Envelopes and principal component regression

Xin Zhang,* Kai Deng, and Qing Mai

Florida State University

Abstract

Envelope methods offer targeted dimension reduction for various models. The overarching 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 ultra 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 novel principal components 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 NIECE 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.

Keywords: Envelope Methods; Penalized Matrix Decomposition; Principal Component; Sufficient Dimension Reduction.

*Correspondence: Xin Zhang (xzhang8@fsu.edu)
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 et al. (2010), envelope methods have demonstrated promising performances in various multivariate statistical problems. Different envelope structures are proposed in the multivariate linear regression model, for example, partial envelope (Su & Cook 2011), inner envelope (Su & Cook 2012), scaled envelope (Cook & Su 2013), and simultaneous envelope (Cook & Zhang 2015b). Connections between envelopes and classical multivariate analysis methods are intensively studied: envelope and partial least squares (Cook et al. 2013a), envelope and reduced-rank regression (Cook et al. 2015), envelope and Bayesian statistics (Khare et al. 2017), envelope and discriminant analysis (Zhang & Mai 2019), and among others. Recent extensions of envelopes beyond the standard linear models is also an emerging field of research. Cook & Zhang (2015a) 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 et al. 2020), matrix-variate (Ding & Cook 2018) and tensor-variate regressions (Li & Zhang 2017). See Cook (2018) for a detailed introduction, and Cook (2020) and Lee & 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 $M$ and $U$ in a model-free setting of envelopes: For NIECE, we extract the eigenvectors $v_i$ of $M$ and examine the “eigen-gaps” of the quadratic form $v_i^T U v_i$.

In what follows, we outline in Section 1.1 the population construct of an envelope $E_M(U)$ following the same notation in Cook & Zhang (2015a) 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 Section 1.2. The population-level connections between envelopes and principal components are given in Section 1.3 to motivate our methodology. The specific goals and organization of this article are outlined in Section 1.4.
1.1 Envelopes: definition and working mechanism

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

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

![Figure 1: The working mechanism of envelope methodology when estimating the mean difference of $X$ across two groups (indicated by red and blue colors). The plot illustrates the efficiency gain by envelope methods in estimating the mean difference $\theta = \mu_2 - \mu_1$; representative projection paths, labeled ‘E’ for envelope analysis and ‘S’ for standard analysis.](image)

The construction and estimation of envelope $\mathcal{E}_M(\mathcal{U})$ is the center of our future development. Cook et al. (2010) showed that the envelope $\mathcal{E}_M(\mathcal{U})$ is unique and always exists. In multivariate analysis, $\mathcal{U}$ comes from the parameter of interest; and $M$ represents a helpful nuisance parameter. The specific choices of $M$ and $\mathcal{U}$ 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 \( X = (X_1, X_2)^T \in \mathbb{R}^2 \). The target parameter is the mean difference \( \theta = \mu_2 - \mu_1 = E(X | G = 2) - E(X | G = 1) \) or equivalently the matrix form \( \Theta = \theta \theta^T \) (i.e. the \( U \) matrix in Definition 1). The nuisance parameter is the marginal covariance \( \Sigma_X = \text{cov}(X) \) (i.e. the \( M \) matrix in Definition 1); and it helps improving the estimation efficiency in the mean difference as illustrated in Figure 1. The two ellipses represent the contours of the conditional distributions of \( X \) 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 Figure 1 we can see that the two empirical distributions of \( X_1 | G = 1 \) and \( X_1 | G = 2 \), 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 \( X \) to compare the two classes. The elimination of immaterial variation leads to much well-separated two distributions as shown in Figure 1 and therefore massive gain in estimation accuracy of \( \theta = \mu_2 - \mu_1 \).

1.2 A brief overview of envelope estimation

Without loss of generality, we assume \( U \) is symmetric positive semi-definite henceforth because we can always replace \( U \) with \( UU^T \geq 0 \) without changing the column subspace \( \text{span}(U) \) and the envelope \( E_M(U) \). The dimension of \( E_M(U) \), denoted by \( u \), \( 0 \leq u \leq p \), is important for all envelope methods and can be estimated consistently (Zhang & 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 \( \hat{\Gamma} \in \mathbb{R}^{p \times u} \) from the following constrained optimization and letting \( \hat{E}_M(U) = \text{span}(\hat{\Gamma}) \),

\[
\hat{\Gamma} = \arg\min_{\Gamma \Gamma^T = I_u} J_n(\Gamma), \quad J_n(\Gamma) = \log |\Gamma^T \hat{M} \Gamma| + \log |\Gamma^T (\hat{M} + \hat{U})^{-1} \Gamma|,
\]

where the symmetric matrices \( \hat{M} > 0 \) and \( \hat{U} \geq 0 \) are finite sample estimators of their population counterparts \( M \) and \( U \). From the orthogonality constraint \( \Gamma^T \Gamma = I_u \), 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 & Zhang (2015a) 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, \( \theta \in \mathbb{R}^p \), and some standard \( \sqrt{n} \)-consistent estimator \( \hat{\theta} \), the particular choice of \( U = \theta \theta^T \) and \( M \) being the asymptotic covariance of \( \hat{\theta} \) reproduces these likelihood-based envelope methods in the literature. The envelope estimator \( \hat{\theta}_{\text{env}} = \hat{\Gamma} \hat{\Gamma}^T \hat{\theta} \) is asymptotically more efficient than the standard estimator \( \hat{\theta} \) in various contexts such as linear and generalized linear models. Such envelope estimation solves for \( \hat{\Gamma} \) based on (1) and then plugs-in \( \hat{\theta}_{\text{env}} = \hat{\Gamma} \hat{\Gamma}^T \hat{\theta} \), 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 et al. (1998), Absil et al. (2009) and Wen & Yin (2013) for more background. By exploiting the geometry of envelopes, Cook & Zhang (2016) developed the 1D algorithm that approximately solves (1) by sequential optimization over one-dimensional vectors than over \( p \times u \) matrices. Relatedly, Cook & 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 et al. (2016). The authors suggest to remove the orthogonality constraint by rotating \( \Gamma \) to as \( \Gamma = (I_u, A^T)_G \) for some coordinates \( A \in \mathbb{R}^{(p-u)\times u} \) and a full rank matrix \( G \in \mathbb{R}^{u \times u} \). Then the iterative Grassmannian optimization is transformed into unconstrained iterative optimization over matrix \( A \).

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 \( \hat{\Theta}_{\text{env}} \) and thus makes the envelope estimator \( \hat{\Theta}_{\text{env}} = \Gamma \Gamma^T \hat{\Theta} \) 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 & 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 \( \Gamma \). This type of coordinate-free penalty (Chen et al. 2010) takes the form of \( P_{\lambda}(\Gamma) = \lambda_i \|\gamma_i\|_2, i = 1, \ldots, p \), where \( \gamma_i \in \mathbb{R}^n \) is the \( i \)-th row of \( \Gamma \). 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 & 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 & Bishop (1999) provided a model-based probabilistic formulation for principal component analysis where only the leading PCs are relevant in the analysis; Welling et al. (2004) generalized such approach to probabilistic models where the maximum likelihood estimation for the dimension reduction sub-
space is a combination of PCs with largest and smallest variances; Zhang & 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 & Zou (2020) provided a response-guided formulation for regression with PCs that unifies the usual principal component regression with ridge regression (Hoerl & 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 $M$ is a reducing subspace of $M$. In this article, to ensure identifiability, we assume that the first $d \ll n$ eigenvalues of $M$ are distinct and their span contains $\text{span}(U)$. In practice, we will specify $d \approx n$ 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 $\mathcal{E}_M(U)$, similar to Proposition 4 in Cook & Zhang (2018).

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

$$\mathcal{E}_M(U) = \sum_{i=1}^{d} \text{span}(v_i),$$

where $v_i \in \mathbb{R}^p$ is the $i$-th eigenvector of $M = \Sigma_X$ 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 (Witten et al. 2009, PMD). 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 et al. 2006, Shen & Huang 2008, Amini & Wainwright 2009, Ma 2013, Vu & 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 $\Sigma_X = V\Sigma_f V^T + \sigma^2 I_p$, where $\Sigma_f \in \mathbb{R}^{u \times u}$ is the covariance of the $u$ latent variables $f$ and $V \in \mathbb{R}^{p \times u}$ is the basis of the $u$-dimensional target subspace. Franks & Hoff (2019) studied a shared subspace spiked model that $\Sigma_{X|Y=k} = \sigma_k^2 (V \Psi_k V^T + I_p)$ for groups of observations indexed by $k = 1, \ldots, K$, where $\Psi_k$ symmetric positive definite matrix and $V \in \mathbb{R}^{p \times u}$ 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 & 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 et al. (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 & Su 2020). Our method can thus also be viewed as an extension of the sparse PLS methods (Chun & Keles 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 $M$, we will let $\lambda_j(M)$ be the $j$-th eigenvalue of $M$, e.g. $\lambda_1(M)$ is the largest eigenvalue of $M$. We use the Frobenius norm, operator norm, $\ell_1$ norm and the maximum norm of matrices. For a matrix $\Omega \in \mathbb{R}^{p_1 \times p_2}$, let $\omega_{ij}$ denote its $(i, j)$-th element, then its Frobenius norm is $\|\Omega\|_F = \sqrt{\sum_{i,j} \omega_{ij}^2}$; its operator norm $\|\Omega\|_{op}$ is its largest eigenvalue; its $\ell_1$ norm is $\|\Omega\|_1 = \max_i \sum_j |\omega_{ij}|$; and its maximum norm is $\|\Omega\|_{\max} = \max_{i,j} |\omega_{ij}|$.

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

**Lemma 1.** $\|P_\mathcal{E} - P_{\widehat{\mathcal{E}}}\|_F = \sqrt{2} \| \sin \Theta(\hat{\Gamma}, \Gamma) \|_F$.

Unless otherwise specified, we will be using $\|P_\mathcal{E} - P_{\widehat{\mathcal{E}}}\|_F$ 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 Section 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 (Witten et al. 2009; PMD). 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. Sections 5 and 6, respectively, contain a simulation study and a real data example. 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 M and 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 $M > 0$ and $U \geq 0$; number of principal components $d$; envelope dimension $u$. Note that $0 \leq u \leq d \leq p$.

2. Obtain the first $d$ eigenvectors of $M$: $V_d = (v_1, \ldots, v_d) \in \mathbb{R}^{p \times d}$ that satisfies $V_d^T V_d = I_d$. The corresponding $d$ eigenvalues are $\lambda_1(M) \equiv v_1^T M v_1 > \cdots > \lambda_d(M) \equiv v_d^T M v_d > 0$.

3. Calculate the envelope scores: $\phi_j \equiv v_1^T U v_j$ for $j = 1, \ldots, d$, and organize them in descending order $\phi_1 \geq \cdots \geq \phi_d$ and define $v_{(j)}$ such that $\phi_{(j)} = v_{(j)}^T U v_{(j)}$.

4. Output: envelope is $E_M(U) = \text{span}(v_{(1)}, \ldots, v_{(u)})$.

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

Lemma 2. Suppose that $E_M(U) \subseteq \text{span}(V_d)$ for some $d \leq p$ and that $\lambda_j(M) \neq \lambda_k(M)$ for any $j, k \in \{1, \ldots, d\}$ and $j \neq k$. Then (2) holds, the envelope can be written as $E_M(U) = \text{span}(v_{(1)}, \ldots, v_{(u)})$, 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 et al. (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 $E_M(U) \subseteq \text{span}(V_d)$ 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 $E_M(U)$ is not estimable. Intuitively, this is because there is only a limited number of eigenvectors of $M$ estimable in high dimensions. In all our theoretical studies, we treat $d$ as fixed for simplicity and allow both $n, p \rightarrow \infty$.

Second, the distinct eigenvalue assumption that $\lambda_j(M) \neq \lambda_k(M)$ 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 et al. 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 & Lei 2013) instead of vectors. We define the minimal eigen-gap of the first $d$ eigenvalues of $M$ as follows,

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

Unless otherwise noted, we will make the assumptions as in Lemma 2. The envelope is thus constructed by the eigenvectors $v_{(1)}, \ldots, v_{(u)}$. 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 $v_{(j)}$ is $\pi(j)$-th eigenvector of $M$ and $\lambda_{\pi(j)}(M) = v_{(j)}^T M v_{(j)}$. 


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

$$\Delta_U \equiv \phi(u) - \phi(u+1) = v_{(u)}^T Uv_{(u)} > 0.$$  \hspace{1cm} (4)

In the next section, we show how the minimal eigen-gap $\Delta$ and the envelope score gap $\Delta_U$ would affect the subspace estimation in finite sample.

### 2.2 Theory for the sample NIECE algorithm

For the sample NIECE procedure, we simply replace $M$ and $U$ with their sample counterparts (e.g. sample covariance matrices) and let $\widehat{v}_{(j)}$, $j = 1, \ldots, u$, be the $\widehat{\pi}(j)$-th eigenvectors of $\widehat{M}$, where $\widehat{\pi}(\cdot)$ is the permutation of indices in $\{1, \ldots, d\}$ according the the estimated envelope scores. Specifically, we can write $\lambda_{\widehat{\pi}(j)}(\widehat{M}) = \widehat{v}_{(j)}^T \widehat{M}v_{(j)}$. 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.

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 $\hat{U}$ and $\hat{M}$; (iii) the minimal eigen-gap $\Delta$ defined in (3); (iv) the envelope score gap $\Delta_U$ defined in (4); and (v) the largest eigenvalue of $U$ denoted as $\nu = \|U\|_{op}$. The following theory is for the analysis in relatively low-dimensional data, and it is not easy to find estimates satisfying (5) in high dimensions.

**Theorem 1.** For any $\epsilon > 0$ and $\epsilon \leq \frac{\Delta}{4}$, if we have

$$\|\hat{M} - M\|_{op} \leq \epsilon, \quad \|\hat{U} - U\|_{op} \leq \epsilon, \quad \left(\frac{\Delta U}{\Delta} \right)^{1 - \nu} \frac{2\sqrt{2}\nu}{\Delta} \epsilon,$$

then $\|P_{\hat{E}} - P_E\|_F \leq \frac{2\sqrt{2}\nu}{\Delta} \epsilon$, where $\hat{E}$ and $E$ are the envelope $\hat{E}_M(U)$ and $E_M(U)$, respectively.

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


In the Supplementary Materials, Lemma [S4] we consider the ideal setting where \( \{ \hat{\pi}(j) \mid j = 1, \ldots, u \} = \{ \pi(j) \mid j = 1, \ldots, u \} \). Then we may replace the minimal eigen-gap \( \Delta \) of the \( d \) eigenvalues of \( \hat{M} \) 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 \( \Delta_j = \min\{ \lambda_{\pi(j)-1}(M) - \lambda_{\pi(j)}(M), \lambda_{\pi(j)}(M) - \lambda_{\pi(j)+1}(M) \} \) 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 \( \Delta \), 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 \( E \) 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 & Zhang 2015b). The simultaneous envelopes are constructed, by jointly estimating the response envelope (Cook et al. 2010) and the predictor envelope (Cook et al. 2013a), to simultaneous reduce the multivariate response \( \hat{Y} \in \mathbb{R}^r \) and predictor \( \hat{X} \in \mathbb{R}^p \) in the following regression model,

\[
\hat{Y} = \alpha + \beta \hat{X} + \varepsilon,
\]

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

In the classical likelihood-based envelope methods, the response is reduced by projecting onto the response envelope \( E_\Sigma(\beta) \), whose likelihood-based estimation is derived from the normal error \( \varepsilon \sim N(0, \Sigma) \). On the other hand, the predictor reduction is achieved by projection onto the predictor envelope \( E_{\Sigma X}(\beta^T) \), whose likelihood-based estimation is derived from further assuming \( \hat{X} \sim N(0, \Sigma_X) \). Thus, for predictor reduction and simultaneous reduction, the joint normality of \( (\hat{X}^T, \hat{Y}^T)^T \) 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 \( w \in \mathbb{R}^{p+r} \), \( \|w\|_2 = 1 \), we have

\[
P(|w^T \begin{pmatrix} \hat{X} \\ \hat{Y} \end{pmatrix} | \geq t) \leq 2 \exp(-\frac{t^2}{\sigma^2}).
\]

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 \( E_\Sigma(\beta) = E_{\Sigma Y}(\Sigma_{YX}) \), where \( \Sigma_Y \in \mathbb{R}^{r \times r} \) is the covariance matrix of \( \hat{Y} \) and \( \Sigma_{YX} \equiv \Sigma_{XY}^T \in \mathbb{R}^{r \times p} \) is the cross-covariance matrix. For predictor envelope, we exploit the symmetry by noticing \( E_{\Sigma X}(\beta^T) = E_{\Sigma X}(\Sigma_{XY}) \). Given the data, the NIECE estimator is obtained by using the sample covariances \( \hat{M} = \hat{\Sigma}_Y \) and \( \hat{U} = \hat{\Sigma}_{XY} \hat{\Sigma}_{XY} \) for response reduction, and \( \hat{M} = \hat{\Sigma}_X \) and \( \hat{U} = \hat{\Sigma}_{XY} \hat{\Sigma}_{XY} \) for predictor reduction.

We have the following result that applies to both response and predictor envelopes (i.e. simultaneous envelopes), where \( E_M(U) = E_{\Sigma Y}(\Sigma_{YX}) \subseteq \mathbb{R}^r \) for response envelope and is re-defined as \( E_M(U) = E_{\Sigma X}(\Sigma_{XY}) \subseteq \mathbb{R}^p \) for predictor envelope. Other quantities such as \( \hat{E}, \Delta, \Delta_U, u \) and \( \nu \) also alters when we switch from response envelope to predictor envelope.
Theorem 2. For any $\epsilon > 0$ such that $\epsilon \leq \max \left\{ \frac{\Delta}{4}, \frac{\Delta_U}{2} \left( 1 + \frac{10\nu}{\Delta} \right)^{-1} \right\}$ and $\epsilon \leq \|\Sigma_{XY}\|_1 + \|\Sigma_{YX}\|_1 < c_1$ for some constant $c_1$, we have

$$\|P_\hat{E} - P_E\|_F \leq \frac{2\sqrt{2}\mu \epsilon}{\Delta},$$

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

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 $\max \left( pr\sqrt{\log p + \log r}, pr\Delta_U^{-1} \sqrt{\log p + \log r} \right) = o(\sqrt{n})$ and $\|\Sigma_{XY}\|_1 + \|\Sigma_{YX}\|_1 < c_1$ for some constant $c_1$, then $\|P_\hat{E} - P_E\|_F \to 0$ 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 et al. 2016) and ECD (Cook & 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 et al. 1998, 2011).
and sequential 1D algorithm (Cook & 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 \( \hat{\Theta} = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T \), \( \hat{U} = \Delta_{\hat{U}} \cdot \Gamma \Phi \Gamma^T \), where \( \Gamma = (v_2, v_3, v_{10}, v_{11}, v_{19}) \) and \( \Omega \) and \( \Omega_0 \) are diagonal matrix of eigenvalues \( \lambda_k = k^3 \) for \( k = 1, \ldots, 20 \) and \( \lambda_k = 0.05 \) for \( k > 20 \), \( \Phi = O_u DO_u^T \) with orthonormal matrix \( O_u \in \mathbb{R}^{u \times u} \) and \( D = \text{diag}(1, \ldots, u) \). We tried three different signal strength settings with \( \Delta_{\hat{U}} = 0.01, 1, \) and 100. For each setting, we randomly generated 100 replicates of \( \hat{M} \) and \( \hat{U} \) from Wishart distributions with degrees of freedom \( n \) and means \( M \) and \( U \) to mimic the sample covariances in regression models. We reported the subspace estimation error \( D(\hat{\Theta}, \Theta) = D(\hat{\Theta}, \Theta) = \|P_e - P_{\hat{\Theta}}\|_F / \sqrt{2u} \), 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 \( \Delta \) in \( M \) 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 \( \Delta_{\hat{U}} \).

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 \( \hat{M} \) 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 \( \hat{M} = X_n^T X_n \) from some data matrix \( X_n \in \mathbb{R}^{n \times p} \). As we have seen from the previous Section, the data matrix \( X_n \) can be either the \( n \) samples of \( X \) or \( Y \). For envelopes in GLM and Cox model, \( X_n \) is also the \( n \) samples of \( X \). Another example is the weighted least squares envelope (Cook & Zhang 2015a), where the \( i \)-th row of \( X_n \) is the squared-root weight \( \sqrt{W_i} \) times the predictor vector \( \tilde{X}_i \).

For simplicity, the data matrix \( X_n \in \mathbb{R}^{n \times p} \) is henceforth defined as the \( n \) i.i.d. samples of mean zero random variable \( X \in \mathbb{R}^p \). We obtain the \( d \) sparse eigenvectors of \( \hat{M} \geq 0 \) based on the penalized matrix decomposition (PMD) method from Witten et al. (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 \( X_n \) instead of the much bigger matrix \( \hat{M} \) as follows.

Let \( X^1 = X_n \) be the original data matrix. For \( k = 1, \ldots, d \), we sequentially solve for

\[
(\hat{u}_k, \hat{v}_k) = \underset{u \in \mathbb{R}^n, v \in \mathbb{R}^p}{\text{argmax}} \ u^T X^k v \quad \text{subject to} \quad \|v\|_1 \leq c, \ |\|v\|_2^2 \leq 1, \ |\|u\|_2^2 \leq 1; \tag{9}
\]

and then deflate the data matrix as \( X^{k+1} = X^k - \hat{\sigma}_k \hat{u}_k \hat{v}_k^T \), where \( \hat{\sigma}_k = \hat{u}_k^T X^k \hat{v}_k \).

To solve (9), we need alternating updates between \( u \in \mathbb{R}^n \) and \( v \in \mathbb{R}^p \). Since there is no penalty on \( u \), we can see that the solution for \( \hat{u}_k \) has to be

\[
\hat{u}_k = \frac{X^k \hat{v}_k}{\|X^k \hat{v}_k\|_2}, \tag{10}
\]
which leads to the following equivalent presentation.

**Lemma 3.** The optimization in (9) is equivalent to (10) and

\[ \hat{v}_k = \arg \max_v v^T \hat{M}_k v \quad \text{subject to } \|v\|_1 \leq \tau \text{ and } \|v\|_2^2 \leq 1, \]

where \( \hat{M}_k = (X^k)^T X^k \) and \( X^k = X^{k-1}(I_p - \hat{v}_{k-1} \hat{v}_{k-1}^T) \).

As noted in [Witten et al. (2009)], by the Karush-Kuhn-Tucker conditions in convex optimization, the solution to (11) is also the solution to

\[ \hat{v}_k = \arg \max_v v^T \hat{M}_k v \quad \text{subject to } \|v\|_1 \leq c, \text{ and } \|v\|_2^2 = 1, \]

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

### 3.2 A general high-dimensional theory

Notice that we have re-defined \( \hat{v}_k \) as the PMD solution for the high-dimensional sparse NIECE. Moreover, we re-define \( \hat{V}_d = (\hat{v}_1, \ldots, \hat{v}_d) \) and other estimates for ease of presentation. However, it is important to note that \( \hat{v}_k \) is no longer the eigenvector of \( \hat{M} \) and that orthogonality among eigenvectors no longer holds, i.e. \( \hat{V}_d^T \hat{V}_d \neq I_d \). Furthermore, \( \hat{\sigma}_k^2 = \hat{v}_k^T \hat{M} \hat{v}_k \) is different from the eigenvalue \( \lambda_k(\hat{M}) \). Due to the non-orthogonality among \( \hat{v}_k \)'s and related issues, the deflation of data matrix is not easy to handle. Nevertheless, we are able to show consistency of \( \hat{v}_1, \ldots, \hat{v}_d \) by analyzing \( X^k = X_n \prod_{j=1}^{k-1}(I_d - \hat{v}_j \hat{v}_j^T) \) 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 \( v_k \) (e.g. sparsity in principal component loadings) is imposed as follows.

\[ \tau_0 = \max_{1 \leq k \leq d} \|v_k\|_1. \]

where a small value of \( \tau_0 \) implies that the first \( d \) eigenvectors of \( M \) are all reasonably sparse in high dimensions.

**Theorem 3.** If \( \tau > \tau_0 \) and \( \|\hat{M} - M\|_{\max} \leq \epsilon \), then there exists a constant \( c_0 > 0 \) that does not depend on \( n, p \) or \( \epsilon \), such that,

\[ \sin^2 \Theta(v_k, \hat{v}_k) \leq c_0 \epsilon \tau^2, \quad k = 1, \ldots, d. \]

Hence, to show the consistency of NIECE in high-dimensional settings, we only need to verify \( \|\hat{M} - M\|_{\max} \leq \epsilon \), which is weaker than the assumption of \( \|\hat{M} - M\|_{\op} \leq \epsilon \) in the non-sparse settings. This is because we apply regularization to obtain accurate estimates of \( v_k \) in high dimensions. Also, note that \( \|\hat{M} - M\|_{\op} \leq \epsilon \) generally cannot hold in high dimensions, while later we will show that \( \|\hat{M} - M\|_{\max} \leq \epsilon \) holds with a probability tending to 1 under mild conditions.
Theorem 4. Assume that $\tau > \tau_0$, $\|\hat{M} - M\|_{\text{max}} \leq \epsilon$ and $\|\hat{U} - U\|_{\text{max}} \leq \epsilon$, then if $\Delta_U > \sqrt{8c_0 \nu^2 \tau^2 \epsilon} + (2c_0 \nu + 1) \tau^2 \epsilon$ with constant $c_0$ defined in (14), there exists a constant $C > 0$ that does not depend on $n$, $p$ or $\epsilon$ such that $\|P_{\hat{E}} - P_E\|_F \leq C\epsilon \tau^2$.

Theorem 4 is a general result that guarantees that as long as we start with a reasonably accurate estimator $\hat{M}$, 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 (6) with jointly sub-Gaussian data (7). 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 $\hat{M} = \hat{\Sigma}_Y$ and $\hat{U} = \hat{\Sigma}_Y \hat{\Sigma}_X \hat{\Sigma}_X$. In high-dimensional sparse setting, e.g. the eigenvectors of $M = \Sigma_Y$ satisfy (13), we have the following results.

**Theorem 5.** For any $\epsilon$ such that $0 < \epsilon \leq 1/p$, assume that $\tau > \tau_0$ and $\Delta_U > \sqrt{8c_0 \nu^2 \tau^2 \epsilon} + (2c_0 \nu + 1) \tau^2 \epsilon$ with constant $c_0$ defined in (14), then there exists a constant $C > 0$ that does not depend on $n$, $r$, $p$ or $\epsilon$ such that $\|P_{\hat{E}} - P_E\|_F \leq C\epsilon \tau^2$ (15) 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 $\Delta_U$ and $\tau$ satisfy that $\max(\tau^2 \sqrt{\log p + \log r}, \Delta_U^{-2} \tau^2 \sqrt{\log p + \log r}) = o(\sqrt{n})$, then we have $\|P_{\hat{E}} - P_E\|_F \to 0$ 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 $\Delta_U$ and $\tau$ satisfy that $\max(\tau^2 \sqrt{\log r}, \Delta_U^{-2} \tau^2 \sqrt{\log r}) = o(\sqrt{n})$, we have $\|P_{\hat{E}} - P_E\|_F \to 0$ 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 $(\log p)/n \to 0$. 

14
### 3.4 Envelope in generalized linear model

Cook & Zhang (2015a) 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 \((X^T, Y^T)^T\) with the following sub-Gaussian tail assumption on \(X\). Similar to (7), 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 \(w \in \mathbb{R}^p, \|w\|_2 = 1\), we have

\[
P(|w^TX| \geq t) \leq 2 \exp(-\frac{t^2}{\sigma^2}).
\]

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 \(\vartheta\) is set to be \(\vartheta = \beta^TX\) by the canonical link function, \(\varphi\) is the dispersion parameter, and functions \(a, b, c\) 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, \(\varphi\) is set to be 1). As such, we can write the log-likelihood as

\[
\ell_n(\beta) = \sum_{i=1}^n \log f(Y_i \mid \beta, X_i) = \sum_{i=1}^n \{Y_i(\beta^TX_i) - b(\beta^TX_i)\},
\]

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 \(E_{\Sigma_X}(\beta) \subseteq \mathbb{R}^p\) where \(\beta \in \mathbb{R}^{p \times 1}\) 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 \(\ell_n(\beta) = \sum_{i=1}^n \{Y_i(\beta^TX_i) - \log(1 + e^{\beta^TX_i})\}\). For normally distributed \(Y\), the GLM reduces to the predictor envelope with a univariate response. We have shown that \(\hat{\Sigma}_X\) is an accurate estimator for \(\Sigma_X\). For \(\beta\), note that the sparsity of the envelope implies the sparsity of \(\beta\). We can obtain \(\hat{\beta}\) by \(\ell_1\) logistic regression with a tuning parameter \(\lambda_n\),

\[
\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^n \{-Y_i(\beta^TX_i) + \log(1 + e^{\beta^TX_i})\} + \lambda_n\|\beta\|_1,
\]

which is based on minimizing the negative log-likelihood plus an \(\ell_1\)-regularization term. Then the NIECE procedure is based on \(\hat{M} = \hat{\Sigma}_X\) and \(\hat{U} = \hat{\beta}\beta^T\).

**Theorem 6.** If \(\tau > \tau_0, \lambda_n = 4\epsilon/\sqrt{n}\), and \(\Delta_U > \sqrt{8c_0n^2\tau^2\epsilon} + (2c_0\nu + 1)\tau^2\epsilon\) with constant \(c_0\) defined in (14), then there exists a constant \(C > 0\) that does not depend on \(n, p\) or \(\epsilon\) such that

\[
\|P_{\hat{\mathcal{E}}} - P_{\mathcal{E}}\|_F \leq C\tau^2\epsilon
\]
with a probability greater than \(1 - C \rho^p \exp(-Cn_\epsilon^2) - C \rho \exp(-Cn_\epsilon^2/s)\). Furthermore, if \(s \log p = o(n)\) and that \(\lambda_n, \tau\) and \(\Delta_\mathbf{U}\) satisfy \(\left\{ \tau^2 \sqrt{\frac{s \log p}{n}}, \frac{\Delta_\mathbf{U}}{\tau^2} \sqrt{\frac{s \log p}{n}}, \lambda_n^{-1} \sqrt{\frac{\log p}{n}} \right\} = o(1)\), we have \(\|\hat{\mathbf{P}_\mathbf{\hat{U}}} - \mathbf{P}_\mathbf{\hat{U}}\|_F \to 0\) 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 \(h(t \mid X) = h_0(t) \exp(\mathbf{\beta}^T X)\), where \(h_0(t)\) is the baseline hazard function and \(\mathbf{\beta} \in \mathbb{R}^{p \times 1}\) is the regression coefficients of the \(p\)-dimensional covariate vector \(X \in \mathbb{R}^p\). Let \(Y\) and \(C\) be the failure time and censoring time, then we assume \(Y\) and \(C\) are conditionally independent given \(X\). Define \(\delta = I(Y \leq C)\) and \(T = \min(Y, C)\), then the observed data consists of independent copies of \(\{T_i, \delta_i, X_i\}, i = 1, \ldots, n\).

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

\[
\ell_n(\mathbf{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \delta_i \left[ \mathbf{\beta}^T X_i - \log \left\{ \sum_{j=1}^{n} I(T_j \geq T_i) \exp(\mathbf{\beta}^T X_j) \right\} \right].
\]

When the covariates in \(X\) are highly correlated with each other, Cook & Zhang (2015a) proposed to construct envelope estimator

\[
\hat{\mathbf{\beta}}_{\text{Env}} = \arg \min_{\mathbf{\beta} = \Gamma} \ell_n(\mathbf{\beta}) - \frac{1}{n} M_n(\Gamma),
\]

where \(M_n(\Gamma) = -\frac{n}{2} \left( \log |\Gamma^T \mathbf{S}_X \Gamma| + \log |\Gamma^T \mathbf{S}_X^{-1} \Gamma| \right)\) is the partially maximized log likelihood of \(X \sim N(\mu_X, \Sigma_X)\) under the envelope structure \(\text{span}(\Gamma) = \mathcal{E}_{\Sigma_X}(\mathbf{\beta})\). 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 \(\ell_1\)-penalized maximum (partial) likelihood estimator,

\[
\hat{\mathbf{\beta}} = \arg \min_{\mathbf{\beta}} \ell_n(\mathbf{\beta}) + \lambda_n \|\mathbf{\beta}\|_1,
\]

where \(\lambda_n > 0\) is the tuning parameter. Similar to the envelope GLM, the sparsity of the envelope \(\mathcal{E}_{\Sigma_X}(\mathbf{\beta})\) implies the sparsity of \(\mathbf{\beta}\). Then the NIECE procedure is based on \(\hat{M} = \hat{\Sigma}_X\) and \(\hat{\mathbf{U}} = \hat{\mathbf{\beta}} \hat{\mathbf{\beta}}^T\). The convergence of \(\hat{\Sigma}_X\) is guaranteed by sub-Gaussian distribution of \(X\), i.e. (16). Properties of \(\hat{\mathbf{U}}\) is based on the estimation consistency of the Lasso estimator \(\hat{\mathbf{\beta}}\) that is derived in Huang et al. (2013). Similar results for using the SCAD penalty (Fan & Li 2001) instead of Lasso can be found in Bradic et al. (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 $\max_{i \leq n} \|X_i - X_i^\prime\|_\infty \leq C$.

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

These conditions are derived from Theorem 4.1 of [Huang et al. (2013)](https://doi.org/10.1080/01621459.2012.695864), where the introduction of $\Sigma(t, M)$ is rather technical and relegated to the Supplementary Materials. The lower bound $\rho^\ast$ on the smallest eigenvalue of $\Sigma(t, M)$ 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 $C_\rho$ and $C_\lambda$ such that if $s \sqrt{\log p \over n} < C_\rho$ and $\lambda_n = C_\lambda \epsilon$, and $\Delta_U > \sqrt{8c_0 \nu^2 \tau^2 \epsilon} + (2c_0 \nu + 1) \tau^2 \epsilon$ with constant $c_0$ defined in (14), then there exists a constant $C > 0$ that does not depend on $n$, $\tau$, $p$ or $\epsilon$ such that

$$\|P_{\hat{\epsilon}} - P_{\epsilon}\|_F \leq C \tau^2 \epsilon$$

with a probability greater than $1 - C p^2 \exp(-C n \epsilon^2 / C p \exp(-C n \epsilon^2 / s))$. Furthermore, if $s \log p = o(n)$ and that $\lambda_n$, $\tau$ and $\Delta_U$ satisfy $\max \{\tau^2 \sqrt{s \log p}, \Delta_U^{-2} \tau^2 \sqrt{s \log p}, \lambda_n^{-1} \sqrt{\log p}\} = o(\sqrt{n})$, we have $\|P_{\hat{\epsilon}} - P_{\epsilon}\|_F \to 0$ 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 Section 1.2 that the envelope estimator for $\theta \in \mathbb{R}^p$ is the projected estimator $\hat{\theta}_{\text{Env}} = \hat{\Gamma}^\top \hat{\theta}$ that improves over the standard estimator $\hat{\theta}$, where $\hat{\Gamma} \in \mathbb{R}^{p \times u}$ is the estimated envelope basis. In many multivariate parameter estimation problems (see [Cook & Zhang 2015a](https://doi.org/10.1214/09-AOAS311) 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 $\theta = \hat{\Gamma}^\top \eta$ for some $\eta \in \mathbb{R}^u$. In low-dimensional settings, [Cook & Zhang 2015a](https://doi.org/10.1214/09-AOAS311) 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 Section 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 $\hat{\Gamma}^\top X \in \mathbb{R}^u$ to obtain the parameters $\hat{\eta}$ and $\hat{\theta}_{\text{Env}} = \hat{\Gamma} \hat{\eta}$. 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 (9), where we follow Witten et al. (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 & 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 & Husson 2012). In simulation studies (Section 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 (Section 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 (Section 2) and the high-dimensional sparse NIECE (SNIECE; Section 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 \(\Gamma = (\Gamma_s^T, 0^T)^T \in \mathbb{R}^{p \times u}\), where entries in \(\Gamma_s \in \mathbb{R}^{s \times u}\) are generated from \(\text{Uniform}[0, 1]\) and then orthogonalized such that \(\Gamma_s^T \Gamma = I_s\).

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.

- \(\Sigma_1 = VDV^T\), where \(V = (v_1, \ldots, v_s) \in \mathbb{R}^{s \times s}\) is a randomly generated orthogonal matrix and \(D = \text{diag}(D_{11}, \ldots, D_{ss}) \in \mathbb{R}^{s \times s}\) 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 \(\Gamma_s = (v_7, v_8, v_9)\).

- \(\Sigma_2 = \Gamma_s^T \Omega \Gamma_s^T + \Gamma_0 \Omega_0 \Gamma_0^T\), where \(\Gamma_0 \in \mathbb{R}^{s \times (s-u)}\) is the orthogonal completion of \(\Gamma_s\) and \(\Omega, \Omega_0\) are symmetric positive definite matrices. In multivariate linear models, we generate
\( \Omega \) as \( \text{ODO}^\top \), where \( \text{O} \) is a randomly generated orthogonal matrix and \( \text{D} \) is diagonal with entries \((k+1)^3, k = 1, \ldots, u; \); \( \Omega_0 \) is a diagonal matrix with entries 50, 1, 1, \ldots, 1. In logistic regression and Cox model, we change the entries in \( \text{D} \) to \((k+1)^2, k = 1, \ldots, u, \) and the diagonals of \( \Omega_0 \) to 50, 0.01, \ldots, 0.01.

- \( \Sigma_3 \) is generated similar to \( \Sigma_2 \). We set the eigenvalues in \( \Omega \) (i.e. diagonals of \( \text{D} \)) as \((k+1)^2, k = 1, \ldots, u \) and let \( \Omega_0 = 0.01I_{s-u} \).

The covariance \( \Sigma_1 \) contains eigenvalues that are very well-separated, where the largest one is many magnitudes larger than the smallest ones. Both the covariances \( \Sigma_1 \) and \( \Sigma_2 \) 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 \( \Sigma_3 \) 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 \( \beta^* \in \mathbb{R}^{p \times q} \) as \( \beta^* = \Gamma \eta \) and then standardized to \( \beta = 10 \cdot \beta^*/\|\beta^*\|_F \), where the entries in \( \eta \in \mathbb{R}^{u \times q} \) are sampled from \textit{Uniform}[0, 1]. The response is generated as \( Y_i = \beta^T X_i + \epsilon_i, i = 1, \ldots, n \), where the error \( \epsilon_i \sim N(0, \text{diag}(\Sigma_A, I_{p-s})) \) for \( A = 1, 2, 3 \) three different covariance structures. The predictors are generated from \( N(0, 30I_q) \) when \( \Sigma_A = \Sigma_1 \) due to high variance in \( \Sigma_1 \), and are generated from \( N(0, I_q) \) for \( \Sigma_A = \Sigma_2 \) and \( \Sigma_3 \).

**M2: Predictor envelope model in multivariate linear regression.** For response dimension \( q = 5 \), we generate \( \beta \in \mathbb{R}^{p \times q} \) in the same way as in Model M1. Then \( Y_i = \beta^T X_i + \epsilon_i, i = 1, \ldots, n \), where the predictor \( X_i \sim N(0, \text{diag}(\Sigma_A, 0.01I_{p-s})) \) and the error \( \epsilon_i \sim N(0, \sigma^2 I_q) \). The magnitude of error is controlled by \( \sigma^2 \) 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 \( \beta = \Gamma \eta \in \mathbb{R}^{p} \) is generated in the same way as in Model M1 where we set \( q = 1 \). The predictor \( X_i \) is also generated in the same way except that we change \( \Sigma_A \) to its scaled version \( \Sigma_A/\|\Sigma_A\| \), 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 \( \exp(\beta^T X_i)/(1 + \exp(\beta^T X_i)) \).

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

We focus on the parameter estimation error \( \|\beta - \hat{\beta}\|_F \) and the subspace estimation error \( \|P_{\Gamma} - P_{\hat{\Gamma}}\|_F/\sqrt{2u} \in [0, 1] \). 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 et al. 2009] that was used in SNIECE. We also include partial least squares regression (PLS, Wold [1966]), sparse partial least square regression (SPLS,
Chun & Keles (2010), sparse reduced-rank regression (SRRR, Chen & Huang 2012), reduced rank stochastic regression (RSSVD, Chen et al. 2012). In multivariate linear models, we also include ordinary least squares (OLS) and Lasso (Tibshirani 1996) estimators as benchmarks for estimating regression parameter $\beta$; in logistic regression model and Cox proportional hazards model, we include $\ell_1$-penalized MLE (PMLE, Friedman et al. 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.

Table 1 summarizes the results for all the four models across the three covariance structures. We have confirmed the following theoretical findings for linear, logistic, and Cox models through numerical experiments. (i) Under covariance structure $\Sigma_3$, 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 estimator; and the SNIECE coincided with the SPCR. (ii) Under covariance structures $\Sigma_1$ and $\Sigma_2$, the first principal component is orthogonal to the envelope, our NIECE procedure correctly identified the first principal component as irrelevant information in regression and eliminated it in subsequent estimation. As a result, the NIECE and SNIECE estimators substantially improved their 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 $\Sigma_3$) and SNIECE converged much faster than their un-penalized counterparts (PCR and NIECE).

Although we did not establish the parameter estimation consistency of $\hat{\beta}_{\text{Env}}$ 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 $\ell_1$-penalized MLE in logistic and Cox models. The significant improvements 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 $\beta$ and failed to select the important predictors when there is a large variation in the error term.

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 $\ell_1$-penalized (generalized) linear model (i.e. LASSO estimator). Following the practical suggestions in Section 4, we use $d = 2u$ principal components for each envelope dimension $u \in \{1, \ldots, 20\}$ in SNIECE.
| $\Sigma_A$ | $\Sigma_1$ | $\Sigma_2$ | $\Sigma_3$ | $\Sigma_1$ | $\Sigma_2$ | $\Sigma_3$ |
|---|---|---|---|---|---|---|
| $p$ | 400 | 1600 | 400 | 1600 | 400 | 1600 | 400 | 1600 | 400 | 1600 | 400 | 1600 |
| OLS | $\Delta_\beta$ | 2.89 | 3.24 | 5.38 | 9.63 | 4.72 | 9.26 | – | – | – | – | – | – |
| Lasso | $\Delta_\beta$ | – | – | – | – | – | – | – | – | – | – | – | – |
| PLS | $\Delta_\beta$ | 2.80 | 2.90 | 3.69 | 5.55 | 3.03 | 5.19 | 0.28 | 0.32 | 0.50 | 0.67 | 0.51 | 0.69 |
| SRRR | $\Delta_\beta$ | 2.42 | 2.54 | 3.61 | 5.63 | 2.74 | 5.11 | 0.26 | 0.30 | 0.47 | 0.65 | 0.47 | 0.69 |
| RSSVD | $\Delta_\beta$ | 2.41 | 2.41 | 2.65 | 2.65 | 1.36 | 1.39 | 0.24 | 0.24 | 0.26 | 0.27 | 0.11 | 0.12 |
| PCR | $\Delta_\beta$ | 3.20 | 3.23 | 3.80 | 4.57 | 2.22 | 3.91 | 0.58 | 0.58 | 0.59 | 0.62 | 0.31 | 0.53 |
| NIECE | $\Delta_\beta$ | 2.36 | 2.40 | 2.99 | 4.08 | 2.22 | 3.91 | 0.20 | 0.21 | 0.24 | 0.41 | 0.31 | 0.53 |
| SPLS | $\Delta_\beta$ | – | – | – | – | – | – | – | – | – | – | – | – |
| SPCR | $\Delta_\beta$ | 3.21 | 3.21 | 3.50 | 3.51 | 1.30 | 1.34 | 0.58 | 0.59 | 0.60 | 0.65 | 0.06 | 0.07 |
| SNIECE | $\Delta_\beta$ | 2.35 | 2.37 | 2.62 | 2.63 | 1.30 | 1.34 | 0.19 | 0.19 | 0.15 | 0.15 | 0.06 | 0.07 |
| S.E.$\leq$ | $\Delta_\beta$ | (0.01) | (0.02) | (0.01) | (0.01) | (0.02) | (0.01) | (0.01) | (0.01) | (0.01) | (0.02) | (0.02) | (0.01) |
| | $\Delta_\Gamma$ | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| M3 (Logistic regression) | | | | | | | | | | | | |
| $\Sigma_A$ | $\Sigma_1$ | $\Sigma_2$ | $\Sigma_3$ | $\Sigma_1$ | $\Sigma_2$ | $\Sigma_3$ |
| $p$ | 400 | 1600 | 400 | 1600 | 400 | 1600 | 400 | 1600 | 400 | 1600 | 400 | 1600 |
| PMLE | $\Delta_\beta$ | 7.13 | 7.35 | 8.13 | 8.20 | 9.25 | 9.17 | 5.46 | 5.72 | 7.96 | 7.96 | 10.48 | 10.69 |
| PCR | $\Delta_\beta$ | 5.80 | 7.66 | 9.35 | 9.48 | 3.17 | 5.28 | 0.65 | 0.76 | 0.63 | 0.73 | 0.23 | 0.42 |
| NIECE | $\Delta_\beta$ | 5.27 | 7.63 | 5.07 | 7.73 | 3.19 | 5.28 | 0.54 | 0.75 | 0.39 | 0.64 | 0.23 | 0.42 |
| SPCR | $\Delta_\beta$ | 3.92 | 3.91 | 9.68 | 9.53 | 1.81 | 1.63 | 0.59 | 0.59 | 0.58 | 0.58 | 0.08 | 0.08 |
| SNIECE | $\Delta_\beta$ | 2.23 | 2.44 | 2.08 | 2.07 | 1.72 | 1.56 | 0.17 | 0.19 | 0.13 | 0.14 | 0.07 | 0.07 |
| S.E.$\leq$ | $\Delta_\beta$ | (0.02) | (0.02) | (0.01) | (0.01) | (0.02) | (0.02) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| | $\Delta_\Gamma$ | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |

Table 1: The median estimation errors for the parameter $\Delta_\beta = \|\beta - \hat{\beta}\|_F$ and for the envelope subspace $\Delta_\Gamma = \|P_\Gamma - \hat{P}_\Gamma\|_F/\sqrt{2u}$. 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 $\Sigma_1$, SRRR had very big estimation errors ($\Delta_\beta > 100$) and was excluded in comparison; similarly, RSSVD also failed due to extremely high variability in data.
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 et al. (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 classify 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 et al. (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 et al. (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 \( n_{\text{train}} \) and a testing set of size \( n_{\text{test}} \), the averaged mis-classification error \( \sum_{i=1}^{n_{\text{test}}} I(Y_i \neq \hat{Y}_i)/n_{\text{test}} \times 100\% \) or the averaged prediction mean squared error \( \sum_{i=1}^{n_{\text{test}}} \| Y_i - \hat{Y}_i \|^2_2/n_{\text{test}} \) 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 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 3 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 $\ell_1$-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 Figure 3 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 Figure 3 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.
**S1 Additional numerical 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 S2 to explain the same amount of variation in data, SNIECE achieves the lowest classification error.

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

Table S2: 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.

**S2 Proofs for Lemmas [1-3]**

**S2.1 Proof for Lemma [1]**

For vectors, a natural measure of the distance between two vectors is

$$
\sin \Theta(\tilde{\gamma}_j, \gamma_j) = \sin \{\angle(\tilde{\gamma}_j, \gamma_j)\} = \sqrt{1 - (\tilde{\gamma}_j^T \gamma_j)^2}, \quad \text{(S1)}
$$

which in fact is equivalent to the Frobenius norm of the projection matrices $\|P_{\tilde{\gamma}_j} - P_{\gamma_j}\|_F$. To see this,

$$
\|P_{\tilde{\gamma}_j} - P_{\gamma_j}\|_F^2 = \text{trace}\{(P_{\tilde{\gamma}_j} - P_{\gamma_j})^T(P_{\tilde{\gamma}_j} - P_{\gamma_j})\}
= 2 - 2\text{trace}\{P_{\gamma_j}P_{\tilde{\gamma}_j}\}
= 2 - 2(\tilde{\gamma}_j^T \gamma_j)^2
= 2\{1 - (\tilde{\gamma}_j^T \gamma_j)^2\}.
$$

Therefore, for vectors,

$$
\sqrt{2} \sin \Theta(\tilde{\gamma}_j, \gamma_j) = \|P_{\tilde{\gamma}_j} - P_{\gamma_j}\|_F. \quad \text{(S2)}
$$
For matrices, a natural measure of the distance between two subspaces is

\[ \|P_{\hat{e}} - P_e\|_F = \|\hat{\Gamma}^T \hat{\Gamma} - \Gamma^T \Gamma\|_F \]

\[ = \sqrt{\text{trace}\{\hat{\Gamma}^T \hat{\Gamma} + \Gamma^T \Gamma - 2\hat{\Gamma}^T \Gamma \Gamma^T\}} \]

\[ = \sqrt{2u - 2\text{trace}\{\hat{\Gamma}^T \Gamma \cdot \Gamma^T \hat{\Gamma}\}} \]

\[ = \sqrt{2 \sum_{j=1}^{u} \{1 - \lambda_j^2(\hat{\Gamma}^T \Gamma)\}}, \]

which equals to \( \sqrt{2} \sqrt{\sum_{j=1}^{u} \sin^2(\theta_j)} \) with \( \theta_j \) being the \( j \)-th principal angle between the two subspaces and the \( j \)-th diagonal element of \( \Theta(\hat{\Gamma}, \Gamma) \). Therefore, we have \( \|P_{\hat{e}} - P_e\|_F = \sqrt{2} \| \sin(\Theta(\hat{\Gamma}, \Gamma)) \|_F \).

### S2.2 Proof for Lemma 2

The Lemma is straightforward to prove by applying Proposition 2.2 in [Cook et al. (2010)](https://www.jstor.org/stable/10.1017/CBO9781139037645). We thus omit the proof.

### S2.3 Proof for Lemma 3

Since there is no penalty on \( u \), we can re-write the optimization in (9) equivalently as,

\[ \hat{u}_k = \frac{X^k \hat{v}_k}{\|X^k \hat{v}_k\|_2}, \]

\[ \hat{v}_k = \arg \max_v \hat{u}_k^T X^k v \quad \text{subject to } \|v\|_1 \leq c \text{ and } \|v\|_2^2 \leq 1, \]

from which we can plug in \( \hat{u}_k \) into the optimization of \( v \) and get

\[ \hat{v}_k = \arg \max_v \frac{v^T (X^k)^T X^k v}{\|X^k v\|_2} \quad \text{subject to } \|v\|_1 \leq c \text{ and } \|v\|_2^2 \leq 1, \]

where the last equality is because maximizing \( v^T (X^k)^T X^k v / \|X^k v\|_2 \) is equivalent to maximizing \( \{v^T (X^k)^T X^k v / \|X^k v\|_2\}^2 = v^T (X^k)^T X^k v \). Therefore, the optimization in (9) is equivalent to

\[ \hat{v}_k = \arg \max_v v^T \hat{M}_k v \quad \text{subject to } \|v\|_1 \leq c \text{ and } \|v\|_2 \leq 1, \quad \text{(S3)} \]

where \( \hat{M}_k = (X^k)^T X^k \).

Next, we have \( \hat{\sigma}_k = \hat{u}_k^T X^k \hat{v}_k = \hat{v}_k^T (X^k)^T \hat{u}_k = \hat{v}_k^T (X^k)^T X^k \hat{v}_k / \|X^k \hat{v}_k\|_2 \). Thus \( X^{k+1} \) can be
written as,

\[
X^{k+1} = X^k - \hat{\sigma}_k \hat{u}_k \hat{v}_k^T \\
= X^k - \frac{\hat{v}_k^T(X^k)^T X^k \hat{v}_k \cdot \hat{u}_k \hat{v}_k^T}{\|X^k \hat{v}_k\|_2^2} \\
= X^k - \frac{\hat{v}_k^T(X^k)^T X^k \hat{v}_k \cdot X^k \hat{v}_k \hat{v}_k^T}{\|X^k \hat{v}_k\|_2^2} \\
= X^k (I_p - \hat{v}_k \hat{v}_k^T).
\]

This implies that we can also write \(\hat{M}_{k+1}\) as

\[
\hat{M}_{k+1} = (X^{k+1})^T X^{k+1} = (I_p - \hat{v}_k \hat{v}_k^T) \hat{M}_k (I_p - \hat{v}_k \hat{v}_k^T).
\]

### S3 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

\[
\Delta = \min_{j=1,\ldots,u} \Delta_j, \quad \Delta_j \equiv \min\{\lambda_{\pi(j)-1}(M) - \lambda_{\pi(j)}(M), \lambda_{\pi(j)}(M) - \lambda_{\pi(j)+1}(M)\}.
\]

(S4)

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 \(\Delta_j\) as

\[
\Delta_j \equiv \min\{\lambda_{\pi_1-1}(M) - \lambda_{\pi_1}(M), \lambda_{\pi_m}(M) - \lambda_{\pi_{m+1}}(M)\}.
\]

(S5)

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

**Lemma S4.** Assume \(\Delta > 0\), and let \(\epsilon > 0\) be a constant such that \(\|\hat{M} - M\|_{\text{op}} \leq \epsilon\), assume that \(\{\pi(j) | j = 1, \ldots, u\} = \{\pi(j) | j = 1, \ldots, u\}\), then

\[
\|\sin \Theta(\Gamma, \hat{\Gamma})\|_F \leq \frac{2\sqrt{\epsilon} \Delta}{\Delta}.
\]

(S7)

**Proof.** From Corollary 1 of [Yu et al. (2014)]{}, we have that for distinct eigenvalues,

\[
\sin \Theta(v_j, \hat{v}_j) \leq \frac{2\|\hat{M} - M\|_{\text{op}}}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})} \leq \frac{2\|\hat{M} - M\|_{\text{op}}}{\Delta_j} \leq \frac{2\|\hat{M} - M\|_{\text{op}}}{\Delta}.
\]

(S8)

For common eigenvalues, \(\lambda_{\pi_1} = \cdots = \lambda_{\pi_m}\) with \(\pi_1 > \cdots > \pi_m\), we let \(W_j = (v_{\pi_1}, \ldots, v_{\pi_m})\) and \(\hat{W}_j = (\hat{\nu}_{\pi_1}, \ldots, \hat{\nu}_{\pi_m})\). Then we have,

\[
\|\sin \Theta(W_j, \hat{W}_j)\|_F \leq \frac{2\sqrt{m}\|\hat{M} - M\|_{\text{op}}}{\Delta_j} \leq \frac{2\sqrt{m}\|\hat{M} - M\|_{\text{op}}}{\Delta}.
\]

(S9)
Since \( \| \sin \Theta(\mathbf{M}, \hat{\mathbf{M}}) \|_F^2 = \sum_{j=1}^u \sin^2(\theta_j) \), where \( \theta_j \)'s are the principal angles, we have

\[
\| \sin \Theta(\mathbf{M}, \hat{\mathbf{M}}) \|_F = \sqrt{\sum_{j \in J_1} \sin^2(\theta_j)} + \sum_{j \in J_2} \| \sin \Theta(\mathbf{W}_j, \hat{\mathbf{W}}_j) \|_F^2 \leq \frac{2\sqrt{u\|\hat{\mathbf{M}} - \mathbf{M}\|_\text{op}}}{\Delta},
\]

where the index sets \( J_1 \) is for distinct eigenvalues and index set \( J_2 \) is for common eigenvalues. The conclusion follows from \( \|\hat{\mathbf{M}} - \mathbf{M}\|_\text{op} \leq \epsilon \).

### S4 Proofs for Theorem 1

Throughout this subsection, we assume that \( \|\hat{\mathbf{M}} - \mathbf{M}\|_\text{op} \leq \epsilon, \|\hat{\mathbf{U}} - \mathbf{U}\|_\text{op} \leq \epsilon \). First, we introduce two technical Lemmas that are intermediate results for proving Theorem 1. Recall that we have defined the envelope scores \( \hat{\phi}_j = \hat{\mathbf{v}}_j^T \hat{\mathbf{U}} \hat{\mathbf{v}}_j \) for \( j = 1, \ldots, d \), and organize them in descending order \( \hat{\phi}(1) \geq \cdots \geq \hat{\phi}(d) \) and define \( \hat{\mathbf{v}}(j) \) correspondingly from \( \hat{\phi}(j) = \hat{\mathbf{v}}(j)^T \hat{\mathbf{U}} \hat{\mathbf{v}}(j) \).

**Lemma S5.** If \( \Delta > 0 \), then \( \sin \Theta(\hat{\mathbf{v}}_i, \mathbf{v}_i) \leq \frac{2\epsilon}{\Delta}, i = 1, \ldots, d \).

**Proof.** This result is a direct consequence of Corollary 1 in Yu et al. (2014).

**Lemma S6.** For \( i = 1, \ldots, d \), we have the following conclusions:

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

\[
|\hat{\phi}_i - \phi_i| \leq \left( 1 + \frac{10\nu}{\Delta} \right) \epsilon, \quad (S11)
\]

where \( \nu = \|\mathbf{U}\|_\text{op} \) is the largest eigenvalue of \( \mathbf{U} \geq 0 \).

2. If \( \|\hat{\mathbf{U}} - \mathbf{U}\|_\text{max} \leq \epsilon \), \( \sin^2 \Theta(\hat{\mathbf{v}}_i, \mathbf{v}_i) \leq c_0 \tau^2 \epsilon \) and \( \|\hat{\mathbf{v}}_i\|_1 \leq \tau \), then

\[
|\hat{\phi}_i - \phi_i| \leq \left\{ (c_0 + 1) \tau^2 + \frac{2\nu}{\Delta} \right\} \epsilon. \quad (S12)
\]

**Proof.** First of all, note that, because \( \sin \Theta(\hat{\mathbf{v}}_i, \mathbf{v}_i) \leq \frac{2\epsilon}{\Delta} \), we have

\[
\|\hat{\mathbf{v}}_i - \mathbf{v}_i\|_2 = \sqrt{2 - 2\cos \Theta(\hat{\mathbf{v}}_i, \mathbf{v}_i)} \leq \sqrt{2 - 2\sqrt{1 - \frac{4\epsilon^2}{\Delta^2}}} \leq \frac{2\sqrt{2\epsilon}}{\Delta}.
\]

Further notice that

\[
|\hat{\phi}_i - \phi_i| = |\hat{\mathbf{v}}_i^T \hat{\mathbf{U}} \hat{\mathbf{v}}_i - \mathbf{v}_i^T \mathbf{U} \mathbf{v}_i| \leq |\hat{\mathbf{v}}_i^T (\hat{\mathbf{U}} - \mathbf{U}) \hat{\mathbf{v}}_i| + |(\hat{\mathbf{v}}_i - \mathbf{v}_i)^T \mathbf{U} (\hat{\mathbf{v}}_i - \mathbf{v}_i)| + 2|\mathbf{v}_i^T \mathbf{U} (\hat{\mathbf{v}}_i - \mathbf{v}_i)|. \quad (S13)
\]
For the last two terms, note that
\[
| (\hat{v}_i - v_i)^T U (\hat{v}_i - v_i) | \leq \| U \|_{op} \| \hat{v}_i - v_i \|_2^2 \leq 8 \epsilon^2 \nu / \Delta^2 \leq 2 \epsilon \nu / \Delta, \tag{S14}
\]
\[
2| v_i^T U (\hat{v}_i - v_i) | \leq 2 \sqrt{v_i^T U^2 v_i} \| \hat{v}_i - v_i \|_2 \leq 4 \sqrt{2} \epsilon \nu / \Delta \leq 8 \epsilon \nu / \Delta, \tag{S15}
\]
where we have used \( \epsilon / \Delta \leq 1 / 4 \).

Similarly, if we assume that \( \| \hat{U} - U \|_{op} \leq \epsilon \) and \( \sin^2 \Theta(\hat{v}_i, v_i) \leq c_0 \tau^2 \epsilon \). Then
\[
\| \hat{v}_i - v_i \|_{\infty} \leq \| \hat{v}_i - v_i \|_2 \leq 2 - 2 \cos \Theta(\hat{v}_i, v_i) \leq 2 - 2 \sqrt{1 - c_0 \tau^2 \epsilon} = 2 \frac{c_0 \tau^2 \epsilon}{1 + \sqrt{1 - c_0 \tau^2 \epsilon}} \leq 2 c_0 \tau^2 \epsilon.
\]

Then the first term in (S13) is
\[
| \hat{v}_i^T (\hat{U} - U) \hat{v}_i | \leq \| \hat{U} - U \|_{max} \| \hat{v}_i \|_1^2 \leq \tau^2 \epsilon.
\]
Then the last two terms in (S13) become
\[
| (\hat{v}_i - v_i)^T U (\hat{v}_i - v_i) | \leq \| U \|_{op} \| \hat{v}_i - v_i \|_2^2 \leq 2 c_0 \nu \tau^2 \epsilon, \tag{S16}
\]
\[
2 | v_i^T U (\hat{v}_i - v_i) | \leq 2 \sqrt{v_i^T U^2 v_i} \| \hat{v}_i - v_i \|_2 \leq 2 \nu \sqrt{2 c_0 \tau^2 \epsilon}. \tag{S17}
\]
And we have the desired conclusions.

Next, to prove Theorem 1, we apply these two Lemmas. By Lemma S6, if \( \Delta_U > 2(1 + 10 \nu / \Delta) \epsilon \), we keep all the \( \hat{v}_i \)'s in the envelope. Combine this fact with Lemma S5 and we have
\[
| \phi_{\pi(j)} - \phi_{\pi(j)} | < \Delta_U / 2 \quad \text{and thus for } j = 1, \ldots, u, \quad \hat{\phi}_{\pi(j)} > \phi_{\pi(j)} - \Delta_U / 2 > \Delta_U / 2, \quad \text{and}
\]
\[
0 \leq \hat{\phi}_{\pi(j)} < \Delta_U / 2 \quad \text{for } j = u + 1, \ldots, d. \quad \text{Therefore we have selected the correct components}
\]
\[
\hat{\pi}(j) \in \{ \pi(k) | k = 1, \ldots, u \} \text{ for all } j = 1, \ldots, u. \quad \text{Therefore,}
\]
\[
\| P_{\hat{\pi}} - P_{\pi} \|_F = \sqrt{2} \sum_{i : v_i \in \pi} \sin^2 \Theta(\hat{v}_i, v_i) \leq \frac{2 \sqrt{2} \epsilon \nu}{\Delta}.
\]

S5  Proofs for Theorem 2 and Corollary 1

Note that the response and predictor envelopes are constructed in symmetry. The NIECE estimator is obtained by using the sample covariances \( \hat{M} = \hat{\Sigma}_Y \) and \( \hat{U} = \hat{\Sigma}_{XY} \hat{\Sigma}_{XY} \) for response reduction, and \( \hat{M} = \hat{\Sigma}_X \) and \( \hat{U} = \hat{\Sigma}_{XY} \hat{\Sigma}_{XY} \) for predictor reduction. We therefore only lay out the proof for response envelope. For predictor envelope, we will need to interchange \( p \) and \( r \).

Recall from Theorem 1, in order to show \( \| P_{\hat{\pi}} - P_{\pi} \|_F \leq \frac{2 \sqrt{2} \epsilon \nu}{\Delta} \), we need to verify
\[
\epsilon \leq \max \left\{ \Delta / 4, \Delta_U / 2 \left( 1 + \frac{10 \nu}{\Delta} \right)^{-1} \right\}, \quad \| \hat{M} - M \|_{op} \leq \epsilon, \quad \| \hat{U} - U \|_{op} \leq \epsilon. \tag{S18}
\]
By the following Lemma S7 and Theorem 1 and we have the desired conclusion. Corollary 1 follows from Theorem 2 by letting \( \epsilon = pr \sqrt{\log p + \log r \over n} \to 0 \).
Lemma S7. 1. For $\epsilon > 0$, we have

$$\|\hat{\Sigma}_Y - \Sigma_Y\|_{op} \leq \epsilon$$  \hspace{1cm} (S19)

with a probability greater than $1 - C r^2 \exp(-C n \epsilon^2 / r^2)$.

2. For $0 < \epsilon < \|\Sigma_{XY}\|_1 + \|\Sigma_{YX}\|_1$, we have

$$\|\hat{\Sigma}_{YX} \hat{\Sigma}_{XY} - \Sigma_{YX} \Sigma_{XY}\|_{op} \leq 2(\|\Sigma_{XY}\|_1 + \|\Sigma_{YX}\|_1) \epsilon$$  \hspace{1cm} (S20)

with a probability greater than $1 - C pr \exp(-C n \epsilon^2 / r^2) - C pr \exp(-C n \epsilon^2 / p^2)$. In addition, if we assume $\|\Sigma_{XY}\|_1 + \|\Sigma_{YX}\|_1 < c_1$ for some constant $c_1$, we can claim

$$\|\hat{\Sigma}_{YX} \hat{\Sigma}_{XY} - \Sigma_{YX} \Sigma_{XY}\|_{op} \leq \epsilon$$  \hspace{1cm} (S21)

with a probability greater than $1 - C pr \exp(-C n \epsilon^2 / r^2) - C pr \exp(-C n \epsilon^2 / p^2)$.

Proof. We first show (S19). Note that

$$\Pr(\|\hat{\Sigma}_Y - \Sigma_Y\|_{op} > \epsilon) \leq \Pr(\|\hat{\Sigma}_Y - \Sigma_Y\|_1 > \epsilon)$$

$$\leq \Pr(\cup_{i,j} |\hat{\sigma}_{Y,ij} - \sigma_{Y,ij}| > \epsilon / r) \leq \sum_{i,j} \Pr(|\hat{\sigma}_{Y,ij} - \sigma_{Y,ij}| > \epsilon / r)$$

$$\leq r^2 \exp(-C n \epsilon^2 / r^2)$$

(S24)

where the last inequality follows from Lemma S13 and joint sub-Gaussian of $(X, Y)$ in (7).

Now we show (S20). Straightforward calculation shows that

$$\|\hat{\Sigma}_{YX} \hat{\Sigma}_{XY} - \hat{\Sigma}_{YX} \hat{\Sigma}_{XY}\|_{op}$$

$$\leq \|\hat{\Sigma}_{YX} \hat{\Sigma}_{XY} - \hat{\Sigma}_{YX} \hat{\Sigma}_{XY}\|_1$$

$$\leq \|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_1 + \|\hat{\Sigma}_{YX} - \Sigma_{YX}\|_1 + \|\Sigma_{XY} (\hat{\Sigma}_{XY} - \Sigma_{XY})\|_1$$

$$\leq \max\{\|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_1, \|\hat{\Sigma}_{YX} - \Sigma_{YX}\|_1\} \|\Sigma_{XY} (\hat{\Sigma}_{XY} - \Sigma_{XY})\|_1$$

which is less than $2(\|\Sigma_{XY}\|_1 + \|\Sigma_{YX}\|_1) \epsilon$ if we assume that

$$\max\{\|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_1, \|\hat{\Sigma}_{YX} - \Sigma_{YX}\|_1\} \leq \epsilon \leq \|\Sigma_{XY}\|_1 + \|\Sigma_{YX}\|_1.$$  \hspace{1cm} (S25)

Hence, it suffices to calculate the probability bound for $\max\{\|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_1, \|\hat{\Sigma}_{YX} - \Sigma_{YX}\|_1\} \leq \epsilon$ as follows.

$$\Pr(\|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_1 > \epsilon) \leq \Pr(\cup_{i,j} |\hat{\sigma}_{XY,ij} - \sigma_{XY,ij}| > \epsilon / p) \leq Cpr \exp(-C n \epsilon^2 / p^2).$$

Similarly, $\Pr(\|\hat{\Sigma}_{YX} - \Sigma_{YX}\|_1 > \epsilon) \leq Cpr \exp(-C n \epsilon^2 / r^2)$. And the conclusion follows. \hfill \Box

29
S6  Proofs for Theorems 3 and 4

We first introduce several technical lemmas, followed by proofs for the two main theorems.

We use $\lambda_k$ for the $k$-th eigenvalue of $M$ and $\hat{\lambda}_k = \hat{v}_k^T \hat{M} \hat{v}_k$, which is not exactly the $k$-th eigenvalue of $\hat{M}$ but is a good estimate for $\lambda_k$. We also define the following quantities throughout this section,

$$
\tau_0 = \max_{1 \leq k \leq d} \|v_k\|_1
$$

$$
\eta_k = \|\hat{v}_k\|_2^2, \quad \eta_k \leq 1
$$

$$
\hat{v}_k = \eta_k^{-1/2} \hat{v}_k, \quad \|\hat{v}_k\|_2 = 1
$$

$$
\eta_0 = \min_{1 \leq k \leq d} \eta_k.
$$

Recall that, for some $\epsilon > 0$ and $\tau > 0$, we assumed in Theorems 3 and 4 that

$$
\|\hat{M} - M\|_{\text{max}} \leq \epsilon, \quad \tau_0 < \tau. \quad (S26)
$$

S6.1  Technical Lemmas

Lemma S8. For any $w \in \mathbb{R}^p$ such that $w^Tw = 1$ we have

$$
\sin^2 \Theta(w, v_k) \leq \frac{\lambda_k - w^T M_k w}{\Delta}, \quad k = 1, \ldots, d. \quad \text{(S27)}
$$

Moreover, if $\sin^2 \Theta(w, v_k) = t$, then

$$
w^T M_k w \leq \lambda_k - \Delta t,
$$

$$
\sin^2 \Theta(w, v_j) \geq 1 - 2t, \quad j \neq k,
$$

$$
(w^T v_j)^2 \leq 2t, \quad j \neq k,
$$

$$
w^T M w \geq \lambda_k (1 - t).
$$

Proof. Recall that we can write $M_k = \sum_{i=k}^p \lambda_i v_i v_i^T$, therefore,

$$
\lambda_k - w^T M_k w = \lambda_k - \sum_{i=k}^p \lambda_i (w^T v_i)^2
$$

$$
= \lambda_k - \sum_{i=k}^p \lambda_i \{1 - \sin^2 \Theta(w, v_i)\}
$$

$$
= \lambda_k - \lambda_k \{1 - \sin^2 \Theta(w, v_k)\} - \sum_{i=k+1}^p \lambda_i \{1 - \sin^2 \Theta(w, v_i)\}
$$

$$
= \lambda_k \sin^2 \Theta(w, v_k) - \sum_{i=k+1}^p \lambda_i \{1 - \sin^2 \Theta(w, v_i)\}.
$$

30
Because \( 1 = \sum_{i=1}^{p} \{1 - \sin^2 \Theta(w, v_i)\} \), \( \sum_{i=k+1}^{p} \{1 - \sin^2 \Theta(w, v_i)\} \leq \sin^2 \Theta(w, v_k) \). Therefore, we have,

\[
\lambda_k - w^T M_k w = \lambda_k \sin^2 \Theta(w, v_k) - \sum_{i=k+1}^{p} \lambda_i \{1 - \sin^2 \Theta(w, v_i)\} \\
\geq \lambda_k \sin^2 \Theta(w, v_k) - \lambda_{k+1} \sum_{i=k+1}^{p} \{1 - \sin^2 \Theta(w, v_i)\} \\
\geq \lambda_k \sin^2 \Theta(w, v_k) - \lambda_{k+1} \sin^2 \Theta(w, v_k) \\
\geq \Delta \sin^2 \Theta(w, v_k).
\]

Then we have \( \sin^2 \Theta(w, v_k) \leq (\lambda_k - w^T M_k w)/\Delta \) and also \( w^T M_k w \leq \lambda_k - \Delta t \), where \( t = \sin^2 \Theta(w, v_k) \).

For \( j \neq k \), we write \( \sin^2 \Theta(w, v_j) = 1 - (w^T v_j)^2 \), so it is equivalent to show that \( (w^T v_j)^2 \leq 2t \). Without loss of generality, suppose \( w^T v_k \geq 0 \) (if \( w^T v_k < 0 \) then we can simply replace \( w - v_k \) with \( w + v_k \) in the following derivation and reach the same conclusion). It follows from \( v_j^T v_j = 0 \) that

\[
(w^T v_j)^2 = \{(w - v_k)^T v_j\}^2 \\
\leq \{(w - v_k)^T (w - v_k)\} \{v_j^T v_j\} \\
= (2 - 2w^T v_k) \\
= 2 - 2\sqrt{1 - t}.
\]

It is then straightforward to verify \( 2 - 2\sqrt{1 - t} \leq 2t \) since \( t = \sin^2 \Theta(w, v_k) \in [0, 1] \), and the conclusion follows.

To show that \( w^T M w \geq \lambda_k (1 - t) \), we notice that

\[
w^T M w = w^T (\sum_{j=1}^{p} \lambda_j v_j v_j^T) w \\
\geq \lambda_k (w^T v_k)^2 \\
= \lambda_k \cos^2 \Theta(w, v_k) \\
= \lambda_k (1 - t).
\]

\[\square\]

**Lemma S9.** For \( k = 1, \ldots, d \),

\[
|v_k^T \hat{M} v_k - \lambda_k| \leq C \epsilon_0^2 \leq C \epsilon^2. \tag{S28}
\]

**Proof.** The conclusion follows immediately from the fact that \( |v_k^T \hat{M} v_k - \lambda_k| = |v_k^T (\hat{M} - M) v_k| \leq \|v_k\|^2 \|\hat{M} - M\|_{\text{max}} \leq \epsilon_0^2 \leq \epsilon^2 \).

\[\square\]
Lemma S10. If $\epsilon \tau^2 < \min \left\{ \frac{1}{4}, \frac{\lambda_1}{4} \right\}$, then the following inequalities hold,

\[
\begin{align*}
\sin^2 \Theta(\hat{v}_1, v_1) &\leq \frac{3\epsilon \tau^2}{\Delta}, \\
\hat{v}_1^T M \hat{v}_1 &\geq \lambda_1 - 2\epsilon \tau^2, \\
\|\hat{v}_1\|^2 &\geq 1 - 2\lambda^{-1}_1 \epsilon \tau^2, \\
|\lambda_1 - \hat{\lambda}_1| &\leq \epsilon \tau^2, \\
\|\hat{v}_1 - v_1\|^2 &\leq (2\lambda^{-1}_1 + 6)\epsilon \tau^2.
\end{align*}
\]

Proof. First, we prove (S31).

Since $\|\hat{v}_1\|^2 = \eta_1 \leq 1$, we have

\[
\hat{v}_1^T M \hat{v}_1 = \hat{v}_1^T M \hat{v}_1 + \hat{v}_1^T (\hat{M} - M) \hat{v}_1 \leq \lambda_1 \eta_1 + \epsilon \tau^2,
\]

where the inequality follows from $\hat{v}_1^T M \hat{v}_1 \leq \|M\|_{op}\|\hat{v}_1\|^2 \leq \lambda_1 \eta_1$ and $\hat{v}_1^T (\hat{M} - M) \hat{v}_1 \leq \|\hat{M} - M\| \|\hat{v}_1\|_2^2 \leq \epsilon \tau^2$.

Analogously, we also have

\[
v_1^T \hat{M} v_1 = v_1^T M v_1 + v_1^T (\hat{M} - M) v_1 \geq \lambda_1 - \epsilon \tau^2.
\]

Since $\hat{v}_1$ is the maximizer of the optimization problem, where the population truth $v_1$ is in the feasible set since $\tau \geq \tau_0$, we must have

\[
\lambda_1 \eta_1 + \epsilon \tau^2 \geq \lambda_1 - \epsilon \tau_0^2,
\]

therefore,

\[
\eta_1 \geq \frac{\lambda_1 - \epsilon (\tau_0^2 + \tau^2)}{\lambda_1} = 1 - \frac{\epsilon (\tau_0^2 + \tau^2)}{\lambda_1} \geq 1 - \frac{2\epsilon \tau^2}{\lambda_1}.
\]

Second, we prove (S29).

From Lemma S8 we have

\[
\sin^2 \Theta(\hat{v}_1, v_1) \leq \frac{\lambda_1 - \eta_1^-1 \hat{v}_1^T M \hat{v}_1}{\Delta}.
\]

We can write

\[
\begin{align*}
\lambda_1 - \eta_1^-1 \hat{v}_1^T M \hat{v}_1 &= v_1^T M v_1 - \eta_1^-1 \hat{v}_1^T M \hat{v}_1 \\
&= v_1^T M v_1 - \eta_1^-1 \hat{v}_1^T \hat{M} \hat{v}_1 + \eta_1^-1 \hat{v}_1^T (\hat{M} - M) \hat{v}_1 \\
&\leq v_1^T M v_1 - \eta_1^-1 \hat{v}_1^T \hat{M} \hat{v}_1 + \eta_1^-1 \|\hat{M} - M\|_{max} \|\hat{v}_1\|_2^2 \\
&\leq v_1^T (M - \hat{M}) v_1 + v_1^T \hat{M} v_1 - \eta_1^-1 \hat{v}_1^T \hat{M} \hat{v}_1 + \eta_1^-1 \epsilon \tau^2 \\
&\leq \|M - \hat{M}\|_{max} \|v_1\|_2^2 + v_1^T \hat{M} v_1 - \eta_1^-1 \hat{v}_1^T \hat{M} \hat{v}_1 + \eta_1^-1 \epsilon \tau^2 \\
&\leq \epsilon (\tau_0^2 + \eta_1^-1 \tau^2) + v_1^T \hat{M} v_1 - \eta_1^-1 \hat{v}_1^T \hat{M} \hat{v}_1 \\
&\leq \epsilon (\tau_0^2 + \eta_1^-1 \tau^2),
\end{align*}
\]

where the last inequality follows from

\[
v_1^T \hat{M} v_1 \leq \eta_1^-1 \hat{v}_1^T \hat{M} \hat{v}_1.
\]
because $\sqrt{\eta_1}v_1$ is in the feasible set of the optimization where $\hat{v}_1$ is the maximizer. Hence, by (S31), we have

$$\lambda_1 - \eta_1^{-1}\hat{v}_1^T\hat{M}\hat{v}_1 \leq \epsilon\{\tau_0^2 + \frac{\tau^2}{1 - 2\epsilon\tau^2\lambda_1^{-1}}\} \leq 3\epsilon^2; \quad (S40)$$

if $2\epsilon\tau^2\lambda_1^{-1} \leq 1/2$. Therefore,

$$\sin^2 \Theta(\hat{v}_1, v_1) \leq \frac{3\epsilon\tau^2}{\Delta}. \quad (S41)$$

Thirdly, we prove (S30).

Specifically, we have the following derivations to get (S30),

$$\hat{v}_1^T\hat{M}\hat{v} = \hat{v}_1^T\hat{M}\hat{v} + \hat{v}_1^T(M - \hat{M})\hat{v}_1 \geq \hat{v}_1^T\hat{M}\hat{v}_1 - \epsilon\tau^2 \geq v_1^T\hat{M}v_1 - \epsilon\tau^2 = \lambda_1 - v_1^T(M - \hat{M})v_1 - \epsilon\tau^2 \geq \lambda_1 - \epsilon\tau_0^2 - \epsilon\tau^2 \geq \lambda_1 - 2\epsilon\tau^2.$$

Fourth, we prove (S32).

We notice that, on one hand,

$$\hat{\lambda}_1 - \lambda_1 = \hat{v}_1^T\hat{M}\hat{v}_1 - v_1^T\hat{M}v_1 \geq v_1^T\hat{M}v_1 - v_1^T\hat{M}v_1 \geq -\epsilon\tau_0^2,$$

while on the other hand,

$$\hat{\lambda}_1 - \lambda_1 = \hat{v}_1^T\hat{M}\hat{v}_1 + \hat{v}_1^T(\hat{M} - M)\hat{v}_1 - v_1^T\hat{M}v_1 \leq \hat{v}_1^T(\hat{M} - M)\hat{v}_1 \leq \epsilon\tau^2.$$

Finally, we prove (S33).

$$\|\hat{v}_1 - v_1\|^2_2 = \hat{v}_1^T\hat{v}_1 + v_1^Tv_1 - 2\hat{v}_1^Tv_1 = 1 + \eta_1 - 2\sqrt{\eta_1 \cos^2 \Theta(\hat{v}_1, v_1)}.$$

Let $\beta \equiv \cos^2 \Theta(\hat{v}_1, v_1)$, then

$$\|\hat{v}_1 - v_1\|^2_2 = 1 - \eta_1 + 2\eta_1 - 2\sqrt{\eta_1 \beta} \leq \frac{2\epsilon\tau^2}{\lambda_1} + 2\sqrt{\eta_1 \frac{\eta_1 - \beta}{\sqrt{\eta_1 + \beta}}},$$

$$\leq \frac{2\epsilon\tau^2}{\lambda_1} + 2\cdot 1 \cdot \frac{3\epsilon\tau^2}{1},$$

33
where the last inequality is because of $\eta_1 \leq 1$ and $\eta_1 - \beta \leq 1 - \beta = \sin^2 \Theta(\hat{v}_1, v_1) \leq 3\epsilon \tau^2$ from (S29), moreover, $\sqrt{\eta_1} + \sqrt{\beta} \geq 1$ if $\max(3\epsilon \tau^2, \frac{3}{4}) \leq \frac{3}{4}$.

Lemma S11. Assume that $\epsilon \tau^2 \leq 1$. If for all $j = 1, \ldots, k - 1$, we have

$$\sin^2 \Theta(\hat{v}_j, v_j) \leq C\epsilon \tau^2, \quad \|\hat{v}_j - v_j\|^2 \leq C\epsilon \tau^2,$$

then for any $w \in \mathbb{R}^p$, $\|w\|^2_2 = 1$ and $\|w\|_1 \leq \tau$,

$$w^T \hat{M}_k w \leq \lambda_k - \Delta \alpha + C\sqrt{\epsilon \tau^2 \alpha} + C\epsilon \tau^2,$$

where $\alpha = \sin^2 \Theta(w, v_k)$.

Proof. From Lemma S8, we have

$$w^T M_k w \leq \lambda_k - \Delta \alpha,$$

thus we can write

$$w^T \hat{M}_k w = w^T M_k w + w^T (\hat{M}_k - M_k) w \leq \lambda_k - \Delta \alpha + w^T (\hat{M}_k - M_k) w.$$

It remains to show that

$$w^T (\hat{M}_k - M_k) w \leq C\sqrt{\epsilon \tau^2 \alpha} + C\epsilon \tau^2.$$  \hfill (S44)

First of all, we decompose the matrix $\hat{M}_k - M_k$ into several terms as follows. We write $M_k$ as

$$M_k = M - \sum_{j<k} \lambda_j v_j v_j^T = M - P_{k-1} M = M - P_{k-1} M - MP_{k-1} + P_{k-1} MP_{k-1},$$

where $P_{k-1}$ is the projection matrix onto the subspace spanned by the first $(k-1)$-eigenvectors of $M$, i.e. $P_{k-1} = \sum_{j<k} v_j v_j^T$.

From Lemma S3, we have

$$\hat{M}_k = (I_p - \hat{v}_{k-1} \hat{v}_{k-1}^T) M_{k-1} (I_p - \hat{v}_{k-1} \hat{v}_{k-1}^T) = \left\{ \prod_{j<k} (I_p - \hat{v}_j \hat{v}_j^T) \right\} \hat{M} \left\{ \prod_{j<k} (I_p - \hat{v}_j \hat{v}_j^T) \right\}.$$

Similarly to the definition of $P_{k-1} = \sum_{j<k} v_j v_j^T$, we define $\hat{P}_{k-1} = \sum_{j<k} \hat{v}_j \hat{v}_j^T$. It is worth mentioning that $\hat{P}_{k-1}$ is not a projection matrix because $\|\hat{v}_j\|^2_2$ may not be exactly 1 and also that $\hat{v}_j$ may not be orthogonal to $\hat{v}_{j-1}$. Therefore, we write

$$\prod_{j<k} (I_p - \hat{v}_j \hat{v}_j^T) = I_p - \sum_{j<k} \hat{v}_j \hat{v}_j^T + \sum_{s=2}^{k-1} \sum_{l_1 < \cdots < l_s < k} (-1)^s (\hat{v}_{l_1} \hat{v}_{l_1}^T) \cdots (\hat{v}_{l_s} \hat{v}_{l_s}^T) = I_p - \hat{P}_{k-1} + \hat{P}_{k-1},$$

and

$$\hat{P}_{k-1} = \sum_{j<k} \hat{v}_j \hat{v}_j^T.$$
where \( \hat{P}_{k-1} \) is not zero in general because \( \hat{v}_1, \ldots, \hat{v}_{k-1} \) are not orthogonal to each other. Also, \( \hat{P}_{k-1} = 0 \) for \( k = 1 \), and \( \hat{P}_{k-1} = 0 \) for \( k = 1, 2 \). Get back to \( \hat{M}_k \), we have

\[
\hat{M}_k = \left\{ I_p - \hat{P}_{k-1} + \hat{P}_{k-1} \right\} \hat{M} \left\{ I_p - \hat{P}_{k-1} + \hat{P}_{k-1} \right\}
\]

\[
= \hat{M} - \hat{P}_{k-1} \hat{M} - \hat{M} \hat{P}_{k-1} + \hat{P}_{k-1} \hat{M} \hat{P}_{k-1} + \hat{P}_{k-1} \hat{M} (I_p - \hat{P}_{k-1} + \hat{P}_{k-1}) + (I_p - \hat{P}_{k-1} + \hat{P}_{k-1}) \hat{M} \hat{P}_{k-1}.
\]  

(S46)

Comparing (S45) and (S46), we have

\[
\begin{align*}
\text{w}^T (\hat{M}_k - M_k)w &= \text{w}^T (\hat{M} - M)w + \text{w}^T (\hat{P}_{k-1} \hat{M} \hat{P}_{k-1} - P_{k-1} MP_{k-1})w \\
&+ \text{w}^T (P_{k-1} M - \hat{P}_{k-1} \hat{M})w + \text{w}^T (MP_{k-1} - \hat{M} \hat{P}_{k-1})w \\
&+ \text{w}^T \left\{ \hat{P}_{k-1} \hat{M} (I_p - \hat{P}_{k-1} + \hat{P}_{k-1}) \right\} w + \text{w}^T \left\{ (I_p - \hat{P}_{k-1} + \hat{P}_{k-1}) \hat{M} \hat{P}_{k-1} \right\} w \\
&\equiv Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6.
\end{align*}
\]

To show (S44), we need to show that

\[
Q_m \leq C \sqrt{\tau^2 \alpha} + C \epsilon \tau^2, \quad m = 1, \ldots, 6.
\]

First, we compute \( Q_1 \),

\[
Q_1 = \text{w}^T (\hat{M} - M)w \leq \| \hat{M} - M \|_{\text{max}} \| w \|_1^2 \leq \epsilon \tau^2.
\]

Second, \( Q_3 = Q_4 \) by symmetry, we only need to compute \( Q_3 \),

\[
Q_3 = \text{w}^T (P_{k-1} M - \hat{P}_{k-1} \hat{M})w \\
= \text{w}^T (P_{k-1} M - \hat{P}_{k-1} \hat{M})w + \text{w}^T (\hat{P}_{k-1} M - \hat{P}_{k-1} \hat{M})w \\
\equiv Q_{31} + Q_{32}.
\]

For \( Q_{32} \), we have

\[
Q_{32} = \text{w}^T \hat{P}_{k-1} (M - \hat{M})w = \sum_{j<k} (\text{w}^T \hat{v}_j) \{ \text{w}^T (M - \hat{M})w \}
\]

\[
\leq \sum_{j<k} (\text{w}^T \hat{v}_j) \cdot \epsilon \tau^2 \leq k \epsilon \tau^2 = C \epsilon \tau^2.
\]

For \( Q_{31} \), we have

\[
Q_{31} = \text{w}^T (P_{k-1} - \hat{P}_{k-1}) Mw = \sum_{j<k} \text{w}^T (v_j v_j^T - \hat{v}_j \hat{v}_j^T) Mw \\
= \sum_{j<k} \text{w}^T (v_j v_j^T - \hat{v}_j \hat{v}_j^T) Mw + \sum_{j<k} \text{w}^T (\hat{v}_j \hat{v}_j^T - \hat{v}_j \hat{v}_j^T) Mw \\
= \sum_{j<k} \text{w}^T (v_j - \hat{v}_j) \cdot (v_j^T Mw) + \sum_{j<k} (\text{w}^T \hat{v}_j) \cdot (v_j - \hat{v}_j)^T Mw,
\]

35
where the two terms are computed as follows.

\[
\sum_{j<k} w^T(v_j - \hat{v}_j) \cdot (v_j^T M w) = \sum_{j<k} w^T(v_j - \hat{v}_j) \cdot (\lambda_j v_j^T w)
\]

\[
\leq \sum_{j<k} (\|w\|_2 \|v_j - \hat{v}_j\|_2) \cdot \left\{ \lambda_j \sqrt{1 - (v_j^T w)^2} \right\}
\]

\[
\leq Ck \sqrt{\epsilon T^2} \alpha = C \sqrt{\epsilon T^2} \alpha.
\]

\[
\sum_{j<k} (w^T \hat{v}_j) \cdot (v_j - \hat{v}_j)^T M w = \sum_{j<k} \left\{ w^T v_j + w^T (\hat{v}_j - v_j) \right\} : (v_j - \hat{v}_j)^T M w
\]

\[
\leq \sum_{j<k} (\sqrt{\alpha} + C \sqrt{\epsilon T^2}) \cdot \sqrt{(v_j - \hat{v}_j)^T M(v_j - \hat{v}_j) \sqrt{w^T M w}}\quad (S47)
\]

\[
\leq \sum_{j<k} (\sqrt{\alpha} + C \sqrt{\epsilon T^2}) \cdot \lambda_1 \sqrt{\|v_j - \hat{v}_j\|^2_2 \cdot \|w\|^2_2}
\]

\[
\leq Ck (\sqrt{\alpha} + \sqrt{\epsilon T^2}) \cdot (\sqrt{\epsilon T^2}) = C \epsilon T^2 + C \sqrt{\epsilon T^2} \alpha,
\]

where the last inequality is because of \((S42)\).

Together, we have \(Q_3 = Q_4 \leq C \epsilon T^2 + C \sqrt{\epsilon T^2} \alpha\).  

Next, for \(Q_2\), we have

\[
Q_2 = w^T (\hat{P}_{k-1} \tilde{M} \hat{P}_{k-1} - P_{k-1} M P_{k-1}) w
\]

\[
= w^T \hat{P}_{k-1} (\tilde{M} - M) \hat{P}_{k-1} w + w^T \hat{P}_{k-1} M (\hat{P}_{k-1} - P_{k-1}) w
\]

\[
+ w^T (\hat{P}_{k-1} - P_{k-1}) M P_{k-1} w,
\]

\[
\equiv Q_{21} + Q_{22} + Q_{23}
\]

where the three terms are computed as follows.

\[
Q_{21} = w^T \hat{P}_{k-1} (\tilde{M} - M) \hat{P}_{k-1} w
\]

\[
\leq \epsilon \| \hat{P}_{k-1} w \|^2_1 = \epsilon \sum_{j<k} (w^T \hat{v}_j) \| \hat{v}_j \|^2_1
\]

\[
\leq k \epsilon T^2 = C \epsilon T^2.
\]

\[
Q_{22} = w^T \hat{P}_{k-1} M (\hat{P}_{k-1} - P_{k-1}) w
\]

\[
= w^T (\sum_{l<k} \hat{v}_l \hat{v}_l^T) M \sum_{j<k} \{ v_j (\hat{v}_j - v_j)^T w + (\hat{v}_j - v_j) \hat{v}_j^T w \}
\]

\[
= \sum_{l,j<k} (w^T \hat{v}_l) \{ \hat{v}_l^T M v_j (\hat{v}_j - v_j)^T w + \hat{v}_l^T M (\hat{v}_j - v_j) \hat{v}_j^T w \}
\]

\[
\leq \sum_{l,j<k} (\sqrt{\epsilon T^2} + \sqrt{\alpha}) \{ \hat{v}_l^T M v_j (\hat{v}_j - v_j)^T w + \hat{v}_l^T M (\hat{v}_j - v_j) \hat{v}_j^T w \},
\]

36
where the last inequality is because \( w^T\hat{v}_l \leq \sqrt{\alpha} + \sqrt{\epsilon r^2} \) as computed in (S47). Moreover, \((\hat{v}_j - v_j)^Tw \leq C\sqrt{\epsilon r^2}\) and \(\hat{v}_j^TM(\hat{v}_j - v_j) \leq C\sqrt{\epsilon r^2}\) follows from previous calculation in (S48). Thus,

\[
Q_{22} \leq C(\sqrt{\epsilon r^2} + \sqrt{\alpha})\sqrt{\epsilon r^2}(\hat{v}_j^TMv_j + \hat{v}_j^Tw),
\]

\[
\leq C(\sqrt{\epsilon r^2} + \sqrt{\alpha})\sqrt{\epsilon r^2}(\lambda_1 + 1)
\]

\[
= C\epsilon r^2 + C\sqrt{\epsilon r^2} \alpha.
\]

\[
Q_{23} = w^T(\hat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1})\mathbf{M}\mathbf{P}_{k-1}w
\]

\[
= w^T \sum_{j<k} \{(\hat{v}_j - v_j)v_j^T + \hat{v}_j(\hat{v}_j - v_j)^T\} \mathbf{M}\mathbf{P}_{k-1}w
\]

\[
= \sum_{j<k} \{w^T(\hat{v}_j - v_j) \cdot (v_j^T\mathbf{M}\mathbf{P}_{k-1}w) + (w^T\hat{v}_j) \cdot (\hat{v}_j - v_j)^T\mathbf{M}\mathbf{P}_{k-1}w\}
\]

\[
= \sum_{j<k} \left\{ w^T(\hat{v}_j - v_j) \cdot (\lambda_jv_j^Tw) + (w^T\hat{v}_j) \cdot \sum_{l<k} \lambda_l(\hat{v}_j - v_j)^Tv_l \cdot v_l^Tw \right\}
\]

\[
\leq \sum_{j<k} \left\{ \sqrt{\epsilon r^2} \cdot (C\sqrt{\alpha}) + (w^Tv_j + w^T\hat{v}_j - w^Tv_j) \cdot \sum_{l<k} C\sqrt{\epsilon r^2} \cdot \sqrt{\alpha} \right\}
\]

\[
\leq C\sqrt{\epsilon r^2} \alpha + C(\sqrt{\alpha} + \sqrt{\epsilon r^2})\sqrt{\epsilon r^2} \alpha
\]

\[
= C\sqrt{\epsilon r^2} \alpha,
\]

where the last equality is because \( \epsilon r^2 \leq 1 \) and \( \alpha \leq 1 \).

Finally, for \( Q_5 \) and \( Q_6 \), we have

\[
Q_5 = Q_6 = w^T(I_p - \hat{\mathbf{P}}_{k-1} + \hat{\mathbf{P}}_{k-1})\hat{\mathbf{M}}\hat{\mathbf{P}}_{k-1}w,
\]

which is zero for \( k \leq 2 \) since \( \hat{\mathbf{P}}_{k-1} = 0 \) for \( k \leq 2 \). For \( k \geq 2 \), we have

\[
\hat{\mathbf{P}}_{k-1}w = \sum_{s=2}^{k-1} \sum_{l_1 \cdots \leq l_s \leq k} (-1)^s(\hat{v}_{l_1}\hat{v}_{l_1}^T) \cdots (\hat{v}_{l_s}\hat{v}_{l_s}^T)w
\]

\[
= \sum_{s=2}^{k-1} \sum_{l_1 \cdots \leq l_s \leq k} (-1)^s(\hat{v}_{l_1}\hat{v}_{l_1}^T) \cdots (\hat{v}_{l_s}\hat{v}_{l_s}^T) \cdot (\hat{v}_{l_1}\hat{v}_{l_1}^T)(\hat{v}_{l_2}\hat{v}_{l_2}^T)w,
\]

where

\[
(\hat{v}_{l_1}\hat{v}_{l_1}^T)(\hat{v}_{l_2}\hat{v}_{l_2}^T)w = (\hat{v}_{l_1}\hat{v}_{l_1}^T)w\hat{v}_{l_1}.
\]

Notice that \( \hat{v}_{l_2}^Tw = w^Tv_{l_2} + w^T(\hat{v}_{l_2} - v_{l_2}) \leq \sqrt{\alpha} + \sqrt{\epsilon r^2} \) for \( l_2 < k \). Also, notice that

\[
\hat{v}_{l_1}^Tv_{l_2} = \hat{v}_{l_1}^Tv_{l_2} + \hat{v}_{l_1}^T(\hat{v}_{l_2} - v_{l_2}) \leq C\sqrt{\epsilon r^2},
\]

since \( l_1 < l_2 < k \). Therefore, for any \( u \in \mathbb{R}^p \),

\[
u^T\hat{\mathbf{P}}_{k-1}w \leq (C\epsilon r^2 + C\sqrt{\epsilon r^2} \alpha) \sum_{s=2}^{k-1} \sum_{l_1 \cdots \leq l_s \leq k} (-1)^s(\hat{v}_{l_1}\hat{v}_{l_1}^T) \cdots (\hat{v}_{l_s}\hat{v}_{l_s}^T)u^T\hat{v}_{l_1}.
\]
\[ |Q_6| = |w^T(I_p - \hat{P}_{k-1} + \hat{P}_{k-1})\hat{M}\hat{P}_{k-1}w| \]
\[
\leq (C\epsilon \tau^2 + C\sqrt{\epsilon \tau^2}) \left| \sum_{s=2}^{k-1} \sum_{l_1 < \cdots < l_s < k} (-1)^s (\hat{v}_{l_s} \hat{v}_{l_s}^T) \cdot \{w^T(I_p - \hat{P}_{k-1} + \hat{P}_{k-1})\hat{M}\hat{v}_{l_1}\} \right| \\
\leq (C\epsilon \tau^2 + C\sqrt{\epsilon \tau^2}) \{w^T(I_p - \hat{P}_{k-1} + \hat{P}_{k-1})\hat{M}\hat{v}_{l_1}\} \\
= (C\epsilon \tau^2 + C\sqrt{\epsilon \tau^2}),
\]
where the last equality is because \(w^T\hat{M}\hat{v}_j\), \(-w^T\hat{P}_{k-1}\hat{M}\hat{v}_j\) and \(w^T{\hat{P}}_{k-1}\hat{M}\hat{v}_j\) are all less than or equals to \(C(\lambda_1 + C\epsilon \tau^2)\).

**Lemma S12.** Assume that \(\epsilon \tau^2 \leq 1\). If for all \(j = 1, \ldots, k - 1\), we have

\[ \sin^2 \Theta(\hat{v}_j, v_j) \leq C\epsilon \tau^2, \quad \|\hat{v}_j - v_j\|^2 \leq C\epsilon \tau^2, \quad \tag{S49} \]

then we must have

\[ \hat{v}_k^T M_k \hat{v}_k \geq \lambda_k \cos^2 \Theta(\hat{v}_k, v_k), \tag{S50} \]
\[ v_k^T \hat{M}_k v_k \geq \lambda_k - C\epsilon \tau^2. \tag{S51} \]

**Proof.** First, to obtain (S50), notice that

\[ \hat{v}_k^T M_k \hat{v}_k \geq \hat{v}_k^T (\lambda_k v_k v_k^T) \hat{v}_k = \lambda_k \frac{\cos^2 \Theta(\hat{v}_k, v_k)}{\|\hat{v}_k\|^2_2}, \tag{S52} \]

where \(\|\hat{v}_k\|^2_2 \leq 1\). Thus \(\hat{v}_k^T M_k \hat{v}_k \geq \lambda_k \cos^2 \Theta(\hat{v}_k, v_k)\).

Next, to get (S51), notice that

\[ v_k^T \hat{M}_k v_k = v_k^T M v_k + v_k^T (\hat{M}_k - M) v_k \]
\[ = \lambda_k - v_k^T (M - \hat{M}) v_k - v_k^T (\hat{M} - \hat{M}_k) v_k \]
\[ \geq \lambda_k - \epsilon \tau^2 - v_k^T (\hat{M} - \hat{M}_k) v_k, \]

which means that it remains to prove \(v_k^T (\hat{M} - \hat{M}_k) v_k \leq C\epsilon \tau^2\). Recall from (S46) that,

\[ \hat{M} - \hat{M}_k \]
\[ = \hat{M} - \{I_p - \hat{P}_{k-1} + \hat{P}_{k-1}\} \hat{M} \{I_p - \hat{P}_{k-1} + \hat{P}_{k-1}\} \]
\[ = \hat{P}_{k-1} \hat{M} + \hat{M}\hat{P}_{k-1} - \hat{P}_{k-1} \hat{M}\hat{P}_{k-1} \]
\[ - \hat{P}_{k-1} \hat{M}(I_p - \hat{P}_{k-1} + \hat{P}_{k-1}) - (I_p - \hat{P}_{k-1} + \hat{P}_{k-1}) \hat{M}\hat{P}_{k-1}. \tag{S53} \]

Follow the same calculation in the proof for Lemma [33] for \(Q_5 = Q_0\), we can see that

\[ v_k^T (\hat{M} - \hat{M}_k) v_k \leq v_k^T (\hat{P}_{k-1} \hat{M} + \hat{M}\hat{P}_{k-1} - \hat{P}_{k-1} \hat{M}\hat{P}_{k-1}) v_k + C\epsilon \tau^2 \]
\[ \leq v_k^T (\hat{P}_{k-1} \hat{M} + \hat{M}\hat{P}_{k-1}) v_k + C\epsilon \tau^2 \]
\[ = 2v_k^T \hat{P}_{k-1} \hat{M} v_k + C\epsilon \tau^2, \]

38
where \( \mathbf{v}_k^T \hat{\mathbf{P}}_{k-1} \mathbf{Mv}_k = \sum_{j<k}(\mathbf{v}_k^T \hat{\mathbf{v}}_j) \cdot (\mathbf{v}_j^T \mathbf{Mv}_k) \). Because \( \mathbf{v}_k^T \hat{\mathbf{v}}_j = \mathbf{v}_k^T (\hat{\mathbf{v}}_j - \mathbf{v}_j) \leq \sqrt{\epsilon \tau^2} \) and that
\[
\mathbf{v}_k^T \hat{\mathbf{P}}_{k-1} \mathbf{Mv}_k = \mathbf{v}_j^T \mathbf{Mv}_k + \mathbf{v}_j^T (\mathbf{M} - \mathbf{M}) \mathbf{v}_k \\
\leq \lambda_k (\mathbf{v}_j^T \mathbf{v}_k) + \epsilon \tau^2 \\
= C \sqrt{\epsilon \tau^2} \epsilon \tau^2 \\
\leq (C + 1) \sqrt{\epsilon \tau^2},
\]
we have \( \mathbf{v}_k^T \hat{\mathbf{P}}_{k-1} \mathbf{Mv}_k \leq C \epsilon \tau^2 \). This completes the proof.

\[ \square \]

### S6.2 Proof for Theorem 3

We prove Theorem 3 by induction.

For \( j = 1 \), the following results is obtained in Lemma S10,
\[
\sin^2 \Theta(\mathbf{v}_1, \hat{\mathbf{v}}_1) \leq C \epsilon \tau^2, \quad \eta_1 \geq 1 - C \epsilon \tau^2, \quad ||\mathbf{v}_1 - \hat{\mathbf{v}}_1||^2 \leq C \epsilon \tau^2.
\]

Suppose the following statements are true for all \( j < k \), and we prove it is also true for \( j = k \),
\[
\sin^2 \Theta(\mathbf{v}_j, \hat{\mathbf{v}}_j) \leq C \epsilon \tau^2, \quad \eta_j \geq 1 - C \epsilon \tau^2, \quad ||\mathbf{v}_j - \hat{\mathbf{v}}_j||^2 \leq C \epsilon \tau^2.
\]

First, we establish \( \sin^2 \Theta(\mathbf{v}_k, \hat{\mathbf{v}}_k) \leq C \epsilon \tau^2 \).

From Lemma S12 and the definition of \( \hat{\mathbf{v}}_k \), we know that
\[
\mathbf{v}_k^T \hat{\mathbf{M}}_k \hat{\mathbf{v}}_k \geq \mathbf{v}_k^T \mathbf{M} \mathbf{v}_k \geq \lambda_k - C \epsilon \tau^2.
\]

From Lemma S11 we let \( \alpha = \sin^2 \Theta(\hat{\mathbf{v}}_k, \mathbf{v}_k) \) and have
\[
\hat{\mathbf{v}}_k^T \hat{\mathbf{M}}_k \hat{\mathbf{v}}_k \leq \hat{\mathbf{v}}_k^T \hat{\mathbf{M}}_k \mathbf{v}_k \leq \lambda_k - \Delta \alpha + C \sqrt{\epsilon \tau^2} \alpha + C \tau^2 \epsilon.
\]

Comparing the lower and upper bound of \( \hat{\mathbf{v}}_k^T \hat{\mathbf{M}}_k \hat{\mathbf{v}}_k \), we have
\[
\lambda_k - C \epsilon \tau^2 \leq \lambda_k - \Delta \alpha + C \sqrt{\epsilon \tau^2} \alpha + C \tau^2 \epsilon,
\]
or equivalently,
\[
0 \geq \alpha - C \sqrt{\epsilon \tau^2} \sqrt{\alpha} - C \epsilon \tau^2 = (\sqrt{\alpha} - \sqrt{C \epsilon \tau^2})^2 - C \epsilon \tau^2,
\]
which implies
\[
\alpha \leq C \epsilon \tau^2.
\]

Next, we establish \( \eta_k = ||\hat{\mathbf{v}}_k||^2 \geq 1 - C \epsilon \tau^2 \).

From Lemma S12 we have
\[
\mathbf{v}_k^T \hat{\mathbf{M}}_k \mathbf{v}_k \geq \lambda_k - C \epsilon \tau^2.
\]
From Lemma S11, we have
\[ \hat{v}_k^T \hat{M}_k \hat{v}_k = \eta_k \hat{v}_k^T \hat{M}_k \hat{v}_k \leq \eta_k (\lambda_k + C \epsilon \tau^2), \]
where we have applied \( \sin^2 \Theta(\hat{v}_k, v_k) \leq C \epsilon \tau^2 \) to obtain the last inequality. Comparing the two equality and noticing that \( v_k^T \hat{M}_k v_k \leq \hat{v}_k^T \hat{M}_k \hat{v}_k \), we have
\[ \eta_k \geq \frac{\lambda_k - C \epsilon \tau^2}{\lambda_k + C \epsilon \tau^2} = 1 - \frac{C \epsilon \tau^2}{\lambda_k + C \epsilon \tau^2} \geq 1 - C \epsilon \tau^2. \]
Finally, we establish \( \|v_k - \hat{v}_k\|^2 \leq C \epsilon \tau^2 \).

S6.3 Proof for Theorem 4
By Lemma S6, if \( \Delta U > 2(\tau^2 + \frac{10 \nu}{\Delta}) \epsilon \), we keep all the \( \hat{v}_i \)'s in the envelope because the high-dimensional NIECE procedure keeps the correct ordering of directions \( \hat{v}_i \). Then the conclusion of Theorem 4 follows directly from Theorem 3 and Lemma S6.

S7 Proofs for Theorem 5 and Corollaries 2 and 3
Combine Lemma S14 with Theorem 4 and we have the desired conclusion. We have the Corollary 2 by letting \( \frac{1}{p} \geq \epsilon \gg \sqrt{\frac{\log p + \log r}{n}} \to 0 \) and \( \tau^2 \epsilon \to 0 \). Then Corollary 3 follows from Corollary 2 by fix \( p \) as constant and letting \( n, r \) diverge.

Lemma S13. Under sub-Gaussian condition, i.e. (16), for any \( \epsilon > 0 \) and some constant \( C > 0 \), we have
\[ \Pr(|\hat{\sigma}_{X,ij} - \sigma_{X,ij}| > \epsilon) \leq C \exp(-Cn \epsilon^2), \quad (S54) \]
where \( n \) is the sample size \( \hat{\sigma}_{X,ij} \) and \( \sigma_{X,ij} \) are the \((ij)\)-th element in \( p \times p \) sample and population covariance of \( X \in \mathbb{R}^p \).

Under joint sub-Gaussian of \((X, Y)\) in (7), for any \( \epsilon > 0 \) and some constant \( C > 0 \), we have
\[ \Pr(|\hat{\sigma}_{ij} - \sigma_{ij}| > \epsilon) \leq C \exp(-Cn \epsilon^2), \quad (S55) \]
where \( n \) is the sample size \( \hat{\sigma}_{ij} \) and \( \sigma_{ij} \) are the \((ij)\)-th element in \((p + r) \times (p + r)\) sample and population covariance of \( Z = (X^T, Y^T)^T \in \mathbb{R}^{p+r} \).
Proof. Let $U$ and $V$ be the $i$-th and $j$-th elements in $X$ or $Z = (X^T, Y^T)^T$. Then we want to show that
\[
\Pr(|\hat{\sigma}_{UV} - \sigma_{UV}| > \varepsilon) \leq \exp(-Cn\varepsilon^2),
\]
where $\hat{\sigma}_{UV} = n^{-1}\sum_{i=1}^{n} U_i V_i$. We can write
\[
\sigma_{UV} = \text{cov}(U, V) = \{\text{cov}(U + V) - \text{var}(U) - \text{var}(V)/2 \equiv (\sigma_{U+V} - \sigma_U - \sigma_V)/2,
\]
and, similarly,
\[
\hat{\sigma}_{UV} = (\hat{\sigma}_{U+V} - \hat{\sigma}_U - \hat{\sigma}_V)/2.
\]
Therefore,
\[
\Pr(|\hat{\sigma}_{UV} - \sigma_{UV}| > \varepsilon) \leq \Pr(|\hat{\sigma}_{U+V} - \sigma_{U+V}| > \varepsilon) + \Pr(|\hat{\sigma}_U - \sigma_U| > \varepsilon) + \Pr(|\hat{\sigma}_V - \sigma_V| > \varepsilon).
\]
Because $U$, $V$ and $(U + V)$ are all sub-Gaussian, the random variables $\hat{\sigma}_U$, $\hat{\sigma}_V$ and $\hat{\sigma}_{U+V}$ are sums of i.i.d. subexponential random variables. By equation (2.4) in [3], we have
\[
\Pr(|\hat{\sigma}_W - \sigma_W| > \varepsilon) \leq 2 \exp(-cn\varepsilon^2),
\]
where $W$ is either $U$, $V$ or $(U + V)$. This completes the proof.

Lemma S14. 1. For $\varepsilon > 0$, we have
\[
\|\hat{\Sigma}_Y - \Sigma_Y\|_{\max} \leq \varepsilon
\]
with a probability greater than $1 - C r^2 \exp(-Cn\varepsilon^2)$.

2. For $0 < \varepsilon \leq 1/p$, we have
\[
\|\hat{\Sigma}_{YX}\hat{\Sigma}_{XY} - \Sigma_{YX}\Sigma_{XY}\|_{\max} \leq \varepsilon
\]
with a probability greater than $1 - C pr \exp(-Cn\varepsilon^2) - C pr \exp(-Cn/p^2)$.

Proof. We first show (S57). Note that
\[
\Pr(\|\hat{\Sigma}_Y - \Sigma_Y\|_{\max} > \varepsilon) = \Pr(\cup_{i,j}|\hat{\sigma}_{Y,i,j} - \sigma_{Y,i,j}| > \varepsilon) \leq Cr^2 \exp(-Cn\varepsilon^2),
\]
where the last inequality follows from joint sub-Gaussian of $(X, Y)$ in (7) and Lemma S13.

Now we show (S58). Straightforward calculation shows that
\[
\|\hat{\Sigma}_{YX}\hat{\Sigma}_{XY} - \Sigma_{YX}\Sigma_{XY}\|_{\max}
\leq \|\Sigma_{YX}(\hat{\Sigma}_{XY} - \Sigma_{XY})\|_{\max} + \|\Sigma_{YX}(\hat{\Sigma}_{XY} - \Sigma_{XY})\|_{\max}
\leq 2\|\Sigma_{YX}(\hat{\Sigma}_{XY} - \Sigma_{XY})\|_{\max} + \|\Sigma_{YX}(\hat{\Sigma}_{XY} - \Sigma_{XY})\|_{\max}.
\]
Under the event that $\|\hat{\Sigma}_{YX} - \Sigma_{YX}\|_{\max} \leq \varepsilon$, we have
\[
\|\Sigma_{YX}(\hat{\Sigma}_{XY} - \Sigma_{XY})\|_{\max} \leq \|\hat{\Sigma}_{YX} - \Sigma_{YX}\|_{\max}\|\Sigma_{XY}\|_{1} \leq \|\Sigma_{XY}\|_{1}\varepsilon,
\]
41
and if we further assume that $\| \hat{\Sigma}_{YX} - \Sigma_{YX} \|_{\max} = \max_{ij} | \hat{\sigma}_{YX,ji} - \sigma_{YX,ji} | \leq \frac{1}{p}$,

$$
\| (\hat{\Sigma}_{YX} - \Sigma_{YX})(\hat{\Sigma}_{XY} - \Sigma_{XY}) \|_{\max} = \sup_{j,k} \left\{ \sum_{i=1}^{p} (\hat{\sigma}_{YX,ji} - \sigma_{YX,ji})(\hat{\sigma}_{YX,ki} - \sigma_{YX,ki}) \right\} \\
\leq \sup_{j,k} \left\{ \sum_{i=1}^{p} (\hat{\sigma}_{YX,ji} - \sigma_{YX,ji}) \cdot \frac{1}{p} \right\} \\
\leq \| \hat{\Sigma}_{YX} - \Sigma_{YX} \|_{\max} \leq \epsilon.
$$

The probability of the event $\| \hat{\Sigma}_{YX} - \Sigma_{YX} \|_{\max} \leq \min\{\epsilon, \frac{1}{p}\}$ can be calculated as,

$$
\Pr\left( | \hat{\sigma}_{YX,jk} - \sigma_{YX,jk} | \leq \min\{\epsilon, \frac{1}{p}\} \text{ for all } j,k \right) \geq 1 - Cpr \exp\left(-Cn\epsilon^2\right) - Cpr \exp\left(-Cn/p^2\right).
$$

And we have the desired conclusion.

\[\square\]

**S8 Proofs for Theorem 6**

Theorem 6 follows directly from the following Lemma S15 and Theorem 4.

Recall from Theorem 4 that we need to verify the following conditions,

$$
\| \hat{M} - M \|_{\max} \leq \epsilon, \quad \| \hat{U} - U \|_{\max} \leq \epsilon, \quad \Delta_U > \sqrt{8c_0\nu^2\tau^2\epsilon} + (2c_0\nu + 1)\tau^2\epsilon.
$$

From the sub-Gaussian distribution of $X$, i.e. (16), we have $\| \hat{M} - M \|_{\max} \leq \epsilon$ holds with probability at least $1 - C\epsilon^2\exp(-Cn\epsilon^2)$. From the following Lemma S15, we have $\| \hat{U} - U \|_{\max} \leq \epsilon$ holds with probability at least $1 - C\epsilon^2\exp(-Cn\epsilon^2/s)$ if $\lambda_n = 4\epsilon/\sqrt{s}$.

Finally, to achieve $\| P_{\hat{\epsilon}} - P_{\epsilon} \|_F \rightarrow 0$ with probability going to 1, we could let

$$
\frac{s \log p}{n\epsilon^2} \rightarrow 0, \quad \frac{ne^2}{s} \rightarrow \infty, \quad \lambda_n = \frac{4\epsilon}{\sqrt{s}}, \quad \Delta_U \rightarrow 0, \quad \tau^2\epsilon \rightarrow 0,
$$

which are satisfied as we let

$$
\sqrt{\frac{s \log p}{n}} \rightarrow 0, \quad \sqrt{\frac{s \log p}{n}} \ll \epsilon \ll \tau^{-2} \ll 1, \quad \sqrt{\frac{\log p}{n}} \ll \lambda_n, \quad \Delta_U^{-\tau^2} \sqrt{\frac{s \log p}{n}} \rightarrow 0.
$$

**Lemma S15.** If $\lambda_n = 4\epsilon$, with a probability greater than $1 - C\epsilon^2\exp(-Cn\epsilon^2)$, we have

$$
\| \hat{\beta} - \beta \|_2^2 \leq C\epsilon^2, \quad \| \hat{U} - U \|_{\max} \leq C(\sqrt{s}\epsilon^2 + s\epsilon^2),
$$

(S60)

where $s$ is the number of important predictors.

**Proof.** By Corollary 1 in Negahban et al. (2012) and the results in Section 4.4 in their paper, it suffices to verify that

$$
\lambda_n \geq 2\| \nabla \{-\frac{1}{n}\ell_n(\beta)\} \|_{\infty} = 2\| \left( -\frac{1}{n} \sum_{i=1}^{n} \left\{ Y_iX_i - \frac{e^{X_i^T\beta}}{1 + e^{X_i^T\beta}}X_i \right\} \right) \|_{\infty}
$$

(S61)
with the specified probability.

Note that \( -\frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i X_i - \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}} X_i \right\} \) is the derivative of the (negative) log likelihood function. It follows that its expectation is 0. Hence,

\[
\Pr\left( -\frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i X_i - \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}} X_i \right\} \geq \epsilon \right) \leq \Pr\left( |\frac{1}{n} \sum_{i=1}^{n} Y_i X_i - E(Y_i X_i)| \geq \epsilon/2 \right) + \Pr\left( |\frac{1}{n} \sum_{i=1}^{n} e^{X_i^T \beta} X_i - E(e^{X_i^T \beta} X_i)| \geq \epsilon/2 \right) \leq C_p \exp(-C n \epsilon^2)
\]

where we use the fact that, since \( X_i \) is sub-Gaussian, and \( Y_i \) and \( \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}} \) are bounded, we have

\( Y_i X_i \) and \( \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}} X_i \) are also sub-Gaussian.

Since \( U = \beta \beta^T \) and \( \hat{U} = \hat{\beta} \hat{\beta}^T \) where \( \hat{\beta} \) is the solution to the \( \ell_1 \)-penalized maximum likelihood estimator (18).

\[
\| \hat{U} - U \|_{\text{max}} = \| (\hat{\beta} - \beta) (\hat{\beta} - \beta)^T \|_{\text{max}} + 2 \| (\hat{\beta} - \beta) \beta^T \|_{\text{max}} \\
\leq \| \hat{\beta} - \beta \|_2^2 + 2 \| \hat{\beta} - \beta \|_\infty \| \beta \|_\infty \\
\leq \| \hat{\beta} - \beta \|_2^2 + 2 \| \hat{\beta} - \beta \|_2 \| \beta \|_2 \\
= \| \hat{\beta} - \beta \|_2^2 + 2 \sqrt{\nu} \| \hat{\beta} - \beta \|_2,
\]

where recall that \( \nu = \| U \|_{\text{op}} = \beta^T \beta \).

\[\Box\]

### S9 Proofs for Theorem 7

Analogous to Lemma S15 for the GLM, we have the following Lemma S16 for the Cox model. Then the proof of Theorem 7 follows the same argument as in the proof for Theorem 6.

**Lemma S16.** There exists some constant \( C_\rho \) and \( C_\lambda \) such that if \( s \sqrt{\frac{\log p}{n}} < C_\rho \) and \( \lambda_n = C_\lambda \epsilon \), with a probability greater than \( 1 - C_p \exp(-C n \epsilon^2) \), we have

\[
\| \hat{\beta} - \beta \|_2^2 \leq C s \epsilon^2, \quad \| \hat{U} - U \|_{\text{max}} \leq C(\sqrt{s} \epsilon^2 + s \epsilon^2),
\]

(S62)

where \( s \) is the number of important predictors.

**Proof.** By Theorem 3.2. in [Huang et al. (2013)](http://example.com), we have the following results for Lasso penalized Cox model under the uniformly bounded covariates and restricted eigenvalue lower bounds conditions. Replacing \( \epsilon \) in Theorem 3.2 of [Huang et al. (2013)](http://example.com) by \( 2 p \exp(-C n \epsilon^2) \), we have the following results.
Let $\lambda_n = C\epsilon$ for some constant that does not depend on $n$, $p$, or $s$. For any constant $C_F > 0$, there exists some constant $C$ such that,

$$\|\hat{\beta} - \beta\|_2 \leq C\sqrt{s\epsilon/C_F},$$

with probability at least $\Pr(A) - 2p\exp(-Cn\epsilon^2)$. The event $A$ is the event that the compatibility and cone invertibility factors are bounded below by the constants $C_F$. Then by Theorem 4.1 of Huang et al. (2013), we can have event $A$ by setting $C_F = \rho^*/2$. More specifically, replacing $\epsilon$ in Theorem 4.1 of Huang et al. (2013) by $p\exp(-cn\epsilon^2)$, we have the following results: event $A$ holds with probability at least $1 - C\rho^*\exp(-Cn\epsilon^2) + 2p\exp(-Cn\epsilon^2)$ if $\rho^*$ is a constant greater than 0 and $s\sqrt{\log p/n} < C\rho$. Finally, $\rho^*$ is defined as the smallest eigenvalue of the population matrix $\Sigma(t, M)$ for some positive constants $t$ and $M$. The accurate definition of $\Sigma(t, M)$ is given as follows.

Following Andersen & Gill (1982), let $N^{(n)}(t) = (N_1(t), \ldots, N_n(t)), t \geq 0$, be the $n$-dimensional counting process where $N_i(t)$ counts the number of observed events for the $i$-th individual and has predictable compensator $\Lambda^{(n)} = (\Lambda_1, \ldots, \Lambda_n)$ defined as $d\Lambda_i(t) = Y_i(t) \exp(\beta^T X_i) d\Lambda_0(t)$, where $\Lambda_0(t)$ is the unknown baseline cumulative hazard function and $Y_i(t) \in \{0, 1\}$ is a predictable at risk indicator process. The matrix $\Sigma(t, M)$ for some constants $t$ and $M$ is essentially a truncated approximation to the Hessian $\nabla^2 \ell_n(\beta)$ up to time $t$, defined as,

$$\Sigma(t, M) = E \int_0^t G_n(s; M) d\Lambda_0(s) \quad G_n(t, M) = \frac{1}{n} \sum_{i=1}^n \{X_i - \mu(t, M)\}^T U_i(t)\{X_i - \mu(t, M)\}$$

$$U_i(t) = Y_i(t) \min\{M, \exp(\beta^T X_i)\} \quad \mu(t, M) = E\{XU(t)\}/E\{U(t)\}.$$  

As noted in Huang et al. (2013), since $\Sigma(t, M)$ is a population matrix, it is reasonable to assume its smallest eigenvalue is a positive constants. Moreover, we only need this to be true for some $t$ and $M$ in the condition.

Then follow the proof of Lemma [S15] we have

$$\|\hat{U} - U\|_{\max} \leq \|\hat{\beta} - \beta\|_2^2 + 2\sqrt{\nu}\|\hat{\beta} - \beta\|_2,$$

which completes the proof. \qed
References

Absil, P.-A., Mahony, R. & Sepulchre, R. (2009), *Optimization algorithms on matrix manifolds*, Princeton University Press.

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

Andersen, P. K. & Gill, R. D. (1982), ‘Cox’s regression model for counting processes: a large sample study’, *The annals of statistics* pp. 1100–1120.

Bradic, J., Fan, J. & Jiang, J. (2011), ‘Regularization for cox’s proportional hazards model with np-dimensionality’, *The Annals of Statistics 39*(6), 3092–3120.

Bro, R., Kjeldahl, K., Smilde, A. K. & Kiers, H. (2008), ‘Cross-validation of component models: a critical look at current methods’, *Analytical and bioanalytical chemistry 390*(5), 1241–1251.

Bühlmann, P., Kalisch, M. & 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*(6), 3074–3110.

Chen, K., Chan, K.-S. & 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*(2), 203–221.

Chen, L. & Huang, J. Z. (2012), ‘Sparse reduced-rank regression for simultaneous dimension reduction and variable selection’, *Journal of the American Statistical Association 107*(500), 1533–1545.

Chen, X., Zou, C. & Cook, R. D. (2010), ‘Coordinate-independent sparse sufficient dimension reduction and variable selection’, *The Annals of Statistics 38*(6), 3696–3723.

Chun, H., Ballard, D. H., Cho, J. & Zhao, H. (2011), ‘Identification of association between disease and multiple markers via sparse partial least-squares regression’, *Genetic epidemiology 35*(6), 479–486.

Chun, H. & Kele¸s, 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*(1), 3–25.

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

Cook, R. D. (2018), *An introduction to envelopes: dimension reduction for efficient estimation in multivariate statistics*, Vol. 401, John Wiley & Sons.

Cook, R. D. (2020), ‘Envelope methods’, *Wiley Interdisciplinary Reviews: Computational Statistics 12*(2), e1484.

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

Cook, R. D., Forzani, L. & Zhang, X. (2015), ‘Envelopes and reduced-rank regression’, *Biometrika 102*(2), 439–456.

Cook, R. D., Forzani, L. et al. (2019), ‘Partial least squares prediction in high-dimensional regression’, *The Annals of Statistics 47*(2), 884–908.
Cook, R. D., Helland, I. S. & Su, Z. (2013a), ‘Envelopes and partial least squares regression’, J. R. Stat. Soc. Ser. B. Stat. Methodol. 75(5), 851–877.

Cook, R. D., Li, B. & Chiromonte, F. (2010), ‘Envelope models for parsimonious and efficient multivariate linear regression’, Statist. Sinica 20(3), 927–960.

Cook, R. D. & Su, Z. (2013), ‘Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression’, Biometrika 100(4), 939–954.

Cook, R. D. & Zhang, X. (2015a), ‘Foundations for envelope models and methods’, Journal of the American Statistical Association 110(510), 599–611.

Cook, R. D. & Zhang, X. (2015b), ‘Simultaneous envelopes for multivariate linear regression’, Technometrics 57(1), 11–25.

Cook, R. D. & Zhang, X. (2016), ‘Algorithms for envelope estimation’, Journal of Computational and Graphical Statistics 25(1), 284–300.

Cook, R. D. & Zhang, X. (2018), ‘Fast envelope algorithms’, Statistica Sinica 28, 1179–1197.

Cook, R., Helland, I. & Su, Z. (2013b), ‘Envelopes and partial least squares regression’, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75(5), 851–877.

Cox, D. R. (1972), ‘Regression models and life-tables’, Journal of the Royal Statistical Society. Series B (Methodological) 34(2), 87–22.

Cox, D. R. (1975), ‘Partial likelihood’, Biometrika 62(2), 269–276.

De Jong, S. (1993), ‘Simpls: an alternative approach to partial least squares regression’, Chemometrics and intelligent laboratory systems 18(3), 251–263.

Ding, S. & Cook, R. (2018), ‘Matrix variate regressions and envelope models’, Journal of the Royal Statistical Society. Series B: Statistical Methodology 80(2), 387–408.

Ding, S., Su, Z., Zhu, G. & Wang, L. (2021), ‘Envelope quantile regression’, Statistica Sinica 31, 79–106.

Edelman, A., Arias, T. A. & Smith, S. T. (1998), ‘The geometry of algorithms with orthogonality constraints’, SIAM journal on Matrix Analysis and Applications 20(2), 303–353.

Fan, J. & Li, R. (2001), ‘Variable selection via nonconcave penalized likelihood and its oracle properties’, Journal of the American statistical Association 96(456), 1348–1360.

Flury, B. (1998), Common principal components & related multivariate models, John Wiley & Sons, Inc.

Flury, B. N. (1984), ‘Common principal components in k groups’, Journal of the American Statistical Association 79(388), 892–898.

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

Franks, A. M. & Hoff, P. (2019), ‘Shared subspace models for multi-group covariance estimation.’, Journal of Machine Learning Research 20(171), 1–37.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), ‘Regularization paths for generalized linear models via coordinate descent’, Journal of statistical software 33(1), 1.
Helland, I. S. (1990), ‘Partial least squares regression and statistical models’, *Scand. J. Statist.* **17**(2), 97–114.

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

Huang, J., Sun, T., Ying, Z., Yu, Y. & Zhang, C.-H. (2013), ‘Oracle inequalities for the lasso in the cox model’, *Annals of statistics* **41**(3), 1142.

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

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**(3), 300–303.

Jolliffe, I. T. (1986), Principal components in regression analysis, in ‘Principal component analysis’, Springer, pp. 129–155.

Jolliffe, I. T. & Cadima, J. (2016), ‘Principal component analysis: a review and recent developments’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**(2065), 20150202.

Josse, J. & Husson, F. (2012), ‘Selecting the number of components in principal component analysis using cross-validation approximations’, *Computational Statistics & Data Analysis* **56**(6), 1869–1879.

Khare, K., Pal, S., Su, Z. et al. (2017), ‘A bayesian approach for envelope models’, *The Annals of Statistics* **45**(1), 196–222.

Lang, W. & Zou, H. (2020), ‘A simple method to improve principal components regression’, *Stat* p. e288.

Lee, M. & Su, Z. (2020), ‘A review of envelope models’, *International Statistical Review* **88**(3), 658–676.

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

Li, L. & Zhang, X. (2017), ‘Parsimonious tensor response regression’, *Journal of the American Statistical Association* **112**(519), 1131–1146.

Ma, Z. (2013), ‘Sparse principal component analysis and iterative thresholding’, *Ann. Statist.* **41**, 772–801.

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

Nygård, S., Borgan, Ø., Lingjærde, O. C. & Størvold, H. L. (2008), ‘Partial least squares cox regression for genome-wide data’, *Lifetime Data Analysis* **14**(2), 179–195.

Oja, E. (1992), ‘Principal components, minor components, and linear neural networks’, *Neural networks* **5**(6), 927–935.

Sæbø, S., Almøy, T., Aarøe, J. & 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**(1), 54–62.

Schott, J. R. (1999), ‘Partial common principal component subspaces’, *Biometrika* **86**(4), 899–908.
Shen, H. & Huang, J. Z. (2008), ‘Sparse principal component analysis via regularized low rank matrix approximation’, *Journal of multivariate analysis* **99**(6), 1015–1034.

Su, Z. & Cook, R. D. (2011), ‘Partial envelopes for efficient estimation in multivariate linear regression’, *Biometrika* **98**(1), 133–146.

URL: [http://dx.doi.org/10.1093/biomet/asq063](http://dx.doi.org/10.1093/biomet/asq063)

Su, Z. & Cook, R. D. (2012), ‘Inner envelopes: efficient estimation in multivariate linear regression’, *Biometrika* **99**(3), 687–702.

Su, Z., Zhu, G., Chen, X. & Yang, Y. (2016), ‘Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression’, *Biometrika* **103**(3), 579–593.

URL: [http://dx.doi.org/10.1093/biomet/asw036](http://dx.doi.org/10.1093/biomet/asw036)

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

Tipping, M. E. & Bishop, C. M. (1999), ‘Probabilistic principal component analysis’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(3), 611–622.

Vershynin, R. (2012), ‘How close is the sample covariance matrix to the actual covariance matrix?’, *Journal of Theoretical Probability* **25**(3), 655–686.

Vu, V. Q. & Lei, J. (2013), ‘Minimax sparse principal subspace estimation in high dimensions’, *The Annals of Statistics* **41**(6), 2905–2947.

Welling, M., Williams, C. & Agakov, F. V. (2004), Extreme components analysis, *in ‘Advances in neural information processing systems’*, pp. 137–144.

Wen, Z. & Yin, W. (2013), ‘A feasible method for optimization with orthogonality constraints’, *Mathematical Programming* **142**(1-2), 397–434.

Witten, D. M., Tibshirani, R. & Hastie, T. (2009), ‘A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis’, *Biostatistics* p. kxp008.

Wold, H. (1966), ‘Estimation of principal components and related models by iterative least squares’, *Multivariate analysis* pp. 391–420.

Yu, Y., Wang, T. & Samworth, R. J. (2014), ‘A useful variant of the davis–kahan theorem for statisticians’, *Biometrika* **102**(2), 315–323.

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

Zhang, X. & Mai, Q. (2018), ‘Model-free envelope dimension selection’, *Electronic Journal of Statistics* **12**(2), 2193–2216.

Zhang, X. & Mai, Q. (2019), ‘Efficient integration of sufficient dimension reduction and prediction in discriminant analysis’, *Technometrics* **61**(2), 259–272.

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

Zhou, L., Cook, R. D. & Zou, H. (2020), ‘Enveloped huber regression’, *arXiv preprint arXiv:2011.00119*.

Zhu, G. & Su, Z. (2020), ‘Envelope-based sparse partial least squares’, *The Annals of Statistics* **48**(1), 161–182.

Zou, H., Hastie, T. & Tibshirani, R. (2006), ‘Sparse principal component analysis’, *Journal of Computational and Graphical Statistics* **15**, 265–286.