PENALIZED LINEAR UNBIASED SELECTION

Cun-Hui Zhang[1]

Rutgers University, Department of Statistics and Biostatistics

Technical Report #2007-003

April 20, 2007

We introduce MC+, a fast, continuous, nearly unbiased, and accurate method of penalized variable selection in high-dimensional linear regression. The LASSO is fast and continuous, but biased. The bias of the LASSO interferes with variable selection. Subset selection is unbiased but computationally costly. The MC+ has two elements: a minimax concave penalty (MCP) and a penalized linear unbiased selection (PLUS) algorithm. The MCP provides the minimum non-convexity of the penalized loss given the level of bias. The PLUS computes multiple local minimizers of a possibly non-convex penalized loss function in certain main branch of the graph of such solutions. Its output is a continuous piecewise linear path encompassing from the origin to an optimal solution for zero penalty. We prove that for a universal penalty level, the MC+ has high probability of correct selection under much weaker conditions compared with existing results for the LASSO for large n and p, including the case of $p \gg n$. We provide estimates of the noise level for proper choice of the penalty level. We choose the sparsest solution within the PLUS path for a given penalty level. We derive degrees of freedom and C_p-type risk estimates for general penalized LSE, including the LASSO estimator, and prove their unbiasedness. We provide necessary and sufficient conditions for the continuity of the penalized LSE under general sub-square penalties. Simulation results overwhelmingly support our claim of superior variable selection properties and demonstrate the computational efficiency of the proposed method.

Key words: Variable selection, model selection, penalized estimation, least squares, correct selection, unbiasedness, non-convex minimization, risk estimation, degrees of freedom, selection consistency.

[1] Research partially supported by NSF Grants DMS 0504387 and DMS 0604571 and NSA Grant MDS 904-02-1-0063.

Address: Department of Statistics and Biostatistics, Busch campus, Rutgers University, Piscataway, New Jersey 08845, U.S.A.
E-mail: czhang@stat.rutgers.edu

1. Introduction. Variable selection is fundamental in statistical analysis of high-dimensional data. With a proper selection method and under suitable conditions, we are able to build consistent models which are easy to interpret, to avoid over fitting in prediction and estimation, and to identify relevant variables for applications or further study. Consider a linear model in which a response vector $\boldsymbol{y} \in \mathbb{R}^n$ depends on p predictors $\boldsymbol{x}_j \in \mathbb{R}^n$, $j = 1, \ldots, p$, through a linear combination $\sum_{j=1}^p \beta_j \boldsymbol{x}_j$. For small p, subset selection methods can be used to find a good guess of the pattern

$$A^o \equiv \big\{j : \beta_j \neq 0\big\}. \tag{1.1}$$

For example, one may impose a proper penalty on the number of selected variables based on the AIC (Akaike, 1973), C_p (Mallows, 1973), BIC (Schwarz, 1978), RIC (Foster and George, 1994) or a data driven method. For large p, subset selection is not computationally feasible, so that continuous penalized or gradient threshold methods are typically used.

Let $\|\cdot\|$ be the Euclidean norm. Consider penalized least squares estimators (LSE)

$$\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(\lambda) \equiv \arg\min_{\boldsymbol{b}} \left\{ \frac{1}{2n} \big\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\big\|^2 + \sum_{j=1}^p \rho\big(|b_j|; \lambda\big) \right\}, \tag{1.2}$$

with a penalty $\rho(t; \lambda)$ indexed by $\lambda \geq 0$, in the linear regression model

$$\boldsymbol{y} = \sum_{j=1}^p \beta_j \boldsymbol{x}_j + \boldsymbol{\varepsilon}, \tag{1.3}$$

where $\boldsymbol{X} \equiv (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)'$, and $\boldsymbol{\varepsilon}$ is an error vector. Assume the penalty $\rho(t; \lambda)$ is nondecreasing in t and has a continuous derivative $\dot\rho(t; \lambda) = (\partial/\partial t)\rho(t; \lambda)$ in $(0, \infty)$. Assume further $\dot\rho(0+; \lambda) > 0$, so that (1.2) has variable selection features with the possibility of $\widehat{\beta}_j = 0$ (Donoho, Johnstone, Hoch and Stern, 1992). Changing the index λ if necessary, we assume $\dot\rho(0+; \lambda) = \lambda$ whenever $\dot\rho(0+; \lambda) < \infty$, so that λ has the interpretation as the threshold level for the individual regression coefficients β_j under the standardization $\|\boldsymbol{x}_j\|^2/n = 1$. In what follows, we treat the set of variables selected by $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(\lambda)$ as

$$\widehat{A} \equiv \widehat{A}(\lambda) \equiv \Big\{j : \widehat{\beta}_j \neq 0\Big\}. \tag{1.4}$$

A widely used procedure of form (1.2) is the LASSO (Tibshirani, 1996) with $\rho(t; \lambda) = \lambda|t|$, which is easy to compute (Osborne, Presnell and Turlach, 2000a, 2000b; Efron, Hastie, Johnstone and Tibshirani, 2004) and has the interpretation as boosting (Schapire, 1990; Freund and Schapire, 1996; Friedman, Hastie and Tibshirani, 2000). Meinshausen and Buhlmann (2006) and Zhao and Yu (2006) showed that the LASSO is variable selection consistent

$$P\Big\{\widehat{A} = A^o\Big\} \to 1 \tag{1.5}$$

under a strong irrepresentable condition on the covariance matrix $\boldsymbol{X}'\boldsymbol{X}$ and some additional regularity conditions on $\{n, p, \boldsymbol{\beta}, \boldsymbol{\varepsilon}\}$. However, the strong irrepresentable condition is quite restrictive, and that due to the estimation bias, the condition is also necessary for LASSO to be selection consistent. Under a relatively mild sparse Riesz condition on the covariance matrix $\boldsymbol{X}'\boldsymbol{X}$, Zhang and Huang (2006) proved that the dimension $|\widehat{A}|$ for the LASSO selection is of the same order as the size

$$d^o \equiv |A^o| = \#\{j : \beta_j \neq 0\} \tag{1.6}$$

of the unknown pattern (1.1) and that the LASSO selects all variables with absolute coefficients above certain separation zone of the order $\sqrt{d^o}\lambda$ under the standardization $\|\boldsymbol{x}_j\|^2/n = 1$. These results are still not satisfactory in view of the possibility of selecting some irrelevant variables and the extra factor $\sqrt{d^o}$ for the separation zone, compared with the intended threshold level λ. Again, due to the estimation bias of the LASSO, the extra factor $\sqrt{d^o}$ cannot be removed under either the sparse Riesz or strong irrepresentable conditions. From these points of view, the bias of the LASSO severely interferes with variable selection when p and d^o are both large.

Prior to the above mentioned studies about the interference of the bias of the LASSO with accurate variable selection, Fan and Li (2001) raised the concern of the effect of the bias of more general penalized estimators on estimation efficiency. They pointed out that the bias of penalized estimators can be removed almost completely by choosing a constant penalty beyond a second threshold level $\gamma\lambda$, and carefully developed the SCAD method (Fan, 1997) for $p > 1$ with the penalty $\lambda\int_0^t \min\big\{1, (\gamma - x/\lambda)^+/(\gamma - 1)\big\}dx$, $\gamma > 2$. Iterative algorithms were developed there and in Hunter and Li (2005) and Zou and Li (2006) to approximate a local minimizer of the SCAD penalized loss for fixed (λ, γ). For penalized methods with unbiasedness and selection features, Fan and Peng (2004) proved the existence, variable selection consistency (1.5) and asymptotic estimation efficiency of some local minimizer of the penalized loss under the dimensionality constraint $p = o(n^r)$ with $r = 1/3, 1/4$ or $1/5$ depending on regularity conditions. Their results apply to general classes of loss and penalty functions but do not address the uniqueness of the solution or provide methodologies for finding the local minimizer with the stated properties. A major cause of computational and analytical difficulties in these studies of unbiased selection methods is the non-convexity of the minimization problem.

The main purpose of this paper is to introduce and study an MC+ methodology which has two components: a *minimax concave penalty* (MCP) and a *penalized linear unbiased selection* (PLUS) algorithm.

The MCP, defined as

$$\rho(t;\lambda) = \lambda \int_0^t \Big(1 - \frac{x}{\gamma\lambda}\Big)^+ dx \tag{1.7}$$

with a regularization parameter $\gamma > 0$, is the minimizer of the maximum concavity

$$\kappa(\rho;\lambda) \equiv \sup_{0<t_1<t_2} \frac{\dot\rho(t_1;\lambda) - \dot\rho(t_2;\lambda)}{t_2 - t_1}, \tag{1.8}$$

subject to the following unbiasedness and selection features:

$$\dot\rho(t;\lambda) = 0 \ \forall \ t \geq \gamma\lambda, \quad \dot\rho(0+;\lambda) = \lambda. \tag{1.9}$$

The PLUS computes a piecewise linear path of critical points for the possibly non-convex minimization problem (1.2). It differs from existing non-convex minimization algorithms in three important aspects: (i) It computes the exact value of local minimizers instead of iteratively approximating them; (ii) It computes a path of possibly multiple local minimizers for the entire range of the penalty level $\lambda \geq 0$ instead of a single solution for a fixed λ; (iii) It computes multiple local minimizers for individual penalty levels λ by tracking along its path of critical points for different values of λ instead of trying to jump from the domain of attractions of one solution to another for a fixed λ. In each step, the PLUS computes one line segment in its path between two turning points, and its computational cost is the same as the LARS (Efron et al 2004) per step. The MC+ with larger regularization parameter γ provides smoother predictors and computationally less complex path, but larger bias and less accurate variable selection. The MC+ path converges to the LASSO path as $\gamma \to \infty$.

The proposed MC+ provides fast, continuous, nearly unbiased, and accurate variable selection in high-dimensional linear regression, as our theoretical and numerical results support. Table 1 presents the results of a simulation experiment to demonstrate the superior selection accuracy and competitive computational complexity of the MC+, compared with the LASSO and SCAD. We measure the selection accuracy by the proportion $\overline{CS}$ of replications with the correct selection $CS \equiv I\{\widehat{A} = A^o\}$ and the computational complexity by the average $\overline{k}$ of the

Table 1: Performance of LASSO, MC+ and SCAD in Experiment 1

100 replications, $n = 300$, $p = 200$, $\beta_* = 1/2$, $\gamma = 2/(1 - \max_{j \neq k} |\boldsymbol{x}_j' \boldsymbol{x}_k|/n) = 2.652$

CS $\equiv I\{\widehat{A} = A^o\}$, $k \equiv \#$(steps); Nearly identical results for known σ

$\overline{CS} \leq 0.14$ for $\lambda/\widehat{\sigma} = \sqrt{(\log p)/n}$; $\overline{CS} \leq 0.01$ for $\lambda/\widehat{\sigma} = 1.96/\sqrt{n}$ or $\sqrt{16(\log p)/n}$

		$d^o = 10$			$d^o = 20$			$d^o = 40$		
$\lambda/\widehat{\sigma}$		lasso	mc+	scad	lasso	mc+	scad	lasso	mc+	scad
$\sqrt{2(\log p)/n}$	$\overline{CS}$	0.34	**0.76**	**0.70**	0.06	**0.78**	**0.61**	0.01	**0.84**	0.24
$= 0.1879$	$\overline{k}$	12	16	26	23	32	51	48	65	132
$\sqrt{4(\log p)/n}$	$\overline{CS}$	**0.88**	**0.97**	**0.93**	0.41	**0.81**	0.49	0.01	0.11	0.00
$= 0.2658$	$\overline{k}$	11	11	14	21	21	27	42	41	57
$\sqrt{8(\log p)/n}$	$\overline{CS}$	0.39	0.40	0.39	0.07	0.08	0.07	0.00	0.00	0.00
$= 0.3759$	$\overline{k}$	10	10	10	17	17	17	31	28	32

number of the PLUS steps. In this experiment, $\boldsymbol{y}$ is generated with $\beta_j = \pm\beta_*$ for $j \in A^o$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}_n)$ in (1.3), and $\boldsymbol{x}_j$ are generated by greedy sequential sampling of groups 10 most correlated vectors from a pool of 600 iid vectors. See Section 3.5 for details. The design $\boldsymbol{X}$ is fixed, with the maximum absolute correlation 0.2459, $\|\boldsymbol{x}_j\| = \sqrt{n}$, and the minimum eigenvalue 0.0374 for $\boldsymbol{X}'\boldsymbol{X}/n$, while A^o and $\boldsymbol{\varepsilon}$ are drawn independently for the 100 replications with $d^o = |A^o| \in \{10, 20, 40\}$. The $\widehat{\sigma}^2$ is the residual mean squares with 100 degrees of freedom in the full 200-dimensional model. Bold face entries indicate $P\{\widehat{A} = A^o\} \approx \overline{CS} > 0.5$.

Why is the MC+ able to avoid both the interference of estimation bias with variable selection and the computational difficulties with non-convex minimization? A short, heuristic explanation is that for sparse $\boldsymbol{\beta}$ and carefully selected γ, the condition

$$\beta_* \equiv \min\Big\{|\beta_j| : j \in A^o\Big\} > \gamma\lambda, \ \ \lambda \geq \sigma\sqrt{2\log p}\Big(\max_{j \leq p} \|\boldsymbol{x}_j\|/n\Big), \tag{1.10}$$

provides the MC+ with unbiasedness at sufficiently large threshold λ, while a moderate maximum concavity $\kappa(\rho; \lambda) = 1/\gamma$ provides certain sparse convexity of the penalized loss

$$L(\boldsymbol{\beta}; \lambda) \equiv \frac{1}{2n}\big\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\big\|^2 + \sum_{j=1}^{p} \rho(|\beta_j|; \lambda), \quad \boldsymbol{\beta} \in \mathbb{R}^p. \tag{1.11}$$

The first inequality of (1.10) allows unbiased selection of all $j \in A^o$. The second one prevents selection of variables outside A^o given the selection of all variables in A^o in the linear model

(1.3) with $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$. Finally, the sparse convexity of (1.11) is needed for computational simplicity. We observe that the penalty function must satisfy

$$\lim_{t\to\infty} \dot\rho(t;\lambda) = 0, \quad \dot\rho(0+;\lambda) \equiv \lim_{t\to 0+} \dot\rho(t;\lambda) > 0, \tag{1.12}$$

for the unbiasedness and selection features. Since this excludes convex penalties, to ensure the convexity of the penalized loss, the convexity of the squared loss in (1.11) must overcome the concavity of the penalty as functions in $\mathbb{R}^p$, at least in sparse regions that matter.

2. A sketch of main results. In this section, we provide a brief description of our results, along with certain crucial concepts, conditions, and necessary notation.

In Section 3 we introduce the PLUS algorithm and discuss the choice of penalties and the regularization parameter γ for the MC+. The PLUS computes a piecewise linear path

$$\widehat{\boldsymbol{\beta}}(\lambda) \equiv \widehat{\boldsymbol{\beta}}^{(k)}(\lambda) = \frac{\lambda^{(k)} - \lambda}{\lambda^{(k)} - \lambda^{(k-1)}} \widehat{\boldsymbol{\beta}}^{(k-1)} + \frac{\lambda - \lambda^{(k-1)}}{\lambda^{(k)} - \lambda^{(k-1)}} \widehat{\boldsymbol{\beta}}^{(k)}, \ k = 1, \ldots, k^*, \tag{2.1}$$

with possibly non-increasing $\lambda^{(k)} \neq \lambda^{(k-1)}$ to accommodate multiple local minimizers. It begins with an initial segment $\widehat{\boldsymbol{\beta}}(\lambda) = \boldsymbol{\beta}^{(0)} = 0 \in \mathbb{R}^p$, $\lambda^{(0)} \leq \lambda < \infty$. For each segment k, $\widehat{\boldsymbol{\beta}}(\lambda) \equiv (\widehat{\beta}_1(\lambda), \ldots, \widehat{\beta}_p(\lambda))'$ satisfies the Kuhn-Tucker-type condition

$$\begin{cases} \boldsymbol{x}_j'\big(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\big)/n = \text{sgn}(\widehat{\beta}_j(\lambda))\dot\rho(|\widehat{\beta}_j(\lambda)|;\lambda), & \widehat{\beta}_j(\lambda) \neq 0 \\ \big|\boldsymbol{x}_j'\big(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\big)/n\big| \leq \lambda, & \widehat{\beta}_j(\lambda) = 0 \end{cases} \tag{2.2}$$

for (1.2). For almost all designs $\boldsymbol{X}$, segments of the PLUS path collectively form certain main branch of the solutions of (2.2). The main branch encompasses continuously from the origin to an optimal solution satisfying $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0$, while other branches of the solutions of (2.2) form separate continuous loops. We assume that the penalty function is of the form $\rho(t;\lambda) = \lambda^2\rho(t/\lambda)$ with a fixed quadratic spline $\rho(t)$, including the ℓ_1 penalty $\rho(t) = t$ for the LASSO as a special case. For such penalty functions, the PLUS has the following geometric interpretation: A transformation of all the branches of the solutions of (2.2) is the intersection of a single ray from the origin and a collection of adjacent and possibly overlapping parallelepipeds in $\mathbb{R}^p$, and each segment of the PLUS path corresponds to the intersection of the ray and one of these parallelepipeds. In this sense, the PLUS path is linear. The computational complexity of the PLUS path, depending on how the parallelepipeds fold, is easily manageable in our simulation experiments.

We prove that the PLUS provides the entire path of the global minimizer of (1.2) under a *global convexity* condition and the *sparsest solution* of (2.2) under a *sparse convexity* condition up to certain rank. The design matrix and penalty satisfy the global convexity condition if

$$c_{\min}\big(\boldsymbol{\Sigma}\big) + \frac{\dot\rho(t_2;\lambda) - \dot\rho(t_1;\lambda)}{t_2 - t_1} > 0, \quad \forall\, 0 < t_1 < t_2, \tag{2.3}$$

where $\boldsymbol{\Sigma} \equiv \boldsymbol{X}'\boldsymbol{X}/n$ and $c_{\min}(\boldsymbol{M})$ denotes the minimum eigenvalue of $\boldsymbol{M}$. For $A \subseteq \{1,\ldots,p\}$, define sub-design matrices and their standardized covariance as

$$\boldsymbol{X}_A \equiv (\boldsymbol{x}_j, j \in A)_{n\times|A|}, \quad \boldsymbol{\Sigma}_A \equiv \boldsymbol{X}_A'\boldsymbol{X}_A/n. \tag{2.4}$$

For $p > n$ or small $c_{\min}(\boldsymbol{\Sigma})$, we introduce the sparse convexity condition with *rank* d^* as

$$\kappa(\rho;\lambda) < c_* \le c_*(d^*) \equiv \min_{|A|\le d^*} c_{\min}\big(\boldsymbol{\Sigma}_A\big), \tag{2.5}$$

where $\kappa(\rho;\lambda)$ is the maximum concavity in (1.8). For $p > n$, $c_{\min}(\boldsymbol{\Sigma}) = 0$ and (2.3) does not hold. However, for variable selection (2.5) is as good as (2.3) when $|\widehat{A} \cup A^o| \le d^*$. Thus, (2.5) is very useful for sparse $\boldsymbol{\beta}$ when $p > n$ and for (nearly) singular $\boldsymbol{\Sigma}$ when $p \le n$. Since $c_*(2) = 1 - \max_{j\ne k}\big|\boldsymbol{x}_k'\boldsymbol{x}_j\big|/n$ for $\|\boldsymbol{x}_j\|^2/n = 1$, (2.5) holds for $d^* = 2$ in Table 1.

In Section 4, we consider the estimations of the mean squared error (MSE) of linear functionals of $\widehat{\boldsymbol{\beta}}$ and the noise level in the linear model (1.3). We derive estimators for the MSE and the *degrees of freedom* of the penalized LSE (1.2) via the SURE method of Stein (1981) and provide sufficient conditions for their unbiasedness. We prove that for full rank designs, the penalized LSE is continuous in $\boldsymbol{y} \in \mathbb{R}^n$ if and only if (iff) the global convexity (2.3) holds, iff the penalized loss function (1.11) is convex in the entire $\mathbb{R}^p$.

In Section 5, we study the probability of correct selection under the global convexity condition (2.3) for $p \le n$ and under the sparse Riesz condition (SRC)

$$\kappa(\rho;\lambda) < c_* \le c_*(d^*) \le c^*(d^*) \equiv \max_{|A|\le d^*} c_{\max}\big(\boldsymbol{\Sigma}_A\big) \le c^* \tag{2.6}$$

for general p and suitable $\{c_*, c^*, d^*\}$, where $c_{\max}(\boldsymbol{M})$ is the largest eigenvalue of $\boldsymbol{M}$, c_* is as in (2.5), and $\kappa(\rho;\lambda)$ is the maximum concavity in (1.8). For $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I}_n)$, $\|\boldsymbol{x}_j\|^2/n = 1$ and the penalty level $\lambda \ge \sigma\sqrt{2(\log p)/n}$, we prove the selection consistency (1.5) for the MC+ method under the SRC, provided that both $\beta_*/(\gamma\lambda)$ and d^*/d^o are greater than certain constants depending on $\{c_*, c^*, \gamma\}$ only, where β_*, γ and d^o are as in (1.10), (1.7) and (1.6) respectively. Under the global convexity condition (2.3), we obtain the explicit bound

$$P\Big\{\widehat{A} \ne A^o\Big\} \le (2p - d^o)\Phi\Big(-\sqrt{n}\frac{\lambda}{\sigma}\Big), \;\; \Phi(t) \equiv \int_{-\infty}^{t} \frac{e^{-x^2/2}}{(2\pi)^{1/2}}dx, \tag{2.7}$$

for $\beta_* \geq (\gamma + \sqrt{\gamma})\lambda$. An interesting aspect of this theory is its validity for p as large as $e^{a_0 n}$ with certain $a_0 > 0$, the *universal penalty level* $\lambda = \sigma\sqrt{2(\log p)/n}$ (Donoho and Johnston, 1994), and general penalty functions satisfying (1.9).

In Section 6, we briefly discuss adaptive penalty, an extension of the PLUS algorithm to generalized linear models, and penalized estimation.

3. The PLUS algorithm and quadratic spline penalties. We divide this section into 5 subsections to cover quadratic spline penalties, a geometric description of the PLUS algorithm, an algebraic description of the PLUS algorithm, penalized LSE for orthonormal designs, and the effects of the regularization parameter γ of the MC+ on the computational complexity and bias.

3.1. Quadratic spline penalties and the MCP. The PLUS algorithm assumes that the penalty function is of the form $\rho(t;\lambda) = \lambda^2\rho(t/\lambda)$, where $\rho(t)$ is an increasing quadratic spline in $[0,\infty)$. Such $\rho(t)$ must have a piecewise linear nonnegative continuous derivative $\dot\rho(t)$ for $t \geq 0$. We index such $\rho(t)$ by the number of threshold levels m, or equivalently the number of knots in $[0,\infty)$, including zero as a knot. Thus,

$$\rho(t;\lambda) = \lambda^2\rho_m(t/\lambda), \quad \dot\rho_m(t) \equiv \frac{d\rho_m}{dt}(t) = \sum_{i=1}^{m}(u_i - v_i t)I\{t_i \leq t < t_{i+1}\} \tag{3.1}$$

with $u_1 = 1$, $v_m = 0$, $t_{m+1} = \infty$ and knots $t_1 = 0 < t_2 < \cdots < t_m = \gamma$, satisfying $u_i - v_i t_{i+1} = u_{i+1} - v_{i+1}t_{i+1} \geq 0$, $1 \leq i < m$. Since $\kappa(\rho;\lambda)$ does not depend on λ for $\rho(t;\lambda) = \lambda^2\rho(t/\lambda)$, we denote the concavity measure (1.8) as $\kappa(\rho)$ in such cases.

We set $\dot\rho_m(0+) = u_1 = 1$ to match the standardization $\dot\rho(0+;\lambda) = \lambda$ in (1.9), and $v_m = 0$ for the uniform boundedness of $\dot\rho(t;\lambda)$. The unbiasedness parts of (1.12) or (1.9) demand $t_m = \gamma > 0$ and thus $m > 1$, but the PLUS includes the LASSO with $m = 1$. For $\|\boldsymbol{x}_j\|^2/n = 1$, $c_{\min}(\boldsymbol{\Sigma}_A) \leq 1$, so that (2.5) requires $\kappa(\rho_m) = \max_{i\leq m} v_i < c_* \leq 1$ for (3.1).

The penalty class (3.1) includes the ℓ_1 penalty with $m = 1$ and $\kappa(\rho_1) = 0$, the MCP with $m = 2$ and $\kappa(\rho_2) = v_1 = 1/\gamma$, and the SCAD penalty with $m = 3$, $v_1 = 0$, $t_2 = 1$ and $\kappa(\rho_3) = v_2 = 1/(\gamma - 1)$. We plot these three penalty functions ρ_m, $m = 1, 2, 3$ and their derivatives in Figure 1, with $\gamma = 5/2$ for the MCP and SCAD penalty.

As mentioned in the introduction, we propose the MCP (1.7) as the default penalty for the PLUS, and thus the acronym MC+. The MCP corresponds to (3.1) with

$$\rho_2(t) = \min\Big\{t - t^2/(2\gamma), \gamma/2\Big\}, \quad \dot\rho_2(t) = (1 - t/\gamma)^+. \tag{3.2}$$

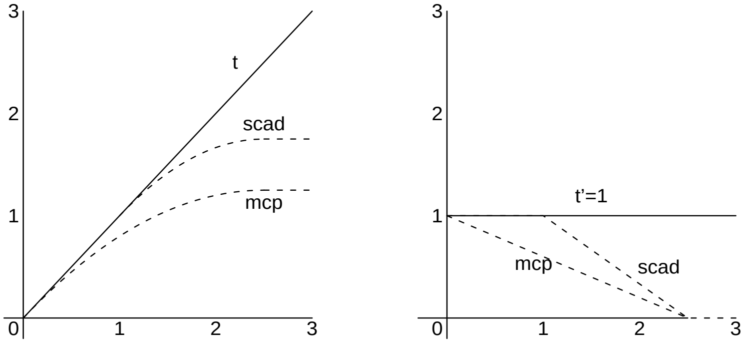

Figure 1: *The ℓ_1 penalty $\rho_1(t) = t$ for the LASSO along with the MCP $\rho_2(t)$ and the SCAD penalty $\rho_3(t)$, $t > 0$, $\gamma = 5/2$. Left: penalties $\rho_m(t)$. Right: their derivatives $\dot{\rho}_m(t)$.*

Among spline penalties satisfying (1.9), the MCP has the smallest number of threshold levels $m = 2$. Since the PLUS path makes a turn when $|\widehat{\beta}_j(\lambda)/\lambda|$, $j \le p$, hit one of the m thresholds, MC+ is the simplest for the PLUS to compute except for the LASSO with $m = 1$. For continuously differentiable penalty $\rho(t;\lambda)$, define

$$\ddot{\rho}(t;\lambda) \equiv \lim_{\epsilon\to 0+} \inf_{0<|x|\le\epsilon} \Big(\dot{\rho}(t+x;\lambda) - \dot{\rho}(t;\lambda)\Big)\Big/ x, \quad t > 0, \tag{3.3}$$

so that $-\ddot{\rho}(t;\lambda)$ measures the local concavity of $\rho(\cdot;\lambda)$ at $t > 0$ and $\kappa(\rho;\lambda) = \sup_{t>0}\big\{-\ddot{\rho}(t)\big\}$ in (1.8) measures the maximum concavity of the penalty $\rho(t;\lambda)$. We call (1.7) the minimax-concave penalty since it has the smallest maximum concavity $\kappa(\rho;\lambda)$ given the threshold level $\gamma\lambda$ for the "complete unbiasedness" in (1.9). Since $\kappa(\rho;\lambda)$ is minimized at $\kappa(\rho_1)$ under (1.9), the MCP offers the global convexity (2.3) of the penalized loss with the smallest possible $c_{\min}(\mathbf{\Sigma})$ and the sparse convexity (2.5) with the highest rank d^*. Thus, it fulfills our main smoothness conditions with the greatest stability. Moreover, for a given level of concavity $\kappa(\rho;\lambda)$, the MCP ensures the unbiasedness above the smallest second threshold $\gamma\lambda$ in (1.9) and thus the smallest separation zone $\beta_* > \gamma\lambda$ in (1.10). Therefore, in our program, the MCP provides computational simplicity, smoothness, unbiasedness and accurate selection for the penalized LSE to the greatest extent. Moreover, the MC+ allows the regularization parameter γ to be set in the entire continuum of $(0,\infty]$. Subsection 3.4 contains further discussion about MCP and other penalty functions and their relationship to threshold estimators for $p = 1$. Subsection 3.5 discusses the effects of regularization parameter γ in (1.7) and (3.2).

3.2. A geometric description of the PLUS algorithm. Let $\widetilde{\boldsymbol{z}} \equiv \boldsymbol{X}'\boldsymbol{y}/n$. For penalty functions of the form $\rho(t;\lambda) = \lambda^2\rho(t/\lambda)$, the optimization problem (1.2) is equivalent to

$$\boldsymbol{b}(\boldsymbol{z}) \equiv \arg\min_{\boldsymbol{b}} \Big\{ -\boldsymbol{b}'\boldsymbol{z} + \frac{1}{2}\boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{b} + \sum_{j=1}^{p} \rho(|b_j|) \Big\} \tag{3.4}$$

through the scale change $\widetilde{\boldsymbol{z}} \to \lambda\boldsymbol{z}$ and $\boldsymbol{\beta} \to \lambda\boldsymbol{b}$, where $\boldsymbol{\Sigma} \equiv \boldsymbol{X}'\boldsymbol{X}/n$. The solution of (3.4) along the ray $\{\widetilde{\boldsymbol{z}}/\lambda, \lambda > 0\}$ provides the solution of (1.2) with the inverse transformation $\widehat{\boldsymbol{\beta}}(\lambda) = \lambda\boldsymbol{b}(\widetilde{\boldsymbol{z}}/\lambda)$. In this subsection, we describe the PLUS algorithm in the rescaled problem (3.4) through a geometric interpretation of its path via the following rescaled version of (2.2):

$$\begin{cases} z_j - \boldsymbol{\chi}_j'\boldsymbol{b} = \mathrm{sgn}(b_j)\dot{\rho}_m(|b_j|), & b_j \neq 0, \\ \big|z_j' - \boldsymbol{\chi}_j'\boldsymbol{b}\big| \le 1 = \dot{\rho}_m(0+), & b_j = 0, \end{cases} \tag{3.5}$$

where ρ_m is as in (3.1) and $\boldsymbol{\chi}_j \equiv \boldsymbol{X}'\boldsymbol{x}_j/n$ are the columns of $\boldsymbol{\Sigma}$.

We shall "plot" the solution $\boldsymbol{b}(\boldsymbol{z})$ against $\boldsymbol{z}$ to allow multiple solutions, instead of directly solving (3.5) for a given $\boldsymbol{z} = \widetilde{\boldsymbol{z}}/\lambda = \boldsymbol{X}'\boldsymbol{y}/(n\lambda)$. In the univariate case $p = 1$, we plot functions in $\mathbb{R}^2$. For $p > 1$, we need to consider $\boldsymbol{b}$ versus $\boldsymbol{z}$ in $\mathbb{R}^{2p}$. Let $H = \mathbb{R}^p$, H^* be its dual, and $\boldsymbol{z} \oplus \boldsymbol{b}$ be members of $H \oplus H^* = \mathbb{R}^{2p}$. Define

$$u(i) \equiv u_{|i|}, \quad v(i) \equiv v_{|i|}, \quad t(i) \equiv \begin{cases} t_i, & 0 < i \le m+1 \\ -t_{|i|+1}, & -m \le i \le 0, \end{cases} \tag{3.6}$$

with the u_i, v_i and t_i in (3.1). For indicators $\boldsymbol{\eta} \in \{-m,\ldots,m\}^p$, let

$$S(\boldsymbol{\eta}) \;\equiv\; \text{the set of all } \boldsymbol{z} \oplus \boldsymbol{b} \text{ satisfying } \begin{cases} z_j - \boldsymbol{\chi}_j'\boldsymbol{b} = \mathrm{sgn}(\eta_j)u(\eta_j) - b_j v(\eta_j), & \eta_j \neq 0 \\ -1 \le z_j - \boldsymbol{\chi}_j'\boldsymbol{b} \le 1, & \eta_j = 0 \\ t(\eta_j) \le b_j \le t(\eta_j + 1), & \eta_j \neq 0 \\ b_j = 0, & \eta_j = 0. \end{cases} \tag{3.7}$$

Since $\mathrm{sgn}(b_j)\dot{\rho}_m(|b_j|) = \mathrm{sgn}(\eta_j)u(\eta_j) - b_j v(\eta_j)$ for $t(\eta_j) \le b_j \le t(\eta_j + 1)$, (3.5) holds iff (3.7) holds for certain $\boldsymbol{\eta}$. For each $\boldsymbol{\eta}$, the linear system in (3.7) is of rank $2p$, since one can always uniquely solve for $\boldsymbol{b}$ and then $\boldsymbol{z}$ if the inequalities are replaced by equations. Thus, since (3.7) has p equations and p pairs of parallel inequalities, $S(\boldsymbol{\eta})$ are p-dimensional parallelepipeds living in $H \oplus H^* = \mathbb{R}^{2p}$. Due to the continuity of $\dot{\rho}_m(t) = (d/dt)\rho_m(t)$ in t by (3.1) and that of $z_j - \boldsymbol{\chi}_j'\boldsymbol{b}$ in both $\boldsymbol{z}$ and $\boldsymbol{b}$, the solutions of (3.7) are identical in the intersection of any given pair

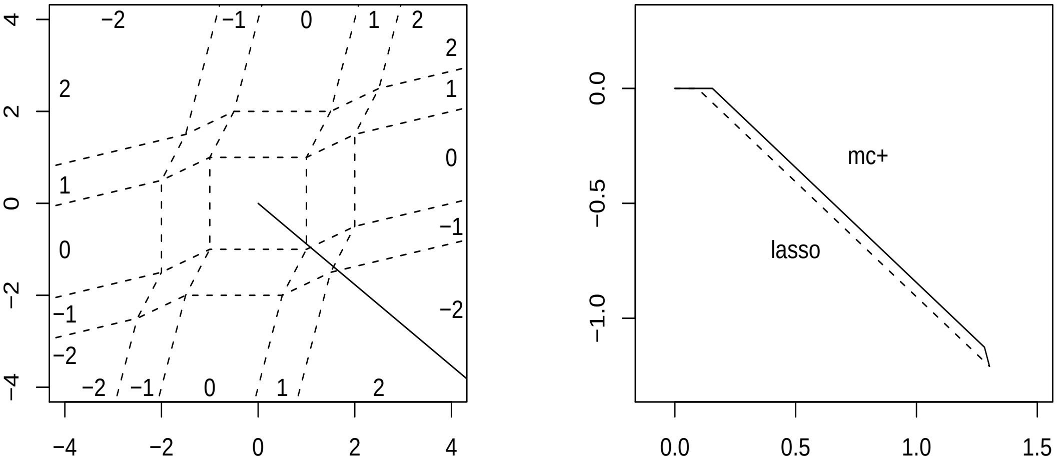

Figure 2: *Left: The solid ray as $\tau\widetilde{\boldsymbol{z}}$ and the projections of the $5^2 = 25$ parallelograms $S(\boldsymbol{\eta})$ for the MCP $\gamma = 2$ to the $\boldsymbol{z}$-space H with dashed-edges, labeled by η_1 and η_2 along the margins inside the box. Right: The MC+ path (solid) as the entire solution set of (2.2) in the $\boldsymbol{\beta}$-space, along with the LASSO path (dashed). Data: $\|\boldsymbol{x}_j\|^2/2 = 1$, $\boldsymbol{x}_1'\boldsymbol{x}_2/2 = 1/4$, $(\widetilde{z}_1, \widetilde{z}_2) = (1, -0.883)$, and $p = 2$.*

of $S(\boldsymbol{\eta})$ with adjacent $\boldsymbol{\eta}$. Furthermore, the p-dimensional interiors of different $S(\boldsymbol{\eta})$ are disjoint in view of the constraints of (3.7) on $\boldsymbol{b}$. Thus, the union of all the p-parallelepipeds $S(\boldsymbol{\eta})$ forms a continuous p-dimensional surface $S \equiv \cup\big\{S(\boldsymbol{\eta}) : \boldsymbol{\eta} \in \{-m, \ldots, m\}^p\big\}$ in $H \oplus H^* = \mathbb{R}^{2p}$. The solution set of (3.5) for all $\boldsymbol{z} = \tau\widetilde{\boldsymbol{z}}$, $\tau = 1/\lambda$, or equivalently the solution set of (2.2) for all $\lambda > 0$, is identical to the intersection of this surface S and the $(p+1)$-dimensional open half subspace $\big\{(\tau\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b} : \tau > 0, \boldsymbol{b} \in H^*\big\}$ in $\mathbb{R}^{2p}$. Figure 2 depicts the projections of $S(\boldsymbol{\eta})$ to H and the MC+ and LASSO solutions for $p = 2$ under the global convexity condition (2.3).

The rescaled PLUS path in $H \oplus H^*$ is a union of connected line segments

$$\cup_{k=0}^{k^*}\ell(\boldsymbol{\eta}^{(k)}|\widetilde{\boldsymbol{z}}), \quad \ell(\boldsymbol{\eta}|\boldsymbol{z}) \equiv S(\boldsymbol{\eta}) \cap \Big\{(\tau\boldsymbol{z}) \oplus \boldsymbol{b} : \tau \geq 0, \boldsymbol{b} \in H^*\Big\}, \tag{3.8}$$

beginning with $\ell(\boldsymbol{\eta}^{(0)}|\widetilde{\boldsymbol{z}}) = \big\{(\tau\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}^{(0)} : 0 \leq \tau \leq \tau^{(0)}\big\}$, $\boldsymbol{\eta}^{(0)} = \boldsymbol{b}^{(0)} = 0$, and connected at

$$\big\{(\tau^{(k-1)}\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}^{(k-1)}\big\} = \ell(\boldsymbol{\eta}^{(k-1)}|\widetilde{\boldsymbol{z}}) \cap \ell(\boldsymbol{\eta}^{(k)}|\widetilde{\boldsymbol{z}}), \quad \widetilde{\boldsymbol{z}} \equiv \boldsymbol{X}'\boldsymbol{y}/n. \tag{3.9}$$

We initialize (3.9) with $\boldsymbol{b}^{(0)} = 0$ and $\tau^{(0)} = 1/\max_{j\leq p}|\widetilde{z}_j|$. Given $(\tau^{(k-1)}\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}^{(k-1)}$, we find a *new line segment* $\ell(\boldsymbol{\eta}^{(k)}|\widetilde{\boldsymbol{z}})$ and compute the other end of it as $(\tau^{(k)}\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}^{(k)}$, $k \geq 1$. The PLUS path ends at step k^* if $(\tau^{(k^*)}\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}^{(k^*)}$ provides an optimal fit satisfying $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}^{(k^*)}/\tau^{(k^*)}) = 0$ with $\tau^{(k^*)} > 0$. If an optimal fit can not be found in the current pass, the PLUS searches through the existing turning points to find a connected new line segment to restart. In this case, the PLUS path can still be written as a single sequence satisfying (3.8) and (3.9) with some repeating segments. This provides the geometric description of the PLUS algorithm.

Non-degenerate designs: *We say that the design matrix $\boldsymbol{X}$ in (1.3) is non-degenerate if for all $A \subset \{1, \ldots, p\}$ of size $|A| = n \wedge p - 1$ and $\eta_j \in \{-1, 0, 1\}$, $j \le p$, the $n \wedge p$ vectors*

$$\boldsymbol{x}_j, j \in A, \ \sum_{k \notin A} \eta_k \boldsymbol{x}_k \text{ are linearly independent.} \tag{3.10}$$

For $p \le n$, $\boldsymbol{X}$ is non-degenerate iff $\text{rank}(\boldsymbol{X}) = p$.

Theorem 1. (i) *Suppose the design matrix $\boldsymbol{X}$ is non-degenerate. Given $\boldsymbol{X}$, there exist a finite set $\Gamma_0(\boldsymbol{X})$ such that for $\gamma \notin \Gamma_0(\boldsymbol{X})$ the MC+ path is composed of sequentially connected line segments $\ell(\boldsymbol{\eta}^{(k)}|\widetilde{\boldsymbol{z}})$ with turning points $(\tau^{(k)}\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}^{(k)}$, $\tau^{(k)} > 0$, $k = 1, \ldots, k^* < \infty$, and ends with an optimal fit satisfying $\boldsymbol{X}'\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}^{(k^*)}/\tau^{(k^*)}\big) = 0$. Consequently, the PLUS path $\widehat{\boldsymbol{\beta}}(\lambda)$ in (2.1) ends with the optimal fit $\widehat{\boldsymbol{\beta}}^{(k^*)} = \boldsymbol{b}^{(k^*)}/\tau^{(k^*)}$ as a solution of (2.2) for all $0 \le \lambda \le \lambda^{(k^*)} = 1/\tau^{(k^*)}$.*
(ii) *For fixed $\gamma > 0$, the design matrix $\boldsymbol{X}$ is non-degenerate and $\gamma \notin \Gamma_0(\boldsymbol{X})$ almost everywhere in $\mathbb{R}^{n\times p}$ under the Lebesgue measure.*
(iii) *For fixed positive $\gamma \ne 1$, the design matrix $\boldsymbol{X}$ is non-degenerate and $\gamma \notin \Gamma_0(\boldsymbol{X})$ almost everywhere under the product of p Haar measures in the $(n-1)$-sphere $\{\boldsymbol{x} : \|\boldsymbol{x}\|^2/n = 1\}$.*

Under the conditions of Theorem 1 (i), the MC+ path forms a main branch from $\widehat{\boldsymbol{\beta}} = 0$ to a point of optimal fit in the graph of the solution set of (2.2). We actually prove that the MC+ algorithm finishes in one pass almost everywhere in $\widetilde{\boldsymbol{z}} \in \mathbb{R}^p$. Theorem 1 (ii) and (iii) assert that the conditions of Theorem 1 (i) hold almost everywhere in $\boldsymbol{X}$ for all fixed $\{n, p, \gamma\}$. Conditions of Theorem 1 (i) is not necessary for the MC+ path to end with an optimal fit. For example, if $\boldsymbol{x}_j = \pm\boldsymbol{x}_k$, the PLUS path uses at most one design vector $\boldsymbol{x}_j$ or $\boldsymbol{x}_k$ in any step, so that it behaves as if one of them never exists. For simplicity, we omit an extension of Theorem 1 to the PLUS with general quadratic penalty (3.1).

Theorem 1 does not guarantee that the PLUS path contains all solutions of (2.2), but Theorem 2 below does under the global convexity condition (2.3). Figure 3 depicts an example in which the complete solution set of (2.2) contains the main branch covered by the MC+ path and a loop not covered. Still, Theorem 7 in Section 5 shows that under the SRC (2.6), the PLUS path provides variable selection consistency in such cases.

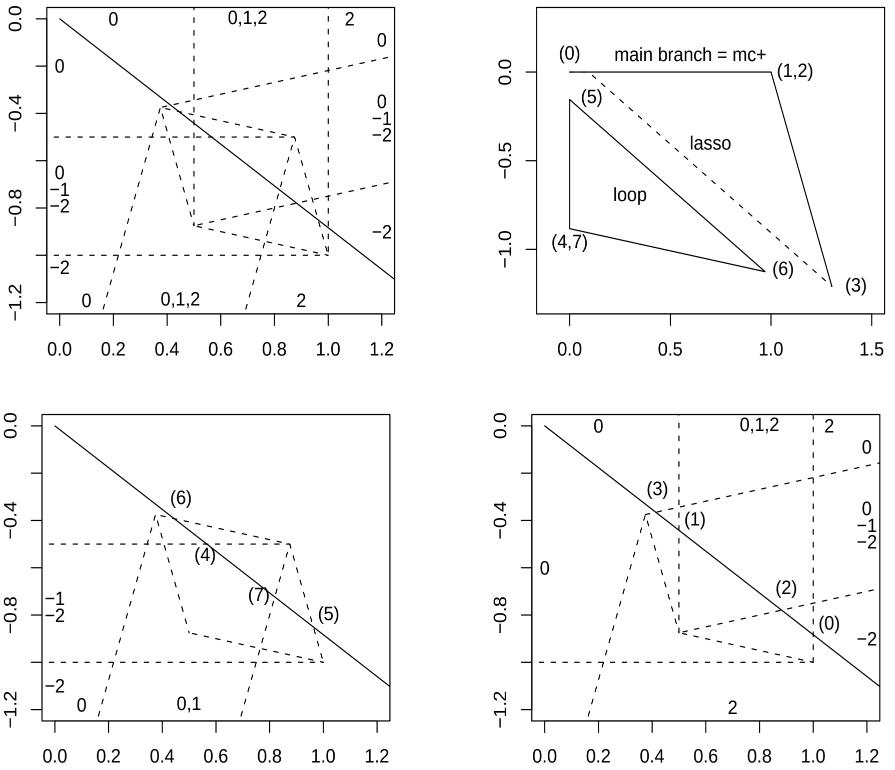

Figure 3: *Plots for the same data as in Figure 2 with $\gamma = 1/2$ for the MCP. Clockwise from the top left: the $\boldsymbol{z}$-space plot with overlapping areas marked by multiple values of η_j; the main branch and one loop as the entire MCP solution set of (2.2) in the $\boldsymbol{\beta}$-space, along with the LASSO; the segments of the main branch with $\tau^{(k)}\widetilde{\boldsymbol{z}}$, $k = 0, 1, 2, 3$, representing transitions $\boldsymbol{\eta} = \binom{0}{0} \to \binom{1}{0} \to \binom{2}{0} \to \binom{2}{-1} \to \binom{2}{-2}$; the loop with $\tau^{(k)}\widetilde{\boldsymbol{z}}$, $k = 4, 5, 6, 7$, representing transitions $\boldsymbol{\eta} = \binom{0}{-2} \to \binom{0}{-1} \to \binom{1}{-1} \to \binom{1}{-2} \to \binom{0}{-2}$. For $\boldsymbol{\eta} \in \{-2, 0, 2\}^p$, $\boldsymbol{z}$-segments turn into $\boldsymbol{\beta}$-points in the MC+ path.*

Let $\boldsymbol{\Sigma}_A$ be as in (2.4). Define $\boldsymbol{\Sigma}(\boldsymbol{\eta}) \equiv \boldsymbol{\Sigma}_{\{j:\eta_j \neq 0\}}$ and

$$\boldsymbol{Q}(\boldsymbol{\eta}) \equiv \boldsymbol{\Sigma}(\boldsymbol{\eta}) - \mathrm{diag}\big(v(\eta_j), \eta_j \neq 0\big), \quad d(\boldsymbol{\eta}) \equiv \#\{j : \eta_j \neq 0\}. \tag{3.11}$$

Since the $\boldsymbol{\chi}_j$ in (3.5) are the columns of $\boldsymbol{\Sigma}$, the first equation of (3.7) can be written as

$$\boldsymbol{Q}(\boldsymbol{\eta})\boldsymbol{P}(\boldsymbol{\eta})\boldsymbol{b} = \boldsymbol{P}(\boldsymbol{\eta})\big(\boldsymbol{z} - \mathrm{sgn}(\boldsymbol{\eta})u(\boldsymbol{\eta})\big), \tag{3.12}$$

where $\boldsymbol{P}(\boldsymbol{\eta}) : \boldsymbol{b} \to (b_j, \eta_j \neq 0)'$ are projections and $u(\cdot)$ is as in (3.6).

Proof of Theorem 1. Let $\boldsymbol{X}$ be fixed. Define $d_k(\boldsymbol{\eta}) \equiv \#\{j : |\eta_j| = k\}$, $k = 1, 2$. We consider three types of indicators $\boldsymbol{\eta} \in \{-2, -1, 0, 1, 2\}^p$ with $\boldsymbol{\eta} = 0$ as Type-1.

Type-2: $d_2(\boldsymbol{\eta}) \ge n \wedge p$. Let $(\tau\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b} \in S(\boldsymbol{\eta})$ as in (3.7), so that (3.5) holds with $z_j = \tau\widetilde{z}_j = \tau\boldsymbol{x}_j'\boldsymbol{y}/n$. Since $\dot\rho_2(|b_j|) = 0$ for $|\eta_j| = 2$, (3.5) implies $\boldsymbol{x}_j'(\tau\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = 0$ for all $|\eta_j| = 2$. Since $\tau\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \in \mathbb{R}^n$ and $\{\boldsymbol{x}_j, |\eta_j| = 2\}$ contains at least $n \wedge p$ linearly independent vectors, by (3.8)

$$\begin{cases} d_2(\boldsymbol{\eta}) \ge n \wedge p \\ (\tau\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b} \in \ell(\boldsymbol{\eta}|\widetilde{\boldsymbol{z}}) \end{cases} \quad \Rightarrow \quad \boldsymbol{X}'(\tau\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = 0. \tag{3.13}$$

Type-3: $d_2(\boldsymbol{\eta}) < n \wedge p$ and $\boldsymbol{\eta} \ne 0$. If (3.5) holds for $\boldsymbol{z} = 0$, then $b_j\boldsymbol{x}_j'\boldsymbol{X}\boldsymbol{b}/n = b_j\boldsymbol{\chi}_j'\boldsymbol{b} = -|b_j|\dot\rho_2(|b_j|)$ for all $b_j \ne 0$, so that $\|\boldsymbol{X}\boldsymbol{b}\|^2/n = -\sum_j |b_j|\dot\rho_2(|b_j|) = 0$ due to $\dot\rho_2(|b_j|) \ge 0$. Since $\dot\rho_2(|b_j|) = 1/\gamma > 0$ for $|b_j| < \gamma$ and $|b_j| \le \gamma$ for $|\eta_j| = 1$, $b_j = \gamma\eta_j$ for $|\eta_j| = 1$ in such cases. Therefore, $\boldsymbol{X}\boldsymbol{b} = \sum_{|\eta_j|=2} b_j\boldsymbol{x}_j + \gamma\sum_{|\eta_k|<2}\eta_k\boldsymbol{x}_k = 0$. This is impossible for non-degenerate $\boldsymbol{X}$ since $\gamma > 0$ and $d_2(\boldsymbol{\eta}) < n \wedge p$. Thus, $0 \oplus \boldsymbol{b} \notin S(\boldsymbol{\eta})$ for indicators $\boldsymbol{\eta}$ of Type-3.

We now consider the choice of γ for the MC+. It follows from (3.2) and (3.11) that the determinant $\det(\boldsymbol{Q}(\boldsymbol{\eta}))$ is a polynomial of $v_1 = 1/\gamma$ with $\det(\boldsymbol{\Sigma}_{j:|\eta_j|=2})(-v_1)^{d_1(\boldsymbol{\eta})} \ne 0$ as the leading term. Let $\Gamma_0(\boldsymbol{X})$ be the finite set of all reciprocals of the roots of such polynomials. We choose $\gamma \notin \Gamma_0(\boldsymbol{X})$ here after, so that $\det(\boldsymbol{Q}(\boldsymbol{\eta})) \ne 0$ for all $\boldsymbol{\eta}$ of Type-3. Since $\det(\boldsymbol{Q}(\boldsymbol{\eta})) \ne 0$, in $S(\boldsymbol{\eta})$ the vector $(b_j, \eta_j \ne 0)'$ is a linear function of $\boldsymbol{z}$ by (3.12) , so that by (3.8) and the discussion in the previous paragraph

$$\begin{cases} d_2(\boldsymbol{\eta}) < n \wedge p \\ \boldsymbol{\eta} \ne 0 \end{cases} \quad \Rightarrow \quad \begin{cases} \ell(\boldsymbol{\eta}|\boldsymbol{z}) \ \text{ is a generalized line segment} \\ 0 \oplus \boldsymbol{b} \notin \ell(\boldsymbol{\eta}|\boldsymbol{z}) \ \forall \boldsymbol{b}. \end{cases} \tag{3.14}$$

Here a generalized line segment includes the empty set, single points in $H \oplus H^* = \mathbb{R}^{2p}$, and line segments of finite or infinite length.

For each nonzero $\boldsymbol{z} \in H \equiv \mathbb{R}^p$, we define a graph $G(\boldsymbol{z})$ with $\ell(\boldsymbol{\eta}|\boldsymbol{z})$ of positive length and Type-3 $\boldsymbol{\eta}$ as edges and the end points of edges as vertexes. The graph $G(\boldsymbol{z})$ is not necessarily connected. A vertex in $G(\boldsymbol{z})$ is terminal if it is also a boundary point of $S(\boldsymbol{\eta})$ for some $\boldsymbol{\eta}$ of Type-2. If the MC+ path reaches a terminal vertex $(\tau\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}$, then $\boldsymbol{b}/\tau$ provides an optimal fit by (3.13) and (3.14). The degree of a vertex in $G(\boldsymbol{z})$ is the number of edges connected to it.

Suppose $\widetilde{\boldsymbol{z}} \ne 0$. At step $k = 0$, the MC+ path reaches $(\tau^{(0)}\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}^{(0)}$ as a boundary point of $S(0)$. Since the p-parallelepipeds (3.7) are contiguous, $(\tau^{(0)}\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}^{(0)}$ is also a boundary point of $S(\boldsymbol{\eta}^{(1)})$ for some $\boldsymbol{\eta}^{(1)}$ satisfying either (3.13) or (3.14) with $\boldsymbol{z} = \widetilde{\boldsymbol{z}}$. If $\boldsymbol{\eta}^{(1)}$ is of Type-2, then $\boldsymbol{b}^{(0)}/\tau^{(0)}$ gives an optimal fit and the MC+ path ends with $k^* = 0$. Otherwise, the MC+ path enters the graph $G(\widetilde{\boldsymbol{z}})$ at the initial vertex $(\tau^{(0)}\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}^{(0)}$. If the degree of the initial vertex is odd and the degrees of all other non-terminal vertexes are even, then the MC+ path traverses through $G(\widetilde{\boldsymbol{z}})$ and eventually reaches a terminal vertex in one pass. This is simply an Euler's Konigsberg problem.

Let S_0 be the union of all intersections of three or more distinct p-parallelepipeds $S(\boldsymbol{\eta})$, $\boldsymbol{\eta} \in \{-2,-1,0,1,2\}^p$, and $H_0 \equiv \{\boldsymbol{z} : (\tau\boldsymbol{z}) \oplus \boldsymbol{b} \in S_0 \text{ for some } \tau \text{ and } \boldsymbol{b}\}$. Since the interiors of the p-parallelepipeds $S(\boldsymbol{\eta})$ do not intersect, the intersections of three distinct $S(\boldsymbol{\eta})$ are $(p-2)$-parallelepipeds, so that the projection of S_0 to the $(p-1)$-sphere $\{\boldsymbol{z} : \|\boldsymbol{z}\| = 1\}$ along the rays $\{\tau\boldsymbol{z}, \tau > 0\}$ has Haar measure zero. Consequently, H_0 has Lebesgue measure zero in $H \equiv \mathbb{R}^p$.

For $\boldsymbol{z} \notin H_0$, each vertex in $G(\boldsymbol{z})$ is a boundary point of exactly two p-parallelepipeds $S(\boldsymbol{\eta})$, so that the initial vertex has degree 1 and other non-terminal vertexes have degree 2 in $G(\boldsymbol{z})$. Thus, the initial vertex is connected to a terminal vertex in $G(\widetilde{\boldsymbol{z}})$ for $\widetilde{\boldsymbol{z}} \notin H_0$. For $\widetilde{\boldsymbol{z}} \in H_0$, the initial vertex is still connected to at lease one terminal vertex in $G(\widetilde{\boldsymbol{z}})$ since H_0^c is dense in $H \equiv \mathbb{R}^p$ and the limits of $G(\boldsymbol{z})$ as $\boldsymbol{z} \to \widetilde{\boldsymbol{z}}$ are subgraphs of $G(\widetilde{\boldsymbol{z}})$. Hence, the PLUS path reaches a terminal vertex in either cases. □

3.3. An algebraic description of the PLUS algorithm. We provide formulas for a simplified version of the PLUS algorithm (3.8) and (3.9), in which we only look for line segments $\ell(\boldsymbol{\eta}^{(k)}|\widetilde{\boldsymbol{z}})$ with positive length, new $\boldsymbol{\eta}^{(k)}$, and $\det(\boldsymbol{Q}(\boldsymbol{\eta}^{(k)})) \neq 0$. This simplification has no impact in our simulation experiments, since it is identical to the full version of the PLUS algorithm for almost all datasets according to Theorem 1 and its proof.

For $\det(\boldsymbol{Q}(\boldsymbol{\eta}^{(k)})) \neq 0$, the map $\boldsymbol{z} \to \boldsymbol{b}$ is unique in $S(\boldsymbol{\eta}^{(k)})$ by (3.5), (3.7) and (3.12). Thus, $\tau^{(k)} \neq \tau^{(k-1)}$ for the k-th segment $\ell(\boldsymbol{\eta}^{(k)}|\widetilde{\boldsymbol{z}})$. We write the PLUS path (2.1) as

$$\widehat{\boldsymbol{\beta}}^{(k)} = \lambda^{(k)}\boldsymbol{b}^{(k)},\ \lambda^{(k)} = 1/\tau^{(k)},\ \boldsymbol{b}(\tau\widetilde{\boldsymbol{z}}) = \boldsymbol{b}^{(k-1)} + \big(\tau - \tau^{(k-1)}\big)\boldsymbol{s}^{(k)}, \tag{3.15}$$

with $\widehat{\boldsymbol{\beta}}(\lambda) = \lambda\boldsymbol{b}(\widetilde{\boldsymbol{z}}/\lambda)$ and $\tau = 1/\lambda$ between $\tau^{(k-1)}$ and $\tau^{(k)}$, $\widetilde{\boldsymbol{z}} \equiv \boldsymbol{X}'\boldsymbol{y}/n$ as in (3.9), the initial segment $\boldsymbol{b}(\tau\widetilde{\boldsymbol{z}}) = \boldsymbol{b}^{(0)} \equiv 0$ for all $0 \leq \tau \leq \tau^{(0)} \equiv 1/\max_j |\widetilde{z}_j|$, turning points $\boldsymbol{b}^{(k)} \equiv (b_1^{(k)}, \ldots, b_p^{(k)})' \in H^* \equiv \mathbb{R}^p$, "slopes" $\boldsymbol{s}^{(k)} \equiv (s_1^{(k)}, \ldots, s_p^{(k)})'$, and hitting times $\tau^{(k)} > 0$. We note that the map $\tau \to \boldsymbol{b}(\tau\widetilde{\boldsymbol{z}})$ is potentially many-to-one and that τ may not be a monotone function as (3.15) traverses through the solution set of (3.5).

As in (3.8), let $\boldsymbol{\eta}^{(k)}$ be the indicator of the p-parallelepiped (3.7) in which the k-th piece of the rescaled PLUS path $(\tau\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}(\tau\widetilde{\boldsymbol{z}})$ lives. In the k-th step, we compute $\boldsymbol{\eta}^{(k)}$, $\boldsymbol{s}^{(k)}$, $\tau^{(k)}$ and $\boldsymbol{b}^{(k)}$ given $\boldsymbol{\eta}^{(0)}, \ldots, \boldsymbol{\eta}^{(k-1)}$, $\tau^{(k-1)}$ and $\boldsymbol{b}^{(k-1)}$. Let

$$C^{(k)} \equiv \Big\{j : |b_j^{(k)}| \in \{t_1, \ldots t_m\} \text{ with } \eta_j^{(k)} \neq 0 \text{ or } \big|\tau^{(k)}\widetilde{z}_j - \boldsymbol{\chi}_j'\boldsymbol{b}^{(k)}\big| = 1\Big\}$$

be the set of critical indices j at which (3.15) hits the boundary of the inequalities in (3.7) in step k at $\tau = \tau^{(k)}$. We try all $2^{|C^{(k-1)}|} - 1$ nonempty subsets of $C^{(k-1)}$ for the path of (3.15) to cross the boundaries at the beginning of its k-th segment, or equivalently all $2^{|C^{(k-1)}|} - 1$ possible candidates for $\boldsymbol{\eta}^{(k)}$. We note that the "one-at-a-time" condition $|C^{(k-1)}| = 1$ holds almost everywhere. Here, $\boldsymbol{\eta}^{(k)}$ is required to be new, so that $\boldsymbol{\eta}^{(k)} \neq \boldsymbol{\eta}^{(\ell)}$ for $0 \leq \ell < k$.

Given $\boldsymbol{\eta}^{(k)}$, the identity (3.12) must hold for $\boldsymbol{\eta} = \boldsymbol{\eta}^{(k)}$, $\boldsymbol{z} = \tau\widetilde{\boldsymbol{z}}$ and $\boldsymbol{b} = \boldsymbol{b}(\tau\widetilde{\boldsymbol{z}})$. Differentiating this identity with respect to τ, we find by (3.12) and (3.15) that

$$\boldsymbol{Q}(\boldsymbol{\eta}^{(k)})\boldsymbol{P}(\boldsymbol{\eta}^{(k)})\boldsymbol{s}^{(k)} = \boldsymbol{P}(\boldsymbol{\eta}^{(k)})\widetilde{\boldsymbol{z}}, \quad \eta_j^{(k)} = 0 \Rightarrow s_j^{(k)} = 0. \tag{3.16}$$

Moreover, the j-th negative gradient of the loss in (3.4) and its derivative are

$$g_j(\tau\widetilde{\boldsymbol{z}}) = \tau\widetilde{z}_j - \boldsymbol{\chi}_j'\boldsymbol{b}(\tau\widetilde{\boldsymbol{z}}), \quad \frac{d}{d\tau}g_j(\tau\widetilde{\boldsymbol{z}}) = \widetilde{z}_j - \boldsymbol{\chi}_j'\boldsymbol{s}^{(k)}, \tag{3.17}$$

by (3.5). Given a tentative choice $\boldsymbol{\eta}^{(k)} \neq \boldsymbol{\eta}^{(\ell)}$, $0 \leq \ell < k$, we check if the solution $\boldsymbol{s}^{(k)}$ of (3.16) indeed carries the k-th segment of the path from the p-parallelepiped $S(\boldsymbol{\eta}^{(k-1)})$ into $S(\boldsymbol{\eta}^{(k)})$ according to (3.7), or equivalently (3.5), for two possible signs of

$$\xi^{(k)} \equiv \operatorname{sgn}(\tau^{(k)} - \tau^{(k-1)}). \tag{3.18}$$

It follows from (3.15) and (3.17) that for either $\xi^{(k)} = \pm 1$ this amounts to verifying

$$\begin{cases} \xi^{(k)}(\eta_j^{(k)} - \eta_j^{(k-1)})s_j^{(k)} \geq 0, & \eta_j^{(k-1)} \neq \eta_j^{(k)} \neq 0 \\ \xi^{(k)}(\widetilde{\eta}_j^{(k)} - \eta_j^{(k-1)})s_j^{(k)} \leq 0, & \eta_j^{(k-1)} = \eta_j^{(k)} \neq 0,\ \widetilde{\eta}_j^{(k)} \neq \eta_j^{(k-1)} \\ \xi^{(k)}\eta_j^{(k-1)}(\widetilde{z}_j - \boldsymbol{\chi}_j'\boldsymbol{s}^{(k)}) \leq 0, & \eta_j^{(k-1)} \neq \eta_j^{(k)} = 0 \\ \xi^{(k)}\widetilde{\eta}_j^{(k)}(\widetilde{z}_j - \boldsymbol{\chi}_j'\boldsymbol{s}^{(k)}) \leq 0, & \eta_j^{(k-1)} = \eta_j^{(k)} = 0 \neq \widetilde{\eta}_j^{(k)}, \end{cases} \tag{3.19}$$

where $\widetilde{\boldsymbol{\eta}}^{(k)}$ is the "neighbor" of $\boldsymbol{\eta}^{(k-1)}$ with $\widetilde{\eta}_j^{(k)} \neq \eta_j^{(k-1)}, \forall\ j \in C^{(k-1)}$. We note that (3.19) checks all indices j with $\widetilde{\eta}_j^{(k)} \neq \eta_j^{(k-1)}$. If the tentative choices of $\boldsymbol{\eta}^{(k)}$ and $\xi^{(k)}$ and their associated $\boldsymbol{s}^{(k)}$ pass the test (3.19), we move on to find $\tau^{(k)}$.

Let $\Delta^{(k)} \equiv \left|\tau^{(k)} - \tau^{(k-1)}\right|$ be the length of the k-th segment of the PLUS path measured in τ. Given the slope $\boldsymbol{s}^{(k)}$ and the sign $\xi^{(k)}$ of $d\tau$ for the segment, there are at most p possible ways for $(\tau\widetilde{\boldsymbol{z}}) \oplus \boldsymbol{b}(\tau\widetilde{\boldsymbol{z}})$ to hit a new side of the boundary of the p-parallelepiped $S(\boldsymbol{\eta}^{(k)})$ in (3.7).

If it first hits the boundary indexed by $\eta_j^{(k)}$, $\Delta^{(k)}$ would be

$$\Delta_j^{(k)} = \begin{cases} \xi_j^{(k)}\{t(\eta_j^{(k)}+1) - b_j^{(k-1)}\}/s_j^{(k)}, & \xi_j^{(k)} s_j^{(k)} > 0 \neq \eta_j^{(k)} \\ \xi_j^{(k)}\{t(\eta_j^{(k)}) - b_j^{(k-1)}\}/s_j^{(k)}, & \xi_j^{(k)} s_j^{(k)} < 0 \neq \eta_j^{(k)} \\ \xi_j^{(k)}\{1 - g_j^{(k-1)}\}/\{\widetilde{z}_j - \boldsymbol{\chi}_j'\boldsymbol{s}^{(k)}\}, & \xi_j^{(k)}(\widetilde{z}_j - \boldsymbol{\chi}_j'\boldsymbol{s}^{(k)}) > 0 = \eta_j^{(k)} \\ \xi_j^{(k)}\{-1 - g_j^{(k-1)}\}/\{\widetilde{z}_j - \boldsymbol{\chi}_j'\boldsymbol{s}^{(k)}\}, & \xi_j^{(k)}(\widetilde{z}_j - \boldsymbol{\chi}_j'\boldsymbol{s}^{(k)}) < 0 = \eta_j^{(k)} \end{cases} \tag{3.20}$$

by (3.7), (3.15) and (3.17), where the function $t(\cdot)$ is as in (3.6) and $g_j^{(k-1)} \equiv g_j(\tau^{(k-1)}\widetilde{\boldsymbol{z}})$ is the j-th negative gradient (3.17) at $\tau = \tau^{(k-1)}$. It follows that

$$\tau^{(k)} = \tau^{(k-1)} + \xi^{(k)}\Delta^{(k)}, \quad \Delta^{(k)} = \min_{1\le j\le p} \Delta_j^{(k)}. \tag{3.21}$$

We note that (3.19) guarantees $\Delta_j^{(k)} > 0, \forall j$ in (3.20). We formally write the PLUS as follows.

The PLUS Algorithm.

Initialization: $\boldsymbol{b}^{(0)} \leftarrow 0$, $\tau^{(0)} \leftarrow 1/\max_{j\le p}|\widetilde{z}_j|$, $k=1$

Iteration:

$$\text{Find } \boldsymbol{\eta}^{(k)} \neq \boldsymbol{\eta}^{(\ell)},\ 0 \le \ell < k, \text{ such that (3.19) holds}$$
$$\text{for the solution } \boldsymbol{s}^{(k)} \text{ of (3.16) and some } \xi^{(k)} = \pm 1 \tag{3.22}$$
$$\text{Find } \tau^{(k)} \text{ according to (3.21)} \tag{3.23}$$
$$\boldsymbol{b}^{(k)} \leftarrow \boldsymbol{b}^{(k-1)} + \big(\tau^{(k)} - \tau^{(k-1)}\big)\boldsymbol{s}^{(k)} \tag{3.24}$$
$$k \leftarrow k+1 \tag{3.25}$$

Termination: (3.22) has no solution for $k = k^*+1$ or $\tau^{(k^*)} = \infty$.

Output: $\tau^{(0)}$, $\boldsymbol{b}^{(0)}$, $\boldsymbol{\eta}^{(k)}$, $\boldsymbol{s}^{(k)}$, $\tau^{(k)}$, $\boldsymbol{b}^{(k)}$, $k = 1, 2, \ldots, k^*$

Theorem 2. *Let $\widehat{\boldsymbol{\beta}}(\lambda)$ be defined by (2.1) and (3.15) with the output of the PLUS.*
(i) *If $\boldsymbol{Q}(\boldsymbol{\eta}^{(k)})$ in (3.11) is positive-definite and $\ell(\boldsymbol{\eta}^{(k)}|\widetilde{\boldsymbol{z}})$ does not live on the boundary of $S(\boldsymbol{\eta}^{(k)})$ in (3.8), then $\widehat{\boldsymbol{\beta}}(\lambda)$ is a local minimizer of the penalized loss (1.11) in the k-th segment of its path strictly between the turning points $\widehat{\boldsymbol{\beta}}^{(k-1)} \equiv \lambda^{(k-1)}\boldsymbol{b}^{(k-1)}$ and $\widehat{\boldsymbol{\beta}}^{(k)} \equiv \lambda^{(k)}\boldsymbol{b}^{(k)}$.*
(ii) *Suppose the sparse convexity condition (2.5) holds with rank d^*. Then, for any given $A \subseteq \{1,\ldots,p\}$ with $|A| \le d^*$, the penalized loss (1.11) is strictly convex and has a unique minimizer under the constraint $\beta_j = 0\ \forall j \notin A$. In particular, if $d(\boldsymbol{\eta}^{(k)}) \le d^*/2$, then $\widehat{\boldsymbol{\beta}}(\lambda)$ in the k-th segment provides the unique solution of (2.2) satisfying $\#\{j : \widehat{\beta}_j \neq 0\} \le d^*/2$.*

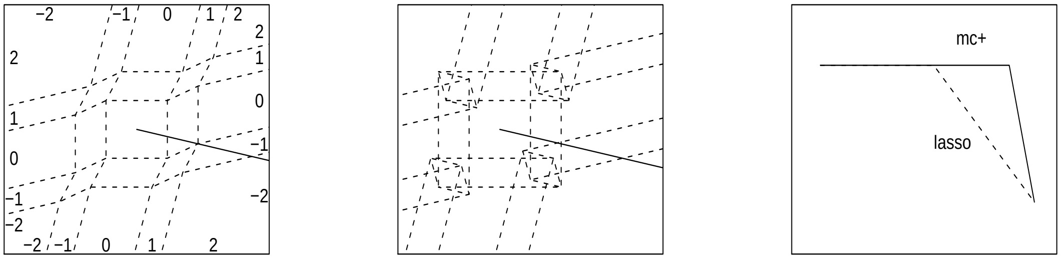

Figure 4: *The same type of plots as in Figures 2 and 3 for the same* $\boldsymbol{X}$ *and more sparse* $(\widetilde{z}_1, \widetilde{z}_2) = (1, -1/2)$. *From the left: the* $\boldsymbol{z}$*-space plot for MC+ with* $\gamma = 2$*; MC+ with* $\gamma = 1/2$*; the MC+ (same for both* $\gamma = 2$ *and* $\gamma = 1/2$*) and LASSO paths in the* $\boldsymbol{\beta}$*-space.*

(iii) *If the global convexity condition (2.3) holds, then* $\lambda^{(k-1)} > \lambda^{(k)}$*, the PLUS path ends with the LSE* $\widehat{\boldsymbol{\beta}}^{(k^*)} = \boldsymbol{\Sigma}^{-1}\boldsymbol{X}'\boldsymbol{y}/n$*, and* $\widehat{\boldsymbol{\beta}}(\lambda)$ *is always the unique solution of (2.2), i.e. the global minimizer of (1.11) and the solution of (1.2).*

The significance of Theorem 2 (ii) is as follows. If $\max(d(\boldsymbol{\eta}^{(k)}), d^o) \le d^*/2$ and the pattern A^o in (1.1) is identified by a solution of (2.2), then the solution is given by $\widehat{\boldsymbol{\beta}}(\lambda)$ in the k-th segment of the PLUS path. Thus, the effect of non-convexity is less pronounced for sparse data. In the example in Figure 4, the convex penalized loss with $\gamma = 2$ yields identical MC+ path as the non-convex one with $\gamma = 1/2$ for sparse data outside regions where the the projections of the parallelograms $S(\boldsymbol{\eta})$ fold severely in the $\boldsymbol{z}$-space for $\gamma = 1/2$. This should be compared with the dramatic difference between $\gamma = 2$ and $\gamma = 1/2$ in Figures 2 and 3 for dense data.

Choice among multiple solutions: Theorem 2 suggests that among multiple solutions in the PLUS path for a specific λ, we choose the sparsest one, the solution in the segment with the smallest $d(\boldsymbol{\eta}^{(k)}) = \mathrm{rank}(\boldsymbol{P}(\boldsymbol{\eta}^{(k)}))$, subject to the positive-definiteness of $\boldsymbol{Q}(\boldsymbol{\eta}^{(k)})$. We break the ties by further minimizing the "degrees of freedom" as the trace of $\boldsymbol{Q}(\boldsymbol{\eta}^{(k)})\boldsymbol{\Sigma}(\boldsymbol{\eta}^{(k)})$ and then maximizing $c_{\min}\big(\boldsymbol{Q}(\boldsymbol{\eta}^{(k)})\big)$ for the stability of the estimator. See (3.11) for the notation and Subsection 4.2 for the justification of the degrees of freedom and stability measures.

Proof of Theorem 2. (i) Since $(\widetilde{\boldsymbol{z}} \oplus \widehat{\boldsymbol{\beta}}(\lambda))/\lambda$ is in the interior of $S(\boldsymbol{\eta}^{(k)})$, (2.2) holds with strict inequality. Thus, by (3.11), the directional derivative of the penalized loss (1.11)

$$\frac{\partial}{\partial t} L(\widehat{\boldsymbol{\beta}}(\lambda) + t\boldsymbol{b}; \lambda) = t\boldsymbol{b}_1'\boldsymbol{Q}(\boldsymbol{\eta}^{(k)})\boldsymbol{b}_1 + \sum_{\eta_j^{(k)}=0} |b_j|\Big(\lambda - \mathrm{sgn}(b_j)\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda))/n + O(t)\Big)$$

is positive for small $t > 0$, where $\boldsymbol{b}_1 = \boldsymbol{P}(\boldsymbol{\eta}^{(k)})\boldsymbol{b}$. Thus, $\widehat{\boldsymbol{\beta}}(\lambda)$ is a local minimizer.

(ii) The strict convexity and uniqueness follow from (iii), since the sparse convex condition implies the global convex condition under the constraint $\beta_j = 0$ for all $j \notin A$. Let λ be fixed. For any solutions $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ of (2.2) with $\#\{j : \widehat{\beta}_j \neq 0\} \le d^*/2$ and $\#\{j : \widetilde{\beta}_j \neq 0\} \le d^*/2$, the constrained uniqueness implies $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}$.

(iii) Let λ be fixed and define $h(t) \equiv \kappa(\rho;\lambda)t^2/2 + \rho(|t|;\lambda) - \lambda|t|$. Since $\dot{\rho}(0+;\lambda) = \lambda$, $h(t)$ is a continuously differentiable convex function in view of the definition of $\kappa(\rho;\lambda)$ in (1.8). It follows that the penalized loss

$$L(\boldsymbol{\beta};\lambda) = \Big\{\frac{1}{2n}\big\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\big\|^2 - \frac{\kappa(\rho;\lambda)}{2}\|\boldsymbol{\beta}\|^2\Big\} + \sum_{j=1}^{p}\Big\{|b_j| + h(|b_j|)\Big\} \tag{3.26}$$

is a sum of two convex functions, with the first one being strictly convex for $c_{\min}(\boldsymbol{\Sigma}) > \kappa(\rho;\lambda)$ and the second one being strictly convex otherwise. This gives the uniqueness. □

3.4. Orthonormal designs and more discussion on penalties. For orthonormal designs $\boldsymbol{x}_j'\boldsymbol{x}_k/n = I\{j = k\}$, the penalized estimation problem (1.2) is reduced to the case of $p = 1$. For $\rho(t;\lambda) = \lambda^2\rho_m(t/\lambda)$ with the quadratic spline penalties (3.1), (3.4) becomes

$$\widehat{\beta}_j = \lambda b\big(\boldsymbol{x}_j'\boldsymbol{y}/(n\lambda)\big), \quad b(z) \equiv \arg\min_b \Big\{(z-b)^2/2 + \rho_m(|b|)\Big\}. \tag{3.27}$$

For $p = 1$ and the MCP with $\kappa(\rho_2) = 1/\gamma < 1$, the solution of (3.27) is

$$b_f(z) = \operatorname{sgn}(z)\min\Big(|z|, \frac{\gamma(|z|-\lambda)^+}{\gamma - 1}\Big),$$

which turns out to be the firm threshold estimator of Gao and Bruce (1997). The firm threshold estimator is always between the soft threshold estimator $b_s(z) \equiv \operatorname{sgn}(z)(|z|-\lambda)^+$ and the hard threshold estimator $b_h(z) \equiv zI\{|z| > \lambda\}$. Actually, $b_s(z) \le b(z) \le b_f(z) \le b_h(z)$ for $z > 0$ and the opposite inequalities hold for $z < 0$ for all solutions of (3.27), given a fixed $\gamma\lambda$ in (1.9) or a fixed maximum concavity $\kappa(\rho_m) = 1/\gamma$ with $\gamma > 1$. For example, the univariate SCAD

$$b_{SCAD}(z) \equiv \operatorname{sgn}(z)\min\Big[|z|, \max\Big\{(|z|-\lambda)^+, \frac{(\gamma-1)|z| - \gamma\lambda}{\gamma - 2}\Big\}\Big], \quad \gamma > 2,$$

satisfies these inequalities and has the concavity measure $\kappa(\rho_3) = 1/(\gamma-1) > \kappa(\rho_2) = 1/\gamma$. We plot these univariate estimators in Figure 5.

For $p = 1$ and $\kappa(\rho_2) = 1/\gamma \ge 1$, the MC+ path has three segments and its sparsest solution gives the hard threshold estimator. See Figure 5 on the left. Antoniadis and Fan (2001) observed that in the orthonormal case, the global minimizer (3.27) for the penalty (1.7)

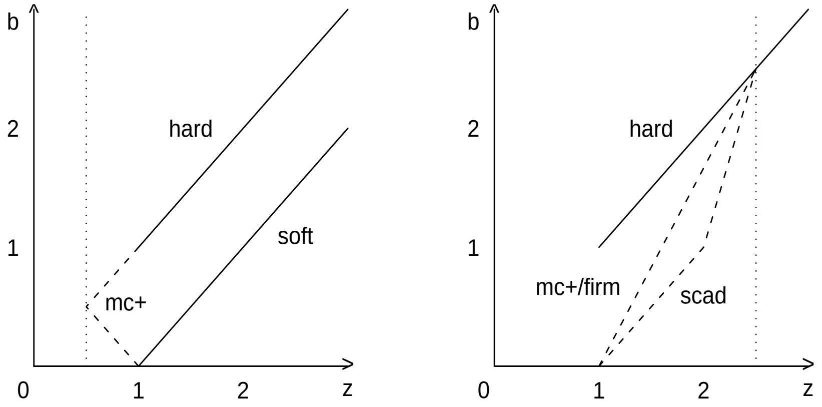

Figure 5: *Left: The univariate hard, soft and MC+ paths in* $z \oplus b \in H \oplus H^* = \mathbb{R}^2$ *with a vertical dotted line at* $z = \gamma = 1/2$*. Right: The hard, MC+/firm and SCAD paths for* $p = 1$ *with* $\gamma = 5/2$*. Hard and soft path in solid, and additional segments of MC+ and SCAD in dashed lines.*

with $\gamma = 1/2$ yields the hard threshold estimator. In fact, in the univariate case, any penalty function with concave derivative $\dot{\rho}(t;\lambda)$ and $\gamma \le 1$ in (1.9) yields the hard threshold estimator as the global minimizer in (3.27).

A crucial difference between the MC+ and hard threshold estimator for $p = 1$, and more generally between the PLUS and other non-convex minimization algorithms for $p > 1$, is in the ways the multiple solutions of (2.2) are treated. Unlike existing iterative algorithms which search for a local minimizer of the penalized loss (1.11) for fixed λ with a given initial guess or tries to jump from the domain of attraction of one local minimizer to another, the PLUS continuously tracks multiple solutions of (2.2) for the entire range of λ and thus is computationally more efficient. In Figure 5 on the left, the dashed segments of the MC+ path continuously connect the two segments of the hard threshold estimator. In Figure 3, the MC+ path has 4 segments labeled by $\boldsymbol{\eta}^{(k)} = \binom{0}{0}$, $\binom{1}{0}$, $\binom{2}{0}$, and $\binom{2}{-1}$, respectively for $k = 0, \ldots, 3$, and terminates with a point of optimal fit at the boundary of the parallelogram indexed by $\boldsymbol{\eta} = \binom{2}{-2}$. The line segments in the rescaled path $\boldsymbol{b}(\lambda)$ turn into single points $\widehat{\boldsymbol{\beta}}(\lambda)$ for $k \in \{0, 2\}$, corresponding to the selection of label sets $A = \emptyset$ and $\{1\}$, while the segments for $k \in \{1, 3\}$ connect these solutions and the terminal point. Given a set of solutions of (2.2) for a given λ, we choose the sparsest local minimizer instead of the one with the smallest penalized loss.

The analytical and computational properties of penalized estimation and selection for general correlated $\boldsymbol{X}$ and concave penalty is much more complicated than the case of $p = 1$, since they are determined in many ways by the interplay between the penalty and the design. To a large extent, the effects of the penalty can be summarized by the threshold factor γ for the unbiasedness in (1.9), the maximum concavity $\kappa(\rho; \lambda)$ in (1.8) and their relationships to the correlations of the design vectors. This naturally leads to our choice of the MCP as the minimizer of $\kappa(\rho; \lambda)$ given the threshold factor γ and the role of $\gamma = 1/\kappa(\rho_1)$ as the regularization parameter for the bias and computational complexity of the MC+.

For $p > 1$ with correlated $\boldsymbol{x}_j$, the interpretations of "hard" or "firm" estimators are not clear. If "hard" means discontinuity of the global minimizer (1.2) in $\boldsymbol{y}$ or the non-uniqueness of (2.2), then by Theorem 5 in Subsection 4.3 the criterion is the failure of (2.3), which means $\kappa(\rho_m) \geq c_{\min}(\boldsymbol{\Sigma})$ for (3.1) and is not a property of the penalty function alone.

3.5. Computational complexity and bias. For the MC+, the tuning parameter γ regulates computational complexity and bias level. Here we study its effects through three simulation experiments, say Experiments 1, 2 and 3.

Experiment 1, summarized in Table 1 in Section 1, illustrates the superior selection accuracy of the MC+ for sparse $\boldsymbol{\beta}$, compared with the LASSO and SCAD. Experiment 2, summarized in Table 2 here, shows the effects of the regularization parameter γ on selection accuracy and computational complexity of the MC+. Experiment 3, summarized in Table 3, demonstrates the scalability of the MC+ methodology for large p. The design vectors $\boldsymbol{x}_j$ are identical in Experiments 1 and 2. We generate a 300×600 random matrix as the difference of two independent random matrices, the first with iid unit exponential entries and the second iid χ^2_1 entries. We normalize the 600 columns of this difference matrix to summation zero and Euclidean length $\sqrt{n}$. We then sequentially sample groups of 10 vectors from this pool of normalized columns. For the m-th group, we sample from the remaining $610-10m$ columns one member as $\boldsymbol{x}_{10m-9}$ and 9 more to maximize the absolute correlation $|\boldsymbol{x}_j' \boldsymbol{x}_{10m-9}|/n$, $j = 10m-8, \ldots, 10m$. In Experiment 3, $\boldsymbol{x}_j$ are generated in the same way with groups of size 50 from a pool of 6000 iid columns, yielding an $\boldsymbol{X}$ with maximum absolute correlation $\max_{j<k} |\boldsymbol{x}_j' \boldsymbol{x}_k|/n = 0.3041$. In all the three experiments, $\beta_j = \pm\beta_*$ for $j \in A^o$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}_n)$.

Strong effects of bias on selection accuracy is observed in all three tables. Heuristically, condition (1.10) provides an unbiased solution of (2.2) for the MC+ and SCAD, while the

Table 2: Performance of MC+ with Different γ in Experiment 2

100 replications, $n = 300$, $p = 200$, $d^o = 30$, $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}_n)$

TM $= |\widehat{A} \setminus A^o| + |A^o \setminus \widehat{A}|$; LASSO for $\gamma = \infty$; rows with $\overline{CS} \leq 0.03$ not reported

+(++): 12 (18) replications fail to reach $\lambda = 0.1$ up to 5000 steps for $\gamma = 1/2$

β_*	λ	γ	0.50	0.99	1.01	1.40	2.652	5.00	∞
3/8	$\sqrt{(\log p)/n}$ $= 0.1329$	$\overline{CS}$	0.03+	0.05	0.05	0.05	0.03	0.00	0.00
		$\overline{TM}$	3.73+	2.45	2.45	2.29	2.45	4.53	9.36
		$\overline{k}$	1504+	668	558	103	53	35	39
3/8	$\sqrt{2(\log p)/n}$ $= 0.1879$	$\overline{CS}$	**0.76+**	**0.84**	**0.84**	**0.83**	0.32	0.11	0.00
		$\overline{TM}$	1.83+	**0.26**	**0.26**	**0.26**	1.58	2.82	4.77
		$\overline{k}$	1495+	660	550	99	32	31	33
	$\sqrt{4(\log p)/n}$	$\overline{TM}$	8.14+	7.23	7.43	9.36	7.76	7.45	7.64
3/8	$= 0.2658$	$\overline{k}$	1319+	462	369	56	23	24	25
3/8	any ≥ 0.1	$\overline{CS}$	**0.87+**	**0.97**	**0.97**	**0.97**	**0.59**	0.20	0.06
3/8		β_*/γ	0.75	0.38	0.37	0.27	0.14	0.08	0.00
1/4	0.1329	$\overline{TM}$	6.73++	5.58	5.77	**5.56**	7.91	9.40	11.44
1/4	0.1879	$\overline{TM}$	11.99++	11.70	11.81	12.29	11.29	11.26	11.41
1/4	0.2658	$\overline{TM}$	21.63++	21.62	21.62	21.39	19.54	19.00	18.67
1/4	any ≥ 0.1	$\overline{CS}$	0.08++	0.11	0.10	0.08	0.01	0.00	0.00
1/4		β_*/γ	0.50	0.25	0.25	0.18	0.09	0.05	0.00

strength of the signal overcomes the bias for all PLUS procedures when β_* is of the order $\sqrt{d^o}\lambda$ or larger. These rules can be easily verified in the three tables via $\lambda \leq \beta_*/\gamma$ or $\lambda \leq M\beta_*/\sqrt{d^o}$ for a moderate M, with $\beta_*/\gamma = 1.885$ and $\beta_*/\sqrt{d^o} = (0.1581, 0.1118, 0.0791)$ in Table 1. They explain the behavior of $P\{\widehat{A} = A^o\} \approx \overline{CS}$ in most entries in the three tables.

In Tables 2 and 3, we report two additional measures of selection accuracy. The second measure is the total miss (TM), the sum of the total false discovery $|\widehat{A} \setminus A^o|$ and the total false negative $|A^o \setminus \widehat{A}|$. Again, the average $\overline{TM}$ over 100 replications demonstrates the superior performance of the MC+ in our simulation experiments. The third measure of selection accuracy is the correct selection CS for any point in the PLUS path up to the stopping rule $\lambda^{(k^*)} \leq 0.1$. Comparison of this measure for the entire path with the $\overline{CS}$ for individual values of λ shows that the universal penalty level $\lambda = \sigma\sqrt{2(\log p)/n}$ is a good choice for variable selection in the MC+ path for standardized designs with $\|\boldsymbol{x}_j\|^2/n = 1$, although the MC+ is

Table 3: Performance of MC+ with $p > n$ in Experiment 3

100 replications, $n = 300$, $p = 2000$, $d^o = 30$, $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}_n)$; *: SCAD with $\gamma = 2.4$

		$\beta_* = 1/2$					$\beta_* = 3/8$				
λ	γ	1.1	1.4	1.7	2.4*	∞	1.1	1.4	1.7	2.4*	∞
$\sqrt{(\log p)/n}$	$\overline{CS}$	0.02	0.02	0.02	0.01	0.00	0.02	0.02	0.02	0.00	0.00
$= 0.1592$	$\overline{TM}$	4.07	4.02	4.56	6.41	48.79	8.03	5.81	**5.23**	29.46	47.32
	$\overline{k}$	366	104	78.3	240	80.0	680	167	100	231	75.2
$\sqrt{2(\log p)/n}$	$\overline{CS}$	**0.93**	**0.93**	**0.93**	0.01	0.00	0.26	0.14	0.05	0.00	0.00
$= 0.2251$	$\overline{TM}$	**0.07**	**0.07**	**0.07**	14.77	25.24	7.61	7.06	7.77	24.68	25.41
	$\overline{k}$	353	98.4	72.7	128	53.9	392	111	61.4	56.4	46.4
$\sqrt{4(\log p)/n}$	$\overline{CS}$	**0.53**	0.16	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$= 0.3183$	$\overline{TM}$	2.14	4.15	6.59	14.82	15.08	17.87	17.58	16.77	18.15	18.16
	$\overline{k}$	342	80.8	42.5	40.2	36.1	152	35.3	19.9	23.8	23.8
any ≥ 0.15	$\overline{CS}$	**1.00**	**1.00**	**0.99**	0.22	0.00	**0.65**	**0.64**	0.41	0.00	0.00
	β_*/γ	0.45	0.36	0.29	0.21	0.00	0.34	0.27	0.22	0.16	0.00

somewhat confused with the choice of $\lambda^{(k)}$ in its path in the most difficult cases.

As expected, we observe in Tables 2 and 3 that MC+ with smaller γ is computationally more costly. Dramatic rise in the number of needed PLUS steps is observed when γ decreases to 1/2. We avoid $\gamma = 1$, since it produces singular $\boldsymbol{Q}(\boldsymbol{\eta}) = 0$ for (3.12) whenever $\sum_{j=1}^{p} |\eta_j| = 1$ for the MC+ with the standardization $\|\boldsymbol{x}_j\|^2/n = 1$.

An interesting phenomenon exhibited in Experiments 2 and 3 is that the observed selection accuracy $\overline{CS}$ is always decreasing in γ. Despite the computational complexity for small γ, the MC+ still recovers the true A^o among so many line segments it traverses through. This suggests that the interference of the bias, not the complexity of the path or the lack of the convexity of the penalized loss, is a dominant factor in variable selection. Of course, bias reduction does not always provide accurate variable selection. When the separation zone shrinks to $\beta_* = 1/4$ from $\beta_* = 3/8$ in Table 2, the selection accuracy suddenly drops to $\overline{CS} \leq 0.11$ for all values of (λ, γ). This seems to indicate that for $\beta_* = 1/4$, the data simply do not contain sufficient information for the identification of A^o due to the failure of (1.10).

Table 3 shows that the MC+ methodology scales well for $p > n$. Comparisons between Tables 2 and 3 demonstrate that for similar d^o and signal strength, the computational complexity of the MC+ is insensitive to p as measured by the average number of steps $\overline{k}$.

4. The MSE, degrees of freedom, and noise level.

In this section, we consider the estimation of the MSE of $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\boldsymbol{\mu}}(\lambda) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)$ under the normality assumption on the errors in (1.3). Formulas for unbiased estimation of risks are derived and justified via the SURE. Necessary and sufficient conditions are provided for the continuity of the penalized LSE. We first consider the estimation of the noise level as it is needed for both risk estimation and proper choice of λ.

4.1. Estimation of noise level.

Consider throughout this subsection standardized designs with $\|\boldsymbol{x}_j\|^2/n = 1$ for all $j \le p$. We have shown in Tables 1, 2 and 3 that the MC+ with the universal penalty level $\lambda = \sigma\sqrt{2(\log p)/n}$ works well for variable selection in the linear model (1.3) with $N(0, \sigma^2)$ errors. In practice, this requires a reasonable estimate of the noise level σ. For $p < n$, the mean residual squares $\|\boldsymbol{y} - \widetilde{\boldsymbol{\mu}}\|^2/\{n - \text{rank}(\boldsymbol{X})\}$ for the full model provides an unbiased estimator of σ^2 as in Table 1, where $\widetilde{\boldsymbol{\mu}}$ is the projection of $\boldsymbol{y}$ to the linear span of the design vectors $\{\boldsymbol{x}_j, j \le p\}$. However, the estimation of σ^2 is a more delicate problem for $p > n$ or small $n - p > 0$. In this subsection, we present some simulation results for a simple estimator of σ^2 in the case of $p > n$.

If (2.2) provides consistent estimates $\widehat{\boldsymbol{\mu}}(\lambda)$ of the mean $\boldsymbol{\mu} \equiv \boldsymbol{X}\boldsymbol{\beta}$, we may estimate σ^2 by

$$\widehat{\sigma}^2(\lambda) \equiv \frac{\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}(\lambda)\|^2}{n - \widehat{\text{df}}(\lambda)} \tag{4.1}$$

with certain $\widehat{\text{df}}(\lambda)$ for the adjustment of degrees of freedom. We will provide a formula of $\widehat{\text{df}}(\lambda)$ in (4.12) below. Still, good $\widehat{\sigma}^2(\lambda)$ requires a consistent $\widehat{\boldsymbol{\mu}}(\lambda)$, which depends on the choice of a suitable λ of the order $\sigma\sqrt{(\log p)/n}$. This circular estimation problem can be solved with

$$\widehat{\sigma} \equiv \widehat{\sigma}(\widehat{\lambda}), \quad \widehat{\lambda} \equiv \min\Big\{\lambda \ge \lambda_* : \widehat{\sigma}^2(\lambda) \le \frac{n\lambda^2}{r_0(\log p)}\Big\}, \tag{4.2}$$

for suitable $r_0 \le 2$ and $\lambda_* > 0$. Here λ_* could be preassigned or determined by upper bounds on $\widehat{\text{df}}(\lambda)$ or the dimension $\#\{j : \widehat{\beta}_j(\lambda) \ne 0\}$. In principle, we may also use in (4.2) estimates $\widehat{\sigma}^2(\lambda)$ based on cross-validation or bootstrap, but the computationally much simpler (4.1) turns out to have the best overall performance in our experiments with $r_0 = 1$ in (4.2).

In Figures 6 and 7, we present simulation results for the estimation of σ in Experiments 4 and 5. In Experiment 4, $\gamma = 1.7$, $\beta_* = 1/2$, and $\boldsymbol{\beta}$ is generated every 10 replications. Its configurations are otherwise identical to that of Experiment 3 reported in Table 3. In Experiment 5, $\boldsymbol{x}_j$ are normalized columns from a Gaussian random matrix with iid rows and the correlation $\sigma_{j,k} = \sigma_{1,2}^{|k-j|}$ among entries within each row, $\gamma = 2/(1 - \max_{j>k}|\boldsymbol{x}_k'\boldsymbol{x}_j|/n)$

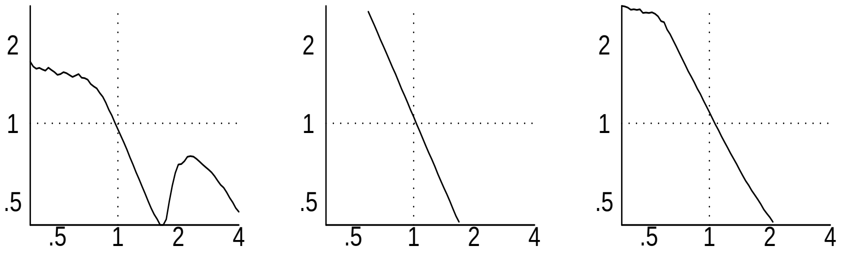

Figure 6: *The median of $\widehat{\sigma}^2(\lambda)/(n\lambda^2/\log p)$ as function of $\lambda/\sqrt{(\log p)/n} \in [2^{-3/2}, 4]$ based on 100 replications. Left: Experiment 4 with $n = 300$, $p = 2000$ and $d^o = 30$. Middle and right: Experiment 5 with high and low correlations respectively, $n = 600$, $p = 3000$ and $d^o = 35$.*

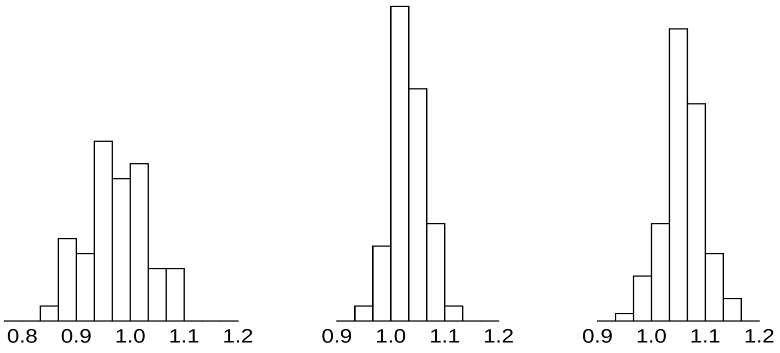

Figure 7: *Histograms of $\widehat{\sigma}$ for the same simulations as in Figure 6.*

as in Experiment 1, the nonzero β_j are composed of 5 blocks of $\beta_*(1, 2, 3, 4, 3, 2, 1)'$ centered at random multiples $j_1, \ldots, j_5$ of 25, β_* sets $\|\boldsymbol{X}\boldsymbol{\beta}\|^2/n = 3$, $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}_n)$, and $\{\boldsymbol{X}, \boldsymbol{\beta}\}$ are generated every 10 replications. It has two settings: $\sigma_{1,2} = 0.9$ for high correlation and $\sigma_{1,2} = 0.1$ for low correlation. We set $\lambda_* = \{2^{-3}(\log p)/n\}^{1/2}$ in both Experiments 4 and 5.

Figure 6 plots the median of $\widehat{\sigma}^2(\lambda)/(n\lambda^2/\log p)$ versus $\lambda/\sqrt{(\log p)/n}$ in the simulations described above. Since all three curves cross the level $\widehat{\sigma}^2(\lambda)/(n\lambda^2/\log p) = 1$ at approximately $\lambda/\sqrt{(\log p)/n} = 1$, the estimation equation (4.2) provides approximately the right answer $\widehat{\sigma}^2 \approx 1$ for $r_0 = 1$. We solve (4.2) per replication and plot the histograms of $\widehat{\sigma}$ in Figure 7. The means and standard deviations are 0.971 ± 0.057, 1.033 ± 0.032 and 1.060 ± 0.039 respectively from the left to the right in Figure 7. Thus, the MSE for $\widehat{\sigma}$ is of the same order as $n^{-1/2}$ in these simulations.

4.2. The estimation of MSE and degrees of freedom. The formulas derived here are based on Stein's (1981) theorem for the unbiased estimation of the MSE of almost differentiable estimators of a mean vector. A map $\boldsymbol{h} : \mathbb{R}^p \to \mathbb{R}^p$ is almost differentiable if

$$\boldsymbol{h}(\boldsymbol{z}+\boldsymbol{v}) = \boldsymbol{h}(\boldsymbol{z}) + \Big\{ \int_0^1 \boldsymbol{H}(\boldsymbol{z}+x\boldsymbol{v})dx \Big\}\boldsymbol{v},\ \forall\, \boldsymbol{v} \in \mathbb{R}^p, \tag{4.3}$$

for certain map $\boldsymbol{H} : \mathbb{R}^p \to \mathbb{R}^{p\times p}$. Suppose in this subsection that $\rho(t;\lambda)$ is almost twice differentiable in $t > 0$, or equivalently $\dot{\rho}(t;\lambda) \equiv (\partial/\partial t)\rho(t;\lambda)$ is almost differentiable with

$$\dot{\rho}(t;\lambda) \equiv \frac{\partial}{\partial t}\rho(t;\lambda) = \dot{\rho}(1;\lambda) + \int_1^t \ddot{\rho}(x;\lambda)dx, \forall\ t > 0, \tag{4.4}$$

for certain function $\ddot{\rho}(x;\lambda)$. Under this condition, $\ddot{\rho}(t;\lambda) = (\partial/\partial t)\dot{\rho}(t;\lambda)$ almost everywhere in $(0,\infty)$. Since (3.3) is the minimum of the left- and right-derivatives, the $\ddot{\rho}(x;\lambda)$ in (3.3) and (4.4) are identical almost everywhere whenever (4.4) holds.

For multivariate normal vectors $\boldsymbol{z} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, Stein's theorem can be stated as

$$E\boldsymbol{h}(\boldsymbol{z})(\boldsymbol{z}-\boldsymbol{\mu})' = E\boldsymbol{H}(\boldsymbol{z})\boldsymbol{V}, \tag{4.5}$$

provided (4.3) and the integrability of all the elements of $\boldsymbol{H}(\boldsymbol{z})$. This can be applied to the penalized LSE (1.2). Let $\boldsymbol{\Sigma}_A$ be as in (2.4). We extend (3.11) to general penalty functions $\rho(t;\lambda)$ as follows:

$$\boldsymbol{Q}(\boldsymbol{\beta};\lambda) \equiv \boldsymbol{\Sigma}_{\{j:\beta_j\neq 0\}} + \text{diag}\Big(\ddot{\rho}(|\beta_j|;\lambda), \beta_j \neq 0\Big), \quad d(\boldsymbol{\beta}) \equiv \#\{j : \beta_j \neq 0\}. \tag{4.6}$$

Theorem 3. *Let $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(\lambda)$ be the penalized LSE in (1.2) for a fixed $\lambda > 0$. Let $\boldsymbol{\Sigma} \equiv \boldsymbol{X}'\boldsymbol{X}/n$ be as in (2.3) and $\widehat{\boldsymbol{P}}$ be the $d(\widehat{\boldsymbol{\beta}}) \times p$ matrix giving the projection $\widehat{\boldsymbol{P}}\boldsymbol{b} = (b_j : \widehat{\beta}_j \neq 0)'$ as in (3.12). Suppose $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I}_n)$ in (1.3) and (2.5) holds with $d^* = p$. Then,*

$$E\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big)\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big)' = E\Big\{\big(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\big)\big(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\big)' + \frac{2\sigma^2}{n}\widehat{\boldsymbol{P}}'\boldsymbol{Q}^{-1}(\widehat{\boldsymbol{\beta}};\lambda)\widehat{\boldsymbol{P}}\Big\} - \frac{\sigma^2}{n}\boldsymbol{\Sigma}^{-1}, \tag{4.7}$$

where $\widetilde{\boldsymbol{\beta}} \equiv \boldsymbol{\Sigma}^{-1}\boldsymbol{X}'\boldsymbol{y}/n$ is the ordinary LSE of $\boldsymbol{\beta}$. In particular, for all $\boldsymbol{a} \in \mathbb{R}^p$,

$$\big|\boldsymbol{a}'\big(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\big)\big|^2 + \frac{2\widehat{\sigma}^2}{n}(\widehat{\boldsymbol{P}}\boldsymbol{a})'\boldsymbol{Q}^{-1}(\widehat{\boldsymbol{\beta}};\lambda)(\widehat{\boldsymbol{P}}\boldsymbol{a}) - \frac{\widehat{\sigma}^2}{n}\boldsymbol{a}'\boldsymbol{\Sigma}^{-1}\boldsymbol{a} \tag{4.8}$$

is an unbiased estimator of the MSE $E\big|\boldsymbol{a}'\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big)\big|^2$, provided $\widehat{\sigma}^2 = \sigma^2$ in the case of known σ^2 or $\widehat{\sigma}^2 = \|\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}\|^2/(n-p)$ in the case of $p < n$. Consequently,

$$E\Big\{\|\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\|^2 + \frac{2\widehat{\sigma}^2}{n}\text{trace}\Big(\boldsymbol{Q}^{-1}(\widehat{\boldsymbol{\beta}};\lambda)\Big) - \frac{\widehat{\sigma}^2}{n}\text{trace}\Big(\boldsymbol{\Sigma}^{-1}\Big)\Big\} = E\big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big\|^2. \tag{4.9}$$

Remark. Condition (2.5) with $d^* = p$ asserts $c_{\min}(\boldsymbol{\Sigma}) > \kappa(\rho;\lambda)$, which is slightly stronger than the global convexity condition (2.3). We prove in Subsection 4.3 that (2.3) is a necessary and sufficient condition for the continuity of $\widehat{\boldsymbol{\beta}}$, which is weaker than the almost differentiability of $\widehat{\boldsymbol{\beta}}$. Thus, the conditions of Theorem 3 are nearly sharp for the application of the SURE. In the k-th segment of the PLUS path (2.1), $\boldsymbol{Q}(\widehat{\boldsymbol{\beta}}(\lambda);\lambda) = \boldsymbol{Q}(\boldsymbol{\eta}^{(k)})$ as in (3.11).

Let $\boldsymbol{\mu} \equiv E\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ with the penalized LSE (1.2). Let $\widetilde{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\mu}}^o$ be the projections of $\boldsymbol{y}$ to the linear spans of $\{\boldsymbol{x}_j, j \le p\}$ and $\{\boldsymbol{x}_j, \beta_j \ne 0\}$ respectively. For uncorrelated errors with common variance σ^2, the degrees of freedom for $\widehat{\boldsymbol{\mu}}^o$ is $\sum_{j=1}^p \mathrm{Cov}(\widetilde{\mu}_j, \widehat{\mu}_j^o)/\sigma^2 = \mathrm{rank}\big(\boldsymbol{x}_j : \beta_j \ne 0\big)$. Thus, since $E\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \sigma^2 \mathrm{rank}(\boldsymbol{X})$ and

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 + \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}\|^2 = 2(\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}),$$

the notion of "degrees of freedom" is extended to $\widehat{\boldsymbol{\mu}}$ as

$$\mathrm{df}(\widehat{\boldsymbol{\mu}}) \equiv \sum_{j=1}^p \frac{\mathrm{Cov}(\widetilde{\mu}_j, \widehat{\mu}_j)}{\sigma^2} = \frac{1}{2} E\Big(\mathrm{rank}(\boldsymbol{X}) - \frac{\|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}\|^2}{\sigma^2} + \frac{\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2}{\sigma^2}\Big). \tag{4.10}$$

This also provides the C_p-type risk estimate

$$\widehat{C}_p \equiv \widehat{C}_p(\lambda) \equiv \|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}\|^2 + \widehat{\sigma}^2\big\{2\,\widehat{\mathrm{df}} - \mathrm{rank}(\boldsymbol{X})\big\} \approx \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2. \tag{4.11}$$

Theorem 3 suggests the unbiased estimator for the degrees of freedom (4.10) as

$$\widehat{\mathrm{df}} \equiv \widehat{\mathrm{df}}(\lambda) \equiv \mathrm{trace}\Big(\boldsymbol{Q}^{-1}(\widehat{\boldsymbol{\beta}};\lambda)\widehat{\boldsymbol{P}}\boldsymbol{\Sigma}\widehat{\boldsymbol{P}}'\Big) \tag{4.12}$$

and the related C_p-type estimator of the MSE $E\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ via (4.11). We refer to Efron (1986) and Meyer and Woodroofe (2000) for more discussions about (4.10) and (4.11).

We summarize in Figure 8 the performance of (4.11) for the MC+ in Experiments 4 and 5, with the $\widehat{\mathrm{df}}$ in (4.12) and the $\widehat{\sigma}$ in (4.2). For each of the three settings, $E\|\widehat{\boldsymbol{\mu}}(\lambda) - \boldsymbol{\mu}\|^2$ and $E\widehat{C}_p(\lambda)$ are approximated by the averages in 100 replications and the expected conditional variance $E\mathrm{Var}(\widehat{C}_p(\lambda)|\boldsymbol{X},\boldsymbol{\beta})$ is approximated by the within-group variance, since $(\boldsymbol{X},\boldsymbol{\beta})$ is unchanged in every 10 replications in each of the three settings. From Figure 8, we observe that the MSE $E\|\widehat{\boldsymbol{\mu}}(\lambda) - \boldsymbol{\mu}\|^2$ is reasonably approximated by $\widehat{C}_p(\lambda)$ for $p > n$, at least before the MC+ starts to over fit with small λ. The following theorem asserts the unbiasedness of (4.11).

Theorem 4. *Under the conditions of Theorem 3, (4.12) is unbiased for (4.10):*

$$E\big(\widehat{\mathrm{df}}\big) = \mathrm{df}(\widehat{\boldsymbol{\mu}}). \tag{4.13}$$

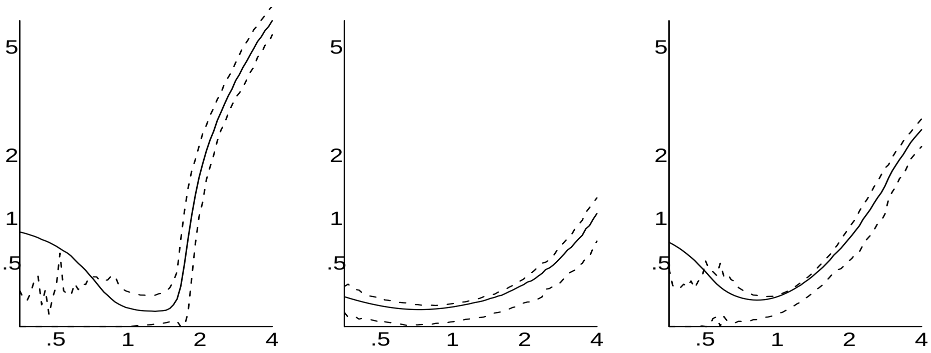

Figure 8: *Approximations of* $E\|\widehat{\boldsymbol{\mu}}(\lambda)-\boldsymbol{\mu}\|^2/n$ *(solid) and* $E\widehat{C}_p(\lambda)/n \pm 2\big\{E\,Var\big(\widehat{C}_p(\lambda)/n\big|\boldsymbol{X},\boldsymbol{\beta}\big)\big\}^{1/2}$ *(dashed) as functions of* $\lambda/\sqrt{(\log p)/n}$ *for the MC+ based on the same simulations as in Figure 6.*

Consequently, (4.11) is an unbiased estimator of the risk $\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2$ *with the* $\widehat{\sigma}^2$ *in (4.8). Furthermore, if* $\rho(t;\lambda)=\lambda t$ *for the LASSO or* $|\widehat{\beta}_j|>\gamma\lambda$ *for all* $\widehat{\beta}_j\neq 0$ *under (1.9), then*

$$\widehat{\mathrm{df}} = \#\big\{j:\widehat{\beta}_j\neq 0\big\}. \tag{4.14}$$

Under a positive cone condition on $\boldsymbol{X}$, Efron *et al* (2004) proved the unbiasedness of $\#\big\{j:\widehat{\beta}_j\neq 0\big\}$ as an estimator for the degrees of freedom for the LARS estimator (not the LASSO) at a fixed k. Our definition of the degrees of freedom and C_p is slightly different, since we use $\|\widetilde{\boldsymbol{\mu}}-\widehat{\boldsymbol{\mu}}\|^2$ and $\mathrm{rank}(\boldsymbol{X})$ in (4.10) and (4.11) for variance reduction, instead of $\|\boldsymbol{y}-\widehat{\boldsymbol{\mu}}\|^2$ and n. We prove $E\#\big\{j:\widehat{\beta}_j\neq 0\big\}=\mathrm{df}(\widehat{\boldsymbol{\mu}})$ for the LASSO for fixed λ, not for fixed k with a stochastic λ. We defer the proofs of Theorems 3 and 4 to Subsection 4.4 as we first need to prove the continuity of the penalized LSE. The performance of (4.11) for the LASSO is similar to that of the MC+ as reported in Figure 8. Figure 9 compares the simulated MSE $E\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2/n$ between the MC+ and LASSO in Experiments 4 and 5.

4.3. Continuity and convexity. The continuity of $\widehat{\boldsymbol{\beta}}$, demanded by Stein (1981), is a property of independent interest on its own right for robust estimation. Here we prove the equivalence of the continuity of the penalized LSE and the global convexity condition for full rank designs. We have considered (1.12) for unbiased selection and the slightly stronger (1.9). For the continuity of (1.2), we only need

$$\lim_{t\to\infty}\rho(t;\lambda)/t^2=0,\quad 0\le\dot{\rho}(0+;\lambda)<\infty. \tag{4.15}$$

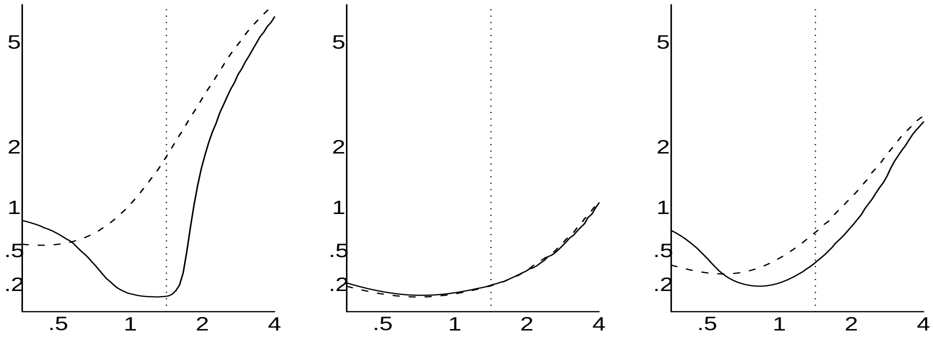

Figure 9: *Average of $\|\widehat{\boldsymbol{\mu}}(\lambda) - \boldsymbol{\mu}\|^2/n$ for the MC+ (solid) and LASSO (dashed) as functions of $\lambda/\sqrt{(\log p)/n}$, with dotted verticals at $\lambda/\sqrt{(\log p)/n} = \sqrt{2}$, based on the same simulations as in Figure 6.*

Theorem 5. *Let λ be fixed. Suppose that $\rho(t;\lambda)$ is a continuously differentiable function of t satisfying (4.15) and that $\boldsymbol{X}$ is of rank p. Then, the following three statements are equivalent to each other:*

(i) *The global minimizer of (1.2) is continuous in $\boldsymbol{y} \in \mathbb{R}^n$;*

(ii) *The global convexity condition (2.3) holds;*

(iii) *The penalized loss (1.11) is strictly convex in $\boldsymbol{\beta} \in \mathbb{R}^p$.*

Proof. Since (ii) ⇒ (iii) has been done in the proof of Theorem 2 (iii), we have two steps.

(iii) ⇒ (i): Since the penalized loss is $\|\boldsymbol{y}\|^2/(2n)$ for $\boldsymbol{\beta} = 0$, $\boldsymbol{y} \to \widehat{\boldsymbol{\beta}}$ maps bounded sets of $\boldsymbol{y}$ in $\mathbb{R}^n$ to bounded sets of $\widehat{\boldsymbol{\beta}}$ in $\mathbb{R}^p$. Since the penalized loss $L(\boldsymbol{\beta};\lambda)$ is continuous in both $\boldsymbol{y}$ and $\boldsymbol{\beta}$ and strictly convex in $\boldsymbol{\beta}$ for each $\boldsymbol{y}$, its global minimum is unique and continuous in $\boldsymbol{y}$.

(i) ⇒ (ii): Since $\widehat{\boldsymbol{\beta}}$ depends on $\boldsymbol{y}$ only through $\widetilde{\boldsymbol{z}} = \boldsymbol{X}'\boldsymbol{y}/n$ and $\boldsymbol{X}$ is of rank p, the map $\widetilde{\boldsymbol{z}} \to \widehat{\boldsymbol{\beta}}$ is continuous from $\mathbb{R}^p$ to its range $\mathscr{I}$. Since $\widehat{\boldsymbol{\beta}}$ is the global minimum, (2.2) must hold and the inverse

$$\widehat{\boldsymbol{\beta}} \to \widetilde{\boldsymbol{z}} = \boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}} + \mathrm{sgn}(\widehat{\boldsymbol{\beta}})\dot{\rho}(|\widehat{\boldsymbol{\beta}}|;\lambda)$$

is continuous for $\widehat{\boldsymbol{\beta}} \in (0,\infty)^p \cap \mathscr{I}$, with per component application of functions and the product operation. It follows that $(0,\infty)^p \cap \mathscr{I}$ is open and does not have a boundary point in $(0,\infty)^p$. Let $\mathbf{1} \equiv (1,\ldots,1)' \in \mathbb{R}^p$. For $\widetilde{\boldsymbol{z}} = x\boldsymbol{\Sigma}\mathbf{1}$ with $x > 0$, $L(x\mathbf{1};\lambda) = o(x^2)$ for the ordinary LSE $x\mathbf{1}$ by the first condition of (4.15), and $L(\boldsymbol{\beta};\lambda)$ is at least $c_{\min}(\boldsymbol{\Sigma})x^2$ for any $\boldsymbol{\beta}$ outside $(0,\infty)^p$. Thus, $(0,\infty)^p \cap \mathscr{I}$ is not empty. As the only nonempty set without any boundary point in $(0,\infty)^p$, $(0,\infty)^p \cap \mathscr{I} = (0,\infty)^p$. Moreover, the map $\widetilde{\boldsymbol{z}} \to \widehat{\boldsymbol{\beta}}$ is one-to-one for $\widehat{\boldsymbol{\beta}} \in (0,\infty)^p$.

We have proved that all points $\boldsymbol{\beta}$ in $(0,\infty)^p$ are unique global minimum of (1.11) for some $\boldsymbol{z} \in \mathbb{R}^p$. Let $\widehat{\boldsymbol{\beta}} = x\mathbf{1} \in (0,\infty)^p$ and $\boldsymbol{b}$ be the eigenvector with $\boldsymbol{\Sigma}\boldsymbol{b} = c_{\min}(\boldsymbol{\Sigma})\boldsymbol{b}$ and $\|\boldsymbol{b}\| = 1$. The quantity

$$\begin{aligned} t^{-1}\frac{\partial}{\partial t}L(\widehat{\boldsymbol{\beta}}+t\boldsymbol{b};\lambda) &= \big\|\boldsymbol{X}\boldsymbol{b}\big\|^2 + \sum_{j=1}^p t^{-1}\mathrm{sgn}(\widehat{\beta}_j)b_j\Big\{\dot\rho(|\widehat{\beta}_j+tb_j|;\lambda)-\dot\rho(|\widehat{\beta}_j|;\lambda)\Big\} \\ &= c_{\min}(\boldsymbol{\Sigma}) + \sum_{j=1}^p t^{-1}b_j\Big\{\dot\rho(x+tb_j;\lambda)-\dot\rho(x;\lambda)\Big\} \qquad (4.16) \end{aligned}$$

must have nonnegative lower limit as $t \to 0+$. Integrating over $x \in [t_1,t_2]$ and then taking the limit, we find

$$c_{\min}(\boldsymbol{\Sigma})(t_2-t_1)+\dot\rho(t_2;\lambda)-\dot\rho(t_1;\lambda) = \lim_{t\to 0+}\int_{t_1}^{t_2} t^{-1}\frac{\partial}{\partial t}L(x\mathbf{1}+t\boldsymbol{b};\lambda)dx \ge 0. \qquad (4.17)$$

It remains to prove that (4.17) holds with strict inequality. If (4.17) holds with equality for certain $0 < t_1 < t_2$, then for $t_1 < x < t_2$ and small t (4.16) becomes

$$t^{-1}\frac{\partial}{\partial t}L(\widehat{\boldsymbol{\beta}}+t\boldsymbol{b};\lambda) = c_{\min}(\boldsymbol{\Sigma}) + \sum_{j=1}^p t^{-1}b_j\Big\{-c_{\min}(\boldsymbol{\Sigma})tb_j\Big\} = 0.$$

This is contradictory to the uniqueness of $\widehat{\boldsymbol{\beta}}$. □

4.4. Almost differentiability and the proofs of Theorems 3 and 4. The proofs of Theorems 3 and 4 requires the following proposition, which provides the almost differentiability of $\widehat{\boldsymbol{\beta}}$. In fact, we prove the stronger Liptchitz condition for $\widehat{\boldsymbol{\beta}}$ under the conditions of Theorem 3.

Proposition 2. *Let λ and $\boldsymbol{X}$ be fixed and treat $\widehat{\boldsymbol{\beta}}$ as a function of $\boldsymbol{y}$. Suppose (2.5) holds with $d^* = p$. Then, $\widehat{\boldsymbol{\beta}} = \boldsymbol{h}(\widetilde{\boldsymbol{z}})$ for $\widetilde{\boldsymbol{z}} = \boldsymbol{X}'\boldsymbol{y}/n \in \mathbb{R}^p$ and certain almost differentiable function $\boldsymbol{h} : \mathbb{R}^p \to \mathbb{R}^p$, such that for all $\boldsymbol{z}$ and $\boldsymbol{v}$ in $\mathbb{R}^p$*

$$\boldsymbol{h}(\boldsymbol{z}+\boldsymbol{v}) = \boldsymbol{h}(\boldsymbol{z}) + \Big\{\int_0^1 \big(\boldsymbol{P}'\boldsymbol{Q}^{-1}\boldsymbol{P}\big)(\boldsymbol{h}(\boldsymbol{z}+x\boldsymbol{v});\lambda)dx\Big\}\boldsymbol{v}, \qquad (4.18)$$

where $\boldsymbol{Q}$ is as in (4.6) and $\boldsymbol{P}(\boldsymbol{\beta};\lambda) : \boldsymbol{b} \to (b_j : \beta_j \neq 0)'$ is the projection as in (3.12) and Theorem 3. Moreover, $c_{\min}\big(\boldsymbol{Q}(\boldsymbol{b};\lambda)\big) \ge c_{\min}(\boldsymbol{\Sigma}) - \kappa(\rho;\lambda) > 0$ for all $0 \neq \boldsymbol{b} \in \mathbb{R}^p$, so that $\boldsymbol{h}(\boldsymbol{z})$ satisfies the Lipschitz condition: for all $\boldsymbol{z}$ and $\boldsymbol{v}$ in $\mathbb{R}^p$

$$\Big\|\boldsymbol{h}(\boldsymbol{z}+\boldsymbol{v})-\boldsymbol{h}(\boldsymbol{z})\Big\| \le \frac{\|\widehat{\boldsymbol{P}}\boldsymbol{v}\|}{c_{\min}(\boldsymbol{\Sigma})-\kappa(\rho;\lambda)} \le \frac{\|\boldsymbol{v}\|}{c_{\min}(\boldsymbol{\Sigma})-\kappa(\rho;\lambda)}. \qquad (4.19)$$

Proof of Proposition 2. Let $\widehat{\boldsymbol{P}}$ be as in Theorem 3. We write (2.2) as

$$\begin{cases} \widehat{\boldsymbol{P}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{P}}\mathrm{sgn}(\widehat{\boldsymbol{\beta}})\dot{\rho}(|\widehat{\boldsymbol{\beta}}|;\lambda) = \widehat{\boldsymbol{P}}\widetilde{\boldsymbol{z}} \\ \big|\widetilde{z}_j - \boldsymbol{x}_j'\boldsymbol{X}\widehat{\boldsymbol{\beta}}/n\big| \le \lambda \end{cases} \tag{4.20}$$

with per component application of univariate functions and operations. Let $\boldsymbol{\delta} \in \{-1,0,1\}^p$ be fixed. It follows from Theorem 5 that the map $\widehat{\boldsymbol{P}}\widetilde{\boldsymbol{z}} \to \widehat{\boldsymbol{P}}\widehat{\boldsymbol{\beta}}$ is continuous in $\widetilde{\boldsymbol{z}} \in \mathbb{R}^p$ and continuously invertible given a fixed $\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\delta}$. Let $H(\boldsymbol{\delta}) \equiv \{\widetilde{\boldsymbol{z}} : \mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\delta}\}$. The boundary of $H(\boldsymbol{\delta})$ has zero Lebesgue measure, since it is contained in the set of $\widetilde{\boldsymbol{z}}$ satisfying $\delta_j\widehat{\beta}_j = 0+$ for $\delta_j \neq 0$ or $\widetilde{z}_j - \boldsymbol{x}_j'\boldsymbol{X}\widehat{\boldsymbol{\beta}}/n = \pm\lambda$ for $\delta_j = 0$, $j = 1,\ldots,p$, according to (4.20). In the interior of $H(\boldsymbol{\delta})$, (4.20) gives $(\partial/\partial\widetilde{z}_j)\widehat{\boldsymbol{\beta}} = 0$ and $(\partial/\partial\widetilde{\boldsymbol{z}})\widehat{\beta}_j = 0$ for $\delta_j = 0$ and

$$\widehat{\boldsymbol{P}}\frac{\partial}{\partial\widehat{\boldsymbol{\beta}}}\Big(\widehat{\boldsymbol{P}}\widetilde{\boldsymbol{z}}\Big)' = \widehat{\boldsymbol{P}}\boldsymbol{\Sigma}\widehat{\boldsymbol{P}} + \widehat{\boldsymbol{P}}\mathrm{diag}\big(\ddot{\rho}(|\widehat{\beta}_j|;\lambda)\big)\widehat{\boldsymbol{P}}' = \boldsymbol{Q}(\widehat{\boldsymbol{\beta}};\lambda).$$

Since (2.5) holds with $d^* = p$, $c_{\min}\big(\boldsymbol{Q}(\boldsymbol{\beta};\lambda)\big) \ge c_{\min}(\boldsymbol{\Sigma}) - \kappa(\rho;\lambda) > 0$ for all $\boldsymbol{\beta} \neq 0$. Thus, the differentiation of the inverse map yields $(\partial/\partial\widetilde{\boldsymbol{z}})\widehat{\boldsymbol{\beta}}' = \widehat{\boldsymbol{P}}'\boldsymbol{Q}^{-1}(\widehat{\boldsymbol{\beta}};\lambda)\widehat{\boldsymbol{P}}$. □

Proof of Theorem 3. It follows from Proposition 2 that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{z}}$ is almost differentiable in $\widetilde{\boldsymbol{z}}$ with derivative

$$\frac{\partial}{\partial\widetilde{\boldsymbol{z}}}\Big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{z}}\Big)' = \widehat{\boldsymbol{P}}'\boldsymbol{Q}^{-1}(\widehat{\boldsymbol{\beta}};\lambda)\widehat{\boldsymbol{P}} - \boldsymbol{\Sigma}^{-1}.$$

Since $\widetilde{\boldsymbol{z}} \equiv \boldsymbol{X}'\boldsymbol{y}/n \sim N(\boldsymbol{\Sigma\beta}, \boldsymbol{\Sigma}\sigma^2/n)$, this and (4.5) imply

$$\begin{aligned} E\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{z}}\big)\big(\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{z}} - \boldsymbol{\beta}\big)' &= E\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{z}}\big)\big(\widetilde{\boldsymbol{z}} - \boldsymbol{\Sigma\beta}\big)'\boldsymbol{\Sigma}^{-1} \\ &= \frac{\sigma^2}{n}\Big\{E\widehat{\boldsymbol{P}}'\boldsymbol{Q}^{-1}(\widehat{\boldsymbol{\beta}};\lambda)\widehat{\boldsymbol{P}} - \boldsymbol{\Sigma}^{-1}\Big\}. \end{aligned}$$

Since the ordinary LSE is $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{z}} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}^{-1}\sigma^2/n)$, it follows that

$$\begin{aligned} & E\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)' \\ = \; & E\big(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\big)\big(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\big)' - E\big(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\big)\big(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\big)' + 2E\big(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\big)\big(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)' \\ = \; & E\big(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\big)\big(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\big)' + \frac{2\sigma^2}{n}\Big\{E\widehat{\boldsymbol{P}}'\boldsymbol{Q}^{-1}(\widehat{\boldsymbol{\beta}};\lambda)\widehat{\boldsymbol{P}} - \boldsymbol{\Sigma}^{-1}\Big\} + \frac{\sigma^2}{n}\boldsymbol{\Sigma}^{-1}. \end{aligned}$$

This proves (4.7). The rest of the theorem follows immediately. □

Proof of Theorem 4. Since $\mathrm{trace}(\boldsymbol{b}\boldsymbol{b}') = \|\boldsymbol{b}\|^2$, (4.7) gives

$$E\big\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\big\|^2 = E\Big\{\big\|\widehat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{\mu}}\big\|^2 + \frac{2\sigma^2}{n}\mathrm{trace}\Big(\boldsymbol{X}\widehat{\boldsymbol{P}}'\boldsymbol{Q}^{-1}(\widehat{\boldsymbol{\beta}};\lambda)\widehat{\boldsymbol{P}}\boldsymbol{X}'\Big)\Big\} - \frac{\sigma^2}{n}\mathrm{trace}\Big(\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}'\Big)$$

$$= \quad E\Big\{\big\|\widehat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{\mu}}\big\|^2 + 2\sigma^2\widehat{\mathrm{df}} - \sigma^2\mathrm{rank}(\boldsymbol{X})\Big\},$$

which implies (4.13) via (4.10). For (4.14), we observe that $\boldsymbol{Q}(\widehat{\boldsymbol{\beta}};\lambda) = \widehat{\boldsymbol{P}}\boldsymbol{\Sigma}\widehat{\boldsymbol{P}}'$ by (4.6) when $\ddot{\rho}(|\widehat{\beta}_j|;\lambda) = 0$ for all $\widehat{\beta}_j \neq 0$. □

5. Variable selection consistency. In this section, we provide two lower bounds for the probability of correct selection $P\{\widehat{A} = A^o\}$ for the sparsest PLUS solution: one for $p \le n$ under the global convexity condition and one for general p under the sparse Riesz condition. These lower bounds imply the selection consistency (1.5) as $\max(n,p) \to \infty$. In fact, we prove the stronger sign consistency in the sense of

$$P\Big\{\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta})\Big\} \to 1, \tag{5.1}$$

where sgn(x) is the sign of x with the convention sgn$(0) = 0$ and per component application on vectors. An analytic upper bound for the dimension $\#\{j : \widehat{\beta}_j(\lambda) \neq 0\}$ of the PLUS selection is also provided.

5.1. Consistency and non-asymptotic probability bounds. Our consistency results are proved by showing that the sparsest PLUS solution is identical to an oracle LSE with high probability. Let $\boldsymbol{X}_A$ and $\boldsymbol{\Sigma}_A \equiv \boldsymbol{X}'_A\boldsymbol{X}_A/n$ be as in (2.4). Given the knowledge of the pattern A^o in (1.1), the oracle LSE $\widehat{\boldsymbol{\beta}}^o \equiv (\widehat{\beta}^o_1, \ldots, \widehat{\beta}^o_p)'$ is given by

$$\big(\widehat{\beta}^o_j, j \in A^o\big)' = \boldsymbol{\Sigma}^{-1}_{A^o}\boldsymbol{X}'_{A^o}\boldsymbol{y}/n, \quad \big(\widehat{\beta}^o_j, j \notin A^o\big)' = 0, \tag{5.2}$$

provided rank$(\boldsymbol{X}_{A^o}) = d^o$ with the d^o in (1.6). Let

$$(w^o_j, j \in A^o)' = \text{the diagonal elements of } \boldsymbol{\Sigma}^{-1}_{A^o}, \tag{5.3}$$

so that Var$(\widehat{\beta}^o_j) = w^o_j\sigma^2/n, j \in A^o$. We first present non-asymptotic bounds for $p \le n$.

Theorem 6. *Let $\lambda > 0$ be fixed and $\widehat{\boldsymbol{\beta}}$ be the penalized LSE in (1.2) with a penalty $\rho(t;\lambda)$ satisfying (1.9). Suppose (2.3) holds and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I}_n)$ in (1.3). Let A^o, d^o, $\widehat{\boldsymbol{\beta}}^o$, β_* and $\Phi(\cdot)$ be as in (1.1), (1.6), (5.2), (1.10) and (2.7) respectively. Suppose $\beta_* \ge \gamma\lambda$. Then,*

$$\begin{aligned} P\Big\{\widehat{A} \neq A^o\Big\} &\le P\Big\{\widehat{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}^o \text{ or } \mathrm{sgn}(\widehat{\boldsymbol{\beta}}) \neq \mathrm{sgn}(\boldsymbol{\beta})\Big\} \\ &\le \sum_{j \in A^o} \Phi\Big(\frac{\gamma\lambda - |\beta_j|}{\sigma(w^o_j/n)^{1/2}}\Big) + 2\sum_{j \notin A^o} \Phi\Big(-\frac{n\lambda}{\sigma\|\boldsymbol{x}_j\|}\Big). \end{aligned} \tag{5.4}$$

In particular, if $\|\boldsymbol{x}_j\|^2/n = 1$ *and* $|\beta_j| \ge \gamma\lambda + \sqrt{w_j^o}\lambda$ *with* $\lambda = \sigma\sqrt{2(1+\epsilon_n)(\log p)/n}$, *then*

$$\begin{aligned} P\Big\{\widehat{\boldsymbol{\beta}} \ne \widehat{\boldsymbol{\beta}}^o \text{ or } \mathrm{sgn}(\widehat{\boldsymbol{\beta}}) \ne \mathrm{sgn}(\boldsymbol{\beta})\Big\} &\le (2p - d^o)\Phi\big(-\sqrt{n}\lambda/\sigma\big) \\ &\le \frac{1}{p^{\epsilon_n}(\pi(1+\epsilon_n)\log p)^{1/2}}, \quad \forall \epsilon_n > -1. \end{aligned} \tag{5.5}$$

It follows from Theorem 2 (iii) that the solution of (2.2) is unique under (2.3), so that Theorem 6 is applicable to the PLUS solution. For the MC+, (5.3) and (2.3) implies $w_j^o \le c_{\max}(\boldsymbol{\Sigma}_{A^o}^{-1}) \le 1/c_{\min}(\boldsymbol{\Sigma}) < \gamma$, so that (5.5) implies (2.7) for $\beta_* \ge (\gamma + \sqrt{\gamma})\lambda$. For the SCAD, we need $\gamma > 1 + 1/c_{\min}(\boldsymbol{\Sigma})$ for (2.7). For $d^o \ll p$, the right-hand side of (5.4) is small for $\lambda \ge \sqrt{2\log p}\max_j \|\boldsymbol{x}_j\|/n$ and $\beta_* \approx \gamma\lambda$. Thus, Theorem 6 provides theoretical support to the heuristic condition (1.10) for selection consistency.

We now consider selection consistency for general p, including $p \gg n$. For positive constants $\{c_*, c^*\}$ and differentiable ρ, define

$$C\big(c_*, c^*, \rho\big) \equiv \max_{0\le w\le 1}\inf_{t\ge 0}\Big\{w\dot\rho^2(t)/c^* + c_*(1-w)t^2\Big\}. \tag{5.6}$$

For quadratic spline penalty (3.1) and $c^* \ge 1$ in (2.6), $C(c_*, c^*, \rho) \le \dot\rho^2(0+) = 1$. Set

$$M_1 \equiv M_1\big(c_*, c^*, \rho\big) \equiv 4/\sqrt{C(c_*, c^*, \rho)}, \quad M_2 \equiv M_2\big(c_*, c^*, \rho\big) \equiv 2 + M_1^2/c_*. \tag{5.7}$$

Theorem 7. *For each* λ *let* $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(\lambda)$ *be the sparsest PLUS solution of (2.2) with a quadratic spline penalty* $\rho(t;\lambda) = \lambda^2\rho_m(t/\lambda)$ *with* $t_m = \gamma < \infty$ *in (3.1). Let* A^o, d^o, $\widehat{\boldsymbol{\beta}}^o$, β_* *and* $\Phi(\cdot)$ *be as in Theorem 6. Suppose the sparse convexity condition (2.5) holds with* $M_2 d^o + 1 \le d^*$ *and* $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I}_n)$ *in (1.3). Let* $a_{n,p}$ *be positive constants and define*

$$\lambda_{n,p} \equiv M_1\sigma\sqrt{(1 + a_{n,p} + 2\log p)/n}.$$

Suppse $\beta_* \ge \gamma\lambda_{n,p}$. *Then, for* $\lambda \le \lambda_{n,p}$ *and* $\epsilon_{n,p} = e^{-a_{n,p}/2}\sqrt{1 + a_{n,p} + 2\log p}$,

$$\begin{aligned} P\Big\{\widehat{A} \ne A^o\Big\} &\le P\Big\{\widehat{\boldsymbol{\beta}} \ne \widehat{\boldsymbol{\beta}}^o \text{ or } \mathrm{sgn}(\widehat{\boldsymbol{\beta}}) \ne \mathrm{sgn}(\boldsymbol{\beta})\Big\} \\ &\le \sum_{j\in A^o}\Phi\Big(\frac{\gamma\lambda_{n,p} - |\beta_j|}{\sigma(w_j^o/n)^{1/2}}\Big) + 2\sum_{j\notin A^o}\Phi\Big(-\frac{n\lambda}{\sigma\|\boldsymbol{x}_j\|}\Big) + e^{\epsilon_{n,p}} - 1. \end{aligned} \tag{5.8}$$

In particular, if $\|\boldsymbol{x}_j\|^2/n = 1$, $|\beta_j| \ge \gamma\lambda_{n,p} + \sqrt{w_j^o}\lambda$ *with* $\lambda = \sigma\sqrt{2(1+\epsilon_n)(\log p)/n}$, *and* $a_{n,p} = 2\{\epsilon_n\log p + \log((1+\epsilon_n)\log p)\} \to \infty$ *with* $\epsilon_n \ge 0$, *then*

$$P\Big\{\widehat{\boldsymbol{\beta}} \ne \widehat{\boldsymbol{\beta}}^o \text{ or } \mathrm{sgn}(\widehat{\boldsymbol{\beta}}) \ne \mathrm{sgn}(\boldsymbol{\beta})\Big\} \le \frac{\pi^{-1/2} + \sqrt{2} + o(1)}{p^{\epsilon_n}((1+\epsilon_n)\log p)^{1/2}} = o(1). \tag{5.9}$$

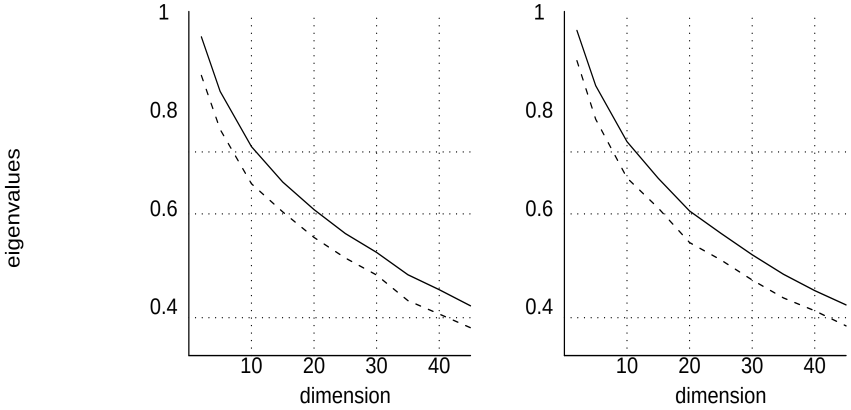

Figure 10: *The mean (solid) of the minimum eigenvalue $c_{\min}(\boldsymbol{X}'_A\boldsymbol{X}_A/n)$ for a random set A of design vectors and the mean minus two standard deviations (dashed) as functions of the dimension $|A|$, each point based on 100 replications, with horizontal dotted lines at $\kappa(\rho_2) = 1/\gamma$ for $\gamma \in \{1.4, 1.7, 2.652\}$. Left: the design $\boldsymbol{X}$ in Experiments 1 and 2. Right: the design $\boldsymbol{X}$ in Experiments 3 and 4.*

Theorems 6 and 7 immediately implies the following asymptotic result.

Theorem 8. *Under the conditions of either Theorem 6 or Theorem 7, (1.5) and (5.1) hold, provided that $\epsilon_n = 0$ for $p \to \infty$ and $\epsilon_n \to \infty$ for fixed p in (5.5) and (5.9).*

We prove Theorems 6 and 7 after providing our upper bound for the dimension of the PLUS selection in Subsection 5.2. Since (2.5) allows $c_* > c_{\min}(\boldsymbol{\Sigma})$, Theorem 7 provides greater level of concavity $\kappa(\rho; \lambda)$ for the penalty than Theorem 6 does, e.g. smaller value of γ for the MC+, and thus requires smaller separation zone β_*.

The SRC and constant factors in Theorems 6 and 7 are quite conservative compared with our simulation results. Technically this is due to the following two reasons: (a) The sparse minimum and maximum eigenvalues, or c_* and c^* respectively in (2.6), are used to bound the effects of matrix operations in the worst case scenario given the dimension/rank of the matrix; (b) The proofs do not consider the possibility that the paths of individual $b_j = \widehat{\beta}_j(\lambda)/\lambda$ pass the bias/concave zone $\{b : 0 < |b| < \gamma\}$ in different steps, which is harder to track analytically but intuitively requires sparse convexity of smaller ranks. This suggests that the penalized loss with the MCP possesses sufficient convexity if

$$P^*\Big\{c_{\min}(\boldsymbol{\Sigma}_A) \geq \kappa(\rho_2) = 1/\gamma \Big| \ |A| = d, \boldsymbol{X}\Big\} \approx 1 \tag{5.10}$$

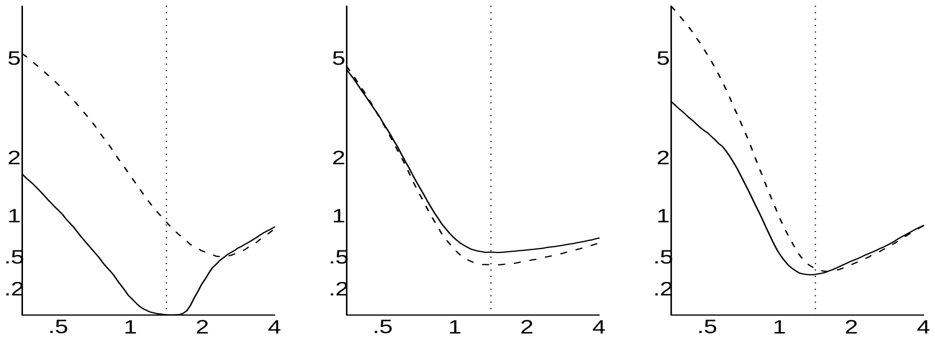

Figure 11: *The normalized total miss $\overline{TM}/d^o$ of the MC+ (solid) and LASSO (dashed) as functions of $\lambda/\sqrt{(\log p)/n}$ based on the same simulations as in Figure 6.*

at a reasonable dimension d, where P^* is the probability under which A is a random subset of $\{1, \ldots, p\}$. In practice, we may substitute the SRC (2.6) with (5.10) and a similar probabilistic upper bound on $c_{\max}(\mathbf{\Sigma}_A)$ under P^*, which are weaker and easy to check. Figure 10 plots the mean and a lower confidence bound of $c_{\min}(\mathbf{\Sigma}_A)$ under P^* as functions of given $|A| = d$. We observe that (5.10) holds for quite a few possible combinations of (d, γ) in our experiments.

Theorem 7 compares favorably with the existing results for the selection consistency of the LASSO (Meinshausen and Buhlmann, 2006; Zhao and Yu, 2006), which require

$$\beta_* \geq M n^{\epsilon_0} \sqrt{d^o} \sigma \sqrt{(\log p)/n} \tag{5.11}$$

for a finite constant M and a small $\epsilon_0 > 0$ and the strong irrepresentable condition

$$\max_{j \notin A^o} \left| n^{-1} \boldsymbol{x}_j' \boldsymbol{X}_{A^o} \mathbf{\Sigma}_{A^o}^{-1} \boldsymbol{s}^o \right| < 1 - \eta_0$$

for a small $\eta_0 > 0$, where $\boldsymbol{s}^o \equiv (\mathrm{sgn}(\beta_j) : \beta_j \neq 0)'$. Zhang and Huang (2006) proved the rate consistency in variable selection for the LASSO under the SRC (2.6) and (5.11) with $\epsilon_0 = 0$. The adverse effects of the factor $\sqrt{d^o}$ in (5.11) on the LASSO selection are evident in our simulation studies.

Figure 11 plots the average of the total miss TM$= |\widehat{A} \setminus A^o| + |A^o \setminus \widehat{A}|$ of the MC+ and LASSO in Experiments 4 and 5. The MC+ performs well in Experiment 4 as the signal is strong. In Experiment 5, the performance of the MC+ is slightly worse than the LASSO for high correlation and slightly better for low correlation, but variable selection is most probably infeasible in this experiment due to weak signal. The two procedures have nearly identical

MSE for high correlation and the MC+ has significantly smaller MSE for low correlation in Experiment 5, shown in Figure 9. As exhibited in Figure 11 and Tables 1, 2 and 3, the universal penalty level $\lambda = \sigma\sqrt{2(\log p)/n}$ is nearly the optimal choice for variable selection in all our simulation experiments. This further confirms the results in Theorems 6 and 7. Other aspects of the simulation results in Experiments 4 and 5 have been reported in Figures 6, 7 and 8.

5.2. The sparsity of the selected model. In this subsection we provide an analytic upper bound on the dimension $|\widehat{A}| = \#\{j : \widehat{\beta}_j(\lambda) \neq 0\}$ of the selected model (1.4). This allows the sparse Riesz condition (2.6) to apply. Define

$$\zeta^* = \max_{1\le m\le p} \zeta^*_m, \quad \zeta^*_m \equiv \max\Big\{\frac{\|\boldsymbol{P}_1(A)\boldsymbol{\varepsilon}\|}{(mn)^{1/2}} : A \subseteq \{1,\dots p\},\ |A| = m\Big\}. \tag{5.12}$$

Theorem 9. *Let d^o, $\{M_1, M_2\}$ and ζ^* be as in (1.6), (5.7) and (5.12) respectively. Suppose the sparse Riesz condition (2.6) holds with $M_2 d^o + 1 \le d^*$. Let $\widehat{\boldsymbol{\beta}}(\lambda)$ be the PLUS path as in (2.1) and (3.15) with a quadratic penalty $\rho(t;\lambda) = \lambda^2\rho(t/\lambda)$ in (3.1). Then,*

$$\#\{j : \widehat{\beta}_j(\lambda) \neq 0 \ \textit{or}\ \beta_j \neq 0\} \le M_2 d^o$$

always holds before λ first reaches the level $\lambda = M_1\zeta^*$.

We need a lemma for the proof of Theorem 9. For $A \subseteq \{1,\dots,p\}$, define

$$\boldsymbol{P}_1(A) \equiv \text{the projection of } \mathbb{R}^n \text{ to the linear span of } \{\boldsymbol{x}_j : j \in A\}. \tag{5.13}$$

Lemma 1. *Let $d^o \equiv |A^o|$ be as in (1.6). Let $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(\lambda)$ be a solution of (2.2) and*

$$\big\{j : \widehat{\beta}_j \neq 0\big\} \cup A^o \subseteq A_1 \subseteq \big\{j : \widehat{\beta}_j \neq 0 \text{ or } |\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})| = \lambda\big\} \cup A^o. \tag{5.14}$$

Let $\boldsymbol{X}_1 = \boldsymbol{X}_{A_1}$ as in (2.4) and $\boldsymbol{\Sigma}_{11} \equiv \boldsymbol{X}_1'\boldsymbol{X}_1/n$. Then, for all $0 \le w \le 1$,

$$\frac{w\|\boldsymbol{X}_1'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|^2}{nc_{\max}(\boldsymbol{\Sigma}_{11})} + \frac{(1-w)\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{c_{\min}^{-1}(\boldsymbol{\Sigma}_{11})} \le \Big\{2\Big(\frac{\lambda^2 d^o}{c_{\min}(\boldsymbol{\Sigma}_{11})}\Big)^{1/2} + 2\frac{\|\boldsymbol{P}_1(A_1)\boldsymbol{\varepsilon}\|}{n^{1/2}}\Big\}^2. \tag{5.15}$$

Proof. Assume $c_{\min}(\boldsymbol{\Sigma}_{11}) > 0$. Set $A_2 \equiv \{1,\dots,p\} \setminus A_1$, $A_3 \equiv A^o$ and $A_4 \equiv A_1 \setminus A^o$. Define $\boldsymbol{b}_k = (b_j, j \in A_k)$, $\forall \boldsymbol{b} \in \mathbb{R}^p$. For $k = 3, 4$, let $\boldsymbol{Q}_k$ be the matrix representing the selection of variables in A_k from A_1, defined as $\boldsymbol{Q}_k\boldsymbol{b}_1 = \boldsymbol{b}_k$. Define matrices $\boldsymbol{\Sigma}_{jk} \equiv n^{-1}\boldsymbol{X}_j'\boldsymbol{X}_k$.

Since $\widehat{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_2 = 0$, the A_1 components of the gradient $\boldsymbol{g} \equiv \boldsymbol{g}(\lambda) \equiv \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda))/n$ satisfy $\boldsymbol{g}_1 = \boldsymbol{X}_1'(\boldsymbol{y} - \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1)/n = \boldsymbol{X}_1'\boldsymbol{\varepsilon}/n + \boldsymbol{\Sigma}_{11}(\boldsymbol{\beta}_1 - \widehat{\boldsymbol{\beta}}_1)$, so that

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{g}_1 + (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1'\boldsymbol{\varepsilon}/n. \tag{5.16}$$

Since $\boldsymbol{Q}_4\boldsymbol{\beta}_1 = 0$ and $\boldsymbol{g}_4'\boldsymbol{Q}_4\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{g}_4'\widehat{\boldsymbol{\beta}}_4 \geq 0$, it follows that

$$\boldsymbol{g}_4'\boldsymbol{Q}_4\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{g}_1 \leq \boldsymbol{g}_4'\boldsymbol{Q}_4\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{g}_1 + \boldsymbol{g}_4'\widehat{\boldsymbol{\beta}}_4 = \boldsymbol{g}_4'\boldsymbol{Q}_4\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1'\varepsilon/n.$$

Let $\boldsymbol{v}_1 \equiv \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{g}_1$ and $\boldsymbol{v}_k \equiv \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{Q}_k'\boldsymbol{g}_k$, $k = 3, 4$. Since $\boldsymbol{g}_4'\boldsymbol{Q}_4\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1'\varepsilon/n = \boldsymbol{v}_4'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{X}_1'\varepsilon/n \leq \|\boldsymbol{v}_4\| \cdot \|\boldsymbol{P}_1(A_1)\boldsymbol{\varepsilon}\|/\sqrt{n}$ and $\boldsymbol{v}_1 = \boldsymbol{v}_3 + \boldsymbol{v}_4$,

$$\big\|\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{g}_1\big\| = \|\boldsymbol{v}_1\| \leq 2\|\boldsymbol{v}_3\| + \frac{\boldsymbol{v}_4'\boldsymbol{v}_1}{\|\boldsymbol{v}_4\|} \leq 2\|\boldsymbol{v}_3\| + \|\boldsymbol{P}_1(A_1)\boldsymbol{\varepsilon}\|/\sqrt{n}. \tag{5.17}$$

Insert this bound to the product of $\boldsymbol{\Sigma}_{11}^{1/2}$ and (5.16), we find

$$\big\|\boldsymbol{\Sigma}_{11}^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)\big\| \leq \|\boldsymbol{v}_1\| + \big\|\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{X}_1'\boldsymbol{\varepsilon}\big\|/n \leq 2\|\boldsymbol{v}_3\| + 2\|\boldsymbol{P}_1(A_1)\boldsymbol{\varepsilon}\|/\sqrt{n}. \tag{5.18}$$

Since $c_{\min}(\boldsymbol{\Sigma}_{11})\|\boldsymbol{v}_3\|^2 \leq \|\boldsymbol{g}_3\|^2 \leq \|\boldsymbol{g}\|_\infty^2 d^o \leq \lambda^2 d^o$ by (2.2), the summation of the squares of (5.17) and (5.18) with weights $w \geq 0$ and $1 - w \geq 0$ yields (5.15). □

Proof of Theorem 9. Since $\widehat{\boldsymbol{\beta}} = 0$ in the initial section with $\lambda \geq \lambda^{(0)}$ and $M_2 \geq 1$, we only need to consider the case $\lambda^{(0)} > M_1\zeta^*$. Let w be the optimal choice in (5.6). It follows from (2.2) and (5.14) that $\beta_j = 0$ for $j \notin A_1$ and $\lambda|\dot{\rho}(\widehat{\beta}_j/\lambda)| = |\dot{\rho}(\widehat{\beta}_j;\lambda)| = |\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n|$ for $j \in A_1 \setminus A^o$, so that by (5.6)

$$\begin{aligned} C\big(c_*, c^*, \rho\big)\lambda^2\big|A_1 \setminus A^o\big| &\leq \sum_{j \in A_1 \setminus A^o} w\{\dot{\rho}(\widehat{\beta}_j;\lambda)\}^2/c^* + c_*(1-w)\widehat{\beta}_j^2 \\ &\leq w\|\boldsymbol{X}_1'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|^2/(nc^*) + c_*(1-w)\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2. \end{aligned} \tag{5.19}$$

By (2.6), $c_{\min}(\boldsymbol{\Sigma}_{11}) \geq c_*$ and $c_{\max}(\boldsymbol{\Sigma}_{11}) \leq c^*$ for $|A_1| \leq d^*$. Thus, by Lemma 1 and (5.19)

$$C\big(c_*, c^*, \rho\big)\lambda^2\big(|A_1| - d^o\big) \leq \Big\{2\Big(\frac{\lambda^2 d^o}{c_*}\Big)^{1/2} + 2\zeta^*\sqrt{|A_1|}\Big\}^2 \leq \frac{8}{c_*}\lambda^2 d^o + 8\big(\zeta^*\big)^2|A_1|$$

in the event $\{|A_1| \leq d^*\}$. For $\lambda \geq M_1\zeta^*$, this implies

$$|A_1| \leq \frac{C(c_*, c^*, \rho) + 8/c_*}{C(c_*, c^*, \rho) - 8/M_1^2} d^o = M_2 d^o.$$

Now, beginning from $\lambda = \lambda^{(0)}$, the set A_1 is allowed to change one-at-a-time in the PLUS path due to the continuity of the path and the flexibility in the choice of A_1 in (5.14), in view of (2.2). Thus, since $|A_1| \leq d^*$ implies $|A_1| \leq M_2 d^o$ and $M_2 d^o + 1 \leq d^*$, $|A_1|$ can never jump from $[0, M_2 d^o]$ to (d^*, ∞) before λ reaches $M_1\zeta^*$. □

5.3. Proofs of Theorems 6 and 7. We are now ready to prove the theorems stated in Subsection 5.1. The proof of Theorem 6 is relatively simple due to the uniqueness result in Theorem 2 (iii). The proof of Theorem 7 requires both Theorems 6 and 8.

Proof of Theorem 6. Since $\widehat{\boldsymbol{\beta}}^o$ is the oracel LSE, $\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o) = 0$ for $j \in A^o$. If $|\widehat{\beta}_j^o| \ge \lambda\gamma$, then $\dot{\rho}(|\widehat{\beta}_j^o|; \lambda) = 0$ by (1.9). Thus, $\widehat{\boldsymbol{\beta}}^o$ is a solution of (2.2) in the event

$$\Omega^o(\lambda) \equiv \Big\{ \min_{j\in A^o} \mathrm{sgn}(\beta_j)\widehat{\beta}_j^o > \gamma\lambda \Big\} \cap \Big\{ \max_{j\notin A^o} |\boldsymbol{x}_j(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n| < \lambda \Big\}. \tag{5.20}$$

Moreover, $\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}^o$ by the uniqueness of $\widehat{\boldsymbol{\beta}}(\lambda)$ in Theorem 2 (iii). Thus, it suffices to show that the right-hand sides of (5.4) and (5.5) are upper bounds for $1 - P\{\Omega^o(\lambda)\}$.

By (5.2), (5.3) and the normality assumption, $\widehat{\beta}_j^o \sim N(\beta_j, \sigma^2 w_j^o/n)$ for all $j \in A^o$. Since $|\beta_j| \ge \beta_* \ge \gamma\lambda$, this implies

$$P\Big\{ \mathrm{sgn}(\beta_j)\widehat{\beta}_j^o \le \gamma\lambda \Big\} \le \Phi\Big(\frac{\gamma\lambda - |\beta_j|}{\sigma(w_j^o/n)^{1/2}} \Big), \quad j \in A^o.$$

Let $\boldsymbol{P}_1^o$ be the projection from $\mathbb{R}^n$ to the linear span of $\{\boldsymbol{x}_j, j \in A^o\}$. Since $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o = (\boldsymbol{I}_n - \boldsymbol{P}_1^o)\boldsymbol{\varepsilon}$, $\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n$ are normal variables with zero mean and

$$\mathrm{Var}\big(\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n\big) = \mathrm{Var}\big(\boldsymbol{x}_j'(\boldsymbol{I}_n - \boldsymbol{P}_1^o)\boldsymbol{\varepsilon}/n\big) \le \mathrm{Var}(\boldsymbol{x}_j'\boldsymbol{\varepsilon}/n) = \sigma^2\|\boldsymbol{x}_j\|^2/n^2,$$

we have $P\big\{ |\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n| \ge \lambda \big\} \le 2\Phi(-n\lambda/(\sigma\|\boldsymbol{x}_j\|))$ for all $j \notin A^o$. Thus, (5.4) follows by summing the above two probability bounds over all $j \le p$. The inequalities in (5.5) follows from (5.4) and the fact that $\Phi(-t) \le t^{-1}e^{-t^2/2}/\sqrt{2\pi}$. □

Proof of Theorem 7. It follows from Theorem 9 that before λ first reaches $[0, M_1\zeta^*]$,

$$|A_1(\lambda)| \equiv \#\{j : \widehat{\beta}_j(\lambda) \ne 0 \ \text{ or } \beta_j \ne 0\} \le M_2 d^o \le d^* - 1 \tag{5.21}$$

By the proof of Theorem 6, the oracle LSE $\widehat{\boldsymbol{\beta}}^o$ is a solution of (2.2) for $\beta_* > \gamma\lambda$ in the event $\Omega^o(\lambda)$ in (5.20). Thus, since $\beta_* \ge \gamma\lambda_{n,p}$, in the event $\Omega^o(\lambda_{n,p}) \cap \{M_1\zeta^* < \lambda_{n,p}\}$, both $\widehat{\boldsymbol{\beta}}(\lambda_{n,p})$ and $\widehat{\boldsymbol{\beta}}^o$ are solutions of (2.2) when λ first hit $\lambda_{n,p}$ in the PLUS path, say in segment k, with $\lambda^{(k)} \le \lambda_{n,p} \le \lambda^{(k-1)}$ necessarily. By (5.21) and the sparse uniqueness of (2.2) in Theorem 2 (ii), we have $\widehat{\boldsymbol{\beta}}(\lambda_{n,p}) = \widehat{\boldsymbol{\beta}}^o$, so that $(\widetilde{\boldsymbol{z}} \oplus \widehat{\boldsymbol{\beta}}^o)/\lambda_{n,p}$ is a point in $\ell(\boldsymbol{\eta}^{(k)}|\widetilde{\boldsymbol{z}})$ in (3.8). Moreover, since the inequalities in (5.20) are strict, $(\widetilde{\boldsymbol{z}} \oplus \widehat{\boldsymbol{\beta}}^o)/\lambda_{n,p}$ is not a boundary point of $\ell(\boldsymbol{\eta}^{(k)}|\widetilde{\boldsymbol{z}})$ and by (2.2), $(\widetilde{\boldsymbol{z}} \oplus \widehat{\boldsymbol{\beta}}^o)/\lambda \in \ell(\boldsymbol{\eta}^{(k)}|\widetilde{\boldsymbol{z}})$ until λ hits $\lambda^{(k)} = \max_{j\notin A^o} |\boldsymbol{x}_j(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n|$. It follows that

$$\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}^o \ \text{ in } \ \Omega^o(\lambda) \cap \Omega^o(\lambda_{n,p}) \cap \{M_1\zeta^* < \lambda_{n,p}\}.$$

Therefore, the inequalities in the proof of Theorem 6 yield

$$\begin{aligned}
& P\Big\{\widehat{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}^o \text{ or } \operatorname{sgn}(\widehat{\boldsymbol{\beta}}) \neq \operatorname{sgn}(\boldsymbol{\beta})\Big\} \\
\leq \quad & P\Big\{M_1\zeta^* \geq \lambda_{n,p}\Big\} + \sum_{j\in A_o} \Phi\Big(\frac{\gamma\lambda_{n,p} - |\beta_j|}{\sigma(w_j^o/n)^{1/2}}\Big) + 2\sum_{j\notin A_o} \Phi\Big(-\frac{n\lambda}{\sigma\|\boldsymbol{x}_j\|}\Big). \qquad (5.22)
\end{aligned}$$

It remains to bound $P\big\{M_1\zeta^* \geq \lambda_{n,p}\big\}$. By (5.12), $\zeta^* = \max_{1\leq m\leq p}\zeta_m^*$ and $mn(\zeta_m^*)^2/\sigma^2$ is the maximum of $\binom{p}{m}$ chi-square variables with m degrees of freedom. Thus, since $\lambda_{n,p} = M_1\sigma\sqrt{(1+a_{n,p}+2\log p)/n}$,

$$P\big\{M_1\zeta^* \geq \lambda_{n,p}\big\} \leq \sum_{m=1}^{p}\binom{p}{m}P\Big\{\chi_m^2 > m(1+a_{n,p}+2\log p)\Big\}.$$

It follows from the standard large deviation method that for $x > 0$ and $t = x/\{2(1+x)\}$,

$$P\Big\{\chi_m^2 > m(1+x)\Big\} \leq Ee^{t\chi_m^2 - tm(1+x)} = e^{-mx/2}(1+x)^{m/2}.$$

Thus, for $\epsilon_{n,p} = e^{-a_{n,p}/2}\sqrt{1+a_{n,p}+2\log p}$,

$$\begin{aligned}
P\big\{M_1\zeta^* \geq \lambda_{n,p}\big\} \quad & \leq \quad \sum_{m=1}^{p}\binom{p}{m}e^{-m(a_{n,p}/2+\log p)}(1+a_{n,p}+2\log p)^{m/2} \\
& \leq \quad \sum_{m=1}^{p}(m!)^{-1}\epsilon_{n,p}^m = \exp(\epsilon_{n,p}) - 1.
\end{aligned}$$

This and (5.22) imply (5.8).

Finally, for $a_{n,p} = 2\{\epsilon_n\log p + \log((1+\epsilon_n)\log p)\} \to \infty$, $1 + a_{n,p} + 2\log p = (2+o(1))(1+\epsilon_n)\log p$, so that

$$e^{\epsilon_{n,p}} - 1 = (1+o(1))\epsilon_{n,p} = \frac{(1+o(1))\sqrt{2(1+\epsilon_n)\log p}}{p^{\epsilon_0}(1+\epsilon_n)\log p}.$$

This and the proof of (5.5) give (5.9). □

6. Discussion. We have introduced and studied the MC+ methodology for unbiased penalized selection. Our theoretical and simulation results have shown the superior selection accuracy of this new method and the computational efficiency of the PLUS algorithm. We have also provided formulas for the estimation of the MSE and the noise level. In this section, we briefly discuss adaptive penalty, general loss, and the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$.

6.1. Adaptive penalty. The PLUS algorithm applies to the penalized loss

$$\frac{1}{2n}\Big\|\boldsymbol{y}-\boldsymbol{X\beta}\Big\|^2+\sum_{j=1}^p \lambda^2\rho_m\big(|\beta_j|r_j/\lambda\big), \quad r_j>0\ \forall\ j, \tag{6.1}$$

through the scale change $\{\boldsymbol{x}_j,\beta_j\}\to\{\boldsymbol{x}_j r_j,\beta_j/r_j\}$. It can be easily modified to accommodate different quadratic ρ_m of the form (3.1) for different j . For example, different $\gamma=\gamma_j$ can be used with the MP+, so that the j-th path $\widehat{\beta}_j(\lambda)$ reaches the unbiased region when $|\widehat{\beta}_j(\lambda)|r_j/\lambda\geq\gamma_j$. This allows r_j and γ_j to be data dependent. For $r_j=1$, the unbiasedness condition $\gamma_j\lambda\leq|\beta_j|$ allows a higher level of convexity than (1.10) does.

Zou (2006) proposed an adaptive LASSO with $\lambda^2\rho_m\big(|\beta_j|r_j/\lambda\big)=\lambda r_j|\beta_j|$, where r_j is a decreasing function of certain consistent initial estimate of β_j. This approach also reduces the estimation bias of the LASSO and was proven for fixed p to provide consistent selection (1.5) and efficient estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}=\boldsymbol{X\beta}$. For $p>n$, the choice of the initial estimate is unclear for the adaptive LASSO.

6.2. General loss functions. Consider the general penalized loss

$$L(\boldsymbol{\beta};\lambda)\equiv\psi(\boldsymbol{\beta})+\sum_{j=1}^p \lambda^2\rho(|\beta_j|/\lambda), \tag{6.2}$$

where $\psi(\boldsymbol{v})\equiv\psi_n(\boldsymbol{v};\boldsymbol{X},\boldsymbol{y})$ is a convex function of $\boldsymbol{v}\in\mathbb{R}^n$ given data $(\boldsymbol{X},\boldsymbol{y})$. In generalized linear models, $n\psi_n(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})$ is the negative log-likelihood. Let $\dot{\boldsymbol{\psi}}\in\mathbb{R}^n$ and $\ddot{\boldsymbol{\Psi}}\in\mathbb{R}^{n\times n}$ be the gradient vector and Hessian matrix of ψ. With $\tau=1/\lambda$ and $\boldsymbol{b}(\tau)=\widehat{\boldsymbol{\beta}}(\lambda)/\lambda$, (3.5) becomes

$$\begin{cases}\tau\dot{\psi}_j\big(\boldsymbol{b}(\tau)/\tau\big)+\mathrm{sgn}\big(b_j(\tau)\big)\dot{\rho}\big(|b_j(\tau)|\big)=0, & b_j(\tau)\neq 0\\ \tau|\dot{\psi}_j\big(\boldsymbol{b}(\tau)/\tau\big)|\leq 1, & b_j(\tau)=0.\end{cases} \tag{6.3}$$

Let $\boldsymbol{P}_\tau$ be the projection $\boldsymbol{\beta}\to(\beta_j,b_j(\tau)\neq 0)'$ and

$$\boldsymbol{Q}(\tau)\equiv\boldsymbol{P}_\tau'\ddot{\boldsymbol{\Psi}}\big(\boldsymbol{b}(\tau)/\tau\big)\boldsymbol{P}_\tau+\mathrm{diag}\Big(\ddot{\rho}(|b_j(\tau)|),b_j(\tau)\neq 0\Big).$$

Let $\boldsymbol{s}(\tau)\equiv(d/d\tau)\boldsymbol{b}(\tau)$. Differentiation of (6.3) with respect to τ yields

$$\boldsymbol{Q}(\tau)\boldsymbol{s}(\tau)=\boldsymbol{w}(\tau),\quad \boldsymbol{w}(\tau)=\boldsymbol{P}_\tau\Big(\ddot{\boldsymbol{\Psi}}\big(\boldsymbol{b}(\tau)/\tau\big)\boldsymbol{b}(\tau)/\tau-\dot{\boldsymbol{\psi}}\big(\boldsymbol{b}(\tau)/\tau\big)\Big). \tag{6.4}$$

With $\boldsymbol{Q}^{(k)}\equiv\boldsymbol{Q}(\tau^{(k)})$ and $\boldsymbol{w}^{(k)}\equiv\boldsymbol{w}(\tau^{(k)})$, (6.4) leads to the recursion

$$\boldsymbol{Q}^{(k-1)}\boldsymbol{s}^{(k)}=\boldsymbol{w}^{(k-1)},\ \boldsymbol{b}^{(k)}=\boldsymbol{b}^{(k-1)}+\xi^{(k)}\boldsymbol{s}^{(k)}\Delta^{(k)},\ \tau^{(k)}=\tau^{(k-1)}+\xi^{(k)}\Delta^{(k)}, \tag{6.5}$$

where $\xi^{(k)} = -1$ if $\text{sgn}(\boldsymbol{s}^{(k)})\text{sgn}(\boldsymbol{s}^{(k-1)}) \in \{-1, 0\}^p$ and $\xi^{(k)} = 1$ otherwise. We set the entrance and exist policy for the active set of variables according to (6.3). We may invert the matrix $\boldsymbol{Q}^{(k-1)}$ in (6.5) or update $\boldsymbol{s}^{(k)}$ one component at a time. This extends the PLUS algorithm to (6.2). The main difference of (6.5) from the PLUS algorithm for (1.11) is that $\Delta^{(k)}$ has to be small when $\psi(t)$ is not a quadratic spline. The main difference of (6.5) from the existing algorithms for computing the LASSO for the generalized linear models (Genkin, Lewis and Madigan 2004; Zhao and Yu, 2004) is the possibility of the sign change $\xi^{(k)} = \pm 1$ to allow the path to traverse from one local minimum to another. Extensions of the LARS with large step size $\Delta^{(k)}$ have been considered by Rosset and Zhu (2003) for support vector machine and by Zhang (2005) for continuous generalized gradient descent.

6.3. Penalized estimation. Although we considered both variable selection and the estimation of the regression coefficients $\boldsymbol{\beta}$ and the mean vector $\boldsymbol{\mu} \equiv \boldsymbol{X}\boldsymbol{\beta}$, our theoretical results are focused on selection accuracy and risk estimation. Donoho and Johnstone (1994) showed that selection and estimation demand different optimal penalty levels for orthonormal designs. This is also the case in Figures 9 and 11 for Experiment 5. Thus, selection and estimation are closely related but different problems. We note that our selection consistency results do imply the estimation efficiency in the sense of $P\{\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^o\} \to 1$. For recent advances in the LASSO or LASSO-like estimations of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$, we refer to Greenshtein and Ritov (2004), Candés and Tao (2005), van de Geer (2006), and Meinshausen and Yu (2006).

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory,* V. Petrov and F. Csáki, eds. 267281. Akadmiai Kiadó, Budapest.

[2] Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations (with discussion). *Journal of American Statistical Association* **96** 939-967.

[3] Candes, E. and Tao, T. (2005) The Dantzig selector: statistical estimation when p is much larger than n. *Preprint*, Department of Computational and Applied Mathematics, Caltech.

[4] Donoho, D.L. and Johnstone, I. (1994). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields* **99** 277-303.

[5] Donoho, D.L., Johnstone, I.M., Hoch, J.C. and Stern, A.S. (1992). Maximum entropy and the nearly black object. *J. R. Statist. Soc. B* **54** 41-81.

[6] Efron, B. (1986). How biased is the apparent error of a prediction rule? *J. Amer. Statist. Assoc.* **81** 461-470.

[7] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407-499.

[8] Fan, J. (1997). Comments on "Wavelets in statistics: a review" by A. Antoniadis. *J. Italian Statist. Assoc.* **6** 131138.

[9] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 13481360.

[10] Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Annals of Statistics* **32** 928-961.

[11] Foster, D.P. and George, E.I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947-1975.

[12] Freund, Y. and Schapire, R.E. (1996). Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, Morgan Kauffmann, San Francisco, 148-156.

[13] Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.* **28** 337-307.

[14] Gao, H.-Y. and Bruce, A.G. (1997). Waveshrink with firm shrinkage. *Statistica Sinica* **7** 855-874.

[15] Genkin, A. Lewis, D.D. and Madigan, D. (2004). Large-scale Bayesian logistic regression for text categorization. Preprint.

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

[17] Hunter, D.R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617-1642.

[18] Mallows, C.L. (1973). Some comments on Cp. *Technometrics* **12** 661675.

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

[20] Meinshausen, N. and Yu, B. (2006) Lasso-type recovery of sparse representations for high-dimensional data. Technical report, Department of Statistics, University of California, Berkeley.

[21] Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28** 1083-1104.

[22] Osborne, M., Presnell, B. and Turlach, B. (2000a). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20** 389-404.

[23] Osborne, M., Presnell, B. and Turlach, B. (2000b). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9** (2) 319-337.

[24] Rosset, S. and Zhu, J. (2003). Piecewise linear regularized solution paths. *Ann. Statist.*, to appear.

[25] Schapire, R. E. (1990). The strength of weak learnability. Machine Learning **5**, 197-227.

[26] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461464.

[27] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135-1151.

[28] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267-288.

[29] van de Geer, S. (2006). High-dimensional generalized linear models and the Lasso. Technical report, ETH, Zuerich.

[30] Zhao, P. and Yu, B. (2004). Boosted lasso. Technical Report 678, Department of Statistics, University of California, Berkeley.

[31] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. Technical report No. 702. Department of Statistics, University of California, Berkeley.

[32] Zhang, C.-H. (2005). Continuous generalized gradient descent. *Journal of Computational and Graphical Statistics*, to appear.

[33] Zhang, C.-H. and Huang, J. (2006). Model-selection consistency of the LASSO in high-dimensional linear regression. Technical Report 2006-003, Department of Statistics, Rutgers University.

[34] Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. Preprint.

[35] Zou, H. and Li, R. (2006). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, to appear.