Some Statistical Methods for Dimension Reduction

A thesis submitted for degree of

Doctorate of Philosophy

by

Ali J. Kadhim Al-Kenani

B.Sc., M.Sc.

Supervised by

Dr. Keming Yu

Department of Mathematical Sciences

School of Information System, Computing and Mathematics

September 2013

Abstract

The aim of the work in this thesis is to carry out dimension reduction (DR) for high dimensional (HD) data by using statistical methods for variable selection, feature extraction and a combination of the two. In Chapter 2, the DR is carried out through robust feature extraction. Robust canonical correlation (RCCA) methods have been proposed. In the correlation matrix of canonical correlation analysis (CCA), we suggest that the Pearson correlation should be substituted by robust correlation measures in order to obtain robust correlation matrices. These matrices have been employed for producing RCCA. Moreover, the classical covariance matrix has been substituted by robust estimators for multivariate location and dispersion in order to get RCCA.

In Chapter 3 and 4, the DR is carried out by combining the ideas of variable selection using regularisation methods with feature extraction, through the minimum average variance estimator (MAVE) and single index quantile regression (SIQ) methods, respectively. In particular, we extend the sparse MAVE (SMAVE) reported in (Wang and Yin, 2008) by combining the MAVE loss function with different regularisation penalties in Chapter 3. An extension of the SIQ of Wu et al. (2010) by considering different regularisation penalties is proposed in Chapter 4.

In Chapter 5, the DR is done through variable selection under Bayesian framework. A flexible Bayesian framework for regularisation in quantile regression (QR) model has been proposed. This work is different from Bayesian Lasso quantile regression (BLQR), employing the asymmetric Laplace error distribution (ALD). The error distribution is assumed to be an infinite mixture of Gaussian (IMG) densities.

Certificate of Originality

I hereby certify that the work presented in this thesis is my own research and has not been presented for a higher degree at any other university or institute. Any material that could be construed as the work of others is fully cited and appears in the references.

Ali J. Kadhim Al-Kenani

Acknowledgements

After the completion of this thesis at Brunel University, I wish to thank everyone who made this thesis possible. I wish to thank my supervisor Dr. Keming Yu for his supervision of this work. Also, I would like to express my great gratitude for his useful advice and encouragement in my work through our scientific discussions.

I thank Prof. Xiangrong Yin and Dr. Qin Wang for sending us the code for the SMAVE method in (Wang and Yin, 2008) and for some suggestions. Also, I thank Prof. Yan Yu for sending us the code for the SIQ method in (Wu et al., 2010).

Special thanks and deepest gratitude to the staff at Brunel University who have made my time at the university enjoyable and stimulating. Sincere gratitude goes out to my dear friends in Brunel University and outside Brunel University. Specifically, I would like to thank (alphabetically), Dr. Abdallah Ally, Fatmir Qirezi, Hakim Mezali, Dr. Hamied Alhashimi, Hussein Hashem, Dr. Majed Altemimi, Mortadah Almamoory, Dr. Rahim Al-Hamzawi and Dr. Tahir Reisan.

Last, but by no means least, I would like to thank my family for the unwavering support during my PhD study.

Author's Publications

1. Alkenani, A. and Yu, K. (2013). A comparative study for robust canonical correlation methods, Journal of Statistical Computation and Simulation 83, 690–718. (http://dx.doi.org/10.1080/00949655.2011.632775).

2. Alkenani, A. and Yu, K. (2013). Sparse MAVE with oracle penalties. Advances and Applications in Statistics 34, 85–105. (http://www.pphmj.com/abstract/7662.htm).

3. Alkenani, A. and Yu, K. (2013). Penalized single-index quantile regression. International Journal of Statistics and Probability 2, 12–30.

(http://dx.doi.org/10.5539/ijsp.v2n3p12).

4. Alkenani, A., Alhamzawi, R. and Yu, K. (2012). Penalized Flexible Bayesian Quantile Regression. Applied Mathematics 3, 2155–2168.

(http://dx.doi.org/10.4236/am.2012.312A296)

5. Alkenani, A. and Yu, K. (2012). New Bandwidth selection for kernel quantile estimators, Journal of Probability and Statistics. (http://dx.doi.org/10.1155/2012/138450).

Table of Contents

List of abbreviations

ACN	Asymmetric contamination
adaptive Lasso	Adaptive least absolute shrinkage and selection operator
AIC	Akaike information criterion
ALD	Asymmetric Laplace distribution
ALMAVE	Sparse MAVE with adaptive Lasso penalty
ALQR	Adaptive Lasso quantile regression
ALSIQ	Single index quantile regression with adaptive Lasso penalty
AP	Air pollution
ARP	Asymptotic rejection probability
BAL	Bayesian adaptive Lasso
BALQR	Bayesian adaptive Lasso quantile regression
BF	Body fat data
BH	Boston housing (BH) data
BL	Bayesian Lasso
BLQR	Bayesian Lasso quantile regression
BQR	Bayesian quantile regression
CCA	Canonical correlation analysis
CCC	Constrained canonical correlation
CD	Curse of dimensionality
CL	The classical canonical correlation estimators
CM	The canonical correlation estimators based on the constrained M
CMF	Conditional mean function
CMS	Central mean subspace
DGK	Devlin, Gnanadesikan and Kettenring estimator
DR	Dimension reduction
FBQR	Flexible Bayesian Quantile Regression
FC	The canonical correlation estimators based on the fast consistent high

	breakdown estimator.
FCH	Fast consistent high breakdown
FCD	Full conditional distribution
FMCD	Fast minimum covariance determinant
FMVE	Fast minimum volume ellipsoid
GS	Gibbs sampling
HD	High dimensional
ICDF	Inverse cumulative distribution function
IHT	Iterative Hessian Transformation
IMG	Infinite mixture of Gaussian
LAD	Least absolute deviation
LARS	Least angle regression
Lasso	Least absolute shrinkage and selection operator
LC's	Linear combinations
LD	Limiting distributions
LQR	Lasso quantile regression
LSIQ	Single index quantile regression with Lasso penalty
MAD	Median absolute deviations
MAVE	Minimum average variance estimator
MB	Median ball estimator
MC	The canonical correlation estimators based on the minimum covariance determinant
MCD	Minimum covariance determinant
MCP	Minimax concave penalty
MCPMAVE	Sparse MAVE with MCP penalty
MMSE	Median mean squared error
MSE	Mean squared error
MV	The canonical correlation estimators based on the minimum volume ellipsoid
MVE	Minimum volume ellipsoid
NLCCA	Nonlinear canonical correlation analysis
NOR	Normal distribution
OPG	Outer product of gradients
OP's	Oracle properties
pHd	principal Hessian directions

QR	Quantile regression
RARs	Robust alternating regressions
RCCA	Robust canonical correlation
RF	The canonical correlation estimators based on the reweighted fast consistent high breakdown estimator.
RFCH	Reweighted fast consistent high breakdown
RK	The canonical correlation estimators based on Kendall's tau correlation
RM	The canonical correlation estimators based on biweight midcorrelation
RMLD	Robust multivariate location and dispersion
RMV	The canonical correlation estimators based on the reweighted multivariate normal estimator
RMVN	Reweighted multivariate normal
ROP	Rate of penalization
RP	The canonical correlation estimators based on percentage bend correlation
RS	The canonical correlation estimators based on Spearman's rho correlation
RW	The canonical correlation estimators based on winsorised correlation
SAVE	Sliced average variance estimation
SCAD	Smoothly clipped absolute deviation
SCADMAVE	Sparse MAVE with SCAD penalty
SCN	Symmetric contamination
SD	Standard deviation
SDR	Sufficient dimension reduction
SI	Single index
SIQ	Single index quantile regression
SIR	Sliced inverse regression
SMAVE	Sparse MAVE
SMN	Scale mixture of normals
SSIR	Sparse sliced inverse regression method
T	Multivariate t distribution
WCC	Weighted canonical correlation
WGCNA	Weighted gene co-expression network analysis
WM	The canonical correlation estimators based on the weighted minimum covariance determinant

Chapter 1

Introduction

Data appears throughout society and trends show that the size of the data sets is becoming larger all the time. Recent developments in data gathering and storage capacities have resulted in huge amounts of multivariate data being collected at a rapid rate. For such large amounts of multivariate data, the well known "Curse of Dimensionality" (CD) poses a challenge to most statistical methods. Richard Bellman (1961) introduced the concept of the CD. The reason for the CD is the exponential increase in volume associated with adding extra dimensions to an associated mathematical space. This means that the increasing of the sparsity will be exponential given a fixed amount of data points. This problem causes the standard statistical methods fail in high dimensional (HD) data.

The number of the variables refers to the dimension of the data. The operation of reducing the number of random variables with as little loss of information as possible is called the dimension reduction (DR). It is one of the main solutions for the CD. The main two ways to shorten the dimensionality of the data are the subset selection and the feature extraction. The subset selection is the process of selecting a subset of the important variables and the feature extraction is the process of transforming (projecting) the variables into a fewer number of new ones.

1.1. Subset selection

Subset selection has become a popular topic of research in many fields. It is the process of choosing a subset of important variables for use in model building. All unimportant variables that have not been chosen are then implicitly assigned coefficients with a value of zero. The main assumption when using a variable selection technique is that the data contains many unimportant variables. Unimportant variables are those which provide no more information than the chosen variable, or that provide no useful information in any context.

Improving the performance of the model's prediction, providing faster and lower cost models and giving a good understanding of the dataset are the central aim of subset selection (Guyon and Elisseeff, 2003). Ranging from simple to sophisticated, many approaches have been developed for the sake of doing variable selection.

Traditional variable selection techniques, such as stepwise selection and best subset regression may suffer from instability, due to their inherent discreteness (Brieman, 1996). To tackle the instability, regularisation methods can also carry out variable selection, as long as the penalty term is appropriately chosen. Regularisation methods are usually formed by adding penalty terms onto the model parameters with respect to the standard loss functions, such as the squared error loss. Compared to traditional subset selection methods, which are discrete procedures, hence with high variance, regularisation methods supply a tool with which we can develop the model's interpretation ability and prediction precision via continuous shrinkage and automatic variable selection, where variable selection is carried out during the process of parameter estimation.

The first use of regularisation idea for variable selection is made by Donoho and Johnstone (1994) and then further developed by Tibshirani (1996) and many other

researchers. For example, Zou and Hastie (2005), Yuan and Lin (2006), Fan and Li (2001), Tibshirani et al. (2005), Zou (2006), Zou and Zhang (2009), Park and Casella (2008), Hans (2009, 2010), Scheipl and Kneib (2009) and Kyung et al. (2010), among others. Although the quadratic loss has some nice mathematical properties, it is very sensitive to non normal errors. Least absolute deviation (LAD) and quantile regression (QR) have lately been used in variable selection approaches as robust regressions.

Koenker and Bassett (1978) introduced the QR. It becomes a widespread approach to characterise the distribution of an outcome of interest, given a set of covariates. In many applications, the extreme conditional quantiles based on the predictors completely different from the centre. Therefore, QR provides a comprehensive analysis of the relationships among variables. It can be seen as an expansion for regression analysis in order to get a more complete and robust analysis (Koenker, 2005). QR has been employed in many real world applications such as finance, microarrays and ecological studies, see Koenker (2005) and Yu et al. (2003) for an overview. For the regularisation methods in the QR, see Koenker (2004), Wang et al. (2007), Li and Zhu (2008), Zou and Yuan (2008), Wu and Liu (2009), Yuan and Yin (2010), Li et al. (2010), Bradic et al. (2011), Alhamzawi et al. (2011), Alhamzawi and Yu (2012), Alkenani et al. (2012) and Alkenani and Yu (2013).

1.2. Feature extraction

Feature extraction shares the objective of subset selection, with the difference that the results must be explained in terms of all of the variables. It denotes the process of finding the transformation that projects the data from the original space to the feature space.

A vast number of feature extraction techniques have emerged in the literature for reducing the dimensionality, without the loss of as much information as possible from the data. These include principal component analysis (see Jolliffe, 2002; Zhang and Olive, 2009), factor analysis (see Gorsuch, 1983), independent component analysis (Comon, 1994), canonical correlation analysis (Hotelling, 1936; Fung et al., 2002; Branco et al., 2005; Zhou, 2009; Zhang, 2011; Alkenani and Yu, 2013), single index models (Powell et al., 1989; Härdle and Stoker, 1989; Ichimura, 1993; Delecroix et al. 2003), the sliced inverse regression (SIR) (Li, 1991), the sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), the principal Hessian directions (pHd) (Li, 1992), the minimum average variance estimator (MAVE) and the outer product of gradients (OPG) methods (Xia et al., 2002, see also Xia 2007, 2008) and successive direction estimation (Yin and Cook, 2005; Yin et al, 2008), among others. On the other hand, there are a number of investigations that have used the feature extraction techniques to solve the CD problem in QR models. For example, Chaudhuri (1991), Gannoun et al. (2004), Wu et al. (2010), Jiang et al. (2012) and Hua et al. (2012). Recently, many studies have been done on combining subset selection and feature extraction. This feature has greatly enhanced the power of DR in applications. For example, see Li et al. (2005), Ni et al. (2005), Zou et al. (2006), Li and Nachtsheim (2006), Li (2007), Zhou and He (2008), Li and Yin (2008), Wang and Yin (2008) and Zeng et al. (2012).

1.3. Thesis outline

This thesis consists of a number of published journal papers that are organised into chapters. Therefore, each chapter can be understood separately and any linkages to other chapters have been clarified. The outline of the thesis is given as follows:

In Chapter 2, robust canonical correlation (RCCA) methods have been proposed. In the correlation matrix, the Pearson correlation has been substituted with the percentage bend correlation and the winsorised correlation in order to get robust correlation matrices. The resulting matrices have been employed to produce RCCA methods. Moreover, the fast consistent high breakdown (FCH), reweighted fast consistent high breakdown (RFCH) and reweighted multivariate normal (RMVN) estimators are employed to estimate the covariance matrix in the canonical correlation analysis (CCA) in order to obtain RCCA methods. After that, these estimators are compared with the existing estimators. The practical precision of the proposed methods is studied by means of simulation experiments under different sampling schemes. Furthermore, to assess the robustness of the estimators, we make use of the breakdown plots and apply the test of independence.

In Chapter 3, we combine MAVE method (Xia et al., 2002) with smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), Adaptive least absolute shrinkage and selection operator (adaptive Lasso) (Zou, 2006) and the minimax concave penalty (MCP) (Zhang, 2010). Our proposed methods have merits over the sparse MAVE (SMAVE) (Wang and Yin, 2008) because all of these regularisation methods have the oracle properties (OP's) and have preferences over sparse inverse DR methods (Li, 2007), in that there is no need for any particular distribution on $\mathbf{x}$ and it is able to estimate the dimensions in the conditional mean function (CMF). The proposed

methods are studied via simulation and real dataset examples in order to examine their performance.

In Chapter 4, we propose an extension of the single index quantile regression (SIQ) method of Wu et al. (2010) by considering the least absolute shrinkage and selection operator (Lasso) and the adaptive Lasso methods for estimation and variable selection. In addition, computational algorithms have been evolved in order to calculate the penalised SIQ estimates. The performance of the proposed methods is verified by both simulation and real data analysis.

In Chapter 5, we develop a flexible Bayesian framework for regularisation in the QR model. Similar to Reich et al. (2010), the error distribution is assumed to be an infinite mixture of Gaussian (IMG) densities. This work is different from Bayesian Lasso employing asymmetric Laplace distribution (ALD) for the error. In fact, the use of the ALD is undesirable due to the lack of coherency. For example, for different τ we have a different distribution for the y_i's and it is difficult to reconcile these differences.

In Chapter 6, the conclusions of the thesis and recommendations for potential future work are summarised.

References

Alhamzawi, R., Yu, K. and Benoit, D. (2012). Bayesian adaptive Lasso quantile regression. Statistical Modelling 12, 279–297

Alhamzawi, R. and Yu, K. (2012). Bayesian Lasso mixed quantile regression. Journal of Statistical Computation and Simulation, to appear.

Alkenani, A., Alhamzawi, R. and Yu, K. (2012). Penalized Flexible Bayesian Quantile Regression. Applied Mathematics 3, 2155–2168.

Alkenani, A. and Yu, K. (2013). A Comparative Study for Robust Canonical Correlation Methods. Journal of Statistical Computation and Simulation 83, 690–718.

Alkenani, A. and Yu, K. (2013). Penalized single-index quantile regression. International Journal of Statistics and Probability 2, 12–30.

Bellman, R. E. (1961). Adaptive Control Processes. Princeton University Press, Princeton, New Jersey.

Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. Journal of Royal Statistics Society, Series B 73, 325–349.

Branco, J. A., Croux, C., Filzmoser, P. and Olivera, M. R. (2005). Robust canonical correlations: a comparative study. Computational Statistics 20, 203–229.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. The Annals of Statistics 24, 2350–2383.

Chaudhuri, P. (1991). Global nonparametric estimation of conditional quantile functions and their derivative. Journal of Multivariate Analysis 39, 246–269.

Common, P. (1984). Independent Component Analysis, a new concept?. Signal Processing 36, 287–314.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). Journal of the American Statistical Association 86, 328–332.

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. Journal of Multivariate Analysis 86, 213–226.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, 425–455.

Fan, J. and Li, R. Z. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Fung, W. K., He, X., Liu, L. and Shi, P. (2002). Dimension reduction based on canonical correlation. Statistica Sinica 12, 1093–1113.

Gannoun, A., Girard, S. and Saracco, J. (2004). Sliced inverse regression in reference curves estimation . Computational Statistics and Data Analysis 46, 103–122.

Gorsuch, R. L. (1983). Factor analysis. Hillsdale, New Jersey: Lawrence Erlbaum.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Hans, C. (2009). Bayesian lasso regression. Biometrika 96, 835–845.

Hans, C. (2010). Model uncertainty and variable selection in Bayesian lasso. Statistics and Computing 20, 221–229.

Härdle, W. and Stoker, T. (1989). Investing smooth multiple regression by the method of average derivatives. Journal of the American Statistical Association 84, 986–995.

Hotelling, H. (1936). Relations between two sets of variates. Biometrika 28, 321–377.

Hua, Y., Gramacy, R. B. and Lian, H. (2012). Bayesian quantile regression for single-index models. Statistics and Computing, to appear; preprint on arXiv:1110.0219.

Ichimura, H. (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models. Journal of Econometrics 58, 71–120.

Jiang, R., Zhou, Z. G., Qian, W. M. and Shao, W. Q (2012). Single-index composite quantile regression. Journal of the Korean Statistical Society 3, 323–332.

Jolliffe, I. T. (2002). Principal component analysis. 2nd ed. Berlin, Germany: Springer Verlag.

Koenker, R. (2004). Quantile regression for longitudinal data. Journal of Multivariate Analysis 91, 74–89.

Koenker, R. (2005). Quantile Regression, Cambridge, U.K.: Cambridge University Press.

Koenker, R. and Bassett, G. (1978). Regression quantiles. Econometrica 46, 33–50.

Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis 5, 369–412.

Li, K. (1991). Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association 86, 316–342.

Li, K. (1992). On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma. Journal of the American Statistical Association 87, 1025–1039.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. Journal of the American Statistical Association 87, 1025–1039.

Li, L. (2007). Sparse sufficient dimension reduction. Biometrika 94, 603–613.

Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free variable selection. Journal of the Royal Statistical Society, Series B 67, 285–299.

Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. Technometrics 48, 503–510.

Li, L. and Yin, X. (2008). Sliced Inverse Regression with regularizations. Biometrics 64, 124–131.

Li, Q., Xi, R. and Lin, N. (2010). Bayesian Regularized Quantile Regression. Bayesian Analysis 5, 1–24.

Li, Y. and Zhu, J. (2008). l_1-norm quantile regressions. Journal of Computational and Graphical Statistics 17, 163–185.

Ni, L., Cook, R. D. and Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. Biometrika 92, 242–247.

Park, T. and Casella, G. (2008). The Bayesian lasso. Journal of the American Statistical Association 103, 681–686.

Powell, J. L., Stock, J. M. and Stoker, T. M. (1989). Semi-parametric estimation of index coefficients. Econometrica 57, 1403–1430.

Reich, B., Bondell, H. and Wang H. (2010). Flexible Bayesian quantile regression for independent and clustered data. Biostatistics 2, 337–352.

Scheipl, F. and Kneib, T. (2009). Locally adaptive Bayesian P-splines with a normal exponential gamma prior. Computational Statistics and Data Analysis 53, 3533–3552.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society, Series B 67, 91–108.

Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. The Journal of Business and Economic Statistics 25, 347–355.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high-dimensional data: sparse MAVE. Computational Statistics and Data Analysis 52, 4512–4520.

Wu, T. Z., Yu, K. and Yu, Y. (2010). Single-index quantile regression. Journal of Multivariate Analysis 101, 1607–1621.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. Statistica Sinica 19, 801–817.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. The Annals of Statistics 35, 2654–2690

Xia, Y. (2008). A multiple-index model and dimension reduction. Journal of the American Statistical Association 103, 1631–1640.

Xia, Y., Tong, H., Li, W. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society, Series B 64, 363–410.

Yin, X. and Cook, R. D. (2005). Direction estimation in single-index regressions. Biometrika 92, 371–384.

Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. Journal of Multivariate Analysis 99, 1733–1757.

Yu, K., Lu, Z. and Stander, J. (2003). Quantile regression: Applications and current research areas. The Statistician 52, 331–350.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68, 49–67.

Yuan, Y. and Yin, G. (2010). Bayesian quantile regression for longitudinal studies with non-ignorable missing data. Biometrics 66, 105–114.

Zeng, P., He, T. and Zhu Y. (2012). A Lasso-type approach for estimation and variable selection in single index models. Journal of Computational and Graphical Statistics 21, 92–109.

Zhang, J. (2011). Applications of a Robust Dispersion Estimator. Ph.D. Thesis, Southern Illinois University. Available at (www.math.siu.edu/olive/szhang.pdf).

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics 38, 894–942.

Zhang, J. and Olive, D. J. (2009). Applications of a Robust Dispersion Estimator. Available at (www.math.siu.edu/olive/pprcovm.pdf).

Zhou, J. (2009). Robust dimension reduction based on canonical correlation. Journal of Multivariate Analysis 100, 195–209.

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. Annals of Statistics 36, 1649–1668.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, Series B 67, 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics 15, 265–286.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. Annals of Statistics 36, 1108–1126.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics 37, 1733–1751.

Chapter 2

A Comparative study for robust canonical correlation methods[1]

The purpose of this chapter is to get robust canonical correlation (RCCA) methods. In the correlation matrix, an approach that substitutes the Pearson correlation with the percentage bend correlation and the winsorised correlation in order to obtain robust correlation matrices is presented. Moreover, the fast consistent high breakdown (FCH), reweighted fast consistent high breakdown (RFCH) and reweighted multivariate normal (RMVN) estimators are employed to obtain robust covariance matrices in the canonical correlation analysis (CCA). Simulation studies are conducted and real data is employed in order to compare the performance of the proposed approaches with the existing methods.

The breakdown plots and independent tests are employed as criteria of the robustness and performance of the estimators. Based on the computational studies and real data example, suggestions on the practical implications of the results are proposed.

[1]This chapter is based on: Alkenani, A. and Yu, K. (2013). A comparative study for robust canonical correlation methods, Journal of Statistical Computation and Simulation 83, 690–718. http://dx.doi.org/10.1080/00949655.2011.632775.

2.1. Introduction

The CCA, originally proposed by Hotelling (1936), is a method that is used for gauging the linear relationship between two sets of variables. The aim of this method is to find basis vectors for two groups of variables achieve the correlations between the projections of the variables into these basis vectors are mutually maximised.

The CCA has been widely applied in many statistical areas and a major advantage of the CCA is its application for dimension reduction (DR) and thus, it acts as a valuable tool that facilitates the understanding of complicated relationships among multidimensional variables (Das and Sen, 1998). The CCA is routinely discussed in many multivariate statistical analysis textbooks. For example, see Anderson (2003), Johnson and Wichern (2003) and Mardia et al. (1979).

Suppose that $\boldsymbol{X}$ is a p-dimensional random variable and $\boldsymbol{Y}$ is a q-dimensional random variable, with $p \leq q$. Furthermore, suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the covariance matrix (if it exists)

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{XX}} & \boldsymbol{\Sigma_{XY}} \\ \boldsymbol{\Sigma_{YX}} & \boldsymbol{\Sigma_{YY}} \end{pmatrix}, \tag{2.1}$$

where $\boldsymbol{\Sigma_{XX}}$ and $\boldsymbol{\Sigma_{YY}}$ are non-singular. The objective of the CCA is to explore the linear relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$, as measured by the correlation between the linear combination (LC) of both groups of variables. Specifically, we look for

$$(\boldsymbol{\beta}_1, \boldsymbol{\eta}_1) = \operatorname{argmax}_{\mathbb{b},\mathbb{c}} \operatorname{Corr}(\mathbb{b}^T \mathbf{X}, \mathbb{c}^T \mathbf{Y}), \tag{2.2}$$

where Corr is the Pearson correlation and the vectors $\boldsymbol{\beta}_1 \in \mathbb{R}^p$ and $\boldsymbol{\eta}_1 \in \mathbb{R}^q$ are called the first pair of canonical vectors. Let $\mathrm{U}_1 = \boldsymbol{\beta}_1^T \mathbf{X}$ and $\mathrm{V}_1 = \boldsymbol{\eta}_1^T \mathbf{Y}$, which are the first pair of canonical variates. According to Equation (2.2), the vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\eta}_1$ are not unique. The normalisation constraint $Var(\boldsymbol{\beta}_1^T \mathbf{X}) = Var(\boldsymbol{\eta}_1^T \mathbf{Y}) = 1$ is required in order to identify $\boldsymbol{\beta}_1$ and $\boldsymbol{\eta}_1$ uniquely (up to a sign).

While the U_1 and V_1 are useful, they do not capture the full dependence structure between $\mathbf{X}$ and $\mathbf{Y}$. To this end, higher order canonical vectors defined for $k = 2,3, \ldots, p$ as

$$(\boldsymbol{\beta}_k, \boldsymbol{\eta}_k) = \text{argmax}_{\mathbb{b},\mathbb{c}} \, \text{Corr}(\mathbb{b}^{\text{T}} \mathbf{X}, \mathbb{c}^{\text{T}} \mathbf{Y}), \tag{2.3}$$

are used where the pairs of canonical variates of order k are $U_k = \boldsymbol{\beta}_k^T \mathbf{X}$ and $V_k = \boldsymbol{\eta}_k^T \mathbf{Y}$ and

$$\text{Cov}(U_k, U_j) = \boldsymbol{\beta}_k^T \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}_j = \text{Cov}(V_k, V_j) = \boldsymbol{\eta}_k^T \boldsymbol{\Sigma}_{\mathbf{YY}} \boldsymbol{\eta}_j = \begin{cases} 1 & \text{if} \quad k = j \\ 0 & \text{if} \quad 1 \leq j < k. \end{cases} \tag{2.4}$$

The correlation ρ_k between the canonical variates of the kth pair, $\rho_k = \text{Corr}(U_k, V_k)$, is the kth canonical correlation. Moreover, the canonical vectors $\boldsymbol{\beta}_k$ and $\boldsymbol{\eta}_k$ are the eigenvectors corresponding to the eigenvalues $\rho_1^2 \geq \cdots \geq \rho_p^2 > 0$ of the matrices

$$\boldsymbol{\Sigma}_{\mathbf{A}} = \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}} \boldsymbol{\Sigma}_{\mathbf{YY}}^{-1} \boldsymbol{\Sigma}_{\mathbf{YX}} \text{ and } \boldsymbol{\Sigma}_{\mathbf{B}} = \boldsymbol{\Sigma}_{\mathbf{YY}}^{-1} \boldsymbol{\Sigma}_{\mathbf{YX}} \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}}, \tag{2.5}$$

or

$$\boldsymbol{R}_{\mathbf{A}} = \boldsymbol{R}_{\mathbf{XX}}^{-1} \boldsymbol{R}_{\mathbf{XY}} \boldsymbol{R}_{\mathbf{YY}}^{-1} \boldsymbol{R}_{\mathbf{YX}} \text{ and } \mathbf{R}_{\mathbf{B}} = \boldsymbol{R}_{\mathbf{YY}}^{-1} \boldsymbol{R}_{\mathbf{YX}} \boldsymbol{R}_{\mathbf{XX}}^{-1} \boldsymbol{R}_{\mathbf{XY}}, \tag{2.6}$$

where $\mathbf{R} = \begin{pmatrix} \boldsymbol{R}_{\mathbf{XX}} & \boldsymbol{R}_{\mathbf{XY}} \\ \boldsymbol{R}_{\mathbf{YX}} & \boldsymbol{R}_{\mathbf{YY}} \end{pmatrix}$ is the correlation matrix. The matrices in Equations (2.5) and (2.6) have the same eigenvalues which correspond to the squared canonical correlations.

Hsu (1941) derived the limiting distributions (LD) of the canonical correlations in the case of a multivariate normal distribution. His result is valid under some very general assumptions regarding the population's canonical correlations. The LD of the canonical vectors have been considered in several papers, see Anderson (1999) for an overview. Kettenring (1971) has generalised CCA to more than two sets of variables. Beaghen (1997) has used a canonical variate approach to analyse the means of repeated measurements. Anderson (1999) gave the complete LD of the canonical correlations and vectors assuming that the nonzero population correlations are distinct.

In order to estimate the canonical correlations and canonical vectors of the population, we first estimate $\mathbf{\Sigma}$ by the sample covariance matrix followed by the computation of the eigenvalues and eigenvectors of the matrices $\mathbf{\Sigma_A}$ and $\mathbf{\Sigma_B}$ as given by Equation (2.5). This procedure works best when $\mathbf{X}$ and $\mathbf{Y}$ are from a multivariate normal distribution; however, it appears to be less efficient with respect to outlying observations. From a practical point of view, it is well known that the sample covariance matrix is not resistant to outliers and thus the CCA based on this matrix will result in uncertain and misleading results. Similarly, Romanazzi (1992) showed that the classical canonical vectors and correlations are also sensitive to outliers. Consequently, in order to obtain accuracy and robustness, there is a need to estimate the population covariance matrix using robust approaches.

An apparent procedure to make CCA more robust, is to estimate a sample covariance or correlation matrix using methods that can account for outliers. One such approach was presented by Karnel (1991), who considered M-estimators as robust estimator of $\mathbf{\Sigma}$ and then followed the classical approach. However, the robustness properties of the M-estimators are poor in high dimensions (Kent and Tyler, 1996).

There are many estimators for robust multivariate location and dispersion (RMLD). The minimum covariance determinant (MCD) estimator is the fastest estimator of the RMLD that has been shown to be both consistent and having a high breakdown point. It has $O(n^{\mathfrak{t}})$ complexity, where $\mathfrak{t} = 1 + \frac{p(p+3)}{2}$ (see Bernholt and Fischer, 2004). The complexity of the minimum volume ellipsoid (MVE) is far higher and there may be no known method for computing the $S, \mathcal{T},$ projection based, constrained M, M-estimate of the scale of the residuals and the M-estimate of the parameters and Stahel–Donoho estimators (Olive and Hawkins, 2010).

Since the mentioned estimators are computationally time consuming, these estimators have been replaced by practical estimators which strike a balance between accuracy and computing cost. However, none of the workable estimators have been proved to be consistent and having a high breakdown point. For example, the fast minimum covariance determinant (FMCD) estimator, which is given in (Rousseeuw and Van Driessen, 1999), is used to replace the MCD estimator. The robust multivariate techniques (one of which is the robust canonical correlation) that claim to use the impractical MCD estimator actually use Rousseeuw and Van Driessen (1999) FMCD estimator.

Taskinen et al. (2006) obtained the influence function and asymptotic properties for CCA based on robust covariance matrix estimates. Following the approach suggested by Wold (1966), Filzmoser et al. (2000) devised a robust method for getting U_1 and V_1 by using robust alternating regressions (RARs).

Branco et al. (2005) compared and discussed a number of approaches for robust canonical correlation analysis (RCCA). The authors proposed a robust method for obtaining all of the canonical variates using the RARs. Also, they stated that the CCA based on the FMCD estimator for the covariance matrix, is predominantly preferred due to its high breakdown point.

Zhou (2009) studied a weighted canonical correlation (WCC) method and its asymptotic properties. In the WCC, each observation is weighted based on its Mahalanobis distance. The author used the FMCD estimator to compute the Mahalanobis distance.

Jiao and Jian (2010) derived the asymptotic normal distributions of estimators of the projection pursuit method based on the CCA. Recently, Kudraszow and Maronna (2011) proposed a method for the RCCA based on the prediction approach.

Olive and Hawkins (2010) showed that the FMCD estimator is not a high breakdown estimator. The authors proposed practical $\sqrt{n}$ consistent, outlier resistant estimators for multivariate location and dispersion. They suggested the FCH, RFCH and RMVN estimators. The authors suggested employing the RMVN estimator for CCA, discrimination, factor analysis, principal components and regression. The RMVN estimator uses a slightly modified method for reweighting such that it gives good estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate normal data, even when there are outliers in the data. Zhang and Olive (2009) used the RMVN estimator with principle component analysis. They suggested employing the RMVN estimator with the classical multivariate procedures. Zhang (2011) used the RMVN estimator for CCA.

Estimators with high complexity require considerable computing time and therefore, their usage will be seldom. The FCH, RFCH and RMVN estimators are roughly 100 times faster than the FMCD estimator (Olive, 2013).

Cannon and Hsieh (2008) suggested robust nonlinear canonical correlation analysis (NLCCA) to deal effectively with data sets with that have low signal-to-noise ratios. To achieve this, they employed a neural network model architecture of standard NLCCA. The authors substituted the cost functions, which were used to set the model parameters using more robust variants. The Pearson correlation was replaced by a biweight midcorrelation.

Wilcox (2004) studied the percentage bend correlation (ρ_{pb}) which is based on the M-estimators of location and the percentage bend measure of scale.

Wilcox (2005) stated that robust versions of the Pearson correlation are divided into two types. The first type consists of those that are robust against outliers, without taking into account the general structure of the data, whereas the second type takes into account the general structure of the data when dealing with outliers. In the literature,

the first and second types are referred to as the M correlation and O correlation, respectively. Moreover, Wilcox (2005) described the four types of M correlations as the ρ_{pb}, biweight midcorrelation (ρ_b), winsorised correlation (ρ_{win}), and Kendall's tau correlation ($\rho_{\tau au}$). Similarly, the author also presented a number of O correlation methods, such as the fast minimum volume ellipsoid (FMVE), FMCD and skipped measures of correlations. The FMVE and FMCD measures employ the central half of the data to estimate location, scatter, covariance and correlation. Skipped correlations are obtained by detecting the outliers using one of the multivariate outlier detection methods and then removing these outliers and applying some of the correlation coefficients to the remaining data (see Wilcox, 2005).

To the author's knowledge, there is no study that has focused on replacing the Pearson correlation in the correlation matrix of the CCA with the ρ_{pb} and ρ_{win}. However, Olive and Hawkins (2010) recommended to employ the FCH, RFCH and RMVN estimators for the CCA, discrimination, factor analysis, principal components and regression and Zhang (2011) used the RMVN for CCA. Until now there has been no research employed regarding the FCH and RFCH estimators for estimating the covariance matrix in the CCA. To this end, the goal of this chapter is to get RCCA methods that depend on the ρ_{pb} and ρ_w in the correlation matrix. Furthermore, we aim to employ the FCH and RFCH estimators in order to estimate the covariance matrix in the CCA to obtain RCCA and then compare these estimators with other known estimators.

In this chapter, we conduct a comparative study to explore the performance of 13 different estimators for canonical vectors and correlation. Simulation studies have been used in order to compare the numerical performances of the 13 different estimators

under different sampling schemes. To assess the robustness of the estimators, we use the breakdown plots and apply the test of independence.

In Section 2.2, different robustifications of CCA are discussed. In Section 2.3, the different estimators are compared using simulation studies. In Section 2.4, the breakdown plots in order to study the robustness of the estimators are used. In Section 2.5, tests of independence are done for the different estimators. An application is used to evaluate the methods in Section 2.6. The conclusions are summarised in Section 2.7.

2.2. RCCA based on robust correlation and robust covariance matrices.

2.2.1. The percentage bend correlation (ρ_{pb})

Let a special case of Huber's function be defined as

$$\psi(x) = max[-1, min(1, x)].$$

Furthermore, let θ_x and θ_y be the respective population medians for the random variables x and y and then define w_x as the solution to the following equation:

$$P(|x - \theta_x| < w_x) = 1 - \gamma, \tag{2.7}$$

where $0 \leq \gamma \leq 0.5$.

Let φ_{pbx} and φ_{pby} denote the percentage bend measure of the location for x and y, respectively. Furthermore, let $\mathfrak{U} = \frac{(x-\varphi_{pbx})}{w_x}$ and $\mathfrak{L} = \frac{(y-\varphi_{pby})}{w_y}$, such that $E[\psi(\mathfrak{U})] = E[\psi(\mathfrak{L})] = 0$.

The percentage bend correlation between x and y is:

$$\rho_{pb} = \frac{E\{\psi(\mathfrak{U})\psi(\mathfrak{Ł})\}}{\sqrt{E\{\psi^2(\mathfrak{U})\}E\{\psi^2(\mathfrak{Ł})\}}}, \tag{2.8}$$

where $-1 \leq \rho_{pb} \leq 1$ and ρ_{pb} is a robust measure of the linear association between x and y, such that the variables x and y are said to be independent when $\rho_{pb} = 0$. The ρ_{pb} depends, in part, on w_x which is a generalisation of the median of the absolute deviations from the median (MAD).

The Huber's function is selected to be used in the percentage bend correlation for a number of reasons. Firstly, Huber's function is a monotonic function. Secondly, Huber's function gives a consistent estimator of location. Thirdly, it has the convenient feature of a single iteration being sufficient in the application. Finally, when $\psi(x) = max[-1, min(1, x)]$, the resulting gauge of scale is a gauge of dispersion (Wilcox, 1994). This means, w_x is a measure of dispersion when $\psi(x) = max[-1, min(1, x)]$.

In order to estimate the percentage bend correlation,

1) Let (x_1, y_1),…., (x_n, y_n), be a random sample. Let M_x be the sample median for the observations $x_1, \ldots\ldots, x_n$. Select a value for $\mathfrak{y}$, where $0 \leq \mathfrak{y} \leq 0.5$.

2) Compute $W_i = |x_i - M_x|$ and $\mathfrak{m} = [(1 - \mathfrak{y})n]$ and let $\hat{w}_x = W_{(\mathfrak{m})}$, where $W_{(1)} \leq \cdots \leq W_{(n)}$ are the W_i values written in ascending order.

3) Compute $S_x = \sum_{i=i_1+1}^{n-i_2} x_{(i)}$ and $\hat{\varphi}_x = \frac{\hat{w}_x(i_2 - i_1) + S_x}{n - i_1 - i_2}$, where i_1 is the number of x_i values, such that $\frac{(x_i - M_x)}{\hat{w}_x} < -1$ and i_2 is the number of x_i values, such that $\frac{(x_i - M_x)}{\hat{w}_x} > 1$.

4) Set $\mathfrak{U}_i = \frac{(x_i - \widehat{\varphi}_x)}{\widehat{w}_x}$. Repeat these computations for the y_i values, $Ł_i = \frac{(y_i - \widehat{\varphi}_y)}{\widehat{w}_y}$.

5) The estimated percentage bend correlation (r_{pb}) between x and y is:

$$r_{pb} = \frac{\sum A_i B_i}{\sqrt{\sum {A_i}^2 \sum {B_i}^2}}, \tag{2.9}$$

where,

$A_i B_i = \psi(\mathfrak{U}_i)\, \psi(Ł_i)$, $A_i = \psi(\mathfrak{U}_i)$, $B_i = \psi(Ł_i)$ and $\psi\ (x) = max[-1, min(1, x)]$.

In order to test the hypothesis $H_0: \rho_{pb} = 0,$ (2.10)

when x and y are independent, we need to compute:

$$T_{pb} = r_{pb}\sqrt{\frac{n-2}{1-r_{pb}^2}}. \tag{2.11}$$

H_0 is rejected if $|T_{pb}| > t_{1-\alpha}$, the $1-\alpha$ quantile of t distribution with degrees of freedom (D.F) $v = n - 2$ (Wilcox, 2005). Here, α is a significance level.

2.2.2. The biweight midcorrelation (ρ_b)

Let ψ be any odd function and let μ_x and μ_y be any measure of location for random variables x and y, respectively. Let ŧ_x and ŧ_y be some measure of scale for random variables x and y, respectively. Let ƙ be some constant and let:

$\mathrm{U} = (x - \mu_x)/(\text{ƙ}\,\text{ŧ}_x)$ and $\mathrm{V} = (y - \mu_y)/(\text{ƙ}\,\text{ŧ}_y)$. Then, a measure of covariance between x and y is:

$$\gamma_{xy} = \frac{\text{ƙ}^2 \text{ŧ}_x \text{ŧ}_y E\{\psi(\mathrm{U})\psi(\mathrm{V})\}}{E\{\acute{\psi}(\mathrm{U})\}\, E\{\acute{\psi}(\mathrm{V})\}}, \tag{2.12}$$

where $\acute{\psi}(.)$ is the derivative of $\psi(.)$.

and the corresponding measure of correlation is given by

$$\rho_b = \frac{\gamma_{xy}}{\sqrt{\gamma_{xx}\,\gamma_{yy}}} \quad -1 \leq \rho_b \leq 1. \tag{2.13}$$

Wilcox (2005) chose ψ as the biweight function and $\text{ƙ} = 9$, where the biweight function is defined as follows:

$$\psi(x) = \begin{cases} x(1 - x^2)^2 & if \quad |x| < 1, \\ 0 & if \quad |x| \geq 1. \end{cases} \tag{2.14}$$

Let M_x and M_y denote the respective medians calculated from the random sample $(x_1, y_1), \ldots, (x_n, y_n)$.

Define $\mathrm{U}_i = (x_i - M_x)/(9\ \mathrm{MAD}_x)$ and $\mathrm{V}_i = (y_i - M_y)/(9\mathrm{MAD}_y)$ then the MAD_x and MAD_y are the values of MAD for the x and y values.

Let $\text{ƛ}_i = \begin{cases} 1 & if \quad -1 \le \mathrm{U}_i \le 1, \\ 0 & \text{otherwise.} \end{cases}$ and $\text{ƀ}_i = \begin{cases} 1 & if \quad -1 \le \mathrm{V}_i \le 1, \\ 0 & \text{otherwise.} \end{cases}$

It follows that the sample biweight midcovariance between x and y is

$$bicov(x,y) = \frac{n\sum \text{ƛ}_i(x_i - M_x)\left(1-\mathrm{U}_i^2\right)^2 \text{ƀ}_i\left(y_i - M_y\right)\left(1-\mathrm{V}_i^2\right)^2}{\left(\sum \text{ƛ}_i(1-\mathrm{U}_i^2)(1-5\mathrm{U}_i^2)\right)\left(\sum \text{ƀ}_i(1-\mathrm{V}_i^2)(1-5\mathrm{V}_i^2)\right)}, \tag{2.15}$$

and the bi-weight mid-correlation is then given by:

$$r_b = \frac{bicov(x,y)}{\sqrt{bicov(x,x)bicov(y,y)}}. \tag{2.16}$$

To test the null hypothesis $$H_0: \rho_b = 0, \tag{2.17}$$

when x and y are independent, we need to compute the test statistic

$$T_b = r_b\sqrt{\frac{n-2}{1-r_b^2}}. \tag{2.18}$$

Under (2.18), we reject H_0 if $|T_b| > t_{1-\alpha/2}$, the $1-\alpha/2$ quantile of t distribution with D.F $v = n-2$.

2.2.3. The winsorised correlation (ρ_{win})

Let x_1 and x_2 be two random variables. Then, the population winsorised correlation between x_1 and x_2 is:

$$\rho_{win} = \frac{E_{win}\{(\mathrm{x}_1 - \mu_{win,1})(\mathrm{x}_2 - \mu_{win,2})\}}{\sigma_{win,1}\,\sigma_{win,2}} = \frac{\sigma_{win,12}}{\sigma_{win,1}\,\sigma_{win,2}}, \quad -1 \le \rho_{win} \le 1, \tag{2.19}$$

where $\sigma_{win,j}$, $j = 1, 2$, is the population winsorised standard deviation of x_j and $E_{win}(\mathrm{x})$ is the winsorised expected value of x. We can obtain the winsorised standard deviation and the winsorised expected value by computing the usual standard deviation and expected value, based on the winsorised observations.

In order to estimate ρ_{win}, based on the random sample $(\mathrm{x}_{11}, \mathrm{x}_{12}), \ldots, (\mathrm{x}_{n1}, \mathrm{x}_{n2})$, first winsorise the observations by computing the ɣ_{ij} values as follows:

$$\text{ɣ}_{ij} = \begin{cases} \mathrm{x}_{(g+1)j} & if \quad \mathrm{x}_{ij} \leq \mathrm{x}_{(g+1)j}, \\ \mathrm{x}_{ij} & if \ \mathrm{x}_{(g+1)j} < \mathrm{x}_{ij} < \mathrm{x}_{(n-g)j}, \\ \mathrm{x}_{(n-g)j} & if \quad \mathrm{x}_{ij} \geq X_{(n-g)j}, \end{cases} \tag{2.20}$$

where g is the number of observations trimmed, or winsorised, from each end of the distribution, corresponding to the jth group. Then ρ_{win} is estimated by computing the Pearson's correlation with the ɣ_{ij} values:

$$r_{win} = \frac{\sum(\text{ɣ}_{i1} - \bar{\text{ɣ}}_1)(\text{ɣ}_{i2} - \bar{\text{ɣ}}_2)}{\sqrt{\sum(\text{ɣ}_{i1} - \text{ɣ}_1)^2 \sum(\text{ɣ}_{i2} - \bar{\text{ɣ}}_2)^2}}. \tag{2.21}$$

To test the null hypothesis

$$H_0: \rho_{win} = 0, \tag{2.22}$$

we need to compute:

$$T_{win} = r_{win}\sqrt{\frac{n-2}{1-r_{win}^2}}. \tag{2.23}$$

Under (2.23), we reject H_0 if $|T_{win}| > t_{1-\alpha/2}$, the $1 - \alpha/2$ quantile of t distribution with D.F $v = \hbar - 2$, where $\hbar$ is the effective sample size and equal to the number of pairs of observations that are not winsorised.

2.2.4. Kendall's tau correlation ($\rho_{\tau au}$)

Kendall's tau correlation is a nonparametric M-type correlation. Because of being resistant to outlying observations, it is often said to be robust. Consider two pairs of observations (x_1, y_1) and (x_2, y_2), such that $x_1 < x_2$ and with the assumption that tied values never occur. If $y_1 < y_2$, then (x_1, y_1) and (x_2, y_2) will be concordant; otherwise (x_1, y_1) and (x_2, y_2) are discordant.

For n pairs of points, let

$$\mathcal{S}_{ij} = \begin{cases} 1 & if \ \ i\text{th and } j\text{th are concordant,} \\ -1 & \text{otherwise.} \end{cases}$$

Kendall's tau correlation formula is

$$r_{\tau au} = \frac{2\sum_{i<j}\mathcal{S}_{ij}}{n(n-1)}. \tag{2.24}$$

Although Kendall's tau correlation provides resistance against outliers, the presence of outliers can substantially change its value if the percentage of outliers is greater than 0.05.

Under independence, the population Kendall's tau correlation $\rho_{\tau au} = 0$.

To test the null hypothesis

$$H_0: \rho_{\tau au} = 0, \tag{2.25}$$

we compute:

$$Z = \frac{6\sum_{i<j}\mathcal{S}_{ij}}{\sqrt{2n(n-1)(2n+5)}}. \tag{2.26}$$

If $|Z| > Z_{1-\alpha/2}$, our decision will be rejecting H_0.

For the sake of comparison the canonical correlation estimators based on Kendall's tau correlation with other canonical correlation estimators, we apply the transformation $\sin\left(\frac{\pi}{2}\,\rho_{\tau au}\right)$ to obtain a consistent estimator under normality.

2.2.5. Spearman's rho correlation (ρ_s)

Spearman's rank correlation ρ_s is the most popular non-parametric correlation, which is a Pearson correlation based on the ranks of the observations. This correlation provides resistance against outliers; however, outliers that are properly placed can alter

its value considerably. In applications, a simple procedure can be used to calculate r_s. In order to estimate ρ_s based on a random sample the formula is given by:

$$r_s = 1 - \frac{6\sum đ_i^2}{n(n^2-1)}, \tag{2.27}$$

where $đ_i = rank\ x_i - rank\ y_i$, which is the difference between the ranks of each observation on the two variables (Myers and Arnold, 2003).

If the sampling from a bivariate normal distribution, r_s does not estimate the same quantity as the Pearson correlation. To compare the estimators of canonical correlations based on spearman's rho correlation with other estimators, we need to apply the transformation $2\sin\left(\frac{\pi}{6}\ \rho_s\right)$ in order to obtain a consistent estimator under normality.

Under the statistical independence, $\rho_s = 0$. To test the following hypothesis

$$H_0: \rho_s = 0, \tag{2.28}$$

We need to calculate the statistic:

$$T_s = r_s\sqrt{\frac{n-2}{1-r_s^2}}. \tag{2.29}$$

H_0 should be rejected if $|T_s| > t_{1-\alpha/2}$, the $1-\alpha/2$ quantile of t distribution with D.F $v = n-2$.

2.2.6. The MVE estimator

The MVE estimator is an affine equivariant estimator that has a high breakdown point (see Rousseuw and Leroy, 1987). Assume any ellipsoid containing 50% of the data. The idea is to find the ellipsoid having the smallest volume among all the ellipsoids. When this ellipsoid is found, the mean and covariance matrix of its points are taken as the estimated measures of location and scatter, respectively. In the multivariate normal model, the covariance matrix needs to be rescaled for consistency. In general, the group of all of the ellipsoids containing half of the data is very large, therefore the approximation must be used to find the MVE.

Let $\hbar^* = \frac{n}{2} + 1$, rounded down to the nearest integer. The approach for computing the FMVE estimator is summarised as follows:

1. Select $\hbar^*$ random points from the available n points without replacement.

2. Compute the volume of the ellipse containing these points.

3. Repeat step 1 and 2 many times.

The FMVE ellipsoid is the set of points giving the smallest volume (Wilcox, 2005).

2.2.7. The MCD estimator

The MCD estimator is also an affine equivariant estimator that has a high breakdown point. The difference between the MCD and MVE estimators is that rather than searching for the subset of 50% of the data that has the smallest volume, the MCD estimator searches for the 50% of the data that has the smallest generalised variance.

The MCD estimator searches for 50% of the data that is most tightly clustered together among all of the subsets containing 50% of the data, as measured by the generalised variance. Like the MVE estimator, the group of all subsets of 50% of the data is very large, hence an approximate method must be used. Rousseeuw and Van Driessen (1999) described an FMCD algorithm employed to achieve this aim. After we find an approximation of the subset of 50% of the data that minimise the generalised variance, we can obtain the MCD estimate of location and scatter by computing the usual mean and covariance matrix, based on its points. The MCD estimator has several merits over the MVE. The MCD estimator is more efficient than the MVE estimator because the MCD is asymptotically normal, whereas the MVE has a lower rate of convergence (Rousseeuw and Van Driessen, 1999). In our comparative study, we used the FMCD and reweighted MCD (WMCD) measures as practical approximations for the MCD.

2.2.8. The constrained M-estimators

Rocke (1996) suggested a modified biweight estimator, which is a constrained M-estimator, where values of g and q are to be determined and the non-decreasing function $\xi(d)$ is defined as:

$$\xi(d) =$$

$$\begin{cases} \frac{g^2}{2} - \frac{g^2(g^4-5g^2q^2+15q^4)}{30q^4} + d^2\left(0.5 + \frac{g^4}{2q^4} - \frac{g^2}{q^2}\right) & if \;\; g \le d \le g+q \\ +d^3\left(\frac{4g}{3q^2} - \frac{4g^3}{3q^4}\right) + d^4\left(\frac{3g^2}{2q^4} - \frac{1}{2q^2}\right) - \frac{4gd^5}{5q^4} + \frac{d^6}{6q^4} & \\ \frac{d^2}{2} & if \;\; 0 \le d < g \\ \frac{g^2}{2} + \frac{q(5q+16g)}{30} & if \;\; d > g+q \end{cases} \tag{2.30}$$

The values of a and q can be selected to obtain the wanted breakdown point and the asymptotic rejection probability (ARP). The ARP is the probability that an observation will obtain weight equals to zero when the size of the sample is huge. If the ARP is α, then a and q are determined by $E_{\chi_p^2}(\xi(d)) = b_0$ and

$$a + q = \sqrt{\chi^2_{p,1-\alpha}},$$

where b_0 is a constant and $\chi^2_{p,1-\alpha}$ is the $1-\alpha$ quantile of a chi-squared distribution with p D.F. Rocke (1996) showed that this estimator can be computed iteratively.

2.2.9. The FCH estimator

Olive and Hawkins (2010) proposed the FCH estimator. The FCH estimator uses the $\sqrt{n}$ consistent DGK estimator in (Devlin et al., 1981) and the high breakdown median ball (MB) estimator in (Olive, 2004) as attractors. An attractor is one of the trial fits used by the robust estimator. Therefore if the robust estimator draws $\mathcal{K}$ elemental sets and then refines them with concentration, then the $\mathcal{K}$ refined elemental sets are the attractors. The FCH estimator also uses a location criterion to choose the attractors. If DGK location estimator $T_{\mathcal{K},D}$ has a greater Euclidean distance from MED($\boldsymbol{X}$) than 50% of the data, where MED($\boldsymbol{X}$) is the coordinate-wise median, then FCH uses the MB attractor. The FCH estimator uses only the attractor with the smallest determinant if

$$\|T_{\mathcal{K},D} - MED(\boldsymbol{X})\| \leq MED\left(D_i(MED(\boldsymbol{X}), \boldsymbol{I}_p)\right), \qquad (2.31)$$

where $D_i(MED(\boldsymbol{X}), \boldsymbol{I}_p)$ is the Euclidean distance from $MED(\boldsymbol{X})$ and $\boldsymbol{I}_p$ is $p \times p$ identity matrix. Here $\|.\|$ refers to the Euclidean distance.

Let $(T_{\mathbb{A}}, \boldsymbol{C}_{\mathbb{A}})$ be the attractor that is used, where $T_{\mathbb{A}}$ and $\boldsymbol{C}_{\mathbb{A}}$ are the location and dispersion estimators, respectively. Then, the estimator $(T_F, \boldsymbol{C}_F)$ takes $T_F = T_{\mathbb{A}}$ and

$$\boldsymbol{C}_F = \frac{MED\left(D_i^2(T_{\mathbb{A}}, \boldsymbol{C}_{\mathbb{A}})\right)}{\chi^2_{p,0.5}} \boldsymbol{C}_{\mathbb{A}}, \tag{2.32}$$

where $D_i^2(T_{\mathbb{A}}, \boldsymbol{C}_{\mathbb{A}})$ is the ith squared sample Mahalanobis distance, which takes the form $D_i^2(T_{\mathbb{A}}, \boldsymbol{C}_{\mathbb{A}}) = \left(\mathbf{x}_i - T_{\mathbb{A}}(\boldsymbol{X})\right)^T \boldsymbol{C}_{\mathbb{A}}^{-1}(\boldsymbol{X})\left(\mathbf{x}_i - T_{\mathbb{A}}(\boldsymbol{X})\right)$ for each observation, the $\chi^2_{p,0.5}$ is the $0.50th$ percentile of a chi-squared distribution and F is the FCH estimator. Olive and Hawkins (2010) showed that the FCH estimator is a high breakdown estimator and $\boldsymbol{C}_F$ is non-singular, even with up to nearly 50% outliers.

2.2.10. The RFCH estimator

Olive and Hawkins (2010) used two standard reweighting steps to produce the RFCH estimator. Let $\left(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1\right)$ be the traditional estimator computed to n_1 cases with $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$ and let

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{MED\left(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)\right)}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_1. \tag{2.33}$$

Then, let $\left(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2\right)$ be the traditional estimator computed to the cases with $D_i^2\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1\right) \leq \chi^2_{p,0.975}$ and let

$$\boldsymbol{C}_{RFCH} = \frac{MED\left(D_i^2(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)\right)}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_2. \tag{2.34}$$

Olive and Hawkins (2010) showed that the RFCH is also a $\sqrt{n}$ consistent estimator.

2.2.11. The RMVN estimator

Olive and Hawkins (2010) suggested the RMVN estimator as a RMLD estimator and they showed this estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, r\boldsymbol{\Sigma})$ where $r > 0$. The RMVN estimator uses a slight modification to a standard reweighting method, such that the RMVN estimator produces good estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate normal data, even if outliers are present (Olive and Hawkins, 2010).

The RMVN estimator uses $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ with $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$ based on n_1 cases. Let $\varsigma_1 = min\{0.5(0.975)n/n_1, 0.995\}$ and $\hat{\boldsymbol{\Sigma}}_1 = \frac{MED\left(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)\right)}{\chi^2_{p,\varsigma_1}} \tilde{\boldsymbol{\Sigma}}_1$. Then, let $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$ be the traditional estimator computed to n_2 cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \leq \chi^2_{p,0.975}$.

$$\boldsymbol{C}_{RMVN} = \frac{MED\left(D_i^2(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)\right)}{\chi^2_{p,\varsigma_2}} \tilde{\boldsymbol{\Sigma}}_2,$$

where $\varsigma_2 = min\{0.5(0.975)n/n_2, 0.995\}$.

Olive (2013) shows that the FCH, RFCH and RMVN methods of RCCA produce consistent estimators of the kth canonical correlation ρ_k on a wide category of elliptically contoured distributions, see (Olive, 2013).

2.3. Simulation study

In this section, we employ a simulation study to compare the different methods. We considered the following:

CL is the classical CCA based on eigenvalues and eigenvectors of the matrices (2.5), which were estimated using the sample covariance matrix.

RP, RM, RW, RK and RS are the CCA based on eigenvalues and eigenvectors of the matrices (2.6) after we used ρ_{pb}, ρ_b, ρ_{win}, ρ_{tau} and ρ_s, respectively, instead of the Pearson correlation.

MV, MC, WM, CM, FC, RF and RMV are the CCA based on eigenvalues and eigenvectors of the matrices (2.5), which are estimated using the FMVE, FMCD, WMCD, CM, FCH, RFCH and RMVN estimators, respectively, instead of the classical sample covariance matrix.

The functions *pball* and *winall* from the Wilcox package at (http://www.unt.edu/rss/class/mike/Rallfun-v9_2.txt) have been used to compute the correlation matrices of ρ_{pb} and ρ_{win}, respectively. The function *bicor* from the package (weighted gene co-expression network analysis) (WGCNA) has been used in order to compute the midcorrelation matrix. The base functions *cor(,method = c("kendall"))* and *cor(, method = c("spearman"))* have been used to calculate Kendall and Spearman correlation matrices, respectively.

The base functions *cov.mve* and *cov.mcd* have been used for computing the FMVE and FMCD covariance matrices. The functions *covRob (,estim="weighted")* and *covRob (,estim="M")* from the package (robust) have been used to calculate the weighted MCD (WM) and constrained M (CM) covariance matrices, respectively. The function *covfch* from the package (rpack.txt) at (www.math.siu.edu/olive/rpack.txt) has

been used for calculating the FCH and RFCH covariance matrices and the function *covrmvn* has been used to calculate the RMVN covariance matrix.

We follow the simulation settings given in Branco et al. (2005). $\mathcal{R} = 500$ samples with size $n = 500$ have been generated. We have assumed $\mathbf{\Sigma_{XX}} = \mathbf{I}_p$ and $\mathbf{\Sigma_{YY}} = \mathbf{I}_q$. The choices for $\mathbf{\Sigma_{XY}}$ are summarised in Table 2.1.

Following the work of Branco et al. (2005), the following sampling distributions were assumed:

1) Normal distribution (NOR), $N_{p+q}(0, \mathbf{\Sigma})$.

2) Multivariate t distribution with three D.F ($\mathbb{T}$).

3) Symmetric contamination (SCN), where 95% of the observations have been generated from $N_{p+q}(0, \mathbf{\Sigma})$ and 5% have been generated from $N_{p+q}(0,9\mathbf{\Sigma})$.

4) Asymmetric contamination (ACN), where 95% of the observations have been generated from $N_{p+q}(0, \mathbf{\Sigma})$ and 5% of the observations equals the point $tr(\mathbf{\Sigma})\mathbf{1}^T$ (where $tr(\mathbf{\Sigma})$ is the trace of $\mathbf{\Sigma}$).

The estimated parameters for a replication ℓ ($\ell = 1, \dots, \mathcal{R}$) are denoted by $\hat{\rho}_k^{\ell}$, $\widehat{\boldsymbol{\beta}}_k^{\ell}$, and $\widehat{\boldsymbol{\eta}}_k^{\ell}$ for $k = 1, \dots, p$. We compare the estimated parameters with the "true" parameters ρ_k, $\boldsymbol{\beta}_k$, and $\boldsymbol{\eta}_k$. The true parameters were computed from the specific matrix $\mathbf{\Sigma}$. The mean squared error (MSE) has the following forms:

$$\mathrm{MSE}(\hat{\rho}_k) = \frac{1}{\mathcal{R}} \sum_{\ell=1}^{\mathcal{R}} \left(\phi\left(\hat{\rho}_k^{\ell}\right) - \phi(\rho_k) \right)^2, \qquad (2.35)$$

where $\phi(\rho_k) = \tanh^{-1}(\rho_k)$ is the Fisher transformation of ρ_k .

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_k) = \frac{1}{\mathcal{R}}\sum_{\ell=1}^{\mathcal{R}} \cos^{-1}\left(\frac{|\boldsymbol{\beta}_k^T \widehat{\boldsymbol{\beta}}_k^{\ell}|}{\|\widehat{\boldsymbol{\beta}}_k^{\ell}\|.\|\boldsymbol{\beta}_k\|}\right), \mathrm{MSE}(\widehat{\boldsymbol{\eta}}_k) = \frac{1}{\mathcal{R}}\sum_{\ell=1}^{\mathcal{R}} \cos^{-1}\left(\frac{|\boldsymbol{\eta}_k^T \widehat{\boldsymbol{\eta}}_k^{\ell}|}{\|\widehat{\boldsymbol{\eta}}_k^{\ell}\|.\|\boldsymbol{\eta}_k\|}\right) \qquad (2.36)$$

Table 2.1. Simulation Setup. $\boldsymbol{\Sigma}_{\mathbf{XX}} = \mathbf{I}_p$ and $\boldsymbol{\Sigma}_{\mathbf{YY}} = \mathbf{I}_q$.

p	q	$\boldsymbol{\Sigma}_{\mathbf{XY}}$
2	2	$\begin{bmatrix} 0.90 & 0 \\ 0 & 0.50 \end{bmatrix}$
4	4	$\begin{bmatrix} 0.90 & 0 & 0 & 0 \\ 0 & 0.50 & 0 & 0 \\ 0 & 0 & 0.33 & 0 \\ 0 & 0 & 0 & 0.25 \end{bmatrix}$

Table 2.2. The $\mathrm{MSE_S}$ of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ multiplied by 1000 for 13 different methods, when the data is from NOR, $p = 2$ and $q = 2$.

	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\eta}}_1$	$\widehat{\boldsymbol{\eta}}_2$	$\phi(\hat{\rho}_1)$	$\phi(\hat{\rho}_2)$
CL	22.33	44.07	22.41	44.29	2.02	2.21
RP	23.96	45.83	24.08	43.67	5.46	3.22
RM	22.92	44.71	23.23	43.62	2.44	2.39
RW	25.74	47.34	27.04	45.49	21.19	6.74
RK	21.91	39.74	21.75	38.69	2.30	2.57
RS	23.75	45.72	23.82	43.97	4.37	2.89
MV	28.48	56.15	28.73	54.79	3.35	3.68
MC	27.78	54.07	28.78	53.02	2.76	3.32
WM	27.12	55.78	27.99	54.04	3.11	2.90
CM	28.12	52.62	29.52	56.39	3.24	3.24
FC	62.89	123.70	60.29	121.65	16.26	16.72
RF	26.59	50.94	24.88	50.03	2.66	2.35
RMV	26.57	51.30	24.98	50.14	2.64	2.36

Table 2.3. The MSE_S of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ multiplied by 1000 for 13 different methods, when the data is from SCN, $p = 2$ and $q = 2$.

	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\eta}}_1$	$\widehat{\boldsymbol{\eta}}_2$	$\phi(\hat{\rho}_1)$	$\phi(\hat{\rho}_2)$
CL	35.38	69.18	35.32	70.82	5.28	5.09
RP	25.11	46.32	25.82	46.21	9.04	3.46
RM	26.14	47.38	26.99	46.46	8.57	3.48
RW	26.24	46.69	27.79	47.76	25.73	7.09
RK	22.71	41.48	23.55	41.28	2.93	2.45
RS	25.29	46.71	26.01	46.71	7.93	3.29
MV	29.54	56.49	27.89	55.89	3.49	3.27
MC	28.59	55.61	28.34	54.68	3.11	3.03
WM	27.89	54.50	28.13	56.11	3.16	3.37
CM	28.82	58.54	26.75	56.45	3.39	2.99
FC	62.54	119.45	60.79	124.31	18.30	7.94
RF	25.73	49.93	25.97	49.13	2.42	2.74
RMV	26.25	51.18	26.08	49.89	2.43	2.88

Table 2.4. The MSE_S of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ multiplied by 1000 for 13 different methods, when the data is from $\mathbb{T}$, $p = 2$ and $q = 2$.

	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\eta}}_1$	$\widehat{\boldsymbol{\eta}}_2$	$\phi(\hat{\rho}_1)$	$\phi(\hat{\rho}_2)$
CL	67.9	124.8	65.3	125.4	21.4	17.0
RP	27.5	48.3	28.1	49.7	15.0	5.0
RM	29.6	48.3	28.9	49.0	22.8	7.1
RW	28.2	46.7	27.5	47.5	37.3	8.6
RK	25.4	43.4	25.0	44.5	3.6	3.1
RS	27.5	48.2	28.0	49.5	14.9	4.9
MV	35.5	71.7	37.4	72.0	5.7	5.6
MC	32.1	62.0	32.3	64.7	4.1	4.4
WM	35.7	68.6	34.1	66.9	4.9	4.7
CM	33.9	66.7	36.3	67.9	5.2	4.4
FC	54.3	107.4	53.8	105.3	11.9	11.0
RF	33.9	68.7	34.3	66.9	4.7	4.1
RMV	34.8	69.4	35.3	67.7	4.8	4.4

Table 2.5. The MSE_S of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ multiplied by 1000 for 13 different methods, when the data is from ACN, $p = 2$ and $q = 2$.

	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\eta}}_1$	$\widehat{\boldsymbol{\eta}}_2$	$\phi(\hat{\rho}_1)$	$\phi(\hat{\rho}_2)$
CL	103.36	482.04	103.80	483.60	113.22	44.62
RP	37.09	159.47	37.49	163.25	3.94	5.11
RM	39.72	175.93	39.07	179.34	8.54	8.16
RW	33.89	118.39	34.30	122.33	7.52	2.92
RK	70.08	162.09	70.95	165.02	15.96	12.00
RS	39.64	174.70	40.14	178.36	4.71	5.63
MV	29.47	56.70	29.55	55.85	3.32	3.29
MC	29.58	55.49	28.29	53.65	3.16	2.89
WM	27.53	55.40	27.54	53.51	3.12	3.00
CM	29.14	55.79	28.23	55.02	3.17	2.93
FC	66.19	133.87	64.49	136.49	19.01	19.50
RF	25.64	50.06	26.24	48.01	2.46	2.66
RMV	26.59	50.89	27.01	49.27	2.56	2.79

Table 2.6. The MSE_S of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\beta}}_4, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \widehat{\boldsymbol{\eta}}_3, \widehat{\boldsymbol{\eta}}_4, \phi(\hat{\rho}_1), \phi(\hat{\rho}_2), \phi(\hat{\rho}_3)$ and $\phi(\hat{\rho}_4)$ multiplied by 1000 for 13 different methods, when the data is from NOR, $p = 4$ and $q = 4$.

	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\beta}}_3$	$\widehat{\boldsymbol{\beta}}_4$	$\widehat{\boldsymbol{\eta}}_1$	$\widehat{\boldsymbol{\eta}}_2$	$\widehat{\boldsymbol{\eta}}_3$	$\widehat{\boldsymbol{\eta}}_4$	$\phi(\hat{\rho}_1)$	$\phi(\hat{\rho}_2)$	$\phi(\hat{\rho}_3)$	$\phi(\hat{\rho}_4)$
CL	40.2	189.3	370.0	350.8	42.1	189.5	367.5	345.0	2.0	2.3	1.7	2.0
RP	43.0	203.0	396.6	372.7	45.0	203.6	395.7	369.3	5.1	2.6	1.7	2.6
RM	41.0	194.5	379.4	357.4	43.4	195.6	377.5	353.0	2.4	2.2	1.7	2.1
RW	46.0	223.1	431.2	398.8	48.9	220.1	429.8	397.6	20.7	5.6	2.6	4.0
RK	38.3	196.4	391.9	366.7	40.4	197.4	388.9	362.1	2.4	2.5	1.9	2.2
RS	42.0	201.2	393.1	368.9	44.8	201.5	391.3	364.6	4.3	2.5	1.7	2.4
MV	47.7	223.0	445.7	419.9	47.9	224.8	439.9	411.0	2.9	3.4	2.6	2.9
MC	45.7	212.7	412.2	388.6	47.7	213.2	412.5	387.8	2.7	2.9	2.3	2.5
WM	45.8	219.4	414.0	376.5	45.5	219.9	418.7	383.5	2.9	2.8	2.4	2.7
CM	47.2	222.1	434.5	411.4	45.8	222.2	434.9	406.5	2.3	2.6	2.2	2.6
FC	86.0	440.4	744.2	651.4	90.4	446.3	746.0	660.0	13.0	9.9	7.0	10.8
RF	43.6	205.9	427.1	402.4	43.4	207.6	426.6	401.7	2.5	2.5	2.1	2.0
RMV	43.9	206.8	426.4	401.4	43.5	208.6	426.9	402.0	2.6	2.5	2.1	2.0

Table 2.7. The MSE_S of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\beta}}_4, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \widehat{\boldsymbol{\eta}}_3, \widehat{\boldsymbol{\eta}}_4, \phi(\hat{\rho}_1), \phi(\hat{\rho}_2), \phi(\hat{\rho}_3)$ and $\phi(\hat{\rho}_4)$ multiplied by 1000 for 13 different methods, when the data is from SCN, $p = 4$ and $q = 4$.

	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\beta}}_3$	$\widehat{\boldsymbol{\beta}}_4$	$\widehat{\boldsymbol{\eta}}_1$	$\widehat{\boldsymbol{\eta}}_2$	$\widehat{\boldsymbol{\eta}}_3$	$\widehat{\boldsymbol{\eta}}_4$	$\phi(\hat{\rho}_1)$	$\phi(\hat{\rho}_2)$	$\phi(\hat{\rho}_3)$	$\phi(\hat{\rho}_4)$
CL	63.6	322.7	594.3	523.1	61.6	321.5	585.4	515.4	5.3	5.7	4.1	4.5
RP	46.3	217.6	435.3	400.1	45.3	213.4	428.9	395.5	7.1	2.9	2.0	2.6
RM	47.6	218.3	439.0	406.4	46.5	214.8	432.0	401.0	6.8	2.9	2.0	2.7
RW	49.2	232.7	450.9	413.0	48.6	228.2	443.3	407.9	22.7	5.3	2.9	3.8
RK	42.1	209.8	433.0	398.6	41.1	209.5	425.9	392.9	2.6	2.6	2.3	2.2
RS	46.6	218.2	435.2	399.0	45.3	215.1	427.7	394.2	6.1	2.9	2.0	2.6
MV	48.4	232.7	465.5	428.7	46.7	228.4	459.1	423.5	2.8	2.7	2.5	2.6
MC	45.6	224.4	454.2	422.4	46.0	218.4	449.8	417.7	2.6	2.7	2.4	2.5
WM	46.5	212.2	448.7	423.3	48.3	213.4	451.2	423.2	3.0	2.9	2.2	2.3
CM	47.6	221.8	448.2	411.7	47.0	223.2	448.5	417.5	2.3	2.4	2.3	2.6
FC	88.5	453.2	707.3	616.0	88.7	452.5	707.5	628.8	10.9	10.4	6.8	10.9
RF	44.2	200.7	393.7	369.0	43.0	200.7	395.1	374.9	2.5	2.4	2.1	2.2
RMV	44.6	203.6	400.1	373.5	43.4	204.0	401.9	379.5	2.5	2.5	2.1	2.2

Table 2.8. The MSE_S of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\beta}}_4, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \widehat{\boldsymbol{\eta}}_3, \widehat{\boldsymbol{\eta}}_4, \phi(\hat{\rho}_1), \phi(\hat{\rho}_2), \phi(\hat{\rho}_3)$ and $\phi(\hat{\rho}_4)$ multiplied by 1000 for 13 different methods, when the data is from $\mathbb{T}$, $p = 4$ and $q = 4$.

	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\beta}}_3$	$\widehat{\boldsymbol{\beta}}_4$	$\widehat{\boldsymbol{\eta}}_1$	$\widehat{\boldsymbol{\eta}}_2$	$\widehat{\boldsymbol{\eta}}_3$	$\widehat{\boldsymbol{\eta}}_4$	$\phi(\hat{\rho}_1)$	$\phi(\hat{\rho}_2)$	$\phi(\hat{\rho}_3)$	$\phi(\hat{\rho}_4)$
CL	102.3	499.4	767.9	695.0	104.5	489.6	761.3	676.9	44.4	24.0	9.2	10.7
RP	48.3	226.9	474.6	445.3	50.0	223.6	476.3	448.8	15.4	3.4	2.4	3.4
RM	51.6	242.5	510.3	472.5	53.5	239.2	511.1	475.1	7.9	4.7	2.8	4.3
RW	49.0	231.9	479.5	450.9	50.7	228.9	487.8	456.4	38.7	6.2	3.4	4.5
RK	43.8	223.0	461.7	435.9	45.2	218.9	466.1	440.0	3.6	3.2	2.5	2.6
RS	48.3	227.5	468.6	441.4	49.7	223.7	472.0	445.1	15.6	3.4	2.4	3.4
MV	58.3	282.7	548.7	505.2	57.8	289.2	559.5	510.2	4.2	5.0	4.0	4.5
MC	54.8	267.1	528.9	493.1	54.8	271.7	534.0	493.2	4.1	4.3	3.5	3.8
WM	60.6	293.2	533.8	484.4	59.1	291.0	532.5	477.8	4.3	5.0	3.6	4.0
CM	60.0	298.4	556.0	511.5	56.7	393.5	547.9	494.4	4.7	4.7	3.6	4.1
FC	78.6	380.5	686.6	612.5	77.5	295.6	685.8	614.9	23.7	7.8	5.5	7.0
RF	59.2	267.3	555.1	506.1	57.6	271.4	557.5	517.9	4.1	4.5	3.2	3.8
RMV	59.9	273.3	549.4	505.3	57.9	273.9	548.3	507.9	4.3	4.8	3.5	3.8

Table 2.9. The MSE_S of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\beta}}_4, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \widehat{\boldsymbol{\eta}}_3, \widehat{\boldsymbol{\eta}}_4, \phi(\hat{\rho}_1), \phi(\hat{\rho}_2), \phi(\hat{\rho}_3)$ and $\phi(\hat{\rho}_4)$ multiplied by 1000 for 13 different methods, when the data from ACN, $p = 4$ and $q = 4$.

	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\beta}}_3$	$\widehat{\boldsymbol{\beta}}_4$	$\widehat{\boldsymbol{\eta}}_1$	$\widehat{\boldsymbol{\eta}}_2$	$\widehat{\boldsymbol{\eta}}_3$	$\widehat{\boldsymbol{\eta}}_4$	$\phi(\hat{\rho}_1)$	$\phi(\hat{\rho}_2)$	$\phi(\hat{\rho}_3)$	$\phi(\hat{\rho}_4)$
CL	237.7	1101.	962.4	693.3	238.5	1101.1	960.8	690.8	777.1	198.5	15.4	3.1
RP	62.0	497.7	711.6	579.4	63.5	497.8	710.8	576.7	4.1	12.7	6.1	2.2
RM	40.9	190.5	404.5	383.8	41.5	190.2	403.0	379.3	2.4	2.0	1.9	2.1
RW	57.6	429.8	666.6	566.1	59.2	431.9	665.1	564.1	7.2	3.2	2.6	2.3
RK	117.9	583.9	756.7	595.2	118.9	583.9	755.2	592.9	17.9	32.5	10.2	3.0
RS	66.0	529.5	735.6	592.1	67.4	530.4	734.6	590.1	4.9	15.0	6.4	2.3
MV	45.5	219.4	454.2	428.8	46.8	214.3	450.0	427.5	2.7	2.8	2.4	2.6
MC	45.2	211.5	436.6	410.6	46.1	210.1	433.3	409.1	2.7	2.7	2.3	2.4
WM	47.0	211.4	434.4	411.7	47.5	211.5	438.0	415.4	2.7	2.8	2.3	2.4
CM	46.2	221.5	440.9	409.3	47.1	225.2	444.3	411.4	2.5	2.9	2.5	2.5
FC	91.3	461.0	742.7	645.6	92.3	456.8	734.4	643.1	11.4	11.4	7.7	10.0
RF	44.5	196.6	406.6	384.6	44.5	198.6	405.1	387.7	2.5	2.6	2.1	2.2
RMV	44.7	199.6	414.1	391.0	44.9	201.0	413.3	395.5	2.5	2.7	2.2	2.2

The findings of the simulation are reported in Tables 2.2–2.9 and Figures 2.1–2.4.

From Table 2.2, the data from the NOR shows that the lowest MSEs for the estimated canonical vectors $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2$, $\widehat{\boldsymbol{\eta}}_1$ and $\widehat{\boldsymbol{\eta}}_2$ were achieved from the RK method, while the largest MSEs were achieved from the FC method. For canonical correlations, the lowest MSEs for the transformed estimated canonical correlation $\phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ were achieved from the CL method, while the largest MSEs for $\phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ were achieved from the RW and FC methods, respectively.

From Table 2.3, the data from the SCN shows that the best estimates for $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1$ and $\widehat{\boldsymbol{\eta}}_2$ were achieved from the RK method, while the worst estimates were achieved from the FC method, with respect to the MSE. For canonical correlations, the best estimate for $\phi(\hat{\rho}_1)$ was achieved from the RF method, while the worst estimate was achieved from the RW. The best estimate for $\phi(\hat{\rho}_2)$ was achieved from the RK, while the worst estimate was achieved from the FC method.

From Table 2.4, the data from $\mathbb{T}$ distribution shows that the best estimates for $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1$ and $\widehat{\boldsymbol{\eta}}_2$ were achieved from the RK method, while the worst estimates were

achieved from the CL method, with respect to the MSE. For canonical correlations, the best estimates for $\phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ were achieved from the RK method, while the worst estimates were achieved from RW and CL methods, respectively.

From Table 2.5, the data from the ACN shows that the lowest MSEs for $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1$ and $\widehat{\boldsymbol{\eta}}_2$ were achieved from the RF method, while the biggest MSEs were achieved from the CL method. For canonical correlations, the lowest MSEs for $\phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ were achieved from the RF method, while the biggest MSEs were achieved from the CL method.

Figure 2.1 shows the MSEs for dimensions $p = 2$ and $q = 2$. The first picture from the left and that from the right show the MSEs for $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$. The second picture from the left and that from the right present the MSEs for $\widehat{\boldsymbol{\eta}}_1$ and $\widehat{\boldsymbol{\eta}}_2$. The third picture from the left and that from the right present the MSEs for $\phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$. The horizontal axis refers to the 13 different methods and the vertical axis refers to the MSEs of the estimators. From Figure 2.1, it is clear that the largest MSEs are for the estimators in the case of ACN and then for those in the case of $\mathbb{T}$ distribution. When considering the ACN, the lowest MSEs for $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ were achieved from the RF and RMV methods, while the biggest MSEs were achieved from the CL and RK methods for $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\eta}}_1$, or the CL and RM methods for $\widehat{\boldsymbol{\beta}}_2$ and $\widehat{\boldsymbol{\eta}}_2$, or the CL and FC methods for $\phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$. For the case of the $\mathbb{T}$ distribution, the lowest MSEs for $\widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1$ and $\widehat{\boldsymbol{\eta}}_2$ were achieved from the RK and RW methods, while the lowest MSEs for $\widehat{\boldsymbol{\beta}}_1$, $\phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ were achieved from the RK and RP methods, RK and MC methods and Rk and RF methods, respectively. The biggest MSEs for $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2$ and $\phi(\hat{\rho}_2)$ were achieved from the CL and FC methods, while the biggest MSEs were achieved from the RW and RM methods for $\phi(\hat{\rho}_1)$.

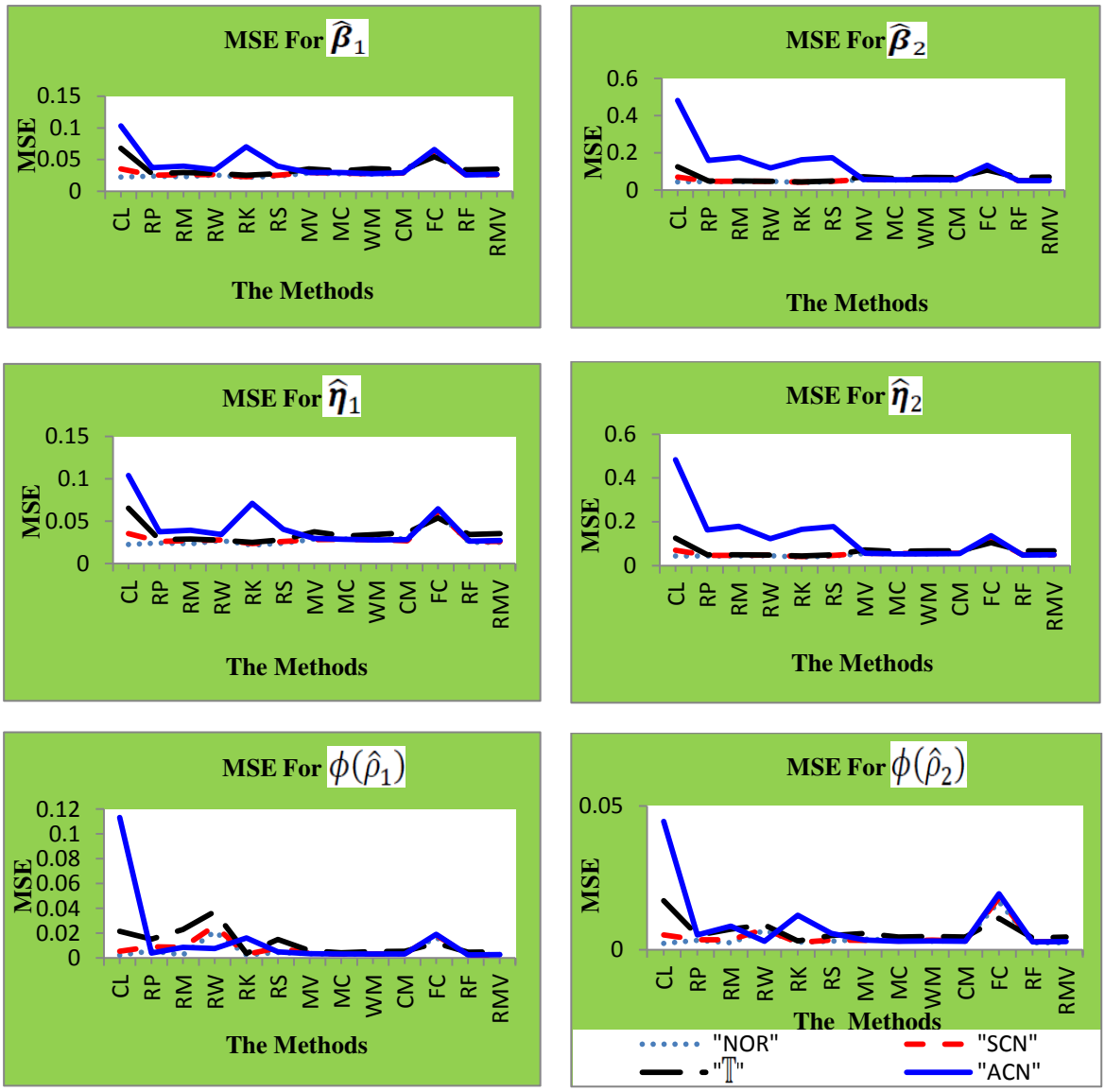


Figure 2.1. The MSEs for the canonical correlations and vectors for 13 estimators and under 4 sampling settings for $p = 2$ and $q = 2$.

From Table 2.6, for the dimensions $p = 4$ and $q = 4$, the data from the NOR shows that the lowest MSEs for $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2$ and $\widehat{\boldsymbol{\eta}}_1$ were achieved from the RK method, while the biggest MSEs were achieved from the FC method. The lowest MSEs for $\widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\beta}}_4, \widehat{\boldsymbol{\eta}}_2, \widehat{\boldsymbol{\eta}}_3$ and $\widehat{\boldsymbol{\eta}}_4$ were achieved from the CL method, while the biggest MSEs were achieved from the FC method. For canonical correlations, the lowest MSE for $\phi(\hat{\rho}_1)$ was achieved from the CL method, while the biggest MSE was achieved from the RW method. The lowest MSEs for $\phi(\hat{\rho}_2)$, $\phi(\hat{\rho}_3)$ and $\phi(\hat{\rho}_4)$ were achieved from the RM

method. The biggest MSEs for $\phi(\hat{\rho}_2)$, $\phi(\hat{\rho}_3)$ and $\phi(\hat{\rho}_4)$ were achieved from the FC method.

From Table 2.7, for the dimensions $p = 4$ and $q = 4$, the data from the SCN shows that the lowest MSEs for $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\eta}}_1$ were achieved from the RK method, while the biggest MSEs were achieved from the FC method. The lowest MSEs for $\widehat{\boldsymbol{\beta}}_2$, $\widehat{\boldsymbol{\beta}}_3$, $\widehat{\boldsymbol{\beta}}_4$, $\widehat{\boldsymbol{\eta}}_2$, $\widehat{\boldsymbol{\eta}}_3$ and $\widehat{\boldsymbol{\eta}}_4$ were achieved from the RF method, while the biggest MSEs were achieved from the FC method. For canonical correlations, the lowest MSEs for $\phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ were achieved from the CM method, while the biggest MSE for $\phi(\hat{\rho}_1)$ was achieved from the RW method. The lowest MSE for $\phi(\hat{\rho}_3)$ was achieved from the RM method. The lowest MSE for $\phi(\hat{\rho}_4)$ was achieved from the RK method. The biggest MSEs for $(\hat{\rho}_2)$, $\phi(\hat{\rho}_3)$ and $\phi(\hat{\rho}_4)$ were achieved from the FC method.

From Table 2.8, for the dimensions $p = 4$ and $q = 4$, the data from $\mathbb{T}$ distribution shows that the lowest MSEs for $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\beta}}_4, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \widehat{\boldsymbol{\eta}}_3$ and $\widehat{\boldsymbol{\eta}}_4$ were achieved from the RK method, while the biggest MSEs were achieved from the CL method. For canonical correlations, the lowest MSEs for $\phi(\hat{\rho}_1)$, $\phi(\hat{\rho}_2)$ and $\phi(\hat{\rho}_4)$ were achieved from the RK method, while the biggest MSEs were achieved from the CL method. The lowest MSE for $\phi(\hat{\rho}_3)$ was achieved from the RS method, while the biggest MSE was achieved from the CL method.

From Table 2.9, for the dimensions $p = 4$ and $q = 4$, the data from the ACN shows that the lowest MSEs for $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\beta}}_4, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \widehat{\boldsymbol{\eta}}_3$ and $\widehat{\boldsymbol{\eta}}_4$ were achieved from the RM method, while the biggest MSEs were achieved from the CL method. For the canonical correlations, the lowest MSEs for $\phi(\hat{\rho}_1)$, $\phi(\hat{\rho}_2)$, $\phi(\hat{\rho}_3)$ and $\phi(\hat{\rho}_4)$ were achieved from the RM method. The biggest MSEs for $\phi(\hat{\rho}_1)$, $\phi(\hat{\rho}_2)$ and $\phi(\hat{\rho}_3)$ were achieved from the CL method and for $\phi(\hat{\rho}_4)$ was achieved from the FC method.

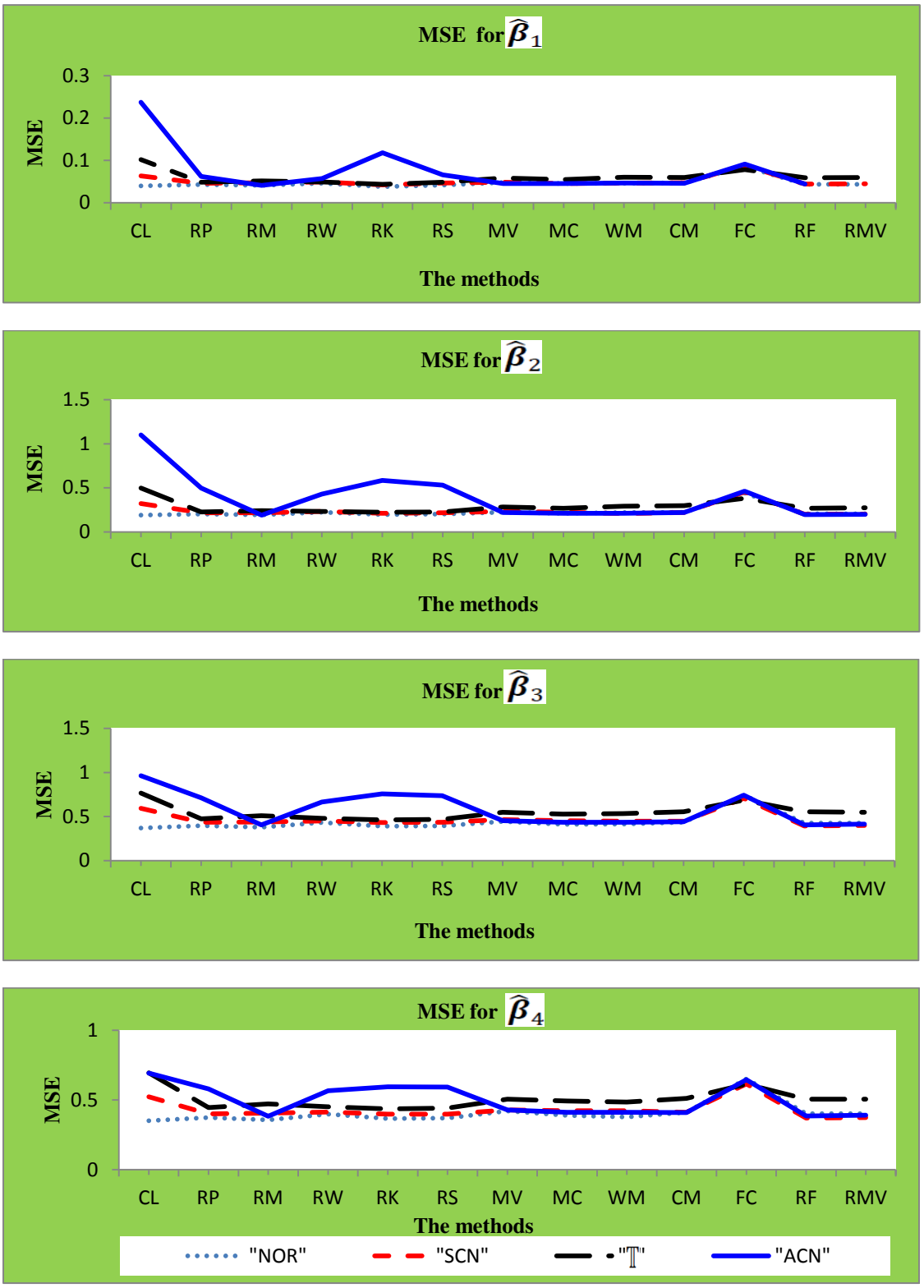


Figure 2.2. The MSEs for $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3$ and $\widehat{\boldsymbol{\beta}}_4$ for 13 estimators and under 4 sampling settings for $p = 4$ and $q = 4$.

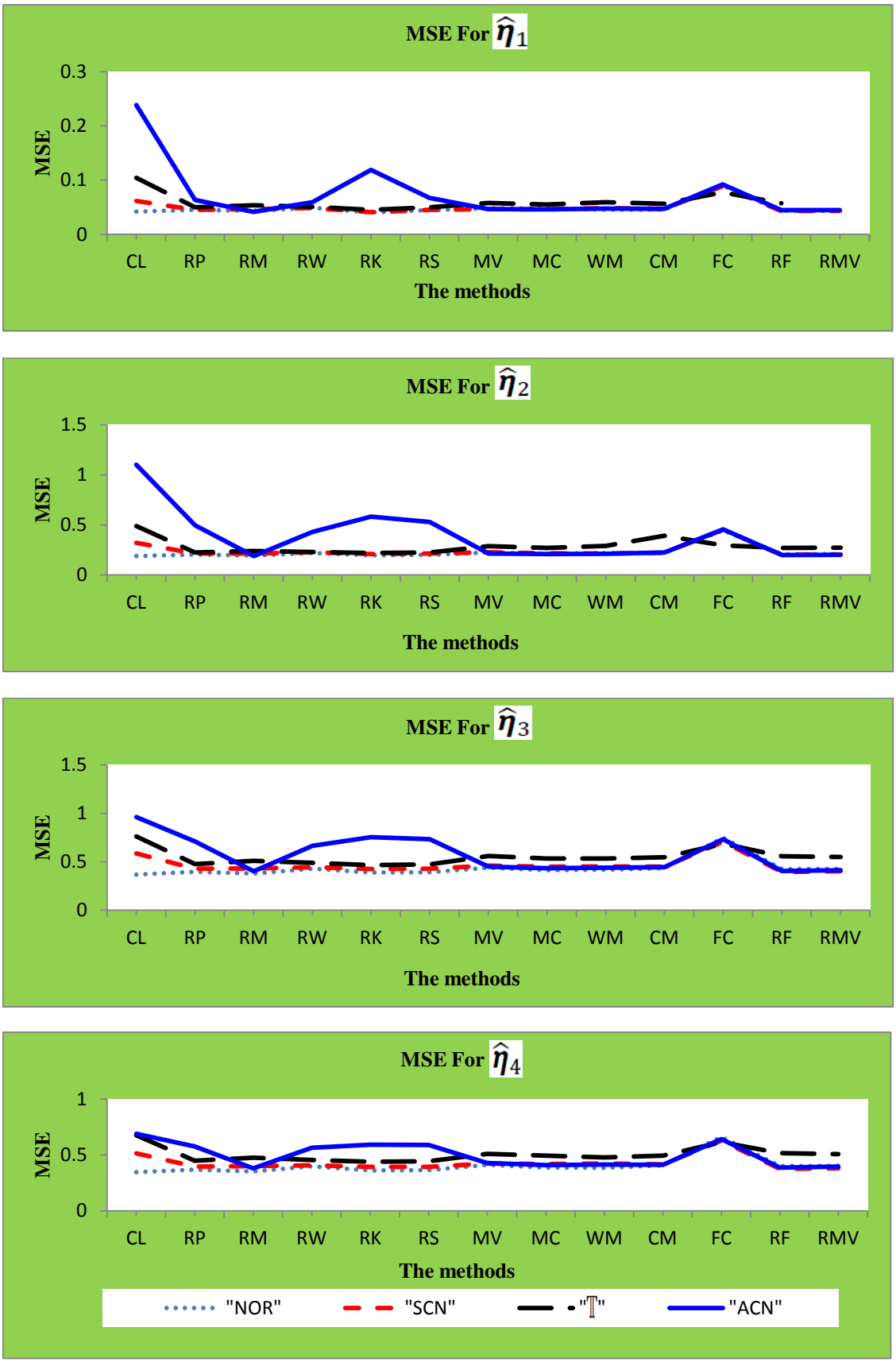


Figure 2.3. The MSEs for $\widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \widehat{\boldsymbol{\eta}}_3$ and $\widehat{\boldsymbol{\eta}}_4$ for 13 estimators and under 4 sampling settings for $p = 4$ and $q = 4$.

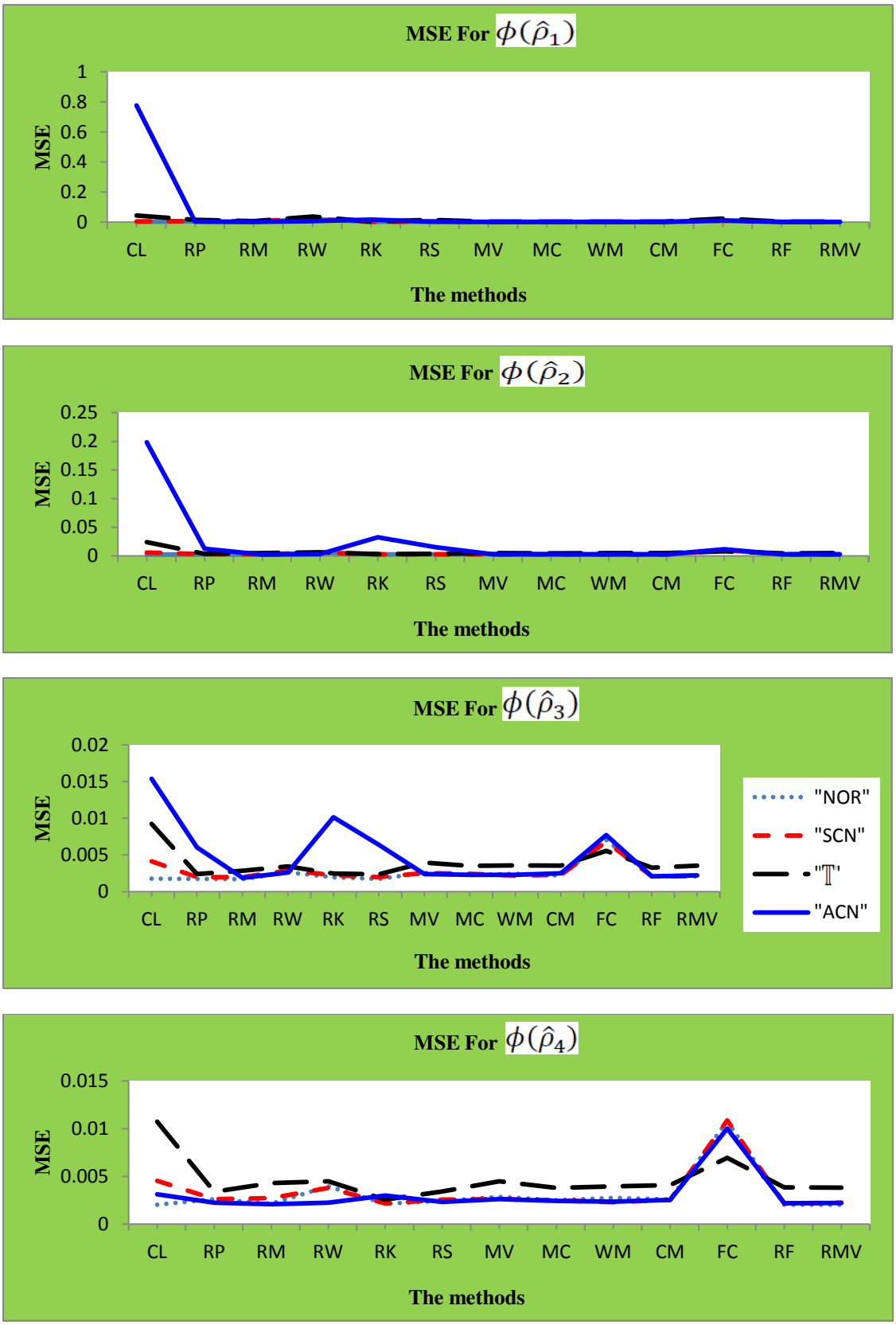


Figure 2.4. The MSEs for $\phi(\hat{\rho}_1), \phi(\hat{\rho}_2), \phi(\hat{\rho}_3)$ and $\phi(\hat{\rho}_4)$ for 13 estimators and under 4 sampling settings for $p = 4$ and $q = 4$.

Figures 2.2–2.4, show the MSEs for $\widehat{\boldsymbol{\beta}}_1$ to $\widehat{\boldsymbol{\beta}}_4$ and $\widehat{\boldsymbol{\eta}}_1$ to $\widehat{\boldsymbol{\eta}}_4$ and $\phi(\hat{\rho}_1)$ to $\phi(\hat{\rho}_4)$ for $p = 4$ and $q = 4$. In general, it is clear that ACN leads to the largest MSEs, followed by $\mathbb{T}$ distribution and SCN.

2.4. Breakdown plots

A simulation was carried out in order to study the robustness of the estimators, when considering outliers. We assumed two groups of variables. Each of them has three variables ($p = q = 3$) and the data was generated from $N_{p+q}(0, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}_{\mathbf{XX}} = \mathbf{I}_3$, $\boldsymbol{\Sigma}_{\mathbf{YY}} = \mathbf{I}_3$ and

$$\boldsymbol{\Sigma}_{\mathbf{XY}} = \begin{bmatrix} 0.90 & 0 & 0 \\ 0 & 0.50 & 0 \\ 0 & 0 & 0.33 \end{bmatrix}$$

The values of the contamination $\boldsymbol{\epsilon}$ were 0%, 5%, 10%, 15%, 20%, 25%, 30%, 35% and 40%, where $\boldsymbol{\epsilon}$ is the percentage of contamination. The contaminated observations were from the ACN distribution. We chose $n = 500$ and the MSEs were computed over $\mathcal{R} = 500$. The results are summarised in Figures 2.5–2.7.

In the figures, each line refers to different estimator. The breakdown plots show how the robustness of the estimator is under increasing $\boldsymbol{\epsilon}$.

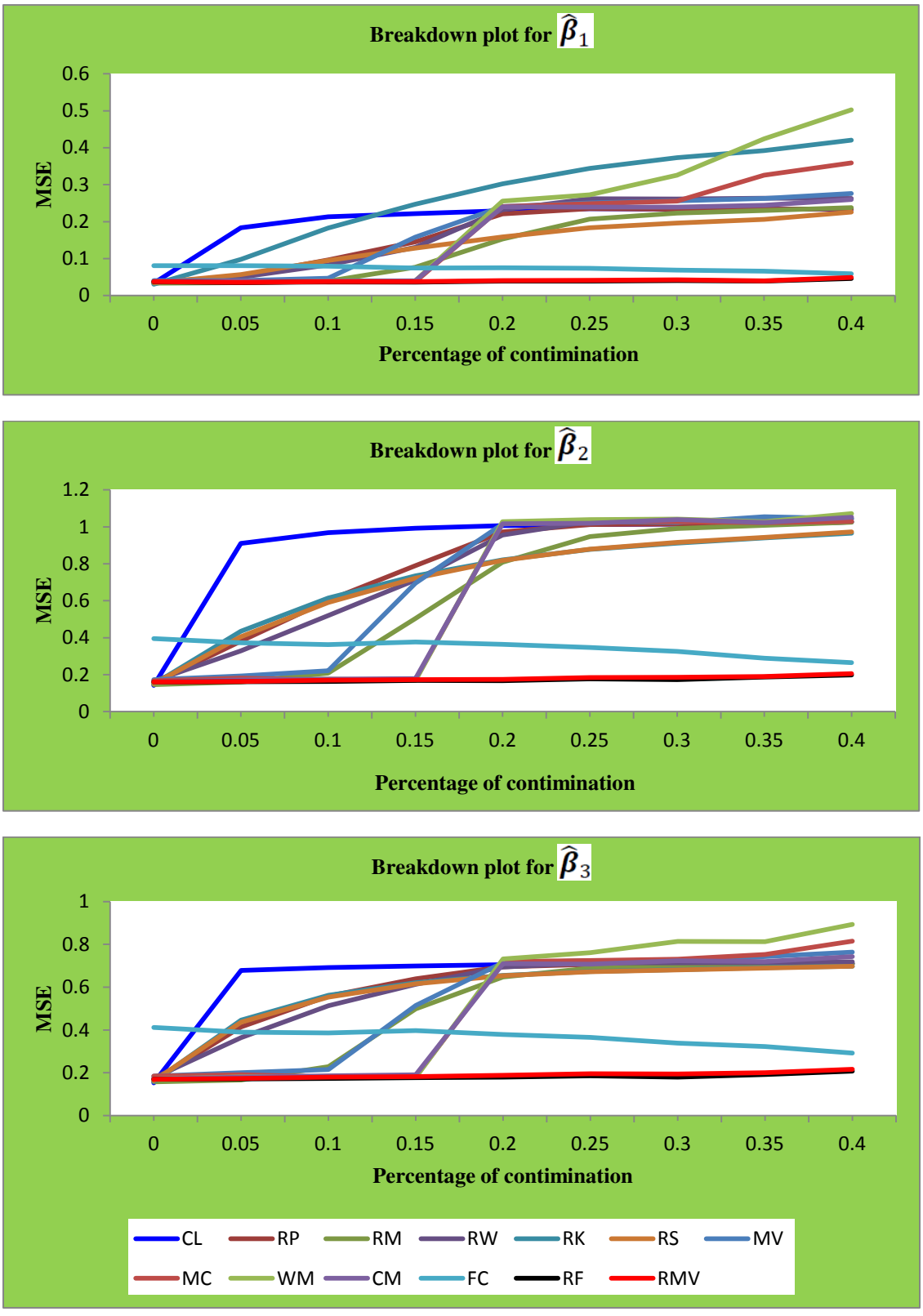


Figure 2.5. Breakdown plot: MSE for $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_3$ as a function of $\boldsymbol{\epsilon}$, from 0% to 40 %. The lines represent the different methods.

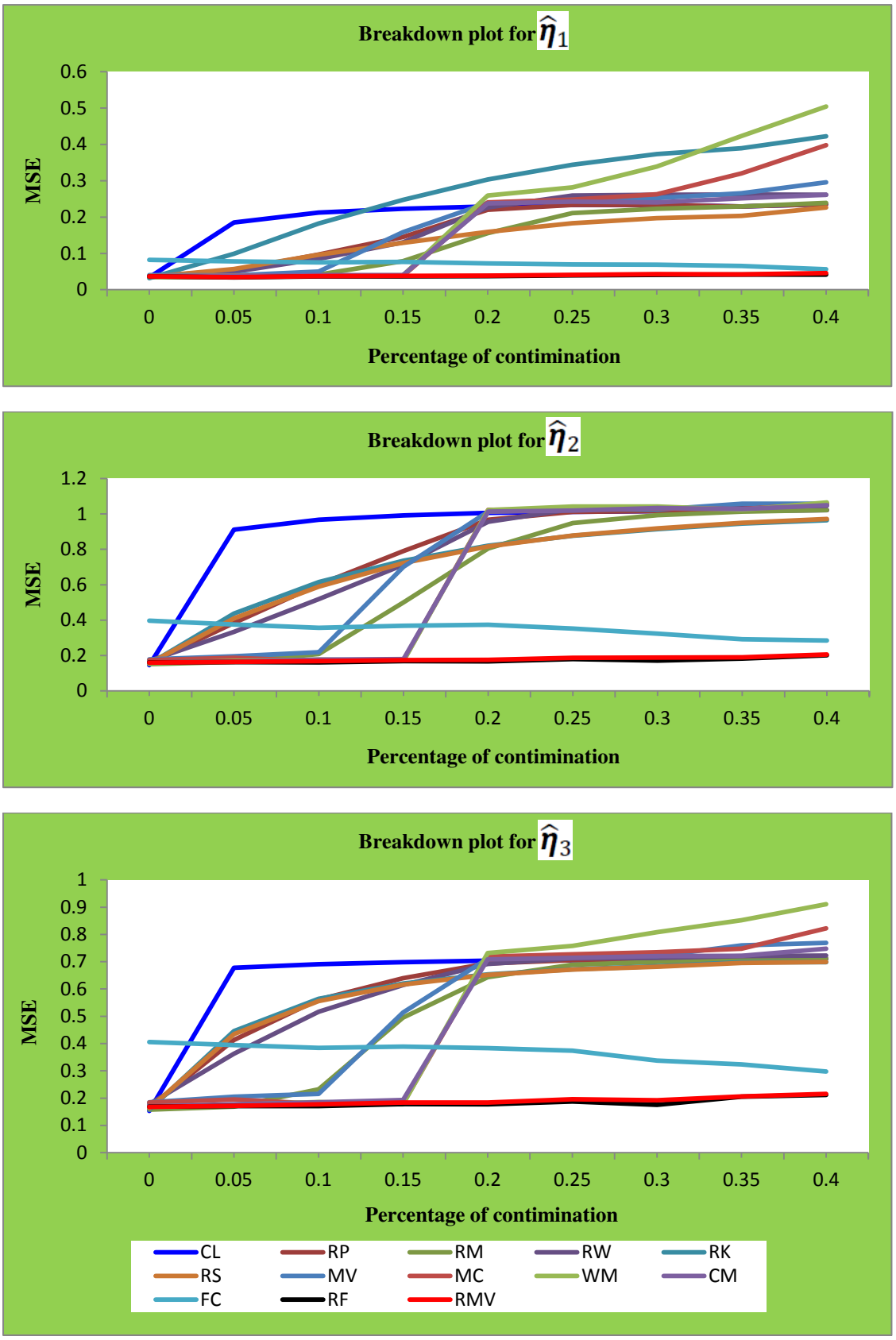


Figure 2.6. Breakdown plot: MSE for $\hat{\boldsymbol{\eta}}_1$, $\hat{\boldsymbol{\eta}}_2$ and $\hat{\boldsymbol{\eta}}_3$ as a function of $\boldsymbol{\epsilon}$, from 0% to 40 %. The lines represent the different methods.

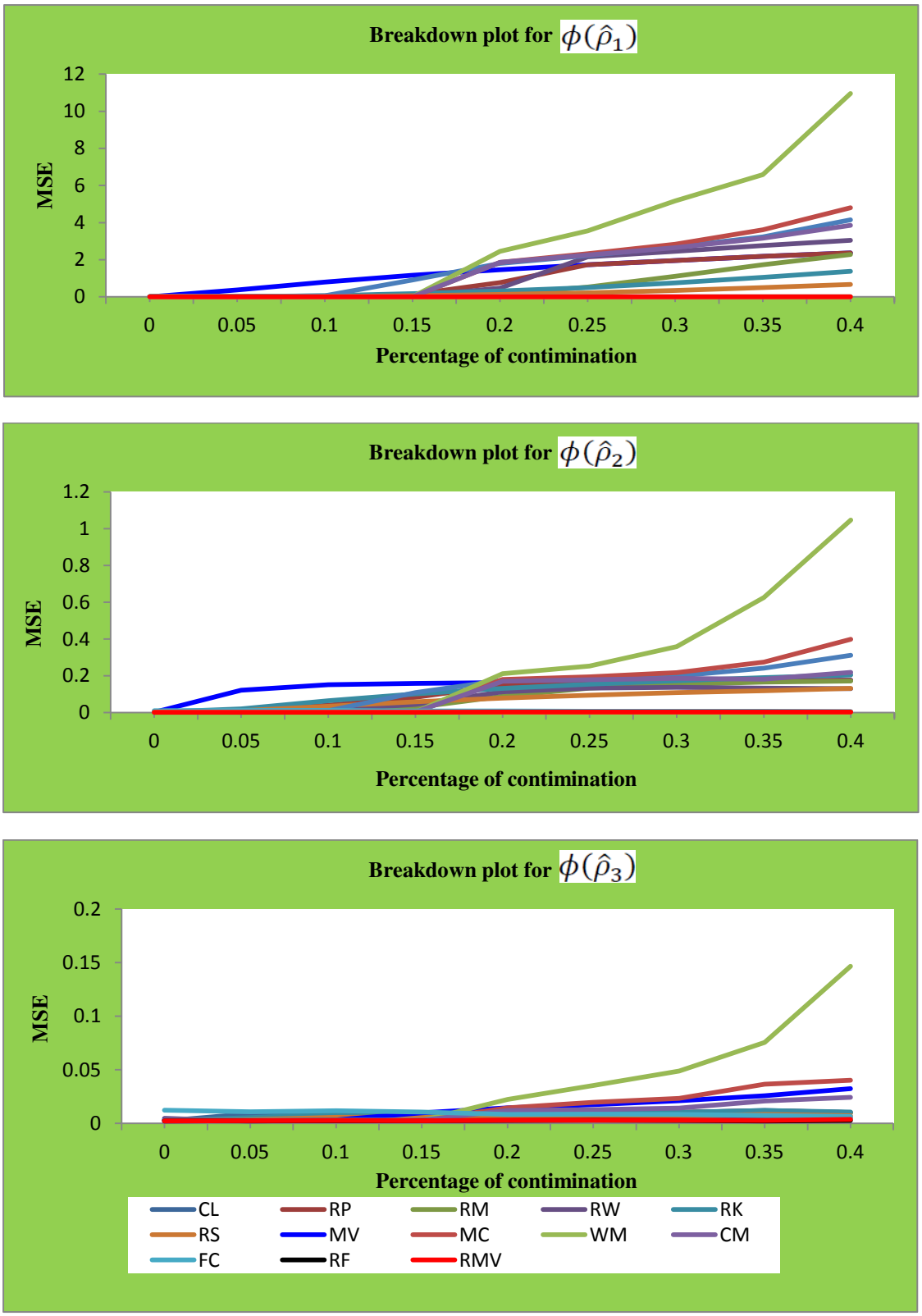


Figure 2.7. Breakdown plot: MSE for $\phi(\hat{\rho}_1), \phi(\hat{\rho}_2)$ and $\phi(\hat{\rho}_3)$ as a function of ϵ, from 0% to 40%. The lines represent the different methods.

Figure 2.5 and Figure 2.6 show the resistance of the MSE of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2$ and $\widehat{\boldsymbol{\beta}}_3$ and $\widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2$ and $\widehat{\boldsymbol{\eta}}_3$ for the different methods, respectively. It is obvious that the MSE of the CL quickly increases in the existence of outliers and the results in Figure 2.5 and Figure 2.6 confirm that the CL is very sensitive to the presence of the outliers. It is clear that the resistance of the methods based on RP, RW, RK and RS estimators, to the existence of outliers is less than that of other robust methods, where the performance of these estimators decreases as ϵ is increased beyond 5%. Similarly, it can be noted that the performance of RM and MV estimators reduces as the ϵ increases beyond 10%.

The performance of the methods based on the MC, WM and CM estimators become worst when ϵ is 15% or more, while that of the methods based on the RF and RMV estimators is still the best for all ϵ. The effectiveness of the method based on the FC estimator is better when the ϵ increases.

Figure 2.7 represents the breakdown plots for $\phi(\hat{\rho}_1), \phi(\hat{\rho}_2)$ and $\phi(\hat{\rho}_3)$. Generally, the MSEs become less for high order canonical correlations. The CL is very sensitive to the outliers and its performance is the worst out of all of the canonical correlations and at all ϵ. For the first canonical correlations, the performance of the method based on the RK estimator becomes worst when ϵ is 5% or more. The efficiency of the methods based on the RP, RS and MV estimators become worst when ϵ is 10% or more. The performance of the methods based on the RM and RW estimators become worst when ϵ is 15% or more. The effectiveness of the methods based on the MC, WM and CM estimators becomes worst when ϵ is 20% or more. The performance of the methods based on the RF and RMV estimators is still the best for all of ϵ, while the performance of the method based on the FC estimator becomes better as ϵ increases. For the second canonical correlation, the performance of the method based on the RW estimator

becomes worst when ϵ is 10% or more. The performance of the methods based on the RK, RP, RS, MV, RM, MC, WM, CM, RF, RMV and FC estimators is still similar to their performance in the case of the first canonical correlation. For the third canonical correlation, the effectiveness of the methods based on the RP and RK estimators becomes worst when ϵ is 25% or more. The performance of the method based on the RM estimator becomes worst when ϵ is 35% or more. The performance of the methods based on the RF, RMV, RW and RS estimators is still good for all the percentages of contamination. The performance of the methods based on the MV, MC, WM and CM estimators become worst when ϵ is 20% or more. The performance of the method based on the FC estimator becomes better when ϵ increases.

2.5. Tests of Independence

Assuming that $(\boldsymbol{X}, \boldsymbol{Y})$ is a multivariate normally distributed and set the independence hypothesis is given by

$$H_0: \boldsymbol{\Sigma_{XY}} = 0 \qquad \text{against} \qquad H_1: \boldsymbol{\Sigma_{XY}} \neq 0.$$

If the above H_0 holds, this means $H_0: \rho_1 = \cdots = \rho_p = 0$.

A simulation study was implemented in order to check the impact of the outliers in tests of independence. We assumed that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. After that, the frequency of rejecting H_0 at the 5% significance level is computed. We assumed that each of $\boldsymbol{X}$ and $\boldsymbol{Y}$ has two variables $p = q = 2$, $\boldsymbol{\Sigma_{XX}} = \boldsymbol{\Sigma_{YY}} = \mathbf{I}_2$ and $\boldsymbol{\Sigma_{XY}} = \text{Diag}(0.05, 0.01)$. Data from NOR, SCN and ACN have been generated. The CL, RP, RM, RW, RK and RS are considered.

The functions *pball, winall* and *spear,* which are available from Wilcox package at (http://www.unt.edu/rss/class/mike/Rallfun-v9_2.txt) have been used in order to conduct the test for the RP, RW and RS in Equations (2.10), (2.22) and (2.28), respectively. We have used the functions *bicorAndPvalue* from the package WGCNA and *Kendall* from the package Kendall in order to test RM in Equation (2.17) and the RK in Equation (2.25), respectively. P-values associated with the above functions have been calculated for $\mathcal{R} = 1000$ replications.

Table 2.10. The percentage of rejection H_0 in 1000 simulations.

	CL	RP	RM	RW	RK	RS
NOR	0.007	0.011	0.010	0.010	0.008	0.009
SCN	0.089	0.014	0.009	0.009	0.017	0.014
ACN	1.000	0.785	0.447	0.447	0.928	0.939

From Table 2.10, in the case of the NOR data, the test with the CL gave good results. In the case of SCN and ACN, the test with the RM and RW gave the best results. The test with the CL estimates was rejected in all 1000 simulations for the ACN data.

2.6. Real data

An example of CCA on the dataset of 3 psychological variables ($\boldsymbol{X}$), 4 academic variables (standardised test scores) and gender ($\boldsymbol{Y}$) for $n = 600$ students, which have been provided by Academic Technology Services (UCLA). The CCA example is available at (http://www.ats.ucla.edu/stat/R/dae/canonical.htm.) and the dataset is available at (http://www.ats.ucla.edu/stat/R/dae/mmreg.csv.). The first group of variables $\boldsymbol{X}$ are locus of control (x_1), self- notion (x_2) and stimulus (x_3). The second group of variables $\boldsymbol{Y}$ are standardised tests in reading (y_1), writing (y_2), maths (y_3) and

science (y_4). Additionally, the sex variable (y_5), where y_5=1 for a female student and y_5=0 for a male student. In our analysis, the categorical variable (sex) was excluded. The aim is to determine how $\boldsymbol{X}$ is related to $\boldsymbol{Y}$.

In the first case, we computed the canonical correlation methods based on the RM, FMCD, RFCH, RMVN and CL estimators with the above data. In the second case, we contaminated the data with 10% data from $\mathbb{T}$ distribution. Then, all the previous methods have been computed.

Table 2.11. $\hat{\rho}_1$, $\hat{\rho}_2$ and $\hat{\rho}_3$ for the non-contaminated and contaminated data.

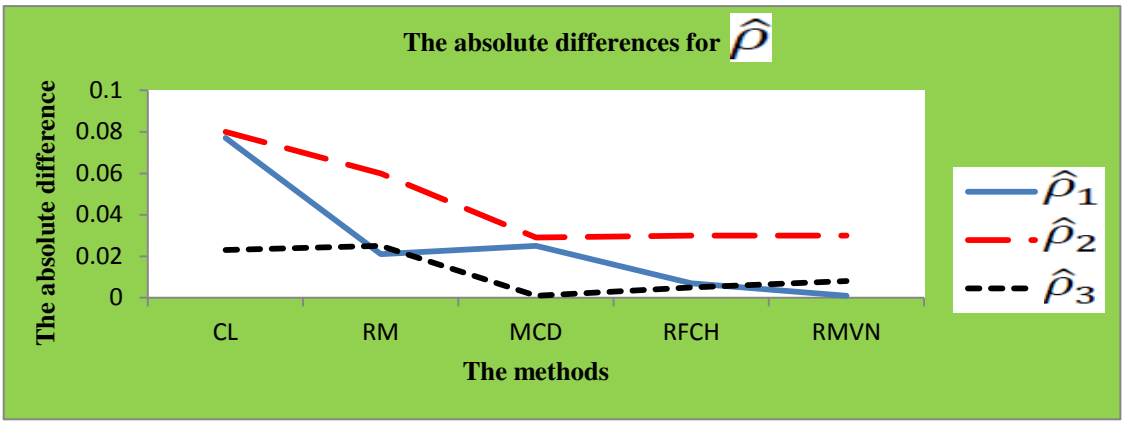
		$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
CL	No contamination	0.446	0.153	0.023
	10% contamination	0.369	0.073	0.046
	$\lvert\mathfrak{D}\rvert$	0.077	0.080	0.023
RM	No contamination	0.449	0.161	0.034
	10% contamination	0.428	0.101	0.009
	$\lvert\mathfrak{D}\rvert$	0.021	0.060	0.025
MCD	No contamination	0.469	0.168	0.036
	10% contamination	0.494	0.139	0.035
	$\lvert\mathfrak{D}\rvert$	0.025	0.029	0.001
RFCH	No contamination	0.459	0.167	0.034
	10% contamination	0.452	0.137	0.029
	$\lvert\mathfrak{D}\rvert$	0.007	0.030	0.005
RMVN	No contamination	0.462	0.174	0.034
	10% contamination	0.461	0.144	0.042
	$\lvert\mathfrak{D}\rvert$	0.001	0.030	0.008

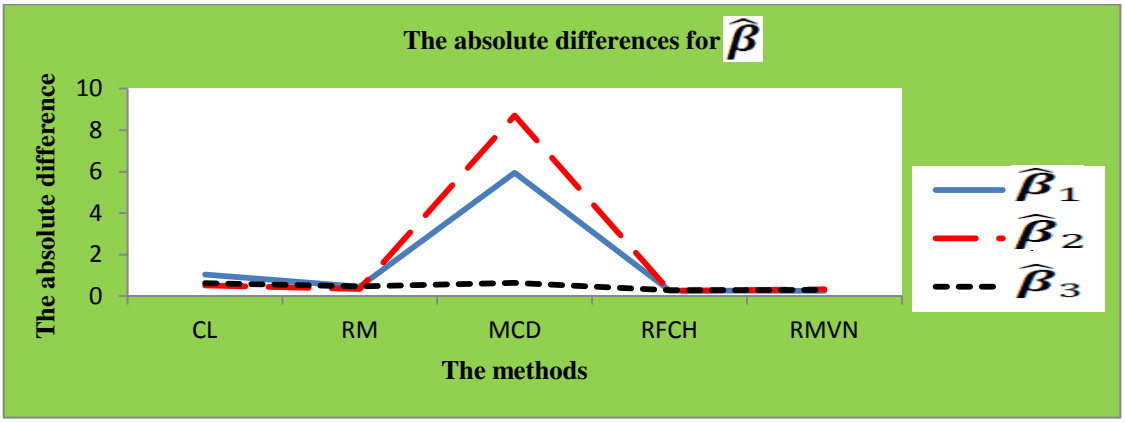
Table 2.12. $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\boldsymbol{\beta}}_2$ and $\widehat{\boldsymbol{\beta}}_3$ for the non-contaminated and contaminated data.

Method			x_1	x_2	x_3	$\sum \lvert\mathfrak{D}\rvert$
CL	$\widehat{\boldsymbol{\beta}}_1$	no contamination	-0.838	0.167	-0.428	
		10% contamination	-0.425	-0.081	-0.809	
		$\lvert\mathfrak{D}\rvert$	0.413	0.248	0.381	1.042
	$\widehat{\boldsymbol{\beta}}_2$	no contamination	0.513	0.594	-0.903	
		10% contamination	0.545	0.716	-0.548	
		$\lvert\mathfrak{D}\rvert$	0.032	0.122	0.355	0.509
	$\widehat{\boldsymbol{\beta}}_3$	no contamination	0.333	-0.850	-0.375	
		10% contamination	0.757	-0.714	-0.307	
		$\lvert\mathfrak{D}\rvert$	0.424	0.136	0.068	0.628
RM	$\widehat{\boldsymbol{\beta}}_1$	no contamination	-0.839	0.229	-0.428	
		10% contamination	-0.677	0.131	-0.617	
		$\lvert\mathfrak{D}\rvert$	0.162	0.098	0.189	0.449
	$\widehat{\boldsymbol{\beta}}_2$	no contamination	0.527	0.599	-0.861	
		10% contamination	0.519	0.739	-0.657	
		$\lvert\mathfrak{D}\rvert$	0.008	0.140	0.204	0.352
	$\widehat{\boldsymbol{\beta}}_3$	no contamination	-0.326	0.824	0.448	
		10% contamination	-0.592	0.704	0.536	
		$\lvert\mathfrak{D}\rvert$	0.266	0.120	0.088	0.474
MCD	$\widehat{\boldsymbol{\beta}}_1$	no contamination	-1.383	0.441	-1.285	
		10% contamination	1.323	-0.349	1.170	
		$\lvert\mathfrak{D}\rvert$	2.706	0.790	2.455	5.951
	$\widehat{\boldsymbol{\beta}}_2$	no contamination	0.798	1.114	-2.411	
		10% contamination	-0.763	-0.930	2.698	
		$\lvert\mathfrak{D}\rvert$	1.561	2.044	5.109	8.714
	$\widehat{\boldsymbol{\beta}}_3$	no contamination	-0.608	1.307	1.626	
		10% contamination	-0.457	1.395	1.226	
		$\lvert\mathfrak{D}\rvert$	0.151	0.088	0.400	0.639
RFCH	$\widehat{\boldsymbol{\beta}}_1$	no contamination	-1.259	0.348	-1.238	
		10% contamination	-1.215	0.251	-1.107	
		$\lvert\mathfrak{D}\rvert$	0.044	0.097	0.131	0.272
	$\widehat{\boldsymbol{\beta}}_2$	no contamination	0.756	1.041	-2.399	
		10% contamination	0.708	0.905	-2.474	
		$\lvert\mathfrak{D}\rvert$	0.048	0.136	0.075	0.259
	$\widehat{\boldsymbol{\beta}}_3$	no contamination	-0.551	1.242	1.408	
		10% contamination	-0.489	1.250	1.196	
		$\lvert\mathfrak{D}\rvert$	0.062	0.008	0.212	0.282
RMVN	$\widehat{\boldsymbol{\beta}}_1$	no contamination	-1.247	0.369	-1.292	
		10% contamination	-1.269	0.276	-1.149	
		$\lvert\mathfrak{D}\rvert$	0.022	0.093	0.143	0.258
	$\widehat{\boldsymbol{\beta}}_2$	no contamination	0.784	1.032	-2.404	
		10% contamination	0.747	0.954	-2.611	
		$\lvert\mathfrak{D}\rvert$	0.037	0.078	0.207	0.322
	$\widehat{\boldsymbol{\beta}}_3$	no contamination	-0.548	1.252	1.392	
		10% contamination	-0.489	1.326	1.225	
		$\lvert\mathfrak{D}\rvert$	0.059	0.074	0.167	0.300

Table 2.13. $\widehat{\boldsymbol{\eta}}_1$, $\widehat{\boldsymbol{\eta}}_2$ and $\widehat{\boldsymbol{\eta}}_3$ for the non-contaminated and contaminated data.

Method			y_1	y_2	y_3	y_4	$\sum\|\mathfrak{D}\|$
CL	$\widehat{\boldsymbol{\eta}}_1$	no contamination	-0.4450	-0.5358	-0.1827	0.0369	
		10% contamination	-0.5816	-0.4628	-0.0739	0.0953	
		$\|\mathfrak{D}\|$	0.1366	0.0730	0.1088	0.0584	0.3768
	$\widehat{\boldsymbol{\eta}}_2$	no contamination	-0.0161	-0.8794	-0.0278	1.2056	
		10% contamination	1.6120	-2.5241	0.4485	0.4895	
		$\|\mathfrak{D}\|$	1.6281	1.6447	0.4763	0.7161	4.4652
	$\widehat{\boldsymbol{\eta}}_3$	no contamination	-0.8924	0.9349	-0.8268	0.8589	
		10% contamination	2.0993	0.6399	-2.5597	-0.3026	
		$\|\mathfrak{D}\|$	2.9917	0.2950	1.7329	1.1615	6.1811
RM	$\widehat{\boldsymbol{\eta}}_1$	no contamination	-0.4282	-0.5806	-0.1214	0.0058	
		10% contamination	-0.4063	-0.6342	-0.1682	0.1304	
		$\|\mathfrak{D}\|$	0.0219	0.0536	0.0468	0.1246	0.2469
	$\widehat{\boldsymbol{\eta}}_2$	no contamination	-0.1154	-0.7279	-0.2124	1.3314	
		10% contamination	0.1018	-1.1584	0.1847	1.1768	
		$\|\mathfrak{D}\|$	0.2172	0.4305	0.3971	0.1546	1.1994
	$\widehat{\boldsymbol{\eta}}_3$	no contamination	0.8458	-1.0293	0.8655	-0.6467	
		10% contamination	1.6446	-0.3097	-0.9924	-0.4579	
		$\|\mathfrak{D}\|$	0.7988	0.7196	1.8579	0.1888	3.5651
MCD	$\widehat{\boldsymbol{\eta}}_1$	no contamination	-0.0496	-0.0594	-0.0078	0.0019	
		10% contamination	0.0502	0.0575	0.0002	0.0051	
		$\|\mathfrak{D}\|$	0.0998	0.1169	0.0080	0.0032	0.2279
	$\widehat{\boldsymbol{\eta}}_2$	no contamination	-0.0255	-0.0773	-0.0084	0.1403	
		10% contamination	0.0214	0.0953	-0.0228	-0.1213	
		$\|\mathfrak{D}\|$	0.0469	0.1726	0.0144	0.2616	0.4955
	$\widehat{\boldsymbol{\eta}}_3$	no contamination	0.0771	0.0356	-0.1635	0.0264	
		10% contamination	0.0301	0.0391	-0.1648	0.0727	
		$\|\mathfrak{D}\|$	0.0470	0.0035	0.0013	0.0463	0.0981
RFCH	$\widehat{\boldsymbol{\eta}}_1$	no contamination	-0.0426	-0.0613	-0.0085	0.0015	
		10% contamination	-0.0414	-0.0553	-0.0067	-0.0039	
		$\|\mathfrak{D}\|$	0.0012	0.0060	0.0018	0.0054	0.0144
	$\widehat{\boldsymbol{\eta}}_2$	no contamination	-0.0162	-0.0757	-0.0139	0.1345	
		10% contamination	-0.0072	-0.0971	0.0217	0.1066	
		$\|\mathfrak{D}\|$	0.0090	0.0214	0.0356	0.0279	0.0939
	$\widehat{\boldsymbol{\eta}}_3$	no contamination	0.1020	0.0137	-0.1464	0.0043	
		10% contamination	0.1014	-0.0069	-0.1384	0.0202	
		$\|\mathfrak{D}\|$	0.0006	0.0206	0.0080	0.0159	0.0451
RMVN	$\widehat{\boldsymbol{\eta}}_1$	no contamination	-0.0445	-0.0619	-0.0086	0.0046	
		10% contamination	-0.0449	-0.0573	-0.0020	-0.0069	
		$\|\mathfrak{D}\|$	0.0004	0.0046	0.0066	0.0115	0.0231
	$\widehat{\boldsymbol{\eta}}_2$	no contamination	-0.0216	-0.0716	-0.0131	0.1373	
		10% contamination	-0.0295	-0.0903	0.0253	0.1209	
		$\|\mathfrak{D}\|$	0.0079	0.0187	0.0384	0.0164	0.0814
	$\widehat{\boldsymbol{\eta}}_3$	no contamination	0.1032	0.0106	-0.1460	0.0059	
		10% contamination	0.0871	-0.0022	-0.1534	0.0432	
		$\|\mathfrak{D}\|$	0.0161	0.0128	0.0074	0.0373	0.0736





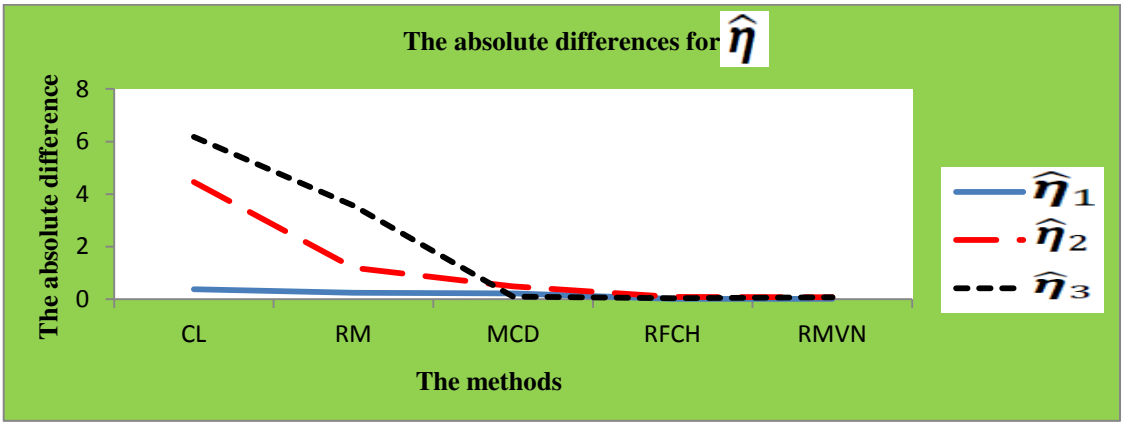


Figure 2.8. The first picture: the absolute differences for $\hat{\rho}_1, \hat{\rho}_2$ and $\hat{\rho}_3$ for the non-contaminated and contaminated data. The second picture: the absolute differences for $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_3$ for the non-contaminated and contaminated data. The third picture: the absolute differences for $\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2$ and $\hat{\boldsymbol{\eta}}_3$ for the non-contaminated and contaminated data.

Table 2.14. The computing time is measured in seconds for different estimation procedures for $\mathcal{R} = 500$ samples with size, $n = 500$.

	NOR 2×2	SCN2×2	$\mathbb{T}2 \times 2$	ACN2×2	NOR4×4	SCN4×4	$\mathbb{T}4 \times 4$	ACN4×4
RP	25	325	26	27	110	437	111	112
RM	27	327	28	29	89	418	91	100
RW	30	329	31	30	70	401	71	86
RK	81	381	81	81	284	618	285	282
RS	9	311	8	8	10	347	10	10
MV	107	411	106	106	256	611	256	249
MC	60	368	60	60	124	428	123	125
WM	63	386	63	63	124	474	124	129
CM	78	396	78	78	132	448	132	138
FC	16	345	16	16	19	341	19	21
RF	16	348	16	16	19	348	19	22
RMV	16	343	16	16	20	349	20	22

Table 2.15. The MSEs of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \phi(\hat{\rho}_1)$ and $\phi(\hat{\rho}_2)$ multiplied by 1000 for the FMCD method by using cov.mcd and covMcd functions when the data are from NOR, SCN, $\mathbb{T}$ and ACN, $p = 2$ and $q = 2$. The computing time is measured in seconds for $\mathcal{R} = 500$ samples with size, $n = 500$.

	NOR		SCN		$\mathbb{T}$		ACN	
	cov.mcd	covMcd	cov.mcd	covMcd	cov.mcd	covMcd	cov.mcd	covMcd
$\widehat{\boldsymbol{\beta}}_1$	27.78	27.24	28.59	26.53	32.12	32.41	29.58	28.29
$\widehat{\boldsymbol{\beta}}_2$	54.07	54.32	55.61	52.26	61.98	65.97	55.49	53.25
$\widehat{\boldsymbol{\eta}}_1$	28.78	27.58	28.34	28.08	32.30	36.39	28.29	26.37
$\widehat{\boldsymbol{\eta}}_2$	53.02	56.70	54.68	55.94	64.74	69.76	53.65	52.55
$\phi(\hat{\rho}_1)$	2.76	3.03	3.11	2.84	4.12	4.279	3.16	2.92
$\phi(\hat{\rho}_2)$	3.32	3.17	3.03	2.96	4.38	4.50	2.89	3.23
The compute time	702	60	1011	368	703	60	703	60

Table 2.16. The MSEs of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\beta}}_4, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2, \widehat{\boldsymbol{\eta}}_3, \widehat{\boldsymbol{\eta}}_4, \phi(\hat{\rho}_1), \phi(\hat{\rho}_2), \phi(\hat{\rho}_3)$ and $\phi(\hat{\rho}_4)$ multiplied by 1000 for the FMCD method using cov.mcd and covMcd functions when the data are from NOR, SCN, $\mathbb{T}$ and ACN, $p = 4$ and $q = 4$, and the computing time, measured in seconds, for $\mathcal{R} = 500$ samples with size, $n = 500$.

	NOR		SCN		$\mathbb{T}$		ACN	
	cov.mcd	covMcd	cov.mcd	covMcd	cov.mcd	covMcd	cov.mcd	covMcd
$\widehat{\boldsymbol{\beta}}_1$	45.66	45.89	45.64	44.99	54.79	58.92	45.23	46.31
$\widehat{\boldsymbol{\beta}}_2$	212.69	215.43	224.39	218.43	267.12	284.09	211.48	217.86
$\widehat{\boldsymbol{\beta}}_3$	412.24	444.68	454.15	437.79	528.93	547.40	436.61	433.56
$\widehat{\boldsymbol{\beta}}_4$	388.61	412.85	422.36	409.68	493.06	513.04	410.57	409.35
$\widehat{\boldsymbol{\eta}}_1$	47.65	46.05	45.97	44.78	54.83	59.34	46.06	45.16
$\widehat{\boldsymbol{\eta}}_2$	213.21	214.77	218.40	218.35	271.66	282.99	210.11	213.12
$\widehat{\boldsymbol{\eta}}_3$	412.51	437.16	449.79	445.43	533.95	543.88	433.29	428.15
$\widehat{\boldsymbol{\eta}}_4$	387.79	412.50	417.72	412.08	493.19	502.59	409.14	404.48
$\phi(\hat{\rho}_1)$	2.66	2.42	2.58	2.66	4.05	5.07	2.71	2.56
$\phi(\hat{\rho}_2)$	2.91	2.77	2.65	2.78	4.33	5.00	2.71	2.95
$\phi(\hat{\rho}_3)$	2.25	2.32	2.41	2.52	3.51	3.62	2.29	2.28
$\phi(\hat{\rho}_4)$	2.50	2.44	2.48	2.15	3.78	4.09	2.43	2.47
The compute time	1808	124	2152	428	1808	123	1800	125

The absolute differences $|\mathfrak{D}|$ between the values of $\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\eta}}_1, \widehat{\boldsymbol{\eta}}_2$ and $\widehat{\boldsymbol{\eta}}_3$ in the non contaminated and contaminated data have been used to measure the changes. From Tables 2.11–2.13 and Figure 2.8, we can observe that the results of the methods based on the RFCH and RMVN estimators are stable and less sensitive to outliers. However, the results of the method based on the RM and FMCD estimators are changeable and unstable. As expected, the results of the classical method were highly affected by the outliers.

Later, we took into account the computation time, along with robustness and efficiency of estimation. Table 2.14 shows the computation time, measured in seconds, for different estimation methods for $\mathcal{R} = 500$ samples with size $n = 500$. From this table, we can see that the computing time for the RS, FC, RF and RMV methods is

significantly lower than that of the other methods. Also, it is obvious that the MV, CM, WM and MC methods are time consuming.

From Tables 2.15 and 2.16, we can see that the *covMcd* estimator from the *roustbase library* is a much faster implementation of FMCD than *cov.mcd* from the *MASS library*, but the MSEs for the canonical coefficients and canonical correlations are larger in many cases.

2.7. Chapter Summary

In this chapter, a number of canonical correlations methods have been compared**.** From our simulation study and real data, we can conclude that the canonical vectors and correlations based on the RFCH and RMVN estimators perform better than the canonical vectors and correlations based on the FMCD estimator or the weighted FMCD estimator. Furthermore, from studying the breakdown plots of different estimators, we clearly observe that the effectiveness of the methods based on the RFCH and RMVN estimators is unrivalled for all percentages of contamination.

Moreover, from the ACN data the simulation study indicated that the performance of the canonical vectors and canonical correlation based on the RM is very promising; this fact is especially emphasised in the case when $p = q = 4$ than in that when $p = q = 2$. Additionally, the breakdown plot indicated that the canonical vectors and canonical correlations based on the RM estimator are higher than those of other M-type correlations. We also observed that although the breakdown plot showed that the FCH estimator had a high breakdown point, this estimator was one of the worst estimators for all cases.

From examining the simulation results of the study, we make a number of practical recommendations. Firstly, in the presence of outliers, we advise the usage of CCA based on the RFCH and RMVN estimators. Secondly, when the percentage of outliers is pre-determined to be less that 15%, we suggest the employment of CCA based on the RM estimators due to the fact that it has performed very well and that the computing time remains very reasonable. Thirdly, in the case of contamination above 20%, we do not recommend the usage of the FMCD estimators. Finally, we recommend the use of the *covMcd* function from the roustbase library to compute FMCD, if computation time is taken into account.

References

Anderson, T. W. (1999). Asymptotic Theory for Canonical Correlation Analysis. Journal of Multivariate Analysis 70, 1–29.

Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. 3rd ed., New Jersey, Wiley.

Beaghen, M.(1997). Canonical Variate Analysis and Related Methods with Longitudinadinal Data. Ph.D. Dissertation, Virginia Polytechnic Institute and State University. Available at (http://scholar.lib.vt.edu/theses/available/etd-11997-212717/unrestricted/etd.pdf).

Bernholt, T. and Fischer, P. (2004). The Complexity of Computing the MCD-Estimator. Theoretical Computer Science 326, 383–398.

Branco, J. A., Croux, C., Filzmoser, P. and Olivera, M. R. (2005). Robust canonical correlations: A comparative study. Computational Statistics 20, 203–229.

Cannon, A. J. and Hsieh, W. W. (2008). Robust nonlinear canonical correlation analysis: Application to seasonal climate forecasting. Nonlinear Processes in Geophysics 15, 221–232.

Das, S. and Sen, P. K. (1998). Canonical correlations. In Encyclopedia of Biostatistics 1, P. Armitage and T. Colton, eds., New York, Wiley, 468–482.

Devlin, S. J., Gnanadesikan, R. and Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. Journal of the American Statistical Association 76, 354–362.

Filzmoser, P., Dehon, C. and Croux, C. (2000). Outlier resistant estimators for canonical correlation analysis. In COMPSTAT: Proceedings in computational Statistics, J.G. Betlehem and P.G.M. van der Heijden, eds., Physica-Verlag, Heidelberg, 301–306.

Hotelling, H. (1936). Relations between two sets of variates. Biometrika 28, 321–377.

Hsu, P. L. (1941). On the limiting distribution of the canonical correlations. Biometrika 32, 38–45.

Jiao, J. and Jian, C. H. (2010). Asymptotic distributions in the projection pursuit based canonical correlation analysis. Science China press and Springer, Berlin, Heidelberg.

Johnson, R. A and Wichern, D. W. (2003). Applied Multivariate Statistical Analysis. New Jersey, Prentice-Hall.

Karnel, G. (1991). Robust canonical correlation and correspondence analysis. In The Frontiers of Statistical Scientific and Industrial Applications, Proceeding of ICOSCO-I, The First International Conference on Statistical Computing, Vol. II, P.R. Nelson, ed., Syracuse, New York, American Sciences Press, 335–354.

Kent, J. T. and Tyler, D. E. (1996). Constrained M-estimation for multivariate location and scatter. Annals of Statistics 24, 1346–1370.

Kettenring, J. R (1971). Canonical analysis of several sets of variables. Biometrika 58,433–451.

Kudraszow, N. L. and Maronna, R. A. (2011). Robust canonical correlation analysis: A predictive approach. Preprint submitted to Elsevier.

Mardia, K., Kent, J. and Bibby, J. (1979). Multivariate Analysis. New York, Academic press.

McCallum, B.T. (2008). Determinacy, learnability, and plausibility in monetary policy analysis: additional results. NBER Working Paper No 14164.

Myers, J. L. and Arnold D. W. (2003). Research Design and Statistical Analysis. 2nd ed., Mahwah, New Jersey, Lawrence Erlbaum, 1–508.

Olive, D. J. (2004). A resistant estimator of multivariate location and dispersion. Computational Statistics and Data Analysis 46, 99–102.

Olive, D. J. (2013). Robust Multivariate Analysis. Available at (http://www.math.siu.edu/olive/multbk.htm).

Olive, D. J. and Hawkins, D. M. (2010). Robust Multivariate Location and Dispersion. Available at (www.math.siu.edu/olive/pphbmld.pdf).

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. Annals of Statistics 24, 1327–1345.

Romanazzi, M. (1992). Influence in canonical correlation. Psychometrika 57, 237–259.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In Mathematical statistics and applications, Vol. B, eds. W. Grossman, G. P flug, I. Vincze, and W. Wertz, Dordrecht: Reidel, 283–297.

Rousseeuw, P. J. and Leroy, A. M. (1987). Robust regression and outlier detection. New York , Wiley.

Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Taskinen, S., Croux, C., Kankainen, A., Ollila, E. and Oja, H. (2006). Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. Journal of Multivariate Analysis 97, 359–384.

Wilcox, R. R. (1994). The percentage bend correlation coefficient. Psychometrika 59, 601–616.

Wilcox, R. R. (2005). Introduction to Robust Estimation and Hypothesis Testing. 2nd ed., San Diego, CA, Academic press.

Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. In A Festschrift for J. Neyman, F. N. David, ed., New York, Wiley, 411–444.

Zhang, J. and Olive, D. J. (2009). Applications of a Robust Dispersion Estimator. Available at (www.math.siu.edu/olive/pprcovm.pdf).

Zhang, J. (2011). Applications of a Robust Dispersion Estimator. Ph.D. Thesis, Southern Illinois University. Available at (www.math.siu.edu/olive/szhang.pdf).

Chapter 3

Sparse MAVE via the adaptive Lasso, SCAD and MCP penalties[2]

The well-known sufficient dimension reduction (SDR) methods supply a tool to find sufficient dimensions without needing to pre-specify a model or an error distribution. These methods replace the original p predictors with d-dimensional linear combinations (LC's) of predictors where $d < p$ without losing of any regression information. However, the explanation of the resulting estimates is not simple because each dimension reduction (DR) component is an LC of all the original predictors.

In this chapter, we propose to combine the shrinkage ideas of the adaptive Lasso, SCAD and MCP with the MAVE, to give sparse and precise solutions. The performance of the proposed methods is assessed by both simulation and real data analysis.

[2]This chapter is based on: Alkenani, A. and Yu, K. (2013). Sparse MAVE with oracle penalties. Advances and Applications in Statistics 34, 85–105. http://www.pphmj.com/abstract/7662.htm.

3.1. Introduction

In many statistical applications, statistical analysis is very complicated due to the dimension p of predictor vector $\mathbf{x}$ is large. A familiar approach that is used to cope with the high dimensional (HD) data in the regression model is to take down shorthand the dimension of the predictors without losing of any information and without the need for a pre-assigned parametric model. This has been obtained via the SDR.

The SDR theory (Cook, 1998) has been introduced to minimise the HD of the predictors, while keeping the regression information and making few assumptions. For regression models, assume y is a scalar response variable and $\mathbf{x} = \left(\mathrm{x}_1, \ldots, \mathrm{x}_p\right)^T$ is a $p \times 1$ predictor vector. The SDR investigates a $p \times d$ matrix $\mathbf{B}$, such that $y \perp\!\!\!\perp \mathbf{x} | \mathbf{x}^T\mathbf{B}$, where $\perp\!\!\!\perp$ refers to independence. The column space spanned by $\mathbf{B}$ is called the DR subspace. The intersection of all of the DR subspaces is called the central subspace if it is a DR subspace, which is denoted by $\mathrm{S}_{y|\mathbf{x}}$. Finding a $\mathrm{S}_{y|\mathbf{x}}$ is an essential goal in SDR because the $\mathrm{S}_{y|\mathbf{x}}$ contains all of the regression information of y, given $\mathbf{x}$. The dimension (d) of the $\mathrm{S}_{y|\mathbf{x}}$ is called the structural dimension (Yu and Zhu, 2013). Knowledge of the $\mathrm{S}_{y|\mathbf{x}}$ is beneficial to answer the question, "how does the distribution of $y|\mathbf{x}$ alter with the value of $\mathbf{x}$?". Various approaches have been proposed to estimate $\mathrm{S}_{y|\mathbf{x}}$. For example, the SIR method (Li, 1991), SAVE method (Cook and Weisberg, 1991), pHd method (Li, 1992), see Cook (1998) for more details.

In many situations, the regression analysis focuses on deducing the conditional mean of the dependent variable that is given to the explanatory variables. Cook and Li (2002) presented the idea of the Central Mean Subspace, where a natural inferential object is used for DR when the mean function is of attention. Also, the authors proposed the Iterative Hessian Transformation (IHT) method. There are a number of

DR approaches have been prposed to estimate $S_{E(y|\mathbf{x})}$, for example the IHT (Cook and Li, 2002) and the MAVE method (Xia et al., 2002). However, all of the SDR methods suffer because each DR component is an LC of all of the explanatory variables, therefore making it very difficult to explain the resulting estimates. As mentioned in Section 1.1, the variable selection is very important in building a multiple regression model. The choice of an appropriate subset of predictors can help to develop prediction accuracy. Also, the interpretation of a smaller subset of predictors is often easier to understand and interpret than a large subset of predictors in practice. Variable selection using the regularisation methods in the ordinary least squares has attracted considerable research interest. For example, the Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), fused Lasso (Tibshirani et al., 2005), adaptive Lasso (Zou, 2006) and MCP (Zhang, 2010).

Under the framework of the SDR, the work of Li et al. (2005) has produced good outcomes. For example, Ni et al. (2005) suggested a shrinkage SIR; Li and Nachtsheim (2006) proposed the sparse SIR method; and Li (2007) unified the inverse DR procedures to obtain sparse SDR. Zhou and He (2008) suggested constrained canonical correlation (CCC). The CCC method uses CANCOR method which is reported in (Fung et al., 2002) with an l_1 norm constraint. However, Fung et al. (2002) demonstrated that CANCOR method is based on the matrix of SIR; thus the CCC can be considered as an alternative method to that used in Li (2007). The major focus of the methods mentioned concentrates on the distribution of $\mathbf{x}|y$ without assuming any specific model. However, these methods do need particular assumptions on $\mathbf{x}$, such as the linearity condition. Li and Yin (2008) suggested a regularised SIR method to adapt SIR to deal with the cases when $n < p$ and highly correlated covariates. Wang and Yin (2008) suggested adding Lasso penalty to the MAVE loss function in order to get

sparse MAVE (SMAVE) estimate. Fan and Li (2001) considered a number of regularisation methods. The authors stated that a good penalty function should have three properties, namely unbiasedness, sparsity and continuity. Fan and Li (2001) conjectured that the (oracle properties) OP's do not hold for the Lasso.

In this chapter, extensions for SMAVE (Wang and Yin, 2008) are proposed. We combine the DR method MAVE (Xia et al., 2002) with the regularisation methods SCAD (Fan and Li, 2001), adaptive Lasso (Zou, 2006) and the MCP (Zhang, 2010). The proposed methods have merits over the SMAVE and the sparse sliced inverse regression method (SSIR) (Li, 2007) because the proposed methods use penalisation which benefits from OP's, while SMAVE and SSIR use Lasso which does not. Also, the proposed methods have advantages over SSIR in that these methods do not need any certain distribution on $\mathbf{x}$ and are able to estimate the dimensions in the conditional mean function (CMF).

The remainder of the chapter is arranged as follows. In Section 3.2, a brief review of SDR for the mean function and MAVE method is given. SMAVE method is reviewed in Section 3.3. Sparse MAVE with adaptive Lasso penalty, SCAD and MCP penalties are introduced in Sections 3.4, 3.5 and 3.6, respectively. Simulation studies are conducted in Section 3.7. The methods have been applied to two sets of real data in Section 3.8. Finally, the conclusions are summarised in Section 3.9.

3.2. SDR for the mean function and MAVE

For regression problems with a scalar response variable y on a $p \times 1$ predictor vector $\mathbf{x}$ assume the following model:

$$y = f(\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_p) + \varepsilon, \tag{3.1}$$

where $f(\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_p) = E(y|\mathbf{x})$ and $E(\varepsilon\,|\mathbf{x}) = 0$. The aim of SDR for the mean function is to explore a subset S of the predictor space such that

$$y \perp\!\!\!\perp E(y|\mathbf{x})|P_S\mathbf{x}, \tag{3.2}$$

where $\perp\!\!\!\perp$ denotes statistical independence and $P_{(.)}$ refers to a projection operator. Subspaces satisfying (3.2) are called mean DR subspaces (Cook and Li, 2002). Thus if $d = dim(S)$ and $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_d)$ is a basis for S, $\mathbf{x}$ can be substituted by LC's $\mathbf{x}^T\boldsymbol{\beta}_1, \mathbf{x}^T\boldsymbol{\beta}_2, \dots, \mathbf{x}^T\boldsymbol{\beta}_d$, $d \leq p$ without losing any information on the CMF. That is, $f(\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_p) = f(\mathbf{x}^T\mathbf{B})$. If the intersection of all subspaces satisfies (3.2), this is called the central mean subspace (CMS) (Cook and Li, 2002) and is denoted by $S_{E(y|\mathbf{x})}$. $S_{E(y|\mathbf{x})}$ is assumed existent over this chapter. Several methods are available for estimating $S_{E(y|\mathbf{x})}$ and one of these methods is the MAVE (Xia et al., 2002). The MAVE is described in detail, as follows:

Xia et al. (2002) proposed the MAVE such that the matrix $\mathbf{B}$ is the solution of

$$\min_{\mathbf{B}}\{\mathrm{E}[y - \mathrm{E}(y|\mathbf{x}^T\mathbf{B})]^2\}, \tag{3.3}$$

where $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$. The conditional variance given $\mathbf{x}^T\mathbf{B}$ is

$$\boldsymbol{\sigma}^2_{\mathbf{B}}(\mathbf{x}^T\mathbf{B}) = \mathrm{E}[\{y - \mathrm{E}(y|\mathbf{x}^T\mathbf{B})\}^2\,|\mathbf{x}^T\mathbf{B}]. \tag{3.4}$$

Thus,

$$\min_{\mathbf{B}} \mathrm{E}[y - \mathrm{E}(y|\mathbf{x}^T\mathbf{B})]^2 = \min_{\mathbf{B}} \mathrm{E}\{\boldsymbol{\sigma}^2_{\mathbf{B}}(\mathbf{x}^T\mathbf{B})\}. \tag{3.5}$$

For any given $\mathbf{x}_0$, $\boldsymbol{\sigma}^2_{\mathbf{B}}(\mathbf{x}^T\mathbf{B})$ can be locally approximated as follows

$$\boldsymbol{\sigma}_{\mathbf{B}}^{2}(\mathbf{x}_0^T\mathbf{B}) \approx \sum_{i=1}^{n}\{y_i - \mathrm{E}(y_i|\mathbf{x}_i^T\mathbf{B})\}^2\,\omega_{i0}$$

$$\approx \textstyle\sum_{i=1}^{n}[y_i - \{a_0 + (\mathbf{x}_i - \mathbf{x}_0)^T\mathbf{B}\,\boldsymbol{b}_0\}]^2\,\omega_{i0},$$

where $a_0 + (\mathbf{x}_i - \mathbf{x}_0)^T\mathbf{B}\,\boldsymbol{b}_0$ is the local linear expansion of $\mathrm{E}(y_i|\mathbf{x}_i^T\mathbf{B})$ at $\mathbf{x}_0$, and $\omega_{i0} \geq 0$ are the kernel weights centred at $\mathbf{x}_0^T\mathbf{B}$ with $\sum_{i=1}^{n}\omega_{i0} = 1$. So the problem of finding $\mathbf{B}$ is the same as solving the following minimisation:

$$\min_{\mathbf{B}:\,\mathbf{B}^T\mathbf{B} = \mathbf{I}_m}\left(\textstyle\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - \left\{a_j + \left(\mathbf{x}_i - \mathbf{x}_j\right)^T\mathbf{B}\,\boldsymbol{b}_j\right\}\right]^2\omega_{ij}\right). \qquad (3.6)$$

3.3. The SMAVE method

Wang and Yin (2008) suggested the SMAVE method. They add an l_1 penalty to the loss function of MAVE in (3.6) to produce a sparse estimate. The authors proposed SMAVE minimises:

$$\textstyle\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - \left\{a_j + \left(\mathbf{x}_i - \mathbf{x}_j\right)^T\mathbf{B}\,\boldsymbol{b}_j\right\}\right]^2\omega_{ij} + \lambda\sum_{k=1}^{p}|\boldsymbol{\beta}_{m,k}|, \qquad (3.7)$$

for $m = 1, \ldots, d$.

They assumed that d is known, then suggested that they could estimate d according to a new version of the Bayesian information criterion (BIC). The algorithm for SMAVE is as follows:

1. Initialise $m = 1$, and set $\mathbf{B} = \boldsymbol{\beta}_0$, any arbitrary $p \times 1$ vector.

2. For a given $\mathbf{B}$, obtain $(a_j, \boldsymbol{b}_j)$ where $j = 1, \ldots, n$, by solving the following problem:

$$\min_{a_j,\boldsymbol{b}_j,j=1,\ldots,n}\left(\textstyle\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - \left\{a_j + \left(\mathbf{x}_i - \mathbf{x}_j\right)^T\mathbf{B}\,\boldsymbol{b}_j\right\}\right]^2\omega_{ij}\right). \qquad (3.8)$$

3. For a given $(\hat{a}_j, \hat{\boldsymbol{b}}_j)$, $j = 1, \ldots, n$, solve $\boldsymbol{\beta}_{mLasso}$ from the following minimisation:

$$\min_{\mathbf{B}:\,\mathbf{B}^T\mathbf{B}=\mathbf{I}_m}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i-\left\{\hat{a}_j+\left(\mathbf{x}_i-\mathbf{x}_j\right)^T\left(\widehat{\boldsymbol{\beta}}_1,\widehat{\boldsymbol{\beta}}_2,\ldots,\widehat{\boldsymbol{\beta}}_{m-1},\boldsymbol{\beta}_m\right)\widehat{\boldsymbol{b}}_j\right\}\right]^2\omega_{ij}+\lambda\sum_{k=1}^{p}\left|\boldsymbol{\beta}_{m,k}\right|\right). \quad (3.9)$$

4. Replace the mth column of $\mathbf{B}$ by $\widehat{\boldsymbol{\beta}}_{mLasso}$ and repeat steps 2 and 3 until convergence.

5. Update $\mathbf{B}$ by $\left(\widehat{\boldsymbol{\beta}}_{1Lasso},\widehat{\boldsymbol{\beta}}_{2Lasso},\ldots,\widehat{\boldsymbol{\beta}}_{mLasso},\boldsymbol{\beta}_0\right)$, and set m to be $m+1$.

6. If $m<d$, continue steps 2 to 5 until $m=d$.

Wang and Yin (2008) used the same refined multidimensional Gaussian Kernel that was proposed by Xia et al. (2002) for MAVE

$$\omega_{ij}=K_h\left\{\left(\mathbf{x}_i-\mathbf{x}_j\right)^T\widehat{\mathbf{B}}\right\}/\sum_{i=1}^{n}K_h\left\{\left(\mathbf{x}_i-\mathbf{x}_j\right)^T\widehat{\mathbf{B}}\right\},$$

and the optimal bandwidth selected in order to minimise the mean integrated squared errors. Also, they used the Gaussian product kernel and $h_{opt}=\text{Ⱥ}(d)n^{-1/(4+d)}$, where $\text{Ⱥ}(d)=\left\{\frac{4}{(d+2)}\right\}^{1/(4+d)}$, where d is the dimension of the kernel function.

3.4. Sparse MAVE with adaptive Lasso penalty (ALMAVE)

Fan and Li (2001) considered a number of regularisation methods and one of these methods is the Lasso. The authors explained that the Lasso produces biased estimates for the large coefficients. Consequently, the Lasso does not have the OP's. As an extension for Lasso, Zou (2006) proposed the adaptive Lasso. The idea of the adaptive Lasso is to allow the penalization for the coefficients of different predictors by using adaptive weights. The authors proved that the OP's are achieved for the adaptive Lasso. Zou (2006) defined the adaptive Lasso minimises

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i{}^T\boldsymbol{\beta})^2 + \lambda\sum_{k=1}^{p}\widetilde{\omega}_k|\beta_k|, \tag{3.10}$$

where $\lambda > 0$ is the tuning parameter. The weights are set to be $\widetilde{\omega}_k = 1/|\tilde{\beta}_k|^{\delta}$, $k = 1, \dots, p$, $\tilde{\beta}$ is a non-penalised regression estimate and $\delta > 0$.

The ALMAVE has been proposed as follows:

$$\min_{\mathbf{B}}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - \left\{a_j + (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{B}\,\boldsymbol{b}_j\right\}\right]^2\omega_{ij} + \lambda\sum_{k=1}^{p}\widetilde{\omega}_k|\boldsymbol{\beta}_{m,k}|\right), \tag{3.11}$$

for $m = 1, \dots, d$.

The algorithm for the ALMAVE is similar to the algorithm in Section 3.3, except in step 3, for a given $(\hat{a}_j, \hat{\boldsymbol{b}}_j)$, $j = 1, \dots, n$, solve $\boldsymbol{\beta}_{mALasso}$ from the following problem:

$$\min_{\mathbf{B}:\,\mathbf{B}^T\mathbf{B} = \mathbf{I}_m}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - \left\{\hat{a}_j + (\mathbf{x}_i - \mathbf{x}_j)^T(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \dots, \hat{\boldsymbol{\beta}}_{m-1}, \boldsymbol{\beta}_m)\,\hat{\boldsymbol{b}}_j\right\}\right]^2\omega_{ij} + \lambda\sum_{k=1}^{p}\widetilde{\omega}_k\,|\boldsymbol{\beta}_{m,k}|\right), \tag{3.12}$$

and then we follow the same steps of the algorithm in Section 3.3.

3.5. Sparse MAVE with SCAD penalty (SCADMAVE)

Fan and Li (2001) suggested the SCAD penalty. The authors proved that the SCAD estimator has the OP's. The SCAD penalty Fan and Li (2001) defined on $[0, \infty)$ is given by

$$p_{SCAD\,\lambda,c}(\Theta) = \begin{cases} \lambda\Theta & if\ \ \Theta \le \lambda \\ \frac{c\lambda\Theta - 0.5(\Theta^2 + \lambda^2)}{c-1} & if\ \ \lambda < \Theta \le c\,\lambda \\ \frac{\lambda^2(c+1)}{2} & if\ \ \Theta > c\lambda, \end{cases} \tag{3.13}$$

and its first derivative is given by

$$p'_{SCAD\ \lambda}(\Theta) = \begin{cases} \lambda & if\ \ \Theta \le \lambda \\ \frac{c\lambda - \Theta}{c-1} & if\ \ \lambda < \Theta \le c\,\lambda \\ 0 & if\ \ \Theta > c\lambda, \end{cases} \tag{3.14}$$

where $c > 2$ and $\lambda \geq 0$ are tuning parameters.

The SCAD penalised regression minimises

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i{}^T\boldsymbol{\beta})^2 + n\sum_{k=1}^{p} p_{SCAD\ \lambda,c}(|\beta_k|). \tag{3.15}$$

The SCADMAVE has been suggested as follows:

$$\min_{\mathbf{B}} \left(\sum_{j=1}^{n}\sum_{i=1}^{n} \left[y_i - \left\{ a_j + (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{B}\, \boldsymbol{b}_j \right\} \right]^2 \omega_{ij} + n\sum_{k=1}^{p} p_{SCAD\ \lambda,c}\left(|\boldsymbol{\beta}_{m,k}|\right) \right). \tag{3.16}$$

The algorithm for the SCADMAVE is similar to the algorithm in Section 3.3, except in step 3, for a given $(\hat{a}_j, \hat{\boldsymbol{b}}_j)$, $j = 1, \ldots, n$, solve $\boldsymbol{\beta}_{mSCAD}$ from the following problem:

$$\min_{\mathbf{B}:\, \mathbf{B}^T\mathbf{B} = \mathbf{I}_m} \left(\sum_{j=1}^{n}\sum_{i=1}^{n} \left[y_i - \left\{ \hat{a}_j + (\mathbf{x}_i - \mathbf{x}_j)^T (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_{m-1}, \boldsymbol{\beta}_m)\, \hat{\boldsymbol{b}}_j \right\} \right]^2 \omega_{ij} + n\sum_{k=1}^{p} p_{SCAD\ \lambda,c}\left(|\boldsymbol{\beta}_{m,k}|\right) \right), \tag{3.17}$$

and then we follow the same steps of the algorithm in Section 3.3.

3.6. Sparse MAVE with MCP penalty (MCPMAVE)

Zhang (2010) proposed a minimax concave penalty (MCP). It supplies the convexity of the penalised loss in sparse areas to a great extent, given particular thresholds for variable selection and unbiasedness.

The MCP Zhang (2010) defined on $[0, \infty)$ is given by

$$p_{MCP\ \lambda,c}(\Theta) = \begin{cases} \lambda\Theta - \frac{\Theta^2}{2\,c} & if\ \ \Theta \le c\,\lambda \\ \frac{1}{2} c\,\lambda^2 & if\ \ \Theta > c\lambda, \end{cases} \tag{3.18}$$

and its first derivative is given by

$$p'_{MCP\ \lambda,c}(\Theta) = \begin{cases} \lambda - \frac{\Theta}{c} & if\ \ \Theta \le c\,\lambda \\ 0 & if\ \ \Theta > c\lambda, \end{cases} \tag{3.19}$$

where $c > 1$ and $\lambda \ge 0$ are tuning parameters. The logic behind the $p_{MCP\ \lambda,c}(\Theta)$ can be understood through $p'_{MCP\ \lambda,c}(\Theta)$. The MCP starts with the rate of penalization (ROP) equivalent to that in the Lasso, but continuously reduces that penalization until $\Theta > c\lambda$ and the ROP goes down to 0.

The MCP penalised regression minimises

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i{}^T\boldsymbol{\beta})^2 + n\sum_{k=1}^{p} p_{MCP\ \lambda,c}(|\beta_k|). \tag{3.20}$$

The MCPMAVE has been suggested as follows:

$$\min_{\mathbf{B}} \left(\sum_{j=1}^{n}\sum_{i=1}^{n} \left[y_i - \left\{ a_j + (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{B}\, \boldsymbol{b}_j \right\} \right]^2 \omega_{ij} + n\sum_{k=1}^{p} p_{MCP\ \lambda,c}\left(|\boldsymbol{\beta}_{m,k}|\right) \right). \tag{3.21}$$

The algorithm for the MCPMAVE is similar to the algorithm in Section 3.3, except in step 3, for a given $(\hat{a}_j, \widehat{\boldsymbol{b}}_j)$, $j = 1, \ldots, n$, $\boldsymbol{\beta}_{mMCP}$ can be obtained by solving the following problem:

$$\min_{\mathbf{B}:\,\mathbf{B}^T\mathbf{B}=\mathbf{I}_m}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i-\left\{\hat{a}_j+\left(\mathbf{x}_i-\mathbf{x}_j\right)^T\left(\widehat{\boldsymbol{\beta}}_1,\widehat{\boldsymbol{\beta}}_2,\ldots,\widehat{\boldsymbol{\beta}}_{m-1},\boldsymbol{\beta}_m\right)\widehat{\boldsymbol{b}}_j\right\}\right]^2\omega_{ij}+\right.$$

$$\left.n\sum_{k=1}^{p}p_{MCP\,\lambda,c}\left(\left|\boldsymbol{\beta}_{m,k}\right|\right)\right), \tag{3.22}$$

and then we follow the same steps of the algorithm in Section 3.3.

3.7. A simulation study

In this section, we demonstrate the behaviour of the suggested methods using many simulation examples and some typical examples are given below:

Example 1: $\mathcal{R}=200$ datasets were generated with size $n=200$ from the model $y=\frac{(\mathbf{x}^T\boldsymbol{\beta}_1)}{\{0.5+(\mathbf{x}^T\boldsymbol{\beta}_2+1.5)^2\}}+0.2\,\varepsilon$, where $\mathbf{x}=(\mathrm{x}_1,\ldots,\mathrm{x}_{10})^T$, x_i and ε are independent and are identically distributed from an $N(0,1)$, $\boldsymbol{\beta}_1=(1,0,\ldots,0)^T$ and $\boldsymbol{\beta}_2=(0,1,0,\ldots,0)^T$ with $S_{E(y|\mathbf{X})}=span(\mathbf{B}_2)$. This means, the model is $y=\frac{\mathrm{x}_1}{\{0.5+(\mathrm{x}_2+1.5)^2\}}+0.2\,\varepsilon$.

Example 2: $\mathcal{R}=200$ data sets were generated with size $n=60$ and 120 from the linear model $y=\mathbf{x}^T\boldsymbol{\beta}+0.5\,\varepsilon$, where $\mathbf{x}=(\mathrm{x}_1,\ldots,\mathrm{x}_{24})^T$, x_i and ε are independent and are identically distributed from an $N(0,1)$ and $\boldsymbol{\beta}=(1,1,1,0,\ldots,0)^T$ with $S_{E(y|\mathbf{X})}=span(\mathbf{B}_1)$. In order to evaluate the performance of the proposed methods when the predictors are correlated, we generate $\mathbf{x}$ from a $N(0,\boldsymbol{\Sigma})$ with $cov(\mathrm{x}_i,\mathrm{x}_j)=0.5^{|i-j|}$ for this model. This means, the model is $y=\mathrm{x}_1+\mathrm{x}_2+\mathrm{x}_3+0.5\,\varepsilon$.

Example 3: $\mathcal{R} = 200$ datasets were generated with size $n = 200$ observations from the model $y = \text{sign}\,(\mathbf{x}^T\boldsymbol{\beta}_1)\log\,(|\mathbf{x}^T\boldsymbol{\beta}_2 + 5|) + 0.2\,\varepsilon$, where $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_{20})^T$, x_i and ε are independent and are identically distributed from an $N(0,1)$. There are three different forms for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, namely:

(1) $\boldsymbol{\beta}_1 = (1,1,1,1,0, \ldots,0)^T$ and $\boldsymbol{\beta}_2 = (0, \ldots,0,1,1,1,1)^T$.

(2) $\boldsymbol{\beta}_1 = (1,1,0.1,0.1,0, \ldots,0)^T$ and $\boldsymbol{\beta}_2 = (0, \ldots,0,0.1,0.1,1,1)^T$.

(3) $\boldsymbol{\beta}_1 = (1, \ldots,1,0, \ldots,0)^T$ and $\boldsymbol{\beta}_2 = (0, \ldots,0,1, \ldots,1)^T$, where each $\boldsymbol{\beta}$ has 10 elements equal to 1 with $S_{E(y|\mathbf{x})} = span(\mathbf{B}_2)$. This means, the models are

$y = \text{sign}\,(\mathrm{x}_1 + \mathrm{x}_2 + \mathrm{x}_3 + \mathrm{x}_4)\log\,(|\mathrm{x}_{17} + \mathrm{x}_{18} + \mathrm{x}_{19} + \mathrm{x}_{20} + 5|) + 0.2\,\varepsilon$ for case(1).

$y = \text{sign}\,(\mathrm{x}_1 + \mathrm{x}_2 + 0.1\mathrm{x}_3 + 0.1\mathrm{x}_4)\log\,(|0.1\mathrm{x}_{17} + 0.1\mathrm{x}_{18} + \mathrm{x}_{19} + \mathrm{x}_{20} + 5|) + 0.2\,\varepsilon$ for case(2).

$y = \text{sign}\,(\mathrm{x}_1 + \mathrm{x}_2 + \cdots + \mathrm{x}_{10})\log\,(|\mathrm{x}_{11} + \mathrm{x}_{12} + \cdots + \mathrm{x}_{20} + 5|) + 0.2\,\varepsilon$ for case(3).

Example 4: $\mathcal{R} = 200$ datasets were generated with size $n = 60$ and 120 from the linear model $y = \mathbf{x}^T\boldsymbol{\beta} + 0.5\,\varepsilon$, where $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_{24})^T$, x_i and ε are independent and are identically distributed from $N(0,1)$, $\boldsymbol{\beta} = (1, \ldots,1)^T$ with $S_{E(y|\mathbf{x})} = span(\mathbf{B}_1)$. This means, the model is $y = \mathrm{x}_1 + \mathrm{x}_2 + \cdots + \mathrm{x}_{24} + 0.5\,\varepsilon$.

Example 5: $\mathcal{R} = 200$ data sets were generated with size $n = 200$ from the linear model $y = \mathbf{x}^T\boldsymbol{\beta} + 0.5\,\varepsilon$, where $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_{100})^T$, x_i and ε are independent and are identically distributed from an $N(0,1)$ and $\boldsymbol{\beta} = (1,1,1,0, \ldots,0)^T$with $S_{E(y|\mathbf{X})} = span(\mathbf{B}_1)$. In order to evaluate the performance of the proposed methods when the predictors are correlated, we generate $\mathbf{x}$ from a $N(0, \boldsymbol{\Sigma})$ with $cov(\mathrm{x}_i, \mathrm{x}_j) = 0.5^{|i-j|}$ for this model. This means, the model is $y = \mathrm{x}_1 + \mathrm{x}_2 + \mathrm{x}_3 + 0.5\,\varepsilon$.

After we write the first term in Equations (3.12), (3.17) and (3.22) in the least squares form, we use the functions (adalasso) from Package 'parcor' (Kraemer and Schaefer, 2012), ncvreg(, penalty=c("SCAD")) and ncvreg(, penalty=c("MCP")) from Package 'ncvreg' in R (Breheny and Huang, 2011) to do the computations in Equations (3.12), (3.17) and (3.22), respectively. To evaluate the precision of the estimation, we compute the average number of zero coefficients (Ave 0's), mean and standard deviation (SD) of the absolute correlation $|r_m|$ between $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_m$ and $\boldsymbol{X}^T\boldsymbol{\beta}_m$ and the mean and SD of the mean squared error (MSE), $E\left(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_m - \boldsymbol{X}^T\boldsymbol{\beta}_m\right)^2$.

Table 3.1. Simulation results for the methods which are studied based on the model in Example 1.

	$\widehat{\boldsymbol{\beta}}_1$					$\widehat{\boldsymbol{\beta}}_2$				
Method	Ave 0's	$\lvert r_1\rvert$ Mean	$\lvert r_1\rvert$ SD	MSE Mean	MSE SD	Ave 0's	$\lvert r_2\rvert$ Mean	$\lvert r_2\rvert$ SD	MSE Mean	MSE SD
SMAVE	5.67	0.9516	0.0497	0.0005	0.0005	2.33	0.8090	0.1266	0.0060	0.0080
ALMAVE	8.67	0.9791	0.0422	0.0003	0.0001	8.00	0.9840	0.0338	0.0003	0.0006
SCADMAVE	5.00	0.9619	0.0425	0.0004	0.0004	2.67	0.8511	0.1103	0.0028	0.0065
MCPMAVE	5.00	0.9590	0.0479	0.0004	0.0005	2.33	0.8263	0.1146	0.0058	0.0075

Table 3.2. Simulation results for the methods which are studied based on the model in Example 2.

	Independent predictors					Correlated predictors				
Method	Ave 0's	$\lvert r\rvert$ Mean	$\lvert r\rvert$ SD	MSE Mean	MSE SD	Ave 0's	$\lvert r\rvert$ Mean	$\lvert r\rvert$ SD	MSE Mean	MSE SD
n=60										
SMAVE	12.67	0.9796	0.0075	0.0155	0.0400	12.00	0.9479	0.1085	0.0119	0.0088
ALMAVE	19.33	0.9918	0.0074	0.0147	0.0360	17.00	0.9866	0.0348	0.0112	0.0074
SCADMAVE	12.00	0.9919	0.0074	0.0149	0.0363	8.00	0.9866	0.0349	0.0111	0.0074
MCPMAVE	12.33	0.9920	0.0100	0.0157	0.0380	8.33	0.9865	0.0350	0.0111	0.0088
n=120										
SMAVE	19.50	0.9899	0.0047	0.0050	0.0100	17.00	0.9934	0.0022	0.0062	0.0087
ALMAVE	21.00	0.9969	0.0031	0.0043	0.0065	20.75	0.9988	0.0007	0.0057	0.0081
SCADMAVE	19.50	0.9956	0.0046	0.0044	0.0065	17.50	0.9982	0.0015	0.0060	0.0082
MCPMAVE	19.50	0.9956	0.0049	0.0045	0.0066	17.75	0.9978	0.0019	0.0061	0.0085

Table 3.3. Simulation results for the methods which are studied based on the model in Example 3.

	$\widehat{\boldsymbol{\beta}}_1$					$\widehat{\boldsymbol{\beta}}_2$				
Method	Ave 0's	$\lvert r\rvert$ Mean	$\lvert r\rvert$ SD	MSE Mean	MSE SD	Ave 0's	$\lvert r\rvert$ Mean	$\lvert r\rvert$ SD	MSE Mean	MSE SD
Case(1)										
SMAVE	9.25	0.9760	0.0071	0.0062	0.0087	3.50	0.8603	0.1045	0.0095	0.0096
ALMAVE	15.00	0.9938	0.0029	0.0061	0.0080	15.75	0.9641	0.0882	0.0045	0.0028
SCADMAVE	10.25	0.9924	0.0062	0.0062	0.0082	3.75	0.8674	0.1075	0.0088	0.0076
MCPMAVE	10.25	0.9934	0.0043	0.0062	0.0080	3.75	0.8729	0.0889	0.0086	0.0074
Case(2)										
SMAVE	11.00	0.9798	0.0078	0.0009	0.0012	2.00	0.6903	0.2194	0.0036	0.0066
ALMAVE	16.00	0.9954	0.0024	0.0007	0.0008	12.75	0.8049	0.1202	0.0015	0.0032
SCADMAVE	11.00	0.9920	0.0055	0.0008	0.0010	2.00	0.7348	0.1213	0.0016	0.0038
MCPMAVE	11.00	0.9918	0.0059	0.0009	0.0011	2.00	0.6972	0.2359	0.0039	0.0069
Case(3)										
SMAVE	2.50	0.9159	0.0408	0.0192	0.0165	2.50	0.9313	0.0398	0.0239	0.0355
ALMAVE	5.75	0.9409	0.0360	0.0179	0.0145	6.50	0.9545	0.0309	0.0231	0.0354
SCADMAVE	2.25	0.9189	0.0360	0.0190	0.0163	2.50	0.9352	0.0309	0.0238	0.0355
MCPMAVE	2.25	0.9158	0.0365	0.0192	0.0163	2.50	0.9332	0.0337	0.0240	0.0356

According to the mean and SD of the $|r_m|$ between the $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_m$ and $\boldsymbol{X}^T\boldsymbol{\beta}_m$ and the mean and SD of the MSE, $E\left(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_m - \boldsymbol{X}^T\boldsymbol{\beta}_m\right)^2$. From Tables 3.1, 3.2, 3.3 and 3.5, it can be seen that the ALMAVE and SCADMAVE show a better performance than the other methods for all cases under consideration, except in Example 3, case (1) where the ALMAVE and MCPMAVE were the best two methods among all of the methods.

Table 3.4. Simulation results for the methods which are studied based on the model in Example 4.

Method	Ave 0's	$\|r\|$ Mean	$\|r\|$ SD	MSE Mean	MSE SD
n=60					
SMAVE	0.00	0.9959	0.0101	0.0564	0.0645
ALMAVE	0.00	0.9466	0.1895	0.0587	0.0653
SCADMAVE	0.00	0.9958	0.0100	0.0563	0.0645
MCPMAVE	0.00	0.9958	0.0100	0.0563	0.0645
n=120					
SMAVE	0.00	0.9976	0.0008	0.0619	0.0557
ALMAVE	0.00	0.9976	0.0009	0.0619	0.0559
SCADMAVE	0.00	0.9975	0.0008	0.0619	0.0557
MCPMAVE	0.00	0.9975	0.0008	0.0619	0.0557

Table 3.5. Simulation results for the methods which are studied based on the model in Example 5.

	Independent predictors					Correlated predictors				
Method	Ave 0's	$\|r\|$ Mean	$\|r\|$ SD	MSE Mean	MSE SD	Ave 0's	$\|r\|$ Mean	$\|r\|$ SD	MSE Mean	MSE SD
SMAVE	87.83	0.9736	0.0073	0.0019	0.0014	80.00	0.9881	0.0024	0.0036	0.0037
ALMAVE	96.83	0.9979	0.0014	0.0016	0.0016	97.00	0.9985	0.0004	0.0031	0.0043
SCADMAVE	87.50	0.9743	0.0073	0.0019	0.0014	80.00	0.9888	0.0022	0.0035	0.0042
MCPMAVE	87.67	0.9741	0.0072	0.0019	0.0014	79.50	0.9881	0.0023	0.0036	0.0042

Also, we can see from Table 3.4 that the SCADMAVE and MCPMAVE have a better performance than the other methods. In general, this shows that the ALMAVE, SCADMAVE and MCPMAVE produce more accurate estimates and these methods are more efficient than the SMAVE method.

It can be seen that in all of the examples, the proposed methods produce a lower MSE and a bigger $|r|$ than the SMAVE method. The variations in the ALMAVE, SCADMAVE and MCPMAVE estimates are approximately similar in the majority of cases and are less than the variations in the estimate of the SMAVE method.

3.8. Real data

To explain the performance of the methods which are studied in this chapter we use two data sets, namely air pollution (AP) data and body fat (BF) data.

3.8.1. Air pollution (AP) data

In this section, we illustrated the methods via an analysis of the AP data. The data contains $n = 500$ observations. The dataset is available from the website (http://lib.stat.cmu.edu/datasets/NO2.dat). The response y is the logarithm (LOG) values of the concentration of Nitrogen dioxide per hour measured in the period from 10/2001 to 08/2003. The seven predictors are the LOG of the number of cars /hour (x_1), temperature 2m above ground (x_2), wind velocity (x_3), the temperature difference between 25 and 2m above ground (x_4), wind trend (x_5), hour of the day (x_6) and day number from 01/10/2001 (x_7).

Table 3.5. The values of the adjusted R-squared for the model fit based on the AP data.

		SMAVE	ALMAVE	SCADMAVE	MCPMAVE
Model fit	Linear	0.76	0.93	0.76	0.76
	Quadratic	0.90	0.94	0.90	0.90
	Cubic	0.93	0.94	0.93	0.93
	Quartic	0.93	0.94	0.93	0.93

Table 3.6. The prediction error of the cubic fit for the methods which are studied based on the AP data.

Method	Prediction error
SMAVE	0.7768
ALMAVE	0.6692
SCADMAVE	0.7740
MCPMAVE	0.7741

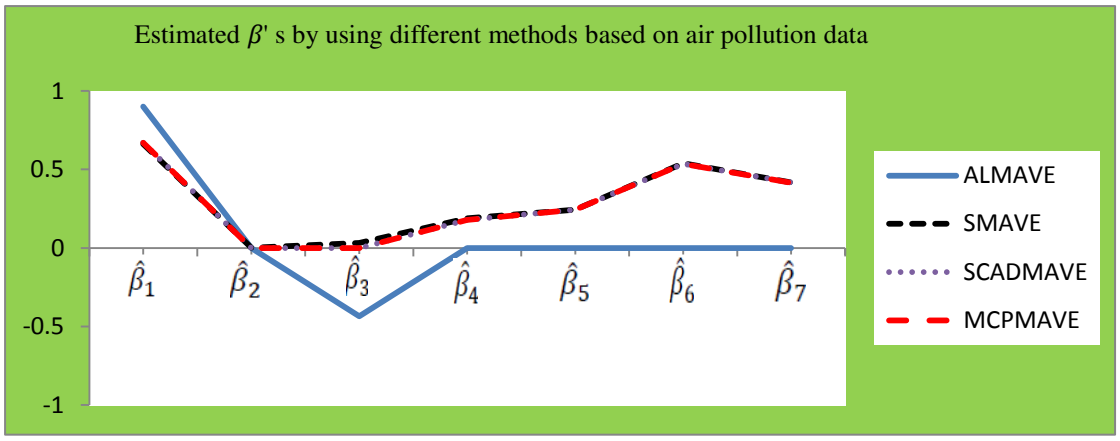


Figure 3.1. A plot explaining the estimated coefficients $\hat{\beta}$'s which are estimated by the different methods based on the AP data.

Table 3.5 reports the values of the adjusted R-squared for the model fit, based on the AP data for all the studied methods. The studied methods discover a nonlinear structure, which can be approximated by a cubic fit. Also, it can be seen that the adjusted R-squared is slightly larger than the SMAVE (Wang and Yin, 2010) for the ALMAVE method (adjusted R-squared= 0.94) and it is similar to the SMAVE for the other methods (adjusted R-squared= 0.93).

Table 3.6 presents the prediction error of the cubic fit for the methods which are studied based on the AP data. It is clear that the ALMAVE, SCADMAVE and MCPMAVE methods have a lower prediction error than the SMAVE method. This means that these methods show a better performance than the SMAVE method.

From Figure 3.1, it can be seen that the estimated cofficients for the SMAVE, SCADMAVE and MCPMAVE methods were approximatly similar, which could be because these methods have the same value for the adjusted R-squared.

3.8.2. Body fat (BF) data

Percentage of BF is a substantial gauge of health. It can be precisely estimated by underwater weighing methods. These methods oftentimes need particular tools and are at times not available, so fitting the percentage of BF to simple measurements of body is an appropriate path to predict the BF. Johnson (1996) presented a dataset in which the percentage of BF and 13 simple measurements about the body (like weight, height and abdomen circumference) were recorded for 252 men. The data set is available in the package ('mfp') in R. The response variable y is the percent of BF (%) . The predictors are the age (x_1), the weight (x_2), the height (x_3), the neck circumference (x_4), the chest circumference (x_5), the abdomen circumference (x_6), the hip circumference (x_7), the thigh circumference (x_8), the knee circumference (x_9), the ankle circumference (x_{10}), the extended biceps circumference (x_{11}), the forearm circumference (x_{12}) and the wrist circumference (x_{13}).

Table 3.7. The values of the adjusted R-squared for the model fit based on the BF data

		SMAVE	ALMAVE	SCADMAVE	MCPMAVE
Model fit	linear	0.92	0.92	0.92	0.92
	Quadratic	0.95	0.95	0.95	0.95
	Cubic	0.96	0.96	0.96	0.96
	Quartic	0.96	0.96	0.96	0.96

Table 3.8. The prediction error of the cubic fit for the methods which are studied based on the BF data.

Method	Prediction error
SMAVE	24.4095
ALMAVE	22.6263
SCADMAVE	23.5089
MCPMAVE	23.0635

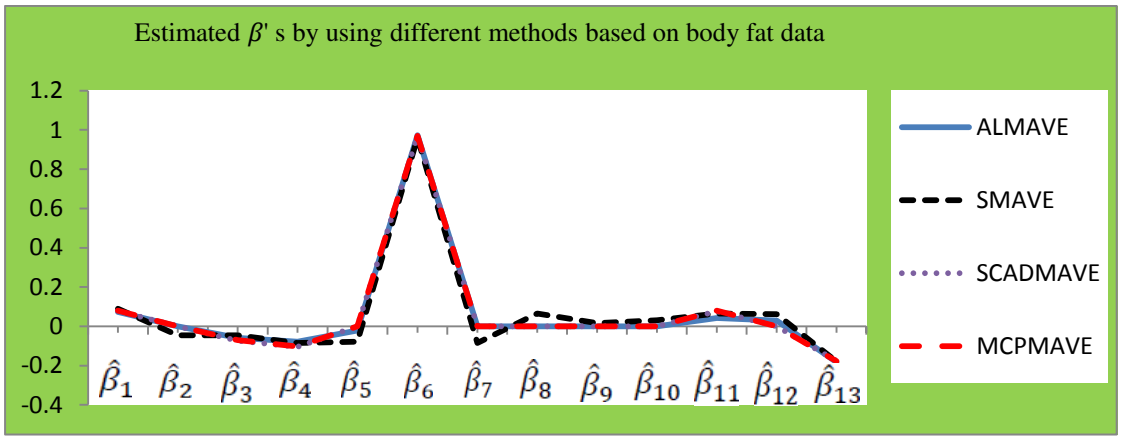


Figure 3.2. A plot explaining the estimated coefficients $\hat{\beta}$'s which are estimated by studied methods based on the BF data.

Table 3.7 reports the values of the adjusted R-squared for the model fit based on the BF data for all the studied methods. All of these methods discover the nonlinear structure better than the linear and the adjusted R-squared is same for all of the methods and for all of the fitted models.

Table 3.8 presents the prediction error of the cubic fit for the methods which are studied based on the BF data. It is clear that all of the proposed methods show a better performance than the SMAVE method. In general, the results are similar to those which are based on the AP data in Table 3.6.

Figure 3.2 presents plots and explains the estimated $\hat{\beta}'s$, which are estimated by studied methods based on the BF data. It can be seen from this figure that there are no big differences among the estimated cofficients for all of the methods.

3.9. Chapter Summary

In this chapter, we merge the shrinkage ideas of the adaptive Lasso, SCAD, MCP with well known sufficient dimension reduction method MAVE, to produce sparse and accurate solutions based on MAVE method. Sparse MAVE based on the adaptive Lasso, SCAD and MCP has been compared with the sparse MAVE method (Wang and Yin, 2008). In order to assess the numerical performance, a simulation study was conducted based on the models in the Examples 1, 2, 3 and 4, as described in Section 3.7. From the simulation study and the real data examples, it can be concluded that the sparse MAVE based on the adaptive Lasso, SCAD and MCP behaves well in comparison to the sparse MAVE (Wang and Yin, 2008), which is based on Lasso penalty, and thus we believe that the sparse MAVE based on the adaptive Lasso, SCAD and MCP is useful practically.

References

Breheny, P. and Huang, J. (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Annals of Applied Statistics 5, 232–253.

Cook, R. (1998). Regression graphics: ideas for studying the regression through graphics. New York, Wily.

Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. The Annals of Statistics 30, 455–474.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). Journal of the American Statistical Association 86**,** 328–332.

Fan, J. and Li, R. Z. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96,1348–1360.

Fung, W. K., He, X., Liu, L. and Shi, P. (2002). Dimension reduction based on canonical correlation. Statistica Sinica 12, 1093–1113.

Kraemer, N. and Schaefer, J. (2012). Package "parcor" .cran.r- project.org/web/packages/parcor/parcor.pdf

Li, K. (1991). Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association 86, 316–342.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. Journal of the American Statistical Association 87, 1025–1039.

Li, L. (2007). Sparse sufficient dimension reduction. Biometrika 94, 603–613.

Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free variable selection. Journal of the Royal Statistical Society Series B, 67, 285–299.

Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. Technometrics 48, 503–510.

Li, L. and Yin, X. (2008). Sliced Inverse Regression with regularizations. Biometrics 64, 124–131.

Ni, L., Cook, R. D. and Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. Biometrika 92, 242–247.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society, Series B 67**,** 91–108.

Xia, Y., Tong, H., Li, W. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society Series B 64, 363–410.

Wang, Q. and Yin, X. (2008). A Nonlinear Multi-Dimensional Variable Selection Method for High Dimensional Data: Sparse MAVE. Computational Statistics and Data Analysis 52, 4512–4520.

Yu, Z. and Zhu, L. (2013). Dimension reduction and predictor selection in semiparametric models. Biometrika, forthcoming.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics 38, 894–942.

Zhou, J. and He, X. M. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. Annals of Statistics 36, 1649–1668.

Zou, H. (2006). The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–142.

Chapter 4

Penalised single-index quantile regression[3]

The single-index (SI) regression and single-index quantile regression (SIQ) estimation methods provide linear combinations of all the original predictors. However, it is possible that there are many unimportant predictors within the original predictors. Thus, the accuracy of parameter estimation and the precision of prediction will be affected by the existence of those unimportant predictors when the mentioned methods are used.

In this chapter, an extension of the SIQ method of Wu et al. (2010) has been proposed, which considers Lasso and adaptive Lasso for estimation and variable selection. Computational algorithms have been developed in order to calculate the penalised SIQ estimates. A simulation study and a real data application have been used to assess the effectiveness of the methods under consideration.

[3]This chapter is based on: Alkenani, A. and Yu, K. (2013). Penalised single-index quantile regression. International Journal of Statistics and Probability 2, 12–30. http://dx.doi.org/10.5539/ijsp.v2n3p12.

4.1. Introduction

In many applications the linear relationship does not hold. Thus, the use of linear regression to describe the relations in these cases is not suitable. The SI model is an extension of the linear regression to deal with nonlinear relationships. It is more elastic than the parametric models and retains their good properties. Besides its ability to reduce the risk of misspecifying the link function, it helps to overcome the "course of dimensionality" (CD). Due to the index $\mathbf{x}^T\boldsymbol{\beta}$ aggregates the high dimensionality of $\mathbf{x}$, many researchers have been used the single index model to deal with the CD problem. The SI technique has been proven over the years to be an active and efficient method to deal with estimation for high-dimensional regression issues. It has gained much attention in recent years because of its usage in many fields. For example, qualitative choice models in econometrics and exposure–response models in biometrics (Härdle et al, 1993). It has the following form:

$$y = f(\mathbf{x}^T\boldsymbol{\beta}) + \varepsilon, \tag{4.1}$$

where y is a real valued response variable and $\mathbf{x}$ is a vector of p-dimensional predictors, $f(.)$ is an unknown univariable measurable function, the error ε is independent of $\mathbf{x}$ with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$, and $\boldsymbol{\beta}$ is the unknown SI vector coefficient satisfying $\|\boldsymbol{\beta}\| = 1$ and the first component β_1 is positive for the sake of model identifiability. Here $\|.\|$ denotes the Euclidean norm.

There are three types of procedures that have been suggested to estimate $\boldsymbol{\beta}$ in the statistical literature. The first type utilises the truth that $\boldsymbol{\beta}$ is proportional to the $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \boldsymbol{\beta} f'(\mathbf{x}^T\boldsymbol{\beta})$, which includes the average derivative estimation method (Härdle and Stoker, 1989), the structure adaptive method (Hristache et al., 2001) and the outer

product of gradients (OPG) method (Xia et al., 2002). The second type contains methods that estimate $f(.)$ and $\boldsymbol{\beta}$ in the same time. For example, the semiparametric least squares estimation method (Ichimura, 1993) and the minimum average variance estimator (MAVE) method (Xia et al., 2002). The third type consists of methods that use regressing $\mathbf{x}$ on y instead of regressing y on $\mathbf{x}$ and were primarily proposed to deal with the sufficient dimension reduction (SDR). For example, the sliced inverse regression (SIR) (Li, 1991), the sliced average variance estimation (SAVE) (Cook and Weisberg, 1991) and the directional regression (Li and Wang, 2007).

The majority of known estimation approaches for model (4.1) were constructed on either least squares or likelihood based methods. Thus, these approaches are expected to be sensitive to outliers. In contrast to the stated approaches, quantile regression (QR) (Koenker and Bassett, 1978) provides a robust alternative. As mentioned in Section 1.1, it supplies us with a full analysis of the relationships among the predictors and the response variable. Also, the QR has been applied in many different fields such as econometrics, finance, microarrays, medical and agricultural studies, see Koenker (2005) and Yu et al. (2003) for more details. A lot of work exists on QR; see for example, He and Shi (1996), He et al. (2002), Lee (2003), Cai and Xu (2009), Wang et al. (2010) and Kai et al. (2011), among others.

Although, a lot of work exists on nonparametric standard mean regression, however, very little exists on nonparametric QR. Nonparametric QR includes local linear methods and the spline methods. The local linear QR method for univariate QR is proposed by Yu and Jones (1998). Theoretically, while the extension of nonparametric conditional quantiles from univariate to higher dimension cases is quite clear, its practical success is impeded by the CD. Therefore, the challenge is to decrease the p-

dimensional predictor vector $\mathbf{x}$ without the loss of any information and without needing a pre-specified parametric model.

Recently, dimension reduction (DR) methods for nonparametric QR models have received a great interest in the statistical literature. Many approaches attempt to reduce the p-dimensional predictor vector $\mathbf{x}$ without losing information and then estimate the conditional quantile. Chaudhuri (1991), Gooijer and Zerom (2003), Yu and Lu (2004), Horowitz and Lee (2005), Dette and Scheder (2011) and Yebin et al. (2011) used variants of the adaptive model to reduce the dimension and thereafter estimate the conditional quantiles. In order to introduce a more efficient estimator of conditional quantiles, Gannon et al. (2004) used the SIR to reduce the dimensionality of the covariates. Recently, Wu et al. (2010) proposed SIQ method. A practical algorithm is introduced where the authors used the local linear QR to estimate $f(.)$ and linear QR to estimate the parametric index. Jiang et al. (2012) proposed the local linear composite QR estimator for a SI model. Hua et al. (2012) developed a Bayesian method for fitting models with a SI using conditional QR. Kong and Xia (2012) suggested an adaptive estimation method for SIQ model.

As mentioned in Section 1.1 and Section 3.1, variable selection is fundamental and very crucial to select important predictors in the high dimensional (HD) data analysis. It can save money and time used to collect unessential information, reduce computation time and improve efficiency and stability. A lot of articles are existed on subset selection by penalising the ordinary least squares; see for example, Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), fused Lasso (Tibshirani et al., 2005) and adaptive Lasso (Zou, 2006).

Because SI methods produce linear combinations (LC's) of all of the predictors, the variable selection approaches become needful for SI modelling when the number of

predictor variables is large and when there are unimportant predictors. Many researchers suggested to generalise a number of classical variable selection methods from linear regression to the SI model, such as the Akaike information criterion (AIC) and cross-validation and others, see for example, Naik and Tsai (2001) and Kong and Xia (2007). These methods are computationally intensive and unstable.

Some research has proposed to generalise the Lasso (Tibshirani, 1996) under the SI model assumptions. Under the scope of the SDR, Li and Yin (2008) combined the idea of Lasso with the SIR. Recently, Wang and Yin (2008) suggested the SMAVE. The authors proposed to add an l_1 penalty term $\lambda \sum_{k=1}^{p} |\beta_k|$ to the MAVE loss function to obtain the SMAVE. The idea of merging MAVE and Lasso, which is proposed in Wang and Yin (2008), was exploited by Zeng et al. (2012) by proposing an l_1 penalty that penalises the $\boldsymbol{\beta}$ and the norm of the $\frac{\partial f(x)}{\partial x}$ together.

Koenker (2004) proposed to use the regularisation in QR. In order to shrink individual effects towards a common value, the author put an l_1 penalty on the random effects in a mixed-effects QR model. Li and Zhu (2008) evolved a piecewise linear solution path for the l_1 penalised QR. Moreover, Wu and Liu (2009) proposed penalised QR with the SCAD and the adaptive Lasso penalties. Yuan and Yin (2010) proposed a Bayesian approach to shrink the random effects towards a common value by introducing an l_2 penalty to the usual QR check loss function. Li et al. (2010) suggested Bayesian regularized QR. The authors proposed different penalties such as Lasso, group Lasso and elastic net penalties. Alhamzawi et al. (2012) extended the Bayesian Lasso quantile regression (BLQR) reported in Li et al. (2010) to Bayesian adaptive Lasso quantile regression (BALQR) by using different penalization parameters for different regression coefficients.

In this chapter, we propose an extension of the SIQ model of Wu et al. (2010) by considering Lasso and adaptive Lasso for estimation and variable selection. Computational algorithms have been developed in order to calculate the penalised SIQ estimates. Our motivating example is an analysis of the Boston housing (BH) data which is available in ('MASS') package in R. The aim of this study is to investigate the relationship between the median value of owner-occupied homes and 13 statistical measurements on the 506 census tracts. In this study, we are interested in choosing the most significant statistical measurements of the 13 statistical measurements for the SIQ model, relating to the median value of owner-occupied homes. A certain correlation is present between the predictors in the BH data. For example, the correlation coefficient is (-0.7692) between the nitric oxides concentration and the weighted mean of distances to five Boston employment centres, (0.7636) between the nitric oxides concentration and the proportion of non-retail business acres/town, (-0.7478) between the weighted mean of distances and proportion of owner-occupied units, (0.7315) between the nitric oxides concentration and proportion of owner-occupied units built and so on. The selection of variables is important in this application, in order to know which predictors have coefficients that vary among subjects. The high correlation between the predictors is an argument to use the adaptive Lasso because the procedure deals with correlated predictors by using adaptive weights for the different predictors.

The remainder of the chapter is organized as follows. A brief review of the SIQ method is given in Section 4.2. Penalised SIQ with Lasso and adaptive Lasso are introduced in Section 4.3 and Section 4.4, respectively. Simulation studies are conducted under different settings in Section 4.5. The applications of the methods using real data are reported in Section 4.6. Lastly, the conclusions are summarised in Section 4.7.

4.2. Single-index quantile regression (SIQ) method

Given $\tau \in (0,1)$, Wu et al. (2010) proposed a SIQ model for the τth conditional quantile $\theta_\tau(\mathbf{x})$ of y given $\mathbf{x}$, as follows

$$\theta_\tau(\mathbf{x}) = f(\mathbf{x}^T\boldsymbol{\beta}), \tag{4.2}$$

where y is a real valued response variable and $\mathbf{x}$ is a vector of p-dimensional predictors, $f(.)$ is an unknown univariable measurable function, $\boldsymbol{\beta}$ is the unknown SI vector coefficient satisfying $\|\boldsymbol{\beta}\| = 1$ and β_1 is positive for the sake of model identifiability.

By replacing the nonparametric counterpart $f(\mathbf{x}^T\boldsymbol{\beta})$ in model (4.2) with $\mathbf{x}^T\boldsymbol{\beta}$, we obtain the linear QR of Koenker and Basset (1978). For the SIQ model (4.2), note $f(.)$ should be $f_\tau(.)$ and $\boldsymbol{\beta}$ should be $\boldsymbol{\beta}_\tau$. For notational convenience the subscript τ was omitted.

Let $\{\mathbf{x}_i, y_i\}$ be an independent identically distributed (i.i.d) sample from $(\mathbf{x}, y)$. For $\mathbf{x}_i^T\boldsymbol{\beta}$ close enough to u, $f(\mathbf{x}_i^T\boldsymbol{\beta})$ can be locally approximated by

$$f(\mathbf{x}_i^T\boldsymbol{\beta}) \approx f(u) + f'(u)(\mathbf{x}_i^T\boldsymbol{\beta} - u) = a + b(\mathbf{x}_i^T\boldsymbol{\beta} - u), \tag{4.3}$$

where $a \stackrel{\text{def}}{=} f(u)$ and $b \stackrel{\text{def}}{=} f'(u)$. Wu et al. (2010) proposed an estimation procedure for estimating $\boldsymbol{\beta}$ and $f(.)$ as follows:

Step 0. Find the initial $\widehat{\boldsymbol{\beta}}^{(0)}$ from the average derivative estimate (ADE) of Chaudhuri et al. (1997). The $\widehat{\boldsymbol{\beta}}^{(0)}$ will be standardised such that $\|\widehat{\boldsymbol{\beta}}\| = 1$ and $\hat{\beta}_1 > 0$.

Step 1. Given $\widehat{\boldsymbol{\beta}}$, obtain $\{\hat{a}_j, \hat{b}_j\}_{j=1}^{n}$ by solving the following

$$\min_{a_j, b_j} \sum_{i=1}^{n} \rho_\tau \left(y_i - a_j - b_j(\mathbf{x}_i - \mathbf{x}_j)^T \widehat{\boldsymbol{\beta}}\right) \omega_{ij}, \tag{4.4}$$

where $\rho_\tau(.)$ is the check loss function defined by $\rho_\tau(u) = \tau\, u\, I_{[0,\infty)}(u) - (1-\tau)\, u\, I_{(-\infty,0)}(u)$, the weight function $\omega_{ij} = K\left(\frac{\mathbf{x}_i^T\widehat{\boldsymbol{\beta}} - \mathbf{x}_j^T\widehat{\boldsymbol{\beta}}}{h}\right) \Big/ \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i^T\widehat{\boldsymbol{\beta}} - \mathbf{x}_j^T\widehat{\boldsymbol{\beta}}}{h}\right)$ and $K(.)$ is a kernel function with the bandwidth h chosen to be optimal.

Step 2. Given $\{\hat{a}_j, \hat{b}_j\}_{j=1}^{n}$, obtain $\widehat{\boldsymbol{\beta}}$ by solving

$$\min_{\boldsymbol{\beta}} \sum_{j=1}^{n} \sum_{i=1}^{n} \rho_\tau \left(y_i - \hat{a}_j - \hat{b}_j(\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\beta}\right) \omega_{ij}$$

$$= \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{n} \sum_{i=1}^{n} \rho_\tau \left(y_{ij}^* - \mathbf{x}_{ij}^{*T} \boldsymbol{\beta}\right) \omega_{ij}^*, \tag{4.5}$$

where $y_{ij}^* = y_i - \hat{a}_j$, $\mathbf{x}_{ij}^* = \hat{b}_j(\mathbf{x}_i - \mathbf{x}_j)$ and $\omega_{ij}^* = \omega_{ij}$ evaluated at the current estimate of $\boldsymbol{\beta}$. In step 2, $\boldsymbol{\beta}$ is estimated via the linear QR without an intercept on n^2 observations $\{y_{ij}^*, \mathbf{x}_{ij}^*\}_{i,j=1}^{n}$ with known weights $\{\omega_{ij}^*\}_{i,j=1}^{n}$ evaluated at the estimate of $\boldsymbol{\beta}$ from the former iteration.

Step 3. Continue repeating the steps 1 and 2 until convergence.

The standardisation of $\widehat{\boldsymbol{\beta}}$ is done as $\boldsymbol{\beta} = sign_{1\beta}/\|\boldsymbol{\beta}\|$, where $sign_{1\beta}$ is the sign of the β_1. The final estimate of $f(.)$ is $\hat{f}(u; h, \widehat{\boldsymbol{\beta}}) = \hat{a}$ where

$$(\hat{a}, \hat{b}) = arg \min_{(a,b)} \sum_{i=1}^{n} \rho_\tau \left(y_i - a - b(\mathbf{x}_i^T\widehat{\boldsymbol{\beta}} - u)\right) K\left(\frac{\mathbf{x}_i^T\widehat{\boldsymbol{\beta}} - u}{h}\right). \tag{4.6}$$

4.3. Single-index quantile regression with Lasso penalty (LSIQ)

The Lasso is proposed by Tibshirani (1996) for simultaneous variable selection and parameter estimation. According to the Lasso, the residual sum of squares is minimised subject to $\sum_{k=1}^{p}|\beta_k|$ being less than a constant. By assuming this constraint, the Lasso shrinks some coefficients and sets other to 0. As an extension to Lasso Tibshirani (1996), Li and Zhu (2008) suggested Lasso quantile regression (LQR) minimises

$$\sum_{i=1}^{n}\rho_\tau(y_i-\mathbf{x}_i^T\boldsymbol{\beta})+\lambda\sum_{k=1}^{p}|\beta_k|, \tag{4.7}$$

where $\lambda>0$ is the tuning parameter. The term $\lambda\sum_{k=1}^{p}|\beta_k|$ in (4.7) is the l_1 penalty QR, which is important for the Lasso.

The LSIQ is proposed here according to an algorithm similar to the algorithm in Section 4.2, except in the initial step where we obtain the $\widehat{\boldsymbol{\beta}}^{(\mathbf{0})}$ from the Lasso linear QR from Li and Zhu (2008). Also, in step 2, given $\{\hat{a}_j,\hat{b}_j\}_{j=1}^{n}$, we obtain $\widehat{\boldsymbol{\beta}}_{Lasso}$ by solving

$$\min_{\boldsymbol{\beta}}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\hat{a}_j-\hat{b}_j(\mathbf{x}_i-\mathbf{x}_j)^T\boldsymbol{\beta}\right)\omega_{ij}+\lambda\sum_{k=1}^{p}|\beta_k|\right)$$

$$=\arg\min_{\boldsymbol{\beta}}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\rho_\tau\left(y_{ij}^*-\mathbf{x}_{ij}^{*\,T}\boldsymbol{\beta}\right)\omega_{ij}^*+\lambda\sum_{k=1}^{p}|\beta_k|\right). \tag{4.8}$$

The final estimate of $f(.)$ is $\hat{f}(u;h,\widehat{\boldsymbol{\beta}}_{Lasso})=\hat{a}$, where

$$(\hat{a},\hat{b})=arg\min_{(a,b)}\sum_{i=1}^{n}\rho_\tau\left(y_i-a-b(\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}_{Lasso}-u)\right)K\left(\frac{\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}_{Lasso}-u}{h}\right). \tag{4.9}$$

4.4. Single-index quantile regression with adaptive Lasso penalty(ALSIQ)

Under specific situations, Lasso has been shown to be consistent by Zou (2006), who derived a necessary condition for the Lasso to be consistent. Consequently, the Lasso is inconsistent in other certain conditions. A flexible version of Lasso method has been suggested by Zou (2006) via assigning different weights for shrinkage the different coefficients of predictors. The author explained that the main merit of his method compared to the Lasso estimator is that the adaptive Lasso has the OP's. Zou (2006) stated that the LARS (Least angle regression) algorithm can be used for solving the adaptive Lasso. Wu and Liu (2009) suggested the adaptive Lasso quantile regression (ALQR) minimises

$$\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda \sum_{k=1}^{p} \tilde{\omega}_k |\beta_k|, \tag{4.10}$$

where the weights are set to be $\tilde{\omega}_k = 1/|\tilde{\beta}_k|^{\delta}$, $k = 1, \ldots, p$, $\tilde{\beta}$ is the non-penalised QR estimate and $\delta > 0$.

The ALSIQ has been suggested according to the algorithm similar to the algorithms in Section 4.2 and 4.3, except in the initial step we obtained the $\widehat{\boldsymbol{\beta}}^{(\mathbf{0})}$ from the ALQR of Wu and Liu (2009). Also, in step 2, given $\{\hat{a}_j, \hat{b}_j\}_{j=1}^{n}$, we got $\widehat{\boldsymbol{\beta}}_{ALasso}$ by solving

$$\min_{\boldsymbol{\beta}} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \rho_\tau \left(y_i - \hat{a}_j - \hat{b}_j (\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\beta} \right) \omega_{ij} + \lambda \sum_{k=1}^{p} \tilde{\omega}_k |\beta_k| \right)$$

$$= \arg\min_{\boldsymbol{\beta}} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \rho_\tau \left(y_{ij}^* - \mathbf{x}_{ij}^{*T} \boldsymbol{\beta} \right) \omega_{ij}^* + \lambda \sum_{k=1}^{p} \tilde{\omega}_k |\beta_k| \right). \tag{4.11}$$

Thus, we can obtain $\widehat{\boldsymbol{\beta}}_{ALasso}$ by solving the minimisation problem in (4.11) as ALQR by using LARS algorithm, see Wu and Liu (2009).

The final estimate of $f(.)$ is $\hat{f}\left(u; h, \widehat{\boldsymbol{\beta}}_{ALasso}\right) = \hat{a}$, where

$$\left(\hat{a}, \hat{b}\right) = arg \min_{(a,b)} \sum_{i=1}^{n} \rho_\tau \left(y_i - a - b\left(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{ALasso} - u\right)\right) K\left(\frac{\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{ALasso} - u}{h}\right). \qquad (4.12)$$

4.5. A simulation study

Many simulations have been implemented in order to check the behaviour of the suggested methods and some examples are reported below:

Example 1: $\mathcal{R} = 200$ datasets were generated with size $n = 300$ observations from the following model:

$$y = 5\cos(\mathbf{x}^T\boldsymbol{\beta}) + \exp(-(\mathbf{x}^T\boldsymbol{\beta})^2) + \varepsilon,$$

where $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_5)^T$, $\boldsymbol{\beta} = (1,2,0,0,0)^T/\sqrt{5}$, x_i i. i. d. $\sim$ Unif(0,1); $i = 1,2,\ldots,5$, the error term $\varepsilon \sim$Exp(0.5), x_i 's and ε are mutually independent. The $\boldsymbol{\beta}$ is estimated for $\tau = (0.10,0.25,0.50,0.75,0.90)$.

Example 2: $\mathcal{R} = 200$ data-sets were generated with size $n = 300$ observations from the following model with homoscedastic errors.

$$y = sin\left\{\frac{\pi\,(\mathbf{x}^T\boldsymbol{\beta} - \mathcal{A}\,)}{\mathcal{C} - \mathcal{A}}\right\} + 0.5\,\varepsilon,$$

where $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_6)^T$ and $\boldsymbol{\beta} = (1,1,0,0,0,0)^T/\sqrt{3}$. Here, $\mathcal{A} = \frac{\sqrt{3}}{2} - \frac{1.645}{\sqrt{12}}$, $\mathcal{C} = \frac{\sqrt{3}}{2} + \frac{1.645}{\sqrt{12}}$, x_i i. i. d. $\sim$ Unif(0,1), $\varepsilon \sim$ N(0,1) and x_i 's and ε are mutually independent. The $\boldsymbol{\beta}$ is estimated for $\tau = (0.10,0.25,0.50,0.75,0.90)$.

Example 3: $\mathcal{R} = 200$ datasets were generated with size $n = 300$ observations from the model $y = exp(\mathbf{x}^T\boldsymbol{\beta}) + \varepsilon$, where $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_{10})^T$ are generated as i.i.d standard normals. The error term is assumed to be $\varepsilon \sim \mathrm{N}(0,1)$ and that it is independent of $\mathbf{x}$. $\boldsymbol{\beta} = (1,1,1,0,0,0,0,0,0,0)^T/\sqrt{3}$ is used.

The $\boldsymbol{\beta}$ is estimated for $\tau = (0.10, 0.25, 0.50, 0.75, 0.90)$.

We analysed each simulated data set using three methods. The LSIQ and ALSIQ methods, which are described in Sections 4.3 and 4.4, respectively, are compared with the SIQ method. The $rq(y^* \sim \mathbf{X}^*, tau, \text{method} = \text{"lasso"})$ function in the quantreg package is used to obtain $\widehat{\boldsymbol{\beta}}_{ALasso}$ in Equation (4.8). The ALassoQR function from the code of Wu and Liu (2009) (Personal communication with Wu) is used to obtain $\widehat{\boldsymbol{\beta}}_{ALasso}$ in Equation (4.11). Similar to Wu and Liu (2009), λ was chosen via a grid search based on the tuning error in terms of the mean squared error (MSE) evaluated on the data. This means that the λ value has been chosen to minimise the MSE.

Table 4.1. The mean and standard deviation (SD) of MSE for $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ based on the model in example 1.

		Method		
τ		SIQ	LSIQ	ALSIQ
	M.MSE	0.0014	0.0006	0.0005
0.10	SD.MSE	0.0011	0.0005	0.0004
	M.MSE	0.0046	0.0022	0.0020
0.25	SD.MSE	0.0049	0.0026	0.0022
	M.MSE	0.0138	0.0046	0.0046
0.50	SD.MSE	0.0128	0.0064	0.0065
	M.MSE	0.0467	0.0335	0.0311
0.75	SD.MSE	0.0593	0.0454	0.0443
	M.MSE	0.0661	0.0581	0.0509
0.90	SD.MSE	0.0857	0.0734	0.0702

Table 4.2.The mean and SD of MSE for $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ based on the model in example 2.

		Method		
τ		SIQ	LSIQ	ALSIQ
	M.MSE	0.0294	0.0372	0.0136
0.10	SD.MSE	0.1025	0.0396	0.0220
	M.MSE	0.0077	0.0067	0.0047
0.25	SD.MSE	0.0086	0.0070	0.0045
	M.MSE	0.0044	0.0043	0.0042
0.50	SD.MSE	0.0048	0.0048	0.0048
	M.MSE	0.0169	0.0072	0.0031
0.75	SD.MSE	0.0198	0.0108	0.0048
	M.MSE	0.0197	0.0070	0.0018
0.90	SD.MSE	0.0230	0.0080	0.0025

Table 4.3.The mean and SD of MSE for $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ based on the model in example 3.

		Method		
τ		SIQ	LSIQ	ALSIQ
	M.MSE	0.0688	0.0565	0.0412
0.10	SD.MSE	0.0434	0.0479	0.0343
	M.MSE	0.0494	0.0452	0.0367
0.25	SD.MSE	0.0325	0.0278	0.0197
	M.MSE	0.0403	0.0336	0.0330
0.50	SD.MSE	0.0455	0.0300	0.0206
	M.MSE	0.0495	0.0370	0.0360
0.75	SD.MSE	0.0747	0.0298	0.0272
	M.MSE	0.0489	0.0453	0.0406
0.90	SD.MSE	0.0298	0.0345	0.0285

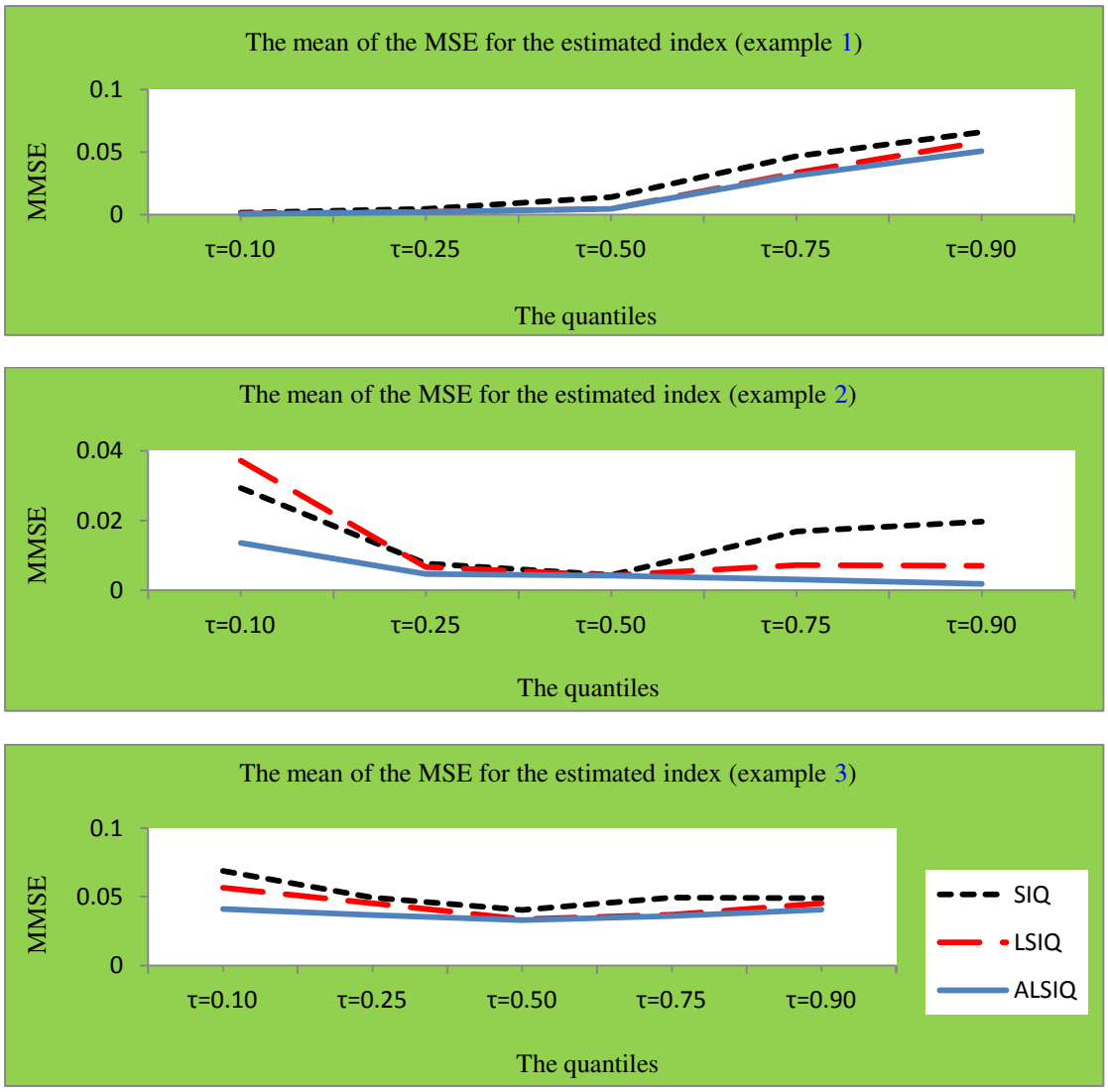


Figure 4.1. Plots explain the mean of MSE for $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ based on the model in examples 1, 2 and 3 respectively.

According to the mean and the SD of the MSE for $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$, from Tables 4.1, 4.2 and 4.3 and Figure 4.1, it can be seen that the proposed methods (ALSIQ and LSIQ) perform better than the SIQ method described in Wu et al. (2010) for all the models under consideration. This indicates that the proposed methods give precise estimates even when the error distribution is asymmetric. Most noticeably, when $\tau = 0.10$ and $\tau = 0.90$ the ALSIQ and LSIQ are significantly more efficient than the SIQ method.

Table 4.4. The mean and MSE for single-index coefficient estimates based on the model in example 1.

τ	**Method**		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
0.10	SIQ	Mean	0.4750	0.8776	0.0081	0.0049	0.0122
		MSE	0.0017	0.0006	0.0005	0.0016	0.0012
	LSIQ	Mean	0.4609	0.8859	0.0037	0.0021	0.0037
		MSE	0.0016	0.0005	0.0003	0.0006	0.0003
	ALSIQ	Mean	0.4667	0.8829	0.0021	0.0050	0.0010
		MSE	0.0021	0.0006	0.0002	0.0002	0.0002
0.25	SIQ	Mean	0.4865	0.8651	-0.0019	-0.0117	0.0097
		MSE	0.0038	0.0016	0.0018	0.0042	0.0069
	LSIQ	Mean	0.4737	0.8759	-0.0009	-0.0050	0.0010
		MSE	0.0039	0.0013	0.0010	0.0016	0.0021
	ALSIQ	Mean	0.4727	0.8766	0.0013	-0.0020	0.0029
		MSE	0.0042	0.0013	0.0009	0.0013	0.0020
0.50	SIQ	Mean	0.4482	0.8664	-0.0149	0.0029	-0.0083
		MSE	0.0098	0.0032	0.0114	0.0111	0.0154
	LSIQ	Mean	0.4658	0.8725	-0.0026	0.0043	0.0016
		MSE	0.0089	0.0027	0.0049	0.0027	0.0044
	ALSIQ	Mean	0.4654	0.8727	-0.0028	0.0060	-0.0008
		MSE	0.0088	0.0027	0.0048	0.0026	0.0045
0.75	SIQ	Mean	0.5429	0.7331	0.0158	0.0058	-0.0230
		MSE	0.0247	0.0531	0.0426	0.0592	0.0323
	LSIQ	Mean	0.5053	0.7881	0.0308	-0.0265	-0.0020
		MSE	0.0359	0.0290	0.0303	0.0268	0.0227
	ALSIQ	Mean	0.5486	0.7660	0.0407	-0.0433	-0.0059
		MSE	0.0412	0.0339	0.0305	0.0211	0.0181
0.90	SIQ	Mean	0.5659	0.6591	-0.0229	-0.0099	-0.0299
		MSE	0.1015	0.0824	0.0456	0.0401	0.0202
	LSIQ	Mean	0.5943	0.6474	-0.0016	-0.0078	-0.0418
		MSE	0.1012	0.0923	0.0332	0.0385	0.0114
	ALSIQ	Mean	0.6029	0.6443	0.0140	0.0017	-0.0420
		MSE	0.0988	0.1017	0.0288	0.0324	0.0123

Table 4.5. The mean and MSE for single-index coefficient estimates based on the model in example 2.

τ	**Method**		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
0.10	SIQ	Mean	0.6702	0.6924	-0.0253	0.0163	0.0132	-0.0421
		MSE	0.0217	0.0175	0.0136	0.0053	0.0189	0.0192
	LSIQ	Mean	0.7686	0.4655	-0.0068	0.0072	-0.0077	0.0054
		MSE	0.1217	0.1222	0.0011	0.0011	0.0016	0.0027
	ALSIQ	Mean	0.8211	0.5174	-0.0014	-0.0007	0.0086	0.0143
		MSE	0.0723	0.0447	0.0001	0.0003	0.0012	0.0052
0.25	SIQ	Mean	0.7753	0.5937	0.0008	0.0238	0.0108	-0.0021
		MSE	0.0517	0.0172	0.0031	0.0054	0.0021	0.0076
	LSIQ	Mean	0.7771	0.5915	-0.0040	0.0160	0.0031	0.0008
		MSE	0.0551	0.0211	0.0017	0.0040	0.0014	0.0042
	ASIQ	Mean	0.6971	0.7016	0.0074	0.0165	0.0228	-0.0032
		MSE	0.0168	0.0180	0.0031	0.0043	0.0054	0.0045
0.50	SIQ	Mean	0.6884	0.7125	0.0059	0.0125	0.0211	-0.0102
		MSE	0.0143	0.0198	0.0028	0.0045	0.0056	0.0024
	LSIQ	Mean	0.7750	0.6099	0.0010	0.0197	0.0114	0.0062
		MSE	0.0472	0.0118	0.0003	0.0035	0.0022	0.0032
	ALSIQ	Mean	0.7738	0.6115	0.0015	0.0202	0.0115	0.0061
		MSE	0.04667	0.01183	0.0003	0.0035	0.0022	0.0032
0.75	SIQ	Mean	0.7154	0.6355	0.0187	0.0330	0.0114	0.0232
		MSE	0.0368	0.0247	0.0058	0.0121	0.0110	0.0182
	LSIQ	Mean	0.6981	0.7016	0.0193	0.0103	0.0137	-0.0042
		MSE	0.0163	0.0172	0.0035	0.0044	0.0043	0.0052
	ALSIQ	Mean	0.7343	0.6680	0.0056	0.0115	0.0060	0.0074
		MSE	0.0274	0.0113	0.0009	0.0028	0.0017	0.0035
0.90	SIQ	Mean	0.7045	0.6420	0.0264	0.0447	0.0246	0.0247
		MSE	0.0819	0.0843	0.0079	0.0160	0.0143	0.0189
	LSIQ	Mean	0.7019	0.6936	0.0087	-0.0038	0.0072	-0.0045
		MSE	0.0176	0.0155	0.0061	0.0057	0.0053	0.0057
	ALSIQ	Mean	0.7606	0.6430	0.0015	0.0068	-0.0023	0.0024
		MSE	0.0357	0.0075	0.0003	0.0015	0.0004	0.0008

Table 4.6. The mean and MSE for single-index coefficient estimates based on the model in example 3.

τ	**Method**		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
0.10	SIQ	Mean	0.5775	0.5570	0.5365	-0.0112	-0.0030	0.0008	0.0208	0.0047	0.0092	0.0023
		MSE	0.0078	0.0070	0.0099	0.0042	0.0056	0.0049	0.0069	0.0072	0.0119	0.0075
	LSIQ	Mean	0.5564	0.5784	0.5483	-0.0152	-0.0236	0.0188	0.0226	-0.0175	0.0221	-0.0275
		MSE	0.0051	0.002	0.0027	0.0088	0.0058	0.0068	0.0078	0.0074	0.0061	0.0061
	ALSIQ	Mean	0.5939	0.5380	0.5650	0.0013	-0.0068	-0.0168	0.0006	0.0033	0.0015	0.0127
		MSE	0.0089	0.0050	0.0062	0.0031	0.0019	0.0027	0.0032	0.0027	0.0032	0.0050
0.25	SIQ	Mean	0.5716	0.5540	0.5633	0.0058	0.0008	0.0171	-0.0108	0.0130	-0.0035	0.0207
		MSE	0.0042	0.0038	0.0032	0.0045	0.0040	0.0084	0.0057	0.0038	0.0069	0.0071
	LSIQ	Mean	0.5698	0.5716	0.5519	-0.0272	-0.0142	-0.0234	0.0187	0.0105	-0.0089	-0.0215
		MSE	0.0032	0.0024	0.0033	0.0055	0.0039	0.0060	0.0048	0.0060	0.0072	0.0042
	ALSIQ	Mean	0.5685	0.5506	0.5810	0.0025	-0.0051	-0.0170	0.0061	0.0114	-0.0125	0.0025
		MSE	0.0050	0.0066	0.0042	0.0023	0.0034	0.0026	0.0047	0.0041	0.0032	0.0023
0.50	SIQ	Mean	0.5574	0.5779	0.5627	-0.0073	0.0013	0.0236	0.0041	-0.0028	-0.0054	-0.0058
		MSE	0.0033	0.0012	0.0077	0.0026	0.0046	0.0042	0.0018	0.0042	0.0048	0.0072
	LSIQ	Mean	0.5958	0.5735	0.5330	0.0045	0.0053	-0.0137	0.0351	0.0003	-0.0157	-0.0112
		MSE	0.0027	0.0012	0.0038	0.0019	0.0014	0.0029	0.0079	0.0040	0.0057	0.0046
	ALSIQ	Mean	0.5737	0.5512	0.5770	-0.0047	-0.0084	0.0094	-0.0025	0.0023	-0.0095	0.0152
		MSE	0.0018	0.0036	0.0021	0.0027	0.0037	0.0032	0.0049	0.0056	0.0051	0.0033
0.75	SIQ	Mean	0.5788	0.5446	0.5645	-0.0036	-0.0162	-0.0020	-0.0001	-0.0056	-0.0004	-0.0006
		MSE	0.0020	0.0083	0.0026	0.0102	0.0034	0.0032	0.0036	0.0049	0.0061	0.0084
	LSIQ	Mean	0.5833	0.5600	0.5558	0.0166	-0.0081	0.0100	0.0106	-0.0140	-0.0128	0.0046
		MSE	0.0026	0.0042	0.0026	0.0022	0.0050	0.0058	0.0046	0.0065	0.0052	0.0021
	ASIQ	Mean	0.5747	0.5514	0.5742	-0.0110	0.0069	0.0125	0.0195	-0.0015	-0.0093	0.0071
		MSE	0.0019	0.0071	0.0041	0.0033	0.0039	0.0054	0.0029	0.0028	0.0045	0.0022
0.90	SIQ	Mean	0.5577	0.5646	0.5671	0.0043	-0.0235	-0.0154	0.0265	0.0087	0.0116	-0.0235
		MSE	0.0041	0.0054	0.0031	0.0041	0.0067	0.0052	0.0066	0.0033	0.0051	0.0074
	LSIQ	Mean	0.5866	0.5553	0.5512	0.0099	0.0060	0.0074	0.0072	-0.0184	-0.0169	0.0092
		MSE	0.0040	0.0054	0.0045	0.0039	0.0037	0.0039	0.0057	0.0041	0.0089	0.0025
	ALSIQ	Mean	0.5939	0.5559	0.5457	-0.0177	0.0034	-0.0067	-0.0043	-0.0097	-0.0111	0.0020
		MSE	0.0065	0.0064	0.0090	0.0039	0.0018	0.0017	0.0031	0.0048	0.0028	0.0036

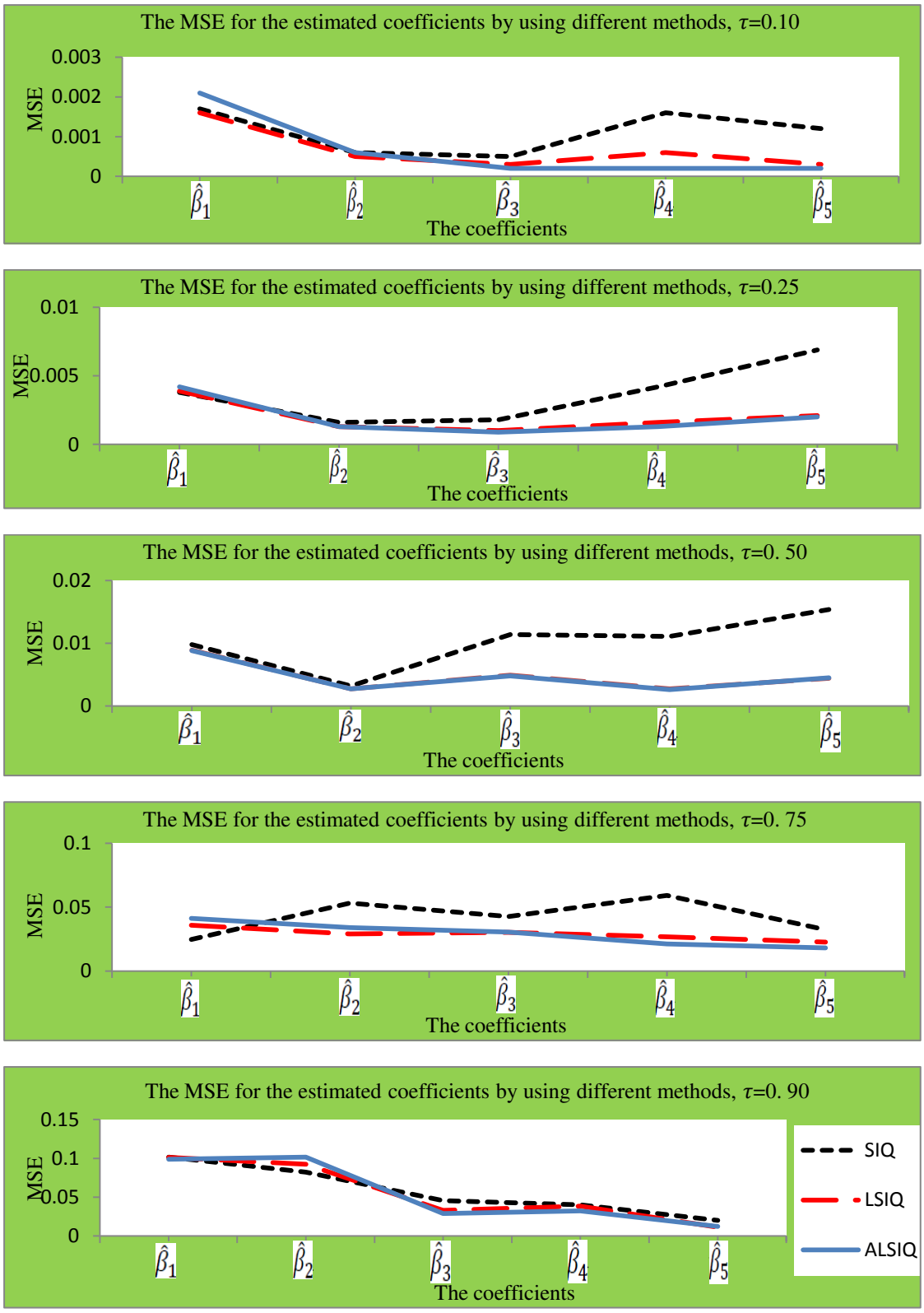


Figure 4.2. Plots explain the MSE for single-index coefficient estimates based on the model in example 1.

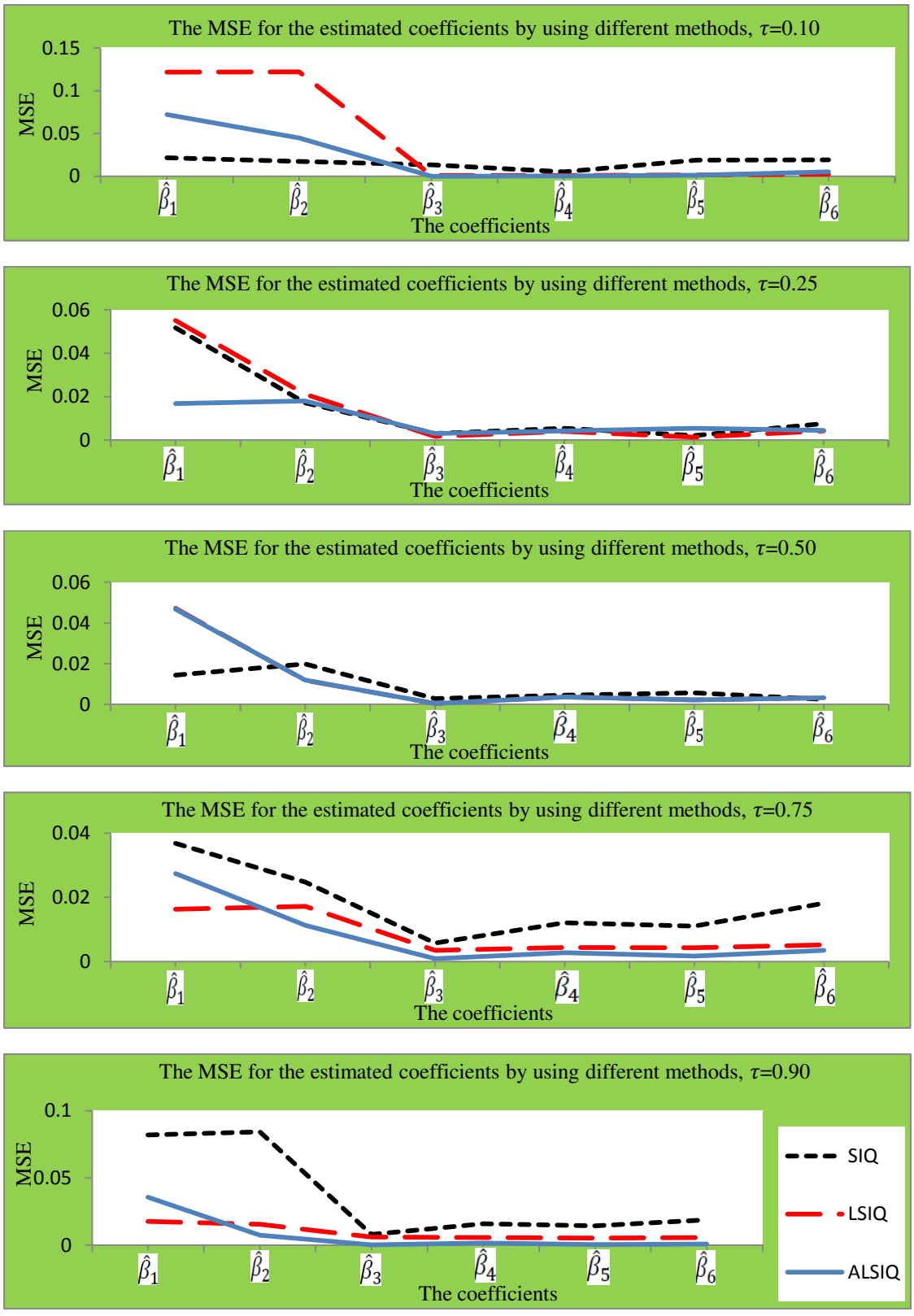


Figure 4.3. Plots explain the MSE for single-index coefficient estimates based on the model in example 2.

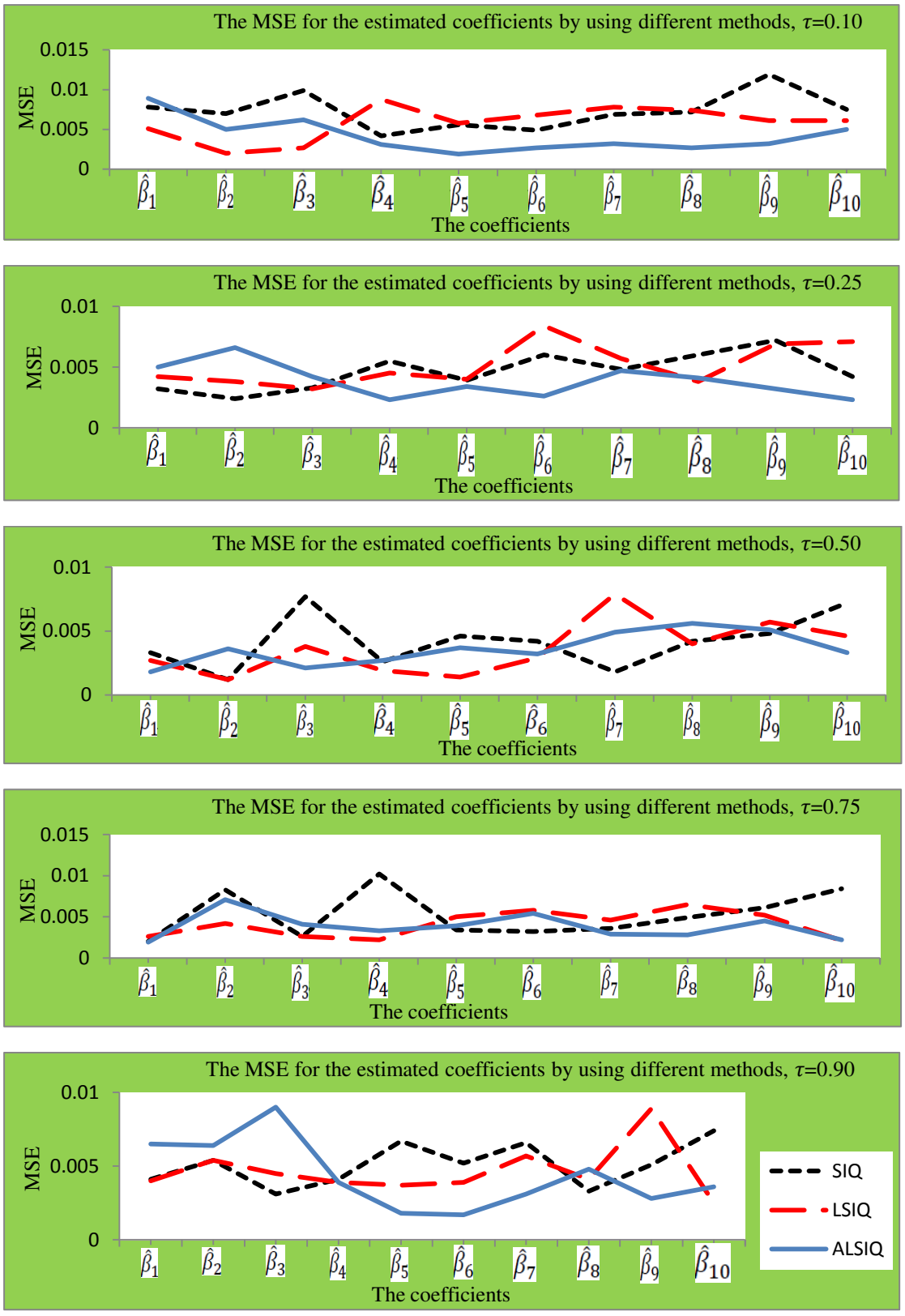


Figure 4.4. Plots explain the MSE for single-index coefficient estimates based on the model in example 3.

According to the MSE for the single-index coefficient estimates, from Tables 4.4, 4.5 and 4.6 and Figures 4.2, 4.3 and 4.4, it can be observed that in the majority of the estimated coefficients, the proposed methods produce a lower MSE than the SIQ method. Furthermore, one can see that the coefficients estimators of the proposed methods are close to the true values.

The variations in the ALSIQ and LSIQ estimates are similar in the majority of cases and less than the variations in the estimate of the SIQ method.

4.6. Boston housing (BH) data

In this section, the methods are illustrated through an analysis of the BH data. The data consist of $n = 506$ observations on 14 variables; medv is the median value of owner-occupied homes and it refers to the response variable. The dataset consist of 13 predictors on the 506 census tracts, which is available in the package ('MASS') in R. In our analysis, the dummy variable (chas) and the categorical variable (rad) were excluded. The predictors under consideration are crime average (x_1), ratio of residential land (x_2), ratio of non-retail business acres/town (x_3), nitric oxides concentration (x_4), rate number of rooms/dwelling (x_5), ratio of owner-occupied units (x_6), weighted mean of distances (x_7), tax average of the property (x_8), pupil-teacher proportion by town (x_9), black population ratio town (x_{10}), and lower status of the population (x_{11}). The response variable medv and the predictor variables were also standardised.

Table 4.7. Single-index coefficient estimates for Boston housing data based on the BH data.

τ	**Method**	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$
0.10	SIQ	0.351	0.012	-0.104	0.169	-0.494	0.191	0.228	0.311	0.220	-0.139	0.584
	LSIQ	0.342	-0.022	-0.059	0.296	-0.372	0.079	0.279	0.250	0.227	-0.202	0.644
	ALSIQ	0.446	0	0	0	-0.354	0	0.013	0.181	0.175	-0.158	0.767
0.25	SIQ	0.647	-0.028	-0.030	0.031	-0.489	0.177	0.213	0.067	0.166	-0.166	0.451
	LSIQ	0.153	0	0	0.243	-0.513	0.042	0.228	0.266	0.314	-0.246	0.609
	ALSIQ	0.123	0	0	-0.252	0.659	-0.146	-0.254	-0.325	-0.328	0.250	-0.354
0.50	SIQ	0.335	-0.009	-0.026	0.055	-0.500	0.130	0.217	0.059	0.206	-0.246	0.681
	LSIQ	0.110	-0.014	0	0.198	-0.597	0.092	0.246	0.165	0.325	-0.225	0.583
	ALSIQ	0.108	-0.014	0	0.198	-0.597	0.093	0.247	0.165	0.325	-0.224	0.583
0.75	SIQ	0.234	-0.032	-0.006	0.085	-0.585	0.109	0.283	-0.002	0.214	-0.308	0.601
	LSIQ	0.084	-0.046	0	0.155	-0.715	0.090	0.282	0.063	0.295	-0.192	0.490
	ALSIQ	0.112	-0.003	0	0.190	-0.656	0.069	0.235	0.009	0.338	-0.217	0.547
0.90	SIQ	0.174	-0.042	0.065	0.165	-0.461	-0.029	0.302	-0.090	0.204	-0.235	0.726
	LSIQ	0.033	-0.016	0.045	0.155	-0.722	0	0.187	0	0.379	-0.132	0.505
	ALSIQ	0.001	-0.057	0.053	0.069	-0.781	0	0.219	0	0.355	-0.135	0.432

Table 4.8. MSE for estimated quantiles curves $\hat{f}(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}})$ based on the BH data.

	τ				
Method	$\mathbf{0.10}$	$\mathbf{0.25}$	$\mathbf{0.50}$	$\mathbf{0.75}$	$\mathbf{0.90}$
SIQ	0.5083	0.3780	0.0901	0.0144	0.3117
LASSO-SIQ	0.4509	0.2957	0.0406	0.0141	0.2833
ALASSO -SIQ	0.4404	0.2611	0.0392	0.0130	0.2922

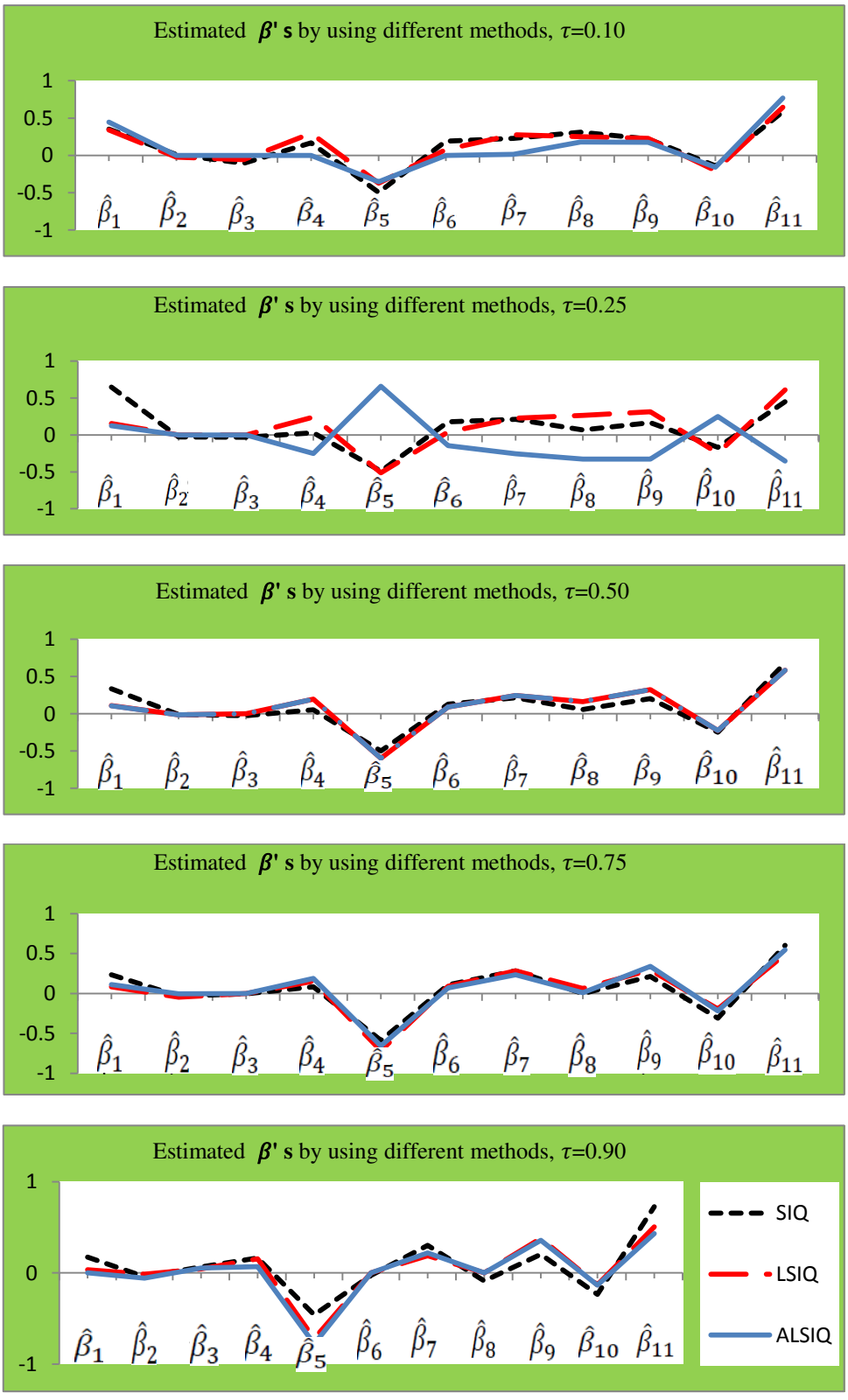


Figure 4.5. Plots explain the single-index coefficient estimates based on the BH data.

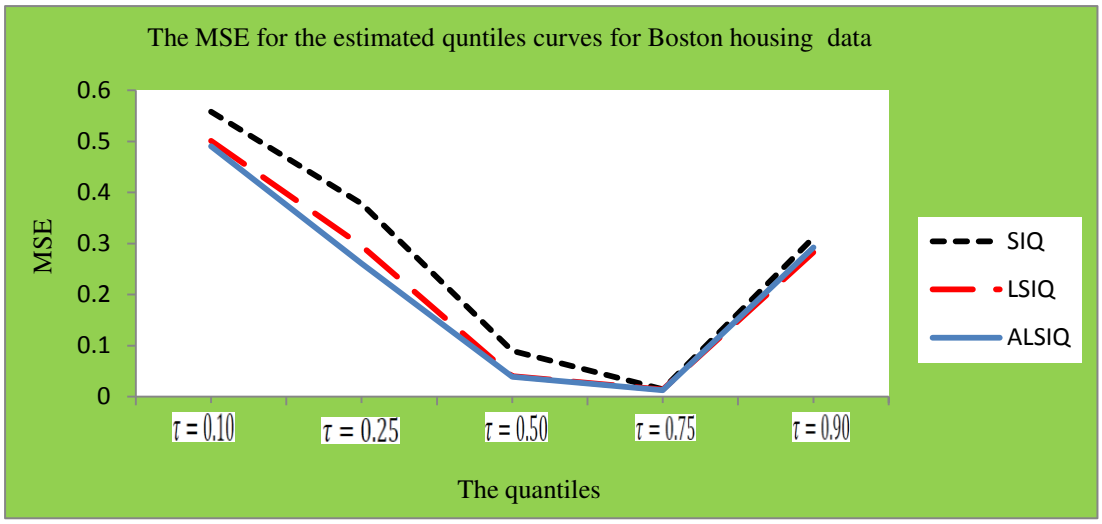


Figure 4.6. MSE for the smooth estimated quantiles curves $\hat{f}(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}})$ based on the BH .

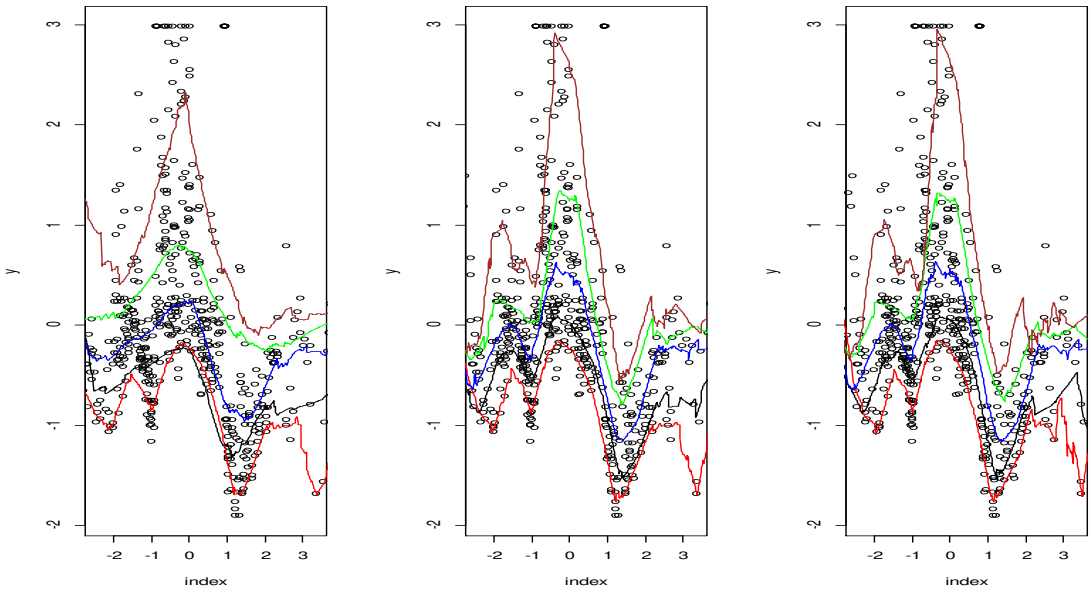


Figure 4.7. Plots for the smooth estimated quantiles curves $\hat{f}(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}})$ which are estimated by the ALSIQ, LSIQ and SIQ, respectively from the right to the left based on the BH data.

The estimated $\widehat{\boldsymbol{\beta}}$ using all the methods under consideration based on the BH data are given in Table 4.7 and explained in Figure 4.5. The estimated coefficient is treated as zero if its absolute value is smaller than 10^{-12}.

Table 4.8 and Figure 4.6 present the MSEs for estimated quantile curves $\hat{f}(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}})$ which are estimated by the proposed methods and the SIQ method based on the BH data for different quantile values. From Table 4.8 and Figure 4.6, it is clear that the proposed methods outperform the SIQ method in fitting the BH data set. Again, it can be seen that when $\tau = 0.10$ and $\tau = 0.90$ the proposed methods are significantly more efficient than the other methods. Figure 4.7 shows the smooth estimated quantile curves $\hat{f}(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}})$ which are estimated by all the methods under consideration based on the BH data for different quantile values.

Similar to Wu et al. (2010) possible quantile curves crossing at both tails can be seen, which due to the sparsity of data in the region concerned. The results of the real data example confirm the results of the simulation studies that the suggested methods behave well.

4.7. Chapter Summary

In this chapter, an extension of the SIQ method of Wu et al. (2010) has been proposed, which considers Lasso and adaptive Lasso for estimation and variable selection. The effectiveness of the proposed extensions is explained via many simulation examples, as well as a real data analysis. From the simulation study and the real data example, it can be concluded that the proposed extensions perform well in comparison to the SIQ method. We believe that the proposed extensions would supply helpful dimension reduction tools. Also, it would support the applicability of shrinkage methods to SIQ models.

References

Alhamzawi, R., Yu, K. and Benoit, D. (2012). Bayesian adaptive LASSO quantile regression. Statistical Modelling 12, 279–297.

Cai, Z. and Xu, X. (2009). Nonparametric quantile estimations for dynamic smooth coefficient models. Journal of the American Statistical Association 104, 371–383.

Chaudhuri, P. (1991). Global nonparametric estimation of conditional quantile functions and their derivative. Journal of Multivariate Analysis 39, 246–269.

Chaudhuri, Doksum, P. K. and Samarov, A. (1997). On average derivative quantile regression. Annals of Statistics 25, 715–744.

Cook, R. D. and Weisberg, S. (1991), Comment on "Sliced Inverse Regression for Dimension Reduction," by K.-C. Li, Journal of the American Statistical Association 86, 328–332.

Dette, H. and Scheder., R.(2011). Estimation of additive quantile regression. Annals of the Institute of Statistical Mathematics 63, 245–265.

Fan, J. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Gannoun, A., Girard, S. and Saracco, J. (2004). Sliced inverse regression in reference curves estimation. Computational Statistics and Data Analysis 46, 103–122.

Gooijer, J. G. De and Zerom, D. (2003). On additive conditional quantiles with high-dimensional covariates. Journal of the American Statistical Association 98, 135–146.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. Annals of Statistics 21, 157–178.

Härdle, W. and Stoker, T. (1989). Investing smooth multiple regression by the method of average derivatives. Journal of the American Statistical Association 84, 986–995.

He, X. and Shi, P. (1996). Bivariate tensor-product B-splines in a partly linear model. Journal of Multivariate Analysis 58, 162–181.

He, X., Zhu, Z. and Fung, W. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. Biometrika 89, 579–590.

Hristache, M., Juditski, A. and Spokoiny, V. (2001). Direct estimation of the index coefficients in a single-index model. Annals of Statistics 29, 595–623.

Horowitz, J. L and Lee S. (2005). Nonparametric Estimation of an Additive Quantile Regression Model. Journal of the American Statistical Association 100, 1238–1249.

Hua, Y., Gramacy, R. B. and Lian, H. (2012). Bayesian quantile regression for single-index models. Statistics and Computing, to appear; preprint on arXiv:1110.0219.

Ichimura, H. (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models. Journal of Econometrics 58, 71–120.

Jiang, R., Zhou, Z. G., Qian, W. M. and Shao, W. Q (2012). Single-index composite quantile regression. Journal of the Korean Statistical Society 3, 323–332.

Kai, B., Li, R. and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. Annals of Statistics 39, 305–332.

Koenker, R. (2004). Quantile regression for longitudinal data. Journal of Multivariate Analysis 91, 74–89.

Koenker, R. and Bassett, G. (1978). Regression quantiles. Econometrica 46, 33–50.

Koenker, R. (2005). Quantile Regression, Cambridge, U.K.: Cambridge University Press.

Kong, E. and Xia, Y. (2007). Variable Selection for the Single-Index Model. Biometrika 94, 217–229.

Kong, E. and Xia, Y. (2012). A single-index quantile regression model and its estimation. Econometric Theory 28, 730–768.

Lee, S. (2003). Efficient semi parametric estimation of a partially linear quantile regression model. Econometric Theory 19, 1–31.

Li, B. and Wang, S. L. (2007). On directional regression for dimension reduction. Journal of the American Statistical Association 102, 997–1008.

Li, K. C. (1991). Sliced Inverse Regression for Dimension Reduction (with discussion). Journal of the American Statistical Association 86, 316–342.

Li, L. and Yin, X. (2008). Sliced Inverse Regression with Regularizations. Biometrics 64, 124–131.

Li, Q., Xi, R. and Lin, N. (2010). Bayesian Regularized Quantile Regression. Bayesian Analysis 5, 1–24.

Li, Y. and Zhu, J. (2008). l_1-norm quantile regressions. Journal of Computational and Graphical Statistics 17, 163–185.

Naik, P. A. and Tsai, C.-L. (2001). Single-Index Model Selections. Biometrika 88, 821–832.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Ser. B 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society, Series B 67**,** 91–108.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: sparse MAVE. Computational Statistics and Data Analysis 52, 4512–4520.

Wang, J. L., Xue, L. G., Zhu, L. X. and Chong, Y. S. (2010). Estimation for a partial-linear single index model. Annals of Statistics 38, 246–274.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. Statistica Sinica 19, 801–817.

Wu, T. Z., Yu, K. and Yu, Y. (2010). Single index quantile regression. Journal of Multivariate Analysis 101, 1607–1621.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002), "An Adaptive Estimation of Dimension Reduction Space" (with discussion), Journal of the Royal Statistical Society, Ser B 64, 363–410.

Yuan, Y. and Yin, G. (2010). Bayesian quantile regression for longitudinal studies with non-ignorable missing data. Biometrics 66, 105–114.

Yebin, C., Gooijer , J. G. D. and Zerom, D. (2011). Efficient estimation of an additive quantile regression model. Scandinavian Journal of Statistics 38, 46–62.

Yu, K., Lu, Z. and Stander, J. (2003). Quantile regression: Applications and current research areas. The Statistician 52, 331–350.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. Journal of the American Statistical Association 93, 228–237.

Yu, K. and Lu, Z. (2004). Local linear additive quantile regression. Scandinavian Journal of Statistics 31, 333–346.

Zeng, P., He, T. and Zhu Y. (2012). A Lasso-type approach for estimation and variable selection in single index models. Journal of Computational and Graphical Statistics 21, 92–109.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

Chapter 5

Penalised Flexible Bayesian quantile regression[4]

Selecting an appropriate subset of predictors can help to develop prediction precision and interpretation. In this chapter, we proposed two regularisation approaches, the flexible Bayesian Lasso quantile regression and its adaptive version. The proposed methods have been compared with three existing methods. Extensive simulation studies and a study based on real data using the body fat dataset are conducted in order to examine the performance of the methods under consideration. The proposed methods perform well in comparison to the other methods in terms of the median mean squared error (MMSE), mean and the standard deviation (SD) criteria of the absolute correlation $|r|$, where the median, mean and SD are taken over the number of simulations. The results suggest that the proposed methods are useful practically.

[4]This chapter is based on: Alkenani, A., Alhamzawi, R. and Yu, K. (2012). Penalized Flexible Bayesian Quantile Regression. Applied Mathematics 3, 2155–2168. http://dx.doi.org/10.4236/am.2012.312A

296

5.1. Introduction

As pointed out in Section 1.1 and Section 4.1, quantile regression (QR) has become a widespread technique which can be used to describe the distribution of an outcome variable, given a set of predictors. It has been employed in many areas such as econometrics, social sciences, microarrays and agricultural studies, see Koenker (2005) for an overview.

Let y_i be a response variable and $\mathbf{x}_i$ a $p\times1$ vector of predictors for the ith obsevation, $q_\tau(\mathbf{x}_i)$ is the inverse cumulative distribution function (ICDF) of y_i given $\mathbf{x}_i$. Then, the relationship between $q_\tau(\mathbf{x}_i)$ and $\mathbf{x}_i$ can be modelled as $q_\tau(\mathbf{x}_i) = \mathbf{x}_i^T\boldsymbol{\beta}_\tau$, where $\boldsymbol{\beta}_\tau$ is a vector of p unknown parameters and τ determines the quantile level.

According to Koenker and Bassett (1978), $\boldsymbol{\beta}_\tau$ can be estimated by

$$\min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{n} \rho_\tau\,(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_\tau), \tag{5.1}$$

where $\rho_\tau(.)$ is the check loss function defined by

$$\rho_\tau(u) = \tau\, u\, I_{[0,\infty)}(u) - (1-\tau)\, u\, I_{(-\infty,0)}(u). \tag{5.2}$$

As a possible parametric link with minimising the check loss function (5.1), Koenker and Machado (1999) showed that the maximum likelihood solution of the asymmetric Laplace distribution (ALD) is equivalent to the minimisation problem in (5.1). Later, this idea was exploited by Yu and Moyeed (2001). The authors suggested a fully Bayesian approach for QR (BQR) under ALD error distribution. Recently, Bayesian approaches for QR have attracted much significant interest in the literature. For example, Tsionas (2003) developed a Gibbs sampling (GS) algorithm for the QR model, while Yu and Standard (2007) proposed Bayesian Tobit QR. Additionally, Geraci and Bottai (2007) considered BQR for longitudinal data using an ALD.

Likewise, Reed and Yu (2009) and Kozumi and Kobayashi (2009) proposed a GS algorithm based on a location-scale mixture representation of the ALD while Benoit and Poel (2011) proposed Bayesian binary QR.

Some researchers suggested nonparametric methods in order to avert the restrictive assumptions of the parametric approaches. See for example Walker and Mallick (1999), Kottas and Gelfand (2001), Hanson and Johnson (2002), Hjort (2003), Hjort and Petrone (2007), Taddy and Kottas (2007) and Kottas and Krnjajic (2009). Recently, Reich et al. (2010) proposed the Flexible Bayesian Quantile Regression (FBQR) approach. The authors assumed that the distribution of the error is an infinite mixture of Gaussian (IMG) densities. They called their method "flexible" because it does not impose parametric assumptions (e.g., ALD) or shape restrictions on the residual distribution (e.g., mode at the quantile of interest), as with other approaches (personal communication with Reich).

As pointed out in the previous chapters, selection of the important predictors from the original predictors is crucial for building a good multiple regression models. Subset selection by penalising the ordinary least squares has attracted considerable research interest. For example see, Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001) and adaptive Lasso (Zou, 2006).

Although the well-known classical least squares approach has many good mathematical properties, it is sensitive and is not robust to outliers (Bradic et al., 2011; Koenker and Bassett, 1978). However, robust variable selection can be achieved using a rigorous method, such as QR.

Koenker (2004) suggested using the regularisation method in conjunction with the QR model. The author placed an l_1 penalty term on the random effects in a mixed-effect QR model. Yuan and Yin (2010) suggested a Bayesian method to shrink the

random effects via adding an l_2 penalty term to the QR check function. In addition, Wang et al. (2007) suggested merging the LAD and the Lasso in order to obtain robust parameter estimation and variable selection simultaneously. Li and Zhu (2008) developed the piecewise linear solution approach of the l_1 penalised QR. Furthermore, Wu and Liu (2009) considered regularised QR with the SCAD and adaptive Lasso. Li et al. (2010) suggested Bayesian regularised QR. The authors proposed different regularisation methods from a Bayesian viewpoint, such as Lasso, elastic net and group Lasso. Alhamzawi et al. (2012) proposed Bayesian adaptive Lasso quantile regression (BALQR), which gives different penalisation parameters to different regression coefficients.

In this chapter, we evolve a flexible Bayesian framework for regularisation in the QR model. Similar to Reich et al. (2010), we assume the error distribution to be the IMG densities. This work is different from Bayesian Lasso quantile regression (BLQR) employing the ALD for the error. In fact, the use of the ALD is unfavourable due to the lack of coherence (Kottas and Krnjaji´c, 2009). For example, for a different τ we have a different distribution for y_i's and it is difficult to resolve these differences. Our motivating example is an analysis of the Body fat (BF) data which is previously analysed by Johnson (1996) and is available in "mfp" package. This study included total body measurements of 252 men. The aim is to explore the relationship between the percentage of the BF and 13 simple body measurements. In this study, we want to choose the most important simple body measurements for the QR model, relating to the percentage BF. High correlations are existent between the predictors in the BF data. For example, the correlation is 0.943 between the weight and the hip circumference, 0.916 between the chest circumference and the abdomen circumference, 0.894 between the hip circumference and the thigh circumference, 0.894 between the weight and the chest

circumference, 0.887 between the weight and the abdomen circumference, 0.874 between the abdomen circumference and the hip circumference and so on. The subset selection is significant in this data, in order to know which predictors have coefficients that vary among subjects. The high correlation between the predictors is an excuse in favour of using Bayesian adaptive Lasso because it deals with correlated predictors by using different weights for the different predictors.

The rest of this chapter is organised as follows. A short review of the FBQR model for independent data is given in Section 5.2. The FBLQR and FBALQR are proposed in Section 5.3 and Section 5.4, respectively. Simulations studies are implemented and applications of the proposed methods on real data are given in Section 5.5. Finally, the conclusions are summarised in Section 5.6.

5.2. Flexible Bayesian Quantile Regression (FBQR)

Following He (1997), Reich et al. (2010) considered the heteroscedastic linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{x}_i^T \boldsymbol{\gamma}\, \varepsilon_i, \tag{5.3}$$

where $\mathbf{x}_i^T \boldsymbol{\gamma} > 0$ for all $\mathbf{x}_i$ and ε_i are independent and identically distributed. The authors rewrote the above model as a QR model:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_{(\tau)} + \mathbf{x}_i^T \boldsymbol{\gamma}_{(\tau)}\, \varepsilon_{i(\tau)}, \tag{5.4}$$

where $\varepsilon_{i(\tau)} = \varepsilon_i - q_\tau(\varepsilon)$ has τth quantile equal to 0, $q_\tau(\varepsilon)$ is the ICDF of ε_i. In order to analyse y_i's τth quantile $\mathbf{x}_i^T \boldsymbol{\beta}_{(\tau)}$, the authors only considered distributions for ε_i with τth quantile equal to 0. Also, they fixed the element of $\boldsymbol{\gamma}_{(\tau)}$, corresponding to the

intercept at 1, in order to separate out the scale of the errors from $\boldsymbol{\gamma}_{(\tau)}$. The subscript τ is omitted in the rest of the chapter for notational convenience.

Reich et al. (2010) suggested a fully Bayesian approach for QR inference. The authors proposed a flexible residual distribution as an IMG densities. They assumed ε_i distribution as follows:

$$\hbar(\varepsilon|\mu,\sigma^2) = \sum_{\mathfrak{m}=1}^{\infty} P_{\mathfrak{m}}\, f\left(\varepsilon\middle|\mu_{\mathfrak{m}},\sigma_{\mathfrak{m}}^2,\mathfrak{q}_{\mathfrak{m}}\right), \tag{5.5}$$

where $P_{\mathfrak{m}}$ are mixture proportions with $\sum_{\mathfrak{m}=1}^{\infty} P_{\mathfrak{m}} = 1$ and $f\left(\varepsilon\middle|\mu_{\mathfrak{m}},\sigma_{\mathfrak{m}}^2,\mathfrak{q}_{\mathfrak{m}}\right)$ is given by

$$f\left(\varepsilon\middle|\mu_{\mathfrak{m}},\sigma_{\mathfrak{m}}^2,\mathfrak{q}_{\mathfrak{m}}\right) = \mathfrak{q}_{\mathfrak{m}}\ N\left(\mu_{1\mathfrak{m}},\sigma_{1\mathfrak{m}}^2\right) + \left(1-\mathfrak{q}_{\mathfrak{m}}\right)\ N\left(\mu_{2\mathfrak{m}},\sigma_{2\mathfrak{m}}^2\right), \tag{5.6}$$

where the mixture proportion $\mathfrak{q}_{\mathfrak{m}} \in (0,1)$ is given by

$$\mathfrak{q}_{\mathfrak{m}} = \frac{\tau-\Phi(-\mu_{2\mathfrak{m}}/\sigma_{2\mathfrak{m}})}{\Phi(-\mu_{1\mathfrak{m}}/\sigma_{1\mathfrak{m}})-\Phi(-\mu_{2\mathfrak{m}}/\sigma_{2\mathfrak{m}})}. \tag{5.7}$$

Here, Φ refers to the $N(0,1)$ distribution, $\sigma_{1\mathfrak{m}},\sigma_{2\mathfrak{m}}$~uniform$(0,c_1)$ for large constant c_1 and $\mu_{1\mathfrak{m}},\mu_{2\mathfrak{m}}$~ALD $(0,\vartheta,\tau)$, where the parameters are the location, scale and the skewness, respectively. The prior for the scale parameter ϑ is Gamma(0.1,0.1). The proportions $P_{\mathfrak{m}}$ are defined via the latent variables Ʉ_m which are independently and identically distributed from beta$(1,Đ)$, where Đ controls the strength of the prior for $P_{\mathfrak{m}}$. The first proportion is $P_1 = \text{Ʉ}_1$ and the others are given by $P_{\mathfrak{m}} = \text{Ʉ}_{\mathfrak{m}} \prod_{\mathfrak{d}<\mathfrak{m}}(1-\text{Ʉ}_{\mathfrak{d}})$ for $\mathfrak{m} \geq 2$.

Reich et al. (2010) rewrite the described model as a mixture model by introducing latent variables $G_i \in \{1,2,\ldots\}$ and $H_i \in \{1,2\}$ as follows

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{x}_i^T\boldsymbol{\gamma}\,\varepsilon_i,$$

where $\varepsilon_i \sim \mathrm{N}\left(\mu_{H_iG_i}, \sigma^2_{H_iG_i}\right)$, $G_i \sim$Categorical$(P_1, P_2, \dots)$ and $H_i \sim$ Categorical$\left(\mathrm{q}_{G_i}, 1 - \mathrm{q}_{G_i}\right)$.

They assign a normal prior distribution for each $\beta_k, (k = 1,2, \dots, p)$ with mean zero and variance c_2. The prior for γ_k is a vague normal prior subject to $\mathbf{x}_i^T \boldsymbol{\gamma} > 0$ for all $\mathbf{x}_i$. Under these assumptions, the conditional distribution of y_i, given $\mu_{H_iG_i}$ and $\sigma^2_{H_iG_i}$ is as follows:

$$f\left(y_i \middle| \mu_{H_iG_i}, \sigma^2_{H_iG_i}\right) \propto \left(\mathbf{x}_i^T \boldsymbol{\gamma} \sigma_{H_iG_i}\right)^{-1} exp\left\{-\frac{\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{x}_i^T \boldsymbol{\gamma} \mu_{H_iG_i}\right)^2}{2\left(\mathbf{x}_i^T \boldsymbol{\gamma} \sigma_{H_iG_i}\right)^2}\right\}, \qquad (5.8)$$

5.3. Flexible Bayesian Quantile Regression with Lasso penalty (FBLQR)

As mentioned in Section 4.3, Tibshirani (1996) proposed the Lasso for simultaneous variable selection and parameter estimation. As a possible link with Bayesian inference, the author showed that if the regression coefficients have independent and identical Laplace priors, the Lasso estimates can be interpreted as posterior mode estimates. This connection motivated Park and Casella (2008) and Hans (2009) to suggest Lasso-based models from a Bayesian perspective. Li et al. (2010) extended the idea of Bayesian Lasso (BL) regression to Bayesian Lasso quantile regression (BLQR). In BLQR, the ALD for the error is employed. Kottas and Krnjaji´c (2009) and Reich et al. (2010) showed that the use of the ALD is undesirable because of the deficiency of coherence.

In this section, a flexible Bayesian framework for regularisation in QR is developed. Similar to Reich et al. (2010), the error distribution is assumed to be the IMG densities. We propose FBLQR minimises

$$(\text{Ų} - \mathbf{x}_i^T\boldsymbol{\beta})^T\boldsymbol{W}^{-1}(\text{Ų} - \mathbf{x}_i^T\boldsymbol{\beta}) + \lambda\sum_{k=1}^{p}|\beta_k|, \tag{5.9}$$

where $\text{Ų} = y_i - \mathbf{x}_i^T\boldsymbol{\gamma}\mu_{H_iG_i}$ and $\boldsymbol{W}$ is a diagonal matrix with the element $\left(\mathbf{x}_i^T\boldsymbol{\gamma}\sigma_{H_iG_i}\right)^2$ on the diagonal i.

A Laplace prior on β_k has been considered, taking the form of $f(\beta_k/\lambda) = \frac{\lambda}{2}\, e^{-\lambda\,|\beta_k|}$, which can be represented as a member of a scale mixture of normals (SMN) (Andrews and Mallows, 1974).

$$\frac{\lambda}{2}e^{-\lambda|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}}e^{-z^2/(2s)}\frac{\lambda^2}{2}e^{-\lambda^2 s/2}ds,\ \lambda > 0 \tag{5.10}$$

Then the $f(\beta_k/\lambda)$ can be written as:

$$f(\beta_k/\lambda) = \frac{\lambda}{2}\, e^{-\lambda\,|\beta_k|} = \int_0^\infty \frac{1}{\sqrt{2\pi s_k}}e^{-\beta_k^2/(2s_k)}\frac{\lambda^2}{2}e^{-\lambda^2 s_k/2}ds_k. \tag{5.11}$$

We consider gamma priors, $f(\lambda^2) \propto \lambda^{2(c_3-1)}e^{(-c_4\lambda^2)}$, on λ^2 (not λ). Then, we have the following hierarchical model:

$$y_i/\mu_{H_iG_i}, \sigma^2_{H_iG_i} \sim N\left(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{x}_i^T\boldsymbol{\gamma}\mu_{H_iG_i},\ \left(\mathbf{x}_i^T\boldsymbol{\gamma}\sigma_{H_iG_i}\right)^2\right),$$

$$\beta_k, s_k/\lambda^2 \sim \frac{1}{\sqrt{2\pi s_k}}e^{-\beta_k^2/(2s_k)}\frac{\lambda^2}{2}e^{-\lambda^2 s_k/2},$$

$$\lambda^2 \sim \lambda^{2(c_3-1)}e^{(-c_4\lambda^2)},$$

$$\mu_{1\text{ŋ}}, \mu_{2\text{ŋ}} \sim \text{ALD}\ (0, \vartheta, \tau\ ),$$

$$\vartheta\ \sim \text{Gamma}(0.1, 0.1),$$

$$\sigma_{1\text{ŋ}}, \sigma_{2\text{ŋ}} \sim \text{Uniform}(0, c_1),$$

$$G_i \sim \text{Categorical}(P_1, P_2, \dots),$$

$$H_i \sim \text{Categorical}\left(\text{q}_{G_i}, 1 - \text{q}_{G_i}\right),$$

$$q_m = \frac{\tau - \Phi(-\mu_{2m}/\sigma_{2m})}{\Phi(-\mu_{1m}/\sigma_{1m}) - \Phi(-\mu_{2m}/\sigma_{2m})},$$

where $P_1 = V_1$, $P_m = V_m \prod_{k<m}(1 - V_k)$ and the latent variables V_m are independently and identically distributed from beta$(1, \text{Đ})$. The details of the Gibbs sampler are given in the appendix.

5.4. Flexible Bayesian quantile regression with Adaptive Lasso penalty (FBALQR)

As pointed out in Sections 3.4 and 4.4, Zou (2006) suggested the adaptive version of the l_1 norm via employing different weights onto different regression coefficients. From a Bayesian viewpoint, Bayesian adaptive Lasso (BAL) was considered by Griffin and Brown (2007) and Sun et al. (2010). Later, Alhamzawi et al. (2012) proposed the Bayesian adaptive Lasso QR (BALQR) using the ALD for the errors. Employing an ALD in Bayesian QR is undesirable, therefore we propose the FBALQR minimises:

$$(Y - \mathbf{x}_i^T \boldsymbol{\beta})^T \boldsymbol{W}^{-1} (Y - \mathbf{x}_i^T \boldsymbol{\beta}) + \sum_{k=1}^{p} \lambda_k |\beta_k| \qquad (5.12)$$

In this section, a Laplace prior on β_k has been proposed taking the formula, $f(\beta_k/\lambda_k) = \frac{\lambda_k}{2} e^{-\lambda_k|\beta_k|}$, which can be interpreted as the SMN (Andrews and Mallows, 1974)

$$\frac{\lambda_k}{2} e^{-\lambda_k|z_k|} = \int_0^\infty \frac{1}{\sqrt{2\pi s_k}} e^{-z_k^2/(2s_k)} \frac{\lambda_k^2}{2} e^{-\lambda_k^2 s_k/2} ds_k. \qquad (5.13)$$

Then the $f(\beta_k/\lambda_k)$ can be written as:

$$f(\beta_k/\lambda_k) = \frac{\lambda_k}{2} e^{-\lambda_k|\beta_k|} = \int_0^\infty \frac{1}{\sqrt{2\pi s_k}} e^{-\beta_k^2/(2s_k)} \frac{\lambda_k^2}{2} e^{-\lambda_k^2 s_k/2} ds_k. \qquad (5.14)$$

Furthermore, we assume a gamma prior on λ_k^2, $f(\lambda_k^2) \propto \lambda_k^{2(c_3-1)} e^{(-c_4\lambda_k^2)}$. To summarise, we propose the following hierarchical Bayesian model:

$$y_i/\mu_{H_iG_i}, \sigma^2_{H_iG_i} \sim N\left(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{x}_i^T\boldsymbol{\gamma}\mu_{H_iG_i},\ \left(\mathbf{x}_i^T\boldsymbol{\gamma}\sigma_{H_iG_i}\right)^2\right),$$

$$\beta_k, s_k/\lambda_k^2 \sim \frac{1}{\sqrt{2\pi s_k}} e^{-\beta_k^2/(2s_k)} \frac{\lambda_k^2}{2} e^{-\lambda_k^2 s_k/2},$$

$$\lambda_k^2 \sim \lambda_k^{2(c_3-1)} e^{(-c_4\lambda_k^2)},$$

$$\mu_{1ɱ}, \mu_{2ɱ} \sim \text{ALD}\ (0, \vartheta, \tau\ ),$$

$$\vartheta\ \sim \text{Gamma}(0.1,0.1),$$

$$\sigma_{1ɱ}, \sigma_{2ɱ} \sim \text{Uniform}(0, c_1),$$

$$G_i \sim \text{Categorical}(P_1, P_2, \ldots),$$

$$H_i \sim \text{Categorical}\left(\mathrm{q}_{G_i}, 1 - \mathrm{q}_{G_i}\right),$$

$$\mathrm{q}_{ɱ} = \frac{\tau - \Phi(-\mu_{2ɱ}/\sigma_{2ɱ})}{\Phi(-\mu_{1ɱ}/\sigma_{1ɱ}) - \Phi(-\mu_{2ɱ}/\sigma_{2ɱ})},$$

where $P_1 = Ⱶ_1$, $P_ɱ = Ⱶ_ɱ \prod_{đ<ɱ}\left(1 - Ⱶ_ɱ\right)$ and the latent variables $Ⱶ_ɱ$ are independently and identically distributed from $\text{beta}(1, Đ)$.

5. 5. A simulation study

A numerical study was implemented in order to assess the behaviour of the proposed methods. We have generated $\mathcal{R} = 200$ data-sets with size $n = 300$ observations from $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{x}_i^T \boldsymbol{\gamma}\, \varepsilon_i$, where $\mathbf{x}_i$ are generated as independently and identically distributed standard normals. The error ε_i is simulated from three possible error distributions: $N(0,1)$, a $t_{(3)}$ distribution with 3 D.F and $\chi^2_{(3)}$ with 3 D.F. The following designs for the vector $\boldsymbol{\beta}$ are assumed:

Design 1: $\boldsymbol{\beta} = (1; -1; -1; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0)^T$

Design 2: $\boldsymbol{\beta} = (3; -3; -3; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0)^T$

Design 3: $\boldsymbol{\beta} = (1; -1; -1; 0; 0; 0; 0; 0; 0; 0; 0; 0; 1; -1; -1)^T$

Design 4: $\boldsymbol{\beta} = (3; -3; -3; 0; 0; 0; 0; 0; 0; 0; 0; 0; 3; -3; -3)^T$,

where the first element in $\boldsymbol{\beta}$ corresponds to the intercept.

Each simulated data set is analysed via five methods. The FBLQR and FBALQR, which are proposed in Sections 5.3 and 5.4 respectively, are compared with the Lasso quantile regression (LQR), the standard frequentist (QR) and the FBQR. The LQR and the standard frequentist (QR) are implemented using the "quantreg" package in R. We run our algorithm for 15000 iteration discarding the first 5000. We set $Đ = 1$, $c_1 = 10$, $c_2 = 100$ and $c_3 = c_4 = 0.1$.

To compare the performance of the estimators, we report the mean and SD of $|r|$ between $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}^T\boldsymbol{\beta}$ and the median of the mean squared error (MMSE) of $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$.

Table 5.1. Simulation results for the FBALQR, FBLQR, FBQR, LQR and QR based on design 1.

τ	**Method**	Error Distribution								
		$N(0,1)$			$t_{(3)}$			$\chi^2_{(3)}$		
		$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE	$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE	$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE
0.25	FBALQR	0.9972	0.0013	0.0034	0.9949	0.0022	0.0118	0.9946	0.0017	0.0154
	FBLQR	0.9968	0.0012	0.0036	0.9946	0.0024	0.01495	0.9921	0.0016	0.0167
	FBQR	0.9966	0.0014	0.0045	0.9935	0.0025	0.0191	0.9905	0.0014	0.0262
	LQR	0.9959	0.0014	0.0045	0.9936	0.0039	0.0185	0.9908	0.0014	0.0226
	QR	0.9957	0.0014	0.0070	0.9932	0.0037	0.0201	0.9881	0.0020	0.0330
0.5	FBALQR	0.9971	0.0013	0.0033	0.9955	0.0011	0.0053	0.9854	0.0082	0.0315
	FBLQR	0.9970	0.0013	0.0034	0.9954	0.0011	0.0055	0.9846	0.0073	0.0326
	FBQR	0.9969	0.0014	0.0048	0.9951	0.0024	0.0054	0.9797	0.0073	0.0552
	LQR	0.9958	0.0018	0.0041	0.9953	0.0021	0.0053	0.9822	0.0060	0.0393
	QR	0.9957	0.0018	0.0056	0.9949	0.0013	0.0056	0.9785	0.0066	0.0570
0.75	FBALQR	0.9962	0.0010	0.0050	0.9934	0.0025	0.0157	0.9453	0.0179	0.0534
	FBLQR	0.9960	0.0010	0.0053	0.9933	0.0024	0.0163	0.9433	0.0166	0.0544
	FBQR	0.9958	0.0009	0.0065	0.9932	0.0032	0.0179	0.9385	0.0176	0.0751
	LQR	0.9952	0.0014	0.0055	0.9931	0.0034	0.0169	0.9398	0.0178	0.0612
	QR	0.9950	0.0014	0.0067	0.9930	0.0031	0.0190	0.9362	0.0229	0.0829

Table 5.2. Simulation results for FBALQR, FBLQR, FBQR, LQR and QR based on design 2.

τ	**Method**	Error Distribution								
		$N(0,1)$			$t_{(3)}$			$\chi^2_{(3)}$		
		$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE	$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE	$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE
0.25	FBALQR	0.99964	0.00021	0.00415	0.99335	0.00264	0.02084	0.99949	0.00035	0.03962
	FBLQR	0.99963	0.00013	0.00473	0.99293	0.00271	0.02313	0.99926	0.00055	0.04691
	FBQR	0.99962	0.00022	0.00544	0.99190	0.00233	0.06672	0.99903	0.00036	0.06922
	LQR	0.99951	0.00021	0.00693	0.99241	0.00210	0.05682	0.99913	0.00043	0.06550
	QR	0.99941	0.00023	0.00694	0.99134	0.00312	0.07380	0.99874	0.00043	0.07351
0.5	FBALQR	0.99973	0.00013	0.00377	0.99952	0.00024	0.00254	0.99885	0.00061	0.01026
	FBLQR	0.99972	0.00014	0.00412	0.99951	0.00024	0.00272	0.99873	0.00062	0.01094
	FBQR	0.99971	0.00016	0.00503	0.99943	0.00025	0.00355	0.99790	0.00070	0.02183
	LQR	0.99962	0.00021	0.00605	0.99942	0.00030	0.00471	0.99861	0.00095	0.01263
	QR	0.99961	0.00020	0.00675	0.99941	0.00050	0.00493	0.99781	0.00095	0.02541
0.75	FBALQR	0.99972	0.00022	0.00342	0.99923	0.00033	0.01215	0.99589	0.00212	0.01207
	FBLQR	0.99961	0.00022	0.00391	0.99915	0.00033	0.01342	0.99574	0.00157	0.01765
	FBQR	0.99960	0.00024	0.00471	0.99882	0.00036	0.04641	0.99432	0.00243	0.05400
	LQR	0.99953	0.00022	0.00614	0.99911	0.00050	0.03240	0.99515	0.00291	0.05099
	QR	0.99951	0.00023	0.00690	0.99878	0.00051	0.04671	0.99391	0.00292	0.07122

Table 5.3. Simulation results for FBALQR, FBLQR, FBQR, LQR and QR based on design 3.

τ	**Method**	Error Distribution								
		$N(0,1)$			$t_{(3)}$			$\chi^2_{(3)}$		
		$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE	$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE	$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE
0.25	FBALQR	0.99868	0.00042	0.00602	0.99755	0.00128	0.01001	0.99699	0.00066	0.01391
	FBLQR	0.99855	0.00044	0.00603	0.99744	0.00123	0.01043	0.99654	0.00091	0.01661
	FBQR	0.99754	0.00060	0.00624	0.99713	0.00132	0.01180	0.99603	0.00163	0.02303
	LQR	0.99853	0.00051	0.00613	0.99731	0.00130	0.01122	0.99615	0.00100	0.02275
	QR	0.99732	0.00063	0.00643	0.99701	0.00142	0.01263	0.99586	0.00157	0.02370
0.5	FBALQR	0.99875	0.00052	0.00435	0.99852	0.00072	0.00593	0.99541	0.00173	0.01253
	FBLQR	0.99872	0.00054	0.00484	0.99824	0.00074	0.00664	0.99511	0.00180	0.01727
	FBQR	0.99823	0.00056	0.00543	0.99811	0.00077	0.00690	0.99270	0.00196	0.02465
	LQR	0.99863	0.00055	0.00495	0.99815	0.00076	0.00685	0.99491	0.00190	0.01965
	QR	0.99821	0.00059	0.00563	0.99802	0.00079	0.00688	0.99225	0.00201	0.03368
0.75	FBALQR	0.99835	0.00052	0.00434	0.99765	0.00080	0.01025	0.98729	0.00471	0.04346
	FBLQR	0.99833	0.00055	0.00430	0.99757	0.00100	0.01044	0.98661	0.00505	0.04612
	FBQR	0.99812	0.00060	0.00476	0.99752	0.00112	0.01699	0.98354	0.00547	0.05250
	LQR	0.99815	0.00058	0.00441	0.99755	0.00111	0.01652	0.98453	0.00521	0.04792
	QR	0.99803	0.00061	0.00478	0.99754	0.00115	0.01910	0.98255	0.00574	0.05293

Table 5.4. Simulation results for FBALQR, FBLQR, FBQR, LQR and QR based on design 4.

τ	**Method**	Error Distribution								
		$N(0,1)$			$t_{(3)}$			$\chi^2_{(3)}$		
		$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE	$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE	$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	MMSE
0.25	FBALQR	0.99994	0.00011	0.00435	0.99977	0.00010	0.02063	0.99978	0.00011	0.01440
	FBLQR	0.99991	0.00014	0.00482	0.99975	0.00012	0.03141	0.99976	0.00013	0.01622
	FBQR	0.99983	0.00018	0.00510	0.99973	0.00017	0.03340	0.99973	0.00022	0.01678
	LQR	0.99985	0.00015	0.00495	0.99974	0.00014	0.03272	0.99974	0.00021	0.01673
	QR	0.99982	0.00022	0.00530	0.99969	0.00019	0.04525	0.99963	0.00023	0.01701
0.50	FBALQR	0.99996	0.00011	0.00114	0.99989	0.00012	0.00844	0.99962	0.00024	0.02352
	FBLQR	0.99993	0.00013	0.00122	0.99987	0.00014	0.00864	0.99941	0.00031	0.02796
	FBQR	0.99984	0.00022	0.00262	0.99982	0.00017	0.01826	0.99921	0.00036	0.03619
	LQR	0.99992	0.00017	0.00127	0.99985	0.00015	0.00958	0.99933	0.00034	0.02842
	QR	0.99982	0.00023	0.00300	0.99980	0.00021	0.02028	0.99906	0.00037	0.03653
0.75	FBALQR	0.99987	0.00009	0.00286	0.99978	0.00012	0.02270	0.99876	0.00050	0.04221
	FBLQR	0.99986	0.00011	0.00445	0.99976	0.00013	0.02281	0.99852	0.00062	0.04271
	FBQR	0.99980	0.00020	0.00563	0.99970	0.00017	0.02403	0.99810	0.00073	0.04899
	LQR	0.99983	0.00014	0.00460	0.99972	0.00015	0.02322	0.99821	0.00065	0.04754
	QR	0.99977	0.00024	0.00670	0.99964	0.00020	0.02470	0.99791	0.00076	0.05685

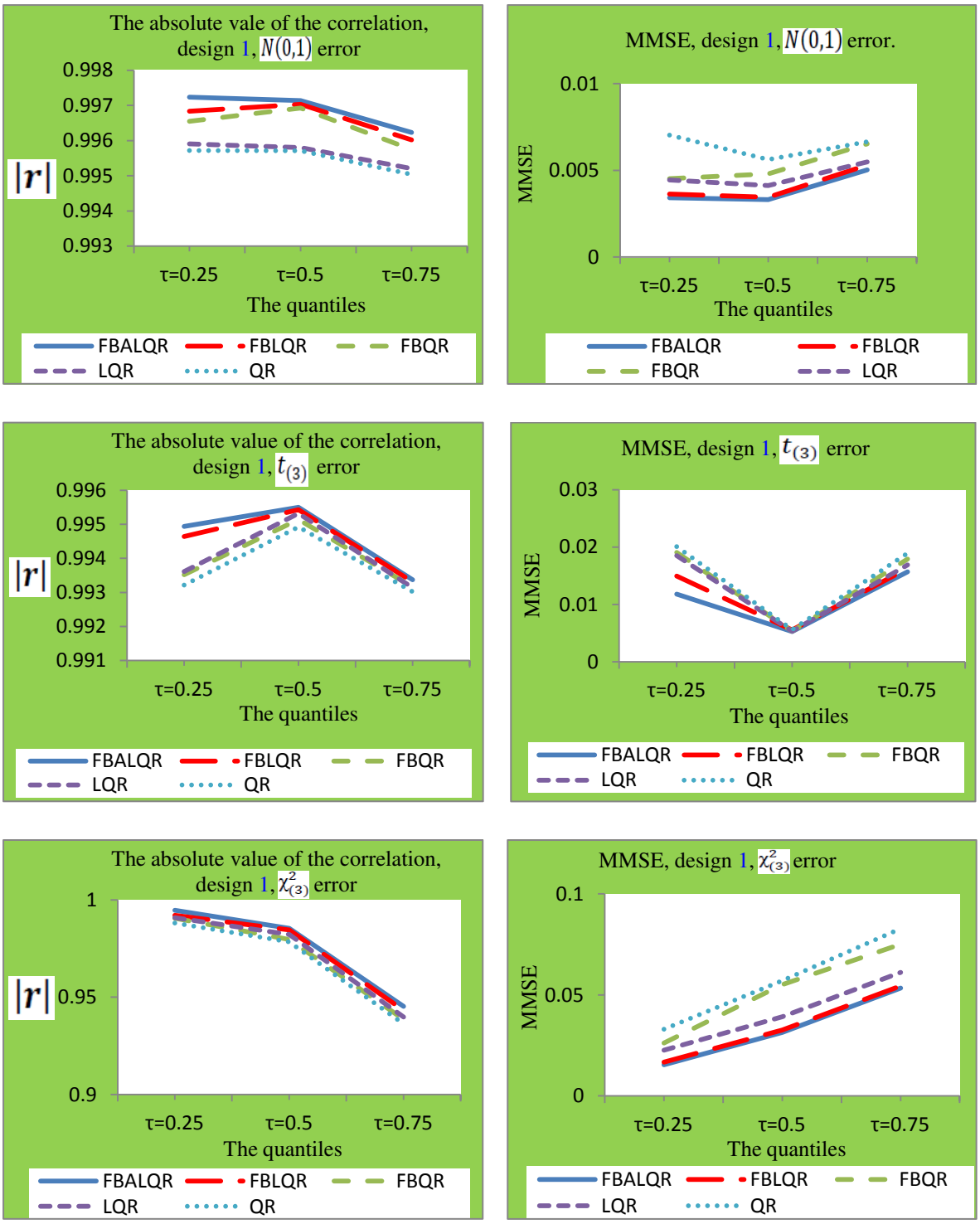


Figure 5.1. The left column explains the plots for the $|r|$ between $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}^T\boldsymbol{\beta}$ for design 1, where the error distributions are $N(0,1)$, $t_{(3)}$ and $\chi^2_{(3)}$, respectively. The right column explains the plots for the MMSE of $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ for design 1, where the error distributions are $N(0,1)$, $t_{(3)}$ and $\chi^2_{(3)}$, respectively.

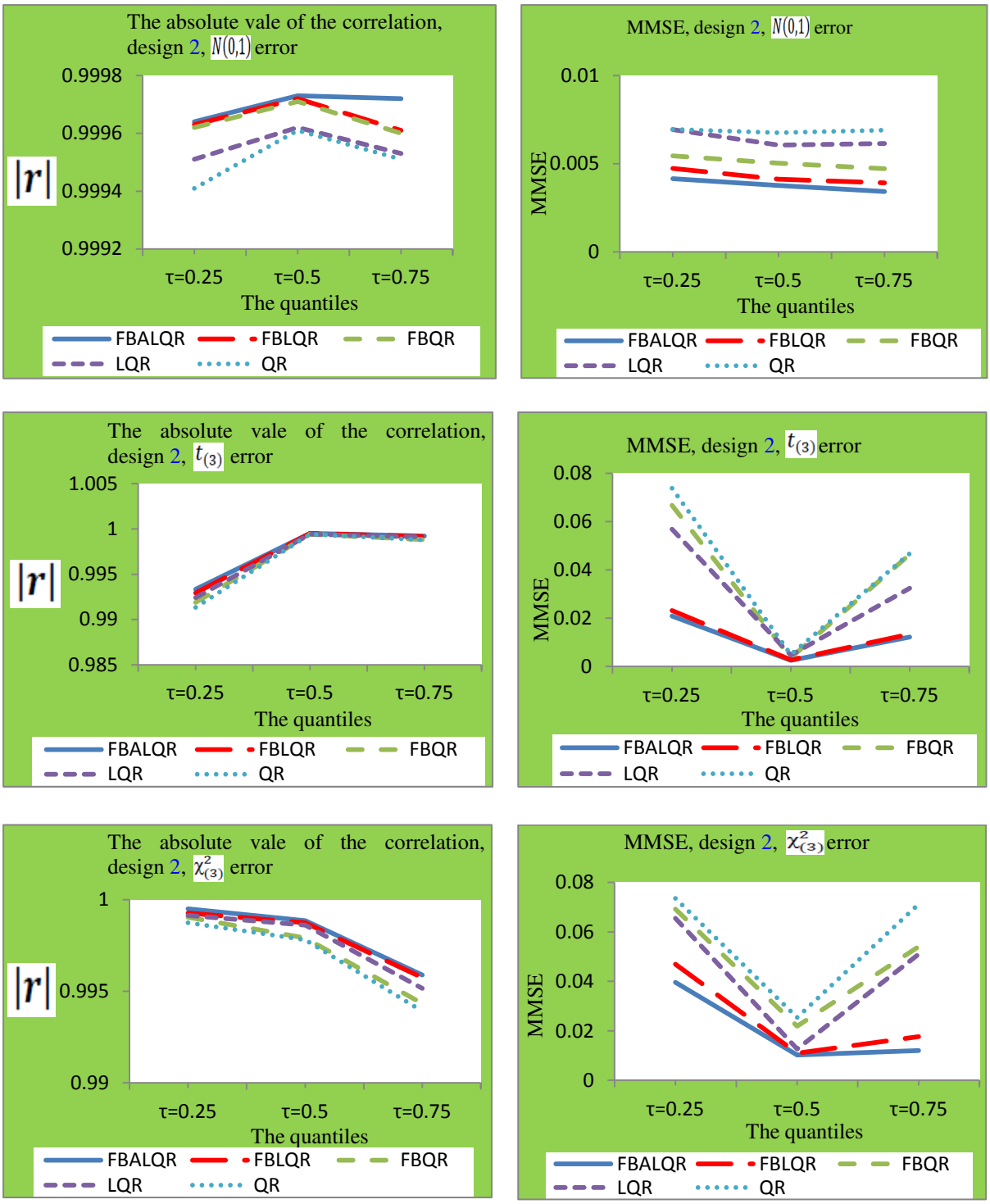


Figure 5.2. The left column explains the plots for $|r|$ between $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}^T\boldsymbol{\beta}$ for design 2, where the error distributions are $N(0,1)$, $t_{(3)}$ and $\chi^2_{(3)}$, respectively. The right column explains the plots for the MMSE of $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ for design 2, where the error distributions are $N(0,1)$, $t_{(3)}$ and $\chi^2_{(3)}$, respectively.

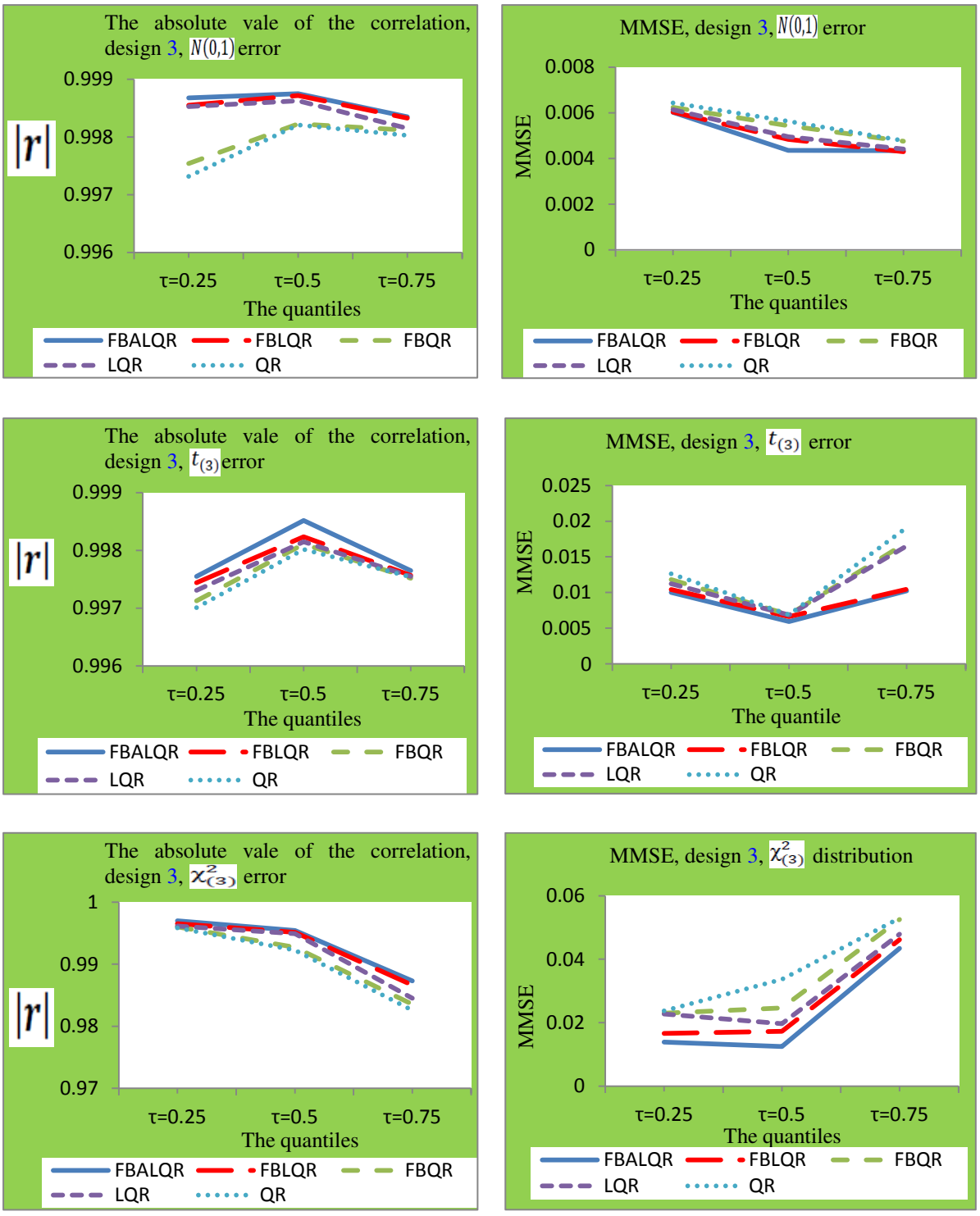


Figure 5.3. The left column explains the plots for $|r|$ between $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}^T\boldsymbol{\beta}$ for design 3, where the error distributions are $N(0,1)$, $t_{(3)}$ and $\chi^2_{(3)}$, respectively. The right column explains the plots for the MMSE of $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ for design 3, where the error distributions are $N(0,1)$, $t_{(3)}$ and $\chi^2_{(3)}$, respectively.

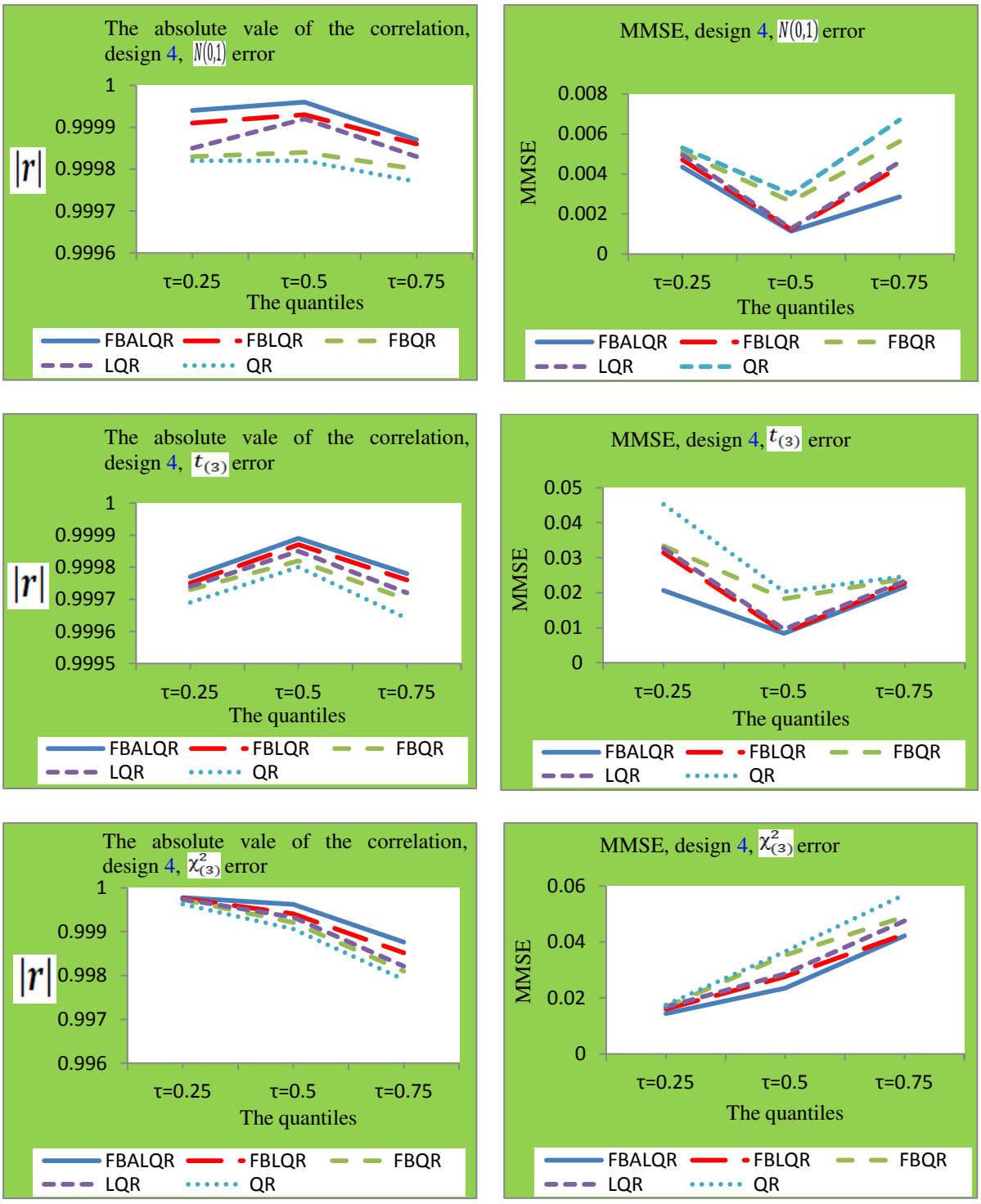


Figure 5.4. The left column explains the plots for $|r|$ between $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}^T\boldsymbol{\beta}$ for design 4, where the error distributions are $N(0,1)$, $t_{(3)}$ and $\chi^2_{(3)}$, respectively. The right column explains the plots for the MMSE of $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ for design 4, where the error distributions are $N(0,1)$, $t_{(3)}$ and $\chi^2_{(3)}$, respectively.

The results of the simulation are presented in Tables 5.1–5.4 and Figures 5.1–5.4. From Tables 5.1–5.4 and Figures 5.1–5.4 and for all of the distributions under consideration, it can be observed that the results of $|r|$ between $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}^T\boldsymbol{\beta}$ for the proposed methods are higher than the other methods, suggesting a good performance from the FBALQR and FBLQR. Instead of looking at $|r|$, we may also look at the MMSE for $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$. The results of the MMSE also support the good characteristics of the FBALQR and FBLQR. This shows that the FBALQR and FBLQR produce accurate estimates even when the distribution of the error is asymmetric. Most noticeably, when $\tau = 0.25$ and $\tau = 0.75$ the FBALQR and FBLQR are significantly more efficient than the other methods. In addition, we can observe that the worst estimators for all of the τ values are QR.

We have illustrated the practical performance of the methods which are discussed in this chapter by using the BF data which is described in subsection 3.8.2.

Table 5.5. MSE for $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$, which is estimated by FBALQR, FBLQR, FBQR, LQR and QR based on the BF data for $\tau = 0.25, \tau = 0.5$ and $\tau = 0.75$.

Method	τ		
	0.25	**0.50**	**0.75**
FBALQR	0.1084258	1.133324e-05	0.1370678
FBLQR	0.1241451	7.436673e-05	0.1423978
FBQR	0.1377442	9.326444e-05	0.1477935
LQR	0.1270168	0.001221321	0.1497241
QR	0.1414865	0.0009612835	0.1639030

Table 5.6. The estimated coefficients $\widehat{\boldsymbol{\beta}}$, which are estimated by FBALQR, FBLQR, FBQR, LQR and QR based on the BF data for $\tau = 0.25, \tau = 0.5$ and $\tau = 0.75$.

	τ														
	0.25					0.50					0.75				
	FBALQR	FBLQR	FBQR	LQR	QR	FBALQR	FBLQR	FBQR	LQR	QR	FBALQR	FBLQR	FBQR	LQR	QR
$\widehat{\beta}_0$	-0.329	-0.352	-0.371	-0.356	-0.376	-0.003	-0.009	-0.009	-0.035	-0.031	0.370	0.377	0.384	0.387	0.405
$\widehat{\beta}_1$	0.145	0.088	0.067	0.089	0.071	0.111	0.105	0.108	0.095	0.106	0.123	0.124	0.145	0.099	0.127
$\widehat{\beta}_2$	-0.001	-0.058	-0.047	-0.324	-0.018	-0.119	-0.099	-0.105	-0.001	-0.127	-0.097	-0.163	-0.120	-0.307	-0.726
$\widehat{\beta}_3$	-0.060	-0.028	-0.027	-0.007	-0.028	-0.030	-0.039	-0.037	-0.039	-0.042	-0.056	-0.039	-0.043	-0.010	-0.091
$\widehat{\beta}_4$	-0.129	-0.202	-0.225	-0.199	-0.245	-0.160	-0.155	-0.171	-0.110	-0.147	-0.082	-0.095	-0.094	-0.018	-0.062
$\widehat{\beta}_5$	0.022	0.040	0.024	0.128	0.017	-0.037	-0.020	-0.046	-0.089	-0.110	-0.076	-0.104	-0.122	-0.112	-0.034
$\widehat{\beta}_6$	1.010	1.085	1.149	1.169	1.162	1.163	1.122	1.167	1.190	1.181	1.089	1.175	1.168	1.180	1.353
$\widehat{\beta}_7$	-0.176	-0.237	-0.324	-0.236	-0.370	-0.211	-0.172	-0.235	-0.296	-0.348	-0.072	-0.101	-0.123	-0.039	-0.003
$\widehat{\beta}_8$	0.140	0.159	0.189	0.233	0.208	0.165	0.136	0.170	0.224	0.212	0.078	0.089	0.132	0.074	0.129
$\widehat{\beta}_9$	0.001	0.054	0.059	-0.008	0.072	-0.009	-0.004	-0.004	-0.043	-0.057	-0.028	-0.045	-0.065	-0.066	-0.129
$\widehat{\beta}_{10}$	-0.001	-0.076	-0.080	-0.025	-0.090	-0.003	-0.011	-0.004	-0.011	-0.011	0.054	0.077	0.085	0.110	0.122
$\widehat{\beta}_{11}$	0.018	0.018	0.024	-0.001	0.019	0.052	0.051	0.062	0.016	0.034	0.082	0.106	0.101	0.043	0.141
$\widehat{\beta}_{12}$	0.132	0.084	0.079	0.207	0.091	0.081	0.064	0.075	0.105	0.105	0.047	0.049	0.054	0.074	0.078
$\widehat{\beta}_{13}$	-0.244	-0.137	-0.116	-0.183	-0.118	-0.174	-0.169	-0.172	-0.193	-0.210	-0.213	-0.215	-0.225	-0.216	-0.204

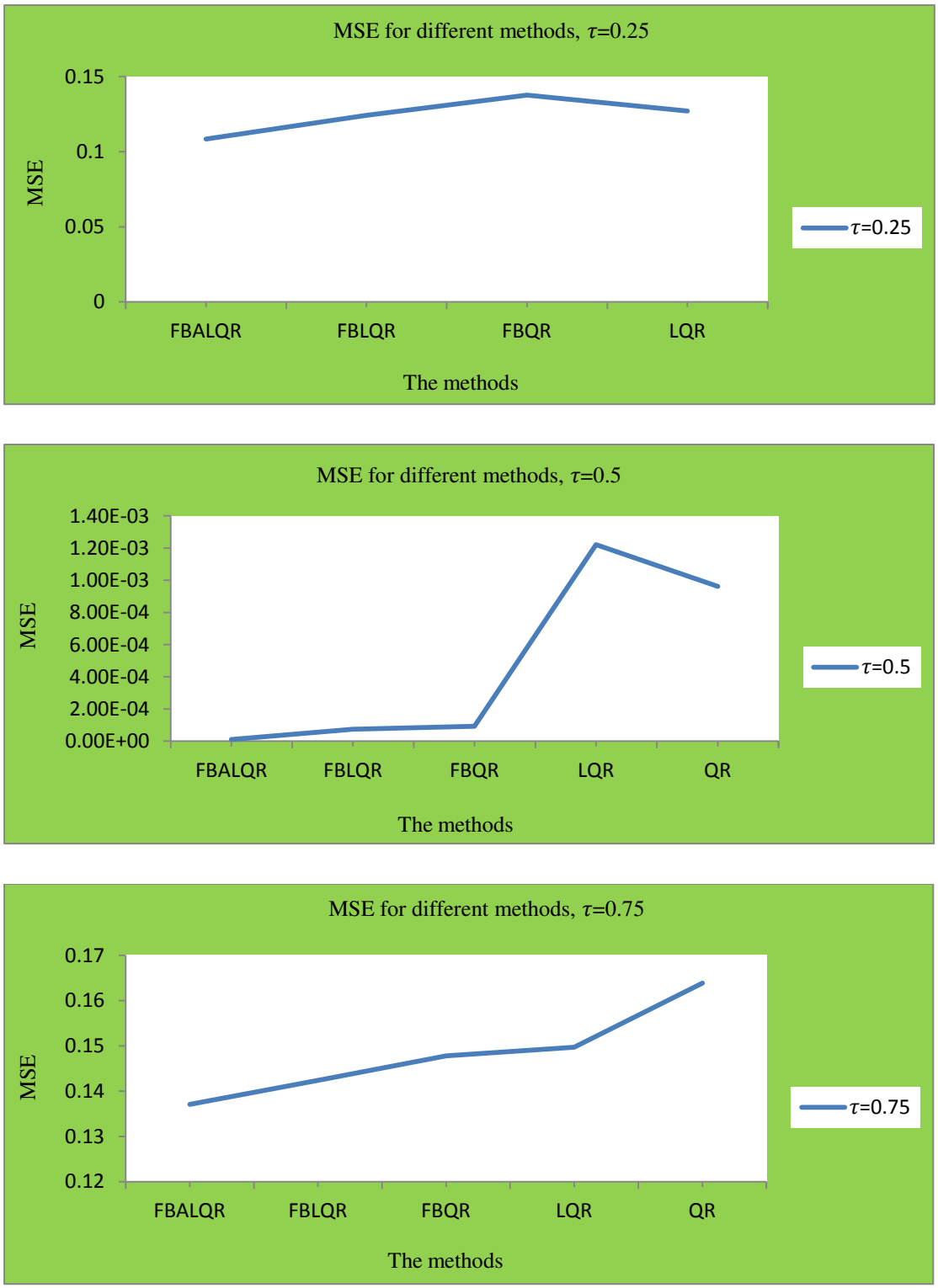


Figure 5.5. Plots explaining MSE for $\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}$, which is estimated by FBALQR, FBLQR, FBQR, LQR and QR based on the BF data for $\tau = 0.25, \tau = 0.5$ and $\tau = 0.75$.

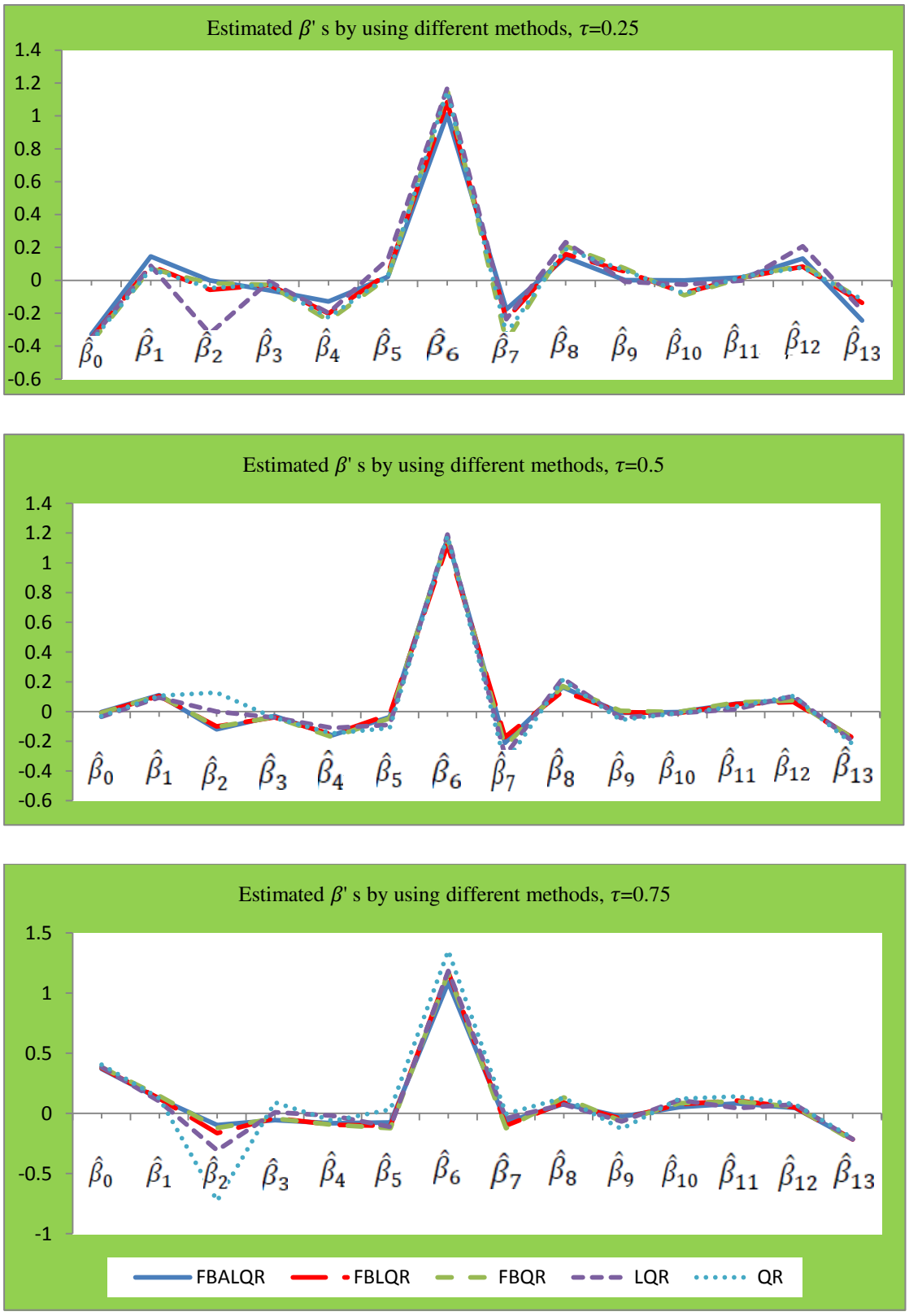


Figure 5.6. Plots explaining the estimated coefficients $\widehat{\boldsymbol{\beta}}$, which are estimated by FBALQR, FBLQR, FBQR, LQR and QR based on the BF data for $\tau = 0.25, \tau = 0.5$ and $\tau = 0.75$.

The results of the BF data analysis are reported in Tables 5.5–5.6 and Figures 5.5–5.6. From Table 5.5 and Figure 5.5, we have made the following observations. According to the MSE criterion, it can be seen that the performance of the FBALQR and FBLQR is better than the performance of the other methods. Also, it is clear that the FBALQR and FBLQR give accurate estimates. Again, we can see that when $\tau = 0.25$ and $\tau = 0.75$, the FBALQR and FBLQR are significantly more efficient than the other methods. The results of the simulation studies and the real data example suggest that the suggested methods perform well.

5.6. Chapter Summary

In this chapter, we have suggested the FBLQR and FBALQR by suggesting a hierarchical model framework. These methods have been compared with FBQR, LQR and the standard frequentist QR methods. In order to assess the numerical performance, simulation studies have been carried based on the model $y_i = \mathbf{x}_i^T \beta + \mathbf{x}_i^T \boldsymbol{\gamma}\, \varepsilon_i$, as described in Section 5.5. From the simulation studies and body fat data, we can conclude that the FBALQR and FBLQR perform well in comparison with the other methods and thus we believe that the proposed methods are practically useful.

References

Andrews, D. F. and Mallows, C. L. (1974). Scale Mixtures of Normal Distributions. Journal of the Royal Statistical Society, Series B 36, 99–102.

Alhamzawi, R., Yu, K. and Benoit, D. (2012). Bayesian adaptive LASSO quantile regression. Statistical Modelling 12, 279–297.

Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. Journal of Royal Statistics Society, Series B 73, 325–349.

Benoit, D. F. and Van den Poel, D. (2011). Binary quantile regression: A Bayesian approach based on the asymmetric Laplace distribution. Journal of Applied Econometrics 26, n/a. doi: 10.1002/jae.1216.

Fan, J. and Li, R. Z. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. Biostatistics 8, 140–154.

Griffin, J. and Brown, P. (2007). Bayesian adaptive lassos with non-convex penalization. Tech. Rep., University of Warwick.

Hans, C. (2009). Bayesian lasso regression. Biometrika 96, 835–45.

Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of Polya trees. Journal of the American Statistical Association 97, 1020–1033.

He, X. (1997). Quantile curves without crossing. The American Statistician 51, 186–191.

Hjort, N. (2003). Topics in non-parametric Bayesian statistics. In Highly structured Stochastic Systems, edited by Green, Hjort and Richardson.

Hjort, N. and Petrone, S. (2007). Nonparametric quantile inference using Dirichlet processes. In Advances in Statistical Modeling and Inference: Essays in Honor of Kjell A. Doksum, Edited by V Nair.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. Statistical Science 14, 382–417.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. Journal of Statistics Education 4, 236–237.

Koenker, R. and Bassett, G. (1978). Regression quantiles. Econometrica 46, 33–50.

Koenker, R. and Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association 94, 1296–1310.

Koenker, R. (2004). Quantile regression for longitudinal data. Journal of Multivariate Analysis 91, 74–89.

Koenker, R. (2005). Quantile Regression, Cambridge, U.K.: Cambridge University Press.

Kottas, A. and Gelfand, A. (2001). Bayesian semi-parametric median regression modeling. Journal of the American Statistical Association 96, 1458–1468.

Kottas, A. and Krnjaji´c, M. (2009). Bayesian nonparametric modelling in quantile regression. Scandinavian Journal of Statistics 36, 297–319.

Kozumi, H. and Kobayashi, G. (2009). Gibbs sampling methods for Bayesian quantile regression, Tech. Rep. 0058, Graduate School of Business Administration, Kobe University, 2009. Available at (http://ci.nii.ac.jp/naid/120001096632).

Leng, C., Tran, M. N. and Nott, D. (2010). Bayesian Adaptive Lasso. Technical report, Available at (arxiv.org/pdf/1009.2300).

Li, Q., Xi, R. and Lin, N. (2010). Bayesian Regularized Quantile Regression. Bayesian Analysis 5, 1–24.

Li, Y. and Zhu, J. (2008). l1-norm quantile regressions. Journal of Computational and Graphical Statistics 17, 163–185.

Park, T. and Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association 103, 681–686.

Reed, C. and Yu, K. (2009). An efficient Gibbs sampler for Bayesian quantile regression. Technical report, Department of Mathematical Sciences, Brunel University.

Reich, B., Bondell, H. and Wang H. (2010). Flexible Bayesian quantile regression for independent and clustered data. Biostatistics 2, 337–352.

Sun, W., Ibrahim, J.G. and Zou, F. (2010). Genome-wide multiple loci mapping in experimental crosses by the iterative adaptive penalized regression. Genetics 185, 349–59.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Tsionas, E. (2003). Bayesian quantile inference. Journal of statistical computation and simulation 73, 659–674.

Taddy, M. and Kottas, A. (2007). A Nonparametric Model-based Approach to Inference for Quantile Regression. Technical report ams2007-21, UCSC Department of Applied Math and Statistics.

Walker, S. and Mallick, B. (1999). A Bayesian Semi-parametric Accelerated Failure Time Model. Biometrics 55, 477–483.

Wang, H., Li, G. and Jiang, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection through the LAD. Journal of Business and Economic Statistics 25, 347–355.

Wu, Y. and Liu, Y. (2009). Variable Selection in Quantile Regression. Statistica Sinica 19, 801–817.

Yu, K. and Moyeed R. A. (2001). Bayesian quantile regression. Statistics and Probability Letters 54, 437–447.

Yu, K., Lu, Z. and Stander, J. (2003). Quantile regression: Applications and current research areas. The Statistician 52, 331–350.

Yu, K. and Stander, J. (2007). Bayesian analysis of a Tobit quantile regression model. Journal of Econometrics 137, 260–276.

Yuan, Y. and Yin, G. (2010). Bayesian quantile regression for longitudinal studies with non-ignorable missing data. Biometrics 66, 105–114.

Zou, H. (2006). The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

Appendix

The details of the Gibbs sampler for the FBLQR method are given as follows:

1- The full conditional distribution (FCD) of β_k is a $\mathrm{N}(\bar{\beta}_k, \hat{\sigma}_k^2)$, where

$$\hat{\sigma}_k^2 = \left(\sum_{i=1}^{n} x_{ik}^2 \left(\mathbf{x}_i^T \boldsymbol{\gamma}\sigma_{H_iG_i}\right)^{-2} + s_k^{-1}\right)^{-1},$$

and

$$\bar{\beta}_k = \hat{\sigma}_k^2 \sum_{i=1}^{n} x_{ik} \left(\mathbf{x}_i^T \boldsymbol{\gamma}\sigma_{H_iG_i}\right)^{-2} \left(y_i - \sum_{j=1, j\neq k}^{p} x_{ij}\beta_j - \mathbf{x}_i^T \boldsymbol{\gamma}\mu_{H_iG_i}\right).$$

2- The FCD of s_k is inverse Gaussian $IG(\mu', \lambda')$, $k = 1, \ldots, p$, where $\mu' = \sqrt{\lambda^2/\beta_k^2}$ and $\lambda' = \lambda^2$.

3- The FCD of λ^2 is $Gamma\left(p + c_3,\ \sum_{k=1}^{p} {}^{s_k}/_{2} + c_4\right)$

Given $N = max\{G_1, \ldots, G_n\}$ and $\mathfrak{m} = 1, \ldots, N$, the parameters $\mu_{1\mathfrak{m}}, \mu_{2\mathfrak{m}}, \text{Ч}_{\mathfrak{m}}, \sigma_{1\mathfrak{m}}, \sigma_{2\mathfrak{m}}$ and the standard deviation parameters can be updated using a Gaussian distribution. The group indicators G_i are also updated using Metropolis-Hasting sampling (see Reich et al. (2010) for more details).

The FCD for all parameters in the FBALQR method is similar to the above description, except for the FCD for s_k and $\lambda_k^2, k = 1, \ldots, p$ which are given by

1- The FCD of s_k is $IG(\mu', \lambda')$, $k = 1, \ldots, p$, where $\mu' = \sqrt{\lambda_k^2/\beta_k^2}$ and $\lambda' = \lambda_k^2$.

2- The FCD of λ_k^2 is $Gamma\left(1 + c_3,\ {}^{s_k}/_{2} + c_4\right)$.

Chapter 6

Conclusions and Future Research

The work in this thesis focuses on some statistical methods relating to variable selection, feature extraction and a combination of the two. The major contributions of the thesis and possible future research are summarised as follows.

6.1. Main Contributions

In Chapter 2, the main contributions are from proposing a number of robust canonical correlation (RCCA) methods. In the correlation matrix of the CCA, we suggest an approach that replaces the Pearson correlation with the percentage bend correlation and the winsorized correlation in order to obtain robust correlation matrices. The resulting correlation matrices have been employed to produce the RCCA methods. Moreover, the FCH, RFCH and RMVN estimators are employed to estimate the covariance matrix in the CCA. After that, these estimators are compared with the existing estimators. Researches on robust estimators such as the FCH, RFCH and RMVN, which are backed by theory, are needed to oppose large amount of material

available in the literature on zero breakdown estimators, such as Fast-MCD and Fast-MVE estimators that are not backed by theory, which were used instead of the MCD and MVE estimators.

In Chapter 3, we extended the Sparse MAVE (SMAVE) (Wang and Yin, 2008) by combining the MAVE method with the variable selection methods SCAD, adaptive Lasso and the MCP. The proposed methods have merits over the SMAVE and SSIR method (Li, 2007) because the proposed methods use penalisation, which benefits from oracle properties, while SMAVE and SSIR use Lasso, which does not. Also, the proposed methods have advantages over SSIR in that these methods do not need any certain distribution on $\mathbf{x}$ and are able to estimate the dimensions in the conditional mean function.

Extensions of the SIQ model of Wu et al. (2010) via considering Lasso and adaptive Lasso are proposed in Chapter 4. In addition, the practical algorithms have been suggested in order to calculate the penalised SIQ estimates.

In Chapter 5, a flexible Bayesian framework for regularisation in quantile regression models is developed. The error distribution is assumed to be an infinite mixture of Gaussian densities. This work is different from Bayesian lasso quantile regression, employing the ALD for the error. In fact, the use of the ALD is undesirable because of the deficiency of coherence (Kottas and Krnjaji´c, 2009).

6.2. Recommendations for Future Research

The topic of Chapter 2 offers the possibility of using robust multivariate location and dispersion RFCH and RMVN to estimate the covariance matrix in the classical multivariate procedures, such as discriminant analysis, factor analysis, principal components and sliced inverse regression in order to obtain robust estimators because the classical multivariate procedures are sensitive to the outliers.

The work is presented in Chapter 3 motivates us to recommend a number of interesting future work recommendations. Two of these are:

1. To study the MAVE with group variable selection-(group Lasso, group MCP and group Bridge).
2. It is possible to extend the MAVE method to the MAVE-QR method. The MAVE-QR method will inherit the same advantages as the MAVE method. Also, we are planning to study the sparse MAVE- QR method with Lasso, adaptive Lasso and other regularisation penalties.

Although the QR has become very popular as a comprehensive extension of classical mean regression, it nonetheless sometimes suffers from two problems. The first problem is the crossing of regression functions estimated for different orders of quantiles. The second problem is its practical success suffers from the "curse of dimensionality" (CD) in HD data. As we have pointed out in Chapter 4, the SIQ method reported by Wu et al. (2010) solved the CD problem, but still suffers from the crossing of regression functions at different orders of quantiles. It is possible to develop the SIQ method from Wu et al. (2010) to deal with this problem.

From Chapter 5, we can recommend the following future work:

1. The proposed methods can be extended to binary and left censored response variables.
2. To study other penalties, like the fused Lasso (Tibshirani et al., 2005), group Lasso (Yaun and Lin, 2006) and Elastic Net (Zou and Hastie, 2005).

References

Kottas, A. and Krnjaji´c, M. (2009). Bayesian nonparametric modelling in quantile regression. Scandinavian Journal of Statistics 36, 297–319.

Li, L. (2007). Sparse sufficient dimension reduction. Biometrika 94, 603–613.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society, Series B 67, 91–108.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high-dimensional data: sparse MAVE. Computational Statistics and Data Analysis 52, 4512–4520.

Wu, T. Z., Yu, K. and Yu, Y. (2010). Single-index quantile regression. Journal of Multivariate Analysis 101, 1607–1621.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68, 49–67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, Series B 67, 301–320.