DIRECTION IDENTIFICATION AND MINIMAX ESTIMATION IN HIGH-DIMENSIONAL SPARSE REGRESSION VIA A GENERALIZED EIGENVALUE APPROACH

Mathieu Sauvenier, Sébastien Van Bellegem

REPRINT | 3351

UCLouvain

CORE | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

Econometric Theory, **0**, 2026, 1–39.
doi:10.1017/S0266466626100334

DIRECTION IDENTIFICATION AND MINIMAX ESTIMATION IN HIGH-DIMENSIONAL SPARSE REGRESSION VIA A GENERALIZED EIGENVALUE APPROACH

MATHIEU SAUVENIER
LIDAM/CORE, UCLouvain, Louvain-la-Neuve, Belgium

SÉBASTIEN VAN BELLEGEM
LIDAM/CORE, UCLouvain, Louvain-la-Neuve, Belgium

In high-dimensional (HD) sparse linear regression, parameter selection and estimation are addressed using a constraint l_0 on the direction of the parameter vector. We begin by establishing a general result that identifies this direction through the leading generalized eigenspace of specific measurable matrices. Using this result, we propose a novel approach to the selection of the best subsets by solving an empirical generalized eigenvalue problem to estimate the direction of the HD parameter. We then introduce a new estimator based on the RIFLE algorithm, providing a non-asymptotic bound for the estimation risk, minimax convergence, and a central limit theorem. Simulations demonstrate the superiority of our method over existing l_0-constrained estimators.

1. INTRODUCTION

Consider the high-dimensional (HD) linear model

$$Y = X\beta + U, \tag{1}$$

where the rows of the *response random vector* $Y \in \mathbb{R}^n$ are independent and identically distributed (i.i.d.) and square-integrable. The rows X_j, $1 \leq j \leq n$, of the *random covariate matrix* $X \in \mathbb{R}^{n\times p}$ are also i.i.d. and square-integrable, with the number of observations smaller than the number of covariates ($n < p$). Assume that $\mathbb{E}(X_1 U_1) = 0$, so that the model expresses Y_1 as the sum of its projection onto the L_2 space of equivalence classes of linear functions of X_1 and an orthogonal residual U_1.

The contribution of this article is twofold. First, we establish a new identification result showing that β can be expressed as a solution to a specific *generalized*

Address correspondence to Mathieu Sauvenier, LIDAM/CORE, UCLouvain, Louvain-la-Neuve, Belgium, e-mail: mathieu.sauvenier@uclouvain.be.

eigenvalue problem (GEP). Second, we leverage this result to develop a novel estimator of β and analyze its statistical properties.

GEPs involve two matrices $A, B \in \mathbb{R}^{p\times p}$. Their solutions consist of scalars λ_i and vectors v_i that satisfy

$$Av_i = \lambda_i Bv_i, \quad \text{for} \quad i = 1, \ldots, p, \tag{2}$$

where the λ_i are called *generalized eigenvalues* and the v_i are the corresponding *generalized eigenvectors*. For a given λ_i, the subspace spanned by all associated generalized eigenvectors is called the *generalized eigenspace* corresponding to λ_i. When B is the identity matrix, the GEP reduces to the standard eigenvalue problem (Ghojogh, Karray, and Crowley, 2019).

In Theorem 1, we show that the *direction of the parameter* β in model (1)—that is, the subspace spanned by all vectors collinear with β—is identified as the generalized eigenspace corresponding to the unique nonzero generalized eigenvalue of a pair of measurable matrices.

A key consequence of this identification result is that it expands the class of estimators for linear models to include those available for the GEP. As the remainder of the article shows, this is particularly advantageous in HD settings.

In HD models, the dimension of the parameter β, denoted by p, exceeds the number of observations n, often because p grows with n. Estimating β in HD contexts is challenging and of significant theoretical and practical interest. The high dimensionality implies that the empirical second-moment matrix of the covariates is singular, rendering consistent estimation of β by ordinary least squares impossible. Nevertheless, cases with $p > n$ arise frequently in fields, such as economics, finance, health studies, and machine learning (Fan and Li, 2006), motivating the development of efficient estimation methods.

There is no unique way to address the problem of consistently estimating the parameters of an HD linear model. One major line of research builds on the idea of *sparsity*. Box and Daniel Meyer (1986) originally coined the term *factor sparsity* to describe models in which most covariates have no effect. Since then, substantial work has focused on estimating parameters in HD *sparse* linear models (HDSLMs), that is, models in which only a subset of elements in β are nonzero (e.g., Belloni, Chernozhukov, and Hansen, 2013; van de Geer, 2016). In this context, saying that a model is (exactly) *sparse* means that

$$\|\beta\|_0 = s < n < p,$$

where s denotes the *sparsity level* and $\|v\|_0 = \sum_{i=1}^{p} \mathbf{1}_{|v_i|>0}$ is the l_0-pseudonorm, which counts the number of nonzero entries in a vector.

To estimate sparse linear models, the *best subset selection problem* (3) arises by combining the sparsity assumption with the least squares principle:

$$\hat{\beta} \in \operatorname{argmin}_{b\in\mathbb{R}^p} \frac{1}{n}\|Y - Xb\|_2^2 \quad \text{subject to} \quad \|b\|_0 \le k, \tag{3}$$

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

where $k \in \mathbb{N}$ is the maximum number of nonzero elements in b, and $\|v\|_2 = \left(\sum_{i=1}^{p} |v_i|^2\right)^{1/2}$ denotes the l_2 norm of a vector v. Because of the l_0 constraint, this optimization is NP-hard (Natarajan, 1995), so no polynomial-time algorithm is known to solve it. Despite these computational challenges, theoretical results suggest that best subset selection remains an attractive estimator for sparse regression (e.g., Greenshtein, 2006; Raskutti, Wainwright, and Yu, 2011; Zhang and Zhang, 2012; Shen et al., 2013; Zhang, Wainwright, and Jordan, 2014; Bertsimas, King, and Mazumder, 2016). Recently, optimal or near-optimal solutions have been proposed by Bertsimas et al. (2016), who reformulate problem (3) as a mixed integer optimization problem. Implementing this method relies on the licensed GUROBI solver and becomes increasingly demanding in computing time as the dimension of the model grows. For a critical review, see, for example, Hastie, Tibshirani, and Tibshirani (2020).

Other approaches to the best subset selection problem exist, in particular by solving a Lagrangian version of (3). However, this is not an equivalent problem because of the l_0 constraint. Within this line of research, Huang et al. (2018) propose a method that, although it approximates the *penalized* version of (3), still allows one to fix the number of nonzero coordinates in the estimates. Hazimeh and Mazumder (2020) follow a similar approach but combine the l_0 penalty with an additional l_q penalty to achieve improved algorithmic performance.

Following an approach similar to that leading to the formulation of the best subset selection, we leverage our identification result to define an estimator for the direction of β under a sparsity constraint. We begin by noting that the generalized eigenvector associated with the largest generalized eigenvalue of the GEP (2) is the solution to the maximization of the following Rayleigh quotient:

$$\operatorname{argmax}_{v \in \mathbb{R}^p} \frac{v^\top A v}{v^\top B v},$$

see, e.g., Golub and Van Loan (2013). Our identification result in Theorem 1 combines this fact with the sparsity assumption to define a sparse estimator:

$$\operatorname{argmax}_{\phi \in \mathbb{R}^p} \frac{\phi^\top X^\top Y Y^\top X \phi}{n \phi^\top X^\top X \phi} \quad \text{subject to} \quad \|\phi\|_0 \leq k, \tag{4}$$

where $k \in \mathbb{N}$. We refer to (4) as the *directional best subset selection* problem. Our results show that for any solution of (4), say v, the product of v with the orthogonal projection of Y onto Xv also solves the best subset selection problem (3). Consequently, solving (4) allows us to solve (3). Furthermore, Proposition 2 below implies that both problems share the same minimax convergence rate in the l_2 norm, specifically $(s \log(p/s)/n)^{1/2}$ in the sub-Gaussian setting (Raskutti et al., 2011).

Building on these ideas, we propose an efficient estimator for the direction of β in model (1), specifically designed for HDSLMs. Our estimator is a modified version of the well-established RIFLE algorithm, originally introduced by Tan et al. (2018) to approximate the solution to the GEP under an l_0 constraint—

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

an NP-hard problem (Moghaddam, Weiss, and Avidan, 2006)—in settings where $p > n$. We show that our estimator attains the minimax convergence rate with high probability in the sub-Gaussian setting. Finally, we establish a *central limit theorem* (*CLT*) for our estimator, addressing the nontrivial challenges arising from its nonlinearity. To do so, we prove a theorem providing sufficient conditions for the convergence in probability of a *Cauchy product*, which may be of independent interest.

GEPs have previously been applied to models with many covariates. In particular, the *sufficient dimension reduction* framework, introduced by Dennis Cook (1994), uses moment-based methods that can be formulated as GEPs to replace the original covariates with a minimal set of linear combinations while preserving the relevant information (Li, 2007). Although these methods were not specifically designed for HDSLMs or for the best subset selection problem, Chen, Zou, and Cook (2010) proposed a method to induce sparsity in the estimated linear combinations of covariates, thereby achieving variable selection *de facto*.

Like the RIFLE algorithm, our estimator requires an initial solution, which it then iteratively refines until convergence. We conclude with a series of Monte Carlo experiments demonstrating the superior performance of our estimator when initialized with the Lasso method, which is the Lagrangian form of the l_1-relaxation of problem (3) Tibshirani, 1996; Chen, Donoho, and Saunders, 1998. Although the Lasso often selects an excessive number of variables when optimized for prediction accuracy Bertsimas et al., 2016; Hastie et al., 2020, our experiments show that for realistic signal-to-noise ratios (SNRs), our estimator achieves more reliable variable selection and estimation than existing l_0-constrained estimators.

The performance of the proposed inference method, termed Loaded-RIFLE (LR), is benchmarked against three alternatives: the Lasso and two l_0-constrained or penalized estimators. The first is the two-stage iterative hard thresholding (two-stage IHT; Jain, Tewari, and Kar, 2014), which extends the well-known iterative hard thresholding to approximate problem (3). The second is the support detection and root finding (SDAR) method of Huang et al. (2018), which approximates the solution to the Lagrangian form of problem (3).

These existing methods have well-established minimax convergence rates and the sure screening property Jain et al., 2014; Huang et al., 2018; Guo, Zhu, and Fan, 2020. However, to the best of our knowledge, no CLT has yet been established for any of them.

Section 4 reports Monte Carlo experiments comparing these estimators in terms of estimation, variable selection, and prediction. Our results show that LR consistently outperforms its competitors across these metrics in many cases. In particular, LR achieves more accurate variable selection, recovering a substantial share of the true support even under low SNR conditions. Moreover, its prediction accuracy, measured by relative risk, remains robust in low-SNR regimes, highlighting its practical relevance for applications such as financial analysis, where low SNRs are common. Overall, these findings suggest that LR offers a promising alternative for reliable estimation and prediction in noisy HD settings.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

The organization of the article is as follows. Section 2 establishes the identification of the direction of β as the solution to a particular GEP. Section 3 builds on this result to estimate the direction of β under a sparsity assumption. We first introduce our algorithm, LR. We then provide a nonasymptotic bound on its l_2 estimation error and derive its convergence rate in sub-Gaussian settings. Next, we show that this rate is minimax optimal in the sub-Gaussian case. Finally, we establish a CLT for the estimator. Section 4 presents Monte Carlo experiments that illustrate the competitiveness of LR in practice. Section 5 concludes. The Appendix collects the proofs of all results, except for the proof of the identification theorem, which is presented in Section 2. It also includes a description of the RIFLE algorithm. Technical lemmas and several standard results used throughout the proofs are provided in the Supplementary Material, where they are proved in full detail. For the convenience of the reader, a Notational Glossary is provided after the Appendix.

2. IDENTIFICATION OF THE DIRECTION OF β

The purpose of this section is to identify the direction of β in model (1) in terms of the moments of Y_1 and X_1.

First, observe that in the model, since the rows of Y and X are i.i.d. in L_2, and because $\mathbb{E}(X_1U_1) = 0$, one easily obtains *normal equations* $\mathbb{E}(X_1Y_1) = \mathbb{E}(X_1X_1^\top)\beta$. For every $\beta \notin \mathrm{Ker}(\Sigma_X) := \{v \in \mathbb{R}^p; \|\Sigma_X v\|_2 = 0\}$, that is for every β not belonging to the *kernel* of *the second moment of* X_1, the latter denoted hereafter $\Sigma_X := \mathbb{E}(X_1X_1^\top)$, one implication of these normal equations is

$$\|\beta\|_2 = \frac{\phi^\top \Sigma_{XY}}{\phi^\top \Sigma_X \phi}$$

with $\Sigma_{XY} := \mathbb{E}(X_1Y_1)$ and $\phi := \beta/\|\beta\|_2$. Hence, rewriting β in terms of ϕ and $\|\beta\|_2$, one obtains after standard computations that

$$\mathbb{E}[Y_1 - X_1^\top \beta]^2 = \mathbb{E}(Y_1^2) - \frac{\phi^\top \Sigma_{XY}\Sigma_{YX}\phi}{\phi^\top \Sigma_X \phi} \tag{5}$$

with $\Sigma_{YX} := \Sigma_{XY}^\top$. The condition $\mathbb{E}(X_1U_1) = 0$ means that $X_1\beta$ is the orthogonal projection of Y_1 onto the L^2 subspace of linear functions of X_1, therefore the *Projection theorem* (e.g., Parthasarathy, 1977) implies that

$$\beta \in \mathrm{argmin}_{b\in\mathbb{R}^p}\, \mathbb{E}[Y_1 - X_1^\top b]^2 \tag{6}$$

and thus

$$\phi \in \mathrm{argmax}_{b\in\mathbb{R}^p}\, \frac{b^\top \Sigma_{XY}\Sigma_{YX} b}{b^\top \Sigma_X b}. \tag{7}$$

As appears readily, if ϕ satisfies Equation (7), so is every vector in $\Phi = \{\alpha\beta | \alpha \in \mathbb{R}\}$ (the *direction* of β) that is not the zero vector. Hence, the identification of β, for instance via the assumption $\mathrm{Ker}(\Sigma_X) = \{0\}$, ensures the identification of Φ.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

In contrast, the identification of β is obtained from the identification of Φ by orthogonally projecting Y_1 onto the L_2 subspace of linear functions of $X_1^\top \phi$ for any $\phi \in \Phi$. This subspace is closed and convex, which ensures the uniqueness of the orthogonal projection. It is also contained in the subspace of the linear function of X_1. Hence, considering the orthogonal projection of Y_1 as an element of this latter space uniquely identifies the solution of Equation (6), that is, β in model (1).

The previous paragraph implies the following identification theorem.

THEOREM 1. *In model (1), assume that Σ_X is positive definite, and let $\Phi = \{\alpha\beta | \alpha \in \mathbb{R}\}$ be the direction of β. Then,*

$$\Phi = \mathrm{Ker}(\Sigma_{XY}\Sigma_{YX} - \Sigma_X \lambda_{\max}), \tag{8}$$

where $\lambda_{\max}$ denotes the largest generalized eigenvalue, defined by

$$\lambda_{\max} = \max_{\phi \in \mathbb{R}^p} \frac{\phi^\top \Sigma_{XY}\Sigma_{YX}\phi}{\phi^\top \Sigma_X \phi}.$$

Equation (8) can be obtained by explicitly solving the maximization (7). The calculations are given in the Appendix. Hence, in this section, we have identified Φ as the solution set of a generalized Rayleigh quotient given in the argument of Equation (7), or equivalently, as the generalized eigenspace associated with the largest generalized eigenvalue of the matrix pair $(\Sigma_{XY}\Sigma_{YX}, \Sigma_X)$ (Equation (8)).

This result highlights a general geometric property of linear models, which, to the best of our knowledge, has not been explicitly noted in the literature. However, we mention that a similar equation appears in the context of partial least squares in Sun, Ji, and Ye (2009).

3. ESTIMATION UNDER SPARSITY

In this section, we use the identification result of Section 2 to produce a theory of estimation of the parameter β in model (1) *under the assumption of sparsity*. Although the literature has proposed solving specific GEPs using various least-squares formulations (e.g., Moghaddam et al., 2006, Sun et al., 2009, Golub and Van Loan, 2013), an original aspect of our approach is to employ a GEP to estimate the parameters of a linear model, a task typically achieved by solving least-squares problems.

Theorem 1 suggests a *directional estimation* of β involving two steps. First, obtain an estimation of the direction β, say $\hat{\Phi}$, via GEP. Second, projecting Y to the span of $X\hat{\phi}$, for $\hat{\phi} \in \hat{\Phi}$ and $\|\hat{\phi}\|_2 = 1$, to obtain (say) $\hat{\nu}$, an estimate of $\|\beta\|_2$, and thus obtaining an estimator of β by applying the Slutsky lemma to the product $\hat{\phi}\hat{\nu}$.

If model (1) is sparse, that is, $\|\beta\|_0 = s < n < p$, one may, in principle, employ the directional best subset section defined in (4) to achieve the first step of the

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

directional estimation. As detailed in Proposition A1 in the Appendix, the solution of the best subsets selection (3) would then be exactly recovered after the second step of the procedure. Nevertheless, as stated in the Introduction, both of these problems are NP-Hard, thus, up to the time when the conjecture $P = NP$ has been (dis)proved, the best one can generally hope for in polynomial time is to approximate their solution. This is exactly our suggestion.

The key point relative to estimation is that Theorem 1 allows to leverage the computational ideas developed in the GEP literature to estimate β in model (1). We introduce a competitive estimator based on the *truncated Rayleigh flow method* (called *RIFLE*) introduced by Tan et al. (2018) to estimate the solution of GEP in HD under a sparsity constraint. This new estimator is described in Algorithm 1 below and is called *Loaded-RIFLE (LR)* because it is a modified version of RIFLE.

Starting from an initial solution, the RIFLE as introduced by Tan et al. (2018) approximates the solution of a GEP under a sparsity constraint when $p > n$. The approach involves iteratively updating the estimated vector in its ascending direction, normalizing it, truncating it so that only the k coefficients with the largest absolute values remain, and then normalizing again. The resulting algorithm is detailed in the Appendix and is referred to in the main text as Algorithm 2. Our experience with RIFLE suggests that under certain conditions (specifically when the step size parameter denoted by η is small, as expected, at least according to the theory of Tan et al. (2018)), the selection of variables performed by the algorithm heavily depends on the order of the absolute values in the initial vector ϕ_0. In most cases, we find that this dependency limits the accuracy of the estimate. To mitigate this effect, we introduce a variant of RIFLE, that is LR.

Hereafter, let $\hat{\Sigma}_X := X^\top X/n$, $\hat{\Sigma}_{XY} := X^\top Y/n$, and $\hat{\Sigma}_{YX} := (\hat{\Sigma}_{XY})^\top$ serve as estimators for Σ_X, Σ_{XY}, and Σ_{YX}, respectively. Moreover, for any vector $v \in \mathbb{R}^p$ and subset $\mathcal{K} \subset \{1, \ldots, p\}$, we define Truncate$(v, \mathcal{K})$ as the vector in $\mathbb{R}^p$ given by

$$(\text{Truncate}(v, \mathcal{K}))_i = \begin{cases} v_i, & \text{if } i \in \mathcal{K}, \\ 0, & \text{otherwise.} \end{cases}$$

ALGORITHM 1. Loaded and truncated Rayleigh flow method (Loaded-RIFLE).
Input: *Matrices* $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ *and* $\hat{\Sigma}_X$*; initial vector* ϕ_0*; initial and final sparsity levels* $\bar{k}, \underline{k} \in \{1, \ldots, p\}$ *with* $\bar{k} > \underline{k}$*; step size* η*; and number of iterations* T.
START

1. *Set* $\mathcal{F}_{0,\bar{k}} =$ *indices of the* $\bar{k}$ *highest elements of* $|\phi_0|$.
2. *Compute* $\phi_{0,\bar{k}} = Truncate(\phi_0, \mathcal{F}_{0,\bar{k}})$.
3. *Set* $k = \bar{k}$.
4. ***While*** $k > \underline{k}$*, repeat the following steps:*
   (a) *Set* $t = 1$.
   (b) ***While*** $t \leq T$*, repeat the following steps:*
      i. *Compute* $\rho_{t-1,k} = \frac{\phi_{t-1,k}^\top \hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1,k}}{\phi_{t-1,k}^\top \hat{\Sigma}_X \phi_{t-1,k}}$.
      ii. *Compute* $C_{t-1,k} = I_{p\times p} + \frac{\eta}{\rho_{t-1,k}}(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \rho_{t-1,k}\hat{\Sigma}_X)$.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

iii. *Compute* $\phi'_{t,k} = \frac{C_{t-1,k}\phi_{t-1,k}}{\|C_{t-1,k}\phi_{t-1,k}\|_2}$.
iv. *Set* $\mathcal{F}_{t,k}$ = *indices of the* k *highest elements of* $|\phi'_{t,k}|$.
v. ***If*** $Card(\mathcal{F}_{t,k} \cup \mathcal{F}_{t-1,k}) > \overline{k}$, *then:*
A. *Set* $\phi_{T,k} = \phi_{0,k}$.
B. *Set* $t = T + 1$ *(exit the inner while loop on t).*
vi. ***Else****:*
A. *Compute* $\hat{\phi}_{t,k} = Truncate(\phi'_{t,k}, \mathcal{F}_{t,k})$.
B. *Compute* $\phi_{t,k} = \frac{\hat{\phi}_{t,k}}{\|\hat{\phi}_{t,k}\|_2}$.
C. *Increment* $t = t + 1$.

(a) *Set* $\phi_{0,k-1} = \phi_{T,k}$.
(b) *Decrement* $k = k - 1$.

5. *Apply Algorithm 2 with* $\phi_0 = \phi_{T,\underline{k}+1}$, $k = \underline{k}$, *and* η. *Let* ϕ_t *be the output of Algorithm 2 after t iterations*

END
Output*:* $\hat{\phi}(t) = \phi_t$.

LR can be seen as a kind of backward stepwise RIFLE. It basically runs RIFLE for $k = \overline{k}$ and then uses the output as input of a new run of RIFLE with $k = \overline{k} - 1$. The algorithm proceeds as such until $k = \underline{k}$. The only difference lies in the IF condition in Step 4.(b)v., which ensures a somewhat smooth update of the set of selected variables. This step is important in extending the RIFLE theory developed by Tan et al. (2018) to LR.

The parameter $\underline{k}$ is a tuning parameter, typically expected to be larger than the true sparsity level s. In the absence of prior knowledge about s, $\underline{k}$ can be selected using standard tuning procedures (see, e.g., Chetverikov, 2024). In the next section, we provide an example and a simulation study illustrating the effectiveness of cross-validation for this purpose.

The theoretical results of the following section suggest that $\overline{k}$ should be chosen such that $\overline{k} = Cs$, for some sufficiently large constant C. However, since s is unknown in practice, we propose the following heuristic: $\overline{k}$ should be selected to balance the computational cost and the model's capacity to include all relevant variables. For example, when assuming the prior $s \ll n < p$, the choice $\overline{k} \approx \frac{3}{4}n$ performs well in practice.

3.1. Minimax Rate of Convergence

Using techniques from Tan et al. (2018) to study the convergence of RIFLE, we derive a non-asymptotic bound on the l_2 distance between the output of LR and the population target vector, defined hereafter $\phi^\star := \beta/\|\beta\|_2$ for β in model (1). Based on this bound, we establish the convergence rate of the estimator in a sub-Gaussian setting and demonstrate that it is minimax. Before presenting the theorems, we introduce additional notations, some of which are specific to GEP and some non-standard but useful within the context of this article.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

For any $Z \in \mathbb{R}^{p\times p}$ and $k \in \mathbb{N}$ s.t. $k \leq p$, let

$$\rho(Z,k) := \sup_{\|v\|_2=1,\, \|v\|_0\leq k} |v^\top Z v|.$$

For fixed k, $\rho(\cdot,k) : \mathbb{R}^{p\times p} \to \mathbb{R}$ is a norm on $\mathbb{R}^{p\times p}$ over $\mathbb{R}$, if Z is symmetric, then $\rho(Z,k)$ is the supremum on a set of maximum eigenvalues of submatrices of Z.

For any $k \in \mathbb{N}$ such that $k \leq p$ let $\mathcal{F}(p,k) := \{F \subset \{1,\ldots,p\}; \text{Card}(F) = k\}$ be the set of sets containing k different integers between 1 and p. Moreover, let $\{e_i\}_{i=1}^p$ be the canonical basis in $\mathbb{R}^p$ (i.e., e_i has zeros on all its elements except on the i^{th} position which contains a one) and let $\mathcal{P}\big(\mathcal{F}(p,k)\big)$ be

$$\left\{ P_{(F)} := (e_{\alpha_1},\ldots,e_{\alpha_k}) \in \mathbb{R}^{p\times k} \middle| F \in \mathcal{F}(p,k); \alpha_i \in F; \alpha_1 < \alpha_2 < \cdots < \alpha_k \right\},$$

that is, the set of orthogonal projectors from $\mathbb{R}^p$ to its k-dimensional subspaces that are spanned by the subsets of $\{e_1,\ldots,e_p\}$ whose cardinalities equal k and that preserve the order of the indexes. This set is useful since for any $v \in \mathbb{R}^p$ such that $\|v\|_0 = k$ then $\mathcal{P}\big(\mathcal{F}(p,k)\big)$ contains a matrix P such that $w = P^\top v$, $w \in \mathbb{R}^k$ and w is equal to v if the former is embodied in $\mathbb{R}^p$. It is easy to check that any of these P satisfies $P^\top P = I_{k\times k}$, $\|PP^\top\|_{\text{spec}} = 1$ and $\|P\|_{\text{spec}} = 1$, where for any $Z \in \mathbb{R}^{p\times p}$, $\|Z\|_{\text{spec}} := \max_{\|v\|_2=1} \|Zv\|_2$ is the *spectral norm* of Z.

For any $m \in \mathbb{N}$ and $Z \in \mathbb{R}^{m\times m}$, let $\lambda_{\max}(Z), \lambda_{\min}(Z)$, and $\lambda_i(Z)$ be, respectively, the maximum, minimum, and i^{th} (in descending order) *eigenvalues* of Z. If Z is positive definite, let $\kappa(Z) = \lambda_{\max}(Z)/\lambda_{\min}(Z)$ be the *condition number* of Z (e.g., Golub and Van Loan, 2013).

For any $k \in \mathbb{N}$ s.t. $k \leq p$ and any $F \in \mathcal{F}(k,p)$, let λ_i, $\lambda_i(F)$, $\hat{\lambda}_i$, and $\hat{\lambda}_i(F)$ be the i^{th} (in descending order) *generalized* eigenvalues of, respectively, the pairs of matrices $\big(\Sigma_{XY}\Sigma_{YX}, \Sigma_X\big)$, $\big(P_{(F)}^\top \Sigma_{XY}\Sigma_{YX}P_{(F)}, P_{(F)}^\top \Sigma_X P_{(F)}\big)$, $\big(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}, \hat{\Sigma}_X\big)$, and finally $\big(P_{(F)}^\top \hat{\Sigma}_{XY}\hat{\Sigma}_{YX}P_{(F)}, P_{(F)}^\top \hat{\Sigma}_X P_{(F)}\big)$.

For any $p \in \mathbb{N}$ and any symmetric matrices $A, B \in \mathbb{R}^{p\times p}$, let

$$Cr(A,B) = \min_{\|v\|_2=1} \left(\big(v^\top A v\big)^2 + \big(v^\top B v\big)^2 \right)^{1/2}$$

be the *Crawford number* of (A,B). Stewart (1979) shows that if $Cr(A,B) > 0$, the GEP associated with (A,B) is *definite*, namely, it has a complete system of generalized eigenvectors and all its generalized eigenvalues are real.

For any $k \in \mathbb{N}$ such that $k \leq p$ define

$$\Delta(k) := \inf_{P\in\{\cup_{i=1}^k \mathcal{P}(\mathcal{F}(p,i))\}} Cr(P^\top \Sigma_{XY}\Sigma_{YX}P, P^\top \Sigma_X P).$$

and

$$\varepsilon(k) := \left(\rho(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \Sigma_{XY}\Sigma_{YX}, k)^2 + \rho(\hat{\Sigma}_X - \Sigma_X, k)^2 \right)^{1/2}.$$

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

In the next results, we have employed the techniques developed in Tan et al. (2018) to provide a non-asymptotic bound on the l_2 error of LR.

THEOREM 2. *In model (1), assume Σ_X is positive definite and that $\|\beta\|_0 = s$. Let $\hat{\phi}(t)$ denote the output of Algorithm 1 after t iterations of the last step, ϕ_0 represent the initial truncated vector as specified in Step 2 of Algorithm 1 and $\phi^\star = \beta/\|\beta\|_2$ stand for the population target vector. Consider the following parameterization for Algorithm 1: $\underline{k} = C's < p$ with $C' > 1$, $\bar{k} = C''\underline{k} < p - s$ with $C'' > 1$, and $k' = \bar{k} + s$. Under this parametrization, select a constant $c > 0$ and a step size η such that*

$$\eta\lambda_{\max}(\Sigma_X) < \frac{1}{1+c}, \tag{9}$$

$$\nu = \left(1 + \sqrt{\frac{s}{\underline{k}}}\right)\left(1 - \frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond \kappa(\Sigma_X)}\right)^{1/2} < 1, \tag{10}$$

where $C_\diamond = (1+c)/(1-c)$. Define $\Lambda := \lambda_{\max}/(1+\lambda^2_{\max})$. Then, conditioned on the events where

$$\frac{\varepsilon(k')}{\Delta(k')} < \frac{\Lambda(1-\nu)\sqrt{\xi}}{4\sqrt{10} + \Lambda(1-\nu)\sqrt{\xi}}, \tag{11}$$

$$\rho(\hat{\Sigma}_X - \Sigma_X, k') \le c\lambda_{\min}(\Sigma_X), \tag{12}$$

$$|\langle \phi^\star, \phi_0\rangle_2| \ge 1 - \nu^2\xi, \tag{13}$$

with ξ is defined as

$$\xi := \min\left\{\frac{1}{8C_\diamond\kappa(\Sigma_X)}, \frac{1}{30(1+c)C^3_\diamond\eta\lambda_{\max}\kappa^3(\Sigma_X)}\right\},$$

the upper bound

$$\sqrt{1 - |(\phi^\star)^\top\hat{\phi}(t)|} \le \nu^t\sqrt{\xi} + \frac{1}{(1-\nu)^2}\frac{2\sqrt{5}}{\Lambda\big(\Delta(k') - \varepsilon(k')\big)}\varepsilon(k') \tag{14}$$

is satisfied almost surely.

Since the theorem extends a known result from Tan et al. (2018) (originally developed for RIFLE) to LR, we provide a complete proof in the Supplementary Material. This proof also fixes the proof of the theorem for RIFLE given in Tan et al. (2018). (Our proof is given in the context of our article, but is readily generalizable to the setup of Tan et al. (2018).)

Theorem 2 enables us to bound the l_2 norm between $\phi^\star$ and $\hat{\phi}(t)$. Specifically, when the inner product $\langle \phi^\star, \hat{\phi}(t)\rangle_2 \ge 0$, we have $\sqrt{1 - |(\phi^\star)^\top\hat{\phi}(t)|} = 1/\sqrt{2}\|\phi^\star - \hat{\phi}(t)\|_2$.

The first term on the right-hand side, $\nu^t\sqrt{\xi}$, represents the optimization error, while the second term represents the statistical error (Tan et al., 2018).

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

Assumptions (11) and (12) control the estimation error for the (generalized) spectrum of the matrix pair. Specifically, these assumptions provide bounds on the condition number of Σ_X, on the spectra of submatrices of $\hat{\Sigma}_X$, and on $\hat{\lambda}_i(F)$ for any $F \in \mathcal{F}(p, k')$ and $i = 1, \ldots, k'$; details can be found in Tan et al. (2018). As shown later, if Y_1 and X_1 are sub-Gaussian, there exist ratios among n, $p(n)$, and $s(n)$ such that the probability of events where these assumptions fail converges to zero as $n \to \infty$.

Theorem 2 also requires certain theoretical conditions on the input parameters of Algorithm 1.

Inequalities (9) and (10) impose technical restrictions on *step size* η. Currently, there is no known method to select η that consistently meets these theoretical criteria. However, both experience and Assumption (9) from Tan et al. (2018) suggest selecting η small enough to satisfy $\eta\lambda(\hat{\Sigma}_X) < 1$.

Furthermore, Equation (13) requires that the cosine between the target vector (population) $\phi^\star$ and the input vector ϕ_0 be sufficiently large. We discuss the choice of the initial vector in the following result and example. This result gives an explicit convergence rate depending on the initialization of Algorithm 1. The next example shows how this result applies in practice.

PROPOSITION 1. *In model (1) assume that* $Y_1, X_{1,1}, \ldots, X_{1,p(n)}$ *are sub-Gaussian, that* $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ *is positive definite and that* $\|\beta\|_0 = s$. *Let* $\hat{\phi}(t)$ *denote the output of Algorithm 1 after t iterations in the last step. Let* ϕ_0 *be the initial truncated vector as specified in Step* 2 *of Algorithm 1 and* $\phi^\star = \beta/\|\beta\|_2$ *represent the population target vector. Suppose that* $\langle \phi^\star, \hat{\phi}(t) \rangle_2 \geq 0$ *a.s. and that* $s\log(p/s)/n = o(1)$ *and* $s/p = o(1)$. *Then, under assumptions (9) and (10) of Theorem 2 if*

$$\|\phi^\star - \phi_0\|_2 = O_p(\alpha_n)$$

with $\alpha_n = o(1)$, *then there exists a choice of* $t = t(n)$ *such that*

$$\|\phi^\star - \hat{\phi}(t)\|_2 = O_p\left(\sqrt{\frac{s\log(p/s)}{n}}\right).$$

The proof is given in the Appendix. An estimate of the magnitude of β can then be obtained by regressing Y onto the span of $X\hat{\phi}(t)$. Let $\hat{\upsilon}$ be this estimator, and let $\hat{\beta}$ denote the estimator of β obtained as the product of $\hat{\phi}(t)$ and $\hat{\upsilon}$. Proposition 2 below tells us that for $t = t(n)$ growing sufficiently fast, we have $\|\hat{\beta} - \beta\|_2 = O_p\big((s\log(p/s)/n)^{1/2}\big)$, which is minimax (Raskutti et al., 2011).

Example 1 [Loaded-RIFLE-post-Lasso]. To obtain a reliable initial vector, one might follow the heuristic suggested in Tan et al. (2018) by using a solution from an l_1 relaxation of the l_0 constrained problem as the initial vector, specifically the Lasso solution for the problem (3) Bertsimas et al., 2016; Hastie et al., 2020.

This choice of initialization is particularly suitable since the Lasso is known to select a large number of covariates when optimized for predictive accuracy. Using

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

its (normalized) solution as the initial vector for Algorithm 1 enables us to reduce the size of the selected support while improving the accuracy of the estimation. As demonstrated in the Monte Carlo experiments in Section 5, this procedure outperforms the other l_0-constrained estimators considered in this article.

The theoretical validity of this initialization procedure depends on how the Lasso is tuned. For example, Chetverikov, Liao, and Chernozhukov (2021) established that the l_2 error of cross-validated Lasso is $O_p\big((s\log(p/s)/n)^{1/2}\times\log(pn)\big)$. This error rate ensures that the probability of events where the assumptions on ϕ_0 in Theorem 2 are violated converges to zero. Other tuning methods for the Lasso that provide similar guarantees are available. For example, one alternative to cross-validation could be to set the penalty parameter based on the self-normalized moderate deviation theory, as proposed in Belloni et al. (2012).

To finalize the theoretical analysis of the convergence of LR, we demonstrate that the convergence rate obtained in Proposition 1 in the sub-Gaussian setting is actually optimal, in the sense that it is the minimax convergence rate of the selection of the directional best subset selection defined by Equation (4). To demonstrate this statement, the following proposition connects the estimation error of β in the model (1) under the assumption of sparsity to the estimation error of $\phi^\star$, thus linking the minimax convergence rates of the two estimators.

PROPOSITION 2. *In model (1) suppose that $\Sigma_X := \mathbb{E}(X_1X_1^\top)$ is positive definite and that $\|\beta\|_0 = s$. Let $\hat\phi \in \mathbb{R}^p$ be an estimator of $\phi^\star := \beta/\|\beta\|_2$ that satisfies the following conditions a.s.:*

$$\langle\hat\phi,\phi^\star\rangle_2 \geq 0, \quad \|\hat\phi\|_2 = 1, \quad \|\hat\phi\|_0 = Cs \text{ with } C \geq 1,$$

the convergence rate $\|\hat\phi - \phi^\star\|_2 = O_p(\alpha_n)$, where $\alpha_n = o(1)$ and $(2s/n)^{1/2}/\alpha_n = o(1)$, and let $\hat\upsilon = \operatorname{argmin}_{m\in\mathbb{R}}\|Y - X\hat\phi m\|_2^2$ be an estimator of $\|\beta\|_2$. Assume that there exist constants $0 < \underline{\kappa} \leq \overline{\kappa} < \infty$ such that, for all $u \in \mathbb{R}^p$ with $\|u\|_2 = 1$ and $\|u\|_0 = \min\{2s,p\}$, the inequality

$$\underline{\kappa} \leq u^\top\hat\Sigma_X u \leq \overline{\kappa} \tag{15}$$

holds true almost surely. Then, if $\hat\beta \in \mathbb{R}^p$ is an estimator of β satisfying $\hat\beta = \hat\phi\hat\upsilon$ a.s., we have

$$\|\hat\beta - \beta\|_2^2 = c\|\hat\phi - \phi^\star\|_2^2 + O_p\big(\alpha_n^2\big) \qquad a.s.,$$

where $c = O_p(1)$.

The proof is given in the Appendix. The proposition provides sufficient conditions to ensure that the l_2 convergence rate of $\hat\beta$ is dominated by the l_2 convergences of the direction estimator $\hat\phi$.

This result is important because it implies that the minimax convergence rate in the loss l_2 of the directional best subset selection (4) is the same as that of the best subset selection (3). Raskutti et al. (2011) have proven that in a Gaussian

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

fixed design (that is, restricting oneself to an analysis conditionally on X), if a version of assumption (15) holds, the minimax convergence rate of $\|\hat{\beta} - \beta\|_2$ is $(s\log(p/s)/n)^{1/2}$ (since $\hat{\beta}$ is the best selection of subsets). Given that this rate is obtained in *fixed design*, it is actually a lower bound on the unconditioned convergence of the best subset selection defined in Equation (3), and thus by Proposition 2 of the directional best subset selection in (4). Since by Proposition 1, LR actually achieves this rate in the sub-Gaussian, unconditioned setting, it must be that the rate obtained in Raskutti et al. (2011) is actually the minimax rate of convergence of the directional best subset selection, thus concluding the demonstration of our claim of optimality of LR.

3.2. Central Limit Theorem

In this section, we establish a CLT for the output of LR. Before presenting the results, we introduce the following definition, which is essential for the statement and proof of the theorem.

DEFINITION 1. *In model (1), let* $\phi^\star := \beta/\|\beta\|_2$ *as above, and* $\|\beta\|_0 < k \in \mathbb{N}$. *Hereafter, we reserve the notation* $\mathcal{Q}^\star$ *for the set of matrices defined by*

$$\mathcal{Q}^\star := \{Q = PP^\top | P \in \cup_{i=\|\phi\|_0}^{k} \mathcal{P}\big(\mathcal{F}(p,i)\big), Q\phi^\star = \phi^\star\}.$$

Moreover, hereafter, we define for any output of Algorithm 2 noted $\hat{\phi}(t)$

$$Q\big(\hat{\phi}(t)\big) = Q_t = P_t P_t^\top,$$

where P_t *is the element of* $\mathcal{P}(\mathcal{F}(p, \|\hat{\phi}(t)\|_0))$ *that is such that if* $w = P_t^\top \hat{\phi}(t)$, *then* w *is equal to* $\hat{\phi}(t)$ *if the former is embodied in* $\mathbb{R}^p$.

We note that any output of Algorithm 1, say $\hat{\phi}(t)$, that is such that $Q\big(\hat{\phi}(t)\big) \in \mathcal{Q}^\star$, has a support containing the support of $\phi^\star$. From Lemma A5 in the Appendix, one can show that in the sub-Gaussian setting, for $t = t(n)$ fast enough, $Q\big(\hat{\phi}(t)\big) \in \mathcal{Q}^\star$ with probability tending to one, i.e., LR possesses *sure screening property* Fan and Lv, 2008; Guo et al., 2020.

In the following, we make use of the following convention.

NOTATION 1. *Let* $\{a_\alpha\}_{\alpha\in\mathbb{N}}$ *be a family of operators in a vector space, if* $j < i$, *then* $\prod_{\alpha=i}^{j} a_\alpha = 1$. *where in the above expression,* 1 *represents the unitary operator in the vector space. If* $j \geq i$, *the symbol* $\prod$ *is used conventionally.*

We now state the CLT for the output of Algorithm 1.

THEOREM 3. *In model (1), suppose that* $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ *is positive definite and* $\|\beta\|_0 = s$.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

Let $\sigma^2 = \mathrm{Var}(U_1)$, $\phi^\star := \beta/\|\beta\|_2$ *and* $\hat{\phi}(t)$ *be the output of Algorithm 1 after t iteration of the last step. Moreover, let* η, ϕ_{t-1} *and* C_{t-1} *be as in Algorithm 2 and let* $\tilde{\phi}_t := C_{t-1}\phi_{t-1}$. *Assume t is an increasing function of n and that* $\langle \phi^\star, \hat{\phi}(t)\rangle_2 \geq 0$ *holds almost surely.*

Suppose there exist constants $\xi > 0$ *and* $\nu < 1$ *such that*

$$\|\phi^\star - \hat{\phi}(t)\|_2 \leq \nu^t \sqrt{\xi} + O_p\left(\sqrt{\frac{s\log(p/s)}{n}}\right), \tag{16}$$

and that $\inf_{\|v\|_2=1, \|v\|_0=k} |v^\top \Sigma_{XY}| > 0$ *with* $\log(n)/t(n) = o(1)$.

Then, conditionally on X, there exists a matrix $Q^\star \in \mathcal{Q}^\star$ *and a constant* $K > 0$ *such that if* $\eta < K$, *then for any unit vector* $w \in \mathbb{R}^p$, $\|w\|_0 \leq \|\hat{\phi}(t)\|_0$, *we have*

$$\sqrt{n}\langle w, \hat{\phi}(t) - \phi^\star + \gamma_n\rangle_2$$

is asymptotically normal,

$$\mathcal{AN}\left(0, w^\top[(I - Q^\star) + \eta\hat{\Sigma}_X Q^\star]^{-1}\eta^2\hat{\Sigma}_X[(I - Q^\star) + \eta\hat{\Sigma}_X Q^\star]^{-1}w\frac{\sigma^2}{\|\beta\|_2^2}\middle| X\right),$$

with $\|\gamma_n\|_2 = o_p(1)$, *where* $\gamma_n := \kappa_n + \tau_n$.

Here,

$$\kappa_n = \sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\right)(\phi_{t-l} - \tilde{\phi}_{t-l}),$$

and

$$\tau_n = \left[\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}g_{t-1-m}\right)\left((I - Q^\star) + \eta\hat{\Sigma}_X Q^\star\right) - I_{p\times p}\right]\phi^\star,$$

with

$$g_{t-1-m} = \frac{\phi_{t-1-m}^\top\hat{\Sigma}_X\phi_{t-1-m}}{\phi_{t-1-m}^\top\hat{\Sigma}_X\phi^\star + \phi_{t-1-m}^\top\frac{X^\top U}{\|\beta\|_2 n}}.$$

The proof of this result is provided in the Appendix. The bias term γ_n consists of two parts: κ_n, which is observable (and therefore correctable), and τ_n, which is unobservable. In the proof, we show that both terms converge to zero, although we cannot determine their exact rates of convergence. From Assumption (16) (which can be verified in the sub-Gaussian, see Proposition 1), we conclude that γ_n converges to zero at a slower rate than $n^{-1/2}$.

The term κ_n accounts for the effects of normalization in Step 3.c) and truncation in Step 3.e) of Algorithm 2. The term τ_n arises primarily from the difference in supports between $\hat{\phi}(t)$ and $\phi^\star$ and can be seen as a bias introduced by regularizing the inverse problem posed by the ill-conditioned matrix $\hat{\Sigma}_X$ when estimating $\phi^\star$ in HD settings.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

The proof of Theorem 3 is based on three propositions: Propositions A3–A5 stated and proved in the Appendix. Proposition A3 provides conditions for the convergence in probability of a stochastic Cauchy product and may be of independent interest. Propositions A4 and A5 analyze LR in depth and are also discussed in the Appendix.

4. MONTE CARLO EXPERIMENTS

The performances of LR are compared against the cross-validated Lasso, as well as two well-studied l_0-constrained estimators of β: the SDAR (Huang et al., 2018) and the two-stage IHT (Jain et al., 2014).

4.1. Setup

The following experiment is reproduced 100 times. In each experiment, we set $n = 100, p = 1,000$ and generate $X \in \mathbb{R}^{n\times p}$ with i.i.d. rows such that $X_1 \sim \mathcal{N}_p(0, \Sigma)$. We generate $Y|X \in \mathbb{R}^n$ according to $Y \sim \mathcal{N}_n(X\beta, \beta^\top \Sigma\beta/m)$, where m represents SNR (Hastie et al., 2020). The vector β is generated with sparsity $\|\beta\|_0 = s = 10$, where each element of β has an equal probability of being assigned a value of 1.

We consider three design setups, ranging from the least correlated design to a highly correlated design.

SETUP 1. *The covariates are uncorrelated, with an identity covariance matrix* $\Sigma = I_{p\times p}$.

SETUP 2. *The covariates have an exponentially decaying correlation structure, with* $\Sigma_{i,j} = 0.5^{|i-j|}$.

SETUP 3. *The covariates follow a high-correlation structure. To construct* Σ, *we set* $W \in \mathbb{R}^{p\times p}$ *such that* $W_{i,j} = 0.9^{|i-j|}$ *and define* $\Sigma = U^\top DU$, *where* U *is the matrix of eigenvectors of* W, $D_{j,j} = 10/\sqrt{j}$, *and* $D_{i,j} = 0$ *for* $i \neq j$.

As noted in Hastie et al. (2020), the performance of estimators can vary significantly with the SNR. An estimator may outperform others at high SNR levels, while underperforming in low SNR settings. Therefore, we consider a range of ten SNRs across the three scenarios. Since the relative performance of the algorithms often changes around an SNR level of 1, we select five values between 0.1 and 1, and four values strictly greater than 1 and below 6 (see Table 1). For each competitor, the sparsity level of the estimated vector is chosen using a ten-fold cross-validation.

4.2. Accuracy metrics

The performance of the estimators was evaluated on three features. First, the performance of the estimators as approximations of the population target vector

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

Table 1. Signal-to-noise ratios considered in the four scenarios

0.1	0.2	0.6	0.8	1.00	2.00	3.00	4.00	5.00	6.00

(the *estimation metric*). Second, the performance of the estimators as predictors (*the prediction metric*). Third, the performance of the estimator as a classifier, that is, their ability to recover the support of the population target vector (the *selection metrics*). The following describes each metric for $\hat{\beta}$ an estimate of β the population target vector.

- The *estimation metric* measures the l_2 error of estimation of the direction target vector $\phi^\star = \beta/\|\beta\|_2$ by measuring $\hat{\beta}/\|\hat{\beta}\|_2 - \beta/\|\beta\|_2$. The results for this metric are provided in Table 2.
- The *prediction metric* measures the relative risk (e.g., Bertsimas et al., 2016, Hastie et al., 2020), that is, $\mathbb{E}[X_1^\top \hat{\beta} - X_1^\top \beta]^2/\mathbb{E}[X_1^\top \beta]^2 = (\hat{\beta} - \beta)^\top \Sigma(\hat{\beta} - \beta)/\beta^\top \Sigma\beta$. The results for these metrics are provided in Table 3.
- For *selection metrics*, we acknowledge the trade-off between type I and type II errors and consider two metrics. Let $S^\star$ be the support of β and $\hat{S}$ be the support of $\hat{\beta}$. Let TP be the cardinality of $S^\star \cap \hat{S}$ (the number of true positives), let FP be the cardinality of $(S^\star)^C \cap \hat{S}$ (the number of fake positives) and FN be the cardinality of $S^\star \cap \hat{S}^C$ (the number of fake negatives). The *precision* metric records $TP/(TP + FP)$. This metric measures the precision of the selection: It gives a perfect score of 1 to every estimated support that is a subset of the target support. The *recall* metric records $TP/(TP + FN)$. This second metric measures how complete was the selection, giving a perfect score of 1 if the target support is a subset of the estimated support. Both metrics are recorded for each replication of the experiments.

The performances of the estimators are compared in terms of the frequency at which they attained a given score. In Tables 4–6, the results for Setups 1–3 are provided. In each table, the frequency of a perfect score, as well as the score attained at the third, second, and first quantiles are exposed.

4.3. Implementations of the algorithms

For reproducibility, the code that generated the results of the experiment is provided on the website of the corresponding author. In the following paragraphs, the particular implementation of each algorithm is described.

- Loaded-Rifle. As suggested theoretically in Example 1, Algorithm 1 is initialized with the solution of a ten-fold cross-validated Lasso if the solution of the former is the nonzero vector. When the cross-validated Lasso is the zero vector, Algorithm 1 is initialized with a solution of Lasso tuned for having $\overline{k}$ nonzero components (in the experiment, we fixed $\overline{k} = 36$). The sparsity level of the output

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

TABLE 2. Summary of direction error per method and SNR

Method	SNR=0.1	SNR=0.2	SNR=0.6	SNR=0.8	SNR=1.0	SNR=2.0	SNR=3.0	SNR=4.0	SNR=5.0	SNR=6.0
Scenario 1										
LR	1.391 (0.033)	1.342 (0.064)	1.235 (0.089)	1.192 (0.100)	1.149 (0.121)	0.947 (0.201)	**0.778** (0.236)	**0.623** (0.266)	**0.457** (0.252)	**0.372** (0.248)
Lasso	**1.058** (0.143)	**1.038** (0.110)	**1.030** (0.082)	**1.038** (0.088)	**1.028** (0.098)	**0.932** (0.162)	0.820 (0.201)	0.688 (0.204)	0.568 (0.182)	0.531 (0.207)
SDAR	1.397 (0.063)	1.373 (0.091)	1.296 (0.122)	1.256 (0.134)	1.238 (0.140)	1.087 (0.213)	0.942 (0.281)	0.745 (0.337)	0.512 (0.334)	0.401 (0.323)
IHT2	1.409 (0.034)	1.400 (0.058)	1.365 (0.098)	1.332 (0.116)	1.314 (0.125)	1.188 (0.167)	1.075 (0.223)	0.925 (0.297)	0.759 (0.340)	0.663 (0.328)
Scenario 2										
LR	1.386 (0.043)	1.342 (0.068)	1.251 (0.090)	1.197 (0.099)	1.164 (0.126)	0.978 (0.208)	**0.773** (0.214)	**0.654** (0.221)	**0.480** (0.251)	**0.422** (0.232)
Lasso	**1.052** (0.134)	**1.056** (0.127)	**1.049** (0.115)	**1.049** (0.090)	**1.045** (0.096)	**0.945** (0.172)	0.816 (0.190)	0.722 (0.189)	0.576 (0.200)	0.527 (0.190)
SDAR	1.405 (0.045)	1.363 (0.100)	1.285 (0.126)	1.260 (0.125)	1.251 (0.118)	1.130 (0.193)	0.925 (0.302)	0.822 (0.330)	0.551 (0.354)	0.454 (0.308)
IHT2	1.404 (0.048)	1.395 (0.067)	1.352 (0.107)	1.332 (0.116)	1.304 (0.126)	1.205 (0.177)	1.090 (0.210)	0.959 (0.249)	0.770 (0.318)	0.636 (0.285)
Setup 3										
LR	1.394 (0.032)	1.361 (0.061)	1.296 (0.102)	1.256 (0.117)	1.231 (0.139)	1.070 (0.159)	**0.972** (0.210)	**0.875** (0.200)	**0.770** (0.215)	**0.677** (0.216)
Lasso	**1.040** (0.122)	**1.059** (0.134)	**1.166** (0.171)	**1.119** (0.148)	**1.143** (0.151)	**1.063** (0.152)	0.975 (0.193)	0.887 (0.175)	0.789 (0.199)	0.713 (0.196)
SDAR	1.409 (0.034)	1.385 (0.080)	1.317 (0.120)	1.312 (0.123)	1.296 (0.136)	1.195 (0.142)	1.103 (0.175)	0.996 (0.212)	0.895 (0.259)	0.824 (0.248)
IHT2	1.407 (0.042)	1.385 (0.080)	1.346 (0.108)	1.348 (0.109)	1.345 (0.110)	1.241 (0.128)	1.185 (0.190)	1.042 (0.251)	0.921 (0.308)	0.826 (0.287)

Note: Bold indicates best per column.

TABLE 3. Summary of relative risk per method and SNR

Method	SNR=0.1	SNR=0.2	SNR=0.6	SNR=0.8	SNR=1.0	SNR=2.0	SNR=3.0	SNR=4.0	SNR=5.0	SNR=6.0
Setup 1										
LR	1.080 (0.028)	1.038 (0.053)	**0.952** (0.068)	**0.920** (0.075)	**0.890** (0.087)	**0.764** (0.111)	0.676 (0.114)	0.612 (0.103)	0.553 (0.083)	0.531 (0.080)
Lasso	**1.069** (0.329)	**1.006** (0.045)	0.984 (0.048)	0.967 (0.077)	0.937 (0.109)	0.777 (0.190)	**0.651** (0.192)	0.532 (0.206)	0.421 (0.176)	0.385 (0.205)
SDAR	3.317 (1.319)	1.476 (0.251)	1.158 (0.205)	1.089 (0.186)	1.089 (0.233)	0.875 (0.190)	0.726 (0.278)	**0.531** (0.340)	**0.314** (0.303)	**0.218** (0.280)
IHT2	2.729 (0.879)	1.315 (0.155)	1.135 (0.123)	1.102 (0.130)	1.081 (0.162)	0.952 (0.152)	0.838 (0.210)	0.683 (0.289)	0.532 (0.334)	0.429 (0.310)
Setup 2										
LR	1.067 (0.042)	1.020 (0.059)	**0.940** (0.075)	**0.901** (0.074)	**0.877** (0.089)	0.764 (0.120)	0.658 (0.096)	0.606 (0.086)	0.551 (0.077)	0.535 (0.066)
Lasso	**1.052** (0.310)	**1.014** (0.081)	0.979 (0.065)	0.954 (0.079)	0.931 (0.106)	**0.751** (0.190)	**0.632** (0.210)	**0.526** (0.186)	0.386 (0.176)	0.360 (0.189)
SDAR	3.347 (1.397)	1.481 (0.393)	1.114 (0.176)	1.079 (0.164)	1.063 (0.227)	0.924 (0.216)	0.701 (0.313)	0.585 (0.325)	**0.334** (0.320)	**0.238** (0.264)
IHT2	2.660 (1.668)	1.338 (0.186)	1.115 (0.137)	1.081 (0.123)	1.039 (0.121)	0.960 (0.148)	0.836 (0.212)	0.707 (0.254)	0.517 (0.299)	0.387 (0.270)
Setup 3										
LR	**1.002** (0.069)	**0.917** (0.071)	**0.834** (0.079)	**0.789** (0.070)	**0.788** (0.094)	0.671 (0.099)	0.626 (0.105)	0.561 (0.064)	0.533 (0.072)	0.500 (0.056)
Lasso	1.070 (0.432)	0.996 (0.103)	0.884 (0.137)	0.840 (0.182)	0.826 (0.171)	**0.596** (0.178)	**0.524** (0.177)	**0.420** (0.122)	**0.351** (0.151)	**0.301** (0.117)
SDAR	2.704 (0.871)	1.294 (0.256)	0.999 (0.173)	0.962 (0.171)	0.955 (0.159)	0.798 (0.154)	0.703 (0.193)	0.555 (0.208)	0.440 (0.237)	0.382 (0.210)
IHT2	2.351 (0.788)	1.240 (0.247)	1.048 (0.167)	1.016 (0.143)	0.995 (0.113)	0.874 (0.140)	0.816 (0.179)	0.648 (0.228)	0.516 (0.262)	0.424 (0.237)

Note: Bold indicates best per column.

TABLE 4. Experiment results for the two selection metrics (first setup)

Setup 1								
Method	LR		Lasso		SDAR		IHT2	
Metric	Precision	Recall	Prec.	Reca.	Prec.	Reca.	Prec.	Reca.
Signal-to-noise ratio = 0.1								
$f(1)$					**0.07**		0.02	
Q_3	**0.032**	**0.1**						
Q_2								
Q_1								
Signal-to-noise ratio = 0.6								
$f(1)$	0.11		0.06		**0.45**		0.2	
Q_3	0.148	**0.4**	**1**		**1**	0.1		
Q_2	0.111	**0.3**	**0.4**					
Q_1	0.078	**0.2**	**0.308**					
Signal-to-noise ratio = 1								
$f(1)$	0.13		0.04		**0.56**		0.39	
Q_3	0.25	**0.5**	0.596	0.35	**1**	0.1	**1**	0.1
Q_2	0.155	**0.4**	0.5		**1**	0.1		
Q_1	0.116	**0.3**	**0.237**					
Signal-to-noise ratio = 3								
$f(1)$	0.33	**0.15**	0.07	0.13	**0.75**	0.04	0.61	
Q_3	**1**	**0.9**	0.545	**0.9**	**1**	0.55	**1**	0.4
Q_2	0.596	**0.75**	0.4	0.7	**1**	0.2	**1**	0.1
Q_1	0.295	**0.5**	0.304	**0.5**	**0.95**	0.1	0.667	0.1
Signal-to-noise ratio = 5								
$f(1)$	0.45	**0.64**	0.01	0.57	0.55	0.37	**0.57**	0.09
Q_3	**1**	**1**	0.414	**1**	**1**	**1**	**1**	0.8
Q_2	0.75	**1**	0.345	**1**	**1**	0.85	**1**	0.6
Q_1	0.297	**0.9**	0.276	0.8	**0.866**	0.6	0.76	0.2

Note: Rows $f(1)$ and Q_i $(i = 1, 2, 3)$ show the frequency of reaching 1 and the third, second, and first quantiles. Blank cells = 0, and bold cells are the best values.

of Algorithm 1 is then selected through an additional ten-fold cross-validation. For each instance of LR, we set the *step size* $\eta = 0.5/\|\hat{\Sigma}_X\|_{\text{spec}}$.

- Two-Stage IHT. The two-stage IHT depends on a parameter fixing the dimension of a larger support on which the solution is pre-optimized. For comparability, we also fixed this parameter to $k < \bar{k} = 38$, where k is selected by cross-validation.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

TABLE 5. Experiment results for the two selection metrics (second setup)

Setup 2								
Method	LR		Lasso		SDAR		IHT2	
Metric	Precision	Recall	Prec.	Reca.	Prec.	Reca.	Prec.	Reca.
Signal-to-noise ratio = 0.1								
$f(1)$	0.01		0.02		0.03		**0.04**	
Q_3	0.033	**0.1**	**0.048**					
Q_2	**0.024**	**0.1**						
Q_1								
Signal-to-noise ratio = 0.6								
$f(1)$	0.06				**0.51**		0.25	
Q_3	0.145	**0.4**	0.388		**1**	0.1	0.667	0.1
Q_2	0.106	**0.3**	0.275		**1**	0.1		
Q_1	0.076	**0.2**	**0.113**					
Signal-to-noise ratio = 1								
$f(1)$	0.17		0.08		**0.63**		0.43	
Q_3	0.351	**0.5**	0.679	0.3	**1**	0.1	**1**	0.1
Q_2	0.181	**0.4**	0.4		**1**	0.1		
Q_1	0.118	**0.2**	**0.296**					
Signal-to-noise ratio = 3								
$f(1)$	0.25	0.15	0.03	**0.16**	**0.63**	0.05	**0.63**	
Q_3	0.955	**0.9**	0.5	**0.9**	**1**	0.6	**1**	0.4
Q_2	0.353	**0.8**	0.333	0.7	**1**	0.3	**1**	0.1
Q_1	0.25	**0.65**	0.244	0.6	**0.732**	0.1	0.683	0.1
Signal-to-noise ratio = 5								
$f(1)$	0.37	**0.62**		0.6	0.53	0.34	**0.68**	0.06
Q_3	**1**	**1**	0.405	**1**	**1**	**1**	**1**	0.8
Q_2	0.388	**1**	0.328	**1**	**1**	0.9	**1**	0.6
Q_1	0.28	**0.9**	0.246	**0.9**	0.8	0.5	**0.882**	0.3

Note: Rows $f(1)$ and Q_i $(i = 1, 2, 3)$ show the frequency of reaching 1 and the third, second, and first quantiles. Blank cells = 0, and bold cells are the best values.

- SDAR. Following the theoretical and practical recommendations in Huang et al. (2018), SDAR is initialized with the solution of a ridge regression with a tuning parameter $2s\log(p)/n$. Moreover, following the same recommendations, we preprocessed SDAR by dividing X and Y by $\sqrt{n}$ (it appears in our experiments that SDAR is particularly sensitive to this transformation).

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

TABLE 6. Experiment results for the two selection metrics (third setup)

Setup 3								
Method	LR		Lasso		SDAR		IHT2	
Metric	Precision	Recall	Prec.	Reca.	Prec.	Reca.	Prec.	Reca.
Signal-to-noise ratio = 0.1								
$f(1)$					0.02		**0.03**	
Q_3	**0.032**	**0.1**						
Q_2								
Q_1								
Signal-to-noise ratio = 0.6								
$f(1)$	0.1		0.02		**0.32**		0.24	
Q_3	0.195	**0.2**	0.333	0.15	**1**	0.1	0.5	0.1
Q_2	0.098	**0.1**	**0.133**					
Q_1			**0.031**					
Signal-to-noise ratio = 1								
$f(1)$	0.1		0.01		**0.32**		0.26	
Q_3	0.477	**0.3**	0.333	**0.3**	**1**	0.1	**1**	0.1
Q_2	0.17	**0.2**	**0.227**	0.1				
Q_1	0.065	**0.1**	**0.143**					
Signal-to-noise ratio = 3								
$f(1)$	0.04			**0.02**	0.29		**0.3**	
Q_3	0.6	0.7	0.311	**0.8**	**1**	0.4	**1**	0.2
Q_2	0.375	0.5	0.25	**0.6**	**0.667**	0.2	0.5	0.1
Q_1	0.279	0.3	0.179	**0.4**	**0.4**	0.1		
Signal-to-noise ratio = 5								
$f(1)$	0.05	0.23	0.01	**0.27**	0.14	0.02	**0.32**	0.03
Q_3	0.588	0.9	0.309	**1**	0.833	0.7	**1**	0.6
Q_2	0.405	**0.8**	0.25	**0.8**	0.643	0.5	**0.764**	0.3
Q_1	0.302	**0.6**	0.189	**0.6**	0.408	0.3	**0.5**	0.2

Note: Rows $f(1)$ and Q_i $(i = 1, 2, 3)$ show the frequency of reaching 1 and the third, second, and first quantiles. Blank cells = 0, and bold cells are the best values.

4.4. Results

Our evaluation of the LR method in estimating, prediction, and selection metrics in various experimental scenarios positions it as an effective, competitive, and often superior option for l_0 estimation, particularly in settings requiring both reliable variable selection and accurate estimation, even under limited SNR conditions.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

In terms of estimation accuracy, LR generally exhibits a favorable trend, with relative errors decreasing as SNR increases. In particular, for SNR values greater than 2, LR consistently achieves the lowest estimation errors, underscoring its effectiveness compared to other methods. Furthermore, the robustness of the LR estimates is reflected in its lower standardized errors compared to other estimators l_0. These results highlight the suitability of LR variants for accurate parameter estimation tasks when moderate to high SNRs are attainable.

Prediction Accuracy: Prediction accuracy, measured by relative risk, reveals another strength of the LR. As soon as the SNR becomes sufficiently high for Lasso to outperform the null model (i.e., achieving a relative risk lower than that of the zero vector), but remains below one, LR surpasses all competing methods. This advantage is even more pronounced when covariates are highly correlated: in the third experimental setup, LR outperforms all competitors across all SNRs below one. These results make LR particularly appealing for predictive tasks in low-SNR and high-correlation environments, such as financial modeling (Hastie et al., 2020).

Selection Accuracy: The selection metrics clearly show that LR frequently outperforms its competitors in terms of recall. While other methods tend to prioritize precision, reducing false positives at the cost of increasing false negatives, LR adopts a more balanced strategy. Starting from a Lasso solution, it simultaneously reduces false positives and increases true positives. This behavior enables LR to recover part or all of the true support, even in low SNR regimes. Its superior performance in variable selection probably contributes to its advantages in both estimation and prediction. These findings support the adoption of LR in sparse recovery applications, where the accurate identification of relevant variables is critical.

In summary, LR demonstrates significant improvements in the accuracy of estimation, prediction, and selection over standard estimators l_0 and cross-validated Lasso. Its robustness in multiple metrics and SNR conditions suggests that LR variants offer substantial benefits for applications requiring strong statistical performance, particularly in domains characterized by low SNRs.

5. CONCLUSION

This article proposes interpreting the parameter vector of the HDSLM as the product of a *direction of interest*, identified as the generalized eigenspace of a pair of measurable matrices, and a magnitude relative to this direction. This perspective offers new insights into linear regression and introduces a parameterization naturally aligned with several core ideas in economic theory.

For example, in a *Cobb–Douglas production function*, the logarithmic output Y can be written as $Y = X^\top \phi \nu$, where ϕ specifies factor shares summing to one and ν reflects returns to scale. In the *Cambridge equation* (Fisher and Brown, 1911; Pigou, 1917), Y may represent money demand, X a vector of goods, ϕ a price vector, and ν the inverse of the velocity of money. In a currency context, Y could

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

be a dollar amount, X quantities of goods, and ϕ prices in euros, with ν representing the exchange rate. These examples illustrate that the choice of ϕ is context-specific and that normalizations, such as $\|\phi\| = 1$, are possible but not unique.

Beyond its interpretive value, our approach connects linear regression to the GEP, expanding the class of estimators for linear models to include those suited to HD sparse settings. We exploit this connection by addressing the best subset selection problem through what we call the *directional best subset selection*, for which we establish a theoretical link to the classical formulation and derive the minimax convergence rate in the l_2 norm. Building on this, we adapt the RIFLE algorithm Tan et al., 2018 to construct an efficient estimator, LR, which achieves the minimax rate under sub-Gaussian assumptions and admits a CLT. To obtain this result, Proposition A3 in the Appendix provides sufficient conditions for the convergence in probability of a stochastic Cauchy product; this result may be of independent interest. These results are derived under the assumption of an *exact* sparsity structure for the parameter of interest β. While common in the theoretical literature, this assumption can be overly restrictive in practical applications. A natural direction for future research is to extend our results to more general forms of approximate sparsity. Another important avenue is to address the unobservable bias identified in the CLT, either by characterizing its exact rate of convergence or by constructing a debiased version of the estimator.

Simulation results show that LR performs competitively, often outperforming existing l_0-constrained estimators in estimation, variable selection, and prediction, while remaining computationally efficient even for large-scale problems. This combination of theoretical optimality and practical speed is rare in the current literature.

While our focus has been on l_0 constraints and the RIFLE-based estimator, the identification result in Theorem 1 has broader implications. It suggests that other GEP estimators (e.g., Safo et al., 2018; Jung, Ahn, and Jeon, 2019; Yuan, Shen, and Zheng, 2019) could be adapted to least squares problems constrained by other norms. More generally, the connection between least squares and GEP opens avenues for incorporating advanced numerical methods into HD estimation.

The scope of the theory presented in this article relies on several strong assumptions. Most notably, we work under an i.i.d. linear model. These assumptions are central to the spectral structure exploited by the GEP and to the identification arguments underlying our main results. Relaxing them—for instance, to allow for dependence, heteroskedasticity, or nonlinear structural relationships—would raise nontrivial theoretical challenges and falls outside the scope of the present analysis. Finally, although we have assumed exogeneity of the model throughout, extending Theorem 1 to instrumental variable (IV) settings remains a promising direction. Recent work (e.g., Belloni et al., 2012; Fan and Liao, 2014; Florens and Van Bellegem, 2015; Chernozhukov, Hansen, and Spindler, 2015; Gold, Lederer, and Tao, 2020) highlights the challenges of endogeneity in high dimensions. Extending our framework to IV models or systems of simultaneous equations, possibly allowing for vector-valued dependent variables and higher-rank GEPs, could substantially broaden its scope and impact.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

COMPETING INTEREST STATEMENT

The authors declare no competing interests exist.

FUNDING STATEMENT

The authors received no specific funding for the work described in this article.

SUPPLEMENTARY MATERIAL

Sauvenier, M., & Van Bellegem, S. (2025). Supplementary Material on "Direction Identification and Minimax Estimation in High-Dimensional Sparse Regression via a Generalized Eigenvalue Approach." To view, please visit: https://doi.org/10.1017/S0266466626100334.

REFERENCES

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6), 2369–2429.

Belloni, A., Chernozhukov, V., & Hansen, C. B. (2013). *Inference for high-dimensional sparse econometric models*, volume 3 of Econometric Society Monographs (pp. 245–295). Cambridge University Press. https://doi.org/10.1017/CBO9781139060035.008

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2), 813–852. https://doi.org/10.1214/15-AOS1388

Box, G. E. P., & Daniel Meyer, R. (1986). An analysis for unreplicated fractional factorials. *Technometrics. A Journal of Statistics for the Physical, Chemical and Engineering Sciences*, 28(1), 11–18. https://doi.org/10.2307/1269599

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1), 33–61. https://doi.org/10.1137/S1064827596304010

Chen, X., Zou, C., & Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Annals of Statistics*, 38(6), 3696–3723. https://doi.org/10.1214/10-AOS826

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1), 649–688.

Chetverikov, D. (2024). *Tuning parameter selection in econometrics*. Preprint, arXiv:2405.03021.

Chetverikov, D., Liao, Z., & Chernozhukov, V. (2021). On cross-validated Lasso in high dimensions. *Annals of Statistics*, 49(3), 1300–1317. https://doi.org/10.1214/20-aos2000

Dennis Cook, R. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425), 177–189.

Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Proceedings of the international congress of mathematicians* (Vol. 3, pp. 595–622). Zurich : European Mathematical Society.

Fan, J., & Liao, Y. (2014). Endogeneity in high dimensions. *Annals of Statistics*, 42(3), 872.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849–911.

Fisher, I., & Brown, H. G. (1911). *The purchasing power of money*. Macmillan.

Florens, J.-P., & Van Bellegem, S. (2015). Instrumental variable estimation in functional linear models. *Journal of Econometrics*, 186(2), 465–476.

van de Geer, S. (2016). *Estimation and testing under sparsity*, volume 2159 of Lecture Notes in Mathematics. Springer. https://doi.org/10.1007/978-3-319-32774-7. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

Ghojogh, B., Karray, F., & Crowley, M. (2019). *Eigenvalue and generalized eigenvalue problems: Tutorial*. Preprint.

Gold, D., Lederer, J., & Tao, J. (2020). Inference for high-dimensional instrumental variables regression. *Journal of Econometrics*, 217(1), 79–111.

Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press.

Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under $l_1$constraint. *Annals of Statistics*, 34(5), 2367–2386. https://doi.org/10.1214/009053606000000768.

Guo, Y., Zhu, Z., & Fan, J. (2020). *Best subset selection is robust against design dependence*. Preprint, arXiv:2007.01478.

Hastie, T., Tibshirani, R., & Tibshirani, R. (2020). Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4), 579–592.

Hazimeh, H., & Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5), 1517–1537.

Huang, J., Jiao, Y., Liu, Y., & Lu, X. (2018). A constructive approach to L_0 penalized regression. *Journal of Machine Learning Research*, 19, Article no. 10, 37 pp.

Jain, P., Tewari, A., & Kar, P. (2014). On iterative hard thresholding methods for high-dimensional m-estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc.

Jung, S., Ahn, J., & Jeon, Y. (2019). Penalized orthogonal iteration for sparse estimation of generalized eigenvalue problem. *Journal of Computational and Graphical Statistics*, 28(3), 710–721. https://doi.org/10.1080/10618600.2019.1568014

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, 94(3), 603–613. https://doi.org/10.1093/biomet/asm044

Moghaddam, B., Weiss, Y., & Avidan, S. (2006). Generalized spectral bounds for sparse LDA. In *Proceedings of the 23rd international conference on machine learning, ICML '06* (pp. 641–648). Association for Computing Machinery. https://doi.org/10.1145/1143844.1143925

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2), 227–234. https://doi.org/10.1137/S0097539792240406

Parthasarathy, K. R. (1977). *Introduction to probability and measure* (1st ed.) The Macmillan Press.

Pigou, A. C. (1917). The value of money. *The Quarterly Journal of Economics*, 32(1), 38–65. https://doi.org/10.2307/1885078

Raskutti, G., Wainwright, M. J., & Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10), 6976–6994. https://doi.org/10.1109/TIT.2011.2165799

Safo, S. E., Ahn, J., Jeon, Y., & Jung, S. (2018). Sparse generalized eigenvalue problem with application to canonical correlation analysis for integrative analysis of methylation and gene expression data. *Biometrics. Journal of the International Biometric Society*, 74(4), 1362–1371. https://doi.org/10.1111/biom.12886

Shen, X., Pan, W., Zhu, Y., & Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5), 807–832. https://doi.org/10.1007/s10463-012-0396-3

Stewart, G. W. (1979). Pertubation bounds for the definite generalized eigenvalue problem. *Linear Algebra and its Applications*, 23, 69–85. https://doi.org/10.1016/0024-3795(79)90094-6.

Sun, L., Ji, S., & Ye, J. (2009). A least squares formulation for a class of generalized eigenvalue problems in machine learning. In *Proceedings of the 26th annual international conference on*

machine learning, ICML '09 (pp. 977–984). Association for Computing Machinery. https://doi.org/10.1145/1553374.1553499.

Tan, K. M., Wang, Z., Liu, H., & Zhang, T. (2018). Sparse generalized eigenvalue problem: Optimal statistical rates via truncated Rayleigh flow. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 80(5), 1057–1086. https://doi.org/10.1111/rssb.12291

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1), 267–288.

Yuan, G., Shen, L., & Zheng, W.-S. (2019). A decomposition algorithm for the sparse generalized eigenvalue problem. In G. Yuan, L. Shen & W.-S. Zheng (Eds.), *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (pp. 6113–6122).

Zhang, C.-H., & Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 27(4), 576–593. https://doi.org/10.1214/12-STS399

Zhang, Y., Wainwright, M. J., & Jordan, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In Balcan, M. F., Feldman, V. & Szepesvári, C. (Eds.), *Proceedings of the 27th conference on learning theory* (volume 35, pp. 921–948). PMLR.

A. APPENDIX

This Appendix presents the proofs of all the results, except for the proof of Theorem 1, which is provided in Section 2. It also includes a detailed description of the RIFLE algorithm. The material is organized in the same order as in the main text. Technical lemmas and several standard auxiliary results used in the proofs are stated here, while their complete proofs are deferred to the Supplementary Material.

We note that the proof of Theorem 2, while correcting and extending the results in Tan et al. (2018), employs the same techniques as in that paper and is therefore provided in the Supplementary Material. The proof of Proposition A2 relies on standard arguments as well, and is likewise presented in the Supplementary Material.

A.1. Proof of Theorem 1

Let $\mathcal{S}$ be the solution set of Problem (7) and observe that it is closed under scalar multiplication. Let $\|\phi\|_{\Sigma_X} := \phi^\top \Sigma_X \phi$, that is, the norm induced by Σ_X and $\mathcal{S}_1 := \{\phi \in \mathcal{S} \text{ s.t. } \phi = 1\}$ and $\mathcal{S}_2 := \{\phi \in \mathcal{S} \text{ s.t. } \|\phi\|_{\Sigma_X} \neq 1\}$, hence we have

$\mathcal{S} = \mathcal{S}_1 \sqcup \mathcal{S}_2$ (disjoint union),

if $\phi \in \mathcal{S}_2$, then $\dfrac{\phi}{\|\phi\|_{\Sigma_X}} \in \mathcal{S}_1$, and

if $\phi \in \mathcal{S}_1$, then $\alpha\phi \in \mathcal{S}_2$ for all $\alpha \neq 1$.

Consider now the following problem whose solution set is equal to $\mathcal{S}_1$:

$$\text{argmax}_{\phi \in \mathbb{R}^p} \phi^\top \Sigma_{XY} \Sigma_{YX} \phi \qquad \text{such that } \|\phi\|^2_{\Sigma_X} = 1. \tag{A.1}$$

By the definition of $\|.\|_{\Sigma_X}$, the Lagrangian of problem (A.1) is

$$\mathcal{L} = \phi^\top \Sigma_{XY} \Sigma_{YX} \phi - \lambda(\phi^\top \Sigma_X \phi - 1).$$

Computing $\nabla_\phi \mathcal{L} = 0$ leads to

$$\Sigma_{XY} \Sigma_{YX} \phi = \Sigma_X \phi \lambda, \tag{A.2}$$

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

which is a scalar equation of the GEP of the matrix pair $\left(\Sigma_{XY}\Sigma_{YX}, \Sigma_X\right)$. Pre-multiplying each side of this equation by $\phi^\top$ identifies λ as the maximum generalized eigenvalue, which is why we symbolize it by $\lambda_{\max}$. Observe that if ϕ satisfies (A.2), then $\alpha\phi$ also satisfies this equation for any $\alpha \in \mathbb{R}$. This last remark allows us to conclude that

$$\mathrm{Ker}\left(\Sigma_{XY}\Sigma_{YX} - \lambda_{\max}\Sigma_X\right) = \mathcal{S}_1 \sqcup \mathcal{S}_2 = \mathcal{S},$$

which is the result to be proven. □

A.2. Statement and Proof of Proposition A1

The following proposition confirms that the solution set $\hat{\Phi}$ identifies the direction of the best subset selection solution.

PROPOSITION A1. *Let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$. For a fixed integer $k \in \mathbb{N}$, let $\hat{\beta}$ denote a solution of the best subset selection problem defined in (3), and let $\hat{\phi}$ denote a solution of the directional best subset selection defined in (4). Then, conditionally on the uniqueness of $\hat{\beta}$ and on $\min_{v\in\mathbb{R}^p,\, \|v\|_0\leq k} v^\top \hat{\Sigma}_X v > 0$, it holds $\hat{\beta}/\|\hat{\beta}\|_2 = \hat{\phi}/\|\hat{\phi}\|_2$ almost surely. Furthermore, if $\hat{\upsilon} = \mathrm{argmin}_{m\in\mathbb{R}} n^{-1}\|Y - X\hat{\phi}m\|_2^2$, then $\hat{\phi}\hat{\upsilon} = \hat{\beta}$ almost surely.*

Proof. Let $\mathcal{O}(p,k) := \{v \in \mathbb{R}^p; \|v\|_0 = k\}$. Observe that for any $v \in \mathcal{O}(p,k)$, there exists an infinity of $(\phi, m) \in \mathcal{O}(p,k)\times\mathbb{R}$ such that $\phi m = v$. Hence, note that if $(\hat{\phi}, \hat{m}) \in \mathrm{argmin}_{(\phi,m)\in\mathcal{O}(p,k)\times\mathbb{R}} n^{-1}\|Y - X\phi m\|_2^2$, then $\hat{\beta} = \hat{\phi}\hat{m}$ almost surely. Now, observe that *for any $\phi \in \mathcal{O}(p,k)$*, we have

$$\mathrm{argmin}_{m\in\mathbb{R}} \frac{1}{n}\|Y - X\phi m\|_2^2 = \frac{\phi^\top X^\top Y}{\phi^\top X^\top X\phi}.$$

Then, in particular, if $\hat{\phi} \in \mathrm{argmin}_{\phi\in\mathcal{O}(p,k)} n^{-1}\|Y - (X\phi\phi^\top XY)/\phi^\top X^\top X\phi)\|_2^2$, we have $\hat{\beta}/\|\hat{\beta}\|_2 = \hat{\phi}/\|\hat{\phi}\|_2$ *a.s.* Finally, computations similar to the ones in the proof of Theorem 1 lead to

$$\mathrm{argmin}_{\hat{\phi}\in\mathcal{O}(p,k)} \frac{1}{n}\|Y - \frac{X\hat{\phi}\hat{\phi}^\top XY}{\hat{\phi}^\top X^\top X\hat{\phi}}\|_2^2 = \mathrm{argmax}_{\hat{\phi}\in\mathcal{O}(p,k)} \frac{\hat{\phi}^\top X^\top YY^\top X\hat{\phi}}{\hat{\phi}^\top X^\top X\hat{\phi}n},$$

that gives the result. □

A.3. Algorithm: Truncated Rayleigh flow method (RIFLE)

ALGORITHM 2. Truncated Rayleigh flow method (RIFLE).

Input: *matrices $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ and $\hat{\Sigma}_X$, initial vector ϕ_0, sparsity level $k \in \{1,\ldots,p\}$ and step size η.*

START:

1. $\mathcal{F}_0 =$ *the set of index of the k highest elements of* $|\phi_0|$;
2. $\phi_0 = Truncate(\phi_0, \mathcal{F}_0)$;
3. *Let* $t = 1$ *and repeat the following steps until convergence:*
   (a) $\rho_{t-1} = (\phi_{t-1}^\top \hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1})/(\phi_{t-1}^\top \hat{\Sigma}_X\phi_{t-1})$;
   (b) $C_{t-1} = I_{p\times p} + (\eta/\rho_{t-1})(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \rho_{t-1}\hat{\Sigma}_X)$;

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

(c) $\phi_t' = C_{t-1}\phi_{t-1}/\|C_{t-1}\phi_{t-1}\|_2$;
(d) $\mathcal{F}_t =$ *the set of index of the k highest elements of* $|\phi_t'|$;
(e) $\hat{\phi}_t = Truncate(\phi_t', \mathcal{F}_t)$;
(f) $\phi_t = \hat{\phi}_t/\|\hat{\phi}_t\|_2$;
(g) $t = t+1$;

END
Output: ϕ_t

Tan et al. (2018) provide the computational complexity of the algorithm, that is, the sum between two terms: First, $O(p)$ for selecting the k largest elements of a p-dimensional vector and second, $O(kp)$ to take the product between a truncated vector and a matrix with columns restricted to the set F_t and to calculate the difference between two matrices with columns restricted to the set F_t (Tan et al., 2018).

A.4. Proof of Proposition 1

In order to prove Proposition 1, we need the following proposition.

A.4.1. *Proposition A2.*

PROPOSITION A2. *In model (1), assume that* $Y_1, X_{1,1}, \ldots, X_{1,p(n)}$ *are sub-Gaussian, that* $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ *is positive definite, and that* $\|\beta\|_0 = s$. *Let* $\phi^\star = \beta/\|\beta\|_2$, $\hat{\Sigma}_X = X^\top X/n$, *and* $k' = Cs$ *for* $C > 1$. *If* $s\log(p/s)/n = o(1)$ *and* $s/p = o(1)$, *then*

$$\varepsilon(k') = O_p\Big(\sqrt{\frac{s\log(p/s)}{n}}\Big)$$

and

$$\rho(\hat{\Sigma}_X - \Sigma_X, k') = O_p\Big(\sqrt{\frac{s\log(p/s)}{n}}\Big).$$

Proof. The result follows directly from Lemmas 4 and 5 stated and proved in the Supplementary Material. □

A.4.2. *Proof of Proposition 1.*

First, we note that since $\langle \phi^\star, \hat{\phi}(t)\rangle_2 \geq 0$, then $\sqrt{1-|(\phi^\star)^\top \hat{\phi}(t)|} = 1/\sqrt{2}\|\phi^\star - \hat{\phi}(t)\|_2$. Second, we observe that since we can choose $t = t(n)$, we can choose it sufficiently fast to ensure that for any positive constant M_1, there is a positive constant M and $N_1 \in \mathbb{N}$ such that for all $n > N_1$,

$$\mathbb{P}\Big(\|\phi^\star - \hat{\phi}(t)\|_2 > \nu^t\sqrt{\xi} + \sqrt{\frac{s\log(p/s)}{n}}M_1\Big) \leq \mathbb{P}\Big(\|\phi^\star - \hat{\phi}(t)\|_2 > \sqrt{\frac{s\log(p/s)}{n}}M\Big),$$

where $\nu^t\sqrt{\xi}$ is as in Theorem 2.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

Hence, concentrating on $\mathbb{P}\Big(\|\phi^\star - \hat{\phi}(t)\|_2 > \sqrt{\frac{s\log(p/s)}{n}}M\Big)$, by the Law of Total Probability, for all positive constants c,

$$\mathbb{P}\Big(\|\phi^\star - \hat{\phi}(t)\|_2 > \sqrt{\frac{s\log(p/s)}{n}}M\Big)$$
$$\leq \mathbb{P}\Big(\|\phi^\star - \hat{\phi}(t)\|_2 > \sqrt{\frac{s\log(p/s)}{n}}M \Big| \|\phi^\star - \phi_0\|_2 < c\Big) + \mathbb{P}\Big(\|\phi^\star - \phi_0\|_2 \geq c\Big).$$

If c is such that $|\langle \phi^\star, \phi_0\rangle_2| \geq 1 - \upsilon^2\xi$ as in Theorem 2, then we can choose $t = t(n)$ such that by Proposition A2 and Theorem 2 for every $\epsilon_2 > 0$, there exists an N_2 such that for every $n > N_2$,

$$\mathbb{P}\Big(\|\phi^\star - \hat{\phi}(t)\|_2 > \sqrt{\frac{s\log(p/s)}{n}}M \Big| \|\phi^\star - \phi_0\|_2 < c\Big) \leq \epsilon_1.$$

Moreover, since $\alpha_n = o(1)$, for every $\epsilon_3 > 0$, there exists N_3 such that for every $n \geq N_3$,

$$\mathbb{P}\Big(\|\phi^\star - \phi_0\|_2 \geq c\Big) < \epsilon_3.$$

Hence, for every $\epsilon > 0$, choose $\epsilon_2 = \epsilon_3 = \epsilon/2$, and let $N = N_1 + N_2 + N_3$, and observe that, as intended, for every $n \geq N$,

$$\mathbb{P}\Big(\|\phi^\star - \hat{\phi}(t)\|_2 > \sqrt{\frac{s\log(p/s)}{n}}M\Big) < \epsilon.$$

□

A.5. Proof of Proposition 2

Let $\upsilon = \|\beta\|_2$ and $U := Y - X\beta$, straightforward computations give

$$\hat{\upsilon} = \frac{\hat{\phi}^\top X^\top Y}{\hat{\phi}^\top X^\top X\hat{\phi}} = \upsilon + \frac{\upsilon}{\hat{\phi}^\top \hat{\Sigma}_X \hat{\phi}}\big(\hat{\phi}^\top \hat{\Sigma}_X(\phi^\star - \hat{\phi}) + \frac{\hat{\phi}^\top X^\top U}{n\upsilon}\big),$$

where the second equality is obtained by adding zero and simplifying. Hence, we have an equality for $\hat{\upsilon} - \upsilon$ that we now use to obtain an inequality. Let $P \in \mathcal{P}(\mathcal{F}(p, 2s))$ be defined as in Section 3.1, and $PP^\top\hat{\phi} = \hat{\phi}$ and $PP^\top(\phi^\star - \hat{\phi}) = (\phi^\star - \hat{\phi})$.

Observe that

$$\hat{\phi}^\top \hat{\Sigma}_X(\phi^\star - \hat{\phi}) = \hat{\phi}^\top PP^\top \hat{\Sigma}_X PP^\top(\phi^\star - \hat{\phi}) \leq \overline{\kappa}\|\phi^\star - \hat{\phi}\|_2,$$

where the second equation is obtained by the Cauchy–Schwarz inequality, the submultiplicativity of the spectral norm, and $\|P\|_{\text{spec}} = 1$. Hence,

$$|\hat{\upsilon} - \upsilon| \leq \frac{\upsilon}{\underline{\kappa}}\Big(\overline{\kappa}\|\phi^\star - \hat{\phi}\|_2 + \frac{1}{\upsilon}\|\frac{P^\top X^\top U}{n}\|_2\Big),$$

which implies $|\hat{\upsilon} - \upsilon| = O_p(\alpha_n) + O_p\big(\sqrt{2s/n}\big) = O_p(\alpha_n)$. And since

$$\begin{aligned}\|\hat{\beta} - \beta\|_2^2 &= \|\hat{\upsilon}\hat{\phi} - \upsilon\phi^\star\|_2^2 \\ &= \hat{\upsilon}^2 - 2\hat{\upsilon}\upsilon\langle\hat{\phi}, \phi^\star\rangle_2 + \upsilon^2 \qquad \text{(Since } \|\hat{\phi}\|_2 = \|\phi^\star\|_2 = 1\text{)} \\ &= (\hat{\upsilon} - \upsilon)^2 + 2\hat{\upsilon}\upsilon(1 - \langle\hat{\phi}, \phi^\star\rangle_2)\end{aligned}$$

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

$$= (\hat{\upsilon} - \upsilon)^2 + \hat{\upsilon}\upsilon \|\hat{\phi} - \phi^\star\|_2^2 \qquad \text{(Since } \langle \hat{\phi}, \phi^\star \rangle_2 \geq 0\text{)}$$
$$= (\hat{\upsilon} - \upsilon)^2 + \big((\hat{\upsilon} - \upsilon)\upsilon + \upsilon^2\big)\|\hat{\phi} - \phi^\star\|_2^2.$$

Since $|\hat{\upsilon} - \upsilon| = O_p(\alpha_n)$, $(\hat{\upsilon} - \upsilon)^2 = O_p(\alpha^2)$ and $(\hat{\upsilon} - \upsilon)\upsilon + \upsilon^2 = O_p(1)$, the result follows. □

A.6. Statement and proof of A3

PROPOSITION A3. *Let $\{a_i\}_{i=0}^m$ and $\{b_i\}_{i=0}^m$ be two sequences of positive random variables such that $m = m(n)$, where $m \to \infty$ as $n \to \infty$. If $\{a_i\}_{i=0}^m = o_p(1)$, $\sup_{i\in\{0,\dots,m\}} |a_i| = O_p(1)$ and it exists a finite $M \in \mathbb{R}$ such that $\lim_{n\to\infty} \sum_{i=0}^{m(n)} b_i = M$ almost surely. Then, $\sum_{i=0}^m a_{m-i} b_i = o_p(1)$.*

Proof. Since $\sup_{i\in\{0,\dots,m\}} a_i = O_p(1)$, for all $\epsilon > 0$, there exist $N_1 \in \mathbb{N}$ and $C > 0$ such that for all $m > m(N_1)$ and all $\gamma > 0$, we have

$$\mathbb{P}\Big(\sum_{i=0}^m a_{m-i} b_i > \gamma\Big) \leq \mathbb{P}\Big(\sum_{i=0}^m a_{m-i} b_i > \gamma \,\Big|\, \sup_{i\in\{0,\dots,m\}} a_i \leq C\Big) + \mathbb{P}\Big(\sup_{i\in\{0,\dots,m\}} a_i > C\Big)$$
$$= \mathbb{P}\Big(\sum_{i=0}^m a_{m-i} b_i > \gamma \,\Big|\, \sup_{i\in\{0,\dots,m\}} a_i \leq C\Big) + \epsilon/2.$$

Now, since $\lim_{n\to\infty} \sum_{i=0}^m b_i = M$ a.s., for all $\alpha > 0$, there exists $N_2 \in \mathbb{N}$ such that for all $m \geq m(N_2)$, we have $\sum_{m\geq m(N_2)} b_i \leq \alpha/(2C)$ almost surely. Consequently, we have for any $\gamma > 0$ and all $m > M(N_2)$ that

$$\mathbb{P}\Big(\sum_{i=0}^m a_{m-i} b_i > \gamma \,\Big|\, \sup_{i\in\{0,\dots,m\}} a_i \leq C\Big) \leq \mathbb{P}\Big(\sum_{i=0}^{m(N_2)} a_{m-i} b_i + \frac{\alpha}{2} > \gamma \,\Big|\, \sup_{i\in\{0,\dots,m\}} a_i \leq C\Big).$$

Note now that since $\{a_i\}_{i=0}^m = o_p(1)$, for any $\alpha, \tilde{\epsilon} > 0$, there exists $N_3 \in \mathbb{N}$ such that for all $m > m(N_3)$, we have

$$\mathbb{P}\Big(a_m > \frac{\alpha}{2m(N_2)M}\Big) < \frac{\tilde{\epsilon}}{(m(N_2)+1)},$$

which implies by a union bound that for all $m > m(N_2) + m(N_3)$,

$$\mathbb{P}\Big(\sup_{i\in\{0,\dots,m(N_2)\}} a_{m-i} > \frac{\alpha}{2m(N_2)M}\Big) \leq \tilde{\epsilon}.$$

Hence, for all $m > m(N_2) + m(N_3)$,

$$\mathbb{P}\Big(\sum_{i=0}^{m(N_2)} a_{m-i} b_i + \frac{\alpha}{2} > \gamma \,\Big|\, \sup_{i\in\{0,\dots,m\}} a_i \leq C\Big)$$
$$\leq \mathbb{P}\Big(\sum_{i=0}^{m(N_2)} a_{m-i} b_i + \frac{\alpha}{2} > \gamma \,\Big|\, \{\sup_{i\in\{0,\dots,m\}} a_i \leq C\} \cap \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \frac{\alpha}{2m(N_2)M}\}\Big)$$
$$+ \mathbb{P}\Big(\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} > \frac{\alpha}{2m(N_2)M} \,\Big|\, \{\sup_{i\in\{0,\dots,m\}} a_i \leq C\}\Big).$$

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

And since for any events A and B such that $\mathbb{P}(B) > 0$, we have

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|\bar{B})\mathbb{P}(\bar{B})$$
$$\mathbb{P}(A) \geq \mathbb{P}(A|B)\mathbb{P}(B)$$
$$\frac{\mathbb{P}(A)}{1-\mathbb{P}(\bar{B})} \geq \mathbb{P}(A|B)$$

and by choosing the events $A := \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \alpha/(2m(N_2)M\}$ and $B = \{\sup_{i\in\{0,\dots,m\}} a_i \leq C\}$, we have for any $\epsilon, \tilde{\epsilon}$, that there exist N_2 and C such that

$$\mathbb{P}\left(\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} > \frac{\alpha}{2m(N_2)M} \,\middle|\, \{\sup_{i\in\{0,\dots,m\}} a_i \leq C\}\right) \leq \frac{\tilde{\epsilon}}{1-\epsilon/2}.$$

Hence, for every ϵ, we choose $\tilde{\epsilon}$ such that $\tilde{\epsilon}/(1-\epsilon/2) < \epsilon/2$. Observe that this choice only affects N_3. As a consequence, we have for all $m > m(N_2) + m(N_3)$,

$$\mathbb{P}\left(\sum_{i=0}^{m(N_2)} a_{m-i}b_i + \frac{\alpha}{2} > \gamma \,\middle|\, \sup_{i\in\{0,\dots,m\}} a_i \leq C\right)$$
$$\leq \mathbb{P}\left(\sum_{i=0}^{m(N_2)} a_{m-i}b_i + \frac{\alpha}{2} > \gamma \,\middle|\, \{\sup_{i\in\{0,\dots,m\}} a_i \leq C\} \cap \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \frac{\alpha}{2m(N_2)M}\}\right) + \epsilon/2,$$

where

$$\mathbb{P}\left(\sum_{i=0}^{m(N_2)} a_{m-i}b_i + \frac{\alpha}{2} > \gamma \,\middle|\, \{\sup_{i\in\{0,\dots,m\}} a_i \leq C\} \cap \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \frac{\alpha}{2m(N_2)M}\}\right)$$
$$\leq \mathbb{P}\left(\frac{\alpha}{2} + \frac{\alpha}{2} > \gamma \,\middle|\, \{\sup_{i\in\{0,\dots,m\}} a_i \leq C\} \cap \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \frac{\alpha}{2m(N_2)M}\}\right)$$
$$\leq \mathbb{P}\left(\alpha > \gamma \,\middle|\, \{\sup_{i\in\{0,\dots,m\}} a_i \leq C\} \cap \{\sup_{i\in\{0,\dots,m(N_2)\}} a_{m(N_2)-i} < \frac{\alpha}{2m(N_2)M}\}\right).$$

And since we can pick $\alpha > 0$ as small as we want, for any fixed $\gamma > 0$, we can make the probability in the last inequality equal to zero. Hence, for all $\gamma, \epsilon > 0$, there exists $N_1, N_2, N_3 \in \mathbb{N}$ such that for all $m > m(N_1) + m(N_2) + m(N_3)$, we have $\mathbb{P}(\sum_{i=0}^{m} a_{m-i}b_i > \gamma) < \epsilon$, which is what is needed to be proven. □

A.7. Statement, Comments, and Proof of Proposition A4

PROPOSITION A4. *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions i.i.d. across rows. For all $n \in \mathbb{N}$, assume that $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top \beta$. Let $\phi^\star = \beta/\|\beta\|_2$ and $s = \|\beta\|_0$. Let t, η, ϕ_t, C_{t-1}, and ϕ_0 be as in Algorithm 1. Finally, let $\tilde{\phi}_t = C_{t-1}\phi_{t-1}$. Then, for every $t \geq 2$,*

$$\phi_t = \left(\prod_{m=1}^{t} \Gamma_{t-m}\right)\phi_0 + \sum_{m=0}^{t-1}\left(\prod_{l=1}^{m} \Gamma_{t-l}\right)(\phi_{t-l} - \tilde{\phi}_{t-l})$$

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

$$+\eta\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}\Gamma_{t-l}\right)\frac{\phi_{t-1-m}^{\top}\hat{\Sigma}_X\phi_{t-1-m}}{\phi_{t-1-m}^{\top}\hat{\Sigma}_X\phi^{\star}+\phi_{t-1-m}^{\top}\frac{X^{\top}U}{\|\beta\|_2 n}}\left(\hat{\Sigma}_X\phi^{\star}+\frac{X^{\top}U}{\|\beta\|_2 n}\right)$$

almost surely, where $\Gamma_i=(I_{p\times p}-\eta\hat{\Sigma}_X)Q_i$ *for any* $i\in\mathbb{N}$.

This proposition analyzes the output of Algorithm 2. The first term explains the numerical dependence of ϕ_t on ϕ_0. From Lemma A1, one can show that there exists η, making this dependence go to zero in probability as $t\to\infty$. The second term, which is a Cauchy product, explains the impact of the normalization in Step 3.c) and of the truncation in Step 3.e) of Algorithm 2. Using Proposition A3 to prove that this term is $o_p(1)$ is a key step in the proof of Theorem 3. The third term, which is also a Cauchy product, is shown to converge to $\phi^{\star}$ by using Proposition A5.

Proof. Let $f(\phi_t):=\rho_t=(\phi_t^{\top}\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_t)/(\phi_t^{\top}\hat{\Sigma}_X\phi_t)$. As a consequence, we have $\nabla_{\phi_t}f=(2/\phi_t^{\top}\hat{\Sigma}_X\phi_t)\big(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_t-f(\phi_t)\hat{\Sigma}_X\phi_t\big)$. Moreover, note that

$$\begin{aligned}\phi_t^{\top}\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_t&=\left(\frac{\phi_t^{\top}X^{\top}Y}{n}\right)^2\\&=\left(\frac{\phi_t^{\top}X^{\top}X\beta}{n}+\frac{\phi^{\top}X^{\top}U}{n}\right)^2\\&=\left(\frac{\phi_t^{\top}X^{\top}X\phi_t\|\beta\|_2}{n}+\frac{\phi_t^{\top}X^{\top}X(\phi^{\star}-\phi_t)\|\beta\|_2}{n}+\frac{\phi_t^{\top}X^{\top}U}{n}\right)^2\\&=\left(\frac{\|X\phi_t\|_2\|\beta\|_2}{n}\left\{1+\frac{\phi_t^{\top}X^{\top}X(\phi^{\star}-\phi_t)}{\|X\phi_t\|_2^2}+\frac{\phi_t^{\top}X^{\top}U}{\|X\phi_t\|_2^2\|\beta\|_2}\right\}\right)^2.\end{aligned}$$

Hence, we have for $\tilde{\phi}_t$ as in Algorithm 2

$$\begin{aligned}\tilde{\phi}_t&=\phi_{t-1}+\eta\frac{1}{f(\phi_{t-1})}\left(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1}-f(\phi_{t-1})\hat{\Sigma}_X\phi_{t-1}\right)\\&=\phi_{t-1}+\frac{\eta}{2}\left(\phi_{t-1}^{\top}\hat{\Sigma}_X\phi_{t-1}\right)\frac{\nabla_{\phi_{t-1}}f}{f(\phi_{t-1})}\\&=\phi_{t-1}+\frac{\eta}{2}\left(\phi_{t-1}^{\top}\hat{\Sigma}_X\phi_{t-1}\right)\nabla_{\phi_{t-1}}\log(f)\\&=\phi_{t-1}+\frac{\eta}{2}\left(\phi_{t-1}^{\top}\hat{\Sigma}_X\phi_{t-1}\right)\nabla_{\phi_{t-1}}\left\{\log(\phi_{t-1}^{\top}\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1})-\log(\phi_{t-1}^{\top}\hat{\Sigma}_X\phi_{t-1})\right\}\\&=\phi_{t-1}+\frac{\eta}{2}\frac{\|X\phi_{t-1}\|_2^2}{n}\nabla_{\phi_{t-1}}\left\{\log(\phi_{t-1}^{\top}\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1})-\log\left(\frac{\|X\phi_{t-1}\|_2^2}{n}\right)\right\},\end{aligned}$$

which, given the previous remark, is equivalent to

$$\begin{aligned}\tilde{\phi}_t=\phi_{t-1}+\frac{\eta}{2}\frac{\|X\phi_{t-1}\|_2^2}{n}\nabla_{\phi_{t-1}}\Bigg\{&2\log(\|\beta\|_2)+2\log\left(\frac{\|X\phi_{t-1}\|_2^2}{n}\right)+\\&2\log\left(1+\frac{\phi_{t-1}^{\top}X^{\top}X(\phi^{\star}-\phi_{t-1})}{\|X\phi_{t-1}\|_2^2}+\frac{\phi_{t-1}^{\top}X^{\top}U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2}\right)-\log\left(\frac{\|X\phi_{t-1}\|_2^2}{n}\right)\Bigg\}.\end{aligned}$$

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

Simplifying, we get

$$\tilde{\phi}_t = \phi_{t-1} + \frac{\eta}{2}\frac{\|X\phi_{t-1}\|_2^2}{n}\nabla_{\phi_{t-1}}\left\{\log(\frac{\|X\phi_{t-1}\|_2^2}{n}) + 2\log\left(1 + \frac{\phi_{t-1}^\top X^\top X(\phi^\star - \phi_{t-1})}{\|X\phi_{t-1}\|_2^2} + \frac{\phi_{t-1}^\top X^\top U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2}\right)\right\}.$$

Computing the gradient for the first terms inside the brackets and simplifying further leads to

$$\tilde{\phi}_t = \phi_{t-1} + \eta\hat{\Sigma}_X\phi_{t-1} + \eta\frac{\|X\phi_{t-1}\|_2^2}{n}\nabla_{\phi_{t-1}}\log\left(1 + \frac{\phi_{t-1}^\top X^\top X(\phi^\star - \phi_{t-1})}{\|X\phi_{t-1}\|_2^2} + \frac{\phi_{t-1}^\top X^\top U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2}\right).$$

Leaving $g_{t-1} := 1 + \frac{\phi_{t-1}^\top X^\top X(\phi^\star - \phi_{t-1})}{\|X\phi_{t-1}\|_2^2} + \frac{\phi_{t-1}^\top X^\top U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2}$, we compute the gradient in the last term

$$\begin{aligned}
&\nabla_{\phi_{t-1}}\log\left(1 + \frac{\phi_{t-1}^\top X^\top X(\phi^\star - \phi_{t-1})}{\|X\phi_{t-1}\|_2^2} + \frac{\phi_{t-1}^\top X^\top U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2}\right)\\
&= \frac{n}{g_{t-1}}\left(\frac{1}{\|X\phi_{t-1}\|_2^2}\left(2\hat{\Sigma}_X\phi_{t-1} + \hat{\Sigma}_X\phi^\star - 2\hat{\Sigma}_X\phi_{t-1} + \frac{X^\top U}{\|\beta\|_2}\right) - g_{t-1}\frac{2\hat{\Sigma}_X\phi_{t-1}}{\|X\phi_{t-1}\|_2^2}\right)\\
&= \frac{n}{g_{t-1}}\left(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_2}\right)\frac{1}{\|X\phi_{t-1}\|_2^2} - \frac{2n\hat{\Sigma}_X\phi_{t-1}}{\|X\phi_{t-1}\|_2^2}.
\end{aligned}$$

And since for all t, $\phi_t = Q_t\phi_t$, we have

$$\begin{aligned}
\tilde{\phi}_t &= \phi_{t-1} - \eta\hat{\Sigma}_X\phi_{t-1} + \eta\frac{1}{g_{t-1}}\hat{\Sigma}_X\phi^\star + \eta\frac{1}{g_{t-1}}\frac{X^\top U}{\|\beta\|_2 n}\\
&= (I_{p\times p} - \eta\hat{\Sigma}_X)\phi_{t-1} + \eta\frac{1}{g_{t-1}}\left(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_2 n}\right)\\
&= (I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-1}\phi_{t-1} + \eta\frac{1}{g_{t-1}}\left(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_2 n}\right), \qquad \textbf{(A.3)}
\end{aligned}$$

which implies that for all $t \in \mathbb{N}$,

$$\begin{aligned}
\tilde{\phi}_t &= (I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-1}\tilde{\phi}_{t-1}\\
&+ (I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-1}(\phi_{t-1} - \tilde{\phi}_{t-1})\\
&+ \frac{\eta}{g_{t-1}}\left(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_n n}\right) \qquad \textbf{(A.4)}
\end{aligned}$$

by letting $\tilde{\phi}_0 = \phi_0$.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

Then, by recursively applying Equation (A.4), we have for all $t \geq 2$,

$$\begin{aligned}\tilde{\phi}_t =& \left(\prod_{k=1}^{t}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-k}\right)\phi_0 \\ &+ \sum_{m=1}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\right)(\phi_{t-m} - \tilde{\phi}_{t-m}) \\ &+ \eta\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\right)\frac{1}{g_{t-1-m}}\left(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_2 n}\right).\end{aligned}$$

By adding zero, on the left-hand side and rearranging the terms and noting that

$$g_{t-1} = \frac{\phi_{t-1}^\top\hat{\Sigma}_X\phi^\star + \frac{\phi_{t-1}^\top X^\top U}{\|\beta\|_n n}}{\phi_{t-1}^\top\hat{\Sigma}_X\phi_{t-1}},$$

we obtain the thesis. □

A.8. Statement, Comments, and Proof of Proposition A5

The following proposition studies the convergence of the Cauchy product in the last term of the decomposition of Proposition A4.

PROPOSITION A5. *Under the same assumptions as in Theorem 3, conditionally on X, there exists $\eta > 0$ and $Q \in \mathcal{Q}^\star$ such that*

$$\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}\Gamma_{t-l}\right)\frac{\phi_{t-1-m}^\top\hat{\Sigma}_X\phi_{t-1-m}}{\phi_{t-1-m}^\top\hat{\Sigma}_X\phi^\star + \phi_{t-1-m}^\top\frac{X^\top U}{\|\beta\|_2 n}} = [(I - Q^\star) + \eta\hat{\Sigma}_X Q^\star]^{-1} + o_p(1),$$

where $\Gamma_m = (I_{p\times p} - \eta\hat{\Sigma}_X)Q_m$ for any $m \in \mathbb{N}$.

This proposition is the key to treating the third term of Proposition A4 and thus obtaining the exact asymptotic variance given in Theorem 3.

A.8.1. *Lemmas.*

LEMMA A1. *For all $n \in \mathbb{N}$, let $X \in \mathbb{R}^{n\times p}$ be a random matrix i.i.d. across row. Let $\hat{\Sigma}_X = X^\top X/n$ and $P \in \mathcal{P}(\mathcal{F}(p,k))$ (as defined in Section 3.1.) for some $k \in \mathbb{N}$, $k < p$. If $P^\top\hat{\Sigma}_X P$ is positive definite a.s. then conditionally on X, there exists $\eta > 0$ such that $\|(I_{p\times p} - \eta\hat{\Sigma}_X)P\|_{\text{spec}} < 1$ a.s.*

Proof. The proof is given in the Supplementary Material. □

LEMMA A2. *Under the same set of assumptions as in Theorem 3, we have*

$$\left|\frac{\phi_t^\top\hat{\Sigma}_X\phi_t}{\phi_t^\top\hat{\Sigma}_X\phi^\star + \phi_t^\top\frac{X^\top U}{\|\beta\|_2 n}} - 1\right| = o_p(1).$$

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

Proof. The result follows from the facts that $\phi_t \xrightarrow{p} \phi^\star$ and $\frac{X^\top U}{\|\beta\|_2 n} \xrightarrow{p} 0$. □

LEMMA A3. *Under the same set of assumptions as in Theorem 3 (and in particular if* $\inf_{\|v\|_2=1,\,\|v\|_0=k} |v^\top \Sigma_{XY}| > 0$*), we have*

$$\sup_{\tilde{t}\in\{1,\dots,t(n)\}} \Big|\frac{\phi_{\tilde{t}}^\top \hat{\Sigma}_X \phi_{\tilde{t}}}{\phi_{\tilde{t}}^\top \hat{\Sigma}_X \phi^\star + \phi_{\tilde{t}}^\top \frac{X^\top U}{\|\beta\|_2 n}} - 1\Big| = O_p(1).$$

Proof. The proof is given in the Supplementary Material. □

LEMMA A4. *For all* $n \in \mathbb{N}$*, let* $X \in \mathbb{R}^{n\times p}$ *be a random matrix i.i.d. across rows. Let* $\hat{\Sigma}_X = X^\top X/n$*,* $k \in \mathbb{N}$ *with* $k < p$*, and* t *st.* $t \to \infty$ *as* $n \to \infty$*. For all* $t = 0, 1, 2, \dots$*, let* $P_t \in \mathcal{P}(\mathcal{F}(p,k))$ *(as defined in Section 3.1.) and* $Q_t = P_t P_t^\top$*. If for all* t*,* $P_t^\top \hat{\Sigma}_X P_t$ *is positive definite almost surely, then, conditionally on* X*, there exists* $\eta > 0$ *and* $M_\eta < \infty$ *such that conditionally on* X*,*

$$\lim_{n\to\infty} \|\sum_{m=0}^{t-1}\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\|_{\text{spec}} = M_\eta \qquad a.s.$$

Proof. Note that

$$\|\sum_{m=0}^{t-1}\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\|_{\text{spec}} \le \sum_{m=0}^{t-1}\sup_{l\in\{1,\dots,t-1\}}\|(I_{p\times p} - \eta\hat{\Sigma}_X)P_{t-l}\|_{\text{spec}}^m \|P_{t-l}\|_{\text{spec}}^m$$

and since for all $l \in \{1,\dots,t-1\}$, by construction, we have $\|P_{t-l}\|_{\text{spec}} = 1$ and by Lemma A1, there exists η such that $\|(I_{p\times p} - \eta\hat{\Sigma}_X)P_{t-l}\|_{\text{spec}} < 1$ a.s., conditionally on X. Hence, for such η, the series on the right-hand side is a convergent geometric series. This proves the thesis since $t \to \infty$ as $n \to \infty$. □

LEMMA A5. *Under the same assumptions as in Theorem 3, there exists* $\alpha > 0$ *such that we have* $\mathbb{P}\big(Q(\phi_t) \in \mathcal{Q}^\star \big| \|\phi_{\tilde{t}} - \phi^\star\|_2 \le \alpha\big) = 1$ *for all* $t \ge \tilde{t}$*, where* Q *and* $\mathcal{Q}^\star$ *are defined in Definition 1 and thereafter.*

Proof. Let x be any output of Algorithm 1 or 2, then following Definition 1, we have

$$Q_{i,j}(x) = \begin{cases} 0 \text{ if } i \neq j \text{ or if } x_i = 0 \\ 1 \text{ else,} \end{cases}$$

where $Q_{i,j}(x)$ is the element of the i^{th} row and j^{th} column of $Q(x)$.

There exists $\epsilon_1 > 0$ s.t. for any x as above if $\|x - \phi^\star\| < \epsilon_1$, then $Q(x) \in \mathcal{Q}^\star$ almost surely. Indeed, suppose that it is not the case. Hence, there exists $i \in \{1,\dots,p\}$ s.t. $x_i = 0$ and $\phi_i^\star \neq 0$ with $|\phi_i^\star| < \|x - \phi^\star\|_2 < \epsilon_1$. However, by choosing ϵ_1 to be the lowest non-null element of $|\phi^\star|_2$, we have reached a contradiction. Finally, since conditioning on $\|x - \phi^\star\|_2 < \epsilon_1$ for this choice of ϵ_1 ensures that for all $h > 0 \|\phi_{t+h} - \phi^\star\| < \|\phi_t - \phi^\star\|$, the thesis is proven. □

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

A.8.2. *Proof of Proposition A5.* For the remainder of the proof, we consider that η is such that $\|(I_{p\times p} - \eta\hat{\Sigma}_X)P_t\|_{\text{spec}} < 1$ a.s. is conditional on X. The existence of such an η follows from Lemma A1.

Let $\Gamma_m := \prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}$ (recall that our convention on $\prod$ implies $\Gamma_0 = I_{p\times p}$) and let

$$g_{t-1-m} := \frac{\phi_{t-1-m}^{\top}\hat{\Sigma}_X\phi_{t-1-m}}{\phi_{t-1-m}^{\top}\hat{\Sigma}_X\phi^{\star} + \phi_{t-1-m}^{\top}\frac{X^{\top}U}{\|\beta\|_2 n}}.$$

Define $L := g_{t-1}I_{p\times p} + g_{t-2}\Gamma_1 + \cdots + g_0\Gamma_{t-1}$ and $M := I_{p\times p} + \Gamma_1 + \cdots + \Gamma_{t-1}$ as a consequence $L - M = (g_{t-1} - 1)I_{p\times p} + (g_{t-2} - 1)\Gamma_1 + \cdots + (g_0 - 1)\Gamma_{t-1}$. Lemmas A2–A4 ensure that by Proposition A3, we have $\|L - M\|_{\text{spec}} = o_p(1)$. Hence, we have shown that $L = M + o_p(1)$. Now, consider M. For any $\alpha > 0$ and $t' \in \mathbb{N}$, let $\mathcal{E}_{\alpha t'} := \{\|\phi_{t'} - \phi^{\star}\|_2 \le \alpha\}$, then we can write $M = M \times \mathbf{1}_{\mathcal{E}_{\alpha\bar{t}}} + M \times \mathbf{1}_{\mathcal{E}^{\mathrm{c}}_{\alpha t'}}$. Given our assumptions about the convergence of ϕ_t, we know that for any $\alpha > 0$ and some t' high enough, we have $M \times \mathbf{1}_{\mathcal{E}^{\mathrm{c}}_{\alpha t'}} = o_p(1)$. Furthermore, by Lemma A5, by choosing α low enough, we see that there exists $Q^{\star} \in \mathcal{Q}^{\star}$ such that

$$M \times \mathbf{1}_{\mathcal{E}_{\alpha t'}} = \left[\sum_{m=0}^{t-t'}\left((I - \eta\hat{\Sigma}_X)Q^{\star}\right)^m + \sum_{m=t-t'+1}^{t-1}\prod_{l=1}^{m}(I - \eta\hat{\Sigma}_X)Q_{t-l}\right] \times \mathbf{1}_{\mathcal{E}_{\alpha t'}}.$$

By Lemma A5 and dealing with indices, the right-hand side can be rewritten as

$$\left[\sum_{m=0}^{t-t'}\left((I - \eta\hat{\Sigma}_X)Q^{\star}\right)^m + \sum_{m=1}^{t'-1}\left((I - \eta\hat{\Sigma}_X)Q^{\star}\right)^{(t-t')}\prod_{l=1}^{m}(I - \eta\hat{\Sigma}_X)Q_{t'-l}\right] \times \mathbf{1}_{\mathcal{E}_{\alpha\bar{t}}}.$$

Since by submutiplicativity of the spectral norm and Lemma A1 $\|(I - \eta\hat{\Sigma}_X)Q^{\star}\|_{\text{spec}} \le \|(I - \eta\hat{\Sigma}_X)P^{\star}\|_{\text{spec}} < 1$, we have (geometric series) that

$$\sum_{l=0}^{t-t'}\left((I - \eta\hat{\Sigma}_X)Q^{\star}\right)^m = [(I - Q^{\star}) + \eta\hat{\Sigma}_X Q^{\star}]^{-1}[(I - \left((I - \eta\hat{\Sigma}_X)Q^{\star}\right)^{t-t'}].$$

Moreover, $\left((I - \eta\hat{\Sigma}_X)Q^{\star}\right)^{t-t'} = o_p(1)$ and

$$\sum_{m=0}^{t'}\left((I - \eta\hat{\Sigma}_X)Q^{\star}\right)^{t-t'}\prod_{l=0}^{m}(I - \eta\hat{\Sigma}_X)Q_{t'-l} = o_p(1)$$

since $t - t' \to \infty$ as $n \to \infty$. Hence, we have shown that

$$L = M + o_p(1) = [(I - Q^{\star}) + \eta\hat{\Sigma}_X Q^{\star}]^{-1} + o_p(1),$$

which is what we needed to prove.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

A.9. Proof of Theorem 3

A.9.1. *Lemmas.*

LEMMA A6. *Under the same set of assumptions as in Theorem 3, we have* $\|\phi_t - \tilde{\phi}_t\|_2 = o_p(1)$.

Proof. The proof is given in the Supplementary Material. □

LEMMA A7. *Under the same set of assumptions as in Theorem 3 (and, in particular, if* $\inf_{\|v\|_2=1,\,\|v\|_0=k} |v^\top \Sigma_{XY}| > 0$ *and by the definition of* $\tilde{\phi}$*), we have* $\sup_{t\in\{1,\dots,t(n)\}} \|\phi_t - \tilde{\phi}_t\|_2 = O_p(1)$.

Proof. The proof is given in the Supplementary Material. □

A.9.2. *Proof of Theorem 3.* For the remainder of the proof, we consider that η is such that $\|(I_{p\times p} - \eta\hat{\Sigma}_X)P_t\|_{\text{spec}} < 1$ a.s. is conditional on X. The existence of η follows from Lemma A1.

By Proposition A4, we have for all $t \geq 2$,

$$\phi_t - \phi^\star = \left(\prod_{m=1}^{t}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-m}\right)\phi_0 \tag{A.5}$$

$$+\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\right)(\phi_{t-l} - \tilde{\phi}_{t-l}) \tag{A.6}$$

$$+\eta\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\right)g_{t-1-m}\left(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_2 n}\right) - \phi^\star, \tag{A.7}$$

where

$$g_{t-1-m} = \frac{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi_{t-1-m}}{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi^\star + \phi_{t-1-m}^\top \frac{X^\top U}{\|\beta\|_2 n}}.$$

We treat the term on the right-hand side of (A.5). By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\sqrt{n}w^\top\left(\prod_{k=1}^{t}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-k}\right)\phi_0 &\leq \sqrt{n}\|\left(\prod_{k=1}^{t}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-k}\right)\phi_0\|_2 \\
&\leq \sqrt{n}\prod_{m=1}^{t}\|(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-m}\|_{\text{spec}} \\
&\leq \sqrt{n}\prod_{m=1}^{t}\|(I_{p\times p} - \eta\hat{\Sigma}_X)P_{t-m}\|_{\text{spec}}\|P_{t-m}\|_{\text{spec}} \\
&\leq \sqrt{n}\sup_{m\in\{1,\dots,t\}}\|(I_{p\times p} - \eta\hat{\Sigma}_X)P_{t-m}\|_{\text{spec}}^t
\end{aligned}$$

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

and by Lemma A1,

$$\mathbb{P}\left(\lim_{n\to\infty}\sqrt{n}\sup_{m\in\{1,\dots,t\}}\|(I_{p\times p}-\eta\hat{\Sigma}_X)P_{t-m}\|_{\text{spec}}^t=0\middle|X\right)=1.$$

Now, we treat the terms (A.6). Lemmas A4, A6, and A7 ensure that by Proposition A3, conditionally on X, we have

$$\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p}-\eta\hat{\Sigma}_X)Q_{t-l}\right)(\phi_{t-l}-\tilde{\phi}_{t-l})=o_p(1).$$

We now treat the term (A.7). First, consider the following:

$$\eta\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p}-\eta\hat{\Sigma}_X)Q_{t-l}\right)g_{t-1-m}\hat{\Sigma}_X\phi^\star-\phi^\star.$$

Since $(I-Q^\star)\phi^\star=0$, we can write

$$\left[\left(\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p}-\eta\hat{\Sigma}_X)Q_{t-l}\right)g_{t-1-m}\right)\left((I_{p\times p}-Q^\star)+\eta\hat{\Sigma}_XQ^\star\right)-I_{p\times p}\right]\phi^\star.$$

Then, by Proposition A5, we have that this quantity converges to zero in probability conditionally on X. Finally, Proposition A5, Slutsky's Lemma, and the CLT imply that

$$w^\top\eta\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p}-\eta\hat{\Sigma}_X)Q_{t-l}\right)g_{t-1-m}\frac{X^\top U}{\|\beta\|_2\sqrt{n}}$$

is

$$\mathcal{AN}\left(0,w^\top[(I-Q^\star)+\eta\hat{\Sigma}_XQ^\star]^{-1}\eta^2\hat{\Sigma}_X[(I-Q^\star)+\eta\hat{\Sigma}_XQ^\star]^{-1}w\frac{\mathbb{V}\text{ar}(U_1)}{\|\beta\|_2^2}\middle|X\right)$$

hence we have proven the thesis.

NOTATIONAL GLOSSARY

HD High-dimensional.

GEP Generalized eigenvalue problem.

RIFLE *Truncated Rayleigh Flow Method* introduced by Tan et al. (2018), referred to as Algorithm 2 in this article.

Loaded-RIFLE Modified version of RIFLE introduced in this article and referred to as Algorithm 1.

(two-stage) IHT *(Two-stage) iterative hard thresholding* algorithm described in Jain et al. (2014).

SDAR Support detection and root finding, algorithm introduced in Huang et al. (2018)

β parameter of interest $\in\mathbb{R}^p$.

$\phi^\star$ direction of the parameter of interest, i.e., $\phi^\star:=\beta/\|\beta\|_2$.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press

$\Sigma_X := \mathbb{E}(X_1 X_1^\top) \in \mathbb{R}^{p\times p}$, $\hat{\Sigma}_X := X^\top X/n \in \mathbb{R}^{p\times p}$, $\Sigma_{XY} := \mathbb{E}(X_1 Y_1) \in \mathbb{R}^p$, $\hat{\Sigma}_{XY} := X^\top Y/n \in \mathbb{R}^p$, some moments of X_1 and Y_1 and their estimators.

$\rho(Z,k) := \sup_{\|v\|_2=1,\,\|v\|_0\le k} |v^\top Z v|$ with $Z \in \mathbb{R}^{p\times p}$ and $k \in \mathbb{N}$ s.t. $k \le p$, some norm of the matrix Z for a given k.

$\mathcal{F}(p,k) := \{F \subset \{1,\dots,p\}; \mathrm{Card}(F) = k\}$, for any $k \in \mathbb{N}$ such that $k \le p$, the set of sets containing k different integers between 1 and p.

$\mathcal{P}\big(\mathcal{F}(p,k)\big)$, that is, the set of orthogonal projectors from $\mathbb{R}^p$ to its k-dimensional subspaces that are spanned by the subsets of $\{e_1,\dots,e_p\}$ whose cardinalities equal k and that preserve the order of the indexes.

For any $m \in \mathbb{N}$ and $Z \in \mathbb{R}^{m\times m}$, $\lambda_{\max}(Z), \lambda_{\min}(Z)$, and $\lambda_i(Z)$ are, respectively, the maximum, minimum, and i^{th} (in descending order) *eigenvalues* of Z. If Z is positive definite, $\kappa(Z) = \lambda_{\max}(Z)/\lambda_{\min}(Z)$ is the *condition number* of Z.

For any $k \in \mathbb{N}$ s.t. $k \le p$ and any $F \in \mathcal{F}(k,p)$, λ_i, $\lambda_i(F)$, $\hat{\lambda}_i$, and $\hat{\lambda}_i(F)$ are the i^{th} (in descending order) *generalized* eigenvalues of, respectively, the pairs of matrices $\big(\Sigma_{XY}\Sigma_{YX}, \Sigma_X\big)$, $\big(P_{(F)}^\top \Sigma_{XY}\Sigma_{YX}P_{(F)},\ P_{(F)}^\top \Sigma_X P_{(F)}\big)$, $\big(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}, \hat{\Sigma}_X\big)$, and finally $\big(P_{(F)}^\top \hat{\Sigma}_{XY}\hat{\Sigma}_{YX}P_{(F)}, P_{(F)}^\top \hat{\Sigma}_X P_{(F)}\big)$.

$Cr(A,B) = \min_{\|v\|_2=1}\left(\big(v^\top A v\big)^2 + \big(v^\top B v\big)^2\right)^{1/2}$ is the *Crawford number* of (A,B) for any $p \in \mathbb{N}$ and any symmetric matrices $A, B \in \mathbb{R}^{p\times p}$.

$\Delta(k) := \inf_{P\in\{\cup_{i=1}^k \mathcal{P}(\mathcal{F}(p,i))\}} Cr(P^\top \Sigma_{XY}\Sigma_{YX}P, P^\top \Sigma_X P)$ for any $k \in \mathbb{N}$ such that $k \le p$.

$\varepsilon(k) := \left(\rho(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \Sigma_{XY}\Sigma_{YX}, k)^2 + \rho(\hat{\Sigma}_X - \Sigma_X, k)^2\right)^{1/2}$ for any $k \in \mathbb{N}$ such that $k \le p$.

The input of Algorithm 1 is: initial vector ϕ_0; initial and final sparsity levels $\bar{k}, \underline{k} \in \{1,\dots,p\}$ with $\bar{k} > \underline{k}$; step size η; and number of iterations T.

$\hat{\phi}(t)$ is the output of Algorithm 1.

$\mathcal{Q}^\star := \{Q = PP^\top | P \in \cup_{i=\|\phi\|_0}^k \mathcal{P}\big(\mathcal{F}(p,i)\big), Q\phi^\star = \phi^\star\}$, a set of matrices identifying the support of $\phi^\star$.

https://doi.org/10.1017/S0266466626100334 Published online by Cambridge University Press