CONFIDENCE REGION AND INTERVALS FOR SPARSE PENALIZED REGRESSION USING VARIATIONAL INEQUALITY TECHNIQUES

Liang Yin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2015

Approved by:

Shu Lu

Yufeng Liu

Amarjit Budhiraja

Scott Provan

Kai Zhang

ABSTRACT

LIANG YIN: CONFIDENCE REGION AND INTERVALS FOR SPARSE PENALIZED REGRESSION USING VARIATIONAL INEQUALITY TECHNIQUES.

(Under the direction of Shu Lu and Yufeng Liu.)

With the abundance of large data, sparse penalized regression techniques are commonly used in data analysis due to the advantage of simultaneous variable selection and prediction. By introducing biases on the estimators, sparse penalized regression methods can often select a simpler model than unpenalized regression. A number of convex as well as non-convex penalties have been proposed in the literature to achieve sparsity. Despite intense work in this area, it remains unclear on how to perform valid inference for sparse penalized regression with a general penalty. In this work, by making use of state-of-the-art optimization tools in variational inequality theory, we propose a unified framework to construct confidence intervals for sparse penalized regression with a wide range of penalties, including the well-known least absolute shrinkage and selection operator (LASSO) penalty and the minimax concave penalty (MCP). We study the inference for two types of parameters: the parameters under the population version of the penalized regression and the parameters in the underlying linear model. Theoretical convergence properties of the proposed methods are obtained. Simulated and real data examples are presented to demonstrate the validity and effectiveness of the proposed inference procedure.

ACKNOWLEDGMENTS

Over the past five years I have received support and encouragement from a great number of individuals. First and foremost I want to thank my two advisors Shu Lu and Yufeng Liu. They provided unreserved support during my PhD and generously paved the way for my development as a research scientist. I appreciate all they contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. The joy and enthusiasm they have for their research were contagious and motivational for me, even during tough times in the Ph.D. pursuit. The members of the Dr Liu's group have contributed immensely to my personal time. The group has been a source of friendships as well as good advice and collaboration. I would like to acknowledge past and present group members Wonyul Lee, Qiang Sun, Chong Zhang, Guan Yu, Patrick Kimes, Yuying Xie and Junlong Zhao. We had valuable discussions of research, life and professional career.

I am also greatly indebted to the many people who in some way contributed to the progress of the work contained herein. I would like to thank my dissertation committee members: Kai Zhang, Amarjit Budhiraja and Scott Provan for their time, interest, helpful comments and insightful questions. I especially appreciate the help and comments provided by Michael Lamm. We had lots of great discussions for variational inequality techniques and shared our research work and experience.

My time at UNC was made enjoyable in large part due to the many friends that became a part of my life. I am grateful for time spent with roommates and friends, Tao Wang, Xinchun Shen, Xuzhe Shen, Minghui Liu, Dong Wang, Jie Xiong, Zhe Wang, Haojin Zhai, Zhankun Sun, Haifeng Lin, Nelson Lee, Di Miao, and for many other people and memories.

Lastly, I would like to thank my parents who raised me with a love of science and supported me in all my pursuits.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

1 INTRODUCTION 1

1.1 Sparse penalized regression and inference for statistical modeling 1

1.2 A population penalized approach for inference after penalized regression 3

1.3 New contributions and key techniques 5

1.4 Some preliminaries and notations on variational inequalities 8

1.5 Outline of the dissertation 12

2 INFERENCE FOR THE LASSO 14

2.1 Introduction 14

2.2 Problem transformations 15

2.2.1 Conversion to a standard quadratic program 15

2.2.2 The variational inequality and normal map formulation 18

2.2.3 Transformations of the SAA problem 21

2.3 Confidence intervals for the population LASSO parameters with fixed λ 24

2.3.1 The convergence and distributions of SAA solutions 24

2.3.2 Estimation of the B-derivative $d\Pi_S(z_0)$ and the normal map L_K 29

2.3.3 Confidence intervals for the normal map solutions 37

2.3.4 Confidence intervals for LASSO parameters 38

2.4 Confidence intervals for the population LASSO parameters with varying λ 39

2.4.1 Properties of z_N and $\Phi_N(z_N)$ 39

2.4.2 Properties of individual confidence bands for z_0 44

2.4.3 Properties of simultaneous confidence bands for z_0 46

2.4.4 Algorithms of LASSO confidence intervals along the path 53

2.5 Confidence intervals for the true parameters in the underlying linear model . . . 57

2.6 Numerical examples 68

2.6.1 Example 2.6.1: Asymptotic distribution of LASSO parameters 68

2.6.2 Example 2.6.2: Low dimensional simulation 71

2.6.3 Example 2.6.3: High dimensional simulation 74

2.6.4 Example 2.6.4: Coverage test for confidence bands 76

2.6.5 Example 2.6.5: Prostate cancer data 78

2.7 Summary 81

3 INFERENCE FOR GENERAL PENALIZED REGRESSIONS 83

3.1 Introduction 83

3.2 Problem transformations 84

3.2.1 Transformations of the population penalized regression 84

3.2.2 Transformations of the SAA problem 92

3.3 Confidence intervals for population penalized parameters 94

3.3.1 Convergence and distribution of SAA solutions 95

3.3.2 Estimators of Σ_0 and L_K 96

3.3.3 Confidence intervals for penalized parameters 100

3.4 Confidence intervals for the true parameters in the underlying linear model . . . 101

3.5 Numerical examples 106

3.5.1 Example 3.5.1: Low dimensional simulation 107

3.5.2 Example 3.5.2: High dimensional simulation 109

3.5.3 Example 3.5.3: Prostate cancer data 111

3.6 Summary 112

4 DISCUSSION 114

4.1 Hypothesis testing for sparse penalized regression 114

4.2 Inference for population constrained linear regression 114

REFERENCES 117

LIST OF TABLES

2.1 Cells in the normal manifold of S_i and the associated critical cones 30

2.2 Matrix representations of ψ_k for $k = 0, \cdots, 8$ 31

2.3 Distances between (β_i, t_i) and cells in the normal manifold of S_i 32

2.4 $E_i^0, \cdots, E_i^8$ in the plane (β_i, t_i) 40

2.5 90% CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.2. 72

2.6 90% individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 2.6.2. 72

2.7 Coverage of the individual CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.2. 73

2.8 Coverage of the individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 2.6.2. 73

2.9 Coverage and length of 95% individual CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.3. 75

2.10 Coverage and length of 95% individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 2.6.3. . . 75

2.11 95% CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.5. 78

2.12 95% individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 2.6.5. 79

3.1 Estimates and 95% CIs for $(\tilde{\beta}_0, \tilde{\beta})$ and $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 3.5.1. 109

3.2 Coverage and length of 95% individual CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 3.5.1. 110

3.3 Coverage and length of 95% individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 3.5.1. . . 110

3.4 Coverage and length of 95% individual CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 3.5.2. 111

3.5 Coverage and length of 95% individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 3.5.2. . . 111

3.6 Estimates and 95% CIs of $(\tilde{\beta}_0, \tilde{\beta})$ and $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 3.5.3. 112

LIST OF FIGURES

1.1 The normal and tangent cones of the polyhedron S. 9

2.1 The normal manifold of S_i. 30

2.2 $E_i^0, \cdots, E_i^8$ in the plane (β_i, t_i) 40

2.3 Distribution of SAA solutions in Example 2.6.1 71

2.4 Boxplot of coverage rates of 95% individual CIs for z_0 in Example 2.6.4 77

2.5 Boxplot of coverage rates of 95% individual CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.4 . . . 77

2.6 95% Confidence bands for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.5. 80

3.1 95% individual CIs of $(\tilde{\beta}_0, \tilde{\beta})$ and $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 3.5.1. 108

CHAPTER 1: INTRODUCTION

1.1 Sparse penalized regression and inference for statistical modeling

In recent years, significant developments have been made in high dimensional data analysis driven by the great needs in different scientific disciplines. Theory and methodology that are developed in modern research are generally guided by the following two aspects: (1) It is desirable for investigators to understand the mechanism in the data with a parsimonious model, to be found through data-driven model selection; and (2) Investigators often need to make statistical inference from the model they select. These two aspects, variable selection and inference, are two central issues in statistical modeling, which are particularly important when a large set of candidate explanatory variables is available for the model.

Regarding data-driven model selection procedures, traditional statistical methods such as ordinary least squares regression often give poor prediction accuracy and are weak in model interpretation for high dimensional problems. With the advantage of simultaneous variable selection and prediction, sparse penalized regression has been widely used. By introducing biases on the resulting estimators through sparse penalization, these methods can often produce estimators with much smaller variances and consequently lower mean square errors than unpenalized estimators. Furthermore, because of the built-in sparsity on the estimators, model selection and parameter estimation can be achieved in a single step. There is a large literature in this area including the L_1 regularized technique LASSO [11; 49], as well as many other extensions with different settings or penalties, see [18; 15; 60; 28; 59; 9; 51; 29; 33; 46; 55; 47], and many more. Lots of these extensions aim to obtain estimators with better properties such as lower bias [15; 55] and with structure [5; 53; 57; 58]. For computation, fast implementations have been proposed to handle data of very high dimensions. For example, the LARS algorithm by [13], the Coordinate-Descent algorithm by [52], and the Glmnet algorithm by [17] are three popular algorithms in practice.

After the data-driven selection, one common practice is to carry out conventional inference on the selected model. Despite its prevalence, this practice is problematic because it ignores the fact that the inference is conditional on the model selection that is itself stochastic. The stochastic nature of the selection process affects and distorts sampling distributions of the post-selection parameter estimates, leading to invalid post-selection inference. The problems of post-selection inference have long been recognized and have been discussed recently by [2; 6; 25; 26; 27].

In recent years, many methods have been developed to achieve valid inference after LASSO. We refer to [8] for a comprehensive review on these developments. We categorize these methods into the following three types of approaches:

- The simultaneous inference approach. This approach is guided by a general heuristic to consider all possible outcomes of the selected model and protect the valid inference for the worst scenario. Papers along this line include [3; 10; 36].

- The bias-correction approach. This approach considers adjusting for the bias that is introduced by the regularization step to achieve valid inference. Papers along this line include [7; 21; 56; 50].

- The conditional sampling distribution approach. This approach aims at understanding the asymptotic or exact distributions of some pivots conditional on the selected model and developing inference methods based on these distributions. Papers along this line include [24; 30].

Although so far there have been many methods can do inference after LASSO, the inference for other penalized regression is still untouched. Fan and Li [15] pointed out three properties for a good regularization penalty. The first one is sparsity. In order to reduce model complexity, regularized regression estimators should automatically set small coefficients to zero. Penalties with singularity at zero, such as LASSO, fulfill this requirement. The second one is the nearly unbiasedness. Although we must introduce biases for sparsity, we want the resulting estimators to be unbiased when the true coefficient is large. Common convex penalties can not achieve the above two properties together. Consequently, a number of non-convex penalties have been proposed to reduce the model bias, such as SCAD [15] and MCP [55] penalties. These penalties do not over penalize coefficients when the true coefficients are large. The last property is that the resulting estimators should be continuous with respect to the tuning parameter to improve stability in model prediction. A penalty function must be singular at the origin if it satisfies the first and third conditions. Therefore, the general penalized regressions with sparse penalties which may have these three properties deserve their own inference method.

1.2 A population penalized approach for inference after penalized regression

In this dissertation, we take a different view of the penalized regression and utilize the state-of-the-art stochastic variational inequality theory in optimization to construct confidence regions and confidence intervals. Consider the standard linear regression setting in which the penalized regression solves

$$\min_{\beta_0,\beta} \frac{1}{N}\big|\big|\mathbf{y} - \beta_0 1_N - \mathbf{X}\beta\big|\big|_2^2 + \sum_{j=1}^{p} P_{\lambda_j}(|\beta_j|), \tag{1.1}$$

where

$$\mathbf{y} \triangleq \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \mathbf{X} \triangleq \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^1 \\ \mathbf{x}^2 \\ \vdots \\ \mathbf{x}^N \end{bmatrix}, \quad 1_N = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^N,$$

and $(\mathbf{x}^1, y_1), \cdots, (\mathbf{x}^N, y_N)$ are independent samples. For each $i = 1, \cdots, N$ and $j = 1, \cdots, p$, $x_{ij} \in \mathbb{R}$, $y_i \in \mathbb{R}$ and $\mathbf{x}^i \in \mathbb{R}^p$. We use bold font to present data vectors and matrices. $\beta_0 \in \mathbb{R}$ and $\beta = (\beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$ are the regression parameters. $P_{\lambda_j}(|\cdot|)$ is a general penalty for β_j with the regularization parameter $\lambda_j > 0$. This general penalty covers the L_1 penalty, the adaptive LASSO penalty [59], or any other nonconvex penalty such as SCAD or MCP. Our interest is on the corresponding inference. To that end, we study the following population version of the penalized regression by solving

$$\min_{\beta_0,\beta} E\big[Y - \beta_0 - \sum_{i=1}^{p} \beta_i X_i\big]^2 + \sum_{j=1}^{p} P_{\lambda_j}(|\beta_j|), \tag{1.2}$$

where $X = (X_1, \cdots, X_p)^T \in \mathbb{R}^p$ is an explanatory random vector, and $Y \in \mathbb{R}$ is a response random variable. We refer to (1.1) as the sample average approximation (SAA) problem of the population penalized problem (1.2). Denote the solution to the SAA problem (1.1) as $(\hat{\beta}_0, \hat{\beta})$, which we refer to as penalized estimators. We will make use of the penalized estimators $(\hat{\beta}_0, \hat{\beta})$ to derive confidence intervals and regions for the population penalized parameters $(\tilde{\beta}_0, \tilde{\beta})$ as the solution of (1.2).

The population penalized approach is closely related to the traditional least squares approach. When penalty terms $P_{\lambda_j}(|\beta_j|)$ all take the value of 0, the problem (1.2) becomes the following population least squares problem:

$$\min_{\beta_0,\beta} \; E[Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j]^2, \tag{1.3}$$

which has a unique minimizer $(E[XX^T])^{-1}E[XY]$ when $E[XX^T]$ is invertible. If additionally X and Y are related by the following linear model

$$Y = \beta_0^{true} + X^T \beta^{true} + \varepsilon \tag{1.4}$$

with $E[\varepsilon|X] = 0$, then this solution to the population least squares problem (1.3) is exactly $(\beta_0^{true}, \beta^{true})$. When penalty terms $P_{\lambda_j}(|\beta_j|) > 0$, the solution to (1.2) is not exactly $(\beta_0^{true}, \beta^{true})$, but is related to $(\beta_0^{true}, \beta^{true})$ in a different way. We will also develop a method which utilizes that relation to construct confidence intervals for the true parameters above in the linear model (1.4).

Why could the minimizer from a population penalized regression be a reasonable target for scientific research? While it is apparent that a selection procedure such as LASSO is necessary when $p > N$, the population penalized approach is also meaningful when $N > p$. In the latter case, although the least squares inference of all coefficients in the model are readily available, it is well-known that including nearly collinear redundant variables in a regression model can "adjust away" some of the causal variables of interest (see discussions in Section 2 in [3]). Moreover, using the full model could be questionable in areas such as social science [1]. In these areas, it is common that when the question of "which variables should be included in the regression model" is asked, the scientific theory is not sufficient enough to dictate the inclusion or exclusion of variables for the inference (even when $N > p$). In this case, a data-driven model from sparse penalized regression would be helpful and more compelling. However, under this situation, the goal of the inference is slightly changed from that of the least squares approach: The investigator is no longer looking for the least squares coefficients that minimize the squared error loss in the population. Instead, she wants to find the least squares estimate subject to certain regularization on the model. Thus, this application of penalized regression leads naturally to the consideration of the population penalized parameters as the target of inference. On the other hand, under appropriately chosen nonconvex penalties, the difference between the population penalized parameters and the least squares regression parameters would be very small. Therefore in this case the inference for population penalized parameters is approximately valid for the least squares regression parameters.

The regularization scheme mentioned above relates closely to the regularization terms $P_{\lambda_j}(|\beta_j|)$. The major difference between the population penalized approach and the least squares approach is the incorporation of constraint information about the model/parameters. Though the source of such information can be from different perspectives, they can all be reflected in the penalty terms with λ_j as a measure of the strength of such information. Thus, the parameters in the population penalized approach are both scientifically and statistically meaningful: They lead to the best approximations to the response when external information is available. This interpretation is valid both for $N < p$ and for $N > p$.

1.3 New contributions and key techniques

In the work for this dissertation, the first contribution is to make use of the penalized (include nonconvex penalization) estimators $(\hat{\beta}_0, \hat{\beta})$ to derive confidence intervals and regions for the population penalized parameters $(\tilde{\beta}_0, \tilde{\beta})$. Our study on the inference of the population penalized parameters is based on study of the asymptotic distribution of penalized estimators (i.e., solutions to (1.1)), as they converge to the population penalized parameters (solution to (1.2)). A good understanding of such asymptotics around the population penalized parameters (as the right asymptotic target) will in turn provide important insights for the inference of true parameters in the linear model (1.4). We also note here that inferences for the population penalized parameters are by themselves meaningful probabilistic statements that are of practical use.

- Since penalized estimators are obtained by solving (1.1), they depend on random samples and are subject to uncertainty. Our inference results provide quantitative measures about the level of such uncertainty, by estimating the distance between the population penalized parameters and the computed penalized estimators. Sizes of those intervals are jointly determined by sample variability and sensitivity of penalized estimators with respect to random samples. Wide intervals indicate low reliability of the estimators, which can be caused by large sample variability or high sensitivity. Thus, these inference results can be used as quantitative assessments on the reliability and uncertainty level of penalized estimators obtained from (1.1).

- The inference results of this work can be used to assess the relative importance of predictors. For nonzero penalized estimators, conclusions can be made regarding whether the corresponding parameters are truly nonzero by checking if the corresponding intervals contain zero or not. For zero penalized estimators, the inference results can be highly informative as well. For example, if the confidence intervals of some penalized parameters are singletons of zero, then we have strong evidence to conclude that the corresponding population penalized parameters are zero.

Besides inference for the penalized parameters, the second contribution of this work is to develop an inference method for the true parameters in the linear model (1.4) via the penalized regression. Our method is based on a relationship between $\tilde{\beta}$ and β^{true} as well as their sample counterparts. To help explain our method, we can take the viewpoint of the following decomposition:

$$\hat{\beta} - \beta^{true} = \underbrace{\hat{\beta} - \tilde{\beta}}_{(*)} + \underbrace{\tilde{\beta} - \beta^{true}}_{(**)}. \tag{1.5}$$

In a sense, the decomposition in (1.5) is similar as the bias-variance decomposition. Through the population penalized approach, we are able to quantify the uncertainty in $(*)$ (or the "variance" part). Since the population penalized parameters $\tilde{\beta}$ is the asymptotic limit of the penalized estimators $\hat{\beta}$, the limiting distribution of $(*)$ characterizes the variation around $\tilde{\beta}$. Through a connection between $\tilde{\beta}$ and β^{true} that corrects the "bias" in $(**)$, we are able to provide valid inference for the true parameters. This method belongs to bias-correction category and is especially useful when the biases introduced by the penalization are large. Simulation results show that under LASSO regression our method performs competitively with existing methods with some gains on the width of confidence intervals for inactive variables in high dimensions.

In this dissertation, we develop the theories based on the fixed dimension p, although it is possible to extend this idea to the case of growing dimensions. The development of our method takes the following steps. First, we transform the problems (1.1) and (1.2) into their corresponding *normal map formulations*, which are equations with a $(2p+1)$-dimensional variable vector z. Next, we obtain the asymptotic distribution of solutions to the normal map formulation of (1.1), and find reliable estimates for quantities that appear in the asymptotic distribution. We then provide methods to compute simultaneous and individual confidence intervals for the solution to the normal map formulation of (1.2). Finally, we convert these confidence intervals into confidence intervals for the solution to (1.2). Note that our inference method is developed for fixed penalties $P_{\lambda_j}(|\beta_j|)$. In practice, the tuning parameters in the penalty terms can be chosen by various criteria or through cross validation.

At last, inspired by existing LASSO path algorithms such as [13], we are interested in the confidence band constructed by consecutively computing confidence intervals along the LASSO solution path with respect to tuning parameter λ. The third contribution of this work is to point out that our confidence intervals for the population LASSO parameters along their solution path have the "piecewise Lipschitz property" under some mild assumptions (That is, the endpoints of the confidence interval between two consecutive knots on a grid of λ are

Lipschitz continuous in λ), and to propose a linear approximation algorithm to track the entire confidence band. We only calculate CIs on the two ends of a λ interval on which the boundaries of the confidence band are Lipschitz, then we link the corresponding boundaries of these two confidence intervals to make an approximated confidence band on this interval. There are two key issues for this algorithm: Finding the λ knots which are cut-off points for piecewise Lipschitz property and calculating confidence interval on these knots. According to our experience, the number of such λ knots is $O(p)$, but unfortunately the computation for confidence intervals at some knots is very expensive. For the computational reason, we suggest a way to modify this tracking algorithm into a more efficient version for computing confidence intervals on a grid of λ values, by avoiding those computationally expensive λ knots. The tracking algorithm provides computational advantage when the confidence intervals are desired at many values of λ.

1.4 Some preliminaries and notations on variational inequalities

This section introduces some preliminary knowledge about variational inequalities, their relation with optimization problems, the normal map formulation, and normal manifolds. The book [14] provides a comprehensive treatment on finite dimensional variational inequalities. The normal map formulation for variational inequalities and normal manifolds for polyhedrons were introduced in [39; 40]. Detailed discussions on normal and tangent cones, faces, and relative interiors are contained in [42] and [43].

We start with definitions of normal cones and tangent cones. Let S be a closed, convex set in $\mathbb{R}^n$, and let $x \in S$. The normal cone to S at x is denoted by $N_S(x)$ and is defined as

$$N_S(x) = \{v \in \mathbb{R}^n \mid \langle v, s - x\rangle \leq 0 \text{ for each } s \in S\}.$$

The tangent cone to S at x is denoted by $T_S(x)$ and is defined as

$$T_S(x) = \{w \in \mathbb{R}^n | \; \exists \{x_k\} \subset S \text{ and } \{\tau_k\} \subset \mathbb{R} \text{ such that } x_k \to x, \tau_k \to 0, \text{ and } (x_k - x)/\tau_k \to w\}.$$

Roughly speaking, $T_S(x)$ contains all the directions along which x can be approached by a sequence of points in S, and $N_S(x)$ contains all the "normal" vectors to S at x. It is easy to see that $N_S(x)$ and $T_S(x)$ are indeed *cones* (a subset of $\mathbb{R}^n$ is a cone if a positive multiple of any element of it still belongs to it). In fact, $N_S(x)$ and $T_S(x)$ are the polar cones of each other, in the sense that the inner product of any element in $N_S(x)$ and any element in $T_S(x)$ is nonpositive, see, e.g., Proposition 1.3.2 of [14].

To illustrate these concepts, consider the polyhedron S in Figure 1.1, defined as $S = \{x \in \mathbb{R}^2 \mid x_1 + x_2 \leq 1, x_1 \geq 0, x_2 \geq 0\}$. Let $x^0 = (1, 0)$. For the moment, ignore z^0 in the figure. The middle graph shows the tangent cone $T_S(x^0)$, which is $\{w \in \mathbb{R}^2 \mid w_1 + w_2 \leq 0, w_2 \geq 0\}$. The right graph shows the normal cone $N_S(x^0) = \{v \in \mathbb{R}^2 \mid v_1 - v_2 \geq 0, v_1 \geq 0\}$.

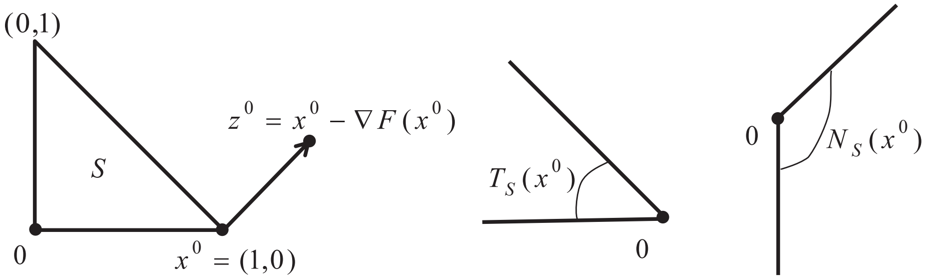


Figure 1.1: The normal and tangent cones of the polyhedron S.

Given a closed, convex set $S \subset \mathbb{R}^n$, and a function $f : \mathbb{R}^n \to \mathbb{R}^n$, the variational inequality associated with (f, S) is the problem of finding $x \in S$ such that

$$0 \in f(x) + N_S(x). \tag{1.6}$$

Here, $f(x) + N_S(x)$ is a set consisting of n-dim vectors of the form $f(x) + v$ for $v \in N_S(x)$. If the set $f(x) + N_S(x)$ contains the origin of $\mathbb{R}^n$, then x is a solution of (1.6).

To see how a variational inequality is related to an optimization problem, consider the problem of minimizing a function $F : \mathbb{R}^n \to \mathbb{R}$ over a closed and convex set S. If $x^0 \in S$ is a local solution to this minimization problem and F is differentiable at x^0, then x^0 satisfies the following variational inequality:

$$0 \in \nabla F(x^0) + N_S(x^0). \tag{1.7}$$

To prove (1.7), choose a point $s \in S$ and consider the line segment connecting x^0 and s. Since x^0 is a local minimum of F we have $\langle \nabla F(x^0), s - x \rangle \geq 0$. The latter inequality holds for any $s \in S$, which gives (1.7) in view of the definition of $N_S(x^0)$. Conversely, if x^0 satisfies (1.7) and F is a convex function, then x^0 is a global minimizer of F over the set S, because for each $s \in S$ one has $F(s) - F(x^0) \geq \langle \nabla F(x^0), s - x \rangle \geq 0$.

A variational inequality can be equivalently formulated as an equation using a concept called the *normal map*. To introduce this concept, let us first consider, for a fixed point $z \in \mathbb{R}^n$, the problem of minimizing $F(x) = \frac{1}{2}\|z - x\|^2$ over the set S. Applying the relation between optimization and variational inequalities, and noting $\nabla F(x) = x - z$, we find that the Euclidean projection $\Pi_S(z)$ is exactly the solution of the following inclusion

$$z - x \in N_S(x).$$

Now, we define the normal map induced by f and S, denoted by f_S, to be a function from $\mathbb{R}^n$ to $\mathbb{R}^n$ given by

$$f_S(z) = f(\Pi_S(z)) + (z - \Pi_S(z)) \text{ for each } z \in \mathbb{R}^n, \tag{1.8}$$

where $\Pi_S(\cdot)$ denotes the Euclidean projector onto S. One can then show for any solution x of (1.6) that the point $z = x - f(x)$ satisfies $\Pi_S(z) = x$ and

$$f_S(z) = 0. \tag{1.9}$$

Conversely, for any solution z of (1.9), the point $x = \Pi_S(z)$ is a solution of (1.6) and satisfies $z = x - f(x)$. Equation (1.9) is the normal map formulation of (1.6).

Let us revisit the example in Figure 1.1 to illustrate above concepts. Suppose $F(x) = \frac{1}{2}(x_1 - 1.5)^2 + \frac{1}{2}(x_2 - 0.5)^2$. It follows that $\nabla F(x^0) = (-0.5, -0.5)$, with $-\nabla F(x^0) = (0.5, 0.5) \in N_S(x^0)$. Hence, x^0 satisfies (1.7). Let $z^0 = x^0 - \nabla F(x^0) = (1.5, 0.5)$. Then $\Pi_S(z^0) = x^0$, and z^0 satisfies

$$\nabla F(\Pi_S(z^0)) + z^0 - \Pi_S(z^0) = \nabla F(x^0) + z^0 - x^0 = (-0.5, -0.5) + (1.5, 0.5) - (1, 0) = 0,$$

which means that z^0 is a solution to (1.9) with ∇F in place of f.

If the set S is a polyhedron in $\mathbb{R}^n$ (a set defined by finitely many affine constraints), then the Euclidean projector Π_S is a *piecewise affine* function from $\mathbb{R}^n$ to $\mathbb{R}^n$: it coincides with an affine function on each of finitely many n-dimensional polyhedrons whose union is $\mathbb{R}^n$ (the dimension of a convex set is defined to be the dimension of its affine hull, which is the smallest affine set containing the set). Those polyhedrons, along with their faces, are called cells in the normal manifold of S. We call a cell with dimension k a k-cell. The relative interiors of all cells in the normal manifold form a partition of $\mathbb{R}^n$ (the relative interior of a convex set is its interior relative to its affine hull). For the set S in Figure 1.1, Π_S is a piecewise affine function with 7 pieces. For example, $\Pi_S(z) = z$ for points z belonging to S, $\Pi_S(z) = (0, z_2)$ for points in the set $\{z \in \mathbb{R}^2 \mid z_1 \leq 0, 0 \leq z_2 \leq 1\}$, and $\Pi_S(z) = (0, 0)$ for points in the set $\{z \in \mathbb{R}^2 \mid z_1 \leq 0, z_2 \leq 0\}$. Those sets are 2-cells in the normal manifold of S. The halfline $\{z \in \mathbb{R}^2 \mid z_1 \leq 0, z_2 \leq 0\}$ and the edge $\{z \in \mathbb{R}^2 \mid z_1 = 0, 0 \leq z_2 \leq 1\}$ are 1-cells. In total, there are seven 2-cells, nine 1-cells, and three 0-cells (vertices of S).

Throughout this dissertation, we use $\|\cdot\|$ to denote the norm of an element in a normed space; unless explicitly stated otherwise, it can be any norm, as long as the same norm is used in all related contexts. We use $\mathcal{N}(0, \Sigma)$ to denote a Normal random vector with covariance matrix Σ. Weak convergence of random variables Y_n to Y will be denoted as $Y_n \Rightarrow Y$. A function $g : \mathbb{R}^n \to \mathbb{R}^m$ is said to be B-differentiable at a point $x_0 \in \mathbb{R}^n$ if there is a positively homogeneous function $G : \mathbb{R}^n \to \mathbb{R}^m$, such that

$$g(x_0 + v) = g(x_0) + G(v) + o(v).$$

The above function G is the B-derivative of g at x_0 and will be written as $dg(x_0)$. For each $h \in \mathbb{R}^n$, $dg(x_0)(h)$ is exactly the directional derivative of g at x_0 for the direction h. In general, B-differentiability is stronger than directional differentiability, as it requires $dg(x_0)(\cdot)$ to be a first order approximation of $g(x_0 + \cdot)$ uniformly in all directions.

1.5 Outline of the dissertation

In this dissertation, we will discuss how to use variational inequality techniques to compute confidence intervals for sparse penalized regression based on the penalty term. The main outline of this dissertation is as follows:

- In Chapter 2, we consider the LASSO regression and transform LASSO problems into variational inequalities to derive confidence intervals and regions for the population LASSO parameters. In terms of the true parameters in the underlying linear model, we propose a method to derive confidence intervals and compare them with existing methods in the literature. Moreover, we study the confidence bands for the population LASSO parameters along the LASSO solution path. We point out that the entire confidence band is neither piecewise linear nor continuous with respect to λ, if we construct confidence band pointwisely by using techniques described in this Chapter. We also propose a linear approximation tracking algorithm to compute confidence intervals.

- In Chapter 3, we consider a general penalized regression with the penalty term satisfying the three properties suggested by [15]. We propose a unified method to construct confidence intervals of the population penalized parameters for these penalized regressions, such as LASSO and MCP regression. For the true parameters in the underlying linear model, by correcting the bias introduced by the penalty term, we obtain asymptotic distribution of the true model estimator to construct the confidence intervals. Technically, we propose another problem transformation approach for the penalized regression optimization problem with general penalties, and extend those asymptotic results obtained in Chapter 2.

- In Chapter 4, we discuss two possible future directions. For the first direction, we point out that it is not trivial to conduct hypothesis testing and find the corresponding p-value for the population penalized parameters and the true model parameters using the asymptotic results in Chapter 3. Therefore it deserves further investigation. The second direction in Section 4.2 is to do inference for population constrained linear regression using variational inequality techniques.

CHAPTER 2: INFERENCE FOR THE LASSO

2.1 Introduction

In this Chapter, we discuss the inference after the LASSO regression, which is probably the most popular method in the family of sparse penalized regression with convex penalties. We consider the following population version of the random design LASSO problem

$$\min_{\beta_0,\beta} E\big[Y-\beta_0-\sum_{i=1}^{p}\beta_i X_i\big]^2+\lambda\sum_{i=1}^{p}|\beta_i|, \tag{2.1}$$

where $X=(X_1,\cdots,X_p)^T\in\mathbb{R}^p$ is an explanatory random vector, $Y\in\mathbb{R}$ is a response random variable, $\lambda>0$ is the regularization parameter, and $\beta_0\in\mathbb{R}$ and $\beta=(\beta_1,\cdots,\beta_p)\in\mathbb{R}^p$ are the regression parameters. The random design is commonly used to select a well performed model for out-of-sample prediction of population, which is of primary concern in many applications. The solution of (2.1) can be estimated by the solution of the corresponding SAA problem

$$\min_{\beta_0,\beta}\frac{1}{N}\big|\big|\mathbf{y}-\beta_0 1_N-\mathbf{X}\beta\big|\big|_2^2+\lambda\sum_{i=1}^{p}|\beta_i|, \tag{2.2}$$

where

$$\mathbf{y}\triangleq\begin{bmatrix} y_1\\ y_2\\ \vdots\\ y_N\end{bmatrix},\quad \mathbf{X}\triangleq\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p}\\ x_{21} & x_{22} & \cdots & x_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ x_{N1} & x_{N2} & \cdots & x_{Np}\end{bmatrix}=\begin{bmatrix}\mathbf{x}^1\\ \mathbf{x}^2\\ \vdots\\ \mathbf{x}^N\end{bmatrix},\quad 1_N=\begin{bmatrix}1\\ 1\\ \vdots\\ 1\end{bmatrix}\in\mathbb{R}^N,$$

and $(\mathbf{x}^1,y_1),\cdots,(\mathbf{x}^N,y_N)$ are independent samples of (X,Y). For each $i=1,\cdots,N$ and $j=1,\cdots,p$, $x_{ij}\in\mathbb{R}$, $y_i\in\mathbb{R}$ and $\mathbf{x}^i\in\mathbb{R}^{1\times p}$. For convenience, we write $\check{\mathbf{X}}=[1_N,\mathbf{X}]$. It is well known that the LASSO estimator $(\hat{\beta}_0,\hat{\beta})$ (the SAA solution) will almost surely converge to the population LASSO parameter $(\tilde{\beta}_0, \tilde{\beta})$ (the solution of (2.1)) as the sample size N goes to ∞. In order to indicate the reliability of this LASSO estimator, we construct confidence interval (CI) for the population LASSO parameter. For the linear model (1.4), we also propose a method to produce confidence intervals for the true parameters $(\beta_0^{true}, \beta^{true})$ (which solves (1.3)).

In Section 2.2, one can see how we transform the population LASSO problem (2.1) and its corresponding SAA problem (2.2) to their normal map formulations. The assumptions needed in this Chapter are also listed in this section. In Section 2.3, we show the methodology of producing confidence intervals for the population LASSO parameters at a fixed value of the regularization parameter λ. When λ changes, in Section 2.4 we study the properties of the confidence bands along the LASSO solution path, and propose sufficient algorithms to construct such bands. In Section 2.5, we establish a connection between the population LASSO parameters $(\tilde{\beta}_0, \tilde{\beta})$ and the true parameters $(\beta_0^{true}, \beta^{true})$. We use this connection to give an estimator of $(\beta_0^{true}, \beta^{true})$, which we denote as $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$, and obtain the asymptotic distribution of $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$ with fixed dimension p. Numerical results are presented in Section 2.6 to illustrate the performance of the proposed methods.

2.2 Problem transformations

In this section, we describe how to transform (2.2) and (2.1) into variational inequalities and normal map formulations, from where we obtain the asymptotic distribution of SAA solutions.

2.2.1 Conversion to a standard quadratic program

In this subsection, we transform the population LASSO problem into a standard quadratic programming problem. We need Assumption 2.1(a) below to guarantee the objective function of (2.1) to be finite valued. We will use the stronger Assumption 2.1(b) in proving convergence results.

Assumption 2.1. *(a) The expectations $E[X_1^2], \cdots, E[X_p^2]$, and $E[Y^2]$ are finite.*

(b) The expectations $E[X_1^4], \cdots, E[X_p^4]$, and $E[Y^4]$ are finite.

To eliminate the nonsmooth term $\sum_{j=1}^{p} |\beta_j|$ from the objective function of (2.1), we introduce a new variable $t \in \mathbb{R}^p$ into (2.1). The transformed problem is

$$\begin{aligned} &\min_{\beta_0,\beta,t} \ E[Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j]^2 + \lambda \sum_{j=1}^{p} t_j \\ &t_j - \beta_j \geq 0, j = 1, \cdots, p \\ &t_j + \beta_j \geq 0, j = 1, \cdots, p. \end{aligned} \tag{2.3}$$

We use S to denote the feasible set of (2.3):

$$S = \{(\beta_0, \beta, t) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \mid t_j - \beta_j \geq 0, t_j + \beta_j \geq 0, j = 1, \cdots, p\}. \tag{2.4}$$

If we write

$$(\beta_0, \beta, t) = (\beta_0, \beta_1, t_1, \beta_2, t_2, \cdots, \beta_p, t_p), \tag{2.5}$$

then we can treat the set S as a Cartesian product:

$$S = \mathbb{R} \times \prod_{i=1}^{p} S_i, \tag{2.6}$$

where for each $i = 1, \cdots, p$ the set S_i is a subset of $\mathbb{R}^2$ defined as

$$S_i = \{(\beta_i, t_i) \mid t_i - \beta_i \geq 0, t_i + \beta_i \geq 0\}. \tag{2.7}$$

Note that in equation (2.5) two ways of ordering elements in (β_0, β, t) are used. We refer to the ordering on the right hand side in (2.5) as "cross" ordering, and the ordering on the left hand side as "block" ordering. Unless explicitly stated otherwise, vectors and matrices are ordered using "block" ordering.

In the next subsection we will transform (2.3) into a variational inequality. This requires writing down the gradient of its objective function. To this end, define a function $F : \mathbb{R} \times \mathbb{R}^p \times$

$\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^{2p+1}$ by

$$F(\beta_0, \beta, t, X, Y) = \begin{bmatrix} -2(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j) \\ -2(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j) X_1 \\ \vdots \\ -2(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j) X_p \\ \lambda e_p \end{bmatrix}, \tag{2.8}$$

where e_p is the p-dimensional vector with all entries being 1. Clearly, F is a continuously differentiable function, and its derivative with respect to (β_0, β, t) at $(\beta_0, \beta, t, X, Y)$ is given by

$$d_1 F(\beta_0, \beta, t, X, Y) = \begin{bmatrix} 2 & 2X^T & 0 \\ 2X & 2XX^T & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.9}$$

Next, define a function $f_0 : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{2p+1}$ by

$$f_0(\beta_0, \beta, t) = E[F(\beta_0, \beta, t, X, Y)]. \tag{2.10}$$

Assumption 1(a) guarantees f_0 to be well defined and finite valued. Moreover, f_0 is an affine function, with its Jacobian matrix being

$$L = E[d_1 F(\beta_0, \beta, t, X, Y)] = \begin{bmatrix} 2 & 2E[X^T] & 0 \\ 2E[X] & 2E[XX^T] & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.11}$$

The following lemma is relatively straightforward and its proof is omitted.

Lemma 2.1. *Suppose Assumption 2.1(a) holds. Then, the objective function of* (2.3) *is a finite valued, convex quadratic function on* $\mathbb{R}^{2p+1}$*, its gradient at each* $(\beta_0, \beta, t) \in \mathbb{R}^{2p+1}$ *is* $f_0(\beta_0, \beta, t)$*, and its Hessian matrix is* L.

We now introduce the second assumption.

Assumption 2.2. *Let $(\tilde{\beta}_0, \tilde{\beta})$ be an optimal solution of (1.2), define $\tilde{t} \in \mathbb{R}^p$ and $\tilde{q} \in \mathbb{R}^p$ by*

$$\tilde{t}_i = |\tilde{\beta}_i| \text{ and } \tilde{q}_i = E[-2(Y - \tilde{\beta}_0 - \sum_{j=1}^{p} \tilde{\beta}_j X_j) X_i] \text{ for each } i = 1, \cdots, p.$$

Let $\mathfrak{I}$ be a subset of $\{1, \cdots, p\}$ defined as

$$\mathfrak{I} = \Big\{ i \in \{1, \cdots, p\} \mid \tilde{\beta}_i \neq 0 \quad or \quad (\tilde{\beta}_i = 0 \text{ and } |\tilde{q}_i| = \lambda) \Big\},$$

and let $L_{\mathfrak{I}}$ be the submatrix of L in (2.11) *that consists of intersections of columns and rows of L with indices in $\{1\} \cup \{i+1, i \in \mathfrak{I}\}$. Assume that $L_{\mathfrak{I}}$ is nonsingular.*

In the above assumption, the vector $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is indeed a solution of (2.3), and Q is a submatrix of the upperleft $(p+1) \times (p+1)$ submatrix of L. Lemma 2.2 of the next subsection will show that the non-singularity of Q guarantees $(\tilde{\beta}_0, \tilde{\beta})$ to be the global unique solution of (2.1).

2.2.2 The variational inequality and normal map formulation

In view of Lemma 2.1, we can rewrite (2.3) as the following variational inequality:

$$-f_0(\beta_0, \beta, t) \in N_S(\beta_0, \beta, t). \tag{2.12}$$

If we would introduce multipliers for constraints defining S in (2.4), we could write down an explicit expression for $N_S(\beta_0, \beta, t)$ and accordingly rewrite (2.12) into the well-known Karush-Kuhn-Tucker conditions. However, that approach would lead to more variables (the multipliers) in the formulation and we would need additional assumptions to ensure the uniqueness of multipliers. For this reason, we choose to deal with (2.12) directly.

Let $(f_0)_S$ be the normal map induced by f_0 and S, as defined in (1.8) with f_0 in place of f. The normal map formulation for (2.12) is

$$(f_0)_S(z) = 0, \tag{2.13}$$

where z is a variable of dimension $2p+1$.

As noted right below Assumption 2.2, the vector $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a solution of (2.3). It is therefore a solution of (2.12) as well. By the relation between variational inequalities and normal maps, the point $z_0 \in \mathbb{R}^{2p+1}$ defined as

$$z_0 = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) - f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \tag{2.14}$$

is a solution to (2.13) and satisfies $\Pi_S(z_0) = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$. Let K be the *critical cone* to S associated with z_0, defined as

$$\begin{aligned} K &= \{w \in T_S(\Pi_S(z_0)) \mid \langle z_0 - \Pi_S(z_0), w\rangle = 0\} \\ &= \{w \in T_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \mid \langle f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}), w\rangle = 0\}. \end{aligned} \tag{2.15}$$

Using the special polyhedral structure of S, we will give an explicit expression of K in the proof of Lemma 2.2 below. Critical cones are commonly used in optimization to define conditions on optimality and local uniqueness of solutions, see, e.g., [38]. We use critical cones here for the same purposes, but also for writing down an expression of the asymptotic distribution of SAA solutions. Let L_K be the normal map induced by the linear function L as in (2.11) and the cone K, defined as in (1.8) with L and K in place of f and S respectively. In Lemma 2.2 below, we show that L_K is a global homeomorphism from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$, that is, a continuous bijective function from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$ whose inverse function is also continuous. The inverse function of L_K will appear in an expression for the asymptotic distribution of SAA solutions.

Lemma 2.2. *Suppose that Assumptions 2.1(a) and 2.2 hold. Then the normal map L_K is a global homeomorphism from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$, and $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is the unique optimal solution of (2.3).*

Proof of Lemma 2.2. We start by examining the structure of K. In view of (2.6), the tangent and normal cones to S at $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ can be written as

$$T_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = \mathbb{R} \times T_{S_1}(\tilde{\beta}_1, \tilde{t}_1) \times \cdots \times T_{S_p}(\tilde{\beta}_p, \tilde{t}_p),$$

and

$$N_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = \{0\} \times N_{S_1}(\tilde{\beta}_1, \tilde{t}_1) \times \cdots \times N_{S_p}(\tilde{\beta}_p, \tilde{t}_p).$$

Let $\tilde{q}$ be as defined in Assumption 2.2, and let $\tilde{q}_0 = E[-2(Y - \tilde{\beta}_0 - \sum_{j=1}^{p} \tilde{\beta}_j X_j)]$. Since $f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = (\tilde{q}_0, \tilde{q}, \lambda e_p)$ and $-f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \in N_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, we have

$$\tilde{q}_0 = 0 \text{ and } -(\tilde{q}_i, \lambda) \in N_{S_i}(\tilde{\beta}_i, \tilde{t}_i) \text{ for each } i = 1, \cdots, p. \tag{2.16}$$

Now choose an arbitrary $v \in T_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, and write it as

$$v = (v_0, v_1, \cdots, v_p)$$

with $v_0 \in \mathbb{R}$ and $v_i \in T_{S_i}(\tilde{\beta}_i, \tilde{t}_i)$ for each $i = 1, \cdots, p$. It is not hard to see that v belongs to K if and only if $\langle -(\tilde{q}_i, \lambda), v_i \rangle = 0$ for each $i = 1, \cdots, p$. We can therefore write K as

$$K = \mathbb{R} \times K_1 \times \cdots \times K_p$$

where

$$K_i = \{v_i \in T_{S_i}(\tilde{\beta}_i, \tilde{t}_i) \mid \langle -(\tilde{q}_i, \lambda), v_i \rangle = 0\} \text{ for each } i = 1, \cdots, p.$$

From (2.45), for each $i = 1, \cdots, p$ we have

$$K_i = \begin{cases} \{(0,0)\} & \text{if } (\tilde{\beta}_i = 0 \text{ and } |\tilde{q}_i| < \lambda), \\ \{(\beta_i, t_i) \in \mathbb{R}^2_+ \mid \beta_i - t_i = 0\} & \text{if } (\tilde{\beta}_i = 0 \text{ and } \tilde{q}_i = -\lambda), \\ \{(\beta_i, t_i) \in \mathbb{R}^2 \mid \beta_i - t_i = 0\} & \text{if } \tilde{\beta}_i > 0, \\ \{(\beta_i, t_i) \in \mathbb{R}_- \times \mathbb{R}_+ \mid \beta_i + t_i = 0\} & \text{if } (\tilde{\beta}_i = 0 \text{ and } \tilde{q}_i = \lambda), \\ \{(\beta_i, t_i) \in \mathbb{R}^2 \mid \beta_i + t_i = 0\} & \text{if } \tilde{\beta}_i < 0. \end{cases} \tag{2.17}$$

We can now give an explicit expression for the affine hull of K. Define two matrices M and

N as follows:

$$M = \begin{bmatrix} 1 & 0 \\ 0 & I_p \\ 0 & I_p \end{bmatrix} \text{ and } N = \begin{bmatrix} 1 & 0 \\ 0 & I_p \\ 0 & -I_p \end{bmatrix},$$

where I_p is the $p \times p$ identity matrix. Construct a matrix Ξ by first adding the common first column of M and N and then adding the $i+1$'th column of M (N) if the condition in the second or third (fourth or fifth) row of (3.20) is satisfied. Columns of Ξ form a basis of the affine hull of K. It is not hard to check that $\Xi^T L \Xi = Q$, where Q is defined in Assumption 2.2. The latter assumption ensures Q to be nonsingular, so it is positive definite. It follows from an application of [39, Theorem 4.3] that L_K is a global homeomorphism. By [40, Theorem 3], $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a locally unique solution to (2.12). Thus, it is a locally unique solution to (2.3). But the objective function of (2.3) is convex, so $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is indeed the global unique solution of (2.3).

□

In the rest of this chapter, we use Σ_0 to denote the covariance matrix of $F(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}, X, Y)$, and let Σ_0^1 be the upper left $(p+1) \times (p+1)$ submatrix of Σ_0. Since the last p elements of $F(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}, X, Y)$ are fixed at λ, we have

$$\Sigma_0 = \begin{bmatrix} \Sigma_0^1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{2.18}$$

In addition, we make the following non-degeneracy condition

Assumption 2.3. *The determinant of Σ_0^1 defined in (2.18) is strictly positive.*

2.2.3 Transformations of the SAA problem

So far we have reformulated (2.1) as a quadratic program (2.3), a variational inequality (2.12), and an equation involving the normal map (2.13). We can reformulate the SAA problem (2.2) in a similar way. By introducing the variable vector t, we rewrite (2.2) as the following problem:

$$\min_{(\beta_0,\beta,t)\in S} \frac{1}{N}\big|\big|\mathbf{y} - \beta_0 1_N - \mathbf{X}\beta\big|\big|_2^2 + \lambda \sum_{j=1}^{p} t_j, \tag{2.19}$$

where S is as defined in (2.4). We define the *SAA function*

$$f_N(\beta_0, \beta, t) = N^{-1} \sum_{i=1}^{N} F(\beta_0, \beta, t, \mathbf{x}^i, y_i),$$

where F is as in (2.8). By noting that $f_N(\beta_0, \beta, t)$ is exactly the gradient of the objective function of (2.19) at (β_0, β, t), we can rewrite (2.19) as a variational inequality

$$0 \in f_N(\beta_0, \beta, t) + N_S(\beta_0, \beta, t). \tag{2.20}$$

The above f_N is an affine function with its Jacobian matrix given by

$$L_N = df_N(\beta_0, \beta, t) = \begin{bmatrix} 2 & 2\sum_{i=1}^{N} \mathbf{x}^i/N & 0 \\ 2\sum_{i=1}^{N} (\mathbf{x}^i)^T/N & 2\sum_{i=1}^{T} (\mathbf{x}^i)^T(\mathbf{x}^i)/N & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.21}$$

Finally, we let $(f_N)_S$ be the normal map induced by f_N and S, and write the normal map formulation of (2.20) as

$$(f_N)_S(z) = 0. \tag{2.22}$$

In Section 2.3 we will discuss the asymptotic distributions and convergence rates of solutions of (2.20) and (2.22), and generate confidence regions and confidence intervals for solutions of (2.12) and (2.13). While Assumptions 2.1 and 2.2 are sufficient for the asymptotic distribution results to hold, the results on convergence rates require additional assumptions, which are introduced below. Assumption 2.4(a) imposes conditions on the random variable $F(\beta_0, \beta, t, X, Y)$ to ensure the SAA function f_N to converge to f_0 in probability at an exponential rate. These conditions will hold, for example, if (X, Y) is a bounded random variable. The other parts impose the same type of assumptions on different random variables.

Assumption 2.4. *(a) For each $h \in \mathbb{R}^{2p+1}$ and $(\beta_0, \beta, t) \in \mathbb{R}^{2p+1}$, let*

$$M_{\beta_0,\beta,t}(h) = E\big[\exp\{\langle h, F(\beta_0, \beta, t, X, Y) - f_0(\beta_0, \beta, t)\rangle\}\big]$$

be the moment generating function of the random variable $F(\beta_0, \beta, t, X, Y) - f_0(\beta_0, \beta, t)$. Let $\mathcal{C}$ be a compact set in $\mathbb{R}^{2p+1}$ that contains $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ in its interior. Assume the following conditions.

1. *There exists a constant $\zeta > 0$ such that $M_{\beta_0,\beta,t}(h) \le \exp\{\zeta^2\|h\|^2/2\}$ for each $h \in \mathbb{R}^{2p+1}$ and $(\beta_0, \beta, t) \in \mathcal{C}$.*

2. *There exists a nonnegative random variable $\iota(X, Y)$ such that*

   $$\|F(\beta_0, \beta, t, X, Y) - F(\beta_0', \beta', t', X, Y)\| \le \iota(X, Y)\|(\beta_0, \beta, t) - (\beta_0', \beta', t')\|$$

   for all (β_0, β, t) and (β_0', β', t') in $\mathcal{C}$ and almost every (X, Y).

3. *The moment generating function of ι is finite valued in a neighborhood of zero.*

(b) The same conditions as in (a) for $d_1F(\beta_0, \beta, t, X, Y)$ instead of $F(\beta_0, \beta, t, X, Y)$. Accordingly, use $E[d_1F(\beta_0, \beta, t, X, Y)]$ to replace $f_0(\beta_0, \beta, t)$ in the conditions.

(c) The same conditions as in (a) for $F(\beta_0, \beta, t, X, Y)F(\beta_0, \beta, t, X, Y)^T$. Accordingly, use $E[F(\beta_0, \beta, t, X, Y)F(\beta_0, \beta, t, X, Y)^T]$ to replace $f_0(\beta_0, \beta, t)$ in the conditions.

Assumption 2.4(a-b) will enable us to show that solutions of (2.22) converge to the solution of (2.13) in probability at an exponential rate (see Theorem 2.1 in Section 2.3.1). We need such an exponential convergence rate to construct reliable estimates for an unknown quantity in an expression of the asymptotic distribution of solutions of (2.13). Assumption 2.4(c) will be needed only for the situations in which the matrix Σ_0^1 defined in (2.18) is singular (see Theorem 2.3); for such situations we will use Assumption 2.4(c) to derive the exponential convergence rate of an estimate of Σ_0^1.

2.3 Confidence intervals for the population LASSO parameters with fixed λ

This section proposes a method to compute confidence intervals and regions for solutions of the population LASSO problem (2.1) with fixed λ based on the solutions to the SAA problem (2.2). Section 2.3.1 below provides convergence properties and asymptotic distributions of solutions to the variational inequality (2.20) and normal map formulation (2.22) of the SAA problem. Section 2.3.2 explains more details on how to estimate quantities that appear in the asymptotic distributions. Following that, Section 2.3.3 shows how to compute confidence intervals for the solution to the normal map formulation (2.13) of (2.1). Finally, Section 2.3.4 discusses how to convert the latter confidence intervals to confidence intervals for solutions of (2.1).

2.3.1 The convergence and distributions of SAA solutions

Theorem 2.1 below provides convergence properties and asymptotic distributions of solutions of the SAA problems (2.20) and (2.22). It shows under Assumptions 2.1 and 2.2 that (2.22) has a unique solution z_N for sufficiently large N, and that z_N converges almost surely to z_0 defined in (2.14). Correspondingly, the projection $\Pi_S(z_N)$ is the unique solution of (2.20), which converges almost surely to $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$. This theorem also provides asymptotic distributions of z_N and $\Pi_S(z_N)$, and gives their convergence rate in probability under Assumption (2.4)(a-b).

Theorem 2.1. *Suppose that Assumptions 2.1 and 2.2 hold. Then, for almost every $\omega \in \Omega$, there exists an integer N_ω, such that for each $N \geq N_\omega$, the equation* (2.22) *has a unique solution z_N in $\mathbb{R}^{2p+1}$, and the variational inequality* (2.20) *has a unique solution in $\mathbb{R}^{2p+1}$ given by $(\hat{\beta}_0, \hat{\beta}, \hat{t}) = \Pi_S(z_N)$. Moreover,*

$$\lim_{N\to\infty} z_N = z_0 \text{ a.e.}, \qquad \lim_{N\to\infty} (\hat{\beta}_0, \hat{\beta}, \hat{t}) = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \text{ a.e.}, \tag{2.23}$$

$$\sqrt{N}(z_N - z_0) \Rightarrow (L_K)^{-1}(\mathcal{N}(0, \Sigma_0)), \tag{2.24}$$

$$\sqrt{N}(\Pi_S(z_N) - \Pi_S(z_0)) \Rightarrow \Pi_K \circ (L_K)^{-1}(\mathcal{N}(0, \Sigma_0)), \tag{2.25}$$

and

$$\sqrt{N} L_K(z_N - z_0) \Rightarrow \mathcal{N}(0, \Sigma_0). \tag{2.26}$$

Suppose in addition that Assumption 2.4(a-b) holds. Then there exist positive real numbers ϵ_0, δ_0*,* μ_0*,* M_0 *and* σ_0*, such that the following inequality holds for each* $\epsilon \in (0, \epsilon_0]$ *and each* N*:*

$$\begin{aligned} &\text{Prob}\left\{\|(\hat{\beta}_0, \hat{\beta}, \hat{t}) - (\tilde{\beta}_0, \tilde{\beta}, \tilde{t})\| < \epsilon\right\} \geq \text{Prob}\left\{\|z_N - z_0\| < \epsilon\right\} \\ \geq &1 - \delta_0 \exp\{-N\mu_0\} - \frac{M_0}{\epsilon^{2p+1}} \exp\left\{-\frac{N\epsilon^2}{\sigma_0}\right\}. \end{aligned} \tag{2.27}$$

Proof of Theorem 2.1. The conclusions will follow from an application of [32, Theorem 7]. First, we verify assumptions of the latter theorem. It can be seen from equations (2.8) and (2.9) that Assumption 1 in [32] holds under Assumption 2.1 of this paper. Assumption 2 in [32] holds as a result of Lemma 2.2. Finally, let $\mathcal{C}$ be a compact set in $\mathbb{R}^{2p+1}$ that contains $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ in its interior. If Assumptions 2.4(a-b) of this paper are satisfied for this $\mathcal{C}$, then Assumption 4 in [32] is satisfied.

By [32, Theorem 7], there exist neighborhoods Z of z_0 and $\mathcal{C}_0$ of $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, and an integer N_ω for almost every $\omega \in \Omega$, such that for each $N \geq N_w$, the equation (2.22) has a unique solution z_N in Z, and the variational inequality (2.20) has a unique solution in $\mathcal{C}_0$ given by $\Pi_S(z_N)$. Equations (3.27), (3.28), (3.30) and (3.31) follow from this theorem. Because the objective function in (2.19) is convex, $\Pi_S(z_N)$ is in fact the globally unique solution for (2.19). From the equivalence between (2.19), (2.20) and (2.22), it follows that z_N and $\Pi_S(x_N)$ are the globally unique solutions to (2.22) and (2.20) respectively.

It remains to prove equation (3.29). Note that the function Π_S is B-differentiable, and its B-derivative at z_0 is exactly Π_K. In view of (3.28), we can apply the Delta theorem (see, for example, [32, Theorem 6]) to Π_S to obtain (3.29).

□

In the above theorem, L_K is the normal map induced by the linear function L in (2.11) and the critical cone K in (2.15). Since K is a polyhedral convex cone, the Euclidean projector Π_K is a piecewise linear function (a function that coincides with a linear function on each of finitely many polyhedral convex cones whose union is the entire space). The normal map L_K is therefore a piecewise linear function as well. If K happens to be a subspace, then Π_K and L_K are linear functions. By Lemma 2.2, L_K is a global homeomorphism under Assumptions 2.1(a) and 2.2. The inverse function $(L_K)^{-1}$ is again a piecewise linear function, and it is a linear function if K is a subspace. Equation (3.28) implies that $\sqrt{N}(z_N - z_0)$ asymptotically follows a normal distribution if K is a subspace, and that the asymptotic distribution is not normal if K is not a subspace. Equation (3.29) gives the asymptotic distribution of $(\hat{\beta}_0, \hat{\beta}, \hat{t}) = \Pi_S(z_N)$, which is the solution of (2.20) or equivalently (2.19). Equation (3.31) shows that z_N converges to z_0 in probability at an exponential rate, as N goes to ∞.

In this section, our objective is to develop a method to compute confidence regions and confidence intervals for z_0 and $(\tilde{\beta}_0, \tilde{\beta})$. After solving the LASSO (2.2) to find its solution $(\hat{\beta}_0, \hat{\beta})$, we let $\hat{t} = |\hat{\beta}|$ so that $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ solves (2.19) and equivalently (2.20). We can then compute z_N by

$$z_N = (\hat{\beta}_0, \hat{\beta}, \hat{t}) - f_N(\hat{\beta}_0, \hat{\beta}, \hat{t}), \tag{2.28}$$

which solves (2.22) and satisfies $(\hat{\beta}_0, \hat{\beta}, \hat{t}) = \Pi_S(z_N)$. Now that z_N is known, from (3.30) one can readily write down an expression for the confidence region of z_0 by using the χ^2 distribution. That expression contains unknown objects Σ_0 and L_K, and we describe below how to estimate those objects.

We will substitute Σ_0 by Σ_N, the sample covariance matrix of $\{F(\hat{\beta}_0, \hat{\beta}, \hat{t}, \mathbf{x}^i, y_i)\}_{i=1}^N$. Let Σ_N^1 be the upperleft $(p+1) \times (p+1)$ submatrix of Σ_N; we have $\Sigma_N = \begin{bmatrix} \Sigma_N^1 & 0 \\ 0 & 0 \end{bmatrix}$. The following lemma shows that Σ_N converges to Σ_0 almost surely, and provides a rate of the convergence of Σ_N in probability.

Lemma 2.3. *Suppose that Assumptions 2.1, 2.2 and 2.4(a-b) hold. Then Σ_N converges to Σ_0 almost surely. If Assumption 2.4(c) holds additionally then there exist positive real numbers δ_1,*

μ_1, M_1 and σ_1, such that the following inequality holds for each $\epsilon > 0$ and each N:

$$\mathrm{Prob}\left\{\|\Sigma_N - \Sigma_0\| < \epsilon\right\} \geq 1 - \delta_1 \exp\{-N\mu_1\} - \frac{M_1}{\min(\epsilon^{(2p+1)^2}, \epsilon^{2p+1})} \exp\left\{-\frac{N\epsilon^2}{\sigma_1}\right\}. \tag{2.29}$$

Proof of Lemma 2.3. Define a function $\Theta : \mathbb{R}^{3p+2} \to \mathbb{R}^{(2p+1)\times(2p+1)}$ by

$$\Theta(\beta_0, \beta, t, X, Y) = F(\beta_0, \beta, t, X, Y)F(\beta_0, \beta, t, X, Y)^T,$$

let $\theta_0(\beta_0, \beta, t) = E[\Theta(\beta_0, \beta, t, X, Y)]$, and for each $N \in \mathbb{N}$ define the sample average function as

$$\theta_N(\beta_0, \beta, t) = \frac{1}{N}\sum_{i=1}^{N} F(\beta_0, \beta, t, x_i, y_i)F(\beta_0, \beta, t, x_i, y_i)^T = \frac{1}{N}\sum_{i=1}^{N} \Theta(\beta_0, \beta, t, x_i, y_i).$$

Note that entries of $\Theta(\beta_0, \beta, t, X, Y)$ are linear combinations of terms $Y^2 X_i X_j$, $\beta_i\beta_j X_i X_j X_k X_l$, and terms of lower degrees, where $i, j, k, l = 1, \cdots, p$. Assumption 2.1 guarantees that θ_0 is finite valued. Moreover, applying [32, Theorem 3(a)] to Θ, we see that θ_0 is a continuous function and that θ_N converges uniformly to θ_0 on compact sets almost surely.

The covariance matrices Σ_N and Σ_0 are given by

$$\Sigma_N = \theta_N(\hat{\beta}_0, \hat{\beta}, \hat{t}) - f_N(\hat{\beta}_0, \hat{\beta}, \hat{t})f_N(\hat{\beta}_0, \hat{\beta}, \hat{t})^T$$

and

$$\Sigma_0 = \theta_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) - f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})^T.$$

It was shown in Theorem 3.1 that $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ converges to $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ almost surely. Consequently, $\theta_N(\hat{\beta}_0, \hat{\beta}, \hat{t})$ converges almost surely to $\theta_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$. Similarly $f_N(\hat{\beta}_0, \hat{\beta}, \hat{t}) \to f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ almost surely. Consequently Σ_N converges to Σ_0 almost surely.

Let $\mathcal{C}$ be a compact set that contains $\tilde{\beta}_0, \tilde{\beta}, \tilde{t}$ in its interior. If Assumption 2.4(c) holds, we can apply [45, Theorem 7.67] (see also [32, Theorem 4(a)]) to find positive real numbers δ_2, μ_2,

M_2 and σ_2, such that the following holds for each $\epsilon > 0$ and each N:

$$\text{Prob}\left\{\sup_{(\beta_0,\beta,t)\in\mathcal{C}} \|\theta_N(\beta_0,\beta,t) - \theta_0(\beta_0,\beta,t)\| \geq \epsilon\right\} \leq \delta_2 \exp\{-N\mu_2\} + \frac{M_2}{\epsilon^{(2p+1)^2}} \exp\left\{-\frac{N\epsilon^2}{\sigma_2}\right\}. \tag{2.30}$$

Similarly, under Assumption 2.4(a) we can apply [45, Theorem 7.67] to f_N to obtain positive real numbers δ_3, μ_3, M_3 and σ_3, such that the following holds for each $\epsilon > 0$ and each N:

$$\text{Prob}\left\{\sup_{(\beta_0,\beta,t)\in\mathcal{C}} \|f_N(\beta_0,\beta,t) - f_0(\beta_0,\beta,t)\| \geq \epsilon\right\} \leq \delta_3 \exp\{-N\mu_3\} + \frac{M_3}{\epsilon^{2p+1}} \exp\left\{-\frac{N\epsilon^2}{\sigma_3}\right\}. \tag{2.31}$$

Since $\|\Sigma_N - \Sigma_0\|$ is not greater than the sum of $\|\theta_N(\hat{\beta}_0,\hat{\beta},\hat{t}) - \theta_0(\hat{\beta}_0,\hat{\beta},\hat{t})\|$, $\|\theta_0(\hat{\beta}_0,\hat{\beta},\hat{t}) - \theta_0(\tilde{\beta}_0,\tilde{\beta},\tilde{t})\|$, $\|f_N(\hat{\beta}_0,\hat{\beta},\hat{t})f_N(\hat{\beta}_0,\hat{\beta},\hat{t})^T - f_0(\hat{\beta}_0,\hat{\beta},\hat{t})f_0(\hat{\beta}_0,\hat{\beta},\hat{t})^T\|$ and $\|f_0(\hat{\beta}_0,\hat{\beta},\hat{t})f_0(\hat{\beta}_0,\hat{\beta},\hat{t})^T - f_0(\tilde{\beta}_0,\tilde{\beta},\tilde{t})f_0(\tilde{\beta}_0,\tilde{\beta},\tilde{t})^T\|$, and θ_0 and f_0 are Lipschitz continuous on compact sets under the assumptions, we obtain (2.29) by combining (2.30), (2.31) and (3.31).

□

Estimation of the normal map L_K requires more understanding of its structure. It was shown in [40] that L_K is exactly $d(f_0)_S(z_0)$, the B-derivative of the normal map $(f_0)_S$ at z_0 (recall the definition of B-derivative at the end of Section 1.4). Applying the chain rule of B-differentiability, one has

$$L_K(h) = d(f_0)_S(z_0)(h) = L\ d\Pi_S(z_0)(h) + h - d\Pi_S(z_0)(h) \text{ for each } h \in \mathbb{R}^{2p+1},$$

where $L = df_0(x_0)$ is defined in (2.11), and $d\Pi_S(z_0)$ is the B-derivative of the Euclidean projector Π_S at z_0 and satisfies $d\Pi_S(z_0) = \Pi_K$ [41]. Note that $d\Pi_S(z)$ is not continuous with respect to z at those points z on the boundary of any $(2p+1)$-cell in the normal manifold of S. This results in the discontinuity of $d(f_0)_S(\cdot)$ at these points. If $d(f_0)_S(z)$ is not continuous with respect to z at $z = z_0$, $d(f_N)_S(z_N)$ may not converge to $d(f_0)_S(z_0)$ even though z_N converges to z_0. Consequently, in general we need to find another estimator of L_K instead of $d(f_N)_S(z_N)$. The following subsection provides more details on the estimation of $d\Pi_S(z_0)$ and L_K.

2.3.2 Estimation of the B-derivative $d\Pi_S(z_0)$ and the normal map L_K

In this subsection, we will define two functions Λ_N and Φ_N from $\mathbb{R}^{2p+1} \times \mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$; for each fixed $z \in \mathbb{R}^{2p+1}$, $\Lambda_N(z)$ and $\Phi_N(z)$ are functions from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$. Theorem 2.2 will prove that $\Lambda_N(z_N)$ and $\Phi_N(z_N)$ converge to $d\Pi_S(z_0)$ and L_K respectively. We will then replace the unknown object L_K in (3.30) by the computable object $\Phi_N(z_N)$, to establish a computable formula for confidence regions of z_0.

Before we introduce Λ_N and Φ_N, we have to discusses the computation of $d\Pi_S(z)$, the B-derivative of the projector Π_S at a given point $z \in \mathbb{R}^{2p+1}$. The fact that S is a polyhedron implies that Π_S is piecewise affine, so for each point z the B-derivative $d\Pi_S(z)$ is a piecewise linear function. Moreover, S has a very special structure, in that it is a Cartesian product as shown in (2.6). Consequently, Π_S is a product of individual projectors, and $d\Pi_S(z)$ is the product of B-derivatives of those individual projectors. That is, for each $z = (\beta_0, \beta, t)$ and $h = (\breve{\beta}_0, \breve{\beta}, \breve{t})$ with "cross" ordering, $d\Pi_S(z)(h) = (\breve{\beta}_0, d\Pi_{S_1}(\beta_1, t_1)(\breve{\beta}_1, \breve{t}_1), \cdots, d\Pi_{S_p}(\beta_p, t_p)(\breve{\beta}_p, \breve{t}_p))$, where each S_i is a subset of $\mathbb{R}^2$ defined in (2.7).

To give specific formulas for $d\Pi_{S_i}$ for each $i = 1, \cdots, p$, we need to examine the structure of the normal manifold of S_i. The set S_i is a convex cone in the (β_i, t_i) space, illustrated by the shaded area in Figure 2.1. Its normal manifold consists of 9 cells, which we denote by $C_i^0, \cdots, C_i^8$. Among those cells, C_i^0 is the singleton $\{0\}$, C_i^1, C_i^2, C_i^3,C_i^4 are half rays as illustrated in Figure 2.1, and C_i^5, C_i^6, C_i^7, C_i^8 are convex cones illustrated in the same figure (C_i^5 is just S_i). The left side of Table 2.1 gives the equality/inequality constraints that define each of those cells. The union of all those cells is $\mathbb{R}^2$, and the relative interiors of those cells form a partition of $\mathbb{R}^2$: each point in $\mathbb{R}^2$ lies in the relative interior of exactly one of $C_i^0, \cdots, C_i^8$ (The relative interiors of the 2-cells C_i^5, C_i^6, C_i^7, C_i^8 are exactly their interiors. The relative interiors of the 1-cells C_i^1, C_i^2, C_i^3,C_i^4 are open half rays excluding the origin. The relative interior of C_i^0 is itself).

The Euclidean projector Π_{S_i} is a piecewise linear function that coincides with a linear

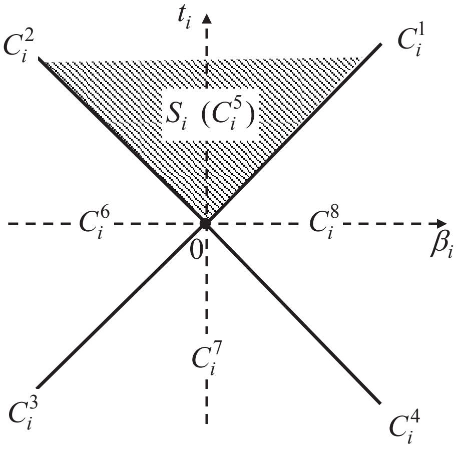


Figure 2.1: The normal manifold of S_i.

Cell	Defining constraints	Critical cone	Defining constraints
C_i^0	$t_i = 0,\ \beta_i = 0$	K_i^0	$t_i - \beta_i \geq 0,\ t_i + \beta_i \geq 0$
C_i^1	$t_i = \beta_i,\ t_i \geq 0$	K_i^1	$t_i - \beta_i \geq 0$
C_i^2	$t_i = -\beta_i,\ t_i \geq 0$	K_i^2	$t_i + \beta_i \geq 0$
C_i^3	$t_i = \beta_i,\ t_i \leq 0$	K_i^3	$t_i = -\beta_i,\ t_i \geq 0$
C_i^4	$t_i = -\beta_i,\ t_i \leq 0$	K_i^4	$t_i = \beta_i,\ t_i \geq 0$
C_i^5	$t_i - \beta_i \geq 0,\ t_i + \beta_i \geq 0$	K_i^5	None
C_i^6	$t_i - \beta_i \geq 0,\ t_i + \beta_i \leq 0$	K_i^6	$t_i = -\beta_i$
C_i^7	$t_i - \beta_i \leq 0,\ t_i + \beta_i \leq 0$	K_i^7	$t_i = 0,\ \beta_i = 0$
C_i^8	$t_i - \beta_i \leq 0,\ t_i + \beta_i \geq 0$	K_i^8	$t_i = \beta_i$

Table 2.1: Cells in the normal manifold of S_i and the associated critical cones

function on each 2-cell C_i^5, C_i^6, C_i^7, C_i^8. More specifically, we define four 2×2 matrices

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ A_2 = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix},\ A_3 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \text{ and } A_4 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

which represent the linear functions coinciding with Π_{S_i} on C_i^5, C_i^6, C_i^8, C_i^7 respectively. On the (relative) interior of each of those 2-cells, the B-derivative $d\Pi_{S_i}(\beta_i, t_i)$ is a linear function. On the relative interior of each 1-cell, the B-derivative $d\Pi_{S_i}(\beta_i, t_i)$ is a piecewise linear function with 2 pieces. The B-derivative $d\Pi_{S_i}(0, 0)$ at the origin is the same as the projector Π_S itself, and is a piecewise linear function with 4 pieces. Note that the B-derivative $d\Pi_{S_i}(\beta_i, t_i)$ at all points (β_i, t_i) in the relative interior of C_i^j for a fixed $j = 0, \cdots, 8$ is the same function, which we denote by ψ_j. Table 2.2 provides the representations of each ψ_j. For example, ψ_1 (the B-derivative $d\Pi_{S_i}(\beta_i, t_i)$ at a point (β_i, t_i) in the relative interior of C_i^1) is a piecewise linear function with two pieces, and it coincides with the linear function A_1 on $C_i^5 \cup C_i^6$ and with A_3 on $C_i^7 \cup C_i^8$.

At each point (β_i, t_i) in $\mathbb{R}^2$, the critical cone to S_i associated with (β_i, t_i) is $T_{S_i}(\Pi_{S_i}(\beta_i, t_i)) \cap \{(\beta_i, t_i) - \Pi_{S_i}(\beta_i, t_i)\}^{\perp}$, which is the same definition for K in (2.15) with S_i and (β_i, t_i) in place of S and z_0. At all points (β_i, t_i) in the relative interior of C_i^j for a fixed $j = 0, \cdots, 8$, the critical cone to S_i associated with (β_i, t_i) is the same, which we denote by K_i^j. The right side of Table 2.1 lists the constraints defining each K_i^j. The cone K_i^j is related with the function ψ_j through the equality $\psi_j = \Pi_{K_i^j}$ [41].

	C_i^5	C_i^6	C_i^7	C_i^8
ψ_0	A_1	A_2	A_4	A_3
ψ_1	A_1	A_1	A_3	A_3
ψ_2	A_1	A_2	A_2	A_1
ψ_3	A_2	A_2	A_4	A_4
ψ_4	A_3	A_4	A_4	A_3
ψ_5	A_1	A_1	A_1	A_1
ψ_6	A_2	A_2	A_2	A_2
ψ_7	A_4	A_4	A_4	A_4
ψ_8	A_3	A_3	A_3	A_3

Table 2.2: Matrix representations of ψ_k for $k = 0, \cdots, 8$

By now we can write down the specific formula for $d\Pi_{S_i}(\beta_i, t_i)$ for each point $(\beta_i, t_i) \in \mathbb{R}^2$: each such point belongs to the relative interior of exactly one cell C_i^j, and $d\Pi_{S_i}(\beta_i, t_i) = \psi_j$. We are ready to give the formula for $d\Pi_S(z)$ for each $z \in \mathbb{R}^{2p+1}$. In view of (2.6), each cell in the normal manifold of S is the product of cells in the normal manifolds of the individual sets, i.e., of the form $\mathbb{R} \times \Pi_{i=1}^p C_i^{\gamma(i)}$, where $\gamma(i) = 0, \cdots, 8$ for each $i = 1, \cdots, p$. For each z in the relative interior of one cell $\mathbb{R} \times \Pi_{i=1}^p C_i^{\gamma(i)}$, $d\Pi_S(z)$ is the same function. We denote the latter function as Ψ_γ, which is a function from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$ and is given by

$$\Psi_\gamma(h) = (\breve{\beta}_0, \psi_{\gamma(1)}(\breve{\beta}_1, \breve{t}_1), \cdots, \psi_{\gamma(p)}(\breve{\beta}_p, \breve{t}_p)) \text{ for each } h = (\breve{\beta}_0, \breve{\beta}, \breve{t}). \tag{2.32}$$

If we use $K(\gamma) = \mathbb{R} \times \Pi_{i=1}^p K_i^{\gamma(i)}$ to denote the critical cone to S associated with a point in the relative interior of $\mathbb{R} \times \Pi_{i=1}^{p} C_i^{\gamma(i)}$, then $\Psi_\gamma(h) = \Pi_{K(\gamma)}(h)$ for each $h \in \mathbb{R}^{2p+1}$.

For technical reasons, we define a function g from the set of integers to $\mathbb{R}$. The function g can be any linear combination of finite many terms of the form aN^b with $a > 0$ and $b \in (0, 1/2)$. Other choices are also possible; for more details see [32]. Among other requirements, the function g needs to satisfy $g(N) \to \infty$ as $N \to \infty$.

Next, we equip the (β_0, β, t) space with a norm, which will be used to compute distances between points in $\mathbb{R}^{2p+1}$ and cells in the normal manifold of S. Theoretically, this can be any norm. For convenience of computation, we use in each individual (β_i, t_i) space the norm, $\|(\beta_i, t_i)\| = \max(|\beta_i + t_i|, |\beta_i - t_i|)$, and use the norm, $\|(\beta_0, \beta, t)\| = \max(|\beta_0|, \max_{i=1,\cdots,p}\{\|(\beta_i, t_i)\|\})$ in the overall (β_0, β, t) space. Table 2.3 provides formulas on distances between a point (β_i, t_i) and each cell in the normal manifold of S_i. The distance between $z = (\beta_0, \beta, t)$ and $\mathbb{R} \times \Pi_{i=1}^{p} C_i^{\gamma(i)}$, a cell in the normal manifold of S, is

$$d(z, \mathbb{R} \times \Pi_{i=1}^{p} C_i^{\gamma(i)}) = \max_{i=1,\cdots,p} d\left((\beta_i, t_i), C_i^{\gamma(i)}\right).$$

Cell	Distance from (β_i, t_i)
C_i^0	$\max(\lvert\beta_i - t_i\rvert, \lvert\beta_i + t_i\rvert)$
C_i^1	$\max(-(\beta_i + t_i), \lvert\beta_i - t_i\rvert)$
C_i^2	$\max(\beta_i - t_i, \lvert\beta_i + t_i\rvert)$
C_i^3	$\max(\beta_i + t_i, \lvert\beta_i - t_i\rvert)$
C_i^4	$\max(-(\beta_i - t_i), \lvert\beta_i + t_i\rvert)$
C_i^5	$\max(-(\beta_i + t_i), \beta_i - t_i, 0)$
C_i^6	$\max(\beta_i + t_i, \beta_i - t_i, 0)$
C_i^7	$\max(\beta_i + t_i, -(\beta_i - t_i), 0)$
C_i^8	$\max(-(\beta_i + t_i), -(\beta_i - t_i), 0)$

Table 2.3: Distances between (β_i, t_i) and cells in the normal manifold of S_i

Let N be a given integer. For each $z \in \mathbb{R}^{2p+1}$, find a cell in the normal manifold of S, that has the smallest dimension among all cells whose distances from z are no more than $1/g(N)$. Let this cell be denoted as $\mathbb{R} \times \Pi_{i=1}^{p} C_i^{\gamma(i)}$. Define the function $\Lambda_N(z) : \mathbb{R}^{2p+1} \to \mathbb{R}^{2p+1}$ by

$$\Lambda_N(z)(h) = \Psi_\gamma(h) \quad \text{for each} \quad h \in \mathbb{R}^{2p+1}, \tag{2.33}$$

where Ψ_γ is defined in (2.32). In other words, $\Lambda_N(z)$ is defined to be the B-derivative $d\Pi_S(z')$ for a point z' that belongs to the relative interior of a cell that has the smallest dimension among all cells whose distances from z are no more than $1/g(N)$. When z' lies in the interior of a full-dimensional cell, $d\Pi_S(z')$ is a linear map from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$; otherwise it is a piecewise linear map.

Next, define the function $\Phi_N : \mathbb{R}^{2p+1} \times \mathbb{R}^{2p+1} \to \mathbb{R}^{2p+1}$ as

$$\Phi_N(z)(h) = L_N \; \Lambda_N(z)(h) + h - \Lambda_N(z)(h) \tag{2.34}$$

for each $z \in \mathbb{R}^{2p+1}$ and $h \in \mathbb{R}^{2p+1}$, where L_N is defined in (2.21). For a given N, Λ_N is a fixed function, while Φ_N depends on sample data since L_N does. If $\Lambda_N(z)$ is a linear map from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$, then $\Phi_N(z)$ is a linear map as well; otherwise it is a piecewise linear map.

Theorem 2.2 below shows that $\Lambda_N(z_N)$ and $\Phi_N(z_N)$ are asymptotically exact estimators of $d\Pi_S(z_0)$ and L_K respectively.

Theorem 2.2. *Suppose that Assumptions 2.1, 2.2 and 2.4(a-b) hold. Then*

$$\lim_{N\to\infty} \operatorname{Prob}\left[\Lambda_N(z_N)(h) = d\Pi_S(z_0)(h) \textit{ for all } h \in \mathbb{R}^{2p+1}\right] = 1,$$

and there exists a positive real number ϕ such that

$$\lim_{N\to\infty} \operatorname{Prob}\left[\sup_{h\in\mathbb{R}^{2p+1}} \frac{\|\Phi_N(z_N)(h) - L_K(h)\|}{\|h\|} < \frac{\phi}{g(N)}\right] = 1. \tag{2.35}$$

Proof of Theorem 2.2. The conclusions follow from Theorem 2.1 and Corollary 3.2 of [31].

□

We can now replace the normal map L_K in (3.30) by $\Phi_N(z_N)$, without changing the weak convergence, see Theorem 2.3 below.

Theorem 2.3. *Suppose that Assumptions 2.1, 2.2 and 2.4(a-b) hold. Then*

$$\sqrt{N}\Phi_N(z_N)(z_N - z_0) \Rightarrow \mathcal{N}(0, \Sigma_0). \tag{2.36}$$

If Assumption 2.3 holds, then

$$\sqrt{N}\begin{bmatrix}(\Sigma_N^1)^{-1/2} & 0\\ 0 & I_p\end{bmatrix}(\Phi_N(z_N))(z_N - z_0) \Rightarrow \mathcal{N}(0, I_{p+1}) \times 0. \tag{2.37}$$

Otherwise, if Assumption 2.4(c) holds, then let l be the number of positive eigenvalues of Σ_0^1 counted with regard to their algebraic multiplicities, and decompose Σ_N^1 as

$$\Sigma_N^1 = U_N^T \Delta_N U_N, \tag{2.38}$$

where U_N is an orthogonal $(p+1)\times(p+1)$ matrix, and Δ_N is a diagonal matrix with monotonically decreasing elements. Let D_N be the upper-left submatrix of Δ_N whose diagonal elements are at least $1/g(N)$. Let l_N be the number of rows in D_N, and let $(U_N)_1$ be the submatrix of U_N that consists of its first l_N rows, and let $(U_N)_2$ consist of the remaining rows of U_N. Then $\text{Prob}\{l_N = l\} \to 1$ *as $N \to \infty$, and*

$$N\big[(\Phi_N(z_N))(z_N - z_0)\big]^T \begin{bmatrix}(U_N)_1^T D_N^{-1}(U_N)_1 & 0\\ 0 & 0\end{bmatrix}\big[(\Phi_N(z_N))(z_N - z_0)\big] \Rightarrow \chi_l^2, \tag{2.39}$$

and

$$N\big[(\Phi_N(z_N))(z_N - z_0)\big]^T \begin{bmatrix}(U_N)_2^T (U_N)_2 & 0\\ 0 & I_p\end{bmatrix}\big[(\Phi_N(z_N))(z_N - z_0)\big] \Rightarrow 0. \tag{2.40}$$

Proof of Theorem 2.3. Equation (3.37) follows from [31, Corollary 3.3]. If Σ_0^1 is nonsingular, then (3.38) follows from the fact that Σ_N^1 converges to Σ_0^1 almost surely, as shown in Lemma 2.3.

Now suppose Σ_0^1 is singular and Assumption 2.4(c) holds. Decompose Σ_0^1 as

$$\Sigma_0^1 = U_0^T \begin{bmatrix} D_0 & 0\\ 0 & 0\end{bmatrix} U_0 \tag{2.41}$$

where U_0 is an orthogonal $(p+1)\times(p+1)$ matrix and D_0 is a diagonal $l \times l$ matrix with strictly positive, monotonically decreasing diagonal elements. Let $(U_0)_1$ be the submatrix of U_0 that consists of its first l rows, and let $(U_0)_2$ consist of the remaining rows of U_0. From (3.37) and (2.41) we have

$$\sqrt{N}\begin{bmatrix}\begin{bmatrix}D_0^{-1/2} & 0\\ 0 & I_{p+1-l}\end{bmatrix}U_0 & 0\\ 0 & I_p\end{bmatrix}(\Phi_N(z_N))(z_N-z_0)\Rightarrow\mathcal{N}(0,I_l)\times 0,$$

which implies

$$N\big[(\Phi_N(z_N))(z_N-z_0)\big]^T\begin{bmatrix}(U_0)_1^T D_0^{-1}(U_0)_1 & 0\\ 0 & 0\end{bmatrix}\big[(\Phi_N(z_N))(z_N-z_0)\big]\Rightarrow\chi_l^2, \tag{2.42}$$

and

$$N\big[(\Phi_N(z_N))(z_N-z_0)\big]^T\begin{bmatrix}(U_0)_2^T(U_0)_2 & 0\\ 0 & I_p\end{bmatrix}\big[(\Phi_N(z_N))(z_N-z_0)\big]\Rightarrow 0. \tag{2.43}$$

According to Lemma 2.3, there exist positive real numbers δ_1, μ_1, M_1 and σ_1, such that (2.29) holds for each $\epsilon>0$ and each N. It follows from the Lipschtiz continuity of eigenvalues that there exist positive numbers δ_2, μ_2, M_2 and σ_2 such that the following holds for each $\epsilon>0$ and each N :

$$\begin{aligned}&\text{Prob}\left\{\left\|\Delta_N-\begin{bmatrix}D_0 & 0\\ 0 & 0\end{bmatrix}\right\|<\epsilon\right\}\\ \geq&1-\delta_2\exp\{-N\mu_2\}-\frac{M_2}{\min(\epsilon^{(2p+1)^2},\epsilon^{2p+1})}\exp\left\{-\frac{N\epsilon^2}{\sigma_2}\right\}.\end{aligned}$$

Denote the right hand side of the above inequality by $\eta_N(\epsilon)$, and let r be the smallest diagonal element in D_0. For N large enough to satisfy $g(N)\geq 2/r$, we have

$$\text{Prob}\left\{(\Delta_N)_{ii}>\frac{1}{g(N)}\text{ for all }i=1,\cdots,l\right\}\geq\text{Prob}\left\{(\Delta_N)_{ii}>\frac{r}{2}\text{ for all }i=1,\cdots,l\right\}\geq\eta_N(r/2)$$

On the other hand for each such N we have

$$\begin{aligned}&\text{Prob}\left\{(\Delta_N)_{ii} < \frac{r}{2} \text{ for all } i = l+1, \cdots, p+1\right\}\\ \geq\, &\text{Prob}\left\{(\Delta_N)_{ii} < \frac{1}{g(N)} \text{ for all } i = l+1, \cdots, p+1\right\} \geq \eta_N\left(\frac{1}{g(N)}\right).\end{aligned}$$

Thus, for large N, the equality $l = l_N$ holds with probability at least $\eta_N(r/2)+\eta_N(1/g(N))-1$, which converges to 1 as $N \to \infty$. It follows that D_N converges to D_0 in probability.

Let $\mathcal{S}$ be the family of $(p+1)\times(p+1)$ matrices A, such that A is symmetric and positive semi-definite, with its largest l eigenvalues strictly larger than $r/2$ and its remaining eigenvalues strictly smaller than $r/2$. Each matrix $A \in S$ has a unique approximation $\hat{A}$ in Frobenius norm of rank no more than l, and the rank of $\hat{A}$ is exactly l. Let $W(A)$ be the pseudo-inverse of $\hat{A}$. Note that W is a continuous function on the set $\mathcal{S}$.

Note that Σ_0^1 belongs to $\mathcal{S}$ with $W(\Sigma_0^1) = (U_0)_1^T D_0^{-1}(U_0)_1$. On the other hand, the probability for Σ_N^1 to belong to $\mathcal{S}$ converges to 1 as $N \to \infty$, and the probability for the equality $(U_N)_1^T D_N^{-1}(U_N)_1 = W(\Sigma_N^1)$ to hold also converges to 1 as $N \to \infty$. Since Σ_N^1 converges to Σ_0^1 almost surely, $(U_N)_1^T D_N^{-1}(U_N)_1$ converges to $(U_0)_1^T D_0^{-1}(U_0)_1$ in probability. This and (2.42) implies (3.40).

To prove (3.41), conduct a spectral decomposition $A = V^T\Lambda V$ for each matrix $A \in \mathcal{S}$, with V being orthogonal and Λ being a diagonal matrix with monotonically decreasing diagonal elements. Let V_2 be the submatrix of V that consists of its last $p+1-l$ rows, and let $H(A) = V_2^T V_2$. The function H is continuous on $\mathcal{S}$. Consequently, the matrix $(U_N)_2^T(U_N)_2$ converges to $(U_0)_2^T(U_0)_2$ in probability. This together with (2.43) implies (3.41).

□

The above theorem deals with two cases separately, depending on whether Σ_0^1 is nonsingular or not. In practice, since Σ_0^1 is unknown, we will always start by decomposing Σ_N^1 as in (2.38). If some eigenvalues of Σ_N^1 (i.e., diagonal elements of Δ_N) are less than $1/g(N)$, then D_N is a proper submatrix of Δ_N, and we will use (3.40) and (3.41) to establish confidence intervals for z_0 (more details will be given in the following subsections). Otherwise, if all eigenvalues of Σ_N^1 are greater than or equal to $1/g(N)$, then D_N equals Δ_N and (3.40) and (3.41) are equivalent to (3.38).

2.3.3 Confidence intervals for the normal map solutions

In this subsection, we discuss how to obtain individual and simultaneous confidence intervals for z_0, based on results in Theorem 2.3.

The computation of individual confidence intervals is based on (3.37). Recall from Lemma 2.2 that the normal map L_K is a global homeomorphism under our assumptions. As a piecewise linear function, both L_K itself and its inverse function are globally Lipschitz continuous. We can apply [41, Lemma 3.1] to conclude from (2.35) that the probability for $\Phi_N(z_N)$ to be a global homeomorphism converges to 1 as $N \to \infty$. If $\Phi_N(z_N)$ is a global homeomorphism, we can then use

$$(\Phi_N(z_N))^{-1}(\mathcal{N}(0, \Sigma_N)) \tag{2.44}$$

to approximate the distribution of $\sqrt{N}(z_N - z_0)$. We discuss how to construct individual confidence intervals from (2.44) depending on whether $\Phi_N(z_N)$ is linear or not.

When $\Phi_N(z_N)$ is a linear map from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$, the distribution in (2.44) is normal. In such situations, we let m_i be the ith diagonal element of the matrix $(\Phi_N(z_N))^{-1}\Sigma_N(\Phi_N(z_N))^{-T}$, and use

$$\left[\ (z_N)_i - \sqrt{\frac{\chi_1^2(\alpha) m_i}{N}},\ (z_N)_i + \sqrt{\frac{\chi_1^2(\alpha) m_i}{N}}\ \right]$$

as an approximate $(1-\alpha)100\%$ individual confidence interval for $(z_0)_i$. Here and in what follows, $\chi_n^2(\alpha)$ is the number that satisfies $P(U > \chi_1^2(\alpha)) = \alpha$ for a χ^2 random variable U with n degrees of freedom.

When $\Phi_N(z_N)$ is not a linear map, we simulate data based on the distribution in (2.44), and find individual confidence intervals by ordering the data by each component and finding bounds on each component that cover a specified percentage of data points. To simulate the distribution of (2.44), let $\mathbb{R} \times \Pi_{i=1}^p C_i^{\gamma(i)}$ be the cell that is used to define $\Lambda_N(z_N)$; it follows that

$$\Lambda_N(z_N) = \Psi_\gamma = \Pi_{K(\gamma)},$$

where $K(\gamma)$ is defined below (2.32). From (3.35) it can be seen that $\Phi_N(z_N)$ is exactly the normal map induced by L_N and $K(\gamma)$. To find $(\Phi_N(z_N))^{-1}(q)$ for a given $q \in \mathbb{R}^{2p+1}$, we first find the vector h that solves the following optimization problem

$$\min_{h \in K(\gamma)} \frac{1}{2} h^T L_N h - q^T h,$$

and then let $(\Phi_N(z_N))^{-1}(q) = h - L_N(h) + q$.

Below we discuss the computation of simultaneous confidence intervals for all components of $(z_0)_i$. From (3.40), the set of $z \in \mathbb{R}^{2p+1}$ satisfying the following constraints

$$\begin{aligned} N\big[\Phi_N(z_N)(z_N - z)\big]^T \begin{bmatrix} (U_N)_1^T D_N^{-1} (U_N)_1 & 0 \\ 0 & 0 \end{bmatrix} \big[\Phi_N(z_N)(z_N - z)\big] &\leq \chi^2_{l_N}(\alpha) \\ \begin{bmatrix} (U_N)_2 & 0 \\ 0 & I_p \end{bmatrix} \big[\Phi_N(z_N)(z_N - z)\big] &= 0 \end{aligned}$$

is an approximate $(1-\alpha)100\%$ confidence region for z_0. The set is an ellipsoid in a subspace of $\mathbb{R}^{2p+1}$, if $\phi_N(z_N)$ is linear. Otherwise it is the union of fractions of different ellipsoids.

To obtain simultaneous confidence intervals, we find the maximal and minimal values of z_i under the above constraints, for each $i = 1, \cdots, 2p+1$. When $\phi_N(z_N)$ is a piecewise linear function with multiple pieces, we treat each of its pieces separately, and then combine the results.

2.3.4 Confidence intervals for LASSO parameters

Having computed confidence intervals for z_0, we transform them into confidence intervals for the population LASSO parameters $(\tilde{\beta}_0, \tilde{\beta})$.

Let $\tilde{q}$ be as defined in Assumption 2.2, and let $\tilde{q}_0 = E[-2(Y - \tilde{\beta}_0 - \sum_{j=1}^p \tilde{\beta}_j X_j)]$. By the definitions of (2.8) and (2.10), we have $f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = (\tilde{q}_0, \tilde{q}, \lambda e_p)$. It follows from (2.14) that $z_0 = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) - (\tilde{q}_0, \tilde{q}, \lambda e_p)$. Since $-f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \in N_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, we know that $\tilde{q}_0 = 0$ which gives $\tilde{\beta}_0 = (z_0)_1$. Thus, confidence intervals of $(z_0)_1$ are exactly those of $\tilde{\beta}_0$.

From the definition of S_i and the fact $-(\tilde{q}_i, \lambda) \in N_{S_i}(\tilde{\beta}_i, \tilde{t}_i)$, we have for each $i = 1, \cdots, p$

$$-\lambda \leq \tilde{q}_i \leq \lambda \text{ if } \tilde{\beta}_i = 0, \quad \tilde{q}_i = -\lambda \text{ if } \tilde{\beta}_i > 0, \text{ and } \tilde{q}_i = \lambda \text{ if } \tilde{\beta}_i < 0. \tag{2.45}$$

This relation between $\tilde{\beta}_i$ and $\tilde{q}_i$ with the fact that $(z_0)_{i+1} = \tilde{\beta}_i - \tilde{q}_i$ imply the following equality for each $i = 1, \cdots, p$:

$$\tilde{\beta}_i = \begin{cases} (z_0)_{i+1} - \lambda & \text{if } (z_0)_{i+1} > \lambda, \\ 0 & \text{if } (z_0)_{i+1} \in [-\lambda, \lambda], \\ (z_0)_{i+1} + \lambda & \text{if } (z_0)_{i+1} < -\lambda. \end{cases} \tag{2.46}$$

Let us denote the right hand side of (3.42) as $\Gamma((z_0)_{i+1})$, which is a nondecreasing piecewise linear function of $(z_0)_{i+1}$. We can then use images of confidence intervals of $(z_0)_{i+1}$ under the map Γ as confidence intervals of $\tilde{\beta}_i$. Because $\Gamma(\cdot)$ takes the constant value of 0 on $[-\lambda, \lambda]$, the confidence interval for $\tilde{\beta}_i$ computed from this method will contain the true solution of (2.1) with a probability larger than the prescribed level, when the confidence interval for $(z_0)_{i+1}$ meets a part of the interval $[-\lambda, \lambda]$.

2.4 Confidence intervals for the population LASSO parameters with varying λ

We have introduced in Section 2.3 on how to construct confidence intervals for z_0 and $(\tilde{\beta}_0, \tilde{\beta})$ when λ is fixed. In this section, we study properties of confidence intervals for z_0 as λ varies with fixed sample size N. Section 2.4.1 provides a condition to ensure the confidence intervals to be computationally tractable. Following that, Sections 2.4.2 and 2.4.3 discuss properties of these confidence intervals, and show that their dependence on λ is Lipschitz continuous when λ is restricted on certain intervals. In Section 2.4.4, we propose algorithms to track the confidence bands for $(\tilde{\beta}_0, \tilde{\beta})$ along the LASSO solution path.

2.4.1 Properties of z_N and $\Phi_N(z_N)$

Recall that each cell in the normal manifold of S has the form $\mathbb{R} \times \Pi_{i=1}^{p} C_i^{\gamma(i)}$, where $\gamma(i) = 0, 1, \cdots, 8$ for each $i = 1, 2, \cdots, p$. We divide the plane (β_i, t_i) into 9 pieces $E_i^0, \cdots, E_i^8$, as illustrated in Figure 2.2. Table 2.4 lists the constraints that define each of the sets $E_i^0, \cdots, E_i^8$.

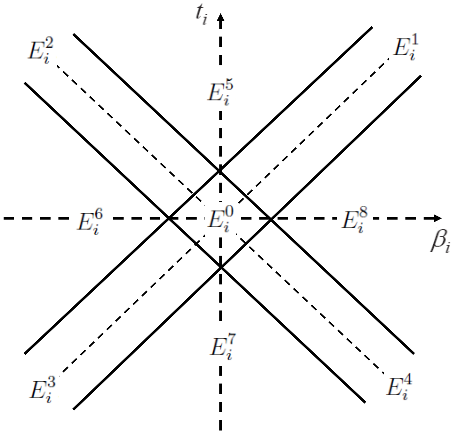


Figure 2.2: $E_i^0, \cdots, E_i^8$ in the plane (β_i, t_i)

Piece	Defining constraints
E_i^0	$\lvert t_i - \beta_i\rvert \leqslant 1/g(N), \quad \lvert t_i + \beta_i\rvert \leqslant 1/g(N)$
E_i^1	$\lvert t_i - \beta_i\rvert \leqslant 1/g(N), \quad t_i + \beta_i > 1/g(N)$
E_i^2	$t_i - \beta_i > 1/g(N), \quad \lvert t_i + \beta_i\rvert \leqslant 1/g(N)$
E_i^3	$\lvert t_i - \beta_i\rvert \leqslant 1/g(N), \quad t_i + \beta_i < -1/g(N)$
E_i^4	$t_i - \beta_i < -1/g(N), \quad \lvert t_i + \beta_i\rvert \leqslant 1/g(N)$
E_i^5	$t_i - \beta_i > 1/g(N), \quad t_i + \beta_i > 1/g(N)$
E_i^6	$t_i - \beta_i > 1/g(N), \quad t_i + \beta_i < -1/g(N)$
E_i^7	$t_i - \beta_i < -1/g(N), \quad t_i + \beta_i < -1/g(N)$
E_i^8	$t_i - \beta_i < -1/g(N), \quad t_i + \beta_i > 1/g(N)$

Table 2.4: $E_i^0, \cdots, E_i^8$ in the plane (β_i, t_i)

Each partition $\mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)}$ is associated with the cell $\mathbb{R} \times \Pi_{i=1}^p C_i^{\gamma(i)}$. Let

$$\boldsymbol{\gamma}(z_N) \triangleq \big(\gamma(1), \cdots, \gamma(p)\big) \text{ such that } z_N \in \mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)}.$$

This $\boldsymbol{\gamma}(z_N)$ identifies the partition that contains z_N. In view of the definition of Λ_N, if z_N is in the partition $\mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)}$, then

$$\Lambda_N(z_N)(h) = \Psi_\gamma(h) \text{ for each } h \in \mathbb{R}^{2p+1}.$$

The "cross" ordered $\Lambda_N(z_N)$ has following representation

$$\Lambda_N(z_N) = \begin{bmatrix} 1 & & & & \\ & \psi_{\gamma(1)} & & & \\ & & \psi_{\gamma(2)} & & \\ & & & \ddots & \\ & & & & \psi_{\gamma(p)} \end{bmatrix},$$

where each $\psi_{\gamma(i)}$, $i = 1, \cdots, p$ is a (piecewise) linear map defined in Table 2.2. We write $\Phi_N(z_N)$ as

$$\Phi_N(z_N) = \begin{bmatrix} \Phi_1 & \Phi_2 \\ \Phi_3 & \Phi_4 \end{bmatrix},$$

where Φ_1 and Φ_4 are respectively functions from $\mathbb{R}^{p+1}$ to $\mathbb{R}^{p+1}$ and from $\mathbb{R}^p$ to $\mathbb{R}^p$.

Lemma 2.4. *Suppose that we have chosen $g(N)$ s.t.*

$$\frac{1}{g(N)} < \lambda, \tag{2.47}$$

z_N is a solution of (2.22). Then for any $i \in \{1, 2, \cdots, p\}$, the components $((z_N)_{i+1}, (z_N)_{i+1+p})$ can only be in E_i^3, E_i^4, E_i^6, E_i^7, or E_i^8. Furthermore, the matrix representations of Φ_4 are all diagonally invertible matrices.

Proof of Lemma 3.1. We know that the SAA optimal solution $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ always lies on the boundary of S, i.e. $\hat{t}_i = |\hat{\beta}_i|$ for any $i = 1, \cdots, p$. From (2.28) and the fact that $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ is the projection of z_N on S, we have

$$(z_N)_{i+1+p} = \hat{t}_i - \lambda,$$

$$(z_N)_{i+1} = \begin{cases} \hat{\beta}_i + \lambda & \text{if } \hat{\beta}_i > 0, \\ \tau_i \in [-\lambda, \lambda] & \text{if } \hat{\beta}_i = 0, \\ \hat{\beta}_i - \lambda & \text{if } \hat{\beta}_i < 0, \end{cases}$$

for each $i = 1, \cdots, p$. According to the defining constraints in Table 1 and (2.47), one can see that $((z_N)_{i+1}, (z_N)_{i+1+p})$ can only be in E_i^3, E_i^4, E_i^6, E_i^7, or E_i^8. This means that $\psi_{\gamma(i)}$ never coincides with A_1, the 2×2 identity matrix.

One can check that $\Lambda_N(z_N)$ has the following form

$$\left[\begin{array}{c|c|c} 1 & 0 & 0 \\ \cline{2-3} 0 & W_1 & W_2 \\ \cline{2-3} 0 & W_3 & W_4 \end{array}\right], \tag{2.48}$$

in which each W_j is a (piecewise) linear function represented by $p \times p$ diagonal matrices. Moreover, we know that

$$\Phi_N(z_N) = L_N \, \Lambda_N(z_N) + I - \Lambda_N(z_N).$$

From (2.21) and (2.48), we obtain

$$\Phi_4 = I_{p \times p} - W_4.$$

Because $\psi_{\gamma(i)}$ never coincides with A_1, the matrix representations of W_4 have diagonal elements 0 or $\frac{1}{2}$. Consequently, all the matrix representations of Φ_4 are diagonally invertible matrices.

□

Lemma 2.5. *Suppose z_N is a solution of (2.22). Let*

$$\mathfrak{L} = \Big\{ i \in \{1, \cdots, p\} \ \big| \ ((z_N)_{i+1}, (z_N)_{i+1+p}) \in E_i^3, E_i^4, E_i^6 \text{ or } E_i^8 \Big\},$$

and $(L_N)_{\mathfrak{L}}$ be the submatrix of L_N that consists of columns and rows of L_N with indices in $\{1\} \cup \{i+1, i \in \mathfrak{L}\}$. Then the following two statements are equivalent.

1. *$\Phi_N(z_N)$ is a global homeomorphism.*

2. *$(L_N)_{\mathfrak{L}}$ is nonsingular, and $((z_N)_{i+1}, (z_N)_{i+1+p}) \notin E_i^0, E_i^1$ or E_i^2 for all $i \in \{1, \cdots, p\}$.*

Proof of Lemma 3.2. For any $i = 1, \cdots, p$ and $j = 0, \cdots, 8$, the critical cone to S_i associated with any point (β_i, t_i) is the same, which are denoted by K_i^j and defined in Table 2. Let

$K(\boldsymbol{\gamma}(z_N)) = \mathbb{R} \times \Pi_{i=1}^{p} K_i^{\gamma(i)}$, then one can see that

$$\Psi_{\boldsymbol{\gamma}(z_N)}(h) = \Pi_{K(\boldsymbol{\gamma}(z_N))}(h) \quad \text{for each } h \in \mathbb{R}^{2p+1}.$$

From (3.35) and (3.34), $\Phi_N(z_N)$ is a normal map induced by linear function L_N and critical cone $K(\boldsymbol{\gamma}(z_N))$.

Next, we give an explicit expression for the affine hull of $K(\boldsymbol{\gamma}(z_N))$. Define two matrices Q_1 and Q_2 as

$$Q_1 = \begin{bmatrix} 1 & 0 \\ 0 & I_p \\ 0 & I_p \end{bmatrix} \text{ and } Q_2 = \begin{bmatrix} 1 & 0 \\ 0 & I_p \\ 0 & -I_p \end{bmatrix},$$

where I_p is the p dimensional identity matrix. Construct a matrix Ξ by adding columns from Q_1, Q_2 or I_{2p+1} according to the following rule: At first add the first column of Q_1, then for $i = 1, \cdots, p$, add $(i+1)$'th column of Q_1 if $((z_N)_{i+1}, (z_N)_{i+1+p}) \in E_i^4$ or E_i^8, or add $(i+1)$'th column of Q_2 if $((z_N)_{i+1}, (z_N)_{i+1+p}) \in E_i^3$ or E_i^6, or add $(i+1)$'th and $(i+1+p)$'th columns of I_{2p+1} if $((z_N)_{i+1}, (z_N)_{i+1+p}) \in E_i^0, E_i^1$ or E_i^2. Note that z_N can not be in E_i^5 and $K_i^7 = \{(0,0)\}$, so columns of Ξ form a basis of the affine hull of $K(\boldsymbol{\gamma}(z_N))$. From Proposition 2.5 and Theorem 4.3 of [39], we find that $\Phi_N(z_N)$ is a global homeomorphism if and only if $\Xi^T L_N \Xi$ is nonsingular. We prove the latter statement is equivalent to the statement 2 in the Lemma.

If $\Xi^T L_N \Xi$ is nonsingular, one can check that $((z_N)_{i+1}, (z_N)_{i+1+p}) \notin E_i^0, E_i^1$ or E_i^2 for all i. Furthermore, in this case $\Xi^T L_N \Xi = (L_N)_{\mathfrak{L}}$. So $(L_N)_{\mathfrak{L}}$ is nonsingular.

On the other hand, if statement 2 is true, then one can check that $\Xi^T L_N \Xi = (L_N)_{\mathfrak{L}}$. Thus $\Xi^T L_N \Xi$ is nonsingular.

□

Lemma 3.2 gives a sufficient and necessary condition for checking if $\Phi_N(z_N)$ is a global homeomorphism in a given SAA problem with fixed N and λ. It shows that it is possible for $\Phi_N(z_N)$ to be a global homeomorphism even when $N < p$. Combining Lemmas 3.1 and 3.2, we obtain a sufficient condition which guarantees the function $\Phi_N(z_N)$ to be homeomorphism for a SAA problem.

Corollary 2.1. *Suppose z_N is a solution of (2.22). For a specific SAA problem, $\Phi_N(z_N)$ is a global homeomorphism if the following condition holds:*

$(L_N)_{\mathfrak{L}}$ is nonsingular and $g(N)$ is chosen to satisfy (2.47).

We assume this condition holds for the SAA problem (2.2) in the rest of Section 2.4.

2.4.2 Properties of individual confidence bands for z_0

If $\Phi_N(z_N)$ is an invertible linear map, then $\sqrt{N}(z_N - z_0)$ asymptotically follows the distribution of $(\Phi_N(z_N))^{-1}Y_N$, where $Y_N \sim \mathcal{N}(0, \Sigma_N)$. Thus, each component of $z_N - z_0$ approximately follows a normal distribution, and we can give an explicit expression for the approximate individual confidence interval for each component of z_0.

Define

$$\tilde{X} \triangleq -2\breve{\mathbf{X}}^T \begin{bmatrix} y_1 - \hat{\beta}_0 - \mathbf{x}^1\hat{\beta} & & \\ & \ddots & \\ & & y_N - \hat{\beta}_0 - \mathbf{x}^N\hat{\beta} \end{bmatrix},$$

then we have

$$\Sigma_N^1 = \frac{1}{N-1}\tilde{X}H\tilde{X}^T = \frac{4}{N-1}\breve{\mathbf{X}}^T\breve{H}\breve{\mathbf{X}},$$

where

$$H = I - \frac{1}{N}1_N 1_N^T,$$

and

$$\breve{H} = \begin{bmatrix} y_1 - \hat{\beta}_0 - \mathbf{x}^1\hat{\beta} & & \\ & \ddots & \\ & & y_N - \hat{\beta}_0 - \mathbf{x}^N\hat{\beta} \end{bmatrix} H \begin{bmatrix} y_1 - \hat{\beta}_0 - \mathbf{x}^1\hat{\beta} & & \\ & \ddots & \\ & & y_N - \hat{\beta}_0 - \mathbf{x}^N\hat{\beta} \end{bmatrix}.$$

We write $\Phi_N(z_N)^{-1}$ as

$$\Phi_N(z_N)^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

in which A_{11} and A_{22} are square matrices with dimensions $p+1$ and p respectively. Then the sample covariance matrix of $(\Phi_N(z_N))^{-1}Y_N$ is

$$\begin{aligned}\widehat{\mathrm{Var}}\big[(\Phi_N(z_N))^{-1}Y_N\big] &= \big(\Phi_N(z_N)\big)^{-1}\begin{bmatrix}\Sigma_N^1 & 0\\ 0 & 0\end{bmatrix}\big(\Phi_N(z_N)\big)^{-T}\\ &= \frac{4}{N-1}\begin{bmatrix}A_{11}\breve{\mathbf{X}}^T\breve{H}\breve{\mathbf{X}}A_{11}^T & A_{11}\breve{\mathbf{X}}^T\breve{H}\breve{\mathbf{X}}A_{21}^T\\ A_{21}\breve{\mathbf{X}}^T\breve{H}\breve{\mathbf{X}}A_{11}^T & A_{21}\breve{\mathbf{X}}^T\breve{H}\breve{\mathbf{X}}A_{21}^T\end{bmatrix}.\end{aligned}$$

For convenience, we introduce a matrix

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1N}\\ b_{21} & b_{22} & \cdots & b_{2N}\\ \vdots & \vdots & \ddots & \vdots\\ b_{(p+1)1} & b_{(p+1)2} & \cdots & b_{(p+1)N}\end{bmatrix} \triangleq A_{11}\breve{\mathbf{X}}^T,$$

which is a $(p+1)\times N$ constant matrix. After defining

$$c_{ik} \triangleq (y_k - \hat{\beta}_0\mathbf{x}^k\hat{\beta})b_{ik}, \text{ for each } k = 1, \cdots, N, \tag{2.49}$$

we can explicitly express the ith $(i = 1, \cdots, p+1)$ diagonal entry of $B\breve{H}B^T$ as

$$diag(i) \triangleq \sum_{k=1}^{N} c_{ik}^2 - \frac{1}{N}(\sum_{k=1}^{N} c_{ik})^2. \tag{2.50}$$

Since each component of $z_N - z_0$ asymptotically follows a normal distribution, the approximate $(1-\alpha)100\%$ individual confidence interval for the ith component of z_0 $(i = 1, \cdots, p+1)$ is

$$\left[(z_N)_i - 2\sqrt{\frac{\chi_1^2(\alpha)diag(i)}{N(N-1)}},\ (z_N)_i + 2\sqrt{\frac{\chi_1^2(\alpha)diag(i)}{N(N-1)}}\right], \tag{2.51}$$

where α is the significance level. From (3.42), this interval is also an approximate $(1-\alpha)100\%$ confidence interval for β_{i-1}.

The following theorem gives properties of such confidence intervals.

Theorem 2.4. *With a fixed sample size N, suppose $\Phi_N(z_N)$ is an invertible linear map and the SAA solution $(\hat{\beta}_0, \hat{\beta})$ is a linear function of λ on an interval $[\lambda_1, \lambda_2]$. Then the square of the width of the interval (2.51) is a quadratic function of λ on $[\lambda_1, \lambda_2]$, and the endpoints of this interval are Lipschitz functions of λ on $[\lambda_1, \lambda_2]$.*

Proof of Theorem 2.4. From (2.49) and (2.50), it is obvious that the square of the width of (2.51) for each $i = 0, \cdots, p$ is a quadratic function of λ on $[\lambda_1, \lambda_2]$.

Since (2.50) is always nonnegative, we can express the width of (2.51) as

$$p_i' \sqrt{(\lambda - b_i')^2 + c_i'}$$

where p_i', b_i' and c_i' are constants and $c_i' \geqslant 0$. If it is strictly positive on the entire interval $[\lambda_1, \lambda_2]$, then it is Lipschitz in λ on this interval. Otherwise, if it is zero for some $\lambda' \in [\lambda_1, \lambda_2]$, then we must have $\lambda' = b_i'$ and $c_i' = 0$. In either case, the width is a Lipschitz function.

From the fact that $\hat{t}_i = |\hat{\beta}_i|$ for any $i = 1, \cdots, p$, equation (2.28) and the expression of f_N, one can see that z_N is a piecewise linear function of $\hat{\beta}$ and is therefore a Lipschitz continuous function of λ. Thus, we conclude that endpoints of (2.51) are Lipschitz in λ on the interval $[\lambda_1, \lambda_2]$.

□

By assuming $\Phi_N(z_N)$ to be an invertible linear map on $[\lambda_1, \lambda_2]$, we assume that z_N stays in the same partition $\mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)}$ with a fixed value of $\boldsymbol{\gamma}(z_N)$ on $[\lambda_1, \lambda_2]$. As λ changes, z_N moves along a piecewise linear path. When z_N enters another partition, $\boldsymbol{\gamma}(z_N)$ and $\Phi_N(z_N)$ will change, and the confidence band will change its course. In Section 2.4.4 we will describe how to find those cut-off points.

2.4.3 Properties of simultaneous confidence bands for z_0

In this subsection, we assume the sample covariance matrix Σ_N^1 to be nonsingular, which is satisfied by sufficiently large N under Assumption 2.3. From (3.38), we can express the asymptotically exact $(1-\alpha)100\%$ confidence region for z_0 as

$$\left\{ z \in \mathbb{R}^{2p+1} \left| \begin{array}{rcl} N[\Phi_N(z_N)(z_N - z)]^T \begin{bmatrix} (\Sigma_N^1)^{-1}, & 0 \\ 0, & 0 \end{bmatrix} [\Phi_N(z_N)(z_N - z)] & \leqslant & \chi^2_{p+1}(\alpha) \\ {[0, I_p]}\,[\Phi_N(z_N)(z_N - z)] & = & 0 \end{array} \right. \right\} \tag{2.52}$$

Moreover, we choose $g(N)$ to satisfy (2.47). By Lemma 3.1, $\Phi_N(z_N)$ is a piecewise linear map with at least two pieces only if z_N is in the partition $\mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)}$ with $\gamma(i) = 3$ or 4 for some i. We define a set

$$\mathcal{G}(z_N) = \left\{ i \,\middle|\, \gamma(i) = 3 \text{ or } 4,\ i = 1, \cdots, p, \text{ where } \mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)} \text{ is the partition containing } z_N \right\},$$

and denote the total number of pieces of $\Phi_N(z_N)$ as T_P. Then we have

$$T_P \triangleq 2^{|\mathcal{G}(z_N)|}. \tag{2.53}$$

The approximate $(1-\alpha)$ 100% confidence region (3.44) is the union of T_P fractions of different ellipsoids. On each fraction, $\Phi_N(z_N)$ has a fixed matrix representation.

To find the explicit expression for a specific ellipsoid fraction, we specify the constraints that define this fraction. Let

$$z - z_N = \begin{bmatrix} r \\ \eta \end{bmatrix}, \quad r \in \mathbb{R}^{p+1},\ \eta \in \mathbb{R}^p,$$

and define a diagonal matrix $D \in \mathbb{R}^{(p+1)\times(p+1)}$ as

$$D_{k+1,k+1} \triangleq \begin{cases} 1 \text{ or } -1, & \big((z_N)_{k+1}, (z_N)_{k+1+p}\big) \in E_k^3 \cup E_k^4 \\ 0, & \big((z_N)_{k+1}, (z_N)_{k+1+p}\big) \in E_k^6 \cup E_k^7 \cup E_k^8 \end{cases}$$

for $k = 1, 2, \cdots, p$, while all the other elements of D are 0's. Note that D can take T_P possible values resulting from the different combinations of ± 1, each of which can be used to define an ellipsoid fraction. From Lemma 3.1, the matrix representations of Φ_4 are all invertible.

Consequently, after some algebraic manipulations we can eliminate the variable η and add some constraints to obtain an explicit representation for the projection of an ellipsoid fraction onto the r space as

$$\left\{ r \in \mathbb{R}^{p+1} \,\middle|\, \begin{array}{rcl} r^T Q r & \leqslant & \frac{1}{N}\chi^2_{p+1}(\alpha) \\ Dr & \leqslant & 0 \end{array} \right\}$$

where

$$Q = K^T(\Sigma^1_N)^{-1}K, \quad K = \Phi_1 - \Phi_2\Phi_4^{-1}\Phi_3 \tag{2.54}$$

and D takes one of the T_P possible values. Note that K is the Schur complement of $\Phi_N(z_N)$, so K is nonsingular if the condition in Corollary 2.1 is satisfied.

To obtain simultaneous confidence intervals, we find the maximal and minimal values of z_i for each $i = 1, \cdots, p+1$ in every piece of $\Phi_N(z_N)$, by solving the following optimization problems

$$\begin{array}{ll} & \max\ r_i \\ \text{s.t.} & \left\{ \begin{array}{rcl} r^T Q r & \leqslant & \frac{1}{N}\chi^2_{p+1}(\alpha), \\ Dr & \leqslant & 0, \end{array} \right. \end{array} \qquad \begin{array}{ll} & \min\ r_i \\ \text{s.t.} & \left\{ \begin{array}{rcl} r^T Q r & \leqslant & \frac{1}{N}\chi^2_{p+1}(\alpha), \\ Dr & \leqslant & 0, \end{array} \right. \end{array} \tag{2.55}$$

for each $i = 1, \cdots, p+1$ and each ellipsoid fraction. In the case in which $\mathcal{G}(z_N) = \varnothing$ (i.e., $\Phi_N(z_N)$ is a linear map) and Q is nonsingular, the constraint $Dr \leqslant 0$ disappears and we can find explicit optimal values of (2.55) as

$$\sqrt{\frac{1}{N}\chi^2_{p+1}(\alpha)(Q^{-1})_{i,i}} \qquad \text{and} \qquad -\sqrt{\frac{1}{N}\chi^2_{p+1}(\alpha)(Q^{-1})_{i,i}} \tag{2.56}$$

for the maximization and minimization problems respectively.

In general, the matrix D will change as one changes to a different ellipsoid fraction. Let the optimal value of (2.55) be R^j_i for the maximization problem and r^j_i for the minimization problem in the jth piece of $\Phi_N(z_N)$. Then we combine the optimal values to obtain the two endpoints of the approximate simultaneous confidence interval for $(z_0)_i$ as

$$\begin{aligned} L_i(\lambda) &= (z_N)_i + \min_{1\leqslant j\leqslant T_P}\{r_i^j\}, \\ U_i(\lambda) &= (z_N)_i + \max_{1\leqslant j\leqslant T_P}\{R_i^j\}. \end{aligned} \tag{2.57}$$

Next, we introduce the definitions that will be used to show the Lipschitz continuity of $L_i(\lambda)$ and $U_i(\lambda)$.

For any points $x, x' \in \mathbb{R}^n$, nonempty and closed sets $C, D \subset \mathbb{R}^n$, we denote the Euclidean distance between x and x' as

$$d_E(x, x') = ||x - x'||_2,$$

and the Euclidean distance between x and C as

$$d_E(x, C) = \inf_{y\in C} ||x - y||_2.$$

The Hausdorff distance between C and D is defined as

$$d_\infty(C, D) = \sup_{x\in\mathbb{R}^n} |d_E(x, C) - d_E(x, D)|.$$

Suppose the feasible set is

$$\Omega = \{x \in \mathbb{R}^n \mid g_i(x) = 0, i \in \mathcal{E};\ h_j(x) \leqslant 0, j \in \mathcal{I}\}.$$

We say that Mangasarian-Fromovitz constraint qualification (MFCQ) [34] holds at a feasible point $\bar{x} \in \Omega$ when the equality constraint gradients are linearly independent and there exists a vector $d \in \mathbb{R}^n$ such that

$$\nabla g_i(\bar{x})^T d = 0, i \in \mathcal{E}; \quad \text{and} \quad \nabla h_j(\bar{x})^T d < 0, \text{ for all } j \in A(\bar{x}) \cap \mathcal{I},$$

where $A(\bar{x})$ represents the active index set of the constraints at $\bar{x}$.

The graph of a set-valued mapping $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is defined as

$$\text{gph}\mathcal{F} = \{(x,u) | u \in \mathcal{F}(x)\}.$$

A set-valued mapping $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is outer semicontinuous (*osc*) relative to $\mathcal{X}$ at $\bar{x}$ if

$$\limsup_{x \to \bar{x},\ x \in \mathcal{X}} \mathcal{F}(x) = \mathcal{F}(\bar{x}).$$

$\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ has the Aubin property relative to $\mathcal{X}$ at $\bar{x}$ for $\bar{u}$, where $\bar{x} \in \mathcal{X}$ and $\bar{u} \in \mathcal{F}(\bar{x})$, if gph$\mathcal{F}$ is locally closed at $(\bar{x}, \bar{u})$ and there are neighborhoods $\mathcal{V} \in \mathcal{N}(\bar{x}), \mathcal{W} \in \mathcal{N}(\bar{u})$, and a positive constant κ such that

$$\mathcal{F}(x') \cap \mathcal{W} \subset \mathcal{F}(x) + \kappa |x' - x| \mathbb{B} \ \text{ for all } \ x, x' \in \mathcal{X} \cap \mathcal{V},$$

where $\mathbb{B}$ denotes the closed unit ball.

Note that the feasible set of (2.55) is changing with respect to λ, we can define a set-valued mapping for a specific ellipsoid fraction as

$$\begin{array}{rcl} \mathcal{F}: \ (0, +\infty) & \rightrightarrows & \mathbb{R}^{p+1} \\ \lambda & \mapsto & \text{feasible set of (2.55).} \end{array}$$

We state two facts about the mapping $\mathcal{F}(\lambda)$.

Lemma 2.6. *Suppose on an interval $[\lambda_1, \lambda_2] \subseteq (0, \infty)$, Σ_N^1 is nonsingular, $\frac{1}{g(N)} < \lambda$ holds and z_N stays in the same partition $\mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)}$ with the value of $\boldsymbol{\gamma}(z_N)$ fixed. Then for any $\bar{\lambda} \in [\lambda_1, \lambda_2]$ and $\bar{r} \in \mathcal{F}(\bar{\lambda})$, $\mathcal{F}(\lambda)$ has Aubin property relative to $[\lambda_1, \lambda_2]$ at $\bar{\lambda}$ for $\bar{r}$.*

Proof of Lemma 2.6. Since $\mathcal{F}(\lambda)$ is just for one ellipsoid fraction and z_N stays in the same partition $\mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)}$, D and $\Phi_N(z_N)$ will not change according to λ on $[\lambda_1, \lambda_2]$. For $\forall \bar{\lambda} \in [\lambda_1, \lambda_2]$ and $\forall \bar{r} \in \mathcal{F}(\bar{\lambda})$, let $M = \frac{1}{N}\chi_{p+1}^2(\alpha)$ and

$$I(\bar{r}) = \{i | \ (D\bar{r})_i = 0\}$$

be the index set of the active constraints in $Dr \leqslant 0$ at $\bar{r}$. Let $D_{I(\bar{r})}$ denote the submatrix of D which consists of the corresponding active rows. Then we have $D_{I(\bar{r})}\bar{r} = 0$. It is well known that the Aubin property follows the MFCQ condition for feasible set mapping [12]. We only need to show the MFCQ condition holds in (2.55) for $\bar{\lambda}$ and $\bar{r}$.

If $\bar{r}^T Q \bar{r} < M$, the first constraints in (2.55) is not active at $\bar{r}$. Since the nonzero entries of different rows in $D_{I(\bar{r})}$ would not appear in the same column, there exists some vector $\vec{d} \in \mathbb{R}^{p+1}$ such that $D_{I(\bar{r})}\vec{d} < 0$. Thus the MFCQ condition holds.

If $\bar{r}^T Q \bar{r} = M$, let

$$h(r) = \begin{bmatrix} r^T Q(\bar{\lambda}) r - M \\ D_{I(\bar{r})} r \end{bmatrix},$$

then $h(\bar{r}) = 0$ and $h(0) = \begin{bmatrix} -M \\ 0 \end{bmatrix}$. Since $h(\cdot)$ is continuous and $D_{I(\bar{r})}$ has a special structure, we can always change $r = 0$ a little bit to find a $\tilde{r}$ such that $h(\tilde{r}) < 0$. Note that matrix $Q(\bar{\lambda})$ is positive semidefinite on $[\lambda_1, \lambda_2]$, so $h(r)$ is a convex function. Thus we have

$$h(\bar{r}) + \nabla h(\bar{r})(\bar{r} - \tilde{r}) \leqslant h(\tilde{r}) < 0.$$

Let $d = \bar{r} - \tilde{r}$, then the above inequality becomes $\nabla h(\bar{r}) d < 0$, which implies the MFCQ condition.

□

Lemma 2.7. *Suppose on an interval $[\lambda_1, \lambda_2] \subseteq (0, \infty)$, Σ_N^1 and K are nonsingular, $\frac{1}{g(N)} < \lambda$ holds, the SAA solution $(\hat{\beta}_0, \hat{\beta})$ is a linear function of λ and z_N stays in the same partition $\mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)}$ with the value of $\gamma(z_N)$ fixed. Then for any $\bar{\lambda} \in [\lambda_1, \lambda_2]$, $\mathcal{F}(\lambda)$ is osc relative to $[\lambda_1, \lambda_2]$ at $\bar{\lambda}$. In addition, $\mathcal{F}(\lambda)$ is bounded on $[\lambda_1, \lambda_2]$.*

Proof of Lemma 2.7. From the proof of Theorem 2.4, we know that each element of Σ_N^1 is a quadratic function of λ. Since Σ_N^1 is nonsingular on $[\lambda_1, \lambda_2]$, according to the matrix inverse formula, each element of $(\Sigma_N^1)^{-1}$ has a form $\frac{P_1(\lambda)}{P_2(\lambda)}$, where $P_1(\lambda)$ and $P_2(\lambda)$ are polynomials and $P_2(\lambda) \neq 0$ on $[\lambda_1, \lambda_2]$. Also, each entry of Q is a rational function of λ with nonzero denominator on $[\lambda_1, \lambda_2]$, because $Q = K^T (\Sigma_N^1)^{-1} K$ and nonsingular matrix K is independent of λ on $[\lambda_1, \lambda_2]$. Moreover, Q is positive definite on $[\lambda_1, \lambda_2]$ since we assume Σ_N^1 and K are nonsingular on $[\lambda_1, \lambda_2]$.

Consider two arbitrary sequence $\{\lambda^n\}_{n \geqslant 1}$ and $\{r^n\}_{n \geqslant 1}$ such that $\lambda^n, \bar{\lambda} \in [\lambda_1, \lambda_2]$ and $r^n \in \mathcal{F}(\lambda^n)$ for all n, $\lambda^n \to \bar{\lambda}$, $r^n \to \bar{r}$ as $n \to \infty$. In order to show *osc* at $\bar{\lambda}$, we must verify that the limit point $\bar{r} \in \mathcal{F}(\bar{\lambda})$. Since $r^n \in \mathcal{F}(\lambda^n)$, we have

$$\begin{cases} (r^n)^T Q(\lambda^n) r^n & \leqslant & M, \\ D r^n & \leqslant & 0. \end{cases} \tag{2.58}$$

Since each element of $Q(\lambda)$ is a rational function with domain $[\lambda_1, \lambda_2]$, $\lim_{n \to \infty} Q(\lambda^n) = Q(\bar{\lambda})$. Taking limit as $n \to \infty$ on the inequalities of (2.58), we get

$$\begin{cases} \bar{r}^T Q(\bar{\lambda}) \bar{r} & \leqslant & M, \\ D\bar{r} & \leqslant & 0, \end{cases}$$

i.e. $\bar{r} \in \mathcal{F}(\bar{\lambda})$. So the *osc* property follows.

On the other hand, in order to see that $\mathcal{F}(\lambda)$ is bounded on $[\lambda_1, \lambda_2]$, we do eigenvalue decomposition for $Q(\bar{\lambda})$. We get

$$Q(\bar{\lambda}) = U(\bar{\lambda}) V(\bar{\lambda}) U(\bar{\lambda})^T,$$

where $U(\bar{\lambda})$ is an orthogonal matrix and $V(\bar{\lambda})$ is the eigenvalue matrix with increasing positive diagonal entries. We always can do this because $Q(\bar{\lambda})$ is positive definite on $[\lambda_1, \lambda_2]$. For convenience, we dismiss the variable $\bar{\lambda}$ from now on. So we have

$$r^T Q r = \big(U^T r\big)^T V \big(U^T r\big). \tag{2.59}$$

Since Q is continuous in λ on $[\lambda_1, \lambda_2]$, all the eigenvalues of Q are also continuous with respect to λ on $[\lambda_1, \lambda_2]$. So all the eigenvalues of Q are positive and bounded on $[\lambda_1, \lambda_2]$. It is obvious that $\big\{U^T r \,\big|\, r^T Q r \leqslant M\big\}$ is bounded on $[\lambda_1, \lambda_2]$ from (2.59), and so is $\big\{||U^T r|| \,\big|\, r^T Q r \leqslant M\big\}$. Since U is an orthogonal matrix, $\big\{r \,\big|\, r^T Q r \leqslant M\big\}$ is bounded on $[\lambda_1, \lambda_2]$, and so is $\big\{r \,\big|\, r^T Q r \leqslant M$ and $D r \leqslant 0\big\}$. I.e., $\mathcal{F}(\lambda)$ is bounded on $[\lambda_1, \lambda_2]$.

□

From Lemma 2.6 and 2.7, we are ready to prove the piecewise Lipschitz property of $L_i(\lambda)$ and $U_i(\lambda)$, when λ is restricted to certain sections.

Theorem 2.5. *With a fixed sample size N, suppose that Σ_N^1 and K are nonsingular on an interval $[\lambda_1, \lambda_2] \subseteq (0, \infty)$. Additionally, suppose (2.47) holds and the SAA solution $(\hat{\beta}_0, \hat{\beta})$ is a linear function of λ and z_N stays in the same partition $\mathbb{R} \times \Pi_{i=1}^{p} E_i^{\gamma(i)}$ with the value of $\boldsymbol{\gamma}(z_N)$ fixed on $[\lambda_1, \lambda_2]$. Then the approximate confidence region (3.44) is Lipschitz continuous on $[\lambda_1, \lambda_2]$ in Hausdorff distance and the endpoints of the approximate simultaneous confidence interval for $(z_0)_i$ in (2.57) are Lipschitz functions of λ on $[\lambda_1, \lambda_2]$.*

The Lipschitz continuity we showed for the end points of confidence intervals is restricted to certain intervals. At some λ, the partition containing z_N changes abruptly, which causes a dramatic change in $\Phi_N(z_N)$ and $\mathcal{G}(z_N)$. Hence a sudden jump in the boundaries of the confidence band may appear at such λ points. Consequently, it is important to track those λ points where the value of $\boldsymbol{\gamma}(z_N)$ changes, in order to correctly characterize the entire confidence band. The next section will describe how to track those discontinuity points and compute confidence bands.

2.4.4 Algorithms of LASSO confidence intervals along the path

In this section, we describe an algorithm to establish confidence bands for the LASSO parameters along the LASSO solution paths. In order to obtain the confidence bands for $(\tilde{\beta}_0, \tilde{\beta})$, we first find the confidence bands for z_0, then transfer them onto (β_0, β) using the projector Γ defined in (3.42). In Sections 2.4.2 and 2.4.3, we showed that the endpoints of the approximate individual and simultaneous confidence intervals for z_0 are Lipschitz continuous in λ on certain λ segments. On those segments, the SAA solution $(\hat{\beta}_0, \hat{\beta})$ is a linear function and z_N stays in a fixed partition. To approximate the confidence bands along the entire path, we first obtain all the break points of the above segments, and then compute the confidence intervals for each of those break points. We use linear approximations for the confidence intervals on the intervals between the break points, which are reasonable given the properties proved in Theorems 2.4 and 2.5.

Although Theorems 2.4 and 2.5 assume $\Phi_N(z_N)$ to be a linear map and Σ_N^1 to be nonsingular, our techniques work even if those assumptions fail. We can compute the confidence intervals as long as the condition in Corollary 2.1 holds. In practice, we can choose $1/g(N)$ to be very small such that most of λ values fall into segments on which $\Phi_N(z_N)$ is linear. Next, we give algorithms to find the knots of λ that result in these λ segments.

First, we modify Rosset and Zhu's path tracking algorithm [44] by using KKT conditions of (2.2) to find the λ segments where $(\hat{\beta}_0(\lambda), \hat{\beta}(\lambda))$ is linear. Recall that our SAA problem is

$$\min_{\beta_0,\beta} \frac{1}{N}||\mathbf{y} - \beta_0 1_N - X\beta||_2^2 + \lambda \sum_{i=1}^{p} |\beta_i|,$$

in which the intercept β_0 is included. Since the original LASSO path tracking algorithm in [44] did not consider the intercept, we adapt that algorithm to obtain Algorithm 1 below.

Denote $\mathbf{x}_i = (x_{1i}, x_{2i}, \cdots, x_{Ni})^T,\ i = 1, \cdots, p$ and $\vec{\boldsymbol{\beta}} = (\beta_0, \beta)$. Let $\mathcal{A} = \{i \mid \beta_i \neq 0,\ i = 1, \cdots, p\}$ be the active set, and

$$\mathbf{X}_{\mathcal{A}} = [\mathbf{x}_j]_{j\in\mathcal{A}}.$$

Assuming $\mathbf{X}_{\mathcal{A}}$ is always full column rank, we present the modified LASSO path tracking algorithm as follows.

Algorithm 1 (Tracking the LASSO solution path for the SAA problem)

1. Inputs: a vector $\mathbf{y}$ in $\mathbb{R}^N$, a matrix $\mathbf{X}$ in $\mathbb{R}^{N\times p}$.

2. Initialization: Set $\beta = 0$, $\beta_0 = \frac{1}{N}1_N^T\mathbf{y}$, $\lambda = ||\frac{2}{N}\mathbf{X}^T(\mathbf{y} - \beta_0 1_N)||_\infty$; active set $\mathcal{A} = \{\ i : |\frac{2}{N}\mathbf{x}_i^T(\mathbf{y} - \beta_0 1_N)| = \lambda\}$ and inactive set $\mathcal{I} = \{1, 2, \cdots, p\}\backslash\mathcal{A}$.

3. While $(\lambda > 0)$

   (a) Compute the direction $\vec{\nu} \triangleq (\nu_0, \nu) \in \mathbb{R} \times \mathbb{R}^p$ of the solution path as λ decreases.

$$\nu_{\mathcal{A}} = \frac{N}{2}\big[\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}} - \frac{1}{N}\mathbf{X}_{\mathcal{A}}^T 1_N 1_N^T \mathbf{X}_{\mathcal{A}}\big]^{-1} sgn\big(\mathbf{X}_{\mathcal{A}}^T(\mathbf{y} - \beta_0 1_N - \mathbf{X}\beta)\big),$$

$$\nu_{\mathcal{I}} = 0, \quad \text{and} \quad \nu_0 = -\frac{1}{N}1_N^T\mathbf{X}_{\mathcal{A}}\nu_{\mathcal{A}}.$$

(b) Set $d_1 = \min\{d > 0 : (\beta + d\nu)_j = 0,\ j \in \mathcal{A}\}$,

$d_2 = \min\{d > 0 : \left|\frac{2}{N}\mathbf{x}_j^T[\mathbf{y} - (\beta_0 + d\nu_0)1_N - \mathbf{X}(\beta + d\nu)]\right| = \lambda - d,\ j \in \mathcal{I}\}$.

Find the step length: $d = \min(d_1, d_2)$.

(c) If $d = d_1$ then remove the variable attaining 0 at d from $\mathcal{A}$ and add it to $\mathcal{I}$.

If $d = d_2$ then add the variable attaining equality at d to $\mathcal{A}$ and remove it from $\mathcal{I}$.

(d) Update β_0, β, λ: $\beta_0 \leftarrow \beta_0 + d\nu_0$, $\beta \leftarrow \beta + d\nu$, $\lambda \leftarrow \lambda - d$.

Record λ, (β_0, β) and $\vec{\nu}$.

4. Return: Sequences of recorded values of λ, (β_0, β) and $\vec{\nu}$.

After running algorithm 1, we obtain a sequence of consecutive λ segments on which the SAA solution $(\hat{\beta}_0(\lambda), \hat{\beta}(\lambda))$ is linear in λ. Our next task is to divide each such segment into smaller pieces on which $\boldsymbol{\gamma}(z_N)$ is fixed. Assuming $\boldsymbol{\gamma}(z_N)$ changes by one component at a time when λ decreases, we present the following algorithm to find all such pieces on any segment on which $(\hat{\beta}_0(\lambda), \hat{\beta}(\lambda))$ is linear.

Algorithm 2 (Locating λ's at which $\boldsymbol{\gamma}(z_N)$ changes)

1. Inputs: a vector $\mathbf{y}$ in $\mathbb{R}^N$, a matrix $\mathbf{X}$ in $\mathbb{R}^{N \times p}$;

   an interval $[\lambda_1, \lambda_2]$ and the direction $\vec{\nu}$ from Algorithm 1;

   the SAA solution $\vec{\beta}(\lambda_2) \triangleq \left(\hat{\beta}_0(\lambda_2), \hat{\beta}(\lambda_2)\right)$ with parameter λ_2;

   the parameter $\frac{1}{g(N)}$ in $\mathbb{R}$ which satisfies (2.47).

2. Initialization: Set $\lambda = \lambda_2$;

   $z_N = \vec{\beta}(\lambda_2) + \frac{2}{N}\breve{\mathbf{X}}^T(\mathbf{y} - \breve{\mathbf{X}}\vec{\beta}(\lambda_2))$ (only components associated with β);

   direction of z_N in decreasing λ: $\nu_z = \vec{\nu} - \frac{2}{N}\breve{\mathbf{X}}^T\breve{\mathbf{X}}\vec{\nu}$;

   find $\boldsymbol{\gamma}(z_N)$ in $\mathbb{R}^p$ using Table 1.

3. While $(\lambda > \lambda_1)$

(a) For $i = 1, \cdots, p$, compute the shortest step length d_i such that $\left((z_N)_{i+1}, (z_N)_{i+1+p}\right)$ meets a boundary of E_i^j for some $j \in \{0, 1, \cdots, 8\}$.

If $\boldsymbol{\gamma}(z_N)_i = 6$ then $d_i = \min\{d > 0 : (z_N)_{i+1} + (\nu_z)_{i+1}d = -(\lambda - d) - \frac{1}{2g(N)}\}$;

else if $\boldsymbol{\gamma}(z_N)_i = 8$ then $d_i = \min\{d > 0 : (z_N)_{i+1} + (\nu_z)_{i+1}d = \lambda - d + \frac{1}{2g(N)}\}$;

else if $\boldsymbol{\gamma}(z_N)_i = 7$ then $d_i = \min\{d > 0 :$

$(z_N)_{i+1} + (\nu_z)_{i+1}d = -(\lambda - d) + \frac{1}{g(N)}$, or $(z_N)_{i+1} + (\nu_z)_{i+1}d = \lambda - d - \frac{1}{g(N)}\}$;

else if $\boldsymbol{\gamma}(z_N)_i = 3$ then $d_i = \min\{d > 0 :$

$(z_N)_{i+1} + (\nu_z)_{i+1}d = -(\lambda - d) - \frac{1}{2g(N)}$, or $(z_N)_{i+1} + (\nu_z)_{i+1}d = -(\lambda - d) + \frac{1}{g(N)}\}$;

else if $\boldsymbol{\gamma}(z_N)_i = 4$ then $d_i = \min\{d > 0 :$

$(z_N)_{i+1} + (\nu_z)_{i+1}d = \lambda - d - \frac{1}{g(N)}$, or $(z_N)_{i+1} + (\nu_z)_{i+1}d = \lambda - d + \frac{1}{2g(N)}\}$.

(b) Find the step length: $d = \min(d_1, d_2, \cdots, d_p)$.

(c) If $d = d_i$ $(i = 1, \cdots, p)$, assume the old value of $\boldsymbol{\gamma}(z_N)_i$ is j $(j = 3, 4, 6, 7,$ or $8)$ and $(z_N)_{i+1} + (\nu_z)_{i+1}d$ achieves the boundary between E_i^j and E_i^l, then we update $\boldsymbol{\gamma}(z_N)_i \leftarrow l$.

(d) Update λ, $\vec{\beta}$, z_N: $\lambda \leftarrow \lambda - d$, $\vec{\beta} \leftarrow \vec{\beta} + \vec{\nu}d$, $z_N = \vec{\beta} + \frac{2}{N}\breve{\mathbf{X}}^T(\mathbf{y} - \breve{\mathbf{X}}\vec{\beta})$.

Record λ, $\vec{\beta}$, z_N and $\boldsymbol{\gamma}(z_N)$.

4. Return: Sequences of recorded values of λ, $\vec{\beta}$, z_N and $\boldsymbol{\gamma}(z_N)$.

Applying Algorithm 2 to every λ segment obtained from Algorithm 1, we obtain the knots of λ between which $\boldsymbol{\gamma}(z_N)$ is of a fixed value. Then we find confidence intervals for z_0 from (2.51) or (2.57) at each λ knot and link their corresponding boundaries linearly to obtain the confidence band on each λ segment. It should be noted that the confidence band is usually not continuous at a λ knot that has different $\boldsymbol{\gamma}(z_N)$ values of the λ segments on its left and right sides. We summarize the main algorithm below.

Algorithm 3 (Main algorithm: computing the confidence bands for $(\tilde{\beta}_0, \tilde{\beta})$)

1. Inputs: a vector $\mathbf{y}$ in $\mathbb{R}^N$, a matrix $\mathbf{X}$ in $\mathbb{R}^{N\times p}$, the parameter $\frac{1}{g(N)}$ in $\mathbb{R}$ which satisfies (2.47).

2. Run Algorithm 1. Obtain s consecutive λ segments with $s+1$ knots $\lambda_1 < \lambda_2 < \cdots < \lambda_{s+1}$ and values of $\vec{\beta}$, $\vec{\nu}$ on $\lambda_2, \cdots, \lambda_{s+1}$ except $\lambda_1 = 0$.

3. For $i = 1, \cdots, s$,

   (a) Run Algorithm 2 on segment $[\lambda_i, \lambda_{i+1}]$. Obtain $s_i + 1$ knots of λ: $\lambda_i = \lambda^1 < \lambda^2 < \cdots < \lambda^{s_i} < \lambda^{s_i+1} = \lambda_{i+1}$ and values of $\boldsymbol{\gamma}(z_N)$ on $\lambda^2, \cdots, \lambda^{s_i+1}$.

   (b) For $j = 1, \cdots, s_i$, compute simultaneous CIs for the first $p+1$ components of z_0 from (2.56) or (2.57), and individual CIs from (2.51) or simulation discussed in Section 2.3.3 at λ^j and λ^{j+1} by using the value of $\boldsymbol{\gamma}(z_N)$ at λ^{j+1}. Then link the corresponding endpoints of CIs for the same component of z_0 linearly between λ^j and λ^{j+1}.

4. Use the projector Γ (3.42) to transform the confidence bands for z_0 into confidence bands for $(\tilde{\beta}_0, \tilde{\beta})$ on $[\lambda_1, \lambda_{s+1}]$.

Using this algorithm, we can efficiently compute confidence intervals at a large number of λ's, if they all belong to one segment. However, when the number of λ segments is large, it can take a long time to compute the entire confidence band. The main computational cost is on computing the confidence intervals at the λ knots where $\mathcal{G}(z_N) \neq \varnothing$. When $\mathcal{G}(z_N) \neq \varnothing$, we use simulation and (2.57) to compute individual and simultaneous confidence intervals respectively, which are computationally expensive. For simplification, we can choose $1/g(N)$ to be very small such that almost all of λ values fall in segments where $\mathcal{G}(z_N) = \varnothing$, and then use (2.51) and (2.56) to compute confidence intervals at the endpoints of these segments.

2.5 Confidence intervals for the true parameters in the underlying linear model

In this section, we derive asymptotic results for the true parameters in an underlying linear model based on the convergence theorems in Section 2.3, and aim to obtain individual confidence intervals for the true parameters.

Suppose the true linear model between X and Y is

$$Y = \beta_0^{true} + X^T \beta^{true} + \varepsilon, \tag{2.60}$$

where $\beta_0^{true} \in \mathbb{R}$ and $\beta^{true} = (\beta_1^{true}, \cdots, \beta_p^{true}) \in \mathbb{R}^p$ are the true parameters. The random error ε has mean zero and variance σ_ε^2. Moreover, ε is independent of X_i for each $i = 1, \cdots, p$. In this section, we assume that $E(X_i) = 0$ for each $i = 1, \cdots, p$, hence $E(Y) = \beta_0^{true}$. Denote the covariance matrix of X as Σ, i.e., $\Sigma = E(XX^T)$.

Plugging (3.46) into (2.14), we have

$$z_0 = \begin{bmatrix} \tilde{\beta}_0 + 2E(Y - \tilde{\beta}_0 - X^T\tilde{\beta}) \\ \tilde{\beta} + 2E\left[(Y - \tilde{\beta}_0 - X^T\tilde{\beta})X\right] \\ \tilde{t} - \lambda e_p \end{bmatrix} = \begin{bmatrix} 2\beta_0^{true} - \tilde{\beta}_0 \\ \tilde{\beta} + 2\Sigma(\beta^{true} - \tilde{\beta}) \\ \tilde{t} - \lambda e_p \end{bmatrix} \tag{2.61}$$

If Σ is nonsingular, then from (3.47) and the fact that $\tilde{\beta}_0 = (z_0)_1$ in Section 2.3.4 we obtain

$$\beta_0^{true} = (z_0)_1, \quad \beta^{true} = \frac{1}{2}\Sigma^{-1}(z_0)_{2:(p+1)} + \left[I_p - \frac{1}{2}\Sigma^{-1}\right]\tilde{\beta}, \tag{2.62}$$

where $(z_0)_{2:(p+1)}$ denotes the vector that consists of the second to $p+1$'th entries of z_0. Expression (3.48) suggests the following estimators

$$\hat{\beta}_0^{true} = (z_N)_1, \quad \hat{\beta}^{true} = \frac{1}{2}\hat{\Theta}(z_N)_{2:(p+1)} + \left[I_p - \frac{1}{2}\hat{\Theta}\right]\hat{\beta}, \tag{2.63}$$

where $\hat{\Theta}$ is an estimator of the precision matrix Σ^{-1}. From (2.28) and (3.47) one may notice that the estimators in (3.49) essentially have the same expression as the de-biased estimator in [56] and [50], but we will show below that our construction of confidence intervals for the true parameters is different from theirs.

Let G be a map from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{p+1}$ defined as

$$G = \frac{1}{2}\left(\begin{bmatrix} 1 & 0 \\ 0 & \Sigma^{-1} \end{bmatrix} B + \begin{bmatrix} 1 & 0 \\ 0 & 2I - \Sigma^{-1} \end{bmatrix} B \circ \Pi_K\right), \tag{2.64}$$

and $\hat{G}$ be the following map

$$\hat{G} = \frac{1}{2}\left(\begin{bmatrix} 1 & 0 \\ 0 & \hat{\Theta} \end{bmatrix} B + \begin{bmatrix} 1 & 0 \\ 0 & 2I - \hat{\Theta} \end{bmatrix} B \circ d\Pi_S(z_N)\right), \tag{2.65}$$

where B is a $(p+1)$ by $(2p+1)$ matrix defined as $B = \begin{bmatrix} I_{p+1} & 0 \end{bmatrix}$. Since Π_S is positively homogeneous, we know that $\Pi_S(z_0) = d\Pi_S(z_0)(z_0) = \Pi_K(z_0)$ and $\Pi_S(z_N) = d\Pi_S(z_N)(z_N)$. Then according to (3.48), (3.49), $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = \Pi_S(z_0)$ and $(\hat{\beta}_0, \hat{\beta}, \hat{t}) = \Pi_S(z_N)$, we can rewrite (3.48) and (3.49) as

$$(\beta_0^{true}, \beta^{true}) = G(z_0) \quad \text{and} \quad (\hat{\beta}_0^{true}, \hat{\beta}^{true}) = \hat{G}(z_N).$$

The following theorem shows that (3.49) gives a consistent estimator of the true parameter $(\beta_0^{true}, \beta^{true})$, and provides an asymptotic distribution from which we can derive a confidence region for $(\beta_0^{true}, \beta^{true})$.

Theorem 2.6. *Suppose that Assumptions 2.1 and 2.2 hold, and the true covariance matrix Σ is nonsingular. Let $\hat{\Theta}$ be a $\sqrt{N}$-consistent estimator of Σ^{-1}. Then $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$ is a consistent estimator of $(\beta_0^{true}, \beta^{true})$ and*

$$\sqrt{N}\left((\hat{\beta}_0^{true}, \hat{\beta}^{true}) - (\beta_0^{true}, \beta^{true})\right) \Rightarrow G \circ (L_K)^{-1}(\mathcal{N}(0, \Sigma_0)), \tag{2.66}$$

where G is the map defined in (3.50) and Σ_0 is defined in (2.18).

Proof of Theorem 2.6. Let $\eta_0^i = \big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$ and $\eta_N^i = \big((z_N)_{i+1}, (z_N)_{i+1+p}\big)$ for all $i = 1, \cdots, p$. From $\tilde{t}_i = |\tilde{\beta}_i|$ and $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = \Pi_S(z_0)$, one can check that η_0^i can only be in riC_i^3, riC_i^4, riC_i^6, riC_i^7 or riC_i^8 (Here "ri" before a set denotes the relative interior of the set). The special structure of S and the locations of z_0 ensure that the equality $d\Pi_S(z_0)z_N - d\Pi_S(z_0)z_0 = d\Pi_S(z_0)(z_N - z_0)$ always holds, thus

$$\begin{aligned}
\sqrt{N}\left((\hat{\beta}_0^{true}, \hat{\beta}^{true}) - (\beta_0^{true}, \beta^{true})\right) &= \sqrt{N}(\hat{G}z_N - Gz_0) \\
&= \sqrt{N}(\hat{G}z_N - Gz_N) + \sqrt{N}(Gz_N - Gz_0) \\
&= \sqrt{N}(\hat{G} - G)z_N + \sqrt{N}G(z_N - z_0).
\end{aligned}$$

Because G is a continuous map, from (3.28) we have

$$\sqrt{N}G(z_N - z_0) \Rightarrow G \circ (L_K)^{-1}(\mathcal{N}(0, \Sigma_0)).$$

To show (3.52) it suffices to prove

$$\lim_{N\to\infty} \text{Prob}\left\{\sqrt{N}||(\hat{G} - G)z_N|| < \epsilon\right\} = 1 \tag{2.67}$$

for each $\epsilon > 0$.

Denote the conical subdivision of Π_{S_i} as $\mathfrak{B}_\mathfrak{i} = \{C_i^5, C_i^6, C_i^7, C_i^8\}$. Let

$$\mathfrak{B}_\mathfrak{i}(\eta^i) = \{C_i \in \mathfrak{B}_\mathfrak{i} \mid \eta^i \in C_i\},$$

and let $|\mathfrak{B}_\mathfrak{i}(\eta^i)|$ be the union of all sets in $\mathfrak{B}_\mathfrak{i}(\eta^i)$. We define two sets

$$I_1 = \left\{i \in \{1, \cdots, p\} \mid \eta_0^i \in \text{ri}C_i^6, \text{ri}C_i^7 \text{ or } \text{ri}C_i^8\right\},$$

and

$$I_2 = \left\{i \in \{1, \cdots, p\} \mid \eta_0^i \in \text{ri}C_i^3 \text{ or } \text{ri}C_i^4\right\}.$$

For a given index $i \in I_1$, since z_N converges to z_0 almost surely, for almost every $\omega \in \Omega$ there exists a positive integer N_ω^i, such that for all $N > N_\omega^i$, $d\Pi_{S_i}(\eta_N^i)$ and $d\Pi_{S_i}(\eta_0^i)$ are the same linear function. Therefore we have $d\Pi_{S_i}(\eta_N^i)\eta_N^i = d\Pi_{S_i}(\eta_0^i)\eta_N^i$.

Similarly, for a given index $i \in I_2$ and almost every $\omega \in \Omega$, there exists a positive integer N_ω^i, such that for all $N > N_\omega^i$, $\eta_N^i \in |\mathfrak{B}_\mathfrak{i}(\eta_0^i)|$, hence $d\Pi_{S_i}(\eta_N^i)\eta_N^i = d\Pi_{S_i}(\eta_0^i)\eta_N^i$.

In summary, the fact that $S = \mathbb{R} \times \Pi_{i=1}^p S_i$ implies that for almost every $\omega \in \Omega$ there exists a positive integer $N_\omega^* = \max\{N_\omega^i, \cdots, N_\omega^p\}$, such that

$$d\Pi_S(z_N)z_N = d\Pi_S(z_0)z_N = \Pi_K(z_N), \quad \text{for all } N > N_\omega^*.$$

Therefore for sufficiently large N,

$$\begin{aligned} & \sqrt{N}||(\hat{G}-G)z_N|| \qquad (2.68) \\ \leqslant \quad & \frac{\sqrt{N}}{2}\left|\left|\left(\begin{bmatrix}1 & 0\\ 0 & \hat{\Theta}\end{bmatrix} - \begin{bmatrix}1 & 0\\ 0 & \Sigma^{-1}\end{bmatrix}\right) Bz_N\right|\right| \\ & + \frac{\sqrt{N}}{2}\left|\left|\left(\begin{bmatrix}1 & 0\\ 0 & 2I-\hat{\Theta}\end{bmatrix} - \begin{bmatrix}1 & 0\\ 0 & 2I-\Sigma^{-1}\end{bmatrix}\right) B\circ\Pi_K(z_N)\right|\right| \\ \leqslant \quad & \frac{\sqrt{N}}{2}\left|\left|\left(\begin{bmatrix}1 & 0\\ 0 & \hat{\Theta}\end{bmatrix} - \begin{bmatrix}1 & 0\\ 0 & \Sigma^{-1}\end{bmatrix}\right)\right|\right| \; ||Bz_N|| \\ & + \frac{\sqrt{N}}{2}\left|\left|\left(\begin{bmatrix}1 & 0\\ 0 & 2I-\hat{\Theta}\end{bmatrix} - \begin{bmatrix}1 & 0\\ 0 & 2I-\Sigma^{-1}\end{bmatrix}\right)\right|\right| \; ||B\circ\Pi_K(z_N)|| \end{aligned}$$

By Theorem 2.1 we know that z_N converges to z_0 almost surely. Furthermore, since $B\circ\Pi_K$ is a continuous map and $\hat{\Theta}$ is a $\sqrt{N}$-consistent estimator of Σ^{-1}, we have the following four equalities hold for each $\epsilon > 0$:

$$\lim_{N\to\infty} \text{Prob}\,\{||Bz_N|| \leqslant ||Bz_0|| + 1\} = 1,$$

$$\lim_{N\to\infty} \text{Prob}\,\{||B\circ\Pi_K(z_N)|| \leqslant ||B\circ\Pi_K(z_0)|| + 1\} = 1,$$

$$\lim_{N\to\infty} \text{Prob}\left\{\sqrt{N}\left|\left|\left(\begin{bmatrix}1 & 0\\ 0 & \hat{\Theta}\end{bmatrix} - \begin{bmatrix}1 & 0\\ 0 & \Sigma^{-1}\end{bmatrix}\right)\right|\right| < \frac{\epsilon}{||Bz_0||+1}\right\} = 1,$$

and

$$\lim_{N\to\infty} \text{Prob}\left\{\sqrt{N}\left|\left|\left(\begin{bmatrix}1 & 0\\ 0 & 2I-\hat{\Theta}\end{bmatrix} - \begin{bmatrix}1 & 0\\ 0 & 2I-\Sigma^{-1}\end{bmatrix}\right)\right|\right| < \frac{\epsilon}{||B\circ\Pi_K(z_0)||+1}\right\} = 1.$$

Combining (2.68) and the above four equalities proves (2.67).

Similarly, from $\left((\hat{\beta}_0^{true}, \hat{\beta}^{true}) - (\beta_0^{true}, \beta^{true})\right) = (\hat{G}-G)z_N + G(z_N - z_0)$, one can show that for each $\epsilon > 0$,

$$\lim_{N \to \infty} \text{Prob}\left\{ ||(\hat{\beta}_0^{true}, \hat{\beta}^{true}) - (\beta_0^{true}, \beta^{true})|| < \epsilon \right\} = 1,$$

i.e., $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$ is a consistent estimator of $(\beta_0^{true}, \beta^{true})$.

□

There are many choices for $\hat{\Theta}$ in real applications. Some common choices are the inverse of sample covariance matrix and the estimate of precision matrix computed by the banding method [4] or the penalized likelihood method [54; 16]. It is well known from the literature that under some regularity conditions, these estimators of the precision matrix have $\sqrt{N}$-consistency when p is fixed [22].

From (2.14) and the definition of f_0 (2.10) we note that $\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$ can be only in the relative interior of C_i^3, C_i^4, C_i^6, C_i^7 or C_i^8 for all i from 1 to p. Below, we consider two cases based on the location of z_0, which correspond to the two situations in which the random variable $(L_K)^{-1}(\mathcal{N}(0, \Sigma_0))$ is normally distributed, or is a combination of more than one normal random variables. We refer to these two cases as the single-piece case and the multiple-piece case respectively.

- Case I (single-piece case). In this case, $\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$ is in the relative interior of C_i^6, C_i^7 or C_i^8 for all $i \in \{1 \cdots p\}$, and the normal map L_K and the B-derivative $d\Pi_S(z_0)$ are linear functions. Note that

  $$d(f_N)_S(z_N)(h) = L_N \; d\Pi_S(z_N)(h) + h - d\Pi_S(z_N)(h) \text{ for each } h \in \mathbb{R}^{2p+1}.$$

  In this case, $d\Pi_S(z_N)$ converges to $d\Pi_S(z_0)$ almost surely, and $d(f_N)_S(z_N)$ converges to L_K almost surely, so we can use $d\Pi_S(z_N)$ and $d(f_N)_S(z_N)$ as the estimators of $d\Pi_S(z_0)$ and L_K respectively.

- Case II (multiple-piece case). In this case, $\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$ is in the relative interior of C_i^3 or C_i^4 for some index $i \in \{1 \cdots p\}$, and L_K and $d\Pi_S(z_0)$ are piecewise linear functions. This is the case in which $d\Pi_S(z)$ is discontinuous at z_0, and we use $\Phi_N(z_N)$ and $\Lambda_N(z_N)$

to estimate L_K and $d\Pi_S(z_0)$ respectively.

To use (3.52) to compute confidence intervals, we replace G and L_K there by their estimators. For Case I, the following theorem gives an approach to compute the asymptotically exact individual confidence intervals for $(\beta_0^{true}, \beta^{true})$.

Theorem 2.7. *Suppose that Assumptions 2.1, 2.2 and 2.4(a-b) hold, the true covariance matrix* Σ *is nonsingular, and the solution to the normal map formulation* (2.13) *satisfies the conditions for Case I. Let* $\hat{\Theta}$ *be a* $\sqrt{N}$*-consistent estimator of* Σ^{-1}*, and define* $H = G(L_K)^{-1}$ *and* $H_N = \hat{G}\,[d(f_N)_S(z_N)]^{-1}$*. If* $(H\Sigma_0 H^T)_{i+1,i+1} \neq 0$*, then*

$$\frac{\sqrt{N}(\hat{\beta}_i^{true} - \beta_i^{true})}{\sqrt{(H_N \Sigma_N H_N^T)_{i+1,i+1}}} \Rightarrow \mathcal{N}\left(0, 1\right), \tag{2.69}$$

for all $i = 0, 1, \cdots, p$*.*

Proof of Theorem 2.7. In Case I, G, L_K and H are linear maps, and $\hat{G}$, $d(f_N)_S(z_N)$ and H_N are all linear for sufficiently large N. To prove (3.53), we will show that $(H_N \Sigma_N H_N^T)_{i,i}$ converges to $(H\Sigma_0 H^T)_{i,i}$ in probability for all $i \in \{1, 2, \cdots, p+1\}$. Then the results follow from (3.52) and Slutsky's Theorem.

From (3.50) and (3.51), one can see that $\hat{G}$ converges to G in probability, since $\hat{\Theta}$ converges to Σ^{-1} in probability and $d\Pi_S(z_N)$ is the same as Π_K for sufficiently large N. Note that $[d(f_N)_S(z_N)]^{-1}$ converges to $(L_K)^{-1}$ almost surely in Case I, which implies for each $\epsilon > 0$,

$$\lim_{N\to\infty} \text{Prob}\left\{||\hat{G} - G||\;||\,[d(f_N)_S(z_N)]^{-1}\,|| < (||L_K^{-1} + 1||)\frac{\epsilon}{2}\right\} = 1, \tag{2.70}$$

and

$$\lim_{N\to\infty} \text{Prob}\left\{||G||\;||L_K^{-1} - [d(f_N)_S(z_N)]^{-1}\,|| < \frac{\epsilon}{2}\right\} = 1. \tag{2.71}$$

Since

$$||\hat{G}\,[d(f_N)_S(z_N)]^{-1} - G(L_K)^{-1}|| \leqslant ||\hat{G} - G||\;||\,[d(f_N)_S(z_N)]^{-1}\,|| + ||G||\;||L_K^{-1} - [d(f_N)_S(z_N)]^{-1}\,||,$$

(2.70) and (2.71) imply

$$\lim_{N\to\infty} \text{Prob}\left\{||H_N - H|| < \epsilon\right\} = 1, \quad \text{for each } \varepsilon > 0. \tag{2.72}$$

By Lemma 2.3, Σ_N converges to Σ_0 almost surely, so $H_N\Sigma_N$ converges to $H\Sigma_0$ in probability. From the following inequality

$$\begin{aligned}
& \text{Prob}\left\{||H_N\Sigma_N H_N^T - H\Sigma_0 H^T|| < \epsilon\right\} \\
\geqslant\ & \text{Prob}\left\{||H_N\Sigma_N - H\Sigma_0||\ ||H_N|| < \frac{\epsilon}{2}\right\} + \text{Prob}\left\{||H\Sigma_0||\ ||H_N - H|| < \frac{\epsilon}{2}\right\} - 1 \\
\geqslant\ & \text{Prob}\left\{||H_N\Sigma_N - H\Sigma_0|| < \frac{\epsilon}{2(||H||+1)}\right\} + \text{Prob}\left\{||H_N|| \leqslant (||H||+1)\right\} - 1 \\
& + \text{Prob}\left\{||H\Sigma_0||\ ||H_N - H|| < \frac{\epsilon}{2}\right\} - 1,
\end{aligned}$$

one can see that $H_N\Sigma_N H_N^T$ converges to $H\Sigma_0 H^T$ in probability, which implies that $(H_N\Sigma_N H_N^T)_{i,i}$ converges to $(H\Sigma_0 H^T)_{i,i}$ in probability for all $i \in \{1, 2, \cdots, p+1\}$.

□

Theorem 2.7 suggests constructing an asymptotically exact individual confidence interval for β_i^{true} with the significance level α as

$$\left[\ \hat{\beta}_i^{true} - \sqrt{\frac{\chi_1^2(\alpha)\bar{m}_i}{N}},\ \hat{\beta}_i^{true} + \sqrt{\frac{\chi_1^2(\alpha)\bar{m}_i}{N}}\ \right],$$

where $\bar{m}_i$ is the ith diagonal element of the matrix $H_N\Sigma_N H_N^T$.

For Case II, to show how to compute asymptotically exact individual confidence intervals for $(\beta_0^{true}, \beta^{true})$, we consider the image of normal random vectors under certain functions. Let $f : \mathbb{R}^{2p+1} \to \mathbb{R}$ be a continuous function and Z be a random variable in $\mathbb{R}^{2p+1}$ with $Z \sim \mathcal{N}(0, I_{p+1}) \times \vec{\mathbf{0}}$. Define $a^r(f) \in (0, \infty)$ as

$$a^r(f) = \inf\left\{c \geqslant 0 \mid \text{Prob}\left\{-c \leqslant f(Z) - r \leqslant c\right\} \geqslant 1 - \alpha\right\}. \tag{2.73}$$

Suppose that $\text{Prob}\left\{f(Z) = b\right\} = 0$ for all $b \in \mathbb{R}$. Then for any given $r \in \mathbb{R}$ and $\alpha \in (0, 1)$, $a^r(f)$

as defined in (3.54) is the smallest value that satisfies

$$\text{Prob}\left\{-a^r(f) \leqslant f(Z) - r \leqslant a^r(f)\right\} = 1 - \alpha.$$

Define two functions R and $\hat{R}$ from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{p+1}$ as

$$R = G \circ (L_K)^{-1} \begin{bmatrix} (\Sigma_0^1)^{\frac{1}{2}} & 0 \\ 0 & I_p \end{bmatrix} \quad \text{and} \quad \hat{R} = \hat{G}' \circ (\Phi_N(z_N))^{-1} \begin{bmatrix} (\Sigma_N^1)^{\frac{1}{2}} & 0 \\ 0 & I_p \end{bmatrix}, \tag{2.74}$$

where

$$\hat{G}' = \frac{1}{2}\left(\begin{bmatrix} 1 & 0 \\ 0 & \hat{\Theta} \end{bmatrix} B + \begin{bmatrix} 1 & 0 \\ 0 & 2I - \hat{\Theta} \end{bmatrix} B \circ \Lambda_N(z_N) \right). \tag{2.75}$$

We denote the jth component function of R and $\hat{R}$ as R_j and $\hat{R}_j$ respectively for each $j = 1, 2, \cdots, p+1$.

Note that the map G is a piecewise linear function in Case II. From the expression (3.50) and the matrix representations of the piecewise linear function Π_K based on the location of z_0 (see Section 2.3.2), one can check that G has the following form

$$\left[\begin{array}{cc|c} 1 & 0 & \\ & & * \\ 0 & \frac{1}{2}\Sigma^{-1}(I - W) + W & \end{array}\right], \tag{2.76}$$

in which W is a piecewise linear function represented by $p \times p$ diagonal matrices with diagonal elements 0 or $\frac{1}{2}$. If Σ and Σ_0^1 are nonsingular, then the matrix representation of each piece of the map G has full row rank. Because L_K is a global homeomorphism under Assumptions 2.1(a) and 2.2, it follows that $\text{Prob}\left\{R_j(Z) = b\right\} = 0$ for all $b \in \mathbb{R}$. The following theorem gives a way of computing individual confidence intervals for $(\beta_0^{true}, \beta^{true})$.

Theorem 2.8. *Suppose that Assumptions 2.1, 2.2 and 2.4(a-b) hold, and the population covariance matrices Σ and Σ_0^1 are nonsingular. Let $\hat{\Theta}$ be a $\sqrt{N}$-consistent estimator of Σ^{-1}, $\alpha \in (0,1)$ and $a^r(\cdot)$ be as in (3.54). Then for every $r \in \mathbb{R}$ and all $j = 0, 1, \cdots, p$, we have*

$$\lim_{N\to\infty} \text{Prob}\left\{ |\sqrt{N}(\hat{\beta}_j^{true} - \beta_j^{true}) - r| \leqslant a^r(\hat{R}_{j+1}) \right\} = 1 - \alpha, \tag{2.77}$$

where R and $\hat{R}$ are defined in (3.55).

The proof of Theorem 2.8 uses the following two lemmas.

Lemma 2.8. *Let $C(\mathbb{R}^{2p+1}, \mathbb{R})$ denote the space of continuous functions from $\mathbb{R}^{2p+1}$ to $\mathbb{R}$, $\{u_N\}_{N=1}^{\infty}$ be a sequence of $C(\mathbb{R}^{2p+1}, \mathbb{R})$ valued random variables which converges to u in probability uniformly on compact sets, and $\{Z_N\}_{N=1}^{\infty}$ be a sequence of real valued random variables that converges to $u(Z)$ in distribution. Then for every $r \in \mathbb{R}$,*

$$\lim_{N\to\infty} \text{Prob}\left\{-a^r(u_N) \leqslant Z_N - r \leqslant a^r(u_N)\right\} = 1 - \alpha.$$

Proof of Lemma 2.8. By Lemma 1 in [23] and the assumption that $u_N \to u$ in probability uniformly on compact sets, it follows that $a^r(u_N) \to a^r(u)$ in probability. Since $a^r(u) > 0$,

$$\frac{1}{a^r(u_N)} \mathbb{1}_{a^r(u_N)>0} \to \frac{1}{a^r(u)}$$

in probability, where $\mathbb{1}_{a^r(u_N)>0}$ is the indicator random variable for the event $a^r(u_N) > 0$. Let A_N denote the event $a^r(u_N) > 0$. Then

$$\begin{aligned} \text{Prob}\left\{-a^r(u_N) \leq Z_N - r \leqslant a^r(u_N)\right\} &= \text{Prob}\left\{A_N;\ -1 \leqslant \frac{Z_N - r}{a^r(u_N)} \leqslant 1\right\} \\ &\quad + \text{Prob}\left\{A_N^c;\ -a^r(u_N) \leqslant Z_N - r \leqslant a^r(u_N)\right\}. \end{aligned}$$

By $a^r(u_N) \to a^r(u)$ in probability and $a^r(u) > 0$ it follows that $\text{Prob}\{A_N\} \to 1$. Therefore

$$\lim_{N\to\infty} \text{Prob}\left\{A_N^c;\ -a^r(u_N) \leqslant Z_N - r \leqslant a^r(u_N)\right\} = 0.$$

Let B_N be the event $-1 \leqslant \frac{Z_N - r}{a^r(u_N)} \mathbb{1}_{a^r(u_N)>0} \leqslant 1$; then we have

$$\text{Prob}\left(B_N\right) \to \text{Prob}\left\{-1 \leqslant \frac{u(Z) - r}{a^r(u)} \leqslant 1\right\} = \text{Prob}\left\{-a^r(u) \leqslant u(Z) - r \leqslant a^r(u)\right\} = 1 - \alpha.$$

Consequently,

$$\lim_{N\to\infty} \text{Prob}\left\{-a^r(u_N) \leqslant Z_N - r \leqslant a^r(u_N)\right\} = \lim_{N\to\infty} \text{Prob}\left\{A_N \cap B_N\right\} = 1-\alpha.$$

□

Lemma 2.9. *Suppose that Assumptions 2.1, 2.2 and 2.4(a-b) hold, and the population covariance matrices Σ and Σ_0^1 are nonsingular. Let $\hat{\Theta}$ be a consistent estimator of Σ^{-1}. Then $\hat{R}$ converges to R in probability uniformly on compact sets.*

Proof of Lemma 2.9. Let

$$T = (L_K)^{-1}\begin{bmatrix} (\Sigma_0^1)^{\frac{1}{2}} & 0 \\ 0 & I_p \end{bmatrix} \quad \text{and} \quad T_N = (\Phi_N(z_N))^{-1}\begin{bmatrix} (\Sigma_N^1)^{\frac{1}{2}} & 0 \\ 0 & I_p \end{bmatrix}.$$

According to the proof of Proposition 2 in [23], we know that T_N converges to T in probability uniformly on compact sets. From Theorem 2.2, (3.50) and (3.56) one can see that $\hat{G}'$ converges to G in probability uniformly on compact sets. Hence, for any $\epsilon > 0$ we have

$$\begin{aligned}
& \lim_{N\to\infty} \text{Prob}\left\{\sup_{h\in\mathbb{R}^{2p+1},h\neq 0} \frac{||\hat{R}h - Rh||}{||h||} < \epsilon\right\} \\
\geqslant \;& \lim_{N\to\infty} \text{Prob}\left\{\sup_{h\in\mathbb{R}^{2p+1},h\neq 0} \frac{||\hat{G}'T_N h - GT_N h||}{||h||} + \sup_{h\in\mathbb{R}^{2p+1},h\neq 0} \frac{||GT_N h - GTh||}{||h||} < \epsilon\right\} \\
\geqslant \;& \lim_{N\to\infty} \text{Prob}\left\{\sup_{h\in\mathbb{R}^{2p+1},h\neq 0} \frac{||\hat{G}'(T_N h) - G(T_N h)||}{||T_N h||} \sup_{h\in\mathbb{R}^{2p+1},h\neq 0} \frac{||T_N h||}{||h||} < \frac{\epsilon}{2}\right\} + \\
& \lim_{N\to\infty} \text{Prob}\left\{||G|| \sup_{h\in\mathbb{R}^{2p+1},h\neq 0} \frac{||T_N h - Th||}{||h||} < \frac{\epsilon}{2}\right\} - 1 \\
= \;& 1,
\end{aligned}$$

i.e., $\hat{R}$ converges to R in probability uniformly on compact sets.

□

Proof of Theorem 2.8. By Lemma 2.9, $\hat{R}_j$ converges to R_j in $C(\mathbb{R}^{2p+1}, \mathbb{R})$ in probability uniformly on compact sets. Let

$$Z_N = \sqrt{N}\left((\hat{\beta}_0^{true}, \hat{\beta}^{true}) - (\beta_0^{true}, \beta^{true})\right)_j$$

for $j = 1, \cdots, p+1$. From (3.52), Z_N converges to $R_j(Z)$ in distribution. Then the conclusions follow from Lemma 2.8 with $u_N = \hat{R}_j$ and $u = R_j$.

□

In practice, for a fixed choice of r we can find the empirical individual confidence intervals for $(\beta_0^{true}, \beta^{true})$ by simulating data from $\hat{R}(Z)$. We first generate data from $\mathcal{N}(0, \Sigma_N)$, then compute $(\Phi_N(z_N))^{-1}(q)$ for a given vector q as described in Section 2.3.3, and based on that obtain an empirical distribution of $\hat{R}(q) = \hat{G}' \circ (\Phi_N(z_N))^{-1}(q)$ since $\hat{G}'$ is computable.

2.6 Numerical examples

This section contains five examples. The first four examples are based on simulated data, and we use them to illustrate the distribution of SAA solutions, examine coverage of confidence intervals and confidence bands computed from the proposed methods and algorithms. The last one uses real data from the literature. We use

$$\frac{1}{g(N)} = \frac{\min(\lambda, \theta)}{N^{1/3}}, \ \theta = 0.0001, \tag{2.78}$$

which satisfy (2.47).

2.6.1 Example 2.6.1: Asymptotic distribution of LASSO parameters

This subsection uses a small example to demonstrate the asymptotic distribution of SAA solutions. We generate data from the model $Y = \beta^{*T}X + \sigma\varepsilon$, where $\beta^* = (2, 1)$, X is a 2-dimensional normal random variable with mean 0 and covariance matrix $\Sigma = 0.5I_2 + 0.5J_2$, with I_2 being the 2×2 identity matrix and J_2 being the 2×2 matrix of 1's, ε follows $\mathcal{N}(0, 1)$, and $\sigma = 3$. Here X and ε are independent of each other.

Consider the LASSO problem with $\lambda = 3$. From the above distributions of X and Y, the first term in the objective function of (2.1) is given by

$$E[Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j]^2 = (\beta^* - \beta)^T \Sigma (\beta^* - \beta) + \beta_0^2 + \sigma^2.$$

We find the following closed-form expression for f_0 defined in (2.10) as

$$f_0(\beta_0, \beta_1, \beta_2, t_1, t_2) = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ t_1 \\ t_2 \end{bmatrix} + \begin{bmatrix} 0 \\ -5 \\ -4 \\ 3 \\ 3 \end{bmatrix}.$$

One can check that $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{t}_1, \tilde{t}_2) = (0, 1, 0, 1, 0)$ satisfies (2.12), and that $z_0 = (0, 4, 3, -2, -3)$. Note that $((z_0)_3, (z_0)_5) \in C_2^4$, and this example is one of Case II.

Specializing (3.30) to this example, we find

$$(z_N)_1 \Rightarrow \mathcal{N}(0, 12N),$$

and

$$\sqrt{N} \begin{bmatrix} 2((z_N)_2 - 4) + \max(0, (z_N)_3 - 3) \\ (z_N)_2 + (z_N)_3 - 7 + \max(0, (z_N)_3 - 3) \end{bmatrix} \Rightarrow \mathcal{N}(0, \begin{bmatrix} 57 & 33 \\ 33 & 57 \end{bmatrix}),$$

with $(z_N)_1$ being independent of $((z_N)_2, (z_N)_3)$. If we write

$$w(u, v) = \begin{bmatrix} 2(u - 4) + \max(0, v - 3) \\ u + v - 7 + \max(0, v - 3) \end{bmatrix},$$

then the set

$$\left\{ (u, v) \in \mathbb{R}^2 \mid N w(u, v)^T \begin{bmatrix} 57 & 33 \\ 33 & 57 \end{bmatrix}^{-1} w(u, v) \leq \chi_2^2(\alpha) \right\} \tag{2.79}$$

contains $((z_N)_2, (z_N)_3)$ with probability approximate $(1 - \alpha)100\%$, for any $\alpha \in (0, 1)$.

Specializing (3.29) to this example, we find that $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$, the solution to (2.2), asymptotically follows the distribution of the following random vector:

$$\left(\frac{s_1}{2\sqrt{N}}, \quad 1 + \frac{-\max(0, (2s_3 - s_2)/3) + s_2}{2\sqrt{N}}, \quad \frac{\max(0, (2s_3 - s_2)/3)}{\sqrt{N}} \right),$$

where $s = (s_1, s_2, s_3)$ is a normal random vector with covariance matrix

$$\begin{bmatrix} 48 & 0 & 0 \\ 0 & 57 & 33 \\ 0 & 33 & 57 \end{bmatrix}.$$

In particular, the probabilities for $\hat{\beta}_2$ to be exactly zero and strictly positive are both 1/2. Moreover, each of the following two regions contain $(\hat{\beta}_1, \hat{\beta}_2)$ with probability about $0.5(1 - \alpha)100\%$:

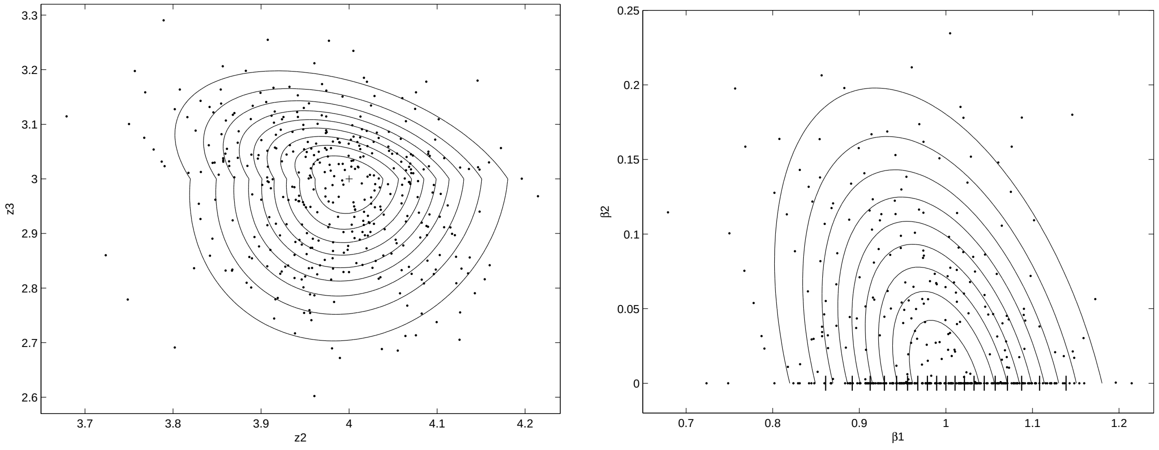
$$\left\{ (u, v) \in \mathbb{R} \times \mathbb{R}_{++} \mid N[u-1, v]^T \begin{bmatrix} 17 & -7 \\ -7 & 17 \end{bmatrix}^{-1} [u-1, v] \leq \chi_2^2(\alpha) \right\} \tag{2.80}$$

and

$$\left\{ (u, v) \in \mathbb{R} \times \{0\} \mid 1 - \sqrt{\frac{57}{4N}\chi_1^2(\alpha)} \leq u \leq 1 + \sqrt{\frac{57}{4N}\chi_1^2(\alpha)} \right\}. \tag{2.81}$$

To demonstrate the distributions graphically, we generate 400 replications, each with sample size $N = 2000$, and compute the SAA solutions z_N and $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ for each of them. The left panel of Figure 2.3 plots the 400 points $((z_N)_2, (z_N)_3)$, and also displays boundaries of regions defined in (2.79), with $\alpha = 0.1, 0.2, \cdots, 0.9$. The nine boundaries divide the plane into ten divisions, with around 40 points (min 34, max 45, mean 40, std 3.62) in each division. The true solution $((z_0)_2, (z_0)_3) = (4, 3)$ is marked with a "+" sign.

The right panel of Figure 2.3 plots the corresponding 400 points $(\hat{\beta}_1, \hat{\beta}_2)$. The curves shown in the graph are boundaries of regions defined in (2.80) with $\alpha = 0.1, 0.2, \cdots, 0.9$. Short vertical lines on the horizontal axis are markers of the endpoints of intervals defined in (2.81) with the same α values. The markers are not located at intersections between the curves and



(a) $((z_N)_2, (z_N)_3)$ (b) $(\hat{\beta}_1, \hat{\beta}_2)$

Figure 2.3: Distribution of SAA solutions in Example 2.6.1

the horizontal axis, because the two regions (2.80) and (2.81) come from different distributions. An extra short vertical line is plotted at the true solution $(\tilde{\beta}_1, \tilde{\beta}_2) = (1, 0)$. The 19 short vertical lines on the horizontal axis divide the axis into 20 intervals. There are 208 points out of the total 400 that lie on the horizontal axis, with about 10 points (min 5, max 18, mean 10.4, std 3.10) in each interval. The other 192 points lie above the horizontal axis, with about 20 points (min 15, max 25, mean 19.2, std 3.01) in each of the ten divisions divided by the nine curves.

2.6.2 Example 2.6.2: Low dimensional simulation

For this example, we simulate data using the model in Example 1 of [49]. The model is the same as that of Example 2.6.1, with $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)$. Here X is normal with mean 0 and covariance $\Sigma_{ij} = \rho^{|i-j|}$ for $\rho = 0.5$, and ε is a standard normal random variable independent of X. We set $\sigma = 1$.

We generate 100 replications, each of sample size $N = 300$, and compute two types of confidence intervals for three fixed λ values 0.5, 1, 2. The first type of confidence intervals is for the population LASSO parameters $(\tilde{\beta}_0, \tilde{\beta})$, and the second is for the true parameters $(\beta_0^{true}, \beta^{true})$ in the underlying linear model (3.46), both of significance levels α =0.1, 0.05, 0.01. For the second type intervals, we also compare our method with two other approaches in the literature. One is the LDPE method [56; 50]. The other is a recent method introduced by

[21], which we call "JM" method. We use nodewise LASSO regression introduced by [35] to compute the estimate of the precision matrix $\hat{\Theta}$ (Graphical LASSO method [54; 16] also works well), which is the same approach used in the LDPE method. Both LDPE and JM methods need to estimate the error variance σ_ε^2 in their asymptotic distributions, and they use scaled LASSO [48] to estimate it. In contrast, our method does not need to estimate σ_ε^2. In terms of the tuning parameter λ, we check the performance of our method using GIC [37] with a weight $\alpha_n = n$ in front of the penalization term of model complexity. In the LDPE method, the model parameters are estimated by scaled LASSO which does not require the specification of a tuning parameter λ. The JM method uses $\lambda = 4\hat{\sigma}_\varepsilon\sqrt{(2\log p)/n}$ as the tuning parameter, where $\hat{\sigma}_\varepsilon$ is the scaled LASSO estimator of the noise level.

	$\lambda = 0.5$			$\lambda = 1$			$\lambda = 2$		
	Est	Ind CI	Sim CI	Est	Ind CI	Sim CI	Est	Ind CI	Sim CI
β_0	-0.08	[-0.18, 0.02]	[-0.31, 0.15]	-0.10	[-0.21, 0.02]	[-0.37, 0.17]	-0.14	[-0.30, 0.03]	[-0.53, 0.25]
β_1	2.82	[2.72, 2.93]	[2.58, 3.07]	2.65	[2.52, 2.78]	[2.35, 2.95]	2.30	[2.11, 2.50]	[1.85, 2.76]
β_2	1.43	[1.32, 1.53]	[1.18, 1.67]	1.28	[1.16, 1.40]	[1.00, 1.56]	0.99	[0.81, 1.16]	[0.58, 1.40]
β_3	0	[0, 0]	[0, 0.17]	0	[0, 0]	[0, 0.12]	0	[0, 0]	[0, 0.08]
β_4	0	[0, 0]	[0, 0.22]	0	[0, 0]	[0, 0.09]	0	[0, 0]	[0, 0]
β_5	1.74	[1.64, 1.83]	[1.52, 1.96]	1.51	[1.40, 1.63]	[1.24, 1.79]	1.06	[0.88, 1.24]	[0.64, 1.48]
β_6	0	[0, 0.08]	[0, 0.31]	0	[0, 0]	[0, 0.14]	0	[0, 0]	[0, 0]
β_7	0	[0, 0.02]	[0, 0.30]	0	[0, 0]	[0, 0.04]	0	[0, 0]	[0, 0]
β_8	0	[0, 0]	[0, 0.17]	0	[0, 0]	[0, 0]	0	[0, 0]	[0, 0]

Table 2.5: 90% CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.2.

		LDPE method		JM method		$\lambda = 0.49$ tuned by GIC	
	True	Est	Ind CI	Est	Ind CI	Est	Ind CI
β_0^{true}	0	–	–	–	–	-0.06	[-0.16, 0.03]
β_1^{true}	3	3.00	[2.90, 3.11]	3.00	[2.91, 3.10]	3.00	[2.90, 3.10]
β_2^{true}	1.5	1.59	[1.48, 1.70]	1.59	[1.49, 1.69]	1.59	[1.48, 1.70]
β_3^{true}	0	-0.06	[-0.18, 0.05]	-0.08	[-0.17, 0.02]	-0.06	[-0.16, 0.04]
β_4^{true}	0	0.05	[-0.06, 0.17]	0.06	[-0.04, 0.16]	0.05	[-0.06, 0.16]
β_5^{true}	2	1.91	[1.79, 2.02]	1.91	[1.81, 2.00]	1.91	[1.79, 2.02]
β_6^{true}	0	0.08	[-0.03, 0.20]	0.08	[-0.02, 0.18]	0.08	[-0.02, 0.19]
β_7^{true}	0	0.03	[-0.09, 0.14]	0.01	[-0.09, 0.11]	0.03	[-0.08, 0.13]
β_8^{true}	0	0.03	[-0.07, 0.14]	0.03	[-0.06, 0.12]	0.03	[-0.07, 0.14]

Table 2.6: 90% individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 2.6.2.

Tables 2.5 and 2.6 show the first and second types of confidence intervals respectively, both of which are computed from a specific replication with significance level 0.1. The "Est" columns in Table 2.5 and Table 2.6 contain values of the SAA solution $(\hat{\beta}_0, \hat{\beta})$ and the true parameter estimates $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$ respectively. The "True" column in Table 2.6 contains the true parameter $(\beta_0^{true}, \beta^{true})$ and the "$\tilde{\beta}$" columns in Table 2.7 contain the solutions to the population LASSO problem $(\tilde{\beta}_0, \tilde{\beta})$ for different λ values. The "Ind CI" and "Sim CI" columns in Table 2.5 give individual and simultaneous confidence intervals respectively. The intervals are not always symmetric around the estimates, a result of the non-normality.

In Table 2.5, the value 0 appears as an endpoint for many intervals, and in some cases the entire interval shrinks to the singleton $\{0\}$. In Table 2.6, the confidence intervals for the intercept β_0 are not available for the LDPE and JM methods. In our method, there is no need to center each replication. The estimates and individual confidence intervals for true parameters computed from these three methods as shown in Table 2.6 are quite similar.

	$\lambda = 0.5$				$\lambda = 1$				$\lambda = 2$			
	$\tilde{\beta}$	α =0.1	0.05	0.01	$\tilde{\beta}$	α =0.1	0.05	0.01	$\tilde{\beta}$	α =0.1	0.05	0.01
β_0	0	93	98	100	0	92	97	100	0	89	96	100
β_1	2.83	97	97	99	2.67	95	97	100	2.33	96	99	100
β_2	1.36	90	96	100	1.22	88	94	100	0.94	86	93	98
β_3	0	100	100	100	0	100	100	100	0	100	100	100
β_4	0	100	100	100	0	100	100	100	0	100	100	100
β_5	1.78	88	95	99	1.56	90	97	100	1.11	89	97	100
β_6	0	100	100	100	0	100	100	100	0	100	100	100
β_7	0	100	100	100	0	100	100	100	0	100	100	100
β_8	0	100	100	100	0	100	100	100	0	100	100	100

Table 2.7: Coverage of the individual CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.2.

		LDPE method			JM method			Our method with GIC		
	True	α =0.1	0.05	0.01	α =0.1	0.05	0.01	α =0.1	0.05	0.01
β_0^{true}	0	–	–	–	–	–	–	93	99	100
β_1^{true}	3	92	96	98	90	92	97	92	96	98
β_2^{true}	1.5	92	94	100	85	91	98	93	96	100
β_3^{true}	0	88	95	99	92	99	100	88	95	99
β_4^{true}	0	87	95	99	96	99	100	88	95	99
β_5^{true}	2	90	96	100	90	94	97	90	95	100
β_6^{true}	0	85	90	95	85	97	100	85	91	94
β_7^{true}	0	90	95	99	95	99	100	91	94	99
β_8^{true}	0	90	96	100	93	99	100	89	97	100

Table 2.8: Coverage of the individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 2.6.2.

To test the coverage of the first type confidence intervals, we compute the population LASSO parameters $(\tilde{\beta}_0, \tilde{\beta})$ for each λ. We first check if the population LASSO parameters are contained in the high-dimensional boxes formed by the simultaneous confidence intervals. We observe 100% coverage from all SAA problems, even with simultaneous confidence intervals of significance level 0.1. This is not very surprising, since these boxes are much larger than the confidence regions of the specified probability levels enclosed in them. Next, we check if components of the population LASSO parameters are contained in the corresponding individual confidence intervals. Table 2.7 lists the numbers of individual confidence intervals that cover the population LASSO parameters, for each λ and each α. For example, the second entry in row 1 means that 93 individual confidence intervals of significance level 0.1 computed from the 100 replications with $\lambda = 0.5$ cover the population LASSO parameter $\tilde{\beta}_0$. As shown in the table, the individual confidence intervals for $\tilde{\beta}_i$ are conservative when $\tilde{\beta}_i$ equals zero with $i \neq 0$. This is consistent with the discussion in Section 2.3.4. For the second type confidence intervals, we check if components of the true parameters $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)$ in the underlying model are contained in the corresponding individual confidence intervals. Table 2.8 lists the numbers of individual confidence intervals that cover β^* for the three methods. Those three methods perform similarly and fairly well.

2.6.3 Example 2.6.3: High dimensional simulation

In this example, we consider a case in which the dimension p is larger than the sample size. The simulation model is the same as that of Example 2.6.1, with β^* being a 300-dimensional vector: $\beta^*_1 = 3$, $\beta^*_2 = \beta^*_{100} = \beta^*_{200} = \beta^*_{300} = 1.5$, $\beta^*_5 = \beta^*_{95} = 2$, and all the other components are 0. Again X is normal with mean 0 and covariance $\Sigma_{ij} = \rho^{|i-j|}$ for $\rho = 0.9$, ε is standard normal and independent of X, and $\sigma = 1$.

We generate 100 replications of sample size $N = 100$, and consider three fixed λ values, 0.5, 1 and 2, as well as the λ value chosen by GIC for each SAA problem. As in Example 2.6.2, we compute two types of individual confidence intervals both with the significance level 0.05: the first type is for the population LASSO parameters $(\tilde{\beta}_0, \tilde{\beta})$, and the second type is for the true parameters $(\beta_0^{true}, \beta^{true})$ in the underlying linear model (3.46). Define the active set as $\mathcal{A} = \{j : \beta^*_j \neq 0\} = \{1, 2, 5, 95, 100, 200, 300\}$ and $\mathcal{A}^c = \{1, 2, \cdots, p\} \backslash \mathcal{A}$. For each type of individual confidence intervals, we report the average coverage, median coverage, average length and median length of the individual confidence intervals corresponding to parameters in either $\mathcal{A}$ or $\mathcal{A}^c$:

$$
\begin{aligned}
\text{Avgcov } \mathcal{A} &= |\mathcal{A}|^{-1} \sum_{j \in \mathcal{A}} \mathrm{CP}_j, & \text{Avgcov } \mathcal{A}^c &= |\mathcal{A}^c|^{-1} \sum_{j \in \mathcal{A}^c} \mathrm{CP}_j, \\
\text{Avglen } \mathcal{A} &= |\mathcal{A}|^{-1} \sum_{j \in \mathcal{A}} \mathrm{ALen}_j, & \text{Avglen } \mathcal{A}^c &= |\mathcal{A}^c|^{-1} \sum_{j \in \mathcal{A}^c} \mathrm{ALen}_j, \\
\text{Medcov } \mathcal{A} &= \operatorname*{median}_{j \in \mathcal{A}} \{\mathrm{CP}_j\}, & \text{Medcov } \mathcal{A}^c &= \operatorname*{median}_{j \in \mathcal{A}^c} \{\mathrm{CP}_j\}, \\
\text{Medlen } \mathcal{A} &= \operatorname*{median}_{j \in \mathcal{A}} \{\mathrm{ALen}_j\}, & \text{Medlen } \mathcal{A}^c &= \operatorname*{median}_{j \in \mathcal{A}^c} \{\mathrm{ALen}_j\},
\end{aligned}
$$

where CP_j and ALen_j respectively represent the empirical coverage probability and average interval length of the confidence intervals for β_j among the 100 replications (see Tables 3.4 and 2.10). For the second type intervals, we compare the above measures computed from our method with those from the LDPE and JM methods (Table 2.10).

	$\lambda = 0.5$				$\lambda = 1$				$\lambda = 2$			
	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen
$\mathcal{A}$	91.86	94.00	0.92	0.92	91.57	94.00	1.17	1.18	90.43	92.00	1.59	1.65
$\mathcal{A}^c$	99.92	100.00	0.07	0.06	99.96	100.00	0.04	0.02	99.96	100.00	0.04	0.01

Table 2.9: Coverage and length of 95% individual CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.3.

As shown in Table 3.4, the first type confidence intervals are often conservative for the inactive variables. The same phenomena is observed in Example 2.6.2. On the other hand, the interval lengths for the inactive variables are very short compared to the lengths for active variables. This is related to the nature of LASSO: For a large λ, many population LASSO parameters are exactly 0's. Thus the SAA solutions of LASSO of these parameters concentrate closely around 0's. This fact also leads to shorter confidence intervals for true parameters of the inactive set, as will be discussed later.

Our	$\lambda = 0.5$				$\lambda = 1$				$\lambda = 2$			
method	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen
$\mathcal{A}$	93.86	94.00	0.81	1.03	95.86	97.00	1.08	1.39	95.86	98.00	1.72	2.28
$\mathcal{A}^c$	92.85	93.00	0.75	0.74	93.44	94.00	1.04	1.02	93.81	94.00	1.71	1.69
	LDPE method				JM method				Our method with GIC			
	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen
$\mathcal{A}$	88.43	89.00	1.04	1.06	84.71	83.00	0.84	0.86	93.14	94.00	1.03	1.02
$\mathcal{A}^c$	95.13	95.00	1.07	1.07	98.94	99.00	0.87	0.87	92.61	93.00	0.72	0.72

Table 2.10: Coverage and length of 95% individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 2.6.3.

The top rows of Table 2.10 report results of our method with different λ values. One may notice that the length of confidence intervals for the true parameters $(\beta_0^{true}, \beta^{true})$ increases when λ increases. For an intuitive explanation, recall that the estimator $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$ in (3.49) is a bias correction version of the LASSO solution $(\hat{\beta}_0, \hat{\beta})$. Large λ brings the LASSO solution close to zero, which causes an increase of the correction part, and the latter leads to wide confidence intervals. On the other hand, if λ is too small, the SAA solution lacks sparsity and the corresponding LASSO estimates are less reliable. This suggests choosing an intermediate value of λ to achieve the best overall performance.

The bottom rows of Table 2.10 show the results calculated from the LDPE method, the JM method and our method (with λ chosen by GIC) respectively. For the active variables, our method performs considerably better than the other two. For the inactive variables, the coverage from the LDPE method is closest to 95%, and the coverage from the JM method are even better. However, their confidence intervals are comparatively wider than those from our method on average. The coverage for inactive variables computed from our method is in line with the coverage for active variables, and they both will be better with larger sample size. Moreover, with the same significance level, intuitively the confidence intervals of the inactive variables can be narrower than the confidence intervals of the active variables on average, because the involved prediction error of a model parameter with large magnitude is larger than that of a model parameter with small magnitude.

2.6.4 Example 2.6.4: Coverage test for confidence bands

In this example, the simulation model is the same as that of Example 2.6.1, with β^* being a 100-dimensional vector: β_1^*, β_{31}^*, β_{61}^* and β_{91}^* are 3; β_2^*, β_{20}^*, β_{38}^*, β_{56}^*, β_{74}^* and β_{92}^* are 1; β_5^*, β_{15}^*, β_{45}^* and β_{70}^* are 2; all the other components are 0. We generate 100 replications and set $N = 300$, $\rho = 0.5$, $\sigma = 3$. Using Algorithm 3, we compute 95% individual confidence intervals for z_0 and $(\tilde{\beta}_0, \tilde{\beta})$ at 25 values of λ as $\frac{6.1}{25}i$, $i = 1, \cdots, 25$. The confidence intervals for z_0 can be obtained by dropping Step 4 in Algorithm 3.

To show the overall performance of coverage, we draw a boxplot for coverage rates in the 101 coordinates at each λ value. $\Phi_N(z_N)$ is close to being singular when λ is small. With λ increasing, more and more coordinates of $\gamma(z_N)$ become 7. Therefore $(L_N)_{\mathfrak{L}}$ is more likely

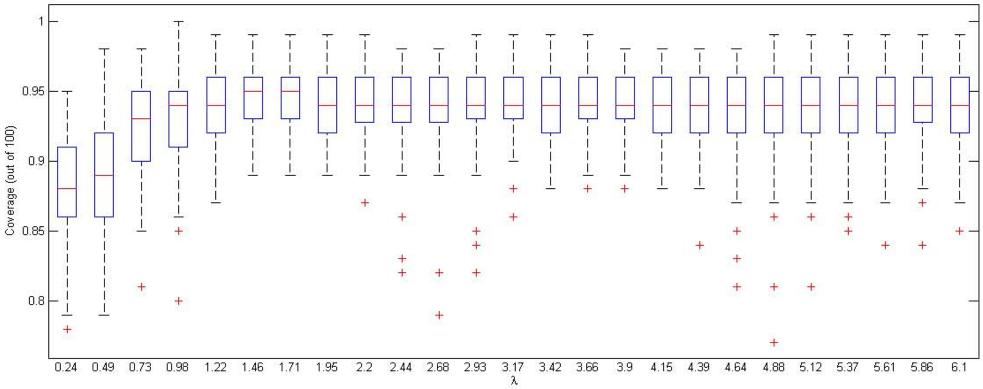


Figure 2.4: Boxplot of coverage rates of 95% individual CIs for z_0 in Example 2.6.4

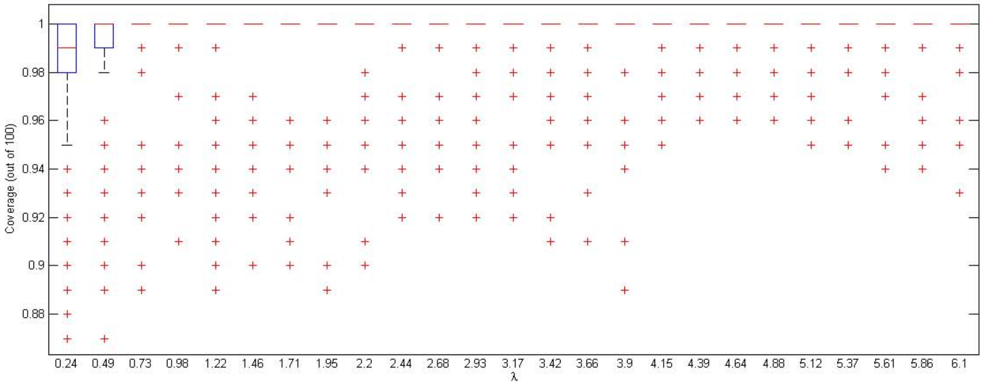


Figure 2.5: Boxplot of coverage rates of 95% individual CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.4

to be nonsingular with large value of λ. This is a general phenomenon. When λ becomes infinity, each $\hat{\beta}_i$ $(i = 1, 2, \cdots, p)$ will decrease to 0. Consequently, from Lemma 3.1 and (2.78), $((z_N)_{i+1}, (z_N)_{i+1+p})$ will approach to E_i^7.

From Figure 2.5, we note that coverage rates in most of coordinates for $(\tilde{\beta}_0, \tilde{\beta})$ tend to be 1 when λ increases. This can be explained by the fact that the affect of projector Γ (3.42) becomes clearer when λ increases. The larger the tuning parameter λ is, the more coordinates of $\tilde{\beta}_i$ are equal to 0.

2.6.5 Example 2.6.5: Prostate cancer data

This subsection considers the prostate cancer example used in [19]. There are eight covariates, log cancer volume, log prostate weight, age, log of the amount of benign prostatic hyperplasia, seminal vesicle invasion, log of capsular penetration, Gleason score, and percent of Gleason scores 4 or 5. The parameters corresponding to these covariates are denoted by $\beta_1, \beta_2, \cdots, \beta_8$. We standardize the data and split observations into two parts. One part consists of 67 observations, which is the training set in [19]. We only use these 67 observations in our computation. Table 2.11 shows simultaneous and individual confidence intervals for population LASSO parameters of significance level 0.05 for λ values $0.45, 0.88$ and 1.49. The value $\lambda = 0.45$ corresponds to $s \approx 0.36$, where s is the standardized turning parameter involved with an alternative formulation of LASSO defined in Section 3.4.2 of [19], and the value $s \approx 0.36$ was chosen in [19] by 10-fold cross-validation. The value $\lambda = 0.88$ is tuned by GIC as in Example 2, and the value $\lambda = 1.49$ is chosen to represent large λ values. For the true model parameters, we compare the confidence intervals computed from our method with those from the LDPE and JM methods, as shown in Table 2.12.

	$\lambda = 0.45$			$\lambda = 0.88$ tuned by GIC			$\lambda = 1.49$		
	Est	Ind CI	Sim CI	Est	Ind CI	Sim CI	Est	Ind CI	Sim CI
β_0	2.47	[2.29, 2.65]	[2.08, 2.85]	2.47	[2.25, 2.68]	[2.01, 2.92]	2.46	[2.20, 2.72]	[1.92, 3.00]
β_1	0.53	[0.30, 0.77]	[0.05, 1.02]	0.42	[0.20, 0.65]	[0, 0.90]	0.16	[0, 0.44]	[0, 0.75]
β_2	0.18	[0.02, 0.33]	[0, 0.50]	0.05	[0, 0.22]	[0, 0.41]	0	[0, 0.13]	[0, 0.63]
β_3	0	[0, 0]	[-0.20, 0.37]	0	[0, 0]	[0, 0.32]	0	[0, 0]	[0, 0.20]
β_4	0	[0, 0.30]	[0, 0.66]	0	[0, 0.02]	[0, 0.43]	0	[0, 0]	[0, 0.13]
β_5	0.08	[0, 0.32]	[0, 0.59]	0	[0, 0.28]	[0, 0.69]	0	[0, 0.09]	[0, 0.57]
β_6	0	[0, 0]	[0, 0.22]	0	[0, 0]	[0, 0.23]	0	[0, 0]	[0, 0.16]
β_7	0	[0, 0.10]	[0, 0.40]	0	[0, 0]	[0, 0.13]	0	[0, 0]	[0, 0]
β_8	0	[0, 0.27]	[0, 0.57]	0	[0, 0.13]	[0, 0.52]	0	[0, 0]	[0, 0.30]

Table 2.11: 95% CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.5.

Table 2.11 shows how the confidence intervals for population LASSO parameters change as λ changes. In particular, when λ takes the value of 0.45, the individual confidence intervals of β_1 and β_2 do not contain zero, while the individual confidence intervals of all other variables (except β_0) include zero in them. This suggests that the first two predictors are the most useful ones in predicting the response. Furthermore, although the LASSO estimator for β_5 is not 0, its interval contains 0 and indicates that the corresponding LASSO parameter is not significantly different from 0. When λ becomes 0.88, the individual confidence interval of β_1 does not contain zero, while the individual confidence intervals of all other variables (except β_0) include zero in them. This change suggests that the first predictor is more important than the second one. At $\lambda = 1.49$, all the individual confidence intervals (except those of β_0) include zero in them, reflecting the shrinking feature of LASSO. Some of the confidence intervals are singletons that contains only zero, which implies that the corresponding variables are not important in predicting the response. The confidence intervals of β_0 are insensitive with changes of λ. Overall, we can compute confidence intervals for population LASSO parameters for a wide range of λ, to obtain information not only about the significance at a particular λ but also the relatively importance of the predictors.

	LDPE method		JM method		$\lambda = 0.88$ tuned by GIC		$\lambda = 0.45$		$\lambda = 1.49$	
	Est	Ind CI	Est	Ind CI	Est	Ind CI	Est	Ind CI	Est	Ind CI
β_1^{true}	0.69	[0.46, 0.93]	0.68	[0.03, 1.33]	0.71	[0.41, 1.01]	0.70	[0.44, 0.95]	0.74	[0.37, 1.11]
β_2^{true}	0.28	[0.09, 0.46]	0.26	[-0.22, 0.75]	0.29	[0.08, 0.49]	0.28	[0.10, 0.46]	0.32	[0.09, 0.55]
β_3^{true}	-0.09	[-0.29, 0.11]	-0.14	[-0.66, 0.38]	-0.08	[-0.34, 0.17]	-0.09	[-0.29, 0.10]	-0.02	[-0.35, 0.31]
β_4^{true}	0.21	[0.01, 0.41]	0.21	[-0.31, 0.73]	0.22	[-0.02, 0.45]	0.21	[-0.00, 0.42]	0.22	[-0.05, 0.49]
β_5^{true}	0.31	[0.08, 0.54]	0.31	[-0.33, 0.94]	0.33	[0.04, 0.63]	0.31	[0.04, 0.58]	0.38	[0.05, 0.71]
β_6^{true}	-0.21	[-0.48, 0.06]	-0.29	[-1.08, 0.50]	-0.19	[-0.47, 0.09]	-0.21	[-0.45, 0.04]	-0.10	[-0.41, 0.21]
β_7^{true}	-0.01	[-0.27, 0.25]	-0.02	[-0.76, 0.72]	-0.02	[-0.28, 0.24]	-0.01	[-0.24, 0.22]	0.04	[-0.25, 0.32]
β_8^{true}	0.24	[-0.03, 0.51]	0.27	[-0.52, 1.05]	0.25	[-0.06, 0.56]	0.24	[-0.01, 0.48]	0.28	[-0.05, 0.61]

Table 2.12: 95% individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 2.6.5.

Table 2.12 lists the estimates of the true model parameters and their individual confidence intervals, computed from the LDPE and JM methods as well as our methods with $\lambda = 0.88$, 0.45 and 1.49. The estimate of the precision matrix $\hat{\Theta}$ is computed by nodewise LASSO except for the JM method (the JM method uses its own procedure). Results from the three methods are generally comparable, except that confidence intervals computed from the JM method are overall wider than the other intervals. Based on results in Table 2.12, confidence intervals for β_1, β_2, β_4 and β_5 from the LDPE method do not contain zero. In contrast, the only confidence interval that does not contain zero from the JM method is the one for β_1. Across all three values of λ, our methods always select β_1, β_2 and β_5 with their confidence intervals not covering zero. For comparison, the 95% ordinary least squares regression confidence intervals for the true parameters are quite similar to confidence intervals computed from LDPE and our method.

We also construct 95% confidence bands for population LASSO parameters. In Figure 2.6,

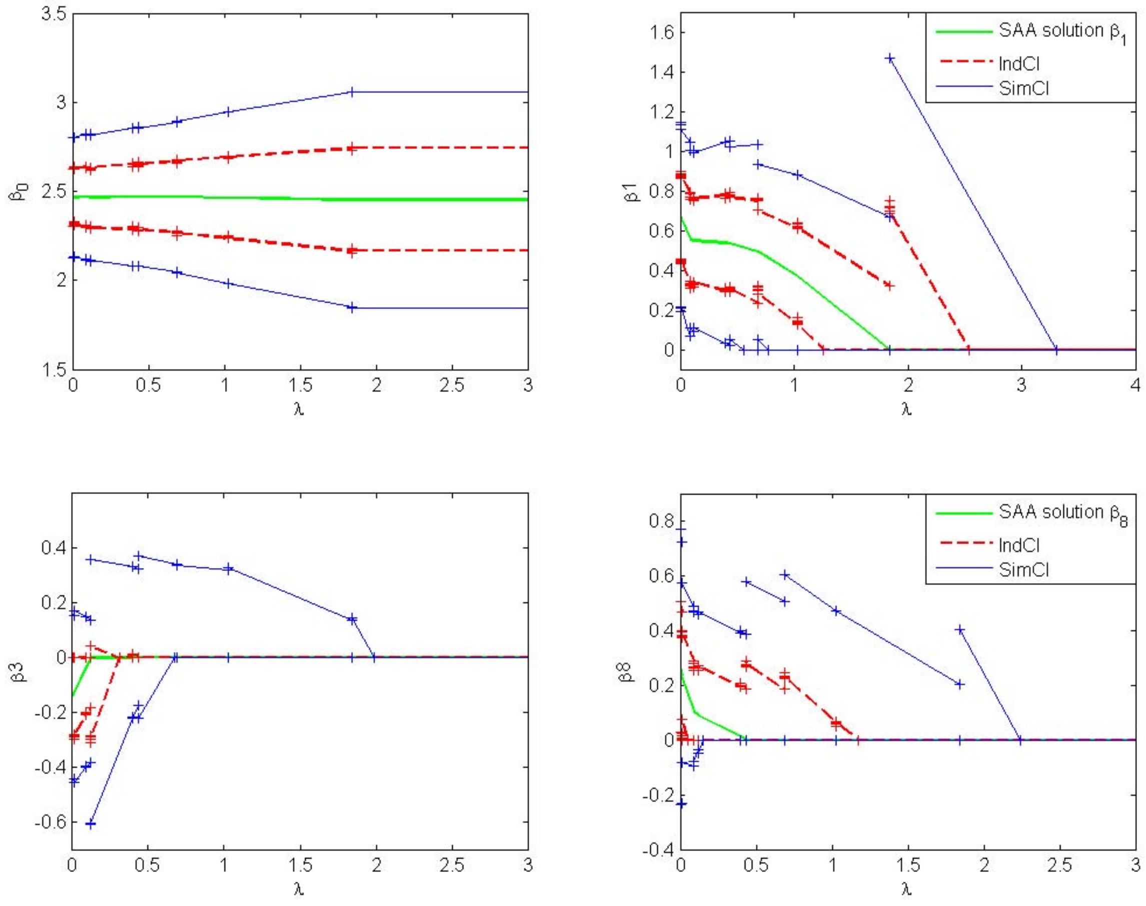


Figure 2.6: 95% Confidence bands for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 2.6.5.

the simultaneous and individual confidence bands for some selected components of β are showed by blue line and red dashed line respectively. We mark the end points of λ segments by "+". The green line represents the LASSO solution path of (2.2) for corresponding β components. Each confidence band consists of 26 λ segments. The behavior of the confidence bands in Figure 2.6 is similar to that of simulation Example 2 in Figure 3.

As expected, individual confidence bands are narrower than simultaneous ones. Each confidence band consists of 26 λ segments, some of which are very short. We mark the end points by "+". Although there are "jumps" at some end points, every confidence band will eventually shrink to zero except that for β_0. This is expected since the true solution of (3.60) goes to zero except $\tilde{\beta}_0$ when λ increases. For β_0, its confidence band does not change if λ is large enough, since the solution $\tilde{\beta}_0(\lambda)$ remains at a fixed value for all large λs.

2.7 Summary

In this chapter, we consider a prevalent sparse penalized regression: the LASSO regression. We transform LASSO problems into variational inequalities and make use of the asymptotic convergence results to derive confidence intervals and regions for the population LASSO parameters. In view of (2.44), the lengths of confidence intervals for population LASSO parameters are affected by two factors. The first is Σ_N, the sample covariance of $F(\hat{\beta}_0, \hat{\beta}, \hat{t}, x_i, y_i)\}_{i=1}^N$. The second is $(\Phi_N(z_N))^{-1}$, which characterizes the sensitivity of solution to (2.2) with respect to random samples. In general, large variance and high sensitivity lead to wide confidence intervals, and small variance and low sensitivity lead to short intervals. Thus, the lengths of confidence intervals for population LASSO parameters reflect the effect of sample variance on the parameter estimates computed from the LASSO. In terms of the true parameters in the underlying linear model, we also propose methods to derive confidence intervals and compare them with existing methods in the literature. Both our theoretical and numerical results confirm the validity and effectiveness of the proposed methods.

Moreover, we study the confidence bands for the population LASSO parameters along the LASSO solution path. We point out that the entire confidence band is neither piecewise linear nor continuous in λ, if we construct confidence band pointwisely by using techniques described in Section 2.3. We propose the linear approximation tracking algorithm in Section 2.4.4 to compute confidence intervals. Theoretically, we justify this algorithm by proving the piecewise Lipschitz property for both individual and simultaneous confidence bands under some mild conditions. Besides, we develop a sufficient and necessary condition in Section 2.4.1 to check the global homomorphism for $\Phi_N(z_N)$, which is crucial for construction of confidence intervals. Corollary 2.1 is a more convenient sufficient condition. As long as (2.47) holds, the global homomorphism of $\Phi_N(z_N)$ implies that the matrix K defined in (3.17) is nonsingular. Furthermore, Σ_N^1 is nonsingular for sufficiently large N. Both are indispensable for using (2.56) to construct simultaneous confidence interval. Finally, we want to point out that Theorems 2.4 and 2.5 do not cover the cases using simulation to find individual confidence intervals and using pseudo-inverse instead of $(\Sigma_N^1)^{-1}$ in (3.44). These two cases deserve further theoretical study in the future.

CHAPTER 3: INFERENCE FOR GENERAL PENALIZED REGRESSIONS

3.1 Introduction

In this chapter, we study a generalization of the methods discussed in Chapter 2 for the LASSO confidence intervals at fixed values of tuning parameters. We propose similar methods to construct confidence intervals for penalized regression parameters for a wide range of penalties, including commonly used penalties such as LASSO and MCP. The requirements for the penalties are consistent with the three properties proposed by [15].

We consider a general population penalized regression problem

$$\min_{\beta_0,\beta} E\big[Y-\beta_0-\sum_{i=1}^{p}\beta_i X_i\big]^2+\sum_{j=1}^{p}P_{\lambda_j}(|\beta_j|). \tag{3.1}$$

For $j = 1, 2, \cdots, p$, $P_{\lambda_j}(|\cdot|)$ is a general penalty for β_j with the regularization parameter λ_j. This general penalty covers the L_1 penalty, the adaptive LASSO penalty [59], or any non-convex penalty such as SCAD or MCP. The conditions of the penalties discussed in this chapter are listed in Section 3.2. We denote the solution of 3.1 as $(\tilde{\beta}_0, \tilde{\beta})$, which we refer to as the population penalized parameters. The solution of (3.1) can be estimated by the solution of the corresponding SAA problem

$$\min_{\beta_0,\beta} \frac{1}{N}\big|\big|\mathbf{y}-\beta_0 1_N-\mathbf{X}\beta\big|\big|_2^2+\sum_{j=1}^{p}P_{\lambda_j}(|\beta_j|), \tag{3.2}$$

where

$$\mathbf{y} \triangleq \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \mathbf{X} \triangleq \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^1 \\ \mathbf{x}^2 \\ \vdots \\ \mathbf{x}^N \end{bmatrix}, \quad 1_N = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^N,$$

and $(\mathbf{x}^1, y_1), \cdots, (\mathbf{x}^N, y_N)$ are independent samples of (X, Y). For each $i = 1, \cdots, N$ and $j = 1, \cdots, p$, $x_{ij} \in \mathbb{R}$, $y_i \in \mathbb{R}$ and $\mathbf{x}^i \in \mathbb{R}^{1\times p}$. We denote its solution as $(\hat{\beta}_0, \hat{\beta})$

In Section 3.2, we state the assumptions and the problem transformations used in this chapter. Sections 3.3 and Section 3.4 discusses how to obtain the confidence intervals for the population penalized parameters and the true model parameters respectively. To illustrate the performance of the proposed method, numerical results are presented in Section 3.5.

3.2 Problem transformations

3.2.1 Transformations of the population penalized regression

In this subsection, we change the appearance of the optimization problem (3.1) gradually to obtain its normal map formulation. Before penetrating into the process, we propose conditions on the penalties $P_{\lambda_i}(\cdot)$.

Assumption 3.1. *(a) For each $i = 1, 2, \cdots, p$, $P_{\lambda_i}(\cdot)$ is nonnegative, nondecreasing and continuously differentiable on $[0, +\infty)$ with $P'_{\lambda_i}(0) > 0$.*

(b) For any optimal solution $(\tilde{\beta}_0, \tilde{\beta})$ (local or global) to (3.1), the second derivative of $P_{\lambda_i}(t_i)$ is Lipchitz continuous on a neighborhood of $t_i = |\tilde{\beta}_i|$ for every i from 1 to p.

Most well-known non-convex penalties satisfy Assumption 3.1(a), as well as convex ones. We list four penalty families as examples.

(a) The adaptive LASSO penalty [59] defined as $P_{\lambda_i}(\beta_i) = \lambda_i|\beta_i|$, where λ_i is the weight for the i^{th} coordinate.

(b) The combination of power penalties, such as elastic net penalty [60] given by $P_\lambda(\beta_i) = \lambda_1|\beta_i| + \lambda_2|\beta_i|^2$.

(c) The SCAD penalty [15] defined via $P_\lambda(0) = 0$ and

$$P'_\lambda(\beta_i) = \lambda \mathbb{1}_{|\beta_i|\leqslant\lambda} + \frac{(a\lambda - |\beta_i|)_+}{a-1}\mathbb{1}_{|\beta_i|>\lambda} \text{ for } a > 2. \tag{3.3}$$

(d) The MCP penalty [55] defined as

$$P_\lambda(\beta_i) = \lambda(|\beta_i| - \frac{\beta_i^2}{2a\lambda})\mathbb{1}_{|\beta_i|<a\lambda} + \frac{a\lambda^2}{2}\mathbb{1}_{|\beta_i|\geqslant a\lambda} \text{ for } a > 0. \tag{3.4}$$

Assumption 3.1(b) is a mild condition for most of penalties. Take SCAD penalty for example. It corresponds to a quadratic spline with two knots, at which it is not continuously twice differentiable. Assumption 3.1(b) requires no optimal solution to (3.1) locates at these two knots for each i. It is not a strong assumption in the sense that the set on which Assumption 3.1(b) does not hold has measure zero.

In the assumption below, part (a) is to ensure the objective function of (3.1) to be finite valued, and part (b) will be used in proving convergence results.

Assumption 3.2. *(a) The expectations $E[X_1^2], \cdots, E[X_p^2]$ and $E[Y^2]$ are finite.*

(b) The expectations $E[X_1^4], \cdots, E[X_p^4]$ are finite.

Next, we are going to transform the problem (3.1) to a normal map formulation by three steps. First, we introduce an equivalent problem, in which a new variable $t \in \mathbb{R}^p$ is employed to eliminate the non-smooth term $\sum_{i=1}^p P_{\lambda_i}(|\beta_i|)$ from the objective function (3.1). This new problem is presented as follows:

$$\begin{array}{ll} \min\limits_{\beta_0,\beta,t} & \mathrm{E}\big[Y - \beta_0 - \sum_{i=1}^p \beta_i X_i\big]^2 + \sum_{i=1}^p P_{\lambda_i}(t_i) + m(||t||_2^2 - ||\beta||_2^2) \\ \text{s.t.} & \qquad t_i - \beta_i \geqslant 0, \quad i = 1, \cdots, p, \\ & \qquad t_i + \beta_i \geqslant 0, \quad i = 1, \cdots, p, \end{array} \tag{3.5}$$

where m is a non-negative constant. If we define $S_i \subset \mathbb{R}^2$ as

$$S_i = \{(\beta_i, t_i) \mid t_i - \beta_i \geqslant 0,\ t_i + \beta_i \geqslant 0\}, \quad i = 1, \cdots, p. \tag{3.6}$$

and write

$$(\beta_0, \beta, t) = (\beta_0, \beta_1, t_1, \beta_2, t_2, \cdots, \beta_p, t_p) \tag{3.7}$$

then we can treat the feasible set of (3.5), denoted by S, as a Cartesian product

$$S = \mathbb{R} \times \Pi_{i=1}^{p} S_i. \tag{3.8}$$

We will use two ways of ordering in (β_0, β, t) as showed in (3.7) interchangeably for notational convenience.

We can choose $m = 0$ if the penalty functions $P_{\lambda_i}(\cdot)$ are all strictly increasing on $[0, +\infty)$ such as Lasso penalties, otherwise we must use a positive m. In general, under Assumption 3.1(a), the third term with any positive coefficient m can guarantee problems (3.1) and (3.5) to be equivalent in the sense that there is an one-to-one correspondence between the optimal solutions of the two problems. It is worth to note that the third term in the objective of (3.5) is necessary for the case that the penalties are not strictly increasing on $[0, +\infty)$. For instance, some non-convex penalties such as SCAD and MCP are "flat" when the variables are larger than some positive thresholds, say d_i for the i^{th} penalty $(i = 1, \cdots, p)$. In other words, $P_{\lambda_i}(t_i)$ takes the same value on $[d_i, +\infty)$ for each i. Without the third term in the objective of (3.5), if $(\tilde{\beta}_0, \tilde{\beta})$ is an optimal solution to (3.1) and $|\tilde{\beta}_i| \geqslant d_i$ for some i, then $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is an optimal solution to (3.5) for all $\tilde{t}_i \geqslant |\tilde{\beta}_i|$. Therefore, without specification we assume $m > 0$ in this chapter (We use $m = \frac{1}{2}$ in the numerical examples).

Second, we transform problem (3.5) into a variational inequality formulation. To this end, we need to write down the gradient of its objective function. Let $P(t) = \Big(P'_{\lambda_1}(t_1), \cdots, P'_{\lambda_p}(t_p)\Big)^T$. Define a function $F : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^{2p+1}$ as

$$F(\beta_0, \beta, t, X, Y) = \begin{bmatrix} -2(Y - \beta_0 - \sum_{i=1}^{p} \beta_i X_i) \\ -2(Y - \beta_0 - \sum_{i=1}^{p} \beta_i X_i)X - 2m\beta \\ P(t) + 2mt \end{bmatrix}. \tag{3.9}$$

Furthermore, we define $f_0 : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{2p+1}$ as

$$f_0(\beta_0, \beta, t) = \mathrm{E}[F(\beta_0, \beta, t, X, Y)]. \tag{3.10}$$

f_0 is well defined and finite valued under Assumption 3.2(a). If $P_{\lambda_i}(t_i)$ is twice differentiable at t_i for every i from 1 to p, then we can write down the derivative of F w.r.t. (β_0, β, t) as

$$d_1F(\beta_0, \beta, t, X, Y) = \begin{bmatrix} 2 & 2X^T & 0 \\ 2X & 2XX^T - 2mI_p & 0 \\ 0 & 0 & \nabla P(t) + 2mI_p \end{bmatrix}, \tag{3.11}$$

where

$$\nabla P(t) = \begin{bmatrix} P''_{\lambda_1}(t_1) & & \\ & \ddots & \\ & & P''_{\lambda_p}(t_p) \end{bmatrix} \tag{3.12}$$

and I_p is the $p \times p$ identity matrix. Moreover, we can write down the Jacobian matrix of f_0 as

$$L(t) = E[d_1F(\beta_0, \beta, t, X, Y)] = \begin{bmatrix} 2 & 2E[X^T] & 0 \\ 2E[X] & 2E[XX^T] - 2mI_p & 0 \\ 0 & 0 & \nabla P(t) + 2mI_p \end{bmatrix}. \tag{3.13}$$

The lemma below shows that there is an one-to-one correspondence between the optimal solutions of problems (3.1) and (3.5).

Lemma 3.1. *Suppose Assumption 3.1(a) and 3.2(a) hold. If $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is an (local) optimal solution to (3.5), then $\tilde{t}_i = |\tilde{\beta}_i|$ for all i from 1 to p, and $(\tilde{\beta}_0, \tilde{\beta})$ is an (local) optimal solution to (3.1). Conversely, if $(\tilde{\beta}_0, \tilde{\beta})$ is an (local) optimal solution to (3.1), then $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is an (local) optimal solution to (3.5), where $\tilde{t}_i = |\tilde{\beta}_i|$ for all i from 1 to p.*

Moreover, the objective function of (3.5) is a finite valued function on $\mathbb{R}^{2p+1}$, and its gradient at each $(\beta_0, \beta, t) \in \mathbb{R}^{2p+1}$ is $f_0(\beta_0, \beta, t)$. In addition, if Assumption 3.1(b) also holds, then its Hessian matrix at $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is $L(\tilde{t})$.

Proof of Lemma 3.1. Without loss of generality, suppose $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a local optimal solution to (3.5). Since $P_{\lambda_i}(\cdot)$ is nondecreasing and m is positive, it is obvious that $\tilde{t}_i = |\tilde{\beta}_i|$ for all i from 1 to p. Denote the objective function in (3.1) by $g_1(\beta_0, \beta)$ and the objective function in

(3.5) by $g_2(\beta_0, \beta, t)$. Then there exists a neighborhood $\mathcal{B}_1$ at $(\tilde{\beta}_0, \tilde{\beta})$ in $\mathbb{R}^{p+1}$, such that

$$g_2(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \leqslant g_2(\beta_0, \beta, t) \quad \text{for } \forall(\beta_0, \beta) \in \mathcal{B}_1 \text{ and } t_i = |\beta_i|,\ i = 1 \cdots, p.$$

That is,

$$g_1(\tilde{\beta}_0, \tilde{\beta}) \leqslant g_1(\beta_0, \beta) \quad \text{for } \forall(\beta_0, \beta) \in \mathcal{B}_1.$$

Therefore, $(\tilde{\beta}_0, \tilde{\beta})$ is a local optimal solution to (3.1).

Conversely, suppose $(\tilde{\beta}_0, \tilde{\beta})$ is a local optimal solution to (3.1). Then there exists a neighborhood $\mathcal{B}_2$ at $(\tilde{\beta}_0, \tilde{\beta})$ in $\mathbb{R}^{p+1}$, such that

$$g_1(\tilde{\beta}_0, \tilde{\beta}) \leqslant g_1(\beta_0, \beta) \quad \text{for } \forall(\beta_0, \beta) \in \mathcal{B}_2.$$

Let $\tilde{t}_i = |\tilde{\beta}_i|$ for all i from 1 to p, then we have

$$g_2(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \leqslant g_2(\beta_0, \beta, t) \quad \text{for } \forall(\beta_0, \beta) \in \mathcal{B}_2 \text{ and } t_i = |\beta_i|,\ i = 1 \cdots, p.$$

Consequently,

$$g_2(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \leqslant g_2(\beta_0, \beta, t) \quad \text{for } \forall(\beta_0, \beta) \in \mathcal{B}_2 \text{ and } \forall t_i \geqslant |\beta_i|,\ i = 1 \cdots, p.$$

Thus, $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a local optimal solution to (3.5).

The second part of Lemma 3.1 is straightforward and we omit its proof.

□

In view of Lemma 3.1, we can transform (3.5) to the following variational inequality:

$$-f_0(\beta_0, \beta, t) \in N_S(\beta_0, \beta, t). \tag{3.14}$$

Third, we state the normal map formulation for (3.14). Let $(f_0)_S$ be the normal map induced by f_0 and S. Then the normal map formulation for (3.14) is

$$(f_0)_S(z) = 0, \quad z \in \mathbb{R}^{2p+1}. \tag{3.15}$$

Let $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ be an (local) optimal solution to (3.5), then $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is also a solution to (3.14). So the point $z_0 \in \mathbb{R}^{2p+1}$ defined as

$$z_0 = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) - f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \tag{3.16}$$

is a solution to (3.15) and satisfies $\Pi_S(z_0) = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$. Let K be the *critical cone* to S associated with z_0, defined as

$$\begin{aligned} K &= \{w \in T_S(\Pi_S(z_0)) \mid \langle z_0 - \Pi_S(z_0), w\rangle = 0\} \\ &= \{w \in T_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \mid \langle f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}), w\rangle = 0\}. \end{aligned} \tag{3.17}$$

At last, we introduce the third assumption and the second lemma.

Assumption 3.3. *Let $(\tilde{\beta}_0, \tilde{\beta})$ be a locally optimal solution of (3.1), define $\tilde{t} \in \mathbb{R}^p$ and $\tilde{q} \in \mathbb{R}^p$ by*

$$\tilde{t}_i = |\tilde{\beta}_i| \text{ and } \tilde{q}_i = E[-2(Y - \tilde{\beta}_0 - \sum_{j=1}^{p} \tilde{\beta}_j X_j)X_i] \text{ for each } i = 1, \cdots, p.$$

Let $\mathfrak{I}$ be a subset of $\{1, \cdots, p\}$ defined as

$$\mathfrak{I} = \left\{ i \in \{1, \cdots, p\} \mid \tilde{\beta}_i \neq 0 \text{ or } (\tilde{\beta}_i = 0 \text{ and } |\tilde{q}_i| = |P'_{\lambda_i}(\tilde{t}_i)|) \right\},$$

and denote $L(\tilde{t})$ in (3.13) by L. Let Q_1 be the submatrix of L that consists of intersections of columns and rows of L with indices in $\{1\} \cup \{i+1, i \in \mathfrak{I}\}$, and let Q_2 be the submatrix of L that consists of intersections of columns and rows of L with indices in $\{i+p+1, i \in \mathfrak{I}\}$. Define matrix Q as

$$Q = Q_1 + \begin{bmatrix} 0 & 0 \\ 0 & Q_2 \end{bmatrix}. \tag{3.18}$$

Assume that Q is nonsingular.

In the above assumption, Q_1 is a submatrix of the upper left $(p+1) \times (p+1)$ submatrix of L, and Q_2 is a submatrix of the lower right $p \times p$ submatrix of L. It is well known that L_K, the normal map induced by L and K in (3.17), is the same as the B-derivative of the normal map

$(f_0)_S$ at z_0 [40]. L_K and its estimator play important roles in this chapter. Lemma 3.2 below shows that L_K is a global homeomorphism from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$, that is, a continuous bijective function from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$ whose inverse function is also continuous.

Lemma 3.2. *Suppose that Assumptions 3.1, 3.2(a) and 3.3 hold. Then the normal map L_K is a global homeomorphism from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$, and $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a locally unique optimal solution to* (3.5)*, where $\tilde{t}_i = |\tilde{\beta}_i|$ for all i from 1 to p.*

Proof of Lemma 3.2. According to Assumption 3.3 and Lemma 3.1 we know that $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a locally optimal solution to (3.5). We will prove it is also a locally unique optimal solution by showing that L_K is a global homeomorphism.

From(3.8), we can write the normal and tangent cone to S at $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ as

$$N_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = \{0\} \times N_{S_1}(\tilde{\beta}_1, \tilde{t}_1) \times \cdots \times N_{S_p}(\tilde{\beta}_p, \tilde{t}_p),$$

and

$$T_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = \mathbb{R} \times T_{S_1}(\tilde{\beta}_1, \tilde{t}_1) \times \cdots \times T_{S_p}(\tilde{\beta}_p, \tilde{t}_p).$$

Let $\tilde{q}$ be as defined in Assumption 3.3, and let $\tilde{q}_0 = E[-2(Y - \tilde{\beta}_0 - \sum_{j=1}^{p} \tilde{\beta}_j X_j)]$. Since $-f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \in N_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, we have

$$\tilde{q}_0 = 0 \text{ and } -(\tilde{q}_i - 2m\tilde{\beta}_i, P'_{\lambda_i}(\tilde{t}_i) + 2m\tilde{t}_i) \in N_{S_i}(\tilde{\beta}_i, \tilde{t}_i) \text{ for each } i = 1, \cdots, p. \tag{3.19}$$

If $\tilde{\beta}_i > 0$ for some $i = 1, \cdots, p$, from the definition of S_i and (3.19) we have

$$\tilde{q}_i - 2m\tilde{\beta}_i = -P'_{\lambda_i}(\tilde{t}_i) - 2m\tilde{t}_i.$$

That is

$$\tilde{q}_i = -P'_{\lambda_i}(\tilde{t}_i),$$

because $\tilde{t}_i = |\tilde{\beta}_i| = \tilde{\beta}_i$. Similarly, if $\tilde{\beta}_i < 0$, then

$$\tilde{q}_i = P'_{\lambda_i}(\tilde{t}_i);$$

if $\tilde{\beta}_i = 0$, then

$$|\tilde{q}_i| \leqslant P'_{\lambda_i}(\tilde{t}_i).$$

According to (3.17), for each $i = 1, \cdots, p$ we have

$$K_i = \begin{cases} \{(0,0)\} & \text{if } \big(\tilde{\beta}_i = 0 \text{ and } |\tilde{q}_i| < |P'_{\lambda_i}(\tilde{t}_i)|\big), \\ \{(\beta_i, t_i) \in \mathbb{R}^2_+ \mid \beta_i - t_i = 0\} & \text{if } \big(\tilde{\beta}_i = 0 \text{ and } \tilde{q}_i = -P'_{\lambda_i}(\tilde{t}_i)\big), \\ \{(\beta_i, t_i) \in \mathbb{R}^2 \mid \beta_i - t_i = 0\} & \text{if } \tilde{\beta}_i > 0, \\ \{(\beta_i, t_i) \in \mathbb{R}_- \times \mathbb{R}_+ \mid \beta_i + t_i = 0\} & \text{if } \big(\tilde{\beta}_i = 0 \text{ and } \tilde{q}_i = P'_{\lambda_i}(\tilde{t}_i)\big), \\ \{(\beta_i, t_i) \in \mathbb{R}^2 \mid \beta_i + t_i = 0\} & \text{if } \tilde{\beta}_i < 0. \end{cases} \tag{3.20}$$

and

$$K = \mathbb{R} \times K_1 \times \cdots \times K_p.$$

Next, we give an explicit expression for the affine hull of K. Define two matrices M and N as follows:

$$M = \begin{bmatrix} 1 & 0 \\ 0 & I_p \\ 0 & I_p \end{bmatrix} \text{ and } N = \begin{bmatrix} 1 & 0 \\ 0 & I_p \\ 0 & -I_p \end{bmatrix}.$$

Construct a matrix Ξ by first adding the common first column of M and N and then adding the $(i+1)^{th}$ column of M (N) if the condition in the second or third (fourth or fifth) row of (3.20) is satisfied. Columns of Ξ form a basis of the affine hull of K. Note that $\Xi^T L \Xi = Q$, where Q is defined in Assumption 3.3. From Proposition 2.5 and Theorem 4.3 of [39], L_K is a global homeomorphism. Under Assumption 3.1(b), it easy to see that the partial derivative of f_0 at $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is strong. An application of [40, Theorem 3] implies that z_0 is a locally unique solution to (3.15), therefore $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a locally unique optimal solution to (3.5).

□

In the above lemma, the non-singularity of Q in (3.18) guarantees $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ to be a locally unique optimal solution to (3.5), so $(\tilde{\beta}_0, \tilde{\beta})$ is also a locally unique solution to (3.1) according to Lemma 3.1. For the details, we refer the reader to its proof in the Appendix B.

As before, we use Σ_0 to denote the covariance matrix of $F(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}, X, Y)$ and let Σ_0^1 be the upper left $(p+1) \times (p+1)$ submatrix of Σ_0. Since the past p elements of $F(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}, X, Y)$ are constants at $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, we have $\Sigma_0 = \begin{bmatrix} \Sigma_0^1 & 0 \\ 0 & 0 \end{bmatrix}$.

3.2.2 Transformations of the SAA problem

We follow the same steps in Subsection 3.2.2 to formulate the SAA problem (3.2) as a normal map equation. First, by introducing the variable $t \in \mathbb{R}^p$ we transform (3.5) to the following equivalent problem:

$$\min_{(\beta_0,\beta,t)\in S} \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right]^2 + \sum_{i=1}^{p} P_{\lambda_i}(t_i) + m(||t||_2^2 - ||\beta||_2^2). \tag{3.21}$$

Second, we rewrite (3.21) as a variational inequality

$$0 \in f_N(\beta_0, \beta, t) + N_S(\beta_0, \beta, t), \tag{3.22}$$

where $f_N(\beta_0, \beta, t) = N^{-1}\sum_{i=1}^{N} F(\beta_0, \beta, t, \mathbf{x}^i, y_i)$. If $P_{\lambda_i}(t_i)$ is twice differentiable at t_i for every i from 1 to p, then the Jacobian matrix of f_N is given by

$$L_N(t) = df_N(\beta_0, \beta, t) = \begin{bmatrix} 2 & 2\sum_{i=1}^{N}\mathbf{x}^i/N & 0 \\ 2\sum_{i=1}^{N}(\mathbf{x}^i)^T/N & 2\sum_{i=1}^{T}(\mathbf{x}^i)^T(\mathbf{x}^i)/N - 2mI_p & 0 \\ 0 & 0 & \nabla P(t) + 2mI_p \end{bmatrix}. \tag{3.23}$$

Third, denoting the normal map induced by f_N and S by $(f_N)_S$, we obtain the normal map formulation of (3.22) as

$$(f_N)_S(z) = 0. \tag{3.24}$$

Let $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ be an (local) optimal solution to (3.21), then $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ is also a solution to (3.22). So the point $z_N \in \mathbb{R}^{2p+1}$ defined as

$$z_N = (\hat{\beta}_0, \hat{\beta}, \hat{t}) - f_N(\hat{\beta}_0, \hat{\beta}, \hat{t}) \tag{3.25}$$

is a solution to (3.24) and satisfies $\Pi_S(z_N) = (\hat{\beta}_0, \hat{\beta}, \hat{t})$. In fact, under Assumptions 3.1, 3.2 and 3.3, this z_N is a locally unique solution to (3.24) when N is large enough and it converges to z_0. This result will be shown in Subsection 3.3.1. Consequently, $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ is a locally unique optimal solution to (3.21) and converges to $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$. Let Σ_N be the sample covariance matrix of $\{F(\hat{\beta}_0, \hat{\beta}, \hat{t}, \mathbf{x}^i, y_i)\}_{i=1}^N$ and Σ_N^1 be the upperleft $(p+1) \times (p+1)$ submatrix of Σ_N, then we have $\Sigma_N = \begin{bmatrix} \Sigma_N^1 & 0 \\ 0 & 0 \end{bmatrix}$. Lemma 2.3 shows that Σ_N converges to Σ_0 almost surely as N goes to infinity for LASSO penalty. One can similarly prove the same convergence result with general penalty in this chapter under Assumptions 3.1-3.4.

Finally, we introduce the last set of assumptions below.

Assumption 3.4. *(a) For each $h \in \mathbb{R}^{2p+1}$ and $(\beta_0, \beta, t) \in \mathbb{R}^{2p+1}$, let*

$$M_{\beta_0,\beta,t}(h) = E\big[\exp\{\langle h, F(\beta_0, \beta, t, X, Y) - f_0(\beta_0, \beta, t)\rangle\}\big]$$

be the moment generating function of the random variable $F(\beta_0, \beta, t, X, Y) - f_0(\beta_0, \beta, t)$. Let $\mathcal{C}$ be a compact set in $\mathbb{R}^{2p+1}$ that contains $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ in its interior, and on which the second derivative of $P_{\lambda_i}(t_i)$ is Lipchitz continuous for each i from 1 to p. Assume the following conditions.

1. *There exists a constant $\zeta > 0$ such that $M_{\beta_0,\beta,t}(h) \leq \exp\{\zeta^2\|h\|^2/2\}$ for each $h \in \mathbb{R}^{2p+1}$ and $(\beta_0, \beta, t) \in \mathcal{C}$.*

2. *There exists a nonnegative random variable $\kappa(X, Y)$ such that*

$$\|F(\beta_0, \beta, t, X, Y) - F(\beta_0', \beta', t', X, Y)\| \leq \kappa(X, Y)\|(\beta_0, \beta, t) - (\beta_0', \beta', t')\|$$

   for all (β_0, β, t) and (β_0', β', t') in $\mathcal{C}$ and almost every (X, Y).

3. *The moment generating function of κ is finite valued in a neighborhood of zero.*

(b) The same conditions as in (a) for $d_1F(\beta_0, \beta, t, X, Y)$ instead of $F(\beta_0, \beta, t, X, Y)$. Accordingly, use $E[d_1F(\beta_0, \beta, t, X, Y)]$ to replace $f_0(\beta_0, \beta, t)$ in the conditions.

(c) The same conditions as in (a) for $F(\beta_0, \beta, t, X, Y)F(\beta_0, \beta, t, X, Y)^T$. Accordingly, use $E[F(\beta_0, \beta, t, X, Y)F(\beta_0, \beta, t, X, Y)^T]$ to replace $f_0(\beta_0, \beta, t)$ in the conditions.

Assumption 3.4(a) imposes conditions on the random variable $F(\beta_0, \beta, t, X, Y)$ as well as the penalty terms. It will hold if (X, Y) is a bounded random variable and Assumption 3.1(b) holds. Assumption 3.4(a) is used to ensure the SAA function f_N to converge to f_0 in probability at an exponential rate. We state the result in the following lemma.

Lemma 3.3. *Suppose that Assumptions 3.1, 3.2 and 3.4(a) hold. Then there exist positive real numbers δ_1, μ_1, M_1 and σ_1 such that the following holds for each $\epsilon > 0$ and each N:*

$$\text{Prob}\left\{\sup_{(\beta_0,\beta,t)\in\mathcal{C}} ||f_N(\beta_0, \beta, t) - f_0(\beta_0, \beta, t)|| \geqslant \epsilon\right\} \leqslant \delta_1 \exp\{-N\mu_1\} + \frac{M_1}{\epsilon^{2p+1}} \exp\left\{-\frac{N\epsilon^2}{\sigma_1}\right\}. \tag{3.26}$$

Proof of Lemma 3.3. The conclusion follows from an application of [32, Theorem 4]. We verify the assumptions of the latter theorem as follows. From equations (4.2) and (3.11) we can see that the Assumption 1 in [32] holds under Assumptions 3.1 and 3.2 of this paper. Moreover, Assumption 3.4(a) in [32] is satisfied for the compact set $\mathcal{C}$ under Assumption 3.4(a) of this paper.

□

The parts (b) and (c) of Assumption 3.4 impose the same type of assumptions on different random variables. Assumption 3.4(a-b) are needed in part of Theorem 3.1 and they enable us to construct reliable estimates for an unknown quantity in the asymptotic distribution of (3.16) (see Theorem 3.2). Assumption 3.4(c) is only required when the matrix Σ_0^1 is singular.

3.3 Confidence intervals for population penalized parameters

In this section, we develop the method to construct confidence intervals for a (locally) optimal solution of the population penalized regression problem (3.1) according to a (locally) optimal solution of the SAA problem (3.2).

3.3.1 Convergence and distribution of SAA solutions

Under Assumptions 3.1-3.3, from Lemma 3.2 we know that z_0 defined in (3.16) is a unique solution to (3.15) in some neighborhood. Furthermore, we can show that (3.24) has a unique solution z_N in a sub-neighborhood for sufficiently large N, and z_N converges almost surely to z_0. The results are summarized in Theorem 3.1 below.

Theorem 3.1. *Suppose that Assumptions 3.1, 3.2 and 3.3 hold. Then, for almost every $\omega \in \Omega$, there exists an integer N_ω and neighborhoods $\mathcal{Z}$ of z_0 and $\mathcal{C}_0$ of $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, such that for each $N \geq N_\omega$, the equation* (3.24) *has a unique solution z_N in $\mathcal{Z}$, and the variational inequality* (3.22) *has a unique solution in $\mathcal{C}_0$ given by $(\hat{\beta}_0, \hat{\beta}, \hat{t}) = \Pi_S(z_N)$. Moreover,*

$$\lim_{N\to\infty} z_N = z_0 \ a.e., \qquad \lim_{N\to\infty} (\hat{\beta}_0, \hat{\beta}, \hat{t}) = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \ a.e., \tag{3.27}$$

$$\sqrt{N}(z_N - z_0) \Rightarrow (L_K)^{-1}(\mathcal{N}(0, \Sigma_0)), \tag{3.28}$$

$$\sqrt{N}(\Pi_S(z_N) - \Pi_S(z_0)) \Rightarrow \Pi_K \circ (L_K)^{-1}(\mathcal{N}(0, \Sigma_0)), \tag{3.29}$$

and

$$\sqrt{N} L_K(z_N - z_0) \Rightarrow \mathcal{N}(0, \Sigma_0). \tag{3.30}$$

Suppose in addition that Assumption 3.4(a-b) holds. Then there exist positive real numbers ϵ_0, δ_0, μ_0, M_0 and σ_0, such that the following holds for each $\epsilon \in (0, \epsilon_0]$ and each N:

$$\begin{aligned} &\text{Prob}\left\{\|(\hat{\beta}_0, \hat{\beta}, \hat{t}) - (\tilde{\beta}_0, \tilde{\beta}, \tilde{t})\| < \epsilon\right\} \geq \text{Prob}\left\{\|z_N - z_0\| < \epsilon\right\} \\ \geq &1 - \delta_0 \exp\{-N\mu_0\} - \frac{M_0}{\epsilon^{2p+1}} \exp\left\{-\frac{N\epsilon^2}{\sigma_0}\right\}. \end{aligned} \tag{3.31}$$

Proof of Theorem 3.1. Follow the proof of Theorem 2.1.

□

From (3.30) we can readily derive an expression for the confidence region of z_0, which will depend on Σ_0 and L_K. However, both of these two are unknown in real applications, since we can not obtain the true solution of (3.1) in advance. In order to obtain computable confidence regions, we need to find reliable estimators of Σ_0 and L_K.

3.3.2 Estimators of Σ_0 and L_K

One can show that Σ_N converges to Σ_0 almost surely under Assumptions 3.1-3.3, therefore we can use Σ_N as a good estimator of Σ_0. Our main task in this subsection is to introduce an estimator of the normal map L_K. Since L_K is exactly the same as $d(f_0)_S(z_0)$, one may thus attempt to use $d(f_0)_S(z_N)$ as an estimate of L_K. However, this is problematic because the function $d(f_0)_S(\cdot)$ may not be continuous with respect to variable z in a neighborhood of z_0. This discontinuity can be seen from the chain rule of B-differentiability:

$$d(f_0)_S(z)(h) = L(t)\ d\Pi_S(z)(h) + h - d\Pi_S(z)(h) \text{ for each } z \in \mathbb{R}^{2p+1},\ h \in \mathbb{R}^{2p+1},$$

where $d\Pi_S(z)$ is the B-derivative of the Euclidean projector Π_S at z. Note that $d\Pi_S(z)$ is not continuous with respect to z at those points z on the boundary of any $(2p+1)$-cell in the normal manifold of S. This results in the discontinuity of $d(f_0)_S(\cdot)$ at these points. If $d(f_0)_S(\cdot)$ is not continuous at z_0, $d(f_0)_S(z_N)$ may not converge to $d(f_0)_S(z_0)$, in which case $d(f_0)_S(z_N)$ is not a good estimator of L_K.

Denote each cell in the normal manifold of S_i as C_i^j for indices from 1 to p. According to (3.6) we can derive the constraints defining each C_i^j which are listed in Table 2.1 in Section 2.3.2. Therefore each $(2p+1)$-cell in the normal manifold of S can be written as $\mathbb{R} \times \Pi_{i=1}^p C_i^{\gamma(i)}$, where $\gamma(i) = 0, \cdots, 8$ for each $i = 1, \cdots, p$. From (3.16) and Lemma 3.2 we note that $\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$ can be only in the relative interior of C_i^3, C_i^4, C_i^6, C_i^7 or C_i^8 for all i. Consequently, $d(f_0)_S(\cdot)$ is not continuous at z_0 only when $\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$ is in the relative interior of C_i^3 or C_i^4 for some index i. We consider two cases based on the location of z_0, which correspond to the two situations in which the random variable $(L_K)^{-1}(\mathcal{N}(0, \Sigma_0))$ is normally distributed, or is a combination of more than one normal random variables..

- Case I: In this case, $\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$ is in the relative interior of C_i^6, C_i^7 or C_i^8 for all $i \in \{1 \cdots p\}$, and the normal map L_K and the B-derivative $d\Pi_S(z_0)$ are linear functions.

We can use $d\Pi_S(z_N)$ and $d(f_N)_S(z_N)$ as the estimators of $d\Pi_S(z_0)$ and L_K respectively.

- Case II: In this case, $\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$ is in the relative interior of C_i^3 or C_i^4 for some index $i \in \{1 \cdots p\}$, and L_K and $d\Pi_S(z_0)$ are piecewise linear functions. In this case, we have to derive an estimator of L_K other than $d(f_N)_S(z_N)$.

To deal with Case II, first we give the expression of $d\Pi_S(z)$, and then construct an asymptotically exact approximation of $d\Pi_S(z_0)$. According to (3.8), we have

$$d\Pi_S(z)(h) = \big(\breve{\beta}_0, d\Pi_{S_1}(\beta_1, t_1)(\breve{\beta}_1, \breve{t}_1), \cdots, d\Pi_{S_p}(\beta_p, t_p)(\breve{\beta}_p, \breve{t}_p)\big), \tag{3.32}$$

for each $z = (\beta_0, \beta, t)$ and $h = (\breve{\beta}_0, \breve{\beta}, \breve{t})$. We denote $d\Pi_{S_i}(\beta_i, t_i)$ in the relative interior of each C_i^j by a function $\psi_j : \mathbb{R}^2 \to \mathbb{R}^2$. Since $d\Pi_{S_i}(\beta_i, t_i)$ is the same function for all (β_i, t_i) in the relative interior of each C_i^j $(j = 0, 1, \cdots, 8)$, $d\Pi_{S_i}(\beta_i, t_i)$ has 9 different expressions. Define four matrices

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Table 2.2 shows the expression of each ψ_j using these matrices. Consequently we can denote $d\Pi_S(z)$ for all z in the relative interior of $\mathbb{R} \times \Pi_{i=1}^p C_i^{\gamma(i)}$ as

$$\Psi_{\boldsymbol{\gamma}(z)}(h) = \big(\breve{\beta}_0, \psi_{\gamma(1)}(\breve{\beta}_1, \breve{t}_1), \cdots, \psi_{\gamma(p)}(\breve{\beta}_p, \breve{t}_p)\big) \ \text{ for each } \ h = (\breve{\beta}_0, \breve{\beta}, \breve{t}), \tag{3.33}$$

where $\boldsymbol{\gamma}(z) \triangleq \big(\gamma(1), \cdots, \gamma(p)\big)$ such that $z \in \text{ ri}\left(\mathbb{R} \times \Pi_{i=1}^p C_i^{\gamma(i)}\right)$.

Next, we construct an estimator of $d\Pi_S(z_0)$. We divide the plane (β_i, t_i) into 9 pieces $E_i^0, \cdots, E_i^8$. The constraints that define each of these sets $E_i^0, \cdots, E_i^8$ are listed in Table 2.4. The function $g(N)$ has many chooses. It can be any combination of finite many terms of the form aN^b with $a > 0$ and $b \in (0, 1/2)$. Each partition $\mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)}$ is related to the $(2p+1)$-cell $\mathbb{R} \times \Pi_{i=1}^p C_i^{\gamma(i)}$. Let

$$\boldsymbol{\gamma}(z) \triangleq \big(\gamma(1), \cdots, \gamma(p)\big) \text{ such that } z \in \mathbb{R} \times \Pi_{i=1}^p E_i^{\gamma(i)}.$$

Given a sample size N and a fixed z, we define a function $\Lambda_N(z) : \mathbb{R}^{2p+1} \to \mathbb{R}^{2p+1}$ as

$$\Lambda_N(z)(h) = \Psi_{\gamma(z)}(h), \text{ for each } h \in \mathbb{R}^{2p+1}. \tag{3.34}$$

One can show that $\Lambda_N(z_N)$ converges to $d\Pi_S(z_0)$ in probability under Assumptions 3.1-3.4.

Based on (3.23), (3.25) and (3.34), we define a function $\Phi_N(z_N) : \mathbb{R}^{2p+1} \to \mathbb{R}^{2p+1}$ as

$$\Phi_N(z_N)(h) = L_N(\hat{t})\ \Lambda_N(z_N)(h) + h - \Lambda_N(z_N)(h) \tag{3.35}$$

for each $h \in \mathbb{R}^{2p+1}$. This $\Phi_N(z_N)$ converges to L_K in probability under Assumptions 3.1-3.4 and hence asymptotically exact estimator of L_K.

Under Assumptions 3.1-3.4, a key result to compute confidence regions is that the weak convergence in (3.30) still holds after substituting $\Phi_N(z_N)$ for L_K. Consequently, if Σ_0^1 is nonsingular, then we have

$$\sqrt{N} \begin{bmatrix} (\Sigma_N^1)^{-1/2} & 0 \\ 0 & I_p \end{bmatrix} (\Phi_N(z_N))(z_N - z_0) \Rightarrow \mathcal{N}(0, I_{p+1}) \times 0. \tag{3.36}$$

If Σ_0^1 is singular, then we can expect Σ_N^1 to be also singular when N is sufficiently large. Let l be the number of positive eigenvalues of Σ_0^1 counted with regard to their algebraic multiplicities, and decompose Σ_N^1 as

$$\Sigma_N^1 = U_N^T \Delta_N U_N$$

where U_N is an orthogonal $(p+1)\times(p+1)$ matrix, and Δ_N is a diagonal matrix with monotonically decreasing elements. Let D_N be the upper-left submatrix of Δ_N whose diagonal elements are at least $1/g(N)$, and let l_N be the number of rows in D_N. Moreover, let $(U_N)_1$ be the submatrix of U_N that consists of its first l_N rows, and let submatrix $(U_N)_2$ consist of the remaining rows of U_N. Then we can present the weak convergence results in the following theorem.

Theorem 3.2. *Suppose that Assumptions 3.1, 3.2, 3.3 and 3.4(a-b) hold. Then*

$$\sqrt{N}\Phi_N(z_N)(z_N - z_0) \Rightarrow \mathcal{N}(0, \Sigma_0). \tag{3.37}$$

If Σ_0^1 is nonsingular, then

$$N\big[(\Phi_N(z_N))(z_N - z_0)\big]^T \begin{bmatrix} (\Sigma_N^1)^{-1} & 0 \\ 0 & I_p \end{bmatrix} \big[(\Phi_N(z_N))(z_N - z_0)\big] \Rightarrow \chi^2_{p+1}, \tag{3.38}$$

and

$$N\big[(\Phi_N(z_N))(z_N - z_0)\big]^T \begin{bmatrix} 0 & I_p \end{bmatrix} \big[(\Phi_N(z_N))(z_N - z_0)\big] \Rightarrow 0. \tag{3.39}$$

If Σ_0^1 is singular and Assumption 3.4(c) holds, then $\mathrm{Prob}\{l_N = l\} \to 1$ *as* $N \to \infty$,

$$N\big[(\Phi_N(z_N))(z_N - z_0)\big]^T \begin{bmatrix} (U_N)_1^T D_N^{-1} (U_N)_1 & 0 \\ 0 & 0 \end{bmatrix} \big[(\Phi_N(z_N))(z_N - z_0)\big] \Rightarrow \chi^2_l, \tag{3.40}$$

and

$$N\big[(\Phi_N(z_N))(z_N - z_0)\big]^T \begin{bmatrix} (U_N)_2^T (U_N)_2 & 0 \\ 0 & I_p \end{bmatrix} \big[(\Phi_N(z_N))(z_N - z_0)\big] \Rightarrow 0. \tag{3.41}$$

Proof of Theorem 3.2. The conclusions follows from Theorem 2.3.

□

We can treat (3.38) and (3.39) as a special case of (3.40) and (3.41). For Case I, the following theorem shows that $d(f_N)_S(z_N)$ is a strongly consistent estimator of L_K.

Theorem 3.3. *Suppose that Assumptions 3.1, 3.2 and 3.3 hold. Moreover, the solution of (3.15) z_0 satisfies the conditions for I. Then $d\Pi_S(z_N)$ defined in (3.33) converges to $d\Pi_S(z_0)$ almost surely, and*

$$d(f_N)_S(z_N) = L_N(\hat{t})\ d\Pi_S(z_N) + I - d\Pi_S(z_N)$$

converges to L_K almost surely. Therefore, for sufficiently large N, $[d(f_N)_S(z_N)]^{-1}$ converges to $(L_K)^{-1}$ almost surely.

Proof of Theorem 3.3. The conclusions follow from Theorem 2.1.

In Case I, we also have Theorem 3.2 hold by substituting $d(f_N)_S(z_N)$ for $\Phi_N(z_N)$.

3.3.3 Confidence intervals for penalized parameters

At the beginning of this subsection, we investigate the relationship between a solution of normal map formulation (3.15) and the corresponding (locally) optimal solution of problem (3.1). Let $\tilde{q}$ be as defined in Assumption 3.3 and $\tilde{q}_0 = E[-2(Y - \tilde{\beta}_0 - \sum_{j=1}^{p} \tilde{\beta}_j X_j)]$. $f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = (\tilde{q}_0, \tilde{q}, \lambda e_p)$. It follows from (3.16) that $z_0 = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) - (\tilde{q}_0, \tilde{q}, \lambda e_p)$. Since $-f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \in N_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, we know that $\tilde{q}_0 = 0$ which gives $\tilde{\beta}_0 = (z_0)_1$. Thus, confidence intervals of $\tilde{\beta}_0$ are exactly those of $(z_0)_1$.

On the other hand, according to the fact $(\tilde{\beta}_i, \tilde{t}_i) = \Pi_{S_i}\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$ for each $i = 1, \cdots, p$, we have the following relationship between $\tilde{\beta}_i$ and $\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$:

$$\tilde{\beta}_i = \begin{cases} \frac{(z_0)_{i+1} + (z_0)_{i+1+p}}{2}, & \text{if } (z_0)_{i+1} + (z_0)_{i+1+p} > 0 \text{ and } (z_0)_{i+1} - (z_0)_{i+1+p} \geqslant 0, \\ 0, & \text{if } (z_0)_{i+1} + (z_0)_{i+1+p} \leqslant 0 \text{ and } (z_0)_{i+1} - (z_0)_{i+1+p} \geqslant 0, \\ \frac{(z_0)_{i+1} - (z_0)_{i+1+p}}{2}, & \text{if } (z_0)_{i+1} + (z_0)_{i+1+p} \leqslant 0 \text{ and } (z_0)_{i+1} - (z_0)_{i+1+p} < 0. \end{cases} \tag{3.42}$$

Let us denote the right hand side of (3.42) as $\Gamma\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$. Note that the above three cases include all the possible situations for the location of $\big((z_0)_{i+1}, (z_0)_{i+1+p}\big)$. This map Γ can be used to obtain confidence intervals for $\tilde{\beta}_i$ $(i = 1, \cdots, p)$ as long as we have confidence intervals for $\big((z_0)_{i+1} + (z_0)_{i+1+p}\big)$ and $\big((z_0)_{i+1} - (z_0)_{i+1+p}\big)$ in hand. For a fixed i, we denote the confidence intervals for $\big((z_0)_{i+1} + (z_0)_{i+1+p}\big)$ and $\big((z_0)_{i+1} - (z_0)_{i+1+p}\big)$ as $[L^i_{\text{plus}}, U^i_{\text{plus}}]$ and $[L^i_{\text{minus}}, U^i_{\text{minus}}]$ respectively. Then the confidence intervals for $\tilde{\beta}_i$ is

$$\big[\Gamma\big(L^i_{\text{plus}}, L^i_{\text{minus}}\big), \Gamma\big(U^i_{\text{plus}}, U^i_{\text{minus}}\big)\big] . \tag{3.43}$$

Here we treat the inputs of Γ as $\big((z_0)_{i+1} + (z_0)_{i+1+p}\big)$ and $\big((z_0)_{i+1} - (z_0)_{i+1+p}\big)$.

Now we focus on how to find confidence intervals for $(z_0)_1$, $\big((z_0)_{i+1} + (z_0)_{i+1+p}\big)$ and $\big((z_0)_{i+1} - (z_0)_{i+1+p}\big)$. Under Assumptions 3.1-3.4, from Theorem 3.2 we can express the asymptotically exact $(1-\alpha)100\%$ confidence region for z_0 as

$$\left\{ z \in \mathbb{R}^{2p+1} \;\middle|\; \begin{array}{rcl} N[\Phi_N(z_N)(z_N - z)]^T \begin{bmatrix} (U_N)_1^T D_N^{-1} (U_N)_1 & 0 \\ 0 & 0 \end{bmatrix} [\Phi_N(z_N)(z_N - z)] & \leqslant & \chi^2_{l_N}(\alpha) \\ \begin{bmatrix} (U_N)_2^T (U_N)_2 & 0 \\ 0 & I_p \end{bmatrix} [\Phi_N(z_N)(z_N - z)] & = & 0 \end{array} \right\} \tag{3.44}$$

for sufficiently large N, where $\chi^2_{l_N}(\alpha)$ is the critical value associated with significant level α of a χ^2 distribution with l_N degrees of freedom. If $\Phi_N(z_N)$ is a linear map, then the set in (3.44) is an ellipsoid in a subspace of $\mathbb{R}^{2p+1}$. Otherwise it is the union of different ellipsoid fractions. To obtain simultaneous confidence intervals, we find the maximal and minimal values of $(z_0)_1$, $\big((z_0)_{i+1} + (z_0)_{i+1+p}\big)$ and $\big((z_0)_{i+1} - (z_0)_{i+1+p}\big)$ in the set of (3.44) by solving optimization problems.

On the other hand, it can be shown that $\Phi_N(z_N)$ is a global homeomorphism with probability 1 as $N \to \infty$. If $\Phi_N(z_N)$ is a global homeomorphism, we can use

$$(\Phi_N(z_N))^{-1}(\mathcal{N}(0, \Sigma_N)) \tag{3.45}$$

to approximate the distribution of $\sqrt{N}(z_N - z_0)$. When $\Phi_N(z_N)$ is a linear map, the distribution in (3.45) is normal. Therefore $(z_0)_1$, $\big((z_0)_{i+1} + (z_0)_{i+1+p}\big)$ and $\big((z_0)_{i+1} - (z_0)_{i+1+p}\big)$ also follow normal distributions, from which we can construct individual confidence intervals. When $\Phi_N(z_N)$ is not a linear map, we simulate data based on the distribution in (3.45), and find empirical individual confidence intervals for $(z_0)_1$, $\big((z_0)_{i+1} + (z_0)_{i+1+p}\big)$ and $\big((z_0)_{i+1} - (z_0)_{i+1+p}\big)$.

3.4 Confidence intervals for the true parameters in the underlying linear model

In this section, we derive asymptotic results for the true parameters in the underlying linear model based on the convergence theorems in Section 3.3, and aim to obtain the corresponding individual confidence intervals.

Suppose our underlying linear model is

$$Y = \beta_0^{true} + X^T\beta^{true} + \varepsilon, \tag{3.46}$$

where $\beta_0^{true} \in \mathbb{R}$ and $\beta^{true} = (\beta_1^{true}, \cdots, \beta_p^{true}) \in \mathbb{R}^p$ are the true parameters. The random error ε has mean zero and variance σ_ε^2. Moreover, ε is independent with X_i for each $i = 1, \cdots, p$. In this section, we assume that $E(X_i) = 0$ for each $i = 1, \cdots, p$, hence $E(Y) = \beta_0^{true}$. Denote the covariance matrix of X as Σ, i.e., $\Sigma = E(XX^T)$.

Plugging (3.46) into (3.16), we have

$$z_0 = \begin{bmatrix} \tilde{\beta}_0 + 2E(Y - \tilde{\beta}_0 - X^T\tilde{\beta}) \\ \tilde{\beta} + 2E(Y - \tilde{\beta}_0 - X^T\tilde{\beta})X + 2m\tilde{\beta} \\ \tilde{t} - P(\tilde{t}) - 2m\tilde{t} \end{bmatrix} = \begin{bmatrix} 2\beta_0^{true} - \tilde{\beta}_0 \\ (1+2m)\tilde{\beta} + 2\Sigma(\beta^{true} - \tilde{\beta}) \\ (1-2m)\tilde{t} - P(\tilde{t}) \end{bmatrix} \tag{3.47}$$

If Σ is invertible, then from (3.47) we obtain

$$\beta_0^{true} = (z_0)_1, \quad \beta^{true} = \frac{1}{2}\Sigma^{-1}(z_0)_{2:(p+1)} + \left[I_p - \frac{1}{2}(1+2m)\Sigma^{-1}\right]\tilde{\beta}, \tag{3.48}$$

where $(z_0)_{2:(p+1)}$ denotes a vector that consists of the second to $(p+1)^{th}$ entries of z_0. Expression (3.48) suggests the following corresponding estimators

$$\hat{\beta}_0^{true} = (z_N)_1, \quad \hat{\beta}^{true} = \frac{1}{2}\hat{\Theta}(z_N)_{2:(p+1)} + \left[I_p - \frac{1}{2}(1+2m)\hat{\Theta}\right]\hat{\beta}, \tag{3.49}$$

where $\hat{\Theta}$ is an estimator of the precision matrix Σ^{-1}. From (3.25) and (3.47) one may notice that (3.49) is essentially the same estimator as in [56] and [50], when dealing with Lasso penalty with $m = 0$. Let G be a map from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{p+1}$ defined as

$$G = \frac{1}{2}\left(\begin{bmatrix} 1 & 0 \\ 0 & \Sigma^{-1} \end{bmatrix} B + \begin{bmatrix} 1 & 0 \\ 0 & 2I - (1+2m)\Sigma^{-1} \end{bmatrix} B \circ \Pi_K\right), \tag{3.50}$$

and $\hat{G}$ be the following map

$$\hat{G} = \frac{1}{2}\left(\begin{bmatrix} 1 & 0 \\ 0 & \hat{\Theta} \end{bmatrix} B + \begin{bmatrix} 1 & 0 \\ 0 & 2I - (1+2m)\hat{\Theta} \end{bmatrix} B \circ d\Pi_S(z_N)\right), \tag{3.51}$$

where B is a $(p+1)$ by $(2p+1)$ matrix defined as $B = \begin{bmatrix} I_{p+1} & 0 \end{bmatrix}$. Since Π_S is homogeneous, we know that $\Pi_S(z_0) = d\Pi_S(z_0)(z_0)$ and $\Pi_S(z_N) = d\Pi_S(z_N)(z_N)$. Then according to (3.48), (3.49), $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = \Pi_S(z_0)$ and $(\hat{\beta}_0, \hat{\beta}, \hat{t}) = \Pi_S(z_N)$, we can rewrite (3.48) and (3.49) as

$$(\beta_0^{true}, \beta^{true}) = G(z_0) \quad \text{and} \quad (\hat{\beta}_0^{true}, \hat{\beta}^{true}) = \hat{G}(z_N).$$

The following theorem shows that (3.49) gives a consistent estimator of the true parameter $(\beta_0^{true}, \beta^{true})$, and states an asymptotic distribution from which we can derive the confidence region for $(\beta_0^{true}, \beta^{true})$.

Theorem 3.4. *Suppose that Assumptions 3.1, 3.2 and 3.3 hold, and the true covariance matrix Σ is nonsingular. Let $\hat{\Theta}$ be a $\sqrt{N}$-consistent estimator of Σ^{-1} and m be a positive constant used in (3.5). Then $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$ is a consistent estimator of $(\beta_0^{true}, \beta^{true})$ and*

$$\sqrt{N}\left((\hat{\beta}_0^{true}, \hat{\beta}^{true}) - (\beta_0^{true}, \beta^{true})\right) \Rightarrow G \circ (L_K)^{-1}(\mathcal{N}(0, \Sigma_0)), \tag{3.52}$$

where G is the map defined in (3.50).

Proof of Theorem 3.4. Follow the proof of Theorem 2.6.

□

There are many choices for $\hat{\Theta}$ in real applications. What people usually used are the inverse of sample covariance matrix and the estimate of precision matrix computed by banding method [4] or penalized likelihood method [16]. From literature, it is well known that these estimators of precision matrix have $\sqrt{N}$-consistency when p is fixed [22].

To use (3.52) to compute confidence intervals, we replace G and L_K there by their estimators. For Case I, the following theorem gives an approach to compute the asymptotically exact individual confidence intervals for $(\beta_0^{true}, \beta^{true})$.

Theorem 3.5. *Suppose that Assumptions 3.1, 3.2 and 3.3 hold, the true covariance matrix* Σ *is nonsingular, and the solution to the normal map formulation* (3.15) *satisfies the conditions for Case I. Let* $\hat{\Theta}$ *be a* $\sqrt{N}$*-consistent estimator of* Σ^{-1}*, and define* $H = G(L_K)^{-1}$ *and* $H_N = \hat{G}\left[d(f_N)_S(z_N)\right]^{-1}$*. If* $(H\Sigma_0 H^T)_{i+1,i+1} \neq 0$*, then*

$$\frac{\sqrt{N}(\hat{\beta}_i^{true} - \beta_i^{true})}{\sqrt{(H_N \Sigma_N H_N^T)_{i+1,i+1}}} \Rightarrow \mathcal{N}(0,1)\,, \tag{3.53}$$

for all $i = 0, 1, \cdots, p$.

Proof of Theorem 3.5. Follow from the proof of Theorem 2.7

□

For Case II, to show how to compute the asymptotically exact individual confidence intervals for $(\beta_0^{true}, \beta^{true})$, we consider the image of normal random vectors under certain functions. Let $f : \mathbb{R}^{2p+1} \to \mathbb{R}$ be a continuous function and Z be a $\mathbb{R}^{2p+1}$ dimensional random variable with $Z \sim \mathcal{N}(0, I_{p+1}) \times \vec{\mathbf{0}}$. Define $a^r(f) \in (0, \infty)$ as

$$a^r(f) = \inf\left\{c \geqslant 0 \mid \text{Prob}\left\{-c \leqslant f(Z) - r \leqslant c\right\} \geqslant 1 - \alpha\right\}. \tag{3.54}$$

Suppose that $\text{Prob}\left\{f(Z) = b\right\} = 0$ for all $b \in \mathbb{R}$. Then for any given $r \in \mathbb{R}$ and $\alpha \in (0,1)$, $a^r(f)$ as defined in (3.54) is the smallest value that satisfies

$$\text{Prob}\left\{-a^r(f) \leqslant f(Z) - r \leqslant a^r(f)\right\} = 1 - \alpha.$$

Define two functions R and $\hat{R}$ from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{p+1}$ as

$$R = G \circ (L_K)^{-1} \begin{bmatrix} (\Sigma_0^1)^{\frac{1}{2}} & 0 \\ 0 & I_p \end{bmatrix} \quad \text{and} \quad \hat{R} = \hat{G}' \circ (\Phi_N(z_N))^{-1} \begin{bmatrix} (\Sigma_N^1)^{\frac{1}{2}} & 0 \\ 0 & I_p \end{bmatrix}, \tag{3.55}$$

where

$$\hat{G}' = \frac{1}{2}\left(\begin{bmatrix} 1 & 0 \\ 0 & \hat{\Theta} \end{bmatrix} B + \begin{bmatrix} 1 & 0 \\ 0 & 2I - (1+2m)\hat{\Theta} \end{bmatrix} B \circ \Lambda_N(z_N)\right). \tag{3.56}$$

We denote the jth component function of R and $\hat{R}$ as R_j and $\hat{R}_j$ respectively for each $j = 1, 2, \cdots, p+1$.

Note that the map G is a piecewise linear function in Case II. From the expression (3.50) and the matrix representations of the piecewise linear function Π_K based on the location of z_0, one can check that G has the following form with $m = \frac{1}{2}$

$$\left[\begin{array}{cc|c} 1 & 0 & \\ 0 & \frac{1}{2}\Sigma^{-1}(I-2W)+W & * \end{array}\right], \tag{3.57}$$

in which W is a piecewise linear function represented by $p \times p$ diagonal matrices with diagonal elements 0 or $\frac{1}{2}$. If Σ is nonsingular, then the submatrix $\frac{1}{2}\Sigma^{-1}(I-2W)+W$ has full row rank. This can be seen by writing down an equivalent expression $\frac{1}{2}\Sigma^{-1}[I-(2-\delta)W]+(I-\frac{1}{2}\delta\Sigma^{-1})W$ with sufficient small positive constant δ. Furthermore, if Σ and Σ_0^1 are both nonsingular, then the matrix representation of each piece of the map G has full row rank. Because L_K is a global homeomorphism under Assumptions 3.2(a) and 3.3, it follows that $\text{Prob}\,\{R_j(Z)=b\}=0$ for all $b \in \mathbb{R}$. The following theorem gives a way of computing individual confidence intervals for $(\beta_0^{true}, \beta^{true})$.

Theorem 3.6. *Suppose that Assumptions 3.1, 3.2, 3.3 and 3.4(a-b) hold, $m = \frac{1}{2}$ and the population covariance matrices Σ and Σ_0^1 are nonsingular. Let $\hat{\Theta}$ be a $\sqrt{N}$-consistent estimator of Σ^{-1}, and $\alpha \in (0,1)$, $a^r(\cdot)$ be as in (3.54). Then for every $r \in \mathbb{R}$ and all $j = 0, 1, \cdots, p$, we have*

$$\lim_{N\to\infty} \text{Prob}\left\{|\sqrt{N}(\hat{\beta}_j^{true} - \beta_j^{true}) - r| \leqslant a^r(\hat{R}_{j+1})\right\} = 1-\alpha, \tag{3.58}$$

where R and $\hat{R}$ are defined in (3.55).

We introduce two lemmas that will be used in the proof of Theorem 3.6.

Lemma 3.4. *Let $C(\mathbb{R}^{2p+1},\mathbb{R})$ denote the space of continuous functions from $\mathbb{R}^{2p+1}$ to $\mathbb{R}$, $\{u_N\}_{N=1}^{\infty}$ be a sequence of $C(\mathbb{R}^{2p+1},\mathbb{R})$ valued random variables which converges to u in probability uniformly on compact sets, and $\{Z_N\}_{N=1}^{\infty}$ be a sequence of real valued random variables* *that converges to $u(Z)$ in distribution. If $m = \frac{1}{2}$, then for every $r \in \mathbb{R}$,*

$$\lim_{N\to\infty} \text{Prob}\left\{-a^r(u_N) \leqslant Z_N - r \leqslant a^r(u_N)\right\} = 1 - \alpha.$$

Proof of Lemma 3.4. Follow the proof of Lemma 2.8

□

Lemma 3.5. *Suppose that Assumptions 3.1, 3.2, 3.3 and 3.4(a-b) hold, and the population covariance matrices Σ and Σ_0^1 are nonsingular. Let $\hat{\Theta}$ be a consistent estimator of Σ^{-1}. Then $\hat{R}$ converges to R in probability.*

Proof of Lemma 3.5. Follow the proof of Lemma 2.9

□

Proof of Theorem 3.6. By Lemma 3.5, $\hat{R}_j$ converges to R_j in $C(\mathbb{R}^{2p+1}, \mathbb{R})$ in probability uniformly on compact sets. Let

$$Z_N = \sqrt{N}\left((\hat{\beta}_0^{true}, \hat{\beta}^{true}) - (\beta_0^{true}, \beta^{true})\right)_j$$

for $j = 1, \cdots, p+1$. From (3.52), Z_N converges to $R_j(Z)$ in distribution. Then the conclusions follow from Lemma 3.4 with $u_N = \hat{R}_j$ and $u = R_j$.

□

In practice, for a fixed choice of r we can find the empirical individual confidence intervals for $(\beta_0^{true}, \beta^{true})$ by simulating data from $\hat{R}(Z)$.

3.5 Numerical examples

In this section, we use MCP penalized regression defined in (3.4) to illustrate the performance of our method proposed in Section 3. We implement it using Matlab and GAMS, and choose $\frac{1}{g(N)} = \frac{0.001}{N^{1/3}}$, $m = \frac{1}{2}$ in (3.5), for all examples in this section. We use the MIQCP (Mixed Integer Quadratically Constrained Program) solver in GAMS to solve optimization problems such as (3.2).

In the first two examples, we generate the data using the following model:

$$Y = \bar{\beta}^T X + \sigma\epsilon \tag{3.59}$$

where $\bar{\beta} \in \mathbb{R}^p$, X is a p-dimensional normal random variable with mean 0 and covariance $\bar{\Sigma}_{ij} = \rho^{|i-j|}$ for $\rho = 0.5$, ϵ is a standard normal random error which is independent of X. We set the noise level $\sigma = 1$. The random design regularized regression problem under model (3.59) is

$$\min_{\beta_0, \beta} (\bar{\beta} - \beta)^T \bar{\Sigma} (\bar{\beta} - \beta) + \beta_0^2 + \lambda \sum_{i=1}^{p} |\beta_i|. \tag{3.60}$$

In simulation we compute the empirical coverage probability, i.e. the fraction of total replications in which the confidence intervals contain the corresponding population penalized parameters or true parameters in the linear model.

3.5.1 Example 3.5.1: Low dimensional simulation

We choose $p = 8$, $\bar{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and generate a $(\mathbf{y}, \mathbf{X})$ dataset with 100 replications of sample size $N = 300$. We consider six MCP penalties, which have parameters (λ, a) as the following values: $\lambda = 0.5$, 1 and 2, $a = 2$ and 2000. In each replication, by solving SAA problem for every MCP penalty, we compute two types of individual confidence intervals. The first type confidence intervals are for the solution to the problem (3.60), while the second type confidence intervals are for the true parameters $\bar{\beta}$, both with confidence level 0.95 ($\alpha = 0.05$). Figure 3.1 shows the 95% individual confidence intervals computed from the first replication for each MCP penalty, which are also listed in Table 3.1. In Figure 3.1, red and green intervals represent the first and second type of confidence intervals respectively, which are given in the "Ind CI1" and "Ind CI2" columns of Table 3.1. The solution to the problem (3.60) is showed as blue dots on the left of red intervals and $\bar{\beta}$ is showed as blue dots on the left of green intervals in Figure 3.1. The estimators $(\hat{\beta}_0, \hat{\beta})$ and $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$ are listed in the "Est1" and "Est2" columns respectively in Table 3.1. From Figure 3.1 and Table 3.1, we observe comparatively short first type confidence intervals as well as singleton $\{0\}$ when the estimate $\hat{\beta}_i = 0$. In addition, the first type confidence intervals are not always symmetric around the estimates,

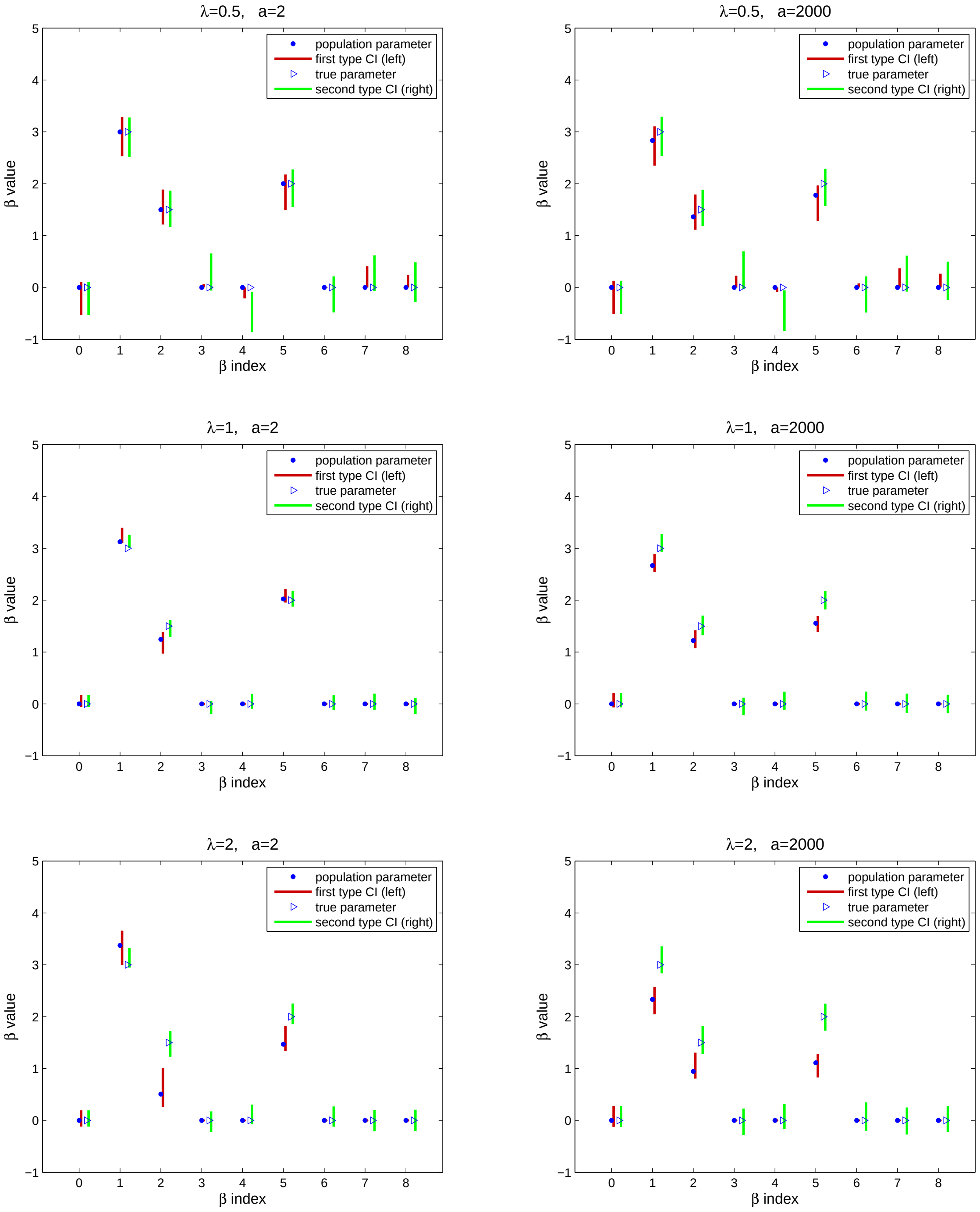


Figure 3.1: 95% individual CIs of $(\tilde{\beta}_0, \tilde{\beta})$ and $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 3.5.1.

except confidence intervals for $\tilde{\beta}_0$. These two phenomena are due to the contraction effect of projection Γ (3.42) on confidence intervals of z_0.

$\lambda = 0.5$	$a = 2$				$a = 2000$			
	Est1	Ind CI1	Est2	Ind CI2	Est1	Ind CI1	Est2	Ind CI2
β_0	-0.21	[-0.53, 0.11]	-0.21	[-0.53, 0.11]	-0.19	[-0.51, 0.13]	-0.19	[-0.51, 0.13]
β_1	2.91	[2.53, 3.29]	2.90	[2.52, 3.28]	2.73	[2.35, 3.11]	2.91	[2.53, 3.29]
β_2	1.55	[1.21, 1.89]	1.52	[1.17, 1.87]	1.45	[1.11, 1.79]	1.53	[1.18, 1.89]
β_3	-0.00	[0, 0.06]	0.30	[-0.06, 0.66]	0	[0, 0.23]	0.34	[-0.02, 0.70]
β_4	-0.00	[-0.21, 0]	-0.47	[-0.86, -0.08]	0	[-0.09, 0]	-0.45	[-0.84, -0.05]
β_5	1.83	[1.49, 2.18]	1.91	[1.55, 2.28]	1.63	[1.29, 1.97]	1.93	[1.57, 2.29]
β_6	0.00	[-0.01, 0]	-0.13	[-0.48, 0.21]	0	[0, 0.08]	-0.13	[-0.48, 0.21]
β_7	0.04	[0, 0.41]	0.27	[-0.07, 0.62]	0.08	[0, 0.37]	0.27	[-0.08, 0.61]
β_8	0.00	[0, 0.25]	0.10	[-0.28, 0.49]	0.00	[0, 0.27]	0.13	[-0.24, 0.50]

$\lambda = 1$	$a = 2$				$a = 2000$			
	Est1	Ind CI1	Est2	Ind CI2	Est1	Ind CI1	Est2	Ind CI2
β_0	0.06	[-0.06, 0.17]	0.06	[-0.06, 0.17]	0.07	[-0.07, 0.22]	0.07	[-0.07, 0.22]
β_1	3.25	[3.10, 3.40]	3.13	[3.00, 3.26]	2.71	[2.54, 2.89]	3.11	[2.93, 3.28]
β_2	1.18	[0.97, 1.39]	1.46	[1.29, 1.62]	1.25	[1.08, 1.42]	1.51	[1.32, 1.70]
β_3	0	[0, 0]	-0.07	[-0.20, 0.06]	0	[0, 0]	-0.05	[-0.22, 0.12]
β_4	0	[0, 0.]	0.05	[-0.09, 0.20]	0	[0, 0]	0.06	[-0.11, 0.24]
β_5	2.09	[1.95, 2.22]	2.03	[1.88, 2.18]	1.54	[1.39, 1.70]	2.00	[1.82, 2.18]
β_6	0	[0, 0]	0.03	[-0.11, 0.17]	0	[0, 0]	0.05	[-0.13, 0.24]
β_7	0	[0, 0]	0.04	[-0.12, 0.20]	0	[0, 0]	0.01	[-0.17, 0.20]
β_8	0	[0, 0]	-0.04	[-0.19, 0.12]	0	[0, 0]	0	[-0.18, 0.18]

$\lambda = 2$	$a = 2$				$a = 2000$			
	Est1	Ind CI1	Est2	Ind CI2	Est1	Ind CI1	Est2	Ind CI2
β_0	0.04	[-0.12, 0.19]	0.04	[-0.12, 0.19]	0.08	[-0.12, 0.28]	0.08	[-0.12, 0.28]
β_1	3.33	[2.99, 3.66]	3.14	[2.95, 3.33]	2.31	[2.05, 2.57]	3.10	[2.84, 3.36]
β_2	0.63	[0.25, 1.01]	1.48	[1.23, 1.73]	1.06	[0.81, 1.31]	1.55	[1.28, 1.82]
β_3	0	[0, 0]	-0.02	[-0.22, 0.17]	0	[0, 0]	-0.02	[-0.28, 0.23]
β_4	0	[0, 0]	0.12	[-0.07, 0.31]	0	[0, 0]	0.08	[-0.17, 0.32]
β_5	1.58	[1.34, 1.82]	2.05	[1.86, 2.25]	1.05	[0.83, 1.28]	1.99	[1.73, 2.25]
β_6	0	[0, 0]	0.07	[-0.12, 0.27]	0	[0, 0]	0.07	[-0.20, 0.35]
β_7	0	[0, 0]	0	[-0.21, 0.20]	0	[0, 0]	-0.01	[-0.27, 0.25]
β_8	0	[0, 0]	0	[-0.20, 0.21]	0	[0, 0]	0.03	[-0.22, 0.27]

Table 3.1: Estimates and 95% CIs for $(\tilde{\beta}_0, \tilde{\beta})$ and $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 3.5.1.

Table 3.2 and Table 3.3 show the empirical coverage probabilities (CP) and average interval lengths (ALen) for 95% individual confidence intervals among the 100 replications. In Table 3.2, the "$\tilde{\beta}$" column contains the solution to the problem (3.60) for different MCP penalties, which we expect to be covered by the first type confidence intervals. In Table 3.3, the "True" column contains the true model parameters $\bar{\beta}$, which we expect to be covered by the second type confidence intervals. Note that the coverage is 100% for the first type confidence interval when $\tilde{\beta}_i = 0$, $i = 1, \cdots, 8$. This is because the contraction effect of projection Γ (3.42) makes the confidence intervals more conservative from z_0 to $\tilde{\beta}$.

3.5.2 Example 3.5.2: High dimensional simulation

In this example, we consider a case that the dimension is much larger than the sample size. We choose $p = 300$ and set $\bar{\beta}$ as the following 300-dimensional vector: $\bar{\beta}_1 = 3$, $\bar{\beta}_2 = \bar{\beta}_{100} =$

$a=2$	$\lambda=0.5$			$\lambda=1$			$\lambda=2$		
	$\tilde{\beta}$	CP	ALen	$\tilde{\beta}$	CP	ALen	$\tilde{\beta}$	CP	ALen
β_0	0	96	0.68	0	99	0.23	0	98	0.34
β_1	3	96	0.79	3.13	98	0.31	3.37	90	0.75
β_2	1.5	94	0.82	1.25	93	0.42	0.51	89	0.70
β_3	0	100	0.17	0	100	0	0	100	0
β_4	0	100	0.20	0	100	0	0	100	0
β_5	2	94	0.73	2.02	93	0.26	1.47	97	0.50
β_6	0	100	0.22	0	100	0	0	100	0
β_7	0	99	0.22	0	100	0	0	100	0
β_8	0	100	0.26	0	100	0	0	100	0

$a=2000$	$\lambda=0.5$			$\lambda=1$			$\lambda=2$		
	$\tilde{\beta}$	CP	ALen	$\tilde{\beta}$	CP	ALen	$\tilde{\beta}$	CP	ALen
β_0	0	96	0.68	0	97	0.28	0	98	0.40
β_1	2.83	97	0.79	2.67	98	0.33	2.33	98	0.49
β_2	1.36	93	0.82	1.22	93	0.33	0.94	92	0.48
β_3	0	99	0.25	0	100	0.01	0	100	0
β_4	0	100	0.26	0	100	0	0	100	0
β_5	1.78	93	0.74	1.56	95	0.30	1.11	93	0.45
β_6	0	100	0.24	0	100	0	0	100	0
β_7	0	100	0.20	0	100	0	0	100	0
β_8	0	100	0.23	0	100	0	0	100	0

Table 3.2: Coverage and length of 95% individual CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 3.5.1.

		$a=2$						$a=2000$					
		$\lambda=0.5$		$\lambda=1$		$\lambda=2$		$\lambda=0.5$		$\lambda=1$		$\lambda=2$	
	True	CP	ALen	CP	ALen	CP	ALen	CP	ALen	CP	ALen	CP	ALen
β_0^{true}	0	96	0.68	99	0.23	98	0.34	96	0.68	97	0.28	98	0.40
β_1^{true}	3	96	0.78	95	0.27	98	0.42	96	0.79	98	0.33	99	0.49
β_2^{true}	1.5	91	0.86	96	0.30	100	0.51	93	0.87	96	0.36	98	0.52
β_3^{true}	0	96	0.84	91	0.28	95	0.42	95	0.85	94	0.34	98	0.50
β_4^{true}	0	91	0.84	97	0.28	99	0.42	90	0.85	99	0.34	100	0.50
β_5^{true}	2	94	0.84	94	0.29	100	0.44	94	0.85	98	0.36	100	0.54
β_6^{true}	0	93	0.84	97	0.28	99	0.41	91	0.84	96	0.34	100	0.49
β_7^{true}	0	96	0.83	94	0.28	98	0.41	96	0.84	99	0.34	100	0.49
β_8^{true}	0	93	0.76	97	0.26	100	0.39	92	0.76	100	0.32	100	0.46

Table 3.3: Coverage and length of 95% individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 3.5.1.

$\bar{\beta}_{200} = \bar{\beta}_{300} = 1.5$, $\bar{\beta}_5 = \bar{\beta}_{95} = 2$, $\bar{\beta}_{10} = 1$, $\bar{\beta}_{25} = 0.5$, and all the other components are 0. We generate a $(\mathbf{y}, \mathbf{X})$ dataset with 100 replications of sample size $N = 100$. We consider six MCP penalties with parameters $\lambda = 0.5, 1$ or 2, and $a = 2$ or 2000. In each replication, we compute two types of individual confidence intervals both with confidence level 0.95 as before. One is for the population penalized parameters, i.e. the solution to the problem (3.60); and the other is for the true parameters $(\beta_0^{true}, \beta^{true})$ in the underlying linear model (3.46). Define the active set as $\mathcal{A} = \{j : \bar{\beta}_j \neq 0\} = \{1, 2, 5, 10, 25, 95, 100, 200, 300\}$ and $\mathcal{A}^c = \{0, 1, 2, \cdots, p\} \backslash \mathcal{A}$. For each type of confidence intervals, we report the average coverage, median coverage, average length and median length of the individual confidence intervals corresponding to coefficients in

$a=$	$\lambda=0.5$				$\lambda=1$				$\lambda=2$			
2	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen
$\mathcal{A}$	79.78	89.00	0.43	0.41	93.00	93.00	0.67	0.72	85.67	90.00	1.14	1.20
$\mathcal{A}^c$	100.00	100.00	0.02	0.02	100.00	100.00	0.00	0.00	100.00	100.00	0.01	0.00
$a=$	$\lambda=0.5$				$\lambda=1$				$\lambda=2$			
2000	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen
$\mathcal{A}$	88.11	88.00	0.52	0.53	92.11	92.00	0.71	0.75	92.11	92.00	0.94	0.99
$\mathcal{A}^c$	99.97	100.00	0.05	0.05	100.00	100.00	0.03	0.03	100.00	100.00	0.02	0.02

Table 3.4: Coverage and length of 95% individual CIs for $(\tilde{\beta}_0, \tilde{\beta})$ in Example 3.5.2.

$a=$	$\lambda=0.5$				$\lambda=1$				$\lambda=2$			
2	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen
$\mathcal{A}$	89.78	90.00	0.44	0.44	93.22	93.00	0.60	0.57	92.33	94.00	1.23	1.19
$\mathcal{A}^c$	93.53	94.00	0.38	0.38	93.90	94.00	0.50	0.50	94.04	94.00	1.05	1.05
$a=$	$\lambda=0.5$				$\lambda=1$				$\lambda=2$			
2000	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen
$\mathcal{A}$	90.00	90.00	0.56	0.56	93.89	94.00	0.81	0.82	94.33	94.00	1.33	1.33
$\mathcal{A}^c$	92.43	93.00	0.45	0.45	93.27	94.00	0.70	0.70	93.64	94.00	1.19	1.19

Table 3.5: Coverage and length of 95% individual CIs for $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 3.5.2.

either $\mathcal{A}$ or $\mathcal{A}^c$:

$$\begin{aligned}
\text{Avgcov } \mathcal{A} &= |\mathcal{A}|^{-1} \sum_{j \in \mathcal{A}} \text{CP}_j, & \text{Avgcov } \mathcal{A}^c &= |\mathcal{A}^c|^{-1} \sum_{j \in \mathcal{A}^c} \text{CP}_j, \\
\text{Avglen } \mathcal{A} &= |\mathcal{A}|^{-1} \sum_{j \in \mathcal{A}} \text{ALen}_j, & \text{Avglen } \mathcal{A}^c &= |\mathcal{A}^c|^{-1} \sum_{j \in \mathcal{A}^c} \text{ALen}_j, \\
\text{Medcov } \mathcal{A} &= \underset{j \in \mathcal{A}}{\text{median}} \{\text{CP}_j\}, & \text{Medcov } \mathcal{A}^c &= \underset{j \in \mathcal{A}^c}{\text{median}} \{\text{CP}_j\}, \\
\text{Medlen } \mathcal{A} &= \underset{j \in \mathcal{A}}{\text{median}} \{\text{ALen}_j\}, & \text{Medlen } \mathcal{A}^c &= \underset{j \in \mathcal{A}^c}{\text{median}} \{\text{ALen}_j\},
\end{aligned}$$

where CP_j and ALen_j respectively represent the empirical coverage probability and average interval length of the confidence intervals for β_j among the 100 replications. Results are listed in Table 3.4 and 3.5

3.5.3 Example 3.5.3: Prostate cancer data

In this example, we consider the prostate cancer dataset [49] and compute confidence intervals of confidence level 0.95 for six MCP penalties with parameters (λ, a) as $\lambda = 0.14$, 0.45 and 1.49, $a = 2$ and 2000. We standardized the data and split observations into two parts. One part consists of 67 observations, which is the training set used in [20]. We only use these 67

$\lambda = 0.14$	$a = 2$				$a = 2000$			
	Est1	Ind CI1	Est2	Ind CI2	Est1	Ind CI1	Est2	Ind CI2
β_0	2.47	[2.30, 2.64]	2.47	[2.30, 2.64]	2.46	[2.30, 2.63]	2.46	[2.30, 2.63]
β_1	0.60	[0.33, 0.87]	0.65	[0.42, 0.89]	0.55	[0.34, 0.76]	0.66	[0.43, 0.89]
β_2	0.25	[0, 0.59]	0.26	[0.05, 0.47]	0.22	[0.04, 0.41]	0.26	[0.08, 0.45]
β_3	-0.02	[-0.29, 0]	-0.12	[-0.30, 0.06]	0	[-0.13, 0.02]	-0.11	[-0.28, 0.06]
β_4	0.17	[0, 0.59]	0.21	[-0.03, 0.46]	0.13	[0, 0.34]	0.21	[0, 0.43]
β_5	0.23	[0, 0.72]	0.30	[0.02, 0.59]	0.19	[0, 0.42]	0.31	[0.06, 0.55]
β_6	0	[-0.08, 0]	-0.22	[-0.42, -0.03]	0	[-0.03, 0]	-0.21	[-0.41, -0.01]
β_7	0	[-0.01, 0.05]	-0.01	[-0.23, 0.21]	0	[0, 0.06]	-0.01	[-0.23, 0.20]
β_8	0.06	[0, 0.35]	0.22	[0.01, 0.44]	0.08	[0, 0.26]	0.23	[0.02, 0.43]

$\lambda = 0.45$	$a = 2$				$a = 2000$			
	Est1	Ind CI1	Est2	Ind CI2	Est1	Ind CI1	Est2	Ind CI2
β_0	2.48	[2.29, 2.66]	2.48	[2.29, 2.66]	2.47	[2.28, 2.65]	2.47	[2.28, 2.65]
β_1	0.76	[0.50, 1.01]	0.70	[0.47, 0.94]	0.53	[0.30, 0.77]	0.70	[0.44, 0.95]
β_2	0.16	[0, 0.38]	0.27	[0.08, 0.46]	0.18	[0.02, 0.33]	0.28	[0.10, 0.46]
β_3	0	[0, 0]	-0.11	[-0.29, 0.07]	0	[0, 0]	-0.09	[-0.29, 0.11]
β_4	0	[0, 0.13]	0.20	[-0.01, 0.41]	0	[0, 0.15]	0.21	[-0.01, 0.42]
β_5	0	[0, 0.10]	0.29	[0.03, 0.55]	0.08	[0, 0.32]	0.31	[0.04, 0.58]
β_6	0	[0, 0]	-0.24	[-0.48, 0.01]	0	[0, 0]	-0.20	[-0.44, 0.04]
β_7	0	[0, 0]	-0.05	[-0.30, 0.20]	0	[0, 0.05]	-0.01	[-0.24, 0.22]
β_8	0	[0, 0.09]	0.25	[-0.01, 0.51]	0	[0, 0.14]	0.24	[0, 0.48]

$\lambda = 1.49$	$a = 2$				$a = 2000$			
	Est1	Ind CI1	Est2	Ind CI2	Est1	Ind CI1	Est2	Ind CI2
β_0	2.46	[2.21, 2.71]	2.46	[2.21, 2.71]	2.46	[2.20, 2.72]	2.46	[2.20, 2.72]
β_1	0.21	[0, 0.56]	0.73	[0.36, 1.10]	0.16	[0, 0.45]	0.73	[0.35, 1.11]
β_2	0	[0, 0.04]	0.31	[0.08, 0.54]	0	[0, 0.07]	0.31	[0.08, 0.55]
β_3	0	[0, 0]	-0.05	[-0.37, 0.27]	0	[0, 0]	-0.04	[-0.38, 0.29]
β_4	0	[0, 0]	0.22	[-0.05, 0.49]	0	[0, 0]	0.22	[-0.06, 0.50]
β_5	0	[0, 0.02]	0.37	[0.04, 0.70]	0	[0, 0.05]	0.37	[0.04, 0.71]
β_6	0	[0, 0]	-0.15	[-0.47, 0.17]	0	[0, 0]	-0.13	[-0.45, 0.19]
β_7	0	[0, 0]	0.01	[-0.29, 0.30]	0	[0, 0]	0.02	[-0.28, 0.31]
β_8	0	[0, 0]	0.26	[-0.08, 0.60]	0	[0, 0]	0.27	[-0.08, 0.61]

Table 3.6: Estimates and 95% CIs of $(\tilde{\beta}_0, \tilde{\beta})$ and $(\tilde{\beta}_0^{true}, \tilde{\beta}^{true})$ in Example 3.5.3.

observations to compute parameter estimates and two types of individual confidence intervals, which are listed in Table 3.6. As we know about the MCP penalty, parameter a controls the degree of non-convexity and λ controls the level of penalization. As λ increasing, more and more parameter estimates and confidence intervals should shrink to 0 and singleton $\{0\}$ respectively. This is consistent with what we have observed in Table 3.1.

3.6 Summary

In this chapter we propose a unified method to construct confidence intervals of the population penalized parameters and the true model parameters, for a wide range of penalties which satisfy the three properties suggested by [15]. By transforming the problems (3.1) and (3.2) to their equivalent problems (3.5) and (3.21) respectively, we exclude the non-smoothness in the objectives. Therefore, we can obtain their normal map formulations and use the asymptotic results to derive confidence intervals. By correcting the bias introduced by the penalty term, we obtain asymptotic distribution of the true model estimator $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$ from the asymptotic result of the normal map solution z_N. The validity and effectiveness of the proposed method are proved by our theoretical and numerical results.

In practice, we solve for a SAA solution $(\hat{\beta}_0, \hat{\beta})$ and use (3.25) to obtain a solution to (3.24). Even if we can only find a locally optimal solution of the SAA problem (3.2) when the objective is non-convex, our method is still meaningful as long as the sample size N is large enough. The confidence intervals we computed are then for a locally optimal solution of the random design regularized regression problem (3.1). It is still challenging nowadays to find the globally optimal solution of a general non-convex optimization problem. In the literature, most of algorithms used to solve the SAA problem with non-convex penalties are approximation algorithms, such as MC+ algorithm [55]. Their goal is to efficiently find an approximate solution that is close to the global optimum. The problem is that (3.25) usually does not hold at these approximate solutions due to sensitivity problem of f_N. This refers us to MIQCP solver in GAMS. Study for a more efficient procedure computing a SAA solution with (3.25) satisfied could be a valuable future work, since MIQCP is not suitable in high dimensional cases.

CHAPTER 4: DISCUSSION

In this chapter, we will discuss two future research directions. The first direction in Section 4.1 is to conduct hypothesis testing for the population penalized parameters and the true model parameters. The second direction in Section 4.2 is to do inference for population constrained linear regression using variational inequality techniques.

4.1 Hypothesis testing for sparse penalized regression

In this dissertation, we obtained confidence intervals both for the population penalized parameters and the true model parameters. A nature question is that can we also do hypothesis tests for the population penalized parameters and for the true model parameters using same techniques, or even find p-value for those tests? Suppose we are interested in testing an individual null hypothesis $H_{0,i} : \beta_i^{true} = 0$ versus the alternative $H_{A,i} : \beta_i^{true} \neq 0$. From (3.5), we are ready to conduct this individual test and find the corresponding p-values when the asymptotic distribution is normal in Case I. In Case II, since the asymptotic distribution is not normal, how to do the hypothesis testing needs further investigation. Similarly, it is not trivial to study the hypothesis tests for the population penalized parameters, since its asymptotic distribution in (3.29) is not normal too.

4.2 Inference for population constrained linear regression

In linear regression problems, linear constraints are often added to the minimization problem for minimizing the mean squared error. For example, applications in finance and hyperspectral imaging often require the model parameter $\beta \in \mathbb{R}^p$ to be non-negative, i.e., $\beta \geqslant 0$. This example fits into the more general framework where the parameter β is subject to a set of inequality linear constrains which can be written as

$$C\beta \leqslant b,$$

where $C \in \mathbb{R}^{q\times p}$ and $b \in \mathbb{R}^q$ are constants. Therefore, we consider the following population version of the constrained linear regression by solving

$$\begin{aligned} \min_{\beta_0,\beta}\ & E[Y - \beta_0 - \textstyle\sum_{j=1}^p \beta_j X_j]^2, \\ \text{s.t.}\ \ & C\beta \leqslant b. \end{aligned} \tag{4.1}$$

Denote the feasible set of (4.1) as S, and define a function $F : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^{p+1}$ by

$$F(\beta_0, \beta, X, Y) = \begin{bmatrix} -2(Y - \beta_0 - \sum_{j=1}^p \beta_j X_j) \\ -2(Y - \beta_0 - \sum_{j=1}^p \beta_j X_j)X_1 \\ \vdots \\ -2(Y - \beta_0 - \sum_{j=1}^p \beta_j X_j)X_p \end{bmatrix}. \tag{4.2}$$

Clearly, F is a continuously differentiable function, and its derivative with respect to (β_0, β) at (β_0, β, X, Y) is given by

$$d_1F(\beta_0, \beta, t, X, Y) = \begin{bmatrix} 2 & 2X^T \\ 2X & 2XX^T \end{bmatrix}, \tag{4.3}$$

Next, define a function $f_0 : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^{p+1}$ by

$$f_0(\beta_0, \beta) = E[F(\beta_0, \beta, X, Y)]. \tag{4.4}$$

Then we can rewrite (4.1) as the following variational inequality:

$$-f_0(\beta_0, \beta) \in N_S(\beta_0, \beta). \tag{4.5}$$

Let $(f_0)_S$ be the normal map induced by f_0 and S. The population version of normal map formulation for (4.5) is

$$(f_0)_S(z) = 0, \tag{4.6}$$

where z is a variable of dimension $p+1$.

We can do similar transformation for the sample version of problem (4.1). Our goal is to obtain the asymptotic distribution, such as (3.28), for the solution to the sample version of normal map formulation for (4.5). Based on that we will further construct confidence intervals for the solution to (4.1) via substituting unknown quantities in the obtained asymptotic distribution by their reliable estimates. The difficulty of this problem is that it is hard to compute $d\Pi_S(\cdot)$ and its estimate due to the general form of the feasible set S.

REFERENCES

[1] R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, K. Zhang, and L. Zhao. What you can learn from wrong causal models. handbook of causal analysis for social research. s. morgan, ed. pages 403–424, 2013b.

[2] R. Berk, L. B. Brown, and L. Zhao. Statistical inference after model selection. *Journal of Quantitative Criminology*, 26:217–236, 2010.

[3] R. Berk, L. D. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Annals of Statistics*, 41:802–837, 2013.

[4] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36:199–227, 2008.

[5] H.D. Bondell and B.J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2007.

[6] L. Breiman. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87:738–754, 1992.

[7] P. Buhlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 2012.

[8] P. Buhlmann and S. van de Geer. Statistics for high-dimensional data: Methods, theory and applications. 2011. Springer.

[9] E. J. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

[10] S. Chatterjee and P. Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106:608–625, 2011.

[11] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[12] A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis.* Springer Monographs in Mathematics. Springer, 2009.

[13] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics (with discussion)*, 32:407–499, 2004.

[14] F Facchinei and J. S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems.* Springer Series in Operations Research. Springer, New York, 2003.

[15] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[16] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

[17] J. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

[18] Wenjiang J. Fu. Penalized regression: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

[19] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag: New York, 2001.

[20] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, New York, 2001.

[21] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 2014. To appear.

[22] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37:4254–4278, 2009.

[23] M. Lamm, S. Lu, and A. Budhiraja. Individual confidence intervals for true solutions to stochastic variational inequalities. *Submitted*, 2014.

[24] J. Lee, D. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference with the lasso. 2014. Published online before print at http://arxiv.org/abs/1311.6238.

[25] H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:21–59, 2005.

[26] H. Leeb and B. M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics*, 34:2554–2591, 2006b.

[27] H. Leeb and B. M. Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24:338–376, 2008b.

[28] Y. Lin and H. H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models. *The Annals of Statistics*, 34:2272–2297, 2006.

[29] Y. Liu and Y. Wu. Variable selection via a combination of the l_0 and l_1 penalties. *Journal of Computional and Graphical Statistics*, 16:782–798, 2007.

[30] R. Lockhart, J. Taylor, R. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Annals of Statistics*, 42:413–468, 2014.

[31] Shu Lu. A new method to build confidence regions for solutions of stochastic variational inequalities. *Optimization*, 63(9):1431–1443, 2014.

[32] Shu Lu and Amarjit Budhiraja. Confidence regions for stochastic variational ienqualities. *Mathematics of Operations Research*, 38(3):545–568, 2013.

[33] J. Lv and Y. Fan. A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37:3498–3528, 2009.

[34] O. L. Mangasarian and S. Fromovitz. The fritz john necessary optimality conditions in the presence of equality and inequality constraints. *Journal of Mathematical Analysis and Applications*, 17:37–47, 1967.

[35] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[36] J. Minnier, L. Tian, and T. Cai. A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106:1371–1382, 2011.

[37] R. Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 99:758–765, 1984.

[38] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer Series in Operations Research. Springer-Verlag, New York, 1999.

[39] S. M. Robinson. Normal maps induced by linear transformations. *Mathematics of Operations Research*, 17:691–714, 1992.

[40] S. M. Robinson. Sensitivity analysis of variational inequalities by normal-map techniques. In Giannessi, F. and Maugeri, A., editor, *Variational Inequalities and Network Equilibrium Problems*, pages 257–269, New York, 1995. Plenum Press.

[41] Stephen M. Robinson. An implicit-function theorem for a class of nonsmooth functions. *Mathematics of Operations Research*, 16:292–309, 1991.

[42] R. Tyrrell Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

[43] R. Tyrrell Rockafellar and Roger J–B Wets. *Variational Analysis.* Springer-Verlag, Berlin, 1998.

[44] S. Rosset and J. Zhu. piecewise linear regularized solution paths. *The Annals of Statistics*, 35:1012–1030, 2007.

[45] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory.* Society for Industrial and Applied Mathematics and Mathematical Programming Society, Philadelphia, PA, 2009.

[46] X. Shen and H. Huang. Grouping pursuit in regression. *Journal of American Statistical Association*, 105:727–739, 2010.

[47] X. Shen, W. Pan, and R. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of American Statistical Association*, 107:223–232, 2012.

[48] T. Sun and C. H. Zhang. Scaled sparse linear regression. *Biometrika*, 99:879–898, 2012.

[49] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

[50] S. Van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202, 2014.

[51] H. Wang and C. Leng. Unified lasso estimation by least squares approximation. *Journal of American Statistical Association*, 102:1039–1048, 2007.

[52] T. T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2:224–244, 2008.

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

[54] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94:19–35, 2007.

[55] C. H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942, 2010.

[56] C. H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76:217–242, 2014.

[57] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37:3468–3497, 2009.

[58] Nengfeng Zhou and Ji Zhu. Group variable selection via a hierarchical lasso and its oracle property. *Statistics and Its Interface*, 3:557–574, 2010.

[59] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

[60] H. Zou and T Hastie. Regularization and variable selection via the elastic net. *Annals of Statistics*, 67:301–320, 2005.