Iowa Research Online
The University of Iowa's Institutional Repository

University of Iowa
Iowa Research Online

Theses and Dissertations

2012

Concave selection in generalized linear models

Dingfeng Jiang
University of Iowa

Copyright 2012 Dingfeng Jiang

This dissertation is available at Iowa Research Online: http://ir.uiowa.edu/etd/2902

Recommended Citation
Jiang, Dingfeng. "Concave selection in generalized linear models." PhD (Doctor of Philosophy) thesis, University of Iowa, 2012.
http://ir.uiowa.edu/etd/2902.

Follow this and additional works at: http://ir.uiowa.edu/etd

Part of the Biostatistics Commons

CONCAVE SELECTION IN GENERALIZED LINEAR MODELS

by

Dingfeng Jiang

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics
in the Graduate College of
The University of Iowa

May 2012

Thesis Supervisor: Professor Jian Huang

ABSTRACT

A family of concave penalties, including the smoothly clipped absolute deviation (SCAD) and minimax concave penalties (MCP), has been shown to have attractive properties in variable selection. The computation of concave penalized solutions, however, is a difficult task. We propose a majorization minimization by coordinate descent (MMCD) algorithm to compute the solutions of concave penalized generalized linear models (GLM). In contrast to the existing algorithms that uses local quadratic or local linear approximation of the penalty, the MMCD majorizes the negative log-likelihood by a quadratic loss, but does not use any approximation to the penalty. This strategy avoids the computation of scaling factors in iterative steps, hence improves the efficiency of coordinate descent. Under certain regularity conditions, we establish the theoretical convergence property of the MMCD algorithm. We implement this algorithm in a penalized logistic regression model using the SCAD and MCP penalties. Simulation studies and a data example demonstrate that the MMCD works sufficiently fast for the penalized logistic regression in high-dimensional settings where the number of covariates is much larger than the sample size.

Grouping structure among predictors exists in many regression applications. We first propose an ℓ_2 grouped concave penalty to incorporate such group information in a regression model. The ℓ_2 grouped concave penalty performs group selection and includes group Lasso (Yuan and Lin (2006)) as a special case. An efficient algorithm is developed and its theoretical convergence property is established under certain

regularity conditions.

The group selection property of the ℓ_2 grouped concave penalty is desirable in some applications; while in other applications selection at both group and individual levels is needed. Hence, we propose an ℓ_1 grouped concave penalty for variable selection at both individual and group levels. An efficient algorithm is also developed for the ℓ_1 grouped concave penalty.

Simulation studies are performed to evaluate the finite-sample performance of the two grouped concave selection methods. The new grouped penalties are also used in analyzing two motivation datasets. The results from both the simulation and real data analyses demonstrate certain benefits of using grouped penalties. Therefore, the proposed concave group penalties are valuable alternatives to the standard concave penalties.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

CONCAVE SELECTION IN GENERALIZED LINEAR MODELS

by

Dingfeng Jiang

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics
in the Graduate College of
The University of Iowa

May 2012

Thesis Supervisor: Professor Jian Huang

Copyright by
DINGFENG JIANG
2012
All Rights Reserved

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Dingfeng Jiang

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Biostatistics at the May 2012 graduation.

Thesis Committee: _____
Jian Huang, Thesis Supervisor

_____
Joseph Cavanaugh

_____
Kung-Sik Chan

_____
Michael Jones

_____
Ying Zhang

ACKNOWLEDGEMENTS

I want to thank my professors in both the biostatistics and statistics departments for their excellent education to train me for research in this fascinating field. Specifically, I would like to thank my advisor Dr. Jian Huang for introducing me to the field of variable selection. For the past three years, he guided me though many difficulties by provoking conversation and great ideas. His open-mindedness and profound knowledge serves an inspiring model to me. I also want to thank Dr. Jane Pendergast for her support by providing the Wisewoman dataset.

I am grateful to the CPHS ITS support team for their tremendous help to maintain an excellent platform of Linux lab.

ABSTRACT

A family of concave penalties, including the smoothly clipped absolute deviation (SCAD) and minimax concave penalties (MCP), has been shown to have attractive properties in variable selection. The computation of concave penalized solutions, however, is a difficult task. We propose a majorization minimization by coordinate descent (MMCD) algorithm to compute the solutions of concave penalized generalized linear models (GLM). In contrast to the existing algorithms that uses local quadratic or local linear approximation of the penalty, the MMCD majorizes the negative log-likelihood by a quadratic loss, but does not use any approximation to the penalty. This strategy avoids the computation of scaling factors in iterative steps, hence improves the efficiency of coordinate descent. Under certain regularity conditions, we establish the theoretical convergence property of the MMCD algorithm. We implement this algorithm in a penalized logistic regression model using the SCAD and MCP penalties. Simulation studies and a data example demonstrate that the MMCD works sufficiently fast for the penalized logistic regression in high-dimensional settings where the number of covariates is much larger than the sample size.

Grouping structure among predictors exists in many regression applications. We first propose an ℓ_2 grouped concave penalty to incorporate such group information in a regression model. The ℓ_2 grouped concave penalty performs group selection and includes group Lasso (Yuan and Lin (2006)) as a special case. An efficient algorithm is developed and its theoretical convergence property is established under certain

regularity conditions.

The group selection property of the ℓ_2 grouped concave penalty is desirable in some applications; while in other applications selection at both group and individual levels is needed. Hence, we propose an ℓ_1 grouped concave penalty for variable selection at both individual and group levels. An efficient algorithm is also developed for the ℓ_1 grouped concave penalty.

Simulation studies are performed to evaluate the finite-sample performance of the two grouped concave selection methods. The new grouped penalties are also used in analyzing two motivation datasets. The results from both the simulation and real data analyses demonstrate certain benefits of using grouped penalties. Therefore, the proposed concave group penalties are valuable alternatives to the standard concave penalties.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

1.1 Ridge, Bridge and Lasso Penalties 1
1.2 Concave Penalty: SCAD and MCP 2
1.3 Computation for Penalized Linear Models 4
1.4 Majorization Minimization Algorithm 7
1.5 Group Lasso for Group Variable Selection 8
1.6 Overview of the Dissertation 10

2 MAJORIZATION MINIMIZATION BY COORDINATE DESCENT ALGORITHM FOR CONCAVE PENALIZATED GLM 12

2.1 Majorization Minimization by Coordinate Descent 12
2.2 Convergence Analysis 18
2.3 Comparison with Other Algorithms 18
2.4 The MMCD for Logistic Regression 20
2.4.1 Computation of solution surface 20
2.4.2 Design of simulation study 23
2.4.3 Comparison of computational efficiency 24
2.4.4 Comparison of Lasso, SCAD and MCP 25
2.4.5 Comparison under misspecification 27
2.4.6 Comparison in a cancer study 29
2.4.7 Results using tuning parameter selection 30
2.5 Further Example of the MMCD Algorithm 31

3 ℓ_2 GROUPED CONCAVE PENALTY IN GLM 34

3.1 ℓ_2 Grouped Concave Penalty in Linear Regression 34
3.2 Computation of the ℓ_2 Grouped Concave Penalty 37
3.3 Convergence Analysis of Proposed Algorithms 40
3.4 Extension of the ℓ_2 Grouped Concave Penalty in GLM 41
3.5 Simulation Studies in Linear and Logistic Models 42
3.6 Simulation Studies in Poisson Models 45

4 ℓ_1 GROUPED CONCAVE PENALTY IN GLM 49

4.1 ℓ_1 Grouped Concave Penalty in Linear Regression 49
4.2 Computation of ℓ_1 Grouped Concave Penalty 51
4.3 Convergence Analysis of Proposed Algorithms 54
4.4 Extension of Grouped Concave Penalty in GLM 54
4.5 Simulation Studies in Linear and Logistic Models 56

5 APPLICATION OF GROUPED PEANALTY IN HEALTH CARE AND GENOMIC DATASETS 61

5.1 Comparison Based on Random Partition 62
5.2 Results Using Tuning Parameter Selection 63

6 CONCLUSION AND DISCUSSION 66

APPENDIX 68

A PROOF OF THEOREM IN CHAPTER 2 68

A.1 Proof of Theorem 2.1 70

B PROOF OF THEOREM IN CHAPTER 3 74

B.1 Proof of Theorem 3.1 76
B.2 Outline of Proof of Theorem 3.2 79

C PROOF OF THEOREM IN CHAPTER 4 80

C.1 Proof of Theorem 4.1 82
C.2 Outline of Proof of Theorem 4.2 85

REFERENCES 87

LIST OF TABLES

Table

2.1 Comparison of computational efficiency between the adaptive rescaling and MMCD algorithms in MCP penalized logistic regressions, $n = 100$ and $p = 1,000$. 24

2.2 Comparison of Lasso, SCAD and MCP in terms of model size (MS), false discover rate (FDR) and predictive AUC (PAUC), $n = 100$, $p = 1,000$. . 26

2.3 Comparison of Lasso, SCAD and MCP in terms of model size (MS), false discover rate (FDR) and predictive AUC (PAUC), $n = 300$, $p = 1,000$. . 27

2.4 Comparison of Lasso, SCAD and MCP in terms of model size (MS), false discover rate (FDR) and predictive AUC (PAUC) under misspecification, $n = 100$, $p = 1,000$. 28

2.5 Comparison of Lasso, SCAD and MCP in terms of model size (MS), false discover rate (FDR) and predictive AUC (PAUC) under misspecification, $n = 300$, $p = 1,000$. 29

2.6 Comparison of Lasso, SCAD and MCP based on 900 random partition of a breast cancer microarray dataset. 30

3.1 Comparison of the ℓ_2 grouped vs. the ungrouped concave penalties in linear models, $n = 300$, $p = 500$ and $\rho = 0.6$. 46

3.2 Comparison of the ℓ_2 grouped vs. the ungrouped concave penalties in logistic models, $n = 300$, $p = 500$ and $\rho = 0.6$. 47

3.3 Comparison of the group Lasso and Lasso penalties in Poisson models, $n = 300$, $p = 500$ and $\rho = 0.6$. 48

4.1 Comparison of the ℓ_1 grouped vs. the ungrouped concave penalties in linear models, $n = 300$, $p = 500$, and $\rho = 0.6$. 57

4.2 Comparison of the ℓ_1 grouped vs. the ungrouped concave penalties in logistic models, $n = 300$, $p = 500$, and $\rho = 0.6$. 58

4.3 Comparison of the ℓ_1 GSCAD and GMCP vs. the ℓ_2 GSCAD and GMCP in linear models, $n = 300$, $p = 500$ and $\rho = 0.6$. 59

4.4 Comparison of the ℓ_1 GSCAD and GMCP vs. the ℓ_2 GSCAD and GMCP in logistic models, $n = 300$, $p = 500$ and $\rho = 0.6$. 60

5.1 Comparison of the grouped vs. the ungrouped concave penalties in WW dataset. 63

5.2 Comparison of the grouped vs. the ungrouped concave penalties in BC dataset. 64

5.3 Data analysis results for the BC dataset using the CV-AUC tuning parameter selection approach. 65

LIST OF FIGURES

Figure

1.1 Penalty and thresholding operator functions of Lasso (left), SCAD (middle) and MCP (right). 5

2.1 SCAD penalty and its majorizations, LQA, Perturbed LQA (PLQA) and LLA. 19

2.2 Solution paths along κ for a causal variable (plot a), and a null variable (plot b). 22

3.1 Boundary of the ℓ_1 and ℓ_2 norm for a vector with two elements 36

3.2 Solution paths of the group Lasso and the ℓ_2 GMCP. Group Lasso (left), ℓ_2 GMCP (center and right). 40

4.1 Solution paths of the proportional Lasso and the ℓ_1 GMCP. Proportional Lasso (left), ℓ_1 GMCP (center and right). 53

CHAPTER 1
INTRODUCTION

Variable selection is a fundamental problem in statistics. When a model is built, a subset of variables is often pursued to reduce complexity and increase interpretability. Subset selection criteria such as AIC (Akaike (1974)), BIC (Schwarz (1978)), or C_p (Mallows (1973)) is generally adequate for small p, the number of variables. However, when p is large, subset selection is computationally infeasible and lacks stability (Breiman (1996)).

For high-dimensional problems with $p \gg n$, the penalization method is an important approach for variable selection. This chapter introduces several important penalty functions $\rho(t;\lambda)$, and provides a general background for the subsequent chapters.

1.1 Ridge, Bridge and Lasso Penalties

Consider a linear model with a response vector $\boldsymbol{y} \in \mathbb{R}^n$ depending on p predictors $\boldsymbol{x}_j$ through a linear combination $\sum_{j=1}^p \beta_j \boldsymbol{x}_j$. Then the criteria function of a penalized linear model is defined as

$$Q(\boldsymbol{\beta};\lambda) = (2n)^{-1}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^p \rho(|\beta_j|;\lambda), \tag{1.1}$$

with a penalty $\rho(t;\lambda)$ indexed by $\lambda \geq 0$. Notation $\|\cdot\|_2$ is the Euclidean norm, $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$. In the linear model, $y_i = \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. In the 1970s', Hoerl and Kennard (1970) proposed the ridge regression with $\rho(t;\lambda) = \lambda t^2/2$ to handle the ill-conditioned design matrix. The trade

off between unbiasness and variance improves the overall accuracy of prediction. The ridge penalty, however, does not perform variable selection. Frank and Friedman (1993) proposed a family of bridge penalty with $\rho(t;\lambda,\gamma) = \lambda|t|^\gamma$, $\gamma > 0$. The bridge penalty has the variable selection feature when $\gamma \leq 1$. Note that when $\gamma = 2$, the bridge penalty is the ridge penalty; when $\gamma = 1$, the bridge penalty is the Least absolute shrinkage and selection operator (Lasso) (Donoho and Johnstone (1994); Tibshirani (1996)) or ℓ_1 penalty with $\rho(t;\lambda) = \lambda|t|$.

The ℓ_1 penalty applies the same degree of penalization to all coefficients, leading to the biased estimation of large coefficients. The biased estimation interferes with the selection. Under the ir-representable condition on the design matrix, Zhao and Yu (2006) proved the selection consistency of the ℓ_1 penalty. Meinshausen and Bühlmann (2006) showed similar results under the neighborhood stability condition. However, under a milder sparse Riesz condition, Zhang and Huang (2008) proved that Lasso is rate consistent only and tends to over select. The biased estimation of large coefficients motivates the concave penalties introduced below.

1.2 Concave Penalty: SCAD and MCP

Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty, and Zhang (2010) proposed the minimum concave penalty (MCP). The SCAD is defined as

$$\rho(t;\lambda,\gamma) = \lambda|t|\mathbf{1}_{\{|t|\leq\lambda\}} + \frac{\gamma\lambda|t| - 0.5(t^2+\lambda^2)}{\gamma-1}\mathbf{1}_{\{\lambda<|t|\leq\gamma\lambda\}} + \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}\mathbf{1}_{\{|t|>\gamma\lambda\}}, \quad (1.2)$$

with $\lambda \geq 0$ and $\gamma > 2$. Here $\mathbf{1}_{x \in A}$ is the indicator function. The MCP is defined as

$$\rho(t; \lambda, \gamma) = \frac{2\lambda\gamma|t| - |t|^2}{2\gamma} \mathbf{1}_{\{|t| \leq \lambda\gamma\}} + \frac{1}{2}\lambda^2\gamma \mathbf{1}_{\{|t| > \lambda\gamma\}}, \tag{1.3}$$

for $\lambda \geq 0$ and $\gamma > 1$. Both penalties gradually reduce the penalization rate to zero as $|t|$ gets larger. The regularization parameter γ controls the degree of concavity, with smaller γ corresponding to more concave penalty. When $\gamma \to \infty$, SCAD and MCP converge to the ℓ_1 penalty. The SCAD and MCP penalties are illustrated in the middle and right panel of Figure 1.1. The (nearly) unbiased estimation of coefficients enables SCAD and MCP to correctly select important variables and estimate their coefficients with high probabilities as if the model were known in advance under certain sparsity conditions and other appropriate regularity conditions. This property is known as the oracle property in the literature.

To illustrate the penalization effect of SCAD and MCP, consider a thresholding operator defined as the solution to a penalized univariate linear regression,

$$\hat{\theta}(\lambda, \gamma) = \underset{\theta}{\operatorname{argmin}} \Big\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i\theta)^2 + \rho(\theta; \lambda, \gamma) \Big\}.$$

Denote the univariate least squares solution by $\hat{\theta}_{LS} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$. Define the soft-thresholding operator as $S(t, \lambda) = \operatorname{sgn}(t)(|t| - \lambda)_+$ for $\lambda > 0$ (Donoho and Johnstone (1994)). Here $x_+ = x1\{x \geq 0\}$ denotes the non-negative part of x. Then $\hat{\theta}(\lambda, \gamma)$ has a closed form solution for SCAD and MCP as follows.

$$\begin{aligned}
\text{For } \gamma > 2, \quad \hat{\theta}_{SCAD}(\lambda, \gamma) &= \begin{cases} S(\hat{\theta}_{LS}, \lambda), & |\hat{\theta}_{LS}| \leq 2\lambda, \\ \frac{\gamma-1}{\gamma-2} S(\hat{\theta}_{LS}, \lambda\gamma/(\gamma-1)), & 2\lambda < |\hat{\theta}_{LS}| \leq \gamma\lambda, \\ \hat{\theta}_{LS}, & |\hat{\theta}_{LS}| > \lambda\gamma. \end{cases} \\
\text{For } \gamma > 1, \quad \hat{\theta}_{MCP}(\lambda, \gamma) &= \begin{cases} \frac{\gamma}{\gamma-1} S(\hat{\theta}_{LS}, \lambda), & |\hat{\theta}_{LS}| \leq \lambda\gamma, \\ \hat{\theta}_{LS}, & |\hat{\theta}_{LS}| > \lambda\gamma. \end{cases}
\end{aligned} \tag{1.4}$$

Observe that both SCAD and MCP use the LS solution when $|\hat{\theta}_{LS}| > \lambda\gamma$; MCP only applies a scaled soft-thresholding operation for $|\hat{\theta}_{LS}| \leq \lambda\gamma$ while SCAD applies a soft-thresholding operation to $|\hat{\theta}_{LS}| < 2\lambda$ and a scaled soft-thresholding operation to $2\lambda < |\hat{\theta}_{LS}| \leq \lambda\gamma$.

Figure 1.1 shows the penalty and thresholding functions for Lasso (left panel), SCAD (middle panel) and MCP (right panel), respectively, with the first row showing the penalty functions and the second showing the thresholding operator functions. Lasso penalizes all the variables. SCAD and MCP gradually reduce the rate of penalization for larger coefficients. MCP reduces to Lasso when $\gamma \to +\infty$ and converges to hard-threshold penalty when $\gamma \to 1$. SCAD converges to Lasso when $\gamma \to +\infty$.

1.3 Computation for Penalized Linear Models

Considerable progress has been made on the computational algorithms of penalized regressions. Efron *et al* (2004) introduced the least angle regression (LARS) algorithm, a variant of which can efficiently compute an entire Lasso solution path in a linear regression model. This modified LARS algorithm coincides with the earlier work of homotopy algorithm by Osborne, Presnell and Turlach (2000). Fan and Li (2001) proposed a local quadratic approximation (LQA) algorithm for computing the SCAD solutions. A drawback of LQA is that once a coefficient is set to zero at any iteration step, it permanently stays at zero and the corresponding variable is then removed from final model. Hunter and Li (2005) suggested using the majorization-minimization (MM) algorithm to optimize a perturbed version of LQA by bounding

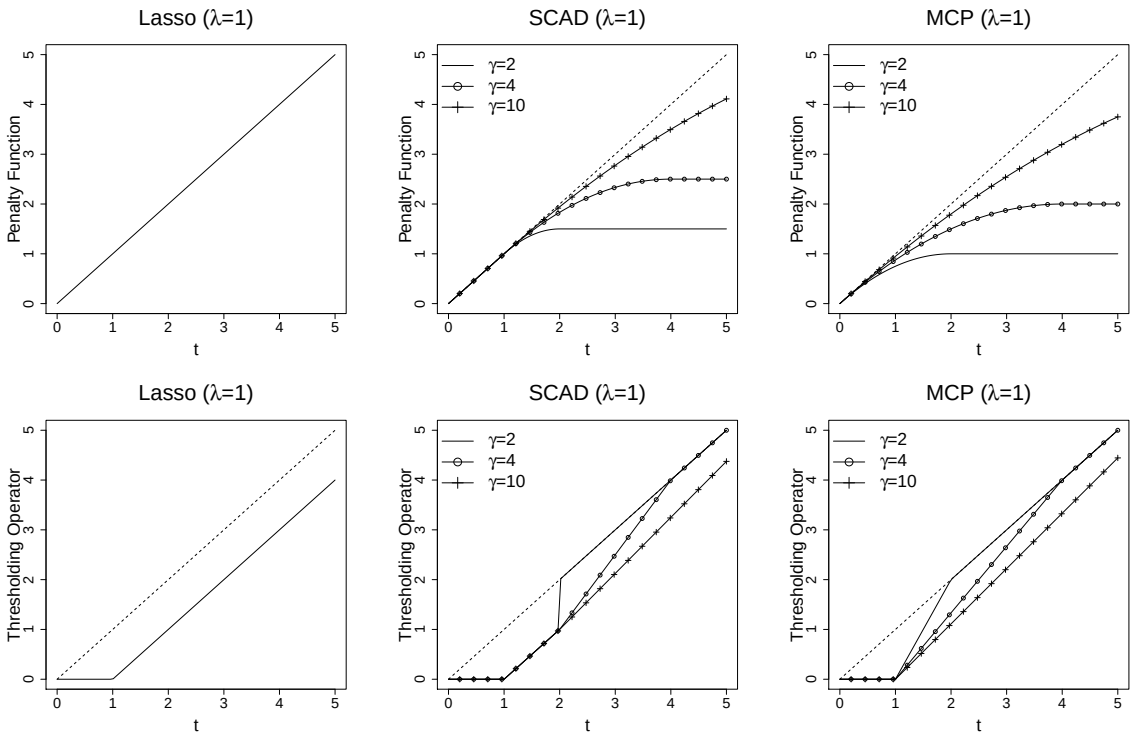

Figure 1.1: Penalty and thresholding operator functions of Lasso (left), SCAD (middle) and MCP (right).

the denominator away from zero. How to choose the size of perturbation and how the perturbation affects the sparsity need to be determined in specific models. Zou and Li (2008) proposed a local linear approximation (LLA) algorithm for computing the concave penalized solutions of SCAD. With LLA, the algorithms designed for Lasso can be repeatedly used to approximate the SCAD solutions. Schifano, Strawderman and Wells (2010) extended the idea of LLA to multiple penalties and proved the convergence properties of their minimization by iterated soft thresholding (MIST) algorithm. Zhang (2010) developed the PLUS algorithm for computing the concave

penalized solutions, including the MCP solutions, in linear models.

In the past few years, the coordinate descent algorithm (CDA) has been recognized as an efficient approach to compute the Lasso solutions in $p \gg n$ models (Friedman, Hastie, Höfling and Tibshirani (2007); Wu and Lange (2008); Friedman, Hastie and Tibshirani (2010)). This algorithm has a long history in applied mathematics and have roots in the Gauss-Siedel method for solving linear systems (Warge (1963); Ortega and Rheinbold (1970); Tseng (2001)). The CDA optimizes an objective function by updating one coordinate (or a group of coordinates) at a time, iteratively cycling through all the coordinates until convergence is reached. It is particularly suitable for the problems that has a closed form solution for each coordinate, but lack one in higher dimensions. CDA for a Lasso penalized linear regression has shown to be very competitive with LARS, especially in high-dimensional cases (Friedman, Hastie, Höfling and Tibshirani (2007); Wu and Lange (2008); Friedman, Hastie and Tibshirani (2010)). The efficiency of CDA may be rooted in three facts: (1) It only takes $O(np)$ operations to cycle through all coordinates; while the algorithms involving in matrix inversion requires $O(np^2)$ operations, whose computational burden increases dramatically when p is large. (2) The closed form solution of each coordinate obtains further efficiency without iterative search. (3) In the computation of a continuous solution surface, if the initial values are properly chosen, then convergence can be reached within a few steps since by continuity, the solution should not be far away from the initial values.

CDA has also been used in computing solutions for models with concave penal-

ties. Breheny and Huang (2011) observed that the CDA converges much faster than the LLA for various combinations of (n, p) and various designs of covariate matrices they considered. Mazumder, Friedman and Hastie (2011) also demonstrated that the CDA has better convergence properties than the LLA. Breheny and Huang (2011) proposed an adaptive rescaling technique to overcome the difficulty due to the constantly changing scaling factors in the computation of GLM models with MCP penalty. However, the adaptive rescaling approach can not be applied to SCAD and the degree of effective concavity applied to the model is unknown until the algorithm is converged.

1.4 Majorization Minimization Algorithm

Majorization Minimization (MM) algorithm is a class of optimization techniques that majorize the objective function when minimizing it. The MM principle can be stated as follows. Let $f(\theta)$ be a real-valued function, then $g(\theta|\theta^m)$ is said to majorize the function $f(\theta)$ at the point θ^m provided

$$\begin{aligned} g(\theta|\theta^m) &\geq f(\theta), \forall\theta, \\ g(\theta^m|\theta^m) &= f(\theta^m). \end{aligned} \tag{1.5}$$

This means the map $\theta \mapsto g(\theta|\theta^m)$ lies above the map $\theta \mapsto f(\theta)$ and is tangent to it at the point of $\theta = \theta^m$.

In general, θ^m is the current estimate in a search for the solution of $f(\theta)$. In a MM algorithm, the optimization is to minimize the majorization function $g(\theta|\theta^m)$ rather than the actual function $f(\theta)$. If θ^{m+1} is the minimizer of $g(\theta|\theta^m)$, it can be

shown that MM approach forces the function $f(\theta)$ move toward its minimum. This can be shown as follows.

$$\begin{aligned} f(\theta^{m+1}) &= g(\theta^{m+1}|\theta^m) + f(\theta^{m+1}) - g(\theta^{m+1}|\theta^m) \\ &\leq g(\theta^m|\theta^m) + f(\theta^m) - g(\theta^m|\theta^m) \\ &= f(\theta^m). \end{aligned} \tag{1.6}$$

This is generally referred as the descent property of MM algorithm.

From the perspective of MM algorithm, the well-known EM algorithm in statistics also falls to this framework. In that case, it is Minorization Maximization with the same acronym MM algorithm. We refer to Lange, Hunter, and Yang (2000); Hunter and Lange (2004) for more details about the MM algorithm.

1.5 Group Lasso for Group Variable Selection

A drawback of the standard concave penalties mentioned above is that they do not consider the structures among predictors. This could potentially affect model selection and prediction. A natural group structure of predictors exists in many applications of regression models. In an ANOVA setting, the dummy variables representing the multi-levels of categorical predictors consist of a group of predictors. In health care studies, the variables measuring the intake of fat-rich foods can be considered as one group. In genomics, the genes involved in the same biological pathway, multiple single-nucleotide polymorphism (SNP) from the same exon can be regarded as members of one group as well. Presumably, the variables in the same group share similar characteristics and tend to be more correlated to each other than the ones outside

the group.

Another situation in which to consider group structure is the case where the effect of predictors in a group is small, but the effect of the entire group is strong. Direct application of an ungrouped penalty could miss these individual-negligible but group-strong predictors. Scientifically, these weak-effect genetic variants are the main contributors of complex diseases such as cardiovascular diseases (CVD) and cancers, which have higher impact at the population level than Mendelian diseases caused by a few rare, highly deleterious genetic mutations.

Group Lasso (Yuan and Lin (2006)) attempts to incorporate the group information into the penalty function. Specifically the group Lasso imposes the ℓ_1 penalty on the Euclidean (ℓ_2) norm of the coefficients of a grouped variables. Meier, van de Geer and Bühlmann (2008) extended the group lasso to a logistic regression to detect the splice site in DNA sequences. The drawback of the Lasso carries over to the group Lasso. It is of great interest to extend the idea to the concave penalties and compare them with the group Lasso.

A distinctive feature of the group Lasso is that the predictors in one group are treated as one unit; they are either selected or dropped at the same time. An advantage of group selection is that it reduces the dimensionality of the model. In other applications, group selection only procedure, however, is not sufficient, particularly when potential null variables exist within the groups. A selection procedure at both group and individual levels would be more appropriate for such applications.

1.6 Overview of the Dissertation

Chapter 2 applies the concave penalties in a generalized linear models (GLM). A majorization minimization by coordinate descent (MMCD) algorithm is specifically designed to compute the solutions for a GLM with concave penalties. The MMCD algorithm seeks closed form solutions for each coordinate by majorizing the loss function. Under reasonable regularity conditions, the convergence property of the MMCD algorithm is established. Simulation studies and a data example show that the MMCD algorithm works sufficiently fast in high-dimensional settings with $p \gg n$.

Chapter 3 proposes an ℓ_2 grouped concave penalty and develops an efficient algorithm based on the (group) coordinate descent approach for computing the ℓ_2 grouped concave penalized solutions. The ℓ_2 grouped concave penalty performs group selection and includes the group Lasso as a special case. Under reasonable conditions, the convergence of the algorithm is established. The ℓ_2 grouped concave penalty is further extended to GLM models and an efficient algorithm sharing similar majorization as MMCD is proposed, whose convergence property is also established. Simulation studies are performed to compare the finite-sample behavior of the ℓ_2 grouped concave penalty and the ungrouped concave penalty.

Chapter 4 proposes an ℓ_1 grouped concave penalty, which selects variables at both group and individual levels. An efficient algorithm is also proposed and the convergence of the algorithm is established under reasonable conditions. The extension of the ℓ_1 grouped concave penally to GLM models is also made in the

chapter. Simulation studies are carried out to evaluate its finite-sample performance.

Chapter 5 applies the ℓ_1 and ℓ_2 grouped concave penalties to two real data examples. The ungrouped and the grouped concave penalties are compared by both simulation studies and real data examples. Chapter 6 concludes the dissertation with remarks and a discussion of future research.

CHAPTER 2
MAJORIZATION MINIMIZATION BY COORDINATE DESCENT ALGORITHM FOR CONCAVE PENALIZATED GLM

Chapter 2 covers the majorization minimization by coordinate descent (MMCD) algorithm in details. Section 2.1 explains the main idea of the MMCD algorithm in computing the solutions for a GLM model with concave penalties. Section 2.2 establishes the theoretical convergence of the MMCD algorithm. Section 2.3 compares the MMCD algorithm to the existing algorithms. Section 2.4 implements the MMCD in a penalized logistic regression model. Computational efficiency of the MMCD algorithm and the adaptive rescaling approach is compared by simulation studies. We also compare the finite-sample performance of Lasso, SCAD and MCP using simulation and real data analyses. Section 2.5 discusses a further extension of the MMCD algorithm in a multinomial regression model.

2.1 Majorization Minimization by Coordinate Descent

Let $\{(y_i, \boldsymbol{x}^i)_{i=1}^n\}$ be the observed data, with $y_i \in \mathbb{R}$ the response and $\boldsymbol{x}^i \in \mathbb{R}^{p+1}$ the predictor vector. A GLM model assumes that y_i depends on $\boldsymbol{x}^i$ through a linear combination $\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}^i$, with $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T \in \mathbb{R}^{p+1}$. The density function of y_i given $\boldsymbol{x}^i$ is

$$f_i(y_i) = \exp\{\frac{y_i\theta_i - \psi(\theta_i)}{\phi_i} + c(y_i, \phi_i)\}. \tag{2.1}$$

Here $\phi_i > 0$ is a dispersion parameter. The form of $\psi(\theta)$ depends on the specified model. For example, in a logistic regression, $\psi(\theta) = \log(1 + \exp(\theta))$.

Consider the (scaled) negative log-likelihood as a loss function $\ell(\boldsymbol{\beta})$, under the canonical link function with $\theta_i = \eta_i$, we have

$$\ell(\boldsymbol{\beta}) \propto \frac{1}{n}\sum_{i=1}^{n}\{\psi(\boldsymbol{\beta}^T\boldsymbol{x}^i) - y_i\boldsymbol{\beta}^T\boldsymbol{x}^i\}. \tag{2.2}$$

Throughout the chapter, we assume $\{(x_{i0})_{i=1}^n = 1\}$ and β_0 is the intercept. The rest p variables are assumed to be (column-wise) standardized, i.e. $\|\boldsymbol{x}_j\|_2^2/n = 1$ with $\boldsymbol{x}_j = (x_{1j}, ..., x_{nj})^T, 1 \leq j \leq p$. The standardization applies a fair penalization to each variable regardless of its scale. Then we define the concave penalized criterion as

$$Q(\boldsymbol{\beta}; \lambda, \gamma) = \frac{1}{n}\sum_{i=1}^{n}\{\psi(\boldsymbol{\beta}^T\boldsymbol{x}^i) - y_i\boldsymbol{\beta}^T\boldsymbol{x}^i\} + \sum_{j=1}^{p}\rho(|\beta_j|; \lambda, \gamma). \tag{2.3}$$

Here we do not penalize the intercept.

Applying a quadratic approximation to $\ell(\boldsymbol{\beta})$ for a given estimation $\tilde{\boldsymbol{\beta}}$ leads to the iteratively reweighed least squares (IRLS) form of the loss function,

$$\ell(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}) = \frac{1}{2n}\sum_{i=1}^{n} w_i(z_i - \boldsymbol{\beta}^T\boldsymbol{x}^i)^2, \tag{2.4}$$

with $w_i(\tilde{\boldsymbol{\beta}}) = \ddot{\psi}(\tilde{\boldsymbol{\beta}}^T\boldsymbol{x}^i)$ and $z_i(\tilde{\boldsymbol{\beta}}) = \ddot{\psi}(\tilde{\boldsymbol{\beta}}^T\boldsymbol{x}^i)^{-1}\{y_i - \dot{\psi}(\tilde{\boldsymbol{\beta}}^T\boldsymbol{x}^i)\} + \tilde{\boldsymbol{\beta}}^T\boldsymbol{x}^i$, where $\dot{\psi}(\theta)$ and $\ddot{\psi}(\theta)$ are the first and second derivatives of $\psi(\theta)$ with respect to (w.r.t.) θ. Let $\hat{\boldsymbol{\beta}}_j^m = (\hat{\beta}_0^{m+1}, ..., \hat{\beta}_j^{m+1}, \hat{\beta}_{j+1}^m, ..., \hat{\beta}_p^m)^T$. For the loss function (2.4), CDA updates $\hat{\boldsymbol{\beta}}_{j-1}^m$ to $\hat{\boldsymbol{\beta}}_j^m$ by minimizing the following criterion

$$\begin{aligned}
\hat{\beta}_j^{m+1} &= \underset{\beta_j}{\operatorname{argmin}}\, Q(\beta_j|\hat{\boldsymbol{\beta}}_{j-1}^m) \\
&= \underset{\beta_j}{\operatorname{argmin}} \frac{1}{2n}\sum_{i=1}^{n} w_i(z_i - \sum_{s<j} x_{ij}\hat{\beta}_s^{m+1} - x_{ij}\beta_j - \sum_{s>j} x_{ij}\hat{\beta}_s^m)^2 \\
&+ \rho(|\beta_j|; \lambda, \gamma),
\end{aligned} \tag{2.5}$$

which treats $\beta_k, k \neq j$ as fixed values with w_i and z_i depending on $(\hat{\boldsymbol{\beta}}_{j-1}^m, \boldsymbol{x}^i, y_i)$. Then the jth coordinate-wise minimizer is obtained by solving the equation,

$$\frac{1}{n}\sum_{i=1}^{n} w_i x_{ij}^2 \beta_j + \rho^{'}(|\beta_j|)\text{sgn}(\beta_j) - \frac{1}{n}\sum_{i=1}^{n} w_i x_{ij}(z_i - (\boldsymbol{x}^i)^T \hat{\boldsymbol{\beta}}_{j-1}^m) - \frac{1}{n}\sum_{i=1}^{n} w_i x_{ij}^2 \hat{\beta}_j^m = 0, \tag{2.6}$$

where $\rho^{'}(|t|)$ is the first derivative of $\rho(|t|)$ w.r.t. $|t|$ and $\text{sgn}(x) = 1, -1$ or $\in [-1, 1]$ for $x > 0, < 0$ or $x = 0$.

For the MCP penalty, directly solving (2.6) gives

$$\hat{\beta}_j^{m+1} = \frac{S(\tau_j, \lambda)}{\delta_j - 1/\gamma}\mathbf{1}_{\{|\tau_j| \leq \delta_j \gamma \lambda\}} + \frac{\tau_j}{\delta_j}\mathbf{1}_{\{|\tau_j| > \delta_j \gamma \lambda\}}, \tag{2.7}$$

where $\delta_j = n^{-1}\sum_{i=1}^n w_i x_{ij}^2$ and $\tau_j = n^{-1}\sum_{i=1}^n w_i x_{ij}(z_i - (\boldsymbol{x}^i)^T \hat{\boldsymbol{\beta}}_{j-1}^m) + \delta_j \hat{\beta}_j^m$. In a linear regression, $w_i = 1$ for $i = 1, ..., n$, thus the scaling factor $\delta_j \triangleq n^{-1}\sum_{i=1}^n w_i x_{ij}^2 = 1$ for standardized predictors. In a GLM, however, the dependence of w_i on $(\hat{\boldsymbol{\beta}}_{j-1}^m, \boldsymbol{x}^i, y_i)$ causes the scaling factor δ_j to change from iteration to iteration. This is problematic because $\delta_j - 1/\gamma$ can be very small and is not guaranteed to be positive. Thus direct application of CDA is not numerically stable and can lead to unreasonable solutions.

To overcome the difficulty, Breheny and Huang (2011) proposed an adaptive rescaling approach, which uses

$$\hat{\beta}_j^{m+1} = \frac{S(\tau_j, \lambda)}{\delta_j(1 - 1/\gamma)}\mathbf{1}_{\{|\tau_j| \leq \gamma \lambda\}} + \frac{\tau_j}{\delta_j}\mathbf{1}_{|\tau_j| > \gamma \lambda\}}, \tag{2.8}$$

for the jth coordinate-wise solution. This is equivalent to apply a new regularization parameter $\gamma^* = \gamma/\delta_j$. The effective regularization parameters are not the same across the penalized variables and not known until the algorithm is converged. Numerically,

the computation of δ_j is not desirable when p is large. Furthermore, the adaptive rescaling cannot be adopted for SCAD because the scaled soft-thresholding operation only applies to the middle clauses for SCAD as show in equation (1.4).

The MMCD algorithm seeks a majorization of the scaling factors δ_j. For standardized predictors, this is equivalent to finding a uniform upper bound of the weights $w_i = \ddot{\psi}(\boldsymbol{\beta}^T \boldsymbol{x}^i), 1 \leq i \leq n$. In principle, we can have a sequence of constants C_i such that $C_i \geq w_i$ and use $M_j = \sum C_i x_{ij}^2/n$ to majorize the scaling factors δ_j. Due to the standardization, we need only a single M to majorize all the p scaling factors. Note that in a GLM model, $\nabla_j^2 \ell(\boldsymbol{\beta}) = \sum \ddot{\psi}(\boldsymbol{\beta}^T \boldsymbol{x}^i) x_{ij}^2/n = \sum w_i x_{ij}^2/n$. Hence, a majorization of w_i results in the majorization of $\nabla_j^2 \ell(\boldsymbol{\beta})$. For simplicity, we put the boundedness condition, $\delta_j \leq M$ on the term $\nabla_j^2 \ell(\boldsymbol{\beta})$ rather than the individual w_i.

From the perspective of the MM algorithm, the majorization of the scaling factors δ_j is equivalent to use a surrogate function $\ell^s(\beta_j | \hat{\boldsymbol{\beta}}_{j-1}^m)$, with

$$\ell^s(\beta_j | \hat{\boldsymbol{\beta}}_{j-1}^m) = \ell(\hat{\boldsymbol{\beta}}_{j-1}^m) + \nabla_j \ell(\hat{\boldsymbol{\beta}}_{j-1}^m)(\beta_j - \hat{\beta}_j^m) + \frac{1}{2} M (\beta_j - \hat{\beta}_j^m)^2, \tag{2.9}$$

when optimizing $\ell(\boldsymbol{\beta})$ w.r.t jth coordinate, where the second partial derivative $\nabla_j^2 \ell(\boldsymbol{\beta})$ in the Taylor expansion is replaced by its upper bound M. Observe that the majorization is applied coordinate-wisely to fit the CDA. The descent property of the MM algorithm ensures that iteratively minimizing $\ell^s(\beta_j | \hat{\boldsymbol{\beta}}_{j-1}^m)$ leads to a descent sequence of the loss function $\ell(\boldsymbol{\beta})$. More details about the MM algorithm can be found in Lange, Hunter, and Yang (2000); Hunter and Lange (2004).

The coordinate-wise solutions for the jth penalized variable using the MMCD

approach are

$$\text{SCAD: } \hat{\beta}_j^{m+1} = \begin{cases} \frac{1}{M}S(\tau_j, \lambda), & |\tau_j| \leq (1+M)\lambda, \\ \frac{S(\tau_j, \gamma\lambda/(\gamma-1))}{M-1/(\gamma-1)}, & (1+M)\lambda < |\tau_j| \leq M\gamma\lambda, \\ \frac{1}{M}\tau_j & |\tau_j| > M\gamma\lambda, \end{cases} \tag{2.10}$$

$$\text{MCP: } \hat{\beta}_j^{m+1} = \begin{cases} \frac{S(\tau_j, \lambda)}{M-1/\gamma} & |\tau_j| \leq M\gamma\lambda, \\ \frac{1}{M}\tau_j & |\tau_j| > M\gamma\lambda, \end{cases} \tag{2.11}$$

with $\tau_j = M\hat{\beta}_j^m + n^{-1}\sum_{i=1}^n x_{ij}(y_i - \dot{\psi}((\boldsymbol{x}^i)^T\hat{\boldsymbol{\beta}}_{j-1}^m))$. The solution of the intercept is

$$\hat{\beta}_0^{m+1} = \tau_0/M, \tag{2.12}$$

with $\tau_0 = M\hat{\beta}_0^m + n^{-1}\sum_{i=1}^n x_{i0}(y_i - \dot{\psi}((\boldsymbol{x}^i)^T\hat{\boldsymbol{\beta}}^m))$ and $\hat{\boldsymbol{\beta}}^m = (\hat{\beta}_0^m, \hat{\beta}_1^m, ..., \hat{\beta}_p^m)^T$. We want to ensure the denominators in (2.10) and (2.11) are positive, that is, $M - 1/(\gamma-1) > 0$ and $M - 1/\gamma > 0$. This naturally leads to the constraint of the penalty, $\inf_t \rho''(|t|; \lambda, \gamma) > -M$, where $\rho''(|t|; \lambda, \gamma)$ is the second derivative of $\rho(|t|; \lambda, \gamma)$ w.r.t $|t|$. For SCAD and MCP, this condition is satisfied by choosing a proper γ. For SCAD, $\inf_t \rho''(|t|; \lambda, \gamma) = -1/(\gamma-1)$; for MCP, $\inf_t \rho''(|t|; \lambda, \gamma) = -1/\gamma$. Therefore, we require $\gamma > 1 + 1/M$ for SCAD and $\gamma > 1/M$ for MCP.

The MMCD algorithm can gain further efficiency by adopting the following tip. Let $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^T$ and $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$, and $\hat{\boldsymbol{\eta}}_j^m = X\hat{\boldsymbol{\beta}}_j^m$ be the linear component corresponding to $\hat{\boldsymbol{\beta}}_j^m$. Observe that

$$\hat{\boldsymbol{\eta}}_{j+1}^m = \hat{\boldsymbol{\eta}}_j^m + \boldsymbol{x}_{j+1}(\hat{\beta}_{j+1}^{m+1} - \hat{\beta}_{j+1}^m) = \hat{\boldsymbol{\eta}}_j^m + (\hat{\boldsymbol{\beta}}_{j+1}^m - \hat{\boldsymbol{\beta}}_j^m)\boldsymbol{x}_{j+1}. \tag{2.13}$$

This equation turns a $O(np)$ operation into a $O(n)$ one. Since this step is involved in each iteration for each coordinate, this simple step turns out to be significant in reducing the computational cost.

We summarize the MMCD algorithm for a given (λ, γ) as follows by assuming the conditions below hold.

- Condition (i): the second partial derivative of $\ell(\boldsymbol{\beta})$ w.r.t β_j is uniformly bounded for standardized X, i.e. there exists a real number $M > 0$ such that $\nabla_j^2 \ell(\boldsymbol{\beta}) \leq M$ for $j = 0, ..., p$.
- Condition (ii): $\inf_t \rho''(|t|; \lambda, \gamma) > -M$, with $\rho''(|t|; \lambda, \gamma)$ being the second derivative of $\rho(|t|; \lambda, \gamma)$ w.r.t $|t|$.

Algorithm 2.1 MMCD Algorithm for Concave Penalty in GLM

1. Given an initial value $\hat{\boldsymbol{\beta}}^0$, compute the corresponding linear component $\hat{\boldsymbol{\eta}}^0$.
2. For $m = 0, 1, ...,$ update $\hat{\boldsymbol{\beta}}_j^m$ to $\hat{\boldsymbol{\beta}}_{j+1}^m$ by using the solution form of (2.10) or (2.11) for the penalized variables and (2.12) for the intercept. After each iteration, also compute the corresponding linear component $\hat{\boldsymbol{\eta}}_{j+1}^m$ using (2.13). Cycle through all the coordinates from $j = 0, ..., p$ such that $\hat{\boldsymbol{\beta}}^m$ is updated to $\hat{\boldsymbol{\beta}}^{m+1}$.
3. Check the convergence criterion. If converges then stop iterations, otherwise repeat step 2 until converges.

The convergence criterion we use is $\|\hat{\boldsymbol{\beta}}^{m+1} - \hat{\boldsymbol{\beta}}^m\|_2 / (\|\hat{\boldsymbol{\beta}}^m\|_2 + \delta) < \varepsilon$, with $\varepsilon = 0.001$ and $\delta = 0.01$.

2.2 Convergence Analysis

Theorem 2.1 establishes that under certain regularity conditions, the MMCD algorithm always converges to a minimum of the criterion function.

Theorem 2.1. *Consider the criterion function (2.3), where the given data* $(\boldsymbol{y}, X)$ *lies on a compact set and no two columns of* X *are identical. Suppose the penalty* $\rho(|t|; \lambda, \gamma) \equiv \rho(t)$ *satisfies* $\rho(t) = \rho(-t)$, $\rho^{'}(|t|)$ *is non-negative, uniformly bounded, with* $\rho^{'}(|t|)$ *being the first derivative (assuming existence) of* $\rho(|t|)$ *w.r.t.* $|t|$. *Also assume that conditions (i) and (ii) of the MMCD algorithm hold.*

Then the sequence $\{\boldsymbol{\beta}^m\}$ *generated by the MMCD algorithm converges to a minimum of the function* $Q(\boldsymbol{\beta})$.

Note that the condition on $(\boldsymbol{y}, X)$ is a mild assumption. The standardization of columns of X can be performed as long as the columns are not identically zero. The proof of theorem (2.1) is provided in Appendix A. It extends the work of Mazumder, Friedman and Hastie (2011) to cover more general loss functions other than the least squares.

2.3 Comparison with Other Algorithms

The LQA (Fan and Li (2001)), perturbed LQA (Hunter and Li (2005)), LLA (Zou and Li (2008)) and MIST (Schifano, Strawderman and Wells (2010)) algorithms share the same spirit in the sense that they all optimize a majorization function instead of the original penalty $\rho(|t|; \lambda, \gamma)$. Figure 2.1 illustrates the three majorizations of SCAD. The left panel of Figure 2.1 is majorized at $t = 3$, while the

right one is majorized at $t = 1$. For perturbed LQA, we choose $\tau_0 = 0.5$. In both plots, $\gamma = 4$ and $\lambda = 2$ are chosen for better illustration purpose.

To apply these methods to GLM, we need to approximate both the loss and penalty. This does not take full advantage of CDA. Indeed, the approximation of the penalty requires additional iterations for convergence and is not necessary, since exact coordinate-wise solution exists. Thus MMCD uses the exact form of the penalty and only majorizes the loss to avoid the computation of the scaling factors. Breheny and Huang (2011) reported that the adaptive rescaling technique is at least 100 times faster than the LLA approach. Therefore, we focus on the comparison between the MMCD algorithm and the adaptive rescaling approach.

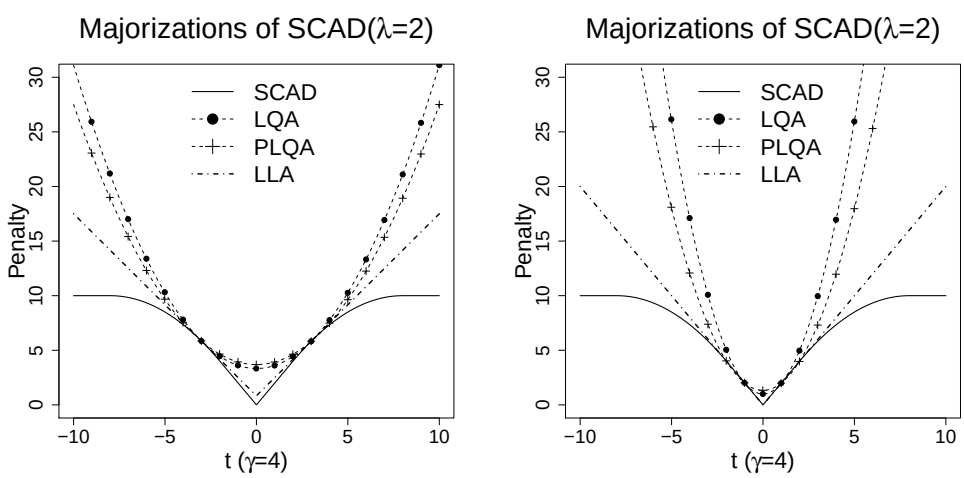

Figure 2.1: SCAD penalty and its majorizations, LQA, Perturbed LQA (PLQA) and LLA.

2.4 The MMCD for Logistic Regression

We implement the MMCD algorithm in the logistic regression, one of the most widely used models in biostatistical applications. Then $\boldsymbol{y}$ is a vector of 0 or 1 with 1 indicating the event of interest. Observe that $\nabla_j \ell(\hat{\boldsymbol{\beta}}) = -(\boldsymbol{x}_j)^T(\boldsymbol{y} - \hat{\boldsymbol{\pi}})/n$ and $\nabla_j^2 \ell(\hat{\boldsymbol{\beta}}) = n^{-1} \sum w_i x_{ij}^2$, with $w_i = \hat{\pi}_i(1 - \hat{\pi}_i)$ and $\hat{\pi}_i$ being the estimated probability of ith observation given estimate $\hat{\boldsymbol{\beta}}$, i.e. $\hat{\pi}_i = 1/(1 + \exp(-\hat{\boldsymbol{\beta}}^T \boldsymbol{x}^i))$. For any $0 \leq \pi \leq 1$, we have $\pi(1 - \pi) \leq 1/4$. Hence the upper bound of $\nabla_j^2 \ell(\hat{\boldsymbol{\beta}})$ is $M = 1/4$ for standardized $\boldsymbol{x}_j$. Correspondingly $\tau_j = 4^{-1}\hat{\beta}_j + n^{-1}(\boldsymbol{x}_j)^T(\boldsymbol{y} - \hat{\boldsymbol{\pi}})$ for $j = 0, ..., p$. By condition (ii), we require $\gamma > 5$ for SCAD and $\gamma > 4$ for MCP.

2.4.1 Computation of solution surface

A common practice in applying the SCAD and MCP penalties is to compute a solution path along a sequence of λ for a fixed value of γ. For example, in a linear regression, it has been suggested one uses $\gamma \approx 3.7$ in SCAD (Fan and Li (2001)) and $\gamma \approx 2.7$ (Zhang (2010)) in MCP. However, in a GLM including the logistic regression, these values may not be appropriate. Therefore, We use a data driven procedure to choose γ together with λ. This requires the computation of solution surface over a two-dimensional grids of (λ, γ). We reparameterize $\kappa = 1/\gamma$ to facilitate the description of the approach for computing the solution surface. By condition (ii) of MMCD algorithm, we have $\kappa \in [0, \kappa_{\max}]$, with $\kappa_{\max} = 1/5$ for SCAD and $\kappa_{\max} = 1/4$ for MCP. Note that when $\kappa = 0$, both SCAD and MCP simplify to Lasso.

Define the grid values for a rectangle in $[0, \kappa_{\max}) \times [\lambda_{\min}, \lambda_{\max}]$ to be $0 = \kappa_1 \leq \kappa_2 \leq \cdots \leq \kappa_K < \kappa_{\max}$ and $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_V = \lambda_{\min}$. The number of grid points, K and V are pre-specified. In our implementation, the κ points are uniform in normal scale while those for λ are uniform in log scale. The $\lambda_{\max}$ is the smallest value of λ such that $\hat{\beta}_j = 0, j = 1, ..., p$. For logistic regression, $\lambda_{\max} = n^{-1}\max_j|(\boldsymbol{x}_j)^T(\boldsymbol{y} - \hat{\boldsymbol{\pi}})|$ with $\hat{\boldsymbol{\pi}} = \bar{y}\boldsymbol{J}$ and $\boldsymbol{J}$ being a unit vector, for every κ_k. We let $\lambda_{\min} = \epsilon\lambda_{\max}$, with $\epsilon = 0.0001$ if $n > p$ and $\epsilon = 0.01$ otherwise. The solution surface is then calculated over the rectangle $[0, \kappa_{\max}) \times [\lambda_{\min}, \lambda_{\max}]$. Denote the MMCD solution for a given (κ_k, λ_v) as $\hat{\boldsymbol{\beta}}_{\kappa_k, \lambda_v}$. We follow the approach of Mazumder, Friedman and Hastie (2011) to compute the solution surface by initializing the algorithm at Lasso solutions along a grid of λ values. Then for each λ, we compute the solutions along the grid of κ starting from $\kappa = 0$, using the solution at the previous point as the initial value for the current point. Algorithm 2.2 details the approach.

Algorithm 2.2 Computation of Solution Surface for Concave Penalty in GLM

1. First compute the Lasso solution along λ. When computing $\hat{\boldsymbol{\beta}}_{\kappa_0, \lambda_{v+1}}$, using $\hat{\boldsymbol{\beta}}_{\kappa_0, \lambda_v}$ as the initial value in the MMCD algorithm.
2. For a given λ_v, compute the solution along κ. That is using $\hat{\boldsymbol{\beta}}_{\kappa_k, \lambda_v}$ as the initial value to compute the solution $\hat{\boldsymbol{\beta}}_{\kappa_{k+1}, \lambda_v}$.
3. Cycle through $v = 1, ..., V$ for step (2) to complete the solution surface.

Define a variable to be a causal predictor if $\beta_j \neq 0$; otherwise define it to be a null predictor. Figure 2.2 presents the solution paths of a causal predictor (plot a) and a null predictor (plot b) along κ using the MCP penalty. Observe that although Lasso tends to over-select in some cases, it could fail to select certain variables, which are selected by MCP (dash line in plot a). This could be a serious problem for Lasso if the missing predictor is a causal one. Furthermore, we observe that the estimates change substantially when κ crosses certain threshold. This justifies our treatment of κ as a tuning parameter and usage of a data-driven procedure to choose both κ and λ, since a pre-specified κ might not give an optimal result.

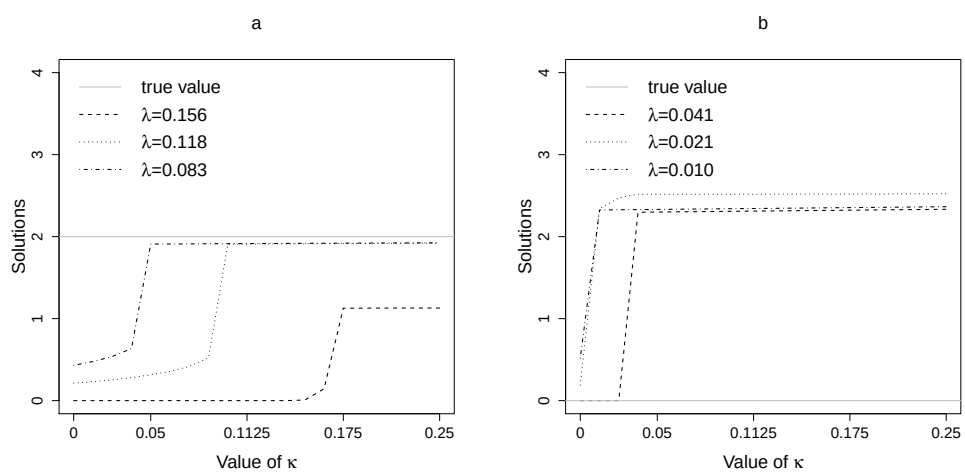

Figure 2.2: Solution paths along κ for a causal variable (plot a), and a null variable (plot b).

2.4.2 Design of simulation study

The binary outcome y_i is generated from the model $\text{logit}(\pi(y_i = 1)) = \boldsymbol{\beta}^T \boldsymbol{x}^i$. The non-intercept p covariates $(x_{i1}, ..., x_{ip})^T \sim \text{MVN}(\mathbf{0}, \sigma^2 \Sigma_{p\times p})$, with $\sigma^2 = 1$. Denote the causal and the null variable sets as $A_0 \equiv \{j : \beta_j \neq 0, 1 \leq j \leq p\}$ with dimension p_0, and $A_1 \equiv \{j : \beta_j = 0, 1 \leq j \leq p\}$ with dimension p_1, respectively. We set $\beta_0 = 0.01$ and $p_0 = 10$ with coefficients $(0.6, -0.6, 1.2, -1.2, 2.4, -0.6, 0.6, -1.2, 1.2, -2.4)^T$ such that the signal-to-noise ratio (SNR), defined as $\text{SNR} = \sqrt{\boldsymbol{\beta}^T X^T X \boldsymbol{\beta}/n\sigma^2}$, is approximately in the range of $(3, 4)$. In the simulation, $p = 1,000$, $K = 20$ and $V = 100$. Two sample size are studied with $n = 100$ and $n = 300$.

We consider five types of Σ with $\rho = 0.5$ to present a median level of correlation,

1. Independent structure (IN), i.e. $\Sigma = I_p$, with I_p being the identity matrix of dimension $p \times p$.
2. Separate structure (SP), i.e. the causal and the null variables are independent, $\Sigma = \text{block diagonal}(\Sigma_0, \Sigma_1)$, with Σ_i being the covariance matrix of predictors in $A_i, i = 0, 1$. For Σ_i, we assume a compound symmetry structure with $\rho(x_{ij}, x_{ik}) = \rho$ for $j \neq k$, $j, k \in A_i, i = 0, 1$.
3. Partial Correlated structure (PC), i.e. part of the causal variables are correlated with part of the null variables. Specifically, $\Sigma = \text{block diagonal}(\Sigma_a, \Sigma_b, \Sigma_c)$, with Σ_a being the covariance matrix for the first 5 causal variables; Σ_b being the covariance matrix for the remaining 5 causal variables and 5 null variables; Σ_c being the covariance matrix of the remaining null variables. A compound

symmetry structure is assumed within Σ_s, $s = a, b, c$.

4. First-order Autoregressive structure (AR), that is $\rho(x_{ij}, x_{ik}) = \rho^{(|j-k|)}$, for $j \neq k, j, k = 1, ..., p$.

5. Compound Symmetry structure (CS) for all the p penalized variables.

2.4.3 Comparison of computational efficiency

Table 2.1 compares the adaptive rescaling approach and the MMCD algorithm in terms of the average time and standard error (SE) of computing the whole solution surface for the MCP penalty. The computation is done on an Inter Xeon CPU (E5440@2.83GHZ) machine with Ubuntu system (Linux version 2.6). The time is measured in seconds with 100 replicates for train datasets with $n = 100$ and $p = 1,000$. In all the models explored, the MMCD is twice faster than the adaptive rescaling approach. We expect that the MMCD will gain more efficiency when p gets larger.

Table 2.1: Comparison of computational efficiency between the adaptive rescaling and MMCD algorithms in MCP penalized logistic regressions, $n = 100$ and $p = 1,000$.

Algorithm	IN(SE)	SP(SE)	PC(SE)	AR(SE)	CS(SE)
SNR	4.34(0.03)	3.10(0.02)	3.90(0.02)	3.19(0.02)	3.05(0.02)
MMCD	188.64(0.89)	105.98(0.56)	107.78(0.49)	119.51(0.55)	107.08(0.42)
Adap Resca	374.18(1.38)	201.93(1.10)	206.11(1.14)	223.13(1.13)	206.74(1.40)

2.4.4 Comparison of Lasso, SCAD and MCP

When comparing the three penalties, we want to remove the interference introduced by the tuning parameter selection processes. Hence, three penalties are compared based on the models with the best predictive performance rather than the models chosen by any tuning parameter selection approach. To fulfill this purpose, each training dataset is generated together with a validation set with $n^* = 2,000$ observations. Based on the training dataset, we compute the solution $\hat{\boldsymbol{\beta}}_{\kappa_k,\lambda_v}$ by the MMCD algorithm. Then, we calculate the predictive Area Under ROC Curve (PAUC) for the validation set, $AUC_{(\kappa_k,\lambda_v)}$ for each $\hat{\boldsymbol{\beta}}_{\kappa_k,\lambda_v}$. The well-known connection between AUC and the Mann-Whitney U statistics, Bamber (1975) is used in computing the AUC:

$$AUC = max\left\{1 - \frac{U_1}{n_1 n_2}, \frac{U_1}{n_1 n_2}\right\},$$

with $U_1 = R_1 - (n_1(n_1+1)/2)$, where n_1 is the number of observations with outcome $y_i = 1$ in the validation set, R_1 is the sum of ranks for the observations with $y_i = 1$ in the validation set. The rank is based on the predictive probability of validation samples with $\hat{\boldsymbol{\pi}}_(\kappa_k, \lambda_v)$ computed from $\hat{\boldsymbol{\beta}}_{\kappa_k,\lambda_v}$. The solution $\hat{\boldsymbol{\beta}}_{\kappa_k,\lambda_v}$ corresponding to the maximum predictive $AUC_{(\kappa_k,\lambda_v)}$ is selected as the final model for comparison.

Denote the estimated causal variables set as $\hat{A}_0 = \{l : \hat{\beta}_l \neq 0, l = 1, ..., p\}$ with dimension $\hat{p}_0$. Denote the set of false positive causal variables as $F = \{l : \hat{\beta}_l \neq 0 \,\&\, \beta_l = 0, l = 1, ..., p\}$ with dimension $\hat{s}$. Define model size (MS) as MS $= \hat{p}_0$, correct size (CS) as CS $= \hat{p}_0 - \hat{s}$ and false discovery rate (FDR) as FDR $= \hat{s}/\hat{p}_0$. We compare the results in terms of MS, FDR and the maximum PAUC of the validation dataset.

Table 2.2 and 2.3 present the average and standard error of MS, FDR and PAUC based on $1,000$ replicates for $n = 100$ and $n = 300$. As a selection tool, Lasso seems to be inferior to the concave penalties in the sense that it favors a larger model size, with a lower PAUC and a higher FDR. This observation is consistent with the theoretical results from Zhang and Huang (2008). Our results also suggest that SCAD has a similar PAUC but with a slightly larger MS and a higher FDR, compared to MCP.

Table 2.2: Comparison of Lasso, SCAD and MCP in terms of model size (MS), false discover rate (FDR) and predictive AUC (PAUC), $n = 100$, $p = 1,000$.

Structure	Penalty	SNR	MS(SE)	FDR(SE*10^2)	PAUC(SE*10^2)
IN	Lasso	4.33	13.63 (0.34)	0.5908 (0.67)	0.8315 (0.12)
	SCAD		9.92 (0.25)	0.4451 (0.78)	0.8558 (0.10)
	MCP		7.95 (0.23)	0.3143 (0.88)	0.8562 (0.10)
SP	Lasso	3.07	18.40 (0.47)	0.6841 (0.63)	0.7712 (0.17)
	SCAD		7.14 (0.19)	0.3942 (0.73)	0.8177 (0.12)
	MCP		6.10 (0.17)	0.2983 (0.76)	0.8185 (0.12)
PC	Lasso	3.88	8.60 (0.22)	0.4330 (0.67)	0.8726 (0.06)
	SCAD		6.30 (0.14)	0.3311 (0.68)	0.8806 (0.05)
	MCP		5.78 (0.13)	0.2743 (0.69)	0.8807 (0.05)
AR	Lasso	3.20	6.01 (0.15)	0.4774 (0.79)	0.8182 (0.12)
	SCAD		4.83 (0.13)	0.3497 (0.87)	0.8391 (0.09)
	MCP		3.78 (0.11)	0.2214 (0.81)	0.8394 (0.09)
CS	Lasso	3.05	17.72 (0.49)	0.6792 (0.63)	0.7723 (0.16)
	SCAD		8.70 (0.28)	0.4468 (0.83)	0.8086 (0.15)
	MCP		7.32 (0.25)	0.3481 (0.89)	0.8098 (0.14)

Table 2.3: Comparison of Lasso, SCAD and MCP in terms of model size (MS), false discover rate (FDR) and predictive AUC (PAUC), $n = 300$, $p = 1,000$.

Structure	Penalty	SNR	MS(SE)	FDR(SE*10^2)	PAUC(SE*10^2)
IN	Lasso	4.32	31.66 (0.34)	0.7111 (0.28)	0.9190 (0.04)
	SCAD		13.75 (0.24)	0.3137 (0.71)	0.9340 (0.03)
	MCP		11.63 (0.21)	0.2183 (0.71)	0.9340 (0.03)
SP	Lasso	3.06	29.60 (0.36)	0.7118 (0.31)	0.8761 (0.05)
	SCAD		11.41 (0.21)	0.2902 (0.62)	0.8942 (0.04)
	MCP		10.26 (0.20)	0.2220 (0.63)	0.8941 (0.04)
PC	Lasso	3.88	34.25 (0.43)	0.7407 (0.42)	0.9074 (0.04)
	SCAD		11.41 (0.23)	0.2953 (0.68)	0.9230 (0.04)
	MCP		10.68 (0.22)	0.2450 (0.70)	0.9231 (0.04)
AR	Lasso	3.21	16.96 (0.53)	0.5810 (0.80)	0.8558 (0.03)
	SCAD		16.54 (0.34)	0.5324 (0.74)	0.8879 (0.05)
	MCP		13.75 (0.31)	0.4447 (0.82)	0.8875 (0.06)
CS	Lasso	3.06	29.21 (0.34)	0.7093 (0.30)	0.8773 (0.05)
	SCAD		12.83 (0.25)	0.3258 (0.73)	0.8926 (0.05)
	MCP		11.37 (0.23)	0.2566 (0.73)	0.8927 (0.04)

2.4.5 Comparison under misspecification

Misspecification happens when the fitted model is different from the underlying generating model. In applications, the robustness to misspecification is desirable since the true underlying model is unknown to the analysts. We compare the selection performance of Lasso, SCAD and MCP under misspecified models. In our simulation, the data is generated from the probit model, i.e. $P(y_i = 1) = \Phi(\boldsymbol{\beta}^T \boldsymbol{x}^i)$, with Φ being the CDF function of the standard normal distribution. The fitted model is, however still the logistic model. We use the same design as subsection (2.4.2) to generate the covariates.

The behavior of Lasso, SCAD and MCP under misspecification is similar to the case when the model is correct specified. Table 2.4 and 2.5 show the average and standard error of MS, FDR and PAUC based on $1,000$ replicates for $n = 100$ and $n = 300$. Lasso seems to prefer a larger model size with a lower PAUC and a higher FDR. SCAD and MCP have similar selection results with MCP being a better tool.

Table 2.4: Comparison of Lasso, SCAD and MCP in terms of model size (MS), false discover rate (FDR) and predictive AUC (PAUC) under misspecification, $n = 100$, $p = 1,000$.

Structure	Penalty	SNR	MS(SE)	FDR(SE*10^2)	PAUC(SE*10^2)
IN	Lasso	4.33	16.36 (0.34)	0.6269 (0.59)	0.8698 (0.10)
	SCAD		10.41 (0.22)	0.4083 (0.80)	0.8986 (0.09)
	MCP		8.92 (0.21)	0.3120 (0.86)	0.8990 (0.09)
SP	Lasso	3.07	18.76 (0.39)	0.6702 (0.57)	0.8396 (0.12)
	SCAD		8.56 (0.17)	0.3726 (0.74)	0.8813 (0.09)
	MCP		7.71 (0.16)	0.3083 (0.76)	0.8812 (0.09)
PC	Lasso	3.88	9.67 (0.23)	0.4398 (0.71)	0.9084 (0.05)
	SCAD		7.15 (0.14)	0.3270 (0.68)	0.9184 (0.05)
	MCP		6.56 (0.13)	0.2761 (0.69)	0.9185 (0.05)
AR	Lasso	3.20	5.88 (0.13)	0.4595 (0.75)	0.8685 (0.08)
	SCAD		4.60 (0.11)	0.3047 (0.83)	0.8847 (0.05)
	MCP		3.78 (0.09)	0.2065 (0.74)	0.8847 (0.04)
CS	Lasso	3.05	17.42 (0.39)	0.6509 (0.63)	0.8415 (0.12)
	SCAD		9.54 (0.21)	0.4109 (0.80)	0.8756 (0.11)
	MCP		8.32 (0.20)	0.3415 (0.84)	0.8755 (0.11)

Table 2.5: Comparison of Lasso, SCAD and MCP in terms of model size (MS), false discover rate (FDR) and predictive AUC (PAUC) under misspecification, $n = 300$, $p = 1,000$.

Structure	Penalty	SNR	MS(SE)	FDR(SE*10^2)	PAUC(SE*10^2)
IN	Lasso	4.32	36.06 (0.37)	0.7237 (0.28)	0.9569 (0.03)
	SCAD		14.46 (0.22)	0.2681 (0.71)	0.9725 (0.02)
	MCP		13.34 (0.20)	0.2174 (0.71)	0.9725 (0.02)
SP	Lasso	3.06	35.22 (0.38)	0.7303 (0.26)	0.9361 (0.03)
	SCAD		13.08 (0.20)	0.2704 (0.63)	0.9530 (0.03)
	MCP		12.01 (0.17)	0.2272 (0.60)	0.9529 (0.03)
PC	Lasso	3.88	41.94 (0.40)	0.7844 (0.20)	0.9498 (0.03)
	SCAD		11.44 (0.17)	0.2385 (0.61)	0.9666 (0.02)
	MCP		10.53 (0.14)	0.1989 (0.57)	0.9666 (0.02)
AR	Lasso	3.21	41.59 (0.76)	0.7786 (0.56)	0.9030 (0.04)
	SCAD		13.19 (0.23)	0.3609 (0.75)	0.9497 (0.03)
	MCP		12.06 (0.22)	0.3070 (0.76)	0.9495 (0.03)
CS	Lasso	3.06	35.25 (0.36)	0.7316 (0.25)	0.9372 (0.03)
	SCAD		13.60 (0.23)	0.2761 (0.72)	0.9528 (0.02)
	MCP		12.78 (0.21)	0.2409 (0.72)	0.9527 (0.02)

2.4.6 Comparison in a cancer study

The concave penalized logistic regression is further applied to a cancer study to compare the Lasso, SCAD and MCP. The purpose of the study is to discover biomarkers associated with the prognosis of breast cancer (van't Veer *et al* (2002); Van de Vijver *et al* (2002)). Approximately $25,000$ genes were scanned using microarrays for $n = 295$ patients. Metastasis within five years is modeled as the outcome. A subset of $1,000$ genes with the highest Spearman correlation to the outcome are used to stabilize the computation.

For the same reason as the simulation study, we do not resort to any tuning parameter selection procedure to choose the model for comparison in this subsection. Instead, we randomly split the whole dataset with $n = 295$ into a training (approximately 1/3 of the observations) and validation datasets (approximately 2/3 of the observations). The model fitting is solely based on the training dataset; the solution corresponding to the maximum PAUC of the validation dataset is chosen as the final model for comparison. This partition process is repeated for 900 times.

The results presented in table 2.6 are consistent with those from simulation as Lasso tends to select a larger model. The SCAD and MCP perform very similarly since the PAUCs of SCAD and MCP are close to each other with similar MS.

Table 2.6: Comparison of Lasso, SCAD and MCP based on 900 random partition of a breast cancer microarray dataset.

Penalty	PAUC(SE*10^2)	MS(SE)
Lasso	0.7523 (0.10)	46.97 (0.53)
SCAD	0.7563 (0.10)	34.48 (0.51)
MCP	0.7565 (0.10)	34.85 (0.50)

2.4.7 Results using tuning parameter selection

We now present the results of the breast cancer study by using the cross-validated area under the ROC (CV-AUC) as an tuning parameter selection method.

This method uses a combination of cross validation and ROC methodology. Cross validation creates the training and test samples, with training samples being used for model fitting, and test samples for computing the predictive AUC of the fitted model. Repeat the process for k times to compute the average predictive AUC, which is defined as the CV-AUC. The models with the highest CV-AUC are chosen as the final model. For details of using the CV-AUC for tuning parameter selection in penalized logistic regression, we refer to Jiang, Huang, and Zhang (2011). We use 5-fold cross validation to compute the CV-AUC.

For this dataset, Lasso penalty selects 101 variables with CV-AUC=0.7797, SCAD penalty selects 26 variables with CV-AUC=0.7859 and MCP selects 24 variables with CV-AUC=0.7886. All the 24 variables selected by MCP are also selected by SCAD. Among the 26 variables selected by SCAD, only 2 are not selected by Lasso. The results are consistent with those of simulation. In particular, the MCP selects a model with the highest CV-AUC with the smallest model size.

2.5 Further Example of the MMCD Algorithm

When the outcome variable has $K > 2$ levels, a logistic model can be extended to a baseline-category logit model. Let y_{ik} be the indicator of the outcome of the ith observation in the kth level, $k = 1, ..., K$ and $\boldsymbol{x}^i$ be the corresponding covariates. The baseline-category logit model assumes that

$$\log(\frac{\pi_k(\boldsymbol{x})}{\pi_K(\boldsymbol{x})}) = \boldsymbol{x}^T \boldsymbol{\beta}_k, \tag{2.14}$$

with $\pi_k(\boldsymbol{x})$ being the probability of the outcome in the kth level, and $\boldsymbol{\beta}_k$ being the corresponding coefficients. As in the case of Logistic regression, we assume $\boldsymbol{\beta}_k \in \mathbb{R}^{p+1}$ and β_{k0} being the intercept and not penalized.

Denote $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_{K-1}^T)$ as the vector of regression coefficients. Given the structure of (2.14), we have $\pi_k(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta}_k)/\{1 + \sum_{k=1}^{K-1}\exp(\boldsymbol{x}^T\boldsymbol{\beta}_k)\}$. Hence the loss function for the multinomial case is

$$\ell(\boldsymbol{\beta}) \quad = \quad \frac{1}{n}\{\sum_{i=1}^{n}\log\{1+\sum_{k=1}^{K-1}\exp((\boldsymbol{x}^i)^T\boldsymbol{\beta}_k)\} - \sum_{i=1}^{n}\sum_{k=1}^{K-1} y_{ik}(\boldsymbol{x}^i)^T\boldsymbol{\beta}_k\}. \tag{2.15}$$

Correspondingly, the penalized regression model for the multinomial outcome is

$$Q(\boldsymbol{\beta}) = \frac{1}{n}\{\sum_{i=1}^{n}\log\{1+\sum_{k=1}^{K-1}\exp((\boldsymbol{x}^i)^T\boldsymbol{\beta}_k)\} - \sum_{i=1}^{n}\sum_{k=1}^{K-1} y_{ik}(\boldsymbol{x}^i)^T\boldsymbol{\beta}_k\} + \sum_{k=1}^{K-1}\sum_{j=1}^{p}\rho(|\beta_{kj}|;\lambda,\gamma). \tag{2.16}$$

Take second derivative of $\ell(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}_k$, we have

$$\nabla_k^2\ell(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{k=1}^{K-1}\exp((\boldsymbol{x}^i)^T\boldsymbol{\beta}_k)}{[1+\sum_{k=1}^{K-1}\exp((\boldsymbol{x}^i)^T\boldsymbol{\beta}_k)]^2}(\boldsymbol{x}^i)^T\boldsymbol{x}^i \tag{2.17}$$

Therefore, for the jth component in $\boldsymbol{\beta}_k$, the upper bound can be easily identified as

$$\nabla_{kj}^2\ell(\boldsymbol{\beta}) \leq \sum_{i=1}^{n} 1/4\boldsymbol{x}_{ij}^2 = 1/4.$$

Thus, we could still use $M = 1/4$ to meet the condition (ii) of the MMCD algorithm for the model. However, because of the multinomial outcome, we need two levels of cycling in the implementation of MMCD algorithm, first cycling through all the jth coordinates within $\boldsymbol{\beta}_k$, then cycling through the $k = 1, ...K-1$ to update $\boldsymbol{\beta}$.

We below outline the MMCD approach for the concave penalized baseline-category logit model.

Algorithm 2.3 MMCD Algorithm for the concave penalized baseline-category logit model

1. Given any initial value of $\hat{\boldsymbol{\beta}}^0$, computing the corresponding $\hat{\boldsymbol{\eta}}^1$.

2. Outer cycling: At step $m = 0, 1, ...,$ update $\hat{\boldsymbol{\beta}}_k^m$ to $\hat{\boldsymbol{\beta}}_k^{m+1}$ by the inner cycling.

   Inner cycling:

   (a) Given the current estimate of $\hat{\boldsymbol{\beta}}_{kj}^m = (\hat{\beta}_{k0}^{m+1}, ..., \hat{\beta}_{kj}^{m+1}, \hat{\beta}_{k(j+1)}^m, ..., \hat{\beta}_{kp}^m)$, update the estimate to $\hat{\boldsymbol{\beta}}_{k(j+1)}^m = (\hat{\beta}_{k0}^{m+1}, ..., \hat{\beta}_{kj}^{m+1}, \hat{\beta}_{k(j+1)}^{m+1}, ..., \hat{\beta}_{kp}^m)$ by using the solution in (2.10 or 2.11) for the penalized variables and (2.12) for the intercept, with τ_j being replace by τ_{kj},

   $$\tau_{kj} = \frac{\hat{\beta}_{kj}^m}{4} + \frac{1}{n}\sum_{i=1}^{n}\{\sum_{k=1}^{K-1} y_{ik} - \frac{\sum_{k=1}^{K-1}\exp((\boldsymbol{x}^i)^T\hat{\boldsymbol{\beta}}_k)}{[1+\sum_{k=1}^{K-1}\exp((\boldsymbol{x}^i)^T\hat{\boldsymbol{\beta}}_k)]^2}\}x_{ij},$$

   with $\hat{\boldsymbol{\beta}}_k$ being the latest estimate of $\boldsymbol{\beta}_k$. After each iteration, also update the corresponding linear component.

   (b) Cycle through all the coordinate $j = 0, ..., p$ such that $\hat{\boldsymbol{\beta}}_k^m$ is updated to $\hat{\boldsymbol{\beta}}_k^{m+1}$.

3. Repeat the inner cycling and cycle through the $k = 1, ..., K-1$ blocks of $\boldsymbol{\beta}$, update $\hat{\boldsymbol{\beta}}^m$ to $\hat{\boldsymbol{\beta}}^{m+1}$.

4. Check the convergence criterion. If converges then stop the iteration, otherwise repeat step 2 and 3 until converge.

CHAPTER 3
ℓ_2 GROUPED CONCAVE PENALTY IN GLM

Section 3.1 introduces the family of ℓ_2 grouped concave penalties, discusses the selection properties. The ℓ_2 grouped concave penalty performs group selection, which includes the group Lasso as a special case. Section 3.2 develops an efficient algorithm for the penalty and establishes its theoretical convergence properties under certain regularity conditions. Section 3.3 extends the ℓ_2 group penalties to GLM models and develops corresponding numeric algorithm, which shares the same spirit of the MMCD algorithm. Section 3.4 performs several comparison between the grouped and ungrouped concave penalties by simulation. The results show certain advantages of grouped concave penalties over the ungrouped ones.

3.1 ℓ_2 Grouped Concave Penalty in Linear Regression

Let $\{(y_i, \boldsymbol{x}^i)_{i=1}^n\}$ be the observed data, with $y_i \in \mathbb{R}$ be the response and $\boldsymbol{x}^i \in \mathbb{R}^{p+1}$ be the vector of predictors. Assuming the underlying generating model is $y_i = \boldsymbol{\beta}^T \boldsymbol{x}^i + \epsilon_i$, with $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T \in \mathbb{R}^{p+1}$ and ϵ_i are i.i.d $N(0, \sigma^2)$, we want to recover the oracle set $A_0 = \{l : \beta_l \neq 0, l = 1, ..., p\}$. Let $\boldsymbol{x}_l = (x_{1l}, x_{2l}, ..., x_{nl})^T$ be the vector of variable $l, l = 0, ..., p$ and denote the design matrix as $\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_p)$. Similar to chapter 2, we assume β_0 is the intercept and is not penalized. For the rest p penalized predictors, we assume a pre-defined group structure, i.e. the p variables are grouped into a total of J groups, with group size $d_j, j = 1, ..., J$ for the jth group, whose design matrix is denoted as $X_j = (\boldsymbol{x}_{j_1}, ..., \boldsymbol{x}_{j_k}, ..., \boldsymbol{x}_{j_{d_j}})$. Given the group structure,

an unique index l matches to the index j_k in the grouped structure. Correspondingly, the design matrix can also be written as $\boldsymbol{X} = (\boldsymbol{x}_0, X_1, X_2, ..., X_J)$. In general, the predictors from the same group are more correlated. Direct application of the concave penalty, which ignores the group structure, tends to select only one predictor from a group. The motivation of the grouped concave penalty is to incorporate the group information in a regression model and improve its performance.

Throughout chapter 3, we assume the penalized predictors are group-wise standardized, i.e. $X_j^T X_j = n\boldsymbol{I}$, with $\boldsymbol{I}$ being the identity matrix of dimension $d_j \times d_j$. Given any design matrix, the group-wise standardization can be done through Cholesky decomposition, provided $d_j < n$ for $j = 1, ..., J$. To ease notation, we assume the standardization process is done properly beforehand. Consider the (scaled) sum of squares as the loss function $\ell(\boldsymbol{\beta})$, i.e.

$$\ell(\boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \boldsymbol{x}^i)^2. \tag{3.1}$$

Then given the aforementioned grouped structure, the ℓ_2 grouped concave penalized criterion is defined as

$$Q^2(\boldsymbol{\beta}; \lambda, \gamma) \quad = \quad \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \boldsymbol{x}^i)^2 + \sum_{j=1}^{J} \rho(\|\boldsymbol{\beta}_j\|_2; \sqrt{d_j}\lambda, \gamma), \tag{3.2}$$

where $\boldsymbol{\beta}_j$ is the regression coefficients corresponding to the jth group. The notation $\|\mathbf{v}\|_2$ is the ℓ_2 norm of a k-dimensional vector $\mathbf{v}$. The penalty parameters $\sqrt{d_j}\lambda$ accounts for the group size such that large size groups and small size groups are fairly penalized. We brief the ℓ_2 grouped SCAD and MCP as the ℓ_2 GSCAD and ℓ_2 GMCP for convenience. Note that when $\gamma = +\infty$, both the ℓ_2 GSCAD and GMCP reduce to

the group Lasso (Yuan and Lin (2006); Meier, van de Geer and Bühlmann (2008)).

To understand the behavior of the ℓ_2 grouped concave penalty, consider a simple group structure with two members. Let $\boldsymbol{\beta}_1 = (b_1, b_2)^T$, Figure 3.1 illustrates the boundary of the ℓ_1 and ℓ_2 norm of $\boldsymbol{\beta}_1$. The linear edge of the ℓ_1 norm encourages sparsity within the group members; while the circular edge of the ℓ_2 norm does not. This is exactly the distinction between the ℓ_1 penalty (Lasso) and ridge penalty. Therefore, the ℓ_2 grouped concave penalty performs selection at group level, while the ℓ_1 grouped concave penalty introduced in chapter 4 performs selection at both group and individual level.

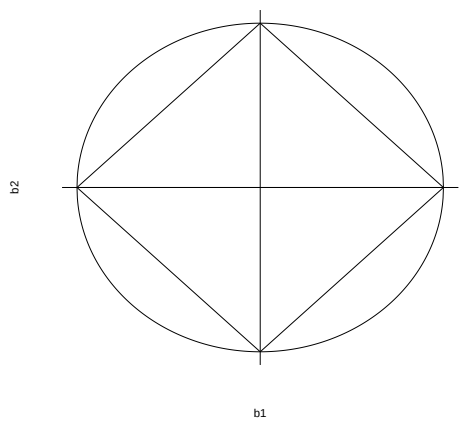

Figure 3.1: Boundary of the ℓ_1 and ℓ_2 norm for a vector with two elements

The following proposition, a direct consequence of the Karush-Kuhm-Tucker conditions, also shows the group selection property of the ℓ_2 grouped concave penalty.

Proposition 3.1. *Let $\hat{\boldsymbol{\beta}}$ be the solution of the ℓ_2 grouped concave penalty regression as defined in (3.2), then a sufficient and necessary condition for $\hat{\boldsymbol{\beta}}$ to be a (local) minimizer is that*

$$\begin{cases} \frac{1}{n}X_j^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \dot{\rho}(\|\hat{\boldsymbol{\beta}}_j\|_2; \sqrt{d_j}\lambda, \gamma)\frac{\hat{\boldsymbol{\beta}}_j}{\|\hat{\boldsymbol{\beta}}_j\|_2}, \forall \hat{\boldsymbol{\beta}}_j \neq 0, \\ \|\frac{1}{n}X_j^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\|_2 \leq \sqrt{d_j}\lambda, \forall \hat{\boldsymbol{\beta}}_j = 0, \end{cases} \tag{3.3}$$

The invariance property of the ℓ_2 GMCP states that certain adjustments are needed in order to achieve the fairness among different group sizes if we want to use the same rate of penalization λ. Note that the ℓ_2 GSCAD does not have such invariance property.

Proposition 3.2. ***(Invariance property of the ℓ_2 GMCP)*** *Given a group of group-wise standardized variables with size of d_j, the ℓ_2 GMCP has the following invariance property,*

$$\rho(\|\boldsymbol{\beta}_j\|_2; \sqrt{d_j}\lambda, \gamma) = \rho(\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2; \lambda, d_j\gamma) \tag{3.4}$$

The proof of the invariance property is easy by substituting into the MCP penalty form (1.3), and thus is skipped. The right hand side of invariance property equation suggests that for the same level of penalty λ, multiplying $\sqrt{d_j}$ to $\|\boldsymbol{\beta}_j\|_2$ is needed in order to standardize the group size.

3.2 Computation of the ℓ_2 Grouped Concave Penalty

Let $\hat{B}_j^m = (\hat{\beta}_0^{m+1}, ..., (\hat{\boldsymbol{\beta}}_j^{m+1})^T, (\hat{\boldsymbol{\beta}}_{j+1}^m)^T, ..., (\hat{\boldsymbol{\beta}}_J^m)^T)^T$ and $\hat{\boldsymbol{\beta}}_j^m = (\hat{\beta}_{j1}^m, ..., \hat{\beta}_{jd_j}^m)^T$. By group CDA approach, we want to update $\hat{\boldsymbol{\beta}}_j^m$ to $\hat{\boldsymbol{\beta}}_j^{m+1}$, i.e. update $\hat{B}_{j-1}^m$ to $\hat{B}_j^m$ if

using the new notation. Group CDA minimizes the criterion function

$$\begin{aligned}\boldsymbol{\beta}_j(\lambda,\gamma) &= \underset{\boldsymbol{\beta}_j}{\operatorname{argmin}}\, Q^2(\boldsymbol{\beta}_j|\hat{B}^m_{j-1}) \\ &= \underset{\boldsymbol{\beta}_j}{\operatorname{argmin}}\{\frac{1}{2n}\sum_{i=1}^n (y_i - (\hat{B}^m_{j-1})^T\boldsymbol{x}^i + (\hat{\boldsymbol{\beta}}^m_j - \boldsymbol{\beta}_j)^T\boldsymbol{x}^i_j)^2 \\ &+ \rho(\|\boldsymbol{\beta}_j\|_2; \sqrt{d_j}\lambda,\gamma)\}, \end{aligned} \tag{3.5}$$

by treating (3.2) as a function of $\boldsymbol{\beta}_j$ while fixing the rest group coordinates, with $\boldsymbol{x}^i_j = (x_{i,j_1}, ..., x_{i,j_{d_j}})^T$ being the covariates in the jth group for ith observation. Simple algebra gives the solution form as follows

$$\ell_2 \text{ GSCAD: } \hat{\boldsymbol{\beta}}^{m+1}_j = \begin{cases} \boldsymbol{u}_j, & \|\boldsymbol{u}_j\|_2 \geq \sqrt{d_j}\lambda\gamma, \\ \frac{1}{1-1/(\gamma-1)}\left[1 - \frac{\sqrt{d_j}\lambda\gamma}{(\gamma-1)\|\boldsymbol{u}_j\|_2}\right]_+ \boldsymbol{u}_j, & 2\sqrt{d_j}\lambda \leq \|\boldsymbol{u}_j\|_2 < \sqrt{d_j}\lambda\gamma, \\ \left[1 - \frac{\sqrt{d_j}\lambda}{\|\boldsymbol{u}_j\|_2}\right]_+ \boldsymbol{u}_j, & \|\boldsymbol{u}_j\|_2 \leq 2\sqrt{d_j}\lambda \end{cases} \tag{3.6}$$

$$\ell_2 \text{ GMCP: } \hat{\boldsymbol{\beta}}^{m+1}_j = \begin{cases} \boldsymbol{u}_j, & \|\boldsymbol{u}_j\|_2 \geq \sqrt{d_j}\lambda\gamma, \\ \frac{1}{1-1/\gamma}\left[1 - \frac{\sqrt{d_j}\lambda}{\|\boldsymbol{u}_j\|_2}\right]_+ \boldsymbol{u}_j, & \|\boldsymbol{u}_j\|_2 < \sqrt{d_j}\lambda\gamma, \end{cases} \tag{3.7}$$

where $\boldsymbol{u}_j = \hat{\boldsymbol{\beta}}^m_j + n^{-1}\sum_{i=1}^n X^T_j\{y_i - (\hat{B}^m_{j-1})^T\boldsymbol{x}^i\} = \hat{\boldsymbol{\beta}}^m_j + n^{-1}X^T_j(\boldsymbol{y} - \boldsymbol{X}\hat{B}^m_{j-1})$. For the non-penalized β_0, it is easy to show the solution form is

$$\hat{\beta}^{m+1}_0 = \hat{\beta}^m_0 + n^{-1}\boldsymbol{x}^T_0(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^m), \tag{3.8}$$

with $\hat{\boldsymbol{\beta}}^m = (\hat{\beta}^m_0, (\hat{\boldsymbol{\beta}}^m_1)^T, ..., (\hat{\boldsymbol{\beta}}^m_J)^T)^T$.

Algorithm 3.1 summarizes the (group) CDA for the computation of the ℓ_2 grouped concave penalty for a given (λ,γ). We also treat γ and λ both as tuning parameters and compute the solution surface over a rectangle of (γ,λ). Let $\kappa = 1/\gamma$, then $\kappa \in [0, \kappa_{\max} = 1/2)$ for the ℓ_2 GSCAD and $\kappa \in [0, \kappa_{\max} = 1)$ for the ℓ_2 GMCP. Note that when $\kappa = 0$, both the ℓ_2 GSCAD and GMCP simplify to the group Lasso.

Algorithm 3.1 Algorithm for the ℓ_2 Grouped Concave Penalty in Linear Regression

1. Given an initial value $\hat{\boldsymbol{\beta}}^0$, compute the corresponding linear part $\hat{\boldsymbol{\eta}}^0$.
2. For $m = 0, 1, ...,$ use (3.8) to update the intercept. For the penalized variables, update $\hat{B}_j^m$ to $\hat{B}_{j+1}^m$ by (3.6) or (3.7) for the jth group coordinate, $j = 1, ..., J$. After each update, compute the corresponding linear component using the latest estimate by the same tips used in MMCD. This step updates $\hat{\boldsymbol{\beta}}^m$ to $\hat{\boldsymbol{\beta}}^{m+1}$.
3. Check the convergence criterion. If algorithm converges then stop iterations, otherwise repeat step 2 until converges.

The solution surface of the ℓ_2 grouped concave penalty is also computed along κ. That is we first compute the solutions of group Lasso along γ. Then for a given κ, the solutions of group Lasso is used to initialize the computation of solutions along grids of κ. Based on the solution form of (3.6, 3.7), some calculations show that $\lambda_{max} = \max_j n^{-1}|X_j^T \boldsymbol{r}|_2/\sqrt{d_j}$ for the ℓ_2 grouped concave penalty, with $\boldsymbol{r} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{int})$ and $\hat{\boldsymbol{\beta}}_{int}$ being the estimate of the intercept-only model.

Figure 3.2 shows the solution path plots for the group Lasso and the ℓ_2 GMCP ($\gamma = 1.2$ and $\gamma = 2.7$) using a simulated example. The example has four groups with group size $(2, 3, 2, 3)$. The first two groups are the causal predictors with coefficients $(\sqrt{2}, -\sqrt{2})$ and $(0.5, 1, -0.5)$. Top row shows the ℓ_2 norm of coefficients, bottom row shows the individual coefficients. Solid and dash lines are the trajectories of causal predictors, dotted lines are the ones of null predictors. From the figure, we see that the group Lasso has a smoother trajectories.

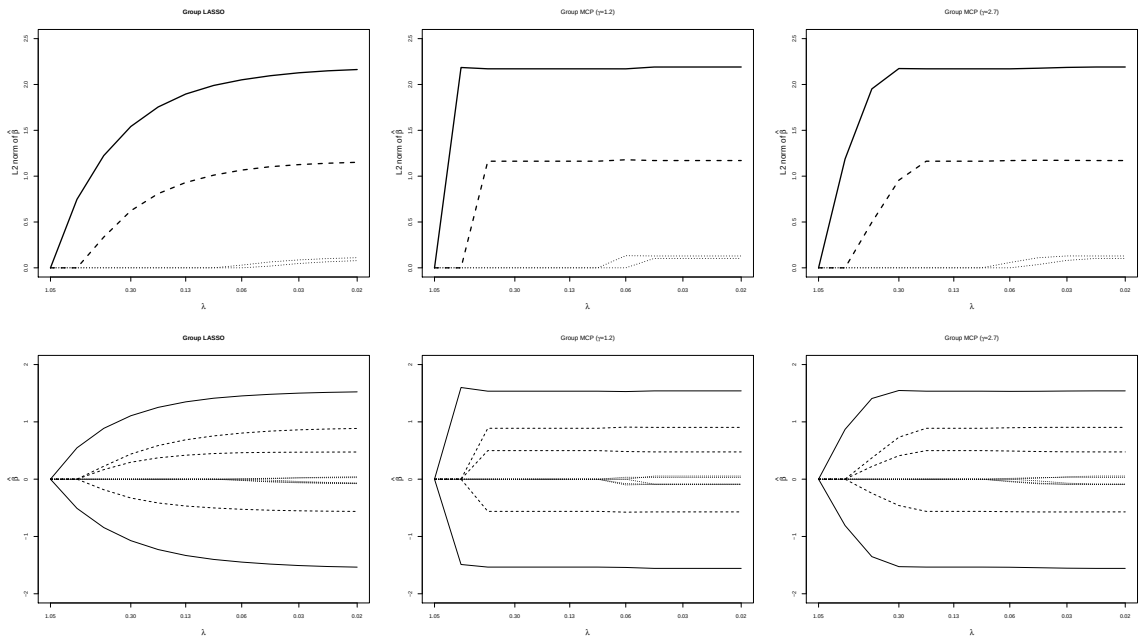

Figure 3.2: Solution paths of the group Lasso and the ℓ_2 GMCP. Group Lasso (left), ℓ_2 GMCP (center and right).

3.3 Convergence Analysis of Proposed Algorithms

Theorem 3.1, whose proof is provided in Appendix B, establishes that under certain regularity conditions, the algorithm 3.1 converges to a minimum of the objective functions for the ℓ_2 grouped concave penalties.

Theorem 3.1. *Consider the objective function in (3.2), where the given data* $(\boldsymbol{y}, \boldsymbol{X})$ *lies on a compact set and no two columns of* $\boldsymbol{X}$ *are identical. Suppose the penalty* $\rho(|t|; \lambda, \gamma) \equiv \rho(t)$ *satisfies* $\rho(t) = \rho(-t)$, $\rho^{'}(|t|)$ *is non-negative, uniformly bounded, with* $\rho^{'}(|t|)$ *being the first derivative (assuming existence) of* $\rho(|t|)$ *w.r.t* $|t|$.

Then the sequence $\{\boldsymbol{\beta}^m\}$ *generated by the aforementioned algorithm converges*

to a minimum of the function $Q^2(\boldsymbol{\beta})$ defined in (3.2).

3.4 Extension of the ℓ_2 Grouped Concave Penalty in GLM

Using the negative log-likelihood, we can easily extend the grouped concave penalty to GLM models. Mathematically, the ℓ_2 grouped concave penalized GLM criterion is

$$Q^2(\boldsymbol{\beta};\lambda,\gamma) \;=\; \frac{1}{n}\sum_{i=1}^{n}\{\psi(\boldsymbol{\beta}^T\boldsymbol{x}^i)-y_i\boldsymbol{\beta}^T\boldsymbol{x}^i\}+\sum_{j=1}^{J}\rho(\|\boldsymbol{\beta}_j\|_2;\sqrt{d_j}\lambda,\gamma). \tag{3.9}$$

Algorithm 3.1 can be modified to compute the solution of (3.9) by using the majorization approach as we do in MMCD. For the ℓ_2 grouped concave penalty, we assume the following conditions hold:

(i) There exists a real number $M>0$ such that $\boldsymbol{a}^T\nabla_j^2\ell(\boldsymbol{\beta})\boldsymbol{a}\le M\boldsymbol{a}^T\boldsymbol{a}$, $j=1,...,J$ for group-wise standardized $\boldsymbol{X}$, with $\nabla_j^2\ell(\boldsymbol{\beta})$ being the partial derivative of $\ell(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}_j$ and $\boldsymbol{a}$ being a non-zero vector with dimension d_j.

(ii) $\inf_t \rho''(|t|;\lambda,\gamma)>-M$, with $\rho''(|t|;\lambda,\gamma)$ being the second derivative of $\rho(|t|;\lambda,\gamma)$ w.r.t $|t|$.

For the ℓ_2 GSCAD and GMCP, the condition (ii) can be satisfied by choosing $\gamma>1+1/M$ for the SCAD and $\gamma>1/M$ for the MCP. Replacing the solution forms (3.6) and (3.7) by the following forms in GLM, the algorithm 3.1 can be used to computed the solutions for the ℓ_2 grouped concave penalty in GLM.

$$\ell_2\,\text{GSCAD: }\hat{\boldsymbol{\beta}}_j^{m+1}=\begin{cases}\frac{\boldsymbol{u}_j}{M}, & \|\boldsymbol{u}_j\|_2\ge M\sqrt{d_j}\lambda\gamma,\\ \frac{1}{M-1/(\gamma-1)}\left[1-\frac{\sqrt{d_j}\lambda\gamma}{(\gamma-1)\|\boldsymbol{u}_j\|_2}\right]_+\boldsymbol{u}_j, & (1+M)\sqrt{d_j}\lambda\le\|\boldsymbol{u}_j\|_2<M\sqrt{d_j}\lambda\gamma,\\ \frac{1}{M}\left[1-\frac{\sqrt{d_j}\lambda}{\|\boldsymbol{u}_j\|_2}\right]_+\boldsymbol{u}_j, & \|\boldsymbol{u}_j\|_2\le(1+M)\sqrt{d_j}\lambda,\end{cases} \tag{3.10}$$

$$\ell_2\,\text{GMCP}:\ \hat{\boldsymbol{\beta}}_j^{m+1} = \begin{cases} \frac{\boldsymbol{u}_j}{M}, & \|\boldsymbol{u}_j\|_2 \geq M\sqrt{d_j}\lambda\gamma, \\ \frac{1}{M-1/\gamma}\left[1-\frac{\sqrt{d_j}\lambda}{\|\boldsymbol{u}_j\|_2}\right]_+ \boldsymbol{u}_j, & \|\boldsymbol{u}_j\|_2 < M\sqrt{d_j}\lambda\gamma, \end{cases} \tag{3.11}$$

with $\boldsymbol{u}_j = M\hat{\boldsymbol{\beta}}_j^m + n^{-1}\sum_{i=1}^n X_j^T\{y_i - \dot{\psi}\left[(\hat{B}_{j-1}^m)^T\boldsymbol{x}^i\right]\}$.

Theorem 3.2 establishes the convergence of the algorithm in a GLM model under certain regularity conditions. It states that the proposed algorithm converges to a minimum of the objective function defined in (3.9) for the ℓ_2 grouped concave penalized GLM. We provided the details of the proof in Appendix B,

Theorem 3.2. *Consider the objective function defined in (3.9), where the given data* $(\boldsymbol{y}, \boldsymbol{X})$ *lies on a compact set and no two columns of* $\boldsymbol{X}$ *are identical. Suppose the penalty* $\rho(|t|;\lambda,\gamma) \equiv \rho(t)$ *satisfies* $\rho(t) = \rho(-t)$, $\rho^{'}(|t|)$ *is non-negative, uniformly bounded, with* $\rho^{'}(|t|)$ *being the first derivative (assuming existence) of* $\rho(|t|)$ *w.r.t* $|t|$. *Also assume the two conditions required in the computation hold.*

Then the sequence $\{\boldsymbol{\beta}^m\}$ *generated by the aforementioned algorithm converges to a minimum of the function* $Q^2(\boldsymbol{\beta})$ *defined in (3.9).*

3.5 Simulation Studies in Linear and Logistic Models

We compare the empirical performance of the ℓ_2 grouped concave penalty with the ungrouped concave penalty in linear and logistic models in this section. For linear models, the generating model is $y_i = \boldsymbol{\beta}^T\boldsymbol{x}^i + \epsilon_i$, with ϵ_i i.i.d $N(0,1)$. For logistic models, the generating model is $\text{logit}(P(y_i = 1)) = \boldsymbol{\beta}^T\boldsymbol{x}^i$. The condition (i) of the ℓ_2 grouped concave penalty in a logistic model can be easily met by choosing $M = 1/4$. The condition (ii) sets the proper range of $\kappa \in [0, 1/5)$ for GSCAD and $\kappa \in [0, 1/4)$

for GMCP.

For both linear and logistic regression, we set $\beta_0 = 0.01$. The p penalized covariates $(x_{i1}, ..., x_{ip})^T \sim \text{MVN}(\mathbf{0}, \sigma^2\Sigma_{p\times p})$, with $\sigma^2 = 1$. The covariance matrix $\Sigma_{p\times p} = \Sigma_{base} + \text{block diagonal}(\Sigma_1, ..., \Sigma_J)$. The Σ_{base} is a $p \times p$ matrix having a compound symmetry structure with $\rho = 0.1$, representing a baseline correlation among predictors. The Σ_j is the covariance matrix for the jth group with dimension $d_j \times d_j$. Within Σ_j, we choose a compound symmetry structure with $\rho = 0.6$ to represent a median level within group correlation.

We set $p = 500$ and consider two scenarios (1) equal group size (EQU) and (2) unequal group size (UNE). For equal group size, set $d_j = 10$ for $j = 1, ..., 50$; for unequal group size, set $d_j = 10$ for $j = 1, 2, 5, 6, ..., 33, 34, 37, 38$, and $d_j = 15$ for $j = 3, 4, 7, 8, ..., 35, 36, 39, 40$. On the coefficient side, we set $\beta_l = 0, l = 61, ..., 500$ and $(\beta_1, ..., \beta_{60}) = (\boldsymbol{a}, -\boldsymbol{a}, \boldsymbol{a}, -\boldsymbol{a}, \boldsymbol{a}, -\boldsymbol{a})$, with $\boldsymbol{a}$ being a vector of length 10. The value of vector $\boldsymbol{a}$ is chosen such that SNR is approximately in the range of $(2, 4)$. We consider 3 types of $\boldsymbol{a}$ as listed below to representing 3 settings,

- 1. $\boldsymbol{a} = (0.2, 0.2, 0.25, 0.25, 0.3, 0.2, 0.2, 0.25, 0.25, 0.3)^T$ to represent the situation that the effects of group members are relative small but similar.
- 2. $\boldsymbol{a} = (0.02, 0.02, 0.2, 0.25, 0.5, 0.02, 0.02, 0.2, 0.25, 0.5)^T$ to represent the situation that the effects of some group members are small but not zero.
- 3. $\boldsymbol{a} = (0.02, 0.02, 0.1, 0.1, 0.7, 0.02, 0.02, 0.1, 0.1, 0.7)^T$ to represent the situation that the only one or two members within the group have strong effect with some members have small effect.

The validation approach is used again to select final models for comparison. Denote these validation samples as $y_i^*, \boldsymbol{x}^{*i}, i = 1, ..., n^*$, with $n^* = 3,000$. For linear regression, compute the predictive mean square error (PMSE) defined as $\text{PMSE}_{\kappa_k,\lambda_v} = \frac{1}{n^*}\sum_{i=1}^{n^*}(y_i^* - \hat{\boldsymbol{\beta}}_{\kappa_k,\lambda_v}^T \boldsymbol{x}^{*i})^2$ corresponding to the solution $\hat{\boldsymbol{\beta}}_{\kappa_k,\lambda_v}$. For logistic regression, compute the PAUC as described in chapter 2. The solutions corresponding to the smallest PMSE in linear models and the largest PAUC in logistic models are selected for comparison.

In addition to model size (MS), correct size (CS), false discovery rate (FDR), we also use the following measurements for comparison purpose. Denote the causal group set as $\mathbb{A}_0 = \{j : \boldsymbol{\beta}_j \neq 0, j = 1, ..., J\}$ with dimension P_0, and denote its estimated version as $\hat{\mathbb{A}}_0 = \{j : \boldsymbol{\beta}_j \neq 0, j = 1, ..., J\}$ with dimension $\hat{P}_0$. Denote the set of false positive groups as $\mathbb{F}_0 = \{j : \hat{\boldsymbol{\beta}}_j \neq 0 \,\&\, \boldsymbol{\beta}_j = 0, j = 1, ..., J\}$ with dimension S. We define group model size (GMS) as $\text{GMS} = \hat{P}_0$, correct group size (CGS) as $\text{CGS} = \hat{P}_0 - S_0$ and group false discovery rate (GFDR) $\text{GFDR} = S_0/\hat{P}_0$. The PMSE in linear models or the PAUC in logistic models of the validation samples are also used. The results reported below are based on 500 replicates.

Table 3.1 shows the comparison between the ℓ_2 grouped concave penalties and the ungrouped concave penalties in linear models and table 3.2 shows the results in logistic models. For both comparisons, we set $n = 300$. We observe similar results in both linear and logistic models. The ℓ_2 GSCAD and GMCP have smaller GMS and GFDR in three settings. At individual level, the ℓ_2 GSCAD and GMCP may not necessary have smaller model sizes when comparing to SCAD and MCP. But the

grouped penalties tend to have smaller FDR. In terms of prediction, the ℓ_2 GSCAD and GMCP seems to outperform the rest approaches. Under the setting 1, where the effect of group members are small and similar, the ℓ_2 grouped concave penalties seems to work best. Among the ℓ_2 GSCAD and GMCP, the ℓ_2 GMCP works slightly better than the ℓ_2 GSCAD.

3.6 Simulation Studies in Poisson Models

Poisson regression model is another important member in the GLM family. It appears that in Poisson regression model it is difficult to apply the majorization approach since there is no simple majorization function exists. However, due to the convex property, group Lasso could be implemented. So in this section, we study the finite-sample performance of Lasso and group Lasso in Poisson regression model. We use the same design to generate the covariates. For the outcome variable, we use the generating model as $E(y_i) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}^i)$.

Table 3.3 presents the results based on 500 replicates. It seems that group Lasso has smaller GMS and smaller GFDR, while Lasso has smaller MS and FDR. In terms of PMSE, these two approaches are very similar with group lasso have a bit higher PMSE. This observation is similar to the cases in linear and logistic models.

Table 3.1: Comparison of the ℓ_2 grouped vs. the ungrouped concave penalties in linear models, $n = 300$, $p = 500$ and $\rho = 0.6$.

Setting	Group (SNR)	Method	PMSE (SE*10^3)	GMS (SE*10)	GFDR (SE*10^3)	MS (SE*10)	FDR (SE*10^3)
1	EQU 3.56	Lasso	1.57 (5.2)	37.76 (1.8)	0.84 (0.8)	112.95 (5.1)	0.47 (2.3)
		SCAD	1.57 (5.2)	37.76 (1.8)	0.84 (0.8)	112.93 (5.1)	0.47 (2.3)
		MCP	1.57 (5.2)	37.76 (1.8)	0.84 (0.8)	112.95 (5.1)	0.47 (2.3)
		ℓ_2Lasso	1.64 (6.3)	27.91 (0.6)	0.78 (0.5)	279.10 (6.2)	0.78 (0.5)
		ℓ_2GSCAD	1.25 (2.8)	7.57 (1.0)	0.15 (8.2)	75.70 (10.2)	0.15 (8.2)
		ℓ_2GMCP	1.25 (2.8)	7.03 (0.7)	0.11 (6.7)	70.28 (7.2)	0.11 (6.7)
	UNE 2.88	Lasso	1.91 (6.5)	36.82 (1.0)	0.86 (0.4)	133.76 (7.7)	0.59 (2.0)
		SCAD	1.91 (6.5)	36.80 (1.0)	0.86 (0.4)	133.48 (7.7)	0.59 (2.0)
		MCP	1.91 (6.5)	36.78 (1.1)	0.86 (0.5)	133.53 (7.8)	0.59 (2.1)
		ℓ_2Lasso	1.56 (5.2)	22.22 (0.6)	0.77 (0.7)	277.37 (7.4)	0.78 (0.6)
		ℓ_2GSCAD	1.24 (2.6)	6.40 (1.0)	0.16 (8.5)	77.03 (11.9)	0.16 (8.7)
		ℓ_2GMCP	1.24 (2.6)	5.62 (0.6)	0.08 (6.1)	67.33 (7.0)	0.08 (6.1)
2	EQU 3.13	Lasso	1.42 (4.1)	35.72 (1.9)	0.83 (1.0)	93.25 (4.9)	0.51 (2.6)
		SCAD	1.40 (3.9)	34.17 (2.1)	0.82 (1.1)	87.03 (5.0)	0.49 (2.8)
		MCP	1.40 (3.9)	31.22 (3.1)	0.79 (3.1)	80.41 (8.0)	0.45 (4.1)
		ℓ_2Lasso	1.62 (6.0)	27.79 (0.7)	0.78 (0.6)	277.94 (6.9)	0.78 (0.6)
		ℓ_2GSCAD	1.25 (2.8)	7.74 (1.2)	0.16 (8.8)	77.44 (11.9)	0.16 (8.8)
		ℓ_2GMCP	1.25 (2.8)	7.15 (0.9)	0.12 (7.5)	71.54 (8.7)	0.12 (7.5)
	UNE 2.61	Lasso	1.59 (5.2)	35.60 (1.1)	0.86 (0.5)	109.39 (6.1)	0.62 (2.0)
		SCAD	1.51 (4.6)	32.35 (1.6)	0.84 (0.9)	89.33 (6.9)	0.57 (2.3)
		MCP	1.51 (4.7)	26.86 (3.2)	0.79 (4.4)	72.28 (10.5)	0.47 (5.5)
		ℓ_2Lasso	1.54 (5.2)	22.15 (0.7)	0.77 (0.8)	276.42 (8.0)	0.78 (0.7)
		ℓ_2GSCAD	1.24 (2.6)	6.40 (0.9)	0.16 (8.5)	77.11 (11.9)	0.16 (8.7)
		ℓ_2GMCP	1.24 (2.6)	5.74 (0.7)	0.09 (6.7)	68.89 (8.8)	0.09 (6.8)
3	EQU 3.21	Lasso	1.35 (3.4)	33.93 (1.9)	0.82 (1.1)	82.28 (4.6)	0.51 (2.6)
		SCAD	1.29 (2.5)	28.28 (2.3)	0.78 (1.9)	63.65 (5.5)	0.48 (3.0)
		MCP	1.30 (2.5)	24.63 (3.5)	0.73 (4.6)	55.54 (8.5)	0.42 (4.7)
		ℓ_2Lasso	1.62 (6.0)	27.80 (0.7)	0.78 (0.7)	277.98 (7.3)	0.78 (0.7)
		ℓ_2GSCAD	1.25 (2.8)	7.69 (1.2)	0.16 (8.8)	76.88 (11.7)	0.16 (8.8)
		ℓ_2GMCP	1.25 (2.8)	7.22 (0.9)	0.12 (7.7)	72.20 (9.1)	0.12 (7.7)
	UNE 2.76	Lasso	1.40 (3.6)	32.44 (1.4)	0.84 (0.7)	83.70 (4.9)	0.61 (2.2)
		SCAD	1.27 (2.0)	21.71 (2.5)	0.75 (4.7)	47.44 (5.9)	0.47 (4.3)
		MCP	1.28 (2.0)	15.99 (2.7)	0.62 (8.7)	35.15 (6.2)	0.36 (6.3)
		ℓ_2Lasso	1.55 (5.3)	22.03 (0.7)	0.77 (0.8)	275.23 (8.5)	0.78 (0.8)
		ℓ_2GSCAD	1.24 (2.6)	6.30 (0.9)	0.15 (8.6)	75.86 (11.4)	0.15 (8.7)
		ℓ_2GMCP	1.24 (2.6)	5.74 (0.7)	0.09 (7.0)	68.95 (9.3)	0.08 (7.1)

Table 3.2: Comparison of the ℓ_2 grouped vs. the ungrouped concave penalties in logistic models, $n = 300$, $p = 500$ and $\rho = 0.6$.

Setting	Group (SNR)	Method	PAUC (SE*10^3)	GMS (SE*10)	GFDR (SE*10^3)	MS (SE*10)	FDR (SE*10^3)
1	EQU 3.56	Lasso	0.865(0.55)	27.75(2.6)	0.77(2.2)	68.87(6.6)	0.44(3.5)
		SCAD	0.865(0.55)	27.77(2.6)	0.77(2.2)	68.85(6.2)	0.44(3.5)
		MCP	0.865(0.55)	27.77(2.6)	0.77(2.2)	68.85(6.2)	0.44(3.5)
		ℓ_2Lasso	0.855(0.76)	15.70(1.7)	0.60(4.4)	157.04(16.8)	0.60(4.4)
		ℓ_2GSCAD	0.888(0.56)	6.99(0.7)	0.11(6.5)	69.86(7.3)	0.11(6.5)
		ℓ_2GMCP	0.888(0.56)	6.70(0.7)	0.07(5.9)	66.98(7.2)	0.07(5.9)
	UNE 2.88	Lasso	0.832(0.63)	21.42(2.3)	0.75(3.0)	52.10(5.5)	0.43(4.1)
		SCAD	0.832(0.63)	21.42(2.3)	0.75(3.0)	52.10(5.5)	0.43(4.1)
		MCP	0.832(0.63)	21.42(2.3)	0.75(3.0)	52.10(5.5)	0.43(4.1)
		ℓ_2Lasso	0.838(0.74)	12.73(1.5)	0.58(5.2)	155.92(19.6)	0.59(5.3)
		ℓ_2GSCAD	0.857(0.70)	7.89(1.5)	0.28(10.3)	95.40(19.3)	0.29(10.5)
		ℓ_2GMCP	0.857(0.70)	6.78(1.2)	0.20(9.6)	81.55(15.2)	0.20(9.7)
2	EQU 3.13	Lasso	0.856(0.62)	25.05(2.6)	0.75(2.8)	56.10(5.4)	0.44(3.8)
		SCAD	0.857(0.62)	25.03(2.6)	0.75(2.8)	56.02(5.5)	0.44(3.8)
		MCP	0.857(0.62)	24.97(2.6)	0.75(2.8)	55.90(5.5)	0.44(3.8)
		ℓ_2Lasso	0.837(0.80)	15.86(1.7)	0.60(4.7)	158.56(17.4)	0.60(4.7)
		ℓ_2GSCAD	0.869(0.65)	7.33(1.0)	0.13(7.6)	73.30(10.5)	0.13(7.6)
		ℓ_2GMCP	0.869(0.65)	6.88(0.9)	0.09(6.7)	68.78(8.5)	0.09(6.7)
	UNE 2.61	Lasso	0.820(0.72)	20.44(2.6)	0.73(3.6)	45.05(6.1)	0.46(4.7)
		SCAD	0.820(0.71)	20.34(2.6)	0.73(3.7)	44.78(6.1)	0.46(4.7)
		MCP	0.820(0.71)	20.31(2.6)	0.73(3.8)	44.67(6.2)	0.46(4.7)
		ℓ_2Lasso	0.820(0.82)	12.42(1.6)	0.57(5.9)	151.71(20.8)	0.57(6.1)
		ℓ_2GSCAD	0.841(0.76)	7.61(1.4)	0.27(10.1)	91.75(17.9)	0.27(10.2)
		ℓ_2GMCP	0.841(0.77)	6.56(1.2)	0.17(9.5)	78.48(14.8)	0.17(9.6)
3	EQU 3.21	Lasso	0.878(0.67)	20.69(2.4)	0.69(3.8)	41.63(4.5)	0.43(4.4)
		SCAD	0.880(0.72)	17.80(2.9)	0.61(7.7)	34.34(6.0)	0.39(6.0)
		MCP	0.880(0.72)	17.21(3.1)	0.57(9.6)	33.53(6.3)	0.36(6.9)
		ℓ_2Lasso	0.843(0.83)	15.16(1.7)	0.58(4.7)	151.58(17.4)	0.58(4.7)
		ℓ_2GSCAD	0.873(0.64)	7.15(1.0)	0.12(7.2)	71.50(9.5)	0.12(7.2)
		ℓ_2GMCP	0.873(0.63)	6.74(0.8)	0.08(6.1)	67.38(8.3)	0.08(6.1)
	UNE 2.76	Lasso	0.843(0.73)	20.57(2.9)	0.73(4.7)	41.63(6.7)	0.50(5.2)
		SCAD	0.850(0.85)	16.08(3.1)	0.62(8.2)	29.71(6.5)	0.43(7.1)
		MCP	0.850(0.85)	14.96(3.3)	0.56(10.5)	27.96(6.8)	0.38(8.4)
		ℓ_2Lasso	0.830(0.80)	12.20(1.4)	0.56(5.3)	149.62(18.2)	0.57(5.4)
		ℓ_2GSCAD	0.852(0.68)	7.01(1.3)	0.21(9.7)	84.53(15.8)	0.21(9.8)
		ℓ_2GMCP	0.852(0.68)	6.35(1.1)	0.14(8.8)	76.44(13.8)	0.14(8.9)

Table 3.3: Comparison of the group Lasso and Lasso penalties in Poisson models, $n = 300$, $p = 500$ and $\rho = 0.6$.

Setting	Group (SNR)	Method	log(PMSE) (SE*10^2)	GMS (SE*10)	GFDR (SE*10^3)	MS (SE)	FDR (SE*10^3)
1	EQU	Lasso	17.78 (9.2)	25.13 (2.8)	0.75 (3.5)	64.35 (0.9)	0.39 (2.9)
	3.56	Glasso	17.84 (9.4)	13.30 (4.7)	0.68 (9.0)	132.99 (4.7)	0.68 (9.0)
	UNE	Lasso	13.06 (7.8)	28.78 (2.7)	0.82 (2.6)	82.23 (1.0)	0.49 (3.3)
	2.88	Glasso	13.19 (7.8)	13.06 (3.7)	0.64 (8.3)	160.50 (4.7)	0.66 (8.4)
2	EQU	Lasso	14.74 (7.5)	27.50 (2.7)	0.77 (3.0)	68.08 (0.8)	0.43 (2.9)
	3.13	Glasso	14.88 (7.5)	17.03 (4.9)	0.69 (7.4)	170.27 (4.9)	0.69 (7.4)
	UNE	Lasso	11.09 (6.4)	31.62 (2.4)	0.84 (1.8)	90.27 (1.0)	0.57 (3.2)
	2.61	Glasso	11.24 (6.3)	15.32 (3.6)	0.66 (8.2)	189.29 (4.6)	0.67 (8.3)
3	EQU	Lasso	15.27 (7.5)	26.99 (2.7)	0.77 (2.8)	60.44 (0.7)	0.48 (2.9)
	3.21	Glasso	15.44 (7.4)	16.15 (4.8)	0.68 (8.2)	161.54 (4.8)	0.68 (8.2)
	UNE	Lasso	12.02 (6.7)	30.08 (2.2)	0.83 (1.6)	75.10 (0.9)	0.59 (2.7)
	2.76	Glasso	12.30 (6.5)	14.26 (3.7)	0.65 (8.7)	175.98 (4.7)	0.66 (8.7)

CHAPTER 4
ℓ_1 GROUPED CONCAVE PENALTY IN GLM

Section 4.1 introduces a new family of grouped concave penalties, the ℓ_1 grouped concave penalties and discusses their selection properties. The ℓ_1 grouped concave penalty performs selection at both group and individual levels, a desirable feature in some applications. Section 4.2 develops an efficient algorithm to compute the solution of the ℓ_1 grouped concave penalized criterion and establishes the theoretical convergence property under certain regularity conditions. Section 4.3 extends the ℓ_1 grouped concave penalty to GLM models and develops corresponding numeric algorithm. Section 4.4 first compares the performance of the ℓ_1 grouped concave penalty with the ungrouped concave penalty, then compares the two grouped concave penalties based on simulation study.

4.1 ℓ_1 Grouped Concave Penalty in Linear Regression

Consider a linear model, $y_i = \boldsymbol{\beta}^T \boldsymbol{x}^i + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$. Here, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T \in \mathbb{R}^{p+1}$ and denote the design matrix $\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_p)$. We also assume a pre-defined group structure with group size $d_j, j = 1, ..., J$ for the design matrix as the ℓ_2 grouped concave penalty. Throughout chapter 4, the penalized predictors are assumed to be column-wise standardized, i.e. $\|\boldsymbol{x}_l\|_2^2 = n$. The ℓ_1 grouped concave penalized linear model is defined as

$$Q^1(\boldsymbol{\beta}; \lambda, \gamma) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \boldsymbol{x}^i)^2 + \sum_{j=1}^{J} \rho(\|\boldsymbol{\beta}_j\|_1; d_j \lambda, \gamma), \tag{4.1}$$

with $\boldsymbol{\beta}_j$ the regression coefficients of jth group. The notation $\|\mathbf{v}\|_1$ is the ℓ_1 norm of a k-dimensional vector $\mathbf{v}$. The penalty parameters $d_j\lambda$ accounts for the group size in group structure. We brief the ℓ_1 grouped SCAD and MCP as the ℓ_1 GSCAD and ℓ_1 GMCP for convenience. Note that when $\gamma = +\infty$, both the ℓ_1 GSCAD and GMCP reduce to the ℓ_1 penalty with penalty parameter $d_j\lambda$ proportional to the group size. We call this special case as proportional Lasso or ℓ_1 Lasso.

The following proposition states that the ℓ_1 grouped concave penalty could have zero and non-zero coefficients within the same group, which means that ℓ_1 grouped concave penalty performs selection both at group and individual levels.

Proposition 4.1. *Let $\hat{\boldsymbol{\beta}}$ be the solution to the ℓ_1 grouped concave penalty regression as defined in (4.1), then a sufficient and necessary condition for $\hat{\boldsymbol{\beta}}$ to be a (local) minimizer is that*

$$\begin{cases} \frac{1}{n}\boldsymbol{x}_{j_k}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \dot{\rho}(\|\hat{\boldsymbol{\beta}}_j\|_1; d_j\lambda, \gamma) sgn(\hat{\beta}_{j_k}), \ \forall \hat{\beta}_{j_k} \neq 0, \\ |\frac{1}{n}\boldsymbol{x}_{j_k}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})| \leq (d_j\lambda - \frac{\sum_{s\neq k}|\hat{\beta}_{j_s}|}{\gamma})_+, \ \forall \hat{\beta}_{j_k} = 0. \end{cases} \tag{4.2}$$

where $\dot{\rho}(|t|)$ is the first derivative of $\rho(|t|)$ w.r.t $|t|$.

Note that the right hand side of second expression could be negative. Under that situation, the ℓ_1 grouped concave penalty performs group selection only as the ℓ_2 grouped concave penalty. Therefore, the bi-level selection feature of the ℓ_1 grouped concave penalty requires a proper λ to achieve the purpose.

Proposition 4.2. ***(Invariance property of the ℓ_1 GMCP)*** *Given a group of column-wise standardized variables with size of d_j, the ℓ_1 GMCP has the following*

invariance property,

$$\rho(\|\boldsymbol{\beta}_j\|_1; d_j\lambda, \gamma) = \rho(d_j\|\boldsymbol{\beta}_j\|_1; \lambda, d_j^2\gamma) \tag{4.3}$$

The invariance property of the ℓ_1 GMCP suggests that for the same level of penalty λ, a modification of penalty and regularization parameters are necessary in order to account for different group size.

4.2 Computation of ℓ_1 Grouped Concave Penalty

Let $\hat{B}^m_{j_k} = (\hat{\beta}^{m+1}_0, ..., (\hat{\boldsymbol{\beta}}^{m+1}_{j-1})^T, \hat{\beta}^{m+1}_{j_1}, ..., \hat{\beta}^{m+1}_{j_k}, \hat{\beta}^m_{j_{k+1}}, ..., \hat{\beta}^m_{j_{d_j}}, (\hat{\boldsymbol{\beta}}^m_{j+1})^T, ..., (\hat{\boldsymbol{\beta}}^m_J)^T)^T$, with $\hat{\boldsymbol{\beta}}^m_j = (\hat{\beta}^m_{j_1}, ..., \hat{\beta}^m_{j_{d_j}})^T$. We want to update $\hat{\boldsymbol{\beta}}^m_j$ to $\hat{\boldsymbol{\beta}}^{m+1}_j$ for $j = 1, ..., J$ and within jth group, we want to update $\hat{\beta}^m_{j_k}$ to $\hat{\beta}^{m+1}_{j_k}$ for $k = 1, ..., d_j$. Using the new notation, we want to update $\hat{B}^m_{j_{k-1}}$ to $\hat{B}^m_{j_k}$. CDA minimizes the criterion function

$$\begin{aligned}
\beta_{j_k}(\lambda, \gamma) &= \underset{\beta_{j_k}}{\operatorname{argmin}}\, Q^1(\beta_{j_k}|\hat{B}^m_{j_{k-1}}) \\
&= \underset{\beta_{j_k}}{\operatorname{argmin}}\{\frac{1}{2n}\sum_{i=1}^n (y_i - (\hat{B}^m_{j_{k-1}})^T\boldsymbol{x}^i + (\hat{\beta}^m_{j_k} - \beta_{j_k})x^i_{j_k})^2 \\
&+ \rho(\sum_{1\le s<j_k} |\hat{\beta}^{m+1}_{j_s}| + |\beta_{j_k}| + \sum_{j_k<s\le d_j} |\hat{\beta}^m_{j_s}|; d_j\lambda, \gamma)\},
\end{aligned} \tag{4.4}$$

by treating (4.1) as a function of β_{j_k} while fixing the rest coordinates. Some simple algebra shows the solution of β_{j_k} for the ℓ_1 GSCAD and GMCP are

$$\ell_1 \text{ GSCAD: } \hat{\beta}^{m+1}_{j_k} = \begin{cases} u_{j_k}, & |u_{j_k}| \ge d_j\lambda\gamma - v_{j_k}, \\ \frac{S\{u_{j_k}, (d_j\lambda\gamma - v_{j_k})/(\gamma-1)\}}{1-1/(\gamma-1)}, & 2d_j\lambda - v_{j_k} \le |u_{j_k}| < d_j\lambda\gamma - v_{j_k}, \\ S(u_{j_k}, d_j\lambda), & |u_{j_k}| \le 2d_j\lambda - v_{j_k}, \end{cases} \tag{4.5}$$

$$\ell_1 \text{ GMCP: } \hat{\beta}^{m+1}_{j_k} = \begin{cases} u_{j_k}, & |u_{j_k}| \ge d_j\lambda\gamma - v_{j_k}, \\ \frac{S\{u_{j_k}, d_j\lambda - v_{j_k}/\gamma\}}{1-1/\gamma}, & |u_{j_k}| < d_j\lambda\gamma - v_{j_k}, \end{cases} \tag{4.6}$$

where $u_{j_k} = \hat{\beta}^m_{j_k} + n^{-1}\sum_{i=1}^n x_{j_k}\{y_i - (\hat{B}^m_{j_{k-1}})^T\boldsymbol{x}^i\} = \hat{\beta}^m_{j_k} + n^{-1}\boldsymbol{x}^T_{j_k}(\boldsymbol{y} - \boldsymbol{X}\hat{B}^m_{j_{k-1}})$, $v_{j_k} = \sum_{1\le s<j_k} |\hat{\beta}^{m+1}_{j_s}| + \sum_{j_k<s\le d_j} |\hat{\beta}^m_{j_s}|$.

Algorithm 4.1 summarizes how to compute the solution of the ℓ_1 grouped concave penalty using the CDA approach.

Algorithm 4.1 Algorithm for ℓ_1 Grouped Concave Penalty in Linear Regression

1. Given an initial value $\hat{\boldsymbol{\beta}}^0$, compute the corresponding linear part $\hat{\boldsymbol{\eta}}^0$.
2. For $m = 0, 1, ...,$ use the form (3.8) in chapter 3 to update the intercept. For the penalized variables, update $\hat{B}^m_{j_k}$ to $\hat{B}^m_{j_{k+1}}$ by using the solution in (4.5) or (4.6) for kth coordinate, $k = 1, ..., d_j$ of jth group. Then repeat the same process for jth group, $j = 1, ..., J$ such that $\hat{\boldsymbol{\beta}}^m$ is updated to $\hat{\boldsymbol{\beta}}^{m+1}$. After each iteration, also compute the corresponding linear component using the latest estimate by the same tips used in MMCD.
3. Check the convergence criterion. If algorithm converges then stop iterations, otherwise repeat step 2 until converges.

For the ℓ_1 grouped concave penalty, the solution surface is also computed along κ. Note that when $\kappa = 0$, both the ℓ_1 GSCAD and GMCP simplify to the Lasso with the penalization proportional to the group sizes (proportional Lasso or ℓ_1 Lasso for short). In the computation, we use the proportional Lasso solution to initialize the solutions surface along grids of κ for a given γ. Based on the solution form of (4.5, 4.6), it is can be shown that $\lambda_{max} = \max_j\{\max_k n^{-1}|\boldsymbol{x}^T_{j_k}(\boldsymbol{y} - \boldsymbol{r})|/d_j\}$ for the ℓ_1 grouped concave penalty, with $\boldsymbol{r} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{int})$ and $\hat{\boldsymbol{\beta}}_{int}$ being the estimate of

the intercept-only model.

Figure 4.1 shows the solution path plots for the proportional Lasso and the ℓ_1 GMCP ($\gamma = 1.2$ and $\gamma = 2.7$) using the same simulated example used in chapter 3. Top row shows the ℓ_1 norm of coefficients, bottom row shows the individual coefficients. Solid and dash lines are the trajectories of causal predictors, dotted lines are the ones for null predictors. The proportional lasso has a smoother trajectories as well.

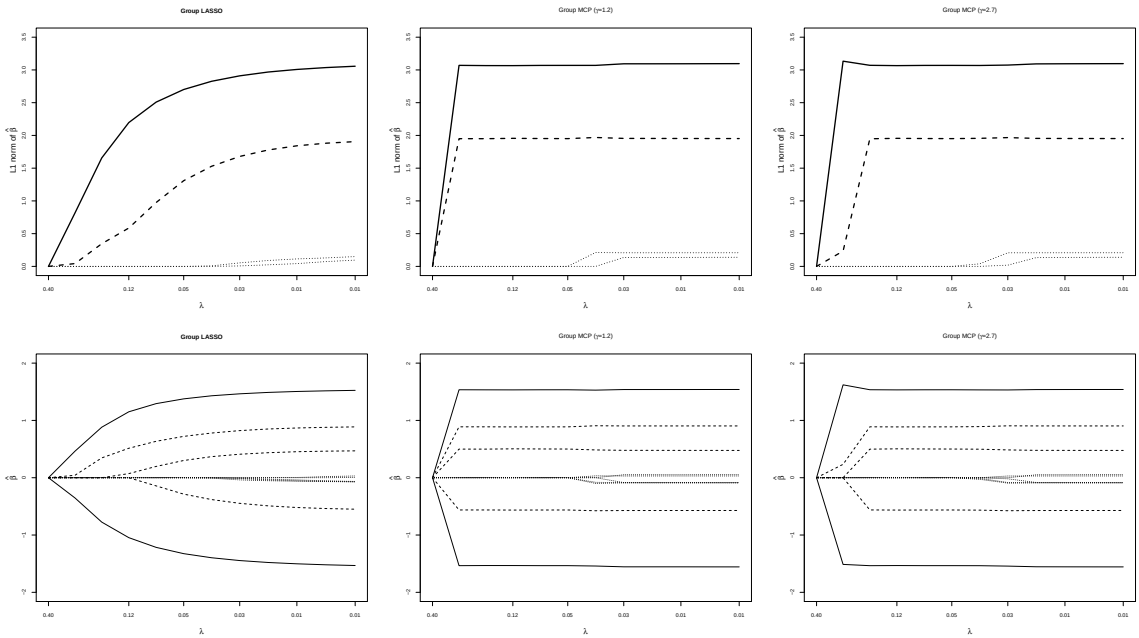

Figure 4.1: Solution paths of the proportional Lasso and the ℓ_1 GMCP. Proportional Lasso (left), ℓ_1 GMCP (center and right).

4.3 Convergence Analysis of Proposed Algorithms

Theorem 4.1, whose proof is provided in Appendix C, establishes that under certain regularity conditions, algorithm 4.1 converge to a minimum of the objective functions for a ℓ_1 grouped concave penalized linear model.

Theorem 4.1. *Consider the objective function in (4.1), where the given data* $(\boldsymbol{y}, \boldsymbol{X})$ *lies on a compact set and no two columns of* $\boldsymbol{X}$ *are identical. Suppose the penalty* $\rho(|t|; \lambda, \gamma) \equiv \rho(t)$ *satisfies* $\rho(t) = \rho(-t)$, $\rho^{'}(|t|)$ *is non-negative, uniformly bounded, with* $\rho^{'}(|t|)$ *being the first derivative (assuming existence) of* $\rho(|t|)$ *w.r.t* $|t|$.

Then the sequence $\{\boldsymbol{\beta}^m\}$ *generated by the algorithm 4.1 converges to a minimum of the function* $Q^1(\boldsymbol{\beta})$ *defined in (4.1).*

4.4 Extension of Grouped Concave Penalty in GLM

The criterion for the ℓ_1 grouped concave penalized GLM model is defined as

$$Q^1(\boldsymbol{\beta}; \lambda, \gamma) \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\{\psi(\boldsymbol{\beta}^T\boldsymbol{x}^i) - y_i\boldsymbol{\beta}^T\boldsymbol{x}^i\} + \sum_{j=1}^{J}\rho(\|\boldsymbol{\beta}_j\|_1; d_j\lambda, \gamma). \tag{4.7}$$

We still apply the MM approach to compute the solutions for the criterion (4.7). For the ℓ_1 grouped concave penalty, assume the following conditions hold:

(i) The second partial derivative of $\ell(\boldsymbol{\beta})$ *w.r.t.* β_{j_k} *is uniformly bounded for column-wise standardized* $\boldsymbol{X}$*, i.e. there exists a real number* $M > 0$ *such that* $\nabla^2_{j_k}\ell(\boldsymbol{\beta}) \leq M$ *for* $j = 1, ..., J$ *and* $k = 1, ..., d_j$.

(ii) $\inf_t \rho^{''}(|t|; \lambda, \gamma) > -M$*, with* $\rho^{''}(|t|; \lambda, \gamma)$ *being the second derivative of* $\rho(|t|; \lambda, \gamma)$ *w.r.t* $|t|$.

The condition (ii) can be met by choosing $\gamma > 1 + 1/M$ for the ℓ_1 GSCAD and

$\gamma > 1/M$ for the ℓ_1 GMCP. Under these two conditions, algorithm 4.1 could be used to compute the solution of (4.7), with coordinate-wise solution forms (4.5, 4.6) being replaced by the follow two forms in GLM models.

$$\ell_1 \text{ GSCAD: } \hat{\beta}_{j_k}^{m+1} = \begin{cases} \frac{u_{j_k}}{M}, & |u_{j_k}| \geq M\left[d_j\lambda\gamma - v_{j_k}\right], \\ \frac{S\{u_{j_k},(d_j\lambda\gamma - v_{j_k})/(\gamma-1)\}}{M-1/(\gamma-1)}, & (1+M)d_j\lambda - Mv_{j_k} \leq |u_{j_k}| < M\left[d_j\lambda\gamma - v_{j_k}\right], \\ \frac{S(u_{j_k},d_j\lambda)}{M}, & |u_{j_k}| \leq (1+M)d_j\lambda - Mv_{j_k}, \end{cases} \tag{4.8}$$

$$\ell_1 \text{ GMCP: } \hat{\beta}_{j_k}^{m+1} = \begin{cases} \frac{u_{j_k}}{M}, & |u_{j_k}| \geq M\left[d_j\lambda\gamma - v_{j_k}\right], \\ \frac{S\{u_{j_k},d_j\lambda - v_{j_k}/\gamma\}}{M-1/\gamma}, & |u_{j_k}| < M\left[d_j\lambda\gamma - v_{j_k}\right], \end{cases} \tag{4.9}$$

where $u_{j_k} = M\hat{\beta}_{j_k}^m + n^{-1}\sum_{i=1}^n x_{j_k}\{y_i - \dot{\psi}\left[(\hat{B}_{j_{k-1}}^m)^T\boldsymbol{x}^i\right]\}$, $v_{j_k} = \sum_{1\leq s<j_k}|\hat{\beta}_{j_s}^{m+1}| + \sum_{j_k<s\leq d_j}|\hat{\beta}_{j_s}^m|$.

Similarly to theorem 4.1, the convergence results could also be established under certain regularity conditions for the GLM model. The following theorem, whose proof is also provided in Appendix C, states that the proposed algorithms converge to a minimum of the objective functions for both ℓ_1 grouped concave penalties in GLM.

Theorem 4.2. *Consider the objective function in (4.7), where the given data* $(\boldsymbol{y}, \boldsymbol{X})$ *lies on a compact set and no two columns of* $\boldsymbol{X}$ *are identical. Suppose the penalty* $\rho(|t|;\lambda,\gamma) \equiv \rho(t)$ *satisfies* $\rho(t) = \rho(-t)$, $\rho^{'}(|t|)$ *is non-negative, uniformly bounded, with* $\rho^{'}(|t|)$ *being the first derivative (assuming existence) of* $\rho(|t|)$ *w.r.t* $|t|$. *Also assume the two conditions required in the computation hold.*

Then the sequence $\{\boldsymbol{\beta}^m\}$ *generated by the aforementioned algorithm converges to a minimum of the function* $Q^1(\boldsymbol{\beta})$ *defined in (4.7).*

4.5 Simulation Studies in Linear and Logistic Models

We use the same design in the ℓ_2 grouped concave penalty in this section. However, we consider two different settings for $\boldsymbol{a}$ as we examine the finite-sample performance of the ℓ_1 grouped concave penalty.

- 4. $\boldsymbol{a} = (0, 0, 0.2, 0.25, 0.5, 0, 0, 0.2, 0.25, 0.5)^T$ to represent the situation that the effects of group members are small to median with some groups having zero effects.
- 5. $\boldsymbol{a} = (0, 0, 0.1, 0.1, 0.7, 0, 0, 0.1, 0.1, 0.7)^T$ to represent the situation that the only one or two members within the group have strong effect with some members have null effect.

We also use the validation approach to select the final model for comparison. Table 4.1 shows the comparison between the ℓ_1 grouped concave penalty vs. the ungrouped concave penalty in linear models; while table 4.2 shows the results in logistic models. The result of both models are similar. Under both settings, the ℓ_1 GSCAD and GMCP have smaller GMS and GFDR than the SCAD and MCP. Under setting 4 when no strong effect variable dominating the group effect, the ℓ_1 GSCAD and GMCP have smaller MS and FDR. While under setting 5 when some strong effect variables dominating the group effect, the SCAD and MCP tend to select a smaller model with similar or slightly better FDR than the corresponding grouped version. In terms of predictive performance, the ℓ_1 GSCAD and GMCP in general outperform the SCAD and MCP under both settings, with more obvious out performance under setting 4.

Table 4.1: Comparison of the ℓ_1 grouped vs. the ungrouped concave penalties in linear models, $n = 300$, $p = 500$, and $\rho = 0.6$.

Setting	Group (SNR)	Method	PMSE (SE*10^3)	GMS (SE*10)	GFDR (SE*10^3)	MS (SE*10)	FDR (SE*10^3)
4	EQU 3.04	Lasso	1.41 (4.1)	35.71 (1.9)	0.83 (1.0)	90.44 (4.9)	0.60 (2.2)
		SCAD	1.39 (3.8)	34.10 (2.0)	0.82 (1.1)	84.09 (5.0)	0.57 (2.5)
		MCP	1.39 (3.8)	30.74 (3.1)	0.79 (3.1)	76.69 (7.8)	0.52 (4.7)
		ℓ_1Lasso	1.41 (4.1)	35.71 (1.9)	0.83 (1.0)	90.44 (4.9)	0.60 (2.2)
		ℓ_1GSCAD	1.22 (2.7)	8.61 (1.1)	0.25 (8.7)	54.08 (2.4)	0.33 (2.9)
		ℓ_1GMCP	1.23 (2.7)	6.16 (0.5)	0.01 (3.4)	55.63 (3.0)	0.34 (3.8)
	UNE 2.55	Lasso	1.58 (5.1)	35.62 (1.1)	0.86 (0.5)	107.89 (6.1)	0.68 (1.6)
		SCAD	1.50 (4.5)	32.40 (1.6)	0.84 (0.9)	88.00 (6.6)	0.62 (2.4)
		MCP	1.50 (4.7)	26.44 (3.4)	0.78 (5.1)	69.95 (10.3)	0.50 (6.4)
		ℓ_1Lasso	1.66 (5.4)	35.83 (1.1)	0.86 (0.5)	112.89 (6.7)	0.70 (1.6)
		ℓ_1GSCAD	1.24 (2.7)	6.05 (0.8)	0.12 (7.9)	60.96 (1.1)	0.41 (1.0)
		ℓ_1GMCP	1.24 (2.7)	5.57 (0.6)	0.07 (6.4)	60.66 (0.8)	0.41 (0.7)
5	EQU 3.13	Lasso	1.33 (3.4)	33.49 (1.9)	0.82 (1.1)	77.77 (4.5)	0.61 (2.3)
		SCAD	1.26 (2.3)	26.43 (2.5)	0.76 (2.7)	56.02 (5.6)	0.54 (3.3)
		MCP	1.27 (2.3)	21.28 (3.4)	0.68 (6.0)	45.38 (7.9)	0.45 (5.6)
		ℓ_1Lasso	1.33 (3.4)	33.49 (1.9)	0.82 (1.1)	77.77 (4.5)	0.61 (2.3)
		ℓ_1GSCAD	1.19 (2.4)	9.73 (1.1)	0.34 (7.9)	48.76 (2.5)	0.33 (2.6)
		ℓ_1GMCP	1.21 (2.6)	6.24 (1.1)	0.01 (3.5)	48.11 (4.1)	0.30 (4.0)
	UNE 2.71	Lasso	1.39 (3.5)	32.30 (1.4)	0.84 (0.7)	80.96 (5.0)	0.68 (1.8)
		SCAD	1.25 (1.9)	19.50 (2.6)	0.70 (6.9)	41.16 (5.8)	0.49 (5.8)
		MCP	1.25 (1.9)	13.89 (2.8)	0.54 (11.1)	29.79 (5.8)	0.34 (7.9)
		ℓ_1Lasso	1.43 (3.8)	32.45 (1.3)	0.84 (0.7)	85.66 (5.3)	0.71 (1.7)
		ℓ_1GSCAD	1.23 (2.5)	7.29 (1.3)	0.23 (10.4)	56.15 (3.5)	0.40 (2.2)
		ℓ_1GMCP	1.24 (2.6)	5.55 (1.1)	0.05 (6.1)	59.62 (1.9)	0.40 (1.4)

We study the performance of ℓ_1 GSCAD and GMCP vs. the ℓ_2 GSCAD and GMCP under the setting 1, 2 and 3 in order to understand their empirical performance. Table 4.3 shows the comparison of the ℓ_1 GSCAD and GMCP vs. the ℓ_2 GSCAD and GMCP in linear models, and table 4.4 shows the comparison of two grouped concave penalty in logistic models. We observe similar results in two models. Two grouped concave penalties have similar predictive performance under setting (1) and (2). The ℓ_1 grouped concave penalties seems to have a slightly advantage over the

Table 4.2: Comparison of the ℓ_1 grouped vs. the ungrouped concave penalties in logistic models, $n = 300$, $p = 500$, and $\rho = 0.6$.

Setting	Group (SNR)	Method	PAUC (SE*10^3)	GMS (SE*10)	GFDR (SE*10^3)	MS (SE*10)	FDR (SE*10^3)
4	EQU 3.04	Lasso	0.852(0.62)	24.96(2.6)	0.75(2.9)	54.70(5.3)	0.55(3.4)
		SCAD	0.852(0.62)	24.90(2.6)	0.74(3.0)	54.57(5.3)	0.55(3.5)
		MCP	0.852(0.62)	24.91(2.6)	0.75(2.9)	54.60(5.3)	0.55(3.4)
		ℓ_1Lasso	0.852(0.62)	24.96(2.6)	0.75(2.9)	54.70(5.3)	0.55(3.4)
		ℓ_1GSCAD	0.874(0.50)	8.70(1.7)	0.23(9.9)	56.08(4.6)	0.40(2.7)
		ℓ_1GMCP	0.873(0.50)	7.97(1.7)	0.17(8.9)	56.88(4.5)	0.39(3.2)
	UNE 2.55	Lasso	0.814(0.73)	20.48(2.8)	0.73(3.9)	44.57(6.7)	0.55(4.7)
		SCAD	0.814(0.72)	20.20(2.8)	0.73(4.4)	43.72(6.7)	0.55(4.9)
		MCP	0.814(0.72)	20.10(2.8)	0.72(5.0)	43.59(6.9)	0.54(5.2)
		ℓ_1Lasso	0.811(0.71)	20.62(2.3)	0.74(2.9)	45.73(5.7)	0.58(3.7)
		ℓ_1GSCAD	0.828(0.84)	10.41(2.7)	0.41(11.7)	56.74(7.5)	0.49(4.0)
		ℓ_1GMCP	0.827(0.84)	9.72(2.7)	0.36(12.0)	56.97(8.2)	0.48(4.1)
5	EQU 3.13	Lasso	0.875(0.68)	20.31(2.5)	0.68(4.0)	40.16(4.6)	0.53(4.2)
		SCAD	0.878(0.73)	17.19(2.9)	0.59(8.1)	32.51(5.8)	0.45(7.0)
		MCP	0.878(0.73)	16.31(3.1)	0.55(10.2)	31.15(6.1)	0.42(8.4)
		ℓ_1Lasso	0.875(0.68)	20.31(2.5)	0.68(4.0)	40.16(4.6)	0.53(4.2)
		ℓ_1GSCAD	0.885(0.57)	10.94(1.9)	0.38(9.6)	40.78(5.9)	0.38(3.9)
		ℓ_1GMCP	0.884(0.57)	9.89(2.2)	0.29(10.4)	40.49(6.7)	0.36(5.0)
	UNE 2.71	Lasso	0.839(0.73)	20.90(2.8)	0.73(4.5)	41.80(6.6)	0.59(4.8)
		SCAD	0.847(0.87)	16.37(3.0)	0.63(8.0)	29.89(6.4)	0.49(7.5)
		MCP	0.847(0.87)	15.23(3.3)	0.57(10.6)	28.15(6.9)	0.44(9.2)
		ℓ_1Lasso	0.832(0.73)	21.79(2.8)	0.75(3.9)	45.55(7.3)	0.63(4.3)
		ℓ_1GSCAD	0.843(0.81)	11.67(3.2)	0.43(12.2)	55.94(7.6)	0.50(4.8)
		ℓ_1GMCP	0.842(0.82)	11.38(3.4)	0.39(13.0)	56.64(7.2)	0.50(5.1)

ℓ_2 concave grouped concave penalties under setting (3), where some group members have much strong effect than the rest members. The ℓ_2 GSCAD and GMCP tend to have a larger model size than the ℓ_1 GSCAD and GMCP. The FDR of the ℓ_1 GSCAD and GMCP is generally smaller than the ℓ_2 grouped version due to their relative smaller model size. In terms of GMS and GFDR, two grouped concave penalties are similar without a clear winner.

Table 4.3: Comparison of the ℓ_1 GSCAD and GMCP vs. the ℓ_2 GSCAD and GMCP in linear models, $n = 300$, $p = 500$ and $\rho = 0.6$.

Setting	Group (SNR)	Method	PMSE (SE*10^3)	GMS (SE*10)	GFDR (SE*10^3)	MS (SE*10)	FDR (SE*10^3)
1	EQU 3.56	ℓ_1Lasso	1.57 (5.2)	37.76 (1.8)	0.84 (0.8)	112.95 (5.1)	0.47 (2.3)
		ℓ_1GSCAD	1.24 (2.8)	6.57 (0.8)	0.05 (6.1)	60.61 (1.0)	0.01 (1.4)
		ℓ_1GMCP	1.24 (2.8)	6.19 (0.4)	0.02 (3.6)	60.19 (0.4)	0.00 (0.7)
		ℓ_2Lasso	1.64 (6.3)	27.91 (0.6)	0.78 (0.5)	279.10 (6.2)	0.78 (0.5)
		ℓ_2GSCAD	1.25 (2.8)	7.57 (1.0)	0.15 (8.2)	75.70 (10.2)	0.15 (8.2)
		ℓ_2GMCP	1.25 (2.8)	7.03 (0.7)	0.11 (6.7)	70.28 (7.2)	0.11 (6.7)
	UNE 2.88	ℓ_1Lasso	2.01 (6.3)	35.96 (1.6)	0.86 (0.8)	130.49 (10.2)	0.59 (2.6)
		ℓ_1GSCAD	1.24 (2.7)	6.18 (0.9)	0.13 (8.4)	61.37 (1.2)	0.02 (1.7)
		ℓ_1GMCP	1.24 (2.6)	5.49 (0.6)	0.06 (6.0)	60.57 (0.7)	0.01 (1.0)
		ℓ_2Lasso	1.56 (5.2)	22.22 (0.6)	0.77 (0.7)	277.37 (7.4)	0.78 (0.6)
		ℓ_2GSCAD	1.24 (2.6)	6.40 (1.0)	0.16 (8.5)	77.03 (11.9)	0.16 (8.7)
		ℓ_2GMCP	1.24 (2.6)	5.62 (0.6)	0.08 (6.1)	67.33 (7.0)	0.08 (6.1)
2	EQU 3.13	ℓ_1Lasso	1.42 (4.1)	35.72 (1.9)	0.83 (1.0)	93.25 (4.9)	0.51 (2.6)
		ℓ_1GSCAD	1.22 (2.6)	8.39 (1.0)	0.24 (8.3)	55.45 (1.9)	0.05 (1.9)
		ℓ_1GMCP	1.23 (2.7)	6.11 (0.3)	0.01 (2.7)	56.24 (2.4)	0.00 (0.6)
		ℓ_2Lasso	1.62 (6.0)	27.79 (0.7)	0.78 (0.6)	277.94 (6.9)	0.78 (0.6)
		ℓ_2GSCAD	1.25 (2.8)	7.74 (1.2)	0.16 (8.8)	77.44 (11.9)	0.16 (8.8)
		ℓ_2GMCP	1.25 (2.8)	7.15 (0.9)	0.12 (7.5)	71.54 (8.7)	0.12 (7.5)
	UNE 2.61	ℓ_1Lasso	1.67 (5.5)	35.83 (1.1)	0.86 (0.5)	114.29 (6.6)	0.64 (1.8)
		ℓ_1GSCAD	1.24 (2.7)	5.99 (0.8)	0.11 (7.9)	60.97 (1.0)	0.02 (1.4)
		ℓ_1GMCP	1.24 (2.7)	5.53 (0.6)	0.06 (6.1)	60.61 (0.7)	0.01 (1.1)
		ℓ_2Lasso	1.54 (5.2)	22.15 (0.7)	0.77 (0.8)	276.42 (8.0)	0.78 (0.7)
		ℓ_2GSCAD	1.24 (2.6)	6.40 (0.9)	0.16 (8.5)	77.11 (11.9)	0.16 (8.7)
		ℓ_2GMCP	1.24 (2.6)	5.74 (0.7)	0.09 (6.7)	68.89 (8.8)	0.09 (6.8)
3	EQU 3.21	ℓ_1Lasso	1.35 (3.4)	33.93 (1.9)	0.82 (1.1)	82.28 (4.6)	0.51 (2.6)
		ℓ_1GSCAD	1.20 (2.4)	9.35 (1.1)	0.31 (8.0)	51.03 (2.3)	0.07 (2.2)
		ℓ_1GMCP	1.21 (2.6)	6.20 (1.0)	0.01 (3.2)	50.42 (3.6)	0.00 (1.8)
		ℓ_2Lasso	1.62 (6.0)	27.80 (0.7)	0.78 (0.7)	277.98 (7.3)	0.78 (0.7)
		ℓ_2GSCAD	1.25 (2.8)	7.69 (1.2)	0.16 (8.8)	76.88 (11.7)	0.16 (8.8)
		ℓ_2GMCP	1.25 (2.8)	7.22 (0.9)	0.12 (7.7)	72.20 (9.1)	0.12 (7.7)
	UNE 2.76	ℓ_1Lasso	1.44 (3.9)	32.62 (1.3)	0.85 (0.7)	88.31 (5.2)	0.63 (2.1)
		ℓ_1GSCAD	1.23 (2.5)	6.92 (1.1)	0.20 (10.1)	56.57 (3.1)	0.04 (2.5)
		ℓ_1GMCP	1.24 (2.6)	5.44 (0.9)	0.04 (5.5)	59.69 (1.7)	0.01 (2.0)
		ℓ_2Lasso	1.55 (5.3)	22.03 (0.7)	0.77 (0.8)	275.23 (8.5)	0.78 (0.8)
		ℓ_2GSCAD	1.24 (2.6)	6.30 (0.9)	0.15 (8.6)	75.86 (11.4)	0.15 (8.7)
		ℓ_2GMCP	1.24 (2.6)	5.74 (0.7)	0.09 (7.0)	68.95 (9.3)	0.08 (7.1)

Table 4.4: Comparison of the ℓ_1 GSCAD and GMCP vs. the ℓ_2 GSCAD and GMCP in logistic models, $n = 300$, $p = 500$ and $\rho = 0.6$.

Setting	Group (SNR)	Method	PAUC (SE*10^3)	GMS (SE*10)	GFDR (SE*10^3)	MS (SE*10)	FDR (SE*10^3)
1	EQU 3.56	ℓ_1Lasso	0.865(0.55)	27.75(2.6)	0.77(2.2)	68.87(6.6)	0.44(3.5)
		ℓ_1GSCAD	0.896(0.46)	7.68(1.1)	0.16(8.4)	61.25(2.2)	0.03(2.1)
		ℓ_1GMCP	0.896(0.47)	7.17(0.8)	0.13(6.9)	61.24(1.6)	0.02(1.6)
		ℓ_2Lasso	0.855(0.76)	15.70(1.7)	0.60(4.4)	157.04(16.8)	0.60(4.4)
		ℓ_2GSCAD	0.888(0.56)	6.99(0.7)	0.11(6.5)	69.86(7.3)	0.11(6.5)
		ℓ_2GMCP	0.888(0.56)	6.70(0.7)	0.07(5.9)	66.98(7.2)	0.07(5.9)
	UNE 2.88	ℓ_1Lasso	0.831(0.61)	22.29(2.1)	0.77(2.4)	55.48(5.1)	0.47(3.7)
		ℓ_1GSCAD	0.852(0.77)	9.58(2.1)	0.39(10.6)	61.54(8.1)	0.17(6.7)
		ℓ_1GMCP	0.851(0.78)	8.80(2.1)	0.34(10.8)	61.97(8.7)	0.16(7.0)
		ℓ_2Lasso	0.838(0.74)	12.73(1.5)	0.58(5.2)	155.92(19.6)	0.59(5.3)
		ℓ_2GSCAD	0.857(0.70)	7.89(1.5)	0.28(10.3)	95.40(19.3)	0.29(10.5)
		ℓ_2GMCP	0.857(0.70)	6.78(1.2)	0.20(9.6)	81.55(15.2)	0.20(9.7)
2	EQU 3.13	ℓ_1Lasso	0.856(0.62)	25.05(2.6)	0.75(2.8)	56.10(5.4)	0.44(3.8)
		ℓ_1GSCAD	0.878(0.49)	8.67(1.7)	0.22(9.7)	56.86(4.7)	0.06(3.8)
		ℓ_1GMCP	0.877(0.50)	7.89(1.7)	0.16(8.7)	57.97(4.2)	0.04(3.6)
		ℓ_2Lasso	0.837(0.80)	15.86(1.7)	0.60(4.7)	158.56(17.4)	0.60(4.7)
		ℓ_2GSCAD	0.869(0.65)	7.33(1.0)	0.13(7.6)	73.30(10.5)	0.13(7.6)
		ℓ_2GMCP	0.869(0.65)	6.88(0.9)	0.09(6.7)	68.78(8.5)	0.09(6.7)
	UNE 2.61	ℓ_1Lasso	0.817(0.70)	20.88(2.3)	0.75(2.8)	47.23(5.6)	0.50(4.1)
		ℓ_1GSCAD	0.833(0.82)	10.38(2.6)	0.41(11.5)	57.19(8.1)	0.20(7.8)
		ℓ_1GMCP	0.832(0.83)	9.98(2.7)	0.37(12.0)	58.32(8.7)	0.19(8.3)
		ℓ_2Lasso	0.820(0.82)	12.42(1.6)	0.57(5.9)	151.71(20.8)	0.57(6.1)
		ℓ_2GSCAD	0.841(0.76)	7.61(1.4)	0.27(10.1)	91.75(17.9)	0.27(10.2)
		ℓ_2GMCP	0.841(0.77)	6.56(1.2)	0.17(9.5)	78.48(14.8)	0.17(9.6)
3	EQU 3.21	ℓ_1Lasso	0.878(0.67)	20.69(2.4)	0.69(3.8)	41.63(4.5)	0.43(4.4)
		ℓ_1GSCAD	0.889(0.55)	10.45(1.8)	0.35(9.3)	41.33(5.9)	0.14(5.2)
		ℓ_1GMCP	0.887(0.55)	9.36(2.0)	0.26(9.9)	41.32(6.8)	0.11(5.8)
		ℓ_2Lasso	0.843(0.83)	15.16(1.7)	0.58(4.7)	151.58(17.4)	0.58(4.7)
		ℓ_2GSCAD	0.873(0.64)	7.15(1.0)	0.12(7.2)	71.50(9.5)	0.12(7.2)
		ℓ_2GMCP	0.873(0.63)	6.74(0.8)	0.08(6.1)	67.38(8.3)	0.08(6.1)
	UNE 2.76	ℓ_1Lasso	0.836(0.73)	21.52(2.8)	0.75(3.9)	45.21(7.1)	0.54(4.8)
		ℓ_1GSCAD	0.847(0.77)	11.47(3.1)	0.42(12.0)	55.43(7.1)	0.22(9.8)
		ℓ_1GMCP	0.847(0.77)	11.17(3.4)	0.38(13.0)	56.35(6.8)	0.22(10.3)
		ℓ_2Lasso	0.830(0.80)	12.20(1.4)	0.56(5.3)	149.62(18.2)	0.57(5.4)
		ℓ_2GSCAD	0.852(0.68)	7.01(1.3)	0.21(9.7)	84.53(15.8)	0.21(9.8)
		ℓ_2GMCP	0.852(0.68)	6.35(1.1)	0.14(8.8)	76.44(13.8)	0.14(8.9)

CHAPTER 5
APPLICATION OF GROUPED PEANALTY IN HEALTH CARE AND GENOMIC DATASETS

To further explore the performance of the ℓ_1 and ℓ_2 grouped concave penalties, we apply the grouped concave penalties together with the ungrouped concave penalties on two motivation datasets. The first dataset, named as WW with $n = 964$, comes from a community intervention project. The project provides education sections on risk factors of cardiovascular diseases (CVD), nutrition of food, etc. to low-income Iowan women in order to increase their awareness of CVD and lower their risk of CVD. For our purpose, we model the standardized relative change of body mass index (BMI) from the baseline as the continuous outcome. We put the intercept, baseline BMI, intervention status (participation in education sections or not), age and race as the unpenalized variables. The set of penalized variables includes $p = 79$ predictors, which are further classified into 11 groups describing social economics status (6 variables), disease and medication history (8 variables), family structure (9 variables), work load (7 variables), intension of weight control (3 variables), high fiber and vitamin food consumption (7 variables), high fat food consumption (10 variables), high energy food consumption(10 variables), smoking status (4 variables), blood test at baseline (6 variables), knowledge of nutrition (9 variables). It should be noted that such group structure is loosely defined by design.

The second dataset, named as BC with $n = 295$, is the same dataset used in chapter 2. A subset of 500 genes with the highest Spearman correlation to the

metastasis within five years status are used in the computation. We use the Gap statistic to determine the optimal group structure. For details about Gap, we refer to Tibshirani, Walther and Hastie (2001); Ma and Huang (2007). We choose the hierarchical cluster method to compute the Gap statistic. For our 500 genes, we end up with 33 groups with the maximum group size of 68 and the minimum group size of 2.

5.1 Comparison Based on Random Partition

For both studies, we randomly partition the whole dataset into a training (approximately 1/3 of the observations) and validation datasets (approximately 2/3 of the observations). The solution surface is computed solely based on the training dataset; the solution corresponding to the minimum PMSE or maximum PAUC of the validation dataset is chosen for comparison. The partition process is repeated for $1,500$ times.

Table 5.1 presents the comparison results for the WW dataset. We observe that all the methods have very similar PMSE. The ℓ_2 grouped concave penalties have the largest model size at individual level, but the smallest model size at group level. Table 5.2 presents the comparison results for the BC dataset. We see that the ℓ_1 GSCAD and GMCP have the highest PAUC. The ℓ_2 GSCAD and GMCP have the smallest GMS, but at individual level, their model size is about twice of those for the ℓ_1 GSCAD and GMCP.

Table 5.1: Comparison of the grouped vs. the ungrouped concave penalties in WW dataset.

Method	PMSE(SE*10^2)	GMS (SE*10)	MS (SE*10)
Lasso	0.987 (0.194)	2.718 (0.549)	6.283 (0.835)
SCAD	0.987 (0.194)	2.727 (0.551)	6.300 (0.840)
MCP	0.987 (0.194)	2.453 (0.489)	5.854 (0.723)
ℓ_1Lasso	0.987 (0.194)	3.127 (0.536)	6.955 (0.868)
ℓ_1SCAD	0.987 (0.194)	3.131 (0.535)	6.984 (0.884)
ℓ_1MCP	0.986 (0.194)	2.937 (0.505)	6.742 (0.834)
ℓ_2Lasso	0.988 (0.194)	2.293 (0.567)	10.305 (2.918)
ℓ_2SCAD	0.988 (0.194)	2.299 (0.568)	10.320 (2.921)
ℓ_2MCP	0.988 (0.194)	2.109 (0.506)	9.433 (2.622)

5.2 Results Using Tuning Parameter Selection

We present the results for the WW and BC datasets using the whole dataset as the training samples. For linear models, we use the cross validation to select the tuning parameters (λ, κ). For logistic models, we use the CV-AUC approach introduced in chapter 2 for tuning parameter selection. We use 5-fold cross validation procedure for both datasets.

For WW dataset, the ungrouped concave penalty (Lasso, SCAD and MCP) all identified that the possession of a health insurance at baseline is helpful in controlling an increased BMI at follow-up, with cross-validated PMSE 0.979. Using the ℓ_1 grouped concave penalties (including the ℓ_1 Lasso, GSCAD and GMCP), we identified that the health insurgence status, LDL level and triglycerides level are protective factors from an increased BMI at follow-up, while the target weight and consumption

Table 5.2: Comparison of the grouped vs. the ungrouped concave penalties in BC dataset.

Method	PMSE(SE*10^3)	GMS (SE*10)	MS (SE*10)
Lasso	0.756 (0.66)	21.96 (1.34)	41.23 (5.58)
SCAD	0.757 (0.66)	20.51 (1.52)	36.29 (5.10)
MCP	0.757 (0.66)	20.40 (1.54)	36.07 (5.17)
ℓ_1Lasso	0.801 (0.63)	15.33 (1.19)	31.06 (3.63)
ℓ_1GSCAD	0.806 (0.60)	12.16 (1.20)	35.24 (5.46)
ℓ_1GMCP	0.806 (0.60)	12.11 (1.23)	35.44 (5.44)
ℓ_2Lasso	0.753 (0.81)	8.80 (0.63)	63.84 (6.16)
ℓ_2GSCAD	0.759 (0.78)	7.75 (0.69)	54.59 (6.71)
ℓ_2GMCP	0.759 (0.78)	7.44 (0.74)	51.50 (7.14)

of cigarettes at baseline are positively associated with the increased BMI at follow-up, with cross-validated MSE 0.977. These 5 identified factors belongs to 4 groups: social economics status (possession of health insurance), blood test at baseline(LDL and triglycerides level), intension of weight control (target weight) and smoking status (consumption of cigarettes). Using the ℓ_2 grouped concave penalty (including the ℓ_2 Lasso, GSCAD and GMCP), we identified no predictors from the pool of penalized variables, with cross-validated MSE 0.979.

Table 5.3 presents the results for BC dataset. The ungrouped concave penalties (Lasso, SCAD and MCP) end up with the same model. The ℓ_1 GSCAD and GMCP have the best CV-AUC, followed by ℓ_2 GSCAD and GMCP. In terms of group model size, we see a consistent result with simulation: the ℓ_2 GSCAD and GMCP prefer a small group model size.

In the application of WW dataset, the grouped concave penalties give a different set of causal variables but similar cross-validated PMSE, while in the BC dataset, we observe that the grouped concave penalties achieve a higher CV-AUC with smaller model size.

Table 5.3: Data analysis results for the BC dataset using the CV-AUC tuning parameter selection approach.

Method	CV-AUC	GMS	MS
Lasso	0.757	30	66
SCAD	0.757	30	66
MCP	0.757	30	66
ℓ_1Lasso	0.815	12	27
ℓ_1SCAD	0.830	7	28
ℓ_1MCP	0.831	7	28
ℓ_2Lasso	0.796	12	71
ℓ_2SCAD	0.821	5	20
ℓ_2MCP	0.821	5	20

CHAPTER 6
CONCLUSION AND DISCUSSION

The MMCD algorithm seeks a closed form solution for each coordinate by using the exact penalty term. This makes it differ from the existing algorithms, such as the LQA, LLA and MIST that approximate the penalty by a majorization function. The majorization of MMCD is, however, applied to the loss function coordinate-wisely to avoid the computation of scaling factors. This approach increases the efficiency of CDA in high-dimensional settings. The theoretical convergence property of the MMCD algorithm is also established under certain regularity conditions.

The comparison with the adaptive rescaling approach indicates that the MMCD is more efficient in $p \gg n$ settings. Simulation and data analysis also reveal the adequacy of the MMCD algorithm. Based on the MMCD solutions of penalized logistic regression, we compare Lasso and the concave penalties including SCAD and MCP for their empirical performance. The MCP seems to be a better choice in terms of predictive AUC and FDR in the simulated models we considered.

The application of the MMCD algorithm to the logistic regression is facilitated by the fact that a simple and effective majorization function can be constructed for the likelihood. However, in some other important GLM models such as the log-linear model, it appears that no simple majorization function exists. One possible approach is to design a sequence of majorization functions according to the solutions at each iteration. This is an interesting problem that requires further investigation.

The grouped concave penalties incorporate the group information of predictors

into the regression. Our results suggest that incorporating the group information improves the selection and estimation under the situations when no members dominate the group effect.

With the proposed grouped penalties, a practical question is how to determine the group structure appropriately. This is still an open question. We offer several possibilities listed below. The first one is based on the design of questionnaire as we did for the WW study. The group information is loosely defined by the design of questionnaires, i.e. a block of questions measuring similar quantity of the study subject. Such group structure is more consistent to the perception of scholars and delivers an easier interpretation, although no statistical procedure justifies the structure. A second approach is to perform a numerical exploration using index like the Gap statistic as we did for BC dataset. Such structure offers more calibration of correlation among predictors, therefore provides an improved overall performance of penalized regression. The group structure, however may not be easy to interpret. A third way, which is more specific to the genomic data, is to use the available information on genes. The Gene Oncology (GO) and multiple databases on biological pathway information would be a good start to collect such group information.

Given current framework, the grouped penalties can be easily generalized to other regression models, such as GEE, mixed models and Cox's proportional hazard model. Another generalization of current group penalty is to use the $\ell_p, 1 \leq p \leq 2$ norm. These extensions point to future researches.

APPENDIX A
PROOF OF THEOREM IN CHAPTER 2

Our proof is similar to that of Mazumder, Friedman and Hastie (2011). The main differences are that we need to take care of the intercept in lemma A.1 and theorem 2.1, the quadratic approximation to the loss function and the coordinate-wise majorization in theorem 2.1. We first present lemma A.1, which is used in the proof of theorem 2.1.

Lemma A.1. *Suppose the data* $(\boldsymbol{y}, X)$ *lies on a compact set and the following conditions hold:*

1. The loss function $\ell(\boldsymbol{\beta})$ *is (total) differentiable w.r.t.* $\boldsymbol{\beta}$ *for any* $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$.

2. The penalty function $\rho(t)$ *is symmetric around 0 and is differentiable on* $t \geq 0$*;* $\rho^{'}(|t|)$ *is non-negative, continuous and uniformly bounded, where* $\rho^{'}(|t|)$ *is the derivative of* $\rho(|t|)$ *w.r.t.* $|t|$.

3. The sequence $\{\boldsymbol{\beta}^k\}$ *is bounded.*

4. For every convergent subsequence $\{\boldsymbol{\beta}^{n_k}\} \subset \{\boldsymbol{\beta}^n\}$*, the successive differences converge to zero:* $\boldsymbol{\beta}^{n_k} - \boldsymbol{\beta}^{n_k - 1} \to 0$.

Then if $\boldsymbol{\beta}^\infty$ *is any limit point of the sequence* $\{\boldsymbol{\beta}^k\}$*, then* $\boldsymbol{\beta}^\infty$ *is a minimum for the function* $Q(\boldsymbol{\beta})$*; i.e.*

$$\liminf_{\alpha \downarrow 0+} \{ \frac{Q(\boldsymbol{\beta}^\infty + \alpha \boldsymbol{\delta}) - Q(\boldsymbol{\beta}^\infty)}{\alpha} \} \geq 0, \tag{A.1}$$

for any $\boldsymbol{\delta} = (\delta_0, ..., \delta_p) \in \mathbb{R}^{p+1}$.

Proof. For any $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^T$ and $\boldsymbol{\delta}_j = (0, ..., \delta_j, ..., 0) \in \mathbb{R}^{p+1}$, we have

$$\begin{aligned} \liminf_{\alpha \downarrow 0+} \{ \frac{Q(\boldsymbol{\beta} + \alpha \boldsymbol{\delta}_j) - Q(\boldsymbol{\beta})}{\alpha} \} &= \nabla_j \ell(\boldsymbol{\beta}) \delta_j + \liminf_{\alpha \downarrow 0+} \{ \frac{\rho(|\beta_j + \alpha \delta_j|) - \rho(|\beta_j|)}{\alpha} \} \\ &= \nabla_j \ell(\boldsymbol{\beta}) \delta_j + \partial \rho(\beta_j; \delta_j), \end{aligned} \tag{A.2}$$

for $j \in \{1, ..., p\}$, with

$$\partial \rho(\beta_j; \delta_j) = \begin{cases} \rho^{'}(|\beta_j|) \mathrm{sgn}(\beta_j) \delta_j, & |\beta_j| > 0; \\ \rho^{'}(0) |\delta_j|, & |\beta_j| = 0, \end{cases} \tag{A.3}$$

where

$$\mathrm{sgn}(x) = \begin{cases} 1, & \text{if } x > 0; \\ -1, & \text{if } x < 0; \\ \text{any } u \in (-1, 1), & \text{if } x = 0. \end{cases}$$

Assume $\boldsymbol{\beta}^{n_k} \to \boldsymbol{\beta}^{\infty} = (\beta_0^{\infty}, ..., \beta_p^{\infty})$, and by condition 4, as $k \to \infty$

$$\boldsymbol{\beta}_j^{n_k - 1} = (\beta_0^{n_k}, ..., \beta_{j-1}^{n_k}, \beta_j^{n_k}, \beta_{j+1}^{n_k - 1}, ..., \beta_p^{n_k - 1}) \to (\beta_0^{\infty}, ..., \beta_{j-1}^{\infty}, \beta_j^{\infty}, \beta_{j+1}^{\infty}, ..., \beta_p^{\infty}) \tag{A.4}$$

By (A.3) and (A.4), we have the results below for $j \in \{1, ..., p\}$.

$$\partial \rho(\beta_j^{n_k}; \delta_j) \to \partial \rho(\beta_j^{\infty}; \delta_j), \text{ if } \beta_j^{\infty} \neq 0; \;\; \partial \rho(\beta_j^{\infty}; \delta_j) \geq \liminf_k \partial \rho(\beta_j^{n_k}; \delta_j), \text{ if } \beta_j^{\infty} = 0. \tag{A.5}$$

By the coordinate-wise minimum of jth coordinate $j \in \{1, ..., p\}$, we have

$$\nabla_j \ell(\boldsymbol{\beta}_j^{n_k - 1}) \delta_j + \partial \rho(\beta_j^{n_k}; \delta_j) \geq 0, \text{ for all } k. \tag{A.6}$$

Thus (A.5, A.6) implies that for all $j \in \{1, ..., p\}$,

$$\nabla_j \ell(\boldsymbol{\beta}^{\infty}) \delta_j + \partial \rho(\beta_j^{\infty}; \delta_j) \geq \liminf_k \{ \nabla_j \ell(\boldsymbol{\beta}_j^{n_k - 1}) \delta_j + \partial \rho(\beta_j^{n_k}; \delta_j) \} \geq 0. \tag{A.7}$$

By (A.2,A.7), for $j \in \{1, ..., p\}$, we have

$$\liminf_{\alpha \downarrow 0+} \{ \frac{Q(\boldsymbol{\beta}^{\infty} + \alpha \boldsymbol{\delta}_j) - Q(\boldsymbol{\beta}^{\infty})}{\alpha} \} \geq 0. \tag{A.8}$$

Following the above arguments, it is easy to see that for $j = 0$

$$\nabla_0 \ell(\boldsymbol{\beta}^\infty)\delta_0 \geq 0. \tag{A.9}$$

Hence for $\boldsymbol{\delta} = (\delta_0, ..., \delta_p) \in \mathbb{R}^{p+1}$, by the differentiability of $\ell(\boldsymbol{\beta})$, we have

$$\begin{aligned}
&\liminf_{\alpha \downarrow 0+} \{\frac{Q(\boldsymbol{\beta}^\infty + \alpha\boldsymbol{\delta}) - Q(\boldsymbol{\beta}^\infty)}{\alpha}\} = \nabla_0 \ell(\boldsymbol{\beta}^\infty)\delta_0 \\
&+ \sum_{j=1}^{p} [\nabla_j \ell(\boldsymbol{\beta}^\infty)\delta_j + \liminf_{\alpha \downarrow 0+} \{\frac{\rho(|\beta_j^\infty + \alpha\delta_j|) - \rho(|\beta_j^\infty|)}{\alpha}\}] \\
&= \nabla_0 \ell(\boldsymbol{\beta}^\infty)\delta_0 + \sum_{j=1}^{p} \liminf_{\alpha \downarrow 0+} \{\frac{Q(\boldsymbol{\beta}^\infty + \alpha\boldsymbol{\delta}_j) - Q(\boldsymbol{\beta}^\infty)}{\alpha}\} \geq 0,
\end{aligned} \tag{A.10}$$

by (A.8, A.9). This completes the proof.

A.1 Proof of Theorem 2.1

Proof. For the sake of notational convenience, let $\chi^j_{\beta_0,...,\beta_{j-1},\beta_{j+1},...,\beta_p} \equiv \chi(u)$ for $Q(\boldsymbol{\beta})$ as a function of the jth coordinate with $(\beta_0, ..., \beta_{j-1}, \beta_{j+1}, ..., \beta_p)$ being fixed. We first deal with the $j \in \{1, ..., p\}$ coordinates, then the intercept (0th coordinate) in the following arguments.

For $j \in \{1, ..., p\}$th coordinate, observe that

$$\begin{aligned}
\chi(u + \delta) - \chi(u) &= \ell(\beta_0, ..., \beta_{j-1}, u + \delta, \beta_{j+1}, ..., \beta_p) \\
&- \ell(\beta_0, ..., \beta_{j-1}, u, \beta_{j+1}, ..., \beta_p) + \rho(|u + \delta|) - \rho(|u|) \\
&= \nabla_j \ell(\beta_0, ..., \beta_{j-1}, u, \beta_{j+1}, ..., \beta_p)\delta \\
&+ \frac{1}{2}\nabla_j^2 \ell(\beta_0, ..., \beta_{j-1}, u, \beta_{j+1}, ..., \beta_p)\delta^2 + o(\delta^2) \\
&+ \rho'(|u|)(|u + \delta| - |u|) + \frac{1}{2}\rho''(|u^*|)(|u + \delta| - |u|)^2,
\end{aligned} \tag{A.11}$$

with $|u^*|$ being some number between $|u+\delta|$ and $|u|$. Notation $\nabla_j \ell(\beta_0, ..., \beta_{j-1}, u, \beta_{j+1}, ..., \beta_p)$ and $\nabla_j^2 \ell(\beta_0, ..., \beta_{j-1}, u, \beta_{j+1}, ..., \beta_p)$ denote the first and second derivative of the function ℓ w.r.t. the jth coordinate (assuming to be existed by condition (1)).

We re-write the RHS of (A.11) as follows:

$$\begin{aligned} RHS(\text{of A.11}) &= \nabla_j \ell(\beta_0, ..., \beta_{j-1}, u, \beta_{j+1}, ..., \beta_p)\delta \\ &+ (\nabla_j^2 \ell(\beta_0, ..., \beta_{j-1}, u, \beta_{j+1}, ..., \beta_p) - M)\delta^2 + \rho'(|u|)\text{sgn}(u)\delta \\ &+ \rho'(|u|)(|u+\delta| - |u|) - \rho'(|u|)\text{sgn}(u)\delta + \frac{1}{2}\rho''(|u^*|)(|u+\delta| - |u|)^2 \\ &+ (M - \frac{1}{2}\nabla_j^2 \ell(\beta_0, ..., \beta_{j-1}, u, \beta_{j+1}, ..., \beta_p))\delta^2 + o(\delta^2). \end{aligned} \tag{A.12}$$

On the other hand, the solution of the jth coordinate ($j \in \{1, ..., p\}$) is to minimize the following function,

$$Q_j(u|\tilde{\boldsymbol{\beta}}) = \ell(\tilde{\boldsymbol{\beta}}) + \nabla_j \ell(\tilde{\boldsymbol{\beta}})(u - \tilde{\beta}_j) + \frac{1}{2}\nabla_j^2 \ell(\tilde{\boldsymbol{\beta}})(u - \tilde{\beta}_j)^2 + \rho(|u|), \tag{A.13}$$

By majorization, we bound $\nabla_j^2 \ell(\tilde{\boldsymbol{\beta}})$ by a constant M for standardized variables. So the actual function being minimized is

$$\tilde{Q}_j(u|\tilde{\boldsymbol{\beta}}) = \ell(\tilde{\boldsymbol{\beta}}) + \nabla_j \ell(\tilde{\boldsymbol{\beta}})(u - \tilde{\beta}_j) + \frac{1}{2}M(u - \tilde{\beta}_j)^2 + \rho(|u|). \tag{A.14}$$

Since u is to minimize (A.14), we have, for the jth ($j \in \{1, ..., p\}$) coordinate ,

$$\nabla_j \ell(\tilde{\boldsymbol{\beta}}) + M(u - \tilde{\beta}_j) + \rho'(|u|)\text{sgn}(u) = 0, \tag{A.15}$$

Because $\chi(u)$ is minimized at u_0, by (A.15), we have

$$\begin{aligned} 0 &= \nabla_j \ell(\beta_0, ..., \beta_{j-1}, u_0 + \delta, \beta_{j+1}, ..., \beta_p) + M(u_0 - u_0 - \delta) + \rho'(|u_0|)\text{sgn}(u_0) \\ &= \nabla_j \ell(\beta_0, ..., \beta_{j-1}, u_0, \beta_{j+1}, ..., \beta_p) + \nabla_j^2 \ell(\beta_0, ..., \beta_{j-1}, u_0, \beta_{j+1}, ..., \beta_p)\delta + o(\delta) \\ &- M\delta + \rho'(|u_0|)\text{sgn}(u_0), \end{aligned} \tag{A.16}$$

if $u_0 = 0$ then the above holds true for some value of $\text{sgn}(u_0) \in (-1, 1)$.

Observe that $\rho'(|x|) \geq 0$, then

$$\rho'(|u|)(|u+\delta| - |u|) - \rho'(|u|)\text{sgn}(u)\delta = \rho'(|u|)[(|u+\delta| - |u|) - \text{sgn}(u)\delta] \geq 0 \quad \text{(A.17)}$$

Therefore using (A.16, A.17) in (A.12) at u_0, we have, for $j \in \{1, ..., p\}$,

$$\begin{aligned} \chi(u_0+\delta) - \chi(u_0) &\geq \frac{1}{2}\rho''(|u^*|)(|u+\delta| - |u|)^2 \\ &+ \delta^2(M - \frac{1}{2}\nabla_j^2 \ell(\beta_0, ..., \beta_{j-1}, u_0, \beta_{j+1}, ..., \beta_p)) + o(\delta^2) \\ &\geq \frac{1}{2}M\delta^2 + \frac{1}{2}\rho''(|u^*|)(|u+\delta| - |u|)^2 + o(\delta^2). \end{aligned} \quad \text{(A.18)}$$

By condition (ii) of the MMCD algorithm $\inf_t \rho''(|t|; \lambda, \gamma) > -M$ and $(|u+\delta| - |u|)^2 \leq \delta^2$. Hence there exist $\theta_2 = \frac{1}{2}(M + \inf_x \rho''(|x|) + o(1)) > 0$, such that for the jth coordinate, $j \in \{1, ..., p\}$,

$$\chi(u_0+\delta) - \chi(u_0) \geq \theta_2 \delta^2. \quad \text{(A.19)}$$

Now consider β_0, observe that

$$\begin{aligned} \chi(u+\delta) - \chi(u) &= \ell(u+\delta, \beta_1, ..., \beta_p) - \ell(u, \beta_1, ..., \beta_p) \\ &= \nabla_0 \ell(u, \beta_1, ..., \beta_p)\delta + \frac{1}{2}\nabla_0^2 \ell(u, \beta_1, ..., \beta_p)\delta^2 + o(\delta^2) \\ &= \nabla_0 \ell(u, \beta_1, ..., \beta_p)\delta + (\nabla_0^2 \ell(u, \beta_1, ..., \beta_p) - M)\delta^2 \\ &+ (M - \frac{1}{2}\nabla_0^2 \ell(u, \beta_1, ..., \beta_p))\delta^2 + o(\delta^2), \end{aligned} \quad \text{(A.20)}$$

By similar arguments to (A.16), we have

$$\begin{aligned} 0 &= \nabla_0 \ell(u_0+\delta, \beta_1, ..., \beta_p) + M(u_0 + \delta - u_0) \\ &= \nabla_0 \ell(u_0, \beta_1, ..., \beta_p) + \nabla_0^2 \ell(u_0, \beta_1, ..., \beta_p)\delta + o(\delta) - M\delta. \end{aligned} \quad \text{(A.21)}$$

Therefore, by (A.20, A.21), for the first coordinate of $\boldsymbol{\beta}$

$$\begin{aligned}\chi(u_0+\delta)-\chi(u_0) &= (M-\frac{1}{2}\nabla_0^2\ell(u_0,\beta_1,...,\beta_p))\delta^2+o(\delta^2)\\ &= \frac{1}{2}M\delta^2+\frac{1}{2}(M-\nabla_0^2\ell(u_0,\beta_1,...,\beta_p))\delta^2+o(\delta^2)\\ &\geq \frac{1}{2}\delta^2(M+o(1)). \end{aligned} \tag{A.22}$$

Hence there exists a $\theta_1=\frac{1}{2}(M+o(1))>0$, such that for the first coordinate of $\boldsymbol{\beta}$

$$\chi(u_0+\delta)-\chi(u_0)\geq\theta_1\delta^2. \tag{A.23}$$

Let $\theta=\min(\theta_1,\theta_2)$, using (A.19,A.23), we have for all the coordinates of $\boldsymbol{\beta}$,

$$\chi(u_0+\delta)-\chi(u_0)\geq\theta\delta^2, \tag{A.24}$$

By (A.24) we have

$$Q(\boldsymbol{\beta}_{j-1}^{m-1})-Q(\boldsymbol{\beta}_j^{m-1})\geq\theta(\beta_j^m-\beta_j^{m-1})^2=\theta\parallel\boldsymbol{\beta}_{j-1}^{m-1}-\boldsymbol{\beta}_j^{m-1}\parallel_2^2, \tag{A.25}$$

where $\boldsymbol{\beta}_j^{m-1}=(\beta_1^m,...,\beta_j^m,\beta_{j+1}^{m-1},...,\beta_p^{m-1})$. The (A.25) establishes the boundedness of the sequence $\{\boldsymbol{\beta}^m\}$ for every $m>1$ since the starting point of $\{\boldsymbol{\beta}^1\}\in\mathbb{R}^{p+1}$.

Apply (A.25) over all the coordinates, we have for all m

$$Q(\boldsymbol{\beta}^m)-Q(\boldsymbol{\beta}^{m+1})\geq\theta\parallel\boldsymbol{\beta}^{m+1}-\boldsymbol{\beta}^m\parallel_2^2. \tag{A.26}$$

Since the (decreasing) sequence $Q(\boldsymbol{\beta}^m)$ converges, (A.26) shows that the sequence $\{\boldsymbol{\beta}^m\}$ have a unique limit point. This completes the proof of the convergence of $\{\boldsymbol{\beta}^m\}$.

The condition (3) and (4) of Lemma A.1 holds by (A.26). Hence, the limit point of $\{\boldsymbol{\beta}^m\}$ is a minimum of $Q(\boldsymbol{\beta})$ by Lemma A.1. This completes the proof of the Theorem 2.1.

APPENDIX B
PROOF OF THEOREM IN CHAPTER 3

Theorem 3.1 is dealing with the ℓ_2 norms of coefficients rather than each individual coefficient in theorem 2.1. Lemma B.1 is needed in the proof of theorem 3.1.

Lemma B.1. *Suppose the data* $(\boldsymbol{y}, \boldsymbol{X})$ *lies on a compact set and the condition (1) - (4) in Lemma A.1 hold.*

Then if $\boldsymbol{\beta}^{\infty}$ *is any limit point of the sequence* $\{\boldsymbol{\beta}^k\}$*, then* $\boldsymbol{\beta}^{\infty}$ *is a minimum for the function* $Q^2(\boldsymbol{\beta})$*, i.e.*

$$\liminf_{\alpha \downarrow 0+} \{ \frac{Q^2(\boldsymbol{\beta}^{\infty} + \alpha \boldsymbol{\delta}) - Q^2(\boldsymbol{\beta}^{\infty})}{\alpha} \} \geq 0, \tag{B.1}$$

for any $\boldsymbol{\delta} = (\delta_0, ..., \delta_p) \in \mathbb{R}^{p+1}$.

Proof. For any $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J)^T$ and $\boldsymbol{\delta}_j = (0, ..., \boldsymbol{\delta}_j, ..., 0) \in \mathbb{R}^{p+1}$, we have

$$\begin{aligned} \liminf_{\alpha \downarrow 0+} \{ \frac{Q^2(\boldsymbol{\beta} + \alpha \boldsymbol{\delta}_j) - Q^2(\boldsymbol{\beta})}{\alpha} \} &= \boldsymbol{\delta}_j^T \nabla_j \ell(\boldsymbol{\beta}) \\ &+ \liminf_{\alpha \downarrow 0+} \{ \frac{\rho(\|\boldsymbol{\beta}_j + \alpha \boldsymbol{\delta}_j\|_2) - \rho(\|\boldsymbol{\beta}_j\|_2)}{\alpha} \} \\ &= \boldsymbol{\delta}_j^T \nabla_j \ell(\boldsymbol{\beta}) + \partial \rho(\|\boldsymbol{\beta}_j\|_2; \boldsymbol{\delta}_j), \end{aligned} \tag{B.2}$$

with

$$\partial \rho(\|\boldsymbol{\beta}_j\|_2; \boldsymbol{\delta}_j) = \begin{cases} \rho^{'}(\|\boldsymbol{\beta}_j\|_2) \frac{\boldsymbol{\delta}_j^T \boldsymbol{\beta}_j}{\|\boldsymbol{\beta}_j\|_2}, & \boldsymbol{\beta}_j \neq \mathbf{0}; \\ \rho^{'}(\mathbf{0}) \|\boldsymbol{\delta}_j\|_2, & \boldsymbol{\beta}_j = \mathbf{0}. \end{cases} \tag{B.3}$$

Assume $\boldsymbol{\beta}^{n_k} \to \boldsymbol{\beta}^{\infty} = (\beta_0^{\infty}, ..., \beta_p^{\infty})$, and by condition (4), as $m \to \infty$

$$\begin{aligned} B_j^{n_m} &= (\beta_0^{n_m+1}, (\boldsymbol{\beta}_1^{n_m+1})^T, ..., (\boldsymbol{\beta}_j^{n_m+1})^T, (\boldsymbol{\beta}_{j+1}^{n_m})^T, ..., (\boldsymbol{\beta}_J^{n_m})^T)^T \\ &\to (\beta_0^{\infty}, (\boldsymbol{\beta}_1^{\infty})^T, ..., (\boldsymbol{\beta}_j^{\infty})^T, ..., (\boldsymbol{\beta}_J^{\infty})^T)^T \end{aligned} \tag{B.4}$$

By (B.3) and (B.4), since $\boldsymbol{\delta}_j^T\boldsymbol{\beta}_j \leq \|\boldsymbol{\delta}_j\|_2\|\boldsymbol{\beta}_j\|_2$ we have the results below for $j \in \{1, ..., J\}$.

$$\partial\rho(\|\boldsymbol{\beta}_j^{n_m}\|_2; \boldsymbol{\delta}_j) \to \partial\rho(\|\boldsymbol{\beta}_j^{\infty}\|_2; \boldsymbol{\delta}_j), \text{ if } \boldsymbol{\beta}_j^{\infty} \neq \mathbf{0};$$

$$\rho(\|\boldsymbol{\beta}_j^{\infty}\|_2; \boldsymbol{\delta}_j) \geq \liminf_m \partial\rho(\|\boldsymbol{\beta}_j^{n_m}\|_2; \boldsymbol{\delta}_j), \text{ if } \boldsymbol{\beta}_j^{\infty} = \mathbf{0}. \tag{B.5}$$

By the group coordinate-wise minimum of jth coordinate, we have

$$\boldsymbol{\delta}_j^T\nabla_j\ell(B_j^{n_m-1}) + \partial\rho(\|\boldsymbol{\beta}_j^{n_m}\|_2; \boldsymbol{\delta}_j) \geq 0. \tag{B.6}$$

Thus (B.5, B.6) implies that

$$\boldsymbol{\delta}_j^T\nabla_j\ell(\boldsymbol{\beta}^{\infty}) + \partial\rho(\|\boldsymbol{\beta}_j^{\infty}\|_2; \boldsymbol{\delta}_j) \quad \geq$$

$$\liminf_m\{\boldsymbol{\delta}_j^T\nabla_j\ell(B_j^{n_m-1}) + \partial\rho(\|\boldsymbol{\beta}_j^{n_m}\|_2; \boldsymbol{\delta}_j)\} \geq 0. \tag{B.7}$$

By (B.2,B.7), for $j \in \{1, ..., J\}$, we have

$$\liminf_{\alpha\downarrow 0+}\{\frac{Q^2(\boldsymbol{\beta}^{\infty} + \alpha\boldsymbol{\delta}_j) - Q^2(\boldsymbol{\beta}^{\infty})}{\alpha}\} \geq 0. \tag{B.8}$$

By a similar argument, it is easy to show that $\nabla_0\ell(\boldsymbol{\beta}^{\infty})\delta_0 \geq 0$. Then by the differentiability of $\ell(\boldsymbol{\beta})$, we have

$$\begin{aligned}
&\liminf_{\alpha\downarrow 0+}\{\frac{Q^2(\boldsymbol{\beta}^{\infty} + \alpha\boldsymbol{\delta}) - Q^2(\boldsymbol{\beta}^{\infty})}{\alpha}\} = \nabla_0\ell(\boldsymbol{\beta}^{\infty})\delta_0 \\
&+ \sum_{j=1}^{J}\{\boldsymbol{\delta}_j^T\nabla_j\ell(\boldsymbol{\beta}^{\infty}) + \liminf_{\alpha\downarrow 0+}\{\frac{\rho(\|\boldsymbol{\beta}_j^{\infty} + \alpha\boldsymbol{\delta}_j\|_2) - \rho(\|\boldsymbol{\beta}_j^{\infty}\|_2)}{\alpha}\}\} \\
&= \nabla_0\ell(\boldsymbol{\beta}^{\infty})\delta_0 + \sum_{j=1}^{J}\liminf_{\alpha\downarrow 0+}\{\frac{Q^2(\boldsymbol{\beta}^{\infty} + \alpha\boldsymbol{\delta}_j) - Q^2(\boldsymbol{\beta}^{\infty})}{\alpha}\} \geq 0,
\end{aligned} \tag{B.9}$$

by (B.8). This completes the result for $Q^2(\boldsymbol{\beta})$.

B.1 Proof of Theorem 3.1

Proof. Write $\chi^{j}_{\beta_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J} \equiv \chi(\boldsymbol{u})$ for $Q^2(\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}_j$ with the rest group coordinates $(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J)$ being fixed. We first deal with the penalized coordinates, then the intercept (0th coordinate) in the following arguments.

Let $\nabla_j \ell(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \boldsymbol{u}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J)$ and $\nabla_j^2 \ell$ denote the first and second derivative of the function ℓ w.r.t. $\boldsymbol{\beta}_j$ (assuming to be existed by condition (1)), for the jth group coordinate, observe

$$\begin{aligned}
\chi(\boldsymbol{u}+\boldsymbol{\delta}) - \chi(\boldsymbol{u}) &= \ell(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \boldsymbol{u}+\boldsymbol{\delta}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J) \\
&\quad - \ell(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \boldsymbol{u}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J) + \rho(\|\boldsymbol{u}+\boldsymbol{\delta}\|_2) - \rho(\|\boldsymbol{u}\|_2) \\
&= \boldsymbol{\delta}^T \nabla_j \ell(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \boldsymbol{u}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J) + \frac{1}{2}\boldsymbol{\delta}^T \nabla_j^2 \ell \boldsymbol{\delta} \\
&\quad + \rho'(\|\boldsymbol{u}\|_2)(\|\boldsymbol{u}+\boldsymbol{\delta}\|_2 - \|\boldsymbol{u}\|_2) + \frac{1}{2}\rho''(\|\boldsymbol{u}^*\|_2)(\|\boldsymbol{u}+\boldsymbol{\delta}\|_2 - \|\boldsymbol{u}\|_2)^2,
\end{aligned} \tag{B.10}$$

with $\|\boldsymbol{u}^*\|_2$ being a number between $\|\boldsymbol{u}+\boldsymbol{\delta}\|_2$ and $\|\boldsymbol{u}\|_2$. For group-wise standardized predictors, we have $\nabla_j^2 \ell = \boldsymbol{I}_{d_j \times d_j}$ for $j = 1, .., J$.

We re-write the RHS of (B.10) as follows:

$$\begin{aligned}
RHS(\text{of B.10}) &= \boldsymbol{\delta}^T \nabla_j \ell(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \boldsymbol{u}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J) + \rho'(\|\boldsymbol{u}\|_2)\frac{\boldsymbol{\delta}^T \boldsymbol{u}}{\|\boldsymbol{u}\|_2} \\
&+ \rho'(\|\boldsymbol{u}\|_2)(\|\boldsymbol{u}+\boldsymbol{\delta}\|_2 - \|\boldsymbol{u}\|_2 - \frac{\boldsymbol{\delta}^T \boldsymbol{u}}{\|\boldsymbol{u}\|_2}) \\
&+ \frac{1}{2}\{\boldsymbol{\delta}^T \nabla_j^2 \ell \boldsymbol{\delta} + \rho''(\|\boldsymbol{u}^*\|_2)(\|\boldsymbol{u}+\boldsymbol{\delta}\|_2 - \|\boldsymbol{u}\|_2)^2\}
\end{aligned} \tag{B.11}$$

On the other hand, since $\boldsymbol{u}$ is to minimize $Q^2(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}_j$, we have,

$$\nabla_j \ell(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \boldsymbol{u}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J) + \rho'(\|\boldsymbol{u}\|_2)\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} = 0, \ \forall \boldsymbol{\beta}_j \neq 0$$

$$\|\nabla_j \ell(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \boldsymbol{u}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J)\|_2 \leq \sqrt{d_j}\lambda, \ \forall \boldsymbol{\beta}_j = 0 \qquad \text{(B.12)}$$

Because $\chi(\boldsymbol{u})$ is minimized at $\boldsymbol{u}_0$, by (B.12), for $\boldsymbol{u}_0 \neq 0$ we have

$$\begin{aligned}
\chi(\boldsymbol{u}_0 + \boldsymbol{\delta}) - \chi(\boldsymbol{u}_0) &= \rho'(\|\boldsymbol{u}\|_2)(\|\boldsymbol{u} + \boldsymbol{\delta}\|_2 - \|\boldsymbol{u}\|_2 - \frac{\boldsymbol{\delta}^T \boldsymbol{u}}{\|\boldsymbol{u}\|_2}) \\
&+ \frac{1}{2}\{\boldsymbol{\delta}^T \nabla_j^2 \ell \boldsymbol{\delta} + \rho''(\|\boldsymbol{u}^*\|_2)(\|\boldsymbol{u} + \boldsymbol{\delta}\|_2 - \|\boldsymbol{u}\|_2)^2\} \\
&\geq \frac{1}{2}\boldsymbol{\delta}^T\{\nabla_j^2 \ell + \text{diag}(\rho''(\|\boldsymbol{u}^*\|_2))\}\boldsymbol{\delta} \\
&\geq \theta_2 \|\boldsymbol{\delta}\|_2^2, \qquad \text{(B.13)}
\end{aligned}$$

since $\|\boldsymbol{\delta} + \boldsymbol{u}\|_2 \|\boldsymbol{u}\|_2 \geq \boldsymbol{u}^T(\boldsymbol{u} + \boldsymbol{\delta})$, $\rho''(|t|) < 0$ for SCAD and MCP, and $\|\boldsymbol{\delta} + \boldsymbol{u}\|_2 \leq \|\boldsymbol{\delta}\|_2 + \|\boldsymbol{u}\|_2$ with $\theta_2 = (1 + \inf_t \rho''(|t|))/2 > 0$.

For $\boldsymbol{u}_0 = 0$, by (B.10) we have

$$\begin{aligned}
\chi(\boldsymbol{\delta}) - \chi(\mathbf{0}) &= \boldsymbol{\delta}^T \nabla_j \ell(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \mathbf{0}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J) + \frac{1}{2}\boldsymbol{\delta}^T \nabla_j^2 \ell \boldsymbol{\delta} \\
&+ \rho'(0)\|\boldsymbol{\delta}\|_2 + \frac{1}{2}\rho''(0)\|\boldsymbol{\delta}\|_2^2 \\
&\geq \theta_2 \|\boldsymbol{\delta}\|_2^2, \qquad \text{(B.14)}
\end{aligned}$$

since $\boldsymbol{\delta}^T \nabla_j \ell(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \mathbf{0}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J) \leq \|\boldsymbol{\delta}\|_2 \|\nabla_j \ell(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{j-1}, \mathbf{0}, \boldsymbol{\beta}_{j+1}, ..., \boldsymbol{\beta}_J)\|_2$, by (B.12) this term is less than $\|\boldsymbol{\delta}\|_2 \sqrt{d_j}\gamma = \|\boldsymbol{\delta}\|_2 \rho'(0)$. Therefor, by (B.13) and (B.14), we have the following results hold for any $\boldsymbol{u}$ for $j = 1, ..., J$.

$$\chi(\boldsymbol{u}_0 + \boldsymbol{\delta}) - \chi(\boldsymbol{u}_0) \geq \theta_2 \|\boldsymbol{\delta}\|_2^2 \qquad \text{(B.15)}$$

By similar arguments, we have a resembling result for the intercept, i.e.

$$\chi(u_0 + \delta) - \chi(u_0) \geq \theta_1 \delta^2, \qquad \text{(B.16)}$$

with $\theta_1 = \frac{1}{2}\nabla_0^2\ell > 0$.

Let $\theta = \min(\theta_1, \theta_2)$, using (B.15,B.16), we have

$$\begin{aligned} Q^2(B_{j-1}^m) - Q^2(B_j^m) &\geq \theta(\boldsymbol{\beta}_j^m - \boldsymbol{\beta}_j^{m+1})^2 \\ &= \theta \parallel B_{j-1}^m - B_j^m \parallel_2^2, \end{aligned} \tag{B.17}$$

where $B_j^m = (\beta_0^{m+1}, ..., (\boldsymbol{\beta}_{j-1}^{m+1})^T, (\boldsymbol{\beta}_j^{m+1})^T, ..., (\boldsymbol{\beta}_J^m)^T)^T$ and $\boldsymbol{\beta}_j^m = (\beta_{j1}^m, ..., \beta_{jd_j}^m)^T$. The (B.17) establishes the boundedness of the sequence $\{\boldsymbol{\beta}^m\}$ with $\boldsymbol{\beta}^m = (\beta_0^m, \boldsymbol{\beta}_1^m, ..., \boldsymbol{\beta}_J^m)^T$ for every $m > 1$ since the starting point of $\{\boldsymbol{\beta}^1\} \in \mathbb{R}^{p+1}$. Apply (B.17) over all the coordinates, we then establish the unique limit point for the ℓ_2 grouped concave penalty,

$$Q^2(\boldsymbol{\beta}^m) - Q^2(\boldsymbol{\beta}^{m+1}) \geq \theta \parallel \boldsymbol{\beta}^{m+1} - \boldsymbol{\beta}^m \parallel_2^2, \text{for all } m. \tag{B.18}$$

due to the fact that the (decreasing) sequence $Q^2(\boldsymbol{\beta}^m)$ converges. This completes the proof of the convergence of $\{\boldsymbol{\beta}^m\}$ for the ℓ_2 grouped concave penalty.

The assumption (3) and (4) of Lemma B.1 holds by (B.18). Hence, the limit point of $\{\boldsymbol{\beta}^m\}$ is a minimum of $Q^2(\boldsymbol{\beta})$ by Lemma B.1. This completes the proof of the convergence of the ℓ_2 grouped concave penalty.

B.2 Outline of Proof of Theorem 3.2

Proof. The proof of theorem 3.2 follows the same logic as theorem 2.1. Below we outline the main difference. In the case of GLM, we have a similar version as (B.10)

$$\begin{aligned}\chi(\boldsymbol{u}+\boldsymbol{\delta})-\chi(\boldsymbol{u}) &= \ell(\beta_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{u}+\boldsymbol{\delta},\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J)\\ &\quad -\ell(\beta_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{u},\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J)+\rho(\|\boldsymbol{u}+\boldsymbol{\delta}\|_2)-\rho(\|\boldsymbol{u}\|_2)\\ &= \boldsymbol{\delta}^T\nabla_j\ell(\beta_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{u},\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J)\\ &\quad +\frac{1}{2}\boldsymbol{\delta}^T\nabla_j^2\ell(\beta_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{u},\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J)\boldsymbol{\delta}+o(\|\boldsymbol{\delta}\|_2^2)\\ &\quad +\rho'(\|\boldsymbol{u}\|_2)(\|\boldsymbol{u}+\boldsymbol{\delta}\|_2-\|\boldsymbol{u}\|_2)+\frac{1}{2}\rho''(\|\boldsymbol{u}^*\|_2)(\|\boldsymbol{u}+\boldsymbol{\delta}\|_2-\|\boldsymbol{u}\|_2)^2,\end{aligned} \tag{B.19}$$

Because $\chi(\boldsymbol{u})$ is minimized at $\boldsymbol{u}_0$ for the surrogate function by the minimization majorization approach, we then have

$$\begin{aligned}\mathbf{0} &= \nabla_j\ell(\beta_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{u}_0+\boldsymbol{\delta},\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J)+M(\boldsymbol{u}_0-\boldsymbol{u}_0-\boldsymbol{\delta})+\rho'(\|\boldsymbol{u}_0\|_2)\frac{\boldsymbol{u}_0}{\|\boldsymbol{u}_0\|_2}\\ &= \nabla_j\ell(\beta_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{u}_0,\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J)+\nabla_j^2\ell(\beta_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{u}_0,\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J)\boldsymbol{\delta}\\ &- M\boldsymbol{\delta}+\rho'(\|\boldsymbol{u}_0\|_2)\frac{\boldsymbol{u}_0}{\|\boldsymbol{u}_0\|_2}+o(\|\boldsymbol{\delta}\|_2^2),\end{aligned} \tag{B.20}$$

for $\boldsymbol{u}_0 \neq \mathbf{0}$. For $\boldsymbol{u}_0 = \mathbf{0}$, we have $\|\nabla_j\ell(\beta_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{u}_0+\boldsymbol{\delta},\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J)\|_2 \leq \sqrt{d_j}\lambda$.

Follow similar argument as in the linear models, then we still have

$$\chi(\boldsymbol{u}_0+\boldsymbol{\delta})-\chi(\boldsymbol{u}_0) \quad \geq \quad \theta_2\|\boldsymbol{\delta}\|_2^2, \tag{B.21}$$

with $\theta_2 = (M+\inf_t\rho''(|t|)+o(1))/2 > 0$ for $j = 1,...,J$. The rest arguments follows as the proof for the ℓ_2 grouped concave penalty in linear models.

APPENDIX C
PROOF OF THEOREM IN CHAPTER 4

The proof of Theorem 4.1 is similar to Theorem 3.1. In this case, we are dealing with the ℓ_1 norm of coefficients. Lemma C.1 is needed in the proof of Theorem 4.1 and Theorem 4.2.

Lemma C.1. *Suppose the data* $(\boldsymbol{y}, \boldsymbol{X})$ *lies on a compact set and the conditions (1) - (4) in Lemma A.1 hold.*

Then if $\boldsymbol{\beta}^\infty$ *is any limit point of the sequence* $\{\boldsymbol{\beta}^k\}$*, then* $\boldsymbol{\beta}^\infty$ *is a minimum for the function* $Q^1(\boldsymbol{\beta})$*i.e.*

$$\liminf_{\alpha \downarrow 0+} \{ \frac{Q^1(\boldsymbol{\beta}^\infty + \alpha \boldsymbol{\delta}) - Q^1(\boldsymbol{\beta}^\infty)}{\alpha} \} \geq 0, \tag{C.1}$$

for any $\boldsymbol{\delta} = (\delta_0, ..., \delta_p) \in \mathbb{R}^{p+1}$.

Proof. For any $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J)^T$ and $\boldsymbol{\delta}_{j_k} = (0, ..., \delta_{j_k}, ..., 0) \in \mathbb{R}^{p+1}$, we have

$$\begin{aligned} \liminf_{\alpha \downarrow 0+} \{ \frac{Q^1(\boldsymbol{\beta} + \alpha \boldsymbol{\delta}_{j_k}) - Q^1(\boldsymbol{\beta})}{\alpha} \} &= \nabla_{j_k} \ell(\boldsymbol{\beta}) \delta_{j_k} \\ + \liminf_{\alpha \downarrow 0+} \{ \frac{\rho(\|\beta_{j_1}, ..., \beta_{j_k} + \alpha \delta_{j_k}, ..., \beta_{j_{d_j}}\|_1) - \rho(\|\beta_{j_1}, ..., \beta_{j_k}, ..., \beta_{j_{d_j}}\|_1)}{\alpha} \} & \\ &= \nabla_{j_k} \ell(\boldsymbol{\beta}) \delta_{j_k} + \partial \rho(\|\beta_{j_1}, ..., \beta_{j_k}, ..., \beta_{j_{d_j}}\|_1; \delta_{j_k}), \end{aligned} \tag{C.2}$$

with

$$\partial \rho(\|\beta_{j_1}, ..., \beta_{j_k}, ..., \beta_{j_{d_j}}\|_1; \delta_{j_k}) = \begin{cases} \rho^{'}(\|\beta_{j_1}, ..., \beta_{j_k}, ..., \beta_{j_{d_j}}\|_1) \mathrm{sgn}(\beta_{j_k}) \delta_{j_k}, & |\beta_{j_k}| > 0; \\ \rho^{'}(\|\beta_{j_1}, ..., 0, ..., \beta_{j_{d_j}}\|_1) |\delta_{j_k}|, & |\beta_{j_k}| = 0, \end{cases} \tag{C.3}$$

where

$$\text{sgn}(x) = \begin{cases} 1, & \text{if } x > 0; \\ -1, & \text{if } x < 0; \\ \text{any } u \in (-1, 1), & \text{if } x = 0. \end{cases}$$

Assume $\boldsymbol{\beta}^{n_k} \to \boldsymbol{\beta}^{\infty} = (\beta_0^{\infty}, ..., \beta_p^{\infty})$, and by assumption 4, as $m \to \infty$

$$B_{j_k}^{n_m} = (\beta_0^{n_m+1}, ..., (\boldsymbol{\beta}_{j-1}^{n_m+1})^T, \beta_{j_1}^{n_m+1}, ..., \beta_{j_k}^{n_{m+1}}, \beta_{j_{k+1}}^{n_m}, ..., \beta_{j_{d_j}}^{n_m}, (\boldsymbol{\beta}_{j+1}^{n_m})^T, ..., (\boldsymbol{\beta}_J^{n_m})^T)^T$$

$$\to (\beta_0^{\infty}, ..., (\boldsymbol{\beta}_{j-1}^{\infty})^T, \beta_{j_1}^{\infty}, ..., \beta_{j_k}^{\infty}, \beta_{j_{k+1}}^{\infty}, ..., \beta_{j_{d_j}}^{\infty}, (\boldsymbol{\beta}_{j+1}^{\infty})^T, ..., (\boldsymbol{\beta}_J^{\infty})^T)^T \quad \text{(C.4)}$$

By (C.3) and (C.4), we have the results below for $k \in \{1, ..., d_j\}$, $j \in \{1, ..., J\}$.

$$\partial\rho(\|\beta_{j_1}^{n_m}, ..., \beta_{j_k}^{n_m}, ..., \beta_{j_{d_j}}^{n_m}\|_1; \delta_{j_k}) \to \partial\rho(\|\beta_{j_1}^{\infty}, ..., \beta_{j_k}^{\infty}, ..., \beta_{j_{d_j}}^{\infty}\|_1; \delta_{j_k}), \text{ if } \beta_{j_k}^{\infty} \neq 0;$$

$$\rho(\|\beta_{j_1}^{\infty}, ..., \beta_{j_k}^{\infty}, ..., \beta_{j_{d_j}}^{\infty}\|_1; \delta_{j_k}) \geq \liminf_m \partial\rho(\|\beta_{j_1}^{n_m}, ..., \beta_{j_k}^{n_m}, ..., \beta_{j_{d_j}}^{n_m}\|_1; \delta_{j_k}), \text{ if } \beta_{j_k}^{\infty} = 0. \quad \text{(C.5)}$$

By the coordinate-wise minimum of kth coordinate in jth group, we have

$$\nabla_{j_k}\ell(B_{j_k}^{n_m-1})\delta_{j_k} + \partial\rho(\|\beta_{j_1}^{n_m}, ..., \beta_{j_k}^{n_m}, \beta_{j_{k+1}}^{n_m-1} ..., \beta_{j_{d_j}}^{n_m-1}\|_1; \delta_{j_k}) \geq 0. \quad \text{(C.6)}$$

Thus (C.5, C.6) implies that

$$\nabla_{j_k}\ell(\boldsymbol{\beta}^{\infty})\delta_{j_k} + \partial\rho(\|\beta_{j_1}^{\infty}, ..., \beta_{j_k}^{\infty}, ..., \beta_{j_{d_j}}^{\infty}\|_1; \delta_{j_k}) \geq$$

$$\liminf_m \{\nabla_{j_k}\ell(B_{j_k}^{n_m-1})\delta_{j_k} + \partial\rho(\|\beta_{j_1}^{n_m}, ..., \beta_{j_k}^{n_m}, \beta_{j_k}^{n_m-1}, ..., \beta_{j_{d_j}}^{n_m-1}\|_1; \delta_{j_k})\} \geq 0. \quad \text{(C.7)}$$

By (C.2,C.7), for $k \in \{1, ..., d_j\}$ and $j \in \{1, ..., J\}$, we have

$$\liminf_{\alpha \downarrow 0+} \{\frac{Q^1(\boldsymbol{\beta}^{\infty} + \alpha\boldsymbol{\delta}_{j_k}) - Q^1(\boldsymbol{\beta}^{\infty})}{\alpha}\} \geq 0. \quad \text{(C.8)}$$

Following the above arguments, it is easy to see that for the intercept

$$\nabla_0\ell(\boldsymbol{\beta}^{\infty})\delta_0 \geq 0. \quad \text{(C.9)}$$

Hence for $\boldsymbol{\delta} = (\delta_0, ..., \delta_p) \in \mathbb{R}^{p+1}$, by the differentiability of $\ell(\boldsymbol{\beta})$, we have

$$\liminf_{\alpha \downarrow 0+} \{ \frac{Q^1(\boldsymbol{\beta}^\infty + \alpha \boldsymbol{\delta}) - Q^1(\boldsymbol{\beta}^\infty)}{\alpha} \} = \nabla_0 \ell(\boldsymbol{\beta}^\infty)\delta_0 + \sum_{j=1}^{J} \sum_{k=1}^{d_j} \{ \nabla_{j_k} \ell(\boldsymbol{\beta}^\infty)\delta_{j_k}$$
$$+ \liminf_{\alpha \downarrow 0+} \{ \frac{\rho(\|\beta_{j_1}^\infty, ..., \beta_{j_k}^\infty + \alpha\delta_{j_k}, ..., \beta_{j_{d_j}}^\infty\|_1) - \rho(\|\beta_{j_1}^\infty, ..., \beta_{j_k}^\infty, ..., \beta_{j_{d_j}}^\infty\|_1)}{\alpha} \} \}$$
$$= \nabla_0 \ell(\boldsymbol{\beta}^\infty)\delta_0 + \sum_{j=1}^{J} \sum_{k=1}^{d_j} \liminf_{\alpha \downarrow 0+} \{ \frac{Q^1(\boldsymbol{\beta}^\infty + \alpha \boldsymbol{\delta}_{j_k}) - Q^1(\boldsymbol{\beta}^\infty)}{\alpha} \} \geq 0, \qquad \text{(C.10)}$$

by (C.8, C.9). This completes the result for $Q^1(\boldsymbol{\beta})$.

C.1 Proof of Theorem 4.1

Proof. Let $\chi^{j_k}_{\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ... \boldsymbol{\beta}_J} \equiv \chi(u)$ for $Q^1(\boldsymbol{\beta})$ as a function of β_{j_k} with $(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ... \boldsymbol{\beta}_J)$ being fixed. We first deal with the penalized coordinates, then the intercept (0th coordinate) in the following arguments.

Denote the first and second derivative of the function ℓ w.r.t. β_{j_k} (assuming to be existed by condition (1)) as $\nabla_{j_k} \ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ... \boldsymbol{\beta}_J)$ and $\nabla^2_{j_k} \ell$. Then for the $k$the coordinate in jth group, observe that

$$\begin{aligned}
\chi(u+\delta) - \chi(u) &= \ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u+\delta, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ... \boldsymbol{\beta}_J) \\
&- \ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ... \boldsymbol{\beta}_J) \\
&+ \rho(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u+\delta, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1) \\
&- \rho(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1) \\
&= \nabla_{j_k} \ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ... \boldsymbol{\beta}_J)\delta + \frac{1}{2}\delta^2 \nabla^2_{j_k} \ell \\
&+ \rho'(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1)(|u+\delta| - |u|) \\
&+ \frac{1}{2}\rho''(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u^*, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1)(|u+\delta| - |u|)^2, \qquad \text{(C.11)}
\end{aligned}$$

with $|u^*|$ being some number between $|u+\delta|$ and $|u|$, and $\rho^{'}(|t|)$ and $\rho^{''}(|t|)$ are the first and second derivative of $\rho(|t|)$ w.r.t. $|t|$. For column-wise standardized predictors, we have $\nabla^2_{j_k}\ell = 1$ for $j = 1, .., J$ and $k = 1, ..., d_j$.

We re-write the RHS of (C.11) as follows:

$$\begin{aligned} RHS(\text{of C.11}) = {} & \nabla_{j_k}\ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ...\boldsymbol{\beta}_J)\delta \\ & + \delta\rho'(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1)\text{sgn}(u) \\ & + \rho'(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1)(|u+\delta| - |u| - \delta\text{sgn}(u)) \\ & + \frac{1}{2}\{\rho^{''}(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u^*, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1)(|u+\delta| - |u|)^2 + \delta^2\nabla^2_{j_k}\ell\} \end{aligned} \tag{C.12}$$

On the other hand, since u is to minimize $Q^1(\boldsymbol{\beta})$ w.r.t. β_{j_k}, we have,

$$\begin{aligned} & \nabla_{j_k}\ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ...\boldsymbol{\beta}_J) + \\ & \qquad \rho'(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1)\text{sgn}(u) = 0, \end{aligned} \tag{C.13}$$

Because $\chi(u)$ is minimized at u_0, by (C.13), we have

$$\begin{aligned} & \nabla_{j_k}\ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u_0, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ...\boldsymbol{\beta}_J) \\ & + \rho'(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u_0, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1)\text{sgn}(u_0) = 0, \end{aligned} \tag{C.14}$$

if $u_0 = 0$ then the above holds true for some value of $\text{sgn}(u_0) \in (-1, 1)$.

Observe that $\rho'(|t|) \geq 0$, then

$$\rho'(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1)[(|u+\delta| - |u|) - \text{sgn}(u)\delta] \geq 0 \tag{C.15}$$

Therefore using (C.14, C.15) in (C.12) at u_0, we have

$$\chi(u_0+\delta)-\chi(u_0) \geq \frac{1}{2}\{\rho^{''}(\|\beta_{j_1},...,\beta_{j_{k-1}},u^*,\beta_{j_{k+1}},...,\beta_{j_{d_j}}\|_1)(|u+\delta|-|u|)^2$$
$$+\delta^2\nabla^2_{j_k}\ell\} \geq \theta_2\delta^2, \tag{C.16}$$

since $(|u+\delta|-|u|)^2 \leq \delta^2$ and $\rho^{''}(|t|) < 0$ for SCAD and MCP with $\theta_2 = (\nabla^2_{j_k}\ell + \inf_t\rho^{''}(|t|))/2 > 0$ for the kth coordinate in jth group.

By similar arguments, we have a resembling result for the intercept, i.e.

$$\chi(u_0+\delta)-\chi(u_0) \quad \geq \quad \theta_1\delta^2, \tag{C.17}$$

with $\theta_1 = \frac{1}{2}\nabla^2_0\ell > 0$.

Let $\theta = \min(\theta_1,\theta_2)$, using (C.16,C.17), we have for all the coordinates of $\boldsymbol{\beta}$,

$$\chi(u_0+\delta)-\chi(u_0) \geq \theta\delta^2, \tag{C.18}$$

By (C.18) we have

$$\begin{aligned} Q^1(B^m_{j_{k-1}}) - Q^1(B^m_{j_k}) &\geq \theta(\beta^m_{j_k}-\beta^{m+1}_{j_k})^2 \\ &= \theta \parallel B^m_{j_{k-1}} - B^m_{j_k} \parallel_2^2, \end{aligned} \tag{C.19}$$

where $B^m_{j_k} = (\beta^{m+1}_0,...,(\boldsymbol{\beta}^{m+1}_{j-1})^T,\beta^{m+1}_{j_1},...,\beta^{m+1}_{j_k},\beta^m_{j_{k+1}},...,\beta^m_{j_{d_j}},(\boldsymbol{\beta}^m_{j+1})^T,...,(\boldsymbol{\beta}^m_J)^T)^T$. The (C.19) establishes the boundedness of the sequence $\{\boldsymbol{\beta}^m\}$ for every $m > 1$ since the starting point of $\{\boldsymbol{\beta}^1\} \in \mathbb{R}^{p+1}$. Apply (C.19) over all the coordinates, we have for all m

$$Q^1(\boldsymbol{\beta}^m) - Q^1(\boldsymbol{\beta}^{m+1}) \geq \theta \parallel \boldsymbol{\beta}^{m+1} - \boldsymbol{\beta}^m \parallel_2^2 . \tag{C.20}$$

Since the (decreasing) sequence $Q^1(\boldsymbol{\beta}^m)$ converges, (C.20) shows that the sequence $\{\boldsymbol{\beta}^m\}$ have a unique limit point. This completes the proof of the convergence of $\{\boldsymbol{\beta}^m\}$.

The assumption (3) and (4) of Lemma C.1 holds by (C.20). Hence, the limit point of $\{\boldsymbol{\beta}^m\}$ is a minimum of $Q^1(\boldsymbol{\beta})$ by Lemma C.1. This completes the proof of the convergence of the ℓ_1 grouped concave penalty.

C.2 Outline of Proof of Theorem 4.2

Proof. The proof of Theorem 4.2 follows the same logic as Theorem 4.1. Below we outline the major difference. In the case of GLM, we have a similar version as (C.11)

$$\begin{aligned}
\chi(u+\delta)-\chi(u) &= \ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u+\delta, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ...\boldsymbol{\beta}_J) \\
&- \ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ...\boldsymbol{\beta}_J) \\
&+ \rho(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u+\delta, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1) \\
&- \rho(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1) \\
&= \nabla_{j_k}\ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ...\boldsymbol{\beta}_J)\delta \\
&+ \frac{1}{2}\delta^2\nabla^2_{j_k}\ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ...\boldsymbol{\beta}_J) \\
&+ \rho'(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1)(|u+\delta|-|u|) \\
&+ \frac{1}{2}\rho''(\|\beta_{j_1}, ..., \beta_{j_{k-1}}, u^*, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}\|_1)(|u+\delta|-|u|)^2 + o(\delta^2),
\end{aligned} \tag{C.21}$$

Because $\chi(u)$ is minimized at u_0 for the surrogate function by the minimization

majorization approach, we then have

$$\begin{aligned}
0 &= \nabla_j \ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u_0 + \delta, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ...\boldsymbol{\beta}_J)) + M(u_0 - u_0 - \delta) \\
&+ \rho'(|u_0|)\text{sgn}(u_0) \\
&= \nabla_j \ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u_0, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ...\boldsymbol{\beta}_J) \\
&+ \nabla_j^2 \ell(\beta_0, \boldsymbol{\beta}_1, ..., \beta_{j_1}, ..., \beta_{j_{k-1}}, u_0, \beta_{j_{k+1}}, ..., \beta_{j_{d_j}}, ...\boldsymbol{\beta}_J)\delta \\
&- M\delta + \rho'(|u_0|)\text{sgn}(u_0) + o(\delta), \qquad \text{(C.22)}
\end{aligned}$$

if $u_0 = 0$ then the above holds true for some value of $\text{sgn}(u_0) \in (-1, 1)$.

Follow similar arguments as in the linear models, by (C.22), we still have

$$\chi(u_0 + \delta) - \chi(u_0) \geq \theta_2 \delta^2, \qquad \text{(C.23)}$$

with $\theta_2 = (M + \inf_t \rho''(|t|) + o(1))/2 > 0$ for $k = 1, ..., d_j,\ j = 1, ..., J$. The rest arguments follows as the proof for the ℓ_1 grouped concave penalty in linear models.

REFERENCES

Akaike, H. A new look at the statistical model identification. *IEEE T Automatic Control*. **1974**, *19(6)*, 716–723.

Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. **1975**, *12(4)*, 387–415.

Breheny, P.; Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with application to biological feature selection. *Annals of Applied Statistics*. **2011**, *5(1)*, 232–253.

Breiman, L. Heuristics of instability and stabilization in model selection. *Annals of Applied Statistics*. **1996**, *24(6)*, 2350–2383.

Donoho, D. L.; Johnstone, J. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*. **1994**, *81(3)*, 425–455.

Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Annals of Statistics*. **2004**, *32(2)*, 407–451.

Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*. **2001**, *96(456)*, 1348–13608.

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

Friedman, J.; Hastie, T.; Höfling, H.; Tibshirani, R. Pathwise coordinate optimization. *Annals of Applied Statistics*. **2007**, *1(2)*, 302–332.

Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. **2010**, *33(1)*, 1–22.

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

Hunter, D. R.; Lange, K. A tutorial on MM algorithms. *Journal of American Statistical Association*. **2004**, *58(1)*, 30–37.

Hunter, D. R.; Li, R. Variable selection using MM algorithms. *Annals of Statistics*. **2005**, *33(4)*, 1617–1642.

Jiang, D.; Huang, J.; Zhang, Y. The cross-validated AUC for MCP-Logistic regression with high-dimensional data. *Statistical Methods in Medical Research*. **2011**, Accepted.

Lange, K.; Hunter, D.; Yang, I. Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphics Statistics*. **2000**, *9(1)*, 1–59.

Ma, S.; Huang, J. Clustering threshold gradient descent regularization: with applications to microarray studies. *Bioinformatics*. **2007**, *23(4)*, 466–472.

Mallows, C. L. Some comments on Cp. *Technometrics*. **1973**, *15(4)*, 661–675.

Mazumder, R.; Friedman, J.; Hastie, T. SparseNet Coordinate descent with non-convex penalties. *Journal of American Statistical Association*. **2011**, *106(495)*, 1125–1138.

Meier, L.; van de Geer, S.; Bühlmann, P. The group lasso for logistic regression *Journal of Royal Statistical Society Series B*. **2008**, *70(1)*, 53–71.

Meinshausen, N.; Buhlmann, P. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*. **2006**, *34(3)*, 1436–1462.

Ortega, J. M.; Rheinbold, W. C. Iterative solution of nonlinear equations in several variables. *Academic Press, New York, NY*. **1970**.

Osborne, M. R.; Presnell, B.; Turlach, B. A. A new approach to variable selection in least square problems. *IMA Journal of Numerical Analysis*. **2000**, *20(3)*: 389-403.

Schifano, E. D.; Strawderman, R. L.; Wells, M. T. Majorization-minimization algorithms for nonsmoothly penalized objective functions. *Electronic Journal of Statistics*. **2010**, *4*, 1258–1299.

Schwarz, G. Estimation the dimension of a model. *Annals of Statistics*. **1978**, *6(2)*: 461–464.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society Series B*. **1996**, *58(1)*, 267–288.

Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of Royal Statistical Society Series B*. **2001**, *63(2)*, 411-423.

Tseng, P. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications*. **2001**, *109(3)*, 475–494.

van't Veer, L. J.; Dai, H.; van de Vijver, M. J.; *et al* Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. **2002**, *415(31)*, 530–536.

van de Vijver, M. J.; He, Y. D.; van't Veer, L. J.; *et al* A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*. **2002**, *347(25)*, 1999–2009.

Warge, J. Minimizing certain convex functions. *SIAM Journal on Applied Mathematics*. **1963**, *11(3)*, 588-593.

Wu, T. T.; Lange K. Coordinate descent algorithms for Lasso penalized regression. *Annals of Applied Statistics*. **2008**, *2(1)*, 224–244.

Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables *Journal of Royal Statistical Society Series B*. **2006**, *68(1)*, 49–67.

Zhang, C. H.; Huang, J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*. **2008**, *36(4)*, 1567–1594.

Zhang, C. H. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*. **2010**, *38(2)*, 894–942.

Zhao, P.; Yu, B. On model selection consistency of LASSO. *Journal of Machine Learning Research*. **2006**, *7*, 2541–2567.

Zou, H.; Li,R. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*. **2008**, *36(4)*, 1509–1533.