Electronic Journal of Statistics
Vol. 17 (2023) 2447–2484
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2154

Envelopes and principal component regression*†

Xin Zhang, Kai Deng and Qing Mai

Department of Statistics
Florida State University
Tallahassee, FL, 32306
e-mail: xzhang8@fsu.edu; kd18h@stat.fsu.edu; mai@stat.fsu.edu

Abstract: Envelope methods offer targeted dimension reduction for various statistical models. The goal is to improve efficiency in multivariate parameter estimation by projecting the data onto a lower-dimensional subspace known as the envelope. Envelope approaches have advantages in analyzing data with highly correlated variables, but their iterative Grassmannian optimization algorithms do not scale very well with high-dimensional data. While the connections between envelopes and partial least squares in multivariate linear regression have promoted recent progress in high-dimensional studies of envelopes, we propose a more straightforward way of envelope modeling from a new principal component regression perspective. The proposed procedure, Non-Iterative Envelope Component Estimation (NIECE), has excellent computational advantages over the iterative Grassmannian optimization alternatives in high dimensions. We develop a unified theory that bridges the gap between envelope methods and principal components in regression. The new theoretical insights also shed light on the envelope subspace estimation error as a function of eigenvalue gaps of two symmetric positive definite matrices used in envelope modeling. We apply the new theory and algorithm to several envelope models, including response and predictor reduction in multivariate linear models, logistic regression, and Cox proportional hazard model. Simulations and illustrative data analysis show the potential for NIECE to improve standard methods in linear and generalized linear models significantly.

MSC2020 subject classifications: Primary 62H25; secondary 62J12.
Keywords and phrases: Envelope methods, penalized matrix decomposition, principal component, sufficient dimension reduction.

Received September 2022.

1. Introduction

The idea of envelope modeling is to exploit the covariance structure in variables and identify and eliminate the part of data tangent to the parameter space of interest but only brings extraneous variability to model fitting. Firstly introduced in Cook, Li and Chiaromonte (2010), envelope methods have demonstrated

*Research is partly supported by grants CCF-1908969, DMS-2053697, and DMS-2113590 from the US National Science Foundation.
†The authors appreciate all the insightful comments and valuable suggestions from the Editor, the Associate Editor and two anonymous reviewers.

promising performances in various multivariate statistical problems. Different envelope structures are proposed in the multivariate linear regression model, for example, partial envelope (Su and Cook, 2011), inner envelope (Su and Cook, 2012), scaled envelope (Cook and Su, 2013), and simultaneous envelope (Cook and Zhang, 2015a). Connections between envelopes and classical multivariate analysis methods are intensively studied: envelope and partial least squares (Cook, Helland and Su, 2013a), envelope and reduced-rank regression (Cook, Forzani and Zhang, 2015), envelope and Bayesian statistics (Khare et al., 2017), envelope and discriminant analysis (Zhang and Mai, 2019), and among others. Recent extensions of envelopes beyond the standard linear models is also an emerging field of research. Cook and Zhang (2015b) proposed a general constructive framework for adapting envelope methods to any estimation procedure, and applied this to generalized linear models and Cox regression. More recently, envelope methodology has been extended to quantile regression (Ding et al., 2021), Huber regression (Zhou, Cook and Zou, 2020), matrix-variate (Ding and Cook, 2018) and tensor-variate regressions (Li and Zhang, 2017). See Cook (2018) for a detailed introduction, and Cook (2020) and Lee and Su (2020) for recent reviews, on envelopes.

In this paper, we will study two challenging and fundamental questions in high-dimensional envelopes: (i) How to overcome the computational bottleneck of envelope estimation in ultra high-dimensions; and (ii) How to quantify the envelope subspace estimation error in a generic model-free setting, especially when the dimension p diverges much faster than the sample size n. To address the computational issue (i) and the theoretical issue (ii) of high-dimensional envelopes, we propose a scalable and straightforward envelope estimation procedure based on a novel principal components regression formulation. Computationally, the proposed novel procedure, Non-Iterative Envelope Component Estimation (NIECE), is a perfect complement to the more delicate and much more expensive iterative Grassmannian optimization approaches that were used in the literature (i.e. all the envelope methods as mentioned earlier). Theoretically, a unified theory for NIECE shows that we can estimate the envelope subspace consistently when the dimension diverges exponentially fast as the sample size in a wide range of models and applications. To the best of our knowledge, this paper is also the first in (a) establishing the finite sample connections between envelope estimation and principal components; (b) providing non-asymptotic analysis of envelope subspace estimation error; (c) devising a novel "eigenvalue gap" argument for theoretical analysis on two positive semi-definite matrices $\mathbf{M}$ and $\mathbf{U}$ in a model-free setting of envelopes: For NIECE, we extract the eigenvectors $\mathbf{v}_i$ of $\mathbf{M}$ and examine the "eigen-gaps" of the quadratic form $\mathbf{v}_i^{\mathsf{T}}\mathbf{U}\mathbf{v}_i$.

In what follows, we outline in Sect. 1.1 the population construct of an envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ following the same notation in Cook and Zhang (2015b) for a general-purpose multivariate parameter estimation. We provide a literature review on the computational aspects, algorithms, and existing high-dimensional theory of envelopes in Sect. 1.2. The population-level connections between envelopes and principal components are given in Sect. 1.3 to motivate our methodology. The specific goals and organization of this article are outlined in Sect. 1.4.

1.1. Envelopes: definition and working mechanism

For a matrix $\mathbf{B} \in \mathbb{R}^{p\times d}$ with full column rank d, let $\mathcal{B} = \text{span}(\mathbf{B}) \subseteq \mathbb{R}^p$ denote the subspace spanned by the columns of $\mathbf{B}$. Let $\mathbf{P}_{\mathcal{B}} = \mathbf{P}_{\mathbf{B}} = \mathbf{B}(\mathbf{B}^{\mathsf{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathsf{T}}$ denote the projection onto $\mathcal{B}$, and $\mathbf{Q}_{\mathcal{B}} = \mathbf{Q}_{\mathbf{B}} = \mathbf{I}_p - \mathbf{P}_{\mathbf{B}}$ denote the projection onto the orthogonal complement of $\mathcal{B}$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix. Formally, an envelope is defined as follows (see, Cook and Zhang, 2015b, for example), where the notion of reducing subspace is from functional analysis (e.g. Conway, 1990).

Definition 1. A reducing subspace of a symmetric matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ is defined as the subspace $\mathcal{R} \subseteq \mathbb{R}^p$ such that $\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$. An $\mathbf{M}$-envelope of a subspace $\mathcal{U} = \text{span}(\mathbf{U}) \subseteq \mathbb{R}^p$, denoted as $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ or $\mathcal{E}_{\mathbf{M}}(\mathcal{U})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{U}$.

The construction and estimation of envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ is the center of our future development. Cook, Li and Chiaromonte (2010) showed that the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ is unique and always exists. In multivariate analysis, $\mathbf{U}$ comes from

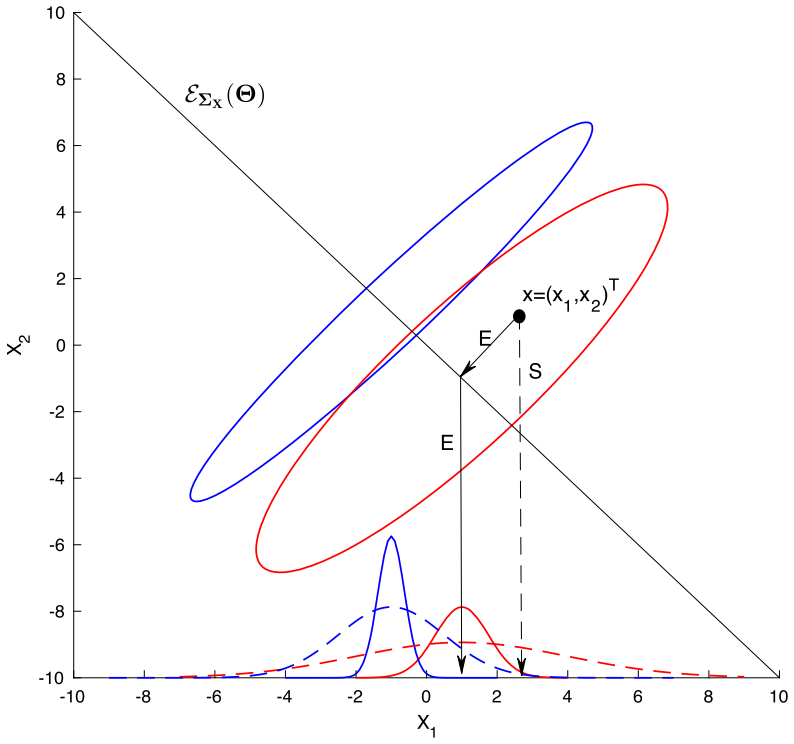

FIG 1. *The working mechanism of envelope methodology when estimating the mean difference of* $\mathbf{X}$ *across two groups (indicated by red and blue colors). The plot illustrates the efficiency gain by envelope methods in estimating the mean difference* $\boldsymbol{\theta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$*; representative projection paths, labeled 'E' for envelope analysis and 'S' for standard analysis.*

the parameter of interest; and $\mathbf{M}$ represents a helpful nuisance parameter. The specific choices of $\mathbf{M}$ and $\mathbf{U}$ depend on the context of applications and the goals of studies. They may vary from one model to another, as we see later in analyzing several envelope models in later sections (including response and predictor envelopes in linear regression, envelopes in logistic regression and Cox model).

Figure 1 illustrates the working mechanism of envelope methods. Consider a simple case with group indicator $G \in \{1, 2\}$ and predictor $\mathbf{X} = (X_1, X_2)^\mathsf{T} \in \mathbb{R}^2$. The target parameter is the mean difference $\boldsymbol{\theta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 = \mathrm{E}(\mathbf{X} \mid G = 2) - \mathrm{E}(\mathbf{X} \mid G = 1)$ or equivalently the matrix form $\boldsymbol{\Theta} = \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}$ (i.e. the $\mathbf{U}$ matrix in Definition 1). The nuisance parameter is the marginal covariance $\boldsymbol{\Sigma}_\mathbf{X} = \mathrm{cov}(\mathbf{X})$ (i.e. the $\mathbf{M}$ matrix in Definition 1); and it helps improving the estimation efficiency in the mean difference as illustrated in Fig. 1. The two ellipses represent the contours of the conditional distributions of $\mathbf{X}$ within each of the two classes. We can see that the mean difference lies in the shorter axis of the ellipses, while the long axis brings large variability but does not contribute to comparing the means. Therefore, if we calculate the difference in the sample means, the differences will also be blurred. For X_1, as shown in Fig. 1, we can see that the two empirical distributions of $X_1 \mid (G = 1)$ and $X_1 \mid (G = 2)$, as the two lower flatten curves represented, are almost indistinguishable. In contrast, the envelope method can identify the ellipses' longer axis as *immaterial variation*. The envelope estimation procedure will project all the data first onto the *envelope*, which contains all the *material variation*, and then onto each axis of $\mathbf{X}$ to compare the two classes. The elimination of immaterial variation leads to much well-separated two distributions as shown in Fig. 1, and therefore massive gain in estimation accuracy of $\boldsymbol{\theta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

1.2. A brief overview of envelope estimation

Without loss of generality, we assume $\mathbf{U}$ is symmetric positive semi-definite henceforth because we can always replace $\mathbf{U}$ with $\mathbf{U}\mathbf{U}^\mathsf{T} \geq 0$ without changing the column subspace $\mathrm{span}(\mathbf{U})$ and the envelope $\mathcal{E}_\mathbf{M}(\mathbf{U})$. The dimension of $\mathcal{E}_\mathbf{M}(\mathbf{U})$, denoted by u, $0 \leq u \leq p$, is important for all envelope methods and can be estimated consistently (Zhang and Mai, 2018). In this paper, we assume the envelope dimension $u \geq 1$ is known or pre-specified.

Given the dimension of an envelope, envelope estimation generally reduces to solving for $\widehat{\boldsymbol{\Gamma}} \in \mathbb{R}^{p \times u}$ from the following constrained optimization and letting $\widehat{\mathcal{E}}_\mathbf{M}(\mathbf{U}) = \mathrm{span}(\widehat{\boldsymbol{\Gamma}})$,

$$\widehat{\boldsymbol{\Gamma}} = \underset{\boldsymbol{\Gamma}^\mathsf{T}\boldsymbol{\Gamma} = \mathbf{I}_u}{\mathrm{argmin}}\, \mathrm{J}_n(\boldsymbol{\Gamma}), \quad \mathrm{J}_n(\boldsymbol{\Gamma}) = \log|\boldsymbol{\Gamma}^\mathsf{T}\widehat{\mathbf{M}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^\mathsf{T}(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1}\boldsymbol{\Gamma}|, \tag{1}$$

where the symmetric matrices $\widehat{\mathbf{M}} > 0$ and $\widehat{\mathbf{U}} \geq 0$ are finite sample estimators of their population counterparts $\mathbf{M}$ and $\mathbf{U}$. From the orthogonality constraint $\boldsymbol{\Gamma}^\mathsf{T}\boldsymbol{\Gamma} = \mathbf{I}_u$, we note that (1) is a non-convex optimization on Stiefel manifolds.

Equivalently, if we consider the subspace as the argument in the above optimization, then (1) is also equivalent to a non-convex optimization on Grassmann manifolds (aka Grassmannian).

Almost all envelope methods are connected to this type of optimization. For a broad class of envelope estimators, it was shown in Cook and Zhang (2015b) that the partially maximized log-likelihood function is $(-n/2)$ times certain sample version of (1). Therefore, normal likelihood-based envelope methods are usually required to solve a version of (1). Furthermore, given a parameter vector of interests, $\boldsymbol{\theta} \in \mathbb{R}^p$, and some standard $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\theta}}$, the particular choice of $\mathbf{U} = \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}$ and $\mathbf{M}$ being the asymptotic covariance of $\widehat{\boldsymbol{\theta}}$ reproduces these likelihood-based envelope methods in the literature. The envelope estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^{\mathsf{T}}\widehat{\boldsymbol{\theta}}$ is asymptotically more efficient than the standard estimator $\widehat{\boldsymbol{\theta}}$ in various contexts such as linear and generalized linear models. Such envelope estimation solves for $\widehat{\boldsymbol{\Gamma}}$ based on (1) and then plugs-in $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^{\mathsf{T}}\widehat{\boldsymbol{\theta}}$, is essentially a two-stage projection pursuit multivariate parameter estimation relying on this generic objective function of envelope basis. In this paper, we focus on the envelope subspace estimation problem and characterizing the estimation error in terms of the distance between the true and estimated envelope subspaces when p is allowed to diverge much faster than n.

Most computational methods for manifold optimizations can be directly used to solve (1), see Edelman, Arias and Smith (1998); Absil, Mahony and Sepulchre (2009) and Wen and Yin (2013) for more background. By exploiting the geometry of envelopes, Cook and Zhang (2016) developed the 1D algorithm that approximately solves (1) by sequential optimization over u one-dimensional vectors than over $p \times u$ matrices. Relatedly, Cook and Zhang (2018) developed the envelope coordinate descent algorithm that sequentially solves the same 1D algorithm objective functions and is shown to be even faster than the original 1D algorithm. Another approach for solving (1) is proposed by Cook, Forzani and Su (2016). The authors suggest to remove the orthogonality constraint by rotating $\boldsymbol{\Gamma}$ to as $\boldsymbol{\Gamma} = (\mathbf{I}_u, \mathbf{A}^{\mathsf{T}})^{\mathsf{T}}\mathbf{G}$ for some coordinates $\mathbf{A} \in \mathbb{R}^{(p-u)\times u}$ and a full rank matrix $\mathbf{G} \in \mathbb{R}^{u\times u}$. Then the iterative Grassmannian optimization is transformed into unconstrained iterative optimization over matrix $\mathbf{A}$. All such computational methods adopt gradient-based iterative schemes for a highly non-convex objective function, which involves log-determinant of symmetric positive definite matrices, and are quite sensitive to initialization when dimensions (p, u) are high.

Extending envelope methods to high dimensional data analysis is challenging, mainly due to the delicate and complicated optimization in (1) that provides no closed-form for $\widehat{\boldsymbol{\Gamma}}$ and thus makes the envelope estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^{\mathsf{T}}\widehat{\boldsymbol{\theta}}$ almost mysterious. Nevertheless, some progress on this problem has been made. Su et al. (2016) proposed a sparse envelope estimator for response reduction and variable selection; and Zhu and Su (2020) proposed a sparse envelope estimator for predictor reduction and variable selection. Both methods rely on a penalized version of (1) with sparsity inducing penalties on rows of $\boldsymbol{\Gamma}$. This type of coordinate-

free penalty (Chen, Zou and Cook, 2010) takes the form of $P_\lambda(\mathbf{\Gamma}) = \lambda_i \|\boldsymbol{\gamma}_i\|_2$, $i = 1, \ldots, p$, where $\boldsymbol{\gamma}_i \in \mathbb{R}^u$ is the i-th row of $\mathbf{\Gamma}$. Their approaches come at the cost of assuming that the dimension p (i.e. the total number of response or predictor variables) can not diverge too quickly: $(p+s)\log(p)/n \to 0$, where s is the sparsity level. Moreover, the p tuning parameters are assumed to satisfy $\sqrt{n}\lambda_i \to 0$ for the s truly active variables and $\sqrt{n}\lambda_i \to \infty$ for the $(p-s)$ inactive variables.

To overcome the computational bottleneck of envelope estimation in the ultra high-dimensional regimes, i.e. $\log(p)/n \to 0$, we need alternative objectives of envelope estimation. Indeed, we propose a sparse principal component approach to envelope estimation that does not involve the usual manifold optimizations. The new approach aims not to replace the existing methods when applicable but to provide a feasible approach when they are not. It also fills the gap in theoretical analysis and provides valuable insights about principal components in regression.

1.3. Envelope and principal components

Principal component analysis (Jolliffe, 2002; Jolliffe and Cadima, 2016) is routinely used in exploratory data analysis as a dimension reduction and visualization method and in regression to improve prediction. In some cases, the PCs with the smallest variances (also known as the minor components, see Oja (1992) for example) are kept for subsequent analysis in addition to the PCs with the largest variances. Using latent variable and normal likelihood, Tipping and Bishop (1999) provided a model-based probabilistic formulation for principal component analysis where only the leading PCs are relevant in the analysis; Welling, Williams and Agakov (2004) generalized such approach to probabilistic models where the maximum likelihood estimation for the dimension reduction subspace is a combination of PCs with largest and smallest variances; Zhang and Chen (2020) discussed scenarios that the maximum likelihood is attained by any combinations of PCs.

On principal component regression, Jolliffe (1982) noted that the PCs with smaller variability could be as useful as PCs with the largest variability; more recently, Lang and Zou (2020) provided a response-guided formulation for regression with PCs that unifies the usual principal component regression with ridge regression (Hoerl and Kennard, 1970).

Our approach is conceptually closely related to principal components regression but is formulated in more general settings (i.e. beyond linear regression) with rigorous theoretical justification using envelopes. From Definition 1, the subspace spanned by any set of eigenvectors of $\mathbf{M}$ is a reducing subspace of $\mathbf{M}$. In this article, to ensure identifiability, we assume that the first $d \lesssim n$ eigenvalues of $\mathbf{M}$ are distinct and their span contains $\text{span}(\mathbf{U})$. In practice, we will specify $d \simeq n$ in the proposed algorithm so that we lose little information by focusing on the first d eigenvectors. When some of the eigenvalues coincide, we will be targeting an upper bound of the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$, similar to Proposition 4 in Cook and Zhang (2018).

In many applications, $\mathbf{M} = \mathbf{\Sigma}_{\mathbf{X}} \equiv \text{cov}(\mathbf{X}) > 0$. Then the population level connection between envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ and the principal components (PCs) of $\mathbf{X}$ can be easily shown as follows,

$$\mathcal{E}_{\mathbf{M}}(\mathbf{U}) = \sum_{\substack{i=1, \\ \mathbf{v}_i^{\mathsf{T}}\mathbf{U}\mathbf{v}_i \neq 0}}^{d} \text{span}(\mathbf{v}_i), \tag{2}$$

where $\mathbf{v}_i \in \mathbb{R}^p$ is the i-th eigenvector of $\mathbf{M} = \mathbf{\Sigma}_{\mathbf{X}}$ and hence the i-th PC. Our algorithm is an efficient realization of the above characteristics of envelopes and is thus straightforward to implement based on the penalized matrix decomposition (PMD, Witten, Tibshirani and Hastie, 2009). In our theoretical study, we first establish the consistency of PMD, which is of independent interest. It is also straightforward to adopt other sparse PC methods in our algorithm (Zou, Hastie and Tibshirani, 2006; Shen and Huang, 2008; Amini and Wainwright, 2009; Ma, 2013; Vu and Lei, 2013).

Several recent studies have connected envelopes with PCs. In particular, Li et al. (2016) studied the supervised dimension reduction problem, where the envelope is formed as the leading PCs in a latent variable model and estimated by EM algorithm. The covariance can be written as $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{V}\mathbf{\Sigma}_{\mathbf{f}}\mathbf{V}^{\mathsf{T}} + \sigma^2\mathbf{I}_p$, where $\mathbf{\Sigma}_{\mathbf{f}} \in \mathbb{R}^{u\times u}$ is the covariance of the u latent variables $\mathbf{f}$ and $\mathbf{V} \in \mathbb{R}^{p\times u}$ is the basis of the u-dimensional target subspace. Franks and Hoff (2019) studied a shared subspace spiked model that $\mathbf{\Sigma}_{\mathbf{X}|Y=k} = \sigma_k^2(\mathbf{V}\mathbf{\Psi}_k\mathbf{V}^{\mathsf{T}} + \mathbf{I}_p)$ for groups of observations indexed by $k = 1, \ldots, K$, where $\mathbf{\Psi}_k$ symmetric positive definite matrix and $\mathbf{V} \in \mathbb{R}^{p\times u}$ is the basis of the u-dimensional target subspace. More recently, Franks (2020) extended the model of Li et al. (2016) to large-p-small-n setting by adopting the Monte Carlo EM algorithm similar to that in Franks and Hoff (2019). These models, as well as the common principal component analysis models (Flury, 1984, 1988; Schott, 1999), can be viewed as special forms of the more general envelope structure in this paper. However, none of these methods has established consistency in subspace estimation when p diverges with n. Although the population level connections between the envelope and principal components, including (2), are noticed by several recent studies, we provide for the first time a unified computational and theoretical approach for ultra high-dimensional envelope estimation.

It is also worth mentioning that envelope methods in multivariate linear regression are regarded as a likelihood-based alternative to the partial least squares (PLS) regression: Cook, Helland and Su (2013a) showed that the sequence of Krylov subspaces in PLS algorithm (De Jong, 1993; Helland, 1990) converges in population to the same envelope subspace. This connection between envelopes and partial least squares has promoted recent progress in high-dimensional studies of envelopes and PLS (Cook et al., 2019; Zhu and Su, 2020). Our method can thus also be viewed as an extension of the sparse PLS methods (Chun and Keleş, 2010; Chun et al., 2011) because it applies to a wide range of multivariate analysis problems beyond linear regression.

1.4. Notation and organization

For any matrix $\mathbf{M}$, we will let $\lambda_j(\mathbf{M})$ be the j-th eigenvalue of $\mathbf{M}$, e.g. $\lambda_1(\mathbf{M})$ is the largest eigenvalue of $\mathbf{M}$. We use the Frobenius norm, operator norm, ℓ_1 norm and the maximum norm of matrices. For a matrix $\mathbf{\Omega} \in \mathbb{R}^{p_1 \times p_2}$, let ω_{ij} denote its (i,j)-th element, then its Frobenius norm is $\|\mathbf{\Omega}\|_F = \sqrt{\sum_{i,j} \omega_{ij}^2}$; its operator norm $\|\mathbf{\Omega}\|_{op}$ is its largest eigenvalue; its ℓ_1 norm is $\|\mathbf{\Omega}\|_1 = \max_j \sum_i |\omega_{ij}|$; and its maximum norm is $\|\mathbf{\Omega}\|_{\max} = \max_{i,j} |\omega_{ij}|$.

We will study the estimation error of envelopes, characterized by the distance between two subspaces $\mathcal{E} = \text{span}(\mathbf{\Gamma})$ and $\widehat{\mathcal{E}} = \text{span}(\widehat{\mathbf{\Gamma}})$ For $\mathbf{\Gamma}, \widehat{\mathbf{\Gamma}} \in \mathbb{R}^{p \times u}$ such that $\mathbf{\Gamma}^\mathsf{T}\mathbf{\Gamma} = \mathbf{I}_u = \widehat{\mathbf{\Gamma}}^\mathsf{T}\widehat{\mathbf{\Gamma}}$, we let $\boldsymbol{\gamma}_i, \widehat{\boldsymbol{\gamma}}_i \in \mathbb{R}^p$, $i = 1, \ldots, u$, be the i-th columns of $\mathbf{\Gamma}$ and $\widehat{\mathbf{\Gamma}}$, respectively. Similar to Yu, Wang and Samworth (2014), we define the $u \times u$ diagonal matrix $\mathbf{\Theta}(\widehat{\mathbf{\Gamma}}, \mathbf{\Gamma})$ such that the j-th principal angle between the two subspace $\text{span}(\mathbf{\Gamma})$ and $\text{span}(\widehat{\mathbf{\Gamma}})$ is at the j-th diagonal element of $\mathbf{\Theta}(\widehat{\mathbf{\Gamma}}, \mathbf{\Gamma})$, and let $\sin \mathbf{\Theta}(\widehat{\mathbf{\Gamma}}, \mathbf{\Gamma}) \in \mathbb{R}^{u \times u}$ be defined entrywise. The principal angles between the two subspace $\text{span}(\mathbf{\Gamma})$ and $\text{span}(\widehat{\mathbf{\Gamma}})$ are denoted as θ_j, $j = 1, \ldots, u$, then $\cos(\theta_j) = \sqrt{1 - \sin^2(\theta_j)} = \lambda_j(\mathbf{\Gamma}^\mathsf{T}\widehat{\mathbf{\Gamma}})$, which is the j-th eigenvalue of $\mathbf{\Gamma}^\mathsf{T}\widehat{\mathbf{\Gamma}}$. In sufficient dimension reduction literature, distance between two subspace $\mathcal{E} = \text{span}(\mathbf{\Gamma})$ and $\widehat{\mathcal{E}} = \text{span}(\widehat{\mathbf{\Gamma}})$ are also commonly characterized as $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F$ instead of $\sin \mathbf{\Theta}(\widehat{\mathbf{\Gamma}}, \mathbf{\Gamma})$. A simple (and somewhat well-known) equivalence is provided in the following.

Lemma 1. $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F = \sqrt{2} \|\sin \mathbf{\Theta}(\widehat{\mathbf{\Gamma}}, \mathbf{\Gamma})\|_F$.

Unless otherwise specified, we will be using $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F$ as the natural measure for the distance between the true and estimated envelope subspaces.

The rest of the article is organized as follows. Section 2 introduces the proposed Non-Iterative Envelope Component Estimation (NIECE) and a general approach to connect the subspace estimation error with the newly introduced "envelope scores". The general theory is applied to the multivariate linear model in Sect. 2.3, where we allow both the numbers of predictors and responses to diverge with the sample size. Section 3 provides the sparse NIECE for high-dimensional data analysis, using the penalized matrix decomposition (PMD, Witten, Tibshirani and Hastie, 2009). The general theory for the sparse NIECE is established first, including the high-dimensional consistency result of PMD, and then applied to the linear and generalized linear models and the Cox model. Section 4 discusses some practical considerations of using NIECE in high-dimensional regression as an alternative to principal component regression. Simulations and real data analysis are presented in Sects. 5 and 6, followed by a brief discussion in Sect. 7. The Appendix contains additional numerical studies and a discussion on generalizing NIECE to the situation where $\mathbf{M}$ has common eigenvalues. Finally, the Supplementary Materials contains all technical details and additional numerical results.

2. Non-iterative envelope component estimation

2.1. Population algorithm and envelope scores

We propose the following NIECE procedure in population. The sample algorithm is readily available by replacing $\mathbf{M}$ and $\mathbf{U}$ with their sample counterparts and is shown to be consistent in terms of envelope subspace estimation when p grows at a relatively slow rate of the sample size n.

1. Input: symmetric $p \times p$ matrices $\mathbf{M} > 0$ and $\mathbf{U} \geq 0$; number of principal components d; envelope dimension u. Note that $0 \leq u \leq d \leq p$.
2. Obtain the first d eigenvectors of $\mathbf{M}$: $\mathbf{V}_d = (\mathbf{v}_1, \ldots, \mathbf{v}_d) \in \mathbb{R}^{p \times d}$ that satisfies $\mathbf{V}_d^\mathsf{T} \mathbf{V}_d = \mathbf{I}_d$. The corresponding d eigenvalues are $\lambda_1(\mathbf{M}) \equiv \mathbf{v}_1^\mathsf{T} \mathbf{M} \mathbf{v}_1 > \cdots > \lambda_d(\mathbf{M}) \equiv \mathbf{v}_d^\mathsf{T} \mathbf{M} \mathbf{v}_d > 0$.
3. Calculate the *envelope scores*: $\phi_j \equiv \mathbf{v}_j^\mathsf{T} \mathbf{U} \mathbf{v}_j$ for $j = 1, \ldots, d$, and organize them in descending order $\phi_{(1)} \geq \cdots \geq \phi_{(d)}$ and define $\mathbf{v}_{(j)}$ such that $\phi_{(j)} = \mathbf{v}_{(j)}^\mathsf{T} \mathbf{U} \mathbf{v}_{(j)}$.
4. Output: envelope is $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) = \operatorname{span}(\mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(u)})$.

Formally, we provide the following fundamental property of envelopes and, more importantly, the newly introduced envelope scores.

Lemma 2. *Suppose that* $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) \subseteq \operatorname{span}(\mathbf{V}_d)$ *for some* $d \leq p$ *and that* $\lambda_j(\mathbf{M}) \neq \lambda_k(\mathbf{M})$ *for any* $j, k \in \{1, \ldots, d\}$ *and* $j \neq k$. *Then* (2) *holds, the envelope can be written as* $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) = \operatorname{span}(\mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(u)})$, *and the envelope scores satisfy* $\phi_{(1)} \geq \cdots \geq \phi_{(u)} > \phi_{(u+1)} = \cdots = \phi_{(d)} = 0$.

Lemma 2 is a consequence of Proposition 2.2 in Cook, Li and Chiaromonte (2010), but it allows easy construction of envelope estimation in a non-iterative manner. We make the following remarks regarding the assumptions in Lemma 2.

First of all, the assumption that $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) \subseteq \operatorname{span}(\mathbf{V}_d)$ for some $d \leq p$ is trivially true if we take $d = p$, which is a reasonable choice in low-dimensional settings. However, when $p > n$, we have to make such an assumption for some $d < n$; otherwise, the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ is not estimable. Intuitively, this is because there is only a limited number of eigenvectors of $\mathbf{M}$ estimable in high dimensions. In all our theoretical studies, we treat d as fixed for simplicity and allow both $n, p \to \infty$.

Second, the distinct eigenvalue assumption that $\lambda_j(\mathbf{M}) \neq \lambda_k(\mathbf{M})$ for any $j, k \in \{1, \ldots, d\}$ and $j \neq k$ is to ensure the identifiability of each eigenvectors. Moreover, this assumption is crucial to the high-dimensional theoretical analysis of the penalized matrix decomposition (PMD, Witten, Tibshirani and Hastie, 2009) method, which sequentially obtains the sparse estimates for the eigenvectors and is adopted in our sparse NIECE approach. The distinct eigenvalue assumption may be relaxed if we estimate the principal subspaces (e.g., Vu and Lei, 2013) instead of vectors. We define the minimal eigen-gap of the first d eigenvalues of $\mathbf{M}$ as follows,

$$0 < \Delta = \min_{j=1,\ldots,d} \{\lambda_j(\mathbf{M}) - \lambda_{j+1}(\mathbf{M})\}. \tag{3}$$

Unless otherwise noted, we will make the assumptions as in Lemma 2. The envelope is thus constructed by the eigenvectors $\mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(u)}$. Let $\pi(\cdot) : \{1, \ldots, d\} \mapsto \{1, \ldots, d\}$ be the permutation of indices according the envelope scores (cf. Step 3 of the NIECE algorithm), so that $\mathbf{v}_{(j)}$ is $\pi(j)$-th eigenvector of $\mathbf{M}$ and $\lambda_{\pi(j)}(\mathbf{M}) = \mathbf{v}_{(j)}^{\mathsf{T}} \mathbf{M} \mathbf{v}_{(j)}$.

Finally, we note that the envelope scores, which satisfy $\phi_{(1)} \geq \cdots \geq \phi_{(u)} > \phi_{(u+1)} = \cdots = \phi_{(d)} = 0$, do not need to be distinct from each other. We only need a gap between the u-th and $(u+1)$-th envelope scores. In our theoretical analysis, the following envelope score gap plays a crucial role

$$\Delta_{\mathbf{U}} \equiv \phi_{(u)} - \phi_{(u+1)} = \mathbf{v}_{(u)}^{\mathsf{T}} \mathbf{U} \mathbf{v}_{(u)} > 0. \tag{4}$$

In the next section, we show how the minimal eigen-gap Δ and the envelope score gap $\Delta_{\mathbf{U}}$ would affect the subspace estimation in finite sample.

For the sample NIECE procedure, we simply replace $\mathbf{M}$ and $\mathbf{U}$ with their sample counterparts (e.g. sample covariance matrices). We let $\widehat{\mathbf{v}}_{(j)}$, $j = 1, \ldots, u$, be the $\widehat{\pi}(j)$-th eigenvectors of $\widehat{\mathbf{M}}$, where $\widehat{\pi}(\cdot)$ is the permutation of indices in $\{1, \ldots, d\}$ according the the estimated envelope scores $\widehat{\phi}_j = \widehat{\mathbf{v}}_j^{\mathsf{T}} \widehat{\mathbf{U}} \widehat{\mathbf{v}}_j$ for $j = 1, \ldots, d$. One critical condition that is necessary for the consistency of the envelope estimation is that the index set $\{\widehat{\pi}(j) \mid j = 1, \ldots, u\}$ converges to the true set $\{\pi(j) \mid j = 1, \ldots, u\}$. In fact, we prove a slightly stronger condition $\widehat{\pi}(j) \to \pi(j)$ for all $j = 1, \ldots, u$ in our theoretical studies. This requires accurate envelope scores estimation. The following proposition summarizes what aspects of $\widehat{\mathbf{v}}_i$ are important to ensure the accurate envelope scores estimation.

Proposition 1. *For $i = 1, \ldots, d$, we have the following conclusions:*

1. *If $\|\widehat{\mathbf{U}} - \mathbf{U}\|_{op} \leq \epsilon$, $\epsilon \leq \Delta/4$, $\sin \Theta(\widehat{\mathbf{v}}_i, \mathbf{v}_i) \leq \frac{2\epsilon}{\Delta}$, and and $\|\widehat{\mathbf{v}}_i\|_2 = 1$, then*

$$|\widehat{\phi}_i - \phi_i| \leq \left(1 + \frac{10\nu}{\Delta}\right) \epsilon, \tag{5}$$

   where $\nu = \|\mathbf{U}\|_{op}$ is the largest eigenvalue of $\mathbf{U} \geq 0$.
2. *If $\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} \leq \epsilon$, $\sin^2 \Theta(\widehat{\mathbf{v}}_i, \mathbf{v}_i) \leq c_0 \tau^2 \epsilon$, and $\|\widehat{\mathbf{v}}_i\|_1 \leq \tau$, then*

$$|\widehat{\phi}_i - \phi_i| \leq \left\{(c_0 + 1)\tau^2 + \frac{2\nu}{\Delta}\right\} \epsilon. \tag{6}$$

The first part of Proposition 1 states the relationship between the accuracy of sample eigenvectors and the accuracy of sample envelope scores. The accuracy of sample eigenvector $\sin \Theta(\widehat{\mathbf{v}}_i, \mathbf{v}_i) \leq 2\epsilon/\Delta$, where $\Delta > 0$ is the eigenvalue gap, is a direct consequence of Corollary 1 in Yu, Wang and Samworth (2014). This part of Proposition 1 leads to the estimation results of the sample NIECE algorithm in Theorem 1. The second part of Proposition 1 is useful for high-dimensional sparse settings: Theorem 1 for sparse NIECE based on penalized matrix decomposition. In high-dimensional sparse settings, we only need to modify Step 2 of the algorithm, by adopting the penalized eigen-decomposition or penalized principal component analysis.

2.2. Theory for the sample NIECE algorithm

We first present a generic theory to characterize how the envelope subspace estimation error is affected by: (i) the envelope dimension; (ii) the estimation errors, measured in operator norms, of sample matrices $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{M}}$; (iii) the minimal eigen-gap Δ defined in (3); (iv) the envelope score gap $\Delta_{\mathbf{U}}$ defined in (4); and (v) the largest eigenvalue of $\mathbf{U}$ denoted as $\nu = \|\mathbf{U}\|_{op}$. The following theory is for the analysis in relatively low-dimensional data, and it is not easy to find estimates satisfying (7) in high dimensions.

Theorem 1. *For any* $\epsilon > 0$ *and* $\epsilon \leq \max\left\{\frac{\Delta}{4}, \frac{\Delta_{\mathbf{U}}}{2}\left(1+\frac{10\nu}{\Delta}\right)^{-1}\right\}$, *if we have*

$$\|\widehat{\mathbf{M}} - \mathbf{M}\|_{op} \leq \epsilon, \quad \|\widehat{\mathbf{U}} - \mathbf{U}\|_{op} \leq \epsilon, \tag{7}$$

then $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F \leq \frac{2\sqrt{2u}}{\Delta}\epsilon$, *where* $\widehat{\mathcal{E}}$ *and* $\mathcal{E}$ *are the envelope* $\widehat{\mathcal{E}}_{\mathbf{M}}(\mathbf{U})$ *and* $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$, *respectively.*

In Theorem 1 and all our theoretical analysis later, the envelope subspace estimation error is measured by $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F$, which is bounded between 0 and $\sqrt{2u}$ by definition. Therefore, it is natural to see $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F$ proportional to $\sqrt{2u}$ in our results. From Theorem 1, the subspace estimation error is proportional to the estimation error of the matrices $\mathbf{M}$ and $\widehat{\mathbf{U}}$, which is indicated by ϵ; and is inversely proportional to the minimal eigen-gap Δ, which directly affects the estimation of eigenvectors $\mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(u)}$ that lie within the envelope. On the other hand, the envelope score gap $\Delta_{\mathbf{U}}$ together ν, which indicates the magnitude of $\mathbf{U}$, affects the envelope subspace estimation error indirectly because $\widehat{\mathbf{U}}$ is only used in Step 3 of NIECE procedure to determine the index ordering. When the matrices $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ are close enough to their population counterparts, i.e. $\epsilon \leq \frac{\Delta_{\mathbf{U}}}{2}\left(1+\frac{10\nu}{\Delta}\right)^{-1}$, the estimated index set $\{\widehat{\pi}(j) \mid j = 1, \ldots, u\}$ converges to the true set $\{\pi(j) \mid j = 1, \ldots, u\}$.

In the Supplementary Materials, Lemma 4, we consider the ideal setting where $\{\widehat{\pi}(j) \mid j = 1, \ldots, u\} = \{\pi(j) \mid j = 1, \ldots, u\}$. Then we may replace the minimal eigen-gap Δ of the d eigenvalues of $\mathbf{M}$ to the following quantity that only involves $u \leq d$ eigenvalues. For simplicity, suppose the eigenvalues $\lambda_{\pi(j)}$'s, $j = 1, \ldots, u$, are all distinct, then we may re-define $\Delta = \min_{j=1,\ldots,u} \Delta_j$, where $\Delta_j = \min\{\lambda_{\pi(j)-1}(\mathbf{M}) - \lambda_{\pi(j)}(\mathbf{M}),\ \lambda_{\pi(j)}(\mathbf{M}) - \lambda_{\pi(j)+1}(\mathbf{M})\}$ and $\lambda_{-1} \equiv -\infty$, $\lambda_{p+1} \equiv \infty$. In the Supplementary Materials, we show that the similar results of Theorem 1 still holds by re-defining Δ, even when some of these $\lambda_{\pi(j)}$'s are not distinct.

To apply the sample NIECE in different contexts, we only need to calculate the probability of (7) to verify the consistency of envelope subspace estimation. This theory allows us to study the un-penalized envelopes with (slowly) diverging dimensions. We next demonstrate that such results hold for the simultaneous predictor and response reduction.

2.3. Simultaneous envelopes in multivariate linear regression

To illustrate the application of Theorem 1, we consider the simultaneous envelopes model (Cook and Zhang, 2015a). The simultaneous envelopes are constructed, by jointly estimating the response envelope (Cook, Li and Chiaromonte, 2010) and the predictor envelope (Cook, Helland and Su, 2013a), to simultaneous reduce the multivariate response $\mathbf{Y} \in \mathbb{R}^r$ and predictor $\mathbf{X} \in \mathbb{R}^p$ in the following regression model,

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}, \tag{8}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^r$ is the intercept vector, $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ is the regression coefficient matrix, and $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ is independent of $\mathbf{X}$.

In the classical likelihood-based envelope methods, the response is reduced by projecting onto the response envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$, whose likelihood-based estimation is derived from the normal error $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$. On the other hand, the predictor reduction is achieved by projection onto the predictor envelope $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}^\mathsf{T})$, whose likelihood-based estimation is derived from further assuming $\mathbf{X} \sim N(0, \boldsymbol{\Sigma}_\mathbf{X})$. Thus, for predictor reduction and simultaneous reduction, the joint normality of $(\mathbf{X}^\mathsf{T}, \mathbf{Y}^\mathsf{T})^\mathsf{T}$ is required. We consider the following sub-Gaussian tail distribution assumption that is weaker than the joint normality assumption.

We assume that there exists $\sigma > 0$ such that for any $\mathbf{w} \in \mathbb{R}^{p+r}$, $\|\mathbf{w}\|_2 = 1$, we have

$$P\left(\left|\mathbf{w}^\mathsf{T}\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{\sigma^2}\right). \tag{9}$$

While alternating updates obtain the likelihood-based simultaneous envelopes estimators, we apply the NIECE to the predictor and response envelopes separately. For response envelope, note that $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) = \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{Y}}(\boldsymbol{\Sigma}_{\mathbf{YX}})$, where $\boldsymbol{\Sigma}_\mathbf{Y} \in \mathbb{R}^{r\times r}$ is the covariance matrix of $\mathbf{Y}$ and $\boldsymbol{\Sigma}_{\mathbf{YX}} \equiv \boldsymbol{\Sigma}_{\mathbf{XY}}^\mathsf{T} \in \mathbb{R}^{r\times p}$ is the cross-covariance matrix. For predictor envelope, we exploit the symmetry by noticing $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}^\mathsf{T}) = \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\Sigma}_{\mathbf{XY}})$. Given the data, the NIECE estimator is obtained by using the sample covariances $\widehat{\mathbf{M}} = \widehat{\boldsymbol{\Sigma}}_\mathbf{Y}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}$ for response reduction, and $\widehat{\mathbf{M}} = \widehat{\boldsymbol{\Sigma}}_\mathbf{X}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}}$ for predictor reduction.

We have the following result that applies to both response and predictor envelopes (i.e. simultaneous envelopes), where $\mathcal{E}_\mathbf{M}(\mathbf{U}) = \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{Y}}(\boldsymbol{\Sigma}_{\mathbf{YX}}) \subseteq \mathbb{R}^r$ for response envelope and is re-defined as $\mathcal{E}_\mathbf{M}(\mathbf{U}) = \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\Sigma}_{\mathbf{XY}}) \subseteq \mathbb{R}^p$ for predictor envelope. Other quantities such as $\widehat{\mathcal{E}}$, Δ, $\Delta_\mathbf{U}$, u and ν also alters when we switch from response envelope to predictor envelope.

Theorem 2. *For any $\epsilon > 0$ such that $\epsilon \leq \max\left\{\frac{\Delta}{4}, \frac{\Delta_\mathbf{U}}{2}\left(1 + \frac{10\nu}{\Delta}\right)^{-1}\right\}$ and $\epsilon \leq \|\boldsymbol{\Sigma}_{\mathbf{XY}}\|_1 + \|\boldsymbol{\Sigma}_{\mathbf{YX}}\|_1 < c_1$ for some constant c_1, we have*

$$\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_\mathcal{E}\|_F \leq \frac{2\sqrt{2u}\epsilon}{\Delta}, \tag{10}$$

with a probability greater than $1 - Cr^2\exp(-C\frac{n\epsilon^2}{r^2}) - Cpr\exp(-C\frac{n\epsilon^2}{r^2}) - Cpr\exp(-C\frac{n\epsilon^2}{p^2})$.

To the best of our knowledge, the above theorem is the first result in establishing non-asymptotic properties of any envelope estimators without sparsity. It also leads to the first known consistency result of envelope subspace estimation with diverging n, p, and r. Specifically, we have the following Corollary by letting $\epsilon = pr\sqrt{(\log p + \log r)/n} \to 0$.

Corollary 1. *Suppose* $\max\left(pr\sqrt{\log p + \log r}, pr\Delta_{\mathbf{U}}^{-1}\sqrt{\log p + \log r}\right) = o(\sqrt{n})$ *and* $\|\mathbf{\Sigma}_{\mathbf{XY}}\|_1 + \|\mathbf{\Sigma}_{\mathbf{YX}}\|_1 < c_1$ *for some constant* c_1*, then* $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F \to 0$ *in probability as* $n, p, r \to \infty$.

The above result holds for both predictor and response envelopes. We have thus established the consistency of the simultaneous envelopes. Analogous to Theorem 2 and Corollary 1, it is straightforward to show similar results for other types of envelopes by applying Theorem 1. We omit details for such extensions and study some of them (envelopes in the generalized linear model and Cox model) in the more challenging high-dimensional settings in the following sections.

2.4. Comparing envelope algorithms

We first compare NIECE with two fast envelope algorithms, CFS (Cook, Forzani and Su, 2016) and ECD (Cook and Zhang, 2018) that are both recently proposed state-of-the-art algorithms and were shown to be much faster and more stable than the full Grassmannian optimization (Edelman, Arias and Smith, 1998; Absil, Mahony and Sepulchre, 2009) and sequential 1D algorithm (Cook and Zhang, 2016) in the literature. We consider the moderately high-dimensions, where $(n, p, u) = (200, 100, 5)$, because the CFS and ECD are not applicable to high-dimensional settings. We set $\mathbf{M} = \mathbf{\Gamma\Omega\Gamma}^{\mathsf{T}} + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^{\mathsf{T}}$, $\mathbf{U} = \Delta_{\mathbf{U}} \cdot \mathbf{\Gamma\Phi\Gamma}^{\mathsf{T}}$, where $\mathbf{\Gamma} = (\mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_{10}, \mathbf{v}_{11}, \mathbf{v}_{19})$ and $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ are diagonal matrix of eigenvalues $\lambda_k = k^3$ for $k = 1, \ldots, 20$ and $\lambda_k = 0.05$ for $k > 20$, $\mathbf{\Phi} = \mathbf{O}_u\mathbf{D}\mathbf{O}_u^{\mathsf{T}}$ with orthonormal matrix $\mathbf{O}_u \in \mathbb{R}^{u\times u}$ and $\mathbf{D} = \text{diag}(1, \ldots, u)$. We tried three different signal strength settings with $\Delta_{\mathbf{U}} = 0.01$, 1, and 100. For each setting, we randomly generated 100 replicates of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ from Wishart distributions with degrees of freedom n and means $\mathbf{M}$ and $\mathbf{U}$ to mimic the sample covariances in regression models. We reported the subspace estimation error $\mathcal{D}(\widehat{\mathbf{\Gamma}}, \mathbf{\Gamma}) = \mathcal{D}(\widehat{\mathcal{E}}, \mathcal{E}) = \|\mathbf{P}_{\mathcal{E}} - \mathbf{P}_{\widehat{\mathcal{E}}}\|_F/\sqrt{2u}$, which is a number between 0 and 1, and the logarithm of CPU time.

Figure 2 shows the advantages of NIECE in both subspace estimation accuracy and computational speed. This numerical study is an illustration of our theoretical analysis in the previous section. When the eigenvalue gap Δ in $\mathbf{M}$ is large enough, the NIECE procedure is very fast (due to the computational advantage of SVD in NIECE versus manifold optimizations in others) and also very accurate for a wide range of signal strengths measured by $\Delta_{\mathbf{U}}$.

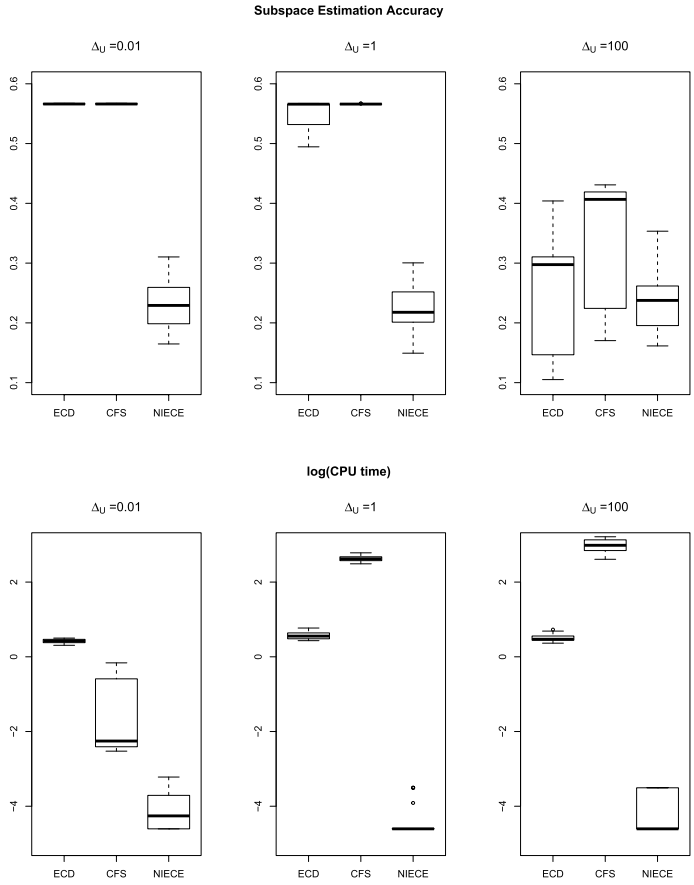

FIG 2. *Advantages of NIECE over the state-of-the-art envelope algorithms in both accuracy (left three panels; smaller value $\mathcal{D}$ indicates more accurate subspace estimation) and time (right three panels; logarithm of CPU time in seconds), and across a wide range of signal strengths: weak signal (left, $\Delta_{\mathbf{U}} = 0.01$), moderate signal (middle, $\Delta_{\mathbf{U}} = 1$), and strong signal (right, $\Delta_{\mathbf{U}} = 100$).*

3. NIECE for high-dimensional data analysis

3.1. High-dimensional sparse NIECE

In high-dimensional sparse envelope settings, we only need to modify Step 2 of NIECE sample algorithm, by replacing the eigen-decomposition of sample matrix $\widehat{\mathbf{M}}$ with the penalized eigen-decomposition or penalized principal component analysis methods that are more suitable for such scenarios. Without loss of generality, we assume $d < n$ and obtain the d sparse eigenvectors. For almost all envelope problems, we can write $\widehat{\mathbf{M}} = \mathbf{X}_n^{\mathsf{T}}\mathbf{X}_n$ from some data matrix $\mathbf{X}_n \in \mathbb{R}^{n\times p}$. As we have seen from the previous Section, the data matrix $\mathbf{X}_n$ can

be either the n samples of $\mathbf{X}$ or $\mathbf{Y}$. For envelopes in GLM and Cox model, $\mathbf{X}_n$ is also the n samples of $\mathbf{X}$. Another example is the weighted least squares envelope (Cook and Zhang, 2015b), where the i-th row of $\mathbf{X}_n$ is the squared-root weight $\sqrt{W_i}$ times the predictor vector $\mathbf{X}_i$.

For simplicity, the data matrix $\mathbf{X}_n \in \mathbb{R}^{n\times p}$ is henceforth defined as the n i.i.d. samples of mean zero random variable $\mathbf{X} \in \mathbb{R}^p$. We obtain the d sparse eigenvectors of $\widehat{\mathbf{M}} \geq 0$ based on the penalized matrix decomposition (PMD) method from Witten, Tibshirani and Hastie (2009). Because $n \ll p$ is very common in high-dimensional data analysis, we adopt the special type of PMD named PMD$(\cdot, L_1)$ on $\mathbf{X}_n$ instead of the much bigger matrix $\widehat{\mathbf{M}}$ as follows.

Let $\mathbf{X}^1 = \mathbf{X}_n$ be the original data matrix. For $k = 1, \ldots, d$, we sequentially solve for

$$(\widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k) = \underset{\mathbf{u}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^p}{\operatorname{argmax}} \mathbf{u}^{\mathsf{T}}\mathbf{X}^k\mathbf{v} \quad \text{subject to } \|\mathbf{v}\|_1 \leq c,\ \|\mathbf{v}\|_2^2 \leq 1,\ \|\mathbf{u}\|_2^2 \leq 1; \tag{11}$$

and then deflate the data matrix as $\mathbf{X}^{k+1} = \mathbf{X}^k - \widehat{\sigma}_k\widehat{\mathbf{u}}_k\widehat{\mathbf{v}}_k^{\mathsf{T}}$, where $\widehat{\sigma}_k = \widehat{\mathbf{u}}_k^{\mathsf{T}}\mathbf{X}^k\widehat{\mathbf{v}}_k$.

To solve (11), we need alternating updates between $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^p$. Since there is no penalty on $\mathbf{u}$, we can see that the solution for $\widehat{\mathbf{u}}_k$ has to be

$$\widehat{\mathbf{u}}_k = \frac{\mathbf{X}^k\widehat{\mathbf{v}}_k}{\|\mathbf{X}^k\widehat{\mathbf{v}}_k\|_2}, \tag{12}$$

which leads to the following equivalent presentation.

Lemma 3. *The optimization in* (11) *is equivalent to* (12) *and*

$$\widehat{\mathbf{v}}_k = \arg\max_{\mathbf{v}} \mathbf{v}^T\widehat{\mathbf{M}}_k\mathbf{v} \quad \text{subject to } \|\mathbf{v}\|_1 \leq \tau \text{ and } \|\mathbf{v}\|_2^2 \leq 1, \tag{13}$$

where $\widehat{\mathbf{M}}_k = (\mathbf{X}^k)^T\mathbf{X}^k$ *and* $\mathbf{X}^k = \mathbf{X}^{k-1}(\mathbf{I}_p - \widehat{\mathbf{v}}_{k-1}\widehat{\mathbf{v}}_{k-1}^T)$.

As noted in Witten, Tibshirani and Hastie (2009), by the Karush-Kuhn-Tucker conditions in convex optimization, the solution to (13) is also the solution to

$$\widehat{\mathbf{v}}_k = \arg\max_{\mathbf{v}} \mathbf{v}^{\mathsf{T}}\widehat{\mathbf{M}}_k\mathbf{v} \quad \text{subject to } \|\mathbf{v}\|_1 \leq c, \text{ and } \|\mathbf{v}\|_2^2 = 1, \tag{14}$$

if c is chosen so that the maximizer of $\mathbf{v}^{\mathsf{T}}\widehat{\mathbf{M}}_k\mathbf{v}$ subject to only one constraint $\|\mathbf{v}\|_1 \leq c$ has L_2 norm greater than or equals to 1.

3.2. A general high-dimensional theory

Notice that we have re-defined $\widehat{\mathbf{v}}_k$ as the PMD solution for the high-dimensional sparse NIECE. Moreover, we re-define $\widehat{\mathbf{V}}_d = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_d)$ and other estimates for ease of presentation. However, it is important to note that $\widehat{\mathbf{v}}_k$ is no longer the eigenvector of $\widehat{\mathbf{M}}$ and that orthogonality among eigenvectors no longer holds, i.e. $\widehat{\mathbf{V}}_d^{\mathsf{T}}\widehat{\mathbf{V}}_d \neq \mathbf{I}_d$. Furthermore, $\widehat{\sigma}_k^2 = \widehat{\mathbf{v}}_k^{\mathsf{T}}\widehat{\mathbf{M}}\widehat{\mathbf{v}}_k$ is different from the eigenvalue $\lambda_k(\widehat{\mathbf{M}})$. Due to the non-orthogonality among $\widehat{\mathbf{v}}_k$'s and related issues, the deflation of data matrix is not easy to handle. Nevertheless, we are able to show

consistency of $\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_d$ by analyzing $\mathbf{X}^k = \mathbf{X}_n \prod_{j=1}^{k-1}(\mathbf{I}_d - \widehat{\mathbf{v}}_j \widehat{\mathbf{v}}_j^{\mathsf{T}})$ based on Lemma 3. Because we assume that the eigenvalues are distinct, PMD offers a computationally efficient way to estimate the eigenvectors. If we use principal subspace estimation (e.g., Cai et al., 2013), the rate may potentially be improved but with higher computation costs.

We first show a general theory of PMD, which is of independent interest. The sparsity in the true eigenvectors $\mathbf{v}_k$ (e.g. sparsity in principal component loadings) is imposed as follows.

$$\tau_0 = \max_{1 \le k \le d} \|\mathbf{v}_k\|_1, \tag{15}$$

where a small value of τ_0 implies that the first d eigenvectors of $\mathbf{M}$ are all reasonably sparse in high dimensions.

Theorem 3. *If $\tau > \tau_0$ and $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \le \epsilon$, then there exists a constant $c_0 > 0$ that does not depend on n, p or ϵ, such that,*

$$\sin^2 \Theta(\mathbf{v}_k, \widehat{\mathbf{v}}_k) \le c_0 \epsilon \tau^2, \quad k = 1, \ldots, d. \tag{16}$$

Hence, to show the consistency of NIECE in high-dimensional settings, we only need to verify $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \le \epsilon$, which is weaker than the assumption of $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{op} \le \epsilon$ in the non-sparse settings. This is because we apply regularization to obtain accurate estimates of $\mathbf{v}_k$ in high dimensions. Also, note that $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{op} \le \epsilon$ generally cannot hold in high dimensions, while later we will show that $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \le \epsilon$ holds with a probability tending to 1 under mild conditions.

Theorem 4. *Assume that $\tau > \tau_0$, $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \le \epsilon$ and $\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} \le \epsilon$, then if $\Delta_{\mathbf{U}} > \sqrt{8c_0\nu^2\tau^2\epsilon} + (2c_0\nu + 1)\tau^2\epsilon$ with constant c_0 defined in (16), then there exists a constant $C > 0$ that does not depend on n, p or ϵ such that $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F^2 \le C\epsilon\tau^2$.*

Theorem 4 is a general result that guarantees that as long as we start with a reasonably accurate estimator $\widehat{\mathbf{M}}$, we can estimate the envelope with the specified error bound. Next, we demonstrate that Theorem 4 leads to the consistency of sparse NIECE under several important envelope models.

3.3. Envelope in multivariate linear model

In this section, we consider the multivariate linear model (8) with jointly sub-Gaussian data (9). To avoid redundancy, we only consider the response envelope. The predictor envelope in linear model is analogous to the response envelope and similar to the predictor envelope in the generalized linear model in the next section.

Recall that the NIECE estimator for response envelope is obtained by using the sample covariances $\widehat{\mathbf{M}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}$. In high-dimensional sparse setting, e.g. the eigenvectors of $\mathbf{M} = \boldsymbol{\Sigma}_{\mathbf{Y}}$ satisfies (15), we have the following results.

Theorem 5. *For any ϵ such that $0 < \epsilon \leq 1/p$, assume that $\tau > \tau_0$ and $\Delta_{\mathbf{U}} > \sqrt{8c_0\nu^2\tau^2\epsilon} + (2c_0\nu + 1)\tau^2\epsilon$ with constant c_0 defined in (16), then there exists a constant $C > 0$ that does not depend on n, r, p or ϵ such that*

$$\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F \leq C\epsilon\tau^2 \tag{17}$$

with a probability greater than $1 - Cr^2\exp(-Cn\epsilon^2) - Cpr\exp(-Cn\epsilon^2)$.

The above non-asymptotic result leads to the following consistency of sparse NIECE.

Corollary 2. *If $\sqrt{\log p + \log r} = o(\sqrt{n})$, $\sqrt{\log p + \log r} < \sqrt{n}/p$ and $\Delta_{\mathbf{U}}$ and τ satisfy that*

$$\max\left(\tau^2\sqrt{\log p + \log r},\ \Delta_{\mathbf{U}}^{-2}\tau^2\sqrt{\log p + \log r}\right) = o(\sqrt{n}),$$

then we have $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F \to 0$ in probability as $n, p, r \to \infty$.

The requirement that $\sqrt{\log p + \log r} < \sqrt{n}/p$ implies that the number of predictors can not diverge quickly (p grow slower than $\sqrt{n}$), while the response envelope allows the number of responses to diverge much faster. To see this clearly, we let p fixed in the following Corollary.

Corollary 3. *If $\log r = o(n)$, and $\Delta_{\mathbf{U}}$ and τ satisfy that $\max(\tau^2\sqrt{\log r}, \Delta_{\mathbf{U}}^{-2}\tau^2\sqrt{\log r}) = o(\sqrt{n})$, we have $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F \to 0$ in probability as $n, r \to \infty$.*

This justifies applying the NIECE approach for response envelope reduction in ultra high-dimensional settings, i.e. $(\log r)/n \to 0$. Similarly, for the predictor envelope in the linear model, if we fix the number of responses, consistency is achieved when $(\log p)/n \to 0$.

3.4. Envelope in generalized linear model

Cook and Zhang (2015b) extended the envelope model from multivariate linear regression context to generalized linear model with canonical link functions, Cox model, and general multivariate parameter estimation problems. We consider the envelope generalized linear model and the envelope Cox model in this and the following sections. The sparse NIECE procedure extends these non-standard envelope models from low dimensional settings to high-dimensional.

We replace the joint sub-Gaussian tail distribution of $(\mathbf{X}^\mathsf{T}, \mathbf{Y}^\mathsf{T})^\mathsf{T}$ with the following sub-Gaussian tail assumption on $\mathbf{X}$. Similar to (9), this is weaker than assuming the normality and is widely used to prove the consistency of penalized methods (e.g., Negahban et al., 2012). Suppose that there exists $\sigma > 0$ such that for any $\mathbf{w} \in \mathbb{R}^p$, $\|\mathbf{w}\|_2 = 1$, we have

$$P(|\mathbf{w}^\mathsf{T}\mathbf{X}| \geq t) \leq 2\exp\left(-\frac{t^2}{\sigma^2}\right). \tag{18}$$

In the generalized linear model with canonical link functions, the response Y follows some exponential family distributions with probability density (or mass) function as

$$f(y \mid \vartheta, \varphi) = \exp\left(\frac{y\vartheta - b(\vartheta)}{a(\varphi)} + c(y, \varphi)\right),$$

where the canonical parameter ϑ is set to be $\vartheta = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}$ by the canonical link function, φ is the dispersion parameter, and functions a, b and c can be specified for different families of the distribution of Y. For simplicity, we focus on one-parameter family so that the dispersion parameter is not included (or, equivalently, φ is set to be 1). As such, we can write the log-likelihood as

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(Y_i \mid \boldsymbol{\beta}, \mathbf{X}_i) = \sum_{i=1}^{n} \{Y_i(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i) - b(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)\}, \tag{19}$$

where the function $b(t) = t^2/2$ for normally distributed Y; it is $\exp(t)$ for Poisson regression and is $\log\{1 + \exp(t)\}$ for Logistic regression. The envelope is $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}) \subseteq \mathbb{R}^p$ where $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is the regression coefficient vector from above.

To avoid redundancy, we only show the results for high-dimensional penalized Logistic regression where $Y = 0$ or 1, and $\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}\{Y_i(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i) - \log(1 + e^{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i})\}$. For normally distributed Y, the GLM reduces to the predictor envelope with a univariate response. We have shown that $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ is an accurate estimator for $\boldsymbol{\Sigma}_{\mathbf{X}}$. For $\boldsymbol{\beta}$, note that the sparsity of the envelope implies the sparsity of $\boldsymbol{\beta}$. We can obtain $\widehat{\boldsymbol{\beta}}$ by ℓ_1 logistic regression with a tuning parameter λ_n,

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}\{-Y_i(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i) + \log(1 + e^{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i})\} + \lambda_n\|\boldsymbol{\beta}\|_1, \tag{20}$$

which is based on minimizing the negative log-likelihood plus an ℓ_1-regularization term. Then the NIECE procedure is based on $\widehat{\mathbf{M}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\mathsf{T}}$.

Theorem 6. *If $\tau > \tau_0$, $\lambda_n = 4\epsilon/\sqrt{s}$, and $\Delta_{\mathbf{U}} > \sqrt{8c_0\nu^2\tau^2\epsilon} + (2c_0\nu + 1)\tau^2\epsilon$ with constant c_0 defined in (16), then there exists a constant $C > 0$ that does not depend on n, p or ϵ such that*

$$\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F \le C\tau^2\epsilon \tag{21}$$

with a probability greater than $1 - Cp^2\exp(-Cn\epsilon^2) - Cp\exp(-Cn\epsilon^2/s)$. Furthermore, if $s\log p = o(n)$ and that λ_n, τ and $\Delta_{\mathbf{U}}$ satisfy $\max\{\tau^2\sqrt{\frac{s\log p}{n}}, \Delta_{\mathbf{U}}^{-2}\tau^2\sqrt{\frac{s\log p}{n}}, \lambda_n^{-1}\sqrt{\frac{\log p}{n}}\} = o(1)$, we have $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F \to 0$ in probability as $n, p \to 0$.

Theorem 6 thus summarizes both the non-asymptotic and asymptotic results for the sparse NIECE procedure in logistic regression.

3.5. Envelope in Cox proportional hazards model

We consider the Cox regression (Cox, 1972) that is widely used in survival data analysis. The hazard function is assumed to be $h(t \mid \mathbf{X}) = h_0(t)\exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})$, where $h_0(t)$ is the baseline hazard function and $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is the regression coefficients of the p-dimensional covariate vector $\mathbf{X} \in \mathbb{R}^p$. Let Y and C be the failure time and censoring time, then we assume Y and C are conditionally independent given $\mathbf{X}$. Define $\delta = I(Y \le C)$ and $T = \min(Y, C)$, then the observed data consists of independent copies of $\{T_i, \delta_i, \mathbf{X}_i\}$, $i = 1, \ldots, n$.

Estimation of $\boldsymbol{\beta}$ is typically achieved by maximizing the Cox's partial likelihood (Cox, 1975), or equivalently by minimizing the following negative logarithm of that,

$$\ell_n(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\left[\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i - \log\left\{\sum_{j=1}^{n} I(T_j \ge T_i)\exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_j)\right\}\right]. \tag{22}$$

When the covariates in $\mathbf{X}$ are highly correlated with each other, Cook and Zhang (2015b) proposed to construct envelope estimator

$$\widehat{\boldsymbol{\beta}}_{\mathrm{Env}} = \arg\min_{\boldsymbol{\beta}=\boldsymbol{\Gamma}\boldsymbol{\eta}} \ell_n(\boldsymbol{\beta}) - \frac{1}{n}M_n(\boldsymbol{\Gamma}), \tag{23}$$

where $M_n(\boldsymbol{\Gamma}) = -\frac{n}{2}(\log|\boldsymbol{\Gamma}^{\mathsf{T}}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^{\mathsf{T}}\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}|)$ is the partially maximized log likelihood of $\mathbf{X} \sim N(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$ under the envelope structure $\mathrm{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$. It is shown to be more efficient than the partial least squares in Cox regression Nygård et al. (2008). However, the optimization is similar to (or even more challenging than) the Grassmannian optimization (1) discussed in the Introduction section and is thus not feasible in the high-dimensional Cox model.

In the high-dimensional setting, to incorporate variable selection, we consider the following ℓ_1-penalized maximum (partial) likelihood estimator,

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) + \lambda_n\|\boldsymbol{\beta}\|_1, \tag{24}$$

where $\lambda_n > 0$ is the tuning parameter. Similar to the envelope GLM, the sparsity of the envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$ implies the sparsity of $\boldsymbol{\beta}$. Then the NIECE procedure is based on $\widehat{\mathbf{M}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\mathsf{T}}$. The convergence of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ is guaranteed by sub-Gaussian distribution of $\mathbf{X}$, i.e. (18). Properties of $\widehat{\mathbf{U}}$ is based on the estimation consistency of the Lasso estimator $\widehat{\boldsymbol{\beta}}$ that is derived in Huang et al. (2013). Similar results for using the SCAD penalty (Fan and Li, 2001) instead of Lasso can be found in Bradic, Fan and Jiang (2011).

Following Huang et al. (2013), we consider the following two conditions.

(i) *Uniformly bounded covariates.* There exists some constant C such that the covariates are bounded as $\max_{i<i'\le n}\|\mathbf{X}_i - \mathbf{X}_{i'}\|_\infty \le C$.

(ii) *Restricted eigenvalue lower bound.* There exists positive constants t and M such that the smallest eigenvalue of the population matrix $\boldsymbol{\Sigma}(t, M)$ is greater than some constant ρ^*.

These conditions are derived from Theorem 4.1 of Huang et al. (2013), where the introduction of $\mathbf{\Sigma}(t, M)$ is rather technical and relegated to the Supplementary Materials. The lower bound ρ^* on the smallest eigenvalue of $\mathbf{\Sigma}(t, M)$ provides a lower bound for the compatibility and cone invertibility factors and the restricted eigenvalue in high-dimensional Cox regression.

Theorem 7. *If $\tau > \tau_0$, there exists some constant C_ρ and C_λ such that if $s\sqrt{\frac{\log p}{n}} < C_\rho$ and $\lambda_n = C_\lambda \epsilon$, and $\Delta_{\mathbf{U}} > \sqrt{8c_0\nu^2\tau^2\epsilon} + (2c_0\nu + 1)\tau^2\epsilon$ with constant c_0 defined in (16), then there exists a constant $C > 0$ that does not depend on n, r, p or ϵ such that*

$$\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F \leq C\tau^2\epsilon \tag{25}$$

with a probability greater than $1 - Cp^2\exp(-Cn\epsilon^2) - Cp\exp(-Cn\epsilon^2/s)$. Furthermore, if $s\log p = o(n)$ and that λ_n, τ and $\Delta_{\mathbf{U}}\}$ satisfy $\max\left\{\tau^2\sqrt{s\log p}, \Delta_{\mathbf{U}}^{-2}\tau^2\sqrt{s\log p}, \lambda_n^{-1}\sqrt{\log p}\right\} = o(\sqrt{n})$, we have $\|\mathbf{P}_{\widehat{\mathcal{E}}} - \mathbf{P}_{\mathcal{E}}\|_F \to 0$ in probability as $n, p \to \infty$.

These asymptotic and non-asymptotic properties for sparse NIECE in the Cox model are analogous to Theorem 6 for logistic regression. Hence, we have illustrated the widely applicable Theorem 4 of NIECE in high-dimensional multivariate analysis.

4. Practical considerations

The theory and methods presented so far in this paper are enough to justify NIECE procedures' applications in high-dimensional settings. When implementing the algorithms, however, several additional practical issues arise. This section considers these practical issues and how to deal with them in a sensible data-driven fashion. Because NIECE is a flexible framework for envelope estimation in various settings, we only provide general guidance on dealing with these practical issues, while additional information from specific models and problems would also be helpful.

First, we consider the constrained and the projected envelope estimators. Recall from Sect. 1.2 that the envelope estimator for $\boldsymbol{\theta} \in \mathbb{R}^p$ is the projected estimator $\widehat{\boldsymbol{\theta}}_{\text{Env}} = \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}^{\mathsf{T}}\widehat{\boldsymbol{\theta}}$ that improves over the standard estimator $\widehat{\boldsymbol{\theta}}$, where $\widehat{\mathbf{\Gamma}} \in \mathbb{R}^{p\times u}$ is the estimated envelope basis. In many multivariate parameter estimation problems (see Cook and Zhang, 2015b, for more background), this projected envelope estimator coincides with the constrained envelope estimator, which is obtained from optimizing a likelihood-based objective function under the constraint $\boldsymbol{\theta} = \widehat{\mathbf{\Gamma}}^{\mathsf{T}}\boldsymbol{\eta}$ for some $\boldsymbol{\eta} \in \mathbb{R}^u$. In low-dimensional settings, Cook and Zhang (2015b) showed that the constrained and the projected envelope estimators are often asymptotically equivalent if not exactly the same. However, in the high-dimension estimation, the two estimators could be very different. In our experience (e.g. simulations in Sect. 5), the constrained envelope estimator is generally more robust and accurate in parameter estimation than the projected estimator. As such, we suggest using the constrained envelope estimator

in practice by refitting the model (linear, generalized linear, or Cox proportional hazard model) on the reduced data $\widehat{\boldsymbol{\Gamma}}^{\mathsf{T}}\mathbf{X} \in \mathbb{R}^u$ to obtain the parameters $\widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}}$. We implemented this procedure in all of our numerical studies.

Another important issue in practice is the tuning parameter selection. The sparsity level in high-dimensional NIECE procedure arises from the penalized matrix decomposition step (11), where we follow Witten, Tibshirani and Hastie (2009) to use cross-validation to select the parameter $c > 0$, which is the maximal L_1 norm of the sparse singular vectors. Specifically, we choose c that minimizes the cross-validated prediction mean squared error in the linear model; and in the logistic regression model and Cox model, we choose c that minimizes the cross-validated negative log-likelihood.

Finally, the envelope dimension selection problem is still an open question in high dimensions. Selecting the envelope dimension u is a crucial step in all envelope methods. Zhang and Mai (2018) recently proposed a versatile BIC-type criterion to select u consistently without restricting to a specific model. However, the theory and optimization techniques are only applicable to low-dimensional problems. This paper assumes the envelope dimension u is known and leaves the dimension selection problem for future studies. In practice, the envelope dimension u and the number of principal components d may be selected by cross-validation, which is widely used in practice (Bro et al., 2008; Josse and Husson, 2012). In simulation studies (Sect. 5), we use the true dimension u for our methods and others and compare their subspace and parameter estimation accuracies, where the number of principal components d is fixed at a much bigger number than u. In real data analysis (Sect. 6), we consider tuning d and u together as $d = 2u$ to reduce the computational cost. The idea behind this practice is to view the envelope as an alternative to principal component regression: Instead of reducing the data into k principal components un-supervised, we choose k envelope components from the first $2k$ principal components.

5. Simulation studies

In this section, we study the empirical performances of the sample NIECE procedure (Sect. 2) and the high-dimensional sparse NIECE (SNIECE; Sect. 3) procedure through simulations. We consider four envelope models in high dimensions (response and predictor envelopes in linear regression, envelope in logistic regression, and envelope in Cox proportional hazards model), where we reduce the p-dimensional predictor (or r-dimensional response) onto the u-dimensional envelopes without loss of relevant information in regression. In all simulation models, we set the sample size $n = 200$, the envelope dimension $u = 3$, and the dimension $p = 400$ or 1600. Note that for response reduction in the linear models, we let p be the number of response variables and q be the predictor dimension (in contrast to the r and p notation in the previous sections). The main goal is to estimate a three-dimensional sparse subspace in high dimensions, where we set the sparsity level $s = 10$. Specifically, the key parameter is $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_s^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{p\times u}$, where entries in $\boldsymbol{\Gamma}_s \in \mathbb{R}^{s\times u}$ are generated

from Uniform$[0,1]$ and then orthogonalized such that $\mathbf{\Gamma}^\mathsf{T}\mathbf{\Gamma} = \mathbf{\Gamma}_s^\mathsf{T}\mathbf{\Gamma}_s = \mathbf{I}_u$. To introduce complex correlations among the s relevant variables, we consider the following three types of envelope covariance structures that will be used in each of the four models.

- $\mathbf{\Sigma}_1 = \mathbf{V}\mathbf{D}\mathbf{V}^\mathsf{T}$, where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_s) \in \mathbb{R}^{s\times s}$ is a randomly generated orthogonal matrix and $\mathbf{D} = \mathrm{diag}(D_{11}, \ldots, D_{ss}) \in \mathbb{R}^{s\times s}$ is a diagonal matrix consists of distinct eigenvalues. In multivariate linear models, the $D_{kk} = (k+1)^3$, $k = 1, \ldots, s$; In logistic regression and Cox model, $D_{kk} = 3^{k+1}$, $k = 1, \ldots, s$. The envelope is constructed from $\mathbf{\Gamma}_s = (\mathbf{v}_7, \mathbf{v}_8, \mathbf{v}_9)$.
- $\mathbf{\Sigma}_2 = \mathbf{\Gamma}_s\mathbf{\Omega}\mathbf{\Gamma}_s^\mathsf{T} + \mathbf{\Gamma}_{0s}\mathbf{\Omega}_0\mathbf{\Gamma}_{0s}^\mathsf{T}$, where $\mathbf{\Gamma}_{0s} \in \mathbb{R}^{s\times(s-u)}$ is the orthogonal completion of $\mathbf{\Gamma}_s$ and $\mathbf{\Omega}, \mathbf{\Omega}_0$ are symmetric positive definite matrices. In multivariate linear models, we generate $\mathbf{\Omega}$ as $\mathbf{O}\mathbf{D}\mathbf{O}^\mathsf{T}$, where $\mathbf{O}$ is a randomly generated orthogonal matrix and $\mathbf{D}$ is diagonal with entries $(k+1)^3$, $k = 1, \ldots, u$; $\mathbf{\Omega}_0$ is a diagonal matrix with entries $50, 1, 1, \ldots, 1$. In logistic regression and Cox model, we change the entries in $\mathbf{D}$ to $(k+1)^2$, $k = 1, \ldots, u$, and the diagonals of $\mathbf{\Omega}_0$ to $50, 0.01, \ldots, 0.01$.
- $\mathbf{\Sigma}_3$ is generated similar to $\mathbf{\Sigma}_2$. We set the eigenvalues in $\mathbf{\Omega}$ (i.e. diagonals of $\mathbf{D}$) as $(k+1)^2$, $k = 1, \ldots, u$ and let $\mathbf{\Omega}_0 = 0.01\mathbf{I}_{s-u}$.

The covariance $\mathbf{\Sigma}_1$ contains eigenvalues that are very well-separated, where the largest one is many magnitudes larger than the smallest ones. Both the covariances $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are created such that the largest eigenvalue associated eigenvector is orthogonal to the envelope. Therefore, as seen in the specific models, the first principal component of the data is irrelevant to the regression but only brings substantial amounts of estimative variability. In contrast, the covariance $\mathbf{\Sigma}_3$ is created in favor of principal component regression – the leading u eigenvalues are well-separated and more than a hundred times larger than the remaining eigenvalues. Then these covariance structures will be used in the following four models.

M1: Response envelope model in multivariate linear regression. For predictor dimension $q = 10$, we generate the sparse coefficient parameter $\boldsymbol{\beta}^* \in \mathbb{R}^{p\times q}$ as $\boldsymbol{\beta}^* = \mathbf{\Gamma}\boldsymbol{\eta}$ and then standardized to $\boldsymbol{\beta} = 10 \cdot \boldsymbol{\beta}^*/\|\boldsymbol{\beta}^*\|_F$, where the entries in $\boldsymbol{\eta} \in \mathbb{R}^{u\times q}$ are sampled from Uniform$[0,1]$. The response is generated as $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, where the error $\boldsymbol{\epsilon}_i \sim N(0, \mathrm{diag}(\mathbf{\Sigma}_A, \mathbf{I}_{p-s}))$ for $A = 1, 2, 3$ three different covariance structures. The predictors are generated from $N(0, 30\mathbf{I}_q)$ when $\mathbf{\Sigma}_A = \mathbf{\Sigma}_1$ due to high variance in $\mathbf{\Sigma}_1$, and are generated from $N(0, \mathbf{I}_q)$ for $\mathbf{\Sigma}_A = \mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$.

M2: Predictor envelope model in multivariate linear regression. For response dimension $q = 5$, we generate $\boldsymbol{\beta} \in \mathbb{R}^{p\times q}$ in the same way as in Model M1. Then $\mathbf{Y}_i = \boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i + \boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, where the predictor $\mathbf{X}_i \sim N(0, \mathrm{diag}(\mathbf{\Sigma}_A, 0.01\mathbf{I}_{p-s}))$ and the error $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2\mathbf{I}_q)$. The magnitude of error is controlled by σ^2 to be comparable with the predictor variability. Specifically, $\sigma = 200, 20, 10$ for $A = 1, 2, 3$ respectively.

M3: Envelope logistic regression. The sparse coefficient vector $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta} \in \mathbb{R}^p$ is generated in the same way as in Model M1 where we set $q = 1$. The predictor $\mathbf{X}_i$ is also generated in the same way except that we change $\boldsymbol{\Sigma}_A$ to its scaled version $\boldsymbol{\Sigma}_A/\|\boldsymbol{\Sigma}_A\|$, so that the classification accuracy by envelope logistic regression is around 20% (neither too challenging nor too easy). The Y_i follows Bernoulli distribution with probability $\exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)/(1+\exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i))$.

M4: Envelope Cox proportional hazards model. The procedure for generating $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\mathbf{X}_i$ are the same as in Model M3. Then the failure time Y_i and censoring time C_i are generated from exponential distribution $\text{Exp}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)$ and $\text{Exp}(0.5)$, respectively. Finally, we let $T_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \le C_i)$.

We focus on the parameter estimation error $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_F$ and the subspace estimation error $\|\mathbf{P}_{\boldsymbol{\Gamma}} - \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\|_F/\sqrt{2u} \in [0,1]$. For comparison purpose, we focus on principal component regression (PCR, Jolliffe, 1986) and sparse principal component regression (SPCR) based on the same penalized matrix decomposition algorithm (Witten, Tibshirani and Hastie, 2009) that was used in SNIECE. We also include partial least squares regression (PLS, Wold, 1966), sparse partial least square regression (SPLS, Chun and Keleş, 2010), sparse reduced-rank regression (SRRR, Chen and Huang, 2012), reduced rank stochastic regression (RSSVD, Chen, Chan and Stenseth, 2012), supervised principal component regression (SupPCR, Bair et al., 2006). In multivariate linear models, we also include ordinary least squares (OLS) and Lasso (Tibshirani, 1996) estimators as benchmarks for estimating regression parameter $\boldsymbol{\beta}$; in logistic regression model and Cox proportional hazards model, we include ℓ_1-penalized MLE (PMLE, Friedman, Hastie and Tibshirani, 2010) for comparison while PLS and SPLS are not directly applicable. In all models, we set $d = 10 > u = 3$ in the NIECE and SNIECE procedures. In our experience, moderately increase d will not change the performance of our methods under these simulation settings. The implementation of SupPCR provided by Bair et al. (2006) is limited to univariate response $\mathbf{Y}$, which pre-screens some predictors based on their correlations with the response and performs classical principal component analysis on the selected predictors. To compare supervised principal component regression under our simulated models with predictor envelope (**M2**–**M4**), we implement the pre-screening step using Lasso for multivariate linear model and penalized MLE for logistic regression and Cox proportional hazards model.

Figures 3–6 summarize the results for all the four models across the three covariance structures. In Sect. A of the Appendix, we provide the average and median estimation errors in Tables 2 and 3, respectively.

We have confirmed the following theoretical findings for linear, logistic, and Cox models through numerical experiments. (i) Under covariance structure $\boldsymbol{\Sigma}_3$, where the leading principal components span the envelope subspace, our NIECE procedure correctly identified these principal components as relevant information in regression. As a result, the NIECE estimator coincided with the PCR

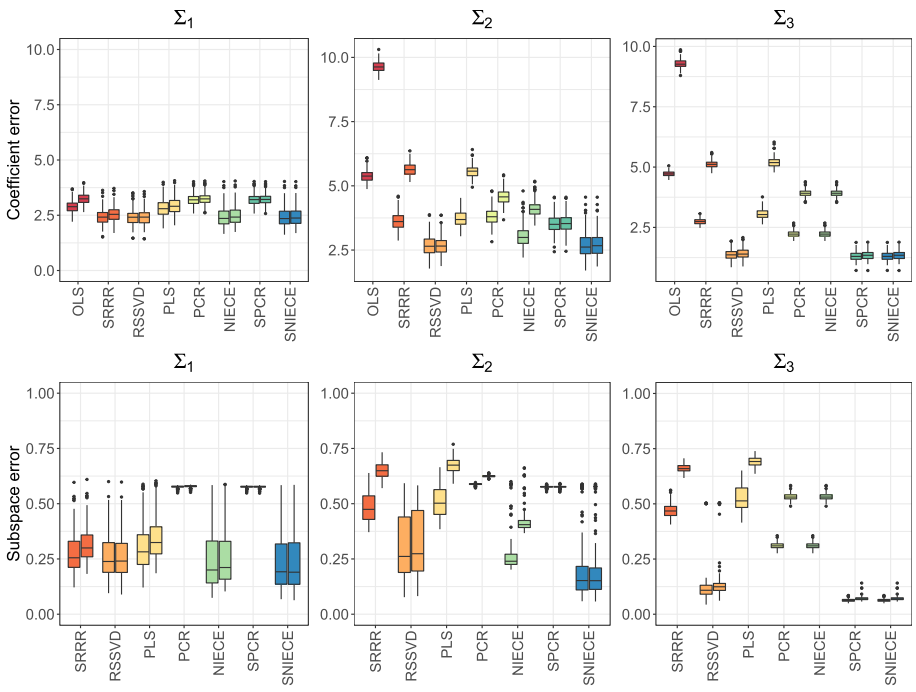

FIG 3. *Summary plot for* ***M1****: response envelope model in multivariate linear regression. Reported are estimation errors of parameter* $\Delta_\beta = \|\beta - \widehat{\beta}\|_F$ *and envelope subspace* $\Delta_{\boldsymbol{\Gamma}} = \|\mathbf{P}_{\boldsymbol{\Gamma}} - \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\|_F/\sqrt{2u}$. *For each method, the left and right bars correspond to* $r = 400$ *and* $r = 1600$, *respectively. Results are based on 200 replications.*

estimator; and the SNIECE coincided with the SPCR. However, SNIECE considers the information in the response $\mathbf{Y}$ while selecting the tuning parameter and can perform more stable than SPCR as shown by Fig. 4. (ii) Under covariance structures $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, either the first or the second principal component is orthogonal to the envelope. Our NIECE procedure correctly identified the corresponding leading principal components as irrelevant information in regression and eliminated it in subsequent estimation. As a result, the NIECE and SNIECE estimators substantially improved over the unsupervised counterparts PCR and SPCR, respectively. The advantages were reflected in both subspace estimation and parameter estimation. (iii) In these high-dimensional sparse settings, both the SPCR (under covariance structure $\boldsymbol{\Sigma}_3$) and SNIECE converged much faster than their un-penalized counterparts (PCR and NIECE).

Although we did not establish the parameter estimation consistency of $\widehat{\boldsymbol{\beta}}_{\text{Env}}$ in theoretical analysis, the overall simulation results showed that the NIECE estimators (penalized or un-penalized) had a promising performance. They outperformed the standard solutions such as OLS, Lasso in multiple linear regression, and ℓ_1-penalized MLE in logistic and Cox models. The significant im-

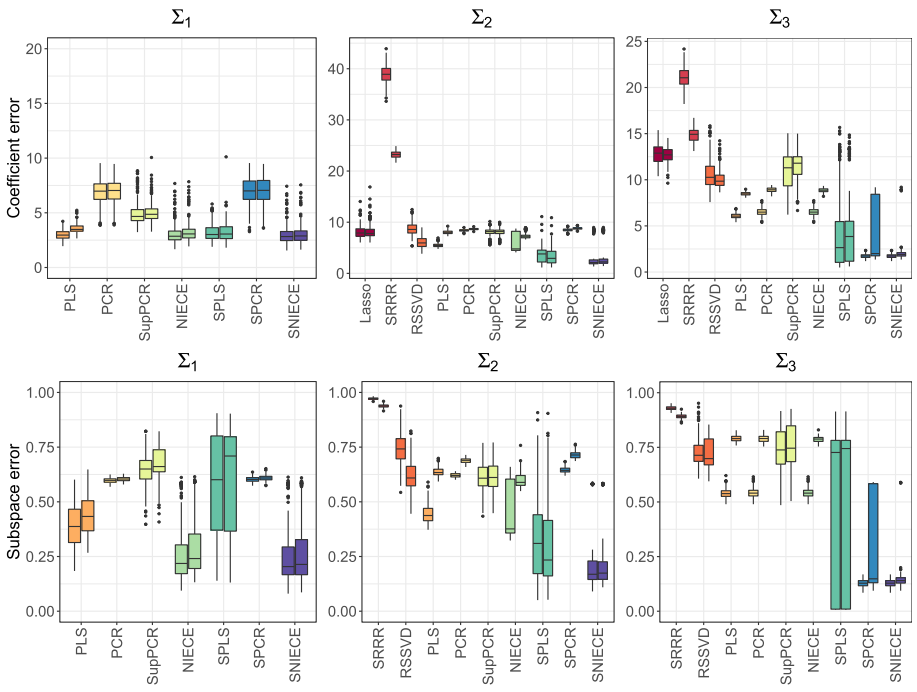

FIG 4. *Summary plot for* ***M2****: Predictor envelope model in multivariate linear regression. Reported are estimation errors of parameter* $\Delta_{\boldsymbol{\beta}} = \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_F$ *and envelope subspace* $\Delta_{\boldsymbol{\Gamma}} = \|\mathbf{P}_{\boldsymbol{\Gamma}} - \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\|_F/\sqrt{2u}$*. For each method, the left and right bars correspond to* $p = 400$ *and* $p = 1600$*, respectively. Results are based on 200 replications. Under covariance structure* $\boldsymbol{\Sigma}_1$*, Lasso and SRRR had very big estimation errors (*$\Delta_{\boldsymbol{\beta}} > 30$ *and* $\Delta_{\boldsymbol{\beta}} > 100$*, respectively) and were excluded in comparison for better visualization; similarly, RSSVD also failed due to extremely high variability in data.*

provements are due to our design of highly correlated variables. On the other hand, while PCR and SPCR were widely used in such high correlation data applications, the NIECE and SNIECE can be viewed as practical and straightforward improvements. Finally, we have also seen significant improvements in SNIECE over popular multivariate regression methods such as SPLS, SRRR, and RSSVD. We want to remark that these simulations were designed to be very challenging and in favor of NIECE procedures to confirm their theoretical properties and potential advantages in a wide range of applications. The NIECE and those methods are not directly comparable because of their different focuses in the regression. For example, SRRR failed under Model **M2** because it focused on imposing group sparsity on column vectors of the coefficient matrix $\boldsymbol{\beta}$ and failed to select the important predictors when there is a large variation in the error term.

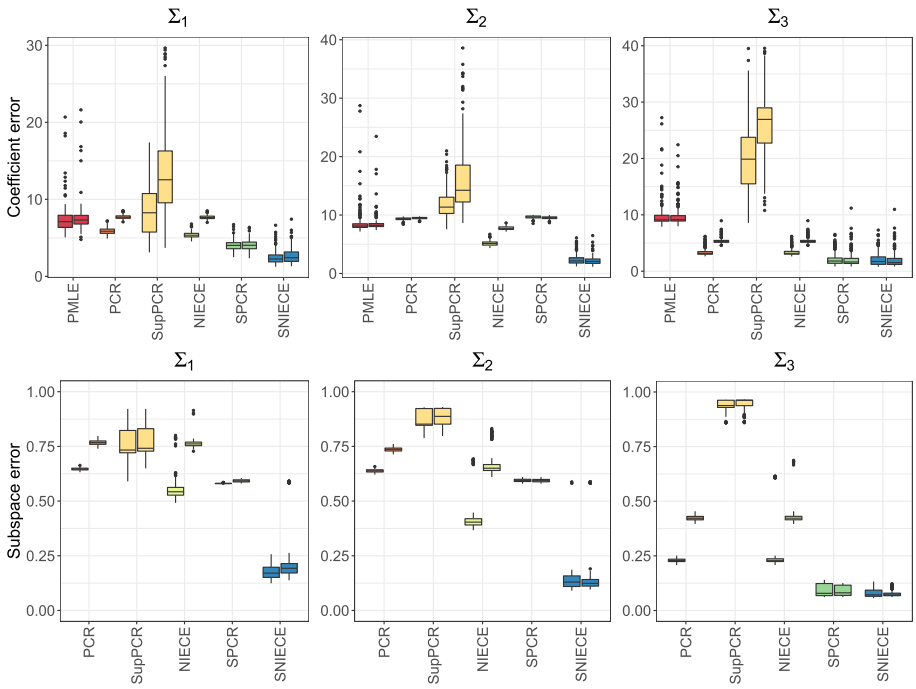

FIG 5. *Summary plot for* ***M3****: Envelope logistic regression. Reported are estimation errors of parameter* $\Delta_{\beta} = \|\beta - \widehat{\beta}\|_F$ *and envelope subspace* $\Delta_{\mathbf{\Gamma}} = \|\mathbf{P}_{\mathbf{\Gamma}} - \mathbf{P}_{\widehat{\mathbf{\Gamma}}}\|_F/\sqrt{2u}$*. For each method, the left and right bars correspond to* $p = 400$ *and* $p = 1600$*, respectively. Results are based on 200 replications.*

6. Real data illustration

In this section, we include three datasets to illustrate the various applications of the proposed high-dimensional sparse NIECE procedure in logistic regression and linear regression with either univariate response or multivariate response. We focus on the predictive results of SNIECE in comparison to SPCA and ℓ_1-penalized (generalized) linear model (i.e. LASSO estimator). Following the practical suggestions in Sect. 4, we use $d = 2u$ principal components for each envelope dimension $u \in \{1, \ldots, 20\}$ in SNIECE.

The first study is the meat property data from Sæbø et al. (2008), where they collected the Near-infrared (NIR) spectroscopy measurements and water, fat, protein compositions for $n = 103$ meat samples. Cook, Helland and Su (2013b) took spectral measurements at every fourth wavelength between 850 nm and 1050 nm as predictors, yielding $p = 50$; and they showed promising performance of the predictor envelope in the linear regression model to predict the protein content. We use the spectral measurements at every two wavelengths as predictors, yielding a higher predictor dimension of $p = 100$. We then use this to illustrate the envelope logistic regression model with NIECE to clas-

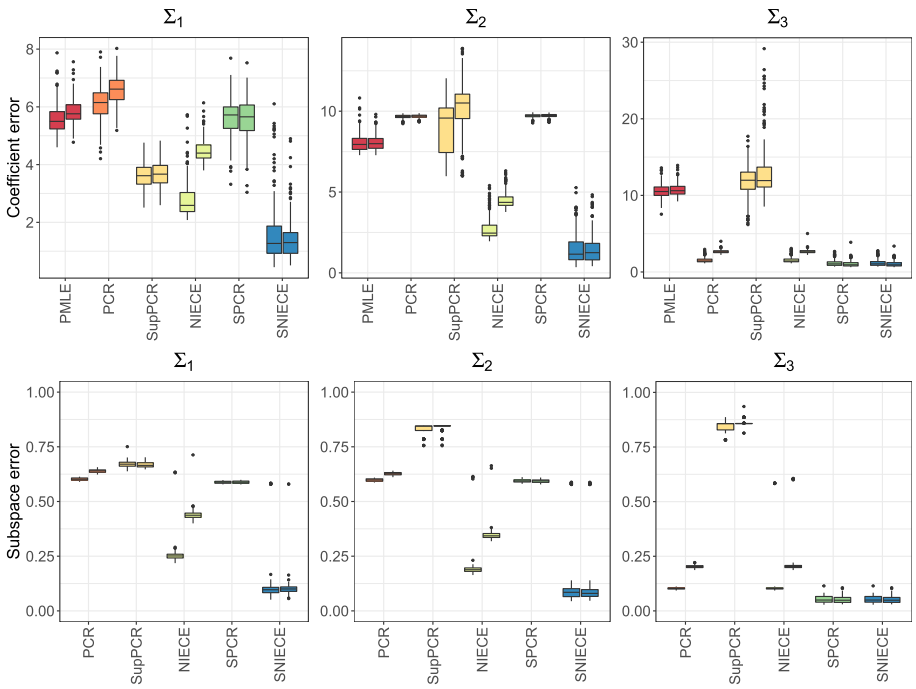

FIG 6. *Summary plot for* ***M4****: Envelope Cox proportional hazards model. Reported are estimation errors of parameter* $\Delta_{\boldsymbol{\beta}} = \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_F$ *and envelope subspace* $\Delta_{\boldsymbol{\Gamma}} = \|\mathbf{P}_{\boldsymbol{\Gamma}} - \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\|_F/\sqrt{2u}$. *For each method, the left and right bars correspond to* $p = 400$ *and* $p = 1600$, *respectively. Results are based on 200 replications.*

sify the $n = 103$ meat samples into 49 beef samples and 54 pork samples. The binary response and highly correlated predictors brought additional challenges than the previous analysis. The second study is the riboflavin production data from Bühlmann, Kalisch and Meier (2014). It contains the riboflavin production of $n = 71$ Bacillus subtilis samples. The univariate continuous response is log-transformed riboflavin production; the high-dimensional predictor variables measure the logarithm of the expression level of $p = 4088$ genes. The third study is the music data from Zhou, Claire and King (2014) and downloaded from the UCI machine learning repository (`https://archive.ics.uci.edu/ml/datasets/Geographical+Original+of+Music`). It includes $n = 1059$ music tracks with $p = 116$ audio features, and the $r = 2$ bivariate response describes the origin location of a track, represented by standardized latitude and longitude.

For each dataset, we randomly split the data 100 times into a training set of size n_{train} and a testing set of size n_{test}, the averaged mis-classification error $\sum_{i=1}^{n_{\text{test}}} I(Y_i \neq \widehat{Y}_i)/n_{\text{test}} \times 100\%$ or the averaged prediction mean squared error $\sum_{i=1}^{n_{\text{test}}} \|\mathbf{Y}_i - \widehat{\mathbf{Y}}_i\|_2^2/n_{\text{test}}$ is then recorded for each number of component u ranges from 1 to 20. For the meat data and the riboflavin data, the training and testing

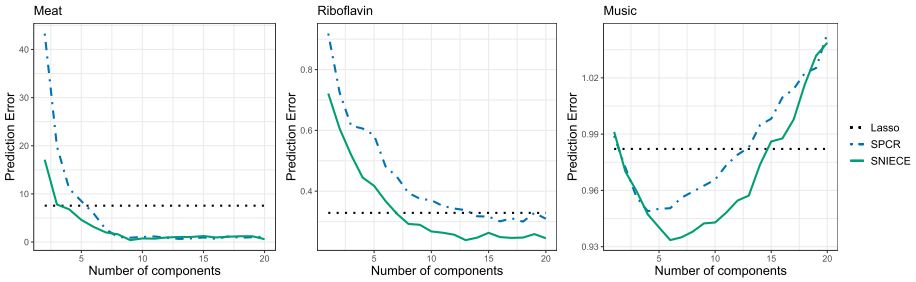

FIG 7. *The average prediction errors using* 100 *random data splittings where we vary the number of components in PCR and the envelope dimension in NIECE. For the meat data (left panel), the prediction error is defined as the misclassification rate in percentage; for the riboflavin data and the music data, the prediction error is the prediction mean squared error.*

sample size ratio is four to one (i.e. five-fold cross-validation); for the music data, we set $n_{\text{train}} = 100$ to produce the high-dimension low-sample size scenario.

Figure 7 summarizes the prediction error. In the meat data, the NIR spectral measurements make the predictors extremely highly-correlated. Therefore, the classification errors of both SPCR and SNIECE decrease as we increase the number of components and eventually converge to nearly perfect classification. With a small number of components, the SNIECE is more effective than SPCR. This phenomenon can be explained by noticing that the first principal component is not useful for classification at all – a situation resembles our simulations. For this data set, the ℓ_1-penalized logistic regression ignores the predictor correlation and has an error rate of about 8%. In the riboflavin production data, the prediction error of SNIECE is uniformly smaller than that of SPCR for the whole range of dimensions. For this very high-dimensional data with much weaker correlations than in the meat data, the envelope approach still significantly improved over the lasso regression while SPCR fails to achieve so. These are very encouraging results and, to the best of our knowledge, the first real data application of envelope regression with p in thousands. Finally, in the music data, the prediction error curves of SNIECE and SPCA in Fig. 7 is another typical situation: the SNIECE can achieve a much smaller error than SPCA and is also uniformly better than both SPCA and Lasso regression for a wide range of dimensions, $2 \leq u \leq 14$. Based on the results in Fig. 7, we have confirmed the potential advantages of NIECE over PCR in the high-dimensional setting. The proposed estimator can be widely adopted as a simple, unified, and effective alternative to SPCR.

7. Discussion

In this article, we develop a new method and theory for envelope subspace estimation by connecting envelope method with principal component regression. We establish a general theory for the non-iterative envelope component estimation (NIECE) algorithm in both sparse and non-sparse settings, where the

dimensionality diverges with the sample size. More importantly, the NIECE algorithm is computationally straightforward and easily generalizable to various supervised learning problems. Our numerical studies show the method performs well in regression problems with high-dimensional highly correlated predictors.

A direction for future research direction is the selection of envelope dimension u when p is large. For NIECE procedure, the envelope dimension selection problem is more challenging than the likelihood-based envelope estimation. This because it also relies on the specification of d, the number of top principal components (PCs) to be used in the NIECE algorithm. We suggest using $d = 2u$ as a conservative way of improving over PCR: Instead of reducing the data into u PCs un-supervised, we choose u components from the first $2u$ PCs. In all the three real datasets, we observe a robust and consistent improvement over PCR for a wide range of choices $u \in \{1, \ldots, 20\}$. Because our methodology is very general and easy to modify, one can adopt other ways of tuning u and d in practice and study the theoretical properties in the future.

Throughout the paper we assume that the first d eigenvalues of $\mathbf{M}$ are distinct. In presence of common eigenvalues, we can re-define the eigengap Δ and relax this assumption to some extent. See Appendix B for a more detailed discussion. For future studies, we plan to incorporate principal subspace estimation into the NIECE algorithm and fully investigate the algorithm and theory in such a more challenging settings.

Appendix A: Additional numerical results

A.1. Additional real data analysis results

Furthermore, it is common in PCA to choose number of components that can explain a specific proportion of variability in data. Thus in meat property data, we also compare the classification performance by first applying PCA and sparse PCA on the original data and observe the number of components u_1 needed for PCA and u_2 needed for sparse PCA, to explain 90%, 95% and 99% of total variance. The comparison results are summarized in Table 1, to explain the same amount of variation in data, SNIECE achieves the lowest classification error.

A.2. Additional simulation results

Tables 2 & 3 demonstrate that the proposed SNIECE achieved the lowest median and average estimation errors. However, due to the challenging model settings such that the variation in data can be very large, the average errors are slightly higher than the median errors for SNIECE in most settings. Notably, under Model **M2** and covariance structure $\mathbf{\Sigma}_2$, SNIECE yielded a considerably higher average error than median error. This observation is also supported by the outliers of SNIECE displayed in Fig. 4. For predictor envelope in linear regression, note that $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T) = \mathcal{E}_{\mathbf{\Sigma_X}}(\mathbf{\Sigma_{XY}})$ at the population level, which gives additional

TABLE 1

Envelope logistic regression on meat property data. Reported are the mean and standard error (in parenthesis) of mis-classification rates (%) for 100 random data splits. Var. denotes total variance explained. u_1 is number of components for PCR and NIECE, u_2 is number of components for SPCR and SNIECE.

Var.	u_1	u_2	PMLE	PCR	NIECE	SPCR	SNIECE
90%	2	5	7.55 (0.61)	40.15 (1.06)	34.90 (1.64)	8.50 (0.61)	5.90 (0.51)
95%	2	7	7.55 (0.61)	40.15 (1.06)	34.90 (1.64)	2.65 (0.39)	2.15 (0.37)
99%	3	12	7.55 (0.61)	8.20 (0.51)	8.20 (0.51)	0.90 (0.22)	0.55 (0.17)

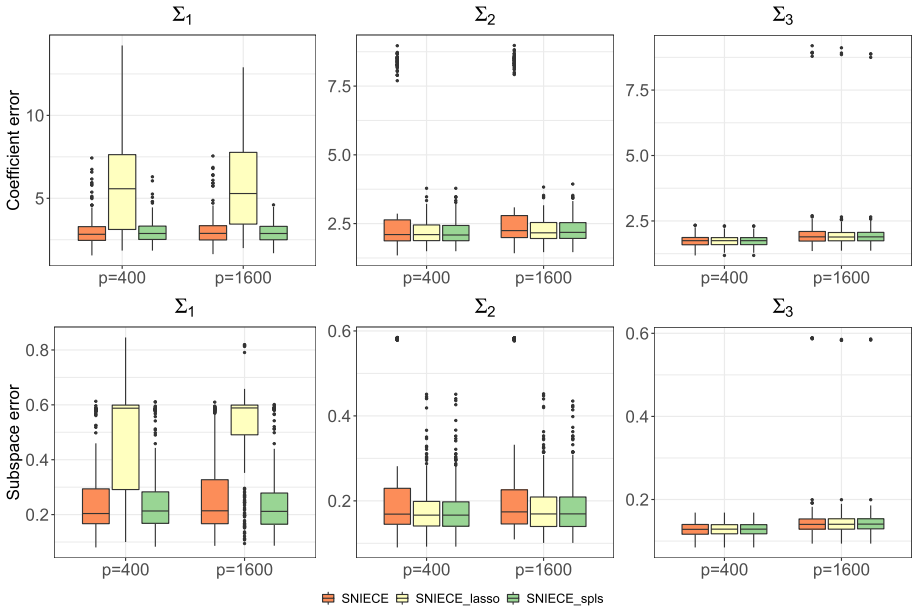

FIG 8. *Summary plot for **M2**: Predictor envelope model in multivariate linear regression. Reported are estimation errors of parameter* $\Delta_{\boldsymbol{\beta}} = \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_F$ *and envelope subspace* $\Delta_{\boldsymbol{\Gamma}} = \|\mathbf{P}_{\boldsymbol{\Gamma}} - \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\|_F/\sqrt{2u}$. *SNIECE uses* $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}}$; *SNIECE_lasso and SNIECE_spls use* $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\beta}}$. *For SNIECE_lasso,* $\widehat{\boldsymbol{\beta}}$ *is the sparse estimate from Lasso; for SNIECE_spls,* $\widehat{\boldsymbol{\beta}}$ *is the sparse estimate from SPLS.*

flexibility regarding the choice of $\widehat{\mathbf{U}}$ in the proposed NIECE procedure. In the present implement of SNIECE in simulations, $\widehat{\mathbf{U}}$ was set to $\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}}$. However, an alternative choice is to set $\widehat{\mathbf{U}}$ to $\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is some sparse estimate from high-dimensional regression methods including Lasso or sparse partial least square regression. Figure 8 summarizes the performance of SNIECE when $\widehat{\mathbf{U}}$ is specified by sample covariance, Lasso, and sparse partial least square under

Model **M2**. The performance of SNIECE varied significantly depending on the choice of $\widehat{\mathbf{U}}$. Specifically, under covariance structure $\mathbf{\Sigma}_1$, Lasso failed to select important predictors due to large variation in the error term. Consequently, initializing $\widehat{\mathbf{U}}$ with Lasso resulted in poor estimation performance for SNIECE; under covariance $\mathbf{\Sigma}_2$, SNIECE initializing $\widehat{\mathbf{U}}$ with sparse estimates Lasso or SPLS obtained more stable estimation compared to initializing with sample covariance $\widehat{\mathbf{\Sigma}}_{\mathbf{XY}}$; under covariance $\mathbf{\Sigma}_3$, all three initialization methods for $\widehat{\mathbf{U}}$ gave similar results. Thus, while implementing the proposed NIECE for linear regression with high-dimensional predictors, users have the flexibility to specify different $\widehat{\mathbf{U}}$ based on the particular problem and data at hand.

Appendix B: A brief discussion on common eigenvalues

If the eigenvalues $\lambda_{\pi(j)}$'s, $j = 1, \ldots, u$, are all distinct, we define the following population quantities

$$\Delta = \min_{j=1,\ldots,u} \Delta_j, \quad \Delta_j \equiv \min\{\lambda_{\pi(j)-1}(\mathbf{M}) - \lambda_{\pi(j)}(\mathbf{M}),\ \lambda_{\pi(j)}(\mathbf{M}) - \lambda_{\pi(j)+1}(\mathbf{M})\}. \tag{26}$$

In presence of common eigenvalues for some $j = 1, \ldots, u$, denoted as $\lambda_{\pi_1} = \cdots = \lambda_{\pi_m}$ with $\pi_1 > \cdots > \pi_m$, $m \leq u$, we re-define Δ_j as

$$\Delta_j \equiv \min\{\lambda_{\pi_1-1}(\mathbf{M}) - \lambda_{\pi_1}(\mathbf{M}),\ \lambda_{\pi_m}(\mathbf{M}) - \lambda_{\pi_m+1}(\mathbf{M})\}. \tag{27}$$

Note that we need to assume these eigenvalues are distinct from the eigenvalues associated with eigenvectors in $\mathcal{E}^{\perp}_{\mathbf{M}}(\mathbf{U})$ (an extreme case will be $\mathbf{M} = \mathbf{I}$, then the NIECE algorithm needs modification), otherwise we need to modify/replace $\mathbf{M}$ by $\mathbf{M} + \mathbf{U}$.

Lemma 4. *Assume $\Delta > 0$, and let $\epsilon > 0$ be a constant such that $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\text{op}} \leq \epsilon$, assume that $\{\widehat{\pi}(j) \mid j = 1, \ldots, u\} = \{\pi(j) \mid j = 1, \ldots, u\}$, then*

$$\|\sin \mathbf{\Theta}(\mathbf{\Gamma}, \widehat{\mathbf{\Gamma}})\|_F \leq \frac{2\sqrt{u}\epsilon}{\Delta}. \tag{28}$$

Proof. From Corollary 1 of Yu, Wang and Samworth (2014), we have that for distint eigenvalues,

$$\sin \Theta(\mathbf{v}_j, \widehat{\mathbf{v}}_j) \leq \frac{2\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\text{op}}}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})} \leq \frac{2\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\text{op}}}{\Delta_j} \leq \frac{2\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\text{op}}}{\Delta}. \tag{29}$$

For common eigenvalues, $\lambda_{\pi_1} = \cdots = \lambda_{\pi_m}$ with $\pi_1 > \cdots > \pi_m$, we let $\mathbf{W}_j = (\mathbf{v}_{\pi_1}, \ldots, \mathbf{v}_{\pi_m})$ and $\widehat{\mathbf{W}}_j = (\widehat{\mathbf{v}}_{\pi_1}, \ldots, \widehat{\mathbf{v}}_{\pi_m})$. Then we have,

$$\|\sin \mathbf{\Theta}(\mathbf{W}_j, \widehat{\mathbf{W}}_j)\|_F \leq \frac{2\sqrt{m}\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\text{op}}}{\Delta_j} \leq \frac{2\sqrt{m}\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\text{op}}}{\Delta}. \tag{30}$$

TABLE 2

The average estimation errors for the parameter $\Delta_{\boldsymbol{\beta}} = \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_F$ and for the envelope subspace $\Delta_{\boldsymbol{\Gamma}} = \|\mathbf{P}_{\boldsymbol{\Gamma}} - \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\|_F/\sqrt{2u}$. Results are based on 200 replications. The maximum standard error among all estimators in each setting (i.e. each column of the Table) are included. In Model M2 with covariance structure $\boldsymbol{\Sigma}_1$, SRRR had very big estimation errors ($\Delta_{\boldsymbol{\beta}} > 100$) and was excluded in comparison; similarly, RSSVD also failed due to extremely high variability in data.

		M1 (Response envelope in linear model)						**M2** (Predictor envelope in linear model)					
$\boldsymbol{\Sigma}_A$		$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$		$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$	
p		400	1600	400	1600	400	1600	400	1600	400	1600	400	1600
OLS	$\Delta_{\boldsymbol{\beta}}$	2.89	3.24	5.38	9.64	4.72	9.29	–	–	–	–	–	–
Lasso	$\Delta_{\boldsymbol{\beta}}$	–	–	–	–	–	–	30.00	29.00	8.08	8.22	12.80	12.69
PLS	$\Delta_{\boldsymbol{\beta}}$	2.81	2.92	3.72	5.56	3.05	5.19	3.00	3.57	5.50	8.02	6.08	8.48
	$\Delta_{\boldsymbol{\Gamma}}$	0.30	0.35	0.51	0.67	0.53	0.69	0.39	0.44	0.44	0.64	0.54	0.79
SRRR	$\Delta_{\boldsymbol{\beta}}$	2.42	2.54	3.61	5.63	2.74	5.12	–	–	38.84	23.23	21.12	14.90
	$\Delta_{\boldsymbol{\Gamma}}$	0.28	0.31	0.49	0.65	0.47	0.66	1.00	1.00	0.97	0.94	0.93	0.89
RSSVD	$\Delta_{\boldsymbol{\beta}}$	2.40	2.39	2.65	2.66	1.36	1.42	–	–	8.64	6.05	10.65	10.12
	$\Delta_{\boldsymbol{\Gamma}}$	0.26	0.26	0.31	0.32	0.12	0.15	–	–	0.74	0.62	0.73	0.72
PCR	$\Delta_{\boldsymbol{\beta}}$	3.20	3.24	3.81	4.57	2.23	3.92	6.92	6.95	8.39	8.66	6.52	8.91
	$\Delta_{\boldsymbol{\Gamma}}$	0.58	0.58	0.59	0.62	0.31	0.53	0.58	0.59	0.61	0.67	0.54	0.79
SupPCR	$\Delta_{\boldsymbol{\beta}}$	–	–	–	–	–	–	5.89	9.57	8.13	8.08	10.96	11.37
	$\Delta_{\boldsymbol{\Gamma}}$	–	–	–	–	–	–	0.64	0.68	0.61	0.61	0.74	0.75
NIECE	$\Delta_{\boldsymbol{\beta}}$	2.44	2.48	3.04	4.11	2.23	3.92	3.11	3.30	5.72	7.42	6.52	8.87
	$\Delta_{\boldsymbol{\Gamma}}$	0.26	0.27	0.27	0.42	0.31	0.53	0.27	0.30	0.44	0.60	0.54	0.79
SPLS	$\Delta_{\boldsymbol{\beta}}$	–	–	–	–	–	–	3.23	3.28	3.58	3.28	4.38	4.52
	$\Delta_{\boldsymbol{\Gamma}}$	–	–	–	–	–	–	0.59	0.59	0.36	0.32	0.45	0.48
SPCR	$\Delta_{\boldsymbol{\beta}}$	3.21	3.21	3.51	3.52	1.30	1.34	6.87	6.86	8.47	8.52	1.74	3.75
	$\Delta_{\boldsymbol{\Gamma}}$	0.58	0.58	0.58	0.58	0.06	0.07	0.58	0.58	0.58	0.58	0.13	0.26
SNIECE	$\Delta_{\boldsymbol{\beta}}$	2.43	2.44	2.73	2.72	1.30	1.34	3.04	3.08	3.40	3.49	1.74	2.08
	$\Delta_{\boldsymbol{\Gamma}}$	0.25	0.25	0.21	0.21	0.06	0.07	0.26	0.28	0.25	0.25	0.13	0.15
S.E.$\leq$	$\Delta_{\boldsymbol{\beta}}$	(0.13)	(0.13)	(0.07)	(0.05)	(0.11)	(0.09)	(0.09)	(0.09)	(0.19)	(0.19)	(0.30)	(0.29)
	$\Delta_{\boldsymbol{\Gamma}}$	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)

		M3 (Logistic regression)						**M4** (Cox hazards model)					
$\boldsymbol{\Sigma}_A$		$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$		$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$	
p		400	1600	400	1600	400	1600	400	1600	400	1600	400	1600
PMLE	$\Delta_{\boldsymbol{\beta}}$	7.45	7.56	8.73	8.57	9.94	9.84	5.57	5.83	8.05	8.05	10.62	10.75
PCR	$\Delta_{\boldsymbol{\beta}}$	5.88	7.71	9.33	9.48	3.36	5.39	6.10	6.60	9.67	9.69	1.55	2.66
	$\Delta_{\boldsymbol{\Gamma}}$	0.65	0.76	0.63	0.73	0.23	0.42	0.59	0.63	0.59	0.62	0.10	0.20
SupPCR	$\Delta_{\boldsymbol{\beta}}$	8.41	14.11	11.98	17.14	20.40	94.61	3.64	3.67	9.10	10.16	11.74	13.11
	$\Delta_{\boldsymbol{\Gamma}}$	0.76	0.77	0.87	0.89	0.94	0.94	0.67	0.67	0.83	0.84	0.85	0.86
NIECE	$\Delta_{\boldsymbol{\beta}}$	5.38	7.67	5.17	7.76	3.41	5.39	2.80	4.48	2.74	4.50	1.56	2.67
	$\Delta_{\boldsymbol{\Gamma}}$	0.56	0.75	0.42	0.66	0.25	0.43	0.25	0.43	0.19	0.34	0.12	0.22
SPCR	$\Delta_{\boldsymbol{\beta}}$	4.37	4.26	9.38	9.43	2.09	2.08	5.76	5.66	9.71	9.72	0.97	0.99
	$\Delta_{\boldsymbol{\Gamma}}$	0.59	0.59	0.58	0.58	0.09	0.09	0.58	0.58	0.58	0.58	0.05	0.06
SNIECE	$\Delta_{\boldsymbol{\beta}}$	2.47	2.65	2.12	2.08	2.03	2.04	1.62	1.44	1.49	1.48	0.97	0.98
	$\Delta_{\boldsymbol{\Gamma}}$	0.19	0.22	0.21	0.26	0.09	0.08	0.12	0.10	0.11	0.11	0.05	0.06
S.E.$\leq$		(0.14)	(0.13)	(0.18)	(0.11)	(0.18)	(0.14)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)

Since $\|\sin\boldsymbol{\Theta}(\boldsymbol{\Gamma}, \widehat{\boldsymbol{\Gamma}})\|_F^2 = \sum_{j=1}^{u} \sin^2(\theta_j)$, where θ_j's are the principal angles, we have

$$\|\sin\boldsymbol{\Theta}(\boldsymbol{\Gamma}, \widehat{\boldsymbol{\Gamma}})\|_F = \sqrt{\sum_{j\in\mathcal{J}_1} \sin^2\Theta(\mathbf{v}_j, \widehat{\mathbf{v}}_j) + \sum_{j\in\mathcal{J}_2} \|\sin\boldsymbol{\Theta}(\mathbf{W}_j, \widehat{\mathbf{W}}_j)\|_F^2}$$

TABLE 3

The median estimation errors for the parameter $\Delta_{\boldsymbol{\beta}} = \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_F$ and for the envelope subspace $\Delta_{\boldsymbol{\Gamma}} = \|\mathbf{P}_{\boldsymbol{\Gamma}} - \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\|_F/\sqrt{2u}$. Results are based on 200 replications. The maximum standard error among all estimators in each setting (i.e. each column of the Table) are included. In Model M2 with covariance structure $\boldsymbol{\Sigma}_1$, SRRR had very big estimation errors ($\Delta_{\boldsymbol{\beta}} > 100$) and was excluded in comparison; similarly, RSSVD also failed due to extremely high variability in data.

		M1 (Response envelope in linear model)						**M2** (Predictor envelope in linear model)					
$\boldsymbol{\Sigma}_A$		$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$		$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$	
p		400	1600	400	1600	400	1600	400	1600	400	1600	400	1600
OLS	$\Delta_{\boldsymbol{\beta}}$	2.89	3.24	5.38	9.63	4.72	9.26	–	–	–	–	–	–
Lasso	$\Delta_{\boldsymbol{\beta}}$	–	–	–	–	–	–	26.18	25.48	7.89	7.79	12.88	12.70
PLS	$\Delta_{\boldsymbol{\beta}}$	2.80	2.90	3.69	5.55	3.03	5.19	2.96	3.48	5.43	8.02	6.06	8.47
	$\Delta_{\boldsymbol{\Gamma}}$	0.28	0.32	0.50	0.67	0.51	0.69	0.39	0.43	0.44	0.63	0.54	0.79
SRRR	$\Delta_{\boldsymbol{\beta}}$	2.42	2.54	3.61	5.63	2.74	5.11	–	–	38.91	23.23	21.06	14.93
	$\Delta_{\boldsymbol{\Gamma}}$	0.26	0.30	0.47	0.65	0.47	0.66	1.00	1.00	0.97	0.94	0.93	0.89
RSSVD	$\Delta_{\boldsymbol{\beta}}$	2.41	2.41	2.65	2.65	1.36	1.39	–	–	8.58	5.95	10.27	9.85
	$\Delta_{\boldsymbol{\Gamma}}$	0.24	0.24	0.26	0.27	0.11	0.12	–	–	0.74	0.61	0.71	0.70
PCR	$\Delta_{\boldsymbol{\beta}}$	3.21	3.21	3.50	3.51	1.30	1.34	6.98	7.04	8.40	8.66	6.49	8.92
	$\Delta_{\boldsymbol{\Gamma}}$	0.58	0.58	0.59	0.62	0.31	0.53	0.58	0.59	0.61	0.67	0.54	0.79
SupPCR	$\Delta_{\boldsymbol{\beta}}$	–	–	–	–	–	–	4.67	4.86	8.16	8.16	11.30	11.80
	$\Delta_{\boldsymbol{\Gamma}}$	–	–	–	–	–	–	0.65	0.66	0.61	0.61	0.74	0.75
NIECE	$\Delta_{\boldsymbol{\beta}}$	2.36	2.40	2.99	4.08	2.22	3.91	2.86	3.07	4.78	7.16	6.49	8.87
	$\Delta_{\boldsymbol{\Gamma}}$	0.20	0.21	0.24	0.41	0.31	0.53	0.22	0.24	0.38	0.59	0.54	0.79
SPLS	$\Delta_{\boldsymbol{\beta}}$	–	–	–	–	–	–	3.02	3.06	3.80	2.93	2.65	3.85
	$\Delta_{\boldsymbol{\Gamma}}$	–	–	–	–	–	–	0.60	0.71	0.31	0.23	0.73	0.74
SPCR	$\Delta_{\boldsymbol{\beta}}$	3.21	3.21	3.50	3.51	1.30	1.34	6.93	6.95	8.49	8.52	1.75	2.00
	$\Delta_{\boldsymbol{\Gamma}}$	0.58	0.58	0.58	0.58	0.06	0.07	0.58	0.58	0.58	0.58	0.13	0.15
SNIECE	$\Delta_{\boldsymbol{\beta}}$	2.35	2.37	2.62	2.63	1.30	1.34	2.83	2.88	2.10	2.24	1.75	1.89
	$\Delta_{\boldsymbol{\Gamma}}$	0.19	0.19	0.15	0.15	0.06	0.07	0.20	0.21	0.17	0.17	0.13	0.14
S.E.$\leq$	$\Delta_{\boldsymbol{\beta}}$	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.17)	(0.13)	(0.02)	(0.02)	(0.05)	(0.07)
	$\Delta_{\boldsymbol{\Gamma}}$	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)	(0.01)

		M3 (Logistic regression)						**M4** (Cox hazards model)					
$\boldsymbol{\Sigma}_A$		$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$		$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$	
p		400	1600	400	1600	400	1600	400	1600	400	1600	400	1600
PMLE	$\Delta_{\boldsymbol{\beta}}$	7.10	7.32	8.11	8.24	9.11	9.15	5.50	5.76	7.94	7.99	10.50	10.61
PCR	$\Delta_{\boldsymbol{\beta}}$	5.86	7.68	9.37	9.51	3.17	5.28	6.15	6.61	9.68	9.70	1.43	2.64
	$\Delta_{\boldsymbol{\Gamma}}$	0.65	0.76	0.63	0.73	0.23	0.42	0.59	0.63	0.59	0.62	0.10	0.20
SupPCR	$\Delta_{\boldsymbol{\beta}}$	8.26	12.55	11.36	14.24	19.92	26.93	3.62	3.67	9.57	10.51	11.98	11.93
	$\Delta_{\boldsymbol{\Gamma}}$	0.73	0.74	0.85	0.89	0.94	0.96	0.67	0.67	0.84	0.85	0.86	0.86
NIECE	$\Delta_{\boldsymbol{\beta}}$	5.35	7.66	5.14	7.72	3.19	5.28	2.59	4.40	2.46	4.36	1.43	2.64
	$\Delta_{\boldsymbol{\Gamma}}$	0.54	0.75	0.39	0.64	0.23	0.42	0.24	0.43	0.18	0.33	0.10	0.20
SPCR	$\Delta_{\boldsymbol{\beta}}$	4.34	4.22	9.42	9.47	1.81	1.64	5.82	5.72	9.72	9.74	0.84	0.81
	$\Delta_{\boldsymbol{\Gamma}}$	0.58	0.59	0.58	0.58	0.08	0.08	0.58	0.58	0.58	0.58	0.05	0.05
SNIECE	$\Delta_{\boldsymbol{\beta}}$	2.22	2.37	1.83	1.86	1.65	1.51	1.27	1.30	1.16	1.25	0.84	0.81
	$\Delta_{\boldsymbol{\Gamma}}$	0.17	0.19	0.11	0.11	0.07	0.07	0.10	0.10	0.08	0.08	0.05	0.04
S.E.$\leq$		(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

$$\leq \frac{2\sqrt{u}\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\text{op}}}{\Delta}, \tag{31}$$

where the index sets $\mathcal{J}_1$ is for distinct eigenvalues and index set $\mathcal{J}_2$ is for common eigenvalues. The conclusion follows from $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\text{op}} \leq \epsilon$. □

Supplementary Material

Supplementary material for "Envelopes and principal component regression"
(doi: 10.1214/23-EJS2154SUPP; .pdf). The supplementary material contains detailed proofs of lemmas and theorems and is provided in a separate file.

References

Absil, P.-A., Mahony, R. and Sepulchre, R. (2009). *Optimization Algorithms on Matrix Manifolds*. Princeton University Press. MR2364186

Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* **101** 119–137. MR2252436

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics* **39** 3092–3120. MR3012402

Bro, R., Kjeldahl, K., Smilde, A. K. and Kiers, H. (2008). Cross-validation of component models: a critical look at current methods. *Analytical and Bioanalytical Chemistry* **390** 1241–1251.

Bühlmann, P., Kalisch, M. and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology.

Cai, T. T., Ma, Z., Wu, Y. et al. (2013). Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics* **41** 3074–3110. MR3161458

Chen, K., Chan, K.-S. and Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 203–221. MR2899860

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association* **107** 1533–1545. MR3036414

Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics* **38** 3696–3723. MR2766865

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 3–25. MR2751241

Chun, H., Ballard, D. H., Cho, J. and Zhao, H. (2011). Identification of association between disease and multiple markers via sparse partial least-squares regression. *Genetic Epidemiology* **35** 479–486.

Conway, J. (1990). *A Course in Functional Analysis*. 2nd edition. Springer, New York. MR1070713

Cook, R. D. (2018). *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics* **401**. John Wiley & Sons. MR3774758

Cook, R. D. (2020). Envelope methods. *Wiley Interdisciplinary Reviews: Computational Statistics* **12** e1484. MR4072465

Cook, R. D., Forzani, L. and Zhang, X. (2015). Envelopes and reduced-rank regression. *Biometrika* **102** 439–456. MR3371015

Cook, R. D., Forzani, L. and Su, Z. (2016). A note on fast envelope estimation. *Journal of Multivariate Analysis* **150** 42–54. MR3534901

Cook, R. D., Forzani, L. et al. (2019). Partial least squares prediction in high-dimensional regression. *The Annals of Statistics* **47** 884–908. MR3909954

Cook, R. D., Helland, I. S. and Su, Z. (2013a). Envelopes and partial least squares regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 851–877. MR3124794

Cook, R., Helland, I. and Su, Z. (2013b). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 851–877. MR3124794

Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica* **20** 927–960. MR2729839 (2012a:62186)

Cook, R. D. and Su, Z. (2013). Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression. *Biometrika* **100** 939–954. MR3142342

Cook, R. D. and Zhang, X. (2015a). Simultaneous envelopes for multivariate linear regression. *Technometrics* **57** 11–25. MR3318345

Cook, R. D. and Zhang, X. (2015b). Foundations for envelope models and methods. *Journal of the American Statistical Association* **110** 599–611. MR3367250

Cook, R. D. and Zhang, X. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics* **25** 284–300. MR3474048

Cook, R. D. and Zhang, X. (2018). Fast envelope algorithms. *Statistica Sinica* **28** 1179–1197. MR3821000

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34** 87–22. MR0341758

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509

De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18** 251–263.

Ding, S. and Cook, R. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **80** 387–408. MR3763697

Ding, S., Su, Z., Zhu, G. and Wang, L. (2021). Envelope quantile regression. *Statistica Sinica* **31** 79–106. MR4270379

Edelman, A., Arias, T. A. and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications* **20** 303–353. MR1646856

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likeli-

hood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association* **79** 892–898. MR0770284

Flury, B. (1988). *Common Principal Components & Related Multivariate Models*. John Wiley & Sons, Inc. MR0986245

Franks, A. (2020). Reducing subspace models for large-scale covariance regression. *arXiv preprint arXiv:2010.00503*. MR4534382

Franks, A. M. and Hoff, P. (2019). Shared subspace models for multigroup covariance estimation. *Journal of Machine Learning Research* **20** 1–37. MR4048982

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.* **17** 97–114. MR1085924 (92e:62108)

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

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013). Oracle inequalities for the lasso in the Cox model. *Annals of Statistics* **41** 1142. MR3113806

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **31** 300–303. MR0841268

Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal Component Analysis* 129–155. Springer. MR0841268

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Science & Business Media. MR2036084

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374** 20150202. MR3479904

Josse, J. and Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis* **56** 1869–1879. MR2892383

Khare, K., Pal, S., Su, Z. et al. (2017). A bayesian approach for envelope models. *The Annals of Statistics* **45** 196–222. MR3611490

Lang, W. and Zou, H. (2020). A simple method to improve principal components regression. *Stat* e288. MR4116322

Lee, M. and Su, Z. (2020). A review of envelope models. *International Statistical Review* **88** 658–676. MR4180672

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association* **112** 1131–1146. MR3735365

Li, G., Yang, D., Nobel, A. B. and Shen, H. (2016). Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis* **146** 7–17. MR3477645

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding.

Ann. Statist. **41** 772–801. MR3099121

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *The Annals of Statistics* **27** 538–557. MR3025133

Nygård, S., Borgan, Ø., Lingjærde, O. C. and Størvold, H. L. (2008). Partial least squares Cox regression for genome-wide data. *Lifetime Data Analysis* **14** 179–195. MR2398971

Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks* **5** 927–935.

Sæbø, S., Almøy, T., Aarøe, J. and Aastveit, A. H. (2008). ST-PLS: a multi-directional nearest shrunken centroid type classifier via PLS. *Journal of Chemometrics: A Journal of the Chemometrics Society* **22** 54–62.

Schott, J. R. (1999). Partial common principal component subspaces. *Biometrika* **86** 899–908. MR1741985

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis* **99** 1015–1034. MR2419336

Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika* **98** 133–146. MR2804215

Su, Z. and Cook, R. D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika* **99** 687–702. MR2966778

Su, Z., Zhu, G., Chen, X. and Yang, Y. (2016). Sparse envelope model: Efficient estimation and response variable selection in multivariate linear regression. *Biometrika* **103** 579–593. MR3551785

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. MR1379242

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61** 611–622. MR1707864

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics* **41** 2905–2947. MR3161452

Welling, M., Williams, C. and Agakov, F. V. (2004). Extreme components analysis. In *Advances in Neural Information Processing Systems* 137–144.

Wen, Z. and Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming* **142** 397–434. MR3127080

Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* kxp008.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis* 391–420. MR0220397

Yu, Y., Wang, T. and Samworth, R. J. (2014). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006

Zhang, J. and Chen, X. (2020). Principal envelope model. *Journal of Statistical Planning and Inference* **206** 249–262. MR4036706

Zhang, X. and Mai, Q. (2018). Model-free envelope dimension selection. *Elec-*

tronic Journal of Statistics **12** 2193–2216. MR3829139

Zhang, X. and Mai, Q. (2019). Efficient integration of sufficient dimension reduction and prediction in discriminant analysis. *Technometrics* **61** 259–272. MR3957146

Zhou, F., Claire, Q. and King, R. D. (2014). Predicting the geographical origin of music. In *2014 IEEE International Conference on Data Mining* 1115–1120. IEEE.

Zhou, L., Cook, R. D. and Zou, H. (2020). Enveloped Huber regression. *arXiv preprint arXiv:2011.00119*.

Zhu, G. and Su, Z. (2020). Envelope-based sparse partial least squares. *The Annals of Statistics* **48** 161–182. MR4065157

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15** 265–286. MR2252527