Estimating the Central Kth Moment Space via an Extension of Ordinary Least Squares

Yuexiao Dong∗ Zhou Yu †

Abstract
Various sufficient dimension reduction methods have been proposed to find linear combinations of predictor $X$, which contain all the regression information of $Y$ versus $X$. If we are only interested in the partial information contained in the mean function or the $k$th moment function of $Y$ given $X$, estimation of the central mean space (CMS) or the central $k$th moment space (CKMS) becomes our focus. However, existing OLS-type estimators for CMS and CKMS require a linearity assumption on the predictor distribution. In this paper, we relax this stringent limitation via the notion of central solution space (CSS). Central $k$th moment solution space is introduced and its estimators are compared with existing methods by simulation.

Key Words: Central $k$th moment space, central solution space, dimension reduction subspace, non-elliptical distribution.

1. Introduction

Let $X$ be a $p$-dimensional random vector representing the predictor, and $Y$ be a random variable representing the response. The goal of dimension reduction (Li, 1991, 1992; Cook and Weisberg, 1991; Cook 1998) is to seek $\beta \in \mathbb{R}^{p \times d}$ ($d < p$), such that $Y \perp X | \beta^T X$. The column space of $\beta$ is called a dimension reduction subspace (DRS). Under very mild conditions (Yin et al., 2008), the intersection of two DRSs is still a DRS. The central space, or the smallest DRS, is the intersection of all DRSs, and is denoted by $S_{Y|X}$ (Cook, 1994, 1996). The dimension of $S_{Y|X}$ is called the structure dimension.

In many situations, regression analysis is mostly concerned about inferring the conditional mean of the response given the predictors. Central mean subspace (CMS; Cook and Li, 2002) is designed to address this kind of problem.

Definition 1 If $Y \perp E(Y|X) | \alpha^T X$, then $S(\alpha) = \text{span}(\alpha)$ is a mean dimension reduction subspace for the regression of $Y$ versus $X$. Let $S_{E(Y|X)} = \cap S_m$, where the intersection is over all mean dimension reduction subspaces $S_m$. If $S_{E(Y|X)}$ is itself a mean dimension reduction subspace, it is called the central mean subspace.

Because $Y \perp X | \alpha^T X$ implies $Y \perp E(Y|X) | \alpha^T X$, a DRS is also a mean dimension reduction subspace. It follows that $S_{E(Y|X)} \subset S_{Y|X}$, since the former is the intersection of at least the same, if not more, subspaces. Following the idea of the central mean space, central $k$th moment dimension reduction space (CKMS; Yin and Cook,

∗Temple University, Philadelphia, PA 19122, U.S.A
†Hong Kong Baptist University, Hong Kong, China
2002) is designed to aim dimension reduction at reducing the mean function, the variance function and up to the kth moment function, leaving the rest of regression $Y$ versus $X$ as the nuisance parameter. Let $M^{(k)}(Y|X) = E\{|Y - E(Y|X)|^k\}$ for $k \geq 2$ and $M^{(1)}(Y|X) = E(Y|X)$.

**Definition 2** If $Y \in \{M^{(1)}(Y|X), \ldots, M^{(k)}(Y|X)\}|\eta^T X$, then $S(\eta) = \text{span}(\eta)$ is a kth moment DRS for the regression of $Y$ versus $X$. Let $S^{(k)}_{Y|X} = \cap S^{(k)}$, where the intersection is over all kth moment DRSs $S^{(k)}$. If $S^{(k)}_{Y|X}$ is itself a kth moment DRS, it is called the central kth moment DRS, or CKMS for short.

By definition, CMS is a special case of CKMS with $k = 1$.

2. Estimators in CKMS and their limitations

In this section, we will study the existing OLS-type estimators in the CKMS and show their limitations. Let $\Sigma = \text{Var}(X)$ and $\text{span}(\beta) = S^{(k)}_{Y|X}$. Without loss of generality, we will assume $E(X) = 0$ and $E(Y) = 0$ in the following discussions. Yin and Cook (2002) proposed the following Theorem.

**Theorem 1** $\Sigma^{-1}E(XY^k) \subseteq S^{(k)}_{Y|X}$ if $E(X|\beta^T X)$ is linear in $\beta^T X$.

The linearity assumption above is known as the linear conditional mean (LCM) assumption, which is very common in the dimension reduction literature. A special case with $k = 1$ reduces to the ordinary least squares (OLS; Li and Duan, 1989) estimator $\beta_{\text{OLS}}$. For $k > 1$, we denote $\Sigma^{-1}E(XY^k)$ by $\beta_{\text{OLS}}^k$. Since $\beta$ is unknown, LCM is assumed to hold for all possible $\beta$ in practice. Therefore the distribution of $X$ has to be elliptically-contoured (Eaton, 1986). This is a serious limitation to the conventional CKMS estimators, as we can easily see from the following illustrative examples.

**Example 1.** $Y = g(X_1) + \epsilon$, $X = (X_1, X_2)^T \sim N(0, I_2)$, $\epsilon \sim N(0, 1)$ and $\epsilon \independent X$. In this model $S_{Y|X} = S^{(1)}_{Y|X} = S_{E(Y|X)} = (1, 0, 0)^T$. The OLS estimator thus satisfies

$$\beta_{\text{OLS}} = E^{-1}(XX^T)E(YY) = [E(X_1g(X_1), 0)]^T \subseteq S_{E(Y|X)} = S_{Y|X}.$$  

Now we consider a case when there is possible non-linearity between the predictors. Assume $X_1 \sim N(0, 1)$, $X_2 = h(X_1) + \delta$ and $\delta \independent (X_1, \epsilon)$. Now the OLS estimator becomes

$$\beta_{\text{OLS}} = E^{-1}(XX^T)E(YY) = c \begin{pmatrix} E(X_2^2) & -E(X_1X_2) \\ -E(X_1X_2) & E(X_1^2) \end{pmatrix} \begin{pmatrix} E[X_1g(X_1)] \\ E[X_2g(X_1)] \end{pmatrix},$$

where $c = \text{det}^{-1}[E(XX^T)]$ is a constant. Thus $\beta_{\text{OLS}} \subseteq S_{Y|X}$ if and only if $E[X_2g(X_1)]E(X_1^2) = E[X_1g(X_1)]E(X_1X_2)$. Plug in $X_2 = h(X_1) + \delta$ and we have

$$E[h(X_1)g(X_1)|E(X_1^2) = E[X_1g(X_1)]E[X_1h(X_1)].$$  

(1)
We can immediately see some special cases. For example, (1) is satisfied when \( h(X_1) \) is linear in \( X_1 \), which implies OLS works for elliptical \( X \). The legitimacy of (1) can be easily checked when the link function \( g(X_1) \) is linear in \( X_1 \). This implies OLS is robust to non-ellipticity in \( X \) when the link function is linear in \( \beta^T X \), where \( \text{span}(\beta) = S_{Y|X} \). Some other scenarios can be constructed for condition (1) to be true. However, generally speaking, OLS can not successfully recover the central space when there is non-ellipticity in \( X \).

**Example 2.** \( Y = X_1 + X_1 X_2 + \epsilon \), \( X = (X_1, X_2, X_3)^T \sim N(0, I_3) \), \( \epsilon \sim N(0, 1) \) and \( \epsilon \perp \perp X \). This example has been examined by Yin and Cook (2002). OLS alone is not able to fully recover the central space since \( \beta_{OLS} = E(XX^T) = (1, 0, 0)^T \). They suggested to include \( \beta_{OLS2} = E(XX^T) = (0, 2, 0)^T \) to exhaustively estimate \( S_{Y|X} \). However, if we introduce non-ellipticity in \( X \), their proposed method will fail. For example, let \( (X_1, X_2)^T \sim N(0, I_2) \), \( X_3 = X_1^2 + X_1 - 1 + \delta \) and \( \delta \perp \perp (X_1, X_2, \epsilon) \). Then while \( \beta_{OLS} = E^{-1}(XX^T)E(XY) = (1, 0, 0)^T \) is still unbiased, \( \beta_{OLS2} = E^{-1}(XX^T)E(XY^2) = (-4/3, 2, 4/3)^T \) no longer belongs to the central space.

**Example 3.** We keep all setup the same as in Example 2 except modifying the link function to be \( Y = X_1 + 1 + (X_2 + 1)\epsilon \). In the normal case, simple calculations show \( \beta_{OLS} = (1, 0, 0)^T \) and \( \beta_{OLS2} = (2, 2, 0)^T \). Together they fully recover \( S_{Y|X} \). In the non-elliptical case, we have \( \beta_{OLS} = (1, 0, 0)^T \). However, \( \beta_{OLS2} = (4/3, 2, 2/3)^T \) is biased.

From the above examples, we can see that when the LCM assumption holds, \( \Sigma^{-1}E(XY^k) \) help recover either the central mean space (in Example 2) or the higher order CKMS (in Example 3). However, these estimators, like OLS, will fail in the presence of non-elliptical predictors.

### 3. Estimators in CKMS without LCM

Classical CKMS estimators require the LCM assumption for the predictor. To loosen this restrictive linearity assumption, we propose the central \( k \)th moment solution space (CKMSS). The idea of using estimating equations for sufficient dimension reduction is originated in Li and Dong (CSS: 2009), where first order dimension reduction methods are generalized without the LCM assumption. Second-order CSS methods are discussed in Dong and Li (2010).

#### 3.1 Central mean solution space

To estimate the central mean space without the LCM assumption, we introduce an estimating equation scheme. Let’s focus on the following equation,

\[
E(YX) = E[YE(X|\beta^TX)] \quad \text{a.s..} \tag{2}
\]

**Definition 3** The central mean solution space is defined to be \( S_{CMSS} = \cap \text{span}(\beta) \), where the intersection is over all \( \beta \) that satisfies equation (2).
In Li and Dong (2009), it was shown that $S_{\text{CMSS}} \subseteq S_{Y|X}$. In the context of CKMS, we know more precisely that $S_{\text{CMSS}} \subseteq S_{E(Y|X)}$. Because equation (2) targets at the conditional independence between $E(Y|X)$ and $Y$ directly, it can be used to estimate the central mean space for non-elliptically distributed predictors. Denote OLS estimator by $S_{\text{OLS}} = \Sigma^{-1}E(YX)$. Then we have

**Theorem 2** (i) $S_{\text{CMSS}} \subseteq S_{E(Y|X)}$.
(ii) If LCM assumption holds, then $S_{\text{CMSS}} = S_{\text{OLS}} \subseteq S_{E(Y|X)}$.

This theorem tells us that the CMSS belongs to the CMS without the LCM assumption. CMSS becomes the conventional OLS with the LCM assumption.

### 3.2 Central $k$th moment solution space

Central mean space is just a special case of the central $k$th moment space with $k = 1$. The generalization to $k > 1$ cases is straightforward. Let $Y_k = (Y, Y^2, \ldots, Y^k)^T$ and $U_{\beta}(X) = E(X|\beta^T X)$. Denote $\otimes$ as Kronecker product and we are now interested in the following equation,

$$E\{|Y_k - E(Y_k)| \otimes X\} = E\{|Y_k - E(Y_k)| \otimes U_{\beta}(X)\} \quad \text{a.s.}$$

**Definition 4** The central $k$th moment solution space is defined to be $S^{(k)}_{\text{CKMSS}} = \cap \text{span} (\beta)$, where the intersection is over all $\beta$ that satisfies equation (3).

In Yin and Cook (2002), the $k$th order covariance subspace $S^{(k)}_{\text{COV}} = S(K)$ is defined as the column space of $K = \Sigma^{-1}E(|Y_k - E(Y_k)|^T \otimes X)$. The following theorem reveals the relations between the $k$th order covariance subspace $S^{(k)}_{\text{COV}}$, the central $k$th moment solution space $S^{(k)}_{\text{CKMSS}}$ and the central $k$th moment space $S^{(k)}_{Y|X}$.

**Theorem 3** (i) $S^{(k)}_{\text{CMSS}} \subseteq S^{(k)}_{Y|X}$.
(ii) If LCM assumption holds, then $S^{(k)}_{\text{CMSS}} = S^{(k)}_{\text{COV}} \subseteq S^{(k)}_{Y|X}$.

### 3.3 Sample estimation for CKMSS

The estimation equations (2) and (3) described previously all have the same form of $g(\beta) = 0$, where $g(\beta)$ is a $p \times p$ random matrix. This is equivalent to $E[\|g(\beta)\|^2] = 0$, where $\| \cdot \|$ is the Frobenious matrix norm. Given the structure dimension $d$ the central solution space, Li and Dong (2009) have shown that the minimizer of $L(\eta) = E[\|g(\eta)\|^2]$, where $\eta \in \mathbb{R}^{p \times d}$ spans the corresponding CSS under a very general identifiable condition.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an independent sample of $(X, Y)$. For a function $r(X, Y)$, let $E_n[r(X, Y)]$ denote the sample average $n^{-1} \sum_{i=1}^n r(X_i, Y_i)$. Center $X_i$ and $Y_i$ to be $\tilde{X}_i = X_i - E_n X$ and $\tilde{Y}_i = Y_i - E_n Y$. Then, for example, the OLS estimator for CMS is $E_n^{-1}(\tilde{X}\tilde{X}^T)E_n(\tilde{X}\tilde{Y})$. For the sample estimation of CSS based estimators, we need to select basis functions $G(\eta^T X) = \{f_1(\eta^T X), \ldots, f_s(\eta^T X)\}$ that are sufficiently flexible to describe $E(X|\eta^T X)$. Dong and Li (2010) discussed
different choice of $G(\eta^T X)$. In our simulation, we use cubic polynomial functions of $\eta_i^T X, i = 1, \ldots, d$, where $\eta_i$’s are columns of $\eta$. Take CMSS as an example, we minimize the following objective function

$$L_n(\eta) = E_n \| \hat{Y} - \hat{f}(\eta^T \hat{X}) \|^2,$$

where

$$\hat{f}(\eta^T \hat{X}) = E_n(\hat{X}|\eta^T \hat{X}) = E_n \{ \hat{X} G(\eta^T \hat{X}) \}[E_n \{ G(\eta^T \hat{X}) G^T(\eta^T \hat{X}) \}^{-1} G(\eta^T \hat{X})].$$

Many computer programs are available for minimizing $L_n(\eta)$. We use the OPTIM function in R. Such numerical minimization needs an initial value of $\eta$, for which we choose outer product gradient (OPG; Xia et al., 2002). OPG estimators do not require the LCM assumption and works reasonably well as initial values in our simulation.

4. Simulation studies

We study the following four models: (I) $Y = \exp(X_1) + 0.5 \epsilon$, (II) $Y = (X_1 + 0.5)^3 + 0.5 \epsilon$, (III) $Y = (X_1 + 3)(X_2 + 1) + 0.5 \epsilon$, and (IV) $Y = (X_2 + 3) + \exp(X_1 + 1) \epsilon$. Here $\epsilon_i \sim t_3$, $\epsilon \sim N(0,1)$, and they are independent of $X$. We introduce nonlinearity in the predictor as follows: $X_1 \sim N(0, 1)$, $X_2 \sim N(0, 1)$, $X_3 = X_1 + 2X_1^2 + 0.5 \delta$, where $\delta \sim N(0, 1)$ and $\delta \perp (X, Y)$. When $j > 3$, $X_j$ are taken to be independent $N(0,1)$, and to be independent of $(X_1, \ldots, X_3)$. We fix $p = 4$ and study sample sizes $n = 50, 100, 200, 300$. Model (I) and (II) are 1-dimensional, while model (III) and (IV) are 2-dimensional.

Table 1: Average and its standard error of $r^2(\beta^T X, \hat{\beta}^T X)$ based on 100 repetitions. OPG: outer product gradient; OLS: ordinary least squares; CMSS: central mean solution space; COV2: covariance subspace estimator with $k = 2$; CKMSS2: central $k$th moment solutions space with $k = 2$. $X$ is non-elliptical with $p = 4$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>$n = 50$</th>
<th>$n = 100$</th>
<th>$n = 200$</th>
<th>$n = 300$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>OPG</td>
<td>0.616 (.030)</td>
<td>0.807 (.023)</td>
<td>0.913 (.010)</td>
<td>0.953 (.005)</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>0.723 (.010)</td>
<td>0.719 (.008)</td>
<td>0.691 (.007)</td>
<td>0.682 (.007)</td>
</tr>
<tr>
<td></td>
<td>CMSS</td>
<td>0.840 (.022)</td>
<td>0.923 (.013)</td>
<td>0.947 (.006)</td>
<td>0.958 (.004)</td>
</tr>
<tr>
<td>II</td>
<td>OPG</td>
<td>0.695 (.031)</td>
<td>0.894 (.016)</td>
<td>0.971 (.004)</td>
<td>0.987 (.002)</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>0.772 (.011)</td>
<td>0.781 (.012)</td>
<td>0.775 (.010)</td>
<td>0.770 (.009)</td>
</tr>
<tr>
<td></td>
<td>CMSS</td>
<td>0.844 (.033)</td>
<td>0.987 (.010)</td>
<td>0.999 (.000)</td>
<td>0.999 (.000)</td>
</tr>
<tr>
<td>III</td>
<td>OPG</td>
<td>0.774 (.031)</td>
<td>0.878 (.023)</td>
<td>0.962 (.008)</td>
<td>0.977 (.005)</td>
</tr>
<tr>
<td></td>
<td>COV2</td>
<td>0.773 (.023)</td>
<td>0.842 (.020)</td>
<td>0.889 (.014)</td>
<td>0.883 (.014)</td>
</tr>
<tr>
<td></td>
<td>CKMSS2</td>
<td>0.802 (.037)</td>
<td>0.906 (.025)</td>
<td>0.980 (.005)</td>
<td>0.986 (.003)</td>
</tr>
<tr>
<td>IV</td>
<td>OPG</td>
<td>0.520 (.026)</td>
<td>0.544 (.021)</td>
<td>0.583 (.029)</td>
<td>0.586 (.021)</td>
</tr>
<tr>
<td></td>
<td>COV2</td>
<td>0.621 (.033)</td>
<td>0.622 (.028)</td>
<td>0.640 (.027)</td>
<td>0.659 (.024)</td>
</tr>
<tr>
<td></td>
<td>CKMSS2</td>
<td>0.623 (.032)</td>
<td>0.678 (.023)</td>
<td>0.729 (.032)</td>
<td>0.741 (.027)</td>
</tr>
</tbody>
</table>
To compare accuracy of different OLS-type estimators, we use sample estimate of the squared vector correlation coefficient (Hotelling, 1936) between \(\beta^T X\) and \(\hat{\beta}^T X\), and denote it by \(r^2(\beta^T X, \hat{\beta}^T X)\). The closer the correlation \(r^2(\beta^T X, \hat{\beta}^T X)\) is to 1, the better the estimators are. Each entry of Table 1 is formatted as \(a(b)\), where \(a\) is the average of the correlation across the 100 simulated samples, and \(b\) is the standard error of this average. Across all 4 models, we can clearly see the improvement of CMSS and CKMSS over conventional OLS and COV estimators. The improvement is most significant when the sample size is large. While CSS based methods are consistent and get better with larger sample size, conventional OLS and COV are biased and won’t necessarily get better. OPG estimators do not require the LCM assumption and actually work decently for model (I), (II) and (III). For these three models, CSS based estimators use OPG as initial values and always yield better estimation than OPG. For model (IV), \(S_{Y|X} = S_{Y|Y}^{(2)}\). OPG fails in this case since it can not estimate the direction in the variance component. Numerical minimization is sensitive to the choice of initial values in this case. Thus we use the COV estimators as initial values and still observe significant improvement.

5. An empirical study

In this section we study the 1985 automobile data set, which can be downloaded from the UCI machine learning repository. After removing the categorical variables and missing values from the original data, the remaining data contains 195 observations. We now have 14 continuous variables, which are Wheelbase, Length, Width, Height, Curb Weight, Engine size, Bore, Stroke, Compression ratio, Horsepower, Peak rpm, City mpg, Highway mpg and Price. To study the factors that affect the prices of automobiles, we take the logarithm of Price to be the response and the other 13 variables to be the predictors.

Figure 1: On the left is the scatter plot of \(Y\) versus the first CKMSS predictor, and on the right is \(Y\) versus the second CKMSS predictor.

Zhu and Zeng (2006) analyzed this data set by Fourier method, and we use their estimator as our benchmark. We standardize the predictor by its mean and
standard deviation and work on this standardized $X$. The structure dimension was determined to be 2 and the Fourier estimators are $\hat{\beta}_T^1 = (0.05, -0.19, 0.08, 0.09, 0.75, -0.24, 0.00, -0.11, 0.17, 0.51, 0.04, -0.08, 0.12)$ and $\hat{\beta}_T^2 = (0.08, -0.38, 0.09, 0.08, 0.03, 0.70, -0.17, -0.20, -0.06, -0.08, 0.06, 0.49, -0.17)$ respectively. It is reasonable to assume that $S_{Y|X} = S_{Y|X}^{(2)}$ and we first try to estimate $S_{Y|X}^{(2)}$ by COV and denote the corresponding estimator by $\hat{\beta}_{\text{COV}} = (\hat{\beta}_{\text{COV}1}, \hat{\beta}_{\text{COV}2})$. The Pearson correlation between $\hat{\beta}_{\text{COV}1}^T X$ and $\hat{\beta}_{T}^1 X$ is 0.961, and the correlation between $\hat{\beta}_{\text{COV}2}^T X$ and $\hat{\beta}_{T}^2 X$ is 0.960. The scatter plots (not shown here) of the standardized non-linear trend, which means CKMSS estimators should be similar to CKMS estimators according to Theorem 3. Next we minimize $L_n(\eta) = E_n \left( |(Y, Y^2) - E_n(Y, Y^2)| \otimes |X - E_n(X|\eta^T X)| \right)^2$ over $\eta \in \mathbb{R}^{13 \times 2}$, where $E_n(X|\eta^T X)$ is estimated by cubic polynomials of $\eta^T X$. We perform numeric minimization with $\hat{\beta}_{\text{COV}}$ as the initial value. The resulting estimator is $\hat{\beta}_{\text{CKMSS}} = (\hat{\beta}_{\text{CKMSS}1}, \hat{\beta}_{\text{CKMSS}2})$.

Now the Pearson correlation between $\hat{\beta}_{\text{CKMSS}1}^T X$ and $\hat{\beta}_{T}^1 X$ is 0.965, and the correlation between $\hat{\beta}_{\text{CKMSS}2}^T X$ and $\hat{\beta}_{T}^2 X$ becomes 0.940. The perspective plots of the response versus two predictors are shown in Figure 1, where the first direction shows a strong linear trend and the second direction shows a parabolic relationship. For detailed examination of the linear trend as well as the parabolic trend, please refer to Zhu and Zeng (2006). Zhu and Zeng made similar findings as ours by Fourier methods, and the CKMSS approach works well at detecting relevant trend for this particular data set.

REFERENCES


