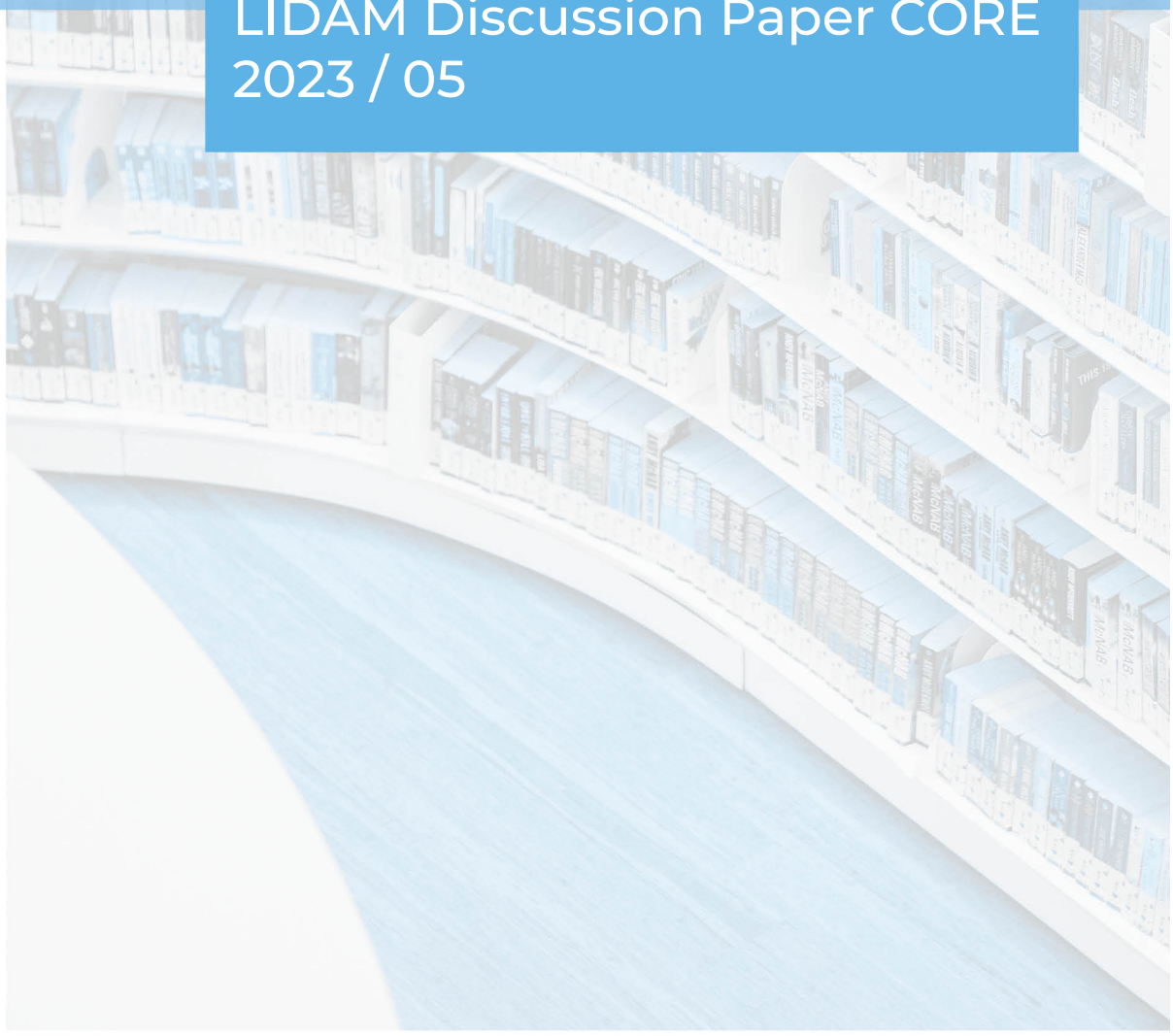
DIRECTION IDENTIFICATION AND MINIMAX ESTIMATION BY GENERALIZED EIGENVALUE PROBLEM IN HIGH DIMENSIONAL SPARSE REGRESSION

Mathieu Sauvenier, Sébastien Van Bellegem

LIDAM Discussion Paper CORE
2023 / 05



UCLouvain

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

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: immaq-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/core/core-discussion-papers.html

Direction Identification and Minimax Estimation by Generalized Eigenvalue Problem in High Dimensional Sparse Regression

Mathieu Sauvenier*[1] and Sébastien Van Bellegem[2]

[1,2]Center for Operations Research and Econometrics (CORE), Voie du Roman Pays, 34 , 1348 Louvain-la-Neuve, Belgium
[1]mathieu.sauvenier@uclouvain.be
[2]sebastien.vanbellegem@uclouvain.be

February 9, 2023

Abstract

In high-dimensional sparse linear regression, the selection and the estimation of the parameters are studied based on an l_0-constraint on the direction of the vector of parameters. We first establish a general result for the direction of the vector of parameters, which is identified through the leading generalized eigenspace of measurable matrices. Based on this result, we suggest addressing the best subset selection problem from a new perspective by solving an empirical generalized eigenvalue problem to estimate the direction of the high-dimensional vector of parameters. We then study a new estimator based on the RIFLE algorithm and demonstrate a nonasymptotic bound of the L^2 risk, the minimax convergence of the estimator and a central limit theorem. Simulations show the superiority of the proposed inference over some known l_0 constrained estimators.

Keywords: High-dimensional model, sparsity, generalized eigenvalue problem, identification, best subset selection, minimax l_0 estimation, central limit theorem.
JEL Codes: C30, C55, C59

*corresponding author

1 Introduction

This paper connects two well-known statistical problems. First, the *best subset selection* problem that is a solution of

$$\tilde{\beta} \in \underset{\hat{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\|Y - X\hat{\beta}\|_2^2 \text{ s.t. } \|\hat{\beta}\|_0 \leq k, \tag{1.1}$$

for $k \in \mathbb{N}$ [Miller, 2002], where $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ and where for any $v \in \mathbb{R}^p$, $\|v\|_2 = \sum_{i=1}^p |v_i|^2$ is the l_2-norm of v and $\|v\|_0 = \sum_{i=1}^m \mathbf{1}_{|v_i|>0}$ is the $l_0-pseudonorm$ of v. Second, the *generalized eigenvalue problem* (GEP), that is, given two square matrices $A, B \in \mathbb{R}^{p\times p}$, solving

$$Av_i = \lambda_i Bv_i \text{ for } i = 1, ..., p \tag{1.2}$$

for λ_i and v_i, which are called *generalized eigenvalues* and *generalized eigenvectors*, respectively. We observe that if B is the identity matrix, then the GEP is the standard eigenvalue problem (for an introduction to GEP, see *e.g.,* Ghojogh, Karray, and Crowley [2019]).

The connection between the two problems is based on the idea that in the linear regression $Y = X\beta+U$, where $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ are random and *iid* across rows and $U = (Y_1 - \mathbb{E}(Y_1|X_1), ..., Y_n - \mathbb{E}(Y_n|X_n))^\top$, the parameter β is completely characterized by two objects. First, its *direction*, $\Phi \subset \mathbb{R}^p$, defined as the vector subspace of $\mathbb{R}^p$ of all the vectors collinear to β, and second, its *magnitude* that its norm in $\mathbb{R}^p$ under a suitable choice of basis for Φ. In Theorem 1 below, we identify the direction Φ with the *generalized eigenspace* associated with the only nonnull *generalized eigenvalue* of a pair of measurable matrices.

One important consequence of our identification result is that by exposing the connection between the linear regression problem and the GEP, we extend the class of estimators available for the former to the one available for the latter. This is especially useful in the context of *high dimensions* (HD).

When the dimension of the target parameter β, namely, p, is greater than the number of observations, n, for instance, because the former is an increasing function of the latter, we say that the model is *high dimensional*. Estimating the parameter β in HD is a problem of great interest from both theoretical and practical points of view. Since the high dimensionality of the model implies that the empirical second moment of the covariates is not invertible, the parameter β cannot be consistently estimated via ordinary least squares. However, situations where $p > n$ arise in practice, for instance, in biology, health studies, economics or machine learning [Fan and Li, 2006], and call for efficient estimation procedures.

There is no unique way to tackle the problem of consistently estimating the parameters of a high-dimensional linear model. One line of research

is based on the notion of *sparsity*. According to Stodden [2006], Box and Meyer [1986] coined the term *factor sparsity* to describe a model where the majority of covariates have zero effect. Since then, a wide and active area of research has specialized in the estimation of the parameters of high-dimensional *sparse* linear models *i.e.* high-dimensional linear models where it is assumed that only a subset of the elements of β differs from zeros (*e.g.* Belloni, Chernozhukov, and Hansen [2013], van de Geer [2016]). Sparsity assumes that

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

where s is the *sparsity level*.

The best subset selection problem (1.1) originates from the combination of the sparsity assumption with the least squares principle. Due to the l_0 constraint, this program is NP-hard [Natarajan, 1995]. Hence, solving (1.1) in polynomial time is impossible unless $P = NP$. Despite this computational difficulty, a set of theoretical results indicates that the best subset selection is a desirable estimator for sparse regression (*e.g.* Greenshtein [2006], Raskutti, Wainwright, and Yu [2011], Zhang and Zhang [2012], Shen, Pan, Zhu, and Zhou [2013], Zhang et al. [2014], Bertsimas, King, and Mazumder [2016]). Recently, optimal or near-optimal solutions have been proposed in Bertsimas et al. [2016] via a reformulation of problem (1.1) as a mixed integer optimization (MIO) problem. The implementation of this method is performed via the licensed GUROBI solver and is demanding in computing time as the dimension of the model increases. For a critical review, see *e.g.* Hastie, Tibshirani, and Tibshirani [2020].

By following the same line of thought that leads to the formulation of the best subset selection, we formulate an estimator of Φ over a sparsity constraint. To do so, we observe that the generalized eigenvector associated with the maximum generalized eigenvalue of the GEP (1.2) is the solution of the maximization of the following Rayleigh quotient

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

see *e.g.* Golub and Van Loan [2013]. Our identification result in Theorem 1 allows combining this fact with the sparsity assumption to obtain the sparse estimator of Φ, that is,

$$\hat{\Phi} := \underset{\phi \in \mathbb{R}^p}{\operatorname{argmax}} \frac{\phi^\top X^\top Y Y^\top X \phi}{n \phi^\top X^\top X \phi} \text{ s.t. } \|\phi\|_0 \leq k \tag{1.3}$$

for $k \in \mathbb{N}$. We call (1.3) the *directional best subset selection* problem.

Below, we demonstrate that for any solution of problem (1.3), say v, the product of v with the orthogonal projection of Y on Xv is a solution of the

best subset selection (1.1). Hence, solving problem (1.3) allows solving (1.1). We also demonstrate that the two problems have the same minimax rate of convergence in the l_2 norm, that is, $(s\log(p/s)/n)^{1/2}$ in the sub-Gaussian setting [Raskutti et al., 2011].

Next, we show how these ideas may be fruitful for the estimation of high-dimensional sparse linear models (HDSLM) by proposing an efficient estimator of Φ, which gives as a byproduct a novel estimator of β. This estimator is a modification of a well-established algorithm called RIFLE that was proposed by Tan, Wang, Liu, and Zhang [2018] to approximate the solution of GEP over a l_0 constraint (which are also NP-Hard [Moghaddam, Weiss, and Avidan, 2006]) when $p > n$. We demonstrate that our estimator converges at the minimax rate of convergence with high probability in the sub-Gaussian setting and that it possesses the sure screening property.

Finally, we provide a central limit theorem for the estimator, which is a nontrivial task given its nonlinearity. To this aim, we prove a theorem that gives sufficient conditions for the convergence in probability of a *Cauchy product*. This theorem may be of independent interest.

Generalized eigenvalue problems have already been employed to tackle models with many covariates. In particular, the *sufficient dimension reduction* approach introduced in Cook [1994] employs moment-based methods that can be formulated as GEP to replace the original covariate with a minimal set of their linear combinations without loss of information [Li, 2007]. Those methods are not designed specifically to tackle high-dimensional sparse linear models and are not an answer to the best subset selection problem. We still mention the research of Chen, Zou, and Cook [2010] that introduced a method to induce sparsity in the estimated linear combination of the covariate and thus to *de facto* select variables.

Like RIFLE, our estimator requires an initial solution that is then iteratively improved toward convergence. The article concludes with a set of Monte Carlo experiments that show the superiority of the proposed inference when our estimator is initialized by using lasso, that is, the Lagrangian version of the l_1-relaxation of problem (1.1) (Tibshirani [1996], Chen, Donoho, and Saunders [1998]). Lasso is known to select too many random variables when tuned for prediction accuracy (Bertsimas et al. [2016], Hastie et al. [2020]). Our experiment shows that for signal-to-noise ratios higher than two, our estimator improves on the initial solution provided by lasso, *while* reducing the number of variables associated with nonnull coefficients.

The performances of the proposed inference are compared with those of three other l_0 constrained/penalized estimators. First and second are the Iterative Hard Thresholding (IHT) and the two-stage IHT (Jain, Tewari, and Kar [2014]) that are designed to approximate problem (1.1). The third competitor is the support detection and root finding (SDAR), proposed by Huang, Jiao, Liu, and Lu [2018] to approximate the Lagrangian version of the best subset selection (which is not equivalent to problem (1.1) because

of the l_0-constraint) but still allows fixing the sparsity level of the estimate.

We summarize our comparison of the four algorithms in Table 1 where *Loaded-RIFLE* is the name of our estimator. Minimax convergence rates and the sure screening property have been established for the IHT and two-stage IHT (Jain et al. [2014] and Guo, Zhu, and Fan [2020]). Minimax convergence rates have been established for SDAR (Huang et al. [2018]). To the best of our knowledge, no CLT has been established for these algorithms. The numerical performances are compared via Monte Carlo experiments in Section 5. Accuracy in estimation, variable selection and prediction are compared. The computing time given in Table 1 is the rounded mean computing time of 100 replications in the second setup described in Section 5. The computational time of the Loaded-RIFLE *includes* the time for computing the initial vector via lasso. Our implementation of the IHT optimizes the step size at each iteration (as in Blumensath and Davies [2009a]), which greatly improves estimation but is costly in computational time.

Estimator	Efficiency	CLT	Num. perf.	Time (s)
IHT	near minimax			$\approx 5e^{-1}$
2 steps-IHT	near minimax			$\approx \mathbf{3e^{-2}}$
SDAR	**minimax**			$\approx 1e^{-1}$
Loaded-RIFLE	**minimax**	**yes**	**best overall**	$\approx 2e^{-1}$

Table 1

The organization of the paper is as follows. Section 2 provides the structural model and the main objects with which we are working. Section 3 addresses the identification of Φ with the solution of a particular GEP. In Section 4, we use the result of the preceding section to treat the estimation of Φ under sparsity. First, we introduce the directional best subset selection (1.3) and establish a theory connecting it with the best subset selection (1.1). Second, we show how these results may be applied by presenting RIFLE and Loaded-RIFLE, which are two algorithms that produce solutions for the directional best subset selection problem. Third, we provide nonasymptotic bounds as well as the convergence rate in the sub-Gaussian setting for the relative error of these two estimators. Fourth, we present a central limit theorem for these estimators and discuss the theorem. This discussion allows us to present intermediate results that provide deep insights into the functioning of RIFLE. In particular, one of these results gives sufficient conditions for the convergence in the probability of the stochastic Cauchy product and may be of independent interest. In Section 5, we present numerical experiments to compare the performance of the proposed algorithms that demonstrate the competitiveness of Loaded-RIFLE. Section 6 contains concluding remarks.

2 Structural Model

We consider a sequence of statistical experiments indexed by $n \in \mathbb{N}$. *For each* n we assume that we observe a random vector of *explanatory variables*, $Y^{(n)} \in \mathbb{R}^n$ and a random matrix of *covariates* $X^{(n)} \in \mathbb{R}^{n \times p}$. We also assume that *for each* $n \in \mathbb{N}$ the rows of $Y^{(n)}$ and $X^{(n)}$ are independent and identically distributed. To model high dimensionality, we assume $p = p(n)$, that is, p, the *number of explanatory variables*, is a function of n, *the number of observations* and then consider situations where $p(n)$ may increase faster than n (for instance, below, we consider situations where the dimension p is of order $O(n^\alpha)$ with $\alpha > 1$).

We assume that for each $n \in \mathbb{N}$ there exists $\beta^{(n)} \in \mathbb{R}^p$ s.t. $\mathbb{E}^{(n)}(Y_1^{(n)}|X_1^{(n)}) = (X_1^{(n)})^\top \beta^{(n)}$ where $Y_1^{(n)}$ and $X_1^{(n)}$ represent the transpose of the first rows of $Y^{(n)}$ and $X^{(n)}$.

This paper is only indirectly interested in $\beta^{(n)}$ and focuses instead on the identification and estimation of the *direction of* $\beta^{(n)}$, which is the one-dimensional subspace of $\mathbb{R}^p$ containing all the vectors collinear to β.

Definition 1 *(Direction space) For any vector* $v \in \mathcal{V}$*, we call the* direction *of* v *the one-dimensional subspace of* $\mathcal{V}$ *defined by*

$$\mathcal{D}(v) := \{mv | m \in \mathbb{R}\}.$$

For the remainder of this paper, we let $\Phi^{(n)} := \mathcal{D}(\beta^{(n)})$ and choose $\phi^{(n)\star} := \beta^{(n)}/\|\beta^{(n)}\|_2$ as the basis for $\Phi^{(n)}$, where for any and $u, v \in \mathbb{R}^p$, the l_2-norm of v, $\|v\|_2$, is induced by the *canonical inner product* on $\mathbb{R}^p$, namely, $\langle u, v\rangle_2 := \sum_{i=1}^p u_i v_i$. We observe that this choice of basis for $\Phi^{(n)}$ implies that the magnitude of $\beta^{(n)}$ is its l_2-norm.

Since p may increase faster than n, we may encounter situations where the dimension of the parameter space is larger than the sample size. Hence, we require our high-dimensional linear model to be sparse; that is, for each $n \in \mathbb{N}$, we assume that $\|\phi^{\star(n)}\|_0 = s(n) < p(n)$.

In the remainder of the paper, we drop the upper script (n) to simplify the notation.

3 Identification of Φ

This section provides a novel theorem that identifies Φ to some functions of moments of Y_1 and X_1. The result suggests a natural estimator that is exploited in Section 4. Before stating the theorem, we introduce some notation. Let $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ be the second moment of X_1; moreover, let $\Sigma_{XY} := \mathbb{E}(X_1 Y_1)$ and $\Sigma_{YX} := (\Sigma_{XY})^\top$. For all $u, v \in \mathbb{R}^p$, let $\langle u, v\rangle_{\Sigma_X} := \langle u, \Sigma_X v\rangle_2$ be the inner product induced by Σ_X on $\mathbb{R}^p$. For any $m \in \mathbb{N}$ and $Z \in \mathbb{R}^{m \times m}$, let $\mathrm{Ker}(Z) = \{v \in \mathbb{R}^m; \|Zv\|_2 = 0\}$ be the *kernel* of Z.

Theorem 1 *Let $Y_1 \in \mathbb{R}$ and $X_1 \in \mathbb{R}^p$ be random. If there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top \beta$ and if Σ_X is positive definite then*

$$\Phi = \underset{\phi \in \mathbb{R}^p}{\operatorname{argmax}} \frac{\phi^\top \Sigma_{XY} \Sigma_{YX} \phi}{\phi^\top \Sigma_X \phi} \tag{3.1}$$

$$= \underset{\phi \in \mathbb{R}^p}{\operatorname{argmax}} \left(\left(\frac{\phi}{\|\phi\|_{\Sigma_X}} \right)^\top \Sigma_{XY} \right)^2 \tag{3.2}$$

$$= \operatorname{Ker}(\Sigma_{XY}\Sigma_{YX} - \Sigma_X \lambda_{\max}) \tag{3.3}$$

where $\lambda_{\max} = \max_{\phi \in \mathbb{R}^p} \frac{\phi^\top \Sigma_{XY} \Sigma_{YX} \phi}{\phi^\top \Sigma_X \phi}$ and $\Phi = \mathcal{D}(\beta)$.

The proof is given in Appendix. Theorem 1 identifies Φ as the solution set of a generalized Rayleigh quotient (equations (3.1) and (3.2)) or equivalently as the generalized eigenspace associated with the maximum generalized eigenvalue of the matrix pair $(\Sigma_{XY}\Sigma_{YX}, \Sigma_X)$ (equation (3.3)).

This result is a general property of the geometry of linear models that, to the best of our knowledge, has not yet been pointed out as such in the literature. We still mention that a somewhat similar equation is given in the context of partial least squares in Sun, Ji, and Ye [2009].

The literature sometimes suggests solving specific GEP via various least squares programs (*e.g.* Moghaddam et al. [2006], Sun et al. [2009], Golub and Van Loan [2013]). One original aspect of the present paper lies in the fact that we propose to estimate the parameters of linear models, a task that is commonly achieved via least squares, through a generalized eigenvalue problem.

In addition to the linearity of the conditional expectation, the only assumption required by Theorem 1 is the positive definitiveness of Σ_X. This assumption is well known to be an identification condition for β (*e.g.,* Florens, Marimoutou, Peguin-Feissolle, Perktold, and Carrasco [2007]).

4 Estimation under sparsity

4.1 Directional Best Subset Selection

We introduce the *directional best subset selection*, prove that its solution has the same direction as the solution of the *best subset selection*, prove that the latter can be recovered by multiplying the former with an estimation of the magnitude, ν, and connect the rate of convergence of the two estimators. Let $\hat{\Sigma}_X := X^\top X/n$, $\hat{\Sigma}_{XY} := X^\top Y/n$ and $\hat{\Sigma}_{YX} := (\hat{\Sigma}_{XY})^\top$ be estimators of Σ_X, Σ_{XY} and Σ_{YX}, respectively.

We consider the empirical version of problem (3.1) to estimate Φ over a sparsity constraint, that is, problem (1.3). The fact that its solution set, $\hat{\Phi}$, finds the direction of the solution of the best subset selection is established in the following proposition.

Proposition 1 *Let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions. For a given $k \in \mathbb{N}$, let $\tilde{\beta}$ and $\hat{\Phi}$ be as in problems (1.1) and (1.3) respectively. If $\tilde{\beta}$ is unique almost surely, then conditionally on the set of events where $\min_{v\in\mathbb{R}^p,\|v\|_0\leq k} v^\top \hat{\Sigma}_X v > 0$ we have $\mathcal{D}(\tilde{\beta}) = \hat{\Phi}$ almost surely.*

The proof is given in the Appendix and leads to the following Corollary.

Corollary 1 *Suppose the assumptions of Proposition 1 hold true. Let $\tilde{\phi} \in \hat{\Phi}$ and $\tilde{v} = \operatorname{argmin}_{m\in\mathbb{R}} n^{-1}\|Y - X\tilde{\phi}m\|_2^2$ then conditionally on the same set of events than in Proposition 1 we have $\tilde{\beta} = \tilde{\phi}\tilde{v}$ almost surely.*

Proposition 1 and Corollary 1 exemplify how Theorem 1 can be employed to extend the class of estimators of the high-dimensional sparse linear model (HDSLM) defined in the preceding section to those of high-dimensional sparse GEP. Here, we focus on l_0-constrained estimators, but the proofs and results can be straightforwardly generalized.

The strategy of estimation relies on two steps: First, we estimate Φ via GEP, and second, if the parameter of interest is β, we choose any basis in $\hat{\Phi}$, say $\tilde{\phi}$, and then regress Y on the span of $X\tilde{\phi}$ to estimate the magnitude of β. Corollary 1 ensures that if in the first step the directional best subset selection is employed, this procedure gives an estimator of β (the direction times the magnitude) that is actually the solution of the best subset selection.

The next proposition expresses the error of estimation of β as a function of the error of the estimation of $\phi^\star$ to connect the minimax rate of convergence of the former with that of the latter.

Proposition 2 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$, assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite, that there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top\beta$, $\|\beta\|_0 = s$. Let $\phi^\star := \beta/\|\beta\|_2$ and assume $\hat{\phi} \in \mathbb{R}^p$ is an estimator of $\phi^\star$ for which $\langle\hat{\phi},\phi^\star\rangle_2 \geq 0$, $\|\hat{\phi}\|_0 = s$ and $\|\hat{\phi}\|_2 = 1$ hold almost surely and whose convergence is such that $\|\hat{\phi} - \phi^\star\|_2 = O_p(\alpha_n)$ where $\alpha_n = o(1)$ satisfies $(2s/n)^{1/2}/\alpha_n = o(1)$. Finally, let $\hat{v} = \operatorname{argmin}_{m\in\mathbb{R}} \|Y - X\hat{\phi}m\|_2^2$ be an estimator of the magnitude of β.*

Conditionally on the set of events where there are $0 < \underline{\kappa} \leq \overline{\kappa} < \infty$ such that for all $u \in \mathbb{R}^p$ with $\|u\|_2 = 1$, $\|u\|_0 = \max\{2s, p\}$ we have

$$\underline{\kappa} \leq u^\top \hat{\Sigma}_X u \leq \overline{\kappa}, \tag{4.1}$$

if $\hat{\beta} \in \mathbb{R}^p$ is an estimator of β that satisfies $\hat{\beta} = \hat{\phi}\hat{v}$ almost surely, we have

$$\|\hat{\beta} - \beta\|_2^2 = c\|\hat{\phi} - \phi\|_2^2 + O_p(\alpha_n^2).$$

where $c = O_p(1)$.

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

This result is important since it implies that the minimax rate of convergence in the l_2 loss of the directional best subset selection (1.3) is the same as that of the best subset selection (1.1). Raskutti et al. [2011] have proven that in a Gaussian and fixed design (where the matrix X is assumed to be nonrandom), if a version of assumption (4.1) holds, the minimax rate of convergence of $\|\tilde{\beta} - \beta\|_2$ is $(s \log(p/s)/n)^{1/2}$. In the next section, we propose algorithms that *aim* to solve problem (1.3) and to achieve this statistical rate of convergence with high probability.

We note that we assumed $\|\hat{\phi}\|_0 = s$ to parallel the setting of Raskutti et al. [2011]. By a slight adjustment in the statement of assumption (4.1) and in the proof, we can replace this assumption by $\|\hat{\phi}\|_0 = k$ for any $k \in \{1, ..., p\}$.

4.2 Algorithms for Directional Best Subset Selection

We aim to solve problem (1.3). However, this problem is NP-hard and hence cannot be exactly solved in polynomial time unless $P = NP$. Nevertheless, Tan et al. [2018] propose an algorithm that approximates the solution of GEP over a sparsity constraint when $p > n$. In essence, their proposition consists of iteratively updating the estimated vector in its ascent direction, normalizing it, truncating it such that only the k coefficients with the highest absolute magnitude are nonzero and finally normalizing it again. The resulting algorithm applied to the problem (1.3) is detailed here.

Notation 1 *When we write Truncate$(v, \mathcal{K})$ for $v \in \mathbb{R}^p$ and $\mathcal{K} \subset \{1, ..., p\}$ we refer to the vector of $\mathbb{R}^p$ defined by*

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

for all $i = 1, ..., p$.

Algorithm 1 Truncated Rayleight flow method (RIFLE)

Input*: matrices $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ and $\hat{\Sigma}_X$, initial vector ϕ_0, sparsity level $k \in \{1, ..., p\}$ 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) $C_{t-1} = I_{p\times p} + (\eta/\rho_{t-1})(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \rho_{t-1}\hat{\Sigma}_X)$;

(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] provides 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, $O(kp)$ for taking the product between a truncated vector and a matrix with columns restricted to the set F_t and for computing the difference between two matrices with columns restricted to set F_t [Tan et al., 2018].

The parameter k is understood as a regularization (or tuning) parameter that is chosen.

Our experience with Algorithm 1 leads us to believe that when η is small (as suggested by the theory in Tan et al. [2018]), the variable selection operated by the algorithm is highly dependent on the ordering of the absolute value of the element of ϕ_0. In many situations, this constrains the quality of the estimation. To circumvent this feature, we introduce a variation of Algorithm 1 that we call Loaded-RIFLE and that is described here.

Algorithm 2 Loaded truncated Rayleight flow method (Loaded-RIFLE)

Input*: matrices* $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ *and* $\hat{\Sigma}_X$*, initial vector* ϕ_0*, initial and final sparsity level* $\overline{k}, \underline{k} \in \{1, ..., p\}$ *with* $\overline{k} > \underline{k}$*, step size* η *and number of iteration* T*.*

START

1. $\mathcal{F}_{0,\overline{k}} =$ *the set of index of the* $\underline{k}$ *highest elements of* $|\phi_0|$;
2. $\phi_{0,\overline{k}} = Truncate(\phi_0, \mathcal{F}_{0,\overline{k}})$;
3. *Let* $k = \overline{k}$;
4. *WHILE* $k > \underline{k}$ *repeat the following steps:*
   (a) *Let* $t = 1$
   (b) *WHILE* $t \leq T$ *repeat the followings steps*
      i. $\rho_{t-1,k} = (\phi_{t-1,k}^\top \hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1,k})/(\phi_{t-1,k}^\top \hat{\Sigma}_X \phi_{t-1,k})$;

ii. $C_{t-1,k} = I_{p\times p} + (\eta/\rho_{t-1,k})(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \rho_{t-1,k}\hat{\Sigma}_X)$;

iii. $\phi'_{t,k} = C_{t-1,k}\phi_{t-1,k}/\|C_{t-1,k}\phi_{t-1,k}, k\|_2$;

iv. $\mathcal{F}_{t,k}$ = *the set of index of the* k *highest elements of* $|\phi'_t|$;

v. IF $Card\big(\mathcal{F}_{t,k} \cup \mathcal{F}_{t-1,k}\big) > \overline{k}$

A. $\phi_{T,k} = \phi_{0,k}$

B. set $t = T + 1$ *(BREAK the while loop on* t*);*

vi. ELSE

A. $\hat{\phi}_{t,k} = Truncate(\phi'_{t,k}, \mathcal{F}_{t,k})$;

B. $\phi_{t,k} = \hat{\phi}_{t,k}/\|\hat{\phi}_{t,k}\|_2$;

C. $t = t + 1$;

(c) $\phi_{0,k-1} = \phi_{T,k}$;

(d) $k = k - 1$;

5. *Apply Algorithm 1 with* $\phi_0 = \phi_{t,\underline{k}+1}$, $k = \underline{k}$, η;

END

Output ϕ_t.

Loaded RIFLE 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 to extend the theory of RIFLE developed by Tan et al. [2018] to Loaded-RIFLE.

4.3 Consistency of the estimation and the rate of convergence

By using the techniques employed in Tan et al. [2018] to study the convergence of Algorithm 1, we derive a nonasymptotic bound on the l_2 distance between the output of Algorithm 2 and the population target vector (which is $\phi^\star$ in this paper). Based on this result, we derive the convergence rate of the estimator in the sub-Gaussian setting. Before stating the theorems, we introduce more notation where some are specific to GEP and others are nonstandard but useful in the context of this paper.

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

$$\rho(Z, k) := \sup_{\|v\|_{\mathbb{R}^p}, \|v\|_0 \leq k} |v^\top Z v|.$$

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 $\mathcal{F}(p,k) := \{F \subset \{1,...,p\}; \text{Card}(F) = k\}$ 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 zero's on all its elements except on the i^{th} position which contains a one) and let

$$\mathcal{P}(\mathcal{F}(p,k)) := \{P_{(F)} := (e_{\alpha_1},...,e_{\alpha_k}) \in \mathbb{R}^{p\times k} | \\ F \in \mathcal{F}(p,k); \alpha_i \in F; \alpha_1 < \alpha_2 < ... < \alpha_k\}$$

be a set of orthogonal projectors from $\mathbb{R}^p$ to its k dimensional subspaces that are spanned by the subsets of $\{e_1,...,e_p\}$ whose cardinalities equal k and that preserve the order of the indexes. This set is useful since if $v \in \mathbb{R}^p$ such that $\|v\|_0 = k$ then $\mathcal{P}(\mathcal{F}(p,k))$ 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 $\mathbb{R}^p$. Observe we have that $P^\top P = I_{k\times k}$, $\|PP^\top\|_{\text{spec}} = 1$ 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 $(\Sigma_{XY}\Sigma_{YX}, \Sigma_X)$, $(P_{(F)}^\top \Sigma_{XY}\Sigma_{YX}P_{(F)}, P_{(F)}^\top \Sigma_X P_{(F)})$, $(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}, \hat{\Sigma}_X)$ and finally $\left(P_{(F)}^\top \hat{\Sigma}_{XY}\hat{\Sigma}_{YX}P_{(F)}, P_{(F)}^\top \hat{\Sigma}_X P_{(F)}\right)$.

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

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

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

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

In the next results, we have employed the techniques developed in Tan et al. [2018] to provide a nonasymptotic bound on the l_2 error of Algorithm 2.

Theorem 2 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top \beta$ and $\|\beta\|_0 = s$. Let $\phi^\star = \beta/\|\beta\|_2$ and $\Lambda := \lambda_1/(1+\lambda_1^2)$.*

Let $\underline{k}$ in Algorithm 2 be such that $\underline{k} = C's$ for $C' > 1$ and let $\overline{k} = C''\underline{k}$ for $C'' > 1$ and $k' = \overline{k} + s$.

Let $\phi_{t,\underline{k}}$ be the output of Algorithm 2 (i.e. after t iterations of the last step of the algorithm). Suppose that $t \geq T$ where T is like in Algorithm 2. Let ϕ_0 be the initial vector in Algorithm 2 after truncation (as in step 2.).

Choose a constant $c > 0$ and η in Algorithm 2 such that

$$\eta \lambda_{\max}(\Sigma_X) < 1/(1+c), \tag{4.2}$$

and

$$\nu = (1 + \sqrt{\frac{s}{\underline{k}}})\left(1 - \frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond \kappa(\Sigma_X)}\right)^{1/2} < 1 \tag{4.3}$$

hold true, where $C_\diamond = (1+c)/(1-c)$.

Then, conditionally on the set of events where

$$\frac{\varepsilon(k')}{\Delta(k')} < \frac{\Lambda(1-\nu)\sqrt{\xi}}{4\sqrt{10} + \Lambda(1-\nu)\sqrt{\xi}}, \tag{4.4}$$

$$\rho(\hat{\Sigma}_X - \Sigma_X, k') \leq c\lambda_{\min}(\Sigma_X) \tag{4.5}$$

$$|\langle \phi^\star, \phi_0 \rangle_2| \leq 1 - \nu^2\xi, \tag{4.6}$$

hold almost surely with ξ defined as

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

we have

$$\sqrt{1 - |(\phi^\star)^\top \phi_{t,\underline{k}}|} \leq \nu^t\sqrt{\xi} + \frac{1}{(1-\nu)^2}\sqrt{10}\frac{\sqrt{2}}{\Lambda\big(\Delta(k') - \varepsilon(k')\big)}\varepsilon(k')$$

almost surely.

Since the theorem is an extension to Algorithm 2 of a known result given in Tan et al. [2018] for Algorithm 1, we give a complete proof in Supplementary Material. We note that this proof provides as a byproduct a version of the theorem for Algorithm 1 that fixes the one given in Tan et al. [2018].

Theorem 2 indeed allows one to bound the l_2 norm between $\phi^\star$ and ϕ_t since if $\langle \phi^\star, \phi_t \rangle_2 \geq 0$, then $\sqrt{1 - |(\phi^\star)^\top \phi_t|} = 1/\sqrt{2}\|\phi^\star - \phi_t\|_2$.

The first term on the right-hand side, namely, $\nu^t\sqrt{\xi}$, is the optimization error. The second term on the right-hand side is the statistical error [Tan et al., 2018].

Assumptions (4.4) and (4.5) control for the error of estimation of the (generalized) spectrum of the (pair of) matrices at hand. In particular, they provide upper and lower bounds on the condition number of Σ_X, on the spectrum of submatrices of $\hat{\Sigma}_X$ and on $\hat{\lambda}_i(F)$ for any $F \in \mathcal{F}(p, k')$ and $i = 1, ..., k'$; Tan et al. [2018] contains the details. As we show later, if Y_1 and X_1 are sub-Gaussian, there exist ratios between n, $p(n)$ and $s(n)$ such that the probability of the events where these assumptions do not hold converges to zero as $n \to \infty$.

Theorem 2 also requires theoretical conditions on the input parameters of Algorithm 2. Inequality (4.2) and (4.3) are technical constraints on the value of the *step size* η. Equation (4.6) requires the cosine between the (population) target vector $\phi^\star$ and the input vector ϕ_0 to be sufficiently high. We discuss this assumption in the last paragraph of this section.

To obtain a convergence rate and to give explicit constraints on the asymptotic behavior of $p(n), s(n)$ and $k(n)$, we provide the following result.

Proposition 3 *For every $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows where $Y_1, X_{1,1}, ..., X_{1,p(n)}$ are sub-gaussian random variables. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top\beta$. Let $\phi^\star = \beta/\|\beta\|_{\mathbb{R}^p}$, $s = \|\beta\|_0$, $\hat{\Sigma}_X = X^\top X/n$ and $k' = Cs$ for $C > 0$.*

Assume $s\log(p/s)/n = o(1)$ and $s/p = o(1)$ then

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

and

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

Proposition 3 implies that if ϕ_t is the output of Algorithm 2 after t iterations of step 5, then for an appropriate choice of η and ϕ_0, $|(1-|(\phi^\star)^\top\phi_t|)^{1/2} - \nu^t\sqrt{\xi}| = O_p((s\log(p/s)/n)^{1/2})$. Moreover, by assuming that t is a *fast enough* (we give a precise sense to this notion in the next result) function of n and that $(\phi^\star)^\top\phi_t \geq 0$, we have $\|\phi^\star - \phi_t\|_2 = O_p((s\log(p/s)/n)^{1/2})$. As we mention above, by combining Theorem 1 in Raskutti et al. [2011] with Proposition 2, we know that in the Gaussian fixed design, $(s\log(p/s)/n)^{1/2}$ is the minimax rate of convergence for the l_0 constrained estimation of $\phi^\star$.

One can obtain from ϕ_t an estimate of the magnitude of β by regressing Y on the span of $X\phi_t$. Let $\hat{\upsilon}$ be this estimator and $\hat{\beta}$ be the estimator of β obtained as the product of ϕ_t and $\hat{\upsilon}$. Proposition 3 tells us that for $t = t(n)$

fast enough $\|\hat{\beta} - \beta\|_2 = O_p\big((s\log(p/s)/n)^{1/2}\big)$, which is minimax [Raskutti et al., 2011].

Theorem 2 relies heavily on the quality of the initial vector ϕ_0. To obtain a good initial vector, we follow the same heuristic as in Tan et al. [2018], that is, by using a solution of an l_1 relaxation of the l_0-constrained problem as the initial vector, which is the lasso for problem (1.1) (Bertsimas et al. [2016], Hastie et al. [2020]). This choice of initialization makes particular sense since the lasso is known to select too many covariates when tuned for predictive accuracy. Using its (normalized) solution as the initial vector for Algorithm 2 allows us to reduce the size of the selected support while improving the accuracy of the estimation. As the Monte Carlo experiment in Section 5 demonstrates, this procedure outperforms the other l_0-constrained estimators considered in this paper. This initialization procedure is theoretically justified by the fact that it has been established in Chetverikov et al. [2021] that the l_2 error of the cross-validated lasso is $O_p\big((s\log(p/s)/n)^{1/2} \times \log(pn)\big)$. This error ensures that the events on which the assumption on ϕ_0 in Theorems 2 is not verified have a probability shrinking to zero.

Unfortunately, we are not aware of any method to select η that has an analog quality. Experience and Assumption (4.2) Tan et al. [2018] suggest selecting η small enough such that $\eta\lambda(\hat{\Sigma}_X) < 1$.

4.4 Central Limit Theorem

In this section, we provide a central limit theorem for the output of Algorithm 2. In the discussion that follows the results, we state three propositions that analyze in depth the behavior of the last step of Algorithm 2, that is, Algorithm 1. One of these propositions gives sufficient conditions for the convergence of a stochastic Cauchy product and may be of independent interest.

Before stating the results, we give the following definition that is necessary for the statement and proof of the theorem.

Definition 2 *Given $\phi^\star$ and p, let*

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

Moreover, for any ϕ_t that is the output of Algorithm 1 or 2, let

$$Q(\phi_t) = Q_t = P_t P_t^\top$$

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

We note that all ϕ_t s.t. $Q(\phi_t) \in \mathcal{Q}^\star$ have a support containing the support of $\phi^\star$. From Lemma 5 in the Appendix, one can show that in the

sub-Gaussian setting, for $t = t(n)$ fast enough, $Q(\phi_t) \in \mathcal{Q}^\star$ with probability tending to one, i.e., Algorithm 2 possesses the *sure screening property* (Fan and Lv [2008], Guo et al. [2020]).

In the remainder of this section, we make use of the following convention.

Notation 2 *Let $\{a_\alpha\}_{\alpha\in\mathbb{N}}$ be a family of operators on a vector space, if $j < i$ then $\prod_{\alpha=i}^{j} a_\alpha = 1$. where in the above expression 1 symbolize the unitary operator on the vector space. If $j \geq i$ the symbol $\prod$ is used conventionally.*

We now state the central limit theorem for the output of Algorithm 1. Since the last step of Algorithm 2 is an application of Algorithm 1, the extension of the result to the output of Loaded-RIFLE is direct.

Theorem 3 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is finite and positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top\beta$. Let $U = Y - X\beta$, $\sigma^2 = \mathbb{V}ar(U_1)$, $\phi^\star = \beta/\|\beta\|_2$ and $s = \|\beta\|_0$. Let t, η, ϕ_t and C_{t-1} be as in Algorithm 1. Let $\tilde{\phi}_t = C_{t-1}\phi_{t-1}$, let t be an increasing function of n and assume that $\langle\phi^\star, \phi_t\rangle_2 \geq 0$ almost surely.*

If for some constants $\xi > 0$ and $\nu < 1$

$$\|\phi^\star - \phi_t\|_2 \leq \nu^t\sqrt{\xi} + O_p\Big(\sqrt{\frac{s\log(p/s)}{n}}\Big) \tag{4.7}$$

and if $\inf_{\|v\|_2=1,\|v\|_0=k} |v^\top\Sigma_{XY}| > 0$, and $\log(n)/t(n) = o(1)$, then conditionally on X, there exists $Q^\star \in \mathcal{Q}^\star$ and $K > 0$ such that if $\eta < K$ then for all $w \in \mathbb{R}^p$ such that $\|w\|_2 = 1$ we have that

$$\sqrt{n}\langle w, \phi_t - \phi^\star + \gamma_n\rangle_2$$

is

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

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

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

and

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

where for all $m \in \mathbb{N}$

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

The proof of the result is given in Appendix. The bias γ_n is the sum of κ_n, which is observable (and hence can be corrected), and τ_n, which cannot be observed. In the proof, we show that both terms converge to zero but without being able to pin down their rates of convergence. From Assumption (4.7) (which can be shown to hold in the sub-Gaussian setting by combining Proposition 3 with Proposition 7 in the Supplementary Material), we conclude that γ_n must converge to zero at a slower rate than $n^{-1/2}$. The term κ_n accounts for the effect of the normalization in step 3.c) and from the truncation in step 3.e) of Algorithm 1. The term τ_n is mainly driven by the difference between the supports of ϕ_t and $\phi^\star$ and can be understood as a bias introduced by the regularization of the inverse problem posed by the ill-conditioned matrix $\hat{\Sigma}_X$ in regard to estimating $\phi^\star$ in high dimensions.

The keys to the proof of Theorem 3 are three propositions that we now state and comment on. The proofs of the three propositions are given in Appendix.

The first proposition decomposes the estimator ϕ_t to clarify its behavior.

Proposition 4 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ 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 $t \geq 2$*

$$\begin{aligned}\phi_t =& \left(\prod_{m=1}^{t} \Gamma_{t-m}\right)\phi_0 + \sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}\Gamma_{t-l}\right)(\phi_{t-l} - \tilde{\phi}_{t-l}) \\ &+ \eta \sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}\Gamma_{t-l}\right)\frac{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi_{t-1-m}}{\phi_{t-1-m}^\top \hat{\Sigma}_X \phi^\star + \phi_{t-1-m}^\top \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p} n}}\left(\hat{\Sigma}_X \phi^\star + \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p} n}\right)\end{aligned}$$

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

The first term explains the numerical dependence of ϕ_t on ϕ_0. From Lemma 1 in the Appendix, 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 1. Using Proposition 5 below 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 6.

The next proposition, which might be of independent interest, gives conditions for the convergence in probability of a stochastic Cauchy product.

Proposition 5 *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* $\sum_{i=0}^{m} a_{m-i}b_i = o_p(1)$.

The following and last proposition study the convergence of the Cauchy product in the last term of the decomposition of Proposition 4.

Proposition 6 *Under the same assumptions than in Theorem 3. Conditionally on* X*, there exists* $\eta > 0$ *and* $Q \in \mathcal{Q}^\star$ *such that*

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

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 4 and thus obtaining the exact asymptotic variance given in Theorem 3.

5 Monte Carlo experiments

In our experiments, we use Algorithm 2 initialized by a solution of lasso to estimate $\phi^\star$. From this estimator, we calculate an estimate of β by using the strategy described in Section 4.1.

We compare the performances of Loaded-RIFLE against those of three well-studied l_0-constrained estimators of β that we normalized to find estimators of $\phi^\star$. The competitors are the IHT (Blumensath and Davies [2008], Blumensath and Davies [2009b], Jain et al. [2014]) that we implement as in [Blumensath and Davies, 2009a], the two-stage IHT Jain et al. [2014] and the SDAR Huang et al. [2018].

5.1 Implementation

In each experiment we let $n = 100, p = 1000$ and generate $X \in \mathbb{R}^{n\times p}$ *iid* across rows such that $X_1 \sim \mathcal{N}_p(0, \Sigma)$ and $Y|X \in \mathbb{R}^n$ such that $Y \sim \mathcal{N}_n(X\beta, \beta^\top \Sigma_X \beta/m)$ where m is the *signal-to-noise ratio* (SNR) and β is generated such that $\|\beta\|_0 = s = 10$ with every element of β having an equal probability to be assigned a value of 1. We consider four scenarios that go from the least correlated design to a highly correlated design.

Scenario 1 *We took* $\Sigma = I_{p\times p}$.

Scenario 2 *We took* $\Sigma_{i,j} = 0.5^{|i-j|}$.

Scenario 3 *We let* $W \in \mathbb{R}^{p\times p}$ *such that* $W_{i,j} = 0.5^{|i-j|}$ *and took* $\Sigma = U^\top DU$ *where* U *is the matrix of eigenvectors of* W*,* $D_{j,j} = 10/\sqrt{j}$ *and* $D_{i,j} = 0$ *when* $i \neq j$*.*

Scenario 4 *We let* $W \in \mathbb{R}^{p\times p}$ *such that* $W_{i,j} = 0.5^{|i-j|}$ *and took* $\Sigma = U^\top DU$ *where* U *is the matrix of eigenvectors of* W*,* $D_{j,j} = 10/j$ *and* $D_{i,j} = 0$ *when* $i \neq j$*.*

As pointed out in Hastie et al. [2020], the performances of estimators vary greatly with the SNR, and it may happen that one estimator outperforms another for a high SNR while underperforming in low SNR settings. Hence, we consider a range of eleven signal-to-noise ratios for the four scenarios. Since the relative performances of the algorithms often change before and after an SNR level of 1, we choose 5 equally spaced values between 0.05 and 1 and between 1 and 6.

0.05	0.24	0.43	0.62	0.81	1.00	2.00	3.00	4.00	5.00	6.00

Table 2: Signal-to-noise ratios considered in the three scenarios.

We fix the sparsity level of the estimated vectors at $k = 12$ for all the estimators. Hence, this experiment assumes that a (good) prior on the sparsity level s is available. This means we do not problematize here the selection of the parameter k. There exist other experimental paradigms where the tuning parameter is selected by cross-validation (as in Hastie et al. [2020] or via BIC-type criteria (as in Huang et al. [2018]).

SDAR and pseudo-lasso have been initialized with the solution of a ridge regression with tuning parameter $2s\log(p)/n$. Loaded-RIFLE has been initialized with a solution from lasso that has approximately $3k = 36$ nonzero components. In addition to k, Loaded-RIFLE and two-stage IHT both depend on $\overline{k}$, a parameter that plays different technical functions in the two algorithms but that encompasses the same heuristic, namely, preoptimizing on a support of dimension $\overline{k} > k$ before selecting a support of k on which the final optimization takes place [Jain et al., 2014]. For both algorithms, we set $\overline{k} = 4k = 48$. For loaded-RIFLE, we set the *step-size* $\eta = 0.5/\|\hat{\Sigma}_X\|_{spec}$. For IHT, the step size changes at each iteration following the rules proposed in Blumensath and Davies [2009a]. For Loaded-RIFLE, we set $T = \log(n) \approx 5$. Following the theory and practice in Huang et al. [2018], we fed SDAR with X and Y, after dividing them by $\sqrt{n}$.

Let $\hat{\beta}$ be an estimate of β. For each experiment, we record three accuracy metrics.

Estimation metric: measuring the (relative) l_2 error of estimation of $\phi^\star$, namely $\hat{\beta}/\|\hat{\beta}\|_2 - \beta/\|\beta\|_2$.

Selection metric: measuring the recovery of the true support, that is a binary variable taking value 1 if the support of β is contained in the support

of $\hat{\beta}$, and 0 otherwise.

Prediction metric: The relative risk (e.g. Bertsimas et al. [2016], Hastie et al. [2020]) that is, $\mathbb{E}[X_1^\top \hat{\beta} - X_1^\top \beta]^2 / \mathbb{E}[X_1^\top \beta]^2 = (\hat{\beta} - \beta)^\top \Sigma (\hat{\beta} - \beta) / \beta^\top \Sigma \beta$.

5.2 Results

For each scenario and each signal-to-noise ratio, we repeated the experiment 100 times. The results of the *estimation metric* are displayed in Figure 1. The results of the *selection metric* are displayed in Table 3 for SNRs higher than 2 (for SNRs strictly lower than two, no algorithm was able to recover complete true support at least once). The results for the *prediction metric* are displayed in Table 4 for three SNRs.

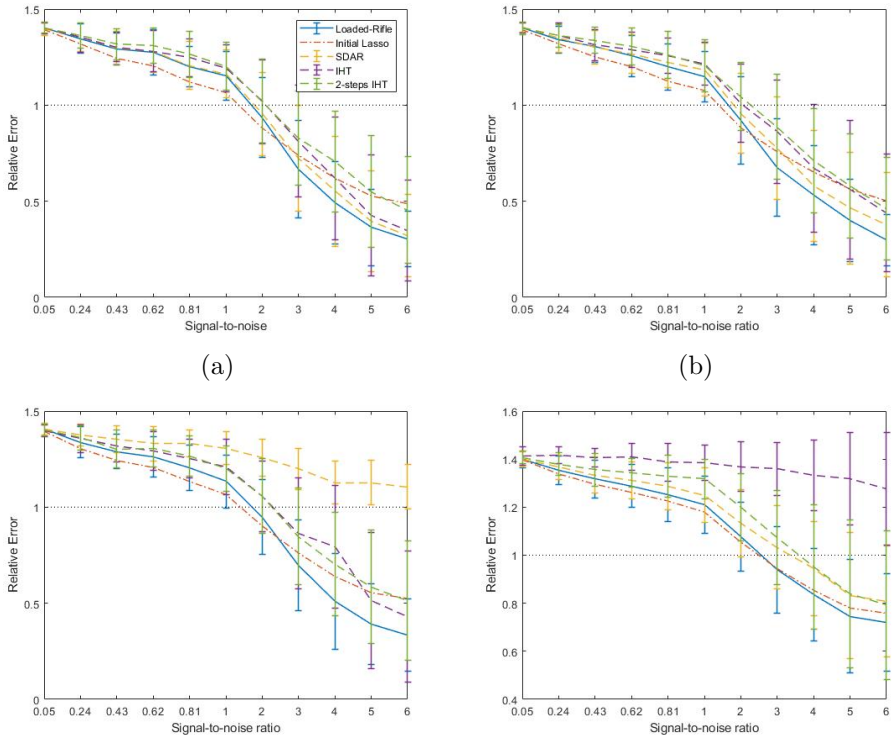


Figure 1: Mean and standard deviation of the relative error of estimation of $\phi^\star = \beta / \|\beta\|_2$ across various signal-to-noise ratios in (a) Scenario 1, (b) Scenario 2, (c) Scenario 3 and (d) Scenario 4. In all the figures, the horizontal dot line represents the score of the null vector.

We draw four main conclusions from the experiment. First, Loaded-RIFLE initialized with lasso outperforms its competitors both in terms of

	Signal-to-noise ratio	2	3	4	5	6
Scenario 1	Loaded-RIFLE	**0.02**	**0.18**	0.4	**0.72**	**0.82**
	SDAR	0.01	0.16	**0.41**	0.68	0.81
	IHT	**0.02**	0.09	0.33	**0.72**	**0.82**
	Two-stages IHT	*	0.01	0.09	0.28	0.43
Scenario 2	Loaded-RIFLE	**0.03**	**0.19**	**0.41**	**0.61**	**0.84**
	SDAR	*	0.11	0.36	0.54	0.69
	IHT	0.01	0.08	0.3	0.47	0.63
	Two-stages IHT	*	0.05	0.1	0.22	0.41
Scenario 3	Loaded-RIFLE	**0.01**	**0.14**	**0.43**	**0.65**	**0.75**
	SDAR	*	*	*	*	*
	IHT	*	0.08	0.16	0.58	0.69
	Two-stages IHT	*	0.05	0.1	0.23	0.36
Scenario 4	Loaded-RIFLE	*	*	0.02	**0.08**	**0.1**
	SDAR	*	*	0.02	0.07	0.03
	IHT	*	*	*	0.01	0.01
	Two-stages IHT	*	*	**0.03**	0.05	**0.1**

Table 3: Frequency at which the support selected by the algorithms contained the true support of β for signal-to-noise ratio greater or equal to 2. The asterisk are zeros. The best result for each scenario are written in bold.

Signal-to-noise ratio		0.05		1		6	
		mean	sd	mean	sd	mean	sd
Scenario 1	Loaded-RIFLE	15.878	**2.8**	**1.632**	**0.3**	**0.109**	**0.1**
	SDAR	**14.627**	4.5	1.666	0.3	0.145	0.2
	IHT	18.509	3.8	1.895	0.4	0.181	0.2
	Two-stages IHT	15.848	2.9	1.737	0.3	0.255	0.2
Scenario 2	Loaded-RIFLE	16.013	**2.8**	**1.546**	**0.3**	**0.096**	**0.1**
	SDAR	**15.196**	4.6	1.644	0.4	0.181	0.2
	IHT	18.136	4.6	1.938	0.3	0.248	0.2
	Two-stages IHT	16.586	3.7	1.743	0.4	0.245	0.2
Scenario 3	Loaded-RIFLE	15.734	2.8	1.477	0.3	**0.131**	0.1
	SDAR	**4.679**	**0.9**	**1.175**	**0.1**	0.856	**0.1**
	IHT	16.337	4.9	1.830	0.5	0.267	0.3
	Two-stages IHT	16.354	3.1	1.710	0.3	0.312	0.3
Scenario 4	Loaded-RIFLE	**11.691**	**2.2**	**0.962**	**0.2**	**0.258**	**0.1**
	SDAR	14.912	3.3	1.239	0.3	0.336	0.2
	IHT	20.168	10.3	2.895	1.1	1.908	1.0
	Two-stages IHT	20.707	5.3	1.831	0.3	0.381	0.2

Table 4: Mean and standard deviation (sd) of the Relative Risk of the four estimators in the four scenarios for three signal-to-noise ratio. The best results for each scenario are written in bold.

variable selection and estimation as well as for prediction in most scenarios. This conclusion is particularly observable for SNRs higher than two and in the more correlated designs (scenarios 3 and 4). Second, for SNRs higher than 2, Loaded-RIFLE improves on the initial lasso solution *while* significantly reducing the size of the estimated support. Third, for SNRs lower than 2, no estimator was able to recover the true support at least once. Fourth, for SNRs lower than 2, all the estimators have a relative error higher than one, that is, the score of a null vector.

6 Conclusion

We propose to interpret the parameter vector of the high-dimensional sparse linear model as the product between a basis in *the direction of interest*, which is identified as the generalized eigenspace of a pair of measurable matrices, and the magnitude relative to this basis.

On the one hand, this approach sheds new light on the well-studied linear regression while inducing a parameterization that emerges naturally in economic theory. To exemplify this, let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$ be random variables such that $\mathbb{E}(Y|X) = X^\top \phi\nu$ where ϕ is some basis in Φ and $\nu \in \mathbb{R}$ is the *magnitude relative to the basis*. As a first example, let Y be the random log-output of a Cobb-Douglass production function, X be the random vector of the quantity of factors of production, and ϕ such that $\sum_{j=1}^p \phi_j = 1$. Then, the return to scale of the production function is decreasing, constant or increasing when ν is lower, equal to or higher than 1, respectively. A second example is the *Cambridge Equation* (Pigou [1917]) where Y is the random money demand, X is the random vector of good to be exchanged, ϕ is the vector of prices, and ν the proportion of money detained for convenience by individuals or, if equilibrium on the money market is assumed, the inverse of the velocity of money (Fisher and Brown [1911]). Third, we observe that the parametrization is induced by the relativity of prices. In the easiest example of this, let Y be a random quantity of dollars, X be a random vector of quantity of goods and ϕ be the vector of price expressed in terms of Euro; then, ν is the exchange rate between the two currencies. As it appears from these examples, the choice of the basis $\phi \in \Phi$ should depend on the context to study; for instance, one could choose ϕ such that $\|\phi\| = 1$ for some norm; nevertheless, infinitely many other specifications are available.

On the other hand, our approach connects the linear regression with the generalized eigenvalue problem and extends the class of estimators available for the former to the one of the latter. This is especially fruitful in the context of high dimensions where standard optimal procedures of estimation are not available.

We show how to apply this last idea by tackling the best subset selection problem via a novel estimator called directional best subset selection. We

provide a theory connecting the solutions of the two problems as well as the minimax rate of convergence of their l_2 error of estimation. This leads us to approximate the solution of the directional best subset selection via a novel estimator based on RIFLE, which is a well-established algorithm developed by Tan et al. [2018] and designed to solve sparse generalized eigenvalue problems in a high-dimensional context. We demonstrate that this proposition leads to an estimator that has the minimax rate of convergence in l_2 for the sub-Gaussian setting. Finally, we provide a central limit theorem for the estimator. To do so, we demonstrate a theorem that gives sufficient conditions for the convergence of a stochastic Cauchy product, which might be of independent interest. Our simulations show that our method is competitive and can outperform existing competitors.

Our estimator is an answer to the important best-subset selection problem of selecting k out of p covariates given n observations. While our estimator does not equal the best subset selection, it still allows an l_0-*constrained* estimation while attaining the minimax rate of convergence of the best subset selection and providing solutions in less than a second, even for large problems. This is one of the rare propositions in the literature to possess these qualities altogether.

Although this paper focuses on the l_0 constrained regression and concentrates on the estimator derived from RIFLE, the result of the second section applies or may be generalized more broadly. In particular, other estimators of GEP exist (*e.g.,* Safo, Ahn, Jeon, and Jung [2018], Jung, Ahn, and Jeon [2019] and Yuan, Shen, and Zheng [2019]) and might be of interest, for instance, to solve least squares problems constrained by norms other than l_0. More generally, the connection between least squares programs and GEP opens the possibilities of bridging new numerical techniques into high-dimensional estimation theory.

Finally, in this paper, we have assumed the exogeneity of the parameters. However, the theoretical study of endogeneity, as well as of instrumental variables (IV) regressions, in high dimensions are questions at the edge of the research frontier in econometrics (*e.g.* Belloni, Chen, Chernozhukov, and Hansen [2012], Fan and Liao [2014], Chernozhukov, Hansen, and Spindler [2015b], Chernozhukov, Hansen, and Spindler [2015a], Gold, Lederer, and Tao [2020]). Further research might explore the possibility of providing an extension of Theorem 1 to the case of an IV regression and study the opportunity it could open for inference. It might also be fruitful to explore the possibility of estimating simultaneous linear equations by using the GEP approach. Since the general form of the GEP of a pair of matrices (A, B) allows the matrix A to have a rank higher than 1, it might be of interest to explore an extension of Theorem 1 to the case where the explanatory variable is not a scalar but a vector.

A Appendix

The proof of Theorems 2, while correcting and extending results given in Tan et al. [2018], employ the same techniques as in the aforementioned paper and hence are given in the Supplementary Material. The proof of Proposition 3 employs known techniques hence is also given in the Supplementary Material.

A.1 Proof of Theorem 1

Let $\upsilon = \|\beta\|_2$. The linearity X_1 of the conditional expectation implies $\Sigma_{XY} = \Sigma_X \phi^\star \upsilon$. Hence we have

$$\frac{(\phi^\star)^\top \Sigma_{XY}}{\|\phi^\star\|^2_{\Sigma_X}} = \upsilon. \tag{A.1}$$

Substituting equation (A.1) into the functional form of the conditional expectation we obtain

$$\mathbb{E}(Y_1|X_1) = X_1^\top \frac{\phi^\star (\phi^\star)^\top \Sigma_{XY}}{\|\phi^\star\|^2_{\Sigma_X}} \tag{A.2}$$

Since $\mathbb{E}(Y_1|X_1)$ is a projection onto the vector space of real-valued linear functions of X that are square integrable, we have that

$$\begin{aligned}
\phi^\star &\in \operatorname*{argmin}_{\phi\in\mathbb{R}^p} \mathbb{E}\left[Y_1 - X_1^T \frac{\phi\phi^\top \Sigma_{XY}}{\|\phi\|^2_{\Sigma_X}}\right]^2 \\
&\in \operatorname*{argmin}_{\phi\in\mathbb{R}^p} \mathbb{E}\left(Y_1^2 - 2Y_1 X_1^T \frac{\phi\phi^\top \Sigma_{XY}}{\|\phi\|^2_{\Sigma_X}} + \frac{\Sigma_{YX}\phi}{\|\phi\|^2_{\Sigma_X}} \phi^\top X_1 X_1^\top \phi \frac{\phi^\top \Sigma_{XY}}{\|\phi\|^2_{\Sigma_X}}\right) \\
&\in \operatorname*{argmin}_{\phi\in\mathbb{R}^p} \mathbb{E}(Y_1^2) - 2\frac{\Sigma_{YX}\phi\phi^\top \Sigma_{XY}}{\|\phi\|^2_{\Sigma_X}} + \frac{\Sigma_{YX}\phi}{\|\phi\|^2_{\Sigma_X}} \|\phi\|^2_{\Sigma_X} \frac{\phi^\top \Sigma_{XY}}{\|\phi\|^2_{\Sigma_X}} \\
&\in \operatorname*{argmax}_{\phi\in\mathbb{R}^p} \left(\frac{\phi}{\|\phi\|_{\Sigma_X}}\right)^\top \Sigma_{XY}\Sigma_{YX} \frac{\phi}{\|\phi\|_{\Sigma_X}}.
\end{aligned}$$

Let $\mathcal{S}$ be the solution set of the previous problem and observe that it is closed under scalar multiplication. Let $\mathcal{S}_1 := \{\phi \in \mathcal{S} \text{ s.t. } \|\phi\|_{\Sigma_X} = 1\}$ and $\mathcal{S}_2 := \{\phi \in \mathcal{S} \text{ s.t. } \|\phi\|_{\Sigma_X} \neq 1\}$, hence we have

$$\begin{aligned}
&\mathcal{S} = \mathcal{S}_1 \sqcup \mathcal{S}_2 \text{ (disjoint union)}, \\
&\text{if } \phi \in \mathcal{S}_2 \text{ then } \frac{\phi}{\|\phi\|_{\Sigma_X}} \in \mathcal{S}_1, \text{ and} \\
&\text{if } \phi \in \mathcal{S}_1 \text{ then } \alpha\phi \in \mathcal{S}_2 \text{ for all } \alpha \neq 1.
\end{aligned}$$

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

$$\operatorname*{argmax}_{\phi\in\mathbb{R}^p} \phi^\top \Sigma_{XY}\Sigma_{YX}\phi \qquad \text{such that } \|\phi\|^2_{\Sigma_X} = 1. \tag{A.3}$$

By definition of $\|.\|_{\Sigma_X}$ the Lagrangian of problem (A.3) 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

$$\Sigma_{XY}\Sigma_{YX}\phi = \Sigma_X \phi\lambda \tag{A.4}$$

which is a scalar equation of the generalized eigenvalue problem of the matrix pair $\big(\Sigma_{XY}\Sigma_{YX}, \Sigma_X\big)$. Pre-multiplying each side of this equation by $\phi^\top$ identifies λ as the maximum generalized eigenvalue which is why we symbolize it by $\lambda_{\max}$. Remark that if ϕ satisfies (A.4) then $\alpha\phi$ also satisfies this equation for any $\alpha \in \mathbb{R}$. This last remark allows us to conclude that

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

which is the result to be proven.□

A.2 Proof of Proposition 1

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 $(\hat{\phi}, \hat{m}) \in \mathrm{argmin}_{(\phi,m)\in\mathcal{O}(p,s)\times\mathbb{R}} n^{-1}\|Y - X\phi m\|_2^2$ then $\tilde{\beta} = \hat{\phi}\hat{m}$ almost surely. Now observe that *for any* $\phi \in \mathcal{O}(p,k)$ we have

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

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 $\mathcal{D}(\tilde{\beta}) = \mathcal{D}(\hat{\phi})$ almost surely. Finally, computations similar to ones in the proof of Theorem 1 lead to

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

that gives the result. □

A.3 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

$P \in \mathcal{P}(F(p,l))$ be defined as in section 4.3, such that $l \leq 2s$ 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 Cauchy–Schwarz inequality, the sub-multiplicativity of the spectral norm and $\|P\|_{\text{spec}} = 1$. Hence

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

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

$$\begin{aligned}
\|\hat{\beta} - \beta\|_2^2 &= \|\hat{\upsilon}\hat{\phi} - \upsilon\phi^\star\|_2^2 \\
&= \hat{\upsilon}^2 - 2\hat{\upsilon}\upsilon\langle\hat{\phi}, \phi^\star\rangle_2 + \upsilon^2 && \text{(Since } \|\hat{\phi}\|_2 = \|\phi^\star\|_2 = 1\text{)} \\
&= (\hat{\upsilon} - \upsilon)^2 + 2\hat{\upsilon}\upsilon(1 - \langle\hat{\phi}, \phi^\star\rangle_2) \\
&= (\hat{\upsilon} - \upsilon)^2 + \hat{\upsilon}\upsilon\|\hat{\phi} - \phi^\star\|_2^2 && \text{(Since } \langle\hat{\phi}, \phi^\star\rangle_2 \geq 0\text{)} \\
&= (\hat{\upsilon} - \upsilon)^2 + \big((\hat{\upsilon} - \upsilon)\upsilon + \upsilon^2\big)\|\hat{\phi} - \phi^\star\|_2^2,
\end{aligned}$$

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.4 Proof of Proposition 4

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 consequences 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

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

Hence we have

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

which is equivalent to

$$\begin{aligned}
\phi_t' = \phi_{t-1} + \frac{\eta}{2}\frac{\|X\phi_{t-1}\|_2^2}{n}\nabla_{\phi_{t-1}}\big\{&2\log(\|\beta\|_2) + 2\log(\frac{\|X\phi_{t-1}\|_2^2}{n}) + \\
&2\log(1 + \frac{\phi_{t-1}^\top X^\top X(\phi^\star - \phi_{t-1})}{\|X\phi_{t-1}\|_2^2} + \frac{\phi_{t-1}^\top X^\top U}{\|X\phi_{t-1}\|_2^2\|\beta\|_2}) - \log(\frac{\|X\phi_{t-1}\|_2^2}{n})\big\}
\end{aligned}$$

simplifying we get

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

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

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

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

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

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

$$\begin{aligned}\tilde{\phi}_t &= \phi_{t-1} - \eta\hat{\Sigma}_X\phi_{t-1} + \eta\frac{1}{g_{t-1}}\hat{\Sigma}_X\phi^\star + \eta\frac{1}{g_{t-1}}\frac{X^\top U}{\|\beta\|_{\mathbb{R}^p}n} \\ &= (I_{p\times p} - \eta\hat{\Sigma}_X)\phi_{t-1} + \eta\frac{1}{g_{t-1}}(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p}n}) \\ &= (I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-1}\phi_{t-1} + \eta\frac{1}{g_{t-1}}(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p}n}),\end{aligned} \tag{A.5}$$

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

$$\begin{aligned}\tilde{\phi}_t = (I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-1}\tilde{\phi}_{t-1} \\ + (I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-1}(\phi_{t-1} - \tilde{\phi}_{t-1}) \\ + \frac{\eta}{g_{t-1}}(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_{\mathbb{R}^p}n})\end{aligned} \tag{A.6}$$

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

Then by recursively applying equation (A.6) 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

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

we obtain the thesis. □

A.5 Proof of Proposition 5

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

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

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 all $\gamma > 0$ and all $m > M(N_2)$ that

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

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

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

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

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

Hence we have for all $m > m(N_2) + m(N_3)$

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

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

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

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 every $\epsilon, \tilde{\epsilon}$

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

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

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

where

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

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 be 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 $\mathbb{P}\Big(\sum_{i=0}^{m} a_{m-i}b_i > \gamma \Big) < \epsilon$ which is what needed to be proven. □

A.6 Proof of Proposition 6

A.6.1 Lemmas

Lemma 1 *For all $n \in \mathbb{N}$ let $X \in \mathbb{R}^{n\times p}$ be a random matrix* iid *across and such that $\mathbb{E}(\|X_1\|^2) < \infty$. Let $\hat{\Sigma}_X = X^\top X/n$, and $P \in \mathcal{P}(\mathcal{F}(p,k))$ for some $k \in \mathbb{N}$, $k < p$.*

Then conditionally on X, if $P^\top \hat{\Sigma}_X P$ is positive definite a.s. then there exists $\eta > 0$ such that

$$\|(I_{p\times p} - \eta\hat{\Sigma}_X)P\|_{\text{spec}} < 1 \qquad a.s.$$

Proof: The proof is given in the Supplementary Material

Lemma 2 *Under the same set of assumptions as in Theorem 3 we have*

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

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

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

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

Proof The proof is given in the Supplementary Material.

Lemma 4 *For all* $n \in \mathbb{N}$*, let* $Y \in \mathbb{R}^n$ *and* $X \in \mathbb{R}^{n\times p}$ *be random functions* iid *across rows. For all* $n \in \mathbb{N}$ *assume* $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ *is finite and positive definite and there exists* $\beta \in \mathbb{R}^p$ *s.t.* $\mathbb{E}(Y_1|X_1) = X_1^\top \beta$ *and* $\|\beta\|_0 = s$*. Let* $\phi^\star = \beta/\|\beta\|_2$*. Let* t *be as in Algorithm 1 and assume that* $t \to \infty$ *as* $n \to \infty$*.*

If conditionally X*, for given* k *and* ϕ_0*, for any* η*, for all* $\phi_t \in \Phi(\phi_0, k, \eta)$ *we have that* $P_t^\top \hat{\Sigma}_X P_t$ *is positive definite a.s. then there exists* η *and* $M_\eta < \infty$ *such that, conditionally on* X*,*

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

Proof. Note that

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

and since for all $l \in \{1, ..., t-1\}$, by construction we have $\|P_{t-l}\|_{\text{spec}} = 1$ and by Lemma 1, 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 5 *Under the same assumptions than in Theorem 3, there exists* $\alpha > 0$ *such that we have* $\mathbb{P}\big(Q(\phi_t) \in \mathcal{Q}^\star \big| \|\phi_{\tilde{t}} - \phi^\star\|_2 \leq \alpha\big) = 1$ *for all* $t \geq \tilde{t}$*, where* Q *and* $\mathcal{Q}^\star$ *are defined in Equation* (**??**) *and thereafter.*

Proof: The proof is given in the Supplementary Material

A.6.2 Proof of Proposition 6

The existence of η follows from Lemma 1. For the remaining 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. conditionally on X.

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 that $\Gamma_0 = I_{p\times p}$) 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 $L := g_{t-1}I_{p\times p} + g_{t-2}\Gamma_1 + ... + g_0\Gamma_{t-1}$ and $M := I_{p\times p} + \Gamma_1 + ... + \Gamma_{t-1}$ as a consequence $L - M = (g_{t-1} - 1)I_{p\times p} + (g_{t-2} - 1)\Gamma_1 + ... + (g_0 - 1)\Gamma_{t-1}$. Lemmas 2, 3 and 4 ensures that by Proposition 5 we have $\|L - M\|_{\text{spec}} = o_p(1)$. Hence we have shown that $L = M + o_p(1)$. Now let's consider M. For any $\alpha > 0$ and $t' \in \mathbb{N}$, let $\mathcal{E}_{\alpha t'} := \{\|\phi_{t'} - \phi^\star\|_2 \leq \alpha\}$, then we can write $M = M \times \mathbf{1}_{\mathcal{E}_{\alpha\tilde{t}}} + M \times \mathbf{1}_{\mathcal{E}^c_{\alpha t'}}$. Given our assumptions about the convergence of ϕ_t, we know that for any $\alpha > 0$ and some t' high enough we have $M \times \mathbf{1}_{\mathcal{E}^c_{\alpha t'}} = o_p(1)$. Furthermore by Lemma 5, by choosing α low enough we have 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 5 and dealing with indices, the right hand-side can be rewritten as

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

Since by submutiplicativity of the spectral norm and Lemma 1 $\|(I-\eta\hat{\Sigma}_X)Q^\star\|_{\text{spec}} \leq \|(I - \eta\hat{\Sigma}_X)P^\star\|_{\text{spec}} < 1$, 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 since $\left((I - \eta\hat{\Sigma}_X)Q^\star\right)^{t-t'} = o_p(1)$ and since

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

because $t - t' \to \infty$ as $n \to \infty$, 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. □

A.7 Proof of Theorem 3

A.7.1 Lemmas

Lemma 6 *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 7 *Under the same set of assumptions as in Theorem 3 we have (and in particular if* $\inf_{\|v\|_2=1,\|v\|_0=k} |v^\top \Sigma_{XY}| > 0$ *and given the definition of* $\tilde{\phi}$*), then* $\sup_{t\in\{1,\dots,t(n)\}} \|\phi_t - \tilde{\phi}_t\|_2 = O_p(1)$.

Proof: The proof is given in the Supplementary Material

A.7.2 Proof of Theorem 3

The existence of η follows from Lemma 1. For the remaining 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. conditionally on X.

By Proposition 4 we have for all $t \geq 2$

$$\phi_t - \phi^\star = \left(\prod_{m=1}^{t}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-m}\right)\phi_0 \tag{A.7}$$

$$+ \sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\right)(\phi_{t-l} - \tilde{\phi}_{t-l}) \tag{A.8}$$

$$+ \eta\sum_{m=0}^{t-1}\left(\prod_{l=1}^{m}(I_{p\times p} - \eta\hat{\Sigma}_X)Q_{t-l}\right)g_{t-1-m}\left(\hat{\Sigma}_X\phi^\star + \frac{X^\top U}{\|\beta\|_2 n}\right) - \phi^\star \tag{A.9}$$

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.7). By Cauchy–Schwarz

inequality we have

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

and by Lemma 1

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

Now we treat terms (A.8). Lemmas 4, 6 and 7 ensure that by Proposition 5, 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.9). First consider

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

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

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

Then by Proposition 6, we have that this quantity converges to zero in probability conditionally on X. Finally, Proposition 6, Slutsky's Lemma and the Central Limit Theorem 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. □

References

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Alexandre Belloni, Victor Chernozhukov, and Christian B. Hansen. *Inference for High-Dimensional Sparse Econometric Models*, volume 3 of *Econometric Society Monographs*, page 245–295. Cambridge University Press, 2013. doi: 10.1017/CBO9781139060035.008.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *Ann. Statist.*, 44(2):813–852, 2016. ISSN 0090-5364. doi: 10.1214/15-AOS1388. URL https://doi.org/10.1214/15-AOS1388.

Thomas Blumensath and Michael E Davies. How to use the iterative hard thresholding algorithm. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009a.

Thomas Blumensath and Mike E Davies. Iterative thresholding for sparse approximations. *Journal of Fourier analysis and Applications*, 14(5):629–654, 2008.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009b.

George E. P. Box and R. Daniel Meyer. An analysis for unreplicated fractional factorials. *Technometrics*, 28(1):11–18, 1986. ISSN 0040-1706. doi: 10.2307/1269599. URL https://doi.org/10.2307/1269599.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998. ISSN 1064-8275. doi: 10.1137/S1064827596304010. URL https://doi.org/10.1137/S1064827596304010.

Xin Chen, Changliang Zou, and R. Dennis Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.*, 38(6):3696–3723, 2010. ISSN 0090-5364. doi: 10.1214/10-AOS826. URL https://doi.org/10.1214/10-AOS826.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–90, 2015a.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, 7(1):649–688, 2015b.

Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated Lasso in high dimensions. *Ann. Statist.*, 49(3):1300–1317, 2021. ISSN 0090-5364. doi: 10.1214/20-aos2000. URL https://doi.org/10.1214/20-aos2000.

R. Dennis Cook. On the interpretation of regression plots. *J. Amer. Statist. Assoc.*, 89(425):177–189, 1994. ISSN 0162-1459. URL http://links.jstor.org/sici?sici=0162-1459(199403)89:425<177:OTIORP>2.0.CO;2-Q&origin=MSN.

Jianqing Fan and Runze Li. Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *International Congress of Mathematicians. Vol. III*, pages 595–622. Eur. Math. Soc., Zürich, 2006.

Jianqing Fan and Yuan Liao. Endogeneity in high dimensions. *Annals of statistics*, 42(3):872, 2014.

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

I. Fisher and H.G. Brown. *The purchasing power of money*. Macmillan, 1911. URL https://books.google.be/books?id=IOwJAAAAIAAJ.

J.P. Florens, V. Marimoutou, A. Peguin-Feissolle, J. Perktold, and M. Carrasco. *Econometric Modeling and Inference*. Themes in Modern Econometrics. Cambridge University Press, 2007. ISBN 9781139466776. URL https://books.google.nl/books?id=MrkUeviIvYkC.

Benyamin Ghojogh, Fakhri Karray, and Mark Crowley. Eigenvalue and generalized eigenvalue problems: Tutorial, 2019.

David Gold, Johannes Lederer, and Jing Tao. Inference for high-dimensional instrumental variables regression. *Journal of Econometrics*, 217(1):79–111, 2020.

G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013. ISBN 9781421407944. URL https://books.google.no/books?id=X5YfsuCWpxMC.

Eitan Greenshtein. Best subset selection, persistence in high-dimensional statistical learning and optimization under $l_1$constraint. *Ann. Statist.*, 34 (5):2367–2386, 2006. ISSN 0090-5364. doi: 10.1214/009053606000000768. URL https://doi.org/10.1214/009053606000000768.

Yongyi Guo, Ziwei Zhu, and Jianqing Fan. Best subset selection is robust against design dependence. *arXiv preprint arXiv:2007.01478*, 2020.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.

Jian Huang, Yuling Jiao, Yanyan Liu, and Xiliang Lu. A constructive approach to L_0 penalized regression. *J. Mach. Learn. Res.*, 19:Paper No. 10, 37, 2018. ISSN 1532-4435.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/218a0aefd1d1a4be65601cc6ddc1520e-Paper.pdf.

Sungkyu Jung, Jeongyoun Ahn, and Yongho Jeon. Penalized orthogonal iteration for sparse estimation of generalized eigenvalue problem. *J. Comput. Graph. Statist.*, 28(3):710–721, 2019. ISSN 1061-8600. doi: 10.1080/10618600.2019.1568014. URL https://doi.org/10.1080/10618600.2019.1568014.

Lexin Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007. ISSN 0006-3444. doi: 10.1093/biomet/asm044. URL https://doi.org/10.1093/biomet/asm044.

Alan Miller. *Subset selection in regression*, volume 95 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2002. ISBN 1-58488-171-2. doi: 10.1201/9781420035933. URL https://doi.org/10.1201/9781420035933.

Baback Moghaddam, Yair Weiss, and Shai Avidan. Generalized spectral bounds for sparse lda. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 641–648, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143925. URL https://doi.org/10.1145/1143844.1143925.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995. ISSN 0097-5397. doi: 10.1137/S0097539792240406. URL https://doi.org/10.1137/S0097539792240406.

A. C. Pigou. The Value of Money. *The Quarterly Journal of Economics*, 32 (1):38–65, 11 1917. ISSN 0033-5533. doi: 10.2307/1885078. URL https://doi.org/10.2307/1885078.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2165799. URL https://doi.org/10.1109/TIT.2011.2165799.

Sandra E. Safo, Jeongyoun Ahn, Yongho Jeon, and Sungkyu Jung. Sparse generalized eigenvalue problem with application to canonical correlation analysis for integrative analysis of methylation and gene expression data. *Biometrics*, 74(4):1362–1371, 2018. ISSN 0006-341X. doi: 10.1111/biom.12886. URL https://doi.org/10.1111/biom.12886.

Xiaotong Shen, Wei Pan, Yunzhang Zhu, and Hui Zhou. On constrained and regularized high-dimensional regression. *Ann. Inst. Statist. Math.*, 65(5):807–832, 2013. ISSN 0020-3157. doi: 10.1007/s10463-012-0396-3. URL https://doi.org/10.1007/s10463-012-0396-3.

G.W. Stewart. Pertubation bounds for the definite generalized eigenvalue problem. *Linear Algebra and its Applications*, 23:69 – 85, 1979. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(79)90094-6. URL http://www.sciencedirect.com/science/article/pii/0024379579900946.

Victoria Stodden. *Model selection when the number of variables exceeds the number of observations*. PhD thesis, 2006. URL http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:3235357. Thesis (Ph.D.)–Stanford University.

Liang Sun, Shuiwang Ji, and Jieping Ye. 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, page 977–984, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553499. URL https://doi.org/10.1145/1553374.1553499.

Kean Ming Tan, Zhaoran Wang, Han Liu, and Tong Zhang. Sparse generalized eigenvalue problem: optimal statistical rates via truncated Rayleigh flow. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 80(5):1057–1086, 2018. ISSN 1369-7412. doi: 10.1111/rssb.12291. URL https://doi.org/10.1111/rssb.12291.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1996)58:1<267:RSASVT>2.0.CO;2-G&origin=MSN.

Sara van de Geer. *Estimation and testing under sparsity*, volume 2159 of *Lecture Notes in Mathematics*. Springer, [Cham], 2016. ISBN 978-3-319-32773-0; 978-3-319-32774-7. doi: 10.1007/978-3-319-32774-7. URL 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].

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

Ganzhao Yuan, Li Shen, and Wei-Shi Zheng. A decomposition algorithm for the sparse generalized eigenvalue problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6113–6122, 2019.

Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.*, 27(4):576–593, 2012. ISSN 0883-4237. doi: 10.1214/12-STS399. URL https://doi.org/10.1214/12-STS399.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948. PMLR, 2014.

B Supplementary Materials

B.1 Proof of Theorems 2

Remark 1 *In the proofs of this section, it will always be understood that the fact that $\phi^\star$ is the leading eigenvector of the pair of matrices $(\Sigma_X, \Sigma_{XY}\Sigma_{YX})$ is proven in Theorem 1.*

Remark 2 *Note that $\Sigma_{XY}\Sigma_{YX}$ is symmetric, hence by Corollary 8.7.2 in Golub and Van Loan [2013] there exists a non-singular $W \in \mathbb{R}^{p\times p}$ such that*

$$W^\top \Sigma_{XY}\Sigma_{YX} W = diag(a_1, ..., a_p) \quad \text{and} \quad W^\top \mathbb{E}(X_1 X_1^\top) W = diag(b_1, ..., b_p)$$

and such that, for all $i \in \{1, ..., p\}$

$$\Sigma_{XY}\Sigma_{YX} W_i = \lambda_i \Sigma_X W_i \qquad \text{where } \lambda_i = a_i/b_i.$$

But since $\Sigma_{XY}\Sigma_{YX}$ is of rank 1, only one a_i is different from zero. Hence only one λ_i is different from zero. This implies that in Tan et al. [2018] $\gamma = 0$. We note that Lemma 3 in Tan et al. [2018] remains valid by interpreting the divisions by zero implied by $\gamma = 0$ as infinities.

Note that the same reasoning holds (when the associated GEP are definite) for $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ and every sub-matrices one can construct by pre- and post-multiplying $\Sigma_{XY}\Sigma_{YX}$ or $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ by any $P \in \mathcal{P}(\mathcal{F}(p, k)$ for any $k \leq p$.

Remark 3 *In order to prove the results we need to define the following quantity. For any vector $v \in \mathbb{R}^p$, let $\text{supp}(v)$ be the set of index of non-zero elements of v. For any set $F \subset \{1, ..., p\}$ let $\phi(F)$ be defined as*

$$\phi(F) := \underset{\phi\in\mathbb{R}^p}{\text{argmax}} \frac{\phi^\top X^\top Y Y^\top X \phi}{n\phi^\top X^\top X \phi} \qquad s.t.\ \phi^\top \hat{\Sigma}_X \phi = 1,\ \text{supp}(\phi) \subseteq F.$$

B.1.1 Lemmas

The next lemma specialized the result of Lemma 3 in Tan et al. [2018] to the setup of this paper.

Lemma 8 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top \beta$ and $\|\beta\|_0 = s$.*

For $k' \in \mathbb{N}$ s.t. $k' > s$, let $F \in \mathcal{F}(p, k')$. Given any $\tilde{\phi} \in \mathbb{R}^k$ such that $\|\tilde{\phi}\|_2 = 1$ and $\tilde{\phi}^\top P_{(F)}^\top \phi(F) > 0$ a.s, let

$$\rho = \frac{\tilde{\phi}^\top P_{(F)}^\top \hat{\Sigma}_{XY}\hat{\Sigma}_{YX} P_{(F)} \tilde{\phi}}{\tilde{\phi}^\top P_{(F)}^\top \hat{\Sigma}_X P_{(F)} \tilde{\phi}}$$

where $P_{(F)}$ *is defined as in section 4.3. Moreover let* $\phi' = C_F\tilde{\phi}/\|C_F\tilde{\phi}\|_2$ *where*

$$C_F = P_{(F)}^\top P_{(F)} + (\eta/\rho)P_{(F)}^\top(\Sigma_{XY}\Sigma_{YX} - \rho\Sigma_X)P_{(F)}$$

and $\eta > 0$ *is some positive constant. Let* $\delta = 1 - \tilde{\phi}^\top P_{(F)}^\top \phi(F)$. *Pick* $c > 0$ *and* η *(sufficiently small) such that*

$$\eta\lambda_{\max}(\Sigma) < 1/(1+c),$$

Then, conditionally on the set of events where for some $b > 0$

$$\begin{gathered}\frac{\varepsilon(k')}{\Delta(k')} < b,\\ \rho(\hat{\Sigma}_X - \Sigma_X, k') \leq c\lambda_{\min}(\Sigma_X),\\ 1-\delta \geq 1-\xi\end{gathered}$$

hold almost surely with ξ *defined as*

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

where $C_\diamond = (1+c)/(1-c)$, *we have*

$$\phi(F)^\top P_{(F)}\phi' \geq \phi(F)^\top P_{(F)}\tilde{\phi} + \frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)(1 - \phi(F)^\top P_{(F)}\tilde{\phi})\frac{1}{C_\diamond\kappa(\Sigma_X)}$$

almost surely,

Proof: The result is a direct consequence of Lemma 3 in Tan et al. [2018] and of Remark 2. □

The following lemma follows directly from Lemma 4 in Tan et al. [2018].

Lemma 9 [Lemma 4 in Tan et al. [2018]] *Consider a vector* v' *with* $F' = \text{supp}(v')$ *and* $\text{Card}(F) = \bar{k}$. *Let* F *be the indices the element of an other vector* v *that have the largest* k *absolutes values, with* $\text{Card}(F) = k$. *If* $\|v'\|_2 = \|v\|_2 = 1$ *then*

$$\begin{aligned}\left|\text{Truncate}(v,F)^\top v'\right| \geq |v^\top v'|\\ - (\bar{k}/k)^{1/2}\min\{\sqrt{1-(v^\top v')^2}, (1+(\bar{k}/k)^{1/2})(1-(v^\top v')^2)\}\end{aligned}$$

The next lemma specialized the result of Lemma 5 in Tan et al. [2018] to the setup of this paper.

Lemma 10 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top\beta$ and $\|\beta\|_0 = s$. Let $\phi^\star = \beta/\|\beta\|_2$, $V = \mathrm{supp}(\phi^\star)$ and $F \subset \{1, ..., p\}$ such that $V \subset F$ with $\mathrm{Card}(F) = k' > s$.*

Conditionally on the set of events where $\Lambda := \lambda_{\max}/(1+\lambda_{\max}^2) > \varepsilon(k')/\Delta(k')$ a.s., we have that

$$\sqrt{1 - |\phi(F)^\top \phi^\star|} \leq \frac{\sqrt{2}}{\Lambda\big(\Delta(k') - \varepsilon(k')\big)}\varepsilon(k').$$

holds almost surely, where $\Delta(k')$ and $\varepsilon(k')$ are defined as in section 4.3.

Proof: The result follows directly from Lemma 5 in Tan et al. [2018] and from Remark 2. □

The following proposition originates from an application of Corollary 1 in Tan et al. [2018] in order to demonstrate a non-asymptotic bound for Algorithm 2. We provide our own proof since the hypothesis given in Tan et al. [2018] are not sufficient to prove the results the authors gave in their paper.

Note that a non-asymptotic bound on the l_2 error of the output of Algorithm 1 follows directly from this proposition.

Proposition 7 *For all $n \in \mathbb{N}$, let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ be random functions* iid *across rows. For all $n \in \mathbb{N}$ assume $\Sigma_X := \mathbb{E}(X_1 X_1^\top)$ is positive definite and there exists $\beta \in \mathbb{R}^p$ s.t. $\mathbb{E}(Y_1|X_1) = X_1^\top\beta$ and $\|\beta\|_0 = s$. Let $\phi^\star = \beta/\|\beta\|_2$ and $\Lambda := \lambda_1/(1+\lambda_1^2)$.*

Let k in Algorithm 1 (respectively $\underline{k}$ in Algorithm 2) be such that $k = Cs$ for $C > 1$ (respectively $\underline{k} = C's$ for $C' > 1$) and let $k' = 2k + s$ (respectively let $\overline{k} = C''\underline{k}$ for $C'' > 1$ and $k' = \overline{k} + s$).

Let $\bar{\phi}$ be equal to ϕ_t in step 3.f) *of Algorithm 1 for any $t \in \mathbb{N}$ (respectively to $\phi_{t,k}$ in step* 4.b)vi.B. *in Algorithm 2 for any $t \in \mathbb{N}$ and $k \in \{\underline{k}+1, ..., \overline{k}\}$). Let $\phi_{(0)}$ be equal to ϕ_0 in step 2 of Algorithm 1 (respectively to $\phi_{0,k}$ in step* 2 *of Algorithm 2 if $k = \overline{k}$, and else to $\phi_{T,k+1}$, as in step* 4.c.*).*

Choose a constant $c > 0$ and choose η in Algorithm 1 (respectively Algorithm 2) such that

$$\eta\lambda_{\max}(\Sigma_X) < 1/(1+c), \tag{B.1}$$

and

$$\nu = [1 + 2(\frac{s}{k_\diamond})^{1/2} + \frac{s}{k_\diamond})]^{1/2}\left(1 - \frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond\kappa(\Sigma_X)}\right)^{1/2} < 1 \tag{B.2}$$

holds true, where $C_\diamond = (1+c)/(1-c)$ and $k_\diamond$ is equal to k (respectively to $\underline{k}$).

Then, conditionally on the set of events where

$$\frac{\varepsilon(k')}{\Delta(k')} < \frac{\Lambda(1-\nu)\sqrt{\xi}}{4\sqrt{10}+\Lambda(1-\nu)\sqrt{\xi}}, \tag{B.3}$$

$$|\langle \phi^\star, \phi_{(0)} \rangle_2| \leq 1-\nu^2\xi, \tag{B.4}$$

$$\rho(\hat{\Sigma}_X - \Sigma_X, k') \leq c\lambda_{\min}(\Sigma_X) \tag{B.5}$$

hold almost surely (with ξ defined as in Lemma 8), we have that

$$\sqrt{1-|(\phi^\star)^\top \bar{\phi}|} \leq \nu^t \sqrt{1-|(\phi^\star)^\top \phi_{(0)}|} + \frac{1}{1-\nu}\sqrt{10}\frac{\sqrt{2}}{\Lambda\big(\Delta(k')-\varepsilon(k')\big)}\varepsilon(k')$$

almost surely.

Proof: Firstly we demonstrate the proposition for $\phi_{t,k}$ in Algorithm 2.

For any $k \in \mathbb{N}$ let ν_k be defined as

$$\nu_k := [1+2(\frac{s}{k})^{1/2}+\frac{s}{k})]^{1/2}\left(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond \kappa(\Sigma_X)}\right)^{1/2}.$$

For any $k \in \{\underline{k}+1, \overline{k}\}$ let $\phi_{t,k}$ be as in step 4.*b*)*vi.B.* in Algorithm 2. Let $F_{t,k} = \text{supp}(\phi_{t,k})$, $F_{t-1,k} = \text{supp}(\phi_{t-1,k})$ and $V = \text{supp}(\phi^\star)$. Moreover let $F^{(t,k)} = F_{t,k} \cup F_{t-1,k} \cup V$. Note that thanks to step 4.*b*)*v.* in Algorithm 2, the cardinality of $F^{(t,k)}$ is at most $\overline{k}+s$.

Now we operate like in the proof of Theorem 1 in Tan et al. [2018] in order to bound $(1-|\langle \phi^\star, \phi_{t,k}|)^{1/2}$ for any $k \in \{\underline{k}+1, \overline{k}\}$ and $t \in \{1,...,T\}$ (where T, $\underline{k}$ and $\overline{k}$ are as in Algorithm 2).

Let $\phi'_{t,k}$ be the as in step 4.*b*)*iii.* of Algorithm 2. Let $F^{(t,k)}$ be defined as above. Note that

$$P_{(F^{(t,k)})}P^\top_{(F^{(t,k)})}\phi_{t-1,k} = \phi_{t-1,k},$$

and that same equality holds for $\phi(F^{(t,k)})$. Moreover, in the notations of Lemma 8, observe that if $\tilde{\phi} = P^\top_{(F^{(t,k)})}\phi_{t-1,k}$ then $\phi' = P^\top_{(F^{(t,k)})}\phi'_{t,k}$. Hence we let $\tilde{\phi} = P^\top_{(F^{(t,k)})}\phi_{t-1,k}$ and apply Lemma 8. To this aim we need to verify the condition on the initial vector $\tilde{\phi}$. To do so we apply inductive reasoning. Suppose that the condition on the initial vector $\tilde{\phi}$ holds, then by Lemma 8 we have

$$\phi(F^{(t,k)})^\top \phi'_{t,k} \geq$$
$$\phi(F^{(t,k)})^\top \phi_{t-1,k} + \frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)(1-\phi(F^{(t,k)})^\top\phi_{t-1,k})(\frac{1}{C_\diamond\kappa(\Sigma_X)})$$

which implies (by subtracting 1 and rearranging the terms)

$$1-\phi(F^{(t,k)})^\top\phi'_{t,k}\le\big(1-\phi(F^{(t,k)})^\top\phi_{t-1,k}\big)\big(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond\kappa(\Sigma_X)}\big). \tag{B.6}$$

We prove a series of inequalities. Without loss of generality, we assume the inner products between $\phi(F^{(t,k)})$, $\phi^\star$ and $\phi'_{t,k}$ are positive everywhere on the set of events we are working on (because otherwise we can simply treat the events where this inner product is negative by making appropriate sign changes in the proof). Hence inequality (B.6) is equivalent to

$$\|\phi(F^{(t,k)})-\phi'_{t,k}\|_2\le\|\phi(F^{(t,k)})-\phi_{t-1,k}\|_2\big(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond\kappa(\Sigma_X)}\big)^{1/2}. \tag{B.7}$$

Moreover by the triangle inequality we have

$$\begin{aligned}
&\|\phi^\star-\phi'_{t,k}\|_2 &\text{(B.8)}\\
&\le\|\phi(F^{(t,k)})-\phi'_{t,k}\|_2+\|\phi(F^{(t,k)})-\phi^\star\|_2\\
&\le\|\phi(F^{(t,k)})-\phi_{t-1,k}\|_2\big(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond\kappa(\Sigma_X)}\big)^{1/2}+\|\phi(F^{(t,k)})-\phi^\star\|_2\\
&\le\|\phi^\star-\phi_{t-1,k}\|_2\big(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond\kappa(\Sigma_X)}\big)^{1/2}+2\|\phi(F^{(t,k)})-\phi^\star\|_2 &\text{(B.9)}
\end{aligned}$$

where the second inequality comes from inequality (B.7). Inequality (B.8) is equivalent to

$$\sqrt{1-|(\phi^\star)^\top\phi'_{t,k}|}\le\sqrt{1-|(\phi^\star)^\top\phi_{t-1,k}|}\bigg(1-\frac{1+c}{8}\eta\lambda_{\min}(\Sigma_X)\frac{1}{C_\diamond\kappa(\Sigma_X)}\bigg)^{1/2}+2\sqrt{1-|\phi(F^{(t,k)})^\top\phi^\star|} \tag{B.10}$$

Let $\hat\phi_{t,k}$ be as in step 4.b).vi.A. of Algorithm 2. By Lemma 9, we have

$$\begin{aligned}
\sqrt{1-|(\phi^\star)^\top\hat\phi_{t,k}|}&\le\Big(1-|(\phi^\star)^\top\phi'_{t,k}|+\big((s/k)^{1/2}+s/k\big)\big(1-|(\phi^\star)^\top\phi'_{t,k}|^2\big)\Big)^{1/2}\\
&\le\sqrt{1-|(\phi^\star)^\top\phi'_{t,k}|}\Big(1+\big((s/k)^{1/2}+s/k\big)\big(1+|(\phi^\star)^\top\phi'_{t,k}|\big)\Big)^{1/2}\\
&\le\sqrt{1-|(\phi^\star)^\top\phi'_{t,k}|}\Big(1+2\big((s/k)^{1/2}+s/k\big)\Big)^{1/2}\\
&\le\nu_k\sqrt{1-|(\phi^\star)^\top\phi_{t-1,k}|}+\sqrt{20}\sqrt{1-|\phi(F^{(t,k)})^\top\phi^\star|}
\end{aligned} \tag{B.11}$$

where the third inequality holds since $|(\phi^\star)^\top \phi'_{t,k}| \leq 1$ and the fourth inequality holds by inequality (B.10).

Finally, since $\phi_{t,k} = \hat{\phi}_{t,k}/\|\hat{\phi}_{t,k}\|_2$ and that by construction $\|\hat{\phi}_{t,k}\|_2 \leq 1$, we have

$$\begin{aligned}\sqrt{1-|(\phi^\star)^\top \phi_{t,k}|} &\leq \sqrt{1-|(\phi^\star)^\top \hat{\phi}_{t,k}|} \\ &\leq \nu_k \sqrt{1-|(\phi^\star)^\top \phi_{t-1,k}|} + \sqrt{20}\sqrt{1-|\phi(F^{(t,k)})^\top \phi^\star|} \\ &\leq \nu_k \sqrt{1-|(\phi^\star)^\top \phi_{t-1,k}|} + \sqrt{20}\frac{\sqrt{2}}{\Lambda\big(\Delta(k')-\varepsilon(k')\big)}\varepsilon(k')\end{aligned} \tag{B.12}$$

where the second inequality holds by inequality (B.11) and the third inequality holds by Lemma 10 that we can apply since inequality (B.3) implies that $\Lambda > \varepsilon(k')/\Delta(k')$. Starting from now, we differ from Tan et al. [2018] and verify that if $\phi_{t-1,k}$ satisfies the condition of Lemma 8, so does $\phi_{t,k}$. Let

$$\omega(k') := \frac{\sqrt{2}}{\Lambda\big(\Delta(k')-\varepsilon(k')\big)}\varepsilon(k')$$

By the triangle inequality, we have

$$\|\phi(F^{(t,k)}) - \phi_{t,k}\|_2 \leq \|\phi^\star - \phi_{t,k}\|_2 + \|\phi(F^{(t,k)}) - \phi^\star\|_2$$

this implies that

$$\begin{aligned}\sqrt{1-|\phi(F^{(t,k)})^\top \phi_{t,k}|} &\leq \sqrt{1-|(\phi^\star)^\top \phi_{t,k}|} + \sqrt{1-\phi(F^{(t,k)})^\top \phi^\star} \\ &\leq \nu_k\sqrt{\xi} + 2\sqrt{20}\omega(k') \\ &\leq \sqrt{\xi}\end{aligned}$$

where the second inequality holds by inequality (B.12) and Lemma 10, and the third inequality holds by inequality (B.3). This proves that if the assumption of Lemma 8 is satisfied for $t-1$ then it is for t. Since we have assumed that $|\phi_{(0)}^\top \phi^\star| \geq 1-\nu^2\xi$, namely that the initial vector satisfies the assumption of Lemma 8, it implies by induction, that the assumption is satisfied for all $t \in \mathbb{N}$. Hence, by applying inequality (B.12) recursively, we obtain

$$\sqrt{1-|(\phi^\star)^\top \phi_{t,k}|} \leq \nu_k^t \sqrt{1-|(\phi^\star)^\top \phi_{0,k}|} + \frac{1-\nu_k^t}{1-\nu_k}\sqrt{20}\omega(k')$$

and since $\nu \geq \nu_k$ we have

$$\sqrt{1-|(\phi^\star)^\top \phi_{t,k}|} \leq \nu^t \sqrt{1-|(\phi^\star)^\top \phi_{0,k}|} + \frac{1}{1-\nu}\sqrt{20}\omega(k')$$

which was needed to be proven.

The proof for ϕ_t in Algorithm 1 follows by replacing $\phi_{t,k}$ by ϕ_t, $\phi'_{t,k}$ by ϕ'_t, $\hat{\phi}_{t,k}$ by $\hat{\phi}_t$, $\phi_{0,k}$ by ϕ_0 (where ϕ'_t, $\hat{\phi}_t$, ϕ_0 are as in Algorithm 1) and $F_{t,k}$ by $F_t := \mathrm{supp}(\phi_t)$, $F_{t-1,k}$ by $F_{t-1} := (\phi_{t-1,k})$ and $F^{(t,k)}$ by $F^{(t)} := F_t \cup F_{t-1} \cup V$ (and noting that the cardinality of $F^{(t)}$ is at most $2k+s$).

□

B.1.2 Proof of Theorem 2

Let T be as in Algorithm 2 and let

$$\omega(k') := \frac{\sqrt{2}}{\Lambda\big(\Delta(k') - \varepsilon(k')\big)}\varepsilon(k').$$

We aim to apply Proposition 7 in order to bound $\sqrt{1-|(\phi^\star)^\top \phi_{t,k}|}$ for any $t \in \mathbb{N}$ and $k \in \{\underline{k}, \overline{k}\}$.

Suppose that for some $k \in \{\underline{k}+1, \overline{k}-1\}$, in the notation of Proposition 7, the condition on $\phi_{(0)}$ is satisfied (that is a condition on $\phi_{0,k}$). Hence by inequality (B.12) and the reasoning thereafter in the proof of Proposition (7), we know that $\phi_{T,k}$ also satisfies this condition. Hence since by Algorithm 2, $\phi_{0,k-1} = \phi_{T,k}$, Proposition 7 may be applied to bound $\sqrt{1-|(\phi^\star)^\top \phi_{T,k-1}|}$. Hence, by induction, our assumption on $\phi_{0,\overline{k}}$, ensures that we can apply Proposition 7 in order to bound $\sqrt{1-|(\phi^\star)^\top \phi_{t,k}|}$ for any $t \in \mathbb{N}$ and $k \in \{\underline{k}, \overline{k}\}$.

Finally, since $\omega(k_1) \geq \omega(k_2)$ if $k_1 > k_2$ and since $t \geq T$ by recursively applying Proposition 7, we have

$$\sqrt{1-|(\phi^\star)^\top \phi_{t,\underline{k}}|} \leq \nu^{t+[\overline{k}-\underline{k}]T}\sqrt{1-|(\phi^\star)^\top \phi_{0,\overline{k}}|} + \frac{1-\nu^{[\overline{k}-\underline{k}]T}}{1-\nu^T}\frac{1}{1-\nu}\sqrt{20}\omega(k')$$

from which the result is obtained. □

B.2 Proof of Proposition 3

B.2.1 Lemmas

Remark 4 *It is common in the literature to characterize sub-gaussian and sub-exponential random variable through the finiteness of their* Orlicz norm *defined as*

$$\|Z\|_{\Psi_\alpha} := \{C > 0; \mathbb{E}\exp(\frac{Z^\alpha}{C^\alpha}) \leq 2\}.$$

From Definitions 5.7 and 5.13 in Vershynin [2012] one easily verify that Z is sub-gaussian if and only if $\|Z\|_{\Psi_2} < \infty$, and is sub-exponential if and only if $\|Z\|_{\Psi_1} < \infty$.

In the proofs we need the following definition.

Definition 3 (Definition 5.1 in Vershynin [2012]) (Nets, covering numbers)

Let (U, d) be a metric space and let $\epsilon > 0$. A subset $\mathcal{N}_\epsilon$ of U is called a ϵ-net of U if every point $u \in U$ can be approximated to within ϵ by some point $v \in \mathcal{N}_\epsilon$, i.e. so that $d(u, v) \leq \epsilon$. The minimal cardinality of a ϵ-net of U, if finite, is denoted $\mathcal{N}(U, \epsilon)$ and is called the covering number of U (at scale ϵ).

Lemma 11 *Assume that $X_{11}, ..., X_{1p}$ are sub-gaussian random variables and that $p \to \infty$ as $n \to \infty$. Assume that $X = (X_1, ..., X_p) \in \mathbb{R}^{n \times p}$ where each row are independent and identically to $(X_{11}, ..., X_{1p})$. Let $s \in \mathbb{N}$, $k = Cs$ for $C > 0$, $\Sigma_X = \mathbb{E}(X_1 X_1^\top)$ and $\hat{\Sigma}_X = X^\top X/n$.*

Then if $s \log(p/s)/n = o(1)$ and if $s/p = o(1)$ then

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

where $\rho(\cdot, \cdot)$ is defined as in section 4.3..

Proof.

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\|v\|_0 \leq k, \|v\|_2 = 1} |\langle (\hat{\Sigma}_X - \Sigma_X)v, v\rangle_2| > t\Big) \\
&= \mathbb{P}\Big(\sup_{P \in \mathcal{P}(\mathcal{F}(p,k)), \|v\|_2 = 1} |\langle (\hat{\Sigma}_X - \Sigma_X)Pv, Pv\rangle_2| > t\Big) \\
&\leq \sum_{P \in \mathcal{P}(\mathcal{F}(p,k))} \mathbb{P}\Big(\sup_{\|v\|_2 = 1} |\langle (\hat{\Sigma}_X - \Sigma_X)Pv, Pv\rangle_2| > t\Big) \qquad \text{(B.13)} \\
&\leq \sum_{P \in \mathcal{P}(\mathcal{F}(p,k))} \mathbb{P}\Big(2 \sup_{v \in \mathcal{N}_{\frac{1}{4}}} |\langle (\hat{\Sigma}_X - \Sigma_X)Pv, Pv\rangle_2| > t\Big) \qquad \text{(B.14)} \\
&\leq \sum_{P \in \mathcal{P}(\mathcal{F}(p,k))} \sum_{v \in \mathcal{N}_{\frac{1}{4}}} \mathbb{P}\big(2|\langle (\hat{\Sigma}_X - \Sigma_X)Pv, Pv\rangle_2| > t\big) \qquad \text{(B.15)} \\
&= \sum_{P \in \mathcal{P}(\mathcal{F}(p,k))} \sum_{v \in \mathcal{N}_{\frac{1}{4}}} \mathbb{P}\Big(2\Big|\frac{1}{n}\langle XPv, XPv\rangle_2 - \langle \Sigma_X Pv, Pv\rangle_2\Big| > t\Big) \\
&= \sum_{P \in \mathcal{P}(\mathcal{F}(p,k))} \sum_{v \in \mathcal{N}_{\frac{1}{4}}} \mathbb{P}\Big(2\Big|\frac{1}{n}\|XPv\|_2^2 - \langle \Sigma_X Pv, Pv\rangle_2\Big| > t\Big). \qquad \text{(B.16)}
\end{aligned}$$

Where inequalities (B.13) and (B.15) follow by union bounds, and where inequality (B.14) follows by Lemma 5.2 in Vershynin [2012] with scale $\epsilon = 1/4$. Now we treat the terms inside the probability on the right-hand side of inequality (B.16). Remark that for all $P \in \mathcal{P}(\mathcal{F}(p, k))$ and all $v \in \mathcal{N}_{\frac{1}{4}}$ we have $\|XPv\|_2^2 = \sum_{i=1}^n \langle P^\top X_i, v\rangle_2^2$ (by definition of the canonical norm on

$\mathbb{R}^n$), that

$$\begin{aligned}\mathbb{E}\langle P^\top X_i, v\rangle_2^2 &= \mathbb{E}(\langle P^\top X_i, v\rangle_2 \langle P^\top X_i, v\rangle_2) \\ &= \mathbb{E}(\langle P^\top X_i \langle P^\top X_i, v\rangle_2, v\rangle_2) \\ &= \langle \mathbb{E}(P^\top X_i \langle P^\top X_i, v\rangle_2), v\rangle_2 \\ &= \langle P^\top \Sigma_X P v, v\rangle_2\end{aligned}$$

and that $\|\langle P^\top X_i, v\rangle_2\|_{\Psi_2} \leq \sum_{j=1}^k \|X_{i,\alpha_j}\|_{\Psi_2} < \infty$ since we have assumed that the components of X_i are sub-gaussian. Hence we have that

$$\frac{1}{n}\sum_{i=n}^{n}(\langle P^\top X_i, v\rangle_2^2 - \langle P^\top \Sigma_X P v, v\rangle_2)$$

is an iid sum of centered sub-exponential. Then by Corollary 5.17 in Vershynin [2012], for all $P \in \mathcal{P}(\mathcal{F}(p,k))$ and for all $v \in \mathcal{N}_{1/4}$ there exists an absolute constant $c > 0$ and $K_{P,v} = \sup_{l \leq 1} l^{-1}(\mathbb{E}(\langle P^\top X_i, v\rangle_2^2))^{1/l}$ such that

$$\mathbb{P}(|\frac{1}{n}\sum_{i=n}^{n}(\langle P^\top X_i, v\rangle_2^2 - \langle P^\top \Sigma P v, v\rangle_2)| > t/2) \leq 2e^{-\frac{nc}{2}\min\{\frac{t}{K}, \frac{t^2}{2K^2}\}}.$$

Let $K = \min_{P \in \mathcal{P}(\mathcal{F}(p,k)), v \in \mathcal{N}_{1/4}} K_{P,v}$, then by inequality (B.16), we conclude that there exists an absolute constant $c > 0$ such that

$$\begin{aligned}\mathbb{P}(\sup_{\|v\|=1, \|v\|_0 \leq 0} |\langle(\hat{\Sigma}_X - \Sigma_X)v, v\rangle_2| > t) &\leq 9^k C_p^k 2e^{-\frac{nc}{2}\min\{\frac{t}{K}, \frac{t^2}{2K^2}\}} && \text{(B.17)} \\ &\leq (\frac{9ep}{k})^k 2e^{-\frac{nc}{2}\min\{\frac{t}{K}, \frac{t^2}{2K^2}\}} && \text{(B.18)} \\ &= 2e^{k\log(\frac{9ep}{k}) - \frac{nc}{2}\min\{\frac{t}{K}, \frac{t^2}{2K^2}\}}.\end{aligned}$$

Where inequality (B.17) follows by Lemma 5.4 in Vershynin [2012] and the fact that $\text{Card}(\mathcal{P}(\mathcal{F}(p,k))) = C_p^k$, and inequality (B.18) follows since $C_p^k \leq (\frac{ep}{k})^k$.

Now we show the bound in probability at the desired rate. For $C > 0$, we have

$$\begin{aligned}\mathbb{P}(\sup_{\|v\|=1, \|v\|_0 \leq 0} |\langle(\hat{\Sigma}_X - \Sigma_X)v, v\rangle_2| > C\sqrt{\frac{k\log(p/k)}{n}}) \\ \leq 2e^{k\log(\frac{9ep}{k}) - \frac{nc}{2}\min\{\frac{C}{K}\sqrt{\frac{k\log(p/k)}{n}}, \frac{C^2}{2K^2}\frac{k\log(p/k)}{n}\}}\end{aligned}$$

and it is easy to check that for all $C > 0$ if $\frac{4K^2}{C^2} > \frac{k\log(p/k)}{n}$ then

$$\min\{\frac{C}{K}\sqrt{\frac{k\log(p/k)}{n}}, \frac{C^2}{2K^2}\frac{k\log(p/k)}{n}\} = \frac{C^2}{2K^2}\frac{k\log(p/k)}{n}$$

and the bound becomes

$$\mathbb{P}(\sup_{\|v\|=1,\|v\|_0\leq 0}|\langle(\hat{\Sigma}_X-\Sigma_X)v,v\rangle_2|>C\sqrt{\frac{\log(p/k)}{n}})\leq 2e^{k\left(\log(\frac{9ep}{k})-\frac{c}{4}\frac{C^2}{K^2}\log(p/k)\right)}$$
$$\leq 2e^{k\left(\log(9e)+\left(1-\frac{c}{4}\frac{C^2}{K^2}\right)\log(p/k)\right)}$$

And for C such that $1<\frac{c}{4}\frac{C^2}{K^2}$ the right hand-side can be made arbitrarily small by increasing p/k. Hence since $p\to\infty$ as $n\to\infty$, $k/p=o(1)$, $\frac{k\log(p/k)}{n}=o(1)$ and $k=C's$ for some $C'>0$, the thesis is proven. □

Lemma 12 *Assume that $Y_1, X_{11}, ..., X_{1p}$ are sub-gaussian random variables. Assume that $(Y,X)=(Y,X_1,...,X_p)\in\mathbb{R}^{n\times 1+p}$ where each row are independent and identically distributed to $(Y_1,X_{11},...,X_{1p})$. Let $s\in\mathbb{N}$, $k=Cs$ for $C>0$, $\Sigma_{XY}=\mathbb{E}(X_1Y_1)$, $\Sigma_{YX}=\Sigma_{XY}^\top$, $\hat{\Sigma}_{XY}=X^\top Y/n$ and $\hat{\Sigma}_{YX}=\hat{\Sigma}_{XY}^\top$.*

Then if $s\log(p/s)/n=o(1)$ and if $s/p=o(1)$ then

$$\rho(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}-\Sigma_{XY}\Sigma_{YX},k)=O_p(\sqrt{\frac{k\log(p/k)}{n}}).$$

where $\rho(\cdot,\cdot)$ is defined as in section 4.3..

Proof. We have

$$\begin{aligned}
&\sup_{\|v\|_2=1;\|v\|_0=k}|v^\top\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}v-v^\top\Sigma_{XY}\Sigma_{YX}^\top v|\\
&=\sup_{\|v\|_2=1;\|v\|_0=k}|(v^\top\hat{\Sigma}_{XY})^2-(v^\top\Sigma_{XY})^2|\\
&=\sup_{\|v\|_2=1;\|v\|_0=k}|v^\top\hat{\Sigma}_{XY}-v^\top\Sigma_{XY}||v^\top\hat{\Sigma}_{XY}+v^\top\Sigma_{XY}|\\
&=\sup_{\|v\|_2=1;\|v\|_0=k}|v^\top\hat{\Sigma}_{XY}-v^\top\Sigma_{XY}|\sup_{\|v\|_2=1;\|v\|_0=k}|v^\top\hat{\Sigma}_{XY}+v^\top\Sigma_{XY}|
\end{aligned}$$

And it is easy to check (using the same techniques that in the proof of Lemma 11) that if $p\to\infty$ as $n\to\infty$, $k/p=o(1)$, $\frac{k\log(p/k)}{n}=o(1)$ and $k=Cs$ for some $C>0$

$$\sup_{\|v\|_2=1;\|v\|_0=k}|v^\top\hat{\Sigma}_{XY}-v^\top\Sigma_{XY}|=O_p(\sqrt{\frac{s\log(p/s)}{n}})$$

and that

$$\sup_{\|v\|_2=1;\|v\|_0=k}|v^\top\hat{\Sigma}_{XY}+v^\top\Sigma_{XY}|=O_p(1).$$

then the proof follows from the fact that

$$O_p(R_n)O_p(1)=O_p(R_n).$$

□

B.2.2 Proof of Proposition 3

The result follows directly from Lemmas 11 and 12. □

B.3 Proof of Lemma 1

$$\|(I_{p\times p} - \eta\hat{\Sigma}_X)P\|^2_{\text{spec}} = \sup_{\|v\|_2=1} v^\top P^\top (I_{p\times p} - \eta\hat{\Sigma}_X)^2 Pv.$$

Since for all v s.t. $\|v\|_2 = 1$ we have

$$v^\top P^\top (I_{p\times p} - \eta\hat{\Sigma}_X)^2 Pv = 1 - \eta\bigg(2v^\top P^\top \hat{\Sigma}_X Pv - \eta v^\top P^\top \hat{\Sigma}_X^2 Pv\bigg),$$

since $P^\top\hat{\Sigma}_X P$ is positive definite by assumption, there exists $\eta_v^\alpha > 0$ small enough such that

$$2v^\top P^\top \hat{\Sigma}_X Pv - \eta^\alpha v^\top P^\top \hat{\Sigma}_X^2 Pv > 0$$

and hence there exists $\eta_v^\beta > 0$ small enough such that

$$0 \le 1 - \eta_v^\beta\bigg(2v^\top P^\top \hat{\Sigma}_X Pv - \eta_v^\alpha v^\top P^\top \hat{\Sigma}_X^2 Pv\bigg) < 1$$

the thesis follows from choosing $\eta = \inf_{\|v\|_2=1}\eta_v$ where $\eta_v = \inf\{\eta_v^\alpha, \eta_v^\beta\}$. □

B.4 Proof of Lemma 3

For some $M > 0$, consider

$$\mathbb{P}\bigg(\sup_{\tilde{t}\in\{1,\dots,t(n)\}} \bigg|\frac{\phi_{\tilde{t}}^\top\hat{\Sigma}_X\phi_{\tilde{t}}}{\phi_{\tilde{t}}^\top\hat{\Sigma}_X\phi^\star + \phi_{\tilde{t}}^\top\frac{X^\top U}{\|\beta\|_2 n}} - 1\bigg| > M\bigg)$$

which is bounded above by

$$\mathbb{P}\bigg(\frac{\sup_{\tilde{t}\in\{1,\dots,t(n)\}}|\phi_{\tilde{t}}^\top\hat{\Sigma}_X(\phi_{\tilde{t}} - \phi^\star) - \phi_{\tilde{t}}^\top\frac{X^\top U}{\|\beta\|_2 n}|}{\inf_{\tilde{t}\in\{1,\dots,t(n)\}}|\phi_{\tilde{t}}^\top\hat{\Sigma}_{XY}|} > \frac{M}{\|\beta\|_2}\bigg)$$

let $M' = M/\|\beta\|_2$, then this probability is bounded above by

$$\mathbb{P}\bigg(\sup_{\tilde{t}\in\{1,\dots,t(n)\}}|\phi_{\tilde{t}}^\top\hat{\Sigma}_X(\phi_{\tilde{t}} - \phi^\star)| + \|\frac{X^\top U}{\|\beta\|_2 n}\|_2 - M'\inf_{\tilde{t}\in\{1,\dots,t(n)\}}|\phi_{\tilde{t}}^\top\hat{\Sigma}_{XY}| > 0\bigg) \tag{B.19}$$

and for all $\tilde{t}$ we have

$$\begin{aligned}|\phi_{\tilde{t}}^\top \hat{\Sigma}_{XY}| &= |\phi_{\tilde{t}}^\top \Sigma_{XY} - \phi_{\tilde{t}}^\top (\Sigma_{XY} - \hat{\Sigma}_{XY})| \\ &\geq |\phi_{\tilde{t}}^\top \Sigma_{XY}| - |\phi_{\tilde{t}}^\top (\Sigma_{XY} - \hat{\Sigma}_{XY})| \\ &\geq \inf_{\|v\|_2=1;\|v\|_0=k} |v^\top \Sigma_{XY}| - \sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} - v^\top \Sigma_{XY}|\end{aligned}$$

hence the probability in (B.19) is bounded above by

$$\begin{aligned}&\mathbb{P}\Bigg(\sup_{\tilde{t}\in\{1,\dots,t(n)\}} |\phi_{\tilde{t}}^\top \hat{\Sigma}_X (\phi_{\tilde{t}} - \phi^\star)| + \|\frac{X^\top U}{\|\beta\|_2 n}\|_2 \\ &+ M' \sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} - v^\top \Sigma_{XY}| - M' \inf_{\|v\|_2;\|v\|_0=k} |v^\top \Sigma_{XY}| > 0 \Bigg).\end{aligned}$$

And for M and n high enough this probability can be bounded by any $\epsilon > 0$ since

$$\begin{aligned}\sup_{\tilde{t}\in\{1,\dots,t(n)\}} |\phi_{\tilde{t}}^\top \hat{\Sigma}_X (\phi_{\tilde{t}} - \phi^\star)| &= O_p(1), \\ \|\frac{X^\top U}{\|\beta\|_2 n}\|_2 &= o_p(1), \\ \sup_{\|v\|_2=1;\|v\|_0=k} |v^\top \hat{\Sigma}_{XY} - v^\top \Sigma_{XY}| &= o_p(1).\end{aligned}$$

Which is what we needed to prove. □

B.5 Proof of Lemma 5

Let x be any output of Algorithm 1 or 2 2, then following Definition 2 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 $Q(x)$.

Firstly, we show 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.

Suppose the opposite. This implies there exists $i \in \{1, ..., p\}$ s.t. $x_i = 0$ and $\phi_i^\star \neq 0$. However $|\phi_i^\star| < \|x - \phi^\star\|_2 < \epsilon_1$. Hence by choosing ϵ_1 to be the lowest non-null element of $|\phi^\star|_2$ we have reached a contradiction.

Secondly, our assumption on the convergence of ϕ_t implies that for all $h > 0$ $\|\phi_{t+h} - \phi^\star\| < \epsilon_2$ and the same reasoning applies. Hence the thesis is proven. □

B.6 Proof of Lemma 6

By adding zero and the triangle inequality we have

$$\|\phi_t - \tilde{\phi}_t\|_2 \leq \|\phi_t - \phi^\star\|_2 + \|\phi^\star - \phi_{t-1}\|_2 + \|\phi_{t-1} - \tilde{\phi}_t\|_2.$$

By assumptions on the convergence of ϕ_t we have $\|\phi_t - \phi^\star\|_2 = o_p(1)$ and $\|\phi_{t-1} - \phi^\star\|_2 = o_p(1)$. It remains to show that $\|\tilde{\phi}_t - \phi_{t-1}\|_2 = o_p(1)$.

By definition of $\tilde{\phi}_t$ and Equation (A.5) in the proof of Proposition 4 gives us

$$\begin{aligned}\tilde{\phi}_t - \phi_{t-1} &= \eta\frac{1}{g_{t-1}}\hat{\Sigma}_X\phi^\star - \eta\hat{\Sigma}_X\phi_{t-1} + \eta\frac{1}{g_{t-1}}\frac{X^\top U}{\|\beta\|_{\mathbb{R}^p} n} \\ &= \eta\hat{\Sigma}_X([\frac{1}{g_{t-1}} - 1]\phi^\star + \phi^\star - \phi_{t-1}) + \eta\frac{1}{g_{t-1}}\frac{X^\top U}{\|\beta\|_{\mathbb{R}^p} n}.\end{aligned}$$

Hence since we have $\phi_{t-1} \xrightarrow{p} \phi^\star$, that by Lemma 2 we have $\frac{1}{g_{t-1}} \xrightarrow{p} 1$ and that by our assumption we have $\|\frac{X^\top U}{\|\beta\|_2 n}\|_2 = o_p(1)$ then the boundedness in probability of $\hat{\Sigma}_X$ ensures the thesis follows. □.

B.7 Proof of Lemma 7

By the triangle inequality we have

$$\sup_{t\in\{1,\dots,t(n)\}} \|\phi_t - \tilde{\phi}_t\|_2 \leq 1 + \sup_{t\in\{1,\dots,t(n)\}} \|\tilde{\phi}_t\|_2.$$

By Equation (A.5) in the proof of Proposition 4 we have

$$\begin{aligned}\sup_{t\in\{1,\dots,t(n)\}} \|\tilde{\phi}_t\|_2 &\leq 1 + \sup_{t\in\{1,\dots,t(n)\}} \frac{\eta}{\rho_{t-1}}\|(\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} - \rho_{t-1}\hat{\Sigma}_X)\phi_{t-1}\|_2 \\ &\leq 1 + \sup_{t\in\{1,\dots,t(n)\}} \frac{\eta}{\rho_{t-1}}\|\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\|_{\text{spec}} + \eta\|\hat{\Sigma}_X\|_{\text{spec}}.\end{aligned}$$

Since $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ and $\hat{\Sigma}_X$ are bounded in probability since $\mathbb{E}\|X_1\|_2^2 < \infty$ by assumption. It remains to note that

$$\begin{aligned}\sup_{t\in\{1,\dots,t(n)\}} \frac{1}{\rho_{t-1}} &\leq \sup_{t\in\{1,\dots,t(n)\}} \frac{\|\hat{\Sigma}_X\|_{\text{spec}}}{\phi_{t-1}^\top\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\phi_{t-1}} \\ &\leq \frac{\|\hat{\Sigma}_X\|_{\text{spec}}}{\inf_{t\in\{1,\dots,t(n)\}}|\phi_{t-1}^\top\hat{\Sigma}_{XY}|^2}\end{aligned}$$

which can be shown to be $O_p(1)$ by a similar strategy than in Lemma 3 since

$$\begin{aligned}\inf_{t\in\{1,\dots,t(n)\}} |\phi_{t-1}^\top\hat{\Sigma}_{XY}|^2 \geq &\inf_{\|v\|_2;\|v\|_0=k} |v\Sigma_{XY}| \\ &- 2\sup_{\|v\|_2;\|v\|_0=k} |v\Sigma_{XY}| \sup_{\|v\|_2;\|v\|_0=k} |v(\Sigma_{XY} - \hat{\Sigma}_{XY})|\end{aligned}$$

and $\sup_{\|v\|_2;\|v\|_0=k} |v(\Sigma_{XY} - \hat{\Sigma}_{XY})| = o_p(1)$. Hence the thesis is proven. □