Fourier Methods for Sufficient Dimension Reduction without Distributional Assumptions

Michael Y. Zhu and Peng Zeng

Department of Statistics
Purdue University

August 11, 2005
Outline

1. Generalized multiple index model
2. Central mean subspace and central subspace
3. Fourier method for estimating central subspace
4. Asymptotics
5. Implementation and example
Generalized Multiple Index Model

Let $Y$ be a univariate response and $X$ a $p$-dimensional predictor vector

- Linear model: $E(Y|X) = \alpha^\top X$.
- Single index model: $E(Y|X) = g(\alpha^\top X)$.
- Multiple index model: $E(Y|X) = g(\alpha_1^\top X, \ldots, \alpha_k^\top X)$
- Generalized multiple index model (Li, 1991)

$$Y = h(\beta_1^\top X, \ldots, \beta_q^\top X, \epsilon)$$

Goal: Identify the dimension reduction subspace (DRS) spanned by $\beta_1, \ldots, \beta_q$ denoted by $\mathcal{S}(\beta_1, \ldots, \beta_q)$. 
Central Subspace

Let $\mathbf{B} = (\beta_1, \ldots, \beta_q)$, equivalently,

- Conditional independence, $Y \perp X | \mathbf{B}^T \mathbf{X}$ (Cook, 1996)
- Conditional distribution, $F(Y | X) = F(Y | \mathbf{B}^T \mathbf{X})$
- The minimal dimension reduction subspace is called Central Subspace (Cook, 1996)

$$S_{Y|X} = \bigcap_{\text{all DRS}} S$$
Central Mean Subspace

- Multiple index model (for mean response):
  \[ E(Y|X) = g(\alpha_1^T X, \ldots, \alpha_k^T X) \]
  Let \( A = (\alpha_1, \alpha_2, \ldots, \alpha_k) \).
- Equivalently
  \[ Y \perp E[Y | X] | A^T X \quad \text{(Cook and Li 2002)} \]
  \( S(A) \) is called mean dimension reduction (MDRS) subspace
- Central mean subspace (Cook and Li 2002)
  \[ S_{E[Y|X]} = \bigcap_{\text{all MDRS}} S \]
- \( S_{E[Y|X]} \subset S_{Y|X} \).
Example

Suppose \( \mathbf{X} = (X_1, \ldots, X_5)^\tau \in \mathbb{R}^5 \). Consider model

\[
Y = X_1 + (X_1 + X_3)^2 + \varepsilon X_4
\]

\[
= g(\beta_1^T \mathbf{X}, \beta_2^T \mathbf{X}) + \varepsilon h(\beta_3^T \mathbf{X})
\]

\[
E[Y | \mathbf{X}] = X_1 + (X_1 + X_3)^2
\]

\[
= g(\beta_1^T \mathbf{X}, \beta_2^T \mathbf{X})
\]

Therefore,

\[
\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\beta_1, \beta_2, \beta_3\}
\]

\[
\mathcal{S}_{E[Y|\mathbf{X}]} = \text{span}\{\beta_1, \beta_2\}
\]

where \( \varepsilon \perp \mathbf{X} \) and \( E[\varepsilon] = 0 \),

\[
\beta_1 = (1, 0, 0, 0, 0)^\tau
\]

\[
\beta_2 = (1, 0, 1, 0, 0)^\tau
\]

\[
\beta_3 = (0, 0, 0, 1, 0)^\tau
\]
**Brief Comment on Existing Methods**

- **Nonparametric methods:** only estimate central mean subspace and require estimation of link function and/or derivatives

- **Link-free methods:**
  - Principal Hessian Direction (Li, 1992), Iterative Hessian Transformation (Cook and Li, 2002), etc. for central mean subspace
  - Sliced Inverse Regression (Li, 1991), Sliced Average Variance Estimate (Cook and Weisberg, 1991), etc. for central subspace

Require distributional assumptions: linearity assumption and/or constant variance assumption

Do not guarantee to recover the space exhaustively.
General Procedure for Link Free Methods

- **Key step:**
  - Find a candidate matrix \( M \) depending on \( X \) and \( Y \), such that

\[
S(M) \subseteq S_{Y|X} \quad (\text{or } S_{E[Y|X]})
\]

- Given a sample \((x_i, y_i), i = 1, 2, \ldots, n,\)
  1. Find an estimate \( \hat{M} \) of \( M \).
  2. Perform spectral decomposition of \( \hat{M} \).
  3. Estimate CS (or CMS) by the space spanned by the eigenvectors of \( \hat{M} \) corresponding to the largest \( q \) eigenvalues.
Fourier Method for Central Mean Subspace

**Heuristics:** Find some vectors that belong to CMS, and let them span the whole CMS.

- $m(x) = E(Y|X = x)$ is a function of $u = A^T x$, then
  \[
  \frac{\partial m}{\partial x}(x) = A \frac{\partial g}{\partial u}(u) \in S_{E[Y|X]}
  \]

- For any $\omega \in \mathbb{R}^p$,
  \[
  \psi(\omega) = \int \exp\{\iota \omega^T x\} \frac{\partial m}{\partial x}(x) f_X(x) \, dx
  \]
  \[
  = -E_{(X,Y)}[Y(\iota \omega + G(X)) \exp\{\iota \omega^T X\}] \in S_{E[Y|X]}
  \]
  where $f_X$ is the density function of $X$, and $G(x) = \frac{\partial}{\partial x} \log f_X(x)$.
Fourier Method for Central Mean Subspace

Because $\psi(\omega) \in \mathcal{S}_{E[Y|X]}$, then

$$S(\psi(\omega)\bar{\psi}(\omega)^\tau) \subseteq \mathcal{S}_{E[Y|X]}$$

for any $\omega \in \mathbb{R}^p$

**Theorem 1.** Define matrix

$$M_{FM} = \text{Re} \int \psi(\omega)\bar{\psi}(\omega)^\tau K(\omega) \, d\omega$$

where $K(\omega)$ is a positive weight function on $\mathbb{R}^p$. Then $M_{FM}$ is nonnegative definite, and

$$S(M_{FM}) = \mathcal{S}_{E[Y|X]}.$$
Heuristics for Estimating Central Subspace

Relationship between central subspace and central mean subspaces. $T(Y)$ and $G(Y)$ are two transformations of $Y$.

It is possible to estimate $S_{Y|X}$ by all possible central mean subspaces.

$$S_{Y|X} = \sum_{all\ possible\ T} S_{E[T(Y)|X]}$$
Represent CS in Terms of CMSs

- A family of transformations,

\[ T(y, t) = \exp\{i\, ty\} = \cos(ty) + i\sin(ty), \quad t \in \mathbb{R}. \]

- \( m(x, t) \) is the Fourier transform (characteristic function) of \( f_{Y|X} \).

\[ m(x, t) = E[T(Y, t) \mid X = x] = \int \exp\{i\, ty\} f_{Y|X}(y \mid x) \, dy. \]

**Lemma 1.** CS can be represented as the sum of a family of CMSs.

\[ S_{Y|X} = \sum_{t \in \mathbb{R}} S_{E[T(Y,t)|X]} \]
Fourier Method for Central Subspace

Define

\[ \phi(\omega, t) = \int \exp\{i \omega^\tau x\} \frac{\partial m}{\partial x} (x, t) f_x(x) \, dx \]

\[ = -E(x,Y)[(i\omega + G(X)) \exp\{itY + i \omega^\tau X\}] \in S_{Y|x} \]

It is obtained by substituting \( Y \) in \( \psi(\omega) \) by \( \exp\{itY\} \).

**Theorem 2.** Define matrix

\[ M_{FC} = \text{Re} \int \int \phi(\omega, t) \overline{\phi}(\omega, t)^\tau K(\omega)k(t) \, d\omega dt \]

where \( K(\omega) \) and \( k(t) \) are positive weight functions. Then \( M_{FC} \) is nonnegative definite, and

\[ S(M_{FC}) = S_{Y|x} \]
When $K(\omega)$ and $k(t)$ are Gaussian Functions

- When $k(t) = (2\pi\sigma_t^2)^{-1/2} \exp\{-t^2/2\sigma_t^2\}$, and
  $K(\omega) = (2\pi\sigma_\omega^2)^{-p/2} \exp\{-\|\omega\|^2/2\sigma_\omega^2\}$.

\[
M_{FC} = E\left[a_{12} \left[ \sigma_\omega^2 I_p + (G(U_1) - \sigma_\omega^2 U_{12})(G(U_2) + \sigma_\omega^2 U_{12})^\top \right] \right]
\]

where $a_{12} = \exp\{-\sigma_t^2(V_1 - V_2)^2/2 - \sigma_\omega^2\|U_1 - U_2\|^2/2\}$ and $(U_1, V_1)$ and $(U_2, V_2)$ are iid as $(X, Y)$.

- $\sigma_\omega^2$ and $\sigma_t^2$ are tuning parameters (constants).
  - They are different from bandwidth in kernel estimation.
  - Theorem 2 is valid for any $\sigma_\omega^2$ and $\sigma_t^2$.

- Other weight functions can also be used.
Estimation of $M_{FC}$

Given a sample $(x_i, y_i), \ i = 1, 2, \ldots, n$, $M_{FC}$ can be estimated by sample average,

$$\hat{M}_{FC} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \left[ \sigma^2 \mathbf{I}_p + (G(x_i) - \sigma^2 x_{ij})(G(x_j) + \sigma^2 x_{ij})^T \right]$$

and the only unknown component is

$$G(x_i) = \frac{\partial}{\partial x} \log f_{\mathbf{X}}(x_i) = \frac{\partial}{\partial x} \frac{f_{\mathbf{X}}(x_i)}{f_{\mathbf{X}}(x_i)}$$
Pugging in Kernel Density Estimate

- Estimate \( G(x_i) \) by plugging in kernel estimate

\[
\hat{G}(x_i) = \frac{\partial}{\partial x} \hat{f}_h(x_i) \\
\hat{f}_h(x_i) = \frac{1}{nh^p} \sum_{\ell=1}^{n} W \left( \frac{x_i - x_\ell}{h} \right)
\]

where

\[
\hat{f}_h(x_i) = \frac{1}{nh^p+1} \sum_{\ell=1}^{n} W' \left( \frac{x_i - x_\ell}{h} \right)
\]

and \( W(\cdot) \) is a kernel function, \( W'(\cdot) \) is the derivative of \( W(\cdot) \), and \( h \) is the bandwidth.
Final Estimate

We have an estimate

\[
\hat{M}_{\text{FCk}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \left[ \sigma_\omega^2 I_p + (\hat{G}(x_i) - \sigma_\omega^2 x_{ij})(\hat{G}(x_j) + \sigma_\omega^2 x_{ij})^\tau \right] \hat{I}_i \hat{I}_j
\]

where \( a_{ij} = \exp\{-\sigma_i^2 y_{ij}^2 / 2 - \sigma_\omega^2 x_{ij}^\tau x_{ij} / 2\} \), \( x_{ij} = x_i - x_j \), and \( y_{ij} = y_i - y_j \), \( \hat{I}_i = I[\hat{f}_h(x_i) > b_n] \), \( I[.] \) is an indicator function, and \( b_n \) is a threshold.

The technique of using \( I[.] \) is called trimming. Its purpose is to trim the points whose estimated densities are extremely small.
Asymptotic Result

**Theorem 3.** Under some regularity conditions, if $f_{\mathbf{X}}(\mathbf{x})$ has partial derivatives up to order $r \geq p + 2$, and

1. $n \to \infty$, $h \to 0$, $b \to 0$ and $b^{-1}h \to 0$;
2. for some $\varepsilon > 0$, $b^4 n^{1-\varepsilon} h^{2p+2} \to \infty$;
3. $nh^{2r-2} \to 0$

then

$$\sqrt{n} (\text{vec}(\widehat{\mathbf{M}}_{Fck}) - \text{vec}(\mathbf{M}_{FC})) \overset{\mathcal{L}}{\to} N(0, \Sigma)$$

where $\Sigma$ is a positive definite matrix.
General Procedure for Estimating Subspaces

Suppose we have observations \((x_i, y_i), i = 1, \ldots, n\).

1. Specify parameters: \(q, \sigma_w^2 = 0.1, \sigma_t^2 = 1.0, h,\) and \(b_n\), if applicable.

2. Standardize data by \(\tilde{x}_i = \hat{\Sigma}^{-1/2} (x_i - \bar{x})\) and \(\tilde{y}_i = (y_i - \bar{y})/s_y\)

3. Calculate an estimate \(\hat{M}\) of \(M_{FC}\) (or \(M_{FM}\)) using data \((\tilde{x}_i, \tilde{y}_i)\).

4. Perform spectral decomposition of \(\hat{M}\). The eigenvalues are \(\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0\), and their corresponding eigenvectors are \(\hat{e}_1, \ldots, \hat{e}_p\).

5. Estimate \(S_{Y|x}\) (or \(S_{E[Y|x]}\)) by \(\hat{S} = \text{span}\{\hat{\Sigma}^{-1/2} \hat{e}_1, \ldots, \hat{\Sigma}^{-1/2} \hat{e}_q\}\).
Simulation Example

Assume \( X \in \mathbb{R}^5 \),

\[
Y = \frac{b_1^T X}{3 + (2 + b_2^T X)^2} + 0.2 \epsilon
\]

where \( b_1 = (1, 1, 0, 0, 0) \), \( b_2 = (0, 0, 0, 1, 1) \), \( \epsilon \) is a random error, and \( X \) follows a mixture of multivariate distributions

\[
X \sim 0.4N(a_1, I_5) + 0.6N(a_2, I_5)
\]

Clearly, \( S_{Y|X} = S(b_1, b_2) \).

- A random sample of 250 observations \( \{(y_i, x_i)\}_{i=1}^{250} \) is generated.

- Plots of \( y_i \) versus \( b_1^T x_i \) and \( b_2^T x_i \):
**Estimated Directions**

- Estimate $M_{FC}$ using $\sigma_\omega^2 = 0.1$, $\sigma_t^2 = 1.0$ and $h = 1$.

- Obtain the first two eigenvectors of $M_{FC}$ and use them to span a space as the estimate of $S_{Y|X}$.

- Plots of $y_i$ versus the estimated directions
August 2005

y

-1.0 -0.5 0.0 0.5 1.0

-4 -2 0 2

first direction

y

-1.0 -0.5 0.0 0.5 1.0

-2 -1 0 1 2 3

second direction