Foundations for Envelope Models and Methods

R. Dennis Cook* and Xin Zhang†

October 6, 2014

Abstract

Envelopes were recently proposed by Cook, Li and Chiaromonte (2010) as a method for reducing estimative and predictive variations in multivariate linear regression. We extend their formulation, proposing a general definition of an envelope and a general framework for adapting envelope methods to any estimation procedure. We apply the new envelope methods to weighted least squares, generalized linear models and Cox regression. Simulations and illustrative data analysis show the potential for envelope methods to significantly improve standard methods in linear discriminant analysis, logistic regression and Poisson regression.

Key Words: Generalized linear models; Grassmannians; Weighted least squares.

1 Introduction

The overarching goal of envelope models and methods is to increase efficiency in multivariate parameter estimation and prediction. The development has so far been restricted to the multivariate linear model,

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i,\ i = 1, \ldots, n, \tag{1.1}$$

where $\boldsymbol{\epsilon}_i \in \mathbb{R}^r$ is a normal error vector that has mean 0, constant covariance $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} > 0$ and is independent of $\mathbf{X}$, $\boldsymbol{\alpha} \in \mathbb{R}^r$ and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ is the regression coefficient matrix in which we are primarily interested. Efficiency gains in the estimation of $\boldsymbol{\beta}$ are achieved by reparameterizing this model in terms of special projections of the response vector $\mathbf{Y} \in \mathbb{R}^r$ or the predictor vector $\mathbf{X} \in \mathbb{R}^p$. In this article we propose extensions of envelope methodology beyond the linear

*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).

model to quite general multivariate contexts. Since envelopes represent nascent methodology that is likely unfamiliar to many, we outline in Section 1.1 response envelopes as developed by Cook, Li and Chiaromonte (2010). An example is given in Section 1.2 to illustrate their potential advantages in application. A review of the literature on envelopes is provided in Section 1.3 and the specific goals of this article are outlined in Section 1.4.

1.1 Response envelopes

Response envelopes for model (1.1) gain efficiency in the estimation of $\boldsymbol{\beta}$ by incorporating the projection $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ onto the smallest subspace $\mathcal{E} \subseteq \mathbb{R}^r$ with the properties (1) the distribution of $\mathbf{Q}_{\mathcal{E}}\mathbf{Y} \mid \mathbf{X}$ does not depend on the value of the non-stochastic predictor $\mathbf{X}$, where $\mathbf{Q}_{\mathcal{E}} = \mathbf{I}_r - \mathbf{P}_{\mathcal{E}}$, and (2) $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ is independent of $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ given $\mathbf{X}$. These conditions imply that the distribution of $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ is not affected by $\mathbf{X}$ marginally or through an association with $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$. Consequently, changes in the predictor affect the distribution of $\mathbf{Y}$ only via $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$. We refer to $\mathbf{P}_{\mathcal{E}}\mathbf{Y}$ informally as the material part of $\mathbf{Y}$ and to $\mathbf{Q}_{\mathcal{E}}\mathbf{Y}$ as the immaterial part of $\mathbf{Y}$.

The notion of an envelope arises in the formal construction of the response projection $\mathbf{P}_{\mathcal{E}}$ as guided by the following two definitions, which are not restricted to the linear model context. Let $\mathbb{R}^{m\times n}$ be the set of all real $m \times n$ matrices and let $\mathbb{S}^{k\times k}$ be the set of all real and symmetric $k \times k$ matrices. If $\mathbf{A} \in \mathbb{R}^{m\times n}$, then $\mathrm{span}(\mathbf{A}) \subseteq \mathbb{R}^m$ is the subspace spanned by columns of $\mathbf{A}$.

Definition 1. *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 $\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$ reduces $\mathbf{M}$.*

This definition of a reducing subspace is equivalent to that given by Cook, Li and Chiaromonte (2010). It is common in the literature on invariant subspaces and functional analysis (Conway 1990), although the underlying notion of "reduction" differs from the usual understanding in statistics.

Definition 2. *(Cook et al. 2010) Let $\mathbf{M} \in \mathbb{S}^{p\times p}$ 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 $\mathcal{B}$.*

Definition 2 is the formal definition of an envelope, which is central to our developments. We will often identify the subspace $\mathcal{B}$ as the span of a specified matrix $\mathbf{U}$: $\mathcal{B} = \mathrm{span}(\mathbf{U})$. To

avoid proliferation of notation in such cases, we will also use the matrix as the argument to $\mathcal{E}_{\mathbf{M}}$: $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) := \mathcal{E}_{\mathbf{M}}(\mathrm{span}(\mathbf{U}))$.

The response projection $\mathbf{P}_{\mathcal{E}}$ is then defined formally as the projection onto $\mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta})$, which by construction is the smallest reducing subspace of $\mathbf{\Sigma_{Y|X}}$ that contains $\mathrm{span}(\boldsymbol{\beta})$ (or "envelopes" $\mathrm{span}(\boldsymbol{\beta})$ and hence the name "envelope"). Model (1.1) can be parameterized in terms of $\mathbf{P}_{\mathcal{E}}$ by using a basis. Let $u = \dim(\mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta}))$ and let $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix with $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$ and $\mathrm{span}(\mathbf{\Gamma}) = \mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta})$. This leads directly to the envelope version of model (1.1),

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \mathbf{\Gamma}\boldsymbol{\eta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \text{ with } \mathbf{\Sigma_{Y|X}} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T, \ i = 1, \ldots, n, \tag{1.2}$$

where $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ gives the coordinates of $\boldsymbol{\beta}$ relative to basis $\mathbf{\Gamma}$, and $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ are positive definite matrices. While $\boldsymbol{\eta}$, $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ depend on the basis $\mathbf{\Gamma}$ selected to represent $\mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta})$, the parameters of interest $\boldsymbol{\beta}$ and $\mathbf{\Sigma_{Y|X}}$ depend only on $\mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta})$ and not on the basis. All parameters in (1.2) can be estimated by maximizing the likelihood from (1.2) with the envelope dimension u determined by standard methods like likelihood ratio testing, information criteria, cross-validation or a hold-out sample, as described by Cook et al. (2010). In particular, the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ of $\boldsymbol{\beta}$ is just the projection of the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ onto the estimated envelope, $\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = \mathbf{P}_{\widehat{\mathbf{\Gamma}}}\widehat{\boldsymbol{\beta}}$, and $\sqrt{n}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{env}}) - \mathrm{vec}(\boldsymbol{\beta}))$ is asymptotically normal with mean 0 and covariance matrix given by Cook et al. (2010), where vec is the vectorization operator that stacks the columns of a matrix and u is assumed to be known.

1.2 Cattle data and response envelopes

The data for this illustration resulted 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 and their weights (in kilograms) were recorded at weeks 2, 4, ..., 18 and 19 after treatment (Kenward 1987). Because of the nature of a cow's digestive system, the treatments were not expected to have an immediate measurable affect on weight. The objectives of the study were to find if the treatments had differential effects on weight and, if so, about when were they first manifested.

We begin by consider the multivariate linear model (1.1), where $\mathbf{Y}_i \in \mathbb{R}^{10}$ is the vector of cattle weights from week 2 to week 19, and the binary predictor X_i is either 0 or 1 indi-

cating the two treatments. Then $\boldsymbol{\alpha} = \mathrm{E}(\mathbf{Y}|X = 0)$ is the mean profile for one treatment and $\boldsymbol{\beta} = \mathrm{E}(\mathbf{Y}|X = 1) - \mathrm{E}(\mathbf{Y}|X = 0)$ is the mean profile difference between treatments. Fitting by maximum likelihood yields the profile plot of the fitted response vectors given in the top panel of Figure 1.1. The maximum absolute ratio of an element in $\widehat{\boldsymbol{\beta}}$ to its bootstrap standard error is only about 1.3, suggesting that there were no inter-treatment difference at any time. However, the likelihood ratio test statistics for the hypothesis $\boldsymbol{\beta} = 0$ is about 27 with 10 degrees of freedom, which indicates that the two treatments did have different effects on cattle weights. Further analysis is necessary to gain a better understanding of the treatment effects. The literature on longitudinal data is rich with ways of modeling mean profiles as a function of time and structuring $\boldsymbol{\Sigma}_{\mathbf{Y}|X}$ to reflect dependence over time. Although not designed specifically for longitudinal data, an envelope analysis offers a viable alternative that does not require prior selection of models for the mean profile or the covariance structure.

Turning to a fit of the envelope model (1.2), likelihood ratio testing and Bayesian information criterion (BIC) both give a strong indication that $u = 1$, so only a single linear combination of the elements of $\mathbf{Y}$ is relevant to comparing treatments. Details for determining an envelope dimension can be found in Cook et al. (2010), and we illustrate dimension selection in the real data applications of Section 6. The corresponding fitted weight profiles are given in the bottom plot of Figure 1.1. The two profile plots in Figure 1.1 are quite similar, except the envelope profiles are notably closer in the early weeks, supporting the notion that there is a lag between treatment application and effect. The absolute ratios of elements in $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ to their bootstrap standard errors were all smaller than 2.5 before week 10 and all larger than 3.4 from week 10 and on. This finding gives an answer the original question: the two treatments have different effects on cattle weight growth starting no later than week 10.

To illustrate the working mechanism of the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\beta}}$ graphically, as shown in Figure 1.2, we use only the weights at week 12 and 14 as the bivariate response $\mathbf{Y} = (Y_1, Y_2)^T$. The maximum likelihood estimator of the first element β_1 of $\boldsymbol{\beta}$ under model (1.1) is just the difference between the treatment means at week 12. In terms of Figure 1.2 this corresponds to projecting each data point onto the horizontal axis and comparing the means of the resulting univariate samples, as represented schematically by the two relatively flat densities for $Y_1|(X = 0)$ and $Y_1|(X = 1)$ along the horizontal axis of Figure 1.2. These densities are close, indicating that it may take a very large sample size to detect the difference between the

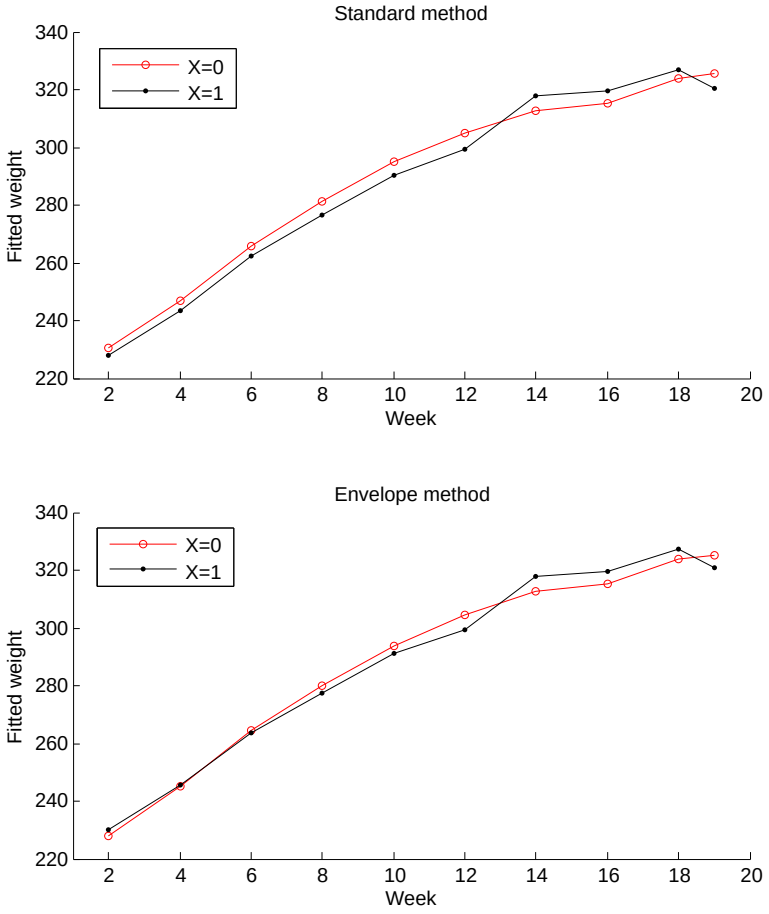


Figure 1.1: Cattle data: The top plot is obtained by a standard likelihood analysis and the bottom plot is obtained by the corresponding envelope analysis.

marginal means that is evident in the figure. In contrast, the envelope estimator for β_1 first projects the data onto the estimated envelope $\text{span}(\widehat{\boldsymbol{\Gamma}})$ and then onto the horizontal axis, as illustrated in Figure 1.2. The result is represented schematically by two peaked densities on the horizontal axis. These densities are relatively well separated, reflecting the efficiency gains provided by an envelope analysis. The standard estimator $\widehat{\boldsymbol{\beta}} = (5.5, -4.8)^T$ with bootstrap standard errors $(4.2, 4.4)^T$, while the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}} = (5.4, -5.1)^T$ with bootstrap standard errors $(1.12, 1.07)^T$.

1.3 Overview of available envelope methods

Envelopes were introduced by Cook, Li and Chiaromonte (2010) for response reduction in the multivariate linear model with normal errors. They proved that the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_{\text{env}}$ of $\boldsymbol{\beta}$ from model (1.2) is $\widehat{\boldsymbol{\beta}}_{\text{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\beta}}$ and that the asymptotic variance of $\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{env}})$ is never larger than that for $\text{vec}(\widehat{\boldsymbol{\beta}})$. The reduction in variation achieved by the envelope estimator can be substantial when the immaterial variation $\boldsymbol{\Omega}_0 = \text{var}(\boldsymbol{\Gamma}_0^T\mathbf{Y})$ is large relative to the material variation $\boldsymbol{\Omega} = \text{var}(\boldsymbol{\Gamma}^T\mathbf{Y})$. For instance, using the Frobenius norm, we

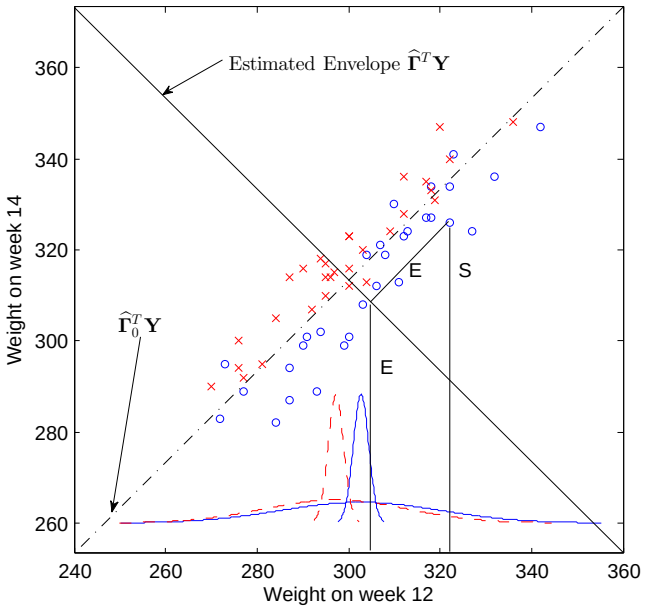


Figure 1.2: Cattle data with the 30 animals receiving one treatment marked as o's and the 30 animals receiving the other marked as x's. Representative projection paths, labeled 'E' for envelope analysis and 'S' for standard analysis, are shown on the plot. Along the 'E' path only the material information $\mathbf{P}_{\widehat{\Gamma}}\mathbf{Y}$ is used for the envelope estimator.

typically see substantial gains when $||\mathbf{\Omega}_0||_F \gg ||\mathbf{\Omega}||_F$, as happens in Figure 1.2.

When some predictors are of special interest, Su and Cook (2011) proposed the partial envelope model, with the goal of improving efficiency of the estimated coefficients corresponding to these particular predictors. They used the $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$-envelope of $\text{span}(\boldsymbol{\beta}_1)$, $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$, to develop a partial envelope estimator of $\boldsymbol{\beta}_1$ in the partitioned multivariate linear regression

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 X_{1i} + \boldsymbol{\beta}_2 \mathbf{X}_{2i} + \boldsymbol{\varepsilon}_i, \ i = 1, \ldots, n, \tag{1.3}$$

where $\boldsymbol{\beta}_1 \in \mathbb{R}^r$ is the parameter vector of interest, $\mathbf{X} = (X_1, \mathbf{X}_2^T)^T$ and the remaining terms are as defined for model (1.1). The partial envelope estimator $\widehat{\boldsymbol{\beta}}_{1,\text{env}} = \mathbf{P}_{\widehat{\Gamma}}\widehat{\boldsymbol{\beta}}_1$ has the potential to yield efficiency gains beyond those for the full envelope, particularly when $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}) = \mathbb{R}^r$ so the full envelope offers no gain.

Cook, Helland and Su (2013) used envelopes to study predictor reduction in multivariate linear regression (1.1), where the predictors are stochastic with $\text{var}(\mathbf{X}) = \mathbf{\Sigma}_{\mathbf{X}}$. Their reasoning led them to parameterize the linear model in terms of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$, which again achieved substantial efficiency gains in the estimation of $\boldsymbol{\beta}$ and in prediction. They also showed that the SIMPLS algorithm (de Jong 1993; see also ter Braak and de Jong , 1998) for partial least squares provides a $\sqrt{n}$-consistent estimator of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ and demonstrated that the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}} = \widehat{\boldsymbol{\beta}}\mathbf{P}^T_{\widehat{\Gamma}(\mathbf{S}_{\mathbf{X}})}$ typically outperforms the SIMPLS estimator in practice, where we

use $\mathbf{P}_{\mathcal{A}(\mathbf{V})} := \mathbf{P}_{\mathbf{A}(\mathbf{V})} = \mathbf{A}(\mathbf{A}^T\mathbf{V}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{V}$ to denote the projection onto $\mathcal{A} = \text{span}(\mathbf{A})$ in the $\mathbf{V}$ inner product, and $\mathbf{Q}_{\mathbf{A}(\mathbf{V})} = \mathbf{I} - \mathbf{P}_{\mathbf{A}(\mathbf{V})}$.

Given the dimension of the envelope, the envelope estimators in these three articles are all maximum likelihood estimators based on normality assumptions, but are also $\sqrt{n}$-consistent with only finite fourth moments. It can be shown that the partial maximized log-likelihood functions $L_n(\mathbf{\Gamma}) = -(n/2)J(\mathbf{\Gamma})$ for estimation of a basis $\mathbf{\Gamma}$ of $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$, $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$ or $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ all have the same form with

$$J(\mathbf{\Gamma}) = \log|\mathbf{\Gamma}^T\widehat{\mathbf{M}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1}\mathbf{\Gamma}|, \tag{1.4}$$

where the positive definite matrix $\widehat{\mathbf{M}}$ and the positive semi-definite matrix $\widehat{\mathbf{U}}$ depend on context. The estimated basis is $\widehat{\mathbf{\Gamma}} = \arg\min J(\mathbf{\Gamma})$, where the minimization is carried out over a set of semi-orthogonal matrices whose dimensions depend on the envelope being estimated. Estimates of the remaining parameters are then simple functions of $\widehat{\mathbf{\Gamma}}$. We represent sample covariance matrices as $\mathbf{S}_{(\cdot)}$ and defined with the divisor n: $\mathbf{S}_{\mathbf{X}} = \sum_{i=1}^n(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T/n$, $\mathbf{S}_{\mathbf{XY}} = \sum_{i=1}^n(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{Y}_i - \overline{\mathbf{Y}})^T/n$ and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ denotes the covariance matrix of the residuals from the linear fit of $\mathbf{Y}$ on $\mathbf{X}$. To determine the estimators of the response envelope $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$, the predictor envelope $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ and the partial envelope $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$, we have $\{\widehat{\mathbf{M}}, \widehat{\mathbf{M}} + \widehat{\mathbf{U}}\} = \{\mathbf{S}_{\mathbf{Y}|\mathbf{X}}, \mathbf{S}_{\mathbf{Y}}\}$, $\{\mathbf{S}_{\mathbf{X}|\mathbf{Y}}, \mathbf{S}_{\mathbf{X}}\}$ and $\{\mathbf{S}_{\mathbf{Y}|\mathbf{X}}, \mathbf{S}_{\mathbf{Y}|\mathbf{X}_2}\}$, respectively. For instance, a basis of the response envelope for the cattle data was estimated by maximizing $L_n(\mathbf{\Gamma})$ with $\{\widehat{\mathbf{M}}, \widehat{\mathbf{M}} + \widehat{\mathbf{U}}\} = \{\mathbf{S}_{\mathbf{Y}|\mathbf{X}}, \mathbf{S}_{\mathbf{Y}}\}$.

Still in the context of multivariate linear regression, Schott (2013) used saddle point approximations to improve a likelihood ratio test for the envelope dimension. Su and Cook (2013) adapted envelopes for the estimation of multivariate means with heteroscedastic errors, and Su and Cook (2012) introduced a different type of envelope construction, called inner envelopes, that can produce efficiency gains when envelopes offer no gains. Cook and Zhang (2014b) introduced envelopes for simultaneously reducing the predictors and responses and showed synergetic effects of simultaneous envelopes in improving both estimation efficiency and prediction. Cook and Zhang (2014a) proposed a fast and stable 1D algorithm for envelope estimation.

1.4 Organization and notation

The previous studies on envelope models and methods are all limited to multivariate linear regression. While envelope constructions seem natural and intuitive in that setting, nothing is available to guide the construction of envelopes in other contexts like generalized linear models. In this article, we introduce the constructive principle that an asymptotically normal estimator $\widehat{\boldsymbol{\phi}}$ of a parameter vector $\boldsymbol{\phi}$ may be improved by enveloping $\mathrm{span}(\boldsymbol{\phi})$ with respect to the asymptotic covariance of $\widehat{\boldsymbol{\phi}}$. This principle recovers past estimators and allows for extensions to many other contexts. While the past envelope applications are rooted in normal likelihood-based linear regression, the new constructive principle not only broadens the existing methods beyond linear regression but also allows generic moment-based estimation alternatives.

The rest of this article is organized as follows. In Section 2 we give a general constructive envelope definition. Based on this definition, we study envelope regression in Section 3 for generalized linear models and other applications. In Section 4, we extend envelope estimation beyond the scope of regression applications and lay general estimation foundations. We propose moment-based and objective-function-based estimation procedures in Section 4.1 and turn to envelopes in likelihood-based estimation in Section 4.2. Simulation results are given in Section 5, and illustrative analyses are given in Section 6.

Although our focus in this article is on vector-valued parameters, we describe in a Supplement to this article how envelopes for matrix parameters can be constructed generally. Due to space limitations, we focus our new applications in generalized linear models. The Supplement also contains applications to weighted least squares and Cox regression, along with additional simulation results, proofs and other technical details.

The following additional notations and definitions will be used in our exposition. The Grassmannian consisting of the set of all u dimensional subspaces of $\mathbb{R}^r$, $u \leq r$, is denoted as $\mathcal{G}_{u,r}$. If $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to a normal random vector with mean 0 and covariance matrix $\mathbf{V}$ we write its asymptotic covariance matrix as $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\theta}}) = \mathbf{V}$. We will use operators $\mathrm{vec} : \mathbb{R}^{a\times b} \to \mathbb{R}^{ab}$, which vectorizes an arbitrary matrix by stacking its columns, and $\mathrm{vech} : \mathbb{R}^{a\times a} \to \mathbb{R}^{a(a+1)/2}$, which vectorizes a symmetric matrix by stacking the unique elements of its columns. Let $\mathbf{A} \otimes \mathbf{B}$ denote the Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$. Envelope estimators will be written with the subscript env, unless the estimator is uniquely

associated with the envelope construction itself, in which case the subscript designation is unnecessary. We use $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}$ to denote an estimator of a parameter $\boldsymbol{\theta}$ given another parameter $\boldsymbol{\alpha}$. For instance, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\boldsymbol{\Gamma}}$ is an estimator of $\boldsymbol{\Sigma}_{\mathbf{X}}$ given $\boldsymbol{\Gamma}$.

2 A constructive definition of envelopes

The following proposition summarizes some key algebraic properties of envelopes. For a matrix $\mathbf{M} \in \mathbb{S}^{p\times p}$, let λ_i and $\mathbf{P}_i$, $i = 1, \ldots, q$, be its distinct eigenvalues and corresponding eigenprojections so that $\mathbf{M} = \sum_{i=1}^{q} \lambda_i \mathbf{P}_i$. Define $f^* : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ as $f^*(\mathbf{M}) = \sum_{i=1}^{q} f(\lambda_i)\mathbf{P}_i$, where $f : \mathbb{R} \to \mathbb{R}$ is a function such that $f(x) = 0$ if and only if $x = 0$.

Proposition 1. *(Cook et al. 2010)*

1. *If $\mathbf{M} \in \mathbb{S}^{p\times p}$ has $q \leq p$ eigenspaces, then the $\mathbf{M}$-envelope of $\mathcal{B} \subseteq \operatorname{span}(\mathbf{M})$ can be constructed as $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) = \sum_{i=1}^{q} \mathbf{P}_i\mathcal{B}$;*

2. *With f and f^* as previously defined, $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) = \mathcal{E}_{\mathbf{M}}(f^*(\mathbf{M})\mathcal{B})$.*

3. *If f is strictly monotonic then $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) = \mathcal{E}_{f^*(\mathbf{M})}(\mathcal{B}) = \mathcal{E}_{\mathbf{M}}(f^*(\mathbf{M})\mathcal{B})$.*

From the first part of this proposition, we see that the $\mathbf{M}$-envelope of $\mathcal{B}$ is the sum of the projections of $\mathcal{B}$ onto the eigenspaces of $\mathbf{M}$; the second part of this proposition gives a variety of substitutes for $\mathcal{B}$ without changing the envelope; and part 3 implies that we may consider replacing $\mathbf{M}$ by $f^*(\mathbf{M})$ for a monotonic function f without affecting the envelope.

Envelopes arose in the studies reviewed in Section 1.3 as natural consequences of postulating reductions in $\mathbf{Y}$ or $\mathbf{X}$. However, they provide no guidance on how to employ parallel reasoning in more complex settings like generalized linear models, or in settings without a clear regression structure. In terms of Definition 2, the previous studies offer no guidance on how to choose the matrix $\mathbf{M}$ and the subspace $\mathcal{B}$ for use in general multivariate problems, particularly since there are many ways to represent the same envelope, as indicated in parts 2 and 3 of Proposition 1. We next propose a broad criterion to guide these selections.

Let $\widehat{\boldsymbol{\theta}}$ denote an estimator of a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^m$ based on a sample of size n. Let $\boldsymbol{\theta}_t$ denote the true value of $\boldsymbol{\theta}$ and assume, as is often the case, that $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t)$ converges in distribution to a normal random vector with mean 0 and covariance matrix $\mathbf{V}(\boldsymbol{\theta}_t) > 0$ as $n \to$

∞. To accommodate the presence of nuisance parameters we decompose $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\boldsymbol{\psi}^T, \boldsymbol{\phi}^T)^T$, where $\boldsymbol{\phi} \in \mathbb{R}^p$, $p \leq m$, is the parameter vector of interest and $\boldsymbol{\psi} \in \mathbb{R}^{m-p}$ is the nuisance parameter vector. The asymptotic covariance matrix of $\widehat{\boldsymbol{\phi}}$ is represented as $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$, which is the $p \times p$ lower right block of $\mathbf{V}(\boldsymbol{\theta}_t)$. Then we construct an envelope for improving $\widehat{\boldsymbol{\phi}}$ as follows.

Definition 3. *The envelope for the parameter $\boldsymbol{\phi} \in \mathbb{R}^p$ is defined as $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t) \subseteq \mathbb{R}^p$.*

This definition of an envelope expands previous approaches reviewed in Section 1.3 in a variety of ways. First, it links the envelope to a particular pre-specified method of estimation through the covariance matrix $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$, while normal-theory maximum likelihood is the only method of estimation allowed by the previous approaches. The goal of an envelope is to improve that pre-specified estimator, perhaps a maximum likelihood, least squares or robust estimator. Second, the matrix to be reduced – here $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$ – is dictated by the method of estimation. Third, the matrix to be reduced can now depend on the parameter being estimated, in addition to perhaps other parameters.

Definition 3 reproduces the partial envelopes for $\boldsymbol{\beta}_1$ reviewed in Section 1.3 and the envelopes for $\boldsymbol{\beta}$ when it is a vector; that is, when $r = 1$ and $p > 1$ or when $r > 1$ and $p = 1$. It also reproduces the partially maximized log likelihood function (1.4) by setting $\mathbf{M} = \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$ and $\mathbf{U} = \boldsymbol{\phi}_t\boldsymbol{\phi}_t^T$. To apply Definition 3 for the partial envelope of $\boldsymbol{\beta}_1$ based on model (1.3), the asymptotic covariance matrix of the maximum likelihood estimator of $\boldsymbol{\beta}_1$ is $\mathbf{V}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1} = (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})_{11}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$, where $(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})_{11}$ is the $(1,1)$ element of the inverse of $\boldsymbol{\Sigma}_{\mathbf{X}} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T$. Consequently, by Proposition 1, $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}}(\boldsymbol{\beta}_1) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$, and thus Definition 3 recovers the partial envelopes of Su and Cook (2011). To construct the partially maximized log likelihood (1.4) we set $\mathbf{M} = \mathbf{V}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}$ and $\mathbf{U} = \boldsymbol{\beta}_1\boldsymbol{\beta}_1^T$. Then using sample versions gives

$$\begin{aligned} J(\boldsymbol{\Gamma}) &= \log|(\mathbf{S}_{\mathbf{X}}^{-1})_{11}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\{(\mathbf{S}_{\mathbf{X}}^{-1})_{11}\mathbf{S}_{\mathbf{Y}|\mathbf{X}} + \widehat{\boldsymbol{\beta}}_1\widehat{\boldsymbol{\beta}}_1^T\}^{-1}\boldsymbol{\Gamma}| \\ &= \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}_2}^{-1}\boldsymbol{\Gamma}|, \end{aligned}$$

which is the partially maximized log-likelihood of Su and Cook (2011). It is important to note that, although $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1) = \mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}}(\boldsymbol{\beta}_1)$, Definition 3 requires that we use $\mathbf{V}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1} = (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})_{11}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ and not $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ alone. The response envelope by Cook et al. (2010) can be seen as

a special case of the partial envelope model with absence of $\mathbf{X}_2$. This implies that Definition 3 can also reproduce the response envelope objective function, which is obtained by replacing $\mathbf{S}_{\mathbf{Y}|\mathbf{X}_2}$ with $\mathbf{S}_{\mathbf{Y}}$ in the $J(\mathbf{\Gamma})$ function (1.4).

As another illustration, consider $\mathbf{X}$ reduction in model (1.1) with $r = 1$ and $p > 1$. To emphasize the scalar response, let $\sigma^2_{Y|\mathbf{X}} = \mathrm{var}(\varepsilon)$ with sample residual variance $s^2_{Y|\mathbf{X}}$. The ordinary least squares estimator of $\boldsymbol{\beta}$ has asymptotic variance $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sigma^2_{Y|\mathbf{X}}\mathbf{\Sigma}^{-1}_{\mathbf{X}}$. Direct application of Definition 3 then leads to the $\sigma^2_{Y|\mathbf{X}}\mathbf{\Sigma}^{-1}_{\mathbf{X}}$-envelope of $\mathrm{span}(\boldsymbol{\beta})$, $\mathcal{E}_{\sigma^2_{Y|\mathbf{X}}\mathbf{\Sigma}^{-1}_{\mathbf{X}}}(\boldsymbol{\beta})$. However, it follows from Proposition 1 that this envelope is equal to $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$, which is the envelope used by Cook, et al., (2013) when establishing connections with partial least squares. To construct the corresponding version of (1.4), let $\mathbf{M} = \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ and $\mathbf{U} = \boldsymbol{\beta}\boldsymbol{\beta}^T$. Then substituting sample quantities

$$\begin{aligned} J(\mathbf{\Gamma}) &= \log|s^2_{Y|\mathbf{X}}\mathbf{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T\{s^2_{Y|\mathbf{X}}\mathbf{S}^{-1}_{\mathbf{X}} + \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T\}^{-1}\mathbf{\Gamma}| \\ &= \log|\mathbf{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T(\mathbf{S}_{\mathbf{X}} - \mathbf{S}_{\mathbf{X}Y}\mathbf{S}^T_{\mathbf{X}Y}/s^2_Y)^{-1}\mathbf{\Gamma}|, \end{aligned} \tag{2.1}$$

which is the partially maximized log-likelihood of Cook et al. (2013). While $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}) = \mathcal{E}_{\sigma^2_{Y|\mathbf{X}}\mathbf{\Sigma}^{-1}_{\mathbf{X}}}(\boldsymbol{\beta})$, we must still use $\mathbf{M} = \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sigma^2_{Y|\mathbf{X}}\mathbf{\Sigma}^{-1}_{\mathbf{X}}$ in the construction of $J(\mathbf{\Gamma})$.

Definition 3 in combination with $J(\mathbf{\Gamma})$ can also be used to derive envelope estimators for new problems. For example, consider enveloping for a multivariate mean $\boldsymbol{\mu}$ in the model $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(0, \mathbf{\Sigma})$. We take $\boldsymbol{\phi} = \boldsymbol{\mu}$ and $\widehat{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^n \mathbf{Y}_i$. Then $\mathbf{M} = \mathbf{V}_{\boldsymbol{\mu}\boldsymbol{\mu}} = \mathbf{\Sigma}$, which is the asymptotic covariance matrix of $\widehat{\boldsymbol{\mu}}$, $\mathbf{U} = \boldsymbol{\mu}\boldsymbol{\mu}^T$, and $\mathbf{M} + \mathbf{U} = \mathrm{E}(\mathbf{Y}\mathbf{Y}^T)$. Substituting sample versions of $\mathbf{M}$ and $\mathbf{U}$ leads to the same objective function $J(\mathbf{\Gamma})$ as that obtained when deriving the envelope estimator from scratch.

3 Envelope regression

In this section we study envelope estimators in a general regression setting. We first discuss the likelihood-based approach to envelope regression in Section 3.1 and then discuss specific applications including generalized linear models (GLM) in Sections 3.2 – 3.4.

3.1 Conditional and unconditional inference in regression

Let $Y \in \mathbb{R}^1$ and $\mathbf{X} \in \mathbb{R}^p$ have a joint distribution with parameters $\boldsymbol{\theta} := (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{\psi}^T)^T \in \mathbb{R}^{q+p+s}$, so the joint density or mass function can be written as $f(Y, \mathbf{X}|\boldsymbol{\theta}) = g(Y|\boldsymbol{\alpha}, \boldsymbol{\beta}^T\mathbf{X})h(\mathbf{X}|\boldsymbol{\psi})$

and the observed data are $\{Y_i, \mathbf{X}_i\}$, $i = 1, \ldots, n$. We take $\boldsymbol{\beta}$ to be the parameter vector of interest and, prior to the introduction of envelopes, we restrict $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ to a product space. Let $L_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(Y_i, \mathbf{X}_i|\boldsymbol{\theta})$ be the full log-likelihood, let the log-likelihood conditional on $\mathbf{X}$ be represented by $C_n(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^n \log g(Y_i|\boldsymbol{\alpha}, \boldsymbol{\beta}^T\mathbf{X}_i)$ and let $M_n(\boldsymbol{\psi}) = \sum_{i=1}^n \log h(\mathbf{X}_i|\boldsymbol{\psi})$ be the marginal log-likelihood for $\boldsymbol{\psi}$. Then we can decompose $L_n(\boldsymbol{\theta}) = C_n(\boldsymbol{\alpha}, \boldsymbol{\beta}) + M_n(\boldsymbol{\psi})$. Since our primary interest lies in $\boldsymbol{\beta}$ and $\mathbf{X}$ is ancillary, estimators are typically obtained as

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} C_n(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{3.1}$$

Our goal here is to improve the pre-specified estimator $\widehat{\boldsymbol{\beta}}$ by introducing the envelope $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$, where $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_t) = \text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}})$, according to Definition 3.

Let $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{p\times p}$ denote an orthogonal matrix where $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$ is a basis for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$. Since $\boldsymbol{\beta}_t \in \mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$, we can write $\boldsymbol{\beta}_t = \boldsymbol{\Gamma}\boldsymbol{\eta}_t$ for some $\boldsymbol{\eta}_t \in \mathbb{R}^u$. Because $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_t)$ typically depends on the distribution of $\mathbf{X}$ and $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$ reduces $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_t)$, the marginal M_n will depend on $\boldsymbol{\Gamma}$. Then the log-likelihood becomes $L_n(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\psi}_1, \boldsymbol{\Gamma}) = C_n(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\Gamma}) + M_n(\boldsymbol{\psi}_1, \boldsymbol{\Gamma})$, where $\boldsymbol{\psi}_1$ represents any parameters remaining after incorporating $\boldsymbol{\Gamma}$. Since both C_n and M_n depend on $\boldsymbol{\Gamma}$, the predictors are no longer ancillary after incorporating the envelope structure and estimation must be carried out by maximizing $\{C_n(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\Gamma}) + M_n(\boldsymbol{\psi}_1, \boldsymbol{\Gamma})\}$.

The relationship between $\mathbf{X}$ and $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$ that is embodied in $M_n(\boldsymbol{\psi}_1, \boldsymbol{\Gamma})$ could be complicated, depending on the distribution of $\mathbf{X}$. However, as described in the following proposition, it simplifies considerably when $\mathrm{E}(\mathbf{X}|\boldsymbol{\Gamma}^T\mathbf{X})$ is a linear function of $\boldsymbol{\Gamma}^T\mathbf{X}$. This is well-known as the linearity condition in the sufficient dimension reduction literature where there $\boldsymbol{\Gamma}$ denotes a basis for the central subspace (Cook 1996). Background on the linearity condition, which is widely regarded as restrictive but nonetheless rather mild, is available from Cook (1998), Li and Wang (2007) and many other articles in sufficient dimension reduction. For instance, if $\mathbf{X}$ follows an elliptically contoured distribution, the linearity condition will be guaranteed for any $\boldsymbol{\Gamma}$ (Eaton 1986).

Proposition 2. *Assume that* $\mathrm{E}(\mathbf{X}|\boldsymbol{\Gamma}^T\mathbf{X})$ *is a linear function of* $\boldsymbol{\Gamma}^T\mathbf{X}$. *Then* $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t)$.

A implication of Proposition 2 is that, for some positive definite matrices $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-u)\times(p-u)}$, $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ and thus M_n must depend on $\boldsymbol{\Gamma}$ through the

marginal covariance $\boldsymbol{\Sigma}_{\mathbf{X}}$. Consequently, we can write $M_n(\boldsymbol{\Sigma}_{\mathbf{X}}, \boldsymbol{\psi}_2) = M_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\psi}_2)$, where $\boldsymbol{\psi}_2$ represents any remaining parameters in the marginal function. If $\mathbf{X}$ is normal with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and variance $\boldsymbol{\Sigma}_{\mathbf{X}}$, then $\boldsymbol{\psi}_2 = \boldsymbol{\mu}_{\mathbf{X}}$ and $M_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\psi}_2) = M_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\mu}_{\mathbf{X}})$ is the marginal normal log-likelihood. In this case, it is possible to maximize M_n over all its parameters except $\boldsymbol{\Gamma}$, as stated in the following proposition.

Proposition 3. *Assume that $\mathbf{X} \in \mathbb{R}^p$ is multivariate normal $N(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$ and that $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$ is a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t)$. Then $\widehat{\boldsymbol{\mu}}_{\mathbf{X},\boldsymbol{\Gamma}} = \overline{\mathbf{X}}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\boldsymbol{\Gamma}} = \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{S}_{\mathbf{X}}\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{S}_{\mathbf{X}}\mathbf{Q}_{\boldsymbol{\Gamma}}$ and*

$$\begin{aligned} M_n(\boldsymbol{\Gamma}) &:= \max_{\boldsymbol{\Omega},\boldsymbol{\Omega}_0,\boldsymbol{\mu}_{\mathbf{X}}} M_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\mu}_{\mathbf{X}}) \qquad (3.2) \\ &= -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}_0|\right\} \\ &= -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}| + \log|\mathbf{S}_{\mathbf{X}}|\right\}, \qquad (3.3) \end{aligned}$$

where $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{p\times p}$ is an orthogonal basis for $\mathbb{R}^p$. Moreover, the global maximum of $M_n(\boldsymbol{\Gamma})$ is attained at all subsets of u eigenvectors of $\mathbf{S}_{\mathbf{X}}$.

This proposition indicates that if $\mathbf{X}$ is marginally normal, the envelope estimators are

$$(\widehat{\boldsymbol{\alpha}}_{\text{env}}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\Gamma}}) = \arg\max\{C_n(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\Gamma}) + M_n(\boldsymbol{\Gamma})\}. \qquad (3.4)$$

In particular, the envelope estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_{\text{env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}}$ and, from Proposition 3, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\text{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{\mathbf{X}}\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{\mathbf{X}}\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}$. It follows also from Proposition 3 that one role of M_n is to pull $\text{span}(\widehat{\boldsymbol{\Gamma}})$ toward the reducing subspaces of $\mathbf{S}_{\mathbf{X}}$, although it will not necessarily coincide with any such subspace.

3.2 Generalized linear models with canonical link

In the generalized linear model setting (Agresti 2002), Y belongs to an exponential family with probability mass or density function $f(Y_i|\vartheta_i, \varphi) = \exp\{[Y_i\vartheta_i - b(\vartheta_i)]/a(\varphi) + c(Y, \varphi)\}$, $i = 1, \ldots, n$, where ϑ is the natural parameter and $\varphi > 0$ is the dispersion parameter. We consider the canonical link function $\vartheta(\alpha, \boldsymbol{\beta}) = \alpha + \boldsymbol{\beta}^T\mathbf{X}$, which is a monotonic differentiable function of $\mathrm{E}(Y|\mathbf{X}, \vartheta, \varphi)$. We also restrict discussion to one-parameter families so that the dispersion parameter φ is not needed.

	$\mu := \mathrm{E}(Y\|\mathbf{X}, \alpha, \boldsymbol{\beta})$	$\mathcal{C}(\vartheta)$	$\mathcal{C}'(\vartheta)$	$-\mathcal{C}''(\vartheta)\ (\propto W)$
Normal	ϑ	$Y\vartheta - \vartheta^2/2$	$Y - \vartheta$	1
Poisson	$\exp(\vartheta)$	$Y\vartheta - \exp(\vartheta)$	$Y - \exp(\vartheta)$	$\exp(\vartheta)$
Logistic	$\exp(\vartheta)/A(\vartheta)$	$Y\vartheta - \log A(\vartheta)$	$Y - \exp(\vartheta)/A(\vartheta)$	$\exp(\vartheta)/A^2(\vartheta)$
Exponential	$-\vartheta^{-1} > 0$	$Y\vartheta - \log(-\vartheta)$	$(Y - 1/\vartheta)$	ϑ^{-2}

Table 1: A summary for the mean functions, the conditional log-likelihoods and their derivatives of various exponential family distributions. $A(\vartheta) = 1 + \exp(\vartheta)$.

The conditional log likelihood takes the form of $\log f(y|\vartheta) = y\vartheta - b(\vartheta) + c(y) := \mathcal{C}(\vartheta)$, where $\vartheta = \alpha + \boldsymbol{\beta}^T\mathbf{X}$ is the canonical parameter. Then the Fisher information matrix for $(\alpha, \boldsymbol{\beta})$ evaluated at the true parameters is

$$\mathbf{F}(\alpha_t, \boldsymbol{\beta}_t) = \mathrm{E}\left(-\mathcal{C}''(\alpha_t + \boldsymbol{\beta}_t^T\mathbf{X})\begin{pmatrix} 1 & \mathbf{X} \\ \mathbf{X}^T & \mathbf{X}\mathbf{X}^T \end{pmatrix}\right), \tag{3.5}$$

where $\mathcal{C}''(\alpha_t + \boldsymbol{\beta}_t^T\mathbf{X})$ is the second derivative of $\mathcal{C}(\vartheta)$ evaluated at $\alpha_t + \boldsymbol{\beta}_t^T\mathbf{X}$.

To construct the asymptotic covariance of $\widehat{\boldsymbol{\beta}}$ from the Fisher information matrix and to introduce notation, we transform $(\alpha, \boldsymbol{\beta})$ into orthogonal parameters (Cox and Reid 1987). Specifically, we transform $(\alpha, \boldsymbol{\beta})$ to $(a, \boldsymbol{\beta})$ so that a and $\boldsymbol{\beta}$ have asymptotically independent maximum likelihood estimators. Define weights $W(\vartheta) = \mathcal{C}''(\vartheta)/\mathrm{E}(\mathcal{C}''(\vartheta))$ so that $\mathrm{E}(W) = 1$, and write $\vartheta = \alpha + \boldsymbol{\beta}^T\mathrm{E}(W\mathbf{X}) + \boldsymbol{\beta}^T\{\mathbf{X} - \mathrm{E}(W\mathbf{X})\} := a + \boldsymbol{\beta}^T\{\mathbf{X} - \mathrm{E}(W\mathbf{X})\}$. Then the new parameterization $(a, \boldsymbol{\beta})$ has Fisher information matrix

$$\mathbf{F}(a_t, \boldsymbol{\beta}_t) \;=\; \mathrm{E}(-\mathcal{C}'')\begin{pmatrix} 1 & 0 \\ 0 & \boldsymbol{\Sigma}_{\mathbf{X}(W)} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{\mathbf{X}(W)} = \mathrm{E}\{W[\mathbf{X} - \mathrm{E}(W\mathbf{X})][\mathbf{X} - \mathrm{E}(W\mathbf{X})]^T\}$. We now have orthogonal parameters and $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}) = \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(a_t, \boldsymbol{\beta}_t) = \left\{\mathrm{E}(-\mathcal{C}'') \cdot \boldsymbol{\Sigma}_{\mathbf{X}(W)}\right\}^{-1}$, while $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$ is the corresponding envelope. The parameterizations $(a, \boldsymbol{\beta})$ and $(\alpha, \boldsymbol{\beta})$ lead to equivalent implementation since we are interested only in $\boldsymbol{\beta}$. Therefore, we use $(\alpha, \boldsymbol{\beta})$ in the estimation procedure that follows and use orthogonal parameters $(a, \boldsymbol{\beta})$ to derive asymptotic properties in Section 3.3.

The conditional log-likelihood, which varies for different exponential family distributions of $Y|\mathbf{X}$, can be written as $C_n(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^n \mathcal{C}(\vartheta_i)$, where $\vartheta_i = \alpha + \boldsymbol{\beta}^T\mathbf{X}_i$ and different functions $\mathcal{C}(\vartheta)$ are summarized in Table 1. We next briefly review Fisher scoring, which is the standard iterative method for maximizing $C_n(\alpha, \boldsymbol{\beta})$, as background for the alternating envelope algorithm (Algorithm 1). At each iteration of the Fisher scoring method, the update step for $\widehat{\boldsymbol{\beta}}$

can be summarized in the form of a weighted least squares (WLS) estimator as follows:

$$\widehat{\boldsymbol{\beta}} \longleftarrow \mathbf{S}^{-1}_{\mathbf{X}(\widehat{W})}\mathbf{S}_{\mathbf{X}\widehat{V}(\widehat{W})}, \tag{3.6}$$

where $\widehat{W} = W(\widehat{\vartheta}) = W(\widehat{\alpha} + \widehat{\boldsymbol{\beta}}^T\mathbf{X})$ is the weight at the current iteration, $\widehat{V} = \widehat{\vartheta} + \{Y - \mu(\widehat{\vartheta})\}/\widehat{W}$ is a pseudo-response variable at the current iteration, the weighted covariance $\mathbf{S}_{\mathbf{X}(\widehat{W})}$ is the sample estimator of $\boldsymbol{\Sigma}_{\mathbf{X}(W)}$ and the sample weighted cross-covariance $\mathbf{S}_{\mathbf{X}\widehat{V}(\widehat{W})}$ is defined similarly. Upon convergence of the Fisher scoring process, the estimator $\widehat{\boldsymbol{\beta}} = \mathbf{S}^{-1}_{\mathbf{X}(\widehat{W})}\mathbf{S}_{\mathbf{X}\widehat{V}(\widehat{W})}$ is a function of the estimators $\widehat{\alpha}$, $\widehat{\vartheta}$, $\widehat{W}$ and $\widehat{V}$ at the final iteration.

Assuming normal predictors, it follows from Section 3.1 that the full log-likelihood can be written as $L_n(\alpha, \boldsymbol{\Gamma}, \boldsymbol{\eta}) = C_n(\alpha, \boldsymbol{\beta}) + M_n(\boldsymbol{\Gamma})$, where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ is the coefficient vector of interest and $M_n(\boldsymbol{\Gamma})$ is given in Proposition 3. For fixed $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$, the Fisher scoring method of fitting the GLM of Y on $\boldsymbol{\Gamma}^T\mathbf{X}$ leads to $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}} = (\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(\widehat{W})}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}\widehat{V}(\widehat{W})}$. Therefore, the partially maximized log-likelihood for $\boldsymbol{\Gamma}$ based on Fisher scoring algorithm is

$$\begin{aligned} L_n(\boldsymbol{\Gamma}) &= C_n(\boldsymbol{\Gamma}) + M_n(\boldsymbol{\Gamma}) \\ &= C_n(\widehat{\alpha}, \boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}) - \frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}}\boldsymbol{\Gamma}| + \log|\mathbf{S}_{\mathbf{X}}|\right\}, \end{aligned} \tag{3.7}$$

where the optimization over $\boldsymbol{\Gamma}$ treats $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$ as a function of $\boldsymbol{\Gamma}$ instead of as fixed. Since we are able to compute the analytical form of the matrix derivative $\partial L_n(\boldsymbol{\Gamma})/\partial\boldsymbol{\Gamma}$, which is summarized in Lemma 1 of Supplement Section C, the alternating update of $\boldsymbol{\Gamma}$ based on $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$ is more efficient than updating $\boldsymbol{\Gamma}$ with fixed $\widehat{\boldsymbol{\eta}}$. We summarize this alternating algorithm for fitting GLM envelope estimators in Algorithm 1, where the alternating between (2a) and (2b) typically converges in only a few rounds. Step (2a) is solved by using the *sg_min* Matlab package by Ross A. Lippert (`http://web.mit.edu/~ripper/www/software/`) with the analytical form of $\partial L_n(\boldsymbol{\Gamma})/\partial\boldsymbol{\Gamma}$, and Fisher scoring was the default method in the Matlab package *glmfit*.

3.3 Asymptotic properties with normal predictors

In this section we describe the asymptotic properties of envelope estimators in regression when $\boldsymbol{\alpha}$ (i.e. orthogonal parameter a in GLM) and $\boldsymbol{\beta}$ are orthogonal parameter vectors and $\mathbf{X}$ is normally distributed. We also contrast the asymptotic behavior of the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}}$ with that of the estimator $\widehat{\boldsymbol{\beta}}$ from $C_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$, and comment on other settings at the end of the

Algorithm 1 The alternating algorithm for GLM envelope estimator.

1. Initialize $\widehat{\alpha}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\vartheta}$, $\widehat{W}$ and $\widehat{V}$ using the standard estimators. Initialize $\widehat{\boldsymbol{\Gamma}}$ using the 1D algorithm (Algorithm 2) discussed in Section 4.1.
2. Alternating the following steps until the convergence of $\widehat{\boldsymbol{\beta}}_{\text{env}}$ or the maximum number of iterations is reached.
   (2a) Update $\widehat{\boldsymbol{\Gamma}}$ by maximizing (3.7) over Grassmann manifold $\mathcal{G}_{u,p}$, where $C_n(\boldsymbol{\Gamma})$ is calculated based on $\widehat{\alpha}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\vartheta}$, $\widehat{W}$ and $\widehat{V}$ at current step.
   (2b) Update $\widehat{\alpha}$ and $\widehat{\boldsymbol{\eta}}$ by the Fisher scoring method of fitting GLM of Y on $\widehat{\boldsymbol{\Gamma}}^T\mathbf{X}$. Then set $\widehat{\boldsymbol{\beta}}_{\text{env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}}$ and simultaneously update $\widehat{\vartheta}$, $\widehat{W}$ and $\widehat{V}$.

section. The results in this section apply to any likelihood-based envelope regression, including the envelope estimators in GLMs .

The parameters involved in the coordinate representation of the envelope model are $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Gamma}$. Since the parameters $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ depend on the basis $\boldsymbol{\Gamma}$ and the objective function is invariant under orthogonal transformations of $\boldsymbol{\Gamma}$, the estimators of these parameters are not unique. Hence, we consider only the asymptotic properties of the estimable functions $\boldsymbol{\alpha}$, $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, which are invariant under choice of basis $\boldsymbol{\Gamma}$ and thus have unique maximizers. Under the normality assumption, $\mathbf{X} \sim N(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$, we neglect the mean vector $\boldsymbol{\mu}_{\mathbf{X}}$ since it is orthogonal to all of the other parameters. We define the following parameters $\boldsymbol{\xi}$ and estimable functions $\boldsymbol{h}$.

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \\ \boldsymbol{\xi}_3 \\ \boldsymbol{\xi}_4 \\ \boldsymbol{\xi}_5 \end{pmatrix} := \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\eta} \\ \text{vec}(\boldsymbol{\Gamma}) \\ \text{vech}(\boldsymbol{\Omega}) \\ \text{vech}(\boldsymbol{\Omega}_0) \end{pmatrix}, \ \boldsymbol{h} = \begin{pmatrix} \boldsymbol{h}_1(\boldsymbol{\xi}) \\ \boldsymbol{h}_2(\boldsymbol{\xi}) \\ \boldsymbol{h}_3(\boldsymbol{\xi}) \end{pmatrix} := \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\Sigma}_{\mathbf{X}}) \end{pmatrix}.$$

Since the number of free parameters in $\boldsymbol{h}$ is $q+p+p(p+1)/2$ and the number of free parameters in $\boldsymbol{\xi}$ is $q + u + (p-u)u + u(u+1)/2 + (p-u)(p-u+1)/2 = q + u + p(p+1)/2$, the envelope model reduces the total number of parameters by $p - u$.

Proposition 4. *Assume that for $i = 1, \ldots, n$, the predictors $\mathbf{X}_i$ are independent copies of a normal random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and variance $\boldsymbol{\Sigma}_{\mathbf{X}} > 0$, and that the data $(Y_i, \mathbf{X}_i)$ are independent copies of $(Y, \mathbf{X})$ with finite fourth moments. Assume also that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are orthogonal parameter vectors. Then, as $n \to \infty$, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{env}} - \boldsymbol{\beta}_t)$ converges to a normal vector*

with mean 0 and covariance matrix

$$\begin{aligned}\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}}) &= \operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}) + \operatorname{avar}(\sqrt{n}\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}}) \\ &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{P}_{\boldsymbol{\Gamma}} + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T) \\ &\leq \operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}),\end{aligned}$$

where $\mathbf{M} = (\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T)\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0) + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_{u(p-u)}$, $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}$ *and* $\boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}_0$.

The conditional log-likelihood is reflected in $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}})$ primarily through the asymptotic variance $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}$, while $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ stem from the normal marginal likelihood of $\mathbf{X}$. The span of $\boldsymbol{\Gamma}$ reduces both $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$, so the envelope serves as a link between the conditional and marginal likelihoods in the asymptotic variance. The first addend $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$ is the asymptotic covariance of the estimator given the envelope. The second addend $\operatorname{avar}(\sqrt{n}\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})$ reflects the asymptotic cost of estimating the envelope and this term is orthogonal to the first. Moreover, the envelope estimator is always more efficient than or equally efficient as the usual estimator $\widehat{\boldsymbol{\beta}}$.

A vector $\operatorname{SE}(\widehat{\boldsymbol{\beta}}_{\text{env}})$ of asymptotic standard errors for the elements of $\widehat{\boldsymbol{\beta}}_{\text{env}}$ can be constructed by first using the plug-in method to obtain an estimate of $\widehat{\operatorname{avar}}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}})$ of $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}})$ and then $\operatorname{SE}(\widehat{\boldsymbol{\beta}}_{\text{env}}) = \operatorname{diag}^{1/2}\{\widehat{\operatorname{avar}}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}})/n\}$.

An important special case of Proposition 4 is given in the following corollary.

Corollary 1. *Under the same conditions as in Proposition 4, if we assume further that* $\boldsymbol{\Sigma}_{\mathbf{X}} = \sigma^2\mathbf{I}_p$, $\sigma^2 > 0$ *then* $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t) = \operatorname{span}(\boldsymbol{\beta}_t)$ *and* $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}}) = \operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}) = \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}$.

Corollary 1 tells us that, if we have normal predictors with isotropic covariance, then the envelope estimator is asymptotically equivalent to the standard estimator and enveloping offers no gain, but there is no loss, either. This implies that there must be some degree of co-linearity among the predictors before envelopes can offer gains. We illustrate this conclusion in the simulations of Section 5.2 for logistic and Poisson regressions and in Supplement Section B for least squares estimators.

Experience has shown that (3.4) provides a useful envelope estimator when the predictors satisfy the linearity condition but are not multivariate normal. In this case there is a connection between the desired envelope $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$ and the marginal distribution of $\mathbf{X}$, as shown in

Proposition 2, and $\widehat{\boldsymbol{\beta}}_{\text{env}}$ is still a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$. If the linearity condition is substantially violated, we can still use the objective function $\{C_n(\boldsymbol{\alpha}, \boldsymbol{\beta}) + M_n(\boldsymbol{\Gamma})\}$ to estimate $\boldsymbol{\beta}$ within $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t)$ but this envelope may no longer equal $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$. Nevertheless, as demonstrated later in Corollary 2, this still has the potential to yield an estimator of $\boldsymbol{\beta}$ with smaller asymptotic variance than $\widehat{\boldsymbol{\beta}}$, although further work is required to characterize the gains in this setting. Alternatively, the general estimation procedure proposed in Section 4.1 yields $\sqrt{n}$-consistent estimators without requiring the linearity condition.

3.4 Other regression applications

Based on our general framework, the envelope model can be adapted to other regression types. In Supplement Sections A.1 and A.2 we give derivation and implementation details for envelope in WLS regression and in Supplement Section A.3, we include details for envelopes in Cox regression. Theoretical results in this section could also apply to the WLS and the Cox regression models, as we discussed in the Supplement Section A.

The envelope methods proposed here are based on sample estimators of asymptotic covariance matrices, which may be sensitive to outliers. However, the idea of envelopes can be extended to robust estimation procedures (see, for example, Yohai et al. (1991)).

4 Envelope estimation beyond regression

Having seen that Definition 3 recovers past envelopes and having applied it in the context of GLMs, we next turn to its general use in estimation. We propose three generic estimation procedures – one based on moments, one based on a generic objective function and one based on a likelihood – for envelope models and methods in general. The estimation frameworks in this section includes previous sections as special cases, and the algorithms in this section can also be applied to any regression context.

4.1 Moment-based and objective-function-based estimation

Definition 3 combined with the 1D algorithm (Cook and Zhang 2014a), which is restated as Algorithm 2 in its population version, gives us a generic moment-based estimator of the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$, requiring only matrices $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$. Setting $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\phi}}\widehat{\boldsymbol{\phi}}^T \in \mathbb{S}^{q\times q}$ and $\widehat{\mathbf{M}}$ equal to a $\sqrt{n}$-

Algorithm 2 The 1D algorithm.

1. Set initial value $\mathbf{g}_0 = \mathbf{G}_0 = 0$.
2. For $k = 0, \ldots, u-1$, $\mathbf{g}_{k+1} \in \mathbb{R}^q$ is obtained direction in the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ as follows,
   (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 $\mathbb{R}^q$.
   (b) Define the stepwise objective function

$$J_k(\mathbf{w}) = \log(\mathbf{w}^T \mathbf{A}_k \mathbf{w}) + \log(\mathbf{w}^T \mathbf{B}_k^{-1} \mathbf{w}), \tag{4.1}$$

   where $\mathbf{A}_k = \mathbf{G}_{0k}^T \mathbf{M} \mathbf{G}_{0k}$, $\mathbf{B}_k = (\mathbf{G}_{0k}^T \mathbf{M} \mathbf{G}_{0k} + \mathbf{G}_{0k}^T \mathbf{U} \mathbf{G}_{0k})$ and $\mathbf{w} \in \mathbb{R}^{q-k}$.
   (c) Solve $\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}} J_k(\mathbf{w})$ 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.

consistent estimator of $\mathbf{V}_{\phi\phi}(\boldsymbol{\theta}_t) \in \mathbb{S}^{q\times q}$, it follows from Cook and Zhang (2014a; Proposition 6) that the 1D algorithm provides a $\sqrt{n}$-consistent estimator $\mathbf{P}_{\widehat{\mathbf{G}}_u} = \widehat{\mathbf{G}}_u \widehat{\mathbf{G}}_u^T$ of the projection onto $\mathcal{E}_{\mathbf{V}_{\phi\phi}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t) \subseteq \mathbb{R}^p$, where $\widehat{\mathbf{G}}_u$ is obtained from the sample version of the 1D algorithm and $u = \dim(\mathcal{E}_{\mathbf{V}_{\phi\phi}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t))$ is assumed to be known. The *moment-based envelope estimator* $\widehat{\boldsymbol{\phi}}_{\text{env}} = \mathbf{P}_{\widehat{\mathbf{G}}_u}\widehat{\boldsymbol{\phi}}$ is then a $\sqrt{n}$-consistent estimator of $\boldsymbol{\phi}_t$. For example, the moment-based estimator for GLMs is obtained by letting $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T$ and $\widehat{\mathbf{M}} = \mathbf{S}^{-1}_{\mathbf{X}(\widehat{W})} / \left\{ n^{-1} \sum_{i=1}^n (-\mathcal{C}''(\widehat{\vartheta}_i)) \right\}$ in the 1D algorithm. Consequently, a distributional assumption, such as normality, is not a requirement for estimators based on the 1D algorithm to be useful, a conclusion that is supported by previous work and by our experience. However, the weight $W(\vartheta)$ depends on the parameter $\boldsymbol{\beta}$ so that iterative updates of weights and estimators could be used to refine the final estimator, as we will discuss at the end of Section 4.2.

The 1D algorithm can also play the role of finding starting values for other optimizations. For instance, in envelope GLMs, we use the 1D algorithm to get starting values for maximizing (3.7). Moreover, the likelihood-based objective function in (3.7) has the property that $L_n(\boldsymbol{\Gamma}) = L_n(\boldsymbol{\Gamma}\mathbf{O})$ for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{u\times u}$. Maximization of L_n is thus over the Grassmannian $\mathcal{G}_{u,p}$. Since $u(p-u)$ real numbers are required to specify an element of $\mathcal{G}_{u,p}$ uniquely, optimization of L_n is essentially over $u(p-u)$ real dimensions, and can be time consuming and sensitive to starting values when this dimension is large. The 1D algorithm mitigates these computational issues: to our experience, with the 1D estimator as the starting value for step (2a) in Algorithm 1, the iterative Grassmann manifold optimization of $L_n(\boldsymbol{\Gamma})$ in

(3.7) typically converges after only a few iterations.

To gain intuition about the potential advantages of the moment-based envelope estimator, assume that a basis $\boldsymbol{\Gamma}$ for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ is known and write the envelope estimator as $\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\phi}}$. Then since the envelope reduces $\mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t)$, we have

$$\begin{aligned}
\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}) &= \mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t) = \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t)\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t)\mathbf{Q}_{\boldsymbol{\Gamma}}, \\
\operatorname{avar}(\sqrt{n}\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\phi}}) &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t)\mathbf{P}_{\boldsymbol{\Gamma}} \leq \mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t).
\end{aligned}$$

These relationships allow some straightforward intuition by writing $\sqrt{n}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_t) = \sqrt{n}(\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_t) + \sqrt{n}\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\phi}}$. The second term $\sqrt{n}\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\phi}}$ is asymptotically normal with mean 0 and variance $\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t)\mathbf{Q}_{\boldsymbol{\Gamma}}$, and is asymptotically independent of $\sqrt{n}(\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_t)$. Consequently, we think of $\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\phi}}$ as the immaterial information in $\widehat{\boldsymbol{\phi}}$. The envelope estimator then achieves efficiency gains by essentially eliminating the immaterial variation $\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t)\mathbf{Q}_{\boldsymbol{\Gamma}}$, the greatest gains being achieved when $\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t)\mathbf{Q}_{\boldsymbol{\Gamma}}$ is large relative to $\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\phi\phi}}(\boldsymbol{\theta}_t)\mathbf{P}_{\boldsymbol{\Gamma}}$. Of course, we will typically need to estimate $\boldsymbol{\Gamma}$ in practice, which will mitigate the asymptotic advantages when $\boldsymbol{\Gamma}$ is known. But when the immaterial variation is large compared to the cost of estimating the envelope, substantial gain will still be achieved. Although we do not have an expression for the asymptotic variance of the moment-based estimator $\widehat{\boldsymbol{\phi}}_{\text{env}} = \mathbf{P}_{\widehat{\mathbf{G}}_u}\widehat{\boldsymbol{\phi}}$, the bootstrap can be used to estimate it. Depending on context, cross validation, an information criteria like AIC and BIC, or sequential hypothesis testing can be used to aid selection of u, as in Cook et al. (2010, 2013) and Su and Cook (2011).

If $\widehat{\boldsymbol{\phi}}$ is obtained by minimizing an objective function $\widehat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}\in\mathbb{R}^q} F_n(\boldsymbol{\phi})$ then, as an alternative to the moment-based estimator $\widehat{\boldsymbol{\phi}}_{\text{env}} = \mathbf{P}_{\widehat{\mathbf{G}}_u}\widehat{\boldsymbol{\phi}}$, an envelope estimator can be constructed as $\widehat{\boldsymbol{\phi}}_{\text{env}} = \widehat{\mathbf{G}}_u\widehat{\boldsymbol{\eta}}$, where $\widehat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}\in\mathbb{R}^u} F_n(\widehat{\mathbf{G}}_u\boldsymbol{\eta})$. We refer to this as the *objective-function-based estimator*. Sometimes these two approaches are identical, i.e., response envelopes (Cook et al. 2010) and partial envelopes (Su and Cook 2011). In Section 4.2 we specialize the objective function approach to maximum likelihood estimators.

4.2 Envelopes for maximum likelihood estimators

In this section, we broaden the envelope estimators in regression to generic likelihood-based estimators. Consider estimating $\boldsymbol{\theta}$ as $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta})$, where $L_n(\boldsymbol{\theta})$ is a log-likelihood that is twice continuously differentiable in an open neighborhood of $\boldsymbol{\theta}_t$. Then under standard

conditions $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t)$ is asymptotically normal with mean 0 and covariance matrix $\mathbf{V}(\boldsymbol{\theta}_t) = \mathbf{F}^{-1}(\boldsymbol{\theta}_t)$, where $\mathbf{F}$ is the Fisher information matrix for $\boldsymbol{\theta}$. The asymptotic covariance matrix of $\widehat{\boldsymbol{\phi}}$, $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$, is the lower right block of $\mathbf{V}(\boldsymbol{\theta}_t)$.

Recall that for envelope models, we can write the log-likelihood as $L_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\psi}, \boldsymbol{\phi}) = L_n(\boldsymbol{\psi}, \boldsymbol{\Gamma}\boldsymbol{\eta})$. To obtain MLE, we need to maximize $L_n(\boldsymbol{\psi}, \boldsymbol{\Gamma}\boldsymbol{\eta})$ over $\boldsymbol{\psi}$, $\boldsymbol{\eta}$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$, where $\boldsymbol{\Gamma}$ is semi-orthogonal basis for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ and $u = \dim(\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t))$. We first discuss the potential advantages of envelopes by considering the maximization of $L_n(\boldsymbol{\psi}, \boldsymbol{\Gamma}\boldsymbol{\eta})$ over $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$ with known $\boldsymbol{\Gamma}$. Since $\boldsymbol{\phi}_t \in \mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$, we have that $\boldsymbol{\phi}_t = \boldsymbol{\Gamma}\boldsymbol{\eta}_t$ for $\boldsymbol{\eta}_t \in \mathbb{R}^u$. Consequently, for know $\boldsymbol{\Gamma}$, the envelope estimators become

$$
\begin{aligned}
(\widehat{\boldsymbol{\psi}}_{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}) &= \arg\max_{\boldsymbol{\psi},\boldsymbol{\eta}} L_n(\boldsymbol{\psi}, \boldsymbol{\Gamma}\boldsymbol{\eta}), && (4.2)\\
\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}} &= \boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}} && (4.3)\\
\widehat{\boldsymbol{\theta}}_{\boldsymbol{\Gamma}} &= (\widehat{\boldsymbol{\psi}}_{\boldsymbol{\Gamma}}^T, \widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}^T)^T, && (4.4)
\end{aligned}
$$

Any basis for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ will give the same solution $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}$: for an orthogonal matrix $\mathbf{O}$, write $\boldsymbol{\phi} = \boldsymbol{\Gamma}\mathbf{O}\mathbf{O}^T\boldsymbol{\eta}$. Then $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}\mathbf{O}} = \mathbf{O}^T\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$.

The estimator $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}$ given in (4.3) is in general different from the moment-based estimator $\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\phi}}$ discussed near the end of Section 4.1. However, as implied by the following proposition, these two estimators have the same asymptotic distribution, which provides some support for the simple projection estimator of Section 4.1.

Proposition 5. *As $n \to \infty$, $\sqrt{n}(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}} - \boldsymbol{\phi}_t)$ converges to a normal random vector with mean 0 and asymptotic covariance*

$$
\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}) = \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\mathbf{P}_{\boldsymbol{\Gamma}} \le \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t).
$$

The $\sqrt{n}$-consistent nuisance parameter estimator also satisfies $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\psi}}_{\boldsymbol{\Gamma}}) \le \operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\psi}})$.

The following corollary to Proposition 5 characterizes the asymptotic variance when an arbitrary envelope is used.

Corollary 2. *If $\boldsymbol{\Gamma}$ is a basis for an arbitrary envelope $\mathcal{E}_{\mathbf{M}}(\boldsymbol{\phi}_t)$, where $\mathbf{M}$ is a symmetric positive definite matrix, then*

$$
\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}) = \boldsymbol{\Gamma}\{\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1}(\boldsymbol{\theta}_t)\boldsymbol{\Gamma}\}^{-1}\boldsymbol{\Gamma}^T \le \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t). \quad (4.5)
$$

Algorithm 3 The iterative envelope algorithm for MLE.

1. Initialize $\widehat{\boldsymbol{\theta}}_0 = \widehat{\boldsymbol{\theta}}$, and let $\widehat{\mathbf{M}}_k$ and $\widehat{\mathbf{U}}_k$ be the estimators of $\mathbf{M}$ and $\mathbf{U}$ based on the k-th envelope estimator $\widehat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\theta}$, so $\widehat{\mathbf{M}}_0$ and $\widehat{\mathbf{U}}_0$ are based only on the pre-specified estimator.
2. For $k = 0, 1, \ldots$ iterate as follows until a measure of the change between $\widehat{\boldsymbol{\phi}}_{k+1}$ and $\widehat{\boldsymbol{\phi}}_k$ is sufficiently small
   (a) Using the 1D algorithm, construct an estimated basis $\widehat{\boldsymbol{\Gamma}}_k$ for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ using $\widehat{\mathbf{M}}_k = \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\widehat{\boldsymbol{\theta}}_k)$ and $\widehat{\mathbf{U}}_k = \widehat{\boldsymbol{\phi}}_k\widehat{\boldsymbol{\phi}}_k^T$.
   (b) Set $\widehat{\boldsymbol{\theta}}_{k+1} = \widehat{\boldsymbol{\theta}}_{\widehat{\boldsymbol{\Gamma}}_k}$ from (4.4).

The above expression shows that $\boldsymbol{\Gamma}$ is intertwined with $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$ in the asymptotic covariance of the envelope estimator $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}$, while the envelope by Definition 3 makes $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}})$ more interpretable because the material and the immaterial variations are separable.

As formulated, this likelihood context allows us to construct the envelope estimator $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}$ when a basis $\boldsymbol{\Gamma}$ for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ is known, but it does not by itself provide a basis estimator $\widehat{\boldsymbol{\Gamma}}$. However, a basis can be estimated by using the 1D algorithm (Algorithm 2 in Section 4.1), setting $\mathbf{M} = \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta})$ and $\mathbf{U} = \boldsymbol{\phi}\boldsymbol{\phi}^T$ and plugging-in the pre-specified estimator $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta})$ to get $\sqrt{n}$-consistent estimators of $\mathbf{M}$ and $\mathbf{U}$. The envelope estimator is then $\widehat{\boldsymbol{\phi}}_{\text{env}} = \widehat{\boldsymbol{\phi}}_{\widehat{\boldsymbol{\Gamma}}}$, where $\widehat{\boldsymbol{\Gamma}} = \widehat{\mathbf{G}}_u$. Then we have the following proposition.

Proposition 6. *If the estimated basis $\widehat{\boldsymbol{\Gamma}}$ for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ is obtained by the 1D algorithm (Algorithm 2), then the envelope estimator $\widehat{\boldsymbol{\phi}}_{\text{env}} = \widehat{\boldsymbol{\phi}}_{\widehat{\boldsymbol{\Gamma}}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\phi}$.*

The envelope estimator $\widehat{\boldsymbol{\phi}}_{\text{env}}$ depends on the pre-specified estimator $\widehat{\boldsymbol{\theta}}$ through $\mathbf{M} = \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta})$ and $\mathbf{U} = \boldsymbol{\phi}\boldsymbol{\phi}^T$. Although it is $\sqrt{n}$-consistent, we have found empirically that, when $\mathbf{M}$ depends non-trivially on $\boldsymbol{\theta}_t$, it can often be improved by iterating so the current estimate of $\widehat{\boldsymbol{\theta}}$ is use to construct estimates of $\mathbf{M}$ and $\mathbf{U}$. The iteration can be implemented as Algorithm 3.

5 Simulations

In this section we present a few simulations to support and illustrate the foundations discussed in previous sections. In Section 5.1 we simulate two datasets to illustrate the workings of envelope estimation in logistic regression. In F 5.2 and 5.3, we simulated 100 datasets for each of the three sample sizes: $n = 50$, 200 and 800. The true dimension of the envelope was always

used in estimation. We mainly focus on the performance of Algorithm 1 in this section, since the predictors were simulated from normal distributions.

5.1 Envelope in logistic regression

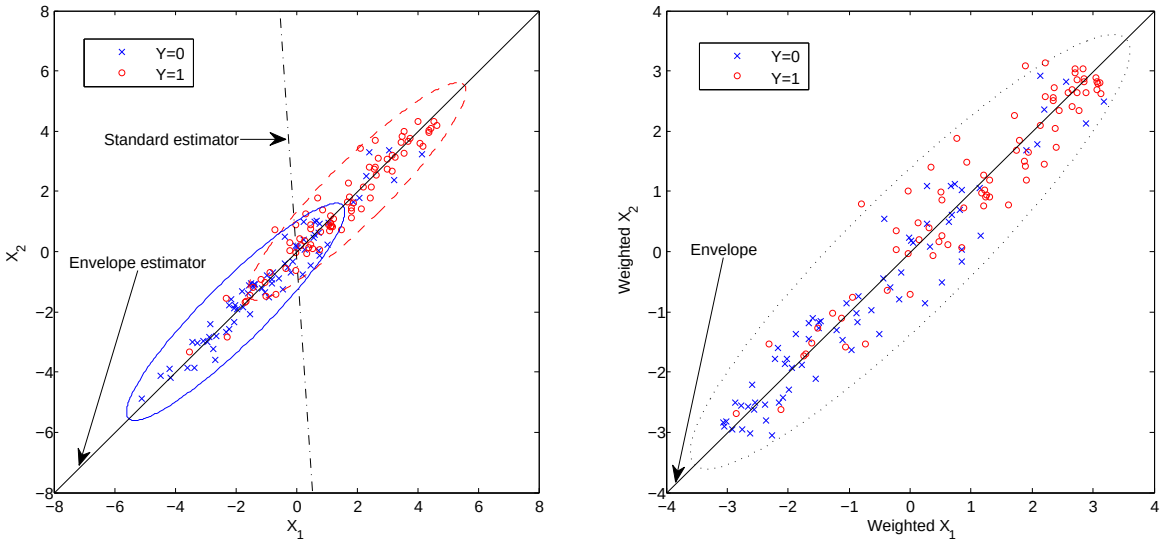
Following Definition 3, we derived in Section 3.2 that the envelope in logistic regression is $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}(W)}}(\boldsymbol{\beta})$, where $\mathbf{\Sigma}_{\mathbf{X}(W)} = \mathrm{cov}(\sqrt{W}\mathbf{X})$ is the covariance of the weighted predictors $\sqrt{W}\mathbf{X}$ and $W = \exp(\boldsymbol{\beta}^T\mathbf{X})/\{1+\exp(\boldsymbol{\beta}^T\mathbf{X})\}^2$. We now use a small simulation study, where $p = 2$ and $u = 1$, to illustrate graphically the advantages of envelopes.

We generated 150 independent observations as follows: $Y_i|\mathbf{X}_i \sim \mathrm{Bernoulli}(\mathrm{logit}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{\mathrm{i}}))$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T = (0.25, 0.25)^T$ and $\mathbf{X}_i \sim N(0, \mathbf{\Sigma}_{\mathbf{X}})$. We let $\mathrm{span}(\boldsymbol{\beta})$ be the principal eigenvector of $\mathbf{\Sigma}_{\mathbf{X}}$ with eigenvalue 10 and let the other eigenvalue be 0.1, which led to co-linearity and made the estimation challenging. We plotted the data in Figure 5.1(a), in which we used two lines to indicate the directions of the two estimators: standard estimator $\widehat{\boldsymbol{\beta}} = (-0.047, 0.692)^T$ with asymptotic standard errors $\mathrm{SE}(\widehat{\boldsymbol{\beta}}) = (0.418, 0.429)^T$, and envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = (0.321, 0.319)^T$ (based on Algorithm 1) with asymptotic standard errors $\mathrm{SE}(\widehat{\boldsymbol{\beta}}_{\mathrm{env}}) = (0.058, 0.057)^T$. The components of $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ are large relative to their standard errors, while the components of $\widehat{\boldsymbol{\beta}}$ are not. The alternative moment-based envelope estimators discussed in Sections 4.1 and 4.2, $\widehat{\boldsymbol{\beta}}_{\mathrm{env},2} = (0.323, 0.317)^T$ (based on Algorithm 2 as indicated by the added subscript) and $\widehat{\boldsymbol{\beta}}_{\mathrm{env},3} = (0.320, 0.320)^T$ (based on Algorithm 3), were both very close to the likelihood-based estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$. We also plotted the weighted predictors $\sqrt{W}\mathbf{X}$ in Figure 5.1 (b) so that we can see from the figure that the envelope is the principal eigenvector of $\mathbf{\Sigma}_{\mathbf{X}(W)}$.

To further demonstrate the point that the standard logistic regression estimator is highly variable in the presence of co-linearity, we simulated another data set according to the same model. Again, the envelope estimator stayed close to the truth, $\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = (0.208, 0.209)^T$ with asymptotic standard errors $\mathrm{SE}(\widehat{\boldsymbol{\beta}}_{\mathrm{env}}) = (0.047, 0.047)^T$. And again the moment-based envelope estimators were both very similar to the likelihood-based estimator, $\widehat{\boldsymbol{\beta}}_{\mathrm{env},2} = (0.207, 0.208)^T$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{env},3} = (0.210, 0.208)^T$. But the standard estimator varied substantially, $\widehat{\boldsymbol{\beta}} = (0.665, -0.248)^T$ with asymptotic standard errors $\mathrm{SE}(\widehat{\boldsymbol{\beta}}) = (0.401, 0.399)^T$.

We use the Australia Institute of Sport (AIS) data to illustrate the phenomenon in this simulation setting. This data set, originally from Cook and Weisberg (1994; page 98), contains

various measurements on 102 male and 100 female athletes collected at the Australian Institute of Sport. We take height (cm) to be X_1, weight (kg) to be X_2 and Y as a binary indicator for male or female athletes. A plot of the data (not shown) is similar to that in Figure 5.1(a) and the envelope dimension was selected as $u = 1$ by five-fold cross-validation so we expected some gain from envelopes. The standard estimator $\widehat{\boldsymbol{\beta}} = (0.123, 0.062)^T$ has asymptotic standard errors $\mathrm{SE}(\widehat{\boldsymbol{\beta}}) = (0.034, 0.022)^T$, while the likelihood-based envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = (0.085, 0.086)^T$, which was obtained by Algorithm 1, has asymptotic standard errors $\mathrm{SE}(\widehat{\boldsymbol{\beta}}_{\mathrm{env}}) = (0.014, 0.014)^T$. The envelope estimator compared to the standard estimator, had 40% and 60% smaller standard errors for the two coefficients in $\boldsymbol{\beta}$.



(a) Estimated envelope and standard estimators.

(b) Visualization of $\boldsymbol{\Sigma}_{\mathbf{X}(W)} = \mathrm{cov}(\sqrt{W}\mathbf{X})$.

Figure 5.1: Illustration of envelopes in logistic regression: (a) envelope and standard estimators in one simulated data set, the true envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}(W)}}(\boldsymbol{\beta}_t)$ and its sample estimate are indistinguishable in the plot; (b) visualization of the covariance $\boldsymbol{\Sigma}_{\mathbf{X}(W)} = \mathrm{cov}(\sqrt{W}\mathbf{X})$ by plotting the weighted predictors $\sqrt{W}\mathbf{X}$ on axes, where the true weights were used: $W_i = \exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)/(1+\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i))^2$.

5.2 Generalized linear models

In this section we report the result of numerical studies to compare three estimators in the contexts of logistic and Poisson regression: (1) the standard GLM estimators obtained by the Fisher scoring method, (2) iteratively re-weighted partial least squares estimators for GLMs (IRPLS; Marx 1996), and (3) envelope estimators based on Algorithm 1. The 1D algorithm (Algo-

rithm 2) was used to provide $\sqrt{n}$-consistent starting values for Grassmannian optimizations in Algorithm 1. We used the Matlab function *glmfit* to obtain the standard GLM estimators. The IRPLS algorithm is a widely recognized extension of PLS algorithms for GLMs. It embeds a weighted partial least squares algorithm in the Fisher scoring method and iteratively updates between the reduced predictors $\widehat{\mathbf{\Gamma}}^T\mathbf{X}$ and the other parameters $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\vartheta}, \widehat{V}, \widehat{W})$, similar to our envelope method described in Section 3.2. The original IRPLS algorithm by Marx (1996) is based on the NIPALS algorithm (Wold 1966) for PLS, but our simulations all used the SIMPLS algorithm (de Jong 1993) within IRPLS. This is because SIMPLS is more widely used and is implemented in the *plsregress* function in Matlab.

In all these simulations, the envelope dimension was equal to 2, $p = 10$ and $\mathbf{X} \sim N(0, \mathbf{\Sigma}_{\mathbf{X}})$, where $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$, $\mathbf{\Gamma} \in \mathbb{R}^{p\times 2}$ was generated by filling with independent uniform $(0, 1)$ variates and then standardized so that $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ is an orthogonal matrix. The positive definite symmetric matrices $\mathbf{\Omega} \in \mathbb{R}^{2\times 2}$ and $\mathbf{\Omega}_0 \in \mathbb{R}^{8\times 8}$ were generated as $\mathbf{ODO}^T$ where $\mathbf{O}$ is an orthogonal matrix and $\mathbf{D}$ is a diagonal matrix of positive eigenvalues. We chose the eigenvalues of $\mathbf{\Omega}$ to be 0.1 and 0.5, while the eigenvalues of $\mathbf{\Omega}_0$ range from 0.002 to 20. The true parameter vector $\boldsymbol{\beta}_t = \mathbf{\Gamma}\boldsymbol{\eta}$ with $\boldsymbol{\eta} = (1, 1)^T$. The intercepts were $\alpha = 2$ for Poisson regression and $\alpha = 1$ for logistic regression.

The averaged angles and length ratios are summarized in Table 2. From this table, we found that the envelope estimator was always the best for various sample sizes. The advantages of envelope estimator over its competitors were clear in both logistic and Poisson regressions. The improvements of envelope estimators over the standard estimators became more substantial with increased sample sizes. At the same time, the standard estimator always did better than the IRPLS estimator in terms of angle.

Recall that, according to Corollary 1, we need some degree of co-linearity among the predictors before envelope can offer gains. To illustrate this conclusion, we repeated the sets of simulations by using $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{I}_p$ for Poisson regression and $\mathbf{\Sigma}_{\mathbf{X}} = 4\mathbf{I}_p$ for logistic regression. From the results summarized in Table 3, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{\text{env}}$ had similar performance, as expected.

5.3 Cox regression

In the study of survival time T on the p-dimensional covariates $\mathbf{Z} \in \mathbb{R}^p$, Cox's proportional hazards model (Cox 1972, 1975) assumes the hazard function $h(t|\mathbf{Z})$ of a subject with covariates $\mathbf{Z}$

			Standard	IRPLS	Envelope
Poisson	$n = 50$	$\angle$	38 (1.5)	55 (0.8)	**25 (2.0)**
		LR	1.4	0.21	1.4
	$n = 200$	$\angle$	21 (1.1)	48 (0.6)	**7 (0.8)**
		LR	1.09	0.25	1.03
	$n = 800$	$\angle$	9 (0.5)	46 (0.4)	**2 (0.3)**
		LR	1.02	0.28	1.01
Logistic	$n = 50$	$\angle$	75 (1)	84 (1)	**74 (2)**
		LR	10.2	0.22	8.1
	$n = 200$	$\angle$	62 (2)	81 (1)	**36 (3)**
		LR	3.0	0.12	2.0
	$n = 800$	$\angle$	45 (2)	77 (0.4)	**9 (1)**
		LR	1.6	0.087	1.0

Table 2: GLM simulations with $\mathbf{\Sigma_X} = \mathbf{\Gamma\Omega\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$. We summarize averaged angles $\angle = \angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$ in degrees and length ratios $LR = ||\widehat{\boldsymbol{\beta}}||/||\boldsymbol{\beta}_t||$between the true parameter vector and various estimators. The best performances are marked in bold. Standard errors are included in the parentheses. Standard errors for the length ratios are less than or equal to 0.01.

			Standard	IRPLS	Envelope
Poisson	$n = 50$	$\angle$	8.9 (0.3)	46 (0.8)	9.2 (0.9)
		LR	1.02	0.58	1.02
	$n = 200$	$\angle$	3.7 (0.1)	38 (0.5)	3.7 (0.1)
		LR	1.00	0.70	1.00
	$n = 800$	$\angle$	1.74 (0.04)	36 (0.3)	1.74 (0.04)
		LR	1.00	0.77	1.00
Logistic	$n = 50$	$\angle$	12.6 (1.6)	17.9 (2.3)	13.7 (1.8)
		LR	1.18	0.27	1.09
	$n = 200$	$\angle$	16.2 (0.4)	29.2 (0.7)	16.9 (0.4)
		LR	1.15	0.79	1.14
	$n = 800$	$\angle$	8.0 (0.2)	17.3 (0.4)	8.1 (0.2)
		LR	1.03	0.90	1.03

Table 3: GLM simulations with isotropic covariance matrices: $\mathbf{\Sigma_X} = 4\mathbf{I}_p$ for logistic regression and $\mathbf{\Sigma_X} = \mathbf{I}_p$ for Poisson regression. We summarize averaged angles $\angle = \angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$ in degrees and length ratios $\mathrm{LR} = ||\widehat{\boldsymbol{\beta}}||/||\boldsymbol{\beta}_t||$ between the true parameter vector and various estimators. Standard errors are included in the parentheses. Standard errors for the length ratio are less than or equal to 0.01.

			Standard	PLS	Envelope
Cox	$n = 50$	$\angle$	50 (0.9)	23 (0.8)	**15 (1)**
		LR	1.95	0.99	0.97
	$n = 200$	$\angle$	25 (0.6)	15 (0.5)	**6 (0.2)**
		LR	1.16	1.00	1.00
	$n = 800$	$\angle$	13 (0.3)	10 (0.3)	**3 (0.07)**
		LR	1.03	1.00	1.00

Table 4: Cox regression: Averaged angles and length ratios between the true parameter vector and various estimators. The best performances are marked in bold. Standard errors are included in parentheses. Standard errors for the length ratio are less than or equal to 0.01.

has the form $h(t|\mathbf{Z}) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{Z})$, where $h_0(t)$ is the unspecified baseline hazard function and $\boldsymbol{\beta}$ is an unknown vector of regression coefficients in which we are primarily interested. Let X_i and C_i be the failure time and the censoring time of the i-th subject, $i = 1, \ldots, n$. Define $\delta_i = I(X_i \leq C_i)$ and $T_i = \min(X_i, C_i)$. The data then consists of $(T_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$. The failure time and the censoring time are assumed to be independent given the covariates. We assume that there are no ties for the observed times.

In our setting, $\mathbf{Z} \in \mathbb{R}^{10}$ followed the multivariate normal distribution $N(0, \boldsymbol{\Sigma}_\mathbf{Z})$ with $\boldsymbol{\Sigma}_\mathbf{Z} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + 0.2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Omega}$ were generated in the same way as in Section 5.2, except that eigenvalues of $\boldsymbol{\Omega}$ were 1 and 5. The true parameter vector $\boldsymbol{\beta}_t = \boldsymbol{\Gamma}\boldsymbol{\eta}$ where $\boldsymbol{\eta} = (1, 1)^T$. Then we followed the simulation model in Nygard and Borgan (2008), where the survival time followed a Weibull distribution with scale parameter $\exp(\boldsymbol{\beta}^T\mathbf{Z}/5)$ and shape parameter 5, which was a Weibull distribution with hazard rate $h(Y|\mathbf{Z}) = 5Y^4 \cdot \exp(\boldsymbol{\beta}^T\mathbf{Z})$. The censoring variable δ was generated as the integer part of a uniform$(0, 2)$ random variable, which gave censoring rates of approximately 50%.

We considered three different estimators: the standard estimator without envelope structure, the likelihood-based envelope estimator (see Supplement Section A.3 for implementation details), and partial least squares for Cox regression (Nygard and Borgan 2008). From Table 4, we found that the envelope estimator was always the best, while the PLS estimator also improved over the standard estimator.

6 Illustrative data analyses

We present three small examples in this section to illustrate selected aspects of the proposed foundations.

6.1 Logistic regression: Colon cancer diagnosis

We use these data to illustrate the classification power of the envelope estimator as well as its estimation efficiency. The objective of this study was to distinguish normal tissues and adenomas (precancerous tissues) that were found during colonoscopy. When the normal and adenomatous tissues are illuminated with ultraviolet light, they fluoresce at different wavelengths. The $p = 22$ predictors, which are laser-induced fluorescence (LIF) spectra measured at 8 nm spacing from 375 to 550 nm, reflect this phenomenon. The data set consists of $n = 285$ tissue samples, comprised of 180 normal tissues and 105 adenomas. Studies on a similar data set can be found in Hawkins and Maboudou-Tchao (2013).

The envelope dimension was chosen to be 2 by five-fold cross-validation. Compared to standard logistic regression, the bootstrap standard errors of the 22 elements in $\widehat{\boldsymbol{\beta}}$ were 2.2 to 12.5 times larger than the bootstrap standard errors of the elements in $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$, which was obtained from Algorithm 1. On the other hand, five-fold cross-validation classification error rate for the standard logistic regression estimator was 21% while it was 17.5% for the envelope estimator with $u = 2$ or $u = 3$. And the error rate for envelope estimator with any dimension between 1 and 10 was no larger than that of the standard estimator. To gain some idea about the level of intrinsic (non-estimative) variation in the envelope error rate, we took $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ to be the true $\boldsymbol{\beta}$, sampled the predictors 1000 times and predicted the response using $\boldsymbol{\beta}_t = \widehat{\boldsymbol{\beta}}_{\mathrm{env}}$. The resulting classification error was 16.7%, which is not far from the observed error rate of 17.5%.

6.2 Linear discriminant analysis: Wheat protein data

In this data set (Cook et al. 2010), we have a binary response variable Y indicating high or low protein content and six highly correlated predictors which are the logarithms of near infrared reflectance at six wavelengths across the range 1680-2310 nm. The correlations between these predictors range from 0.9118 to 0.9991. We standardized the predictors marginally so that each had mean 0 and variance 1. We illustrate the possibility of using envelopes to improve Fisher's

	Standard	PLS	Envelope
Error rate	10.2	57.1	5.4
S.E.	0.4	0.6	0.2

Table 5: Five-fold cross-validation error rates based on Fisher's LDA and expressed as percentages.

	Standard	IRPLS	Envelope
Pearson's χ^2	642.1	556.1	553.7
S.E.	7.8	1.6	1.6

Table 6: Performance of four methods applied to the horseshoe crab data.

linear discriminant analysis (LDA).

It has been shown (Ye 2007) that two-class LDA is equivalent to a least squares problem: no matter how we code Y for two classes, the least squares fitted coefficient vector $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is always proportional to the LDA direction. Hence, we used $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}^T\mathbf{X}$ as the discriminant direction and fitted the LDA classification rule based on it. We similarly replaced the OLS coefficient vector with the coefficient vector $\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ estimated by SIMPLS algorithm and the coefficient vector and $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ estimated by using the 1D algorithm (Algorithm 2) with respect to objective function (2.1) to improve OLS. We estimated the error rates for these three classifiers as the average error rate over 100 replications of five-fold cross-validations on the $n = 50$ samples. The average error rates are shown in Table 5. Clearly the 1D algorithm had the best classification performance.

6.3 Poisson regression: Horseshoe crab data

This is a textbook Poisson regression dataset about nesting horseshoe crabs (see Agresti (2002), Section 4.3). The response is the number of satellite male crabs residing near a female crab. Explanatory variables included the female crab's color, spine condition, weight, and carapace width. Since color is a factor with four levels, we use three indicator variables to indicate color. Also, spine condition is a factor with three levels, so we used two indicator variables for spine condition, for a total of seven predictors. The sample size is $n = 173$.

Five-fold cross-validation was used to determine the dimension of the envelope, giving $u = 1$, which was also the answer given by BIC. To assess the gain in envelope reduction with $u = 1$, we repeated the five-fold cross-validation procedure 100 times and obtained the averaged

Pearson's χ^2 statistic $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2/\widehat{Y}_i$ and its standard error shown in Table 6. Clearly, the likelihood-based envelope estimator based on Algorithm 1 had significant improvement over the standard method and a slight edge over IRPLS.

7 Discussion

With our general definition and algorithms, we laid foundations for envelope models and methods. Our general framework subsumes previous envelope models in multivariate linear regression and provides guidance for future studies. Building upon this framework, we extended the scope of envelope methods to GLMs, Cox regression and weighted least squares. Although we focused on vector-valued parameters, the extension in the Supplement for matrix-valued parameters suggests that our general construction can be straightforwardly extend to matrix-valued or even higher-order tensor-valued regressions (see, for example, Zhou and Li 2014; Zhou, Li and Zhu 2014).

We propose three estimation procedures ranging from specific to general. Estimation based on Algorithm 1 is proposed for normal predictors and is also preferable in practice when the linearity condition holds. Algorithm 3 outlines an iterative procedure that is more flexible and that provides $\sqrt{n}$-consistent estimators even without linearity or normality. We can use this procedure as a substitute for Algorithm 1 when the distribution of $\mathbf{X}$ is far from elliptical. Finally, the most generic envelope estimator for an arbitrary pre-specified estimator $\widehat{\boldsymbol{\beta}}$ is proposed as $\widehat{\boldsymbol{\beta}}_{\text{env}} = \mathbf{P}_{\widehat{\mathbf{G}}_u}\widehat{\boldsymbol{\beta}}$ or $\widehat{\boldsymbol{\beta}}_{\text{env}} = \widehat{\mathbf{G}}_u\widehat{\boldsymbol{\eta}}_{\widehat{\mathbf{G}}_u}$ if $\widehat{\boldsymbol{\beta}}$ is obtained based on objective function, where $\widehat{\mathbf{G}}_u$ is estimated from the 1D algorithm (Algorithm 2). These estimators can deal with a variety of problems, ranging from likelihood-based estimators to problems that involve heuristic non-parametric or semi-parametric estimation.

Inner envelopes (Su and Cook, 2012) are not covered by these foundations because they correspond to a different type of envelope construction. Extensions to inner envelopes are possible, but the analysis and methods involved are different than those for envelopes and so are outside the scope of this article.

Another future research direction is to extend the likelihood-based and the moment-based envelope estimation to high-dimensional settings. One possible approach is to develop penalized likelihood estimation procedure based on the likelihood-based estimation in Section 4.2.

Another more straightforwardly implemented way is to use the ridge-type covariance estimators proposed by Ledoit and Wolf (2004) for $\widehat{\mathrm{M}}$ and $\widehat{\mathrm{U}}$ in our algorithms. Properties of such extensions are yet to be studied.

Acknowledgments

Research for this article was supported in part by grant DMS-1007547 from the National Science Foundation. The authors thank Doug Hawkins for providing the Colon Cancer Diagnosis dataset. The authors are grateful to the Editor, an Associate Editor and a referee for comments that led to significant improvements in this article and to Lexin Li for helpful discussions.

References

[1] AGRESTI, A. (2002), Categorical data analysis, 2nd Edition. *New York: Wiley*.

[2] TER BRAAK, C.J.F. AND DE JONG, S. (1998), The objective function of partial least squares regression, *Journal of Chemometrics*, **12**, 41–54.

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

[4] COOK, R.D. (1996), Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983–992.

[5] COOK, R.D. (1998), Regression Graphics: Ideas for Studying Regressions Through Graphics. *New York: Wiley*.

[6] COOK, R.D. AND WEISBERG, S. (1994), An Introduction to Regression Graphics. *New York: Wiley*.

[7] COOK, R.D., HELLAND, I.S. AND SU, Z. (2013), Envelopes and partial least squares regression. *To appear in JRSS-B*.

[8] 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.

[9] COOK, R.D. AND ZHANG, X. (2014a). Algorithms for envelope estimation. *arXiv:1403.4138*.

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

[11] COX, D.R. (1972), Regression models and life-tables (with Discussi1987on), *J. R. Statist. Soc. B*, **34**, 187–220.

[12] COX, D.R. (1975), Partial likelihood, *Biometrika*, **62**, 269–276.

[13] COX, D.R. AND REID, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society B*, **49**, 1–39.

[14] EATON, M.L. (1986), A characterization of spherical distributions, *Journal of Multivariate Analysis*, **20**, 272–276.

[15] HAWKINS, D.M. AND MABOUDOU-TCHAO, E.M. (2013), Smoothed Linear Modeling for Smooth Spectral Data. *To appear in International Journal of Spectroscopy*.

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

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

[18] LEDOIT, O. AND WOLF, M. (2004), A well-conditioned estimator for large-dimensional covariance matrices. *J. Multiv. Anal.*, **88**, 365–411.

[19] LI, B. AND WANG, S. (2007). On directional regression for dimension reduction. *Journal of American Statistical Association*, **102**, 997–1008.

[20] MARX, B. D. (1996), Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*,**38**, 374–381.

[21] NYGARD, S. AND BORGAN, O. (2008), Partial least squares Cox regression for genome-wide data. *Lifetime Data Analysis*,**14**, 179–195.

[22] SCHOTT, J.R. (2013), On the likelihood ratio test for envelope models in multivariate linear regression. *Biometrika*, **100**, 531–537.

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

[24] SU, Z. AND COOK, R.D. (2012), Inner envelopes: efficient estimation in multivariate linear models. *Biometrika*, **99**, 687–702.

[25] SU, Z. AND COOK, R.D. (2013), Estimation of multivariate means with heteroscedastic errors using envelope models. *Statistica Sinica*, **23**, 213–203.

[26] WOLD, H. (1966), Estimation of Principal Components and Related Models by Iterative Least Squares. *New York: Academic Press.*

[27] YE, J. (2007), Least squares linear discriminant analysis. *Proceedings of the 24th international conference on Machine Learning.*, 1087–1093.

[28] YOHAI, V.J., STAHEL, W.A. AND ZAMAR R.H. (1991), A procedure for robust estimation and inference in linear regression, *Directions in Robust Statistics and Diagnostics, Part II, W. A. Stahel and S. W. Weisberg, Eds., Springer*, 365–374.

[29] ZHOU, H. AND LI, L. (2014), Regularized matrix regression, *JRSS-B*, **76**, 463–483.

[30] ZHOU, H., LI, L. AND ZHU, H. (2004), Tensor regression with applications in neuroimaging data analysis, *JASA*, **108**, 540–552.

Supplement: Proofs and Technical Details for "Foundations for Envelope Models and Methods"

R. Dennis Cook and Xin Zhang

A Regression applications

A.1 Envelopes for weighted least squares

Consider a heteroscedastic linear model with data consisting of n independent copies of $(Y, \mathbf{X}, W)$, where $Y \in \mathbb{R}^1$, $\mathbf{X} \in \mathbb{R}^p$, and $W > 0$ is a weight with $\mathrm{E}(W) = 1$:

$$Y = \mu + \boldsymbol{\beta}^T \mathbf{X} + \varepsilon/\sqrt{W}, \tag{A.1}$$

where $\varepsilon \perp\!\!\!\perp (\mathbf{X}, W)$ and $\mathrm{var}(\varepsilon) = \sigma^2$. The constraint $\mathrm{E}(W) = 1$ is without loss of generality and serves to normalize the weights so they have mean 1 and to make subsequent expressions a bit simpler. Fitting under this model is typically done by using weighted least squares (WLS):

$$(\widehat{\mu}, \widehat{\boldsymbol{\beta}}) = \arg\min n^{-1} \sum_{i=1}^{n} W_i (Y_i - \mu - \boldsymbol{\beta}^T \mathbf{X}_i)^2, \tag{A.2}$$

where we normalize the sample weights so that $\overline{W} = 1$. Let $\boldsymbol{\Sigma}_{\mathbf{X}(W)} = \mathrm{E}\{W(\mathbf{X} - \mathrm{E}(W\mathbf{X}))(\mathbf{X} - \mathrm{E}(W\mathbf{X}))^T\}$ be the weighted covariance matrix of $\mathbf{X}$, and let $\boldsymbol{\Sigma}_{\mathbf{X}Y(W)} = \mathrm{E}[W\{\mathbf{X} - \mathrm{E}(W\mathbf{X})\}\{Y - \mathrm{E}(WY)\}]$ be the weighted covariance between $\mathbf{X}$ and Y. Then $\boldsymbol{\beta}_t = \boldsymbol{\Sigma}_{\mathbf{X}(W)}^{-1} \boldsymbol{\Sigma}_{\mathbf{X}Y(W)}$ and $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_t)$ converges to a normal random vector with mean 0 and variance $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sigma^2 \boldsymbol{\Sigma}_{\mathbf{X}(W)}^{-1}$. According to Definition 3, we should now strive to estimate the $\sigma^2 \boldsymbol{\Sigma}_{\mathbf{X}(W)}^{-1}$-envelope of $\mathrm{span}(\boldsymbol{\beta}_t)$, which can be done for a specified dimension u by using the 1D algorithm with $\widehat{\mathbf{M}} = s^2 \mathbf{S}_{\mathbf{X}(W)}^{-1}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\beta}} \widehat{\boldsymbol{\beta}}^T$, where s^2 is the estimator of σ^2 from the WLS fit (A.2), and $\mathbf{S}_{\mathbf{X}(W)}$ is the sample version of $\boldsymbol{\Sigma}_{\mathbf{X}(W)}$. Once an estimated basis $\widehat{\boldsymbol{\Gamma}}$ for $\mathcal{E}_{\sigma^2 \boldsymbol{\Sigma}_{\mathbf{X}(W)}^{-1}}(\boldsymbol{\beta}_t)$ is obtained, we re-fit the WLS regression model, replacing $\mathbf{X}$ with $\widehat{\boldsymbol{\Gamma}}^T \mathbf{X}$, to estimate the coordinate vector $\boldsymbol{\eta}$, :

$$(\widehat{\mu}, \widehat{\boldsymbol{\eta}}) = \arg\min n^{-1} \sum_{i=1}^{n} W_i (Y_i - \mu - \boldsymbol{\eta}^T \widehat{\boldsymbol{\Gamma}}^T \mathbf{X}_i)^2. \tag{A.3}$$

The resulting envelope estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = \widehat{\boldsymbol{\Gamma}} \widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\Gamma}} (\widehat{\boldsymbol{\Gamma}}^T \mathbf{S}_{\mathbf{X}(W)} \widehat{\boldsymbol{\Gamma}})^{-1} \widehat{\boldsymbol{\Gamma}}^T \mathbf{S}_{\mathbf{X}Y(W)} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{\mathbf{X}(W)})} \widehat{\boldsymbol{\beta}}. \tag{A.4}$$

Cross validation or a hold out sample could now be used straightforwardly as an aid to selecting u. When the weights are constant, this method has the same structure as partial least squares regression with the important exception that partial least squares uses the SIMPLS algorithm in place of the 1D algorithm that we recommend.

To take a likelihood approach with the envelope $\mathcal{E}_{\sigma^2\mathbf{\Sigma}^{-1}_{\mathbf{X}(W)}}(\boldsymbol{\beta}_t)$, we follow Section 3.1 to decompose the full log-likelihood into the sum of the conditional log-likelihood from $Y|(\mathbf{X}, W)$, denoted by $C_n(\mu, \boldsymbol{\beta}, \sigma^2) := C_n(\mu, \mathbf{\Gamma}, \boldsymbol{\eta}, \sigma^2)$, and the conditional log-likelihood from $\mathbf{X}|W$, denoted by $M_n(\boldsymbol{\psi}_1, \mathbf{\Gamma})$, where $\boldsymbol{\psi}_1$ denotes additional nuisance parameters. We condition on the observed values of W, assuming that it is ancillary. Then holding $\mathbf{\Gamma}$ fixed and partially maximizing $C_n(\mu, \mathbf{\Gamma}, \boldsymbol{\eta}, \sigma^2) + M_n(\boldsymbol{\psi}_1, \mathbf{\Gamma})$, we get a likelihood-based objective function $L_n(\mathbf{\Gamma})$ that plays the same role as the moment-based objective function $J(\mathbf{\Gamma})$. Once an basis $\widehat{\mathbf{\Gamma}}$ is obtained by maximizing $L_n(\mathbf{\Gamma})$, the final estimators follow from (A.3) and (A.4).

Two kinds of assumptions on the distribution of $\mathbf{X}|W$ could be made depending on the sampling model. One is to assume that the predictors and the weights are independently sampled and that the predictor vector follows a multivariate normal distribution. By Proposition 2, we have $\mathcal{E}_{\sigma^2\mathbf{\Sigma}^{-1}_{\mathbf{X}(W)}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t)$ and the partially maximized log-likelihood up to a constant is (see derivation in the Supplement):

$$L_n(\mathbf{\Gamma}) \;=\; -\frac{n}{2}\left\{\log(s^2_{\mathbf{\Gamma}}) + \log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}}\mathbf{\Gamma}|\right\}, \tag{A.5}$$

where $s^2_{\mathbf{\Gamma}} = s^2_{Y(W)} - \mathbf{S}^T_{\mathbf{X}Y(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}Y(W)}$ is the estimator of σ^2 from (A.3) with fixed $\mathbf{\Gamma}$ and $s^2_{Y(W)} = \sum_{i=1}^n W_i(Y_i - \sum_{j=1}^n W_jY_j/n)/n$ is the sample weighted variance for Y.

Alternatively, it might be reasonable to assume that $\mathbf{X}|W$ has a weighed variance similar to the error variance from $Y|(\mathbf{X}, W)$: $\mathbf{X}|W \sim N(\boldsymbol{\mu}_{\mathbf{X}}, W^{-1}\mathbf{\Sigma})$. Then the partially maximized log-likelihood up to a constant is (see derivation in Section A.2):

$$\begin{aligned} L_n(\mathbf{\Gamma}) \;&=\; -\frac{n}{2}\left\{\log|\mathbf{\Gamma}^T s^2\mathbf{S}^{-1}_{\mathbf{X}(W)}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T(s^2\mathbf{S}^{-1}_{\mathbf{X}(W)} + \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T)^{-1}\mathbf{\Gamma}|\right\} \\ &=\; -\frac{n}{2}\left\{\log|\mathbf{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}|Y(W)}\mathbf{\Gamma}|\right\}, \end{aligned} \tag{A.6}$$

where $\mathbf{S}_{\mathbf{X}|Y(W)} = \mathbf{S}_{\mathbf{X}(W)} - \mathbf{S}_{\mathbf{X}Y(W)}\mathbf{S}^T_{\mathbf{X}Y(W)}/s^2_{Y(W)}$ and the first equality shows equivalence between maximizing this likelihood-based objective function of $\mathbf{\Gamma}$ and minimizing the moment based objective function $J(\mathbf{\Gamma})$ with $\widehat{\mathbf{M}} = s^2\mathbf{S}^{-1}_{\mathbf{X}(W)}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T$.

The asymptotic results in Proposition 4 hold for $\widehat{\boldsymbol{\beta}}_{\text{env}}$ if (i) $\mathbf{X}$ is normal and independent of W, and we use (A.5) to get $\widehat{\boldsymbol{\Gamma}}$; or (ii) $\mathbf{X}|W \sim N(\boldsymbol{\mu}_{\mathbf{X}}, W^{-1}\boldsymbol{\Sigma})$ and we use (A.6) to get $\widehat{\boldsymbol{\Gamma}}$. Then the asymptotic covariance of the envelope estimator will be asymptotically less than or equal to that of the standard WLS estimator. Moreover, as long as the model (A.1) holds and both ε and $\mathbf{X}|W$ follow a distribution with finite fourth moments, asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\text{env}}$ is guaranteed no matter which $L_n(\boldsymbol{\Gamma})$ we choose to maximize. Finally, if we fix $W = 1$, then both objective functions in (A.5) and in (A.6) reduces to (2.1), which corresponds to the envelope model for reducing $\mathbf{X}$ in the homoscedastic linear models (Cook et al. 2013).

A.2 Derivations for equation (A.6)

The normality assumption $Y|(\mathbf{X}, W) \sim N(\mu+\boldsymbol{\beta}^T\mathbf{X}, \sigma^2/W)$ leads to the following conditional log-likelihood:

$$C_n(\mu, \boldsymbol{\beta}, \sigma^2) \;=\; -n/2\log(\sigma^2) - \sum_{i=1}^{n} W_i(Y_i - \mu - \boldsymbol{\beta}^T\mathbf{X}_i)/(2\sigma^2). \tag{A.7}$$

By straightforward computation, the maximum likelihood estimators is obtained as:

$$\widehat{\boldsymbol{\beta}} \;=\; \mathbf{S}^{-1}_{\mathbf{X}(W)}\mathbf{S}_{\mathbf{X}Y(W)} \tag{A.8}$$

$$\widehat{\mu} \;=\; \sum_{i=1}^{n} W_iY_i/n - \widehat{\boldsymbol{\beta}}^T\sum_{i=1}^{n} W_i\mathbf{X}_i/n := \overline{Y}_{(W)} - \widehat{\boldsymbol{\beta}}^T\overline{\mathbf{X}}_{(W)} \tag{A.9}$$

$$\widehat{\sigma}^2 \;=\; s^2_{Y(W)} - \mathbf{S}^T_{\mathbf{X}Y(W)}\mathbf{S}^{-1}_{\mathbf{X}(W)}\mathbf{S}_{\mathbf{X}Y(W)} := s^2 \tag{A.10}$$

Given $\boldsymbol{\Gamma}$, which is a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$, we then only need to maximize $C_n(\mu, \boldsymbol{\beta}, \sigma^2)$ under the constraint that $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$. The resulting estimators are:

$$\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}} \;=\; (\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}Y(W)} \tag{A.11}$$

$$\widehat{\mu}_{\boldsymbol{\Gamma}} \;=\; \overline{Y}_{(W)} - \widehat{\boldsymbol{\eta}}^T_{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}^T\overline{\mathbf{X}}_{(W)} \tag{A.12}$$

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}} \;=\; \boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}} = \mathbf{P}_{\boldsymbol{\Gamma}(\mathbf{S}_{\mathbf{X}(W)})}\widehat{\boldsymbol{\beta}} \tag{A.13}$$

$$\widehat{\sigma}^2_{\boldsymbol{\Gamma}} \;=\; s^2_{Y(W)} - \mathbf{S}^T_{\mathbf{X}Y(W)}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}Y(W)} := s^2_{\boldsymbol{\Gamma}}. \tag{A.14}$$

If we replace $\boldsymbol{\Gamma}$ by its $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\Gamma}}$ from either a moment-based objective function or a likelihood-based objective function, then the above estimators will be our envelope estimators.

To obtain a likelihood-based objective function for $\mathbf{\Gamma}$, we need to compute the partially maximized conditional likelihood of $Y|(\mathbf{X}, W)$: $C_n(\mathbf{\Gamma}) = C_n(\widehat{\mu}_{\mathbf{\Gamma}}, \widehat{\boldsymbol{\beta}}_{\mathbf{\Gamma}}, \widehat{\boldsymbol{\sigma}}_{\mathbf{\Gamma}})$, plus the partially maximized marginal likelihood $M_n(\mathbf{\Gamma})$ of $\mathbf{X}|W$. First, we can get $C_n(\mathbf{\Gamma})$ from above derivations as:

$$C_n(\mathbf{\Gamma}) \;=\; -n/2\log(s^2_{\mathbf{\Gamma}}) - n/2. \tag{A.15}$$

Assuming that $\mathbf{X}$ and W are independent and that $\mathbf{X} \sim N(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{\Sigma}_{\mathbf{X}})$, we have by Proposition 2 that $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t)$. The partially maximized log-likelihood $M_n(\mathbf{\Gamma})$ follows from Proposition 3:

$$M_n(\mathbf{\Gamma}) \;=\; -\frac{n}{2}\left\{\log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}}\mathbf{\Gamma}| + \log|\mathbf{S}_{\mathbf{X}}|\right\}. \tag{A.16}$$

Therefore, the final objective function for $\mathbf{\Gamma}$ is

$$\begin{aligned} L_n(\mathbf{\Gamma}) &= C_n(\mathbf{\Gamma}) + M_n(\mathbf{\Gamma}) \\ &= -\frac{n}{2}\left\{1 + \log(s^2_{\mathbf{\Gamma}}) + \log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}}\mathbf{\Gamma}| + \log|\mathbf{S}_{\mathbf{X}}|\right\} \\ &= -\frac{n}{2}\log\left\{s^2_{Y(W)} - \mathbf{S}^T_{\mathbf{X}Y(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}Y(W)}\right\} \\ &\quad -\frac{n}{2}\left\{\log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}}\mathbf{\Gamma}|\right\} + \text{constant}. \end{aligned} \tag{A.17}$$

Alternatively, if we assume $\mathbf{X}|W \sim N(\boldsymbol{\mu}_{\mathbf{X}}, W^{-1}\mathbf{\Sigma})$, $\mathbf{\Sigma} > 0$ and use the fact that $\mathrm{E}(W) = 1$, then $\boldsymbol{\mu}_{\mathbf{X}} = \mathrm{E}(W\mathbf{X}) = \mathrm{E}(\mathbf{X})$. Moreover, the covariance

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{X}(W)} &= \mathrm{E}\left[W(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T\right] \\ &= \mathrm{E}\left\{\mathrm{E}\left[W(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})|W\right]\right\} \\ &= \mathrm{E}\left\{W \cdot \mathrm{var}(\mathbf{X}|W)\right\} = \mathrm{E}(W \cdot W^{-1}\mathbf{\Sigma}) \\ &= \mathbf{\Sigma}. \end{aligned}$$

The maximum likelihood estimators for $\boldsymbol{\mu}_{\mathbf{X}}$ and $\mathbf{\Sigma}$ are then $\widehat{\boldsymbol{\mu}}_{\mathbf{X}} = \overline{\mathbf{X}}_{(W)}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{X}} = \mathbf{S}_{\mathbf{X}(W)}$. Follow the same calculation in the proof of Proposition 3, we have

$$M_n(\mathbf{\Gamma}) = -\frac{n}{2}\left\{\log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\mathbf{\Gamma}| + \log|\mathbf{S}_{\mathbf{X}(W)}|\right\}.$$

Therefore, the final objective function for $\boldsymbol{\Gamma}$ up to a constant is

$$\begin{aligned} L_n(\boldsymbol{\Gamma}) &= C_n(\boldsymbol{\Gamma}) + M_n(\boldsymbol{\Gamma}) \\ &= -\frac{n}{2}\left\{\log(s^2_{\boldsymbol{\Gamma}}) + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\boldsymbol{\Gamma}|\right\} \\ &= -\frac{n}{2}\log\left\{s^2_{Y(W)} - \mathbf{S}^T_{\mathbf{X}Y(W)}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}Y(W)}\right\} \\ &\quad -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\boldsymbol{\Gamma}|\right\} \\ &= -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\left(\mathbf{S}_{\mathbf{X}(W)} - \mathbf{S}_{\mathbf{X}Y(W)}s^{-2}_{Y(W)}\mathbf{S}^T_{\mathbf{X}Y(W)}\right)\boldsymbol{\Gamma}|\right\} \\ &:= -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}|Y(W)}\boldsymbol{\Gamma}|\right\}. \end{aligned} \tag{A.18}$$

A.3 Envelopes for Cox regression

Continuing from the discussion of Section A.3, the log partial likelihood for $\boldsymbol{\beta}$ is

$$C_n(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_i\left\{\boldsymbol{\beta}^T\mathbf{Z}_i - \log\left(\sum_{j=1}^n I(T_j \geq T_i)\exp(\boldsymbol{\beta}^T\mathbf{Z}_j)\right)\right\}, \tag{A.19}$$

whose first-order derivative (score function) and second-order derivative (Hessian matrix) are

$$\begin{aligned} C'_n(\boldsymbol{\beta}) &= \sum_{i=1}^n \delta_i\left\{\mathbf{Z}_i - \sum_{j=1}^n W_{ij}\mathbf{Z}_j\right\}, \\ C''_n(\boldsymbol{\beta}) &= -\sum_{i=1}^n \delta_i\left\{\left(\sum_{j=1}^n W_{ij}\mathbf{Z}_j\mathbf{Z}_j^T\right) - \left(\sum_{j=1}^n W_{ij}\mathbf{Z}_j\right)\left(\sum_{j=1}^n W_{ij}\mathbf{Z}_j^T\right)\right\}, \end{aligned}$$

where $W_{ij} = I(T_j \geq T_i)\exp(\boldsymbol{\beta}^T\mathbf{Z}_j)/\left\{\sum_{j=1}^n I(T_j \geq T_i)\exp(\boldsymbol{\beta}^T\mathbf{Z}_j)\right\}$ is the weight with respect to subject i. The Fisher information matrix is $\mathbf{J}(\boldsymbol{\beta}) = -\lim_{n\to\infty} n^{-1}C''_n(\boldsymbol{\beta})$. The implementation of envelope Cox model is similar to that of envelope GLM in Section 3.2 by adding the same $M_n(\boldsymbol{\Gamma})$ to the objective function. Theoretical properties of the envelope Cox model is included in Section 3.1.

B Additional simulations: least squares regression

In this section, we report the results of numerical studies on the robustness of three estimators in linear regression: (1) the ordinary least squares (OLS) estimator, (2) the PLS estimator from the SIMPLS algorithm, and (3) the moment-based envelope estimator from the 1D algorithm in

t_6	$\mathbf{\Sigma_X}$			$0.01\mathbf{\Sigma_X}$		
	Standard	SIMPLS	1D	Standard	SIMPLS	1D
$n = 50$	57	34	6	85	44	48
S.E.	1.6	0.08	0.003	0.6	1.2	2.7
$n = 200$	40	32	3	80	35	14
S.E.	1.8	0.8	0.09	1.0	0.8	0.6
$n = 800$	23	27	1	71	30	7
S.E.	1.4	1.0	0.5	1.5	0.8	0.2

$U(0, 1)$	$\mathbf{\Sigma_X}$			$0.01\mathbf{\Sigma_X}$		
	Standard	SIMPLS	1D	Standard	SIMPLS	1D
$n = 50$	27	33	1.8	74	33	12.6
S.E.	2	1	0.05	2	1	0.5
$n = 200$	14	29	0.8	59	31	5.4
S.E.	1	1	0.03	2	1	0.2
$n = 800$	6.8	24	0.4	44	25	2.8
S.E.	0.4	1	0.01	2	1	0.1

χ^2_4	$\mathbf{\Sigma_X}$			$0.01\mathbf{\Sigma_X}$		
	Standard	SIMPLS	1D	Standard	SIMPLS	1D
$n = 50$	73	35	13	88	60	75
S.E.	1.4	1.0	0.8	0.3	2	2
$n = 200$	57	33	6.0	85	47	52
S.E.	2	1	0.2	1	1	2
$n = 800$	38	27	2.9	79	35	13
S.E.	2	1	0.1	1	1	1

Table 7: Simulations for heterogeneous covariance matrix $\mathbf{\Sigma_X} = \mathbf{\Gamma\Omega\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$. Averaged angles between the true parameter vector and three estimators for the least squares simulation. Standard error is included below each value.

t_6		Standard	SIMPLS	1D
$n = 50$	$\angle$	27.8 (0.7)	33.3 (0.7)	30.9 (0.9)
	LR	1.32	0.79	1.22
$n = 200$	$\angle$	13.1 (0.3)	18.1 (0.4)	14.0 (0.4)
	LR	1.06	0.91	1.05
$n = 800$	$\angle$	6.6 (0.2)	9.0 (0.2)	6.7 (0.2)
	LR	1.02	0.98	1.02

$U(0, 1)$		Standard	SIMPLS	1D
$n = 50$	$\angle$	6.8 (0.2)	22.2 (0.5)	6.9 (0.2)
	LR	1.02	0.71	1.01
$n = 200$	$\angle$	3.3 (0.08)	11.7 (0.3)	3.3 (0.08)
	LR	1.00	0.89	1.00
$n = 800$	$\angle$	1.7 (0.04)	5.9 (0.1)	1.7 (0.04)
	LR	1.00	0.97	1.00

χ^2_4		Standard	SIMPLS	1D
$n = 50$	$\angle$	51 (1.4)	50 (1.2)	53 (1.4)
	LR	1.42	1.21	1.39
$n = 200$	$\angle$	29.9 (0.8)	30.6 (0.7)	32.6 (0.9)
	LR	1.39	1.16	1.28
$n = 800$	$\angle$	15.0 (0.4)	16.2 (0.4)	15.9 (0.4)
	LR	1.08	1.04	1.07

Table 8: Simulations for isotropic covariance matrix $\mathbf{\Sigma_X} = 0.01\mathbf{I}_p$. Averaged angles and length ratios between the true parameter vector and various estimators. Standard errors of the averaged angles are included in parentheses.

Cook and Zhang (2014a). Same as in Section 5, for each of the simulation settings, we simulated 100 datasets for each of the three sample sizes: $n = 50$, 200 and 800. The true dimension of the envelope was always used in estimation. We used the Matlab function *plsregress* to compute the PLS estimator. From Proposition 4.1 in Cook et al. (2013), we know that the SIMPLS algorithm produces $\sqrt{n}$-consistent envelope estimators when the number of component in the algorithm is chosen to be the dimension of the envelope. The univariate response was generated as $Y = \boldsymbol{\beta}^T\mathbf{X} + \varepsilon$, where the elements of ε were $U(0, 1)$ variates. The covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, was generated as it was for logistic and the Poisson regression, except the eigenvalues of $\boldsymbol{\Omega}$ were 1 and 5, while the eigenvalues of $\boldsymbol{\Omega}_0$ ranged from 0.0002 to 20. These choices introduce a degree of collinearity in the immaterial variation $\mathrm{var}(\boldsymbol{\Gamma}_0^T\mathbf{X})$, which is the kind of regression that PLS estimators were designed to handle. We also used $0.01\boldsymbol{\Sigma}_{\mathbf{X}}$ instead of $\boldsymbol{\Sigma}_{\mathbf{X}}$ to compare performance with a weaker signal.

The angles are shown in Table 7, the lengths being excluded because they produced no comparative conclusion beyond those available from the results shown. The 1D algorithm did the best in all situations, except when a relatively small sample sizes was combined with a weak signal ($0.01\boldsymbol{\Sigma}_{\mathbf{X}}$) and t_6 or χ^2_4 errors. In those situations the SIMPLS might have a slight edge. Because there was strong collinearity in $\mathbf{X}$, OLS did notably worse than the other two methods, as expected. One exception to this was when $n = 800$ with uniformly distributed predictors.

We also repeated the simulations of Table 7 when $\boldsymbol{\Sigma}_{\mathbf{X}} = 0.01\mathbf{I}_p$, so envelope methods reduce to the standard OLS estimator. The OLS and 1D estimators performed essentially the same in all cases, while the SIMPLS algorithm did a bit worse with t_6 and χ^2_4 errors, and notably worse with $U(0, 1)$ errors. Table 8 gives the results of the least squares simulation with predictor covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}} = 0.01\mathbf{I}_p$.

C Implementation details for envelope GLM in Section 3.2

Let $\mathbb{X}_n$ be the $n \times p$ data matrix of $\mathbf{X}_i$'s and let $\mathbb{C}_n$ be the $n \times 1$ data matrix of $\mathcal{C}'(\vartheta_i)$, where $\mathcal{C}'(\vartheta)$ is the first order derivative from Table 1. Then the $p \times u$ gradient matrix of $\partial C_n(\boldsymbol{\Gamma})/\partial\boldsymbol{\Gamma}$ can be computed conveniently as follows.

Lemma 1.

$$\begin{aligned}
\frac{\partial C_n(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} &= \mathbb{X}_n^T \mathbb{C}_n \boldsymbol{\eta}^T + \mathbf{S}_{\mathbf{X}V(W)} \mathbb{C}_n^T \mathbb{X}_n \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \\
&\quad - \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbb{X}_n^T \mathbb{C}_n \mathbf{S}_{\mathbf{X}V(W)}^T \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \\
&\quad - \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}V(W)} \mathbb{C}_n^T \mathbb{X}_n \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1}; \\
\frac{\partial M_n(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} &= -n \left\{ \mathbf{S}_{\mathbf{X}} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}} \mathbf{\Gamma})^{-1} + \mathbf{S}_{\mathbf{X}}^{-1} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}}^{-1} \mathbf{\Gamma})^{-1} \right\}.
\end{aligned}$$

To check this Lemma, we compared the analytical form of the derivatives to the numerical derivatives. The numerical derivatives of a function $f(\mathbf{\Gamma})$ is computed by varying each element of $\mathbf{\Gamma}$ by a small number $\delta = 1.0 \times 10^{-6}$ and evaluate the function:

$$\left[\frac{\partial f(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} \right]_{ij} = \frac{f(\mathbf{\Gamma} + \delta \mathbf{G}_{ij}) - f(\mathbf{\Gamma} - \delta \mathbf{G}_{ij})}{2\delta}, \tag{C.1}$$

where $\mathbf{G}_{ij} \in \mathbb{R}^{p \times u}$ is zero matrix except its ij-th entry is one. We found the difference between our analytical derivative and the numerical derivative is around 10^{-7} for every element.

Proof. This proof of Lemma 1 mainly involves rather complicated matrix differentiation.

$$\begin{aligned}
\frac{\partial \vartheta_i}{\partial \mathrm{vec}(\mathbf{\Gamma})} &= \frac{\partial}{\partial \mathrm{vec}(\mathbf{\Gamma})} \left\{ \mathbf{X}_i^T \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}V(W)} \right\} \\
&= \frac{\partial}{\partial \mathrm{vec}(\mathbf{\Gamma})} \left\{ \left(\mathbf{S}_{\mathbf{X}V(W)}^T \otimes \mathbf{X}_i^T \right) \mathrm{vec}[\mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T] \right\} \\
&= \left[\frac{\partial}{\partial \mathrm{vec}(\mathbf{\Gamma})} \mathrm{vec}(\mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T) \right]^T \left(\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i \right).
\end{aligned} \tag{C.2}$$

Apply the following chain rule:

$$\partial \mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^T \mathbf{B}^T \otimes \mathbf{I}) \partial \mathrm{vec}(\mathbf{A}) + (\mathbf{C}^T \otimes \mathbf{A}) \partial \mathrm{vec}(\mathbf{B}) + (\mathbf{I} \otimes \mathbf{AB}) \partial \mathrm{vec}(\mathbf{C}),$$

we get

$$\begin{aligned}
\partial \mathrm{vec}(\mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T) / \partial \mathrm{vec}(\mathbf{\Gamma}) &= \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 \\
&= \left[\mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \otimes \mathbf{I}_p \right] \mathbf{I}_{pu} \\
&\quad + (\mathbf{\Gamma} \otimes \mathbf{\Gamma}) \frac{\partial \mathrm{vec}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1}}{\partial \mathrm{vec}(\mathbf{\Gamma})} \\
&\quad + \left[\mathbf{I}_p \otimes \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \right] \mathbf{K}_{pu}.
\end{aligned}$$

The second term after the last equal sign is computed by noticing that

$$\begin{aligned}
\partial \mathrm{vec} \left\{ (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \right\} &= -\mathrm{vec} \left\{ (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \left[\partial (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma}) \right] (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \right\} \\
&= -\left((\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \otimes (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \right) \partial \mathrm{vec}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma}),
\end{aligned}$$

and that

$$\begin{aligned}\frac{\partial \mathrm{vec}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})}{\partial \mathrm{vec}(\mathbf{\Gamma})} &= (\mathbf{I}_u \otimes \mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)})\frac{\partial \mathrm{vec}(\mathbf{\Gamma})}{\partial \mathrm{vec}(\mathbf{\Gamma})} + (\mathbf{\Gamma}^T \otimes \mathbf{I}_u)\frac{\partial \mathrm{vec}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)})}{\partial \mathrm{vec}(\mathbf{\Gamma})} \\ &= (\mathbf{I}_u \otimes \mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}) + (\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)} \otimes \mathbf{I}_u)\mathbf{K}_{pu}.\end{aligned}$$

Hence the second term equals to

$$\begin{aligned}\mathbf{T}_2 &= -\left(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1} \otimes \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\right) \\ &\quad -\left(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)} \otimes \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\right)\mathbf{K}_{pu}.\end{aligned}$$

Finally, substitute the three terms into (C.2), we have the three terms from

$$\begin{aligned}\frac{\partial \vartheta_i}{\partial \mathrm{vec}(\mathbf{\Gamma})} &= (\mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3)^T\left(\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i\right) \\ \mathbf{T}_1^T\left(\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i\right) &= \left[(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i\right] \\ &= [\boldsymbol{\eta} \otimes \mathbf{X}_i] \\ \mathbf{T}_3^T\left(\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i\right) &= \left\{\left[\mathbf{I}_p \otimes \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\right]\mathbf{K}_{pu}\right\}^T\left(\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i\right) \\ &= \mathbf{K}_{pu}^T\left[\mathbf{S}_{\mathbf{X}V(W)} \otimes (\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\right]\end{aligned}$$

$$\begin{aligned}\mathbf{T}_2^T\left(\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i\right) &= -\left[(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\right] \\ &\quad -\mathbf{K}_{pu}^T\left[\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}V(W)} \otimes (\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\right] \\ &= -\left[\boldsymbol{\eta} \otimes \mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\right] \\ &\quad -\mathbf{K}_{pu}^T\left[\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}\boldsymbol{\eta} \otimes (\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\right]\end{aligned}$$

Notice that each of the above terms is a $pu \times 1$ vector, and then the $\mathbf{K}_{pu}^T$ serves only to change $\mathrm{vec}(\mathbf{A}) \to \mathrm{vec}(\mathbf{A}^T)$, therefore

$$\begin{aligned}\frac{\partial \vartheta_i}{\partial \mathrm{vec}(\mathbf{\Gamma})} &= \mathrm{vec}(\mathbf{M}_1 + \mathbf{M}_3^T) + \mathrm{vec}(\mathbf{M}_2 + \mathbf{M}_4^T) \\ \mathbf{M}_1 &= \left[\mathbf{X}_i \cdot \boldsymbol{\eta}^T\right] \\ \mathbf{M}_3 &= \left[(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i \cdot \mathbf{S}_{\mathbf{X}V(W)}^T\right] \\ \mathbf{M}_2 &= -\left[\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i \cdot \boldsymbol{\eta}^T\right] \\ \mathbf{M}_4 &= -\left[(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i \cdot \boldsymbol{\eta}^T\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\right],\end{aligned}$$

where $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3^T$, $\mathbf{M}_4^T$ are all $p \times u$ matrices.

To implement, we have the gradient matrix as

$$\frac{\partial C_n(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} + \frac{\partial M_n(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} = \frac{\partial C_n(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} - n\{\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma})^{-1} + \mathbf{S}_{\mathbf{X}}^{-1}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{\Gamma})^{-1}\}, \tag{C.3}$$

where $\frac{\partial C_n(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}}$ is re-arranged as a $p \times u$ matrix

$$\begin{aligned}
\frac{\partial C_n(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} &= \left\{ \sum_{i=1}^{n} \mathcal{C}'(\vartheta_i) \frac{\partial \vartheta_i}{\partial \mathrm{vec}(\mathbf{\Gamma})} \right\}_{p\times u} \\
&= \sum_{i=1}^{n} \mathcal{C}'(\vartheta_i)\{\mathbf{X}_i \cdot \boldsymbol{\eta}^T + \mathbf{S}_{\mathbf{X}V(W)} \cdot \mathbf{X}_i^T\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\} \\
&\quad - \sum_{i=1}^{n} \mathcal{C}'(\vartheta_i) \left[\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i \cdot \boldsymbol{\eta}^T\right] \\
&\quad - \sum_{i=1}^{n} \mathcal{C}'(\vartheta_i) \left[(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i \cdot \boldsymbol{\eta}^T\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\right]^T \\
&= \mathbb{X}_n^T\mathbb{C}_n\boldsymbol{\eta}^T + \mathbf{S}_{\mathbf{X}V(W)}\mathbb{C}_n^T\mathbb{X}_n\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1} \\
&\quad - \mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbb{X}_n^T\mathbb{C}_n\boldsymbol{\eta}^T \\
&\quad - \mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}\boldsymbol{\eta}\mathbb{C}_n^T\mathbb{X}_n\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}.
\end{aligned}$$

□

D Enveloping a matrix-valued parameter

Definition 3 is sufficiently general to cover a matrix-valued parameter $\boldsymbol{\phi} \in \mathbb{R}^{r\times c}$ by considering the $\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\phi}}))$-envelope of $\mathrm{vec}(\boldsymbol{\phi})$. However, matrix-valued parameters come with additional structure that is often desirable to maintain during estimation. For instance, in the multivariate linear model reviewed in Section 1.3, envelope construction was constrained to reflect separate row and column reduction of the matrix parameter $\boldsymbol{\beta}$. The advantages of row and column reductions have been discussed by Li et al. (2010) and Hung et al. (2012). Although our primary focus is on vector-valued parameters, in this section we indicate how to adapt for a matrix-valued parameter.

Suppose that $\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_t)$ converges to a matrix normal distribution with mean 0, column variance $\mathbf{\Delta}_L \in \mathbb{S}^{r\times r}$ and row variance $\mathbf{\Delta}_R \in \mathbb{S}^{c\times c}$. (See Dawid (1981) for background on the matrix normal distribution.) Then

$$\sqrt{n}(\mathrm{vec}(\widehat{\boldsymbol{\phi}}) - \mathrm{vec}(\boldsymbol{\phi}_t)) \to N(0, \mathbf{\Delta}_R \otimes \mathbf{\Delta}_L), \tag{D.1}$$

and direct application of Definition 3 yields the envelope $\mathcal{E}_{\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L}(\mathrm{vec}(\boldsymbol{\phi}_t))$. However, this envelope may not preserve the intrinsic row-column structure of $\boldsymbol{\phi}_t$. In the following definition we introduce a restricted class of envelopes that maintain the matrix structure of $\boldsymbol{\phi}_t$. We define the Kronecker product of two subspaces as $\mathcal{A} \otimes \mathcal{B} = \{\mathbf{a} \otimes \mathbf{b} | \mathbf{a} \in \mathcal{A};\ \mathbf{b} \in \mathcal{B}\}$.

Definition 4. *Assume that $\widehat{\boldsymbol{\phi}}$ is asymptotically matrix normal as give in (D.1). Then the tensor envelope for $\boldsymbol{\phi}_t$, denoted by $\mathcal{K}_{\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t)$, is defined as the intersection of all reducing subspaces $\mathcal{E}$ of $\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L$ that contain $\mathrm{span}(\mathrm{vec}(\boldsymbol{\phi}_t))$ and can be written as $\mathcal{E} = \mathcal{E}_R \otimes \mathcal{E}_L$ with $\mathcal{E}_R \subseteq \mathbb{R}^c$ and $\mathcal{E}_L \subseteq \mathbb{R}^r$.*

We see from this definition that $\mathcal{K}_{\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t)$ always exist and is the smallest subspace with the required properties. Let $\mathbf{L} \in \mathbb{R}^{r \times d_c}$, $d_c < r$, and $\mathbf{R} \in \mathbb{R}^{c \times d_r}$, $d_r < c$ be semi-orthogonal matrices such that $\mathcal{K}_{\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t) = \mathrm{span}(\mathbf{R} \otimes \mathbf{L})$. By definition, $\mathrm{span}(\boldsymbol{\phi}_t) \subseteq \mathrm{span}(\mathbf{L})$ and $\mathrm{span}(\boldsymbol{\phi}_t^T) \subseteq \mathrm{span}(\mathbf{R})$. Hence we have $\boldsymbol{\phi}_t = \mathbf{L}\boldsymbol{\eta}\mathbf{R}^T$ and $\mathrm{vec}(\boldsymbol{\phi}_t) = (\mathbf{R} \otimes \mathbf{L})\mathrm{vec}(\boldsymbol{\eta})$ for some $\boldsymbol{\eta} \in \mathbb{R}^{d_c \times d_r}$.

The next proposition shows how to factor $\mathcal{K}_{\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t) \in \mathbb{R}^{rc}$ into the tensor product of envelopes $\mathcal{E}_{\boldsymbol{\Delta}_R}(\boldsymbol{\phi}_t^T) \in \mathbb{R}^r$ and $\mathcal{E}_{\boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t) \in \mathbb{R}^c$ for the row and column spaces of $\boldsymbol{\phi}_t$. These tensor factors are envelopes in smaller spaces that preserve the row and column structure and can facilitate analysis and interpretation.

Proposition 7. $\mathcal{K}_{\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t) = \mathcal{E}_{\boldsymbol{\Delta}_R}(\boldsymbol{\phi}_t^T) \otimes \mathcal{E}_{\boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t)$.

For example, in reference to model (1.1), the distribution of $\widehat{\boldsymbol{\beta}} = \mathbf{S}_{\mathbf{XY}}^T \mathbf{S}_{\mathbf{X}}^{-1}$ satisfies (D.1) with $\boldsymbol{\Delta}_R = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ and $\boldsymbol{\Delta}_L = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. The tensor envelope is then

$$\mathcal{K}_{\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}}(\boldsymbol{\beta}^T) \otimes \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}).$$

If we are interested in reducing only the column space of $\boldsymbol{\beta}$, which corresponds to response reduction, we would use $\mathbb{R}^p \otimes \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$ for constructing an envelope estimator of $\boldsymbol{\beta}$, and then $\boldsymbol{\beta} = \mathbf{L}\boldsymbol{\eta}\mathbf{I}_p = \mathbf{L}\boldsymbol{\eta}$ where $\mathbf{L}$ is a semi-orthogonal basis for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$, which reproduces the envelope construction in Cook et al. (2010). Similarly, if we are interested in only predictor reduction, we would take $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T) \otimes \mathbb{R}^r$, which reproduces the envelope construction in Cook et al. (2013). More generally, the definition of the tensor envelope and Proposition 7 straightforwardly connects and combines the envelope models in the predictor space and in the response space, leading to the simultaneous envelope method in Cook and Zhang (2014b).

E Proofs

E.1 Proposition 5 and Corollary 2

Proof. The asymptotic normality and the asymptotic covariance of $\widehat{\boldsymbol{\phi}}_n$ can be see by definition. We write the Fisher information $\mathbf{F}(\boldsymbol{\theta}_t)$ as its natural block matrix structure

$$\mathbf{F}(\boldsymbol{\theta}_t) = \mathbf{F}(\boldsymbol{\psi}_t, \boldsymbol{\phi}_t) = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix},$$

where $\mathbf{A} = -\mathrm{E}_{\boldsymbol{\theta}_t}\left\{\left(\partial^2 L(\boldsymbol{\psi}, \boldsymbol{\phi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T\right)\right\}$, $\mathbf{B} = -\mathrm{E}_{\boldsymbol{\theta}_t}\left\{\left(\partial^2 L(\boldsymbol{\psi}, \boldsymbol{\phi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\phi}^T\right)\right\}$ and $\mathbf{C} = -\mathrm{E}_{\boldsymbol{\theta}_t}\left\{\left(\partial^2 L(\boldsymbol{\psi}, \boldsymbol{\phi})/\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T\right)\right\}$. Then $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t) = \mathbf{F}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t) = (\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1}$ by block matrix inversion.

The Hessian matrix for the constrained optimization (4.2)-(4.4) is

$$\mathbf{F}(\boldsymbol{\psi}_t, \boldsymbol{\eta}_t) = \begin{pmatrix} \mathbf{A} & \mathbf{B}\boldsymbol{\Gamma} \\ \boldsymbol{\Gamma}^T\mathbf{B}^T & \boldsymbol{\Gamma}^T\mathbf{C}\boldsymbol{\Gamma} \end{pmatrix}, \tag{E.1}$$

since $\partial/\partial\boldsymbol{\eta} = \boldsymbol{\Gamma}^T \cdot \partial/\partial\boldsymbol{\phi}$ under the constraint that $\boldsymbol{\phi} = \boldsymbol{\Gamma}\boldsymbol{\eta}$. Therefore the asymptotic variance for $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$ is

$$\begin{aligned} \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}) &= \boldsymbol{\Gamma}\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})\boldsymbol{\Gamma}^T \\ &= \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{C}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T \\ &= \boldsymbol{\Gamma}\left\{\boldsymbol{\Gamma}^T\mathbf{V}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\boldsymbol{\Gamma}\right\}^{-1}\boldsymbol{\Gamma}^T \\ &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}\mathbf{P}_{\boldsymbol{\Gamma}} - \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0\mathbf{V}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\mathbf{V}\mathbf{P}_{\boldsymbol{\Gamma}}, \end{aligned}$$

where the last equality is obtained by applying the equality (Cook and Forzani 2009),

$$(\boldsymbol{\Gamma}^T\mathbf{V}^{-1}\boldsymbol{\Gamma})^{-1} = \boldsymbol{\Gamma}^T\mathbf{V}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^T\mathbf{V}\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\mathbf{V}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\mathbf{V}\boldsymbol{\Gamma}. \tag{E.2}$$

Then by requiring that $\mathrm{span}(\boldsymbol{\Gamma})$ reduces $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$, we have $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}) = \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}\mathbf{P}_{\boldsymbol{\Gamma}}$.

Similarly for the asymptotic variance of $\widehat{\boldsymbol{\psi}}_n$ and $\widehat{\boldsymbol{\psi}}_{\boldsymbol{\Gamma}}$ we have the following results from block matrix inversion and (E.2).

$$\begin{aligned} \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\psi}}_{\boldsymbol{\Gamma}}) &= \left(\mathbf{A} - \mathbf{B}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{C}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{B}^T\right)^{-1} \\ &\leq \left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T\right)^{-1} = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\psi}}_n). \end{aligned}$$

□

E.2 Proposition 6

From Cook and Zhang (2014a; Proposition 6), we know that $\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}^T$ is $\sqrt{n}$-consistent for the projection onto the envelope $\mathcal{E}_{\mathbf{V}_{\phi\phi}}(\boldsymbol{\phi}_t)$. Then the proof follows from the proof of $\sqrt{n}$-consistency of the k-th direction in the 1D algorithm, where we replace the sample objective function $J_k(\mathbf{g})$ with $L_n(\boldsymbol{\psi}, \widehat{\mathbf{\Gamma}}\boldsymbol{\eta})$.

E.3 Proposition 2

Proof. The envelope of interest is $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\boldsymbol{\beta}_t)$. The Fisher information per observation is then obtained by taking the expectation of the second derivative of $\log f(y|\boldsymbol{\alpha}, \boldsymbol{\beta}^T\mathbf{X}) + \log g(x|\boldsymbol{\psi})$ with respect to $\boldsymbol{\theta}$. Clearly, the information matrix is block diagonal in $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\boldsymbol{\psi}$ so we need only the second derivatives of $\log f(y|\boldsymbol{\alpha}, \boldsymbol{\beta}^T\mathbf{X})$. Then the Fisher information matrix for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ can be written unambiguously in the form

$$\mathbf{F}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \mathrm{E}\begin{pmatrix} \mathbf{A}(Y, \boldsymbol{\beta}_t^T\mathbf{X}, \boldsymbol{\alpha}_t) & \mathbf{a}(Y, \boldsymbol{\beta}_t^T\mathbf{X}, \boldsymbol{\alpha}_t)\mathbf{X}^T \\ \mathbf{X}\mathbf{a}^T(Y, \boldsymbol{\beta}_t^T\mathbf{X}, \boldsymbol{\alpha}_t) & c(Y, \boldsymbol{\beta}_t^T\mathbf{X}, \boldsymbol{\alpha}_t)\mathbf{X}\mathbf{X}^T \end{pmatrix},$$

where $\mathbf{A}$ is the second derivative matrix with respect to $\boldsymbol{\alpha}$, the off diagonal blocks are the cross derivatives $\partial^2/\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta}^T$ with all factors collected into $\mathbf{a}$ except $\mathbf{X}$, and $c(Y, \boldsymbol{\beta}_t^T\mathbf{X}, \boldsymbol{\alpha}_t)\mathbf{X}\mathbf{X}^T$ is the second derivative with respect to $\boldsymbol{\beta}$. The asymptotic covariance matrix for $\widehat{\boldsymbol{\beta}}_n$ is then

$$\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \{\mathrm{E}(c\mathbf{X}\mathbf{X}^T) - \mathrm{E}(\mathbf{X}\mathbf{a}^T)\mathrm{E}^{-1}(\mathbf{A})\mathrm{E}(\mathbf{a}\mathbf{X}^T)\}^{-1}.$$

Since $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\mathbf{V}_{\beta\beta}^{-1}}(\boldsymbol{\beta}_t)$, it is sufficient to consider, making the additional assumption that $E(\mathbf{X}|\mathbf{\Gamma}^T\mathbf{X})$ is a linear function of $\mathbf{X}$ and requiring that $\mathrm{span}(\boldsymbol{\beta}_t) \subseteq \mathrm{span}(\mathbf{\Gamma})$,

$$\begin{aligned}
\mathbf{P}_{\mathbf{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}_t)\mathbf{Q}_{\mathbf{\Gamma}} &= \mathbf{P}_{\mathbf{\Gamma}}\{\mathrm{E}(c\mathbf{X}\mathbf{X}^T) - \mathrm{E}(\mathbf{X}\mathbf{a}^T)\mathrm{E}^{-1}(\mathbf{A})\mathrm{E}(\mathbf{a}\mathbf{X}^T)\}\mathbf{Q}_{\mathbf{\Gamma}} \\
&= \mathbf{P}_{\mathbf{\Gamma}}\{\mathrm{E}(c\mathbf{\Gamma}\mathbf{\Gamma}^T\mathbf{X}\cdot\mathbf{X}^T) - \mathrm{E}(\mathbf{X}\mathbf{a}^T)\mathrm{E}^{-1}(\mathbf{A})\mathrm{E}(\mathbf{a}\mathbf{X}^T)\}\mathbf{Q}_{\mathbf{\Gamma}} \\
&= \mathbf{P}_{\mathbf{\Gamma}}\{\mathrm{E}(c\mathbf{X}\mathrm{E}(\mathbf{X}^T|Y, \mathbf{\Gamma}^T\mathbf{X})) - \mathrm{E}(\mathbf{X}\mathbf{a}^T)\mathrm{E}^{-1}(\mathbf{A})\mathrm{E}(\mathbf{a}\mathrm{E}(\mathbf{X}^T|Y, \mathbf{\Gamma}^T\mathbf{X}))\}\mathbf{Q}_{\mathbf{\Gamma}}.
\end{aligned}$$

This follows since c and $\mathbf{a}$ are functions of only Y and $\boldsymbol{\beta}^T\mathbf{X}$ and $\text{span}(\boldsymbol{\beta}_t) \subseteq \text{span}(\boldsymbol{\Gamma})$. Next, using the fact that we are in a regression and thus $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}$ implies $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\Gamma}^T\mathbf{X}$, we have

$$\begin{aligned}
\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}\mathbf{Q}_{\boldsymbol{\Gamma}} &= \mathbf{P}_{\boldsymbol{\Gamma}}\{\text{E}(c\mathbf{X}\text{E}(\mathbf{X}^T|\boldsymbol{\Gamma}^T\mathbf{X})) - \text{E}(\mathbf{X}\mathbf{a}^T)\text{E}^{-1}(\mathbf{A})\text{E}(\mathbf{a}\text{E}(\mathbf{X}^T|\boldsymbol{\Gamma}^T\mathbf{X}))\}\mathbf{Q}_{\boldsymbol{\Gamma}} \\
&= \mathbf{P}_{\boldsymbol{\Gamma}}\{\text{E}(c\mathbf{X}\mathbf{X}^T\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_\mathbf{X})}) - \text{E}(\mathbf{X}\mathbf{a}^T)\text{E}^{-1}(\mathbf{A})\text{E}(\mathbf{a}\mathbf{X}^T\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_\mathbf{X})})\}\mathbf{Q}_{\boldsymbol{\Gamma}} \\
&= \mathbf{P}_{\boldsymbol{\Gamma}}\{\text{E}(c\mathbf{X}\mathbf{X}^T) - \text{E}(\mathbf{X}\mathbf{a}^T)\text{E}^{-1}(\mathbf{A})\text{E}(\mathbf{a}\mathbf{X}^T)\}\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_\mathbf{X})}\mathbf{Q}_{\boldsymbol{\Gamma}} \\
&= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_\mathbf{X})}\mathbf{Q}_{\boldsymbol{\Gamma}} \\
&= \boldsymbol{\Gamma}\{\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}\boldsymbol{\Gamma}\}\{\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_\mathbf{X}\boldsymbol{\Gamma}\}^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_\mathbf{X}\mathbf{Q}_{\boldsymbol{\Gamma}}.
\end{aligned}$$

The second equality above uses the linearity condition and the rest is algebra. Since $\boldsymbol{\Gamma}$ has full column rank and the next two matrix factors in $\{\cdot\}$ are positive definite, we see that $\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}\mathbf{Q}_{\boldsymbol{\Gamma}} = 0$ if and only if $\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_\mathbf{X}\mathbf{Q}_{\boldsymbol{\Gamma}} = 0$; that is, if and only if $\text{span}(\boldsymbol{\Gamma})$ reduces $\boldsymbol{\Sigma}_\mathbf{X}$. Consequently, $\text{span}(\boldsymbol{\Gamma})$ reduces $\mathbf{V}_{\boldsymbol{\beta\beta}}$ if and only if $\text{span}(\boldsymbol{\Gamma})$ reduces $\boldsymbol{\Sigma}_\mathbf{X}$. Note that in this proof we did not require $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be orthogonal parameters, nor did we require an exponential family regression. The only requirements are the regression structure and the linearity condition. □

E.4 Proposition 3

Proof. The multivariate normal log-likelihood up to a constant is

$$M_n(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X}) = -\frac{n}{2}\{\log|\boldsymbol{\Sigma}_\mathbf{X}| + \text{trace}[\boldsymbol{\Sigma}_\mathbf{X}^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X})^T/n]\}, \tag{E.3}$$

Since $\text{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_t)$ reduces $\boldsymbol{\Sigma}_\mathbf{X}$, we need to maximize $M_n(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X})$ under the constraint that $\text{span}(\boldsymbol{\Gamma})$ reduces $\boldsymbol{\Sigma}_\mathbf{X}$. This can be done by substitute $\boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ into the objective function, where $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are two symmetric positive definite matrices. The resulting objective function $M_n(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Gamma})$ can be first maximized over $\boldsymbol{\mu}_\mathbf{X}$ to get $\widehat{\boldsymbol{\mu}}_\mathbf{X} = \overline{\mathbf{X}}$. Then the partially maximized objective function can be expressed as

$$\begin{aligned}
-\frac{2}{n}M_n(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Gamma}) &= \log|\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T| + \text{trace}[(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)^{-1}\mathbf{S}_\mathbf{X}] \\
&= \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0| + \text{trace}\{(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)\mathbf{S}_\mathbf{X}\} \\
&= \log|\boldsymbol{\Omega}| + \text{trace}(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_\mathbf{X}) + \log|\boldsymbol{\Omega}_0| + \text{trace}\{(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\mathbf{S}_\mathbf{X}) \\
&= \log|\boldsymbol{\Omega}| + \text{trace}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_\mathbf{X}\boldsymbol{\Gamma}) + \log|\boldsymbol{\Omega}_0| + \text{trace}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\mathbf{S}_\mathbf{X}\boldsymbol{\Gamma}_0).
\end{aligned}$$

For any symmetric positive definite matrix $\mathbf{M}$, we apply the following result to optimize over $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ in $M_n(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Gamma})$,

$$\arg\min_{\mathbf{A}} \{\log|\mathbf{A}| + \operatorname{trace}(\mathbf{A}^{-1}\mathbf{M})\} = \mathbf{M}, \tag{E.4}$$

where the minimization is over all positive definite symmetric matrices. Therefore we have $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}$, $\widehat{\boldsymbol{\Omega}}_{0,\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}_0$ and

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\widehat{\boldsymbol{\Omega}}_{0,\boldsymbol{\Gamma}}\boldsymbol{\Gamma}_0^T = \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{S}_{\mathbf{X}}\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{S}_{\mathbf{X}}\mathbf{Q}_{\boldsymbol{\Gamma}}. \tag{E.5}$$

The remaining partially maximized form is $M_n(\boldsymbol{\Gamma}) = -n/2\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}_0|\right\}$ where we have omitted a constant p from $\operatorname{trace}\{\mathbf{I}_u\} + \operatorname{trace}\{\mathbf{I}_{p-u}\}$. The rest of the proof is a direct consequence of the following implication from Cook et al. (2013): Suppose $\mathbf{A} \in \mathbb{S}^{t\times t}$ is non-singular and that the column partition $(\mathbf{B}, \mathbf{B}_0) \in \mathbb{R}^{t\times t}$ is orthogonal. Then (1) $|\mathbf{B}^T\mathbf{A}\mathbf{B}| = |\mathbf{A}| \cdot |\mathbf{B}_0^T\mathbf{A}^{-1}\mathbf{B}_0|$ and (2) $|\mathbf{A}| \leq |\mathbf{B}^T\mathbf{A}\mathbf{B}| \cdot |\mathbf{B}_0^T\mathbf{A}\mathbf{B}_0|$ if and only if $\operatorname{span}(\mathbf{B})$ reduces $\mathbf{A}$. □

E.5 Proposition 4

Proof. Let $\mathbf{E}_p \in \mathbb{R}^{p^2\times p(p+1)/2}$ denote the expansion operator and let $\mathbf{C}_p \in \mathbb{R}^{p(p+1)/2\times p^2}$ denote the contraction operator such that they connect vec and vech operators as $\operatorname{vec}(\mathbf{A}) = \mathbf{E}_p\operatorname{vech}(\mathbf{A})$ and $\operatorname{vech}(\mathbf{A}) = \mathbf{C}_p\operatorname{vec}(\mathbf{A})$ for any $\mathbf{A} \in \mathbb{S}^{p\times p}$.

Because the standard estimator $(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} = \mathbf{S}_{\mathbf{X}}$ were obtained separately from $C_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and the marginal multivariate normal likelihood of $\mathbf{X}$, and because that the parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are orthogonal, the asymptotic variance of $\widehat{\boldsymbol{h}}_n$ is block diagonal as $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\alpha}}_n) = \mathbf{V}(\boldsymbol{\alpha}_t) = \mathbf{F}^{-1}(\boldsymbol{\alpha}_t)$, $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_n) = \mathbf{V}(\boldsymbol{\beta}_t) = \mathbf{F}^{-1}(\boldsymbol{\beta}_t)$ and $\operatorname{avar}[\sqrt{n}\operatorname{vech}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}})] = \mathbf{F}_{\boldsymbol{\Sigma}_{\mathbf{X}}}^{-1}$, where the Fisher information matrix $\mathbf{F}_{\boldsymbol{\Sigma}_{\mathbf{X}}} = \mathbf{E}_p^T(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{E}_p/2$. (Cook et al. 2013). Hence,

$$\begin{aligned}\mathbf{F}(\boldsymbol{h}_t) &= \mathrm{E}\left\{\left(\frac{\partial L_n(\boldsymbol{h})}{\partial \boldsymbol{h}}\right)_{\boldsymbol{h}=\boldsymbol{h}_t}\left(\frac{\partial L_n(\boldsymbol{h})}{\partial \boldsymbol{h}^T}\right)_{\boldsymbol{h}=\boldsymbol{h}_t}\right\} = -\mathrm{E}\left\{\left(\frac{\partial^2 L_n(\boldsymbol{h})}{\partial \boldsymbol{h}\partial \boldsymbol{h}^T}\right)_{\boldsymbol{h}=\boldsymbol{h}_t}\right\} \\ &= \operatorname{diag}\left(\mathbf{F}(\boldsymbol{\alpha}_t), \mathbf{F}(\boldsymbol{\beta}_t), \mathbf{F}_{\boldsymbol{\Sigma}_{\mathbf{X}}}\right).\end{aligned}$$

Following Cook et al. (2013), the asymptotic variance for $\widehat{\boldsymbol{h}}_{\mathrm{env}} = \boldsymbol{h}(\widehat{\boldsymbol{\phi}}_{\mathrm{env}})$ can be expressed as $\operatorname{avar}(\widehat{\boldsymbol{h}}_{\mathrm{env}}) = \mathbf{H}(\mathbf{H}^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H})^{\dagger}\mathbf{H}^T$, where $\mathbf{H} = \partial\boldsymbol{h}/\partial\boldsymbol{\xi} = \left(\partial\boldsymbol{h}_i/\partial\boldsymbol{\xi}_j\right)_{i=1,\ldots,3,j=1,\ldots,5}$ is the gradient matrix and † denotes the Moore-Penrose generalized inverse. By direct computation,

we found

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_{m-p} & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \mathbf{I}_p & 0 & 0 \\ 0 & 0 & 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & \mathbf{C}_p(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix}. \tag{E.6}$$

Since $\mathrm{avar}(\widehat{\boldsymbol{h}}_{\mathrm{env}}) = \mathbf{H}(\mathbf{H}^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H})^\dagger\mathbf{H}^T$ depends only on $\mathrm{span}(\mathbf{H})$, we transform $\mathbf{H} \to \mathbf{H}_1$ similar to Cook et al. (2013) so that $\mathrm{span}(\mathbf{H}) = \mathrm{span}(\mathbf{H}_1)$ and

$$\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\mathrm{env}}) = \mathbf{H}_1(\mathbf{H}_1^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H}_1)^\dagger\mathbf{H}_1^T = \sum_{j=1}^{5} \mathbf{H}_{1j}(\mathbf{H}_{1j}^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H}_{1j})^\dagger\mathbf{H}_{1j}^T, \tag{E.7}$$

where $\mathbf{H}_1 = (\mathbf{H}_{11}, \ldots, \mathbf{H}_{15})$ has the blockwise form of

$$\begin{aligned} \mathbf{H}_1 &= \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \mathbf{H}_{13} & \mathbf{H}_{14} & \mathbf{H}_{15} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I}_{m-p} & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0 & 0 & 0 \\ 0 & 0 & 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) & \mathbf{C}_p(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix}. \end{aligned}$$

Because of the zeros blocks in $\mathbf{H}_{11}$, $\mathbf{H}_{14}$ and $\mathbf{H}_{15}$, they have no contribution to the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$, which is the middle $p \times p$ block of $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\mathrm{env}})$:

$$\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{env}}) = \boldsymbol{\Gamma}(\mathbf{H}_{12}^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H}_{12})^\dagger\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)(\mathbf{H}_{13}^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H}_{13})^\dagger(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T). \tag{E.8}$$

Then by straightforward calculations similar to Cook et al. (2010), we have

$$\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{env}}) = \boldsymbol{\Gamma}\left\{\boldsymbol{\Gamma}^T\mathbf{F}(\boldsymbol{\beta}_t)\boldsymbol{\Gamma}\right\}^{-1}\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T),$$

where $\mathbf{F}(\boldsymbol{\beta}_t) = \left\{\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_n)\right\}^{-1}$ and $\mathbf{M} = (\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T)\mathbf{F}(\boldsymbol{\beta}_t)(\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0) + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_{u(p-u)}$.

From Proposition 5, we recognize the first term in $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{env}})$ is $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$ which equals to $\mathrm{avar}(\sqrt{n}\boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})$. To further interpret the second term, we follow the same calculation as in Cook et al. (2010) by making the following substitutes: $\boldsymbol{\Sigma} \to \boldsymbol{\Sigma}_{\mathbf{X}}$, $\boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1} \to \mathbf{F}(\boldsymbol{\beta}_t)$ and the dimension $r \to p$. Then the conclusion follows.

□

E.6 Corollary 1

Proof. From Proposition 2, we can see $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\sigma^2\mathbf{I}_p}(\boldsymbol{\beta}_t) = \mathrm{span}(\boldsymbol{\beta}_t)$. Therefore, the envelope is one-dimensional and $\boldsymbol{\beta}_t = \boldsymbol{\Gamma}\eta$ with $\boldsymbol{\Gamma} = \boldsymbol{\beta}_t/||\boldsymbol{\beta}_t|| \in \mathbb{R}^{p\times 1}$ and

$\eta = ||\boldsymbol{\beta}_t|| \in \mathbb{R}^1$. The covariances $\Omega = \sigma^2$ and $\boldsymbol{\Omega}_0 = \sigma^2 \mathbf{I}_{p-1}$. The matrix $\mathbf{M}$ in Proposition 4 becomes $\mathbf{M} = (\eta\boldsymbol{\Gamma}_0^T)\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\eta\boldsymbol{\Gamma}_0)$ because $\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 = \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} = \mathbf{I}_{p-1}$ cancel with the last term. Hence,

$$\begin{aligned}
\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}}) &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{P}_{\boldsymbol{\Gamma}} + (\eta\boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\eta\boldsymbol{\Gamma}_0^T) \\
&= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{P}_{\boldsymbol{\Gamma}} + \eta^2\boldsymbol{\Gamma}_0(\eta^2\boldsymbol{\Gamma}_0^T\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T \\
&= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{Q}_{\boldsymbol{\Gamma}} \\
&= \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}),
\end{aligned}$$

where the last two equalities have used the fact that $\boldsymbol{\Gamma}$ reduces $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}$. □

E.7 Proposition 7

Proof. Let $(\mathbf{R} \otimes \mathbf{L})_0 := (\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L})$ is a $pr \times (pr - d_X d_Y)$ semi-orthogonal matrix so that $\mathbf{O} := (\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L}, \mathbf{R} \otimes \mathbf{L})$ is an orthogonal basis for $\mathbb{R}^{pr}$, where $(\mathbf{R}, \mathbf{R}_0)$ and $(\mathbf{L}, \mathbf{L}_0)$ are orthogonal bases for $\mathbb{R}^p$ and $\mathbb{R}^r$.

We first prove that the following statements are equivalent.

1. $\operatorname{span}(\mathbf{R} \otimes \mathbf{L})$ reduce $\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L$.
2. $\operatorname{span}(\mathbf{R})$ reduces $\boldsymbol{\Delta}_R$ and $\operatorname{span}(\mathbf{L})$ reduces $\boldsymbol{\Delta}_L$.
3. $\mathbf{O}(\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L)\mathbf{O}^T = \operatorname{diag}\{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \mathbf{M}_4\}$, where $\mathbf{M}_1 = \mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R} \otimes \mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L}$, $\mathbf{M}_2 = \mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R}_0^T \otimes \mathbf{L}_0\boldsymbol{\Delta}_L\mathbf{L}_0^T$, $\mathbf{M}_3 = \mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R}^T \otimes \mathbf{L}_0\boldsymbol{\Delta}_L\mathbf{L}_0^T$ and $\mathbf{M}_4 = \mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R}_0^T \otimes \mathbf{L}\boldsymbol{\Delta}_L\mathbf{L}^T$.
4. $\mathbf{O}(\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L)\mathbf{O}^T = \operatorname{diag}\{\mathbf{M}_1, \mathbf{M}_0\}$, where $\mathbf{M}_0 = (\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L})(\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L)(\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L})^T$.

By definition, we know that Statement 1 is equivalent to 4 and that 2 is equivalent to 3. Also, 3 implies 4. We then only need to show that 4 implies 3. From 4, we know that the off-diagonal blocks of $\mathbf{O}(\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L)\mathbf{O}^T$ are all zero matrices. More explicitly,

$$\begin{aligned}
0 &= \mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R} \otimes \mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L}_0 = \mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R} \otimes \mathbf{L}_0^T\boldsymbol{\Delta}_L\mathbf{L} \\
&= \mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L} = \mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R} \otimes \mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L} \\
&= \mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L}_0 = \mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R} \otimes \mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L}_0. \qquad \text{(E.9)}
\end{aligned}$$

Without loss of generality, we assume that $\mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R}$, $\mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L}$, $\mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R}_0$ and $\mathbf{L}_0^T\boldsymbol{\Delta}_L\mathbf{L}_0$ are all nonzero. Otherwise, the problem degenerates to the case of reducing either $\boldsymbol{\Delta}_R$ or $\boldsymbol{\Delta}_L$ and the proof would be trivial. Because $\mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R}$, $\mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L}$, $\mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R}_0$ and $\mathbf{L}_0^T\boldsymbol{\Delta}_L\mathbf{L}_0$ are all nonzero, we must have from (E.9) that

$$0 \;=\; \mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L}_0 = \mathbf{L}_0^T\boldsymbol{\Delta}_L\mathbf{L} = \mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R} = \mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R}_0,$$

which implies the following blocks within $\mathbf{M}_0$ are all zero matrices:

$$\begin{aligned} 0 \;&=\; \mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L}_0 = \mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}_0^T\boldsymbol{\Delta}_L\mathbf{L} \\ &=\; \mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}_0^T\boldsymbol{\Delta}_L\mathbf{L}_0 = \mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R} \otimes \mathbf{L}_0^T\boldsymbol{\Delta}_L\mathbf{L}_0 \\ &=\; \mathbf{R}_0^T\boldsymbol{\Delta}_R\mathbf{R} \otimes \mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L}_0 = \mathbf{R}^T\boldsymbol{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}^T\boldsymbol{\Delta}_L\mathbf{L}_0. \end{aligned}$$

Hence, $\mathbf{M}_0 = \operatorname{diag}\{\mathbf{M}_2, \mathbf{M}_3, \mathbf{M}_4\}$, which leads to the equivalence of the four statements. From Definition 3, Definition 4 and the equivalence between statement 1 and 2, we reach the conclusion of Proposition 7.

□

References

[1] COOK, R. D. AND FORZANI, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104**, 197–208.

[2] COX, D.R. (1972), Regression models and life-tables (with Discussion), *J. R. Statist. Soc. B*, **34**, 187–220.

[3] COX, D.R. (1975), Partial likelihood, *Biometrika*, **62**, 269–276.

[4] DAWID, A.P. (1981), Some matrix-variate distribution theory: Notational considerations and a Bayesian application, *Biometrika*, **68**, 265–274.

[5] EATON, M.L. (1986), A characterization of spherical distributions, *Journal of Multivariate Analysis*, **20**, 272–276.

[6] HUNG, H., WU, P. S., TU, I. P. AND HUNG, S. Y. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika*, **99**, 569–583.

[7] LI, B., KIM, M.K. AND ALTMAN, M. (2010). On dimension folding of matrix or array valued statistical objects. *Annals of Statistics*, **38**, 1097–1121.

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