Some Recent Developments in Sufficient Dimension Reduction

Lexin Li

University of Minnesota, September 17, 2009
Outline

• A motivating example: diffuse large-B-cell lymphoma (DLBCL) data

• New challenges to SDR methodology
  – small-\(n\)-large-\(p\)
  – simultaneous variable selection
  – missing data
  – censored response

• Some solutions to address those challenges

• Concluding remarks

Joint work with Howard Bondell, Wenbin Lu, Xiangrong Yin
A Microarray Survival Data

Diffuse large-B-cell lymphoma:

- DLBCL is the most common type of lymphoma in adults
- potentially curable by anthracycline-based chemotherapy
- however, the survival rate is only 35% to 40%
- it is important to understand the factors that influence the patients’ survival after the chemotherapy


- response: survival time (right censored) of 240 DLBCL patients
- predictors: gene expression profiles of 7399 genes
A Microarray Survival Data

Goal: use gene expression profiles to predict patients’ survival time

Data setup:

• $T$: survival time, $C$: censoring time
• $\tilde{Y} = \min(T, C)$, $\delta = I(T < C)$
• $X = (X_1, \ldots, X_p)^T$: gene expression levels of $p$ genes
• observed sample data: $\{y_i, \delta_i, x_i\}_{i=1}^n$

A Cox proportional hazards model:

$$\lambda(t|X) = \lambda_0(t) \exp\{f(X)\} = \lambda_0(t) \exp\{\beta_1 X_1 + \ldots + \beta_p X_p\}$$
A Microarray Survival Data

Challenges:

• high-dimensional predictors – curse of dimensionality

• $n << p$, where $p = 7399$, $n = 240$, no unique solution for Cox proportional hazards model

Sufficient dimension reduction solution:

• find $d$ surrogate predictors, $U_1, \ldots, U_d$, that contain all information about patients’ survival time given predictors

• $d << p$ and $d < n$

• fit a model using $U_1, \ldots, U_d$ as predictors, e.g.,

$$\lambda(t|X) = \lambda_0(t) \exp\{f(U_1, \ldots, U_d)\}$$
Introduction to SDR

General framework of SDR:

- study 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
  \[ Y \perp X \mid \Gamma^T X \quad \Leftrightarrow \quad Y \mid X \overset{\mathcal{D}}{=} Y \mid \Gamma^T X \]
- replace $X$ with $\Gamma^T X = (\gamma_1^T X, \ldots, \gamma_d^T X)$ without losing any regression information of $Y \mid X$

Key concept: central subspace $S_{Y \mid X}$

- $Y \perp X \mid \Gamma^T X \quad \Rightarrow \quad S_{DRS} = \text{Span}(\Gamma) \quad \Rightarrow \quad S_{Y \mid X} = \cap S_{DRS}$
- $S_{Y \mid X}$ is a parsimonious population parameter that captures all regression information of $Y \mid X$; main object of interest in SDR
Introduction to SDR

Known regression models in SDR:

- multi-index model: \[ Y = f_1(\gamma_1^T X) + \ldots + f_d(\gamma_d^T X) + \varepsilon \]
- heteroscedastic model: \[ Y = f(\gamma_1^T X) + g(\gamma_2^T X)\varepsilon \]
- logit model: \[ \log \left( \frac{P(Y=1|X)}{1-P(Y=1|X)} \right) = f(\gamma_1^T X) \]

Existing SDR estimation methods:

- sliced inverse regression (SIR), sliced average variance estimation (SAVE), principal Hessian directions (PHD), minimum average variance estimation (MAVE), inverse regression estimation (IRE), covariance inverse regression estimation (CIRE), \ldots
Sliced Inverse Regression

Original formulation:

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

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

- 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 \( \text{Span}(\gamma_1, \ldots, \gamma_d) \subseteq S_{Y|X} \)

- assumes \( Y \) is categorical or slice \( Y \) to estimate \( E(X|Y) \)

- asymptotic / permutation test available to determine \( d \)

- requires the linearity condition
Recall: find surrogate predictors, \((U_1, \ldots, U_d)\), that contain all regression information about survival time \(T\) – use \((\gamma_1^T X, \ldots, \gamma_d^T X)\) obtained from applying SIR to regression of \(T\) on \(X\)

**Some challenges:**

- \(n < p\) and highly correlated predictors, the sample covariance matrix \(\hat{\Sigma}_x\) is not invertible, or, the matrix inversion becomes unstable
- variable (gene) selection along with dimension reduction
- missing predictor information
- true survival time \(T\) is not observable (right censored)
Ridge Sliced Inverse Regression

Targeting problem: \( n < p \) and highly correlated predictors

Motivation: ridge regression for homoscedastic linear model

Key ideas:

- develop a least squares formulation of SIR
- add ridge penalty
A least squares formulation of SIR:

$$\min_{A,C} \sum_{y=1}^{h} \hat{f}_y \| (\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y \|^2$$

where $A \in \mathbb{R}^{p\times d}$, $C = (C_1, \ldots, C_h) \in \mathbb{R}^{d\times h}$, $Y$ is assumed to take values in $\{1, 2, \ldots, h\}$, $\bar{X}_y$ denotes the average of $X$ in the $y$-th slice, and $\hat{f}_y = n_y/n$. Let $\hat{A}$ denote its minimizer. Then $\hat{A}$ gives the SIR solution.
Ridge Sliced Inverse Regression

Ridge sliced inverse regression (I):

\[
\min_{A,C} \sum_{y=1}^{h} \hat{f}_y \| (\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y \|^2 + \tau \text{vec}(A)^T \text{vec}(A)
\]

Let \( \hat{A} \) denote its minimizer. Then \( \hat{A} \) gives the ridge sliced inverse regression solution.
Ridge Sliced Inverse Regression

Alternating least squares optimization:

- Given $A$, $\hat{C} = (\hat{C}_1, \ldots, \hat{C}_h)$, with $\hat{C}_y = (A^T\hat{\Sigma}^2_x A)^{-1} A^T\hat{\Sigma}_x (\bar{X}_y - \bar{X})$

- Given $C$, $\text{vec}(\hat{A}) = \left( C D_f C^T \otimes \hat{\Sigma}^2_x + \tau I_{pd} \right)^{-1} \left( C D_f \otimes \hat{\Sigma}_x \right) \tilde{Y}$, where $\tilde{Y} = \text{vec}(\bar{X}_1 - \bar{X}, \ldots, \bar{X}_h - \bar{X})$ and $D_f = \text{diag}(\hat{f}_1, \ldots, \hat{f}_h)$.

Tuning parameter selection via GCV:

$$\text{GCV} = \frac{\| (I_{ph} - S_\tau) (D_f^{1/2} \otimes I_p) \tilde{Y} \|^2}{ph \{1 - \text{trace}(S_\tau)/ph\}^2},$$

where

$$S_\tau = \left( D_f^{1/2} \hat{C}^T \otimes \hat{\Sigma}_x \right) \left( \hat{C} D_f \hat{C}^T \otimes \hat{\Sigma}^2_x + \tau I_{pd} \right)^{-1} \left( \hat{C} D_f^{1/2} \otimes \hat{\Sigma}_x \right).$$
Ridge sliced inverse regression (II):

$$\min_{A,C} \sum_{y=1}^{h} f_y \| (\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y \|^2 + \tau \text{vec}(AC')^T (D_f \otimes \hat{\Sigma}_x) \text{vec}(AC')$$

Let $\hat{A}$ denote its minimizer. Then $\hat{A}$ gives the ridge sliced inverse regression solution.
Ridge Sliced Inverse Regression

Closed form optimization solution: \( \hat{A} \) consists of the first \( d \) eigenvectors of the spectral decomposition

\[
\hat{\Sigma}_x|_y \gamma_j = \lambda_j (\hat{\Sigma}_x + \tau I_p) \gamma_j, \quad j = 1, \ldots, p,
\]

and this is the ridge SIR proposal of Zhong, et al (2005).

Tuning parameter selection via GCV:

\[
\text{GCV} = \frac{\| (I_{ph} - S^*_\tau)(D_f^{1/2} \otimes I_p)\tilde{Y} \|^2}{ph\{1 - \text{trace}(S^*_\tau)/ph\}^2},
\]

where

\[
S^*_\tau = \left( D_f^{1/2} \hat{C}^T (\hat{C} D_f \hat{C}^T)^{-1} \hat{C} D_f^{1/2} \right) \otimes \left( \hat{\Sigma}_x^{1/2} (\hat{\Sigma}_x + \tau I_p)^{-1} \hat{\Sigma}_x^{1/2} \right).
\]
Shrinkage Sliced Inverse Regression

Targeting problem: simultaneous variable selection along with dimension reduction

Motivation: Lasso type of penalized regression

Key ideas:

• a two-step procedure: obtain a usual SDR estimator, then introduce shrinkage

• to achieve variable selection, shrinkage is imposed on the entire rows of the dimension reduction basis
Shrinkage sliced inverse regression: introduce a $p \times 1$ shrinkage vector $\alpha = (\alpha_1, \ldots, \alpha_p)^\top$; given $(\hat{A}, \hat{C})$

$$\hat{\alpha} = \arg\min_{\alpha} \left\{ \sum_{y=1}^{h} f_y \| (\bar{X}_y - \bar{X}) - \hat{\Sigma}_x \text{diag}(\alpha) \hat{A} \hat{C}_y \|^2 \right\}$$

subject to $\sum_{j=1}^{p} |\alpha_j| \leq \lambda$. Then $\text{Span} \left( \text{diag}(\hat{\alpha}) \hat{A} \right)$ is called a shrinkage sliced inverse regression estimator of $S_{Y|X}$.

Variable selection by shrinkage:

- when $\lambda \geq p$, $\hat{\alpha}_j = 1$ for all $j$’s
- when $\lambda$ decreases, some $\hat{\alpha}_j$’s are shrunken to zero, which in turn shrinking the entire rows of $\hat{A}$
**Shrinkage Sliced Inverse Regression**

“Standard” Lasso type optimization:

- rewrite the optimization problem as

\[
\hat{\alpha} = \arg \min_{\alpha} \left\{ \| \tilde{Y} - (\text{diag}(\hat{A}\hat{C}_1)\hat{\Sigma}_x, \ldots, \text{diag}(\hat{A}\hat{C}_h)\hat{\Sigma}_x)^\top \alpha \|_2^2 \right\}
\]

where \( \tilde{Y} = \text{vec}(\bar{X}_1 - \bar{X}, \ldots, \bar{X}_h - \bar{X}) \in \mathbb{R}^{ph \times 1}. \)

- the solution can be obtained by any Lasso algorithm.

**Tuning parameter selection via AIC/BIC**

**Remarks:**

- Closely related to nonnegative garroote (Breiman, 1995):
  \[ P(\hat{\alpha}_j > 0) \to 1. \]
- Ni, Cook and Tsai (2005)
**Variable Selection without a Model**

**Goal:** to seek the smallest subset of the predictors $X_A$, with partition $X = (X^T_A, X^T_I)^T$, such that

$$Y \perp X_I | X_A.$$ 

**Existence and uniqueness:** Given the existence of the central subspace $S_{Y|X}$, $A$ uniquely exists.

**Relation between $A$ and basis of $S_{Y|X}$:** (Cook, 2004, Proposition 1)

$$\Gamma^{p \times d} = \begin{pmatrix} \Gamma_A \\ \Gamma_I \end{pmatrix} = \begin{pmatrix} \Gamma_A \\ 0 \end{pmatrix}, \Gamma_A \in \mathbb{R}^{(p-p_I) \times d}, \Gamma_I \in \mathbb{R}^{p_I \times d}.$$ 

The rows of a basis of the central subspace corresponding to $X_I$ are all zero vectors; and all the predictors whose corresponding rows of the $S_{Y|X}$ basis equal zero belong to $X_I$. 

19
Theoretical properties:

• show that the resulting estimators achieve consistency in variable selection, i.e., can select all the truly active predictors with probability approaching one as $n \to \infty$

• show that the resulting estimators retain the $\sqrt{n}$-consistency of estimating the dimension reduction basis for all the truly active predictors

Remarks: since SDR estimators do not require any traditional models, the proposed method achieves consistent variable selection without assuming any model
Application to DLBCL Data

Dimension reduction estimation:
- principal components regression
- partial least squares based sliced inverse regression
- ridge sliced inverse regression
- shrinkage ridge sliced inverse regression

Analysis:
- pre-select 329 genes out of 7399 genes using a univariate Cox model
- double slicing for SIR to handle the censored response

Model and prediction of patient’s survival:
- fit a Cox proportional hazards model to training data
- assign patients to low, medium, and high risk groups
SDR with Missing Predictors

**Targeting problem:** develop SDR estimation methods that come as close as possible to the analysis *had data not been missing*

**Motivation:** *(augmented)* inverse probability weighted estimation (Robins, Rotnitzky and Zhao, 1994)

**Key ideas:**

- decompose SIR components
- obtained the consistent *(augmented)* inverse probability weighted estimator for those involving missing data
SDR with Missing Predictors

Current solution: the complete-case analysis

- may be inefficient if a large portion of the data are missing
- is valid only when the missing mechanism follows missing completely at random (MCAR), i.e., missingness is independent of all the observed and unobserved quantities

A more general missing mechanism: missing at random (MAR)

- missing depends on the observed data, but not on the unobserved
- partition \( X = (X_1^T, X_2^T)^T \), where \( X_1 \in \mathbb{R}^{p_1} \) has complete observations for all subjects, \( X_2 \in \mathbb{R}^{p_2} \) contains predictors with missing; introduce a missing indicator \( R \), which equals 1 if there is no missingness for all predictors, and 0 otherwise

\[ R \perp X_2 | (Y, X_1) \]
Recall SIR: $\Sigma_{x|y} \gamma_j = \lambda_j \Sigma_x \gamma_j$, where $\Sigma_{x|y} = \text{Cov}[E\{X - E(X)\}|Y]$

Partition:

$$\Phi_0 = E(X)E(X)^\top = \begin{pmatrix} E(X_1)E(X_1)^\top & E(X_1)E(X_2)^\top \\ E(X_2)E(X_1)^\top & E(X_2)E(X_2)^\top \end{pmatrix},$$

$$\Phi_1 = E( E(X|Y)E(X|Y)^\top ) = \begin{pmatrix} E(E(X_1|Y)E(X_1|Y)^\top) & E(E(X_2|Y)E(X_1|Y)^\top) \\ E(E(X_2|Y)E(X_1|Y)^\top)^\top & E(E(X_2|Y)E(X_2|Y)^\top) \end{pmatrix},$$

$$\Phi_2 = E(XX^\top) = \begin{pmatrix} E(X_1X_1^\top) & E(X_2X_1^\top) \\ E(X_2X_1^\top)^\top & E(X_2X_2^\top) \end{pmatrix}.$$ 

Note that $\Sigma_{x|y} = \Phi_1 - \Phi_0$ and $\Sigma_x = \Phi_2 - \Phi_0$. 
SIR with Missing Predictors

Inverse probability weighted estimation:

\[ X_{2}^{\text{IPW}} = \frac{R}{\pi(\beta)} X_2, \quad \text{and} \quad (X_2 X_2^T)^{\text{IPW}} = \frac{R}{\pi(\beta)} X_2 X_2^T \]

where \( \pi = P(R = 1|Y, X_1) \) denotes the probability of no missing given the observed data, and \( \pi(\beta) = \pi(Y, X_1; \beta) \) denotes a parametric model, indexed by the parameter \( \beta \), we posit on \( \pi \).

**Lemma:** Given that \( \pi(\beta) \) is correctly specified and MAR, we have

\[ E(X_{2}^{\text{IPW}}) = E(X_2), \quad E((X_2 X_2^T)^{\text{IPW}}) = E(X_2 X_2^T), \quad E(X_2^{\text{IPW}} X_1^T) = E(X_2 X_1^T), \]

and \( E(X_2^{\text{IPW}}|Y) = E(X_2|Y) \).

\[
E(X_{2}^{\text{IPW}}) = E \left[ E \left\{ \frac{R}{\pi(\beta)} X_2 | Y, X_1 \right\} \right] = E \left\{ \frac{E(R|Y, X_1) E(X_2|Y, X_1)}{\pi(\beta)} \right\} = E(X_2)
\]
Augmented inverse probability weighted estimation:

\[ X_2^{\text{AIPW}} = \frac{R}{\pi(\beta)} X_2 - \frac{R - \pi(\beta)}{\pi(\beta)} \psi_1(\theta_1), \]

\[ (X_2 X_2^T)^{\text{AIPW}} = \frac{R}{\pi(\beta)} X_2 X_2^T - \frac{R - \pi(\beta)}{\pi(\beta)} \psi_2(\theta_2). \]

where \( \psi_1 = \mathbb{E}(X_2 | Y, X_1) \), \( \psi_2 = \mathbb{E}(X_2 X_2^T | Y, X_1) \), and \( \psi_1(\theta_1), \psi_2(\theta_2) \) are parametric models we posit on \( \psi_1 \) and \( \psi_2 \), respectively.

**Lemma: (double robustness)** Given that either \( \pi(\beta) \) is correctly specified, or \( \psi(\theta_1) \) and \( \psi(\theta_2) \) are correctly specified, and MAR, we have

- \( \mathbb{E}(X_2^{\text{AIPW}}) = \mathbb{E}(X_2) \)
- \( \mathbb{E}((X_2 X_2^T)^{\text{AIPW}}) = \mathbb{E}(X_2 X_2^T) \)
- \( \mathbb{E}(X_2^{\text{AIPW}} X_1^T) = \mathbb{E}(X_2 X_1^T) \)
- \( \mathbb{E}(X_2^{\text{AIPW}} | Y) = \mathbb{E}(X_2 | Y) \).
**SIR with Missing Predictors**

**Theoretical properties:** under suitable regularity conditions, the AIPW-SIR estimator is **doubly robust** and $\sqrt{n}$-consistent

**Remarks:**

- can generalize to other SDR estimators, e.g., SAVE, PHD
- thanks to the flexible structure of $S_{Y|X}$, the new method can work with more flexible regression models other than the homoscedastic linear regression model
- the proposed estimator loses some nonparametric flavor by positing models for $\pi(\beta)$, $\psi_1(\theta_1)$ and $\psi_2(\theta_2)$; however the new method works well empirically even when simple models like logistic model or linear model were used
A Simulation Example

Simulation setup:

\[ Y = \gamma_1^T X (\gamma_1^T X + \gamma_2^T X + 3) + 0.1 \varepsilon, \]

- \( X = (V_1, \ldots, V_p)^T \) is multivariate normal with mean 0, and the correlation between \( V_i \) and \( V_j \) equals \( 0.3|i-j| \)
- The error \( \varepsilon \) is standard normal and is independent of \( X \)
- \( S_{Y|X} = \text{Span}(\gamma_1, \gamma_2) \), with dimension \( d = 2 \), \( \gamma_1 = (1, 0, 0, \ldots, 0)^T \), and \( \gamma_2 = (0, 1, 0, \ldots, 0)^T \).
- \( n = 200 \) and \( p = 5 \).
- For illustration, we only had missingness in the first predictor \( V_1 \):

\[
\pi = P(R = 1|Y, V_2, \ldots, V_p) = \frac{\exp(c_0 + 0.25Y + 0.5V_2 - V_p)}{1 + \exp(c_0 + 0.25Y + 0.5V_2 - V_p)},
\]

where \( c_0 \) is a scalar constant that controls the proportion of missing.
A Simulation Example

![Graphs showing median and mad vector correlations against missing proportion]
Discussions

Remarks: this high-dimensional (microarray) data analysis has prompted new methodological and theoretical development of sufficient dimension reduction

Not the end of the story:

• asymptotics of variable selection within SDR when $p \rightarrow \infty$
• SDR for survival data
• SDR for longitudinal data
• outlier and robust estimation
• dimension reduction that accounts for predictor group information
• …
References


Thank You!