# A Selective Review of Sufficient Dimension Reduction

**Lexin Li**

**Department of Statistics**
**North Carolina State University**

# Outline

1. General Framework
   - Sufficient Dimension Reduction
   - Key Concepts
   - Estimation Approaches
2. Basic Dimension Reduction Approaches
   - Sliced Inverse Regression
   - Other Dimension Reduction Approaches
3. Some Potentially Useful Extensions
   - Classification
   - Variable Selection
   - Complex Responses
   - Complex Predictors
   - Nonlinear Sufficient Dimension Reduction
4. Conclusion and Discussion

# Sufficient dimension reduction

**Basic regression (supervised learning) setup:**

- study the conditional distribution of $Y \in \mathbb{R}^r$ given $X \in \mathbb{R}^p$
- find a $p \times d$ matrix $\gamma = (\gamma_1, \ldots, \gamma_d)$, $d \leq p$, such that

$$Y \perp\!\!\!\perp X | \gamma^{\mathsf{T}} X \ \Leftrightarrow \ Y|X = Y|\gamma^{\mathsf{T}} X \ \Leftrightarrow \ X|(\gamma^{\mathsf{T}} X, Y) = X|\gamma^{\mathsf{T}} X$$

- replace $X$ with $\gamma^{\mathsf{T}} X = (\gamma_1^{\mathsf{T}} X, \ldots, \gamma_d^{\mathsf{T}} X)$ without losing any regression information of $Y|X$
- $(\gamma_1^{\mathsf{T}} X, \ldots, \gamma_d^{\mathsf{T}} X)$ are called the *sufficient predictors*
- $\gamma$ is not unique!

# Key concepts

**Central subspace:**

$$Y|X = Y|\gamma^{\top}X \;\Rightarrow\; \mathcal{S}_{DRS} = \mathrm{Span}(\gamma) \;\Rightarrow\; \mathcal{S}_{Y|X} = \cap\, \mathcal{S}_{DRS}$$

**Examples:**

$$\begin{aligned}
Y &= f(\gamma_1^{\top}X) + \sigma\varepsilon \\
Y &= f_1(\gamma_1^{\top}X) + f_2(\gamma_2^{\top}X) \times \varepsilon \\
\text{logit} &= \gamma_1^{\top}X, \ \text{where logit} = \log\left\{\frac{P(Y=1|X)}{1-P(Y=1|X)}\right\}
\end{aligned}$$

# Key concepts

**Central mean subspace:**

$$E(Y|X) = E(Y|\gamma^{\top} X) \;\; \Rightarrow \;\; \mathcal{S}_{E(Y|X)}$$

For many models, $\mathcal{S}_{Y|X} = \mathcal{S}_{E(Y|X)}$

**Examples:**

$$\begin{aligned} Y &= f_1(\gamma_1^{\top} X) + \ldots + f_d(\gamma_d^{\top} X) + \varepsilon \\ Y &= f_1(\gamma_1^{\top} X) + f_2(\gamma_2^{\top} X) \times \varepsilon \end{aligned}$$

# Estimation approaches

**Inverse moment based:**

- sliced inverse regression (Li, 1991) and many variants: $E(X|Y)$
- sliced average variance estimation (Cook and Weisberg, 1991) $\mathrm{Cov}(X|Y)$
- directional regression (Li and Wang, 2007)

**Kernel smoothing based:**

- minimum average variance estimation (Xia et al. 2002): estimation of the derivative of $E(Y|X)$
- variants: Xia (2007), Wang and Xia (2008)

**Others:** (not complete)

- ordinary least squares (Li and Duan, 1991)
- reproducing kernel Hilbert space (Fukumizu, Bach and Jordan, 2004, 2009)
- contour based (Li, Zha and Chiramonte, 2005, Li, Artemious and Li, 2010)

# Estimation approaches

**Comparison:**

- Inverse moment based:
  - very easy and fast to compute
  - requires a relatively large sample size
  - requires conditions on the distribution of $X$ (*linearity condition*)
- Kernel smoothing based:
  - works well for small sample size
  - requires no condition on $X$
  - requires kernel smoothing
  - relatively slow

# Sliced inverse regression

**Foundation:** under the linearity condition,

$$\Sigma_x^{-1} E\{X - E(X)|Y\} \in \mathcal{S}_{Y|X}$$

**Spectral decomposition formulation:**

$$\Sigma_{x|y}\gamma_j = \lambda_j \Sigma_x \gamma_j, \quad j = 1, \ldots, p,$$

where $\Sigma_{x|y} = \mathrm{Cov}[E\{X - E(X)|Y\}]$ and $\Sigma_x = \mathrm{Cov}(X)$.

- obtain the first $d$ eigenvectors $(\gamma_1, \ldots, \gamma_d)$ corresponding to the largest $d$ positive eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d > 0$, then $\mathrm{Span}(\gamma_1, \ldots, \gamma_d) \subseteq \mathcal{S}_{Y|X}$
- assumes $Y$ is categorical or slice $Y$ to estimate $E(X|Y)$
- asymptotic test / permutation test / BIC to determine $d$

# Sliced inverse regression

**The linearity condition:**

- $E(X|\gamma^\top X)$ is a linear function of $\gamma^\top X$ for a $\mathcal{S}_{Y|X}$ basis $\gamma$
- $X$ is elliptically symmetric; $X$ is normally distributed
- approximately true as $p \to \infty$ with a fixed $d$
- involves no $Y$ or $Y|X$, so *nonparametric* or *model-free*

**Some important variants:**

- canonical correlation analysis: $\max \mathrm{Corr}^2\{h(Y), b^\top X\}$ over $h(\cdot)$ and $b$
- letting $\beta \equiv E[h(Y)\Sigma_x^{-1}E\{X - E(X)|Y\}] = \Sigma_x^{-1}\mathrm{Cov}\{h(Y), X\}$, then $\beta \in \mathcal{S}_{Y|X}$

**Beyond SIR:**

- sliced average variance estimation: 2nd inverse moment; exhaustive
- directional regression: 1st and 2nd inverse moments; exhaustive

# Other dimension reduction approaches

**Principal components analysis:**

- spectral decomposition of $\Sigma_x$
- unsupervised; linear combinations of $X$

**Partial least squares:**

- at the population level, PLS = OLS; under the linearity condition, PLS estimates $\mathcal{S}_{E(Y|X)}$ (Li, Cook and Tsai, 2007)
- supervised; linear combinations of $X$

**Multidimensional scaling and nonlinear dimension reduction:**

- unsupervised; nonlinear combinations of $X$

**Indepedent components analysis:**

- unsupervised; linear combinations of $X$

# Classification

**Discriminant analysis:**

- directly applicable to categorical $Y$
- at the population level, SIR $\Leftrightarrow$ LDA $\Leftrightarrow$ Fisher's discriminant analysis; SAVE $\Leftrightarrow$ QDA
- SIR/SAVE produce sufficient predictors instead of classification rule; LDA/QDA produce probability estimate of $Y = g|X$ and a classification rule
- SIR/SAVE require the linearity condition (normality) on $X$; LDA/QDA require the normality assumption on $X|Y$

**Why useful:**

- of course ...

# Variable selection

**Basic ideas:**

- rewrite the SDR estimation in least squares, then apply $L_1$ type penalty (adaptive group Lasso, SCAD)
- foundation: $Y \perp\!\!\!\perp X_A | X_I \Leftrightarrow$ corresponding rows of $\gamma = 0$
- differ from most model-based variable selection approaches in that no parametric model on $Y|X$ is imposed

**Consistency in selection:**

- fixed $p$, $n \to \infty$ (Ni, Cook and Tsai, 2005, Bondell and Li, 2009)
- diverging $p \to \infty$, $n \to \infty$, $p < n$ (Wu and Li, 2010)
- $p = o(a^n)$ for any fixed $a > 1$ (Zhu, Li, Li and Zhu, 2010)

**Why useful:**

- help interpretation, e.g., identifying regions of brain that are relevant to phenotype

# Multivariate and complex responses

**Basic ideas:**

- dimension reduction is still on $X$ instead of $Y$
- key observations:

$$\begin{aligned} \mathcal{S}_{E(Y|X)} &= \mathcal{S}_{E(Y_1|X)} \oplus \ldots \oplus \mathcal{S}_{E(Y_r|X)} \\ \mathcal{S}_{Y|X} &\supseteq \mathcal{S}_{Y_1|X} \oplus \ldots \oplus \mathcal{S}_{Y_r|X} \end{aligned}$$

- multivariate reduced rank model (Cook and Setodji, 2003): one response at a time
- projective sampling (Li, Wen and Zhu, 2008): sample $a$ on a unit ball $O(n)$ times, and regress $a^\mathsf{T} Y$ on $X$

**Why useful:**

- e.g., voxel-wise imaging genetics (ignore the spatial information)
- what if $Y$ has structures, such as spatial information in MRI, or positive definiteness in DTI? — open question

# Predictors with structures

**Basic ideas:**

- predictors have group structure, and dimension reduction (linear combinations) should be within groups (Li, 2009, Li, Li and Zhu, 2010)
- direct sum structure: $\gamma_1 \oplus \ldots \oplus \gamma_g$
- partial dimension reduction, e.g., genetic / imaging information plus clinical / demographical information

**Why useful:**

- fusion of different data modalities
- what if $X$ has, e.g., network structures? — open question

# Matrix or array valued predictors

**Basic ideas:**

- predictor is a matrix or an array instead of a vector, and dimension reduction wishes to preserve interpretation
- dimension folding (Li, Kim and Altman, 2010):

$$\gamma^{\mathsf{T}} \mathrm{X} \eta = (\eta \otimes \gamma)^{\mathsf{T}} \mathrm{vec}(\mathrm{X})$$

- tensor PCA / tensor ICA

**Why useful:**

- MRI or fMRI

# Functional predictors

**Basic ideas:**

- predictor is a functional curve (dense / sparse)
- sliced inverse regression in functional space (Ferré and Yao, 2003, 2005, Hsing and Ren, 2009, Li and Hsing, 2010):
- functional PCA

**Why useful:**

- common nowadays in genetics and imaging data

# Nonlinear sufficient dimension reduction

**Basic ideas:**

- map $X$ to $\phi(X)$, then do linear SDR in the $\phi(X)$ space (Wu, Liang and Mukherjee, 2008, Zhu and Li, 2010)
- the optimal separating hyperplane (Li, Artemious and Li, 2010)

**Why useful:**

- categorical predictors
- $n < p$
- predictors with complex structures
- how to do variable selection in this setup? — open question

# Conclusion and discussion

- application of existing SDR solutions to imaging data
- motivate new methodology development for dimension reduction

# Thank You!