Algorithms for Envelope Estimation

R. Dennis Cook* and Xin Zhang†

September 19, 2014

Abstract

Envelopes were recently proposed as methods for reducing estimative variation in multivariate linear regression. Estimation of an envelope usually involves optimization over Grassmann manifolds. We propose a fast and widely applicable one-dimensional (1D) algorithm for estimating an envelope in general. We reveal an important structural property of envelopes that facilitates our algorithm, and we prove both Fisher consistency and $\sqrt{n}$-consistency of the algorithm.

Key Words: Envelopes; Grassmann manifold; reducing subspaces.

1 Introduction

Envelope methods aim to reduce estimative variation in multivariate linear models. The reduction is typically associated with predictors or responses, and can generally be interpreted as effective dimensionality reduction in the parameter space. Such reduction is achieved by enveloping the variation in the data that is material to the goals of the analysis while simultaneously excluding the immaterial variation. Efficiency gains are then achieved by essentially basing estimation on the material variation alone. The improvement in estimation and prediction can be quite substantial when the immaterial variation is large, sometimes equivalent to taking thousands of additional observations.

*R. Dennis Cook is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: dennis@stat.umn.edu).

†Xin Zhang is Assistant Professor, Department of Statistics, Florida State University, Tallahassee, FL, 32306 (Email: henry@stat.fsu.edu).

The novel notion of an envelope, which is a subspace of the predictor or response spaces containing all of the material variation, was first introduced by Cook et al. (2010) for response reduction in multivariate linear models, subsequently studied by Su and Cook (2011) for partial reduction and recently studied by Cook et al. (2013) for predictor reduction. In particular, Cook et al. (2013) found that the commonly used PLS algorithm, SIMPLS (de Jong 1993), is in fact based on a $\sqrt{n}$-consistent envelope estimator, while the corresponding likelihood-based approach produces a better estimator.

The likelihood-based approach to envelope estimation requires, for a given envelope dimension u, optimizing an objective function of the form $f(\mathbf{\Gamma})$, where $\mathbf{\Gamma}$ is a $k \times u$, $k > u$, semi-orthogonal basis matrix for the envelope. The objective function satisfies $f(\mathbf{\Gamma}) = f(\mathbf{\Gamma O})$ for any $u \times u$ orthogonal matrix $\mathbf{O}$. Hence the optimization is essentially over the set of all u-dimensional subspaces of $\mathbb{R}^k$, which is a Grassmann manifold denoted as $\mathcal{G}_{u,k}$. Since $u(k-u)$ real numbers are required to specify an element of $\mathcal{G}_{u,k}$ uniquely, the optimization is essentially over $u(k-u)$ real dimensions. In multivariate linear regression, k can be either the number of responses r or the number of predictors p, depending on whether one is pursuing response or predictor reduction.

All present envelope methods rely on the Matlab package `sg_min` by Ross A. Lippert (`http://web.mit.edu/~ripper/www/software/`) to optimize $f(\mathbf{\Gamma})$. This package provides iterative optimization techniques on Stiefel and Grassmann manifolds, including non-linear conjugate gradient (PRCG and FRCG) iterations, dog-leg steps and Newton's method. To implement an envelope estimation procedure, one needs to specify the objective function $f(\mathbf{\Gamma})$ and its analytical first-order derivative function. Then given an initial value of $\mathbf{\Gamma}$, this package will compute numerical second-order derivatives and iterate until convergence or the maximum number of iterations is reached. The Matlab toolbox `envlp` by R. D. Cook, Z. Su and Y. Yang (`http://code.google.com/p/envlp/`) uses `sg_min` to implement a variety of envelope estimators along with associated inference methods. The `sg_min` package works well for envelope estimation when $u(k-u)$ is not too large, but optimization is often

computationally difficult for large values of $u(k-u)$. At higher dimensions, each iteration becomes exponentially slower, local minima can become a serious issue and good starting values are essential. The `envlp` toolbox implements a different version of $f(\mathbf{\Gamma})$ for each type of envelope, along with tailored starting values.

In this article we present two advances in envelope computation. First, we propose in Section 3 a model-free objective function $J_n(\mathbf{\Gamma})$ for estimating an envelope and show that the three major envelope methods are based on special cases of J_n. This unifying objective function is to be optimized over the Grassmann manifold $\mathcal{G}_{u,k}$, which for larger values of $u(k-u)$ will be subject to the same computational limitations associated with speed, local minima and starting values. Second, we propose in Section 4 a fast one-dimensional (1D) algorithm that mitigates these computational issues. To adapt the envelope construction for relatively large values of $u(k-u)$, we break down Grassmann optimization into a series of one-dimensional optimizations so that the estimation procedure is speeded up greatly, and starting values and local minima are no longer an issue. Although it may be impossible to break down a general u-dimensional Grassmann optimization problem, we rely on special characteristics of envelopes in statistical problems to achieve the breakdown. The resulting 1D algorithm, which is easy-to-implement, stable and requires no initial value input, can be tens to hundreds times faster than the general Grassmann manifold optimization for $u > 1$, while still providing a desirable $\sqrt{n}$-consistent envelope estimator. Very recently, Cook and Zhang (2014) introduced simultaneous reduction of the predictors and the response by envelopes. The objective function in Cook and Zhang (2014) has the form of $f(\mathbf{L}, \mathbf{R})$ where $\mathbf{L}$ and $\mathbf{R}$ are both semi-orthogonal matrices and the optimization is over two Grassmann manifolds. They used special forms of the 1D algorithm to find initial values for $\mathbf{L}$ and $\mathbf{R}$. The 1D algorithm we introduce in Section 4 is much more general and is directly applicable beyond the multivariate linear regression context.

The rest of this article is organized as follows. In Section 2, we review briefly key algebraic foundations of envelopes, and also review concepts and methodology in the context of an example. Because envelopes are nascent methodology, the level of detail in this example is

somewhat greater than what might be considered traditional. Section 5 consists of simulation studies and a data example to further demonstrate the advantages of the 1D algorithm. Proofs and technical details are included in the Appendix.

The following notations and definitions will be used in our exposition. Let $\mathbb{R}^{m\times n}$ be the set of all real $m \times n$ matrices and let $\mathbb{S}^k$ be the set of all real and symmetric $k \times k$ matrices. Suppose $\mathbf{M} \in \mathbb{R}^{m\times n}$, then $\mathrm{span}(\mathbf{M}) \subseteq \mathbb{R}^m$ is the subspace spanned by columns of $\mathbf{M}$. We use $\mathbf{P}_{\mathbf{A}(\mathbf{V})} = \mathbf{A}(\mathbf{A}^T\mathbf{V}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{V}$ to denote the projection onto $\mathrm{span}(\mathbf{A})$ with the $\mathbf{V}$ inner product and use $\mathbf{P}_{\mathbf{A}}$ to denote projection onto $\mathrm{span}(\mathbf{A})$ with the identity inner product. Let $\mathbf{Q}_{\mathbf{A}(\mathbf{V})} = \mathbf{I} - \mathbf{P}_{\mathbf{A}(\mathbf{V})}$. Sample covariance matrices are represented as $\mathbf{S}_{(\cdot)}$ and defined with the divisor n. For instance, $\mathbf{S}_{\mathbf{X}} = \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T/n$, $\mathbf{S}_{\mathbf{XY}} = \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T/n$ and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ denotes the covariance matrix of the residuals from the linear fit of $\mathbf{Y}$ on $\mathbf{X}$: $\mathbf{S}_{\mathbf{Y}|\mathbf{X}} = \mathbf{S}_{\mathbf{Y}} - \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{XY}}$.

2 Review of envelopes

2.1 Definition of an envelope

The following definition of a reducing subspace is equivalent to the usual definition found in functional analysis (Conway 1990) and in the literature on invariant subspaces, but the underlying notion of reduction is incompatible with how it is usually understood in statistics. Nevertheless, it is common terminology in those areas and is the basis for the definition of an envelope (Cook, et al., 2010) which is central to our developments.

Definition 1. *A subspace $\mathcal{R} \subseteq \mathbb{R}^d$ is said to be a reducing subspace of $\mathbf{M} \in \mathbb{R}^{d\times d}$ if $\mathcal{R}$ decomposes $\mathbf{M}$ as $\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$ reduces $\mathbf{M}$.*

The next definition shows how to construct an envelope in terms of reducing subspaces.

Definition 2. *Let $\mathbf{M} \in \mathbb{S}^d$ and let $\mathcal{B} \subseteq \mathrm{span}(\mathbf{M})$. Then the $\mathbf{M}$-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{B}$.*

The intersection of two reducing subspaces of $\mathbf{M}$ is still a reducing subspace of $\mathbf{M}$. This means that $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$, which is unique by its definition, is the smallest reducing subspace containing $\mathcal{B}$. Also, the $\mathbf{M}$-envelope of $\mathcal{B}$ always exist because of the requirement $\mathcal{B} \subseteq \mathrm{span}(\mathbf{M})$. In regression applications, when the subspace $\mathcal{B}$ is represented as $\mathcal{B} = \mathrm{span}(\mathbf{U})$ for some matrix $\mathbf{U}$, we write $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) := \mathcal{E}_{\mathbf{M}}(\mathrm{span}(\mathbf{U})) = \mathcal{E}_{\mathbf{M}}(\mathcal{B})$ to avoid notation proliferation. Let $\mathcal{E}_{\mathbf{M}}^{\perp}(\mathbf{U})$ denote the orthogonal complement of $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$.

The following proposition from Cook, et al. (2010) gives a characterization of envelopes.

Proposition 1. *Let $q \leq d$ denote the number of eigenspaces of $\mathbf{M} \in \mathbb{S}^d$. Then $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) = \sum_{i=1}^{q} \mathbf{P}_i \mathcal{B}$, where $\mathbf{P}_i$ is the projection onto the i-th eigenspace of $\mathbf{M}$.*

From this proposition, we see that the $\mathbf{M}$-envelope of $\mathcal{B}$ is the sum of the eigenspaces of $\mathbf{M}$ that are not orthogonal to $\mathcal{B}$; that is, the eigenspaces of $\mathbf{M}$ onto which $\mathcal{B}$ projects non-trivially. This implies that the envelope is the span of some subset of the eigenspaces of $\mathbf{M}$. In the regression context, $\mathcal{B}$ is typically the span of a regression coefficient matrix or a matrix of cross-covariances, and $\mathbf{M}$ is chosen as a covariance matrix which is usually positive definite. We next illustrate the potential gain of envelope method using a linear regression example.

2.2 Concepts and the response envelope method

We use Kenward's (1987) data to illustrate the working mechanism of envelopes in multivariate linear regression. These data came from an experiment to compare two treatments for the control of an intestinal parasite in cattle. Thirty animals were randomly assigned to each of the two treatments. Their weights (in kilograms) were recorded at the beginning of the study prior to treatment application and at 10 times during the study corresponding to weeks 2, 4, 6, ..., 18 and 19; that is, at two-weeks intervals except the last which was over a one-week interval. The goal was to find if there is a detectable difference between the two treatments and, if such a difference exists, the time at which it first occurred. As emphasized by Kenward (1987), although these data have a typical longitudinal structure, the nature of the disease means that growth during the experiment is not amenable to modeling as a smooth function of time, and

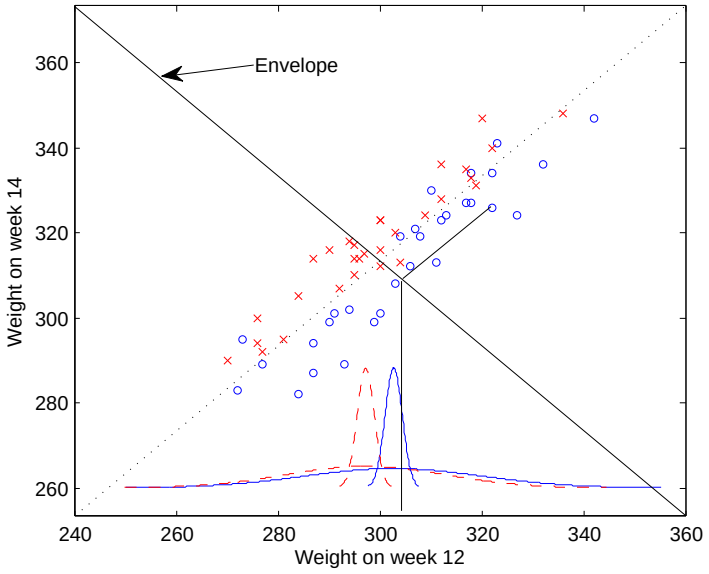

Figure 2.1: Kenward's cow data with the 30 animals receiving one treatment marked as o's and the 30 animals receiving the other marked as x's.The curves on the bottom are densities of $Y_6|(X=0)$ and $Y_6|(X=1)$: the flat two curves are obtained by projecting the data onto the Y_6 axis (standard analysis), and the two other densities are obtained by first project the data onto the envelope and then onto the Y_6 axis (envelope analysis). One representative projection path is shown on the plot.

that fitting growth profiles with a low degree polynomial may hide interesting features of the data because the mean growth curves for the two treatment groups are very close relative to their variation from animal to animal. Kenward modeled the data using a multivariate linear model with an "ante-dependence" covariance structure. Here we proceed with an envelope analysis based on a multivariate linear model, following the response envelope structure outlined by Cook et al. (2010). Later in Section 3 we demonstrate how various envelope methods, including response envelopes, partial envelopes and predictor envelopes, can be unified into a general objective function framework.

Neglecting the basal measurement for simplicity, let $\mathbf{Y}_i \in \mathbb{R}^{10}$, $i = 1, \ldots, 60$, be the vector of weight measurements of each animal over time and let $X_i = 0$ or 1 indicate the two treatments. Our interest lies in the regression coefficient $\boldsymbol{\beta}$ from the multivariate linear regression $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta} X + \boldsymbol{\epsilon}$, where it is assumed that $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$. Let $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ denote the ordinary least squares estimator of $\boldsymbol{\beta}$, which is also the maximum likelihood estimator. The

estimates and their residual bootstrap standard errors are shown in Table 1. The maximum absolute t-value over the elements of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is 1.30, suggesting that the treatments do not have a differential affect on animal weight. However, with a value of 26.9 on 10 degrees of freedom, the likelihood ratio statistics for the hypothesis $\boldsymbol{\beta} = 0$ indicates otherwise. We next turn to an envelope analysis.

We first consider the regression of the 6th and 7th element of $\mathbf{Y}$, corresponding to weeks 12 and 14, on X, and now letting $\mathbf{Y} = (Y_6, Y_7)^T$. This allows us to represent the regression graphically and thereby provide intuition on the working mechanism of an envelope analysis. Figure 2.1 shows a plot of Y_6 versus Y_7 with the points marked by treatment. Since $\boldsymbol{\beta} = \mathrm{E}(\mathbf{Y}|X = 1) - \mathrm{E}(\mathbf{Y}|X = 0) \in \mathbb{R}^{2\times 1}$, the standard estimator for $\boldsymbol{\beta}$ is obtained as the difference in the marginal means after projecting the data onto the horizontal and vertical axes of the plot. The two densities estimates with the larger variation shown along the horizontal axes of the plot represent this operation. These density estimates are nearly identical, which explains the relatively small t-values from the standard model mentioned previously. However, it is clear from the figure that the treatments do differ. An envelope analysis infers that $\boldsymbol{\beta} = (\beta_6, \beta_7)^T$ is parallel to the second eigenvector of $\boldsymbol{\Sigma} = \mathrm{cov}(Y_6, Y_7)$. Hence by Proposition 1, the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) = \mathrm{span}(\boldsymbol{\beta})$, as shown on the plot. The envelope represents the subspace in which the populations differ, which seems consistent with the pattern of variation shown in the plot. The orthogonal complement of the envelope, represented by a dashed line on the plot, represents the immaterial variation which contains no information about the difference $\boldsymbol{\beta}$. The two populations are inferred to be the same when projected onto this subspace, which also seems consistent with the pattern of variation in the plot. The envelope estimator of a mean difference is obtained by first projecting the points onto the envelope and thus removing the immaterial variation, and then projecting the points onto the horizontal or vertical axis. The two density estimates with the smaller variation represent this operation for inference on β_6. These densities are well separated, leading to increased efficiency.

Back to the original problem of ten responses, we first formally incorporate a basis for

the response envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) \subseteq \mathbb{R}^{10}$ into the multivariate linear model $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta} X + \boldsymbol{\epsilon}$. Let $\boldsymbol{\Gamma} \in \mathbb{R}^{10\times u}$ be a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ and let $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ be an orthogonal matrix. Then $\mathrm{span}(\boldsymbol{\beta}) \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ and we can express $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^{u\times 1}$ carries the coordinates of $\boldsymbol{\beta}$ relative to the basis $\boldsymbol{\Gamma}$ and $1 \leq u \leq 10$. The envelope version of the multivariate linear model can now be written as $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta} X + \boldsymbol{\epsilon}$, with $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Omega} \in \mathbb{S}^u$ and $\boldsymbol{\Omega}_0 \in \mathbb{S}^{(10-u)}$ are positive definite matrices. Under this envelope model, $\boldsymbol{\Gamma}_0^T\mathbf{Y}|X \sim \boldsymbol{\Gamma}_0^T\mathbf{Y}$ and $\boldsymbol{\Gamma}_0^T\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\Gamma}^T\mathbf{Y}|X$. Consequently, $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ does not respond to changes in X either marginally or because of an association with $\boldsymbol{\Gamma}^T\mathbf{Y}$. For these reasons we regard $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ as the immaterial information and $\boldsymbol{\Gamma}^T\mathbf{Y}$ as the material information. In Figure 2.1, the immaterial variation is along the dashed diagonal line and the material variation is along the solid diagonal line. Envelope analysis are particularly effective when the immaterial variation $\mathrm{var}(\boldsymbol{\Gamma}_0^T\mathbf{Y})$ is large relative to the material variation $\mathrm{var}(\boldsymbol{\Gamma}^T\mathbf{Y})$, which can be seen from Figure 2.1.

Using notation that connects with later developments, the Grassmann objective function for estimating a basis $\boldsymbol{\Gamma}$ of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ is $\log|\boldsymbol{\Gamma}^T\widehat{\mathbf{M}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1}\boldsymbol{\Gamma}|$, where $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|X}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}}$ with population versions $\mathbf{M} = \boldsymbol{\Sigma}$ and $\mathbf{U} = \boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\beta}^T$. After finding a value $\widehat{\boldsymbol{\Gamma}}$ of $\boldsymbol{\Gamma}$ that minimizes this objective function, which will be discussed in Section 3, over all semi-orthogonal matrices $\boldsymbol{\Gamma} \in \mathbb{R}^{10\times u}$, the envelope estimator of $\boldsymbol{\beta}$ is given by $\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. Because $u(10-u) \leq 25$, the real dimensions involved in this optimization are small and the `envlp` code does not encounter computational issues. Standard methods like BIC and likelihood ratio testing can be used to guide the choice of the envelope dimension u. Both methods indicate clearly that $u = 1$ in this illustration. In other words, the treatment difference is manifested in only one linear combination $\boldsymbol{\Gamma}^T\mathbf{Y}$ of the response vector.

The envelope estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ is shown in Table 1 along with bootstrap standard errors and standard errors obtained from the asymptotic normal distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{env}} - \boldsymbol{\beta})$ by the plug-in method (See Cook et al. (2010) for the asymptotic covariance matrix). We see that the asymptotic standard errors are a bit smaller than the bootstrap standard errors. Using either set of standard errors and using a Bonferroni adjustment for multiple testing, we see that there

OLS estimator										
Week	2	4	6	8	10	12	14	16	19	19
$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$	2.4	3.3	3.1	4.7	4.7	5.5	-4.8	-4.5	-2.8	5.0
Bootstrap SE	2.9	3.2	3.5	3.6	4.0	4.2	4.4	4.5	5.4	6.0
Envelope estimator										
$\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$	-2.2	-0.5	0.9	2.4	2.9	5.4	-5.1	-4.6	-3.7	4.2
Bootstrap SE	1.13	0.84	1.07	1.03	0.81	1.12	1.07	1.04	1.08	1.02
Asymptotic SE/$\sqrt{n}$	0.88	0.74	0.72	0.84	0.70	1.02	0.92	0.86	0.90	0.85
Bootstrap SE ratios of OLS estimator over envelope estimator										
SE ratios	2.6	3.8	3.3	3.5	5.0	3.7	4.1	4.3	5.0	5.9

Table 1: Bootstrap standard errors of the 10 elements in $\widehat{\boldsymbol{\beta}}$ under the OLS estimator and the envelope estimator with $u = 1$. The bootstrap standard errors were estimated using 100 bootstrap samples.

is a difference between the treatments and that the difference is first manifested around week 10 and remains thereafter. As shown in the final row of Table 1, the bootstrap standard errors for the elements of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ were 2.2 to 5.9 times those of $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$. Hundreds of additional samples would be needed to reduce the standard errors of the elements of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ by these amounts.

In this illustration we used the response envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ to improve efficiency by effectively reducing the multivariate response. Other types of envelopes have been studied in the literature for various settings. We next propose a unified objective function for estimating an arbitrary envelope, denoted by $\mathcal{E}_{\mathrm{M}}(\mathrm{U})$, and show how it recovers previous constructs and inspires new envelope methods.

3 Objective functions for estimating an envelope

3.1 The objective function and its properties

In this section we propose a generic objective function for estimating a basis $\boldsymbol{\Gamma}$ of an arbitrary envelope $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) \subseteq \mathbb{R}^d$, where $\mathbf{M} \in \mathbb{S}^d$ is a positive definite matrix. Let $\mathcal{B}$ be spanned by a $d \times d$ matrix $\mathbf{U}$ so that $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) = \mathcal{E}_{\mathbf{M}}(\mathbf{U})$. Because $\mathrm{span}(\mathbf{U}) = \mathrm{span}(\mathbf{U}\mathbf{U}^T)$, we can always denote the envelope by $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ for some symmetric matrix $\mathbf{U} \geq 0$. We propose the following

generic population objective function for estimating $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$:

$$J(\mathbf{\Gamma}) = \log|\mathbf{\Gamma}^T\mathbf{M}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{\Gamma}|, \tag{3.1}$$

where $\mathbf{\Gamma} \in \mathbb{R}^{d\times u}$ denotes a semi-orthogonal basis for elements in Grassmann manifold $\mathcal{G}_{u,d}$, u is the dimension of the envelope, and $u < d$. We refer to the operation of optimizing (3.1) or its sample version given later in (3.2) as full Grassmann (FG) optimization. Since $J(\mathbf{\Gamma}) = J(\mathbf{\Gamma}\mathbf{O})$ for any orthogonal $u \times u$ matrix $\mathbf{O}$, the minimizer $\widetilde{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} J(\mathbf{\Gamma})$ is not unique. But we are interested only in $\mathrm{span}(\widetilde{\mathbf{\Gamma}})$, which is unique as shown in the following proposition.

Proposition 2. *Let $\widetilde{\mathbf{\Gamma}} \in \mathbb{R}^{d\times u}$ be any minimizer of $J(\mathbf{\Gamma})$. Then $\mathrm{span}(\widetilde{\mathbf{\Gamma}}) = \mathcal{E}_{\mathbf{M}}(\mathbf{U})$.*

To gain intuition on how $J(\mathbf{\Gamma})$ is minimized by any $\widetilde{\mathbf{\Gamma}}$ that spans the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$, we let $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{d\times d}$ be an orthogonal matrix and decompose the objective function into two parts: $J(\mathbf{\Gamma}) = J^{(1)}(\mathbf{\Gamma}) + J^{(2)}(\mathbf{\Gamma})$, where

$$\begin{aligned} J^{(1)}(\mathbf{\Gamma}) &= \log|\mathbf{\Gamma}^T\mathbf{M}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}_0^T\mathbf{M}\mathbf{\Gamma}_0|, \\ J^{(2)}(\mathbf{\Gamma}) &= \log|\mathbf{\Gamma}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{\Gamma}| - \log|\mathbf{\Gamma}_0^T\mathbf{M}\mathbf{\Gamma}_0| \\ &= \log|\mathbf{\Gamma}_0^T(\mathbf{M}+\mathbf{U})\mathbf{\Gamma}_0| - \log|\mathbf{\Gamma}_0^T\mathbf{M}\mathbf{\Gamma}_0| - \log|\mathbf{M}+\mathbf{U}|. \end{aligned}$$

The first function $J^{(1)}(\mathbf{\Gamma})$ is minimized by any $\mathbf{\Gamma}$ that spans a reducing subspace of $\mathbf{M}$. Minimizing the second function $J^{(2)}(\mathbf{\Gamma})$ is equivalent to minimizing $\log|\mathbf{\Gamma}_0^T(\mathbf{M}+\mathbf{U})\mathbf{\Gamma}_0| - \log|\mathbf{\Gamma}_0^T\mathbf{M}\mathbf{\Gamma}_0|$, which is no less than zero and equals to zero when $\mathbf{\Gamma}_0^T\mathbf{U}\mathbf{\Gamma}_0 = 0$. Thus $J^{(2)}(\mathbf{\Gamma})$ is minimized by any $\mathbf{\Gamma}$ such that $\mathbf{\Gamma}_0^T\mathbf{U}\mathbf{\Gamma}_0 = 0$, or equivalently, $\mathrm{span}(\mathbf{U}) \subseteq \mathrm{span}(\mathbf{\Gamma})$. These properties of $J^{(1)}(\mathbf{\Gamma})$ and $J^{(2)}(\mathbf{\Gamma})$ are combined by $J(\mathbf{\Gamma})$ to get a reducing subspace of $\mathbf{M}$ that contains $\mathrm{span}(\mathbf{U})$. In the context of multivariate linear regression, minimizing $J^{(2)}(\mathbf{\Gamma})$ is related to minimizing the residual sum of squares while minimizing $J^{(1)}$ is in effect pulling the solution towards principal components of responses or predictors. Finally, because u is the dimension of the envelope, the minimizer $\mathrm{span}(\widetilde{\mathbf{\Gamma}})$ is unique by Definition 2.

The sample version J_n of J based on a sample of size n is constructed by substituting

estimators $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ of $\mathbf{M}$ and $\mathbf{U}$:

$$J_n(\mathbf{\Gamma}) = \log|\mathbf{\Gamma}^T\widehat{\mathbf{M}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1}\mathbf{\Gamma}|. \tag{3.2}$$

Proposition 2 shows Fisher consistency of minimizers from optimizing the population objective function. Furthermore, $\sqrt{n}$-consistency of $\widehat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} J_n(\mathbf{\Gamma})$ is stated in the following proposition.

Proposition 3. *Let $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ denote $\sqrt{n}$-consistent estimators for $\mathbf{M} > 0$ and $\mathbf{U} \geq 0$. Let $\widehat{\mathbf{\Gamma}} \in \mathbb{R}^{d\times u}$ be a minimizer of $J_n(\mathbf{\Gamma})$, then $\mathbf{P}_{\widehat{\mathbf{\Gamma}}}$ is $\sqrt{n}$-consistent for the projection onto $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$.*

When we connect the objective function $J_n(\mathbf{\Gamma})$ with multivariate linear models in Section 3.2, we will find that previous likelihood-based envelope objective functions can be written in form (3.2). The likelihood approach to envelope estimation is based on normality assumptions for the conditional distribution of the response given the predictors or the joint distribution of the predictors and responses. The envelope objective function arising from this approach is a partially maximized log-likelihood obtained broadly a follows. After incorporating the envelope structure into the model, partially maximize the normal log-likelihood function $L_n(\boldsymbol{\psi}, \mathbf{\Gamma})$ over all the other parameters $\boldsymbol{\psi}$ with $\mathbf{\Gamma}$ fixed. This leads to a likelihood-based objective function $L_n(\mathbf{\Gamma})$, which equals a constant plus $-(n/2)J_n(\mathbf{\Gamma})$ with $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ depending on context. Proposition 3 indicates that the function $J_n(\mathbf{\Gamma})$ can be used as a generic moment-based objective function requiring only $\sqrt{n}$-consistent matrices $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$. Consequently, normality is not a requirement for estimators based on $J_n(\mathbf{\Gamma})$ to be useful, a conclusion that is supported by previous work and by our experience. FG optimization of $J_n(\mathbf{\Gamma})$ can be computationally intensive and can require a good initial value. The 1D algorithm in Section 4 mitigates the computational issues.

3.2 Connections with previous work

Envelope applications have so far been mostly restricted to the homoscedastic multivariate linear model

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n, \tag{3.3}$$

where $\mathbf{Y} \in \mathbb{R}^r$, the predictor vector $\mathbf{X} \in \mathbb{R}^p$, $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$, $\boldsymbol{\alpha} \in \mathbb{R}^r$ and the errors $\boldsymbol{\varepsilon}_i$ are independent copies of the normal random vector $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$. The maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are then $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{X}}^{-1}$ and $\widehat{\boldsymbol{\Sigma}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$.

3.2.1 Response envelopes

Cook, et al. (2010) studied response envelopes for estimation of the coefficient matrix $\boldsymbol{\beta}$. They conditioned on the observed values of $\mathbf{X}$ and motivated their developments by allowing for the possibility that some linear combinations of the response vector $\mathbf{Y}$ are immaterial to the estimation of $\boldsymbol{\beta}$, as described previously in Section 2.2. Reiterating, suppose that there is an orthogonal matrix $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r\times r}$ so that (i) $\mathrm{span}(\boldsymbol{\beta}) \subseteq \mathrm{span}(\boldsymbol{\Gamma})$ and (ii) $\boldsymbol{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}$. This implies that $(\mathbf{Y}, \boldsymbol{\Gamma}^T\mathbf{X}) \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{X}$ and thus that $\boldsymbol{\Gamma}_0^T\mathbf{X}$ is immaterial to the estimation of $\boldsymbol{\beta}$. The smallest subspace $\mathrm{span}(\boldsymbol{\Gamma})$ for which these conditions hold is the $\boldsymbol{\Sigma}$-envelope of $\mathrm{span}(\boldsymbol{\beta})$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$.

To determine the FG estimator of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$, we let $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}}$ in the objective function $J_n(\boldsymbol{\Gamma})$ to reproduce the likelihood-based objective function in Cook et al. (2010). Then the maximum likelihood envelope estimators are $\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}$, where $\widehat{\boldsymbol{\Gamma}} = \arg\min J_n(\boldsymbol{\Gamma})$. Assuming normality for $\boldsymbol{\varepsilon}_i$, Cook et al. (2010) showed that the asymptotic variance of the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ is no larger than that of the usual least squares estimator $\widehat{\boldsymbol{\beta}}$. Under the weaker condition that $\boldsymbol{\varepsilon}_i$ are independent and identically distributed with finite fourth moments, the sample covariance matrices $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ are $\sqrt{n}$-consistent for $\mathbf{M} = \boldsymbol{\Sigma}$ and $\mathbf{U} = \boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma} = \boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\beta}^T$. By Proposition 3, we have $\sqrt{n}$-consistency of the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ under this weaker condition.

3.2.2 Partial envelopes

Su and Cook (2011) used the $\boldsymbol{\Sigma}$-envelope of $\mathrm{span}(\boldsymbol{\beta}_1)$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}_1)$, to develop a partial envelope estimator of $\boldsymbol{\beta}_1$ in the partitioned multivariate linear regression

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}_1\mathbf{X}_{1i} + \boldsymbol{\beta}_2\mathbf{X}_{2i} + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n, \tag{3.4}$$

where $\boldsymbol{\beta}_1 \in \mathbb{R}^{r\times p_1}$, $p_1 \le p$, is the parameter vector of interest, $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and the remaining terms are as defined for model (3.3). In this formulation, the immaterial information is $\boldsymbol{\Gamma}_0^T\mathbf{Y}$, where $\boldsymbol{\Gamma}_0$ is a basis for $\mathcal{E}_{\boldsymbol{\Sigma}}^{\perp}(\boldsymbol{\beta}_1)$. Since $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}_1) \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$, the partial envelope estimator $\widehat{\boldsymbol{\beta}}_{1,\mathrm{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\beta}}_1$ has the potential to yield efficiency gains beyond those for the full envelope, particularly when $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) = \mathbb{R}^r$ so the full envelope offers no gain. In the maximum likelihood estimation of $\boldsymbol{\Gamma}$, the same forms of $\widehat{\mathbf{M}}$, $\widehat{\mathbf{U}}$ and $J_n(\boldsymbol{\Gamma})$ are used for partial envelopes $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}_1)$, except the roles of $\mathbf{Y}$ and $\mathbf{X}$ in the usual response envelopes are replaced with the residuals: $\mathbf{R}_{\mathbf{Y}|\mathbf{X}_2}$, residuals from the linear fits of $\mathbf{Y}$ on $\mathbf{X}_2$, and $\mathbf{R}_{\mathbf{X}_1|\mathbf{X}_2}$, the residuals of $\mathbf{X}_1$ on $\mathbf{X}_2$. Setting $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{R}_{\mathbf{Y}|\mathbf{X}_2}|\mathbf{R}_{\mathbf{X}_1|\mathbf{X}_2}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}_2}$ in the objective function $J_n(\boldsymbol{\Gamma})$ reproduces the likelihood objective function of Su and Cook. Again, Proposition 3 gives $\sqrt{n}$-consistency without normality.

3.2.3 Predictor envelopes

Cook, et al. (2013) studied predictor reduction in model (3.3), except the predictors are now stochastic with $\mathrm{var}(\mathbf{X}) = \boldsymbol{\Sigma}_{\mathbf{X}}$ and $(\mathbf{Y}, \mathbf{X})$ was assumed to be normally distributed for the construction of maximum likelihood estimators. Their reasoning, which parallels that for response envelopes, led them to parameterize the linear model in terms of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ and to achieve similar substantial gains in the estimation of $\boldsymbol{\beta}$ and in prediction. The immaterial information in this setting is given by $\boldsymbol{\Gamma}_0^T\mathbf{X}$, where $\boldsymbol{\Gamma}_0$ is now a basis for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}^{\perp}(\boldsymbol{\beta}^T)$. They also showed that the SIMPLS algorithm for partial least squares provides a $\sqrt{n}$-consistent estimator of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ and demonstrated that the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = \widehat{\boldsymbol{\beta}}\mathbf{P}^T_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{\mathbf{X}})}$ typically outperforms the SIMPLS estimator in practice. For predictor reduction in model (3.3), the envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ is esti-

mated with $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{X}|\mathbf{Y}}$, $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{X}}$. As with response and partial envelopes, Proposition 3 gives us $\sqrt{n}$-consistency without requiring normality for $(\mathbf{Y}, \mathbf{X})$.

Techniques for estimating the dimension of an envelope are discussed in the parent articles of these methods, including use of an information criterion like BIC, cross validation or a hold-out sample.

3.3 New envelope estimators inspired by the objective function

The objective function $J_n(\mathbf{\Gamma})$ can also be used for envelope estimation in new problems. For example, to estimate the multivariate mean $\boldsymbol{\mu} \in \mathbb{R}^r$ in the model $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, we can use the $\mathbf{\Sigma}$-envelope of $\mathrm{span}(\boldsymbol{\mu})$ by taking $\mathbf{M} = \mathbf{\Sigma}$ and $\mathbf{U} = \boldsymbol{\mu}\boldsymbol{\mu}^T$, whose sample versions are: $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}}$, $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T$ and $\widehat{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^n \mathbf{Y}_i$. Then substituting $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ leads to the same objective function $J_n(\mathbf{\Gamma})$ as that obtained when deriving the likelihood-based envelope estimator from scratch.

For the second example, let $\mathbf{Y}_i \sim N_r(\boldsymbol{\mu}, \mathbf{\Sigma})$, $i = 1, \ldots, n$, consist of longitudinal measurements of n subjects over r fixed time points. Suppose we are not interested in the overall mean $\bar{\mu} = \mathbf{1}_r^T\boldsymbol{\mu}/r \in \mathbb{R}^1$ but rather interest centers on the deviations at each time point $\boldsymbol{\alpha} = \boldsymbol{\mu} - \bar{\mu}\mathbf{1}_r \in \mathbb{R}^r$. Let $\mathbf{Q_1} = \mathbf{I}_r - \mathbf{1}_r\mathbf{1}_r^T/r$ denote the projection onto the orthogonal complement of $\mathrm{span}(\mathbf{1}_r)$. Then $\boldsymbol{\alpha} = \mathbf{Q_1}\boldsymbol{\mu}$ and we consider estimating the constrained envelope: $\mathcal{E}_{\mathbf{Q_1}\mathbf{\Sigma}\mathbf{Q_1}}(\mathbf{Q_1}\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{Q_1}) := \mathcal{E}_{\mathbf{M}}(\mathbf{U})$. Optimizing $J_n(\mathbf{\Gamma})$ with $\widehat{\mathbf{M}} = \mathbf{Q_1}\mathbf{S}_{\mathbf{Y}}\mathbf{Q_1}$ and $\widehat{\mathbf{U}} = \mathbf{Q_1}\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\mathbf{Q_1}$ will again lead to the maximum likelihood estimator and to $\sqrt{n}$-consistency without normality. Later from Proposition 4, we will see that $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) = \mathbf{Q_1}\mathcal{E}_{\mathbf{\Sigma}}(\boldsymbol{\mu}\boldsymbol{\mu}^T)$ and the optimization can be simplified.

Proper choices for $\mathbf{M}$ and $\mathbf{U}$ are important for the success of any envelope method. So far these choices have been guided by likelihood considerations, but other methods must be developed when a likelihood is not available. For instance, we have had considerable success with the following idea. Suppose that an estimator $\widehat{\boldsymbol{\theta}}$ of a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^m$ is asymptotically normal, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(0, \mathbf{V}(\boldsymbol{\theta}))$. Then, to improve $\widehat{\boldsymbol{\theta}}$, the FG estimator can be applied with

$\widehat{\mathbf{U}} = \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^T$ and $\widehat{\mathbf{M}}$ equal to a $\sqrt{n}$-consistent estimator of $\mathbf{V}(\boldsymbol{\theta})$. Justifications and asymptotic results for this and other methods will be reported elsewhere, including applications to generalized linear models. For instance, it can be shown that this approach with its extension to matrix valued parameters reproduces all of the envelope estimators discussed in Section 3. We propose a 1D algorithm in the next section that is applicable to the previous models in Section 3.2 and also to future envelope methods.

4 A 1D algorithm

In this section we propose a method for estimating a basis $\boldsymbol{\Gamma}$ of an arbitrary envelope $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) \subseteq \mathbb{R}^d$ based on a series of one-dimensional optimizations. The resulting algorithm is fast and stable, does not require carefully chosen starting values and the estimator it produces converges at the root-n rate. The estimator can be used as it stands, or as a $\sqrt{n}$-consistent starting value for (3.2). In the latter case, one Newton-Raphson step from the starting value provides an estimator that is asymptotically equivalent under normality to the maximum likelihood estimators discussed in Section 3.2 (Lehmann and Casella, 1998, p. 454.)

The population algorithm described in this section extracts one dimension at a time from $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) = \mathcal{E}_{\mathbf{M}}(\mathbf{U})$ until a basis is obtained. It requires only $\mathbf{M} > 0$, $\mathbf{U} \geq 0$ and $u = \dim(\mathcal{E}_{\mathbf{M}}(\mathcal{B}))$ as previously defined in Section 3. Sample versions are obtained by substituting $\sqrt{n}$-consistent estimators $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ for $\mathbf{M}$ and $\mathbf{U}$. Otherwise, the algorithm itself does not depend on a statistical context, although the manner in which the estimated basis is used subsequently does.

The following proposition is the basis for a sequential breakdown of a u-dimensional FG optimization (see also Cook and Zhang (2014; Lemma 5)).

Proposition 4. *Let* $(\mathbf{B}, \mathbf{B}_0)$ *denote an orthogonal basis of* $\mathbb{R}^d$*, where* $\mathbf{B} \in \mathbb{R}^{d\times q}$*,* $\mathbf{B}_0 \in \mathbb{R}^{d\times(d-q)}$ *and* $\operatorname{span}(\mathbf{B}) \subseteq \mathcal{E}_{\mathbf{M}}(\mathcal{B})$*. Then* $\mathbf{v} \in \mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{B})$ *implies that* $\mathbf{B}_0\mathbf{v} \in \mathcal{E}_{\mathbf{M}}(\mathcal{B})$*.*

Suppose we know an orthogonal basis $\mathbf{B}$ for a subspace of the envelope $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$. Then by

Algorithm 1 The 1D algorithm.

1. Set initial value $\mathbf{g}_0 = \mathbf{G}_0 = 0$.
2. For $k = 0, \ldots, u-1$,
   (a) Let $\mathbf{G}_k = (\mathbf{g}_1, \ldots, \mathbf{g}_k)$ if $k \geq 1$ and let $(\mathbf{G}_k, \mathbf{G}_{0k})$ be an orthogonal basis for $\mathbb{R}^d$.
   (b) Define the stepwise objective function

$$D_k(\mathbf{w}) = \log(\mathbf{w}^T \mathbf{M}_k \mathbf{w}) + \log\{\mathbf{w}^T (\mathbf{M}_k + \mathbf{U}_k)^{-1} \mathbf{w}\}, \tag{4.1}$$

   where $\mathbf{M}_k = \mathbf{G}_{0k}^T \mathbf{M} \mathbf{G}_{0k}$, $\mathbf{U}_k = \mathbf{G}_{0k}^T \mathbf{U} \mathbf{G}_{0k}$ and $\mathbf{w} \in \mathbb{R}^{d-k}$.
   (c) Solve $\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}} D_k(\mathbf{w})$ subject to a length constraint $\mathbf{w}^T \mathbf{w} = 1$.
   (d) Define $\mathbf{g}_{k+1} = \mathbf{G}_{0k} \mathbf{w}_{k+1}$ to be the unit length $(k+1)$-th stepwise direction.

Proposition 4 we can find the rest of $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$ by looking into $\mathcal{E}_{\mathbf{B}_0^T \mathbf{M} \mathbf{B}_0}(\mathbf{B}_0^T \mathcal{B})$, which is a lower dimensional envelope. This then provides a motivation for Algorithm 1, which sequentially constructs vectors $\mathbf{g}_k \in \mathcal{E}_{\mathbf{M}}(\mathcal{B})$, $k = 1, \ldots, u$, until a basis is obtained, $\mathrm{span}(\mathbf{g}_1, \ldots, \mathbf{g}_u) = \mathcal{E}_{\mathbf{M}}(\mathcal{B})$. This algorithm follows the structure implied by Proposition 4 and the stepwise objective functions J_k are each one-dimensional versions of (3.1). The first direction $\mathbf{g}_1$ requires optimization in $\mathbb{R}^d$, while the optimization dimension is reduced by 1 in each subsequent step.

Remark 1. At step 2(c) of Algorithm 1, we need to minimize the stepwise objective function $D_k(\mathbf{w})$ under the constraint that $\mathbf{w}^T \mathbf{w} = 1$. The `sg_min` package can still be used to deal with this constraint since we are optimizing over one-dimensional Grassmann manifolds. An alternative way is to integrate the constraint $\mathbf{w}^T \mathbf{w} = 1$ into the objective function in (4.1), so that we only need to minimize the unconstrained function

$$\widetilde{D}_k(\mathbf{w}) = \log(\mathbf{w}^T \mathbf{M}_k \mathbf{w}) + \log\{\mathbf{w}^T (\mathbf{M}_k + \mathbf{U}_k)^{-1} \mathbf{w}\} - 2\log(\mathbf{w}^T \mathbf{w}), \tag{4.2}$$

with an additional normalization step for its minimizer $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_{k+1} / ||\mathbf{w}_{k+1}||$. This unconstrained objective function $D_k(\mathbf{w})$ can be solved by any standard numerical methods such as conjugate gradient or Newton's method. We have implemented this idea with the general purpose optimization function `optim` in R and obtained good results.

Remark 2. We have also considered other types of sequential optimization methods for envelope estimation. For example, we considered minimizing $D_1(\mathbf{w})$ at each step under orthogonality constraints such as $\mathbf{w}_{k+1}^T\mathbf{w}_j = 0$ or $\mathbf{w}_{k+1}^T\mathbf{M}\mathbf{w}_j = 0$ for $j \leq k$. These types of orthogonality constraints are used widely in PLS algorithms and principal components analysis. We find that the statistical properties of these sequential methods are inferior to those of the 1D algorithm. For instance, they are clearly inferior in simulations and we doubt that they lead to consistent estimators.

The next two propositions establish the Fisher consistency of Algorithm 1 in the population and the $\sqrt{n}$-consistency of its sample version.

Proposition 5. *Assume that* $\mathbf{M} > 0$*, and let* $\mathbf{G}_u$ *denote the end result of the algorithm. Then* $\mathrm{span}(\mathbf{G}_u) = \mathcal{E}_{\mathbf{M}}(\mathcal{B})$.

Proposition 6. *Assume that* $\mathbf{M} > 0$ *and let* $\widehat{\mathbf{M}} > 0$ *and* $\widehat{\mathbf{U}}$ *denote* $\sqrt{n}$*-consistent estimators for* $\mathbf{M}$ *and* $\mathbf{U}$*. Let* $\widehat{\mathbf{G}}_u$ *denote the estimator obtained from the 1D algorithm using* $\widehat{\mathbf{M}}$ *and* $\widehat{\mathbf{U}}$ *instead of* $\mathbf{M}$ *and* $\mathbf{U}$*. Then* $\mathbf{P}_{\widehat{\mathbf{G}}_u}$ *is* $\sqrt{n}$*-consistent for the projection onto* $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$.

The algorithm discussed in this section can be used straightforwardly in the contexts of the three envelopes reviewed in Section 3.2 and the extensions sketched in Section 3.3. The statistical properties of the 1D algorithm estimator stated in Propositions 5 and 6 are exactly parallel to the properties of FG optimization in Propositions 2 and 3.

5 Simulations

In this section, we compare the 1D algorithm to FG (full Grassmann manifold) optimization, focusing on computational cost. For fair comparisons, the implementation of our 1D algorithm was based on minimizing the length-constrained objective function (4.1) using the `sg_min` package. Implementation of the 1D algorithm with other computing packages using the unconstrained objective function (4.2) may offer even faster estimation procedures.

5.1 Simulations

We considered the response envelope model in Cook et al. (2010) with univariate predictor $X \sim N(0,1)$ and multivariate response $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}X + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_r(0, \boldsymbol{\Sigma})$, and we were interested in estimation of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$. We generated $\mathbf{M} = \boldsymbol{\Sigma}$ and $\mathbf{U} = \boldsymbol{\beta}\boldsymbol{\beta}^T$ in accordance with an envelope structure: $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ for some positive definite matrices $\boldsymbol{\Omega} \in \mathbb{S}^u$ and $\boldsymbol{\Omega}_0 \in \mathbb{S}^{r-u}$ and a vector of ones $\boldsymbol{\eta} = \mathbf{1}_u \in \mathbb{R}^u$. The semi-orthogonal basis $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ for $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ was randomly generated and $\boldsymbol{\Gamma}_0$ was then obtained so that $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ was an orthogonal basis for $\mathbb{R}^r$. The two covariance matrices $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$ were generated as $\mathbf{A}\mathbf{A}^T > 0$, where $\mathbf{A}$ was a square matrix with corresponding dimensions and was filled with uniform $(0,1)$ random numbers.

We first examined the performances of our 1D algorithm in the population. We generated 100 pairs of $\mathbf{M}$ and $\mathbf{U}$ for each of three dimension configurations, $(r,u) = (10,3)$, $(r,u) = (30,10)$ and $(r,u) = (70,20)$. These dimensions correspond to the real optimization dimensions $u(r-u) = 21$, 200 and 1000 for FG optimization, while the 1D algorithm optimizes over at most $r-1$ real dimensions at each iteration. We recorded the CPU time T for estimating an envelope and the Frobenius norm between the true envelope and an estimated envelope defined as $\text{dist}(\boldsymbol{\Gamma}, \widetilde{\boldsymbol{\Gamma}}) = ||\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T - \widetilde{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{\Gamma}}^T||_F$. The results for running the 1D algorithm (Algorithm 1) and the FG optimization of (3.2) are given in the first three rows of Table 2. Apparently the 1D algorithm achieved the same accuracy as FG optimization and was much less time-consuming, especially at the large dimension $(r,u) = (30,10)$ and $(r,u) = (70,20)$.

We next generated 100 replicated data sets for one pairs of $\mathbf{M}$ and $\mathbf{U}$, and used the sample estimator $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|X}$ and $\widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}} - \mathbf{S}_{\mathbf{Y}|X}$ for envelope estimation. We let $n = 400$ and kept the same dimensions. From Table 2, we can see the 1D algorithm outperformed FG optimization in terms of computational efficiency.

For FG optimization, we chose initial values according to the approach described in Su and Cook (2011; Section 3.5), first optimizing the objective function over the $2r$ eigenvectors of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. This initial value search procedure alone could be computationally costly, but

	1D algorithm		FG optimization	
	T	$\text{dist}(\mathbf{\Gamma}, \widehat{\mathbf{\Gamma}})$	T	$\text{dist}(\mathbf{\Gamma}, \widehat{\mathbf{\Gamma}})$
$(n, r, u) = (\infty, 10, 3)$	2.0 (0.2)	$< 1.0 \times 10^{-8}$	6.6 (0.3)	$< 1.0 \times 10^{-8}$
$(n, r, u) = (\infty, 30, 10)$	2.6 (0.1)	$< 1.0 \times 10^{-4}$	127 (11)	$< 1.0 \times 10^{-4}$
$(n, r, u) = (\infty, 70, 20)$	447 (11)	$< 1.0 \times 10^{-2}$	5084 (1283)	$< 1.0 \times 10^{-2}$
$(n, r, u) = (400, 10, 3)$	0.6 (0.04)	1.1 (0.05)	1.2 (0.09)	1.0 (0.05)
$(n, r, u) = (400, 30, 10)$	30.7 (0.6)	2.8 (0.02)	121 (7)	3.1 (0.02)
$(n, r, u) = (400, 70, 20)$	534 (5)	4.6 (0.04)	4187 (68)	4.7 (0.03)

Table 2: Comparisons between the 1D algorithm and FG optimization: T and $\text{dist}(\mathbf{\Gamma}, \widehat{\mathbf{\Gamma}})$ denote the average running time in seconds and the average angle between the true and estimated envelopes, each over 100 simulations. Standard error are given in parentheses. The population algorithms with $\mathbf{M}$ and $\mathbf{U}$ were indicated with $n = \infty$ and the sample algorithms had $n = 400$.

we did not include the time spent on this when we summarized the computing time T for the FG optimization algorithm in Table 2. Additionally, we used only the true value of u in each simulation. The performance of optimizations at other than the true value of u, as necessary in the application of BIC, need not follow those of Table 2, as we illustrate in the next section.

5.2 Starting values

As mentioned previously, good starting values can be crucial to the performance of FG optimization. To highlight this point, we used the meat data analyzed previously by Cook et al. (2013) for envelope predictor reduction in multivariate linear regression. This data set consists of spectral measurements from infrared transmittance for fat, protein and water for 103 meat samples. Following Cook et al. (2013), we used the protein percentage as the univariate response. The $p = 50$ predictors were spectral measurements at every fourth wavelength between 850nm and 1050nm. Using five-fold cross-validation prediction error as their criterion and u varying from 1 to 25, Cook et al. (2013) compared the maximum likelihood envelope estimator to the OLS and SIMPLS estimators. In their context, the maximum likelihood and FG estimators are the same, as described in Section 3.2.3, so we refer to it as the FG estimator. The starting value for their FG envelope estimator was the SIMPLS estimator, which is $\sqrt{n}$-

consistent in the context of predictor envelopes and had better performance than OLS. SIMPLS was designed specifically for predictor reduction and is not applicable to response or partial reduction or to the extensions discussed in Section 3.3. In this study we used the same setup as Cook et al. (2013), except we focused on comparisons between the 1D algorithm and the FG envelope estimator with starting values again chosen following the approach described in Su and Cook (2011; Section 3.5), since the $2r$ eigenvectors of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$ may be all that is generally available without recourse to the 1D algorithm.

We plotted in Figure 5.1 (top two plots) the five-fold cross-validation squared prediction error and the elapsed CPU time (in seconds) for computing the FG envelope estimators with dimensions $u = 1, \ldots, 25$. Although we had five-folds and thus estimated five envelopes for each dimension u, the time reported is the average for estimating one envelope. The number of real optimization dimensions $u(50 - u)$ varied between 49 and 625. For the larger values of u, FG optimization took a very long time to compute, so we capped the number of allowed iterations at 5000. For small dimensions, $u \leq 3$, FG optimization and the 1D algorithm had close prediction performance, and there were no convergence issues. For $u = 4$ and 5, FG optimizations tended to become trapped into local minima, as indicated by the prediction error. For larger dimensions, $u > 10$, FG optimization began bumping into the iteration limit. The computation time for the 1D method was almost linearly increasing in u because of the sequential manner of the algorithm. With increasing number of components, the prediction errors of both methods converged towards that of the ordinary least squares estimator as expected, since they both reduce to ordinary least squares when $u = 50$. However, the 1D algorithm provided better estimators, consistently over u, than the OLS estimator and the FG envelope estimator.

This difference in the results reported by Cook et al. (2013) and the results shown in the top plot of Figure 5.1 arises because of the different staring values. In Cook et al. (2013), the SIMPLS starting values were $\sqrt{n}$-consistent, while here we chose initial values from the eigenvectors of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. When using these starting values, FG optimizations tended to get trapped by local minima that were close to the initial values, which accounts for the inferior

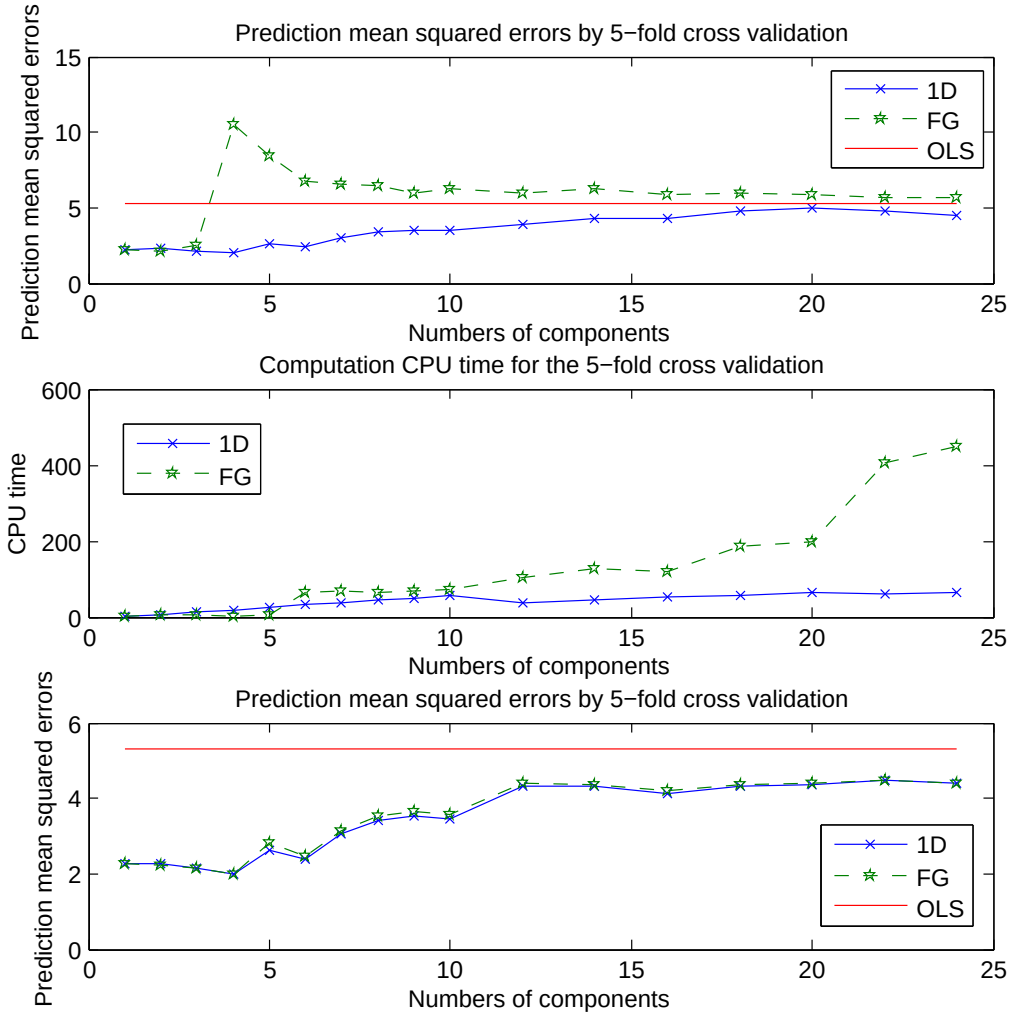

Figure 5.1: Meat protein data. The FG optimizations shown in the top and the middle plots were based on starting values suggested by Cook and Su (2011; Section 5.3). The FG optimization in the bottom plot used the 1D algorithm for starting values.

performance of the FG envelope estimator in this setting. From Lehmann and Casella (1998; Theorem 4.3), we know that one Newton-Raphson iteration from any $\sqrt{n}$-consistent estimator, the 1D algorithm estimator for instance, will be asymptotically equivalent to the MLE, even if there were local minima. We used 100 iterations (instead of one) for the FG optimization with 1D algorithm estimators as initial values. The cross-validation prediction errors, shown in the bottom plot of Figure 5.1, were very close to those of the 1D algorithm. The FG algorithm did a little bit worse than the 1D algorithm at some u because with 100 iterations it occasionally got trapped in a local minimum as it tried to improve the starting value.

6 Conclusion

Our study led to the following conclusions. The FG envelope estimator (3.2) can be computed straightforwardly when the number of real dimensions $u(k-u)$ is relatively small, say less than 150, as illustrated in the example of Section 2.2. When this dimension is large, computing time and local minima can become serious issues, and then root-n consistent starting values become crucial. The 1D algorithm can be used confidently for starting values, or as a stand-alone algorithm for envelope estimation.

Acknowledgments

Research for this article was supported in part by grant DMS-1007547 from the National Science Foundation.

Supplementary materials

Additional proofs: We provided an online supplementary file that included additional proofs for Propositions 3, the $\sqrt{n}$-consistency of the generic objective function (3.2), for Proposition 6, the $\sqrt{n}$-consistency of the 1D algorithm, and for Proposition 5, the Fisher consistency of the 1D algorithm. ("Supplement.pdf", PDF file)

Matlab codes: Matlab codes containing basic functions for Grassmann manifold optimizations and codes for fitting various envelope models with both full Grassmannian (FG) optimization and 1D algorithm. ("EnvelopeModels_1DAlgorithm.zip", zipped files)

References

[1] ABSIL, P. A., MAHONY, R., AND SEPULCHRER, R. (2008), Optimization Algorithms on Matrix Manifolds. *Princeton University Press*.

[2] AMEMIYA, T. (1985), Advanced Econometrics, *Harvard University Press*.

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

[4] COOK, R.D., HELLAND, I.S. AND SU, Z. (2013), Envelopes and partial least squares regression. *JRSS-B*, **75**,851–877.

[5] COOK, R.D., LI, B. AND CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica*, **20**,927–1010.

[6] COOK, R.D. AND ZHANG, X. (2014). Simultaneous envelopes for multivariate linear regression. *Technometrics.* DOI:10.1080/00401706.2013.872700

[7] DE JONG, S. (1993), SIMPLS: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab. Syst.*,**18**, 251–26.

[8] KENWARD, M. G. (1987), A method for comparing profiles of repeated measurements. *JRSS-C*, **36**, 296–308.

[9] LEHMANN, E. L. AND CASELLA, G. (1998). Theory of Point Estimation. Second edition. *Springer, New York.*

[10] SEBER, G.A.F. (2008), A matrix handbook for statisticians, *Wiley-Interscience*.

[11] SU, Z. AND COOK, R.D. (2011), Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, **98**, 133–146.

A Appendix: Proofs and Technical Details

A.1 Proposition 2

The proof of this proposition is very similar to the proof of Proposition 4.2 in Cook et al. (2013), thus is omitted.

A.2 Proposition 3

The proof follows in the same way as the proof of Proposition 6 in Section A.2. Thus we omit the details.

A.3 Proposition 4

Proof. From our set-up, we know that $\mathbf{B}^T\mathbf{M}\mathbf{B} > 0$ thus $\mathcal{E}_{\mathbf{B}^T\mathbf{M}\mathbf{B}}(\mathbf{B}^T\mathcal{B})$ exists. Let $\mathbf{\Gamma}$ be a basis of $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$, and $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ be a orthogonal basis of $\mathbb{R}^d$, then $\mathbf{M} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$ and $\mathcal{B} \subseteq \text{span}(\mathbf{\Gamma})$ for some symmetric matrices $\mathbf{\Omega} > 0$ and $\mathbf{\Omega}_0 > 0$. Therefore,

$$\begin{aligned} \mathbf{B}_0^T\mathbf{M}\mathbf{B}_0 &= (\mathbf{B}_0^T\mathbf{\Gamma})\mathbf{\Omega}(\mathbf{B}_0^T\mathbf{\Gamma})^T + (\mathbf{B}_0^T\mathbf{\Gamma}_0)\mathbf{\Omega}_0(\mathbf{B}_0^T\mathbf{\Gamma}_0)^T \\ \mathbf{B}_0^T\mathcal{B} &\subseteq \text{span}(\mathbf{B}_0^T\mathbf{\Gamma}), \end{aligned} \tag{A1}$$

where $\text{span}(\mathbf{B}_0^T\mathbf{\Gamma})$ is the orthogonal compliment of $\text{span}(\mathbf{B}_0^T\mathbf{\Gamma}_0)$ in $\mathbb{R}^{d-q}$ since $\text{span}(\mathbf{B}) \subseteq \text{span}(\mathbf{\Gamma})$. Then we see that

$$\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0 = \mathbf{P}_{\mathbf{B}_0^T\mathbf{\Gamma}}\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0\mathbf{P}_{\mathbf{B}_0^T\mathbf{\Gamma}} + \mathbf{Q}_{\mathbf{B}_0^T\mathbf{\Gamma}}\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0\mathbf{Q}_{\mathbf{B}_0^T\mathbf{\Gamma}}, \tag{A2}$$

which implies that $\text{span}(\mathbf{B}_0^T\mathbf{\Gamma})$ is a reducing subspace of $\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0$ which also contains $\mathbf{B}_0^T\mathcal{B}$ by (A1). By definition, we know that $\mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{B})$ is the smallest reducing subspace of $\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0$ that contains $\mathbf{B}_0^T\mathcal{B}$. Hence $\mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{B}) \subseteq \text{span}(\mathbf{B}_0^T\mathbf{\Gamma})$. Thus $\mathbf{v} \in \mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{B})$ implies $\mathbf{B}_0\mathbf{v} \in \mathcal{E}_{\mathbf{M}}(\mathcal{B})$.

□

A.4 Proposition 5

Proof. We first write

$$\begin{aligned} \mathbf{M} &= \mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T, \\ \mathbf{M} + \mathbf{U} &= \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T, \end{aligned}$$

where $\mathbf{\Omega}_0 > 0$, $\mathbf{\Omega} > 0$, $\mathbf{\Phi} > 0$, $\mathbf{\Omega} - \mathbf{\Phi} \geq 0$, $\mathbf{\Gamma}$ is semi-orthogonal basis for $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$ and $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^d$ is orthogonal basis for $\mathbb{R}^d$.

We begin by considering optimization for the first direction $\mathbf{g}_1 = \arg\min_{\mathbf{g}\in\mathbb{R}^d} J_0(\mathbf{g})$, where $J_0(\mathbf{g}) = \log|\mathbf{g}^T\mathbf{M}\mathbf{g}| + \log|\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}|$ and the minimization is subject to the constraint $\mathbf{g}^T\mathbf{g} = 1$. Let $\mathbf{g} = \mathbf{\Gamma}\mathbf{h} + \mathbf{\Gamma}_0\mathbf{h}_0$ for some $\mathbf{h} \in \mathbb{R}^u$ and $\mathbf{h}_0 \in \mathbb{R}^{(d-u)}$. Consider the optimization problem as the unconstrained problem,

$$\mathbf{g}_1 = \arg\min_{\mathbf{g}\in\mathbb{R}^d} \left\{\log|\mathbf{g}^T\mathbf{M}\mathbf{g}| + \log|\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}| - 2\log|\mathbf{g}^T\mathbf{g}|\right\}.$$

Then we will have the same solution as the original problem up to an arbitrary scaling constant. Next, we plug-in these expressions for $\mathbf{g}$, $\mathbf{M}+\mathbf{U}$ and $\mathbf{M}$,

$$\begin{aligned}
& \log|\mathbf{g}^T\mathbf{M}\mathbf{g}| + \log|\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}| - 2\log|\mathbf{g}^T\mathbf{g}| \\
= \;& \log\{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0\} + \log\{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0\} - 2\log\{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0\} \\
\equiv \;& f(\mathbf{h}, \mathbf{h}_0).
\end{aligned}$$

Taking partial derivative with respect to $\mathbf{h}_0$, it is shown in the Supplement that $f(\mathbf{h}, \mathbf{h}_0)$ is minimized uniquely at $\mathbf{h}_0 = 0$ and consequently $\mathbf{g}_1$ is in the envelope.

For the $(k+1)$-th direction, $\mathbf{g}_{k+1} = \mathbf{G}_{0k}\mathbf{w}_{k+1}$ where $\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d-k}} J_k(\mathbf{w})$, subject to $\mathbf{w}^T\mathbf{w} = 1$. Because the form of

$$J_k(\mathbf{w}) = \log|\mathbf{w}^T\mathbf{G}_{0k}^T\mathbf{M}\mathbf{G}_{0k}\mathbf{w}| + \log|\mathbf{w}^T\left\{\mathbf{G}_{0k}^T(\mathbf{M}+\mathbf{U})\mathbf{G}_{0k}\right\}^{-1}\mathbf{w}|$$

is the same as that for the first direction, this gives $\mathbf{w}_{k+1} \in \mathcal{E}_{\mathbf{G}_{0k}^T\mathbf{M}\mathbf{G}_{0k}}(\mathbf{G}_{0k}^T\mathcal{B})$. Therefore $\mathbf{g}_{k+1} = \mathbf{G}_{0k}\mathbf{w}_{k+1} \in \mathcal{E}_{\mathbf{M}}(\mathcal{B})$ by Proposition 4.

A.5 Proposition 6

Our proof of $\sqrt{n}$-consistency follows from Amemiya's (1985) results on the asymptotic properties of extremum estimators. We show in the Supplement that the conditions necessary for application of Amemiya's results hold in our context, particularly his Propositions 4.1.1 and 4.1.3.

□

Supplement: Additional Proofs and Technical Details for "Algorithms for Envelope Estimation"

R. Dennis Cook and Xin Zhang

A Proof for $\sqrt{n}$-consistency of the algorithm

A.1 Proposition 5

Proof. Recall that

$$\begin{aligned} & \log|\mathbf{g}^T\mathbf{M}\mathbf{g}| + \log|\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}| - 2\log|\mathbf{g}^T\mathbf{g}| \\ = \;& \log\{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0\} + \log\{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0\} - 2\log\{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0\} \\ \equiv \;& f(\mathbf{h}, \mathbf{h}_0). \end{aligned}$$

Taking partial derivative with respect to $\mathbf{h}_0$, we have

$$\frac{\partial}{\partial \mathbf{h}_0} f(\mathbf{h}, \mathbf{h}_0) = \frac{2\mathbf{\Omega}_0\mathbf{h}_0}{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}_0^{-1}\mathbf{h}_0}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0} - \frac{4\mathbf{h}_0}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}.$$

To get local minimums we need to set $\frac{\partial}{\partial \mathbf{h}_0} f(\mathbf{h}, \mathbf{h}_0) = 0$ which gives the following equality.

$$\left\{ \frac{2\mathbf{\Omega}_0}{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}_0^{-1}}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0} \right\} \mathbf{h}_0 = \left\{ \frac{4}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0} \right\} \mathbf{h}_0.$$

Define

$$\mathbf{A}_0 = \left\{ \frac{2\mathbf{\Omega}_0}{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}_0^{-1}}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0} \right\} / \left\{ \frac{4}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0} \right\}.$$

Since $\mathbf{\Omega}_0 > 0$, we know $\mathbf{A}_0 > 0$. Then $\mathbf{A}_0\mathbf{h}_0 = \mathbf{h}_0$ has solutions only as eigenvectors of $\mathbf{A}_0$. The eigenvectors of $\mathbf{A}_0$ are the same as those of $\mathbf{\Omega}_0$. Hence, $\mathbf{h}_0$ equals 0 or any eigenvector $\ell_k(\mathbf{\Omega}_0)$ of $\mathbf{\Omega}_0$. Therefore, the minimum value of $f(\mathbf{h}, \mathbf{h}_0)$ has to be obtained by 0 or $\ell_k(\mathbf{\Omega}_0)$ (since $\mathbf{h}_0 = \infty$ can be easily eliminated). If $\mathbf{h}_0 = 0$ then our conclusion follows.

Assume $\mathbf{h}_0 \neq 0$ and $\mathbf{\Omega}_0\mathbf{h}_0 = \lambda_k\mathbf{h}_0$. Then,

$$\begin{aligned} f(\mathbf{h}, \mathbf{h}_0) &= \log\{\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \lambda_k\mathbf{h}_0^T\mathbf{h}_0}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}\} + \log\{\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \frac{1}{\lambda_k}\mathbf{h}_0^T\mathbf{h}_0}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}\} \\ &= \log\{\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}W_h + \lambda_k(1-W_h)\} + \log\{\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}W_h + \frac{1}{\lambda_k}(1-W_h)\}, \end{aligned}$$

where $W_h = \frac{\mathbf{h}^T\mathbf{h}}{\mathbf{h}^T\mathbf{h}+\mathbf{h}_0^T\mathbf{h}_0}$ is the weight between 0 and 1. Because $\log()$ is concave, we have $\log(aW_h + b(1-W_h)) \geq W_h\log(a) + (1-W_h)\log(b)$. Hence,

$$\begin{aligned} f(\mathbf{h}, \mathbf{h}_0) &\geq W_h\left\{\log\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}\right\} + (1-W_h)\left\{\log(\lambda_k) + \log(\frac{1}{\lambda_k})\right\} \\ &= W_h\left\{\log\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}\right\} \\ &\geq W_h \cdot \min_{\mathbf{h}\in\mathbb{R}^d}\left\{\log\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}\right\} \\ &\geq \min_{\mathbf{h}\in\mathbb{R}^d}\left\{\log\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}\right\}. \end{aligned}$$

The last inequality holds because (see below)

$$\min_{\mathbf{h}\in\mathbb{R}^d}\left\{\log\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}\right\} < 0, \tag{A1}$$

Moreover, the lower bound of $f(\mathbf{h}, \mathbf{h}_0)$, which is negative, will be attained if we let $W_h = 1$ and let $\mathbf{h} = \arg\min_{\mathbf{h}\in\mathbb{R}^d}\left\{\log\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}\right\}$. So we have the minimum found at $W_h = \frac{\mathbf{h}^T\mathbf{h}}{\mathbf{h}^T\mathbf{h}+\mathbf{h}_0^T\mathbf{h}_0} = 1$, or equivalently, $\mathbf{g} = \mathbf{\Gamma}\mathbf{h} \in \text{span}(\mathbf{\Gamma})$.

To show that (A1) holds, we first show that $\min_{\mathbf{h}\in\mathbb{R}^d}\left\{\log\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}\right\} \leq 0$, then we assume the equality to conduct the proof by contradiction. Define the following two functions,

$$\begin{aligned} \mathrm{F}(\mathbf{h}; \mathbf{\Phi}, \mathbf{\Omega}^{-1}) &:= \log\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}, \\ \mathrm{F}(\mathbf{h}; \mathbf{\Omega}, \mathbf{\Omega}^{-1}) &:= \log\frac{\mathbf{h}^T\mathbf{\Omega}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}, \end{aligned}$$

Recall that $\mathbf{\Omega} - \mathbf{\Phi} \geq 0$, hence $\mathrm{F}(\mathbf{h}; \mathbf{\Phi}, \mathbf{\Omega}^{-1}) \leq \mathrm{F}(\mathbf{h}; \mathbf{\Omega}, \mathbf{\Omega}^{-1})$ for any $\mathbf{h}$. Consider the minimum of both $\mathrm{F}(\mathbf{h}; \mathbf{\Phi}, \mathbf{\Omega}^{-1})$ and $\mathrm{F}(\mathbf{h}; \mathbf{\Omega}, \mathbf{\Omega}^{-1})$, we have

$$\min_{\mathbf{h}} \mathrm{F}(\mathbf{h}; \mathbf{\Phi}, \mathbf{\Omega}^{-1}) \leq \min_{\mathbf{h}} \mathrm{F}(\mathbf{h}; \mathbf{\Omega}, \mathbf{\Omega}^{-1}) = 0,$$

where the minimum of the right hand side is zero by taking $\mathbf{h}$ equals to any eigenvector of $\mathbf{\Omega}$.

Now we assume that $\min_{\mathbf{h}} \mathrm{F}(\mathbf{h}; \mathbf{\Phi}, \mathbf{\Omega}^{-1}) = 0$. Then for an arbitrary $\mathbf{h}$,

$$0 \leq \mathrm{F}(\mathbf{h}; \mathbf{\Phi}, \mathbf{\Omega}^{-1}) \leq \mathrm{F}(\mathbf{h}; \mathbf{\Omega}, \mathbf{\Omega}^{-1}).$$

Let $\mathbf{h}_i = \ell_i(\mathbf{\Omega})$, $i = 1, \ldots, u$, be the i-th unit eigenvector of $\mathbf{\Omega}$ and plug $\mathbf{h}_i$ into the above inequalities, we have

$$0 \leq \mathrm{F}(\mathbf{h}_i; \mathbf{\Phi}, \mathbf{\Omega}^{-1}) \leq \mathrm{F}(\mathbf{h}_i; \mathbf{\Omega}, \mathbf{\Omega}^{-1}) = 0,\ i = 1, \ldots, u,$$

which implies

$$0 = \mathrm{F}(\mathbf{h}_i; \mathbf{\Phi}, \mathbf{\Omega}^{-1}) = \mathrm{F}(\mathbf{h}_i; \mathbf{\Omega}, \mathbf{\Omega}^{-1}) = 0,\ i = 1, \ldots, u,$$

and more explicitly,

$$\log(\mathbf{h}_i^T \mathbf{\Phi} \mathbf{h}_i) = \log(\mathbf{h}_i^T \mathbf{\Omega} \mathbf{h}_i),\ i = 1, \ldots, u,$$

which implies $\mathbf{\Phi} = \mathbf{\Omega}$ because that $\mathbf{\Phi},\ \mathbf{\Omega} \in \mathbb{R}^{u \times u}$ and $\mathbf{h}_i$, $i = 1, \ldots, u$, are u linear independent vectors. Then by definition $\mathbf{U} = \mathbf{\Gamma}^T(\mathbf{\Phi} - \mathbf{\Omega})\mathbf{\Gamma} = 0$ leads to contradiction with the dimension of the envelope. □

A.2 Proposition 6

Proof. Our proof of $\sqrt{n}$-consistency hinges on Amemiya's (1985) results on the asymptotic properties of extremum estimators. Proposition 4.1.1 and Proposition 4.1.3 in Amemiya (1985) can be applied to our context. We first state these results and then sketch how they can be used to prove the $\sqrt{n}$-consistency for our algorithm.

Let $Q_n(\mathbf{y}, \boldsymbol{\theta})$ be a real-valued function of the random variables $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ and the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)^T$. We shall sometimes write $Q_n(\mathbf{y}, \boldsymbol{\theta})$ more compactly as $Q_n(\boldsymbol{\theta})$. Let the parameter space be $\boldsymbol{\Theta}$ and let the true value of $\boldsymbol{\theta}$ be $\boldsymbol{\theta}_t$ which is in $\boldsymbol{\Theta}$. Then Proposition 4.1.1 and Proposition 4.1.3 in Amemiya (1985) give asymptotic properties of the extremum estimator, $\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q_n(\mathbf{y}, \boldsymbol{\theta})$. We summarize the conditions in Amemiya's Propositions as follows.

(A) The parameter space $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^K$;

(B) $Q_n(\mathbf{y}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$; for all $\mathbf{y}$ and is a measurable function of $\mathbf{y}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$;

(C) $n^{-1}Q_n(\boldsymbol{\theta})$ converges to a nonstochastic function $Q(\boldsymbol{\theta})$ in probability uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as n goes to infinity, and $Q(\boldsymbol{\theta})$ attains a unique global maximum at $\boldsymbol{\theta}_t$;

(D) $\partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ exists and is continuous in an open, convex neighborhood of $\boldsymbol{\theta}_0$;

(E) $n^{-1}\left\{\partial^2 Q_n(\mathbf{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T\right\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*}$ converges to a finite nonsingular matrix

$$\mathbf{A}(\boldsymbol{\theta}_t) = \lim_{n\to\infty} \mathrm{E}_{\boldsymbol{\theta}_t}\left\{n^{-1}\left\{\partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T\right\}\right\},$$

for any random sequences $\boldsymbol{\theta}_n^*$ such that $\mathrm{plim}(\boldsymbol{\theta}_n^*) = \boldsymbol{\theta}_t$;

(F) $n^{-1/2}\left\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \to N(0, \mathbf{B}(\boldsymbol{\theta}_t))$, where

$$\mathbf{B}(\boldsymbol{\theta}_t) = \lim_{n\to\infty} \mathrm{E}_{\boldsymbol{\theta}_t}\left\{n^{-1}\left\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right\}\left\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\right\}\right\}.$$

Proposition A1. *Under assumptions (A)-(C), $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_t$ in probability.*

Proposition A2. *Under assumptions (A)-(F), $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_t) \to N(0, \mathbf{A}(\boldsymbol{\theta}_t)^{-1}\mathbf{B}(\boldsymbol{\theta}_t)\mathbf{A}(\boldsymbol{\theta}_t)^{-1})$.*

In our adaptation of Proposition A1 and Proposition A2, we let $\boldsymbol{\theta} \equiv \mathbf{g}$ whose true value is denoted by $\mathbf{g}_t$ and let the random variables $\mathbf{y} = \mathrm{vech}(\widehat{\mathbf{M}}, \widehat{\mathbf{U}})$. The parameter space is the 1D manifold $\boldsymbol{\Theta} = \mathcal{G}_{(d,1)}$ which is a compact subset of $\mathbb{R}^d$, so condition (A) in Proposition A1 is satisfied. The function to be maximized is defined as follows.

$$Q_n(\mathbf{g}) = -n/2\log(\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g}) - n/2\log(\mathbf{g}^T(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1}\mathbf{g}) + n\log(\mathbf{g}^T\mathbf{g}). \tag{A2}$$

Condition (B) then holds. We next verify condition (C) that $n^{-1}Q_n(\mathbf{g})$ converges uniformly to

$$Q(\mathbf{g}) = -1/2\log(\mathbf{g}^T\mathbf{M}\mathbf{g}) - 1/2\log(\mathbf{g}^T(\mathbf{M} + \mathbf{U})^{-1}\mathbf{g}) + \log(\mathbf{g}^T\mathbf{g}). \tag{A3}$$

We have shown that the population objective function $Q(\mathbf{g})$ attains the unique global maximum at $\mathbf{g}_t$. For simplicity, we assume $\mathbf{M}$ and $\mathbf{M}+\mathbf{U}$ both have distinct eigenvalues so that $\mathbf{g}_t$ is the unique maximum of $Q(\mathbf{g})$ in the 1D manifold $\mathbf{\Theta}$. For the case where there are multiple local maxima of $Q(\mathbf{g})$, we can obtain similar results by applying Proposition 4.1.2 in Amemiya (1985) as an alternative of Proposition A1. Since $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ are $\sqrt{n}$-consistent for $\mathbf{M}$ and $\mathbf{U}$, the eigenvectors and eigenvalues of $\widehat{\mathbf{M}}$ and $(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}$ are $\sqrt{n}$-consistent for the eigenvectors and eigenvalues of their population counterparts.

Then $n^{-1}Q_n(\mathbf{g})$ converge in probability to $Q(\mathbf{g})$ uniformly in $\mathbf{g}$, as can be seen from the following argument.

$$\begin{aligned} n^{-1}Q_n(\mathbf{g}) - Q(\mathbf{g}) &= -1/2\left(\log(\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}) - \log(\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g})\right) \\ &\quad -1/2\left(\log(\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g}) - \log(\mathbf{g}^T\mathbf{M}\mathbf{g})\right) \\ &= -1/2\log\left[\frac{\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}}{\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}}\right] - 1/2\log\left[\frac{\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g}}{\mathbf{g}^T\mathbf{M}\mathbf{g}}\right]. \end{aligned}$$

Hence, $\sup_{\mathbf{g}\in\mathbf{\Theta}}\log(\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g}/\mathbf{g}^T\mathbf{M}\mathbf{g}) = \sup_{\mathbf{g}\in\mathbf{\Theta}}\log(\mathbf{g}^T\mathbf{M}^{-1/2}\widehat{\mathbf{M}}\mathbf{M}^{-1/2}\mathbf{g}/\mathbf{g}^T\mathbf{g})$, which equals to the logarithm of the largest eigenvalue of $\mathbf{M}^{-1/2}\widehat{\mathbf{M}}\mathbf{M}^{-1/2}$ and converges to 0 in probability. Similarly, $\sup_{\mathbf{g}\in\mathbf{\Theta}}\log[\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}/\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}]$ converges to zero in probability. Therefore, $n^{-1}Q_n(\mathbf{g})$ converges to $Q(\mathbf{g})$ in probability uniformly in $\mathbf{g}\in\mathbf{\Theta}$. Note that we have assumed $\mathbf{M}+\mathbf{U}>0$ and $\mathbf{M}^{-1}>0$, so their eigenvalues will be bounded away from zero.

We next verify conditions $(D)-(F)$. By straightforward calculation, condition (D) follows from the second derivative matrix

$$\begin{aligned} n^{-1}\frac{\partial^2 Q_n(\mathbf{g})}{\partial\mathbf{g}\partial\mathbf{g}^T} &= 2(\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g})^{-2}(\widehat{\mathbf{M}}\mathbf{g}\mathbf{g}^T\widehat{\mathbf{M}}) - (\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g})^{-1}\widehat{\mathbf{M}} \\ &\quad +2\left[\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}\right]^{-2}\left[(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\right] \\ &\quad -\left[\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}\right]^{-1}(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1} \\ &\quad -2(\mathbf{g}^T\mathbf{g})^{-2}\mathbf{P}_{\mathbf{g}} + (\mathbf{g}^T\mathbf{g})^{-1}\mathbf{I}_p. \end{aligned} \tag{A4}$$

Condition (E) holds because the above quantity is a smooth function of $\mathbf{g}$, $\widehat{\mathbf{M}}$ and $(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}$.

Last, we need to verify condition (F). From the proof of Proposition A2, we need only show that $n^{-1}\left\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = O_p(1/\sqrt{n})$ for $\sqrt{n}$-consistency of the estimator $\widehat{\boldsymbol{\theta}}_n$. The derivative $n^{-1}\left\{\partial Q_n(\mathbf{g})/\partial\mathbf{g}\right\}_{\mathbf{g}=\mathbf{g}_t}$ equals

$$-(\mathbf{g}_t^T\widehat{\mathbf{M}}\mathbf{g}_t)^{-1}\widehat{\mathbf{M}}\mathbf{g}_t - (\mathbf{g}_t^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}_t)^{-1}(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}_t + 2\mathbf{g}_t. \tag{A5}$$

Following the derivation for the population objective function, we know that $\left\{\partial Q(\mathbf{g})/\partial\mathbf{g}\right\}_{\mathbf{g}=\mathbf{g}_t} = 0$. Then the result follows from the fact that $n^{-1}\partial Q_n(\mathbf{g})/\partial\mathbf{g}$ is a smooth function of $\widehat{\mathbf{M}}$ and $(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}$ which are $\sqrt{n}$-consistent estimators.

So far, we have verified the conditions $(A)-(F)$ so that the sample estimator $\widehat{\mathbf{g}}_1$ will be $\sqrt{n}$-consistent for the population estimator. For the $(k+1)$-th direction, $k<u$, let $\widehat{\mathbf{G}}_k$ denote an $\sqrt{n}$-consistent estimator of the first k directions and let $(\widehat{\mathbf{G}}_k, \widehat{\mathbf{G}}_{0k})$ be an orthogonal matrix. The $(k+1)$-th direction is defined by $\mathbf{g}_{k+1} = \widehat{\mathbf{G}}_{0k}\mathbf{w}_{k+1}$ where the parameters are $\mathbf{w}_{k+1}\in\boldsymbol{\Theta}_{k+1}\subset\mathbb{R}^{p-k}$ and the parameter space is $\boldsymbol{\Theta}_{k+1}=\mathcal{G}_{p-k,1}$. We show that we can obtain a $\sqrt{n}$-consistent estimator $\widehat{\mathbf{w}}_{k+1}$, so the $\sqrt{n}$-consistency of $\widehat{\mathbf{g}}_{k+1}=\widehat{\mathbf{G}}_{0k}\widehat{\mathbf{w}}_{k+1}$ then follows. We define our objective functions $Q_n(\mathbf{w})$ and $Q(\mathbf{w})$ as

$$Q_n(\mathbf{w}) = -n/2\log(\mathbf{w}^T(\widehat{\mathbf{G}}_{0k}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})\widehat{\mathbf{G}}_{0k})^{-1}\mathbf{w}) - n/2\log(\mathbf{w}^T\widehat{\mathbf{G}}_{0k}^T\widehat{\mathbf{M}}\widehat{\mathbf{G}}_{0k}\mathbf{w}) + n\log(\mathbf{w}^T\mathbf{w})$$
$$Q(\mathbf{w}) = -1/2\log(\mathbf{w}^T(\mathbf{G}_{0k}^T(\mathbf{M}+\mathbf{U})\mathbf{G}_{0k})^{-1}\mathbf{w}) - 1/2\log(\mathbf{w}^T\mathbf{G}_{0k}^T\mathbf{M}\mathbf{G}_{0k}\mathbf{w}) + \log(\mathbf{w}^T\mathbf{w})$$

Following the same logic as verifying the conditions for the first direction, we can see that $\widehat{\mathbf{w}} = \arg\max Q_n(\mathbf{w})$ will be $\sqrt{n}$-consistent for $\mathbf{v}_t = \arg\max Q(\mathbf{w})$ by noticing that $(\widehat{\mathbf{G}}_{0k}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})\widehat{\mathbf{G}}_{0k})^{-1}$ and $\widehat{\mathbf{G}}_{0k}^T\widehat{\mathbf{M}}\widehat{\mathbf{G}}_{0k}$ are $\sqrt{n}$-consistent estimators for $(\mathbf{G}_{0k}^T(\mathbf{M}+\mathbf{U})\mathbf{G}_{0k})^{-1}$ and $\mathbf{G}_{0k}^T\mathbf{M}\mathbf{G}_{0k}$. Since all the u directions will be $\sqrt{n}$-consistent, the projection onto $\widehat{\mathbf{G}}_u = (\widehat{\mathbf{g}}_1,\ldots,\widehat{\mathbf{g}}_u)$ will be a $\sqrt{n}$-consistent estimator for the projection onto the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$. □

References

[1] AMEMIYA, T. (1985), Advanced Econometrics, *Harvard University Press*.