8-2007

Robust and Misspecification Resistant Model Selection in Regression Models with Information Complexity and Genetic Algorithms

Yan Liu
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Yan Liu entitled "Robust and Misspecification Resistant Model Selection in Regression Models with Information Complexity and Genetic Algorithms." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Business Administration.

, Major Professor

We have read this dissertation and recommend its acceptance:

Hamparsum Bozdogan

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Yan Liu entitled "Robust and Misspecification Resistant Model Selection in Regression Models with Information Complexity and Genetic Algorithms." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Business Administration.

Hamparsum Bozdogan
Major Professor

We have read this dissertation
and recommend its acceptance:

Kenneth Gilbert

Mary Leitnaker

Mohammed Mohsin

Accepted for the Council:

Carolyn R. Hodges
Vice Provost and Dean of the
Graduate School

(Original signatures are on file with official student records.)

Robust and Misspecification Resistant Model Selection in Regression Models with Information Complexity and Genetic Algorithms

A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Yan Liu
August 2007

Dedication

This dissertation is dedicated to my husband, Rui Zhang, and to my parents. Their endless love and support are my sources of spirit and encouragement.

Acknowledgments

I would like to thank all those who supported and helped me to complete this dissertation.

I am deeply indebted to my advisor, Dr. Hamparsum Bozdogan, for his guidance, advice and encouragement throughout my doctoral studies and research at the University of Tennessee, Knoxville.

I would like to thank Dr. Kenneth Gilbert and Dr. Leitnaker, who served as my committee members and gave me valuable comments on my dissertation. Also, I would like to thank them for giving me the flexibility in my teaching assistantship assignments, which helped me to have more time to concentrate on my research. I would like to thank Dr. Mohammed Mohsin, my external committee member, for his support and valuable comments.

Thanks to Dr. Frank Guess and Dr. Ramón Leon, for introducing me to General Electric for a great internship opportunity.

Specially, I would like to express my gratitude to my husband Rui Zhang. His endless love and support gave me the opportunity to complete this work.

Abstract

In this dissertation, we develop novel computationally efficient model subset selection methods for multiple and multivariate linear regression models which are both robust and misspecification resistant. Our approach is to use a three-way hybrid method which employs the information theoretic measure of complexity (ICOMP) computed on robust M-estimators as model subset selection criteria, integrated with genetic algorithms (GA) as the subset model searching engine.

Despite the rich literature on the robust estimation techniques, bridging the theoretical and applied aspects related to robust model subset selection has been somewhat neglected. A few information criteria in the multiple regression literature are robust. However, none of them is model misspecification resistant and none of them could be generalized to the misspecified multivariate regression. In this dissertation, we introduce for the first time both robust and misspecification resistant information complexity (ICOMP) criterion to fill in the gap in the literature.

More specifically in multiple linear regression, we introduce robust M-estimators with misspecification resistant ICOMP and use the new information criterion as the fitness function in GA to carry out the model subset selection. For multivariate linear regression, we derive the two-stage robust Mahalanobis distance (RMD) estimator and introduce this RMD estimator in the computation of information criteria. The new information criteria are used as the fitness function in the GA to perform the model subset selection.

Comparative studies on the simulated data for both multiple and multivariate regression show that the robust and misspecification resistant ICOMP outperforms the other robust information criteria and the non-robust ICOMP computed using OLS (or MLE) when the data contain outliers and error terms in the model deviate from a normal distribution. Compared with the all possible model subset selection, GA combined with the robust and misspecification resistant information criteria is proved to be an effective method which can quickly find the a near optimal subset, if not the best, without having to search the whole subset model space.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Model subset selection in linear regression has been playing an important role since the 1960's. Over the past a few decades of development, many model selection algorithms and criteria came to existence in the literature. Among them are some classical model selection procedures and the information theoretic based criteria. However, some critical issues related to the subset selection have been ignored, such as the misspecification of the model and robustness of the selection criteria. In this chapter, we will briefly introduce these subset selection methods in regression and the motivation and contribution of this dissertation. More details for the model selection algorithms and criteria are discussed in the next few chapters.

1.1 Subset Selection in Regression Models

Model subset selection in regression is to find out the relationship between the response variable(s) of interest and the potential predictor variables among the candidates of competing subset models.

1.1.1 Classical Model Subset Selection

Usually, the classical model selection methods are performed through the hypothesis tests. An arbitrary significance level is selected by the practitioners beforehand to decide whether the resulting model should include or exclude a certain predictor variable. However, many statisticians and other scientists have long been aware that the so-called significance levels used by the subset selection packages are totally without foundation (Linhart and Zucchini, 1986; Burnham and Anderson, 2002). Some scientists find that all the various hypothesis-testing approaches have no theoretical justification and may often perform poorly (Burnham and Anderson, 2002).

Many classical model selection procedures exist in almost all of the popular statistics packages, such as the forward selection, backward elimination and stepwise selection. However, both forward and backward procedures cannot deal with the collinearity in the predictor variables. Boyce et al. (1974) criticizes the backward, forward and stepwise selection that "little or no theoretical justification exists for the order in which variables enter or exit the algorithm." A major criticism on stepwise selection is that "it rarely finds the overall best model or even the best subset of a particular size of the model" (Mantel, 1970; Hocking, 1976, 1983; Moses, 1986). Another criticism on stepwise selection is that "it, at the very best, can only produce an 'adequate' model" (Sokal and Rohlf, 1981).

All these shortcomings inherent in different classical model selection procedures put limitations on selecting the optimal subset or nearly optimal subset in the regression models. Some statisticians and researchers thus prefer the all-possible subset selection procedure to choose the best model. However, in many cases, this method is not computationally feasible in a reasonable time and is rather expensive. The total possible number of subsets can reach over millions (if we have more than 20 predicator variables) or even billions (if we have more than 30 predictor variables) of models to evaluate.

One big disadvantage of the forward selection, backward elimination, stepwise selection and all possible subset selection is that the performances of all the procedures depend on whether the selection criteria are appropriately chosen. The performances will be vulnerable to the existence of the extreme values or outliers in the data if the selection criteria are not "robust"; they will be easily contaminated by the misspecified functional form of the model (such as the departure from the assumption of Gaussian distribution of the residuals) if the selection criteria are not resistant to the model misspecification.

1.1.2 Information Theoretic Model Selection

In the regression model context, different criteria corresponding to different assumptions are used to select subset models, such as minimizing the mean squared error, maximizing the likelihood function, Mallow's C_p, etc., which are all motivated by the reduction of model space.

Motivated from a very different point of view, information theoretic criteria are introduced and developed in the model selection literature. The basic idea of these types of criteria is trying to minimize the Kullback-Leibler distance between the distribution of response variable(s) under the optimal subset model and under the true model. Akaike (1973) firstly derived the information theoretic criterion known as Akaike's information criterion (AIC), which provides a new paradigm of model selection in the analysis of empirical data. Many

other information criteria have been proposed since then, such as Bayesian information criterion (BIC or SBC) (Schwarz, 1978), Generalized Akaike's information criterion (GAIC) (Shibata, 1989; Bozdogan, 2000), Bozdogan's informational complexity (ICOMP) (Bozdogan, 1988a,b, 1990, 1994a, 2000; Bozdogan and Bearse, 2003) etc. Burnham and Anderson (1998; 2002) recommend the information-theoretic approach for the analysis of data from observational studies. They state that "Inference from multiple models, or the selection of a single 'best' model, by methods based on the Kullback-Leibler distance are almost certainly better than other methods commonly in use now (e.g., null hypothesis testing of various sorts, the use of R^2, or merely the use of just one available model)."

The form of different information criteria will be discussed in detail in the following chapter.

1.1.3 Robust Model Selection

Over the past 30 years, many robust estimation procedures are devised as alternatives to the classical least squares procedures. The purpose of the robustness is to make the estimation insensitive to small deviations from the model assumptions and resistant to unusual observations in the data. Numerous works have been done in this area, such as (Andrews et al., 1972; Huber, 1981, 1996, 2004; Holland and Welsch, 1977; Hampel et al., 1986; Rousseeuw and Leroy, 1987; Olive, 2005).

Despite the rich literature on the robust estimation techniques, bridging the theoretical and applied aspects related to robust model subset selection has been somewhat neglected (Ronchetti, 1985, 1997). An incomplete list of the robust model selection literature can be briefly discussed as follows. Both Ronchetti (1985) and Hampel (1983) introduced a robust version of Akaike's information criterion on model selection procedures for regression models. The difference between the two robust versions of AIC is that they used different penalty terms. Ronchetti and Staudte (1994) presented a robust version of Mallows's C_p for regression models. Ronchetti, et al. (1997) presented a cross-validation method for the robust model selection. Machado (1993) derives a robust version of BIC or SBC by defining M-estimators on the objective function for a parametric model. Qian and Künsch (1996) presented a robust criterion on Rissanen's stochastic complexity (SC).

1.1.4 Model Selection under Misspecification

Researchers may misspecify the regression models in a number of possible ways (Godfrey, 1988). These are: the incorrect functional form of the model; the multicollinearity among the predictor variables; the skewness and kurtosis in the variables which cause the

non-normality of the disturbances; the autocorrelation and heteroskedasticity of the disturbances. In many circumstances, one may not be sure whether the model is "correctly specified". The incorrectly specified models may result in inconsistent estimates of parameters and the standard inferential techniques can be invalidated (White, 1982).

Some studies have been done to detect model misspecification and investigate the consequences of it. Berk (1966; 1970) considers the consistency question of maximum likelihood estimators in the frame of Bayesian method and emphasizes the information theoretic interpretation. Huber (1967) considers the same question independently from Berk's work in a more classical way under general conditions. He concludes that the maximum likelihood estimator converges to a well-defined limit even when the probability model is not correctly specified. However he does not explicitly discuss the information theoretic interpretation of this limit (White, 1982). Akaike (1973) emphasizes this information theoretic interpretation and indicates that when the true distribution is unknown, the maximum likelihood estimator is a natural parameter estimator which minimizes the Kullback-Leibler information criterion (Kullback and Leibler, 1951). White (1982) studies the consequences and detection of model misspecification when maximum likelihood techniques are used in less general conditions than Huber's. He provides specification robust procedures based on both inner-product form (also known as "Hessian form") and outer-product form of the Fisher information matrix.

Bozdogan as well as his colleagues (Bozdogan, 2004a,b; Howe and Bozdogan, 2007; Magnus, 2007) develop new ICOMP-type criteria for model subset selection based on his original work (Bozdogan, 1988a,b, 1990, 1994b, 2000, 2004a). This misspecification resistant version of ICOMP, known as $\text{ICOMP}_{\text{misspec}}$, allows for non-Gaussian errors. They indicate in their paper (Howe and Bozdogan, 2007; Magnus, 2007) that "based on the existing literature, to our knowledge, there is no other criteria to date which penalizes the presence of skewness and the inflation of kurtosis in the model selection process except ICOMP."

1.2 Motivation

The linear model subset selection did not make much progress since it first began in the 1960s. In the first edition of his book, Miller (1990) provides a good source of comprehensive summary of subset selection approaches prior to 1990. But in the second edition of his book (2002), he indicates that "there has been very little real progress" in the subset selection since 1990.

The robust model selection problems have been ignored in the literature compared with the prosperous development in the robust estimation methods in the past three decades (Ronchetti, 1985, 1997).

Furthermore, in the literature, there is not much attention paid to the misspecification of the fitted model (Howe and Bozdogan, 2007) despite its significance. Chatfield (1995) states the seriousness of the misspecification as "Model misspecification is a major, if it is not the dominant, source of error in the quantification of most scientific analysis."

Because of such background, new model selection criteria which can solve both the robustness and misspecification problems are invoked. In this dissertation, we are to introduce and develop a new hybridized method of model subset selection, which is robust and at the same time misspecification resistant in both multiple linear regression (MLR) and multivariate regression (MVR) models.

Our model selection criteria are generated from Bozdogan's information-theoretic measure of complexity (Bozdogan, 1988a,b, 1990, 1994a, 2000; Bozdogan and Bearse, 2003) but put in the context of robust estimators.

For computational efficiency, Genetic Algorithm (GA) is applied to select the optimal or near optimal subset of predictor variables, in which the robust version of information-theoretic measure of complexity (ICOMP) is used as the fitness function.

1.3 Contributions

Some critical contributions are made by this dissertation to the model subset selection literature, which include the following.

Firstly, in this dissertation, we introduce the robust estimators to Bozdogan's (Bozdogan and Bearse, 2003) information theoretic measure of complexity (ICOMP) for misspecified model, so that the new criterion is robust and at the same time misspecification resistant for the model subset selection. This criterion is proved to be an effective method in the simulation studies and on real world applications in this dissertation. To our knowledge, there has not been any work done both on robustness and model misspecification at the same time in the model selection literature.

Secondly, we generalize the robust and misspecification resistant information criterion to the multivariate regression (MVR) model selection. No such criterion has been applied to the multivariate models so far.

Thirdly, the robust and misspecification resistant information criterion is further hybridized with genetic algorithm (GA), which drastically speeds up the model selection processes and make them feasible in a timely manner that is not costly. In other words, genetic algorithm is able to find the optimal or near optimal subset model without having to search the full model space. The new version of information complexity serves as our fitness function in GA.

1.4 Organization of Dissertation

This dissertation consists of 7 chapters. In Chapter 1, we introduce the current model subset selection techniques in regression models and their pros and cons. The motivation, contribution and organization of this dissertation are proposed. In Chapter 2, different robust estimation methods which exist in the literature are reviewed with emphasis on four types of robust M-estimators and their properties. In Chapter 3, we discuss the information criteria for model subset selection. ICOMP and the misspecification resistant version of ICOMP are introduced. In Chapter 4, genetic algorithms are developed. The advantages and disadvantages of GA are discussed. The graphical user interface (GUI) of GA used in our MATLAB programming for this dissertation is given to illustrate how we hybridize GA with robust and misspecification resistant information criteria. In Chapter 5, we develop the robust and misspecification resistant of ICOMP(IFIM), namely RICOMP(IFIM)$_{\text{misspec}}$. Comparative study is performed on a Monte Carlo simulation data using different robust information criteria. RICOMP(IFIM)$_{\text{misspec}}$ outperforms the other robust information criteria and the non-robust ICOMP(IFIM) computed using OLS estimator. Two real world data examples are presented to show the effectiveness of the GA subset selection. In Chapter 6, we derive an iteratively robust Mahalanobis Distance (RMD) estimator for the multivariate linear regression model. Robust and misspecification resistant ICOMP are developed using RMD estimator. A three-way hybrid method is presented when we take this robust and misspecification resistant ICOMP as the fitness function in GA. Comparative study is carried out on the multivariate Monte Carlo simulation data. In our study, the robust and misspecification resistant ICOMP outperforms the others. GA for model subset selection method is proved to be efficient. A real data example is presented at the end. Lastly Chapter 7, consists of conclusions and suggestions for further future research.

Chapter 2

Robust Estimators

For a very long time, the least squares approach and its generalizations have been the main estimation methods for regression models. Without the presence of "outliers," they have nice properties and serve us very well. However, these types of estimators are too sensitive to the outliers so that the resulting residual analysis may be misleading. Outliers are not uncommon as obtaining data becomes easier nowadays. These outliers are either from the heavy tailed distributions of the model or the extreme data observations which mostly results from the errors. Under such circumstances, new alternative approaches were created to substitute least squares estimation, which are more resistant to outliers and would have been influenced much less by the outliers. These new approaches are referred to as the robust estimation methods or techniques.

Since the first creation of the term "robustness" by Box (1953), numerous works have been done in this area. Huber was among those who contributed the most to this literature. His fundamental paper (Huber, 1964) can be taken as the milestone of the robust estimation. Since then, he has provided great details on both mathematical aspects and summaries in the following articles (Huber, 1972, 1973, 1977). Two of his books (Huber, 1977, 1981) were republished in 1996 and 2004, respectively. A number of books and articles besides Huber's are propagating since the 1960's. Among them are Tukey (1962; 1972), Hampel (1974), Andrews et al. (1972), Andrews (1974), Hogg (1974; 1979a; 1979b), Rousseeuw and Leroy (1987), Olive (2005) and most recently, Maronna et al. (2006) etc., to mention a few.

Numerous robust procedures are available to the researchers and practitioners, such as the M-estimator, L-estimator, R-estimator, S-estimator and various generalizations such as the adaptive versions. In this dissertation, we focus on the M-estimator.

2.1 M-Estimator (Maximum Likelihood Type Estimates)

M-estimator is introduced by Huber (1964). It is based on the modification of the principle of maximum likelihood estimation.

Let $x_1, x_2, \ldots, x_n$ be a random sample arising from the probability density of $f(x \mid \theta)$, where θ is a location parameter. The logarithm of the likelihood function of θ is

$$\ln L(\theta \mid x) = \sum_{i=1}^{n} \ln f(x_i \mid \theta). \tag{2.1}$$

The maximum likelihood estimate (MLE) of θ, denoted by $\hat{\theta}_{MLE}$ is given by

$$\hat{\theta}_{MLE} = \arg\min_{\hat{\theta}} \sum_{i=1}^{n} \left[-\ln f(x_i \mid \theta) \right]. \tag{2.2}$$

The performance of MLE depends on the assumed distribution of the data. It can be biased and inefficient when distributions depart from normality or are heavy tailed with outliers.

Huber (1964) generalized the maximum likelihood estimation by using the function $\rho(x_i \mid \theta)$ to substitute $[-\ln f(x_i \mid \theta)]$ in equation 2.1 and take the latter as a special case.

The generalized maximum likelihood estimation, or in short M-estimation is to minimize

$$\sum_{i=1}^{n} \rho(x_i \mid \theta).$$

The M-estimator is given by

$$\hat{\theta}_M = \arg\min_{\hat{\theta}} \sum_{i=1}^{n} \rho(x_i \mid \theta), \tag{2.3}$$

where ρ is a symmetric, positive-definite function with a unique minimum at zero (Rousseeuw and Yohai, 1984). The properties of ρ are therefore given as follows:

1. $\rho(x) = \rho(-x)$
2. $\rho(x) \geq 0$
3. $\rho(0) = 0$
4. if $|x_i| > |x_j|, \rho(x_i) \geq \rho(x_j)$.

Suppose that the minimization ($\min \sum_{i=1}^{n} \rho(x_i \mid \theta)$) can be achieved by differentiating the ρ function with respect to θ, which gives

$$\sum_{i=1}^{n} \psi(x_i \mid \theta) = 0, \tag{2.4}$$

where $\psi(x) = \rho'(x)$, is the first derivative of the ρ function.

When the ψ function is monotonic, the solution to equation 2.4 is called monotonic M-estimator. When the ψ function is non-monotonic (or "redescending"), the solution to equation 2.4 is called redescending M-estimator.

The solution of the M-estimators is not equivariant with respect to scale. We need to find the scale invariant version of the M-estimators, which is to find the solution of

$$\sum_{i=1}^{n} \psi\left(\frac{x_i \mid \theta}{\hat{\sigma}}\right) = 0, \tag{2.5}$$

where $\hat{\sigma}$ is a robust estimate of scale. Two possibilities are (Hogg, 1979a):

$$\hat{\sigma}_1 = 1.4826 \times MAD, \text{ where } MAD = \text{median}\,|x_i - \text{median}(x_i)| \tag{2.6}$$

or

$$\hat{\sigma}_2 = 0.7413 \times IQR, \text{ where IQR} = 75^{th} \text{ percentile - } 25^{th} \text{ percentile.} \tag{2.7}$$

In robust M-estimation, we need to determine a ψ function, so that the resulting estimator will guard against some percentages of outliers and produce efficient estimators (Hogg, 1979a).

For many choices of robust functions ρ and ψ, there are not closed forms for the solutions. Optimization algorithms (such as Newton-Raphson) or iterative methods are required to compute the M-estimators. In the iterative procedure, a weight function is used to recalculate the weight on the observations, which is defined as

$$w_i(x_i) = \begin{cases} \psi\left(\frac{x_i}{\hat{\sigma}}\right)/\left(\frac{x_i}{\hat{\sigma}}\right), & \text{if } x_i \neq 0; \\ 1, & \text{if } x_i = 0. \end{cases} \tag{2.8}$$

Various M-estimators have been proposed in the literature. Here, we propose the four of them used in this dissertation.

2.1.1 Huber's Minimax Function

Huber (1964) derives the robust ρ and ψ functions as follows:

$$\rho(z) = \begin{cases} \frac{1}{2}z^2, & \text{for } |z| \leq k; \\ k|z| - \frac{1}{2}k^2, & \text{for } |z| > k. \end{cases} \tag{2.9}$$

$$\psi(z) = \begin{cases} z, & \text{if } |z| \leq k; \\ k\text{sgn}(z), & \text{if } |z| > k. \end{cases} \tag{2.10}$$

In the above equations, k is a given constant, called "tuning constant".

The ρ and ψ functions are associated with a distribution that is "normal" in the middle and "double exponential" in both tails (Hogg, 1979a). The corresponding M-estimator is the minimax solution of the asymptotic variance of the estimator T.

$$\min_T \max_F [asym.\ var(T)],$$

where F ranges over the set of all $F = (1-\varepsilon)\Phi + \varepsilon H$ distributions for fixed ε and symmetric H (i.e., ε-contaminated normal distribution) and $T = T_n(x_1, x_2, \ldots, x_n)$ is the estimator over the sample (Huber, 1964).

The related weight function can be calculated as

$$w(z) = \begin{cases} 1, & \text{if } |z| \leq k; \\ \frac{k}{|z|,} & \text{if } |z| > k. \end{cases} \tag{2.11}$$

One reasonable value for the tuning constant suggested is $k = 2$ in the Princeton Study (Andrews et al., 1972). Another reasonable suggestion for k is $k = 1.5$ (Hogg, 1979b).

2.1.2 Andrews' Sine Wave Function

Andrews presents the sine wave function of M-estimate in Andrews et al. (1972) and Andrews (1974):

$$\rho(z) = \begin{cases} c[1 - \cos(z/c)], & \text{if } |z| \leq c\pi; \\ 2c, & \text{if } |z| > c\pi. \end{cases} \tag{2.12}$$

$$\psi(z) = \begin{cases} \sin(z/c), & \text{if } |z| \leq c\pi; \\ 0, & \text{if } |z| > c\pi. \end{cases} \tag{2.13}$$

The related weight function is given by

$$w(z) = \begin{cases} \frac{\sin(z/c)}{z/c,} & \text{if } |z| \leq c\pi; \\ 0, & \text{if } |z| > c\pi. \end{cases} \tag{2.14}$$

The tuning constant $c = 1.5$ or $c = 2.1$ is suggested for using this function (Hogg, 1979b).

2.1.3 Tukey's Biweight Function

Tukey's biweight function is given by (Beaton and Tukey, 1974)

$$\rho(z) = \begin{cases} \frac{z^2}{2} - \frac{z^4}{2c^2} + \frac{z^6}{6c^4}, & \text{if } |z| \leq c; \\ \frac{c^2}{6,} & \text{if } |z| > c. \end{cases} \tag{2.15}$$

$$\psi(z) = \begin{cases} z\Big(1 - (z/c)^2\Big)^2, & \text{if } |z| \leq c; \\ 0, & \text{if } |z| > c. \end{cases} \tag{2.16}$$

The related weight function is

$$w(z) = \begin{cases} \Big(1 - (z/c)^2\Big)^2, & \text{if } |z| \leq c; \\ 0, & \text{if } |z| > c. \end{cases} \tag{2.17}$$

Tuning constant suggested is $c = 6.0$ (Hogg, 1979a).

2.1.4 Hampel's Function

Hampel's ψ function (Andrews et al., 1972; Hampel, 1974) is given by

$$\rho(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| \leq a; \\ a|z| - \frac{1}{2}a^2, & \text{if } a < |z| \leq b; \\ \frac{a\left(c|z| - \frac{1}{2}z^2\right)}{c-b} - (7/6)a^2, & \text{if } b < |z| \leq c; \\ a(b + c - a), & \text{if } |z| > c. \end{cases} \tag{2.18}$$

$$\psi(z) = \begin{cases} z, & \text{if } |z| \leq a; \\ a\text{sgn}(z), & \text{if } a < |z| \leq b; \\ \frac{a\text{sgn}(z)(c - |z|)}{c-b}, & \text{if } b < |z| \leq c; \\ 0, & \text{if } |z| > c. \end{cases} \tag{2.19}$$

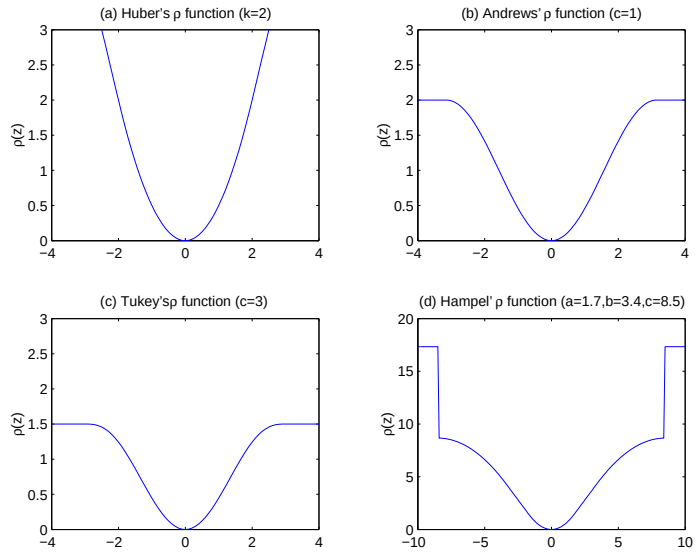

Figure 2.1: Example ρ Functions of M-estimation: (a) Huber's ρ function with k=2; (b) Andrews' ρ function with c=1; (c) Tukey's ρ function with c=3; (d) Hampel's ρ function with a=1.7, b=3.4, c=8.5.

The related weight function is

$$w(z) = \begin{cases} 1, & \text{if } |z| \le a; \\ a/|z|, & \text{if } a < |z| \le b; \\ \frac{a(c-|z|)}{|z|(c-b)}, & \text{if } b < |z| \le c; \\ 0, & \text{if } |z| > c, \end{cases} \tag{2.20}$$

where $a = 1.7$, $b = 3.4$, $c = 8.5$ refers to as Hampel's 17A function and $a = 1.2$, $b = 3.5$, $c = 8.0$ refers to as Hampel's 12A function in the Princeton Study (Andrews et al., 1972).

Interested readers can find more examples of reasonable tuning constant values in Holland and Welsch (1977).

Among the four M-estimates, Huber's minimax function gives the monotone M-estimators. The other three robust functions give the redescending M-estimators.

Figures 2.1 and 2.2 show examples of ρ and ψ functions of the four M-estimators with fixed tuning constants.

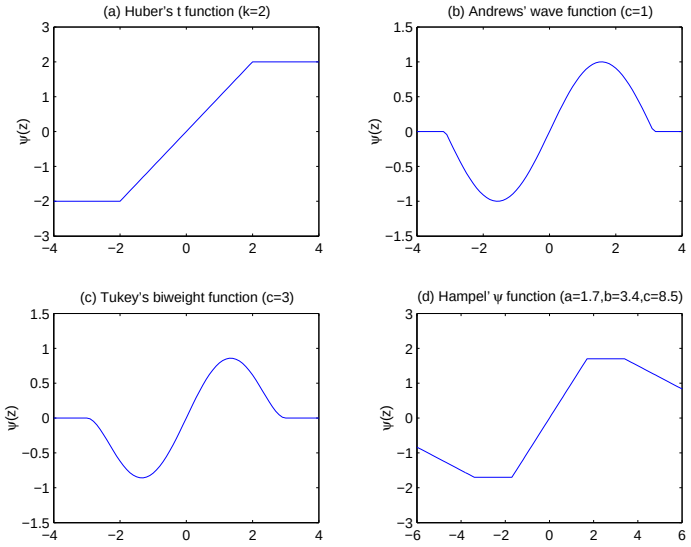

Figure 2.2: Example ψ Functions of M-estimation: (a) Huber's ψ function with k=2; (b) Andrews' ψ function with c=1; (c) Tukey's ψ function with c=3; (d) Hampel's ψ function with a=1.7, b=3.4, c=8.5.

2.2 Properties of Robust M-estimators

2.2.1 Influence Function

The influence function (IF) or influence curve (IC) is essentially the first derivative of an estimator (Hampel, 1974). It is used to derive the asymptotic variance and study the robustness properties of an estimator. It is an asymptotic version of its sensitivity curve.

The influence function of an M-estimator is proportional to the ψ function (Huber, 2004, page 45). Let T be an M-estimator and let the ψ function exist. Let F be a probability distribution and $T(F)$ has definition. The influence function IF is given by

$$IF(x; F, T) = \frac{\psi\left(x; T(F)\right)}{-\int \left(\partial/\partial\theta\right) \psi\left(x; T(F)\right) F(dx)}. \tag{2.21}$$

2.2.2 Breakdown Point

Simply speaking, the breakdown point (BDP) of an estimator is the largest proportion of unusual points in the data before destroying the analysis. Intuitively, the largest BDP is 50% since we can not distinguish the good data points and bad data points if more than half of the data are contaminated. The closer the BDP is to 50%, the more robust the estimator is.

Hampel (1971) gives the definition of BDP. Donoho and Huber (1983) give the definition of the breakdown point for the finite sample (FBP). We will focus on the FBP and follow the definition from Hampel et al. (1986).

The definition of finite breakdown point (FBP) is as follows.

Definition 2.2.1. Let $(x_1, \ldots, x_n)$ be a random sample, the finite-sample breakdown point ε_n^* of the estimator T_n is given by

$\varepsilon_n^*\left(T_n; x_1, \ldots, x_n\right) = \frac{1}{n} \max\left\{m; \max \sup_{y_1,\ldots,y_m} |T_n(z_1, \ldots, z_n)| < \infty\right\}$,

where the sample $(z_1, \ldots, z_n)$ is obtained by replacing the m data points $x_{i1}, \ldots, x_{im}$ by $y_1, \ldots, y_m$.

The FBP usually does not depend on the random sample and slightly depends on the sample size.

It is shown that the FBP of monotonic M-estimate is zero (Maronna et al., 2006).

2.2.3 Asymptotic Normality

It is shown in Huber (2004) that M-estimators are asymptotically normally distributed.

Definition 2.2.2. Assume $x_1, \ldots, x_n$ are independent random variables with common distribution P. Suppose that sequence $T_n = T_n(x_1, \ldots, x_n)$ satisfies $\frac{1}{\sqrt{n}}\Sigma\psi(x_i, T_n) \to 0$ and

its consistency has already been proved by some other means. Then, T_n in probability is asymptotically normal.

2.3 Other Robust Estimators

2.3.1 L-Estimator (Linear Combinations of Order Statistics)

Consider a random sample of n observations from a continuous type distribution. The order statistics of the sample are given by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. An L-estimator is defined to be a linear combination of these order statistics.

Sample median is the simplest L-estimator. The other examples of L-estimators are: α-trimmed mean, Gastwirth's estimator and Tukey's trimean (Hogg, 1979b). More L-estimators can be found in Andrews, et al. (1972).

Knoenker and Basset (1978) generalize L-estimators to the regression situation by using the following definition of quantiles

$$\rho(r_i) = \begin{cases} -(1-p)r_i, & r_i < 0; \\ pr_i, & r_i \geq 0, \end{cases} \tag{2.22}$$

where r_i is the residual from the i^{th} data observation to the location estimate. It is obvious that when $p = 1/2$, the quantile corresponds to the sample median.

The major disadvantage for L-estimators is that they are relying on the value of noise contamination rate, $1-p$, and are not easy to optimize. Since they ignore part of the data, they are among the least efficient estimators (Nasraoui, 2004).

2.3.2 R-Estimator (Estimates Derived from Rank Tests)

We illustrate R-estimator by taking linear regression model as an example. R-estimator is to replace one factor in the residual squares used by OLS estimator by the rank of the residuals. Mathematically, instead of minimizing the sum of squared residuals ($\min \sum_{i=1}^{n} r_i^2$), we minimize the sum of the product of the residual and the rank of the residual ($\min \sum_{i=1}^{n} r_i R_i$), where r_i is the residual for the i^{th} observation; R_i is the rank of the i^{th} residual ($R_i = 1, 2, \ldots, n$). In more generalized form, the rank of the residual in the optimization function is replaced by a score function of the residual rank. That is, we wish to find $\min \sum_{i=1}^{n} r_i a(R_i)$, where $a(\cdot)$ denotes a nondecreasing score function of the rank of the residuals, such that $a_{(1)} \leq a_{(2)} \leq \ldots a_{(n)}$. Two examples of the scores are (Hogg, 1979b):

1. Wilcoxon scores: $a(R_i) = R_i$, $R_i = 1, 2, \ldots, n$ is the rank

2. Median scores:

$$a(R_i) = \begin{cases} 1, & \text{if } R_i > (n+1)/2; \\ -1, & \text{if } R_i \leq (n+1)/2. \end{cases}$$

One disadvantage of R-estimators is that they are not easy to optimize and the practitioners need prior information about the noise contamination rate (Nasraoui, 2004).

Jurečková (1977) proves that under certain conditions, the R-estimators and M-estimators are asymptotically equivalent. Because of this, it seems more reasonable to use M-estimators instead of R-estimators, since they are much easier to compute.

2.3.3 S-Estimator (Estimates Derived from Scale Estimation)

S-estimators for regression were first introduced by Rousseeuw and Yohai (1984). They are so called "S-estimators" because they are based on estimators of scale. S-estimators are created by the motivation of seeking high Breakdown Point (BDP) regression estimators which at the same time share the nice asymptotic properties of robust M-estimators (Rousseeuw and Yohai, 1984). Simply speaking, breakdown point (BDP) refers to the fractions of contaminated data. The highest BDP we can achieve is 50%, since the data can not be discriminated by "good" or "bad" if more than half (50%) of the data are contaminated. The rigorous asymptotic definition for the BDP of large samples is given by Hampel (1971). Donoho and Huber (1983) introduced another version of BDP for finite samples.

S-estimators are obtained through the one-dimensional estimators of scale defined by a function ρ satisfying (Rousseeuw and Yohai, 1984):

- (R1) ρ is symmetric, continuously differentiable and $\rho(0) = 0$;
- (R2) There exists $c > 0$, such that ρ is strictly inceasing on $[0, c]$ and constant on $[c, \infty]$.

For any sample $\{r_1, r_2, \ldots, r_n\}$ of real numbers, the scale estimate $s(r_1, r_2, \ldots, r_n)$ is defined as the solution of

$$\frac{1}{n} \sum_{i=1}^{n} \rho(r_i/s) = K, \tag{2.23}$$

where $K = E_\phi[\rho]$, where ϕ is the standard normal distribution.

The formal definition of S-estimators for regression is given by Rousseeuw and Yohai (1984) as follows.

Definition. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a sample from regression data with p-dimensional x_i. For each vector $\boldsymbol{\theta}$, residuals $r_i(\boldsymbol{\theta}) = y_i - x_i^t \boldsymbol{\theta}$ of which the dispersion $s\left(r_i(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta})\right)$ is calculated by the equation $\frac{1}{n} \sum_{i=1}^{n} \rho(r_i/s) = K$, where ρ satisfies (R1) and (R2).

The S-estimator $\hat{\boldsymbol{\theta}}$ is defined by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\hat{\boldsymbol{\theta}}} s\left(r_1(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta})\right). \tag{2.24}$$

And the final scale estimator is given by

$$\hat{\sigma} = s\left(r_1(\hat{\boldsymbol{\theta}}), \ldots, r_n(\hat{\boldsymbol{\theta}})\right). \tag{2.25}$$

Rousseeuw and Yohai (1984) also give an example of the ρ-function

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4}, & \text{for } |x| \leq c; \\ \frac{c^2}{6,} & \text{for } |x| \geq c. \end{cases} \tag{2.26}$$

The derivative of which is Tukey's biweight function.

In their paper, Rousseeuw and Yohai (1984) also give the breakdown point for S-estimators along with their asymptotic behavior.

Despite the attractive high BDP and other asymptotic properties of S-estimators, the main disadvantage of S-estimators is how to find an effective algorithm to calculate them. Ruppert (1992) is one researcher who tried to make his algorithm (which he called SIR-REAL) computationally feasible.

2.3.4 Others

Besides the above robust estimators we discussed, there are a lot of other robust estimators with certain asymptotic properties in the literatures, such as the generalized M-estimators (or "GM-estimators" in short) (Huber, 1981), the least median of squares (LMS) estimator (Rousseeuw, 1984), the least trimmed sum of squares (LTS) estimator (Rousseeuw, 1983), the MM-estimator (Yohai, 1987) and the τ-estimator (Yohai and Zamar, 1988), to mention a few.

2.4 Numerical Example - Stack Loss Data

To show the working of M-estimators effectively, we will use the stack loss data as an example.

The stack loss data is taken from Brownlee (1965, page 454) and was analyzed as a benchmark data for robust regression and outlier detection by a large number of researchers (Rousseeuw and Leroy, 1987). The data are measurements of a plant oxidizing ammonia to nitric acid on 21 consecutive days. The variables are:

- response y: stack loss,

- predictor x_1: air flow,
- predictor x_2: cooling water inlet temperature,
- predictor x_3: acid concentration.

According to the literature, most researchers conclude that observations 1, 3, 4 and 21 are outliers in this data set. Daniel and Wood (1971) find the unusually large residual of observation 21, which has considerable influence on the estimated coefficients of the fitted model. They delete observation 21 and three other observations 1, 3 and 4, and fit the variables x_1, x_2 and x_1^2 to the remaining 17 observations. Andrews (1974) analyzes these four unusual observations and fit the variables x_1, x_2 and x_3 by his robust method (Andrews sine wave function) with and without these four points and compares his results with the models fitted by OLS method with and without these points.

Here, we analyze the data in a similar way to that of Andrews'. We fit the variables x_1, x_2 and x_3 by OLS method to all the observations and to the remaining observations (with points 1, 3, 4 and 21 deleted). Then, we fit the variables x_1, x_2 and x_3 to all the observations by robust methods (Huber, Andrews, Tukey and Hampel) and compare the results with that of OLS. The fitted equations for the stack loss data by both OLS methods and robust methods are summarized in Table 2.1. As seen in the table, the four robust methods fit the equations consistent with the OLS method with observations 1, 3, 4 and 21 removed, which is denoted as OLS_2 in the table. Figure 2.3 shows the plots of the standardized residuals of all the fitted models. It suggests that the residual plot from original OLS only indicate observation 21 as a suspicious point. The residual plot from the OLS_2 (with four outliers deleted) and those residual plots from the robust methods are able to identify all the four unusual data points. Figure 2.4 shows the Quantile-Quantile plots (QQ plots) of the residuals for all the fitted models. Once again, the QQ plots from robust methods are identical to the one obtained from the OLS_2 analysis with observations 1, 3, 4 and 21 deleted (Figure 2.4b).

Other than M-estimation, there are some other robust methods and generalizations in the literature that provide good alternatives to least squares estimation.

Table 2.1: Stackloss Data: Fitted Equations

	OLS	OLS_2[a]	Huber	Andrews	Tukey	Hampel
Tuning constants			k=1.5	c=1.5	c=6.0	a=1.2; b=3.5; c=8.0;
$\hat{\beta}_0$	-39.92	-37.65	-38.79	-37.27	-37.68	-38.42
$\hat{\beta}_1$	0.716	0.798	0.833	0.813	0.822	0.882
$\hat{\beta}_2$	1.295	0.577	0.724	0.535	0.545	0.517
$\hat{\beta}_3$	-0.152	-0.067	-0.110	-0.071	-0.075	-0.098

[a]OLS with observations 1, 3, 4 and 21 deleted

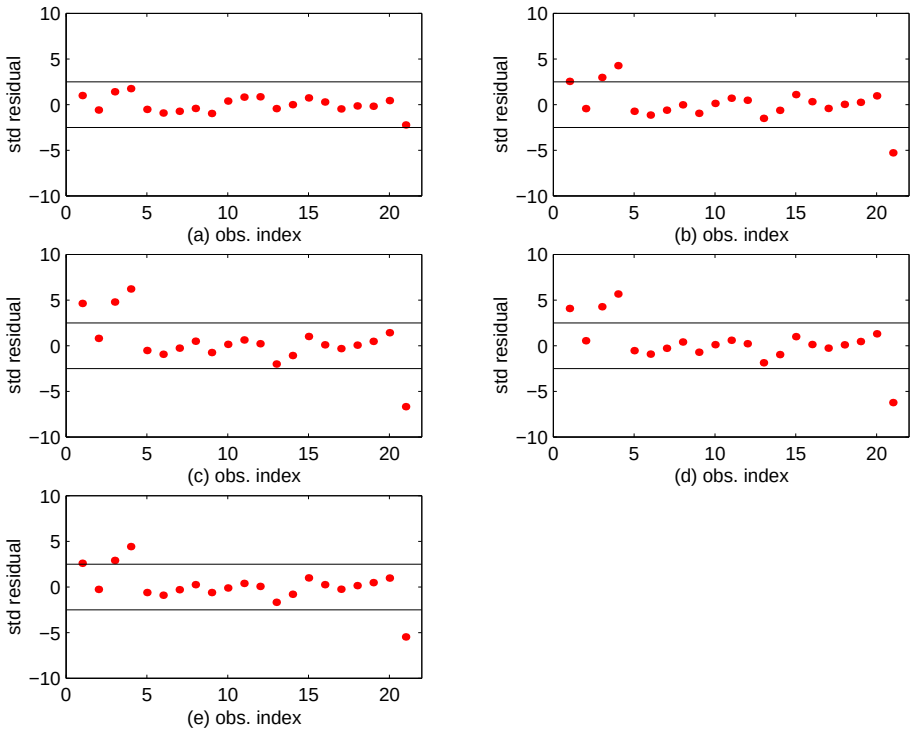

Figure 2.3: Stackloss Data: Plots of Standardized Residuals. (a) standardized residual from least-squares estimation. (b) standardized residual from Huber estimation. (c) standardized residuals from Andrews estimation. (d) standardized residuals from Tukey estimation. (e) standardized residuals from Hampel estimation.

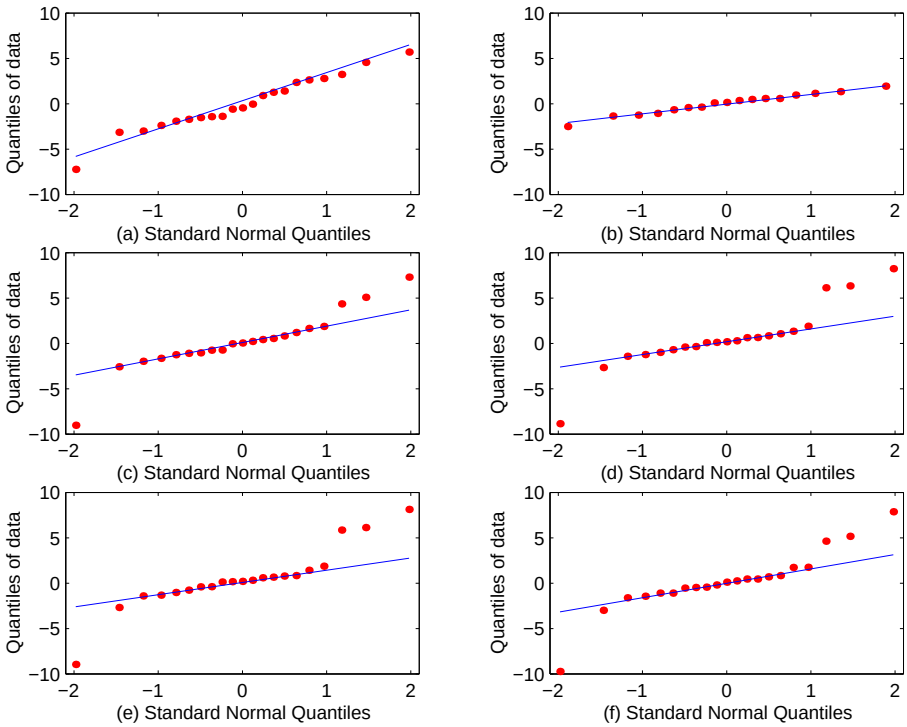

Figure 2.4: Stackloss Data: Quantile-Quantile plots of the residuals. (a) QQplot residuals from least-squares (with all observations). (b) QQplot residuals from least-squares (without observations 1, 3, 4 and 21). (c) QQplot residuals from Huber estimation. (d) QQplot residuals from Andrews estimation. (e) QQplot residuals from Tukey estimation. (f) QQ-plot residuals from Hampel estimation.

Chapter 3

Information Criteria

3.1 Introduction

Many information theoretic based criteria were developed in the model subset selection and evaluation literature since the first derivation of Akaike's information criterion (AIC) (Akaike, 1973).

Akaike's (1973) AIC makes a compromise between the maximized log likelihood (the lack of fit component) and the number of free parameters estimated in the model (the penalty component), which is given by

$$\text{AIC} = -2\text{log}L(\hat{\boldsymbol{\theta}}) + 2k, \tag{3.1}$$

where $\text{log}L(\hat{\boldsymbol{\theta}})$ is the natural logarithm of maximized likelihood function, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the parameter vector $\boldsymbol{\theta}$ and k is the number of free parameters in the model. AIC is an unbiased estimator of minus twice the expected log likelihood. The model with minimum AIC will be chosen to be the best to fit the data.

Based on Akaike's original AIC, many model-selection procedures which take the form of a penalized likelihood (a negative log likelihood plus a penalty term) have been proposed (Sclove, 1987).

Schwarz's (1978) Bayesian information criterion (SBC, also known as BIC) is given by

$$\text{SBC} = -2\text{log}L(\hat{\boldsymbol{\theta}}) + k\text{log}(n), \tag{3.2}$$

where $\text{log}L(\hat{\boldsymbol{\theta}})$, $\hat{\boldsymbol{\theta}}$ and k have the same meanings as in AIC; n is the number of observations.

Generalized Akaike's information criterion (GAIC) (Shibata, 1989; Bozdogan, 2000) is defined by

$$\text{GAIC} = -2\text{log}L(\hat{\boldsymbol{\theta}}) + 2tr(\widehat{\mathcal{F}}^{-1}\widehat{\mathbf{R}}), \tag{3.3}$$

where $\widehat{\mathcal{F}}$ is the estimated Fisher information matrix (FIM) in inner product or Hessian form, $\widehat{\mathbf{R}}$ is the estimated Fisher information matrix in outer product. $tr(\widehat{\mathcal{F}}^{-1}\widehat{\mathbf{R}})$ is the Lagrange Multiplier test (LMT) statistic. GAIC is also known as Takeuchi's (1976) information criterion (TIC), or AIC_T.

There are some other AIC-type information criteria, such as the consistent Akaike's information criterion (CAIC) (Bozdogan, 1987b), consistent AIC with Fisher information (CAICF) (Bozdogan, 1987b), and corrected information criterion (AIC_c) (Sugiura, 1978; Hurvich and Tsai, 1990), to mention a few.

3.2 Bozdogan's Information Theoretic Measure of Complexity - ICOMP

Motivated by the similar considerations in AIC, Bozdogan (1987a; 1988a; 1988b; 1990) introduced a new entropic or information-theoretic measure of complexity called ICOMP as an alternative criterion for model subset selection, which is based on the structural complexity of a set of random vectors via a generalization of the information-based covariance complexity index of van Emden (1971). This criterion is an additive composition of two information-based complexities of the covariance matrix of the parameter estimates of a model, and the covariance matrix of the residuals (Bozdogan, 1990).

ICOMP is designed to estimate a loss function, which is given in the following form

$$\textbf{Loss} = \textbf{lack of fit} + \textbf{lack of parsimony} + \textbf{profusion of complexity}. \tag{3.4}$$

The third term of equation 3.4 refers to the interdependencies or the correlations among the parameter estimates and the random error term of a model. Equation 3.4 provides a trade-off between lack of fit and a scalar measure of the accuracy of the parameter estimates (Bozdogan, 2000).

Instead of penalizing the free parameters directly, ICOMP penalizes the covariance complexity of the model. The definition of ICOMP is given as (Bozdogan, 1988a, 1990)

$$\text{ICOMP} = -2\log L(\hat{\boldsymbol{\theta}}) + 2\text{C}(\widehat{\Sigma}_{\text{Model}}), \tag{3.5}$$

where $\log L(\hat{\boldsymbol{\theta}})$ is the natural logarithm of the maximized likelihood function, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the parameter vector $\boldsymbol{\theta}$. $\text{C}(\cdot)$ represents a real-valued complexity measure and $\widehat{Cov}(\hat{\boldsymbol{\theta}}) = \widehat{\Sigma}_{\text{Model}}$ represents the estimated covariance matrix of the parameter vector of the model. Two forms of $\text{C}(\cdot)$ are defined in Bozdogan's paper (Bozdogan, 1988a, 1990), which are given in next section.

Bozdogan (2000) makes some criticisms of AIC-type information criteria and compares his ICOMP with those criteria. He pointed out that "Although AIC is invariant under parameter transformations, it does not have the virtue of detecting the problem caused by the curvature of a model, especially in univariate and multivariate nonlinear models." This has also been known by other researchers. AIC often overfits the model. "AIC and AIC-type criteria are based on MLE's, which often are biased and they do not fully take into account the concept of parameter redundancy, accuracy, and the parameter interdependencies in model fitting and selection process." The difference between ICOMP class criteria and AIC-type criteria is that "with ICOMP we have the advantage of working with both biased as well as unbiased estimates of the parameters and measure the complexities of their covariances to study the robustness properties of different methods of parameter estimates." "ICOMP provides a more judicious penalty term than AIC and AIC-type criteria, since counting and penalizing the number of parameters in a model is necessary but by no means sufficient." "Model complexity depends intrinsically on many factors other than the model dimension, such as the several forms for parameter redundancy, parameter stability, random error structure of the model, and the linearity and nonlinearity of the parameters of the model, etc."

In some of his recent papers, Bozdogan showed that ICOMP class criteria overwhelmingly agree more often with KL distance than AIC-type criteria (Bozdogan and Haughton, 1998) and that ICOMP class criteria clearly performed better than AIC-type criteria in model selection, prediction and perturbation studies (Bozdogan, 2000).

3.2.1 Definition of Covariance Complexity

Following Van Emden (1971), we give the initial definition of information complexity of a covariance matrix $\mathbf{\Sigma}$ for the multivariate distribution as (Bozdogan, 1988a)

$$\mathrm{C}_0(\mathbf{\Sigma}) = \frac{1}{2}\sum_{i=1}^{p}\log(\sigma_{jj}) - \frac{1}{2}\log|\mathbf{\Sigma}|, \tag{3.6}$$

where $\sigma_{jj} \equiv \sigma_j^2$ is the j^{th} diagonal element of $\mathbf{\Sigma}$, p is the dimension of $\mathbf{\Sigma}$ and $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$. Note that $\mathrm{C}_0(\mathbf{\Sigma}) = 0$ when $\mathbf{\Sigma}$ is a diagonal matrix (i.e., the variables are linearly independent) and $\mathrm{C}_0(\mathbf{\Sigma})$ is infinity when there is linear dependency among the variables. Van Emden (1971) indicated that equation 3.6 is not an effective measure of the amount of complexity in the covariance matrix $\mathbf{\Sigma}$ since:

- $\mathrm{C}_0(\mathbf{\Sigma})$ depends on the marginal and common distributions of the random variables and

- The first term of $C_0(\mathbf{\Sigma})$ in equation 3.6 would change under orthonormal transformations.

To improve this, Bozdogan proposes an information-theoretic maximal measure of complexity of a covariance matrix $\mathbf{\Sigma}$ of the multivariate normal distribution as (Bozdogan, 1990, Proposition 3.1)

$$C_1(\mathbf{\Sigma}) = \frac{p}{2}\log\left[\frac{tr(\mathbf{\Sigma})}{p}\right] - \frac{1}{2}\log|\mathbf{\Sigma}|, \tag{3.7}$$

where $tr(\mathbf{\Sigma})$ refers to the trace of $\mathbf{\Sigma}$, $|\mathbf{\Sigma}|$ denotes the determinant of $\mathbf{\Sigma}$, and $p = dim(\mathbf{\Sigma})$.

$C_1(\mathbf{\Sigma})$ in equation 3.7 is an upper bound of $C_0(\mathbf{\Sigma})$ in equation 3.6. $C_1(\mathbf{\Sigma})$ is independent of the coordinate system associated with the variance $\sigma_j^2 \equiv \sigma_{jj}, j = 1, 2, \ldots, p$. Different from $C_0(\mathbf{\Sigma})$, $C_1(\mathbf{\Sigma})$ is invariant to scalar multiplication and orthonormal transformations. Furthermore, $C_1(\mathbf{\Sigma})$ is a monotonically increasing function of the dimension p of $\mathbf{\Sigma}$ (Magnus and Neudecker, 1999).

3.2.2 ICOMP(IFIM) – The Sum of Two Kullback-Leibler Distances

When we use the inverse-Fisher information matrix (IFIM) to estimate the covariance matrix $\mathbf{\Sigma}$ in equation 3.7, ICOMP gives its most general form, denoted by ICOMP(IFIM).

For a multivariate normal linear or nonlinear structural model, the general form of ICOMP(IFIM) is defined as (Bozdogan, 2004a, Proposition 2.2)

$$\text{ICOMP(IFIM)} = -2\log L(\hat{\boldsymbol{\theta}}) + 2C_1(\widehat{\mathcal{F}}^{-1}(\hat{\boldsymbol{\theta}})), \tag{3.8}$$

where $\widehat{\mathcal{F}}^{-1}(\hat{\boldsymbol{\theta}})$ is the estimated inverse-Fisher information matrix (IFIM) based on the estimated parameter vector $\hat{\boldsymbol{\theta}}$, which is also known as the Cramér Rao lower bound (CRLB) matrix (Cramér, 1946; Rao, 1945, 1947, 1948)) and exploits the asymptotic optimality properties of MLE's (Bozdogan, 2004a); and $C_1(\widehat{\mathcal{F}}^{-1}(\hat{\boldsymbol{\theta}}))$ denotes the maximal informational complexity of $\widehat{\mathcal{F}}^{-1}(\hat{\boldsymbol{\theta}})$. The first component of equation 3.8 measures the lack of fit of the model and the second term is a scale measure of the complexity of the estimated inverse Fisher information matrix, which takes into account of the accuracy of the estimated parameters.

Based on equation 3.7, $C_1(\widehat{\mathcal{F}}^{-1}(\hat{\boldsymbol{\theta}}))$ can be calculated as

$$C_1(\widehat{\mathcal{F}}^{-1}(\hat{\boldsymbol{\theta}})) = \frac{s}{2}\log\left[\frac{tr(\widehat{\mathcal{F}}^{-1})}{s}\right] - \frac{1}{2}\log|\widehat{\mathcal{F}}^{-1}|, \tag{3.9}$$

where $s = \text{rank}(\widehat{\mathcal{F}}^{-1})$ and

$$\mathcal{F}^{-1} = \text{Cov}(\hat{\boldsymbol{\theta}}) = -E\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{(\partial\theta)(\partial\theta')}\right]. \tag{3.10}$$

Under regularity conditions, $\hat{\boldsymbol{\theta}}$, the MLE of $\boldsymbol{\theta}$, asymptotically follows a normal distribution with covariance matrix $\widehat{\mathcal{F}}^{-1}$. Consequently, $\text{C}_1(\widehat{\mathcal{F}}^{-1})$ measures the Kullback-Leibler (1951) distance against the independence of the estimated parameters. Therefore, ICOMP(IFIM) can be viewed as the sum of two KL distances. The first measure of the KL distance is between the true data and the fitted model and the second measure of the KL distance is against the independence of the parameter estimates (Bozdogan and Bearse, 2003).

3.3 ICOMP(IFIM) for Regression Models

3.3.1 ICOMP(IFIM) for Multiple Linear Regression (MLR) Models

Consider a multiple linear regression model given in

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.11}$$

where $\mathbf{y}$ is a vector of $(n \times 1)$ observations on a dependent variable; $\mathbf{X}$ is a matrix of $(n \times p)$ nonstochastic predetermined predictor variables $(p = k + 1)$; $\boldsymbol{\beta}$ is a vector of $(p \times 1)$ coefficients and $\boldsymbol{\varepsilon}$ is a vector of $(n \times 1)$ random errors, which follows $\varepsilon_i \sim N(0, \sigma^2)$, for $i = 1, 2, \ldots, n$, or equivalently, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$.

The density function of the regression model for a particular sample $(x_1, x_2, \ldots, x_n)$ is

$$f(y_i|x_i, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{(y_i - x_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]. \tag{3.12}$$

The likelihood function of the sample is thus

$$L(\boldsymbol{\beta}, \sigma^2|y, \mathbf{X}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{(y - \mathbf{X}\boldsymbol{\beta})'(y - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right], \tag{3.13}$$

and the log likelihood function is

$$l(\boldsymbol{\beta}, \sigma^2|y, \mathbf{X}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{(y - \mathbf{X}\boldsymbol{\beta})'(y - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}. \tag{3.14}$$

The maximum likelihood estimates (MLE's) of $(\boldsymbol{\beta}, \sigma^2)$, $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ are:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}y; \tag{3.15}$$

$$\widehat{\sigma}^2 = \frac{(y - \mathbf{X}\boldsymbol{\beta})'(y - \mathbf{X}\boldsymbol{\beta})}{n} = \frac{RSS}{n}, \tag{3.16}$$

where RSS is the residual sum of squares.

The maximum likelihood covariance matrix of the estimated regression coefficients is

$$\widehat{Cov}(\hat{\boldsymbol{\beta}})_{MLE} = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{3.17}$$

The estimated inverse Fisher information matrix (IFIM) is given by (Bozdogan, 2004a)

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = \widehat{\mathcal{F}}^{-1} = \begin{bmatrix} \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\widehat{\sigma}^4}{n} \end{bmatrix}. \tag{3.18}$$

Using the estimated inverse Fisher information matrix (IFIM) in equation 3.18 and defining $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)$, the ICOMP(IFIM) for multiple regression model is given by (Bozdogan, 2004a)

$$\begin{aligned} \text{ICOMP(IFIM)}_{reg} &= -2\log L(\hat{\boldsymbol{\theta}}) + 2\text{C}_1\left(\widehat{\mathcal{F}}^{-1}(\widehat{\boldsymbol{\theta}})\right) \\ &= n\log(2\pi) + n\log(\widehat{\sigma}^2) + n + 2\text{C}_1\left(\widehat{\mathcal{F}}^{-1}(\widehat{\boldsymbol{\theta}})\right), \end{aligned} \tag{3.19}$$

where

$$\text{C}_1\left(\widehat{\mathcal{F}}^{-1}(\widehat{\boldsymbol{\theta}})\right) = \frac{p}{2}\log\left[\frac{tr\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} + \frac{2\widehat{\sigma}^4}{n}}{p}\right] - \frac{1}{2}\log|\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}| - \frac{1}{2}\log\left(\frac{2\widehat{\sigma}^4}{n}\right). \tag{3.20}$$

In equation 3.18, when $(\mathbf{X}'\mathbf{X})^{-1}$ increases, the variance $\widehat{\sigma}^2$ decreases and when the variance $\widehat{\sigma}^2$ increases, $(\mathbf{X}'\mathbf{X})^{-1}$ decreases. Thus, $\text{C}_1(\widehat{\mathcal{F}}^{-1})$ obtains a trade-off between the two extremes and guards against multicollinearity (Bozdogan, 2004a).

3.3.2 ICOMP(IFIM) for Multivariate Linear Regression (MVR) Models

The classical MVR model in compact matrix form is given by

$$\mathbf{y} = \mathbf{XB} + \mathbf{E}, \tag{3.21}$$

where

- $\mathbf{Y}$ is an $(n \times m)$ matrix of n independent observations on m responses;

- $\mathbf{X}$ is an $(n \times p)$ matrix of n observations on k nonstochastic independent variables augmented by a constant column $\mathbf{1}$, where $p = k + 1$;
- $\mathbf{B}$ is a $(p \times m)$ matrix of regression coefficients;
- $\mathbf{E}$ is an $(n \times m)$ matrix of random errors, which satisfies

$$Cov(\mathbf{E}) = \mathbf{\Sigma} \otimes \mathbf{I_n}, \tag{3.22}$$

where $\otimes$ denotes the Kronecker product.

Similar to the work of Howe and Bozdogan, and Magnus (Howe and Bozdogan, 2007; Magnus, 2007), we do not assume the normal distribution of the observation and get the maximum likelihood function. However, we obtain the quasi-maximum likelihood estimators for $\mathbf{B}$ and $\mathbf{\Sigma}$ by maximizing the normal likelihood. Suppose $\boldsymbol{\theta} = ((vec\mathbf{B})', (vec\mathbf{\Sigma})')'$, the normal likelihood function is given by

$$l(\boldsymbol{\theta}) = -\frac{nm}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}tr(\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{XB})'. \tag{3.23}$$

The quasi-maximum likelihood estimators, $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Sigma}}$ are

$$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}; \tag{3.24}$$

and

$$\widehat{\mathbf{\Sigma}} = \frac{\left(\mathbf{y} - \mathbf{X}\widehat{\mathbf{B}}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\mathbf{B}}\right)}{n}. \tag{3.25}$$

The information matrix $\mathcal{F}$ is given by

$$\mathcal{F} = \begin{pmatrix} \mathbf{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2}D_m'(\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1})D_m \end{pmatrix}. \tag{3.26}$$

Therefore, the inverse inner product form of the information matrix is

$$\mathcal{F}^{-1} = \begin{pmatrix} \mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2}{n}D_m^{+}(\mathbf{\Sigma} \otimes \mathbf{\Sigma})D_m^{+\prime} \end{pmatrix}. \tag{3.27}$$

Where D_m^{+} is the Moore-Penrose inverse of the duplication matrix D_m.

ICOMP(IFIM) for MVR model is defined by

$$\text{ICOMP(IFIM)}_{\text{MVR}} = nm\log 2\pi + n\log\left|\widehat{\mathbf{\Sigma}}\right| + nm + 2C_1(\widehat{\mathcal{F}}^{-1}), \tag{3.28}$$

where

$$C_1\left(\widehat{\mathcal{F}}^{-1}\right) = \frac{s}{2}\log\left(\frac{tr\widehat{\mathcal{F}}^{-1}}{s}\right) - \frac{1}{2}\log|\widehat{\mathcal{F}}^{-1}| \tag{3.29}$$

Here $s = rank(\widehat{\mathcal{F}}^{-1}) = mp + \frac{1}{2}m(m+1)$,

$$tr\widehat{\mathcal{F}}^{-1} = (tr\widehat{\boldsymbol{\Sigma}})\left(tr(\mathbf{X}'\mathbf{X})^{-1}\right) + \frac{1}{2n}\left(tr\widehat{\boldsymbol{\Sigma}}^2 + (tr\widehat{\boldsymbol{\Sigma}})^2 + 2\sum_{j=1}^{m}\sigma_{jj}^2\right), \tag{3.30}$$

and

$$|\widehat{\mathcal{F}}^{-1}| = 2^m n^{-\frac{1}{2}m(m+1)}|\widehat{\boldsymbol{\Sigma}}|^{m+p+1}|\mathbf{X}'\mathbf{X}|^{-m}, \tag{3.31}$$

where σ_{jj}^2 is the j^{th} diagonal element of $\widehat{\boldsymbol{\Sigma}}$.

We refer the readers to Howe and Bozdogan, and Magnus (Howe and Bozdogan, 2007; Magnus, 2007) for proofs.

3.4 Bozdogan's Information Criteria under Model Misspecification

3.4.1 Two Forms of the Fisher Information Matrix

Suppose the probability density function of the underlying model is $f(x|\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}$ be a consistent maximum likelihood estimator (MLE) of $\boldsymbol{\theta}^*$. The "inner product form" (or "Hessian" form) of the Fisher information matrix is defined as

$$\mathcal{F} = -E\left[\frac{\partial^2 \log f(\mathbf{X}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}, \tag{3.32}$$

and the "outer product form" of the Fisher information matrix is defined as

$$\mathcal{R} = E\left[\frac{\partial \log f(\mathbf{X}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]\left[\frac{\partial \log f(\mathbf{X}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}. \tag{3.33}$$

These two forms of the Fisher information matrices are useful to check the misspecification of the model (Bozdogan, 2004a). For an independently and identically distributed sample $x_1, x_2, \ldots, x_n$ and under standard regularity conditions (Lehmann, 1983), we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \sim N(\mathbf{0}, \mathcal{F}^{-1}\mathcal{R}\mathcal{F}^{-1}) \tag{3.34}$$

as $n \to \infty$ (Howe and Bozdogan, 2007; Magnus, 2007).

The covariance matrix

$$Cov(\boldsymbol{\theta}^*) = \mathcal{F}^{-1}\mathcal{R}\mathcal{F}^{-1} \tag{3.35}$$

is called robust covariance matrix (or "sandwich" covariance matrix). The estimation of robust covariance matrix was first introduced by Huber (1967) and White (1982), which produces an asymptotically consistent covariance matrix estimator for dependent data without having to make distributional assumptions. In other words, the robust covariance matrix estimation gives a correct covariance matrix regardless of the correctness of the assumed model $f(x|\boldsymbol{\theta})$.

3.4.2 Bozdogan's Misspecification-Resistant Information Criteria

The general form of Bozdogan's information-theoretic measure of complexity under model misspecification (Bozdogan, 2004a), ICOMP(IFIM)$_{misspec}$, is defined by

$$\text{ICOMP(IFIM)}_{misspec} = -2\log L(\widehat{\boldsymbol{\theta}}) + 2\text{C}_1(\widehat{Cov}(\widehat{\boldsymbol{\theta}})_{misspec}), \tag{3.36}$$

where

$$\widehat{Cov}(\widehat{\boldsymbol{\theta}})_{misspec} = \widehat{\mathcal{F}}^{-1}\widehat{R}\widehat{\mathcal{F}}^{-1}. \tag{3.37}$$

Note that in equation 3.37, when the model is correctly specified and certain regularity conditions hold (White, 1982), the inner product form and outer product form of the Fisher information matrix are the same.

$$\widehat{\mathcal{F}} = \widehat{\mathcal{R}}$$

and the information matrix can be expressed by the inverse of either Hessian form $\widehat{\mathcal{F}}^{-1}$ or outer product form $\widehat{\mathcal{R}}^{-1}$.

$$\widehat{Cov}(\widehat{\boldsymbol{\theta}}) = \widehat{\mathcal{F}}^{-1} = \widehat{\mathcal{R}}^{-1} \tag{3.38}$$

Equation 3.36 thus reduces to equation 3.8.

In the case of model misspecification, the inner and outer product forms are different from each other, i.e.,

$$\widehat{\mathcal{F}} \neq \widehat{\mathcal{R}}$$

3.5 ICOMP(IFIM)$_{\text{misspec}}$ for Regression Models

3.5.1 ICOMP(IFIM)$_{\text{misspec}}$ for Multiple Linear Regression Models

Consider the multiple linear regression (MLR) model given in equation 5.1 or 5.2. The general form of Bozdogan's information criterion under model misspecification, denoted by ICOMP(IFIM)$_{misspec}$ (Bozdogan, 2004a), is given by

$$\text{ICOMP(IFIM)}_{misspec} = n\log(2\pi) + n\log(\widehat{\sigma}^2) + n + 2\text{C}_1(\widehat{Cov}(\hat{\boldsymbol{\theta}})_{misspec}), \tag{3.39}$$

where $\widehat{Cov}(\hat{\boldsymbol{\theta}})_{misspec}$ is calculated by equation 3.37.

In the regression case, the estimated inverse Fisher information matrix in inner-product form is

$$\widehat{\mathcal{F}}^{-1} = \begin{bmatrix} \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\widehat{\sigma}^4}{n} \end{bmatrix} \tag{3.40}$$

and the estimated outer-product form of the Fisher information matrix is

$$\widehat{\mathcal{R}} = \begin{bmatrix} \frac{1}{\widehat{\sigma}^4}\mathbf{X}'\mathbf{D^2}\mathbf{X} & \mathbf{X}'\mathbf{1}\frac{Sk}{2\widehat{\sigma}^3} \\ \left(\mathbf{X}'\mathbf{1}\frac{Sk}{2\widehat{\sigma}^3}\right)' & \frac{(n-p)(Kt-1)}{4\widehat{\sigma}^4} \end{bmatrix}, \tag{3.41}$$

where $\mathbf{D}^2 = diag(\widehat{\varepsilon}_1^2, \widehat{\varepsilon}_2^2, \ldots, \widehat{\varepsilon}_n^2)$ and $\mathbf{X}$ is an $(n \times p)$ matrix of predictor variables. $\mathbf{1}$ is an $(n \times 1)$ vector of ones. Sk and Kt are the coefficients of skewness and Kurtosis respectively, which are given by

$$Sk = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^3\right)}{\widehat{\sigma}^3}, \tag{3.42}$$

and

$$Kt = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^4\right)}{\widehat{\sigma}^4}. \tag{3.43}$$

Substituting equations 3.40 and 3.41 to equation 3.37, the estimated robust covariance matrix is given by

$$\widehat{Cov}(\widehat{\boldsymbol{\theta}})_{misspec} = \begin{bmatrix} \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\widehat{\sigma}^4}{n} \end{bmatrix} \begin{bmatrix} \frac{1}{\widehat{\sigma}^4}\mathbf{X}'\mathbf{D^2}\mathbf{X} & \mathbf{X}'\mathbf{1}\frac{Sk}{2\widehat{\sigma}^3} \\ \left(\mathbf{X}'\mathbf{1}\frac{Sk}{2\widehat{\sigma}^3}\right)' & \frac{(n-p)(Kt-1)}{4\widehat{\sigma}^4} \end{bmatrix} \begin{bmatrix} \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\widehat{\sigma}^4}{n} \end{bmatrix}. \tag{3.44}$$

3.5.2 ICOMP(IFIM)$_{\text{misspec}}$ for Multivariate Linear Regression (MVR) Models

In equation 3.21, we define the standardized $\mathbf{y}$ as $\mathbf{V} = (\mathbf{y} - \mathbf{XB})\boldsymbol{\Sigma}^{-1/2}$ so that $E(\mathbf{V}) = 0$ and $var(vec\mathbf{V}) = I_{mn}$. The matrix generalization of skewness $\boldsymbol{\Gamma}_1$ is given by

$$\boldsymbol{\Gamma}_1 = E(vec\mathbf{V})\left(vec(\mathbf{V}'\mathbf{V} - nI_m)\right)', \tag{3.45}$$

and the matrix generalization of kurtosis $\boldsymbol{\Gamma}_2$ is given by

$$\boldsymbol{\Gamma}_2 = E(vec\mathbf{V}'\mathbf{V})(vec\mathbf{V}'\mathbf{V})'. \tag{3.46}$$

The inner product form of the information matrix $\mathcal{F}$ is given in equation 3.26 and the inverse inner product form of the information matrix is given in equation 3.27 respectively.

The outer product form of the information matrix $\mathcal{R}$ is

$$\mathcal{R} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X} & \frac{1}{2}(\boldsymbol{\Sigma}^{-1/2} \otimes X')\boldsymbol{\Gamma}_1 D_m^{+\prime}\boldsymbol{\Delta} \\ \frac{1}{2}\boldsymbol{\Delta} D_m^{+}\boldsymbol{\Gamma}_1'(\boldsymbol{\Sigma}^{-1/2} \otimes X) & \frac{1}{4}\boldsymbol{\Delta} D_m^{+}\boldsymbol{\Gamma}_2^{*} D_m^{+\prime}\boldsymbol{\Delta} \end{pmatrix}. \tag{3.47}$$

In equation 3.47, $\boldsymbol{\Delta} = D_m' \left(\boldsymbol{\Sigma}^{-1/2} \otimes \boldsymbol{\Sigma}^{-1/2}\right) D_m$ and $\boldsymbol{\Gamma}_2^{*} = \boldsymbol{\Gamma}_2 - n^2 (vecI_m)(vecI_m)'$.

When the model is correctly specified, Γ_1 is reduced to 0 and Γ_2^{*} is reduced to $2nN_m$, consequently, $\mathcal{R} = \mathcal{F}$.

When the model is misspecified, the variance of the quasi-maximum likelihood estimator of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$, can be consistently approximated by $\mathcal{V} = \mathcal{F}^{-1}\mathcal{R}\mathcal{F}^{-1}$ (Gouriéroux, 1995a,b; Hendry, 1995; White, 1996)(see (Howe and Bozdogan, 2007; Magnus, 2007)), which is given by

$$\mathcal{V} = \begin{pmatrix} \boldsymbol{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1} & \frac{1}{n}\left(\boldsymbol{\Sigma}^{1/2} \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\boldsymbol{\Gamma}_1 D_p \boldsymbol{\Delta}^{-1} \\ \frac{1}{n}\boldsymbol{\Delta}^{-1} D_p'\boldsymbol{\Gamma}_1'\left(\boldsymbol{\Sigma}^{1/2} \otimes \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) & \frac{1}{n^2}\boldsymbol{\Delta}^{-1} D_p'\boldsymbol{\Gamma}_2^{*} D_p \boldsymbol{\Delta}^{-1} \end{pmatrix}. \tag{3.48}$$

The ICOMP(IFIM) for the misspecified model, namely ICOMP(IFIM)misspec is given by

$$\text{ICOMP(IFIM)misspec} = nm\log(2\pi) + n\log|\widehat{\Sigma}| + nm + 2C_1\left(\widehat{\mathcal{V}}\right), \tag{3.49}$$

where

$$C_1(\widehat{\mathcal{V}}) = \frac{s}{2}\log\left(\frac{tr\widehat{\mathcal{V}}}{s}\right) - \frac{1}{2}\log|\widehat{\mathcal{V}}|, \tag{3.50}$$

and where $s = rank(\widehat{\mathcal{V}})$.

In equation 6.28,

$$tr\widehat{\mathcal{V}} = tr(\widehat{\boldsymbol{\Sigma}})tr\left((\mathbf{X}'\mathbf{X})^{-1}\right) + \frac{1}{n^2}tr(D_m^{+})\left(\widehat{\boldsymbol{\Sigma}}^{1/2} \otimes \widehat{\boldsymbol{\Sigma}}^{1/2}\right)\widehat{\boldsymbol{\Gamma}}_2^{*}\left(\widehat{\boldsymbol{\Sigma}}^{1/2} \otimes \widehat{\boldsymbol{\Sigma}}^{1/2}\right) D_m^{+\prime}, \tag{3.51}$$

and

$$\left|\widehat{\mathcal{V}}\right| = 2^{-m(m-1)} n^{-m(m+1)} \left|\widehat{\boldsymbol{\Sigma}}\right|^{m+p+1} \left|\mathbf{X}'\mathbf{X}\right|^{-m} \left|D_m'\left(\widehat{\boldsymbol{\Gamma}}_2^{*} - \widehat{\boldsymbol{\Gamma}}_1'(I_m \otimes \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\widehat{\boldsymbol{\Gamma}}_1\right) D_m\right|. \tag{3.52}$$

Chapter 4

Genetic Algorithms

4.1 Introduction to Genetic Algorithms

A genetic algorithm is an adaptive stochastic optimization algorithm that mimics the procedure of biological evolution and natural selection. Genetic algorithms were widely used as optimization methods since they were first introduced by John Holland (1975) in the mid-1970s.

Traditionally, a genetic algorithm encodes information on a binary string (called "chromosome"). For a given problem, one chromosome represents a solution. The implementation of a genetic algorithm starts with a randomly generated population of chromosomes and evolves in generations. In each generation, every chromosome in the population is evaluated by a certain fitness function. Multiple chromosomes are selected from the current population to form a new population by genetic operators such as recombination (or crossover) and mutation. The better fitness the chromosome has, the more chance it has to be selected. The procedure is iterated until the condition of the algorithm is satisfied or a maximum number of generations has been produced.

The GA for model subset selection in regression used in this dissertation follows closely to the work of Bozdogan (2004a), Bozdogan and Bearse (2003), which are in turn based on Goldberg's simple genetic algorithm (SGA) (Goldberg, 1989).

4.2 GA on Model Selection

Before starting the genetic algorithm, we have to define the coding scheme for the possible subset regression models. Each regression model is encoded as a binary string, in which 1 indicates presence and 0 indicates absence of a certain predictor variable. All stings have the same length but contain different combinations of predictor variables. For example, in

a regression model of 6 predictors augmented by a constant term, the binary string 1100110 represents the model including the constant term and predictors x_1, x_4 and x_5, excluding the predictors x_2, x_3 and x_6.

Here are the steps we follow to implement genetic algorithm in subsetting regression models:

Step 1. Generate an initial population of regression models

The initial population is generated by selecting N subset regression models from the model space randomly. Population size N is an important parameter in GA, which represents the number of models to begin with in the algorithm. The population size is problem dependent. There is no general rule for how large the population size should be in the literature. Our GA is flexible to allow one to choose any population size.

Step 2. Evaluate each model in the population

Generally speaking, we may use any model selection criteria described in Chapter 3 as the fitness function to evaluate the models in GA. Particularly in this dissertation, we use the robust version of Bozdogan's ICOMP(IFIM), namely, RICOMP(IFIM) and RICOMP(IFIM)$_{\text{misspec}}$ for both the correctly specified model and misspecified model. To avoid confusion, we will use RICOMP and RICOMP$_{\text{misspec}}$ in short. AIC and AIC-type information criteria can also be used as fitness function in GA for model evaluation.

Step 3. Create a new population

A new population containing the offsprings of the parents from the previous population is created by following the substeps below.

[**selection**] In this dissertation, the parent models are selected by a ratio of the fitness value, RICOMP (or RICOMP$_{\text{misspec}}$), which is similar to Bozdogan (2004a). We illustrate how the models are selected by RICOMP here. The model selected by RICOMPmisspec will follow the same procedure.

Firstly, the RICOMP based on a certain robust estimate for each of the subset models in the population is calculated. We then get the difference RICOMP value by substracting the RICOMP for each model from the maximum RICOMP value in the population. That is

$$\Delta RICOMP_i = RICOMP_{max} - RICOMP_i \tag{4.1}$$

for $i = 1, 2, \ldots, N$, where N is the population size.

Secondly, we compute the average of the difference, which is given by

$$\overline{\Delta RICOMP} = \frac{1}{N}\sum_{i=1}^{N} \Delta RICOMP_i. \tag{4.2}$$

Finally, the ratio of each model's difference value to the average difference value is calculated by

$$R_i = \Delta RICOMP_i / \overline{\Delta RICOMP}. \tag{4.3}$$

The ratio R_i will determine which models will be used to reproduce the models in the next population. The chance of a model being selected is proportional to this ratio. This means a model with $R_i = 2$ will double its chance to be selected to reproduce its offsprings compared to a model with $R_j = 1$. This is called the proportional selection (Bozdogan, 2004a). There is a ranking selection method of ICOMP used by Bozdogan and Bearse (2003).

[reproduction] New offsprings are reproduced from the selected parents by crossover and mutation process.

- (crossover) Selected parents are randomly paired. A crossover (or recombination) process is then used to mate the paired parents and reproduce new offsprings. There are three choices of crossover (or recombination) operations: single point crossover, two point crossover and uniform crossover. A crossover probability between 0 and 1 is used to control the crossover rate, which is chosen by the researcher. The examples of these three crossover methods are given as follows, where | refers to the crossover point.
  - **Single point crossover** - There is one crossover point in each parent. The binary string of the offspring is copied from one parent before crossover point and copied from the other parent after the crossover point.
    Parent A 1000|110010 Offspring A 1000010001
    →
    Parent B 1100|010001 Offspring B 1100110010
  - **Two point crossover** - There are two crossover points in each parent. The binary string of the offspring is copied from one parent before the first crossover point, copied from the other parent between the first and second crossover point and copied the rest from the first parent.
    Parent A 1000|1100|10 Offspring A 1000010010
    →
    Parent B 1100|0100|01 Offspring B 1100110001
  - **Uniform Crossover** - The binary string of the offspring is copied randomly from either parent.
    Parent A 1000110010 Offspring A 1000010001
    ⇕ →

Parent B 1100010001 Offspring B 1100110010

- (mutation) Mutation is used in GA to allow the searching process to jump to another area instead of being constrained in a limit area so that the algorithm is not restricted to a local optimum. The mutation rate is specified by a mutation probability between 0 and 1. Mutation should be used sparingly since it is a random search operator. Otherwise the algorithm with a high mutation rate will become little more than a random search (Lin and Lee, 1996). The mutation probability is selected by the researcher.
- (elitism) The elitism rule is used in our GA. By using elitism rule, at least one best solution is copied without any change from current population to the next. In that way, the best solution can survive to the end of the algorithm.

[new population] A new population with new offsprings is now formed and replaces the old one. This population will be used for further runs of the algorithm in the iteration.

Step 4. Terminate the algorithm and return the best solutions in current population if final condition is satisfied, otherwise go to Step 2

If the solutions in the population satisfy the final condition of the model selection criteria, stop the algorithm and return the best solutions in current population. Otherwise go to Step 2 for a new iteration.

There are certain advantages and disadvantages related to genetic algorithms.

4.3 The Advantages and Disadvantages of GA

4.3.1 The Advantages of GA

GA can solve problems with enormous numbers of possible solutions within a reasonable time. This is particularly valuable in the model subset selection with a large number of predictor variables.

The optimization procedure of GA does not require the search on the gradient of the objective function and thus is not likely to be restricted to a local optima (Goldberg, 1989).

The combination of information criteria as a fitness function with GA is more likely to select "better" models than the classical stepwise selection (Bozdogan, 2004a). The combination of robustness, information criteria and GA enables us to inexpensively select the optimal or near optimal models in a robust and misspecification resistant framework.

It is flexible for the researcher to select GA parameters, such as the population size, the number of generations, crossover method and probability and mutation probability.

4.3.2 The Disadvantages of GA

There is some uncertainty on how to select the parameters in GA. No guidance in the literature is available to direct the selection of the parameters of GA. Understanding the relationships of these factors requires further investigation (Mahfoud, 1994). In this dissertation, we practice some combinations of the parameter choices, such as the population size and the number of generations depending on the problem.

As any other non-exhaustive search methods, GA may not return the overall optima but only the "good" ones in some situations.

4.4 GA Hybridized with Robust and Misspecification Resistant Information Criteria

To calculation the robust and misspecification resistant model subset selection in GA in this dissertation, we implemented an easy to use graphical user interface (GUI) software in Matlab. The GUI used here is developed based on the work of Bozdogan (2004a) under the correctly specified model. In our GUI, we develop and apply the misspecification resistant version of ICOMP. We provide the flexibility to choose from the fitness values of RICOMP(IFIM) or RICOMP(IFIM)$_{\text{misspec}}$ under both correctly specified model or misspecified model assumptions. Furthermore, the parameter estimation method can be chosen from the ordinary least squares (OLS) method and four robust estimation methods: Huber's method, Andrews' method, Tukey's method and Hampel's method, which enable us to select the model on a robust basis. Therefore, the researcher can apply our GUI for robust and misspecification resistant model selection using GA.

The GUI inputs and outputs are illustrated in Tables 4.1 and 4.2.

Table 4.1: GUI Inputs and Descriptions for GA

Inputs	Descriptions
No. of Runs	
No. of Generations	
Population Size	
Estimation Method	to choose one from the five methods: OLS, Huber's, Andrews', Tukey's, Hampel's
Fitness Value	to choose one from the two fitness values: RICOMP(IFIM), RICOMPmisspec
Probability of Crossover	a real number value from 0 to 1
Crossover Method	to choose one from the three methods: single point, two-point, uniform
Probability of Mutation	a real number from 0 to 1
Elitism	check or uncheck the box "Yes"
Input Data Files	Y: input the dependent variable Y X: input the set of independent variables Xs
Go	to start the algorithm
Reset	to reset the parameters
Exit	to exit the algorithm

Table 4.2: GUI Outputs and Descriptions for GA

Outputs	Descriptions
View 2D/3D plot	to show the 2D/3D plot of criterion values versus number of generations
Save Figure	to pop up the 2D/3D plot and save it
Fitness Chromosome	to return the predictor variables in the best model
Fitness Binary String	to return the binary string representing the best model
Fitness Score	to return the best fitness value
Outputs in Matlab	the table of generation results for GA; the success rate to pick up the best model; the fitness chromosome, fitness binary string and fitness score value
Output file	to show the same results as the outputs in Matlab command window

Chapter 5

Robust and Misspecification Resistant Model Selection in Multiple Linear Regression

5.1 Robust Estimates in Multiple Linear Regression

Consider the multiple linear regression model given by

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \end{aligned} \tag{5.1}$$

for the i^{th} observation, where $\mathbf{x}' = (1, x_{i1}, \ldots, x_{ik})$ is the i^{th} row of an $n \times p$ $(p = k + 1)$ design matrix $\mathbf{X}$.

In a compact matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5.2}$$

where $\mathbf{y}$ is a vector of $(n \times 1)$ observations on a dependent variable; $\mathbf{X}$ is a matrix of $(n \times p)$ nonstochastic predetermined predictor variables $(p = k+1)$; $\boldsymbol{\beta}$ is a vector of $(p \times 1)$ coefficients and $\boldsymbol{\varepsilon}$ is a vector of $(n \times 1)$ random errors, which follows $\varepsilon_i \sim N(0, \sigma^2)$, for $i = 1, 2, \ldots, n$, or equivalently, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$.

The least squares (LS) estimator of $\boldsymbol{\beta}$ is to minimize the sum of the squared residuals

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i^2 = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2. \tag{5.3}$$

That is,

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} r_i^2, \tag{5.4}$$

where $r_i = y_i - \hat{y}_i$ is the residual for the i^{th} observation, and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$ is the i^{th} predicted value. In matrix form, the LS estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{5.5}$$

Least squares estimator is very unstable under slight changes of the underlying distribution. When assuming the normal distribution of the underlying model, the method of least squares produces a good estimator of the regression coefficients with good properties. However, in many real world applications, the assumption of normality is not appropriate. Huber (1972) criticizes this "dogma" of assuming measurement error to be normally distributed and suggests that "a more rational action would be to check whether they were compatible with a normal distribution and if not, to develop a different theory of estimation."

Least squares estimator is also very sensitive to the unusual observations (or "outliers") in the data, which either comes from the longer or heavier tailed distribution than normal or simply they are erroneous data points.

The general M-estimation thus is introduced to the robust regression literature as an alternative method for the LS estimation. M-estimation replaces the objective function of minimizing sum of squared residuals used in OLS estimation by another function of the residuals. The function is a less rapidly increasing loss function than that of the OLS estimation, and it assigns less importance to outliers. The M-estimator minimizes the objective function

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i'\boldsymbol{\beta}), \tag{5.6}$$

where $\rho\colon \mathbb{R}^1 \to \mathbb{R}^1$ is a symmetric, positive-definite function with a unique minimum at zero (Rousseeuw and Yohai, 1984).

That is, the M-estimator of regression coefficient $\hat{\boldsymbol{\beta}}_M$ is given by

$$\hat{\boldsymbol{\beta}}_M = \arg\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} \rho(r_i). \tag{5.7}$$

The M-estimator is not necessarily scale invariant. To obtain a scale invariant version of M-estimator, we need to solve

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i}{s}\right) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right), \tag{5.8}$$

where s is a robust estimate of scale ($s = \hat{\sigma}$). In this dissertation, we choose the median absolute deviation (MAD) as our robust scale estimator, which is given by

$$s = 1.4826 \times \text{median}|r_i - \text{median}(r_i)|. \tag{5.9}$$

The tuning constant 1.4826 makes s an approximately unbiased estimator of σ if n is large and the error distribution is normal (Montgomery et al., 2001).

To minimize equation 5.8, let $\psi = \rho'$ be the first partial derivative of ρ. Differentiating equation 5.8 with respect to the coefficient $\beta_j (j = 0, 1, \ldots, k)$, and setting the partial derivatives equal to zero, we get a system of $k + 1$ equations

$$\sum_{i=1}^{n} x_{ij} \psi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right) = 0, \ j = 0, 1, \ldots, k, \tag{5.10}$$

where x_{ij} is the i^{th} observation on the j^{th} predictor and $x_{i0} = 1$.

Rewriting equation 5.10, we have

$$\sum_{i=1}^{n} x_{ij} \psi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right) = \sum_{i=1}^{n} x_{ij} w_i (y_i - \mathbf{x}_i'\boldsymbol{\beta}) = 0, \ j = 0, 1, \ldots, k, \tag{5.11}$$

where

$$w_i = \begin{cases} \psi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right) \Big/ \left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right), & \text{if } y_i \neq \mathbf{x}_i'\hat{\boldsymbol{\beta}}; \\ 1, & \text{if } y_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}. \end{cases} \tag{5.12}$$

Equation 5.11 can be written in matrix form as

$$\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{y}, \tag{5.13}$$

where $\mathbf{W}$ is an $n \times n$ diagonal matrix, whose diagonal elements $(w_1, w_2, \ldots, w_n)$ are the weights given by equation 5.12. Equation 5.13 is referred to as the weighted least-squares (WLS) normal equation. The one-step estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \tag{5.14}$$

In general, the ψ function is nonlinear and iterative methods could be used to solve equation 5.11. In this dissertation, the *iteratively reweighted least-squares*(IRLS) method is employed (Beaton and Tukey, 1974) for solving the ψ function.
The steps of IRLS are as follows:

1. Select initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$, such as the LS estimate.
2. At each iteration t, calculate the weight $w_i^{(t-1)}$ using the equation 5.12.
3. Solve for the new weighted least squares estimate

$$\hat{\boldsymbol{\beta}}^{(t)} = (\mathbf{X}'\mathbf{W}^{(\mathbf{t-1})}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(\mathbf{t-1})}\mathbf{y}$$

4. Repeat steps 2 and 3 until the estimated coefficient ($\hat{\boldsymbol{\beta}}$) converges.

Now, our question is how to determine the covariance matrix of $\hat{\boldsymbol{\beta}}$ since it is important to make model subset selections and other model inferences. Huber (1973) shows that asymptotically $\hat{\boldsymbol{\beta}}$ follows an approximately normal distribution with the covariance matrix

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \frac{E[\psi^2(\varepsilon/\sigma)]}{\{E[\psi'(\varepsilon/\sigma)]\}^2}(\mathbf{X}'\mathbf{X})^{-1}. \tag{5.15}$$

One good choice of estimating the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \frac{(ns^2)\sum_{i=1}^n \psi^2[(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})/s]}{(n-p)\{\sum_{i=1}^n \psi'[(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})/s]\}^2}(\mathbf{X}'\mathbf{X})^{-1}. \tag{5.16}$$

The weighted least-squares (WLS) program automatically produces the estimate of the covariance matrix as

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \frac{\sum_{i=1}^n w_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2}{(n-p)}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}. \tag{5.17}$$

Welsch (1975) suggests the combination of the equation 5.16 and 5.17, which gives

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \frac{(ns^2)\sum_{i=1}^n \psi^2[(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})/s]}{(n-p)\{\sum_{i=1}^n \psi'[(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})/s]\}^2}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}. \tag{5.18}$$

We will use the covariance matrix estimation suggested by Welsch in this dissertation.

5.2 Information Criteria on Robust Regression Model Selection

Ronchetti (1985) proposed a robust regression model selection procedure, known as AICR, which generalized Akaike Information Criterion (AIC) to the robust linear regression. The

AICR is defined as

$$\text{AICR}(p;\alpha,\rho) = 2\sum_{i=1}^{n}\rho_c\left\{\frac{w(x_i)}{\widehat{\sigma}}(y_i - \mathbf{x_i}'\widehat{\boldsymbol{\beta}})\right\} + \alpha p, \tag{5.19}$$

where p is the number of parameters in the model; ρ_c is Huber's ρ function defined in equation 2.9; $\widehat{\boldsymbol{\beta}}$ is the corresponding M-estimator and $\widehat{\sigma}$ is some robust estimate of σ; α is chosen to be

$$\alpha = \alpha_c = 2\frac{E_\Phi\psi_c^2}{E_\Phi\psi_c'},$$

where ψ_c is Huber's ψ function defined in equation 2.10 and Φ is the standard normal distribution function.

Note that $\alpha_c < 2$ and $\alpha_\infty = 2$ so that $AICR(p;\alpha_\infty,\rho_\infty) = AIC(p;2)$, which is the classical AIC statistic under normality.

Hampel (1983) proposed his robust version of AIC, named HAIC, by obtaining another choice for α from Ronchatti's. He chose α to be

$$\alpha = \frac{E_\Phi\psi_c^2}{E_\Phi\psi_c'} + \frac{E_\Phi\psi_c^2}{(E_\Phi\psi_c')^2}.$$

Hampel's HAIC is defined by

$$HAIC = 2\sum_{i=1}^{n}\rho_c\left\{\frac{w(x_i)}{\widehat{\sigma}}(y_i - \mathbf{x_i}'\widehat{\boldsymbol{\beta}})\right\} + \left\{\frac{E_\Phi\psi_c^2}{E_\Phi\psi_c'} + \frac{E_\Phi\psi_c^2}{(E_\Phi\psi_c')^2}\right\}p. \tag{5.20}$$

Machado (1993) derives a robust version of Bayesian Information Criterion (BIC or SBC), based on the objective functions defining M-estimators for parametric models. Its specific Schwarz Bayesian criterion (SBC) based on Huber's M-estimator for robust regression model, RBIC, is given by

$$RBIC = 2\sum_{i=1}^{n}\rho_c\left\{\frac{w(x_i)}{\widehat{\sigma}}(y_i - \mathbf{x_i}'\widehat{\boldsymbol{\beta}})\right\} + p\log n. \tag{5.21}$$

We notationally use RBIC instead of RSBC to be compatible with the literature, although Schwarz Bayesian criterion (SBC) is not grounded on information theoretic developments such as AIC or ICOMP.

In all of Ronchetti's AICR, Hampel's HAIC and Machado's RBIC, they use $w(x) = 1$. In this dissertation, when we derive the robust version of ICOMP(IFIM), namely RICOMP(IFIM), we generalize $w(x) \in [0,1]$ to allow the criterion better take into account the influence of outliers.

The robust version of cross-validation method, RCV, can be defined as (Qian and Künsch, 1996)

$$RCV = 2\sum_{i=1}^{n} \rho_c \left\{ \frac{w(x_i)}{\widehat{\sigma}} (y_i - \mathbf{x_i}'\widehat{\boldsymbol{\beta}}^{(i)}) \right\}, \tag{5.22}$$

where $\widehat{\boldsymbol{\beta}}^{(i)}$ represent the M-estimator based on all the observations in the sample except point (x_i, y_i).

Qian and Künsch (1996) in their paper proved that based on Hampel's (1974) heuristic influence function, RCV is expected to behave similiarly as the AICR if α is a correct choice and the M-estimator has a bounded influence. Ronchetti et al. (1997) also studied the cross-validation method in their paper.

Qian and Künsch (1996; 1998) derived the Stochastic Complexity, SC, using the stochastic complexity theory of Rissanen (1989; 1996), as a robust model selection criterion in linear regression, which is defined as

$$\begin{aligned} SC\left(\mathbf{Y}_n \mid \mathbf{x}_n\right) &= \sum_{i=1}^{n} \rho_c \left\{ \frac{w(x_i)}{\widehat{\sigma}} (y_i - \mathbf{x_i}'\widehat{\boldsymbol{\beta}}) \right\} + \frac{p}{2}\log E_{f_0}\psi_c' + \frac{1}{2}\log\left|\mathbf{X}_n'\mathbf{W}_n^2\mathbf{X}_n\right| \\ &\quad + \log\prod_{j=1}^{p} \frac{|\hat{\beta}_j| + n^{-1/4}}{\sigma}. \end{aligned} \tag{5.23}$$

The first term in equation 5.23 is the robust fitting error which shows the goodness of fit. The other three terms in the equation represent the model complexity.

5.3 Robust ICOMP for Regression Model

In this dissertation, we define a robust version of Bozdogan's information-theoretic measure of complexity as (Bozdogan, 2003; Liu, 2004)

$$\text{RICOMP(IFIM)} = -2\log L(\hat{\boldsymbol{\theta}}_R) + 2\text{C}_1(\widehat{\mathcal{F}}_R^{-1}), \tag{5.24}$$

where $\hat{\boldsymbol{\theta}}_R = (\hat{\boldsymbol{\beta}}_R, \hat{\sigma}_R^2)$ represents the robust estimate of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$, which can be obtained from any of the M-estimation methods we discussed in Section 2.1. $L(\hat{\boldsymbol{\theta}}_R)$ represents the maximized likelihood function on the robust estimate of the parameter vector $\hat{\boldsymbol{\theta}}_R$. $\widehat{\mathcal{F}}_R^{-1}$ represents the robust estimate of the inverse Fisher information matrix.

The result in equation 5.24 can be easily generalized to the regression model. Inspired by QianQian and Künsch (1996; 1998), we employ Huber's Least Favorable Distribution (Huber, 1964) to estimate the density function of the random errors r_i in the regression model

and then build the robust lack of fit part (negative log likelihood function of the robust estimates) in RICOMP(IFIM).

Following Qian and Künsch (1996), the least favorable distribution of r_i given x_i is

$$f(r_i \mid x_i) = \frac{w(x_i)}{\sigma} f_0\left(\frac{w(x_i)r_i}{\sigma}\right) = (1-\lambda)\left(\sqrt{2\pi}\sigma\right)^{-1} w(x_i) \exp\left\{-\rho\left(\frac{w(x_i)r_i}{\sigma}\right)\right\}, \quad (5.25)$$

where $f_0(r) = (1-\lambda)\left(\sqrt{2\pi}\right)^{-1} \exp\{-\rho(r)\}$, $\lambda (0 < \lambda < 1)$ is a constant.

Note that in their paper, they use $\rho(\cdot)$ in equation 5.25 as Huber function. Here, we generalize ρ to all the robust functions stated in Section 2.1.

Based on the above least favorable density function of r_i, the likelihood function for the parameter $(\boldsymbol{\beta}, \sigma)$ is given by

$$L(\boldsymbol{\beta}, \sigma \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} f(r_i \mid x_i). \quad (5.26)$$

The log likelihood function of $\boldsymbol{\theta}$ is

$$l(\boldsymbol{\beta}, \sigma \mid \mathbf{y}, \mathbf{X}) = n\log(1-\lambda) - \frac{n}{2}\log 2\pi - n\log\sigma + \sum_{i=1}^{n}\log w(x_i) - \sum_{i=1}^{n}\rho\left\{\frac{w(x_i)}{\widehat{\sigma}}(y_i - \mathbf{x_i}'\widehat{\boldsymbol{\beta}})\right\}. \quad (5.27)$$

Since the first term $n\log(1-\lambda)$ is a constant, it can be dropped when we construct the RICOMP(IFIM) for the regression model.

The robust version of Bozdogan's ICOMP(IFIM), RICOMP(IFIM) in equation 5.24 can be written in the following form for the regression model

$$\begin{aligned} \text{RICOMP(IFIM)}_{reg} &= -2\log L(\hat{\boldsymbol{\theta}}_R) + 2\text{C}_1(\widehat{\mathcal{F}}_R^{-1}) \quad (5.28)\\ &= n\log 2\pi + 2n\log\sigma - 2\sum_{i=1}^{n}\log w(x_i) \\ &\quad +2\sum_{i=1}^{n}\rho\left\{\frac{w(x_i)}{\widehat{\sigma}}(y_i - \mathbf{x_i}'\widehat{\boldsymbol{\beta}})\right\} + 2\text{C}_1(\widehat{\mathcal{F}}_R^{-1}), \end{aligned}$$

where

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}_R, \widehat{\sigma}_R^2) = \widehat{\mathcal{F}}_R^{-1} = \begin{bmatrix} \widehat{Cov}(\hat{\boldsymbol{\beta}}_R)_R & \mathbf{0} \\ \mathbf{0} & \frac{2\widehat{\sigma}_R^4}{n} \end{bmatrix}, \quad (5.29)$$

and where

$$\text{C}_1(\widehat{\mathcal{F}}_R^{-1}) = \frac{s}{2}\log\left[\frac{tr(\widehat{\mathcal{F}}_R^{-1})}{s}\right] - \frac{1}{2}\log|\widehat{\mathcal{F}}_R^{-1}|, \quad (5.30)$$

and $\hat{\sigma}_R$ is estimated by Median Absolute Deviation (MAD) of residuals.

Note that by asymptotic theory, we have (Bozdogan, 2003)

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \sim N(\mathbf{0}, Cov(\widehat{\boldsymbol{\beta}}_R)), \tag{5.31}$$

where $\widehat{\boldsymbol{\beta}}_R$ is the robust estimate of regression coefficient $\boldsymbol{\beta}$ and $Cov(\widehat{\boldsymbol{\beta}}_R)$ is the robust covariance matrix.

The estimated covariance matrix $\widehat{Cov}(\widehat{\boldsymbol{\beta}}_R)$ is given by

$$\widehat{Cov}(\widehat{\boldsymbol{\beta}}_R) = \frac{(n\widehat{\sigma}^2)\sum_{i=1}^{n}\psi^2[(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})/\widehat{\sigma}]}{(n-p)\{\sum_{i=1}^{n}\psi'[(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})/\widehat{\sigma}]\}^2}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}. \tag{5.32}$$

RICOMP(IFIM) in equation 5.24 as well as that in equation 5.28 provides us with a unified information criterion which combines the robust estimators into Bozdogan's (1988a; 1990) information-theoretic measure of complexity (ICOMP) in model subset selection literature. Besides the nice property of ICOMP(IFIM) over the AIC-type information criterion, this new criterion is also robust to the departure of normality assumption of the model and unusual observations.

The model with minimum RICOMP(IFIM) value will be chosen as the best among the competing candidate models. For conciseness and to avoid confusion, we will use RICOMP to represent RICOMP(IFIM) in the following sections.

5.4 Robust and Misspecification Resistant ICOMP for Regression Model

The robust version for Bozdogan's misspecification-resistant ICOMP(IFIM), denoted by RICOMP(IFIM)$_{\text{misspec}}$, is given by

$$\begin{aligned}\text{RICOMP(IFIM)}_{\text{misspec}} &= n\log 2\pi + 2n\log\sigma - 2\sum_{i=1}^{n}\log w(x_i) \\ &\quad +2\sum_{i=1}^{n}\rho\left\{\frac{w(x_i)}{\widehat{\sigma}}(y_i - \mathbf{x_i}'\widehat{\boldsymbol{\beta}})\right\} + 2\mathrm{C}_1\left(\widehat{Cov}(\widehat{\boldsymbol{\theta}})_{R\_misspec}\right),\end{aligned} \tag{5.33}$$

where

$$\widehat{Cov}(\widehat{\boldsymbol{\theta}})_{R\_misspec} = \widehat{\mathcal{F}}_R^{-1}\widehat{\mathcal{R}}_R\widehat{\mathcal{F}}_R^{-1}, \tag{5.34}$$

and where

$$\mathrm{C}_1\left(\widehat{Cov}(\widehat{\boldsymbol{\theta}})_{R\_misspec}\right) = \frac{s}{2}\log\left(\frac{tr\left(\widehat{Cov}(\widehat{\boldsymbol{\theta}})_{R\_misspec}\right)}{s}\right) - \frac{1}{2}\log\left|\widehat{Cov}(\widehat{\boldsymbol{\theta}})_{R\_misspec}\right|. \quad (5.35)$$

In equation 5.34, $\widehat{\mathcal{F}}_R^{-1}$ and $\widehat{\mathcal{R}}_R$ are the robust versions of the estimates of inner and outer-product forms of the Fisher information matrix. $\widehat{\mathcal{F}}_R^{-1}$ is computed using equation 5.29 and $\widehat{\mathcal{R}}_R$ is computed as

$$\widehat{\mathcal{R}}_{(R)} = \begin{bmatrix} \frac{1}{\widehat{\sigma}_R^4}\mathbf{X'D^2X} & \mathbf{X'1}\frac{Sk}{2\widehat{\sigma}_R^3} \\ \left(\mathbf{X'1}\frac{Sk}{2\widehat{\sigma}_R^3}\right)' & \frac{(n-p)(Kt-1)}{4\widehat{\sigma}_R^4} \end{bmatrix}, \quad (5.36)$$

where $\mathbf{D}^2 = diag(\widehat{\varepsilon}_1^2, \widehat{\varepsilon}_2^2, \ldots, \widehat{\varepsilon}_n^2)$ and $\mathbf{X}$ is an $(n \times p)$ matrix of predictor variables. $\mathbf{1}$ is an $(n \times 1)$ vector of ones. Sk and Kt are the coefficients of skewness and Kurtosis respectively, which are given by $Sk = \frac{\left(\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^3\right)}{\widehat{\sigma}^3}$ and $Kt = \frac{\left(\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^4\right)}{\widehat{\sigma}^4}$. $\hat{\sigma}_R$ is the robust estimate of σ.

Another form of RICOMP(IFIM)$_{misspec}$ is defined as

$$\begin{aligned} \text{RICOMP(IFIM)}_{\text{misspec}} &= n\log 2\pi + 2n\log\sigma - 2\sum_{i=1}^n \log w(x_i) \quad (5.37) \\ &\quad + 2\sum_{i=1}^n \rho\left\{\frac{w(x_i)}{\widehat{\sigma}}(y_i - \mathbf{x_i}'\widehat{\boldsymbol{\beta}})\right\} + 2\left[tr(\widehat{\mathcal{F}}_R^{-1}\widehat{R}_{(R)}) + \mathrm{C}_1(\widehat{\mathcal{F}}_R)\right] \end{aligned}$$

which has a different complexity term from that in equation 5.34.

5.5 Robust Model Selection Algorithm

For all possible subset selection of p-dimensional data, we have $2^p - 1$ models to evaluate. Here are the steps to choose the best model among the competing candidate models:

1. For a certain subset model, select one of the M-estimation methods described in Section 2.1.

2. Use iteratively reweighted least-squares (IRLS) algorithm to compute the robust estimates of the model parameters.

3. Use Welsch's method in 5.18 to obtain the robust estimate of the covariance matrix of the model.

4. Compute the RICOMP(IFIM) or the RICOMP(IFIM)$_{\text{misspec}}$ value given by equation 5.28 and 5.34, respectively, for the model depending on whether the model is misspecified.

5. Repeat steps 1-4 for all possible subset models. Sort the RICOMP(IFIM) or RICOMP(IFIM)$_{\text{misspec}}$ values from the smallest to the largest and choose the best subset model with the minimum RICOMP(IFIM) (or RICOMP(IFIM)$_{\text{misspec}}$ value).

The GA model subset selection should follow the steps described in Section 4.2, where the fitness function is RICOMP(IFIM) or RICOMP(IFIM)$_{\text{misspec}}$.

In the following section, we demonstrate both on a simulation protocol and real world examples to illustrate the effectiveness of this new information criteria RICOMP(IFIM) and RICOMP(IFIM)$_{\text{misspec}}$ and compare them with the robust model selection methods in the current literature, AICR, HAIC, and RBIC.

5.6 Numerical Examples

5.6.1 Simulation Example

In this simulation study, we use the following Monte Carlo protocol following Bozdogan and Haughton (1998) to generate data and calculate the parameter estimates and model subset selection.

Let z_1, z_2, z_3 and z_4 be independent random variables that follow standard normal distribution $N(0,1)$. The first three predictor variables are simulated by

$$x_1 = \sqrt{1-\alpha^2}z_1 + \alpha z_4, \tag{5.38}$$

$$x_2 = \sqrt{1-\alpha^2}z_2 + \alpha z_4, \tag{5.39}$$

$$x_3 = \sqrt{1-\alpha^2}z_3 + \alpha z_4. \tag{5.40}$$

It is obvious that the variance of $x_i (i = 1, 2, 3)$ is 1, and that the covariance of x_i and x_j $(i \neq j, i, j = 1, 2, 3)$ is α^2. By assigning different values for α, we can control the degree of collinearity among the predictor variables. For our simulations, we use $\alpha^2 = 0.5$.

Suppose λ_{max} is the largest eigenvalue of the covariance matrix of x_1, x_2, x_3 with β_{max} as the eigenvector corresponding to λ_{max}. We can show that $E[(x\beta_{max})_i^2] = \lambda_{max}$ (Johnson and Wichern, 1992, page 358). The eigenvector β_{max} yields a high variability for $x\beta$. The response variable y is generated from the first three predictor variables,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_{max} + \boldsymbol{\varepsilon}, \tag{5.41}$$

where $\mathbf{X} = [x_1, x_2, x_3]$. In our simulations, we assign different distributions for ε to illustrate the model subset selection results under the model misspecification (i.e., the error distribution deviates from the normal distribution). Five outliers on the response variable y are arbitrarily introduced to this simulation data, $y(53) = -10; y(62) = 15; y(18) =$

$-12; y(96) = 10; y(40) = 18$. Note that the observations are picked at random to assign the outliers.

Then, we generate seven redundant variables, $x_4, x_5, \ldots, x_{10}$ using the uniform random numbers, which are given by

$$x_4 = 4 * rand(0,1), \cdots, x_{10} = 10 * rand(0,1), \tag{5.42}$$

where rand(0,1) generates the standard Uniform random numbers.

For convenience in showing our model selection results, we define the "full model" (M_f), "true model" (M_t), "other correct model" (M_{oc}), "overfitting model" (M_{of}), "redundant model" (M_r) and "wrong model" (M_w) as follows. The "full model" is the model containing all the predictor variables, which is denoted by $M_f = \{x_1, x_2, \ldots, x_{10}\}$. The "true model" is the one including the first three predictor variables, which is denoted by $M_t = \{x_1, x_2, x_3\}$. We define the "other correct model" M_{oc} as any non-empty strict subset of the true model, which is denoted by $M_{oc} \subset M_t$. We define the "overfitting model" M_{of} as the one containing both the true model and any redundant variable(s). In other words, the true model is a strict subset of the overfitting model, $M_t \subset M_{of}$. The "redundant model" M_r, is defined as the model including both the other correct model and any redundant variable(s). By this definition, $M_{oc} \subset M_r$. Finally, the "wrong model" M_w is defined as any model containing redundant variable(s) only.

Parameter Estimation on the True Model

In this part, we estimate the regression coefficients for the true model by using both the OLS method (with and without the outliers) and the four robust methods assuming the error in equation 5.41 is normally distributed with $N(0, 0.5)$. The OLS estimate with outliers deleted is supposed to work well since we only introduce the outliers, but the error distribution is normal. In this way, we can compare the estimation results of the robust methods with the OLS estimates. The QQ-plot and histogram of the response variable for one simulation are given in Figure 5.1, where both plots show the obvious outliers in the raw data.

The estimation results are summarized in Table 5.1, which also include the tuning constants used for each robust function. We can see from the table that OLS estimates are far away from the true values indicating the large biases. The OLS_2 estimates (OLS without the outliers) are very close to the true values with small biases. The four robust estimates work well too, which are close to the true values and the OLS_2 estimates. The four robust methods dramastically reduce the standard deviation of the model compared with the OLS method. Among them, Huber's and Andrews' estimate give the smallest standard deviation, which may be a sign that Huber or Andrews robust estimations combined with

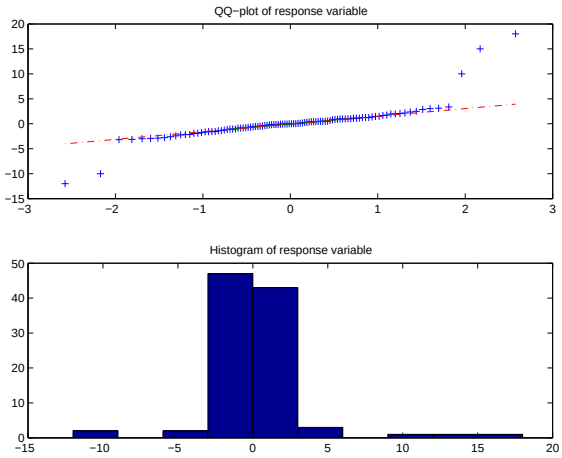

Figure 5.1: Simulated Data: QQ-plot and Histogram of the Response for One Simulation.

information criteria will perform the best in the model subset selection. The RICOMP and $\text{RICOMP}_{\text{misspec}}$ values based on robust estimators are reduced to about one third of those values based on OLS method.

The plots of standard residuals for the OLS estimate and four robust estimates are shown in Figure 5.2. We can see that the OLS method can unambiguously pick up three of the outliers, but mask two outliers. The four robust methods can successfully pick up all the five outliers in the data.

The QQ-plot of the residuals is given in Figure 5.3, which shows the consistency of the robust estimates with the OLS_2 estimate (OLS without the outliers).

All Possible Subset Selection

We simulated data set of size $n = 100$ observations with the random error term given in equation 5.41 from 10 different distributions. We repeated this experiment 100 times on each of the 10 different distributions. Then an all possible subset selection is carried out on each of these simulations using the information criteria AICR, HAIC, RBIC, RICOMP(IFIM) and $\text{RICOMP(IFIM)}_{\text{misspec}}$. Note that we have four forms for each of the RICOMP(IFIM) and $\text{RICOMP(IFIM)}_{\text{misspec}}$ respectively based on four different robust functions. We note that this is a very high level intensive simulation. We use the same tuning constants for each of the robust functions as we did in the previous estimation part. The true model $\{x_1, x_2, x_3\}$ is desired to be selected. We count how many times each information criterion hit the true model, other correct model, overfitting model and redundant model in 100 runs. The results are summarized in Tables 5.2 and 5.3.

Table 5.1: Simulated Data: Parameter Estimates

	TRUE	OLS	OLS_2[a]	Huber	Andrews	Tukey	Hampel
tuning constants				k=1.345	c=1.5	c=6.0	a=1.2 b=3.5 c=8.0
$\hat{\beta}_1$	0.6523	0.6308	0.6271	0.6155	0.6208	0.6225	0.6158
$\hat{\beta}_2$	0.5176	0.66	0.5886	0.5851	0.5586	0.5653	0.5664
$\hat{\beta}_3$	0.5537	0.6636	0.5473	0.5779	0.5706	0.5654	0.5732
bias		-0.0215	-0.0252	-0.0369	-0.0315	-0.0298	-0.0365
		0.1423	0.071	0.0674	0.041	0.0477	0.0487
		0.11	-0.0063	0.0243	0.0169	0.0118	0.0195
std deviation		2.9921	2.9964	**0.4217**	**0.4212**	0.4291	0.4241
RICOMP		503.0906		153.1132	83.1116	115.1337	106.6408
$RICOMP_{misspec}$		512.3352		165.8937	95.869	127.8569	119.1623

[a]OLS with outliers 18, 40, 53, 62 and 96 deleted

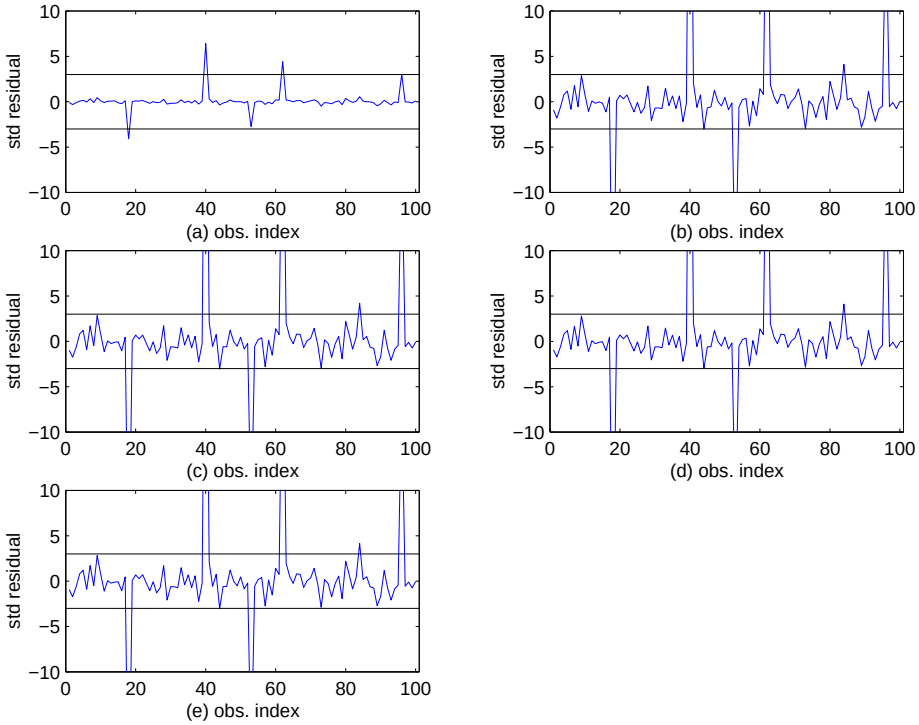

Figure 5.2: Simulated Data: Plots of the Standardized Residuals. (a) standardized residual from OLS estimation. (b) standardized residual from Huber's estimation. (c) standardized residual from Andrews' estimation. (d) standardized residual from Tukey's estimation. (e) standardized residual from Hampel's estimation.

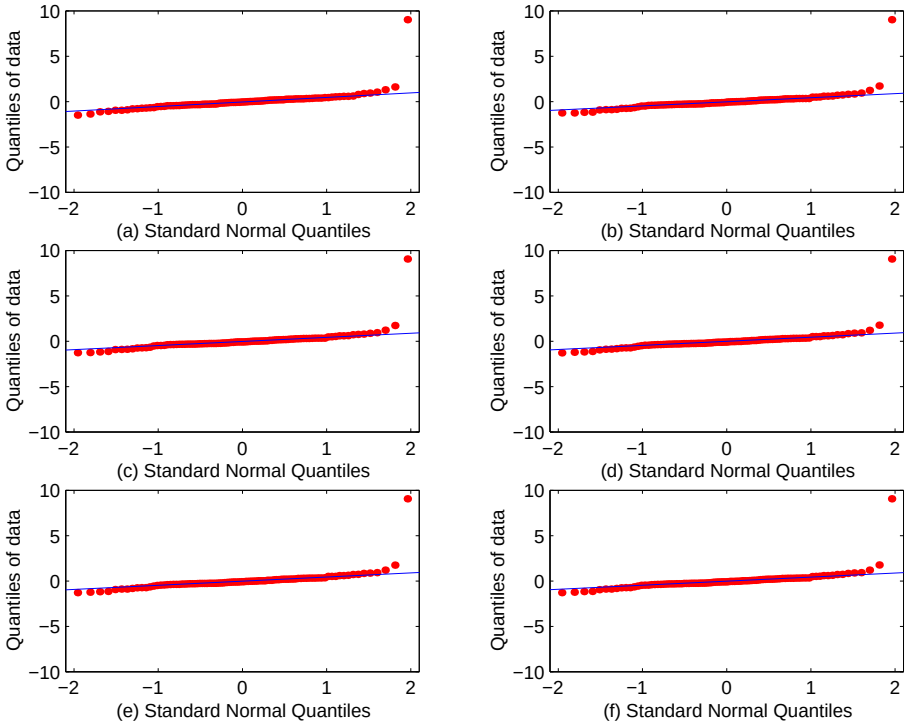

Figure 5.3: Simulated Data: QQ-plot of the Residuals. (a) QQplot of residuals from OLS estimation (with all observations). (b) QQplot of residuals from OLS estimation (without observations 18, 40, 53, 62 and 96). (c) QQplot of residuals from Huber's estimation. (d) QQplot of residuals from Andrews estimation. (e) QQplot of residuals from Tukey estimation. (f) QQplot of residuals from Hampel estimation.

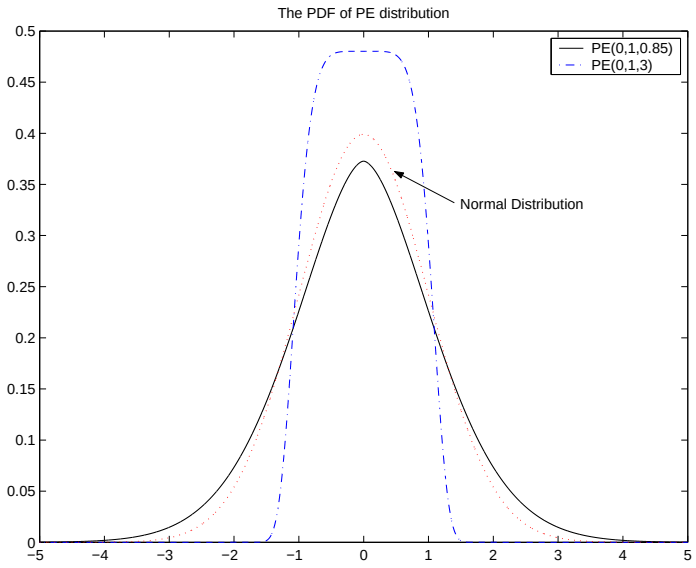

Figure 5.4: Probability Density Function of PE Distribution.

Ten different distributions for the error ε in equation 5.41 are chosen in this study to represent deviations from normality. They are normal distribution $N(0, 0.5)$, contaminated normal distributions: $0.5N(0,1)+0.5N(0,0.5)$ and $0.5N(0,1)+0.5N(0,2)$, normal distribution contaminated with t distribution: $0.5N(0,1) + 0.5t(3)$ and $0.5N(0,1) + 0.5t(5)$, power exponential distribution, $PE(0, 1, 0.85)$, $PE(0, 1, 3)$ (refer to Appendix A for the details), skewed power exponential distribution, $SPE_3(0, 1, 1), SPE_3(0, 1, 1.25), SPE_2(0, 1, 1)$ (refer to Appendix A for the details). The probability plots of power exponential distributions are shown in Figure 5.4 and those of skewed power exponential distributions are shown in Figure 5.5. Note that the SPE distributions shown in Figure 5.5 are not the ones we used in this simulation. Our purpose is to show the skewness feature of the SPE distributions. From both plots, we see that both PE and SPE distributions deviate considerably from the normal distribution.

Table 5.2 contains the all possible subset selection results for the ten models generated from different error distributions using RICOMP(IFIM), AICR, HAIC and RBIC as model selection criteria. Table 5.3 contains the all possible subset selection results for the ten models using $\text{RICOMP(IFIM)}_{\text{misspec}}$ as model selection criteria. All of the information criteria in both tables make model subset selection on the same 100 simulated data of size $n = 100$ so that the results for the information criteria are comparable to each other.

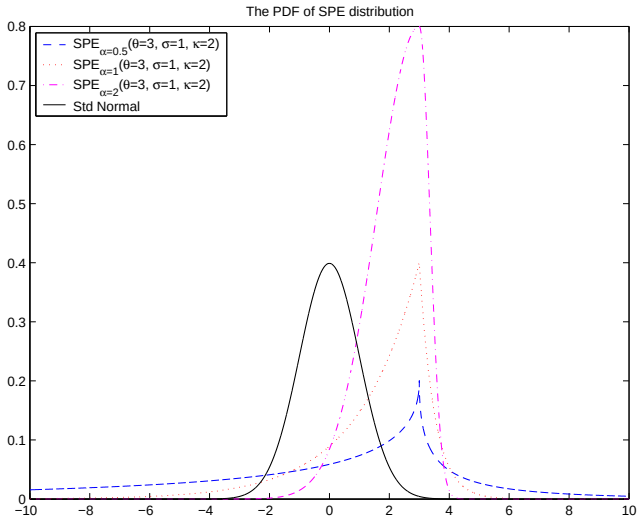

Figure 5.5: Probability Density Function of SPE Distribution.

Table 5.2: Simulated Data: All Possible Subset Model Selection in 100 runs(1)

Error Distribution	Model Category	RICOMP					AICR	HAIC	RBIC
		OLS	Huber	Andrews	Tukey	Hampel			
$\varepsilon \sim$ N(0,0.5)	True model	12	41	13	45	30	12	15	64
	Other Correct Model	54	0	0	0	0	0	0	0
	Overfitting Model	18	59	86	55	70	88	85	36
	Redundant Model	16	0	1	0	0	0	0	0
$\varepsilon \sim$ 0.5N(0,1)+0.5N(0,0.5)	True model	20	51	21	45	36	16	20	63
	Other Correct Model	35	0	1	0	0	0	0	0
	Overfitting Model	25	49	78	55	64	84	80	37
	Redundant Model	20	0	0	0	0	0	0	0
$\varepsilon \sim$ 0.5N(0,1)+0.5N(0,2)	True model	22	24	11	27	10	6	5	28
	Other Correct Model	40	2	2	8	1	0	0	19
	Overfitting Model	18	55	59	53	67	69	67	31
	Redundant Model	20	19	28	12	22	25	28	22
$\varepsilon \sim$ 0.5N(0,1)+0.5t(3)	True model	21	35	24	32	27	15	16	42
	Other Correct Model	35	1	0	1	1	0	0	8
	Overfitting Model	21	52	70	65	64	75	74	43
	Redundant Model	23	12	6	2	8	10	10	7
$\varepsilon \sim$ 0.5N(0,1)+0.5t(5)	True model	27	35	20	33	25	8	10	48
	Other Correct Model	31	3	0	0	0	3	3	10
	Overfitting Model	22	58	76	65	70	81	79	37
	Redundant Model	20	4	4	2	5	8	8	5
$\varepsilon \sim$ PE(0,1,0.85)	True model	15	26	7	30	13	5	4	25
	Other Correct Model	48	6	3	5	4	2	4	19
	Overfitting Model	16	54	74	53	61	64	61	33
	Redundant Model	21	14	16	12	22	29	31	23
$\varepsilon \sim$ PE(0,1,3)	True model	27	41	12	26	21	9	10	55
	Other Correct Model	40	1	0	1	3	0	0	7
	Overfitting Model	21	54	78	69	69	86	85	33
	Redundant Model	12	4	10	4	7	5	5	5
$\varepsilon \sim$ SPE(3,0,1,1)	True model	26	49	15	34	35	15	17	52
	Other Correct Model	33	0	1	0	0	0	0	2
	Overfitting Model	18	50	82	65	64	84	82	45
	Redundant Model	23	1	2	1	1	1	1	1
$\varepsilon \sim$ SPE(2,0,1,1)	True model	19	39	16	26	28	6	8	50
	Other Correct Model	41	2	1	0	0	1	1	8
	Overfitting Model	23	59	80	74	68	92	90	41
	Redundant Model	17	0	3	0	4	1	1	1
$\varepsilon \sim$ SPE(3,0,1,1.25)	True model	22	13	7	6	7	1	1	14
	Other Correct Model	51	0	0	0	0	0	0	1
	Overfitting Model	17	86	88	92	90	98	98	81
	Redundant Model	10	1	5	2	3	1	1	4

Table 5.3: Simulated Data: All Possible Subset Model Selection in 100 runs(2)

Error Distribution	Model Category	$\text{RICOMP}_{\text{misspec}}$				
		OLS	Huber	Andrews	Tukey	Hampel
$\varepsilon \sim$ N(0,0.5)	True model	1	98	96	96	96
	Other Correct Model	93	0	0	0	0
	Overfitting Model	0	2	4	4	4
	Redundant Model	6	0	0	0	0
$\varepsilon \sim$ 0.5N(0,1)+0.5N(0,0.5)	True model	2	97	94	98	98
	Other Correct Model	87	0	3	0	1
	Overfitting Model	1	3	3	2	1
	Redundant Model	10	0	0	0	0
$\varepsilon \sim$ 0.5N(0,1)+0.5N(0,2)	True model	1	67	58	64	54
	Other Correct Model	90	32	35	31	34
	Overfitting Model	3	0	4	2	7
	Redundant Model	6	1	3	3	5
$\varepsilon \sim$ 0.5N(0,1)+0.5t(3)	True model	2	85	83	85	79
	Other Correct Model	87	13	14	13	17
	Overfitting Model	1	2	3	2	4
	Redundant Model	10	0	0	0	0
$\varepsilon \sim$ 0.5N(0,1)+0.5t(5)	True model	0	87	90	93	89
	Other Correct Model	94	13	8	7	9
	Overfitting Model	0	0	2	0	2
	Redundant Model	6	0	0	0	0
$\varepsilon \sim$ PE(0,1,0.85)	True model	1	61	56	63	55
	Other Correct Model	93	38	41	34	43
	Overfitting Model	0	1	3	3	2
	Redundant Model	6	0	0	0	0
$\varepsilon \sim$ PE(0,1,3)	True model	1	89	72	88	81
	Other Correct Model	97	7	24	8	15
	Overfitting Model	0	4	3	4	4
	Redundant Model	2	0	1	0	0
$\varepsilon \sim$ SPE(3,0,1,1)	True model	3	98	90	96	94
	Other Correct Model	92	2	8	3	4
	Overfitting Model	0	0	2	1	2
	Redundant Model	5	0	0	0	0
$\varepsilon \sim$ SPE(2,0,1,1)	True model	2	94	85	97	93
	Other Correct Model	94	6	10	3	6
	Overfitting Model	0	0	5	0	1
	Redundant Model	4	0	0	0	0
$\varepsilon \sim$ SPE(3,0,1,1.25)	True model	2	86	70	78	80
	Other Correct Model	97	2	8	1	3
	Overfitting Model	0	11	18	20	16
	Redundant Model	1	1	4	1	1

Table 5.2 shows that the information criteria based on the ordinary least squares estimator (OLS) is not comparable with those based on the robust estimators. The ICOMP based on OLS estimator tends to both underfit the model by selecting the "other correct models" and overfit the model by selecting the "overfitting models". Generally speaking, the RICOMP based on the robust estimators can pick up the true models more often than the OLS method but tend to overfit the model. Since AICR, HAIC and RBIC are computed based on the Huber's estimation, we focus on comparisons between the RICOMP based on Huber's estimation and AICR, HAIC and RBIC. We see that RICOMP based on Huber's estimation outperform AICR and HAIC. RBIC is doing better than RICOMP in terms of hitting the true models more often. But RICOMP is still comparable with RBIC.

Table 5.3 should give us a better idea how the misspecification resistant version of RICOMP, $\text{RICOMP}_{\text{misspec}}$, performs on the model subset selection. Since all of the ten models contain outliers and most of their error distributions are misspecified (deviated from normal distribution), the results in this table are more reliable. We can see that the $\text{RICOMP}_{\text{misspec}}$ based on the robust estimates definitely outperforms the $\text{ICOMP}_{\text{misspec}}$ based on the OLS estimates. The $\text{ICOMP}_{\text{misspec}}$ based on the OLS estimates severely underfit the model by only selecting the subsets of the true model. The $\text{RICOMP}_{\text{misspec}}$ based on the robust estimates can pick up the true model over 95% of the time for the models with error distributions $N(0, 0.5), 0.5N(0, 1) + 0.5N(0, 0.5), SPE_3(0, 1, 1), SPE_2(0, 1, 1)$. For the models with error distributions of normal contaminated with t distributions ($0.5N(0, 1) + 0.5t(3)$ and $0.5N(0, 1) + 0.5t(5)$) and PE(0,1,3), the $\text{RICOMP}_{\text{misspec}}$ can pick up the true model over 85% of the time.

If we compare the results in both Tables 5.2 and 5.3, we will see that the $\text{RICOMP}_{\text{misspec}}$ based on the robust estimators works the best among all the information criteria. Particularly, although RICOMP based on Huber's estimators is not doing as well as RBIC, $\text{RICOMP}_{\text{misspec}}$ based on Huber's estimators is doing much better than RBIC for all of the models with different distributions. Since RBIC does not have the misspecification resistant version, it is completely not comparable with the $\text{RICOMP}_{\text{misspec}}$. We can also see that both RICOMP and the other three robust information criteria (AICR, HAIC and RBIC) in Table 5.2 do not work well on the models with error distributions $0.5N(0, 1) + 0.5N(0, 2)$ and $PE(0, 1, 0.85)$ because of the heavy tailed distribution. However, $\text{RICOMP}_{\text{misspec}}$ in Table 5.3 performs much better. $\text{RICOMP}_{\text{misspec}}$ can pick up the true model 50% to 60% of the time. And they can pick up the "correct model" (including the "true model" and "other correct model") over 90% of the time.

One interesting error distribution we should notice is the $SPE_3(0, 1, 1.25)$. This SPE distribution introduces both skewness and kurtosis to the model and makes the model selection most difficult. If we look at Table 5.2, none of the information criteria, including

RICOMP does a good job. In particular, all the information criteria tend to overfit the model. The reason is that the penalty term in all the information criteria do not penalize enough for the complexity of the model, so that it does not have a balanced trade off with the lack of fit part of the criteria. After we introduced the misspecification resistant version of RICOMP, $\text{RICOMP}_{\text{misspec}}$, the case changes. We can see from Table 5.3, $\text{RICOMP}_{\text{misspec}}$ picks up the true model over 80% of the time for this distribution, which is a tremendous improvement. That is why we prefer the misspecification resistant information criteria $\text{RICOMP}_{\text{misspec}}$ for the complex model subset selection. Unfortunately, the robust version of AIC and BIC, namely AICR, HAIC and RBIC, are not resistant to the presence of skewness and kurtosis in the model.

The box plots for the values of RICOMP, $\text{RICOMP}_{\text{misspec}}$, AICR, HAIC and RBIC of the true models for four of the distributions as examples are displayed in Figures 5.6 and 5.7 (to save space, we did not output the box plots for all error distributions). In both plots, hub_1 and hub_2 represent the RICOMP and $\text{RICOMP}_{\text{misspec}}$ values based on Huber's estimator; and_1 and and_2 represent the RICOMP and $\text{RICOMP}_{\text{misspec}}$ values based on Andrews' estimator; tuk_1 and tuk_2 represent the RICOMP and $\text{RICOMP}_{\text{misspec}}$ values based on Tukey's estimator; ham_1 and ham_2 represent the RICOMP and $\text{RICOMP}_{\text{misspec}}$ values based on Hampel's estimator. Notice that on both figures, we did not plot ICOMP and $\text{ICOMP}_{\text{misspec}}$ values based on the OLS estimators.

GA Subset Selection

In this section, we carry out the model subset selection using Genetic Algorithm (GA) as the optimization method with the robust and misspecification resistant version of ICOMP, namely $\text{RICOMP}_{\text{misspec}}$, as its fitness function. Our goal is to test if the three-way hybrid method can fulfill the selection of true model when the true model exists.

Simulate $\{x_1, x_2, x_3\}$ from (5.40) and $\{x_4, \ldots, x_{10}\}$ from (5.42) with sample size $n = 100$. Generate y from (5.41), where $\varepsilon \sim SPE_3(0, 1, 1.25)$. The model subset selection is carried out using GA with $\text{RICOMP}_{\text{misspec}}$ as its the fitness function, where the $\text{RICOMP}_{\text{misspec}}$ is computed based on the Huber's minimax function with tuning constant $k = 1.345$. The parameters in GA we used are given in Table 5.4.

We run GA 15 times and see which model is picked up by GA. For each run of GA, new simulation data set was generated.

The results for the 15 GA runs are shown in Table 5.5. 12 out of these 15 runs picked up the true model $\{x_1, x_2, x_3\}$. Three runs overfit the model with one more predictor (in the 5th run, 7th run and 14th run). This result is reasonable considering that the corresponding all possible subset selection can identify the true model 86 times in 100 runs.

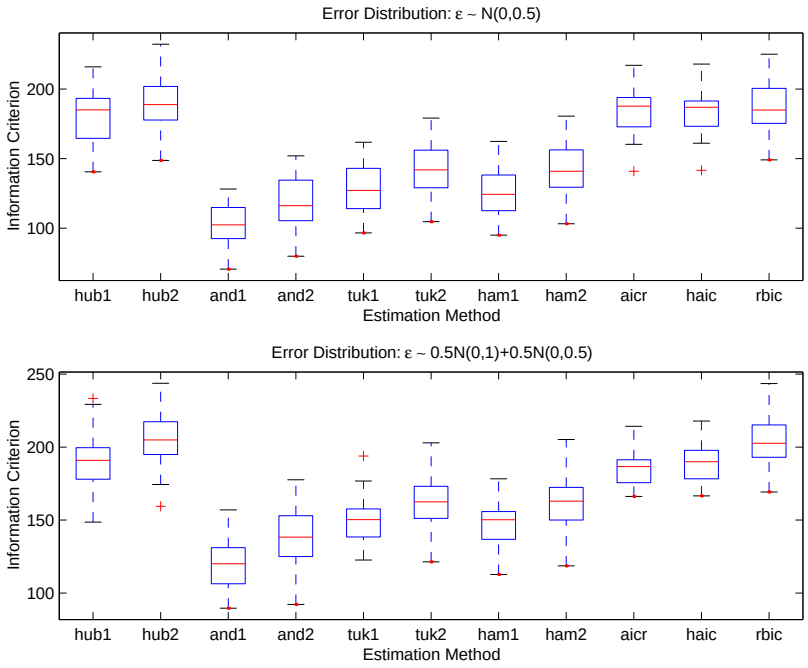

Figure 5.6: Simulated Data: Boxplot for the All Possible Model Selection(1).

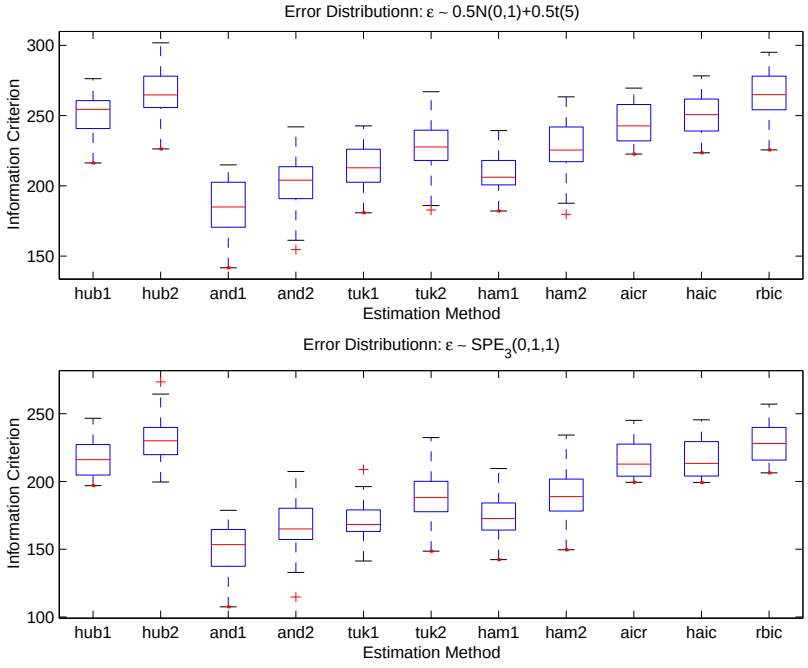

Figure 5.7: Simulated Data: Boxplot for the All Possible Model Selection(2).

Table 5.4: GUI Inputs of GA Parameters for Simulated Data

No. of runs	15
No. of generations	50
Population size	30
Estimation method	Huber's
Fitness value	$\text{RICOMP}_{\text{misspec}}$
Probability of crossover	0.5
Crossover Method	uniform
Probability of Mutation	0.01
Elitism	Yes

Table 5.5: Simulated Data: Model Subset Selection in 15 Runs of the GA

Run Number	Variable Selected	Binary String	$\text{RICOMP}_{\text{misspec}}$
1	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	256.36
2	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	241.7
3	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	259.28
4	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	274.75
5	$\{x_1, x_2, x_3, x_5\}$	1 1 1 0 1 0 0 0 0 0	245.85
6	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	243.06
7	$\{x_1, x_2, x_3, x_6\}$	1 1 1 0 0 1 0 0 0 0	244.35
8	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	262.18
9	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	229.5
10	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	238.76
11	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	216.87
12	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	236.3
13	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	266.08
14	$\{x_1, x_2, x_3, x_4\}$	1 1 1 1 0 0 0 0 0 0	258.71
15	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	257.41

Table 5.6: Simulated Data: Model Subset Selection in One Run of the GA

Generation	variable selected	Binary String	RICOMPmisspe
1	$\{x_1, x_2, x_3, x_8\}$	1 1 1 0 0 0 0 1 0 0	274.99
2	$\{x_1, x_2, x_3, x_8\}$	1 1 1 0 0 0 0 1 0 0	274.99
3	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	257.41
4	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	257.41
5	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	257.41
$\vdots$	$\vdots$	$\vdots$	$\vdots$
50	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	257.41

Table 5.6 shows the result from one run of the GA. The results of 50 generations of GA procedure are recorded. We can see that in this particular run, GA with RCIOMPmisspec as fitness function based on Huber's estimator selected the true model as early as the 3rd generation.

The 2D and 3D plots for this one run of the GA are shown in Figure 5.8 and 5.9, respectively, to show the optimization procedure in GA.

Conclusion and Discussion

From this simulation study, we conclude that the robust and misspecification resistant version of ICOMP, $\text{RICOMP}_{\text{misspec}}$, is an effective information criterion for the model selection. It outperforms the other robust information criteria in the literature, AICR, HAIC and RBIC, which are vulnerable to the model misspecification, especially when the random error departs from the normal distribution. When skewness exists in the error, the three information criteria work poorly in the model subset selection. The information criteria computed based on the OLS estimator is completely non-comparable with those of the robust version.

GA combined with the robust and misspecification resistant information complexity, $\text{RICOMP}_{\text{misspec}}$, is a fast and effective model selection method, which can reach to the optimal model quickly without having to search the entire model space. This is especially valuable when applied to large data sets with huge number of predictor variables in high dimensional data mining and knowledge discovery.

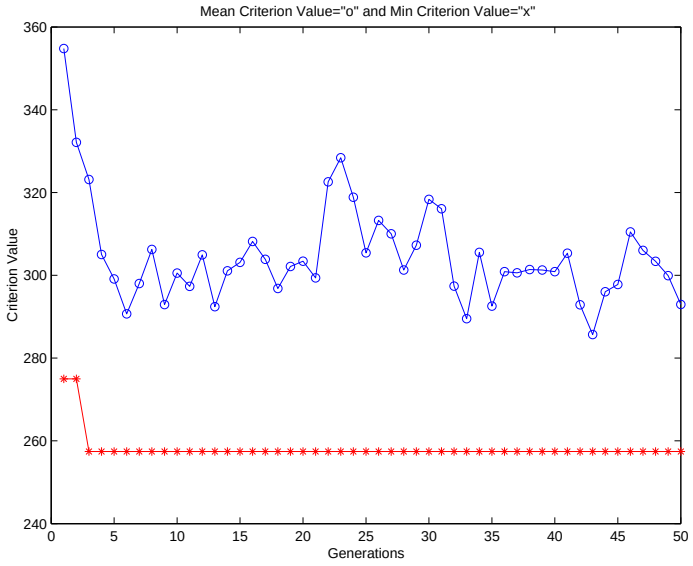

Figure 5.8: Simulated Data: 2D plot for One Run of the GA.

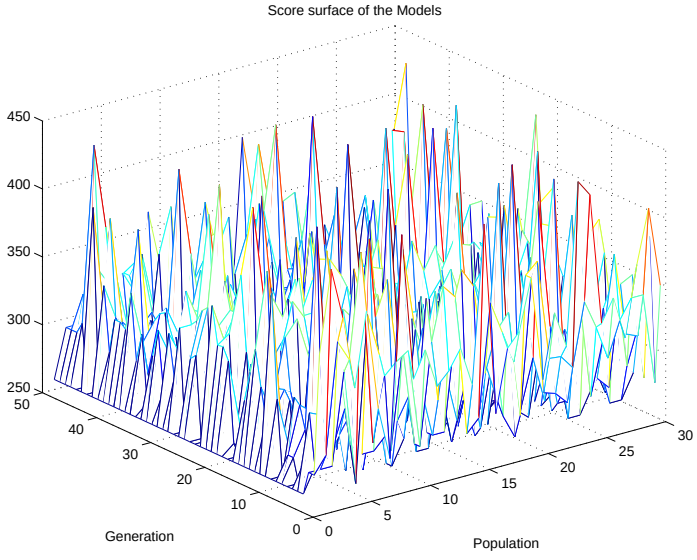

Figure 5.9: Simulated Data: 3D plot for One Run of the GA.

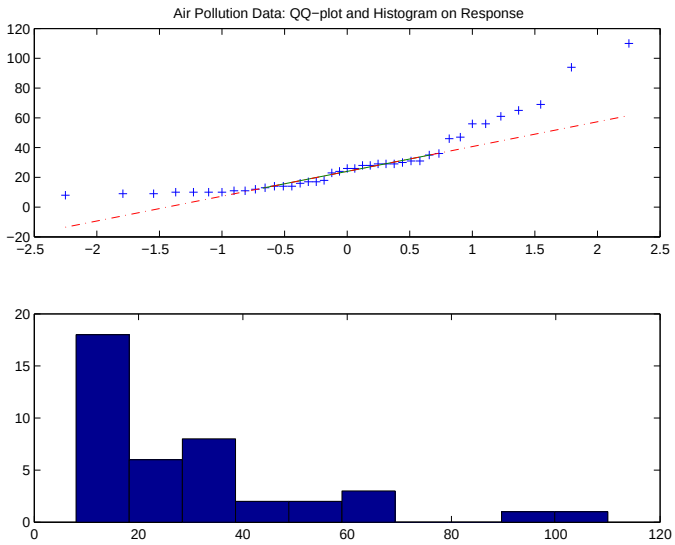

Figure 5.10: Air Pollution Data: QQ-plot and Histogram of the Response.

5.6.2 Real Data Examples

Example 1: Air Pollution Data

In this example, we consider the air pollution data from Sokal and Rohlf (1981), which were collected from the United States government publications. It gives the mean values for the air pollution and related sources for $n = 41$ U.S. cities over the years $1961 - 1971$.

The dependent variable is:

$y = \mathrm{SO}_2$: Sulfur dioxide content of air in micrograms per cubic meter,

and the six predictor variables are:

$x_1 =$ Temp: Average annual temperature in degrees Fahrenheit,

$x_2 =$ Man: Number of manufacturing enterprises employing 20 or more workers

$x_3 =$ Pop: Population size in thousands from the 1970 census,

$x_4 =$ Wind: Average annual wind speed in miles per hour,

$x_5 =$ Rain: Average annual precipitation in inches,

$x_6 =$ RainDays: Average number of days with precipitation per year.

The QQ-plot and histogram shown in Figure 5.10 give a first glance of the response variable y, SO_2. The distribution of the response variable is extremely right shewed. The $\mathrm{RICOMP}_{\mathrm{misspec}}$ is thus preferred for model selection in the analysis of this data. All possible subset selection and GA subset selection will be performed on this data and compared.

The correlation matrix between y (SO_2) and $x's$ is given in Table 5.7. We can see from this table that the response variable is relatively highly correlated with predictor x_2

Table 5.7: Air Pollution Data: Correlation Matrix

	y	x_1	x_2	x_3	x_4	x_5	x_6
y	1						
x_1	-0.4336	1					
x_2	**0.6448**	-0.19	1				
x_3	0.4938	-0.0627	**0.9553**	1			
x_4	0.0947	-0.3497	0.2379	0.2126	1		
x_5	0.0543	0.3863	-0.0324	-0.0261	-0.013	1	
x_6	0.3696	-0.4302	0.1318	0.0421	0.1641	0.4961	1

(Number of manufacturing enterprises) and mildly correlated with x_3 (population size) and x_1 (average annual temperature). The predictors x_2 (Number of manufacturing enterprises) and x_3 (population size) are highly correlated to each other.

The estimation of parameters for the full model is given in Table 5.8. The estimated coefficients and their corresponding t statistic are presented in the first part of the table. We can see that the four robust functions (Huber, Andrews, Tukey and Hampel) give similar estimators, which are different from the OLS estimator. They agree that the predictor variable x_2 is the most important variable in predicting the response variable with the highest absolute t value, which is greater than the approximate critical value of 2. The four robust estimations reduce the standard deviation (root mean squared error) in the model and also lower the RICOMP and $\text{RICOMP}_{\text{misspec}}$ values for the full model. Since the Hampel's estimation gives the smallest model standard deviation, we decide to make model subset selection using $\text{RICOMP}_{\text{misspec}}$ based on Hampel's estimate.

The top 5 subset model selected by the all possible method is given in Table 5.9. As we expected, x_2 is selected as the only predictor for the top 1 subset model and all the rest of the top 5 models include x_2 as a predictor, which is consistent from what we observe from the full model estimation.

We then ran GA for the model subset selection 100 times using the same information criterion to compare the results. The parameters used in GA and the results are summarized in Tables 5.10 and 5.11.

We see from Table 5.11 that the 100 runs of the GA selected only two subset models. It chose the top 1 subset model in all possible selection for 90 times out 100 runs. It selected the second best all possible subset model for 10 times. The GA subset selection is consistent with the all possible subset selection method.

The 3D plot of 100 runs of GA in Figure 5.11 shows the optimization procedure in GA.

Table 5.8: Air Pollution Data: Estimates of the Full Model Parameters

	OLS	Huber	Andrews	Tukey	Hampel
$\hat{\beta}_0$	111.73 (-2.36)[a]	106.84 (-2.21)	103.30 (-2.12)	99.85 (2.01)	103.55 (2.14)
$\hat{\beta}_1$	-1.27 (-2.04)	-1.12 (-1.76)	-1.10 (-1.72)	-1.02 (1.57)	-1.09 (1.71)
$\hat{\beta}_2$	0.06 (-4.12)	0.06 (-3.59)	0.06 (-3.53)	0.05 (3.26)	0.06 (3.47)
$\hat{\beta}_3$	-0.04 (-2.60)	-0.03 (-2.02)	-0.03 (-1.98)	-0.03 (1.71)	-0.03 (1.91)
$\hat{\beta}_4$	-3.18 (-1.75)	-3.78 (-2.02)	-3.60 (-1.92)	-3.79 (1.98)	-3.69 (1.98)
$\hat{\beta}_5$	0.51 (-1.41)	0.40 (-1.07)	0.38 (-1.00)	0.32 (0.83)	0.38 (1.01)
$\hat{\beta}_6$	-0.05 (-0.32)	-0.02 (-0.13)	-0.01 (-0.04)	0.01 (0.07)	-0.01 (0.07)
$\hat{\sigma}$	14.636	11.9899	11.8332	11.8169	11.7078
RICOMP	396.65	367.0384	353.0385	366.737	367.1705
$\text{RICOMP}_{\text{misspec}}$	402.5282	369.2291	354.433	368.079	368.6085

[a]the number in the parenthesis is the t statistic

Table 5.9: Air Pollution Data: Top 5 Subsets from All Possible Subset Model Selection

Ranking	subset model	$\text{RICOMP}_{\text{misspec}}$ with Hampel
1	$\{x_0, -, x_2, -, -, -, -\}$	320.0774
2	$\{x_0, -, x_2, -, x_4, -, x_6\}$	322.5799
3	$\{x_0, x_1, x_2, -, x_4, -, -\}$	323.6904
4	$\{x_0, -, x_2, x_3, x_4, -, x_6\}$	325.6095
5	$\{x_0, -, x_2, x_3, -, -, -\}$	325.6302

Table 5.10: Air Pollution Data: GUI Inputs of GA Parameters

No. of runs	100
No. of generations	30
Population size	20
Estimation method	Hampel
Fitness value	$\text{RICOMP}_{\text{misspec}}$
Probability of crossover	0.7
Crossover Method	uniform
Probability of Mutation	0.01
Elitism	Yes

Table 5.11: Air Pollution Data: Model Subset Selection in 100 Runs of the GA

GA Ranking	Chromosome	Binary String	$\text{RICOMP}_{\text{misspec}}$	Hit ratio[a]
1 (1)[b]	0 - 2 - - - -	1 0 1 0 0 0 0	320.08	90%
2 (2)	0 - 2 - 4 - 6	1 0 1 0 1 0 1	322.58	10%

[a]How many times the subset model is selected in 100 runs of GA

[b]The parenthesis includes the corresponding all possible model selection rankings for comparison purposes.

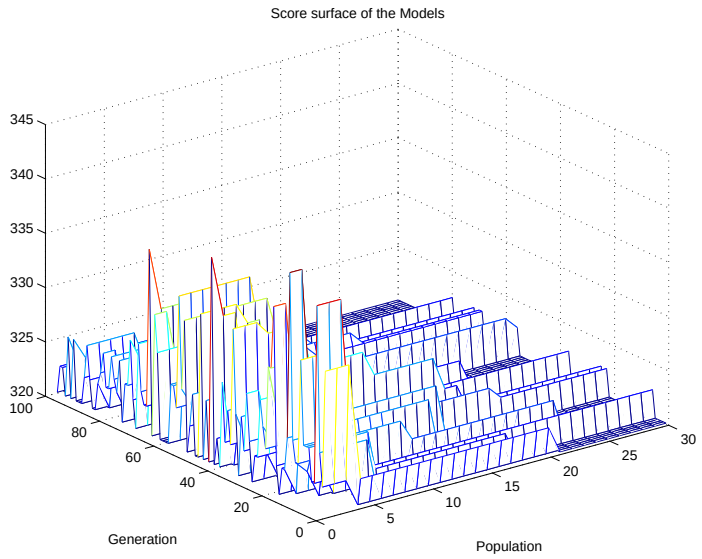

Figure 5.11: Air Pollution Data: 3D-plot of 100 Runs of the GA.

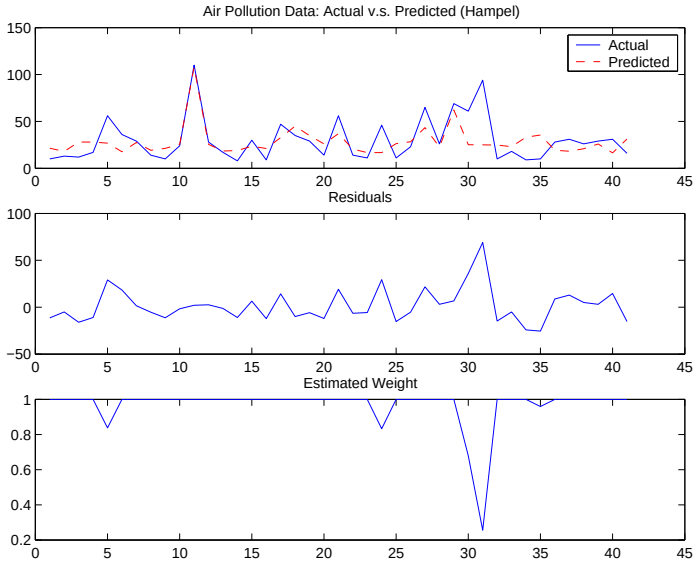

Figure 5.12: Air Pollution Data: Plot of the Best Subset Model.

Interpretation Both all possible subset selection and GA agree in selecting x_2 (number of manufacturing enterprises employing 20 or more workers) as the only variable to predict the response. The final fitted model is

$$\begin{aligned} SO_2 &= 15.50 + 0.0277\text{Man} \\ \text{(t-statistics)} &\quad (4.113) \quad (5.385) \end{aligned}$$

where t statistics are given in the parenthesis. When one more manufacturing enterprise employing 20 or more workers is constructed in the area, the sulfur dioxide content of air in the city will increase 0.0277 micrograms per cubic meter. Figure 5.12 shows the plots of actual versus predicted values, residuals and estimated weights for the best subset model. The larger residuals are assigned lower weights.

Example 2: Body Fat Data

In this example, we analyze body fat data and determine the best subset of the predictors.

This data is obtained from Statlib dataset (http://lib.stat.cmu.edu/datasets/bodyfat). It lists estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men.

The **response** variable is

y: Percent body fat from Siri's (1956) equation;

The 13 **predictor** variables are:

x_1: Age (years),

x_2: Weight (lbs),

x_3: Height (inches),

x_4: Neck circumference (cm),

x_5: Chest circumference (cm),

x_6: Abdomen 2 circumference (cm),

x_7: Hip circumference (cm),

x_8: Thigh circumference (cm),

x_9: Knee circumference (cm),

x_{10}: Ankle circumference (cm),

x_{11}: Biceps (extended) circumference (cm),

x_{12}: Forearm circumference (cm),

x_{13}: Wrist circumference (cm).

The measurement standards are listed in Benhke and Wilmore (1974, page 45-48), where for instance, the abdomen 2 circumference is measured "laterally, at the level of the iliac crests, and anteriorly, at the umbilicus".

These data are used to produce the predictive equations for lean body weight (Penrose et al., 1985), where the equations were constructed from the first 143 of the 252 observations.

The QQ-plot and histogram of the response variable are shown in Figure 5.13. The response is approximately normally distributed with one possible outlier. We then construct the histograms for some of the predictors (x_2, x_3, x_4 and x_7), which are displayed in Figure 5.14. There are obvious outliers in these predictors. To save space, we just show the distributions of four out of thirteen predictors here. We can also observe possible outliers in some of the other predictors ($x_6, x_8, x_9, x_{10}, x_{12}$).

The correlation coefficient matrix of response and predictors are given in Table 5.12. It shows strong collinearities among the predictors of weight and all the body circumference measurements, which make sense in real life. Specifically, weight (x_2) is highly correlated to all the body circumference measurements (x_4 - x_{13}). All the body circumference measurements are positively highly correlated to each other. The age (x_1) and height (x_3) do not show strong correlations with the other predictors. The response is relatively highly correlated to the weight and several body circumference measurements. The heavy correlation structure in the predictor variables make the model subset selection necessary and important. Since the model is misspecified in terms of collinearity and presence of an outlier, we prefer the $\text{RICOMP}_{\text{misspec}}$ as our model selection criterion.

The root mean square error (RMSE) of the full model based on each estimation method is given in Table 5.13. We can see that the RMSE are about the same for both the OLS

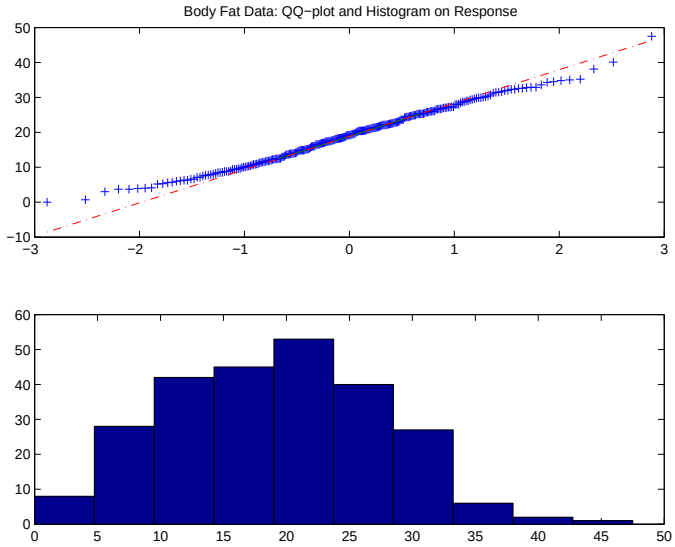

Figure 5.13: Body Fat Data: QQ-plot and Histogram of the Response.

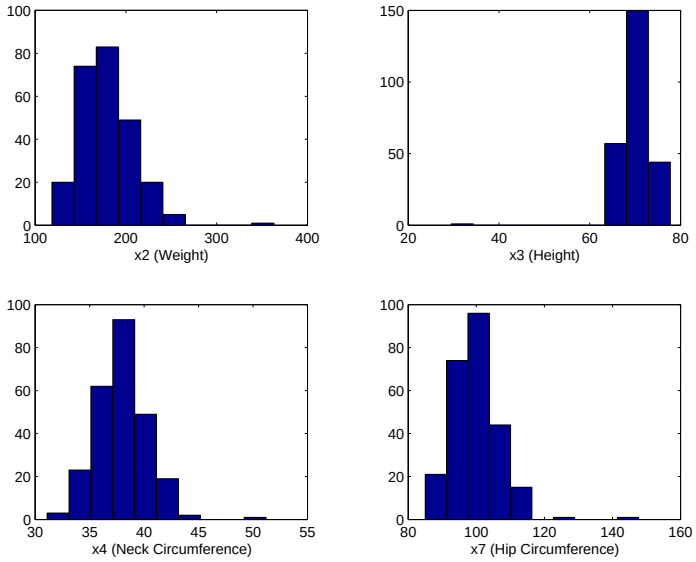

Figure 5.14: Body Fat Data: Histogram of the Predictors.

Table 5.12: Body Fat Data: Correlation Matrix

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
y	1													
x_1	0.29	1												
x_2	0.61	-0.01	1											
x_3	-0.09	-0.17	0.31	1										
x_4	0.49	0.11	0.83	0.25	1									
x_5	0.70	0.18	0.89	0.13	0.78	1								
x_6	0.81	0.23	0.89	0.09	0.75	0.92	1							
x_7	0.63	-0.05	0.94	0.17	0.74	0.83	0.87	1						
x_8	0.56	-0.20	0.87	0.15	0.70	0.73	0.77	0.90	1					
x_9	0.51	0.02	0.85	0.29	0.67	0.72	0.74	0.82	0.80	1				
x_{10}	0.27	-0.11	0.61	0.26	0.48	0.48	0.45	0.56	0.54	0.61	1			
x_{11}	0.49	-0.04	0.80	0.21	0.73	0.73	0.69	0.74	0.76	0.68	0.48	1		
x_{12}	0.36	-0.09	0.63	0.23	0.62	0.58	0.50	0.55	0.57	0.56	0.42	0.68	1	
x_{13}	0.35	0.21	0.73	0.32	0.74	0.66	0.62	0.63	0.56	0.66	0.57	0.63	0.59	1

estimation and robust estimation. The reason may be that the response variable of this data is not severely departed from being a normal distribution. Since we have both potential outliers in the response variable and most of the predictor variables, we prefer the robust estimation here. Among the robust estimation, Andrews' sine wave function gives the relatively smaller RMSE. Therefore, we are going to perform the model subset selection using the $\text{RICOMP}_{\text{misspec}}$ combined with Andrews' robust estimates.

The top 10 subsets by all possible model selection are given in Table 5.14. The parameters used in GA and the subset selection results in 100 runs of the GA are given in Tables 5.15 and 5.16, respectively.

We can see from both tables that the GA subset selection is consistent with the all possible selection. In 100 runs of the GA, it picks up the first ranking model (in all possible selection) for 51 times, the second ranking model for 28 times and the third ranking model for 9 times. It picks up the top 10 models 99 times out of 100 runs of the GA. The last model GA picked is ranking 47^{th} in the all possible selection, which is not bad considering we have $2^{13} - 1 = 8191$ subset models in total for the all possible subset selection. The optimization procedure of GA is shown in the 3D plot in Figure 5.15.

Both all possible and GA model selection methods agree to pick up the parsimonious subset with x_1 (Age), x_6 (Abdomen 2 Circumference in cm) and x_{13} (Wrist circumference in cm).

Interpretation The final fitted model with x_1, x_6 and x_{13} is as follows:

Table 5.13: Body Fat Data: RMSE for the Full Model

	RMSE				
Full Model	OLS	Huber	Andrews	Tukey	Hampel
$\{x_1, x_2, \ldots, x_{13}\}$	4.3053	4.307	**4.3056**	4.306	4.306

Table 5.14: Body Fat Data: Top 10 Subsets by All Possible Model Selection

Ranking	Selected Variables	$\text{RICOMP}_{\text{misspec}}$ with Andrews
1	0 1 - - - - 6 - - - - - - 13	1367.5872
2	0 - 2 - - - 6 - - - - - - -	1369.4238
3	0 1 - - 4 - 6 - - - - - - -	1381.6407
4	0 - 2 3 - - 6 - - - - - - -	1384.4165
5	0 - - - - - 6 - - 9 - - - -	1385.4765
6	0 - - - 4 - 6 7 - - - - - -	1386.8336
7	0 - 2 - 4 - 6 - - - - - - 13	1389.5806
8	0 - - 3 - - 6 - - - - - - -	1390.1929
9	0 1 - - 4 - 6 - - - - - - 13	1390.3188
10	0 1 - - - - 6 - - - - 11 - 13	1390.3694

Table 5.15: Body Fat Data: GUI Inputs of GA Parameters

No. of runs	100
No. of generations	50
Population size	50
Estimation method	Andrews'
Fitness value	$\text{RICOMP}_{\text{misspec}}$
Probability of crossover	0.5
Crossover Method	uniform
Probability of Mutation	0.01
Elitism	Yes

Table 5.16: Body Fat Data: Model Subset Selection in 100 Runs of the GA

GA Ranking	Chromosome	Binary String	$\text{RICOMP}_{\text{misspec}}$	Hit Ratio[a]
1 (1)[b]	0 1 - - - - 6 - - - - - - 13	1 1 0 0 0 0 1 0 0 0 0 0 0 1	1367.6	51
2 (2)	0 - 2 - - - 6 - - - - - - -	1 0 1 0 0 0 1 0 0 0 0 0 0 0	1369.4	28
3 (3)	0 1 - - 4 - 6 - - - - - - -	1 1 0 0 1 0 1 0 0 0 0 0 0 0	1381.6	9
4 (6)	0 - - - 4 - 6 7 - - - - - -	1 0 0 0 1 0 1 1 0 0 0 0 0 0	1386.8	7
5 (8)	0 - - 3 - - 6 - - - - - - -	1 0 0 1 0 0 1 0 0 0 0 0 0 0	1390.2	3
6 (10)	0 1 - - - - 6 - - - - 11 - 13	1 1 0 0 0 0 1 0 0 0 0 1 0 1	1390.4	1
7 (47)	0 - 2 3 - 5 6 - - - - - - 13	1 0 1 1 0 1 1 0 0 0 0 0 0 1	1406.8	1

[a]How many times the subset model is selected in 100 runs of the GA

[b]The parenthesis includes the corresponding all possible model selection ranking for comparison purpose.

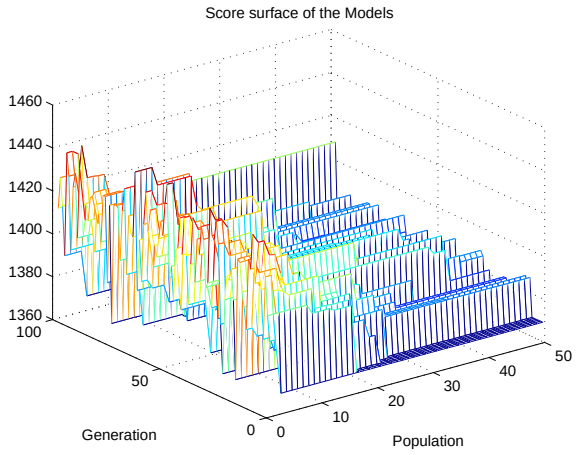

Figure 5.15: Body Fat Data: 3D-plot of 100 Runs of the GA.

$$\begin{array}{lcccccc} \text{Percent Body Fat} & = & -10.25 + 0.083(\text{Age}) + 0.756(\text{Abdomen 2 Circ.}) - 2.43(\text{Wrist Circ.}) \\ \text{(t-statistics)} & & (-1.771) & (3.531) & (21.792) & (6.176) \end{array}$$

where the t statistics are given beneath the fitted equation. When the age increases by 1 year, the body fat estimated in Siri's (1956) equation will increase by 0.083%. When the abdomen 2 circumference increases 1 cm, the body fat estimated in Siri's equation will increase by 0.756%. When the wrist circumference increases 1 cm, the body fat estimated in Siri's equation will decrease by 2.43%. Figure 5.16 shows the plots of actual versus predicted values, residuals and estimated weights for the best subset model. We see that the larger residuals were assigned lower weights.

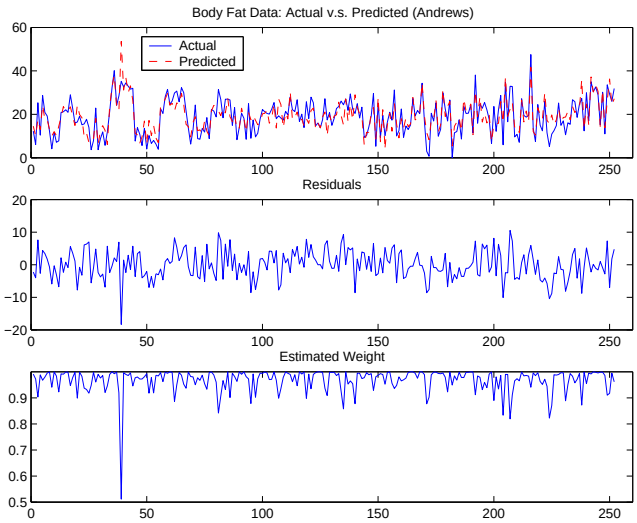

Figure 5.16: Body Fat Data: Plot of the Best Subset Model.

Chapter 6

Robust and Misspecification-Resistant Model Selection in Multivariate Regression

6.1 Robust Estimates in Multivariate Linear Regression

Consider the classical multivariate linear regression (MVR) model given in matrix form by

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{6.1}$$

where

- $\mathbf{Y}$ is an $(n \times m)$ matrix of n independent observations on m responses;
- $\mathbf{X}$ is an $(n \times p)$ matrix of n observations on p nonstochastic independent variables;
- $\mathbf{B}$ is a $(p \times m)$ matrix of regression coefficients;
- $\mathbf{E}$ is an $(n \times m)$ matrix of random errors, which satisfies

$$Cov(\mathbf{E}) = \mathbf{\Sigma} \otimes \mathbf{I_n}, \tag{6.2}$$

  where $\otimes$ denotes the Kronecker product. and $\mathbf{E} \sim N_{nm}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I_n})$.

A few robust estimators for the MVR model have been proposed in the literature. Koenker and Portnoy (1990) first proposed a robust estimate for the MVR model, which

is actually a robust alternative for the seemingly unrelated regression (SUR) estimator of Zellner (1962). Rousseeuw et al. (2004) proposed robust estimates for the MVR model based on a robust estimate of the covariance matrix of the predictor and response variables. Ben et al. (2006) derives the robust estimates for regression coefficients and covariance matrix simultaneously by minimizing the determinant of the covariance matrix estimate, subject to a constraint on a robust scale of the Mahalanobis norms of the residuals.

In this dissertation, we derive the estimators of regression coefficients and the covariance matrix by minimizing the robust Mahalanobis distance on the residuals, which we called the RMD estimator. The two stage estimate procedure is illustrated as follows.

Stage 1. Obtain one-step RMD estimates

1. Calculate the MLE estimate of B and $\mathbf{\Sigma}$

The maximum likelihood estimate for $\mathbf{B}$ of MVR model in equation (6.1) is given by

$$\widehat{\mathbf{B}}_{MLE} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}. \tag{6.3}$$

The estimated response $\mathbf{Y}$ is given by

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}}_{MLE}. \tag{6.4}$$

The estimated error matrix $\widehat{\mathbf{E}}$ is given by

$$\widehat{\mathbf{E}} = \mathbf{Y} - \widehat{\mathbf{Y}}, \tag{6.5}$$

where $\widehat{\mathbf{E}} = (\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n)'$ with $\widehat{\varepsilon}_i'$ $(i = 1, 2, \ldots, n)$ as the estimated i^{th} row of m-variate residuals.

The MLE for the covariance matrix is

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\widehat{\mathbf{E}}'\widehat{\mathbf{E}}. \tag{6.6}$$

2. Compute the Mahalanobis distance for the residuals

The squared Mahalanobis distance of the estimated i^{th} row of residuals $\widehat{\varepsilon}_i'$ is defined by

$$\mathbf{d_i} = \widehat{\varepsilon}_i'\widehat{\mathbf{\Sigma}}^{-1}\widehat{\varepsilon}_i. \tag{6.7}$$

In this step, we project the m-variate residual matrix $\widehat{\mathbf{E}}(n \times m)$ to the one-dimensional Mahalanobis distance $\mathbf{d} = (d_i)(i = 1, 2, \ldots, n)$, denoted by $\hat{\mathbf{E}}_{(n\times m)} \rightarrow \mathbf{d}_{(n\times 1)} = (d_i)$.

3. Compute the one-step RMD estimator for covariance matrix and coefficients

The robust estimator of the covariance matrix using the Mahalanobis distance method, $\widehat{\mathbf{\Sigma}}_{rmd}$ is defined by

$$\widehat{\mathbf{\Sigma}}_{rmd} = \frac{1}{n}\sum_{i=1}^{n} W(d_i)\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i', \tag{6.8}$$

where $W(d_i)$ is the weight function given by

$$W(d_i) = \begin{cases} \psi(d_i)/d_i, & \text{if } d_i \neq 0; \\ 1, & \text{if } d_i = 0 \end{cases} \tag{6.9}$$

in which the $\psi(\cdot)$ are the robust M functions given in Section 2.1.

The robust M-estimator of regression coefficients, $\widehat{\mathbf{B}}_{rmd}$, are defined by

$$\widehat{\mathbf{B}}_{rmd} = (\mathbf{X'W(d)X})^{-\mathbf{1}}\mathbf{X'W(d)Y}. \tag{6.10}$$

Stage 2. Obtain iterative RMD estimates

In this stage, we compute the RMD estimates of the MVR coefficients and the covariance matrix using the *iteratively reweighted least-squares* method similar to what we used for the MLR model.

1. Select initial estimate

Take the one-step RMD estimate as our initial estimate, denoted by $\widehat{\mathbf{B}}^{(0)} = \widehat{\mathbf{B}}_{rmd}$ and $\widehat{\mathbf{\Sigma}}^{(0)} = \widehat{\mathbf{\Sigma}}_{rmd}$.

2. At each iteration t, calculate the estimate of the coefficient and covariance matrix

$$\widehat{\mathbf{E}}^{(t)} = \mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}^{(t-1)},$$

$$\widehat{\mathbf{\Sigma}}^{(t)} = \frac{1}{n}\widehat{\mathbf{E}}'^{(t)}\widehat{\mathbf{E}}^{(t)},$$

$$\mathbf{d_i}^{(t)} = \widehat{\boldsymbol{\varepsilon}}_i'^{(t)}\widehat{\mathbf{\Sigma}}^{(t)-1}\widehat{\boldsymbol{\varepsilon}}_i^{(t)},$$

$$W_{(d_i)}^{(t)} = \begin{cases} \psi(d_i^{(t)})/d_i^{(t)}, & \text{if } d_i^{(t)} \neq 0; \\ 1, & \text{otherwise,} \end{cases}$$

$$\widehat{\mathbf{\Sigma}}_{RMD}^{(t)} = \frac{1}{n}\sum_{i=1}^{n} W_{(d_i)}^{(t)}\widehat{\boldsymbol{\varepsilon}}_i^{(t)}\widehat{\boldsymbol{\varepsilon}}_i'^{(t)},$$

$$\widehat{\mathbf{B}}_{RMD}^{(t)} = (\mathbf{X'W}_{\mathbf{(d)}}^{\mathbf{(t)}}\mathbf{X})^{-\mathbf{1}}\mathbf{X'W}_{\mathbf{(d)}}^{\mathbf{(t)}}\mathbf{y}.$$

3. Repeat step 2 until $\widehat{\mathbf{B}}_{RMD}$ converges.

$\widehat{\mathbf{\Sigma}}^{(t)}_{RMD}$ and $\widehat{\mathbf{B}}^{(t)}_{RMD}$ are then the final iterative RMD estimates for the covariance matrix and regression coefficients.

6.2 Robust Information Criteria for MVR Model Selection

Recall the classic Akaike Information Criterion (AIC) for the multivariate regression model is given by

$$\text{AIC}_{\text{reg}} = nm\log(2\pi) + n\log|\mathbf{\Sigma}| + nm + 2[mp + m(m+1)/2] \tag{6.11}$$

and the classic Bayesian Information Criterion (BIC or SBC) for the multivariate regression model is given by

$$\text{BIC}_{\text{reg}} = nm\log(2\pi) + n\log|\mathbf{\Sigma}| + nm + [mp + m(m+1)/2]\log(n). \tag{6.12}$$

If we substitute the iterative RMD estimate $\widehat{\mathbf{\Sigma}}_{RMD}$ for the covariance matrix $\mathbf{\Sigma}$ in both equation 6.11 and 6.12, we obtain the robust information criteria for the MVR models, which are defined as follows:

$$\text{ROBAIC}_{\text{reg}} = nm\log(2\pi) + n\log|\widehat{\mathbf{\Sigma}}_{RMD}| + nm + 2[mp + m(m+1)/2] \tag{6.13}$$

and the robust Bayesian Information Criterion (ROBBIC) for the multivariate regression model is given by

$$\text{ROBBIC}_{\text{reg}} = nm\log(2\pi) + n\log|\widehat{\mathbf{\Sigma}}_{RMD}| + nm + [mp + m(m+1)/2]\log(n), \tag{6.14}$$

where $\widehat{\mathbf{\Sigma}}_{RMD}$ can be estimated using any of the M functions given in Section 2.1.

However, for the simplicity of comparisons in the following data examples, all the ROBAIC and ROBBIC are computed based on the Huber's estimator.

6.3 Robust ICOMP for MVR Model Selection

The robust version of Bozdogan's ICOMP(IFIM) for multivariate regression (MVR) model, denoted by RICOMP(IFIM)$_{MVR}$, is defined by

$$\text{RICOMP(IFIM)}_{\text{MVR}} = nm\log 2\pi + n\log\left|\widehat{\mathbf{\Sigma}}_{RMD}\right| + nm + 2C_1(\widehat{\mathcal{F}}^{-1}), \tag{6.15}$$

where $\widehat{\mathbf{\Sigma}}_{RMD}$ is the covariance matrix estimated by the Robust Mahalanobis distance (RMD) method. $\widehat{\mathcal{F}}^{-1}$ is the estimated inverse Fisher Information matrix, which is given by

$$\widehat{\mathcal{F}}^{-1} = \begin{pmatrix} \widehat{\mathbf{\Sigma}}_{RMD} \otimes (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2}{n} D_m^+ (\widehat{\mathbf{\Sigma}}_{RMD} \otimes \widehat{\mathbf{\Sigma}}_{RMD}) D_m^{+\prime} \end{pmatrix}, \tag{6.16}$$

where D_m^+ is the Moore-Penrose inverse of the duplication matrix D_m.

The complexity part $C_1(\widehat{\mathcal{F}}^{-1})$ can be calculated by

$$C_1\left(\widehat{\mathcal{F}}^{-1}\right) = \frac{s}{2}\log\left(\frac{tr\widehat{\mathcal{F}}^{-1}}{s}\right) - \frac{1}{2}\log|\widehat{\mathcal{F}}^{-1}|. \tag{6.17}$$

Here $s = rank(\mathcal{F}^{-1}) = mp + \frac{1}{2}m(m+1)$.

$$tr\widehat{\mathcal{F}}^{-1} = (tr\widehat{\mathbf{\Sigma}}_{RMD})\left(tr(\mathbf{X}'\mathbf{X})^{-1}\right) + \frac{1}{2n}\left(tr\widehat{\mathbf{\Sigma}}_{RMD}^2 + (tr\widehat{\mathbf{\Sigma}}_{RMD})^2 + 2\sum_{j=1}^{m}\widehat{\sigma}_{jj}^2\right) \tag{6.18}$$

and

$$|\widehat{\mathcal{F}}^{-1}| = 2^m n^{-\frac{1}{2}m(m+1)} |\widehat{\mathbf{\Sigma}}_{RMD}|^{m+p+1} |\mathbf{X}'\mathbf{X}|^{-m}, \tag{6.19}$$

where $\widehat{\sigma}_{jj}^2$ is the j^{th} diagonal element of $\widehat{\mathbf{\Sigma}}_{RMD}$.

We also introduce the robustness to two other forms of Bozdogan's information criteria based on the estimators of posterior expected utility (PEU) (Bozdogan and Haughton, 1998; Bozdogan, 2007).

$$\text{RICOMP(IFIM)}_{\text{PEU}} = nm\log 2\pi + n\log\left|\widehat{\mathbf{\Sigma}}_{RMD}\right| + nm + s + 2C_1(\widehat{\mathcal{F}}^{-1}) \tag{6.20}$$

$$\text{RICOMP(IFIM)}_{\text{LN\_PEU}} = nm\log 2\pi + n\log\left|\widehat{\mathbf{\Sigma}}_{RMD}\right| + nm + s + \log(n)C_1(\widehat{\mathcal{F}}^{-1}), \tag{6.21}$$

where $s = mp + \frac{1}{2}m(m+1)$ is the number of parameters estimated in the model.

We can see that $\text{RICOMP(IFIM)}_{\text{PEU}}$ and $\text{RICOMP(IFIM)}_{\text{LN\_PEU}}$ in equation 6.20 and 6.21 can be used for more complex model subset selection than the regular $\text{RICOMP(IFIM)}_{\text{MVR}}$ in equation 6.15 since they have more stringent penalty terms.

6.4 Robust and Misspecification-Resistant ICOMP for MVR Model Selection

In equation 6.1, define the standardized $\mathbf{Y}$ as $\mathbf{V} = (\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}^{-1/2}$, where $\mathbf{\Sigma} = \frac{1}{n}\mathbf{E}'\mathbf{E}$, so that $E(\mathbf{V}) = 0$ and $var(vec\mathbf{V}) = I_{mn}$. The matrix generalization of skewness $\mathbf{\Gamma}_1$ is given

by

$$\mathbf{\Gamma}_1 = E(vec\mathbf{V})\left(vec(\mathbf{V}'\mathbf{V} - nI_m)\right)', \tag{6.22}$$

and the matrix generalization of kurtosis $\mathbf{\Gamma}_2$ is given by

$$\mathbf{\Gamma}_2 = E(vec\mathbf{V}'\mathbf{V})(vec\mathbf{V}'\mathbf{V})'. \tag{6.23}$$

The inner product form of the information matrix $\mathcal{F}$ is given by

$$\mathcal{F} = \begin{pmatrix} \mathbf{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2} D_m^{+\prime}(\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1}) D_m^{+} \end{pmatrix}, \tag{6.24}$$

where D_m^+ is the Moore-Penrose inverse of the duplication matrix D_m.

The outer product form of the information matrix $\mathcal{R}$ is given by

$$\mathcal{R} = \begin{pmatrix} \mathbf{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X} & \frac{1}{2}(\mathbf{\Sigma}^{-1/2} \otimes X')\mathbf{\Gamma}_1 D_m^{+\prime}\mathbf{\Delta} \\ \frac{1}{2}\mathbf{\Delta} D_m^{+}\mathbf{\Gamma}_1'(\mathbf{\Sigma}^{-1/2} \otimes X) & \frac{1}{4}\mathbf{\Delta} D_m^{+}\mathbf{\Gamma}_2^{*} D_m^{+\prime}\mathbf{\Delta} \end{pmatrix}. \tag{6.25}$$

In equation 6.25, $\mathbf{\Delta} = D_m'\left(\mathbf{\Sigma}^{-1/2} \otimes \mathbf{\Sigma}^{-1/2}\right) D_m$ and $\mathbf{\Gamma}_2^* = \mathbf{\Gamma}_2 - n^2(vecI_m)(vecI_m)'$.

When the model is correctly specified, Γ_1 is reduced to 0 and Γ_2^* is reduced to $2nN_m$, consequently, $\mathcal{R} = \mathcal{F}$.

When the model is misspecified, the variance of the quasi-maximum likelihood estimator of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$, can be consistently approximated by $\mathcal{V} = \mathcal{F}^{-1}\mathcal{R}\mathcal{F}^{-1}$ (Gouriéroux, 1995a,b; Hendry, 1995; White, 1996)(see (Howe and Bozdogan, 2007; Magnus, 2007)), which is given by

$$\mathcal{V} = \begin{pmatrix} \mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1} & \frac{1}{n}\left(\mathbf{\Sigma}^{1/2} \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{\Gamma}_1 D_p \mathbf{\Delta}^{-1} \\ \frac{1}{n}\mathbf{\Delta}^{-1} D_p'\mathbf{\Gamma}_1'\left(\mathbf{\Sigma}^{1/2} \otimes \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) & \frac{1}{n^2}\mathbf{\Delta}^{-1} D_p'\mathbf{\Gamma}_2^{*} D_p \mathbf{\Delta}^{-1} \end{pmatrix}. \tag{6.26}$$

The robust and misspecification resistant ICOMP(IFIM), $\text{RICOMP(IFIM)}_{\text{misspec}}$ is given by

$$\text{RICOMP(IFIM)}_{\text{misspec}} = nm\log(2\pi) + n\log|\widehat{\Sigma}_{RMD}| + nm + 2C_1\left(\widehat{\mathcal{V}}\right), \tag{6.27}$$

where $\widehat{\Sigma}_{RMD}$ is the covariance matrix estimated by Mahalanobis distance method.

$$C_1(\widehat{\mathcal{V}}) = \frac{s}{2}\log\left(\frac{tr\widehat{\mathcal{V}}}{s}\right) - \frac{1}{2}\log|\widehat{\mathcal{V}}|, \tag{6.28}$$

and where $s = rank(\widehat{\mathcal{V}})$.

In equation 6.28,

$$tr\widehat{\mathcal{V}} = tr(\widehat{\mathbf{\Sigma}}_{RMD})tr\left((\mathbf{X}'\mathbf{X})^{-1}\right) + \frac{1}{n^2}trD_m^+\left(\widehat{\mathbf{\Sigma}}_{RMD}^{1/2}\otimes\widehat{\mathbf{\Sigma}}_{RMD}^{1/2}\right)\widehat{\mathbf{\Gamma}}_2^*\left(\widehat{\mathbf{\Sigma}}_{RMD}^{1/2}\otimes\widehat{\mathbf{\Sigma}}_{RMD}^{1/2}\right)D_m^{+\prime} \tag{6.29}$$

and

$$\begin{aligned}\left|\widehat{\mathcal{V}}\right| &= 2^{-m(m-1)}n^{-m(m+1)}\left|\widehat{\mathbf{\Sigma}}_{RMD}\right|^{m+p+1}\left|\mathbf{X}'\mathbf{X}\right|^{-m} \\ &\quad \left|D_m'\left(\widehat{\mathbf{\Gamma}}_2^* - \widehat{\mathbf{\Gamma}}_1'(I_m\otimes\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\widehat{\mathbf{\Gamma}}_1\right)D_m\right|. \end{aligned}\tag{6.30}$$

Other robust and misspecification resistant versions of ICOMP(IFIM) are given by

$$\begin{aligned}\text{RICOMP(IFIM)}_{misspec\_PEU} &= -2\log L(\widehat{\boldsymbol{\theta}}) + tr(\widehat{\mathcal{F}}^{-1}\widehat{\mathcal{R}}) + 2C_1\left(\widehat{\mathcal{F}}^{-1}\right) \\ &= nm\log(2\pi) + n\log|\widehat{\Sigma}_{RMD}| + nm + tr(\widehat{\mathcal{F}}^{-1}\widehat{\mathcal{R}}) \\ &\quad +2C_1\left(\widehat{\mathcal{F}}^{-1}\right), \end{aligned}\tag{6.31}$$

and

$$\begin{aligned}\text{RICOMP(IFIM)}_{misspec\_LN\_PEU} &= -2\log L(\widehat{\boldsymbol{\theta}}) + tr(\widehat{\mathcal{F}}^{-1}\widehat{\mathcal{R}}) + \log(n)C_1\left(\widehat{\mathcal{F}}^{-1}\right) \\ &= nm\log(2\pi) + n\log|\widehat{\Sigma}_{RMD}| + nm + tr(\widehat{\mathcal{F}}^{-1}\widehat{\mathcal{R}}) \\ &\quad +\log(n)C_1\left(\widehat{\mathcal{F}}^{-1}\right). \end{aligned}\tag{6.32}$$

6.5 Robust MVR Model Selection Algorithm

For the all possible subset selection of the data with p-dimensional predictor variables, we have $2^p - 1$ models to evaluate. Here are the steps to select the best model among the competing candidate models:

1. For a certain subset model, select one of the M-estimation methods described in Section 2.1.

2. Estimate the model covariance matrix and regression coefficients using the iterative Mahalanobis distance method described in Section 6.1.

3. Compute the RICOMP(IFIM) or $\text{RICOMP(IFIM)}_{\text{misspec}}$ for the model depending on whether the model is misspecified.

4. Repeat steps 1-3 for all possible subset models. Sort the RICOMP(IFIM) or $\text{RICOMP(IFIM)}_{\text{misspec}}$ values from the smallest to the largest and choose the best subset model with the minimum RICOMP(IFIM) (or $\text{RICOMP(IFIM)}_{\text{misspec}}$) value.

The GA model subset selection should follow the steps described in Section 4.2, where the fitness function is RICOMP(IFIM) or RICOMP(IFIM)$_{\text{misspec}}$.

The numerical examples are given in the next section to demonstrate the effectiveness of the new proposed information criteria.

6.6 Numerical Examples

6.6.1 Simulation Data Example

This Monte Carlo protocol follows closely to that used in Bozdogan and Haughton (1998) and is generalized to the multivariate linear regression model.

Let z_1, z_2, z_3 and z_4 be independent random variables following standard normal distribution $N(0,1)$. The first three predictor variables x_1, x_2, x_3 are simulated using

$$x_i = \sqrt{1-\alpha^2}z_i + \alpha z_4 \qquad \text{for} \quad i = 1, 2, 3, \tag{6.33}$$

where $\alpha \in [0, 1]$.

It is obvious that the variance of $x_i, i = 1, \ldots, 3$ is 1; the covariance of x_i and x_j is α^2, if $i \neq j, i, j = 1, 2, 3$. Thus, we can control the degree of multicollinearity among the predictor variables by assigning different values on α. In our simulation study, we use $\alpha^2 = 0.5$.

Seven redundant variables $x_4, \ldots, x_{10}$ are generated using the uniform random numbers, which are given by

$$x_4 = 4 * rand(0,1), \cdots, x_{10} = 10 * rand(0,1) \tag{6.34}$$

where rand(0,1) generates the standard Uniform random numbers.

Let $\mathbf{X} = \{x_1, x_2, x_3\}$. Suppose λ_{max} is the largest eigenvalue of the covariance matrix of $\mathbf{X}$ with β_{max} as the eigenvector corresponding to λ_{max}; λ_{min} is the smallest eigenvalue of the covariance matrix of $\mathbf{X}$ with β_{min} as the eigenvector corresponding to λ_{min}. Let

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{max} \\ \beta_{min} \end{bmatrix}$$

The response variable $\mathbf{Y}$ is generated by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E} \tag{6.35}$$

To illustrate the performance of the robustness, we introduced five 2-dimensional outliers in the response variable $\mathbf{Y}$. The observations assigned outliers are selected randomly and given at the same position for the two response variables:

y(19,1) = -12; y(19,2) = 16;
y(40,1) = 10; y(40,2) = 20;
y(53,1) = 10; y(53,2) = -15;
y(62,1) = 18; y(62,2) = 20;
y(96,1) = 15; y(96,2) = 16;

where $y(i,j)(i = 1, \ldots, n, j = 1, 2)$ is the i^{th} observation in the j^{th} response vector.

For this simulation, we assign different error distributions for $\mathbf{E}$ to demonstrate the performance of model subset selection under the model misspecification.

For convenience in showing our model selection results, we define the "full model" (M_f), "true model" (M_t), "other correct model" (M_{oc}), "overfitting model" (M_{of}), "redundant model" (M_r) and "wrong model" (M_w) as follows. The "full model" is the model containing all the predictor variables, which is denoted by $M_f = \{x_1, x_2, \ldots, x_{10}\}$. The "true model" is the one including the first three predictor variables, which is denoted by $M_t = \{x_1, x_2, x_3\}$. We define the "other correct model" M_{oc} as any non-empty strict subset of the true model, which is denoted by $M_{oc} \subset M_t$. We define the "overfitting model" M_{of} as the one containing both the true model and any redundant variable(s). In other words, the true model is a strict subset of the overfitting model, $M_t \subset M_{of}$. The "redundant model" M_r, is defined as the model including both the other correct model and any redundant variable(s). By this definition, $M_{oc} \subset M_r$. Finally, the "wrong model" M_w is defined as any model containing redundant variable(s) only.

Parameter Estimation on the True Model

In this part, we estimate the MVR coefficients and their biases for the true model given in equation 6.35 for one simulation with sample size $n = 200$. The MLE method (with and without outliers) and RMD method based on the four robust functions are used for the estimation. The error distribution is assumed standard normal so that we can compare our RMD estimates with the two MLEs. The 2D plot of the response variables in this simulation is shown in Figure 6.1. The QQ-plot and histogram for each of the response variables are shown in Figure 6.2. Both of them show that we have multivariate outliers and the distribution of the response variables is heavy tailed.

The estimation results are summarized in Table 6.1. The true generated regression coefficients and all estimates are shown in the table for comparison purposes along with the biases. The tuning constants used by each robust estimator are also given in the table. We can see that the MLE_2 (MLE without outliers) and robust estimates produce more accurate prediction for the true coefficients in terms of smaller biases than that of MLE with all observations. In addition, both the MLE_2 and robust estimates reduced the determinant of covariance matrix compare with the MLE. Among them, Tukey's robust

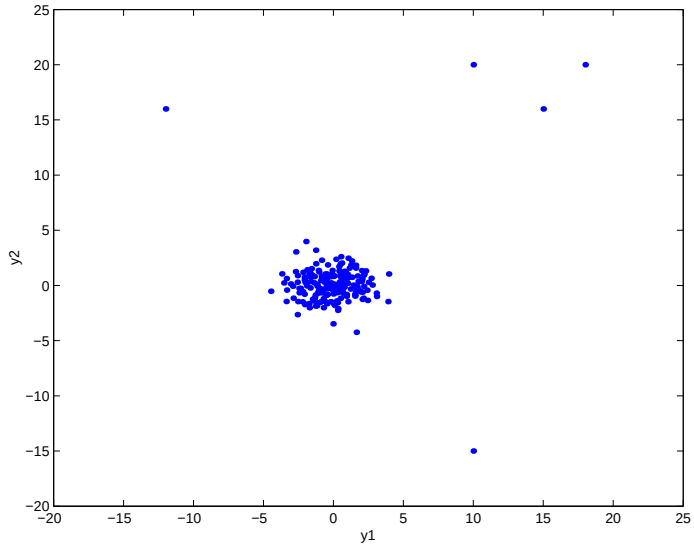

Figure 6.1: Multivariate Simulation: 2D plot of the Response for one simulation.

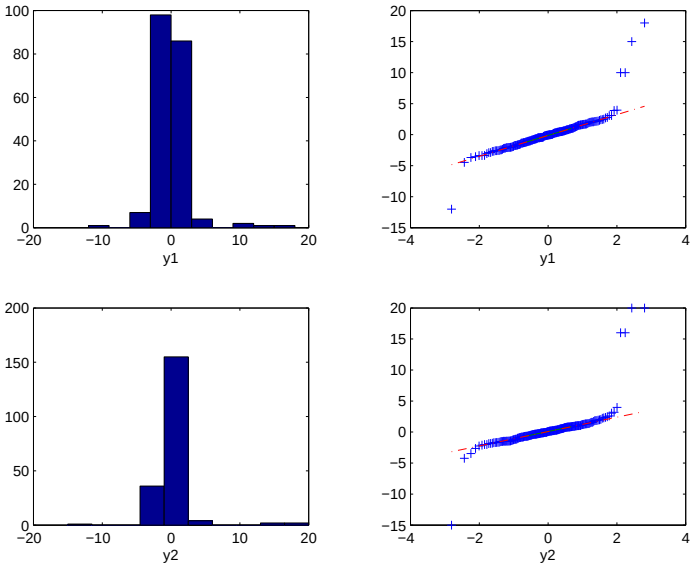

Figure 6.2: Multivariate Simulation: QQ Plot and Histogram of the Response for One Simulation.

Table 6.1: Multivariate Simulation: Parameter Estimates

	True		MLE		MLE_2[a]			
$\{\widehat{\mathbf{B}}_1, \widehat{\mathbf{B}}_2\}$	0.5382	0.6117	0.6186	0.4175	0.4875	0.5706		
	0.5834	0.2261	0.4199	-0.1506	0.599	0.1669		
	0.6083	-0.7581	0.414	-0.35	0.4739	-0.6265		
bias			-0.0804	0.1942	0.0507	0.0411		
			0.1635	0.3767	-0.0155	0.0592		
			0.1943	-0.4081	0.1344	-0.1316		
$\lvert\mathbf{\Sigma}\rvert$			44.6778		1.1487			
	Huber		Andrews		Tukey		Hampel	
Tuning Constant	k = 2		c = 2.1		c = 6.0		a = 1.7 b = 3.4 c = 8.5	
$\{\widehat{\mathbf{B}}_1, \widehat{\mathbf{B}}_2\}$	0.4915	0.5626	0.4869	0.5627	0.4867	0.559	0.486	0.5638
	0.596	0.1572	0.6013	0.1623	0.6024	0.16	0.6046	0.1611
	0.4667	-0.6144	0.4706	-0.615	0.4689	-0.6094	0.4665	-0.6172
bias	0.0467	0.049	0.0513	0.049	0.0515	0.0527	0.0522	0.0478
	-0.0125	0.0689	-0.0179	0.0639	-0.019	0.0661	-0.0212	0.065
	0.1415	-0.1437	0.1377	-0.1431	0.1393	-0.1487	0.1418	-0.1409
$\lvert\mathbf{\Sigma}\rvert$	1.4866		1.0102		**0.9739**		1.0306	

[a]MLE with outliers 19, 40, 53, 62 and 96 deleted

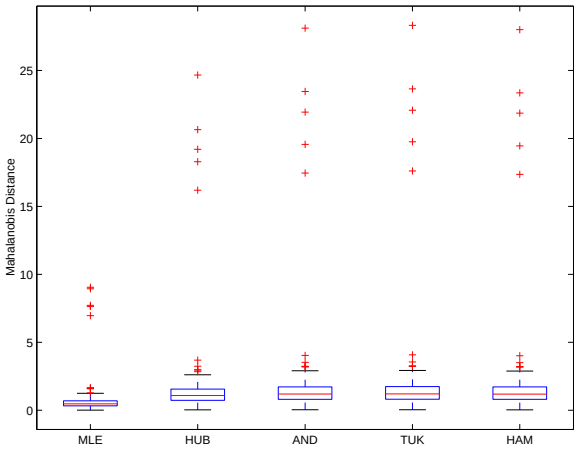

Figure 6.3: Multivariate Simulation: Boxplot of the Mahalanobis Distance of the Residuals.

estimator gives the smallest determinant of the covariance matrix. Therefore, we will use one of the misspecification resistant versions of ICOMP combined with Tukey's RMD estimator to perform the model subset selection in the following section.

The box plots of Mahalanobis distance of the residuals for all the estimators are shown in Figure 6.3. The average Mahalanobis distances of the residuals from robust estimates are higher than those from the MLE. Both robust estimates and MLE can identify all the five outliers.

All Possible Subset Selection

10 simulation data sets are generated with the error term in equation 6.1 from 10 different multivariate distributions. The sample size for each data set is $n = 200$ and the replication is 100. All possible subset selection is done on each of the 10 simulation data sets using the information criteria ROBAIC, ROBBIC, RICOMP(IFIM)$_{LN\_PEU}$ and RICOMP(IFIM)$_{misspec\_LN\_PEU}$. Each of the robust ICOMP values can be calculated based on either the MLE or any four of the robust functions (Huber's, Andrews', Tukey's and Hampel's). Thus, all possible subset selection is performed and compared using 12 information criteria for each of the 10 simulations. The same tuning constants are used for the robust functions as those in the parameter estimation section. The true model $\{x_1, x_2, x_3\}$ is desired to be selected. In the 100 replications of each simulation data, we count how many times each information criterion picks up the true model, other correct model, overfitting model, redundant model and wrong model. The all possible subset selection results are given in Tables 6.2 and 6.3.

Ten different distributions for the error term $\mathbf{E}$ in equation 6.1 are used in this simulation study to represent the multivariate normal distribution and the deviation from normality. The ten error distributions include: multivariate normal (MVN) distribution with non-correlated and correlated error (denoted as MVN1 and MVN2 in the table), multivariate t (MVT) distribution with 3 degrees of freedom with non-correlated and correlated errors (denoted as MVT1 and MVT2 in the table), contaminated multivariate normal distribution with multivariate t distribution (denoted as $0.5\text{MVN1} + 0.5\text{MVT1}$ and $0.5\text{MVN2} + 0.5\text{MVT2}$ in the table), multivariate power exponential (MPE) distribution with kurtosis parameter $\beta = 0.85$ with non-correlated and correlated error (denoted as MPE1 and MPE2 in the table), and multivariate power exponential distribution with kurtosis parameter $\beta = 2$ with non-correlated and correlated error (denoted as MPE3 and MPE4 in the table). For the details of the MPE distribution, refer to the Appendix B.

Table 6.2: Multivariate Simulation: All Possible Subset Model Selection in 100 runs(1)

Error Distribution	Model Category	RICOMP(IFIM)$_{\text{LN\_PEU}}$					ROBAIC	ROBBIC
		MLE	Huber	Andrews	Tukey	Hampel		
MVN1	M_t[a]	65	96	92	92	91	57	98
	M_{oc}[b]	35	0	0	0	0	0	0
	M_{of}[c]	0	4	8	8	9	43	2
MVN2	M_t	68	99	97	97	98	59	100
	M_{oc}	32	0	0	0	0	0	0
	M_{of}	0	1	3	3	2	41	0
MVT1	M_t	37	100	99	99	99	56	95
	M_{oc}	63	0	0	0	0	0	5
	M_{of}	0	0	1	1	1	44	0
MVT2	M_t	36	99	98	98	98	56	92
	M_{oc}	64	0	0	0	0	0	5
	M_{of}	0	1	2	2	2	44	3
0.5MVN1+ 0.5 MVT1	M_t	60	100	99	98	98	66	100
	M_{oc}	40	0	0	0	0	0	0
	M_{of}	0	0	1	2	2	34	0
0.5MVN2+ 0.5 MVT2	M_t	62	100	98	97	98	68	100
	M_{oc}	38	0	0	0	0	0	0
	M_{of}	0	0	2	3	2	32	0
MPE1	M_t	60	99	99	98	97	51	99
	M_{oc}	40	0	0	0	0	0	0
	M_{of}	0	1	1	2	3	49	1
MPE2	M_t	56	97	97	97	97	54	95
	M_{oc}	44	0	0	0	0	0	2
	M_{of}	0	3	3	3	3	46	3
MPE3	M_t	66	99	96	96	97	78	99
	M_{oc}	34	0	0	0	0	0	0
	M_{of}	0	1	4	4	3	22	1
MPE4	M_t	74	99	98	98	98	74	100
	M_{oc}	26	0	0	0	0	0	0
	M_{of}	0	1	2	2	2	26	0

Table 6.3: Multivariate Simulation: All Possible Subset Model Selection in 100 runs(2)

Error Distribution	Model Category	RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$				
		OLS	Huber	Andrews	Tukey	Hampel
MVN1	M_t	65	96	92	92	91
	M_{oc}	35	0	0	0	0
	M_{of}	0	4	8	8	9
MVN2	M_t	68	99	97	97	98
	M_{oc}	32	0	0	0	0
	M_{of}	0	1	3	3	2
MVT1	M_t	37	100	99	99	99
	M_{oc}	63	0	0	0	0
	M_{of}	0	0	1	1	1
MVT2	M_t	36	99	98	98	98
	M_{oc}	64	0	0	0	0
	M_{of}	0	1	2	2	2
0.5MVN1+ 0.5 MVT1	M_t	60	100	99	98	98
	M_{oc}	40	0	0	0	0
	M_{of}	0	0	1	2	2
0.5MVN2+ 0.5 MVT2	M_t	62	100	98	97	98
	M_{oc}	38	0	0	0	0
	M_{of}	0	0	2	3	2
MPE1	M_t	60	99	99	98	97
	M_{oc}	40	0	0	0	0
	M_{of}	0	1	1	2	3
MPE2	M_t	56	97	97	97	97
	M_{oc}	44	0	0	0	0
	M_{of}	0	3	3	3	3
MPE3	M_t	66	99	96	96	97
	M_{oc}	34	0	0	0	0
	M_{of}	0	1	4	4	3
MPE4	M_t	74	99	98	98	98
	M_{oc}	26	0	0	0	0
	M_{of}	0	1	2	2	2

Table 6.2 shows the all possible subset selection results using the ROBAIC, ROBBIC and RICOMP(IFIM)$_{\text{LN\_PEU}}$ based on MLE and four robust estimators. Table 6.3 shows the all possible subset selection results using the RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$ based on MLE and four robust estimators for the same 100 replications. Comparing the results from both tables, we see that the robust version of information criteria outperforms the non-robust version of information criteria (those computed on MLE). Specifically, RICOMP(IFIM)$_{\text{LN\_PEU}}$ and RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$ based on all four robust estimates and ROBBIC outperform ROBAIC and ICOMP computed on MLE. Because of the large sample size ($n = 200$), RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$ asymptotically approaches to RICOMP(IFIM)$_{\text{LN\_PEU}}$. Thus, they give the same model subset selection results. It is worth noticing that when the error distribution is MVN, ROBBIC picks up the true model more often(98% and 100% of the time) than the other robust information criteria although it is significantly better than the robust ICOMP. When the error distribution is MVT, RICOMP(IFIM)$_{\text{LN\_PEU}}$ and RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$ pick up the true model more often than the other criteria. When the error distribution is from contaminated MVN and MVT, both robust ICOMP and ROBBIC performs the best. When the error distribution is from MPE distribution, the robust ICOMP and ROBBIC outperforms the others. ROBAIC does not work well. It tends to overfit the model. Non-robust ICOMP based on MLE does not work well either since it tends to underfit the model. However, when the error distributions deviate from normality, ROBAIC outperforms ICOMP based on MLE. In other words, the information criterion based on MLE is not resistant to the departure from the normality assumption.

The box plots of the information criteria for the true models are given in Figure 6.4. We just show two of the ten error distributions as examples. In the plot, the MLE1 and MLE2 represent the RICOMP(IFIM)$_{\text{LN\_PEU}}$ and the RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$ values computed on MLE; HUB1 and HUB2 represent the RICOMP(IFIM)$_{\text{LN\_PEU}}$ and the RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$ values based on Huber's estimator; AND1 and AND2 represent the RICOMP(IFIM)$_{\text{LN\_PEU}}$ and the RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$ values based on Andrews' estimator; TUK1 and TUK2 represent the RICOMP(IFIM)$_{\text{LN\_PEU}}$ and the RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$ values based on Tukey's estimator; HAM1 and HAM2 represent the RICOMP(IFIM)$_{\text{LN\_PEU}}$ and the RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$ values based on Hampel's estimator. RAIC and RBIC represent ROBAIC and ROBBIC respectively. We can see that the values of information criteria based on the robust estimates are lower than those based on MLE.

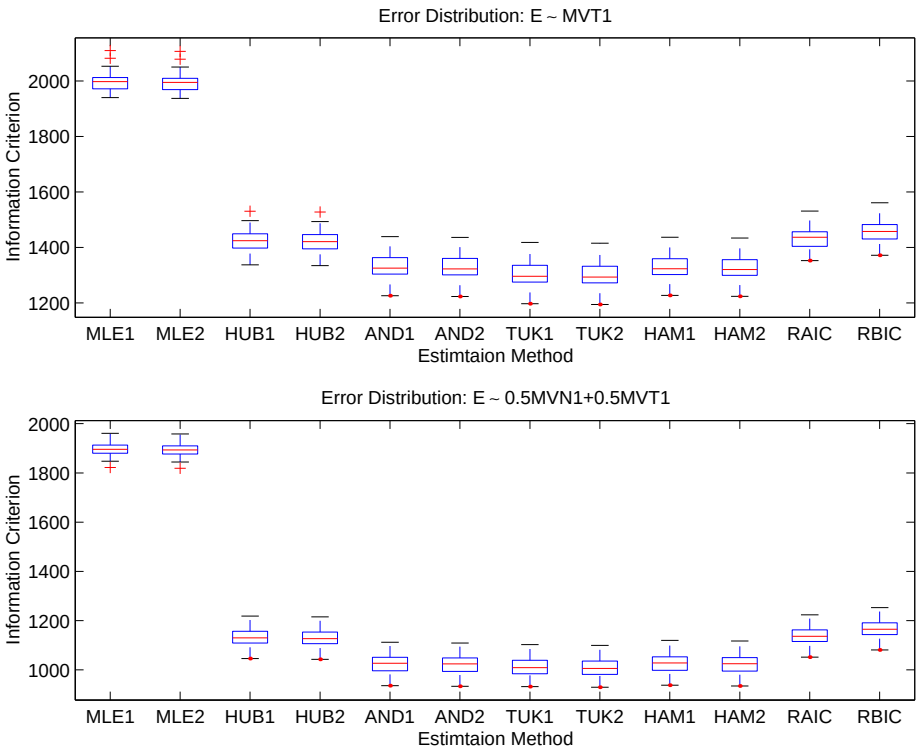

Figure 6.4: Multivariate Simulation: Boxplot for Information Criteria.

Table 6.4: GUI Inputs of GA Parameters for Multivariate Simulated Data

No. of runs	15
No. of generations	60
Population size	30
Estimation method	Tukey's
Fitness value	RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$
Probability of crossover	0.5
Crossover Method	uniform
Probability of Mutation	0.01
Elitism	Yes

GA Subset Selection

In this section, we performed the model subset selection using Genetic Algorithm (GA). Our goal is to conduct the model subset selection via GA and compare the results with those of the all possible subset selection and see if the three-way hybrid method can fulfill the selection of the true model when the true model exists.

Generate the simulation data from the Monte Carlo protocol in equation 6.33 and 6.34 with sample size $n = 200$. Generate response variables $\mathbf{Y}$ from equation 6.35, where the error term $\mathbf{E}$ follows the multivariate t distributions with 3 degrees of freedom and error terms are non-correlated. The fitness function of GA to carry out the subset selection is the RICOMP(IFIM)$_{\text{misspec\_LN\_PEU}}$ based on Tukey's estimator with tuning constant $c = 6.0$. All the parameters used in GA are given in Table 6.4.

We ran GA for 15 times. For each run of the GA, a new simulated data set was generated and the subset model picked by GA was recorded. The results for the 15 runs of the GA are shown in Table 6.5. 14 out of the 15 runs picked up the true model $\{x_1, x_2, x_3\}$. One run out of the 15 runs overfitted the model with one redundant predictor (the 13th run). Recall that the all possible subset selection method with the same information criterion picked up the true model 99 times in the 100 runs. Therefore, this GA result is reasonable.

One run of the GA result is shown in Table 6.6. The optimization procedure for 60 generations of the GA is presented in detail. In this particular run of the GA, the robust and misspecification resistant ICOMP can pick up the true model as early as the 19^{th} generation and retain it until the last 60^{th} generation.

The 2D and 3D plots for this one run of GA are shown in Figures 6.5 and 6.6, respectively, to show this optimization process.

Table 6.5: Multivariate Simulation: Model Subset Selection in 15 Runs of the GA

Run	Variable Selected	Binary String	$\text{RICOMP}_{\text{misspec\_LN\_PEU}}$
1	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1285.6
2	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1283.7
3	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1195.2
4	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1277.6
5	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1297.1
6	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1240.9
7	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1333.8
8	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1311.2
9	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1315.9
10	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1274.1
11	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1315.3
12	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1225.6
13	$\{x_1, x_2, x_3, x_7\}$	1 1 1 0 0 0 1 0 0 0	1349.8
14	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1353.4
15	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1292.7

Table 6.6: Multivariate Simulation: Model Subset Selection in One Run of the GA

Generation	Variable Selected	Binary String	$\text{RICOMP}_{\text{misspec\_LN\_PEU}}$
1	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
2	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
3	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
4	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
5	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
6	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
7	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
8	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
9	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
10	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
11	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
12	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
13	$\{x_1, x_2, x_3, x_4, x_7\}$	1 1 1 1 0 0 1 0 0 0	1311.7
14	$\{x_1, x_2, x_3, x_4\}$	1 1 1 1 0 0 0 0 0 0	1301.3
15	$\{x_1, x_2, x_3, x_4\}$	1 1 1 1 0 0 0 0 0 0	1301.3
16	$\{x_1, x_2, x_3, x_4\}$	1 1 1 1 0 0 0 0 0 0	1301.3
17	$\{x_1, x_2, x_3, x_4\}$	1 1 1 1 0 0 0 0 0 0	1301.3
18	$\{x_1, x_2, x_3, x_4\}$	1 1 1 1 0 0 0 0 0 0	1301.3
19	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1292.7
20	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1292.7
⋮	⋮	⋮	⋮
60	$\{x_1, x_2, x_3\}$	1 1 1 0 0 0 0 0 0 0	1292.7

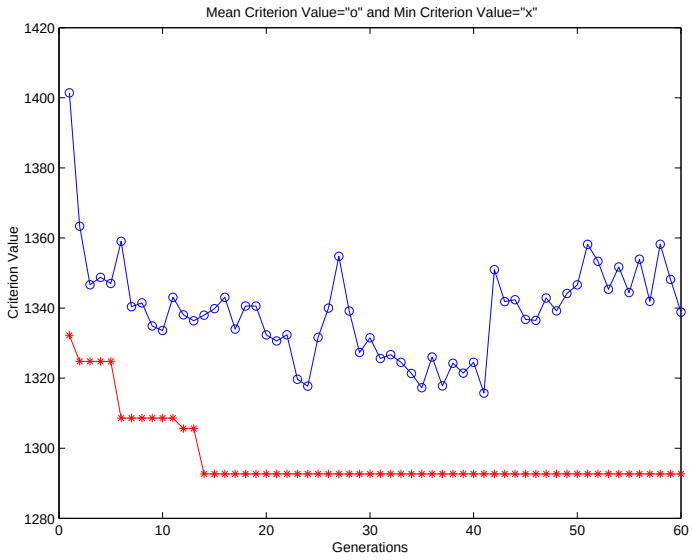

Figure 6.5: Multivariate Simulation: 2D plot for One Run of the GA.

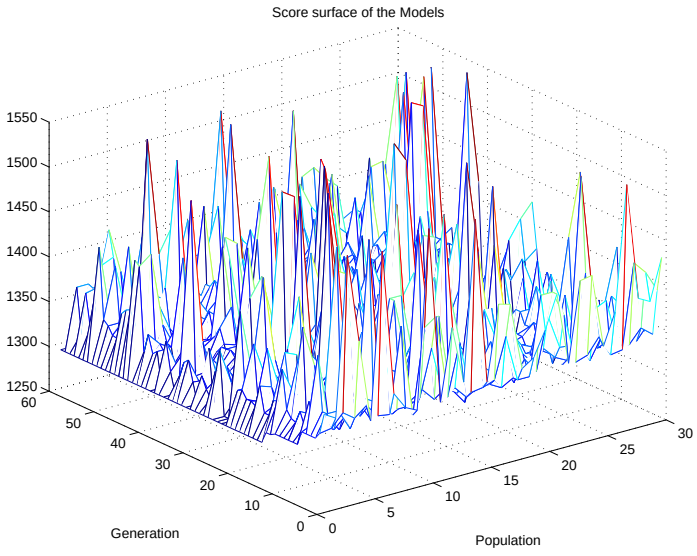

Figure 6.6: Multivariate Simulation: 3D plot for One Run of the GA.

Conclusion and Discussion

From this comparative study on the multivariate simulation data, we conclude that the robust and misspecification resistant version of ICOMP outperforms the other information criteria in the model subset selection, especially for those models with non-normal random errors. The RICOMP, $\text{RICOMP}_{\text{misspec}}$ and ROBBIC work better than ROBAIC and non-robust ICOMP computed on MLE.

GA combined with the robust and misspecification resistant version of ICOMP is a quick and effective model selection method. It can pick up the true model or optimal model very quickly in the early generation and retain it to the final result.

6.6.2 Real Data Example

Plasma-Retinol Data

Observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. Nieremberg et al. (1989) studied the determinant of plasma levels of these micronutrients. In an unpublished study, they collected data on 14 variables of 315 patients to investigate the relationship between the personal characteristics and dietary factors and plasma concentrations of retinol, beta-carotene and other carotenoids. Ben et al. (2006) used this data to illustrate the performance of his robust τ estimation on the multivariate linear regression. The data is available at $http://lib.stat.cmu.edu/datasets/Plasma\_Retinol$.

There are 315 observations (patients) in this data. The two dependent variables are:

Y_1 BETAPLASMA: Plasma beta-carotene (ng/ml)

Y_2 RETPLASMA: Plasma Retinol (ng/ml)

And the fourteen independent variables are:

X_1 AGE: Age (years);

X_2 SEX: Sex (1=Male, 2=Female);

X_3 SMOKSTAT1: Smoking status (1=Never);

X_4 SMOKSTAT2: Smoking status (1=Former);

X_5 QUETELET: Quetelet (weight/(height^2));

X_6 VITUSE1: Vitamin Use (1=Yes, fairly often);

X_7 VITUSE2: Vitamin Use (1Yes, not often);

X_8 CALORIES: Number of calories consumed per day;

X_9 FAT: Grams of fat consumed per day;

X_{10} FIBER: Grams of fiber consumed per day;

X_{11} ALCOHOL: Number of alcoholic drinks consumed per week;

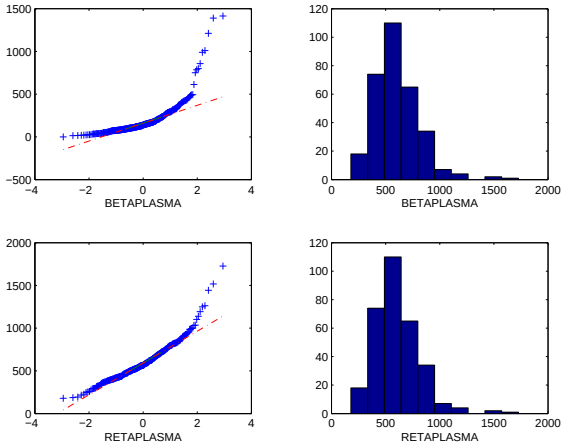

Figure 6.7: Plasma-Retinol Data: QQ-plot and Histogram of the Response Variables.

X_{12} CHOLESTEROL: Cholesterol consumed (mg per day);

X_{13} BETADIET: Dietary beta-carotene consumed (mcg per day);

X_{14} RETDIET: Dietary retinol consumed (mcg per day).

The QQ-plots and histograms for the two response variables are shown in Figure 6.7. Both response variables BETADIET and RETDIET are right skewed. The histograms for the nine continuous predictor variables are given in Figure 6.8. We can see from the plots that almost all of the continuous predictor variables are right skewed to some extent and contain outliers. The robust and misspecified version of ICOMP thus is appropriate for the model subset selection.

The correlation matrix among the response variables and predictor variables is given in Table 6.7. From the table, we see that the two response variables are not correlated to each other. None of the response variables has strong correlation with any of the predictor variables. Predictor x_8 (Calories), x_9 (Fat) and x_{12} (Cholesterol) are highly positively correlated to each other. x_8 (Calories) is mildly correlated to x_{10} (Fiber), x_{11} (Alcohol) and x_{14} (Retdiet). x_{13} (Betadiet) is mildly correlated to x_{10} (Fiber). The lack of correlational relation between the response variables and predictor variables may result from the complicated data structure.

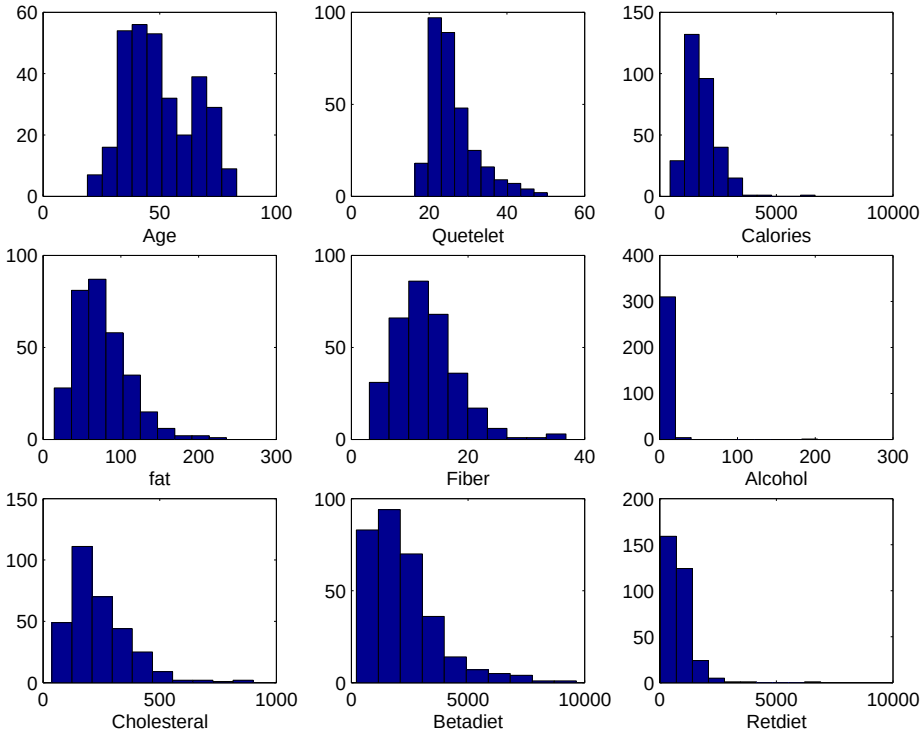

Figure 6.8: Plasma-Retinol Data: Histogram of the Continuous Predictor Variables.

Table 6.7: Plasma Data: Correlation Matrix

	y_1	y_2	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
y_1	1.00															
y_2	0.07	1.00														
x_1	0.10	0.21	1.00													
x_2	0.09	-0.18	-0.28	1.00												
x_3	0.09	-0.09	0.07	0.15	1.00											
x_4	0.01	0.15	0.03	-0.13	**-0.76**	1.00										
x_5	-0.23	0.01	-0.02	-0.01	0.09	-0.03	1.00									
x_6	0.22	0.04	0.09	0.06	0.13	-0.05	-0.09	1.00								
x_7	-0.01	-0.02	-0.17	0.13	-0.01	0.00	0.05	-0.47	1.00							
x_8	-0.02	-0.07	-0.18	-0.21	-0.12	0.07	0.00	0.00	0.02	1.00						
x_9	-0.09	-0.09	-0.17	-0.20	-0.14	0.09	0.05	-0.03	0.00	**0.87**	1.00					
x_{10}	0.24	-0.04	0.04	-0.05	0.07	0.04	-0.09	0.09	-0.01	0.47	0.28	1.00				
x_{11}	-0.02	0.02	0.05	-0.23	-0.13	0.03	-0.07	-0.09	-0.03	0.45	0.19	-0.02	1.00			
x_{12}	-0.13	-0.07	-0.11	-0.26	-0.11	0.05	0.11	-0.03	0.01	**0.66**	**0.71**	0.15	0.18	1.00		
x_{13}	0.22	-0.01	0.07	0.00	0.01	0.09	-0.01	0.10	-0.02	0.24	0.14	0.48	0.04	0.12	1.00	
x_{14}	-0.05	-0.06	-0.01	-0.07	0.01	0.01	0.03	0.02	-0.02	0.40	0.41	0.21	0.04	0.44	0.05	1.00

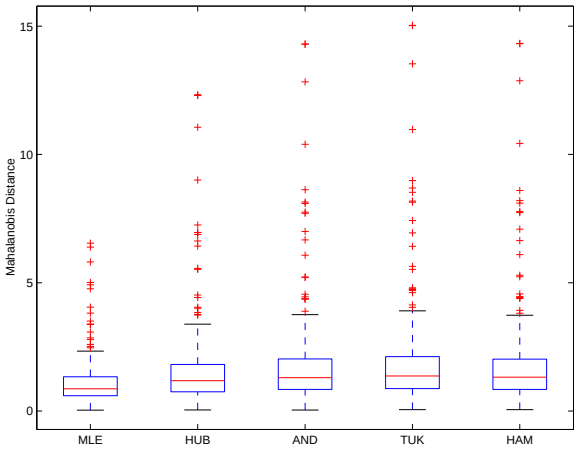

Figure 6.9: Plasma Data: Boxplot of the Mahalanobis Distance of the Residuals.

On the full model, we compute and compare two types of estimates of the regression coefficients: the MLE and multivariate robust estimates. We use the same tuning constants for the robust estimates as we did in the Monte Carlo simulation study. The side-by-side box-plot of the Mahalanobis distance of the residuals $d_i = \widehat{\varepsilon}_i' \widehat{\mathbf{\Sigma}}^{-1} \widehat{\varepsilon}_i$, where $1 \leq i \leq 315$ for both MLE and robust estimates are shown in Figure 6.9. We see that on average, the residuals from the robust estimates are larger than that from the MLE. More outliers are shown in the box plot of the robust estimates than that of the MLE. In other words, the robust estimates can identify more outliers than the MLE. In fact, if we define the outliers as those observations whose Mahalanobis of the residuals $d_i > \sqrt{\chi^2_{2,0.99}}$, the MLE reveals only 12 outliers; multivariate Huber's estimate can reveal 25 outliers; multivariate Andrews' estimate can reveal 31 outliers; multivariate Tukey's estimate can reveal 33 outliers and multivariate Hampel's estimate can reveal 31 outliers. Ben et al. (2006) in his paper claims that his τ estimate reveals 27 outliers, which is in-between the outliers our robust estimates can identify.

The determinant of the covariance matrix of the full model computed on each estimation method are given in Table 6.8. We see that the determinant of the covariance matrix calculated from robust estimates is smaller than that calculated from the MLE. Among all the robust estimates, the Tukey estimate gives the smallest determinant of the covariance. We decided to perform the model subset selection using $\text{RICOMP}_{\text{misspec}}$ based on multivariate Tukey's estimator.

The best subset model selected by $\text{RICOMP}_{\text{misspec}}$ with multivariate Tukey's estimator is given in Table 6.9, along with the subset models selected by multivariate robust BIC and multivariate τ-estimate given in Ben et al. (2006) for the purpose of comparison. We

Table 6.8: Plasma Data: Determinant of Covariance Matrix for the Full Model

	$\lvert\mathbf{\Sigma}\rvert$				
Full Model	MLE	Huber	Andrews	Tukey	Hampel
$\{x_1, \ldots, x_{14}\}$	1.0466×10^9	0.2012×10^9	0.13×10^9	0.1061×10^9	0.1276×10^9

Table 6.9: Plasma Data: Best Subset Model

Best Subset Model	Selection Method
$\{x_1, x_2, x_3, x_4, x_5, x_7, x_{10}\}$	$\text{RICOMP}_{\text{misspec}}$ with Tukey's estimator
$\{x_1, x_2, x_3, x_5, x_{13}\}$	ROBBIC
$\{x_1, x_2, x_5, x_6, x_{10}, x_{13}\}$	τ-estimator

should be aware that Ben et al. (2006) did not perform the "formal" all possible model subset selection in their paper. They simply fitted the full model and picked up all the variables "that are statistically significant at level 0.05 for at least one equation." We can see that all these three methods agree to choose x_1 and x_2 as the best predictors.

The top 16 subset models selected by $\text{RICOMP}_{\text{misspec}}$ with multivariate Tukey's estimator using all possible subset selection technique is given in Table 6.10.

The parameters used in GA are given in Table 6.11 and the subset selection results in 15 runs of GA are given in Table 6.12. For each run of the GA results, the corresponding ranking of all possible subset selection is given in the parenthesis in Table 6.12. We see that 12 out of 15 runs of the GA hit the top 10 models chosen by all possible subset selection. 14 out of 15 runs of the GA hit the top 15 models and 15 out of 15 runs of the GA hit the top 20 models. These results are optimal since we have $2^{14} - 1 = 16,383$ all possible subsets in total in the model space. These results show that GA is an effective way to perform model subset selection.

The 2D plot and 3D plot of GA show the optimization process are given in Figures 6.10 and 6.11.

Interpretation The final fitted subset model with $x_1, x_2, x_3, x_4, x_5, x_7$ and x_{10} is as follows

$$\begin{aligned} y_1 &= -8.13 + 1.62x_1 + 66.59x_2 + 57.89x_3 + 21.49x_4 - 4.51x_5 + 9.38x_7 + 3.14x_{10} \\ y_2 &= 413.58 + 2.77x_1 + 6.70x_2 + 43.16x_3 + 91.96x_4 - 0.48x_5 + 19.53x_7 - 3.07x_{10}. \end{aligned} \tag{6.36}$$

Table 6.10: Plasma Data: Top 16 Subsets by All Possible Model Selection

Ranking	Selected Variables	$\text{RICOMP}_{\text{misspec}}$ with Tukey
1	0 1 2 3 4 5 - 7 - - 10 - - - -	7755.8764
2	0 1 - 3 4 5 6 7 - - 10 - - - -	7758.1327
3	0 1 2 3 4 - 6 - - - 10 - - - -	7759.011
4	0 1 2 3 4 5 6 7 - - - - - - -	7762.8166
5	0 1 2 3 - 5 6 7 - - 10 - - - -	7763.0988
6	0 1 2 3 4 5 6 7 - 9 10 - - - -	7763.8668
7	0 - - 3 4 - 6 7 - - 10 - - - -	7764.4676
8	0 1 2 3 4 - - - - - 10 - - - -	7764.5677
9	0 1 - 3 4 - 6 7 - 9 10 - - - -	7764.9857
10	0 1 2 3 4 5 6 - - 9 10 - - - -	7765.4471
11	0 1 2 - 4 - 6 7 - - 10 - - - -	7766.6372
12	0 1 - 3 - 5 6 7 - - 10 - - - -	7766.6897
13	0 - 2 3 4 5 6 - - - 10 - - - -	7767.6668
14	0 1 - 3 4 - 6 - - 9 10 - - - -	7767.7774
15	0 1 2 3 4 5 - - - - 10 11 - - -	7768.4374
16	0 - 2 - 4 5 6 7 - - 10 - - - -	7768.6484

Table 6.11: Plasma Data: GUI Inputs of GA Parameters

No. of runs	15
No. of generations	120
Population size	50
Estimation method	Tukey's
Fitness value	$\text{RICOMP}_{\text{misspec}}$
Probability of crossover	0.5
Crossover Method	uniform
Probability of Mutation	0.01
Elitism	Yes

Table 6.12: Plasma Data: Model Subset Selection in 15 Runs of the GA

GA Ranking	Variables Selected	Binary String	Scores
1 (1)[a]	0 1 2 3 4 5 - 7 - - 10 - - - -	0 1 1 1 1 1 0 1 0 0 1 0 0 0 0	7755.9
2 (2)	0 1 - 3 4 5 6 7 - - 10 - - - -	0 1 0 1 1 1 1 1 0 0 1 0 0 0 0	7758.1
3 (16)	0 - 2 - 4 5 6 7 - - 10 - - - -	0 0 1 0 1 1 1 1 0 0 1 0 0 0 0	7768.6
4 (7)	0 - - 3 4 - 6 7 - - 10 - - - -	0 0 0 1 1 0 1 1 0 0 1 0 0 0 0	7764.5
5 (11)	0 1 2 - 4 - 6 7 - - 10 - - - -	0 1 1 0 1 0 1 1 0 0 1 0 0 0 0	7766.6
6 (15)	0 1 2 3 4 5 - - - - 10 11 - - -	0 1 1 1 1 1 0 0 0 0 1 1 0 0 0	7768.4
7 (6)	0 1 2 3 4 5 6 7 - 9 10 - - - -	0 1 1 1 1 1 1 1 0 1 1 0 0 0 0	7763.9
8 (7)	0 - - 3 4 - 6 7 - - 10 - - - -	0 0 0 1 1 0 1 1 0 0 1 0 0 0 0	7764.5
9 (6)	0 1 2 3 4 5 6 7 - 9 10 - - - -	0 1 1 1 1 1 1 1 0 1 1 0 0 0 0	7763.9
10 (3)	0 1 2 3 4 - 6 - - - 10 - - - -	0 1 1 1 1 0 1 0 0 0 1 0 0 0 0	7759
11 (3)	0 1 2 3 4 - 6 - - - 10 - - - -	0 1 1 1 1 0 1 0 0 0 1 0 0 0 0	7759
12 (9)	0 1 - 3 4 - 6 7 - 9 10 - - - -	0 1 0 1 1 0 1 1 0 1 1 0 0 0 0	7765
13 (1)	0 1 2 3 4 5 - 7 - - 10 - - - -	0 1 1 1 1 1 0 1 0 0 1 0 0 0 0	7755.9
14 (3)	0 1 2 3 4 - 6 - - - 10 - - - -	0 1 1 1 1 0 1 0 0 0 1 0 0 0 0	7759
15 (1)	0 1 2 3 4 5 - 7 - - 10 - - - -	0 1 1 1 1 1 0 1 0 0 1 0 0 0 0	7755.9

[a]The parenthesis includes the corresponding all possible model selection ranking for the purpose of comparison.

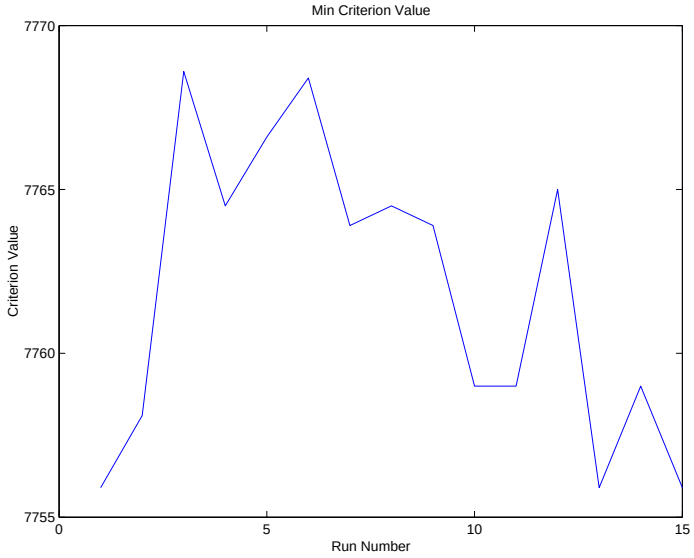

Figure 6.10: Plasma Data: 2D-plot of 15 Runs of the GA.

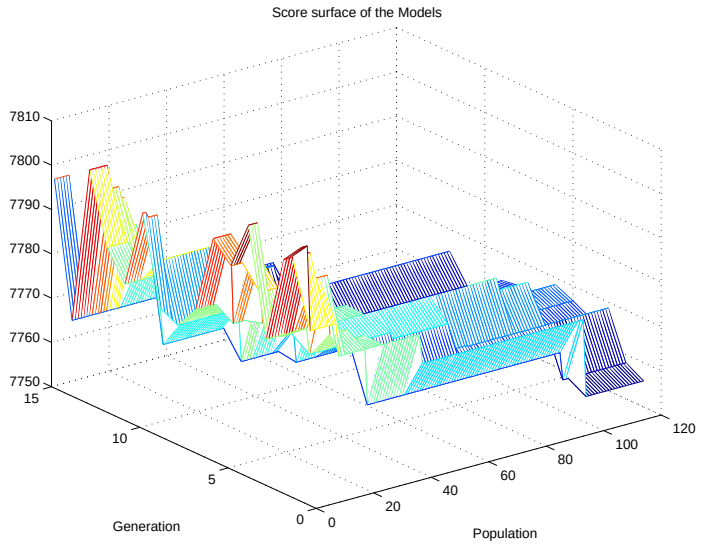

Figure 6.11: Plasma Data: 3D-plot of 15 Runs of the GA.

The contents of plasma beta-carotene (y_1) and plasma retinol (y_2) are affected by the following factors: age (x_1), sex (1=Male, 2=Female) (x_2), smoking status1 (1=never) (x_3), smoking status2 (1=Former) (x_4), QUETELET (weight/(height^2)) (x_5), vitamin use2 (1=Yes, not often) (x_7) and grams of fiber consumed per day (x_{10}).

Holding the other factors constant, both plasma contents increase with the increasing of patients' age. The female patients have higher contents of both plasma than the male patients. Current smokers reduce the contents of both plasma the most. The increasing of QUETELET index results in the decreasing of both plasma contents. There is no difference in plasma contents between the patients who do not use vitamin and those who use vitamin fairly often. However, if the patients use vitamin but not often, both of their plasma contents will increase. The more fiber is consumed by the patients per day, the more plasma beta-carotene is and the less plasma retinol is.

Chapter 7

Future Research

7.1 Conclusions

In this dissertation, we propose a novel computationally efficient model subset selection method for multiple and multivariate linear regression models which is both robust and misspecification resistant. Basically speaking, our approach is a three-way hybrid method which employs the information theoretic measure of complexity (ICOMP) computed on robust M-estimators as model subset selection criteria, and integrates genetic algorithms (GA) as the subset model searching engine.

For the first time in the literature, we introduce both robustness and misspecification resistance to the computation of information criteria, which makes the model subset selection procedure robust to unusual observations in the data and deviations from the assumption of normality.

For the first time in the literature, we introduce robust and misspecification resistant information criteria to genetic algorithms (GA) as its fitness function. The three-way hybrid approach is heuristically shown to be efficient.

For the multiple linear regression, we develop robust versions of Bozdogan's information-theoretic measure of complexity (ICOMP) and name them 'RICOMP' and '$\text{RICOMP}_{\text{misspec}}$' for the correctly specified model and misspecified model respectively. We compare the performance of RICOMP and $\text{RICOMP}_{\text{misspec}}$ in the Monte Carlo simulation data with other robust information criteria existing in the literature, namely AICR, HAIC and RBIC. The conclusion is that our robust versions of ICOMP are more efficient than the robust versions of AIC and BIC in terms of picking up the simulated true model. Particularly, when the error term in the simulation data departs from normal distribution with skewness and kurtosis, the robust and misspecification resistant version of ICOMP outperforms all the other information criteria. Of course, when we compare the robust information criteria

with the non-robust information criteria, or those computed based on OLS estimator, the robust version of information criteria performs much better. The non-robust information criteria are sensitive to the outliers and non-normality assumption of the error term.

For multivariate regression, we derive the two-stage iteratively robust Mahalanobis distance (RMD) estimator and then introduce this RMD estimator to the computation of information criteria. This is the first time in the multivariate regression literature, we combine both robustness and misspecification resistance to the information criteria. Firstly, we introduce this RMD estimator to AIC and BIC and get the robust version of AIC and BIC, called ROBAIC and ROBBIC. Secondly, we introduce this RMD estimator to Bozdogan's ICOMP and get the robust version of ICOMP in the multivariate framework. In the comparative study on the multivariate simulation data, the robust and misspecification resistant version of information criteria outperforms the non-robust version of information criteria.

In the simulation study conducted by the three-way hybrid method, GA combined with the robust and misspecification resistant information criteria is proved to be an effective model selection method. It can reach the optimal, if not the best, solution quickly without having to search the whole subset model space.

7.2 Future Research

A few follow up researches are suggested to be done in the future.

Firstly, more robust estimators other than robust M-estimator can be introduced to the computation of information criteria. There are some robust estimators in the literature with nice properties, such as the high breakdown point, which could be used if one can overcome the computational complexity of these estimators. However, the robust estimators for the multivariate literature is still the big challenge.

Secondly, the robust and misspecification resistant information criteria could be generalized to the robust SUR (Seemingly Unrelated Regression) model, which allows one to select different predictors for different responses. In this way, we can increase the flexibility of the model subset selection techniques.

Finally, the robust and misspecification resistant information criteria for model subset selection could be generalized to other areas of statistics, such as the logistic regression, discriminant analysis, cluster analysis and time series etc.

Bibliography

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *2nd International Symposium on Information Theory*, pages 267–281. Akademia Kiado, Budapest.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics*, 16:523–531.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton Univ. Press.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, 46:199–208.

Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185.

Behnke, A. and Wilmore, J. (1974). *Evaluation and Regulation of Body Build and Composition*. Prentice-Hall, Englewood Cliffs, N.J.

Ben, M. G., Martínes, E., and Yohai, V. (2006). Robust estimation for the multivariate linear model based on a τ-scale. *Journal of Multivariate Analysis*, 97:1600 – 1622.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, 37:51–58.

Berk, R. H. (1970). Consistency a posteriori. *Annals of Mathematical Statistics*, 41:894–906.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40:318–335.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont.

Boyce, D. E., Farhi, A., and Weischedel, R. (1974). *Optimal Subset Selection: Multiple Regression, Interdependence and Optimal Network Algorithms*. Springer-Verlag, Berlin.

Bozdogan (2007). *Information Complexity and Multivariate Learning in High Dimensions with Applications in Data Mining.* In Preparation.

Bozdogan, H. (1987a). Icomp: A new model selection criterion. In Bock, H. H., editor, *Classification and Related Methods of Data Analysis*. Elsevier Science, Amsterdam (North-Holland). Preprint paper.

Bozdogan, H. (1987b). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52:345–370.

Bozdogan, H. (1988a). Icomp: A new model-selection criterion. In Bock, H. H., editor, *Classification and Related Methods of Data Analysis*. Elsevier Science, Amsterdam (North-Holland).

Bozdogan, H. (1988b). The theory and applications of information-theoretic measure of complexity (icomp) as a new model selection criterion. Unpublished research report, the Institute of Statistical Mathematics, Tokyo, Japan, and the Department of Mathematics, University of Virginia, Charlottesville, VA.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics Theory and Methods*, 19(1):221–278.

Bozdogan, H. (1994a). Aic-replacements for some multivariate tests of homogeneity with applications in multisample clustering and variable selection. In Bozdogan, H. e., editor, *Proceedings of the first US/Japan conference on the Frontiers of Statistical Modeling: An Informational Approach*, volume 2, pages 199–232. Kluwer Academic Publishers, Dordrecht.

Bozdogan, H. (1994b). Engineering & scientific applications of informational modeling. In Bozdogan, H. e., editor, *Proceedings of the first US/Japan conference on the Frontiers of Statistical Modeling: An Informational Approach*, volume 3. Kluwer Academic Publishers, Dordrecht.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44:62–91.

Bozdogan, H. (2003). Robust and misspecification resistant model selection with information complexity and genetic algorithms. presentation, Istanbul, Turkey. Invited paper presented at Euro-Informs Conference in the session "Information Theoretic Methods".

Bozdogan, H. (2004a). Intelligent statistical data mining with information complexity and genetic algorithms. In Bozdogan, H., editor, *Statistical Data Mining and Knowledge Discovery*, pages 15–56. Chapman & Hall/CRC.

Bozdogan, H. (2004b). *Statistical Modeling and Model Evaluation: A New Informational Approach*. Chapman & Hall/CRC.

Bozdogan, H. and Bearse, P. M. (2003). Information complexity criteria for detecting influential observations in dynamic multivariate linear models using the genetic algorithm. *Journal of Statistical Planning and Inference*, 114:31–44.

Bozdogan, H. and Haughton, D. (1998). Information complexity criteria for regression models. *Computational Statistics & Data Analysis*, 28:51–76.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*. Wiley, New York, 2 edition.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer, NY.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, NY, 2 edition.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society*, 158:419–466.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

Daniel, C. and Wood, F. S. (1971). *Fitting Equations to Data*. Wiley, New York.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In Bickel, P.J., D. K. and Hodges, J. J., editors, *Festschrift fur Erich L. Lehmann (Wadsworth, Belmont, CA)*, pages 157–184.

Fernandez, C. and Steel, M. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93:359–371.

Godfrey, L. G. (1988). *Misspecification Tests in Econometrics*. Cambridge University Press. Econometric Society monographs No. 16.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, New York.

Gómez, E., Gómez-Villegas, M. A., and Marín, J. M. (1998). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics-Theory and Methods*, 27(3):589–600.

Gouriéroux, C. (1995a). *Statistics and Econometric Models*, volume 1. Cambridge University Press, Cambridge.

Gouriéroux, C. (1995b). *Statistics and Econometric Models*, volume 2. Cambridge University Press, Cambridge.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42:1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *The Journal of the American Statistical Association*, 69(346):383–393.

Hampel, F. R. (1983). Some aspects of model choice in robust statistics. In *Proceedings of the 44th Session of ISI*, pages 767–771, Madrid. Book 2.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics, The Approach Based on Influence Functions*. John Wiley & Sons, New York.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford University Press.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49. A Biometric Invited Paper.

Hocking, R. R. (1983). Developments in linear regression methodology: 1959-1982. *Technometrics*, 25(3):219–230.

Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestion for future applications and theory. *Journal of the American Satistical Association*, 69:909–927.

Hogg, R. V. (1979a). An introduction to robust estimation. In Launer, R. L. and Wilkinson, G. N., editors, *Robustness in Statistics*, pages 1–17. Academic Press, New York.

Hogg, R. V. (1979b). Statistical robustness: One view of its use in applications today. *The American Statistician*, 33:108–115.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics: Theory and Methods*, A6(9):813–827.

Howe, A. and Bozdogan, H. (2007). Misspecified multivariate regression models using genetic algorithm and information complexity. working paper.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *In Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*. University of California Press, Berkeley.

Huber, P. J. (1972). The 1972 wald lecture robust statistics: A review. *Annals of Mathematical Statistics*, 43:1041–1067.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures, and monte carlo. *Annals of Statistics*, 1:799–821.

Huber, P. J. (1977). *Robust Statistical Procedures*. Society of Industrial and Applied Mathematics, Philadelphia.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, NY.

Huber, P. J. (1996). *Robust Statistical Procedures*. Society of Industrial and Applied Mathematics, Philadelphia, 2 edition.

Huber, P. J. (2004). *Robust Statistics*. John Wiley & Sons, New Jersey.

Hurvich, C. M. and Tsai, C. L. (1990). Model selection for least absolute deviations regression in small samples. *Statistics & probability Letters*, 9:259–265.

Johnson, R. and Wichern, D. (1992). *Applied Multivariate Statistical Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 3 edition.

Jurečková, J. (1977). Asymptotic relations of m-estimates and r-estimates in linear regerssion models. *Annals of Statistics*, 5:464–472.

Koenker, R. and Basset, G. J. (1978). Regression quantiles. *Econometrica*, 36:33–50.

Koenker, R. and Portnoy, S. (1990). M estimation of multivariate regressions. *Journal of the American Statistical Association*, 85:1060–1068.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

Lin, C.-T. and Lee, C. S. G. (1996). *Neural Fuzzy Systems*. Prentice Hall, Upper Saddle River.

Linhart, H. and Zucchini, W. (1986). *Model Selection*. John Wiley & Sons, New York.

Liu, Y. (2004). Robust and misspecification resistant model selection with information complexity and genetic algorithms. presentation, Joint Statistical Meetings (JSM), Toronto, Canada. in section "Diagnostics and Robustness".

Machado, J. A. F. (1993). Robust model selection and m-estimation. *Econometric Theory*, (9):478–493.

Magnus, J. R. (2007). The asymptotic variance of the pseudo maximum likelihood estimator. Forthcoming in Econometric Theory, 23.

Magnus, J. R. and Neudecker, H. (1999). *Matrix Differential Calculus*. John Wiley & Sons, New York, 2 edition.

Mahfoud, S. W. (1994). Population sizing for sharing methods. Illinois Genetic Algorithms Laboratory Report 94005, University of Illinois, Champaign-Urbana, IL.

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics*, 12:621–625.

Maronna, R., Martin, R., and Yohai, V. (2006). *Robust Statistics, Theory and Method*. John Wiley & sons, Ltd, England.

Miller, A. (1990). *Subset Selection in Regression*. Chapman & Hall, London.

Miller, A. (2002). *Subset Selection in Regression*. Chapman & Hall, New York, 2 edition.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2001). *Introduction to Linear Regression Analysis*. John Wiley & Sons, New York, 3 edition.

Moses, L. E. (1986). *Think and Explain with Statistics*. Addison-Wesley, MA. Reading.

Nasraoui, O. (2004). A brief overview of robust statistics. Tutorial. URL: http://www.louisville.edu/∼o0nasr01/Websites/tutorials/RobustStatistics/.

Nierenberg, D., Stukel, T., Baron, J., Dain, B., and Greenberg, E. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, 130:511–521.

Olive, D. (2005). Applied robust statistics. URL: http://www.math.siu.edu/olive/olbookp.htm.

Penrose, K., Nelson, A., and Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques (abstract). *Medicine and Science in Sports and Exercise*, 17(2):189.

Qian, G. and Künsch, H. R. (1996). On model selection in robust linear regression. Research Report 80, ETH Zünch.

Qian, G. and Künsch, H. R. (1998). On model selection via stochastic complexity in robust linear regression. *Journal of Statistical Planning and Inference*, 75:91–116.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91.

Rao, C. R. (1947). Minimum variance and the estimation of several parameters. In *Proceedings of the Cambridge Philosophical Society*, volume 43, pages 280–283.

Rao, C. R. (1948). Sufficient statistics and minimum variance estimates. In *Proceedings of the Cambridge Philosophical Society*, volume 45, pages 213–218.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Co. Pte. Ltd., Singapore.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42:40–47.

Ronchetti, E. (1985). Robust model selection in regression. *Statistics & probability Letters*, 3:21–23.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica*, 7:327–338.

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 cp. *Journal of American Statistical Association*, 89:550–559.

Rousseeuw, P., Aelst, S., Driessen, K., and Agullo, J. (2004). Robust multivariate regression. *Technometrics*, 46:293–305.

Rousseeuw, P. J. (1983). Multivariate estimation with high breakdown point. Research Report 192, Centre for Statistics and Operations Research, VUB Brussels.

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*. John Wiley & Sons, NY. Wiley Series in Probability and Mathematical Statistics.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of s-estimators. In Franke, J., H. W. and D., M., editors, *in Robust and Nonlinear Time Series Analysis*, pages 256–272. Springer-Verlag, New York.

Ruppert, D. (1992). Computing s estimators for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics*, 1:253–270.

Sánchez Manzano, E. G., Gómez-Villegas, M. A., and Marín-Diazaraque, J.-M. (2002). A matrix variate generalization of the power exponential family of distributions. *Communications in Statistics-Theory and Methods*, 31(12):2167–2182.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52:333–343.

Shibata, R. (1989). Statistical aspects of model selection. In Williams, J., editor, *From data to Model*, pages 215–240. Springer-Verlag, New York.

Siri, W. (1956). *Advances in Biological and Medical Physics*, volume IV, chapter Gross Composition of the Body. Academic Press, New York.

Sokal, R. R. and Rohlf, F. J. (1981). *Biometry*. W.H. Freeman, San Francisco, 2 edition. Also found in: Hand, D. J., et al. (1994) A Handbook of Small Data Sets, London: Chapman and Hall, 20-21.

Subbotin, M. T. (1923). On the law of frequency of errors. *Matematicheskii Sbornik*, pages 296–300.

Sugiura, N. (1978). Further analysis of the data by akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods*, A7:13–26.

Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, 153:12–18. [In Japanese].

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33:1–67.

Tukey, J. W. (1972). How computing and statistics affect each other. In *In The Babbage Memorial Meeting: Report of Proceedings*, pages 21–37, London. The British Computer Society and The Roy al Statistical Society.

van Emden, M. H. (1971). *An Analysis of Complexity*, volume 35. Mathematical Centre Tracts, Amsterdam.

Welsch, R. E. (1975). *Confidence Regions for Robust Regression.* Statistical Computing Section Proceedings of the American Statistical Association, Washington, D. C.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–26.

White, H. (1996). *Estimation, Inference and Specification Analysis.* Cambridge University Press. Econometric Society Monographs No. 22.

Yohai, V. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15:642–656.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, 83(402):406–413.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association*, 57:348–368.

Appendix

Appendix A

Power Exponential Distribution

The Power Exponential (PE) distribution is a generalized error distribution by introducing a kurtosis parameter to the Normal distribution. It is first developed by Subbotin (1923) and popularized in the 1970's (Box and Tiao, 1973).

Suppose random variable x follows PE distribution with parameters μ, σ, β. The probability density function of x is given by

$$f(x;\mu,\sigma,\beta) = \frac{1}{\sigma\Gamma\left(1+\frac{1}{2\beta}\right)2^{1+\frac{1}{2\beta}}}\exp\left(-\frac{1}{2}\left|\frac{x-\mu}{\sigma}\right|^{2\beta}\right), \tag{A.1}$$

where $\mu(\mu \in \Re)$ is the location parameter , $\sigma(\sigma > 0)$ is the scale parameter and $\beta(\beta > 0)$ is the kurtosis parameter.

Different values of β are related to different unimodal symmetric curves. $\beta = 1$ gives Normal distribution; $\beta = 0.5$ gives Laplace distribution; $\beta \to \infty$ gives the Uniform distribution. The distribution of x is denoted as $x \sim PE(\mu, \sigma, \beta)$.

When the PE distribution is specified with mean $\mu = 0$ and scale $\sigma = 1$, it is called standard PE distribution. Figure A.1 plots the pdf of a few standard PE distributions with different β values and compares the PE distribution with Standard Normal Distribution.

Skewed Power Exponential Distribution

Azzalini (1986) developed the Skewed Power Exponential (SPE) Distribution to introduce both skewness and kurtosis. Here, we present the form of Fernandez and Steel (1998). Suppose x follows SPE distribution with parameters α, θ, σ and κ. The probability density function of x is defined as

$$f_{SPE}(x) = \frac{\alpha}{\sigma}\frac{\kappa}{1+\kappa^2}\exp\left(-\frac{\kappa^\alpha}{\sigma^\alpha}\left[(x-\theta)^+\right]^\alpha - \frac{1}{\sigma^\alpha\kappa^\alpha}\left[(x-\theta)^-\right]^\alpha\right), \tag{A.2}$$

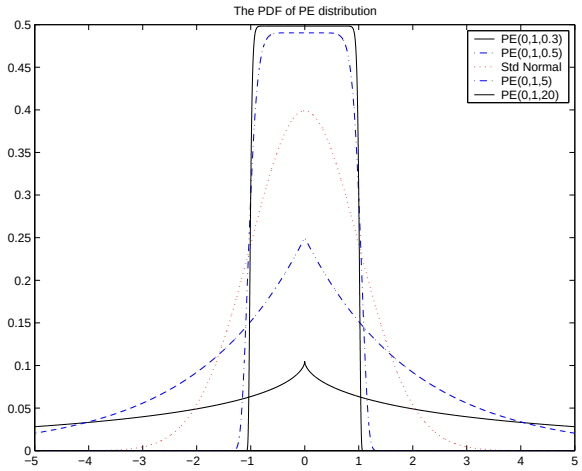

Figure A.1: PDF Plots of Standard PE Distributions.

where $\theta(\theta \in \Re)$ is the location parameter; $\sigma(\sigma > 0)$ is the scale parameter; $\kappa(\kappa > 0)$ is the skewness parameter; and $\alpha(\alpha > 0)$ is the shape parameter, which accommodates the kurtosis. And

$$u^+ = \begin{cases} \text{u}, & \text{if } u \geq 0; \\ 0, & \text{if } u < 0. \end{cases}$$

$$u^- = \begin{cases} \text{-u}, & \text{if } u \leq 0; \\ 0, & \text{if } u > 0. \end{cases}$$

The distribution of x is denoted by $x \sim SPE_\alpha(\theta, \sigma, \kappa)$. For $\kappa = 1$, the distribution is symmetric about θ.

Figure A.2 plots the pdf of a few SPE distributions with the same θ, σ and κ and different α values and compares them with the Standard Normal Distribution. Note that this figure is the same as Figure 5.5 in Chapter 5. We present it here for our convenience.

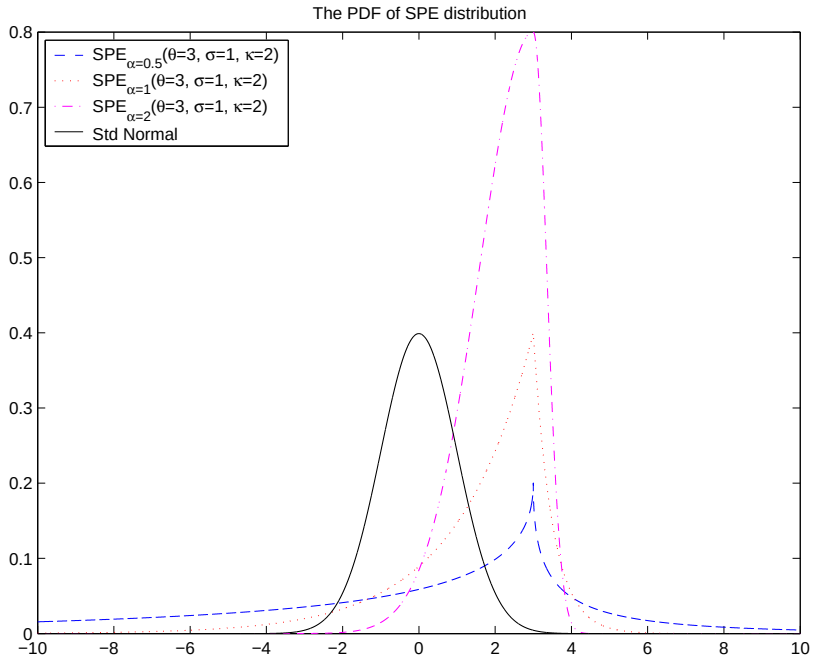

Figure A.2: PDF Plots of SPE Distributions.

Appendix B

Multivariate Power Exponential Distribution

Gómez et al. (1998) generalized the power exponential family of distributions to the multivariate case. A continuous random vector $\mathbf{x}$ is said to have a p-variate power exponential distribution if its probability density function is defined by

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)=\frac{p\Gamma(p/2)}{\pi^{p/2}\Gamma\left(1+\frac{p}{2\beta}\right)2^{1+p/2\beta}}\mid\boldsymbol{\Sigma}\mid^{-1/2}\exp\left\{-\frac{1}{2}\left((\mathbf{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right)^{\beta}\right\}. \tag{B.1}$$

where $\mathbf{x}=(x_1,\ldots,x_p)'$ is the p-dimensional vector, parameters $\boldsymbol{\mu}\in\mathcal{R}^p$, Σ is a $(p\times p)$ positive definite symmetric matrix and $0<\beta<\infty$. We denote $\mathbf{x}\sim PE_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$. When $p=1$, equation B.1 reduces to the density function of univariate PE distribution given A.1.

Sánchez-Manzano et al. (2002) proposed a definition of the matrix variate power exponential distribution . A random $(p\times n)$ matrix $\mathbf{X}$ is said to have a $(p\times n)$-variate power exponential distribution with parameters $\mathbf{M}$, a $(p\times n)$ matrix; $\boldsymbol{\Sigma}$, a $(p\times p)$ definite positive matrix; $\boldsymbol{\Phi}$, a $(n\times n)$ definite positive matrix and $\beta\in(0,\infty)$, if

$$Vec(\mathbf{X}')\sim PE_{pn}\left(Vec(\mathbf{M}'),\boldsymbol{\Sigma}\otimes\boldsymbol{\Phi},\beta\right) \tag{B.2}$$

where $Vec(\cdot)$ refers to the vector operator. We denote $\mathbf{X}\sim MPE_{p\times n}(\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Phi},\beta)$.

The density function of $\mathbf{X}$ is defined by

$$f(\mathbf{X},\boldsymbol{\Sigma},\boldsymbol{\Phi},\beta)=k\mid\boldsymbol{\Sigma}\mid^{n/2}\mid\boldsymbol{\Phi}\mid^{-p/2}\exp\left\{-\frac{1}{2}\left(tr((\mathbf{X}-\mathbf{M})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\boldsymbol{\Phi}^{-1})\right)^{\beta}\right\}, \tag{B.3}$$

where

$$k=\frac{pn\Gamma(pn/2)}{\pi^{pn/2}\Gamma(1+pn/2\beta)2^{1+pn/2\beta}}.$$

If $\mathbf{X}\sim MPE_{p\times n}(\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Phi},\beta)$, then $\mathbf{X}'\sim MPE_{n\times p}(\mathbf{M}',\boldsymbol{\Sigma},\boldsymbol{\Phi},\beta)$.

When $n=1$, the matrix variate PE distribution given in equation B.3 reduces to the multivariate PE distribution in equation B.1.

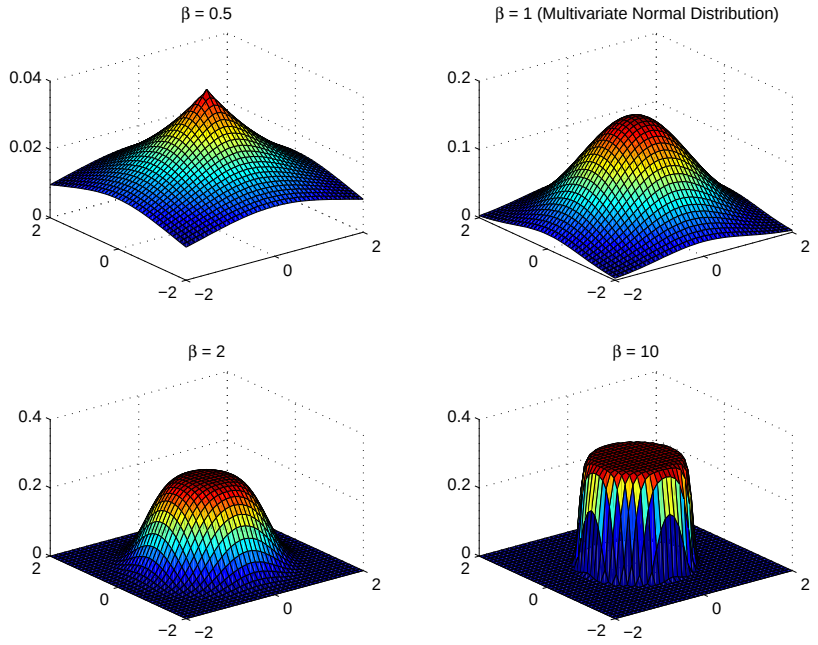

Figure B.1: PDF Plots for Multivariate PE Distributions.

Figure B.1 shows the pdf plots of a few matrix multivariate PE distributions with $p = 2, n = 1, \mathbf{\Phi} = I_1, \mathbf{\Sigma} = I_2$ and different β values.

When $\beta = 1$, the density in equation B.3 is reduced to multivariate normal distribution; when $\beta = 0.5$, the density is reduced to a matrix generalization of the double exponential distribution; when $\beta \to \infty$, the density is reduced to multivariate uniform distribution.

Vita

Yan Liu graduated from Tianjin University in China, where she obtained her Bachelor's degree in Engineering Economics and Master's degree in Management Science and System Engineering with a concentration in Supply Chain. She won several special academic awards throughout her studies at Tianjin University, which honored her outstanding achievements.

She joined the Statistics Department (now the department of Statistics, Operations and Management Science) at the University of Tennessee, Knoxville in 2001. She obtained her Master degree in Statistics in 2003. She has been working as a graduate teaching assistant in the University of Tennessee since 2001. Her work included instructing probability and statistics to the engineering and science students and instructing statistics to the business students. She loved teaching.

She worked as an intern in the Applied Statistics Lab at General Electric Global Research Center at Niskayuna, NY, in summer 2004. There she joined a team working on a financial project for GE Capital. She participated in developing data mining models and completed a green belt project. Both her teammates and she enjoyed their time working together.

She is a member of the American Statistical Association. She was the speaker at the 2004 and 2005 Joint Statistical Meetings.

Her research interests include Data Mining, Time Series, Multivariate Analysis, Regression Analysis, Survival Analysis, Design of Experiment.