Journal of the
Royal Statistical Society

SERIES B
Statistical
Methodology

J. R. Statist. Soc. B (2014)
76, *Part* 1, *pp.* 217–242

Confidence intervals for low dimensional parameters in high dimensional linear models

Cun-Hui Zhang

Rutgers University, Piscataway, USA

and Stephanie S. Zhang

Columbia University, New York, USA

[Received October 2011. Final revision January 2013]

Summary. The purpose of this paper is to propose methodologies for statistical inference of low dimensional parameters with high dimensional data. We focus on constructing confidence intervals for individual coefficients and linear combinations of several of them in a linear regression model, although our ideas are applicable in a much broader context. The theoretical results that are presented provide sufficient conditions for the asymptotic normality of the proposed estimators along with a consistent estimator for their finite dimensional covariance matrices. These sufficient conditions allow the number of variables to exceed the sample size and the presence of many small non-zero coefficients. Our methods and theory apply to interval estimation of a preconceived regression coefficient or contrast as well as simultaneous interval estimation of many regression coefficients. Moreover, the method proposed turns the regression data into an approximate Gaussian sequence of point estimators of individual regression coefficients, which can be used to select variables after proper thresholding. The simulation results that are presented demonstrate the accuracy of the coverage probability of the confidence intervals proposed as well as other desirable properties, strongly supporting the theoretical results.

Keywords: Confidence interval; High dimension; Linear regression model; *p*-value; Statistical inference

1. Introduction

The area of high dimensional data is an area of intense research in statistics and machine learning, owing to the rapid development of information technologies and their applications in scientific experiments and everyday life. Numerous large, complex data sets have been collected and are waiting to be analysed; meanwhile, an enormous effort has been mounted to meet this challenge by researchers and practitioners in statistics, computer science and other disciplines. A great number of statistical methods, algorithms and theories have been developed for the prediction and classification of future outcomes, the estimation of high dimensional objects and the selection of important variables or features for further scientific experiments and engineering applications. However, statistical inference with high dimensional data is still largely untouched, owing to the complexity of the sampling distributions of existing estimators. This is particularly

Address for correspondence: Cun Hui Zhang, Department of Statistics and Biostatistics, Hill Center, Busch Campus, Rutgers University, Piscataway, NJ 08854, USA.
E-mail: czhang@stat.rutgers.edu

1369–7412/14/76217

so in the context of the so-called large p smaller n problem, where the dimension of the data p is greater than the sample size n.

Regularized linear regression is one of the best understood statistical problems in high dimensional data. Important work has been done in problem formulation, methodology and algorithm development, and theoretical analysis under sparsity assumptions on the regression coefficients. This includes l_1 regularized methods (Tibshirani, 1996; Chen *et al.*, 2001; Greenshtein and Ritov, 2004; Greenshtein, 2006; Meinshausen and Bühlmann, 2006; Tropp, 2006; Zhao and Yu, 2006; Candès and Tao, 2007; Zhang and Huang, 2008; Bickel *et al.*, 2009; Koltchinskii, 2009; Meinshausen and Yu, 2009; van de Geer and Bühlmann, 2009; Wainwright, 2009a; Zhang, 2009; Ye and Zhang, 2010; Koltchinskii *et al.*, 2011; Sun and Zhang, 2010), non-convex penalized methods (Frank and Friedman, 1993; Fan and Li, 2001; Fan and Peng, 2004; Kim *et al.*, 2008; Zhang, 2010; Zhang and Zhang, 2012), greedy methods (Zhang, 2011a), adaptive methods (Zou, 2006; Huang *et al.*, 2008; Zhang, 2011b; Zhang and Zhang, 2012), screening methods (Fan and Lv, 2008), and more. For further discussion, we refer to related sections in Bühlmann and van de Geer (2011) and recent reviews in Fan and Lv (2010) and Zhang and Zhang (2012).

Among existing results, variable selection consistency is most relevant to statistical inference. In an l_0 sparse setting, an estimator is variable selection consistent if it selects the oracle model composed of exactly the set of variables with non-zero regression coefficients. In the large p smaller n setting, variable selection consistency has been established under incoherence and other l_∞-type conditions on the design matrix for the lasso (Meinshausen and Bühlmann, 2006; Tropp, 2006; Zhao and Yu, 2006; Wainwright, 2009a), and under sparse eigenvalue or l_2-type conditions for non-convex methods (Zhang, C. H., 2010; Zhang, T., 2011a, b; Zhang and Zhang, 2012). Another approach in variable selection with high dimensional data involves subsampling or randomization, including notably the stability selection method (Meinshausen and Bühlmann, 2010). Since the oracle model is typically assumed to be of smaller order in dimension than the sample size n in selection consistency theory, consistent variable selection allows a great reduction of the complexity of the analysis from a large p smaller n problem to a problem involving the oracle set of variables only. Consequently, taking the least squares estimator on the selected set of variables if necessary, statistical inference can be justified in the smaller oracle model.

However, statistical inference based on selection consistency theory typically requires a *uniform signal strength condition* that all non-zero regression coefficients be greater in magnitude than an inflated level of noise to take model uncertainty into account. This inflated level of noise can be written as $C\sigma\sqrt{\{(2/n)\log(p)\}}$, where σ is the level of noise. It follows from Fano's lemma (Fano, 1961) that $C \geqslant \frac{1}{2}$ is required for variable selection consistency with a general standardized design matrix (Wainwright, 2009b; Zhang, 2010). This uniform signal strength condition is, unfortunately, seldom supported by either the data or the underlying science in applications when the presence of weak signals cannot be ruled out. In such cases, consistent estimation of the distribution of the least squares estimator after model selection is impossible (Leeb and Potscher, 2006). Conservative statistical inference after model selection or classification has been considered in Berk *et al.* (2010) and Laber and Murphy (2011). However, such conservative methods may not yield sufficiently accurate confidence regions or p-values for common applications with a large number of variables.

We propose a low dimensional projection (LDP) approach to constructing confidence intervals for regression coefficients without assuming the uniform signal strength condition. We provide theoretical justifications for the use of the proposed confidence interval for a preconceived regression coefficient or a contrast depending on a small number of regression coefficients.

We believe that, in the presence of potentially many non-zero coefficients of small or moderate magnitude, construction of a confidence interval for such a preconceived parameter is an important problem in and of itself and was open before our paper (Leeb and Potscher, 2006), but the method proposed is not limited to this application.

Our theoretical work also justifies the use of LDP confidence intervals simultaneously after a multiplicity adjustment, in the absence of a preconceived parameter of interest. Moreover, a thresholded LDP estimator can be used to select variables and to estimate the entire vector of regression coefficients.

The most important difference between the LDP and existing variable selection approaches concerns the requirement of the uniform signal strength condition, as we have mentioned earlier. This is a necessity for the simultaneous correct selection of *all zero and non-zero* coefficients. If this criterion is the goal, we cannot do better than technical improvements over existing methods. However, a main complaint about the variable selection approach is the practicality of the uniform signal strength condition, and the crucial difference between the two approaches is precisely in the case where the condition fails to hold. Without the condition, the LDP approach is still able to select correctly all coefficients above a threshold of order $C\surd\{(2/n)\log(p)\}$ and all zero coefficients. The power of the LDP method is small for testing small non-zero coefficients, but this is unavoidable and does not affect the correct selection of other variables. In this sense, the confidence intervals proposed decompose the variable selection problem into multiple marginal testing problems for individual coefficients as Gaussian means.

Most results presented here are available through Zhang and Zhang (2011). Proofs for the current paper are provided in the on-line supplement.

2. Methodology

We develop methodologies and algorithms for the construction of confidence intervals for the individual regression coefficients and their linear combinations in the linear model

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\varepsilon, \qquad \varepsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I}), \tag{1}$$

where $\mathbf{y}\in\mathbb{R}^n$ is a response vector, $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_p)\in\mathbb{R}^{n\times p}$ is a design matrix with columns $\mathbf{x}_j$ and $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^{\mathrm{T}}$ is a vector of unknown regression coefficients. When $\mathrm{rank}(\mathbf{X})<p$, $\boldsymbol{\beta}$ is unique under proper conditions on the sparsity of $\boldsymbol{\beta}$, but not in general. To simplify the discussion, we standardize the design to $\|\mathbf{x}_j\|_2^2=n$. The design matrix $\mathbf{X}$ is assumed to be deterministic throughout the paper, except in Section 3.5.

The following notation will be used. For real numbers x and y, $x\wedge y=\min(x,y)$, $x\vee y=\max(x,y)$, $x_+=x\vee 0$ and $x_-=(-x)_+$. For vectors $\mathbf{v}=(v_1,\ldots,v_m)$ of any dimension, $\mathrm{supp}(\mathbf{v})=\{j:v_j\neq 0\}$, $\|\mathbf{v}\|_0=|\mathrm{supp}(\mathbf{v})|=\#\{j:v_j\neq 0\}$ and $\|\mathbf{v}\|_q=\{\Sigma_j|v_j|^q)^{1/q}$, with the usual extension to $q=\infty$. For $A\subset\{1,\ldots,p\}$, $\mathbf{v}_A=(v_j,j\in A)^{\mathrm{T}}$ and $\mathbf{X}_A=(\mathbf{x}_k,k\in A)$, including $A=-j=\{1,\ldots,p\}\setminus\{j\}$.

2.1. Bias-corrected linear estimators

In the classical theory of linear models, the least squares estimator of an estimable regression coefficient β_j can be written as

$$\hat{\beta}_j^{(\mathrm{lse})}:=(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{y}/(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{x}_j, \tag{2}$$

where $\mathbf{x}_j^{\perp}$ is the projection of $\mathbf{x}_j$ to the orthogonal complement of the column space of $\mathbf{X}_{-j}=(\mathbf{x}_k,k\neq j)$. Since this is equivalent to solving the equations $(\mathbf{x}_j^{\perp})^{\mathrm{T}}(\mathbf{y}-\beta_j\mathbf{x}_j)$ in the score system

$\mathbf{v} \rightarrow (\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{v}$, $\mathbf{x}_j^{\perp}$ can be viewed as the score vector for the least squares estimation of β_j. The score vector $\mathbf{x}_j^{\perp}$ can be defined by $(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{x}_k = 0\ \forall k \neq j$ and $(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{x}_j = \|\mathbf{x}_j^{\perp}\|_2^2$. For estimable β_j and β_k,

$$\mathrm{cov}(\hat{\beta}_j^{(\mathrm{lse})}, \hat{\beta}_k^{(\mathrm{lse})}) = \sigma^2 (\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{x}_k^{\perp} / (\|\mathbf{x}_j^{\perp}\|_2^2 \|\mathbf{x}_k^{\perp}\|_2^2). \tag{3}$$

In the high dimensional setting $p > n$, $\mathrm{rank}(\mathbf{X}_{-j}) = n$ for all j when $\mathbf{X}$ is in general position. In this case $\mathbf{x}_j^{\perp} = 0$ and estimator (2) is undefined. However, it may still be interesting to preserve certain properties of the least squares estimator. This can be done by retaining the main equation $\mathbf{z}_j^{\mathrm{T}}(\mathbf{y} - \beta_j \mathbf{x}_j) = 0$ in a score system $\mathbf{z}_j : \mathbf{v} \rightarrow \mathbf{z}_j^{\mathrm{T}}\mathbf{v}$ and relaxing the constraint $\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k = 0$ for $k \neq j$, resulting in a linear estimator. For any score vector $\mathbf{z}_j$ that is not orthogonal to $\mathbf{x}_j$, the corresponding univariate linear regression estimator satisfies

$$\hat{\beta}_j^{(\mathrm{lin})} = \frac{\mathbf{z}_j^{\mathrm{T}}\mathbf{y}}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j} = \beta_j + \frac{\mathbf{z}_j^{\mathrm{T}}\varepsilon}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j} + \sum_{k \neq j} \frac{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k \beta_k}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j}.$$

The estimator consequently has a covariance structure of form (3). A problem with the new system is its bias. For every $k \neq j$ with $\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k \neq 0$, the contribution of β_k to the bias is linear in β_k. Thus, under the assumption $\|\boldsymbol{\beta}\|_0 \leqslant 2$, which is very strong, the bias of $\hat{\beta}_j^{(\mathrm{lin})}$ is still unbounded when $\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k \neq 0$ for at least one $k \neq j$. We note that, for $\mathrm{rank}(\mathbf{X}_{-j}) = n$, it is impossible to have $\mathbf{z}_j \neq 0$ and $\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k = 0$ for all $k \neq j$, so bias is unavoidable. Nevertheless, this simple analysis of the linear estimator suggests a bias correction with a non-linear initial estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{init})}$:

$$\hat{\beta}_j = \hat{\beta}_j^{(\mathrm{lin})} - \sum_{k \neq j} \frac{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k \hat{\beta}_k^{(\mathrm{init})}}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j} = \frac{\mathbf{z}_j^{\mathrm{T}}\mathbf{y}}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j} - \sum_{k \neq j} \frac{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k \hat{\beta}_k^{(\mathrm{init})}}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j}. \tag{4}$$

We may also interpret equation (4) as a one-step self-bias correction from the initial estimator and write

$$\hat{\beta}_j := \hat{\beta}_j^{(\mathrm{init})} + \mathbf{z}_j^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(\mathrm{init})}) / \mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j.$$

The estimation error of equation (4) can be decomposed as a sum of the noise and the approximation errors:

$$\hat{\beta}_j - \beta_j = \frac{\mathbf{z}_j^{\mathrm{T}}\varepsilon}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j} + \frac{\sum_{k \neq j} \mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k (\beta_k - \hat{\beta}_k^{(\mathrm{init})})}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j}. \tag{5}$$

We require that $\mathbf{z}_j$ be a vector depending on $\mathbf{X}$ only, so that $\mathbf{z}_j^{\mathrm{T}}\varepsilon / \|\mathbf{z}_j\|_2 \sim N(0, \sigma^2)$. A full description of equation (4) still requires the specification of the score vector $\mathbf{z}_j$ and the initial estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{init})}$. These choices will be discussed in the following two subsections.

2.2. Low dimensional projections

We propose to use as $\mathbf{z}_j$ a relaxed orthogonalization of $\mathbf{x}_j$ against other design vectors. Recall that $\mathbf{z}_j$ aims to play the role of $\mathbf{x}_j^{\perp}$, the projection of $\mathbf{x}_j$ to the orthogonal complement of the column space of $\mathbf{X}_{-j} = (\mathbf{x}_k, k \neq j)$. In the trivial case where $\|\mathbf{x}_j^{\perp}\|_2$ is not too small, we may simply take $\mathbf{z}_j = \mathbf{x}_j^{\perp}$. In addition to the case of $\mathrm{rank}(\mathbf{X}_{-j}) = n$, in which $\mathbf{x}_j^{\perp} = 0$, a relaxed projection may be useful when $\|\mathbf{x}_j^{\perp}\|_2$ is positive but small. Since a relaxed projection $\mathbf{z}_j$ is used and estimator (4) is a bias-corrected projection of $\mathbf{y}$ in the direction of $\mathbf{z}_j$, hereafter we call estimator (4) the low dimensional projection estimator (LDPE) for easy reference.

A proper relaxed projection $\mathbf{z}_j$ should control both the noise and the approximation error terms in equation (5), given suitable conditions on $\{\mathbf{X},\boldsymbol{\beta}\}$ and an initial estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{init})}$. The approximation error of the LDPE (4) can be controlled by bounding the numerator of the bias term in equation (5) as follows:

$$\left|\sum_{k\neq j}\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k(\beta_k-\hat{\beta}_k^{(\mathrm{init})})\right|\leqslant\left(\max_{k\neq j}|\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k|\right)\|\hat{\boldsymbol{\beta}}^{(\mathrm{init})}-\boldsymbol{\beta}\|_1. \tag{6}$$

This conservative bound is conveniently expressed as the product of a known function of $\mathbf{z}_j$ and the initial estimation error independent of j. For score vectors $\mathbf{z}_j$, define

$$\begin{aligned}\eta_j&=\max_{k\neq j}|\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k|/\|\mathbf{z}_j\|_2,\\ \tau_j&=\|\mathbf{z}_j\|_2/|\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j|.\end{aligned} \tag{7}$$

We refer to η_j as the bias factor since $\eta_j\|\hat{\boldsymbol{\beta}}^{(\mathrm{init})}-\boldsymbol{\beta}\|_1$ controls the approximation error in constraint (6) relative to the length of the score vector. We refer to τ_j as the noise factor, since $\tau_j\sigma$ is the standard deviation of the noise component in equation (5). Since $\mathbf{z}_j^{\mathrm{T}}\boldsymbol{\varepsilon}\sim N(0,\sigma^2\|\mathbf{z}_j\|_2^2)$, equation (5) yields

$$\eta_j\|\hat{\boldsymbol{\beta}}^{(\mathrm{init})}-\boldsymbol{\beta}\|_1/\sigma=o(1)\Rightarrow\tau_j^{-1}(\hat{\beta}_j-\beta_j)\approx N(0,\sigma^2). \tag{8}$$

Thus, we would like to pick a $\mathbf{z}_j$ with a small η_j for the asymptotic normality and a small τ_j for efficiency of estimation. Confidence intervals for β_j and linear functionals of them can be constructed subject to condition (8) and a consistent estimator of σ.

We still need a suitable $\mathbf{z}_j$, a relaxed orthogonalization of $\mathbf{x}_j$ against other design vectors. When the unrelaxed $\mathbf{x}_j^{\perp}$ is non-zero, it can be viewed as the residual of the least squares fit of $\mathbf{x}_j$ on $\mathbf{X}_{-j}$. A familiar relaxation of the least squares method is the addition of an l_1-penalty. This leads to the choice of $\mathbf{z}_j$ as the residual of the lasso. Let $\hat{\gamma}_j$ be the vector of coefficients from the lasso regression of $\mathbf{x}_j$ on $\mathbf{X}_{-j}$. The lasso-generated score is

$$\mathbf{z}_j=\mathbf{x}_j-\mathbf{X}_{-j}\hat{\gamma}_j,\qquad\qquad \hat{\gamma}_j=\arg\min_{\mathbf{b}}\left\{\frac{\|\mathbf{x}_j-\mathbf{X}_{-j}\mathbf{b}\|_2^2}{2n}+\lambda_j\|\mathbf{b}\|_1\right\}. \tag{9}$$

It follows from the Karush–Kuhn–Tucker conditions for equation (9) that $|\mathbf{x}_k^{\mathrm{T}}\mathbf{z}_j/n|\leqslant\lambda_j$ for all $k\neq j$, so expression (7) holds with $\eta_j\leqslant n\lambda_j/\|\mathbf{z}_j\|_2$. This gives many choices of $\mathbf{z}_j$ with different $\{\eta_j,\tau_j\}$. Explicit choices of such a $\mathbf{z}_j$, or equivalently a λ_j, are described in the next subsection. A rationale for the use of a common penalty level λ_j for all components of $\mathbf{b}$ in equation (9) is the standardization of all design vectors. In an alternative in Section 2.4, called the restricted LDPE (RLDPE), the penalty is set to 0 for certain components of $\mathbf{b}$ in equation (9).

2.3. *Implementation of the low dimensional projection estimator*

We must pick $\hat{\boldsymbol{\beta}}^{(\mathrm{init})}$, $\hat{\sigma}$, and the λ_j in equation (9). Since consistent estimation of σ and fully automatic choices of λ_j are needed, we use methods based on the scaled lasso and the least squares estimator after model selection by the scaled lasso (called the scaled lasso–LSE method). We outline the basic ideas and specific implementations in Table 1.

The scaled lasso (Antoniadis, 2010; Sun and Zhang, 2010, 2012) is a joint convex minimization method given by

$$\{\hat{\boldsymbol{\beta}}^{(\mathrm{init})},\hat{\sigma}\}=\arg\min_{\mathbf{b},\sigma}\left\{\frac{\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2^2}{2\sigma n}+\frac{\sigma}{2}+\lambda_0\|\mathbf{b}\|_1\right\}, \tag{10}$$

Table 1. LDPE $(1-\alpha)$% confidence intervals for a preconceived β_j

Step	*Idea*	*Specific implementation*
1	Find an initial estimate $\hat{\boldsymbol{\beta}}^{(\mathrm{init})}$ of the entire $\boldsymbol{\beta}$ and a consistent estimate $\hat{\sigma}$	Scaled lasso (10) or least squares estimator after scaled lasso selection (11)
2	Find $\mathbf{z}_j$ as an approximate projection of $\mathbf{x}_j$ to $\mathbf{X}_{-j}$ in the sense of near orthogonality to each $\mathbf{x}_k$, $k \neq j$	Residuals, with small bias and noise factors (7), in lasso regression of $\mathbf{x}_j$ against $\mathbf{X}_{-j}$ (Table 2)
3	Calculate the LDPE $\hat{\beta}_j$ bias correction by projecting the residual of $\hat{\boldsymbol{\beta}}^{(\mathrm{init})}$ to $\mathbf{z}_j$	$\hat{\beta}_j = \hat{\beta}_j^{(\mathrm{init})} + \mathbf{z}_j^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(\mathrm{init})})/\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j$
4	Calculate the noise factor τ_j	$\tau_j = \Vert\mathbf{z}_j\Vert_2/\vert\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j\vert$
5	Calculate the LDPE confidence interval	$\hat{\beta}_j \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}\tau_j$

with a preassigned penalty level λ_0. It automatically provides an estimate of the noise level in addition to the initial estimator of $\boldsymbol{\beta}$. We use $\lambda_0 = \lambda_{\mathrm{univ}} = \surd\{(2/n)\log(p)\}$ in our simulation study. Existing error bounds for the estimation of both $\boldsymbol{\beta}$ and σ require $\lambda_0 = A\surd\{(2/n)\log(p/\epsilon)\}$ with certain $A > 1$ and $0 < \epsilon \leqslant 1$ (Sun and Zhang, 2012).

Estimator (10) has appeared in the literature in different forms. The joint minimization formulation was given in Antoniadis (2010), and an equivalent algorithm in Sun and Zhang (2010). If the minimum over $\mathbf{b}$ is taken first in equation (10), the resulting $\hat{\sigma}$ appeared earlier in Zhang (2010). The square-root lasso (Belloni *et al.*, 2011) gives the same $\hat{\boldsymbol{\beta}}^{(\mathrm{init})}$ with a different formulation, but not joint estimation of σ. In addition, the formulations in Zhang (2010) and Sun and Zhang (2010) allow concave penalties and a degrees-of-freedom adjustment.

Like the lasso, the scaled lasso is biased. An alternative initial estimator of $\{\boldsymbol{\beta}, \sigma\}$ can be produced by applying least squares after scaled lasso selection. Let $\hat{S}^{(\mathrm{scl})}$ be the set of non-zero estimated coefficients produced by the scaled lasso. When $\hat{S}^{(\mathrm{scl})}$ catches most large $|\beta_j|$, the bias of estimator (10) can be reduced by the least squares estimator in the selected model $\hat{S}^{(\mathrm{scl})}$ and the corresponding degrees of freedom adjusted estimate of σ:

$$\{\hat{\boldsymbol{\beta}}^{(\mathrm{init})}, \hat{\sigma}\} = \underset{\mathbf{b},\sigma}{\arg\min}\left\{\frac{\Vert\mathbf{y} - \mathbf{X}\mathbf{b}\Vert_2^2}{2\sigma\max(n - |\hat{S}^{(\mathrm{scl})}|, 1)} + \frac{\sigma}{2} : b_j = 0 \quad \forall j \notin \hat{S}^{(\mathrm{scl})}\right\}. \tag{11}$$

This defines the scaled lasso–LSE estimator. We use the same notation in equations (10) and (11) since they both give initial estimates for the LDPE (4) and a noise level estimator for statistical inference based on the LDPE. The specific estimators will henceforth be referred to by their names or as estimators (10) and (11). The scaled lasso–LSE estimator enjoys similar analytical error bounds to those of the scaled lasso and outperformed the scaled lasso in a simulation study (Sun and Zhang, 2012).

The scaled lasso can be also used to determine λ_j for the $\mathbf{z}_j$ in equation (9). However, the penalty level for the scaled lasso, set to guarantee performance bounds for the estimation of regression coefficients and noise level, may not be the best for controlling the bias and the standard error of the LDPE. By equations (7) and (8), it suffices to find a $\mathbf{z}_j$ with a small bias factor η_j and small noise factor τ_j. These quantities always can be easily computed. This is quite different from the estimation of $\{\boldsymbol{\beta}, \sigma\}$ in equation (10) in which the effect of overfitting is unobservable.

We choose λ_j by tracking η_j and τ_j in the lasso path. One of our ideas is to reduce η_j by allowing some overfitting of $\mathbf{x}_j$ as long as τ_j is reasonably small. Ideally, this slightly more conservative

Table 2. Computation of $\mathbf{z}_j$ from lasso (12)†

```
Objective: find a point in the path of lasso regression of $\mathbf{x}_j$ against $\mathbf{X}_{-j}$ such that the residual
  vector has small bias and noise factors

Input    $\eta_j^* = \surd\{2\log(p)\}$                                    {target bound for bias factor}
         $\kappa_0 = \frac{1}{4}$                                            {small tuning parameter}
Step 1   Compute $\mathbf{z}_j(\lambda)$ for $\lambda \geqslant \lambda_*$                        {residual of the lasso in equation (12)}
         Compute $\eta_j(\lambda)$ and $\tau_j(\lambda)$ for $\lambda \geqslant \lambda_*$              {bias and noise factors in equation (12)}
Step 2   If $\eta_j(\lambda_*) \geqslant \eta_j^*$, return $\mathbf{z}_j \leftarrow \mathbf{z}_j(\lambda_*)$;
         otherwise                                                 {controlled bias minimization}
           $\tau_j^* \leftarrow (1+\kappa_0)\min\{\tau_j(\lambda) : \eta_j(\lambda) \geqslant \eta_j^*\}$      {bound for noise factor}
           $\lambda \leftarrow \arg\min\{\eta_j(\lambda) : \tau_j(\lambda) \leqslant \tau_j^*\}$            {bias factor minimization}
           return $\mathbf{z}_j \leftarrow \mathbf{z}_j(\lambda)$
```

†Comments are given in braces; '=' denotes the default value; '←' denotes the assignment operation; λ_* is the smallest non-zero penalty level in the computed lasso path.

approach will lead to confidence intervals with more accurate coverage probabilities. Along the lasso path for regressing $\mathbf{x}_j$ against $\mathbf{X}_{-j}$, let

$$\hat{\gamma}_j(\lambda) = \underset{\mathbf{b}}{\arg\min}\{\|\mathbf{x}_j - \mathbf{X}_{-j}\mathbf{b}\|_2^2/(2n) + \lambda\|\mathbf{b}\|_1\} \tag{12}$$

$$\mathbf{z}_j(\lambda) = \mathbf{x}_j - \mathbf{X}_{-j}\hat{\gamma}_j(\lambda),$$

$$\eta_j(\lambda) = \max_{k \neq j} |\mathbf{x}_k^{\mathrm{T}}\mathbf{z}_j(\lambda)|/\|\mathbf{z}_j(\lambda)\|_2,$$

$$\tau_j(\lambda) = \|\mathbf{z}_j(\lambda)\|_2/|\mathbf{x}_j^{\mathrm{T}}\mathbf{z}_j(\lambda)|$$

be the coefficient estimator $\hat{\gamma}_j$, residual $\mathbf{z}_j$, the bias factor η_j and the noise factor τ_j, as functions of λ. We compute $\mathbf{z}_j$ according to the algorithm in Table 2.

In Table 2, step 2 finds a feasible upper bound η_j^* for the bias factor and the corresponding noise factor τ_j^*. It then seeks $\mathbf{z}_j = \mathbf{z}_j(\lambda_j)$ in equation (12) at a certain level $\lambda = \lambda_j$ with a smaller $\eta_j = \eta_j(\lambda_j)$, subject to the constraint $\tau(\lambda_j) \leqslant (1+\kappa_0)\tau_j^*$ on the noise factor. It follows from proposition 1, part (a), below that $\eta_j(\lambda)$ is non-decreasing in λ, so searching for the smallest $\eta_j(\lambda)$ is equivalent to searching for the smallest λ in step 2, subject to the constraint.

In the search for $\mathbf{z}_j$ with smaller η_j in step 2, the relative increment in the noise factor τ_j is no greater than κ_0. This corresponds to a loss of relative efficiency that is no greater than $1 - 1/(1+\kappa_0)^2$ for the estimation of β_j. In our simulation experiments, $\kappa_0 = \frac{1}{4}$ provides a suitable choice, compared with $\kappa_0 = 0$ and $\kappa_0 = \frac{1}{2}$. We would like to emphasize here that the score vectors $\mathbf{z}_j$ that are computed by the algorithm in Table 2 are completely determined by the design $\mathbf{X}$.

A main objective of the algorithm in Table 2 is to find a $\mathbf{z}_j$ with a bias factor $\eta_j \leqslant C\surd\log(p)$ to allow a uniform bias bound via expressions (6)–(8). It is ideal if $C = \surd 2$ is attainable, but a bounded C also works with the argument. When $\eta_j^* = \surd\{2\log(p)\}$ is not feasible, step 1 finds a larger upper bound η_j^* for the bias factor. When $\sup_\lambda \eta_j(\lambda) < \surd\{2\log(p)\}$, $\eta_j^* < \surd\{2\log(p)\}$ after the adjustment in step 1, resulting in an even smaller η_j in step 2. This does happen in our simulation experiments. The choice of the target upper bound $\surd\{2\log(p)\}$ for η_j is based on its feasibility as well as the sufficiency of $\eta_j \leqslant \surd\{2\log(p)\}$ for the verification of the condition in expression (8) based on the existing l_1-error bounds for the estimation of $\boldsymbol{\beta}$. Proposition 1 below asserts that $\max_{j \leqslant p}\eta_j^* \leqslant C\surd\log(p)$ is feasible when $\mathbf{X}$ allows an optimal rate of sparse recovery. In our simulation experiments, we can use $\eta_j^* \leqslant \surd\{2\log(p)\}$ in all replications and settings for all variables, a total of more than 1 million instances. Moreover, the theoretical results in Section 3.5 prove that, for the η_j^* in Table 2, $\max_{j\leqslant p}\eta_j^*\leqslant 3\surd\log(p)$ with high probability under proper conditions on random $\mathbf{X}$. It is worthwhile to note that both η_j and τ_j are readily computed, and control of $\max_k \eta_k$ is not required for the LDPE to apply to variables with small η_j.

We use the rest of this subsection to present some useful properties of the lasso path (12) for the implementation of the algorithm in Table 2 and some sufficient conditions for the uniform bound $\max_j \eta_j^*\leqslant C\surd\log(p)$ of the bias factors in the output. Let

$$\hat\sigma_j(\lambda)=\arg\min_{\sigma}\min_{\mathbf{b}}\left\{\frac{\|\mathbf{x}_j-\mathbf{X}_{-j}\mathbf{b}\|_2^2}{2n\sigma}+\frac{\sigma}{2}+\lambda\|\mathbf{b}\|_1\right\} \tag{13}$$

be the solution of $\hat\sigma$ in equation (10) with $\{\mathbf{X},\mathbf{y},\lambda_0\}$ replaced by $\{\mathbf{X}_{-j},\mathbf{x}_j,\lambda\}$.

Proposition 1.

(a) In the lasso path (12), $\|\mathbf{z}_j(\lambda)\|_2$, $\eta_j(\lambda)$ and $\hat\sigma_j(\lambda)$ are non-decreasing functions of λ, and $\tau_j(\lambda)\leqslant 1/\|\mathbf{z}_j(\lambda)\|_2$. Moreover, $\hat\gamma_j(\lambda)\neq 0$ implies that $\eta_j(\lambda)=\lambda n/\|\mathbf{z}_j(\lambda)\|_2$.

(b) Let $\lambda_{\text{univ}}=\surd\{(2/n)\log(p)\}$. Then,

$$\hat\sigma_j(C\lambda_{\text{univ}})>0 \text{ iff } \{\lambda>0:\eta_j(\lambda)\leqslant C\surd\{2\log(p)\}\neq\emptyset, \tag{14}$$

and, in this case, the algorithm in Table 2 provides

$$\begin{aligned}\eta_j&\leqslant\eta_j^*\leqslant(1\vee C)\surd\{2\log(p)\},\\ \tau_j&\leqslant n^{-1/2}(1+\kappa_0)/\hat\sigma_j(C\lambda_{\text{univ}}).\end{aligned} \tag{15}$$

Moreover, when $\mathbf{z}_j(0)=\mathbf{x}_j^\perp=0$, $\eta_j(0+)\inf\{\|\gamma_j\|_1:\mathbf{X}_{-j}\gamma_j=\mathbf{x}_j\}=\surd n$.

(c) Let $0<a_0<1\leqslant C_0<\infty$. Suppose that for $s=a_0n/\log(p)$

$$\inf_{\delta}\sup_{\beta}\left\{\|\boldsymbol{\delta}(\mathbf{X},\mathbf{y})-\boldsymbol{\beta}\|_2^2:\mathbf{y}=\mathbf{X}\boldsymbol{\beta},\ \sum_{j=1}^{p}\min(|\beta_j|/\lambda_{\text{univ}},1)\leqslant s+1\right\}\leqslant 2C_0s\log(p)/n.$$

Then, $\max_{j\leqslant p}\eta_j^*\leqslant\surd\{(4C_0/a_0)\log(p)\}$ for the algorithm in Table 2.

The monotonicity of $\|\mathbf{z}_j(\lambda)\|_2$ and $\eta_j(\lambda)$ in proposition 1, part (a), provides directions of search in both steps of the algorithm in Table 2.

Proposition 1, part (b), provides mild conditions for controlling the bias factor at $\eta_j\leqslant\eta_j^*\leqslant C\surd\{2\log(p)\}$ and the standard error to the order $\tau_j=O(n^{-1/2})$. It asserts that $\eta_j^*\leqslant\surd\{2\log(p)\}$ when the scaled lasso (13) with $\lambda=\lambda_{\text{univ}}$ yields a positive $\hat\sigma_j$. In the completely collinear case where $\mathbf{x}_k=\mathbf{x}_j$ for some $k\neq j$, $\inf\{\|\gamma_j\|_1:\mathbf{x}_j=\mathbf{X}_{-j}\gamma_j\}=1$ gives the largest $\eta_j=\surd n$. This suggests a connection between the minimum feasible η_j and 'near estimability' of β_j, with small η_j for nearly estimable β_j. It also provides a connection between the smallest $\eta_j(\lambda)$ and an l_1-recovery problem, leading to proposition 1, part (c).

Proposition 1, part (c), asserts that the validity of the upper bound $\max_j\eta_j^*\leqslant C\surd\log(p)$ for the bias factor is a consequence of the existence of an estimator $\boldsymbol{\delta}$ with the stated l_2-recovery bound in the noiseless case of $\varepsilon=0$. In the more difficult case of $\varepsilon\sim N(0,\sigma^2\mathbf{I})$, l_2-error bounds of the same type have been proven under sparse eigenvalue conditions on $\mathbf{X}$, and by proposition 1, part (c), $\max_j\eta_j^*\leqslant C\surd\log(p)$ is also a consequence of such conditions.

2.4. Restricted low dimensional projection estimator

We have also experimented with an LDPE using a restricted lasso relaxation for $\mathbf{z}_j$. This RLDPE can be viewed as a special case of a more general weighted low dimensional projection with different levels of relaxation for different variables $\mathbf{x}_k$ according to their correlation to $\mathbf{x}_j$. Although we have used equation (6) to bound the bias, the summands with larger absolute correlation $|\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_k/n|$ are likely to have a greater contribution to the bias due to the initial estimation error $|\hat{\beta}_k^{(\mathrm{init})} - \beta_k|$. A remedy for this phenomenon is to force smaller $|\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k/n|$ for large $|\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_k/n|$ with a weighted relaxation. For the lasso (9), this weighted relaxation can be written as

$$\mathbf{z}_j = \mathbf{x}_j - \mathbf{X}_{-j}\hat{\gamma}_j, \qquad \hat{\gamma}_j = \arg\min_{\mathbf{b}} \left\{ \frac{\|\mathbf{x}_j - \mathbf{X}_{-j}\mathbf{b}\|_2^2}{2n} + \lambda_j \sum_{k \neq j} w_k |b_k| \right\},$$

with w_k being a decreasing function of the absolute correlation $|\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_k/n|$. In the RLDPE, we simply set $w_k = 0$ for large $|\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_k/n|$ and $w_k = 1$ for other k.

Here is an implementation of the RLDPE. Let $K_{j,m}$ be the index set of the m largest $|\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_k|$ with $k \neq j$, and let $\mathbf{P}_{j,m}$ be the orthogonal projection to the linear span of $\{\mathbf{x}_k, k \in K_{j,m}\}$. Let $\mathbf{z}_j = f(\mathbf{x}_j, \mathbf{X}_{-j})$ denote the algorithm in Table 2 as a mapping $(\mathbf{x}_j, \mathbf{X}_{-j}) \rightarrow \mathbf{z}_j$. We compute the RLDPE by taking the projection of all design vectors to the orthogonal complement of $\{\mathbf{x}_k, k \in K_{j,m}\}$ before the application of the procedure in equation (12) and Table 2. The resulting score vector can be written as

$$\mathbf{z}_j = f(\mathbf{P}_{j,m}^{\perp}\mathbf{x}_j, \mathbf{P}_{j,m}^{\perp}\mathbf{X}_{-j}). \tag{16}$$

2.5. *Confidence intervals*

In Section 3, we provide sufficient conditions on $\mathbf{X}$ and $\boldsymbol{\beta}$ under which the approximation error in equation (5) is of smaller order than the standard deviation of the noise component. We construct approximate confidence intervals for such configurations of $\{\mathbf{X}, \boldsymbol{\beta}\}$ as follows.

The covariance of the noise component in equation (5) is proportional to

$$\mathbf{V} = (V_{jk})_{p \times p}, \qquad V_{jk} = \frac{\mathbf{z}_j^{\mathrm{T}}\mathbf{z}_k}{|\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j||\mathbf{z}_k^{\mathrm{T}}\mathbf{x}_k|} = \sigma^{-2}\mathrm{cov}\left(\frac{\mathbf{z}_j^{\mathrm{T}}\varepsilon}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j}, \frac{\mathbf{z}_k^{\mathrm{T}}\varepsilon}{\mathbf{z}_k^{\mathrm{T}}\mathbf{x}_k} \right). \tag{17}$$

Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathrm{T}}$ be the vector of LDPEs $\hat{\beta}_j$ in equation (4). For sparse vectors $\mathbf{a}$ with bounded $\|\mathbf{a}\|_0$, e.g. $\|\mathbf{a}\|_0 = 2$ for a contrast between two regression coefficients, an approximate $100(1-\alpha)\%$ confidence interval is

$$|\mathbf{a}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}| \leqslant \hat{\sigma}\,\Phi^{-1}(1-\alpha/2)(\mathbf{a}^{\mathrm{T}}\mathbf{V}\mathbf{a})^{1/2}, \tag{18}$$

where Φ is the standard normal distribution function. We may choose $\{\hat{\boldsymbol{\beta}}^{(\mathrm{init})}, \hat{\sigma}\}$ in equation (10) or (11) and $\mathbf{z}_j$ in Table 2 or equation (16) in the construction of $\hat{\boldsymbol{\beta}}$ and the confidence intervals. An alternative, larger estimate of σ, producing more conservative approximate confidence intervals, is the penalized maximum likelihood estimator of Städler *et al.* (2010).

3. Theoretical results

In this section, we prove that, when the l_1-loss of the initial estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{init})}$ is of an expected magnitude and the noise level estimator $\hat{\sigma}$ is consistent, the LDPE-based confidence interval has approximately the preassigned coverage probability for statistical inference of linear combinations of β_j with sufficiently small η_j. Under proper conditions on X such as those given in proposition 1, the width of such confidence intervals is of the order $\tau_j \asymp n^{-1/2}$. The accuracy of the approximation for the coverage probability is sufficiently sharp to allow simultaneous interval estimation of all β_j.

The LDPE provides a sequence of approximately normal estimates $\hat{\beta}_j$ with errors of order $n^{-1/2}$. This sequence is not sparse but can be thresholded just as in the Gaussian sequence model. Although variable selection and the estimation of the entire vector $\boldsymbol{\beta}$ are not the focus of the paper, the thresholded LDPE has important implications for these closely related topics. They are discussed in Section 3.3.

In Section 3.4, we use existing error bounds to verify the conditions on $\hat{\boldsymbol{\beta}}^{(\text{init})}$ and $\hat{\sigma}$ under regularity conditions on deterministic designs and a capped l_1 relaxation of the sparsity condition $\|\boldsymbol{\beta}\|_0 \leqslant s$, provided that $s\log(p) \ll \surd n$. Random-matrix theory is used in Section 3.5 to check regularity conditions on both deterministic and random designs.

3.1. Confidence intervals for preconceived parameters, deterministic design

Here we establish the asymptotic normality of the LDPE (4) and the validity of the resulting confidence interval (18) for a preconceived parameter. This result is new and useful in and of itself since high dimensional data often present a few effects that are known to be of high interest in advance. Examples include treatment effects in clinical trials, or the effect of education on income in socio-economic studies. Simultaneous confidence intervals for all individual β_j and the thresholded LDPE for the entire vector $\boldsymbol{\beta}$ will be considered in the next subsection as consequences of this result.

Let $\lambda_{\text{univ}} = \surd\{(2/n)\log(p)\}$. Suppose that model (1) holds with a vector $\boldsymbol{\beta}$ satisfying the capped l_1 sparsity condition:

$$\sum_{j=1}^{p} \min\{|\beta_j|/(\sigma\lambda_{\text{univ}}), 1\} \leqslant s. \tag{19}$$

This condition holds if $\boldsymbol{\beta}$ is l_0 sparse with $\|\boldsymbol{\beta}\|_0 \leqslant s$ or l_q sparse with $\|\boldsymbol{\beta}\|_q^q/(\sigma\lambda_{\text{univ}})^q \leqslant s$, $0<q\leqslant 1$. Let $\sigma^* = \|\boldsymbol{\varepsilon}\|_2/\surd n$. A generic condition that we impose on the initial estimator is

$$P[\|\hat{\boldsymbol{\beta}}^{(\text{init})} - \boldsymbol{\beta}\|_1 \geqslant C_1 s\sigma^* \surd\{(2/n)\log(p/\epsilon)\}] \leqslant \epsilon \tag{20}$$

for a certain fixed constant C_1 and all $\alpha_0/p^2 \leqslant \epsilon \leqslant 1$, where $\alpha_0 \in (0,1)$ is a preassigned constant. We also impose a similar generic condition on an estimator $\hat{\sigma}$ for the noise level:

$$P\{|\hat{\sigma}/\sigma^* - 1| \geqslant C_2 s(2/n)\log(p/\epsilon)\} \leqslant \epsilon, \qquad \forall \alpha_0/p^2 \leqslant \epsilon \leqslant 1, \tag{21}$$

with a fixed C_2. We use the same ϵ in condition (20) and (21) without much loss of generality.

By requiring fixed $\{C_1, C_2\}$, we implicitly impose regularity conditions on the design $\mathbf{X}$ and the sparsity index s in condition (19). Existing oracle inequalities can be used to verify condition (20) for various regularized estimators of $\boldsymbol{\beta}$ under different sets of conditions on $\mathbf{X}$ and $\boldsymbol{\beta}$ (Candès and Tao, 2007; Zhang and Huang, 2008; Bickel *et al.*, 2009; van de Geer and Bühlmann, 2009; Zhang, 2009, 2010; Ye and Zhang, 2010; Sun and Zhang, 2012; Zhang and Zhang, 2012). Although most existing results are derived for penalty or threshold levels depending on a known noise level σ and under the l_0-sparsity condition on $\boldsymbol{\beta}$, their proofs can be combined or extended to obtain condition (20) once condition (21) becomes available. For the joint estimation of $\{\boldsymbol{\beta}, \sigma\}$ with estimators (10) or (11), specific sets of sufficient conditions for both condition (20) and condition (21), based on Sun and Zhang (2012), are stated in Section 3.4. In fact, the probability of the union of the two events is smaller than ϵ in the specific case where $\lambda_0 = A\surd\{(2/n)\log(p/\epsilon)\}$ in equation (10) for a certain $A>1$.

Theorem 1. Let $\hat{\beta}_j$ be the LDPE in equation (4) with an initial estimator $\hat{\boldsymbol{\beta}}^{(\text{init})}$. Let η_j and τ_j be the bias and noise factors in equation (7), $\sigma^* = \|\boldsymbol{\varepsilon}\|_2/\surd n$, $\max(\epsilon'_n, \epsilon''_n) \to 0$ and $\eta^* > 0$. Suppose that condition (20) holds with $\eta^* C_1 s\surd\{(2/n)\log(p/\epsilon)\} \leqslant \epsilon'_n$. If $\eta_j \leqslant \eta^*$, then

$$P\{|\tau_j^{-1}(\hat{\beta}_j-\beta_j)-\mathbf{z}_j^{\mathrm{T}}\varepsilon/\|\mathbf{z}_j\|_2|>\sigma^*\epsilon_n'\}\leqslant\epsilon. \tag{22}$$

If in addition condition (21) holds with $C_2 s(2/n)\log(p/\epsilon)\leqslant\epsilon_n''$, then, for all $t\geqslant(1+\epsilon_n')/(1-\epsilon_n'')$,

$$P(|\hat{\beta}_j-\beta_j|\geqslant\tau_j\hat{\sigma}t)\leqslant 2\Phi_{n-1}\{-(1-\epsilon_n'')t+\epsilon_n'\}+2\epsilon, \tag{23}$$

where $\Phi_n(t)$ is the Student t-distribution function with n degrees of freedom. Moreover, for the covariance matrix $\mathbf{V}$ in expression (17) and all fixed m,

$$\lim_{n\to\infty}\inf_{\mathbf{a}\in\mathscr{A}_{n,p,m}} P\{|\mathbf{a}^{\mathrm{T}}\hat{\boldsymbol{\beta}}-\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}|\leqslant\hat{\sigma}\,\Phi^{-1}(1-\alpha/2)(\mathbf{a}^{\mathrm{T}}\mathbf{V}\mathbf{a})^{1/2}\}=1-\alpha, \tag{24}$$

where $\Phi(t)=P\{N(0,1)\leqslant t\}$ and $\mathscr{A}_{n,p,m}=\{\mathbf{a}:\|\mathbf{a}\|_0\leqslant m,\ \max_{j\leqslant p}|a_j|\eta_j\leqslant\eta^*\}$.

Since $(\mathbf{z}_j^{\mathrm{T}}\varepsilon/\|\mathbf{z}_j\|_2, j\leqslant p)$ has a multivariate normal distribution with identical marginal distributions $N(0,\sigma^2)$, condition (22) establishes the joint asymptotic normality of the LDPE for finitely many $\hat{\beta}_j$ under condition (20). This allows us to write the LDPE as an approximate Gaussian sequence

$$\hat{\beta}_j=\beta_j+N(0,\tau_j^2\sigma^2)+o_P(\tau_j\sigma). \tag{25}$$

Under the additional condition (21), conditions (23) and (24) justify the approximate coverage probability of the resulting confidence intervals.

Remark 1. In theorem 1, all conditions on $\mathbf{X}$ and $\boldsymbol{\beta}$ are imposed through conditions (20) and (21), and the requirement of relatively small η_j to work with these conditions. The uniform signal strength condition can be written as

$$\min_{\beta_j\neq 0}|\beta_j|\geqslant C\sigma\surd\{(2/n)\log(p)\}. \tag{26}$$

Here $C\geqslant\frac{1}{2}$, which is required for variable selection consistency (Wainwright, 2009a; Zhang, 2010), is not required for conditions (20) and (21). This is the most important feature of the LDPE in setting it apart from variable selection approaches. More explicit sufficient conditions for conditions (20) and (21) are given in Section 3.4 for the initial estimators (10) and (11).

Remark 2. Although theorem 1 does not require τ_j to be small, the noise factor is proportional to the width of the confidence interval and thus its square is reciprocal to the efficiency of the LDPE. The bias factor η_j is required to be relatively small for equations (1) and (4), but no condition is imposed on $\{\eta_k, k\neq j\}$ for the inference of β_j. Since η_j and τ_j are computed in Table 2, we may apply theorem 1 to a set of the easy-to-estimate β_j with small $\{\eta_j,\tau_j\}$ and leave out some difficult-to-estimate regression coefficients if necessary.

In our implementation in Table 2, $\mathbf{z}_j$ is the residual of the lasso estimator in the regression model for $\mathbf{x}_j$ against $\mathbf{X}_{-j}=(\mathbf{x}_k, k\neq j)$. It follows from proposition 1 that, under proper conditions on the design matrix, $\eta_j\asymp\surd\log(p)$ and $\tau_j\leqslant 1/\|\mathbf{z}_j\|_2\asymp n^{-1/2}$ for the algorithm in Table 2. Such rates are realized in the simulation experiments that are described in Section 4 and further verified for Gaussian designs in Section 3.5. Thus, the dimension constraint for the asymptotic normality and proper coverage probability in theorem 1 is $s\log(p)/\surd n\to 0$.

3.2. Simultaneous confidence intervals

Here we provide theoretical justifications for simultaneous applications of the proposed LDPE confidence interval, after multiplicity adjustments, in the absence of a preconceived parameter of interest. In theorem 1, condition (22) is uniform in $\epsilon\in[\alpha_0/p^2,1]$ and condition (23) is uniform in the corresponding t. This uniformity allows Bonferroni adjustments to control the familywise error rate in simultaneous interval estimation.

Theorem 2. Suppose that condition (20) holds with $\eta^* C_1 s\surd\{(2/n)\log(p/\epsilon)\} \leqslant \epsilon'_n$. Then,

$$P\{\max_{\eta_j\leqslant\eta^*} |\tau_j^{-1}(\hat\beta_j-\beta_j) - \mathbf{z}_j^{\mathrm{T}}\epsilon/\|\mathbf{z}_j\|_2| > \sigma^*\epsilon'_n\} \leqslant \epsilon. \tag{27}$$

If condition (21) also holds with $C_2 s(2/n)\log(p/\epsilon) \leqslant \epsilon''_n$, then, for all $j \leqslant p$ and $t \geqslant (1+\epsilon'_n)/(1-\epsilon''_n)$,

$$P\{\max_{\eta_j\leqslant\eta^*} |\hat\beta_j-\beta_j|/(\tau_j\hat\sigma) > t\} \leqslant 2\,\Phi_n\{-(1-\epsilon''_n)t+\epsilon'_n\}\,\#\{j:\eta_j\leqslant\eta^*\}+2\epsilon. \tag{28}$$

If, in addition to condition (20) and condition (21), $\max_{j\leqslant p}\eta_j \leqslant \eta^* = O\{\surd\log(p)\}$ and $\max\{s\log(p)/n^{1/2},\epsilon\} \to 0$ as $\min(n,p)\to\infty$, then, for fixed $\alpha\in(0,1)$ and $c_0>0$,

$$\liminf_{n\to\infty} P\left[\max_{j\leqslant p}\left|\frac{\hat\beta_j-\beta_j}{\tau_j(\hat\sigma\wedge\sigma)}\right| \leqslant c_0 + \sqrt{\left\{2\log\left(\frac{p}{\alpha}\right)\right\}}\right] \geqslant 1-\alpha. \tag{29}$$

The error bound (27) asserts that the $o_P(1)$ in equation (25) is uniform in j. This uniform central limit theorem and the simultaneous confidence intervals (28) and (29) are valid as long as conditions (20) and (21) hold with $s\log(p) = o(n^{1/2})$. Since conditions (20) and (21) are consequences of condition (19) and proper regularity conditions on $\mathbf{X}$, these results do not require the uniform signal strength condition (26), as discussed in remark 1.

It follows from condition (25) and proposition 1, part (b), that, for a fixed j, the estimation error of $\hat\beta_j$ is of the order $\tau_j\sigma$ and $\tau_j \asymp n^{-1/2}$ under proper conditions. In contrast, with penalty level $\lambda = \sigma\surd\{(2/n)\log(p)\}$, the lasso may have a high probability of estimating β_j by 0 when $\beta_j = \lambda/2$. Thus, in the worst case scenario, the lasso inflates the error by a factor of order $\surd\log(p)$. Of course, the lasso is superefficient when it estimates the actual zero β_j by 0.

3.3. Thresholded low dimensional projection estimator

The raw LDPE (4) is not designed for variable selection or the estimation of the entire vector $\boldsymbol\beta$, although these topics are closely related to statistical inference of individual coefficients. Instead, the thrust of the LDPE approach is to turn regression problem (1) into a Gaussian sequence model (25) with uniformly small approximation error and a consistent estimator of the covariance structure. If the LDPE is used directly to estimate the entire vector $\boldsymbol\beta$, it has an l_2-error of order $\sigma^2 p/n$, compared with $\sigma^2 s\log(p)/n$ for the lasso. Thus, although the raw LDPE is sufficient for statistical inference of a preconceived β_j, it is not optimal for estimation of the entire $\boldsymbol\beta$ or for variable selection. Our recommendation is instead to use a thresholded LDPE. An interesting question is whether this thresholded LDPE has any advantages over existing regularized estimators, such as the lasso, for the estimation of high dimensional objects. This question is addressed here.

Thresholding may be performed by using either the hard or the soft thresholding method, respectively

$$\hat\beta_j^{(\mathrm{thr})} = \begin{cases} \hat\beta_j I(|\hat\beta_j| > \hat t_j), \\ \mathrm{sgn}(\hat\beta_j)(|\hat\beta_j| - \hat t_j)^+, \end{cases} \tag{30}$$

with

$$\hat S^{(\mathrm{thr})} = \{j : |\hat\beta_j| > \hat t_j\},$$

where $\hat{\beta}_j$ is as in theorem 1 and $\hat{t}_j \approx \hat{\sigma}\tau_j \Phi^{-1}\{1-\alpha/(2p)\}$ for some $\alpha > 0$. Although the theory is similar between the two (Donoho and Johnstone, 1994), our explicit analysis focuses on soft thresholding. As was the case for simultaneous confidence intervals, the thresholded LDPE can be justified by the uniformity of condition (22) in $\epsilon \in [\alpha_0/p^2, 1]$ and of condition (23) in the corresponding t in theorem 1. This uniformity applies to the approximation for the Gaussian sequence (25), leading to sharp l_2- and selection error bounds of the thresholded LDPE for the estimation of the entire vector $\boldsymbol{\beta}$.

Theorem 3. Let $L_0 = \Phi^{-1}\{1-\alpha/(2p)\}$, $\tilde{t}_j = \tau_j \sigma L_0$, and $\hat{t}_j = (1+c_n)\hat{\sigma}\tau_j L_0$ with positive constants α and c_n. Suppose that condition (20) holds with $\eta^* C_1 s/\sqrt{n} \leqslant \epsilon_n'$, $\max_{j\leqslant p} \eta_j \leqslant \eta^*$, and

$$P\left\{ \frac{(\hat{\sigma}/\sigma) \vee (\sigma/\hat{\sigma}) - 1 + \epsilon_n'\sigma^*/(\hat{\sigma}\wedge\sigma)}{1-(\hat{\sigma}/\sigma - 1)_+} > c_n \right\} \leqslant 2\epsilon. \tag{31}$$

Let $\hat{\boldsymbol{\beta}}^{(\text{thr})} = (\hat{\beta}_1^{(\text{thr})}, \ldots, \hat{\beta}_p^{(\text{thr})})^{\mathrm{T}}$ be the soft thresholded LDPE (30) with these $\hat{t}_j$. Then, there is an event Ω_n with $P(\Omega_n^{\mathrm{c}}) \leqslant 3\epsilon$ such that

$$E\|\hat{\boldsymbol{\beta}}^{(\text{thr})} - \boldsymbol{\beta}\|_2^2 I_{\Omega_n} \leqslant \sum_{j=1}^{p} \min[\beta_j^2, \tau_j^2\sigma^2\{L_0^2(1+2c_n)^2+1\}] + \frac{\epsilon L_n}{p}\sigma^2 \sum_{j=1}^{p} \tau_j^2, \tag{32}$$

where $L_n = 4/L_0^3 + 4c_n/L_0 + 12c_n^2 L_0$. Moreover, with at least probability $1-\alpha-3\epsilon$,

$$\{j : |\beta_j| > (2+2c_n)\tilde{t}_j\} \subseteq \hat{S}^{(\text{thr})} \subseteq \{j : \beta_j \neq 0\}. \tag{33}$$

Remark 3.

(a) Since $\max_{j\leqslant p} \eta_j \leqslant C\sqrt{\log(p)}$ can be achieved under mild conditions, the sample size requirement of theorem 3 is $s\sqrt{\{\log(p)/n\}} \to 0$ for the estimation and selection error bounds in expressions (32) and (33). This is a weaker requirement than $s\log(p)/\sqrt{n} \to 0$ for the asymptotic normality in equation (24).

(b) The proof of theorem 3 shows that, for proper small constants $c_n > 0$, inequality (31) is a consequence of condition (21). We assume this for the rest of the paper.

As implied by theorem 3, the major difference between expression (33) and existing variable selection consistency theory is again in the signal requirement. Variable selection consistency requires the uniform signal strength condition (26) as discussed in remark 1. Moreover, existing variable selection methods are not guaranteed to select correctly variables with large $|\beta_j|$ or $\beta_j = 0$ in the presence of small $|\beta_j| \neq 0$. In comparison, theorem 3 makes no assumption of condition (26). Under the regularity conditions for expression (33), large $|\beta_j|$ are selected by the thresholded LDPE and $\beta_j = 0$ are not selected, in the presence of possibly many small non-zero $|\beta_j|$.

There is also an analytical difference between the thresholded LDPE and existing regularized estimators that lies in the quantities that are thresholded. For the LDPE, the effect of thresholding on the approximate Gaussian sequence (25) is explicit and requires only univariate analysis to understand. In comparison, for the lasso and some other regularized estimators, thresholding is applied to the gradient $\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n$ via Karush–Kuhn–Tucker-type conditions, leading to more complicated non-linear multivariate analysis.

For the estimation of $\boldsymbol{\beta}$, the order of the l_2-error bound in inequality (32), $\Sigma_{j=1}^{p} \min(\beta_j^2, \sigma^2\lambda_{\text{univ}}^2)$, is slightly sharper than the typical order of $\|\boldsymbol{\beta}\|_0\sigma^2\lambda_{\text{univ}}^2$ or $\sigma\lambda_{\text{univ}}\Sigma_{j=1}^{p} \min\{|\beta_j|, \sigma\lambda_{\text{univ}}\}$ in the literature, where $\lambda_{\text{univ}} = \sqrt{\{(2/n)\log(p)\}}$. However, since the lasso and other regularized estimators are proven to be rate optimal in the maximum l_2-estimation loss for many classes of sparse $\boldsymbol{\beta}$, the main advantage of the thresholded LDPE seems to be the clarity of the effect of thresholding on the individual $\hat{\beta}_j$ in the approximate Gaussian sequence (25).

3.4. Checking conditions by oracle inequalities

Our main theoretical results, which are stated in theorems 1–3 in the above two subsections, justify the LDPE-based confidence interval of a single preconceived linear parameter of $\boldsymbol{\beta}$, simultaneous confidence intervals for all β_j and the estimation and selection error bounds for the thresholded LDPE for the vector $\boldsymbol{\beta}$. These results are based on conditions (20) and (21). We have mentioned that, for proper $\hat{\boldsymbol{\beta}}^{(\text{init})}$ and $\hat{\sigma}$, these two generic conditions can be verified in many ways under condition (19) on the basis of existing results. The purpose of this subsection is to describe a specific way of verifying these two conditions and thus to provide a more definitive and complete version of the theory.

In regularized linear regression, oracle inequalities have been established for different estimators and loss functions. We confine our discussion here to the scaled lasso (10) and the scaled lasso–LSE estimator (11) as specific choices of the initial estimator, since the confidence interval in theorem 1 is based on the joint estimation of regression coefficients and the noise level. We further confine our discussion to bounds for the l_1-error of $\hat{\boldsymbol{\beta}}^{(\text{init})}$ and the relative error of $\hat{\sigma}$ involved in conditions (20) and (21).

We use the results in Sun and Zhang (2012), where properties of estimators (10) and (11) were established on the basis of a compatibility factor (van de Geer and Bühlmann, 2009) and sparse eigenvalues. Let $\xi \geqslant 1$, $S = \{j : |\beta_j| > \sigma\lambda_{\text{univ}}\}$, and $\mathscr{C}(\xi, S) = \{\mathbf{u} : \|\mathbf{u}_{S^c}\|_1 \leqslant \xi \|\mathbf{u}_S\|_1\}$. The compatibility factor is defined as

$$\kappa(\xi, S) = \inf\{\|\mathbf{X}\mathbf{u}\|_2 |S|^{1/2} / (n^{1/2} \|\mathbf{u}_S\|_1) : 0 \neq \mathbf{u} \in \mathscr{C}(\xi, S)\}. \tag{34}$$

Let $\phi_{\min}$ and $\phi_{\max}$ denote the smallest and largest eigenvalues of matrices respectively. For positive integers m, define sparse eigenvalues as

$$\begin{aligned} \phi_-(m, S) &= \min_{B \supset S,\, |B \setminus S| \leqslant m} \phi_{\min}(\mathbf{X}_B^{\mathrm{T}} \mathbf{X}_B / n), \\ \phi_+(m, S) &= \min_{B \cap S = \emptyset,\, |B| \leqslant m} \phi_{\max}(\mathbf{X}_B^{\mathrm{T}} \mathbf{X}_B / n). \end{aligned} \tag{35}$$

The following theorem is a consequence of checking the conditions of theorem 1 by theorems 2 and 3 in Sun and Zhang (2012).

Theorem 4. Let $\{A, \xi, c_0\}$ be fixed positive constants with $\xi > 1$ and $A > (\xi + 1)/(\xi - 1)$. Let $\lambda_0 = A\surd\{(2/n)\log(p/\epsilon)\}$. Suppose that $\boldsymbol{\beta}$ is sparse in the sense of condition (19), $\kappa^2(\xi, S) \geqslant c_0$, and $(s \vee 1)(2/n)\log(p/\epsilon) \leqslant \mu_*$ for a certain $\mu^* > 0$.

(a) Let $\hat{\boldsymbol{\beta}}^{(\text{init})}$ and $\hat{\sigma}$ be the scaled lasso estimator in equation (10). Then, conditions (20) and (21) hold for certain constants $\{\mu_*, C_1, C_2\}$ depending on $\{A, \xi, c_0\}$ only. Consequently, all conclusions of theorems 1–3 hold with $C_1 \eta^*(s\lambda_0/A) \leqslant \epsilon'_n$ and $C_2 s(\lambda_0/A)^2 \leqslant \epsilon''_n$ for certain $\{\epsilon', \epsilon''\}$ satisfying $\max(\epsilon', \epsilon'') \to 0$.

(b) Let $\hat{\boldsymbol{\beta}}^{(\text{init})}$ and $\hat{\sigma}$ be the scaled lasso–LSE estimator (11). Suppose that $\xi^2/\kappa^2(\xi, S) \leqslant K/\phi_+(m, S)$ and $\phi_-(m, S) \geqslant c_1 > 0$ for certain $K > 0$ and integer $m - 1 < K|S| \leqslant m$. Then, conditions (20) and (21) hold for certain constants $\{\mu^*, C_1, C_2\}$ depending on $\{A, \xi, c_0, c_1, K\}$ only. Consequently, all conclusions of theorems 1–3 hold with $C_1 \eta^*(s\lambda_0/A) \leqslant \epsilon'_n$ and $C_2 s(\lambda_0/A)^2 \leqslant \epsilon''_n$ for certain $\{\epsilon', \epsilon''\}$ satisfying $\max(\epsilon', \epsilon'') \to 0$.

Remark 4. Let $A = (\xi + 1)/(\xi - 1)$. Then, there are constants $\{\tau_0, \nu_0\} \subset (0, 1)$ satisfying the condition $(1 - \tau_0^2)A = (\xi + 1)/\{\xi - (1 + \nu_0)/(1 - \nu_0)\}$. For these $\{\tau_0, \nu_0\}$, $n \geqslant 3$, and $p \geqslant 7$, we may take the constants in theorem 4, part (a), as

$$\mu_* = \min\left\{ \frac{2c_0\tau_0^2}{A^2(\xi+1)}, \frac{\tau_0^2/(1/\nu_0-1)}{2A(\xi+1)}, \log\left(\frac{4}{e}\right) \right\}, \; C_2 = \frac{\tau_0^2}{\mu_*}, \; C_1 = \frac{C_2}{A(1-\tau_0^2)}. \tag{36}$$

The main conditions of theorem 4 are

$$\left. \begin{array}{c} \kappa^2(\xi,S) \geqslant c_0, \\ \xi^2/\kappa^2(\xi,S) \leqslant K/\phi_+(m,S), \\ \phi_-(m,S) \geqslant c_1, \end{array} \right\} \tag{37}$$

where m is the smallest integer upper bound of $K|S|$. Whereas theorem 4, part (a), requires only the first inequality in expression (37), theorem 4, part (b), requires all three. Let

$$\mathrm{RE}_2(\xi,S) = \inf\{\|\mathbf{Xu}\|_2/(n^{1/2}\|\mathbf{u}\|_2) : \mathbf{u} \in \mathscr{C}(\xi,S)\},$$
$$F_1(\xi,S) = \inf\{\|\mathbf{X}^{\mathrm{T}}\mathbf{Xu}\|_\infty |S|/(n\|\mathbf{u}_S\|_1) : \mathbf{u} \in \mathscr{C}(\xi,S), u_j \mathbf{x}_j^{\mathrm{T}}\mathbf{Xu} \leqslant 0, j \notin S\}$$

be respectively the restricted eigenvalue and sign-restricted cone invertibility factor for the design matrix. It is worthwhile to note that

$$F_1(\xi,S) \geqslant \kappa^2(\xi,S) \geqslant \mathrm{RE}_2^2(\xi,S) \tag{38}$$

always holds and lower bounds of these quantities can be expressed in terms of sparse eigenvalues (Ye and Zhang, 2010). By Sun and Zhang (2012), we may replace $\kappa^2(\xi,S)$ throughout theorem 4 with $F_1(\xi,S)$. In view of inequality (38), this will actually weaken the condition. However, since more explicit proofs are given in terms of $\kappa(\xi,S)$ in Sun and Zhang (2012), the compatibility factor is used in theorem 4 to facilitate a direct matching of proofs between this paper and Sun and Zhang (2012). By Zhang (2010), condition (37) can be replaced by the sparse Riesz condition,

$$s \leqslant \frac{d^*}{\phi_+(d^*,\emptyset)/\phi_-(d^*,\emptyset) + \frac{1}{2}}. \tag{39}$$

Proposition 2 below provides a way of checking condition (37) for a given design in equation (1).

Proposition 2. Let $\{\xi, M_0, c_*, c^*\}$ be fixed positive constants, $\lambda_1 = M_0\surd\{\log(p)/n\}$ and

$$\hat{\boldsymbol{\Sigma}} = ((\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_k/n)\, I\{|\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_k/n| \geqslant \lambda_1\})_{p\times p}$$

be the thresholded Gram matrix. Suppose that $\phi_{\min}(\hat{\boldsymbol{\Sigma}}) \geqslant c_*$ and $s\lambda_1(1+\xi)^2 \leqslant c_*/2$. Then, for all $|S| \leqslant s$, $\kappa^2(\xi,S) \geqslant c_*/2$. Let $K = 2\xi^2(c^*/c_* + \frac{1}{2})$. If, in addition, $\phi_{\max}(\hat{\boldsymbol{\Sigma}}) \leqslant c^*$ and $s\lambda_1(1+K) + \lambda_1 \leqslant c_*/2$, then $\phi_-(m,S) \geqslant c_*/2$ and condition (37) holds with $c_0 = c_*/2$.

The main condition of proposition 2 is a small $s\surd\{\log(p)/n\}$. This is not restrictive since theorem 1 requires the stronger condition of a small $s\log(p)/\surd n$. It follows from Bickel and Levina (2008) that, after hard thresholding at a level of order λ_1, sample covariance matrices converge to a population covariance matrix in the spectrum norm under mild sparsity conditions on the population covariance matrix. Since convergence in the spectrum norm implies convergence of the minimum and maximum eigenvalues, $\phi_{\min}(\hat{\boldsymbol{\Sigma}}) \geqslant c_*$ and $\phi_{\max}(\hat{\boldsymbol{\Sigma}}) \leqslant c^*$ are reasonable conditions. This and other applications of random-matrix theory are discussed in the next subsection.

3.5. Checking conditions by random-matrix theory

The most basic requirements for our main theoretical results in Sections 3.1 and 3.2 are conditions (20) and (21), and the existence of $\mathbf{z}_j$ with small η_j and τ_j. For deterministic design matrices, sufficient conditions for error bounds (20) and (21) are given in theorem 4 in the form of condition (37), and sufficient conditions for the existence of $\eta_j \leqslant C\surd\log(p)$ and $\tau_j \asymp n^{-1/2}$ are given in proposition 1. These sufficient conditions are all analytical conditions on the design matrix. In this subsection, we use random-matrix theory to check these conditions with more explicit constant factors.

The conditions of theorems 1 and 4 hold for the following classes of design matrices:

$$\begin{aligned}\mathscr{X}_{s,n,p} &= \mathscr{X}_{s,n,p}(c_*, \delta, \xi, K)\\ &= \{\mathbf{X} : \max_{j\leqslant p} \eta_j \leqslant 3\surd\log(p),\ \max_{j\leqslant p} \tau_j^2\sigma_j^2 \leqslant 2/n,\ \min_{|S|\leqslant s} \kappa^2(\xi, S) \geqslant c_*(1-\delta)/4,\\ &\qquad \max_{|S|\leqslant s} \phi_+(m, S)\xi^2/\kappa^2(\xi, S) \leqslant K,\ \min_{|S|\leqslant s} \phi_-(m, S) \geqslant c_*(1-\delta)\}, \qquad (40)\end{aligned}$$

for certain positive $\{s, c_*, \delta, \xi, K\}$, where $\{\eta_j, \tau_j\}$ are computed from $\mathbf{X}$ by the algorithm in Table 2 with $\kappa_0 \leqslant \frac{1}{4}$, and $\kappa(\xi, S)$ and $\phi_\pm(m, S)$ are the compatibility factor and sparse eigenvalues of $\mathbf{X}$ given in expressions (34) and (35), with $m-1 < Ks \leqslant m$.

Let P_Σ be probability measures under which $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_p) \in \mathbb{R}^{n\times p}$ has independent rows with the multivariate normal distribution

$$\tilde{\mathbf{x}}_j \sim N(0, \boldsymbol{\Sigma}). \qquad (41)$$

The column standardized version of $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_p)$ is

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p), \qquad \mathbf{x}_j = \tilde{\mathbf{x}}_j\surd n/\|\tilde{\mathbf{x}}_j\|_2. \qquad (42)$$

Since our discussion is confined to column standardized design matrices for simplicity, we assume without loss of generality that the diagonal elements of $\boldsymbol{\Sigma}$ all equal 1. Under P_Σ, $\mathbf{X}$ does not have independent rows but $\mathbf{x}_j$ is still related to $\mathbf{X}_{-j}$ through

$$\mathbf{x}_j = \mathbf{X}_{-j}\gamma_j + \varepsilon_j\surd n/\|\tilde{\mathbf{x}}_j\|_2, \qquad \varepsilon_j \sim N(0, \sigma_j^2 I_{n\times n}), \qquad (43)$$

where ε_j is independent of $\mathbf{X}_{-j}$. Let Θ_{jk} be the elements of $\boldsymbol{\Sigma}^{-1}$. Since the linear regression of $\tilde{\mathbf{x}}_j$ against $(\tilde{\mathbf{x}}_k, k \neq j)$ has coefficients $-\Theta_{jk}/\Theta_{jj}$ and noise level $1/\Theta_{jj}$, we have

$$\gamma_j = (-\sigma_j^2\Theta_{jk}\|\tilde{\mathbf{x}}_k\|_2/\|\tilde{\mathbf{x}}_j\|_2, k \neq j)^{\mathrm{T}}, \qquad \sigma_j^2 = 1/\Theta_{jj}. \qquad (44)$$

It follows that, when $\phi_{\min}(\boldsymbol{\Sigma}) \geqslant c_*$, $\max_{j\leqslant p}\tau_j^2 \leqslant 2/(nc_*)$ in $\mathscr{X}_{s,n,p}(c_*, \delta, \xi, K)$.

The aim of this subsection is to prove that $P_\Sigma(\mathscr{X}_{s,n,p})$ is uniformly large for a general collection of P_Σ. This result has two interpretations. The first interpretation is that, when $\mathbf{X}$ is indeed generated in accordance with equations (41) and (42), the regularity conditions have a high probability of holding. The second interpretation is that $\mathscr{X}_{s,n,p}$, which is a deterministic subset of $\mathbb{R}^{n\times p}$, is sufficiently large as measured by P_Σ in the collection. Since $\mathscr{X}_{s,n,p}$ does not depend on $\boldsymbol{\Sigma}$ and the probability measures P_Σ are nearly orthogonal for different $\boldsymbol{\Sigma}$, the use of P_Σ does not add the random-design assumption to our results.

The following theorem specifies $\{c_*, c^*, \delta, \xi, K\}$ in expression (40) for which $P_\Sigma\{\mathscr{X}_{s,n,p}(c_*, \delta, \xi, K)\}$ is large when $s\log(p)/n$ is small. This works with the LDPE theory since $s\log(p)/\surd n \to 0$ is required anyway in theorem 1. Define a class of coefficient vectors with small l_q-tail as

$$\mathscr{B}_q(s, \lambda) = \{\mathbf{b} \in \mathbb{R}^p : \sum_{j=1}^{p} \min(|b_j|^q/\lambda^q, 1) \leqslant s\}.$$

We note that $\mathscr{B}_1(s, \sigma, \lambda_{\text{univ}})$ is the collection of all $\boldsymbol{\beta}$ satisfying the capped l_1-sparsity condition (19).

Theorem 5. Suppose that $\mathrm{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_{p\times p}$, eigenvalues $(\boldsymbol{\Sigma}) \subset [c_*, c^*]$, and all rows of $\boldsymbol{\Sigma}^{-1}$ are in $\mathscr{B}_1(s, \lambda_{\text{univ}})$. Then, there are positive numerical constants $\{\delta_0, \delta_1, \delta_2\}$ and K depending only on $\{\delta_1, \xi, c_*, c^*\}$ such that

$$\inf_{(K+1)(s+1)\leqslant \delta_0 n/\log(p)} P_{\boldsymbol{\Sigma}}\{\mathbf{X} \in \mathscr{X}_{s,n,p}(c_*, \delta_1, \xi, K)\} \geqslant 1 - \exp(-\delta_2 n).$$

Consequently, when $\mathbf{X}$ is indeed generated from expressions (41) and (42), all conclusions of theorems 1–3 hold for both estimators (10) and (11) with a probability adjustment of a probability smaller than $2\exp(-\delta_2 n)$, provided that $\boldsymbol{\beta} \in \mathscr{B}_1(s, \sigma\lambda_{\text{univ}})$ and $\lambda_0 = A\surd\{(2/n)\log(p/\epsilon)\}$ in equation (10) with a fixed $A > (\xi+1)/(\xi-1)$.

Remark 5. It follows from theorem II.13 of Davidson and Szarek (2001) that, for certain positive $\{\delta_0, \delta_1, \delta_2\}$,

$$\mathscr{X}'_{n,p} = \{\mathbf{X}: \min_{|S|+m\leqslant\delta_0 n/\log(p)} \phi_-(m,S) \geqslant c_*(1-\delta_1), \max_{|S|+m\leqslant\delta_0 n/\log(p)} \phi_+(m,S) \leqslant c^*(1+\delta_1)\}$$

satisfies $P_{\boldsymbol{\Sigma}}\{\mathscr{X}'_{n,p}\} \geqslant 1 - \exp(-\delta_2 n)$ for all $\boldsymbol{\Sigma}$ in theorem 5 (Candès and Tao, 2005; Zhang and Huang, 2008). Let $K = 4\xi^2(c^*/c_*)(1+\delta_1)/(1-\delta_1)$ and $\{k, l\}$ be positive integers satisfying $4l/k \geqslant K$ and $\max\{k+l, 4l\} \leqslant \delta_0 n/\log(p)$. For $\mathbf{X} \in \mathscr{X}'_{n,p}$, the conditions

$$\kappa(\xi, S) \geqslant \{c_*(1-\delta_1)\}^{1/2}/2,$$
$$\xi^2\phi_+(m,S)/\kappa^2(\xi,S) \leqslant K$$

hold for all $|S| \leqslant k$, where m is the smallest integer upper bound of $K|S|$.

The $P_{\boldsymbol{\Sigma}}$-induced regression model (43) provides a motivation for the use of the lasso in expression (12) and Table 2 to generate score vectors $\mathbf{z}_j$. However, the goal of the procedure is to find $\mathbf{z}_j$ with small η_j and τ_j for controlling the variance and bias of the LDPE (4) as in theorem 1. This is quite different from the usual applications of the lasso for prediction, estimation of regression coefficients or model selection.

4. Simulation results

We set $n = 200$ and $p = 3000$, and run several simulation experiments with 100 replications in each setting. In each replication, we generate an independent copy of $(\tilde{\mathbf{X}}, \mathbf{X}, \mathbf{y})$, where, given a particular $\rho \in (-1, 1)$, $\tilde{\mathbf{X}} = (\tilde{x}_{ij})_{n\times p}$ has independent and identically distributed $N(0, \boldsymbol{\Sigma})$ rows with $\boldsymbol{\Sigma} = (\rho^{|j-k|})_{p\times p}$, $\mathbf{x}_j = \tilde{\mathbf{x}}_j\surd n/|\tilde{\mathbf{x}}_j|_2$, and $(\mathbf{X}, \mathbf{y})$ is as in equation (1) with $\sigma = 1$. Given a particular $\alpha \geqslant 1$, $\beta_j = 3\lambda_{\text{univ}}$ for $j = 1500, 1800, 2100, \ldots, 3000$, and $\beta_j = 3\lambda_{\text{univ}}/j^\alpha$ for all other j, where $\lambda_{\text{univ}} = \surd\{(2/n)\log(p)\}$. This simulation example includes four cases, labelled A, B, C and D, respectively $(\alpha, \rho) = (2, \frac{1}{5})$, $(1, \frac{1}{5})$, $(2, \frac{4}{5})$, $(1, \frac{4}{5})$.

The simulation experiment is designed in the framework of the theory in Section 3. The theory states an asymptotic sample size requirement of $s\log(p)/n^{1/2} \to 0$. However, this condition is a reflection of a conservative bias bound (6) and the compatibility factor (34) of the design. In fact, it is only necessary to have a small $Cs\log(p)/n^{1/2}$, where the factor C refers to a combination of quantities that are treated as constant in the theory. In particular, a smaller C is expected for more orthogonal designs, allowing larger values of $s\log(p)/n^{1/2}$. The simulation experiments explore the behaviour of the LDPE over $(s, s\log(p)/n^{1/2})$ equal to $(8.93, 5.05)$ and $(29.24, 16.55)$ respectively for $\alpha = 2$ and $\alpha = 1$, where $s = \Sigma_j \min(|\beta_j|/\lambda_{\text{univ}}, 1)$. Thus, case A is expected to be the easiest, with the smallest $\{s, s\log(p)/n^{1/2}\}$ and the least correlated design vectors, whereas case D is expected to be the most difficult.

In addition to the lasso with penalty level λ_{univ}, the scaled lasso (10) with penalty level $\lambda_0 = \lambda_{\text{univ}}$ and the scaled lasso–LSE estimator (11), we consider an oracle estimator along with the LDPE (4) and its restricted version derived from equation (16), the RLDPE. The oracle estimator is the least squares estimator of β_j when the β_k are given for all $k \neq j$ except for the three k with the smallest $|k-j|$. It can be written as

$$\begin{aligned} \hat{\beta}_j^{(\text{o})} &= \frac{(\mathbf{z}_j^{(\text{o})})^{\text{T}}}{\|\mathbf{z}_j^{(\text{o})}\|_2^2}\Big(\mathbf{y} - \sum_{k \notin K_j} \mathbf{x}_k \beta_k\Big), \\ \hat{\sigma}^{(\text{o})} &= \|\mathbf{P}^{\perp}_{K_j} \varepsilon\|_2 / \sqrt{n}, \end{aligned} \tag{45}$$

where $K_j = \{j-1, j, j+1\}$ for $1 < j < p$, $K_1 = \{1,2,3\}$, $K_p = \{p-2, p-1, p\}$ and $\mathbf{z}_j^{(\text{o})} = \mathbf{P}^{\perp}_{K_j \backslash \{j\}} \mathbf{x}_j$. Here, $\boldsymbol{P}^{\perp}_{K}$ is the orthogonal projection to the space of n-vectors orthogonal to $\{\mathbf{x}_k, k \in K\}$. Note that the oracular knowledge reduces the complexity of the problem from $(n,p) = (200, 3000)$ to $(n,p) = (200, 3)$, and that the variables $\{\mathbf{x}_k, k \in K_j\}$ also have the highest correlation to $\mathbf{x}_j$. For both the LDPE and the RLDPE, the scaled lasso–LSE (11) is used to generate $\hat{\boldsymbol{\beta}}^{(\text{init})}$ and $\hat{\sigma}$, whereas the algorithm in Table 2, with $\kappa_0 = \frac{1}{4}$, is used to generate $\mathbf{z}_j$. The default $\eta_j^* = \surd\{2\log(p)\}$ passed the test in step 2 of Table 2 without adjustment in all instances during the simulation study. This guarantees $\eta_j \leqslant \surd\{2\log(p)\}$ for the bias factor. For the RLDPE, $m = 4$ is used in equation (16).

The asymptotic normality of the LDPE holds well in our simulation experiments. Table 3 and Fig. 1 demonstrate the behaviour of the LDPE and RLDPE for the largest β_j, compared with that of the other estimation methods. The scaled lasso has more bias and a larger variance than the lasso but is entirely data driven. The bias can be significantly reduced through the scaled lasso–LSE method; however, error resulting from failure to select some maximal β_j remains. This is clearest in the histograms corresponding to the distribution of errors for the scaled lasso–LSE method in settings B and D, where $\alpha = 1$ and the β_j decay at a slower rate. For a small increase in variance, the LDPE and RLDPE further reduce the bias of the scaled lasso–LSE method. This is also the case when $\hat{\boldsymbol{\beta}}^{(\text{init})}$ is a heavily biased estimator such as the lasso or

Table 3. Summary statistics for various estimates of the maximal $\beta_j = |\beta|_\infty$: the lasso, the scaled lasso, the scaled lasso–LSE method, the oracle estimator, the LDPE and the RLDPE

Setting	*Statistic*	*Results for the following estimators:*					
		Lasso	*Scaled lasso*	*Scaled lasso–LSE*	*Oracle*	*LDPE*	*RLDPE*
A	Bias	−0.2965	−0.4605	−0.0064	−0.0045	−0.0038	−0.0028
	Standard deviation	0.0936	0.1360	0.1004	0.0730	0.0860	0.0960
	Median absolute error	0.2948	0.4519	0.0549	0.0507	0.0531	0.0627
B	Bias	−0.2998	−0.5341	−0.0476	0.0049	−0.0160	−0.0167
	Standard deviation	0.1082	0.1590	0.2032	0.0722	0.1111	0.1213
	Median absolute error	0.2994	0.5150	0.0693	0.0500	0.0705	0.0799
C	Bias	−0.3007	−0.4423	−0.0266	−0.0049	−0.0194	−0.0181
	Standard deviation	0.1207	0.1520	0.1338	0.1485	0.1358	0.1750
	Median absolute error	0.3000	0.4356	0.0657	0.0994	0.0902	0.1150
D	Bias	−0.3258	−0.5548	−0.1074	−0.0007	−0.0510	−0.0405
	Standard deviation	0.1367	0.1844	0.2442	0.1455	0.1768	0.2198
	Median absolute error	0.3319	0.5620	0.0857	0.0955	0.1112	0.1411

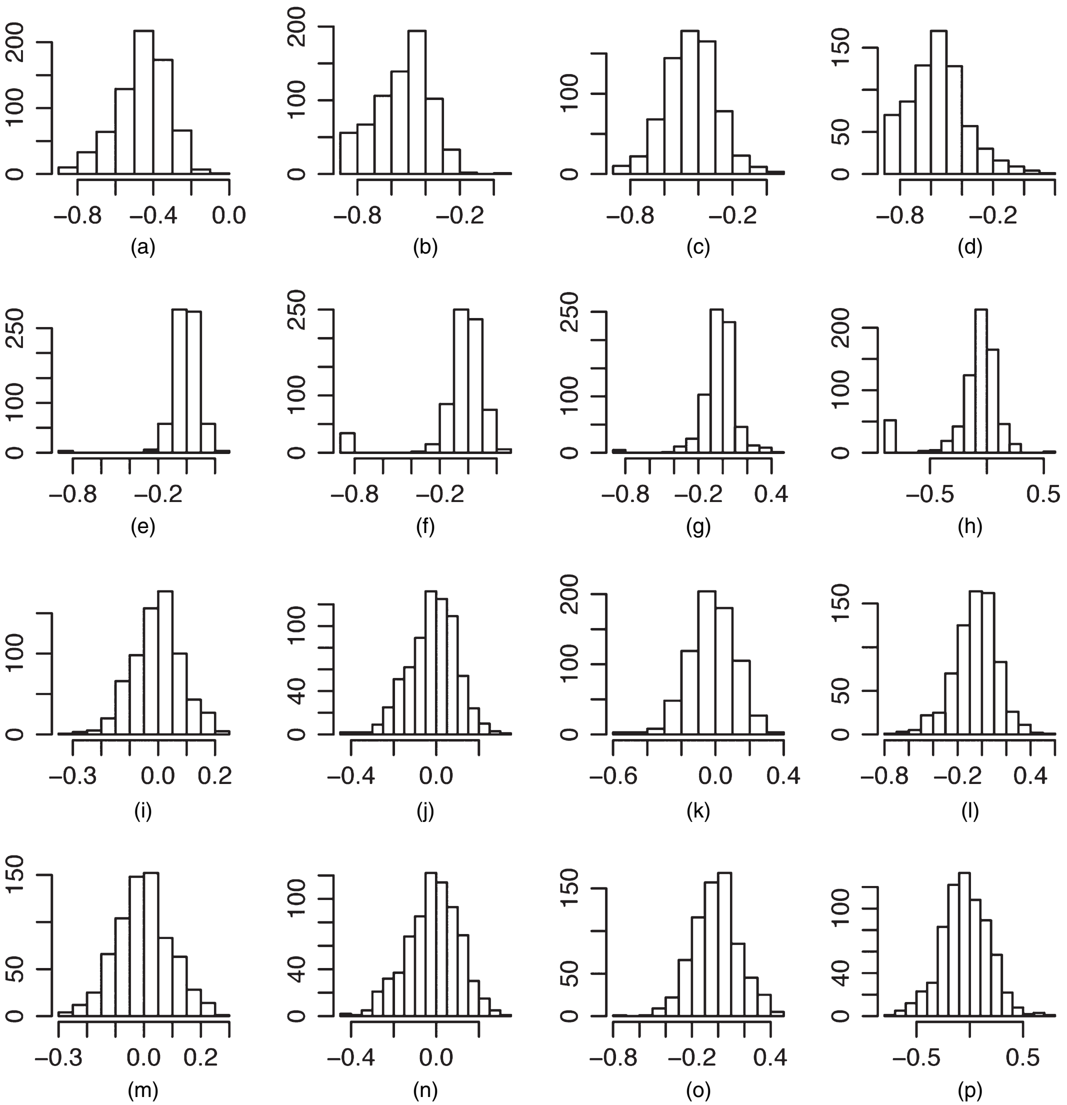

Fig. 1. Histogram of errors when estimating maximal β_j by using (a)–(d) the scaled lasso, (e)–(h) the scaled lasso–LSE method, (i)–(l) the LDP and (m)–(p) the RLDPE: (a), (e), (i), (m) simulation setting A; (b), (f), (j), (n) simulation setting B; (c), (g), (k), (o) simulation setting C; (d), (h), (l), (p) simulation setting D

scaled lasso, and the improvement is most dramatic when estimating large β_j. Although the asymptotic normality of the LDPE holds even better for small β_j in the simulation study, a parallel comparison for small β_j is not meaningful; the lasso typically estimates small β_j by 0, whereas the raw LDPE is not designed to be sparse.

The overall coverage probability of the LDPE-based confidence interval matches relatively well the preassigned level, as expected from our theoretical results. The LDPE and RLDPE create confidence intervals $\hat{\beta}_j \pm 1.96\hat{\sigma}\tau_j$ with approximately 95% coverage in settings A and C and somewhat higher coverage probability in B and D. Refer to Table 4 for precise values. Since the coverage probabilities for each individual β_j are calculated on the basis of a sample of 100

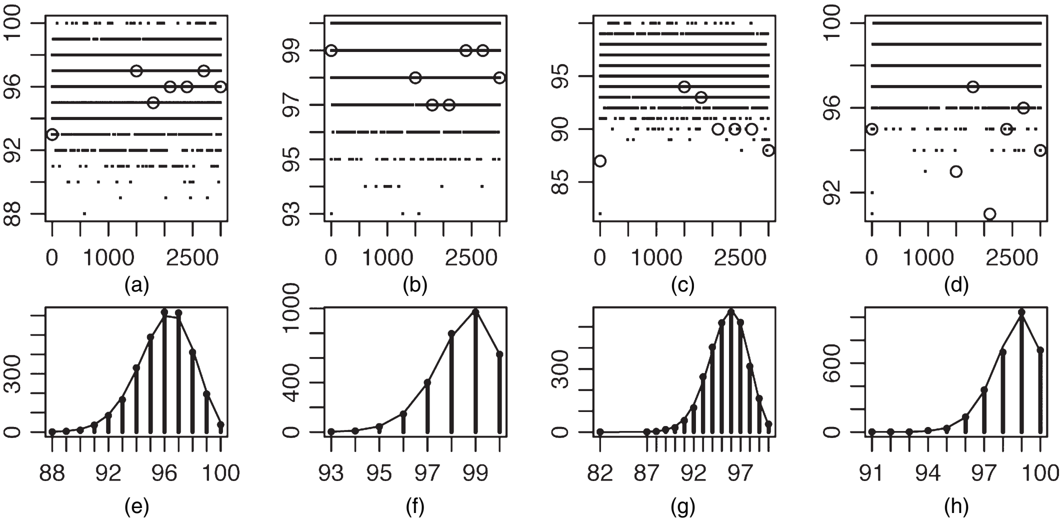

Fig. 2. (a)–(d) Coverage frequencies of the LDPE confidence intervals *versus* the index of β_j (○, maximal β_j) and (e)–(h) the number of variables for given values of the relative coverage frequency, superimposed on the binomial(100, $\tilde{p}$) probability mass function, where $\tilde{p}$ is the simulated mean coverage for the LDPE: (a), (e) simulation setting A; (b), (f) simulation setting B; (c), (g) simulation setting C; (d), (h) simulation setting D

Table 4. Mean coverage probability of the LDPE and RLDPE

		Coverage probabilities for the following settings:			
		A	*B*	*C*	*D*
All β_j	LDPE	0.9597	0.9845	0.9556	0.9855
	RLDPE	0.9595	0.9848	0.9557	0.9885
Maximal β_j	LDPE	0.9571	0.9814	0.9029	0.9443
	RLDPE	0.9614	0.9786	0.9414	0.9786

replications, the empirical distribution of the simulated relative coverage frequencies exhibits some randomness, which matches that of the binomial $(n, \tilde{p})$ distribution, with $n = 100$ and $\tilde{p}$ equal to the simulated mean coverage, as shown in Figs 2 and 3.

Two separate issues may lead to some variability in the coverage. As with settings B and D, the overall coverage may exceed the stated confidence level when the presence of many small signals in β is interpreted as noise, increasing $\hat{\sigma}$ and hence the width of the confidence intervals, along with the coverage; however, this phenomenon will not result in undercoverage. In addition, compared with the overall coverage probability, the coverage probability is somewhat smaller when large values of β_j are associated with highly correlated columns of $\mathbf{X}$. This is most apparent when plotting coverage *versus* index in settings C and D, which are the two settings with higher correlation between adjacent columns of $\mathbf{X}$. For additional clarity, the points corresponding to maximal values of β_j in Figs 2 and 3 are emphasized by larger circles, and the coverage of the LDPE and RLDPE for maximal β_j are listed separately from the overall coverage in the last two rows of Table 4. It can be seen from these details that the RLDPE (16) further eliminates

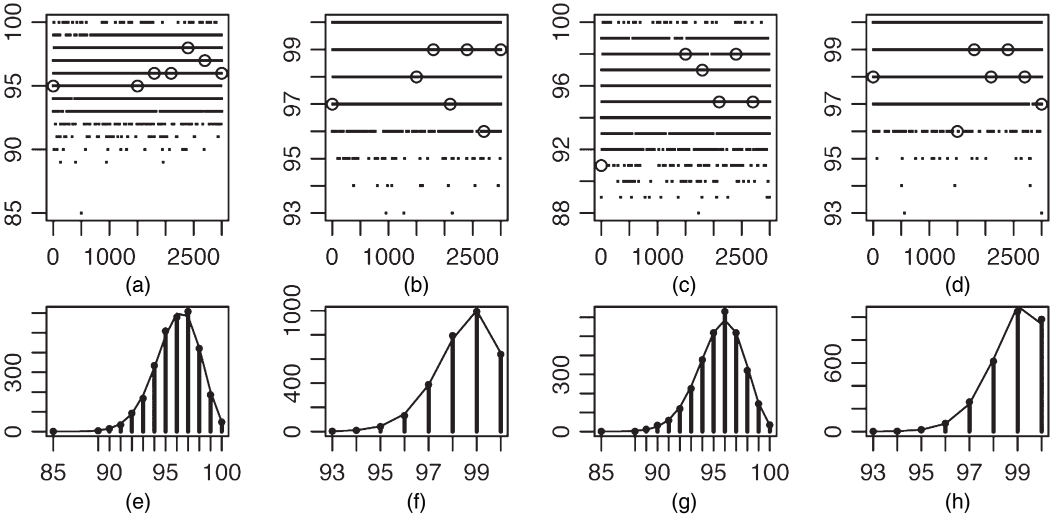

Fig. 3. (a)–(d) Coverage frequencies of the RLDPE confidence intervals *versus* the index of β_j ($\circ$, maximal β_j) and (e)–(h) the number of variables for given values of the relative coverage frequency, superimposed on the binomial(100, $\tilde{p}$) probability mass function, where $\tilde{p}$ is the simulated mean coverage for the RLDPE: (a), (e) simulation setting A; (b), (f) simulation setting B; (c), (g) simulation setting C; (d), (h) simulation setting D

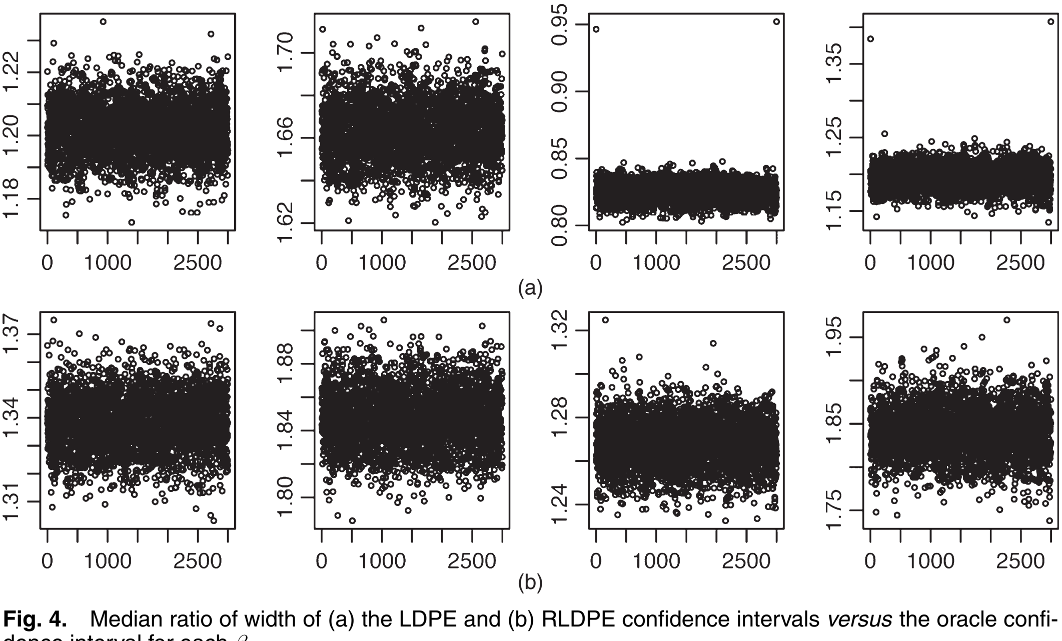

Fig. 4. Median ratio of width of (a) the LDPE and (b) RLDPE confidence intervals *versus* the oracle confidence interval for each β_j

the bias that is caused by the association of relatively large values of β_j with highly correlated columns of $\mathbf{X}$ and improves coverage probabilities. The bias correction effect is also visible in the histograms in Fig. 1 in setting D, but not in C.

The LDPE and RLDPE confidence intervals are of reasonable width, comparable with that of the confidence intervals that are derived from the oracle estimator. Consider the median ratio between the width of the LDPE (and RLDPE) confidence intervals and that of the oracle confidence intervals, which are shown in Fig. 4. The distribution of the median ratio that is associated with each β_j is uniform over the different $j=1,\ldots,3000$ in settings A and B. The anomalies at $j=1$ and $j=3000$ in settings C and D are a result of the structure of $\mathbf{X}$. When the correlation between nearby columns of $\mathbf{X}$ is high, the fact that the first and last columns of $\mathbf{X}$ have fewer highly correlated neighbours gives the oracle a relatively greater advantage. Since the medians of the ratios are uniformly distributed over j, it is reasonable to summarize the ratios in each simulation setting with the median value over every replication of every β_j, as listed in Table 5. Note that the LDPE is more efficient than the oracle estimator in the high correlation settings C and D. This is probably due to the benefit of relaxing the orthogonality constraint of $\mathbf{x}_j^{\perp}$ when the correlation of the design is high and the error of the initial estimator is relatively small. The median ratio between the widths of the LDPE and oracle confidence intervals reaches its highest value of 1.6400 in setting B, where the coverage of the LDPE intervals is high and the benefit of relaxing the orthogonality constraint is small, if any, relative to the oracle.

Recall that the RLDPE improves the coverage probability for large β_j at the cost of an increase in the variance of the estimator; thus, the RLDPE confidence intervals are somewhat wider than the LDPE confidence intervals. Although the improvement in coverage probability is focused on the larger values of β_j, all β_j are affected by the increase in variance and confidence interval width.

We may also consider the performance of the LDPE as a point estimator. Table 6 and Fig. 5 compare the mean-squared errors of the LDPE and RLDPE estimators of β_j with that of the oracle estimator of β_j. This comparison is consistent with the comparison of the median widths of the confidence intervals in Table 5 and Fig. 4 that was discussed earlier.

The lasso and scaled lasso estimators have larger biases for larger values of β_j but perform very well for smaller values. In contrast, the LDPE and the oracle estimators are not designed to be sparse and have very stable errors over all β_j. For the estimation of the entire vector $\boldsymbol{\beta}$ or its support, it is appropriate to compare a thresholded LDPE with the lasso, the scaled lasso, the scaled lasso–LSE method and a matching thresholded oracle estimator. Hard thresholding

Table 5. Medians of the width ratio medians in Fig. 4

Setting	*LDPE*	*RLDPE*
A	1.2020	1.3359
B	1.6400	1.8238
C	0.8209	1.2678
D	1.1758	1.8150

Table 6. Medians of the mean-squared error ratios in Fig. 5

Setting	*LDPE*	*RLDPE*
A	0.7551	0.6086
B	0.5232	0.4232
C	1.5950	0.6656
D	1.1169	0.5049

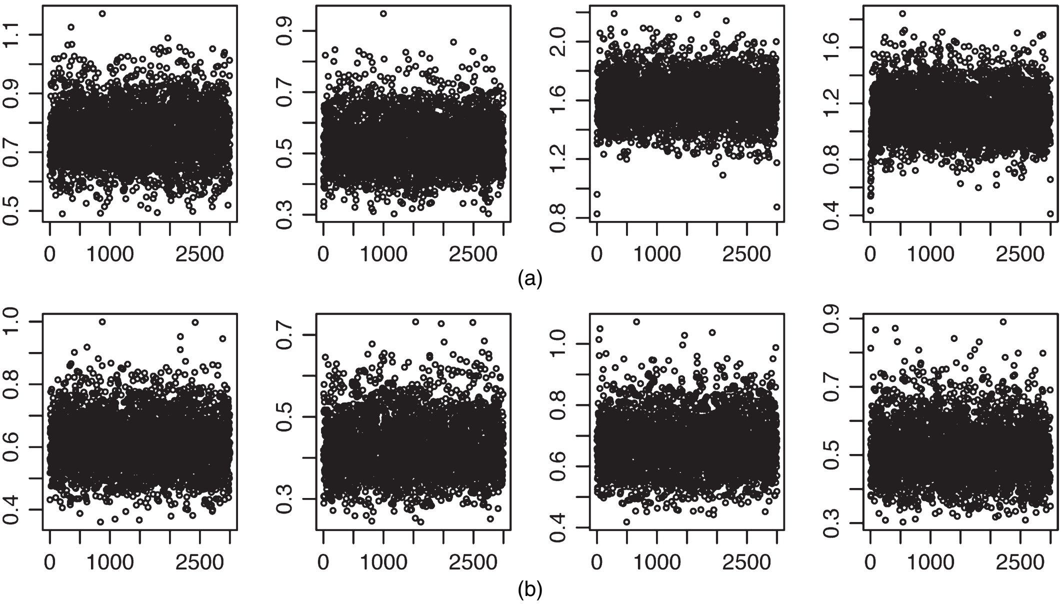

Fig. 5. Efficiency (the ratio of the MSEs) of (a) the LDPE and (b) RLDPE estimators *versus* the oracle estimator for each β_j

Table 7. Summary statistics for the l_2-loss of five estimators of β: the lasso, the scaled lasso, the scaled lasso–LSE, the thresholded oracle estimator and the thresholded LDPE

Setting	*Statistic*	*Results for the following estimators:*				
		Lasso	*Scaled lasso*	*Scaled lasso–LSE*	*Thresholded oracle*	*Thresholded LDPE*
A	Mean	0.8470	1.2706	0.3288	0.3624	0.3621
	Standard deviation	0.1076	0.2393	0.1465	0.0908	0.1884
	Median	0.8252	1.2131	0.3042	0.3577	0.3312
B	Mean	0.9937	1.5837	0.7586	0.5658	0.7969
	Standard deviation	0.1214	0.2624	0.2976	0.0615	0.3873
	Median	0.9820	1.5560	0.6219	0.5675	0.6983
C	Mean	0.8836	1.2411	0.4817	0.6803	0.5337
	Standard deviation	0.1402	0.2208	0.2083	0.2843	0.2164
	Median	0.8702	1.2295	0.4343	0.6338	0.4642
D	Mean	1.0775	1.6303	1.0102	0.9274	1.2627
	Standard deviation	0.1437	0.2381	0.3572	0.2342	0.5576
	Median	1.0570	1.6389	0.9216	0.8716	1.1011

was implemented: $\hat{\beta}_j I(|\hat{\beta}_j| \leqslant \hat{t}_j)$ for the thresholded LDPE with $\hat{t}_j = \hat{\sigma}\tau_j \Phi^{-1}\{1-1/(2p)\}$ and $\hat{\beta}_j^{(\mathrm{o})} I(|\hat{\beta}_j^{(\mathrm{o})}| \leqslant \hat{t}_j^{(\mathrm{o})})$ for the thresholded oracle with $\hat{t}_j^{(\mathrm{o})} = \hat{\sigma}^{(\mathrm{o})} \|\mathbf{z}_j^{(\mathrm{o})}\|_2^{-1} \Phi^{-1}\{1-1/(2p)\}$, where $\{\hat{\beta}_j^{(\mathrm{o})}, \hat{\sigma}^{(\mathrm{o})}, \mathbf{z}_j^{(\mathrm{o})}\}$ are as in expression (45). Since $\beta_j \neq 0$ for all j, the comparison is confined to the l_2-estimation error. Table 7 lists the mean, standard deviation and median of the l_2-loss of these five estimators over 100 replications. Of the five estimators, only the scaled lasso, the scaled lasso–LSE estimator and the thresholded LDPE are purely data driven. The performance of the scaled lasso–LSE, the thresholded LDPE and the thresholded oracle methods are comparable and they always outperform the scaled lasso. They also outperform the lasso in cases A, B and C. In the hardest case, D, which has both a high correlation between adjacent columns of $\mathbf{X}$ and a slower decay in β_j, the thresholded oracle slightly outperforms the lasso and the lasso slightly outperforms the scaled lasso–LSE and thresholded LDPE methods. Generally, the l_2-loss of the thresholded LDPE remains slightly above that of the scaled lasso–LSE method, which improves on the scaled lasso by reducing its bias. Note that our goal is not to find a better estimator for the entire vector $\boldsymbol{\beta}$ since quite a few versions of the estimation optimality of regularized estimators have already been established. What we demonstrate here is that the cost of removing bias with the LDPE, and thus giving up shrinkage, is small.

5. Discussion

We have developed the LDPE method of constructing $\hat{\beta}_1, \ldots, \hat{\beta}_p$ for the individual regression coefficients and estimators for their finite dimensional covariance structure. Under proper conditions on $\mathbf{X}$ and $\boldsymbol{\beta}$, we have proved the asymptotic unbiasedness and normality of finite dimensional distribution functions of these estimators and the consistency of their estimated covariances. Thus, the LDPE yields an approximate Gaussian sequence as the raw LDPE in expression (25), which allows us to assess the level of significance of each unknown coefficient β_j without the uniform signal strength assumption (26). The method proposed applies to making inference about a preconceived low dimensional parameter, which is an interesting practical problem and a primary goal of this paper. It also applies to making inference about all regression coefficients via simultaneous interval estimation and the correct selection of large and zero coefficients in the presence of many small coefficients.

The raw LDPE is not sparse, but it can be thresholded to take advantage of the sparsity of $\boldsymbol{\beta}$, and the sampling distribution of the thresholded LDPE can still be bounded on the basis of the approximate distribution of the raw LDPE. A thresholded LDPE is proven to attain l_2-rate optimality for the estimation of an entire sparse $\boldsymbol{\beta}$.

The focus of this paper is interval estimation and hypothesis testing without the uniform signal strength condition. Another important problem is prediction. Since prediction at a design point $\mathbf{a}$ is equivalent to the estimation of the 'contrast' $\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}$, with possibly large $\|\mathbf{a}\|_0$, the implication of the LDPE for prediction is an interesting future research direction.

The proposed LDP approach is closely related to semiparametric inference. This connection was discussed in Zhang (2011) along with a definition of the minimum Fisher information and a rationale for the asymptotic efficiency of an LDPE in a general setting.

We use the lasso to provide a relaxation of the projection of $\mathbf{x}_j$ to $\mathbf{x}_j^{\perp}$. This choice is primarily due to our familiarity with the computation of the lasso and the readily available scaled lasso method of choosing a penalty level. We have also considered some other methods of relaxing the projection. A particularly interesting method is the following constrained minimization of the variance of the noise term in equation (5):

$$\mathbf{z}_j = \underset{\mathbf{z}}{\arg\min}\{\|\mathbf{z}\|_2^2 : |\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j| = n, \max_{k\neq j} |\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k/n| \leqslant \lambda_j'\}. \tag{46}$$

Similarly to the lasso in equation (9), equation (46) can be solved via quadratic programming; specifically, one may take the minimizer between the solutions involving the two linear constraints $\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j = \pm n$. The lasso solution (9) is feasible in equation (46) with $\lambda_j n/|\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j| = \lambda_j'$.

Acknowledgements

The authors thank a referee and the Associate Editor for their careful reading of the manuscript and for suggestions leading to several improvements. The research of the first author is partially supported by National Science Foundation grants DMS 0906420, DMS 1106753 and DMS 1209014, and National Security Agency grant H98230-11-1-0205.

References

Antoniadis, A. (2010) Comments on: l_1-penalization for mixture regression models. *Test*, **19**, 257–258.

Belloni, A., Chernozhukov, V. and Wang, L. (2011) Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, **98**, 791–806.

Berk, R., Brown, L. B. and Zhao, L. (2010) Statistical inference after model selection. *J. Quant. Crimin.*, **26**, 217–236.

Bickel, P. J. and Levina, E. (2008) Regularized estimation of large covariance matrices. *Ann. Statist.*, **36**, 199–227.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.

Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. New York: Springer.

Candès, E. J. and Tao, T. (2005) Decoding by linear programming. *IEEE Trans. Inform. Theor.*, **51**, 4203–4215.

Candès, E. and Tao, T. (2007) The dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.*, **35**, 2313–2404.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (2001) Atomic decomposition by basis pursuit. *SIAM Rev.*, **43**, 129–159.

Davidson, K. and Szarek, S. (2001) Local operator theory, random matrices and Banach spaces. In *Handbook on the Geometry of Banach Spaces*, vol. 1 (eds W. B. Johnson and J. Lindenstrauss). Amsterdam: North-Holland.

Donoho, D. L. and Johnstone, I. (1994) Minimax risk over l_p-balls for l_q-error. *Probab. Theor. Reltd Flds*, **99**, 277–303.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.

Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. *Statist. Sin.*, **20**, 101–148.

Fan, J. and Peng, H. (2004) On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.*, **32**, 928–961.

Fano, R. (1961) *Transmission of Information; a Statistical Theory of Communications*. Cambridge: Massachusetts Institute of Technology Press.

Frank, I. E. and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109–148.

van de Geer, S. and Bühlmann, P. (2009) On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.*, **3**, 1360–1392.

Greenshtein, E. (2006) Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.*, **34**, 2367–2386.

Greenshtein, E. and Ritov, Y. (2004) Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, **10**, 971–988.

Huang, J., Ma, S. and Zhang, C.-H. (2008) Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sin.*, **18**, 1603–1618.

Huang, J. and Zhang, C.-H. (2012) Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *J. Mach. Learn. Res.*, **13**, 1809–1834.

Kim, Y., Choi, H. and Oh, H.-S. (2008) Smoothly clipped absolute deviation on high dimensions. *J. Am. Statist. Ass.*, **103**, 1665–1673.

Koltchinskii, V. (2009) The dantzig selector and sparsity oracle inequalities. *Bernoulli*, **15**, 799–828.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, **39**, 2302–2329.

Laber, E. and Murphy, S. A. (2011) Adaptive confidence intervals for the test error in classification (with discussion). *J. Am. Statist. Ass.*, **106**, 904–913.

Leeb, H. and Potscher, B. M. (2006) Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.*, **34**, 2554–2591.

Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.

Meinshausen, N. and Yu, B. (2009) Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, **37**, 246–270.

Städler, N., Bühlmann, P. and van de Geer, S. (2010) l_1-penalization for mixture regression models (with discussion). *Test*, **19**, 209–285.

Sun, T. and Zhang, C.-H. (2010) Comments on: l_1-penalization for mixture regression models. *Test*, **19**, 270–275.

Sun, T. and Zhang, C.-H. (2012) Scaled sparse linear regression. *Biometrika*, **99**, 879–898.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Tropp, J. A. (2006) Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theor.*, **52**, 1030–1051.

Wainwright, M. J. (2009a) Sharp thresholds for noisy and high-dimensional recovery of sparsity using l_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theor.*, **55**, 2183–2202.

Wainwright, M. J. (2009b) Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theor.*, **55**, 5728–5741.

Ye, F. and Zhang, C.-H. (2010) Rate minimaxity of the Lasso and Dantzig selector for the l_q loss in l_r balls. *J. Mach. Learn. Res.*, **11**, 3481–3502.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.

Zhang, C.-H. (2011) Statistical inference for high-dimensional data. In *Very High Dimensional Semiparametric Models, Report No. 48/2011*, pp. 28–31. Mathematisches Forschungsinstitut Oberwolfach.

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

Zhang, C.-H. and Zhang, S. S. (2011) Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Preprint arXiv:1110.2563*.

Zhang, C.-H. and Zhang, T. (2012) A general theory of concave regularization for high dimensional sparse estimation problems. *Statist. Sci.*, **27**, 576–593.

Zhang, T. (2009) Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.*, **37**, no. 5A, 2109–2144.

Zhang, T. (2011a) Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Trans. Inform. Theor.*, **57**, 4689–4708.

Zhang, T. (2011b) Multi-stage convex relaxation for feature selection. *Preprint arXiv:1106.0565*.

Zhao, P. and Yu, B. (2006) On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2541–2567.

Zou, H. (2006) The adaptive Lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509–1533.

Supporting information

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

‘Supplement material for Confidence intervals for low-dimensional parameters in high-dimensional linear models’.