Foundations for Envelope Models and Methods

R. Dennis Cook and Xin Zhang

Envelopes were recently proposed by Cook, Li and Chiaromonte 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. Supplementary materials for this article are available online.

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,

\[ Y_i = \alpha + \beta X_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.1} \]

where \( \epsilon_i \in \mathbb{R}^r \) is a normal error vector that has mean 0, constant covariance \( \Sigma_{XY} > 0 \) and is independent of \( X, \alpha \in \mathbb{R}^r \) and \( \beta \in \mathbb{R}^{r \times p} \) is the regression coefficient matrix in which we are primarily interested. Efficiency gains in the estimation of \( \beta \) are achieved by reparameterizing this model in terms of special projections of the response vector \( Y \in \mathbb{R}^r \) or the predictor vector \( X \in \mathbb{R}^p \). In this article we propose extensions of envelope methodology beyond the linear 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 \( \beta \) by incorporating the projection \( P_E Y \) onto the smallest subspace \( E \subseteq \mathbb{R}^r \) with the properties (1) the distribution of \( Q_E Y \mid X \) does not depend on the value of the non stochastic predictor \( X \), where \( Q_E = I - P_E \), and (2) \( P_E Y \) is independent of \( Q_E Y \) given \( X \). These conditions imply that the distribution of \( Q_E Y \) is not affected by \( X \) marginally or through an association with \( P_E Y \). Consequently, changes in the predictor affect the distribution of \( Y \) only via \( P_E Y \). We refer to \( P_E Y \) informally as the material part of \( Y \) and to \( Q_E Y \) as the immaterial part of \( Y \).

The notion of an envelope arises in the formal construction of the response projection \( P_E \) 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 matrices and let \( \mathbb{S}^{k \times k} \) be the set of all real and symmetric matrices. If \( A \in \mathbb{R}^{m \times n} \), then span\( (A) \subseteq \mathbb{R}^m \) is the subspace spanned by columns of \( A \).

Definition 1. A subspace \( R \subseteq \mathbb{R}^p \) is said to be a reducing subspace of \( M \in \mathbb{S}^{p \times p} \) if \( R \) decomposes \( M \) as \( M = P_R M P_R + Q_R M Q_R \). If \( R \) is a reducing subspace of \( M \), we say that \( R \) reduces \( M \).

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 \( M \in \mathbb{S}^{p \times p} \) and let \( B \subseteq \text{span}(M) \). Then the \( M \)-envelope of \( B \), denoted by \( E_M(B) \), is the intersection of all reducing subspaces of \( M \) that contain \( B \).

Definition 2 is the formal definition of an envelope, which is central to our developments. We will often identify the subspace \( B \) as the span of a specified matrix \( U \); \( B = \text{span}(U) \). To avoid proliferation of notation in such cases, we will also use the matrix as the argument to \( E_M \); \( E_M(U) := E_M(\text{span}(U)) \).

The response projection \( P_E \) is then defined formally as the projection onto \( E_{YX} (\beta) \), which by construction is the smallest reducing subspace of \( \Sigma_{YX} \) that contains \( \beta \) (or “envelopes” span\( (\beta) \) and hence the name “envelope”). Model (1.1) can be parameterized in terms of \( P_E \) by using a basis. Let \( u = \dim (E_{YX} (\beta)) \) and let \( (\Gamma, \Gamma_0) \in \mathbb{R}^{r \times u} \) be an orthogonal matrix with \( \Gamma \in \mathbb{R}^{r \times r} \) and \( \Gamma_0 = E_{YX} (\beta) \). This leads directly to the envelope version of model (1.1),

\[ Y_i = \alpha + \Gamma \eta X_i + \epsilon_i, \quad \Sigma_{YX} = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T, \quad i = 1, \ldots, n, \tag{1.2} \]

where \( \beta = \Gamma \eta, \eta \in \mathbb{R}^{u \times p} \) gives the coordinates of \( \beta \) relative to basis \( \Gamma \), and \( \Omega \) and \( \Omega_0 \) are positive definite matrices. While \( \eta \), \( \Omega \) and \( \Omega_0 \) depend on the basis \( \Gamma \) selected to represent \( E_{YX} (\beta) \), the parameters of interest \( \beta \) and \( \Sigma_{YX} \) depend only on \( E_{YX} (\beta) \) 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 \( \hat{\beta}_{\text{env}} \) of \( \beta \) is just the projection of the maximum likelihood estimator \( \hat{\beta} \) onto the estimated envelope, \( \hat{\beta}_{\text{env}} = P_T \hat{\beta} \), and \( \sqrt{n} (\text{vec}(\hat{\beta}_{\text{env}}) - \text{vec}(\beta)) \) 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: 30 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 \( Y_i \in \mathbb{R}^{10} \) is the vector of cattle weights from week 2 to week 19, and the binary predictor \( X_i \) is either 0 or 1 indicating the two treatments. Then \( \alpha = E(Y|X = 0) \) is the mean profile for one treatment and \( \beta = E(Y|X = 1) - E(Y|X = 0) \) 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. The maximum absolute ratio of an element in \( \beta \) to its bootstrap standard error is only about 1.3, suggesting that there were no inter-treatment differences at any time. However, the likelihood ratio test statistics for the hypothesis \( \beta = 0 \) 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 \( \Sigma_{Y|X} \) 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 \( Y \) 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. The two profile plots in Figure 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 \( \hat{\beta}_{\text{env}} \) to their bootstrap standard errors were all smaller than 2.5 before week 10 and all larger than 3.4 from week 10 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 \( \hat{\beta}_{\text{env}} = P_T \hat{\beta} \) graphically, as shown in Figure 2, we use only the weights at week 12 and 14 as the bivariate response \( Y = (Y_1, Y_2)^T \). The maximum likelihood estimator of the first element \( \beta_1 \) of \( \beta \) under model (1.1) is just the difference between the treatment means at week 12. In terms of Figure 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 tight densities for \( Y_1|X = 0 \) and \( Y_1|X = 1 \) along the horizontal axis of Figure 2. These densities are close, indicating that it may take a very large sample size to detect the difference between the marginal means that is evident in the figure. In contrast, the envelope estimator for \( \beta_1 \) first projects the data onto the estimated envelope span(\( \hat{T} \)) and then onto the horizontal axis, as illustrated in Figure 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 \( \hat{\beta} = (5.5, -4.8)^T \) with bootstrap standard errors \( (4.2, 4.4)^T \), while the envelope estimator \( \hat{\beta}_{\text{env}} = (5.4, -5.1)^T \) 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 $\hat{\beta}_{env}$ of $\beta$ from model (1.2) is $\hat{\beta}_{env} = P Y^T \hat{\beta}$ and that the asymptotic variance of $\text{vec}(\hat{\beta}_{env})$ is never larger than that for $\text{vec}(\hat{\beta})$. The reduction in variation achieved by the envelope estimator can be substantial when the immaterial variation $\Omega_0 = \text{var}(\Gamma^T Y)$ is large relative to the material variation $\Omega = \text{var}(\Gamma^T Y)$. For instance, using the Frobenius norm, we typically see substantial gains when $\|\Omega_0\|_F \gg \|\Omega\|_F$, as happens in Figure 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 $\Sigma_{XYX}$-envelope of $\text{span}(\beta_1)$, $\mathcal{E}_{\Sigma_{XYX}}(\beta_1)$, to develop a partial envelope estimator of $\beta_1$ in the partitioned multivariate linear regression

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\beta_1 \in \mathbb{R}^r$ is the parameter vector of interest, $X = (X_1, X_2)^T$ and the remaining terms are as defined for model (1.1). The partial envelope estimator $\hat{\beta}_{env} = P Y^T \hat{\beta}_1$ has the potential to yield efficiency gains beyond those for the full envelope, particularly when $\mathcal{E}_{\Sigma_{XYX}}(\beta_1) = \mathbb{R}^r$, 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 $\text{var}(X) = \Sigma_X$. Their reasoning led them to parameterize the linear model in terms of $\mathcal{E}_{\Sigma_{XYX}}(\beta^T)$, which again achieved substantial efficiency gains in the estimation of $\beta$ and in prediction. They also showed that the SIMPLS algorithm (de Jong 1993; see also ter Braak and de Jong 1999) for partial least squares provides a $\sqrt{n}$-consistent estimator of $\mathcal{E}_{\Sigma_{XYX}}(\beta^T)$ and demonstrated that the envelope estimator $\hat{\beta}_{env} = \hat{\beta}_1 Y^T F \Sigma_X$ typically outperforms the SIMPLS estimator in practice, where we use $P_{\mathcal{A}(V)} := P_{\mathcal{A}(V)} = AA^T V A^T$ to denote the projection onto $\mathcal{A} = \text{span}(A)$ in the $V$ inner product, and $Q_{\mathcal{A}(V)} = I - P_{\mathcal{A}(V)}$.

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 $L_n(\Gamma) = -(n/2 J(\Gamma))$ for estimation of a basis $\Gamma$ of $\mathcal{E}_{\Sigma_{XYX}}(\beta)$, $\mathcal{E}_{\Sigma_{XYX}}(\beta_1)$ or $\mathcal{E}_{\Sigma_{XYX}}(\beta^T)$ all have the same form with

$$J(\Gamma) = \log |\Gamma^T \hat{\Sigma} \Gamma| + \log |\Gamma^T (\hat{\Sigma} + \hat{\Upsilon})^{-1} \Gamma|,$$

where the positive definite matrix $\hat{\Sigma}$ and the positive semidefinite matrix $\hat{\Upsilon}$ depend on context. The estimated basis is $\hat{\Gamma} = \arg\min J(\Gamma)$, where the minimization is carried out over a set of semiorthogonal matrices whose dimensions depend on the envelope being estimated. Estimates of the remaining parameters are then simple functions of $\hat{\Gamma}$. We represent sample covariance matrices as $\Sigma_{11}$ and defined with the divisor $n$: $S_X = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T / n$, $S_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})^T / n$ and $S_{YX}$ denotes the covariance matrix of the residuals from the linear fit of $Y$ on $X$. To determine the estimators of the response envelope $\mathcal{E}_{\Sigma_{YX}}(\beta)$, the predictor envelope $\mathcal{E}_{\Sigma_{XY}}(\beta^T)$ and the partial envelope $\mathcal{E}_{\Sigma_{XYX}}(\beta_1)$, we have $[\hat{\Sigma}, \hat{\Sigma} + \hat{\Upsilon}] = [S_{YX}, S_{Y1}, S_{X1}, S_{YX}, S_{XY}, S_{YX}]$. For instance, a basis of the response envelope for the cattle data was estimated by maximizing $L_n(\Gamma)$ with $[\hat{\Sigma}, \hat{\Sigma} + \hat{\Upsilon}] = [S_{YX}, S_{Y1}, S_{X1}, S_{YX}, S_{XY}, S_{YX}]$. 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 (2015) introduced envelopes for simultaneously reducing the predictors and responses and showed synergistic effects of simultaneous envelopes in improving both estimation efficiency and prediction. Cook and Zhang (2014) 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 $\hat{\phi}$ of a parameter vector $\phi$ may be improved by enveloping $\text{span}(\phi)$ with respect to the asymptotic covariance of $\hat{\phi}$. 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 $R^r$, $u \leq r$, is denoted as $G_u$. If $\sqrt{n}(\hat{\theta} - \theta)$ converges to a normal random vector with mean 0 and covariance matrix $V$ we write its asymptotic covariance matrix as $\text{var}(\sqrt{n}\hat{\theta}) = V$. We will use operators $v : R^{d \times b} \rightarrow R^{a \times b}$, which vectorizes an arbitrary matrix by stacking its columns, and $v^T : R^{a \times b} \rightarrow R^{b \times (a+1)/2}$, which vectorizes a symmetric matrix by stacking the unique elements of its columns. Let $A \otimes B$ denote the Kronecker product of two matrices $A$ and $B$. 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 $\hat{\theta}_x$ to denote an estimator of a parameter $\theta$ given another parameter $x$. For instance, $\Sigma_{X,\Gamma}$ is an estimator of $\Sigma_X$ given $\Gamma$.

2. A CONSTRUCTIVE DEFINITION OF ENVELOPES

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

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

1. If $M \in S^{p \times p}$ has $q \leq p$ eigenvalues, then the $M$-envelope of $B \subseteq \text{span}(M)$ can be constructed as $E_M(B) = \sum_{i=1}^q P_i B P_i$.
2. With $f$ and $f^*$ as previously defined, $E_M(B) = E_M(f^*(M)B)$.
3. If $f$ is strictly monotonic then $E_M(B) = E_{f^*(M)}(B) = E_M(f^*(M)B)$.

From the first part of this proposition, we see that the $M$-envelope of $B$ is the sum of the projections of $B$ onto the eigenspaces of $M$; the second part of this proposition gives a variety of substitutes for $B$ without changing the envelope; and part 3 implies that we may consider replacing $M$ by $f^*(M)$ 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 $Y$ or $X$. 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 $M$ and the subspace $B$ 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 $\hat{\theta}$ denote an estimator of a parameter vector $\theta \in \Theta \subseteq R^m$ based on a sample of size $n$. Let $\theta_0$ denote the true value of $\theta$ and assume, as is often the case, that $(\sqrt{n}(\hat{\theta} - \theta_0))$ converges in distribution to a normal random vector with mean 0 and covariance matrix $V(\theta_0) > 0$ as $n \rightarrow \infty$. To accommodate the presence of nuisance parameters we decompose $\theta = (\phi^T, \phi^T)^T$, where $\phi \in R^p$, $p \leq m$, is the parameter vector of interest and $\psi \in R^{m-p}$ is the nuisance parameter vector. The asymptotic covariance matrix of $\hat{\theta}$ is represented as $V_{\phi}(\theta_0)$, which is the $p \times p$ lower right block of $V(\theta_0)$. Then we construct an envelope for improving $\hat{\phi}$ as follows.

**Definition 3.** The envelope for the parameter $\phi \in R^p$ is defined as $E_{\phi}(\theta_0, \phi_0) \subseteq R^p$.

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 prespecified method of estimation through the covariance matrix $V_{\phi}(\theta_0)$, 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 prespecified estimator, perhaps a maximum likelihood, least squares or robust estimator. Second, the matrix to be reduced – here $V_{\phi}(\theta_0)$ – 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 $\beta_1$ reviewed in Section 1.3 and the envelopes for $\beta$ 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 $M = V_{\phi}(\theta_0)$ and $U = \phi, \phi^T$. To apply Definition 3 for the partial envelope of $\beta_1$ based on model (1.3), the asymptotic covariance matrix of the maximum likelihood estimator of $\beta_1$ is $V_{\beta_1, \beta_1}(\theta_0) = \Sigma_{X,\Gamma}^{-1}(\Sigma_{X,\Gamma}^{-1} + \hat{\Sigma}_{X,\Gamma})^{-1}$. Consequently, by Proposition 1, $E_{\beta_1, \beta_1}(\hat{\theta}_1) = E_{\beta_1, \beta_1}(\beta_1)$, and thus Definition 3 recovers the partial envelopes of Su and Cook (2011). To construct the partially maximized log-likelihood (1.4) we set $M = V_{\beta_1, \beta_1}$, and $U = \beta_1, \beta_1^T$. Then using sample versions gives $J(\Gamma) = \log |(S_{X,\Gamma}^{-1})_{11} (\Gamma^T S_{X,\Gamma} \Gamma) + \log \Gamma^T \left(\left(S_{X,\Gamma}^{-1}\right)_{11} S_{X,\Gamma} + \hat{\Sigma}_{X,\Gamma} + \hat{\beta}_1{\hat{\beta}_1}^T\right)^{-1} \Gamma |$ which is the partially maximized log-likelihood of Su and Cook (2011). It is important to note that, although $E_{\beta_1, \beta_1}(\hat{\theta}_1) = E_{\beta_1, \beta_1}(\beta_1)$, Definition 3 requires that we use $V_{\beta_1, \beta_1} = (\Sigma_{X,\Gamma}^{-1})_{11} S_{X,\Gamma}$ and not $S_{X,\Gamma}$ alone. The response envelope by Cook et al. (2010) can be seen as a special case of the partial envelope model with absence of $X_2$. This implies that Definition 3 can also reproduce the response envelope objective function, which is obtained by replacing $S_{Y,X}$ with $S_Y$ in the $J(\Gamma)$ function (1.4).

As another illustration, consider $X$ reduction in model (1.1) with $r = 1$ and $p > 1$. To emphasize the scalar response, let
\( \sigma^2_{\varepsilon X} = \text{var}(\varepsilon) \) with sample residual variance \( s^2_{\varepsilon X} \). The ordinary least squares estimator of \( \beta \) has asymptotic variance \( V_{\beta \beta} = \sigma^2_{\varepsilon X} \Sigma^{-1} \). Direct application of Definition 3 then leads to the \( \sigma^2_{\varepsilon X} \Sigma^{-1} \) -envelope of \( \text{span}(\beta) \), \( \mathcal{E}_{\sigma^2_{\varepsilon X} \Sigma^{-1}}(\beta) \). However, it follows from Proposition 1 that this envelope is equal to \( \mathcal{E}_{\varepsilon X}(\beta) \), 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

\[
\mathbf{M} = V_{\beta \beta} \quad \text{and} \quad U = \beta \beta^T.
\]

Then substituting sample quantities

\[
J(\Gamma) = \log |s^2_{\varepsilon X} \Gamma^T \Sigma^{-1} \Gamma| + \log |\Gamma^T \left[ s^2_{\varepsilon X} \Sigma^{-1} + \hat{\beta} \hat{\beta}^T \right]^{-1} \Gamma|
\]

which is the partially maximized log-likelihood of Cook et al. (2013). By applying Definition 3 in combination with \( J(\Gamma) \) and \( \mathcal{E}_{\varepsilon X}(\beta) \), we restrict \( \theta = \psi = \beta \) to a product space. Let

\[
\mathbf{M} = V_{\beta \beta} \quad \text{and} \quad U = \beta \beta^T,
\]

so that the marginal log-likelihood for \( Y \in \mathbb{R} \) can be written as \( \log |\mathbf{M}^T \mathbf{X} \mathbf{X}^T + \mathbf{U} | \) and the observed data are \( \{Y_i, X_i\}, i = 1, \ldots, n \). We take \( \hat{\beta} \) to be the parameter vector of interest and, prior to the introduction of envelopes, we restrict \( \alpha, \beta \) and \( \Psi \) to a product space. Let

\[
L_n(\theta) = \sum_{i=1}^n \log g(Y_i|\theta) \text{ be the full log-likelihood, let the log-likelihood conditional on } X \text{ be represented by } C_n(\alpha, \beta) = \sum_{i=1}^n \log g(Y_i|\alpha, \beta^T X_i) \text{ and let } M_n(\psi) = \sum_{i=1}^n \log h(X_i|\psi) \text{ be the marginal log-likelihood for } \psi. \]

Then we can decompose

\[
L_n(\theta) = C_n(\alpha, \beta) + M_n(\psi).
\]

Since our primary interest lies in \( \beta \) and \( X \) is ancillary, estimators are typically obtained as

\[
(\hat{\alpha}, \hat{\beta}) = \arg \max_{\alpha, \beta} C_n(\alpha, \beta).
\]

Our goal here is to improve the prespecified estimator \( \hat{\beta} \) by introducing the envelope \( \mathcal{E}_{\psi}(\beta) \), where

\[
\hat{\beta} = \arg \max_{\beta} C_n(\alpha, \beta),
\]

\[
\text{avar}(\sqrt{n} \hat{\beta}),\text{ according to Definition 3.}
\]

Let \( (\Gamma, \mathbf{\Gamma}_0) \in \mathbb{R}^{p \times p} \) denote an orthogonal matrix where \( \Gamma \in \mathbb{R}^{p \times p} \) is a basis for \( \mathcal{E}_{\psi}(\beta) \). Since \( \beta \in \mathcal{E}_{\psi}(\beta) \), we can write \( \beta_1 = \Gamma \eta \) for some \( \eta \in \mathbb{R}^p \). Because \( V_{\beta \beta}(\theta) \) typically depends on the distribution of \( X \) and \( \mathcal{E}_{\psi}(\beta) \) reduces \( V_{\beta \beta}(\theta) \), the marginal \( M_n \) will depend on \( \Gamma \). Then the log-likelihood becomes

\[
L_n(\alpha, \eta, \psi_1, \Gamma) = C_n(\alpha, \eta, \Gamma) + M_n(\psi_1, \Gamma).
\]

The relationship between \( \mathbf{X} \) and \( \hat{\mathcal{E}}_{\psi}(\beta) \) that is embodied in \( M_n(\psi_1, \Gamma) \) could be complicated, depending on the distribution of \( X \). However, as described in the following proposition, it simplifies considerably when \( \mathbb{E}(X|\Gamma^T \mathbf{X}) \) is a linear function of \( \Gamma^T \mathbf{X} \). This is well-known as the linearity condition in the sufficient dimension reduction literature where \( \Gamma \) 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 \( X \) follows an elliptically contoured distribution, the linearity condition will be guaranteed for any \( \Gamma \) (Eaton 1986).

Proposition 2. Assume that \( \mathbb{E}(X|\Gamma^T \mathbf{X}) \) is a linear function of \( \Gamma^T \mathbf{X} \). Then \( \hat{\mathcal{E}}_{\psi}(\beta) = \mathcal{E}_{\varepsilon X}(\beta) \).

An implication of Proposition 2 is that, for some positive definite matrices \( \Omega \in \mathbb{R}^{p \times p} \) and \( \Sigma \in \mathbb{R}^{p \times p} \), \( \mathcal{E}_{\varepsilon X} = \Gamma^T \Omega \mathcal{E}_{\varepsilon X} \Gamma \) and thus \( M_n \) must depend on \( \Gamma \) through the marginal covariance \( \Sigma \). Consequently, we can write

\[
M_n(\Sigma, \psi_2) = M_n(\Gamma, \Omega, \Omega_0, \psi_2),
\]

which \( \psi_2 \) represents any remaining parameters in the marginal function. If \( X \) is normal with mean \( \mu_X \) and variance \( \Sigma_X \), then \( \psi_2 = \mu_X \) and

\[
M_n(\Gamma, \Omega, \Omega_0, \mu_X) = M_n(\Gamma, \Omega, \Omega_0, \mu_X)
\]

is the marginal normal log-likelihood. In this case, it is possible to maximize \( M_n \) over all its parameters except \( \Gamma \), as stated in the following proposition.

Proposition 3. Assume that \( X \in \mathbb{R}^p \) is multivariate normal \( N(\mu_X, \Sigma_X) \) and that \( \Gamma \in \mathbb{R}^{p \times u} \) is a semiothogonal basis matrix for \( \mathcal{E}_{\varepsilon X}(\beta) \). Then

\[
\Gamma X = \mathbf{X}, \quad \Sigma_X^{-1} = \mathbf{P}_T \Sigma_X \mathbf{P}_T + \mathbf{Q}_T \Sigma_X \mathbf{Q}_T
\]

and

\[
M_n(\Gamma) := \max_{\Omega_0, \mu_X} M_n(\Gamma, \Omega, \Omega_0, \mu_X)
\]

\[
= -n/2 \{ \log |\Gamma^T \Sigma_X \Gamma| + \log |\Gamma^T \Sigma_X \Omega_0| \}
\]

\[
= -n/2 \{ \log |\Gamma^T \Sigma_X \Gamma| + \log |\Gamma^T \Sigma_X^{-1} \Gamma| + \log |\Sigma_X| \}
\]

(3.2)

(3.3)

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

This proposition indicates that if \( X \) is marginally normal, the envelope estimators are

\[
(\hat{\alpha}_{\text{env}}, \hat{\Gamma}) = \arg \max_{\alpha, \Gamma} C_n(\alpha, \eta, \Gamma) + M_n(\Gamma) \]

(3.4)

In particular, the envelope estimator of \( \beta \) is \( \hat{\beta}_{\text{env}} = \mathbf{\Gamma} \eta \) and, from Proposition 3, \( \hat{\Sigma}_{\text{env}} = P_T \Sigma_X P_T + Q_T \Sigma_X Q_T \). It follows also from Proposition 3 that one role of \( M_n \) is to pull span(\( \mathbf{\Gamma} \)) toward the subspace of \( \Sigma_X \), 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 \( f(Y_i | \theta_i, \varphi) = \exp\{[Y_i \theta_i - b(\theta_i)]/\alpha(\varphi) + c(Y, \varphi)\}, \) \( i = 1, \ldots, n, \) where \( \theta \) is the natural parameter and \( \varphi > 0 \) is the dispersion parameter. We consider the canonical link function \( \vartheta(\alpha, \beta) = \alpha + \beta^T X, \) which is a monotonic differentiable function of \( E(Y|X, \theta, \varphi). \) We also restrict discussion to one-parameter families so that the dispersion parameter \( \varphi \) is not needed.

The conditional log-likelihood takes the form of

\[
\log f(y | \theta) = y \theta - b(\theta) + c(y) := C(\theta),
\]

where \( \theta = \alpha + \beta^T X \) is the canonical parameter. Then the Fisher information matrix for \((\alpha, \beta)\) evaluated at the true parameters is

\[
F(\alpha, \beta) = E(\mathbb{C}''(\alpha + \beta^T X) \left( \frac{1}{X^T X} \right)),
\]  

(3.5)

where \( \mathbb{C}''(\alpha + \beta^T X) \) is the second derivative of \( \mathbb{C}(\alpha) \) evaluated at \( \alpha + \beta^T X. \)

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

\[
F(\alpha, \beta) = E(\mathbb{C}''(\alpha) \left( \frac{1}{X^T X} \right)),
\]

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

The conditional log-likelihood, which varies for different exponential family distributions of \( Y|X \), can be written as

\[
\mathbb{C}_n(\alpha, \beta) = \sum_{i=1}^n \mathbb{C}(\theta_i),
\]

where \( \theta_i = \alpha + \beta^T X_i \), and different functions \( \mathbb{C}(\theta) \) are summarized in Table 1. We next briefly review Fisher scoring, which is the standard iterative method for maximizing \( \mathbb{C}_n(\alpha, \beta) \), as background for the alternating envelope algorithm (Algorithm 1). At each iteration of the Fisher scoring method, the update step for \( \hat{\beta} \) can be summarized in the form of a weighted least squares (WLS) estimator as follows:

\[
\hat{\beta} \leftarrow S_{X|\hat{\beta}}^{-1} S_{X|\hat{\theta}} \hat{\theta},
\]

(3.6)

where \( \hat{W} = W(\hat{\theta}) = W(\hat{\alpha} + \hat{\beta}^T X) \) is the weight at the current iteration, \( \hat{V} = \hat{\theta} + [Y - \mu(\hat{\theta})]/\hat{W} \) is a pseudo-response variable at the current iteration, the weighted covariance \( S_{X|\hat{\theta}} \) is the sample estimator of \( \Sigma_{X|W} \) and the sample weighted cross-covariance \( S_{X|\hat{\beta}} \) is defined similarly. Upon convergence of the Fisher scoring process, the estimator \( \hat{\beta} = S_{X|\hat{\beta}}^{-1} S_{X|\hat{\theta}} \hat{\theta} \) is a function of the estimators \( \hat{\alpha}, \hat{\beta}, \hat{W}, \) and \( \hat{V} \) at the final iteration.

Assuming normal predictors, it follows from Section 3.1 that the full log-likelihood can be written as

\[
L_n(\alpha, \Gamma, \eta) = C_n(\alpha, \beta) + M_n(\Gamma),
\]

where \( \beta = \Gamma \eta \) is the coefficient vector of interest and \( M_n(\Gamma) \) is given in Proposition 3. For fixed \( \Gamma \in \mathbb{R}^{p \times a}, \) the Fisher scoring method of fitting the GLM of \( Y \) on \( X \) leads to

\[
\hat{n}_\Gamma = (\Gamma^T S_{X|\hat{\theta}} \Gamma)^{-1} \Gamma^T S_{X|\hat{\theta}} \hat{\theta},
\]

(3.7)

where the optimization over \( \Gamma \) treats \( \hat{n}_\Gamma \) as a function of \( \Gamma \) instead of as fixed. Since we are able to compute the analytical form of the matrix derivative of \( \partial L_n(\Gamma) / \partial \Gamma \), which is summarized in Lemma 1 of Supplement Section C, the alternating update of \( \Gamma \) based on \( \hat{n}_\Gamma \) is more efficient than updating \( \Gamma \) with fixed \( \eta \). 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 \texttt{nlm} Matlab package by Ross A. Lippert (http://web.mit.edu/rippr/www/software/) with the analytical form of \( \partial L_n(\Gamma) / \partial \Gamma \), and Fisher scoring was the default method in the Matlab package glmfit.

**Algorithm 1.** The alternating algorithm for GLM envelope estimator.

1. Initialize \( \hat{\alpha}, \hat{\beta}, \hat{\theta}, \hat{W}, \) and \( \hat{V} \) using the standard estimators. Initialize \( \hat{\Gamma} \) using the 1D algorithm (Algorithm 2) discussed in Section 4.1.

2. Alternating the following steps until the convergence of \( \hat{\beta}_{env} \) or the maximum number of iterations is reached.

   (2a) Update \( \hat{\Gamma} \) by maximizing (3.7) over Grassmann manifold \( \mathcal{G}_a \), where \( C_n(\Gamma) \) is calculated based on \( \hat{\alpha}, \hat{\beta}, \hat{W}, \) and \( \hat{V} \) at current step.

   (2b) Update \( \hat{\alpha} \) and \( \hat{\eta} \) by the Fisher scoring method of fitting GLM of \( Y \) on \( \hat{n}_\Gamma X \). Then set \( \hat{\beta}_{env} = \hat{\Gamma} \hat{\eta} \) and simultaneously update \( \hat{\beta}, \hat{W}, \) and \( \hat{V} \).

3.3 Asymptotic Properties With Normal Predictors

In this section we describe the asymptotic properties of envelope estimators in regression when \( \alpha \) (i.e., orthogonal parameter \( a \) in GLM) and \( \beta \) are orthogonal parameter vectors and \( X \) is normally distributed. We also contrast the asymptotic behavior of the envelope estimator \( \hat{\beta}_{env} \) with that of the estimator \( \hat{\beta} \) from \( C_n(\alpha, \beta) \), and comment on other settings at the end of the 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 \( \alpha, \eta, \Sigma, \Omega, \) and \( \Gamma. \) Since the parameters \( \eta, \Sigma, \) and \( \Omega \) depend on the basis \( \Gamma \) and the objective function is invariant under orthogonal transformations of \( \Gamma \), the estimators of these parameters are not unique. Hence, we consider only the asymptotic properties of the estimable functions \( \alpha, \beta, \Gamma = \Gamma \eta \) and \( \Sigma_X = \Gamma \Sigma \Gamma^T + \Gamma \Sigma \Omega \Gamma^T \), which are invariant under choice of basis \( \Gamma \) and thus have unique maximizers. Under the normality assumption, \( X \sim N(\mu_X, \Sigma_X) \), we neglect the mean vector \( \mu_X \) since it is orthogonal to all of the other parameters. We define
the following parameters \( \xi \) and estimable functions \( h \):\

\[
\xi = \begin{pmatrix}
\xi_1 \\
\xi_2 \\
\xi_3 \\
\xi_4 \\
\xi_5
\end{pmatrix} := \begin{pmatrix}
\alpha \\
\eta \\
\text{vec}(\Gamma) \\
\text{vech}(\Omega)
\end{pmatrix},
\]

\[
h = \begin{pmatrix}
h_1(\xi) \\
h_2(\xi) \\
h_3(\xi)
\end{pmatrix} := \begin{pmatrix}
\alpha \\
\beta \\
\text{vech}(\Sigma_X)
\end{pmatrix}.
\]

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

**Proposition 4.** Assume that for \( i = 1, \ldots, n \), the predictors \( X_i \) are independent copies of a normal random vector \( X \) with mean \( \mu_X \) and variance \( \Sigma_X > 0 \), and that the data \((Y_i, X_i)\) are independent copies of \((Y, X)\) with finite fourth moments. Assume also that \( \alpha \) and \( \beta \) are orthogonal parameter vectors. Then, as \( n \rightarrow \infty \), \( \sqrt{n}(\hat{\beta}_{env} - \beta_i) \) converges to a normal vector with mean 0 and covariance matrix

\[
\text{avar}(\sqrt{n}\hat{\beta}_{env}) = \text{var}(\sqrt{n}\hat{\beta}_T) + \text{var}(\sqrt{n}\Omega_T\hat{\beta}_G) = P_T V_{\beta\beta} P_T + (\eta^T \otimes \Gamma_0) M^{-1} (\eta \otimes \Gamma_0^T) \leq \text{avar}(\sqrt{n}\hat{\beta}),
\]

where

\[
M = (\eta \otimes \Gamma_0^T) V_{\beta\beta}^{-1} (\eta^T \otimes \Gamma_0) + \Omega \otimes \Omega_0^{-1} + \Omega^{-1} \otimes \Omega_0 - 2I_n(p-u), \quad \Omega = \Gamma^T \Sigma_X \Gamma \text{ and } \Omega_0 = \Gamma_0^T \Sigma_X \Gamma_0.
\]

The conditional log-likelihood is reflected in \( \text{avar}(\sqrt{n}\hat{\beta}_{env}) \) primarily through the asymptotic variance \( V_{\beta\beta} \), while \( \Omega \) and \( \Omega_0 \) stem from the normal marginal likelihood of \( X \). The span of \( \Gamma \) reduces both \( V_{\beta\beta} \) and \( \Sigma_X \), so the envelope serves as a link between the conditional and marginal likelihoods in the asymptotic variance. The first addend \( \text{avar}(\sqrt{n}\hat{\beta}_T) \) is the asymptotic covariance of the estimator given the envelope. The second addend \( \text{avar}(\sqrt{n}\Omega_T\hat{\beta}_G) \) 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 \( \hat{\beta} \).

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

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 \( \Sigma_X = \sigma^2 I_p, \sigma^2 > 0 \) then \( \hat{\varepsilon}_{C_{n\hat{\beta}}}(\beta_i) = \text{span}(\beta_i) \) and \( \text{avar}(\sqrt{n}\hat{\beta}_{env}) = \text{var}(\sqrt{n}\hat{\beta}) = V_{\beta\beta} \).

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 \( \hat{\varepsilon}_{C_{n\hat{\beta}}}(\beta_i) \) and the marginal distribution of \( X \), as shown in Proposition 2, and \( \hat{\beta}_{env} \) is still a \( \sqrt{n} \)-consistent estimator of \( \beta \). If the linearity condition is substantially violated, we can still use the objective function \( \{C_n(\alpha, \beta) + M_n(\Gamma)\} \) to estimate \( \beta \) within \( \hat{\varepsilon}_{C_{n\hat{\beta}}}(\beta_i) \) but this envelope may no longer equal \( \hat{\varepsilon}_{C_{n\hat{\beta}}}(\beta_i) \). Nevertheless, as demonstrated later in Corollary 2, this still has the potential to yield an estimator of \( \beta \) with smaller asymptotic variance than \( \hat{\beta} \), 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, e.g., 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 2014), which is restated as Algorithm 2 in its population version, gives us a generic moment-based estimator of the envelope $E_M(U)$, requiring only matrices $\hat{M}$ and $\hat{U}$. Setting $\hat{U} = \hat{\phi}^T \hat{\phi} \in S^{q \times q}$ and $\hat{M}$ equal to a $\sqrt{n}$-consistent estimator of $V_{\phi}(\theta_1)$ in $S^{q \times q}$, it follows from Cook and Zhang (2014, Proposition 6) that the 1D algorithm provides a $\sqrt{n}$-consistent estimator $P_{G_{0}} = G_{0} \hat{G}_{n}$ of the projection onto $E_{\phi}(\theta_1)$. $G_{0}$ is obtained from the sample version of the 1D algorithm and $u = \dim(E_{\phi}(\theta_1))$ is assumed to be known.

The moment-based envelope estimator $\hat{\phi}_{\text{env}} = P_{G_{0}} \hat{\phi}$ is then a $\sqrt{n}$-consistent estimator of $\hat{\phi}$. For example, the moment-based estimator for GLMs is obtained by letting $\hat{U} = \hat{\beta}^T \hat{\beta}$ and $\hat{M} = S_{X \beta}^{-1} / \{n^{-1} \sum_{i=1}^{n} (-C' \hat{\beta}))\}$ 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 $W(\theta)$ depends on the parameter $\beta$ 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.

Algorithm 2 The 1D algorithm.

1. Set initial value $g_{0} = G_{0} = 0$.
2. For $k = 0, \ldots, u - 1$, $G_{k+1}$ in $\mathbb{R}^{q}$ is obtained direction in the envelope $E_M(U)$ as follows,
   (a) Let $G_{k} = \{g_{1}, \ldots, g_{k}\}$ if $k \geq 1$ and let $(G_{k}, G_{0k})$ be an orthogonal basis for $\mathbb{R}^{q}$.
   (b) Define the stepwise objective function
   $$J_{k}(w) = \log(w^{T}A_{k}w) + \log(w^{T}B_{k}^{-1}w),$$
   (4.1)
   where $A_{k} = G_{0k}^{T}M_{G_{0k}}$, $B_{k} = (G_{0k}^{T}M_{G_{0k}} + G_{0k}^{T}UG_{0k})$ and $w \in \mathbb{R}^{q-k}$.
   (c) Solve $w_{k+1} = \arg \min_{w} J_{k}(w)$ subject to a length constraint $w^{T}w = 1$.
   (d) Define $G_{k+1} = G_{0k}w_{k+1}$ to be the unit length $(k+1)$th stepwise direction.

The 1D algorithm can also play the role of 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 $L_{n}(\Gamma) = L_{n}(\Gamma \Omega)$ for any orthogonal matrix $\Omega \in \mathbb{R}^{u \times u}$. Maximization of $L_{n}$ is thus over the Grassmannian $G_{n,p}$. Since $u(p-u)$ real numbers are required to specify an element of $G_{n,p}$ 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 2(a) in Algorithm 1, the iterative Grassmann manifold optimization of $L_{n}(\Gamma)$ 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 $\Gamma$ for $E_{\phi}(\theta_{1})$ is known and write the envelope estimator as $P_{\Gamma} \hat{\phi}$. Then since the envelope reduces $V_{\phi}(\theta_{1})$, we have
$$\text{var}(\sqrt{n} \hat{\phi}) = P_{\Gamma}V_{\phi}(\theta_{1})P_{\Gamma} + \mathbb{Q}_{\Gamma}V_{\phi}(\theta_{1})\mathbb{Q}_{\Gamma},$$
$$\text{var}(\sqrt{n}P_{\Gamma} \hat{\phi}) = P_{\Gamma}V_{\phi}(\theta_{1})P_{\Gamma} \leq V_{\phi}(\theta_{1}).$$

These relationships allow some straightforward intuition by writing $\sqrt{n}(\hat{\phi} - \phi) = \sqrt{n}(P_{\Gamma} \hat{\phi} - \phi) + \sqrt{n}Q_{\Gamma} \hat{\phi}$. The second term $\sqrt{n}Q_{\Gamma} \hat{\phi}$ is asymptotically normal with mean 0 and variance $V_{\phi}(\theta_{1})\mathbb{Q}_{\Gamma}$, and is asymptotically independent of $\sqrt{n}(P_{\Gamma} \hat{\phi} - \phi)$. Consequently, we think of $Q_{\Gamma} \hat{\phi}$ as the immaterial information in $\hat{\phi}$. The envelope estimator then achieves efficiency gains by essentially eliminating the immaterial variation $V_{\phi}(\theta_{1})\mathbb{Q}_{\Gamma}$, the greatest gains being achieved when $Q_{\Gamma}V_{\phi}(\theta_{1})\mathbb{Q}_{\Gamma}$ is large relative to $P_{\Gamma}V_{\phi}(\theta_{1})P_{\Gamma}$. Of course, we will typically need to estimate $\Gamma$ in practice, which will mitigate the asymptotic advantages when $\Gamma$ 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 $\hat{\phi}_{\text{env}} = P_{G_{0}} \hat{\phi}$, 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 $\hat{\phi}$ is obtained by minimizing an objective function $\hat{\phi} = \arg \min_{\phi \in \mathbb{R}^{p}} F_{n}(\phi)$ then, as an alternative to the moment-based estimator $\hat{\phi}_{\text{env}} = P_{G_{0}} \hat{\phi}$, an envelope estimator can be constructed as $\hat{\phi}_{\text{env}} = G_{0} \hat{\eta}$, where $\hat{\eta} = \arg \min_{\eta \in \mathbb{R}^{p}} F_{n}(G_{0} \eta)$. We refer to this as the objective-function-based estimator. Sometimes these two approaches are identical, that is, 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 $\theta$ as $\hat{\theta} = \arg \max_{\theta} L_{n}(\theta)$, where $L_{n}(\theta)$ is a log-likelihood that is twice continuously differentiable in an open neighborhood of $\theta$. Then under standard conditions $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically normal with mean 0 and covariance matrix $V_{\theta}(\theta) = F^{-1}(\theta)$, where $F$ is the Fisher information matrix for $\theta$. The asymptotic covariance matrix of $\hat{\phi}$, $V_{\phi}(\theta)$, is the lower right block of $V_{\theta}(\theta)$.

Recall that for envelope models, we can write the log-likelihood as $L_{n}(\theta) = L_{n}(\psi, \phi) = L_{n}(\psi, \Gamma \phi)$. To obtain MLE, we need to maximize $L_{n}(\psi, \Gamma \phi)$ over $\psi$, $\eta$ and $\Gamma \in \mathbb{R}^{p \times u}$, where $\Gamma$ is semiorthogonal basis for $E_{\phi}(\theta_{1})$ and $u = \dim(E_{\phi}(\theta_{1})/\phi_{1})$. We first discuss the potential advantages of envelopes by considering the maximization of $L_{n}(\psi, \Gamma \phi)$ over $\psi$ and $\eta$ with known $\Gamma$. Since $\phi_{1} \in E_{\phi}(\theta_{1})/\phi_{1})$, we have that $\phi_{1} = \Gamma \eta$, for $\eta \in \mathbb{R}^{u}$. Consequently, for known $\Gamma$, the envelope
estimators become

\[
\hat{\theta}_{\Gamma} = \arg \max_{\theta} L_n(\psi, \Gamma). 
\]

\[
\hat{\phi}_{\Gamma} = \Gamma \hat{\eta}_{\Gamma} 
\]

\[
\hat{\theta} = (\hat{\phi}_{\Gamma}, \hat{\phi}^T_{\Gamma})^T. 
\]

Any basis for \(E_{V_{\phi}(\theta)}(\phi)\) will give the same solution \(\hat{\phi}_{\Gamma}\) for an orthogonal matrix \(O\), write \(\phi = \Gamma O \eta\). Then \(\hat{\eta}_{\Gamma} = O^T \hat{\eta}_{\Gamma}\).

The estimator \(\hat{\phi}_{\Gamma}\) given in (4.3) is in general different from the moment-based estimator \(P_{\phi} \hat{\phi}\) 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 \(n \to \infty\), \(\sqrt{n}(\hat{\phi}_{\Gamma} - \phi)\) converges to a normal random vector with mean 0 and asymptotic covariance

\[
\text{avar}(\sqrt{n} \hat{\phi}_{\Gamma}) = P_{\Gamma} V_{\phi}(\theta) P_{\Gamma} \leq V_{\phi}(\theta). 
\]

The \(\sqrt{n}\)-consistent nuisance parameter estimator also satisfies \(\text{avar}(\sqrt{n} \hat{\phi}_{\Gamma}) \leq \text{avar}(\sqrt{n} \hat{\phi})\).

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

**Corollary 2.** If \(\Gamma\) is a basis for an arbitrary envelope \(E_{M}(\phi)\), where \(M\) is a symmetric positive definite matrix, then

\[
\text{avar}(\sqrt{n} \hat{\phi}_{\Gamma}) = \Gamma \{\Gamma^T V_{\phi}(\theta) \Gamma\}^{-1} \Gamma^T \leq V_{\phi}(\theta). 
\]

The above expression shows that \(\Gamma\) in the asymptotic covariance of the envelope estimator \(\hat{\phi}_{\Gamma}\), while the envelope by Definition 3 makes \(\text{avar}(\sqrt{n} \hat{\phi}_{\Gamma})\) more interpretable because the material and the immaterial variances are separable.

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

**Proposition 6.** If the estimated basis \(\hat{\Gamma}\) for \(E_{V_{\phi}(\theta)}(\phi)\) is obtained by the 1D algorithm (Algorithm 2), then the envelope estimator \(\hat{\phi}_{\text{env}} = \hat{\phi}_{\Gamma}\) is a \(\sqrt{n}\)-consistent estimator of \(\phi\).

The envelope estimator \(\hat{\phi}_{\text{env}}\) depends on the prespecified estimator \(\hat{\theta}\) through \(M = V_{\phi}(\theta)\) and \(U = \phi \phi^T\). Although it is \(\sqrt{n}\)-consistent, we have found empirically that, when \(M\) depends nontrivially on \(\theta\), it can often be improved by iterating so the current estimate of \(\theta\) is used to construct estimates of \(M\) and \(U\). The iteration can be implemented as Algorithm 3.

**Algorithm 3** The iterative envelope algorithm for MLE.

1. Initialize \(\hat{\theta}_0 = \hat{\theta}\), and let \(\hat{M}_0\) and \(\hat{U}_0\) be the estimators of \(M\) and \(U\) based on the 4th envelope estimator \(\hat{\theta}_4\) of \(\theta\), so \(\hat{M}_0\) and \(\hat{U}_0\) are based only on the prespecified estimator.
2. For \(k = 0, 1, \ldots\) iterate as follows until a measure of the change between \(\hat{\phi}_{k+1}\) and \(\hat{\phi}_k\) is sufficiently small.

(a) Using the 1D algorithm, construct an estimated basis \(\hat{\Gamma}_k\) for \(E_{V_{\phi}(\theta)}(\phi)\) using \(\hat{M}_k = V_{\phi}(\theta)\) and \(\hat{U}_k = \hat{\phi}_k \hat{\phi}^T_k\).

(b) Set \(\hat{\theta}_{k+1} = \hat{\theta}_k\) from (4.4).

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 Sections 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 \(E_{\Sigma_{X}(\beta)}\), where \(\Sigma_{X}(\beta) = \text{cov}(\sqrt{W}X)\) is the covariance of the weighted predictors \(\sqrt{W}X\) and \(W = \exp(\beta^T X)/(1 + \exp(\beta^T X))\). 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: \(Y_i | X_i \sim \text{Bernoulli}(\logit(\beta^T X))\), \(\beta = (\beta_1, \beta_2)^T\) \((0.25, 0.25)^T\) and \(X_i \sim N(0, \Sigma_X)\). We let \(\text{span}(\beta)\) be the principal eigenvector of \(\Sigma_X\) 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 3(a), in which we used two lines to indicate the directions of the two estimators: standard estimator \(\hat{\beta} = (-0.047, 0.692)^T\) with asymptotic standard errors \(\text{SE}(\hat{\beta}) = (0.418, 0.429)^T\), and envelope estimator \(\hat{\beta}_{\text{env}} = (0.321, 0.319)^T\) (based on Algorithm 1) with asymptotic standard errors \(\text{SE}(\hat{\beta}_{\text{env}}) = (0.058, 0.057)^T\). The components of \(\hat{\beta}_{\text{env}}\) are large relative to their standard errors, while the components of \(\hat{\beta}\) are not. The alternative moment-based-envelope estimators discussed in Sections 4.1 and 4.2, \(\hat{\beta}_{\text{env}, 2} = (0.323, 0.317)^T\) (based on Algorithm 2 as indicated by the added subscript) and \(\hat{\beta}_{\text{env}, 3} = (0.320, 0.320)^T\) (based on Algorithm 3), were both very close to the likelihood-based estimator \(\hat{\beta}_{\text{env}}\). We also plotted the weighted predictors \(\sqrt{W}X\) in Figure 3(b) so that we can see from the figure that the envelope is the principal eigenvector of \(\Sigma_{X(W)}\).

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

We use the Australia Institute of Sport (AIS) data to illustrate the phenomenon in this simulation setting. This dataset, originally from Cook and Weisberg (1994; p. 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 3(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 has asymptotic standard errors $\text{SE}(\hat{\beta}) = (0.034, 0.022)^T$, while the likelihood-based envelope estimator $\hat{\beta}_{\text{env}} = (0.085, 0.086)^T$, which was obtained by Algorithm 1, has asymptotic standard errors $\text{SE}(\hat{\beta}_{\text{env}}) = (0.014, 0.014)^T$. The envelope estimator compared to the standard estimator, had 40% and 60% smaller standard errors for the two coefficients in $\beta$.

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 reweighted partial least squares estimators for GLMs (IRPLS; Marx 1996), and (3) envelope estimators based on Algorithm 1. The 1D algorithm (Algorithm 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 estimator always did better than the IRPLS estimator in terms of angle with increased sample sizes. At the same time, the standard estimators over the standard estimators became more substantial.

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.

| GLM simulations with $\Sigma_X = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T$ |
|---------------------------------|------|------|---------|
|                                | Standard | IRPLS | Envelope |
| **Poisson**                    |        |      |         |
| $n = 50$                       | ∆     | 38 (1.5) | 55 (0.8) | **25 (2.0)** |
|                               | LR    | 1.4 | 0.21 | 1.4 |
| $n = 200$                      | ∆     | 21 (1.1) | 48 (0.6) | **7 (0.8)** |
|                               | LR    | 1.09 | 0.25 | 1.03 |
| $n = 800$                      | ∆     | 9 (0.5) | 46 (0.4) | **2 (0.3)** |
|                               | LR    | 1.02 | 0.28 | 1.01 |
| **Logistic**                   |        |      |         |
| $n = 50$                       | ∆     | 75 (1) | 84 (1) | **74 (2)** |
|                               | LR    | 10.2 | 0.22 | 8.1 |
| $n = 200$                      | ∆     | 62 (2) | 81 (1) | **36 (3)** |
|                               | LR    | 3.0 | 0.12 | 2.0 |
| $n = 800$                      | ∆     | 45 (2) | 77 (0.4) | **9 (1)** |
|                               | LR    | 1.6 | 0.087 | 1.0 |

**NOTE:** We summarize averaged angles $\angle = \angle(\beta, \hat{\beta})$ in degrees and length ratios $LR = ||\hat{\beta}||/||\beta||$ 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.
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 dataset consists of $n = 285$ tissue samples, comprised of 180 normal tissues and 105 adenomas. Studies on similar datasets 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 $\beta$ were 2.2 to 12.5 times larger than the bootstrap standard errors of the elements in $\hat{\beta}_{env}$, 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 (nonestimative) variation in the envelope error rate, we took $\hat{\beta}_{env}$ to be the true $\beta$, sampled the predictors 1000 times and predicted the response using $\hat{\beta} = \hat{\beta}_{env}$. 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 dataset (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 linear discriminant analysis (LDA).
Table 5. Five-fold cross-validation error rates based on Fisher’s LDA and expressed as percentages

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

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 $\hat{\beta}_{OLS}$ is always proportional to the LDA direction. Hence, we used $\hat{P}_{OLS}X$ as the discriminant direction and fitted the LDA classification rule based on it. We similarly replaced the OLS coefficient vector with the coefficient vector $\hat{P}_{PLS}$ estimated by SIMPLS algorithm and the coefficient vector and $\hat{P}_{env}$ 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, sec. 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 $\chi^2$ statistic $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/\hat{Y}_i$ 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.

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

|                  | Standard | IRPLS | Envelope |
|------------------|----------|-------|----------|
| Pearson’s $\chi^2$ | 642.1    | 556.1 | 553.7    |
| S.E.             | 7.8      | 1.6   | 1.6      |

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, e.g., Zhou and Li 2014; Zhou, Li, and Zhu 2013).

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 $X$ is far from elliptical. Finally, the most generic envelope estimator for an arbitrary prespecified estimator $\hat{\beta}$ is proposed as $\hat{\beta}_{env} = P_{G_u}\hat{\beta}$ or $\hat{\beta}_{env} = \hat{G}_u\hat{G}_u$ if $\hat{\beta}$ is obtained based on objective function, where $\hat{G}_u$ 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 semiparametric 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 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 $\hat{M}$ and $\hat{U}$ in our algorithms. Properties of such extensions are yet to be studied.

SUPPLEMENTARY MATERIALS

The supplementary materials contain additional proofs and technical details.

[Received June 2013. Revised September 2014.]

REFERENCES

Agresti, A. (2002), Categorical Data Analysis, (2nd ed.) New York: Wiley. [603,610]
Conway, J. (1990), A Course in Functional Analysis (2nd ed.), New York: Springer. [399]
Cook, R. D. (1996), “Graphics for Regressions With a Binary Response,” Journal of the American Statistical Association, 91, 983–992. [603]
— (1998), Regression Graphics: Ideas for Studying Regressions Through Graphs, New York: Wiley. [603]
Cook, R. D., Helland, I. S., and Su, Z. (2013), “Envelopes and Partial Least Squares Regression,” Journal of the Royal Statistical Society, Series B, 75, 851–877. [601]
Cook, R. D., Li, B., and Chiaromonte, F. (2010), “Envelope Models for Simposionand and Efficient Multivariate Linear Regression” (with discussion), Statistica Sinica, 20, 927–1010. [599]
Cook, R. D., and Weisberg, S. (1994), An Introduction to Regression Graphics, New York: Wiley. [607]
Cook, R. D., and Zhang, X. (2014), “Algorithms for Envelope Estimation,” arXiv:1403.4138. [601,606]
— (2015), “Simultaneous Envelopes for Multivariate Linear Regression,” Technometrics, 57, 11–25. [601]
Cox, D. R. (1972), “Regression Models and Life-Tables” (with discussion), Journal of the Royal Statistical Society, Series B, 34, 187–220. [609]
— (1975), “Partial Likelihood,” Biometrika, 62, 269–276. [609]
Cox, D. R., and Reid, N. (1987), “Parameter Orthogonality and Approximate Conditional Inference,” *Journal of the Royal Statistical Society*, Series B, 49, 1–39. [604]
de Jong, S. (1993), “SIMPLS: An Alternative Approach to Partial Least Squares Regression,” *Chemometrics and Intelligent Laboratory Systems*, 18, 251–263. [601, 608]

Eaton, M. L. (1986), “A Characterization of Spherical Distributions,” *Journal of Multivariate Analysis*, 20, 272–276. [603]

Hawkins, D. M., and Maboudou-Tchao, E. M. (2013), “Smoothed Linear Modeling for Smooth Spectral Data,” *International Journal of Spectroscopy*, 8 pages. [609]

Kenward, M. G. (1987), “A Method for Comparing Profiles of Repeated Measurements,” *Journal of the Royal Statistical Society*, Series C, 36, 296–308. [600]

Ledoit, O., and Wolf, M. (2004), “A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices,” *Journal of Multivariate Analysis*, 88, 365–411. [610]

Li, B., and Wang, S. (2007), “On Directional Regression for Dimension Reduction,” *Journal of American Statistical Association*, 102, 997–1008. [603]

Marx, B. D. (1996), “Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression,” *Technometrics*, 38, 374–381. [608]

Nygard, S., and Borgan, O. (2008), “Partial Least Squares Cox Regression for Genome-Wide Data,” *Lifetime Data Analysis*, 14, 179–195. [609]

Schott, J. R. (2013), “On the Likelihood Ratio Test for Envelope Models in Multivariate Linear Regression,” *Biometrika*, 100, 531–537. [601]

Su, Z., and Cook, R. D. (2011), “Partial Envelopes for Efficient Estimation in Multivariate Linear Regression,” *Biometrika*, 98, 133–146. [601, 602, 606]

——— (2012) “Inner Envelopes: Efficient Estimation in Multivariate Linear Models,” *Biometrika*, 99, 687–702. [601]

——— (2013) “Estimation of Multivariate Means With Heteroscedastic Errors Using Envelope Models,” *Statistica Sinica*, 23, 213–230. [601]

ter Braak, C. J. F., and de Jong, S. (1998), “The Objective Function of Partial Least Squares Regression,” *Journal of Chemometrics*, 12, 41–54. [601]

Wold, H. (1966), *Estimation of Principal Components and Related Models by Iterative Least Squares*, New York: Academic Press. [608]

Ye, J. (2007), “Least Squares Linear Discriminant Analysis,” in *Proceedings of the 24th International Conference on Machine Learning*, pp. 1087–1093. [610]

Yohai, V. J., Stahel, W. A., and Zamar, R. H. (1991), “A Procedure for Robust Estimation and Inference in Linear Regression,” in *Directions in Robust Statistics and Diagnostics, Part II*, eds. W. A. Stahel and S. W. Weisberg, New York: Springer, pp. 365–374. [605]

Zhou, H., and Li, L. (2014), “Regularized Matrix Regression,” *Journal of the Royal Statistical Society*, Series B, 76, 463–483. [610]

Zhou, H., Li, L., and Zhu, H. (2013), “Tensor Regression With Applications in Neuroimaging Data Analysis,” *Journal of the American Statistical Association*, 108, 540–552. [610]