國　立　交　通　大　學
運輸與物流管理學系

博　士　論　文

VIF 式矩陣擾動法解決線性模式的共線性問題

VIF-based Iterative Matrix Perturbation Method for Linear Models Having Collinearity Problems

研 究 生：黃 建 嘉

指導教授：卓 訓 榮 教授

周 幼 珍 副教授

中　華　民　國　一　〇　四　年　七　月

VIF 式矩陣擾動法解決線性模式的共線性問題

VIF-based Iterative Matrix Perturbation Method for Linear Models Having Collinearity Problems

研 究 生：黃建嘉　　Student：Chien-Chia Huang

指導教授：卓訓榮　　Advisor：Hsun-Jung Cho

周幼珍　　Yow-Jen Jou

國 立 交 通 大 學

運 輸 與 物 流 管 理 學 系

博 士 論 文

A Dissertation

Submitted to Department of Transportation and Logistics Management

College of Management

National Chiao Tung University

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Traffic and Transportation

July 2015

Hsinchu, Taiwan, Republic of China

中華民國一零四年七月

Ackonwledgements

The completion of this dissertation is primarily based upon the continuous support, inspiration and encouragements from many people along the road of my PhD career. The entire spell has been like a roller-coaster, sometimes up and sometimes down. However, only when I make it through can I really enjoy the sweetiness of the process.

First and foremost, I want to thank my entire family who have been there for me for the past years, especially when I went through the hardest spell of my life in the hospital. All the pain and sorrow has now paid off and I believe the PhD degree would be a great present in return.

Second, I want to thank my ex-girlfriend, Yi-Shan Li. She was the one who saved my life when I was very close to death in dorm. Her persistency in sending me to the EMU of a hospital truly became timely and the key to my survival from the diesease. The gratitude is always there for her.

Third, I would like to thank the medical team, inclusive of my two cousins and all the doctors, who have provided their help during my time in ICU in hospital. Weren't it for their help, I would not have been able to walk out the hospital healthy and completed this dissertation.

Fourth, I want to devote my appreciation to my advisors, Dr. Hsun-Jung Cho and Dr. Yow-Jen Jou. Ever since I entered National Chiao Tung University, I have been lucky enough to meet them. I always remember that I was recommended to Dr. Cho's lab by Dr. Jou in the summer of 2003. After that, my academic career totally changed. I was taught to know how to choose the right references for further knowledge in science, especially in mathematics and statistics. Weren't it for this direction, I would have been a head-less fly in an avent garden.

Fifth, I would like to thank Dr. Shu-Cherng Fang, Dr. Yunan Liu, Dr. Salah El-maghraby and all the fellows in the department of Industrial and Systems Engineering (ISE), of North Carolina State university. It was my fortune that I had the opportunity to embark my second PhD career in NCSU. Dr. Fang has been like a mentor to me. During the time in US, what influences me the most is not the acedmic strength but how to be a

person of integrity. That's what I have learned from him. Dr. Liu is the most intelligent person I have ever met in my life. His attitude towards everything has inspired me much. Albeit it was a pity that I failed to be productive with him in academic area, I truly have learned very much from him.

Last but not least, I would also like to thank Dr. Michael Fuchs from the department of applied mathematics, NCTU; Dr. Ming-Jong Yao from my department; my seniors, Guey-Shii Lin, Yu-Kuan Chen, Heng Huang, Chia-Chun Hsu in DSL lab; and those who I have been aquainted with during my time in NCTU. I am grateful that they are there with me during my PhD career.

Abstract

VIF-based Iterative Matrix Perturbation Method for Linear Models having Collinearity Problem

Chien-Chia Liäm Huang

In this thesis, we develop a novel matrix perturbation method for solving the data collinearity problem in linear models. The proposed matrix perturbation method inherits the merit from matrix theory and is shown effective both theoretically and practically. On the one hand, from the theoretical perspective, we provide and prove several important results concerning the quantities pertinent to the method, for instance, the eigenvalue perturbation theorem, boundedness results and so like. On the other hand, from the practical perspective, we carry out real-world applications to validate the effectiveness of the proposed method.

This thesis comprises four pieces of work, each presented in one chapter, from Chapters 2 to 6. In Chapter 2, we develop the base model for the entire thesis, a VIF-based (variance-inflation-factor-based) optimization model for solving the data collinearity problem in linear regression, with the full set of the linear model assumptions fulfilled. Employing the first-order Taylor approximation, we construct a linearly-constrained quadratic programming (LCQP) model. The optimization model turns out to be strictly convex and thence can be solved in polynomial time. We further propose an iterative algorithm for obtaining a dense but biased estimate less influenced by data collinearity. Since the total bias for the new estimate is bounded, we term the estimate the *accumulatively-least-bias* estimate. Real-world applications validate that the variances of the estimates can be improved using the proposed method.

In Chapter 3, for purpose of gaining more improvements in variance, we develop a series of optimization models, based upon the second-order Taylor approximation. Of all developed models, we pay special attention to the one with VIF approximated up to the second order and with the objective unchanged. The resultant model is a (possibly

nonconvex) quadratically constrained quadratic programming (QCQP) problem. QCQPs are a special branch of classic nonlinear programming problems and are of both theoretical and practical interests. Since the underlying QCQP can be nonconvex, we treat the convex and nonconvex cases respectively. For the convex case, the traditional Karush-Kuhn-Tucker (K-K-T) optimality condition guarantees the global optimality of the solution. As to the nonconvex case, to achieve the global optimality, we resort to the linear conic programming (LCOP) technique. Existing results provide us with a solid theoretical support. Also, we provide the generic structure of a solution algorithm for obtaining a new estimate.

In Chapter 4, we relax the homoskedasticity assumption in the linear model assumptions and establish the corresponding results. In general, there are two approaches to address the heteroskedasticity models. The first approach is to stick to the OLSEs and then use the heteroskedasticity-robust covariance estimates for the subsequent hypothesis testings. The other approach goes to the generalized least squares (GLS). However, the use of GLS requires the knowledge of the function representative of the heteroskedasticity, which is in general not known explicity and which requires further estimation. Therefore, in this particular chapter, we shall resort to the first approach and leave the second approach to the later chapter. The main result we prove resides in that the optimal perturbation matrix derived in Chapter 2 can be used to improve the heteroskedasticity-robust covariance estimates. Based upon this result, we establish the revised hypothesis testing procedures, such as perturbed t- or perturbed F-test using the improved covariance estimates. We illustrate the method with a real world application.

In Chapter 5, we proceed on to relax the normality assumption and assume that the error distributions now follow the exponential family, which leads to the notion of generalized linear models (GLMs). The main results are that we propose a new collinearity diagnostic for generalized linear models and that we propose a new estimate for generalized linear models having collinearity problems. The new diagnostic is termed the weighted variance inflation factor (WVIF), which behaves exactly the same as the traditional VIF in regression diagnostic, both indicating how inflated the variances are due to the presence

of data collinearity. The resulting estimate is termed the perturbed maximum likelihood estimate (PMLE). Although the PMLE is no longer unbiased, but it comes with an improved variance. We analyze a real world cancer dataset to show the effectiveness of our method.

In Chapter 6, we extend the results established for the GLMs in Chapter 5 to a unified variance reduction method for theory of generalized least squares (GLS). GLS is involved in various applications, such as heteroskedastic regression models, time-series models having autoregression correlated errors, fixed and random effect models in panel data, the estimation of maximum likelihood estimates in GLMs, the measurement-error regression models, and so on. Particularly, we show that the proposed diagnostic, WVIF, can be used for detecting data collinearity for any regression models involving GLS estimations. Moreover, as a by-product, we propose a new class of estimates, called the *perturbed generalized least squares estimate* (PGLS), for such classes of models. Importants properties and illustrations using real-world applications are given.

Contents

1 Introduction **1**

1.1 Preliminary 3

1.1.1 Basic Notations 3

1.1.2 The Linear Model 4

1.1.3 Assumptions 5

1.2 Organization of the Dissertation 6

2 VIF-based Iterative Matrix Perturbation Method for Linear Regression having Collinearity Problems **8**

2.1 Introduction 9

2.2 Model Derivation 13

2.2.1 Preliminary 13

2.2.2 Objective Function 15

2.2.3 VIF Constraints 16

2.2.4 A Convex Quadratic Programming Model 17

2.3 Statistical Issues 18

2.3.1 Approximate Confidence Bounds 18

2.3.2 Range of Correlation Coefficients 19

2.3.3 Sign Restrictions 19

2.3.4 Goodness of Fit 20

2.3.5 Refined VIF-based LCQP Model 21

2.3.6 On Positive Definiteness of Perturbed Correlation Matrix 21

2.4 Proposed Algorithm, Parameter Settings and Numerical Issues 21
2.4.1 Numerical Issues and Algorithm Inputs 22
2.4.2 A VIF-based Algorithm 23
2.5 Hypothesis Testing 24
2.6 Numerical Examples - Real Data 24
2.6.1 The Diabetes Dataset 25
2.6.2 An application to Collinear Data 26
2.6.3 The Sales Dataset 30
2.6.4 United States Pork Data 31
2.7 Concluding Remarks 35

3 Second-Order VIF-based Optimization Model for Linear Regression having Collinearity Problems 37
3.1 Introduction 37
3.2 The Model 39
3.2.1 Objective Function 40
3.2.2 Feasible Region 41
3.2.3 A Nonlinear Optimization Model and Its Variants 45
3.3 Solution Strategies 46
3.3.1 Convex Case - Simple Case 46
3.3.2 Nonconvex Case - Conic Reformulation 47
3.3.3 Solution Algorithm 48
3.4 Concluding Remarks 50

4 VIF-based Iterative Matrix Perturbation Method for Heteroskedasticity-Robust Covariance Estimators in the Presence of Collinearity 51
4.1 Introduction 51
4.2 Some Properties 53
4.3 Main Results - Perturbed HCEs 57
4.4 Perturbed Test Statistics and Confidence Intervals 60

4.5 An Example 61
4.6 Conclusions 63

5 A New Multicollinearity Diagnostic for Generalized Linear Models 64

5.1 Introduction 64
5.1.1 Our Constributions 67
5.1.2 Organizations 68
5.2 An Example - Cancer Data 68
5.3 Weighted VIF - A New Diagonstic 70
5.3.1 An Alternative Estimate 71
5.3.2 Performance of PMLEs 72
5.3.3 Goodness of Fit 74
5.4 Cancer Data Revisited 75
5.5 Conclusions 78

6 Variance Reduction Method for Generalized Least Squares in the Presence of Multi-collinearity 79

6.1 Introduction 79
6.2 Perturbed Generalized Least Squares 82
6.3 Some Properties 85
6.4 Iterative Algorithm and Examples 86
6.4.1 An Iterative Algorithm 86
6.4.2 Example - Longley's Data 87
6.4.3 Example - Cancer Data 88
6.5 Conclusions 90

7 Conclusions and Future Research Directions 91

7.1 Future Research Directions 92
7.1.1 Optimal Rank-2 Perturbation 92
7.1.2 Nonconvex L^p Regression for $p \in (0, 1)$ 93

Appendices **94**

A Schinnar's CN-Based Model - Review **95**

A.1 Formulation using Condition Number 95

B Application to Weighted Least Squares - A Special Case of Generalized Least Squares **97**

B.1 Properties 97

B.2 Proofs of Theorems 101

List of Figures

2.1 Figures (a) - (j) represent the relationship between the condition number and the *VIF*s for each explanatory variable in the diabetes dataset. (Solid Line – Left Axis; Dashed Line – Right Axis) 27

2.2 Figure (a) shows the relationship between the condition number and the perturbation norm; Figure (b) concerns th positive definiteness property of the perturbed correlation matrix for diabetes data. (Solid Line – Left Axis; Dashed Line – Right Axis) 28

2.3 The relationships between the condition number and *VIF*s for (a) FAM, (b) PEER and (c) SCHOOL for education data. (Solid Line – Left Axis; Dashed Line – Right Axis) 29

2.4 Figure (a) shows the relationship between condition number and the perturbation norm. Figure (b) concerns the positive definiteness of Ω^*_{XX} for education data. (Solid Line – Left Axis; Dashed Line – Right Axis) 30

2.5 The relationship between the condition number and *VIF*s for (a) A_t (b) P_t (c) E_t (d) A_{t-1} and (e) P_{t-1} (From Top to Bottom) (Solid Line – Left Axis; Dashed Line – Right Axis) 32

2.6 The figure shows (i) the relationship between the condition number and the perturbation norm; and (ii) the positive definiteness of Ω^*_{XX} for sales data. (Solid Line – Left Axis; Dashed Line – Right Axis) 33

2.7 The relationship between the condition number and the *VIF*s for (a) *lp* (b) *lq* (c) *ldw* (d) *ldpe* (e) *lcr*4 (f) *d*98 and (g) *t* (From Top to Bottom) (Solid Line – Left Axis; Dashed Line – Right Axis) 35

2.8 Figure (a) shows the relationship between the condition number and the perturbation norm; and Figure (b) concerns the positive definiteness of Ω^*_{XX}. (Solid Line – Left Axis; Dashed Line – Right Axis) 36

4.1 Plots for pairs of indexes; A vs. B with A corresponding to left y-axis (blue, solid line) and B to right y-axis (green, dashed line); x-axis denotes the iteration counts and only those integral values count. 63

5.1 Increasing trend in CN using the ridge estimation in logistic regression . . 69

5.2 Plots for WVIFs resulting from the algorithm 76

5.3 Changes in $\lambda_{\min}(X\widehat{W}_{\infty}X)$ resulting from the algorithm 77

6.1 Plots of the WVIFs 88

6.2 Plots of CN (blue line to left y-axis) versus $\lambda_{\min}$ and $\lambda_{\max}$ (dashed lines to right y-axis) of the matrix $X^T\widehat{W}X$ 89

6.3 Changes in $\lambda_{\min}(X\widehat{W}_{\infty}X)$ resulting from the algorithm 90

7.1 The comparison between l_2-norm (blue line) and $l_{0.5}$-norm (green line) . . . 93

List of Tables

2.1 The summary of the original information for the Diabete data, derived from the OLS. Note: *s.e.* denotes standard error 25

2.2 The summary of the original information for the Diabete data, derived from the proposed algorithm 26

2.3 The summary of the original information for the education data, derived from the ordinary least squares. 27

2.4 The correlation matrix for variables FAM, PEER and SCHOOL. 28

2.5 The summary of the original information for the education data, derived from the proposed algorithm 29

2.6 The summary of the original information for the advertising data, derived from the ordinary least squares 31

2.7 The summary of the new information for the advertising data, derived from the proposed algorithm 31

2.8 The correlation matrix for explanatory variables for US Pork Data 32

2.9 The summary of the original information for (detrended) US pork data, derived from the ordinary least squares 33

2.10 The summary of the original information for (trended) US pork data, derived from the ordinary least squares 34

2.11 The summary of the original information for (detrended) US pork data, derived from the proposed algorithm 34

4.1 Summary for Bayhan's data; V stands for VIF and HC for heteroskedasticity-robust variance. 61

4.2 Estimators and their 1-S.D. interval estimation for Bayhan's data 62

5.1 Summary of OLSE, RMLE and MLE for the cancer data 68

5.2 Comparison of VIFs for the education data Chatterjee and Hadi (2006); The ridge shrinkage parameters $k_1 = 10^{-5}$ and $k_2 = p - 1/\beta_{\text{OLS}}^T \beta_{\text{OLS}} = 0.249$; data are centered and scaled 69

5.3 Results for the centered and scaled cancer data; Note: (i) for OLSE, we have $V_0 = I$ and $\widehat{\pi}_i = 0$; (ii) parameters for obtaining the PMLE are $v_l = 4$, $v_r = 2$ and $\rho = 50$, see Chapter 2; (iii) total additional iterations are 3 and it takes 0.353 CPU seconds (average under 100 trials) for the algorithm to terminate 76

5.4 Summary of the estimates and the associated χ^2-statistics and the deviances 77

5.5 Minimal eigenvalue of $X^T \widehat{W}_\infty X$ at each estimate 77

6.1 Summary of PGLSE for centered and scaled Longley's data; $WVIF$'s are the diagonals of the matrix $(\widehat{X}^T \widehat{X})^{-1}$;$\lambda$'s are the eigenvalues and CN is the condition number of the matrix $\widehat{X}^T \widehat{X}$. 88

6.2 Results for the centered and scaled cancer data; Note: (i) for OLSE, we have $V_0 = I$ and $\widehat{\pi}_i = 0$; (ii) parameters for obtaining the PMLE are $v_l = 4$, $v_r = 2$ and $\rho = 50$, see Chapter 2; (iii) total additional iterations are 3 and it takes 0.353 CPU seconds (average under 100 trials) for the algorithm to terminate 89

6.3 Minimal eigenvalue of $X^T \widehat{W} X$ at each estimate 90

Chapter 1

Introduction

Regression modeling has been one of the most useful and important tools for statistical research. Of all the clans in regression modeling, *linear regression* refers to building a linear relationship between a set of *explanatory* variables and a *response* variable, enabling the researchers to examine how those explantory interact with each other and affect the responses. For instance, in econometrics, by having all other factors fixed, one can perform the *ceteris peribus* analysis to examine how a unit change in a given factor affects the outcome by simply looking at the coefficient associated. Others such as biology, physics, environmental science and the likes are the related applicable fields. Thence, its wide applicability renders regression modeling an important tool for scientific studies.

Linear regression is favorable in the sense that the estimation can be done by using the ordinary least squares (OLS). The theorey of OLS is so well-founded that it provides a systematic way for researchers to obtain an estimate almost automatically. In the meanwhile, supported by a rich body of statistical inference, OLS has become the core of the regression analysis.

Whereas theoretically explanatory variables are assumed independent implicitly in most, if not all, linear models, the analysis proceeds with data given exogenously in practice. Usually, data are collected from a vast unknown population, encompassing certain degree of stochasticity and unpredictability, and thus problems result. For instance, an erroneous input of a data point may likely yield a completely different result for the estimates; missing data points due to certain reasons can be an obstacle for the subsequently

analysis; or high similarity (dependency) in the collected data may lead to unsatisfactory results and so like. Problems of this sort are referred to as the *data problems*.

Of all data problems in linear regression, we are more concerned about the nearly linear dependency among several factors assumed independent implicitly in the model specification. Numerically, high dependency results directly in rank-deficiency problem in the OLS estimation. The resultant problems, such as high variance, low statistical significance, and even incorrect signs, can frustrate the researchers aiming to model a problem using the regression.

Biased regression has been one of the approaches devised in the literature for solving data collinearity problems in linear regression. The most dominant method is this particular category is the ridge regression. Ridge regression has won its fame by its simplicity in implementation in practice and by producing good results. Yet, ridge regression has certain problems. First of all, by definition, the collinearity stems from the dependency between at least two different covariates in the data matrix. Instead of tackling the problem from the root, the ridge method tackles the problem by breaking the intrinsic structure of the correlation matrix, under the assumption that the data matrix is normalized. Second, the breaking the intrinsic structure leads to plausibly good results. In particular, from the persepective of VIF, the results are even infeasible, rendering also the hypothesis testings infeasible, which in turn renders the future inferences infeasible. Third, the ridge method requires a one-dimensional parameter. To find a good parameter for the ridge method, one potentially has to deal with hard optimization problem which is NP-hard in nature. Thence, to even find a local optimal, the computational effort will be quite unfriendly, let alone looking for the global oiptimality. Moreover, a simply local optimal cannot guarantee the performance of the resulting ridge estimate.

Motivated by such frustrations, we aim to develope a alternative and useful method, based on the reliable diagnostic tool, to solve the commonly seen problem in the context of linear regression. The developed method is useful becasue our method inherits the merit from matrix theory and mitigates the data collinearity problem by improving not only the values of the underlying diagnostic but also the very intrinsic eigenstructure of the

correlation matrix. The close connection between the diagnostic used and the correlation matrix leads to the success of our method in dealing with the problem.

We carry out both the real-world applications (when they are accessible) and random instance experiments (when they are inaccessible) to validate the developed method. The success in real-world applications reveals that our method is capable of addressing realistic datasets troubled by the existence of data collinearity. Moreover, from the results of the random instance experiments, we learn that our method has more capability in dealing with arbitrarily generated (with some variabilities well controlled) datasets.

1.1 Preliminary

Next, we shall first introduce the basic notations to be used subsequently in the thesis, followed by the specification of the linear model that serves as the fundamental of the entire analysis. Assumptions are given thereafter.

1.1.1 Basic Notations

All vectors are in boldface and are real column vectors, unless otherwise specified. An m-by-1 vector is denoted by $\mathbf{x} \equiv [x_i]_{i=1}^m \in \mathbb{R}^m$, for x_i the i-th element thereof and for $m \in \mathbb{N}$, the set of natural numbers.

Matrices are real and in capitals. A real m-by-n matrix A, with $a_{i,j}$ as its (i,j)-th element, for $i \in \{1, ..., m\}$ and $j \in \{1, ..., n\}$, is written as $A \equiv [a_{i,j}] \in \mathbb{R}^{m \times n}$. Occasionally, we use $(A)_{i,j} \equiv a_{i,j}$ to indicate the (i,j)-th element of a matrix A. We let $A \equiv [a_{i,j}]_{i \neq j} \in \mathbb{R}^{m \times m}$ be a squared matrix with all diagonals being zero and with off-diagonals $a_{i,j}$'s, $\forall i \neq j$. A diagonal matrix $A \in \mathbb{R}^{m \times m}$ is denoted by $Diag[a_i]$ with a_i the i-th diagonal element and with all off-diagonals zero. We denote by $A \succ (\succeq) 0$ a positive (semi)definite matrix A. The transpose of a matrix $A \in \mathbb{R}^{m \times n}$ is denoted by $A^T \in \mathbb{R}^{n \times m}$. The inverse of a matrix $A \in \mathbb{R}^{m \times m}$, if it exists, is denoted by $A^{-1} \in \mathbb{R}^{m \times m}$; otherwise, we use the Moore-Penrose generalized inverse of a matrix A which is denoted by $A^\dagger$. The determinant of a matrix $A \in \mathbb{R}^{m \times m}$ is denoted by $\det A$. The trace of a matrix $A \in \mathbb{R}^{m \times m}$ is denoted by

$\mathbf{tr}A \equiv \sum_{i=1}^{m} a_{i,i}$, the sum of the diagonals. The (k,l)-th submatrix derived from deleting the k-th row and l-th column of a matrix A is denoted by $A\{k,l\}$. In particulat, when $k = l$, we write the submatrix as $A\{k\}$.

The eigen-system of a matrix $A \in \mathbb{R}^{m\times m}$ is denoted by the pair $(\Lambda(A), \mathbf{V}(A))$, where $\Lambda(A) = [\lambda_i]_{i=1}^{m}$ is the vector of eigenvalues and $\mathbf{V}(A) \equiv (\upsilon_1, ..., \upsilon_m)$ is a matrix composed of the eigenvectors $\upsilon_k \in \mathbb{R}^m$, for $k = 1, ..., m$. We write $\lambda_i(A)$ as the i-th smallest eigenvalue of a matrix A of proper dimension. As a convention, the smallest and largest eigenvalues are denoted by $\lambda_{\min}(= \lambda_1)$ and $\lambda_{\max}(= \lambda_m)$, respectively. The condition number (CN) of a matrix $A \in \mathbb{R}^{m\times m}$ is defined by the ratio $\kappa(A) \equiv |\frac{\lambda_{\min}}{\lambda_{\max}}|(A)$.

We adopt the conventional vector/matrix l_2-norms, meaning that the basic axioms for a norm are satisfied. For an m-by-1 vector $\mathbf{x}$, the l_2-norm is defined by $\|\mathbf{x}\|_2^2 \equiv \sum_{i=1}^{m} x_i^2$. As for a matrix $A \in \mathbb{R}^{m\times m}$, its norm is given by $\|A\|_2^2 \equiv \sum_{i,j=1}^{m} a_{i,j}^2$. Moreover, for any matrix $Q \succ 0$, the induced norm for a vector $\mathbf{x}$ is denoted by $\|\mathbf{x}\|_Q^2 \equiv \mathbf{x}^T Q\mathbf{x}$.

A univariate random variable $Y \in \mathbf{R}$ that is normally distributed with mean $\mathbb{E}[Y]$ and variance $Var(Y)$ is denoted by $Y \sim \mathcal{N}(\mathbb{E}[Y], Var(Y))$. If Y is a standard normal random variable, then it is denoted by $Y \sim \mathcal{N}(0,1)$. For a multi-variate case, we replace the singletons $\mathbb{E}[Y]$ and $Var(Y)$ by a vector $\mathbb{E}[\mathbf{Y}]$ and the covariance matrix σ. Thence, we write $\mathbf{Y} \sim \mathcal{N}(\mathbb{E}[\mathbf{Y}], \sigma)$, for $\mathbf{Y} \in \mathbb{R}^m$, $m \in \mathbb{N}$.

1.1.2 The Linear Model

We will consider a general *centered and scaled* linear model, *viz.*,

$$\mathbf{y} = X\beta + \mathbf{u}. \tag{1.1}$$

In particular, $X \in \mathbb{R}^{n\times k}$ is the centered and scaled data matrix, where n is the number of observations and k is the number of independent variables considered. The matrix product $X^T X\,(\equiv \Omega_{XX}) \in \mathbb{R}^{k\times k}$ is the *correlation matrix* for the independent variables accordingly (Belsley, 1984). The correlation coefficient between variable i and variable j is denoted by $\varpi_{i,j}$, *viz.*, $\Omega_{XX} \equiv [\varpi_{i,j}]$ with $\varpi_{i,j} = 1$, if $i = j$, and $-1 \le \varpi_{i,j} \le 1$, if

$i \neq j$, for $1 \leq i, j \leq k$. Let $\mathbf{y} \in \mathbb{R}^n$ be the response (or the dependent) vector, $\beta \in \mathbb{R}^k$ the vector of regression coefficients and $\mathbf{u} \in \mathbb{R}^n$ be the error vector. When estimated, the *fitted* model is given by

$$\widehat{\mathbf{y}} = X\widehat{\beta} + \widehat{\mathbf{u}}, \tag{1.2}$$

wherein $\widehat{\beta}$ is the estimated regressor vector, $\widehat{\mathbf{u}}$ is the residual vector and $\widehat{\mathbf{y}}$ is the predicted vector.

1.1.3 Assumptions

Whereas, in this dissertation, there is no specific datatype requirement, the two major datatypes in question are *cross-sectional* and *time-series*. The basic assumptions for both datatypes are the traditional linear model assumptions, *viz.*, the Gauss-Markov assumptions. However, the detailed descriptions of the assumptions for each data type differ.

For cross-sectional data, the full set of G-M assumptions state: (i) linear in parameters (β's); (ii) random sampling; (iii) no perfect collinearity; (iv) zero conditional mean $\mathbb{E}(\mathbf{u}|X) = 0$; and (v) homoskedasticity, $Var(u_i|X) = Var(u_i) = \sigma^2$.

For time-series data, we will enforce the full set of the G-M assumptions that state: (i) linear in parameters (β's); (ii) no perfect collinearity; (iii) zero conditional mean $\mathbb{E}(u_t|X) = 0, t = 1, \ldots, n$; (iv) homoskedasticity, $Var(u_t|X) = Var(u_t) = \sigma^2, t = 1, \ldots, n$; and (v) no serial correlations $Corr(u_t, u_s|X) = 0, \forall t \neq s$.

Besides the aformentioned sets of assumptions for the two datatypes, we also require the so-called *normality* assumption on the random error, *viz.*,$(vi-1)$ $u_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n$, for cross-sectional data; and $(vi-2)$ $u_t \sim \mathcal{N}(0, \sigma^2), t = 1, ..., N$ for time-series data.

Lastly, we will assume that the data matrix is exogenous in essence, *viz.*, X is given and fixed.

1.2 Organization of the Dissertation

This is how the dissertation is organized. In Chapter 2, we show how to build the VIF-based optimization model for linear regression having collinearity problem, given the full set of linear model assumptions fulfilled. Based upon the first-order approximation, the resulting optimization model is a convex linearly-constrained quadratic programming (LCQP) problem. Built upon the LCQP, we propose an iterative algorithm for solving linear regression having collinearity problem. We carry out real-world applications to validate the effectiveness of the proposed method.

In Chapter 3, we construct various second-order optimization models. We pay special attention to one variant that is a (possibly non-convex) qudratically-constrained quadratic programming (QCQP) problem. We discuss both the convex and nonconvex cases. For the first case, we can directly obtain the global optimality using the existing results. For the second case, we employ the linear conic programming (LCOP) technique to reformulate the problem. Well-established results ensure us to attain the global optimality.

In Chapter 4, we relax the homoskedasticity assumption and stick to the OLS approach. We prove that the optimal perturbation derived from solving the LCQP in Chapter 2 can be used to reduce the variance portions of the heteroskedasticity-robust covariance estimate (HCE), thereby increasing the statistical significance. Variants of the HCEs can also be obtained. Besides, we provide perturbed versions of the test statistics and the pertinent interval estimations.

In Chapter 5, we relax the normality assumption for the error distributions and assume they all follow the exponential family. Such a class of models is called the generalized linear models (GLMs) and we focus on the collinearity diagnostic problem for this particular class. By far, condition number is still the major diagnostic tool for detection of collinearity in the literature. We then propose a new diagnostic tool, which is called the weighted VIF (WVIF), for detecting the presence of data collinearity. A by-product of the new diagnostic is class of alternative estimates for generalized linear models, which is called the perturbed maximum likelihood estimates (PLMEs), with variances improved.

In Chapter 6, we extend the results in Chapter]5 to a unified framework for the

family of models involving the generalized least squares (GLS) estimations. We extend the perturbation method to the theory of GLS and apply the new diagnostic developed in Chapter 5 to detecting the presence of data collinearity in such a class of models. A new class of alternative estimates, called the perturbed generalized least squares estimates (PGLSEs), then can be formed using the diagnostic. Conclusions and future research directions are given in 7.

Chapter 2

VIF-based Iterative Matrix Perturbation Method for Linear Regression having Collinearity Problems

In this chapter, we address data collinearity problem in multiple linear regression from an optimization perspective. We propose a novel linearly constrained quadratic programming model, based on the concept of the variance inflation factor (VIF). We employ the matrix perturbation method that involves imposing a general symmetric non-diagonal perturbation matrix on the correlation matrix. The proposed VIF-based model reduces the largest VIF by minimizing the resulting biases. The VIF-based model can mitigate the harm from data collinearity through the reduction in both the condition number and VIFs, meanwhile improving the statistical significance. The resulting estimate has bounded biases under an iterative framework and hence is termed the *least accumulative bias* (LABs) estimate. Certain potential statistical properties can be further considered as the side constraints for the proposed model. Various numerical examples validate the proposed method.

2.1 Introduction

Motivation. Collinearity problem describes the situation that the non-orthogonality exists among explanatory variables in regression models. The breakdown in the orthogonality among explanatory variables causes imprecisions in the use of normal equation in the ordinary least squares (OLS) estimations. The imprecision oftentimes leads to high variances, and thus low statistical significance, and even the incorrect signs of the OLS estimates (OLSEs) (Zhen and Wohlgenant, 2010). This renders the OLSEs, even though they still exist, inappropriate in statistical senses. Moreover, the power in prediction of the established regression model is far weakened (Belsley, 1984). Hence, collinearity deserves more attention because its presence has frustrated the researchers endeavoring to establish important relationship among interesting variables using regression models (Shacham and Brauner, 1997; Næs and Mevik, 2001).

Belsley et al. (2005) ascribed collinearity problem to a *data* problem, instead of a statistical one. This is because the specification of the models has implied the independence among the explanatory variables. Inevitably, yet, we work with non-experimental data, pooling from an unknown vast population. Uncertainties and unpredictabilities in the behaviour of the collected samples oftentimes lead to unsatisfactory results. Much effort has been sowed in designing sophisticated experiments as well as new techniques to, at least, mitigate the harm from data collinearity, for example the ridge regression (Hoerl and Kennard, 1970), the LASSO (Tibshirani, 1996) and the bridge regression (Frank and Friedman, 1993).

Spanos and McGuirk (2002) discussed the near-multicollinearity problems by classifying the problems into (i) a structual issue (*systematic volatility*) and (ii) a numerical issue (*erratic volatility*). Systematic volatility refers to high correlations among the explanatory variables, concerning the structure of the correlation matrix that potentially invokes the presence of the data collinearity. The numerical issues concern the ill-conditionedness of the data matrix.

Ridge regression has won its reputation in addressing data collinearity problems successfully by imposing perturbations on the diagonals of a correlation matrix. Ridge re-

gression takes effect by sacrificing the intrinsic structure of the correlation matrix, *viz.*, the correlation between a variable and itself is 1 by definition. The perturbation, no matter how small, on the diagonals of the correlation matrix violates this definition essentially.

We aim at the proposal of a novel approach that incorporates both the systematic and erratic volatilities, as aformentioned. From the systematic perspective, the proposed approach mitigates the effects from data collinearity by preserving the intrinsic structure of the correlation matrix, in which aspect the ridge regression does not. From the erratic perspective, the proposed approach exploits the numerical sensitivity of a matrix suffering from ill-conditionedness, inheriting the essence from matrix theory (Horn and Johnson, 2012). Moreover, our approach retains the intrinsic property of a correlation matrix. Therefore, our approach enjoys both theoretical and numerical merits.

Literature Review. By far, there is no concensus for the detection of the presence of data collinearity. The presence of data collinearity is often detected through the concepts of the condition number (CN), the variance inflation factor (VIF) and so forth (Fox and Monette, 1992; Curto and Pinto, 2007; Kovács et al., 2005). Another criterion is that if there is more than one non-diagonal elements of the correlation matrix having value(s) very close to ± 1, the collinearity can be said to be present. All those concepts are built upon the correlation matrix for the explanatory variables in regression models. Each of these existing diagnostic tools has its own advantages and weaknesses. For a thorough discussion, please refer to Belsley et al. (2005).

The concept of CN relates to the eigenstructure of the correlation matrix. The CN is defined through the ratio of the largest to the smallest eigenvalues of the correlation matrix. A high CN value, usually 30 in the literature (Belsley et al., 2005), suggests the presence of data collinearity. Nevertheless, there is still no rule of thumb for the CN to reveal the presence of data collinearity. One of the drawbacks of CN is that CN tends to be inflated, misleading the researchers to believing the presence of data collinearity (Lazaridis, 2007). The use of CN as the main diagnostic tool should be with caution.

Another widely used diagnostic tool is called the VIF that indicates how the variance of the corresponding coefficient is inflated due to data collinearity (Curto and Pinto, 2011;

Robinson and Schumacker, 2009). Naturally, a high *VIF* value for an explanatory variable suggests the presence of data collinearity. Although there is no rule of thumb for *VIF*s, a value of 10 is often adopted, but with caution (Obrien, 2007). Built upon the concept of *VIF*, Fox and Monette (1992) introduced a generalized diagnostic for the collinearity problems. Curto and Pinto (2011) proposed the corrected *VIF*s that incorporate the R^2 values as the adjustments for the original *VIF*s. Lin et al. (2011) proposed a regression method using the concept of *VIF* as the criterion for variable selection. Liao and Valliant (2012) examined the role of *VIF*s in complex survey data. Increases in the use of *VIF* as the major tool for different purposes reveals the increase in its importance in the literature. Based upon this fact, we adopt *VIF* as the major tool as a natural choice.

There have been a great many of techniques developed as the remedy for the consequent symptoms resulting from data collinearity. To name a few, the ridge regression (Hoerl and Kennard, 1970) and McDonald (2009)), the LASSO, the bridge regression, the principal component regression (Batah et al., 2009), the partial least-squares regression (Abdi, 2010), better estimates, say Liu-type estimates (Liu, 2003)) or other estimates (Bagheri and Midi, 2009), the variable deletion/selection approach (Xin and Zhu, 2012), data adjustments (Echambadi and Hess, 2007; Shieh, 2011) and so forth (Bashtian et al., 2011; Fierro and Bunch, 1992; Lin, 2008) have been widely used and discussed. Others are referred to Soofi (1990); Leung and Yu (2000) and Næs and Mevik (2001).

In light of the above from the literature, we propose a novel approach, built upon the concept of *VIF*, as a new remedy for data collinearity problems in multiple linear regression. More specifically, we develop a novel optimization model based upon the concept of *VIF* to tackle data collinearity problems. Similar to ridge regression but unlike those intended for variable selection, the proposed *VIF*-based model possesses the feature that all the variables in the regression model are kept, which can be of practical interest (McDonald and Schwing, 1973) and (Schwing and McDonald, 1976).

Our Contributions. We address the data collinearity problems in multiple linear regression. Our contributions are summarized as follows.

First of all, we propose a novel *VIF*-based optimization model to overcome the data

collinearity problems. To the best of our knowledge, there is no such model to date in the literature. The established linearly constrained quadratic programming (LCQP) model is convex, so it is computationally efficient. Moreover, on the one hand, our approach resembles LASSO (Tibshirani, 1996) and bridge regression (Frank and Friedman, 1993)) in the sense that all involve solving associated optimization models to obtain estimates. On the other hand, our approach resembles ridge regression in the sense that the established estimate can be representative of a revised normal equation, to be discussed in the next paragraph.

Second, compared to the ridge regression, our approach tackles the collinearity problems in a more reasonable way. Ridge estimates are derived from imposing a diagonal perturbation matrix, λI, for I an identity matrix of appropriate dimension, on the correlation matrix, *viz.*, $\widehat{\beta}_R = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$, given the centered and scaled data matrix X and $\mathbf{y}$. This can be counterintuitive, for the correlation between a variable and itself is by definition 1, no matter how small the perturbation λ is. One may even find out that the ridge regression can still produce a solution by setting $\lambda = 1$. Our approach preserves the intrinsic of the correlation matrix by imposing a perturbation matrix $\mathcal{W}$ on the correlation matrix, *viz.*, $\widehat{\beta}_{\mathcal{W}} = (X^T X + \mathcal{W})^{-1} X^T \mathbf{y}$, for $\mathcal{W}$ symmetric and having zeros on the diagonal. Such an imposition does not change the values on the diagonal of the original correlation matrix.

Third, our approach provides another tool for regression modelling that evades variable selection when data collinearity problem is present. This can be helpful when certain important variables might be ruled out by the variable selection according to certain criteria, as pointed out in McDonald and Schwing (1973) and Schwing and McDonald (1976). Therefore, our approach accompanies the ridge regression in this aspect. Both the LASSO and the bridge regression are intended for variable selections, so we will not make comparisons therewith.

Lastly, various examples validate our approach. The numerical results indicate that (i) our approach can not only improve the *VIF*s, but the *CN* as well; (ii) statistical significance can be potentially improved; and (iii) estimates with different signs can be

corrected.

2.2 Model Derivation

In this section, we aim at the establishment of the LCQP for addressing data collinearity problems. Before that, we first introduce certain notations that will be used subsequently in the model derivation. We then specify the construction of the model.

The key idea of the proposed approach resides in the superposition of a symmetric non-diagonal perturbation matrix on the correlation matrix. This essentially originates from the matrix theory. When a matrix suffers from ill-posedness, *viz.*, the condition number thereof is very high, a slight perturbation in the data can result in a relatively large change in the eigenstructure of the matrix (Horn and Johnson, 2012). Our model inherits the merit of the theory, to show that such perturbations can be obtained through solving a convex LCQP.

We start with the centered and scaled regression model, generally without intercept. In light of Belsley et al. (2005), both the normal equation and the *VIF*s can be considered as a function of the correlation matrix. We employ the first-order Taylor approximation on both the normal equation and the *VIF*s to construct the objective function and the constraints, respectively, for the resultant LCQP. The LCQP has a convex quadratic objective function and a set of linear constraints. The convexity guarantees the computational efficiency in solving the LCQP, as will be shown later in Section 2.6.

2.2.1 Preliminary

We start with the *centered and scaled* regression model

$$\mathbf{y} = X\beta + \mathbf{u}. \tag{2.1}$$

An OLSE can be secured by the normal equation, as a function of the correlation matrix, in particular

$$\widehat{\beta}\left(\Omega_{XX}\right) = \left(X^T X\right)^{-1} X^T \mathbf{y} \equiv \Omega_{XX}^{-1} \Omega_{X\mathbf{y}}, \tag{2.2}$$

where $\Omega_{X\mathbf{y}} \equiv X^T\mathbf{y}$. The population correlation matrix is by definition symmetric and positive definite. The sample correlation matrix, however, may not be so, depending on the behaviour of the collected non-experimental data, for example Knol and ten Berge (1989). In this study, we assume that the sample correlation matrix is positive definite, *viz.*, $\Omega_{XX} \succ 0$. Mathematically, we require the smallest eigenvalue of the correlation matrix to be positive.

In light of Belsley et al. (2005), the *VIF*s of the independent variables can be derived from the diagonal elements of the inverse of the correlation matrix, *viz.*, the diagonals of Ω_{XX}^{-1}. We denote by $\mathcal{V}_i(\Omega_{XX})$ the *VIF* for the i-th independent variable, in particular

$$\mathcal{V}_i(\Omega_{XX}) \equiv \left(\Omega_{XX}^{-1}\right)_{i,i}, \text{ for } i \in \{1, ..., k\}. \tag{2.3}$$

Our aim is to find a symmetric non-diagonal perturbation matrix $\mathcal{W} \equiv [\Delta\varpi_{i,j}]_{i\neq j}$ such that the resulting correlation matrix remains positive definite, *viz.*,

$$\Omega_{XX}^* = \Omega_{XX} + \mathcal{W} \succ 0. \tag{2.4}$$

Note that $(\mathcal{W})_{i,j} \equiv \Delta\varpi_{i,j} = 0$, for $i = j$. Quantities imposed with (2.4) are said to be *perturbed*. In our case, the OLSEs (2.2) imposed with (2.4) become the perturbed OLSEs (POLSEs), or simply perturbed estimates, in particular

$$\widehat{\beta}_{\mathcal{W}}(\Omega_{XX}^*) \equiv (\Omega_{XX}^*)^{-1}\Omega_{X\mathbf{y}}. \tag{2.5}$$

The *VIF*s (2.3) imposed with (2.4) become the perturbed *VIF*s, in particular

$$\mathcal{V}_i(\Omega_{XX}^*) \equiv \left(\Omega_{XX}^{*-1}\right)_{i,i}, \text{ for } i \in \{1, ..., k\}. \tag{2.6}$$

We shall employ the perturbation method by applying the first-order Taylor approximation to the perturbed versions of (2.2) and (2.3), *viz.*, (2.5) and (2.6), to establish the objective function and the constraints, respectively, for the LCQP.

2.2.2 Objective Function

We start with the perturbed normal equation (2.5). Applying the first-order Taylor approximation about Ω_{XX} to $\widehat{\beta}\left(\Omega_{XX}^{*}\right)$ suggests

$$\widehat{\beta}_{\mathcal{W}}\left(\Omega_{XX}^{*}\right)=\widehat{\beta}\left(\Omega_{XX}\right)+\sum\nolimits_{i>j}\widehat{\beta}_{i,j}^{\prime}\left(\Omega_{XX}\right)\triangle\varpi_{i,j}=\widehat{\beta}\left(\Omega_{XX}\right)+B\varpi, \tag{2.7}$$

where $\varpi\in\mathbb{R}^{k(k-1)/2}$ is the *decision vector* consisting of $\triangle\varpi_{i,j}$, $\forall i>j$, and the columns of $B\in\mathbb{R}^{n\times(k(k-1)/2)}$ are made of $\widehat{\beta}_{i,j}^{\prime}\left(\Omega_{XX}\right)$ for each particular pair (i,j), $\forall i\neq j$. We remind the readers of that the perturbation matrix $\mathcal{W}\equiv\left[\triangle\varpi_{i,j}\right]_{i\neq j}$ will be composed of the elements in the decision vector ϖ, by rearranging the elements properly. The matrix B can be derived from taking partial derivative of (2.2) with respect to the nondiagonals. In particular,

$$\widehat{\beta}_{i,j}^{\prime}\left(\Omega_{XX}\right)=\frac{\partial\widehat{\beta}_{i,j}\left(\Omega_{XX}\right)}{\partial\varpi_{i,j}}=-\Omega_{XX}^{-1}\left(\mathbf{e}_{i}\widehat{\beta}_{j}+\mathbf{e}_{j}\widehat{\beta}_{i}\right),\forall i\neq j$$

wherein $\mathbf{e}_i$ is a column zero vector with 1 at the i-th position.

Since we are imposing perturbations on the OLSEs, the resulting estimates become biased. We define the bias by the difference between the perturbed and the original estimates, *viz.*, $b_{\mathcal{W}}\equiv\widehat{\beta}_{\mathcal{W}}\left(\Omega_{XX}^{*}\right)-\widehat{\beta}\left(\Omega_{XX}\right)=B\varpi$. Therefore, our objective here is to minimize the bias incurred from such perturbations. In particular, we aim to minimize the bias in the regression estimate as well as the length of the perturbation, *viz.*,

$$\min_{\varpi\in\mathbb{R}^{k(k-1)/2}}\|B\varpi\|_{2}^{2}+\rho\|\varpi\|_{2}^{2}=\varpi^{T}(B^{T}B+\rho I)\varpi, \tag{2.8}$$

which is convex given any $\rho>0$.

2.2.3 VIF Constraints

We proceed on with the perturbed *VIF*s in (2.6). Applying the first-order Taylor approximation about Ω_{XX} to each $\mathcal{V}_l\left(\Omega_{XX}^*\right)$ suggests

$$\mathcal{V}_l\left(\Omega_{XX}^*\right) = \mathcal{V}_l\left(\Omega_{XX}\right) + \sum_{i>j} \mathcal{V}_l'\left(\Omega_{XX}\right)\Delta\varpi_{i,j}, \text{ for } l = 1, ..., k. \tag{2.9}$$

The keys to obtain the derivative $\mathcal{V}_l'\left(\Omega_{XX}\right)$ are twofold. First, from matrix theory (Horn and Johnson 1990), we know that

$$\mathcal{V}_l\left(\Omega_{XX}\right) = \det\Omega_{XX}\{l\}/\det\Omega_{XX}. \tag{2.10}$$

Second, the determinant of a matrix is the product of its eigenvalues, *viz.*,

$$\det\Omega_{XX} = \prod_{p=1}^{n}\lambda_p\left(\Omega_{XX}\right) \text{ and } \det\Omega_{XX}\{l\} = \prod_{\hat{p}=1}^{n-1}\lambda_{\hat{p}}\left(\Omega_{XX}\{l\}\right), \tag{2.11}$$

where λ_p and $\lambda_{\hat{p}}$ are the p-th and $\hat{p}$-th eigenvalues for the matrices Ω_{XX} and $\Omega_{XX}\{l\}$, respectively.

Based upon (2.10) and (2.11), it can be shown that

$$\mathcal{V}_l'\left(\Omega_{XX}\right) = 2\left[\sum_{\hat{q}=1}^{n-1}\frac{\widehat{\upsilon}_{\hat{q},\hat{\imath}}\widehat{\upsilon}_{\hat{q},\hat{\jmath}}}{\lambda_{\hat{q}}\left(\Omega_{XX}\{l\}\right)} - \sum_{q=1}^{n}\frac{\upsilon_{q,i}\upsilon_{q,j}}{\lambda_q\left(\Omega_{XX}\right)}\right]\mathcal{V}_l\left(\Omega_{XX}\right), \tag{2.12}$$

which is a scalar. Note that υ_q and $\widehat{\upsilon}_{\hat{q}}$ are the q-th and $\hat{q}$-th eigenvectors associated with Ω_{XX} and $\Omega_{XX}\{l\}$, respectively; and $\upsilon_{q,i}$ and $\widehat{\upsilon}_{\hat{q},\hat{\imath}}$ are the i-th and $\hat{\imath}$-th elements of the eigenvectors υ_q and $\widehat{\upsilon}_{\hat{q}}$, respectively. We can thereby write (2.9) as

$$\mathcal{V}_l\left(\Omega_{XX}^*\right) = \mathcal{V}_l\left(\Omega_{XX}\right) + \mathbf{v}_l^T\varpi, \text{ for } l = 1, \ldots, k, \tag{2.13}$$

for $i \neq j$ and $\hat{\imath} \neq \hat{\jmath}$ in (2.12). Note that the vector $\mathbf{v}_l$ in (2.13) is composed of (2.12) for different pairs (i, j).

From (2.13) we define the difference between the perturbed and the original *VIF*s by

$$d_l(\varpi) \equiv \mathcal{V}_l(\Omega^*_{XX}) - \mathcal{V}_l(\Omega_{XX}) = \mathbf{v}_l^T \varpi, \text{ for } l = 1, \ldots, k.$$

Our aim is to decrease the *VIF*s by imposing a perturbation on the correlation matrix, *viz.*, we want $d_l(\varpi) < 0$. To this end, according to the authors' experience, it suffices to consider only the independent variable with the largest *VIF* value, in particular, we consider

$$\mathcal{V}_{k^*}(\Omega^*_{XX}) = \mathcal{V}_{k^*}(\Omega_{XX}) + \mathbf{v}_{k^*}^T \varpi$$

where $k^* = \max_{1 \le j \le k} \{j | \mathcal{V}_j(\Omega_{XX}) \ge \mathcal{V}_i(\Omega_{XX}), \text{ for } i \ne j\}$. We denote such a constraint as

$$\mathcal{V}_{\max}(\Omega^*_{XX}) = \mathcal{V}_{\max}(\Omega_{XX}) + \mathbf{v}_{\max}^T \varpi \text{ so that } d_{\max}(\varpi) < 0 \tag{2.14}$$

Note that (2.14) constitutes a linear constraint, and thus convex. We are now ready to state the LCQP.

2.2.4 A Convex Quadratic Programming Model

Combining (2.8) and (2.14), we form the following norm-minimization model

$$\min_{\varpi \in \mathbb{R}^{k(k-1)/2}} \left\{ \varpi^T (B^T B + \rho I) \varpi : -\mathbf{v}_{\max}^T \varpi = v_r \right\} \tag{2.15}$$

where $\rho(> 0)$ is an exogenous trade-off parameter and $v_r(> 0)$ is the reduction in the *VIF*. An intuitive interpretation of (2.15) is the following. We aim to find a perturbation matrix (symmetric and zeros diagonals) so that the largest *VIF* is reduced, meanwhile minimizing the bias resulting from the perturbation.

The strict convexity of the objective function guarantees that the global minimizer to (2.15) can be found, in light of the K-K-T optimality condition (Bazaraa et al., 2013) which in our case is both necessary and sufficient. In particular, the optimal perturbation

vector ϖ^* is given by

$$\varpi^* = \frac{-v_r}{\|\mathbf{v}_{\max}\|^2_{(B^TB+\rho I)^{-1}}} \left(B^TB + \rho I\right)^{-1} \mathbf{v}_{\max}. \tag{2.16}$$

The trade-off parameter ρ plays two roles in the model. First, we wish the norm of perturbation $\varpi^T\varpi$ is small enough so that the model works well. If the norm were too large, the model could have poorly performed. The trade-off parameter ρ serves as the penalty imposed on the norm of $\varpi^T\varpi$. Second, the addition of the trade-off parameter ensures that the objective function is well-posed, *viz.*, $B^TB + Diag\left[\rho\right] \succ 0$, for $\rho > 0$. The choices of the parameters ρ and v_r will be discussed more in Section 2.4.

2.3 Statistical Issues

There are certain issues for the perturbation matrix derived from (2.15) as well as the goodness of fit for the perturbed model. We discuss the issues in what follows.

2.3.1 Approximate Confidence Bounds

Let $\Omega^*_{XX} \equiv \left[\varpi^*_{i,j}\right]$. We require that the perturbed correlation coefficients should lie within the confidence interval under the null hypothesis $H_0 : \left\{\varpi^*_{i,j} = \varpi_{i,j}\right\}$ for a specific level of confidence. Let the null hypothesis be $H_0 : \left\{\varpi^*_{i,j} = \varpi_{i,j}\right\}$ against the alternative $H_1 : \left\{\varpi^*_{i,j} \neq \varpi_{i,j}\right\}$. We adopt the Fisher's Z-transform on the correlation coefficients, given by

$$Z_{\varpi^*} = \frac{1}{2}\ln\left(\frac{1+\varpi^*_{i,j}}{1-\varpi^*_{i,j}}\right), \text{ for } i \neq j,$$

following approximately an normal distribution with mean Z_ϖ and variance $\sigma^2_Z = \frac{1}{n-3}$, *viz.*, $Z_{\varpi^*} \sim \mathcal{N}\left(Z_\varpi, (n-3)^{-1}\right)$. The Fisher's Z-statistic ($\mathcal{Z}$) is defined as

$$\mathcal{Z} \equiv \frac{(Z_{\varpi^*} - Z_\varpi)}{\sigma_Z} = \frac{\sqrt{n-3}}{2}\ln\left[\frac{\left(1+\varpi^*_{i,j}\right)\left(1-\varpi_{i,j}\right)}{\left(1-\varpi^*_{i,j}\right)\left(1+\varpi_{i,j}\right)}\right] \to \mathcal{N}\left(0,1\right).$$

Since both the standard deviation σ_Z and $\varpi_{i,j}$, $\forall i,j \in \{1,...,k\}$, are known, we can derive the confidence bounds for $\varpi^*_{i,j}$ accordingly. It is not hard to see that the following constraint satisfies the $100 \times (1-\alpha)\,\%$ confidence interval for $\varpi^*_{i,j}$

$$\frac{\frac{(1+\varpi_{i,j})}{(1-\varpi_{i,j})}\exp\left(\frac{2z_{\frac{\alpha}{2}}}{\sqrt{n-3}}\right)-1}{\frac{(1+\varpi_{i,j})}{(1-\varpi_{i,j})}\exp\left(\frac{2z_{\frac{\alpha}{2}}}{\sqrt{n-3}}\right)+1} \leq \varpi^*_{i,j} \leq \frac{\frac{(1+\varpi_{i,j})}{(1-\varpi_{i,j})}\exp\left(\frac{2z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right)-1}{\frac{(1+\varpi_{i,j})}{(1-\varpi_{i,j})}\exp\left(\frac{2z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right)+1}, \text{ for } i \neq j \tag{2.17}$$

Unfortunately, the confidence bounds (2.17) serves merely as an approximation for the perturbed correlation coefficient $\varpi^*_{i,j}$, for $i \neq j$. The use of the Fisher's Z-transform relies on the normality condition. The Fisher's Z-statistic can perform very poorly if the sample size is not large enough, say $n < 500$ (Paul, 1989). Moreover, there is so far no well-founded hypothesis testing procedure for testing the relationship between the perturbed and the original correlation coefficients, *viz.*, $H_0 : \{\varpi^*_{i,j} = \varpi_{i,j}\}$ against the alternative $H_1 : \{\varpi^*_{i,j} \neq \varpi_{i,j}\}$. We bring up (2.17) for purpose of completion.

2.3.2 Range of Correlation Coefficients

We require that the perturbed correlation matrix still retains its own characteristic, *viz.*, every correlation coefficient should be within the -1 to 1 range. Mathematically, we have

$$\varpi^*_{i,j} \in (-1,1)\,, \text{ for } i \neq j, \text{ and } \varpi^*_{i,j} = 1, \text{ for } i = j. \tag{2.18}$$

2.3.3 Sign Restrictions

The signs of the perturbation should satisfy some rules, if any. That is, for some specific (i,j)-th element, the corresponding perturbation is restricted in sign,

$$\triangle\varpi_{i,j} = 0 \text{ or } \pm\triangle\varpi_{i,j} > 0, \text{ for } (i,j) \in \mathcal{I} \tag{2.19}$$

where $\mathcal{I}$ is an index set subject to sign restrictions.

2.3.4 Goodness of Fit

The value of R^2 must be less than unity. The R^2 statistic is defined by $R^2 = 1 - C_1 (\widetilde{\mathbf{y}} - \widehat{\mathbf{y}})^T (\widetilde{\mathbf{y}} - \widehat{\mathbf{y}})$, where $\widetilde{\mathbf{y}} \equiv (\mathbf{y} - \overline{\mathbf{y}})$, $\widehat{\mathbf{y}} = X\widehat{\beta}$ and $C_1 \equiv \left(\widetilde{\mathbf{y}}^T \widetilde{\mathbf{y}}\right)^{-1}$. It is obvious that R^2 is a function of the regression estimate $\widehat{\beta}$, and therefore a function of the correlation matrix. In particular, letting $R^2 \equiv \mathcal{R}(\Omega_{XX})$, we see that

$$\mathcal{R}(\Omega_{XX}) = 1 - C_1 \left(\widetilde{\mathbf{y}} - X\widehat{\beta}(\Omega_{XX})\right)^T \left(\widetilde{\mathbf{y}} - X\widehat{\beta}(\Omega_{XX})\right).$$

Imposing a perturbation matrix $\mathcal{W}$ such that (2.4) holds, the first-order Taylor approximation of $\mathcal{R}(\Omega^*_{XX})$ about Ω_{XX} suggests

$$\mathcal{R}(\Omega^*_{XX}) = \mathcal{R}(\Omega_{XX}) + \sum_{i>j} \mathcal{R}'_{i,j}(\Omega_{XX}) \triangle \varpi_{i,j}, \tag{2.20}$$

in which $\mathcal{R}'(\Omega_{XX})$ is the partial derivative with respect to $\varpi_{i,j}$, $\forall i > j$. It is not hard to see that the $\mathcal{R}'_{i,j}(\Omega_{XX})$ in (2.20) is given by, for $i \neq j$,

$$\mathcal{R}'_{i,j}(\Omega_{XX}) = 2C_1 \left(\widehat{\beta}^T - \widetilde{\mathbf{y}}^T X \Omega^{-1}_{XX}\right) \left(\mathbf{e}_i \widehat{\beta}_j + \mathbf{e}_j \widehat{\beta}_i\right), \tag{2.21}$$

which is a scalar and $\mathbf{e}_i$ is a column zero vector with 1 at the i-th position. Expressed in the vector/matrix notation, (2.21) becomes $\mathcal{R}(\Omega^*_{XX}) = \mathcal{R}(\Omega_{XX}) + \mathbf{r}^T \varpi$. Similarly, we want the difference between the perturbed and original R-squareds to remain non-negative, *viz.*,

$$\mathcal{R}(\Omega^*_{XX}) - \mathcal{R}(\Omega_{XX}) = \mathbf{r}^T \varpi \geq 0. \tag{2.22}$$

Remarkably, the R^2 constraint (2.22) can be helpful when we lose grip on the R^2 for the perturbed model.

As a consequence, at the optimal perturbation ϖ^*, the changes in the goodness of fit is given by

$$\Delta\mathcal{R} = \frac{\mathbf{r}^T (B^T B + \rho I)^{-1} \mathbf{v}_{\max}}{\|\mathbf{v}_{\max}\|^2_{(B^T B + \rho I)^{-1}}} v_r = \mathcal{O}(v_r). \tag{2.23}$$

Numerical evidences suggest that the changes in the goodness of fit is negligible.

2.3.5 Refined VIF-based LCQP Model

The original LCQP (2.15) can be refined by incorporating the side constraints (2.17)-(2.19) and (2.22), *viz.*,

$$\min_{\varpi \in \mathbb{R}^{k(k-1)/2}} \left\{ \varpi^T \left(B^T B + \rho I \right) \varpi : - \mathbf{v}_{\max}^T \varpi = v_r, \text{ (2.17)-(2.19) and (2.22)} \right\}. \tag{2.24}$$

According to our experience, there is no need to impose additional complexity to the original LCQP model (2.15). Thence, the statistical constraints (2.17)-(2.19) and (2.22) will be treated as posterior test.

2.3.6 On Positive Definiteness of Perturbed Correlation Matrix

We now return to (2.4) left assertive. In general, $\mathcal{W}$ is indefinite in essence, so there is no guarantee that (2.4) is true. Computationally, however, if one can ensure that the minimal eigenvalue of Ω_{XX} is greater than or equal to the maximal eigenvalue of $\mathcal{W}$, (2.4) will hold. Specifically, assuming that Ω_{XX}^* is positive definite, then, for any $\mathbf{x} \neq 0$, we see that

$$\mathbf{x}^T \Omega_{XX}^* \mathbf{x} = \mathbf{x}^T \left(\Omega_{XX} + \mathcal{W} \right) \mathbf{x} > \left(\min_{1 \le i \le k} \left\{ \lambda_i \left(\Omega_{XX} \right) \right\} - \max_{1 \le i \le k} \left\{ \lambda_i \left(\mathcal{W} \right) \right\} \right) \mathbf{x}^T \mathbf{x} > 0, \tag{2.25}$$

which implies $\min_{1 \le i \le k} \left\{ \lambda_i \left(\Omega_{XX} \right) \right\} \ge \max_{1 \le i \le k} \left\{ \lambda_i \left(\mathcal{W} \right) \right\}$.

2.4 Proposed Algorithm, Parameter Settings and Numerical Issues

Before introducing the proposed algorithm, we bring up certain numerical issues that relate to the design of the proposed algorithm. The proposed algorithm requires certain exogeneous inputs to begin with. We shall describe the model input and the settings thereof. We state the general structure of the proposed algorithm thereafter.

2.4.1 Numerical Issues and Algorithm Inputs

Recall that the proposed LCQP (2.15) tackles the data collinearity problem by reducing the largest *VIF* while keeping the resultant biases well controlled. The hope to find a perturbation matrix that can reduce the largest *VIF* down to a specified level in a one-step fashion can be of greed and is risky rendering the LCQP (2.15) to collapse. Compared to the ridge regression, designing an iterative algorithm for our approach is of necessity, as the inevitable trade-off for finding a more general symmetric non-diagonal perturbation matrix. We shall show that the computational efficiency of the designed algorithm for the LCQP (2.15) does not frustrate us much, albeit iterative.

Parameters for the VIF-based model. For the LCQP (2.15), we need to specify the trade-off parameter ρ and the reduction in *VIF* v_r. The trade-off parameter ρ is by default set to 5. We will show that the choice of ρ does not affect much the performance of (2.15).

The choice of v_r is more important than that for ρ, because the choice of v_r relates directly to the performance (or quality) of the solution to the LCQP model. A moderate level for v_r, say 5, is oftentimes a good choice. Any value greater than 10 would not be recommended. We choose $v_r = 1$ by default. A more elaborate setting for v_r is the dynamic adjustment. Keeping v_r at a constant level may limit the performance of our approach, especially at few steps before termination. A recommended setting for v_r would be in decreasing order, *viz.*, $v_r^i > v_r^{i+1} > 0$ for i the iteration number. The only trade-off for the dynamic adjustment is, however, that the number of iterations increases. This can be offset by the computational efficiency in solving (2.15).

Parameters for the algorithm. Since the proposed algorithm is iterative in nature, we need to specify the initial values for the algorithm to begin with. The OLSEs $\widehat{\beta}_{OLS}$ for (2.1) are chosen as the initial estimates for the algorithm. The correlation matrix Ω_{XX} as well as the *VIF*s $\{\mathcal{V}_j(\Omega_{XX}),\ j = 1, ..., k\}$ are known from the data beforehand. Alternatively, it suffices to have the knowledge of the collected data, *viz.*, $\mathbf{y}$ and X.

We now specify the stopping criterion for our iterative algorithm. Recall that what

LCQP (2.15) does is to reduce the largest *VIF* while keeping the bias minimized. And since our aim is to mitigate the harm from the presence of data collinearity, it is reasonable to set an appropriately desired level for the largest *VIF* to go to. We denote the desired level by v_l. For the purpose of comparison, we shall let v_l be 3 by default, depending on the examples.

Algorithm Inputs. By and large, the algorithm inputs can be represented as a six-tuple vector $\mathcal{P}$, in particular

$$\mathcal{P} = (\mathbf{y}, X, \rho, v_r, v_l) \tag{2.26}$$

2.4.2 A VIF-based Algorithm

We now present our proposed algorithm. A general algorithmic structure for the algorithm is given below.

Algorithm 1 VIF-Based Iterative Algorithm

Require: Problem input $\mathcal{P}$ as in (2.26)
Ensure: Optimal POLSEs $\widehat{\beta}_{\mathcal{W}}$;
1: **repeat**
2: Generate matrix B for objective function and vector $\mathbf{v}_{k^*}$ for constraint;
3: Use (2.16) to form $\mathcal{W}^* = [\Delta\varpi_{i,j}]_{i>j}$;
4: Update OLSE;
5: **until** $\max_{j=1,\ldots,k} \mathcal{V}_i \leq v_l$

Regarding the **Step 4** of the proposed algorithm, the readers must be aware of the fact that the superposition of the perturbation matrix $\mathcal{W}^*$ has a universal effect on all *VIF*s (and, of course, the estimates). Recall that the constraint of the LCQP (2.15) is built upon the independent variable with the largest *VIF*, and the index thereof is determined by $k^* = \max\limits_{1\leq j\leq k} \{j|\mathcal{V}_j(\Omega_{XX}) \geq \mathcal{V}_i(\Omega_{XX}), i \neq j\}$. The perturbation matrix $\mathcal{W}^*$ is, however, imposed upon each and every non-diagonal element of the correlation matrix. There is no escape that all *VIF*s change due to the superposition. Fortunately, the changes in the *VIF*s are beneficial in the sense that only those pathological *VIF*s are reduced while those good *VIF*s fluctuates slightly around their original values. We shall prove this fact in later chapter. By and large, numerically, we effectively calculate all *VIF*s in **Step 4**

and the index in the constraint $-\mathbf{v}_{\max}^T \varpi = v_r$ (or $\mathbf{r}^T \varpi \geq 0$) varies over iterations.

2.5 Hypothesis Testing

In this section, we discuss the hypothesis testing on the perturbed estimates. As the objective function of (2.15) suggests, the superposition of the perturbation on the correlation matrix results in changes in the OLSEs. It is necessary to perform the standard t-test on the perturbed regression estimates. We also show that the the t-statistic is effectively a function of the *VIF*s.

We implement the standard t-test on the null hypothesis $H_0 : \left\{\widehat{\beta}_j\left(\Omega_{XX}^*\right) = 0\right\}$ against $H_1 : \left\{\widehat{\beta}_j\left(\Omega_{XX}^*\right) \neq 0\right\}$. The purpose is to ensure that the POLSEs still retain their statistical significance.

The t-statistic for the j-th estimate is defined by $t_j = \widehat{\beta}_j / se\left(\hat{\beta}_j\right)$, for $j = 1, ..., k$. The standard errors of the regression estimates are effectively functions of the *VIF*s, in particular,

$$se\left(\widehat{\beta}_j\right) = \sqrt{\sigma^2 \mathcal{V}_j\left(\Omega_{XX}^*\right)} \text{ with } \sigma^2 = \left\|\left(\mathbf{y} - \bar{\mathbf{y}} - X\widehat{\beta}\left(\Omega_{XX}^*\right)\right)\right\|_2^2,$$

for a centered and scaled model (2.1). The perturbed t-statistic therefore is defined as

$$t_j\left(\Omega_{XX}^*\right) \equiv \widehat{\beta}\left(\Omega_{XX}^*\right) / \sqrt{\sigma^2 \mathcal{V}_j\left(\Omega_{XX}^*\right)}. \tag{2.27}$$

From the relationship between $t_j\left(\Omega_{XX}^*\right)$ and $\mathcal{V}_j\left(\Omega_{XX}^*\right)$ in (2.27), it is obvious that the reductions in *VIF*s increase the values of the t-statistics.

2.6 Numerical Examples - Real Data

We apply the LCQP (2.15) and the proposed algorithm in Section 2.4.2 to four examples of two data types. The first two examples are of cross-section. One of them is the diabetes dataset discussed in Efron et al. (2004), and the other is a collinear dataset (Chatterjee and Hadi, 2006) in that the correlation coefficients among independent variables are all

greater than 0.9. The rest are time-series datasets. One of them is drawn from (Chatterjee and Hadi, 2006), and the other is from Zhen and Wohlgenant (2010). All experiments are implemented in MATLAB R2010b with Pentium 4, 3GHz CPU and 1G RAM.

2.6.1 The Diabetes Dataset

The diabetes data, as discussed in Efron et al. (2004), has 10 explanatory variables, inclusive of *age*, *sex*, *bmi*, *map*, *tc*, *ldl*, *hdl*, *tch*, *ltg* and *glu*; and there is one response variable. The context of the dataset is to construct a model to examine the relationship among those variables. The basic regression information is summarized in Table 2.1, consisting of the OLS estimates, *VIF* values and the standard error.

Vars	*age*	*sex*	*bmi*	*map*	*tc*	*CN*
$\widehat{\beta}_{OLS}$	−0.48	−11.42	24.76	15.45	−37.72	470.08
*VIF*s	**1.22**	**1.28**	**1.51**	**1.46**	**59.2**	–
s.e.	59	60.46	65.7	64.6	411.46	–
Vars	*ldl*	*hdl*	*tch*	*ltg*	*glu*	R^2
$\widehat{\beta}_{OLS}$	22.7	4.81	8.43	35.78	3.22	0.52
*VIF*s	**39.19**	**15.4**	**8.89**	**10.08**	**1.49**	–
s.e.	334.79	209.87	159.45	169.75	65.16	–

Table 2.1: The summary of the original information for the Diabete data, derived from the OLS. Note: *s.e.* denotes standard error

Applying the LCQP (2.15) using the algorithm to the model yields the perturbed information in Table 2.2. We see that the R^2 value for the perturbed model remains at the same level as that from the OLS results. The *CN* has dropped down enormously from 470 to 21.

Vars	*age*	*sex*	*bmi*	*map*	*tc*	*CN*
$\widehat{\beta}_{\mathcal{W}}$	-0.2	-11.31	25.56	15.99	-37.12	20.46
*VIF*s	**1.17**	**1.32**	**2.26**	**1.52**	**2.99**	Iters
s.e.	57.88	61.52	80.59	65.99	92.76	93
Vars	*ldl*	*hdl*	*tch*	*ltg*	*glu*	R^2
$\widehat{\beta}_{\mathcal{W}}$	22.92	4.28	9.01	36.45	3.76	0.52
*VIF*s	**2.09**	**1.75**	**2.18**	**2.18**	**1.4**	CPU
s.e.	77.45	70.81	79.1	79.19	63.46	6.48

Table 2.2: The summary of the original information for the Diabete data, derived from the proposed algorithm

Figure 2.6.1 indicates the relationship between the *CN* and the *VIF*s. From the figure, we first observe that the *VIF*s for (a) *age*, (b) *sex*, (c) *bmi*, (d) *map* and (j) *glu* basically fluctuate around their original values, even though the *VIF*s for (c) *bmi* and (d) *map* increase slightly. As for those variables with pathological *VIF*s, their *VIF*s suggest a tendency to decrease over the iterations.

Figure 2.6.1 concerns two things. First, Figure 2.6.1 -(a) shows the relationship between the *CN* and the perturbation norm $\|\varpi\|_2$ over the iterations. By and large, the norm increases as the *CN* decreases, even though there are certain fluctuations after the 60-th iteration. This fact indicates that it indeed takes more effort for the perturbation to reduce to *VIF*s, as the ill-posedness is mitigated. The basically reflects the matrix property (Horn and Johnson, 2012). Second, Figure 2.6.1-(b) gives the numerical validation of the postive definiteness assumption (2.4).

2.6.2 An application to Collinear Data

We make an attempt to deal with collinear dataset as an example. The context of the collinear dataset results from a study on the equal opportunity in public education in United States. The objective was to examine the effect of school inputs on students' achievements. More details are referred to Chatterjee and Hadi (2006) and the related

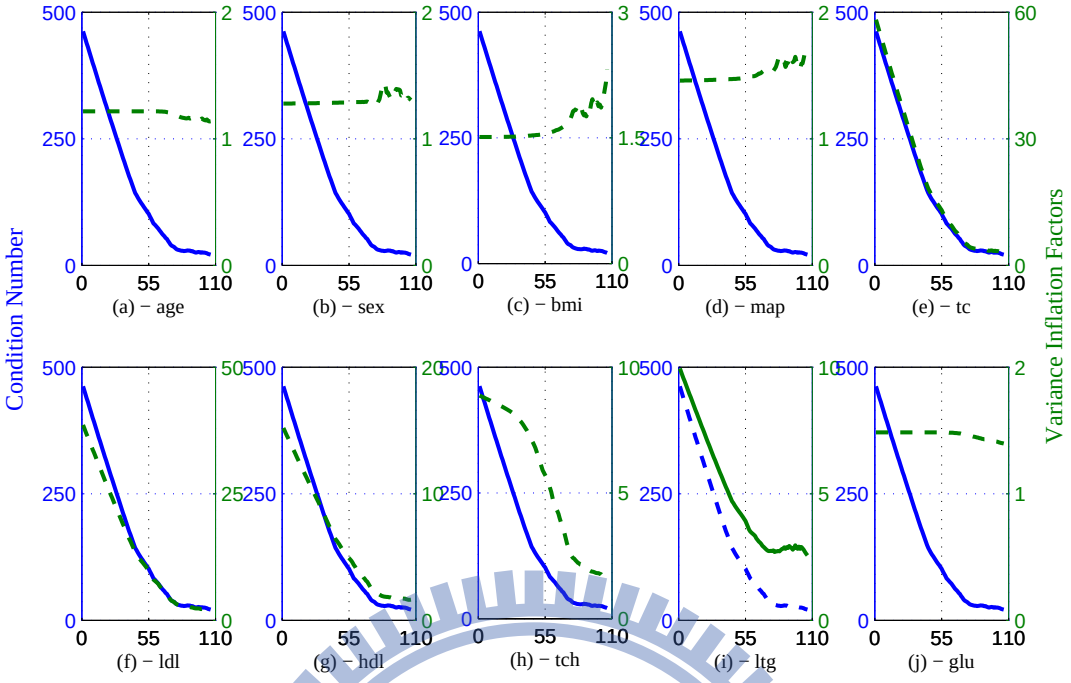

Figure 2.1: Figures (a) - (j) represent the relationship between the condition number and the *VIF*s for each explanatory variable in the diabetes dataset. (Solid Line – Left Axis; Dashed Line – Right Axis)

literature therein. There are 3 independent variables (*FAM*, *PEER* and *SCHOOL*), 1 dependent variable (*ACHV*) and 70 random measurements.

The regression information is summarized in Table 2.3 below. As the authors mentioned, a high F-statistic value (5.72) indicates that the three variables are valid as the explanatory variables, although the t-statistics reveal statistical insignificance individually.

Variables	*FAM*	*PEER*	*SCHOOL*	*CN*
$\widehat{\beta}_{OLS}$	1.16	1.74	−1.91	393.97
*VIF*s	**38.443**	**31.478**	**88.372**	R^2
s.e.	12.52	11.33	18.98	0.19

Table 2.3: The summary of the original information for the education data, derived from the ordinary least squares.

The information in Table 2.3, as well as the correlation matrix below, all suggest the

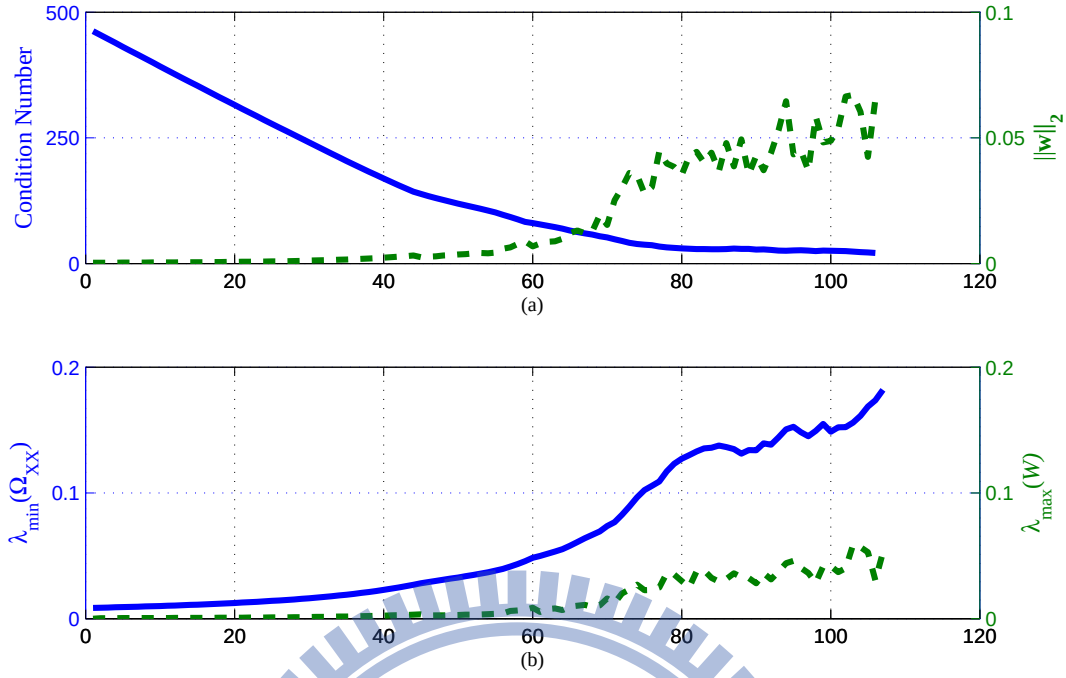

Figure 2.2: Figure (a) shows the relationship between the condition number and the perturbation norm; Figure (b) concerns th positive definiteness property of the perturbed correlation matrix for diabetes data. (Solid Line – Left Axis; Dashed Line – Right Axis)

strong collinearity among independent variables. The strong linear structure between pairs of the three variables does affect the estimates obtained in Table 2.3.

Vars	*FAM*	*PEER*	*SCHOOL*
FAM	1	0.959	0.986
PEER	–	1	0.983
SCHOOL	–	–	1

Table 2.4: The correlation matrix for variables FAM, PEER and SCHOOL.

We see that the original OLSE for *SCHOOL* suggests a negative marginal effect, with respect to standard deviations, on students' achievements, holding other factors constant. However, our approach suggests a more reasonably positive marginal effect for *SCHOOL*, as summarized in Table 2.5. The confliction between the original and pertured results suggest that more investigation may be needed to confirm how those variables are related.

Variables	*FAM*	*PEER*	*SCHOOL*	*CN*	R^2
$\widehat{\beta}_{\mathcal{W}}$	0.46	0.54	0.14	14.39	0.18
*VIF*s	**2.66**	**2.85**	**3.84**	Iter	CPU
s.e.	3.33	3.44	3.99	87	1.73

Table 2.5: The summary of the original information for the education data, derived from the proposed algorithm

Figure 2.6.2 shows that all *VIF*s,together with the *CN*, decreases. More to that, if we compare (c) to (a) and (b) in Figure 2.6.2, the proposed algorithm effectively renders the most pathological *VIF* to decrease stably over the iterations.

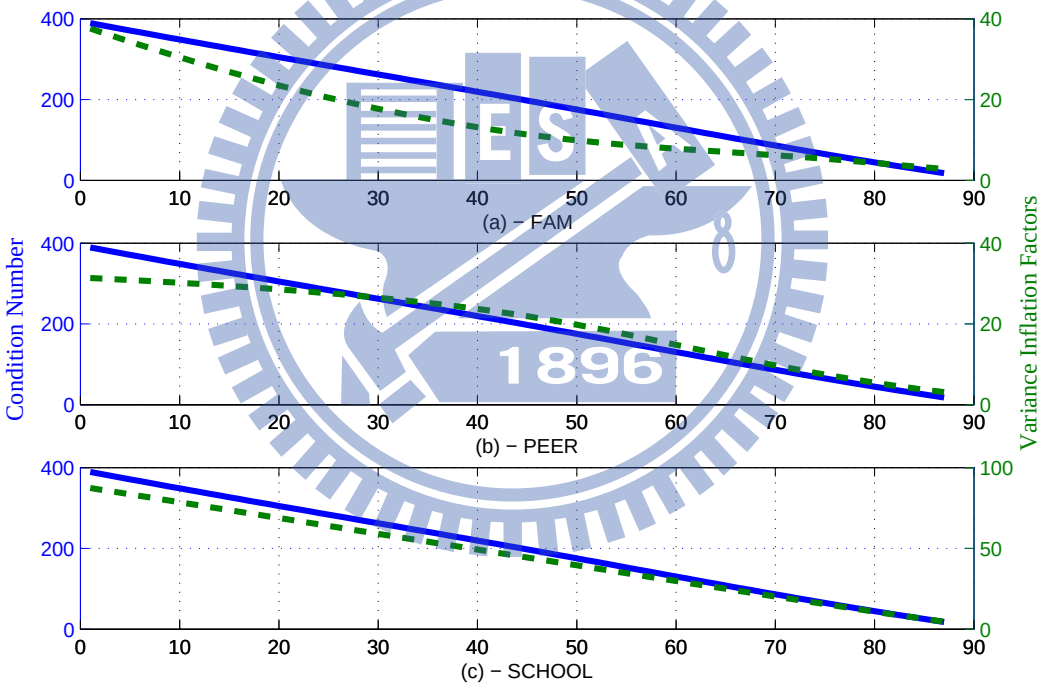

Figure 2.3: The relationships between the condition number and *VIF*s for (a) FAM, (b) PEER and (c) SCHOOL for education data. (Solid Line – Left Axis; Dashed Line – Right Axis)

The next figure shows (i) the relationship between the condition number and the perturbation norm; and (ii) the relationship between the minimal and maximal eigenvalues for Ω^*_{XX} and $\mathcal{W}$. In Figure 2.6.2-(a), the figure again confirms that, as the condition number drops down, the perturbation norm becomes larger. In Figure 2.6.2-(b) the positive

definiteness is firm accordingly.

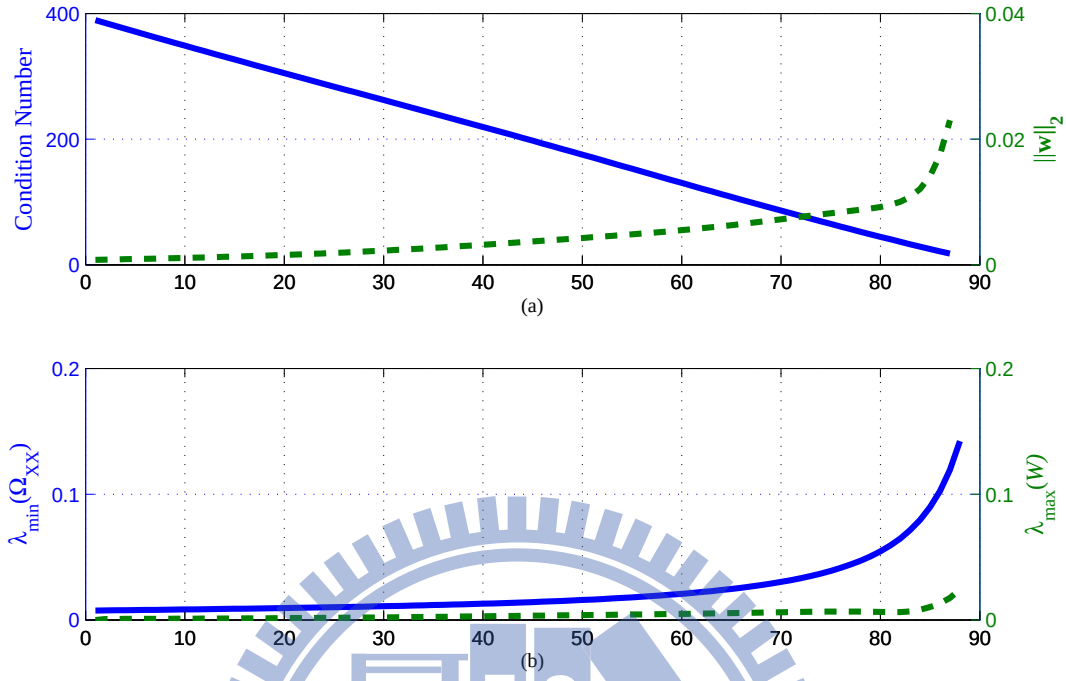

Figure 2.4: Figure (a) shows the relationship between condition number and the perturbation norm. Figure (b) concerns the positive definiteness of Ω^*_{XX} for education data. (Solid Line – Left Axis; Dashed Line – Right Axis)

2.6.3 The Sales Dataset

The following data represents a period of 23 years during which the firm was operating under fairly stable condition. The data shows the effect of the advertising expenditures (A_t and the lagged A_{t-1}), promotion expenditures (P_t and the lagged P_{t-1}), and sales expense (E_t) on the aggregate sales of a firm in period t. For details of the data, please see Chatterjee and Hadi (2006).

We summarize the OLS information is given in Table 2.6.

Vars	A_t	P_t	E_t	A_{t-1}	P_{t-1}	CN
$\widehat{\beta}_{OLS}$	2.35	3.9	3.15	1.6	2.03	233.92
VIFs	**36.94**	**33.47**	**1.08**	**25.92**	**43.52**	R^2
s.e.	6.81	6.48	1.16	5.7	7.39	0.92

Table 2.6: The summary of the original information for the advertising data, derived from the ordinary least squares

Applying the proposed approach to the dataset, the results are summarized in Table 2.7. The CPU time is 0.7 second for 47 iterations.

Vars	A_t	P_t	E_t	A_{t-1}	P_{t-1}	CN	R^2
$\widehat{\beta}_{\mathcal{W}}$	1.98	3.49	3.44	1.21	1.6	11.78	0.92
VIFs	**2.51**	**2.85**	**1.16**	**1.97**	**2.36**	Iters	CPU
s.e.	1.03	1.77	2.61	0.72	0.79	45	0.69

Table 2.7: The summary of the new information for the advertising data, derived from the proposed algorithm

In Figure 2.6.3, we observe that the good VIF for (c) E_t fluctuates upward slightly, while other pathological VIFs drop down enormously. All results reveal the same conclusion as in the previous two examples.

2.6.4 United States Pork Data

The following example illustrates the United States Pork dataset. The dataset contains the retail price (rp), deflated retail price (drp), net farm value (nfv), total price spread (tps), deflated price spread (dps), wage price (wp), wage price index (wpi), fuel price index (fpi), industrial cost (ic), deflated industrial cost (dic), quantity (q), consumer price index (cpi), and top 4 concentration rate ($cr4$) for US pork industry since 1970 to 2008.

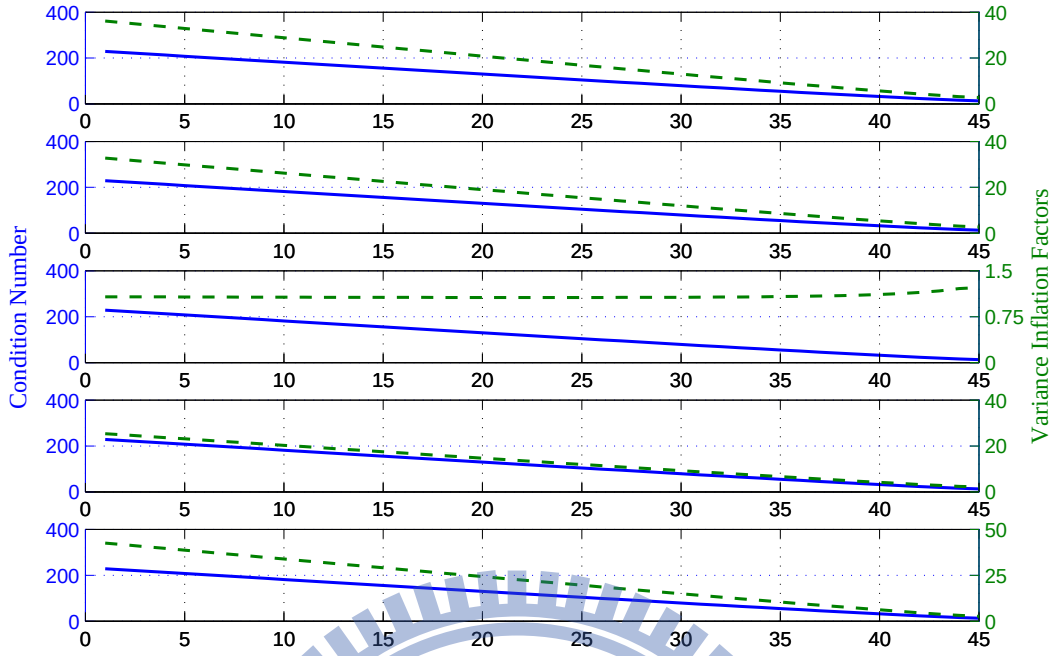

Figure 2.5: The relationship between the condition number and VIFs for (a) A_t (b) P_t (c) E_t (d) A_{t-1} and (e) P_{t-1} (From Top to Bottom) (Solid Line – Left Axis; Dashed Line – Right Axis)

Vars	lp	lq	ldw	$ldpe$	$lcr4$	$d98$	t
lp	1	−0.25	0.89	−.004	−0.79	−0.19	−0.86
lq	–	1	0.13	−.17	−0.04	0.07	−0.11
ldw	–	–	1	−0.05	−0.78	−0.19	−0.91
$ldpe$	–	–	–	1	−0.02	−0.27	0.12
$lcr4$	–	–	–	–	1	0.25	0.95
$d98$	–	–	–	–	–	1	0.19
t	–	–	–	–	–	–	1

Table 2.8: The correlation matrix for explanatory variables for US Pork Data

We refer the detailed context of the dataset to Zhen and Wohlgenant (2010). The model of interest is a detrended one, given by $lr = \beta_1 lp + \beta_2 lq + \beta_3 ldw + \beta_4 ldpe + \beta_5 lcr4 + \beta_6 d98 + \beta_7 t + \epsilon$, where lr is the log of the difference between deflated retail price and deflated price spread, lp is the log of deflated retail price, lq is the log of quantity, ldw is

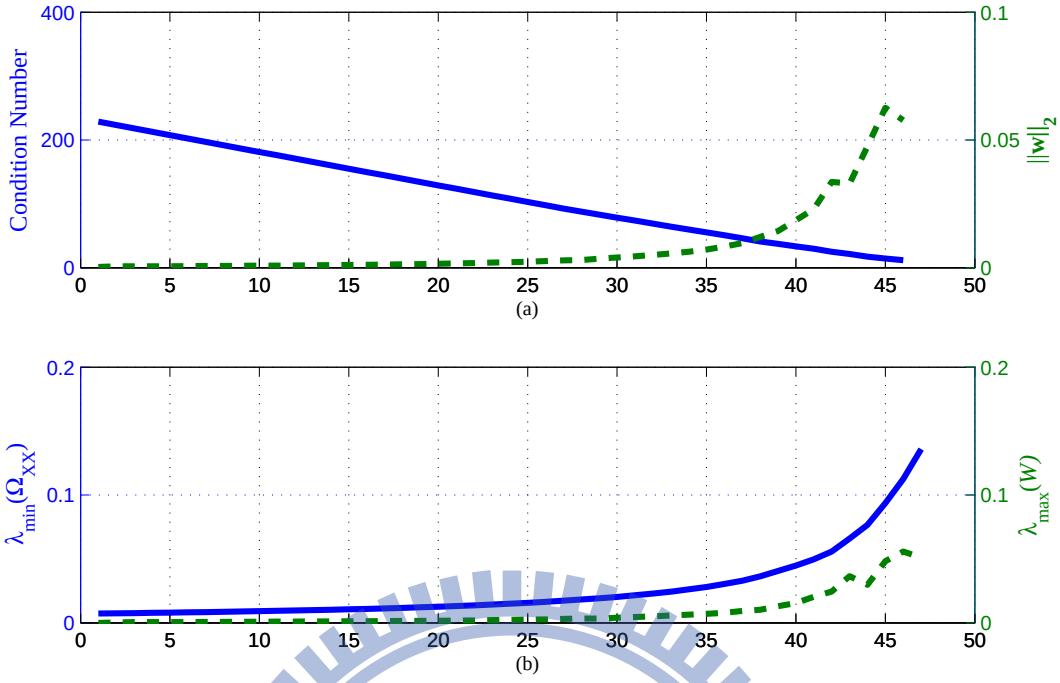

Figure 2.6: The figure shows (i) the relationship between the condition number and the perturbation norm; and (ii) the positive definiteness of Ω^*_{XX} for sales data. (Solid Line – Left Axis; Dashed Line – Right Axis)

the log of deflated wage price index, $ldpe$ is the log of deflated fuel price index, $lcr4$ is the log os top 4 concentration rates, $d98$ is equal 1 if the year is 1998 or 1999; and 0, otherwise, and t is the cardinal of the years. As reported, the results for the trended version have a sign difference between coefficients for ldw. Too see that, we first look at the correlation matrix given below. Table 2.9 contains the OLS information. The trending variable was intended to account for the trending variations existing among the variables. However, it turns out that introducing a detrending variable t invokes serious collinearity problem for the problem.

Vars	lp	lq	ldw	$ldpe$	$lcr4$	$d98$	t	CN
$\widehat{\beta}_{OLS}$	0.19	−0.04	−0.15	0.06	0.06	−0.05	−0.43	336.74
VIFs	18.45	3.61	22.38	1.87	25.19	1.17	58.79	R^2
s.e.	0.28	0.13	0.31	0.09	0.33	0.07	0.51	0.98

Table 2.9: The summary of the original information for (detrended) US pork data, derived from the ordinary least squares

More to that, if we look at the trended version of the problem, we see that the coefficient for ldw has effectively a different sign, see Table 2.10.

Vars	lp	lq	ldw	$ldpe$	$lcr4$	$d98$
$\widehat{\beta}_{OLS}$	0.21	-0.03	-0.01	0.01	0.2	-0.05
VIFs	18.45	3.61	22.38	1.87	25.19	1.17
s.e.	0.37	0.16	0.33	0.09	0.16	0.09

Table 2.10: The summary of the original information for (trended) US pork data, derived from the ordinary least squares

We attempt to see if the proposed approach can mitigate the effect from introducing the detrending variable t. Table 2.11 suggests the perturbed information for the detrended model.

Vars	lp	lq	ldw	$ldpe$	$lcr4$	$d98$	t	CN	R^2
$\widehat{\beta}_W$	0.19	-0.04	0.03	0.01	-0.15	-0.05	-0.14	14.1	0.96
VIFs	2.65	1.05	2.69	1.73	2.95	1.12	2.31	Iters	CPU
s.e.	0.19	0.12	0.19	0.15	0.2	0.12	0.18	197	3.53

Table 2.11: The summary of the original information for (detrended) US pork data, derived from the proposed algorithm

Two things can be observed. First, the proposed approach successfully reverses the sign for variable ldw, agreeing with that in the trended model. Second, however, as we see from Table 2.11, the coefficient for variable $lcr4$ becomes negative, meaning that the marginal percentage effect on lr has a 14% drops, holding others constant.

Figure 2.6.4 reveals the fact that the CN behaves as VIFs, especially those pathological ones, do. In Figure 2.6.4-(a), we observe that there are certain fluctuations on the curve for the perturbation norm over the iterations. Similarly, there is a quick incline in CN at around 120-th iteration. The phenomena may be due to certain implicit numerical issues. By and large, Figure 2.6.4-(a) confirms that as the CN goes down, the perturbation norm goes up. Figure 2.6.4-(b) simply confirms the positive definiteness of Ω_{XX}.

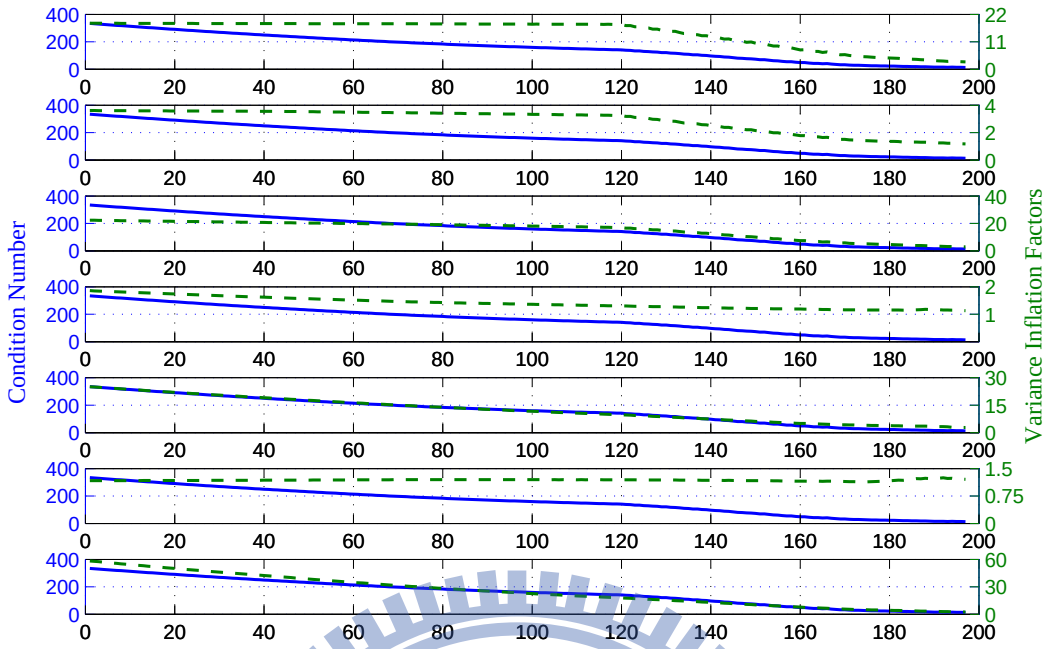

Figure 2.7: The relationship between the condition number and the *VIF*s for (a) *lp* (b) *lq* (c) *ldw* (d) *ldpe* (e) *lcr*4 (f) *d*98 and (g) *t* (From Top to Bottom) (Solid Line – Left Axis; Dashed Line – Right Axis)

2.7 Concluding Remarks

In this study, we propose a novel optimization model, based on the concept of *VIF*, to alleviate data collinearity problems in multiple linear regression. We show that the *VIF*s can decrease through solving the convex LCQP (2.15), using the proposed VIF-based algorithm. Various numerical examples validate the proposed approach.

The comparison between the proposed algorithm and the ridge regression can be unfair, because both approaches require exogenous parameters that directly affect the performance thereof. More to that, the relationship between the settings for ridge regression and our approach is not obviously related. In general, ridge regression outperforms the proposed algorithm in the *VIF* reduction.

There are a few issues needed to be solved. First of all, the development of the testing procedure for testing the hypothesis $H_0 : \{\varpi^*_{i,j} = \varpi_{i,j}\}$ against the alternative $H_1 : \{\varpi^*_{i,j} \neq \varpi_{i,j}\}$ is still missing in the literature. Once the testing procedure has been constructed, the examination of the perturbations generated by the LCQPs (2.15) can be

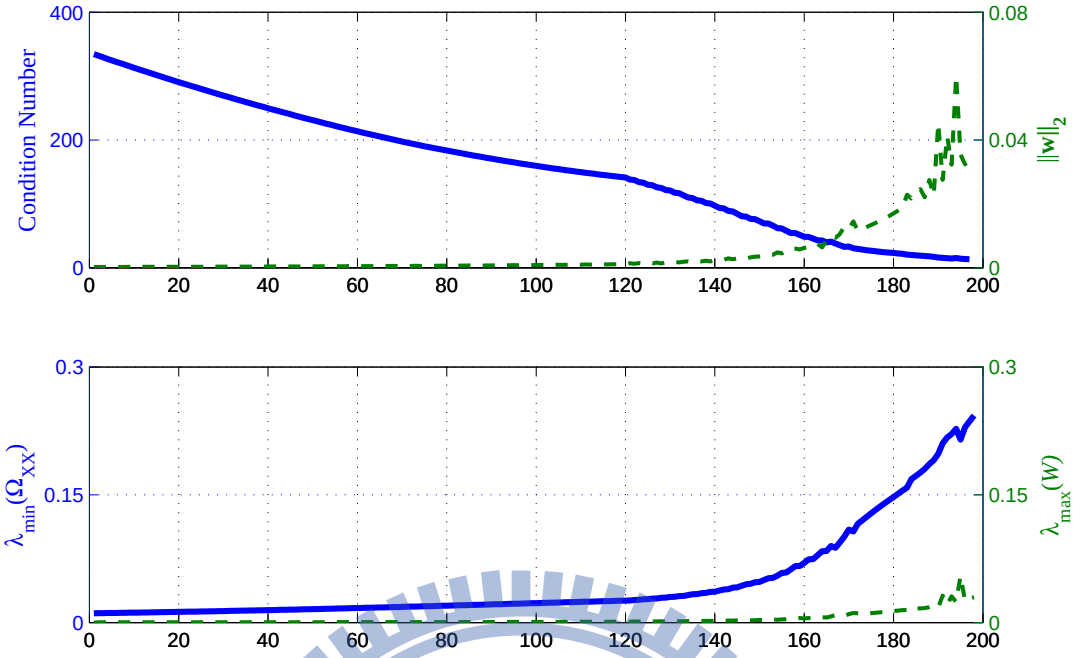

Figure 2.8: Figure (a) shows the relationship between the condition number and the perturbation norm; and Figure (b) concerns the positive definiteness of Ω^*_{XX}. (Solid Line – Left Axis; Dashed Line – Right Axis)

done. Second, as the numerical examples suggest, the algorithm still suffer from numerical instability after certain iterations. This is due to the feature that the proposed algorithm can decrease the *CN* while decreasing the *VIF*s. As the *CN* has dropped down to a certain level, the sensitivity of the matrix becomes weak. So, a stablization of the performance of the algorithm may be needed.

Chapter 3

Second-Order VIF-based Optimization Model for Linear Regression having Collinearity Problems

In this chapter, we study the linear regression having data collinearity problems. In order to achieve more improvements in the variance inflation factors, we propose a family of second-order VIF-based optimization models. Among the family, we pay special attention to one with the VIF constraint approximated up to the second order and with the objective unchanged. This particular model is a quadratically constrained quadratic programming (QCQP), an important branch of nonlinear programming problems. The established second-order model has a wide applicability that it can be used whether or not the existence of heteroskedasticity is observed.

3.1 Introduction

This chapter is sequel to Chpater 2 in that the authors proposed a VIF-based optimization model to treat linear regression models having collinearity problems, based upon the first-order approximation in both the OLSEs and the VIFs. Under the first-order

approximation, the authors showed that the problem is equivalent to solving a linearly-constrained quadratic programming (LCQP) which can be solved quite efficiently using any commercial software. Moreover, strict convexity of the established LCQP ensures an explicit solution to the problem, rendering the solution procedure much easier.

First-order approximation, albeit practically useful, provides merely a *rough* approximation to the problem in question. More precisions, and therefore more information, can be secured if one delves into higher-order approximations, say the second-order approximation. The more information we have in hand, the closer we get to the core of the problem. Therefore, it becomes our interest to derive a preciser second-order model for linear regression having collinearity problems.

Combinations of two criteria (bias and VIF) result in three potential variants for the second-order family. Since the VIF is more important than the bias itself, we shall focus on a particular variant that encompasses a second-order approximation on the VIF constraint, while the bias objective remains unchanged. This particular variant turns out to be a quadratically constrained quadratic programming problem (QCQP), having a possibly nonconvex constraint depending on the VIF associated.

QCQPs are a classic branch of nonlinear programming problems and are of both theoretical importance and practical interest. Various famous nonlinear programming problems belong to its clan. For instance, the Max-Cut problem, binary quadratic programs, trust-rigion method and the likes are all of QCQP.

Essentially, QCQP is NP-hard and cannot be solved in polynomial time. For convex QCQPs, given any precision, computationally efficient algorithms can be used to solve the problems in polynomial time. On the other hand, not surprisingly, the same does not apply to nonconvex QCQPs. For many years, researchers have endeavored themselves in developing computationally efficient algorithms for solving nonconvex QCQPs to approximate the global optimality. To name a few, Al-Khayyal et al. (1995) developed a relaxation method for solving nonconvex QCQPs; Audet et al. (2000) proposed a branch and cut algorithm to solve nonconvex QCQPs and similarly Linderoth (2005) proposed a simiplicial branch and bound algorithm;Anstreicher (2009) compared the semidefinite

programming (Vandenberghe and Boyd, 1996; Boyd and Vandenberghe, 2009) to the reformulation-linearization technique for nonconvex QCQPs; Anstreicher (2012) studied the convex relaxation for QCQPs. For more discussions, see Fang and Rajasekera (1986); Ben-Tal and Teboulle (1996); Nemirovskii and Scheinberg (1996); Zheng et al. (2011); Misener and Floudas (2012) and the references therein. See also Bao et al. (2011) for a thorough review.

The proposed second-order VIF-based model has a wide applicability. In light of Chapter 2, the proposed model is applicable to not only homoskedastic linear regression models but also heteroskedastic ones, both under the influence of data collinearity. When there is no departure from homoskedaticity, the situation is clear, *viz.*, one simply adds higher-order terms to the original VIF-based model given in Chapter 2. When the departure from homoskedasticity is observed, one either follows the traditional OLS-approach, with nothing changed, or one resport to the alternative WLS-approach, with the mere changes in the basic matrices.

This is how the rest of this study is organized. In Section 3.2, we show how to construct the second-order VIF-based model for linear regression having collinearity problems. The resulting optimization problem is nonlinear in essence. Several variants are also discussed. In Section 4.2, we apply the conic reformulation technique to solving the established optimization problem and design a solution procedure to achieve the global optimality. We also carry out numerical examples for validate the effectiveness of the method. Lastly, we give some concluding remarks in Section 3.4.

3.2 The Model

We shall follow Chapter 2, *viz.*, we aim to find a perturbation matrix $\mathcal{W}$ such that the ill-conditioned matrix X^TX becomes well-posed. Specifically, we wish to find such $\mathcal{W}$ so that $X^TX + \mathcal{W} \succ 0$ holds. Quantities imposed with the perturbation are said to be *perturbed*.

3.2.1 Objective Function

The superposition of $\mathcal{W}$ renders the new estimate $\widehat{\beta}$ biased. Thence, we wish that such a perturbation will result in a bias that is minimal.

The bias comes primarily from the ordinary least squares estimate (OLSE) that can be deemed a function of the correlation matrix, given the data matrix is centered and scaled. In particular, we see that

$$\widehat{\beta}_{\text{OLS}}(\Omega_{XX}) = (X^T X)^{-1} X^T \mathbf{y} \equiv \Omega_{XX}^{-1} \Omega_{X\mathbf{y}}, \tag{3.1}$$

where $X^T X \equiv \Omega_{XX}$ is the correlation matrix among the factors and $X^T \mathbf{y} \equiv \Omega_{X\mathbf{y}}$. The superposition of $\mathcal{W}$ onto Ω_{XX} results in the perturbed OLSE (POLSE)

$$\widehat{\beta}_{\text{OLS}}(\Omega_{XX}^*) = (X^T X + \mathcal{W})^{-1} X^T \mathbf{y} = \Omega_{XX}^{*-1} \Omega_{X\mathbf{y}}. \tag{3.2}$$

Now, applying the Taylor series expansion up to the second order to $\widehat{\beta}_{\text{OLS}}(\Omega_{XX}^*)$ about Ω_{XX} suggests

$$\begin{aligned}\widehat{\beta}_{\text{OLS}}(\Omega_{XX}^*) = \widehat{\beta}_{\text{OLS}}(\Omega_{XX}) \\ + \sum_{i>j} \frac{\partial \widehat{\beta}_{\text{OLS}}(\Omega_{XX})}{\partial \varpi_{i,j}} \Delta\varpi_{i,j} + \frac{1}{2!} \sum_{i>j} \sum_{k>l} \frac{\partial^2 \widehat{\beta}_{\text{OLS}}(\Omega_{XX})}{\partial \varpi_{i,j} \varpi_{k,l}} \Delta\varpi_{k,l} \Delta\varpi_{i,j},\end{aligned}$$

where the first-order partial derivative $\frac{\partial \widehat{\beta}_{\text{OLS}}(\Omega_{XX})}{\partial \varpi_{i,j}}$ is given by

$$\frac{\partial \widehat{\beta}_{\text{OLS}}(\Omega_{XX})}{\partial \varpi_{i,j}} = -\Omega_{XX}^{-1} (\mathbf{e}_i \mathbf{e}_j^T + \mathbf{e}_j \mathbf{e}_i^T) \widehat{\beta}_{\text{OLS}}, \tag{3.3}$$

for $i > j$, and

$$\frac{\partial^2 \widehat{\beta}_{\text{OLS}}(\Omega_{XX})}{\partial \varpi_{i,j} \varpi_{k,l}} = (I + \Omega_{XX}^{-1} (\mathbf{e}_k \mathbf{e}_l^T + \mathbf{e}_l \mathbf{e}_k^T)) \Omega_{XX}^{-2} (\mathbf{e}_i \mathbf{e}_j^T + \mathbf{e}_j \mathbf{e}_i^T) \widehat{\beta}_{\text{OLS}}, \tag{3.4}$$

for $i > j$ and $k > l$. Hence, (3.2.1) can be written as

$$\widehat{\beta}_{\text{OLS}}(\Omega^*_{XX}) = \widehat{\beta}_{\text{OLS}}(\Omega_{XX}) + B_1\varpi + B_2\varpi \otimes \varpi, \tag{3.5}$$

where $\otimes$ denotes the Kronecker product so that $\varpi \otimes \varpi \in \mathbb{R}^{(k(k-1)/2)^2}$, $B_1 \in \mathbb{R}^{k\times(k(k-1)/2)}$ is the matrix associated with the first-order approximation (3.3) and $B_2 \in \mathbb{R}^{k\times(k(k-1)/2)^2}$ is the matrix associated with the second-order approximation (3.4).

From (3.5), the difference between $\widehat{\beta}_{\text{OLS}}(\Omega^*_{XX})$ and $\widehat{\beta}_{\text{OLS}}(\Omega_{XX})$ can be seen as the bias and we want to minimize the resulting bias. In particular, we aim at minimizing

$$\min_{\varpi\in\mathbb{W}} \|B_1\varpi\|_2^2 + \|B_2\varpi \otimes \varpi\|_2^2 + \rho_1\|\varpi\|_2^2 + \rho_2\|\varpi \otimes \varpi\|_2^2 \tag{3.6}$$

where $\mathbb{W}$ denotes the feasible region containing those ϖ's feasible to the optimization problem, which soon will be discussed, and ρ_1 and ρ_2 are penalty parameters.

Since the term $\varpi \otimes \varpi$ is quadratic in ϖ, the established objective function can be further simplified as a standard quadratic form

$$\min_{\varpi\in\mathbb{W}} \frac{1}{2}\varpi^T\widetilde{B}\varpi + b^T\varpi \tag{3.7}$$

for symmetric matrix $\widetilde{B}$ and some vector b. Note that $\widetilde{B}$ may or may not be positive (semi)definite in essence. Therefore, the simplified objective function 3.7 can be nonconvex.

3.2.2 Feasible Region

The presence of collinearity renders the variances of the OLSEs unsatisfactorily high, reflecting in the associated *VIF*s. Therefore, we aim to reduce the *VIF*s through the help of the perturbation.

According to Belsley et al. (2005), a *VIF* is a function of the correlation matrix Ω_{XX}, in particular,

$$\mathcal{V}_l(\Omega_{XX}) = \frac{\det \Omega_{XX}\{l\}}{\det \Omega_{XX}}, \text{for } l = 1, \ldots, k, \tag{3.8}$$

where $\Omega_{XX}\{l\}$ is the l-th submatrix deriving from deleting the l-th column and row of Ω_{XX}. The perturbed *VIF* is therefore

$$\mathcal{V}_l(\Omega^*_{XX}) = \frac{\det \Omega^*_{XX}\{l\}}{\det \Omega^*_{XX}}, \text{for } l = 1, \ldots, k. \tag{3.9}$$

Applying Taylor approximation to (3.9) about Ω_{XX} up to the second order suggests

$$\mathcal{V}_l(\Omega^*_{XX}) = \mathcal{V}_i(\Omega_{XX}) + \mathbf{v}^T_{l,1}\varpi + \frac{1}{2!}\mathbf{v}^T_{l,2}\varpi \otimes \varpi, \text{for } l = 1, \ldots, k, \tag{3.10}$$

where $\mathbf{v}_{l,1} \in \mathbb{R}^{k(k-1)/2}$ and $\mathbf{v}_{l,2} \in \mathbb{R}^{(k(k-1)/2)^2}$ contain the first- and the second-order information about the changes in l-th *VIF*, respectively.

The components of $\mathbf{v}_{l,1}$ are the derivative $\frac{\partial \mathcal{V}_l}{\partial \varpi_{i,j}}$ which has been given in Chapter 2 and is given by

$$\frac{\partial \mathcal{V}_l(\Omega_{XX})}{\partial \varpi_{i,j}} = 2\left(\sum_{l=1}^{k} \frac{\upsilon_{l,i}\upsilon_{l,j}}{\lambda_l} - \sum_{l=1}^{k-1} \frac{\nu_{l,i}\nu_{l,j}}{\theta_l}\right)\mathcal{V}_l(\Omega_{XX}), \text{for } i > j \text{ and } l = 1, \ldots, k, \tag{3.11}$$

where $\upsilon_{l,i}$ ($\upsilon_{l,j}$) and $\nu_{l,j}$ ($\nu_{l,j}$) are the $i(j)$-th component of the l-th eigenvectors of the matrices Ω_{XX} and $\Omega_{XX}\{l\}$, respectively. Note that the indexes i and j run through different dimensions of the matrices Ω_{XX} and $\Omega_{XX}\{l\}$ in the terms in the parentheses in (3.11).

The components of $\mathbf{v}_{l,2}$ comprise the second-order derivative $\frac{\partial^2 \mathcal{V}_l}{\partial \varpi_{i',j'}\partial \varpi_{i,j}}$ which is given by

$$\begin{aligned}\frac{\partial^2 \mathcal{V}_l(\Omega_{XX})}{\partial \varpi_{i',j'}\partial \varpi_{i,j}} &= 4\left(\sum_{l=1}^{k} \frac{\upsilon_{l,i}\upsilon_{l,j}}{\lambda_l} - \sum_{l=1}^{k-1} \frac{\nu_{l,i}\nu_{l,j}}{\theta_l}\right)\left(\sum_{l=1}^{k} \frac{\upsilon_{l,i'}\upsilon_{l,j'}}{\lambda_l} - \sum_{l=1}^{k-1} \frac{\nu_{l,i'}\nu_{l,j'}}{\theta_l}\right)\mathcal{V}_l(\Omega_{XX}) \\ &+ \left(\sum_{l=1}^{k} \frac{\frac{\partial \lambda_l}{\partial \varpi_{i',j'}}\frac{\partial \lambda_l}{\partial \varpi_{i,j}} - \lambda_l \frac{\partial^2 \lambda_l}{\partial \varpi_{i,j}\partial \varpi_{i',j'}}}{\lambda_l^2} - \sum_{l=1}^{k-1} \frac{\frac{\partial \theta_l}{\partial \varpi_{i',j'}}\frac{\partial \theta_l}{\partial \varpi_{i,j}} - \theta_l \frac{\partial^2 \theta_l}{\partial \varpi_{i,j}\partial \varpi_{i',j'}}}{\theta_l^2}\right)\mathcal{V}_l(\Omega_{XX}),\end{aligned} \tag{3.12}$$

for $l = 1, \ldots, k$.

It remains to specify the first- and second-order derivatives in (3.12). The first deriva-

tive $\frac{\partial \lambda_l}{\partial \varpi_{i,j}}$ has been given in Chapter 2 and has the following expression

$$\begin{aligned} \frac{\lambda_l}{\partial \varpi_{i,j}} &= 2\upsilon_{l,i}\upsilon_{l,j}; \\ \frac{\upsilon_l}{\partial \varpi_{i,j}} &= (\Omega_{XX} - \lambda I)^{\dagger}(\mathbf{e}_i\mathbf{e}_j^T + \mathbf{e}_j\mathbf{e}_i^T)\upsilon_l, \text{for } l = 1, \ldots, k, \end{aligned} \tag{3.13}$$

where $\upsilon_{l,i}$ denotes the i-th element of the eigenvector associated with the l-th eigenvalue and $(\Omega_{XX} - \lambda I)^{\dagger}$ is the Moore-Penrose generalized inverse of the singular matrix. Similar expressions can be obtained fairly easily for the derivative with respect to $\varpi_{i',j'}$.

To obtain an expression for the second derivative $\frac{\partial^2 \lambda_l}{\partial \varpi_{i,j}\partial \varpi_{i',j'}}$ and $\frac{\partial^2 \upsilon_l}{\partial \varpi_{i,j}\partial \varpi_{i',j'}}$, two things are needed: (i) the eigensystem with respect to the l-th eigenvalue, in particular,

$$\begin{aligned} (\Omega_{XX} - \lambda_l I)\upsilon_l &= 0; \\ \upsilon_l^T \upsilon_l &= 1; \end{aligned} \tag{3.14}$$

and (ii) the derivative of the generalized inverse of a matrix (P. Decell Jr, 1974), described in the following theorem

Theorem 3.1 (Derivative of the Moore-Penrose Generalized Inverse)**.** *Given a singular matrix $A \in \mathbb{R}^{n\times m} \equiv [a_{i,j}]$, denote the Moore-Penrose generalized inverse by $A^{\dagger} \in \mathbb{R}^{m\times n}$. The first-order derivative with respect to $a_{i,j}$, $A^{\dagger\prime}$, is given by*

$$A^{\dagger\prime} = -A^{\dagger}A'A^{\dagger} + (A'^T A^{\dagger T} A^{\dagger} + A^{\dagger}A^{\dagger T}A'^T) - A^{\dagger}A(A'^T A^{\dagger T} A^{\dagger} + A^{\dagger}A^{\dagger T}A'^T)AA^{\dagger}; \tag{3.15}$$

wherein A' is the first-order derivative of A.

Proof. The proof of Theorem 3.1 has a similar argument as that in P. Decell Jr (1974) and therefore is sketched in what follows.

According to the definition of the generalized inverse of a singular matrix, the generalized inverse must satisfy: (i) $AA^{\dagger}A = A$; (ii) $A^{\dagger}AA^{\dagger} = A^{\dagger}$; (iii) $(AA^{\dagger})^T = AA^{\dagger}$; and (iv) $(A^{\dagger}A)^T = A^{\dagger}A$. Then the derivative $A^{\dagger\prime}$ is the solution to the system of equations deriving from taking derivatives, elementwise, on $(i) - (iv)$. □

With the above, we can obtain the expressions for the second derivatives

$$\frac{\partial^2 \lambda_l}{\partial \varpi_{i',j'} \partial \varpi_{i,j}} = 2(\frac{\partial \upsilon_{l,i}}{\partial \varpi_{i',j'}} \upsilon_{l,j} + \upsilon_{l,i} \frac{\partial \upsilon_{l,j}}{\partial \varpi_{i',j'}});$$
$$\frac{\partial^2 \upsilon_l}{\partial \varpi_{i',j'} \partial \varpi_{i,j}} = \frac{\partial (\Omega_{XX} - \lambda I)^\dagger}{\partial \varpi_{i',j'}} (\mathbf{e}_i \mathbf{e}_j^T + \mathbf{e}_j \mathbf{e}_i^T) \upsilon_l + (\Omega_{XX} - \lambda I)^\dagger (\mathbf{e}_i \mathbf{e}_j^T + \mathbf{e}_j \mathbf{e}_i^T) \frac{\partial \upsilon_l}{\partial \varpi_{i',j'}};$$

wherein the first term on the RHS of the second equality exploits (3.15). Note that $\frac{\partial \upsilon_{l,i}}{\partial \varpi_{i',j'}} = \mathbf{1}_{\{i=i'\}}$ is binary and the same for $\frac{\partial \upsilon_{l,j}}{\partial \varpi_{i',j'}}$.

Taking the derivative of (3.14) twice, with respect to $\varpi_{i,j}$ and $\varpi_{i',j'}$ respectively, suggests the following system of equations

$$\frac{-\partial^2 \lambda_l}{\partial \varpi_{i',j'} \partial \varpi_{i,j}} \upsilon_l + \frac{\partial (\Omega_{XX} - \lambda_l I)}{\partial \varpi_{i',j'}} \frac{\partial \upsilon_l}{\partial \varpi_{i,j}} + \frac{\partial (\Omega_{XX} - \lambda_l I)}{\partial \varpi_{i,j}} \frac{\partial \upsilon_l}{\partial \varpi_{i',j'}} + \frac{(\Omega_{XX} - \lambda_l I) \partial^2 \upsilon_l}{\partial \varpi_{i',j'} \partial \varpi_{i,j}} = 0;$$
$$\frac{\partial^2 \upsilon_l}{\partial \varpi_{i',j'} \partial \varpi_{i,j}} + \frac{\partial \upsilon_l}{\partial \varpi_{i',j'}} \frac{\partial \upsilon_l}{\partial \varpi_{i,j}} = 0;$$

where the first-order derivatives are given in (3.13) and the second-order in (3.2.2).

From Chapter 2, it becomes apparent that the difference represents the changes in the *VIF*. In particular, defining v_r as the reduction in the i-th *VIF*, we obtain the following nonlinear constraint

$$\mathbf{v}_{i,1}^T \varpi + \frac{1}{2!} \mathbf{v}_{i,2}^T \varpi \otimes \varpi \leq -v_r, \tag{3.16}$$

for $1 \leq i \leq k$. Now, define

$$k^* \equiv \{1 \leq j \leq k : \mathcal{V}_j(\Omega_{XX}) \geq \mathcal{V}_i(\Omega_{XX}) \text{ for } i \neq j\} \tag{3.17}$$

as the index(es) associated with the OLSE(s) having the largest *VIF* value. Then (3.16) with respect to (3.17) becomes

$$\mathbf{v}_{k^*,1}^T \varpi + \frac{1}{2!} \mathbf{v}_{k^*,2}^T \varpi \otimes \varpi \leq -v_r. \tag{3.18}$$

It then follows that the feasible region comprises those ϖ's feasible to (3.19), *viz.*,

$$\mathbb{W} \equiv \left\{ \varpi \in \mathbb{R}^{k(k-1)/2} : \mathbf{v}_{k^*,1}^T \varpi + \frac{1}{2!} \mathbf{v}_{k^*,2}^T \varpi \otimes \varpi \leq -v_r \right\} \tag{3.19}$$

which is quadratic in essence.

3.2.3 A Nonlinear Optimization Model and Its Variants

It remains to combine (3.6) and (3.19) to obtain the following quadratically-constrained nonlinear optimization model

$$\min_{\varpi \in \mathbb{W}} \left\{ \|B_1 \varpi\|_2^2 + \|B_2 \varpi \otimes \varpi\|_2^2 + \rho_1 \|\varpi\|_2^2 + \rho_2 \|\varpi \otimes \varpi\|_2^2 \right\} \tag{3.20}$$

where ρ_1 and ρ_2 are exogenous parameters.

In conjunction with Chapter 2, various combinations of VIF-based models can be obtained, depending on what the focus is. For instance, compared to the bias, it makes more sense to have better precision in *VIF*, because *VIF* relates directly to the statistical significance. Thence, we may as well consider the variant

$$\min_{\varpi \in \mathbb{W}} \left\{ \frac{1}{2} \left(\|B_1 \varpi\|_2^2 + \rho_1 \|\varpi\|_2^2 \right) \right\}, \tag{3.21}$$

which is now a quadratically-constrained quadratic programming (QCQP) model. In this study, we shall play with the established QCQP (3.21), for the importance of the values of the biases is relatively minor.

To end this section, we note that since the variable $\varpi \otimes \varpi$ in the constraint is composed of the quadratic and bilinear terms of the $\varpi_{i,j}$'s, it then can be expressed as the following general form

$$\mathbf{v}_{k^*,1}^T \varpi + \frac{1}{2!} \mathbf{v}_{k^*,2}^T \varpi \otimes \varpi \leq -v_r \Rightarrow \frac{1}{2} \varpi^T V_{k^*} \varpi + \mathbf{v}_{k^*}^T \varpi \leq -v_r, \tag{3.22}$$

for some appropriate symmetric matrix V_{k^*} and vector $\mathbf{v}_{k^*}$. Consequently, 3.19 becomes

$$\mathbb{W} \equiv \left\{ \varpi \in \mathbb{R}^{k(k-1)/2} : \frac{1}{2} \varpi^T V_{k^*} \varpi + \mathbf{v}_{k^*}^T \varpi \leq -v_r \right\}. \tag{3.23}$$

3.3 Solution Strategies

Since the objective function in (3.21) is strictly convex given any $\rho > 0$, the convexity of the VIF constraint then becomes the key in determining the difficulty of the QCQP (3.21). In what follows, we shall discuss two possible cases, convex and nonconvex, and design an associated solution algorithm for each case.

3.3.1 Convex Case - Simple Case

Let us first consider the case in that the matrix V_{k^*} in the VIF constraint of (3.21) is positive definite. In this case, we have convex QCQP problem that can be solved rather efficiently using any commercial solver.

The convexity of the problem (3.21) ensures that solving the system of K-K-T optimality conditions leads to the global optimality. In particular, the corresponding K-K-T system is given by

$$\begin{aligned}
&\bigtriangledown \mathcal{L}(\varpi, \mu)|_{\varpi=\varpi^*} = \left(B_1^T B_1 + \rho_1 I + \mu V_{k^*}\right)\varpi^* + \mu \mathbf{v}_{k^*} = 0; \\
&\frac{1}{2}\varpi^{*T} V_{k^*}\varpi^* + \mathbf{v}_{k^*}^T \varpi^* + v_r \leq 0, \quad \mu \geq 0; \\
&\mu\left(\frac{1}{2}\varpi^{*T} V_{k^*}\varpi^* + \mathbf{v}_{k^*}^T \varpi^* + v_r\right) = 0,
\end{aligned} \tag{3.24}$$

where μ is the Lagrangian multiplier associated with the VIF constraint. It then is not hard to see that the optimal perturbation ϖ^* has the following expression

$$\varpi^* = -\mu^*(B_1^T B_1 + \rho I + \mu^* V_{k^*})^{-1}\mathbf{v}_{k^*} \tag{3.25}$$

where ϖ^* solves the unconstrained dual problem

$$\max_{\mu \geq 0}\left\{v_r \mu - \frac{\mu^2}{2}\mathbf{v}_{k^*}^T \left(B_1^T B_1 + \rho I + \mu V_{k^*}\right)^{-1}\mathbf{v}_{k^*}\right\}, \tag{3.26}$$

which can be solved using any numerical searching scheme.

3.3.2 Nonconvex Case - Conic Reformulation

Now, consider that case in that the constraint in (3.21) is non-convex, so (3.21) becomes non-convex in essence. As aforementioned, a general nonconvex QCQP is NP-hard and thus cannot be solved in poly-nomial time, unless NP = P (Cormen et al., 2001). Despite the fact, it is still our aim to design a non-heuristic algorithmic procedure to solve the resulting QCQP (3.21). To do this, we shall resort to the LCOP technique to solve the problem.

First, to obtain a conic relaxation for the established QCQP (3.21), let us first define the following sets

$$\begin{aligned} \mathcal{D}_{(k^2-k+2)/2} &= \left\{ U \in \mathcal{M}_{(k^2-k+2)/2} : \begin{bmatrix} 1 \\ \varpi \end{bmatrix}^T U \begin{bmatrix} 1 \\ \varpi \end{bmatrix} \geq 0, \forall \varpi \in \mathbb{W} \right\}; \\ Z &= \left\{ Y \in \mathcal{M}_{(k^2-k+2)/2} : Y = \begin{bmatrix} 1 \\ \varpi \end{bmatrix}^T \begin{bmatrix} 1 \\ \varpi \end{bmatrix}, \text{ for some } \varpi \in \mathbb{W} \right\}; \end{aligned} \tag{3.27}$$

where $\mathcal{D}_{(k^2-k+2)/2}$ is a closed convex cone with dimension $(k(k-1)/2+1)\times(k(k-1)/2+1)$, Z is a set of matrices of the same dimension and $\mathbb{W}$ is the feasible region and has been given in (3.19). The dual of the cone $\mathcal{D}_{(k^2-k+2)/2}$ is given by $Closure(Cone(Z)) \equiv \mathcal{D}^*_{(k^2-k+2)/2}$, the closure of the cone associated with the set Z, according to Lu et al. (2011).

With the above, we obtain the following nonconvex linear conic programming problem (LCOP)

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}(B_1^T B_1 + \rho_1 I) \bullet W \\ \text{subject to} \quad & \frac{1}{2} V_{k^*} \bullet W + \mathbf{v}_{k^*} \varpi + v_r \leq 0. \\ & \begin{bmatrix} 1 & \varpi^T \\ \varpi & W \end{bmatrix} = Y; \\ & Y \in \mathcal{D}^*_{(k^2-k+2)/2}; \text{ and} \\ & \text{rank}(Y) = 1, \end{aligned} \tag{3.28}$$

wherein the rank constraint is nonconvex in essence. By relaxing the rank constraint, we

obtain the following convex LCOP

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}(B_1^T B_1 + \rho_1 I) \bullet W \\ \text{subject to} \quad & \frac{1}{2} V_{k^*} \bullet W + \mathbf{v}_{k^*} \varpi + v_r \leq 0. \\ & \begin{bmatrix} 1 & \varpi^T \\ \varpi & W \end{bmatrix} = Y; \\ & Y \in \mathcal{D}^*_{(k^2-k+2)/2}, \end{aligned} \tag{3.29}$$

associated with the original QCQP (3.21). Now, applying the conic duality theory (Shapiro, 2001), the dual of the LCOP (3.29) is given by

$$\begin{aligned} \text{maximize} \quad & \frac{1}{2}\kappa - \mu v_r \\ \text{subject to} \quad & \begin{bmatrix} -\kappa & \mu \mathbf{v}_{k^*}^T \\ \mu \mathbf{v}_{k^*} & (B_1^T B_1 + \rho_1 I + \mu V_{k^*}) \end{bmatrix} \in \mathcal{D}_{(k^2-k+2)/2}; \\ & \mu \geq 0. \end{aligned} \tag{3.30}$$

We have the following theorem

Theorem 3.2 (Equivalence (Lu et al., 2011)). *If the QCQP (3.21) has a finite optimal value, then the established (3.29), (3.30) and (3.21) are equivalent in the sense that they all have the same optimal value.*

3.3.3 Solution Algorithm

We shall design a solution algorithm for deriving the POLSEs based upon the conic reformulation (3.30).

Algorithm. The basic structure of the algorithm is basically the same as that in Jou et al. (2014), with the only difference residing in the subroutine for getting the optimal

perturbation. A generic structure of the proposed algorithm is given in the following pseudo-code, see Algorithm 2.

Algorithm 2 Iterative Algorithm for POLSEs

Require: Problem input $\mathcal{P} = (X, \mathbf{y}, v_r = 1, v_l = 6, \rho = 50)$
Ensure: Optimal POLSEs $\widehat{\beta}_{\mathcal{W}}$;

```
repeat
    Generate matrix B_1 for objective function, and matrix V_{k*} and vector v_{k*} for constraint;
    (Subroutine) Obtain the optimal perturbation ϖ*
    Update POLSE;
until max_{j=1,...,k} V̂_i ≤ v_l
```

Subroutine. It remains to specify the subroutine for obtaining the optimal perturbation ϖ^*. To do this, we first note that the decision variable of the LCOP (3.30) is not ϖ, while that of (3.29) is. Despite this fact, according to Theorem 3.2, (3.30) and (3.29) share the same optimal value. Thence, an immediate intuition to circumvent the problem is to solve (3.30) for the optimal value, and then impose the optimal-value-constraint on (3.29) to obtain the optimal perturbation ϖ^*.

To be specific, let us assume the pair (κ^*, μ^*) solves (3.30) with the optimal value $\kappa^* - \mu^* v_r$. Then, to obtain the optimal perturbation, we solve the revised QCQP

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{2}(B_1^T B_1 + \rho_1 I) \bullet W \\
\text{subject to} \quad & \frac{1}{2} V_{k^*} \bullet W + \mathbf{v}_{k^*} \varpi + v_r \leq 0. \\
& \begin{bmatrix} 1 & \varpi^T \\ \varpi & W \end{bmatrix} = Y; \\
& Y \in \mathcal{D}^*_{(k^2-k+2)/2}, \\
& \frac{1}{2}(B_1^T B_1 + \rho_1 I) \bullet W = \frac{1}{2}\kappa^* - \mu^* v_r
\end{aligned} \tag{3.31}$$

wherein the last equality is the optimal-value-constraint added to (3.29).

Notably, an LCOP problem may or may not be computable, depending on the intrinsic structure of the underlying cone. The key to solving an LCOP problem then resides in finding a cone so that the problem becomes computable essentially.

Hence, the last step to solving the LCOP (3.30) is to find a computable cone $\mathcal{C}^*_{(k^2-k+2)/2}$ such that $\mathcal{C}^*_{(k^2-k+2)/2} \supseteq \mathcal{D}^*_{(k^2-k+2)/2}$ and $\mathcal{C}_{(k^2-k+2)/2} \subseteq \mathcal{D}_{(k^2-k+2)/2}$ on the dual side. The question as to how to find the computable cone $\mathcal{C}^*_{(k^2-k+2)/2}$ will be referred to Lu et al. (2011) for further details.

The result follows.

Theorem 3.3. *If (3.29) has finite optimal value, then there exists at least one ϖ^* such that $W = \varpi^*\varpi^{*T}$ and that the optimal value is $\frac{1}{2}\kappa^* - \mu^* v_r$.*

We are ready to state the subroutine for deriving the optimal perturbation.

Algorithm 3 Subroutine

Require: Problem input $B_1, \rho_1, V_{k^*}, \mathbf{v}_{k^*}$ and v_r

Ensure: Optimal Perturbation ϖ^*

1: Solve the LCOP (3.30) to obtain (κ^*, μ^*);

2: Solve the revised QCQP (3.31) to obtain the ϖ^*

3.4 Concluding Remarks

We propose a family of second-order VIF-based optimization models for linear models having collinearity problem. Based upon the previous work in Chapter 2, the second-order model can be applied whether or not the heteroskedasticity may be present. More specifically, the departure from homoskedasticity can be dealt with by either constructing new optimization models or delivering the optimal perturbation matrix generated by the original QCQP (3.31) directly to the associated HCEs.

The well-founded theory for linear conic programming facilitates solving the original nonconvex nonlinear programming problem. Notably, the difficulty of the solving the problem remains unchanged, however.

Chapter 4

VIF-based Iterative Matrix Perturbation Method for Heteroskedasticity-Robust Covariance Estimators in the Presence of Collinearity

In this study, we investigate linear regression having both heteroskedasticity and collinearity problems. We discuss the properties related to the perturbation method. Important observations are summarized as theorems. We then prove the main result that states the heteroskedasticity-robust variances can be improved and that the resulting bias is minimized by using the matrix perturbation method. We analyze a practical example for validation of the method.

4.1 Introduction

In the context of linear regression, the major problem of having heteroskedasticity resides in the invalidity of the usual test statistics, say t- or F-statistic, that they no longer follow their original distributions, regardless of the sample size. To circumvent the problem, the

heteroskedasticity-robust covariance estimate (HCE) is employed (White, 1980; Cribari-Neto et al., 2007, 2014). The key idea of the HCE is to take advantage of the ordinary least squares (OLS) residual, by which an HCE is proven able to remove the effect of heteroskedasticity in a consistent sense. Correct inference is then drawn using the variance portion lying on the main diagonal of the HCE associated, instead of the original covariance estimate. For more dicussions, see Andrews (1991); Bollerslev (1986); Engle (2002); MacKinnon and White (1985) and the references therein; see also MacKinnon (2013) for a thorough discussion.

Regression modeling is subject to another source of problems coming primarily from the nature of the collected data, called the *data problem* (Belsley et al., 2005). The most commonly seen data problems include the existence of outliers, missing inputs, the presence of collinearity among covariates and so forth. Of all data problems, we are more concerned about the collinearity problem. High dependency renders the OLS estimation procedure involving matrix inversion problematic, resulting in consequences inclusive of inflated variances, weakened power in prediction and even controversial signs against the hypothesis (Belsley et al., 2005). Biased regressions, such as the ridge regression (Hoerl and Kennard, 1970), the LASSO (Tibshirani, 1996), the bridge regression (Frank and Friedman, 1993) and the likes (Mandel, 1982; Wold et al., 1984; Stewart, 1987; Lin et al., 2011), have been widely accepted methods for dealing with such problems in regression.

Chances are that one encounters the situation in that the regression model suffers from both heteroskedasticity and collinearity problems at the same time. Little work has been done for such a class of models in the literature, though. Most of the existing work is based on the ridge-type estimates. For instance, Alheety and Kibria (2009) discussed such a problem based on the Liu estimate (Liu, 2003) which is a variant of the OLSEs. In a most recent paper, Aslam (2013) studied the problem by directly applying the ridge-type estimates to the HCEs. However, only the numerical evidences were provided. For some recent developments, see Alkhamisi (2012); Khalaf (2012); Aslam (2014); Månsson et al. (2014) and the references therein. When both problems exist, an immediate consequence resides in the possibly inflated variances in the HCE associated, making the subsequent

inference less significant in the statistical sense. The major work then is to employ the matrix perturbation method in Chapter 2 to the new class of linear regression models having both heteroskedasticity and collinearity problems at the same time. We aim to show that the heteroskedasticity-robust variances can be improved by imposing an optimal perturbation matrix on the HCE, when using the OLS approach.

We summarize our contributions in what follows. First, we provide theoretical justifications as the ground support for the matrix perturbation method developed in Chapter 2. Important theorems such as the eigenvalue perturbation theorem are discussed. Second, we prove that the optimal perturbation matrix can be directly applied to improve the heteroskedasticity-robust variances in the presence of data collinearity. Monotonic improvements can be observed under certain condition. Third, because the heteroskedasticity-robust variances are improved, this implies that the perturbations help shorten the predicted 1-S.D. heteroskedasticity-robust intervals for the unknown parameters.

This is how this study is organized. In Section 4.2, we study the properties related to the iterative matrix perturbation. Several important observations are summarized as theorems. Implications are discussed. In Section 4.3, we state the main results on the heteroskedasticity-robust variances and the sketch of proofs are given. Perturbed testing statistics and the associated confidence intervals are given in Section 4.4. A practical example is analyzed using the proposed method and results and discussions are given in Section 4.5. Conclusions and future research directions are in Section 4.6.

4.2 Some Properties

In this section, we study some properties regarding the perturbation method and the sketch of proofs are given.

It has been shown there exists an unique optimal perturbation matrix $\mathcal{W}$ such that $\Omega^*_{XX} \equiv \Omega_{XX} + \mathcal{W} \succ 0$, meaning positive definite, and that the resulting bias is minimized, where $X^T X \equiv \Omega_{XX}$ with X normalized. In particular, $\mathcal{W}$ comprises the components of

the vector $\varpi^* = [\Delta\varpi_{i,j}]_{i>j}$

$$\varpi^* = \frac{-v_r}{\|\mathbf{v}_{k^*}\|^2_{(B^TB+\rho I)^{-1}}} \left(B^TB + \rho I\right)^{-1} \mathbf{v}_{k^*}, \tag{4.1}$$

where v_r is the reduction step in the (largest) VIF; $\mathbf{v}_{k^*}$ is the vector made of the derivative $\partial(\Omega^{-1}_{XX})_{k^*,k^*}/\partial\varpi_{i,j}$, for $i > j$; k^* corresponds to the index of the largest VIF; B is a matrix made of the derivative $\partial\widehat{\beta}_{\text{OLS}}/\partial\varpi_{i,j}$, for $i > j$; and $\rho(> 0)$ is a penalty parameter. The optimal perturbation matrix $\mathcal{W}$ is then formed by the relationship $\mathcal{W} = [\Delta\varpi_{i,j}]_{i\neq j}$, with $\mathcal{W}_{i,i} = 0$, for all $i = 1, \ldots, k$.

The presence of data collinearity renders the matrix Ω_{XX}, with X centered and scaled, ill-conditioned and thus the eigen-structure is very sensitive to small perturbations on the elements (Horn and Johnson, 2012). The perturbation method inherits this merit from matrix theory and works by changing the eigen-structure to achieve the goal. The first theorem concerns the changes in the eigenvalues of the matrix Ω_{XX} through perturbation.

Theorem 4.1 (Eigenvalue Perturbation Theorem). *Let $0 < \lambda_{\min} \leq \lambda_2 \leq \cdots \leq \lambda_{\max}$ and $0 < \lambda^*_{\min} \leq \lambda^*_2 \leq \cdots \leq \lambda^*_{\max}$ be the eigenvalues for Ω_{XX} and Ω^*_{XX}, respectively. Then for ϖ^* in (4.1) such that $\Omega^*_{XX} \succ 0$, we have*

(i) $\lambda^*_{\max} \leq \lambda_{\max}$;

(ii) $\lambda^*_{\min} \geq \lambda_{\min}$;

(iii) Define $s(A) \equiv \max_{i\neq j} |\lambda_i - \lambda_j|$ as the spread of a squared matrix A. Then $s(\Omega^*_{XX}) \leq s(\Omega_{XX})$; and

(iv) Let $\mathcal{I} = \{1, ..., k\}$ be the index set corresponding the increasingly-ordered eigenvalues. Then for ϖ^* in (4.1) such that $\Omega^*_{XX} \succ 0$, there exists a proper index set $\mathcal{J} = \{1, ..., j\} \subset \mathcal{I}$ such that $\lambda^*_j \geq \lambda_j$, for $j \in \mathcal{J}$ and $\lambda^*_{j'} \leq \lambda_{j'}$, for $j' \in \mathcal{I}\backslash\mathcal{J}$.

Proof. To prove (i), one begin with the difference $\lambda^*_{\min} - \lambda_{\min}$ and use Cauchy's interlacing theorem (Horn and Johnson, 2012) to see that the difference is indeed positive. Similar argument applies to prove (ii). As for (iii), simply use (i) and (ii). The last claim is proven by the fact that $\mathbf{tr}\Omega_{XX} = \mathbf{tr}\Omega^*_{XX}$. □

The eigenvalue perturbation theorem possesses several important implications. First of all,

traditional matrix perturbation theory gives two-side bounds on the associated eigenvalues (Stewart and Sun, 1990). Albeit important, the information on how a specific eigenvalue moves can sometimes be inconclusive. Theorem 4.1 tells the directions in which the largest and smallest eigenvalues move.

Second, from (i)–(iii) of Theorem 4.1, one can immediately connect to the CN of the perturbed matrix Ω^*_{XX}. Specifically, the smaller the spread of a matrix is, the smaller the CN thereof will be, which leads to the following results on the reduction in CN.

Theorem 4.2 (Reduction in condition number)**.** *For ϖ^* in (4.1) such that $\Omega^*_{XX} \succ 0$ holds, the following hold:*

(i) The perturbed CN becomes smaller, viz.,

$$\kappa(\Omega^*_{XX}) \leq \kappa(\Omega_{XX}). \tag{4.2}$$

(ii) Let v_r be the reduction in the largest VIF. Then for ϖ^* in (4.1) such that $\Omega_{XX}+\mathcal{W} \succ 0$ holds, we have

$$\kappa(\Omega^*_{XX}) = \kappa(\Omega_{XX}) + \mathcal{O}(v_r). \tag{4.3}$$

To prove Theorem 4.2, we need the following lemma.

Lemma 4.1. *Given any $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$. If $\mathbf{x}$ and $\mathbf{y}$ lie in the same quadrant, then for any matrix $\mathbf{Q} \succ 0$ in $\mathbb{R}^{n\times n}$, we have $\mathbf{x}^T\mathbf{Q}\mathbf{y} > 0$.*

Proof. It follows that

$$\mathbf{x}^T\mathbf{Q}\mathbf{y} = \mathbf{tr}\mathbf{Q}\mathbf{x}\mathbf{y}^T = \mathbf{tr}\mathbf{Q}\mathbf{R} > \lambda_{\min}(\mathbf{Q})\mathbf{tr}\mathbf{R} > 0, \tag{4.4}$$

where the first inequality follows from Fang et al. (1994) and the last inequality holds because the diagonals of $\mathbf{R} \equiv \mathbf{x}\mathbf{y}^T$ are positive and $\lambda_{\min}(\mathbf{Q}) > 0$ due to positive definiteness. □

Proof of Theorem 4.2. The first claim is an immediate corollary of (i) and (ii) in Theorem

4.1. To prove (ii), we start with the following first-order approximation for CN

$$\kappa(\Omega^*_{XX}) = \kappa(\Omega_{XX}) + \mathbf{c}^T \varpi, \tag{4.5}$$

where $\mathbf{c} \in \mathbb{R}^{k(k-1)/2}$ is the vector comprising derivatives $\partial\kappa(\Omega_{XX})/\partial\varpi_{i,j}$ for $i > j$. It is not hard to see that

$$\frac{\partial\kappa(\Omega_{XX})}{\partial\varpi_{i,j}} = \frac{2\lambda_{\min}\upsilon_{\max,i}\upsilon_{\max,j} - 2\lambda_{\max}\upsilon_{\min,i}\upsilon_{\min,j}}{\lambda^2_{\min}}, \text{ for } i > j, \tag{4.6}$$

where $\upsilon_{\max,i}$, for instance, is the i-th entry of the eigenvector associated with $\lambda_{\max}$ and the rest follow. Then, at the optimal perturbation ϖ^* (4.1), we have

$$\begin{aligned}\kappa(\Omega_{XX}) &= \kappa(\Omega^*_{XX}) - \mathbf{c}^T\varpi^* \\ &= \kappa(\Omega^*_{XX}) + \left(\frac{\mathbf{c}^T\left(B^TB+\rho I\right)^{-1}\mathbf{v}_{k^*}}{\|\mathbf{v}_{k^*}\|_{(B^TB+\rho I)^{-1}}}\right)\upsilon_r \\ &= \kappa(\Omega^*_{XX}) + \mathcal{O}(\upsilon_r).\end{aligned} \tag{4.7}$$

It remains to use Lemma B.1 to show the constant associated with υ_r is positive, which is equivalent to prove $\mathbf{c}^T\left(B^TB+\rho I\right)^{-1}\mathbf{v}_{k^*} > 0$. This completes the proof. □

A subtle geometric implication of Theorem 4.2 is when collinearity is present, Ω_{XX} is ill-conditioned and thus the shape of Ω_{XX} is very ellipsoidal. The shape of Ω_{XX} can be refined to be more round-shaped through the optimal perturbation $\mathcal{W}$.

As to *VIF*, in light of Chapter 2, generally the reduction in *VIF*s is not universally monotonic. Only those pathological *VIF*s are improved, while non-pathological ones fluctuate slightly around their original values. By pathological, we mean those *VIF*s with values greater than some critical value, dependent of υ_l and υ_r. By and large, the sum of all *VIF*s decreases. We summarize the observations in the next theorem.

Theorem 4.3 (Reduction in *VIF*s)**.** *For ϖ^* in (4.1) such that $\Omega^*_{XX} \succ 0$ holds, the following are true.*

(i) The sum of all perturbed VIFs is less than or equal to that of the original VIFs, viz.,

$$\mathbf{tr}\Omega_{XX}^{*-1} \leq \mathbf{tr}\Omega_{XX}^{-1}. \tag{4.8}$$

(ii) Let v_l be the target VIF for the largest VIF to achieve. Then if $\min_{1\leq l\leq k} \mathcal{V}_l(\Omega_{XX}) > v_l$ *for any $v_l > 1/\lambda_{\min}(\Omega_{XX}) + v_r$, we have*

$$\left(\Omega_{XX}^{*-1}\right)_{i,i} \leq (\Omega_{XX}^{-1})_{i,i}, \forall i = 1, ..., k. \tag{4.9}$$

Proof. To prove (i), use the fact that the trace of a matrix equals the sum of its eigenvalues and consider the difference $\mathbf{tr}\Omega_{XX}^{*-1} - \mathbf{tr}\Omega_{XX}^{-1}$, it follows that

$$\mathbf{tr}\Omega_{XX}^{*-1} - \mathbf{tr}\Omega_{XX}^{-1} = \sum_{i=1}^{n}\left(\frac{1}{\lambda_i^*} - \frac{1}{\lambda_i}\right) = \sum_{i=1}^{n}\left(\frac{\lambda_i - \lambda_i^*}{\lambda_i^*\lambda_i}\right) \leq \frac{\sum_{i=1}^{n}(\lambda_i - \lambda_i^*)}{\lambda_{\min}^*\lambda_{\min}} = 0 \tag{4.10}$$

The first inequality holds because $\lambda_{\min} \leq \lambda_j$, $\forall j$ and the same for $\lambda_{\min}^*$. The last equality holds for $\mathbf{tr}\Omega_{XX}^* = \mathbf{tr}\Omega_{XX}$.

To prove (ii), given $\min_{1\leq i\leq k} \mathcal{V}_i(\Omega_{XX}) > v_l$, we have $\mathcal{V}_j(\Omega_{XX}) > v_l$, $\forall j = 1, ..., k$. For the case $j = k^*$, this is trivial. Consider any $j \neq k^*$, we see that

$$\left(\Omega_{XX}^{*-1}\right)_{j,j} - \left(\Omega_{XX}^{-1}\right)_{j,j} < \left(\Omega_{XX}^{*-1}\right)_{j,j} - v_l = \frac{\theta_{\min}^*\theta_2^*\cdots\theta_{k-1}^*}{\lambda_{\min}^*\lambda_2^*\cdots\lambda_{\max}^*} - v_l \leq \frac{1}{\lambda_{\min}^*} - v_l \leq 0. \tag{4.11}$$

by using Cauchy interlacing theorem (Horn and Johnson, 2012). Here θ_j^* is the j-th eigenvalue of the submatrix $\Omega_{XX}\{k^*\}$, $j = 1, \ldots, k-1$, and $v_l > \frac{1}{\lambda_{\min}^*} + v_r$ for some $v_r > 0$. The completes the proof. □

4.3 Main Results - Perturbed HCEs

Consider the HCE that has the following expression

$$HC_0 = \sigma^2(X^TX)^{-1}X^TEX(X^TX)^{-1}, \tag{4.12}$$

where $E \equiv Diag\,[\widehat{e}_i^2]$ for which $\widehat{e}_i^2$ is the i-th squared OLS residual. Following the same maneuver in Chapter 2 and given X centerd and scaled, (4.12) is indeed a function of correlation matrix Ω_{XX}, *viz.*,

$$HC_0(\Omega_{XX}) = \sigma^2 \Omega_{XX}^{-1} X^T E X \Omega_{XX}^{-1} \equiv \Omega_{XX}^{-1} \widehat{E} \Omega_{XX}^{-1}, \tag{4.13}$$

for $\widehat{E} \equiv X^T E X$. Note that, under homoskedasticity, we have $E = \sigma^2 I$ so that $HC_0(\Omega_{XX}) = \sigma^2 \Omega_{XX}^{-1}$ is the original OLS covariance matrix.

Remarkably, the residual matrix E is effectively a function of Ω_{XX} but has been implicitly ignored in (4.13). This is because the changes in each element of E is of order $\mathcal{O}(\sup_{1\le j\le k(k-1)/2} \varpi_j)$ and the total change in X is of order $\mathcal{O}(\sup_{1\le i\le n} \sup_{1\le j\le k(k-1)/2} \varpi_j^i)$. Numerically, such changes are negligible in general.

The perturbed version of (4.12) is thus given by

$$HC_0(\Omega_{XX}^*) = \sigma^2 \Omega_{XX}^{*-1} \widehat{E} \Omega_{XX}^{*-1} = (\Omega_{XX} + \mathcal{W})^{-1} \widehat{E} (\Omega_{XX} + \mathcal{W})^{-1}, \tag{4.14}$$

which we shall call *perturbed heteroskedasticity-robust covariance estimate* (PHCE). The main result for PHCE says monotonic improvements in the variance portions can be achieved under certain condition.

Theorem 4.4 (Reduction in Heteroskedasticity-Robust Variances)**.** *If* $\min_{1\le j\le k} \mathcal{V}_j > v_l$, *for any* $v_l > 1/\lambda_{\min}(\Omega_{XX}) + v_r$, *then for* ϖ^* *in (4.1) such that* $\Omega_{XX} + \mathcal{W} \succ 0$ *holds, we have*

$$(HC_0(\Omega_{XX}^*))_{i,i} \le (HC_0(\Omega_{XX}))_{i,i}, \ \textit{for } i = 1, \ldots, k. \tag{4.15}$$

Note that the condition "$\min_{1\le j\le k} \mathcal{V}_j > v_l$, for any $v_l > 1/\lambda_{\min}(\Omega_{XX}) + v_r$" in the above theorem is purely theoretical. A practical condition is

$$\min_{1\le j\le k} \mathcal{V}_j > v_l, \quad \text{for any} \quad v_l > \epsilon + v_r \quad \text{for any} \quad \epsilon > 0. \tag{4.16}$$

To prove Theorem 4.4, we need the following lemmas. The first lemma concerns the determinant of Ω_{XX}^*, while the other specifies the changes in the VIF matrix Ω_{XX}^{-1}.

Lemma 4.2 (Changes in determinant). *If* $\min_{1\le j\le k}\mathcal{V}_j > v_l$, *for any* $v_l > 1/\lambda_{\min}(\Omega_{XX}) + v_r$, *then for* ϖ^* *in (4.1) such that* $\Omega_{XX}+\mathcal{W}\succ 0$ *holds, we have* $\det(\Omega^*_{XX}) \ge \det(\Omega_{XX})$.

Lemma 4.3 (Changes in VIFs). *If* $\min_{1\le j\le k}\mathcal{V}_j > v_l$, *for any* $v_l > 1/\lambda_{\min}(\Omega_{XX}) + v_r$, *then for* ϖ^* *in (4.1) such that* $\Omega_{XX}+\mathcal{W}\succ 0$ *holds, we have*

(i) $\mathcal{V}_{i,j}(\Omega^*_{XX}) \le \mathcal{V}_{i,j}(\Omega_{XX})$, *if* $\mathcal{V}_{i,j}(\Omega_{XX}) > 0$; *and*

(ii) $\mathcal{V}_{i,j}(\Omega^*_{XX}) \ge \mathcal{V}_{i,j}(\Omega_{XX})$, *if* $\mathcal{V}_{i,j}(\Omega_{XX}) < 0$.

Proof. If $i = j$, then the proof is trivial. For any $i \ne j$, we consider two cases, *viz.*, *(i)* $\mathcal{V}_{i,j}(\Omega_{XX}) > 0$ and *(ii)* $\mathcal{V}_{i,j}(\Omega_{XX}) < 0$. For both cases, proofs are completed by starting with the difference $\mathcal{V}_{i,j}(\Omega^*_{XX}) - \mathcal{V}_{i,j}(\Omega_{XX})$ and use Lemma 4.2. □

Proof of Theorem 4.4. Consider the difference $\Delta(HC_0(\Omega_{XX}))_{i,i} \equiv (HC_0(\Omega^*_{XX}))_{i,i} - (HC_0(\Omega_{XX}))_{i,i}$:

$$\begin{aligned}\Delta(HC_0(\Omega_{XX}))_{i,i} &= \sum_{p=1}^{k}\mathcal{V}_{p,i}(\Omega^*_{XX})\sum_{q=1}^{k}\mathcal{V}_{i,q}(\Omega^*_{XX})\widehat{E}_{q,p} - \sum_{p=1}^{k}\mathcal{V}_{p,i}(\Omega_{XX})\sum_{q=1}^{k}\mathcal{V}_{i,q}(\Omega_{XX})\widehat{E}_{q,p}\\ &\le \sum_{p=1}^{k}\mathcal{V}_{p,i}(\Omega_{XX})\sum_{q=1}^{k}(\mathcal{V}_{i,q}(\Omega^*_{XX}) - \mathcal{V}_{i,q}(\Omega_{XX}))\widehat{E}_{q,p}\\ &\le 0,\end{aligned}$$

where the last two inequalities follow from Lemma 4.3. □

To end this section, we mention some variants of HCEs that can be used in the presence of outliers (Cribari-Neto et al., 2007).

Corollary 4.1 (Variants of HCE). *By replacing the weight matrix* E *in (4.12), several variants can be obtained. In particular, we have*

$$HC_m(\Omega^*_{XX}) = \sigma^2(\Omega_{XX}+\mathcal{W})^{-1}\widehat{E}_m(\Omega_{XX}+\mathcal{W})^{-1},\ m = 1,2,3 \tag{4.17}$$

where $\widehat{E}_1 = \frac{n}{n-r}E$, $\widehat{E}_2 = Diag\left[\frac{1}{1-h_i}\right]E$ *and* $\widehat{E}_3 = Diag\left[\frac{1}{(1-h_i)^2}\right]E$. *Here* h_i *is the i-th diagonal of the perturbed hat matrix* $\mathbf{H} \equiv X(X^TX+\mathcal{W})^{-1}X^T$.

4.4 Perturbed Test Statistics and Confidence Intervals

Having established the results above help us construct the perturbed versions of the test statistics. Perturbed test statistics will be used to construct the perturbed confidence intervals under specific level of confidence.

Perturbed t-Statistics and Interval Estimation We are interested in the following test on the estimates, *viz.*,

$$H_0 : \widehat{\beta}_j(\Omega^*_{XX}) = \beta_j \quad \text{against} \quad H_1 : \widehat{\beta}_j(\Omega^*_{XX}) \neq \beta_j, \quad \text{for some} \quad 1 \leq j \leq k. \tag{4.18}$$

The corresponding t-test statistics are therefore

$$\widehat{t}^m_j(\Omega^*_{XX}) = \frac{\widehat{\beta}_j(\Omega^*_{XX}) - \beta_j}{\sigma\sqrt{(HC_m(\Omega^*_{XX}))_{j,j}}}, m = 0, 1, 2, 3. \tag{4.19}$$

Accordingly, we can construct the $(1-\alpha)\%$ confidence interval for the perturbed regressors. In particular,

$$\widehat{\beta}_j \pm t_{1-\frac{\alpha}{2},n-tr(H)}\sigma\sqrt{(HC_m(\Omega^*_{XX}))_{j,j}},$$

for which $H \equiv X(X^TX + \mathcal{W})^{-1}X^T$ is the perturbed hat matrix.

Perturbed F-Statistics We are also interested in the test on some group of estimates, *viz.*,

$$H_0 : \widehat{\beta}_j(\Omega^*_{XX}) = 0 \text{ for } j \in \mathcal{J} \quad \text{against} \quad H_1 : \widehat{\beta}_j(\Omega^*_{XX}) \neq 0 \text{ for at least one } j \in \mathcal{J}, \tag{4.20}$$

wherein $\mathcal{J} \subseteq \mathcal{K} \equiv \{1, ..., k\}$ is an index set. The perturbed F-statistic has the following general form

$$\widehat{F}_m(\Omega^*_{XX}) \equiv \frac{1}{\sigma^2\,|\mathcal{J}|}\widehat{\beta}_{\mathcal{J}}(\Omega^*_{XX})^T(HC_m(\Omega^*_{XX})\{\mathcal{K}\backslash\mathcal{J}\})^{-1}\widehat{\beta}_{\mathcal{J}}(\Omega^*_{XX}), m = 0, 1, 2, 3, \tag{4.21}$$

where, for some matrix A, $A\{j\}$ denotes the submatrix derived from deleting the j-th row and column of A and $|\mathcal{A}|$ denotes the cardinality of some set $\mathcal{A}$.

4.5 An Example

We shall illustrate our established results with a practical example drawn from Bayhan and Bayhan (1998), to which the detailed descriptions are referred. Implementation is run using MATLAB R2012a under the Windows environment with 2.5GHz CPU and 4G RAM.

This example contains two covariates, X_1 and X_2, and a response vector $\mathbf{y}$. The correlation coefficient between X_1 and X_2 is 0.92, indicating strong collinearity relationship in between. Using the White's test under the null hypothesis that the error is homoskedastic, the resulting p-value is $1.3323e^{-14}$, suggesting that we reject the null hypothesis, and hence heteroskedastic.

From the information in the row associated with OLSE in Table 4.1, we see that the variance portions in the HCE are unsatisfactorily high, even though there has been no concensus on a rule of thumb for HCE that indicate the presence of stong collinearity. According to Theorem 4.4, monotonic improvements in the variance portions can be expect, using the practical bound (4.16).

We implement the iterative algorithm developed in Chapter 2 on the Bayhan's data. The parameter setting for the algorithm is $v_l = 4$ (target level for the largest VIF to achieve), $v_r = 2$ (reduction step in VIF) and $\rho = 50$ (penalty) and the OLSE is chosen to initialize the algorithm. Total number of iterations is 3 and it takes 0.387 CPU second on average (100 trials). Since the optimal perturbation is unique, so results are the same in each and every trial. Results are summarized in 4.1.

	V_1	V_2	HC_1	HC_2	$\lambda_{\min}$	$\lambda_{\max}$	CN
OLSE	6.507	6.507	87.119	85.11	0.08	1.92	23.985
POLSE	3.591	3.591	26.989	26.267	0.151	1.849	12.282

Table 4.1: Summary for Bayhan's data; V stands for VIF and HC for heteroskedasticity-robust variance.

From Table 4.1, we first see that both the VIFs (V_1 and V_2) and the heteroskedaticity-robust variances (HC_1 and HC_2) are enormously improved, which implies the new estimate POLSE is less influenced by the presence of data collinearity. Second, the improvements in the eigenvalues, *viz.*, minimal $\lambda_{\min}$ increases and maximal $\lambda_{\min}$ decreases, lead to the improvement of the CN, thereby healing the ill-conditionedness essentially, see Theorem 4.1 in Section 4.2.

Graphical evidences are given in Figure 4.1. The first two subfigures show the relationships between the heteroskedasticity-robust variances and the VIFs. Monotonic improvement is as expected, since the condition in Theorem 4.4 is fulfilled. The rest subfugures simply show the improvements in the eigenvalues lead to the improvement in the CN.

	$\widehat{\beta}_1$	$\widehat{\beta}_2$
OLSE	0.283	1.713
Int Est	(-14.801, 15.336)	(-13.256, 16.683)
POLSE	0.756	1.315
Int Est	(-7.64, 9.151)	(-6.967, 9.597)

Table 4.2: Estimators and their 1-S.D. interval estimation for Bayhan's data

Table 4.2 shows information of the estimates and their associated interval estimations under a 95% confidence level for the problem. It can be observed that, at the outset, the interval estimation has quite big ranges for both estimates, which simply results from the presence of data collinearity, which in turn inflates the corresponding standard deviations. After implementing the algorithm, the ranges of the new interval estimations are improved.

All the results suggest the success of the imposition of the optimal perturbation $\mathcal{W}$ on the correlation matrix X^TX in dealing with linear regression having both the heteroskedasticity and collinearity problems. Moreover, the computational efficiency adds more practical values to the proposed method.

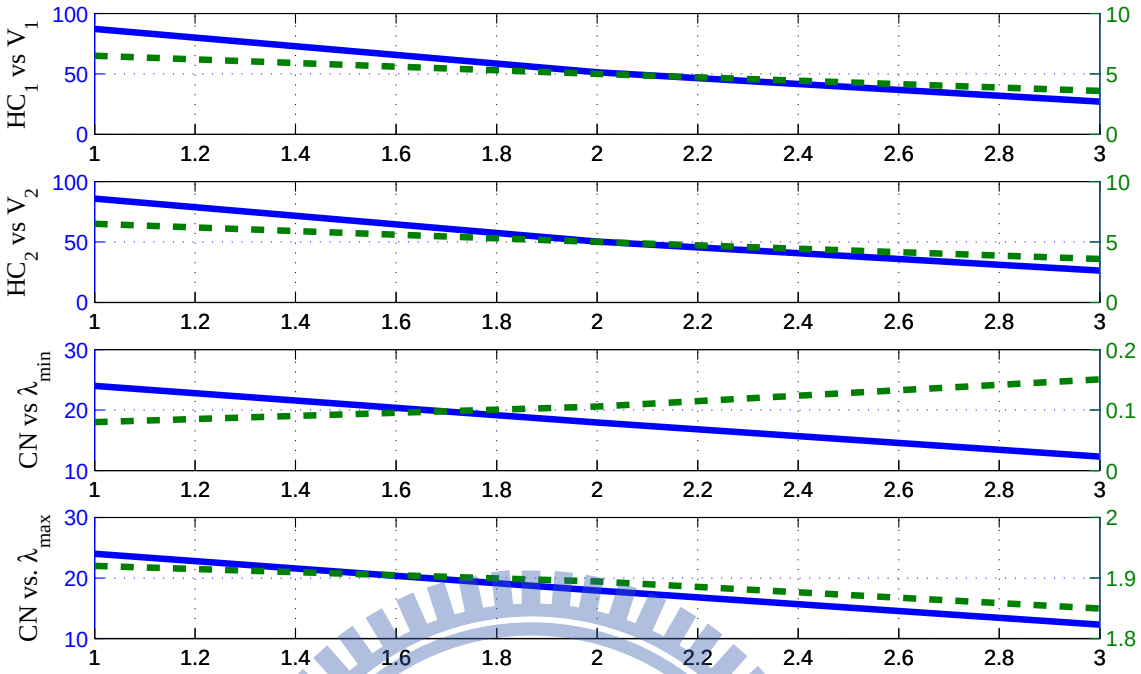

Figure 4.1: Plots for pairs of indexes; A vs. B with A corresponding to left y-axis (blue, solid line) and B to right y-axis (green, dashed line); x-axis denotes the iteration counts and only those integral values count.

4.6 Conclusions

We prove that the optimal perturbation matrix obtained in Chapter 2 can be directly applied to the HCE to improve the heteroskedasticity-robust variances, when heteroskedasticity is present. The application is almost effortless, because the optimal perturbation ϖ^* has an explicit expression and the additional complexity merely involves matrix addition. Unlike ridge-type estimates, one of the advantages of the LABs-type estimate is its capability in improving not only the VIFs but the CN in the meanwhile. However, no comparisons are made between our method and the ridge method, because each method involves different numbers of exogenous parameters. Lack of connection between the parameters renders the comparison somewhat meaningless. We will leave this for an open problem for the moment.

Chapter 5

A New Multicollinearity Diagnostic for Generalized Linear Models

We propose a new collinearity diagnostic tool for generalized linear models (GLMs). The new diagnostic tool is termed the weighted VIF (WVIF) behaving exactly the same as the traditional VIF in the context of regression diagnostic, given data matrix normalized. Compared to the use of CN (CN), WVIF shows more reliable information on how severe the situation is, when data collinearity does exist. An alternative estimate, a by-product of the new diagnostic, outperforms the ridge estimate in the presence of data collinearity in both aspects of WVIF and CN. Evidences are given through analyzing a real-world cancer data using logistic models.

5.1 Introduction

It has been well known that high linear dependency among covariates in any regression model, including not only traditional regression models but generalized linear models (GLMs), can pose frustrating statistical problems, such as unsatisfactorily inflated variances, lowered power in prediction and even incorrect signs, to the estimation results (Belsley et al., 2005). An evident example resides in the ordinary least squares (OLS) estimation. Strong linear dependecy among certain covariates renders the data matrix X rank deficient, and so is the matrix X^TX that plays an important role in the OLS

estimation.

Similar problems can occur in the GLMs. Although the estimation procedure is based on the maximum likelihood (ML) estimation, the ultimate estimate, the ML estimate (MLE), is oftentimes obtained through the iteratively re-weighted least squares (IWLS) algorithm, involving again the computation of matrix inverse (McCullough and Nelder, 1989). When the matrix X^TX is ill-conditioned, the weighted matrix X^TWX can be ill-conditioned as well, for some weight matrix W (Schaefer et al., 1984; Lee and Silvapulle, 1988). Then problems take place.

In the context of linear regression, the detection of the presence of data collinearity has been well studied. For instance, the correlation coefficients, the CN (CN) and the variance inflation factors (VIFs), all which are based upon the information from correlation matrix, have been thoroughly discussed (Belsley et al., 2005). Little work has been done for GLMs, however. In the work of Schaefer et al. (1984), the authors discoursed three aspects that serve as the criteria for collinearity diagnostic, inclusive of (i) the R^2 value when regressing one variable on the others; (ii) the sum of squared residuals in (i); and the minimal eigenvalue of the matrix X^TWX. It is intuitive to see that when collinearity is present, the R^2 tends to 1, the sum of squared residuals tends to 0, and the minimal eigenvalue tends to 0. A similar discussion can be found in Schaefer (1986). In conjunction with (iii) above, the criterion of looking at the minimal eigenvalue is equivalent to looking at the CN of a matrix. The authors (Lee and Silvapulle, 1988) suggested that CN be a formal diagnostic tool for the detection of collinearity and a value exceeding 10 (Belsley et al., 2005) can be used as a threshold to tell if collinearity is present. Similarly, in the work by Lesaffre and Marx (1993), the authors also treated CN as the main tool for collinearity diagnostic. In other work such as Weissfeld and Sereika (1991), the authors developed some collinearity diagnostics using the singular value decomposition (SVD) technique. For more discussions, see Mackinnon and Puterman (1989); Lesaffre and Marx (1993).

Biased estimation has been a widely accepted techique in dealing with non-orthogonal problem in regression. The most famous techniques in the clan are the ridge regression (Hoerl and Kennard, 1970), the LASSO (Tibshirani, 1996) and the principal component

regression (Hadi and Ling, 1998; Batah et al., 2009). As for GLMs, it dates back to the 70's or earlier that researchers began applying the ridge method to GLMs, in particular the logistic models, with collinearity problem and much research has been done on this topic. Schaefer et al. (1984); Schaefer (1986) proposed the ridge logistic estimate as a remedy for logistic regression having data collinearity problem. Follow-up research, such as Lee and Silvapulle (1988); Le Cessie and Van Houwelingen (1992), all disucssed the ridge logistic estimate in the same fashion. Application of LASSO to logistic models has also been studied, see Meier et al. (2008) and the references therein. Besides, similar idea applied to the Poisson regression has been developed as well Månsson and Shukur (2011).

As pointed out in Lazaridis (2007), treating CN as the only main collinearity diagnostic tool can at times be misleading by its disguised high value. Chances are that an extremely high value of CN does not necessarily lead to the conclusion of strong collinearity by only looking at its revised value. Investigations are sometimes needed for confirmation. Thus, CN should be used with caution. Further, given CN be the main diagnostic tool, the ridge method may not be of any help in dealing with the ill-conditioning problem. When dealing with collinearity problems in traditional regression models, the power of the ridge method resides in reducing the VIFs enormously, but not the CN nor the eigenvalues. Nevertheless, in the most recent work (García et al., 2014), the authors pointed out that the VIFs resulting from the ridge regression may pathologically violate the basic definition of a VIF, *viz.*, the range of a VIF is $[0, \infty)$. Hence, the combination of the CN and the ridge method may not be a good choice, as will be evident in Section 5.2. On the other hand, in Chapter 2 we proposed a new estimate, called the LABs estimate, for regression models having collinearity problem. The prominent properties of a LABs estimate resides in the ability of improving both the CN and the VIFs of the estimate, without breaking any basic defintions. As such, we believe using LABs-type estimates to combat collinearity problems in GLMs will be a better alternative.

In this study, it is our aim to propose a new collinearity diagnostic tool for GLMs, based on the perturbation method in Chapter 2. This new diagnostic will be termed the *weighted VIF* (WVIF). When data matrix X is standardized, WVIFs behave exactly

the same as the traditional VIFs, both indicating how inflated a variance is due to the presence of data collinearity. Since the new diagnostic is a weighted version of the VIF, the problem pointed out by García et al. (2014) is then avoided, *viz.*, the range of a WVIF is now $(0, \infty)$. An alternative estimate, with improved variances, can be obtained as a by-product. For ease of presentation, the analysis will be done through analyzing a single-class logistic model. The generalizations to multinomial logistic models and to more generic GLMs can be done with slight modifications.

5.1.1 Our Constributions

Our contributions are as follows. First, we introduce a new multicollinearity diagnostic for GLMs. This newly proposed diagnostic is termed the *weighted VIF* (WVIF) and functions exactly the same as the VIF in the context of the traditional regression diagnostics.

Second, whereas CN of the Fisher's information matrix has been considered the major tool for multicollinearity diagnostic in the literature, the problem pointed out by Lazaridis (2007) may lead to frustratingly misguiding results in telling the presence of data collinearity. On the contrary, the WVIF avoids such a problem and provides direct information on how inflated a variance is because of data collinearity.

Third, since a WVIF is the weighted version of a VIF, the range of a WVIF becomes $(0, \infty)$, in lieu of $[1, \infty)$ for a VIF, assuming that the R^2 value of regressing one variate on the others has the range $[0, 1]$. It is worth mentioning that ridge regression oftentimes yields fantastic results on VIFs, *viz.*, some VIFs are less than 1, impling negative R^2 value of regressing one variate on the others, see García et al. (2014). On the other hand, the perturbation method in Chapter 2 avoids such problems and can keep all definitions intact. Thus, applying the LABs method may be a better alternative.

Last but not least, based upon the WVIF and the LABs method, we propose a new estimate for GLMs under the influence of data collinearity. Whereas this new estimate is obtained through perturbing the original MLE, we shall name it the *perturbed MLE* (PMLE). The PMLE has improved variances and the resulting bias is minimal in an optimal sense. We further show some bounds on the measurements of goodness of fit.

5.1.2 Organizations

This is how the rest of the study is organized. We start with analyzing a practical example and examine the performance of the ridge method applied to the logistic regression, from the perspective of CN of the Fisher's information matrix in Section 5.2. Then, we propose the new diagnostic and study some basic properties of the resulting estimate in Section 5.3. The practical example will be revisited and discussed from the perspective of the proposed diagnostic in Section 5.4. Comparisons of the estimates are also made therein. Conclusions are given in Section 5.5.

5.2 An Example - Cancer Data

We shall begin with a practical example Lee (1974); Lesaffre and Marx (1993). This example contains 5 continuous variables (CELL - X_1, SMEAR - X_2, Infil - X_3, LI - X_4 and TEMP - X_5) and a response vector $\mathbf{y}$ taking on binary values. The goal is to build a logistic model to examine the relationship among the covariates. Details and data are referred to Lesaffre and Marx (1993).

Vars	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	CN
OLSE	-0.037	-0.312	0.357	0.248	-0.077	19.03
RMLE	-0.196	-1.594	1.811	1.304	-0.422	418.11
MLE	-0.196	-1.594	1.811	1.304	-0.422	418.13

Table 5.1: Summary of OLSE, RMLE and MLE for the cancer data

For purpose of illustration, we analyze the full model, with all variables included. In this illustration, data matrix X is standardized so that the linear predictor η constains no intercept, *viz.*, $\beta_0 = 0$. The OLS estimate (OLSE) is chosen to initialize the IWLS algorithm. The initial CN, evaluated at the OLSE, is 19.031, indicating the presence of mild collinearity among the covariates. Two maximum likelihood estimates (MLEs) are derived using the usual IWLS and the ridge method, denoted by MLE and ridge MLE (RMLE) Schaefer et al. (1984); Schaefer (1986); Lee and Silvapulle (1988); Le Cessie and Van Houwelingen (1992). The shrinkage parameter k for the ridge method is set to be 0.001. Results are summarized in Table 5.1.

We can first observe that the RMLE is essentially the same as the MLE, with a slight difference in the CNs which are unsatisfactorily high, suggesting the presence of strong collinearity. As a matter of fact, when applying the IWLS or the ridge method, the CN increases iteration by iteration, as shown in Figure 5.1. This reveals that, upon arriving at the MLE or RMLE, this estimate still has a serious ill-conditioning problem.

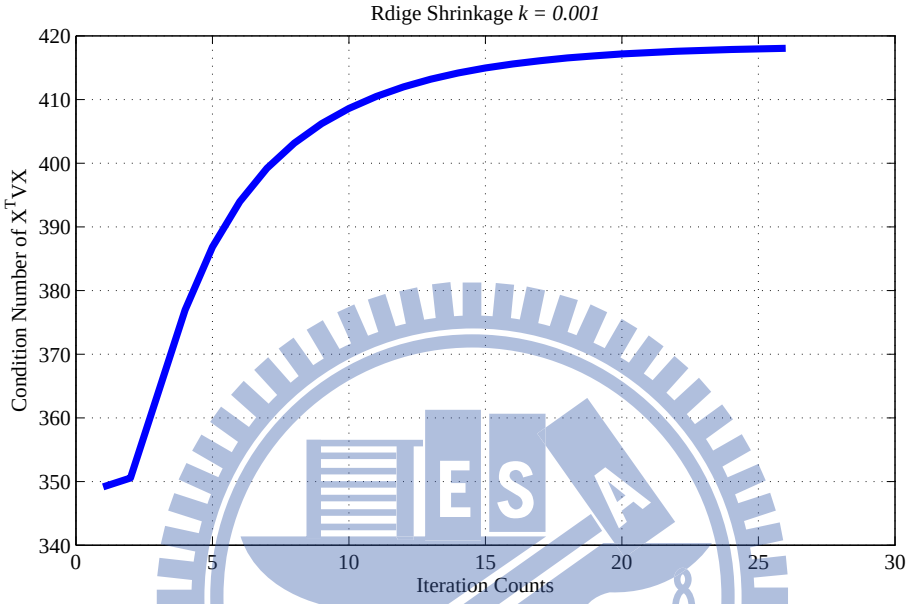

Figure 5.1: Increasing trend in CN using the ridge estimation in logistic regression

This is no surprise, because the ridge method reduces merely the VIFs, but not the CN. On the contrary, the LABs method Jou et al. (2014) is capable of improving both the VIFs and the CN. To provide evidences, we digress to revisit the education data from (see Chatterjee and Hadi, 2006, pp.222-228) (note this is an example for linear regression, not generalized linear models). The results are summarized in Table 5.2.

Vars	VIF_{β_1}	VIF_{β_2}	VIF_{β_3}	CN
OLS	38.44	31.48	88.37	393.95
RDG (k_1)	0.56	0.46	1.28	393.94
RDG (k_2)	0.42	0.35	0.87	266.32
LABs	2.66	2.85	3.84	14.39

Table 5.2: Comparison of VIFs for the education data Chatterjee and Hadi (2006); The ridge shrinkage parameters $k_1 = 10^{-5}$ and $k_2 = p - 1/\beta_{\text{OLS}}^T \beta_{\text{OLS}} = 0.249$; data are centered and scaled

From Table 5.2, the results yielded by the ridge method are either fantastic or unsatisfactory. First, a VIF with value less than 1 implies negative R^2 value. For instance,

in the choice of k_2, the R_j^2 values are -1.38, -1.86 and -0.15, respectively. This totally violates the basic definition of a VIF obviously. For a similar discussion, see García et al. (2014). Moreover, the CNs in the cases of either k_1 or k_2 are worse than what LABs method produces. All these evidences indicate that the LABs method outperforms the ridge method and can be a better alternative in dealing with collinearity problems.

Returning to the cancer data, it is known that the ridge method applied to the IWLS for logistic regression is done through the following recursion

$$\widehat{\beta}_{l+1} = \widehat{\beta}_l + (X^T \widehat{W}_l X + kI)^{-1} X^T (\mathbf{y} - \widehat{\pi}_l), \tag{5.1}$$

at the $(l+1)$-st iteration. The algorithm terminates once the difference between the successive estimates differs negligibly. From Figure 5.1, the increasing trend in CN implies that $(X^T \widehat{W}_l X + kI)$ gets more and more ill-conditioned, indicating strong ill-conditioning problem at the RMLE, which in turn is likely to cause the variances inflated unsatisfactorily.

5.3 Weighted VIF - A New Diagonstic

Let us consider the following generalized linear model for logistic regression

$$\mathbf{y} = \pi + \epsilon, \tag{5.2}$$

where $\mathbf{y} \in \mathbb{R}^{n\times 1}$ is the random component vector taking binary values; π is the systematic component vector, defined by

$$\pi \equiv \mathbb{E}[\mathbf{y}] = (1 + \exp\{-X^T \widehat{\beta}\})^{-1} \in \mathbb{R}^{n\times 1}, \tag{5.3}$$

with X collinear and normalized; and ϵ is the random error vector with mean zero and variance $Var(\epsilon_i) = \pi_i(1-\pi_i)$, for $i = 1, \ldots, n$, with $W \equiv Diag[\pi_i(1-\pi_i)]$. In this study, we shall assume that each component of $\mathbf{y}$ is i.i.d. and follows some distribution in the exponential family. In our case, y_i follows a bernoulli distribution, or equivalently, $\mathbf{y}$

follows a binomial distribution.

Let $\widehat{\beta}_{\text{MLE}}$ be the MLE for (5.2). Asymptotically, we have

$$\widehat{\beta}_{\text{MLE}} \sim \mathcal{N}(\beta, (X^T\widehat{W}_\infty X)^{-1}\phi), \tag{5.4}$$

viz., $\widehat{\beta}_{\text{MLE}}$ follows a multivariate normal distribution with mean β and covariance $(X^T\widehat{W}_\infty X)^{-1}\phi$. When X has rank-deficiency problem, the matrix $X^T\widehat{W}_\infty X$ can be ill-conditioned, thereby inflating the variances of the MLE Lee and Silvapulle (1988); Lesaffre and Marx (1993).

Now, take $\widehat{X} \equiv \widehat{W}_\infty^{1/2} X$, with X collinear and normalized. The asymptotic variance is therefore $(X^T\widehat{W}_\infty X)^{-1}\phi = (\widehat{X}^T\widehat{X})^{-1}\phi$. Under this setting, the situation becomes similar to that in Chapter 2. So, define

$$\widehat{\Omega}_{XX} \equiv \widehat{X}^T\widehat{X}, \tag{5.5}$$

which will be termed the *weighted correlation matrix*, even though it is not, among the covariates. On the main diagonals of the matrix $\widehat{\Omega}_{XX}^{-1}$ lie the *weighted VIFs* (WVIFs) associated with the covariates. The WVIF has a similar function as the conventional VIF does, *viz.*, their values both represent how inflated the variances are due to the presence of data collinearity. Moreover, the use of CN as the diagnostic for detection the presence of data collinearity can be misleading Lazaridis (2007). Therefore, we propose the WVIF as a new diagnostic tool, as it provides direct information on the severity of data collinearity. Hence, one can tell the presence of data collinearity by simply looking at the values of the WVIFs associated, provided there are more than one WVIF having value greater than a threshold, say, 6 or more Belsley et al. (2005), although there has been no rules of thumb as yet.

5.3.1 An Alternative Estimate

Assume that MLE $\widehat{\beta}_{\text{MLE}}$ exists for (5.2) and is a limit point of a convergent sequence, *viz.*,

$$\widehat{\beta}_{\text{MLE}} = (X^T\widehat{W}_\infty X)^{-1}X^T\widehat{W}_\infty \mathbf{Z} = \widehat{\Omega}_{XX}^{-1}\widehat{\Omega}_{X\mathbf{z}}, \tag{5.6}$$

where $\widehat{\Omega}_{X\mathbf{z}} \equiv X^T\widehat{W}_\infty\mathbf{Z}$. The variances of the MLE $\widehat{\beta}_{\text{MLE}}$ may still be inflated due to the presence of data collinearity. Our strategy here is to find a perturbation matrix $\mathcal{W}$ so that the pathological WVIFs can be improved. The resulting estimate then takes the form

$$\widehat{\beta}_{\text{MLE}}(\mathcal{W}) = \widehat{\Omega}_{XX}^{*-1}\widehat{\Omega}_{X\mathbf{z}}, \tag{5.7}$$

for which $\widehat{\Omega}_{XX}^* \equiv X^T\widehat{W}_\infty X + \mathcal{W} \succ 0$. We shall call this new estimate the perturbed MLE (PMLE), which will be obtained through an iterative algorithm as given in Chapter 2.

It has been shown that the optimal perturbation $\varpi^* = [\Delta\varpi_{i,j}]_{i>j} \in \mathbb{R}^{k\times k(k-1)/2}$, where k is the number of covariates, has an explicit expression. The following expression serves as a counterpart in a WVIF case

$$\varpi^* = \frac{-v_r}{\|\widehat{\mathbf{v}}_{k^*}\|^2_{\left(\widehat{B}^T\widehat{B}+\rho I\right)^{-1}}}\left(\widehat{B}^T\widehat{B}+\rho I\right)^{-1}\widehat{\mathbf{v}}_{k^*}, \tag{5.8}$$

where v_r is the reduction step in the (largest) WVIF; $\widehat{\mathbf{v}}_{k^*}$ is the vector made of the derivative $\partial(X^T\widehat{W}_\infty X)^{-1}_{k^*,k^*}/\partial\varpi_{i,j}$, for $i > j$; k^* corresponds to the index of the largest WVIF; $\widehat{B}$ is a matrix made of the derivative $\partial\widehat{\beta}_{\text{MLE}}/\partial\varpi_{i,j}$, for $i > j$; and $\rho(> 0)$ is a penalty parameter. The optimal perturbation matrix $\mathcal{W}$ is then formed by the relationship $\mathcal{W} = [\Delta\varpi_{i,j}]_{i\neq j}$, with $\mathcal{W}_{i,i} = 0$, for all $i = 1, \ldots, k$.

5.3.2 Performance of PMLEs

We study the performance of the proposed estimate in what follows. Results are summarized as theorems without proofs.

The very first concern resides in the question if $\widehat{\Omega}_{XX}^* \equiv \widehat{\Omega}_{XX} + \mathcal{W} \succ 0$ holds. The perturbation matrix $\mathcal{W}$ is generally indefinite, so there is no guarantee the positive definiteness of $\widehat{\Omega}_{XX}^*$ will hold. Nevertheless, from matrix theory, a necesary and sufficient condition for $\widehat{\Omega}_{XX}^* \succ 0$ to hold is to check if $\lambda_{\min}(\widehat{\Omega}_{XX}^{-1}\mathcal{W}) > -1$. For a proof, refer to Horn and Johnson (2012).

The perturbation method takes effect by changing the eigen-strucutre of the matrix $\widehat{\Omega}_{XX}$. The minimal eigenvalue of a matrix is oftentimes an crucial criterion to tell if a

matrix is ill-conditioned Schaefer et al. (1984); Lesaffre and Marx (1993). It has been shown that imposing an optimal perturbation matrix comprising (5.8) can improve the minimal eigenvalue essentially. The following result states how the minimal eigenvalue of the matrix $\widehat{\Omega}^*_{XX}$ changes.

Theorem 5.1 (Minimal Eigenvalue). *Let* $0 < \lambda_{\min} \leq \lambda_2 \leq \cdots \leq \lambda_{\max}$ *and* $0 < \lambda^*_{\min} \leq \lambda^*_2 \leq \cdots \leq \lambda^*_{\max}$ *be the increasingly-ordered eigenvalues for* $\widehat{\Omega}_{XX}$ *and* $\widehat{\Omega}^*_{XX}$*, respectively. Then for* ϖ^* *in (5.8), we have* $\lambda^*_{\min} \geq \lambda_{\min}$.

As was evident in Table 5.2 in Section 5.2, the LABs method heals the ill-conditionedness, reflecting on the improvement in the CN through changing the eigen-structure of the matrix $\widehat{\Omega}_{XX}$. The next result .

Theorem 5.2 (Condition Number). *Let* κ *and* κ^* *be the CN associated with the matrices* $\widehat{\Omega}_{XX}$ *and* $\widehat{\Omega}^*_{XX}$*, repsectively. For the optimal perturbation* ϖ^* *given in (5.8), we have* $\kappa^* \leq \kappa$.

Similar to Chapter 2, the optimal perturbation (5.8) is obtained by constraining on the reduction in the largest WVIF(s). Normally, imposing reduction on the most pathological, or simply the largest, WVIF renders the reduction in other pathological ones, while those non-pathologicals fluctuate slightly around their original values. However, if all WVIFs are pathological, then a monotonic reduction can be observed. We have the following result.

Theorem 5.3 (WVIFs). *For* ϖ^* *given in (5.8), the following are true:*

(i) The sum of all perturbed WVIFs is less than or equal to that of the original WVIFs, viz.,

$$\mathbf{tr}\widehat{\Omega}^{*-1}_{XX} \leq \mathbf{tr}\widehat{\Omega}^{-1}_{X}. \tag{5.9}$$

(ii) Let v_l *be the target WVIF for the largest WVIF to achieve and* $v_r > 0$ *the reduction in the largest WVIF. Then if* $\min_{1\leq i\leq k} \mathcal{V}_i(\widehat{\Omega}_{XX}) > v_l(v_r)$ *for any* $v_l > \lambda^{-1}_{\min}(\widehat{\Omega}_{XX}) + v_r$*, we have*

$$\left(\widehat{\Omega}^{*-1}_{XX}\right)_{i,i} \leq \left(\widehat{\Omega}^{-1}_{XX}\right)_{i,i}, \forall i = 1, ..., k. \tag{5.10}$$

where $v_l(v_r)$ means the parameter v_l is dependent on the other parameter v_r.

Note here that the condition in (ii) of Theorem 6.3 is far too crude and theoretical. A practical bound for (ii) to hold is $\min_{1\le i\le k} \mathcal{V}_i(\widehat{\Omega}_{XX}) > v_l$ for any $v_l > v_r$. In other words, as long as all WVIFs are pathological, monotonic reduction is certain.

5.3.3 Goodness of Fit

When measuring the goodness of fit of the constructed model, one way is to measure the deviance and the other the generalized Pearson's χ^2 statistic, both following a χ^2 distribution with the same degree of freedom asymptotically. The perturbed versions of (5.11) and (5.13) can be obtained by replacing the MLE $\widehat{\beta}_{\text{MLE}}$ by the PMLE $\widehat{\beta}(\mathcal{W})$. We shall look into the measurements of fit in both aspects.

The deviance is defined as two times the difference between log-likelihood functions. In particular, for our case, the deviance has the following expression

$$\mathbf{D}(\mathbf{y};\widehat{\pi}) = -2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) + \log(1-\widehat{\pi}_i)\right\}. \tag{5.11}$$

Consider the difference $\Delta\mathbf{D} \equiv \mathbf{D}(\mathbf{y};\widehat{\pi}(\mathcal{W})) - \mathbf{D}(\mathbf{y};\widehat{\pi})$ and the result follows

Proposition 5.1. *Given the optimal perturbation matrix $\mathcal{W}$ comprising ϖ^*, define the bias $\mathbf{b} \equiv \widehat{\beta}(\mathcal{W}) - \widehat{\beta}_{MLE}$. Then $\Delta\mathbf{D}(\mathcal{W}) \ge -2\sum_{i=1}^{n}\exp\{-\mathbf{x}_i^T\beta(\mathcal{W})\}$.*

Proof. It is not hard to show that

$$\begin{aligned}
\Delta\mathbf{D}(\mathcal{W}) &\ge -2\sum_{i=1}^{n}(\log\widehat{\pi}_i(\mathcal{W}) - \log\widehat{\pi}_i) \\
&\ge -2\sum_{i=1}^{n}\left\{\exp\{-\mathbf{x}_i^T\widehat{\beta}(\mathcal{W})\} - \exp\{-\mathbf{x}_i^T\widehat{\beta}_{\text{MLE}}\}\right\} \\
&\ge -2\sum_{i=1}^{n}\exp\{-\mathbf{x}_i^T\beta(\mathcal{W})\}.
\end{aligned} \tag{5.12}$$

where we have used the facts $1 - 1/y \le \log y \le y - 1$ for any $y > 0$ Topsoe (2004), $e^a \ge 0$ and $e^a - 1 \le e^a, \forall a \in \mathbb{R}$. □

Another measurement of fit is the Pearson's χ^2-statistic defined through the following standardized residual

$$X^2 \equiv \sum_{i=1}^{n} \left(\frac{y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}} \right)^2 \sim \chi^2_{n-(k+1)}, \tag{5.13}$$

where $\widehat{\pi}_i = (1 + \exp\{-\mathbf{x}_i^T \widehat{\beta}\})^{-1}$, for $i = 1, \ldots, n$. Following the same maneuver, let the difference $\Delta X^2 \equiv X^2(\mathcal{W}) - X^2$. We have the following property

Proposition 5.2. *Given the optimal perturbation matrix $\mathcal{W}$ comprising ϖ^* and the bias* $\mathbf{b}$ *as defined previously. Then $\Delta X^2(\mathcal{W}) \leq \sum_{i=1}^{n} \exp\{-\mathbf{x}_i^T \widehat{\beta}(\mathcal{W})\}$.*

Proof. We see that

$$\begin{aligned} \Delta X^2(\mathcal{W}) &\leq \sum_{i=1}^{n} \frac{1-\widehat{\pi}_i(\mathcal{W})}{\widehat{\pi}_i(\mathcal{W}))} - \frac{\widehat{\pi}_i}{(1-\widehat{\pi}_i)} \\ &= \sum_{i=1}^{n} \frac{1}{\widehat{\pi}_i(\mathcal{W}))} - \frac{1}{(1-\widehat{\pi}_i)} \\ &\leq \sum_{i=1}^{n} \exp\{-\mathbf{x}_i^T \widehat{\beta}(\mathcal{W})\}. \end{aligned} \tag{5.14}$$

where we have used the facts $0 \leq y_i \leq 1$, for all $i = 1, ..., n$, and $e^a \geq 0, \forall a \in \mathbb{R}$. □

5.4 Cancer Data Revisited

Let us revisit the cancer data discussed in Section 5.2 and we add the WVIF perspective to the problem.

Table 5.3 summarizes the WVIF information, corresponding to OLSE, MLE, RMLE and PMLE for the cancer data. A few things can be observed in the table. First, at OLSE, high WVIFs suggest strong collinearity existing among the variables. However, since the OLSE serves simply as a starting point for the IWLS, there is no need to worry about the collinearity problem at this moment.

Second, the difference between the MLE and RMLE is negligible, meaning that the ridge method does not take much effect on improving the performance of the alternative estimate RMLE. Values of WVIFs for both estimates merely indicate the presence of mild collinearity among the variables. Yet, from the persepctive of CN, extremely high CN may

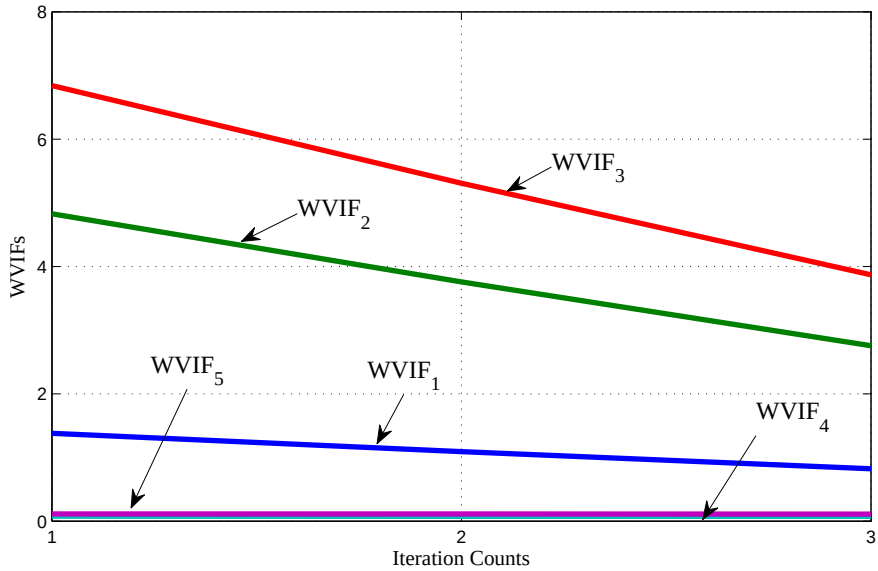

Figure 5.2: Plots for WVIFs resulting from the algorithm

Vars	$WVIF_1$	$WVIF_2$	$WVIF_3$	$WVIF_4$	$WVIF_5$	CN
OLSE	12.145	57.149	81.8	1.136	1.042	19.031
MLE	1.682	5.944	8.447	0.078	0.114	418.13
RMLE	1.684	5.945	8.449	0.079	0.115	418.11
PMLE	0.822	2.755	3.869	0.075	0.111	193.556

Table 5.3: Results for the centered and scaled cancer data; Note: (i) for OLSE, we have $V_0 = I$ and $\widehat{\pi}_i = 0$; (ii) parameters for obtaining the PMLE are $v_l = 4$, $v_r = 2$ and $\rho = 50$, see Chapter 2; (iii) total additional iterations are 3 and it takes 0.353 CPU seconds (average under 100 trials) for the algorithm to terminate

lead the researchers to the conclusion that severe ill-conditioning problem exists, provided the WVIF information is left behind. This observation is consistent with Lazaridis (2007).

Third, the PMLE outperforms both the MLE and RMLE in both the CN and the WVIFs. Even though, the CN is still high (≈ 194), but the WVIF provides more reliable information on the serverity of the data collinearity. Improvement in the WVIF means that the PMLE is less influenced by the presence of data collinearity and so PMLE has lowered variances. A graphical evidence is given in Figure 5.2. Furthermore, in Table 5.4, we see that the χ^2-statistic and the deviance of PMLE do not differ much from that of MLE or of RMLE, which means that the conclusion of goodness of fit for the MLE/RMLE applies to PMLE, under certain level of confidence.

To end this section, we look into another important criterion used to identify the pres-

Vars	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	χ^2	**D**
OLSE	-0.037	-0.312	0.357	0.248	-0.077	23.769	42.157
MLE	-0.196	-1.594	1.811	1.305	-0.421	71.942	29.106
RMLE	-0.196	-1.594	1.811	1.305	-0.421	71.942	29.106
PMLE	-0.023	-1.286	1.438	1.303	-0.417	72.308	29.311

Table 5.4: Summary of the estimates and the associated χ^2-statistics and the deviances

ence of data collinearity, *i.e.*, the minimal eigenvalue of the Fisher's information matrix $X^T\widehat{W}_\infty X$. When the effect of collinearity is strong, the minimal eigenvalue tends to 0, rendering the CN tend to infinity Schaefer et al. (1984); Schaefer (1986). In conjunction with the high value of CNs at MLE and RMLE, we can see that $\lambda_{\min}$ at either estimate remains close to zero, leading to high CN value and indicating the severity of data collinearity. On the other hand, improvement in $\lambda_{\min}$ is manifest at the PMLE, consistent with Theorem B.2. Changes in $\lambda_{\min}$ before reaching the PMLE are $(0.063, 0.078, 0.1004, 0.138)$, as shown in Figure 5.3.

Estimates	MLE	RMLE	PMLE
$\lambda_{\min}(X^T\widehat{W}_\infty X)$	0.063	0.063	0.138

Table 5.5: Minimal eigenvalue of $X^T\widehat{W}_\infty X$ at each estimate

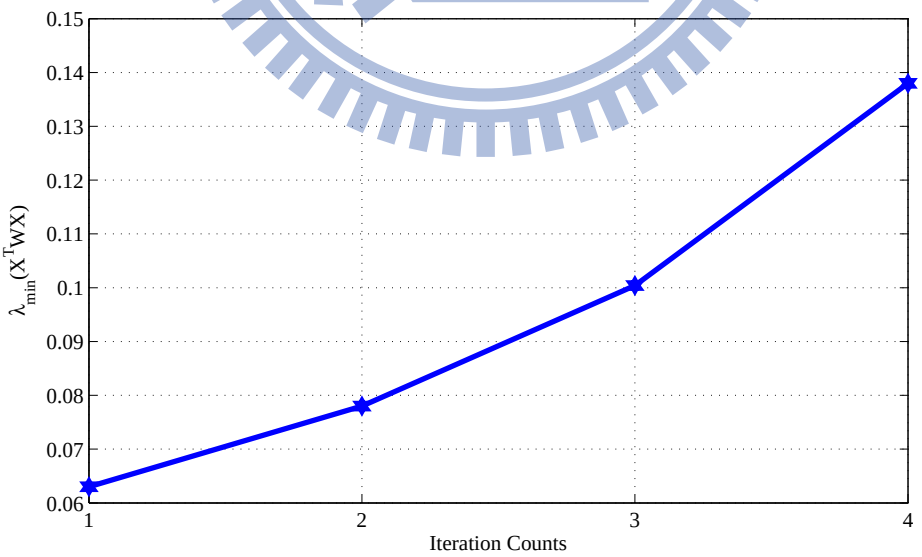

Figure 5.3: Changes in $\lambda_{\min}(X\widehat{W}_\infty X)$ resulting from the algorithm

5.5 Conclusions

In this study, we propose a new collinearity diagnostic tool for generalized linear models, illustrated by analyzing a logistic model. Similar to the VIF in the context of traditional regression models, the WVIF plays the same role in generalized linear models. The success of the application to a real-world problem manifests its practical value. The extension to multi-class binary-response models and the generalization to more generic generalized linear models should be immediate with certain modifications.

Chapter 6

Variance Reduction Method for Generalized Least Squares in the Presence of Multi-collinearity

In this study, we propose a variance reduction method for generalized least squares (GLS) in the presence of multicollinearity. It is well known that the asymptotic variance of a GLS estimate is $(X^TWX)^{-1}\sigma^2$, given X centered and scaled. The key idea of the variance reduction method then resides in reducing the diagonal elements of the weighted matrix $(X^TWX)^{-1}$ by imposing a perturbation matrix on it. The proposed method has a wide applicability in dealing with either regression models having heteroskedasticity, or time-series models with autoregressive serial correlation, or fixed (random) effect models for panel data, or the measurement-error models, or even the generalized linear models, all in the presence of multi-collinearity. We carry out real-world applications to validate the proposed method.

6.1 Introduction

Generalized least squares (GLS), first described by Aitken (1936), is employed when the covariance structure of the response vector is not scalar and is representive of some covariance matrix Σ. It is well known that GLS estimates (GLSEs) are unbiased, efficient,

consistent and asymptotically normal (Amemiya, 1985). The well known ordinary least squares (OLS) is special case of GLS when $\Sigma = I$. There are situations in which parameter estimations are done through the GLS. For instance, in linear regression, when the homoskedasticity is absent, the weighted least squares (WLS) is used. The WLS is another offspring in the GLS family and has a special covariance structure with zero off-diagonals in Σ. In time-series modeling, the existence of correlated error, or sometimes called the autoregressive correlated error, poses some problems on the usual OLS estimate (OLSE), which is no longer the best linear unbiased estimate (BLUE). Under this circumstance, GLS plays an crucial role in the estimation procedure. For generalized linear models (GLMs) (McCullough and Nelder, 1989), even though the estimation is done via the maximum likelihood (ML) estimation, the ML estimates (MLEs) are obtained through the so-called iteratively re-weighted least squares (IRWL) algorithm which is another variant of the GLS. All these examples have indicated the importance and wide applicability of the GLS.

In linear regression, the OLS estimation is problematic when the collected data exhibits collinear behaviour among the covariates. When the data matrix X has rank-deficiency problem, the inverse of the matrix $X^T X$ suffers from numerical instability, thereby rendering the resulting OLSE unsatisfactory. Consequences, such as high variances, weakened prediction power and even incorrect signs of some OLSEs against the hypothesis (Belsley et al., 2005), can be expected. Similar problems can occur in the GLS estimation. GLS estimation involves the inverse of the weighted matrix $X^T W X$, for some weight matrix W. When the matrix $X^T X$ is rank deficient, it is likely that $X^T W X$ suffers from ill-conditioning problem as well (Schaefer et al., 1984; Lee and Silvapulle, 1988).

To circumvent the problem, biased regression techniques have been applied to obtain alternative estimates. Among all biased regressions, the ridge regression (Hoerl and Kennard, 1970) has been the well accepted technique to do the job. To name a few, Schaefer et al. (1984); Schaefer (1986) proposed the ridge logistic estimate as a remedy for logistic regression having data collinearity problem. Follow-up research, such as Lee and Silvapulle (1988); Le Cessie and Van Houwelingen (1992), all disucssed the ridge logistic estimate

in the same fashion. Similar research on Poisson regression and application of LASSO to GLMs (Meier et al., 2008) can be found as well Månsson and Shukur (2011). Most recently, the authors propose another VIF-based estimate to solve the multicollinearity problem in both linear regression and GLMs. The central idea of the method consists in perturbing the off-diagonals of the correlation matrix, contrary to the ridge method. Interesting properties can be found in Chapters 2 and 3.

Comparing the perturbation method, the ridge method has several drawbacks. First, the VIFs resulting from the ridge method may not all be feasible (some VIFs may be less than 1), for a VIF is supposed to be greater than or equal to 1, by definition ($0 \leq R_j^2 \leq 1$ so that $1 \leq (1 - R_j^2)^{-1} < \infty$), see García et al. (2014). Our method preserves this property. Second, the search for the optimal shrinkage parameter in the ridge method is no easy task, which in general involves solving a non-convex optimization problem. Comparatively, our method merely involves solving a convex optimization problem at each stage and the total number of required iterations is finite. Further, it has been proven that the optimal perturbation exists and is unique. Third, the rationale of the perturbation method resides in the fact that multicollinearity stems from the close connection among different pairs of covariates, instead of that between a covariate and itself. Therefore, the perturbations on the off-diagonals make more sense than perturbing the diagonals in addressing multicollinearity problem.

In the traditional regression diagnostics, the detection of the presence of multicollinearity is often done through the correlation-related indexes, such as the correlation coefficients, condition number (CN) and variance inflation factor (VIF) (Belsley et al., 2005). On the other hand, for GLMs, CN of the Fisher's information matrix is by far still the major diagnostic tool. Compared to VIF, the other indexes, especially the CN (Lazaridis, 2007), may not clearly reflect how severe the situation is, because VIF has a direct connection to the variance of the estimate, while the others do not. Therefore, VIF has become a widely accepted index for detection of multicollinearity. In Chapter 5, a new diagnostic tool for generalized linear models, called the weighted VIF (WVIF), is proposed and behaves exactly the same as the traditional VIF in indicating how inflated a variance is

due to the presence of multicollinearity among covariates. In that work, the perturbation method is employed to solve the multicollinearity problem in GLMs and the resulting estimate has its variance enormously improved.

Our Contribution. The major contribution of this work is that we extend the idea of WVIF to the theory of GLS, as the new diagnostic for detecting the multicollinearity problems in various situations in which GLS is applied. The resulting family of estimates is termed the pertured GLS estimates (PGLSEs) that possess the property that the variances are greatly improved in the presence of multicollinearity. Further, since the increase in bias at each stage is minimized, the total bias upon obtaining a PGLSE is also minimized. The ill-conditionedness of the weighted matrix X^TWX can also be healed in the meanwhile.

Organization. This is how this study is organized. In Section 6.2, we show how to construct the PGLSEs by imposing an optimal perturbation vector ϖ^*, to be used to form the optimal perturbation matrix $\mathcal{W}$, on the weighted matrix X^TWX . In Section 6.4, we first introduce an iterative algorithm for obtaining the PGLSEs and then apply the method to analyzing two real-world examples, one with autoregressive serial correlation problem (to the first order) in time-series modeling in Section 6.4.2 while the other with multicollinearity problem in GLMs . Conclusions are given in Section 6.5.

6.2 Perturbed Generalized Least Squares

Consider a GLSE

$$\widehat{\beta}_{\text{GLS}}(\equiv \widehat{\beta}_g) = (X^TWX)^{-1}X^TW\mathbf{y}, \tag{6.1}$$

for X normalized and collinear, and for some weight matrix W. For regression models having heteroskedasticity, the weight matrix is a diagonal matrix, with the j-th diagonal being $h_j(X)^{-2}$ (or $\widehat{h}_j(X)^{-2}$) for heteroskedasticity; for time-series models having autoregressive serial correlation, assumed AR(1), the Prais-Winston's Prais and Winsten (1954) weight matrix has its (i,j)-th element as $(1-\widehat{\phi})^{-1}\widehat{\phi}^{|i-j|}$, where $\widehat{\phi} \equiv \frac{\sum_{i=1}^n \widehat{e}_t\widehat{e}_{t-1}}{\sum_{i=1}^n \widehat{e}_t^2}$ with $\widehat{e}_t = y_t - \widehat{\beta}_1 x_{t_1} - \cdots - \widehat{\beta}_k x_{t_k}$ for $t = 1, ..., n$; as to the logistic models, the weight matrix is

diagonal and has its j-th diagonal as $\pi_j(1-\pi_j)$, where $\pi_j = \mathbb{E}(y_j)$ for $j = 1, \ldots, n$. For more details, see Amemiya (1985).

In general, the asymptotic variance of $\widehat{\beta}_g$ can be written as $(X^TWX)^{-1}\sigma^2$. When X has rank deficiency problem, the matrix X^TX becomes ill-conditioned, which may very likely lead the matrix X^TWX to being ill-conditioned as well Schaefer et al. (1984); Lee and Silvapulle (1988). In this case, the variances tend to be inflated unsatisfactorily. The diagonals of the matrix $(X^TWX)^{-1}$ can then be thought of as a multiplier of the asymptotic variance for each component of $\widehat{\beta}_g$, which we shall term the *weight variance inflation factor* (WVIF). Obviously, A special case of WVIF is the traditional VIF in the regression diagnostic, when $W = I$ Belsley et al. (2005). Thence, a WVIF behaves exactly the same as a VIF, both indicating how inflated a variance is due to the presence of multicollinearity. If the inflated diagonals can be reduced, then the variances are improved.

In Chapter 2, the authors proposed a matrix perturbation method that helps improve the diagonal elements of the matrix $(X^TX)^{-1}$, *viz.*, the VIFs, in the OLS estimation. The proposed perturbation method involves solving a series of linearly-constrained convex quadratic bias-minimization problems, until the largest VIF is below certain level. The major finding in Chapter 2 consists in showing that by solving the following bias-minimizing optimization problem

$$\begin{aligned} \min \quad & \|B\varpi\|_2^2 + \rho\|\varpi\|_2^2 \\ \text{s.t.} \quad & -\mathbf{v}_{\max}^T\varpi = v_r \text{ and } \varpi \in \mathbb{R}^{k(k-1)/2} \end{aligned} \tag{6.2}$$

which has a unique optimal solution $\varpi^* = [\Delta\varpi_{i,j}]_{i>j}$

$$\varpi^* = \frac{-v_r}{\|\mathbf{v}_{k^*}\|^2_{(B^TB+\rho I)^{-1}}} \left(B^TB + \rho I\right)^{-1} \mathbf{v}_{k^*}. \tag{6.3}$$

where B is a matrix comprising the derivative $\partial\widehat{\beta}_{\text{OLS}}/\partial\varpi_{i,j}$, for $i > j$; $\mathbf{v}_{k^*}$ is a vector comprising the derivative $\partial(X^TX)^{-1}_{k^*,k^*}/\partial\varpi_{i,j}$, for $i > j$; k^* corresponds to the index of largest VIF; and $v_r > 0$ and $\rho > 0$ are given parameters, such that the largest VIF is reduced by an amount of at least v_r. Remarkably, $\mathcal{W}$ is formed by properly arranging

$\Delta\varpi_{i,j}$ into the positions corresponding to each (i,j)-pair, for all $i \neq j$. Therefore, $\mathcal{W}$ has zero diagonals ($\mathcal{W}_{i,i} = 0$) and is symmetric ($\mathcal{W}_{i,j} = \mathcal{W}_{j,i}$). The increase in bias is reprentative in terms of B and ϖ^*, *viz.*, $B\varpi^*$. The resulting new estimate then looks like

$$\widehat{\beta}_{\text{OLS}}(\mathcal{W}) = (X^T X + \mathcal{W})^{-1} X^T \mathbf{y} = \widehat{\beta}_{\text{OLS}} + B\varpi^*, \tag{6.4}$$

with $X^T X + \mathcal{W} \succ 0$ and with bias $B\varpi^*$. Since at each stage the increase in bias is minimized, the accumulative bias is also minimized upon obtaining (6.4). This estimate (6.4) is thus termed the *least-accumulative-bias* (LABs) estimate.

Following the same maneuver, we assume that there exists a matrix L such that $W = L^T L$ and take $\widehat{X} \equiv LX$ and $\widehat{\mathbf{y}} \equiv L\mathbf{y}$. Then (6.1) becomes

$$\widehat{\beta}_g = (\widehat{X}^T \widehat{X})^{-1} \widehat{X}^T \widehat{\mathbf{y}}. \tag{6.5}$$

We can immediately harness the results in Chapter 2 and obtain a new optimal perturbation $\widehat{\mathcal{W}}$ comprising the vector $\widehat{\varpi}^* = [\Delta\widehat{\varpi}_{i,j}]_{i>j}$

$$\widehat{\varpi}^* = \frac{-\upsilon_r}{\|\widehat{\mathbf{v}}_{k^*}\|^2_{\left(\widehat{B}^T\widehat{B}+\rho I\right)^{-1}}} \left(\widehat{B}^T\widehat{B} + \rho I\right)^{-1} \widehat{\mathbf{v}}_{k^*}. \tag{6.6}$$

where $\widehat{B}$ is a matrix comprising the derivative $\partial\widehat{\beta}_{\text{GLS}}/\partial\varpi_{i,j}$, for $i > j$; $\widehat{\mathbf{v}}_{k^*}$ is a vector comprising the derivative $\partial(X^T W X)^{-1}_{k^*,k^*}/\partial\varpi_{i,j}$, for $i > j$ and k^* corresponds to the index of largest WVIF, such that the largest WVIF is reduced by an amount of at least υ_r. Consequently, the perturbed GLSE (PGLSE) is given by

$$\widehat{\beta}_g(\widehat{\mathcal{W}}) = (X^T W X + \widehat{\mathcal{W}})^{-1} X^T W \mathbf{y} = (\widehat{X}^T \widehat{X} + \widehat{\mathcal{W}})^{-1} \widehat{X}^T \widehat{\mathbf{y}}. \tag{6.7}$$

for which $\widehat{X}^T \widehat{X} + \widehat{\mathcal{W}} \succ 0$ and with bias $\widehat{B}\widehat{\varpi}^*$.

6.3 Some Properties

In this section, we study the properties of the perturbation method for GLS estimations. For ease of presentation, we let $\widehat{\Omega}_{XX} \equiv X^T W X$ and $\widehat{\Omega}^*_{XX} \equiv X^T W X + \widehat{\mathcal{W}}$.

The question that if the addition of the perturbation matrix $\widehat{\mathcal{W}}$ to the GLSE $\widehat{\beta}_g$ still makes $\widehat{\Omega}^*_{XX} \equiv X^T W X + \widehat{\mathcal{W}}$ positive definite has been answered in Chapter 2.

The next result is the eigenvalue perturbation theorem for GLS. Notice that the improve in the largest eigenvalue is no longer true for GLS. It may be because of the structure of the weighted matrix.

Theorem 6.1 (Minimal Eigenvalue). *Let $0 < \lambda_{\min} \leq \lambda_2 \leq \cdots \leq \lambda_{\max}$ and $0 < \lambda^*_{\min} \leq \lambda^*_2 \leq \cdots \leq \lambda^*_{\max}$ be the increasingly-ordered eigenvalues for $\widehat{\Omega}_{XX}$ and $\widehat{\Omega}^*_{XX}$, respectively. Then for ϖ^* in (5.8), we have $\lambda^*_{\min} \geq \lambda_{\min}$.*

Even without the good results on the largest eigenvalue, we still have the following good property on the condition number.

Theorem 6.2 (Condition Number). *Let κ and κ^* be the CN associated with the matrices $\widehat{\Omega}_{XX}$ and $\widehat{\Omega}^*_{XX}$, repsectively. For the optimal perturbation ϖ^* given in (5.8), we have $\kappa^* \leq \kappa$.*

The next result can be treated as a variant of the VIF results in Chapter 4.

Theorem 6.3 (WVIFs). *For ϖ^* given in (5.8), the following are true:*

(i) The sum of all perturbed WVIFs is less than or equal to that of the original WVIFs, viz.,

$$\mathbf{tr}\widehat{\Omega}^{*-1}_{XX} \leq \mathbf{tr}\widehat{\Omega}^{-1}_{X}. \tag{6.8}$$

(ii) Let v_l be the target WVIF for the largest WVIF to achieve and $v_r > 0$ the reduction in the largest WVIF. Then if $\min_{1\leq i\leq k} \mathcal{V}_i(\widehat{\Omega}_{XX}) > v_l(v_r)$ for any $v_l > \lambda^{-1}_{\min}(\widehat{\Omega}_{XX}) + v_r$, we have

$$\left(\widehat{\Omega}^{*-1}_{XX}\right)_{i,i} \leq \left(\widehat{\Omega}^{-1}_{XX}\right)_{i,i}, \ \forall i = 1, ..., k. \tag{6.9}$$

where $v_l(v_r)$ means the parameter v_l is dependent on the other parameter v_r.

6.4 Iterative Algorithm and Examples

In the first part of this section, we first introduce an iterative algorithm for obtaining the PGLSE. The algorithm serves as a variant of that in Chapter 2 and has a similar structure, see Algorithm 4. The only difference resides in the input for the problem underlying. In the second part, we shall analyze two practical examples, one with serial correlation to the first order and the other a GLM having multicollinearity problem.

6.4.1 An Iterative Algorithm

Similar to the algorithm proposed in Chapter 2, the algorithm for obtaining PGLSEs requires input, given in the following sextuple

$$\mathcal{P} = (X, \mathbf{y}, W, v_r, v_l, \rho), \tag{6.10}$$

where X is the centered and scaled data matrix, $\mathbf{y}$ is the centered response vector, W is the weight matrix, v_r is the reduction step in the largest WVIF, denoted by $\widehat{\mathcal{V}}$, v_l is the target level for the largest WVIF to achieve, and ρ is a penalty parameter.

Algorithm 4 Iterative Algorithm for PGLSEs

Require: Problem input $\mathcal{P} = (X, \mathbf{y}, W, v_r = 2, v_l = 4, \rho = 50)$
Ensure: Optimal PGLSEs $\widehat{\beta}_{\text{GLS}}(\widehat{\mathcal{W}})$;

```
repeat
    Generate matrix B̂ for objective function and vector v̂_{k*} for constraint;
    Use (B.5) to form Ŵ = [Δω̂_{i,j}]_{i≠j};
    Update GLSE by (6.7);
until max_{j=1,...,k} V̂_i ≤ v_l
```

The generic structure of the algorithm is given in Algorithm 4. The algorithm is always initialized with the OLSE with at least one pathological WVIF, or equivalently variances. The algorithm accomplishes the task by reducing the largest WVIF at at least an amount of v_r at each iteration and terminates when the largest WVIF is less than or equal to the target level v_l. The requirement of the iterative setup consists in the numberical stability, because we are doing perturbation. Any attempt of one-step fashion will only crush the algorithm.

The setting of v_r depends on v_l. According to Belsley et al. (2005), an appropriate value for v_l is 6, depending on the cases Craney and Surles (2002); Obrien (2007). A basic criterion for the setting of v_r is $v_r < v_l$. An alternative strategy is an iterative setup, say, $v_r^{i+1} = 0.5v_l^i$ at $(i+1)$-st iteration. As to the penalty ρ, usually a value greater that or equal to 30 will suffice. The value of ρ has a minor effect on the result, since the norms of the increases in bias $\|\widehat{B}\widehat{\varpi}^*\|^2$ are approximately of order $\mathcal{O}(10^{-2})$, depending on W (whereas $\mathcal{O}(10^{-3})$ for the perturbation method in Chapter 2). In general, the value of the penalty is positively related to the norm of the increase in bias $\|\widehat{B}\widehat{\varpi}^*\|^2$, according to our experiences.

6.4.2 Example - Longley's Data

The first example is taken from Faraway (2002) and is known as the Longley's data Longley (1967). The problem contains two covariates, GNP and population, and one response variable, the number of people employed from 1947 to 1962. The errors are assumed to take an autoregressive form up to the first order and the model to be fitted is given by

$$\begin{aligned} \text{Employed}_t &= \beta_1 \text{GNP}_t + \beta_2 \text{Population}_t + \epsilon_t, \\ \epsilon_{t+1} &= \eta\epsilon_t + e_t, \\ e_t &\sim \mathcal{N}(0, \tau^2) \end{aligned} \tag{6.11}$$

It is not hard to see that the estimated coefficient for the autoregressive function is $\widehat{\eta} = 0.31$ and therefore the weight matrix W is formed by the rule $W_{i,j} = \widehat{\eta}^{|i-j|}$. To use the OLS estimation (6.5), one need to first apply the Cholesky decomposition to the weight matrix such that $W = L^T L$ and the obtain the transformed matrices $\widehat{X} \equiv LX$ and $\widehat{y} \equiv L\mathbf{y}$. The following table summarizes the information of fitting the model by GLS.

In Table 6.1, high values of both WVIFs indicate the presence of severe collinearity at the GLSE. The same conclusion can be made by looking at the CN. Applying the proposed method to the GLSE yields the results also in Table 6.1. Total number of iterations for the algorithm to terminate is 10 and the average CPU time is 0.42 second (100 trials).

We first see that the sign of coefficient associated Population_t changes from negative

Vars	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$WVIF_1$	$WVIF_2$	$\lambda_{\min}$	$\lambda_{\max}$	CN
GLSE	120440	-32880	20.246	19.463	0.025	8.518	337.235
PGLSE	5184	3565	3.46	3.326	0.15	8.3937	55.958

Table 6.1: Summary of PGLSE for centered and scaled Longley's data; $WVIF$'s are the diagonals of the matrix $(\widehat{X}^T\widehat{X})^{-1}$;$\lambda$'s are the eigenvalues and CN is the condition number of the matrix $\widehat{X}^T\widehat{X}$.

to positive, which may require more investigations to make certain that the sign-changing makes any sense. Nevertheless, our result indeed provides a different viewpoint on the relationship between the employment and the population. From Table 6.1, the WVIFs have been enormously improved and both eigenvalues are improved as well, the latter of which is consistent with Theorem 1 in Chapter 3. Figure 6.1 shows the monotonic improvement in the WVIFs, since both WVIFs are pathological and Figure 6.2 shows the relationships between the CN and the eigenvalues.

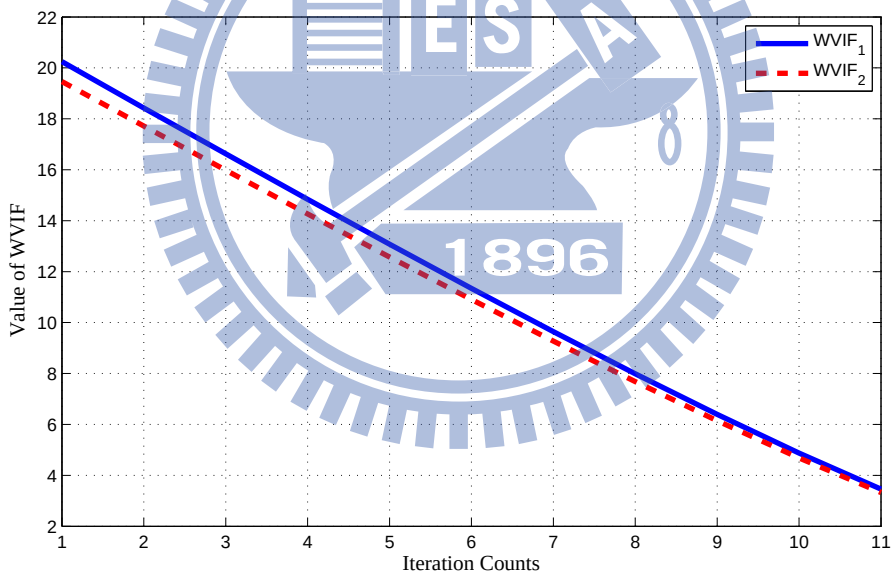

Figure 6.1: Plots of the WVIFs

6.4.3 Example - Cancer Data

The next example is taken from Lee (1974); Lesaffre and Marx (1993). This example contains 5 continuous variables (CELL - X_1, SMEAR - X_2, Infil - X_3, LI - X_4 and TEMP - X_5) and a response vector $\mathbf{y}$ taking on binary values. Details and data are referred to

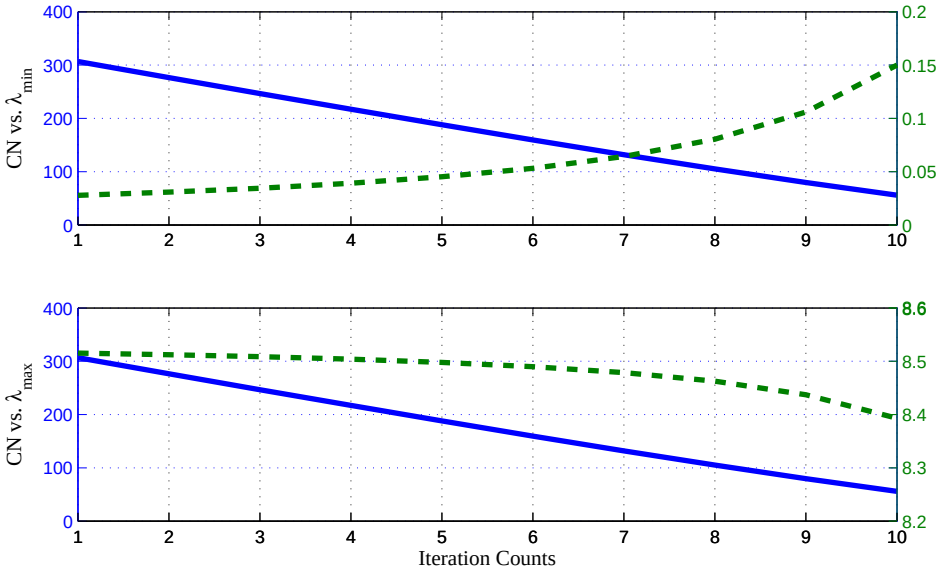

Figure 6.2: Plots of CN (blue line to left y-axis) versus $\lambda_{\min}$ and $\lambda_{\max}$ (dashed lines to right y-axis) of the matrix $X^T\widehat{W}X$

Lesaffre and Marx (1993).

Vars	$WVIF_1$	$WVIF_2$	$WVIF_3$	$WVIF_4$	$WVIF_5$	CN
OLSE	12.145	57.149	81.8	1.136	1.042	19.031
MLE	1.682	5.944	8.447	0.078	0.114	418.13
PMLE	0.822	2.755	3.869	0.075	0.111	193.556

Table 6.2: Results for the centered and scaled cancer data; Note: (i) for OLSE, we have $V_0 = I$ and $\widehat{\pi}_i = 0$; (ii) parameters for obtaining the PMLE are $v_l = 4$, $v_r = 2$ and $\rho = 50$, see Chapter 2; (iii) total additional iterations are 3 and it takes 0.353 CPU seconds (average under 100 trials) for the algorithm to terminate

In this illustration, data matrix X is normalized so that the linear predictor η constains no intercept, *viz.*, $\beta_0 = 0$. The iteratively re-weighted least squares (IRLS) algorithm is initialized with any arbitrary estimate and we choose the OLSE to begin with. Table 6.2 summarizes the WVIF information, corresponding to OLSE, MLE and PMLE for the cancer data. A few things can be observed in the table. First, at OLSE, high WVIFs suggest strong collinearity existing among the variables. However, since the OLSE serves simply as a starting point for the IRLS, there is no need to worry about the collinearity problem at this moment. However, values of WVIFs at the MLE merely indicate the presence of mild collinearity among the variables. Even so, we can still implement out algorithm to obtain a better estimate for this problem.

Another important criterion used to identify the presence of data collinearity is the minimal eigenvalue of $X^T\widehat{W}X$. When the effect of collinearity is strong, the minimal eigenvalue tends to 0, rendering the condition number tend to infinity Schaefer et al. (1984); Schaefer (1986). From Table 6.3, improvement in $\lambda_{\min}$ is manifest at the PMLE, again consistent with Theorem 1 in Chapter 5. Changes in $\lambda_{\min}$ before reaching the PMLE are $(0.063, 0.078, 0.1004, 0.138)$, iteration by iteration, as shown in Figure 6.3.

Estimates	MLE	PMLE
$\lambda_{\min}(X^T\widehat{W}X)$	0.063	0.138

Table 6.3: Minimal eigenvalue of $X^T\widehat{W}X$ at each estimate

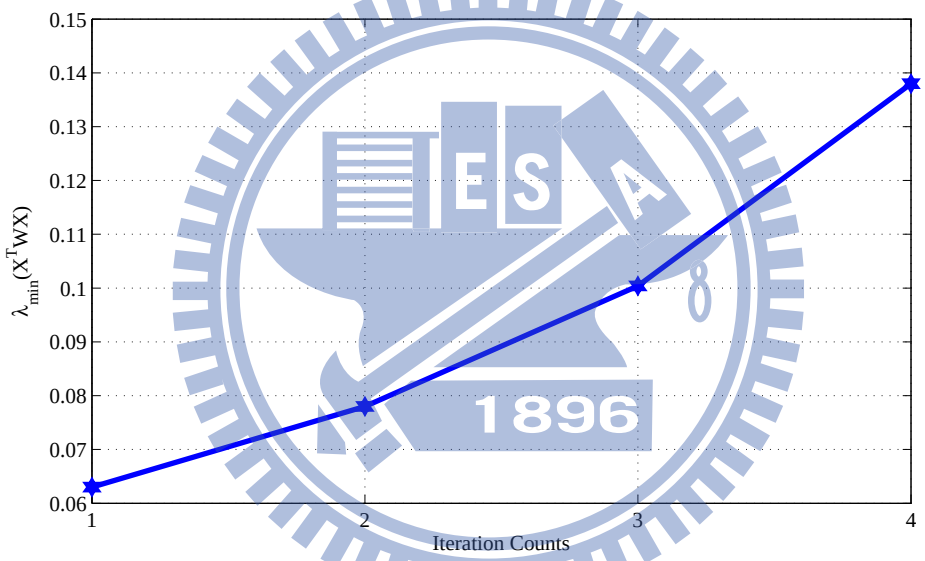

Figure 6.3: Changes in $\lambda_{\min}(X\widehat{W}_{\infty}X)$ resulting from the algorithm

6.5 Conclusions

In this study, we propose a variance reduction method based on the concept of WVIF, which behaves exactly the same as the VIF in the context of regression diagnostics, for the problems that are solved by GLS. The proposed method can be thought of as a generalization of the VIF-based method, when the weight matrix is identity I. Applications to pracitcal problems validates the effectiveness and make the proposed method of practical use.

Chapter 7

Conclusions and Future Research Directions

In this proposal, we have established a VIF-based optimization model, which is convex quadratic, to solve a series of collinearity problems in linear regression. First, we solve the prototype problem that is the collinearity problem in linear regression without departures from any of the full set of Gauss-Markov assumptions. Based upon the establishment, we construct the so-called *least-accumulative-bias* (LABs) estimates.

Second, we extend the LABs estimates to a wider class of linear models having collinearity problems. We study the situations when the departure from homoskedasticity is observed. When the form for the non-constant conditional variance is unknown, we prove that the optimal perturbation matrix generated by the VIF-based model can be directly applied to the associated HCEs to reduce the variance portions (under certain conditions). On the other hand, provided the form for heteroskedasticity is given or well estimated, we show that there exists another optimization model able to solve the collinearity problem.

Third, we propose a conceptual second-order VIF-based optimization model for linear models having collinearity problem. The second-order model serves as a mean to provide a preciser solution for solving the collinearity problem. Moreover, the second-order model can be applied whether or not the homoskedasticity holds. Since the second-order model

is nonlinear and nonconvex in essence, so simplify solving the problem entails the use of the more sophisticated technique, called linear conic programming (LCOP). Existing theory guarantees the global optimality for the perturbation matrix.

7.1 Future Research Directions

As of this moment, we have basically covered the collinearity problem in a linear model, mostly for cross-sectional data. In what follows, we shall discuss some doable extensions for our method.

7.1.1 Optimal Rank-2 Perturbation

Recall that our method produces a series of optimal perturbation matrix $\mathcal{W}$ which is typically full-rank and brings universal impact to the entire correlation matrix, except for the main diagonals, as is shown below.

$$\mathcal{W} = \begin{bmatrix} 0 & \varpi_{1,2} & \dots & \varpi_{1,k} \\ \varpi_{2,1} & 0 & \ddots & \varpi_{2,k} \\ \vdots & \ddots & 0 & \vdots \\ \varpi_{k,1} & \varpi_{k,2} & \dots & 0 \end{bmatrix}.$$

To improve the numerical performance, it is an intuitive idea to find an optimal rank-2 perturbation matrix, resulting from a sum of two rank-1 matrix and its transpose.

$$\widehat{\mathcal{W}} = \begin{bmatrix} 0 & 0 & \varpi_{i,j} & 0 \\ 0 & 0 & \ddots & 0 \\ \varpi_{j,i} & \ddots & 0 & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

It has been known that any rank-1 matrix M is equivalent to the product of two rank-1 vectors, *viz.*, $M = uv^T$, for some vectors u and v. In our case, we then wish to find u and v such that $\widehat{\mathcal{W}} = uv^T + vu^T$.

7.1.2 Nonconvex L^p Regression for $p \in (0, 1)$

Another interesting but aggressive breakthrough is the establishment of the theory of the so-called L^p-*regression* for $0 < p < 1$. Conventionally, people deal with the cases for any $p \geq 1$, building solid and beautiful theories. However, the case for any $0 < p < 1$ has been quite less studied ever since the regression analysis was in its infancy. The main reason resides in the non-convexity of the use of (quasi)-norm $\| \bullet \|_p^2$. Difficulties occur when trying to build even a simple point estimation for such a less-studied branch.

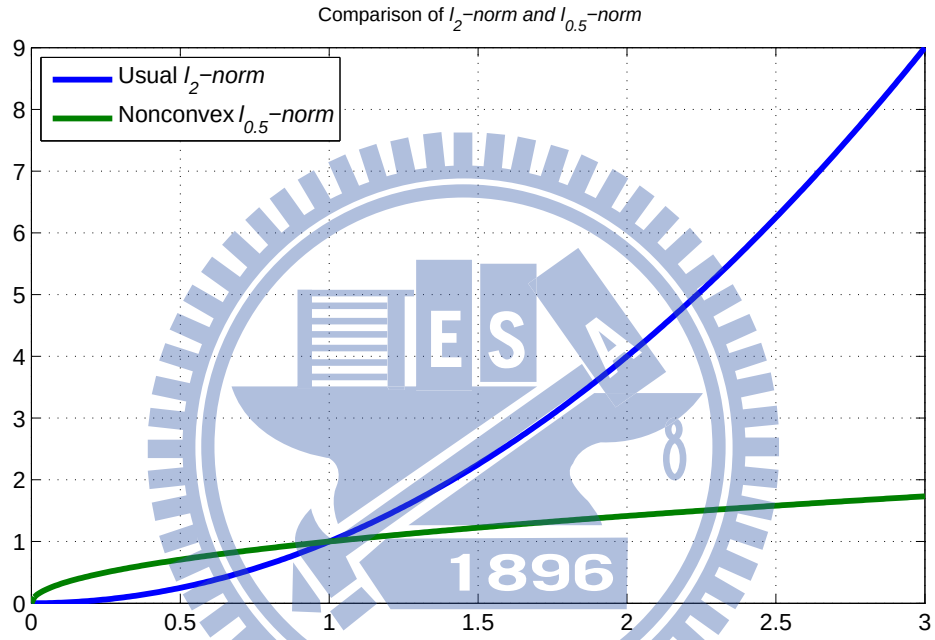

Figure 7.1: The comparison between l_2-norm (blue line) and $l_{0.5}$-norm (green line)

Unlike any usual L^p-norm for $p \geq 1$, an L^p-norm for $p \in (0, 1)$ endows small residuals more weight in the estimation process. For instance, when $u_i = 0.3$, we have $\|u_i\|_{0.5} = 0.548$, while $\|u_i\|_2 = 0.09$. Such a difference will lead to very distinct results for different choice of p. Therefore, we believe the study on the L^p-*regression* for $0 < p < 1$ will be so interesting that many surprising properties can be established to enrich the literature.

Appendices

Appendix A

Schinnar's CN-Based Model - Review

In this appendix, we shall review Schinnar's CN-based optimization model for solving collinearity problem in multiple linear regression in brief (Schinnar and Walters, 1983). The only difference between Schinnar's model and our model resides in the diagnostic used in formulating the models. Therefore, we shall merely review the difference.

A.1 Formulation using Condition Number

Let $\mathcal{C}(A)$ denote the condition number of some matrix $A \in \mathbb{R}^{k \times k}$. By definition, the condition number of the correlation matrix Ω_{XX} is defined by the following ratio of eigenvalues

$$\mathcal{C}(\Omega_{XX}) \equiv \frac{\lambda_{max}(\Omega_{XX})}{\lambda_{min}(\Omega_{XX})}, \tag{A.1}$$

where we assume that $\Omega_{XX} \succ 0$ and $\lambda_{max}(\Omega_{XX})$ ($\lambda_{min}(\Omega_{XX})$) is the largest (smallest) eigenvalue of the correlation matrix Ω_{XX}.

The goal here is to find a perturbation matrix $\mathcal{W}$ such that $\mathcal{C}(\Omega_{XX}+\mathcal{W})$ gets improved in comparison to $\mathcal{C}(\Omega_{XX})$. To do this, we apply the Taylor's expansion up to the first

order to $\mathcal{C}(\Omega_{XX}+\mathcal{W})$ and it follows that

$$\begin{aligned}\mathcal{C}(\Omega_{XX}+\mathcal{W}) &= \mathcal{C}(\Omega_{XX}) + \sum_{i>j} \mathcal{C}'(\Omega_{XX})\varpi_{i,j}, \\ &= \mathcal{C}(\Omega_{XX}) + \mathbf{c}^T\varpi,\end{aligned} \tag{A.2}$$

wherein the term $\mathbf{c}^T\varpi$ can be deemed the difference between the original condition number $\mathcal{C}(\Omega_{XX})$ and the new one $\mathcal{C}(\Omega_{XX}+\mathcal{W})$. Note that the derivative $\mathcal{C}'(\Omega_{XX})$ has the following representation

$$\mathcal{C}'(\Omega_{XX}) = \frac{\lambda'_{max}\lambda_{min} - \lambda'_{min}\lambda_{max}}{\lambda^2_{min}}, \tag{A.3}$$

where $\lambda'_{max} = 2\mu_{max,i}\mu_{max,j}$, the $i-$ and $j-$th element of the eigenvector μ_{max} and the rest follows similarly.

Since the presence of collinearity in the data matrix renders $\mathcal{C}(\Omega_{XX})$ unsatisfactorily high, we want $\mathcal{C}(\Omega_{XX}+\mathcal{W}) < \mathcal{C}(\Omega_{XX})$, or equivalently, $-\mathbf{c}^T\varpi > 0$. With this along with the same maneuver on the estimate $\widehat{\beta}(\Omega_{XX})$, one can formulate the following bias-minimization model

$$\min_{\varpi\in\mathbb{R}^{k(k-1)/2}} \{\varpi^T(B^TB+\rho I)\varpi : -\mathbf{c}^T\varpi = c_r\} \tag{A.4}$$

where $\rho(>0)$ is an exogenous trade-off parameter and $c_r(>0)$ is the reduction in the CN.

Appendix B

Application to Weighted Least Squares - A Special Case of Generalized Least Squares

In this appendix, we study the properties of the perturbation method applied to the weighted least squares estimates (WLSEs) under the influence of collinearity. All proofs will be given in Section B.2.

B.1 Properties

Before stating the results, the relationship among the theorems is as follows. Theorem B.1 serves as the fundamental of Theorem B.2. Theorem B.2, in turn, leads to Theorems B.3 and B.4. We may now state the results.

Because the perturbation matrix $\mathcal{W}$ is in general indefinite, there is no guarantee the matrix $\widehat{\Omega}^*_{XX}$ will be positive definite. A necessary condition for $\widehat{\Omega}^*_{XX} \succ 0$ to be true is to compare the smallest eigenvalue of $\widehat{\Omega}_{XX}$, $\lambda_{\min}\left(\widehat{\Omega}_{XX}\right)$, to the largest eigenvalue of $\mathcal{W}$, $\lambda_{\max}(\mathcal{W})$. That is, if we have $\lambda_{\min}\left(\widehat{\Omega}_{XX}\right) > \lambda_{\max}(\mathcal{W})$, then $\widehat{\Omega}^*_{XX} \succ 0$ holds true.

Specifically, for any $\mathbf{x} \neq 0$, we see that

$$\begin{aligned}\mathbf{x}^T \widehat{\Omega}^*_{XX} \mathbf{x} &= \mathbf{x}^T (\widehat{\Omega}_{XX} + \mathcal{W}) \mathbf{x} \\ &> \left(\min_{1 \leq i \leq n} \left\{ \lambda_i \left(\widehat{\Omega}_{XX} \right) \right\} - \max_{1 \leq i \leq n} \left\{ \lambda_i \left(\mathcal{W} \right) \right\} \right) \mathbf{x}^T \mathbf{x} > 0,\end{aligned}$$

and therefore we need $\lambda_{\min}\left(\widehat{\Omega}_{XX}\right) > \lambda_{\max}(\mathcal{W})$ for $\widehat{\Omega}^*_{XX} \succ 0$ to be true. Theorem B.1 provides a stronger (both necessary and sufficient) condition for the positive definiteness of $\widehat{\Omega}_{XX}$ to be true.

Theorem B.1 (Positive Definiteness of $\widehat{\Omega}^*_{XX}$)**.** *The matrix $\widehat{\Omega}^*_{XX} \equiv \widehat{\Omega}_{XX} + \mathcal{W}$ is positive definite if and only if every eigenvalue of $\widehat{\Omega}^{-1}_{XX}\mathcal{W}$ is greater than -1.*

The perturbation method takes effect by changing the eigenstructure of the ill-conditioned matrix $\widehat{\Omega}_{XX}$, especially on the eigenvalues. Whereas traditional matrix perturbation theory (Stewart and Sun, 1990) provides two-sided bounded for the changes in the eigenvalues, how a specific eigenvalue changes can sometimes be inconclusive. The next theorem specifies the exact directions of the changes in the smallest and largest eigenvalues of $\widehat{\Omega}_{XX}$ after perturbation.

Theorem B.2 (Eigenvalue Perturbation Theorem)**.** *Let $0 < \lambda_{\min} \leq \lambda_2 \leq \cdots \leq \lambda_{\max}$ and $0 < \lambda^*_{\min} \leq \lambda^*_2 \leq \cdots \leq \lambda^*_{\max}$ be the increasingly-ordered eigenvalues for $\widehat{\Omega}_{XX}$ and $\widehat{\Omega}^*_{XX}$, respectively. Then for $\varpi^* \in \mathbb{W}$, we have $\lambda^*_{\min} \geq \lambda_{\min}$.*

Whereas Theorem B.2 only specifies the change in $\lambda_{\min}$ is montonic, there is no guarantee that $\lambda_{\max}$ shares the same property. Nevertheless, the increase in $\lambda_{\min}$ still leads to the reduction in the CN of Ω_{XX}. The result is given in what follows.

Theorem B.3 (Reduction in condition number)**.** *For $\varpi^* \in \mathbb{W}$ such that $\widehat{\Omega}^*_{XX} \succ 0$ holds, the following hold:*

(i) The perturbed CN becomes smaller, viz.,

$$\kappa\left(\widehat{\Omega}^*_{XX}\right) \leq \kappa\left(\widehat{\Omega}_{XX}\right). \tag{B.1}$$

(ii) Let $0 < v_r < \infty$ be the reduction in the largest VIF. Then for $\varpi^* \in \mathbb{W}$ such that $\widehat{\Omega}^*_{XX} \succ 0$ holds, the reduction in CN is of order $\mathcal{O}(v_r)$, viz,

$$\kappa\left(\widehat{\Omega}_{XX}\right) = \kappa\left(\widehat{\Omega}^*_{XX}\right) + \mathcal{O}(v_r). \tag{B.2}$$

Implications of Theorem B.3 are as follows. First, through the changes in the smallest and the largest eigenvalues of the matrix $\widehat{\Omega}_{XX}$, the matrix becomes less and less ill-conditioned iteration after iteration. A geometric interpretation can be made: the shape of the matrix $\widehat{\Omega}_{XX}$, which was very ellipsoidal at the outset, becomes more and more round-shaped iteration after iteration. However, on the other hand, this property becomes the major limitation of the perturbation method, as the sensitivity in the eigenstructure resulting from the ill-conditionedness fades. 1 Second, the ridge-type estimates do not possess such a property in the CN as the LABs estimates do. This reveals the fact that the ridge-type estimates take effect by reducing the WVIF (or the VIF) solely. The close connection between WVIF (or VIF) and the CN becomes the most prominent property of the LABs estimates over the rigde-type estimates.

Third, the connection between the WVIF and the CN leads to the possibility to formulate another optimization problem using the CN. By letting

$$\widetilde{\mathbb{W}} \equiv \left\{\varpi \in \mathbb{R}^{n(n-1)/2} : -\mathbf{c}^T\varpi = c_r\right\}, \tag{B.3}$$

where $\mathbf{c}$ is a vector containing the derivatives $\frac{\partial\kappa(\widehat{\Omega}_{XX})}{\partial\varpi_{i,j}}$ for $i > j$ and $c_r > 0$ is the reduction in CN, be the CN-based feasible region, we can immediately derive the following CN-based LCQP

$$\min\left\{\left\|\widehat{B}\varpi\right\|_2^2 + \rho\left\|\varpi\right\|_2^2 : \varpi \in \widetilde{\mathbb{W}}\right\}, \tag{B.4}$$

where the associated $\mathcal{W}$ satisfies $\widehat{\Omega}^*_{XX} \succ 0$. The details will be clear later in Section B.2. The following results follow.

Proposition B.1 (Alternative Optimal Perturbation)**.** *Given $\widetilde{\mathbb{W}}$ nonvoid, there exists at least one $\varpi^* \in \widetilde{\mathbb{W}}$ that solves the CN-based LCQP (B.4). In particular,*

(i) the optimal perturbation vector ϖ^* has an explicit expression

$$\varpi^* = \frac{-c_r}{\|\mathbf{c}\|^2_{\left(\widehat{B}^T\widehat{B}+\rho I\right)^{-1}}} \left(\widehat{B}^T\widehat{B} + \rho I\right)^{-1} \mathbf{c}, \tag{B.5}$$

where the $\widehat{B}$ above is the same as the one in (6.3) and ρ is the penalty parameter.

(ii) For $\varpi^* \in \widetilde{\mathbb{W}}$ given in (B.4), the reduction in WVIF is of order $\mathcal{O}(c_r)$.

We remark on (ii) of Proposition B.1. For each distinct WVIF, the constant associated with the reduction term depends on the values of the associated constraint vector $\widehat{\mathbf{v}}$. To be more specific, for some WVIF$_j$, the constant C is given by

$$C = \frac{\widehat{\mathbf{v}}_j^T(\widehat{B}^T\widehat{B} + \rho I)^{-1}\mathbf{c}}{\|\mathbf{c}\|^2_{(\widehat{B}^T\widehat{B}+\rho I)^{-1}}}, \forall j = 1, ..., k.$$

Furthermore, when $j = k^*$, the enumerator in the above constant is exactly the same as that for the reduction in CN when using the WVIF-based optimal perturbation (6.3). The only difference in between resides in the denominator, thereby differing in the amount of reductions in the two aspects.

Lastly, we turn our attention to the changes in the WVIFs. Similar to the changes in VIFs in Jou et al. (2014), the reductions in WVIFs are not universal, but only those pathological WVIFs. By pathological, we mean those WVIFs with values higher than a critical value, depending upon the reduction step v_r and the target level v_l. Variables with good WVIFs will have their WVIF values fluctuating around the original values. The reduction is universally monotonic only when all WVIFs are greater than the critical value. By and large, the sum of all WVIFs will be lesser. We summarize the observations in the next theorem.

Theorem B.4 (Reduction in WVIFs)**.** *For $\varpi^* \in \mathbb{W}$ such that $\widehat{\Omega}^*_{XX} \succ 0$ holds, the following are true:*

(i) The sum of all perturbed WVIFs is less than or equal to that of the original WVIFs, viz.,

$$\mathbf{tr}\left(\widehat{\Omega}^{*-1}_{XX}\right) \leq \mathbf{tr}\left(\widehat{\Omega}^{-1}_{XX}\right). \tag{B.6}$$

(ii) Let v_l be the target VIF for the largest VIF to achieve and $v_r > 0$ the reduction in the largest VIF. Then if $\min_{1\le i\le k} \mathcal{V}_i(\widehat{\Omega}_{XX}) > v_l(v_r)$ for any $v_l > \lambda_{\min}^{-1}(\widehat{\Omega}_{XX}) + v_r$, we have

$$\left(\widehat{\Omega}_{XX}^{*-1}\right)_{i,i} \le \left(\widehat{\Omega}_{XX}^{-1}\right)_{i,i}, \ \forall i = 1, ..., k. \tag{B.7}$$

where $v_l(v_r)$ means the parameter v_l is dependent on the other parameter v_r.

B.2 Proofs of Theorems

We now give the proofs of the theorems in Section 4.2 in what follows. The first is on the positive definiteness of the perturbation correlation matrix $\widehat{\Omega}_{XX}$.

Proof of Theorem B.1. Assume $\widehat{\Omega}_{XX}^{*}$ is positive definite. We see that $\widehat{\Omega}_{XX} + \mathcal{W} = \widehat{\Omega}_{XX}(I + \widehat{\Omega}_{XX}^{-1}\mathcal{W})$. Since the matrices I and $\widehat{\Omega}_{XX}^{-1}\mathcal{W}$ commute and are both symmetric, the eigenvalues of the matrix $I + \widehat{\Omega}_{XX}^{-1}\mathcal{W}$ are the sum of eigenvalues of I and $\widehat{\Omega}_{XX}^{-1}\mathcal{W}$, *viz.*, $\Lambda(I + \widehat{\Omega}_{XX}^{-1}\mathcal{W}) = \mathbf{e} + \Lambda(\widehat{\Omega}_{XX}^{-1}\mathcal{W})$, for $\mathbf{e} \equiv [1, ..., 1]^T$. Since $\widehat{\Omega}_{XX}$ positive definite, by assumption, and $(I + \widehat{\Omega}_{XX}^{-1}\mathcal{W})$ is symmetric, the eigenvalues of the product $\widehat{\Omega}_{XX}(I + \widehat{\Omega}_{XX}^{-1}\mathcal{W})$ have the same number of positive, negative and zero eigenvalues as $\widehat{\Omega}_{XX}$. But $\widehat{\Omega}_{XX}$ and $\widehat{\Omega}_{XX}^{*}$ are both positive definite, it follows that $\Lambda(I + \widehat{\Omega}_{XX}^{-1}\mathcal{W}) > 0$, which implies the claimed result.

By the same reasoning as above, if every eigenvalue of $\widehat{\Omega}_{XX}^{-1}\mathcal{W}$ is greater than -1, $(I + \widehat{\Omega}_{XX}^{-1}\mathcal{W})$ is positive definite. The product of two positive definite matrices yields a positive definite matrix (Horn and Johnson, 2012). This concludes the proof. □

Next, we prove the important eigenvalue perturbation theorem.

Proof of Theorem B.2. To prove this claim, we consider the difference $\lambda_{\min}^{*} - \lambda_{\min}$ and use the Cauchy interlacing theorem to get

$$\lambda_{\min}^{*} - \lambda_{\min} \ge \sum\nolimits_{j=1}^{n-1} \theta_j - \sum\nolimits_{j=1}^{n-1} \theta_j^{*} = 0.$$

where θ_j and θ_j^* are the j-th eigenvalue of the submatrix $\widehat{\Omega}_{XX}\{k^*\}$ and $\widehat{\Omega}^*_{XX}\{k^*\}$, respectively. $\square$

Before the proof concerning the reduction in the CN after the optimal perturbation, we need the following lemma.

Lemma B.1. *Given any* $\mathbf{x}$ *and* $\mathbf{y}$ *in* $\mathbb{R}^n$*. If* $\mathbf{x}$ *and* $\mathbf{y}$ *lie in the same quadrant, then for any matrix* $Q \succ 0$ *in* $\mathbb{R}^{n\times n}$*, we have* $\mathbf{x}^T Q\mathbf{y} > 0$*.*

Proof. It follows that

$$\mathbf{x}^T Q\mathbf{y} = \mathbf{tr} Q\mathbf{x}\mathbf{y}^T = \mathbf{tr} QR > \lambda_{\min}(Q)\mathbf{tr} R > 0,$$

where the first inequality follows from Fang et al. (1994) and the last inequality holds because the diagonals of $R \equiv \mathbf{x}\mathbf{y}^T$ are positive and $\lambda_{\min}(Q) > 0$ due to positive definiteness. $\square$

We are ready to prove Theorem B.3.

Proof of Theorem B.3. The first claim is an immediate corollary of (i) and (ii) in Theorem B.2. To prove (ii), we start with the following first-order approximation for CN

$$\kappa(\widehat{\Omega}^*_{XX}) = \kappa(\widehat{\Omega}_{XX}) + \mathbf{c}^T\varpi,$$

where $\mathbf{c} \in \mathbb{R}^{k(k-1)/2}$ is the vector comprising derivatives $\frac{\partial}{\partial \varpi_{i,j}}\kappa(\widehat{\Omega}_{XX})$ for $i > j$. It is not hard to see that

$$\frac{\partial \kappa(\widehat{\Omega}_{XX})}{\partial \varpi_{i,j}} = \frac{2\lambda_{\min}\upsilon_{\max,i}\upsilon_{\max,j} - 2\lambda_{\max}\upsilon_{\min,i}\upsilon_{\min,j}}{\lambda_{\min}^2}, \text{ for } i > j,$$

where $\upsilon_{\max,i}$, for instance, is the i-th entry of the eigenvector associated with $\lambda_{\max}$ and the rest follow. The reduction in the CN is therefore $-\mathbf{c}^T\varpi$, for any $\varpi \in \mathbb{W}$. Let $\varpi^* \in \mathbb{W}$

be the optimal perturbation as given in (6.3). It follows that

$$\begin{aligned}\kappa(\widehat{\Omega}_{XX}) &= \kappa(\widehat{\Omega}^*_{XX}) - \mathbf{c}^T\varpi^* \\ &= \kappa(\widehat{\Omega}^*_{XX}) + \left(\frac{\mathbf{c}^T\left(B^TB+\rho I\right)^{-1}\widehat{\mathbf{v}}_{k^*}}{\|\widehat{\mathbf{v}}_{k^*}\|_{(B^TB+\rho I)^{-1}}}\right)v_r \\ &= \kappa(\widehat{\Omega}^*_{XX}) + \mathcal{O}(v_r).\end{aligned}$$

It remains to show the constant associated with v_r is positive, which is equivalent to prove $\mathbf{c}^T\left(B^TB+\rho I\right)^{-1}\mathbf{v}_{k^*} \geq 0$. To prove this, according to Lemma B.1, it suffices to prove that $\mathbf{c}$ and $\mathbf{v}_{k^*}$ lie in the same quadrant, which then guarantees the positivity. We consider the following product

$$\mathbf{c}\mathbf{v}^T_{k^*} \equiv P = \left[\frac{\partial\kappa(\widehat{\Omega}_{XX})}{\partial\varpi_{i,j}} \times \frac{\partial\mathcal{V}_{k^*}\left(\widehat{\Omega}_{XX}\right)}{\partial\varpi_{i,j}}\right],$$

which is a matrix resulting from the product of derivatives for all (i,j) pair and $i>j$. The diagonals of the above product, $P_{r,r}$, $1 \leq r \leq \frac{n(n-1)}{2}$, are positive because both quantities are decreasing at ϖ^*. This proves the claimed result in (ii). □

Next, we prove Theorem B.4.

Proof of Theorem B.4. (i) To prove (B.6), we use the fact that the trace of a matrix equals the sum of its eigenvalues. Considering the difference $\mathbf{tr}\widehat{\Omega}^{*-1}_{XX} - \mathbf{tr}\widehat{\Omega}^{-1}_{XX}$, it follows that

$$\mathbf{tr}\widehat{\Omega}^{*-1}_{XX} - \mathbf{tr}\widehat{\Omega}^{-1}_{XX} = \sum_{i=1}^{n}\left(\frac{1}{\lambda^*_i} - \frac{1}{\lambda_i}\right) = \sum_{i=1}^{n}\left(\frac{\lambda_i-\lambda^*_i}{\lambda^*_i\lambda_i}\right) \leq \frac{\sum_{i=1}^{n}(\lambda_i-\lambda^*_i)}{\lambda^*_{\min}\lambda_{\min}} = 0$$

The first inequality holds because $\lambda_{\min} \leq \lambda_j$, $\forall j$ and the same for $\lambda^*_{\min}$. The last equality holds for $\mathbf{tr}\widehat{\Omega}^*_{XX} = \mathbf{tr}\widehat{\Omega}_{XX}$.

(ii) Given $\min_{1\leq i\leq k}\mathcal{V}_i(\widehat{\Omega}_{XX}) > v_l$, we have $\mathcal{V}_j(\widehat{\Omega}_{XX}) > v_l$, $\forall j = 1,...,k$. For the case $j=k^*$, this is trivial because of the WVIF constraint. Consider any $j \neq k^*$, we see that

$$\left(\widehat{\Omega}^{*-1}_{XX}\right)_{j,j} - \left(\widehat{\Omega}^{-1}_{XX}\right)_{j,j} < \left(\widehat{\Omega}^{*-1}_{XX}\right)_{j,j} - v_l = \frac{\theta^*_{\min}\theta^*_2\cdots\theta^*_{k-1}}{\lambda^*_{\min}\lambda^*_2\cdots\lambda^*_{\max}} - v_l \leq \frac{1}{\lambda^*_{\min}} - v_l \leq 0.$$

by using Cauchy interlacing theorem (Horn and Johnson, 2012). Here θ_j^* is the j-th eigenvalue of the submatrix $\widehat{\Omega}_{XX}\{k^*\}$, $j = 1, \ldots, k-1$, and $v_l > \frac{1}{\lambda_{\max}^*} + v_r$ for some $v_r > 0$. This proves (B.7). □

Lastly, we prove the boundedness results on the LABs estimates and the algorithm that terminates in a finite number of steps.

Proof. (i) The bias between the PWLSE and an OLSE is given by

$$\|\widehat{\beta}_{\mathcal{W}^*} - \widehat{\beta}_{\text{OLS}}\|_2^2 = \|\widehat{B}\varpi^*\|_2^2 \leq \|\widehat{B}\|_2^2 \|\varpi\|_2^2 = \mathcal{O}(\|\varpi^*\|_2^2).$$

(ii) Let M be the total number of iterations for the algorithm to terminate. It follows that

$$M = \left\lceil \frac{\mathcal{V}_{k^*}\left(\widehat{\Omega}_{XX}\right) - v_l}{v_r} \right\rceil < \infty,$$

given $0 < v_l < \infty$, the target level for the largest WVIF, and $0 < v_r < \infty$, the reduction step in WVIFs, are finite. Note that $\lceil a \rceil$ denotes the least integer greater than or equal to $a \in \mathbb{R}$.

(iii) Let i be the iteration number and M be the maximal iteration number in (ii). We have

$$\|\widehat{\beta}_{\mathcal{W}^*}^i - \widehat{\beta}_{\text{OLS}}\|_2^2 \leq c'\|\varpi^i\|_2^2, \text{ for } i = 1, ..., M,$$

wherein $\widehat{\beta}_{\mathcal{W}^*}^i$ and ϖ^i are the results at the i-th iteration of the algorithm, and $c' \equiv \|\widehat{B}\|_2^2$. Hence, upon termination, the total bias generated by the algorithm is

$$\sum\nolimits_{i=1}^{M} \|\widehat{\beta}_{\mathcal{W}^*}^i - \widehat{\beta}_{\text{OLS}}\|_2^2 \leq \sum\nolimits_{i=1}^{M} c'\|\varpi^i\|_2^2 \leq \widetilde{c}_M,$$

where $\widetilde{c}_M \equiv c' M \sup_{0 \leq i \leq M} \|\varpi^i\|_2^2$ is a constant. This proves the claim. □

Bibliography

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (pls regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):97–106.

Aitken, A. C. (1936). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48.

Al-Khayyal, F. A., Larsen, C., and Van Voorhis, T. (1995). A relaxation method for nonconvex quadratically constrained quadratic programs. *Journal of Global Optimization*, 6(3):215–230.

Alheety, M. and Kibria, B. G. (2009). On the liu and almost unbiased liu estimators in the presence of multicollinearity with heteroscedastic or correlated errors. *Surveys in Mathematics and its Applications*, 4:155–167.

Alkhamisi, M. (2012). Ridge estimation in linear models with heteroskedastic errors. *Sankhya B*, 74(2):302–314.

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858.

Anstreicher, K. M. (2009). Semidefinite programming versus the reformulation-linearization technique for nonconvex quadratically constrained quadratic programming. *Journal of Global Optimization*, 43(2-3):471–484.

Anstreicher, K. M. (2012). On convex relaxations for quadratically constrained quadratic programming. *Mathematical programming*, 136(2):233–251.

Aslam, M. (2013). Using heteroscedasticity consistent standard errors for the linear regression model with correlated regressors. *Communications in Statistics-Simulation and Computation*, (just-accepted).

Aslam, M. (2014). Performance of kibria's method for the heteroscedastic ridge regression model: Some monte carlo evidence. *Communications in Statistics-Simulation and Computation*, 43(4):673–686.

Audet, C., Hansen, P., Jaumard, B., and Savard, G. (2000). A branch and cut algorithm for nonconvex quadratically constrained quadratic programming. *Mathematical Programming*, 87(1):131–152.

Bagheri, A. and Midi, H. (2009). Robust estimations as a remedy for multicollinearity caused by multiple high leverage points. *Journal of Mathematics and Statistics*, 5(4):311.

Bao, X., Sahinidis, N. V., and Tawarmalani, M. (2011). Semidefinite relaxations for quadratically constrained quadratic programming: A review and comparisons. *Mathematical programming*, 129(1):129–157.

Bashtian, M. H., Arashi, M., and Tabatabaey, S. (2011). Using improved estimation strategies to combat multicollinearity. *Journal of Statistical Computation and Simulation*, 81(12):1773–1797.

Batah, F. S. M., Özkale, M. R., and Gore, S. (2009). Combining unbiased ridge and principal component regression estimators. *Communications in StatisticsTheory and Methods*, 38(13):2201–2209.

Bayhan, G. M. and Bayhan, M. (1998). Forecasting using autocorrelated errors and multicollinear predictor variables. *Computers & industrial engineering*, 34(2):413–421.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2013). *Nonlinear programming: theory and algorithms*. John Wiley & Sons.

Belsley, D. A. (1984). Collinearity and forecasting. *Journal of Forecasting*, 3(2):183–196.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons.

Ben-Tal, A. and Teboulle, M. (1996). Hidden convexity in some nonconvex quadratically constrained quadratic programming. *Mathematical Programming*, 72(1):51–63.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327.

Boyd, S. and Vandenberghe, L. (2009). *Convex optimization*. Cambridge university press.

Chatterjee, S. and Hadi, A. S. (2006). *Regression analysis by example*. John Wiley & Sons.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., Stein, C., et al. (2001). *Introduction to algorithms*, volume 2. MIT press Cambridge.

Craney, T. A. and Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, 14(3):391–403.

Cribari-Neto, F., Maria da Glória, A. L., et al. (2014). New heteroskedasticity-robust standard errors for the linear regression model. *Brazilian Journal of Probability and Statistics*, 28(1):83–95.

Cribari-Neto, F., Souza, T. C., and Vasconcellos, K. L. (2007). Inference under heteroskedasticity and leveraged data. *Communications in StatisticsTheory and Methods*, 36(10):1877–1888.

Curto, J. D. and Pinto, J. C. (2007). New multicollinearity indicators in linear regression models. *International Statistical Review*, 75(1):114–121.

Curto, J. D. and Pinto, J. C. (2011). The corrected vif (cvif). *Journal of Applied Statistics*, 38(7):1499–1507.

Echambadi, R. and Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science*, 26(3):438–445.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350.

Fang, S. and Rajasekera, J. (1986). Controlled perturbations for quadratically constrained quadratic programs. *Mathematical Programming*, 36(3):276–289.

Fang, Y., Loparo, K. A., and Feng, X. (1994). Inequalities for the trace of matrix product. *Automatic Control, IEEE Transactions on*, 39(12):2489–2490.

Faraway, J. J. (2002). Practical regression and anova using r.

Fierro, R. D. and Bunch, J. R. (1992). Multicollinearity and total least squares. *Preprint Series*, 977.

Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

García, C., García, J., López Martín, M., and Salmerón, R. (2014). Collinearity: revisiting the variance inflation factor in ridge regression. *Journal of Applied Statistics*, (ahead-of-print):1–14.

Hadi, A. S. and Ling, R. F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician*, 52(1):15–19.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Jou, Y.-J., Huang, C.-C. L., and Cho, H.-J. (2014). A vif-based optimization model to alleviate collinearity problems in multiple linear regression. *Computational Statistics*, pages 1–27.

Khalaf, G. (2012). Improved estimator in the presence of multicollinearity. *Journal of Modern Applied Statistical Methods*, 11(1):12.

Knol, D. L. and ten Berge, J. M. (1989). Least-squares approximation of an improper correlation matrix by a proper one. *Psychometrika*, 54(1):53–61.

Kovács, P., Petres, T., and Tóth, L. (2005). A new measure of multicollinearity in linear regression models. *International statistical review*, 73(3):405–412.

Lazaridis, A. (2007). A note regarding the condition number: the case of spurious and latent multicollinearity. *Quality & Quantity*, 41(1):123–135.

Le Cessie, S. and Van Houwelingen, J. (1992). Ridge estimators in logistic regression. *Applied statistics*, pages 191–201.

Lee, A. and Silvapulle, M. (1988). Ridge estimation in logistic regression. *Communications in Statistics-Simulation and Computation*, 17(4):1231–1257.

Lee, E. T. (1974). A computer program for linear logistic regression analysis. *Computer programs in biomedicine*, 4(2):80–92.

Lesaffre, E. and Marx, B. D. (1993). Collinearity in generalized linear regression. *Communications in Statistics-Theory and Methods*, 22(7):1933–1952.

Leung, S. F. and Yu, S. (2000). Collinearity and two-step estimation of sample selection models: problems, origins, and remedies. *Computational Economics*, 15(3):173–199.

Liao, D. and Valliant, R. (2012). Variance inflation factors in the analysis of complex survey data. *Survey Methodology*, 38(1):53–62.

Lin, D., Foster, D. P., and Ungar, L. H. (2011). Vif regression: A fast regression algorithm for large data. *Journal of the American Statistical Association*, 106(493).

Lin, F.-J. (2008). Solving multicollinearity in the process of fitting regression model using the nested estimate procedure. *Quality & Quantity*, 42(3):417–426.

Linderoth, J. (2005). A simplicial branch-and-bound algorithm for solving quadratically constrained quadratic programs. *Mathematical Programming*, 103(2):251–282.

Liu, K. (2003). Using liu-type estimator to combat collinearity. *Communications in Statistics-Theory and Methods*, 32(5):1009–1020.

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical association*, 62(319):819–841.

Lu, C., Fang, S.-C., Jin, Q., Wang, Z., and Xing, W. (2011). Kkt solution and conic relaxation for solving quadratically constrained quadratic programming problems. *SIAM Journal on Optimization*, 21(4):1475–1490.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pages 437–461. Springer.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Mackinnon, M. J. and Puterman, M. L. (1989). Collinearity in generalized linear models. *Communications in statistics-theory and methods*, 18(9):3463–3472.

Mandel, J. (1982). Use of the singular value decomposition in regression analysis. *The American Statistician*, 36(1):15–24.

Månsson, K. and Shukur, G. (2011). A poisson ridge regression estimator. *Economic Modelling*, 28(4):1475–1481.

Månsson, K., Shukur, G., and Sjölander, P. (2014). A new asymmetric interaction ridge (air) regression method. *Communications in Statistics-Theory and Methods*, 43(3):616–643.

McCullough, P. and Nelder, J. A. (1989). Generalized linear models.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100.

McDonald, G. C. and Schwing, R. C. (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15(3):463–481.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Misener, R. and Floudas, C. A. (2012). Global optimization of mixed-integer quadratically-constrained quadratic programs (miqcqp) through piecewise-linear and edge-concave relaxations. *Mathematical programming*, 136(1):155–182.

Næs, T. and Mevik, B.-H. (2001). Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, 15(4):413–426.

Nemirovskii, A. and Scheinberg, K. (1996). Extension of karmarkar's algorithm onto convex quadratically constrained quadratic problems. *Mathematical Programming*, 72(3):273–289.

Obrien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690.

P. Decell Jr, H. (1974). On the derivative of the generalized inverse of a matrix. *Linear and Multilinear Algebra*, 1(4):357–359.

Paul, S. (1989). Test for the equality of several correlation coefficients. *Canadian Journal of Statistics*, 17(2):217–227.

Prais, S. J. and Winsten, C. B. (1954). Trend estimators and serial correlation. Technical report, Cowles Commission discussion paper.

Robinson, C. and Schumacker, R. E. (2009). Interaction effects: centering, variance inflation factor, and interpretation issues. *Multiple Linear Regression Viewpoints*, 35(1):6–11.

Schaefer, R., Roi, L., and Wolfe, R. (1984). A ridge logistic estimator. *Communications in Statistics-Theory and Methods*, 13(1):99–113.

Schaefer, R. L. (1986). Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation*, 25(1-2):75–91.

Schinnar, A. and Walters, L. (1983). Minimum bias respecification of multicollinearity regressions.

Schwing, R. C. and McDonald, G. C. (1976). Measures of association of some air pollutants, natural ionizing radiation and cigarette smoking with mortality rates. *Science of the Total Environment*, 5(2):139–169.

Shacham, M. and Brauner, N. (1997). Minimizing the effects of collinearity in polynomial regression. *Industrial & engineering chemistry research*, 36(10):4405–4412.

Shapiro, A. (2001). On duality theory of conic linear problems. In *Semi-infinite programming*, pages 135–165. Springer.

Shieh, G. (2011). Clarifying the role of mean centring in multicollinearity of interaction effects. *British Journal of Mathematical and Statistical Psychology*, 64(3):462–477.

Soofi, E. S. (1990). Effects of collinearity on information about regression coefficients. *Journal of Econometrics*, 43(3):255–274.

Spanos, A. and McGuirk, A. (2002). The problem of near-multicollinearity revisited: erratic vs systematic volatility. *Journal of Econometrics*, 108(2):365–393.

Stewart, G. W. (1987). Collinearity and least squares regression. *Statistical Science*, pages 68–84.

Stewart, G. W. and Sun, J.-g. (1990). Matrix perturbation theory.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Topsoe, F. (2004). Some bounds for the logarithmic function. *Inequality Theory and Applications*, 4.

Vandenberghe, L. and Boyd, S. (1996). Semidefinite programming. *SIAM review*, 38(1):49–95.

Weissfeld, L. A. and Sereika, S. M. (1991). A multicollinearity diagnostic for generalized linear models. *Communications in Statistics-Theory and Methods*, 20(4):1183–1198.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

Xin, L. and Zhu, M. (2012). Stochastic stepwise ensembles for variable selection. *Journal of Computational and Graphical Statistics*, 21(2):275–294.

Zhen, Z. and Wohlgenant, M. (2010). Modeling farm-retail price spread in u.s. pork industry. *Mater Thesis*.

Zheng, X., Sun, X., and Li, D. (2011). Nonconvex quadratically constrained quadratic programming: best dc decompositions and their sdp representations. *Journal of Global Optimization*, 50(4):695–712.