Published in final edited form as:

J Am Stat Assoc. 2020 ; 115(530): 794–809. doi:10.1080/01621459.2019.1585251.

Confidence Intervals for Sparse Penalized Regression with Random Designs

Guan Yu^{*}, Liang Yin^{*}, Shu Lu, Yufeng Liu[†]

The State University of New York at Buffalo University of North Carolina at Chapel Hill

Abstract

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 estimation. A number of convex as well as non-convex penalties have been proposed in the literature to achieve sparse estimates. Despite intense work in this area, how to perform valid inference for sparse penalized regression with a general penalty remains to be an active research problem. In this paper, by making use of state-of-the-art optimization tools in stochastic variational inequality theory, we propose a unified framework to construct confidence intervals for sparse penalized regression with a wide range of penalties, including convex and non-convex penalties. We study the inference for parameters under the population version of the penalized regression as well as parameters of the underlying linear model. Theoretical convergence properties of the proposed method are obtained. Several simulated and real data examples are presented to demonstrate the validity and effectiveness of the proposed inference procedure.

Keywords

confidence interval; non-convex penalty; penalized regression; random design; variational inequality

1 Introduction

With the advantage of simultaneous variable selection and estimation, sparse penalized regression techniques have been widely used. By introducing biases on the estimators, sparse penalized regression can often select a simpler model and produce estimators with smaller mean square errors than unpenalized regression. One of the well-known representatives is the L_1 penalized technique LASSO (Donoho and Johnstone, 1994; Tibshirani, 1996). LASSO has become a popular variable selection method due to its good selection performance and computational efficiency. Many other extensions with different penalties have been studied in the literature, see for example Fan and Li (2001); Zou and

[†]Corresponding author (yfliu@email.unc.edu).

^{*}Guan Yu and Liang Yin are both co-first authors. Guan Yu is Assistant Professor, Department of Biostatistics, The State University of New York at Buffalo; Liang Yin was a graduate student, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill; Shu Lu is Associate Professor, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill; Yufeng Liu is Professor, Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599.

Hastie (2005); Candes and Tao (2007); Liu and Wu (2007); Lv and Fan (2009); Zhang (2010).

For computational implementation of these methods, there is a large literature on efficient algorithms. The LARS algorithm by Efron et al. (2004) and the Coordinate-Descent algorithm by Wu and Lange (2008) are two popular examples. Mazumder et al. (2011) proposed the SparseNet algorithm to deal with non-convex penalties. In terms of inference, much less development has been made, especially for estimators from non-convex penalized regression. For the LASSO, one common approach is to first perform model selection and then carry out inference based on some pivot distributions conditional on the selected model. For example, see Lee et al. (2016); Lockhart et al. (2014). This approach does not sufficiently account for the stochastic errors in the model selection step. Another popular approach is to achieve valid inference by adjusting the bias introduced by the L_1 regularization term. Papers along this line include Javanmard and Montanari (2014); Van de Geer et al. (2014); Zhang and Zhang (2014). Recently, Lu et al. (2017) suggested using a variational inequality formulation to establish an asymptotic distribution of LASSO estimators that can be used to construct confidence intervals (CI) for the population LASSO parameters as well as the true model parameters. Some other methods about high dimensional inference include Voorman et al. (2014); Ning et al. (2017); Zhao et al. (2017).

For sparse penalized regression with penalties more complex than the LASSO, not much work has been done on inference. In particular, with a non-convex penalty, the regression problem may have multiple local optimal solutions. This brings up the question of whether one can use a local solution to construct meaningful confidence intervals. The goal of this paper is to provide a unified framework to perform valid inference for sparse penalized regression with general penalties, based on a local solution. Our assumptions about the penalties are in line with the desired properties for regularized penalty functions given in Fan and Li (2001), namely, sparsity, unbiasedness, and continuity. For example, this framework can be applied to the adaptive LASSO penalty (Zou, 2006), the non-convex log penalty (Friedman, 2012), SCAD (Fan and Li, 2001), MCP (Zhang, 2010), the transformed ℓ_1 penalty (Nikolova, 2000), and so on.

In this paper, we consider a general random-design penalized regression problem

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

where $X = (X_1, \dots, X_p)^T \in \mathbb{R}^p$ is an explanatory random vector with mean 0, and $Y \in \mathbb{R}$ is a response random variable. Here $\beta_0 \in \mathbb{R}$ and $\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ are the regression parameters. For $j = 1, 2, \dots, p$, $P_{\lambda_j}(\cdot)$ is a general penalty for β_j with the regularization parameter λ_j . This general penalty covers many convex and non-convex penalties.

The solution of (1) can be estimated by the solution of the corresponding sample average approximation (SAA) problem

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

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}, \quad \mathbf{1}_N = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^N,$$

and $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$ are independent samples of (X, Y) . For each $i = 1, \dots, N$ and $j = 1, \dots, p$, $x_{ij} \in \mathbb{R}, y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^{p \times 1}$. We refer to a local solution to the population penalized problem (1) as a population penalized parameter, and denote it as $(\tilde{\beta}_0, \tilde{\beta})$. We refer to a local solution to the SAA problem (2) as a penalized estimator, and denote it as $(\hat{\beta}_0, \hat{\beta})$.

The population penalized parameter $(\tilde{\beta}_0, \tilde{\beta})$ is closely related to the traditional least squares parameter. When $P_{\lambda_j}(|\beta_j|) = 0$ for $j = 1, \dots, p$, the problem (1) becomes the following population least squares problem:

$$\min_{\beta_0, \beta} E[Y - \beta_0 - \sum_{i=1}^p \beta_i X_i]^2, \quad (3)$$

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

$$Y = \beta_0^{true} + X^T \beta^{true} + \varepsilon \quad (4)$$

with $E[\varepsilon|X] = 0$, then the solution to the population least squares problem (3) is exactly $(\beta_0^{true}, \beta^{true})$, which we refer to as the true model parameter. When $P_{\lambda_j}(|\beta_j|) > 0$, $(\tilde{\beta}_0, \tilde{\beta})$ is not exactly $(\beta_0^{true}, \beta^{true})$, but there is a relation between $(\tilde{\beta}_0, \tilde{\beta})$ and $(\beta_0^{true}, \beta^{true})$, that will be described in Section 3.4.

The idea of our proposed method is to use the penalized estimator $(\hat{\beta}_0, \hat{\beta})$ to derive confidence intervals for the population penalized parameter $(\tilde{\beta}_0, \tilde{\beta})$, and then exploit the relation between $(\tilde{\beta}_0, \tilde{\beta})$ and $(\beta_0^{true}, \beta^{true})$ to derive confidence intervals for the true model parameter $(\beta_0^{true}, \beta^{true})$ in the linear model (4). Therefore, our proposed method could construct confidence intervals for both $(\tilde{\beta}_0, \tilde{\beta})$ and $(\beta_0^{true}, \beta^{true})$. Note that the valid inference of $(\tilde{\beta}_0, \tilde{\beta})$ is also useful for some problems such as the cost-effective linear regression taking into account the cost of collecting variables. For each sample, suppose we need to spend c_j dollars on collecting the value of the j th variable X_j where $j = 1, 2, \dots, p$. For this special linear

regression problem, in order to find a relatively cheap linear model with good prediction performance, we need to estimate the regression coefficient vector that minimizes an objective function which balances between the expected prediction accuracy and the data collation cost, that is,

$$(\tilde{\beta}_0, \tilde{\beta}) \in \arg \min_{\beta_0, \beta} E(Y - \beta_0 - \sum_{j=1}^p X_j \beta_j)^2 + \lambda \sum_{j=1}^p c_j P(|\beta_j|),$$

where $P(x)$ is a continuously differentiable non-convex function, and it is an approximation to the indicator function $\mathcal{I}(x)$ that equals to 1 if $x > 0$ and 0 if $x = 0$. The parameter λ can be selected according to the budget. In this example, the population penalized parameter becomes a reasonable target of the inference and our proposed method could deliver its asymptotically exact confidence intervals. On the other hand, even when a model on the relation between X and Y is not available, the confidence intervals of $(\tilde{\beta}_0, \tilde{\beta})$ can still provide measures on the randomness of the penalized estimators.

The main techniques to construct confidence intervals for $(\tilde{\beta}_0, \tilde{\beta})$ and $(\beta_0^{true}, \beta^{true})$, consist of three steps. First, we transform problems (1) and (2) into their corresponding variational inequality and normal map formulations and obtain an asymptotic distribution of a solution to the normal map formulation of (2). Next, by finding reliable estimates for quantities that describe the asymptotic distribution, we provide methods to compute confidence intervals for $(\tilde{\beta}_0, \tilde{\beta})$ based on a solution to the normal map formulation of (2). Finally, we establish the connection between $(\tilde{\beta}_0, \tilde{\beta})$ and $(\beta_0^{true}, \beta^{true})$, from which we obtain the bias-corrected estimator of $(\beta_0^{true}, \beta^{true})$, the asymptotic distribution and confidence intervals. The methodology in this paper is developed for a fixed dimension p , based on a local solution $(\hat{\beta}_0, \hat{\beta})$ to (2). The confidence intervals we obtain for $(\tilde{\beta}_0, \tilde{\beta})$ are for the local solution to (1) close to $(\hat{\beta}_0, \hat{\beta})$. On the other hand, for any local solution $(\hat{\beta}_0, \hat{\beta})$ of (2), we can always obtain confidence intervals for the true model parameter $(\beta_0^{true}, \beta^{true})$. Indeed, under the setting we consider in Section 3.4, a local solution $(\hat{\beta}_0, \hat{\beta})$ almost surely converges to $(\beta_0^{true}, \beta^{true})$.

Although our method and the method proposed in Lu et al. (2017) use similar techniques, there are important new contributions. First, we propose a unified framework to construct confidence intervals for a large class of penalties, including the method using the LASSO penalty (Lu et al. (2017)) as a special case. Second, for non-convex penalties, the construction of the confidence intervals and the theoretical studies are more complicated. We propose a new transformation of the original optimization problem to its corresponding variational inequality and normal map formulations. Special technical conditions and theoretical results for general penalties are studied. Third, the proposed method based on non-convex penalties could deliver better confidence intervals than the methods using convex penalties. For example, in our numerical studies, we compare the method using the MCP penalty with $a = 2$ and the method using the MCP penalty with $a = 2000$ (the case that MCP penalty is very close to the LASSO penalty). Our numerical results indicate that the lengths of the confidence intervals for $a = 2$ are generally shorter than those for $a = 2000$.

with similar coverage rates. This is possibly due to the smaller bias imposed by the MCP penalty with a small a .

The rest of this paper is organized as follows. In Section 2, we show some background on variational inequality, and present the problem transformations. Section 3 discusses how to obtain the confidence intervals for the population penalized parameters as well as the true model parameters in the linear model. Some theoretical results about convergence properties are shown in this section. In Section 4, we present numerical results to illustrate the performance of the proposed method. Section 5 contains some discussion. The technical details of variational inequalities and proofs are shown in the supplementary materials.

Throughout this paper, we use $\mathcal{N}(0, \Sigma)$ to denote a normal random vector with mean zero and covariance matrix Σ , and $Y_n \Rightarrow Y$ to represent weak convergence of a sequence of random variables $\{Y_n\}$ to Y . The inner product between two vectors x and y is denoted as $\langle x, y \rangle$. For a convex set S and a vector $z \in \mathbb{R}^n$, we use $\Pi_S(z)$ to denote the Euclidean projection onto S .

A function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is said to be B-differentiable at a point $x_0 \in \mathbb{R}^n$ if there exists a positively homogeneous function $df(x_0): \mathbb{R}^n \rightarrow \mathbb{R}^m$, such that $f(x_0 + v) = f(x_0) + df(x_0)(v) + o(v)$. The function $df(x_0)$ is called the B-derivative of f at x_0 .

2 Background and problem transformations

In this section, we first introduce some background on variational inequalities and normal maps. Then we introduce how to transform the problems (1) and (2) to their corresponding variational inequality and normal map formulations. Some assumptions for our theoretical analysis are also given in this section.

2.1 Background on variational inequalities and normal maps

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

$$0 \in f(x) + N_S(x), \quad (5)$$

where $N_S(x)$ is the normal cone to S at x defined as

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

Variational inequalities are closely related to optimization problems. Consider a problem of minimizing an objective function $F: \mathbb{R}^n \rightarrow \mathbb{R}$ over a closed and convex set $S \subseteq \mathbb{R}^n$. A well-known fact is that if $x \in S$ is a local solution to this minimization problem and F is differentiable at x , then x satisfies the variational inequality

$$0 \in \nabla F(x) + N_S(x),$$

where $\nabla F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the gradient of the function F . Conversely, if x satisfies the above variational inequality, and F is a convex function, then x is a global minimizer of F over the set S .

Besides the above connection with the original minimization problem, the variational inequality can be equivalently formulated as an equation, using a concept called the *normal map*. The normal map induced by f and S is a function $f_S: \mathbb{R}^n \rightarrow \mathbb{R}^n$ given by

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

where $\Pi_S(z)$ denotes the Euclidean projector of z onto S . For any solution $x \in S$ to the variational inequality (5), the point $z = x - f(x)$ satisfies $\Pi_S(z) = x$ and

$$f_S(z) = 0. \quad (6)$$

Conversely, for any solution z to (6), the point $x = \Pi_S(z)$ is a solution to (5) and satisfies $z = x - f(x)$. Equation (6) is called the normal map formulation of (5).

To understand the above relations, we consider an example of minimizing $F(x) = \frac{1}{2}\|x - x_0\|_2^2$ for a fixed point $x_0 \in \mathbb{R}^n$ over the convex set S . We know that the solution is the projection of x_0 onto the convex set S , denoted as $\Pi_S(x_0)$. For the function $F(x)$, the gradient is $\nabla F(x) = x - x_0$. Thus, the variational inequality formulation (5) of this minimization problem is the problem of finding $x \in S$ such that $x_0 - x \in N_S(x)$. That is, we need to find $x \in S$ such that $\langle x_0 - x, s - x \rangle \leq 0$ for each $s \in S$. We can show that the solution is $\Pi_S(x_0)$. On the other hand, the normal map induced by $\nabla F(x)$ and S is $f_S(z) = \nabla F(\Pi_S(z)) + z - \Pi_S(z) = \Pi_S(z) - x_0 + z - \Pi_S(z) = z - x_0$. Thus, the normal map formulation (6) of this minimization problem is $z - x_0 = 0$. The solution to this equation is x_0 and the point $x = \Pi_S(x_0)$ is the solution to the original minimization problem. More details about the variational inequality and normal map can be found in the supplementary materials. In the following Sections 2.2 and 2.3, we show how to transform the problems (1) and (2) to their corresponding normal map formulations, respectively.

2.2 Transformations of the population penalized regression

In this subsection, we transform the optimization problem (1) into a normal map formulation. Before discussing details about the transformation, we introduce conditions on the penalties $P_{\lambda_i}(\cdot)$. In this subsection, as well as in Sections 2.3 and 3.1-3.3, $\lambda = (\lambda_1, \dots, \lambda_p) > 0$ is fixed.

Assumption 1.

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

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

Many well-known penalty functions satisfy Assumption 1(a). We list five penalty functions as examples.

- (a) The adaptive LASSO penalty (Zou, 2006) defined as $P_{\lambda_i}(|\beta_i|) = \lambda_i |\beta_i|$, where λ_i is the weight for the i^{th} coordinate.
- (b) The log penalty (Friedman, 2012) defined as $P_{\lambda_i}(|\beta_i|) = \frac{\lambda_i}{\log(1+a)} \log(a|\beta_i| + 1)$, where $a > 0$.
- (c) The transformed ℓ_1 penalty (Nikolova, 2000) defined as $P_{\lambda_i}(|\beta_i|) = \lambda_i \frac{(a+1)|\beta_i|}{a+|\beta_i|}$, where $a > 0$.
- (d) The SCAD penalty (Fan and Li, 2001) defined as $P_{\lambda}(0) = 0$ and

$$P'_{\lambda_i}(|\beta_i|) = \lambda_i \mathbb{1}_{|\beta_i| \leq \lambda_i} + \frac{(a\lambda_i - |\beta_i|)_+}{a-1} \mathbb{1}_{|\beta_i| > \lambda_i} \text{ where } a > 2. \quad (7)$$

- (e) The MCP penalty (Zhang, 2010) defined as

$$P_{\lambda_i}(|\beta_i|) = \lambda_i \left(|\beta_i| - \frac{\beta_i^2}{2a\lambda_i} \right) \mathbb{1}_{|\beta_i| < a\lambda_i} + \frac{a\lambda_i^2}{2} \mathbb{1}_{|\beta_i| \geq a\lambda_i} \text{ where } a > 0. \quad (8)$$

We can check that penalties (a), (b), and (c) satisfy Assumption 1(b). The SCAD and MCP penalties satisfy this assumption almost everywhere. Take the SCAD penalty for example. The function corresponds to a quadratic spline with two knots, at which it is not continuously twice differentiable. Assumption 1(b) requires that no local solution to (1) locates at these two knots for each i . It is a reasonable assumption since the set of points on which twice continuous differentiability fails has measure zero.

In the assumption below, part (a) is to ensure the objective function of (1) to be finite valued, and part (b) will be used in proving convergence results.

Assumption 2.

- (a) The expectations $E[X_1^2], \dots, E[X_p^2]$ and $E[Y^2]$ are finite.
- (b) The expectations $E[X_1^4], \dots, E[X_p^4]$ and $E[Y^4]$ are finite.

Next, we transform the problem (1) into a normal map formulation in three steps. In the first step, we introduce an equivalent problem, in which a new variable $t \in \mathbb{R}^p$ is added to eliminate the non-smooth term $\sum_{i=1}^p P_{\lambda_i}(|\beta_i|)$ from the objective function (1). The new problem is as follows:

$$\begin{aligned}
& \min_{\beta_0, \beta, t} E[Y - \beta_0 - \sum_{i=1}^p \beta_i X_i]^2 + \sum_{i=1}^p P_{\lambda_i}(t_i) + m(\|t\|_2^2 - \|\beta\|_2^2) \\
& \text{s.t.} \quad t_i - \beta_i \geq 0, \quad i = 1, \dots, p, \\
& \quad \quad t_i + \beta_i \geq 0, \quad i = 1, \dots, p,
\end{aligned} \tag{9}$$

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

$$S_i = \{(\beta_i, t_i) | t_i - \beta_i \geq 0, t_i + \beta_i \geq 0\}, \quad i = 1, \dots, p, \tag{10}$$

and write

$$(\beta_0, \beta, t) = (\beta_0, \beta_1, t_1, \beta_2, t_2, \dots, \beta_p, t_p), \tag{11}$$

then we can treat the feasible set of (9), denoted by S , as a Cartesian product

$$S = \mathbb{R} \times \prod_{i=1}^p S_i. \tag{12}$$

We will use the two ways of ordering of (β_0, β, t) in (11) interchangeably for notational convenience.

Note that the above transformation is different from the method shown in Lu et al. (2017) for the LASSO penalty. The term $m(\|t\|_2^2 - \|\beta\|_2^2)$ is added into the objective function of (9) in order to ensure $t_i = |\beta_i|$ in any optimal solution to (9), so that there is an one-to-one correspondence between the optimal solutions to (1) and (9). This is necessary and important when the penalty functions are not strictly increasing on $[0, +\infty)$. For instance, some non-convex penalties such as SCAD and MCP are flat on some intervals $[d_i, +\infty)$.

In the second step, we transform (9) into a variational inequality. To this end, we need to write down the gradient of its objective function. Define a function

$F: \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R} \rightarrow \mathbb{R}^{2p+1}$ as

$$F(\beta_0, \beta, t, X, Y) = \begin{bmatrix} -2(Y - \beta_0 - \sum_{i=1}^p \beta_i X_i) \\ -2(Y - \beta_0 - \sum_{i=1}^p \beta_i X_i)X - 2m\beta \\ (P'_{\lambda_i}(t_i) + 2mt_i)_{i=1}^p \end{bmatrix}. \tag{13}$$

Furthermore, define a function $f_0: \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^{2p+1}$ as

$$f_0(\beta_0, \beta, t) = E[F(\beta_0, \beta, t, X, Y)]. \tag{14}$$

The function f_0 is well defined and finite valued under Assumption 2(a). If $P_{\lambda_i}(t_i)$ is twice differentiable at t_i for every $i = 1, \dots, p$, then we can write down the derivative of F with respect to (β_0, β, t) as

$$d_1 F(\beta_0, \beta, t, X, Y) = \begin{bmatrix} 2 & 2X^T & 0 \\ 2X & 2XX^T - 2mI_p & 0 \\ 0 & 0 & \text{diag}(P''_{\lambda_i}(t_i) + 2m)_{i=1}^p \end{bmatrix}, \quad (15)$$

where $\text{diag}(P''_{\lambda_i}(t_i) + 2m)_{i=1}^p$ represents the diagonal matrix whose i th diagonal element is $P''_{\lambda_i}(t_i) + 2m$ and I_p is the $p \times p$ identity matrix. Moreover, the Jacobian matrix of f_0 is

$$L(t) = E[d_1 F(\beta_0, \beta, t, X, Y)] = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2E[XX^T] - 2mI_p & 0 \\ 0 & 0 & \text{diag}(P''_{\lambda_i}(t_i) + 2m)_{i=1}^p \end{bmatrix}. \quad (16)$$

The lemma below shows that there is an one-to-one correspondence between the (local or global) optimal solutions to (1) and (9).

Lemma 1.—*Suppose Assumptions 1(a) and 2(a) hold. Then the objective function of (9) is finite valued on $\mathbb{R}^{2p+1}$, and its gradient at each $(\beta_0, \beta, t) \in \mathbb{R}^{2p+1}$ is $f_0(\beta_0, \beta, t)$. If $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a (local) optimal solution to (9), then $\tilde{t}_i = |\tilde{\beta}_i|$ for all $i = 1, \dots, p$, and $(\tilde{\beta}_0, \tilde{\beta})$ is a (local) optimal solution to (1). Conversely, if $(\tilde{\beta}_0, \tilde{\beta})$ is a (local) optimal solution to (1), then $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a (local) optimal solution to (9), where $\tilde{t}_i = |\tilde{\beta}_i|$ for all $i = 1, \dots, p$.*

If Assumption 1(b) holds additionally, then the Hessian matrix of the objective function of (9) at $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is $L(\tilde{t})$.

In view of Lemma 1, we can transform (9) to the following variational inequality:

$$-f_0(\beta_0, \beta, t) \in N_S(\beta_0, \beta, t). \quad (17)$$

In the last step, we state the normal map formulation for (17). Let $(f_0)_S$ be the normal map induced by f_0 and S . Then the normal map formulation for (17) is

$$(f_0)_S(z) = 0, \quad z \in \mathbb{R}^{2p+1}. \quad (18)$$

For the rest of the paper, let $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ be a local solution to (9). Then $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is also a solution to (17). Therefore, the point $z_0 \in \mathbb{R}^{2p+1}$ defined as

$$z_0 = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) - f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \quad (19)$$

is a solution to (18) and satisfies $\Pi_S(z_0) = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$.

Let Σ_0 be the covariance matrix of $F(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}, X, Y)$. We can check that Σ_0 is well defined if Assumption 2(b) holds. 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 not random at $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, we have $\Sigma_0 = \begin{bmatrix} \Sigma_0^1 & 0 \\ 0 & 0 \end{bmatrix}$. In our theoretical analysis as shown in Section 3, we found that the B-derivative of the normal map $(f_0)_S$ at z_0 plays an important role in the construction of the confidence intervals. To study the property of the B-derivative of the normal map $(f_0)_S$ at z_0 , we need the following assumption.

Assumption 3.—Let $(\tilde{\beta}_0, \tilde{\beta})$ be a local solution to (1), define $\tilde{t} \in \mathbb{R}^p$ and $\tilde{q} \in \mathbb{R}^p$ by

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

Let $\mathfrak{S}$ be a subset of $\{1, \dots, p\}$ defined as

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

and denote $L(\tilde{t})$ in (16) by L . Let Q_1 be the submatrix of L that consists of intersections of columns and rows of L with indices in $\{1\} \cup \{i+1, i \in \mathfrak{S}\}$, and let Q_2 be the submatrix of L that consists of intersections of columns and rows of L with indices in $\{i+p+1, i \in \mathfrak{S}\}$. Define matrix Q as

$$Q = Q_1 + \begin{bmatrix} 0 & 0 \\ 0 & Q_2 \end{bmatrix}. \quad (20)$$

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 . If $(\tilde{\beta}_0, \tilde{\beta})$ is a solution to the optimization problem (1), then for every

$i \in \{1, \dots, p\}$, we have $\tilde{\beta}_i \neq 0$ and $|\tilde{q}_i| = |P'_{\lambda_i}(\tilde{t}_i)|$, or $\tilde{\beta}_i = 0$ and $|\tilde{q}_i| \leq |P'_{\lambda_i}(\tilde{t}_i)|$. Since we have $\tilde{\beta}_i = 0$ and $|\tilde{q}_i| < |P'_{\lambda_i}(\tilde{t}_i)|$ for some coefficients, $|\mathfrak{S}|$ is generally not equal to p . Furthermore, in our theoretical analysis, we assume that the dimension p is fixed, the matrix Q can be nonsingular in many cases. The nonsingularity of Q is a standard assumption to guarantee that $(\tilde{\beta}_0, \tilde{\beta})$ is a locally unique optimal solution.

As shown in Robinson (1995), the B-derivative of the normal map $(f_0)_S$ at z_0 is the same as the normal map L_K induced by the linear function defined by the matrix L and the critical cone K to S associated with z_0 , defined as

$$\begin{aligned}
 K &= \{w \in T_S(\Pi_S(z_0)) \mid \langle z_0 - \Pi_S(z_0), w \rangle = 0\} \\
 &= \{w \in T_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \mid \langle f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}), w \rangle = 0\},
 \end{aligned} \tag{21}$$

where for each $x \in S$,

$T_S(x) = \{w \in \mathbb{R}^n \mid \exists \{x_k\} \subset S \text{ and } \{\tau_k\} \subset \mathbb{R} \text{ such that } x_k \rightarrow x, \tau_k \rightarrow 0, \text{ and } (x_k - x)/\tau_k \rightarrow w\}$ is the tangent cone to S at x . To be specific, the normal map L_K is defined as $L_K(z) = L \cdot \Pi_K(z) + z - \Pi_K(z)$ for any $z \in \mathbb{R}^{2p+1}$. The tangent cone $T_S(x)$ contains all the directions along which x can be approached by a sequence of points in S converging to x . Lemma 2 below shows that L_K is a global homeomorphism from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$ (a continuous bijective function from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$ whose inverse function is also continuous). In the proof of Lemma 2, we will give the explicit expression of the critical cone K .

Lemma 2.—*Suppose that Assumptions 1, 2(a) and 3 hold. Then the normal map L_K is a global homeomorphism from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$, and there is a neighborhood of $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ in which it is the unique local solution to (9).*

Combining Lemma 1 and 2, we can conclude that the assumptions in Lemma 2 guarantee $(\tilde{\beta}_0, \tilde{\beta})$ to be the unique local solution to (1) in a neighborhood of it.

2.3 Transformations of the SAA problem

We follow the same steps in Subsection 2.2 to formulate the SAA problem (2) as a normal map equation. First, by introducing the variable $t \in \mathbb{R}^p$, we transform the SAA problem (2) to the following equivalent problem:

$$\min_{(\beta_0, \beta, t) \in S} \frac{1}{N} \sum_{i=1}^N \left[y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right]^2 + \sum_{i=1}^p P_{\lambda_i}(t_i) + m(\|t\|_2^2 - \|\beta\|_2^2). \tag{22}$$

Second, we rewrite (22) as a variational inequality

$$0 \in f_N(\beta_0, \beta, t) + N_S(\beta_0, \beta, t), \tag{23}$$

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 $i = 1, \dots, p$, then the Jacobian matrix of f_N is given by

$$L_N(t) = df_N(\beta_0, \beta, t) = \begin{bmatrix} 2 & 2 \sum_{i=1}^T \mathbf{x}_i^T / N & 0 \\ 2 \sum_{i=1}^N \mathbf{x}_i / N & 2 \sum_{i=1}^T \mathbf{x}_i \mathbf{x}_i^T / N - 2mI_p & 0 \\ 0 & 0 & \text{diag}(P''_{\lambda_i}(t_i) + 2m)_{i=1}^p \end{bmatrix}. \quad (24)$$

Third, denoting the normal map induced by f_N and S by $(f_N)_S$, we obtain the normal map formulation of (23) as

$$(f_N)_S(z) = 0. \quad (25)$$

Let $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ be a local solution to (22). Then, $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ is also a solution to (23). So the point $z_N \in \mathbb{R}^{2p+1}$ defined as

$$z_N = (\hat{\beta}_0, \hat{\beta}, \hat{t}) - f_N(\hat{\beta}_0, \hat{\beta}, \hat{t}) \quad (26)$$

is a solution to (25) and satisfies $\Pi_S(z_N) = (\hat{\beta}_0, \hat{\beta}, \hat{t})$.

In fact, under Assumptions 1, 2 and 3, z_N is a locally unique solution to (25) when N is large enough and it converges to a solution z_0 to (18). This result will be shown in Subsection 3.1. Correspondingly, $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ is a locally unique solution to (22) and converges to a local solution $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ to (9). 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 upper left $(p+1) \times (p+1)$ submatrix of Σ_N , then we have $\Sigma_N = \begin{bmatrix} \Sigma_N^1 & 0 \\ 0 & 0 \end{bmatrix}$. Lu et al. (2017) (Lemma 3) shows that Σ_N converges to Σ_0 almost surely

as $N \rightarrow \infty$ for t LASSO penalty. We can similarly prove the same convergence result with a general penalty in this paper under Assumptions 1–4. Assumption 4 is shown as follows.

Assumption 4.

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

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

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 Lipschitz continuous for each $i = 1, \dots, p$. 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 k is finite valued in a neighborhood of zero.

(b) The same conditions as in (a) for $d_1 F(\beta_0, \beta, t, X, Y)$ instead of $F(\beta_0, \beta, t, X, Y)$. Accordingly, use $E[d_1 F(\beta_0, \beta, t, X, Y)]$ 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 $\beta_0(\beta_0, \beta, t)$ in the conditions.

Assumption 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 1(b) holds. Assumption 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.—Suppose that Assumptions 1, 2 and 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 sufficiently large N :

$$\text{Prob} \left\{ \sup_{(\beta_0, \beta, t) \in \mathcal{C}} \|f_N(\beta_0, \beta, t) - f_0(\beta_0, \beta, t)\| \geq \epsilon \right\} \leq \delta_1 \exp\{-N\mu_1\} + \frac{M_1}{\epsilon^{2p+1}} \exp\left\{-\frac{N\epsilon^2}{\sigma_1}\right\}. \quad (27)$$

Parts (b) and (c) of Assumption 4 impose the same type of assumptions on different random variables. Assumption 4(a-b) is needed to construct a reliable estimate for an unknown quantity in the asymptotic distribution in Theorem 1. Assumption 4(c) is only needed when the matrix Σ_0^1 is singular.

3 Construction of confidence intervals using stochastic variational inequality techniques

In this section, we show the proposed method and some related theoretical results to construct confidence intervals using stochastic variational inequality techniques. We first

develop the limiting distribution of SAA solutions in Section 3.1. Then, in Section 3.2, we show how to estimate the unknown quantities in the limiting distribution. The construction of the confidence intervals for the population penalized parameters and the true model parameters in the underlying linear model will be studied in Section 3.3 and Section 3.4, respectively. To present our proposed inference method clearly, we outline the procedures to construct confidence intervals for the true model parameters in Table 1. The extension to the high dimensional case is provided in Section 3.5.

3.1 Convergence and distribution of SAA solutions

Based on Lemma 2 and the relation between (9) and (18), Assumptions 1–3 guarantee z_0 defined in (19) to be a locally unique solution to (18). Furthermore, we show in Theorem 1 below that for sufficiently large N , (25) has a unique solution z_N in a neighborhood of z_0 , and that z_N converges almost surely to z_0 . This theorem also provides results on asymptotic distributions and convergence rates.

Theorem 1.—*Suppose that Assumptions 1, 2 and 3 hold. Then, with probability 1, there exist neighborhoods $\mathcal{X}$ of z_0 and $\mathcal{C}_0$ of $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, such that for sufficiently large N , the equation (25) has a unique solution z_N in $\mathcal{X}$, and the variational inequality (23) has a unique solution in $\mathcal{C}_0$ given by $(\hat{\beta}_0, \hat{\beta}, \hat{t}) = \Pi_S(z_N)$. Moreover,*

$$\lim_{N \rightarrow \infty} z_N = z_0 \text{ a.s.}, \quad \lim_{N \rightarrow \infty} (\hat{\beta}_0, \hat{\beta}, \hat{t}) = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \text{ a.s.}, \quad (28)$$

$$\sqrt{N}(z_N - z_0) \Rightarrow (L_K)^{-1}(\mathcal{N}(0, \Sigma_0)), \quad (29)$$

and

$$\sqrt{N}L_K(z_N - z_0) \Rightarrow \mathcal{N}(0, \Sigma_0). \quad (30)$$

In addition, if Assumption 4(a-b) holds, then there exist positive real numbers $\epsilon_0, \delta_0, \mu_0, M_0$ and σ_0 , such that for each $\epsilon \in (0, \epsilon_0]$ and each sufficiently large N ,

$$\begin{aligned} & \text{Prob}\left\{\|(\hat{\beta}_0, \hat{\beta}, \hat{t}) - (\tilde{\beta}_0, \tilde{\beta}, \tilde{t})\| < \epsilon\right\} \geq \text{Prob}\{\|z_N - z_0\| < \epsilon\} \\ & \geq 1 - \delta_0 \exp\{-N\mu_0\} - \frac{M_0}{\epsilon^{2p+1}} \exp\left\{-\frac{N\epsilon^2}{\sigma_0}\right\}. \end{aligned} \quad (31)$$

In Theorem 1, L_K is the normal map induced by the linear function $L(\tilde{t})$ in (16) and the critical cone K defined in (21). We use L_K^{-1} to denote its inverse function. Functions Π_K, L_K and L_K^{-1} are linear if K is a subspace, otherwise they are piecewise linear. Compared to (Lu et al., 2017, Theorem 1) which considers the LASSO penalty, Theorem 1 here handles general penalties that satisfy Assumption 1. The results shown in Theorem 1 are used in the construction of confidence intervals of the population penalized parameter as well as the true model parameter as shown in the following sections.

3.2 Estimators of Σ_0 and L_K

In order to use (29) and (30) to obtain computable confidence regions and intervals, we need to find reliable estimators of Σ_0 and L_K , as we discuss in this subsection. One can show that Σ_N converges to Σ_0 almost surely under Assumptions 1–4. See the remarks below (26). Therefore, we use Σ_N to estimate Σ_0 . Our main task in this subsection is to introduce an estimator of the normal map L_K , knowing that L_K is exactly $d(f_0)_S(z_0)$ (Robinson, 1995), the B-derivative of $(f_0)_S$ at z_0 . Let $d\Pi_S(z)$ be the B-derivative of the Euclidean projector Π_S at z . Since S is a polyhedral convex set in $\mathbb{R}^{2p+1}$, Π_S coincides with a different affine function on each $(2p+1)$ -cell in the normal manifold of S (see Table 1 in the supplementary materials for definitions of the normal manifold and cells). The B-derivative $d\Pi_S(z)$ is a linear function for points z in the interior of each such cell, and is piecewise linear for z on the boundary. Moreover, $d\Pi_S(z)$ is not continuous with respect to z at points z on the boundary of any $(2p+1)$ -cell. Therefore, the function $d(f_0)_S(z)$ is generally not continuous with respect to z at such points, which can be seen from the chain rule of B-differentiability:

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

If $d(f_0)_S(\cdot)$ is not continuous at z_0 , then $d(f_0)_S(z_N)$ is not guaranteed to converge to $d(f_0)_S(z_0)$ even though z_N converges to z_0 . To introduce the estimators of L_K , we will consider two cases based on the location of z_0 .

For each $i = 1, \dots, p$, denote the 9 cells in the normal manifold of S_i as $C_i^j, j = 0, 1, \dots, 8$ (see Figure 1). According to (10) we derive the constraints defining each C_i^j which are listed in Table 1 in the supplementary materials. That table also lists the critical cones K_i^j to S_i associated with a point in the relative interior of C_i^j . Each $(2p+1)$ -cell in the normal manifold of S can then be written as $\mathbb{R} \times \Pi_{i=1}^p C_i^{\gamma(i)}$, where $\gamma(i) = 0, \dots, 8$ for each $i = 1, \dots, p$. From (19), Assumption 1(a) and Lemma 2, we notice that $((z_0)_{2i}, (z_0)_{2i+1})$ can only appear in the relative interior of $C_i^3, C_i^4, C_i^6, C_i^7$ or C_i^8 for each i . Consequently, $d\Pi_S(z)$ is not continuous at z_0 if and only if $((z_0)_{2i}, (z_0)_{2i+1})$ is in the relative interior of C_i^3 or C_i^4 for some index i . The two cases are defined below, where the first case corresponds to the situation in which the random variable $(L_K)^{-1}(\mathcal{N}(0, \Sigma_0))$ is normally distributed, and the second case is for situations in which L_K is a piecewise linear function.

- Case I: In this case, $((z_0)_{2i}, (z_0)_{2i+1})$ is in the relative interior of C_i^6, C_i^7 or C_i^8 for all $i \in \{1 \dots p\}$, and the normal map L_K and the B-derivative $d\Pi_S(z_0)$ are linear functions. Since $d(f_0)_S(z)$ is continuous at z_0 in this case, 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, $((z_0)_{2i}, (z_0)_{2i+1})$ is in the relative interior of C_i^3 or C_i^4 for some index $i \in \{1 \dots p\}$, and L_K and $d\Pi_S(z_0)$ are piecewise linear functions. Since

$d(f_0)_S(z)$ is generally not continuous at z_0 in this case, we have to derive an estimator of L_K other than $d(f_N)_S(z_N)$.

In both cases, $d(f_N)_S(z_N)$ is an invertible linear map with high probability (Lu, 2014b, Proposition 3.5). While it is reasonable to expect Case I to occur more often than Case II in practice, one cannot identify Case I in advance since z_0 is unknown. To derive an estimator of L_K , 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 (12), we have

$$d\Pi_S(z)(h) = (\check{\beta}_0, d\Pi_{S_1}(\beta_1, t_1)(\check{\beta}_1, \check{t}_1), \dots, d\Pi_{S_p}(\beta_p, t_p)(\check{\beta}_p, \check{t}_p)), \quad (32)$$

for each $z = (\beta_0, \beta, t)$ and $h = (\check{\beta}_0, \check{\beta}, \check{t})$. We denote $d\Pi_{S_i}(\beta_i, t_i)$ in the relative interior of each C_i^j by a function $\psi_j: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ ($j = 0, 1, \dots, 8$). 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 in the supplementary materials shows the expression of each ψ_j using these matrices.

Denote the relative interior of $\mathbb{R} \times \Pi_i^p = {}_1C_i^{\gamma(i)}$ by $\text{ri}(\mathbb{R} \times \Pi_i^p = {}_1C_i^{\gamma(i)})$. For all

$z \in \text{ri}(\mathbb{R} \times \Pi_i^p = {}_1C_i^{\gamma(i)})$, we can write $d\Pi_S(z)$ as

$$\Psi_{\gamma(z)}(h) = (\check{\beta}_0, \psi_{\gamma(1)}(\check{\beta}_1, \check{t}_1), \dots, \psi_{\gamma(p)}(\check{\beta}_p, \check{t}_p)) \text{ for each } h = (\check{\beta}_0, \check{\beta}, \check{t}), \quad (33)$$

where $\gamma(z) = (\gamma(1), \dots, \gamma(p))$ such that $z \in \text{ri}(\mathbb{R} \times \Pi_i^p = {}_1C_i^{\gamma(i)})$.

Next, we construct an estimator of $d\Pi_S(z_0)$. We divide the plane (β, t) into 9 pieces $E_i^0, \dots, E_i^8$ (see Figure 1). The constraints that define each of these sets $E_i^0, \dots, E_i^8$ are listed in Table 3 in the supplementary materials. The function $g(N)$ in that table can be any combination of finite many terms of the form aN^b with $a > 0$ and $b \in (0, 1/2)$, among other choices. For more details, see Lu and Budhiraja (2013). Each partition $\mathbb{R} \times \Pi_i^p = {}_1E_i^{\gamma(i)}$ is related to the $(2p+1)$ -cell $\mathbb{R} \times \Pi_i^p = {}_1C_i^{\gamma(i)}$. Let

$$\hat{\gamma}(z) = (\gamma(1), \dots, \gamma(p)) \text{ such that } z \in \mathbb{R} \times \Pi_i^p = {}_1E_i^{\gamma(i)}.$$

Given a sample size N and a fixed z , we define a function $\Lambda_N(z): \mathbb{R}^{2p+1} \rightarrow \mathbb{R}^{2p+1}$ as

$$\Lambda_N(z)(h) = \Psi_{\hat{\gamma}(z)}(h), \text{ for each } h \in \mathbb{R}^{2p+1}. \quad (34)$$

According to Theorem 3.1 of Lu (2014a), $\Lambda_N(z_N)$ converges to $d\Pi_S(z_0)$ in probability under Assumptions 1–4.

Based on (24), (26) and (34), we define a function $\Phi_N(z_N): \mathbb{R}^{2p+1} \rightarrow \mathbb{R}^{2p+1}$ as

$$\Phi_N(z_N)(h) = L_N(\hat{t})\Lambda_N(z_N)(h) + h - \Lambda_N(z_N)(h) \quad (35)$$

for each $h \in \mathbb{R}^{2p+1}$. The following theorem shows that $d(f_N)_{S(z_N)}$ is a consistent estimator of L_K for Case I, and $\Phi_N(z_N)$ is a consistent estimator of L_K for both Case I and Case II.

Theorem 2.

- (a) Suppose that Assumptions 1, 2 and 3 hold. If z_0 satisfies the conditions for Case I, then $d\Pi_S(z_N)$ defined in (33) converges to $d\Pi_S(z_0)$ almost surely, and

$$d(f_N)_{S(z_N)} = L_N(\hat{t}) d\Pi_S(z_N) + I - d\Pi_S(z_N) \quad (36)$$

converges to L_K almost surely.

- (b) Suppose that Assumptions 1, 2, 3 and 4(a-b) hold. Then $\Phi_N(z_N)$ converges to L_K in probability.

The two functions $d(f_N)_{S(z_N)}$ and $\Phi_N(z_N)$ are generally different when $((z_N)_{2i}, (z_N)_{2i+1})$ belongs to $E_i^0, E_i^1, E_i^2, E_i^3$ or E_i^4 for some i , in which case $\Phi_N(z_N)$ is a piecewise linear function. In contrast, $d(f_N)_{S(z_N)}$ is a piecewise linear function only when $((z_N)_{2i}, (z_N)_{2i+1})$ belongs to C_i^3 or C_i^4 for some i .

Under Assumptions 1–4, we can show that the weak convergence in (30) still holds after L_K is substituted by $\Phi_N(z_N)$. Consequently, if Σ_0^1 is nonsingular, then we have

$$\sqrt{N} \begin{bmatrix} (\Sigma_N^1)^{-1/2} & 0 \\ 0 & I_p \end{bmatrix} (\Phi_N(z_N))(z_N - z_0) \Rightarrow \mathcal{N}(0, I_{p+1}) \times 0. \quad (37)$$

If Σ_0^1 is singular, we 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 l be the number of positive eigenvalues of Σ_0^1 counted with regard to their algebraic multiplicities, 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 . Furthermore, let $(U_N)_1$ be the submatrix of U_N that consists of its first l_N rows, and submatrix $(U_N)_2$ consists of the remaining rows of U_N . We present the weak convergence results in the following theorem, which generalizes Theorem 3 in Lu et al. (2017) to cover all penalties satisfying Assumption 1.

Theorem 3.—Suppose that Assumptions 1, 2, 3 and 4(a-b) hold. Then

$$\sqrt{N} \Phi_N(z_N)(z_N - z_0) \Rightarrow \mathcal{N}(0, \Sigma_0). \quad (38)$$

If Σ_0^1 is nonsingular, then

$$N[(\Phi_N(z_N))(z_N - z_0)]^T \begin{bmatrix} (\Sigma_N^1)^{-1} & 0 \\ 0 & I_p \end{bmatrix} [(\Phi_N(z_N))(z_N - z_0)] \Rightarrow \chi_{p+1}^2, \quad (39)$$

and

$$N[(\Phi_N(z_N))(z_N - z_0)]^T \begin{bmatrix} 0 & 0 \\ 0 & I_p \end{bmatrix} [(\Phi_N(z_N))(z_N - z_0)] \Rightarrow 0. \quad (40)$$

If Σ_0^1 is singular and Assumption 4(c) holds, then $\text{Prob}\{I_N = I\} \rightarrow 1$ as $N \rightarrow \infty$,

$$N[(\Phi_N(z_N))(z_N - z_0)]^T \begin{bmatrix} (U_N)_1^T D_N^{-1} (U_N)_1 & 0 \\ 0 & 0 \end{bmatrix} [(\Phi_N(z_N))(z_N - z_0)] \Rightarrow \chi_I^2, \quad (41)$$

and

$$N[(\Phi_N(z_N))(z_N - z_0)]^T \begin{bmatrix} 0 & 0 \\ 0 & I_p \end{bmatrix} [(\Phi_N(z_N))(z_N - z_0)] \Rightarrow 0. \quad (42)$$

We can treat (39) and (40) as a special case of (41) and (42). In fact, if z_0 satisfies Case I, then Theorem 3 still holds if $\Phi_N(z_N)$ is replaced by $d(f_N)_{S(z_n)}$.

3.3 Confidence intervals for the population penalized parameters

In this subsection, we describe how to obtain confidence interval for $(\tilde{\beta}_0, \tilde{\beta})$ from the asymptotic distribution of z_N . First, we investigate the relationship between a solution to the normal map formulation (18) and the corresponding solution to (1). Let $\tilde{q}$ be as defined in Assumption 3 and $\tilde{q}_0 = E[-2(Y - \tilde{\beta}_0 - \sum_{i=1}^p \tilde{\beta}_i X_i)]$. From (13), (14) and (19), we have $z_0 = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) - (\tilde{q}_0, \tilde{q} - 2m\tilde{\beta}, (P'_{\lambda_i}(\tilde{t}_i) + 2m\tilde{t}_i)^p_1)$. In the supplementary materials, it is shown in (B.1) that $\tilde{q}_0 = 0$, which implies $\tilde{\beta}_0 = (z_0)_1$. Thus, confidence intervals of $\tilde{\beta}_0$ are exactly those of $(z_0)_1$. On the other hand, using the fact $(\tilde{\beta}_i, \tilde{t}_i) = \Pi_{S_i}((z_0)_{2i}, (z_0)_{2i+1})$ for each $i = 1, \dots, p$, we have the following relationship between $\tilde{\beta}_i$ and $((z_0)_{2i}, (z_0)_{2i+1})$:

$$\tilde{\beta}_i = \Gamma(V_+, V_-) = \begin{cases} \frac{1}{2}V_+, & \text{if } V_+ > 0 \text{ and } V_- \geq 0, \\ 0, & \text{if } V_+ \leq 0 \text{ and } V_- \geq 0, \\ \frac{1}{2}V_-, & \text{if } V_+ \leq 0 \text{ and } V_- < 0, \end{cases} \quad (43)$$

where $V_+ = (z_0)_{2i} + (z_0)_{2i+1}$ and $V_- = (z_0)_{2i} - (z_0)_{2i+1}$. The above three cases in (43) include all the possible situations for the location of $((z_0)_{2i}, (z_0)_{2i+1})$. This map Γ can be used to obtain confidence intervals for $\tilde{\beta}_i (i = 1, \dots, p)$ after we calculate confidence intervals for $((z_0)_{2i} + (z_0)_{2i+1})$ and $((z_0)_{2i} - (z_0)_{2i+1})$. For a fixed i , we denote the $(1 - \alpha/2)\%$ confidence

intervals for $((z_0)_{2i} + (z_0)_{2i+1})$ and $((z_0)_{2i} - (z_0)_{2i+1})$ as $[L_+^i, U_+^i]$ and $[L_-^i, U_-^i]$ respectively. Then a $(1 - \alpha)\%$ (conservative) confidence interval for $\tilde{\beta}_i$ is given by

$$[\Gamma(L_+^i, L_-^i), \Gamma(U_+^i, U_-^i)]. \quad (44)$$

Next, we show how to find confidence intervals for $(z_0)_1$, $((z_0)_{2i} + (z_0)_{2i+1})$ and $((z_0)_{2i} - (z_0)_{2i+1})$. Under Assumptions 1–4, from Theorem 3 we can express the asymptotically exact $(1 - \alpha)100\%$ confidence region for z_0 as

$$\left\{ \begin{array}{l} z \in \mathbb{R}^{2p+1} \\ N[\Phi_N(z_N)(z_N - z)]^T \begin{bmatrix} (U_N)_1^T D_N^{-1} (U_N)_1 & 0 \\ 0 & 0 \end{bmatrix} [\Phi_N(z_N)(z_N - z)] \leq \chi_{I_N}^2(\alpha) \\ N[\Phi_N(z_N)(z_N - z)]^T \begin{bmatrix} 0 & 0 \\ 0 & I_p \end{bmatrix} [\Phi_N(z_N)(z_N - z)] = 0 \end{array} \right\} \quad (45)$$

where $\chi_{I_N}^2(\alpha)$ is the critical value associated with significant level α of a χ^2 distribution with I_N degrees of freedom. If $\Phi_N(z_N)$ is a linear map, then the set in (45) 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 $(z_0)_1$, $((z_0)_{2i} + (z_0)_{2i+1})$ and $((z_0)_{2i} - (z_0)_{2i+1})$ in the set of (45) by solving optimization problems.

For individual confidence intervals, first we notice that $\Phi_N(z_N)$ is a global homeomorphism with probability 1 as $N \rightarrow \infty$ (see the proof of Theorem 2 in the supplementary materials). If $\Phi_N(z_N)$ is a global homeomorphism, we can use

$$(\Phi_N(z_N))^{-1}(\mathcal{N}(0, \Sigma_N)) \quad (46)$$

to approximate the distribution of $\sqrt{N}(z_N - z_0)$ as in (29). When $\Phi_N(z_N)$ is a linear map, the distribution in (46) is normal. Therefore $(z_0)_1$, $((z_0)_{2i} + (z_0)_{2i+1})$ and $((z_0)_{2i} - (z_0)_{2i+1})$ 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 (46), and find empirical individual confidence intervals for $(z_0)_1$, $((z_0)_{2i} + (z_0)_{2i+1})$ and $((z_0)_{2i} - (z_0)_{2i+1})$ by taking $\frac{\alpha}{2}\%$ and $(1 - \frac{\alpha}{2})\%$ percentiles of the data as the lower and upper bounds respectively.

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

In this subsection, we develop a method to compute confidence intervals for $(\beta_0^{true}, \beta^{true})$, based on a relation between a population penalized parameter $(\tilde{\beta}_0, \tilde{\beta})$ and the true model parameter $(\beta_0^{true}, \beta^{true})$. Suppose the underlying linear model is

$$Y = \beta_0^{true} + X^T \beta^{true} + \varepsilon, \quad (47)$$

where $\beta_0^{true} \in \mathbb{R}$ and $\beta^{true} = (\beta_1^{true}, \dots, \beta_p^{true}) \in \mathbb{R}^p$ are the true model parameters. Let $t_i^{true} \in \mathbb{R}^p$ be defined as $t_i^{true} = |\beta_i^{true}|$. Denote the covariance matrix of X as Σ , and assume the random error ε has mean zero and variance σ^2 . Moreover, ε is independent with X_i for all $i = 1, \dots, p$. For simplicity we assume $E(X_i) = 0$ for each $i = 1, \dots, p$. Consequently we have $E(Y) = \beta_0^{true}$ and $\Sigma = E(XX^T)$. We assume that Σ is nonsingular and therefore we do not need Assumption 3 in this subsection.

In developing theoretical results of this subsection, we will let $\lambda = (\lambda_1, \dots, \lambda_p)$ converge to 0. Due to this change, assumptions stated in Section 2.2 need to be changed accordingly. We will replace Assumptions 1 by Assumption 1', and keep Assumption 2. We will not need Assumption 4 until the end of this section.

Assumption 1'(a).—For each $i = 1, 2, \dots, p$, $P_0(t) = 0$ for all $t \geq 0$. Moreover, for each positive λ_i in a neighborhood of 0, $P_{\lambda_i}(\cdot)$ is nonnegative, nondecreasing and continuously differentiable on $[0, +\infty)$ with $P_{\lambda_i}(0) = 0$ and $P'_{\lambda_i}(0) > 0$.

Assumption 1'(b).—For each $i = 1, \dots, p$, there exist neighborhoods $\mathcal{T}_i$ of t_i^{true} in $\mathbb{R}_+$ and Λ_i of 0 in $\mathbb{R}_+$, such that $P''_{\lambda_i}(t_i)$ the second derivative of $(\cdot)$ with respect to t_i , exists for each $\lambda_i \in \Lambda_i$ and $t_i \in \mathcal{T}_i$. Moreover, $P'_{\lambda_i}(t_i)$ and $P''_{\lambda_i}(t_i)$ are Lipschitz continuous in (λ_i, t_i) on $\mathfrak{Z}_i \times \Lambda_i$.

Assumption 1'(c).—For each $i = 1, \dots, p$, there exists a neighborhood $\mathcal{T}_i$ of t_i^{true} in $\mathbb{R}_+$, such that the mixed partial derivatives $\frac{\partial^2 P}{\partial \lambda_i \partial t_i}(0, t_i)$ and $\frac{\partial^3 P}{\partial \lambda_i \partial^2 t_i}(0, t_i)$ exist for each $t_i \in \mathcal{T}_i$, with

$$\lim_{\lambda_i \rightarrow 0} \sup_{t_i \in \mathcal{T}_i} \left| \frac{P'_{\lambda_i}(t_i)}{\lambda_i} - \frac{\partial^2 P}{\partial \lambda_i \partial t_i}(0, t_i) \right| + \left| \frac{P''_{\lambda_i}(t_i)}{\lambda_i} - \frac{\partial^3 P}{\partial \lambda_i \partial^2 t_i}(0, t_i) \right| = 0.$$

Besides the convex LASSO penalty, we can check that many non-convex penalty functions such as SCAD, MCP, the log-penalty, and the transformed ℓ_1 penalty satisfy Assumption 1' (a). We can further check that the LASSO penalty, the log-penalty, and the transformed ℓ_1 penalty also satisfy Assumptions 1'(b) and 1'(c). For the SCAD and MCP penalties, Assumptions 1'(b) and 1'(c) are satisfied almost everywhere except that $t_i^{true} = 0$ for some i .

Assumptions 1'(b) and 1'(c) are used to guarantee that the SAA function f_N almost surely converges to the true function f_0 in the space of continuously differentiable functions on a neighborhood of $(\beta_0^{true}, \beta^{true}, t^{true})$, and that $\sqrt{N}(f_N - f_0)$ weakly converges to a random function in that space. This assumption is needed for the techniques based on stochastic variational inequalities to be applicable. It is possible to weaken this assumption by developing techniques for a broader class of problems in which the SAA function f_N (or equivalently, the first order derivative of the penalty function) is not necessarily continuously differentiable, and we will investigate this in future work.

Under Assumption 1'(a), with $\lambda = 0$ the problem (1) becomes the least square problem (3), which has a unique solution $(\beta_0^{true}, \beta^{true})$ in view of the linear model (47). Let $t^{true} \in \mathbb{R}^p$ be defined as if $t_i^{true} = |\beta_i^{true}|$. Then (9) with $\lambda = 0$ has a unique solution $(\beta_0^{true}, \beta^{true}, t^{true})$. By the equivalence between (9) and (17), $(\beta_0^{true}, \beta^{true}, t^{true})$ is also the unique solution to

$$-f_0(\beta_0, \beta, t) \in N_S(\beta_0, \beta, t)$$

where $f_0(\beta_0, \beta, t)$ is as defined in (14) but with $\lambda = 0$:

$$f_0(\beta_0, \beta, t) = \begin{bmatrix} -2E[Y - \beta_0 - \sum_{i=1}^p \beta_i X_i] \\ -2E[(Y - \beta_0 - \sum_{i=1}^p \beta_i X_i)X] - 2m\beta \\ 2mt \end{bmatrix}.$$

Let z_0^* be as defined in (19) with $\lambda = 0$ and $(\beta_0^{true}, \beta^{true}, t^{true})$ replacing $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$. The following lemma presents the relation between $(\beta_0^{true}, \beta^{true})$ and z_0^* .

Lemma 4.—*Suppose that Assumptions 1'(a) and 2 hold. Then we have*

$$(\beta_0^{true}, \beta^{true}) = G^*(z_0^*) \quad (48)$$

where $G^* \in \mathbb{R}^{(p+1) \times (2p+1)}$ is defined as

$$G^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & (1 + 2m)^{-1} I_p & 0 \end{bmatrix}. \quad (49)$$

Lemma 4 above indicates that an estimator of the true parameter $(\beta_0^{true}, \beta^{true})$ is $G^*(z_N)$ where z_N is defined in (26). In the following theorem, we show the asymptotic distribution of the estimator $G^*(z_N)$. Before stating this theorem, we define a matrix

$L^* \in \mathbb{R}^{(2p+1) \times (2p+1)}$ and a cone $K^* \subset \mathbb{R}^{2p+1}$ as follows:

$$L^* = df_0(\beta_0^{true}, \beta^{true}, t^{true}) = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2\Sigma - 2mI_p & 0 \\ 0 & 0 & 2mI_p \end{bmatrix}, \quad (50)$$

and

$$K^* = \{w \in T_{S(\beta_0^{true}, \beta^{true}, t^{true})} | \langle f_0(\beta_0^{true}, \beta^{true}, t^{true}), w \rangle = 0 \}. \quad (51)$$

Note that

$$f_0(\beta_0^{true}, \beta^{true}, t^{true}) = \begin{bmatrix} 0 \\ -2m\beta^{true} \\ 2mt^{true} \end{bmatrix}.$$

This implies that

$$K^* = \mathbb{R} \times \Pi_{i=1}^p K_i^* \quad (52)$$

where

$$K_i^* = \begin{cases} \{(\beta_i, t_i) | \beta_i = t_i\} & \text{if } \beta_i^{true} > 0, \\ \{(\beta_i, t_i) | \beta_i = -t_i\} & \text{if } \beta_i^{true} < 0, \\ S_i & \text{if } \beta_i^{true} = 0. \end{cases} \quad (53)$$

As the setting considered in this subsection is different from previous sections, L^* and K^* here are different from L and K defined in Assumption 3 and (21). The previous L and K are associated with $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, a solution to the population problem with a fixed positive λ , while L^* and K^* are associated with $(\beta_0^{true}, \beta^{true}, t^{true})$ and $\lambda = 0$.

Theorem 4.—Suppose that Assumptions 1 (a-c) and 2 hold. Let $m > 0$ be sufficiently small so that $\Sigma - mI_p$ is nonsingular, L^* and K^* be defined as above, Σ_0^* be the covariance matrix of the random vector $F(\beta_0^{true}, \beta^{true}, t^{true}, X, Y)$ defined in (13), and Σ_0^{*1} be the upper left $(p+1) \times (p+1)$ submatrix of Σ_0^* . Moreover, let λ_i 's be chosen to satisfy $\lim_{N \rightarrow \infty} \sqrt{N} \lambda_i = c_i$ for some constant $c_i \geq 0$, z_N be defined in (26), and define

$(\hat{\beta}_0^{true}, \hat{\beta}^{true}) = G^*(z_N)$ and $h \in \mathbb{R}^p$ by $h_i = c_i \frac{\partial^2 P}{\partial \lambda_i \partial t_i}(0, t_i^{true})$. Then $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$ is a consistent estimator of $(\beta_0^{true}, \beta^{true})$ and

$$\sqrt{N} \left((\hat{\beta}_0^{true}, \hat{\beta}^{true}) - (\beta_0^{true}, \beta^{true}) \right) \Rightarrow G^* \circ (L_{K^*}^*)^{-1}(\mathcal{N}(0, \Sigma_0^{*1}), h). \quad (54)$$

Note that the distribution of $G^{*o}(L_{K^*}^*)^{-1}(\mathcal{N}(0, \Sigma_0^{*1}), h)$ can be normal or non-normal. When the true parameter $\beta_i^{true} \neq 0$ for each i , we can show that K^* is a subspace of $\mathbb{R}^{(2p+1)}$ and $(L_{K^*}^*)^{-1}$ is a linear function. Therefore, the limiting distribution of the true parameter estimator $G^*(z_N)$ is normal in this case. However, if the true parameter $\beta_i^{true} = 0$ for some i , the limiting distribution can be normal or non-normal.

Theorem 5.—*Suppose that the assumptions in Theorem 4 hold. If $h_i = 0$ for each i , then $(\hat{\beta}_0^{true}, \hat{\beta}^{true})$ is a consistent estimator of $(\beta_0^{true}, \beta^{true})$ and*

$$G^* \circ (L_{K^*}^*)^{-1}(\mathcal{N}(0, \Sigma_0^{*1}), h) = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2}\Sigma^{-1} \end{bmatrix}(\mathcal{N}(0, \Sigma_0^{*1})) = \mathcal{N}(0, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2\Sigma^{-1} \end{bmatrix}).$$

Furthermore, let $\hat{\Theta}$ be a consistent estimate of Σ^{-1} . Define $H_N = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2}\hat{\Theta} \end{bmatrix}$. Then,

$$\frac{\sqrt{N}(\hat{\beta}_i^{true} - \beta_i^{true})}{\sqrt{(H_N \Sigma_N^1 H_N^T)_{i+1, i+1}}} \Rightarrow \mathcal{N}(0, 1), \text{ for all } i = 0, 1, \dots, p. \quad (55)$$

Since $h_i = c_i \frac{\partial^2 P}{\partial \lambda_i \partial t_i}(0, t_i^{true})$, we know that $h_i = 0$ if $c_i = 0$. Therefore, if λ_i 's are chosen to be $o(1/\sqrt{N})$ in the penalty function, the limiting distribution will be a multivariate normal distribution. In this normal case, the above Theorem 5 provides a method to compute asymptotically exact individual confidence intervals for $(\beta_0^{true}, \beta^{true})$.

When $h_i \neq 0$ for some i , the asymptotic distribution in Theorem 4 does not necessarily reduce to a normal distribution. To see this, consider the following example. Let $p = 2$, $\beta_0^{true} = 0$, $\beta_1^{true} = 0$ and $\beta_2^{true} = a_0$ for some constant

$$a_0 > 0. \text{ Let } \lambda = \frac{1}{\sqrt{N}}, \text{ so } c_1 = c_2 = 1. \text{ Let } \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, m = 1, \text{ and } \frac{\partial^2 P}{\partial \lambda_1 \partial t_1}(0, t_1^{true}) = \frac{\partial^2 P}{\partial \lambda_2 \partial t_2}(0, t_2^{true}) = 1$$

(which is satisfied by the LASSO penalty function $P(\lambda, t) = \lambda t$). It follows that $h_1 = h_2 = 1$. Let $q_0, q_1, q_2 \in \mathbb{R}$. To find $(L_{K^*}^*)^{-1}(q_0, q_1, q_2, h_1, h_2)$ we consider the following problem

$$\min_{(\beta_0, \beta_1, t_1, \beta_2, t_2) \in K^*} \beta_0^2 + \beta_1^2 + t_1^2 + \beta_2^2 + t_2^2 - q_0\beta_0 - q_1\beta_1 - h_1t_1 - q_2\beta_2 - h_2t_2,$$

whose solution satisfies $(q_0, q_1, q_2, h_1, h_2) \in L^*(\beta_0, \beta_1, \beta_2, t_1, t_2) + N_{K^*}(\beta_0, \beta_1, \beta_2, t_1, t_2)$. Here $K^* = \mathbb{R} \times S_1 \times \{(\beta_2, t_2) | \beta_2 = t_2\}$. The solution to the above problem is given by

$$\beta_0 = \frac{q_0}{2}, \quad \beta_1 = \begin{cases} \frac{q_1}{2} & \text{if } -1 \leq q_1 \leq 1 \\ \frac{q_1 + 1}{4} & \text{if } q_1 \geq 1 \\ \frac{q_1 - 1}{4} & \text{if } q_1 \leq -1 \end{cases}, \quad t_1 = |\beta_1|, \quad \beta_2 = t_2 = \frac{q_2 + 1}{4}.$$

As a result, $(L_{K^*}^*)^{-1}(q_0, q_1, q_2, h_1, h_2) = (\beta_0, \beta_1, \beta_2, t_1, t_2)$ is a piecewise affine function of $(q_0, q_1, q_2, h_1, h_2)$ with three pieces. Furthermore, since $G^*(\cdot)$ is a linear transformation, we conclude that the asymptotic distribution is non-normal in this case.

Next, we show how to estimate $(L_{K^*}^*)^{-1}$ in the situation considered in Theorem 4. We need to make an assumption analogous to Assumption 4.

Assumption 4'(a).—*The same conditions as in Assumption 4(a), with $(\beta_0^{true}, \beta^{true}, t^{true})$ replacing $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$.*

Assumption 4'(b).—*The same conditions as in Assumption 4(b), with $(\beta_0^{true}, \beta^{true}, t^{true})$ replacing $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$.*

To estimate $(L_{K^*}^*)^{-1}$, we can also use $\Phi_N(z_N)$. Under Assumption 4', to show that $\Phi_N(z_N)$ is a consistent estimator of $(L_{K^*}^*)^{-1}$, the key is to show (31) holds (with $(\beta_0^{true}, \beta^{true}, t^{true})$ replacing $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$). To show this, one needs to show that f_N converges to f_0 in probability at an exponential rate in the space of continuous differentiable functions. For the first $p+1$ components, this follows from Assumption 4'. For the last p components, note that the norm of the function P'_{λ_i}/λ_i in the space of continuous differentiable functions is bounded due to Assumption 1'(c). In addition, since $\lim_{N \rightarrow \infty} \sqrt{N}\lambda_i = c_i$ for each i , we have $\sqrt{N}\|P'_{\lambda_i}\| \leq \rho$ for some constant ρ . As a result, for each $\epsilon > 0$, we have $\|P'_{\lambda_i}\| \leq \rho/\sqrt{N} < \epsilon/2$ for sufficiently large N . Therefore, f_N converges to f_0 in probability at an exponential rate in the space of continuous differentiable functions, and $\Phi_N(z_N)$ converges to $(L_{K^*}^*)^{-1}$ in probability.

For the normal case, equation (55) justifies using the diagonal elements of $H_N \Sigma_N H_N^T$, divided by N , as the estimated variances of $(\beta_0^{true}, \beta^{true})$.. For the non-normal case, first we define two functions R and $\hat{R}$ from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{p+1}$ as

$$R = G^* \circ (L_{K^*}^*)^{-1} \begin{bmatrix} (\Sigma_0^*)^{\frac{1}{2}} & 0 \\ 0 & \text{diag}(h_i)_{i=1}^p \end{bmatrix} \quad \text{and} \quad \hat{R} = G^* \circ (\Phi_N(z_N))^{-1} \begin{bmatrix} (\Sigma_N^*)^{\frac{1}{2}} & 0 \\ 0 & \text{diag}(\hat{h}_i)_{i=1}^p \end{bmatrix}, \quad (56)$$

where $\hat{h}_i = \sqrt{N} \lambda_i \frac{\partial^2 P}{\partial \lambda_i \partial t_i}(0, \hat{\beta}_i^{\text{true}})$. Denote the i^{th} component function of R and $\hat{R}$ as R_i and $\hat{R}_i$ respectively for each i . Let $f: \mathbb{R}^{2p+1} \rightarrow \mathbb{R}$ be a continuous function and Z be a $(2p+1)$ -dimensional random variable with $Z \sim \mathcal{N}(0, I_{p+1}) \times \vec{1}$. Define $a^r(f) \in (0, \infty)$ as

$$a^r(f) = \inf\{c \geq 0 \mid \text{Prob}\{-c \leq f(Z) - r \leq c\} \geq 1 - \alpha\}. \quad (57)$$

Suppose that $\text{Prob}\{f(Z) = b\} = 0$ for all $b \in \mathbb{R}$. Then for any given $r \in \mathbb{R}$ and $\alpha \in (0, 1)$, $a^r(f)$ as defined in (57) is the smallest value that satisfies

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

Since the map G^* has full row rank, $L_{K^*}^*$ is a global homeomorphism, and Σ_0^{*1} is nonsingular. If $h_i = 0$ for each i , then the matrix representation of each piece of the map R has full row rank as well. Therefore, $\text{Prob}\{R_i(Z) = b\} = 0$ for all $b \in \mathbb{R}$. The following theorem provides a way to compute individual confidence intervals for $(\beta_0^{\text{true}}, \beta^{\text{true}})$ in the general case where $h_i = 0$ for each i .

Theorem 6.—*Suppose that assumptions in Theorem 4 and Assumptions 4'(a-b) hold, and $h_i = 0$ for each i . Let $\alpha \in (0, 1)$ and $a^r(\cdot)$ be as in (57). Then for every $r \in \mathbb{R}$ and all $i = 0, 1, \dots, p$, we have*

$$\lim_{N \rightarrow \infty} \text{Prob}\{|\sqrt{N}(\hat{\beta}_i^{\text{true}} - \beta_i^{\text{true}}) - r| \leq a^r(\hat{R}_{i+1})\} = 1 - \alpha, \quad (58)$$

where R and $\hat{R}$ are defined in (56).

From (58), one can compute the empirical $(1 - \alpha)$ percentile confidence intervals for $(\beta_0^{\text{true}}, \beta^{\text{true}})$ by simulating data from $\hat{R}(Z)$. The constant r can be used to control the centers of the confidence intervals for all β_i^{true} simultaneously, which may affect the interval lengths. A reasonable choice of r is 0 if the empirical distribution of $\hat{R}(Z)$ is approximately symmetric with respect to 0. Results of Theorems 5 and 6 are applicable to a wide range of general penalty functions which covers the LASSO as a special case (Lu et al., 2017). The procedure to construct confidence intervals of the true regression coefficients is summarized in Table 1.

3.5 Extension to the high dimensional case

In our previous theoretical analysis, we assume that the dimension p is fixed. It is interesting to study the extension of our proposed method to the high dimensional case where the dimension p is also allowed to go to infinity.

As shown in Lemma 4, since $(\beta_0^{true}, \beta^{true}) = G^*(z_0^*)$, our proposed estimate for the true parameter $(\beta_0^{true}, \beta^{true})$ is $G^*(z_N)$. In fact, we can also show that $(\beta_0^{true}, \beta^{true}) = G(z_0)$, where z_0 is defined in (19) with $\lambda > 0$ and G is a map from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{p+1}$ defined as

$$G = \frac{1}{2} \left(\begin{bmatrix} 1 & 0 \\ 0 & \Sigma^{-1} \end{bmatrix} B + \begin{bmatrix} 1 & 0 \\ 0 & 2I - (1 + 2m)\Sigma^{-1} \end{bmatrix} B \circ \Pi_K \right), \quad (59)$$

and the matrix $B \in \mathbb{R}^{(p+1) \times (2p+1)}$ is given by $B = [I_{p+1} \ 0]$. Motivated by this result, we can also estimate the true parameter $(\beta_0^{true}, \beta^{true})$ by $\hat{G}(z_N)$, where $\hat{G}$ is a map from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{p+1}$ defined as

$$\hat{G} = \frac{1}{2} \left(\begin{bmatrix} 1 & 0 \\ 0 & \hat{\Theta} \end{bmatrix} B + \begin{bmatrix} 1 & 0 \\ 0 & 2I - (1 + 2m)\hat{\Theta} \end{bmatrix} B \circ d\Pi_S(z_N) \right), \quad (60)$$

and $\hat{\Theta}$ is a consistent estimate of Σ^{-1} . Theoretically, if $\lim_{N \rightarrow \infty} \sqrt{N} \lambda_i = 0$ for each i , we can show that $G^*(z_N)$ and $\hat{G}(z_N)$ have the same asymptotic distribution.

According to the definition of $\hat{G}$ in (60) and the definition of z_N in (26), we can show that $\hat{G}(z_N) = (\hat{\beta}_0 + \frac{1}{N} 1_N^T (\mathbf{y} - \hat{\beta}_0 1_N - \mathbf{X} \hat{\beta}), \hat{\beta} + \frac{1}{N} \hat{\Theta} \mathbf{X}^T (\mathbf{y} - \hat{\beta}_0 1_N - \mathbf{X} \hat{\beta}))$. Therefore, the estimate of β is the sum of an initial estimate $\hat{\beta}$ (e.g., LASSO, SCAD or MCP estimate) and a bias-correction term. Interestingly, although we have different motivations, $\hat{G}(z_N)$ turns out to be the same as the estimate proposed by Van de Geer et al. (2014). For the high dimensional case with $p \gg N$, if we choose $\lambda = O(\sqrt{\log(p)/N})$ converging to 0 as $N \rightarrow \infty$, and use conditions to guarantee that: (a) $\|\hat{\beta} - \beta^{true}\|_1 = O_p(s_0 \sqrt{\log(p)/N})$ where s_0 is the number of true nonzero regression coefficients; (b) $s_0 = o(\sqrt{N}/\log(p))$, and some sparsity assumptions about the precision matrix Σ^{-1} , we can show that the asymptotic distribution of $\hat{G}(z_N)$ is normal (Van de Geer et al. (2014)). However, the theoretical analysis of the asymptotic distribution of $G^*(z_N)$ and $\hat{G}(z_N)$ using stochastic variational inequality techniques for $p \gg N$ case is challenging. Many fundamental results about variational inequality (e.g., some results used in the proof of Theorem 1) need to be generalized to the high dimensional case.

Although we assume that the dimension p is fixed in the theoretical analysis using stochastic variational inequality techniques, our proposed method is applicable to the large p small N data in practice. For the high dimensional data, we can use the results shown in Theorem 6 to construct confidence intervals. In Section 4, we will use Example 3 to study the performance of our method for the high dimensional data.

4 Numerical examples

In this section, we use the MCP methods in (8) to illustrate the performance of the techniques proposed in Section 3. For all examples in this section, we choose $\frac{1}{g(N)} = \frac{0.001}{N^{1/3}}$ and $m = \frac{1}{2}$ in (9). We use the mixed integer quadratically constrained program (MIQCP) solver in the optimization modeling language GAMS (Brooke et al., 1998) to obtain accurate solutions to (2).

For all simulated examples, we generate the data using the following linear model:

$$Y = X^T \beta^{true} + \sigma \epsilon, \quad (61)$$

where $\beta^{true} \in \mathbb{R}^p$, X is a p -dimensional normal random variable with mean 0 and covariance Σ , ϵ is a standard normal random error which is independent of X . We set the noise level $\sigma = 1$. Under the model (61), the population penalized regression problem (1) can be written as

$$\min_{\beta_0, \beta} (\beta^{true} - \beta)^T \Sigma (\beta^{true} - \beta) + \beta_0^2 + \sum_{j=1}^p P_{\lambda_j}(|\beta_j|). \quad (62)$$

We compute confidence intervals for the population penalized parameter $(\tilde{\beta}_0, \tilde{\beta})$ and the true model parameter β^{true} , which we refer to as the first and second types of confidence intervals respectively. To show their performance in simulation study, we report the following two measures: the empirical coverage rate (the fraction of total replications in which the confidence intervals contain the corresponding population penalized parameters or true model parameters) and the average confidence interval length. For the second type of confidence interval, we compare our proposed method with the LDPE method (Van de Geer et al. (2014); Zhang and Zhang (2014)), the method introduced by Javanmard and Montanari (2014) (denoted as JM method), and the method proposed by Lu et al. (2017) (denoted as SVI-Lasso). In terms of the tuning parameter λ , we study the performance of our proposed method with some fixed values as well as the value of λ chosen by the Generalized Information Criterion (GIC, Konishi and Kitagawa (2008)).

4.1 Example 1: Low dimensional setting with the auto-regressive covariance structure

For this example, we generate a training dataset with 500 replications of sample size $N = 300$, dimension $p = 8$, true model parameter $\beta^{true} = (3, 1.5, 0, 0, 2, 0, 0, 0)$, and true covariance matrix $\Sigma_{ij} = 0.5^{|i-j|}$. We consider six MCP penalties with parameters (λ, a) taking the following values: $\lambda = 0.5, 1$ or 2 , and $a = 2$ or 2000 . When $a = 2000$, the MCP penalties are very close to the LASSO penalty. In each replication, after solving the SAA problem for every MCP penalty, we compute the two types of individual confidence intervals with the confidence level 0.95 ($\alpha = 0.05$).

Tables 2 and 3 show the empirical coverage rates (CR) and average interval lengths (Len) for 95% individual confidence intervals of the 500 replications. In Table 2, the “ $\tilde{\beta}$ ” column contains the population penalized parameters for different MCP penalties, which are

expected to be covered by the first type of confidence intervals. In Table 3, the “True” column contains the true model parameters β^{true} , which are expected to be covered by the second type of confidence intervals. Note that the coverage rate is 100% for the first type of confidence interval when $\tilde{\beta}_i = 0$. This is due to the shrinkage effect of the projection Γ (43) from z_0 to $\tilde{\beta}$, which causes the confidence intervals for $\tilde{\beta}_i$ to be the singleton $\{0\}$. When $\lambda = 0.5$ and $a = 2$, the population penalized parameters coincide with the true model parameters. The second type of confidence intervals are much longer than the first type for the inactive parameters $\beta_3, \beta_4, \beta_6, \beta_7$ and β_8 as expected. In practice, which type of confidence intervals to use depends on the type of parameters of interest. The first type of confidence intervals can be used to assess the randomness of the penalized estimates with a fixed penalty. This type of inference is especially useful when the penalty conveys prior information on the parameters. In contrast, the second type of confidence intervals provide inference information for the underlying true parameters directly.

As a remark, the parameter λ controls the level of penalization and the parameter a in the MCP penalty controls the degree of non-convexity. As shown in Tables 2 and 3, when λ increases to 1 and 2, the differences between the population penalized parameters and the true model parameters become larger. On the other hand, as a gets large, such as $a = 2000$, the MCP penalty becomes close to the LASSO penalty. The lengths of the second type of confidence intervals for $a = 2$ are generally shorter than those for $a = 2000$ with similar coverage rates as shown in Table 3. This may be due to the smaller bias imposed by the MCP penalty with a small a . In addition, as shown in Table 4, our proposed method with λ selected by GIC has very good performance. The comparison between our proposed methods and the LASSO-type methods indicate that our methods perform well for the inference of the true parameters in the linear model.

4.2 Example 2: Low dimensional setting with the equi-correlation covariance structure

In this example, we consider the equi-correlation covariance structure where $\Sigma_{ij} = 0.5$ for all $i \neq j$ and $\Sigma_{jj} = 1$ for all j . The other settings are the same as Example 1.

Table 5 shows the performance of the 95% individual confidence intervals of the population penalized parameters. The results shown in this table are very similar to the results of Example 1 shown in Table 2. As shown in Table 5, for each fixed λ , the proposed method using $a = 2$ delivers better performance than the method using $a = 2000$ in most cases, especially when λ is small. Table 6 shows the comparison of the individual confidence intervals of the true model parameters constructed by our method and the LASSO-type methods. Similar to Example 1, our proposed method using GIC performs well. In addition, for this example, our method (GIC), SVI-Lasso and LDPE methods deliver similar performance. All these three methods perform better than the JM method.

4.3 Example 3: High dimensional example

In this example, we consider a high dimensional case in which the dimension is much larger than the sample size. We choose $p = 300$ with β^{true} being a 300-dimensional vector:

$\beta_1^{true} = 3, \beta_2^{true} = \beta_{100}^{true} = \beta_{200}^{true} = \beta_{300}^{true} = 1.5, \beta_5^{true} = \beta_{95}^{true} = 2, \beta_{10}^{true} = 1, \beta_{25}^{true} = 0.5$, and all the

other components are 0. The true covariance matrix is $\Sigma_{ij} = 0.5^{|i-j|}$. We generate a training dataset with 500 replications of sample size $N = 100$. For this high dimensional example, we consider three MCP penalties with parameters $\lambda = 0.5, 1$ or 2 , and $a = 3$. In each replication, we use the nodewise LASSO regression introduced by Meinshausen and Bühlmann (2006) to compute the estimate of the precision matrix $\hat{\Theta}$, and compute the individual confidence intervals of the true model parameters with the confidence level 0.95. Define the active set as $\mathcal{A} = \{i: \beta_i^{\text{true}} \neq 0\} = \{1, 2, 5, 10, 25, 95, 100, 200, 300\}$ and $\mathcal{A}^c = \{0, 1, 2, \dots, 300\} \setminus \mathcal{A}$. In Table 7, for different methods, we report the average coverage rate, median coverage rate, average length and median length of the individual confidence intervals for true model parameters in $\mathcal{A}$ and $\mathcal{A}^c$ respectively:

$$\text{Avgcov } \mathcal{A} = |\mathcal{A}|^{-1} \sum_{i \in \mathcal{A}} \text{CR}_i, \text{ Avgcov } \mathcal{A}^c = |\mathcal{A}^c|^{-1} \sum_{i \in \mathcal{A}^c} \text{CR}_i,$$

$$\text{Avglen } \mathcal{A} = |\mathcal{A}|^{-1} \sum_{i \in \mathcal{A}} \text{Len}_i, \text{ Avglen } \mathcal{A}^c = |\mathcal{A}^c|^{-1} \sum_{i \in \mathcal{A}^c} \text{Len}_i,$$

$$\text{Medcov } \mathcal{A} = \text{median}\{\text{CR}_i\}_{i \in \mathcal{A}}, \text{ Medcov } \mathcal{A}^c = \text{median}\{\text{CR}_i\}_{i \in \mathcal{A}^c},$$

$$\text{Medlen } \mathcal{A} = \text{median}\{\text{Len}_i\}_{i \in \mathcal{A}}, \text{ Medlen } \mathcal{A}^c = \text{median}\{\text{Len}_i\}_{i \in \mathcal{A}^c},$$

where CR_i and Len_i denote the empirical coverage rate and average interval length of the confidence interval for the parameter β_i^{true} for the 500 replications, respectively.

For our proposed methods, as λ increases to 1 and 2, both the average coverage rates and lengths increase. Compared with LDPE, our proposed method using GIC to choose the tuning parameter has much shorter average lengths while the average coverage rates are only slightly lower. Although the JM method delivers similar average lengths as our proposed method (GIC), the average coverage rates of our proposed method are much closer to the nominal level 95%. Overall, the results shown in Table 7 indicate that our proposed method still delivers comparable performance for the high dimensional case.

4.4 Example 4: ADNI data

In this real data example, we consider the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (<http://www.loni.ucla.edu/ADNI>). The main goal of ADNI was to test whether the serial structural magnetic resonance imaging (MRI), fluorodeoxyglucose positron emission tomography (FDG-PET) images and some other biological markers such as cerebrospinal fluid (CSF) could be combined to measure the progression of mild

cognitive impairment (MCI) and early Alzheimer's Disease (AD). To that end, 800 adults with ages between 55 and 90 were recruited from over 50 sites across the US and Canada. In our analysis, we use data from 199 subjects who have complete baseline MRI, FDG-PET, and CSF data. Using the data processing method shown in Thung et al. (2014), we obtained 93 MRI features, 93 PET features, and 5 CSF features for each subject. The response variable is the Mini-Mental State Examination (MMSE) score (Folstein et al. (1975)) which is often used to screen for cognitive impairment.

The data are standardized at the beginning of our analysis. For our proposed method, we use the MCP penalty with the parameter $a = 3$ and choose the best tuning parameter λ by GIC. Table 8 shows the selected features of different methods, where the selected features are the features whose 95% confidence intervals do not contain 0. The numbers of features selected by our method, SVI-Lasso, LDPE, and JM are 13, 12, 13, and 3, respectively. Among the 13 features selected by our proposed method, 11 features are selected by the SVI-Lasso method, 9 features are selected by the LDPE method and 3 features are selected by the JM method. Table 9 shows the estimates and 95% individual confidence intervals of the 13 features selected by our proposed method. The results of our proposed method and the results of SVI-Lasso and LDPE methods are comparable. As shown in Table 9, for most features among the 13 features, the absolute values of the estimates delivered by the JM method are much smaller than the corresponding values of the other methods. The 95% confidence intervals of the JM method are also very different from the corresponding confidence intervals of the other methods.

5 Discussion

In this paper we propose a unified framework to construct confidence intervals for the population penalized parameters as well as the true model parameters for a large class of penalties. By transforming the population penalized regression problem (1) and its SAA problem (2) to the equivalent problems (9) and (22) respectively, we exclude the non-smoothness in the objectives. Furthermore, we obtain their normal map formulations (18) and (25), and derive the asymptotic distributions and the two types of confidence intervals. Our numerical results show that these methods are effective. When the objective functions in (1) and (2) are non-convex as a result of non-convex penalty functions, most existing algorithms are only guaranteed to obtain a local optimal solution for the SAA problem. Our proposed methods will generate confidence intervals based on that local solution. In practice, we solve for a SAA solution $(\hat{\beta}_0, \hat{\beta})$ and then use (26) to obtain a solution to (25). The first type of confidence intervals we compute are for a local optimal solution of the population penalized regression problem (1). From any local solution of (2), we can always compute confidence intervals for the true model parameters, which are the second type of confidence intervals we compute.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the editors, the associate editor, and referees for their helpful comments and suggestions. This research was supported in part by US National Science Foundation grants DMS-1407241 (Liu, Lu and Yin), and DMS-1109099 (Lu and Yin).

References

- Brooke A, Kendrick D, Meeraus A, and Raman R (1998), GAMS, A User's Guide, Washington, DC: GAMS Development Corporation, available online at <http://www.gams.com>.
- Candes EJ and Tao T (2007), "The Dantzig selector: statistical estimation when p is much larger than n ," *The Annals of Statistics*, 35, 2313–2351.
- Donoho DL and Johnstone IM (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425–455.
- Efron B, Hastie T, Johnstone I, and Tibshirani R (2004), "Least angle regression," *The Annals of Statistics*, 32, 407–499.
- Fan J and Li R (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.
- Folstein MF, Folstein SE, and McHugh PR (1975), "Mini-mental state: a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, 12, 189–198. [PubMed: 1202204]
- Friedman JH (2012), "Fast sparse regression and classification," *International Journal of Forecasting*, 28, 722–738.
- Javanmard A and Montanari A (2014), "Confidence intervals and hypothesis testing for high-dimensional regression," *Journal of Machine Learning Research*, 15, 2869–2909.
- Konishi S and Kitagawa G (2008), *Information criteria and statistical modeling*, Springer Science & Business Media.
- Lee JD, Sun DL, Sun Y, Taylor JE, et al. (2016), "Exact post-selection inference, with application to the lasso," *The Annals of Statistics*, 44, 907–927.
- Liu Y and Wu Y (2007), "Variable selection via a combination of the L_0 and L_1 penalties," *Journal of Computational and Graphical Statistics*, 16, 782–798.
- Lockhart R, Taylor J, Tibshirani R, and Tibshirani R (2014), "A significance test for the lasso," *The Annals of Statistics*, 42, 413–468. [PubMed: 25574062]
- Lu S (2014a), "A new method to build confidence regions for solutions of stochastic variational inequalities," *Optimization: A Journal of Mathematical Programming and Operations Research*, 63, 1431–1443.
- Lu S (2014b), "Symmetric Confidence Regions and Confidence Intervals for Normal Map Formulations of Stochastic Variational Inequalities," *SIAM Journal on Optimization*, 24, 1458–1484.
- Lu S and Budhiraja A (2013), "Confidence regions for stochastic variational inequalities," *Mathematics of Operations Research*, 38, 545–568.
- Lu S, Liu Y, Yin L, and Zhang K (2017), "Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 589–611.
- Lv J and Fan Y (2009), "A unified approach to model selection and sparse recovery using regularized least squares," *The Annals of Statistics*, 37, 3498–3528.
- Mazumder R, Friedman J, and Hastie T (2011), "SparseNet: Coordinate descent with non-convex penalties," *Journal of the American Statistical Association*, 106, 1125–1138. [PubMed: 25580042]
- Meinshausen N and Bühlmann P (2006), "High-dimensional graphs and variable selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462.
- Nikolova M (2000), "Local strong homogeneity of a regularized estimator," *SIAM Journal on Applied Mathematics*, 61, 633–658.

- Ning Y, Liu H, et al. (2017), “A general theory of hypothesis tests and confidence regions for sparse high dimensional models,” *The Annals of Statistics*, 45, 158–195.
- Robinson SM (1995), “Sensitivity analysis of variational inequalities by normal-map techniques,” in *Variational Inequalities and Network Equilibrium Problems*, ed. Giannessi F and Maugeri A, New York: Plenum Press, pp. 257–269.
- Thung K-H, Wee C-Y, Yap P-T, Shen D, Initiative ADN, et al. (2014), “Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion,” *NeuroImage*, 91, 386–400. [PubMed: 24480301]
- Tibshirani R (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B*, 58, 267–288.
- Van de Geer S, Bühlmann P, Ritov Y, and Dezeure R (2014), “On asymptotically optimal confidence regions and tests for high-dimensional models,” *The Annals of Statistics*, 42, 1166–1202.
- Voorman A, Shojaie A, and Witten D (2014), “Inference in high dimensions with the penalized score test,” arXiv preprint arXiv:1401.2678.
- Wu TT and Lange K (2008), “Coordinate descent algorithms for lasso penalized regression,” *The Annals of Applied Statistics*, 2, 224–244.
- Zhang CH (2010), “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of Statistics*, 38, 894–942.
- Zhang CH and Zhang SS (2014), “Confidence intervals for low dimensional parameters in high dimensional linear models,” *Journal of the Royal Statistical Society: Series B*, 76, 217–242.
- Zhao S, Shojaie A, and Witten D (2017), “In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference,” arXiv preprint arXiv:1705.05543.
- Zou H (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101, 1418–1429.
- Zou H and Hastie T (2005), “Regularization and variable selection via the elastic net,” *The Annals of Statistics*, 67, 301–320.

Figure 1:
The normal manifold of S_i (left) and $E_i^0, \dots, E_i^8$ (right).

Table 1:

Construction of the $(1 - \alpha)\%$ confidence intervals of β_0^{true} and β^{true}

Step 1. Find the penalized estimates $\hat{\beta}_0$ and $\hat{\beta}$ by solving the SAA problem (2). The tuning parameters are chosen by the Generalized Information Criterion (GIC).

Step 2. Calculate the solution of the normal map formulation (25),

$$z_N = (\hat{\beta}_0, \hat{\beta}, \hat{t}) - f_N(\hat{\beta}_0, \hat{\beta}, \hat{t}), \text{ where } \hat{t}_i = |\hat{\beta}_i|, i = 1, \dots, p.$$

Step 3. Calculate $(\hat{\beta}_0^{true}, \hat{\beta}^{true}) = G^*(z_N)$, where G^* is the function defined in (59).

Step 4. If for every $i \in \{0, 1, 2, \dots, p\}$, $\hat{h}_i$ is very close to 0, we consider Case I to construct individual confidence intervals approximately. Otherwise, we consider Case II.

Case I: the $(1 - \alpha)\%$ confidence interval of β_i^{true} is $[\hat{\beta}_i^{true} - \Phi^{-1}(1 - \alpha/2) \cdot \tau_i, \hat{\beta}_i^{true} + \Phi^{-1}(1 - \alpha/2) \cdot \tau_i]$, where $\tau_i = \sqrt{H_N \Sigma_N H_N^T / N}$ and H_N is defined in Theorem 5;

Case II: we first use simulation to estimate the $(1 - \alpha/2)\%$ percentile of $\hat{R}_{i+1}(Z)$, where $\hat{R}_{i+1}$ is defined by (56) and $Z \sim \mathcal{N}(0, I_{p+1}) \times \vec{1}$. The estimated percentile is denoted as η_i . The $(1 - \alpha)\%$ confidence interval of β_i^{true} is $[\hat{\beta}_i^{true} - \eta_i / \sqrt{N}, \hat{\beta}_i^{true} + \eta_i / \sqrt{N}]$.

Table 2:

Coverage rates and average lengths of 95% individual CIs for population penalized parameters $(\tilde{\beta}_0, \tilde{\beta})$ for different MCP penalties from 500 replications with sample size $N = 300$ generated in Example 1.

	$a = 2$									$a = 2000$								
	$\lambda = 0.5$			$\lambda = 1$			$\lambda = 2$			$\lambda = 0.5$			$\lambda = 1$			$\lambda = 2$		
	$\tilde{\beta}$	CR	Len	$\tilde{\beta}$	CR	Len	$\tilde{\beta}$	CR	Len	$\tilde{\beta}$	CR	Len	$\tilde{\beta}$	CR	Len	$\tilde{\beta}$	CR	Len
β_0	0	0.99	0.26	0	0.99	0.27	0	0.98	0.39	0	0.98	0.28	0	0.98	0.32	0	0.98	0.46
β_1	3	0.97	0.30	3.13	0.97	0.36	3.37	0.95	0.88	2.83	0.97	0.32	2.67	0.98	0.38	2.33	0.96	0.56
β_2	1.5	0.97	0.30	1.25	0.97	0.49	0.51	0.96	0.84	1.36	0.98	0.33	1.22	0.98	0.38	0.94	0.98	0.56
β_3	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.04	0	1.00	0.01	0	1.00	0.00
β_4	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.04	0	1.00	0.01	0	1.00	0.00
β_5	2	0.98	0.26	2.02	0.98	0.30	1.47	0.98	0.56	1.78	0.99	0.29	1.56	0.98	0.34	1.11	0.98	0.52
β_6	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.02	0	1.00	0.00	0	1.00	0.00
β_7	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.01	0	1.00	0.00	0	1.00	0.00
β_8	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00

Table 3:

Coverage rates and average lengths of 95% individual CIs for true model parameters $(\beta_0^{true}, \beta^{true})$ for different MCP penalties from 500 replications with sample size $N = 300$ generated in Example 1.

	True	$a = 2$						$a = 2000$					
		$\lambda = 0.5$		$\lambda = 1$		$\lambda = 2$		$\lambda = 0.5$		$\lambda = 1$		$\lambda = 2$	
		CR	Len	CR	Len	CR	Len	CR	Len	CR	Len	CR	Len
β_0^{true}	0	0.96	0.23	0.96	0.23	0.95	0.34	0.96	0.24	0.95	0.28	0.95	0.40
β_1^{true}	3	0.95	0.26	0.96	0.27	0.99	0.42	0.96	0.28	0.99	0.33	1.00	0.49
β_2^{true}	1.5	0.95	0.29	0.96	0.30	1.00	0.51	0.96	0.31	0.97	0.36	1.00	0.53
β_3^{true}	0	0.95	0.28	0.96	0.29	0.97	0.42	0.97	0.30	0.97	0.35	0.98	0.50
β_4^{true}	0	0.97	0.28	0.97	0.29	0.97	0.42	0.96	0.30	0.97	0.35	0.98	0.50
β_5^{true}	2	0.96	0.28	0.96	0.29	0.99	0.44	0.97	0.30	1.00	0.36	1.00	0.54
β_6^{true}	0	0.95	0.28	0.95	0.28	0.98	0.42	0.95	0.30	0.97	0.34	0.98	0.49
β_7^{true}	0	0.96	0.28	0.96	0.29	0.98	0.42	0.96	0.30	0.97	0.35	0.99	0.50
β_8^{true}	0	0.96	0.25	0.97	0.26	0.99	0.38	0.97	0.27	0.99	0.32	1.00	0.45

Table 4:

Coverage rates and average lengths of 95% individual CIs for true model parameters $(\beta_0^{true}, \beta^{true})$ for different methods from 500 replications with sample size $N = 300$ generated in Example 1.

	True	Our method ($a = 2$)						Lasso type methods							
		$\lambda = 0.5$		$\lambda = 1$		$\lambda = 2$		GIC		SVI-Lasso		LDPE		JM	
		CR	Len	CR	Len	CR	Len	CR	Len	CR	Len	CR	Len	CR	Len
β_1^{true}	3	0.95	0.26	0.96	0.27	0.99	0.42	0.95	0.26	0.95	0.26	0.94	0.26	0.88	0.25
β_2^{true}	1.5	0.95	0.29	0.96	0.30	1.00	0.51	0.95	0.29	0.95	0.29	0.94	0.28	0.89	0.28
β_3^{true}	0	0.95	0.28	0.96	0.29	0.97	0.42	0.95	0.28	0.95	0.28	0.95	0.28	0.99	0.28
β_4^{true}	0	0.97	0.28	0.97	0.29	0.97	0.42	0.97	0.28	0.97	0.28	0.96	0.28	0.97	0.28
β_5^{true}	2	0.96	0.28	0.96	0.29	0.99	0.44	0.96	0.28	0.97	0.29	0.96	0.28	0.92	0.28
β_6^{true}	0	0.95	0.28	0.95	0.28	0.98	0.42	0.95	0.28	0.94	0.28	0.95	0.28	0.98	0.28
β_7^{true}	0	0.96	0.28	0.96	0.29	0.98	0.42	0.96	0.28	0.95	0.28	0.95	0.28	0.99	0.28
β_8^{true}	0	0.96	0.25	0.97	0.26	0.99	0.38	0.97	0.25	0.96	0.26	0.97	0.26	0.98	0.25

Table 5:

Coverage rates and average lengths of 95% individual CIs for population penalized parameters ($\tilde{\beta}_0, \tilde{\beta}$) for different MCP penalties from 500 replications with sample size $N = 300$ generated in Example 2.

$a = 2$										$a = 2000$									
	$\lambda = 0.5$			$\lambda = 1$			$\lambda = 2$				$\lambda = 0.5$			$\lambda = 1$			$\lambda = 2$		
	$\tilde{\beta}$	CR	Len	$\tilde{\beta}$	CR	Len	$\tilde{\beta}$	CR	Len		$\tilde{\beta}$	CR	Len	$\tilde{\beta}$	CR	Len	$\tilde{\beta}$	CR	Len
β_0	0	0.98	0.26	0	0.98	0.27	0	0.99	0.38	0	0.98	0.27	0	0.97	0.31	0	0.97	0.41	
β_1	3	0.99	0.31	3.10	0.98	0.36	3.57	0.91	0.95	2.88	0.98	0.33	2.75	0.97	0.38	2.50	0.97	0.52	
β_2	1.5	0.98	0.32	1.20	0.98	0.55	0.57	0.95	0.88	1.37	0.97	0.34	1.25	0.97	0.38	1.00	0.96	0.52	
β_3	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.05	0	1.00	0.02	0	1.00	0.01	
β_4	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.05	0	1.00	0.02	0	1.00	0.00	
β_5	2	0.97	0.32	2.10	0.99	0.38	1.57	0.98	0.93	1.88	0.97	0.34	1.75	0.97	0.38	1.50	0.98	0.52	
β_6	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.05	0	1.00	0.02	0	1.00	0.00	
β_7	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.05	0	1.00	0.02	0	1.00	0.00	
β_8	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.05	0	1.00	0.02	0	1.00	0.00	

Table 6:

Coverage rates and average lengths of 95% individual CIs for true model parameters $(\beta_0^{true}, \beta^{true})$ for different methods from 500 replications with sample size $N = 300$ generated in Example 2.

	True	Our method ($a = 2$)						Lasso type methods							
		$\lambda = 0.5$		$\lambda = 1$		$\lambda = 2$		GIC		SVI-Lasso		LDPE		JM	
		CR	Len	CR	Len	CR	Len	CR	Len	CR	Len	CR	Len	CR	Len
β_1^{true}	3	0.97	0.30	0.97	0.31	1.00	0.46	0.97	0.30	0.97	0.30	0.97	0.30	0.91	0.29
β_2^{true}	1.5	0.95	0.30	0.97	0.32	1.00	0.50	0.95	0.30	0.95	0.31	0.95	0.30	0.89	0.29
β_3^{true}	0	0.95	0.30	0.96	0.31	0.99	0.43	0.95	0.30	0.96	0.30	0.96	0.30	0.98	0.29
β_4^{true}	0	0.96	0.30	0.97	0.31	0.99	0.43	0.96	0.30	0.96	0.30	0.96	0.30	0.99	0.29
β_5^{true}	2	0.92	0.30	0.94	0.31	0.99	0.46	0.92	0.30	0.92	0.31	0.92	0.30	0.88	0.29
β_6^{true}	0	0.95	0.30	0.95	0.31	1.00	0.43	0.95	0.30	0.95	0.30	0.94	0.30	0.98	0.29
β_7^{true}	0	0.95	0.30	0.95	0.31	0.99	0.43	0.95	0.30	0.95	0.30	0.95	0.30	0.96	0.29
β_8^{true}	0	0.94	0.30	0.95	0.31	1.00	0.43	0.94	0.30	0.94	0.30	0.94	0.30	0.98	0.29

Table 7:

Average coverage rates and lengths of 95% individual confidence intervals for the true model parameters in the linear model with different methods computed from 500 replications with sample size $N = 100$ and dimension $p = 300$ generated in Example 3.

	Our method ($\lambda = 0.5$)				Our method ($\lambda = 1$)				Our method ($\lambda = 2$)			
	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen
$\mathcal{A}$	92.82	92.60	0.46	0.45	94.64	95.00	0.66	0.65	95.24	95.40	1.26	1.25
$\mathcal{A}^c$	93.26	93.40	0.39	0.39	93.51	93.60	0.56	0.56	93.73	94.00	1.08	1.08
	Our method (GIC)				LDPE				JM			
	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen	Avgcov	Medcov	Avglen	Medlen
$\mathcal{A}$	92.91	93.00	0.57	0.56	93.84	94.40	1.13	1.14	88.07	87.80	0.55	0.55
$\mathcal{A}^c$	93.37	93.40	0.47	0.47	95.31	95.60	1.14	1.14	99.38	99.40	0.55	0.55

Table 8:

Selected features of different methods for the ADNI data.

Method	Selected Features
Our method (GIC)	9, 19, 40, 59, 67, 80, 95, 130, 134, 147, 156, 168, 178
SVI-Lasso	9, 19, 40, 77, 80, 95, 130, 134, 147, 156, 168, 178
LDPE	9, 19, 40, 59, 77, 80, 83, 90, 111, 134, 147, 156, 168
JM	19, 40, 134

Table 9:

Estimates and 95% individual confidence intervals of the 13 features selected by our proposed method for the ADNI data.

	Our method (GIC)		SVI-Lasso		LDPE		JM	
	Est	Ind CI	Est	Ind CI	Est	Ind CI	Est	Ind CI
β_9^{true}	-0.20	[-0.33, -0.07]	-0.20	[-0.33, -0.07]	-0.19	[-0.34, -0.04]	-0.15	[-0.31, 0.01]
β_{19}^{true}	0.24	[0.06, 0.41]	0.23	[0.07, 0.40]	0.24	[0.09, 0.40]	0.25	[0.08, 0.42]
β_{40}^{true}	-0.21	[-0.36, -0.06]	-0.21	[-0.35, -0.06]	-0.20	[-0.35, -0.05]	-0.16	[-0.33, 0.00]
β_{59}^{true}	0.15	[0.01, 0.29]	0.15	[0.00, 0.30]	0.16	[0.01, 0.30]	0.12	[-0.03, 0.28]
β_{67}^{true}	0.13	[0.00, 0.27]	0.13	[0.00, 0.26]	0.12	[-0.02, 0.26]	0.11	[-0.03, 0.26]
β_{80}^{true}	0.23	[0.03, 0.43]	0.23	[0.03, 0.42]	0.21	[0.04, 0.38]	0.15	[-0.01, 0.30]
β_{95}^{true}	0.20	[0.00, 0.40]	0.21	[0.01, 0.41]	0.19	[-0.01, 0.39]	0.04	[-0.11, 0.20]
β_{130}^{true}	0.21	[0.00, 0.42]	0.20	[0.01, 0.40]	0.18	[-0.03, 0.39]	0.04	[-0.12, 0.19]
β_{134}^{true}	0.25	[0.08, 0.43]	0.24	[0.02, 0.45]	0.24	[0.06, 0.43]	0.23	[0.08, 0.38]
β_{147}^{true}	-0.22	[-0.41, -0.02]	-0.22	[-0.41, -0.03]	-0.21	[-0.40, -0.02]	-0.03	[-0.18, 0.13]
β_{156}^{true}	-0.19	[-0.36, -0.02]	-0.19	[-0.36, -0.02]	-0.18	[-0.37, 0.00]	-0.05	[-0.21, 0.11]
β_{168}^{true}	-0.24	[-0.43, -0.04]	-0.24	[-0.43, -0.04]	-0.22	[-0.43, -0.02]	0.01	[-0.14, 0.16]
β_{178}^{true}	-0.19	[-0.34, -0.03]	-0.19	[-0.35, -0.04]	-0.18	[-0.36, 0.00]	-0.07	[-0.22, 0.08]