| Statistica Sinica Preprint No: SS-2018-0507 | |
|---|---|
| **Title** | Calibrated zero-norm regularized LS estimator for high-dimensional error-in-variables regression |
| **Manuscript ID** | SS-2018-0507 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202018.0507 |
| **Complete List of Authors** | Ting Tao<br>Shaohua Pan and<br>Shujun Bi |
| **Corresponding Author** | Shujun Bi |
| **E-mail** | bishj@scut.edu.cn |

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507

Statistica Sinica

# Calibrated zero-norm regularized LS estimator for high-dimensional error-in-variables regression

Ting Tao, Shaohua Pan and Shujun Bi

*School of Mathematics, South China University of Technology, Guangzhou.*

*Abstract:* This study focuses on using a high-dimensional error-in-variables regression to identify a small number of important interpretable factors from corrupted data in applications in which measurement errors or missing data cannot be ignored. Motivated by the convex conditioned Lasso (CoCoLasso) method and the advantage of using a zero-norm regularized LS estimator rather than a Lasso for clean data, we propose a calibrated zero-norm regularized LS (CaZnRLS) estimator. To do so, we construct a calibrated least squares loss with a positive-definite projection of an unbiased surrogate for the covariance matrix of covariates. Then, we use the multi-stage convex relaxation approach to compute the proposed estimator. Under restricted strong convexity on the true covariate matrix, we derive the $\ell_2$-error bound for each iteration. Then, we establish the decreasing error bound sequence and the sign consistency of the iterations after a finite number of steps. Statistical guarantees are also provided for the CaZnRLS estimator under two types of measurement errors. Numerical comparisons with the CoCoLasso and nonconvex Lasso show that the CaZnRLS has a better relative RMSE and correctly identifies more of the predictors.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



*Key words and phrases:* Error-in-variables regression, high-dimensional, multi-stage convex relaxation, zero-norm regularized LS.

## 1. Introduction

High-dimensional regressions are becoming popular in many fields, including genomics, finance, image processing, climate science, sensor networks, and so on. The canonical high-dimensional linear regression model assumes that the number of available predictors $p$ is larger than the sample size $n$, although the number of true relevant predictors $s$ is much less than $p$. This model can be expressed as

$$y = X\beta^* + \varepsilon, \tag{1.1}$$

where $y = (y_1, \ldots, y_n)^{\mathbb{T}}$ is the vector of responses, $X = (x_{ij})$ is an $n \times p$ matrix of covariates, $\beta^* \in \mathbb{R}^p$ is a sparse coefficient vector with $s$ nonzero entries, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathbb{T}}$ is the noise vector. Unless otherwise stated, we assume all covariates are centered such that the intercept term is not included in (1.1) and the matrix $X$ of covariates has normalized columns.

Current popular high-dimensional regression methods include convex-type estimators, such as the Lasso of Tibshirani (1996), adaptive Lasso of Zou (2006), elastic net of Zou and Hastie (2005), and Dantzig selector of Candès and Tao (2007), and nonconvex-type estimators, such as the smoothly clipped absolute deviation (SCAD) of Fan and Li (2001) and

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



minimax concave penalty (MCP) of Zhang (2010). Refer to Fan and Lv (2010) and Bühlmann and van de Geer (2011) for excellent overviews of these methods. To some extent, these methods imitate the performance of the zero-norm penalized LS estimator

$$\beta^{\mathrm{zn}} \in \underset{\|\beta\|_\infty \le R}{\arg\min} \left\{ \frac{1}{2n\lambda} \|y - X\beta\|^2 + \|\beta\|_0 \right\}, \tag{1.2}$$

where the ball constraint $\|\beta\|_\infty \le R$, for some $R > 0$, ensures the well-definedness of $\beta^{\mathrm{zn}}$, and $\lambda > 0$ is the regularization parameter. Recently, by developing a global exact penalty for the equivalent mathematical program with equilibrium constraints (MPEC), Bi and Pan (2018) showed that a global optimal solution can be obtained for (1.2) from the solution of a global exact penalization problem. In addition, the popular SCAD estimator is the result of eliminating the dual part of a global exact penalization problem. By solving such a problem in an alternating way, they proposed a multi-stage convex relaxation approach (GEP-MSCRA), which can be regarded as an adaptive Lasso that incorporates dual information. Note that for the clean design matrix $X$, the zero-norm regularized LS estimator computed using the GEP-MSCRA has a remarkable advantage over the Lasso in terms of reducing the prediction error and capturing the sparsity.

In reality, we often face corrupted data, owing to inaccurate observations for covariates, or missing values. Common examples include sen-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



sor network data (see Slijepcevic, Megerian, and Potkonjak (2002)), high-throughout sequencing (see Benjamini and Speed (2012)), and gene expression data (see Purdom and Holmes (2005)). In this setting, naively applying the high-dimensional regression method for clean data to corrupted data will yield misleading inference results; see Rosenbaum and Tsybakov (2010). Then, it is natural to ask how to modify the zero-norm regularized LS estimator for corrupted data, without losing its advantages. Motivated by the convex conditioned Lasso (CoCoLasso) method of Datta and Zou (2017), we propose a calibrated zero-norm regularized LS (CaZnRLS) estimator. For convenience, we assume that a corrupted covariate matrix $Z = (z_{ij})_{n\times p}$ rather than the true covariate matrix $X$ is observed. As mentioned in Loh (2014) and Datta and Zou (2017), depending on the context, there are various ways to model measurement errors. For example, in an additive noise setting, $Z = X + A$, where $A = (a_{ij})_{n\times p}$ is the additive noise matrix. In a multiplicative errors setup, $Z = X \circ M$, where $M = (m_{ij})_{n\times p}$ is the matrix of multiplicative errors, and "$\circ$" denotes the elementwise multiplication operator. Note that missing values can be viewed as a special case of multiplicative errors.

The loss term $\frac{1}{2n}\|y - X\beta\|^2$ in the clean setting can be rewritten as

$$\frac{1}{2}\beta^{\mathbb{T}}\Sigma\beta - \xi^{\mathbb{T}}\beta + \frac{1}{2n}\|y\|^2 \quad \text{with } \Sigma := \frac{1}{n}X^{\mathbb{T}}X \text{ and } \xi := \frac{1}{n}X^{\mathbb{T}}y. \tag{1.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



By recalling that the covariates are centered, it is easy to check that $(\Sigma, \xi)$ is an unbiased estimator of $(\Sigma_x, \Sigma_x\beta^*)$, where $\Sigma_x$ denotes the covariance matrix of the covariates. Using corrupted $Z$ and $y$, Loh and Wainwright (2012) constructed an unbiased surrogate $(\widehat{\Sigma}, \widehat{\xi})$ of $(\Sigma, \xi)$, and obtained an estimation for the true $\beta^*$ using the following optimization model:

$$\widehat{\beta} \in \underset{\|\beta\|_1 \le R_0}{\arg\min} \left\{ \frac{1}{2}\beta^{\mathbb{T}}\widehat{\Sigma}\beta - \widehat{\xi}^{\mathbb{T}}\beta + \lambda_n\|\beta\|_1 \right\}. \tag{1.4}$$

Note that the unbiased surrogate $\widehat{\Sigma}$ constructed from $Z$ may not be positive semidefinite (PSD); for example, when $x_{ij}$ is corrupted by independent additive errors $a_{ij}$ with mean zero and variance $\tau^2$, the matrix $\widehat{\Sigma} = \frac{1}{n}Z^{\mathbb{T}}Z - \tau^2 I$ is an unbiased surrogate for $\Sigma$, which has a negative eigenvalue because $n < p$. As a result, the objective function of (1.4) may be nonconvex and lower unbounded. Loh and Wainwright imposed the constraint $\|\beta\|_1 \le R_0$ on model (1.4) to guarantee that it has an optimal solution. Through some careful analysis, they showed that if $R_0$ is chosen properly, a projected gradient descent algorithm will converge in polynomial time to a small neighborhood of the set of all global minimizers. However, as remarked in Datta and Zou (2017), the practical performance of the nonconvex Lasso model (1.4) depends greatly on the choice of $R_0$. A similar shortcoming applies to the procedure of Chen and Caramanis (2013).

To overcome the aforementioned shortcoming and benefit from the con-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



vex formulation of the Lasso, Datta and Zou (2017) proposed the CoCo-Lasso method. Let $W \succeq \widehat{\epsilon} I$ mean that $W - \widehat{\epsilon} I$ is PSD and let $\|Z\|_{\max} = \max_{i,j} |z_{ij}|$ denote the elementwise maximum norm of a matrix $Z$. They first solved the following PSD optimization problem:

$$\overline{\Sigma} \in \underset{W \succeq \widehat{\epsilon} I}{\arg\min} \|W - \widehat{\Sigma}\|_{\max} \quad \text{for some } \widehat{\epsilon} > 0, \tag{1.5}$$

to obtain the nearest positive definite (PD) approximation to the unbiased surrogate $\widehat{\Sigma}$ of $\Sigma$, constructed as in Loh and Wainwright (2012) using $Z$. Then, they defined

$$\overline{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\overline{y} - \overline{Z}\beta\|^2 + \lambda \|\beta\|_1 \right\} \tag{1.6}$$

using the Cholesky factor $\overline{Z}/\sqrt{n}$ of $\overline{\Sigma}$ and the vector $\overline{y}$ satisfying $\overline{Z}^{\mathbb{T}} \overline{y} = Z^{\mathbb{T}} y$.

The elementwise maximum norm in model (1.5) plays two roles: measuring the approximation of $\overline{\Sigma}$ to $\Sigma$, and removing a particular noise from $\widehat{\Sigma}$. Compared with other elementwise norms, such as the $\ell_1$-norm and Frobenius norm, the maximum norm yields an approximation with entries that are closer to those of $\widehat{\Sigma}$. However, the computation of $\overline{\Sigma}$ is expensive, because model (1.5) is a convex program of $p^2$ variables involving two nonsmooth terms: the objective function $\|W - \widehat{\Sigma}\|_{\max}$, and the PSD constraint. Figure 1 indicates that when using the alternating direction method of multipliers (ADMM), described in Appendix A of Datta and Zou (2017),

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



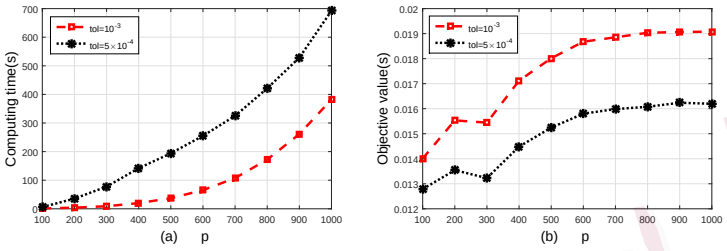


Figure 1: The computing time and objective value of Algorithm 1 of Datta and Zou (2017) with the stopping condition $\max(\|A_{i+1}-A_i\|_F, \|B_{i+1}-B_i\|, \|\Lambda_{i+1}-\Lambda_i\|) \leq \text{tol}$

to solve (1.5), using $\widehat{\Sigma}$ from the data in Subsection 5.1, the computing time increases quickly with $p$ and the improvement in the accuracy of the solution. Now, consider using (1.5) to approximate the covariance matrix $\Sigma_x$ instead of the noisy unbiased surrogate $\widehat{\Sigma}$. Here, it is reasonable to seek a slightly less accurate approximation that can be achieved cheaply, and then to employ a more effective high-dimensional regression method than the Lasso to define the estimator. When the elementwise maximum norm in (1.5) is replaced with the Frobenius norm, the solution is exactly the projection of $\widehat{\Sigma} - \widehat{\epsilon} I$ onto the PSD cone, and can be obtained from one eigenvalue decomposition for $\widehat{\Sigma}$. In addition, when $\widehat{\Sigma} = \frac{1}{n} Z^{\mathbb{T}} Z - \tau^2 I$, this solution matches the structure of $\widehat{\Sigma}$ well. Motivated by this, we replace the objective function of (1.5) with the Frobenius norm of $W - \widehat{\Sigma}$ to obtain an approximation $\widetilde{\Sigma}$. Then, we define a zero-norm regularized LS estimator using its eigenvalue decomposition.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



Note that a Dantzig selector-type estimator and an improved version were proposed in (Rosenbaum and Tsybakov, 2010, 2013) and Belloni, Rosenbaum and Tsybakov (2017), respectively, for additive measurement error models. Because these estimators are defined via an optimization problem with a difference of convexity (D.C.) constraint, it is difficult to obtain these estimators in practice. To overcome this difficulty, Belloni, Rosenbaum, and Tsybakov (2016) relaxed the nonconvex constraint set to a convex set, and proposed two conic programming-based estimators for the same model setup, which can be viewed as a relaxed version of the Dantzig selector for clean data. In addition, Städler and Bühlmann (2012) derived an algorithm for a sparse linear regression with missing data, based on a sparse inverse covariance matrix estimation. In the spirit of Loh and Wainwright (2012) and Datta and Zou (2017), we propose the CaZnRLS estimator, which simultaneously handles additive errors, multiplicative errors, and missing data. Although the CaZnRLS estimator is defined using a nonconvex optimization problem, the GEP-MSCRA in Bi and Pan (2018) (see Section 3) provides an efficient solver this problem that solves a sequence of weighted $\ell_1$-regularized LS problems. As shown in the simulation study in Section 5, the estimator still reduces prediction error and captures the sparsity for the contaminated data, as it does for the clean data.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



The rest of this paper is organized as follows. In Section 2, we define the CaZnRLS estimator and provide a primal-dual view of this estimator. Section 3 describes the GEP-MSCRA solver used to compute the CaZnRLS estimator. In Section 4, under a restricted eigenvalue assumption on the matrix $\Sigma$, we provide the deterministic theoretical guarantees, including the $\ell_2$-error bound for every iteration, decreasing error bound sequence, and sign consistency of the iterations, after a finite number of steps. Here, we also provide the statistical guarantees for the computed estimator under two types of measurement error. In Section 5, we compare the performance of the CaZnRLS estimator with that of the CoCoLasso and nonconvex Lasso (NCL). All proofs and technical details are provided in the online Supplementary Material.

To close this section, we introduce some necessary notation. Let $\mathbb{S}^p$ be the space consisting of all $p \times p$ real symmetric matrices, equipped with the trace inner product $\langle W, Y \rangle = \mathrm{trace}(W^{\mathbb{T}} Y)$ and its induced Frobenius norm $\|\cdot\|_F$, and let $\mathbb{S}^p_+$ be the cone consisting of all PSD matrices in $\mathbb{S}^p$. For any symmetric matrix $W$, let $\lambda_{\min}(W)$ and $\lambda_{\max}(W)$ denote the smallest and largest eigenvalues, respectively, of $W$. For any vector $z$, $\|z\|_\infty$ denotes the infinity norm of $z$. Let $I$ and $e$ denote an identity matrix and a vector of all ones, respectively, with dimensions that are known from the context.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



For a closed set $\Omega$, $\delta_\Omega(\cdot)$ denotes the indicator function on $\Omega$. That is, $\delta_\Omega(x) = 0$ if $x \in \Omega$; otherwise, $\delta_\Omega(x) = +\infty$. When $\Omega$ is convex, $\Pi_\Omega(\cdot)$ denotes the projection operator onto $\Omega$. For an index set $\Lambda \subseteq \{1, \ldots, p\}$, write $\Lambda^c := \{1, \ldots, p\} \backslash \Lambda$, and denote $\mathbb{I}_\Lambda(\cdot)$ as the characterization function on $\Lambda$, and $Y_\Lambda$ as the submatrix of $Y$ consisting of the column $Y_j$, for $j \in \Lambda$. For any nonnegative real number $a$, $\lfloor a \rfloor$ and $\lceil a \rceil$ denote the largest integer less than $a$ and the smallest integer greater than $a$, respectively.

## 2. The CaZnRLS estimator

When the data are corrupted by measurement errors, the observed matrix $Z$ of predictors is a function of the true covariate matrix $X$ and random errors. In this case, one may construct an unbiased surrogate $(\widehat{\Sigma}, \widehat{\xi})$ for the pair $(\Sigma, \xi)$ using $Z$ and $y$, as in Loh and Wainwright (2012). For the specific form of $(\widehat{\Sigma}, \widehat{\xi})$ under various types of measurement errors, refer to Section 2 in Loh and Wainwright (2012), or see the Supplementary Material. Now, assume that an unbiased surrogate $(\widehat{\Sigma}, \widehat{\xi})$ is available. Let $\widehat{\Sigma}$ have the eigenvalue decomposition $\widehat{\Sigma} = P\mathrm{Diag}(\theta_1, \ldots, \theta_p)P^{\mathbb{T}}$, where $P$ is a $p \times p$ orthonormal matrix and $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_p$ are the eigenvalues of $\widehat{\Sigma}$.

Because it is time-consuming to compute a solution for (1.5) when $p$ is large, we replace elementwise maximum norm in (1.5) with Frobenius norm,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



and achieve a nearest PD approximation to $\widehat{\Sigma}$ using the following model:

$$\widetilde{\Sigma} = \underset{W \succeq \widehat{\epsilon} I}{\arg\min} \|W - \widehat{\Sigma}\|_F. \tag{2.1}$$

Note that (2.1) has the same solution set as the problem $\min_{W \succeq \widehat{\epsilon} I} \|W - \widehat{\Sigma}\|_F^2$ does. Thus,

$$\widetilde{\Sigma} = \widehat{\epsilon} I + \Pi_{\mathbb{S}_+^p}(\widehat{\Sigma} - \widehat{\epsilon} I) = P \mathrm{Diag}\big(\max(\theta_1, \widehat{\epsilon}), \ldots, \max(\theta_p, \widehat{\epsilon})\big) P^{\mathbb{T}}. \tag{2.2}$$

Clearly, when $\widehat{\Sigma} = \frac{1}{n} Z^{\mathbb{T}} Z - \tau^2 I$, a composite of a low-rank matrix and an identity matrix, the solution $\widetilde{\Sigma}$ keeps this structure. Furthermore, one eigenvalue decomposition of $\frac{1}{n} Z^{\mathbb{T}} Z$ is enough to formulate the solution $\widetilde{\Sigma}$. Indeed, let

$$\begin{cases} \widetilde{Z} := \sqrt{n} P \mathrm{Diag}\Big(\sqrt{\max(\theta_1, \widehat{\epsilon})}, \ldots, \sqrt{\max(\theta_p, \widehat{\epsilon})}\Big) P^{\mathbb{T}}, \\ \widetilde{y} := \sqrt{n} P \mathrm{Diag}\Big(\frac{1}{\sqrt{\max(\theta_1, \widehat{\epsilon})}}, \ldots, \frac{1}{\sqrt{\max(\theta_p, \widehat{\epsilon})}}\Big) P^{\mathbb{T}} \widehat{\xi}. \end{cases} \tag{2.3}$$

Then, from (2.2), we have that $\widetilde{\Sigma} = \frac{1}{n} \widetilde{Z}^{\mathbb{T}} \widetilde{Z}$ and $\widehat{\xi} = \frac{1}{n} \widetilde{Z}^{\mathbb{T}} \widetilde{y}$.

Although the computation of $\widetilde{\Sigma}$ becomes much cheaper than that of $\overline{\Sigma}$, the accuracy of its approximation to $\widehat{\Sigma}$ is worse, because minimizing the elementwise maximum norm tends to give smaller entries. This requires that we define an estimator using high-dimensional regression methods that are more effective than the Lasso. A natural candidate is a nonconvex-type estimator, such as the SCAD or MCP, because they can remove the bias of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



the Lasso. Note that the SCAD and MCP functions are actually imitating the performance of the zero-norm. We define the zero-norm regularized LS estimator

$$\widetilde{\beta} \in \arg\min_{\beta\in\mathbb{R}^p} \Big\{ \frac{1}{2n\lambda}\|\widetilde{Z}\beta - \widetilde{y}\|^2 + \|\beta\|_0 \Big\}. \tag{2.4}$$

Taking into account that $(\widetilde{Z}, \widetilde{y})$ is a calibrated pair of $(\widehat{\Sigma}, \widehat{\xi})$, we call (2.4) a calibrated version of the zero-norm regularized LS estimator defined using the corrupted observation $Z$, as in (1.2), except that the ball constraint is now removed, owing to the coerciveness of the strong convex $\|\widetilde{Z}\beta - \widetilde{y}\|^2$. Compared with the SCAD estimator, the solution of (2.4) seems to be much more difficult, because the problem in (2.4) is even discontinuous, owing to the combinatorial property of the zero-norm. However, as demonstrated later, the SCAD estimator is actually equivalent to the zero-norm regularized LS.

Next, we provide a primal-dual view of the estimator $\widetilde{\beta}$. Define

$$\phi(t) := \frac{a-1}{a+1}t^2 + \frac{2}{a+1}t \ \ (a>1), \quad \text{for} \ \ t\in\mathbb{R}. \tag{2.5}$$

Using this function, we can immediately check that, for any $\beta \in \mathbb{R}^p$,

$$\|\beta\|_0 = \min_{w\in\mathbb{R}^p} \Big\{ \sum_{i=1}^p \phi(w_i) \colon \langle e-w, |\beta|\rangle = 0, \ 0\le w\le e \Big\}.$$

This shows that the zero-norm is essentially an optimal value function of a parameterized mathematical program with equilibrium constraints (M-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



PEC), because $\langle e-w, |\beta|\rangle = 0$ and $e - w \geq 0$ constitute an equilibrium constraint. Thus, (2.4) is equivalent to the following MPEC:

$$\min_{\beta, w \in \mathbb{R}^p} \left\{ \frac{1}{2n\lambda} \|\widetilde{Z}\beta - \widetilde{y}\|^2 + \sum_{i=1}^{p} \phi(w_i) \colon \langle e-w, |\beta|\rangle = 0,\, 0 \leq w \leq e \right\}, \quad (2.6)$$

in the sense that if $\widetilde{\beta}^{\natural}$ is a global optimal solution of (2.4), then $(\widetilde{\beta}^{\natural}, \mathrm{sign}(|\widetilde{\beta}^{\natural}|))$ is globally optimal to (2.6); conversely, if $(\widetilde{\beta}^{\natural}, \widetilde{w}^{\natural})$ is a global optimal solution of (2.6), then $\widetilde{\beta}^{\natural}$ is globally optimal to (2.4), with $\|\widetilde{\beta}^{\natural}\|_0 = \sum_{i=1}^{p} \phi(\widetilde{\beta}_i^{\natural})$.

The MPEC form (2.6) shows that the difficulty in computing the estimator $\widetilde{\beta}$ arises from the constraint $\langle e - w, |\beta|\rangle = 0$, which introduces the bothersome nonconvexity. Because it is much harder to handle nonconvex constraints than it is to handle a nonconvex objective, we consider its penalized version,

$$\min_{\beta \in \mathbb{R}^p, w \in [0,e]} \left\{ \frac{1}{2n\lambda} \|\widetilde{Z}\beta - \widetilde{y}\|^2 + \sum_{i=1}^{p} \phi(w_i) + \rho\langle e-w, |\beta|\rangle \right\}, \quad (2.7)$$

where $\rho > 0$ is the penalty parameter. By the coerciveness of the function $\beta \mapsto \|\widetilde{Z}\beta - \widetilde{y}\|^2$, there exists a constant $\widehat{R} > 0$ such that (2.6) and (2.7) are equivalent to their respective versions in which the variable $\beta$ is required to lie in the set $\{\beta \in \mathbb{R}^p \mid \|\beta\|_\infty \leq \widehat{R}\}$. Thus, invoking Theorem 2.1 of Bi and Pan (2018), we have the following result.

**Theorem 1.** *Let $L_f$ be the Lipschitz constant of $f(\beta) := \frac{1}{2n}\|\widetilde{Z}\beta - \widetilde{y}\|^2$ on the ball $\{\beta \in \mathbb{R}^p \colon \|\beta\|_\infty \leq \widehat{R}\}$. Then, for every $\rho \geq \overline{\rho} := \frac{4aL_f}{(a+1)\lambda}$, the global*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



*optimal solution set of* (2.7) *associated with* $\rho$ *coincides with that of* (2.6).

Theorem 1 shows that the problem in (2.7) is a global exact penalty of (2.6), in the sense that it has the same global optimal solution set as that of (2.6), once $\rho$ is greater than a threshold. Consequently, $\widetilde{\beta}$ can be achieved by solving the following exact penalty problem with $\rho > \overline{\rho}$:

$$\widetilde{\beta} \in \underset{\beta\in\mathbb{R}^p, w\in[0,e]}{\arg\min} \left\{ \frac{1}{2n}\|\widetilde{Z}\beta - \widetilde{y}\|^2 + \sum_{i=1}^{p} \lambda\Big[\phi(w_i) + \rho(1-w_i)|\beta_i|\Big] \right\}. \tag{2.8}$$

Compared with (2.4), the problem in (2.8) involves an additional variable $w \in \mathbb{R}^p$, which provides part of the dual information on (2.4). Hence, (2.8) can be viewed as a primal-dual equivalent form of (2.4). This form does not involve the combinatorial difficulty, and its nonconvexity is due only to the coupled term $\langle w, |\beta|\rangle$, which is clearly much easier to cope with. In particular, the SCAD function in Fan and Li (2001) is precisely the optimal value of the inner minimization in (2.8) w.r.t. $w$. To see this, we define

$$\psi(t) := \begin{cases} \phi(t) & \text{if } t \in [0,1], \\ +\infty & \text{otherwise.} \end{cases} \tag{2.9}$$

Recalling the conjugate $\psi^*(\omega) = \sup_{t\in\mathbb{R}}\{t\omega - \psi(t)\}$ of $\psi$ by Rockafellar (1970), we can compactly write the inner minimization in (2.8) w.r.t. $w$ as

$$\min_{w\in\mathbb{R}^p}\Big\{\textstyle\sum_{i=1}^{p}\lambda\big[\psi(w_i) + \rho(1-w_i)|\beta_i|\big]\Big\} = \textstyle\sum_{i=1}^{p}\lambda\big[\rho|\beta_i| - \psi^*\big(\rho|\beta_i|\big)\big].$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



After an elementary calculation, the conjugate $\psi^*$ of $\psi$ has the form

$$\psi^*(\omega) = \begin{cases} 0 & \text{if } \omega \le \frac{2}{a+1}, \\ \frac{((a+1)\omega-2)^2}{4(a^2-1)} & \text{if } \frac{2}{a+1} < \omega \le \frac{2a}{a+1}, \\ \omega - 1 & \text{if } \omega > \frac{2a}{a+1}. \end{cases}$$

By comparing this with the expression of the SCAD function $p_\gamma(t)$, the function $\lambda[\rho|t| - \psi^*(\rho|t|)]$, with $\lambda = \frac{(a+1)\gamma^2}{2}$ and $\rho = \frac{2}{(a+1)\gamma}$, reduces to $p_\gamma(t)$. Thus,

$$\widetilde{\beta} \in \arg\min_{\beta\in\mathbb{R}^p} \Big\{ \frac{1}{2n}\|\widetilde{Z}\beta - \widetilde{y}\|^2 + \textstyle\sum_{i=1}^p p_\gamma(|\beta_i|) \Big\}. \tag{2.10}$$

## 3. GEP-MSCRA for computing the estimator $\widetilde{\beta}$

From the previous section, to compute the estimator $\widetilde{\beta}$, one need only solve a single penalty problem (2.8), which is much easier than (2.4) because its nonconvexity is from the coupled term $\langle w, |\beta|\rangle$. The GEP-MSCRA proposed by Bi and Pan (2018) makes good use of the coupled structure, and solves the problem in (2.8) in an alternating way. Because the threshold $\overline{\rho}$ is unknown, though one may obtain an upper estimation for it, a varying $\rho$ is introduced in the GEP-MSCRA. The iterations of the GEP-MSCRA are described below.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



**Algorithm 1 GEP-MSCRA for computing $\widetilde{\beta}$**

**Initialization:** Choose $\lambda > 0, \rho_0 = 1$ and an initial $w^0 \in [0, \frac{1}{2}e]$. Set $k := 1$.

**while** the stopping conditions are not satisfied **do**

1. Compute the following minimization problem

$$\beta^k = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \left\| \widetilde{Z}\beta - \widetilde{y} \right\|^2 + \lambda \sum_{i=1}^{p} (1 - w_i^{k-1}) |\beta_i| \right\}. \tag{3.1}$$

2. When $k = 1$, select a suitable $\rho_1 \geq \rho_0$ in terms of $\|\beta^1\|_\infty$. Otherwise, select $\rho_k$ such that $\rho_k \geq \rho_{k-1}$ for $k \leq 3$; and $\rho_k = \rho_{k-1}$ for $k > 3$.

3. Seek the unique optimal solution $w_i^k$ $(i = 1, \ldots, p)$ of the problem

$$w_i^k = \arg\min_{0 \leq w_i \leq 1} \left\{ \phi(w_i) - \rho_k w_i |\beta_i^k| \right\}. \tag{3.2}$$

4. Let $k \leftarrow k + 1$, and then go to Step 1.

**end while**

**Remark 1. (a)** Because $\phi$ is strongly convex, the problem in (3.2) has a unique optimal solution. From $\phi$, we immediately obtain

$$w_i^k = \min \left[ 1, \max \left( \frac{(a+1)\rho_k |\beta_i^k| - 2}{2(a-1)}, 0 \right) \right] \quad \text{for} \quad i = 1, 2, \ldots, p. \tag{3.3}$$

Thus, the main computation in each step is solving a weighted $\ell_1$-regularized LS. In this sense, the GEP-MSCRA is analogous to the local linear approx-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



imation algorithm of Zou and Li (2008) applied to the problem in (2.10), except for the start-up and the weights. The start-up of the former depends explicitly on the dual variable $w^0$, whereas that of the latter depends implicitly on a good estimator $\beta^0$. Therefore, when computing CaZnRLS using the GEP-MSCRA, one actually obtains an adaptive Lasso estimator. The initial $w^0$ may be an arbitrary vector from the box set $[0, \frac{1}{2}e]$. Here, we restrict $w^0$ to the box set $[0, \frac{1}{2}e]$, rather than the feasible set $[0, e]$ of $w$ in (2.8), so as to achieve a better initial estimator $\beta^1$.

**(b)** Owing to the combinatorial property of $\|\cdot\|_0$, it is almost impossible to obtain $\widetilde{\beta}$ exactly. The popular Lasso of Tibshirani (1996) and adaptive Lasso of Zou (2006), as a one-step and a series of convex relaxations to (2.4), respectively, arise from the primal angle, whereas the series of weighted $\ell_1$-norm regularized LS problems in the GEP-MSCRA arise from the primal-dual reformulation of (2.4).

**(c)** From the formula in (3.3), if $\rho_k|\beta_i^k|$ is larger, then $w_i^k$ has a value close to one, Thus, in the $(k+1)$th iteration, a smaller weight $(1-w_i^k)$ is imposed on the variable $\beta_i$, and consequently a conservative strategy is used for sparsity. Consider that for some difficult problems, the solution $\beta^1$ yielded by the $\ell_1$-regularized LS problem may not have a sharp gap between its nonzero and zero entries. Hence, in order to guarantee that the subsequent $\beta^k$ has a

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



correct sparse support, we increase $\rho_k$, for $k \leq 3$, appropriately; that is, we cut down the smaller nonzero entries conservatively. For $k > 3$, in general, $\beta^k$ has a big difference between its nonzero and zero entries. Therefore, we keep $\rho_k$ unchanged so as to cut down the smaller nonzero entries quickly.

In the Supplementary Material, we implement the GEP-MSCRA by applying the semi-smooth Newton augmented Lagrangian method (ALM) to the dual of (3.1). As discussed in Li, Sun, and Toh (2018), the semi-smooth Newton ALM fully exploits the second-order information and good structure of its dual, and can yield an accurate solution.

## 4. Theoretical guarantees for the GEP-MSCRA

In this section, we denote $S^*$ as the support of the true vector $\beta^*$, and define

$$\mathcal{C}(S^*) := \bigcup_{S \supset S^*, |S| \leq 1.5s} \Big\{ \beta \in \mathbb{R}^p \colon \|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1 \Big\}.$$

We say that $\Sigma$ satisfies the $\kappa$-restricted eigenvalue condition (REC), or $X$ satisfies the $\kappa$-restricted strong convexity on $\mathcal{C}(S^*)$, if $\kappa > 0$ is such that

$$\beta^{\mathbb{T}}\Sigma\beta = \frac{1}{n}\|X\beta\|^2 \geq \kappa\|\beta\|^2 \quad \text{for all } \beta \in \mathcal{C}(S^*).$$

This REC is a little stronger than that used in Negahban et al. (2012) for the clean Lasso and in Datta and Zou (2017) for the CoCoLasso, because $\mathcal{C}(S^*) \supseteq \big\{\beta \in \mathbb{R}^p : \|\beta_{(S^*)^c}\|_1 \leq 3\|\beta_{S^*}\|_1\big\}$, and is different from the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



$(L, S^*, N)$-restricted eigenvalue condition introduced in van de Geer and Bühlmann (2009). We provide the deterministic theoretical guarantees for the GEP-MSCRA under this REC with appropriate $\lambda, \rho_1$, and $\rho_3$. These include the error bound of every iteration $\beta^k$ to the true $\beta^*$, decrease of the error sequence, and sign consistency of $\beta^k$ after a finite number of steps.

### 4.1. Error bound sequence and its decrease

To determine the error bound of iteration $\beta^k$ to the true $\beta^*$, we write

$$D := \widehat{\Sigma} - \Sigma \quad \text{and} \quad \widetilde{\varepsilon} := \widehat{\xi} - \widetilde{\Sigma}\beta^*. \tag{4.4}$$

The following theorem states a deterministic result for the error bound.

**Theorem 2.** *Suppose $\Sigma$ satisfies the $\kappa$-REC on $\mathcal{C}(S^*)$, with $\kappa > 24s\|D\|_{\max}$. If $\lambda$ and $\rho_3$ are chosen such that $\lambda \geq 8\|\widetilde{\varepsilon}\|_\infty$ and $\rho_3 \leq \frac{2(\kappa - 24s\|D\|_{\max})}{5\sqrt{2}\lambda}$, then*

$$\|\beta^k - \beta^*\| \leq \frac{5\sqrt{s}\,\lambda}{2(\kappa - 24s\|D\|_{\max})} \qquad \forall k \in \mathbb{N}. \tag{4.5}$$

The error bound in Theorem 2 has the same order, that is, $O(\lambda\sqrt{s})$, as that established for the clean Lasso by Negahban et al. (2012). From the proof of Theorem 1 in Datta and Zou (2017), $\|D\|_{\max} \leq \frac{\kappa}{64s}$ holds with a high probability. Therefore, there is a high probability that the error bound of $\beta^k$ is not greater than $\frac{4\lambda\sqrt{s}}{\kappa}$, which is a little better than the bound $\frac{4\sqrt{2}\lambda\sqrt{s}}{\kappa}$ in Datta and Zou (2017). However, $\lambda$ is allowed to be greater than $8\|\widetilde{\varepsilon}\|_\infty$, instead of $2\|\widetilde{\varepsilon}\|_\infty$, as in Datta and Zou (2017).

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



Theorem 2 provides an error bound for each iteration, but does not tell us if the error bound of the current $\beta^k$ is better than that of the previous $\beta^{k-1}$. To answer this question, we study the decrease of the error bound sequence by bounding $(1-w_i^k)^2$, for $i \in S^*$. Write $F^0 := S^*$ and, for $k \in \mathbb{N}$, define

$$F^k := \Big\{ i \colon \big| |\beta_i^k| - |\beta_i^*| \big| \geq (\rho_k)^{-1} \Big\} \text{ and } \Lambda^k := \Big\{ i \colon |\beta_i^*| \leq \frac{4a}{(a+1)\rho_k} \Big\}. \quad (4.6)$$

By Lemma 3, $(1-w_i^k)^2$, for $i \in S^*$, can be controlled by $\max(\mathbb{I}_{\Lambda^k}(i), \mathbb{I}_{F^k}(i))$. As a result, we have the following error bound involving $\mathbb{I}_{\Lambda^k}(i)$.

**Theorem 3.** *Suppose $\Sigma$ satisfies the $\kappa$-REC on $\mathcal{C}(S^*)$, with $\kappa > 24s\|D\|_{\max}$. If $\lambda$ and $\rho_3$ are chosen in the same way as in Theorem 2, then*

$$\begin{aligned} \big\|\beta^k - \beta^*\big\| \leq{}& \frac{4+2\sqrt{2}}{\kappa - 24s\|D\|_{\max}} \big\|\widetilde{\varepsilon}_{S^*}\big\| + \Big(\frac{1}{\sqrt{2}}\Big)^{k-1} \big\|\beta^1 - \beta^*\big\| \\ &+ \frac{2\lambda}{\kappa - 24s\|D\|_{\max}} \sum_{j=1}^{k-1} \sqrt{\textstyle\sum_{i\in S^*} \mathbb{I}_{\Lambda^j}(i)} \Big(\frac{1}{\sqrt{2}}\Big)^{k-1-j} \quad \forall k \in \mathbb{N}. \end{aligned}$$

The error bound in Theorem 3 consists of three parts: a statistical error $\|\widetilde{\varepsilon}_{S^*}\|$ induced by noise, an identification error $\sum_{j=1}^{k-1} \sqrt{\sum_{i\in S^*} \mathbb{I}_{\Lambda^j}(i)} (\frac{1}{\sqrt{2}})^{k-1-j}$ related to the choice of $\rho_j$, and a computation error $(\frac{1}{\sqrt{2}})^{k-1} \|\beta^1 - \beta^*\|$. By the definition of $\Lambda^j$, if $\rho_j$ is chosen such that $\rho_j > \frac{4a}{(a+1)\min_{i\in S^*}|\beta_i^*|}$, then the identification error becomes zero. Consequently, the error bound sequence decreases to the statistical error $\|\widetilde{\varepsilon}_{S^*}\|$ as $k$ increases. Clearly, if $\min_{i\in S^*}|\beta_i^*|$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



is not too small, it is easy to choose such $\rho_j$. In the next section, we provide an explicit choice range of $\rho_j$ such that the identification error is zero. From Theorem 3, we also observe that a smaller error bound of $\beta^1$ brings a smaller error bound for $\beta^k$, with $k \geq 2$. The importance of $\beta^1$ also comes from the fact that one may use it to estimate the choice range of $\rho_j$ ($j \geq 1$), because $\|\widetilde{\varepsilon}\|_\infty$ is unknown in practice. In the implementation of the GEP-MSCRA, we choose $\rho_1$ using this strategy.

### 4.2. Sign consistency

We show that if the smallest nonzero component of $\beta^*$ is not so small, then the GEP-MSCRA can deliver $\beta^l$ satisfying $\mathrm{supp}(\beta^l) = \mathrm{supp}(\beta^*)$ within a finite number of steps. To achieve this goal, we need the oracle least squares solution:

$$\beta^{\mathrm{LS}} \in \arg\min_{\beta\in\mathbb{R}^p}\Big\{\frac{1}{2n}\|\widetilde{Z}\beta-\widetilde{y}\|^2:\ \mathrm{supp}(\beta)\subseteq S^*\Big\}. \tag{4.7}$$

Write $\varepsilon^{\mathrm{LS}} := \frac{1}{n}\widetilde{Z}^{\mathbb{T}}(\widetilde{y}-\widetilde{Z}\beta^{\mathrm{LS}})$. Then, $\varepsilon^{\mathrm{LS}}_{S^*} = \widetilde{Z}^{\mathbb{T}}_{S^*}(\widetilde{Z}\beta^{\mathrm{LS}}-\widetilde{y}) = 0$. This implies that $\beta^{\mathrm{LS}}_{S^*} - \beta^*_{S^*} = \widetilde{\Sigma}^{-1}_{S^*S^*}\big[\frac{1}{n}\widetilde{Z}^{\mathbb{T}}_{S^*}\big(\widetilde{Z}\beta^{\mathrm{LS}}-\widetilde{Z}\beta^*\big)\big] = \widetilde{\Sigma}^{-1}_{S^*S^*}\big[\frac{1}{n}\widetilde{Z}^{\mathbb{T}}_{S^*}(\widetilde{y}-\widetilde{Z}\beta^*)\big]$, and

$$\beta^{\mathrm{LS}}_{S^*} - \beta^*_{S^*} = \widetilde{\Sigma}^{-1}_{S^*S^*}\big(\widehat{\xi}_{S^*} - \widetilde{\Sigma}_{S^*S^*}\beta^*_{S^*}\big) = \widetilde{\Sigma}^{-1}_{S^*S^*}\widetilde{\varepsilon}_{S^*} := \widetilde{\varepsilon}^{\dagger}. \tag{4.8}$$

Based on this observation for $\beta^{\mathrm{LS}}$, we establish the following result.

**Theorem 4.** *Suppose $\Sigma$ satisfies the $\kappa$-REC on $\mathcal{C}(S^*)$, with $\kappa > 24s\|D\|_{\max}$. Set $\gamma := \kappa - 24s\|D\|_{\max}$. If $\lambda, \rho_1$, and $\rho_3$ are chosen such that $\lambda \geq 6\|\varepsilon^{\mathrm{LS}}\|_\infty$,*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



$\rho_1 > \max\big(\frac{4a}{(a+1)\min_{i\in S^*}|\beta_i^*|}, \gamma\lambda^{-1}\|\widetilde{\varepsilon}^{\dagger}\|_\infty\big)$, *and* $\rho_3 \le \sqrt{\frac{4\gamma}{9\sqrt{3}\lambda}}$, *respectively, then for all* $k \in \mathbb{N}$,

$$\big\|\beta^k - \beta^{\mathrm{LS}}\big\| \le \frac{2.03\rho_{k-1}\lambda}{\gamma}\sqrt{|F^{k-1}|} \quad \text{and} \quad \sqrt{|F^k|} \le \frac{18.27\sqrt{3}\rho_k\rho_{k-1}\lambda}{(9\sqrt{3}-4)\gamma}\sqrt{|F^{k-1}|}.$$

*In particular, when* $k \ge \overline{k}$ *with* $\overline{k} = \big\lceil \frac{0.5\ln(s)}{\ln[(9\sqrt{3}-4)\gamma\lambda^{-1}]-\ln[18.27\sqrt{3}(\rho_3)^2]} \big\rceil$, *we have*

$$\beta^k = \beta^{\mathrm{LS}} \quad \text{and} \quad \mathrm{sign}(\beta^k) = \mathrm{sign}(\beta^*).$$

**Remark 2. (a)** Note that Datta and Zou (2017) achieved the sign consistency of $\overline{\beta}$ under an irrepresentable condition on $\Sigma$ and the condition $\min_{i\in S^*}|\beta_i^*| > [4\|\Sigma_{S^*S^*}^{-1}\|_\infty + (\lambda_{\min}(\Sigma_{S^*S^*}))^{-1/2}]\lambda$, where $\|A\|_\infty = \max_i \sum_j |A_{ij}|$ means the matrix $\ell_\infty$-norm. Their irrepresentable condition on $\Sigma$ requires that $\|\Sigma_{(S^*)^cS^*}\Sigma_{S^*S^*}^{-1}\|_\infty \le \overline{\gamma} < 1$ and $\lambda_{\min}(\Sigma_{S^*S^*}) \ge C_{\min}$, for some constants $\overline{\gamma} > 0$ and $C_{\min} > 0$. Here, the former restricts the scale of the entries of $\Sigma$, and the latter is precisely the REC of $\Sigma$ on the set $\big\{\beta \in \mathbb{R}^p \colon \beta_{(S^*)^c} = 0\big\}$. We obtain the sign consistency of $\beta^k$, for $k \ge \overline{k}$, under the $\kappa$-REC of $\Sigma$ on $\mathcal{C}(S^*)$, with $\kappa > 24s\|D\|_{\max}$ and $\rho_1 > \max\big(\frac{4a}{(a+1)\min_{i\in S^*}|\beta_i^*|}, \gamma\lambda^{-1}\|\widetilde{\varepsilon}^{\dagger}\|_\infty\big)$. When $\lambda_{\min}(\Sigma_{S^*S^*})$ is large, there is high probability that our $\kappa$-REC holds. In addition, when $\|\Sigma_{(S^*)^cS^*}\Sigma_{S^*S^*}^{-1}\|_\infty \le \overline{\gamma}$ does not hold, our $\kappa$-REC may hold; for example, consider $\Sigma = [1\ 0\ 2; 0\ 1\ 2; 2\ 2\ 9]$ and $S^* = \{1, 2\}$. In fact, to some extent, our $\kappa$-REC also depends on the unbiased surrogate $\widehat{\Sigma}$ of $\Sigma$. If $\|\widehat{\Sigma} - \Sigma\|_{\max}$ is small, there is a high probability that our $\kappa$-REC holds.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



Finally, the condition on $\min_{i\in S^*}|\beta_i^*|$ used by Datta and Zou (2017) implies a large choice range for our parameter $\rho_1$ regardless of whether $\|\Sigma_{S^*S^*}^{-1}\|_\infty$ or $(\lambda_{\min}(\Sigma_{S^*S^*}))^{-1/2}$ is larger or $\lambda$ is larger.

**(b)** Note that $\rho_3 \leq \sqrt{\frac{4\gamma}{9\sqrt{3}\lambda}}$. Together with the definition of $\overline{k}$, we have $\ln[(9\sqrt{3}-4)\gamma\lambda^{-1}] - \ln[18.27\sqrt{3}(\rho_3)^2] \geq \ln(1.4)$, which, with $s \geq 9$, implies that $\overline{k} \leq \widehat{k} := \lceil \frac{0.5\ln(s)}{\ln(1.4)} \rceil$. As one referee pointed out, $\overline{k}$ or $\widehat{k}$ is actually unknown because it depends on the sparsity $s$ of $\beta^*$. In practice, some prior upper estimation on $s$ is usually available; for example, a rough upper estimation on $s$ is the dimension $p$. Thus, one still can obtain a rough upper estimation on $\widehat{k}$. In the practical numerical computation, we identify $\overline{k}$ by monitoring the index change of the nonzero entries in each iteration.

**(c)** By Theorem 4, the choice of $\rho_1$ is crucial for the GEP-MSCRA to yield an oracle solution with a sign that is consistent with that of $\beta^*$ after a finite number of steps. As remarked after Theorem 3, the ease of choosing $\rho_1$ depends on the error bound of $\beta^1$. From Theorem 4 and Theorem 3, we conclude that a smaller $\rho_3$ results in good output from the GEP-MSCRA in terms of the error bound and sign consistency. Furthermore, for those problems with high noise, a large $\lambda$ is needed and, of course, the error bound of $\beta^k$ becomes large.

We have established the deterministic theoretical guarantees of the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



GEP-MSCRA when computing the CaZnRLS estimator under suitable conditions. From (Raskutti, Wainwright, and Yu, 2010, 2011), if $X$ is from the $\Sigma_x$-Gaussian ensemble $\big($i.e., $X$ is formed by independently sampling each row $X^i \sim N(0, \Sigma_x)\big)$, then there exists a constant $\kappa > 0$ (depending on $\Sigma_x$), such that $\Sigma$ satisfies the REC on $\mathcal{C}(S^*)$ with probability greater than $1-c_1 \exp(-c_2 n)$, as long as $n > cs \ln p$, where $c, c_1$, and $c_2$ are absolutely positive constants. It is natural to ask whether such $\kappa$ satisfies the requirements of the above theorems. What is the likelihood of choosing $\lambda, \rho_1$, and $\rho_3$ as required in the above theorems? In the Supplementary Material, we focus on these questions for two specific types of errors-in-variables models.

## 5. Numerical experiments

We use simulated data sets to evaluate the performance of the CaZnRLS estimator, computed using the GEP-MSCRA (see the Supplementary Material for the implementation details). Then, we compare its performance with that of CoCoLasso and NCL in terms of the number of signs identified correctly (NC) and incorrectly (NIC) for the predictors, and in terms of the relative root-mean-square error (RMSE). Let $\beta^f$ be the final output of one

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



of three solvers. Define

$$\mathrm{NC}(\beta^f) := \sum_{i\in S^*} \mathbb{I}\big\{|\mathrm{sign}(\beta_i^f) - \mathrm{sign}(\beta_i^*)| = 0\big\},$$

$$\mathrm{NIC}(\beta^f) := N_{\mathrm{nz}}(\beta^f) - \mathrm{NC}(\beta^f), \quad \text{and} \quad \text{relative RMSE} := \frac{\|\beta^f - \beta^*\|}{\|\beta^*\|},$$

where $N_{\mathrm{nz}}(\beta^f) := \sum_{i=1}^p \mathbb{I}\big\{|\beta_i| > 10^{-8}\big\}$ is the number of nonzero entries of $\beta^f$. All results are obtained using a desktop computer running on 64-bit Windows with an Intel(R) Core(TM) i7-7700 CPU 3.6GHz and 16 GB memory.

For the GEP-MSCRA, we choose $a = 6.0$ for $\phi$, $w^0 = 0$, and $\rho_k$ for $k \le 3$ as

$$\rho_1 = \max\Big(1, \frac{5}{3\|\beta^1\|_\infty}\Big), \ \rho_k = \min\Big(2\rho_{k-1}, \frac{10^8}{\|\beta^k\|_\infty}\Big) \text{ for } k = 2, 3.$$

We terminate GEP-MSCRA at $\beta^k$ once the following condition is satisfied:

$$\begin{cases} |N_{\mathrm{nz}}(\beta^{k-j}) - N_{\mathrm{nz}}(\beta^{k-j-1})| \le 5, \ j = 0, 1, 2; \\ \big|\frac{1}{2n}\|\widetilde{Z}\beta^k - \widetilde{y}\|^2 - \frac{1}{2n}\|\widetilde{Z}\beta^{k-1} - \widetilde{y}\|^2\big| \le 0.1, \end{cases}$$

or the number of iterations reaches the maximum number $k_{\max} = 4$ (Our code is available from `https://github.com/SCUT-OptGroup/ErrorInvar`). This stopping criterion captures a solution $\beta^k$ with a sparsity that tends to be stable, and a predictor error that has a small variation. In addition, from Remark 2(b), we have a rough upper estimation for $\overline{k}$ as $\lceil \frac{0.5\ln(p)}{\ln(1.4)} \rceil$,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



which is equal to 11 for $p = 1000$. As such, we set the maximum number of iterations to four. We solve the dual of (3.1) using Algorithm 2 for $\epsilon^j = 10^{-8}$. For the NCL, we run the code "doProjGrad," solving the model in (1.4) with $\lambda_n = 0$ and $R_0 = \|\beta^*\|_1$, for the test examples. Because the Matlab code for CoCoLasso is not available, we include our implementation in the Supplementary Material. It is time-consuming for Algorithm 4 to use the stopping rule $\max\{\epsilon^k_{\text{pinf}}, \epsilon^k_{\text{dinf}}, \epsilon^k_{\text{gap}}\} \leq 10^{-5}$. Therefore, we use the looser $\max\{\epsilon^k_{\text{pinf}}, \epsilon^k_{\text{dinf}}, 10^{-3}\epsilon^k_{\text{gap}}\} \leq 10^{-4}$ to obtain an approximate solution for (1.5), and then use Algorithm 2 to solve the associated problem (1.6).

From the theoretical results in Section 4, the appropriate $\lambda$ lies in an interval associated with $\|\widetilde{\varepsilon}\|_\infty$. Such a $\lambda$ is also suitable for the CoCoLasso, by the proof of Theorem 1 and 2 in Datta and Zou (2017). As such, we set $\lambda = \max(0.01, \frac{\alpha^*}{n}\|\widetilde{Z}^{\mathbb{T}}\widetilde{y}\|_\infty)$ and $\max(0.01, \frac{\alpha^*}{n}\|\overline{Z}^{\mathbb{T}}\overline{y}\|_\infty)$ for the CaZnRLS and CoCoLasso, respectively, where the appropriate $\alpha^* \in [0.06, 0.32]$ is chosen using the five-fold corrected cross-validation proposed by Datta and Zou (2017).

Throughout this section, all test examples are generated randomly as the triple $(p, s, n)$, consisting of the dimension $p$ of the predicted variable, number of nonzero entries of $\beta^*$, and sample size $n$. Among others, $n = \lfloor \alpha s \ln(p) \rfloor$, with $\alpha = 4{+}0.2(j{-}1)$, for $j = 1, \ldots, 11$. We obtain observation $y$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



from the model (1.1), where the entries of $\varepsilon$ are independent and identically distributed (i.i.d.) $\mathcal{N}(0, \sigma^2)$. We describe how to generate the true $\beta^*_{S^*}$ in the following. The average relative RMSE (respectively, NC and NIC) is the average of the total RMSE (respectively, NC and NIC) for 100 problems, generated randomly.

### 5.1. Random locations of the nonzero entries of $\beta^*$

In this section, we evaluate the performance of the CaZnRLS using randomly generated examples, where $\beta^*_{S^*}$ is an i.i.d. standard normal random vector, with the $s = \lfloor 0.5\sqrt{p} \rfloor$ entries of $S^*$ chosen randomly from $\{1, \ldots, p\}$. First, we test whether the CaZnRLS is stable with respect to the variance $\sigma$ of $\varepsilon$.

**Example 1.** We generate $Z = X + A$ with $p = 500$, where the rows of $X$ are i.i.d. standard normal random vectors with mean zero and covariance matrix $\Sigma_X = I$. The rows of $A$ are i.i.d. $\mathcal{N}(0, I)$.

Figure 2 plots the average relative RMSE, NC, and NIC curves of the CaZnRLS, CoCoLasso, and NCL for Example 1 under different sample sizes, with $\sigma = 0.5$ and 1.0. The subfigures in the first column show that the CaZnRLS is comparable to, or even a little better than the CoCoLasso in terms of the relative RMSE. The second column shows that the NC of the CaZnRLS is at most two fewer than that of the CoCoLasso. Finally, the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



third column indicates that the NIC of the CaZnRLS is much lower than that of the CoCoLasso. From this, we conclude that the CaZnRLS maintains the advantages of the zero-norm regularized LS estimator in the clean data setting. We also see that the CaZnRLS exhibits similar performance for $\sigma = 0.5$ and $\sigma = 1$, indicating that it is insensitive to the variance $\sigma$ of the regression error. Therefore, in the following, we always take $\sigma = 0.5$.

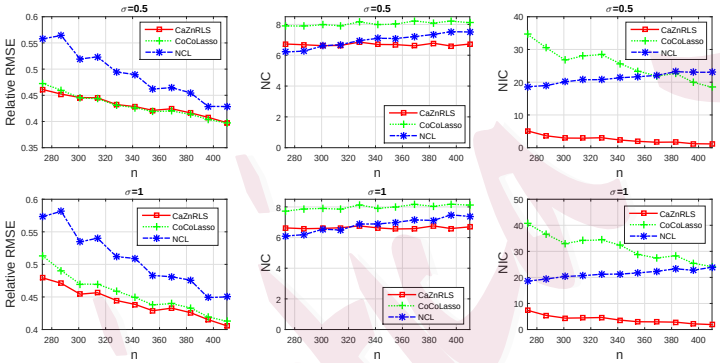


Figure 2: The relative RMSE, NC, and NIC of three solvers under different $\sigma$ for Example 1

Next, we evaluate the performance of the CaZnRLS for three classes of measurement errors using test problems generated with $p = 1000$.

**Case 1. Additive errors**

**Example 2.** We generate $Z = X + A$, where $X$ is defined as in Example 1, and the rows of $A$ are i.i.d. $\mathcal{N}(0, \tau^2 I)$, with $\tau = 0.5$ or 1.0.

**Example 3.** We generate $Z = X + A$, where the entries of $X$ are i.i.d. and follow the uniform distribution on $(0, 1)$, and $A$ is defined as in Example 2.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



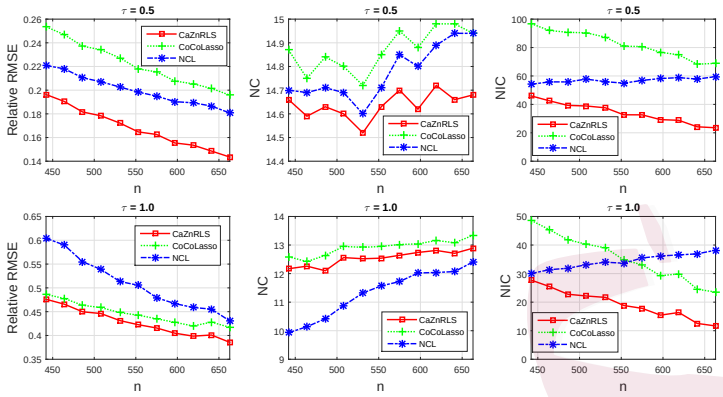


Figure 3: The relative RMSE, NC, and NIC of three solvers under different $n$ for Example 2

Figure 3 plots the average relative RMSE, NC, and NIC curves of three solvers under different sample sizes for Example 2. From this figure, whether $X$ is corrupted by high noise or low noise, the CaZnRLS is the best of the three solvers in terms of the relative RMSE and NIC, though its NC is (at most one) fewer than the NC of the CoCoLasso. The relative RMSE of the CaZnRLS improves on that of the CoCoLasso by at least 20% for the low noise, and by 4% for the high noise when $n \geq \lfloor 5s\ln(p) \rfloor$. We also see that the NCL performs worst in terms of the relative RMSE, NC, and NIC for the high noise.

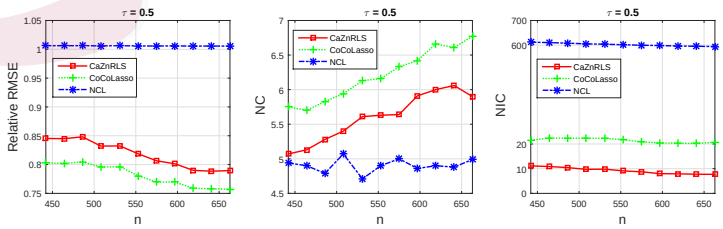


Figure 4: The relative RMSE, NC, and NIC of three solvers under different $n$ for Example 3

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



Figure 4 plots the average relative RMSE, NC, and NIC curves of three solvers under different sample sizes for Example 3. We see that the three solvers have much higher relative RMSEs than those in Example 2. Furthermore, the NCL fails to give the desired estimator. The relative RMSE of the CaZnRLS is a little (about 4%) higher than that of the CoCoLasso. After checking the unbiased estimation $\Sigma$ of the covariance matrix of the true covariates, we find that the irrepresentable and minimum eigenvalue conditions in Datta and Zou (2017) are not satisfied. Now, it is not clear whether our REC on $\mathcal{C}(\beta^*)$ holds. This does not contradict the theoretical analysis in Section 4, because we know only that our REC on $\mathcal{C}(\beta^*)$ holds w.h.p. when $X$ is from the Gaussian ensemble. The first subfigure indicates that it is very likely that our REC does not hold when $X$ is from the uniform distribution.

**Case 2. Multiplicative errors**

**Example 4.** We generate $Z = X \circ M$, where the rows of $X$ are i.i.d. $\mathcal{N}(0, I)$, and the entries of $M$ are i.i.d. and follow the log-normal distribution; that is, $\ln(M_{ij})$ are i.i.d and follow $N(0, \tau^2 I)$, with $\tau = 0.5$ or 0.8.

**Example 5.** We generate $Z$ in the same way as in Example 4, except that the entries of $X$ are i.i.d. and follow the Laplace distribution with mean zero and variance one.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



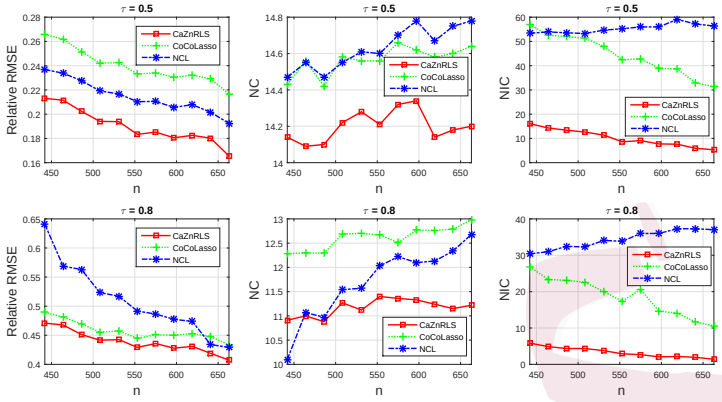


Figure 5: The relative RMSE, NC, and NIC of three solvers under different $n$ for Example 4

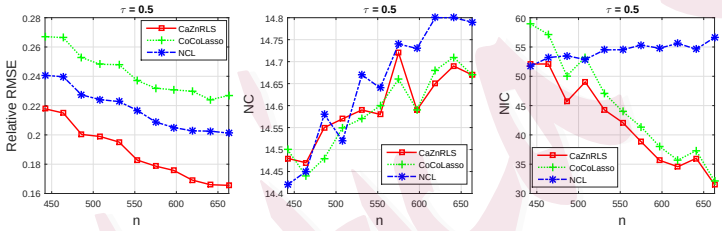


Figure 6: The relative RMSE, NC, and NIC of three solvers under different $n$ for Example 5

Figures 5 and 6 plot the average relative RMSE, NC, and NIC curves of three solvers under different $n$ for Examples 4 and 5, respectively. Comparing Figures 5 with 3, we see that the CaZnRLS and CoCoLasso perform similarly, as they do for the additive errors. That is, the CaZnRLS outperforms the CoCoLasso in terms of the relative RMSE and NIC, whether for $X$ corrupted by high noise or low noise, although its NC is (at most two) fewer than the NC of the CoCoLasso. This, together with Figure 6, leads us to conclude that the CaZnRLS performs similarly when the rows of $X$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



follow the Gaussian and Laplace distributions.

**Case 3. Missing data case**

**Example 6.** We generate $(Z_{ij})_{n\times p}$ for $Z_{ij} = X_{ij}$ with probability $1-\tau$ and, $Z_{ij} = 0$ with probability $\tau$, for $\tau = 0.3$ or $0.5$, where the rows of $X$ are i.i.d. and follow the standard normal distribution $\mathcal{N}(0, I)$.

**Example 7.** We generate $Z$ in the same way as in Example 6 except that $X_{ij}$ are i.i.d. and follow the exponential distribution with mean one and variance one.

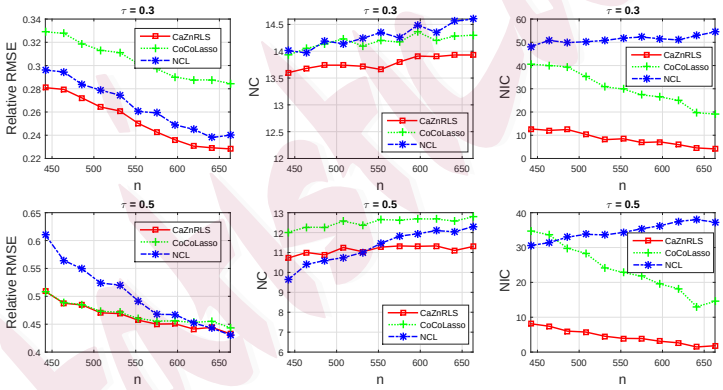


Figure 7: The relative RMSE, NC, and NIC of three solvers under different $n$ for Example 6

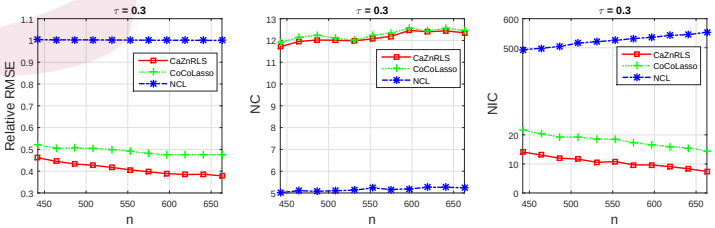


Figure 8: The relative RMSE, NC, and NIC of three solvers under different $n$ for Example 7

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



Figures 7 and 8 plot the average relative RMSE, NC, and NIC curves of three solvers under different $n$ for Examples 6 and 7, respectively. Comparing Figure 7 with Figure 3 or 5, we see that the three solvers perform similarly to the cases of additive and multiplicative errors. In fact, similarly to Example 2, 4, and 5, Example 6 satisfies the irrepresentable and minimum eigenvalue conditions in Datta and Zou (2017) when $n \geq \lfloor 4.4s \ln(p) \rfloor$. Of course, our REC on $\mathcal{C}(\beta^*)$ holds with a high probability for Examples 2 and 4, and Figures 6–8 indicate that our REC holds with a high probability when the rows of $X$ follow the Laplace and exponential distributions. Figure 8 shows that, when the entries of $X$ follow the exponential distribution, the CaZnRLS is superior to the other two solvers in terms of the relative RMSE and NIC, and its RMSE improves on that of the CoCoLasso by at least 11%. Now, the NCL fails to yield the desired estimator. After checking, we find that Example 7 does not satisfy the irrepresentable and minimum eigenvalue conditions in Datta and Zou (2017). Now, it is not clear whether our REC holds for this example.

Motivated by one referee's comments, we next provide an example that does not satisfy the irrepresentable condition, but in which our REC holds w.h.p.

**Example 8.** We generate $Z = X + A$, with $p = 250$, where the entries of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



$X_{S^*}$ are i.i.d. $\mathcal{N}(0,1)$, the entries of $X_{(S^*)^c}$ are i.i.d. $\mathcal{N}(0,5^2)$, and the rows of $A$ are generated in the same way as in Example 2, with $\tau = 0.75$.

Figure 9 plots the average relative RMSE, NC, and NIC curves of the CaZnRLS and CoCoLasso under different $n$ for Example 8. Because the NCL fails in this example, we do not include its results in Figure 9. We see that the relative RMSE of the CaZnRLS is lower than that of the CoCoLasso, and when $n \geq \lfloor 5s\ln(p) \rfloor$, the relative RMSE of the CaZnRLS improves on that of the CoCoLasso by at least 10%. The NC and NIC of the CoCoLasso are still higher than those of the CaZnRLS, but the NC of the latter is at most one fewer than that of the former. This example further confirms the theoretical results in Section 4.

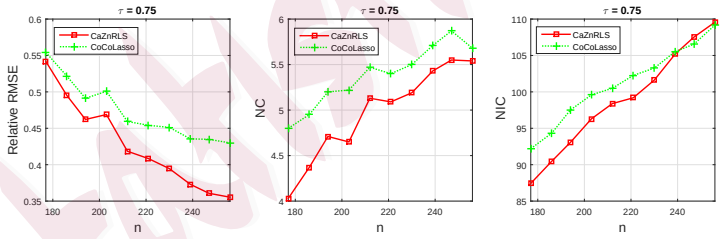


Figure 9: The relative RMSE, NC, and NIC of three solvers under different $n$ for Example 8

## 5.2. Fixed locations of the nonzero entries of $\beta^*$

As one referee pointed out, it would be interesting to show the effects of the correlation between the predictors on the performance of the three solvers. In this section, we test whether this correlation affects the per-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



formance of the three solvers using the examples generated by Datta and Zou (2017), in which the locations of the nonzero entries of $\beta^*$ are fixed. Specifically, $\beta^* = (3, 1.5, 0, 0, 2, 0, \ldots, 0)$, with the number of nonzero entries $s = 3$. The data $X$ are generated using $p = 250$ and $n = 100$, such that the rows of $X$ obey i.i.d. $\mathcal{N}(0, \Sigma_X)$, for $(\Sigma_X)_{ij} = 0.5^{|i-j|}$. Table 1 summarizes the simulation results of the three solvers for the additive errors, multiplicative errors, and missing data, where the error matrices $A$ and $M$ for the additive and multiplicative errors, respectively, are generated in the same way as in Example 2 and 4, respectively. The contaminated matrix $Z$ in the missing data is generated in the same way as in Example 6.

Table 1: The average relative RMSE, NC, and NIC of the three solvers

| | **Additive errors** $\tau = 1$ | | | **Multiplicative errors** $\tau = 0.8$ | | | **Missing data** $\tau = 0.5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | CaZnRLS | CoCoLasso | NCL | CaZnRLS | CoCoLasso | NCL | CaZnRLS | CoCoLasso | NCL |
| RMSE | 0.410 | 0.492 | 0.535 | 0.370 | 0.524 | 0.600 | 0.447 | 0.521 | 0.528 |
| NC | 2.81 | 2.87 | 2.41 | 2.76 | 2.87 | 2.18 | 2.69 | 2.75 | 2.27 |
| NIC | 1.48 | 2.46 | 6.48 | 1.30 | 2.48 | 5.31 | 2.41 | 2.60 | 6.90 |

From Table 1, the CaZnRLS yields the lowest relative RMSE and NIC for the three classes of measurement errors, although its NC is a little lower than that of the CoCoLasso. The NCL yields the highest relative RMSE and NIC. Comparing the numerical results in Section 5.1, we find that the three solvers perform similarly to those examples in which the locations of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



the nonzero entries of $\beta^*$ are not fixed. That is, the correlation between the predictors has little effect on their performance.

The numerical comparisons in the previous two subsections show that when the true covariate matrix $X$ follows the standard normal distribution (i.e., our REC holds with a high probability), or other distributions, such as the Laplace distribution in Example 5 and the exponential distribution in Example 7, the CaZnRLS outperforms the CoCoLasso in terms of the relative RMSE (especially for low noise cases) and NIC. However, its NC is a little lower than that of the CoCoLasso. As shown in Figure 10, the CaZnRLS requires much less computing time.

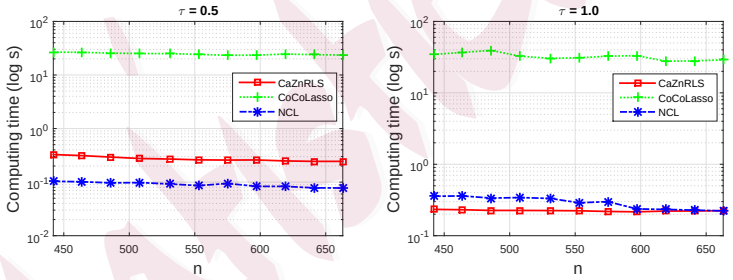


Figure 10: The computing time for the three solvers in Example 2

## Supplementary Material

The online Supplementary Material includes the algorithms, auxiliary lemmas, and proofs of the lemmas, theorems, and corollaries.

## Acknowledgments

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



The authors thank the anonymous referees for their valuable suggestions and comments on the original manuscript. The authors are indebted to Professor Po-Ling Loh for sharing the R and Matlab code used to compute the NCL estimator. The research of Shaohua Pan and Shujun Bi was supported by the National Natural Science Foundation of China under project Nos. 11571120 and 11701186, and by the Natural Science Foundation of Guangdong Province under project No. 2017A030310418.

## References

Belloni, A., Rosenbaum, M. and Tsybakov, A. B. (2017). Linear and conic programming estimators in high-dimensional errors-in-variables models. *Journal of the Royal Statistical Society, Series B 79*, pp. 939–956.

Belloni, A., Rosenbaum, M. and Tsybakov, A. B. (2016). An $\ell_1$, $\ell_2$, $\ell_\infty$-regularization approach to high-dimensional errors-in-variables models. *Electronic Journal of Statistis 10*, pp. 1729–1750.

Bi, S. J. and Pan, S. H. (2018). GEP-MSCRA for the group zero-norm regularized least squares estimator. *arXiv:1804.09887v1.*

Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



content bias in high-throughput sequencing. *Nucleic Acids Research 40*, pp. e72–e72.

Bühlmann, P. and van de Geer, S. A. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.

Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis*. New York: John Wiley and Sons.

Chen, Y. and Caramanis, C. (2013). Noisy and missing data regression: distribution-oblivious support recovery. *Journal of Machine Learning Research 28*, pp. 383–391.

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when $p$ is much larger than $n$. *The Annals of Statistics 35*, pp. 2313–2351.

Datta, A. and Zou, H. (2017). *CoCoLASSO for high-dimensional error-in-variables regression. The Annals of Statistics 45*, pp. 2400–2426.

Duchi, J., Shalev-Shwartz, S., Singer, Y. and Chandra T. (2008). Efficient projections onto the $L_1$-ball for learning in high-dimensions. *In Proceedings of the 25th International Conference on Machine Learning*, pp. 272–279.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistics Association 96*, pp. 1348–1360.

Fazel, M., Pong, T. K., Sun, D. F. and Tseng, P. (2013). Hankel matrix rank minimization with applications in system identification and realization. *SIAM Journal on Matrix Analysis and Applications 34*, pp. 946–977.

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis* (2 ed.). New York: Cambridge University Press.

Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. New York: Cambridge University Press.

Loh, P. L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics 40*, pp. 1637–1664.

Loh, P. L. (2014). High-dimensional statistics with systematically corrupted data. *University of California, PhD thesis, http://escholarship.org/uc/item/8j49c5n4*.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



Li, X. D., Sun, D. F. and Toh, K.-C. (2018). A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. *SIAM Journal on Optimization 28*, pp. 433–458.

Liu, J., Ji, S. W. and Ye, J. P. (2011). SLEP: Sparse Learning with Efficient Projections. *Arizona State University. URL: http://www.public.asu.edu/jye02/Software/SLEP*.

Mifflin, R. (1977). Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization 15*, pp. 959–972.

Nesterov, Y. (2013). Gradient methods for minimizing composite objective function. *Mathematical Programming 140*, pp. 125–161.

Negahban, S., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science 27*, pp. 538–557.

Purdom, E. and Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology 4*, Artical 16.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



Qi, L. and Sun, J. (1993). A nonsmooth version of Newton's method. *Mathematical Programming 58*, pp. 353–367.

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research 11*, pp. 2241–2259.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over $\ell_1$-balls. *IEEE Transactions on Information Theory 57*, pp. 6976–6994.

Rosenbaum, M. and Tsybakov, A. B. (2010). Sparse recovery under matrix uncertainty. *Annals of Statistics 38*, pp. 2620–2651.

Rosenbaum, M. and Tsybakov, A. B. (2013). Improved matrix uncertainty selector. *Institute of Mathematical Statistics Collections 9*, pp. 276–290.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.

Slijepcevic, S., Megerian, S. and Potkonjak, M. (2002). Location errors in wireless embedded sensor networks: sources, models, and effects on applications. *Mobile Computing and Communications Review 6*, pp. 67–78.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



Städler, N. and Bühlmann, P. (2012). Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing 22*, pp. 219–235.

Sun, D. F., Yang, L. Q. and Toh, K.-C. (2016). An efficient inexact ABCD method for least squares semidefinite programming. *SIAM Journal on Optimization 26*, pp. 1072–1100.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B 58*, pp. 267–288.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics 3*, pp. 1360–1392.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*, pp. 894–942.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B 67*, pp. 301–320.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, pp. 1418–1429.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0507



Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*, pp. 1509–1533.

School of Mathematics, South China University of Technology

E-mail: (*201620122022@mail.scut.edu.cn)*

*School of Mathematics, South China University of Technology*

*E-mail: (shhpan@scut.edu.cn)*

*School of Mathematics, South China University of Technology*

*E-mail: (bishj@scut.edu.cn)*

*Phone: +86 13724034152*

*Fax: +86 020 87110448*