# A Note on Cumulative Mean Estimation

Bilin Zeng[1], Zhou Yu[2], and Xuerong Meggie Wen[3]

[1]Department of Mathematics, California State University, Bakersfield, CA, U.S.A.
[2]School of Finance and Statistics, East China Normal University, Shanghai, China
[3]Department of Mathematics and Statistics,Missouri University of Science and Technology, MO, U.S.A.

**Abstract**
For many-valued or continuous $Y$, the standard practice in *sufficient dimension reduction* (Li, 1991; Cook, 1998) of replacing the response $Y$ with a discrete version of $Y$ usually results in the loss of power due to the loss of the intra-slice information. Most of the existing slicing methods highly rely on the choices of the total number of slices $h$. Zhu et al. (2010) proposed a method called the *cumulative slicing estimation* (CUME) which avoids the otherwise subjective selection of $h$. In this paper, we revisit CUME from a different perspective to gain more insights, and then refine its performance by incorporating the intra-slice covariances. We prove that our new method, which we call the *covariance cumulative slicing estimation* (COCUM), under some conditions, is more comprehensive than CUME since it captures a larger part of the central subspace. Simulation studies suggest that our method is comparable to CUME, and outperforms CUME when the predictors are skewed. The asymptotic results of COCUM are also proved.

**Key Words:** Cumulative Slicing Estimation, Dimension Reduction, Sliced Inverse Regression, Intraslice Covariance Estimation, Ensemble Estimator Approach.

## 1. Introduction

For a typical regression problem with a univariate random response $Y$ and a $p$-dimensional random vector $\mathbf{X}$, sufficient dimension reduction (SDR: Li, 1991; Cook, 1998) aims to reduce the dimension of $\mathbf{X}$ without loss of information on the regression and without requiring a pre-specified parametric model. The basic idea of sufficient dimension reduction is to replace the predictors $\mathbf{X} \in \mathbb{R}^p$ with a $d$-dimensional linear vector $\boldsymbol{\eta}^T\mathbf{X}$, where $\boldsymbol{\eta}$ is a $p \times d$ matrix with $d \leq p$, such that $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^T\mathbf{X}$. The column space of $\boldsymbol{\eta}$ is called a dimension reduction subspace, and the intersection of all such subspaces is called the *central subspace*, denoted by $\mathcal{S}_{Y|\mathbf{X}}$. The dimension $d$ of $\mathcal{S}_{Y|\mathbf{X}}$ is called the structural dimension of the regression. Under mild conditions (Yin, Li and Cook, 2008), the central subspace exists and is unique. We assume that $\mathcal{S}_{Y|\mathbf{X}}$ exists throughout this article. The goal of sufficient dimension reduction is to estimate and make statistical inferences about $\mathcal{S}_{Y|\mathbf{X}}$ and $d$.

Many methods have been developed to estimate $\mathcal{S}_{Y|\mathbf{X}}$. Among them, sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook and Weisberg, 1991), minimum average variance estimation (MAVE; Xia et al., 2002), and directional regression (DR; Li and Wang, 2007) are perhaps the most widely investigated methods in the literature. Cook and Li (2002) proposed the central mean subspace where the interest of dimension reduction is restricted to the conditional mean function $\mathrm{E}(Y|\mathbf{X})$, and the central mean subspace $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$ is defined as the smallest column spaces spanned by $\boldsymbol{\eta}$ such that

$\mathrm{E}(Y|\mathbf{X}) \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^T\mathbf{X}$.

For many-valued or continuous $Y$, the standard practice in SDR is to replace the response $Y$ with a discrete version $\check{Y}$ by partitioning the range of $Y$ into $h$ non-overlapping slices, then work on $\check{Y}$ and assume that $\mathcal{S}_{\check{Y}|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$. However, this assumption does not always hold, and the difference between the working and target regressions can be significant. Moreover, even when this assumption holds, we might still face the loss of power since we use only the information retained in $\check{Y}$, discarding all the intra-slice information.

The number of slices $h$ is a tuning parameter much like the tuning parameter encountered in the smoothing literature (Li, 1987; Härdle et al., 1988). Experience indicates that good results are often obtained by choosing $h$ to be somewhat larger than $d + 1$, and sometimes requires several trials of $h$. However, beyond empirical experiences, how to select the optimal $h$ still remains unknown in the literature.

Zhu et al. (2010) proposed a method called the *cumulative slicing estimation* (CUME) which sums up all possible estimations relating to $\mathrm{E}(\mathbf{X}I(Y \leq \tilde{y}))$ for all $\tilde{y}$ in the support of $Y$ to avoid the subjective selection of $h$. They showed that the estimator of CUME enjoys the common $\sqrt{n}$ convergence rate and is more efficient comparing to SIR and other first-moment slicing estimation methods.

In this article, we first revisit CUME from a different perspective to gain more insights, and then refine its performance by incorporating the intra-slice covariances. The rest of this article is organized as follows. In Section 2, we reinterpret CUME via the ensemble estimator approach (Yin and Li, 2011). Our new estimation method, which we call the *covariance cumulative slicing estimation* (COCUM), along with its asymptotic properties are investigated in Section 3. We illustrate the performances of our method via simulation studies in Section 4. A brief summary is given in Section 5. For easy of exposition, we defer all proofs to the Appendix.

## 2. CUME: the Ensemble Estimator Approach

### 2.1 CUME

In this subsection, we give a brief review of CUME (Zhu et al., 2010). For ease of exposition, we assume hereafter that $\mathrm{E}(\mathbf{X}) = 0$. Define

$$\mathbf{m}(\tilde{y}) = \mathrm{E}(\mathbf{X}I(Y \leq \tilde{y})) \tag{2.1}$$

for $\tilde{y} \in \mathbb{R}^1$. To preserve the integrity of $\mathcal{S}_{Y|\mathbf{X}}$, let $\tilde{Y}$ be an independent copy of $Y$ and the kernel matrix for CUME be defined by:

$$\mathbf{M} = \mathrm{E}[\mathbf{m}(\tilde{Y})\mathbf{m}^T(\tilde{Y})w(\tilde{Y})], \tag{2.2}$$

where $w(.)$ is a nonnegative weight function which is often set as 1.

Assuming the linearity condition (see Feng et al., 2013; more details will be given in Section 2.2), which is a mild condition imposed on the marginal distribution of $\mathbf{X}$ only, the column space of $\mathbf{M}$ is a subset of $\boldsymbol{\Sigma}\mathcal{S}_{Y|\mathbf{X}}$, where $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X})$. At the sample level, suppose $(\mathbf{X}_i, Y_i)$, $i = 1, \cdots, n$ are independent copies of $(\mathbf{X}, Y)$, we can estimate $\mathbf{M}$ by $\mathbf{M}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{m}_n(Y_i)\mathbf{m}_n^T(Y_i)w(Y_i)$, where $\mathbf{m}_n(Y_i) = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})I(Y_j \leq Y_i)$, and

$\bar{\mathbf{X}} = \frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{X}_i$. Let $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum\limits_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$ be the sample predictor variance, assuming a known $d$, the CUME estimator of $\mathcal{S}_{Y|\mathbf{X}}$ is constructed by the $d$ eigenvectors of $\hat{\mathbf{\Sigma}}^{-1}\mathbf{M}_n$ corresponding to its $d$ largest eigenvalues. Zhu et al. (2010) studied the asymptotic properties of the CUME estimator, and also provided a data-driven method on the determination of $d$.

### 2.2 The Ensemble Estimator Approach

In this section, we revisit CUME via the ensemble estimator approach (Yin and Li, 2011). The main result of Yin and Li (2011) is summarized by the following lemma.

**Lemma 1** *(Yin and Li, 2011) Let $\mathscr{J}$ be a family of functions $f : \Omega_Y \to \mathbb{F}$, where $\mathbb{F}$ can be the set of real or complex numbers; let $F_Y$ be the distribution function of $Y$, $L_2(F_Y)$ be the class of functions $f(Y)$ with finite variances and $(f_1, f_2) = \mathrm{E}[f_1(Y)f_2(Y)]$ as the inner product. Let $\mathcal{S}_{\mathrm{E}[f(Y)|\mathbf{X}]}$ be the central mean subspace for the conditional mean $\mathrm{E}[f(Y)|\mathbf{X}]$, as defined in Cook and Li (2002). If $\mathscr{J}$ is a subset of $L_2(F_Y)$ that is dense in $\mathscr{B}$, where $\mathscr{B} = \{I_B : B\, is\, a\, Borel\, set\, in\, \Omega_Y\}$, then we have:*

$$\mathrm{Span}\{\mathcal{S}_{[\mathrm{E}(f(Y)|\mathbf{X})]} : f \in \mathscr{J}\} = \mathcal{S}_{Y|\mathbf{X}}. \tag{2.3}$$

Hence, for a sufficiently rich family of $f(Y)$, the conditional mean subspaces $\mathcal{S}_{[\mathrm{E}(f(Y)|\mathbf{X})]}$, when put together, can recover the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. Yin and Li (2011) showed that both the Fourier transformation method proposed by Zhu and Zeng (2006), and the sliced regression (SR) proposed by Wang and Xia (2008) are special examples of the above ensemble estimators. Specifically, Zhu and Zeng (2006) used $\mathscr{J} = \{f_t(y) = e^{\imath ty} : t \in \mathbb{R}^1\}$, where $\imath$ is the imaginary unit; while Wang and Xia (2008) used $\mathscr{J} = \{f_t(y) = I_{(-\infty,t)}(y) : t \in \mathbb{R}^1\}$, and in practice taking those slicing points as the values for $t$, then use MAVE (Xia et al., 2002) to estimate the individual central mean subspace $\mathcal{S}_{\mathrm{E}[f_t(Y)|\mathbf{X}]}$.

Assuming the following two conditions which are commonly used in the sufficient dimension reduction literature (Cook, 2004):

(**a**) linearity condition: $E(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X}) = \mathbf{P}^T_{\boldsymbol{\eta}}(\mathbf{\Sigma})\mathbf{X}$, where $\mathbf{P}_{\boldsymbol{\eta}}(\mathbf{\Sigma}) = \boldsymbol{\eta}(\boldsymbol{\eta}^T\mathbf{\Sigma}\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^T\mathbf{\Sigma}$.

(**b**) coverage condition: $\mathrm{Span}(\mathbf{\Sigma}^{-1}\mathbf{m}(t)) = \mathcal{S}_{\mathrm{E}[f_t(Y)|\mathbf{X}]}$.

We are now ready to demonstrate that CUME also belongs to the family of the ensemble estimators. Let $\mathscr{J} = \{f_t(y) = I_{(-\infty,t)}(y) : t \in \mathbb{R}^1\}$. The population coefficient vector from the ordinary least squares fit of of $f_t(Y)$ on $\mathbf{X}$ is $\boldsymbol{\eta}_t = \mathbf{\Sigma}^{-1}\mathrm{E}(f_t(Y)\mathbf{X}) = \mathbf{\Sigma}^{-1}m(t)$ without the intercept term. Following the result of Duan and Li (1991), it is not hard to prove that, assuming condition (a), the above OLS coefficient falls into the central mean subspace $\mathcal{S}_{\mathrm{E}[f_t(Y)|\mathbf{X}]}$. The following proposition concludes our discussion of CUME in this section.

**Proposition 1** *Let $\mathscr{J} = \{f_t(y) = I_{(-\infty,t)}(y) : t \in \mathbb{R}^1\}$, then assuming conditions (a) and (b), $\mathbf{\Sigma}^{-1}m(t) \in \mathcal{S}_{\mathrm{E}[f_t(Y)|\mathbf{X}]}$, and the column space of $\hat{\mathbf{\Sigma}}^{-1}\mathbf{M}_n$ provides a consistent estimator of $\mathcal{S}_{Y|\mathbf{X}}$.*

According to the above proposition, we can see that CUME is also a special example of the family of the ensemble estimators.

## 3. Covariance Cumulative Slicing Estimation (COCUM)

### 3.1 The Method

As Cook and Ni (2006) pointed out, the use of $\boldsymbol{\eta}_t$ discards the intra-slice information which might result in a loss of power. Instead, in this section, we propose a method called the *covariance cumulative slicing estimation* (COCUM) which incorporates the intra-slice information into the estimation of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$.

To recover the intra-slice covariance information, following Cook and Ni (2006), we take

$$\mathbf{m}_c(\tilde{y}) = \mathrm{E}(\mathbf{X}YI(Y \leq \tilde{y})) \tag{3.1}$$

for $\tilde{y} \in \mathbb{R}^1$. The kernel matrix for COCUM, $\mathbf{M}_c$, is constructed similarly as (2.2) except replacing $\mathbf{m}(\tilde{y})$ with $\mathbf{m}_c(\tilde{y})$.

Denote $\boldsymbol{\beta}_t = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{X}YI(Y \leq t))$, it is easy to show that $\boldsymbol{\beta}_t$ can be decomposed as $F_t\boldsymbol{\Sigma}^{-1}\mathrm{Cov}(\mathbf{X}, Y|I(Y \leq t) = 1) + \mathrm{E}(Y|I(Y \leq t) = 1)\boldsymbol{\eta}_t$, where $F_t = P(Y \leq t)$.

**Theorem 1** *Assuming condition (a), then $\boldsymbol{\beta}_t \in \mathcal{S}_{Y|\mathbf{X}}$, for $t \in \mathbb{R}^1$. Furthermore, assuming condition (b) and a similar coverage condition* $\mathrm{Span}(\boldsymbol{\Sigma}^{-1}\mathbf{m}_c(t)) = S_{\mathrm{E}[g_t(Y)|\mathbf{X}]}$*, where $g_t(y) = yI(y \leq t)$, then, comparing with CUME, the column space of $\boldsymbol{\Sigma}^{-1}\mathbf{M}_c$ always encloses that of $\boldsymbol{\Sigma}^{-1}\mathbf{M}$.*

Theorem 1 suggests that theoretically COCUM always wins over CUME since it recovers more of the central subspace.

### 3.2 The Asymptotic Properties

The following theorem shows that COCUM possesses the same asymptotic property as CUME. Zhu, Miao and Peng (2006) derived the strong consistency for the slicing estimation of the SIR matrix when $p = o(n^{1/4})$. However, the results from both CUME and COCUM are better than Zhu, Miao and Peng (2006). The proof is similar to that of Theorem 2 of CUME (Zhu et al., 2010) and is omitted.

**Theorem 2** *Let $X_i$ be the $i^{th}$ coordinate of $\mathbf{X}$, suppose $max_{1 \leq i \leq p}E(X_i^8 Y^8) < \infty$ uniformly for $p$, and then*

$$||\boldsymbol{\Sigma}_n^{-1}\mathbf{M}_n^c - \boldsymbol{\Sigma}^{-1}\mathbf{M}_c|| = o(pn^{-1/2}\log n)$$

*almost surely where $||\cdot||$ is the Frobenius norm, where $\mathbf{M}_n^c = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_c(Y_i)\mathbf{m}_c^T(Y_i)w(Y_i)$.*

Let $\tilde{Y}$ be an independent copy of $Y$, $T(\mathbf{X}, \mathbf{Y}) = \boldsymbol{\Sigma}^{-1}\{\mathbf{X}\mathbf{X}^T - E\mathbf{X}\mathbf{X}^T - (\mathbf{X} - E\mathbf{X})E\mathbf{X}^T - E\mathbf{X}(\mathbf{X} - E\mathbf{X})^T\}\boldsymbol{\Sigma}^{-1}\mathbf{M}_c - \boldsymbol{\Sigma}^{-1}[2\mathbf{m}_c(Y)\mathbf{m}_c^T(Y)\omega(Y) + 2E\{\mathbf{X}YI(Y \leq \tilde{Y})\mathbf{m}_c^T(\tilde{Y})\omega(\tilde{Y})|\mathbf{X}, Y\} + 2E\{\mathbf{m}_c(\tilde{Y})\mathbf{X}^T Y^T I(Y \leq \tilde{Y})\omega(\tilde{Y})|\mathbf{X}, Y\} - 6\mathbf{M}_c]$. The following theorem states the asymptotic normality of our estimator which is similar to that of Theorem 3 of Zhu et al. (2010). The proof is also omitted.

**Theorem 3** *Assuming the following regularity conditions:*

1. *$max_{1 \leq i \leq p}E(X_i^8 Y^8) < \infty$ uniformly for $p$,*
2. *The minimum eigenvalue of $\boldsymbol{\Sigma}$ satisfies $\lambda_{min}(\boldsymbol{\Sigma}) > 0$,*
3. *The largest eigenvalue of $\mathbf{M}_c$ satisfies $\lambda_{max}(\mathbf{M}_c) < \infty$ holds uniformly for $p$,*

*4.* $E\{\gamma^T T(\mathbf{X}, Y)\gamma\} \to G > 0$ *for any unit length* $\gamma$,

*5.* $p = o(n^{1/2})$;

*then*

$$\sqrt{n}\gamma^T(\mathbf{\Sigma}_n^{-1}\mathbf{M}_n - \mathbf{\Sigma}^{-1}\mathbf{M})\gamma \xrightarrow{\mathcal{D}} N(0, G).$$

### 3.3 The Determination of $d$

One of the goals for sufficient dimension reduction is to estimate the structural dimension. Many methods have been developed to determine the structural dimension, such as Li (1991), Schott (1994), Bura and Cook (2001), Zhu et al. (2006), Zhu et al. (2010). Following Zhu et al. (2010), we use a modified BIC-type method for COCUM. Define

$$G(k) = n\sum_{i=1}^{k}\lambda_{ni}^2/\sum_{i=1}^{p}\lambda_{ni}^2 - C_n k(k+1)/2 \tag{3.2}$$

where $\hat{\lambda}_{n1} \geq \hat{\lambda}_{n2} \geq \ldots \geq \hat{\lambda}_{np}$ are the sample eigenvalues of the kernel matrix. And the estimated dimension $\hat{K}$ is defined as

$$\hat{K} = arg\ max_{1\leq k\leq p}G(k). \tag{3.3}$$

As pointed out in Zhu et al. (2010), smaller value of the penalty constant $C_n$ tends to overestimate the dimension $d$; while larger $C_n$ tends to underestimate the dimension $d$. A data-driven manner is needed to choose an appropriate value for $C_n$ under a certain method. We choose $C_n = \frac{1}{2}log(n)$ which mostly leads to satisfactory results. The following theorem states the consistency of $\hat{K}$. The proof is similar to that of Theorem 6 of Zhu et al. (2010) and is omitted.

**Theorem 4** *Assuming the conditions in Theorem 2, if* $\frac{C_n}{n} \to 0$ *and* $C_n \to \infty$, *as* $n \to \infty$, *the estimated structural dimension* $\hat{K}$ *obtained via* (3.3) *converges to the true structural dimension* $d$ *in probability.*

## 4. Simulation Studies

In this section, we compare the performance of COCUM with CUME. We considered two different models with the design matrix generated from normal and Gamma distributions. To evaluate the performance of different methods, following Li and Dong (2009), we use the ratio of the square multiple correlation coefficient to the dimension $d$:

$$\frac{\rho^2}{d} = \frac{trace\{(\hat{\boldsymbol{\beta}}^T\mathbf{\Sigma}\hat{\boldsymbol{\beta}})^{-1}(\hat{\boldsymbol{\beta}}^T\mathbf{\Sigma}\boldsymbol{\beta})(\boldsymbol{\beta}^T\mathbf{\Sigma}\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta})^{-1}\}}{d} \tag{4.1}$$

as evaluation measurements. Here $\hat{\boldsymbol{\beta}}$ is the estimate for true $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$ is the covariance matrix for predictor $\mathbf{X}$. $\rho^2$ gets closer to the true dimension $d$ if $\hat{\boldsymbol{\beta}}^T\mathbf{X}$ and $\boldsymbol{\beta}^T\mathbf{X}$ have a linear relation and it is 0 if they are uncorrelated. Hence, the closer the ratio is to 1, the better fit of the model. In our simulation studies, the average ratios from 1000 simulation runs are reported. In addition, the frequencies of the estimated structural dimension over 1000 trials are also given to illustrate the performance of the modified BIC method. For CUME, we follow the suggestion of Zhu et al. (2010) and take $C_n = 2n^{3/4}/p$. All numbers reported in the frequency table are multiplied by 10.

**Model I** We first consider a two dimensional model ($d = 2$) with predictors $\mathbf{X} = (X_1, X_2, X_3, \ldots, X_p)$ generated from $N(0, I_p)$, and standard normal $\epsilon$. The univariate response $Y$ is constructed as:

$$Y = 1.5(5 + X_1)(2 + X_2 + X_3) + 0.5\epsilon. \tag{4.2}$$

In this case, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ where $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^T$ and $\boldsymbol{\beta}_2 = (0, 1, 1, 0, \ldots, 0)^T$. As shown in Table 4.1, in terms of the average ratios, COCUM slightly outperforms CUME, while the performances of both methods deteriorate as $p$ gets larger. Table 4.2 suggests that CUME has serious problem of underestimating the structural dimension $d = 2$.

**Model II** We now consider a model with the predictor $\mathbf{X} = (X_1, X_2, X_3, \ldots, X_p)$ generated from $Gamma(0.1, 10)$, and standard normal $\epsilon$. The univariate response $Y$ is constructed as:

$$Y = (10 + (X_2 + 0.5)^2)\sqrt{(0.25 + X_1)} + 0.5\epsilon. \tag{4.3}$$

In this case, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ where $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^T$ and $\boldsymbol{\beta}_2 = (0, 1, 0, \ldots, 0)^T$. Table 4.3 provides the averages of the ratios and standard errors of the averages from 1000 simulation runs for Model II. It is obvious that COCUM wins over CUME. With larger sample size ($n = 400$ or 600), the average ratios are well above 0.9 even when $p = 20$. Table 4.4 again shows that CUME tends to underestimate the structural dimension, while COCUM does a good job most of the time.

## 5. Conclusion

Zhu et al. (2010) proposed a method called the *cumulative slicing estimation* (CUME) which sums up all possible estimations relating to $\mathrm{E}(\mathbf{X}I(Y \leq \tilde{y}))$ for all $\tilde{y}$ in the support of $Y$ to avoid the otherwise subjective selection of $h$. They showed that the estimator of CUME enjoys the common $\sqrt{n}$ convergence rate and is more efficient comparing to SIR and other first-moment slicing estimation methods. In this article, we revisit CUME from the ensemble estimator approach (Yin and Li, 2011) which helps us to gain more insights of CUME. We then propose a new method, which we call COCUM which incorporates the intra-slice information into the estimation of the central subspace. The asymptotic properties of COCUM are also discussed.

Our method can also be applied to multi-dimensional responses. For multi-response $\mathbf{Y}$, we first adapt the projective resampling method (Li, Wen and Zhu, 2008) to our proposed COCUM. Suppose $\mathbf{Y} \in \mathbb{R}^q$ and let $\mathbf{U}$ be a random vector uniformly distributed on the unit sphere $S^q$. Then we can define $\mathbf{m}(\mathbf{U}, y) = \mathrm{E}\{\mathbf{X}\mathbf{U}^T\mathbf{Y}I(\mathbf{U}^T\mathbf{Y} \leq \mathbf{U}^T y)\}$ and construct the candidate matrix $\mathbf{M} = \mathrm{E}\{\mathbf{m}(\mathbf{U}, \tilde{\mathbf{Y}})\mathbf{m}^T(\mathbf{U}, \tilde{\mathbf{Y}})\omega(\tilde{\mathbf{Y}})\}$. The column space spanned by $\boldsymbol{\Sigma}^{-1}\mathbf{M}$ is a subspace of $S_{\mathbf{Y}|\mathbf{X}}$.

We can further adapt the idea of COCUM to principle Hessian direction (Li, 1992) and define $\mathbf{m}(y) = \mathrm{E}(\mathbf{Z}\mathbf{Z}^T Y I(Y \leq y))$ and then we can adopt $\mathbf{M} = \mathrm{E}\{\mathbf{m}(\tilde{Y})\mathbf{m}^T(\tilde{Y})\omega(\tilde{Y})\}$ as the candidate matrix as a second order method, which is a potential competitor of CUME-SAVE or CUME-DR.

In our opinion, the ensemble estimator approach (Yin and Li, 2011) sheds lights on the developments of many existing sufficient dimension reduction methods such as SIR (Li, 1991), $K^{th}$ moment method (Yin and Cook, 2002), Fourier transformation method (Zhu and Zeng, 2006), SR (Wang and Xia, 2008). It is of special interest to study how this approach can be adopted to help developing new dimension reduction methods with functional data.

## Appendix

**Proof of Proposition 1**
Since $f_t(y) = I_{(-\infty,t)}(y)$, $\mathbf{\Sigma}^{-1}m(t)$ is the population coefficient vector from the ordinary least square fit of $f_t(Y)$ on $\mathbf{X}$ without the intercept term. By Theorem 1 of Cook and Li (2002), we can show that $\mathbf{\Sigma}^{-1}m(t) \in \mathcal{S}_{\mathrm{E}[f_t(Y)|\mathbf{X}]}$ assuming condition (a). Take one step further, assuming condition (b), we then have the equality, i.e., $\mathrm{Span}(\mathbf{\Sigma}^{-1}\mathbf{m}(t)) = \mathcal{S}_{\mathrm{E}[f_t(Y)|\mathbf{X}]}$.

Since $\mathscr{J} = \{f_t(y) = I_{(-\infty,t)}(y) : t \in \mathbb{R}^1\}$ is dense in $L_2(F_Y)$, by Lemma 1, we have $\mathrm{Span}\{\mathcal{S}_{[\mathrm{E}(f_t(Y)|\mathbf{X})]} : f \in \mathscr{J}\} = \mathcal{S}_{Y|\mathbf{X}}$. Let $\hat{\mathbf{m}}(t)$ denote the corresponding sample estimator of $\mathbf{m}(t)$, for $t = Y_1, \ldots, Y_n$. Hence $\hat{\mathbf{m}}(Y_i) = \mathbf{m}_n(Y_i)$. By Theorem 2.2 of Yin and Li (2011), we have $\mathrm{Span}(\mathbf{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{m}(Y_i)\mathbf{m}^T(Y_i)w(Y_i)) = \mathcal{S}_{Y|\mathbf{X}}$, a.s.. Since $\mathrm{Span}(\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{m}}(t))$ is a consistent estimator of $\mathrm{Span}(\mathbf{\Sigma}^{-1}\mathbf{m}(Y_i))$, for $t = Y_1, \ldots, Y_n$, $\mathrm{Span}(\hat{\mathbf{\Sigma}}^{-1}\mathbf{M}_n)$ is also a consistent estimator of $\mathcal{S}_{Y|\mathbf{X}}$. □

**Proof of Theorem 1**
Let $f_t(y) = I(y \leq t)$, and $g_t(y) = yI(y \leq t)$, then $\mathbf{\Sigma}^{-1}\mathbf{m}(t) = \mathbf{\Sigma}^{-1}\mathrm{Cov}\{f_t(y), \mathbf{X}\}$, and $\mathbf{\Sigma}^{-1}\mathbf{m}_c(t) = \mathbf{\Sigma}^{-1}\mathrm{Cov}\{g_t(y), \mathbf{X}\}$. Following similar arguments of Theorem 1 of Cook and Li (2002), we can show that $\mathbf{\Sigma}^{-1}\mathbf{m}(t) \in S_{f_t(Y)|X}$, and $\mathbf{\Sigma}^{-1}\mathbf{m}_c(t) \in S_{g_t(Y)|X}$. Assuming that $\mathrm{Span}(\mathbf{\Sigma}^{-1}\mathbf{m}(t)) = S_{f_t(Y)|X}$, and $\mathrm{Span}(\mathbf{\Sigma}^{-1}\mathbf{m}_c(t)) = S_{g_t(Y)|X}$

For any $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d}$,

$$g_t(y) \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\alpha}^T\mathbf{X} \Rightarrow f_t(y) \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\alpha}^T\mathbf{X},\ a.s.$$

which implies that

$$S_{f_t(Y)|\mathbf{X}} \subseteq S_{g_t(Y)|\mathbf{X}}.$$

By the structure of $\mathbf{M}$ and $\mathbf{M}_c$, it is straightforward that $\mathrm{Span}(\mathbf{\Sigma}^{-1}\mathbf{M}) \subseteq \mathrm{Span}(\mathbf{\Sigma}^{-1}\mathbf{M}_c)$.

Also, it is easy to show that both $\mathbf{\Sigma}^{-1}\mathbf{M}$ and $\mathbf{\Sigma}^{-1}\mathbf{M}_c$ are in $\mathcal{S}_{Y|\mathbf{X}}$. Hence, under very strong coverage conditions, $\mathrm{Span}(\mathbf{\Sigma}^{-1}\mathbf{M})$ recovers more of the central subspace. □

## REFERENCES

Bura, E. and Cook, R. D. (2001). Extending sliced inverse regression: The weighted chi-squared test. *Journal of the American Statistical Association*, **96**, 996–1003.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, **32**, 1062–1092.

Cook, R. D. (1998). *Regression Graphics.* Wiley, New York.

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

Cook, R. D. and Ni, L. (2006). Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika*, **93**, 65–74.

Cook, R. D. and Weisberg, S. (1991). Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 328–332.

Duan, N. and Li, K. C. (1991). Slicing Regression: A link-free regression method. *The Annals of Statistics*, **19**, 505–530.

Feng, Z., Wen, X., Yu, Z., and Zhu, L.-X. (2013). On partial sufficient dimension reduction with applications to partially linear multi-index models. *Journal of the American Statistical Association*, **108**, 237–246.

Härdle, W., Hall, P., and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum. *Journal of the American Statistical Association*, **83**, 86–95.

Li, B. and Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. *The Annals of Statistics*, **37**, 1272–1298.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, **102**, 997–1008.

Li, B., Wen, S. Q., and Zhu, L. X. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association*, **103**, 1177–1186.

Li, K. C. (1987). Asymptotic optimality for Cp, CL, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, **15**, 958–975.

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

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another applicationof Stein's lemma. *Journal of the American Statistical Association*, **87**, 1025–1039.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of American Statistical Association*, **89**, 141–148.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, **103**, 811–821.

Xia, Y., Tong, H., Li W. K., and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of Royal Statistical Society*, **64**, 363–410.

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional kth moment in regression. *Journal of Royal Statistical Society*, **64**, 159–175.

Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, **39**, 3392–3416.

Yin, X., Li, B., and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, **99**, 1733–1757.

Zhu, L., Zhu, L. X., and Feng, Z. (2010). Dimension reduction in regression through cumulative slicing estimation. *Journal of American Statistical Association*, **105**, 1455–1466.

Zhu, L.-X., Miao, B. Q., and Peng, H. (2006). Sliced Inverse Regression with large dimensional covariates. *Journal of American Statistical Association*, **101**, 630–643.

Zhu, L., Wang, T., Zhu, L. X. and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, **97**, 295–304.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of American Statistical Association*, **101**, 1638-1651.

**Table 4.1**: Ratio Average (standard error) for Model I

| | | $n = 200$ | | $n = 400$ | | $n = 600$ | |
|---|---|---|---|---|---|---|---|
| | Method | ratio | deviation | ratio | deviation | ratio | deviation |
| $p = 10$ | CUME | 0.71 | 0.22 | 0.80 | 0.18 | 0.85 | 0.14 |
| | COCUM | 0.75 | 0.19 | 0.84 | 0.15 | 0.88 | 0.11 |
| $p = 15$ | CUME | 0.66 | 0.19 | 0.74 | 0.17 | 0.79 | 0.14 |
| | COCUM | 0.69 | 0.18 | 0.78 | 0.14 | 0.83 | 0.12 |
| $p = 20$ | CUME | 0.63 | 0.16 | 0.71 | 0.16 | 0.76 | 0.14 |
| | COCUM | 0.64 | 0.16 | 0.74 | 0.14 | 0.79 | 0.12 |

**Table 4.2**: Dimension estimate for Model I

| $p$ | $n$ | $d = 1$ | | $d = 2$ | | $d > 2$ | |
|---|---|---|---|---|---|---|---|
| | | CUME | COCUM | CUME | COCUM | CUME | COCUM |
| 10 | 200 | 100 | 1 | 0 | 98.5 | 0 | 0.5 |
| 10 | 400 | 100 | 0.3 | 0 | 99.7 | 0 | 0 |
| 10 | 600 | 100 | 0.2 | 0 | 99.8 | 0 | 0 |
| 15 | 200 | 100 | 0 | 0 | 82.9 | 0 | 17.1 |
| 15 | 400 | 100 | 0 | 0 | 98.1 | 0 | 1.9 |
| 15 | 600 | 100 | 0 | 0 | 99.3 | 0 | 0.7 |
| 20 | 400 | 100 | 0 | 0 | 74 | 0 | 26 |
| 20 | 600 | 100 | 0 | 0 | 91 | 0 | 9 |

**Table 4.3**: Ratio Average (standard error) for Model II

| | | $n = 200$ | | $n = 400$ | | $n = 600$ | |
|---|---|---|---|---|---|---|---|
| | Method | ratio | deviation | ratio | deviation | ratio | deviation |
| $p = 10$ | CUME | 0.52 | 0.16 | 0.57 | 0.25 | 0.63 | 0.32 |
| | COCUM | 0.93 | 0.15 | 0.96 | 0.08 | 0.98 | 0.04 |
| $p = 15$ | CUME | 0.49 | 0.08 | 0.51 | 0.11 | 0.54 | 0.19 |
| | COCUM | 0.89 | 0.20 | 0.94 | 0.10 | 0.96 | 0.07 |
| $p = 20$ | CUME | 0.48 | 0.04 | 0.49 | 0.07 | 0.51 | 0.10 |
| | COCUM | 0.85 | 0.22 | 0.92 | 0.14 | 0.95 | 0.08 |

**Table 4.4**: Dimension estimates for Model II

| $p$ | $n$ | $d = 1$ | | $d = 2$ | | $d > 2$ | |
|---|---|---|---|---|---|---|---|
| | | CUME | COCUM | CUME | COCUM | CUME | COCUM |
| 10 | 200 | 100 | 8.3 | 0 | 91 | 0 | 0.7 |
| 10 | 400 | 100 | 1.8 | 0 | 97.2 | 0 | 1 |
| 10 | 600 | 100 | 0.3 | 0 | 99 | 0 | 0.7 |
| 15 | 200 | 100 | 3.6 | 0 | 90.1 | 0 | 6.3 |
| 15 | 400 | 100 | 0.8 | 0 | 95.4 | 0 | 3.8 |
| 15 | 600 | 100 | 0.1 | 0 | 95.9 | 0 | 4 |
| 20 | 400 | 100 | 0.4 | 0 | 91.2 | 0 | 8.4 |
| 20 | 600 | 100 | 0.1 | 0 | 94.3 | 0 | 5.6 |