Confidence Intervals and Regions for the LASSO Using Stochastic Variational Inequality Techniques in Optimization

Shu Lu, Yufeng Liu†, Liang Yin, and Kai Zhang*
University of North Carolina at Chapel Hill

Abstract

Sparse regression techniques have been popular in recent years due to their ability in handling high dimensional data with built-in variable selection. The LASSO (Tibshirani, 1996) is perhaps one of the most well-known examples. Despite intense work in this direction, how to provide valid inference for sparse regularized methods remains to be a challenging statistical problem. In this paper, we take a unique point of view of this problem and propose to make use of stochastic variational inequality techniques in optimization to derive confidence intervals and regions for the LASSO. Some theoretical properties of the procedure are obtained. Both simulated and real data examples are used to demonstrate the performance of the proposed method.

Keywords: Confidence intervals, confidence regions, optimization, random design, shrinkage, sparsity, variational inequality.

† Corresponding author (Email: yfliu@email.unc.edu)
* Shu Lu is Associate Professor, Department of Statistics and Operations Research, Yufeng Liu is Professor, Department of Statistics and Operations Research, Department of Genetics, Carolina Center for Genome Sciences, Department of Biostatistics, Liang Yin is Senior Revenue Science Analyst, FedEx, and Kai Zhang is Assistant Professor, Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599. The authors would like to thank the Editor Professor Ingrid Van Keilegom, the Associate Editor, and reviewers for their constructive comments. This research was partly supported by US National Science Foundation grants DMS-1407241 (Liu, Lu and Yin), DMS-1109099 (Lu and Yin), DMS-1309619 (Zhang), DMS-1127914 (Zhang as a faculty fellow at the Statistical and Applied Mathematical Sciences Institute); and US National Institute of Health grants R01 CA-149569, and P01 CA-142538.
1 Introduction

Variable selection and inference are two central issues in modern research on linear models. Theory and methodology that are developed in recent years are generally guided by the following two aspects: data-driven model selection, and statistical inference on the selected model. These two aspects are particularly important when many covariates are available for the model.

Regarding data-driven model selection procedures, sparse regularized techniques have been popular for simultaneous variable selection and prediction. By introducing biases on the 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 (Donoho and Johnstone, 1994; Tibshirani, 1996), as well as many other extensions (e.g. Fan and Li (2001); Candes and Tao (2007)). See Buhlmann and van de Geer (2011) for a comprehensive review.

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. This problem has long been recognized and discussed recently by Breiman (1992); Berk et al. (2013).

In recent years, many methods have been developed to achieve valid inference after LASSO. We categorize these methods into the following three types of approaches: 1. 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 (Chatterjee and Lahiri, 2011; Minnier et al., 2011; Berk et al., 2013). 2. 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 Buhlmann (2013); Zhang and Zhang (2014); Van de Geer et al. (2014); Javanmard and Montanari (2014). 3. 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 (Lockhart et al., 2014; Lee et al., 2014).

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

$$\min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{1}$$

where \( \{x_{ij}\}_{j=1}^{p}, y_i \) are i.i.d. samples with \( x_{ij}, y_i \in \mathbb{R} \), \( \lambda > 0 \) is the tuning parameter, and \( \beta_0 \in \mathbb{R} \) and \( \beta = (\beta_1, \cdots, \beta_p) \) are the regression parameters. Our interest is on its inference. To that end, we study the following population version of the LASSO by solving

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

where \( X \in \mathbb{R}^p \) is the random input vector, \( Y \in \mathbb{R} \) is the response variable. We refer to (1) as the sample average approximation (SAA) problem of the population LASSO problem (2). We denote the solution to the SAA problem (1) as \((\hat{\beta}_0, \hat{\beta})\), which we refer to as LASSO estimators. Our first contribution is to make use of the LASSO estimators \((\hat{\beta}_0, \hat{\beta})\) to derive confidence intervals and regions for the population LASSO parameters \((\tilde{\beta}_0, \tilde{\beta})\), the solution of (2).

The population LASSO approach is closely related to the traditional least squares approach. When \( \lambda \) is 0, the problem (2) becomes the population least squares problem \( \min_{\beta_0, \beta} E[Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j]^2 \), 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 linear model \( Y = \beta_0^{true} + X^T \beta^{true} + \varepsilon \), with \( E[\varepsilon|X] = 0 \), then the solution to the population least squares problem is exactly \((\beta_0^{true}, \beta^{true})\). When \( \lambda > 0 \), the solution to (2) is not exactly \((\beta_0^{true}, \beta^{true})\), but is related to \((\beta_0^{true}, \beta^{true})\) in a different way. Our second contribution is to develop a method which utilizes that relation to construct confidence intervals for the parameters \((\beta_0^{true}, \beta^{true})\) in the linear model.

Why could the minimizer from a population LASSO 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 LASSO 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. Moreover, using the full model could be
questionable in areas such as social science (Berk et al., 2013b). 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 to dictate the inclusion or exclusion of variables for the inference (even when $n > p$). In this case, a data-driven model from LASSO would be helpful. 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. From a modelling perspective, this consideration means that the loss function has changed from the squared error loss to the sum of the square error and the regularization on $\beta_j$’s, as the objective function in (1). The corresponding risk function, which is the expected loss, is then the objective function in (2). Based on this consideration, the population LASSO parameters defined in (2) are precisely the minimizer of the risk function, and thus become a reasonable target for the inference.

The major difference between the population LASSO 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 regularization term with $\lambda$ as a measure of the strength of such information. Thus, the parameters in the population LASSO approach are both scientifically and statistically meaningful: They lead to the best approximations to the response when external information is available.

Our study on the inference of the population LASSO parameters is based on study of the asymptotic distribution of LASSO estimators (i.e., solutions to (1)), as they converge to the population LASSO parameters (the solution to (2)). A good understanding of such asymptotics around the population LASSO parameters will in turn provide important insights for the inference of true parameters $(\beta_0^{\text{true}}, \beta^{\text{true}})$. We also note here that inferences for the population LASSO parameters are by themselves meaningful probabilistic statements of practical use.

- LASSO estimators from (1) 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 LASSO parameters and the computed LASSO estimators. Sizes of those intervals are jointly determined by sample variability and sensitivity of LASSO 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 LASSO estimators obtained from (1).

- The inference results of this paper can be used to assess the relative importance of predictors. For nonzero LASSO 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 LASSO estimators, the inference results can be highly informative as well. For example, if the confidence intervals of some LASSO parameters are singletons of zero (see Section 3.2 for related discussion), then we have strong evidence to conclude that the corresponding population LASSO parameters are zero.

Besides inference for the LASSO parameters, we also develop an inference method for the true parameters \((\beta_0^{\text{true}}, \beta^{\text{true}})\). Our method is based on a relationship between \(\hat{\beta} - \beta^{\text{true}}\) as well as their sample counterparts. The details of this relationship can be found in Section 4. To help explain our method, we can take the viewpoint of the following decomposition: \(\hat{\beta} - \beta^{\text{true}} = \hat{\beta} - \tilde{\beta} + \tilde{\beta} - \beta^{\text{true}}\). Roughly speaking, this decomposition is similar as the bias-variance decomposition. In Sections 2 and 3, through the population LASSO approach, we are able to quantify the uncertainty in \((\hat{\beta} - \tilde{\beta})\) (or the “variance” part). Since the population LASSO parameters \(\tilde{\beta}\) is the asymptotic limit of the LASSO estimators \(\hat{\beta}\), the limiting distribution of \((\hat{\beta} - \tilde{\beta})\) characterizes the variation around \(\tilde{\beta}\) (see Section 3 for the theory and Figure 1 for an example). In Section 4, through a connection between \(\tilde{\beta}\) and \(\beta^{\text{true}}\) that corrects the “bias” in \((\tilde{\beta} - \beta^{\text{true}})\), we are able to provide valid inference for the true parameters. Simulation results in Section 5 show that our method performs competitively with existing methods.

The key techniques of our method are asymptotical results for solutions of (1) as the sample size grows. In this paper, we develop 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) and (2) into their corresponding normal map formulations (see (15) and (9) in Section 2 for more details), 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), 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 (2). Finally,
we convert these confidence intervals into confidence intervals for the solution to (2). Note that our inference method is developed for a fixed value of \( \lambda \). In practice, the value of \( \lambda \) can be chosen by various criteria or through cross validation. To provide inference for the true parameters \((\beta_{0}^{\text{true}}, \beta_{1}^{\text{true}})\), we establish a connection between the population LASSO parameters (which solves (2)) and the true parameters with (32). Estimators of the true parameters can be obtained in (33), and confidence intervals for the true parameters are then constructed based on the asymptotic distributions of those estimators.

The rest of the article is organized as follows. We first introduce some background on variational inequalities as well as the normal map formulations in Section 2, and then discuss construction of confidence intervals and regions for the population LASSO parameters in Section 3. In Section 4 we provide inference for the true parameters assuming \( X \) and \( Y \) are related by a linear model. Numerical results are presented in Section 5 to illustrate the performance of the proposed methods, followed by some discussion. The appendix collects technical details of variational inequalities, proofs, and additional numerical results.

In this paper, we use \( \langle x, y \rangle \) to denote the inner product between two vectors \( x \) and \( y \). 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. Let \( \mathcal{N}(0, \Sigma) \) denote a Normal random vector with covariance matrix \( \Sigma \). Weak convergence of random variables \( Y_n \) to \( Y \) is denoted as \( Y_n \Rightarrow Y \). A function \( g : \mathbb{R}^n \rightarrow \mathbb{R}^m \) is said to be B-differentiable at a point \( x_0 \in \mathbb{R}^n \) if there is a positively homogeneous function \( G : \mathbb{R}^n \rightarrow \mathbb{R}^m \), such that \( g(x_0 + v) = g(x_0) + G(v) + o(v) \). The 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 a stronger property 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.

2 Problem transformations

In this section, we describe how to transform (1) and (2) into quadratic programs, variational inequalities and their normal map formulations. Based on those transformations, we will obtain the asymptotic distribution of SAA solutions in Section 3. Appendix A gives background information about variational inequalities and normal maps.
2.1 Conversion to a standard quadratic program

In this subsection, we transform the population LASSO problem into a standard quadratic program. We need Assumption 1(a) below to guarantee the objective function of (2) to be finite valued. We will use the stronger Assumption 1(b) in proving convergence results.

**Assumption 1.** (a) The expectations $E[X_1^2], \cdots, E[X_p^2], \text{and} E[Y^2]$ are finite.

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

To eliminate the nonsmooth term $\sum_{j=1}^{p} |\beta_j|$ from the objective function of (2), we introduce a new variable $t \in \mathbb{R}^p$ into (2). The transformed problem is

$$
\min_{(\beta_0, \beta, t) \in S} E[Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j]^2 + \lambda \sum_{j=1}^{p} t_j
$$

where the feasible set $S$ of (3) is given by

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

In Section 2.2 we will transform (3) into a variational inequality. This requires writing down the gradient of its objective function. To this end, we define a continuously differentiable function $F: \mathbb{R}^{3p+2} \to \mathbb{R}^{2p+1}$ below, and write down its derivative with respect to $(\beta_0, \beta, t)$ at $(\beta_0, \beta, t, X, Y)$, denoted as $d_1 F(\beta_0, \beta, t, X, Y)$:

$$
F(\beta_0, \beta, t, X, Y) = 
\begin{bmatrix}
-2(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j) \\
-2(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j) X_1 \\
\vdots \\
-2(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j) X_p \\
\lambda e_p
\end{bmatrix},
\quad
\begin{bmatrix}
2 & 2X^T & 0 \\
2X & 2XX^T & 0 \\
0 & 0 & 0
\end{bmatrix},
\quad
\begin{bmatrix}
-2(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j) \\
-2(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j) X_1 \\
\vdots \\
-2(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j) X_p \\
\lambda e_p
\end{bmatrix}
$$

where $e_p \in \mathbb{R}^p$ is the vector of all ones. Next, define the expectations of $F$ and $d_1 F$ as

$$
f_0(\beta_0, \beta, t) = E[F(\beta_0, \beta, t, X, Y)] \text{ and } L = E[d_1 F(\beta_0, \beta, t, X, Y)].
$$

Assumption 1(a) guarantees that $f_0$ is a well-defined and finite valued function from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$. In fact, it is an affine function, and its Jacobian matrix is $L$.

The following lemma is straightforward and its proof is omitted.
Lemma 1. Suppose Assumption 1(a) holds. Then, the objective function of (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. Let $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ be an optimal solution of (2), define $\tilde{t} \in \mathbb{R}^p$ and $\tilde{q} \in \mathbb{R}^p$ by

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

Define an index set $\mathcal{I}$ as

$$
\mathcal{I} = \left\{ i \in \{1, \ldots, p\} \mid \text{ either } \{\tilde{\beta}_i \neq 0\}, \text{ or } \{\tilde{\beta}_i = 0 \text{ and } |\tilde{q}_i| = \lambda\} \right\},
$$

and let $Q$ be the submatrix of $L$ in (6) that consists of intersections of columns and rows of $L$ with indices in $\{1\} \cup \{i + 1, i \in \mathcal{I}\}$. Assume that $Q$ is nonsingular.

In the above assumption, the vector $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is indeed a solution of (3), and $Q$ is a submatrix of the upperleft $(p + 1) \times (p + 1)$ submatrix of $L$. Lemma 2 of Section 2.2 will show that the non-singularity of $Q$ guarantees $(\tilde{\beta}_0, \tilde{\beta})$ to be the globally unique solution of (2).

2.2 The variational inequality and normal map formulation

In view of Lemma 1, and based on the relation between constrained optimization and variational inequalities discussed in Appendix A, we can rewrite (3) as the following variational inequality:

$$
-f_0(\beta_0, \beta, t) \in N_S(\beta_0, \beta, t), \tag{7}
$$

where $N_S(\beta_0, \beta, t)$ is the normal cone to $S$ at $(\beta_0, \beta, t)$: in general, for a convex set $C$ and a point $x \in C$, the normal cone $N_C(x)$ is defined as $N_C(x) = \{v \in \mathbb{R}^n \mid \langle v, c - x \rangle \leq 0 \text{ for each } c \in C\}$. If we introduce multipliers for constraints defining $S$ in (4), we could write down an explicit expression for $N_S(\beta_0, \beta, t)$ and accordingly rewrite (7) into the well-known Karush-Kuhn-Tucker conditions. That approach would lead to more variables (the multipliers) in the formulation, and we deal with (7) directly in our approach. Next, we introduce the normal map formulation of (7). To this end, we define the normal map $(f_0)_S$ induced by $f_0$ and $S$ as follows

$$
(f_0)_S(z) = f_0(\Pi_S(z)) + (z - \Pi_S(z)) \text{ for each } z \in \mathbb{R}^{2p+1}, \tag{8}
$$

where $\Pi_S(z)$ denotes the Euclidean projection of $z$ onto $S$, namely the point in $S$ nearest to $z$ in Euclidean norm. The normal map $(f_0)_S$ as defined above is a function from $\mathbb{R}^{2p+1}$ and $\mathbb{R}^{2p+1}$.
Because $S$ is a polyhedral convex set in $\mathbb{R}^{2p+1}$, the Euclidean projector $\Pi_S$ is a piecewise affine function from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$: it coincides with an affine function on each of finitely many $(2p+1)$-dimensional polyhedrons whose union is $\mathbb{R}^{2p+1}$ (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 $\mathbb{R}^{2p+1}$ (the relative interior of a convex set is its interior relative to its affine hull).

Since $f_0$ is affine under Assumption 1(a), and $\Pi_S$ is piecewise affine, the normal map $(f_0)_S$ is a piecewise affine function from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$. With $z$ being a variable of dimension $2p+1$, the normal map formulation for (7) is the following equation:

$$(f_0)_S(z) = 0.$$  

(9)

As noted right below Assumption 2, the vector $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a solution of (3). It is therefore a solution of (7) as well. By the relation between variational inequalities and normal maps (see Appendix A), the point $z_0 \in \mathbb{R}^{2p+1}$ defined as

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

(10)

is a solution to (9) 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

$$K = \{w \in T_S(\Pi_S(z_0)) \mid \langle z_0 - \Pi_S(z_0), w \rangle = 0 \} = \{w \in T_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \mid \langle f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}), w \rangle = 0 \},$$  

(11)

where $T_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is the tangent cone to $S$ at the point $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$. In general, for a convex set $C$ and a point $x \in C$, $T_C(x)$ is defined as (for more details see Appendix A)

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

Using the special polyhedral structure of $S$, we will give an explicit expression of $K$ in the proof of Lemma 2 below. Critical cones are commonly used to express optimality conditions. We will use critical cones to express the asymptotic distribution of SAA solutions. Let $L_K$ be the normal map induced by the linear function defined by the matrix $L$ in (6) and the cone $K$, defined as in (8) with $L$ and $K$ in place of $f_0$ and $S$ respectively. In Lemma 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 in Theorem 1.

**Lemma 2.** Suppose that Assumptions 1(a) and 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 (3).

In the rest of this paper we use $\Sigma_0$ to denote the covariance matrix of $F(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}, X, Y)$, and let $\Sigma_1^0$ be the upper left $(p + 1) \times (p + 1)$ submatrix of $\Sigma_0$. Since the last $p$ elements of $F(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}, X, Y)$ are fixed at $\lambda$, entries of $\Sigma_0$ that are not contained in $\Sigma_1^0$ are all zeros.

### 2.3 Transformations of the SAA problems

So far we have reformulated (2) as a quadratic program (3), a variational inequality (7), and an equation involving the normal map (9). We can reformulate the SAA problem (1) in a similar way. By introducing the variable vector $t$, we rewrite (1) as the following problem:

$$
\min_{(\beta_0, \beta, t) \in S} \frac{1}{N} \sum_{i=1}^{N} \left[ y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right]^2 + \lambda \sum_{j=1}^{p} t_j,
$$

where $S$ is as defined in (4). For each $i$, let $x_i = (x_{ij})_{j=1}^{p}$ be the $p$-dim vector consisting of observed features of the $i$th sample. We define the SAA function $f_N(\beta_0, \beta, t) = N^{-1} \sum_{i=1}^{N} F(\beta_0, \beta, t, x_i, y_i)$, where $F$ is as in (5). By noting that $f_N(\beta_0, \beta, t)$ is exactly the gradient of the objective function of (12) at $(\beta_0, \beta, t)$, we can rewrite (12) as a variational inequality

$$
0 \in f_N(\beta_0, \beta, t) + N_S(\beta_0, \beta, t).
$$

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} x_i^T / N & 0

2 \sum_{i=1}^{N} x_i / N & 2 \sum_{i=1}^{N} x_i x_i^T / N & 0

0 & 0 & 0
\end{bmatrix}.
$$

Finally, we let $(f_N)_S$ be the normal map induced by $f_N$ and $S$, and write the normal map formulation of (13) as

$$
(f_N)_S(z) = 0.
$$

In Section 3 we will discuss the asymptotic distributions and convergence rates of solutions of (13) and (15), and generate confidence regions and confidence intervals (CI) for solutions of (7)
and (9). While Assumptions 1 and 2 are sufficient for the asymptotic distribution results to hold, the convergence rate results require additional assumptions, given in Assumption 3(a-c) in Appendix B. Those assumptions will hold, for example, if $X$ and $Y$ are bounded. With those assumptions, we can show that solutions of (15) converge to the solution of (9) in probability at an exponential rate, which is useful in constructing estimates for an unknown quantity in the asymptotic distributions. Without Assumption 3, our method remains valid for situations in which the asymptotic distributions are normal.

3 Inference for population LASSO parameters

This section studies inference for solutions of the population LASSO problem (2) based on solutions to an SAA problem (1). Section 3.1 discusses the convergence behavior of solutions for the variational inequality (13) and the normal map formulation (15) of the SAA problem, and explains how to estimate the asymptotic distributions. Section 3.2 shows how to compute confidence regions and intervals for the solution to the normal map formulation (9) of (2), and how to convert those into confidence intervals for solutions of (2).

3.1 The convergence and distributions of SAA solutions

Theorem 1 below gives convergence properties and asymptotic distributions of solutions of the SAA problems (13) and (15). It shows under Assumptions 1 and 2 that equation (15) has a unique solution $z_N$ for sufficiently large $N$, and that $z_N$ converges almost surely to $z_0$ defined in (10). Correspondingly, the projection $\Pi_S(z_N)$ is the unique solution of (13), which converges almost surely to $(\beta_0, \bar{\beta}, \bar{t})$. This theorem also provides asymptotic distributions of $z_N$ and $\Pi_S(z_N)$, and gives their convergence rate in probability under Assumption 3(a-b) in Appendix B.

**Theorem 1.** Suppose Assumptions 1 and 2 hold. Then, for almost every $\omega \in \Omega$, there exists an integer $N_\omega$, such that for each $N \geq N_\omega$, the equation (15) has a unique solution $z_N$ in $\mathbb{R}^{2p+1}$, and the variational inequality (13) has a unique solution given by $(\beta_0, \bar{\beta}, \bar{t}) = \Pi_S(z_N)$. Moreover,

$$\lim_{N \to \infty} z_N = z_0 \ a.e., \quad \lim_{N \to \infty} (\beta_0, \bar{\beta}, \bar{t}) = (\beta_0, \bar{\beta}, \bar{t}) \ a.e.,$$

$$\sqrt{N}(z_N - z_0) \Rightarrow (L_K)^{-1}(N(0, \Sigma_0)), \quad \sqrt{N}(\beta_0, \bar{\beta}, \bar{t}) \Rightarrow N(0, \Sigma_0),$$

and

$$\sqrt{N}(\Pi_S(z_N) - \Pi_S(z_0)) \Rightarrow \Pi_K \circ (L_K)^{-1}(N(0, \Sigma_0)).$$
Suppose in addition that Assumption 3(a-b) holds. Then there exist positive real numbers $\epsilon_0, \delta_0, \mu_0, M_0$ and $\sigma_0$, such that the following inequality holds for each $\epsilon \in (0, \epsilon_0]$ and each $N$:

$$
\text{Prob}\left\{ \| (\hat{\beta}_0, \hat{\beta}, \hat{\ell}) - (\beta_0, \beta, \ell) \| < \epsilon \right\} \geq \text{Prob}\left\{ \| z_N - z_0 \| < \epsilon \right\}
$$

$$
\geq 1 - \delta_0 \exp\{-N\mu_0\} - \frac{M_0}{\epsilon^{2p+1}} \exp\left\{ -\frac{N\epsilon^2}{\sigma_0} \right\}.
$$

(19)

In the above theorem, $L_K$ is the normal map induced by the linear function $L$ in (6) and the critical cone $K$ in (11). Since $K$ is a polyhedral convex cone, the Euclidean projector $\Pi_K$ 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. By Lemma 2, $L_K$ is a global homeomorphism under Assumptions 1(a) and 2. The inverse function $(L_K)^{-1}$ is again a piecewise linear function. Equations (17) and (18) give the asymptotic distributions of $z_N$ and $\Pi_S(z_N) = (\hat{\beta}_0, \hat{\beta}, \hat{\ell})$ respectively. Equation (19) says that $z_N$ converges to $z_0$ in probability at an exponential rate.

The goal of this section is to develop a method to compute confidence intervals for $z_0$ and $(\hat{\beta}_0, \hat{\beta})$. After solving the LASSO (1) to find its solution $(\hat{\beta}_0, \hat{\beta})$, we let $\hat{\ell} = |\hat{\beta}|$ so that $(\hat{\beta}_0, \hat{\beta}, \hat{\ell})$ solves (12) and equivalently (13). We can then compute $z_N$ by

$$
z_N = (\hat{\beta}_0, \hat{\beta}, \hat{\ell}) - f_N(\hat{\beta}_0, \hat{\beta}, \hat{\ell}),
$$

(20)

which solves (15) and satisfies $(\hat{\beta}_0, \hat{\beta}, \hat{\ell}) = \Pi_S(z_N)$. With $z_N$ known, from (17) one can readily write down an expression for the confidence region of $z_0$ by using the $\chi^2$ distribution. That expression contains unknown objects $\Sigma_0$ and $L_K$, and we describe below how to estimate them.

We will estimate $\Sigma_0$ by $\Sigma_N$, the sample covariance matrix of $\{F(\hat{\beta}_i, \hat{\beta}, \hat{\ell}, x_i, y_i)\}_{i=1}^N$. Let $\Sigma_N^1$ be the upperleft $(p + 1) \times (p + 1)$ submatrix of $\Sigma_N$; we have $\Sigma_N = \begin{bmatrix} \Sigma_N^1 & 0 \\ 0 & 0 \end{bmatrix}$. The following lemma shows that $\Sigma_N$ converges to $\Sigma_0$ almost surely, and provides the convergence rate of $\Sigma_N$.

**Lemma 3.** Suppose that Assumptions 1 and 2 hold. Then $\Sigma_N$ converges to $\Sigma_0$ almost surely. If Assumption 3(a-c) holds additionally, then there exist positive real numbers $\delta_1, \mu_1, M_1$ and $\sigma_1$, such that the following inequality holds for each $\epsilon > 0$ and each $N$:

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

(21)
Estimation of the normal map \( L_K \) requires more understanding of its structure. By Robinson (1995), \( L_K \) is exactly \( d(f_0)_S(z_0) \), the B-derivative (defined at the end of Section 1) of the normal map \( (f_0)_S \) at \( z_0 \). With \( L = df_0(z_0) \) defined in (6), the chain rule of B-differentiability gives

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

where \( d\Pi_S(z_0) \) is the B-derivative of the Euclidean projector \( \Pi_S \) at \( z_0 \) and satisfies \( d\Pi_S(z_0) = \Pi_K \). Depending on the structure of \( K \), there are two different cases:

- **Case I** (the linear case). When \( z_0 \) belongs to the interior of some \((2p+1)\)-cell in the normal manifold of \( S \) (so that \( \Pi_S \) coincides with an affine function in a neighborhood of \( z_0 \)), \( K \) is a subspace of \( \mathbb{R}^{2p+1} \). In this case, both \( \Pi_K = d\Pi_S(z_0) \) and \( L_K = d(f_0)_S(z_0) \) are linear functions, and \( \sqrt{N}(z_N - z_0) \) asymptotically follows a normal distribution based on (17). Note that \( d(f_N)_S(z_N)(h) = L_N \ d\Pi_S(z_N)(h) + h - d\Pi_S(z_N)(h) \) for each \( h \in \mathbb{R}^{2p+1} \). Since \( z_N \) converges to \( z_0 \) almost surely and \( \Pi_S \) coincides with an affine function around \( z_0 \), \( d\Pi_S(z_N) \) and \( d(f_N)_S(z_N) \) almost surely converge to \( d\Pi_S(z_0) \) and \( L_K \) respectively, and we can use \( d\Pi_S(z_N) \) and \( d(f_N)_S(z_N) \) as estimators for \( d\Pi_S(z_0) \) and \( L_K \) respectively.

- **Case II** (the piecewise linear case). When \( z_0 \) lies on the boundary of some \((2p+1)\)-cell in the normal manifold of \( S \), \( K \) is not a subspace of \( \mathbb{R}^{2p+1} \), \( L_K \) and \( d\Pi_S(z_0) \) are piecewise linear functions, and the asymptotic distribution of \( \sqrt{N}(z_N - z_0) \) is not normal. In this case, \( d\Pi_S(z) \) is not continuous with respect to \( z \) at \( z_0 \), and we do not have the almost sure convergence of \( d\Pi_S(z_N) \) and \( d(f_N)_S(z_N) \). Below, we show how to construct estimators \( \Lambda_N(z_N) \) for \( d\Pi_S(z_0) \) and \( \Phi_N(z_N) \) for \( L_K \) by exploiting the exponential convergence rate.

We define two functions \( \Lambda_N \) and \( \Phi_N \) from \( \mathbb{R}^{2p+1} \times \mathbb{R}^{2p+1} \) to \( \mathbb{R}^{2p+1} \) as follows. For each fixed \( z \in \mathbb{R}^{2p+1} \), \( \Lambda_N(z) \) and \( \Phi_N(z) \) are functions from \( \mathbb{R}^{2p+1} \) to \( \mathbb{R}^{2p+1} \). Let \( N \) be a given integer, and define a function \( g(N) = N^3 \). (For other possible choices of \( g(N) \) and details for constructing \( \Lambda_N \), see Appendix C.) For each \( z \in \mathbb{R}^{2p+1} \), consider all cells in the normal manifold of \( S \) that are within a distance of \( 1/g(N) \) from \( z \), and among those cells choose the cell with the smallest dimension. For all points \( z' \) in the relative interior of the latter cell, \( d\Pi_S(z') \) is the same function, which we define as \( \Lambda_N(z) \). We then define \( \Phi_N : \mathbb{R}^{2p+1} \times \mathbb{R}^{2p+1} \rightarrow \mathbb{R}^{2p+1} \) as

\[
\Phi_N(z)(h) = L_N \ \Lambda_N(z)(h) + h - \Lambda_N(z)(h)
\]
for each \( z \in \mathbb{R}^{2p+1} \) and \( h \in \mathbb{R}^{2p+1} \), where \( L_N \) is defined in (14). For a given \( N \), \( \Lambda_N \) is a fixed function, while \( \Phi_N \) depends on random samples since \( L_N \) does. Theorem 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.** Suppose that Assumptions 1, 2 and 3(a-b) hold. Then

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

and there exists a positive real number \( \phi \) such that

\[
\lim_{N \to \infty} \text{Prob} \left\{ \sup_{h \in \mathbb{R}^{2p+1}} \frac{\| \Phi_N(z_N)(h) - L_K(h) \|}{\| h \|} < \frac{\phi}{g(N)} \right\} = 1. \tag{23}
\]

We can now replace \( L_K \) in (17) by \( \Phi_N(z_N) \) without affecting the convergence.

**Theorem 3.** Suppose that Assumptions 1, 2 and 3(a-b) hold. Then

\[
\sqrt{N}\Phi_N(z_N)(z_N - z_0) \Rightarrow \mathcal{N}(0, \Sigma_0). \tag{24}
\]

If \( \Sigma_0^1 \) is nonsingular, then

\[
\sqrt{N} \begin{bmatrix} (\Sigma_0^1)^{-1/2} & 0 \\ 0 & I_p \end{bmatrix} (\Phi_N(z_N))(z_N - z_0) \Rightarrow \mathcal{N}(0, I_{p+1} \times 0). \tag{25}
\]

If \( \Sigma_0^1 \) is singular and Assumption 3(c) holds, then let \( l \) be the number of positive eigenvalues of \( \Sigma_0 \) counted with regard to their algebraic multiplicities, and decompose \( \Sigma_N^1 \) as

\[
\Sigma_N^1 = U_N^T \Delta_N U_N, \tag{26}
\]

where \( U_N \) is an orthogonal \((p+1) \times (p+1)\) matrix, and \( \Delta_N \) is a diagonal matrix with monotonically decreasing elements. Let \( D_N \) be the upper-left submatrix of \( \Delta_N \) 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 \( N \to \infty \), with

\[
N[\Phi_N(z_N)(z_N - z_0)]^T \begin{bmatrix} (U_N)_1^T D_N^{-1}(U_N)_1 & 0 \\ 0 & 0 \end{bmatrix} \left[ ((\Phi_N(z_N))(z_N - z_0)) \right] \Rightarrow \chi_l^2, \tag{27}
\]

and

\[
N[\Phi_N(z_N)(z_N - z_0)]^T \begin{bmatrix} (U_N)_2^T (U_N)_2 & 0 \\ 0 & I_p \end{bmatrix} \left[ ((\Phi_N(z_N))(z_N - z_0)) \right] \Rightarrow 0. \tag{28}
\]
The above theorem deals with two cases separately, depending on whether \( \Sigma_0^1 \) is nonsingular or not. In practice, since \( \Sigma_0^1 \) is unknown, we will always start by decomposing \( \Sigma_N^1 \) as in (26). If some eigenvalues of \( \Sigma_N^1 \) (i.e., diagonal elements of \( \Delta_N \)) are less than \( 1/g(N) \), then \( D_N \) is a proper submatrix of \( \Delta_N \), and we will use (27) and (28) to establish confidence intervals for \( z_0 \) (more details will be given in the following subsections). Otherwise, if all eigenvalues of \( \Sigma_N^1 \) are greater than or equal to \( 1/g(N) \), then \( D_N \) equals \( \Delta_N \) and (27) and (28) are equivalent to (25).

Note that we do not need to differentiate Cases I (single-piece) and II (multiple-piece) in Theorem 3, because \( \Phi_N(z_N) \) can be used as an estimator for \( L_K \) in both cases. In Case I, with high probability \( \Lambda_N(z_N) \) and \( \Phi_N(z_N) \) are linear functions and coincide with \( d\Pi_S(z_N) \) and \( d(f_N)_S(z_N) \) respectively.

### 3.2 Computation of confidence intervals

This subsection discusses how to obtain individual and simultaneous confidence intervals for \( z_0 \) based on (24), (27) and (28), and how to convert those into confidence intervals for \( (\tilde{\beta}_0, \tilde{\beta}) \).

One can show under assumptions in Theorem 2 that \( \Phi_N(z_N) \) is with high probability a global homeomorphism, and use

\[
(\Phi_N(z_N))^{-1}(N(0, \Sigma_N))
\]

(29)

to approximate the distribution of \( \sqrt{N}(z_N - z_0) \) and compute individual confidence intervals for \( z_0 \); see Lamm et al. (2016) for justifications for such an approach.

When \( \Phi_N(z_N) \) is a linear function from \( \mathbb{R}^{2p+1} \) to \( \mathbb{R}^{2p+1} \), the distribution in (29) is normal. In such situations, we let \( m_i \) be the \( i \)th diagonal element of the matrix \( (\Phi_N(z_N))^{-1}\Sigma_N(\Phi_N(z_N))^{-T} \), and use \[
\left[ (z_N)_i - N^{-1/2} \sqrt{\chi^2_1(\alpha)} m_i, (z_N)_i + N^{-1/2} \sqrt{\chi^2_1(\alpha)} m_i \right]
\]
as an approximate \((1 - \alpha)100\%\) confidence interval for \((z_0)_i\). Here and in what follows, \( \chi^2_n(\alpha) \) is the quantile that satisfies \( P(U > \chi^2_1(\alpha)) = \alpha \) for a \( \chi^2 \) random variable \( U \) with \( n \) degrees of freedom.

When \( \Phi_N(z_N) \) is piecewise linear, one way to generate confidence intervals for \( z_0 \) is by simulating data based on the distribution in (29), ordering data by each component and finding intervals to cover a specified percentage of points. See Appendix C.4 for details on how to compute \( (\Phi_N(z_N))^{-1}(q) \) for a given vector \( q \).

Simultaneous confidence intervals for all components of \((z_0)_i\) can be computed based on (27) and (28), by using the \( \chi^2 \) distribution to obtain a confidence region and then finding the minimal bounding box of that region. For details and for an alternative approach, see Appendix C.5.
After computing confidence intervals for \( z_0 \), we transform them into confidence intervals for the population LASSO parameters \((\tilde{\beta}_0, \tilde{\beta})\), based on the relation between \( z_0 \) and \((\tilde{\beta}_0, \tilde{\beta})\) derived in Appendix C.6. It is shown there that \( \tilde{\beta}_0 = (z_0)_1 \), so confidence intervals of \((z_0)_1\) are exactly those of \( \tilde{\beta}_0 \). For each \( i = 1, \cdots, p \), \( \tilde{\beta}_i \) and \((z_0)_{i+1}\) satisfy the following relation:

\[
\tilde{\beta}_i = \begin{cases} 
(z_0)_{i+1} - \lambda & \text{if } (z_0)_{i+1} > \lambda, \\
0 & \text{if } (z_0)_{i+1} \in [-\lambda, \lambda], \\
(z_0)_{i+1} + \lambda & \text{if } (z_0)_{i+1} < -\lambda.
\end{cases} 
\tag{30}
\]

Let us denote the right hand side of (30) 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 \( \Gamma \) as confidence intervals of \( \tilde{\beta}_i \). Note that \( \Gamma(\cdot) \) takes the constant value of 0 on \([-\lambda, \lambda]\). As a result, when the confidence interval for \((z_0)_{i+1}\) is contained entirely in \([-\lambda, \lambda]\), the confidence interval for \( \tilde{\beta}_i \) is a singleton \( \{0\} \). When the confidence interval for \((z_0)_{i+1}\) intersects with a part of \([-\lambda, \lambda]\), the confidence interval for \( \tilde{\beta}_i \) will contain the true solution of (2) with a probability larger than the prescribed level.

4 Inference for true parameters in a linear model

In this section, we derive asymptotic results and individual confidence intervals for the true parameters in an underlying linear model based on the convergence theorems in Section 3.

Suppose the true linear model between \( X \) and \( Y \) is

\[
Y = \beta_0^{true} + X^T \beta^{true} + \varepsilon,
\tag{31}
\]

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 \( \varepsilon \) has mean zero and variance \( \sigma^2_\varepsilon \). Moreover, \( \varepsilon \) 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 \( \Sigma \), i.e., \( \Sigma = E(XX^T) \). If \( \Sigma \) is nonsingular, by plugging (31) into (10), with the fact that \( \tilde{\beta}_0 = (z_0)_1 \), we obtain

\[
\beta_0^{true} = (z_0)_1, \quad \beta^{true} = \frac{1}{2} \Sigma^{-1} (z_0)_{2:(p+1)} + \left[ I_p - \frac{1}{2} \Sigma^{-1} \right] \tilde{\beta},
\tag{32}
\]

where \((z_0)_{2:(p+1)}\) denotes the vector that consists of the second to \((p + 1)\)th entries of \( z_0 \). Expression (32) suggests the following estimators

\[
\hat{\beta}_0^{true} = (z_N)_1, \quad \hat{\beta}^{true} = \frac{1}{2} \hat{\Theta}(z_N)_{2:(p+1)} + \left[ I_p - \frac{1}{2} \hat{\Theta} \right] \hat{\beta},
\tag{33}
\]

15
where \( \hat{\Theta} \) is an estimator of the precision matrix \( \Sigma^{-1} \). From (20) and (33) one may notice that the estimator \((\hat{\beta}_0^{\text{true}}, \hat{\beta}^{\text{true}})\) can be expressed as the sum of the LASSO estimator \( \hat{\beta} \) and a bias-correction term, which is similar to the de-biased estimators in Zhang and Zhang (2014) and Van de Geer et al. (2014). However, our inference method is different. Our approach is based on analysis of the transformed problems, and developed under different assumptions that allow the asymptotic distribution considered to be non-normal. In the simulation, Van de Geer et al. (2014) used scaled LASSO to obtain the LASSO estimate \( \hat{\beta} \) and the estimate \( \hat{\sigma}_e^2 \) of the error variance, while we use the original LASSO to calculate \( \hat{\beta} \) with a tuned parameter \( \lambda \). Numerical results show that our method is competitive with existing methods, and performs better in certain cases especially for high dimensions (see Table 4).

Next, let a matrix \( B \in \mathbb{R}^{(p+1) \times (2p+1)} \) be given by \( B = \begin{bmatrix} I_{p+1} & 0 \end{bmatrix} \), and define two functions \( G \) and \( \hat{G} \) from \( \mathbb{R}^{2p+1} \) to \( \mathbb{R}^{p+1} \) as

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

and

\[
\hat{G} = \frac{1}{2} \left( \begin{bmatrix} 1 & 0 \\ 0 & \hat{\Theta} \end{bmatrix} B + \begin{bmatrix} 1 & 0 \\ 0 & 2I - \hat{\Theta} \end{bmatrix} B \circ d\Pi_S(z_N) \right).
\]

Since \( S \) is a cone, \( \Pi_S \) is positively homogeneous, so we have \( \Pi_S(z_0) = d\Pi_S(z_0) \) and \( \Pi_S(z_N) = d\Pi_S(z_N) \). It follows from (32), (33), \((\hat{\beta}_0, \hat{\beta}, \hat{t}) = \Pi_S(z_0) \) and \((\hat{\beta}_0, \hat{\beta}, \hat{t}) = \Pi_S(z_N) \) that we can rewrite (32) and (33) as \((\beta_0^{\text{true}}, \beta^{\text{true}}) = G(z_0) \) and \((\hat{\beta}_0^{\text{true}}, \hat{\beta}^{\text{true}}) = \hat{G}(z_N) \). The following theorem shows that (33) gives a consistent estimator of the true parameter \((\beta_0^{\text{true}}, \beta^{\text{true}}) \), and provides an asymptotic distribution from which we can conduct inference for \((\beta_0^{\text{true}}, \beta^{\text{true}}) \).

**Theorem 4.** Suppose that Assumptions 1 and 2 hold, and the true covariance matrix \( \Sigma \) is nonsingular. Let \( \hat{\Theta} \) be a \( \sqrt{N} \)-consistent estimator of \( \Sigma^{-1} \), and \( G \) be defined as in (34). Then \((\hat{\beta}_0^{\text{true}}, \hat{\beta}^{\text{true}}) \) is a consistent estimator of \((\beta_0^{\text{true}}, \beta^{\text{true}}) \) and

\[
\sqrt{N} \left( (\hat{\beta}_0^{\text{true}}, \hat{\beta}^{\text{true}}) - (\beta_0^{\text{true}}, \beta^{\text{true}}) \right) \Rightarrow G \circ (L_K)^{-1}(N(0, \Sigma_0)).
\]

There are many choices for \( \hat{\Theta} \). Some common choices are the inverse of the sample covariance matrix and the estimate of the precision matrix computed by the banding method (Bickel and Levina, 2008) or the penalized likelihood method (Yuan and Lin, 2007; Friedman et al., 2008).
It is well known that under some regularity conditions, those estimators have \( \sqrt{N} \)-consistency when \( p \) is fixed (Lam and Fan, 2009). To use (36) to compute confidence intervals, we replace \( G \) and \( L_K \) by their estimators. For Case I (the linear case defined in Section 3.1), the following theorem gives an approach to compute the asymptotically exact individual confidence intervals for \( (\beta_0^{\text{true}}, \beta_i^{\text{true}}) \).

**Theorem 5.** Suppose that Assumptions 1 and 2 hold, the true covariance matrix \( \Sigma \) is nonsingular, and the solution to the normal map formulation (9) satisfies conditions for Case I. Let \( \hat{\Theta} \) be a \( \sqrt{N} \)-consistent estimator of \( \Sigma^{-1} \), and define \( H = G(L_K)^{-1} \) and \( H_N = \hat{G}[d(f_N)s(z_N)]^{-1} \). For each \( i = 0, 1, \cdots, p \), if \( (H\Sigma_0H^T)_{i+1,i+1} \neq 0 \) then

\[
\frac{\sqrt{N}(\hat{\beta}_i^\text{true} - \beta_i^\text{true})}{\sqrt{(H_N\Sigma_NH_N^T)_{i+1,i+1}}} \Rightarrow \mathcal{N}(0,1). \tag{37}
\]

Theorem 5 suggests constructing an asymptotically exact individual confidence interval for \( \beta_i^{\text{true}} \) with the significance level \( \alpha \) as

\[
\left[ \hat{\beta}_i^{\text{true}} - N^{-1/2}\sqrt{\lambda_i^2(\alpha)\overline{m}_i}, \hat{\beta}_i^{\text{true}} + N^{-1/2}\sqrt{\lambda_i^2(\alpha)\overline{m}_i} \right],
\]

where \( \overline{m}_i \) is the \((i+1)\)th diagonal element of the matrix \( H_N\Sigma_NH_N^T \).

Next we discuss how to compute individual confidence intervals for \( (\beta_0^{\text{true}}, \beta_i^{\text{true}}) \) in Case II (the piecewise linear case). 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 \( Z \sim \mathcal{N}(0, I_p+1) \times \mathbf{0} \). Let \( \alpha \in (0,1) \), and define \( a^\ast(f) \in (0, \infty) \) as

\[
a^\ast(f) = \inf \{ c \geq 0 \mid \text{Prob}\{ -c \leq f(Z) - r \leq c \} \geq 1 - \alpha \}. \tag{38}
\]

Suppose that \( \text{Prob}\{ f(Z) = b \} = 0 \) for all \( b \in \mathbb{R} \). Then for any given \( r \in \mathbb{R} \) and \( \alpha \in (0,1) \), \( a^\ast(f) \) as defined in (38) is the smallest value that satisfies \( \text{Prob}\{ -a^\ast(f) \leq f(Z) - r \leq a^\ast(f) \} = 1 - \alpha \).

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

\[
R = G \circ (L_K)^{-1} \begin{bmatrix} (\Sigma_0^1)^{1/2} & 0 \\ 0 & I_p \end{bmatrix} \quad \text{and} \quad \hat{R} = \hat{G}' \circ (\Phi_N(z_N))^{-1} \begin{bmatrix} (\Sigma_N^1)^{1/2} & 0 \\ 0 & I_p \end{bmatrix}, \tag{39}
\]

where

\[
\hat{G}' = \frac{1}{2} \left( \begin{bmatrix} 1 & 0 \\ 0 & \hat{\Theta} \end{bmatrix} B + \begin{bmatrix} 1 & 0 \\ 0 & 2I - \hat{\Theta} \end{bmatrix} B \circ \Lambda_N(z_N) \right). \tag{40}
\]

We denote the \( j \)th component function of \( R \) and \( \hat{R} \) as \( R_j \) and \( \hat{R}_j \) respectively, for \( j = 1, \cdots, p+1 \).

Note that the map \( G \) is a piecewise linear function in Case II. From the expression (34) and the matrix representations of the piecewise linear function \( \Pi_K \) based on the location of \( z_0 \) (see
Appendix C), one can check that $G$ has the following form

$$
\begin{bmatrix}
1 & 0 \\
0 & \frac{1}{2}\Sigma^{-1}(I - W) + W
\end{bmatrix}^*,
$$

in which $W$ is a piecewise linear function represented by $p \times p$ diagonal matrices with diagonal elements 0 or $\frac{1}{2}$. If $\Sigma$ and $\Sigma_0^1$ 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 1(a) and 2, 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 6. Suppose Assumptions 1, 2 and 3(a-b) hold, and the covariance matrices $\Sigma$ and $\Sigma_0^1$ are nonsingular. Let $\hat{\Theta}$ be a $\sqrt{N}$-consistent estimator of $\Sigma^{-1}$, $\alpha \in (0, 1)$ and $r \in \mathbb{R}$. Let $a^r(\cdot)$ be as in (38), and define $R$ and $\hat{R}$ as in (39). Then for all $j = 0, 1, \ldots, p$, we have

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

In practice, for a fixed choice of $r$, we can find individual confidence intervals for $(\beta_0^{true}, \beta^{true})$ by simulating data from $\hat{R}(Z)$. We first generate data from $N(0, \Sigma_N)$, then compute $(\Phi_N(z_N))^{-1}(q)$ for a given vector $q$ as described in Appendix C.4, to obtain an empirical distribution of $\hat{R}(q) = \hat{G}^r \circ (\Phi_N(z_N))^{-1}(q)$ where $\hat{G}^r$ is defined in (40).

## 5 Numerical examples

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

5.1 Example 1: The asymptotic distribution of LASSO solutions

We generate 400 replications of 2000 observations from the model $Y = \beta^*TX + \sigma \varepsilon$, 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 $2 \times 2$ identity matrix and $J_2$ being the $2 \times 2$ matrix of 1’s, $\varepsilon \sim N(0, 1)$, and $\sigma = 3$. Here $X$ and $\varepsilon$ are independent of each other.

In Figure 1, we illustrate the piecewise Gaussian asymptotic distribution for the SAA solutions with $\lambda = 3$. In the left panel, the nine boundaries divide the plane into ten divisions, with
Figure 1: Distributions of SAA solutions in Example 1. The left panel plots the 400 points \(((z_N)_2, (z_N)_3)\) and displays boundaries of regions defined in (F.1) in the appendix with \(\alpha = 0.1, 0.2, \ldots, 0.9\). The true solution \(((z_0)_2, (z_0)_3) = (4, 3)\) is marked with a “+” sign. The right panel plots the corresponding 400 points \((\hat{\beta}_1, \hat{\beta}_2)\) and displays boundaries of regions defined in (F.2) with \(\alpha = 0.1, 0.2, \ldots, 0.9\). Short vertical lines on the horizontal axis are markers of the endpoints of intervals defined in (F.3) with the same \(\alpha\) values.

around 40 points (min 34, max 45, mean 40, std 3.62) in each division. In the right panel, the markers are not located at intersections between the curves and the horizontal axis, because the two regions (F.2) and (F.3) (in Appendix F.1) 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.

5.2 Example 2: Confidence intervals in low dimensional setting

In this example, we simulate data using the model in Example 1 of Tibshirani (1996). The model is the same as that of Example 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 \(\varepsilon\) is a standard normal random variable independent of \(X\). We generate 100 replications of 300 observations with \(\sigma = 1\), and compute two types of confidence intervals for three fixed \(\lambda\) 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 \((\beta_0, \beta_{true})\).
in the underlying linear model (31), both of significance levels \( \alpha = 0.1 \). For the second type intervals, we compare our method with two other approaches in the literature: the LDPE method (Zhang and Zhang, 2014; Van de Geer et al., 2014) and the method introduced by Javanmard and Montanari (2014) (referred to “JM” method). Similar to the LDPE method, we use nodewise LASSO regression introduced by Meinshausen and Bühlmann (2006) to compute the estimate of the precision matrix \( \hat{\Theta} \). In terms of the tuning parameter \( \lambda \), we check the performance of our method using GIC (Nishii, 1984). In the LDPE method, the model parameters are estimated by scaled LASSO without specifying \( \lambda \). 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.

For the first type of confidence intervals, we observe 100% coverage of simultaneous confidence intervals from all SAA problems with significance level 0.1. This is not very surprising, since the boxes formed by the simultaneous confidence intervals are much larger than the confidence regions of the specified probability levels enclosed in them. The coverage of individual confidence intervals is larger than \( (1 - \alpha)100\% \) in general with reasonable lengths (see Tables S.4 and S.5 in Appendix F.2). In Table 1, the intervals are not always symmetric around the estimates, a result of the non-normality. 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, the confidence intervals for the intercept \( \beta_0 \) 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 are quite

<table>
<thead>
<tr>
<th>( \lambda = 0.5 )</th>
<th>( \lambda = 1 )</th>
<th>( \lambda = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-0.08 [-0.18, 0.02] [-0.31, 0.15]</td>
<td>-0.10 [-0.21, 0.02] [-0.37, 0.17]</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>2.82 [2.72, 2.93] [2.58, 3.07]</td>
<td>2.65 [2.52, 2.78] [2.35, 2.95]</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>1.43 [1.32, 1.53] [1.18, 1.67]</td>
<td>1.28 [1.16, 1.40] [1.00, 1.56]</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>0 [0, 0] [0, 0.17]</td>
<td>0 [0, 0] [0, 0.12]</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>0 [0, 0] [0, 0.22]</td>
<td>0 [0, 0] [0, 0.09]</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>1.74 [1.64, 1.83] [1.52, 1.96]</td>
<td>1.51 [1.40, 1.63] [1.24, 1.79]</td>
</tr>
<tr>
<td>( \beta_6 )</td>
<td>0 [0, 0.08] [0, 0.31]</td>
<td>0 [0, 0] [0, 0.14]</td>
</tr>
<tr>
<td>( \beta_7 )</td>
<td>0 [0, 0.02] [0, 0.30]</td>
<td>0 [0, 0] [0, 0.04]</td>
</tr>
</tbody>
</table>

Table 1: 90% confidence intervals for population LASSO parameters computed by our method for different \( \lambda \) values from a typical replication of 300 observations generated in Example 2. The “Est” columns contain values of the SAA solution \( \hat{\beta}_0, \hat{\beta} \). The “Ind CI” and “Sim CI” columns give individual and simultaneous confidence intervals respectively.
covidence intervals for spectively represent the empirical coverage probability and average

(see Table S.7). For the second type

5.3 Example 3: Confidence intervals in high dimensional setting

This example considers a case in which the dimension $p$ is larger than the sample size. The simulation model is the same as that of Example 1, with $\beta^*$ 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 $\Sigma_{ij} = \rho|\xi - j|$ for $\rho = 0.9$, $\varepsilon$ is standard normal and independent of $X$, and $\sigma = 1$. We generate 100 replications of 300 observations, and consider three fixed $\lambda$ values, 0.5, 1 and 2, as well as the $\lambda$ value chosen by GIC for each SAA problem. As in Example 2, we compute the two types of individual confidence intervals both with the significance level 0.05. Define the active set as $\mathcal{A} = \{j : \beta_j^* \neq 0\} = \{1, 2, 5, 95, 100, 200, 300\}$ and $\mathcal{A}^c = \{1, 2, \ldots, p\} \setminus \mathcal{A}$. For each type of individual confidence intervals, we report the average coverage (Avgcov), median coverage (Medcov), average length (Avglen) and median length (Medlen) of the individual confidence intervals corresponding to parameters in either $\mathcal{A}$ or $\mathcal{A}^c$: Avgcov $\mathcal{A} = |\mathcal{A}|^{-1} \sum_{j \in \mathcal{A}} CP_j$, Avgcov $\mathcal{A}^c = |\mathcal{A}^c|^{-1} \sum_{j \in \mathcal{A}^c} CP_j$, Avglen $\mathcal{A} = |\mathcal{A}|^{-1} \sum_{j \in \mathcal{A}} ALen_j$, Avglen $\mathcal{A}^c = |\mathcal{A}^c|^{-1} \sum_{j \in \mathcal{A}^c} ALen_j$, Medcov $\mathcal{A} = \text{median}_{j \in \mathcal{A}}\{CP_j\}$, Medcov $\mathcal{A}^c = \text{median}_{j \in \mathcal{A}^c}\{CP_j\}$, Medlen $\mathcal{A} = \text{median}_{j \in \mathcal{A}}\{ALen_j\}$, Medlen $\mathcal{A}^c = \text{median}_{j \in \mathcal{A}^c}\{ALen_j\}$, where $CP_j$ and $ALen_j$ respectively represent the empirical coverage probability and average length of the confidence intervals for $\beta_j$ among the 100 replications (see Tables 3 and 4). For the second type

\begin{table}
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline
 & LDPE method & & JM method & & $\lambda = 0.49$ tuned by GIC & \\
 & True & Est & Ind CI & CP & Est & Ind CI & CP \\
\hline
$\hat{\beta}_0$ & 0 & - & - & - & -0.06 & [-0.16, 0.03] & 93 \\
\hline
$\beta_1^{true}$ & 3 & 3.00 & [2.90, 3.11] & 92 & 3.00 & [2.90, 3.10] & 92 \\
\hline
$\beta_2^{true}$ & 1.5 & 1.59 & [1.48, 1.70] & 92 & 1.59 & [1.48, 1.70] & 93 \\
\hline
$\beta_3^{true}$ & 0 & -0.06 & [-0.18, 0.05] & 88 & -0.08 & [-0.17, 0.02] & 92 \\
\hline
$\beta_4^{true}$ & 0 & 0.05 & [-0.06, 0.17] & 87 & 0.06 & [-0.04, 0.16] & 96 \\
\hline
$\beta_5^{true}$ & 2 & 1.91 & [1.79, 2.02] & 90 & 1.91 & [1.79, 2.02] & 90 \\
\hline
$\beta_6^{true}$ & 0 & 0.08 & [-0.03, 0.20] & 85 & 0.08 & [-0.02, 0.18] & 85 \\
\hline
$\beta_7^{true}$ & 0 & 0.03 & [-0.09, 0.14] & 90 & 0.01 & [-0.09, 0.11] & 95 \\
\hline
$\beta_8^{true}$ & 0 & 0.03 & [-0.07, 0.14] & 90 & 0.03 & [-0.06, 0.12] & 93 \\
\hline
\end{tabular}
\caption{90% individual confidence intervals and coverage for true parameters in the model (31) computed by different methods from a typical replication of 300 observations generated in Example 2. The “Est” columns contain the true parameter estimates ($\hat{\beta}_0^{true}, \hat{\beta}_1^{true}$), and the “True” column contains the true parameter ($\beta_0^{true}, \beta_1^{true}$). The empirical coverage probability with significance level 0.1 are collected in the “CP” columns.}
\end{table}
intervals, we compare the above measures computed from our method with those from the LDPE and JM methods (Table 4).

<table>
<thead>
<tr>
<th></th>
<th>( \lambda = 0.5 )</th>
<th></th>
<th>( \lambda = 1 )</th>
<th></th>
<th>( \lambda = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avgcov</td>
<td>Medcov</td>
<td>Avglen</td>
<td>Medlen</td>
<td>Avgcov</td>
</tr>
<tr>
<td>( \mathcal{A} )</td>
<td>91.86</td>
<td>94.00</td>
<td>0.92</td>
<td>0.92</td>
<td>91.57</td>
</tr>
<tr>
<td>( \mathcal{A}^c )</td>
<td>99.92</td>
<td>100.00</td>
<td>0.07</td>
<td>0.06</td>
<td>99.96</td>
</tr>
</tbody>
</table>

Table 3: Average coverage (%) and length of 95% individual confidence intervals for the population LASSO parameters computed by our method for different \( \lambda \) values from 100 replications of 100 observations and dimension \( p = 300 \) generated in Example 3.

<table>
<thead>
<tr>
<th></th>
<th>( \lambda = 0.5 )</th>
<th></th>
<th>( \lambda = 1 )</th>
<th></th>
<th>( \lambda = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avgcov</td>
<td>Medcov</td>
<td>Avglen</td>
<td>Medlen</td>
<td>Avgcov</td>
</tr>
<tr>
<td>( \mathcal{A} )</td>
<td>93.86</td>
<td>94.00</td>
<td>0.81</td>
<td>1.03</td>
<td>95.86</td>
</tr>
<tr>
<td>( \mathcal{A}^c )</td>
<td>92.85</td>
<td>93.00</td>
<td>0.75</td>
<td>0.74</td>
<td>93.44</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{A} )</td>
<td>88.43</td>
<td>89.00</td>
<td>1.04</td>
<td>1.06</td>
<td>84.71</td>
<td>83.00</td>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td>( \mathcal{A}^c )</td>
<td>95.13</td>
<td>95.00</td>
<td>1.07</td>
<td>1.07</td>
<td>98.94</td>
<td>99.00</td>
<td>0.87</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 4: Average coverage (%) and length of 95% individual confidence intervals for true parameters in the underlying linear model (31) computed by different methods from 100 replications of 100 observations and dimension \( p = 300 \) generated in Example 3.

As shown in Table 3, the first type confidence intervals are often conservative for the inactive variables. The same phenomena are observed in Example 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 \( \lambda \), many population LASSO parameters are exactly 0’s. Thus the SAA solutions of LASSO of these parameters concentrate closely around 0’s. Similarly, we also observe shorter confidence intervals for true parameters of the inactive set compared with those for true parameters of the active set, as shown in Table 4.

The top rows of Table 4 report results of our method with different \( \lambda \) values. One may notice that the length of confidence intervals for the true parameters \((\beta_0^{true}, \beta^{true})\) increases when \( \lambda \) increases. For an intuitive explanation, recall that the estimator \((\hat{\beta}_0^{true}, \hat{\beta}^{true})\) in (33) is a bias correction version of the LASSO solution \((\hat{\beta}_0, \hat{\beta})\). Large \( \lambda \) 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 \( \lambda \) is too small, the SAA solution lacks sparsity and the corresponding LASSO estimates are less reliable. This suggests choosing an intermediate
value of $\lambda$ to achieve the best overall performance.

The bottom rows of Table 4 show the results calculated from the LDPE method, the JM method and our method (with $\lambda$ 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 close to 95%, and the coverage from the JM method is even higher. 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 get better with larger sample sizes.

5.4 Example 4: Prostate cancer data

This subsection considers the prostate cancer example used in Hastie et al. (2001). We standardize the data and split observations into two parts. One part consists of 67 observations, which is the training set in Hastie et al. (2001). We only use these 67 observations in our computation. 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 5. The results for the population LASSO parameters can be found in Appendix F.3.

<table>
<thead>
<tr>
<th></th>
<th>LDPE method</th>
<th>JM method</th>
<th>$\lambda = 0.88$ tuned by GIC</th>
<th>$\lambda = 0.45$</th>
<th>$\lambda = 1.49$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>Ind CI</td>
<td>Est</td>
<td>Ind CI</td>
<td>Est</td>
</tr>
<tr>
<td>$\beta_{1}^{true}$</td>
<td>0.69</td>
<td>[0.46, 0.93]</td>
<td>0.68</td>
<td>[0.03, 1.33]</td>
<td>0.71</td>
</tr>
<tr>
<td>$\beta_{2}^{true}$</td>
<td>0.28</td>
<td>[0.09, 0.46]</td>
<td>0.26</td>
<td>[-0.22, 0.75]</td>
<td>0.29</td>
</tr>
<tr>
<td>$\beta_{3}^{true}$</td>
<td>-0.09</td>
<td>[-0.29, 0.11]</td>
<td>-0.14</td>
<td>[-0.66, 0.38]</td>
<td>-0.08</td>
</tr>
<tr>
<td>$\beta_{4}^{true}$</td>
<td>0.21</td>
<td>[0.01, 0.41]</td>
<td>0.21</td>
<td>[-0.31, 0.73]</td>
<td>0.22</td>
</tr>
<tr>
<td>$\beta_{5}^{true}$</td>
<td>0.31</td>
<td>[0.08, 0.54]</td>
<td>0.31</td>
<td>[-0.33, 0.94]</td>
<td>0.33</td>
</tr>
<tr>
<td>$\beta_{6}^{true}$</td>
<td>-0.21</td>
<td>[-0.48, 0.06]</td>
<td>-0.29</td>
<td>[-1.08, 0.50]</td>
<td>-0.19</td>
</tr>
<tr>
<td>$\beta_{7}^{true}$</td>
<td>-0.01</td>
<td>[-0.27, 0.25]</td>
<td>-0.02</td>
<td>[-0.76, 0.72]</td>
<td>-0.02</td>
</tr>
<tr>
<td>$\beta_{8}^{true}$</td>
<td>0.24</td>
<td>[-0.03, 0.51]</td>
<td>0.27</td>
<td>[-0.52, 1.05]</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table 5: 95% individual confidence intervals for true parameters in the model (31) computed by different methods from prostate cancer data with sample size $N = 67$ in Example 4. 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 $\beta_{1}^{true}, \beta_{2}^{true}, \ldots, \beta_{8}^{true}$.

Table 5 lists the estimates of the true model parameters and their individual confidence intervals, computed from our methods as well as the LDPE and JM methods with $\lambda = 0.88$ (tuned by GIC as in Example 2), 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 5, confidence intervals for \( \beta_1^{true}, \beta_2^{true}, \beta_4^{true} \) and \( \beta_5^{true} \) 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 \( \beta_1^{true} \). Across all three values of \( \lambda \), our methods always select \( \beta_1^{true}, \beta_2^{true} \) and \( \beta_5^{true} \) 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.

6 Discussion

In this paper, 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 and for the true parameters. Both our theoretical and numerical results confirm the validity and effectiveness of the proposed methods. In view of (29), the lengths of confidence intervals for population LASSO parameters are affected by two factors. The first is \( \Sigma_N \), the sample covariance of \( F(\hat{\beta}_0, \hat{\beta}, \hat{t}, x_i, y_i) \). The second is \( (\Phi_N(z_N))^{-1} \), which characterizes the sensitivity of solution to (1) 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. With respect to confidence intervals for the true regression parameters, our technique performs competitively compared with existing methods. Although our asymptotic results are developed with \( N \to \infty \) and fixed dimensions, our technique appears to perform reasonably well for moderately high dimensional examples. One possible research direction is to explore the high dimensional asymptotic properties by extending the proposed technique. Another research direction is to extend the technique to general sparse penalized methods.

References


