Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization

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

[Received December 2014. Revised February 2016]

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

Keywords: Confidence intervals; Confidence regions; Optimization; Random design; Shrinkage; Sparsity; Variational inequality

1. Introduction

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

Regarding data-driven model selection procedures, sparse regularized techniques have been popular for simultaneous variable selection and prediction. By introducing biases on the estimators through sparse penalization, these methods can often produce estimators with much smaller variances and consequently lower mean-square errors than unpenalized estimators. Furthermore, because of the built-in sparsity on the estimators, model selection and parameter estimation can be achieved in a single step. There is a large literature in this area including the $L_1$-regularized technique the lasso (Donoho and Johnstone, 1994; Tibshirani, 1996), as well as many other extensions (e.g. Fan and Li (2001) and Candès and Tao (2007)). See Bühlmann and van de Geer (2011) for a comprehensive review.

After data-driven selection, one common practice is to carry out conventional inference on the model selected. Despite its prevalence, this practice is problematic because it ignores the fact that the inference is conditional on the model selection that is itself stochastic. The stochastic nature of the selection process affects and distorts sampling distributions of the post-selection
parameter estimates, leading to invalid post-selection inference. This problem has long been recognized and was discussed recently by Breiman (1992) and Berk et al. (2013a).

In recent years, many methods have been developed to achieve valid inference after the lasso. We categorize these methods into the following three types of approach:

(a) the simultaneous inference approach, which is guided by a general heuristic to consider all possible outcomes of the model selected and to protect the valid inference for the worst scenario (Chatterjee and Lahiri, 2011; Minnier et al., 2011; Berk et al., 2013a);

(b) the bias correction approach, which considers adjusting for the bias that is introduced by the regularization step to achieve valid inference (references along this line include Bühlmann (2013), Zhang and Zhang (2014), Van de Geer et al. (2014) and Javanmard and Montanari (2015));

(c) the conditional sampling distribution approach, which aims at understanding the asymptotic or exact distributions of some pivots conditionally on the model selected and developing methods of inference that are based on these distributions (Lockhart et al., 2014; Lee et al., 2014).

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

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

(1)

where \( \{x_{ij}\}_{j=1}^{p}, y_i, i = 1, \ldots, N, \) are independently and identically distributed samples with \( x_{ij}, y_i \in \mathbb{R}, \lambda > 0 \) is the tuning parameter and \( \beta_0 \in \mathbb{R} \) and \( \beta = (\beta_1, \ldots, \beta_p) \) are the regression parameters. Our interest is in its inference. For that, we study the following population version of the lasso by solving

\[ \min_{\beta_0, \beta} \left( \mathbb{E} \left[ Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right]^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \]  

(2)

where \( X \in \mathbb{R}^p \) is the random input vector and \( Y \in \mathbb{R} \) is the response variable. We refer to problem (1) as the sample average approximation (SAA) problem of the population lasso problem (2).

We denote the solution to the SAA problem (1) as the sample average approximation (SAA) problem of the population lasso problem (2). Our interest is in its inference. For that, we study the following population version of the lasso by solving

\[ \min_{\beta_0, \beta} \left( \mathbb{E} \left[ Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right]^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \]  

(2)

where \( X \in \mathbb{R}^p \) is the random input vector and \( Y \in \mathbb{R} \) is the response variable. We refer to problem (1) as the sample average approximation (SAA) problem of the population lasso problem (2). We denote the solution to the SAA problem (1) as \( (\hat{\beta}_0, \hat{\beta}) \), which we refer to as lasso estimators. Our first contribution is to make use of the lasso estimators \( (\hat{\beta}_0, \hat{\beta}) \) to derive confidence intervals and regions for the population lasso parameters \( (\beta_0, \beta) \), the solution of problem (2).

The population lasso approach is closely related to the traditional least squares approach. When \( \lambda \) is 0, problem (2) becomes the population least squares problem \( \min_{\beta_0, \beta} \mathbb{E} \left[ Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right]^2 \), which has a unique minimizer \( \beta_0 \text{true} + \beta \text{true}^T X \), with \( \mathbb{E} [\epsilon | X] = 0 \), then the solution to the population least squares problem is exactly \( (\hat{\beta}_0 \text{true}, \hat{\beta} \text{true}) \). When \( \lambda > 0 \), the solution to problem (2) is not exactly \( (\beta_0 \text{true}, \beta \text{true}) \) but is related to \( (\beta_0 \text{true}, \beta \text{true}) \) in a different way. Our second contribution is to develop a method which utilizes that relationship to construct confidence intervals for the parameters \( (\beta_0 \text{true}, \beta \text{true}) \) in the linear model.

Why could the minimizer from a population lasso be a reasonable target for scientific research? Although it is apparent that a selection procedure such as the lasso is necessary when \( p > n \), the population lasso approach is also meaningful when \( n > p \). In the latter case, although least squares inference of all coefficients in the model is readily available, it is well known that including nearly collinear redundant variables in a regression model can ‘adjust away’ some of the causal
variables of interest. Moreover, using the full model could be questionable in areas such as social science (Berk et al., 2013b). In these areas, it is common that, when the question of “which variables should be included in the regression model?” is asked, scientific theory is not sufficient to dictate the inclusion or exclusion of variables for the inference (even when $n > p$). In this case, a data-driven model from the lasso would be helpful. However, under this situation, the goal of inference is slightly changed from that of the least squares approach: the investigator is no longer looking for the least squares coefficients that minimize the squared error loss in the population. Instead, she wants to find the least squares estimate subject to certain regularization on the model. From a modelling perspective, this consideration means that the loss function has changed from the squared error loss to the sum of the square error and the regularization on $\beta_j$s, as the objective function in expression (1). The corresponding risk function, which is the expected loss, is then the objective function in expression (2). On the basis of this consideration, the population lasso parameters defined in problem (2) are precisely the minimizer of the risk function, and thus they become a reasonable target for inference.

The major difference between the population lasso approach and the least squares approach is the incorporation of constraint information about the model or parameters. Though the source of such information can be from different perspectives, they can all be reflected in the regularization term with $\lambda$ as a measure of the strength of such information. Thus, the parameters in the population lasso approach are both scientifically and statistically meaningful: they lead to the best approximations to the response when external information is available.

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

(a) Lasso estimators from expression (1) depend on random samples and are subject to uncertainty. Our inference results provide quantitative measures about the level of such uncertainty, by estimating the distance between the population lasso parameters and the computed lasso estimators. Sizes of those intervals are jointly determined by sample variability and sensitivity of lasso estimators with respect to random samples. Wide intervals indicate low reliability of the estimators, which can be caused by large sample variability or high sensitivity. Thus, these inference results can be used as quantitative assessments on the reliability and uncertainty level of lasso estimators obtained from expression (1).

(b) The inference results of this paper can be used to assess the relative importance of predictors. For non-zero lasso estimators, conclusions can be made regarding whether the corresponding parameters are truly non-zero by checking whether the corresponding intervals contain zero or not. For zero lasso estimators, the inference results can be highly informative as well. For example, if the confidence intervals of some lasso parameters are singletons of zero (see Section 3.2 for related discussion), then we have strong evidence to conclude that the corresponding population lasso parameters are zero.

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

The key techniques of our method are asymptotic results for solutions of problem (1) as the sample size grows. In this paper, we develop theories based on the fixed dimension $p$, although it is possible to extend this idea to the case of growing dimensions. The development of our method takes the following steps. First, we transform problems (1) and (2) into their corresponding normal map formulations (see expressions (15) and (9) in Section 2 for more details), which are equations with a $(2p + 1)$-dimensional variable vector $z$. Next, we obtain the asymptotic distribution of solutions to the normal map formulation of problem (1) and find reliable estimates for quantities that appear in the asymptotic distribution. We then provide methods to compute simultaneous and individual confidence intervals for the solution to the normal map formulation of problem (2). Finally, we convert these confidence intervals into confidence intervals for the solution to problem (2). Our inference method is developed for a fixed value of $\lambda$. In practice, the value of $\lambda$ can be chosen by various criteria or through cross-validation. To provide inference for the true parameters $(\beta_{0, \text{true}}, \beta_{1, \text{true}})$, we establish a connection between the population lasso parameters (which solves problem (2)) and the true parameters with expression (32) in Section 4. Estimators of the true parameters can be obtained in expression (33), and confidence intervals for the true parameters are then constructed based on the asymptotic distributions of those estimators.

The rest of the paper is organized as follows. We first introduce the assumptions and problem transformations in Section 2 and then discuss construction of confidence intervals and regions for the population lasso parameters in Section 3. In Section 4 we provide inference for the true parameters assuming that $X$ and $Y$ are related by a linear model. Numerical results are presented in Section 5 to illustrate the performance of the methods proposed, followed by some discussion. The on-line appendix collects technical details of variational inequalities, proofs and additional numerical results.

In this paper, we use $(x, y)$ to denote the inner product between two vectors $x$ and $y$. We use $\| \cdot \|$ to denote the norm of an element in a normed space; unless explicitly stated otherwise, it can be any norm, as long as the same norm is used in all related contexts. Let $N(0, \Sigma)$ denote a normal random vector with covariance matrix $\Sigma$. Weak convergence of random variables $Y_n$ to $Y$ is denoted as $Y_n \Rightarrow Y$. A function $g : \mathbb{R}^n \to \mathbb{R}^m$ is said to be $B$ differentiable at a point $x_0 \in \mathbb{R}^n$ if there is a positively homogeneous function $G : \mathbb{R}^n \to \mathbb{R}^m$, such that $g(x_0 + v) = g(x_0) + G(v) + o(v)$. The function $G$ is the $B$-derivative of $g$ at $x_0$ and will be written as $dg(x_0)$. For each $h \in \mathbb{R}^n$, $dg(x_0)(h)$ is exactly the directional derivative of $g$ at $x_0$ for the direction $h$. In general, $B$-differentiability is a stronger property than directional differentiability, as it requires $dg(x_0)(\cdot)$ to be a first-order approximation of $g(x_0 + \cdot)$ uniformly in all directions.

The programs that were used to analyse the data can be obtained from

http://wileyonlinelibrary.com/journal/rss-datasets

2. Problem transformations

In this section, we describe how to transform problems (1) and (2) into quadratic programs, variational inequalities and their normal map formulations. On the basis of those transforma-
tions, we shall obtain the asymptotic distribution of SAA solutions in Section 3. The on-line appendix A gives background information about variational inequalities and normal maps.

2.1. Conversion to a standard quadratic program

In this subsection, we transform the population lasso problem into a standard quadratic program. We need assumption 1, part (a), below to guarantee that the objective function of expression (2) is finite valued. We shall use the stronger assumption 1, part (b), in proving convergence results.

Assumption 1.

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

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

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

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

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

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

$$
F(\beta_0, \beta, t, X, Y) = \begin{pmatrix}
-2 \left( Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right) \\
-2 \left( Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right) X_1 \\
\vdots \\
-2 \left( Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right) X_p \\
\lambda e_p
\end{pmatrix},
\quad d_1 F(\beta_0, \beta, t, X, Y) = \begin{pmatrix}
2 & 2X^T & 0 \\
2X & 2XX^T & 0 \\
0 & 0 & 0
\end{pmatrix},
$$

where $e_p \in \mathbb{R}^p$ is the vector of all 1s. Next, define the expectations of $F$ and $d_1 F$ as

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

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

The following lemma is straightforward and its proof has been omitted.

Lemma 1. Suppose that assumption 1, part (a), holds. Then, the objective function of expression (3) is a finite-valued, convex quadratic function on $\mathbb{R}^{2p+1}$, its gradient at each $(\beta_0, \beta, t) \in \mathbb{R}^{2p+1}$ is $f_0(\beta_0, \beta, t)$ and its Hessian matrix is $L$.

We now introduce the second assumption.
Assumption 2. Let \((\tilde{\beta}_0, \tilde{\beta})\) be an optimal solution of problem (2) and define \(\tilde{t} \in \mathbb{R}^n\) and \(\tilde{q} \in \mathbb{R}^n\) by
\[
\tilde{t}_i = |\tilde{\beta}_i| \quad \text{and} \quad \tilde{q}_i = E \left[ -2 \left( Y - \tilde{\beta}_0 - \sum_{j=1}^{p} \tilde{\beta}_j X_j \right) X_i \right] \quad \text{for each } i = 1, \ldots, p.
\]
Define an index set \(\mathcal{J}\) as \(\mathcal{J} = \{i \in \{1, \ldots, p\} \mid \text{either } |\tilde{\beta}_i| \neq 0 \text{ or } |\tilde{\beta}_i| = 0 \text{ and } \tilde{q}_i = \lambda \}\), and let \(Q\) be the submatrix of \(L\) in expression (6) that consists of intersections of columns and rows of \(L\) with indices in \(\{1\} \cup \{i + 1, i \in \mathcal{J}\}\). Assume that \(Q\) is non-singular.

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

2.2. The variational inequality and normal map formulation

In view of lemma 1, and on the basis of the relationship between constrained optimization and variational inequalities that is discussed in the on-line appendix A, we can rewrite problem (3) as the following variational inequality:

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

\[
(f_0)_S(z) = f_0(\Pi_S(z)) + z - \Pi_S(z) \quad \text{for each } z \in \mathbb{R}^{2p+1},
\]

where \(\Pi_S(z)\) denotes the Euclidean projection of \(z\) onto \(S\), namely the point in \(S\) that is nearest to \(z\) in Euclidean norm. The normal map \((f_0)_S\) as defined above is a function from \(\mathbb{R}^{2p+1}\) and \(\mathbb{R}^{2p+1}\).

Because \(S\) is a polyhedral convex set in \(\mathbb{R}^{2p+1}\), the Euclidean projector \(\Pi_S\) is a piecewise affine function from \(\mathbb{R}^{2p+1}\) to \(\mathbb{R}^{2p+1}\); it coincides with an affine function on each of finitely many \((2p + 1)\)-dimensional polyhedra whose union is \(\mathbb{R}^{2p+1}\) (the dimension of a convex set is defined to be the dimension of its affine hull, which is the smallest affine set containing the set). Those polyhedra, along with their faces, are called cells in the normal manifold of \(S\). We call a cell with dimension \(k\) a \(k\)-cell. The relative interiors of all cells in the normal manifold form a partition of \(\mathbb{R}^{2p+1}\) (the relative interior of a convex set is its interior relative to its affine hull). Since \(f_0\) is affine under assumption 1, part (a), and \(\Pi_S\) is piecewise affine, the normal map \((f_0)_S\) is a piecewise affine function from \(\mathbb{R}^{2p+1}\) to \(\mathbb{R}^{2p+1}\). With \(z\) a variable of dimension \(2p + 1\), the normal map formulation for expression (7) is the equation

\[
(f_0)_S(z) = 0.
\]

As noted immediately below assumption 2, the vector \((\tilde{\beta}_0, \tilde{\beta}, \tilde{t})\) is a solution of problem (3). It is therefore a solution of problem (7) as well. By the relationship between variational inequalities and normal maps (see the on-line appendix A), the point \(z_0 \in \mathbb{R}^{2p+1}\) defined as
is a solution to equation (9) and satisfies $\Pi_S(z_0) = (\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$. Let $K$ be the critical cone to $S$ associated with $z_0$, defined as

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

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

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

Using the special polyhedral structure of $S$, we shall give an explicit expression of $K$ in the proof of lemma 2 below. Critical cones are commonly used to express optimality conditions. We shall use critical cones to express the asymptotic distribution of SAA solutions. Let $L_K$ be the normal map that is induced by the linear function defined by the matrix $L$ in expression (6) and the cone $K$, defined as in expression (8) with $L$ and $K$ in place of $f_0$ and $S$ respectively. In lemma 2 below, we show that $L_K$ is a global homeomorphism from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$, i.e. a continuous bijective function from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$ whose inverse function is also continuous. The inverse function of $L_K$ will appear in an expression for the asymptotic distribution of SAA solutions in theorem 1.

**Lemma 2.** Suppose that assumption 1, part (a), and assumption 2 hold. Then the normal map $L_K$ is a global homeomorphism from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{2p+1}$, and $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is the unique optimal solution of problem (3).

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

### 2.3. Transformations of sample average approximation problems

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

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

where $S$ is as defined in expression (4). For each $i$, let $x_i = (x_{ij})_{j=1}^{p}$ be the $p$-dimensional vector consisting of observed features of the $i$th sample. We define the SAA function

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

where $F$ is as in expression (5). By noting that $f_N(\beta_0, \beta, t)$ is exactly the gradient of the objective function of problem (12) at $(\beta_0, \beta, t)$, we can rewrite problem (12) as a variational inequality

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

The above $f_N$ is an affine function with its Jacobian matrix given by
Finally, we let $f_N/S$ be the normal map induced by $f_N$ and $S$, and write the normal map formulation of expression (13) as

\[
(f_N)_S(z) = 0.
\]

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

3. Inference for population lasso parameters

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

3.1. The convergence and distributions of sample average approximation solutions

Theorem 1 below gives convergence properties and asymptotic distributions of solutions of the SAA problems (13) and (15). It shows under assumptions 1 and 2 that equation (15) has a unique solution $z_N$ for sufficiently large $N$, and that $z_N$ converges almost surely to $z_0$ defined in expression (10). Correspondingly, the projection $\Pi_S(z_N)$ is the unique solution of problem (13), which converges almost surely to $\tilde{\beta}_0, \tilde{\beta}, \tilde{t}$. Theorem 1 also provides asymptotic distributions of $z_N$ and $\Pi_S(z_N)$, and gives their rate of convergence in probability under assumption 3, parts (a) and (b), in the on-line appendix B.

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

\[
\lim_{N \to \infty} z_N = z_0 \quad \text{almost everywhere,}
\]

\[
\lim_{N \to \infty} (\hat{\beta}_0, \hat{\beta}, \hat{t}) = (\beta_0, \hat{\beta}, \hat{t}) \quad \text{almost everywhere},
\]

\[
\sqrt{N}(z_N - z_0) \Rightarrow (L_K)^{-1}(\mathcal{N}(0, \Sigma_0)),
\]

\[
\sqrt{NL_K}(z_N - z_0) \Rightarrow \mathcal{N}(0, \Sigma_0)
\]

(16)
and
\[ \sqrt{N}\{\Pi_S(z_N) - \Pi_S(z_0)\} \Rightarrow \Pi_K \circ (L_K)^{-1}(N(0, \Sigma_0)). \] (18)

Suppose in addition that assumption 3, parts (a) and (b), holds. Then there are positive real numbers \( \epsilon_0, \delta_0, \mu_0, M_0 \) and \( \sigma_0 \) such that the following inequality holds for each \( \epsilon \in (0, \epsilon_0] \) and each \( N \):

\[
\text{Prob}\{\|\{\hat{\beta}_0, \hat{\beta}, \hat{\iota}\} - \{\tilde{\beta}_0, \tilde{\beta}, \tilde{\iota}\}\| < \epsilon\} \geq \text{Prob}\{\|z_N - z_0\| < \epsilon\}
\]
\[
\geq 1 - \delta_0 \exp(-N\mu_0) - \frac{M_0}{\epsilon^{p+1}} \exp\left(-\frac{N\epsilon^2}{\sigma_0}\right).
\] (19)

In theorem 1, \( L_K \) is the normal map induced by the linear function \( L \) in expression (6) and the critical cone \( K \) in equation (11). Since \( K \) is a polyhedral convex cone, the Euclidean projector \( \Pi_K \) is a piecewise linear function (a function that coincides with a linear function on each of finitely many polyhedral convex cones whose union is the entire space). The normal map \( L_K \) is therefore a piecewise linear function as well. By lemma 2, \( L_K \) is a global homeomorphism under assumption 1, part (a), and assumption 2. The inverse function \( (L_K)^{-1} \) is again a piecewise linear function. Equations (17) and (18) give the asymptotic distributions of \( z_N \) and \( \Pi_S(z_N) = (\tilde{\beta}_0, \tilde{\beta}, \tilde{\iota}) \) respectively. Equation (19) says that \( z_N \) converges to \( z_0 \) in probability at an exponential rate.

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

\[ z_N = (\hat{\beta}_0, \hat{\beta}, \hat{\iota}) - f_N(\hat{\beta}_0, \hat{\beta}, \hat{\iota}), \] (20)

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

We shall estimate \( \Sigma_0 \) by \( \Sigma_N \), the sample covariance matrix of \( \{F(\hat{\beta}_0, \hat{\beta}, \hat{\iota}, x_i, y_i)\}_{i=1}^N \). Let \( \Sigma_N^1 \) be the upper left \( (p + 1) \times (p + 1) \) submatrix of \( \Sigma_N \); we have

\[ \Sigma_N = \begin{pmatrix} \Sigma_N^1 & 0 \\ 0 & 0 \end{pmatrix}. \]

The following lemma shows that \( \Sigma_N \) converges to \( \Sigma_0 \) almost surely and provides the convergence rate of \( \Sigma_N \).

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

\[
\text{Prob}\{\|\Sigma_N - \Sigma_0\| < \epsilon\} \geq 1 - \delta_1 \exp(-N\mu_1) - \frac{M_1}{\min\{\epsilon^{2(p+1)}, \epsilon^{2p+1}\}} \exp\left(-\frac{N\epsilon^2}{\sigma_1}\right).
\] (21)

Estimation of the normal map \( L_K \) requires more understanding of its structure. By Robinson (1995), \( L_K \) is exactly \( d(f_0)_S(z_0) \), the \( B \)-derivative (defined at the end of Section 1) of the normal map \( (f_0)_S \) at \( z_0 \). With \( L = d f_0(x_0) \) defined in expression (6), the chain rule of \( B \)-differentiability gives

\[ L_K(h) = d(f_0)_S(z_0)(h) = L d\Pi_S(z_0)(h) + h - d\Pi_S(z_0)(h) \]

for each \( h \in \mathbb{R}^{2p+1} \),

where \( d\Pi_S(z_0) \) is the \( B \)-derivative of the Euclidean projector \( \Pi_S \) at \( z_0 \) and satisfies \( d\Pi_S(z_0) = \Pi_K \). Depending on the structure of \( K \), there are two different cases.
a) Case I (the linear case): when $z_0$ belongs to the interior of some $(2p+1)$-cell in the normal manifold of $S$ (so that $\Pi_S$ coincides with an affine function in a neighbourhood of $z_0$), $K$ is a subspace of $\mathbb{R}^{2p+1}$. In this case, both $\Pi_K = \mathbb{d} \Pi_S(z_0)$ and $L_K = \mathbb{d} (f_0)(z_0)$ are linear functions, and $\sqrt{N}(z_N - z_0)$ asymptotically follows a normal distribution based on expression (17). Note that $d(f_N)(z_N)(h) = L_N \mathbb{d} I_S(z_N)(h) + h - \mathbb{d} I_S(z_N)(h)$ for each $h \in \mathbb{R}^{2p+1}$. Since $z_N$ converges to $z_0$ almost surely and $\Pi_S$ coincides with an affine function around $z_0$, $\mathbb{d} I_S(z_N)$ and $d(f_N)(z_N)$ almost surely converge to $\mathbb{d} I_S(z_0)$ and $L_K$ respectively, and we can use $\mathbb{d} I_S(z_N)$ and $d(f_N)(z_N)$ as estimators for $\mathbb{d} I_S(z_0)$ and $L_K$ respectively.

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

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

$$\Phi_N(z)(h) = L_N \Lambda_N(z)(h) + h - \Lambda_N(z)(h)$$

for each $z \in \mathbb{R}^{2p+1}$ and $h \in \mathbb{R}^{2p+1}$, where $L_N$ is defined in expression (14). For a given $N$, $\Lambda_N$ is a fixed function, whereas $\Phi_N$ depends on random samples since $L_N$ does. Theorem 2 below shows that $\Lambda_N(z_N)$ and $\Phi_N(z_N)$ are asymptotically exact estimators of $\mathbb{d} I_S(z_0)$ and $L_K$ respectively.

**Theorem 2.** Suppose that assumptions 1 and 2 and assumption 3, parts (a) and (b), hold. Then

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

and there is a positive real number $\phi$ such that

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

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

**Theorem 3.** Suppose that assumptions 1 and 2 and assumption 3, parts (a) and (b), hold. Then

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

If $\Sigma_0^1$ is non-singular, then

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

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

$$\Sigma_0^1 = U_N^T \Delta_N U_N,$$  (26)
where \( U_N \) is an orthogonal \((p + 1) \times (p + 1)\) matrix, and \( \Delta_N \) is a diagonal matrix with monotonically decreasing elements. Let \( D_N \) be the upper left submatrix of \( \Delta_N \) whose diagonal elements are at least \( 1/g(N) \). Let \( l_N \) be the number of rows in \( D_N \), and let \((U_N)_1\) be the submatrix of \( U_N \) that consists of its first \( l_N \) rows, and let \((U_N)_2\) consist of the remaining rows of \( U_N \). Then \( \text{Prob}(l_N = l) \to 1 \) as \( N \to \infty \), with

\[
N((\Phi_N(z_N))(z_N - z_0))^T \begin{pmatrix} (U_N)_1^T D_N^{-1} (U_N)_1 & 0 \\ 0 & 0 \end{pmatrix} ((\Phi_N(z_N))(z_N - z_0)) \Rightarrow \chi^2_{l}
\]

and

\[
N((\Phi_N(z_N))(z_N - z_0))^T \begin{pmatrix} (U_N)_2^T (U_N)_2 & 0 \\ 0 & I_p \end{pmatrix} ((\Phi_N(z_N))(z_N - z_0)) \Rightarrow 0.
\]

Theorem 3 deals with two cases separately, depending on whether \( \Sigma_0^{-1} \) is non-singular or not. In practice, since \( \Sigma_0^{-1} \) is unknown, we shall always start by decomposing \( \Sigma_N^{-1} \) as in equation (26). If some eigenvalues of \( \Sigma_N^{-1} \) (i.e. diagonal elements of \( \Delta_N \)) are less than \( 1/g(N) \), then \( D_N \) is a proper submatrix of \( \Delta_N \), and we shall use expressions (27) and (28) to establish confidence intervals for \( z_0 \) (more details will be given in the following subsections). Otherwise, if all eigenvalues of \( \Sigma_N^{-1} \) are greater than or equal to \( 1/g(N) \), then \( D_N = \Delta_N \) and expressions (27) and (28) are equivalent to expression (25).

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

### 3.2. Computation of confidence intervals

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

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

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

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

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

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

Simultaneous confidence intervals for all components of \((z_0)_i\) can be computed on the basis of expressions (27) and (28), by using the \( \chi^2 \)-distribution to obtain a confidence region and then finding the minimal bounding box of that region. For details and for an alternative approach, see appendix C.5.
After computing confidence intervals for $z_0$, we transform them into confidence intervals for the population lasso parameters $(\tilde{\beta}_0, \tilde{\beta})$, on the basis of the relationship between $z_0$ and $(\tilde{\beta}_0, \tilde{\beta})$ derived in appendix C.6. It is shown there that $\tilde{\beta}_0 = (z_0)_{1}$, so confidence intervals of $(z_0)_{1}$ are exactly those of $\tilde{\beta}_0$. For each $i = 1, \ldots, p$, $\tilde{\beta}_i$ and $(z_0)_{i+1}$ satisfy
\[
\tilde{\beta}_i = \begin{cases} 
(z_0)_{i+1} - \lambda & \text{if } (z_0)_{i+1} > \lambda, \\
0 & \text{if } (z_0)_{i+1} \in [-\lambda, \lambda], \\
(z_0)_{i+1} + \lambda & \text{if } (z_0)_{i+1} < -\lambda.
\end{cases}
\] (30)

Denote the right-hand side of expression (30) as $\Gamma\{(z_0)_{i+1}\}$, which is a non-decreasing piecewise linear function of $(z_0)_{i+1}$. We can then use images of confidence intervals of $(z_0)_{i+1}$ under the map $\Gamma$ as confidence intervals of $\tilde{\beta}_i$. Note that $\Gamma(\cdot)$ takes the constant value of 0 on $[-\lambda, \lambda]$. As a result, when the confidence interval for $(z_0)_{i+1}$ is contained entirely in $[-\lambda, \lambda]$, the confidence interval for $\tilde{\beta}_i$ is a singleton $\{0\}$. When the confidence interval for $(z_0)_{i+1}$ intersects with a part of $[-\lambda, \lambda]$, the confidence interval for $\tilde{\beta}_i$ will contain the true solution of problem (2) with a probability that is larger than the prescribed level.

4. Inference for true parameters in a linear model

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

Suppose that the true linear model between $X$ and $Y$ is
\[
Y = \beta_0^{\text{true}} + X^T \beta^{\text{true}} + \varepsilon,
\] (31)
where $\beta_0^{\text{true}} \in \mathbb{R}$ and $\beta^{\text{true}} = (\beta_1^{\text{true}}, \ldots, \beta_p^{\text{true}}) \in \mathbb{R}^p$ are the true parameters. The random error $\varepsilon$ has mean 0 and variance $\sigma^2 \Sigma^{-1}$. Moreover, $\varepsilon$ is independent of $X_i$ for each $i = 1, \ldots, p$. In this section, we assume that $E[X_i] = 0$ for each $i = 1, \ldots, p$; hence $E[Y] = \beta_0^{\text{true}}$. Denote the covariance matrix of $X$ as $\Sigma$, i.e. $\Sigma = E[XX^T]$. If $\Sigma$ is non-singular, by plugging equation (31) into expression (10), with the fact that $\tilde{\beta}_0 = (z_0)_{1}$, we obtain
\[
\beta_0^{\text{true}} = (z_0)_{1},
\]
\[
\beta^{\text{true}} = \frac{1}{2} \Sigma^{-1}(z_0)_{2:(p+1)} + (I_p - \frac{1}{2} \Sigma^{-1})\tilde{\beta},
\] (32)
where $(z_0)_{2:(p+1)}$ denotes the vector that consists of the second to $(p + 1)$th entries of $z_0$. Expression (32) suggests the estimators
\[
\tilde{\beta}_0^{\text{true}} = (z_N)_{1},
\]
\[
\beta^{\text{true}} = \frac{1}{2} \hat{\Theta}(z_N)_{2:(p+1)} + (I_p - \frac{1}{2} \hat{\Theta})\tilde{\beta},
\] (33)
where $\hat{\Theta}$ is an estimator of the precision matrix $\Sigma^{-1}$. From expressions (20) and (33) we note that the estimator $(\beta_0^{\text{true}}, \beta^{\text{true}})$ can be expressed as the sum of the lasso estimator $\tilde{\beta}$ and a bias correction term, which is similar to the debiased estimators in Zhang and Zhang (2014) and Van de Geer et al. (2014). However, our inference method is different. Our approach is based on analysis of the transformed problems and is developed under different assumptions that allow the asymptotic distribution considered to be non-normal. In the simulation, Van de Geer et al. (2014) used the scaled lasso to obtain the lasso estimate $\tilde{\beta}$ and the estimate $\sigma^2$ of the error variance, whereas we use the original lasso to calculate $\tilde{\beta}$ with a tuned parameter $\lambda$. Numerical results show that our method is competitive with existing methods and performs better in certain cases especially for high dimensions (see Table 4 in Section 5.3).
Next, let a matrix $B \in \mathbb{R}^{(p+1) \times (2p+1)}$ be given by $B = (I_{p+1} \ 0)$, and define two functions $G$ and $\hat{G}$ from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{p+1}$ as

$$G = \frac{1}{2} \left\{ \left( \begin{array}{ccc} 1 & 0 & 0 \\ 0 & \Sigma^{-1} & \end{array} \right) B + \left( \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 2I - \Sigma^{-1} & \end{array} \right) B \circ \Pi_K \right\}. \quad (34)$$

$$\hat{G} = \frac{1}{2} \left\{ \left( \begin{array}{ccc} 1 & 0 & 0 \\ 0 & \hat{\Sigma} & \end{array} \right) B + \left( \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 2I - \hat{\Sigma} & \end{array} \right) B \circ \text{d}\Pi_S(zN) \right\}. \quad (35)$$

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

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

$$\sqrt{N} \{ (\beta_0^{\text{true}}, \beta^{\text{true}}) - (\beta_0^{\text{true}}, \beta^{\text{true}}) \} \Rightarrow G \circ (L_K)^{-1}(\mathcal{N}(0, \Sigma_0)). \quad (36)$$

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

**Theorem 5.** Suppose that assumptions 1 and 2 hold, the true covariance matrix $\Sigma$ is nonsingular and the solution to the normal map formulation (9) satisfies conditions for case I. Let $\hat{\Theta}$ be a root-$N$-consistent estimator of $\Sigma^{-1}$, and define $H = G(L_K)^{-1}$ and $H_N = \hat{G}\{d(f_N) \Sigma(zN)^{-1}\}$. For each $i = 0, 1, \ldots, p$, if $(H \Sigma_0 H^T)^{-1}_{i+1,i+1} \neq 0$ then

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

Theorem 5 suggests constructing an asymptotically exact individual confidence interval for $\beta_i^{\text{true}}$ with level of significance $\alpha$ as $[\hat{\beta}_i^{\text{true}} - N^{-1/2} \beta_i^{\text{true}} + N^{-1/2} \sqrt{\chi_1^2(\alpha) \bar{m}_i}]$, where $\bar{m}_i$ is the $(i + 1)$th diagonal element of the matrix $H_N \Sigma_N H_N^T$.

Next we discuss how to compute individual confidence intervals for $(\beta_0^{\text{true}}, \beta^{\text{true}})$ in case II (the piecewise linear case). Let $f : \mathbb{R}^{2p+1} \rightarrow \mathbb{R}$ be a continuous function and $Z$ be a random variable in $\mathbb{R}^{2p+1}$ with $Z \sim \mathcal{N}(0, I_{p+1}) \times \hat{\Theta}$. Let $\alpha \in (0, 1)$, and define $a^*(f) \in (0, \infty)$ as

$$a^*(f) = \inf\{c \geq 0 \mid \text{Prob}\{-c \leq f(Z) - r \leq c \} \geq 1 - \alpha\}. \quad (38)$$

Suppose that $\text{Prob}\{f(Z) = b\} = 0$ for all $b \in \mathbb{R}$. Then, for any given $r \in \mathbb{R}$ and $\alpha \in (0, 1)$, $a^*(f)$ as
defined in equation (38) is the smallest value that satisfies $\text{Prob}\{-a^r(f) \leq f(Z) - r \leq a^r(f)\} = 1 - \alpha$. Define two functions $R$ and $\hat{R}$ from $\mathbb{R}^{2p+1}$ to $\mathbb{R}^{p+1}$ as

$$R = G \circ (L_K)^{-1} \left( \begin{array}{cc} (\Sigma_0^1)^{1/2} & 0 \\ 0 & I_p \end{array} \right),$$

$$\hat{R} = \hat{G}' \circ (\Phi_N(z_N))^{-1} \left( \begin{array}{cc} (\Sigma_N^1)^{1/2} & 0 \\ 0 & I_p \end{array} \right),$$

where

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

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

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

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

in which $W$ is a piecewise linear function represented by $p \times p$ diagonal matrices with diagonal elements 0 or $\frac{1}{2}$. If $\Sigma$ and $\Sigma_0^1$ are non-singular, then the matrix representation of each piece of the map $G$ has full row rank. Because $L_K$ is a global homeomorphism under assumption 1, part (a), and assumption 2, it follows that $\text{Prob}\{R_j(Z) = b\} = 0$ for all $b \in \mathbb{R}$. The following theorem gives a way of computing individual confidence intervals for $(\hat{\beta}_0^\text{true}, \beta^\text{true})$.

**Theorem 6.** Suppose that assumptions 1 and 2 and assumption 3, parts (a) and (b), hold, and the covariance matrices $\Sigma$ and $\Sigma_0^1$ are non-singular. Let $\hat{\Theta}$ be a root-$N$-consistent estimator of $\Sigma^{-1}$, $\alpha \in (0, 1)$ and $r \in \mathbb{R}$. Let $a^r(\cdot)$ be as in equation (38), and define $R$ and $\hat{R}$ as in expression (39). Then, for all $j = 0, 1, \ldots, p$, we have

$$\lim_{N \to \infty} \text{Prob}\{\sqrt{N}(\hat{\beta}_j^\text{true} - \beta_j^\text{true}) - r \leq a^r(\hat{R}_{j+1})\} = 1 - \alpha. \quad (42)$$

In practice, for a fixed choice of $r$, we can find individual confidence intervals for $(\hat{\beta}_0^\text{true}, \beta^\text{true})$ by simulating data from $\hat{R}(Z)$. We first generate data from $N(0, \Sigma_N)$, and then compute $(\Phi_N(z_N))^{-1}(q)$ for a given vector $q$ as described in the on-line appendix C.4, to obtain an empirical distribution of $\hat{R}(q) = \hat{G}' \circ (\Phi_N(z_N))^{-1}(q)$ where $\hat{G}'$ is defined in equation (40).

5. **Numerical examples**

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

5.1. **Example 1: asymptotic distribution of lasso solutions**

We generate 400 replications of 2000 observations from the model $Y = \beta^* X + \sigma \epsilon$, where $\beta^* = (2, 1)$, $X$ is a two-dimensional normal random variable with mean 0 and covariance matrix $\Sigma = 0.5I_2 + 0.5J_2$, with $I_2$ being the $2 \times 2$ identity matrix and $J_2$ being the $2 \times 2$ matrix of 1s, $\epsilon \sim \mathcal{N}(0, 1)$ and $\sigma = 3$. Here $X$ and $\epsilon$ are independent of each other.
Fig. 1. Distribution of SAA solutions in example 1: (a) plot of the 400 points $((z_N)_2, (z_N)_3)$, displaying the boundaries of regions defined in expression (F.1) in the on-line appendix with $\alpha = 0.1, 0.2, \ldots, 0.9$ (+, true solution); (b) corresponding 400 points $((\hat{\beta}_1, \hat{\beta}_2)$, displaying boundaries of regions defined in expression (F.2) with $\alpha = 0.1, 0.2, \ldots, 0.9$ (l, markers of the end points of intervals defined in expression (F.3) with the same $\alpha$-values)
In Fig. 1, we illustrate the piecewise Gaussian asymptotic distribution for the SAA solutions with $\lambda = 3$. In Fig. 1(a), the nine boundaries divide the plane into 10 divisions, with around 40 points (minimum 34; maximum 45; mean 40; standard deviation 3.62) in each division. In Fig. 1(b), the markers are not located at intersections between the curves and the horizontal axis, because the two regions (F.2) and (F.3) (in the on-line appendix F.1) come from different distributions. An extra short vertical line is plotted at the true solution $\tilde{\beta}_1$, $\tilde{\beta}_2 = 1$, 0. The 19 short vertical lines on the horizontal axis divide the axis into 20 intervals. There are 208 points out of the total 400 that lie on the horizontal axis, with about 10 points (minimum 5; maximum 18; mean 10.4; standard deviation 3.10) in each interval. The other 192 points lie above the horizontal axis, with about 20 points (minimum 15; maximum 25; mean 19.2; standard deviation 3.01) in each of the 10 divisions divided by the nine curves.

5.2. Example 2: confidence intervals in low dimensional setting

In this example, we simulate data by using the model in example 1 of Tibshirani (1996). The model is the same as that of example 1, with $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)$. Here $X$ is normal with mean 0 and covariance $\Sigma_{ij} = \rho^{|i-j|}$ for $\rho = 0.5$, and $\varepsilon$ is a standard normal random variable independent of $X$. We generate 100 replications of 300 observations with $\sigma = 1$ and compute two types of confidence interval for three fixed $\lambda$-values 0.5, 1 and 2. The first type of confidence intervals is for the population lasso parameters $(\tilde{\beta}_0, \tilde{\beta})$, and the second is for the true parameters $(\beta^\text{true}_0, \beta^\text{true})$ in the underlying linear model (31), both of levels of significance $\alpha = 0.1$. For the second type of intervals, we compare our method with two other approaches in the literature: the low dimensional projection estimator (LDPE) method (Zhang and Zhang, 2014; Van de Geer et al., 2014) and the method that was introduced by Javanmard and Montanari (JM) (2015). Similarly to the LDPE method, we use nodewise lasso regression introduced by Meinshausen and Bühlmann (2006) to compute the estimate of the precision matrix $\hat{\Theta}$. In terms of the tuning parameter $\lambda$, we check the performance of our method by using the generalized information criterion GIC (Nishii, 1984). In the LDPE method, the model parameters are estimated by the scaled lasso without specifying $\lambda$. The JM method uses $\lambda = 4\hat{\sigma}_\varepsilon/\sqrt{\{2\log(p)/n\}}$ as the tuning parameter, where $\hat{\sigma}_\varepsilon$ is the scaled lasso estimator of the level of noise.

For the first type of confidence intervals, we observe 100% coverage of simultaneous confidence intervals from all SAA problems with level of significance 0.1. This is not very surprising, since the boxes that are formed by the simultaneous confidence intervals are much larger than the confidence regions of the specified probability levels enclosed in them. The coverage of individual confidence intervals is larger than 100(1 − $\alpha$)% in general with reasonable lengths (see Tables S.4 and S.5 in the on-line appendix F.2). In Table 1, the intervals are not always symmetric around the estimates, which is a result of the non-normality. The value 0 appears as an end point for many intervals, and in some cases the entire interval shrinks to the singleton {0}.

In Table 2, the confidence intervals for the intercept $\beta_0$ are not available for the LDPE and JM methods. In our method, there is no need to centre each replication. The estimates and individual confidence intervals for true parameters computed from these three methods are quite similar. They also perform similarly and fairly well in terms of the coverage (see Table S.7 in the on-line supporting information).

5.3. Example 3: confidence intervals in high dimensional setting

This example considers a case in which the dimension $p$ is larger than the sample size. The simulation model is the same as that of example 1, with $\beta^*$ being a 300-dimensional vector: $\beta_1^* = 3$, $\beta_2^* = \beta_{100}^* = \beta_{200}^* = \beta_{300}^* = 1.5$, $\beta_5^* = \beta_{95}^* = 2$ and all the other components are 0.
Table 1. 90% confidence intervals for population lasso parameters computed by our method for various $\lambda$-values from a typical replication of 300 observations generated in example 2†

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Results for $\lambda=0.5$</th>
<th>Results for $\lambda=1$</th>
<th>Results for $\lambda=2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Individual confidence interval</td>
<td>Simultaneous confidence interval</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>-0.08</td>
<td>[-0.18, 0.02]</td>
<td>[-0.31, 0.15]</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>2.82</td>
<td>[2.72, 2.93]</td>
<td>[2.58, 3.07]</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>1.43</td>
<td>[1.32, 1.53]</td>
<td>[1.18, 1.67]</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0</td>
<td>[0, 0]</td>
<td>[0, 0.17]</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0</td>
<td>[0, 0]</td>
<td>[0, 0.22]</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>1.74</td>
<td>[1.64, 1.83]</td>
<td>[1.52, 1.96]</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0</td>
<td>[0, 0.08]</td>
<td>[0, 0.31]</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>0</td>
<td>[0, 0.02]</td>
<td>[0, 0.30]</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>0</td>
<td>[0, 0]</td>
<td>[0, 0.17]</td>
</tr>
</tbody>
</table>

†The ‘Estimate’ columns contain values of the SAA solution $(\hat{\beta}_0, \hat{\beta})$. 
Table 2. 90% individual confidence intervals and coverage for true parameters in model (31) computed by various methods from a typical replication of 300 observations generated in example 2†

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Results for LDPE method</th>
<th></th>
<th></th>
<th>Results for JM method</th>
<th></th>
<th></th>
<th>Results for $\lambda = 0.49$ tuned by GIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Estimate</td>
<td>Individual confidence interval</td>
<td>Coverage probability</td>
<td>Estimate</td>
<td>Individual confidence interval</td>
<td>Coverage probability</td>
<td>Estimate</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>3</td>
<td>3.00</td>
<td>[2.90, 3.11]</td>
<td>92</td>
<td>3.00</td>
<td>[2.91, 3.10]</td>
<td>90</td>
<td>3.00</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>1.5</td>
<td>1.59</td>
<td>[1.48, 1.70]</td>
<td>92</td>
<td>1.59</td>
<td>[1.49, 1.69]</td>
<td>85</td>
<td>1.59</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0</td>
<td>-0.06</td>
<td>[-0.18, 0.05]</td>
<td>88</td>
<td>-0.08</td>
<td>[-0.17, 0.02]</td>
<td>92</td>
<td>-0.06</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>0</td>
<td>0.05</td>
<td>[-0.06, 0.17]</td>
<td>87</td>
<td>0.06</td>
<td>[-0.04, 0.16]</td>
<td>96</td>
<td>0.05</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>2</td>
<td>1.91</td>
<td>[1.79, 2.02]</td>
<td>90</td>
<td>1.91</td>
<td>[1.81, 2.00]</td>
<td>90</td>
<td>1.91</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>0</td>
<td>0.08</td>
<td>[-0.03, 0.20]</td>
<td>85</td>
<td>0.08</td>
<td>[-0.02, 0.18]</td>
<td>85</td>
<td>0.08</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>0</td>
<td>0.03</td>
<td>[-0.09, 0.14]</td>
<td>90</td>
<td>0.01</td>
<td>[-0.09, 0.11]</td>
<td>95</td>
<td>0.03</td>
</tr>
</tbody>
</table>

†The ‘Estimate’ columns contain the true parameter estimates ($\hat{\beta}_0, \hat{\beta}_1, \ldots$), and the ‘True’ column contains the true parameter ($\beta_0^{true}, \beta_1^{true}, \ldots$). The empirical coverage probabilities with significance level 0.1 are collected in the ‘Coverage probability’ columns.
Table 3. Average coverage and length of 95% individual confidence intervals for the population lasso parameters computed by our method for various $\lambda$-values from 100 replications of 100 observations and dimension $p = 300$ generated in example 3

<table>
<thead>
<tr>
<th>Results for $\lambda = 0.5$</th>
<th>Results for $\lambda = 1$</th>
<th>Results for $\lambda = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ Avgcov (%)</td>
<td>Medcov (%)</td>
<td>Avglen (%)</td>
</tr>
<tr>
<td>91.86</td>
<td>94.00</td>
<td>0.92</td>
</tr>
<tr>
<td>$A^c$ Avgcov (%)</td>
<td>Medcov (%)</td>
<td>Avglen (%)</td>
</tr>
<tr>
<td>99.92</td>
<td>100.00</td>
<td>0.07</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Active set</th>
<th>Results for our method, $\lambda = 0.5$</th>
<th>Results for our method, $\lambda = 1$</th>
<th>Results for our method, $\lambda = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ Avgcov (%)</td>
<td>Medcov (%)</td>
<td>Avglen (%)</td>
<td>Medlen (%)</td>
</tr>
<tr>
<td>93.86</td>
<td>94.00</td>
<td>0.81</td>
<td>1.03</td>
</tr>
<tr>
<td>$A^c$ Avgcov (%)</td>
<td>Medcov (%)</td>
<td>Avglen (%)</td>
<td>Medlen (%)</td>
</tr>
<tr>
<td>92.85</td>
<td>93.00</td>
<td>0.75</td>
<td>0.74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results for LDPE method</th>
<th>Results for JM method</th>
<th>Results for our method with GIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ Avgcov (%)</td>
<td>Medcov (%)</td>
<td>Avglen (%)</td>
</tr>
<tr>
<td>88.43</td>
<td>89.00</td>
<td>1.04</td>
</tr>
<tr>
<td>$A^c$ Avgcov (%)</td>
<td>Medcov (%)</td>
<td>Avglen (%)</td>
</tr>
<tr>
<td>95.13</td>
<td>95.00</td>
<td>1.07</td>
</tr>
</tbody>
</table>

Again $X$ is normal with mean 0 and covariance $\Sigma_{ij} = \rho^{|i-j|}$ for $\rho = 0.9$, $\epsilon$ is standard normal and independent of $X$, and $\sigma = 1$. We generate 100 replications of 100 observations and consider three fixed $\lambda$-values, 0.5, 1 and 2, as well as the $\lambda$-value that was chosen by GIC for each SAA problem. As in example 2, we compute the two types of individual confidence intervals both with the level of significance 0.05. Define the active set as $A = \{j : \beta_j^* \neq 0\} = \{1, 2, 5, 95, 100, 200, 300\}$ and $A^c = \{1, 2, \ldots, p\} \setminus A$. For each type of individual confidence intervals, we report the average coverage Avgcov, median coverage Medcov, average length Avglen and median length Medlen of the individual confidence intervals corresponding to parameters in either $A$ or $A^c$: $\text{Avgcov}(A) = |A|^{-1} \sum_{j \in A} CP_j$, $\text{Avgcov}(A^c) = |A^c|^{-1} \sum_{j \in A^c} CP_j$, $\text{Avglen}(A) = |A|^{-1} \sum_{j \in A} \text{ALen}_j$, $\text{Avglen}(A^c) = |A^c|^{-1} \sum_{j \in A^c} \text{ALen}_j$, $\text{Medcov}(A) = \text{median}_{j \in A}(CP_j)$, $\text{Medcov}(A^c) = \text{median}_{j \in A^c}(CP_j)$, $\text{Medlen}(A) = \text{median}_{j \in A}(\text{ALen}_j)$ and $\text{Medlen}(A^c) = \text{median}_{j \in A^c}(\text{ALen}_j)$, where $CP_j$ and $\text{ALen}_j$ respectively represent the empirical coverage probability and average interval length of the confidence intervals for $\beta_j$ among the 100 replications (Tables 3 and 4). For the second type of intervals, we compare the above measures computed from our method with those from the LDPE and JM methods (Table 4).

As shown in Table 3, the first type of confidence intervals are often conservative for the inactive variables. The same phenomena are observed in example 2. In contrast, the interval lengths for the inactive variables are very short compared with the lengths for active variables.
Table 5. 95% individual confidence intervals for true parameters in model (31) computed by various methods from the prostate cancer data with sample size $N = 67$ in example 4†

<table>
<thead>
<tr>
<th></th>
<th>Results for LDPE method</th>
<th></th>
<th>Results for JM method</th>
<th></th>
<th>Results for $\lambda = 0.88$ tuned by GIC</th>
<th></th>
<th>Results for $\lambda = 0.45$</th>
<th></th>
<th>Results for $\lambda = 1.49$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Individual confidence interval</td>
<td>Estimate</td>
<td>Individual confidence interval</td>
<td>Estimate</td>
<td>Individual confidence interval</td>
<td>Estimate</td>
<td>Individual confidence interval</td>
<td>Estimate</td>
<td>Individual confidence interval</td>
</tr>
<tr>
<td>$\beta_{true}^1$</td>
<td>0.69</td>
<td>[0.46, 0.93]</td>
<td>0.68</td>
<td>[0.03, 1.33]</td>
<td>0.71</td>
<td>[0.41, 1.01]</td>
<td>0.70</td>
<td>[0.44, 0.95]</td>
<td>0.74</td>
<td>[0.37, 1.11]</td>
</tr>
<tr>
<td>$\beta_{true}^2$</td>
<td>0.28</td>
<td>[0.09, 0.46]</td>
<td>0.26</td>
<td>[−0.22, 0.75]</td>
<td>0.29</td>
<td>[0.08, 0.49]</td>
<td>0.28</td>
<td>[0.10, 0.46]</td>
<td>0.32</td>
<td>[0.09, 0.55]</td>
</tr>
<tr>
<td>$\beta_{true}^3$</td>
<td>−0.09</td>
<td>[−0.29, 0.11]</td>
<td>−0.14</td>
<td>[−0.66, 0.38]</td>
<td>−0.08</td>
<td>[−0.34, 0.17]</td>
<td>−0.09</td>
<td>[−0.29, 0.10]</td>
<td>−0.02</td>
<td>[−0.35, 0.31]</td>
</tr>
<tr>
<td>$\beta_{true}^4$</td>
<td>0.21</td>
<td>[0.01, 0.41]</td>
<td>0.21</td>
<td>[−0.31, 0.73]</td>
<td>0.22</td>
<td>[−0.02, 0.45]</td>
<td>0.21</td>
<td>[−0.00, 0.42]</td>
<td>0.22</td>
<td>[−0.05, 0.49]</td>
</tr>
<tr>
<td>$\beta_{true}^5$</td>
<td>0.31</td>
<td>[0.08, 0.54]</td>
<td>0.31</td>
<td>[−0.33, 0.94]</td>
<td>0.33</td>
<td>[0.04, 0.63]</td>
<td>0.31</td>
<td>[0.04, 0.58]</td>
<td>0.38</td>
<td>[0.05, 0.71]</td>
</tr>
<tr>
<td>$\beta_{true}^6$</td>
<td>−0.21</td>
<td>[−0.48, 0.06]</td>
<td>−0.29</td>
<td>[−1.08, 0.50]</td>
<td>−0.19</td>
<td>[−0.47, 0.09]</td>
<td>−0.21</td>
<td>[−0.45, 0.04]</td>
<td>−0.10</td>
<td>[−0.41, 0.21]</td>
</tr>
<tr>
<td>$\beta_{true}^7$</td>
<td>−0.01</td>
<td>[−0.27, 0.25]</td>
<td>−0.02</td>
<td>[−0.76, 0.72]</td>
<td>−0.02</td>
<td>[−0.28, 0.24]</td>
<td>−0.01</td>
<td>[−0.24, 0.22]</td>
<td>0.04</td>
<td>[−0.25, 0.32]</td>
</tr>
<tr>
<td>$\beta_{true}^8$</td>
<td>0.24</td>
<td>[−0.03, 0.51]</td>
<td>0.27</td>
<td>[−0.52, 1.05]</td>
<td>0.25</td>
<td>[−0.06, 0.56]</td>
<td>0.24</td>
<td>[−0.01, 0.48]</td>
<td>0.28</td>
<td>[−0.05, 0.61]</td>
</tr>
</tbody>
</table>

†There are eight covariates: log(cancer volume), log(prostate weight), age, log(amount of benign prostatic hyperplasia), seminal vesicle invasion, log(capsular penetration), Gleason score and percentage of Gleason scores 4 or 5. The parameters corresponding to these covariates are denoted by $\beta_{true}^1, \beta_{true}^2, \ldots, \beta_{true}^8$. 
This is related to the nature of the lasso: for a large $\lambda$, many population lasso parameters are exactly 0s. Thus the SAA solutions of the lasso of these parameters concentrate closely around 0s. Similarly, we also observe shorter confidence intervals for true parameters of the inactive set compared with those for true parameters of the active set, as shown in Table 4.

The top rows of Table 4 report results of our method with various $\lambda$-values. One may note that the length of confidence intervals for the true parameters $(\hat{\beta}_0^{\text{true}}, \hat{\beta}_1^{\text{true}}, \ldots)$ increases when $\lambda$ increases. For an intuitive explanation, recall that the estimator $(\hat{\beta}_0^{\text{true}}, \hat{\beta}_1^{\text{true}}, \ldots)$ in expression (33) is a bias correction version of the lasso solution $(\hat{\beta}_0, \hat{\beta}, \ldots)$. Large $\lambda$ brings the lasso solution close to 0, which causes an increase in the correction part, and this leads to wide confidence intervals. In contrast, if $\lambda$ is too small, the SAA solution lacks sparsity and the corresponding lasso estimates are less reliable. This suggests choosing an intermediate value of $\lambda$ to achieve the best overall performance.

The bottom rows of Table 4 show the results that were calculated from the LDPE method, the JM method and our method (with $\lambda$ chosen by GIC). For the active variables, our method performs considerably better than the other two. For the inactive variables, the coverage from the LDPE method is close to 95%, and the coverage from the JM method is even higher. However, their confidence intervals are comparatively wider than those from our method on average. The coverage for inactive variables computed from our method is in line with the coverage for active variables, and they both become better with larger sample sizes.

5.4. Example 4: prostate cancer data
This subsection considers the prostate cancer example that was used in Hastie et al. (2001). We standardize the data and split observations into two parts. One part consists of 67 observations, which are the training set in Hastie et al. (2001). We use only these 67 observations in our computation. For the true model parameters, we compare the confidence intervals that were computed from our method with those from the LDPE and JM methods, as shown in Table 5. The results for the population lasso parameters can be found in the on-line Appendix F.3.

Table 5 lists the estimates of the true model parameters and their individual confidence intervals, computed from our methods as well as the LDPE and JM methods with $\lambda=0.88$ (tuned by GIC as in example 2), 0.45 and 1.49. The estimate of the precision matrix $\hat{\Theta}$ is computed by the nodewise lasso except for the JM method (the JM method uses its own procedure). Results from the three methods are generally comparable, except that confidence intervals that were computed from the JM method are overall wider than those from our method on average. The coverage for inactive variables computed from our method is in line with the coverage for active variables, and they both become better with larger sample sizes.

6. Discussion
In this paper, we transform lasso problems into variational inequalities and make use of the asymptotic convergence results to derive confidence intervals and regions for the population lasso parameters and for the true parameters. Both our theoretical and our numerical results confirm the validity and effectiveness of the methods proposed. In view of expression (29), the lengths of confidence intervals for population lasso parameters are affected by two factors. The first is $\Sigma_N$, the sample covariance of $F\{\hat{\beta}_0, \hat{\beta}, \hat{t}, x_i, y_i\}_{i=1}^N$. The second is $(\Phi_N(z_N))^{-1}$, which
characterizes the sensitivity of the solution to problem (1) with respect to random samples. In general, large variance and high sensitivity lead to wide confidence intervals, and small variance and low sensitivity lead to short intervals. Thus, the lengths of confidence intervals for population lasso parameters reflect the effect of sample variance on the parameter estimates that are computed from the lasso. With respect to confidence intervals for the true regression parameters, our technique performs competitively compared with existing methods. Although our asymptotic results are developed with $N \to \infty$ and fixed dimensions, our technique appears to perform reasonably well for moderately high dimensional examples. One possible research direction is to explore the high dimensional asymptotic properties by extending the technique proposed. Another research direction is to extend the technique to general sparse penalized methods.

Acknowledgements

The authors thank the Joint Editor, the Associate Editor and the reviewers for their constructive comments. This research was partly supported by US National Science Foundation grants DMS-1407241 (Liu, Lu and Yin), DMS-1109099 (Lu and Yin), DMS-1309619 (Zhang), DMS-1127914 (Zhang as a Faculty Fellow at the Statistical and Applied Mathematical Sciences Institute) and US National Institutes of Health grants R01 CA-149569 and P01 CA-142538.

References


*Supporting information*

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Web-based supporting materials for Confidence intervals and regions for the LASSO using stochastic variational inequality techniques in optimization’.