Dimension Reduction Methods with Application to High-dimensional Data with a Censored Response

Tuan Nguyen

under supervision of **Javier Rojo**

Pan-American Statistics Institute

May 1, 2010

SMALL n, LARGE p PROBLEM

- Dimension reduction
- Motivation: Survival analysis using microarray gene expression data
  - few patients, but 10-30K gene expression levels per patient
  - patients' survival times
- Predict patient survival taking into account microarray gene expression data

Outline

- Microarray Data
- Dimension Reduction Methods
  - Random Projection
  - Rank-based Partial Least Squares
- Application to microarray data with censored response
- Conclusions

Outline: Our Contributions

Dimension Reduction Methods

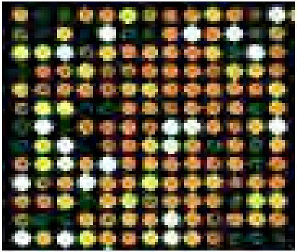
- Random Projection
  - Improvements on the lower bound for k from Johnson-Lindenstrauss (JL) Lemma
    - L_2-L_2: Gaussian and Achlioptas random matrices
    - L_2-L_1: Gaussian and Achlioptas random matrices
- Variant of Partial Least Squares (PLS): Rank-based PLS
  - insensitive to outliers
  - weight vectors as solution to optimization problem

OUTLINE

- Microarray Data
- Dimension Reduction Methods
- Small application
- Conclusions

DNA MICROARRAY

- **Traditional Methods**: one gene, one experiment
- **DNA Microarray**: thousands of genes in a single experiment
  - Interactions among the genes
  - Identification of gene sequences
  - Expression levels of genes

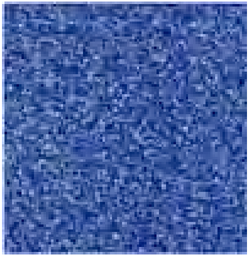


What is DNA Microarray?

- **Medium for matching** known and unknown DNA samples
- **Goal**: derive an expression level of each gene (abundance of mRNA)

Figure: Oligonucleotide array

Figure: Oligonucleotide array



Oligonucleotide Microarray

- **GeneChip (Affymetrix)**
- ~30,000 sample spots
  - each gene is represented by more than 1 spots
  - thousands of genes on array
- **Hybridization** Principle
- glowing spots: **gene expressed**

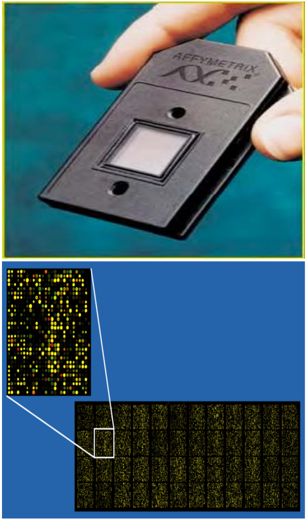


Figure: Oligonucleotide Microarray

Oligonucleotide Microarray

- Intensity of each spot is scanned
  - Expression level for each gene = total expression across all the spots
- **Multiple Samples:** 1 array - 1 sample
  - Matrix of gene expression levels

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \dots & \dots & \dots & \dots \\ X_{N1} & X_{N2} & \dots & X_{Np} \end{pmatrix}$$

Some other Arrays

- 1 flavor in past: Expression Arrays
- Innovation: many flavors
  - SNPs: single nucleotide polymorphisms within or bet. populations
  - Protein: interactions bet. protein-protein, protein-DNA/RNA, protein-drugs
  - TMAs: comparative study bet. tissue samples
  - Exon: alternative splicing and gene expression

Applications of Microarray

- Gene discovery and disease diagnosis
  - Functions of new genes
  - Inter-relationships among the genes
  - Identify genes involved in development of diseases
- **Analyses**
  - Gene selection, classification, clustering, prediction

Difficulties of Microarray

- thousands of genes, few samples ($N \ll p$)
- Survival Information
  - Observe the triple $(\mathbf{X}, \mathrm{T}, \delta)$
  - $\mathbf{X} = (x_1, \ldots, x_p)^T$ gene expression data matrix
  - $T_i = \min(y_i, c_i)$ observed survival times ($i = 1, \ldots, N$)
  - $\delta_i = I(y_i \leq c_i)$ censoring indicators

Analyzing Microarray data

2-stage procedure:

- 1) Dimension reduction methods

$$\mathbf{M}_{N \times k} = \mathbf{X}_{N \times p} \mathbf{W}_{p \times k} \, , \qquad k < N \ll p$$

- 2) Regression model

OUTLINE: OUR CONTRIBUTIONS

Dimension Reduction Methods

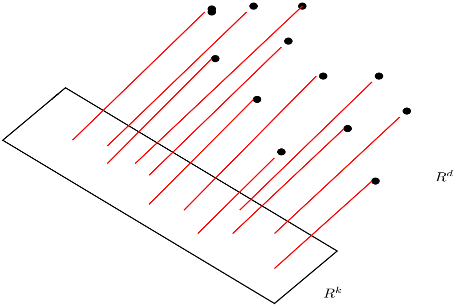
- Random Projection
  - Improvements on the lower bound for k from Johnson-Lindenstrauss (JL) Lemma
    - L_2-L_2: Gaussian and Achlioptas random matrices
    - L_2-L_1: Gaussian and Achlioptas random matrices
- Variant of Partial Least Squares (PLS): Rank-based PLS

Random Projection (RP)

The original matrix X is projected onto M by a random matrix Γ,

$$\mathbf{M}_{N\,\mathbf{x}\,k} = \mathbf{X}_{N\,\mathbf{x}\,p}\ \ \Gamma_{p\,\mathbf{x}\,k} \tag{1}$$

- Johnson-Lindenstrauss Lemma (1984)
  - preserve pairwise dist. among the points (within $1 \pm \epsilon$)
  - k cannot be too small



RP: Johnson-Lindenstrauss (JL) Lemma (1984)

Johnson-Lindenstrauss Lemma

For any $0 < \epsilon < 1$ and integer n, let $k = O(\ln n/\epsilon^2)$. For any set V of n points in $\mathbf{R}^p$, there is a linear map $f : \mathbf{R}^p \to \mathbf{R}^k$ such that for any $\mathbf{u}, \mathbf{v} \in V$,

$$(1 - \epsilon) \left|\left|\mathbf{u} - \mathbf{v}\right|\right|^2 \leq \left|\left|\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v})\right|\right|^2 \leq (1 + \epsilon) \left|\left|\mathbf{u} - \mathbf{v}\right|\right|^2 .$$

- n points in $p-$dimensional space can be projected onto a $k-$dimensional space such that the pairwise distance bet. any 2 points is preserved within $(1 \pm \epsilon)$.
- Euclidean distance in both original and projected spaces (L_2-L_2 projection)
- f is linear, but not specified

JL Lemma: Improvement on lower bound for k

for $\mathbf{x} = \mathbf{u} - \mathbf{v} \in \mathbf{R}^p$

- L_2 distance: $||\mathbf{x}|| = \sqrt{\sum_{j=1}^{p} x_i^2}$
- L_1 distance: $||\mathbf{x}||_1 = \sum_{j=1}^{p} |x_i|$

Γ is of dimension p by k

- Frankl and Maehara (1988):

$$k \geq \left\lceil \frac{27 \ln n}{3\epsilon^2 - 2\epsilon^3} \right\rceil + 1 \tag{2}$$

- Indyk and Motwani (1998):
  - entries to Γ are i.i.d. $N(0, 1)$

JL Lemma: Improvement on lower bound for k

Compare

Dasgupta and Gupta (2003): mgf technique

For any $0 < \epsilon < 1$ and integer n, let k be such that

$$k \geq \frac{24 \ln n}{3\epsilon^2 - 2\epsilon^3}. \tag{3}$$

For any set V of n points in $\mathbf{R}^p$, there is a linear map $f: \mathbf{R}^p \to \mathbf{R}^k$ such that for any $\mathbf{u}, \mathbf{v} \in V$,

$$P\left[(1-\epsilon)\,||\mathbf{u}-\mathbf{v}||^2 \leq ||\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v})||^2 \leq (1+\epsilon)\,||\mathbf{u}-\mathbf{v}||^2\right] \geq 1 - \frac{2}{n^2}.$$

- $\mathbf{f}(\mathbf{x}) = \frac{1}{\sqrt{k}}\mathbf{x}\Gamma$, where $\mathbf{x} = \mathbf{u} - \mathbf{v}$
- Gaussian random matrix: $k\frac{||\mathbf{f}(\mathbf{x})||^2}{||\mathbf{x}||^2} \sim \chi_k^2$
- best available lower bound for k

JL Lemma: Dasgupta and Gupta

Considering the tail probabilities separately

$$P\left[\left|\left|\mathbf{f}(\mathbf{x})\right|\right|^2 \geq (1+\epsilon)\left|\left|\mathbf{x}\right|\right|^2\right] \leq \frac{1}{n^2}, \tag{4}$$

$$P\left[\left|\left|\mathbf{f}(\mathbf{x})\right|\right|^2 \leq (1-\epsilon)\left|\left|\mathbf{x}\right|\right|^2\right] \leq \frac{1}{n^2}. \tag{5}$$

- $\mathbf{f}(\mathbf{x}) = \frac{1}{\sqrt{k}}\mathbf{x}\Gamma$, where $\mathbf{x} = \mathbf{u} - \mathbf{v}$

$$k\frac{\left|\left|\mathbf{f}(\mathbf{x})\right|\right|^2}{\left|\left|\mathbf{x}\right|\right|^2} \sim \chi_k^2 \tag{6}$$

- using mgf technique to bound the tail probabilities

JL Lemma: Dasgupta and Gupta

Sketch of Proof

- define $\mathbf{f}(\mathbf{x}) = \frac{1}{\sqrt{k}}\mathbf{x}\Gamma$, and $\mathbf{y} = \sqrt{k}\frac{\mathbf{f}(\mathbf{x})}{||\mathbf{x}||}$
- Let $y_j = \frac{\mathbf{x}r_j}{||\mathbf{x}||} \sim N(0,1)$, and $y_j^2 \sim \chi_1^2$ with $E(||\mathbf{y}||^2) = k$
- Let $\alpha_1 = k(1+\epsilon)$, and $\alpha_2 = k(1-\epsilon)$, then

JL Lemma: Dasgupta and Gupta

Sketch of Proof

- Right-tail prob:

$$
\begin{aligned}
P\left[||\mathbf{f}(\mathbf{x})||^2 \geq (1+\epsilon)\,||\mathbf{x}||^2\right] &= P\left[||\mathbf{y}||^2 \geq \alpha_1\right] \\
&\leq \left(e^{-s(1+\epsilon)} E\left(e^{s y_j^2}\right)\right)^k, \quad s > 0 \\
&= e^{-s\alpha_1}(1-2s)^{-k/2}, \quad s \in (0, 1/2).
\end{aligned}
$$

- Left-tail prob:

$$
\begin{aligned}
P\left[||\mathbf{f}(\mathbf{x})||^2 \leq (1-\epsilon)\,||\mathbf{x}||^2\right] &= P\left[||\mathbf{y}||^2 \leq \alpha_2\right] \\
&\leq \left(e^{s(1-\epsilon)} E\left(e^{-s y_j^2}\right)\right)^k, \quad s > 0 \\
&= e^{s\alpha_2}(1+2s)^{-k/2} \\
&\leq e^{-s\alpha_1}(1-2s)^{-k/2}, \quad s \in (0, 1/2).
\end{aligned}
$$

JL LEMMA: DASGUPTA AND GUPTA

Sketch of Proof

- Minimize $e^{-s(1+\epsilon)}(1-2s)^{-1/2}$ wrt s, with $s^* = \frac{1}{2}\left(\frac{\epsilon}{1+\epsilon}\right)$,

$$P\left[||\mathbf{y}||^2 \geq \alpha_1\right] \leq exp\left(-\frac{k}{2}(\epsilon - \ln(1+\epsilon))\right)$$
$$\leq exp\left(-\frac{k}{12}(3\epsilon^2 - 2\epsilon^3)\right) \tag{7}$$

- if $k \geq \frac{24\ln n}{3\epsilon^2 - 2\epsilon^3}$, then

$$P\left[||\mathbf{f}(\mathbf{x})||^2 \geq (1+\epsilon)\,||\mathbf{x}||^2\right] \leq 1/n^2$$

and

$$P\left[||\mathbf{f}(\mathbf{x})||^2 \leq (1-\epsilon)\,||\mathbf{x}||^2\right] \leq 1/n^2$$

JL Lemma

Projection of all $\binom{n}{2}$ pairs of distinct points

- Results are given in terms of preserving distances between 1 pair of points
  - lower bound for probability chosen to be $1 - 2/n^2$
- Interest: simultaneously preserve distances among all $\binom{n}{2}$ pairs of distinct points

JL Lemma

Projection of all $\binom{n}{2}$ pairs of distinct points

$$P\left\{ \bigcap_{\substack{\mathbf{u},\mathbf{v}\in V \\ \mathbf{u}\neq\mathbf{v}}} (1-\epsilon)\,||\mathbf{u}-\mathbf{v}||^2 \leq ||\mathbf{f}(\mathbf{u})-\mathbf{f}(\mathbf{v})||^2 \leq (1+\epsilon)\,||\mathbf{u}-\mathbf{v}||^2 \right\} \tag{8}$$

$$\geq 1 - \sum_{\substack{\mathbf{u},\mathbf{v}\in V \\ \mathbf{u}\neq\mathbf{v}}} P\left[\left\{(1-\epsilon)\,||\mathbf{u}-\mathbf{v}||^2 \leq ||\mathbf{f}(\mathbf{u})-\mathbf{f}(\mathbf{v})||^2 \leq (1+\epsilon)\,||\mathbf{u}-\mathbf{v}||^2 \right\}^c\right]$$

$$\geq 1 - \binom{n}{2}\frac{2}{n^2} = \frac{1}{n}.$$

- Introduce $\beta > 0$ (Achlioptas (2001)) so that for each pair $\mathbf{u}, \mathbf{v} \in V$,

$$P\left[(1-\epsilon)\,||\mathbf{u}-\mathbf{v}||^2 \leq ||\mathbf{f}(\mathbf{u})-\mathbf{f}(\mathbf{v})||^2 \leq (1+\epsilon)\,||\mathbf{u}-\mathbf{v}||^2\right] \geq 1 - 2/n^{2+\beta}.$$

- prob. in Eq. (8) is bounded from below by $1 - 1/n^\beta$

Improvement on Dasgupta and Gupta Bound

Rojo and Nguyen (2009a) (*NR_1* Bound): k is smallest even integer satisfying

$$\left(\frac{1+\epsilon}{\epsilon}\right) g(k,\epsilon) \leq \frac{1}{n^{2+\beta}} \tag{9}$$

- $g(k,\epsilon) = e^{-\lambda_1}\frac{\lambda_1^{d-1}}{(d-1)!}$ is decreasing in k
- $\lambda_1 = k(1+\epsilon)/2$ and $d = k/2$.
- Gaussian random matrix
- work directly with the exact distribution of $k\frac{||\mathbf{f}(\mathbf{x})||^2}{||\mathbf{x}||^2}$, where $\mathbf{x} = \mathbf{u} - \mathbf{v}$
- 12%-34% improvement

Outline of Proof

- **Gamma-Poisson Relationship:** Suppose $X \sim Gamma(d, 1)$, for $d = 1, 2, 3, \ldots$, and $Y \sim Poisson(x)$, then

$$P(X \geq x) = \int_x^\infty \frac{1}{\Gamma(d)} z^{d-1} e^{-z} dz = \sum_{y=0}^{d-1} \frac{x^y e^{-x}}{y!} = P(Y \leq d-1) \tag{10}$$

- $||\mathbf{y}||^2 = \sum_{j=1}^{k} y_j^2 \sim \chi_k^2$
- Let $d = k/2$, $\alpha_1 = k(1+\epsilon)$, and $\alpha_2 = k(1-\epsilon)$, then,
- **Right tail prob.:**

$$P[||\mathbf{y}||^2 \geq \alpha_1] = \int_{\alpha_1/2}^\infty \frac{1}{\Gamma(a)} z^{d-1} e^{-z} dz = \sum_{y=0}^{d-1} \frac{(\alpha_1/2)^y e^{-\alpha_1/2}}{y!}$$

- **Left tail prob.:**

$$P[||\mathbf{y}||^2 \leq \alpha_2] = \int_0^{\alpha_2/2} \frac{1}{\Gamma(a)} z^{d-1} e^{-z} dz = \sum_{y=d}^{\infty} \frac{(\alpha_2/2)^y e^{-\alpha_2/2}}{y!}$$

Outline of Proof

Theorem 1: Given d as a positive integer

a) Suppose $1 \leq d < \lambda_1$, then

$$\sum_{y=0}^{d-1} \frac{\lambda_1^y}{y!} \leq \left(\frac{\lambda_1}{\lambda_1 - d}\right)\left(\frac{\lambda_1^{d-1}}{(d-1)!}\right) \tag{11}$$

b) Suppose $0 < \lambda_2 < d$, then

$$\sum_{y=d}^{\infty} \frac{\lambda_2^y}{y!} \leq \left(\frac{\lambda_2}{d - \lambda_2}\right)\left(\frac{\lambda_2^{d-1}}{(d-1)!}\right) \tag{12}$$

Improvement on Dasgupta and Gupta Bound

back

Right-tail prob.

$$
\begin{aligned}
P[||\mathbf{y}||^2 \geq \alpha_1] &= e^{-\lambda_1} \sum_{y=0}^{d-1} \frac{\lambda_1^y}{y!} \\
&\leq \left(\frac{1+\epsilon}{\epsilon}\right)\left(\frac{\lambda_1^{d-1}}{(d-1)!}\right) e^{-\lambda_1} \qquad (13)
\end{aligned}
$$

Left-tail prob.

$$
\begin{aligned}
P[||\mathbf{y}||^2 \leq \alpha_2] &= e^{-\lambda_2} \sum_{y=d}^{\infty} \frac{\lambda_2^y}{y!} \\
&\leq \left(\frac{1-\epsilon}{\epsilon}\right)\left(\frac{\lambda_2^{d-1}}{(d-1)!}\right) e^{-\lambda_2} \\
&\leq \left(\frac{1+\epsilon}{\epsilon}\right)\left(\frac{\lambda_1^{d-1}}{(d-1)!}\right) e^{-\lambda_1}
\end{aligned}
$$

- lower bound for k:
  $P[||\mathbf{y}||^2 \geq \alpha_1] + P[||\mathbf{y}||^2 \leq \alpha_2] \leq 2e^{-\lambda_1}\left(\frac{1+\epsilon}{\epsilon}\right)\left(\frac{\lambda_1^{d-1}}{(d-1)!}\right) \leq 2/n^{2+\beta}$

COMPARISON OF THE LOWER BOUNDS ON k

Prob.

Table: L_2-L_2 distance: NR_1 Bound, and DG Bound.

N(0,1) entries		NR_1 Bound	**DG Bound**	% **Improv.**
n=10	$\epsilon = .1, \beta = 1$	2058	2961	30
	$\epsilon = .3, \beta = 1$	254	384	34
	$\epsilon = .1, \beta = 2$	2962	3948	25
	$\epsilon = .3, \beta = 2$	368	512	28
n=50	$\epsilon = .1, \beta = 1$	3976	5030	21
	$\epsilon = .3, \beta = 1$	494	653	24
	$\epsilon = .1, \beta = 2$	5572	6707	17
	$\epsilon = .3, \beta = 2$	692	870	20
n=100	$\epsilon = .1, \beta = 1$	4822	5921	19
	$\epsilon = .3, \beta = 1$	598	768	22
	$\epsilon = .1, \beta = 2$	6716	7895	15
	$\epsilon = .3, \beta = 2$	834	1024	19
n=500	$\epsilon = .1, \beta = 1$	6808	7991	15
	$\epsilon = .3, \beta = 1$	846	1036	18
	$\epsilon = .1, \beta = 2$	9390	10654	12
	$\epsilon = .3, \beta = 2$	1168	1382	15

Extension of JL Lemma to L_1 norm

- JL Lemma (L_2-L_2 projection): space of original points in L_2, and space of projected points in L_2.
- L_1-L_1 Random Projection
  - JL Lemma cannot be extended to the L_1 norm
    - Charikar & Sahai (2002), Brinkman & Charikar (2003), Lee & Naor (2004), Indyk (2006)

Outline: Our Contributions

Dimension Reduction Methods

- Random Projection
  - Improvements on the lower bound for k from Johnson-Lindenstrauss (JL) Lemma
    - L_2-L_2: Gaussian and Achlioptas random matrices
    - L_2-L_1: Gaussian and Achlioptas random matrices
- Variant of Partial Least Squares (PLS): Rank-based PLS

Extension of JL Lemma to L_1 norm

- L_2-L_1 Random Projection: space of original points in L_2, and space of projected points in L_1.
- Ailon & Chazelle (2006) and Matousek (2007)
  - the original L_2 pair-wise distances are preserved within $(1 \pm \epsilon)\sqrt{2/\pi}$ of the projected L_1 distances with k,

$$k \geq C\epsilon^{-2}(2\ln(1/\delta)) \tag{14}$$

- $\delta \in (0, 1)$, $\epsilon \in (0, 1/2)$, C sufficiently large
- when $\delta = 1/n^{2+\beta}$, then $k = O((4 + 2\beta)\ln n/\epsilon^2)$
- sparse Gaussian random matrix (Ailon & Chazelle, and Matousek), sparse Achlioptas random matrix (Matousek)

L_2-L_1 RP: Improvement to Ailon & Chazelle, and Matousek bound

▸ Proof ▸ Ach

Rojo and Nguyen (2009a) NR_2 Bound

$$k \geq \frac{(2+\beta)\ln n}{-\ln(A(s_\epsilon^*))} \tag{15}$$

- $A(s) = 2e^{-s\sqrt{2/\pi}(1+\epsilon)+s^2/2}\Phi(s)$, $s > 0$
- s_ϵ^* is unique minimizer of $A(s)$
- based on mgf technique
- Gaussian and Achlioptas random matrix
- 36%-40% improvement

Outline of Proof

back

- Define $\mathbf{f}(\mathbf{x}) = \frac{1}{k}\mathbf{x}\Gamma$,
- Let $y_j = \frac{\mathbf{x}r_j}{||\mathbf{x}||_2} \sim N(0,1)$, then $E(||\mathbf{y}||_1) = k\sqrt{2/\pi}$ and $M_{|y_j|}(s) = 2e^{s^2/2}\Phi(s), \forall s$.
- Let $\alpha_1 = k\sqrt{2/\pi}(1+\epsilon)$, and $\alpha_2 = k\sqrt{2/\pi}(1-\epsilon)$, then
- **Right tail prob.:**

$$P[||\mathbf{y}||_1 \geq \alpha_1] \leq \left(2e^{-(s\alpha_1/k)+(s^2/2)}\Phi(s)\right)^k, \quad s > 0.$$

- **Left tail prob.:**

$$\begin{aligned} P[||\mathbf{y}||_1 \leq \alpha_2] &\leq \left(2e^{(s\alpha_2/k)+(s^2/2)}(1-\Phi(s))\right)^k, \quad s > 0 \\ &\leq \left(2e^{-(s\alpha_1/k)+(s^2/2)}\Phi(s)\right)^k, \quad s > 0. \end{aligned} \tag{16}$$

- Eq. (16) is obtained since $e^{2s\sqrt{2/\pi}} < \frac{\Phi(s)}{1-\Phi(s)}$, $s > 0$.
- minimize Eq. (16) wrt s, and plug s^* to get the bound.

Comparison of the lower bounds on k

Table: L_2-L_1 distance: Matousek bound ($C = 1$), and NR_2 Bound.

N(0,1) entries		**Matousek**	NR_2 **Bound**	% **Improv.**
n=10	$\epsilon = .1, \beta = 1$	1382	823	40
	$\epsilon = .3, \beta = 1$	154	99	36
	$\epsilon = .1, \beta = 2$	5527	3290	40
	$\epsilon = .3, \beta = 2$	615	394	36
n=50	$\epsilon = .1, \beta = 1$	2348	1398	40
	$\epsilon = .3, \beta = 1$	261	168	36
	$\epsilon = .1, \beta = 2$	3130	1863	40
	$\epsilon = .3, \beta = 2$	348	223	36
n=100	$\epsilon = .1, \beta = 1$	2764	1645	40
	$\epsilon = .3, \beta = 1$	308	197	36
	$\epsilon = .1, \beta = 2$	3685	2193	40
	$\epsilon = .3, \beta = 2$	410	263	36
n=500	$\epsilon = .1, \beta = 1$	3729	2220	40
	$\epsilon = .3, \beta = 1$	415	266	36
	$\epsilon = .1, \beta = 2$	4972	2960	40
	$\epsilon = .3, \beta = 2$	553	354	36

Outline: Our Contributions

Dimension Reduction Methods

- Random Projection
  - Improvements on the lower bound for k from Johnson-Lindenstrauss (JL) Lemma
    - L_2-L_2: Gaussian and Achlioptas random matrices
    - L_2-L_1: Gaussian and Achlioptas random matrices
- Variant of Partial Least Squares (PLS): Rank-based PLS

Achlioptas-Typed Random Matrices

- Achlioptas (2001): for $q = 1$ or $q = 3$

$$r_{ij} = \sqrt{q} \begin{cases} +1 & \text{with prob. } \frac{1}{2q} \\ 0 & \text{with prob. } 1 - \frac{1}{q} \\ -1 & \text{with prob. } \frac{1}{2q}. \end{cases} \tag{17}$$

- L_2-L_2 projection

- same lower bound for k as in the case of Gaussian random matrix

$$k \geq \frac{(24 + 12\beta) \ln n}{3\epsilon^2 - 2\epsilon^3}. \tag{18}$$

ACHLIOPTAS

- Define $\mathbf{f}(\mathbf{x}) = \frac{1}{\sqrt{k}}\mathbf{x}\Gamma$, and $y_j = \frac{\mathbf{x}r_j}{||\mathbf{x}||_2}$
- **Goal**: bound mgf of y_j^2
- bounding the even moments of y_j by the even moments in the case $r_{ij} \sim N(0, 1)$
  - $\Longrightarrow$ mgf of y_j^2 using r_{ij} from Achlioptas proposal is bounded by mgf using $r_{ij} \sim N(0, 1)$.

L_2-L_2 RP: Improvement on Achlioptas bound

- Rademacher random matrix Eq. (17) with $q = 1$
- $r_{ij}^2 = 1 \quad$ and $\quad r_{lj} r_{mj} \stackrel{D}{=} r_{ij} \quad$ independent

Rojo and Nguyen (2009b)

- Let $y_j = \sum_{i=1}^{p} c_i r_{ij}$, with $c_i = \frac{x_i}{||\mathbf{x}||_2}$, then

$$k\left(\frac{||\mathbf{f}(\mathbf{x})||^2}{||\mathbf{x}||^2}\right) = \sum_{j=1}^{k} y_j^2 \stackrel{D}{=} k + 2\sum_{j=1}^{k}\sum_{l=1}^{p}\sum_{m=l+1}^{p} c_{lm} r_{lmj} \qquad (19)$$

  with $c_{lm} = c_l c_m$, and $r_{lmj} = r_{lj} r_{mj}$

- 3 improvements on Achlioptas bound

L_2-L_2 RP: Improvement on Achlioptas bound

Rojo and Nguyen (2009b): Method 1

- Hoeffding's inequality based on mgf

Let U_i's be independent and bounded random variables such that U_i falls in the interval $[a_i, b_i]$ $(i = 1, \ldots, n)$ with prob. 1. Let $S_n = \sum_{i=1}^n U_i$, then for any $t > 0$,

$$P[S_n - E(S_n) \geq t] \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$$

and

$$P[S_n - E(S_n) \leq t] \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$$

- Method 1: lower bound for k

$$k \geq \left(\frac{(8 + 4\beta)\ln n}{\epsilon^2}\right)\left(\frac{p-1}{p}\right). \tag{20}$$

Outline of Proof

- **Right tail prob.:**

$$P[||\mathbf{y}||^2 \geq k(1+\epsilon)] = P\left[\sum_{j=1}^{p}\sum_{l=1}^{p}\sum_{m=l+1}^{p} c_{lm} r_{lmj} \geq \frac{k\epsilon}{2}\right]$$

$$\leq exp\left(-\frac{k\epsilon^2}{8\sum_{l=1}^{p}\sum_{m=l+1}^{p} c_l^2 c_m^2}\right). \quad (21)$$

- **Left tail prob.:**

$$P[||\mathbf{y}||^2 \leq k(1-\epsilon)] = P\left[\sum_{j=1}^{p}\sum_{l=1}^{p}\sum_{m=l+1}^{p} c_{lm} r_{lmj} \leq -\frac{k\epsilon}{2}\right]$$

$$\leq exp\left(-\frac{k\epsilon^2}{8\sum_{l=1}^{p}\sum_{m=l+1}^{p} c_l^2 c_m^2}\right). \quad (22)$$

Outline of Proof

- $\sum_{l=1}^{p} \sum_{m=l+1}^{p} c_l^2 c_m^2$ is max. at $c_l = c_m = 1/p$

$$4 \sum_{l=1}^{p} \sum_{m=l+1}^{p} c_l^2 c_m^2 \leq 2 \left(\frac{p-1}{p} \right)$$

- **Right tail prob.:**

$$P[||\mathbf{y}||^2 \geq k(1+\epsilon)] \leq exp \left(-\frac{k\epsilon^2}{4} \left(\frac{p-1}{p} \right) \right)$$

- **Left tail prob.:**

$$P[||\mathbf{y}||^2 \leq k(1-\epsilon)] \leq exp \left(-\frac{k\epsilon^2}{4} \left(\frac{p-1}{p} \right) \right)$$

L_2-L_2 RP: Improvement on Achlioptas bound

Rojo and Nguyen (2009b): Method 2

- Berry-Esseen inequality

Let $X_1, \ldots, X_m$ be i.i.d. random variables with $E(X_i) = 0$, $\sigma^2 = E(X_i^2) > 0$, and $\rho = E|X_i|^3 < \infty$. Also, let $\bar{X_m}$ be the sample mean, and F_m the cdf of $\bar{X_m}\sqrt{n}/\sigma$. Then for all x and m, there exists a positive constant C such that

$$|F_m(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{m}}$$

- $\rho/\sigma^3 = 1$, and $C = 0.7915$ for ind. X_i's (Siganov)
- Method 2: lower bound for k: 10%-30% improvement

$$1 - \Phi\left(\epsilon\sqrt{\frac{kp}{2(p-1)}}\right) + \frac{0.7915}{\sqrt{kp(p-1)/2}} \leq 1/n^{2+\beta}. \qquad (23)$$

L_2-L_2 RP: Improvement on Achlioptas bound

Rojo and Nguyen (2009b): Method 3

- Pinelis inequality

Let U_i's be independent rademacher random variables. Let $d_1, \ldots, d_m$ be real numbers such that $\sum_{i=1}^{m} d_i^2 = 1$. Let $S_m = \sum_{i=1}^{m} d_i U_i$, then for any $t > 0$,

$$P[|S_m| \geq t] \leq \min\left(\frac{1}{t^2}, 2(1 - \Phi(t - 1.495/t))\right)$$

- Method 3: lower bound for k: 15% improvement ($\epsilon = 0.1$)

$$k \geq \frac{2(p-1)a_n^2}{p\epsilon^2}, \tag{24}$$

where $a_n = \frac{Q_n + \sqrt{Q_n^2 + 4(1.495)}}{2}$, and $Q_n = \Phi^{-1}\left(1 - \frac{1}{n^{2+\beta}}\right)$.

Comparison of the lower bounds on k: Rademacher random matrix

▸ Simulations

Table: L_2-L_2 distance with $\epsilon = 0.1$: 1) Method 1 (based on Hoeffding's inequality), 2) Method 2 (using Berry-Esseen inequality), 3) Method 3 (based on Pinelis inequality), and 4) Achlioptas Bound (does not depend on p). Last column is the % improv. of Method 2 on Achlioptas Bound.

	β	p	Method 1	Method 3	Method 2	Achlioptas	% Improv.
$n = 10$	0.5	5000	2303	2045	1492		40
		10000	2303	2046	1492	2468	40
$n = 10$	1	5000	2763	2472	1912		35
		10000	2763	2472	1911	2961	35
$n = 50$	0.5	5000	3912	3553	3009		28
		10000	3912	3554	2995	4192	29
$n = 50$	1	5000	4694	4300	3948		22
		10000	4694	4300	3821	5030	24
$n = 100$	0.5	5000	4605	4214	3809		23
		10000	4605	4215	3715	4935	25
$n = 100$	1	30000	5527	5100	4816		19
		70000	5527	5100	4623	5921	22

L_2-L_1: Achlioptas Random Matrix

Gaussian

- same lower bound as in L_2-L_1 case using Gaussian Random Matrix (NR_2 bound)
- bounding the mgf of Achlioptas-typed r.v. by that of a standard Gaussian
- 36%-40% improvement

SUMMARY

Random Projection

- Gaussian random matrix
  - L_2-L_2 projection: improvement of 12%-34% on Dasgupta and Gupta bound
  - L_2-L_1 projection: improvement of 36%-40% on Chazelle & Ailon bound
- Achlioptas-typed random matrix
  - L_2-L_2 projection: improvement of 20%-40% on Achlioptas bound (Rademacher random matrix)
  - L_2-L_1 projection: improvement of 36%-40% on Matousek bound
- lower bound for k is still large for practical purposes: active research

RP vs. PCA, PLS, ...

Random Projection (RP)

- criterion: preserve pairwise distance (within $1 \pm \epsilon$)
- k is large

PCA, PLS

- optimization criterion

Principal Component Analysis (PCA)

- Karl Pearson (1901)

- Variance optimization criteria,

$$w_k = \underset{w'w=1}{\arg\max}\ \mathrm{Var}(Xw) = \underset{w'w=1}{\arg\max}(N-1)^{-1}w'X'Xw$$

s.t. $w_k' X' X w_j = 0$ for all $1 \le j < k$.

- ignores response y
- eigenvalue decomposition of sample cov. matrix

PCA

- Sample covariance matrix $S = (N-1)^{-1}X'X$
- Eigenvalue decomposition: $S = V\Delta V'$
  - $\Delta = \text{diag}(\lambda_1 \geq \cdots \geq \lambda_N)$ eigenvalues
  - $V = (v_1, \ldots, v_N)$ unit eigenvectors
- weight vectors $w_k = v_k$
- PCs are $M_k = Xw_k$, $k = 1, \ldots, N$
- Cumulative variation explained by the 1*st* K PCs is $\sum_{k=1}^{K} \lambda_k$

PCA

How to Select K: number of PC's

- Proportion of Variation Explained
- Cross-validation

Partial Least Squares (PLS)

▸ PCA

- Herman Wold (1960)

- Covariance optimization criteria,

$$w_k = \underset{w'w=1}{\arg\max}\ \mathrm{Cov}(Xw, y) = \underset{w'w=1}{\arg\max} (N-1)^{-1} w'X'y$$

s.t. $w_k' X' X w_j = 0$ for all $1 \leq j < k$.

- incorporates both the covariates and response
- ignores censoring

Incorporating Censoring in PLS

- Nguyen and Rocke (2002):
  - PLS weights: $w_k = \sum_{i=1}^{N} \theta_{ik} v_i$, where v_i eigenvectors of $X'X$
  - θ_{ik} depend on y only through $a_i = u_i' y$, where u_i eigenvectors of XX'
  - a_i is slope coeff. of simple linear regression of y on u_i if X is centered.
- Nguyen and Rocke: Modified PLS (MPLS)
  - replace a_i by slope coeff. from Cox PH regression of y on u_i to incorporate censoring PLS

PLS: Nonparametric Approaches to Incorporate Censoring

- Datta (2007):
  - Reweighting: Inverse Probability of Censoring Weighted ▸ RWPLS
    - replace censored response with 0
    - reweigh the uncensored response by the inverse prob. that it corresponds to a censored obs.
  - Mean Imputation: ▸ MIPLS
    - keep the uncensored response
    - replace the censored response by its expected value given that the true survival time exceeds the censoring time

Outline: Our Contributions

Dimension Reduction Methods

- Random Projection
- Variant of Partial Least Squares (PLS):
  Rank-based PLS
  - insensitive to outliers
  - weight vectors as solution to optimization problem

NOTATIONS:

for vectors $u = (u_1, \ldots, u_n)'$ and $v = (v_1, \ldots, v_n)'$

- Ranks of u_i's: indices of positions in ascending or descending order
- sample Pearson correlation coeff.

$$Cor(u, v) = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2}\sqrt{\sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

- sample Spearman correlation coeff.: corr. on the ranks

Nguyen and Rojo (2009a, b): Rank-based Partial Least Squares (RPLS)

- cov/corr. measure in PLS is influenced by outliers
- replace Pearson corr. by Spearman rank corr.

$$w_k = \underset{w'w=1}{\arg\max}(N-1)^{-1} w' R_X' R_y \tag{25}$$

- use Nguyen and Rocke's procedure (MPLS) and Datta's RWPLS and MIPLS to incorporate censoring

RPLS: Derivation of the weights

$$w_1 = \frac{R_X' R_y}{||R_X' R_y||} \tag{26}$$

$$w_k \propto P_{k-1} w_1, \; k \geq 2 \tag{27}$$

where

$$P_{k-1} = I - \zeta_1 S_{R_X} - \zeta_2 S_{R_X}^2 - \cdots - \zeta_{k-1} S_{R_X}^{k-1} \tag{28}$$

with $S_{R_X}^j = \underbrace{S_{R_X} S_{R_X} \ldots S_{R_X}}_{j \text{ times}}$, $\qquad S_{R_X} = R_X' R_X$, $S_X = X'X$, and

ζ's obtained from

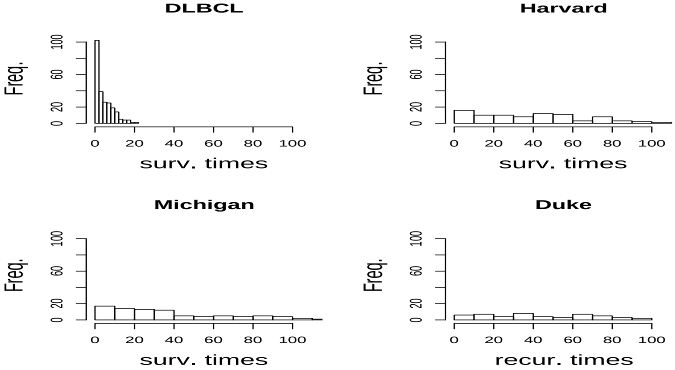
$$\begin{aligned} w_1' P_{k-1} S_X w_1 &= 0 \\ w_1' P_{k-1} S_X w_2 &= 0 \\ \ldots & \\ w_1' P_{k-1} S_X w_{k-1} &= 0 \end{aligned}$$

OUTLINE

- Microarray Data
- Dimension Reduction Methods
- Small application to Microarray data with censored response
- Conclusions

Real Datasets

▸ Simulations



- **Diffuse Large B-cell Lymphoma (DLBCL):** 240 cases, 7399 genes, 42.5% cens.
- **Harvard Lung Carcinoma:** 84 cases, 12625 genes, 42.9% cens.
- **Michigan Lung Adenocarcinoma:** 86 cases, 7129 genes, 72.1% cens.
- **Duke Breast Cancer:** 49 cases, 7129 genes, 69.4% cens.

Cox Proportional Hazards (PH) Model

$$h(t; z_i, \beta) = h_0(t)e^{z_i'\beta} \tag{29}$$

- h_0: unspecified baseline hazard

$$S(t; z_i, \beta) = S_0(t)^{e^{z_i'\beta}} \tag{30}$$

- incorporates the covariates and censored information
- proportional hazards assumption

Accelerated Failure Time (AFT) model

- logarithm of true survival time

$$\log(y_i) = \mu + X_i'\beta + \sigma u_i \tag{31}$$

- y_i's are true survival times
- μ and σ are location and scale parameters
- u_i's are the errors i.i.d. with some distribution

Assessment of the Methods

Selection of K

1) K is fixed across all methods

    - 1^{st} K PCs explain a certain proportion of predictor variability

2) K is chosen by Cross-validation (CV)

    - K is inherent within each method

Assessment of the Methods

Selection of K by CV

- Cox PH Model

$$CV(surv.error) = \frac{1}{sM} \sum_{i=1}^{s} \sum_{m=1}^{M} \sum_{t \in D_m} \left[\hat{\hat{S}}_{-m}(t) - \hat{\hat{S}}_m(t) \right]^2$$

  - $i = 1, \ldots, s$ is index of simulation
  - $m = 1, \ldots, M$ is index for the fold
  - D_m is the set of death times in m^{th} fold
  - $\hat{\hat{S}}_m$ denotes the est. surv. function for m^{th} fold
  - $\hat{\hat{S}}_{-m}$ denotes the est. surv. function when the m^{th} fold is removed

$$\hat{\hat{S}}_m(t) = \frac{1}{N_m} \sum_{n=1}^{N_m} \hat{S}_{m,n}(t)$$

$$\hat{\hat{S}}_{-m}(t) = \frac{1}{N_{-m}} \sum_{n=1}^{N_{-m}} \hat{S}_{-m,n}(t)$$

Assessment of the Methods

Selection of K by CV

- AFT Model

$$CV(fit.error) = \frac{1}{sM}\sum_{i=1}^{s}\sum_{m=1}^{M}\left[\frac{\sum_{l=1}^{N_m}\delta_{m,l}(i)\left(\hat{y}^*_{m,l}(i) - y^*_{m,l}(i)\right)^2}{\sum_{l=1}^{N_m}\delta_{m,l}(i)}\right]$$

  - $i = 1, \ldots, s$ is index of simulation
  - $m = 1, \ldots, M$ is index for the fold
  - $l = 1, \ldots, N_m$ is index for the individual in m^{th} fold
  - $y^*_{m,l}(i) = \ln\left(y_{m,l}(i)\right)$
  - $\hat{y}^*_{m,l}(i)$ are the estimates of $y^*_{m,l}(i)$

$$\hat{y}^*_{m,l}(i) = \hat{\mu}_{-m,AFT}(i) + M_{m,l}(i)'\hat{\beta}_{-m,AFT}(i)$$

PCA, MPLS, RWPLS, MIPLS, RMPLS

- Principal Component Analysis (PCA):
  - variance optimization criterion
  - ignores response
- Modified Partial Least Squares (MPLS), Reweighted PLS (RWPLS), and Mean Imputation PLS (MIPLS)
  - cov/corr. optimization criterion
  - incorporates the response and censoring
- Proposed Method: Rank-based Partial Least Squares (RPLS)
  - cov/corr. optimization criterion
  - incorporates the response and censoring; based on ranks

UNIV, SPCR, CPCR

- Univariate Selection (UNIV): Bolvestad
  1. fit univariate regression model for each gene, and test null hypothesis $\beta_g = 0$ vs. alternative $\beta_g \neq 0$
  2. arrange the genes according to increasing p-values
  3. pick out the top $K-$ranked genes
- Supervised Principal Component Regression (SPCR): Bair and Tibshirani
  1. use UNIV to pick out a subset of original genes
  2. apply PCA to the subsetted genes
- Correlation Principal Component Regression (CPCR): Zhao and Sun
  1. use PCA but retain all PCs
  2. apply UNIV to pick out the top K-ranked PCs

Real Datasets

Table: Cox model: DLBCL and Harvard datasets. K chosen by CV for the different methods. The $min(CV(surv.error))$ of the 1000 repeated runs are shown.

	DLBCL		HARVARD	
Method	K	error	K	error
PCA	7	0.1026	13	0.121
MPLS	1	0.1076	1	0.1304
RMPLS	1	**0.1056**	1	**0.1124**
CPCR	2	0.1014	2	0.1402
SPCR	1	0.1063	3	0.1473
UNIV	11	0.1221	14	0.1663

REAL DATASETS

Table: AFT Lognormal Mixture model: DLBCL, Harvard, Michigan and Duke datasets. K chosen by CV for the different methods. The $min(CV(fit.error))$ of the 1000 repeated runs are shown.

Method	DLBCL		HARVARD		MICHIGAN		DUKE	
	K	error	K	error	K	error	K	error
PCA	5	5.33	6	1.43	4	4.14	4	19.15
MPLS	3	2.56	1	0.51	3	1.04	1	15.91
RMPLS	5	**1.62**	2	**0.36**	4	**0.64**	3	**5.41**
RWPLS	1	5.54	1	1.35	1	4.90	1	13.11
RRWPLS	1	5.32	1	2.06	1	4.07	2	9.60
MIPLS	3	2.80	2	0.58	2	1.72	2	11.79
RMIPLS	4	1.92	2	0.56	2	1.11	1	10.42
CPCR	7	4.59	5	1.11	4	2.47	4	9.88
SPCR	1	5.48	1	2.12	2	5.10	2	22.14
UNIV	5	4.09	6	0.84	6	1.72	7	10.80

Gene Ranking

- Select significant genes
  - Ranking of Genes: Absolute Value of the Estimated Weights on the Genes (AEW)

$$AEW = |W\hat{\beta}_R^*| \tag{32}$$

where $\hat{\beta}_R^* = \frac{\hat{\beta}_R}{se(\hat{\beta}_R)}$

  - R denotes either Cox or AFT model

Gene Ranking

Table: Number of top-ranked genes in common between the ranked versions of PLS and their un-ranked counterparts for DLBCL, Harvard, Michigan and Duke datasets using the absolute of the estimated weights for the genes for 1st component. The first row shows the number of considered top-ranked genes.

	K top-ranked genes	25	50	100	250	500	1000
DLBCL	MPLS and RMPLS	15	33	74	188	397	802
	RWPLS and RRWPLS	0	0	1	32	140	405
	MIPLS and RMIPLS	18	36	76	201	409	843
HARVARD	MPLS and RMPLS	12	28	58	171	368	822
	RWPLS and RRWPLS	0	0	3	15	80	273
	MIPLS and RMIPLS	14	28	69	170	371	804
MICHIGAN	MPLS and RMPLS	10	20	46	117	273	601
	RWPLS and RRWPLS	0	0	0	2	20	126
	MIPLS and RMIPLS	0	0	1	12	45	158
DUKE	MPLS and RMPLS	3	3	3	21	73	210
	RWPLS and RRWPLS	0	3	7	36	105	287
	MIPLS and RMIPLS	0	0	2	18	59	194

Assessment of the Methods

Simulation: Generating Gene Expression values

$$x_{ij} = exp(x^*_{ij}) \tag{33}$$

$$x^*_{ij} = \sum_{k=1}^{d} r_{ki}\tau_{kj} + \epsilon_{ij}, \quad k = 1, \ldots, d \tag{34}$$

- $\tau_{kj} \overset{iid}{\sim} N(\mu_\tau, \sigma^2_\tau)$
- $\epsilon_{ij} \sim N(\mu_\epsilon, \sigma^2_\epsilon)$
- $r_{ki} \sim Unif(-0.2, 0.2)$ fixed for all simulations
- $d = 6$, $\mu_\epsilon = 0$, $\mu_\tau = 5/d$, $\sigma_\tau = 1$, $\sigma_\epsilon = 0.3$

- 5,000 simulations
- $p \in 100, 300, 500, 800, 1000, 1200, 1400, 1600$
- $N = 50$

Simulation Setup

Simulation: Generating Survival Times

- Cox PH model

Exponential distribution

$$y_i = y_{0i} e^{-X_i'\beta}$$

$$c_i = c_{0i} e^{-X_i'\beta}$$

$$y_{0i} \sim Exp(\lambda_y)$$

$$c_{0i} \sim Exp(\lambda_c)$$

Weibull distribution

$$y_i = y_{0i} \left(e^{-X_i'\beta}\right)^{1/\lambda_y}$$

$$c_i = c_{0i} \left(e^{-X_i'\beta}\right)^{1/\lambda_c}$$

$$y_{0i} \sim Weib(\lambda_y, \alpha_y)$$

$$c_{0i} \sim Weib(\lambda_c, \alpha_c)$$

Simulation Setup

Simulation: Generating Survival Times

- AFT model
  - $\ln(y_i) = \mu + X_i'\beta + u_i$ and $\ln(c_i) = \mu + X_i'\beta + w_i$
    - log normal mixture:
      $f_{u_i}(u) = 0.9\phi(u) + \frac{0.1}{10}\phi(u/10)$
    - exponential, lognormal, log-t
  - $w_i \sim Gamma(a_c, s_c)$

Simulation Setup

Simulation: Generating Survival Times

- both Cox and AFT models
  - observed surv. time $T_i = min(y_i, c_i)$
  - censoring indicator $\delta_i = I(y_i < c_i)$
  - censoring rate $P[y_i < c_i]$
- $\beta_j \sim N(0, \sigma_\pi^2)$, $\sigma_\pi = 0.2$ for all p's
  - outliers in the response for large p

Performance Measures

Performance Measures: once K is selected

- Mean squared error of weights on the genes

$$MSE(\beta) = \frac{1}{s} \sum_{i=1}^{s} \sum_{j=1}^{p} (\beta_j - \hat{\beta}_j(i))^2$$

  - $\hat{\beta} = W \hat{\beta}_{Cox\,,\,AFT}$
- AFT Model
  - Mean squared error of fit

$$MSE(fit) = \frac{1}{s} \sum_{i=1}^{s} \left[\frac{\sum_{n=1}^{N} \delta_n(i) \left(\hat{y}_n^*(i) - y_n^*(i)\right)^2}{\sum_{n=1}^{N} \delta_n(i)} \right]$$

  - $y_n^*(i) = log\left(y_n(i)\right)$
  - $\hat{y}_n^*(i) = \hat{\mu}_{AFT}(i) + M_n(i)' \hat{\beta}_{AFT}(i)$

PERFORMANCE MEASURES

Performance Measures:

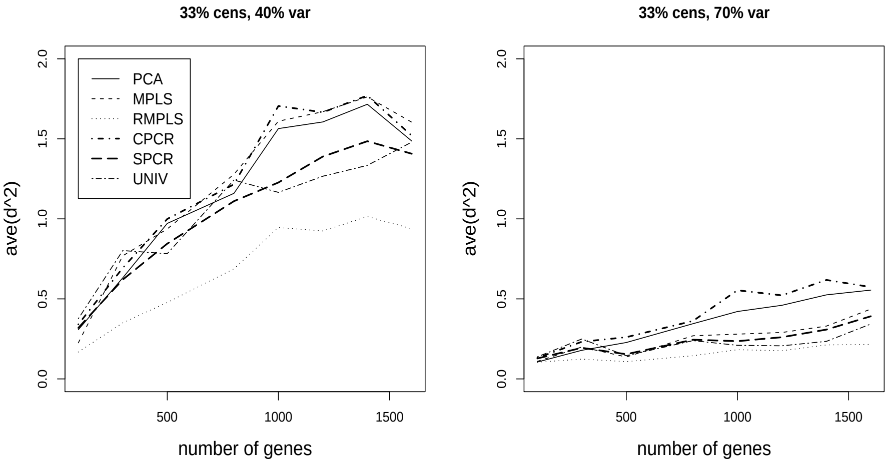
- Cox PH Model
  - MSE of est. surv. function evaluated at the average of the covariates

$$ave(d^2) = \frac{1}{s}\sum_{i=1}^{s}\sum_{t\in D_s}\left(\bar{S}_i(t) - \hat{\bar{S}}_i(t)\right)^2$$

  - MSE of est. surv. function evaluated using the covariates of each individual

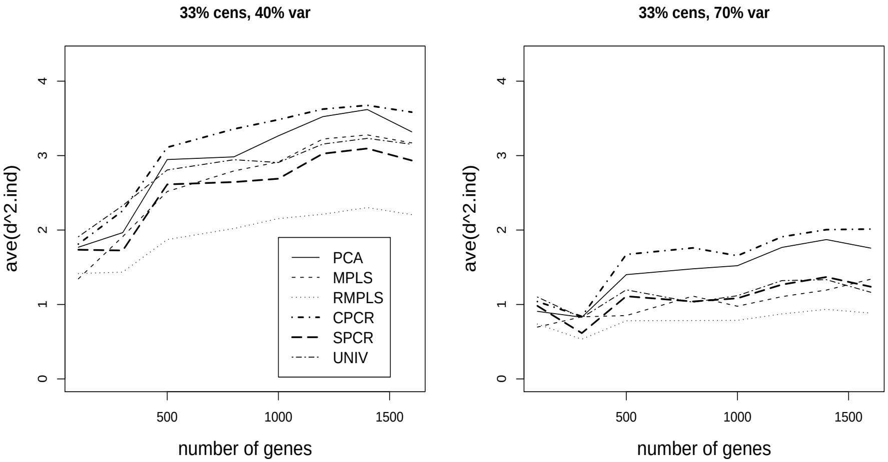
$$ave(d^2.ind) = \frac{1}{sN}\sum_{i=1}^{s}\sum_{n=1}^{N}\sum_{t\in D_s}\left(S_{in}(t) - \hat{S}_{in}(t)\right)^2$$

Simulation Results: Cox model, fix K



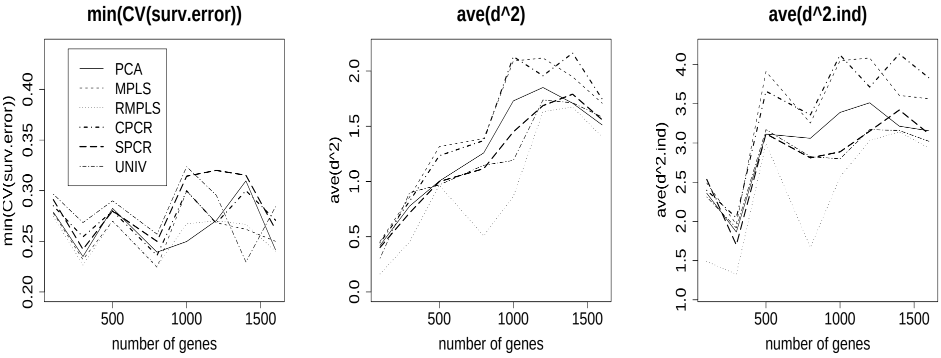
Cox model: $1/3$ censored. **MSE of est. surv. function evaluated at average of covariates ($ave(d^2)$)** for datasets with approximately 40% and 70% TVPE accounted by the first 3 PCs comparing PCA, PLS, MPLS, CPCR, SPCR, and UNIV.

Simulation Results: Cox model, fix K



Cox model: $1/3$ censored. **MSE of est. surv. function evaluated using the covariates of each individual** (*$ave(d^2.ind)$*) for datasets with approximately 40% and 70% TVPE accounted by the first 3 PCs comparing PCA, PLS, MPLS, CPCR, SPCR, and UNIV.

Simulation Results: Cox model, CV



Cox model: $1/3$ censored. K is chosen by CV. **Minimized CV of squared error of est. surv. function** $min(CV(surv.error))$**, MSE of est. surv. function evaluated at average of covariates (**$ave(d^2)$**), and MSE of est. surv. function using the covariates of each individual** ($ave(d^2.ind)$) comparing PCA, MPLS, RMPLS, CPCR, SPCR, and UNIV.

Simulation Results: AFT model, CV

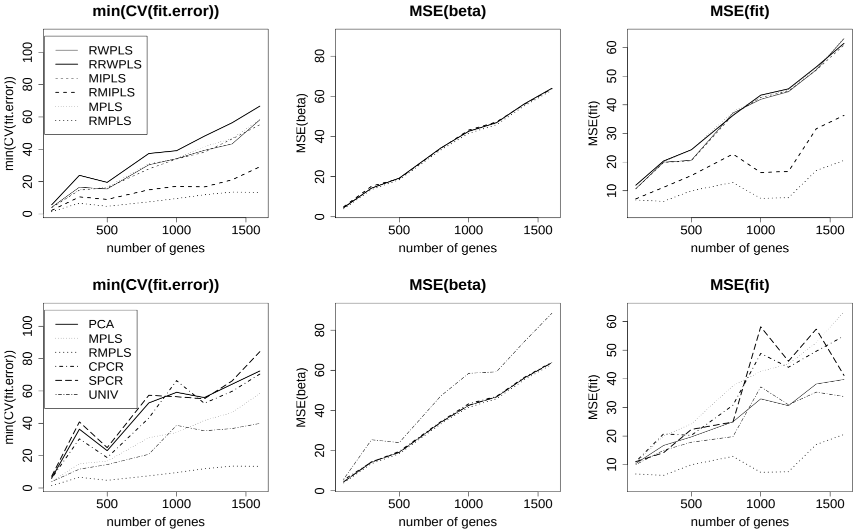


Figure 1: AFT lognormal mixture model: 1/3 censored. K is chosen by CV. $min(CV(fit.error))$, $MSE(\beta)$, and $MSE(fit)$ comparing RWPLS, RRWPLS, MIPLS, RMIPLS, MPLS, and RMPLS (top row), and comparing PCA, MPLS, RMPLS, SPCR, CPCR, and UNIV (bottom row) based on 5000 simulations.

Simulation Results: AFT model, CV

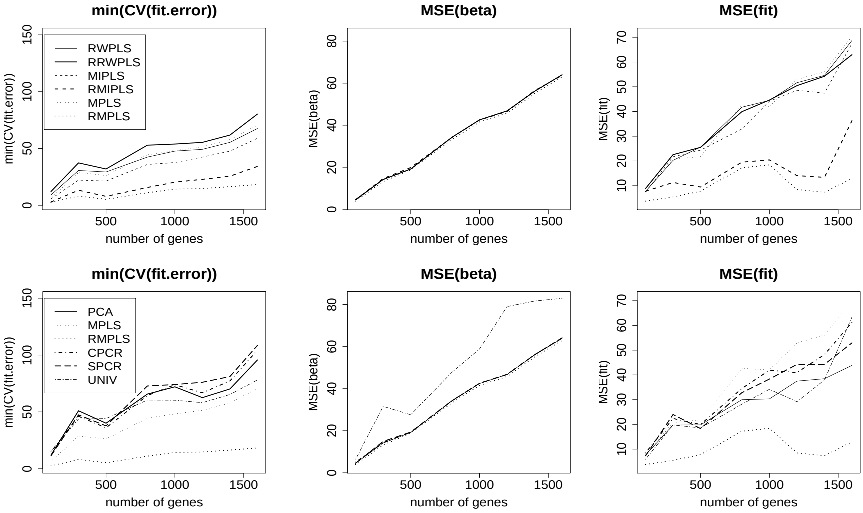


Figure: AFT exponential model: 1/3 censored. K is chosen by CV based on 5000 simulations.

SUMMARY

- Rank-based Partial Least Squares (RPLS)
  - replace Pearson correlation with Spearman rank correlation
  - **incorporate censoring**: Nguyen and Rocke's MPLS, Datta's RWPLS and MIPLS
- Rank-based Partial Least Squares (RPLS) works well
  - in presence of outliers in response
  - comparable to MPLS and PCA in absence of outliers

Outlook

Conclusions

Dimension reduction methods

- Random Projection
  - Johnson-Lindenstrauss (JL) Lemma
  - Improvements on the lower bound for k
    - L_2-L_2: Gaussian and Achlioptas-type random matrices
    - L_2-L_1: Gaussian and Achlioptas-type random matrices
- Rank-based Partial Least Squares (RPLS)
  - competitive dimension reduction method

Thank you!

Special Thanks:

Dr. Rojo, Chair and Advisor, Professor of Statistics, Rice University

ACKNOWLEDGMENT

This work is supported by:

NSF Grant SES-0532346, NSA RUSIS Grant H98230-06-1-0099, and NSF REU Grant MS-0552590

DATTA: REWEIGHTING (RWPLS)

- Kaplan-Meier estimator

$$\hat{S}_c(t) = \prod_{t_i \leq t} \left[1 - \frac{c_i}{n_i} \right]$$

where $t_1 < \cdots < t_m$ are ordered cens. times, $c_i =$ no. cens. observations, $n_i =$ no. still alive at time t_i

- Let $\tilde{y}_i = 0$ for $\delta_i = 0$ and $\tilde{y}_i = T_i / \hat{S}_c(T_i^-)$ for $\delta_i = 1$
- apply PLS to $(\mathbf{X}, \tilde{y})$ back

Datta: Mean Imputation (MIPLS)

- Conditional expectation

$$y^* = \frac{\sum_{t_j > c_i} t_j \Delta \hat{S}(t_j)}{\hat{S}(c_i)}$$

where t_j are ordered death times, $\Delta \hat{S}(t_j)$ is jump size of $\hat{S}$ at t_j, $n_i =$ no. still alive at time t_i

- Let $\tilde{y}_i = y_i$ for $\delta_i = 1$ and $\tilde{y}_i = y_i^*$ for $\delta_i = 0$
- apply PLS to $(\mathbf{X}, \tilde{y})$ back

BIBLIOGRAPHY

- Achlioptas, D. Database-friendly random projections. *Proc. ACM Symp. on the principles of database systems*, 2001. pp. 274-281.
- Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Broldrick, J.C., Sabet, H., Tran, T., and Yu, X. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403*, 2000. pp. 503-511.
- Bair, E., Hastie, T., and Tibshirani, R. Prediction by supervised principal components. *Journal of American Statistical Association 101*. 2006. pp. 119-137.
- Bovelstad, H.M., Nygard, S., Storvold, H.L., Aldrin, M., Borgan, O., Frigessi, A., and Lingjaerde, O.C. Predicting survival from microarray data - a comparative study. *Bioinformatics Advanced Access* 2007.
- Cox, D.R. Regression Models and life tables (with discussion). *Statistical Society Series B34*, 1972. pp. 187.
- Dai, J.J., Lieu, L., and Rocke, D.M. Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology, vol. 5, issue 1, article 6*, 2006.
- Dasgupta, S. and Gupta, A. An elementary proof of the Johson- Lindenstrauss lemma. *Technical report 99-006*, UC Berkeley, March 1999.
- Datta, S., Le Rademacher, J., and Datta, S. Predicting patient survival by accelerated failure time modeling using partial least squares and lasso. *Biometrics 63*, 2007. pp. 259-271.
- Johnson, W. and Lindenstrauss, J. Extensions of Lipschitz maps into a Hilbert space. *Contemp. Math. 26*, 1984, pp. 189-206.
- Kaplan E.L., and Meier, P. Nonparametric estimation from incomplete observations. *Journal of American Statistics Association 53*, 1958. pp. 467-481.
- Klein, J.P., and Moeschberger, M.L. Survival Analysis: techniques for censored and truncated data. *Springer*, second edition. New York, 2003.
- Kohonen, T. et al. Self organization of massive document collection. *IEEE Transactions on Neural Networks, 11 (3)*, May 2000. pp. 574-585.

BIBLIOGRAPHY

- Li, K.C., Wang, J.L., and Chen C.H. Dimension reduction for censored regression data. *The Annals of Statistics 27*, 1999. pp. 1-23.
- Li, L. and Li, H. Dimension reduction methods for microarrays with application to censored survival data. *Center for Bioinformatics and Molecular Biostatistics, Paper surv2*, 2004.
- Li, P., Hastie, T.J., and Church K.W. Very Sparse Random Projections. *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, 2006. pp. 287-296.
- Mardia, K.V., Kent, J.T., and Bibby, J.M. *Multivariate Analysis*. Academic Press, 2003.
- Nguyen, D.V. and Rocke, D.M. Assessing patient survival using microarray gene expression data via partial least squares proportional hazard regression. *Computational Science Statistics, 33*, 2001. pp. 376.
- Nguyen, D.V. and Rocke, D.M. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics, 18, 1625*, 2002.
- Nguyen, D.V. Partial least squares dimension reduction for microarray gene expression data with a censored response. *Mathematical Biosciences, 193*, 2005. pp. 119-137.
- Nguyen TS, Rojo J, Dimension reduction of microarray data in the presence of a censored survival response: a simulation study, *Statistical Applications in Genetics and Molecular Biology* **8.1.4**, 2009.
- Nguyen TS, Rojo J, Dimension reduction of microarray gene expression data: the accelerated failure time model, *Journal of Bioinformatics and Computational Biology*, **7.6**, 2009. pp. 939-954.
- Rojo J, Nguyen TS, Improving the Johnson-Lindenstrauss Lemma, *Journal of Computer and System Sciences*. Submitted, 2009.
- Rojo J, Nguyen TS, Improving the Achlioptas bound for Rademacher random matrices. In Preparation, 2009.
- Sun, J. Correlation principal component regression analysis of NIR data. *Journal of Chemometrics, 9*, 1995. pp. 21-29.
- Wold, H. Estimation of principal components and related models by iterative least squares. In Krishnaiah, P. (ed.), *Multivariate Analysis*. Academic Press, N.Y., 1966. pp. 391-420.
- Zhao, Q., and Sun, J. Cox survival analysis of microarray gene expression data using correlation principal component regression. *statistical applications in genetics and molecular biology vol.6.1, article 16*. Berkeley Electronic Press, 2007.