Biometrika (2020), **00**, 0, *pp.* 1–17
Printed in Great Britain

doi: 10.1093/biomet/asaa036

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 and the central mean envelope, and study their relationships to the standard envelope in the context of response reduction in multivariate linear models. The martingale difference divergence envelope effectively captures the nonlinearity in the conditional mean without imposing any parametric structure or requiring any tuning in estimation. Heteroscedasticity, or nonconstant conditional covariance of $Y \mid X$, is further detected by the central mean envelope based on a slicing scheme for the data. We reveal the nested structure of different envelopes: (i) the central mean envelope contains the martingale difference divergence envelope, with equality when $Y \mid X$ has a constant conditional covariance; and (ii) the martingale difference divergence envelope contains the standard envelope, with equality when $Y \mid X$ has a linear conditional mean. We develop an estimation procedure that first obtains the martingale difference divergence envelope and then estimates the additional envelope components in the central mean envelope. We establish consistency in envelope estimation of the martingale difference divergence envelope and central mean envelope without stringent model assumptions. Simulations and real-data analysis demonstrate the advantages of the martingale difference divergence envelope and the central mean envelope over the standard envelope in dimension reduction.

Some key words: Envelope model; 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 a multivariate linear model of a multivariate response $Y \in \mathbb{R}^r$ on a predictor $X \in \mathbb{R}^p$,

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

where $\alpha \in \mathbb{R}^r$, $\beta \in \mathbb{R}^{r\times p}$, and $\varepsilon_i \sim N(0, \Sigma)$ with $\Sigma > 0$ independent of X_i. Here, $(X_i, Y_i), i = 1, \ldots, n$, are independent and identically distributed. 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 purpose of estimating β. Envelope models and methods have recently been developed for 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) $E(Y \mid X = x) = \alpha + \beta x$ is linear in x, and (ii) $\mathrm{cov}(Y \mid X = x) = \Sigma$ does not depend on x. Most existing envelope methods for multivariate linear models rely on these two assumptions (e.g., Cook et al., 2010, 2013; Cook & Zhang, 2015b; Khare et al., 2017). As a result, the performance of these methods may suffer severely from any violation of the assumptions. While most existing envelope methods concentrate on parametric settings, particularly for multivariate linear models, we aim to address nonlinearity and heteroscedasticity by revisiting model-free envelopes in a more flexible dimension reduction setting similar to that of Cook et al. (2007).

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

$$Q_{\mathcal{S}} Y \mid X \sim Q_{\mathcal{S}} Y, \quad P_{\mathcal{S}} Y \perp\!\!\!\perp Q_{\mathcal{S}} Y \mid X, \tag{2}$$

where $P_{\mathcal{S}}$ is the projection onto $\mathcal{S}$ and $Q_{\mathcal{S}} = I_r - P_{\mathcal{S}}$ is the projection onto $\mathcal{S}^{\perp}$, the orthogonal complement of $\mathcal{S}$. The two statements in (2) imply that the distribution of $Q_{\mathcal{S}} Y$ is not affected by X and that $Q_{\mathcal{S}} Y$ 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 into the estimation. Under the multivariate linear model (1), the conditions in (2) are equivalent to the parametric conditions

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

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 upon varying x, changes in the conditional mean function $E(Y \mid X = x) = \alpha + \beta x$ lie within the subspace $\mathcal{S}$; the second statement in (3) implies that, given X, we obtain 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 behind 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 interest, 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 way to estimate 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 $(P_{\mathcal{S}}Y, X)$ is minimized. This idea is in the same spirit as some recent advances in sufficient dimension reduction with distance covariance (e.g., Sheng & Yin, 2016; Chen et al., 2019; Sheng & Yuan, 2019). 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 central mean subspace in dimension reduction (Cook & Li, 2002), which focuses on the conditional mean function $E(Y \mid X)$ 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, so that our focus is on prediction and inference for the conditional mean function. In particular, in the following we replace the parametric envelope model assumptions in (3) with a more general mean dependence and conditional covariance reduction. We define the central mean envelope as the smallest subspace $\mathcal{S} \subseteq \mathbb{R}^r$ such that

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

This notion of the central mean envelope 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 the central mean envelope is model-free, we can 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 $\operatorname{cov}(Y \mid X = x) = \Sigma(x)$ to depend on x, and we redefine $\Sigma = E\{\operatorname{cov}(Y \mid X)\} = E\{\Sigma(X)\}$, which reduces to the same Σ as in the multivariate linear model (1). While the central mean envelope reduces all $\Sigma(x)$, a weaker condition than (4) leads to the martingale difference divergence envelope, which is the reducing subspace of Σ.

While most existing envelope methods concentrate on parametric models, our work differs in 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 the development. As such, our development of the central mean envelope enriches the set of dimension reduction techniques for the 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, our new definitions of the central mean envelope and the martingale difference divergence envelope are consistent with the standard envelope, and bridge the gap between standard envelopes in multivariate linear models and general conditional independence in sufficient dimension reduction: the central mean envelope implies conditional independence of $Q_{\mathcal{S}}Y \perp\!\!\!\perp (P_{\mathcal{S}}Y, X)$ if $Y \mid X$ is normally distributed; the central mean envelope contains the martingale difference divergence envelope, with equality when $Y \mid X$ has a constant conditional covariance; and the martingale difference divergence envelope contains the standard envelope, with equality when $Y \mid X$ has a linear conditional mean. The notion of the central mean envelope is completely generic for any two random vectors X and Y. Parallel to the recent developments of standard envelopes, our central mean envelope framework is 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 tensor envelopes (Li & Zhang, 2017; Zhang & Li, 2017).

2. Brief review of the martingale difference divergence matrix

Lee & Shao (2018) introduced the martingale difference divergence matrix, which can be viewed as an extension of the martingale difference divergence (Shao & Zhang, 2014; Park et al., 2015) from a scalar to a matrix, and 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 $E(\|Y\|^2 + \|X\|) < \infty$, then

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

where (Y', X') is an independent copy of (Y, X). The notation $M_{Y|X}$ abbreviates the $\text{MDDM}(Y \mid X)$ used in Lee & Shao (2018). From (5), $M_{Y|X} \in \mathbb{R}^{r\times r}$ is a real, symmetric and positive-semidefinite matrix. We assume $E(\|Y\|^2 + \|X\|) < \infty$ 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 that X is independent of Y in mean if $E(X \mid Y) = E(X)$. Clearly, independence in mean is not a symmetric property; X being independent of Y in mean does not imply that 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 $E(Y \mid X = x) - E(Y) \in \text{span}(M_{Y|X})$. Moreover,* $\text{span}[\text{cov}\{E(Y \mid X)\}] = \text{span}(M_{Y|X})$.

The above lemma suggests that we can simply eigendecompose the matrix $M_{Y|X}$ and use the nontrivial eigenvectors to span the subspace $\mathcal{S}$ so that $Q_{\mathcal{S}}Y$ is independent of X in mean, i.e., $E(Q_{\mathcal{S}}Y \mid X) = E(Q_{\mathcal{S}}Y)$ 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 $\hat{M}_{Y|X} = -n^{-2}\sum_{k,l=1}^n (Y_k - \bar{Y}_n)(Y_l - \bar{Y}_n)^{\mathrm{T}}\|X_k - X_l\|$ where $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$.

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 $E(Y \mid X)$, so we naturally want to know the subspace $\mathcal{S} \subseteq \mathbb{R}^r$ such that $E(Y \mid X) = E(P_{\mathcal{S}}Y \mid X) + E(Q_{\mathcal{S}}Y \mid X) = E(P_{\mathcal{S}}Y \mid X) + E(Q_{\mathcal{S}}Y)$ and restrict our attention to $E(P_{\mathcal{S}}Y \mid X)$. The immaterial part $Q_{\mathcal{S}}Y$ is independent of X in mean, and hence the variability in $Q_{\mathcal{S}}Y$ cannot be reduced by regressing on X. As stated in (4), we also want $\text{cov}(Q_{\mathcal{S}}Y, P_{\mathcal{S}}Y \mid X) = 0$ so that the variation in $Q_{\mathcal{S}}Y$ 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 central mean envelope is the following.

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

By construction, the central mean envelope $\mathcal{E}_{E(Y|X)}$ always exists and is unique. Definition 2 also generalizes the classical definition of envelopes in multivariate linear models: it is a generic definition involving two random vectors and is thus model-free; the mean function may no longer have a linear form, and the conditional covariance may depend on x.

Moreover, the central mean envelope also connects to the parametric form of envelopes in Definition 1.

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

From the above proposition, the central mean envelope can be viewed 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 models, the central mean envelope does not have to be associated with a parameter subspace such as $\mathrm{span}(\beta)$ in response reduction or $\mathrm{span}(\beta^{\mathrm{T}})$ in predictor reduction. However, directly estimating the central mean envelope 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 to facilitate the estimation of the central mean envelope.

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

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

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

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

The notion of martingale difference divergence envelope is more intuitive from this proposition. Comparing it with the central mean envelope defined by (4), the only difference is in the second statement: while the central mean envelope requires $Q_{\mathcal{S}}Y$ and $P_{\mathcal{S}}Y$ to be conditionally uncorrelated given X, the martingale difference divergence envelope requires them to be marginally uncorrelated. The following proposition establishes a more explicit connection between the martingale difference divergence envelope and the central mean envelope.

PROPOSITION 3. *The martingale difference divergence envelope is contained in the central mean envelope, $\mathcal{E}_{\Sigma}(M_{Y|X}) \subseteq \mathcal{E}_{E(Y|X)}$, and equality holds if* $\mathrm{cov}(Y \mid X)$ *does not depend on X.*

This proposition implies that the martingale difference divergence envelope and the central mean envelope are identical in the multivariate linear regression setting (1): $\mathcal{E}_{\Sigma}(\beta) =$

$\mathcal{E}_\Sigma(M_{Y|X}) = \mathcal{E}_{E(Y|X)}$. Also, one can improve the estimation of the central mean envelope by focusing on the martingale difference divergence envelope first. The martingale difference divergence envelope is itself of substantial interest since it fully captures the potentially nonlinear dependence in mean, and also retains the marginal uncorrelated material and immaterial information. More importantly, unlike estimation of the central mean envelope in general, estimation of the martingale difference divergence envelope 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 central mean envelope and the martingale difference divergence envelope. Consider the following regression of a multivariate response $Y_i \in \mathbb{R}^r$ on a univariate predictor $X_i \in \mathbb{R}^1$:

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

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

Then the central mean envelope, 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 $\Gamma \equiv (\gamma_1, \gamma_2, \ldots, \gamma_u) \in \mathbb{R}^{r\times u}$ be a semi-orthogonal basis matrix for the central mean envelope and let $(\Gamma, \Gamma_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix; then we have the coordinate representation

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

where $f(X) \in \mathbb{R}^u$ is the lower-dimensional latent function that reflects the nonlinear mean change of Y captured within the central mean envelope, 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 central mean envelope and those orthogonal to the central mean envelope.

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

In Fig. 1 we plot the estimated central mean envelope components $\hat{\Gamma}^{\mathrm{T}}Y = (\hat{\gamma}_1^{\mathrm{T}}Y, \hat{\gamma}_2^{\mathrm{T}}Y)^{\mathrm{T}} \in \mathbb{R}^2$ versus the univariate predictor X, where $\hat{\gamma}_1$ is also the estimated basis matrix for the martingale difference divergence envelope. Because of the nonlinearity, the standard envelope component $\hat{\beta}^{\mathrm{T}}Y$ fails to detect any meaningful information. In contrast, the first central mean envelope component captures the nonlinearity clearly, while the second central mean envelope component reveals the heteroscedasticity.

4. ESTIMATION PROCEDURES

4.1. *Estimating the martingale difference divergence envelope*

Recall from Proposition 2 that the martingale difference divergence envelope $\mathcal{E}_\Sigma(M_{Y|X})$ equals $\mathcal{E}_{\Sigma_Y}(M_{Y|X})$. Since the marginal covariance $\Sigma_Y = \mathrm{cov}(Y)$ is much easier to estimate than

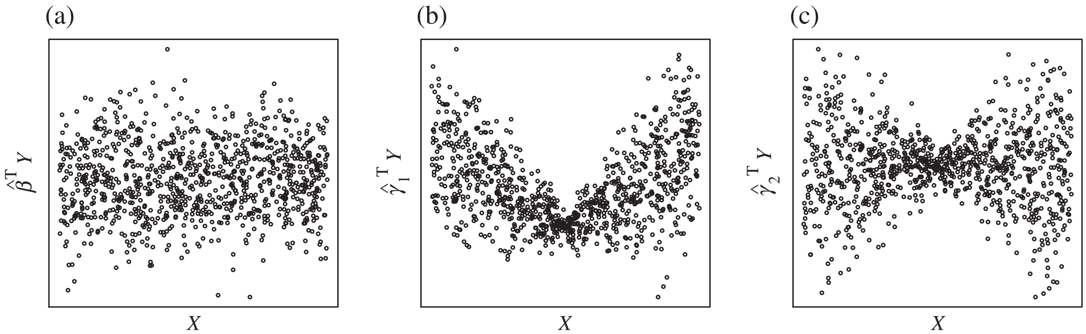

Fig. 1. Estimated envelope components of a multivariate response Y versus a univariate predictor X from simulated data: (a) standard envelope; (b) first component ($\hat{\gamma}_1^{\mathrm{T}} Y$) of the central mean envelope; (c) second component ($\hat{\gamma}_2^{\mathrm{T}} Y$) of the central mean envelope.

$\Sigma = E\{\mathrm{cov}(Y \mid X)\}$ in nonlinear regression, the form $\mathcal{E}_{\Sigma_Y}(M_{Y|X})$ is more useful in estimation. Given the dimension $u_1 = \dim\{\mathcal{E}_{\Sigma_Y}(M_{Y|X})\}$, we propose the following optimization for estimating the martingale difference divergence envelope as $\mathrm{span}(\hat{G})$:

$$\hat{G} = \underset{G^{\mathrm{T}}G=I_{u_1}}{\arg\min} \log\left|G^{\mathrm{T}}(\hat{\Sigma}_Y + \hat{M}_{Y|X})^{-1}G\right| + \log|G^{\mathrm{T}}\hat{\Sigma}_Y G|, \tag{8}$$

where $\hat{M}_{Y|X}$ is the sample martingale difference divergence matrix and $\hat{\Sigma}_Y$ is the sample covariance of Y.

The objective function in (8) can be viewed as the partially optimized pseudolikelihood of model-free envelope estimation (Zhang & Mai, 2018, § 3.1). Since $\hat{M}_{Y|X}$ and $M_{Y|X}$ are both symmetric positive-semidefinite matrices, we can write $M_{Y|X} = VV^{\mathrm{T}}$ and $\hat{M}_{Y|X} = \hat{V}\hat{V}^{\mathrm{T}}$ for some $V, \hat{V} \in \mathbb{R}^{r\times r}$. Then $\mathrm{span}(V) = \mathrm{span}(M_{Y|X})$. The parameter of interest $(\Sigma_Y, M_{Y|X})$ is now reparameterized as (Σ_Y, V). As such, the pseudo-loglikelihood function for jointly estimating Σ_Y and V can be written as

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

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

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

The proposed objective function in (8) for the martingale difference divergence envelope and the corresponding pseudolikelihood 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 $Y \mid X \sim N(\beta X, \Sigma)$. Comparing $\mathcal{E}_\Sigma(\beta) = \mathcal{E}_\Sigma(\beta\Sigma_X\beta^{\mathrm{T}})$ and

$\mathcal{E}_{\Sigma_Y}(M_{Y|X})$, one can see the analogy between the fitted value covariance $\beta\Sigma_X\beta^{\mathrm{T}} = \mathrm{cov}\{E(Y \mid X)\} = \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY} \equiv \Sigma_{\mathrm{fit}}$ and the martingale difference divergence matrix $M_{Y|X}$, where $\Sigma_{YX} = \mathrm{cov}(Y, X)$. Moreover, as we replaced $\hat{\Sigma}_Y$ and $\hat{M}_{Y|X}$ in (8) with the sample least-squares estimates $(\hat{\Sigma}, \hat{\Sigma}_{\mathrm{fit}})$, the maximum likelihood estimator of $\mathcal{E}_\Sigma(\beta)$ is reproduced by the same optimization. Our optimization (8) and pseudolikelihood argument (9) fit into 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 of Cook & Zhang (2016) and the 1D dimension selection procedure of Zhang & Mai (2018).

4.2. *Estimating the central mean envelope: some intuition*

From Proposition 1, $\mathcal{E}_{E(Y|X)} = \sum_x \mathcal{E}_{\Sigma(x)}(M_{Y|X})$. In estimation, we approximate $\Sigma(x)$ for all values of x by a finite number of covariance matrices, Σ_h for $h = 1, \ldots, H$ where $H \geqslant 2$. 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 (Li, 1991). For multivariate X, as in K-means inverse regression (Setodji & Cook, 2004), which extends the sliced inverse regression from univariate to multivariate responses, we construct H clusters of X by the K-means clustering algorithm. In this way we can obtain sample covariance matrices $\hat{\Sigma}_h$ based on the n_h samples in the hth slice/cluster, where $\sum_{h=1}^H n_h = n$.

If we assume normality and a constant mean function within each slice or 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 central mean envelope $\mathcal{E}_{E(Y|X)}$ becomes the smallest subspace that reduces all the Σ_h and contains the mean subspace $\mathrm{span}\{\mu_1 - E(X), \ldots, \mu_H - E(X)\}$. A similar envelope structure has been studied in groupwise regression (Su & Cook, 2013; Park et al., 2017) and in quadratic discriminant analysis (Zhang & Mai, 2019). It can be estimated from the likelihood-based optimization

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

which is used in Park et al. (2017) and Zhang & Mai (2019). However, since we can estimate the martingale difference divergence envelope straightforwardly and accurately, there is no need to approximate $E(Y \mid X)$ by $\mu_1, \ldots, \mu_H$. We propose a more accurate and practical estimation procedure for the central mean envelope in the next subsection.

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

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

where the immaterial variation $\Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}$ is static over h. This is because we want to fully capture the heteroscedasticity with the central mean envelope. By Proposition 3 of Zhang & Mai (2019), we know that (11) and (10) are in fact targeting an upper bound of the central mean envelope. This targeting subspace, called the envelope discriminant subspace, reduces all the Σ_h and contains the mean subspace $\mathrm{span}\{\mu_1 - E(X), \ldots, \mu_H - E(X)\}$ 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 the central mean envelope satisfies (11) or, more generally, $\Sigma(X) = \Gamma\Omega(X)\Gamma^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}$. When this assumption fails, we are effectively targeting 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) derived properties of sufficient dimension reduction subspaces.

4.3. *Estimating the central mean envelope: a novel two-part estimation*

From Proposition 3 we know that the martingale difference divergence envelope is always a portion of the central mean envelope: $\mathcal{E}_\Sigma(M_{Y|X}) \subseteq \mathcal{E}_{E(Y|X)}$. Suppose we know the dimensions $u = \dim\{\mathcal{E}_{E(Y|X)}\}$ and $u_1 = \dim\{\mathcal{E}_\Sigma(M_{Y|X})\}$, where $u_1 \leqslant u$.

When $\mathcal{E}_\Sigma(M_{Y|X}) = \mathcal{E}_{E(Y|X)}$, or equivalently $u_1 = u$, the estimation procedures in § 4.1 and § 4.2 are different, and generally produce different estimators for the same subspace. Even when $\Sigma(x)$ is not constant, the two subspaces may be the same even though they come from different definitions. To address this issue, we propose a two-part estimation approach that involves first obtaining the martingale difference divergence envelope and then extracting the unique part of the central mean envelope that is not in the martingale difference divergence envelope. The two-part estimator of the central mean envelope reduces to the same estimator of the martingale difference divergence envelope in § 4.1 if $\mathcal{E}_\Sigma(M_{Y|X}) = \mathcal{E}_{E(Y|X)}$. For $u_1 < u$, we develop the two-part estimation procedure of the central mean envelope as follows.

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

$$\hat{H} = \underset{H^{\mathrm{T}}H=I_{u_2}}{\arg\min}\ \log\left|H^{\mathrm{T}}(\hat{G}_0^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{G}_0)H\right| + \sum_{h=1}^{H}\frac{n_h}{n}\log\left|H^{\mathrm{T}}(\hat{G}_0^{\mathrm{T}}\hat{\Sigma}_h\hat{G}_0)H\right|, \tag{12}$$

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

LEMMA 3. *Let $\Sigma_h \in \mathbb{R}^{r\times r}$, for $h = 1, \ldots, H$ with $H \geqslant 2$, 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 that $\mathcal{E} \equiv \text{span}(\Gamma) \subseteq \mathbb{R}^r$ is the smallest subspace such that $\Sigma_h = \Gamma\Omega_h\Gamma^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}$ holds for all h for some Ω_h and Ω_0, and that $\dim(\mathcal{E}) = u$. Then $\mathcal{E} = \text{span}(\hat{G})$, where $\hat{G}$ is defined as*

$$\hat{G} = \underset{G\in\mathbb{R}^{r\times u},\, G^{\mathrm{T}}G=I_u}{\arg\min}\left\{\log|G^{\mathrm{T}}\Sigma^{-1}G| + \sum_{h=1}^{H}\pi_h\log|G^{\mathrm{T}}\Sigma_h G|\right\}. \tag{13}$$

The above lemma suggests a more effective objective function than the one 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 central mean envelope consistently; however, their motivations are different. As mentioned earlier, (10) is motivated by the likelihood for estimating the central mean envelope, while the new objective function (13) focuses more on the heterogeneity of the covariance matrices Σ_h, since Σ is the weighted average of all the Σ_h. This is indeed much more desirable in the two-part estimation because the first part, the martingale difference divergence envelope, 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 central mean envelope is almost always better than direct estimation from (10) or (13). This is due to the fact that the martingale difference divergence envelope part is easier to estimate, as the sample matrix $\hat{M}_{Y|X}$ in (8) is more accurately estimated than $\hat{\Sigma}_h$ in (10) and (13), especially when the sample size n_h is small for some slices or clusters. Moreover, the optimization in (8) is more feasible than the optimization in (10) and (13). As 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 standard optimization methods with orthogonality constraints (e.g., Absil et al., 2009; Wen & Yin, 2013). For this reason, we use the two-part estimation of the central mean envelope in all our numerical studies.

4.4. *Consistency*

We establish the $\sqrt{n}$-consistency of the estimator $\text{span}(\hat{G})$ from (8) for the martingale difference divergence envelope, and of the two-part estimator $\text{span}(\hat{G}, \hat{G}_0\hat{H})$ from (8) and (12) for the central mean envelope. 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 assumptions. Instead, we assume that $\hat{\Sigma}_Y$, $\hat{\Sigma}_h$ $(h = 1, \ldots, H)$ and $\hat{M}_{Y|X}$ are $\sqrt{n}$-consistent estimators for their population counterparts $\Sigma_Y > 0$, $\Sigma_h > 0$ and $M_{Y|X} \geqslant 0$. This is a mild assumption that can be satisfied easily: the sample covariance matrices $\hat{\Sigma}_Y$ and $\hat{\Sigma}_h$ are $\sqrt{n}$-consistent when Y and $Y \mid X$ have finite fourth moments, while $\sqrt{n}$-consistency of the sample martingale difference divergence matrix $\hat{M}_{Y|X}$ was established by Lee & Shao (2018). The consistency of the martingale difference divergence envelope estimation is obtained by applying Proposition 3 of Cook & Zhang (2016).

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

We can speed up the computation by employing the 1D algorithm of Cook & Zhang (2016), which sequentially estimates one direction at a time for the martingale difference divergence envelope. The resulting estimator, $\hat{G}^{1\text{D}} \in \mathbb{R}^{r\times u_1}$, is no longer a minimizer of (8), but the $\sqrt{n}$-consistency in Proposition 4 still holds if we replace $\hat{G}$ with $\hat{G}^{1\text{D}}$ and apply 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, $\text{pr}(\hat{u}_1 = u_1) \to 1$ as $n \to \infty$. We will demonstrate this dimension selection procedure in our numerical studies.

For the second part of the estimation of the central mean envelope, 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 let $G_0 \in \mathbb{R}^{r\times(r-u_1)}$ 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 $\hat{G} \in \mathbb{R}^{r\times u_1}$ be any minimizer of* (8) *and $\hat{H} \in \mathbb{R}^{(r-u_1)\times u_2}$ any minimizer of* (12)*. If* (14) *is true, then $P_{\hat{G}_0\hat{H}}$ is $\sqrt{n}$-consistent for the projection onto $\mathcal{D}$. Moreover, the projection onto* $\text{span}(\hat{G}, \hat{G}_0\hat{H})$ *is $\sqrt{n}$-consistent for the projection onto $\mathcal{E}_{E(Y|X)}$.*

The assumption in (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 $h \geqslant 2$ in the example of § 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 estimators from different slicing schemes. The fused estimator in Cook & Zhang (2014) can be directly applied to our optimization in (12), and in a way it also circumvents the issue of choosing an 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 central mean envelope. We compare our method with two closely related methods: the standard envelope estimator of Cook et al. (2010) and the dimension reduction method which decomposes the martingale difference divergence matrix directly (Lee & Shao, 2018). In the presence of a nonlinear mean function, the standard envelope model based on (1) and $\mathcal{E}_\Sigma(\beta)$ is misspecified. However, since $\mathcal{E}_\Sigma(\beta) = \mathcal{E}_\Sigma(\Sigma_{YX})$, the standard envelope remains a well-defined subspace in all of our simulation examples, but is ineffective at detecting nonlinear and heteroscedastic envelope components. The martingale difference divergence matrix method performs eigen-decomposition of $\hat{M}_{Y|X}$ and targets directly the mean subspace $\mathrm{span}(M_{Y|X}) = \mathrm{span}\{E(Y \mid X = x) - E(Y)$ for all $x\}$. In comparison, the proposed two-part estimator, in particular the martingale difference divergence envelope part from (8), has a bigger advantage when the responses Y are highly correlated.

In addition, by interchanging the roles of X and Y we can compare our method with three popular sufficient dimension reduction methods: sliced inverse regression (Li, 1991), sliced average variance estimation (Cook & Weisberg, 1991), and principal fitted components (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 aim 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 $Y \mid X \sim Y \mid \beta^{\mathrm{T}}X$. The central subspace is 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-semidefinite 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, for example by omitting the standardization step in estimation, and interchanging the roles of X and Y, the targeting subspace of these methods becomes a subset of the central mean envelope. Therefore, for a fair comparison 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 to 15 and take the sample size to be $n = 200$ 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., $\|P_\Gamma - P_{\hat{\Gamma}}\|_{\mathrm{F}}$. We also compute the principal angle between each of the central mean envelope directions and the estimated subspace, i.e., $\theta_j = \cos^{-1}\{(\gamma_j^{\mathrm{T}} P_{\hat{\Gamma}} \gamma_j)^{1/2}\}$ $(j = 1, \ldots, u)$, where $\Gamma = (\gamma_1, \ldots, \gamma_u)$. The values of θ_j are bounded between 0 and 90, where $\theta_j = 0$ indicates that γ_j is contained in $\mathrm{span}(\hat{\Gamma})$ and $\theta_j = 90$ indicates that γ_j is orthogonal to $\mathrm{span}(\hat{\Gamma})$. For sliced inverse regression, sliced average variance estimation and the second-part estimation of the central mean envelope, the number of slices H is 5; for principal fitted components, the function basis is the cubic polynomial $(x, x^2, x^3)^{\mathrm{T}}$.

Table 1. *Estimation errors* $\|P_\Gamma - P_{\hat\Gamma}\|_F$ *averaged over* 100 *replicates, with standard errors* ($\times 100$) *in parentheses*

Model	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)

SIR, sliced inverse regression (Li, 1991); SAVE, sliced average variance estimation (Cook & Weisberg, 1991); PFC, principal fitted components (Cook & Forzani, 2008b); MDDM, martingale difference divergence matrix (Lee & Shao, 2018); Standard, the standard envelope of Cook et al. (2010); MDDE, martingale difference divergence envelope, the proposed estimator from (8).

5.2. *The standard envelope coincides with the central mean envelope*

We first consider four models where the standard envelope, the martingale difference divergence envelope and the central mean envelope are all the same in population. Therefore, the two-part central mean envelope estimator is identical to the martingale difference divergence envelope estimator based on (8). Similar to the data-generating process in § 3.3, the central mean envelope basis is randomly generated and orthogonalized, $\Gamma = (\gamma_1, \ldots, \gamma_u) \in \mathbb{R}^{r\times u}$, and the predictor X follows the $\text{Un}(-1, 1)$ distribution unless otherwise specified. Here, we use $\text{Un}(a, b)$ to denote the uniform distribution on (a, b). The errors are generated as $E_i = 0.1\,\Sigma^{1/2}(X_i)\,\varepsilon_i$, where the ε_i are independent and identically distributed as $N(0, I_r)$ and $\Sigma(x) = \Gamma\Omega(x)\Gamma^{\mathrm T} + \Gamma_0\Omega_0\Gamma_0^{\mathrm T}$ is specified as follows. We use $O_k \in \mathbb{R}^{k\times k}$ to denote an arbitrarily generated orthogonal matrix.

Model I. Linear mean and constant covariance: $Y = \gamma_1 X + E_i$, $\Gamma = \gamma_1 \in \mathbb{R}^{r\times 1}$, $\Omega_0 = O_{r-1}\,\text{diag}\{\exp(3.0), \exp(2.5), \ldots, \exp(-3.5)\}O_{r-1}^{\mathrm T}$ and $\Omega = \exp(5)$.

Model II. Nonlinear mean and constant covariance: $Y = \gamma_1 X + \gamma_2 \exp(2|X|) + E_i$, $\Gamma = (\gamma_1, \gamma_2) \in \mathbb{R}^{r\times 2}$, $\Omega_0 = O_{r-2}\,\text{diag}\{\exp(3.0), \exp(2.5), \ldots, \exp(-3.0)\}O_{r-2}^{\mathrm T}$ and $[\Omega]_{ij} = 0.5^{|i-j|}\exp(5)$.

Model III. Linear mean and nonconstant covariance: $Y = \gamma_1 X + 0.1\,\Sigma^{1/2}(X_i)\,\varepsilon_i$, where Γ and $\Sigma = E\{\Sigma(X)\}$ are the same as in Model II, but the off-diagonal of $\Omega(x)$ is $\Omega_{12}(x) = |x|\exp(5)$.

Model IV. Nonlinear mean and 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 $X \sim \text{Un}(-1, 1)$.

The results are summarized in Table 1, where all methods are compared in terms of the overall estimation error $\|P_\Gamma - P_{\hat\Gamma}\|_F$. The martingale difference divergence envelope estimator based on (8) is always the best except in Model I, for which the standard envelope is the maximum likelihood estimator and has a slight advantage. In constant-covariance scenarios, i.e., Models I and II, the martingale difference divergence envelope has similar performance to the better of the standard envelope and the martingale difference divergence matrix estimators. This is because the martingale difference divergence envelope estimator is essentially a hybrid of the

Table 2. *Estimation errors of the second direction, $\theta_2 = \cos^{-1}\{(\gamma_2^{\mathrm{T}} P_{\hat{\Gamma}} \gamma_2)^{1/2}\}$, with standard errors in parentheses*

Model	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)

SIR, sliced inverse regression (Li, 1991); SAVE, sliced average variance estimation (Cook & Weisberg, 1991); PFC, principal fitted components (Cook & Forzani, 2008b); MDDM, martingale difference divergence matrix (Lee & Shao, 2018); Standard, the standard envelope of Cook et al. (2010); MDDE, martingale difference divergence envelope, the proposed estimator from (8).

envelope estimator, which incorporates covariance structural information of Σ, and the martingale difference divergence matrix, which is effective at capturing nonlinear means. In nonconstant-covariance scenarios, i.e., Models III and IV, the martingale difference divergence envelope has better performance than both the standard envelope and the martingale difference divergence matrix.

For models with two-dimensional subspaces, we further compute the principal angles between the second central mean envelope direction γ_2 and the estimated subspace, i.e., $\theta_2 = \cos^{-1}\{(\gamma_2^{\mathrm{T}} P_{\hat{\Gamma}} \gamma_2)^{1/2}\}$. For Model III, the second direction is only estimable from the nonconstant covariance, but does not appear in the mean function. Therefore sliced inverse regression, principal fitted components and the martingale difference divergence matrix fail to estimate γ_2 accurately, while sliced average variance estimation, the standard envelope and the proposed estimator perform well. For Models II and IV, the second direction appears in both the nonlinear mean and the covariance. Therefore the nonlinear methods all work well, while the standard envelope method becomes ineffective; see Table 2 for details. Overall, the martingale difference divergence envelope estimator is the most reliable method.

5.3. *The standard envelope and the martingale difference divergence envelope are proper subsets of the central mean envelope*

In this subsection we consider three models where $\mathcal{E}_\Sigma(\Sigma_{YX}) = \mathcal{E}_\Sigma(\gamma_1) \subsetneq \mathcal{E}_\Sigma(M_{Y|X}) \subset \mathcal{E}_{E(Y|X)} = \mathrm{span}(\Gamma)$. The central mean envelope estimator is now based on the two-part estimation in § 4.2.

Model V. Linear mean and nonconstant covariance: same as Model III except that we change the off-diagonal of $\Omega(X)$ from $|X|\exp(5)$ to X, so $\mathcal{E}_\Sigma(\gamma_1) = \mathcal{E}_\Sigma(M_{Y|X}) = \mathrm{span}(\gamma_1) \subset \mathcal{E}_{E(Y|X)}$.

Model VI. Nonlinear mean and nonconstant covariance: same as Model V except that the mean function is $\gamma_2\exp(2|X|)$ and is linearly independent of X, so $\mathcal{E}_\Sigma(\gamma_1) \subset \mathcal{E}_\Sigma(M_{Y|X}) \subset \mathcal{E}_{E(Y|X)}$.

Model VII. Nonlinear mean, nonconstant covariance and multivariate predictor: like Model IV, we have a nonlinear mean function $\gamma_2\{\exp(2|X_1|)+\exp(|X_2|+|X_3|)+\exp(|X_4|+|X_5|)\}$, where the multivariate predictor is five-dimensional, $X = (X_1,\ldots,X_5)^{\mathrm{T}}$, and each coordinate is uniformly distributed from -1 to 1; the covariance structure is the same as in the previous two models except that $\Omega_{12}(X) = X_4 + X_5$. We have $\mathcal{E}_\Sigma(\gamma_1) \subset \mathcal{E}_\Sigma(M_{Y|X}) \subset \mathcal{E}_{E(Y|X)}$.

Table 3. *Estimation errors* $\|P_\Gamma - P_{\hat{\Gamma}}\|_F$ *averaged over* 100 *replicates, with standard errors* $(\times 100)$ *in parentheses*

Model	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)

SIR, sliced inverse regression (Li, 1991); SAVE, sliced average variance estimation (Cook & Weisberg, 1991); PFC, principal fitted components (Cook & Forzani, 2008b); MDDM, martingale difference divergence matrix (Lee & Shao, 2018); Standard, the standard envelope of Cook et al. (2010); CME, the proposed two-part estimator.

Table 4. *Estimation errors of the second direction,* $\theta_2 = \cos^{-1}\{(\gamma_2^{\mathrm{T}} P_{\hat{\Gamma}} \gamma_2)^{1/2}\}$*, with standard errors in parentheses*

Model	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)

SIR, sliced inverse regression (Li, 1991); SAVE, sliced average variance estimation (Cook & Weisberg, 1991); PFC, principal fitted components (Cook & Forzani, 2008b); MDDM, martingale difference divergence matrix (Lee & Shao, 2018); Standard, the standard envelope of Cook et al. (2010); CME, the proposed two-part estimator.

The overall estimation error $\|P_\Gamma - P_{\hat{\Gamma}}\|_F$ is summarized in Table 3, and the estimation error $\theta_2 = \cos^{-1}\{(\gamma_2^{\mathrm{T}} P_{\hat{\Gamma}} \gamma_2)^{1/2}\}$ of the second central mean envelope direction γ_2 is reported in Table 4. For these more challenging models, the proposed central mean envelope estimator is much more accurate than the competitors, especially in estimating the second direction, which comes from the nonlinearity and heteroscedastic error.

We also tried different numbers of slices/clusters, $H = 2, 5, 10$ and 15, for Models V–VII. The results are indistinguishable from those in Tables 3 and 4, where $H = 5$ was used. The overall performance of the two-part estimation is very encouraging.

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 for the standard envelope, the martingale difference divergence envelope, and the central mean envelope. Implementation details are given in the Supplementary Material. Table 5 summarizes the dimension selection results for the central mean envelope under the previous simulation models. For Models II–IV, because the standard envelope coincides with the central mean envelope, both methods should be able to determine the true dimension consistently as $n \to \infty$. Clearly, the central mean envelope method has much better finite-sample performance. For Models V and VI, the standard envelope failed to detect the second central mean envelope components and, not surprisingly, underselected the dimension. The central mean envelope dimension selection yielded reasonable results for these two very

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

Model	Standard envelope						CME					
	$\hat{u} < u$		$\hat{u} = u$		$\hat{u} > u$		$\hat{u} < u$		$\hat{u} = u$		$\hat{u} > u$	
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

CME, the proposed two-part estimator.

challenging models. There is also a significant improvement in accuracy when the sample size is increased from 200 to 600.

6. Real-data illustration

In an application to real data, we demonstrate the advantages of the central mean envelope over the standard envelope in prediction and subspace estimation. The riboflavin dataset from Bühlmann et al. (2014) contains 71 samples of the riboflavin production rate and expression level of 4088 genes. We use the logarithm of the riboflavin production rate as the predictor and the logarithm of the gene expression level as the response. As in Bühlmann et al. (2014), and because of the high dimensionality, we focus on the top 50 genes that are selected by the martingale difference correlation (Shao & Zhang, 2014). We use the same dimension selection procedures as in § 5.4 for the standard envelope and the central mean envelope. For the standard envelope, the selected dimension is $u = 1$. For the central mean envelope, the selected dimensions are $u_1 = 1$ for the first part, the nonlinear mean, and $u_2 = 0$ for the second part, the heteroscedastic error. In the Supplementary Material, the nonlinearity in $E(\hat{\gamma}^{\mathrm{T}} Y \mid X)$ is clearly demonstrated, where $\hat{\gamma}$ is the estimated central mean envelope, which has dimension one and is the same as the martingale difference divergence envelope.

For numerical comparison, we divide the data into training and testing sets by randomly choosing 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 the response on the predictor, $\hat{E}(\hat{\gamma}^{\mathrm{T}} Y \mid X)$, while the immaterial part of the response is predicted by its unconditional mean $\hat{E}(\hat{\Gamma}_0^{\mathrm{T}} Y)$. Specifically, for kernel regression, we use a Gaussian kernel with the optimal bandwidth from the `ksr` function in Matlab. We repeat this procedure 100 times, and report 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, $B^{-1} \sum_{b=1}^{B} \|P_{\hat{\mathcal{S}}} - P_{\hat{\mathcal{S}}^b}\|_{\mathrm{F}}$, where $\hat{\mathcal{S}}$ is the estimated subspace on the original data and $\hat{\mathcal{S}}^b$ $(b = 1, \ldots, B)$ is the estimated subspace on m bootstrap samples from $B = 200$ bootstrap replicates. This is a commonly used criterion in the sufficient dimension reduction literature (see, e.g., Ye & Weiss, 2003; Luo & Li, 2016). We consider $m = 71$ bootstrap samples; then the covariance can occasionally become ill-conditioned and so we add $0.01 I_r$ to the sample covariance $\hat{\Sigma}_Y$. Alternatively, we consider $m = 150$. From Table 6 it can be seen that the proposed method improves the subspace estimation significantly.

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

	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)

7. Discussion

Although the focus of this paper has been on multivariate response reduction in regression, the two new envelope structures, namely the martingale difference divergence envelope and the central mean envelope, are model-free and can be used in situations beyond regression models. For example, in the Supplementary Material we apply the proposed method to handwritten digit recognition data, and demonstrate the usefulness of the central mean envelope as a data visualization tool in discriminant analysis and classification. Moreover, by interchanging the roles of X and Y, the martingale difference divergence envelope $\mathcal{E}_{\Sigma_X}(M_{X|Y})$ or the central mean envelope $\mathcal{E}_{E(X|Y)}$ can serve as an upper bound for 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 the framework to simultaneous predictor and response reduction and to stationary multivariate time series. The former would be an extension of the simultaneous envelope in a multivariate linear model (Cook & Zhang, 2015b), while the latter can be achieved by using the cumulative version of $\hat{M}_{Y|X}$ (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 the paper, and also thank Professor R. Dennis Cook for his comments and suggestions. This research was partly supported by the U.S. National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online includes details of the dimension selection procedure, further information on the dataset, and proofs of the lemmas and propositions.

References

ABSIL, P.-A., MAHONY, R. & SEPULCHRE, R. (2009). *Optimization Algorithms on Matrix Manifolds*. Princeton, New Jersey: Princeton University Press.

BÜHLMANN, P., KALISCH, M. & MEIER, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annu. Rev. Statist. Appl.* **1**, 255–78.

CHEN, X., YUAN, Q. & YIN, X. (2019). Sufficient dimension reduction via distance covariance with multivariate responses. *J. Nonparam. Statist.* **31**, 268–88.

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

COOK, R. D. (2018). Principal components, sufficient dimension reduction, and envelopes. *Annu. Rev. Statist. Appl.* **5**, 533–59.

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. *Statist. Sci.* **23**, 485–501.

Cook, R. D. & Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *J. Am. Statist. Assoc.* **104**, 197–208.
Cook, R. D., Helland, I. S. & Su, Z. (2013). Envelopes and partial least squares regression. *J. R. Statist. Soc.* B **75**, 851–77.
Cook, R. D. & Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455–74.
Cook, R. D., Li, B. & Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika* **94**, 569–84.
Cook, R. D., Li, B. & Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica* **20**, 927–60.
Cook, R. D. & Weisberg, S. (1991). Comment on ‘Sliced inverse regression for dimension reduction’ by Ker-Chau Li. *J. Am. Statist. Assoc.* **86**, 328–32.
Cook, R. D. & Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *J. Am. Statist. Assoc.* **109**, 815–27.
Cook, R. D. & Zhang, X. (2015a). Foundations for envelope models and methods. *J. Am. Statist. Assoc.* **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. *J. Comp. Graph. Statist.* **25**, 284–300.
Cook, R. D. & Zhang, X. (2018). Fast envelope algorithms. *Statist. Sinica* **28**, 1179–97.
Dawid, A. P. (1979). Conditional independence in statistical theory. *J. R. Statist. Soc.* B **41**, 1–31.
Khare, K., Pal, S. & Su, Z. (2017). A Bayesian approach for envelope models. *Ann. Statist.* **45**, 196–222.
Lee, C. E. & Shao, X. (2018). Martingale difference divergence matrix and its application to dimension reduction for stationary multivariate time series. *J. Am. Statist. Assoc.* **113**, 216–29.
Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R*. Boca Raton, Florida: Chapman and Hall/CRC.
Li, K. C. (1991). Sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.* **86**, 316–27.
Li, L. & Zhang, X. (2017). Parsimonious tensor response regression. *J. Am. Statist. Assoc.* **112**, 1131–46.
Luo, W. & Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **103**, 875–87.
Matteson, D. S. & Tsay, R. S. (2017). Independent component analysis via distance covariance. *J. Am. Statist. Assoc.* **112**, 623–37.
Park, T., Shao, X. & Yao, S. (2015). Partial martingale difference correlation. *Electron. J. Statist.* **9**, 1492–517.
Park, Y., Su, Z. & Zhu, H. (2017). Groupwise envelope models for imaging genetic analysis. *Biometrics* **73**, 1243–53.
Setodji, C. M. & Cook, R. D. (2004). *K*-means inverse regression. *Technometrics* **46**, 421–9.
Shao, X. & Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *J. Am. Statist. Assoc.* **109**, 1302–18.
Sheng, W. & Yin, X. (2016). Sufficient dimension reduction via distance covariance. *J. Comp. Graph. Statist.* **25**, 91–104.
Sheng, W. & Yuan, Q. (2019). Sufficient dimension folding in regression via distance covariance for matrix-valued predictors. *Statist. Anal. Data Mining* **13**, 71–82.
Su, Z. & Cook, R. D. (2013). Estimation of multivariate means with heteroscedastic errors using envelope models. *Statist. Sinica* **23**, 213–30.
Székely, G. J. & Rizzo, M. L. (2009). Brownian distance covariance. *Ann. Appl. Statist.* **3**, 1236–65.
Székely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–94.
Vepakomma, P., Tonde, C. & Elgammal, A. (2018). Supervised dimensionality reduction via distance correlation maximization. *Electron. J. Statist.* **12**, 960–84.
Wang, W., Zhang, X. & Li, L. (2019). Common reducing subspace model and network alternation analysis. *Biometrics* **75**, 1109–20.
Wen, Z. & Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Math. Program.* **142**, 397–434.
Ye, Z. & Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Am. Statist. Assoc.* **98**, 968–79.
Zhang, X. & Li, L. (2017). Tensor envelope partial least-squares regression. *Technometrics* **59**, 426–36.
Zhang, X. & Mai, Q. (2018). Model-free envelope dimension selection. *Electron. J. Statist.* **12**, 2193–216.
Zhang, X. & Mai, Q. (2019). Efficient integration of sufficient dimension reduction and prediction in discriminant analysis. *Technometrics* **61**, 259–72.

[*Received on* 15 *February* 2019. *Editorial decision on* 23 *December* 2019]