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
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

\[
X = \begin{pmatrix}
X_{11} & X_{12} & \ldots & X_{1p} \\
X_{21} & X_{22} & \ldots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
X_{N1} & X_{N2} & \ldots & X_{Np}
\end{pmatrix}
\]
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 \((X, T, \delta)\)
  - \(X = (x_1, \ldots, x_p)^T\) 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

\[ M_{N \times k} = X_{N \times p} W_{p \times k}, \quad k < N \ll p \]

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 $\Gamma$,

$$M_{N \times k} = X_{N \times p} \Gamma_{p \times k}$$  \hspace{1cm} (1)

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

![Diagram](image)
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 $\mathbb{R}^p$, there is a linear map $f : \mathbb{R}^p \to \mathbb{R}^k$ such that for any $u, v \in V$,

$$(1 - \epsilon) \|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon) \|u - v\|^2.$$

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

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

$\Gamma$ is of dimension $p$ by $k$

- Frankl and Maehara (1988):

$$k \geq \left\lfloor \frac{27 \ln n}{3\epsilon^2 - 2\epsilon^3} \right\rfloor + 1$$

- Indyk and Motwani (1998):
  - entries to $\Gamma$ are i.i.d. $N(0, 1)$
Dasgupta and Gupta (2003): mgf technique

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

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

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

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

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

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

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

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

\[ k \frac{\|f(x)\|^2}{\|x\|^2} \sim \chi_k^2 \quad (6) \]

- Using mgf technique to bound the tail probabilities
Sketch of Proof

- Define \( f(x) = \frac{1}{\sqrt{k}} x \Gamma \), and \( y = \sqrt{k} \frac{f(x)}{||x||} \)
- Let \( y_j = \frac{x_{r_j}}{||x||} \sim N(0, 1) \), and \( y_j^2 \sim \chi^2_1 \) with \( E(||y||^2) = k \)
- Let \( \alpha_1 = k(1 + \epsilon) \), and \( \alpha_2 = k(1 - \epsilon) \), then
JL Lemma: Dasgupta and Gupta

Sketch of Proof

- **Right-tail prob:**

  \[ P \left[ \|f(x)\|^2 \geq (1 + \epsilon) \|x\|^2 \right] = P \left[ \|y\|^2 \geq \alpha_1 \right] \]

  \[ \leq \left( e^{-s(1+\epsilon)} E \left( e^{sy_j^2} \right) \right)^k, \quad s > 0 \]

  \[ = e^{-s\alpha_1} (1 - 2s)^{-k/2}, \quad s \in (0, 1/2). \]

- **Left-tail prob:**

  \[ P \left[ \|f(x)\|^2 \leq (1 - \epsilon) \|x\|^2 \right] = P \left[ \|y\|^2 \leq \alpha_2 \right] \]

  \[ \leq \left( e^{s(1-\epsilon)} E \left( e^{-sy_j^2} \right) \right)^k, \quad s > 0 \]

  \[ = e^{s\alpha_2} (1 + 2s)^{-k/2} \]

  \[ \leq e^{-s\alpha_1} (1 - 2s)^{-k/2}, \quad s \in (0, 1/2). \]
Sketch of Proof

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

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

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

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

and

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

\[
P\left\{ \bigcap_{u,v \in V \atop u \neq v} (1 - \epsilon) \|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon) \|u - v\|^2 \right\} \geq 1 - \frac{2}{n^2} + \beta
\]

- Introduce \( \beta > 0 \) (Achlioptas (2001)) so that for each pair \( u, v \in V \),
  \[
P\left[ (1 - \epsilon) \|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon) \|u - v\|^2 \right] \geq 1 - \frac{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) (NR₁ 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 $k \frac{||f(x)||^2}{||x||^2}$, where $x = u - v$
- 12%-34% improvement
OUTLINE OF PROOF

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

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

(10)

▶ $||y||^{2} = \sum_{j=1}^{k} y_{j}^{2} \sim \chi_{k}^{2}$

▶ Let $d = k/2$, $\alpha_{1} = k(1 + \epsilon)$, and $\alpha_{2} = k(1 - \epsilon)$, then,

▶ **Right tail prob.:**

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

▶ **Left tail prob.:**

$$P[||y||^{2} \leq \alpha_{2}] = \int_{0}^{\alpha_{2}/2} \frac{1}{\Gamma(a)} z^{d-1} e^{-z} dz = \sum_{y=d}^{\infty} \frac{\alpha_{2}/2)^{y}e^{-\alpha_{2}/2}}{y!}$$
**Theorem 1:** Given $d$ as a positive integer

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

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

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

$$\sum_{y=d}^{\infty} \frac{\lambda_2^y}{y!} \leq \left( \frac{\lambda_2}{d - \lambda_2} \right) \left( \frac{\lambda_2^{d-1}}{(d - 1)!} \right) \quad (12)$$
**IMPROVEMENT ON DASGUPTA AND GUPTA BOUND**

Right-tail prob.

\[
P[\|y\|^{2} \geq \alpha_{1}] = e^{-\lambda_{1}} \sum_{y=0}^{d-1} \frac{\lambda_{1}^{y}}{y!} \leq \left( \frac{1 + \epsilon}{\epsilon} \right) \left( \frac{\lambda_{1}^{d-1}}{(d-1)!} \right) e^{-\lambda_{1}} \tag{13}\]

Left-tail prob.

\[
P[\|y\|^{2} \leq \alpha_{2}] = e^{-\lambda_{2}} \sum_{y=d}^{\infty} \frac{\lambda_{2}^{y}}{y!} \leq \left( \frac{1 - \epsilon}{\epsilon} \right) \left( \frac{\lambda_{2}^{d-1}}{(d-1)!} \right) e^{-\lambda_{2}} \leq \left( \frac{1 + \epsilon}{\epsilon} \right) \left( \frac{\lambda_{1}^{d-1}}{(d-1)!} \right) e^{-\lambda_{1}}
\]

► lower bound for \(k\):

\[
P[\|y\|^{2} \geq \alpha_{1}] + P[\|y\|^{2} \leq \alpha_{2}] \leq 2e^{-\lambda_{1}} \left( \frac{1+\epsilon}{\epsilon} \right) \left( \frac{\lambda_{1}^{d-1}}{(d-1)!} \right) \leq 2/n^{2+\beta}
\]
Comparison of the Lower Bounds on $k$

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

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

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 $k$,

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

- $\delta \in (0, 1)$, $\epsilon \in (0, 1/2)$, $C$ 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)
Rojo and Nguyen (2009a) \( NR_2 \) Bound

\[
k \geq \frac{(2 + \beta) \ln n}{- \ln(A(s^*_{\epsilon}))}
\]  

\( A(s) = 2e^{-s\sqrt{2/\pi(1+\epsilon)+s^2/2}}\Phi(s), \ s \geq 0 \)

\( s^*_{\epsilon} \) is unique minimizer of \( A(s) \)

based on mgf technique

Gaussian and Achlioptas random matrix

36\%-40\% improvement
Define \( f(x) = \frac{1}{k}x\Gamma \).

Let \( y_j = \frac{x_{rj}}{|x|_2} \sim N(0, 1) \), then \( E(\|y\|_1) = k\sqrt{2/\pi} \) 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.**:

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

**Left tail prob.**:

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

Eq. (16) is obtained since \( e^{2s\sqrt{2/\pi}} < \frac{\Phi(s)}{1 - \Phi(s)}, \quad s > 0 \).

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.

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

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

- $L_2$-$L_2$ projection

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

$$k \geq \frac{(24 + 12\beta) \ln n}{3\epsilon^2 - 2\epsilon^3}. \quad (18)$$
Define $f(x) = \frac{1}{\sqrt{k}} x \Gamma$, and $y_j = \frac{x r_j}{||x||_2}$

**Goal:** bound mgf of $y_j^2$

bounding the even moments of $y_j$ by the even moments in
the case $r_{ij} \sim N(0, 1)$

$\implies$ mgf of $y_j^2$ using $r_{ij}$ from Achlioptas proposal is
bounded by mgf using $r_{ij} \sim N(0, 1)$. 

Achlioptas
**$L_2-L_2$ RP: Improvement on Achlioptas Bound**

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

Rojo and Nguyen (2009b)

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

$$k \left( \frac{\|f(x)\|_2^2}{\|x\|_2^2} \right) = \sum_{j=1}^{k} y_j^2 \overset{D}{=} k + 2 \sum_{j=1}^{k} \sum_{l=1}^{p} \sum_{m=l+1}^{p} c_{lm} r_{lmj}$$  

(19)

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

- 3 improvements on Achlioptas bound
**L₂-L₂ 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 $k$**

\[
k \geq \left( \frac{(8 + 4\beta) \ln n}{\epsilon^2} \right) \left( \frac{p - 1}{p} \right).
\]  (20)
OUTLINE OF PROOF

- Right tail prob.:

\[
P[\|y\|^2 \geq k(1 + \epsilon)] = P \left[ \sum_{j=1}^{p} \sum_{l=1}^{p} \sum_{m=l+1}^{p} c_{lm}r_{lmj} \geq \frac{k\epsilon}{2} \right] \\
\leq \exp \left( - \frac{k\epsilon^2}{8 \sum_{l=1}^{p} \sum_{m=l+1}^{p} c_l^2 c_m^2} \right). \quad (21)
\]

- Left tail prob.:

\[
P[\|y\|^2 \leq k(1 - \epsilon)] = P \left[ \sum_{j=1}^{p} \sum_{l=1}^{p} \sum_{m=l+1}^{p} c_{lm}r_{lmj} \leq -\frac{k\epsilon}{2} \right] \\
\leq \exp \left( - \frac{k\epsilon^2}{8 \sum_{l=1}^{p} \sum_{m=l+1}^{p} c_l^2 c_m^2} \right). \quad (22)
\]
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.:

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

- Left tail prob.:

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

$$1 - \Phi\left(\epsilon \sqrt{\frac{kp}{2(p-1)}}\right) + \frac{0.7915}{\sqrt{kp(p-1)/2}} \leq 1/n^{2+\beta}.$$  (23)
**L₂-L₂ 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 ($\epsilon = 0.1$)

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

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**

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.

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

\[
 w_k = \arg \max_{w'w=1} \text{Var}(Xw) = \arg \max_{w'w=1} (N-1)^{-1}w'X'Xw \\
\text{s.t. } w_k'X'Xw_j = 0 \text{ for all } 1 \leq j < k.
\]

- 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 1st \( K \) PCs is: \( \sum_{k=1}^{K} \lambda_k \)
How to Select $K$: number of PC’s

- Proportion of Variation Explained
- Cross-validation
PARTIAL LEAST SQUARES (PLS)

- Herman Wold (1960)

- Covariance optimization criteria,

\[ w_k = \arg \max_{w'w=1} \text{Cov}(Xw, y) = \arg \max_{w'w=1} (N - 1)^{-1} w'X'y \]

s.t. \( w'_kX'Xw_j = 0 \) 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 $X'X$
- $\theta_{ik}$ 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: Nonparametric Approaches to Incorporate Censoring**

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

\[ w_k = \arg \max_{w'R'_X R_y} (N - 1)^{-1} w'R'_X R_y \quad (25) \]

- use Nguyen and Rocke’s procedure (MPLS) and Datta’s RWPLS and MIPLS to incorporate censoring
RPLS: Derivation of the Weights

\[ w_1 = \frac{R'_X R_y}{||R'_X R_y||} \]  

(26)

\[ w_k \propto P_{k-1} w_1, \quad k \geq 2 \]  

(27)

where

\[ P_{k-1} = I - \zeta_1 S_{R_X} - \zeta_2 S_{R_X}^2 - \cdots - \zeta_{k-1} S_{R_X}^{k-1} \]  

(28)

with \( S_{R_X}^j = S_{R_X} S_{R_X} \cdots S_{R_X} \), \( j \) times, \( S_{R_X} = R'_X R_X \), \( S_X = X'X \), and

\( \zeta \)'s obtained from

\[ w'_1 P_{k-1} S_X w_1 = 0 \]

\[ w'_1 P_{k-1} S_X w_2 = 0 \]

\[ \cdots \]

\[ w'_1 P_{k-1} S_X w_{k-1} = 0 \]
OUTLINE

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

Cox Proportional Hazards (PH) Model

\[ h(t; z_i, \beta) = h_0(t)e^{z_i'\beta} \]  \hspace{1cm} (29)

- \( h_0 \): unspecified baseline hazard

\[ S(t; z_i, \beta) = S_0(t)e^{z_i'\beta} \]  \hspace{1cm} (30)

- incorporates the covariates and censored information
- proportional hazards assumption
Survival Analysis

Accelerated Failure Time (AFT) model

- logarithm of true survival time

\[ \log(y_i) = \mu + X'_i\beta + \sigma u_i \]  \hspace{1cm} (31)

- \( y_i \)'s are true survival times
- \( \mu \) and \( \sigma \) 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

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

- $i = 1, \ldots, s$ 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
- $\hat{S}_m$ denotes the est. surv. function for $m^{th}$ fold
- $\hat{S}_{-m}$ denotes the est. surv. function when the $m^{th}$ fold is removed

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

$$\hat{S}_{-m}(t) = \frac{1}{N_{-m}} \sum_{n=1}^{N_{-m}} \hat{S}_{-m,n}(t)$$
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
- $y_{m,l}^*(i) = \ln (y_{m,l}(i))$
- $\hat{y}_{m,l}^*(i)$ are the estimates of $y_{m,l}^*(i)$

$$\hat{y}_{m,l}^*(i) = \hat{\mu}_{-m,AFT}(i) + M_{m,l}(i)' \hat{\beta}_{-m,AFT}(i)$$
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
**Table:** Cox model: DLBCL and Harvard datasets. $K$ chosen by CV for the different methods. The $\text{min}(CV(\text{surv.error}))$ of the 1000 repeated runs are shown.

<table>
<thead>
<tr>
<th>Method</th>
<th>DLBCL</th>
<th>HARVARDD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K</td>
<td>error</td>
</tr>
<tr>
<td>PCA</td>
<td>7</td>
<td>0.1026</td>
</tr>
<tr>
<td>MPLS</td>
<td>1</td>
<td>0.1076</td>
</tr>
<tr>
<td><strong>RMPLS</strong></td>
<td>1</td>
<td><strong>0.1056</strong></td>
</tr>
<tr>
<td>CPCR</td>
<td>2</td>
<td>0.1014</td>
</tr>
<tr>
<td>SPCR</td>
<td>1</td>
<td>0.1063</td>
</tr>
<tr>
<td>UNIV</td>
<td>11</td>
<td>0.1221</td>
</tr>
</tbody>
</table>
### Real Datasets

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

<table>
<thead>
<tr>
<th>Method</th>
<th>DLBCL ( K )</th>
<th>DLBCL Error</th>
<th>HARVARD ( K )</th>
<th>HARVARD Error</th>
<th>MICHIGAN ( K )</th>
<th>MICHIGAN Error</th>
<th>DUKE ( K )</th>
<th>DUKE Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>5</td>
<td>5.33</td>
<td>6</td>
<td>1.43</td>
<td>4</td>
<td>4.14</td>
<td>4</td>
<td>19.15</td>
</tr>
<tr>
<td>MPLS</td>
<td>3</td>
<td>2.56</td>
<td>1</td>
<td>0.51</td>
<td>3</td>
<td>1.04</td>
<td>1</td>
<td>15.91</td>
</tr>
<tr>
<td>RMPLS</td>
<td>5 1.62</td>
<td>2 0.36</td>
<td>4</td>
<td>0.64</td>
<td>3</td>
<td>5.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RWPLS</td>
<td>1 5.54</td>
<td>1 1.35</td>
<td>1</td>
<td>4.90</td>
<td>1</td>
<td>13.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RRWPLS</td>
<td>1 5.32</td>
<td>1 2.06</td>
<td>1</td>
<td>4.07</td>
<td>2</td>
<td>9.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPLS</td>
<td>3 2.80</td>
<td>2 0.58</td>
<td>2</td>
<td>1.72</td>
<td>2</td>
<td>11.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMIPLS</td>
<td>4 1.92</td>
<td>2 0.56</td>
<td>2</td>
<td>1.11</td>
<td>1</td>
<td>10.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPCR</td>
<td>7 4.59</td>
<td>5 1.11</td>
<td>4</td>
<td>2.47</td>
<td>4</td>
<td>9.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPCR</td>
<td>1 5.48</td>
<td>1 2.12</td>
<td>2</td>
<td>5.10</td>
<td>2</td>
<td>22.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNIV</td>
<td>5 4.09</td>
<td>6 0.84</td>
<td>6</td>
<td>1.72</td>
<td>7</td>
<td>10.80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Select significant genes

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

\[
AEW = |W \hat{\beta}_R^*|
\]  \hspace{1cm} (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.

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

Simulation: Generating Gene Expression values

\[ x_{ij} = \exp(x^*_{ij}) \] (33)

\[ x^*_{ij} = \sum_{k=1}^{d} r_{ki} \tau_{kj} + \epsilon_{ij}, \quad k = 1, \ldots, d \] (34)

- \( \tau_{kj} \sim iid \, N(\mu_\tau, \sigma^2_\tau) \)
- \( \epsilon_{ij} \sim N(\mu_\epsilon, \sigma^2_\epsilon) \)
- \( r_{ki} \sim Unif(-0.2, 0.2) \) 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: 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

▶ \( w_i \sim \text{Gamma}(a_c, s_c) \)
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]
  - $\beta_j \sim N(0, \sigma^2_\pi)$, $\sigma_\pi = 0.2$ 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

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

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

- **Cox PH Model**
  
  - **MSE of est. surv. function evaluated at the average of the covariates**
    \[
    \text{ave}(d^2) = \frac{1}{s} \sum_{i=1}^{s} \sum_{t \in D_s} \left( \bar{S}_i(t) - \hat{S}_i(t) \right)^2
    \]

  - **MSE of est. surv. function evaluated using the covariates of each individual**
    \[
    \text{ave}(d^2 \text{. ind}) = \frac{1}{sN} \sum_{i=1}^{s} \sum_{n=1}^{N} \sum_{t \in D_s} \left( S_{in}(t) - \hat{S}_{in}(t) \right)^2
    \]
**Simulation Results: Cox model, fix K**

Cox model: 1/3 censored. **MSE of est. surv. function evaluated at average of covariates** ($\text{ave}(d^2)$) 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** ($\text{ave}(d^2.\text{ind})$) 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**

**min(CV(surv.error))**

- **PCA**
- **MPLS**
- **RMPLS**
- **CPCR**
- **SPCR**
- **UNIV**

**Number of genes**

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** ($\text{ave}(d^2)$), and **MSE of est. surv. function using the covariates of each individual** ($\text{ave}(d^2\.ind)$) 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. \( \text{min}(CV(\text{fit.error})) \), \( MSE(\beta) \), and \( MSE(\text{fit}) \) 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.
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

Outlet
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, \( c_i = \text{no. cens. observations} \), \( n_i = \text{no. still alive at time } t_i \)

- 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 \((X, \tilde{y})\)
Data: 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 \( t_j \), \( n_i = \text{no. still alive at time } t_i \)

- Let \( \tilde{y}_i = y_i \) for \( \delta_i = 1 \) and \( \tilde{y}_i = y^*_i \) for \( \delta_i = 0 \)

- apply PLS to \( (X, \tilde{y}) \)


