Received: 23 December 2022 | Revised: 28 January 2023 | Accepted: 5 February 2023

DOI: 10.1002/sta4.548

ORIGINAL ARTICLE

WILEY

High-dimensional regression coefficient estimation by nuclear norm plus l_1 norm penalization

Matteo Farnè | Angela Montanari

Department of Statistical Sciences, University of Bologna, Bologna, Italy

Correspondence
Matteo Farnè, Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41, 40126, Bologna, Italy.
Email: matteo.farne@unibo.it

Funding information
RFO 2022, Ministry of University and Research, Italian Republic.

We propose a new estimator of the regression coefficients for a high-dimensional linear regression model, which is derived by replacing the sample predictor covariance matrix in the ordinary least square (OLS) estimator with a different predictor covariance matrix estimate obtained by a nuclear norm plus l_1 norm penalization. We call the estimator ALgebraic Covariance Estimator-regression (ALCE-reg). We make a direct theoretical comparison of the expected mean square error of ALCE-reg with OLS and RIDGE. We show in a simulation study that ALCE-reg is particularly effective when both the dimension and the sample size are large, due to its ability to find a good compromise between the large bias of shrinkage estimators (like RIDGE and least absolute shrinkage and selection operator [LASSO]) and the large variance of estimators conditioned by the sample predictor covariance matrix (like OLS and principal orthogonal complement thresholding [POET]).

KEYWORDS
high dimension, nuclear norm, precision matrix, regression coefficient, sparsity

1 | INTRODUCTION

Estimating the regression coefficients of a high-dimensional linear regression model is a relevant statistical challenge. Let us consider a mean-centered $n \times p$ predictor matrix $\mathbf{X}$ and a mean-centered $n \times 1$ response vector $\mathbf{y}$. The ordinary least square (OLS) estimator

$$\hat{\beta}_{OLS} = \arg\min_{\beta \in \mathbb{R}^p} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \hat{\Sigma}_X^{-1}\hat{\sigma}_{XY}, \tag{1}$$

where $\hat{\Sigma}_X = \frac{\mathbf{X}'\mathbf{X}}{n}$ is the sample covariance matrix of the predictors and $\hat{\sigma}_{XY}$ is the vector of sample covariances between the response variable y and the predictors, is not even computable when $p \geq n$ and numerically very unstable when p is large compared with n, even when $n > p$. Therefore, some strategies to regularize $\hat{\beta}_{OLS}$ have been developed, as, when p is large, $\hat{\beta}_{OLS}$ may present anomalously large absolute values and implausible signs (see Hoerl, 2020 in the Special Issue appeared on Technometrics, ; Joseph, 2020 to celebrate the 50 years of ridge regression). The two best known strategies lead to RIDGE estimator (Hoerl & Kennard, 1970), which is derived as

Abbreviations: ALCE-reg, ALgebraic Covariance Estimator-regression; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least squares; POET, principal orthogonal complement thresholding..

Matteo Farnè and Angela Montanari are equally contributing authors.

Stat. 2023;12:e548.
https://doi.org/10.1002/sta4.548

$$\hat{\beta}_{RIDGE} = \arg\min_{\beta \in \mathbb{R}^p} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + k\|\beta\|_2^2, \quad (2)$$

where $\|\beta\|_2^2 = \beta'\beta = \sum_{j=1}^p \beta_j^2$ and LASSO estimator (Tibshirani, 1996), which is derived as

$$\hat{\beta}_{LASSO} = \arg\min_{\beta \in \mathbb{R}^p} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + k\|\beta\|_1, \quad (3)$$

where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. As clearly highlighted by Hastie (2020), RIDGE and LASSO estimators can be rephrased as constrained optimization problems, where the constraint is $\|\beta\|_2 < C_k$ in the case of RIDGE regression and $\|\beta\|_1 < C_k$ in the case of LASSO regression, for some $C_k > 0$.

In Hastie (2020), the link between ridge regression and the spectral decomposition of the matrix $\mathbf{X}'\mathbf{X}$ is elegantly pointed out, while Le et al. (2020) describe the relationship between ridge regression and covariance matrix regularization. These results show that, when $p \geq n, \hat{\beta}_{RIDGE}$ may be extremely biassed, as also reported in Zou (2020). Although $\hat{\beta}_{LASSO}$ tends to be slightly less biassed and a bit more variable, it is also subject to several drawbacks in high dimensions, particularly when the coefficient vector β is not element-wise sparse. It follows that, when p is large, $\hat{\beta}_{OLS}$ is not feasible or extremely variable, while RIDGE and LASSO are very biassed.

In this paper, we explore the possibility to replace the sample covariance matrix of the predictors $\hat{\Sigma}_X$ in the OLS estimator $\hat{\beta}_{OLS}$ by a regularized covariance matrix estimate, obtained by solving the specific regularization problem described in Farnè and Montanari (2020). Therein, a high-dimensional covariance matrix estimator is proposed, under the assumption that the true covariance matrix of the predictors Σ_X follows a low rank plus sparse structure. This assumption is very natural as it results from an approximate factor model (Chamberlain & Rothschild, 1982) imposed to the vector $\mathbf{x}$. Principal orthogonal complement thresholding (POET) estimator (Fan et al., 2013) also assumes a low rank plus sparse structure for Σ_X. That algebraic structure has been analysed and retrieved in exact form in Chandrasekaran et al. (2011) and in approximate form in Chandrasekaran et al. (2010). Following those proposals, in Farnè and Montanari (2020), Σ_X is recast as the solution of a least squares problem penalized by the nuclear norm of the low rank component (see Fazel et al. 2001) and the l_1 norm of the sparse component of Σ_X. The statistical properties of such estimator, called ALCE (ALgebraic Covariance Estimator), have been studied in Farnè and Montanari (2020).

Given these premises, it sounds appropriate to replace the matrix $\hat{\Sigma}_X$ by ALCE estimator in $\hat{\beta}_{OLS}$ and to explore the statistical properties of the resulting estimator of β. Our expectation is that the ALCE estimator of β is able to attain a convenient balance between bias and variance when p is large, thus providing a valid alternative when OLS is too unstable and RIDGE/LASSO are too biassed.

The rest of the paper is structured as follows. Section 2 explores the theoretical framework behind our proposed high-dimensional regression coefficient estimator. Section 3 describes in more detail the statistical properties of our proposed estimator. Section 4 contains a wide simulation study, where a full p-dimensional regression coefficient vector is recovered, under different dimensions and sample sizes, by means of several methods, which are thoroughly compared. Finally, Section 5 provides some concluding remarks.

1.1 | Notation

Given a $p \times p$ symmetric positive semidefinite matrix $\mathbf{M}$, we denote by $\lambda_i(\mathbf{M}), i \in \{1,...,p\}$, the eigenvalues of $\mathbf{M}$ in decreasing order. We recall the following norm definitions:

1. Element-wise:
   a. L_0 norm: $\|\mathbf{M}\|_0 = \sum_{i=1}^p \sum_{j=1}^p \mathbb{1}(\mathbf{M}_{ij} \neq 0)$, which is the total number of nonzeros;
   b. L_1 norm: $\|\mathbf{M}\|_1 = \sum_{i=1}^p \sum_{j=1}^p |\mathbf{M}_{ij}|$;
   c. Frobenius norm: $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^p \mathbf{M}_{ij}^2}$;
   d. Maximum norm: $\|\mathbf{M}\|_\infty = \max_{i \leq p, j \leq p} |\mathbf{M}_{ij}|$.
2. Induced by vector:
   a. $\|\mathbf{M}\|_{0,v} = \max_{i \leq p} \sum_{j \leq p} \mathbb{1}(\mathbf{M}_{ij} \neq 0)$, which is the maximum number of nonzeros per row–column;
   b. Spectral norm: $\|\mathbf{M}\|_2 = \|\mathbf{M}\| = \lambda_1(\mathbf{M})$.
3. Schatten:
   a. Nuclear norm of $\mathbf{M}$, here defined as the sum of the eigenvalues of $\mathbf{M}$: $\|\mathbf{M}\|_* = \sum_{i=1}^p \lambda_i(\mathbf{M})$.

We denote the rank of $\mathbf{M}$ as $\mathrm{rk}(\mathbf{M})$ and the sparsity pattern of $\mathbf{M}$ as $\mathrm{sgn}(\mathbf{M})$, where $\mathrm{sgn}(\mathbf{M})$ is a $p \times p$ matrix whose ij entry is 1 if $\mathbf{M}_{ij} > 0$, 0 if $\mathbf{M}_{ij} = 0$, -1 if $\mathbf{M}_{ij} < 0$. We indicate with $\mathrm{diag}(\mathbf{M})$ a diagonal $p \times p$ matrix containing only the diagonal of $\mathbf{M}$, and we define the matrix $\mathrm{off\_diag}(\mathbf{M}) = \mathbf{M} - \mathrm{diag}(\mathbf{M})$.

2 | THEORETICAL FRAMEWORK

The aim of this paper is to compare the performance of different estimators of the vector of linear regression coefficients in high dimensions and to test the validity of a new proposal. The covariance structure of the vector of predictors $\mathbf{x}$ is crucial when deciding how to replace the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ in $\hat{\beta}_{OLS}$ with a feasible alternative when $p \geq n$. From Hoerl and Kennard (1970), we know that $\hat{\beta}_{RIDGE} = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}$, where the shrinkage term $k\mathbf{I}_p$ has the effect of reconditioning the eigenvalues of $\mathbf{X}'\mathbf{X}$ in a way that avoids singularity and guarantees invertibility, although at the price of a large bias. Instead, the LASSO acts as a variable selector and is thus oriented to identify a restricted set of predictors from the p input ones, even when in the true model β is not element-wise sparse.

When p is large, it is very likely to have a redundant set of predictors, that is, to have predictor multicollinearity, which inevitably affects the conditioning properties of the sample covariance matrix of $\mathbf{x}$. As a consequence, it is not unreasonable to postulate for the vector of predictors $\mathbf{x}$ an approximate factor model of the following kind:

$$\mathbf{x} = \chi + \epsilon, \tag{4}$$

where χ is the common component of $\mathbf{x}$, that is, $\chi = \mathbf{Bf}$, with $\mathbf{B}$ $p \times r$ matrix of factor loadings *s.t.* $\mathbf{B}'\mathbf{B} = \mathbf{I}_r$, and $\mathbf{f}$ $r \times 1$ random vector of common factors *s.t.* $\mathbb{E}(\mathbf{f}) = 0$ and $\mathbb{E}(\mathbf{f}) = \mathbf{I}_r$, while ϵ is the vector of the so called unique factors of $\mathbf{x}$, that is, a $p \times 1$ random vector *s.t.* $\mathbb{E}(\epsilon) = 0$ and $\mathbb{V}(\epsilon) = \mathbf{S}^*$, with $\mathbf{S}^*$ $p \times p$ sparse covariance matrix. From these assumptions,

$$\mathbb{V}(\mathbf{x}) := \Sigma_X = \mathbf{B}\mathbb{V}(\mathbf{f})\mathbf{B}' + \mathbb{V}(\epsilon) = \mathbf{L}^* + \mathbf{S}^*, \tag{5}$$

where $\mathbf{L}^* = \mathbf{BB}'$. In other words, (5) states that the covariance matrix of $\mathbf{x}$, Σ_X, admits a low rank plus sparse decomposition.

Let us analyse the OLS estimator $\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \hat{\Sigma}_X^{-1}\hat{\sigma}_{XY}$, where $\hat{\Sigma}_X = \frac{\mathbf{X}'\mathbf{X}}{n}$ is the sample covariance matrix of $\mathbf{x}$ and $\hat{\sigma}_{XY}$ is the vector of sample covariances between the response variable y and the predictors $x_1, x_2, \ldots, x_p$. In order to obtain a computable estimator when $p \geq n$, RIDGE regression replaces $\mathbf{X}'\mathbf{X}$ by the matrix $\mathbf{X}'\mathbf{X} + k\mathbf{I}_p, k > 0$, in $\hat{\beta}_{OLS}$. This plug-in process has the effect of reconditioning the eigenvalues of $\mathbf{X}'\mathbf{X}$, thus producing a computable and very stable estimator of β, at the price of introducing a systematic bias in the estimate, also due to the inversion of the matrix $\mathbf{X}'\mathbf{X} + k\mathbf{I}_p$. For this reason, the need arises to study an alternative estimator of Σ_X able to limit this inevitable estimation bias, while reconditioning $\hat{\Sigma}_X$, which is not positive definite when $p \geq n$.

For this purpose, we propose to exploit the low rank plus sparse structure of Σ_X displayed in (5). In particular, since we have assumed the covariance matrix of $\mathbf{x}$ to be low rank plus sparse, we can approach the estimation of Σ_X by solving the following problem:

$$\min_{\mathbf{L}, \mathbf{S} \in \mathbb{R}^{p \times p}} \|\hat{\Sigma}_X - (\mathbf{L} + \mathbf{S})\|_F + \psi \mathrm{rk}(\mathbf{L}) + \rho \|\mathbf{S}\|_0, \tag{6}$$

where ψ and ρ are threshold parameters. Unfortunately, this approach is not feasible, because the composite penalty $\psi \mathrm{rk}(\mathbf{L}) + \rho \|\mathbf{S}\|_0$ is nonconvex, so that problem (6) is NP-hard. A possible way to overcome this drawback is by solving the following problem

$$\min_{\mathbf{L}, \mathbf{S} \in \mathbb{R}^{p \times p}} \|\hat{\Sigma}_X - (\mathbf{L} + \mathbf{S})\|_F + \psi \|\mathbf{L}\|_* + \rho \|\mathbf{S}\|_1, \tag{7}$$

since it has been proved that $\|\mathbf{L}\|_*$ is the tightest convex relaxation of $\mathrm{rk}(\mathbf{L})$ and $\|\mathbf{S}\|_1$ is the tightest convex relaxation of $\|\mathbf{S}\|_0$ (see Fazel, 2002). Problem (7) is thus nonsmooth but convex, which means it is solvable in polynomial time.

The pair of estimators $(\hat{\mathbf{L}}, \hat{\mathbf{S}}) = \arg\min_{\mathbf{L}, \mathbf{S} \in \mathbb{R}^{p \times p}} \mathcal{L}(\mathbf{L}, \mathbf{S}) + \mathcal{P}(\mathbf{L}, \mathbf{S})$, where $\mathcal{L}(\mathbf{L}, \mathbf{S}) = \|\hat{\Sigma}_X - (\mathbf{L} + \mathbf{S})\|_F$ and $\mathcal{P}(\mathbf{L}, \mathbf{S}) = \psi \|\mathbf{L}\|_* + \rho \|\mathbf{S}\|_1$ are called ALCE (ALgebraic Covariance Estimator, Farnè & Montanari, 2020). We denote the pair of ALCE estimators as $(\hat{\mathbf{L}}_A, \hat{\mathbf{S}}_A)$, and the overall ALCE covariance estimator as $\hat{\Sigma}_A = \hat{\mathbf{L}}_A + \hat{\mathbf{S}}_A$. $(\hat{\mathbf{L}}_A, \hat{\mathbf{S}}_A)$ has been proved to be both algebraically and parametrically consistent, in the following sense.

Definition 1. A pair of symmetric matrices $(\mathbf{L}, \mathbf{S})$ with $\mathbf{L}, \mathbf{S} \in \mathbb{R}^{p \times p}$ is an algebraically consistent estimate of the low rank plus sparse decomposition (5) for the covariance matrix Σ_X if the following conditions hold:

(i) the low rank estimate $\mathbf{L}$ is positive semidefinite with rank $\mathrm{rk}(\mathbf{L}) = \mathrm{rk}(\mathbf{L}^*) = r$;
(ii) the residual estimate $\mathbf{S}$ is positive definite with the true sparsity pattern $\mathrm{sgn}(\mathbf{S}) = \mathrm{sgn}(\mathbf{S}^*)$;
(iii) $\Sigma = \mathbf{L} + \mathbf{S}$ is positive definite.

Parametric consistency holds if the pair $(\mathbf{L}, \mathbf{S})$ is close to $(\mathbf{L}^*, \mathbf{S}^*)$ in some norm with probability approaching 1.

Definition 2. A pair of symmetric matrices $(\mathbf{L},\mathbf{S})$ with $\mathbf{L},\mathbf{S}\in\mathbb{R}^{p\times p}$ is a parametrically consistent estimate of the low rank plus sparse decomposition (5) for the covariance matrix Σ_X if the norm $g_\gamma=\max\left(\frac{\|\mathbf{L}-\mathbf{L}^*\|_2}{\|\mathbf{L}^*\|_2},\frac{\|\mathbf{S}-\mathbf{S}^*\|_\infty}{\gamma\|\mathbf{S}^*\|_{0,v}}\right)$, with $\gamma\in\mathbb{R}^+$, converges to 0 with probability approaching 1.

Parametric consistency is a usual property in statistical analysis, while algebraic consistency is a typical feature of this approach. The word 'ALgebraic' in the ALCE acronym follows from the need to control the degree of transversality of the following algebraic manifolds:

$$\mathcal{L}(r)=\{\mathbf{L}|\mathbf{L}\succeq 0,\mathbf{L}=\mathbf{UDU}',\mathbf{U}\in\mathbb{R}^{p\times r},\mathbf{U}'\mathbf{U}=\mathbf{I}_r,\mathbf{D}\in\mathbb{R}^{r\times r}\text{diagonal}\}, \tag{8}$$

$$\mathcal{S}(s)=\{\mathbf{S}\in\mathbb{R}^{p\times p}|\mathbf{S}\succ 0,|\text{supp}(\mathbf{S})|\le s\}, \tag{9}$$

where $\mathcal{L}(r)$ is the variety of matrices with at most rank r and $\mathcal{S}(s)$ is the variety of (element-wise) sparse matrices with at most s nonzero elements. The two varieties $\mathcal{L}(r)$ and $\mathcal{S}(s)$ can be disentangled if $\mathbf{L}^*\in\mathcal{L}(r)$ is far from being sparse and $\mathbf{S}^*\in\mathcal{S}(s)$ is far from being low rank. It follows the need to impose them to be close to orthogonality, which is enforced by bounding the following rank-sparsity measures:

$$\xi(\mathcal{T}(\mathbf{L}^*))=\max_{\mathbf{L}\in T(\mathbf{L}^*),\|\mathbf{L}\|_2\le 1}\|\mathbf{L}\|_\infty, \tag{10}$$

$$\mu(\Omega(\mathbf{S}^*))=\max_{\mathbf{S}\in\Omega(\mathbf{S}^*),\|\mathbf{S}\|_\infty\le 1}\|\mathbf{S}\|_2, \tag{11}$$

where $T(\mathbf{L}^*)$ and $\Omega(\mathbf{S}^*)$ are the tangent spaces to $\mathcal{L}(r)$ and $\mathcal{S}(s)$, respectively. Further, the algebraic and parametric consistency of $(\hat{\mathbf{L}}_A,\hat{\mathbf{S}}_A)$ requires to control the magnitude of the eigenvalues of $\mathbf{L}^*$, the sparsity pattern of $\mathbf{S}^*$, the smallest eigenvalue of $\mathbf{L}^*$ and the minimum absolute nonzero element in $\mathbf{S}^*$ with respect to $\xi(\mathcal{T}(\mathbf{L}^*))$ and $\mu(\Omega(\mathbf{S}^*))$. The latent random processes $\mathbf{f}$ and ϵ are imposed to be independent and identically distributed, with sub-Gaussian tails. We stress that the r eigenvalues of $\mathbf{L}^*$ are imposed to scale to $\gamma_\alpha p^\alpha$, with $\gamma_\alpha>0$ and $\alpha\in\left(\frac{1}{2},1\right]$, which corresponds to allowing for weak factors in (4) and that the sparsity pattern of $\mathbf{S}^*$ is controlled by imposing $\|\mathbf{S}^*\|_{0,v}\le\gamma_\delta p^\delta$, with $\gamma_\delta>0$ and $\delta\in\left[0,\frac{1}{2}\right]$, which corresponds to limit the cumulation of residual covariances in a specific row. We refer to Farnè and Montanari (2020) for more technical details.

In this paper, we focus on the ALCE estimator of the regression coefficient (ALCE-reg), defined as $\hat{\beta}_{ALCE}=\hat{\Sigma}_A^{-1}\hat{\sigma}_{XY}$. Following Farnè and Montanari (2020), we also perform the unshrinkage of estimated latent eigenvalues, as this operation improves the sample total loss as much as possible in the finite sample. Once we set $\hat{r}_A=\text{rk}(\hat{\mathbf{L}}_A)$ and we define the spectral decomposition of $\hat{\mathbf{L}}_A$ as $\hat{\mathbf{L}}_A=\hat{\mathbf{U}}_A\hat{\mathbf{D}}_A\hat{\mathbf{U}}_A'$, with $\hat{\mathbf{U}}_A$ $p\times\hat{r}_A$ matrix such that $\hat{\mathbf{U}}_A'\hat{\mathbf{U}}_A=\mathbf{I}_{\hat{r}_A}$ and $\hat{\mathbf{D}}_A$ $\hat{r}_A\times\hat{r}_A$ diagonal matrix, we can get the UNALCE (UNshrunk ALCE) estimates as follows:

$$\hat{\mathbf{L}}_U=\hat{\mathbf{U}}_A(\hat{\mathbf{D}}_A+\psi\mathbf{I}_r)\hat{\mathbf{U}}_A', \tag{12}$$

$$\text{diag}(\hat{\mathbf{S}}_U)=\text{diag}(\hat{\Sigma}_A)-\text{diag}(\hat{\mathbf{L}}_U), \tag{13}$$

$$\text{off\_diag}(\hat{\mathbf{S}}_U)=\text{off\_diag}(\hat{\mathbf{S}}_A), \tag{14}$$

where $\psi>0$ is any chosen eigenvalue threshold parameter. Importantly, it can be proved (Farnè & Montanari, 2020) that it holds

$$\left(\hat{\mathbf{L}}_U,\hat{\mathbf{S}}_U\right)=\arg\min_{\mathbf{L}\in\hat{\mathcal{L}}(\hat{r}_A),\mathbf{S}\in\hat{\mathcal{S}}_{diag}}\frac{1}{2}\|\hat{\Sigma}_X-(\mathbf{L}+\mathbf{S})\|_2, \tag{15}$$

where

$$\hat{\mathcal{L}}(\hat{r}_A)=\{\mathbf{L}|\mathbf{L}\succeq 0,\mathbf{L}=\hat{\mathbf{U}}_A\mathbf{D}\hat{\mathbf{U}}_A',\mathbf{D}\in\mathbb{R}^{r\times r}\text{diagonal}\}, \tag{16}$$

$$\begin{aligned}\hat{\mathcal{S}}_{diag}=\{&\mathbf{S}\in\mathbb{R}^{p\times p}|\text{diag}(\mathbf{L})+\text{diag}(\mathbf{S})=\text{diag}(\hat{\Sigma}_A),\\&\text{off\_diag}(\mathbf{S})=\text{off\_diag}(\hat{\mathbf{S}}_A),\mathbf{L}\in\hat{\mathcal{L}}(\hat{r}_A)\}.\end{aligned} \tag{17}$$

For each threshold pair (ψ,ρ), we can finally compute the overall UNALCE estimate $\hat{\Sigma}_U = \hat{\mathbf{L}}_U + \hat{\mathbf{S}}_U$ and derive the UNALCE estimator of the regression coefficient (UNALCE-reg) as $\hat{\beta}_{UNALCE} = \hat{\Sigma}_U^{-1} \hat{\sigma}_{XY}$.

3 | ESTIMATION FRAMEWORK

We set the standard linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{18}$$

where ε, the residual vector, is assumed to be distributed as $MVN(\mathbf{0},\sigma^2\mathbf{I}_n)$ and uncorrelated with the $p \times 1$ vector of predictors $\mathbf{x}$. First, we consider the OLS coefficient estimator $\hat{\beta}_{OLS} = \hat{\Sigma}_X^{-1}\hat{\sigma}_{XY}$. We know that $\mathrm{VAR}(\hat{\beta}_{OLS}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. We write the sum of squared errors $L^2_{OLS} = (\hat{\beta}_{OLS}-\beta)'(\hat{\beta}_{OLS}-\beta)$. We know from Hoerl and Kennard (1970) that $\mathrm{E}(L^2_{OLS}) = \sigma^2 \mathrm{tr}\left[(\mathbf{X}'\mathbf{X})^{-1}\right]$, which leads to $\mathrm{E}(\hat{\beta}'_{OLS}\hat{\beta}_{OLS}) = \beta'\beta + \sigma^2\mathrm{tr}\left[(\mathbf{X}'\mathbf{X})^{-1}\right]$ and that $\mathrm{V}(L^2_{OLS}) = 2\sigma^4\mathrm{tr}\left[(\mathbf{X}'\mathbf{X})^{-2}\right]$. Following Hoerl and Kennard (1970), we also get that $\mathrm{E}(L^2_{OLS}) = \sigma^2\sum_{j=1}^p \frac{1}{\lambda_j(n\hat{\Sigma}_X)}$ and $\mathrm{V}(L^2_{OLS}) = 2\sigma^4\sum_{j=1}^p \frac{1}{\lambda_j(n\hat{\Sigma}_X)^2}$, whose lower bounds are $\frac{\sigma^2}{\lambda_p(n\hat{\Sigma}_X)}$ and $\frac{2\sigma^4}{\lambda_p(n\hat{\Sigma}_X)^2}$, respectively. Similarly, we know that $\hat{\beta}_{OLS}$ is obtained by minimizing the sum of squares $\phi(\beta) = (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)$. For a generic estimator $\hat{\beta}$, we thus know that

$$\phi(\hat{\beta}) = (\mathbf{y}-\mathbf{X}\hat{\beta}_{OLS})'(\mathbf{y}-\mathbf{X}\hat{\beta}_{OLS}) + (\hat{\beta}-\hat{\beta}_{OLS})'\mathbf{X}'\mathbf{X}(\hat{\beta}-\hat{\beta}_{OLS}). \tag{19}$$

When $p \geq n$, however, $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist, so that $\hat{\beta}_{OLS}$ is unfeasible. Moreover, when p is large, due to the Marcenko–Pastur law (Marčenko & Pastur, 1967), it is likely that $\lambda_p(n\hat{\Sigma}_X)$ is really small, thus making $\mathrm{E}(L^2_{OLS})$ and $\mathrm{V}(L^2_{OLS})$ explode. Therefore, the need to recondition the eigenvalues of $\hat{\Sigma}_X$ rises, in order to limit the expected sum of squared errors and its variance. For this reason, first, we construct an alternative estimator of the coefficient vector with this aim, and second, we compare its statistical properties with the OLS and the RIDGE ones.

Let us consider the ALCE-reg estimator $\hat{\beta}_{ALCE}(\psi,\rho) = \hat{\Sigma}_A^{-1}(\psi,\rho)\hat{\sigma}_{XY}$, where the dependence on the threshold pair (ψ,ρ) is made explicit. Suppose that (5) holds. We note that solving the following problem

$$\left(\hat{\mathbf{L}}_A(\psi,\rho),\hat{\mathbf{S}}_A(\psi,\rho)\right) = \arg\min_{\mathbf{L},\mathbf{S}\in\mathbb{R}^{p\times p}} \|\hat{\Sigma}_X - (\mathbf{L}+\mathbf{S})\|_F + \psi\|\mathbf{L}\|_* + \rho\|\mathbf{S}\|_1 \tag{20}$$

is equivalent to solve the problem

$$\left(\hat{\mathbf{L}}_A(\psi,\rho),\hat{\mathbf{S}}_A(\psi,\rho)\right) = \arg\min_{\mathbf{L},\mathbf{S}\in\mathbb{R}^{p\times p}} \|\hat{\Sigma}_X - (\mathbf{L}+\mathbf{S})\|_F \tag{21}$$

subject to $\|\mathbf{L}\|_* \leq \phi_\psi$ and $\|\mathbf{S}\|_1 \leq \phi_\rho$, for some $\phi_\psi,\phi_\rho > 0$. Then, we can write $\hat{\Sigma}_A(\psi,\rho) = \hat{\mathbf{L}}_A(\psi,\rho) + \hat{\mathbf{S}}_A(\psi,\rho)$.

Theorem 1. *Suppose that $\lambda_p(\Sigma_X) = O(1)$. Under all the assumptions and conditions of Theorem 1 in Farnè and Montanari (2020), there exists a positive ζ_A such that for all $p \in \mathbb{N}$ as $n \to \infty$,* $\mathcal{P}\left(\frac{1}{p^{\alpha+\delta}}\|\hat{\Sigma}_A^{-1}(\psi,\rho) - \Sigma_X^{-1}\|_2 \leq \zeta_A\sqrt{\frac{\log p}{n}}\right) \to 1$.

In light of Theorem 1 (proof reported in Section S1), the definition of $\hat{\beta}_{ALCE}(\psi,\rho)$ is thus well-posed. In practice, the ALCE solution pair $\left(\hat{\mathbf{L}}_A(\psi,\rho),\hat{\mathbf{S}}_A(\psi,\rho)\right)$ is computed by the algorithm in Section S2. At this stage, we need to decide how to optimally select the thresholds ψ and ρ. We select them under a validation set scheme, that is, by selecting the pair $(\psi_{val},\rho_{val}) = \arg\min_{\psi\in\varphi,\rho\in\varrho}(\mathbf{y}-\mathbf{X}\hat{\beta}_{ALCE}(\psi,\rho))'(\mathbf{y}-\mathbf{X}\hat{\beta}_{ALCE}(\psi,\rho))$, where, φ, the vector of candidate eigenvalue thresholds, ψ, is composed by multiples of $\frac{1}{p}$, and $\varrho = \varphi/\sqrt{p}$. It is worth stressing that here, differently from Farnè and Montanari (2020), the tuning parameters ψ and ρ are chosen in order to optimize $\hat{\Sigma}_A(\psi,\rho)$ taking the linear dependence between $\mathbf{X}$ and $\mathbf{y}$ into account. In the same way, we derive $\hat{\beta}_{UNALCE}(\psi_{val},\rho_{val}) = \hat{\Sigma}_U^{-1}(\psi_{val},\rho_{val})\hat{\sigma}_{XY}$, where $\hat{\Sigma}_U(\psi_{val},\rho_{val}) = \hat{\mathbf{L}}_U(\psi_{val},\rho_{val}) + \hat{\mathbf{S}}_U(\psi_{val},\rho_{val})$, with $\hat{\mathbf{L}}_U(\psi_{val},\rho_{val})$ and $\hat{\mathbf{S}}_U(\psi_{val},\rho_{val})$ computed as in (12), (13) and (14).

Under all the assumptions and conditions of Theorem 1 in Farnè and Montanari (2020), $\hat{\Sigma}_A(\psi,\rho)$ is both algebraically and parametrically consistent, in the sense of Definitions 1 and 2, respectively. Under the same conditions, A.7 in Farnè and Montanari (2020) ensures that $\hat{\Sigma}_X$ is also parametrically consistent *wrt* Σ_X in spectral norm. Moreover, imposing $\lambda_p(\Sigma_X) = O(1)$, the same rate also holds for its inverse, although the strict requirement $p < n$ is needed.

Theorem 2. *Suppose that $\lambda_p(\Sigma_X) = O(1)$ and $p < n$. Under all the assumptions and conditions of Theorem 1 in Farnè and Montanari (2020), there exists a positive ζ_X such that, for all $p \in \mathbb{N}$ as $n \to \infty$,* $\mathcal{P}\left(\frac{1}{p^{\alpha}}\|\hat{\Sigma}_X^{-1} - \Sigma_X^{-1}\|_2 \leq \zeta_X\sqrt{\frac{\log p}{n}}\right) \to 1$.

Then, it is possible to prove that, provided that Theorem 1 in Farnè and Montanari (2020) holds, $\hat{\Sigma}_A(\psi,\rho)$ is the estimate with the most concentrated possible eigenvalues around the true ones among the estimators $\Sigma=\mathbf{L}+\mathbf{S}$, under the constraints $\|\mathbf{L}\|_* \leq \phi_\psi$ and $\|\mathbf{S}\|_1 \leq \phi_\rho$.

Theorem 3. *Let us define $\eta_{\Sigma_X}=\mathrm{tr}(\Sigma_X)/p$. Under the assumptions of Theorem 1 in Farnè and Montanari (2020), once fixed $\hat{\Sigma}_X$, the following statement holds for all $p\in\mathbb{N}$ as $n\to\infty$:*

$$\left(\hat{\mathbf{L}}_A(\psi,\rho),\hat{\mathbf{S}}_A(\psi,\rho)\right)=\arg\min_{\mathbf{L},\mathbf{S}\in\mathbb{R}^{p\times p}}\frac{1}{p}\left[\sum\nolimits_{j=1}^{p}(\lambda_j(\mathbf{L}+\mathbf{S})-\eta_{\Sigma_X})^2\right],$$

under the constraints $\|\mathbf{L}\|_* \leq \phi_\psi$ and $\|\mathbf{S}\|_1 \leq \phi_\rho$.

Theorem 3, proved in Section S1, is a guarantee that $\hat{\Sigma}_A(\psi,\rho)$ presents the best possible conditioning properties under the constraints $\|\mathbf{L}\|_* \leq \phi_\psi$ and $\|\mathbf{S}\|_1 \leq \phi_\rho$, for all $p\in\mathbb{N}$ as $n\to\infty$.

Let us consider $L^2_{ALCE}=(\hat{\beta}_{ALCE}-\beta)'(\hat{\beta}_{ALCE}-\beta)$. It naturally holds $\mathrm{E}(L^2_{ALCE})=\sigma^2\sum_{j=1}^{p}\frac{1}{\lambda_j(\hat{\Sigma}_A(\psi,\rho))}$ and $\mathrm{V}(L^2_{ALCE})=2\sigma^4\sum_{j=1}^{p}\frac{1}{\lambda_j(\hat{\Sigma}_A(\psi,\rho))^2}$, whose lower bounds are $\frac{\sigma^2}{\lambda_p(\hat{\Sigma}_A(\psi,\rho))}$ and $\frac{2\sigma^4}{\lambda_p(\hat{\Sigma}_A(\psi,\rho))^2}$, respectively. It follows that, when $p\geq n$, or p is large, ALCE-reg solution provides a clear improvement over OLS, due to the maximum eigenvalue concentration property of Theorem 3. Also, recalling Corollary 5 in Farnè and Montanari (2020), we learn that UNALCE has more stringent requirements for positive definiteness compared with ALCE, such that $\lambda_p(\hat{\Sigma}_U(\psi_{val},\rho_{val}))<\lambda_p(\hat{\Sigma}_A(\psi_{val},\rho_{val}))$ by construction. Therefore, although UNALCE is also improving considerably the explosive value of L^2_{OLS}, it is nonetheless expected to perform worse than ALCE, because it is systematically closer to nonpositive definiteness on average.

We now formally compare the performance of $\hat{\beta}_{ALCE}$ to the one of $\hat{\beta}_{OLS}$ and $\hat{\beta}_{RIDGE}$. Let us define $\hat{\Sigma}_R=\frac{\mathbf{X}'\mathbf{X}}{n}+\frac{k}{n}\mathbf{I}_p$. We can alternatively define RIDGE estimator as $\hat{\beta}_{RIDGE}=\hat{\Sigma}_R^{-1}\hat{\sigma}_{XY}$. Then, (4.6) in Hoerl and Kennard (1970) shows that $\mathrm{E}(L^2_{RIDGE})=\gamma_1^R(k)+\gamma_2^R(k)$, where $\gamma_1^R(k)=\sigma^2\sum_{j=1}^{p}\frac{\lambda_j(n\hat{\Sigma}_X)}{(\lambda_j(n\hat{\Sigma}_X)+k)^2}$ is the variance of $\hat{\beta}_R$, and $\gamma_2^R(k)=k^2\beta'(n\hat{\Sigma}_R)^{-2}\beta$ is the squared bias of $\hat{\beta}_{RIDGE}$. In Hoerl and Kennard (1970), the authors claim that there always exists a value of k such that the overall sum of squared errors L^2_{RIDGE} is lower than L^2_{OLS}. Comparing L^2_{RIDGE} to L^2_{ALCE}, since the variance of $\hat{\beta}_{ALCE}(\psi_{val},\rho_{val}),\gamma_1^A(\psi_{val},\rho_{val})$, can be written as $\gamma_1^A(\psi_{val},\rho_{val})=\mathrm{E}(L^2_{ALCE})=\sigma^2\sum_{j=1}^{p}\frac{1}{\lambda_j(n\hat{\Sigma}_A(\psi_{val},\rho_{val}))}$, because the expected squared bias of $\hat{\beta}_{ALCE}$ is $\gamma_2^A(\psi_{val},\rho_{val})=0$ under the conditions of Theorem 1, we first learn that $\mathrm{E}(L^2_{ALCE})$ can be much lower than $\mathrm{E}(L^2_{OLS})$ when p is large, due to Theorem 3, and, second, that it will be harder to find a value of k ensuring that $\mathrm{E}(L^2_{RIDGE})<\mathrm{E}(L^2_{ALCE})$, because $\gamma_1^A(\psi_{val},\rho_{val})<\gamma_1^O$ for the maximum eigenvalue concentration property of Theorem 3.

We can state the following corollary (proved in Section S1) on the error rate of $\hat{\beta}_{ALCE}$.

Corollary 1. *Under the conditions of Theorem 1, for some positive ζ_β it holds $\frac{1}{p^{\alpha+\delta}}\|\mathbf{E}_A\|\leq\zeta_\beta\sqrt{\frac{\log p}{n}}$ with probability approaching 1, for all $p\in\mathbb{N}$ as $n\to\infty$.*

Corollary 1 provides the error rate of $\hat{\beta}_{ALCE}$, which is related to the spikiness degree of the eigenvalues of $\mathbf{L}^*$ and the sparsity degree of $\mathbf{S}^*$. When $\alpha=1$ and $\delta=0$ (like in Fan et al., 2013), which corresponds to the case of pervasive latent factors and negligible residual sparsity, the rescaling term $\frac{1}{p^{\alpha+\delta}}$ boils down to $\frac{1}{p}$.

Let us finally analyse and compare in detail the estimation errors of the three methods. We define the estimation error matrices $\mathbf{E}_A,\mathbf{E}_R,\mathbf{E}_O$ as $\mathbf{E}_A=\hat{\Sigma}_A^{-1}\hat{\sigma}_{XY}-\Sigma_X^{-1}\sigma_{XY}$, $\mathbf{E}_R=\hat{\Sigma}_R^{-1}\hat{\sigma}_{XY}-\Sigma_X^{-1}\sigma_{XY}$, $\mathbf{E}_O=\hat{\Sigma}_X^{-1}\hat{\sigma}_{XY}-\Sigma_X^{-1}\sigma_{XY}$, respectively. First, we can write

$$\|\mathbf{E}_A\|-\|\mathbf{E}_R\|=O\left(\frac{p^{\alpha+\delta}}{\sqrt{n}}\right)-O\left(\frac{p^\alpha}{\sqrt{n}}+\frac{k}{n}\right), \tag{22}$$

because $\hat{\Sigma}_R=\hat{\mathbf{U}}_X\left(\hat{\mathbf{D}}_X+\frac{k}{n}\mathbf{I}_p\right)\hat{\mathbf{U}}_X'$, with $\hat{\Sigma}_X=\hat{\mathbf{U}}_X\hat{\mathbf{D}}_X\hat{\mathbf{U}}_X'$. Therefore, the comparison as k varies will also depend on the value of n. If n is not that large, it may be the case that $\|\mathbf{E}_A\|-\|\mathbf{E}_R\|<0$, also because $\|\mathbf{E}_R\|$ becomes larger and larger after a certain value of k, due to the increasing estimation bias (see Figure 1 in Hoerl & Kennard, 1970). It follows that, if p is large and n is not, ALCE may overcome RIDGE due to the excessive bias in the RIDGE estimate, provided that Theorem 3 holds.

The difference $\|\mathbf{E}_A\|-\|\mathbf{E}_O\|$ will intrinsically depend on the p/n ratio. When p/n is not smaller than 1, OLS is not feasible. When p/n is slightly below 1, the expected sum of squared errors is such that ALCE is going to prevail, because they are both asymptotically unbiassed, but $\gamma_1^A(\psi_{val},\rho_{val})\ll\gamma_1^O$. Moreover, a high sparsity degree in the residual covariance component $\mathbf{S}^*$ will also certainly favour ALCE, because it leads to even better conditioned covariance matrix estimates. When p is reasonably small and n is large, instead, the situation will be drastically different, with OLS likely to prevail.

Concerning prediction error, we stress that the optimal threshold pair (ψ_{val},ρ_{val}) is specifically chosen by minimizing $\phi(\hat{\beta}_{ALCE}(\psi,\rho))$ in a validation set. Similarly, in practice, the penalization parameter k is chosen by minimizing $\phi(\hat{\beta}_{RIDGE}(k))$ under a cross-validation scheme. Theoretically speaking, it is thus enough to note that $\phi(\hat{\beta}_{ALCE}(\psi,\rho)) = \|\mathbf{XE}_A\|$, $\phi(\hat{\beta}_{RIDGE}(k)) = \|\mathbf{XE}_R\|$, $\phi(\hat{\beta}_{OLS}) = \|\mathbf{XE}_O\|$, to claim that the properties of regression coefficient estimators are directly transmitted to the predictions based on estimated coefficients.

4 | SIMULATION STUDY

4.1 | Data generation

In this section, we describe the simulation study carried out to explore the performance of different estimators of a high-dimensional regression coefficient vector. We set the regression model for $i = 1,\dots,n$:

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i, \tag{23}$$

where we draw $\beta_j \sim N(10,1), j = 1,\dots,p$. The data vector $\mathbf{x}_i$ is generated in order to have a covariance matrix Σ_X respecting (5), which is a typical situation in a real high-dimensional setting. For this purpose, we set $\mathbf{x}_i = \mathbf{Bf}_i + \epsilon_i$, where $\mathbf{f}_i \sim MVN(\mathbf{0},\mathbf{I}_r)$, $\mathbf{B}$ is a semi-orthogonal $p \times r$ matrix such that $\text{tr}(\mathbf{BB}') = \theta, \theta = 0.8$, and $\epsilon_i \sim MVN(\mathbf{0},\mathbf{S}^*)$, where $\mathbf{S}^*$ is element-wise sparse positive definite such that $\text{tr}(\mathbf{S}^*) = 1-\theta$. We set $\varepsilon_i \sim MVN(0,\sigma^2(n,SNR))$, with $\sigma^2(n,SNR) = \frac{\beta'\mathbf{X}'\mathbf{X}\beta}{nSNR^2}, SNR = 10$.

The key simulation parameters are as follows: the dimension p and the sample size n; the rank r and θ, the variance proportion of Σ_X explained by $\mathbf{L}^*$; the number of off-diagonal nonzeros s in the sparse component $\mathbf{S}^*$; the percentage of nonzeros $\pi_{\mathbf{S}^*}$ over the number of off-diagonal elements; the percentage of the (absolute) residual covariance $\varrho_{\mathbf{S}^*}$; the condition number of $\Sigma_X, c(\Sigma_X) = \frac{\lambda_1(\Sigma_X)}{\lambda_p(\Sigma_X)}$; $N = 100$ replicates for each setting.

Table 1 describes the scenarios used to test estimation performance. We set three values of p, that is, $p = 100,250,500$, and two values of n, that is, $n = 100,250$. Apart from Scenario 1, which is a classical $p < n$ scenario, all the other scenarios present a $p \geq n$ situation, where the OLS estimator $\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ does not exist, because $\hat{\Sigma}_X$ is not positive definite. Under all scenarios, Σ_X follows a low rank plus sparse decomposition of type (5), where the nonzero elements of $\mathbf{S}^*$ are extremely small ($\varrho_{\mathbf{S}^*}$ close to 0). The proportion of residual nonzeros π_s is really similar across scenarios and close to 2.5%. The condition number of Σ_X increases as p increases.

4.2 | Performance metrics

For each scenario, we calculate $\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and $\hat{\beta}_{POET} = \hat{\Sigma}_{POET}^{-1}\hat{\sigma}_{XY}$, where $\hat{\Sigma}_{POET}$ is derived by POET as in Fan et al. (2013), with the sparsity threshold selected by cross-validation. Then, we derive $\hat{\beta}_{ALCE}(\psi_{val},\rho_{val})$ by the algorithm in Section S2, $\hat{\beta}_{UNALCE}(\psi_{val},\rho_{val})$ as in (12), (13), (14), and we compute $\hat{\beta}_{RIDGE-\min}$ and $\hat{\beta}_{LASSO-\min}$, which are, respectively, the RIDGE/LASSO estimate with $k = k_{min}$, that is, the value of k returning the minimum cross-validated mean square error of predictions.

On each replicate $t = 1,\dots,N$ of model (23), we calculate the estimates $\hat{\Sigma}_t = \hat{\mathbf{L}}_t + \hat{\mathbf{S}}_t$, obtained by ALCE, UNALCE and POET and the matrix $\hat{\Sigma}_{R,min} = \frac{1}{n}(\mathbf{X}'\mathbf{X} + k_{\min}\mathbf{I}_p)$. We derive the two following metrics: $Loss_{\Sigma,t} = \|\hat{\Sigma}_t - \Sigma_X\|$ and $c(\hat{\Sigma}_t) = \frac{\lambda_1(\hat{\Sigma}_t)}{\lambda_p(\hat{\Sigma}_t)}$. We focus on the estimated coefficient vector $\hat{\beta}$ via all considered methods, that is, OLS, POET, ALCE, UNALCE, RIDGE and LASSO. Then, we measure their estimation performance as follows: $\text{M}(\hat{\beta}) = \frac{1}{N}\sum_{t=1}^{N}\hat{\beta}_t$ and $\mathbf{b}_{\hat{\beta}} = \text{M}(\hat{\beta}) - \beta$; $\text{VAR}_{t,\hat{\beta}} = (\hat{\beta}_t - \text{M}(\hat{\beta}))^2$; $\text{MSE}_{t,\hat{\beta}} = (\hat{\beta}_t - \beta)^2$. We generate for each replicate $t^* = 1,\dots,N$ one test observation, $(\mathbf{x}_{t^*}, y_{t^*})$, from model (23), we calculate the prediction $\hat{y}_{t^*} = \hat{\beta}'\mathbf{x}_{t^*}$, and we derive the prediction mean square error $err_{\hat{y}} = \sum_{t^*=1}^{N}(\hat{y}_{t^*} - y_{t^*})^2$.

Finally, we obtain the following overall performance metrics:

- $\hat{b}_M = \text{Avg}\left\{\mathbf{b}_{\hat{\beta}}\right\}_{j=1,\dots,p}$;

TABLE 1 Scenarios 1–6: key parameters.

	p	n	p/n	SNR	r	s	θ	$\varrho_{\mathbf{S}^*}$	$c(\Sigma_X)$
Scenario 1	100	250	0.4	10	2	256	0.8	4.01E-05	59.82
Scenario 2	100	100	1	10	2	256	0.8	4.01E-05	59.82
Scenario 3	250	250	1	10	5	1563	0.8	6.39E-05	89.97
Scenario 4	250	100	2.5	10	5	1563	0.8	6.39E-05	89.97
Scenario 5	500	250	2	10	10	6372	0.8	9.35E-05	116.77
Scenario 6	500	100	5	10	10	6372	0.8	9.35E-05	116.77

- $\widehat{SD}_M = \text{Avg}\left\{\sqrt{\frac{1}{N}\sum_{t=1}^{N} VAR_{t,\hat{\beta}}}\right\}_{j=1,\ldots,p}$;
- $\widehat{RMSE}_M = \text{Avg}\left\{\sqrt{\frac{1}{N}\sum_{t=1}^{N} MSE_{t,\hat{\beta}}}\right\}_{j=1,\ldots,p}$;

where $\hat{b}_M$, $\widehat{SD}_M$ and $\widehat{RMSE}_M$ are the average bias, standard deviation and root mean square error across all the coefficients, respectively. Two more performance metrics are then derived by just averaging over the N replicates: $Loss_\Sigma = M(Loss_{\Sigma,t}), c(\hat{\Sigma})_M = M(c(\hat{\Sigma}_t))$.

4.3 | Simulation results

Here, we report the simulation results about coefficient estimation and prediction performance. For the results on the performance of low rank and sparse component estimates, we refer to Section S3.

We start by the analysis of Scenario 1, which is the most favourable to OLS. In Table 2, we report the error metrics relative to the overall covariance matrix estimates and the regression coefficients. We can note that OLS is by far the best method to estimate β in this case. LASSO is the second best, due to a very limited variance. Note that LASSO does not estimate any zero coefficient in this case. RIDGE is not doing so well, due to a strong bias. Then, we note that UNALCE, ALCE and POET, that is, the methods based on a low rank plus sparse assumption, work poorly in this case. This happens because, in a $n > p$ case, the unnecessary variance introduced by estimation mechanisms involving thresholding procedures leads to too variable estimates. This is also reflected in the prediction performance.

Concerning Scenario 2, Table 3 shows that OLS cannot be computed when $p \geq n$. RIDGE regression is extremely biassed. UNALCE performs better than the competitors in terms of covariance loss but worse in terms of coefficient estimates. ALCE offers the best compromise between bias and variance in coefficient estimation, apart from LASSO, which anyway reports an average percentage of zero coefficients equal to 27.75%. Focusing on prediction performance, we note that LASSO is the best in this case, followed by RIDGE and ALCE.

Analysing the performance in Scenario 3, we can observe in Table 4 that ALCE is able to overcome in the RMSE even LASSO, which presents an average of 25.71% zero coefficients in the estimated β. Concerning prediction error, ALCE comes first while UNALCE comes second in this case. POET performs instead very badly, due to its excessive variability, coming from bad conditioning properties. In contrast, RIDGE covariance estimate is too regularized and therefore very biassed.

TABLE 2 Scenario 1: Performance metrics on covariance matrix and regression coefficient estimates.

	OLS	POET	ALCE	UNALCE	RIDGE	LASSO
$Loss_\Sigma$	0.1275	0.0288	0.0240	0.1041	0.0249	
$c(\hat{\Sigma})_M$	50.3784	50.5221	4.7703	335.3218	3.2322	
$\hat{b}_M$	0.0329	−1.0325	−1.9723	−0.4451	−2.0337	-0.0649
$\widehat{SD}_M$	2.0002	10.6623	5.2994	6.3342	1.8812	2.0282
$\widehat{\text{RMSE}}_M$	2.0084	10.7909	5.6985	6.3880	2.7806	2.0371
$err_{\hat{y}}$	1.5249	5.0667	2.9282	3.2580	1.6926	1.5501

Abbreviations: ALCE, ALgebraic Covariance Estimator; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least square; POET, principal orthogonal complement thresholding; UNALCE, UNshrunk ALCE.

TABLE 3 Scenario 2: Performance metrics on covariance matrix and regression coefficient estimates.

	OLS	POET	ALCE	UNALCE	RIDGE	LASSO
$Loss_\Sigma$		0.1744	0.1782	0.1660	0.3258	
$c(\hat{\Sigma})_M$		85.3694	54.5152	68.1953	5.4656	
$\hat{b}_M$		−1.0325	−2.1540	−0.5621	−9.3732	−4.2476
$\widehat{SD}_M$		10.6623	8.3666	10.1546	0.5867	6.1503
$\widehat{RMSE}_M$		10.7909	8.7133	10.2351	9.4786	7.4959
$err_{\hat{y}}$		5.0667	4.2558	4.8431	3.9065	3.5836

Abbreviations: ALCE, ALgebraic Covariance Estimator; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least square; POET, principal orthogonal complement thresholding; UNALCE, UNshrunk ALCE.

TABLE 4 Scenario 3: Performance metrics on covariance matrix and regression coefficient estimates.

	OLS	POET	ALCE	UNALCE	RIDGE	LASSO
$Loss_{\Sigma}$		0.0901	0.1444	0.1142	0.6875	
$c(\hat{\Sigma})_M$		110.8359	51.4028	71.1490	2.5419	
$\hat{b}_M$		−0.7465	−4.7092	−2.8010	−9.4508	−4.2132
$\widehat{SD}_M$		10.5739	5.5576	7.6022	0.5517	6.0189
$\widehat{RMSE}_M$		10.6538	7.3308	8.1497	9.5235	7.3738
$err_{\hat{y}}$		4.6145	3.5863	3.6945	4.5236	3.8776

Abbreviations: ALCE, ALgebraic Covariance Estimator; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least square; POET, principal orthogonal complement thresholding; UNALCE, UNshrunk ALCE.

TABLE 5 Scenario 4: Performance metrics on covariance matrix and regression coefficient estimates.

	OLS	POET	ALCE	UNALCE	RIDGE	LASSO
$Loss_{\Sigma}$		0.1464	0.1879	0.1600	0.3518	
$c(\hat{\Sigma})_M$		211.5288	59.4695	85.1038	3.8387	
$\hat{b}_M$		−3.2978	−4.7464	−2.7805	−9.5795	−8.7588
$\widehat{SD}_M$		18.2251	9.0520	12.5209	0.6903	3.9818
$\widehat{RMSE}_M$		18.6312	10.3005	12.9046	9.6712	9.9294
$err_{\hat{y}}$		8.5758	5.0294	6.5145	4.8290	5.1829

Abbreviations: ALCE, ALgebraic Covariance Estimator; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least square; POET, principal orthogonal complement thresholding; UNALCE, UNshrunk ALCE.

TABLE 6 Scenario 5: Performance metrics on covariance matrix and regression coefficient estimates.

	OLS	POET	ALCE	UNALCE	RIDGE	LASSO
$Loss_{\Sigma}$		0.0824	0.1232	0.0987	0.1529	
$c(\hat{\Sigma})_M$		223.6059	59.8279	85.3687	6.0234	
$\hat{b}_M$		−3.9754	−5.6329	−4.0860	−9.4855	−8.2014
$\widehat{SD}_M$		15.5912	6.6111	8.9797	0.6941	4.6687
$\widehat{RMSE}_M$		16.1936	8.7460	9.9287	9.5737	9.6415
$err_{\hat{y}}$		7.3630	4.3897	4.7376	4.3087	4.2394

Abbreviations: ALCE, ALgebraic Covariance Estimator; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least square; POET, principal orthogonal complement thresholding; UNALCE, UNshrunk ALCE.

In Table 5, we observe that under Scenario 4 RIDGE and LASSO are prevailing in the RMSE, but ALCE is really close and is the second best (behind RIDGE) in the prediction error. On the contrary, UNALCE and (even more) POET lie far. This occurs because a more biassed estimate of Σ_X, but with a lower condition number, results to be more effective for estimating β. We stress however the extreme bias of RIDGE and that LASSO in this case produces an average of 83.8% zero coefficients.

Table 6 shows that, under Scenario 5, POET is completely out of target for coefficient estimation, while ALCE is prevailing in the RMSE against RIDGE and LASSO by a good margin, showing the best balance between bias and variance. Concerning prediction error, ALCE, RIDGE and LASSO are really close, although LASSO presents 76.23% zero coefficients on average. It is remarkable that, when comparing the *median* squared prediction error, ALCE is prevailing over all the competitors. This means that, when p is large, the larger variance of ALCE and UNALCE coefficients compared with RIDGE and LASSO may occasionally impact on prediction error, while preserving the goodness of systematic performance.

In the end, concerning Scenario 6, Table 7 shows that the variance of ALCE explodes, in a way that awards RIDGE and LASSO in the RMSE. The gap with RIDGE/LASSO is particularly important in the prediction error, although we must note that 93.95% of coefficients are estimated as zero by LASSO. All in all, the ratio p/n is too large in this case to ensure the effectiveness of Theorems 1 and 3.

TABLE 7 Scenario 6: Performance metrics on covariance matrix and regression coefficient estimates.

	OLS	POET	ALCE	UNALCE	RIDGE	LASSO
$Loss_{\Sigma}$		0.1385	0.1587	0.1475	0.5606	
$c(\hat{\Sigma})_M$		1877789	70.5391	102.8426	2.8386	
$\hat{b}_M$		8.5758	−5.6909	−4.0805	−9.6669	−9.5867
$\widehat{SD}_M$		−3.2978	10.9784	15.2220	0.7840	2.6333
$\widehat{RMSE}_M$		18.2251	12.4715	15.8731	9.7615	10.1526
$err_{\hat{y}}$		18.6312	9.0990	11.1340	4.6971	4.9583

Abbreviations: ALCE, ALgebraic Covariance Estimator; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least square; POET, principal orthogonal complement thresholding; UNALCE, UNshrunk ALCE.

5 | CONCLUSIONS

In this paper, we have proposed a new estimator of a high-dimensional regression coefficient vector, named ALCE-reg, based on estimating the covariance matrix of the predictors by nuclear norm plus l_1 norm penalization under the low rank plus sparse structure assumption. We have shown that, theoretically speaking, ALCE-reg may improve over RIDGE/LASSO when both p and n are large (allowing for $p \geq n$), because of a systematically much lower bias. The new method also relevantly outperforms OLS, which is unfeasible if $p \geq n$ or very unstable when p is large.

A wide simulation study shows that adopting for threshold selection a tailored method which targets prediction error turns out to be an advantage for full regression coefficient vector estimation in high dimensions. Another relevant finding is that a relatively biassed covariance matrix estimate with a low condition number performs better in terms of regression coefficient estimation than a good covariance matrix estimate with a systematically worse conditioning. Additionally, when p/n is slightly smaller or larger than 1, ALCE-reg also systematically improves the prediction error.

In light of these findings, we have discovered that ALCE-reg represents a good compromise between methods like OLS and POET, too much affected by sample eigenvalues, which results in a large estimation variance, and RIDGE/LASSO, characterized by large estimation bias. ALCE-reg is particularly effective compared with competitors when both p and n are large, and the ratio p/n is not too far from 1. In the future, it will be interesting to conduct further experiments under different coefficient or covariance structures and to add to simulation study competitors based on direct eigenvalue shrinkage.

ACKNOWLEDGEMENTS

We thank Trevor Hastie for inspiring us this work with his talk “Ridge Regularization: An Essential Concept in Data Science” delivered at ISI 2021 in a session named “Data Science and Statistics” and two anonymous referees for their kind and constructive comments. Open Access Funding provided by Universita degli Studi di Bologna within the CRUI-CARE Agreement.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in UCI Repository at http://archive.ics.uci.edu/ml/datasets. These data were derived from the following resources available in the public domain: - Arrhythmia Data Set, http://archive.ics.uci.edu/ml/datasets/Arrhythmia.

ORCID

Matteo Farnè https://orcid.org/0000-0002-2403-6599

REFERENCES

Chamberlain, G., & Rothschild, M. (1982). Arbitrage, factor structure, and mean-variance analysis on large asset markets.

Chandrasekaran, V., Parrilo, P. A., & Willsky, A. S. (2010). Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, pp. 1610–1613.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., & Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, *21*(2), 572–596.

Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(4), 603–680.

Farnè, M., & Montanari, A. (2020). A large covariance matrix estimator under intermediate spikiness regimes. *Journal of Multivariate Analysis*, *176*, 104577.

Fazel, M. (2002). Matrix rank minimization with applications. (Ph.D. Thesis), Stanford University.

Fazel, M., Hindi, H., & Boyd, S. P. (2001). A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01ch37148)*, *6*, IEEE, pp. 4734–4739.

Hastie, T. (2020). Ridge regularization: An essential concept in data science. *Technometrics*, *62*(4), 426–433.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hoerl, R. W. (2020). Ridge regression: A historical context. *Technometrics*, *62*(4), 420–425.

Joseph, V. R. (2020). *Celebrating 50 years of ridge regression*, Vol. 62: Taylor & Francis.

Le, C. M., Levin, K., Bickel, P. J., & Levina, E. (2020). Comment: Ridge regression and regularization of large matrices. *Technometrics*, *62*(4), 443–446.

Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, *1*(4), 457.

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

Zou, H. (2020). Comment: Ridge regression–still inspiring after 50 years. *Technometrics*, *62*(4), 456–458.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Farnè, M., & Montanari, A. (2023). High-dimensional regression coefficient estimation by nuclear norm plus l_1 norm penalization. *Stat*, *12*(1), e548. https://doi.org/10.1002/sta4.548