Biometrika (2020), **xx**, x, *pp*. 1–17
Printed in Great Britain

Envelopes in Multivariate Regression Models with Nonlinearity and Heteroscedasticity

BY X. ZHANG

Department of Statistics, Florida State University, 117 N. Woodward Ave, Tallahassee, Florida 32306, U.S.A.

henry@stat.fsu.edu

C. E. LEE

Department of Business Analytics and Statistics, University of Tennessee, Knoxville, 916 Volunteer Blvd, Knoxville, Tennessee 37996, U.S.A.

clee88@utk.edu

AND X. SHAO

Department of Statistics, University of Illinois at Urbana Champaign, 725 South Wright St, Champaign, Illinois 61820, U.S.A.

xshao@illinois.edu

SUMMARY

Envelopes have been proposed in recent years as a nascent methodology for sufficient dimension reduction and efficient parameter estimation in multivariate linear models. We extend the classical definition of envelopes in Cook et al. (2010) to incorporate a nonlinear conditional mean function and a heteroscedastic error. Given any two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^r$, we propose two new model-free envelopes – called the Martingale Difference Divergence Envelope (MDDE) and the Central Mean Envelope (CME) – and study their relationships with the standard envelope in the context of response reduction in the multivariate linear models. The MDDE effectively captures the nonlinearity in the conditional mean without imposing any parametric structure or requiring any tuning in estimation. Heteroscedasticity, or the non-constant conditional covariance of $Y \mid X$, is further detected by the CME based on a slicing scheme for the data. We reveal the nested structure of different envelopes: (1) the CME contains the MDDE, with equality when $Y \mid X$ has a constant conditional covariance; and (2) the MDDE contains the standard envelope, with equality when $Y \mid X$ has a linear conditional mean. We further develop an estimation procedure that obtains the MDDE first and then estimates the additional envelope components in the CME. We establish consistency in envelope estimation of MDDE and CME without stringent model assumptions. Simulations and real data analysis demonstrate the advantages of MDDE and CME over standard envelope in dimension reduction.

Key Words: Envelope Models; Heteroscedasticity; Multivariate Linear Model; Nonlinear Dependence; Sufficient Dimension Reduction.

1. INTRODUCTION

The first envelope model was proposed by Cook et al. (2010), in the context of multivariate linear model of multivariate response $Y \in \mathbb{R}^r$ on predictor $X \in \mathbb{R}^p$,

$$Y_i = \alpha + \beta X_i + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where $\alpha \in \mathbb{R}^r$, $\beta \in \mathbb{R}^{r\times p}$, and $\varepsilon_i \sim N(0, \Sigma)$, $\Sigma > 0$, is independent of X_i. The goal of envelope methodology is to increase efficiency in estimating the regression coefficient β. An envelope is essentially a targeted dimension reduction subspace that contains the material variation in the data for the goal of estimating β. Envelope models and methods have been recently developed in a series of regression problems, see Cook (2018) for an overview, and in general multivariate analysis (Cook & Zhang, 2015a).

The multivariate linear model (1) has restrictive assumptions on the conditional mean and covariance: (i) $\mathrm{E}(Y \mid X = x) = \alpha + \beta x$ is linear in x, and (ii) $\mathrm{cov}(Y \mid X = x) = \Sigma$ does not depend on x. Most of the existing envelope methods in multivariate linear model rely on these two assumptions (e.g., Cook et al., 2010, 2013; Cook & Zhang, 2015b; Khare et al., 2017). As a result, these methods may suffer severely from any violation of the two assumptions. Whilst most existing envelope methods concentrate on parametric settings, particularly for the multivariate linear models, we aim to address nonlinearity and heteroscedasticity by revisiting model-free envelope pursuit in a more flexible dimension reduction setting that is similar to Cook et al. (2007).

The response envelope in Cook et al. (2010) is constructed as the smallest subspace $\mathcal{S} \subseteq \mathbb{R}^r$ such that

$$\text{(i) } Q_{\mathcal{S}}Y \mid X \sim Q_{\mathcal{S}}Y, \quad \text{(ii) } P_{\mathcal{S}}Y \perp\!\!\!\perp Q_{\mathcal{S}}Y \mid X, \tag{2}$$

where $P_{\mathcal{S}}$ is the projection onto $\mathcal{S}$, and $Q_{\mathcal{S}} = I_r - P_{\mathcal{S}}$ is the projection onto the orthogonal complement of $\mathcal{S}$, i.e. $\mathcal{S}^{\perp}$. The two statements in (2) imply that the distribution of $Q_{\mathcal{S}}Y$ is not affected by X and $Q_{\mathcal{S}}Y$ is conditionally independent of $P_{\mathcal{S}}Y$ given X. Therefore, $Q_{\mathcal{S}}Y$ is immaterial because it does not contain useful information about β but only brings extraneous variation in estimation. Under the multivariate linear model (1), the conditions in (2) are equivalent to the following parametric conditions,

$$\text{(i) } \mathrm{span}(\beta) \subseteq \mathcal{S}, \quad \text{(ii) } \Sigma = P_{\mathcal{S}}\Sigma P_{\mathcal{S}} + Q_{\mathcal{S}}\Sigma Q_{\mathcal{S}}, \tag{3}$$

where $\mathrm{span}(\beta) \subseteq \mathbb{R}^r$ is the subspace spanned by the column vectors of $\beta \in \mathbb{R}^{r\times p}$. The first statement in (3) implies that, by varying x, the changes in the conditional mean function $\mathrm{E}(Y \mid X = x) = \alpha + \beta x$ lie within the subspace $\mathcal{S}$; and the second statement in (3) implies that, given X, we have conditionally uncorrelated components by projecting the response onto $\mathcal{S}$ and onto its orthogonal complement $\mathcal{S}^{\perp}$. The smallest such subspace is called the Σ-envelope of $\mathrm{span}(\beta)$ and is formally defined as follows.

DEFINITION 1. *(Cook et al., 2010) A subspace $\mathcal{S} \subseteq \mathbb{R}^r$ is said to be a reducing subspace of $\Sigma \in \mathbb{R}^{r\times r}$ if $\mathcal{S}$ decomposes Σ as $\Sigma = P_{\mathcal{S}}\Sigma P_{\mathcal{S}} + Q_{\mathcal{S}}\Sigma Q_{\mathcal{S}}$. The Σ-envelope of $\mathcal{B} \equiv \mathrm{span}(\beta)$, denoted by $\mathcal{E}_{\Sigma}(\mathcal{B})$ or $\mathcal{E}_{\Sigma}(\beta)$, is the intersection of all reducing subspaces of Σ that contain $\mathcal{B}$.*

By definition, the existence, uniqueness and minimal dimensionality of $\mathcal{E}_{\Sigma}(\beta)$ are guaranteed (Cook et al., 2010, Proposition 2.1). The idea of envelope methodology is to employ this dimension reduction subspace $\mathcal{E}_{\Sigma}(\beta)$ to improve estimation and prediction. Moreover, the envelope establishes a parametric link between the parameter of interests β, and the nuisance parameter Σ.

From the basic properties of conditional independence (Dawid, 1979), (2) is equivalent to $Q_{\mathcal{S}}Y \perp\!\!\!\perp (P_{\mathcal{S}}Y, X)$. A natural approach for estimating such $\mathcal{S}$ without assuming model (1) is to optimize over all subspaces $\mathcal{S} \subseteq \mathbb{R}^r$ such that the distance covariance (Székely et al., 2007; Székely & Rizzo, 2009) between $Q_{\mathcal{S}}Y$ and $(P_{\mathcal{S}}Y, X)$ are minimized. This idea shares the same spirit of some recent advances in sufficient dimension reduction with distance covariance (e.g., Sheng & Yin, 2016; Matteson & Tsay, 2017; Vepakomma et al., 2018). However, we consider an alternative approach to achieve a less ambitious goal that is more relevant to regression analysis.

Our proposal is inspired by the notion of the central mean subspace in dimension reduction (Cook & Li, 2002), which focuses on the conditional mean function $\mathrm{E}(Y \mid X)$ instead of the whole conditional distribution $Y \mid X$ when considering reduction of X. Analogous to the central mean subspace, we consider reduction of Y under the envelope model framework such that our focus is on prediction and inference from the conditional mean function. In particular, we replace the parametric envelope model assumptions in (3) with the more general mean dependence and conditional covariance reduction in the following. We define the central mean envelope (CME) as the smallest subspace $\mathcal{S} \subseteq \mathbb{R}^r$ such that,

$$\text{(i) } \mathrm{E}(Q_{\mathcal{S}}Y \mid X) = \mathrm{E}(Q_{\mathcal{S}}Y), \quad \text{(ii) } \mathrm{cov}(Q_{\mathcal{S}}Y, P_{\mathcal{S}}Y \mid X) = 0. \tag{4}$$

The above notion of CME nicely bridges the gap between the response reduction in (2) and the parametric response envelope in (3): (2) implies (4); (4) implies (3); and the three are equivalent under the multivariate linear model. Moreover, because the definition of CME is model-free, we employ a nonparametric measure called the martingale difference divergence matrix (Lee & Shao, 2018) to capture the general dependence in mean. We also allow the conditional covariance $\mathrm{cov}(Y \mid X = x) = \Sigma(x)$ to depend on x, and re-define Σ as $\Sigma = \mathrm{E}\{\mathrm{cov}(Y \mid X)\} = \mathrm{E}\{\Sigma(X)\}$, which reduces to the same Σ in the multivariate linear model (1). While the CME reduces all $\Sigma(x)$, a weaker condition than (4) leads to the martingale difference divergence envelope (Definition 3), which is the reducing subspace of Σ.

Whilst most existing envelope methods concentrate on parametric models, our work differs obviously by tackling a more challenging problem of model-free envelope estimation without linearity and constant covariance assumptions in linear models. Such an extension is far from trivial, and new techniques are required throughout our development. As such, our development for the central mean envelope enriches the dimension reduction techniques for conditional mean in regression (e.g., Cook & Li, 2002) and dimension reduction techniques in general. It complements the distance covariance type solutions (e.g., Sheng & Yin, 2016; Matteson & Tsay, 2017; Vepakomma et al., 2018). Moreover, the new definitions of central mean envelope (CME) and martingale difference divergence envelope (MDDE) are consistent with the standard envelope and bridges the gap between the standard envelopes in multivariate linear model and the general conditional independence in sufficient dimension reduction: the CME implies conditional independence of $Q_{\mathcal{S}}Y \perp\!\!\!\perp (P_{\mathcal{S}}Y, X)$ if $Y \mid X$ is normally distributed; the CME contains the MDDE with equality when $Y \mid X$ has a constant conditional covariance; the MDDE contains the standard envelope with equality when $Y \mid X$ has a linear conditional mean. It is also worth mentioning that the notion of CME is completely generic for any two random vectors X and Y. Parallel to the recent developments of standard envelopes, our framework of CME are not restricted to response reduction in regression; it can be extended straightforwardly to predictor reduction (Cook et al., 2013), simultaneous reduction (Cook & Zhang, 2015b) and even to tensor envelopes (Zhang & Li, 2017; Li & Zhang, 2017).

2. A BRIEF REVIEW OF MARTINGALE DIFFERENCE DIVERGENCE MATRIX

Lee & Shao (2018) introduced the martingale difference divergence matrix (MDDM), which can be viewed as an extension of martingale difference divergence (Shao & Zhang, 2014; Park et al., 2015) from a scalar to a matrix, and further applied it to the dimension reduction of a stationary multivariate time series. For two real-valued random vectors $Y \in \mathbb{R}^r$ and $X \in \mathbb{R}^p$, if $\mathrm{E}(\|Y\|^2 + \|X\|) < \infty$, then

$$M_{Y|X} \equiv \mathrm{MDDM}(Y \mid X) = -\mathrm{E}\left[\{Y - \mathrm{E}(Y)\}\{Y' - \mathrm{E}(Y')\}^T \|X - X'\|\right], \tag{5}$$

where (Y', X') is an independent copy of (Y, X). We introduced the notation $M_{Y|X}$ as the abbreviation of $\mathrm{MDDM}(Y \mid X)$ in Lee & Shao (2018). From (5), $M_{Y|X} \in \mathbb{R}^{r\times r}$ is a real, symmetric and positive semi-definite matrix. We assume $\mathrm{E}(\|Y\|^2 + \|X\|) < \infty$ in our exposition unless otherwise specified.

The span of $M_{Y|X}$ is closely related to the *mean dependence*: For two random vectors X and Y, we say X is independent of Y in mean if $\mathrm{E}(X \mid Y) = \mathrm{E}(X)$. Clearly, dependence in mean is not symmetric, X is independent of Y in mean does not imply Y is independent of X in mean.

As a direct consequence of Theorem 1 in Lee & Shao (2018), we have the following result.

LEMMA 1. *For all values of $x \in \mathbb{R}^p$ in the support of X, we have* $\mathrm{E}(Y \mid X = x) - \mathrm{E}(Y) \in \mathrm{span}(M_{Y|X})$. *Moreover,* $\mathrm{span}[\mathrm{cov}\{\mathrm{E}(Y \mid X)\}] = \mathrm{span}(M_{Y|X})$.

The above lemma suggests that we can simply eigen-decompose the matrix $M_{Y|X}$ and use the non-trivial eigenvectors to span the subspace $\mathcal{S}$ such that $Q_{\mathcal{S}}Y$ is independent of X in mean, i.e. $\mathrm{E}(Q_{\mathcal{S}}Y \mid X) = \mathrm{E}(Q_{\mathcal{S}}Y)$ in (4). Given a random sample $(X_i, Y_i)_{i=1}^n$ from the joint distribution of (X, Y), the sample version of $M_{Y|X}$ can be straightforwardly calculated as $\widehat{M}_{Y|X} = -\frac{1}{n^2}\sum_{k,l=1}^n (Y_k - \overline{Y}_n)(Y_l - \overline{Y}_n)^T \|X_k - X_l\|$ where $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$.

3. THE CENTRAL MEAN ENVELOPE

3.1. *Formal definition and properties*

In multivariate regression of Y on X, we are primarily interested in the conditional mean function $\mathrm{E}(Y \mid X)$. So we naturally want to know the subspace $\mathcal{S} \subseteq \mathbb{R}^r$ such that $\mathrm{E}(Y \mid X) = \mathrm{E}(P_{\mathcal{S}}Y \mid X) + \mathrm{E}(Q_{\mathcal{S}}Y \mid X) = \mathrm{E}(P_{\mathcal{S}}Y \mid X) + \mathrm{E}(Q_{\mathcal{S}}Y)$, and restrict our attention to $\mathrm{E}(P_{\mathcal{S}}Y \mid X)$. The immaterial part $Q_{\mathcal{S}}Y$ is mean independent of X and thus the variability in $Q_{\mathcal{S}}Y$ can not be reduced by regressing on X. As we stated in (4), we also want to have $\mathrm{cov}(Q_{\mathcal{S}}Y, P_{\mathcal{S}}Y \mid X) = 0$ so that the variation in $Q_{\mathcal{S}}Y$ also does not affect the regression analysis of $P_{\mathcal{S}}Y \mid X$ through correlation.

The next Lemma guarantees the existence of the smallest subspace $\mathcal{S}$ that satisfies (4).

LEMMA 2. *If $\mathcal{S}_1 \subseteq \mathbb{R}^r$ and $\mathcal{S}_2 \subseteq \mathbb{R}^r$ both satisfy* (4)*, then their intersection $\mathcal{S}_1 \cap \mathcal{S}_2$ also satisfies* (4).

The formal definition of the central mean envelope (CME) in the following.

DEFINITION 2. *The central mean envelope of $Y \in \mathbb{R}^r$ on $X \in \mathbb{R}^p$, denoted as $\mathcal{E}_{\mathrm{E}(Y|X)} \subseteq \mathbb{R}^r$, is defined as the intersection of all subspaces $\mathcal{S} \subseteq \mathbb{R}^r$ that satisfy* (4).

By construction, the CME $\mathcal{E}_{\mathrm{E}(Y|X)}$ always exists and is unique. It also generalizes the classical definition of envelopes in multivariate linear models: the definition is a generic definition of two random

vectors and thus model-free; the mean function may no longer have linear form; and the conditional covariance may depend on x.

Moreover, the CME also connects to the parametric form of envelopes in Definition 1.

PROPOSITION 1. *The CME of Y on X reduces $\Sigma(x)$. Moreover, $\mathcal{E}_{\mathrm{E}(Y|X)} = \sum_x \mathcal{E}_{\Sigma(x)}(M_{Y|X})$.*

From the above proposition, the CME can be seen as a subspace that contains $\mathrm{span}(M_{Y|X})$ and also jointly reduces all $\Sigma(x)$ for different values of x. Unlike the standard envelopes in multivariate linear model, the CME does not have to be associated with a parameter subspace such as $\mathrm{span}(\beta)$ in response reduction or $\mathrm{span}(\beta^T)$ in predictor reduction. However, directly estimating the CME as the sum of subspaces $\sum_x \mathcal{E}_{\Sigma(x)}(M_{Y|X})$ is difficult without any additional structural assumptions or simplification. Therefore, we introduce the martingale difference divergence envelope (MDDE) to facilitate the estimation of CME.

3.2. *Martingale difference divergence envelope: a portion of the central mean envelope*

Even when the covariance $\mathrm{cov}(Y \mid X = x) = \Sigma(x)$ is non-constant, it is helpful to first model the mean function $\mathrm{E}(Y \mid X)$ without fully considering $\Sigma(x)$ for all x. We introduce the following definition of martingale difference divergence envelope (MDDE) based on the expectation of conditional covariance $\Sigma = \mathrm{E}\{\mathrm{cov}(Y \mid X)\}$.

DEFINITION 3. *The martingale difference divergence envelope of $Y \in \mathbb{R}^r$ on $X \in \mathbb{R}^p$, denoted as $\mathcal{E}_\Sigma(M_{Y|X})$, is the intersection of all reducing subspaces of $\Sigma = \mathrm{E}\{\mathrm{cov}(Y \mid X)\}$ that contain $\mathrm{span}(M_{Y|X}) = \mathrm{span}[\mathrm{cov}\{\mathrm{E}(Y \mid X)\}]$.*

Because of $\Sigma_Y \equiv \mathrm{cov}(Y) = \mathrm{cov}\{\mathrm{E}(Y \mid X)\} + \mathrm{E}\{\mathrm{cov}(Y \mid X)\} = \mathrm{cov}\{\mathrm{E}(Y \mid X)\} + \Sigma$, we have the following important property of MDDE.

PROPOSITION 2. *The martingale difference divergence envelope $\mathcal{E}_\Sigma(M_{Y|X}) = \mathcal{E}_{\Sigma_Y}(M_{Y|X})$ and is the intersection of all $\mathcal{S} \subseteq \mathbb{R}^r$ such that (i) $\mathrm{E}(Q_\mathcal{S} Y \mid X) = \mathrm{E}(Q_\mathcal{S} Y)$, and (ii) $\mathrm{cov}(Q_\mathcal{S} Y, P_\mathcal{S} Y) = 0$.*

The MDDE is more intuitive from this proposition. Comparing to the CME defined by (4), the only difference is in their second statement: while the CME requires $Q_\mathcal{S} Y$ and $P_\mathcal{S} Y$ to be conditionally uncorrelated given X, the MDDE requires them to be marginally uncorrelated. Then the following proposition establish the more explicit connection between the MDDE and the CME.

PROPOSITION 3. *The MDDE is contained in the CME, $\mathcal{E}_\Sigma(M_{Y|X}) \subseteq \mathcal{E}_{\mathrm{E}(Y|X)}$, where the equality is attained if $\mathrm{cov}(Y \mid X)$ does not depend on X.*

This proposition implies that the MDDE and the CME are identical in the multivariate linear regression setting (1): $\mathcal{E}_\Sigma(\beta) = \mathcal{E}_\Sigma(M_{Y|X}) = \mathcal{E}_{\mathrm{E}(Y|X)}$. Also, we can improve the estimation of the CME by focusing on MDDE first. The MDDE is itself of substantial interests, since it fully captures the potentially nonlinear dependence in mean and also maintains the marginal uncorrelated material and immaterial information. More importantly, unlike the estimation of CME in general, the estimation of MDDE is rather straightforward and does not require slicing or clustering.

3.3. *Coordinate representation and visualization*

To gain more intuition, we use a simulated regression example to visualize the CME and the MDDE. Consider the following regression of multivariate response $Y_i \in \mathbb{R}^r$ on univariate predictor $X_i \in \mathbb{R}^1$,

$$Y_i = m(X_i) + \Sigma(X_i) \cdot \varepsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

where $m(X) = \mathrm{E}(Y \mid X)$, $\Sigma(X) = \mathrm{cov}(Y \mid X)$ and $\varepsilon \sim N(0, I_r)$ is independent of X.

Then the CME, which satisfies (4), is the smallest subspace $\mathcal{S} \subseteq \mathbb{R}^r$ that contains $m(X) \in \mathbb{R}^r$ and reduces $\Sigma(X) \in \mathbb{R}^{r\times r}$. Let $\Gamma \equiv (\gamma_1, \gamma_2, \cdots, \gamma_u) \in \mathbb{R}^{r\times u}$ be a semi-orthogonal basis matrix for the CME and $(\Gamma, \Gamma_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix, then we have the coordinate representation as,

$$m(X) = \Gamma f(X), \quad \Sigma(X) = \Gamma\Omega(X)\Gamma^T + \Gamma_0\Omega_0(X)\Gamma_0^T, \tag{7}$$

where $f(X) \in \mathbb{R}^u$ is the lower-dimensional latent function that reflects the non-linear mean change of Y captured within the CME, and $\Omega(X) \in \mathbb{R}^{u\times u}$ and $\Omega_0(X) \in \mathbb{R}^{(r-u)\times(r-u)}$ are symmetric matrices that reflect the heteroscedastic errors in the CME and those orthogonal to the CME.

We consider $n = 1000$ samples simulated from the above regression model where the response dimension is $r = 20$ and the CME dimension is $u = 2$. Predictor X follows Uniform$(-1, 1)$ distribution and Y is generated as in (6) and (7). We set $f(X) = (\exp(|X|/4), 0)^T \in \mathbb{R}^2$, $\Omega(X) \in \mathbb{R}^{2\times2}$ to be $0.2|X|$ on diagonal and $0.15X$ off diagonal, and Ω_0 to be constant matrix with eigenvalues $\{\exp(-1)/10, \ldots, \exp(-18)/10\}$. Because $\mathrm{E}(Y \mid X) = \exp(|X|/4) \cdot \gamma_1$ is symmetric around the origin, linear regression coefficient $\beta = 0$ and the standard envelope $\mathcal{E}_\Sigma(\beta) = \mathrm{span}(\beta) = \emptyset$ fails to capture anything useful. In this example, the MDDE is $\mathcal{E}_\Sigma(M_{Y|X}) = \mathrm{span}(M_{Y|X}) = \mathrm{span}(\gamma_1)$ because $\mathrm{E}(\Omega(X))$ is a diagonal matrix and $\mathrm{span}(\gamma_1)$ is a reducing subspace of Σ.

In Figure 1, we plot the estimated central mean envelope components $\widehat{\Gamma}^T Y = (\widehat{\gamma}_1^T Y, \widehat{\gamma}_2^T Y)^T \in \mathbb{R}^2$ versus the univariate predictor X, where $\widehat{\gamma}_1$ is also the estimated basis matrix for MDDE. Due to the nonlinearity, the standard envelope component $\widehat{\beta}^T Y$ fails to detect any meaningful information. On the other hand, the first CME component captures the clear nonlinearity, while the second CME component demonstrates the heteroscedasticity. The estimation procedure is introduced in the following.

4. ESTIMATION PROCEDURES

4.1. *Estimating the MDDE*

Recall from Proposition 2 that the MDDE $\mathcal{E}_\Sigma(M_{Y|X}) = \mathcal{E}_{\Sigma_Y}(M_{Y|X})$. Since the marginal covariance $\Sigma_Y = \mathrm{cov}(Y)$ is much easier to estimate than $\Sigma = \mathrm{E}\{\mathrm{cov}(Y \mid X)\}$ in nonlinear regression, the form $\mathcal{E}_{\Sigma_Y}(M_{Y|X})$ is more constructive in estimation. Given the dimension $u_1 = \dim\{\mathcal{E}_{\Sigma_Y}(M_{Y|X})\}$, we propose the following optimization for estimating the MDDE as $\mathrm{span}(\widehat{G})$,

$$\widehat{G} = \arg\min_{G^T G = I_{u_1}} \log|G^T(\widehat{\Sigma}_Y + \widehat{M}_{Y|X})^{-1}G| + \log|G^T\widehat{\Sigma}_Y G|, \tag{8}$$

where $\widehat{M}_{Y|X}$ is the sample MDDM and $\widehat{\Sigma}_Y$ is the sample covariance of Y.

The above objective function can be viewed as the partially optimized pseudo-likelihood of model-free envelope estimation (Zhang & Mai, 2018, Section 3.1). Since $\widehat{M}_{Y|X}$ and $M_{Y|X}$ are both symmetric positive semi-definite matrix, we can write $M_{Y|X} = VV^T$ and $\widehat{M}_{Y|X} = \widehat{V}\widehat{V}^T$ for some $V, \widehat{V} \in \mathbb{R}^{r\times r}$. Then

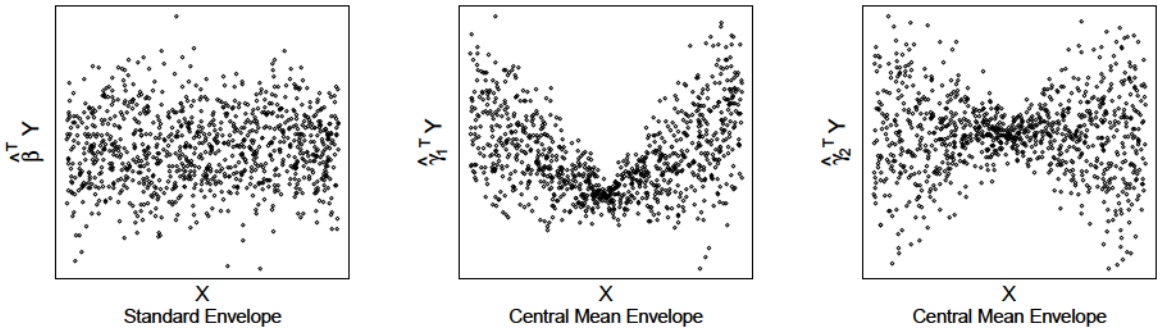

Fig. 1. Estimated envelope components of a multivariate response versus a univaraite predictor from a simulated data. The three plots visualizes standard envelope (left) and the central mean envelope (middle and right) components.

$\text{span}(V) = \text{span}(M_{Y|X})$. The parameter of interest $(\Sigma_Y, M_{Y|X})$ is now reparameterized as (Σ_Y, V). As such, the pseudo-log-likelihood function for joint estimating Σ_Y and V can be written as

$$\ell_n(\Sigma_Y, V) = -\frac{n}{2}\left[\log|\Sigma_Y| + \text{trace}(\Sigma_Y^{-1}\widehat{\Sigma}_Y) + \text{trace}\{(V-\widehat{V})^T\Sigma_Y^{-1}(V-\widehat{V})\}\right], \tag{9}$$

which is to be optimized over the constrained parameter space: $\Sigma_Y = G\Phi G^T + G_0\Phi_0 G_0^T$ and $V = G\eta$ for some $G \in \mathbb{R}^{r\times u_1}$, $\text{span}(G) = \mathcal{E}_{\Sigma_Y}(M_{Y|X}) = \mathcal{E}_{\Sigma_Y}(V)$. By Lemma 3.1 of Zhang & Mai (2018), the unconstrained maximization of (9) leads to the sample covariance $\widehat{\Sigma}_Y$ and sample MDDM $\widehat{M}_{Y|X}$, while the maximum of (9) under envelope constraints is attained at the solution of (8).

An intuitive explaination of (9) is as follows: The first two terms is the negative log-likelihood for Σ_Y under normality, and the last term characterizes the mean function $m(x) = \text{E}(Y \mid X = x) - \text{E}(Y)$. Analogous to the squared Mahalanobis distance of $\{m(x) - \widehat{m}(x)\}^T\Sigma_Y^{-1}\{m(x) - \widehat{m}(x)\}$, we replaced $m(x)$ and $\widehat{m}(x)$ for all values of x with matrices V and $\widehat{V}$. Therefore, the last term in (9) is an overall discrepancy between the constrained $V = G\eta$ and the unconstrained sample estimator $\widehat{V}$.

The proposed objective function in (8) for MDDE and the corresponding pseudo-likelihood in (9) are closely related to the normal likelihood-based estimation in the standard envelope. In Cook et al. (2010), the estimation of $\mathcal{E}_\Sigma(\beta)$ is derived from the conditional normal assumption of $Y \mid X \sim N(\beta X, \Sigma)$. Comparing $\mathcal{E}_\Sigma(\beta) = \mathcal{E}_\Sigma(\beta\Sigma_X\beta^T)$ and $\mathcal{E}_{\Sigma_Y}(M_{Y|X})$, we see the analogy between the fitted value covariance $\beta\Sigma_X\beta^T = \text{cov}\{\text{E}(Y \mid X)\} = \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY} \equiv \Sigma_{\text{fit}}$ and the MDDM $M_{Y|X}$. Moreover, as we replace $\widehat{\Sigma}_Y$ and $\widehat{M}_{Y|X}$ in (8) with the sample least squares estimates $(\widehat{\Sigma}, \widehat{\Sigma}_{\text{fit}})$, the maximum likelihood estimator of $\mathcal{E}_\Sigma(\beta)$ is reproduced by the same optimization. Our optimization (8) and pseudo-likelihood argument (9) fall in the envelope estimation framework of Cook & Zhang (2016) and the envelope dimension inferential framework of Zhang & Mai (2018), respectively. Therefore, we adopt the 1D algorithm (Cook & Zhang, 2016) and 1D dimension selection procedure (Zhang & Mai, 2018).

4.2. *Estimating the CME: some intuitions*

From Propositions 1, $\mathcal{E}_{\mathrm{E}(Y|X)} = \sum_x \mathcal{E}_{\Sigma(x)}(M_{Y|X})$. In estimation, we approximate $\Sigma(x)$ for all values of x with a finite number of covariance matrices: Σ_h, $h = 1, \ldots, H$, $H \geq 2$. Each Σ_h represents the conditional covariance of $\mathrm{cov}\{Y \mid X \in \mathcal{R}_h\}$, where $\mathcal{R}_1, \ldots, \mathcal{R}_H$ is a partition of the support of X. For univariate X, we partition the range of X into H fixed non-overlapping slices similar to the sliced inverse regression procedure (SIR; Li, 1991). For multivariate X, similar to the idea of K-means inverse regression (Setodji & Cook, 2004) that extends SIR from univariate response to multivariate, we construct H clusters of X by the K-means clustering algorithm. As such, we can obtain sample covariance matrices $\widehat{\Sigma}_h$ based on the n_h samples in the h-th slice/cluster, where $\sum_{h=1}^{H} n_h = n$.

If we assume normality and constant mean function within each slice/cluster $\mathcal{R}_h$, then the conditional distribution of $Y \mid X$ is characterized by $Y \mid (X \in \mathcal{R}_h) \sim N(\mu_h, \Sigma_h)$, $h = 1, \ldots, H$. The CME $\mathcal{E}_{\mathrm{E}(Y|X)}$ becomes the smallest subspace that reduces all Σ_h and contains the mean subspace $\mathrm{span}(\mu_1 - \mathrm{E}(X), \ldots, \mu_H - \mathrm{E}(X))$. Similar envelope structure has been studied in groupwise regression (Su & Cook, 2013; Park et al., 2017) as well as in quadratic discriminant analysis (Zhang & Mai, 2019). It can be estimated from the following likelihood-based optimization,

$$\widehat{\Gamma} = \arg\min_{\Gamma^T\Gamma = I_u} \log|\Gamma^T \widehat{\Sigma}_Y^{-1} \Gamma| + \sum_{h=1}^{H} \frac{n_h}{n} \log|\Gamma^T \widehat{\Sigma}_h \Gamma|, \tag{10}$$

which is used in Park et al. (2017) and Zhang & Mai (2019). However, since we can estimate MDDE straightforwardly and accurately, there is no need to approximate $\mathrm{E}(Y \mid X)$ by $\mu_1, \ldots, \mu_H$. As such, we propose a more accurate and practical estimation procedure for the CME in the next section.

Similar to Su & Cook (2013) and Park et al. (2017), in order to derive the likelihood-based estimation (10), we have implicitly assumed that the CME satisfies

$$\Sigma_h = \Gamma \Omega_h \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T, \tag{11}$$

where the immaterial variation $\Gamma_0 \Omega_0 \Gamma_0^T$ is static over h. This is because that we want to fully capture the heteroscedasticity by the CME. Following Proposition 3 of Zhang & Mai (2019), we know that (11) and (10) are in fact targeting at an upper bound of the CME. This targeting subspace, called envelope discriminant subspace, reduces all Σ_h and contains the mean subspace $\mathrm{span}(\mu_1 - \mathrm{E}(X), \ldots, \mu_H - \mathrm{E}(X))$ as well as the inverse covariance subspace $\sum_{h=2}^{H} \mathrm{span}(\Sigma_h^{-1} - \Sigma_1^{-1})$, which is of interest in quadratic discriminant analysis. Without loss of generality, we henceforth assume that CME satisfies (11), or more generally, satisfies $\Sigma(X) = \Gamma \Omega(X) \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T$. When this assumption fails, we are effectively targeting at a bigger subspace without loss of information on heteroscedasticity. To model the heterogeneity among covariance matrices, Cook & Forzani (2008a) and Wang et al. (2019) proposed related subspace models, and Cook & Forzani (2009, Proposition 1) developed properties of sufficient dimension reduction subspaces.

4.3. *Estimating the CME: a novel two-part estimation*

From Proposition 3, we know that the MDDE is always a portion of the CME: $\mathcal{E}_\Sigma(M_{Y|X}) \subseteq \mathcal{E}_{\mathrm{E}(Y|X)}$. Suppose we know the dimensions $u = \dim\{\mathcal{E}_{\mathrm{E}(Y|X)}\}$ and $u_1 = \dim(\mathcal{E}_\Sigma(M_{Y|X}))$, where $u_1 \leq u$.

When $\mathcal{E}_\Sigma(M_{Y|X}) = \mathcal{E}_{\mathrm{E}(Y|X)}$, or equivalently, $u_1 = u$, the estimation procedures in Sections 4.1 and 4.2 are different and generally produce different estimators for the same subspace. Note that even when

$\Sigma(x)$ is not constant, the two subspaces may be the same although they come from different definitions. To address this issue, we propose the two-part estimation approach of first obtaining the MDDE and then obtaining the unique part in CME that is not in MDDE. The two-part estimator of the CME reduces to the same estimator of MDDE in Section 4.1 if $\mathcal{E}_\Sigma(M_{Y|X}) = \mathcal{E}_{\mathrm{E}(Y|X)}$. For $u_1 < u$, we develop the two-part estimation procedure of the CME as follows.

First, we estimate $\mathcal{E}_\Sigma(M_{Y|X})$ as $\mathrm{span}(\widehat{G})$ from (8). Next, we estimate the difference $\mathcal{D} \equiv \mathrm{span}\{v : v \in \mathcal{E}_{\mathrm{E}(Y|X)},\ v \notin \mathcal{E}_\Sigma(M_{Y|X})\}$. Let $(\widehat{G}, \widehat{G}_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix then $\mathcal{D}$ is estimated as $\mathrm{span}(\widehat{G}_0\widehat{H})$, where $\widehat{H} \in \mathbb{R}^{(r-u_1)\times u_2}$ and $u_2 = \dim(\mathcal{D}) = u - u_1$. Specifically, $\widehat{H}$ is estimated as

$$\widehat{H} = \arg\min_{H^TH=I_{u_2}} \log|H^T(\widehat{G}_0^T\widehat{\Sigma}^{-1}\widehat{G}_0)H| + \sum_{h=1}^{H}\frac{n_h}{n}\log|H^T(\widehat{G}_0^T\widehat{\Sigma}_h\widehat{G}_0)H|, \tag{12}$$

which is inspired by the following Lemma. Finally, the two-part estimation of $\mathcal{E}_{\mathrm{E}(Y|X)}$ is $\mathrm{span}(\widehat{G}, \widehat{G}_0\widehat{H})$.

LEMMA 3. *Let $\Sigma_h \in \mathbb{R}^{r\times r}$, $h = 1, \ldots, H$, $H \geq 2$, be a series of symmetric positive definite matrices, and let $\Sigma = \sum_{h=1}^H \pi_h\Sigma_h$, where $\pi_h > 0$ and $\sum_{h=1}^H \pi_h = 1$. Assume $\mathcal{E} \equiv \mathrm{span}(\Gamma) \subseteq \mathbb{R}^r$ is the smallest subspace such that $\Sigma_h = \Gamma\Omega_h\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$ holds for all h for some Ω_h and Ω_0, and that $\dim(\mathcal{E}) = u$, then $\mathcal{E} = \mathrm{span}(\widehat{G})$, where $\widehat{G}$ is defined as follows,*

$$\widehat{G} = \arg\min_{G\in\mathbb{R}^{r\times u},\ G^TG=I_u} \{\log|G^T\Sigma^{-1}G| + \sum_{h=1}^{H}\pi_h\log|G^T\Sigma_hG|\}. \tag{13}$$

The above Lemma suggests a more effective objective function than the objective function in (10). Specifically, we have made a simple but important change in (10) to get (13). We replaced the marginal covariance Σ_Y by the conditional covariance Σ. Lemma 3 implies that both objective functions estimate the CME consistently, however, the motivations are different. As mentioned earlier, (10) is motivated from the likelihood for estimating the CME, while the new objective function (13) focuses more on the heterogeneity of covariance matrices Σ_h since Σ is the weighted average of all Σ_h's. This is indeed much more desirable in the two-part estimation because the first part, the MDDE, already contains the conditional mean function. In practice, we have observed that the two-part estimation based on (13) is more accurate than that based on (10).

In our experience, the two-part estimation procedure of the CME is almost always better than the direct estimation from (10) or (13). This is due to the fact that the MDDE part is easier to estimate, as the sample matrix $\widehat{M}_{Y|X}$ in (8) is more accurately estimated than $\widehat{\Sigma}_h$ in (10) and (13), especially when the sample size n_h is small for some slices/clusters. Moreover, the optimization in (8) is more feasible than the optimization in (10) and (13). As we have mentioned earlier, the objective function in (8) can be solved using more specialized envelope algorithms (e.g., Cook & Zhang, 2016, 2018) that are much faster and more accurate than using standard optimization methods with orthogonality constraints (e.g., Absil et al., 2009; Wen & Yin, 2013). As such, we use the two-part estimation of the CME in all our numerical studies.

4.4. *Consistency*

We establish the $\sqrt{n}$-consistency of our estimator $\mathrm{span}(\widehat{G})$ from (8) for the MDDE and the two-part estimator $\mathrm{span}(\widehat{G}, \widehat{G}_0\widehat{H})$ from (8) and (12) for the CME. Since the subspaces are uniquely defined by the projection matrices onto them, the asymptotic results are stated in terms of projection matrices.

For all the asymptotic results, we require no model or distributional assumption. Instead, we assume $\widehat{\Sigma}_Y$, $\widehat{\Sigma}_h$ and $\widehat{M}_{Y|X}$ are $\sqrt{n}$-consistent estimators for their population counterparts $\Sigma_Y > 0$, $\Sigma_h > 0$, $h = 1, \ldots, H$, and $M_{Y|X} \geq 0$. This is a mild assumption that can be easily satisfied: the sample covariance matrices $\widehat{\Sigma}_Y$ and $\widehat{\Sigma}_h$ are $\sqrt{n}$-consistent when Y and $Y \mid X$ have finite fourth moments; $\sqrt{n}$-consistency of the sample MDDM $\widehat{M}_{Y|X}$ is established in Lee & Shao (2018). The consistency of the MDDE estimation is obtained by applying Proposition 3 in Cook & Zhang (2016).

PROPOSITION 4. *Let $\widehat{G} \in \mathbb{R}^{r\times u_1}$ be any minimizer of* (8)*, then $P_{\widehat{G}}$ is $\sqrt{n}$-consistent for the projection onto $\mathcal{E}_\Sigma(M_{Y|X})$.*

We can speed up the computation by adopting the 1D algorithm (Cook & Zhang, 2016) that sequentially estimates one direction at a time for the MDDE. The resulting estimator, $\widehat{G}^{\mathrm{1D}} \in \mathbb{R}^{r\times u_1}$, is no longer an minimizer of (8) but the $\sqrt{n}$-consistency in Proposition 4 still holds if we replace $\widehat{G}$ with $\widehat{G}^{\mathrm{1D}}$ and apply the Proposition 6 from Cook & Zhang (2016). Moreover, the 1D algorithm is also coupled with a model-free envelope selection criterion (Zhang & Mai, 2018) that estimates the envelope dimension consistently. That is, $\Pr(\widehat{u}_1 = u_1) \to 1$ as $n \to \infty$. We will demonstrate this dimension selection procedure in our numerical studies.

For the second part estimation of the CME, we need the following coverage condition on the slicing scheme. Specifically, the central mean envelope is covered as,

$$\sum_x \mathcal{E}_{\Sigma(x)}(M_{Y|X}) = \sum_{h=1}^{H} \mathcal{E}_{\Sigma_h}(M_{Y|X}). \tag{14}$$

Let $G \in \mathbb{R}^{r\times u_1}$ be a basis matrix for $\mathcal{E}_\Sigma(M_{Y|X})$, and $G_0 \in \mathbb{R}^{r\times(r-u_1)}$ be its orthogonal completion. We show the consistency of the first part $\mathcal{E}_\Sigma(M_{Y|X})$ from (8), and the second part $\mathcal{D}$ from (12).

PROPOSITION 5. *Let $\widehat{G} \in \mathbb{R}^{r\times u_1}$ be any minimizer of* (8) *and $\widehat{H} \in \mathbb{R}^{(r-u_1)\times u_2}$ be any minimizer of* (12)*. If* (14) *is true, then $P_{\widehat{G}_0\widehat{H}}$ is $\sqrt{n}$-consistent for the projection onto $\mathcal{D}$. Moreover, the projection onto $\mathrm{span}(\widehat{G}, \widehat{G}_0\widehat{H})$ is $\sqrt{n}$-consistent for the projection onto $\mathcal{E}_{\mathrm{E}(Y|X)}$.*

The assumption of (14) is a very mild condition and is easily satisfied when $\Sigma(x)$ is a smooth function of x. For example, this condition holds for any slicing scheme with $h \geq 2$ in the example of Section 3.3. In practice, one can combine different slicing/clustering schemes to achieve a more robust and accurate estimator. See Cook & Zhang (2014) for more background on the effect of slicing and constructing fused estimator from different slicing schemes. The fused estimator in Cook & Zhang (2014) can be directly applied to our optimization in (12), and in a way also circumvents the issue of choosing optimal slicing scheme and weakens the assumption (14).

5. SIMULATIONS

5.1. *Comparison*

In this section, we study the finite-sample performance of our two-part estimation of the CME. We compare with two closely related methods: the standard envelope estimator in Cook et al. (2010) and the dimension reduction method by decomposing the MDDM directly (Lee & Shao, 2018). In presence of nonlinear mean function, the standard envelope model based on (1) and $\mathcal{E}_\Sigma(\beta)$ is mis-specified. However, since $\mathcal{E}_\Sigma(\beta) = \mathcal{E}_\Sigma(\Sigma_{YX})$, the standard envelope is still a well-defined subspace in all of our simulation examples but ineffective in detecting nonlinear and heteroscedastic envelope components. The MDDM method performs eigen-decomposition of $\widehat{M}_{Y|X}$ and targets directly at the mean subspace $\mathrm{span}(M_{Y|X}) = \mathrm{span}\{\mathrm{E}(Y \mid X = x) - \mathrm{E}(Y) : \forall x\}$. In comparison, the proposed two-part estimator, specifically the MDDE part from (8), has more advantage when Y are highly correlated.

In addition, we also compare with three popular sufficient dimension reduction methods by interchanging the role of X and Y: the sliced inverse regression (SIR; Li, 1991), the sliced average variance estimation (SAVE; Cook & Weisberg, 1991), and the principal fitted components (PFC; Cook & Forzani, 2008b). When studying the regression of a univariate response $Y \in \mathbb{R}^1$ on a multivariate predictor $X \in \mathbb{R}^p$, these sufficient dimension reduction methods are aiming at the central subspace $\mathcal{S}_{Y|X} \subseteq \mathbb{R}^p$. Let $\beta \in \mathbb{R}^{p\times d}$ be some basis matrix of the central subspace $\mathcal{S}_{Y|X}$, then $Y \mid X \sim Y \mid \beta^T X$. The central subspace are then estimated as $\Sigma_X^{-1}\mathrm{span}(M_{X|Y}^{\mathrm{SIR}})$, $\Sigma_X^{-1}\mathrm{span}(M_{X|Y}^{\mathrm{SAVE}})$, and $\Sigma_X^{-1}\mathrm{span}(M_{X|Y}^{\mathrm{PFC}})$, for some $p \times p$ symmetric positive semi-definite matrices $M_{X|Y}^{\mathrm{SIR}}$, $M_{X|Y}^{\mathrm{SAVE}}$ and $M_{X|Y}^{\mathrm{PFC}}$. By removing the Σ_X^{-1} term in these dimension reduction methods (e.g. removing the standardization step in estimation), and interchanging the roles of X and Y, the targeting subspace of these methods is a subset of the CME. Therefore, to be fairly compared with envelope methods, we estimate $\mathrm{span}(M_{X|Y}^{\mathrm{SIR}})$, $\mathrm{span}(M_{X|Y}^{\mathrm{SAVE}})$ and $\mathrm{span}(M_{X|Y}^{\mathrm{PFC}})$ from these sufficient dimension reduction methods.

In the simulation studies, we set the dimension $r = 15$ and the sample size $n = 200$ or 600. For each example, we replicate the simulation 100 times and compute the Frobenius norm of the difference between the projection matrices onto the true and estimated subspaces, i.e. $\|P_\Gamma - P_{\widehat{\Gamma}}\|_F$. We also compute the principal angles between each CME directions and the estimated subspace, i.e. $\theta_j = \cos^{-1}\{(\gamma_j^T P_{\widehat{\Gamma}}\gamma_j)^{1/2}\}$, $j = 1, \ldots, u$, where $\Gamma = (\gamma_1, \ldots, \gamma_u)$. The values for θ_j are bounded between 0 and 90, where $\theta_j = 0$ indicates that γ_j is contained in $\mathrm{span}(\widehat{\Gamma})$ and $\theta_j = 90$ indicates that γ_j is orthogonal to $\mathrm{span}(\widehat{\Gamma})$. For SIR, SAVE and the second part estimation of CME, the number of slices $H = 5$; and for PFC, the function basis is the cubic polynomials $(x, x^2, x^3)^T$.

5.2. *The standard envelope coincides with the CME*

We first consider the following four models where the standard envelope, the MDDE, and the CME are all the same in population. Therefore, the two-part CME estimator is identical to the MDDE estimator based on (8). Similar to the data generating process in Section 3.3, the CME basis is randomly generated and orthogonalized: $\Gamma = (\gamma_1, \ldots, \gamma_u) \in \mathbb{R}^{r\times u}$; and the predictor X follows Uniform$(-1, 1)$ distribution unless otherwise specified. The errors are generated as $E_i = 0.1 \cdot \Sigma^{1/2}(X_i) \cdot \varepsilon_i$, where ε_i's are i.i.d. $N(0, I_r)$ and $\Sigma(x) = \Gamma\Omega(x)\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$ is specified as follows. We use $O_k \in \mathbb{R}^{k\times k}$ to denote an arbitrarily generated orthogonal matrix.

Models	n	SIR	SAVE	PFC	MDDM	Standard	MDDE
I	200	0.07 (0.02)	0.07 (0.3)	0.07 (0.2)	0.07 (0.2)	0.04 (0.2)	0.05 (0.2)
	600	0.04 (0.2)	0.04 (0.1)	0.04 (0.1)	0.04 (0.1)	0.03 (0.1)	0.03 (0.1)
II	200	0.07 (0.3)	0.25 (1.7)	0.07 (0.3)	0.07 (0.3)	0.58 (6.7)	0.05 (0.2)
	600	0.04 (0.1)	0.15 (0.7)	0.04 (0.2)	0.04 (0.2)	0.53 (6.7)	0.03 (0.1)
III	200	0.57 (2.6)	0.21 (0.6)	0.75 (3.6)	0.41 (1.8)	0.15 (2.6)	0.11 (1.4)
	600	0.56 (2.7)	0.13 (0.4)	0.73 (3.4)	0.41 (2.3)	0.06 (0.2)	0.06 (0.2)
IV	200	0.12 (0.4)	0.14 (0.5)	0.12 (0.4)	0.12 (0.4)	1.08 (5.9)	0.06 (0.2)
	600	0.07 (0.2)	0.09 (0.3)	0.07 (0.2)	0.07 (0.2)	0.88 (6.8)	0.04 (0.1)

Table 1. *Estimation error $\|P_\Gamma - P_{\widehat{\Gamma}}\|_F$ averaged over 100 replicates, where the standard errors are also included in the paratheses after multiplied by* 100*. The SIR, SAVE, PFC, MDDM, Standard, and MDDE refer to the methods in Li (1991), Cook & Weisberg (1991), Cook & Forzani (2008b), Lee & Shao (2018), Cook et al. (2010), and the proposed estimator from* (8)*, respectively.*

- Model (I). Linear mean, constant covariance: $Y = \gamma_1 X + E_i$, $\Gamma = \gamma_1 \in \mathbb{R}^{r\times 1}$, $\Omega_0 = O_{r-1}\mathrm{diag}\{\exp(3.0), \exp(2.5), \ldots, \exp(-3.5)\}O_{r-1}^T$ and $\Omega = \exp(5)$.
- Model (II). Nonlinear mean, constant covariance: $Y = \gamma_1 X + \gamma_2 \exp(2|X|) + E_i$, $\Gamma = (\gamma_1, \gamma_2) \in \mathbb{R}^{r\times 2}$, $\Omega_0 = O_{r-2}\mathrm{diag}\{\exp(3.0), \exp(2.5), \ldots, \exp(-3.0)\}O_{r-2}^T$ and $[\Omega]_{ij} = 0.5^{|i-j|} \cdot \exp(5)$.
- Model (III). Linear mean, nonconstant covariance: $Y = \gamma_1 X + 0.1 \cdot \Sigma^{1/2}(X_i) \cdot \varepsilon_i$, where Γ and $\Sigma = \mathrm{E}\{\Sigma(X)\}$ are the same as in Model (II) but the off-diagonal of $\Omega(x)$ is $\Omega_{12}(x) = |x| \cdot \exp(5)$.
- Model (IV). Nonlinear mean, nonconstant covariance. The nonlinear mean function is the same as in Model (II); and the nonconstant covariance is the same as in Model (III).

In all the simulations, the standard envelope becomes $\mathcal{E}_\Sigma(\Sigma_{YX}) = \mathcal{E}_\Sigma(\gamma_1)$ because the second direction γ_2 either appears only in the covariance or appears with a nonlinear mean function $\gamma_2 \exp(2|X|)$ that is uncorrelated with $X \sim$Uniform$(-1, 1)$.

We summarize the results in Table 1, where we compare each methods in terms of overall estimation error $\|P_\Gamma - P_{\widehat{\Gamma}}\|_F$. The MDDE estimator based on (8) is always the best except for Model (I), where the standard envelope is the maximum likelihood estimator and has a slight advantage. In constant covariance scenarios, i.e. Models (I) and (II), the MDDE has similar performance as the better one of the standard envelope and the MDDM estimators. This is because that the MDDE estimation is essentially a hybrid of the envelope estimation (that incorporates covariance structural information of Σ) and the MDDM (that is effective in capturing nonlinear means). In nonconstant covariance scenarios, i.e. Models (III) and (IV), the MDDE has better performance than both the standard envelope and the MDDM.

For Models with two-dimensional subspace, we further compute the principal angles between the second CME direction γ_2 and the estimated subspace, i.e. $\theta_2 = \cos^{-1}\{(\gamma_2^T P_{\widehat{\Gamma}}\gamma_2)^{1/2}\}$. For Model (III), the second direction is only estimable from the nonconstant covariance but does not appear in the mean function. Therefore, SIR, PFC and MDDM fail to estimate γ_2 accurately, while SAVE, standard envelope, and the proposed estimator perform well. For Model (II) and (IV), the second direction appears in both the (nonlinear) mean and the covariance. Therefore, the nonlinear methods, SIR, SAVE, PFC, MDDM, and MDDE all work well, while the standard envelope method becomes ineffective. Overall, the MDDE estimator is the most reliable one.

Models	n	SIR	SAVE	PFC	MDDM	Standard	MDDE
II	200	0.21 (0.01)	0.11 (0.01)	0.10 (0.01)	0.12 (0.01)	35.10 (4.40)	0.09 (0.01)
	600	0.11 (0.01)	0.05 (0.01)	0.05 (0.01)	0.06 (0.01)	32.39 (4.34)	0.04 (0.01)
III	200	23.38 (1.27)	6.82 (0.26)	33.89 (2.13)	15.81 (0.84)	4.69 (0.86)	3.43(0.44)
	600	23.87 (1.30)	4.35 (0.15)	33.26 (1.91)	16.77 (1.11)	1.69 (0.06)	1.69 (0.06)
IV	200	1.65 (0.06)	1.59 (0.05)	1.49 (0.05)	1.52 (0.05)	59.02 (3.36)	1.35 (0.05)
	600	1.04 (0.04)	0.99 (0.03)	0.92 (0.03)	0.95 (0.04)	47.61 (3.76)	0.79 (0.03)

Table 2. *Estimation error of the second direction* $\theta_2 = \cos^{-1}\{(\gamma_2^T P_{\widehat{\Gamma}}\gamma_2)^{1/2}\}$

Models	n	SIR	SAVE	PFC	MDDM	Standard	CME
V	200	57.46 (2.7)	27.41 (1.0)	74.17 (3.5)	39.66 (1.7)	141.50 (0.1)	8.44 (0.3)
	600	55.74 (2.8)	24.72 (1.1)	73.56 (3.4)	39.68 (2.0)	141.45 (0.1)	4.79 (0.1)
VI	200	58.26 (2.9)	25.33 (1.1)	71.15 (3.5)	65.67 (3.5)	182.27 (3.2)	7.27 (0.2)
	600	54.77 (2.9)	23.81 (0.9)	67.26 (3.3)	60.92 (3.1)	175.19 (3.4)	4.13 (0.1)
VII	200	65.29 (2.9)	34.66 (1.3)	61.05 (3.2)	21.04 (0.9)	181.66 (3.2)	9.26 (1.9)
	600	61.05 (2.87)	34.39 (1.5)	62.32 (3.0)	20.05 (0.8)	183.01 (2.7)	3.98 (0.1)

Table 3. *Estimation error* $\|P_\Gamma - P_{\widehat{\Gamma}}\|_F$ *averaged over 100 replicates, where the standard errors are also included in the paratheses after multiplied by* 100. *The SIR, SAVE, PFC, MDDM, Standard, and CME refer to the methods in Li (1991), Cook & Weisberg (1991), Cook & Forzani (2008b), Lee & Shao (2018), Cook et al. (2010), and the proposed two-part estimation, respectively.*

5.3. *The standard envelope and the MDDE are proper subsets of the CME*

In this section, we consider the following three models where $\mathcal{E}_\Sigma(\Sigma_{YX}) = \mathcal{E}_\Sigma(\gamma_1) \subseteq \mathcal{E}_\Sigma(M_{Y|X}) \subset \mathcal{E}_{\mathrm{E}(Y|X)} = \mathrm{span}(\Gamma)$. The CME estimator is now based on two-part estimation in Section 4.2.

- Model (V). Linear mean, nonconstant covariance: same as Model (III) except that we change the off-diagonal of $\Omega(X)$ is from $|X| \cdot \exp(5)$ to X. Thus $\mathcal{E}_\Sigma(\gamma_1) = \mathcal{E}_\Sigma(M_{Y|X}) = \mathrm{span}(\gamma_1) \subset \mathcal{E}_{\mathrm{E}(Y|X)}$.
- Model (VI). Nonlinear mean, nonconstant covariance. Same as Model (V), except that the mean function is $\gamma_2 \cdot \exp(2|X|)$ and is linearly independent of X. Thus $\mathcal{E}_\Sigma(\gamma_1) \subset \mathcal{E}_\Sigma(M_{Y|X}) \subset \mathcal{E}_{\mathrm{E}(Y|X)}$.
- Model (VII). Nonlinear mean, nonconstant covariance and multivariate predictor. Similar to Model (IV), we have nonlinear mean function $\gamma_2 \cdot \{\exp(2|X_1|) + \exp(|X_2| + |X_3|) + \exp(|X_4| + |X_5|)\}$, where the multivariate predictor is five-dimensional $X = (X_1, \ldots, X_5)^T$ and each coordinate is uniformly distributed from -1 to 1. The covariance structure is the same as the previous two models except for $\Omega_{12}(X) = X_4 + X_5$. We have $\mathcal{E}_\Sigma(\gamma_1) \subset \mathcal{E}_\Sigma(M_{Y|X}) \subset \mathcal{E}_{\mathrm{E}(Y|X)}$.

Similar to Section 5.2, we summarize the overall estimation error $\|P_\Gamma - P_{\widehat{\Gamma}}\|_F$ in Table 3, and summarize second CME direction γ_2 estimation error $\theta_2 = \cos^{-1}\{(\gamma_2^T P_{\widehat{\Gamma}}\gamma_2)^{1/2}\}$ in Table 4. For these more challenging models, the proposed CME estimator is much more accurate than the competitors, especially in estimating the second direction that comes from nonlinearity and heteroscedastic error.

Moreover, we tried different numbers of slices/clusters, $H = 2, 5, 10, 15$, for Models (V)–(VII). The results are indistinguishable from the results in Tables 3 and 4, where we have used $H = 5$. The overall performance of the two-parts estimation is very encouraging.

Models	n	SIR	SAVE	PFC	MDDM	Standard	CME
V	200	23.93 (1.39)	9.52 (0.43)	33.33 (2.05)	15.34 (0.77)	82.91 (0.58)	2.58 (0.10)
	600	23.83 (1.42)	9.38 (0.48)	33.79 (2.04)	16.25 (0.92)	86.33 (0.31)	1.44 (0.05)
VI	200	25.32 (1.50)	10.17 (0.46)	32.50 (1.98)	29.74 (1.95)	76.35 (3.07)	2.59 (0.10)
	600	23.88 (1.52)	9.66 (0.39)	30.57 (1.90)	27.15 (1.74)	79.58 (2.74)	1.44 (0.05)
VII	200	25.72 (1.50)	13.01 (0.60)	23.80 (1.59)	8.38 (0.36)	75.31 (3.16)	4.06 (1.16)
	600	24.17 (1.31)	12.92 (0.65)	25.39 (1.56)	8.11 (0.32)	80.64 (2.65)	1.46 (0.05)

Table 4. *Estimation error of the second direction* $\theta_2 = \cos^{-1}\{(\gamma_2^T P_{\widehat{\Gamma}}\gamma_2)^{1/2}\}$.

Models	Standard Envelope						CME					
	$\widehat{u} < u$		$\widehat{u} = u$		$\widehat{u} > u$		$\widehat{u} < u$		$\widehat{u} = u$		$\widehat{u} > u$	
n	200	600	200	600	200	600	200	600	200	600	200	600
II	54	55	44	45	2	0	0	0	100	100	0	0
III	8	0	84	98	8	2	35	0	65	100	0	0
IV	75	77	23	22	2	1	0	0	100	100	0	0
V	86	98	13	2	1	0	0	0	56	50	44	20
VI	94	99	6	1	0	0	0	0	53	82	47	18

Table 5. *Percentages of correctly selected dimension, under-selection, and over-selection over 100 replicates for each model setting based on the dimension selection procedure in Zhang & Mai (2018).*

5.4. *Selecting the envelope dimensions*

We applied the model-free information criteria of envelope dimension selection proposed by Zhang & Mai (2018) to select the envelope dimensions (i.e. for standard envelope, MDDE, and CME). Implementation details is included in Section S2 of the Supplementary Materials. Table 5 summarizes the dimension selection results for the CME under the previous simulation models. For Models (II)–(IV), because the standard envelope coincides with the CME, both methods should be able to determine the true dimension consistently when $n \to \infty$. Clearly, the CME method has much better finite sample performance. For Models (V) and (VI), the standard envelope fails to detect the second CME components, and, not surprisingly under-selected the dimension. The CME dimension selection has reasonable results for these two very challenging models. There is also a significant improvement of accuracy when the sample size is increased from 200 to 600.

6. REAL DATA ILLUSTRATION

We demonstrate the advantages of the central mean envelope over standard envelope in prediction and subspace estimation. The riboflavin data set from Bühlmann et al. (2014) contains 71 samples of the riboflavin production rate and the expression level of 4,088 genes. We use the logarithm of the riboflavin production rate as the predictor and the logarithm of the gene expression as the response. Similar to Bühlmann et al. (2014), and due to the high dimensionality, we focus the top 50 genes that selected by the martingale difference correlation Shao & Zhang (2014). We applied the same dimension selection procedures as in Section 5.4 for the standard envelope and the CME. For the standard envelope, the selected dimension is $u = 1$. For the CME, the selected dimensions are $u_1 = 1$ for the first part (nonlinear mean) and $u_2 = 0$ for the second part (heteroscedastic error). In the Supplementary Materials,

	Prediction		Estimation	
	Linear	Kernel	$m = 71$	$m = 150$
Standard envelope	37.57 (0.50)	37.22 (0.50)	0.37 (0.014)	0.34 (0.021)
Proposed envelope	35.83 (0.38)	35.30 (0.33)	0.25 (0.007)	0.17 (0.004)

Table 6. *Comparison of envelope methods in prediction (prediction mean squared error) and in estimation (subspace bootstrap variability). The standard envelope and proposed envelope refer to the estimators from Cook et al. (2010) and from optimization* (8)*, respectively.*

the nonlinearity in $\mathrm{E}(\widehat{\gamma}^T Y \mid X)$ is clearly demonstrated in Figure S1, where $\widehat{\gamma}$ is the estimated CME, which has dimension one and is the same as the MDDE.

For numerical comparison, we divide the data into training and testing sets by randomly choose 56 samples as training data and the remaining 15 as testing data. Then we compute the prediction mean squared error based on either linear or kernel regression of the material part of response on predictor, $\widehat{\mathrm{E}}(\widehat{\gamma}^T Y \mid X)$, while the immaterial part of the response is predicted by its unconditional mean $\widehat{\mathrm{E}}(\widehat{\Gamma}_0^T Y)$. Specifically, for the kernel regression, we use Gaussian kernel with the optimal bandwidth from the "ksr" function in Matlab. We repeat this procedure for 100 times and present the average and standard deviation of the prediction mean squared error in Table 6. The improvement is significant. To evaluate the subspace estimation accuracy, we consider the bootstrap variability of subspaces $B^{-1}\sum_{b=1}^{B}\|P_{\widehat{\mathcal{S}}} - P_{\widehat{\mathcal{S}}^b}\|_F$, where $\widehat{\mathcal{S}}$ is the estimated subspace on the original data and $\widehat{\mathcal{S}}^b$, $b = 1, \ldots, B$, is the estimated subspace on m bootstrap samples from $B = 200$ bootstrap replicates. This is a commonly used criterion in sufficient dimension reduction literature (for example, Ye & Weiss, 2003; Luo & Li, 2016). We consider $m = 71$ bootstrap samples, then the covariance may occasionally become ill-conditioned and we hence add $0.01 I_r$ to the sample covariance $\widehat{\Sigma}_Y$. Alternatively, we also consider $m = 150$. From Table 6, the proposed method has improved the subspace estimation significantly.

7. Discussion

Although our focus is on multivariate response reduction in regression, the two new envelope structures, MDDE and CME, are model-free and can be used beyond regression models. For example, in the Supplementary Materials Section S4, we also apply our method on a handwritten digit recognition data and illustrate the CME as a useful data visualization tool in discriminant analysis and classification. Moreover, by interchanging the roles of X and Y, the MDDE $\mathcal{E}_{\Sigma_X}(M_{X|Y})$ or the CME $\mathcal{E}_{\mathrm{E}(X|Y)}$ can serve as an upper bound of the central subspace $\mathcal{S}_{Y|X}$ in sufficient dimension reduction (Cook, 1998; Li, 2018) and potentially improve standard sufficient dimension reduction methods. Two future research directions are to extend our framework to simultaneous predictor and response reduction, and to stationary multivariate time series. The former would be an extension of the simultaneous envelope in multivariate linear model (Cook & Zhang, 2015b), while the latter can be achieved by using the cumulative version of $\widehat{M}_{Y|X}$ (Lee & Shao, 2018). Properties of such extensions are yet to be studied.

Acknowledgement

The authors are grateful to the Editor, Associate Editor and two referees for insightful comments that have led to significant improvements of this paper; and would like to thank Professor R. Dennis Cook

from the University of Minnesota for his comments and suggestions on the manuscript. Research for this paper was partly supported by U.S. National Science Foundation.

REFERENCES

ABSIL, P.-A., MAHONY, R. & SEPULCHRE, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

BÜHLMANN, P., KALISCH, M. & MEIER, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application* **1**, 255–U809.

COOK, R., HELLAND, I. & SU, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 851–877.

COOK, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, vol. 318. John Wiley & Sons.

COOK, R. D. (2018). Principal components, sufficient dimension reduction, and envelopes. *Annual Review of Statistics and Its Application* **5**, 533–559.

COOK, R. D. & FORZANI, L. (2008a). Covariance reducing models: An alternative to spectral modelling of covariance matrices. *Biometrika* **95**, 799–812.

COOK, R. D. & FORZANI, L. (2008b). Principal fitted components for dimension reduction in regression. *Statistical Science* **23**, 485–501.

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

COOK, R. D. & LI, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics* **30**, 455–474.

COOK, R. D., LI, B. & CHIAROMONTE, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika* **94**, 569–584.

COOK, R. D., LI, B. & CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica* , 927–960.

COOK, R. D. & WEISBERG, S. (1991). Comments on "sliced inverse regression for dimension reduction" by kc li. *Journal of the American Statistical Association* **86**, 328–332.

COOK, R. D. & ZHANG, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *Journal of the American Statistical Association* **109**, 815–827.

COOK, R. D. & ZHANG, X. (2015a). Foundations for envelope models and methods. *Journal of the American Statistical Association* **110**, 599–611.

COOK, R. D. & ZHANG, X. (2015b). Simultaneous envelopes for multivariate linear regression. *Technometrics* **57**, 11–25.

COOK, R. D. & ZHANG, X. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics* **25**, 284–300.

COOK, R. D. & ZHANG, X. (2018). Fast envelope algorithms. *Statistica Sinica* **28**, 1179–1197.

DAWID, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)* , 1–31.

KHARE, K., PAL, S. & SU, Z. (2017). A bayesian approach for envelope models. *The Annals of Statistics* **45**, 196–222.

LEE, C. E. & SHAO, X. (2018). Martingale difference divergence matrix and its application to dimension reduction for stationary multivariate time series. *Journal of the American Statistical Association* **113**, 216–229.

LI, B. (2018). *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC.

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.

LI, L. & ZHANG, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association* **112**, 1131–1146.

LUO, W. & LI, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **103**, 875–887.

MATTESON, D. S. & TSAY, R. S. (2017). Independent component analysis via distance covariance. *Journal of the American Statistical Association* **112**, 623–637.

PARK, T., SHAO, X. & YAO, S. (2015). Partial martingale difference correlation. *Electronic Journal of Statistics* **9**, 1492–1517.

PARK, Y., SU, Z. & ZHU, H. (2017). Groupwise envelope models for imaging genetic analysis. *Biometrics* **73**, 1243–1253.

SETODJI, C. M. & COOK, R. D. (2004). K-means inverse regression. *Technometrics* **46**, 421–429.

SHAO, X. & ZHANG, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association* **109**, 1302–1318.

SHENG, W. & YIN, X. (2016). Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics* **25**, 91–104.

SU, Z. & COOK, R. D. (2013). Estimation of multivariate means with heteroscedastic errors using envelope models. *Statistica Sinica* , 213–230.

SZÉKELY, G. J. & RIZZO, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics* , 1236–1265.

SZÉKELY, G. J., RIZZO, M. L., BAKIROV, N. K. et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

VEPAKOMMA, P., TONDE, C., ELGAMMAL, A. et al. (2018). Supervised dimensionality reduction via distance correlation maximization. *Electronic Journal of Statistics* **12**, 960–984.

WANG, W., ZHANG, X. & LI, L. (2019). Common reducing subspace model and network alternation analysis. *Biometrics* **75**, 1109–1120.

WEN, Z. & YIN, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming* **142**, 397–434.

YE, Z. & WEISS, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association* **98**, 968–979.

ZHANG, X. & LI, L. (2017). Tensor envelope partial least-squares regression. *Technometrics* **59**, 426–436.

ZHANG, X. & MAI, Q. (2018). Model-free envelope dimension selection. *Electronic Journal of Statistics* **12**, 2193–2216.

ZHANG, X. & MAI, Q. (2019). Efficient integration of sufficient dimension reduction and prediction in discriminant analysis. *Technometrics* **61**.