Electronic Journal of Statistics
Vol. 12 (2018) 2193–2216
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1449

Model-free envelope dimension selection

Xin Zhang and Qing Mai

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

Abstract: An envelope is a targeted dimension reduction subspace for simultaneously achieving dimension reduction and improving parameter estimation efficiency. While many envelope methods have been proposed in recent years, all envelope methods hinge on the knowledge of a key hyperparameter, the structural dimension of the envelope. How to estimate the envelope dimension consistently is of substantial interest from both theoretical and practical aspects. Moreover, very recent advances in the literature have generalized envelope as a model-free method, which makes selecting the envelope dimension even more challenging. Likelihood-based approaches such as information criteria and likelihood-ratio tests either cannot be directly applied or have no theoretical justification. To address this critical issue of dimension selection, we propose two unified approaches – called FG and 1D selections – for determining the envelope dimension that can be applied to any envelope models and methods. The two model-free selection approaches are based on the two different envelope optimization procedures: the full Grassmannian (FG) optimization and the 1D algorithm [11], and are shown to be capable of correctly identifying the structural dimension with a probability tending to 1 under mild moment conditions as the sample size increases. While the FG selection unifies and generalizes the BIC and modified BIC approaches that existing in the literature, and hence provides the theoretical justification of them under weak moment condition and model-free context, the 1D selection is computationally more stable and efficient in finite sample. Extensive simulations and a real data analysis demonstrate the superb performance of our proposals.

Keywords and phrases: Dimension reduction, envelope models and methods, information criterion, model selection.

Received September 2017.

1. Introduction

Envelope methods provide means to achieve sufficient dimension reduction and estimation efficiency on a wide range of multivariate statistics problems. The first envelope method was introduced by Cook et al. [7] in multivariate linear regression to gain efficiency in parameter estimation. Various types of envelope models have been further proposed in multivariate linear regression [22, 5, 10, 4, etc.]. More recently, Cook and Zhang [9] proposed a new definition and framework of envelope that adapted envelope methods to any multivariate parameter estimation procedure. Envelope methods now can be constructed in the model-free context, and are no longer restricted to likelihood-based estimation or stringent regression model assumptions. This greatly facilitates further adaptations of envelope methods to many potential fields such as tensor decomposition and regression with neuroimaging applications [17, 25], Aster models for life history analysis [15, 14], etc.

Most envelope methods rely on the knowledge of the envelope dimension (perhaps an exception is to apply bootstrap model averaging over the envelope dimension to estimate the regression coefficient [13]). However, selecting envelope dimension is a theoretically challenging but crucial issue that becomes a severe nag in applications. Even for likelihood-based envelope methods, where information criteria and likelihood-ratio tests are widely used, no theoretical justification is known when the likelihood is mis-specified. To the best of our knowledge, all existing envelope dimension selection procedures in the literature fall into two categories – either (1) theoretically justified procedures that relying on strong model and distributional assumptions, or, (2) selection procedures based on heuristics such as cross-validation and heuristic information criteria. For example, Schott [20] provided some pioneering results for likelihood-ratio tests, [10] developed a sequential asymptotic χ^2-test based on rank estimation from Bura and Cook [2], and Cook and Su [8] have shown model selection consistency of BIC under their scaled envelope model and normally distributed errors. All such procedures require the linear model assumption and normality assumptions on either the error or even on the joint distribution of $(\mathbf{Y}, \mathbf{X})$. It is thus difficult to generalize such approaches to the model-free context and to justify such approaches without normality assumptions. On the other hand, information criteria such as AIC [1] and BIC [21] are widely used in envelope literature ever since the first paper in envelope [7]. More recently, Li and Zhang [17] proposed a modified BIC criterion for the more complicated tensor envelope regression models to estimate the dimension of tensor envelopes on each mode of the tensor. Unfortunately, there is no theoretical justification for BIC or modified BIC in envelope models while the normal assumption or the model assumption is violated. Specifically, without the normality assumption, the envelope estimator is still applicable and is $\sqrt{n}$-consistent estimator for the parameter of interest if we know the true dimension of the envelope, but there is no theory or method available (to the best of our knowledge) for selecting the envelope dimension consistently without relying on the normality assumption or the likelihood. One motivation of this paper is to formally address the theoretical challenges in envelope dimension selection without requiring distributional or model assumptions.

Given the dimension of an envelope, envelope estimation generally reduces to solving for $\widehat{\mathbf{\Gamma}}$ that minimizes the objective function of the following form,

$$\mathrm{J}(\mathbf{\Gamma}) = \log \mid \mathbf{\Gamma}^{\top}\mathbf{M}\mathbf{\Gamma} \mid + \log \mid \mathbf{\Gamma}^{\top}(\mathbf{M}+\mathbf{U})^{-1}\mathbf{\Gamma} \mid, \tag{1.1}$$

where $\mathbf{M},\ \mathbf{U} \in \mathbb{R}^{p\times p}$ are symmetric matrices such that $\mathbf{M} > 0$ and $\mathbf{U} \geq 0$, and $\mathbf{\Gamma} \in \mathbb{R}^{p\times u}$ is the basis matrix such that $\mathbf{\Gamma}^{\top}\mathbf{\Gamma} = \mathbf{I}_u$ with u being the dimension of $\mathbf{M}$-envelope of $\mathbf{U}$ (cf. Definition 2.2). We will review the formal definitions and estimation procedures in Section 2.

In multivariate linear envelope models, it can be shown [9] that the partially maximized log-likelihood function is $(-n/2)$ times certain sample version of (1.1). Furthermore, given a parameter vector of interests $\boldsymbol{\theta} \in \mathbb{R}^p$ and some standard $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\theta}}$, the particular choice of $\mathbf{U} = \boldsymbol{\theta}\boldsymbol{\theta}^{\intercal}$ and $\mathbf{M}$ being the asymptotic covariance of $\widehat{\boldsymbol{\theta}}$ reproduces the envelope methods in the literature. The envelope estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^{\intercal}\widehat{\boldsymbol{\theta}}$ is asymptotically more efficient than the standard estimator $\widehat{\boldsymbol{\theta}}$ in various contexts such as linear and generalized linear models. Such envelope estimation, which solves for $\widehat{\boldsymbol{\Gamma}}$ based on (1.1) and then plugs-in $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^{\intercal}\widehat{\boldsymbol{\theta}}$, is essentially a two-stage projection pursuit multivariate parameter estimation relying on a generic objective function of envelope basis. Our new formulation in Section 3.1 offers a way of viewing model-free envelope estimation as an alternative quasi-likelihood approach involving a key matrix $\mathbf{M}$, a parameter of interest $\boldsymbol{\theta}$, and a feasible parametrization set $\mathcal{A}_k$ for optimization. This connection greatly deepens our understanding of model-free envelope methods, and is the key to prove consistency in envelope dimension selection without any stringent normality assumptions as in the literature. It shows that even when no likelihood function is available, we can construct a quasi-likelihood based on methods of moments. We expect that this connection will also facilitate the construction of envelope methods in future research, especially when a likelihood function is not available.

The non-convex optimization with orthogonality constraints in (1.1) is difficult to solve and has no explicit solution, and thus brings both computational and theoretical challenges for envelope model estimation and inference. To address the dimension selection problem, it is desirable to combine the efficient computational methods with the selection criteria. In this paper, we propose two unified and model-free envelope dimension selection procedures that are applicable to any envelope methods, either model-based or model-free, and suitable for any envelope estimation, either likelihood-based or moment-based. Consistency in selecting the envelope dimension is established for both procedures under mild moment conditions and without requiring any particular models.

The first one is called the FG procedure, based on fully optimizing the envelope objective function over a sequence of Grassmannians with increasing dimensions. The FG procedure is closely related to the BIC and is shown to include the BIC and the modified BIC [17] as special cases. Thus it provides solid theoretical justifications for the popular use of BIC in envelope dimension selection under non-normality and potential model mis-specifications.

The second one is called the 1D procedure, based on a recent envelope algorithm [11, cf. 1D algorithm] that sequentially optimizes a series of objective functions derived from (1.1) over one-dimensional Grassmannians. This can lead to faster, more accurate and stable envelope estimation and dimension selection. Moreover, because the FG envelope estimation can not guarantee "nested" envelope subspace estimates with increasing dimensions, the sequentially nested 1D envelope subspace estimates become even more desirable for its computational simplicity and stability. However, as one of our interesting theoretical findings, simply plugging in the results from 1D algorithm into the FG criterion (or BIC/modified BIC) will not guarantee consistency in selecting envelope dimensions. We thus proposed a new 1D criterion and established consistency in envelope dimension selection with it.

The contributions of this paper are multi-fold. First of all, ever since the introduction of envelope methods [7], there lacks a theoretically well justified approach to selecting its structural dimension in practice. Although [7] suggested that an information criterion like AIC or BIC may be used to select the structural dimension, no theoretical results were presented to show that such an approach leads to consistent selection if the normality assumption is dropped. In the later papers, BIC has also been applied or modified [17, e.g.] as a working method to select the dimension beyond linear models, while no study exists on the consistency of the BIC type selection. Our paper closes these theoretical gaps for the first time in this research area. Our results complement the existing papers on envelope methods by providing theoretical support to their data analysis. Our studies overcome some major difficulties since we do not rely on any likelihood or model assumptions. Now all the moment-based and the model-free envelope methods (and even future envelope methods) are finally completed with a properly justified model selection criterion. Secondly, our new formulation in Section 3.1 offers a way of viewing model-free envelope estimation as an alternative quasi-likelihood approach that facilitate the construction of envelope methods in future research, especially when a likelihood function is not available. Thirdly, the FG and 1D selection criteria proposed in this paper are tied to the estimation methods in the sense that the FG criterion must be applied with the FG estimator and the 1D criterion must be applied with the 1D estimator. Plugging in an arbitrary root-n consistent estimator into either criteria will generally not guarantee consistency in envelope dimension selection. The link between the estimation methods and selection criteria offers a crucial guidance in practice.

2. Review of envelopes and the 1D algorithm

We first review the definitions of reducing subspace and envelope. Besides being the basis for envelope methods, the concept of reducing subspace is also commonly used in functional analysis [3], but the notion of "reduction" differs from the usual understanding in statistics.

Definition 2.1. *(Reducing Subspace) A subspace $\mathcal{R} \subseteq \mathbb{R}^p$ is said to be a reducing subspace of $\mathbf{M} \in \mathbb{R}^{p\times p}$ if $\mathcal{R}$ decomposes $\mathbf{M}$ as $\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$, where $\mathbf{P}_{\mathcal{R}}$ is the projection matrix onto $\mathcal{R}$ and $\mathbf{Q}_{\mathcal{R}} = \mathbf{I}_p - \mathbf{P}_{\mathcal{R}}$ is the projection onto $\mathcal{R}^{\perp}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$ reduces $\mathbf{M}$.*

Definition 2.2. *(Envelope) The $\mathbf{M}$-envelope of $\mathrm{span}(\mathbf{U})$, denoted by $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$, is the intersection of all reducing subspaces of $\mathbf{M} > 0$ that contain $\mathrm{span}(\mathbf{U})$.*

It can be shown that $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ is unique and always exists. The dimension of $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$, denoted by u, $0 \leq u \leq p$, is important for all envelope methods. A smaller u usually indicates more efficiency gain can be achieved by taking the advantage of envelope structures.

To see the advantages of envelopes, consider the classical multivariate linear model as an example,

$$\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{Y}_i \in \mathbb{R}^{p\times 1}$ is the multivariate response, $\boldsymbol{\varepsilon}_i \sim N(0_p, \boldsymbol{\Sigma})$ is independent of $\mathbf{X}_i \in \mathbb{R}^q$. To estimate $\boldsymbol{\beta} \in \mathbb{R}^{p\times q}$, Cook et al. [7] seeks the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) \subseteq \mathbb{R}^p$ (cf. Definition 2.2). Let $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$ be a semi-orthogonal basis matrix of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$, whose orthogonal completion is $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p\times(p-u)}$. The definition of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ has two implications: (1) $\boldsymbol{\Gamma}^{\mathsf{T}}\mathbf{Y}$ contains all the information about $\boldsymbol{\beta}$ because $\boldsymbol{\beta}$ resides in $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$; (2) $\boldsymbol{\Gamma}^{\mathsf{T}}\mathbf{Y}$ is independent of $\boldsymbol{\Gamma}_0^{\mathsf{T}}\mathbf{Y}$ given $\mathbf{X}$ because by Definition 2.1, we can write $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathsf{T}} + \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^{\mathsf{T}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\mathsf{T}} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^{\mathsf{T}}$ for some $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. Hence, we can safely reduce immaterial variability in the data by eliminating $\boldsymbol{\Gamma}_0^{\mathsf{T}}\mathbf{Y}$. Consequently, the envelope estimator promotes efficiency in estimation.

We emphasize that the application of envelopes do not rely on the regression model (2.1). Definition 2.2 is generic and only involves two matrices $\mathbf{M}$ and $\mathbf{U}$. In a general statistical estimation problem of some parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$, Cook and Zhang [9] generalized the notion of envelopes as a way to improve some "standard" existing $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\theta}}$. In such general cases where the likelihood function need not be known, they proposed to construct the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ with $\mathbf{U} = \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}$ and $\mathbf{M}$ being the asymptotic covariance of $\widehat{\boldsymbol{\theta}}$. To obtain a semi-orthogonal basis matrix estimate for the envelope, $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) = \mathcal{E}_{\mathbf{M}}(\boldsymbol{\theta})$, we solve for $\widehat{\boldsymbol{\Gamma}} \in \mathbb{R}^{p\times u}$ that minimizes the generic moment-based objective function:

$$\mathrm{J}_n(\boldsymbol{\Gamma}) = \log \mid \boldsymbol{\Gamma}^{\mathsf{T}}\widehat{\mathbf{M}}\boldsymbol{\Gamma} \mid + \log \mid \boldsymbol{\Gamma}^{\mathsf{T}}(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1}\boldsymbol{\Gamma} \mid . \tag{2.2}$$

Given the true envelope dimension u, the $\sqrt{n}$-consistency of the estimated envelope is established in [11], which we review in the following.

Proposition 2.1. *Let $\widehat{\boldsymbol{\Gamma}}_u \in \mathbb{R}^{p\times u}$, $0 \leq u \leq p$, be the minimizer of $\mathrm{J}_n(\boldsymbol{\Gamma})$ in (2.2), where u is the envelope dimension of $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) \subseteq \mathbb{R}^p$. If $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ are $\sqrt{n}$-consistent in (2.2), then $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_u} = \widehat{\boldsymbol{\Gamma}}_u\widehat{\boldsymbol{\Gamma}}_u^{\mathsf{T}}$ is a $\sqrt{n}$-consistent estimate for the projection onto the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$.*

After obtaining $\widehat{\boldsymbol{\Gamma}}$ from minimizing the above objective function, the envelope estimator of $\boldsymbol{\theta}$ is set as $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^{\mathsf{T}}\widehat{\boldsymbol{\theta}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\theta}}$. Therefore, the envelope estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^{\mathsf{T}}\widehat{\boldsymbol{\theta}}$ is $\sqrt{n}$-consistent and can be much accurate than the standard estimator $\widehat{\boldsymbol{\theta}}$.

Different choices of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$ lead to different envelope methods in the literature. Table 1 summarizes some commonly used sample estimators $\{\widehat{\mathbf{M}}, \widehat{\mathbf{U}}\}$ for envelope regression. We use $\mathbf{S}_{\mathbf{A}}$ to denote the sample covariance matrix of a random vector $\mathbf{A}$ and use $\mathbf{S}_{\mathbf{A}|\mathbf{B}}$ to denote the sample conditional covariance of $\mathbf{A} \mid \mathbf{B}$. For the partial envelope method, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ is the important predictor. For the generalized linear model, $\mathbf{S}_{\mathbf{X}(W)}$ is the weighted sample covariance defined in Cook and Zhang [9], where more detailed discussion on the choices of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ can be found.

Table 1

Some commonly used sample estimators for envelope regression: response envelope [7], partial envelope [22] and predictor envelope [5] for linear models, and envelopes for generalized linear models [9].

	Response	Partial	Predictor	Generalized Linear Model
$\widehat{\mathbf{M}}$	$\mathbf{S}_{\mathbf{Y}\mid\mathbf{X}}$	$\mathbf{S}_{\mathbf{Y}\mid\mathbf{X}}$	$\mathbf{S}_{\mathbf{X}\mid\mathbf{Y}}$	$\mathbf{S}_{\mathbf{X}(W)}$ or $\mathbf{S}_{\mathbf{X}}$
$\widehat{\mathbf{M}}+\widehat{\mathbf{U}}$	$\mathbf{S}_{\mathbf{Y}}$	$\mathbf{S}_{\mathbf{Y}\mid\mathbf{X}_2}$	$\mathbf{S}_{\mathbf{X}}$	$\widehat{\mathbf{M}}+\widehat{\boldsymbol{\beta}}\mathbf{S}_{\mathbf{X}(W)}\widehat{\boldsymbol{\beta}}^{\mathsf{T}}$

When the envelope dimension u becomes large, especially when p is not small, the computation based on the full Grassmannian (FG) optimization of (2.2) can be expensive and requires good initial values to circumvent the issue with local minima. When selecting the envelope dimension, this computational issue is even worse: we need to conduct the optimization repeatedly for $k=1,\dots,p$ since the solutions is not nested as we increase k, that is, $\text{span}(\widehat{\boldsymbol{\Gamma}}_k)\not\subseteq\text{span}(\widehat{\boldsymbol{\Gamma}}_{k+1})$. Thus, in Section 3.3 we propose a computationally efficient alternative to the FG envelope dimension selection approach that is based on FG optimization of (2.2). Our new approach is based on the 1D algorithm proposed by Cook and Zhang [11] that breaks down the FG optimization of (2.2) to "one-direction-at-a-time". We review the population 1D algorithm in the following.

For $k=0,\dots,p-1$, let $\mathbf{g}_k\in\mathbb{R}^p$ denote the k-th sequential direction to be obtained. Let $\mathbf{G}_k=(\mathbf{g}_1,\dots,\mathbf{g}_k)$, and $(\mathbf{G}_k,\mathbf{G}_{0k})$ be an orthogonal basis for $\mathbb{R}^p$ and set initial value $\mathbf{g}_0=\mathbf{G}_{00}=0$. Define $\mathbf{M}_k=\mathbf{G}_{0k}^{\mathsf{T}}\mathbf{M}\mathbf{G}_{0k}$, $\mathbf{U}_k=\mathbf{G}_{0k}^{\mathsf{T}}\mathbf{U}\mathbf{G}_{0k}$, and the objective function after k sequential steps

$$\phi_k(\mathbf{w})=\log(\mathbf{w}^{\mathsf{T}}\mathbf{M}_k\mathbf{w})+\log\{\mathbf{w}^{\mathsf{T}}(\mathbf{M}_k+\mathbf{U}_k)^{-1}\mathbf{w}\},\tag{2.3}$$

which has to be minimized over $\mathbf{w}\in\mathbb{R}^{p-k}$ subject to $\mathbf{w}^{\mathsf{T}}\mathbf{w}=1$. The $(k+1)$-th envelope direction is $\mathbf{g}_{k+1}=\mathbf{G}_{0k}\mathbf{w}_{k+1}$, where $\widehat{\mathbf{w}}_{k+1}=\arg\min_{\mathbf{w}^{\mathsf{T}}\mathbf{w}=1}\phi_k(\mathbf{w})$. The 1D algorithm produces a nested solution path that contains the true envelope: $\text{span}(\mathbf{G}_1)\subset\cdots\subset\text{span}(\mathbf{G}_u)=\mathcal{E}_{\mathbf{M}}(\mathbf{U})\subset\text{span}(\mathbf{G}_{u+1})\subset\cdots\subset\text{span}(\mathbf{G}_p)=\mathbb{R}^p$. As we replace $\mathbf{M}$ and $\mathbf{U}$ in the above optimization with some $\sqrt{n}$-consistent $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$, we will obtain sequential $\sqrt{n}$-consistent estimates $\widehat{\mathbf{G}}_k=(\widehat{\mathbf{g}}_1,\dots,\widehat{\mathbf{g}}_k)\in\mathbb{R}^{p\times k}$, $k=1,\dots,p$. The $\sqrt{n}$-consistency of the 1D algorithm is established in [11], which we review in the following.

Proposition 2.2. *Under the same assumptions as Proposition 2.1, $\mathbf{P}_{\widehat{\mathbf{G}}_u}=\widehat{\mathbf{G}}_u\widehat{\mathbf{G}}_u^{\mathsf{T}}$ is a $\sqrt{n}$-consistent estimate for the projection onto the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$.*

This proposition suggests that the 1D algorithm share the same convergence rate as the FG optimization, while faster, stable and nested solutions are produced. Therefore, by Propositions 2.1 and 2.2, both the FG and 1D estimators are $\sqrt{n}$-consistent given the true envelope dimension and $\sqrt{n}$-consistent $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$.

In most applications, $\sqrt{n}$-consistent $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ are easy to obtain, but there lacks a theoretically justified method to choose the crucial hyperparameter u under the generality of the envelope methods. We assume $\sqrt{n}$-consistency of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ to their population counterparts $\mathbf{M}>0$ and $\mathbf{U}\geq 0$ throughout the exposition and focus on the selection of dimension u. For the matrices $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ in Table 1, their $\sqrt{n}$-consistency is easily justified: for response envelope and partial envelope, it suffices to assume the error term in the linear model is i.i.d. with finite fourth moments; and for the predictor envelope and generalized linear model envelope, it suffices to assume the predictor $\mathbf{X}$ is i.i.d. with finite fourth moments.

Very recently, Cook and Zhang [12] proposed an envelope coordinate descent (ECD) algorithm that is even faster than the 1D algorithm without loss of accuracy. In particular, the ECD algortihm and the 1D algorithm solve the same sequence of objective functions (2.3); and they share the same theoretical properties. Therefore, all the theoretical results presented in this paper are all true when we substitute the 1D algorithm for the ECD algorithm.

3. Envelope dimension selection

3.1. A new quasi-likelihood argument for model-free envelope estimation

The generic moment-based envelope estimation of $\boldsymbol{\theta}$ is essentially a two-stage estimator, where the first stage is estimating an envelope basis $\widehat{\boldsymbol{\Gamma}}$ from $\mathrm{J}_n(\boldsymbol{\Gamma})$ and the second stage is projecting the standard estimator onto the estimated envelope subspace: $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}^{\mathsf{T}}\widehat{\boldsymbol{\theta}}$ to eliminate immaterial variation. The objective function $\mathrm{J}_n(\boldsymbol{\Gamma})$ has previously been proposed and studied by Cook and Zhang [11] and Cook and Zhang [9] purely for estimating an envelope basis, but it is still difficult to understand the effect and implication of $\mathrm{J}_n(\boldsymbol{\Gamma})$ on $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}}$ and to study the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}}$.

We show that $\mathrm{J}_n(\boldsymbol{\Gamma})$ can be viewed as a quasi-likelihood function. Moreover, our results connect $\mathrm{J}_n(\boldsymbol{\Gamma})$ with the joint estimation of $\mathbf{M}$ and $\boldsymbol{\theta}$ that leads to both the standard and the envelope estimators. Define

$$\ell_n(\mathbf{M}, \boldsymbol{\theta}) = \log|\mathbf{M}| + \mathrm{trace}\left[\mathbf{M}^{-1}\left\{\widehat{\mathbf{M}} + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathsf{T}}\right\}\right]. \tag{3.1}$$

Then, given a working dimension $k = 0, \ldots, p$, that is not necessarily the true envelope dimension u, the envelope estimation is a constrained minimization of (3.1) over the following feasible parameter set,

$$\begin{aligned}\mathcal{A}_k = \{(\mathbf{M}, \boldsymbol{\theta}) : \mathbf{M} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\mathsf{T}} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^{\mathsf{T}} > 0,\ \boldsymbol{\theta} = \boldsymbol{\Gamma}\boldsymbol{\eta}, \boldsymbol{\eta} \in \mathrm{R}^{k\times 1},\\ \text{and } (\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)^{\mathsf{T}}(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) = \mathbf{I}_p\},\end{aligned} \tag{3.2}$$

where $\mathcal{A}_0$ is defined as $\mathcal{A}_0 = \{(\mathbf{M}, \boldsymbol{\theta}) : \mathbf{M} > 0, \boldsymbol{\theta} = 0\}$, and the standard estimator is achieved at $\mathcal{A}_p$.

Under the envelope parametrization of $\mathbf{M} = \mathbf{M}(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$ and $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\Gamma}, \boldsymbol{\eta})$ in (3.2), $\ell_n(\mathbf{M}, \boldsymbol{\theta})$ in (3.1) is now an over-parametrized objective function for the envelope estimation: $\ell_n(\mathbf{M}, \boldsymbol{\theta}) = \ell_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\eta})$. We show that this constrained optimization problem reproduces $\mathrm{J}_n(\boldsymbol{\Gamma})$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}}$ in Cook and Zhang [9].

Lemma 3.1. *The minimizer of* $\ell_n(\mathbf{M},\boldsymbol{\theta})$ *in* (3.1) *under the envelope parametrization in* (3.2) *is* $\widehat{\mathbf{M}}_{\mathrm{Env}} = \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}^{\mathsf{T}}\widehat{\mathbf{M}}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}^{\mathsf{T}} + \widehat{\mathbf{\Gamma}}_0\widehat{\mathbf{\Gamma}}_0^{\mathsf{T}}\widehat{\mathbf{M}}\widehat{\mathbf{\Gamma}}_0\widehat{\mathbf{\Gamma}}_0^{\mathsf{T}}$ *and* $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}^{\mathsf{T}}\widehat{\boldsymbol{\theta}}$, *where* $\widehat{\mathbf{\Gamma}}$ *is the minimizer of the partially optimized objective function* $\ell_n(\mathbf{\Gamma}) = \min_{\mathbf{\Omega},\mathbf{\Omega}_0,\boldsymbol{\eta}} \ell_n(\mathbf{\Gamma},\mathbf{\Omega},\mathbf{\Omega}_0,\boldsymbol{\eta}) = \mathrm{J}_n(\mathbf{\Gamma}) + \log|\widehat{\mathbf{M}} + \widehat{\mathbf{U}}| + p$ *for* $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^{\mathsf{T}}$.

Lemma 3.1 shows that, although $\mathrm{J}_n(\mathbf{\Gamma})$ is not an objective function for $\boldsymbol{\theta}$, it can be viewed as a partially minimized quasi-likelihood function $\ell_n(\mathbf{M},\boldsymbol{\theta})$ under the envelope parametrization, up to an additive constant difference. Our dimension selection method is based on this quasi-likelihood formulation that is completely generic and model-free. This new finding and formulation will largely facilitate our theoretical derivation of envelope dimension selection consistency in the next two sections.

3.2. Dimension selection based on the full Grassmannian optimization

We first discuss some properties about $\mathrm{J}_n(\mathbf{\Gamma})$ defined in (2.2) to motivate our dimension selection criterion. It can be shown that $\mathrm{J}_n(\mathbf{\Gamma})$ converges uniformly in probability to its population counterpart $\mathrm{J}(\mathbf{\Gamma}) = \log|\mathbf{\Gamma}^{\mathsf{T}}\mathbf{M}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^{\mathsf{T}}(\mathbf{M} + \mathbf{U})^{-1}\mathbf{\Gamma}|$. To distinguish estimators at different envelope working dimensions, let $\mathbf{\Gamma}_k$ and $\widehat{\mathbf{\Gamma}}_k \in \mathbb{R}^{p\times k}$ denote the minimizers of the population objective function $\mathrm{J}(\mathbf{\Gamma})$ and the sample objective function $\mathrm{J}_n(\mathbf{\Gamma})$ at dimension k. The objective functions $\mathrm{J}_n(\mathbf{\Gamma})$ and $\mathrm{J}(\mathbf{\Gamma})$ are well-defined only for envelope dimension $k = 1,\ldots,p$. But (3.1) and (3.2) are well-defined for $k = 0$. For $k = 0$, we can show that $\min_{\mathcal{A}_0} \ell_n(\mathbf{M},\boldsymbol{\theta}) = \log|\widehat{\mathbf{M}} + \widehat{\mathbf{U}}| + p$ is achieved at $\widehat{\mathbf{M}}_{\mathrm{Env},0} = \widehat{\mathbf{M}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{Env},0} = 0$. Therefore, we define $\mathrm{J}_n(\mathbf{\Gamma}_k) = \mathrm{J}(\mathbf{\Gamma}_k) = 0$ for $k = 0$. Consequently, we have the following results.

Lemma 3.2. *If* $u = 0$, *then* $\mathrm{J}(\mathbf{\Gamma}_k) = 0$ *for all* $k = 0,\ldots,p$. *If* $u > 0$, *then* $\mathrm{J}(\mathbf{\Gamma}_u) < \mathrm{J}(\mathbf{\Gamma}_k) < 0$, *for* $0 < k < u$, *and* $\mathrm{J}(\mathbf{\Gamma}_k) = \mathrm{J}(\mathbf{\Gamma}_u) < 0$, *for* $k \geq u$. *Moreover, for* $0 \leq u < k$, $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) \subset \mathrm{span}(\mathbf{\Gamma}_k)$.

Lemma 3.2 shows that, $\mathrm{J}(\mathbf{\Gamma}_k)$ is strictly greater than $\mathrm{J}(\mathbf{\Gamma}_u)$ when $k < u$, and remains constant once k exceeds u. We thus propose to select the envelope dimension via minimizing the following criterion,

$$\mathcal{I}_n(k) \equiv \mathrm{J}_n(\widehat{\mathbf{\Gamma}}_k) + \frac{C \cdot k \cdot \log(n)}{n}, \quad k = 0,1,\ldots,p, \tag{3.3}$$

where $C > 0$ is a constant and $\mathcal{I}_n(0) = 0$. Under envelope linear models [7, 22, 5, e.g.], this FG criterion is closely related to the likelihood-based BIC. More specifically, $\mathcal{I}_n(k)$ is exactly the BIC for envelope linear models if we let Ck matches the number of free parameters. For response envelope model [7], this means $C = p$ is the total number of predictors; and for partial envelope model [22], $C = p_1$ is the number of primary predictors $\mathbf{X}_1$; for predictor envelope model [5], $C = r$ is the number of response variables. We will discuss more about the choice of C later in Section 3.4. The envelope dimension is selected as $\widehat{u}_{\mathrm{FG}} =$

$\arg\min_{0\le k\le p}\mathcal{I}_n(k)$, where we use subscript FG to denote full Grassmannian optimization of $\mathrm{J}_n(\boldsymbol{\Gamma})$. The criterion (3.3) has a form similar to the Bayesian information criterion, but has the fundamental difference that $J_n(\widehat{\boldsymbol{\Gamma}}_k)$ is not a likelihood function. Properties of $\mathcal{I}_n$ are not easy to obtain, as the results for likelihood functions do not apply here. Nevertheless, we can show that (3.3) leads to consistent dimension selection without likelihood arguments.

Theorem 3.1. *For any constant $C > 0$ and $\sqrt{n}$-consistent $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ in (3.3), we have $\Pr(\widehat{u}_{\mathrm{FG}} = u) \to 1$ as $n \to \infty$.*

We have three remarks about the results in Theorem 3.1. First, Theorem 3.1 reveals that the choice of C does not affect the consistency of our proposed dimension selection procedure. We will discuss more on the role of this constant C in Section 3.4. Second, the consistency shown in Theorem 3.1 does not require any model assumptions. Therefore, (3.3) can be applied to any models with the envelope structure. Third, in the heavily-studied case of multivariate linear regression model, $\mathrm{J}_n(\boldsymbol{\Gamma})$ will reproduce the normal likelihood-based objective function if we plug in appropriate choices of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ [Section 1.3; 9]. In such cases, (3.3) will reproduce the Bayesian information criteria for multivariate linear envelope models, where the same criterion in (3.3) has been used without any justification but yielded good results. The following Corollary to Theorem 3.1 confirmed that the envelope dimension $\widehat{u}_{\mathrm{BIC}}$ selected from the Bayesian Information Criterion is indeed consistent.

Corollary 3.1. *Suppose that the sample covariance matrices $\mathbf{S}_{\mathbf{X}}$, $\mathbf{S}_{\mathbf{Y}}$ and $\mathbf{S}_{\mathbf{XY}}$ are $\sqrt{n}$-consistent, then for envelope linear models, we have $\Pr(\widehat{u}_{\mathrm{BIC}} = u) \to 1$ as $n \to \infty$.*

Corollary 3.1 reinforces the message that, although envelope estimates are typically constructed under some normality assumptions, normality is generally not essential for the application of envelope estimates. Previous studies of envelope linear models [7, 22, 5, 4] have shown that the envelope estimators obtained by maximizing the normality-based likelihood function are still $\sqrt{n}$-consistent and asymptotically normal even when the normality assumption is violated and the likelihood is mis-specified. Corollary 3.1 further showed that, even when the likelihood function is mis-specified due to non-normality, it can still help with selecting the dimension correctly. To the best of our knowledge, this is the first time in the literature that an envelope dimension selection criterion is justified without stringent likelihood assumptions. For the same reason, the modified BIC in Li and Zhang [17] is also able to select the tensor envelope dimension consistently since it's also a special case of our FG criterion.

3.3. Dimension selection based on the 1D estimation

As mentioned earlier in Section 2, the FG optimization can not guarantee nested envelope subspace, $\mathrm{span}(\widehat{\boldsymbol{\Gamma}}_k) \not\subseteq \mathrm{span}(\widehat{\boldsymbol{\Gamma}}_{k+1})$, while the 1D algorithm always produces a strictly nested solution path: $\mathrm{span}(\widehat{\mathbf{G}}_k) \subset \mathrm{span}(\widehat{\mathbf{G}}_{k+1})$. Therefore, it is an intuitive practice [10, 17, e.g.] to select envelope dimension based on BIC using the 1D envelope estimator. However, simply replacing $\widehat{\boldsymbol{\Gamma}}_k$ with the 1D estimator $\widehat{\mathbf{G}}_k$ in BIC, or the FG criterion in general (3.3), may not produce asymptotically consistent envelope dimension selection results since $\widehat{\mathbf{G}}_k$ is not a local optimizer of $\mathrm{J}_n(\boldsymbol{\Gamma})$. Therefore, when applying the 1D algorithm, we propose to select the envelope dimension via minimizing the following 1D criterion instead of the FG criterion,

$$\mathcal{I}_n^{\mathrm{1D}}(k) \equiv \sum_{j=1}^{k} \phi_{j,n}(\widehat{\mathbf{w}}_j) + \frac{C \cdot k \cdot \log(n)}{n}, \quad k = 0, 1, \ldots, p, \tag{3.4}$$

where $C > 0$ is a constant, $\mathcal{I}_n^{\mathrm{1D}}(0) = 0$, and the function $\phi_{j,n}(\mathbf{w})$ is the sample version of $\phi_j(\mathbf{w})$ defined in (2.3). We select the envelope dimension selected as $\widehat{u}_{\mathrm{1D}} = \arg\min_{0 \leq k \leq p} \mathcal{I}_n^{\mathrm{1D}}(k)$.

Theorem 3.2. *For any constant $C > 0$ and $\sqrt{n}$-consistent $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ in (3.4), $\Pr(\widehat{u}_{\mathrm{1D}} = u) \to 1$ as $n \to \infty$.*

We have two remarks about the 1D criterion $\mathcal{I}_n^{\mathrm{1D}}(k)$. First, it is easy to see that $\sum_{j=1}^{k} \phi_{j,n}(\widehat{\mathbf{w}}_j)$ serves as the same role as $\mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_k)$ in the full Grassmannian optimization criterion $\mathcal{I}_n(k)$ in (3.3). But the change of criterion here is critical as we have a different optimization problem. In fact, simply replacing $\widehat{\boldsymbol{\Gamma}}_k$ in the FG criterion (3.3) with the 1D solution $\widehat{\mathbf{G}}_k$ will not guarantee consistency in the selection. This is due to the fact that $\widehat{\mathbf{G}}_k$, although a $\sqrt{n}$-consistent envelope basis estimator, is not a local minima of the full Grassmannian objective function $\mathrm{J}_n(\boldsymbol{\Gamma})$. Instead, using $\sum_{j=1}^{k} \phi_{j,n}(\widehat{\mathbf{w}}_j)$ is indeed necessary for envelope dimension selection based on the 1D algorithm. Secondly, the computational cost of obtaining $\mathcal{I}_n^{\mathrm{1D}}(k)$, $k = 1, \ldots, p$, is much less than that of $\mathcal{I}_n(k)$, $k = 1, \ldots, p$. This is not only because the 1D algorithm is much faster and more stable than the FG optimization, but also due to the sequential nature of the 1D algorithm. For the 1D algorithm, we only need to run it once to estimate the $(p-1)$-dimensional envelope $\widehat{\mathbf{G}}_{p-1}$ to obtain all the values of $\mathcal{I}_n^{\mathrm{1D}}(k)$, $k = 1, \ldots, p$. For the full Grassmannian approach, it requires estimation of each envelope basis $\widehat{\boldsymbol{\Gamma}}_1, \ldots, \widehat{\boldsymbol{\Gamma}}_{p-1}$ separately and the computation for $\widehat{\boldsymbol{\Gamma}}_{p-1}$ alone can be more costly than obtaining $\widehat{\mathbf{G}}_{p-1}$. Therefore, in practice, we would strongly recommend using the 1D approach instead of the FG approach, when p is large. Simulation studies in the next section also show that the 1D approach is much more accurate and effective than the FG approach.

3.4. Role of C

Our proposed model-free criteria (3.3) and (3.4) are motivated from the BIC, and as we mentioned earlier in Corollary 3.1, the FG criterion (3.3) indeed includes the BIC for envelope linear models as a special choice. Specifically, the first term $\mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_k)$ in (3.3) will be the $(2/n)$ times the negative normal log-likelihood with appropriate choices of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$, whereas the second term $Ck\log(n)/n$ corresponds to the penalty term on the number of parameters in the linear envelope models. In the 1D criterion, the first term has no likelihood interpretation but is analogous to the first term in the FG criterion, thus the same penalty on the number of parameters were used. Because of these connections and connections with BIC, we suggest to use $C = 1$ for both the FG and the 1D criteria when the parameter $\boldsymbol{\theta}$ is naturally a vector. In other situations, where $\boldsymbol{\theta}$ is naturally a matrix-valued or even tensor-valued parameters, we would try to match Ck with the number of parameters in the model.

Although we focused on a vector-valued parameter $\boldsymbol{\theta} \in \mathbb{R}^p$ in the quasi-likelihood argument of $\ell_n(\mathbf{M}, \boldsymbol{\theta})$, the theoretical results in Lemma 3.1 and Theorems 3.1 & 3.2 do not impose any restriction on $\boldsymbol{\theta}$ being a vector. Our proofs are in fact written for a matrix-valued $\boldsymbol{\theta} \in \mathbb{R}^{p\times q}$ and can be straightforwardly extended to tensor-valued $\boldsymbol{\theta}$. In such cases, matching Ck with the number of parameters would give $C = q$ for $\boldsymbol{\theta} \in \mathbb{R}^{p\times q}$ when we are enveloping the column space of $\boldsymbol{\theta}$. Also, the term $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\intercal}$ in $\ell_n(\mathbf{M}, \boldsymbol{\theta})$ may also be replaced by a weighted version $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\mathbf{W}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\intercal}$ for some $\mathbf{W} \in \mathbb{R}^{q\times q}$ to tie more closely to the likelihood function for potentially improved efficiency. See Cook and Zhang [9] (Definition 4 and Proposition 7 in the Supplement) for a detailed discussion on enveloping a matrix-valued parameter and choices of $\mathbf{W} > 0$.

The proposed envelope dimension selection approaches in this paper are as flexible as possible, since we only require $\sqrt{n}$-consistent $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ for the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ without additional assumptions on distributions of variables or specific models. The theoretical developments, i.e. Theorems 3.1 and 3.2, only require C to be a positive constant to guarantee asymptotically correct selection of the envelope dimension with probability one. However, in finite sample, the selection of envelope dimension may be affected by the choice of C. It is hard to describe qualitatively the effect C on the dimension selection, because that depends on many factors such as the signal-to-noise ratio of the data, the sample size, the total number of parameters, the efficiency and the variance of the $\sqrt{n}$-consistent estimators $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$, etc. Nonetheless, from the proposed criteria (3.3) and (3.4), we know that smaller C leads to a more conservative choice of the envelope dimension, potentially overestimation ($\widehat{u} > u$), and larger C leads to a more aggressive choice and potentially underestimation ($\widehat{u} < u$). From our experience, the number C should be set to its default value $C = 1$ when there is no additional model assumption or prior information. When we know additional model assumption or prior information, C should be set such that Ck best matches the degree-of-freedom or total number of free parameters of the model or estimation procedure. For example,if the envelope is enveloping a vector-valued parameter, e.g. linear or generalized linear regression with univariate response or predictor, then let $C = 1$; if the envelope is enveloping a matrix or tensor valued parameter, then usually the best result comes from $C > 1$, where Ck should be obtained from calculating the total number of free parameters, which relate to the dimension of the matrix/tensor as well as the true rank of the parameter matrix/tensor [4, e.g.]. In the next section, we use

TABLE 2
Frequencies of selected dimensions for a generic $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ with $p = 20$ and $u = 5$.

	1D selection			FG selection					
	(I)	(II)	(III)	(I)		(II)		(III)	
n		$\widehat{u}_{1\mathrm{D}} = 5$		$\widehat{u}_{\mathrm{FG}} = 5$	$\widehat{u}_{\mathrm{FG}} = 6$	$\widehat{u}_{\mathrm{FG}} = 5$	$\widehat{u}_{\mathrm{FG}} = 6$	$\widehat{u}_{\mathrm{FG}} = 5$	$\widehat{u}_{\mathrm{FG}} = 6$
150	98	45	67	59.5	24	66	12.5	28.5	44.5
200	99	75.5	94	66	23	77.5	15.5	39.5	47
250	100	95	97	70	24	82.5	14.5	33	48
300	100	99.5	99	67.5	24	85	13.5	40.5	47
400	100	100	100	75.5	18.5	84.5	14	49	41
800	100	100	100	84.5	13.5	91	8.5	56	39.5

$C = 1$ for generic envelopes, where we have no information about the model, and also use $C = 1$ for envelope models (simulation Section 4.2 and real data Section 4.5) where the parameter of interest is a vector; then in Section 4.3 we study the effect of C for a matrix valued parameter. The numerical results further support our opinion in the above.

4. Numerical studies

4.1. Generic envelopes

In this section, we present numerical studies of dimension selection for a generic envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) = \mathrm{span}(\boldsymbol{\Gamma})$, where $\mathbf{M} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\intercal} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^{\intercal}$ and $\mathbf{U} = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^{\intercal}$ follow the envelope structure. In this section, we use $p = 20$ and $u = 5$. The envelope basis matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$ is a randomly generated semi-orthogonal matrix and then $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p\times(p-u)}$ is the orthogonal completion of $\boldsymbol{\Gamma}$ such that $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ is orthogonal. For $\boldsymbol{\Phi}$, we generate $\mathbf{A} \in \mathbb{R}^{u\times u}$ with each element a_{ij} sampled from the uniform distribution over [0,1]. Then we set $\boldsymbol{\Phi} = \mathbf{A}\mathbf{A}^{\intercal}$. We considered the following three different models for the symmetric positive definite matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. Model (I): both $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are randomly generated independently in the same way as $\boldsymbol{\Phi}$. Model (II): $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are each generated as $\mathbf{O}\mathbf{D}\mathbf{O}^{\intercal}$ with $\mathbf{O}$ being an orthogonal matrix and $\mathbf{D}$ being a diagonal matrix of positive elements on its diagonal. We set the diagonal elements in $\mathbf{D}$ for $\boldsymbol{\Omega}$ as $1, \dots, u$, and the diagonal elements in $\mathbf{D}$ for $\boldsymbol{\Omega}_0$ as $\exp(-4),\ \exp(-3.5), \dots, \exp(3)$. Model (III): all parameters are the same as Model (II) except that $\boldsymbol{\Omega}_0$ is now $0.1\mathbf{I}_{p-u}$.

We simulated 200 pairs of sample matrices from Wishart distributions, $\widehat{\mathbf{M}} \sim W_p(\mathbf{M}/n,\ n)$ and $\widehat{\mathbf{U}} \sim W_p(\mathbf{U}/n,\ n)$ so that they are $\sqrt{n}$-consistent for their population counterparts. We vary the sample size n from 150 to 800. In Table 2, we report the percentages of selecting the envelope dimension correctly by the two proposed approaches. In all three models, the 1D criterion is very effective and provides consistent selection of u: the percentage of correctly selecting the envelope dimension is monotonically approaching 1 as the sample size increases. The FG criterion is less competitive but still gives reasonable results especially because the total number of free parameters in $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Phi}$ is $p(p+1)/2 + u(u+1)/2 = 225$ which is not a small number comparing to n. For

TABLE 3

Selection and estimation results for three different envelope models. Left panel includes percentages of correct selection. Right panel includes means and standard errors of $\lVert\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert_F$ for the standard estimator and the envelope estimators with either true or estimated dimensions.

		Correct Selection %			Estimation Error $\lVert\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert_F$					
					Standard	Envelope				
Model	n	1D	FG	CV		true u	1D	FG	CV	S.E.≤
Linear	150	93	81	65.5	0.49	0.31	0.33	0.33	0.36	0.015
	300	99	92	62.5	0.32	0.19	0.19	0.20	0.27	0.008
	600	99	92.5	68	0.23	0.13	0.14	0.14	0.18	0.007
Logistic	150	72	77.5	24.5	2.16	0.56	0.67	0.60	1.41	0.072
	300	92	89.5	39	1.40	0.34	0.35	0.34	0.82	0.042
	600	98	94	27	0.98	0.22	0.22	0.24	0.56	0.030
Cox	150	58	54	NA	1.33	1.24	1.22	1.23	NA	0.022
	300	83	75.5	NA	0.98	0.90	0.89	0.90	NA	0.013
	600	100	93	NA	0.79	0.72	0.72	0.72	NA	0.008

the FG approach, we also reported the percentage of $\widehat{u}_{\mathrm{FG}} = 6$ in Table 2, which demonstrates a clear tendency for the FG approach to over-estimate the envelope dimension. From Lemma 3.2, the over-estimated envelope dimension will result in a larger subspace that contains the true envelope. Thus over-estimating u eventually still leads to consistent and unbiased envelope estimator for $\boldsymbol{\theta}$ and cause much less harm than under-estimating u.

4.2. Envelope models

In this section, we simulate three different envelope models where the envelope dimension is $u = 2$ for $p = 10$. The three models are: the multivariate linear model (2.1), the logistic regression model and the Cox proportional hazard model. For the linear regression in (2.1), we generated X_i and ϵ_i independently from $N(0,1)$ and $N_p(0,\boldsymbol{\Sigma})$, where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\eta} = (1,1)^{\intercal}$, and $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\intercal} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^{\intercal}$. The covariance $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are each generated as $\mathbf{O}\mathbf{D}\mathbf{O}^{\intercal}$ similar to Model (II) in Section 4.1. We set the eigenvalues as $1, 5$ in $\boldsymbol{\Omega}$, and $\exp(-4), \exp(-3), \ldots, \exp(3)$ in $\boldsymbol{\Omega}_0$. For the logistic regression: $Y_i \sim \mathrm{Bernoulli}(\mathrm{logit}(\boldsymbol{\beta}^{\intercal}\mathbf{X}_i))$, we simulate $\mathbf{X}_i$ from $N_p(0,\boldsymbol{\Sigma}_{\mathbf{X}})$ where the parameters $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\beta}$ are the same as $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ in the linear model. For the Cox model, we follow the simulation model in Cook and Zhang [9] and let the survival time follow a Weibull distribution with scale parameter $\exp(\boldsymbol{\beta}^{\intercal}\mathbf{X}/5)$ and shape parameter 5, which gives hazard rate $h(Y \mid \mathbf{X}) = 5Y^4 \cdot \exp(\boldsymbol{\beta}^{\intercal}\mathbf{X})$. The censoring variable δ_i is generated from Bernoulli(0.5) distributions, which gives censoring rates of approximately 50%. Then the data $(Y_i, \delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, are used to fit the envelope Cox model, where Y_i is the failure time, $\delta_i = 0$ or 1 indicating whether the failure time is censored or observed, $\mathbf{X}_i$ is the predictor vector. Data generation for $\mathbf{X}_i$ is similar to the logistic regression set-up, except for $\boldsymbol{\Omega}_0 = 0.1\mathbf{I}_8$ and $\boldsymbol{\eta} = (0.2, 0.2)^{\intercal}$.

For each of the above models, we consider sample size $n = 150$, 300 and 600 and generated 200 data sets for each of the sample sizes. Table 3 summarizes the

Table 4

Multivariate linear regression, response envelope model with multivariate response of dimension $p = 10$ and envelope dimension $u = 2$. The parameter of interest is the 10×3 regression coefficient matrix, where $q = 3$ is the number of predictors, and hence the best choice of C should be $C = q = 3$. The left panel summarizes percentages of correct selection; and the right panel summarizes the average of selected dimension.

	Correct Selection %						Average selected $\widehat{u}$					
	1D			FG			1D			FG		
n	150	300	600	150	300	600	150	300	600	150	300	600
$C = 1$	38	59	63	23	42	52	2.95	2.53	2.46	3.51	2.90	2.62
$C = 3$	92	100	100	92	100	100	1.94	2.00	2.00	2.00	2.00	2.00
$C = 5$	66	100	100	86	100	100	1.66	2.00	2.00	1.86	2.00	2.00
$C = 10$	5	55	100	19	95	100	1.05	1.55	2.00	1.19	1.95	2.00

percentages of correctly selected envelope dimension and the estimation error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F$ in each of the simulations, where we compare the standard estimators (e.g. least squares, likelihood and partial likelihood estimators) to the envelope estimators using the true dimension or using the selected dimensions. In addition to the proposed 1D and FG selection procedures, we also compare with selecting envelope dimension based on prediction errors from five-fold cross-validation for Linear and Logistic regression models. Regarding dimension selection accuracy, both 1D and FG procedures have produced satisfactory results and are substantially more accurate than the cross-validation procedure. Moreover, the accuracy of both 1D and FG selections improves quickly with increasing sample size, while the improvement in cross-validation with increasing sample size is not guaranteed. For our proposed method, the percentage of correct selection is low only for the Cox model at sample size $n = 150$, which is a small number considering the 50% censoring rate. The cross-validation procedure is not directly applicable (at least not trivially) for Cox model. For all scenarios, there is no significant difference among envelope estimators with the true or the estimated dimension by 1D and FG procedures, which are all significantly better than the standard estimator. However the cross-validation tends to over-select the envelope dimension, which leads to less efficient estimation.

4.3. Matrix-valued parameter

As an illustration of the effect of C, we simulated data from the multivariate linear regression model (2.1), where we considered the response envelope model with multivariate response of dimension $p = 10$ and envelope dimension $u = 2$. The parameter of interest is the 10×3 regression coefficient matrix, where $q = 3$ is the number of predictors, and hence from our discussion in Section 3.4 we expect the best choice of C to be $C = q = 3$. The parameters and data are generated in same way as the single predictor linear model in Section 4.2, where we still set elements in $\boldsymbol{\eta} \in \mathbb{R}^{u \times q}$ as all ones to get $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$. From Table 4, we have the following observations: (1) for all values of C, the percentage of correct selection goes toward 1 when the sample size goes to infinity (even for $C = 1$, the percentage goes slowly but steadily to 90% as we keep increase the sample size to 6000); this also numerically verifies our theoretical results; (2) the "best" choice is apparently $C = 3$ because this lets the penalty Ck in the criteria matches the number of parameters in the model; (3) from the average value of selected $\widehat{u}$, we see that $C > 3$ leads to underestimation of u when the sample size is small and $C < 3$ leads to overestimation; (4) $C = 1$ with the 1D criterion is a "robust" choice, even for the small sample size $n = 150$ the averaged selection is 2.95.

We make two additional remarks on the performance of $C = 1$. On one hand, the average dimension is only slightly larger than the true dimension even for small sample size. In the situations when $C = 1$ is not the optimal choice, the 1D criterion with $C = 1$ may overestimate the dimension by a small amount. On the other hand, overestimation of the dimension slightly is much less of an issue comparing to the issue of underestimating the envelope dimension. If we apply envelope methods with a slightly larger structural dimension, estimation of the parameter is still unbiased. The slightly larger structural dimension will only lead to some efficiency loss. Meanwhile, if the dimension is underestimated, the envelope estimator will be biased and important directions will be missed. Fortunately, underestimation is not likely to happen, according to the simulation results in Table 4.

4.4. Scale invariance

In this section, we investigate the scale invariance property of envelopes. By Definition 2.2, it is easy to see that the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) = \mathcal{E}_{a\mathbf{M}}(b\mathbf{U})$ for any constants $a > 0$ and $b > 0$. Let $\lambda = b/a > 0$ reflects the potential scale changes in $\mathbf{M}$ and $\mathbf{U}$. In a model-free context, the population objective function (1.1) is equivalent to $\mathrm{J}_{\lambda}(\boldsymbol{\Gamma}) = \log|\boldsymbol{\Gamma}^{\mathsf{T}}\mathbf{M}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^{\mathsf{T}}(\mathbf{M} + \lambda\mathbf{U})\boldsymbol{\Gamma}|$, which is minimized by $\mathrm{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\mathbf{M}}(\mathbf{U})$; however, the corresponding sample objective function $\mathrm{J}_{n,\lambda}(\boldsymbol{\Gamma})$ will produce different $\sqrt{n}$-consistent estimators for the envelope. This could potentially lead to different dimension selection results. Nonetheless, our theorems still applies and the 1D and FG procedures are still consistent.

To empirically study the effect of λ on envelope estimation and thus on dimension selection, we use the same response envelope model from Section 4.2. Table 5 summarizes the results for different choices of $\lambda = \{0.01, 0.1, 1, 10, 100\}$ for replacing $\widehat{\mathbf{U}}$ with $\lambda\widehat{\mathbf{U}}$. In our experience, λ introduces the trade-off between information about the envelope from $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$. Heuristically speaking, if $\widehat{\mathbf{U}}$ is estimated more efficiently than $\widehat{\mathbf{M}}$, then we probably can improve the envelope estimation by introducing a $\lambda > 1$. However in most context, especially in envelope regression, the optimal choice seems to be $\lambda = 1$, which reproduces the likelihood-based objective functions in the literature. It is possible that some choices of $\lambda \neq 1$ may produce better finite-sample results, but it is often appropriate to choose $\lambda = 1$ to keep $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ in the same unit. For example, in the response envelope model (cf. Table 1), the natural choice of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ corresponds to the sample covariance matrices of residual and fitted value, respectively. Thus it is natural to keep them in the same scale and unit.

Table 5

Percentage of correct dimension selection in the response envelope model. The original scale is $\lambda = 1$, which corresponds to the likelihood-based objective function.

λ		0.01	0.1	1	10	100
$n = 150$	1D	0	1	93	49.5	55.5
	FG	0	1	81	33.5	11.5
$n = 600$	1D	0	57.5	99	57	54
	FG	0	57.5	92.5	55	19.5

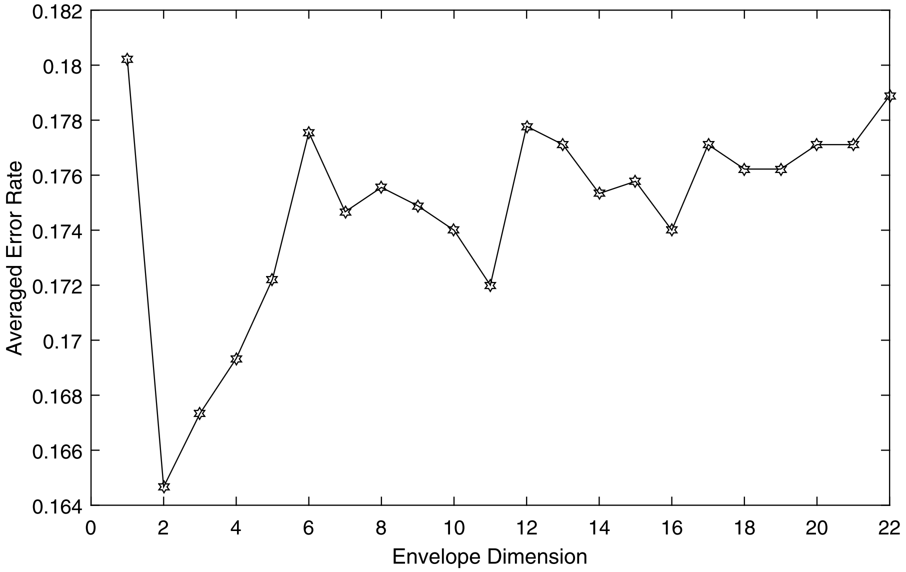

Fig 1. *Colon cancer tissue data: averaged mis-classification error rates for moment-based envelope estimators with various dimension based on 100 random data splitting. The standard logistic regression is the rightmost point, where $u = p = 22$.*

4.5. Real data illustration

For a real data illustration, we revisit the data set for envelope logistic regression in Cook and Zhang [9]. The data is from a colonoscopy study where 105 adenomatous (precancerous) tissue samples and 180 normal tissue samples were illuminated with ultraviolet light so that they fluoresce at difference wavelengths. The purpose of the study is to classify the total $n = 285$ tissue samples into the two classes, i.e. $Y = 1$ (adenomatous) and $Y = 0$ (normal), using the $p = 22$ predictors that are from laser-induced fluorescence spectra measured as 8nm spacing from 375 to 550 nm. More details of such colonoscopy study and a similar data set can be found in Hawkins and Maboudou-Tchao [16].

For this data set, we study the moment-based envelope estimators based on the 1D algorithm as an alternative to the likelihood-based approach demonstrated in Cook and Zhang [9]. Using the model-free dimension selection criterion developed in this paper, envelope dimension $u = 2$ is selected by both the FG approach (3.3) and the 1D approach (3.4) developed in this paper. We then randomly split the data into 80% training samples (228 samples) and 20% testing samples (57 samples) repeatedly for 100 times and fit the 1D moment-based envelope estimator for various dimensions and evaluate its classification power on the testing data set. As a result, the averaged mis-classification error rate is 0.1647 with standard error 0.0051 for the envelope estimator with selected dimension $u = 2$, much better than 0.1802 with standard error 0.0047 from fitting with $u = 1$ (if we underestimate the envelope dimension), and also much better than 0.1789 with standard error 0.0054 from the standard logistic regression. Figure 1 further summarizes the averaged error rate for envelope estimators with various dimensions from $u = 1$ to $u = 22$. Clearly, $u = 2$ is the desirable envelope dimension for this data set that is selected by our model-free criteria.

On the other hand, if we assume the predictor is normal then the envelope MLE, given the envelope dimension u, can be obtained use the iterative algorithm [9, Algorithm 1]. Standard BIC approach for selecting u is then applicable based on the full likelihood of $(Y, \mathbf{X})$. As a result, $u = 1$ is selected. However, envelope MLE with $u = 1$ will give bad classification result and Cook and Zhang [9] also used $u = 2$ for their envelope MLE, where the dimension is selected based on five-fold cross-validation. To further compare $u = 1$ versus $u = 2$, we considered the likelihood-ratio statistics for comparing the corresponding two envelope estimators with the standard estimator (i.e. the full model with $u = p$). For $u = 2$, the likelihood-ratio test gives a p-value of 0.104, which suggest that the envelope estimator and the standard estimator are consistent with each other although the envelope estimator has much improved prediction accuracy. Other other hand, for $u = 1$, the likelihood-ratio test gives a p-value of 1.48×10^{-6}, which is clear evidence against the envelope model with $u = 1$. Clearly, likelihood-ratio, cross-validation, and our proposed 1D and FG selection procedures all agree on $u = 2$, while standard BIC fails and selects $u = 1$. In addition, our 1D selection procedure is more than 7 times faster than the five-fold cross-validation based on 1D algorithm.

For this data set, the most likely reason for the standard BIC to fail to select a “reasonable” envelope dimension is probably due to the non-normality in the predictors. While cross-validation is computationally more expensive and has no theoretical justification, our proposed 1D and FG selection approaches can relax the normality assumption and select the asymptotically consistent and practically useful envelope dimension.

5. Discussion

In this paper, we proposed two envelope dimension selection procedures, based on the most general envelope definition from [9] and the two generic envelope estimation algorithms (FG and 1D algorithms) from [11]. The two dimension selection procedures are widely applicable, easy-to-implement, theoretically jus-

TABLE 6

The values of $\phi_k(\mathbf{w}_k)$, $k = 1,\ldots,p$, when replace $\mathbf{U}$ by $\lambda\mathbf{U}$ in the original response envelope model as in Table 3. The envelope dimension is $u = 2$.

λ	0.01	0.1	1	10	100
$\phi_1(\mathbf{w}_1)$	-0.01	-0.13	-0.83	-2.33	-4.53
$\phi_2(\mathbf{w}_2)$	-0.001	-0.01	-0.09	-0.08	-0.003
$\phi_k(\mathbf{w}_k)$, $k \geq 3$	0	0	0	0	0

tified, and have very encouraging performances in our numerical studies. Compared to k-fold cross-validation, the computational complexity of our procedure is reduced for at least k times. More importantly, the accuracy of the proposed method for selecting the true envelope dimension is shown to be much better than the cross-validation's result. Also, in some cases where cross-validation is not directly applicable, our procedure is still straightforward to implement.

This paper addresses the dimension selection problem and unified theory is provided (Theorems 3.1 and 3.2). In practice, the envelope dimension selection depends on the accuracy of envelope estimation. This motivated us to propose the 1D procedure that is computationally more feasible and is shown to be more accurate than FG procedure in selecting envelope dimension. For future research, we believe it is possible to adjust our criteria according to more efficient envelope estimation, possibly in high-dimensional settings.

In studying the distance between subspaces spanned by sample and population eigenvectors, gaps between the eigenvectors of interests and the rest eigenvectors is often a key ingredient of the theoretical results [23, e.g.]. Although the envelope is indeed a subspace spanned by eigenvectors of $\mathbf{M}$ that intersects with $\mathbf{U}$, the sample estimator of envelope is not obtained via sample eigenvectors $\widehat{\mathbf{M}}$. This makes it difficult to bound the distance between sample and population envelopes. The distance, $D(\widehat{\mathbf{\Gamma}},\mathbf{\Gamma}) \equiv \|\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}^{\top} - \mathbf{\Gamma}\mathbf{\Gamma}^{\top}\|_F$, is shown to be $O_p(n^{-1/2})$ (Propositions 2.1 and 2.2). But it will be difficult to further improve on this rate. However, our 1D criterion motivates us to view the minima of the 1D algorithm objective functions (2.3), $\phi_k(\mathbf{w}_k)$, $k = 1,\ldots,p$, as an analogy to the eigenvalues in eigen-problems. The eigenvalue gap in eigenspace estimation is analogous to $\phi_u(\mathbf{w}_u)$ in envelope estimation because $\phi_k(\mathbf{w}_k) < 0$ for $k \leq u$ and equals 0 for $k > u$. Intuitively, the bigger the gap $\phi_u(\mathbf{w}_u)$, the easier to select envelope dimension in practice. In Table 6, we report the values of $\phi_k(\mathbf{w}_k)$, $k = 1,\ldots,p$, when replace $\mathbf{U}$ by $\lambda\mathbf{U}$ in the original response envelope model as in Table 3. Recall from the finite sample results in Table 5, the best case scenario (i.e. highest percentage in selecting dimension correctly) is $\lambda = 1$ followed by $\lambda = 10$. The gap $\phi_2(\mathbf{w}_2)$ $(u = 2)$ in Table 6 confirms our conjecture that this gap is a good indicator for envelope estimation and dimension selection accuracy. Unfortunately, without prior knowledge of the envelope dimension u, it is impossible to estimate this quantity. Further theoretical analysis on this gap $\phi_u(\mathbf{w}_u)$ is beyond the scope of this paper and left as a future research topic.

It is also worth mentioning that there have been many methods for determining the dimension of a sufficient dimension reduction subspace [24, 29, 18, 26, 6, 19, 28, 27, for example], but the envelope dimension selection problem is very different and arguably more difficult in two aspects. First, sufficient dimension reduction methods are restricted to regression problems, whereas envelope methods can be applied to any multivariate parameter estimation. Our work provides a unified approach to select the structure dimension of envelopes under its full generality. Secondly, many sufficient dimension reduction methods can be formulated as a generalized eigenvalue problem where the dimension of interest is the rank of some kernel matrix. For envelopes, this is not so straightforward, as the envelopes are usually estimated from Grassmannian optimization where no analytic solution can be derived. This is also a part of the reason why we need two different criteria for different envelope optimizing procedures. BIC-type criteria have already been used extensively, with proven selection consistency, in the dimension determination problems for sufficient dimension reduction [26, 27]. While the log-likelihood term in those BIC-type criteria can usually be expressed explicitly as a function of eigenvalues (e.g. equation (10) in Zhu et al. [26]), or modified as the ratio of sums of squared eigenvalues (equation (6.1) in Zhu et al. [27]), the envelope objective function can not be further simplified to derive its asymptotic properties. Hence, studies on the envelop methods such as in this paper requires much more efforts in the technical proofs. The key technical trick is our new quasi-likelihood formulation in Section 3.1 which is useful for future studies on model-free envelopes.

Appendix A: Some useful preparation

Proof for Corollary 3.1 is omitted as it is straightforward from Theorem 3.1. The remaining proofs are provided in this Appendix. We will need to apply the following Proposition A.1 and Lemma A.1, which are obtained from Cook and Zhang [11, Propositions 2, 3, 5 and 6] and Cook et al. [5, Lemmas 6.2 and 6.3], for our proofs.

Proposition A.1. *If* $k = u$, *then* $\mathrm{span}(\boldsymbol{\Gamma}_k) = \mathrm{span}(\mathbf{G}_k) = \mathcal{E}_{\mathbf{M}}(\mathbf{U})$*; if, in addition,* $\widehat{\mathbf{M}}$ *and* $\widehat{\mathbf{U}}$ *are both* $\sqrt{n}$*-consistent, then* $\widehat{\boldsymbol{\Gamma}}_k\widehat{\boldsymbol{\Gamma}}_k^{\mathsf{T}}$ *and* $\widehat{\mathbf{G}}_k\widehat{\mathbf{G}}_k^{\mathsf{T}}$ *are both* $\sqrt{n}$*-consistent for the projection onto* $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$*.*

Lemma A.1. *Suppose that* $\mathbf{M} > 0$ *is a* $p \times p$ *symmetric matrix* $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ *is an orthogonal basis matrix for* $\mathbb{R}^p$*, then* $\log|\mathbf{M}| = \log|\boldsymbol{\Gamma}_0^{\mathsf{T}}\mathbf{M}\boldsymbol{\Gamma}_0| - \log|\boldsymbol{\Gamma}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{\Gamma}| \leq \log|\boldsymbol{\Gamma}_0^{\mathsf{T}}\mathbf{M}\boldsymbol{\Gamma}_0| + \log|\boldsymbol{\Gamma}^{\mathsf{T}}\mathbf{M}\boldsymbol{\Gamma}|$*, where the second equality holds if and only if* $\mathrm{span}(\boldsymbol{\Gamma})$ *is a reducing subspace of* $\mathbf{M}$*.*

Appendix B: Proof for Lemma 3.1

Proof. First, we substitute $\mathbf{M} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\mathsf{T}} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^{\mathsf{T}}$ and $\boldsymbol{\theta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ into $\ell_n(\mathbf{M}, \boldsymbol{\theta})$ and expand it explicitly as

$$\begin{aligned}
\ell_n(\mathbf{M}, \boldsymbol{\theta}) \quad &= \quad \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0| \\
&+ \quad \mathrm{trace}\left[(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^{\mathsf{T}} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1} \cdot \boldsymbol{\Gamma}_0^{\mathsf{T}}) \cdot \left\{\widehat{\mathbf{M}} + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\Gamma}\boldsymbol{\eta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\Gamma}\boldsymbol{\eta})^{\mathsf{T}}\right\}\right] \\
&\equiv \quad \ell_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\eta})
\end{aligned}$$

where the first part is from $\log|\mathbf{M}| = \log|\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\top} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^{\top}| = \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0|$. We next show that $\mathrm{J}_n(\boldsymbol{\Gamma})$ is obtained by partially minimizing $\ell_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\eta})$ over $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. Taking derivative of $\ell_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$ and set it equaling zero, we have

$$0 = (\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^{\top} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\cdot\boldsymbol{\Gamma}_0^{\top})\cdot(2\boldsymbol{\eta} - 2\boldsymbol{\Gamma}^{\top}\widehat{\boldsymbol{\theta}}),$$

which leads to the minimizer $\widehat{\boldsymbol{\eta}}(\boldsymbol{\Gamma}) = \boldsymbol{\Gamma}^{\top}\widehat{\boldsymbol{\theta}}$. As a result, $\widehat{\boldsymbol{\theta}}(\boldsymbol{\Gamma}) = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\top}\widehat{\boldsymbol{\theta}} = \mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\theta}}$. Furthermore, the partially minimized ℓ_n is now

$$\begin{aligned}
\ell_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0) &= \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0| \\
&+ \operatorname{trace}\Big[(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^{\top} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\cdot\boldsymbol{\Gamma}_0^{\top}) \\
&\cdot \Big\{\widehat{\mathbf{M}} + (\widehat{\boldsymbol{\theta}} - \mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\theta}})^{\top}\Big\}\Big] \\
&= \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0| \\
&+ \operatorname{trace}\Big[(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^{\top} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\cdot\boldsymbol{\Gamma}_0^{\top})\cdot\Big\{\widehat{\mathbf{M}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^{\top}\mathbf{Q}_{\boldsymbol{\Gamma}}\Big\}\Big] \\
&= \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0| + \operatorname{trace}(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^{\top}\cdot\widehat{\mathbf{M}}) \\
&+ \operatorname{trace}\Big\{\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^{\top}\cdot(\widehat{\mathbf{M}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^{\top}\mathbf{Q}_{\boldsymbol{\Gamma}})\Big\} \\
&= \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0| + \operatorname{trace}(\boldsymbol{\Omega}^{-1}\cdot\boldsymbol{\Gamma}^{\top}\widehat{\mathbf{M}}\boldsymbol{\Gamma}) \\
&+ \operatorname{trace}\Big\{\boldsymbol{\Omega}_0^{-1}\cdot\boldsymbol{\Gamma}_0^{\top}(\widehat{\mathbf{M}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^{\top}\mathbf{Q}_{\boldsymbol{\Gamma}})\boldsymbol{\Gamma}_0\Big\}.
\end{aligned}$$

It is a well-known fact (from normal likelihood) that $\mathbf{S} = \arg\min_{\boldsymbol{\Sigma}>0}\{\operatorname{trace}(\boldsymbol{\Sigma}^{-1}\mathbf{S}) + \log|\mathbf{S}|\}$. This leads to the minimizers $\widehat{\boldsymbol{\Omega}}(\boldsymbol{\Gamma}) = \boldsymbol{\Gamma}^{\top}\widehat{\mathbf{M}}\boldsymbol{\Gamma}$ and $\widehat{\boldsymbol{\Omega}}_0(\boldsymbol{\Gamma}) = \boldsymbol{\Gamma}_0^{\top}(\widehat{\mathbf{M}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^{\top}\mathbf{Q}_{\boldsymbol{\Gamma}})\boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}_0^{\top}(\widehat{\mathbf{M}} + \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^{\top})\boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}_0^{\top}(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})\boldsymbol{\Gamma}_0$ from the last equality of $\ell_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$ above. The partially minimized objective function of $\boldsymbol{\Gamma}$ is finally

$$\begin{aligned}
\ell_n(\boldsymbol{\Gamma}) &= \log|\widehat{\boldsymbol{\Omega}}(\boldsymbol{\Gamma})| + \log|\widehat{\boldsymbol{\Omega}}_0(\boldsymbol{\Gamma})| + u + p - u \\
&= \log|\boldsymbol{\Gamma}^{\top}\widehat{\mathbf{M}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^{\top}(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})\boldsymbol{\Gamma}_0| + p \\
&= \log|\boldsymbol{\Gamma}^{\top}\widehat{\mathbf{M}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^{\top}(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1}\boldsymbol{\Gamma}| + \log|\widehat{\mathbf{M}} + \widehat{\mathbf{U}}| + p,
\end{aligned}$$

where the last equality is obtained from Lemma A.1. Thus, we have proven that $\ell_n(\boldsymbol{\Gamma}) = \mathrm{J}_n(\boldsymbol{\Gamma}) + \log|\widehat{\mathbf{M}} + \widehat{\mathbf{U}}| + p$ and the minimizer $\widehat{\boldsymbol{\Gamma}}$ for $\mathrm{J}_n(\boldsymbol{\Gamma})$ is also the minimizer of the partially minimized negative quasi-likelihood function $\ell_n(\boldsymbol{\Gamma})$. It is then straightforward to see that $\widehat{\mathbf{M}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\Gamma}})\widehat{\boldsymbol{\Gamma}}^{\top} + \widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\Omega}}_0(\widehat{\boldsymbol{\Gamma}})\widehat{\boldsymbol{\Gamma}}_0^{\top} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\mathbf{M}}\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\mathbf{M}}\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}}(\widehat{\boldsymbol{\Gamma}}) = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\theta}}$. □

Appendix C: Proof for Lemma 3.2

Proof. The Lemma's proof is similar to the proof of Lemma 6.3 of Cook et al. (2013). For completeness, we provide a complete proof here. For $u = 0$, it is clear that $\mathbf{U} = 0$ and thus $\operatorname{span}(\boldsymbol{\Gamma}_k)$ will be any k-dimensional reducing subspace of $\mathbf{M}$ for all k and $\mathrm{J}(\boldsymbol{\Gamma}_k) = 0 = \mathrm{J}(\boldsymbol{\Gamma}_u)$. For $u \geq 1$, we write $\mathrm{J}(\boldsymbol{\Gamma})$ as

$$\begin{aligned} \mathrm{J}(\boldsymbol{\Gamma}) &= \log|\boldsymbol{\Gamma}^{\mathsf{T}}\mathbf{M}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^{\mathsf{T}}(\mathbf{M}+\mathbf{U})^{-1}\boldsymbol{\Gamma}| \\ &= \log|\boldsymbol{\Gamma}^{\mathsf{T}}\mathbf{M}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^{\mathsf{T}}(\mathbf{M}+\mathbf{U})\boldsymbol{\Gamma}_0| - \log|\mathbf{M}+\mathbf{U}| \\ &\geq \log|\boldsymbol{\Gamma}^{\mathsf{T}}\mathbf{M}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^{\mathsf{T}}\mathbf{M}\boldsymbol{\Gamma}_0| - \log|\mathbf{M}+\mathbf{U}| \\ &\geq \log|\mathbf{M}| - \log|\mathbf{M}+\mathbf{U}|, \end{aligned}$$

where the first inequality attains its equality if and only if $\boldsymbol{\Gamma}_0^{\mathsf{T}}\mathbf{U}\boldsymbol{\Gamma}_0 = 0$, which is equivalent to $\operatorname{span}(\mathbf{U}) \subseteq \operatorname{span}(\boldsymbol{\Gamma})$; the second inequality attains its equality if and only if $\operatorname{span}(\boldsymbol{\Gamma})$ is a reducing subspace of $\mathbf{M}$. Since the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ is the smallest subspace satisfying both conditions, $k = u$ is the minimum dimension for $\mathrm{J}(\boldsymbol{\Gamma}_k)$ to achive the minimum $\mathrm{J}(\boldsymbol{\Gamma}_u) = \log|\mathbf{M}| - \log|\mathbf{M}+\mathbf{U}| < 0$. Hence, $\mathrm{J}(\boldsymbol{\Gamma}_u) < \mathrm{J}(\boldsymbol{\Gamma}_k) < 0$, for $0 < k < u$. So far, we only left to show that the minimum value $\mathrm{J}(\boldsymbol{\Gamma}_u)$ is achievable by $\mathrm{J}(\boldsymbol{\Gamma}_k)$ for $k > u$. Consider decomposing $\mathbf{M}$ as $\mathbf{M} = \boldsymbol{\Gamma}_u\boldsymbol{\Omega}\boldsymbol{\Gamma}_u^{\mathsf{T}} + \boldsymbol{\Gamma}_{0u}\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_{0u}^{\mathsf{T}}$, let $\mathbf{B}_{k-u} \in \mathbb{R}^{(p-u)\times(k-u)}$ be a semi-orthogonal basis for a reducing subspace of $\boldsymbol{\Omega}_0$. Then by letting $\boldsymbol{\Gamma}_k$ equal to $(\boldsymbol{\Gamma}_u, \mathbf{A}_{k-u})$, where $\mathbf{A}_{k-u} = \boldsymbol{\Gamma}_{0u}\mathbf{B}_{k-u} \in \mathbb{R}^{p\times(k-u)}$, it is straightforward to see that $\boldsymbol{\Gamma}_k$ is a reducing subspace of $\mathbf{M}$ that contains $\operatorname{span}(\mathbf{U})$ thus the minimum of the objective function is achieved: $\mathrm{J}(\boldsymbol{\Gamma}_k) = \mathrm{J}(\boldsymbol{\Gamma}_u)$. □

Appendix D: Proof for Theorem 3.1

Proof. We need to show that $\Pr(\mathcal{I}_n(k) - \mathcal{I}_n(u) > 0) \to 1$ as $n \to \infty$ for both $0 \leq k < u$ and $0 \leq u < k$ scenarios. By definition of $\mathcal{I}_n(k)$, we have

$$\mathcal{I}_n(k) - \mathcal{I}_n(u) = \mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_k) - \mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_u) + (k-u)\cdot\log(n)/n. \tag{D.1}$$

Firstly, for $0 \leq k < u$, suffice it to show that $\mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_k) - \mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_u) = \mathrm{J}(\boldsymbol{\Gamma}_k) - \mathrm{J}(\boldsymbol{\Gamma}_u) + o_p(1)$, where $\mathrm{J}(\boldsymbol{\Gamma}_u) < \mathrm{J}(\boldsymbol{\Gamma}_k) < 0$ from Lemma 3.2. We have $\mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_j) = \mathrm{J}(\boldsymbol{\Gamma}_j) + o_p(1)$ for all $j = 1, \ldots p$, because both the sample and population objective functions are essentially optimized over Grassmannian, i.e. $\boldsymbol{\Gamma}$ affects the objective functions $\mathrm{J}_n(\boldsymbol{\Gamma})$ and $\mathrm{J}(\boldsymbol{\Gamma})$ only through $\operatorname{span}(\boldsymbol{\Gamma})$. The functions are differentiable and the derivative $\nabla_k \mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_k) = \nabla_u \mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_k) = 0$, where ∇_k and ∇_u are derivatives over the Grassmannians $Gr(p,k)$ and $Gr(p,u)$, respectively.

Next, for $0 \leq u < k$, we show in the following that $\mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_k) - \mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_u) = \mathrm{J}(\boldsymbol{\Gamma}_k) - \mathrm{J}(\boldsymbol{\Gamma}_u) + O_p(n^{-1}) = O_p(n^{-1})$. It follows from (D.1) that the dominant term in $\mathcal{I}_n(k) - \mathcal{I}_n(u)$ is $(k-u)\cdot\log(n)/n$, which is a positive number. Therefore, $\Pr(\mathcal{I}_n(k) - \mathcal{I}_n(u) > 0) \to 1$ as $n \to \infty$ for $0 \leq u < k$. The special case of $u = 0$ is included in the derivation, as $\mathrm{J}(\boldsymbol{\Gamma}_u) = \mathrm{J}(\boldsymbol{\Gamma}_k) = 0$ for all k and $\boldsymbol{\Gamma}_k$ can be any k-dimensional reducing subspace of $\mathbf{M}$.

To show that $\mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_k) - \mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_u) = O_p(n^{-1})$ for $k > u$, we use the negative quasi-likelihood function $\ell_n(\mathbf{M}, \boldsymbol{\theta})$ in (3.1). By Lemma 3.1, we know that $\mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_k) - \mathrm{J}_n(\widehat{\boldsymbol{\Gamma}}_u) = \ell_n(\widehat{\boldsymbol{\Gamma}}_k) - \ell_n(\widehat{\boldsymbol{\Gamma}}_u) = \ell_n(\widehat{\mathbf{M}}_{\mathrm{Env},k}, \widehat{\boldsymbol{\theta}}_{\mathrm{Env},k}) - \ell_n(\widehat{\mathbf{M}}_{\mathrm{Env},u}, \widehat{\boldsymbol{\theta}}_{\mathrm{Env},u})$, where $\widehat{\mathbf{M}}_{\mathrm{Env}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}}$ is defined in Lemma 3.1 and we use additional subscript

k and u to distinguish different envelope basis $\widehat{\boldsymbol{\Gamma}}$ in $\widehat{\mathbf{M}}_{\mathrm{Env}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{Env}}$. We further use $\boldsymbol{\psi} = \{\mathrm{vech}^{\mathsf{T}}(\mathbf{M}), \mathrm{vec}^{\mathsf{T}}(\boldsymbol{\theta})\}^{\mathsf{T}} \in \mathbb{R}^{p(p+1)/2+pq}$ to denote the vector of all unique parameters in the quasi-likelihood function and write $\ell_n(\boldsymbol{\psi}) \equiv \ell_n(\mathbf{M},\boldsymbol{\theta})$, and define $\widehat{\boldsymbol{\psi}}$, $\widehat{\boldsymbol{\psi}}_k$ and $\widehat{\boldsymbol{\psi}}_u$ from the estimators $(\widehat{\mathbf{M}},\widehat{\boldsymbol{\theta}})$, $(\widehat{\mathbf{M}}_{\mathrm{Env},k},\widehat{\boldsymbol{\theta}}_{\mathrm{Env},k})$ and $(\widehat{\mathbf{M}}_{\mathrm{Env},u},\widehat{\boldsymbol{\theta}}_{\mathrm{Env},u})$, respectively. To show $\ell_n(\widehat{\boldsymbol{\psi}}_k) - \ell_n(\widehat{\boldsymbol{\psi}}_u) = O_p(n^{-1})$, we consider Taylor expansion of $\ell_n(\widehat{\boldsymbol{\psi}}_k)$ at $\widehat{\boldsymbol{\psi}}_u$: $\ell_n(\widehat{\boldsymbol{\psi}}_k) = \ell_n(\widehat{\boldsymbol{\psi}}_u) + \ell_n'(\widehat{\boldsymbol{\psi}}_u)(\widehat{\boldsymbol{\psi}}_k - \widehat{\boldsymbol{\psi}}_u) + (1/2)(\widehat{\boldsymbol{\psi}}_k - \widehat{\boldsymbol{\psi}}_u)^{\mathsf{T}}\ell_n''(\widetilde{\boldsymbol{\psi}}_u)(\widehat{\boldsymbol{\psi}}_k - \widehat{\boldsymbol{\psi}}_u)$ and

$$\ell_n(\widehat{\boldsymbol{\psi}}_k) - \ell_n(\widehat{\boldsymbol{\psi}}_u) = \ell_n'(\widehat{\boldsymbol{\psi}}_u)(\widehat{\boldsymbol{\psi}}_k - \widehat{\boldsymbol{\psi}}_u) + (1/2)(\widehat{\boldsymbol{\psi}}_k - \widehat{\boldsymbol{\psi}}_u)^{\mathsf{T}}\ell_n''(\widetilde{\boldsymbol{\psi}}_u)(\widehat{\boldsymbol{\psi}}_k - \widehat{\boldsymbol{\psi}}_u),$$

where $\widetilde{\boldsymbol{\psi}}_u$ is in the neighborhood of $\widehat{\boldsymbol{\psi}}_u$ so that we can find a series of $\widetilde{\boldsymbol{\psi}}_u$ such that $\ell_n''(\widetilde{\boldsymbol{\psi}}_u)$ converge in probability to a positive definite matrix in probability as $n \to \infty$. Since $k > u$, the estimators $\widehat{\boldsymbol{\psi}}_k$ is unbiased and $\sqrt{n}$-consistent. Recall that the objective function $\ell_n(\mathbf{M},\boldsymbol{\theta}) = \log|\mathbf{M}| + \mathrm{trace}[\mathbf{M}^{-1}\{\widehat{\mathbf{M}} + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathsf{T}}\}]$ is smooth and arbitrarily order differentiable with respect to $\mathbf{M} > 0$ and $\boldsymbol{\theta}$, and thus with respect to their unique elements vector $\boldsymbol{\psi}$. Therefore $\ell_n'(\boldsymbol{\psi})$ is a smooth differentiable function of $\boldsymbol{\psi}$ such that $\ell_n'(\widehat{\boldsymbol{\psi}}_u) = \ell_n'(\widehat{\boldsymbol{\psi}}) + O_p(n^{-1/2}) = 0 + O_p(n^{-1/2})$. For some $\widetilde{\boldsymbol{\psi}}_u$ in the neighborhood of $\widehat{\boldsymbol{\psi}}_u$ that $\widetilde{\boldsymbol{\psi}}_u \to \boldsymbol{\psi}_u$ in probability, $\ell_n''(\widetilde{\boldsymbol{\psi}}_u) = O_p(1)$. Since both $\widehat{\boldsymbol{\psi}}_k$ and $\widehat{\boldsymbol{\psi}}_u$ is $\sqrt{n}$-consistent and can be writen as $\boldsymbol{\psi} + O_p(n^{-1/2})$, we have $\ell_n(\widehat{\boldsymbol{\psi}}_k) - \ell_n(\widehat{\boldsymbol{\psi}}_u) = O_p(n^{-1/2}) * O_p(n^{-1/2}) + O_p(n^{-1/2}) * O_p(1) * O_p(n^{-1/2}) = O_p(n^{-1})$. □

Appendix E: Proof for Theorem 3.2

Proof. We re-write $\mathcal{I}_n^{\mathrm{1D}}(k)$, $k = 1, \ldots, p$, as $\mathcal{I}_n^{\mathrm{1D}}(k) = \sum_{j=1}^{k}\{\phi_j(\widehat{\mathbf{w}}_j) + \log(n)/n\}$. The increment $\mathcal{I}_n^{\mathrm{1D}}(k) - \mathcal{I}_n^{\mathrm{1D}}(k-1) = \phi_{k,n}(\widehat{\mathbf{w}}_k) + \log(n)/n$ is exactly the full Grassmannian criterion for the envelope $\mathcal{E}_{\mathbf{M}_k}(\mathbf{U}_k)$ evaluated at the one-dimensional envelope estimator. From the following proof, we will show that the negative term $\phi_{k,n}(\widehat{\mathbf{w}}_k)$ dominates the positive term $\log(n)/n$ for $k < u$ because the envelope $\mathcal{E}_{\mathbf{M}_k}(\mathbf{U}_k)$ has dimension greater than 0, then the positive term $\log(n)/n$ will dominate the negative term for $k > u$ because the envelope $\mathcal{E}_{\mathbf{M}_k}(\mathbf{U}_k)$ has dimension zero. More specifically, we claim that the following two statements are true:

1. for $j \leq u$, $\phi_{j,n}(\widehat{\mathbf{w}}_j) + \log(n)/n$ converges to a negative constant $\phi_j(\mathbf{w}_j) < 0$, in probability, as $n \to \infty$; and
2. for $j > u$, $\phi_{j,n}(\widehat{\mathbf{w}}_j) = O_p(n^{-1})$ and $\Pr\left(\phi_j(\widehat{\mathbf{w}}_j) + \log(n)/n > 0\right) \to 1$ as $n \to \infty$.

Then the first statement implies that, for $j < u$, $\Pr\left(\mathcal{I}_n^{\mathrm{1D}}(k) - \mathcal{I}_n^{\mathrm{1D}}(u) > 0\right) \to 1$ as $n \to \infty$; and the second statement implies that for $j > u$, $\Pr\big(\mathcal{I}_n^{\mathrm{1D}}(k) - \mathcal{I}_n^{\mathrm{1D}}(u) > 0\big) \to 1$ as $n \to \infty$. The conclusion, $\Pr(\widehat{u}_{\mathrm{1D}} = u) \to 1$ as $n \to \infty$, thus follows from the above two statements, which are proved in the following.

From Proposition 4 in Cook and Zhang [11], we know that $\mathbf{w}_{k+1} \in \mathcal{E}_{\mathbf{M}_k}(\mathbf{U}_k)$ implies $\mathbf{g}_{k+1} = \mathbf{G}_{0k}\mathbf{w}_{k+1}/||\mathbf{w}_{k+1}|| \in \mathcal{E}_{\mathbf{M}}(\mathbf{U})$, and that $\mathcal{E}_{\mathbf{M}_k}(\mathbf{U}_k)$ has dimension greater than zero (i.e. $\mathcal{E}_{\mathbf{M}_k}(\mathbf{U}_k)$ not equals to the origin) if and only if

$k \leq u$. Then, for $j \leq u$, the first statement follows because $\phi_{j,n}(\mathbf{w})$ is a smooth differentiable function of $\mathbf{w}$ and $\widehat{\mathbf{w}}_j$ is $\sqrt{n}$-consistent for $\mathbf{w}_j$ (in terms of their projection matrices, upon which the functional value $\phi_{n,j}(\mathbf{w})$ solely depends). The function $\phi_{j,n}(\widehat{\mathbf{w}}_j)$ converges to a negative value $\phi_j(\mathbf{w}_j) < 0$ in probability as shown in the proof of Propositions 5 and 6 in Cook and Zhang [11]. The proof of Theroem 3.1 only requires $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ to be $\sqrt{n}$-consistent estimators. Now $\widehat{\mathbf{M}}_k$ and $\widehat{\mathbf{U}}_k$ are also $\sqrt{n}$-consistent [Proposition 6; 11]. For $j > u$, the second statement $\phi_{j,n}(\widehat{\mathbf{w}}_j) - 0 = O_p(n^{-1})$ can be proved following the lines of proof for Theroem 3.1, by replacing $\mathrm{J}_n(\widehat{\mathbf{\Gamma}})$, $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ with $\phi_{n,j}(\widehat{\mathbf{w}})$ and $\mathcal{E}_{\mathbf{M}_k}(\mathbf{U}_k)$ and by noticing this is the special case of $k = 1 > u = 0$ for Theroem 3.1. □

Acknowledgements

Research for this paper was partly supported by grants CCF-1617691 and DMS-1613154 from the U.S. National Science Foundation. The authors are grateful to the Editor, Associate Editor and two referees for their insightful comments and suggestions.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723. MR0423716

[2] Bura, E. and Cook, R. D. (2003). Rank estimation in reduced-rank regression. *Journal of Multivariate Analysis*, 87(1):159–176. MR2007266

[3] Conway, J. (1990). *A Course in Functional Analysis. Second edition.* Springer, New York. MR1070713

[4] Cook, R. D., Forzani, L., and Zhang, X. (2015). Envelopes and reduced-rank regression. *Biometrika*, 102(2):439–456. MR3371015

[5] Cook, R. D., Helland, I. S., and Su, Z. (2013). Envelopes and partial least squares regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(5):851–877. MR3124794

[6] Cook, R. D. and Li, B. (2004). Determining the dimension of iterative hessian transformation. *The Annals of Statistics*, 32(6):2501–2531. MR2153993

[7] Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica*, 20(3):927–960. MR2729839

[8] Cook, R. D. and Su, Z. (2013). Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression. *Biometrika*, 100(4):939–954. MR3142342

[9] Cook, R. D. and Zhang, X. (2015a). Foundations for envelope models and methods. *Journal of the American Statistical Association*, 110(510):599–611. MR3367250

[10] Cook, R. D. and Zhang, X. (2015b). Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57(1):11–25. MR3318345

[11] Cook, R. D. and Zhang, X. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, 25(1):284–300. MR3474048

[12] Cook, R. D. and Zhang, X. (2018). Fast envelope algorithms. *Statistica Sinica*, 28(3):1179–1197.

[13] Eck, D. J. and Cook, R. D. (2017). Weighted envelope estimation to handle variability in model selection. *Biometrika*, 104(3):743–749. MR3694595

[14] Eck, D. J., Geyer, C. J., and Cook, R. D. (2017). An application of envelope methodology and aster models. *arXiv preprint arXiv:1701.07910*.

[15] Geyer, C. J., Wagenius, S., and Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, 94(2):415–426. MR2380569

[16] Hawkins, D. M. and Maboudou-Tchao, E. M. (2013). Smoothed linear modeling for smooth spectral data. *International Journal of Spectroscopy*.

[17] Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146. MR3735365

[18] Ma, Y. and Zhang, X. (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika*, 102(2):409–420. MR3371013

[19] Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89(425):141–148. MR1266291

[20] Schott, J. R. (2013). On the likelihood ratio test for envelope models in multivariate linear regression. *Biometrika*, 100(2):531–537. MR3068454

[21] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464. MR0468014

[22] Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, 98(1):133–146. MR2804215

[23] Yu, Y., Wang, T., and Samworth, R. (2015). A useful variant of the davis-kahan theorem for statisticians. *Biometrika*, 102(2). MR3371006

[24] Zeng, P. (2008). Determining the dimension of the central subspace and central mean subspace. *Biometrika*, 95(2):469–479. MR2521593

[25] Zhang, X. and Li, L. (2017). Tensor envelope partial least-squares regression. *Technometrics*, 59(4):426–436. MR3740960

[26] Zhu, L., Miao, B., and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474):630–643. MR2281245

[27] Zhu, L.-P., Zhu, L.-X., and Feng, Z.-H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105(492):1455–1466. MR2796563

[28] Zhu, X., Wang, T., and Zhu, L. (2016). Dimensionality determination: a thresholding double ridge ratio criterion. *arXiv preprint arXiv:1608.04457*.

[29] Zou, C. and Chen, X. (2012). On the consistency of coordinate-independent sparse estimation with bic. *Journal of Multivariate Analysis*, 112:248–255. MR2957301