Dimension reduction methods for microarray censored survival data

Lexin Li

University of California, Davis

May 18, 2005 QPRC

Outline

- Diffuse large-B-cell lymphoma data
- Statistical problem and challenges
- Dimension reduction methods
- Application to lymphoma data
- Future work

Lymphoma Microarray Survival Data

- Diffuse large-B-cell lymphoma has an annual incidence in U.S. of more than 25,000 cases.
- Combination chemotherapy, 35% to 40% survival rate
- International prognostic index (age, tumor stage, etc) is a well-established outcome predictor. However, the outcome in patients with identical IPI values varies considerably.
- Hypothesis: gene expression profiles could be used independently of IPI to predict the patients survival after chemotherapy.

Rosenwald et al. (NEJM 2002) Data Set

- 240 patients with diffuse large-B-cell lymphoma
- 42% survival rate, median follow-up 2.8 years overall, and 7.3 years for survivors
- Gene expression profiles of 7399 genes
- 160 patients in the training group, and 80 patients in the testing group
- Our focus: use gene expression to predict censored continuous phenotype, i.e., patients survival time.

Survival Data Analysis

- Notations:
  - 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
  - Observed sample data: $\{y_i, \delta_i, X_i\}_{i=1}^n$
- A general Cox proportional hazards model

$$\lambda(t|X) = \lambda_0(t) \exp\{f(X)\} = \lambda_0(t) \exp\{\beta_1 x_1 + \ldots + \beta_p x_p\}$$

Challenges

- Challenges:
  - Phenotype (survival time) is right-censored.
  - $n << p$, where $p = 7399$, $n = 240$, no unique solution for Cox proportional hazards model
- Goal of dimension reduction: find d surrogate predictors, $s_1, \ldots, s_d$, such that,
  - Contain all the information about patients survival time
  - $d << p$ and $d < n$
  - Fit a model using $s_1, \ldots, s_d$ as predictors, e.g.,

$$\lambda(t|X) = \lambda_0(t) \exp\{f(s_1, \ldots, s_d)\}$$

Sufficient Dimension Reduction

- Goal of sufficient dimension reduction:
  - Find a $p \times d$ matrix $\eta = (\eta_1, \ldots, \eta_d)$, $d \leq p$, such that

$$T \perp\!\!\!\perp X \,|\, \eta^{\mathsf{T}} X$$

  - Replace X with $\eta^{\mathsf{T}} X = (\eta_1^{\mathsf{T}} X, \ldots, \eta_d^{\mathsf{T}} X)$
  - *without* loss of information on regression $T \,|\, X$
  - *without* assuming any model or distribution for $T \,|\, X$
- Key concept – Central subspace: $\mathcal{S}_{T|X}$

$$T \perp\!\!\!\perp X \,|\, \eta^{\mathsf{T}} X \Rightarrow \mathcal{S}_{DRS} = \text{Span}(\eta) \Rightarrow \mathcal{S}_{T|X} = \cap\, \mathcal{S}_{DRS}$$

Sliced Inverse Regression

- Surrogate predictors: $(s_1, \ldots, s_d) = (\eta_1^\mathsf{T} X, \ldots, \eta_d^\mathsf{T} X)$
  - First d eigenvectors of the eigen-decomposition

$$\Sigma_{X|T}\ \eta_i = \lambda_i\ \Sigma_X\ \eta_i$$

    where $\Sigma_{X|T} = \mathrm{Cov}(\mathrm{E}(X \mid T))$, and $\Sigma_X = \mathrm{Cov}(X)$
  - Asymptotic test is available to determine d
- To estimate $\Sigma_{X|T}$, slicing of T is needed, i.e., partitioning T into fixed non-overlapping slices
- Theoretical justification:

$$\mathrm{Span}\{\mathrm{Cov}(\mathrm{E}(X \mid T))\} \subseteq \mathcal{S}_{T \mid X}$$

Modification of SIR to Censored Data

- True survival time T is unobservable
- Since (y, δ) is a function of (T, C), one can show that

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

- Algorithm:
  - Double slicing of (y, δ) (rather than slicing of T)
  - The rest are the same as a standard SIR
- Combine SIR with Principal Component Analysis (PCA)
- Fit *any* model, e.g. a Cox proportional hazards model, using extracted SIR components as predictors

Survival Time *versus* SIR Component

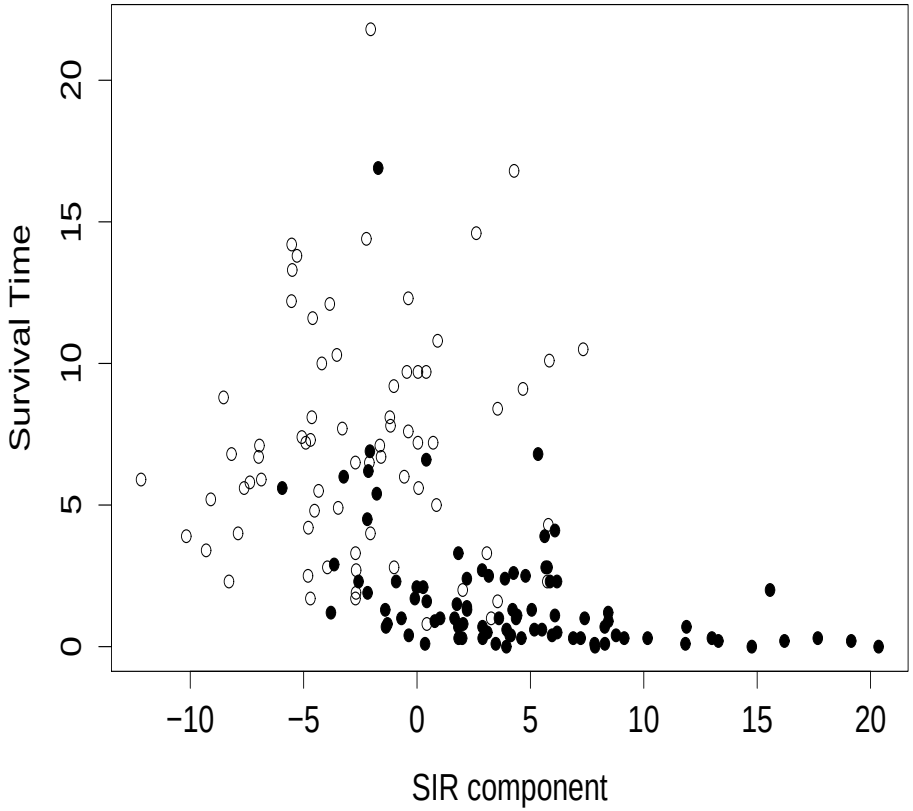

Figure 1: dot: patients who were dead; circle: patients who were alive. A Cox proportional hazards model: $\lambda(t|X) = \lambda_0(t)\exp\{0.242s - 0.005s^2\}$

Overall Survival in Predicted Risk Groups

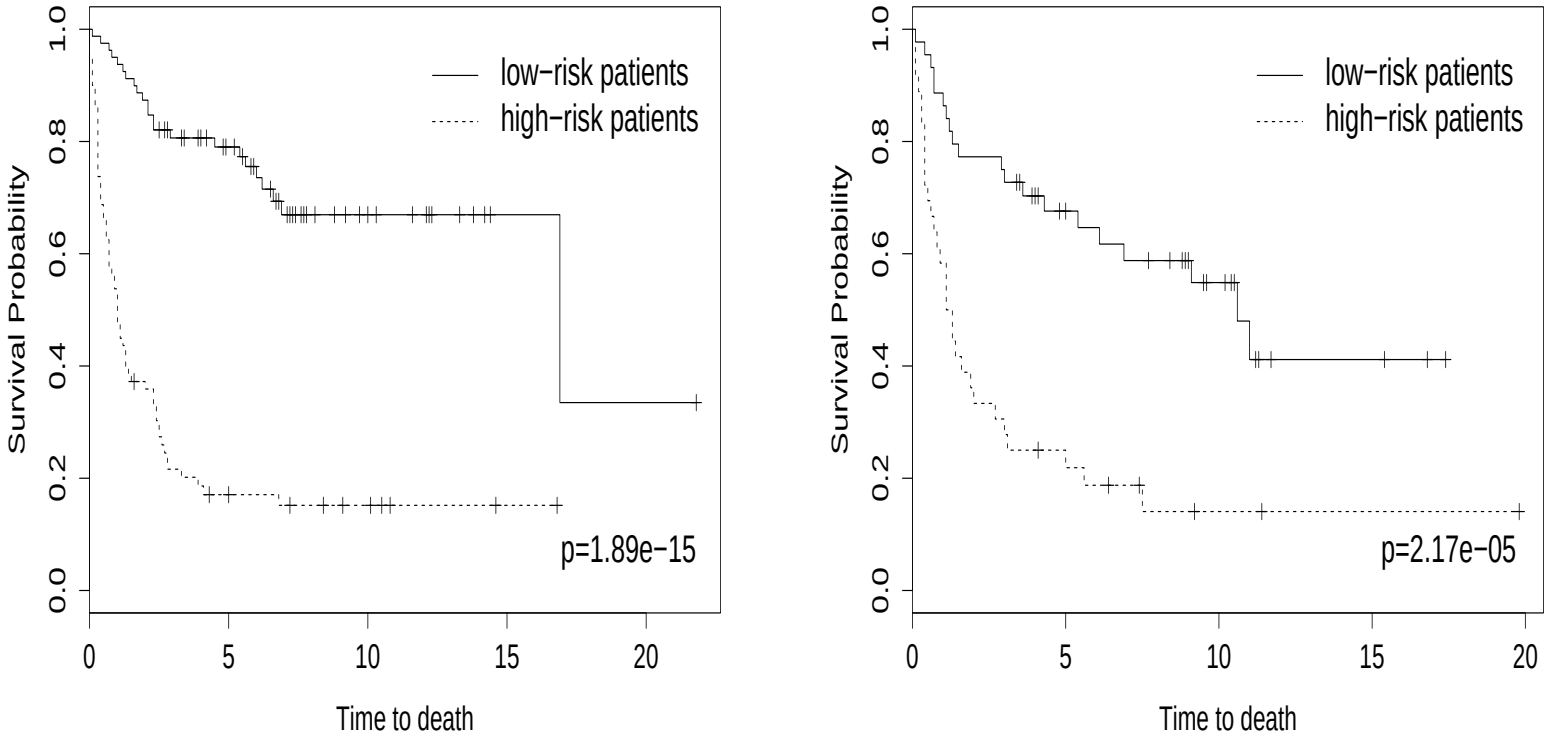

Figure 2: Survival curves for patients in two risk groups with positive and negative estimated scores. Training data (left); Testing data (right)

Area Under ROC Curve

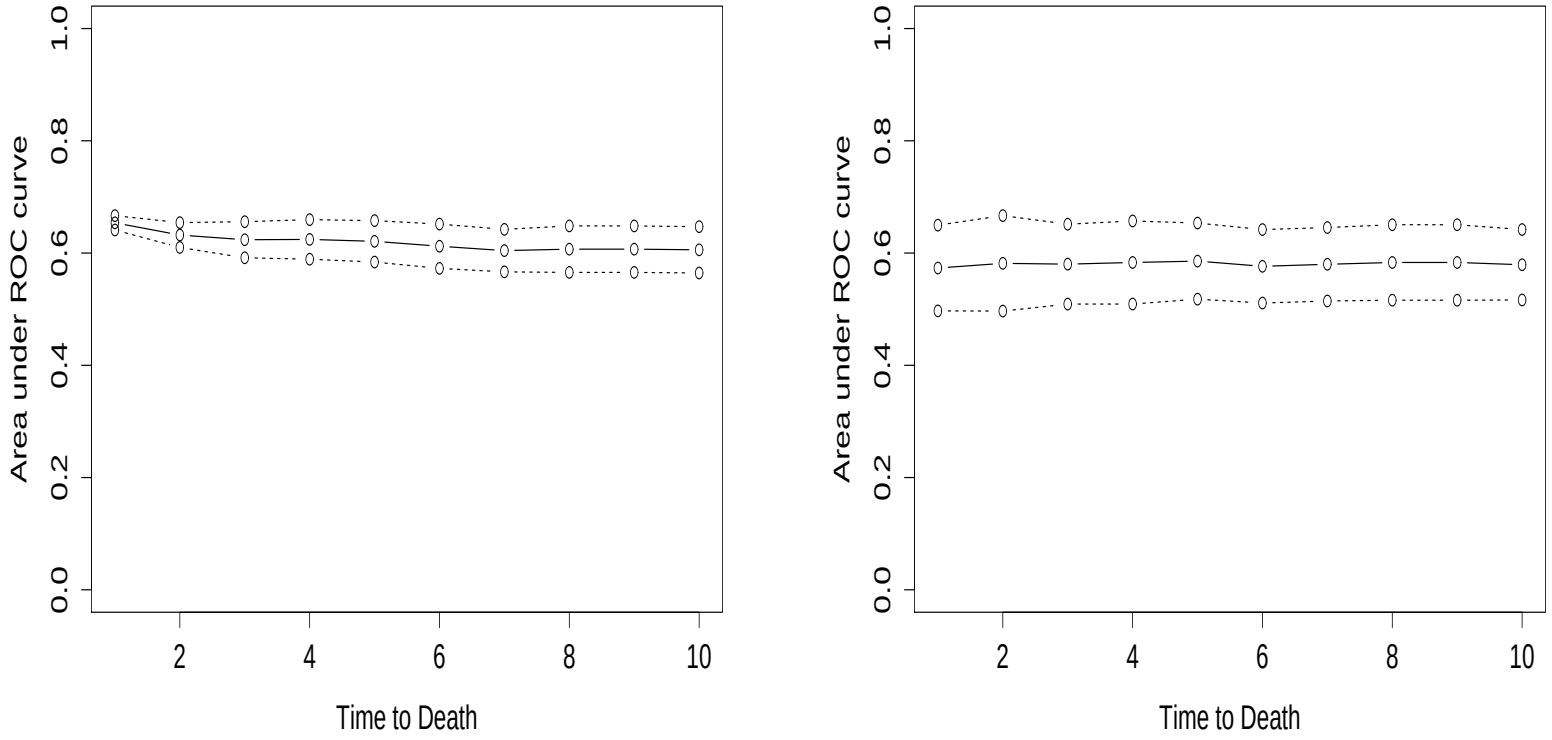

Figure 3: Area under ROC at time 1 year to 10 years for 5-fold cross-validation. Training data (left); Testing data (right)

Comparison with Existing Methods

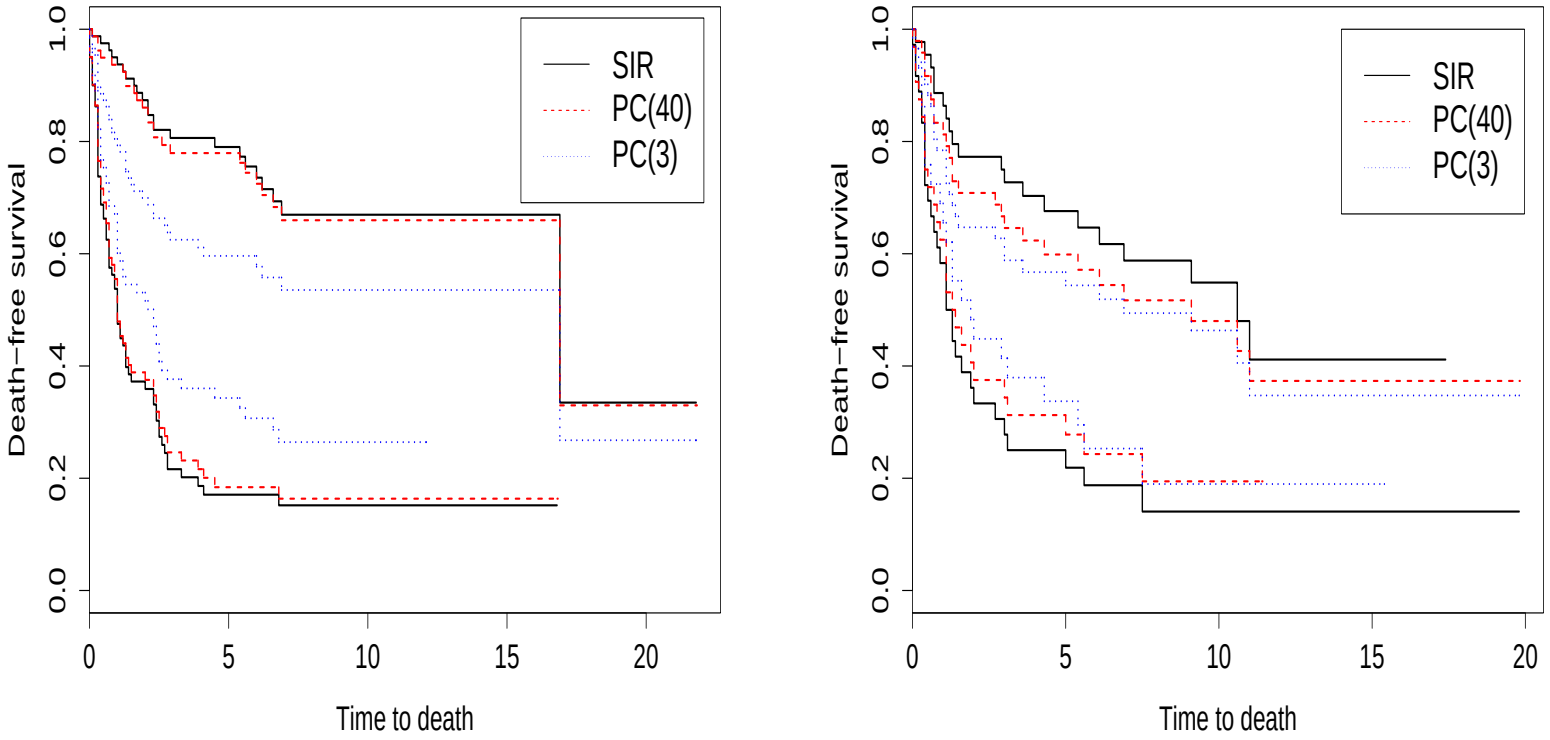

Figure 4: Comparison with principal components Cox models. Training data (left); Testing data (right)

Future Work

- Identify predictive genes based on built model
- Study prediction power by combining IPI and gene expression profiles
- Study treatment effect after adjusting for individual gene expression pattern
- Combine sufficient dimension reduction with gene networks inference

References

- Li and Li, H. (2004) *Bioinformatics*
- Sufficient dimension reduction (general):
  - Li, K-C. (1991) *JASA*
  - Cook (1998) *Regression Graphics*
- Sufficient dimension reduction for survival data:
  - Chen, Wang, and Li, K-C. (1999) *Annals of Statistics*
  - Cook (2002) *Statistics in Medicine*
- Diffuse large-B-cell lymphoma data:
  - Rosenwald et al. (2002) *NEJM*

Acknowledgements

- Joint work with Dr. Hongzhe Li
- Supported by NIH grants ES11269 (L.Li) and ES09911 (H.Li)