# Sufficient Dimension Reduction with Application to Microarray Data Analysis

**Lexin Li**

NC State University
DEPARTMENT OF STATISTICS

Hangzhou, China June 11, 2007

# Outline

- Motivating example – diffuse large-B-cell lymphoma (DLBCL) data
- Statistical problem formulation
- Introduction to sufficient dimension reduction (SDR)
- Applications to DLBCL data analysis
- New challenges to SDR methodology
- Some solutions to answer those challenges
- Concluding remarks and future work

# Motivating Example

**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 predict the patients' survival after chemotherapy

**DLBCL data of Rosenwald, et al. NEJM (2002):**

- response: survival time (right censored) of 240 DLBCL patients
- predictors:
  - IPI: international prognostic index (age, tumor stage, ...)
  - gene expression profiles of 7399 genes

# Motivating Example

**Goal:** use clinic information (IPI) and/or gene expression profiles to predict patients' survival time

**Data setup:**

- $T$: survival time, $C$: censoring time
- $Y = \min(T, C)$, $\delta = I(T < C)$
- $X = (X_1, \ldots, X_p)^\intercal$: gene expression levels of $p$ genes
- $W$: IPI risk-group indicator (low, medium, & high-risk); $\{1, \ldots, K\}$
- observed sample data: $\{y_i, \delta_i, x_i, w_i\}_{i=1}^n$

**A Cox proportional hazards model:**

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

# Motivating Example

**Challenge:**

- high-dimensional predictors; curse of dimensionality
- $n << p$, where $p = 7399 + 1$, $n = 240$, no unique solution for Cox proportional hazards model

**Goals of sufficient dimension reduction:** find $d$ surrogate predictors, $Z_1, \ldots, Z_d$, such that,

- contain all information about patients' survival time given predictors
- $d << p$ and $d < n$
- fit a model using $Z_1, \ldots, Z_d$ as predictors, e.g.,

$$\lambda(t|X) = \lambda_0(t) \exp\{f(Z_1, \ldots, Z_d)\}$$

# 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

$$Y \perp\!\!\!\perp X | \Gamma^\intercal X \qquad \Leftrightarrow \qquad Y|X \overset{\mathcal{D}}{=} Y|\Gamma^\intercal X$$

- replace $X$ with $\Gamma^\intercal X = (\gamma_1^\intercal X, \ldots, \gamma_d^\intercal X)$ without losing any regression information of $Y|X$

**Key concept:** central subspace $\mathcal{S}_{Y|X}$

- $Y \perp\!\!\!\perp X | \Gamma^\intercal X \quad \Rightarrow \quad \mathcal{S}_{DRS} = \text{Span}(\Gamma) \quad \Rightarrow \quad \mathcal{S}_{Y|X} = \cap \mathcal{S}_{DRS}$
- $\mathcal{S}_{Y|X}$ is a parsimonious population parameter that captures all regression information of $Y|X$

# Introduction to SDR

**Known regression models in SDR:**

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

**Estimation methods:** (nonparametric / model-free)

- sliced inverse regression (Li,K-C, 1991)
- SAVE, PHD, IHT, ...
- IRE, CIRE, ...

**Recall:** find surrogate predictors, $(Z_1, \ldots, Z_d)$, that contain all regression information about survival time $T$ – use $(\gamma_1^\intercal X, \ldots, \gamma_d^\intercal X)$, where $\Gamma = (\gamma_1, \ldots, \gamma_d)$ is a basis of $\mathcal{S}_{T|X}$

# Sliced inverse regression for Microarray Data

**Generalized eigen decomposition:**

$$\mathrm{Cov}\{\mathrm{E}(X|T)\}\gamma_i = \lambda_i \Sigma_x \gamma_i, \qquad i = 1, \ldots, p$$

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 $\mathsf{Span}(\gamma_1, \ldots, \gamma_d) \subseteq \mathcal{S}_{Y|X}$

**Population foundation:** under some condition

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

which in turn implies that

$$\mathsf{Span}\left(\Sigma_x^{-1}\mathrm{Cov}\{\mathrm{E}(X|T)\}\right) \subseteq \mathcal{S}_{T|X}$$

# SIR for Microarray Data

**Additional observations:**

- the conditions are typically imposed on the marginal distribution of $X$, rather than $Y|X$ (model-free)
- asymptotic or permutation tests are available to determine the structural dimension $d = \dim(\mathcal{S}_{T|X})$
- sample estimation requires slicing of the response variable $T$

**Computation:**

- compute sample inverse mean within each slice $s$, $s = 1, \ldots, h$, $\widehat{\mathrm{E}}(X|T = s) = \frac{1}{n_s} \sum_{t_i \in \text{slice } s} x_i$
- compute the usual sample estimates $\widehat{\mathrm{Cov}}\{\mathrm{E}(X|T)\}$ and $\hat{\Sigma}_x$

# SIR for Microarray Data

**Sliced inverse regression for censored response:**

- true survival time $T$ is *not* observable (nor is $C$)
- what can be observed are $Y = \min(T, C)$ and $\delta = I(T < C)$
- since $(Y, \delta)$ is a function of $(T, C)$, one can show that

$$\mathcal{S}_{(Y,\delta)|X} \subseteq \mathcal{S}_{(T,C)|X}$$

- under certain conditions, $\mathcal{S}_{(Y,\delta)|X} \subseteq \mathcal{S}_{(T,C)|X} \subseteq \mathcal{S}_{T|X}$
- consequently, do SIR for bivariate response $(Y, \delta)$
- operationally, do double slicing of $(Y, \delta)$, instead of slicing $T$

# SIR for Microarray Data

**Sliced inverse regression in conjunction with principal components:**

- $p$ is large, and $n << p$
- first obtain $q$ leading principal components $\Psi^{\mathsf{T}} X$, where $\Psi = (\psi_1^{\mathsf{T}}, \ldots, \psi_q)$, that capture most variability among the predictor space
- then do SIR based on the principal components
- there will be no information loss if $\mathcal{S}_{T|X} \subseteq \text{Span}(\Psi)$
- the number of principal components $q$ can take a range of values with little effect on prediction
- empirical study shows that SIR+PC is better than PC in prediction

# SIR for Microarray Data

**Sliced inverse regression using both IPI and gene expression:**

- IPI is also a useful predictor of patients's survival time
- let $W$ denote categorical IPI, and $(X_w, T_w)$ denote $(X,T)|(W=w)$
- obtain the first $d$ eigenvectors,

$$\left(\sum_{w=1}^{K} P(W=w)\mathrm{Cov}\{\mathrm{E}(X_w|T_w)\}\right)\ \gamma_j = \lambda_j \left(\sum_{w=1}^{K} P(W=w)\mathrm{Cov}(X_w)\right)\ \gamma_j$$

- Span$(\gamma_1, \ldots, \gamma_d) \subseteq \mathcal{S}_{T|X}^{(W)}$, where the basis $\Gamma$ of $\mathcal{S}_{T|X}^{(W)}$ satisfies that

$$T \perp\!\!\!\perp (X,W)|(\Gamma^{\mathsf{T}}X, W)$$

# Application to DLBCL Data

**Analysis:**

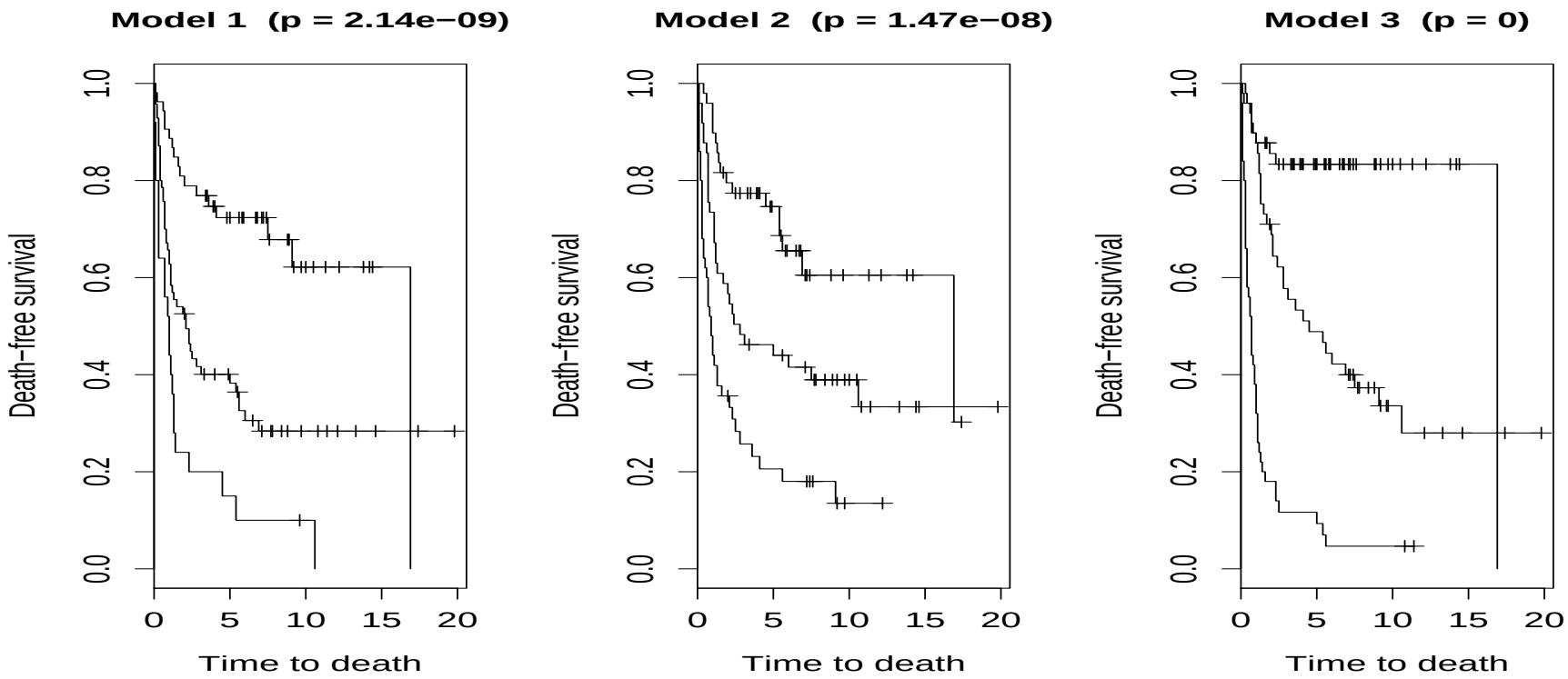
- partition data into a training group of 148 patients and a testing group of 74 patients, following Bair, et al. (2006)
- a Cox model is fitted, after dimension reduction, based on training data, and prediction is evaluated on testing data
- prediction of overall survival is based on the Kaplan-Meier estimates of survival curves for three groups of patients, defined by the fitted model: the low-risk, the medium-risk, and the high-risk patients

**Compare 3 models:**

- model 1: only IPI as predictor
- model 2: only gene expressions as predictors
- model 3: both IPI and gene expressions as predictors

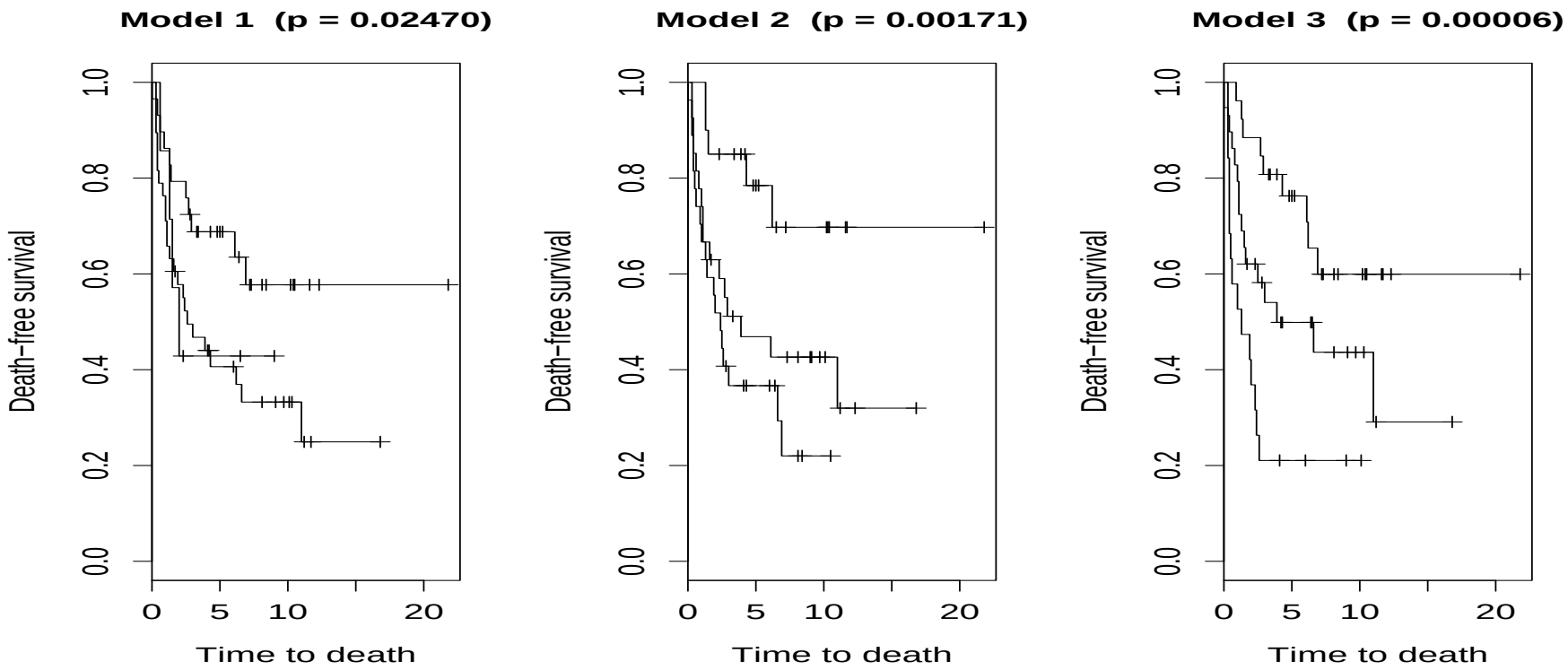
# Application to DLBCL Data

Training data:



# Application to DLBCL Data

Testing data:



# Discussions

**Remarks:** SDR can be useful for microarray data analysis

**References:**


- Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*. 20, 3406-3412.
- Li, L. (2006). Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*. 22, 466-471.


**Not the end of the story:**

- highly correlated predictors
- $n << p$
- variable (gene) selection in the framework of SDR

# Ridge Sliced Inverse Regression

**A least squares formulation of SIR:**

$$(\hat{A}, \hat{C}) = \arg\min_{A,C} \left\{ \sum_{y=1}^{h} \hat{f}_y \parallel (\bar{X}_y - \bar{X}) - \hat{\Sigma}_x A C_y \parallel^2 \right\}$$

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 $\hat{f}_y = n_y/n$, $y = 1, \ldots, h$. Then Span$(\hat{A})$ estimates $\mathcal{S}_{\mathrm{SIR}}$

**Ridge sliced inverse regression:**

$$(\hat{A}, \hat{C}) = \arg\min_{A,C} \left\{ \sum_{y=1}^{h} \hat{f}_y \parallel (\bar{X}_y - \bar{X}) - \hat{\Sigma}_x A C_y \parallel^2 + \tau \mathrm{vec}(A)^{\mathsf{T}} \mathrm{vec}(A) \right\}$$

Span$(\hat{A})$ is called a ridge sliced inverse regression estimator of $\mathcal{S}_{Y|X}$.

# Ridge Sliced Inverse Regression

**Alternating least squares optimization:**

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

**Tuning parameter selection via GCV:**

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

where

$$S_\tau = \left(D_f^{1/2} \hat{C}^\intercal \otimes \hat{\Sigma}_x\right) \left(\hat{C} D_f \hat{C}^\intercal \otimes \hat{\Sigma}_x^2 + \tau I_{pd}\right)^{-1} \left(\hat{C} D_f^{1/2} \otimes \hat{\Sigma}_x\right).$$

# Shrinkage Ridge Sliced Inverse Regression

**Remarks on ridge SIR:**

- can account for highly correlated predictors and $n < p$
- no variable selection yet

**Shrinkage ridge sliced inverse regression:** let $\alpha = (\alpha_1, \ldots, \alpha_p)^\intercal$ denote a $p \times 1$ shrinkage vector; given $(\hat{A}, \hat{C})$

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

subject to $\sum_{j=1}^{p} |\alpha_j| \leq \lambda$. Then Span $(\mathrm{diag}(\hat{\alpha})\, \hat{A})$ is called a shrinkage ridge sliced inverse regression estimator of $\mathcal{S}_{Y|X}$.

# Shrinkage Ridge Sliced Inverse Regression

**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 $\hat{A}$

**"Standard" Lasso type optimization:**

- rewrite the optimization problem as

$$\hat{\alpha} = \arg\min_{\alpha} \left\{ \| \tilde{Y} - \left( \mathrm{diag}(\hat{A}\hat{C}_1)\hat{\Sigma}_x, \ldots, \mathrm{diag}(\hat{A}\hat{C}_h)\hat{\Sigma}_x \right)^{\mathsf{T}} \alpha \|^2 \right\}$$

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

- the solution can be obtained by any Lasso algorithm.

**Tuning parameter selection via AIC/BIC/RIC:** Ni, Cook and Tsai (2005)

# Application to DLBCL Data

**Analysis:**

- focus on using gene expression to predict patients' survival
- pre-select 329 genes among 7399 genes using a univariate Cox model

**Gene selection:**

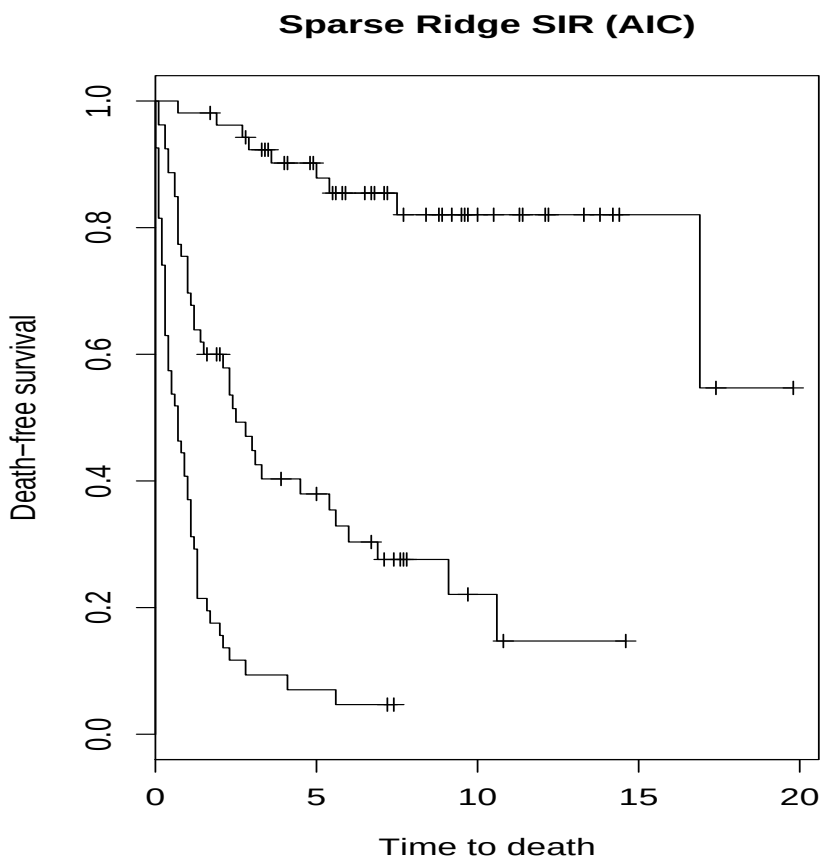
- AIC-based shrinkage ridge SIR selects 34 genes
- BIC/RIC-based shrinkage ridge SIR selects 12 genes

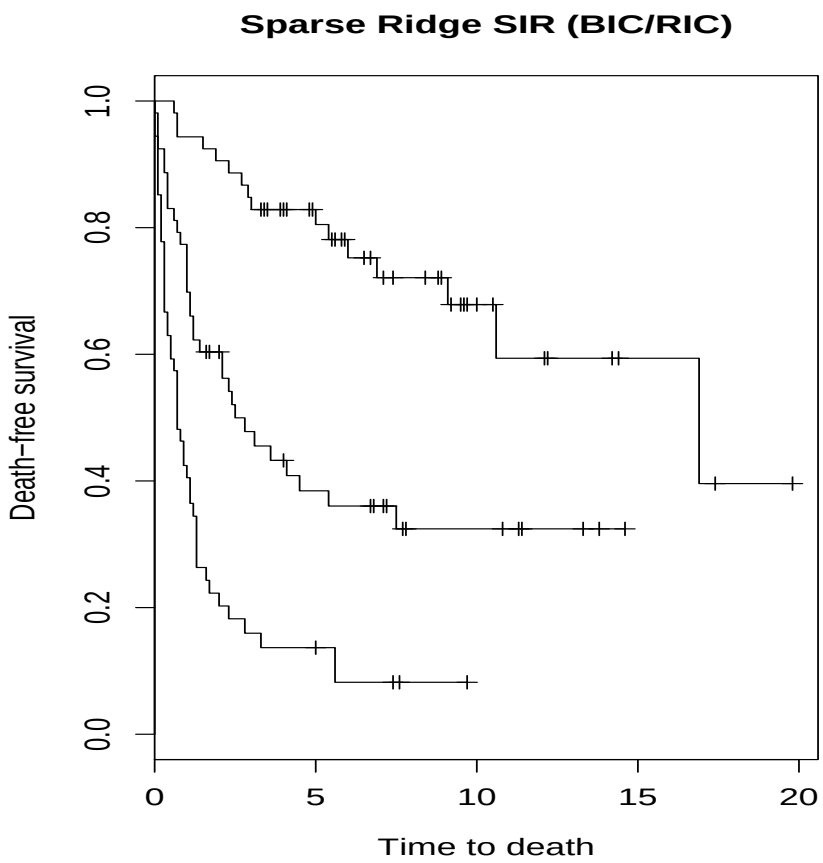
**Model and prediction of patient's survival:**

- fit a Cox proportional hazards model to training data
- assign patients to low, medium, and high risk groups

# Application to DLBCL Data

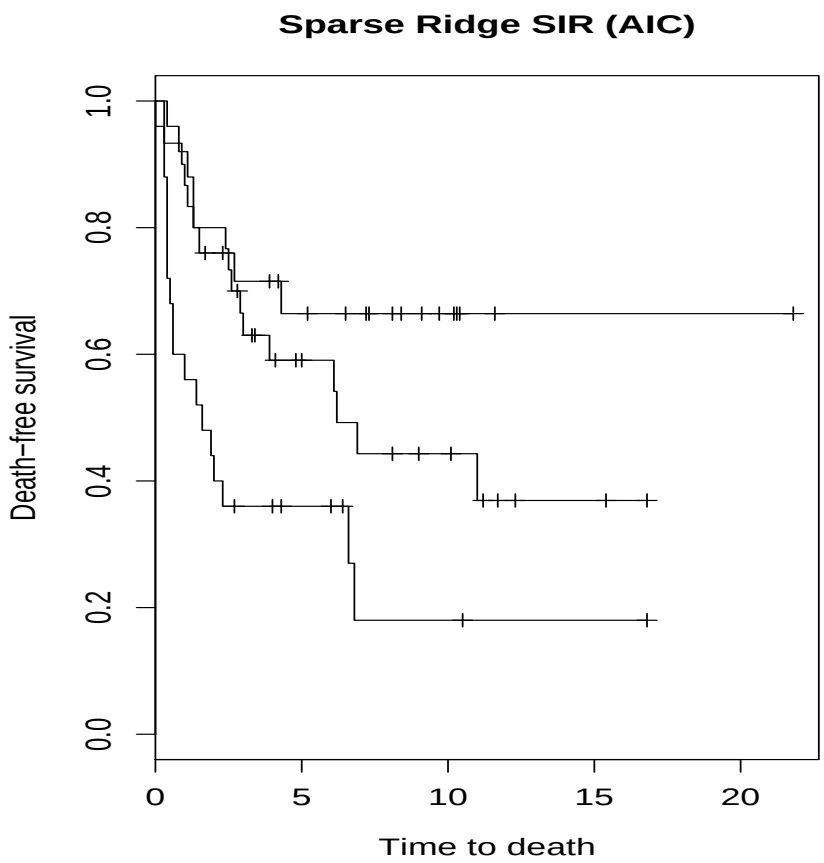
Training data:

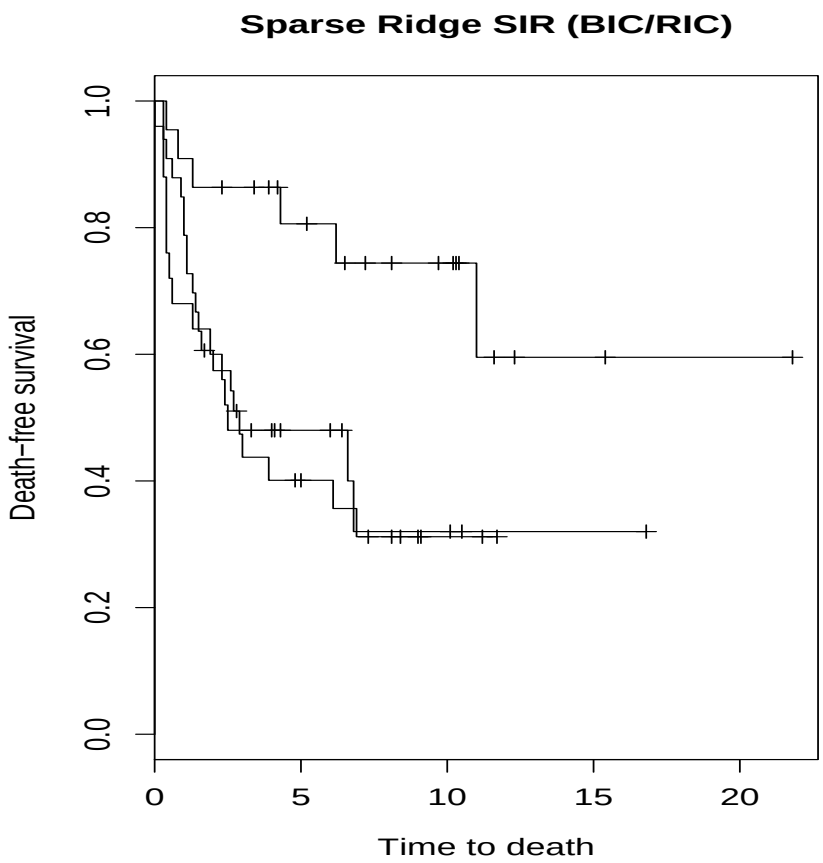


# Application to DLBCL Data

Testing data:



# Discussions

**Remarks:** Microarray data has motivated development of new methodology in SDR

**References:**


- Li, L. and Yin, X. (2007). Sliced inverse regression with regularizations. *Biometrics*, to appear.


**Still, not the end of the story:**

- ridge SIR may not be suitable for very large $p$ computationally
- asymptotic properties of variable selection has not yet been studied

# Follow-up Work

**Problem of $n << p$ as well as highly correlated predictors:**

- establish a link between partial least squares and $\mathcal{S}_{Y|X}$ in SDR framework
- develop a modified version of SIR that can effectively account for $n << p$ as well as highly correlated predictors

**References:**


- Li, L., Cook, R.D., and Tsai, C.L. (2007). A note on partial inverse regression. *Biometrika*, to appear.

# Follow-up Work

**Properties of variable selection in SDR and beyond:**

- show the oracle property of variable selection for a class of shrinkage SDR estimators; i.e., can select all active predictors consistently without assuming a model
- also generalize to a large family of SDR estimators beyond SIR

**References:**


- Bondell, H.D., and Li, L. (2007). Shrinkage inverse regression estimation for model free variable selection. Technical report.
- Li, L. (2007). A note on sparse sufficient dimension reduction. *Biometrika*, to appear.

# Thank You!