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 $\mathbf{X}$ a p-dimensional predictor vector

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

$$Y = h(\beta_1^\tau \mathbf{X}, \ldots, \beta_q^\tau \mathbf{X}, \epsilon)$$

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 $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$, equivalently,

- Conditional indepence, $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^\tau \mathbf{X}$ (Cook, 1996)
- Conditional distribution, $F(Y \mid \mathbf{X}) = F(Y \mid \mathbf{B}^\tau \mathbf{X})$
- The minimal dimension reduction subspace is called **Central Subspace** (Cook, 1996)

$$\mathcal{S}_{Y|\mathbf{X}} = \bigcap_{\text{all DRS}} \mathcal{S}$$

Central Mean Subspace

- Multiple index model (for mean response):

$$E(Y|\mathbf{X}) = g(\alpha_1^\tau \mathbf{X}, \ldots, \alpha_k^\tau \mathbf{X})$$

  Let $\mathbf{A} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$.

- Equivalently

$$Y \perp\!\!\!\perp E[Y \mid \mathbf{X}] \mid \mathbf{A}^\tau \mathbf{X} \qquad \text{(Cook and Li 2002)}$$

  $\mathcal{S}(\mathbf{A})$ is called mean dimension reduction (MDRS) subspace

- Central mean subspace (Cook and Li 2002)

$$\mathcal{S}_{E[Y|\mathbf{X}]} = \bigcap_{\text{all MDRS}} \mathcal{S}$$

- $S_{E[Y|X]} \subset S_{Y|X}$.

Example

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

$$\begin{aligned} Y &= X_1 + (X_1 + X_3)^2 + \varepsilon\, X_4 \\ &= g(\boldsymbol{\beta}_1^\tau \mathbf{X}, \boldsymbol{\beta}_2^\tau \mathbf{X}) + \varepsilon\, h(\boldsymbol{\beta}_3^\tau \mathbf{X}) \end{aligned}$$

$$\begin{aligned} E[Y \mid \mathbf{X}] &= X_1 + (X_1 + X_3)^2 \\ &= g(\boldsymbol{\beta}_1^\tau \mathbf{X}, \boldsymbol{\beta}_2^\tau \mathbf{X}) \end{aligned}$$

Therefore,

$$\mathcal{S}_{Y|\mathbf{X}} = \mathrm{span}\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3\}$$

$$\mathcal{S}_{E[Y|\mathbf{X}]} = \mathrm{span}\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\}$$

where $\varepsilon \perp\!\!\!\perp \mathbf{X}$ and $E[\varepsilon] = 0$,

$$\begin{aligned} \boldsymbol{\beta}_1 &= (1, 0, 0, 0, 0)^\tau \\ \boldsymbol{\beta}_2 &= (1, 0, 1, 0, 0)^\tau \\ \boldsymbol{\beta}_3 &= (0, 0, 0, 1, 0)^\tau \end{aligned}$$

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 Regresssion (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 $\mathbf{M}$ depending on $\mathbf{X}$ and Y, such that

$$\mathcal{S}(\mathbf{M}) \subseteq \mathcal{S}_{Y|\mathbf{x}} \quad (\text{or } \mathcal{S}_{E[Y|\mathbf{x}]})$$

- Given a sample $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots, n$,
  1. Find an estimate $\widehat{\mathbf{M}}$ of $\mathbf{M}$.
  2. Perform spectral decomposition of $\widehat{\mathbf{M}}$.
  3. Estimate CS (or CMS) by the space spanned by the eigenvectors of $\widehat{\mathbf{M}}$ corresponding to the largest q eigenvalues.

Fourier Method for Central Mean Subspace

$\boxed{\textit{Heuristics:}}$ Find some vectors that belong to CMS, and let them span the whole CMS.

- $m(\mathbf{x}) = E(Y|\mathbf{X}=\mathbf{x})$ is a function of $\mathbf{u} = \mathbf{A}^\tau \mathbf{x}$, then

$$\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}) = \mathbf{A}\,\frac{\partial g}{\partial \mathbf{u}}(\mathbf{u}) \in \mathcal{S}_{E[Y|\mathbf{X}]}$$

- For any $\boldsymbol{\omega} \in \mathbb{R}^p$,

$$\begin{aligned}\psi(\boldsymbol{\omega}) &= \int \exp\{\imath\,\boldsymbol{\omega}^\tau \mathbf{x}\}\,\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x})\, f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\ &= -E_{(\mathbf{X},Y)}\Big[Y\,(\imath\boldsymbol{\omega} + G(\mathbf{X}))\,\exp\{\imath\,\boldsymbol{\omega}^\tau \mathbf{X}\}\Big] \quad \in \mathcal{S}_{E[Y|\mathbf{X}]}\end{aligned}$$

where $f_{\mathbf{X}}$ is the density function of $\mathbf{X}$, and $G(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}\log f_{\mathbf{X}}(\mathbf{x})$.

Fourier Method for Central Mean Subspace

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

$$\mathcal{S}(\psi(\boldsymbol{\omega})\bar{\psi}(\boldsymbol{\omega})^\tau) \subseteq \mathcal{S}_{E[Y|\mathbf{X}]}, \quad \text{for any } \boldsymbol{\omega} \in \mathbb{R}^p$$

Theorem 1. *Define matrix*

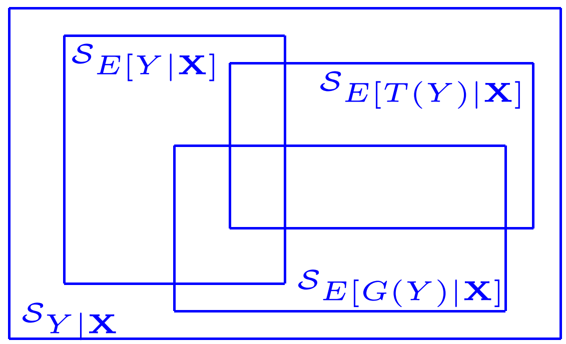
$$\mathbf{M}_{FM} = Re \int \psi(\boldsymbol{\omega})\bar{\psi}(\boldsymbol{\omega})^\tau K(\boldsymbol{\omega})\, d\boldsymbol{\omega}$$

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

$$\mathcal{S}(\mathbf{M}_{FM}) = \mathcal{S}_{E[Y|\mathbf{X}]}.$$

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 $\mathcal{S}_{Y|\mathbf{X}}$ by all possible central mean subspaces.

$$\mathcal{S}_{Y|\mathbf{X}} = \sum_{\text{all possible } T} \mathcal{S}_{E[T(Y)|\mathbf{X}]}$$

Represent CS in Terms of CMSs

- A family of transformations,

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

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

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

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

$$\mathcal{S}_{Y|\mathbf{X}} = \sum_{t \in \mathbb{R}} \mathcal{S}_{E[T(Y,t)|\mathbf{X}]}$$

Fourier Method for Central Subspace

Define

$$\begin{aligned}\phi(\boldsymbol{\omega}, t) &= \int \exp\{\imath\, \boldsymbol{\omega}^\tau \mathbf{x}\} \tfrac{\partial m}{\partial \mathbf{x}}(\mathbf{x}, t)\, f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\ &= -E_{(\mathbf{X},Y)}\left[(\imath\boldsymbol{\omega} + G(\mathbf{X}))\, \exp\{\imath t Y + \imath\, \boldsymbol{\omega}^\tau \mathbf{X}\}\right] \in \mathcal{S}_{Y|\mathbf{X}}\end{aligned}$$

It is obtained by substituting Y in $\psi(\boldsymbol{\omega})$ by $\exp\{\imath t Y\}$.

Theorem 2. *Define matrix*

$$\mathbf{M}_{FC} = Re \iint \phi(\boldsymbol{\omega}, t)\bar{\phi}(\boldsymbol{\omega}, t)^\tau K(\boldsymbol{\omega}) k(t)\, d\boldsymbol{\omega} dt$$

where $K(\boldsymbol{\omega})$ *and* $k(t)$ *are positive weight functions. Then* $\mathbf{M}_{FC}$ *is nonnegative definite, and*

$$\mathcal{S}(\mathbf{M}_{FC}) = \mathcal{S}_{Y|\mathbf{X}}$$

When $K(\boldsymbol{\omega})$ and $k(t)$ are Gaussian Functions

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

$$\boxed{\mathbf{M}_{\mathrm{FC}} = E\left[a_{12}\left[\sigma_\omega^2 \mathbf{I}_p + (G(\mathbf{U}_1) - \sigma_\omega^2 \mathbf{U}_{12})(G(\mathbf{U}_2) + \sigma_\omega^2 \mathbf{U}_{12})^\tau\right]\right]}$$

where $a_{12} = \exp\{-\sigma_t^2(V_1 - V_2)^2/2 - \sigma_\omega^2\|\mathbf{U}_1 - \mathbf{U}_2\|^2/2\}$ and $(\mathbf{U}_1, V_1)$ and $(\mathbf{U}_2, V_2)$ are iid as $(\mathbf{X}, Y)$.

- σ_ω^2 and σ_t^2 are tuning parameters (constants).
  - They are different from bandwidth in kernel estimation.
  - Theorem 2 is valid for any σ_ω^2 and σ_t^2.
- Other weight functions can also be used.

Estimation of $\mathbf{M}_{\text{FC}}$

Given a sample $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots, n$, $\mathbf{M}_{\text{FC}}$ can be estimated by sample average,

$$\boxed{\widehat{\mathbf{M}}_{\text{FC}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \left[\sigma_\omega^2 \mathbf{I}_p + (G(\mathbf{x}_i) - \sigma_\omega^2 \mathbf{x}_{ij})(G(\mathbf{x}_j) + \sigma_\omega^2 \mathbf{x}_{ij})^\tau\right]}$$

and the only unknown component is

$$G(\mathbf{x}_i) = \tfrac{\partial}{\partial \mathbf{x}} \log f_{\mathbf{X}}(\mathbf{x}_i) = \frac{\frac{\partial}{\partial \mathbf{x}} f_{\mathbf{X}}(\mathbf{x}_i)}{f_{\mathbf{X}}(\mathbf{x}_i)}$$

Pugging in Kernel Density Estimate

- Estimate $G(\mathbf{x}_i)$ by plugging in kernel estimate

$$\hat{G}(\mathbf{x}_i) = \frac{\frac{\partial}{\partial \mathbf{x}} \hat{f}_h(\mathbf{x}_i)}{\hat{f}_h(\mathbf{x}_i)}$$

where

$$\hat{f}_h(\mathbf{x}_i) = \frac{1}{nh^p} \sum_{\ell=1}^{n} W\left(\frac{\mathbf{x}_i - \mathbf{x}_\ell}{h}\right)$$

$$\frac{\partial}{\partial \mathbf{x}} \hat{f}_h(\mathbf{x}_i) = \frac{1}{nh^{p+1}} \sum_{\ell=1}^{n} W'\left(\frac{\mathbf{x}_i - \mathbf{x}_\ell}{h}\right)$$

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

$$\boxed{\widehat{\mathbf{M}}_{\text{FCk}} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\left[\sigma_\omega^2\mathbf{I}_p + (\hat{G}(\mathbf{x}_i) - \sigma_\omega^2\mathbf{x}_{ij})(\hat{G}(\mathbf{x}_j) + \sigma_\omega^2\mathbf{x}_{ij})^\tau\right]\hat{I}_i\hat{I}_j}$$

where $a_{ij} = \exp\{-\sigma_t^2 y_{ij}^2/2 - \sigma_\omega^2\mathbf{x}_{ij}^\tau\mathbf{x}_{ij}/2\}$, $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$, and $y_{ij} = y_i - y_j$, $\hat{I}_i = I_{[\hat{f}_h(\mathbf{x}_i) > b_n]}$, $I_{[\cdot]}$ 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

$$\sqrt{n}\,(vec(\widehat{\mathbf{M}}_{FCk}) - vec(\mathbf{M}_{FC})) \xrightarrow{\mathcal{L}} N(0, \mathbf{\Sigma})$$

where $\mathbf{\Sigma}$ *is a positive definite matrix.*

General Procedure for Estimating Subspaces

Suppose we have observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$.

1. Specify parameters: q, $\sigma_\omega^2 = 0.1$, $\sigma_t^2 = 1.0$, h, and b_n, if applicable.
2. Standardize data by $\tilde{\mathbf{x}}_i = \widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$ and $\tilde{y}_i = (y_i - \bar{y})/s_y$
3. Calculate an estimate $\widehat{\mathbf{M}}$ of $\mathbf{M}_{\text{FC}}$ (or $\mathbf{M}_{\text{FM}}$) using data $(\tilde{\mathbf{x}}_i, \tilde{y}_i)$.
4. Perform spectral decomposition of $\widehat{\mathbf{M}}$. The eigenvalues are $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$, and their corresponding eigenvectors are $\hat{\mathbf{e}}_1, \ldots \hat{\mathbf{e}}_p$.
5. Estimate $\mathcal{S}_{Y|\mathbf{X}}$ (or $\mathcal{S}_{E[Y|\mathbf{X}]}$) by $\widehat{\mathcal{S}} = \text{span}\{\widehat{\mathbf{\Sigma}}^{-1/2}\hat{\mathbf{e}}_1, \ldots, \widehat{\mathbf{\Sigma}}^{-1/2}\hat{\mathbf{e}}_q\}$.

Simulation Example

Assume $\mathbf{X} \in R^5$,

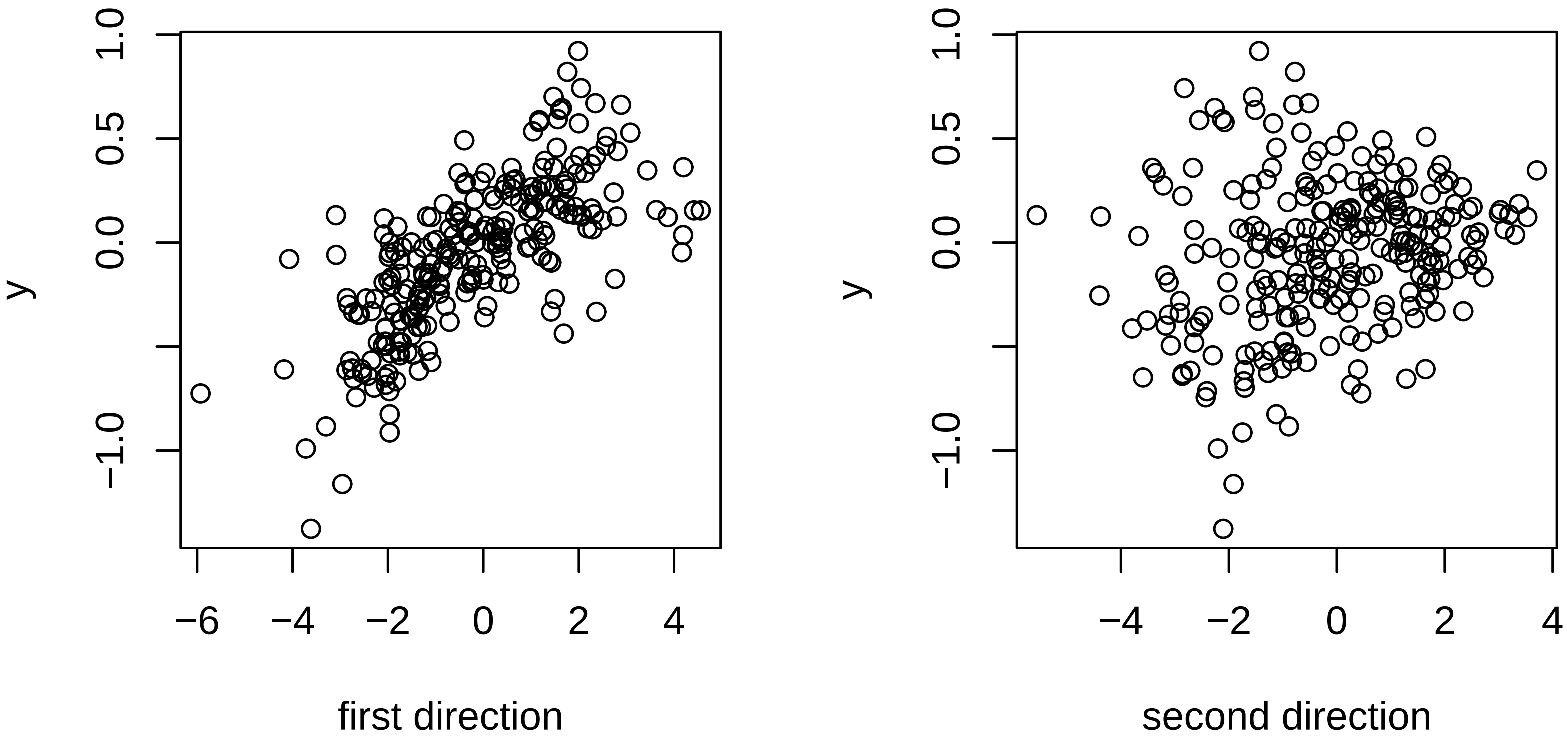
$$Y = \frac{b_1^\tau \mathbf{X}}{3 + (2 + b_2^\tau \mathbf{X})^2} + 0.2\epsilon$$

where $b_1 = (1, 1, 0, 0, 0)$, $b_2 = (0, 0, 0, 1, 1)$, ϵ is a random error, and $\mathbf{X}$ follows a mixture of multivariate distributions

$$\mathbf{X} \sim 0.4N(a_1, I_5) + 0.6N(a_2, I_5)$$

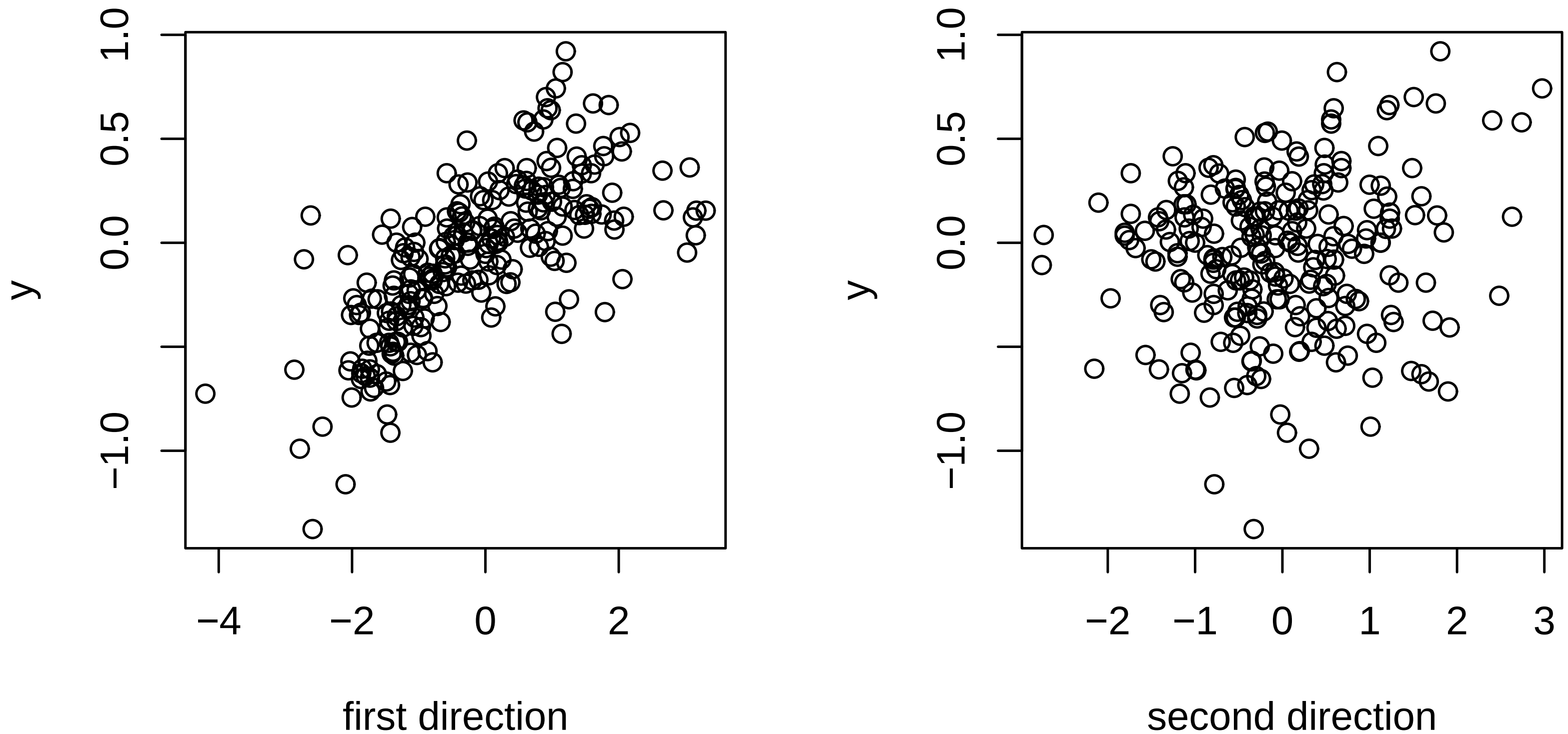
Clearly, $S_{Y|\mathbf{X}} = \mathcal{S}(b_1, b_2)$.

- A random sample of 250 observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^{250}$ is generated
- Plots of y_i versus $b_1^\tau \mathbf{x}_i$ and $b_2^\tau \mathbf{x}_i$:


y
1.0
0.5
0.0
−1.0
−6
−4
−2
0
2
4
first direction
y
1.0
0.5
0.0
−1.0
−4
−2
0
2
4
second direction

Estimated Directions

- Estimate $\mathbf{M}_{\text{FC}}$ using $\sigma_\omega^2 = 0.1$, $\sigma_t^2 = 1.0$ and $h = 1$.
- Obtain the first two eigenvectors of $\mathbf{M}_{\text{FC}}$ and use them to span a space as the estimate of $S_{Y|\mathbf{X}}$.
- Plots of y_i versus the estimated directions


y
1.0
0.5
0.0
−1.0
−4
−2
0
2
first direction
y
1.0
0.5
0.0
−1.0
−2
−1
0
1
2
3
second direction