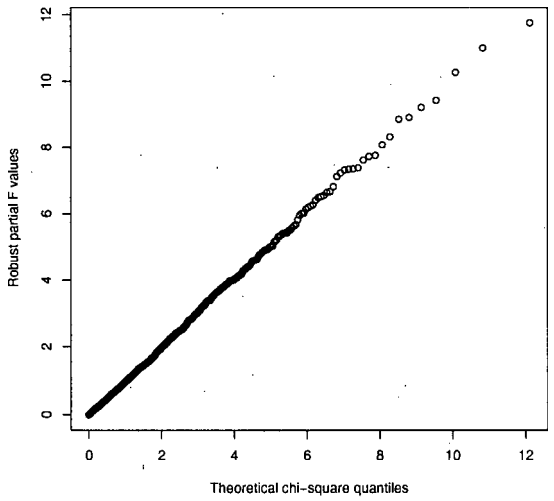


Figure 2.1: QQplot of the robust partial F values against the theoretical χ^2_1 quantiles.

2.4.1 Numerical complexity of the algorithms

If we sequence all d covariates, the standard FS procedure requires $\mathcal{O}(nd^2)$ time. However, when applied with a stopping criterion, the complexity of FS depends on the number of

covariates selected in the model. Assuming that the model size will not exceed a certain number $m < d$, the complexity of FS is less than or equal to $\mathcal{O}(ndm)$. Similarly, the maximum complexity of SW is $\mathcal{O}(n(dm + m^2)) = \mathcal{O}(ndm)$.

Since we used the coordinatewise median as the bivariate location estimate, the correlation based on Maronna's M-estimate can be computed in $\mathcal{O}(n \log n + bn)$ time, where b is the number of iterations required. Assuming that b does not exceed $\mathcal{O}(\log n)$ (convergence was achieved after 3 to 5 iterations in our simulations), the complexity of this estimate is $\mathcal{O}(n \log n)$. As a result, the maximum complexity of robust FS is $\mathcal{O}((n \log n)dm)$, and the maximum complexity of robust SW is $\mathcal{O}((n \log n)(dm + m^2)) = \mathcal{O}((n \log n)dm)$.

Though *all possible subsets regression* is expected to select a better model (with respect to predictive power) than any step-by-step algorithm, its computational burden is extremely high for large values of d, since it requires the fitting of all $2^d - 1$ submodels. The complexity of the classical algorithms of this type is $\mathcal{O}(2^d nd^2)$. Since robust model selection methods proposed so far uses *all possible subsets regression*, the complexity of the existing robust algorithms is $\mathcal{O}(2^d nd^2)$ multiplied by the number of iterations required for the robust fits.

2.4.2 Limitation of the proposed algorithms

The robust FS and SW procedures based on robust pairwise correlations proposed are resistant to bivariate (correlation) outliers. However, they can be sensitive to three- or higher-dimensional outliers, that is, outliers that are not detected by univariate and

bivariate analyses. Also, the correlation matrix obtained from the pairwise correlation approach may not be positive definite, forcing the use of correction for positive definiteness in some cases (see, e.g., Alqallaf et al. 2002).

It should be emphasized here that these are very small prices to pay to make the selection of covariates possible for large values of d. For example, in our simulations (presented later) we used $d = 50$. It is impossible to apply *all possible subsets regression* on a dataset of this dimension. If one robust fit takes 0.001 cpu second, we would need $2^{50} * 0.001/(3600 * 24 * 365)$ years to select the final model.

2.5 A simulation study

To compare our robust methods with the classical ones, we carried out a simulation study similar to Frank and Friedman (1993). The total number of variables is $d = 50$. A small number $a = 9$ or $a = 15$ of them are nonzero covariates. We considered 2 correlation structures of these nonzero covariates: "moderate correlation" case and "no correlation" case, which are described below.

For the moderate-correlation case, we considered 3 independent 'unknown' processes, represented by latent variables L_i, $i = 1, 2, 3$, which are responsible for the systematic variation of both the response and the covariates. The model is

$$Y = 7L_1 + 6L_2 + 5L_3 + \epsilon = \text{Signal} + \epsilon, \tag{2.19}$$

where $L_i \sim N(0, 1)$, and ϵ is a normal error not related to the latent variables. The variance of ϵ is chosen such that the signal-to-noise ratio equals 2, that is $\text{Var}(\epsilon) = 110/4$.

The nonzero covariates are divided in 3 equal groups, with each group related to exactly one of the latent variables by the following relation

$$X_j = L_i + \delta_j,$$

where $\delta_j \sim N(0,1)$. Thus, we have a true correlation of 0.5 between the covariates generated with the same latent variable.

For the no-correlation case (a true correlation of 0 between the covariates), independent predictors $X_j \sim N(0,1)$ are considered, and Y is generated using the a non-zero covariates, with coefficients $(7,6,5)$ repeated three times for $a = 9$, and five times for $a = 15$.

For each case we generated 1000 datasets each of which was randomly divided into a training sample of size 100 and a test sample of size 100.

Contamination of the training data. Each of the $d - a$ noise variables are contaminated independently. Each observation of a noise variables is assigned probability 0.003 of being replaced by a large number. If this observation is contaminated, then the corresponding observation of Y is also replaced by a large number. Thus, the probability that any particular row of the training sample will be contaminated is $1-(1-0.003)^{d-a}$, which is approximately 10% for $a = 15$, and 11.6% for $a = 9$.

For each of the 4 methods (2 classical and 2 robust), we fitted the obtained model on the training data, and then used it to predict the test data outcomes. We used MM-estimator (Yohai 1987) to fit the models obtained by either of the robust methods, because of its high breakdown point and high efficiency at the normal model. For each simulated dataset, we recorded (1) the average squared prediction error on the test sample, (2) the

number of noise variables selected in the model, and (3) the total number of variables selected in the model.

Table 2.1: Performance of the classical and robust methods in clean and contaminated data for moderate-correlation case. The average (SD) of mean squared prediction error (MSPE) on the test set and the number of noise variables (Noise) selected are shown.

		$a = 9$		$a = 15$	
Data	Method	MSPE	Noise	MSPE	Noise
Clean	FS	59.7 (12.0)	4.9 (2.4)	50.2 (9.3)	4.3 (2.2)
	SW	60.3 (12.3)	4.8 (2.3)	51.2 (9.7)	4.2 (2.1)
	Rob FS	60.4 (12.2)	5.1 (2.6)	51.5 (10.3)	4.7 (2.5)
	Rob SW	61.1 (12.8)	5.0 (2.5)	52.8 (10.5)	4.6 (2.4)
Contam	FS	157.6 (40.8)	13.6 (3.1)	134.5 (32.9)	11.7 (2.9)
	SW	158.4 (41.3)	13.4 (3.0)	136.3 (33.3)	11.6 (2.8)
	Rob FS	94.9 (27.9)	2.5 (2.9)	78.9 (23.7)	1.6 (2.9)
	Rob SW	95.1 (27.8)	2.4 (2.8)	79.3 (23.4)	1.5 (2.6)

Table 2.1 shows the average (sd) of the first two quantities mentioned above over all generated datasets for the moderate-correlation case. The average (sd) of the third quantity (total number of variables) is similar for all the methods in the clean data. However, for the contaminated data, the average increases (decreases) for the classical (robust) methods. For example, for $a = 15$, the average for the classical methods increases from 13 to 17 (approximately), while for the robust methods it decreases from 13 to 6.

In general, FS performs as good as SW, and robust FS performs as good as robust

SW. For the clean data, the performance of robust FS (SW) is comparable to standard FS (SW). For the contaminated data, the test errors produced by robust methods are much smaller than the classical ones. Also, the models obtained by robust methods contain fewer noise variables than the classical ones.

Table 2.2: Performance of the classical and robust methods in clean and contaminated data for no-correlation case. The average (SD) of mean squared prediction error (MSPE) on the test set and the average number of noise variables (Noise) selected are shown.

		$a = 9$		$a = 15$	
Data	Method	MSPE	Noise	MSPE	Noise
Clean	FS	55.6 (11.6)	5.0 (2.4)	107.0 (21.7)	4.6 (2.3)
	SW	55.8 (11.8)	4.8 (2.3)	108.1 (22.1)	4.3 (2.1)
	Rob FS	56.5 (12.4)	5.1 (2.6)	109.9 (21.6)	4.8 (2.4)
	Rob SW	56.7 (12.8)	4.9 (2.5)	108.4 (22.4)	4.6 (2.3)
Contam	FS	161.8 (38.1)	13.6 (3.0)	296.7 (75.3)	11.9 (2.8)
	SW	162.5 (37.5)	13.4 (2.8)	297.9 (75.9)	11.7 (2.7)
	Rob FS	72.5 (13.9)	2.1 (2.4)	124.1 (19.9)	1.2 (1.8)
	Rob SW	72.6 (13.8)	2.1 (2.3)	124.2 (20.8)	1.2 (1.7)

Table 2.2 presents the results for the no-correlation case. Here, robust FS and SW more drastically outperform the standard FS and SW, as compared to the moderate-correlation case. Note that the errors presented in this table are not comparable to those of Table 2.1 since Y is generated using the non-zero covariates ($a = 9$ or $a = 15$), instead of the 3 latent variables. Thus, Y has much more variability for $a = 15$ than for $a = 9$.

2.5.1 Model selection with Spearman's ρ and Kendall's τ

In Section 2.3 we expressed the classical FS and SW algorithms in terms of the correlation matrix of the data. In Section 2.4 we replaced these correlations by their robust counerparts to obtain robust FS and SW.

We can also consider replacing the classical correlations in FS and SW by Spearman's ρ or Kendall's τ, since they are standard estimates of association that are invariant to monotone transformations of the data. They may be good options for variable selection when there is skewness in the data and no cluster of multivariate outliers. A small simulation study (not presented here) indicates that the methods based on Spearman's ρ and Kendall's τ may perform better than the classical FS and SW. Further study is required to investigate their performance as model building tools and compare them with classical and robust FS and SW.

It should be mentioned here that Spearman's ρ can be computed in $\mathcal{O}(n \log n)$ time, the same as the adjusted-Winsorized correlation estimate (the new correlation estimate proposed later in this thesis). Though Kendall's τ separately examines each of the $\binom{n}{2}$ (order of n^2) pairs of bivariate observations, there is an algorithm that can calculate Kendall's τ in $\mathcal{O}(n \log n)$ time (Knight 1966).

2.6 Examples

In this section, we used two real-data examples to show the robustness and scalability of our algorithms.

Executive data. This dataset is obtained from Mendenhall and Sincich (2003). The annual salary of 100 executives is recorded as well as 10 potential predictors (7 quantitative and 3 qualitative) such as education, experience etc. We label the candidate predictors from 1 to 10. Classical FS (with $F_{0.9}$ as the inclusion criterion) and SW (with $F_{0.9}$ as both inclusion and deletion criterion) both select the covariates: $(1, 3, 4, 2, 5)$. Robust FS and SW (also with $F_{0.9}$ as inclusion and deletion criterion) select the same model.

We then contaminated the data by replacing one small value of predictor 1 (less than 5) by a large value 100. When FS and SW are applied to the contaminated data, they both now select a larger set of variables: $(7, 3, 4, 2, 1, 5, 10)$. Thus, changing a single number in the data set drastically changes the selected model. On the other hand, robust FS and SW select the same model, $(1, 3, 4, 2, 5)$, when applied to the contaminated dataset.

Particle data. This quantum physics dataset was used for the KDD-Cup 2004. Each of $n = 50000$ data-points (rows) describes one "example" (particle generated in a high energy collider experiment). There are 80 variables in the data: Example ID, class of the example (positive examples are denoted by 1, negative examples by 0), and 78 feature measurements. We considered only the feature variables in our analysis. We deleted 13 of the features (either because they have a large number of missing values, or they are degenerate with all observations equal to 0), and used the first feature as the response. Thus, there are 64 covariates and one response in the selected data. Though this analysis may not be of particular scientific interest, it will demonstrate the scalability and robustness of our algorithms.

We first applied the four algorithms to a training sample of size $n = 5000$. The remaining 45000 cases will be used as a test sample. The classical FS and SW (with $F_{0.9}$ criterion) select the same model. It contains the following 25 covariates:

$$(2, 60, 58, 18, 8, 4, 51, 53, 1, 59, 5, 20, 10, 6, 62, 19, 38, 46, 39, 47, 21, 36, 50, 48, 37).$$

With $F_{0.95}$ criterion, the model has 23 covariates. Interestingly, only one covariate is selected by robust FS and SW (with either $F_{0.9}$ or $F_{0.95}$ criterion): Covariate 1. The reason for this drastic difference is as follows. The robust correlation of Y and Covariate 1 is 0.86, while the classical correlation between these variables is only 0.42. About 86% of the values of the response variable and 88% of the values of Covariate 1 are equal to zero. There are many zeroes in other covriates as well. Classical methods fail to identify this unusual pattern in the data and therefore are unable to select a parsimonious model that fits well the majority the data (as opposed to all the data). The robust methods, on the other hand, successfully detect the unusual pattern and select a model capable of predicting well 90% of the data as explained below.

We fitted the selected classical and robust models using the training data, and then used them to predict the test data outcomes. The 5% and 10% trimmed means of squared prediction errors for the classical and (robust) models are: 0.012 (0.043) and 0.008 (0.005), respectively. That is, the robust model with only one covariate predicts 90% of the data better than the classical model with 25 covariates.

To illustrate the scalability of our algorithm we also used a training sample of size $n = 25000$. This time, classical FS and SW select a model of 30 covariates, and robust FS and SW both select one covariate, in this case covariate 2 instead of covariate 1. (Covariates 1 and 2 have robust correlations 0.82 and -0.85 with Y, respectively.)

2.7 Conclusion

The main contribution of this chapter is that we developed robust step-by-step algorithms as one-step model-building procedures. Classical step-by-step algorithms FS and SW are popular and computationally suitable, but they are sensitive to outliers. We expressed these algorithms in terms of sample means, variances and correlations, and obtained simple robust versions of FS and SW by replacing these sample quantities by their robust counterparts. We used robust partial F-tests for stopping during the implementation of the proposed robust algorithms.

For the construction of the robust correlation matrix of the required covariates we used a pairwise approach, because it is both computationally suitable, and more consistent with the idea of step-by-step algorithms. We used robust correlations derived from a simplified version of Maronna's bivariate M-estimator of the scatter matrix.

Our robust methods have much better performance compared to the standard FS and SW algorithms. Also, they are computationally very suitable, and scalable to large dimensions.

Chapter 3

Two-step Model Building: Robust Sequencing with Least Angle Regression

3.1 Introduction

In this chapter, we will consider the first step (sequencing) of the two-step model building procedure. The candidate covariates will be sequenced to form a list such that the good predictors are likely to appear at the beginning of the list . The first m covariates of the list will form the reduced set which will be studied further in the next chapter.

We need a suitable step-by-step algorithm to sequence the covariates. We focus on the powerful algorithm recently proposed by Efron, Hastie, Johnstone and Tibshirani

(2004), which is called Least Angle Regression (LARS). LARS is computationally efficient and has been shown to have clear statistical advantages over other step-by-step algorithms.

Since LARS is based on sample means, variances and correlations (as will be shown later), it yields poor results when the data are contaminated. This is a potentially serious deficiency. Therefore, we propose several approaches to strengthen the robustness properties of LARS without affecting its computational efficiency too much, and compare their behavior.

The rest of this chapter is organized as follows. In Section 3.2, we review LARS in details. In Section 3.3, we express the LARS procedure in terms of the correlation matrix of the data. In Section 3.4, we illustrate LARS' sensitivity to outliers and introduce two different approaches to robustify LARS. A small simulation study is also presented here to compare the performance and the computing time of LARS to those of the two robust approaches. In Section 3.5, we investigate the selection of the size of the reduced set of candidate predictors. Section 3.6 proposes to use bootstrap to stabilize the results obtained by robust LARS. Section 3.7 introduces "learning curves" as a graphical tool to choose the size of the reduced set. Section 3.8 contains some real-data applications. Section 3.9 concludes and the chapter appendix contains some technical derivations.

3.2 Review: Least Angle Regression (LARS)

Least Angle Regression (LARS), proposed by Efron, Hastie, Johnstone and Tibshirani (2004), is closely related to another new algorithm called Forward Stagewise (Hastie,

Tibshirani and Friedman 2001, Chapter 10). To better understand LARS, we will review the Forward Stagewise procedure in details.

3.2.1 Forward Stagewise procedure (Stagewise)

The Forward Stagewise procedure (Stagewise) is related to the classical algorithm Forward Selection (FS). In FS, when the first predictor (X_1, say) is selected, all other predictors are regressed on X_1, and the residual vectors compete for the next entrance in the model. This causes a problem. Important predictors that happen to be correlated with X_1 are eliminated from the competition in many cases, which the researchers usually want to avoid. In this sense, FS is an aggressive model-building algorithm.

The Forward Stagewise procedure is a less aggressive version of FS. Unlike FS, the Stagewise procedure takes many tiny steps to move towards a final model. We take the zero vector as the initial prediction. If X_1 has the largest absolute correlation with Y, we modify our prediction by moving a 'small' step in the direction of X_1. We obtain the new residual vector and repeat the process, until the required number of predictors are selected. The goal is to obtain the order in which the variables enter the model.

We can assume, without loss of generality, that the covariates have mean 0 and variance 1, and the response has mean 0. Let ϵ be a positive constant, typically small (less than the absolute value of the regression coefficients). The Stagewise algorithm can be described as follows:

1. Set the prediction vector, $\hat{\boldsymbol{\mu}} = \mathbf{0}$.

2. Calculate $\hat{c}_j = X_j'(Y - \hat{\boldsymbol{\mu}}),\ \ j = 1, \ldots, k,$

   where $\hat{c}_j$ is proportional to the correlation between X_j and the current residual.

3. Let $m = \text{argmax}_j\, |\hat{c}_j|$. Modify the current prediction vector as follows:

$$\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} + \epsilon\, \text{sign}(\hat{c}_m)\, X_m,$$

   where ϵ is a positive constant.

4. Repeat steps 2 and 3.

At each step, the algorithm updates the prediction, and keeps track of the sequence of covariates as they enter the model. Notice that at each Stagewise step, we maximize the correlation of the current residual vector with a covariate, which is equivalent to minimizing the 'local loss'

$$\sum_{i=1}^{n} (Y_i - \hat{\mu}_i - \beta_j X_{ji})^2 \tag{3.1}$$

over all j. (Because of standardization, $\hat{\beta}_m$ is proportional to $\hat{c}_m$.)

Stagewise and FS: loss comparison

Both Stagewise and FS select the same variable in their first steps, because they minimize the same loss function. Suppose, the selected variable is X_1, i.e., its loss $\sum_{i=1}^{n} (Y_i - \beta_1 X_{1i})^2$ is minimum. Let us consider that, after many ϵ-steps along X_1, Stagewise is about to choose a second variable from the contenders $X_2, \ldots, X_d$. On the other hand, for the second step, FS considers $Z_2, \ldots, Z_d$, which are the residuals of the corresponding covariates after being adjusted for X_1. To choose the second variable, FS minimizes the

loss

$$\sum_{i=1}^{n} \left(Y_i - \beta_1 X_{1i} - \beta_j Z_{ji}\right)^2,$$

which is same as the loss

$$\sum_{i=1}^{n} \left(Y_i - \beta_1^* X_{1i} - \beta_j^* X_{ji}\right)^2. \tag{3.2}$$

Note that β_1^* is usually different from β_1. The loss used in Stagewise is

$$\sum_{i=1}^{n} \left(Y_i - \beta_1^\epsilon X_{1i} - \tilde{\beta}_j X_{ji}\right)^2. \tag{3.3}$$

This loss depends on our position on the X_1-vector controlled by ϵ. By choosing a small ϵ, we ensure that $\beta_1^\epsilon \leq \beta_1^*$, so that the variables correlated with X_1 have more chance of staying in the competition. (It should be noted that we are not 'fitting' a model. We are 'selecting' the covariates.) The minimizer of the Stagewise loss cannot beat the minimizer of the FS loss (see equation (3.2)), at least at this stage. This means, if FS chooses (X_1, X_2) and Stagewise selects (X_1, X_3), then FS will yield a greater value of R^2. Because, FS technically considers the residual sum of squares of the final fit (equation (3.2)). However, this is not necessarily true for the next stage if FS selects (X_1, X_2, X_4) (for example), and Stagewise selects (X_1, X_3, X_5). (Because, this Stagewise combination has not been considered by FS so far *in any order*. In other words, FS has taken a different path.)

Greater R^2 does not necessarily imply more prediction accuracy. Moreover, FS cannot guarantee greater R^2 for a particular subset size in all cases. Therefore, orthogonalization of the subsequent covariates with respect to the active ones is not usually meaningful. This is why researchers often prefer Stagewise to FS.

Stagewise and Boosting

Boosting, originally developed for classification problems, is a procedure that combines the output of many "weak" classifiers to produce a powerful "committee" (Hastie *et al.* 2001, Chapter 10). In the general setup, boosting is a way of predicting Y_i by combining a set of simple "basis" functions additively:

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \beta_k b(\boldsymbol{x}, \boldsymbol{\alpha}_k), \quad k = 1, \ldots, K, \tag{3.4}$$

where β_k are the expansion coefficients, and $b(\boldsymbol{x}, \boldsymbol{\alpha}_k)$ are real-valued functions of multivariate $\boldsymbol{x}$ with parameters $\boldsymbol{\alpha}_k$. Usually, $f(\boldsymbol{x})$ is fitted by considering a loss function

$$\sum_{i=1}^{n} L\left(Y_i, \sum_{k=1}^{K} \beta_k b(\boldsymbol{x}_i, \boldsymbol{\alpha}_k)\right), \tag{3.5}$$

which is minimized with respect to β_k's and $\boldsymbol{\alpha}_k$'s. Often, this is computationally intensive, and the solution to (3.5) is approximated by sequentially adding new basis functions without adjusting the parameters and expansion coefficients of the existing ones. In this approach, at each iteration k, one updates

$$f_k(\boldsymbol{x}_i) = f_{k-1}(\boldsymbol{x}_i) + \beta_k b(\boldsymbol{x}_i, \boldsymbol{\alpha}_k)$$

by minimizing the loss

$$\sum_{i=1}^{n} L\left(Y_i, f_{k-1}(\boldsymbol{x}_i) + \beta_k b(\boldsymbol{x}_i, \boldsymbol{\alpha}_k)\right). \tag{3.6}$$

In *Regularized boosting*, one uses a parameter ϵ to control the "learning rate" of the boosting procedure:

$$f_k(\boldsymbol{x}_i) = f_{k-1}(\boldsymbol{x}_i) + \epsilon \beta_k b(\boldsymbol{x}_i, \boldsymbol{\alpha}_k). \tag{3.7}$$

The Stagewise algorithm discussed before is very similar to this regularized boosting, where we use squared error loss for L in (3.6), K (see (3.4)) is the number of ϵ-steps

in Stagewise, $b(\boldsymbol{x}_i, \boldsymbol{\alpha}_k) = x_{ki}$ (the ith observation of covariate $X_{(k)}$ chosen in the kth iteration, not necessarily same as X_k), and β_k in (3.7) is replaced by $\text{sign}\{\beta_k\}$.

Choice of an appropriate ϵ for Stagewise

The choice of an appropriate ϵ is a problem with the Stagewise procedure, and is the motivation for LARS. If ϵ is 'small', the number of Stagewise steps to reach a final model may be very large, increasing the computational burden of the algorithm. On the other hand, if ϵ is 'large', we have either or both of the following two problems:

'Incorrect' ordering of predictors: If $\epsilon \to |\hat{c}_m|$, Stagewise will aggressively throw covariates correlated with X_m out of the competition.

A closed loop: In many cases, after the selection of the mth (say) covariate, the remaining predictors (that are not yet selected) have very small correlations with the current residual vector. Suppose that the correlation of the currently selected predictor X_m is positive. If ϵ is large, when we make an 'ϵ-step' in the direction of '$+X_m$', the correlation of X_m with the updated residuals becomes negative and larger in absolute value than that of any other competitor. Thus, the next Stagewise step is to make an 'ϵ-step' in the direction of '$-X_m$'. These back-and-forth movements may go on endlessly.

Even when there is no closed loop, the Stagewise procedure may require hundreds of tiny steps to reach the final model. Therefore, this algorithm is not computationally suitable. LARS overcomes this problem by taking a mathematical approach.

3.2.2 The LARS algorithm

LARS uses a mathematical formula to accelerate the computations in the Stagewise procedure. Suppose, the first selected predictor in Stagewise is X_1 (i.e., X_1 has the largest absolute correlation with Y). If we choose a 'small' ϵ, there will be at least several Stagewise steps in the direction of the vector X_1. A second predictor X_2 (say) will come in the picture as soon as we cross a certain point in the direction of X_1, a point at which both X_1 and X_2 have equal absolute correlation with the residual. LARS uses a mathematical formula to determine that point, and the prediction is modified by making a move up to that point in a single step.

In Stagewise, when a second predictor X_2 enters the model for the first time, we make a few small steps in the direction of X_2, but then X_1 becomes more correlated with the residual, and we move in the direction of X_1. Thus, we alternate between the two directions, technically maintaining approximately equal absolute correlations of X_1 and X_2 with the residual (until a third predictor comes into the picture). LARS, on the other hand, mathematically determines a direction that has equal angle (correlation) with X_1 and X_2, and makes the second LARS move along that direction upto a point (determined mathematically, again) at which a third predictor X_3 has equal absolute correlation with the residual vector, and so on.

For the original LARS algorithm, Efron *et al.* (2004) is referred to, which is designed to get the modified predictions at each step, in addition to the sequence of the covariates as they enter the model. In Section 3.3 we show that, if we are interested in the ordering of the covariates only (and not the modified predictions), the algorithm can be expressed in terms of the correlation matrix of the data (and not the observations themselves).

3.2.3 LARS and Shrinkage methods

LARS and Ridge Regression

If there are many correlated variables in a linear model, the estimates may show high variance. This can be prevented by imposing a restriction on the size of the coefficients. The ridge regression (Hoerl and Kennard 1970) minimizes a penalized residual sum of squares

$$\sum_{i=1}^{n}(Y_i - \beta' \boldsymbol{x}_i)^2 + \lambda \sum_{j=1}^{d} \beta_j^2,$$

where λ is the parameter that controls the amount of shrinkage.

Ridge regression shrinks the coefficients towards zero, but does not set some coefficients exactly equal to zero. Therefore, it is not suitable for subset selection, and cannot be compared to LARS. By imposing a different penalty, the Lasso algorithm (Tibshirani 1996) forces some of the coefficients to zero, which is presented below.

LARS and Lasso

The Lasso (Tibshirani 1996) estimates are obtained by minimizing

$$\sum_{i=1}^{n}(Y_i - \beta' \boldsymbol{x}_i)^2 + \lambda \sum_{j=1}^{d} |\beta_j|. \tag{3.8}$$

Moderate to large λ will cause some of the Lasso coefficients to be exactly zero, others will be smaller in magnitude than the corresponding least squares estimates.

Interestingly, the estimates (and the sequance of the covariates) obtained by LARS and Lasso are usually quite close, if not the same. The reason has not been established

mathematically, though it is clear that both algorithms can be viewed as less aggressive versions of the FS procedure. Efron *et al.* (2004) suggested a modification in the LARS algorithm that will yield the Lasso solution, which is as follows. Let $\hat{\boldsymbol{\beta}}$ be the current Lasso estimate, and $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$. Then, for the Lasso estimates,

$$\text{sign}(\hat{\beta}_j) = \text{sign}\left(\text{corr}(Y - \hat{\boldsymbol{\mu}}, X_j)\right),$$

which is not necessarily true for the LARS estimates. This restriction should be enforced in the LARS algorithm if we want the Lasso solution.

To better understand the LARS-Lasso relationship, let us consider the following definition of the Lasso estimate, which is equivalent to (3.8).

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}'\boldsymbol{x}_i)^2, \quad \text{subject to } \sum_{j=1}^{d} |\beta_j| \leq t. \tag{3.9}$$

The 'tuning parameter' t is varied over a certain range. If $t > \sum_{j=1}^{d} |\hat{\beta}_j^{\text{ls}}|$, where $\hat{\beta}_j^{\text{ls}}$ are the least squares estimates, then the Lasso estimates are the least squares estimates. For a LARS-Lasso comparison, suppose that LARS has selected the first covariate $s_1 X_1$. (LARS considers the 'signed covariates' to determine the equiangular vectors later.) To select the second covariate, LARS mathematically determines the minimum distance γ to move along $s_1 X_1$ so that a new covariate $s_2 X_2$ (say) is equally correlated with the residual vector. We may assume that γ is determined first (as in Stagewise, where we make many ϵ-steps to obtain β_ϵ (see 3.3)) so that LARS loss can be written as

$$\sum_{i=1}^{n} (Y_i - \gamma s_1 X_{1i} - \beta_j X_{ji})^2 = \sum_{i=1}^{n} (Y_i - \gamma_s X_{1i} - \beta_j X_{ji})^2, \tag{3.10}$$

where $j = 2, \ldots, d$, and $\gamma_s = s_1\gamma$ is a restricted regression coeffcient. Since $|\gamma_s|$ is the minimum distance (determined mathematically) to move before a second variable enters

the model, a comparison of (3.10) and (3.9) makes it evident that, in the Lasso algorithm,

$$t < |\gamma_s| \Rightarrow \text{ only } X_1 \text{ is in the model (only } \hat{\beta}_1 \text{ is nonzero).}$$

In LARS, when we have two active covariates s_1X_1 and s_2X_2, we modify our prediction by moving along the equiangular vector B_A upto a point γ_A, so that the LARS loss has the form

$$\begin{aligned} &\sum_{i=1}^{n} (Y_i - \gamma_s X_{1i} - \gamma_A B_{Ai} - \beta_j X_{ji})^2 \\ &= \sum_{i=1}^{n} (Y_i - \gamma_s X_{1i} - \gamma_A (w_1 s_1 X_{1i} + w_2 s_2 X_{2i}) - \beta_j X_{ji})^2 \\ &= \sum_{i=1}^{n} (Y_i - (\gamma + \gamma_A w_1) s_1 X_{1i} - \gamma_A w_2 s_2 X_{2i} - \beta_j X_{ji})^2, \end{aligned} \tag{3.11}$$

where $j = 3, \ldots, d$, and w_1 and w_2 are given by (3.21) (see Chapter Appendix, Section 3.10.2). Again, since γ_A is the minimum distance to move along B_A before a third covariate comes in the picture, by comparing (3.11) and (3.9) we can say

$$|\gamma_s| \leq t < |\gamma_s| + |\gamma_A(w_1 + w_2)| \Rightarrow \text{ only } X_1 \text{ and } X_2 \text{ will be in the model,}$$

and so on. Note that $|(\gamma + \gamma_A w_1)s_1| + |\gamma_A w_2 s_2| = |\gamma_s| + |\gamma_A(w_1 + w_2)|$, since γ, γ_A, w_1 and w_2 are all positive at this stage. (The w_j may not be all positive for higher dimensions.) Thus, it is not surprising that LARS and Lasso sequences agree in most cases. However, Lasso requires a computationally expensive quadratic programming technique to obtain the estimates, while the computational cost of LARS is comparable to the ordinary least squares applied to the full set of covariates.

For a comparison of Stagewise and boosting we refer to Section 3.2.1. Since LARS is a mathematical solution of the Stagewise problem, the LARS algorithm maybe considered as a mathematical alternative to a regularized boosting algorithm.

3.3 LARS expressed in terms of correlations

In this section, we show that the sequence of covariates obtained by LARS can be derived from the correlation matrix of the data (without using the observations themselves).

Let $Y, X_1, \ldots, X_d$ be the variables, standardized using their mean and standard deviation. Let r_{jY} denote the correlation between X_j and Y, and R_X be the correlation matrix of the covariates $X_1, \ldots, X_d$. Suppose that X_m has the maximum absolute correlation r with Y and denote $s_m = \text{sign}(r_{mY})$. Then, X_m becomes the first *active variable* and the current prediction $\hat{\boldsymbol{\mu}} \leftarrow \mathbf{0}$ should be modified by moving along the direction of $s_m X_m$ upto a certain distance γ that can be expressed in terms of correlations between the variables (see Chapter Appendix, Section 3.10.1, for details). By determining γ, LARS simultaneously identifies the new covariate that will enter the model, that is the second active variable.

As soon as we have more than one active variable, LARS modifies the current prediction along the *equiangular direction*, that is the direction that has equal angle (correlation) with all active covariates. Moving along this direction ensure that the current correlation of each active covariate with the residual decreases equally. Let A

be the set of subscripts corresponding to the active variables. In Chapter Appendix (Section 3.10.2) the standardized equiangular vector B_A is derived. Note that we do not need the direction B_A itself to decide which covariate enters the model next. We only need the correlation of all variables (active and inactive) with B_A. These correlations can be expressed in terms of the correlation matrix of the variables as shown in Chapter Appendix (Section 3.10.2). LARS modifies the current prediction by moving along B_A upto a certain distance γ_A which, again, can be determined from the correlations of the variables (see Chapter Appendix, Section 3.10.3).

Thus, the sequence of covariates obtained by the LARS algorithm is a function of the correlation matrix of the standardized data. We now summarize the LARS algorithm in terms of correlations r_{jY} between X_j and Y, and the correlation matrix R_X of the covariates:

1. Set the active set, $A = \emptyset$, and the sign vector $\boldsymbol{s}_A = \emptyset$.

2. Determine $m = \text{argmax}\, |r_{jY}|$, and $s_m = \text{sign}\{r_{mY}\}$. Let $r = s_m r_{mY}$.

3. Put $A \leftarrow A \cup \{m\}$, and $\boldsymbol{s}_A \leftarrow \boldsymbol{s}_A \cup \{s_m\}$.

4. Calculate $a = [\mathbf{1}_A'(D_A R_A D_A)^{-1}\mathbf{1}_A]^{-1/2}$, where $\mathbf{1}_A$ is a vector of 1's, $D_A = \text{diag}(\boldsymbol{s}_A)$, and R_A is the submatrix of R_X corresponding to the active variables. Calculate $\boldsymbol{w}_A = a\ (D_A R_A D_A)^{-1}\mathbf{1}_A$, and $a_j = (D_A \boldsymbol{r}_{jA})'\boldsymbol{w}_A$, for $j \in A^c$, where $\boldsymbol{r}_{jA}$ is the vector of correlations between X_j and the active variables. (Note that, when there is only one active covariate X_m, the above quantities simplify to $a = 1$, $w = 1$, and $a_j = r_{jm}$.)

5. For $j \in A^c$, calculate $\gamma_j^+ = (r - r_{jY})/(a - a_j)$, and $\gamma_j^- = (r + r_{jY})/(a + a_j)$,

and let $\gamma_j = \min(\gamma_j^+, \gamma_j^-)$. Determine $\gamma = \min\{\gamma_j, j \in A^c\}$, and m, the index corresponding to the minimum $\gamma = \gamma_m$. If $\gamma_m = \gamma_m^+$, set $s_m = +1$. Otherwise, set $s_m = -1$. Modify $r \leftarrow r - \gamma a$, and $r_{jY} \leftarrow r_{jY} - \gamma a_j$, for $j \in A^c$.

6. Repeat steps 3, 4 and 5.

3.4 Robustification of LARS

From the results in Section 3.3, it is not surprising to see that LARS is sensitive to contamination in the data. To illustrate this, we use a dataset on executives obtained from Mendenhall and Sincich (2003). The annual salary of 100 executives is recorded as well as 10 potential predictors (7 quantitative and 3 qualitative) such as education, experience etc. We label the candidate predictors from 1 to 10. LARS sequences the covariates in the following order: $(1, 3, 4, 2, 5, 6, 9, 8, 10, 7)$. We contaminate the data by replacing one small value of predictor 1 (less than 5) by the large value 100. When LARS is applied to the contaminated data, we obtain the following completely different sequence of predictors: $(\mathbf{7}, 3, \mathbf{2}, \mathbf{4}, 5, \mathbf{1}, \mathbf{10}, \mathbf{6}, \mathbf{8}, \mathbf{9})$. Predictor 7, which was selected last (10th) in the clean data, now enters the model first. The position of predictor 1 changes from first to sixth. Predictors 2 and 4 interchange their places. Thus, changing a single number in the data set completely changes the predictor sequence, which illustrates the sensitivity of LARS to contamination.

We now introduce two approaches to robustify the LARS procedure which we call the *plug-in* and *cleaning* approaches respectively.

3.4.1 Robust Plug-in

The plug-in approach consists of replacing the non-robust building blocks of LARS (mean, variance and correlation) by robust counterparts. The choices of fast computable robust center and scale measures are straightforward: median (med) and median absolute deviation (mad). Unfortunately, good available robust correlation estimators are computed from the d-dimensional data and therefore are very time consuming (see Rousseeuw and Leroy 1987). Robust pairwise approaches (see Huber 1981) are not affine equivariant and therefore are sensitive to two-dimensional outliers. One solution is to use robust correlations derived from a pairwise affine equivariant covariance estimator. A computationally efficient choice is a bivariate M-estimator as defined by Maronna (1976). Alternatively, a bivariate correlation estimator can be computed from bivariate Winsorized data. Both methods will be explained in detail below.

M Plug-in

Maronna's bivariate M-estimator of the location vector $\boldsymbol{t}$ and scatter matrix $\boldsymbol{V}$ is defined in Chapter 2. It is affine equivariant and computationally efficient, and has breakdown point 1/3 in two dimensions. As before, to further simplify computations, we used the coordinatewise median as the bivariate location estimate and only solved (2.18) to estimate the scatter matrix and hence the correlation. In this equation we used the function $u_2(t) = \min(c/t, 1)$ with $c = 9.21$, the 99% quantile of a χ^2_2 distribution. The bivariate correlations are then ensembled to form a $d \times d$ correlation matrix R. Finally, LARS is applied to this robust correlation matrix. We call this the **M plug-in** method.

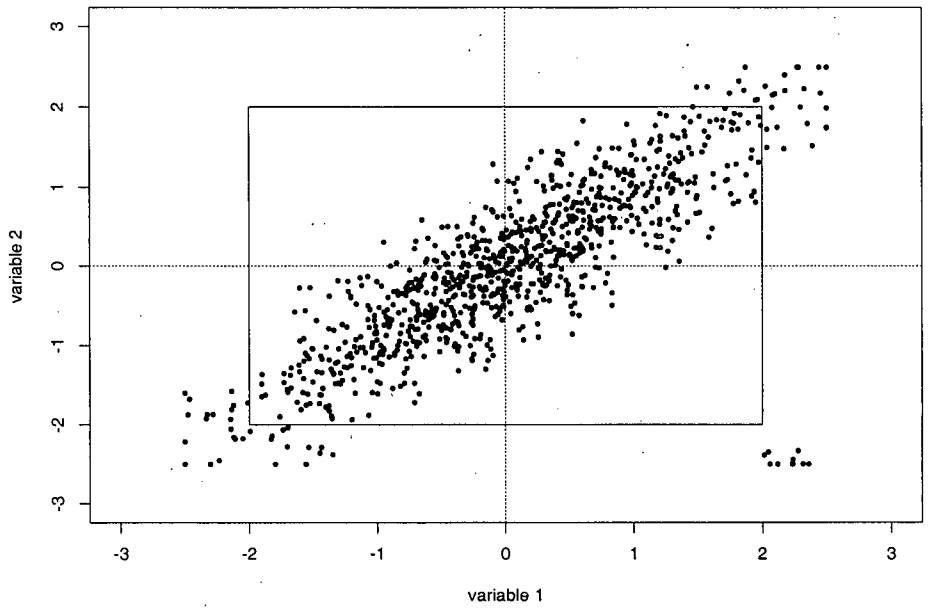


Figure 3.1: Limitation of separate univariate-Winsorizations ($c = 2$). The bivariate outliers are left almost unchanged.

W Plug-in

For very large, high-dimensional data we need an even faster robust correlation estimator. Huber (1981) introduced the idea of one-dimensional Winsorization of the data, and suggested that classical correlation coefficients be calculated from the transformed data. Alqallaf, Konis, Martin and Zamar (2002) re-examined this approach for the estimation of individual elements of a large-dimension correlation matrix. For n univariate observations $x_1, x_2 \ldots, x_n$, the transformation is given by $u_i = \psi_c((x_i - \text{med}(x_i))/\text{mad}(x_i))$, $i = 1, 2, \ldots, n$, where the Huber score function $\psi_c(x)$ is defined as $\psi_c(x) = \min\{\max\{-c, x\}, c\}$, with c a tuning constant chosen by the user, e.g., $c = 2$ or $c = 2.5$. This one-dimensional Winsorization approach is very fast to compute but un-

fortunately it does not take into account the orientation of the bivariate data. It merely brings the outlying observations to the boundary of a $2c \times 2c$ square, as shown in Figure 3.1. This plot clearly shows that the univariate approach does not resolve the effect of the obvious outliers at the bottom right which are shrunken to the corner (2, -2), and thus are left almost unchanged.

To remedy this problem, we propose a *bivariate Winsorization* of the data based on an initial tolerance ellipse for the majority of the data. Outliers are shrunken to the border of this ellipse by using the bivariate transformation $\boldsymbol{u} = \min(\sqrt{c/D(\boldsymbol{x})},\ 1)\ \boldsymbol{x}$ with $\boldsymbol{x} = (x_1, x_2)^t$. Here $D(\boldsymbol{x})$ is the Mahalanobis distance based on an initial bivariate correlation matrix R_0. For the tuning constant c we used $c = 5.99$, the 95% quantile of the χ^2_2 distribution. We call this the **W plug-in** method. The choice of R_0 will be discussed below.

Figure 3.2 shows bivariate Winsorizations for both the complete data set of Figure 3.1 and the data set excluding the outliers. The ellipse for the contaminated data is only slightly larger than that for the clean data. By using bivariate Winsorization the outliers are shrunken to the boundary of the larger ellipsoid.

The initial correlation estimate. Choosing an appropriate initial correlation matrix R_0 is an essential part of bivariate Winsorization. For computational simplicity we can choose the estimate based on univariate Winsorization explained above. However, we propose an adjusted Winsorization method that is more resistant to bivariate outliers. This method uses two tuning constants: a tuning constant c_1 for the two quadrants that contain the majority of the standardized data and a smaller tuning constant c_2 for the other two quadrants. For example, c_1 is taken equal to 2 or 2.5 as before and

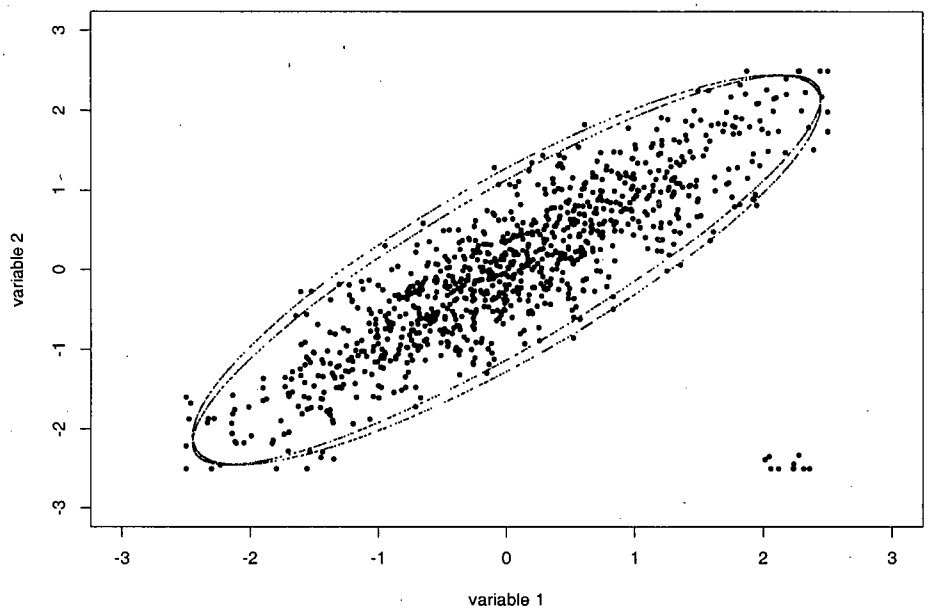


Figure 3.2: Bivariate Winsorizations for clean and contaminated data. The ellipse for the contaminated data is only slightly larger than that for the clean data.

$c_2 = hc_1$ where $h = n_2/n_1$ with n_1 the number of observations in the major quadrants and $n_2 = n - n_1$. We use $c_1 = 2$ in this chapter.

Figure 3.3 shows how the adjusted Winsorization deals with bivariate outliers, which are now shrunken to the boundary of the smaller square. Thus, adjusted Winsorization handles bivariate outliers much better than univariate Winsorization. The initial correlation matrix R_0 is obtained by computing the classical correlation matrix of the adjusted Winsorized data.

It should be mentioned here that, though we used $c_2 = hc_1$ in this study, a more reasonable choice would have been $c_2 = \sqrt{h}\, c_1$ (i.e., $c_2^2 = hc_1^2$), because the areas of the two squares should be proportional to the number of observations they contain.

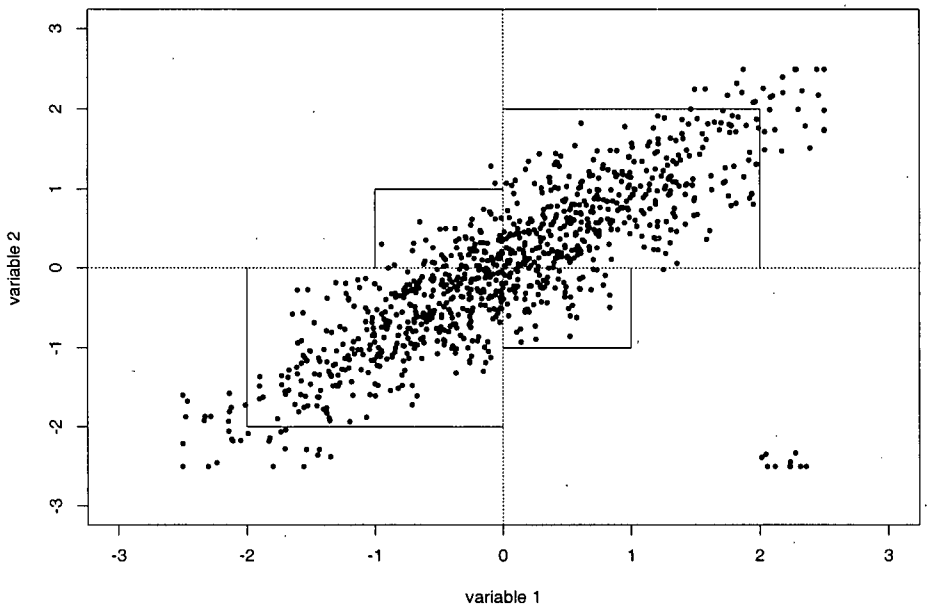


Figure 3.3: Adjusted-Winsorization (for initial estimate R_0) with $c_1 = 2$, $c_2 = 1$. The bivariate outliers are now shrunken to the corner of the smaller square.

Note that the correlations based on both univariate- and adjusted-Winsorized data can be computed in $\mathcal{O}(n \log n)$ time. The adjusted-Winsorized estimate takes slightly more time for a particular n, but is much more accurate in the presence of bivariate outliers as shown above. Bivariate-Winsorized estimate and Maronna's M-estimate also require $\mathcal{O}(n \log n)$ time, but Maronna's M-estimate has a larger multiplication factor depending on the number of iterations required. Thus for large n, the bivariate-Winsorized estimate is much faster to compute than Maronna's M-estimate. Figure 3.4 shows for each of the four correlation estimates the mean cpu times in seconds (based on 100 replicates) for 5 different sample sizes: 10000, 20000, 30000, 40000 and 50000. These results confirm that the bivariate-Winsorized estimate is faster to compute than Maronna's M-estimate and the difference increases with sample size. Numerical results (not presented here)

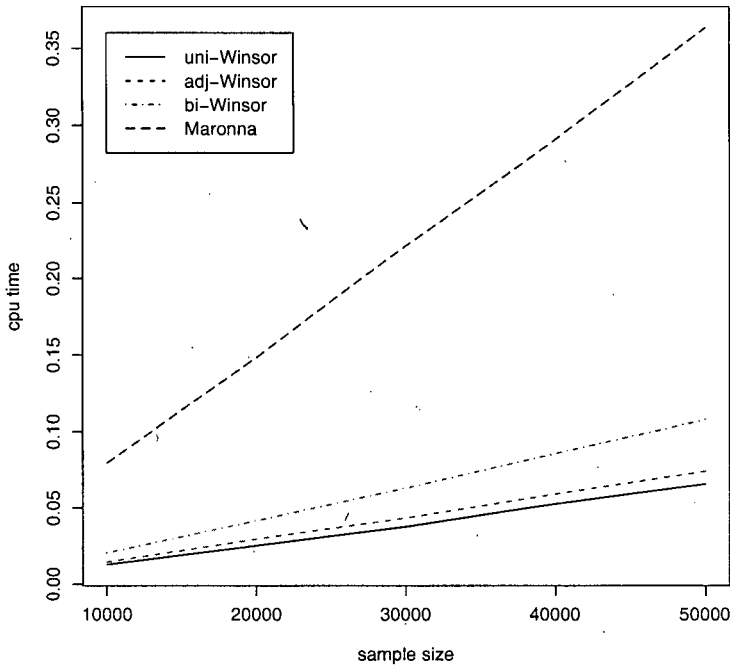


Figure 3.4: Numerical complexity of different correlation estimates. Each estimate can be computed in $\mathcal{O}(n \log n)$ time, but Maronna's estimate has a larger multiplication factor.

showed that the bivariate-Winsorized estimate is almost as accurate as Maronna's M-estimate also in the presence of contamination. Note that both the univariate-Winsorized and adjusted-Winsorized correlations are very fast to compute.

3.4.2 Robust Data Cleaning

If the dimension d is not extremely large, an alternative approach to robustifying LARS is to apply it on cleaned data. For example, each standardized d-dimensional data point $\boldsymbol{x} = (x_1, \ldots, x_d)^t$ can be replaced by its Winsorized counterpart $\boldsymbol{u} = \min(\sqrt{c/D(\boldsymbol{x})},\ 1)\ \boldsymbol{x}$ in the d-dimensional space. Here $D(\boldsymbol{x}) = \boldsymbol{x}^t V^{-1} \boldsymbol{x}$, is the Mahalanobis distance of $\boldsymbol{x}$ based

on V, a fast computable, robust initial correlation matrix. A reasonable choice for the tuning distance c is $c = \chi^2_d(0.95)$, the 95% quantile of the χ^2_d distribution.

The initial correlation matrix V. The choice of the initial correlation matrix V is an essential part of the Winsorization procedure. Most available high-breakdown, affine-equivariant methods are inappropriate for our purposes because they are too computationally intensive. Therefore, we resort to pairwise approaches, that is methods in which each entry of the correlation matrix is estimated separately (see Alqallaf et al. 2002). As before we will use a bivariate M-estimator or the bivariate windsorized estimator to calculate the correlations in V. The resulting methods are called **M cleaning** and **W cleaning**, respectively.

3.4.3 Simulations

To investigate the performance and stability of the four proposed methods we consider a simulation study involving a small number of variables. We used the following design (see Ronchetti *et al.* 1997). The error distributions considered are ($e1$) standard normal, ($e2$) 93% from standard normal and 7% from $N(0, 5^2)$, ($e3$) standard normal divided by a uniform on $(0, 1)$, and ($e4$) 90% from standard normal and 10% from $N(30, 1)$.

Two design matrices are considered: the uniform design for which the columns are generated from a uniform distribution on $(0, 1)$, and the leverage design which is the same as the uniform design except that it contains a leverage point. Six variables are used from which the first three are nonzero and in order of importance. The true regression coefficients for the nonzero variables are 7, 5 and 3, respectively. The sample size equals

Table 3.1: Percentages of correct sequences obtained by classical and robust methods for univariate and leverage designs with 4 different error distributions.

	Uniform				Leverage			
Method	$e1$	$e2$	$e3$	$e4$	$e1$	$e2$	$e3$	$e4$
LARS E	97	86	11	8	0	1	1	2
LARS G	100	89	26	24	0	2	5	7
M plug-in E	95	97	53	87	96	96	49	87
M plug-in G	99	99	74	95	99	99	68	95
W plug-in E	96	97	58	78	92	85	46	59
W plug-in G	99	99	77	89	94	86	61	68
M cleaning E	96	98	55	89	96	97	50	87
M cleaning G	99	99	77	97	100	98	73	97
W cleaning E	96	98	54	82	96	94	52	83
W cleaning G	99	99	76	92	98	96	71	92

$n = 60$ and we generated 200 data sets for each setting. We used two performance measures which we call exact (E) and global (G). The exact measure gives the percentage of times a procedure sequences the important variables in front and in their true order. The global measure gives the percentage of times a procedure sequences the important variables in front in any order.

Table 3.1 shows the simulation results. For error distribution $e1$ (standard normal), the performance of the robust methods is almost as good as that of standard LARS. For the heavy tailed distributions the robust methods drastically outperform LARS.

Overall we see from Table 3.1 that the plug-in approaches are almost as stable as the computationally more expensive data cleaning approaches. Comparing the M and W approaches for both the plug-in and data cleaning procedures, it is reassuring to see that the computationally faster W approach (see Figure 3.5 below) is almost as stable as the M approach.

Numerical complexity of the algorithms

We now compare the computational complexity of the different methods. The standard LARS procedure sequences all d covariates in only $\mathcal{O}(nd^2)$ time. The plug-in and cleaning procedures based on M-estimators both require $\mathcal{O}((n \log n)d^2)$ time. Based on Winsorization these procedures also require $\mathcal{O}((n \log n)d^2)$ time, but with a much smaller multiplication factor. Moreover, if we are only interested in sequencing the top fraction of a large number of covariates, then the plug-in approach will be much faster than the cleaning approach, because the plug-in approach only calculates the required correlations along the way instead of the 'full' correlation matrix. In this case, the complexity for plug-in methods reduces to $\mathcal{O}((n \log n)dm)$, where m is the number of variables being sequenced.

Figure 3.5 shows the mean cpu times based on 10 replicates for LARS, W plug-in and M plug-in for different dimensions d with a fixed sample size $n = 2000$. The times required by the cleaning methods are not shown because they were similar to the plug-in times since we sequenced all the covariates. As in Figure 3.4, we see that the approaches based on M-estimators are more time consuming than the Winsorization approaches. The difference increases fast with dimension.

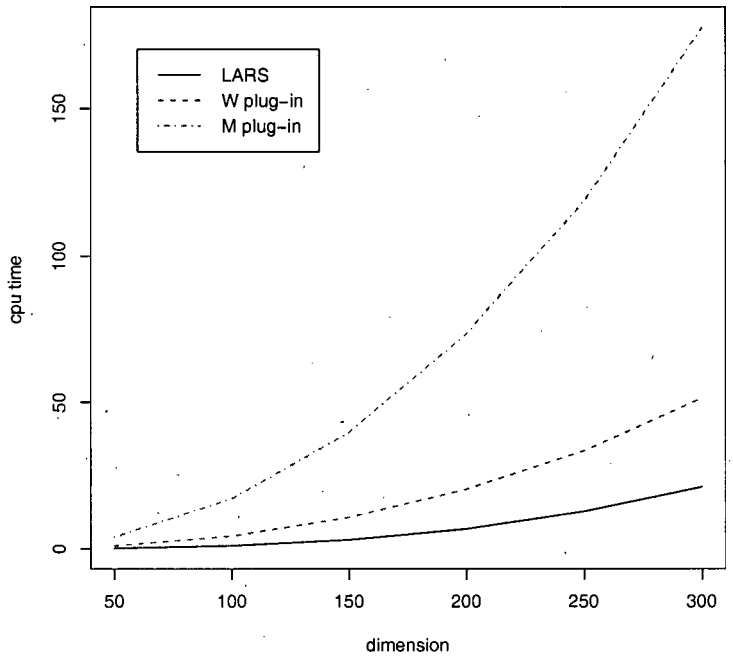


Figure 3.5: Numerical complexity of different techniques. LARS requires $\mathcal{O}(nd^2)$ time. W plug-in and M plug-in both require $\mathcal{O}((n \log n)d^2)$ time, but M plug-in has a larger multiplication factor.

The cleaning approaches perform slightly better than the plug-in approaches when the number of variables is relatively small, and much smaller than the number of cases (see Table 3.1). However, plug-in approaches are less time-consuming when only a part of the predictors are sequenced. Since W plug-in has a reasonable performance compared to the other methods and has favorable computing times, this method is to be preferred for large, high-dimensional datasets. The performance of W plug-in will be studied further in the next sections and we will call this method *robust LARS* from now on.

3.5 Size of the reduced set

To obtain a good final model, it is important to choose an appropriate value of m, the size of the reduced set of covariates kept from the sequencing step. The reduced set should be large enough to include most of the important covariates, but not so large as to make the segmentation step (where we have to evaluate all possible subsets of the reduced set) impractical. Several factors can be important when determining the size m such as d, the total number of variables, the sample size n, the unknown number of non-zero variables in the optimal model, the correlation structure of the covariates, and of course also time and feasibility of the segmentation step. For example, for high-dimensional datasets, including only 1% of the variables in the reduced set may make the segmentation step already infeasible.

To investigate what values of m are appropriate, we carry out a simulation study similar to Frank and Friedman (1993). The total number of variables is $d = 100$. A small number $a = 9$ or $a = 15$ of them are nonzero covariates. We considered 3 correlation structures of these nonzero covariates: "no-correlation" case, "moderate-correlation" case and "high-correlation" case, which are described below.

For the no-correlation case (a true correlation of 0 between the covariates), independent covariates $X_j \sim N(0, 1)$ are considered, and Y is generated using the a non-zero covariates, with coefficients $(7, 6, 5)$ repeated three times for $a = 9$, and five times for $a = 15$. The variance of the error term is chosen such that the signal-to-noise ratio equals 2.

For the moderate-correlation and high-correlation cases, we consider 3 independent

'unknown' processes, represented by latent variables L_i, $i = 1, 2, 3$, which are responsible for the systematic variation of both the response and the covariates. The model is

$$Y = 5L_1 + 4L_2 + 3L_3 + \epsilon = \text{Signal} + \epsilon, \tag{3.12}$$

where $L_i \sim N(0, 1)$, and ϵ is a normal error not related to the latent variables. The variance of ϵ is chosen such that the signal-to-noise ratio equals 2, that is $\text{Var}(\epsilon) = 50/4$. The nonzero covariates are divided in 3 equal groups, with each group related to exactly one of the latent variables by the following relation

$$X_j = L_i + \delta_j,$$

where $\delta_j \sim N(0, \sigma_j^2)$. The value of σ_j^2 determines the correlation structure of the nonzero covariates. The high-correlation case has a true correlation of 0.9 between the covariates generated with the same latent variable, and the moderate-correlation case has a true correlation of 0.5.

For each situation we generated 100 samples of size $n = 150$. Outliers were added by giving the noise term a large positive mean (asymmetric error). We considered four different levels of contamination: $0, 5, 10$ and 20%.

For the high-correlation and moderate-correlation cases, though "a" of the covariates are linked to the response Y through the latent variables, it is not clear which of these covariates should be considered important for explaining Y. Even when the true pairwise correlations of the covariates are zero (no-correlation case), the "best" model not necessarily includes all of the a non-zero coefficients because of the bias-variance trade-off. Therefore, for each simulated dataset we first find the "best" model among all possible subsets of the non-zero covariates that has the minimum prediction error estimated by 5-fold cross-validation.

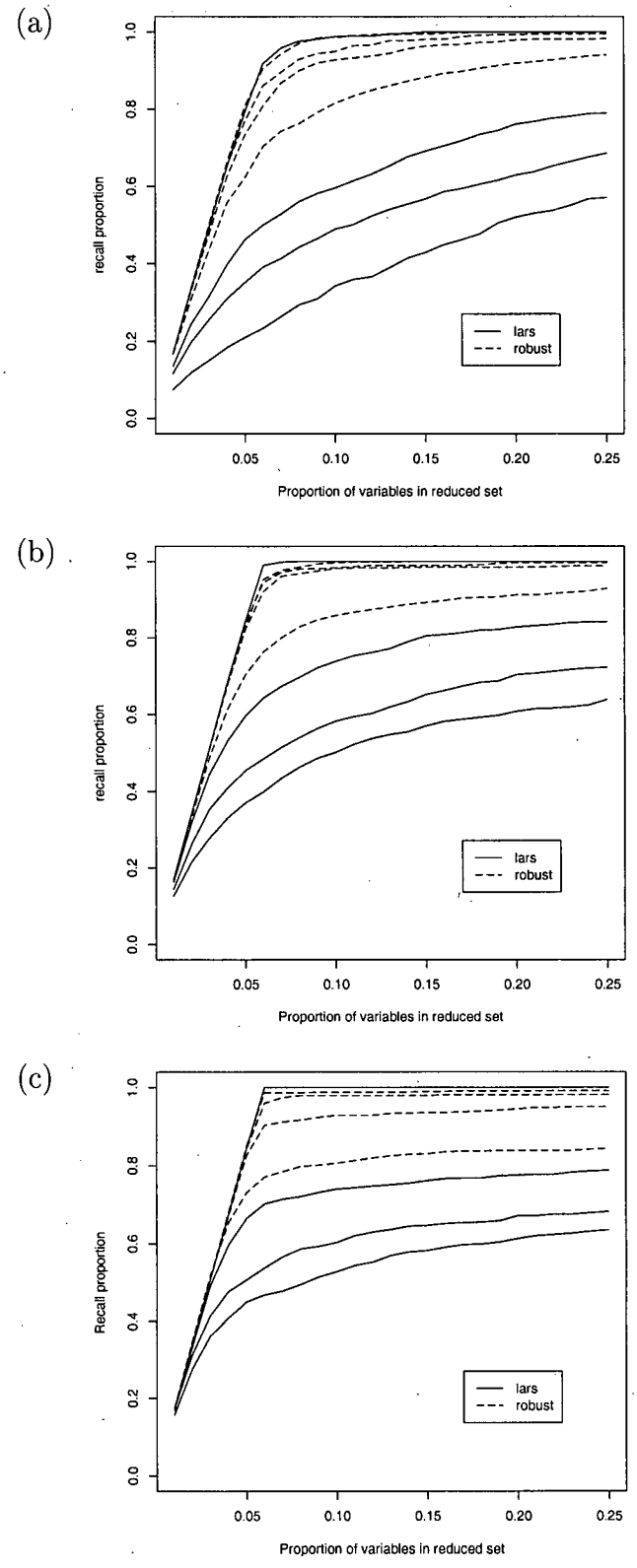


Figure 3.6: Recall curves for $a = 9$; (a) no correlation (b) low correlation (c) high correlation. The 4 curves for (robust) LARS correspond to 4 levels of contamination.

For each simulated dataset, we determine the "recall proportion", i.e., the proportion of important variables (in the sense that they are in the "best" model by cross-validation) that are captured (recalled) by LARS/robust LARS for a fixed size of the reduced sequence.

For $a = 9$, Figure 3.6 plots the average recall proportion against the size of the reduced set for the three correlation structures. In each plot, the 4 curves with the same line type correspond to the 4 levels of contamination, higher curves correspond to lower levels of contamination. These plots show that, for each correlation structure considered, we can capture the important variables if the percentage of variables in the reduced set is 9 or 10. Robust LARS performs as good as LARS for clean data, and much better than LARS for contaminated data.

Figure 3.7 plots the average recall proportion against the size of the reduced set for the moderate-correlation case with $a = 15$. This plot can be compared with Figure 3.6(b) to see how the increase in the number of nonzero variables affects the recall proportions. In both cases, we observe that the average recall proportions stop increasing even before the size m of the reduced set exceeds the number a of non-zero variables.

3.6 Bootstrapped sequencing

To obtain more stable and reliable results we can combine robust LARS with bootstrap. Therefore, we generate a number B of bootstrap samples from the dataset, and use robust LARS to obtain the corresponding sequence of covariates for each of these bootstrap samples. Each sequence ranks the covariates from 1 to d. For each covariate we can take

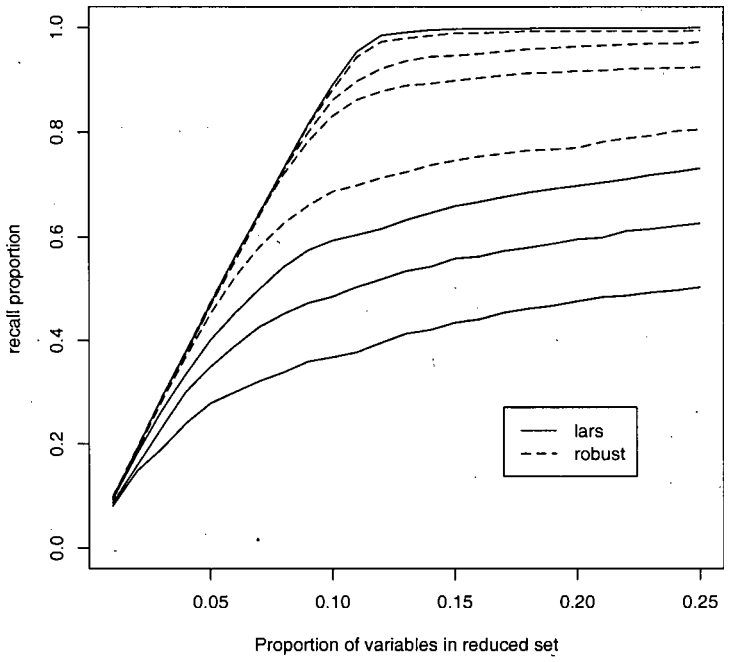


Figure 3.7: Recall curves for $a = 15$ and moderate correlation with 4 different levels of contamination.

the average over these B ranks, and the m covariates with the smallest average ranks then form the reduced set.

When resampling from a high-dimensional dataset (compared to the sample size, e.g., $n = 150, d = 100$) the probability of obtaining singular samples becomes very high. Note that even the original sample may already be singular or the dimension d of the data may exceed the sample size. In these cases it will be impossible to sequence all covariates. We can easily overcome this problem by sequencing only the first $m_0 < d$ of the covariates for each bootstrap sample, where preferably $m_0 \geq m$. We then rank the covariates according to the number of times (out of B) they are actually sequenced. When ties occur, the order of the covariates is determined according to the average rank

in the sequences. In our simulations, we generated $B = 100$ bootstrap samples from each of the 100 simulated datasets. We sequenced the first 25 covariates in each bootstrap sample.

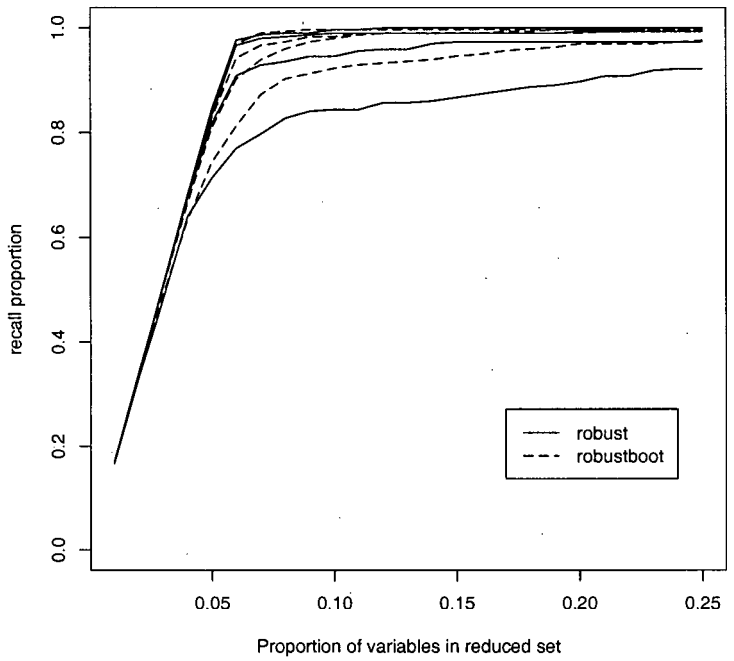
Figure 3.8 shows the recall curves obtained by robust LARS (solid lines) and bootstrapped robust LARS (dotted lines) for covariates with moderate correlation. The recall curves obtained by bootstrapped robust LARS perform better than the initial robust LARS curves for all levels of contamination, the difference being larger with larger contamination proportions. This confirms that by applying the bootstrap we obtain more stable and reliable results. Even with 20% of contamination, bootstrapped robust LARS with $m = 10$ ($a = 9$) or $m = 15$ ($a = 15$) already yields a recall proportion around 90%.

To investigate what minimum number of bootstrap samples is required to obtain significant improvement over robust LARS, we also tried $B = 10$, 20 and 50 in the above setups. In each case, $B = 10$ and $B = 20$ do not yield much improvement, while with $B = 50$ the results obtained are almost as stable as with $B = 100$.

3.7 Learning curves

Although the simulation results in the previous sections suggested that it suffices to select the size of the reduced set equal to or slightly larger than the number of predictors in the final model, we usually have no information about the number of predictors that is needed. Hence, a graphical tool to select the size of the reduced set would be useful. The following plot can be constructed to determine a reasonable size for the reduced set. Starting from a model with only 1 variable (the first one in the sequence), we increase the number of

(a)



(b)

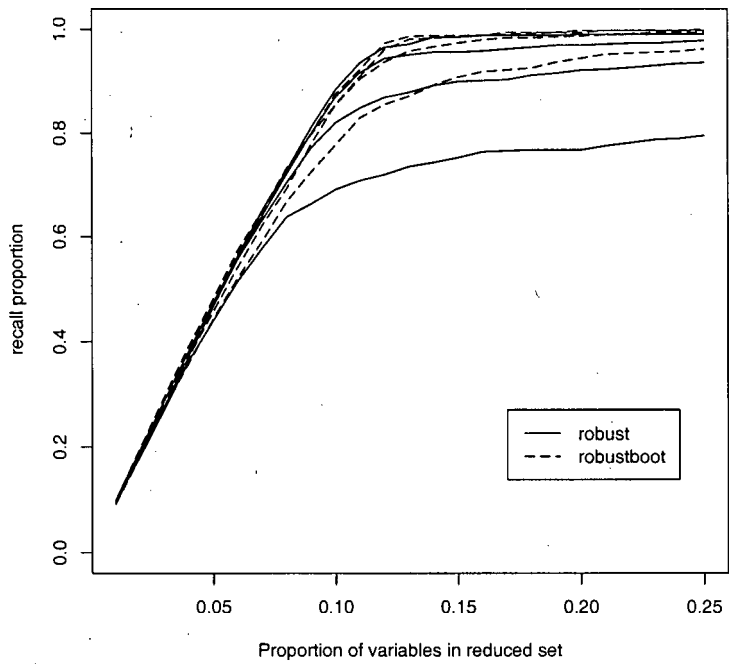


Figure 3.8: Recall curves for robust LARS and bootstrapped robust LARS for covariates with moderate correlation; (a) $a = 9$ (b) $a = 15$. The 4 curves for each method correspond to 4 levels of contamination.

variables according to the sequence obtained and each time fit a robust regression model to compute a robust R^2 measure such as $R^2 = 1 - \text{Median}(\boldsymbol{e}^2)/\text{MAD}^2(Y)$, where $\boldsymbol{e}$ is the vector of residuals from the robust fit. We then plot these robust R^2 values against the number of variables in the model to obtain a *learning curve.* The size of the reduced set can be selected as the point where the learning curve does not have a considerable slope anymore.

A problem that can occur with a robust R^2 measure is that, unlike its classical counterpart, it is not always a nondecreasing function of the number of covariates. This can be resolved as follows. If the robust R^2 at any step is smaller than that of the preceding step, then fit a robust simple linear regression of the residuals from the preceding step on the newly selected covariate. The residuals obtained from this fit can be used to compute another robust R^2 value. We then use the larger of the two values.

To investigate the performance of learning curves, we consider a dataset on air pollution and mortality in 60 Metropolitan areas in the United States. The response variable is the age-adjusted mortality. There are 14 potential predictors, numbered from 1 to 14. Since row 21 contains 2 missing values, we drop this observation from the data. Based on robust data exploration we identified 4 clear outliers that correspond to the four metropolitan areas in California. We applied 5-fold cross-validation (CV) to this dataset without the four outliers, and obtained the "best model" that has the following 7 covariates: $(2, 3, 4, 6, 7, 10, 13)$. (The order of the covariates is not relevant here.)

Bootstrapped robust LARS applied to this dataset (including the outliers) produced the sequence $(7, 5, 13, 4, 6, 3, 2, 10, 9, 1, 14, 11, 8, 12)$. We used this sequence and fitted Least Median of Squares (Rousseeuw 1984) regressions to obtain the robust R^2 values.

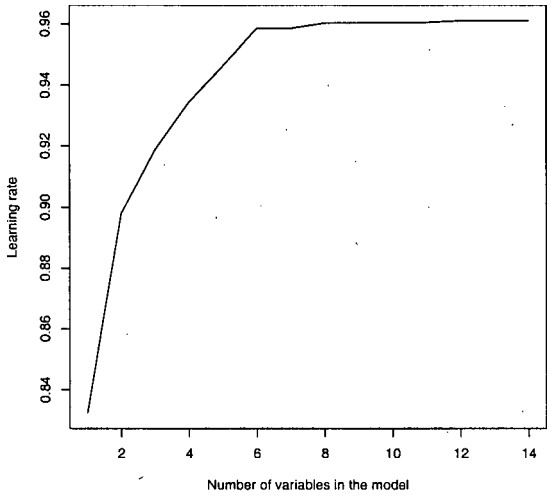


Figure 3.9: Learning curve for Pollution data. A reduced set of 8 covariates is suggested by the plot.

Figure 3.9 shows the corresponding learning curve. This plot suggests a reduced set of size 8. It is encouraging to notice that the reduced set (first 8 covariates in the sequence above) contains all 7 predictors selected in the "best model" obtained by CV.

3.8 Examples

In this section we use two real datasets to evaluate the performance of (bootstrapped) robust LARS. The demographic data example further explores the idea of "learning curves" to choose the size of the reduced set. We then use a large dataset (protein data) to demonstrate the scalability as well as stability of robust LARS.

Demographic data. This dataset contains demographical information on the 50 states of the United States for 1980. The response variable of interest is the murder rate per 100,000 residents. There are 25 predictors which we number from 1 to 25. Exploration of the data using robust estimation and graphical tools revealed one clear outlier. We applied 5-fold CV to this dataset without the outlier, and obtained the "best of 25" model that has the following 15 covariates $(1, 2, 3, 5, 6, 8, 9, 10, 16, 17, 18, 19, 21, 24, 25)$.

Figure 3.10 shows the learning curve for the Demographic data based on bootstrapped robust LARS. This plot suggests a reduced set of size 12 which include the covariates: $(22, 20, 4, 15, \mathbf{10}, \mathbf{2}, \mathbf{19}, \mathbf{25}, \mathbf{8}, \mathbf{18}, \mathbf{6}, \mathbf{24})$. The boldface numbers correspond to covariates in the sequence that are also in the model obtained by CV. The number of "hits" is 8 out of 12.

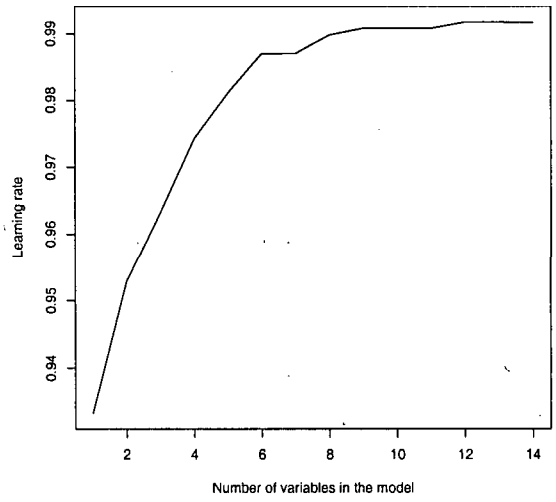


Figure 3.10: Learning curve for Demographic data. A reduced set of 12 covariates is suggested by the plot.

We applied 5-fold CV to the clean data using the reduced set of size 12 obtained by

bootstrapped robust LARS. The model selected in this case has the following 9 covariates: $(22, 20, 4, 15, 2, 10, 25, 18, 24)$. To compare this "best of 12" model with the "best of 25" model above, we estimated the prediction errors of these two models 1000 times using 5-fold CV. The two density curves are shown in Figure 3.11. The "best of 12" model has a mean error of 204.8 (median error 201.5) while the "best of 25" model has a mean error of 215.9 (median error 202.0). Also, the standard deviations (mads) of the errors are 25.6 (22.7) and 74.6 (31.4), respectively. (Some of the "best of 25" errors are very large and not included in the plot.) Thus, bootstrapped robust LARS gives more stable results in this high-variability dataset. It should be mentioned here that we needed almost 10 days to find the "best of 25" model, while "best of 12" model requires less than 5 minutes including the time needed to sequence the covariates by bootstrapped robust LARS. (CV on m covariates is $2^{(d-m)}$ times faster than CV on d covariates.)

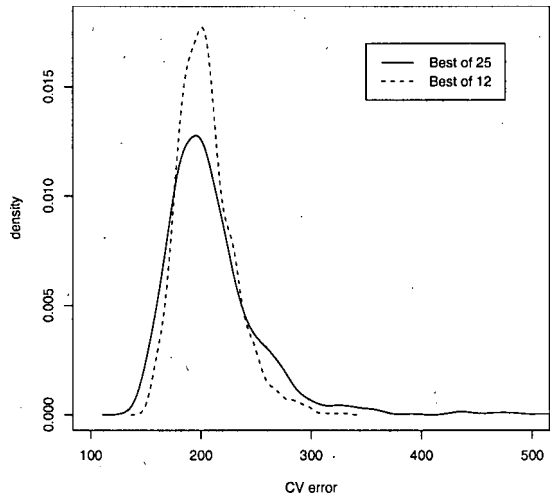


Figure 3.11: Error densities for the two "best" models for Demographic data. The "best of 12" model gives more stable result.

Protein data. This dataset of $n = 145751$ protein sequences was used for the KDD-Cup 2004. Each of the 153 blocks corresponds to a native protein, and each data-point of a particular block is a candidate homologous protein. There are 75 variables in the dataset: the block number (categorical) and 74 measurements of protein features. We replace the categorical variable by block indicator variables, and use the first feature as the response. Though this analysis may not be of particular scientific interest, it will demonstrate the scalability and stability of the robust LARS algorithm.

We used the package R to apply robust LARS to this dataset, and obtained a reduced set of size 25 from $d = 225$ covariates (152 block indicators + 73 features) in only 30 minutes. Given the huge computational burden of other robust variable selection procedures, our algorithm maybe considered extremely suitable for computations of this magnitude.

For a thorough investigation of the performance of robust LARS with this dataset, we select 5 blocks with a total of $n = 4141$ protein sequences. These blocks were chosen because they contain the highest proportions of homologous proteins (and hence the highest proportions of potential outliers). We split the data of each block into two almost equal parts to get a training sample of size $n = 2072$ and a test sample of size $n = 2069$. The number of covariates is $d = 77$, with 4 block indicators (variables $1 - 4$) and 73 features. We apply bootstrapped robust LARS with $B = 100$ bootstrap samples and we sequence the first 25 variables of each bootstrap sample. The resulting learning curve is shown in Figure 3.12.

This plot suggests that a drastic reduction to a small number of predictors can be performed, e.g. m=5 or m=10. The first 10 predictors found by bootstrapped robust

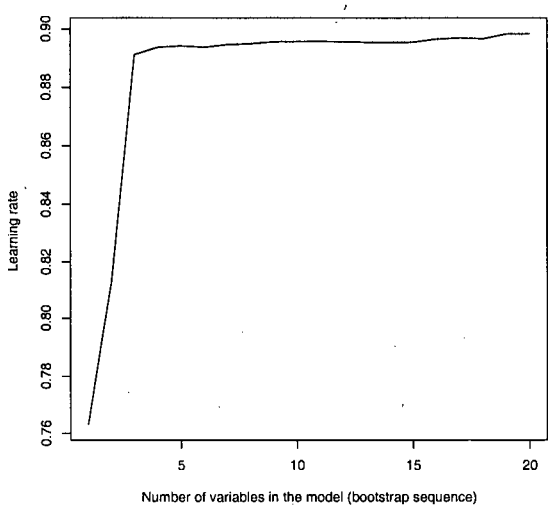


Figure 3.12: Learning curve for Protein data. A reduced set of 5 covariates is suggested by the plot.

LARS are $(14, 13, 5, 76, 73, 8, 7, 40, 46, 51)$. The covariates in this sequence are almost the same as those obtained with the whole dataset (not shown). The standard LARS produced the sequence $(14, 13, 5, 8, 7, 76, 18, 65, 2, 46)$. Note that the two sequences are quite different. For example, if we select a model from the first five predictors, then only 3 predictors are contained in both sequences. Using MM-estimators and robust AIC, the best model selected from the first five variables of the robust sequence contains variables $(14, 13, 5, 76)$ while the best model out of the first 10 predictors contains variables $(14, 13, 5, 76, 40)$. Hence only 1 variable is added.

Using classical AIC, the best model selected from the first 5 variables of the LARS sequence contains variables $(14, 13, 5, 8)$. Variable 76 of the corresponding robust model is replaced by Variable 8. The best model from the first 10 predictors contains variables $(14, 13, 5, 8, 76, 2)$. Note that 2 variables are added to the list compared to 1 variable in

the robust case.

We fitted the 4 best models using the training data, and then used them to predict the test data outcomes. The 1%, 5% and 10% trimmed means of prediction errors for the smaller robust (classical) model are : 114.92 (117.49), 92.77 (95.66) and 74.82 (78.19), respectively. The corresponding quantities for the larger robust (classical) model are: 114.37 (115.46), 92.43 (94.84) and 74.34 (76.50), respectively. Notice that the robust models always outperform the classical models.

3.9 Conclusion

The main contribution of this chapter is that we developed robust versions of LARS to obtain a reduced set of covariates for further investigation. We also introduced the idea of multivariate-Winsorization of the data (when the dimension is not too large). We can perform computationally suitable classical multivariate analyses on the transformed data to obtain reliable results. We also proposed a new robust correlation estimate for bivariate data which we called the "adjusted-Winsorized correlation estimate."

LARS is a very effective, time-efficient model building tool, but is not resistant to outliers. We introduced two different approaches to construct robust versions of the LARS technique. The plug-in approach replaces the classical correlations in LARS by easily computable robust correlation estimates. The cleaning approach first transforms the dataset by shrinking the outliers towards the bulk of the data, and then applies LARS on the transformed data. Both approaches use robust pairwise correlation estimates which can be computed efficiently using bivariate-Winsorization or bivariate M-estimates.

The data cleaning approach is limited in use because the sample size needs to be (much) larger than the number of candidate predictors to ensure that the resulting correlation matrix is positive definite. Moreover, the data cleaning approach is more time consuming than the plug-in approach, certainly when only part of the predictors is being sequenced. Since the plug-in approach has good performance, is faster to compute and more widely applicable, we prefer this method. Comparing bivariate M-estimates with bivariate Winsorization we showed that the latter is faster to compute with important time differences when the number of candidate predictors becomes high.

We propose using the robust LARS technique to sequence the candidate predictors and as such identify a reduced set of most promising predictors from which a more refined model can be selected in a second segmentation step. We recommend combining W plug-in with bootstrap to obtain more stable and reliable results. The reduced sets obtained by bootstrapped robust LARS contain more of the important covariates than the reduced sets obtained by initial robust LARS.

It is important to select the number of predictors to use for the second step. This number is a trade-off between success-rate, that is the number of important predictors captured in the reduced set, and feasibility of the segmentation step. Our simulation study indicated that the reduced set can have size comparable to the actual number of relevant candidate predictors. However, this number is usually unknown. To still get an idea about an appropriate size for the reduced set we introduced a learning curve that plots robust R^2 values versus dimension. An appropriate size can be selected as the dimension corresponding to the point where the curve starts to level off.

3.10 Chapter Appendix

3.10.1 Determination of γ for one active covariate

Assume that the first selected covariate is $+X_m$. The current prediction $\hat{\boldsymbol{\mu}} \leftarrow \mathbf{0}$ should be modified as

$$\hat{\boldsymbol{\mu}} \leftarrow \gamma\, X_m.$$

The distance γ should be such that the modified residual $(Y - \hat{\boldsymbol{\mu}})$ will have equal correlation with $+X_m$ and another signed covariate X_j. We have

$$\text{cor}(Y - \hat{\boldsymbol{\mu}}, X_m) = \frac{X_m'(Y - \gamma X_m)/n}{\text{SD}(Y - \gamma X_m)} = \frac{r - \gamma}{\text{SD}(Y - \gamma X_m)}, \tag{3.13}$$

and

$$\text{cor}(Y - \hat{\boldsymbol{\mu}}, +X_j) = \frac{X_j'(Y - \gamma X_m)/n}{\text{SD}(Y - \gamma X_m)} = \frac{r_{jY} - \gamma r_{jm}}{\text{SD}(Y - \gamma X_m)}. \tag{3.14}$$

Equating (3.13) to (3.14), we have

$$\gamma(+X_j) = \frac{r - r_{jY}}{1 - r_{jm}}. \tag{3.15}$$

Similarly, equating (3.13) with the correlation of modified residual and $-X_j$ we have

$$\gamma(-X_j) = \frac{r + r_{jY}}{1 + r_{jm}}. \tag{3.16}$$

We should take the minimum of (3.15) and (3.16) and minimum over all inactive (not yet selected) j. The signed covariate that will enter the model at this point is determined alongwith.

3.10.2 Quantities related to equiangular vector B_A

Here, A is the set of 'active' subscripts. Let $X_A = (\cdots s_l X_l \cdots)$, $l \in A$, where s_l is the sign of X_l as it enters the model. The standardized equiangular vector B_A is obtained using the following three conditions. B_A is a linear combination of the active signed predictors.

$$B_A = X_A \, \boldsymbol{w}_A \, , \quad \text{where } \boldsymbol{w}_A \text{ is a vector of weights.} \tag{3.17}$$

B_A has unit variance:

$$\frac{1}{n} B_A' B_A = 1. \tag{3.18}$$

B_A has equal correlation (a, say) with each of the active predictors. Since the covariates and B_A are standardized,

$$\frac{1}{n} X_A' B_A = a \, \mathbf{1}_A \, , \quad \mathbf{1}_A \text{ is a vector of 1's.} \tag{3.19}$$

Using equation (3.17) in equation (3.18), we have

$$\frac{1}{n} \boldsymbol{w}_A' X_A' X_A \boldsymbol{w}_A = 1,$$

so that

$$\boldsymbol{w}_A' R_A^{(s)} \boldsymbol{w}_A = 1, \tag{3.20}$$

where $R_A^{(s)}$ is the correlation matrix of the active signed variables. Using (3.17) in (3.19), we have

$$R_A^{(s)} w_A = a \mathbf{1}_A,$$

so that the weight vector w_A can be expressed as

$$w_A = a \, (R_A^{(s)})^{-1} \mathbf{1}_A.$$

Let R_A be the correlation matrix the unsigned active covariates, i.e., R_A is a submatrix of R_X. Let $\boldsymbol{s}_A$ be the vector of signs of the active covariates (we get the sign of each covariate as it enters the model). We have

$$w_A = a\,(D_A R_A D_A)^{-1}\mathbf{1}_A, \tag{3.21}$$

where D_A is the diagonal matrix whose diagonal elements are the elements of s_A. Finally, using equation (3.21) in equation (3.20), we get

$$a = [\mathbf{1}_A'(D_A R_A D_A)^{-1}\mathbf{1}_A]^{-1/2}. \tag{3.22}$$

The correlation of an inactive covariate X_j with B_A, denoted by a_j, can be expressed as follows

$$a_j = \frac{1}{n}X_j' B_A = \frac{1}{n}X_j' X_A \boldsymbol{w}_A = (D_A \boldsymbol{r}_{jA})'\boldsymbol{w}_A, \tag{3.23}$$

where $\boldsymbol{r}_{jA}$ is the vector of correlation coefficients between the inactive covariate X_j and the (unsigned) selected covariates. Thus, we need only (a part of) the correlation matrix of the data (not the observations themselves) to determine the above quantities.

3.10.3 Determination of γ for two or more active covariates

Let us update $r \leftarrow (r - \gamma)$, see (3.13), and $r_{jY} \leftarrow (r_{jY} - \gamma r_{jm})$, see (3.14).

The correlation of an active covariate with the 'current' residual $Y - \hat{\boldsymbol{\mu}}$ is $r/\mathrm{SD}(Y - \hat{\boldsymbol{\mu}})$, and the correlation of the active covariate with the current equiangular vector B_A is 'a'. Therefore, the correlation between an active covariate and the 'modified' residual $(Y - \hat{\boldsymbol{\mu}} - \gamma_A B_A)$ is

$$\frac{r - \gamma_A\, a}{\mathrm{SD}(Y - \hat{\boldsymbol{\mu}} - \gamma_A B_A)}.$$

An inactive covariate $+X_j$, $j \in A^c$, has correlation $r_{jY}/\mathrm{SD}(Y - \hat{\boldsymbol{\mu}})$ with the 'current' residual, and it has correlation a_j with B_A. Therefore, the correlation between $+X_j$, $j \in A^c$, and the 'modified' residual is

$$\frac{r_{jY} - \gamma_A a_j}{\mathrm{SD}(Y - \hat{\boldsymbol{\mu}} - \gamma_A B_A)}.$$

Equating the above two quantities, we get

$$\gamma_A(+X_j) = (r - r_{jY})/(a - a_j). \tag{3.24}$$

Similarly,

$$\gamma_A(-X_j) = (r + r_{jY})/(a + a_j). \tag{3.25}$$

We have to choose the minimum possible γ_A over all inactive covariates. Note that when A has only one covariate, (3.24) and (3.25) reduce to (3.15) and (3.16), respectively.

Chapter 4

Two-step Model Building: Robust Segmentation

4.1 Introduction

In Chapter 3 we developed robust sequencing methods to obtain a reduced set of covariates from which the final prediction model can be selected. According to the notation used before, we have m predictors $X_1, \ldots, X_m$ in the reduced set. In this chapter we consider methods of segmentation (evaluation of all possible subsets of the reduced set of covariates) in order to select the final prediction model.

To compare different subsets of covariates, we require an appropriate robust selection criterion. For this purpose, we review some classical selection criteria in Section 4.2, and their robust counterparts in Section 4.3. We use $\hat{\boldsymbol{\beta}}_p$ to denote the estimate of $\boldsymbol{\beta}_p$ for the

p-parameter submodel (p predictors including the intercept) under consideration. Many of the methods below require and estimate the variance of the error term under the "true" model, σ^2. In such cases, σ^2 is estimated using the full model (with $k = m + 1$ parameters).

4.2 Review: classical selection criteria

In this section, we review some important classical selection criteria: Final Prediction Error (FPE), Akaike Information Criterion (AIC), Mallows' C_p, cross-validation and bootstrap.

4.2.1 Akaike Information Criterion (AIC)

A measure of the similarity between the fitted distribution $f(y|\hat{\boldsymbol{\beta}}_p)$ and the true distribution $g(y|\boldsymbol{\beta})$ is the Kullback-Leibler information number (Kullback and Leibler 1951)

$$\begin{aligned} I(g,f) &= E\left\{\log\frac{g(Y|\boldsymbol{\beta})}{f(Y|\hat{\boldsymbol{\beta}}_p)}\right\} \\ &= \int \log\left(\frac{g(y|\boldsymbol{\beta})}{f(y|\hat{\boldsymbol{\beta}}_p)}\right) g(y|\boldsymbol{\beta})\, dy. \end{aligned}$$

It can be shown that

(i) $I(g,f) \geq 0$,

(ii) $I(g,f) = 0 \iff g(y) = f(y)$ almost everywhere (Lebesgue measure).

Our purpose is to minimize

$$I(g,f) = E\{\log g(Y|\boldsymbol{\beta})\} - E\left\{\log f(Y|\hat{\boldsymbol{\beta}}_p)\right\},$$

where only the second term is important in evaluating the fitted model. This term is unknown, and it seems reasonable to consider the log-likelihood

$$L(\hat{\boldsymbol{\beta}}_p) = \sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\beta}}_p)$$

as an estimate of $nE\left\{\log f(Y|\hat{\boldsymbol{\beta}}_p)\right\}$. However, this estimate has a bias, since the same data are used to find the estimates $\hat{\boldsymbol{\beta}}_p$ and to calculate the log-likelihood. Akaike (1973) showed that the expected value of the bias $\simeq p$. Therefore, the corrected estimate of $n\,E\left\{\log f(Y|\hat{\boldsymbol{\beta}}_p)\right\}$ is

$$L^*(\hat{\boldsymbol{\beta}}_p) = L(\hat{\boldsymbol{\beta}}_p) - p.$$

Based on this, Akaike (1973) proposed to choose the model that minimizes the Akaike Information Criterion:

$$\text{AIC} = -2L(\hat{\boldsymbol{\beta}}_p) + 2p.$$

Bhansali and Downham (1977) proposed to generalize AIC by choosing a model that minimizes, for a chosen fixed α,

$$\text{AIC}(p,\alpha) = -2L(\hat{\boldsymbol{\beta}}_p) + \alpha\, p.$$

For normal errors,

$$\text{AIC}(p,\alpha) = K(n,\hat{\sigma}) + \frac{RSS_p}{\hat{\sigma}^2} + \alpha p, \tag{4.1}$$

where $K(n,\hat{\sigma})$ is a constant depending on the marginal distribution of the covariates, RSS_p is the residual sum of squares, and $\hat{\sigma}^2$ is the estimate of σ^2 from the full model.

4.2.2 Mallows' C_p

Let us consider the following submodel of p parameters:

$$Y_i = \boldsymbol{\beta}_p' \boldsymbol{x}_i + \epsilon_i, \ i = 1, 2, \cdots, n, \tag{4.2}$$

where ϵ_i are independent observations from the distribution F with mean zero and variance σ^2 (when the current submodel is the true model). This subset model may produce biased fitted values, i.e., $\mathrm{E}(\hat{Y}_i) \neq \mathrm{E}(Y_i)$, where $\hat{Y}_i = \hat{\boldsymbol{\beta}}_p' \boldsymbol{x}_i$. The bias may be tolerable if it is offset by a reduced variance. Therefore, Mallows (1973) considered the mean square error for each fitted value, and defined the following criterion for model evaluation:

$$\begin{aligned} J_p &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{mse}(\hat{Y}_i) \\ &= \frac{1}{\sigma^2} \mathrm{E} \left[\sum_{i=1}^{n} (\hat{Y}_i - \mathrm{E}(Y_i))^2 \right]. \end{aligned} \tag{4.3}$$

The value of J_p has to be estimated from the data. Mallows (1973) proposed the following estimate:

$$C_p = \widehat{J}_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n, \tag{4.4}$$

where $\hat{\sigma}^2$ is the estimate of σ^2 from the full model. It can be shown that, for the full model with $k = m + 1$ parameters, $C_k = k$. It is interesting to note that, for normal errors, the C_p statistic is equivalent to $\mathrm{AIC}(p, 2)$ (see (4.1)).

4.2.3 Final Prediction Error (FPE)

Akaike (1969, 1970) proposed a criterion for the selection of predictors in the context of autoregressive processes. The author minimized an estimate of the expected squared

error in predicting the observations that are independent of the available data, but have the same distribution.

Consider the subset model (4.2). Suppose that we are trying to predict, using the estimates $\hat{\boldsymbol{\beta}}_p$, the values Y_i^* satisfying

$$Y_i^* = \boldsymbol{\beta}_p' \boldsymbol{x}_i + \epsilon_i^*, \; i = 1, 2, \cdots, n, \tag{4.5}$$

where ϵ_i^*'s have the same distribution F, but they are independent of the ϵ_i's. The Final Prediction Error (FPE) of the current model is defined as

$$\text{FPE} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{E}\left[(Y_i^* - \hat{\boldsymbol{\beta}}_p' \boldsymbol{x}_i)^2 \right]. \tag{4.6}$$

It is interesting to note that, for the linear regression setup considered above

$$\begin{aligned} \text{FPE} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{E}\left[(Y_i^* - \text{E}(Y_i^*) + \text{E}(Y_i) - \hat{\boldsymbol{\beta}}_p' \boldsymbol{x}_i)^2 \right] \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \sigma^2 + \frac{1}{\sigma^2} \text{E}\left[\sum_{i=1}^{n} (\hat{Y}_i - \text{E}(Y_i))^2 \right] \\ &= J_p + n, \end{aligned}$$

where J_p is defined in (4.3). Therefore, based on (4.4), an estimate of FPE is given by

$$\widehat{\text{FPE}} = \frac{RSS_p}{\hat{\sigma}^2} + 2p, \tag{4.7}$$

where $\hat{\sigma}^2$ is the estimate of σ^2 from the full model. Note that, for the evaluation of linear prediction models, $\widehat{\text{FPE}}$ is equivalent to the C_p statistic.

4.2.4 Cross-validation

Cross-validation (CV) obtains an estimate of the error-rate of a prediction rule by splitting the n data points into a training sample of size n_t (used for fitting the predic-

tion model, i.e., for estimating the model parameters) and a validation sample of size $n_v = n - n_t$ (used for assessing the model). We calculate the average prediction error based on all or some of the $\binom{n}{n_v}$ different validation samples, and use it as a criterion to select a prediction model. It is often called leave-n_v-out cross-validation, or CV(n_v).

The vast majority of papers on this topic deals with leave-one-out cross-validation, denoted by CV(1). Lachenbruch and Mickey (1968) proposed the use of CV(1) in discriminant analysis. The method is furthered by Allen (1974), Stone (1974), and Geisser (1975). The asymptotic equivalence of CV(1) and AIC is shown by Stone (1977).

Efron (1983) used CV(1) to estimate the error rate of a prediction rule in the situation where the response Y is dichotomous. We can easily generalize the author's approach to a continuous response. Suppose that we have an $n \times p$ dataset

$$Z = \{\boldsymbol{z}_i, i = 1, 2, \cdots, n\},$$

where each case (row) $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ is an observation of the random quantity (X, Y), with X being a row-vector of $(p-1)$ covariates and Y being a real-valued response variable. The dataset Z is a random sample from distribution H on the p-dimensional sample space $\mathbb{R}^p$.

We want to evaluate a prediction rule $\eta(\boldsymbol{x}, Z)$ constructed based on the given dataset. An example of $\eta(\boldsymbol{x}, Z)$ is $\hat{\boldsymbol{\beta}}_Z'\boldsymbol{x}$ where $\hat{\boldsymbol{\beta}}_Z$ is the linear regression coefficient of Y on X. We want to estimate the error rate of $\eta(\boldsymbol{x}, Z)$ when $\eta(\boldsymbol{x}_0, Z)$ is used to predict a future value y_0 of Y for a given predictor value $\boldsymbol{x}_0$. Let $Q[y_0, \eta(\boldsymbol{x}_0, Z)]$ denote the "error" in predicting y_0 from $\boldsymbol{x}_0$. For example, we can consider the squared error

$$Q[y_0, \eta(\boldsymbol{x}_0, Z)] = (y_0 - \eta(\boldsymbol{x}_0, Z))^2. \tag{4.8}$$

True error rate

The true error rate $\mathrm{Err}(Z, H)$ of the prediction rule $\eta(\boldsymbol{x}_0, Z)$ can be defined as

$$\mathrm{Err}(Z, H) = E_H \left(Q[Y_0, \eta(X_0, Z)] \right), \tag{4.9}$$

the expectation being taken over $(X_0, Y_0) \sim H$ with Z fixed at its observed value.

Apparent error rate

The most obvious estimate of the true error rate $\mathrm{Err}(Z, H)$ is the apparent error rate $\overline{\mathrm{err}}(Z, H)$:

$$\overline{\mathrm{err}}(Z, H) = \frac{1}{n} \sum_{i=1}^{n} Q[y_i, \eta(\boldsymbol{x}_i, Z)], \tag{4.10}$$

which usually underestimates $\mathrm{Err}(Z, H)$, because the same data have been used both to construct and to evaluate the prediction rule $\eta(\boldsymbol{x}, Z)$.

CV error rate

CV attempts to overcome the problem of underestimation of $\mathrm{Err}(Z, H)$ by dividing the given dataset into the training and the validation parts. For CV(1), let $Z_{(i)}$ be the training set with case $\boldsymbol{z}_i$ removed, and $\eta(\boldsymbol{x}, Z_{(i)})$ be the corresponding prediction rule. The CV(1) estimate of $\mathrm{Err}(Z, H)$ is given by

$$\widehat{\mathrm{Err}}^{(\mathrm{CV})} = \frac{1}{n} \sum_{i=1}^{n} Q[y_i, \eta(\boldsymbol{x}_i, Z_{(i)})]. \tag{4.11}$$

Shao (1993) used CV(n_v) for model selection in regression using a random selection of the $\binom{n}{n_v}$ possible validation samples.

4.2.5 Bootstrap

Efron (1983) used bootstrap to estimate the true error rate $\mathrm{Err}(Z, H)$ (see (4.9)). Since the apparent error rate $\bar{\mathrm{err}}(Z, H)$ (see (4.10)) is an underestimate of $\mathrm{Err}(Z, H)$, a correction is required. Let $\mathrm{op}(Z, H)$ be defined as

$$\mathrm{op}(Z, H) = \mathrm{Err}(Z, H) - \bar{\mathrm{err}}(Z, H). \tag{4.12}$$

The expectation of $\mathrm{op}(Z, H)$, denoted by $\mathrm{w}(H)$, is given by

$$\mathrm{w}(H) = E_H\big\{\mathrm{Err}(Z, H) - \bar{\mathrm{err}}(Z, H)\big\}, \tag{4.13}$$

which could be the ideal correction if it were known. Note that, though the true error rate and the apparent error rate are defined for particular dataset Z, the target correction is the expectation over all datasets. It is not easy to find an estimate for (4.12) which is defined for a particular Z.

The unknown $\mathrm{w}(H)$ can be estimated using the bootstrap procedure to get the bootstrap estimate of $\mathrm{Err}(Z, H)$ as

$$\widehat{\mathrm{Err}}^{(\mathrm{Boot})} = \bar{\mathrm{err}} + \hat{\mathrm{w}}^{(\mathrm{Boot})}. \tag{4.14}$$

To obtain $\hat{\mathrm{w}}^{(\mathrm{Boot})}$, let Z^* be a bootstrap sample, i.e., a random sample of size n from $\hat{H}$. Based on (4.13), $\hat{\mathrm{w}}^{(\mathrm{Boot})}$ can be written as

$$\begin{aligned}
\hat{\mathrm{w}}^{(\mathrm{Boot})} &= E_*\left\{\mathrm{Err}(Z^*, \hat{H})\right\} - E_*\left\{\bar{\mathrm{err}}(Z^*, \hat{H})\right\} \\
&= E_*\left(\frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta(\boldsymbol{x}_i, Z^*)]\right) - E_*\left(\sum_{i=1}^{n} P_i^* Q[y_i, \eta(\boldsymbol{x}_i, Z^*)]\right),
\end{aligned} \tag{4.15}$$

where E_* is the expectations over all bootstrap samples, and P_i^* is the proportion of times a particular case $\boldsymbol{z}_i$ occurs in the bootstrap sample Z^*, i.e.,

$$P_i^* = \frac{\#\{\boldsymbol{z}_j^* = \boldsymbol{z}_i\}}{n},\ i = 1, 2, \cdots, n.$$

The expression inside the first pair of parentheses of (4.15) suggests that the prediction rule be constructed with the bootstrap sample Z^*, and an average error of this rule be calculated on the given dataset Z. The expression inside the second pair of parentheses of (4.15) suggests that the prediction rule be constructed with the bootstrap sample Z^*, and an average error of this rule be calculated on the same bootstap sample Z^*.

4.3 Review: robust selection criteria

In this section we present the robust counterparts of the classical selection criteria AIC, C_p and FPE (in order of appearance in the robustness literature), and discuss their limitations. We discuss robust counterparts of cross-validation and bootstrap procedures in Section 4.4 and Section 4.5, respectively.

4.3.1 Robust AIC

Ronchetti (1985) proposed a robust counterpart of the AIC statistic. The extension of AIC to AICR is inspired by the extension of maximum likelihood estimation to M-estimation. The author derived AICR for an error distribution with density

$$f(\epsilon) = K \exp(-\chi(\epsilon)). \tag{4.16}$$

For a given constant α and a given function χ, we can choose the model that minimizes

$$\text{AICR}(p, \alpha, \chi) = 2 \sum_{i=1}^{n} \chi(r_i) + \alpha p, \tag{4.17}$$

where $r_i = (Y_i - \hat{\boldsymbol{\beta}}_p' \boldsymbol{x}_i)/\hat{\sigma}$, $\hat{\sigma}$ is some robust estimate of σ, and $\hat{\boldsymbol{\beta}}_p'$ is the M-estimator defined as the implicit solution of

$$\sum_{i=1}^{n} \psi(r_i)\boldsymbol{x}_i = \mathbf{0},$$

with $\psi = \chi'$. The author also proposed a choice for the parameter α, which is given by

$$\alpha = 2\,\mathrm{E}\left[\psi^2(\epsilon)\right] / \mathrm{E}\left[\psi'(\epsilon)\right]. \tag{4.18}$$

Limitation. The author considered that the M-estimate was the maximum likelihood estimate for the density in (4.16). Unfortunately, this only hold for unbounded χ functions, and in such cases the breakdown point of the M-estimate is 0.

4.3.2 Robust C_p

Ronchetti and Staudte (1994) pointed out the sensitivity of the classical C_p to outlying points, and proposed a robust C_p statistic denoted by RC_p. Consider an M-estimator $\hat{\boldsymbol{\beta}}_p$ with weights $\hat{w}_i = \psi(r_i)/r_i$, where r_i is the residual for the ith observation. Using these weights, the author defined a weighted version of J_p (see (4.3)) as follows

$$\Gamma_p = \frac{1}{\sigma^2}\mathrm{E}\left[\sum_{i=1}^{n} \hat{w}_i^2(\hat{Y}_i - \mathrm{E}(Y_i))^2\right].$$

The author proposed the estimate of Γ_p, i.e., the robust version of C_p (see (4.4)), as

$$\mathrm{RC_p} = \frac{W_p}{\hat{\sigma}^2} - (U_p - V_p), \tag{4.19}$$

where $W_p = \sum \hat{w}_i^2 r_i^2$ is the weighted residual sum of squares, $\hat{\sigma}^2$ is a robust and consistent estimate of σ^2 from the full model, and U_p and V_p are constants depending on the weight function and the number of parameters p.

When the weights are identically 1, RC_p reduces to Mallows C_p.

4.3.3 Robust FPE

The robust analogue to the classical FPE criterion is proposed by Yohai (1997). Let s be an estimate of the scale σ from the full model, and $\hat{\boldsymbol{\beta}}_p$ be the M-estimator of the particular model under consideration.

$$\hat{\boldsymbol{\beta}}_p = \operatorname{argmin} \sum_{i=1}^{n} \chi((y_i - \boldsymbol{\beta}_p' \boldsymbol{x}_i)/s). \tag{4.20}$$

When we are trying to predict y_i^* (equation 4.5) using the estimate $\hat{\boldsymbol{\beta}}_p$, the robust FPE (RFPE) is defined as

$$\text{RFPE} = \sum_{i=1}^{n} \text{E}\left[\chi\left(\frac{y_i^* - \hat{\boldsymbol{\beta}}_p' \boldsymbol{x}_i}{\sigma}\right)\right], \tag{4.21}$$

where the expectations are taken in the y_i^*'s as well as in $\hat{\boldsymbol{\beta}}_p$. Note that when $\chi(u) = u^2$, RFPE reduces to the classical FPE.

Using second order Taylor expansions with respect to $\boldsymbol{\beta}_p$ (assuming that the current model is the true model), RFPE is expressed as

$$\text{RFPE} \simeq \text{E}\left(\sum_{i=1}^{n} \chi\left(\frac{y_i - \hat{\boldsymbol{\beta}}' \boldsymbol{x}_i}{\sigma}\right)\right) + p\frac{A}{B}, \tag{4.22}$$

where $A = \text{E}(\psi^2(\epsilon/\sigma))$, $B = \text{E}(\psi'(\epsilon/\sigma))$, and $\psi = \chi'$. Therefore, an estimate of RFPE is given by

$$\widehat{\text{RFPE}} \simeq \sum_{i=1}^{n} \chi\left(\frac{y_i - \hat{\boldsymbol{\beta}}' \boldsymbol{x}_i}{s}\right) + p\frac{\hat{A}}{\hat{B}}, \tag{4.23}$$

where $\hat{A} = n^{-1} \sum_{i=1}^{n} (\psi^2(r_i/s))$, $B = n^{-1} \sum_{i=1}^{n} (\psi'(r_i/s))$, and $r_i = y_i - \hat{\boldsymbol{\beta}}_p' \boldsymbol{x}_i$.

The performance of RFPE has not been studied so far. In Section 4.6 we carry out a simulation study to evaluate RFPE.

4.4 Robust cross-validation

Ronchetti, Field and Blanchard (1997) proposed a robust cross-validation procedure which is a robust version of the cross-validation method proposed by Shao (1993). The authors used estimators that have optimal bounded influence for prediction. However, their method is computationally expensive. Hubert and Engelen (2004) proposed a fast cross-validation method in the context of robust covariance estimation with MCD and robust principal component analysis.

In this section we propose a robust CV procedure which is computationally suitable. First, let us consider a simple robustification of the CV procedure achieved by (a) constructing a robust prediction rule, denoted by $\eta^{\mathrm{R}}(\boldsymbol{x}, Z)$, based on the given dataset Z, and (b) calculating robust summary statistics of the prediction errors $Q[y_i, \eta^{\mathrm{R}}(\boldsymbol{x}_i, Z_{(i)})]$. For the construction of a robust prediction rule, we consider the regression MM-estimates proposed by Yohai (1987) because of its high breakdown point and high efficiency at the normal model. This estimate is defined as follows.

Definition 4.1. (Regression MM-estimate) *Let $\chi_0 : \mathbb{R} \to \mathbb{R}$ and $\chi_1 : \mathbb{R} \to \mathbb{R}$ be two score functions such that $\chi_0(u) \leq \chi_1(u)$, $u \in \mathbb{R}$, and each χ satisfies the following set of regularity conditions:*

1. $\chi(-u) = \chi(u)$, $u \in \mathbb{R}$,
2. χ *is non-decreasing on* $[0, \infty)$,
3. $\chi(0) = 0$, *and* $\chi(\infty) = 1$,
4. χ *is continuously differentiable.*

Let $\tilde{\boldsymbol{\beta}}$ be a high-breakdown-point "initial" estimate for $\boldsymbol{\beta}$, and $\hat{\sigma}$ be the estimate of scale of the residuals based on $\tilde{\boldsymbol{\beta}}$ satisfying

$$\frac{1}{n}\sum_{i=1}^{n}\chi_0\left((y_i - \boldsymbol{x}_i^t\tilde{\boldsymbol{\beta}})/\hat{\sigma}\right) = b, \tag{4.24}$$

where $b \in (0,1]$ is the expectation of $\chi_0(.)$ under the central model. Then, the regression MM-estimate $\hat{\boldsymbol{\beta}}$ is defined as the solution of

$$\sum_{i=1}^{n}\chi_1'\left((y_i - \boldsymbol{x}_i^t\hat{\boldsymbol{\beta}})/\hat{\sigma}\right)\boldsymbol{x}_i = \mathbf{0}. \tag{4.25}$$

A reasonable choice for the initial estimate $\tilde{\boldsymbol{\beta}}$ in Definition 4.1 is the regression S-estimate proposed by Rousseeuw and Yohai (1984) because of its high breakdown point. This estimate is defined as follows.

Definition 4.2. (Regression S-estimate) *Let $\chi_0 : \mathbb{R} \to \mathbb{R}$ be the score function described above. The regression S-estimate $\tilde{\boldsymbol{\beta}}$ is defined as*

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ \hat{\sigma}(\boldsymbol{\beta}), \tag{4.26}$$

where $\hat{\sigma}(\boldsymbol{\beta})$ solves

$$\frac{1}{n}\sum_{i=1}^{n}\chi_0\left((y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})/\hat{\sigma}(\boldsymbol{\beta})\right) = b. \tag{4.27}$$

The corresponding S-estimate of scale, $\hat{\sigma}$, is given by

$$\hat{\sigma} = \inf_{\boldsymbol{\beta}}\hat{\sigma}(\boldsymbol{\beta}) = \hat{\sigma}(\tilde{\boldsymbol{\beta}}). \tag{4.28}$$

For a robust summary of the (squared) prediction errors we will use the trimmed means with different amounts of trimming. The α-trimmed mean of X, denoted by $m_\alpha(X)$, is the sample mean obtained after dropping the largest $100\alpha\%$ observations of

X. Let $U_1 < U_2 < \cdots < U_n$ be the ordered observations of X, and $k = [n(1-\alpha)]$, where $[n(1-\alpha)]$ means the integer part of $n(1-\alpha)$. Then,

$$m_\alpha(X) = \frac{1}{n-k}\sum_{j=1}^{n-k} U_j. \tag{4.29}$$

The robust counterpart of $\widehat{\text{Err}}^{(\text{CV})}$ is now given by

$$\widehat{\text{Err}}^{(\text{RCV})} = m_\alpha\left(Q[y_i, \eta^{\text{R}}(\boldsymbol{x}_i, Z_{(i)})]\right). \tag{4.30}$$

Note that $\widehat{\text{Err}}^{(\text{RCV})}$ does not estimate $\text{Err}(Z, H)$ (see equation 4.9). Instead, it estimates

$$\begin{aligned} \text{Err}_{\text{R}}(Z, H) &= E_H^\alpha\left(Q[y_0, \eta^{\text{R}}(\boldsymbol{x}_0, Z)]\right) \\ &= \frac{1}{1-\alpha}\int_0^{H^{-1}(1-\alpha)} Q[y_0, \eta^{\text{R}}(\boldsymbol{x}_0, Z)]dH. \end{aligned} \tag{4.31}$$

The use of α-trimmed mean will help us identify the robust model(s) that can be expected to predict $100(1-\alpha)\%$ of the future data better than other models.

4.4.1 Dealing with numerical complexity

The computation of the MM estimates of regression for each training sample, i.e., the computation of $\hat{\boldsymbol{\beta}}_{(i)}$, $i = 1, 2, \cdots, n$, is very computer intensive. We propose to remedy this problem as follows. We express the MM estimates of regression based on all the observations on the current set of covariates as a weighted least squares fit, and obtain the weighted least squares fit for each training sample by using the selected cases and their corresponding weights. We elaborate the proposed method below.

Let $r_i = y_i - \hat{\boldsymbol{\beta}}^t \boldsymbol{x}_i$ be the residuals obtained from the fit $\hat{\boldsymbol{\beta}}$ (based on all the observations on the current set of covariates). Once the robust fit is complete, $\hat{\boldsymbol{\beta}}$ can be

expressed as a weighted least squares regression coefficient as follows:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} w_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^t \right)^{-1} \sum_{i=1}^{n} w_i \, \boldsymbol{x}_i \, y_i, \tag{4.32}$$

with the weights w_i expressed as

$$w_i = \chi_1' \left(r_i / \hat{\sigma} \right) / r_i, \; i = 1, 2, \cdots, n. \tag{4.33}$$

For further computational ease, we will assume that $\hat{\sigma}_{(i)} \simeq \hat{\sigma}$, where $\hat{\sigma}_{(i)}$ is the S-scale based on the training sample $Z_{(i)}$. Now, a computationally suitable version of the regression MM-estimate $\hat{\boldsymbol{\beta}}_{(i)}$ can be calculated as

$$\hat{\boldsymbol{\beta}}_{(i)}^{(0)} = \left(\sum_{j \neq i}^{n} w_j \, \boldsymbol{x}_j \, \boldsymbol{x}_j^t \right)^{-1} \sum_{j \neq i}^{n} w_j \, \boldsymbol{x}_j \, y_j. \tag{4.34}$$

Note that no robust fitting is needed for the calculation of $\hat{\boldsymbol{\beta}}_{(i)}^{(0)}$.

One-step adjustment

Based on a small simulation study (not presented here), we consider a one-step correction to $\hat{\boldsymbol{\beta}}_{(i)}^{(0)}$ to make it closer to $\hat{\boldsymbol{\beta}}_{(i)}$. Let $r_j^{(1)} = y_j - \boldsymbol{x}_j^t \hat{\boldsymbol{\beta}}_{(i)}^{(0)}$, $j = 1, 2, \cdots, i-1, i+1, \cdots, n$. The updated set of weights $w_j^{(1)}$ can be expressed as

$$w_j^{(1)} = \chi_1' \left(r_j^{(1)} / \hat{\sigma} \right) / r_j^{(1)}, \; j = 1, 2, \cdots, i-1, i+1, \cdots, n. \tag{4.35}$$

Thus, an adjusted estimate of $\boldsymbol{\beta}_{(i)}$ is given by

$$\hat{\boldsymbol{\beta}}_{(i)}^{(1)} = \left(\sum_{j \neq i}^{n} w_j^{(1)} \, \boldsymbol{x}_j \, \boldsymbol{x}_j^t \right)^{-1} \sum_{j \neq i}^{n} w_j^{(1)} \, \boldsymbol{x}_j \, y_j. \tag{4.36}$$

4.5 Robust bootstrap

For the purpose of making robust statistical inferences about the linear regression coefficient $\boldsymbol{\beta}$, Salibián-Barrera (2000), and Salibián-Barrera and Zamar (2002) developed the robust bootstrap procedure. The author(s) considered the regression MM-estimate $\hat{\boldsymbol{\beta}}$, and generated a large number of re-calculated $\hat{\boldsymbol{\beta}}^*$'s to estimate the asymptotic covariance matrix and the distribution function of the robust estimate $\hat{\boldsymbol{\beta}}$. This robust procedure is computationally suitable, because a linear system of equations is solved for each bootstrap sample.

We propose to use a similar approach to develop a computationally suitable robust counterpart of the bootstrap estimate $\widehat{\text{Err}}^{(\text{Boot})}$ of the true prediction error $\text{Err}(Z, H)$. Let $\hat{\boldsymbol{\beta}}$ be the MM-estimate, $\tilde{\boldsymbol{\beta}}$ be the (initial) S-estimate and $\hat{\sigma}$ be the S-scale. The robust counterpart of the apparent error rate $\bar{\text{err}}(Z)$ (see equation 4.10) is given by

$$\bar{\text{err}}_{\text{R}}(Z) = m_\alpha \left(Q[y_i, \eta(\boldsymbol{x}_i, Z)] \right), \tag{4.37}$$

where $m_\alpha(.)$ is the α-trimmed mean defined before, and $\eta^{\text{R}}(\boldsymbol{x}_i, Z)$ uses the MM-estimate $\hat{\boldsymbol{\beta}}$. Let r_i and $\tilde{r}_i$ be the residuals associated with the MM- and S-estimates, respectively. Once the robust fit is complete, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ can be expressed as a weighted least squares fit. Equation (4.32) shows the weighted average representation of $\hat{\boldsymbol{\beta}}$ with the weights w_i defined in Equation (4.33). The scale estimate $\hat{\sigma}$ can be expressed as

$$\hat{\sigma} = \sum_{i=1}^{n} v_i \, (y_i - \tilde{\boldsymbol{\beta}}^t \boldsymbol{x}_i), \tag{4.38}$$

with the weights v_i defined as

$$v_i = \frac{\hat{\sigma}}{nb} \chi_0 \left(\tilde{r}_i / \hat{\sigma} \right) / \tilde{r}_i, \; i = 1, 2, \cdots, n. \tag{4.39}$$

Let $Z^* = \{(\boldsymbol{x}_i, y_i),\, i = 1, 2, \cdots, n\}$ be a bootstrap sample from Z. The unadjusted bootstrap estimates can be calculated as

$$\hat{\boldsymbol{\beta}}_u^* = \left(\sum_{i=1}^{n} w_i^* \boldsymbol{x}_i^* \boldsymbol{x}_i^{*t}\right)^{-1} \sum_{i=1}^{n} w_i^* \boldsymbol{x}_i^* y_i^*, \tag{4.40}$$

$$\hat{\sigma}_u^* = \sum_{i=1}^{n} v_i^* (y_i^* - \tilde{\boldsymbol{\beta}}^t \boldsymbol{x}_i^*), \tag{4.41}$$

where $w_i^* = w_j$ and $v_i^* = v_j$ when $(\boldsymbol{x}_i^*, y_i^*) = (\boldsymbol{x}_j, y_j)$. The corrected bootstrap estimate $\hat{\boldsymbol{\beta}}^*$ can be obtained as

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} + M(\hat{\boldsymbol{\beta}}_u^* - \hat{\boldsymbol{\beta}}) + \boldsymbol{d}(\hat{\sigma}_u^* - \hat{\sigma}), \tag{4.42}$$

where M and $\boldsymbol{d}$ are the linear correction factors (see Salibián-Barrera and Zamar 2002). The robust prediction rules $\eta^{\mathrm{R}}(\boldsymbol{x}, Z^*)$ can be based on the $\hat{\boldsymbol{\beta}}^*$ above. Now the robust counterpart of $\widehat{\mathrm{w}}^{(\mathrm{Boot})}$ (see equation 4.15) is given by

$$\widehat{\mathrm{w}}^{(\mathrm{RBoot})} = E_* \left\{ m_\alpha \left(Q[y_i, \eta^{\mathrm{R}}(\boldsymbol{x}_i, Z^*)] \right) \right\} - E_* \left\{ m_\alpha \left(Q[y_i^*, \eta^{\mathrm{R}}(\boldsymbol{x}_i^*, Z^*)] \right) \right\}. \tag{4.43}$$

Finally, the robust bootstrap estimate of $\mathrm{Err}_{\mathrm{R}}(Z, H)$ (see equation 4.31) can be expressed as

$$\widehat{\mathrm{Err}}^{(\mathrm{RBoot})} = \bar{\mathrm{err}}_{\mathrm{R}}(Z) + \widehat{\mathrm{w}}^{(\mathrm{RBoot})}. \tag{4.44}$$

4.6 Simulation study

At first, we carry out a small simulation (Section 4.6.1) to show that the classical CV and bootstrap estimates of true error rate (see (4.9)) are sensitive to outliers while the robust estimates are resistant to outliers. We then conduct another study (Section 4.6.2) where we use these methods along with FPE and RFPE to select the "best" models, and

compare the predictive powers of these models. Since AIC and C_p are equivalent to FPE for linear regression setup with normal errors, and robust AIC and robust C_p have some limitations, we do not consider these criteria for our simulation study.

4.6.1 Robustness of the estimates

We evaluate the 4 estimates $\widehat{\text{Err}}^{(\text{CV})}$, $\widehat{\text{Err}}^{(\text{Boot})}$, $\widehat{\text{Err}}^{(\text{RCV})}$ and $\widehat{\text{Err}}^{(\text{RBoot})}$ using simulated clean and contaminated datasets. Since the true error rates $\text{Err}(Z, H)$ and $\text{Err}_{\text{R}}(Z, H)$ are different (the latter uses the trimmed mean), we multiply the robust estimates by

$$\lambda = \frac{E_{H_0}\left(Q[y, \eta^{\text{R}}(\boldsymbol{x}, Z)]\right)}{E^{\alpha}_{H_0}\left(Q[y, \eta^{\text{R}}(\boldsymbol{x}, Z)]\right)}$$

to make the results more comparable with the classical results.

We considered two standard normal covariates X_1 and X_2, and generated $Y = 2\,X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0,\, 4)$. We simulated 100 datasets, and for each dataset we calculated the estimates mentioned above. We then contaminated each dataset as follows. Each of the 3 variables (2 covariates and the response) is contaminated independently. Each observation of a variable is assigned probability 0.03 of being replaced by a large number. Therefore, the probability that any particular row of the dataset will be contaminated is $1 - (1 - 0.03)^3$, which means approximately 9% of the rows will be contaminated. For each contaminated dataset we obtained the 4 estimates mentioned above. Table 4.1 presents the results for the first 10 trials.

The average $\text{Err}(Z, H_0)$ (the average true error rate for the clean data) is 4.12, while the average $\text{Err}_{\text{R}}(Z, H_0)$ (multiplied by λ) is 4.13. Table 4.1 shows that, for the clean data both the classical and robust methods estimate the true error rates very well. However, in

Table 4.1: First 10 trials: classical and robust estimates of prediction errors.

Trial	CV		Boot		RCV		RBoot	
	Clean	Contam	Clean	Contam	Clean	Contam	Clean	Contam
1	4.60	12.01	4.61	11.65	4.18	5.80	4.15	5.47
2	4.05	10.20	4.02	8.99	3.48	5.18	3.42	5.05
3	3.86	10.35	3.88	10.64	3.70	5.72	3.67	5.66
4	4.95	11.56	4.97	11.80	5.45	6.14	5.43	6.47
5	3.95	14.92	3.94	11.77	4.29	5.57	4.25	5.33
6	4.54	12.56	4.52	10.91	5.11	6.37	5.07	6.30
7	5.22	10.28	5.16	10.53	4.72	6.37	4.74	6.65
8	4.03	8.54	4.04	8.63	4.04	5.43	4.04	5.45
9	4.16	10.45	4.20	10.60	4.21	6.59	4.21	6.36
10	4.57	9.82	4.53	9.72	4.75	5.90	4.70	6.54
...	...	...	...	...	...	...	...	...
mean	4.14	10.94	4.13	10.28	4.17	5.32	4.16	5.22
(sd)	(0.58)	(2.57)	(0.58)	(2.13)	(0.67)	(0.94)	(0.69)	(0.93)

the contaminated data, robust estimates perform much better than the classical methods.

4.6.2 Final model selection

In this simulation study we use the classical segmentation methods CV, Boot and FPE along with their robust counterparts RCV, RBoot and RFPE to select the "best" models, and compare the predictive powers of these models. The study is similar to Frank and

Friedman (1993). We considered 2 latent variables L_i, $i = 1, 2$, to generate $Y = 6L_1 + 5L_2 + \epsilon$, where $L_i \sim N(0, 1)$, and ϵ is a normal error not related to the latent variables. We considered a total of $m = 8$ covariates. Of them, $a = 4$ are related to the two latent variables, with 2 covariates related to L_1 and the other two related to L_2.

We generated 100 datasets each of which was randomly divided into a training sample of size 100 and a test sample of size 100. Each training dataset was then contaminated as follows. A number of rows (10%) were chosen randomly, and for these rows the covariates values were replaced by large positive numbers while the response values were replaced by large negative numbers.

We used all 6 methods on the clean and contaminated training data to select and fit the final models, and then used them to predict the test data outcomes. For each simulated dataset, we recorded the number of noise variables in the model, and the average squared prediction error on the test sample.

Table 4.2 shows the average test error and the average number of noise variables selected by each method. For the clean data, the robust methods perform as good as the classical methods. For the contaminated data, robust methods produce much smaller test errors than the classical methods. Also, robust models contain less noise variables. The performance of the three robust methods are similar.

Table 4.2: Performance of the classical and robust methods of segmentation (evaluation of all possible subsets of the reduced set).

Method		Test error		Noise	
		Clean	Contam	Clean	Contam
Classical	CV	41.81	56.49	0.00	0.60
	Boot	41.32	54.88	0.00	0.50
	FPE	41.93	55.17	0.02	0.60
Robust	RCV	42.97	43.62	0.06	0.08
	RBoot	41.59	44.80	0.08	0.08
	RFPE	42.73	44.91	0.06	0.06

4.7 Examples

In this section we use two real datasets to evaluate the performance of the classical and robust methods for the segmentation of the reduced set. Both of these datasets were used in Chapter 3 for the evaluation of robust sequencing.

4.7.1 Demographic data

This dataset contains $n = 50$ obsrvations on $d = 25$ covariates and a response. For more details Section 3.8 is referred to. Using the learning curve based on standard LARS, we selected the reduced set $(22, 20, 4, 15, 25, 2, 14, 5, 3, 17, 24, 23)$. The robust bootstrapped LARS produced the reduced set $(22, 20, 4, 15, 10, 2, 19, 25, 8, 18, 6, 24)$.

We applied the classical segmentation methods CV, Boot and FPE on the first reduced set above. The covariates selected by these methods are $(22, 4, 25, 2, 14, 17, 24, 23)$, $(22, 4, 15, 25, 2, 17, 24)$, and $(22, 20, 4, 15, 25, 2, 14, 17, 24, 23)$, respectively. We then applied the robust methods RCV, RBoot and RFPE on the second reduced set. The covariates selected are $(22, 4, 15, 10, 19, 25, 18, 24)$, $(22, 20, 4, 10, 19, 25, 18, 24)$, and $(15, 6, 24)$, respectively. Interestingly, RFPE selects a very small model compared to others.

To compare the models obtained by the classical and robust methods, we used the clean data (dropping one clear outlier) to estimate the prediction errors of these models 1000 times using 5-fold CV. The mean prediction errors for the models are: CV 199.3, Boot 198.2, FPE 207.6, RCV 195.8, RBoot 197.5 and RFPE 246.9. The robust method RCV performs slightly better than RBoot, and both of them perform much better than the classical methods and RFPE.

4.7.2 Protein data

This KDD-Cup 2004 dataset was used in Section 3.8. We considered $n = 4141$ protein sequences from 5 blocks. The number of covariates is $d = 77$, with 4 block indicators (variables $1 - 4$) and 73 features. The data were split to get a training sample of size $n = 2072$ and a test sample of size $n = 2069$.

We considered a reduced set of size 5 using the learning curve based on standard LARS on the training data, which contains the covariates $(14, 13, 5, 8, 7)$. Robust bootstrapped LARS gives the reduced set $(14, 13, 5, 76, 73)$. We applied the 3 classical methods of segmentation on the first reduced set. They all select the same model, and it

includes the covariates $(14, 13, 5, 8)$. The robust methods used on the second reduced set select the covariates $(14, 13, 5, 76)$.

We fitted the 2 models using the training data, and then used them to predict the test data outcomes. The 1%, 5% and 10% trimmed means of prediction errors for the robust (classical) model are : 114.92 (117.49), 92.77 (95.66) and 74.82 (78.19), respectively. It is encouraging to note that the robust methods outperform the classical methods for the majority of the data.

4.8 Conclusion

The main contribution of this chapter is that we developed computationally suitable robust methods of segmentation (evaluation of all possible subsets of the reduced set obtained in Chapter 3) to select the final model.

Classical selection criteria FPE, AIC, C_p, CV and bootstrap are sensitive to outliers. We also identified certain limitations of Robust AIC (Ronchetti 1985) and robust CV (Ronchetti, Field and Blanchard 1997) methods. We proposed computationally suitable robust versions of CV and bootstrap procedures. We evaluated our methods using both simulated and real datasets, and compared them with the classical methods as well as robust FPE proposed by Yohai (1997). According to the simulation study, the performance of the three robust methods are similar, and better than the classical methods. In the real datasets, robust CV (RCV) and robust bootstrap (RBoot) have better performance compared to RFPE.

Chapter 5

Properties of Adjusted-Winsorized Correlation Estimate

5.1 Introduction

In Chapter 3 we proposed a new correlation estimate for bivariate data, which we called the adjusted-Winsorized estimate. Unlike two separate univariate Winsorizations for X and Y (Huber 1981 and Alqallaf 2003), we proposed a joint Winsorization with a larger tuning constant c_1 for the points falling in the two major quadrants, and a smaller constant c_2 for the points in the two minor quadrants.

In this chapter we will establish the consistency and derive the influence function of the proposed correlation estimate. We will then discuss the asymptotic normality of this estimate.

Definition 5.1. (Adjusted-Winsorization) *The adjusted-Winsorization of* $(u, v) \in \mathbb{R}^2$, *denoted by* $\Psi_{\boldsymbol{c}}(u, v)$ *with* $\boldsymbol{c} = (c_1, c_2)$, *is defined as*

$$\Psi_{\boldsymbol{c}}(u, v) = (\psi_{\boldsymbol{c}}(u), \psi_{\boldsymbol{c}}(v)) = \begin{cases} (\psi_{c_1}(u), \psi_{c_1}(v)), & uv \geq 0, \\ (\psi_{c_2}(u), \psi_{c_2}(v)), & uv < 0, \end{cases} \tag{5.1}$$

where ψ *is a non-decreasing symmetric function, and* c_1 *and* c_2 *are chosen constants.*

Definition 5.2. (Adjusted-Winsorized estimate of correlation) *Let* (X_i, Y_i), $i = 1, 2, \cdots, n$, *be a random sample from a bivariate distribution with location parameters* μ_X *and* μ_Y, *and scale parameters* σ_X *and* σ_Y, *respectively. Let* $\boldsymbol{\theta} = (\mu_X, \mu_Y, \sigma_X, \sigma_Y)$, *and* $\hat{\boldsymbol{\theta}} = (\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_X, \hat{\sigma}_Y)$ *be an estimate of* $\boldsymbol{\theta}$. *Denote* $\hat{U}_i = (X_i - \hat{\mu}_X)/\hat{\sigma}_X$, *and* $\hat{V}_i = (Y_i - \hat{\mu}_Y)/\hat{\sigma}_Y$. *Let* $\Psi_{\boldsymbol{c}}\left(\hat{U}_i, \hat{V}_i\right) = \left(\psi_{\boldsymbol{c}}(\hat{U}_i), \psi_{\boldsymbol{c}}(\hat{V}_i)\right)$ *be as defined in (5.1). Then, the adjusted-Winsorized estimate* $\hat{r}_w$ *of the correlation between* X *and* Y *is given by*

$$\hat{r}_w = \frac{\frac{1}{n}\sum_{i=1}^{n} \psi_{\boldsymbol{c}}(\hat{U}_i)\psi_{\boldsymbol{c}}(\hat{V}_i) - \left(\frac{1}{n}\sum_{i=1}^{n} \psi_{\boldsymbol{c}}(\hat{U}_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n} \psi_{\boldsymbol{c}}(\hat{V}_i)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \psi_{\boldsymbol{c}}^2(\hat{U}_i) - \left(\frac{1}{n}\sum_{i=1}^{n} \psi_{\boldsymbol{c}}(\hat{U}_i)\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} \psi_{\boldsymbol{c}}^2(\hat{V}_i) - \left(\frac{1}{n}\sum_{i=1}^{n} \psi_{\boldsymbol{c}}(\hat{V}_i)\right)^2}}. \tag{5.2}$$

For the validity of the results obtained in the subsequent sections, we need some assumptions on the functions ψ_{c_1} and ψ_{c_2} used for the adjusted-Winsorization of the data. Let $\psi : \mathbb{R} \to \mathbb{R}$ satisfy the following set of regularity conditions:

A1. $\psi(-u) = -\psi(u)$, $u \in \mathbb{R}$,

A2. ψ is non-decreasing,

A3. ψ is continuously differentiable,

A4. ψ, ψ' and $\psi'(u)u$ are bounded.

For the adjusted-Winsorization of the data, we will use the S-scales $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ defined in Chapter 4. Let us assume that the score function $\chi : \mathbb{R} \to \mathbb{R}$ used in the S-scales satisfy the following set of regularity conditions:

B1. $\chi(-u) = \chi(u)$, $u \in \mathbb{R}$, $\chi(0) = 0$, and $\chi(\infty) = 1$,

B2. χ is non-decreasing on $[0, \infty)$,

B3. χ is continuously differentiable,

B4. χ, χ' and $\chi'(u)u$ are bounded.

5.2 Consistency of adjusted-Winsorized estimate

The following theorem shows that under certain regularity conditions the adjusted-Winsorized correlation estimates are consistent, provided that the location and scale estimates are consistent.

Theorem 5.1. (Consistency of adjusted-Winsorized estimate) *Let* (X_i, Y_i), $i = 1, 2, \cdots, n$, *be a random sample from a bivariate distribution with location parameters* μ_X *and* μ_Y, *and scale parameters* σ_X *and* σ_Y, *respectively. Let* $U_i = (X_i - \mu_X)/\sigma_X$, *and* $V_i = (Y_i - \mu_Y)/\sigma_Y$ *be the standardized variables. Let* $\boldsymbol{\theta} = (\mu_X, \mu_Y, \sigma_X, \sigma_Y)$, *and* $\hat{\boldsymbol{\theta}} = (\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_X, \hat{\sigma}_Y)$ *be an estimate of* $\boldsymbol{\theta}$. *Then, if*

$$\hat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\theta},$$

then

$$\hat{r}_w \xrightarrow[n\to\infty]{P} r_w,$$

where $\hat{r}_w$ is the adjusted-Winsorized estimate of correlation between X and Y, and

$$r_w = \frac{E\left[\psi_c(U)\psi_c(V)\right] - E\left[\psi_c(U)\right]E\left[\psi_c(V)\right]}{\sqrt{E\left[\psi_c^2(U)\right] - \left(E\left[\psi_c(U)\right]\right)^2}\sqrt{E\left[\psi_c^2(V)\right] - \left(E\left[\psi_c(V)\right]\right)^2}}. \tag{5.3}$$

To prove this theorem, we need an extension of "Serfling's Lemma" (Serfling 1980, page 253).

Lemma 5.1. (Extension of Serfling's Lemma) *Let $Z_i = (X_i, Y_i)$, $i = 1, 2, \cdots, n$, be a sequence of independent random variables having an identical bivariate distribution with parameter vector $\boldsymbol{\theta}$. Let $g(\boldsymbol{z}, \boldsymbol{t}) : \mathbb{R}^2 \times \mathbb{R}^4 \to \mathbb{R}$ be continuous in $\boldsymbol{t}$ uniformly on $\boldsymbol{z} \in A^c(\boldsymbol{\theta}, \Delta)$ for all $\Delta > 0$, and $P(\boldsymbol{z} \in A(\boldsymbol{\theta}, \Delta)) \to 0$ as $\Delta \to 0$, with $P(\boldsymbol{z} \in A(\boldsymbol{\theta}, 0)) = 0$. Assume that $|g(\boldsymbol{z}, \boldsymbol{t})| < K$ for all $\boldsymbol{z} \in \mathbb{R}^2$. Let $\hat{\boldsymbol{\theta}}_n$ be a sequence of random vectors such that $\hat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n) \xrightarrow[n\to\infty]{P} E[g(Z, \boldsymbol{\theta})]. \tag{5.4}$$

Proof. We have to show that, for any given $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n) - E[g(Z, \boldsymbol{\theta})]\right| < \epsilon\right) = 1. \tag{5.5}$$

Now,

$$\left|\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n) - E[g(Z, \boldsymbol{\theta})]\right|$$

$$= \left|\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n) I(\boldsymbol{z}_i \in A(\boldsymbol{\theta}, \Delta)) - E[g(Z, \boldsymbol{\theta}) I(Z \in A(\boldsymbol{\theta}, \Delta))]\right.$$
$$\left. + \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n) I(\boldsymbol{z}_i \in A^c(\boldsymbol{\theta}, \Delta)) - E[g(Z, \boldsymbol{\theta}) I(Z \in A^c(\boldsymbol{\theta}, \Delta))]\right|$$

$$\leq \left|\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n) I(\boldsymbol{z}_i \in A(\boldsymbol{\theta}, \Delta))\right| + \left|E[g(Z, \boldsymbol{\theta}) I(Z \in A(\boldsymbol{\theta}, \Delta))]\right|$$
$$+ \left|\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n) I(\boldsymbol{z}_i \in A^c(\boldsymbol{\theta}, \Delta)) - E[g(Z, \boldsymbol{\theta}) I(Z \in A^c(\boldsymbol{\theta}, \Delta))]\right| \quad (5.6)$$

$$\leq \left|k\,\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{z}_i \in A(\boldsymbol{\theta}, \Delta))\right| + \left|k\,E[I(Z \in A(\boldsymbol{\theta}, \Delta))]\right|$$
$$+ \left|\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n) I(\boldsymbol{z}_i \in A^c(\boldsymbol{\theta}, \Delta)) - \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \boldsymbol{\theta}) I(\boldsymbol{z}_i \in A^c(\boldsymbol{\theta}, \Delta))\right|$$
$$+ \left|\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \boldsymbol{\theta}) I(\boldsymbol{z}_i \in A^c(\boldsymbol{\theta}, \Delta)) - E[g(Z, \boldsymbol{\theta}) I(Z \in A^c(\boldsymbol{\theta}, \Delta))]\right|, \quad (5.7)$$

$$= Q \text{ (say)}.$$

Note that the last expression in (5.6) is bounded by the sum of the last two expressions in (5.7). We will now deal with each of the four parts in (5.7).

As $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{z}_i \in A(\boldsymbol{\theta}, \Delta)) \to P(\boldsymbol{z} \in A(\boldsymbol{\theta}, \Delta))$, and $P(\boldsymbol{z} \in A(\boldsymbol{\theta}, 0)) = 0$. Therefore, for any given $\epsilon > 0$, there exists $\Delta > 0$ such that

$$\lim_{n\to\infty} P\left(k\,\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{z}_i \in A(\boldsymbol{\theta}, \Delta)) < \epsilon/4\right) = 1, \quad (5.8)$$

and

$$k\,E[I(Z \in A(\boldsymbol{\theta}, \Delta))] = k\,P(Z \in A(\boldsymbol{\theta}, \Delta)) < \epsilon/4. \quad (5.9)$$

We now focus on the third part of (5.7). Since $\hat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}$, we have, for any $\delta > 0$,

$$\lim_{n\to\infty} P\left(\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right\| < \delta\right) = 1. \quad (5.10)$$

Now, for any $\epsilon > 0$, we can choose $\delta = \delta(\Delta)$ (where Δ has been chosen before) such that

$$\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right\| < \delta \Longrightarrow \left|g(\boldsymbol{z}, \hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{z}, \boldsymbol{\theta})\right| < \epsilon/4, \quad \boldsymbol{z} \in A^c(\boldsymbol{\theta}, \Delta). \quad (5.11)$$

That is, δ is chosen in such a way that it will ensure the uniform continuity of $g(\boldsymbol{z}, \boldsymbol{\theta})$ in $\boldsymbol{\theta}$ on $\boldsymbol{z} \in A^c(\boldsymbol{\theta}, \Delta)$. Using (5.10) and (5.11), we have, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\left(g(\boldsymbol{z}, \hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{z}, \boldsymbol{\theta})\right) I(\boldsymbol{z} \in A^c(\boldsymbol{\theta}, \Delta))\right| < \epsilon/4\right) = 1,$$

which gives

$$\lim_{n\to\infty} P\Bigg(\Bigg|\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \hat{\boldsymbol{\theta}}_n) I(\boldsymbol{z}_i \in A^c(\boldsymbol{\theta}, \Delta)) - \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \boldsymbol{\theta}) I(\boldsymbol{z}_i \in A^c(\boldsymbol{\theta}, \Delta))\Bigg| < \epsilon/4\Bigg) = 1. \quad (5.12)$$

For the fourth part of (5.7), we can use the Weak Law of Large Numbers. For any $\epsilon > 0$,

$$\lim_{n\to\infty} P\Bigg(\Bigg|\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i, \boldsymbol{\theta}) I(\boldsymbol{z}_i \in A^c(\boldsymbol{\theta}, \Delta)) - E[g(Z, \boldsymbol{\theta}) I(Z \in A^c(\boldsymbol{\theta}, \Delta))]\Bigg| < \epsilon/4\Bigg) = 1. \quad (5.13)$$

Using inequalities (5.8), (5.9), (5.12) and (5.13) in (5.7), we have,

$$\lim_{n\to\infty} P\left(Q < \epsilon\right) = 1,$$

which completes the proof. ■

To prove Theorem 5.1, we also need the following lemma, which is similar to Lemma 7.7 (Salibián-Barrera 2000, page 217), where the author deals with ρ-functions (χ-functions according to our notation).

Lemma 5.2. (Uniform continuity of ψ-functions) *Let $\psi : \mathbb{R} \to \mathbb{R}$ be a continuous function such that $\psi(u) = -c$ for $u \leq -c$, and $\psi(u) = c$ for $u \geq c$, where c is a finite constant. Let $m \in \mathcal{M}$ and $s \in \mathcal{S}$, where $\mathcal{M}$ and $\mathcal{S}$ are bounded real intervals, and* $\inf \mathcal{S} > 0$. *Then*

$$f(u, m, s) = \psi\left(\frac{u - m}{s}\right), \quad u \in \mathbb{R}, \quad m \in \mathcal{M}, \quad s \in \mathcal{S},$$

is conutinuous in m and s uniformly in u.

The proof of this lemma is presented in Chapter Appendix (Section 5.8.1).

Proof of Theorem 5.1

With the use of Lemma 5.1 and Lemma 5.2 the proof is straightforward. We have $Z = (X, Y)$, $\boldsymbol{\theta} = (\mu_X, \mu_Y, \sigma_X, \sigma_Y)$, and $\boldsymbol{t} = (\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_X, \hat{\sigma}_Y)$. First, let us deal with the second term in the numerator of Equation 5.2. We have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}(\hat{U}_i) &= \frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X-\hat{\mu}_X}{\hat{\sigma}_X}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X-\hat{\mu}_X}{\hat{\sigma}_X}\right) I\big((X-\hat{\mu}_X)(Y-\hat{\mu}_Y)>0\big) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X-\hat{\mu}_X}{\hat{\sigma}_X}\right) I\big((X-\hat{\mu}_X)(Y-\hat{\mu}_Y)<0\big).
\end{aligned}$$

Consider

$$\begin{aligned}
g(Z, \boldsymbol{\theta}) &= \psi_{\boldsymbol{c}}\left(\frac{X-\mu_X}{\sigma_X}\right) \\
&= \psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right) I\big((X-\mu_X)(Y-\mu_Y)>0\big) \\
&\quad + \psi_{c_2}\left(\frac{X-\mu_X}{\sigma_X}\right) I\big((X-\mu_X)(Y-\mu_Y)<0\big).
\end{aligned}$$

Since the tuning constant of our score function ψ changes with quadrants, to apply Lemma 5.1 we set

$$A(\boldsymbol{\theta}, \Delta) = \{\boldsymbol{z} = (x, y) : |x-\mu_X| < \Delta \text{ or } |y-\mu_Y| < \Delta\}.$$

We have to choose δ in (5.11) such that if $(x-\mu_X, y-\mu_Y) \in A^c(\boldsymbol{\theta}, \Delta)$ belongs to a particular quadrant, then $(x-\hat{\mu}_X, y-\hat{\mu}_Y)$ belongs to the same quadrant. If, for example, $(x-\mu_X)(y-\mu_Y) > 0$, then $(x-\hat{\mu}_X)(y-\hat{\mu}_Y) > 0$, and, using Lemma 5.2,

$\psi_{\boldsymbol{c}}\left(\frac{x-\hat{\mu}_X}{\hat{\sigma}_X}\right) = \psi_{c_1}\left(\frac{x-\hat{\mu}_X}{\hat{\sigma}_X}\right)$ is continuous in $\hat{\mu}_X$ and $\hat{\sigma}_X$ uniformly on $\boldsymbol{z} \in A^c(\boldsymbol{\theta}, \Delta)$. Therefore, using Lemma 5.1, we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X-\hat{\mu}_X}{\hat{\sigma}_X}\right) \quad \xrightarrow[n\to\infty]{P} \quad E\left[\psi_{\boldsymbol{c}}\left(\frac{X-\mu_X}{\sigma_X}\right)\right].$$

That is,

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}(\hat{U}_i) \quad \xrightarrow[n\to\infty]{P} \quad E\left[\psi_{\boldsymbol{c}}(U)\right]. \tag{5.14}$$

Similarly,

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}(\hat{V}_i) \quad \xrightarrow[n\to\infty]{P} \quad E\left[\psi_{\boldsymbol{c}}(V)\right]. \tag{5.15}$$

Let us now deal with the first term in the numerator of Equation 5.2. We have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}(\hat{U}_i)\psi_{\boldsymbol{c}}(\hat{V}_i) &= \frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{\boldsymbol{c}}\left(\frac{Y-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) I\big((X-\hat{\mu}_X)(Y-\hat{\mu}_Y)>0\big) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{Y-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) I\big((X-\hat{\mu}_X)(Y-\hat{\mu}_Y)<0\big).
\end{aligned}$$

Considering $g(Z, \boldsymbol{\theta}) = \psi_{\boldsymbol{c}}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{\boldsymbol{c}}\left(\frac{Y-\mu_Y}{\sigma_Y}\right) = \psi_{\boldsymbol{c}}(U)\psi_{\boldsymbol{c}}(V)$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}(\hat{U}_i)\ \psi_{\boldsymbol{c}}(\hat{V}_i) \quad \xrightarrow[n\to\infty]{P} \quad E\left[\psi_{\boldsymbol{c}}(U)\psi_{\boldsymbol{c}}(V)\right]. \tag{5.16}$$

Using (5.16), (5.14) and (5.15) in the numerator of (5.2), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}(\hat{U}_i)\ \psi_{\boldsymbol{c}}(\hat{V}_i) - \left(\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}(\hat{U}_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}(\hat{V}_i)\right) \\
&\qquad \xrightarrow[n\to\infty]{P} E\left[\psi_{\boldsymbol{c}}(U)\psi_{\boldsymbol{c}}(V)\right] - E\left[\psi_{\boldsymbol{c}}(U)\right]E\left[\psi_{\boldsymbol{c}}(V)\right].
\end{aligned} \tag{5.17}$$

Let us now focus on the denominator of Equation 5.2. We have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\psi_c^2(\hat{U}_i) &= \frac{1}{n}\sum_{i=1}^{n}\psi_c^2\left(\frac{X-\hat{\mu}_X}{\hat{\sigma}_X}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}^2\left(\frac{X-\hat{\mu}_X}{\hat{\sigma}_X}\right)I\big((X-\hat{\mu}_X)(Y-\hat{\mu}_Y)>0\big)\\ &\quad+\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}^2\left(\frac{X-\hat{\mu}_X}{\hat{\sigma}_X}\right)I\big((X-\hat{\mu}_X)(Y-\hat{\mu}_Y)<0\big).\end{aligned}$$

Consider $g(Z, \boldsymbol{\theta}) = \psi_c^2\left(\frac{X-\mu_X}{\sigma_X}\right) = \psi_c^2(U)$. We then have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_c^2(\hat{U}_i) \xrightarrow[n\to\infty]{P} E\left[\psi_c^2(U)\right]. \tag{5.18}$$

Similarly,

$$\frac{1}{n}\sum_{i=1}^{n}\psi_c^2(\hat{V}_i) \xrightarrow[n\to\infty]{P} E\left[\psi_c^2(V)\right]. \tag{5.19}$$

Using (5.18), (5.19), (5.14) and (5.15), we can say that the denominator of (5.2) converges in probability to the denominator of (5.3). ■

5.3 Influence function of adjusted-Winsorized estimate

The following theorem gives the influence function of the adjusted-Winsorized correlation estimate, when the influence functions of the scale estimates are well-defined.

Theorem 5.2. (Influence function of the adjusted-Winsorized estimate) *Let (X, Y) follow a continuous distribution H. Consider the adjusted-Winsorized correlation functional $r_w(H)$ given by*

$$r_w(H) = \frac{N(H)}{D(H)}, \tag{5.20}$$

with

$$N(H) = E_H\left\{\psi_c\left(\frac{X - m_X(H)}{s_X(H)}\right)\psi_c\left(\frac{Y - m_Y(H)}{s_Y(H)}\right)\right\} - E_H\left\{\psi_c\left(\frac{X - m_X(H)}{s_X(H)}\right)\right\}E_H\left\{\psi_c\left(\frac{Y - m_Y(H)}{s_Y(H)}\right)\right\}, \quad (5.21)$$

and

$$D(H) = \left[E_H\left\{\psi_c^2\left(\frac{X - m_X(H)}{s_X(H)}\right)\right\} - \left(E_H\left\{\psi_c\left(\frac{X - m_X(H)}{s_X(H)}\right)\right\}\right)^2\right]^{1/2} \left[E_H\left\{\psi_c^2\left(\frac{Y - m_Y(H)}{s_Y(H)}\right)\right\} - \left(E_H\left\{\psi_c\left(\frac{Y - m_Y(H)}{s_Y(H)}\right)\right\}\right)^2\right]^{1/2}, \quad (5.22)$$

where $m_X(H)$ and $m_Y(H)$ are location functionals, and $s_X(H)$ and $s_Y(H)$ are dispersion functionals. Suppose that

1. *The central model H_0 is symmetric about $(m_X(H_0), m_Y(H_0))$. We can assume, without loss of generality, that $m_X(H_0) = 0$, $m_Y(H_0) = 0$, $s_X(H_0) = 1$, and $s_Y(H_0) = 1$.*

2. *The influence functions $IF(s_X, u, H_0)$ and $IF(s_Y, v, H_0)$ are well-defined for all $\boldsymbol{z} = (u, v) \in \mathbb{R}^2$.*

The influence function of $r_w(H)$ at H_0 and $\boldsymbol{z}$, denoted by $IF(r_w, \boldsymbol{z}, H_0)$, is given by

$$IF(r_w, \boldsymbol{z}, H_0) = \frac{D_0\dot{N}_0 - N_0\dot{D}_0}{D_0^2},$$

where

$$\begin{aligned} N_0 &= E_{H_0}\{\psi_c(X)\psi_c(Y)\}, \\ D_0 &= \sqrt{E_{H_0}\{\psi_c^2(X)\}E_{H_0}\{\psi_c^2(Y)\}}, \end{aligned}$$

$$\dot{N}_0 = -E_{H_0}\{\psi_c(X)\psi_c(Y)\} + \psi_c(u)\psi_c(v)$$
$$- IF(s_Y, v, H_0)\, E_{H_0}\{\psi_c(X)\,\psi_c'(Y)\,Y\}$$
$$- IF(s_X, u, H_0)\, E_{H_0}\{\psi_c(Y)\,\psi_c'(X)\,X\},$$

and

$$\dot{D}_0 = -\sqrt{\frac{E_{H_0}\{\psi_c^2(Y)\}}{2E_{H_0}\{\psi_c^2(X)\}}}\left(E_{H_0}\{\psi_c^2(X)\} - \psi_c^2(u)\right.$$
$$\left. + IF(s_X, u, H_0)\, E_{H_0}\{2\,\psi_c(X)\,\psi_c'(X)\,X\}\right)$$
$$-\sqrt{\frac{E_{H_0}\{\psi_c^2(X)\}}{2E_{H_0}\{\psi_c^2(Y)\}}}\left(E_{H_0}\{\psi_c^2(Y)\} - \psi_c^2(v)\right.$$
$$\left. + IF(s_Y, v, H_0)\, E_{H_0}\{2\,\psi_c(Y)\,\psi_c'(Y)\,Y\}\right).$$

Proof. Let H be given by

$$H_{t,z} = (1-t)H_0 + t\delta_z. \tag{5.23}$$

For a fixed $z = (u, v)$, using (5.21) we can express $N(H_{t,z}) = N(t, z) = N(t)$ as

$$N(t) = (1-t)\, E_{H_0}\left\{\psi_c\left(\frac{X - m_X(t)}{s_X(t)}\right)\psi_c\left(\frac{Y - m_Y(t)}{s_Y(t)}\right)\right\}$$
$$+ t\,\psi_c\left(\frac{u - m_X(t)}{s_X(t)}\right)\psi_c\left(\frac{v - m_Y(t)}{s_Y(t)}\right)$$
$$- \left[(1-t)\, E_{H_0}\left\{\psi_c\left(\frac{X - m_X(t)}{s_X(t)}\right)\right\} + t\,\psi_c\left(\frac{u - m_X(t)}{s_X(t)}\right)\right]$$
$$\left[(1-t)\, E_{H_0}\left\{\psi_c\left(\frac{Y - m_Y(t)}{s_Y(t)}\right)\right\} + t\,\psi_c\left(\frac{v - m_Y(t)}{s_Y(t)}\right)\right] \tag{5.24}$$

$$
\begin{aligned}
&= (1-t)\, E_{H_0}\left\{\psi_{\boldsymbol{c}}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{\boldsymbol{c}}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\right\} \\
&\qquad + t\,\psi_{\boldsymbol{c}}\left(\frac{u-m_X(t)}{s_X(t)}\right)\psi_{\boldsymbol{c}}\left(\frac{v-m_Y(t)}{s_Y(t)}\right) \\
&\quad - (1-2t)\, E_{H_0}\left\{\psi_{\boldsymbol{c}}\left(\frac{X-m_X(t)}{s_X(t)}\right)\right\} E_{H_0}\left\{\psi_{\boldsymbol{c}}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\right\} \\
&\qquad - t\, E_{H_0}\left\{\psi_{\boldsymbol{c}}\left(\frac{X-m_X(t)}{s_X(t)}\right)\right\}\psi_{\boldsymbol{c}}\left(\frac{v-m_Y(t)}{s_Y(t)}\right) \\
&\qquad - t\, E_{H_0}\left\{\psi_{\boldsymbol{c}}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\right\}\psi_{\boldsymbol{c}}\left(\frac{u-m_X(t)}{s_X(t)}\right) + o(t), \quad (5.25)
\end{aligned}
$$

where $o(t)$ includes the terms involving t^2. Now, since $m_X(0) = 0$, $m_Y(0) = 0$, $s_X(0) = 1$, and $s_Y(0) = 1$, we have

$$
\begin{aligned}
\frac{d}{dt}N(t)\Big|_{t=0} &= -E_{H_0}\left\{\psi_{\boldsymbol{c}}(X)\psi_{\boldsymbol{c}}(Y)\right\} + \psi_{\boldsymbol{c}}(u)\psi_{\boldsymbol{c}}(v) \\
&\qquad + \frac{d}{dt}E_{H_0}\left\{\psi_{\boldsymbol{c}}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{\boldsymbol{c}}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\right\}\Big|_{t=0}. \quad (5.26)
\end{aligned}
$$

The last term in (5.26) can be written as

$$
\begin{aligned}
&\frac{d}{dt}E_{H_0}\left\{\psi_{\boldsymbol{c}}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{\boldsymbol{c}}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\right\}\Big|_{t=0} \\
&= \frac{d}{dt}\Bigg[E_{H_0}\left\{\psi_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{c_1}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\right. \\
&\qquad \left.I\Big((X-m_X(t))(Y-m_Y(t)) > 0\Big)\right\}\Bigg]_{t=0} \\
&\qquad + \frac{d}{dt}\Bigg[E_{H_0}\left\{\psi_{c_2}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{c_2}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\right. \\
&\qquad\qquad \left.I\Big((X-m_X(t))(Y-m_Y(t)) < 0\Big)\right\}\Bigg]_{t=0}. \quad (5.27)
\end{aligned}
$$

Interchanging the operations of differentiation and integration (see Chapter Appendix, Section 5.8.2, for the justification), we can express (5.27) as

$$
\frac{d}{dt}E_{H_0}\left\{\psi_{\boldsymbol{c}}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{\boldsymbol{c}}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\right\}\Big|_{t=0}
$$

$$= E_{H_0}\left[\frac{d}{dt}\left\{\psi_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{c_1}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right) I\big(XY>0\big)\right\}\Big|_{t=0}\right]$$
$$+ E_{H_0}\left[\frac{d}{dt}\left\{\psi_{c_2}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{c_2}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right) I\big(XY<0\big)\right\}\Big|_{t=0}\right]. \quad (5.28)$$

The first term of (5.28) can be expressed as

$$E_{H_0}\left[\frac{d}{dt}\left\{\psi_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{c_1}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right) I\big(XY>0\big)\right\}\Big|_{t=0}\right]$$

$$= E_{H_0}\Bigg\{\psi_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi'_{c_1}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)$$
$$\frac{-s_Y(t)\,\frac{d}{dt}[m_Y(t)] - \frac{d}{dt}[s_Y(t)]\big(Y-m_Y(t)\big)}{s_Y^2(t)}\, I\big(XY>0\big)$$
$$+\psi_{c_1}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\psi'_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right)$$
$$\frac{-s_X(t)\,\frac{d}{dt}[m_X(t)] - \frac{d}{dt}[s_X(t)]\big(X-m_X(t)\big)}{s_X^2(t)}\, I\big(XY>0\big)\Big|_{t=0}\Bigg\}$$

$$= -E_{H_0}\Big\{\psi_{c_1}(X)\psi'_{c_1}(Y)\big[IF(m_Y,v,H_0)+IF(s_Y,v,H_0)\,Y\big]\, I\big(XY>0\big)$$
$$+\psi_{c_1}(Y)\psi'_{c_1}(X)\big[IF(m_X,u,H_0)+IF(s_X,u,H_0)\,X\big]\, I\big(XY>0\big)\Big\}. \quad (5.29)$$

Similarly, the second term of (5.28) can be expressed as

$$E_{H_0}\left[\frac{d}{dt}\left\{\psi_{c_2}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{c_2}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right) I\big(XY<0\big)\right\}\Big|_{t=0}\right]$$

$$= -E_{H_0}\Big\{\psi_{c_1}(X)\psi'_{c_1}(Y)\big[IF(m_Y,v,H_0)+IF(s_Y,v,H_0)\,Y\big]\, I\big(XY<0\big)$$
$$+\psi_{c_1}(Y)\psi'_{c_1}(X)\big[IF(m_X,u,H_0)+IF(s_X,u,H_0)\,X\big]\, I\big(XY<0\big)\Big\}. \quad (5.30)$$

Using (5.29) and (5.30) in (5.28), we have

$$\frac{d}{dt}E_{H_0}\left\{\psi_{\mathbf{c}}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{\mathbf{c}}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\right\}\Big|_{t=0}$$

$$= - IF(m_Y, v, H_0)\; E_{H_0}\{\psi_c(X)\,\psi_c'(Y)\} \;-\; IF(m_X, u, H_0)\; E_{H_0}\{\psi_c(Y)\,\psi_c'(X)\}$$

$$- \; IF(s_Y, v, H_0)\; E_{H_0}\{\psi_c(X)\,\psi_c'(Y)\,Y\} \;-\; IF(s_X, u, H_0)\; E_{H_0}\{\psi_c(Y)\,\psi_c'(X)\,X\}$$

$$= - IF(s_Y, v, H_0)\; E_{H_0}\{\psi_c(X)\,\psi_c'(Y)\,Y\} - IF(s_X, u, H_0)\; E_{H_0}\{\psi_c(Y)\,\psi_c'(X)\,X\}. \quad (5.31)$$

Using (5.31) in (5.26), we have

$$\frac{d}{dt}N(t)\Big|_{t=0} = -E_{H_0}\{\psi_c(X)\psi_c(Y)\} + \psi_c(u)\psi_c(v)$$

$$- \; IF(s_Y, v, H_0)\; E_{H_0}\{\psi_c(X)\,\psi_c'(Y)\,Y\}$$

$$- \; IF(s_X, u, H_0)\; E_{H_0}\{\psi_c(Y)\,\psi_c'(X)\,X\}. \quad (5.32)$$

Using (5.23) in (5.22), we have

$$D(t) = D_1^{\frac{1}{2}}(t)\; D_2^{\frac{1}{2}}(t), \quad (5.33)$$

where

$$D_1(t) = (1-t)\; E_{H_0}\left\{\psi_c^2\left(\frac{X - m_X(t)}{s_X(t)}\right)\right\} + t\,\psi_c^2\left(\frac{u - m_X(t)}{s_X(t)}\right)$$

$$- \left[(1-t)\; E_{H_0}\left\{\psi_c\left(\frac{X - m_X(t)}{s_X(t)}\right)\right\} + t\,\psi_c\left(\frac{u - m_X(t)}{s_X(t)}\right)\right]^2, \quad (5.34)$$

and

$$D_2(t) = (1-t)\; E_{H_0}\left\{\psi_c^2\left(\frac{Y - m_Y(t)}{s_Y(t)}\right)\right\} + t\,\psi_c^2\left(\frac{v - m_Y(t)}{s_Y(t)}\right)$$

$$- \left[(1-t)\; E_{H_0}\left\{\psi_c\left(\frac{Y - m_Y(t)}{s_Y(t)}\right)\right\} + t\,\psi_c\left(\frac{v - m_Y(t)}{s_Y(t)}\right)\right]^2. \quad (5.35)$$

Differentiating both sides of (5.33) w.r.t. t, and setting $t = 0$, we have

$$\frac{d}{dt}D(t)\Big|_{t=0} = \frac{1}{2}\left\{\sqrt{\frac{D_2(t)}{D_1(t)}}\;\frac{d}{dt}D_1(t) + \sqrt{\frac{D_1(t)}{D_2(t)}}\;\frac{d}{dt}D_2(t)\right\}\Big|_{t=0}. \quad (5.36)$$

From (5.34), we have

$$\frac{d}{dt}D_1(t)\Big|_{t=0} = -E_{H_0}\{\psi_c^2(X)\} + \frac{d}{dt}E_{H_0}\left\{\psi_c^2\left(\frac{X-m_X(t)}{s_X(t)}\right)\right\}\Big|_{t=0} + \psi_c^2(u). \qquad (5.37)$$

Using similar arguments as in the case of the numerator (see Chapter Appendix, Section 5.8.2),

$$E_{H_0}\left\{\frac{d}{dt}\psi_c^2\left(\frac{X-m_X(t)}{s_X(t)}\right)\right\}\Big|_{t=0}$$

$$= E_{H_0}\left\{\frac{d}{dt}\psi_{c_1}^2\left(\frac{X-m_X(t)}{s_X(t)}\right) I(XY>0)\Big|_{t=0}\right\} + E_{H_0}\left\{\frac{d}{dt}\psi_{c_2}^2\left(\frac{X-m_X(t)}{s_X(t)}\right) I(XY<0)\Big|_{t=0}\right\} \qquad (5.38)$$

$$= E_{H_0}\Bigg\{2\,\psi_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi'_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right) \frac{-s_X(t)\,\frac{d}{dt}[m_X(t)] - \frac{d}{dt}[s_X(t)](X-m_X(t))}{s_X^2(t)} I(XY>0)$$
$$+ 2\,\psi_{c_2}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi'_{c_2}\left(\frac{X-m_X(t)}{s_X(t)}\right) \frac{-s_X(t)\,\frac{d}{dt}[m_X(t)] - \frac{d}{dt}[s_X(t)](X-m_X(t))}{s_X^2(t)} I(XY<0)\Big|_{t=0}\Bigg\}$$

$$= -E_{H_0}\Big\{2\,\psi_{c_1}(X)\psi'_{c_1}(X)\big[IF(m_X,u,H_0) + IF(s_X,u,H_0)\,X\big]\,I(XY>0)$$
$$+ 2\,\psi_{c_2}(X)\psi'_{c_2}(X)\big[IF(m_X,u,H_0) + IF(s_X,u,H_0)\,X\big]\,I(XY<0)\Big\}$$

$$= -IF(m_X,u,H_0)E_{H_0}\{2\psi_c(X)\psi'_c(X)\} - IF(s_X,u,H_0)E_{H_0}\{2\psi_c(X)\psi'_c(X)X\}$$

$$= -IF(s_X,u,H_0)E_{H_0}\{2\,\psi_c(X)\psi'_c(X)X\}. \qquad (5.39)$$

Using (5.39) in (5.37), we have

$$\frac{d}{dt}D_1(t)\Big|_{t=0} = -E_{H_0}\{\psi_c^2(X)\} - IF(s_X,u,H_0)E_{H_0}\{2\psi_c(X)\psi'_c(X)X\} + \psi_c^2(u). \qquad (5.40)$$

Similarly,

$$\frac{d}{dt}D_2(t)\Big|_{t=0} = -E_{H_0}\{\psi_c^2(Y)\} - IF(s_Y, v, H_0)E_{H_0}\{2\psi_c(Y)\psi_c'(Y)Y\} + \psi_c^2(v). \quad (5.41)$$

Using (5.40) and (5.41) in (5.36), we have

$$\begin{aligned}\frac{d}{dt}D(t)\Big|_{t=0} = &-\sqrt{\frac{E_{H_0}\{\psi_c^2(Y)\}}{2E_{H_0}\{\psi_c^2(X)\}}}\Big(E_{H_0}\{\psi_c^2(X)\} - \psi_c^2(u) \\ &+ IF(s_X, u, H_0)\, E_{H_0}\{2\,\psi_c(X)\,\psi_c'(X)\,X\}\Big) \\ &- \sqrt{\frac{E_{H_0}\{\psi_c^2(X)\}}{2E_{H_0}\{\psi_c^2(Y)\}}}\Big(E_{H_0}\{\psi_c^2(Y)\} - \psi_c^2(v) \\ &+ IF(s_Y, v, H_0)\, E_{H_0}\{2\,\psi_c(Y)\,\psi_c'(Y)\,Y\}\Big). \quad (5.42)\end{aligned}$$

We also have

$$N(t)\big|_{t=0} = E_{H_0}\{\psi_c(X)\psi_c(Y)\} = N_0 \quad \text{(say)}, \quad (5.43)$$

and

$$D(t)\big|_{t=0} = \sqrt{E_{H_0}\{\psi_c^2(X)\}E_{H_0}\{\psi_c^2(Y)\}} = D_0 \quad \text{(say)}. \quad (5.44)$$

Finally, differentiating both sides of

$$r_w(t) = \frac{N(t)}{D(t)} \quad (5.45)$$

w.r.t. t, and setting $t = 0$, we have

$$IF(r_w, \boldsymbol{z}, H_0) = \frac{D_0\dot{N}_0 - N_0\dot{D}_0}{D_0^2}, \quad (5.46)$$

where $\dot{N}_0 = \frac{d}{dt}N(t)\big|_{t=0}$ and $\dot{D}_0 = \frac{d}{dt}D(t)\big|_{t=0}$ are obtained from (5.32) and (5.42), respectively, and N_0 and D_0 are obtained from (5.43) and (5.44), respectively. ■

Figure 5.1 shows a 3D-plot of $IF(r_w, \boldsymbol{z}, H_0)$ against $\boldsymbol{z} = (u, v)$, with $u, v \in [-10, 10]$, $H_0 = N(\mathbf{0}, \Sigma)$, and

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{5.47}$$

We used $\rho = 0.5$ for the bivariate normal distribution, and $(c_1, c_2) = (3, 2)$ for r_w.

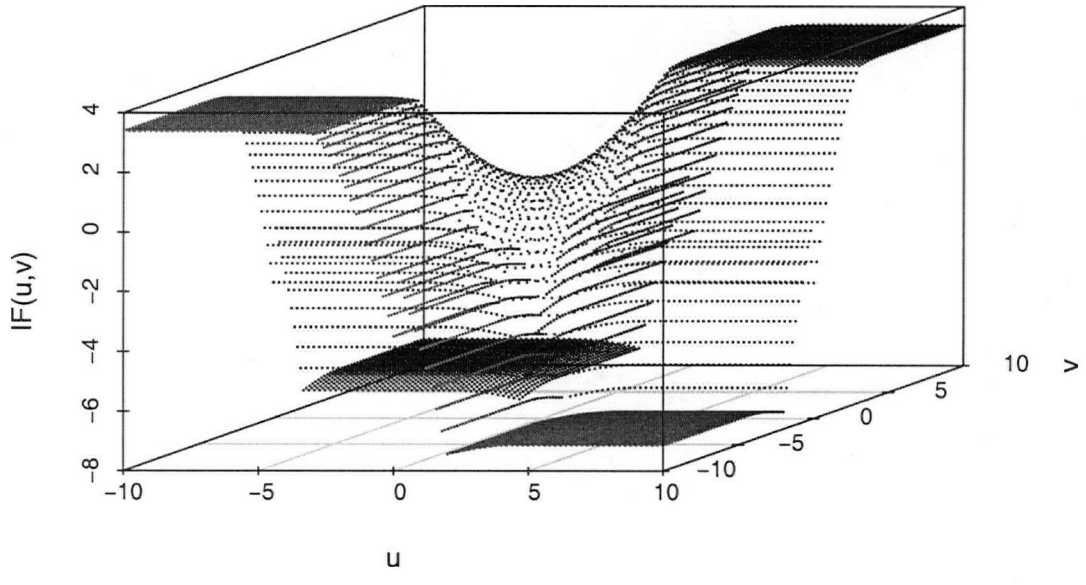


Figure 5.1: Influence curve of adjusted-Winsorized estimate with $(c_1, c_2) = (3, 2)$. The curve is symmetric about $(0, 0)$.

Based on Equation 5.46 and Figure 5.1, we can make the following comments on the influence function of the adjusted-Winsorized estimate:

- Because of Regularity Condition A4, the expectations in (5.46) are bounded. Also, $D_0 > 0$. Therefore, $IF(r_w, \boldsymbol{z}, H_0)$ is bounded when $IF(s_X, u, H_0)$ and $IF(s_Y, v, H_0)$

are bounded. Figure 5.1 also exhibits the boundedness of the influence function.

- Since the ψ-functions are symmetric, the influence function $IF(r_w, \boldsymbol{z}, H_0)$ is symmetric about $(u, v) = (m_X(H_0), m_Y(H_0)) = (0, 0)$ if $IF(s_X, u, H_0)$ is symmetric about $u = 0$, and $IF(s_Y, v, H_0)$ is symmetric about $v = 0$.

- By setting $c_1 = c_2 = c$ in $IF(r_w, \boldsymbol{z}, H_0)$, the influence function of the univariate-Winsorized correlation functional can be obtained.

5.3.1 Standard error of adjusted-Winsorized estimate

Let $Z_i = (X_i, Y_i)$, $i = 1, 2, \cdots, n$, be i.i.d. according to a continuous distribution H, and $r_w(H)$ be the adjusted-Winsorized correlation functional. The influence function of $r_w(H)$ at the central model H_0 and $\boldsymbol{z} \in \mathbb{R}^2$, denoted by $IF(r_w, \boldsymbol{z}, H_0)$, is given by (5.46). Using this, the asymptotic variance of the adjusted-Winsorized correlation estimator $\hat{r}_w$ for the central model H_0 can be obtained as (see, for example, Hampel, Ronchetti, Rousseeuw and Stahel 1986, page 85)

$$AV(r_w, H_0) = \int IF^2(r_w, \boldsymbol{z}, H_0)\, dH_0(\boldsymbol{z}). \tag{5.48}$$

For a sufficiently large n, the standard error of $\hat{r}_w$, denoted by $SE(\hat{r}_w)$, is given by

$$SE(\hat{r}_w) = \sqrt{AV(r_w, H_0)/n}. \tag{5.49}$$

Since it is difficult to get a closed form expression for (5.48), we can use numerical integration to obtain approximations to the asymptotic variance and the standard error of $\hat{r}_w$.

To evaluate the accuracy of the (approximate) standard error of the adjusted-Winsorized correlation estimates, we carried out the following simulation study. We

generated 2000 random samples of size n from a bivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\Sigma = \Sigma(\rho)$ given by (5.47). We considered 3 different sample sizes: $n = 25$, 100 and 400, and 3 different correlation coefficient values: $\rho = 0.1$, 0.5 and 0.9. The values of (c_1, c_2) chosen for these correlation coefficients are $(3, 3)$, $(3, 2)$ and $(3, 1)$, respectively. The choice of c_1 and c_2 is discussed later on in Section 5.4.

Table 5.1: Evaluation of the standard errors of $\hat{r}_w$. The empirical SD and formula-based SE are close.

	$n = 25$		$n = 100$		$n = 400$	
ρ	SD	SE	SD	SE	SD	SE
0.10	0.203	0.199	0.100	0.099	0.049	0.050
0.50	0.136	0.146	0.069	0.073	0.034	0.037
0.90	0.035	0.039	0.018	0.019	0.009	0.010

Table 5.1 presents the obtained results. For each n, the first column gives the empirical standard deviations of the adjusted-Winsorized correlation estimates (based on 2000 samples), while the second column shows the standard errors calculated using (5.49). In general, the differences between the numbers in the two columns are reasonably small, particularly for large sample sizes.

An estimate of the asymptotic variance in (5.48) is given by

$$\widehat{AV}(r_w, H_n) = \frac{1}{n}\sum_{i=1}^{n} IF_n^2(r_w, \boldsymbol{z}_i, H_n), \tag{5.50}$$

and the estimated standard error of $\hat{r}_w$, denoted by $\widehat{SE}(\hat{r}_w)$ is given by

$$\widehat{SE}(\hat{r}_w) = \sqrt{\widehat{AV}(r_w, H_n)/n}. \tag{5.51}$$

5.4 Choice of c_1 and c_2 for $\hat{r}_w$

It is important to note that the robustness, efficiency and intrinsic bias (to be discussed in the next section) of the adjusted-Winsorized correlation estimate depends on the values chosen for c_1 and c_2. Figure 3.3 in Chapter 3 shows how the bivariate outliers are handled by $\hat{r}_w$. If we choose large values for c_1 and c_2, $\hat{r}_w$ will be less resistant to the outliers. On the other hand, this will lead to a decrease in the standard error (increase in efficiency) and a decrease in the intrinsic bias, both of which are desirable as well. Thus, the choice of c_1 and c_2 may depend on our goal in a particular situation.

For application purposes, we can first select an appropriate value for c_1 (the larger tuning constant for the two "major" quadrants, i.e., the quadrants that contain the majority of the standardized data). Then, for the two "minor" quadrants, we can use $c_2 = hc_1$, where h is the ratio of the number of observations in the minor quadrants to the number of observations in the major quadrants. Note that, as $|\rho|$ increases from 0 to 1, the asymptotic value of h decreases from 1 to 0.

5.5 Intrinsic bias in adjusted-Winsorized estimate

Let $Z = (X, Y)$ have a continuous distribution H, and $r_w(H)$ be the adjusted-Winsorized correlation functional. Let the central model H_0 be given by $H_0 = N(\mathbf{0}, \Sigma)$, where $\Sigma = \Sigma(\rho)$ is given by (5.47), and $\rho = \rho(H_0)$ is the true correlation coefficient of X and Y.

The intrinsic bias of r_w occurs at the central model H_0 because of the data transformation. Since the adjusted-Winsorized data have a slightly different correlation coef-

ficient than that of the original data, $r_w(H_0) \neq \rho(H_0)$. Therefore, the intrinsic bias of r_w, denoted by $IB(r_w)$, is given by

$$IB(r_w) = r_w(H_0) - \rho(H_0). \tag{5.52}$$

To compare $r_w(H_0)$ and $\rho(H_0)$ empirically, we generated random samples of size $n = 100000$ from a bivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma = \Sigma(\rho)$. We considered several values of ρ from -1 to 1. To calculate $r_w = r_w(H_0)$ we used Huber score function with different values of c_1 with $c_2 = hc_1$, where h is defined in the last section.

Figure 5.2 displays the plots of r_w against ρ for $c_1 = 0.01$, 1, 2 and 3. Based on these plots we can make the following comments:

- The intrinsic bias of r_w decreases as c_1 increases.
- r_w is a non-decreasing function of ρ.
- Consider $0 \leq \rho \leq 1$. The magnitude of intrinsic bias increases from zero to reach its maximum at $\rho = 0.5$, and then decreases to zero. Similar behavior is observed for $-1 \leq \rho \leq 0$.
- For $-1 \leq \rho \leq 0$ the intrinsic bias is negative, while for $0 \leq \rho \leq 1$ the intrinsic bias is positive. This is the exact opposite of the results obtained for univariate-Winsorized estimate (see Alqallaf 2003, Figure 4.5, page 97), for which the bias is positive when $-1 \leq \rho \leq 0$, and negative when $0 \leq \rho \leq 1$.

To compare the behavior of r_w (with $c_1 \simeq 0$, $c_2 \simeq 0$) with the univariate-Winsorized correlation estimate $r = r(H_0)$ (with $c \simeq 0$), we plotted r with $c = 0.01$ against ρ in

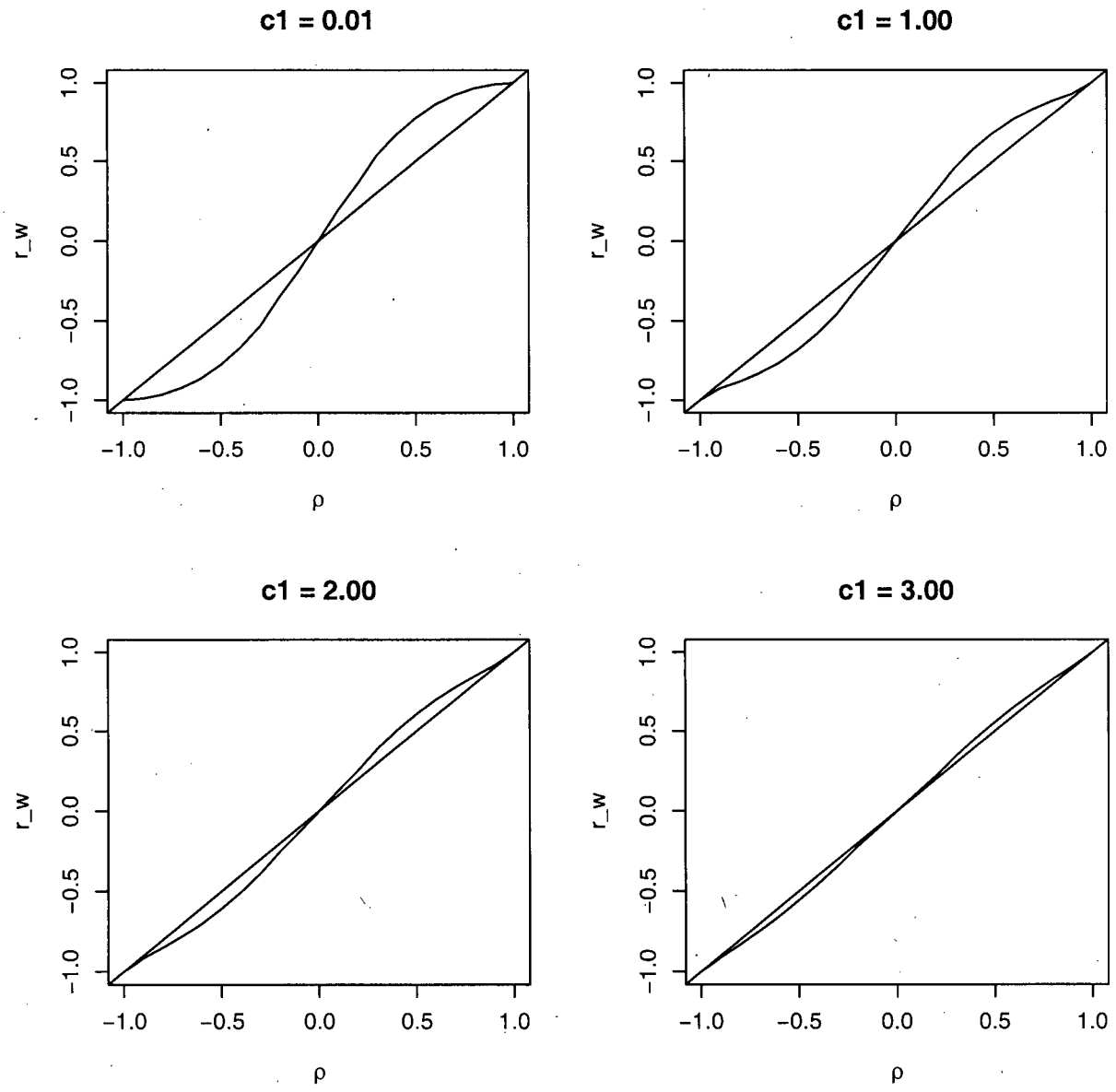


Figure 5.2: Intrinsic bias in adjusted-Winsorized estimates with $c_2 = hc_1$. The bias in r_w decreases as c_1 increases.

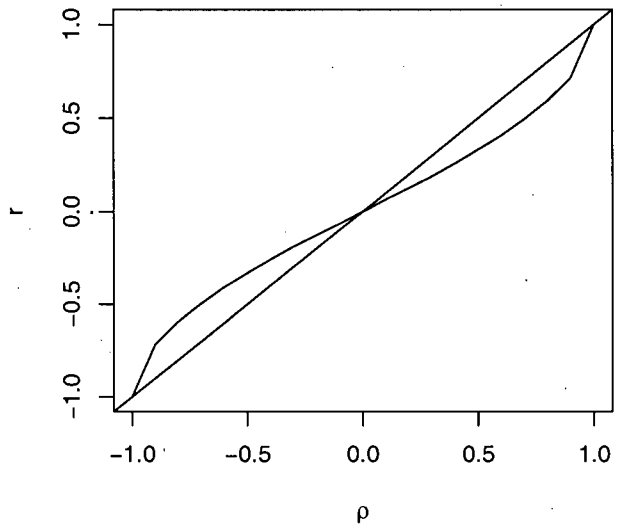


Figure 5.3: Intrinsic bias in univariate-Winsorized estimate (c=0.01).

Figure 5.3 (which is a reproduction of the top left plot of Figure 4.5, Alqallaf 2003). This plot is the exact opposite of the top left plot of Figure 5.2 in terms of the sign of the intrinsic bias. The reason maybe as follows. Though $c_1 = 0 \Rightarrow c_2 = 0$ for the adjusted-Winsorized estimates, but $c_2 = hc_1$ approaches zero faster than c_1, making the limit of r_w different from the limit of r as c tends to zero (the limit in the latter case being the quadrant correlation estimate).

5.5.1 Achieving (approximate) Fisher-consistency for r_w

Let $H_0 = N(\mathbf{0}, \Sigma(\rho))$ and let $r_w(\rho, c_1, c_2)$ be the asymptotic value of $r_w(H_0)$ when we use tuning constants c_1 and c_2. A plot of $r_w(\rho, c_1, c_2)$ against ρ is exhibited by Figure 5.2 for different values of c_1 and $c_2 = hc_1$. Recall that h is the ratio of the number of observations

in the minor quadrants to the number of observations in the major quadrants. To fix ideas and without loss of generality, let us assume that $\rho \geq 0$. We notice that when $c_2 = c_1$ (univariate-Winsorized estimate, Alqallaf 2003) $r_w(\rho, c_1, c_1) \leq \rho$, while $r_w(\rho, c_1, hc_1) \geq \rho$ when $c_2 = hc_1$. Therefore, to achieve Fisher-consistency for a fixed value of ρ, we could use an appropriate value of c_2 between $c_2 = hc_1$ and $c_2 = c_1$. That is, we could use $c_2 = ac_1$, where $a \in (h, 1)$ is such that $r_w(\rho, c_1, ac_1) = \rho$. In practice, we could approximate a by numerical means. For a fixed value of c_1 we could obtain a table relating ρ and $a = g_1(\rho)$. Since h is a decreasing function of ρ we have that $a = g(h)$. Tables relating a and h could be constructed by numerical means for any desired value of c_1. We will not elaborate this approach further, since a simple approach that works remarkably well is presented below.

To avoid the construction and use of numerous tables, we can use a simple alternative that does not require any table. Since $c_2 = hc_1$ and $c_2 = c_1$ give biases of similar (though not same) magnitudes with opposite signs, we can use $c_2 = c_1(h+1)/2$, that is, $a = (h+1)/2$. Figure 5.4 displays the plots of r_w against ρ for $c_1 = 1$ with $c_2 = hc_1$ and $c_2 = c_1(h+1)/2$. Though the first plot ($c_2 = hc_1$) is the same as the top right plot of Figure 5.2, it is presented again here to make its scale more comparable to that of the second plot. The degree of Fisher-consistency achieved by using $c_2 = c_1(h+1)/2$ is quite satisfactory.

Note that $c_2 = c_1(h+1)/2 \leq c_1$. Therefore, with this choice of c_2 the adjusted-Winsorized estimate is still more resistant to bivariate outliers than the univariate-Winsorized estimate. At the same time, the extra tuning constant c_2 allows us to make our estimate approximately Fisher-consistent.

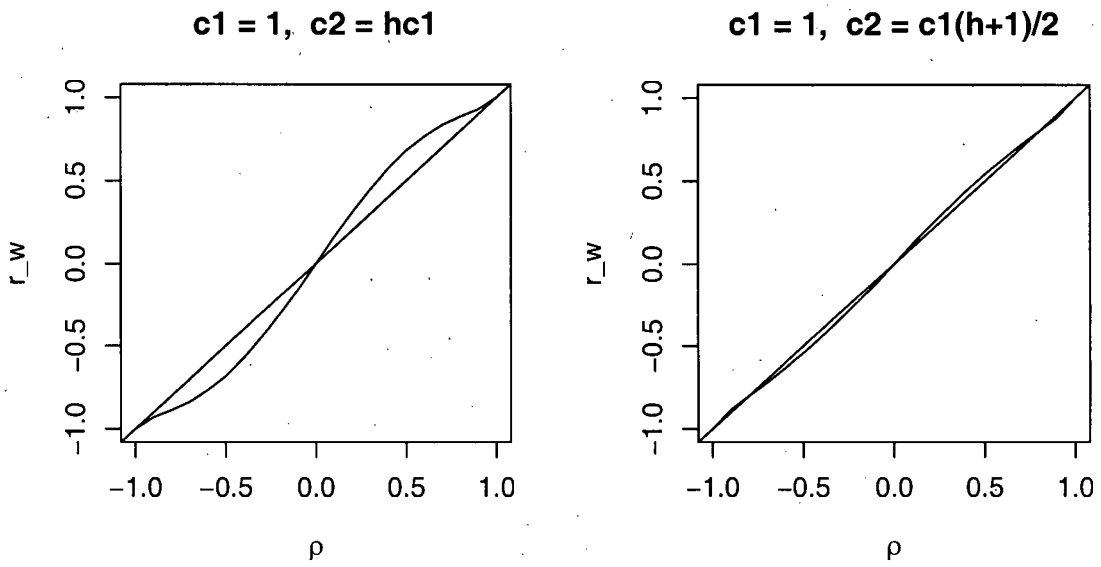


Figure 5.4: Approximate Fisher-consistency for r_w. By using $c_2 = c_1(h+1)/2$ we get less intrinsic bias than $c_2 = hc_1$.

We mentioned in Chapter 3 that though we used $c_2 = hc_1$ in this study, a more reasonable choice would have been $c_2 = \sqrt{h}\,c_1$ (i.e., $c_2^2 = hc_1^2$), because the areas of the two squares should be proportional to the number of observations they contain. Interestingly, $(h+1)/2$ (the shrinkage factor that gives approximate Fisher-consistency) is the first-order Taylor expansion of $\sqrt{h}$.

5.6 Asymptotic normality of adjusted-Winsorized estimate

Since the indicator functions involved in the adjusted-Winsorized correlation estimate are not differentiable, proving the asymptotic normality of this estimate is extremely

difficult. One approach we can consider is to replace the sample indicator functions by the true ones, which we can do only if the amount of error due to the replacement is $o\left(1/\sqrt{n}\right)$. See Chapter Appendix (Section 5.8.3) for more details of this idea, where we use this approach to establish the asymptotic normality of the numerator of (5.2), i.e., the asymptotic normality of the adjusted-Winsorized "covariance" estimates.

Unfortunately, for the denominator of (5.2), the amount of error due to replacing the sample indicator functions by the true ones is $O\left(1/\sqrt{n}\right)$ (see Chapter Appendix, Section 5.8.4), and cannot be ignored. Therefore, the above approach cannot be used to establish the asymptotic normality of the adjusted-Winsorized "correlation" estimates.

As a remedy of this, we can use differentiable versions of the indicator functions, denoted by $\gamma_1\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$ and $\gamma_2(.) = 1-\gamma_1(.)$, respectively. For example, γ_1 can be the distribution function of a continous random variable with support $(-\epsilon, \epsilon)$, for any small $\epsilon > 0$. Using the functions γ_1 and γ_2, we now define the smoothed adjusted-Winsorization of the data, and the smoothed adjusted-Winsorized correlation estimates as follows.

Definition 5.3. (Smoothed adjusted-Winsorization)

The smoothed adjusted-Winsorization of $(u, v) \in \mathbb{R}^2$, *denoted by* $\Psi_{\boldsymbol{c}}^S(u, v)$, *is defined as*

$$\Psi_{\boldsymbol{c}}^S(u, v) = \left(\psi_{\boldsymbol{c}}^S(u),\, \psi_{\boldsymbol{c}}^S(v)\right), \tag{5.53}$$

where

$$\begin{aligned}\psi_{\boldsymbol{c}}^S(u) &= \psi_{c_1}(u)\gamma_1(uv) + \psi_{c_2}(u)\gamma_2(uv),\\ \psi_{\boldsymbol{c}}^S(v) &= \psi_{c_1}(v)\gamma_1(uv) + \psi_{c_2}(v)\gamma_2(uv),\end{aligned}$$

ψ is a non-decreasing symmetric function, and c_1 and c_2 are chosen constants.

Definition 5.4. (Smoothed adjusted-Winsorized estimate of correlation) *Let* (X_i, Y_i), $i = 1, 2, \cdots, n$, *be a random sample from a bivariate distribution with location parameters* μ_X *and* μ_Y, *and scale parameters* σ_X *and* σ_Y, *respectively. Let* $\boldsymbol{\theta} = (\mu_X, \mu_Y, \sigma_X, \sigma_Y)$, *and* $\hat{\boldsymbol{\theta}} = (\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_X, \hat{\sigma}_Y)$ *be an estimate of* $\boldsymbol{\theta}$. *Denote* $\hat{U}_i = (X_i - \hat{\mu}_X)/\hat{\sigma}_X$, *and* $\hat{V}_i = (Y_i - \hat{\mu}_Y)/\hat{\sigma}_Y$. *Then, the smoothed adjusted-Winsorized correlation estimate,* $\hat{r}_S$, *is defined as*

$$\hat{r}_S = \frac{\frac{1}{n}\sum_{i=1}^{n} \psi_{\mathbf{c}}^S(\hat{U}_i)\psi_{\mathbf{c}}^S(\hat{V}_i) - \left(\frac{1}{n}\sum_{i=1}^{n} \psi_{\mathbf{c}}^S(\hat{U}_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n} \psi_{\mathbf{c}}^S(\hat{V}_i)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\{\psi_{\mathbf{c}}^S(\hat{U}_i)\right\}^2 - \left(\frac{1}{n}\sum_{i=1}^{n} \psi_{\mathbf{c}}^S(\hat{U}_i)\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\{\psi_{\mathbf{c}}^S(\hat{V}_i)\right\}^2 - \left(\frac{1}{n}\sum_{i=1}^{n} \psi_{\mathbf{c}}^S(\hat{V}_i)\right)^2}}. \tag{5.54}$$

The following theorem states the asymptotic normality of the smoothed adjusted-Winsorized correlation estimates, provided that the ψ- and χ-functions satisfy the above conditions, and the location estimates are consistent.

Theorem 5.3. (Asymptotic normality of the smoothed adjusted-Winsorized estimate) *Let* (X_i, Y_i), $i = 1, 2, \cdots, n$, *be a random sample from an elliptically symmetric bivariate distribution with location parameters* μ_X *and* μ_Y, *and scale parameters* σ_X *and* σ_Y, *respectively. Let* $\hat{\mu}_X$ *and* $\hat{\mu}_Y$ *be consistent estimates of* μ_X *and* μ_Y, *and* $\hat{\sigma}_X$ *and* $\hat{\sigma}_Y$ *are S-estimates of* σ_X *and* σ_Y *with score functions satisfying conditions B1 – B4. Then,*

$$\sqrt{n}\,(\hat{r}_S - r_S) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, AV),$$

where $\hat{r}_w$ *is defined in (5.54) with* ψ*-functions satisfying conditions A1 – A4,*

$$r_S = \frac{E\left[\psi_{\mathbf{c}}^S\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{\mathbf{c}}^S\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]}{\sqrt{E\left[\left\{\psi_{\mathbf{c}}^S\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}^2\right]}\sqrt{E\left[\left\{\psi_{\mathbf{c}}^S\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}^2\right]}},$$

and the variance AV of the limiting distribution is given by

$$AV = \nabla_g' \Sigma \nabla_g,$$

where $\underset{(3\times3)}{\Sigma} = \{\sigma_{ij}\}$, *with* $\sigma_{ij} = \text{Cov}\,(Q_i, Q_j)$; $i = 1, 2, 3$; $j = 1, 2, 3$;

$$\begin{aligned}
Q_1 &= \psi_c^S\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_c^S\left(\frac{Y-\mu_Y}{\sigma_Y}\right) - \frac{K_9}{D_X}\chi\left(\frac{X-\mu_X}{\sigma_X}\right) - \frac{K_{10}}{D_Y}\chi\left(\frac{Y-\mu_Y}{\sigma_Y}\right), \\
Q_2 &= \left\{\psi_c^S\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}^2 - \frac{K_{17}}{D_X}\chi\left(\frac{X-\mu_X}{\sigma_X}\right) - \frac{K_{18}}{D_Y}\chi\left(\frac{Y-\mu_Y}{\sigma_Y}\right), \\
Q_3 &= \left\{\psi_c^S\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}^2 - \frac{K_{19}}{D_Y}\chi\left(\frac{Y-\mu_Y}{\sigma_Y}\right) - \frac{K_{20}}{D_X}\chi\left(\frac{X-\mu_X}{\sigma_X}\right), \\
\nabla_g &= \left(\frac{1}{\sqrt{V\,W}}, -\frac{U}{2\,V\sqrt{V\,W}}, -\frac{U}{2\,W\sqrt{V\,W}}\right),
\end{aligned}$$

with

$$\begin{aligned}
U &= E\left[\psi_c^S\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_c\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right], \\
V &= E\left[\left\{\psi_c^S\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}^2\right], \\
W &= E\left[\left\{\psi_c^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}^2\right],
\end{aligned}$$

and the constants used in the above expressions are specified below:

$$D_X = E\left[\chi'\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\right], \tag{5.55}$$

$$D_Y = E\left[\chi'\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right], \tag{5.56}$$

$$\begin{aligned}
K_1 = \frac{1}{\sigma_X}E\left[\psi'_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{c_1}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\gamma_1^2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right] \\
+ \frac{2}{\sigma_X}E\left[\psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{c_1}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\gamma_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right. \\
\left.\gamma_1'\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right],
\end{aligned} \tag{5.57}$$

K_2 is obtained by interchanging X and Y in (5.57), K_3 is obtained by replacing c_1 by c_2, and γ_1 by γ_2 in (5.57), K_4 is obtained by interchanging X and Y, and replacing c_1 by c_2, and γ_1 by γ_2 in (5.57),

$$\begin{aligned} K_5 = \frac{1}{\sigma_X} E\Bigg[& \psi'_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{Y-\mu_Y}{\sigma_Y}\right) \\ & \gamma_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\Bigg] \\ + \frac{1}{\sigma_X} E\Bigg[& \psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\gamma'_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right) \\ & \left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\Bigg] \\ + \frac{1}{\sigma_X} E\Bigg[& \psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\gamma_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right) \\ & \gamma'_2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\Bigg], \end{aligned} \tag{5.58}$$

K_6 is obtained by interchanging X and Y, and c_1 and c_2 in (5.58), K_7 is obtained by interchanging c_1 and c_2 in (5.58), K_8 is obtained by interchanging X and Y in (5.58),

$$K_9 = K_1 + K_3 + K_5 + K_7, \tag{5.59}$$

$$K_{10} = K_2 + K_4 + K_6 + K_8, \tag{5.60}$$

$$\begin{aligned} K_{11} = \frac{2}{\sigma_X} E\Bigg[& \psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi'_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\gamma_1^2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\Bigg] \\ + \frac{2}{\sigma_X} E\Bigg[& \psi_{c_1}^2\left(\frac{X-\mu_X}{\sigma_X}\right)\gamma_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right) \\ & \gamma'_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\Bigg], \end{aligned} \tag{5.61}$$

$$\begin{aligned} K_{12} = \frac{2}{\sigma_Y} E\Bigg[& \psi_{c_1}^2\left(\frac{X-\mu_X}{\sigma_X}\right)\gamma_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right) \\ & \gamma'_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\Bigg], \end{aligned} \tag{5.62}$$

K_{13} and K_{14} are obtained by replacing c_1 by c_2, and γ_1 by γ_2 in (5.61) and (5.62) (respectively),

$$K_{15} = \frac{2}{\sigma_X} E\left[\psi'_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{X-\mu_X}{\sigma_X}\right)\right.$$
$$\left.\gamma_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$
$$+\frac{2}{\sigma_X} E\left[\psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\psi'_{c_2}\left(\frac{X-\mu_X}{\sigma_X}\right)\right.$$
$$\left.\gamma_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$
$$+\frac{2}{\sigma_X} E\left[\psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{X-\mu_X}{\sigma_X}\right)\gamma'_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right.$$
$$\left.\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$
$$+\frac{2}{\sigma_X} E\left[\psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{X-\mu_X}{\sigma_X}\right)\gamma_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right.$$
$$\left.\gamma'_2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right], \quad (5.63)$$

$$K_{16} = \frac{2}{\sigma_Y} E\left[\psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{X-\mu_X}{\sigma_X}\right)\gamma'_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right.$$
$$\left.\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$
$$+\frac{2}{\sigma_Y} E\left[\psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{X-\mu_X}{\sigma_X}\right)\gamma_1\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right.$$
$$\left.\gamma'_2\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\right], \quad (5.64)$$

$$K_{17} = K_{11} + K_{13} + K_{15}, \quad (5.65)$$

$$K_{18} = K_{12} + K_{14} + K_{16}, \quad (5.66)$$

and, finally, K_{19} and K_{20} are obtained by interchanging X and Y in (5.65) and (5.66), respectively.

Sketch of the Proof.

The numerator of (5.54) can be written as

$$\hat{N}_S = \frac{1}{n}\sum_{i=1}^{n}\psi_c^S\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_c^S\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) - \frac{1}{n}\sum_{i=1}^{n}\psi_c^S\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\frac{1}{n}\sum_{i=1}^{n}\psi_c^S\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right). \tag{5.67}$$

We can express the first term of (5.67) as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\psi_c^S\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_c^S\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\gamma_1\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) + \psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\gamma_2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\right\}\\
&\times\left\{\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_1\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) + \psi_{c_2}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\right\}\\
&= \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_1^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\\
&\qquad + \frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_2^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\\
&+ \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_1\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\\
&+ \frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_1\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right).
\end{aligned} \tag{5.68}$$

Let us consider the first term of (5.68). Using Taylor expansion about $(\mu_X, \sigma_X, \mu_Y, \sigma_Y)$, we can write

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_1^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$

$$
\begin{aligned}
&\doteq \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_1^2\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)\\
&-\frac{1}{n\tilde{\sigma}_X}\sum_{i=1}^{n}\psi'_{c_1}\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\psi_{c_1}\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\gamma_1^2\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)(\hat{\sigma}_X-\sigma_X)\\
&\quad-\frac{2}{n\tilde{\sigma}_X}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\gamma_1\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\\
&\quad\times\gamma'_1\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}(\hat{\sigma}_X-\sigma_X)\\
&-\frac{1}{n\tilde{\sigma}_Y}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\psi'_{c_1}\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\gamma_1^2\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)(\hat{\sigma}_Y-\sigma_Y)\\
&\quad-\frac{2}{n\tilde{\sigma}_Y}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\gamma_1\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\\
&\quad\times\gamma'_1\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}(\hat{\sigma}_Y-\sigma_Y),
\end{aligned}
$$

since the coefficients of $(\hat{\mu}_X-\mu_X)$ and $(\hat{\mu}_Y-\mu_Y)$ converge to zero in probability. Thus,

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_1^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\\
&\quad\doteq\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_1^2\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)\\
&\qquad-K_1(\hat{\sigma}_X-\sigma_X)-K_2(\hat{\sigma}_Y-\sigma_Y), \quad (5.69)
\end{aligned}
$$

where K_1 and K_2 are as defined in the statement of the theorem. Similarly, the next three terms of (5.68) can be expressed as

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_2^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\\
&\quad\doteq\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_2^2\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)\\
&\qquad-K_3(\hat{\sigma}_X-\sigma_X)-K_4(\hat{\sigma}_Y-\sigma_Y), \quad (5.70)
\end{aligned}
$$

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_1\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$
$$\doteq \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_1\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)$$
$$- K_5(\hat{\sigma}_X-\sigma_X) - K_6(\hat{\sigma}_Y-\sigma_Y), \quad (5.71)$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_1\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$
$$\doteq \frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_1\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)$$
$$- K_7(\hat{\sigma}_X-\sigma_X) - K_8(\hat{\sigma}_Y-\sigma_Y), \quad (5.72)$$

respectively, where K_3, K_4, K_5, K_6, K_7 and K_8 are as defined in the statement of the theorem. Using (5.69), (5.70), (5.71) and (5.72) in (5.68), we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}^{S}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{\boldsymbol{c}}^{S}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$

$$\doteq \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_1^2\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)$$
$$+\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_2^2\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)$$
$$+\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_1\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)$$
$$+\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_1\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)$$
$$- K_9(\hat{\sigma}_X-\sigma_X) - K_{10}(\hat{\sigma}_Y-\sigma_Y), \quad (5.73)$$

where $K_9 = K_1 + K_3 + K_5 + K_7$ and $K_{10} = K_2 + K_4 + K_6 + K_8$.

Now, it is straightforward to show that the second term of (5.67) is $o(1/\sqrt{n})$. Therefore, using (5.73) in (5.67), we have

$$\hat{N}_S \doteq \frac{1}{n}\sum_{i=1}^{n}\psi_c^S\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_c^S\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right) - K_9(\hat{\sigma}_X-\sigma_X) - K_{10}(\hat{\sigma}_Y-\sigma_Y). \quad (5.74)$$

Note that the first 4 terms of the right-hand-side of (5.73) are combined to get a single term in (5.74).

Since $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are S-scales, we have (see Alqallaf 2003, page 156)

$$\hat{\sigma}_X - \sigma_X \doteq \frac{\sum_{i=1}^{n}\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right) - nb}{\sum_{i=1}^{n}\chi'\left(\frac{X_i-\mu_X}{\sigma_X}\right)\left(\frac{X_i-\mu_X}{\sigma_X}\right)}, \quad (5.75)$$

and

$$\hat{\sigma}_Y - \sigma_Y \doteq \frac{\sum_{i=1}^{n}\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right) - nb}{\sum_{i=1}^{n}\chi'\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)}. \quad (5.76)$$

Using (5.75) and (5.76) in (5.74), we have

$$\begin{aligned}\hat{N}_S \doteq{}& \frac{1}{n}\sum_{i=1}^{n}\psi_c^S\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_c^S\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\\ &- \frac{K_9}{D_X}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-b\right) - \frac{K_{10}}{D_Y}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-b\right), \quad (5.77)\end{aligned}$$

where D_X and D_Y are defined in the statement of the theorem (equations 5.55 and 5.56).

The first term of the denominator of (5.54) can be written as

$$\hat{D}_1^S = \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_c^S\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\right\}^2 - \left\{\frac{1}{n}\sum_{i=1}^{n}\psi_c^S\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\right\}^2. \quad (5.78)$$

We can express the first term of (5.78) as

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\psi_c^S\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\right\}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\gamma_1\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\right.$$
$$\left. + \psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\gamma_2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\right\}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\gamma_1^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$
$$+ \frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\gamma_2^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$
$$+ \frac{2}{n}\sum\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\gamma_1\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$
$$\times\,\gamma_2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right). \quad (5.79)$$

As in the case of the numerator, by using Taylor expansion about $(\mu_X, \sigma_X, \mu_Y, \sigma_Y)$, the three terms on the right-hand-side of (5.79) can be expressed as

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\gamma_1^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$
$$\doteq \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}^2\left(\frac{X_i-\mu_X}{\sigma_X}\right)\gamma_1^2\left(\frac{X_i-\mu_X}{\hat{\sigma}_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)$$
$$- K_{11}(\hat{\sigma}_X-\sigma_X) - K_{12}(\hat{\sigma}_Y-\sigma_Y), \quad (5.80)$$

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\gamma_2^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$
$$\doteq \frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}^2\left(\frac{X_i-\mu_X}{\sigma_X}\right)\gamma_2^2\left(\frac{X_i-\mu_X}{\hat{\sigma}_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)$$
$$- K_{13}(\hat{\sigma}_X-\sigma_X) - K_{14}(\hat{\sigma}_Y-\sigma_Y), \quad (5.81)$$

$$\frac{2}{n}\sum\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\gamma_1\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\gamma_2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$
$$\doteq \frac{2}{n}\sum\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\gamma_1\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)\gamma_2\left(\frac{X_i-\mu_X}{\sigma_X}\frac{Y_i-\mu_Y}{\sigma_Y}\right)$$
$$- K_{15}(\hat{\sigma}_X-\sigma_X) - K_{16}(\hat{\sigma}_Y-\sigma_Y), \quad (5.82)$$

where K_{11}, K_{12}, K_{13}, K_{14}, K_{15} and K_{16} are as defined in the statement of the theorem. Using (5.80), (5.81) and (5.82) in (5.79), we have

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\psi_c^S\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\right\}^2 \doteq \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_c^S\left(\frac{X_i-\mu_X}{\sigma_X}\right)\right\}^2 - K_{17}(\hat{\sigma}_X-\sigma_X) - K_{18}(\hat{\sigma}_Y-\sigma_Y), \quad (5.83)$$

where $K_{17} = K_{11} + K_{13} + K_{15}$ and $K_{18} = K_{12} + K_{14} + K_{16}$. Now, it can be shown that the second term of (5.78) is $o(1/\sqrt{n})$. Therefore, using (5.83) in (5.78), and using (5.75) and (5.76), we have

$$\hat{D}_1^S \doteq \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_c^S\left(\frac{X_i-\mu_X}{\sigma_X}\right)\right\}^2 - \frac{K_{17}}{D_X}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-b\right) - \frac{K_{18}}{D_Y}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-b\right). \quad (5.84)$$

Similarly, for the second term of the denominator of (5.54) we have

$$\hat{D}_2^S \doteq \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_c^S\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\right\}^2 - \frac{K_{19}}{D_Y}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-b\right) - \frac{K_{20}}{D_X}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-b\right). \quad (5.85)$$

Using (5.77), (5.84) and (5.85), and ignoring the terms which are $o(1/\sqrt{n})$, we can state that

$$\sqrt{n}\left[\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}\psi_c^S\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_c^S\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) \\ \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_c^S\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\right\}^2 \\ \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_c^2\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\right\}^2 \end{pmatrix} - \begin{pmatrix} E\left[\psi_c^S\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_c\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] \\ E\left[\left\{\psi_c^S\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}^2\right] \\ E\left[\left\{\psi_c^S\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}^2\right] \end{pmatrix}\right] \xrightarrow[n\to\infty]{d} \mathcal{N}(\mathbf{0}, \Sigma), \quad (5.86)$$

where Σ is as defined in the statement of the theorem. Finally, we can use the δ-method (Billingsley 1986) on (5.86) to complete the proof. ■

5.7 Conclusion

The main contribution of this chapter is that we established some asymptotic properties of the adjusted-Winsorized correlation estimate (the new robust correlation estimate proposed in this thesis). Our estimate is consistent and has bounded influence. We obtained its asymptotic variance and intrinsic bias. The tuning constants of this estimate can be chosen such that we have approximate Fisher-consistency. An smoothed version of this estimate is asymptotically normal. The computing time of this estimate is $\mathcal{O}(n \log n)$, the same as that of the univariate-Winsorized correlation estimate, but our estimate is more resistant to bivariate outliers.

5.8 Chapter Appendix

5.8.1 Proof of Lemma 5.2

We need to show that, for any $\epsilon > 0$, there exists $\delta_m > 0$ and $\delta_s > 0$ such that, for all $u \in \mathbb{R}$,

$$|m_1 - m_2| < \delta_m, \quad |s_1 - s_2| < \delta_s \quad \Rightarrow \quad |f(u, m_1, s_1) - f(u, m_2, s_2)| < \epsilon.$$

Following Salibián-Barrera (2000), we will first show that there exists a closed and bounded interval $\mathcal{U}$ such that, for any $m_1, m_2 \in \mathcal{M}$, $s_1, s_2 \in \mathcal{S}$,

$$\psi\left(\frac{u - m_1}{s_1}\right) = \psi\left(\frac{u - m_2}{s_2}\right), \quad u \notin \mathcal{U}. \tag{5.87}$$

It is given that $\underline{m} \leq m \leq \bar{m}$ and $\underline{s} \leq s \leq \bar{s}$. Consider $u \geq \bar{m} + \bar{s}c$. For any $m \leq \bar{m}$ and $s \leq \bar{s}$ we have $(u - m)/s \geq c$. Now, consider $u \leq \underline{m} - \underline{s}c$. For any $m \geq \underline{m}$ and $s \geq \underline{s}$ we

have $(u-m)/s \leq -c$. Thus, for $\mathcal{U} = [\underline{m} - \underline{s}c,\ \bar{m} + \bar{s}c]$, (5.87) holds.

For $u \in \mathcal{U}$, ψ is uniformly continuous. Therefore, we only need to show that

$$\begin{aligned}\left|\frac{u-m_1}{s_1} - \frac{u-m_2}{s_2}\right| &= |u|\frac{|s_2-s_1|}{s_1 s_2} + \frac{|m_2 s_1 - m_1 s_2|}{s_1 s_2} \\ &= |u|\frac{|s_2-s_1|}{s_1 s_2} + |m_2|\frac{|s_2-s_1|}{s_1 s_2} + s_2\frac{|m_2-m_1|}{s_1 s_2} \\ &\leq K_{\mathcal{U}}\frac{|s_2-s_1|}{\underline{s}^2} + \bar{m}\frac{|s_2-s_1|}{\underline{s}^2} + \frac{|m_2-m_1|}{\underline{s}^2},\end{aligned}$$

where $K_{\mathcal{U}} = \sup\{|u| : u \in \mathcal{U}\}$. Thus, $\left|\frac{u-m_1}{s_1} - \frac{u-m_2}{s_2}\right| < \delta$ if $|m_2 - m_1|$ and $|s_2 - s_1|$ are sufficiently small. ■

5.8.2 Influence function: interchanging differentiation and integration

The first term of the right-hand-side of (5.27) can be expressed as

$$\begin{aligned}&\frac{d}{dt}\int_{m_X(t)}^{\infty}\int_{m_Y(t)}^{\infty}\psi_{c_1}\left(\frac{x-m_X(t)}{s_X(t)}\right)\psi_{c_1}\left(\frac{y-m_Y(t)}{s_Y(t)}\right)f(x,y)\,dy\,dx\Big|_{t=0} \\ &\quad + \frac{d}{dt}\int_{-\infty}^{m_X(t)}\int_{-\infty}^{m_Y(t)}\psi_{c_1}\left(\frac{x-m_X(t)}{s_X(t)}\right)\psi_{c_1}\left(\frac{y-m_Y(t)}{s_Y(t)}\right)f(x,y)\,dy\,dx\Big|_{t=0}. \quad (5.88)\end{aligned}$$

The first part of (5.88) can be written as

$$\frac{d}{dt}\int_{m_X(t)}^{\infty}\int_{m_Y(t)}^{\infty}\psi_{c_1}\left(\frac{x-m_X(t)}{s_X(t)}\right)\psi_{c_1}\left(\frac{y-m_Y(t)}{s_Y(t)}\right)f(x,y)\,dy\,dx\Big|_{t=0}$$

$$\begin{aligned}&= \lim_{t\to 0}\frac{1}{t}\left[\int_{m_X(t)}^{\infty}\int_{m_Y(t)}^{\infty}\psi_{c_1}\left(\frac{x-m_X(t)}{s_X(t)}\right)\psi_{c_1}\left(\frac{y-m_Y(t)}{s_Y(t)}\right)f(x,y)\,dy\,dx\right. \\ &\qquad \left. - \int_0^{\infty}\int_0^{\infty}\psi_{c_1}(x)\,\psi_{c_1}(y)\,f(x,y)\,dy\,dx\right]\end{aligned}$$

$$= \lim_{t\to 0} \frac{1}{t} \left[\int_{m_X(t)}^{\infty} \int_{m_Y(t)}^{\infty} \left\{ \psi_{c_1}\left(\frac{x - m_X(t)}{s_X(t)}\right) \psi_{c_1}\left(\frac{y - m_Y(t)}{s_Y(t)}\right) \right.\right.$$
$$\left.\left. - \psi_{c_1}(x)\,\psi_{c_1}(y) \right\} f(x,y)\,dy\,dx \right]$$
$$- \lim_{t\to 0} \frac{1}{t} \int_0^{m_X(t)} \int_0^{\infty} \psi_{c_1}(x)\,\psi_{c_1}(y)\,f(x,y)\,dy\,dx$$
$$- \lim_{t\to 0} \frac{1}{t} \int_0^{\infty} \int_0^{m_Y(t)} \psi_{c_1}(x)\,\psi_{c_1}(y)\,f(x,y)\,dy\,dx$$
$$- \lim_{t\to 0} \frac{1}{t} \int_0^{m_X(t)} \int_0^{m_Y(t)} \psi_{c_1}(x)\,\psi_{c_1}(y)\,f(x,y)\,dy\,dx \quad (5.89)$$

$$= \lim_{t\to 0} \left[\int_0^{\infty} \int_0^{\infty} \frac{1}{t} \left\{ \psi_{c_1}\left(\frac{x - m_X(t)}{s_X(t)}\right) \psi_{c_1}\left(\frac{y - m_Y(t)}{s_Y(t)}\right) - \psi_{c_1}(x)\,\psi_{c_1}(y) \right\} \right.$$
$$\left. I\big(x - m_X(t) > 0\big)\, I\big(y - m_Y(t) > 0\big) f(x,y)\,dy\,dx \right], \quad (5.90)$$

since the last three terms of (5.89) are zero. Now, assuming that $\psi(u)$, $\psi'(u)$ and $\psi'(u)u$ are bounded for all $u \in \mathbb{R}$ (under Regularity Condition A4), we can show that $\frac{d}{dt}\psi_{c_1}\left(\frac{X - m_X(t)}{s_X(t)}\right) \psi_{c_1}\left(\frac{Y - m_Y(t)}{s_Y(t)}\right)\Big|_{t=0}$ is bounded. Therefore, using Lebesgue's Dominated Convergence Theorem (see, for example, Bartle 1995, page 44), we have

$$\frac{d}{dt} \int_{m_X(t)}^{\infty} \int_{m_Y(t)}^{\infty} \psi_{c_1}\left(\frac{x - m_X(t)}{s_X(t)}\right) \psi_{c_1}\left(\frac{y - m_Y(t)}{s_Y(t)}\right) f(x,y)\,dy\,dx \Big|_{t=0}$$

$$= \int_0^{\infty} \int_0^{\infty} \lim_{t\to 0} \frac{1}{t} \left\{ \psi_{c_1}\left(\frac{x - m_X(t)}{s_X(t)}\right) \psi_{c_1}\left(\frac{y - m_Y(t)}{s_Y(t)}\right) - \psi_{c_1}(x)\,\psi_{c_1}(y) \right\} f(x,y)\,dy\,dx$$

$$= \int_0^{\infty} \int_0^{\infty} \frac{d}{dt} \psi_{c_1}\left(\frac{x - m_X(t)}{s_X(t)}\right) \psi_{c_1}\left(\frac{y - m_Y(t)}{s_Y(t)}\right) \Big|_{t=0} f(x,y)\,dy\,dx. \quad (5.91)$$

Similarly,

$$\frac{d}{dt} \int_{-\infty}^{m_X(t)} \int_{-\infty}^{m_Y(t)} \psi_{c_1}\left(\frac{x - m_X(t)}{s_X(t)}\right) \psi_{c_1}\left(\frac{y - m_Y(t)}{s_Y(t)}\right) f(x,y)\,dy\,dx \Big|_{t=0}$$

$$= \int_{-\infty}^{0} \int_{-\infty}^{0} \frac{d}{dt} \psi_{c_1}\left(\frac{x - m_X(t)}{s_X(t)}\right) \psi_{c_1}\left(\frac{y - m_Y(t)}{s_Y(t)}\right) \Big|_{t=0} f(x,y)\,dy\,dx. \quad (5.92)$$

Combining (5.91) and (5.92), we have

$$\frac{d}{dt}\left[E_{H_0}\left\{\psi_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{c_1}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)I\Big((X-m_X(t))(Y-m_Y(t))>0\Big)\right\}\right]_{t=0}$$
$$= E_{H_0}\left[\frac{d}{dt}\left\{\psi_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right)\psi_{c_1}\left(\frac{Y-m_Y(t)}{s_Y(t)}\right)\right\}\Big|_{t=0} I\big(XY>0\big)\right]. \quad (5.93)$$

We can now write (5.28) from (5.27).

In a similar way, we can show that

$$\frac{d}{dt}\left[E_{H_0}\left\{\psi^2_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right)I\Big((X-m_X(t))(Y-m_Y(t))>0\Big)\right\}\right]_{t=0}$$
$$= E_{H_0}\left[\frac{d}{dt}\left\{\psi^2_{c_1}\left(\frac{X-m_X(t)}{s_X(t)}\right)\right\}\Big|_{t=0} I\big(XY>0\big)\right], \quad (5.94)$$

which gives (5.38).

5.8.3 Asymptotic normality of the adjusted-Winsorized "covariance" estimate

The numerator of (5.2) can be written as

$$\hat{N} = \frac{1}{n}\sum_{i=1}^{n}\psi_{\mathbf{c}}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{\mathbf{c}}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) - \frac{1}{n}\sum_{i=1}^{n}\psi_{\mathbf{c}}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\frac{1}{n}\sum_{i=1}^{n}\psi_{\mathbf{c}}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right). \quad (5.95)$$

Now, we can express the first term of (5.95) as

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\mathbf{c}}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{\mathbf{c}}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) = \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\widehat{I}_i(c_1)$$
$$+\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\widehat{I}_i(c_2), \quad (5.96)$$

where $\widehat{I}_i(c_1) = I\big((X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y) > 0\big)$, and $\widehat{I}_i(c_2) = I\big((X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y) < 0\big)$. Denote $I_i(c_1) = I\big((X_i - \mu_X)(Y_i - \mu_Y) > 0\big)$, and $I_i(c_2) = I\big((X_i - \mu_X)(Y_i - \mu_Y) < 0\big)$. Let us focus on the cases when $\widehat{I}_i(c_1)$ and $I_i(c_1)$ take different values. Assuming (without loss of generality) that $\hat{\mu}_X < \mu_X$, we can argue that $\widehat{I}_i(c_1) \neq I_i(c_1)$ when

$$\hat{\mu}_X < X_i < \mu_X.$$

(Or, $\hat{\mu}_Y < Y_i < \mu_Y$.) Now, $\hat{\mu}_X - \mu_X$ is $O(1/\sqrt{n})$, which means $\widehat{I}_i(c_1) - I_i(c_1)$ is $O(1/\sqrt{n})$. Also, for the X_i's above, $\psi_{c_1}\big(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\big)$ is $O(1/\sqrt{n})$ since ψ is linear in the neighborhood of zero, while $\psi_{c_1}\big(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\big)$ is bounded. Therefore,

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{c_1}\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\right)\big(\widehat{I}_i(c_1) - I_i(c_1)\big) = O(1/n). \tag{5.97}$$

Similarly,

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{c_2}\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\right)\big(\widehat{I}_i(c_2) - I_i(c_2)\big) = O(1/n). \tag{5.98}$$

Using (5.97) and (5.98) in (5.96), we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \psi_{\boldsymbol{c}}\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{\boldsymbol{c}}\left(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\right) &\doteq \frac{1}{n}\sum_{i=1}^{n} \psi_{c_1}\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\right) I_i(c_1) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \psi_{c_2}\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\right) I_i(c_2), \end{aligned} \tag{5.99}$$

where "$\doteq$" means "asymptotically equivalent".

Let us now focus on the second term of (5.95). We can write

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\widehat{I}_i(c_1)+\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\widehat{I}_i(c_2)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)I_i(c_1)+\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)I_i(c_2)+O\left(1/\sqrt{n}\right). \quad (5.100)$$

Using Taylor expansion about (μ_X,σ_X) in (5.100), and expressing all the terms that involve $\hat{\mu}_X-\mu_X$ or $\hat{\sigma}_X-\sigma_X$ as $O\left(1/\sqrt{n}\right)$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)I_i(c_1)+\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\mu_X}{\sigma_X}\right)I_i(c_2)+O\left(1/\sqrt{n}\right). \quad (5.101)$$

Since $E\left[\psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)I(c_1)\right]=0$, we have $\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)I_i(c_1)=O\left(1/\sqrt{n}\right)$. Similarly, $\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)I_i(c_1)=O\left(1/\sqrt{n}\right)$. Using these results in (5.101), we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)=O\left(1/\sqrt{n}\right). \quad (5.102)$$

Similarly,

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)=O\left(1/\sqrt{n}\right). \quad (5.103)$$

Therefore, we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)=o\left(1/\sqrt{n}\right). \quad (5.104)$$

Using (5.99) and (5.104) in (5.95), we have

$$\hat{N}\doteq\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)I_i(c_1)$$
$$+\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)I_i(c_2). \quad (5.105)$$

Now, using Taylor expansion about (μ_X, σ_X), we have

$$\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right) = \psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right) - \frac{1}{\tilde{\sigma}_X}\psi'_{c_1}\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)(\hat{\mu}_X-\mu_X) - \frac{1}{\tilde{\sigma}_X}\psi'_{c_1}\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)(\hat{\sigma}_X-\sigma_X), \quad (5.106)$$

for $\hat{\mu}_X < \tilde{\mu}_X < \mu_X$ and $\hat{\sigma}_X < \tilde{\sigma}_X < \sigma_X$.

Similarly,

$$\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) = \psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right) - \frac{1}{\tilde{\sigma}_Y}\psi'_{c_1}\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)(\hat{\mu}_Y-\mu_Y) - \frac{1}{\tilde{\sigma}_Y}\psi'_{c_1}\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)(\hat{\sigma}_Y-\sigma_Y), \quad (5.107)$$

for $\hat{\mu}_Y < \tilde{\mu}_Y < \mu_Y$ and $\hat{\sigma}_Y < \tilde{\sigma}_Y < \sigma_Y$. Using (5.106) and (5.107) in the first part of the right-hand-side of (5.105), we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)I_i(c_1)$$

$$\doteq \frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)I_i(c_1) - \frac{1}{n\tilde{\sigma}_Y}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi'_{c_1}\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)I_i(c_1)\,(\hat{\sigma}_Y-\sigma_Y) - \frac{1}{n\tilde{\sigma}_X}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\psi'_{c_1}\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)I_i(c_1)\,(\hat{\sigma}_X-\sigma_X), \quad (5.108)$$

where the other terms are ignored since they are $o(1/\sqrt{n})$. Using Lemma 5.1, we have

$$\frac{1}{n\tilde{\sigma}_Y}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi'_{c_1}\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)\left(\frac{Y_i-\tilde{\mu}_Y}{\tilde{\sigma}_Y}\right)I_i(c_1) \xrightarrow[n\to\infty]{P} \frac{1}{\sigma_Y}E\left[\psi_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi'_{c_1}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)I(c_1)\right] = A_{c_1}, \quad (5.109)$$

and

$$\frac{1}{n\tilde{\sigma}_X}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\psi'_{c_1}\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)I_i(c_1)$$
$$\xrightarrow[n\to\infty]{P}\frac{1}{\sigma_X}E\left[\psi_{c_1}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi'_{c_1}\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)I(c_1)\right]=B_{c_1}, \quad (5.110)$$

where $I(c_1)=I\big((X-\mu_X)(Y-\mu_Y)>0\big)$. Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)I_i(c_1)$$
$$\doteq\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)I_i(c_1)-A_{c_1}(\hat{\sigma}_Y-\sigma_Y)-B_{c_1}(\hat{\sigma}_X-\sigma_X). \quad (5.111)$$

Since $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are S-scales, using (5.75) and (5.76) in (5.111), we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_1}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)I_i(c_1)$$

$$\doteq\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_1}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)I_i(c_1)$$
$$-\frac{A_{c_1}}{D_Y}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-b\right)-\frac{B_{c_1}}{D_X}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-b\right), \quad (5.112)$$

where

$$D_X=E\left[\chi'\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\right], \quad (5.113)$$

and

$$D_Y=E\left[\chi'\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{Y-\mu_Y}{\dot{\sigma}_Y}\right)\right]. \quad (5.114)$$

Similarly,

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{c_2}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)I_i(c_2)$$

$$\doteq \frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{c_2}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right) I_i(c_2)$$
$$-\frac{A_{c_2}}{D_Y}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-b\right)-\frac{B_{c_2}}{D_X}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-b\right), \quad (5.115)$$

where

$$A_{c_2}=\frac{1}{\sigma_X}E\left[\psi_{c_2}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi'_{c_2}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right) I(c_2)\right], \quad (5.116)$$

$$B_{c_2}=\frac{1}{\sigma_Y}E\left[\psi_{c_2}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi'_{c_2}\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right) I(c_2)\right], \quad (5.117)$$

and D_X and D_Y are as above. Using (5.112) and (5.115) in (5.105), we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi_{\boldsymbol{c}}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$

$$\doteq \frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi_{\boldsymbol{c}}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)$$
$$-\frac{A_{\boldsymbol{c}}}{D_Y}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-b\right)-\frac{B_{\boldsymbol{c}}}{D_X}\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-b\right), \quad (5.118)$$

where $A_{\boldsymbol{c}}=A_{c_1}+A_{c_2}$, and $B_{\boldsymbol{c}}=B_{c_1}+B_{c_2}$.

Using (5.118, we can state that

$$\sqrt{n}\left[\hat{N}-E\left\{\psi_{\boldsymbol{c}}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{\boldsymbol{c}}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}\right] \xrightarrow[n\to\infty]{d} \mathcal{N}(0,Q),$$

where

$$Q=\mathrm{Var}\left[\psi_{\boldsymbol{c}}\left(\frac{X-\mu_X}{\sigma_X}\right)\psi_{\boldsymbol{c}}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)-\frac{A_{\boldsymbol{c}}}{D_Y}\chi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)-\frac{B_{\boldsymbol{c}}}{D_X}\chi\left(\frac{X-\mu_X}{\sigma_X}\right)\right].$$

5.8.4 Difficulty with the denominator of (5.2)

Let us focus on the first term in the denominator of (5.2). It can be expressed as

$$\hat{D}_1=\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)-\left\{\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\right\}^2. \quad (5.119)$$

As in the case of the numerator, we can write

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{c}}^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)=\frac{1}{n}\sum_{i=1}^{n}\psi_{c_1}^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\widehat{I}_i(c_1)+\frac{1}{n}\sum_{i=1}^{n}\psi_{c_2}^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\widehat{I}_i(c_2). \tag{5.120}$$

Unfortunately, this time we cannot replace $\widehat{I}_i(.)$ by $I_i(.)$. The reason is as follows. When $\widehat{I}_i(c_1) \neq I_i(c_1)$, we have either (i) $\hat{\mu}_X < X_i < \mu_X$, or (ii) $\hat{\mu}_Y < Y_i < \mu_Y$. In the first case, as in the case of the numerator (see Section 5.8.3), we can show that the amount of error due to replacing $\widehat{I}_i(.)$ by $I_i(.)$ is $O(1/n)$. However, in the second case, though $\hat{\mu}_Y - \mu_Y$ is $O(1/\sqrt{n})$, the term $\psi_.\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$ (which is also $O(1/\sqrt{n})$ for the Y_i's above) is not there to make the product $O(1/n)$. Therefore, the amount of error in replacing the sample indicator functions by the true ones is $O(1/\sqrt{n})$, and cannot be ignored.

Chapter 6

Conclusion

In this study, we considered the problem of selecting linear prediction models for large high-dimensional datasets that possibly contain a fraction of contaminations. Our goal was to achieve robustness and scalability at the same time. We considered one-step and two-step model building procedures, the latter consisting of sequencing and segmentation steps. We will now summarize the main ideas proposed in this thesis, and the main results obtained.

One-step model building

We proposed robust versions of step-by-step algorithms FS and SW. We expressed these classical algorithms in terms of sample means, variances and correlations, and replaced these sample quantities by their robust counterparts to obtain the robust algorithms. We used robust correlations derived from a simplified version of bivariate M-estimates of the

scatter matrix. We proposed robust partial F-tests for stopping during the implementation of robust FS and SW procedures.

Our robust methods have much better performance compared to the standard FS and SW algorithms. Also, they are computationally very suitable, and scalable to large dimensions.

Two-step model building

Robust sequencing

We considered time-efficient algorithm LARS to sequence (some of) the d covariates to form a list such that the good predictors are likely to appear in the beginning. Since LARS is not resistant to outliers, we proposed two different approaches to robustify LARS. In the plug-in approach, we replaced the classical correlations in LARS by easily computable robust correlation estimates. In the data-cleaning approach, we first transformed the dataset by shrinking the outliers towards the bulk of the data (which we call multivariate-Winsorization), and then applied standard LARS on the transformed data. The data-cleaning approach is more time-consuming than the plug-in approach when only some of the predictors are being sequenced.

For both approaches (plug-in and data-cleaning), we used robust correlations derived from a simplified version of the bivariate M-estimates of the scatter matrix. We also proposed correlation estimates using bivariate-Winsorization of the data. We showed that the latter is faster to compute with important time differences when the number of

candidate predictors becomes large.

We recommend combining robust LARS with bootstrap to obtain more stable and reliable results. The reduced sets obtained by bootstrapped robust LARS contain more of the important covariates than the reduced sets obtained by initial robust LARS.

To obtain a reduced set of m covariates for further investigation, we introduced a learning curve that plots robust R^2 values versus dimension. An appropriate value of m is the dimension corresponding to the point where the curve starts to level off.

Robust segmentation

We perfromed all possible subsets regression on the reduced set of covariates obtained in the first step. Since classical selection criteria FPE, AIC, C_p, CV and bootstrap are sensitive to outliers, we needed robust selection criteria for this purpose. We identified certain limitations of Robust AIC (Ronchetti 1985) and robust CV (Ronchetti, Field and Blanchard 1997) methods. We proposed computationally suitable robust CV and robust bootstrap procedures in this thesis. We evaluated our methods using simulated and real datasets, and compared them with the classical methods as well as robust FPE proposed by Yohai (1997). Our robust CV and robust bootstrap methods have better performance compared to the classical methods and robust FPE.

Adjusted-Winsorized correlation estimate

For the development of robust LARS, we proposed this new correlation estimate for bivariate data. The proposed estimate is consistent and has bounded influence. We obtained its asymptotic variance and intrinsic bias. The tuning constants of this estimate can be chosen such that we have approximate Fisher-consistency. An smoothed version of this estimate is asymptotically normal. The computing time of this estimate is the same (approximately) as that of the univariate-Winsorized correlation estimate, but our estimate is more resistant to bivariate outliers.

Bibliography

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, **21**: 243–247.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, **22**: 203–217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory, Academiai Kiado, Budapest*, pages 267–281.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**: 125–127.

Alqallaf, F. A. (2003). *A new contamination model for robust estimation with large high-dimensional datasets.* PhD thesis, Department of Mathematics (Institute of Applied Mathematics), University of British Columbia.

Alqallaf, F. A., Konis, K. P., Martin, R. D., and Zamar, R. H. (2002). Scalable robust covariance and correlation estimates for data mining. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta*, pages 14–23.

Bartle, R. G. (1995). *The Elements of Integration and Lebesgue Measure.* John Wiley & Sons.

Bhansali, R. J. and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika*, **67**: 546–551.

Billingsley, P. (1986). *Probability and Measure.* John Wiley & Sons, 2nd edition.

Croux, C., Van Aelst, S., and Dehon, C. (2003). Bounded influence regression using high breakdown scatter matrices. *Ann. Inst. Statist. Math.*, **55(2)**: 265–285.

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

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

Frank, I. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**: 109–148.

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

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics.* John Wiley & Sons.

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

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

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Hubert, M. and Engelen, S. (2004). Fast cross-validation of high-breakdown resampling methods for PCA. unpublished manuscript.

Knight, W. R. (1966). A computer method calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association*, **61**: 436–439.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**: 79–86.

Lachenbruch, P. and Mickey, M. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, **10**: 1–11.

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

Mallows, C. L. (1995). More comments on C_p. *Technometrics*, **37**: 362–372.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, **4**: 51–67.

Mendenhall, W. and Sincich, T. (2003). *A Second Course in Statistics: Regression Analysis.* Pearson Education, Inc., New Jersey, 6th edition.

Morgenthaler, S., Welsch, R. E., and Zenide, A. (2003). Algorithms for robust model selection in linear regression. *Theory and Applications of Recent Robust Methods,* eds. M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, Basel (Switzerland): Birkhäuser-Verlag.

Ronchetti, E. (1985). Robust model selection in regression. *Statistics and Probability Letters*, **3**: 21–23.

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

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

Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, **79**: 871–880.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley-Interscience, New York.

Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**: 212–223.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis* (J. Franke, W. Härdle, and R. D. Martin, *eds.*), Lecture Notes in Statistics **26**, Springer Verlag, New York: 256–272.

Salibián-Barrera, M. (2000). *Contributions to the theory of robust inference.* PhD thesis, Department of Statistics, University of British Columbia.

Salibián-Barrera, M. and Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statistics*, **30**: 556–582.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

Shao, J. (1996). Bootstrap Model Selection. *Journal of the American Statistical Association*, **91**: 655–665.

Sommer, S. and Huggins, R. M. (1996). Variable selection using the Wald Test and Robust C_p. *Journal of the Royal Statistical Society, Ser. B*, **45**: 15–29.

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

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

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

Weisberg, S. (1985). *Applied Linear Regression.* Wiley-Interscience, New York, 2nd edition.

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*, **15**: 642–656.

Yohai, V. J. (1997). A new robust model selection criterion for linear models: RFPE. unpublished manuscript.