Journal of Econometrics 186 (2015) 317–324

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

On the residual empirical process based on the ALASSO in high dimensions and its functional oracle property ☆

A. Chatterjee [a], S. Gupta [b], S.N. Lahiri [b,*]

[a] Stat-Math Unit, Indian Statistical Institute, New Delhi, India

[b] Department of Statistics, North Carolina State University, Raleigh, NC, USA

ARTICLE INFO

Article history:
Available online 28 February 2015

Keywords:
Asymptotic uniform linearity
Brownian bridge
Oracle property
Prediction intervals
Regularization
Weak convergence

ABSTRACT

This paper considers post variable-selection inference in a high dimensional penalized regression model based on the ALASSO method of Zou (2006). It is shown that under suitable sparsity conditions, the residual empirical process based on the ALASSO provides valid inference methodology in very high dimensional regression problems where conventional methods fail. It is also shown that the ALASSO based residual empirical process satisfies a *functional oracle property*, i.e., in addition to selecting the set of relevant variables with probability tending to one, the ALASSO based residual empirical process converges to the same limiting Gaussian process as the OLS based residual empirical process under the oracle. The functional oracle property is critically exploited to construct asymptotically valid confidence bands for the error distribution function and prediction intervals for unobserved values of the response variable in the high dimensional set up, where traditional non-penalized methods are known to fail. Simulation results are presented illustrating finite sample performance of the proposed methodology.

1. Introduction

Consider a high dimensional linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where, y_i is the response, $\mathbf{x}_i = \left(x_{i,1}, \ldots, x_{i,p}\right)'$ is a p dimensional (nonrandom) vector of covariates, $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)'$ is the regression parameter and $\{\epsilon_i : i = 1, \ldots, n\}$ are independent and identically distributed (iid) error variables with zero mean and common distribution function F, and where A' denotes the transpose of a matrix A. Here, the dimension of the regression parameter (or equivalently, the number of covariates) $p = p_n$ is allowed to increase with the sample size n and it is assumed that the true regression parameter has sparsity in the sense that only $1 \le p_0 < p$ of the regression parameters are nonzero, where $p_0 = p_{0n}$ may also tend to infinity with n. For notational simplicity, without loss of generality (w.l.o.g.), suppose that the first p_0 components of the true regression parameter $\boldsymbol{\beta}$ are non-zero and the last $(p-p_0)$ components are zero.

A popular method for simultaneous variable selection and inference in such high dimensional regression models is given by the Adaptive LASSO (or ALASSO) method of Zou (2006). Specifically, the ALASSO estimator of $\boldsymbol{\beta}$ is defined as the minimizer of the weighted ℓ_1-penalized least squares criterion function,

$$\widehat{\boldsymbol{\beta}}_n = \underset{\mathbf{u}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\mathbf{u}\right)^2 + \lambda_n \sum_{j=1}^{p} \frac{|u_j|}{|\tilde{\beta}_{j,n}|^{\gamma}}, \tag{1.2}$$

where, $\lambda_n > 0$ is a regularization parameter, $\gamma > 0$ and $\tilde{\beta}_{j,n}$ is the jth component of $\widetilde{\boldsymbol{\beta}}_n$, a consistent, preliminary estimator of $\boldsymbol{\beta}$. For example, when $p \le n$ and the $n \times p$ design matrix X (with its ith row given by $\mathbf{x}_i'$, $i = 1, \ldots, n$) is of full rank, one may choose $\widetilde{\boldsymbol{\beta}}_n$ as the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$. The ALASSO is an improvement over its precursors, the LASSO and related bridge estimators that often require strong regularity conditions on the design vectors $\mathbf{x}_i$'s for consistent variable selection and that have non-trivial bias in the selected nonzero components (cf. Knight and Fu, 2000; Fan and Li, 2001; Yuan and Lin, 2007; Zhao and Yu, 2006). Indeed, Zou (2006) showed that under some mild regularity conditions, for fixed p,

$$\lim_{n\to\infty} \boldsymbol{P}\left(\{j : \widehat{\beta}_{j,n} \ne 0\} = \{1, \ldots, p_0\}\right) = 1, \tag{1.3}$$

☆ Research partially supported by NSF grant nos. DMS-1007703 and DMS 1310068 and NSA grant no. H98230-11-1-0130.

* Corresponding author.
E-mail address: snlahiri@stat.tamu.edu (S.N. Lahiri).

http://dx.doi.org/10.1016/j.jeconom.2015.02.012
0304-4076/

and that the asymptotic distribution of the ALASSO estimators of the selected variables satisfies:

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}^{(1)}\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2 \mathbf{C}_{11}^{-1}\right), \tag{1.4}$$

where $\widehat{\boldsymbol{\beta}}_n^{(1)} = \left(\widehat{\beta}_{1,n}, \ldots, \widehat{\beta}_{p_0,n}\right)$, $\boldsymbol{\beta}^{(1)} = \left(\beta_1, \ldots, \beta_{p_0}\right)$ and $\boldsymbol{C}_{11}$ is the upper left $p_0 \times p_0$ submatrix of $\boldsymbol{C} \equiv \lim_{n\to\infty} n^{-1} \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'$. Note that if we had *a priori* knowledge that $\beta_{p_0+1} = \cdots = \beta_p = 0$ in (1.1), the asymptotic distribution of the OLS estimator of $\boldsymbol{\beta}^{(1)}$ coincides with the limit distribution in (1.4). Thus, in the parlance of Fan and Li (2001), the ALASSO method enjoys the oracle property, *i.e.*, it can correctly identify the set of non-zero components of $\boldsymbol{\beta}$, with probability tending to 1 (cf. (1.3)) and at the same time, estimate the non-zero components accurately, with the same precision as that of the OLS method, in the limit (cf. (1.4)). Extensions of the Oracle property to the case of unbounded p at varying levels of generality are also known; see Huang et al. (2008b), Chatterjee and Lahiri (2013) and the references therein.

Beyond estimation of the non-zero regression parameters, estimation of the error distribution function (d.f.) F is an important problem as it forms the starting point and the basis for a number of inference procedures under the regression model (1.1). There is a well developed body of literature on the empirical process of residuals, both for the case when p is finite (Koul, 1969, 2002) and when p is diverging (Portnoy, 1984, 1985; Mammen, 1996; Chen and Lockhart, 2001). A key result in this context is the uniform asymptotic linear (AUL) representation of the residual empirical process which says that the residual empirical distribution function (edf) can be approximated *uniformly* by the sum of the (unobservable) edf

$$F_n(\cdot) = n^{-1} \sum_{i=1}^{n} \mathbb{1}(\epsilon_i \leq \cdot) \tag{1.5}$$

of the error variables and a (linear) term depending on the normalized estimator of $\boldsymbol{\beta}$, with negligible asymptotic error. Here and in the following, $\mathbb{1}(\cdot)$ denotes the indicator function. However, the best possible rate of p as a function of n for which the AUL representation is known to hold is $p = O(n^{1/3})$ for general designs and $p = o(n^{1/2})$ under some strong conditions on the design vectors. Indeed, Chen and Lockhart (2001) (hereafter referred to as [CL01]) give an example to show that one cannot improve upon the growth rate $p = O(n^{1/3})$, in general.

In this paper, we show that under some regularity conditions, the ALASSO based residual edf enjoys the AUL property in presence of sparsity even when $p \gg n$. This is an artifact of the Oracle property of the ALASSO estimator which enables it to screen out all irrelevant variables as a part of its consistent variable selection feature, thereby drastically reducing the high dimensionality of the original problem to a manageable level. In view of the existing bounds on the growth rate of the dimension p as a function of n, the results here (cf. Theorem 1) extend the AUL property to very high dimensions that do not seem achievable using commonly used (unpenalized) estimators of the full regression vector $\boldsymbol{\beta}$.

As an application of the AUL representation, we establish *Functional Oracle Property* of the ALASSO method. Specifically, we show that in addition to identifying the set of relevant variables with probability tending to one, the ALASSO-based residual empirical process converges to the *same* Gaussian process on $L^\infty([-\infty, \infty])$ as the residual empirical process based on the OLS in the reduced model under the oracle. Thus, the ALASSO enjoys the 'Oracle property' of Fan and Li (2001) in the more general context of estimating a functional parameter than just the estimation of finitely many linear combinations of the $\boldsymbol{\beta}$-vector.

As a second application, we use the ALASSO based empirical process to construct asymptotically valid prediction intervals (PIs) for unobserved values of the response variable. In this case, a naive application of the edf based on the OLS (and related) estimators of the *full* vector $\boldsymbol{\beta}$ would lead to PIs that may not achieve the target coverage probability even in large samples. Indeed, the simulation results from Section 4 show that even for the normal error distribution, the empirical coverage probabilities of the OLS based PIs in the full model perform very poorly at all levels of the sample size (at $n = 60$ and at $n = 200$) whenever the dimension p becomes large (e.g., in the case '$p = 50$ for $n = 60$', among others). In contrast, the empirical coverage probabilities of the ALASSO based PIs in high dimensions are reasonably close to the nominal levels even for a sample size $n = 60$ for normal and non-normal error distributions. As a result, the ALASSO based inference has definite advantages over non-penalized methods in high-dimensional regression models in presence of sparsity.

We conclude this section with a brief literature review. The literature on penalized regression in high dimensions has been growing very rapidly in recent years and here we shall limit ourselves to only a modest account of the work that is most related to the present paper, due to space considerations. In a pioneering work, Tibshirani (1996) introduced the LASSO as an estimation and variable selection method in regression models, using ℓ_1-based penalization. In a seminal work, Zou (2006) introduced the ALASSO method as an improvement over the LASSO and established its variable selection consistency property and asymptotic distribution theory under weaker conditions on the design matrix. Other popular penalized estimation and variable selection methods are given by the SCAD of Fan and Li (2001) and the MCP method of Zhang (2010), among others. Properties of the ALASSO and the related methods have been investigated by many authors, including Knight and Fu (2000), Meinshausen and Bühlmann (2006), Wainwright (2009), Bunea et al. (2007), Bickel et al. (2009), Huang et al. (2008b), Huang et al. (2008a), Zhang and Huang (2008), Meinshausen and Yu (2009), Pötscher and Schneider (2009), Chatterjee and Lahiri (2011b) and Gupta (2012), among others. Fan and Li (2001) introduced the important notion of "oracle property" in the context of penalized estimation and variable selection by the SCAD. Belloni et al. (2011a) give an authoritative introduction to high dimensional problems in Econometric models. For results on post variable selection inference, see Leeb and Pötscher (2008), Bach (2009), Chatterjee and Lahiri (2011a), Belloni et al. (2011b), Gautier and Tsybakov (2011), Liao and Phillips (2012), and the references therein. For a book-length account of the inference issues and recent results in the high dimensional set up, see the recent monograph of Bühlmann and van de Geer (2011).

The rest of the paper is organized as follows. Section 2 sets up the notation and states the conditions. Section 3 describes the main theoretical findings. Simulation results are reported in Section 4, followed by a set of concluding remarks in Section 5. Proofs of all the results are given in the final section.

2. Background and conditions

2.1. Notation and conventions

Note that as we allow the dimension $p = p_n$ of the regression parameters to diverge with n, the quantities, like $\mathbf{x}_i$, $\boldsymbol{\beta}$ and the number p_0 of nonzero components of $\boldsymbol{\beta}$, all depend on n. However, we often suppress the dependence on n to ease notation. As described in Section 1, without loss of generality (w.l.o.g.), let $I_n \equiv \{j : 1 \leq j \leq p_n, \beta_{j,n} \neq 0\} = \{1, \ldots, p_{0n}\}$ be the (population) set of non-zero regression coefficients, where $\beta_j \equiv \beta_{j,n}$ is the jth component of $\boldsymbol{\beta} \equiv \boldsymbol{\beta}_n$. The ALASSO yields an estimator

$$\widehat{I}_n \equiv \{j : 1 \leq j \leq p_n,\ \widehat{\beta}_{j,n} \neq 0\}$$

of I_n. Thus, $\widehat{I}_n$ is the set of variables selected by the ALASSO.

Next, define $\boldsymbol{C}_n = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'$. Define the submatrices of $\boldsymbol{C}_n$ using the partition

$$\boldsymbol{C}_n = \begin{bmatrix} \boldsymbol{C}_{11,n}, \boldsymbol{C}_{12,n} \\ \boldsymbol{C}_{21,n}, \boldsymbol{C}_{11,n} \end{bmatrix}$$

where $\boldsymbol{C}_{11,n}$ is $p_0 \times p_0$. For the development here, we do *not* require that $\boldsymbol{C}_n$ be invertible even when $p \le n$. However, for the rest of the paper, we do suppose that $\boldsymbol{C}_{11,n}$ is invertible. Thus, the rank of X is greater than or equal to p_0, and the non-zero regression coefficients are uniquely determined. For any $p \times 1$ vector $\mathbf{x} \in \mathbb{R}^p$, let $\mathbf{x}^{(1)}$ denote the subvector consisting of its first p_0 components. Thus, $\boldsymbol{\beta}_n = (\boldsymbol{\beta}_n^{(1)'}, \boldsymbol{\beta}_n^{(2)'})'$ where $\boldsymbol{\beta}_n^{(2)'} = \mathbf{0} \in \mathbb{R}^{p-p_0}$. Note that the oracle based OLS concerns the reduced model

$$y_i = \mathbf{x}_i^{(1)'}\boldsymbol{\beta}^{(1)} + \epsilon_i, \quad i = 1, \dots, n. \tag{2.1}$$

Denote the oracle based OLS of $\boldsymbol{\beta}^{(1)}$ by $\widetilde{\boldsymbol{\beta}}_n^{\mathrm{OR}}$. Note that by our assumption on $\boldsymbol{C}_{11,n}$, $\widetilde{\boldsymbol{\beta}}_n^{\mathrm{OR}}$ is well defined.

Next define the ith residual based on the ALASSO as

$$e_i = y_i - \sum_{j=1}^{p} x_{ij}\widehat{\beta}_{n,j}, \quad 1 \le i \le n. \tag{2.2}$$

Let $\widehat{F}_n(x) = n^{-1}\sum_{i=1}^n \mathbb{1}(e_i \le x)$, $x \in \mathbb{R}$ be the ALASSO based residual edf. Also, let $\widetilde{F}_n^{\mathrm{OR}}$ denote the edf of the residuals from the oracle based OLS. Also, let $\bar{\mathbf{x}}_n = n^{-1}\sum_{i=1}^n \mathbf{x}_i$, $\eta_{11,n}$ = the smallest eigenvalue of $\boldsymbol{C}_{11,n}$ and $\kappa_n = \min\{|\beta_{j,n}| : j \in I_n\}$.

2.2. Conditions

We shall make use of the following conditions:

(C.1) There exists $\delta \in (0, 1)$, such that for all $\mathbf{x} \in \mathbb{R}^{p_0}$, $\mathbf{y} \in \mathbb{R}^{p-p_0}$ and $n > \delta^{-1}$,

$$\left(\mathbf{x}'\boldsymbol{C}_{12,n}\mathbf{y}\right)^2 \le \delta^2 \left(\mathbf{x}'\boldsymbol{C}_{11,n}\mathbf{x}\right) \cdot \left(\mathbf{y}'\boldsymbol{C}_{22,n}\mathbf{y}\right).$$

(C.2) $\max\{|\beta_{j,n}| : j \in I_n\} = O(1)$ and $p_0 \log n \ll n\eta_{11,n}^2\kappa_n^2$.

(C.3) (i) $p = O(n^\nu)$ for some integer $\nu \ge 1$.

(ii) $\max\{\mathbf{x}_i^{(1)'}\boldsymbol{C}_{11,n}^{-1}\mathbf{x}_i^{(1)} : 1 \le i \le n\} = O(1)$ and $\bar{\mathbf{x}}_n^{(1)'}\boldsymbol{C}_{11,n}^{-1}\bar{\mathbf{x}}_n^{(1)} \to c_\infty^2$ for some $c_\infty^2 \in [0, \infty)$.

(iii) $\max\left\{n^{-1}\sum_{i=1}^n |x_{i,j}|^{2\nu+2} : 1 \le j \le p\right\} = O(1)$.

(C.4) (i) For some $\gamma \ge 1$,
$\lambda_n p_0 \ll n\eta_{11,n}\kappa_n^{1+\gamma}$ and $\lambda_n^2 p_0 \ll n\eta_{11,n}\kappa_n^{2\gamma}$.

(ii) $p_0^3 \log p_0 = o(n)$.

(C.5) (i) $\boldsymbol{E}(\epsilon_1) = 0$, $\boldsymbol{E}(\epsilon_1^2) = \sigma^2 \in (0, \infty)$ and $\boldsymbol{E}|\epsilon_1|^{2\nu+2} < \infty$, where ν is as in (C.3) above.

(ii) The distribution F is strictly increasing on its support, an interval $[c, d]$ where $-\infty \le c < d \le \infty$. It has a density f with derivative Df such that both f and $|Df|$ are bounded on (c, d) by some constant M_f.

(C.6) (i) There exists a sequence $a_n \to 0+$ such that $n^{1/2}a_n \to \infty$ and

$$\boldsymbol{P}\Big(\max_{1\le j\le p} |\tilde{\beta}_{j,n} - \beta_j| > a_n\Big) = o(1).$$

(ii) $\kappa_n \gg a_n$ and $\lambda_n a_n^{-\gamma} \gg n\sqrt{p_0}\kappa_n$.

We now comment on the conditions. Condition (C.1) requires the multiple correlation between the relevant variables (with $\beta_{j,n} \ne 0$) and the spurious variables ($\beta_{j,n} = 0$) to be strictly less than one, in absolute value. This condition is weaker than assuming orthogonality of the two sets of variables. Variants of this condition have been used in the literature, particularly in the context of the LASSO: see Meinshausen and Yu (2009), Huang et al. (2008a), Chatterjee and Lahiri (2011a), and the references therein. Next consider Condition (C.2). When p is bounded and the regression parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is fixed and $\boldsymbol{C}_n \to \boldsymbol{C}$ (element-wise) for a nonsingular $\boldsymbol{C}$, Condition (C.2) trivially holds with $\eta_{11,n}^{-1} + \kappa_n^{-1} = O(1)$. In the increasing dimensional case, the first part of (C.2) requires the nonzero regression co-efficients to be (uniformly) bounded by some constant, while the second part of (C.2) imposes lower bounds on the smallest eigenvalue of the submatrix $\boldsymbol{C}_{11,n}$ and on the smallest non-zero regression co-efficient. Note that if the smallest eigenvalue of $\boldsymbol{C}_{11,n}$ is bounded away from zero and p_0 is fixed (but p may still diverge to infinity), the smallest nonzero regression parameter is allowed to go to zero at a rate arbitrary close to $n^{-1/2}\sqrt{\log n}$.

Next consider Condition (C.3). Note that we allow the dimension of the regression parameters to grow at an arbitrarily large *polynomial rate* with the sample size. This growth rate is directly related to the moment condition on ϵ_1 in (C.5) and also to the conditions on the (sample) moments of the design vectors. Part (ii) of (C.3) is used for establishing the Functional Central Limit Theorem (FCLT) for the residual empirical process $n^{1/2}(\widehat{F}_n(\cdot) - F(\cdot))$ while part (iii) is needed for certain moderate deviation bounds. Here we point out that the results of the paper can be extended to the case $\log p = o(n)$, provided ϵ_1 has a sub-Gaussian tail and the components of the $\mathbf{x}_i$s are (uniformly) bounded. Condition (C.4) specifies the growth rate of p_0 and also an upper bound on the penalty parameter λ_n that we need here to prove the AUL result. In view of the discussions earlier, the growth rate of p_0 is optimal up to a logarithmic term in general (cf. [CL01]). (C.5) gives the required conditions on the error distribution function. The moment condition in part (i) is somewhat stronger than what is usually used in the context of proving a FCLT for the residual edf based on the OLS, in dimensions p that grow at most as a small fractional power of the sample size. Here we use the stronger condition to prove the AUL and the functional oracle property of the ALASSO method under sparsity, allowing p to grow at an arbitrary polynomial rate in n, provided (C.3)(iii) and (C.5)(i) hold. The second part of (C.5) is a standard condition that has been used by several authors in the past; See Koul (2002), [CL01] and the references therein.

Finally, consider Condition (C.6). Part (i) of (C.6) requires the initial estimator $\widetilde{\boldsymbol{\beta}}_n$ to be consistent for the underlying regression parameter $\boldsymbol{\beta}$ at a suitable rate. Part (ii) of Condition (C.6) specifies a lower bound on the penalty parameter λ_n that, in particular, depends on the tuning parameter γ of the ALASSO criterion and the convergence rate a_n of the initial estimator. It also specifies conditions on the smallest nonzero regression coefficient with respect to the convergence rate a_n of the initial estimator to ensure detectability. Note that the initial estimator need not be root-n consistent for the Functional Oracle property to hold. For $p > n$, LASSO and related bridge estimators have been suggested as potential choices of $\widetilde{\boldsymbol{\beta}}_n$ in the literature. Indeed, results of Zhao and Yu (2006) and Zhang and Huang (2008) can be used to verify (C.6)(i) for LASSO under additional regularity conditions on the design vectors and the error variables. When $p \le n$ and we use the OLS as the initial estimator, we can set $a_n = n^{-1/2}\sqrt{\log n}$, as shown by the following result:

Proposition 2.1. *Suppose that $p \le n$, Condition (C.5)(i) holds with $\nu = 1$, $\boldsymbol{C}_n$ is nonsingular, and $\widetilde{\boldsymbol{\beta}}_n$ is the OLS of $\boldsymbol{\beta}$. Also, let $\max\left\{n^{-1}\sum_{i=1}^n |\tilde{x}_{i,j}|^4 : 1 \le j \le p\right\} = O(1)$ where $\tilde{x}_{i,j}$ is the jth element of $(\mathbf{x}_i'\boldsymbol{C}_n^{-1})$. Then, (C.6)(i) holds with $a_n = K\sqrt{n^{-1}\log n}$.*

3. Main results

3.1. Asymptotic uniform linearity in high dimensions

For $t \in [0, 1]$, define

$$\widehat{Z}_n(t) = \frac{1}{\sqrt{n}}\sum[\mathbb{1}(F(e_i) \le t) - t]$$

$$Z_n(t) = \frac{1}{\sqrt{n}}\sum[\mathbb{1}(F(\epsilon_i) \le t) - t]$$

where recall that $\mathbb{1}(\cdot)$ denotes the indicator function. Set $J(t) = f(F^{-1}(t))$. The first result of this section proves the AUL property of the ALASSO based residual edf for $p = O(n^{\nu})$ as $n \to \infty$, for any integer $\nu \geq 1$.

Theorem 1. *Under Conditions* (C.1)–(C.6),

$$\sup\left\{\left|\widehat{Z}_n(t) - \left[Z_n(t) - \bar{\boldsymbol{x}}_n'\{\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\}J(t)\right]\right| : t \in [0,1]\right\} = o_P(1).$$

Theorem 1 shows that the AUL property holds for the edf of the residuals based on the ALASSO fit even for $p \gg n$, provided the regularity conditions (C.1)–(C.6) hold and the regression model has enough sparsity. In comparison, results of Portnoy (1986) imply that the AUL property for edf based on the residuals from a standard method of estimation (such as M-estimation) of the full vector $\boldsymbol{\beta}$ may fail whenever $p^2/n \to \infty$. As a result, the penalization step of ALASSO plays a crucial role in extending the AUL property to very high dimensional regression problems under sparsity.

3.2. Functional oracle property of the ALASSO

The AUL property has a number of important applications in the context of statistical inference in high dimensional regression (cf. Koul, 2002). Among others, it allows one to establish the asymptotic distribution of the residual edf which, in turn, can be used to carry out goodness of fit tests on F and set confidence bands for F. Note that under Conditions (C.1)–(C.6), Theorem 1 of [CL01] and standard weak convergence theory (cf. van der Vaart and Wellner, 1996) can be used to conclude that

$$\mathcal{Z}_n^{\mathrm{OR}} \equiv n^{1/2}\left(\widehat{F}_n^{\mathrm{OR}} - F\right) \to^w \mathcal{Z}_\infty \quad \text{on } L^\infty([-\infty,\infty]) \tag{3.1}$$

where $\widehat{F}_n^{\mathrm{OR}}(x)$ is the edf of the residuals based on OLS under the oracle and $\mathcal{Z}_\infty$ is a zero mean Gaussian process with covariance function

$$\begin{aligned}\tau(x,y) &= \mathrm{Cov}(\mathcal{Z}_\infty(x), \mathcal{Z}_\infty(y)), \quad x, y \in [-\infty., \infty] \\ &= F(x \wedge y) - F(x)F(y) + c_\infty^2\sigma^2 f(x)f(y) \\ &\quad + \left[\boldsymbol{E}\epsilon_1\mathbb{1}(\epsilon_1 \leq x)f(y) + \boldsymbol{E}\epsilon_1\mathbb{1}(\epsilon_1 \leq y)f(x)\right]c_\infty^2. \end{aligned} \tag{3.2}$$

Here and in the following, $L^\infty([-\infty,\infty])$ denotes the space of all bounded functions on $[-\infty,\infty]$, equipped with the sup-norm and $\to^w$ denotes weak convergence on $L^\infty([-\infty,\infty])$. Define the ALASSO based residual empirical process (cf. (2.2))

$$\mathcal{Z}_n(x) = \sqrt{n}\left[\widehat{F}_n(x) - F(x)\right], \quad x \in [-\infty,\infty].$$

Also, recall that $\widehat{I}_n$ is the set of covariates selected by the ALASSO. With this, we have the following result:

Theorem 2 (*Functional Oracle Property of ALASSO*). *Under Conditions* (C.1)–(C.6), *as* $n \to \infty$,

$$\boldsymbol{P}(\widehat{I}_n = I_n) \to 1 \quad \text{and} \quad \mathcal{Z}_n \to^w \mathcal{Z}_\infty \quad \text{on } L^\infty([-\infty,\infty]),$$

where $\mathcal{Z}_\infty(\cdot)$ *is the Gaussian Process in* (3.1) *with covariance function* (3.2).

Thus, Theorem 2 shows that under conditions (C.1)–(C.6), in addition to selecting the correct set of variables with high probability, the ALASSO based residual empirical process converges to the same limit as the OLS based residual empirical process under the oracle. This shows that the penalization step in ALASSO is highly effective in identifying the true set of relevant variables and as a result, it is capable of capturing the asymptotic distribution of the residual empirical process under the oracle, even when the latter is a complicated probability law on a function space. By an application of the continuous mapping theorem and the functional Delta method, this in turn implies that the ALASSO method can capture the oracle limit of a very large class of functionals of the residual edf, solely under (additional) continuity and Hadamard differentiability conditions on the functional. Thus, compared to the existing results that are limited to considering finitely many linear combinations of the regression parameter vector $\boldsymbol{\beta}$, Theorem 2 extends the scope of ALASSO based inference to a wide class of functionals of F.

Theorem 2 immediately allows us to construct large sample confidence bands for the error d.f. F. To that end, let q_α denote the α quantile of the random variable $\|\mathcal{Z}_\infty\|_\infty = \sup\{|\mathcal{Z}_\infty(x)| : x \in \mathbb{R}\}$, $\alpha \in (0,1)$. (Note that $\mathcal{Z}_\infty(-\infty) = \mathcal{Z}_\infty(\infty) = 0$ with probability one as $\tau(x,x) = 0$ for $x \in \{\pm\infty\}$ and hence, $\|\mathcal{Z}_\infty\|_\infty$ is also the supremum of $|\mathcal{Z}_\infty(x)|$ over all $x \in [-\infty,\infty]$.) Define $J_\alpha = \{G : \|G - \widehat{F}_n\|_\infty \leq q_\alpha\}$. Then we have the following result.

Corollary 1. *Suppose that Conditions* (C.1)–(C.6) *hold. Then, for any* $\alpha \in (0,1)$,

$$\boldsymbol{P}\left(F \in J_\alpha\right) \to \alpha \quad \text{as } n \to \infty.$$

Thus, Corollary 1 shows that the ALASSO based residual edf can be used to construct an asymptotically valid confidence band for F at any given confidence level $\alpha \in (0,1)$ in high dimensional regression model (1.1). Here again, the naive approach of using the OLS based residual edf fails to produce an asymptotically valid confidence band for F when $p^3/n \to \infty$. Indeed, the simulation results in the next section show that the confidence bands based on the OLS based residual edf have very poor coverage probabilities even in large samples.

3.3. Prediction intervals

Next, we consider the problem of constructing large sample prediction intervals for an unobserved value $y_{\mathbf{x}}$ of the response variable Y at a given value $\mathbf{x} \in \mathbb{R}^p$ of the covariates. Let $\widehat{y}_{\mathbf{x}} = \mathbf{x}'\widehat{\boldsymbol{\beta}}_n$ denote the predicted value of Y at $\mathbf{x}$ based on the ALASSO estimator $\widehat{\boldsymbol{\beta}}_n$. Let $\alpha \in (0,1)$. Define a level α PI for $y_{\mathbf{x}}$ as

$$J_\alpha^0 = \left[\widehat{y}_{\mathbf{x}} + \widehat{q}_{[1-\alpha]/2}, \widehat{y}_{\mathbf{x}} + \widehat{q}_{(1+\alpha)/2}\right],$$

where $\widehat{q}_t = \widehat{F}_n^{-1}(t)$ is the t-quantile of the ALASSO based residual edf $\widehat{F}_n$, $t \in (0,1)$. The following result shows that the resulting PI attains the nominal confidence level α asymptotically.

Theorem 3. *Suppose that Conditions* (C.1)–(C.6) *hold and that* $\mathbf{x}^{(1)\prime}C_{11,n}^{-1}\mathbf{x}^{(1)} = o(n)$, *where* $\mathbf{x}^{(1)}$ *is the vector of the first* p_0*-components of* $\mathbf{x}$. *Then, for any* $\alpha \in (0,1)$,

$$\boldsymbol{P}\left(y_{\mathbf{x}} \in J_\alpha^0 \middle| \epsilon_1, \ldots, \epsilon_n\right) = \alpha[1 + o_P(1)].$$

Note that for the validity of Theorem 3, in addition to (C.1)–(C.6), we require conditions on the first p_0 components of the new covariate vector $\mathbf{x}$ at which we want to predict the response variable $y_{\mathbf{x}}$. Note that for the (asymptotic) validity of the PI constructed from the ALASSO-based residual edf, no condition on the last $p - p_0$ components (corresponding to the zero-components of $\boldsymbol{\beta}$) has been imposed. This seems to be a reasonable thing to ask for, since under the sparsity assumption on the model (1.1), the value of $y_{\mathbf{x}}$ does not depend on the last $p - p_0$ components of $\mathbf{x}$, and Theorem 3 delivers a valid large sample PI for $y_{\mathbf{x}}$ for any choices of the last $p - p_0$ components of $\mathbf{x}$.

4. Simulation results

To illustrate finite sample properties of the proposed methodology based on the ALASSO residual empirical process, we carried out a detailed simulation study. We considered the following five combinations of (n, p), setting $p_0 = 10$ in all cases. Also, *the non-zero components of* $\boldsymbol{\beta}$ *are kept fixed in all these five cases.*

(a) $(n = 60,\ p = 20)$, with

$$\boldsymbol{\beta} = (4, -1.5, -8, 0.9, -3, 0.75, 2, -0.5, 10, -6, 0, \ldots, 0)'.$$

(b) $(n = 60,\ p = 50)$.
(c) $(n = 60,\ p = 100)$.
(d) $(n = 200,\ p = 50)$.
(e) $(n = 200,\ p = 500)$.

The ith row of the design matrix $\mathbf{x}_i' = \left(x_{i,1}, \ldots, x_{i,p_0}\right)$ is selected independently from a $N_p(\mathbf{0}, \Sigma)$ population. Once the design matrix is generated, it is kept fixed throughout the simulation. We considered two different choices for the covariance matrix Σ:

(S1) A block diagonal structure: $\Sigma = \mathrm{diag}\left(\Sigma_{1,1}, \mathbf{I}_{p-p_0}\right)$, where $\Sigma_{1,1} = ((\sigma_{i,j}))$ is a $p_0 \times p_0$ matrix with (i, j)th element $\sigma_{i,j} = (0.7)^{|i-j|}$, $1 \le i, j \le p_0$.
(S2) An equi-correlation structure: $\Sigma = (1-0.7) \cdot \mathbf{I}_p + (0.7) \cdot \mathbf{1}\mathbf{1}'$, where $\mathbf{1}$ is a p dimensional of vector of 1's.

It was found that the choice of the covariance matrix Σ, was not influencing the results significantly. Hence we only report the results for the (S1) case.

The error distribution considered was $F = N(0, 1)$. The penalty parameter λ_n was selected using two different approaches, namely, *Theoretical rates* and *Cross Validation*, as described below:

(Th) *Theoretical rates:* here λ_n is selected in such a manner such that it satisfies the theoretical rate conditions given in (C.4) and (C.6). We have used $\lambda_n = cn^{1/4}$, with $c \in \{0.25, 1, 2, 5\}$. If $p > n$, we have used the LASSO as the initial estimator, which requires the choice of a tuning parameter $\lambda_{0,n}$. In our simulations we have used $\lambda_{0,n} = 0.5 \cdot n^{1/2}$.
(CV) *Cross Validation:* here λ_n was selected using 5-fold cross validation. If the LASSO was used as an initial estimator, the corresponding tuning parameter was also selected by 5-fold CV.

Construction of the confidence bands J_α (cf. Corollary 1) requires the knowledge of the quantiles q_α.

(BB) *Brownian Bridge:* Note that in general, the quantiles q_α's are *not* distribution free. However, in the important special case where design vectors x_i's are generated by realizations of zero mean iid random vectors and the smallest eigenvalue of $\boldsymbol{C}_{11,n}$ are bounded away from zero (which happens to be the case in the simulation set up), the constant $c_\infty^2 = 0$. In this case, the quantiles q_αs are given by the quantiles of the supremum of the Brownian Bridge on [0, 1], which does not depend on any population parameters and hence, can be used for any F.

As mentioned above, the ALASSO requires an initial estimator and we have used the following:

(AL) *OLS:* In case of $p < n$: the initial estimator is the OLS and the second stage estimator is the ALASSO.
(AL.L) *LASSO:* In case of both $p < n$ and $p > n$: the initial estimator is the LASSO and the second stage estimator is the ALASSO.

One of the referees has suggested that our methods be compared with post-model selection based estimators, by applying the usual OLS on the model selected by ALASSO. Another suggestion was to compare the performance of our procedures by benchmarking against the Oracle based estimator (by applying the OLS on the known set of relevant covariates). The following list describes these methods:

Table 1
Empirical coverage probabilities for the error d.f. F based on the confidence band J_α with **(S1)** as the covariance matrix and $\alpha = 0.9$. The error df is $F = N(0, 1)$.

(n, p)	Choice of λ_n		Estimation Methods (q_α estimated using BB)				
			AL	AL.L	PM.AL	PM.AL.L	Oracle[b]
(60, 20)	CV		0.81	0.806	0.792	0.806	0.884
	Th[a]	0.25	0.794	0.826			
		1.00	0.898	0.914			
		5.00	0.826	0.742			
(60, 50)	CV		0.516	0.606	0.402	0.464	0.926
	Th[a]	0.25	0.230	0.564			
		1.00	0.916	0.948			
		5.00	0.890	0.888			
(60, 100)	CV			0.202		0.112	0.928
	Th[a]	0.25		0.004			
		1.00		0.818			
		5.00		0.896			
(200, 50)	CV		0.888	0.892	0.878	0.860	0.932
	Th[a]	0.25	0.872	0.854			
		1.00	0.938	0.934			
		5.00	0.924	0.912			
(200, 500)	CV			0.038		0.002	0.912
	Th[a]	0.25		0.000			
		1.00		0.864			
		2.00		0.936			

[a] Using $\lambda_n = c \cdot n^{1/4}$, with $c \in \{1/4, 1, 5\}$ or $\{1/4, 1, 2\}$.
[b] The Oracle based method does not depend on the choice of λ_n.

(PM.AL) For the $p < n$ case: we applied OLS on the model selected by the ALASSO. The tuning parameter was selected by 5-fold CV and q_α were computed using the Brownian Bridge result.
(PM.AL.L) For both $p < n$ and $p > n$ case: we applied OLS on the model selected by the ALASSO (with the LASSO as the initial estimator). The tuning parameter and the quantiles were selected in the same manner as in (PM.AL) above.
(OR) The oracle based estimator, where the true set of non-zero coefficients (*viz.* $\{1, \ldots, p_0\}$) of $\boldsymbol{\beta}$ is known and OLS is applied on these first p_0 components. Here, the quantiles q_α are based on (BB).

The itemization shown above provides the symbols which we will use to describe our results. We do not study all possible combinations as it would be too time consuming. Instead, we focus on selected cases to study the performance of each procedure. In case any value is missing in Tables 1 and 2, then the corresponding cases were not studied.

Table 1 shows the finite sample coverage probabilities of the confidence bands J_α, for the underlying error cdf $F = N(0, 1)$, given by Corollary 1, at nominal level $\alpha = 0.9$ for all five different combinations of (n, p). In all cases, the theoretical choice of λ_n provides good coverage with both ALASSO based and the ALASSO + LASSO based methods. Further, their performance changes with the change of the tuning parameter, sometimes with the appropriate amount of penalization the coverage probabilities improve drastically. The situation is not the same in case of CV based choice of tuning parameter, particularly in the high-dimensional cases. In the low-dimensional cases, the CV based method performs nicely with slightly poor performance in case of $(n = 60, p = 50)$. Similar behavior was found using $(n = 200, p = 150)$, which we have not reported. In all cases, the oracle based method performs very nicely, as expected. The Post-Model selection based methods provide comparable performance in the low-dimensional cases, but their performance deteriorates in case where $p > n$ or $p \sim n$.

Table 2
Prediction interval accuracy based on J_α^0 using one test sample, with $\alpha = 0.9$, for two different error df's.

(n, p)	$F = N(0, 1)$		$F = \chi_1^2 - 1$	
	AL	AL.L	AL	AL.L
(60, 20)	0.778	0.782	0.808	0.814
(60, 50)	0.590	0.610	0.620	0.700
(60, 100)		0.284		0.332
(200, 50)	0.848	0.848	0.814	0.814
(200, 500)		0.504		0.412

Finally, for the ALASSO based prediction, the PI's were computed by using the CV-based choice of the tuning parameter. Overall, the results in Table 2 indicate that the prediction interval procedure performs nicely for low-dimensional p and has worse performance in the high-dimensional cases. To consider the effect of an asymmetric error distribution, we also included $F = \chi_1^2 - 1$ in the study of ALASSO based PIs. In this situation, the performance for $\chi_1^2 - 1$ errors is comparable to the $N(0, 1)$ errors.

5. Conclusion

In this paper, we investigated the properties of the ALASSO method for inference on the error distribution in a high dimensional regression model when the dimension p of the regression model may grow at an arbitrary polynomial rate with the sample size n. The main results of the paper show that under appropriate regularity conditions, the ALASSO method enjoys a Functional Oracle property. Specifically, the ALASSO can identify the true set of nonzero regression coefficients with probability tending to one and at the same time, the centered and scaled residual empirical process based on the ALASSO converges weakly to the *same Gaussian process* as the centered and scaled residual empirical process based on the OLS under the Oracle that specifies *a priori* the nonzero coefficients in the regression model. It is important to note that the Oracle reduces the dimension of the regression model to p_0 which grows only as a fractional power of the sample size. In comparison, the full model may have a dimension p as large as any given polynomial power of the sample size. Thus, the ALASSO can effectively deal with the complexities of a very high dimensional model guaranteeing a level of precision that is comparable to the OLS method in a much lower dimensional model.

A key requirement for the use of the ALASSO method is the choice of a suitable initial estimator. For $p \le n$, Proposition 2.1 shows that the OLS method provides an attractive choice under fairly general regularity conditions. However, for $p > n$, the OLS is no longer unique and one must employ alternative estimators that are well defined in such high dimensional cases. Here, we may choose the initial estimator as the LASSO estimator or the Ridge estimator of the full p-dimensional parameter $\boldsymbol{\beta}$. The simulation results show that the coverage probabilities of the confidence bands for the error distribution based on the ALASSO method with these choices of the initial estimator are very good even when the sample size is not too large, i.e., in the case, $n = 60$ and $p = 50, 100$. In comparison, the OLS based confidence bands, where applicable, perform rather poorly. Thus, the ALASSO method has a definite advantage over traditional methods that target estimating the full regression parameter and fail to exploit model sparsity. It is worth pointing out that the growth rate of p can be generalized to

$$\log p = o(n) \quad \text{as } n \to \infty,$$

provided some stronger conditions on the design vectors and error distribution (such as $E\exp(t\epsilon_1) < \infty$ for all t in an open interval around the origin) hold. However, we preferred to work with a weaker moment condition that is applicable to a larger class of error distributions. Specifically, our results allow for (a potentially arbitrarily large) polynomial growth rate that is directly related to the tails of the error distribution. Further, we expect that under suitable regularity conditions, other popular penalized regression methods with the Oracle property, such as the SCAD and the MCP, can also be shown to have the Functional Oracle property.

6. Proofs

Recall that $\mathbf{x}^{(1)}$ denotes the first p_0 components of a vector $\mathbf{x} \in \mathbb{R}^p$. Also, recall that the rescaled ALASSO based residual empirical process and the unobservable empirical process of the error variables $\epsilon_1, \ldots, \epsilon_n$ are respectively given by

$$\widehat{Z}_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big[\mathbb{1}\Big(F(e_i) \le t \Big) - t \Big], \quad t \in [0, 1] \quad \text{and}$$

$$Z_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big[\mathbb{1}\Big(F(\epsilon_i) \le t \Big) - t \Big] \quad t \in [0, 1].$$

Let $\Gamma_n(t, \mathbf{u}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big[\mathbb{1}\Big(F(\epsilon_i - n^{-1/2}\mathbf{x}_i'\mathbf{u}) \le t \Big) - t \Big]$, $t \in [0, 1]$ and $\mathbf{u} \in \mathbb{R}^p$. Then,

$$\widehat{Z}_n(t) = \Gamma_n\Big(t, \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \Big), \quad t \in [0, 1].$$

Next define the random vector $\boldsymbol{W}_n = n^{-1/2} \sum_{i=1}^n \mathbf{x}_i'\epsilon_i$ and write $\boldsymbol{W}_n' = \Big(\boldsymbol{W}_n^{(1)'}, \boldsymbol{W}_n^{(2)'} \Big)$. Let

$$V_n(\mathbf{u}) = \mathbf{u}'\boldsymbol{C}_n\mathbf{u} - 2\mathbf{u}'\boldsymbol{W}_n + \lambda_n \sum_{j=1}^p |\tilde{\beta}_{j,n}|^{-\gamma} \left(\left| \beta_{j,n} + \frac{u_j}{\sqrt{n}} \right| - |\beta_{j,n}| \right), \quad \mathbf{u} \in \mathbb{R}^p.$$

Note that

$$\widehat{\boldsymbol{\beta}}_n = \underset{\mathbf{u} \in \mathbb{R}^p}{\operatorname{argmin}}\; V_n(\mathbf{u}). \tag{6.1}$$

Let $K, K(\cdot)$ denote generic constants with values in $(0, \infty)$ that do not depend on n, but may depend on the arguments, if any. Also, unless otherwise specified, limits in order symbols are taken by letting $n \to \infty$.

Proof of Proposition 2.1. Note that under the conditions of the proposition, $p \le n$ and $\boldsymbol{C}_n$ is nonsingular. Hence, the OLS estimator of $\boldsymbol{\beta}$ is unique and is given by $\widetilde{\boldsymbol{\beta}}_n = n^{-1}\boldsymbol{C}_n^{-1} \sum_{i=1}^n \mathbf{x}_i y_i$. Hence, it is easy to check that for any $1 \le j \le p$,

$$\beta_{n,j} - \beta_j = n^{-1} \sum_{i=1}^n \tilde{x}_{i,j}\epsilon_i.$$

Hence, using the Fuk–Nagaev inequality (cf. Section 4 of Fuk and Nagaev, 1971), we have

$$\begin{aligned}
&\boldsymbol{P}\Big(\max\{|\beta_{n,j} - \beta_j| : j = 1, \ldots, p\} > Kn^{-1/2}\sqrt{\log n} \Big) \\
&\le \boldsymbol{P}\left(\max\left\{ \left| n^{-1/2} \sum_{i=1}^n \tilde{x}_{i,j}\epsilon_i \right| : 1 \le j \le p \right\} > K\sqrt{\log n} \right) \\
&\le \sum_{j=1}^p \left[Kn^{-2}(\log n)^{-2} \sum_{i=1}^n |\tilde{x}_{i,j}|^4 \boldsymbol{E}|\epsilon_i|^4 + \exp\left(-K \frac{n\log n}{\Big[\sum_{i=1}^n \tilde{x}_{i,j}^2\sigma^2\Big]} \right) \right] \\
&= O\Big(pn^{-1}(\log n)^{-2} \Big) = o(1).
\end{aligned} \tag{6.2}$$

Hence, Condition (C.6)(i) holds.

Proof of Theorem 1. Define the sets

$$A_{1,n} = \left\{\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\|_\infty \le a_n\right\}$$

$$A_{2,n} = \left\{\|\boldsymbol{W}_n^{(1)}\| \le K\sqrt{p_0 \log n}\right\}$$

$$A_{3,n} = \left\{\|\boldsymbol{W}_n^{(2)}\|_\infty \le K\sqrt{\log n}\right\}.$$

Note that by Condition (C.6), $\boldsymbol{P}(A_{1n}^c) = o(1)$. We now show that under Conditions (C.3)(ii) and (C.5)(i), $\boldsymbol{P}(A_{k,n}^c) = o(1)$ for $k = 2, 3$. To that end, consider $A_{2,n}$. As in the proof of Proposition, using the Fuk–Nagaev inequality (cf. Section 4 of Fuk and Nagaev, 1971), with $s = 2\nu + 2$, we have

$$\begin{aligned}\boldsymbol{P}(A_{2,n}^c) &\le \boldsymbol{P}\left(\max\left\{\left|n^{-1/2}\sum_{i=1}^n x_{i,j}\epsilon_i\right| : 1 \le j \le p\right\} > K\sqrt{\log n}\right)\\ &\le \sum_{j=1}^p\Bigg[Kn^{-(s-2)/2}(\log n)^{-s/2}\sum_{i=1}^n |x_{i,j}|^s\boldsymbol{E}|\epsilon_i|^s\\ &\quad + \exp\Bigg(-K\frac{n\log n}{[\sum_{i=1}^n x_{i,j}^2\sigma^2]}\Bigg)\Bigg]\\ &= O\Big(pn^{-(s-2)/2}(\log n)^{-s/2}\Big) = o(1). \end{aligned} \tag{6.3}$$

By similar arguments, $\boldsymbol{P}(A_{3,n}^c) = o(1)$. Hence, writing $A_n = A_{1,n} \cap A_{2,n} \cap A_{3,n}$, it follows that $P(A_n) \to 1$.

Note that by condition (C.1) on the set A_n, for n large,

$$\begin{aligned} V_n(\mathbf{u}) &\ge (1-\delta)\left[\mathbf{u}^{(1)'}\boldsymbol{C}_{11,n}\mathbf{u}^{(1)} + \mathbf{u}^{(2)'}\boldsymbol{C}_{22,n}\mathbf{u}^{(2)}\right]\\ &\quad - 2\mathbf{u}'\boldsymbol{W}_n + \frac{\lambda_n}{\sqrt{n}}\sum_{j=1}^p\frac{|u_j|}{|\tilde\beta_{j,n}|^\gamma} - 2\lambda_n\sum_{j=1}^{p_0}\frac{|\beta_{j,n}|}{|\tilde\beta_{j,n}|^\gamma}\\ &\ge (1-\delta)\|\mathbf{u}^{(1)}\|^2\eta_{11,n} - 2\|\mathbf{u}^{(1)}\|\|\boldsymbol{W}_n^{(1)}\| - 2\lambda_n\sum_{j=1}^{p_0}\frac{|\beta_{j,n}|}{|\tilde\beta_{j,n}|^\gamma}\\ &\quad + \sum_{j=p_0+1}^p |u_j|\left(\frac{\lambda_n}{\sqrt{n}}\cdot\frac{1}{|\tilde\beta_{j,n}|^\gamma} - 2|W_{j,n}|\right)\\ &\ge \|\mathbf{u}^{(1)}\|\left\{(1-\delta)\|\mathbf{u}^{(1)}\|\eta_{11,n} - K\sqrt{p_0\log n}\right\}\\ &\quad - 2\lambda_n\sum_{j=1}^{p_0}|\beta_{j,n}|^{1-\gamma}\Big(1 + K(\gamma)a_n\Big)\\ &\quad + K(\gamma)\cdot\sum_{j=p_0+1}^p |u_j|\left[\lambda_n\cdot n^{-1/2}a_n{}^{-\gamma} - K(\log n)^{1/2}\right]. \end{aligned} \tag{6.4}$$

Next note that for any $a > 0, c > 0$, the quadratic form $Q(x) = ax^2 - bx - c$ is positive whenever $x > (2a)^{-1}[b + \sqrt{b^2 + 4ac}]$. Hence, it follows from (6.4) and Conditions (C.2) and (C.4) that

$$V_n(\mathbf{u}) > 0 \quad \text{for all } \mathbf{u} \in B_n \equiv \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}^{(1)}\| > M_n\} \tag{6.5}$$

where

$$\begin{aligned} M_n &\equiv K(\delta,\gamma)\max\Bigg\{\eta_{11,n}^{-1}(p_0\log n)^{1/2}\\ &\quad + \Big(\eta_{11,n}^{-1}\lambda_n\sum_{j=1}^{p_0}|\beta_{j,n}|^{1-\gamma}\Big)^{1/2}\Bigg\}. \end{aligned}$$

Next, note that for any $\mathbf{u} = (\mathbf{u}^{(1)}, \mathbf{u}^{(2)})$ and $p_0 + 1 \le j \le p$, the absolute value of the jth component of $\mathbf{u}^{(1)'}\boldsymbol{C}_{12,n}$ is bounded above by

$$\begin{aligned}\sum_{l=1}^{p_0}|u_l c_{l,j,n}| &\le \left(\sum_{l=1}^{p_0}u_l^2\right)^{1/2}\left(\sum_{l=1}^{p_0}c_{l,j,n}^2\right)^{1/2}\\ &\le \|\mathbf{u}^{(1)}\|\left[\mathrm{tr}(\boldsymbol{C}_{11,n})\right]^{1/2}c_{j,j,n}^{1/2}.\end{aligned}$$

Since by (C.2), $\max\{c_{j,j,n} : 1 \le j \le p\} = O(1)$, by (C.4) it follows that on the set A_n and for $n \ge n_0$ (for some $n_0 \ge 1$), uniformly over $\{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}^{(1)}\| \le M_n,\ \mathbf{u}^{(2)} \ne \mathbf{0}\}$,

$$\begin{aligned} &V_n(\mathbf{u}) - V_n\left(\mathbf{u}^{(1)}, \mathbf{0}\right)\\ &= \mathbf{u}^{(1)'}\boldsymbol{C}_{12,n}\mathbf{u}^{(2)} + \mathbf{u}^{(2)'}\boldsymbol{C}_{22,n}\mathbf{u}^{(2)}\\ &\quad - 2\mathbf{u}^{(2)'}\boldsymbol{W}_n^{(2)} + \frac{\lambda_n}{\sqrt{n}}\sum_{j=p_0+1}^p\frac{|u_j|}{|\tilde\beta_{j,n}|^\gamma}\\ &\ge \sum_{j=p_0+1}^p |u_j|\left[\frac{\lambda_n}{\sqrt{n}|\tilde\beta_{j,n}|^\gamma} - 2|W_{j,n}| - \left|\left(\mathbf{u}^{(1)'}\boldsymbol{C}_{12,n}\right)_j\right|\right]\\ &\ge K(\gamma)\sum_{j=p_0+1}^p |u_j|\left[\lambda_n n^{-1/2}a_n^{-\gamma} - (\log n)^{1/2} - M_n p_0^{1/2}\right]\\ &> 0. \end{aligned} \tag{6.6}$$

Since $V_n(\mathbf{0}) = 0$, from (6.5) and (6.6), it follows that the minimum of $V_n(\mathbf{u})$ cannot be attained at any $\mathbf{u}$ satisfying either '$\|\mathbf{u}^{(1)}\| > M_n$' or '$\|\mathbf{u}^{(1)}\| \le M_n, \mathbf{u}^{(2)} \ne \mathbf{0}$'. Thus,

$$\operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^p} V_n(\mathbf{u}) \in B_n \equiv \left\{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}^{(1)}\| \le M_n,\ \mathbf{u}^{(2)} = \mathbf{0}\right\},$$

whenever (6.5) and (6.6) hold. Hence, on the set A_n and for $n \ge n_0$,

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right) &\equiv \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^p} V_n(\mathbf{u}) = \operatorname*{argmin}_{\|\mathbf{u}^{(1)}\|\le M_n} V_n\left(\mathbf{u}^{(1)}, \mathbf{0}\right)\\ &= \left(\mathbf{U}_n^{(1)'}, \mathbf{0}'\right)', \end{aligned} \tag{6.7}$$

where $\mathbf{U}_n^{(1)} = \boldsymbol{C}_{11,n}^{-1}\left(\boldsymbol{W}_n^{(1)} - n^{-1/2}\lambda_n\tilde{\mathbf{s}}_n^{(1)}/2\right)$ and $\tilde{\mathbf{s}}_n^{(1)} = (\tilde s_{1,n}, \ldots, \tilde s_{p_0,n})'$ with $\tilde s_{j,n} \equiv \mathrm{sgn}(\beta_{j,n})|\tilde\beta_{j,n}|^{-\gamma}$, $1 \le j \le p_0$. Also, let $\mathbf{U}_n = \left(\mathbf{U}_n^{(1)'}, \mathbf{0}'\right)' \in \mathbb{R}^p$. Hence, it follows that for $n \ge n_0$ and on A_n,

$$\widehat{Z}_n(t) = \Gamma_n\Big(t, \mathbf{U}_n\Big), \quad t \in [0, 1].$$

Next we obtain a stochastic expansion for $\Gamma_n(t, \mathbf{U}_n)$. Note that $\Gamma_n(t, \mathbf{U}_n) = n^{-1/2}\sum_{i=1}^n\Big[\mathbb{1}\Big(F(\epsilon_i - n^{-1/2}\mathbf{x}_i'\mathbf{U}) \le t\Big) - t\Big] = n^{-1/2}\sum_{i=1}^n\Big[\mathbb{1}\Big(F(\epsilon_i - n^{-1/2}\mathbf{x}_i^{(1)\prime}\mathbf{U}^{(1)}) \le t\Big) - t\Big]$. Let $\tilde{\mathbf{b}}_n = -n^{-1/2}\lambda_n\cdot\boldsymbol{C}_{11,n}^{-1}\tilde{\mathbf{s}}_n^{(1)}/2$. Using Conditions (C.2)–(C.4), it can be shown that

$$\begin{aligned} &n^{-1/2}\max\left\{\left|\tilde{\mathbf{b}}_n'\mathbf{x}_i^{(1)}\right| : 1 \le i \le n\right\}\\ &\le \frac{\lambda_n}{2n}\max\left\{\|\boldsymbol{C}_{11,n}^{-1/2}\mathbf{x}_i^{(1)}\|\right\}\cdot\|\boldsymbol{C}_{11,n}^{-1/2}\|\cdot\|\tilde{\mathbf{s}}_n^{(1)}\|\\ &= O_p\Big(\frac{\lambda_n\eta_{11,n}^{-1/2}\kappa_n^{-\gamma}\sqrt{p_0}}{n}\Big)\\ &= o(1) \end{aligned} \tag{6.8}$$

since, by (6.3), $\|\tilde{\mathbf{s}}_n^{(1)}\| \le K\sqrt{p_0}\kappa_n^{-\gamma}$ on A_n. Also, by similar arguments, $\|\bar{\mathbf{x}}_n^{(1)'}\tilde{\mathbf{b}}_n\| = O_p(1)$. This, together with a direct variance calculation and Condition (C.3), shows that

$$|\bar{\mathbf{x}}_n^{(1)'}\mathbf{U}_n^{(1)}| = O_p(1). \tag{6.9}$$

Further, by Condition (C.3),

$$n^{-1}\boldsymbol{E}\left[\mathbf{U}_n^{(1)'}[X^{(1)\prime}X^{(1)}]\mathbf{U}_n^{(1)}\right] = O(p_0), \tag{6.10}$$

where $X^{(1)}$ denotes the $n \times p_0$ submatrix of X consisting of its first p_0 columns, i.e., the design matrix corresponding to the Oracle that sets $\beta_{p_0+1} = \cdots = \beta_p = 0$.

By (6.9) and (6.10), Conditions ($\boldsymbol{E}$.1) and ($\boldsymbol{E}$.2) of [CL01] hold (with $d_n = p_0$). Now the result follows from Theorem 1 of [CL01] by using Conditions (C.2)–(C.6).

Proof of Theorem 2. Note that under (C.1)–(C.6), by (6.7) in the proof of Theorem 1, there exists an $n_0 \geq 1$ such that for all $n \geq n_0$, $A_n \subset \{\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^{(2)} = \mathbf{0}\}$. Hence, for $n \geq n_0$,

$$\boldsymbol{P}\left(\widehat{I}_n = I_n\right) = \boldsymbol{P}\left(\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^{(2)} = \mathbf{0}\right) \tag{6.11}$$

$$\geq \boldsymbol{P}(A_n) \to 1 \quad \text{as } n \to \infty, \tag{6.12}$$

proving the first part of the theorem, i.e., proving the consistency of variable selection under (C.1)–(C.6).

Next consider the second part, i.e.,

$$\mathcal{Z}_n \to^w \mathcal{Z}_\infty \quad \text{on } L^\infty([-\infty, \infty]).$$

Note that by Conditions (C.2)–(C.4) and Taylor's expansion, on A_n,

$$\left|\tilde{\mathbf{s}}_{j,n} - \mathrm{sgn}(\beta_{j,n})|\beta_{j,n}|^{-\gamma}\right| \leq K\gamma\kappa_n^{-(\gamma+1)}a_n \tag{6.13}$$

uniformly in $j \in \{1, \ldots, p_0\}$, for all $n \geq n_0$. Now using Condition (C.3)–(C.5), (6.13), Theorem 1 and Lindeberg Central Limit Theorem, it can be shown that for any $x_1 < x_2 < \cdots < x_k$ $(1 \leq k < \infty)$,

$$\left(\sqrt{n}[\widehat{F}_n(x_1) - F(x_1)], \ldots, \sqrt{n}[\widehat{F}_n(x_k) - F(x_k)]\right)' \to^d N(\mathbf{0}, \Sigma)$$

where $\to^d$ denotes convergence in distribution on $\mathbb{R}^k$ and where Σ is a $k \times k$ matrix with (i, j)th element given by $\tau(x_i, x_j)$, $1 \leq i, j \leq k$. This shows that the finite dimensional distributions of $\mathcal{Z}_n$ converge to those of $\mathcal{Z}_\infty$. It now remains to establish asymptotic tightness of $\mathcal{Z}_n$. In view of Theorem 1, this would follow from the asymptotic tightness of $\{Z_n(F(x)) : x \in [-\infty, \infty]\}$ and the tightness of the sequence of random variables $\{n^{-1/2}\bar{\boldsymbol{x}}_n^{(1)'}\mathbf{U}_n^{(1)}\}$. The first assertion is a standard result on the empirical process of iid random variables. As for the second, note that by Condition (C.3), $\mathrm{Var}(\bar{\boldsymbol{x}}_n^{(1)'}C_{11,n}^{-1}\boldsymbol{W}_n^{(1)}) = \bar{\boldsymbol{x}}_n^{(1)'}C_{11,n}^{-1}\bar{\boldsymbol{x}}_n^{(1)}\sigma^2 = O(1)$. Hence, using (6.8) and (6.13), we have

$$\begin{aligned}\bar{\boldsymbol{x}}_n^{(1)'}\mathbf{U}_n^{(1)} &= \bar{\boldsymbol{x}}_n^{(1)'}C_{11,n}^{-1}\boldsymbol{W}_n^{(1)} + o_P(1)\\ &= O_P(1).\end{aligned}$$

Therefore, $\mathcal{Z}_n$ is asymptotically tight. Since the limiting Gaussian process has continuous sample paths with probability one, the FCLT follows. This completes the proof of Theorem 2.

Proof of Corollary 1. Follows from Theorem 2 and the Continuous mapping theorem.

Proof of Theorem 3. Note that by Theorem 2, for any $\alpha \in (0, 1)$,

$$\widehat{F}_n^{-1}(\alpha) \to_p F^{-1}(\alpha).$$

Next, define the "ideal" interval J_α^* by replacing $\hat{q}_t$ with $q_t \equiv F^{-1}(t)$ for $t = \alpha/2, 1 - \alpha/2$. Then, it is easy to check that

$$P(y_{\mathbf{x}} \in J_\alpha^* | \epsilon_1, \ldots, \epsilon_n) = [1 - \alpha] + o_p(1).$$

Theorem 3 follows by combining the two assertions. We omit the routine details.

Acknowledgments

The authors thank four referees and the handling Co-Editors of the paper for the Annals volume, Professors Marine Carrasco and Silvia Goncalves for a number of constructing comments that significantly improved an earlier version of the paper.

References

Bach, F., 2009. Model-consistent sparse estimation through the bootstrap. Preprint available at http://arxiv.org/abs/0901.3202.
Belloni, A., Chernozhukov, V., Hansen, C., 2011a. Inference for high-dimensional sparse econometric models. Preprint. arXiv:1201.0220.
Belloni, A., Chernozhukov, V., Hansen, C., 2011b. Inference on treatment effects after selection amongst high-dimensional controls. Preprint. arXiv:1201.0224.
Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of lasso and Dantzig selector. Ann. Statist. 37 (4), 1705–1732.
Bühlmann, P., van de Geer, S., 2011. Statistics for high-dimensional data. Methods, theory and applications. In: Springer Series in Statistics, Springer, Heidelberg.
Bunea, F., Tsybakov, A., Wegkamp, M., 2007. Sparsity oracle inequalities for the lasso. Electron. J. Stat. 1, 169–194.
Chatterjee, A., Lahiri, S.N., 2013. Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. Ann. Statist. 41 (3), 1232–1259.
Chatterjee, A., Lahiri, S.N., 2011b. Strong consistency of Lasso estimators. Sankhya A 73 (1), 55–78.
Chatterjee, A., Lahiri, S.N., 2011a. Bootstrapping Lasso estimators. J. Amer. Statist. Assoc. 106 (494), 608–625.
Chen, G., Lockhart, R.A., 2001. Weak convergence of the empirical process of residuals in linear models with many parameters. Ann. Statist. 29, 748–762.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.
Fuk, D.K., Nagaev, S.V., 1971. Probability inequalities for sums of independent random variables. Theory Probab. Appl. 16, 643–660.
Gautier, E., Tsybakov, A., 2011. High-dimensional instrumental variables regression and confidence sets. Preprint. arXiv:1105.2454v2.
Gupta, S., 2012. A note on the asymptotic distribution of lasso estimator for correlated data. Sankhya, Series A 74, 10–28.
Huang, J., Horowitz, J.L., Ma, S., 2008a. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Statist. 36 (2), 587–613.
Huang, J., Ma, S., Zhang, C.-H., 2008b. Adaptive Lasso for sparse high-dimensional regression models. Statist. Sinica 18 (4), 1603–1618.
Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. Ann. Statist. 28 (5), 1356–1378.
Koul, H.L., 1969. Asymptotic behavior of wilcoxon type confidence regions in multiple linear regression. Ann. Math. Statist. 40, 1950–1979.
Koul, H.L., 2002. Weighted Empirical processes in Dynamic Nonlinear Models. In: Lecture Notes in Statistics, Springer, New York.
Leeb, H., Pötscher, B.M., 2008. Sparse estimators and the oracle property, or the return of Hodges' estimator. J. Econometrics 142 (1), 201–211.
Liao, Z., Phillips, P.C., 2012. Automated estimation of vector error correction models. Preprint, (1873). Cowles Foundation for Research in Economics, Yale University.
Mammen, E., 1996. Empirical process of residuals for high-dimensional linear models. Ann. Statist. 24, 307–335.
Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. Ann. Statist. 34 (3), 1436–1462.
Meinshausen, N., Yu, B., 2009. Lasso-type recovery of sparse representations for high-dimensional data. Ann. Statist. 37 (1), 246–270.
Portnoy, S., 1984. Asymptotic behaviour of m-estimators of p regression parameters when p^2/n is large. i. consistency. Ann. Statist. 12, 1298–1309.
Portnoy, S., 1985. Asymptotic behaviour of m-estimators of p regression parameters when p^2/n is large. ii.normal approximation. Ann. Statist. 13, 1403–1417.
Portnoy, S., 1986. Asymptotic behaviour of the empiric distribution of m-estimated residuals from a regression model with many parameters. Ann. Statist. 14, 1152–1170.
Pötscher, B.M., Schneider, U., 2009. On the distribution of the adaptive LASSO estimator. J. Statist. Plann. Inference 139 (8), 2775–2790.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
van der Vaart, A.W., Wellner, J.A., 1996. Weak convergence and empirical processes. In: Springer Series in Statistics, Springer-Verlag, New York, With applications to statistics.
Wainwright, M.J., 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inform. Theory 55 (5), 2183–2202.
Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. Biometrika 94 (1), 19–35.
Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38, 894–942.
Zhang, C.-H., Huang, J., 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. Ann. Statist. 36, 1567.
Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. J. Mach. Learn. Res. 7, 2541–2563.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101 (476), 1418–1429.