Motivation

The era of scientific mass production (Efron, 2011)

- flood of data, primarily because of advances in new technologies (e.g., microarrays)
- a deluge of questions
- thousands of estimates or hypothesis tests that the statistician is asked to tackle
- complex relationships between variables
- unknown or poor measurement property

Acknowledgements

Collaborators:
- Jemila Hamid
- David Tritchler
- Celia Greenwood
- Elena Parkhomenko
- Pingzhao Hu
- Binod Neupane
- Sathish Pichika
- Ashley Bonner

AND all members of http://beyene-sigma-lab.com/

Funding:
- CIHR
- NSERC
- Genome Canada
- MITACS
Microarrays

- A tool for capturing genetic information (genotype, gene expression etc.) at a large scale

Using high-throughput genotype data to answer questions on complex diseases (Sham & Cherney, 2011)

- Genetic variants involved in individual differences in the propensity to develop disease
- Where are these sequence changes located on the 23 chromosomes that constitute the human genome?
- What is the nature of the sequence changes in these variants (e.g., single base pair changes, copy number changes, etc.)?

Questions ...

- What are the frequencies and effect sizes of these changes?
- How important are these changes relative to the environmental variation in explaining individual differences in disease susceptibility?
- And how do the genetic changes interact with each other and with environmental factors?

Applications of Microarray Technology

- Gene expression profiling
  - In different cells/tissues
  - During the course of development
  - Under different environmental or chemical stimuli
  - In disease state versus healthy
- Molecular diagnosis:
  - Molecular classification of diseases
- Drug development
  - Identification of new targets
- Pharmacogenomics
  - Individualized medicine
The “Omics” era

- Genome – genomics
  - Epigenomics
  - Pharmacogenomics
  - Nutrigenomics
- Transcriptome – transcriptomics
- Protein – proteomics
- Metabolome – metabolomics
- Interactomics
- etc

Statistical Issues

- Experimental design
- Image analysis
- Preprocessing (Normalization, filtering, MV imputation)
- Data visualization
- Regulatory network
- Identify differentially expressed genes
- Clustering
- Classification
- Gene enrichment analysis
- Integrative analysis & meta-analysis

Manhattan plot displaying GWA findings with respect to their genomic positions, highlighting signals of particular interest (McCarthy et al, Nature Rev Genet 2008). Type 2 diabetes component of the Welcome Trust Case Control Consortium study. The strongest associations are seen on chromosomes 10 (transcription factor 7-like 2; TCF7L2), 16 (fat mass and obesity associated; FTO) and 6 (CDK5 regulatory subunit associated protein 1-like 1; CDKAL1).
Data visualization - heatmap

Golub et al.,
ALL – acute lymphoblastic leukemia
AML – acute myeloid leukemia

“Biostatomics”
- The art and science of extracting, organizing, analyzing and interpreting “omics”, clinical, lifestyle and other environmental data
- Systemizing diverse data in order to produce useful information
- A crucial tool for an interdisciplinary research on the determinants and impact of complex diseases:
  - molecular-genetic factors, risk modifiers and population health

Why integrate data?
- Class comparison/univariate association
  - Improve power to measure small effects
  - Improve precision of estimated effects
  - Assess heterogeneity
- Association/correlation between different sets of variables
  - Derive structure in each set and maximize their correlation
- Class prediction
  - Improve prediction accuracy
  - Understand (quantify) relative contribution of different sources of data
**Data integration**

- Conceptual framework
- Integrating similar data types
- Integrating heterogeneous data types
- Integrating statistical information with biological domain data

---

**Integrating “Similar” Data**

- Meta-analysis approaches and methods
  - Fixed versus random effect models
  - Weights based on quality scores
  - Different parameterization of association parameter (i.e., different effect size)
  - Modified meta-analytic methods for heterogeneous cohorts

---

**Part-1:**

Class comparison / univariate association

---

Integrating “similar” data types

- Meta-analytic techniques have been used with excellent success to combine similar types of data across different studies that address similar hypotheses.
- We have demonstrated that our ability to detect associations can be greatly enhanced when proper meta-analytic techniques are applied.


Hypotheses
- Suppose there are k cohorts (studies/populations)
- Let the effect in ith cohort = \( \beta_i \)

Fixed effect model
- (Traditional) FE: \( H_0: \beta_1 = \ldots = \beta_k = 0 \) vs. \( H_1: \beta_i = \beta \neq 0 \)
- New FE: \( H_0: \beta_1 = \ldots = \beta_k = 0 \) vs. \( H_1: \beta_i \neq 0 \) for at least one cohort

Random effect model:
- Let \( \beta_1, \beta_2, \ldots, \beta_k \sim iid N(\mu, \tau^2) \); \( \mu \) = overall mean, \( \tau^2 \) = between-cohort variance
- (Traditional) RE: \( H_0: \mu = 0 \) vs. \( H_1: \mu \neq 0 \)
- New RE: \( H_0: \mu = 0 \) and \( \tau^2 > 0 \) vs. \( H_1: \mu \neq 0 \) or \( \tau^2 > 0 \)

Simulation Study
- Genotypes, \( x = 0, 1, 2 \)
- Total sample sizes, \( N \): 2000, 4000, 6000, 8000, 10000
- Number of studies, \( k \): 2, 3, 5, 7, 10 studies
- \( \beta_1, \beta_2, \ldots, \beta_k \) were simulated from \( N(\mu, \tau^2) \)
- 10,000 simulations for each combination of \( (\mu, \tau) \), where
  - Average effect, \( \mu = 0, .05, .10, .15, .20, .25, .30 \)
  - Corresponding ORs: 1.00, 1.05, 1.11, 1.16, 1.22, 1.28, 1.35.
  - Heterogeneity, \( \tau = 0, 0.1, 0.2, 0.3 \)
  - SNPs minor allele frequency (MAF): MAF = 0.05; MAF = 0.20
  - Genetic risk models: Multiplicative, Dominant, Recessive
- Data were simulated/analyzed using logistic regression

5 studies, each with 800 subjects with 400 cases and 400 controls; MAF = 0.20 (equal)

Tau = 0.2, Study sizes, Case-control ratio=1:1 and MAF = 0.20 (equal)

Integrating “heterogeneous” data

- Observations for different types of variables are available on the same subjects in each study
  - Example-1:
    - Sparse canonical correlation analysis (SCCA)
  - Example-2:
    - Weighted Kernel Fisher discriminant analysis (wKFD)
Part-II
Correlation across data types

Classical PCA Review

- Transform the original variables into new set of variables called principal components.

\[
\begin{array}{cccc}
X_1 & X_2 & \ldots & X_p \\
\end{array} \quad \rightarrow \quad \begin{array}{cccc}
Z_1 & Z_2 & \ldots & Z_p \\
\end{array}
\]

Classical PCA Review

- Principal components are nothing but linear combinations of the original variables:

  \[
  (PC 1) \quad Z_1 = v_{11}X_1 + v_{12}X_2 + \cdots + v_{1p}X_p = v'_1X \\
  (PC 2) \quad Z_2 = v_{21}X_1 + v_{22}X_2 + \cdots + v_{2p}X_p = v'_2X \\
  \vdots \\
  (PC p) \quad Z_p = v_{p1}X_1 + v_{p2}X_2 + \cdots + v_{pp}X_p = v'_pX
  \]

- Loading values: \( V = (v_1, v_2, \ldots, v_p) \)

Classical PCA Review

- Three identifying properties:

  1. Maximized Variances:
     \[ \text{Var}(Z_1) \geq \text{Var}(Z_2) \geq \cdots \geq \text{Var}(Z_p) \geq 0. \]
  2. Orthonormal Loading Vectors, Uncorrelated PCs:
     \[ ||v_j||^2 = 1, \quad v'_j v_m = 0 \quad \text{and} \quad \text{Cov}(Z_j, Z_m) = 0 \quad \text{for} \ j \neq m. \]
  3. Total Variance preserved:
     \[ \sum_{j=1}^{p} \text{Var}(Z_j) = \sum_{j=1}^{p} \text{Var}(X_j). \]

- Variance-covariance matrix dictates solutions.

- Easily found with SVD: \( X = UDV' \)
Issues in High-dimensional data

- **All non-zero loadings; cannot interpret.**

  
  \begin{align*}
  (PC 1) & \quad Z_1 = -0.56X_1 - 0.59X_2 - 0.57X_3 + 0.07X_4 + 0.02X_5 - 0.01X_6 + \ldots \\
  (PC 2) & \quad Z_2 = 0.01X_1 + 0.10X_2 - 0.04X_3 + 0.77X_4 - 0.63X_5 + 0.01X_6 + \ldots \\
  (PC 3) & \quad Z_3 = 0.61X_1 - 0.14X_2 - 0.51X_3 - 0.38X_4 - 0.41X_5 - 0.01X_6 + \ldots \\
  (PC 4) & \quad Z_4 = 0.47X_1 - 0.72X_2 + 0.32X_3 + 0.31X_4 + 0.25X_5 - 0.02X_6 + \ldots \\
  (PC 5) & \quad Z_5 = 0.29X_1 + 0.32X_2 - 0.56X_3 + 0.40X_4 + 0.58X_5 + 0.06X_6 + \ldots \\
  (PC 6) & \quad Z_6 = 0.05X_1 - 0.05X_2 - 0.01X_3 - 0.07X_4 - 0.07X_5 + 0.90X_6 + \ldots 
  \end{align*}

- **Unrealistic and impractical** for latent features of the data to be driven by so many variables.

Sparse PCA

- **Force** those small residual loadings to 0

  \begin{align*}
  (PC 1) & \quad Z_1 = \begin{bmatrix} -0.72X_1 - 0.60X_2 - 0.35X_3 \\ 0X_4 + 0X_5 - 0X_6 \end{bmatrix} \\
  (PC 2) & \quad Z_2 = \begin{bmatrix} 0X_1 + 0X_2 + 0X_3 - 0.67X_4 + 0.74X_5 \\ 0X_6 \end{bmatrix} \\
  (PC 3) & \quad Z_3 = \begin{bmatrix} 0X_1 - 0.22X_2 - 0.94X_3 + 0X_4 + 0X_5 + 0X_6 \end{bmatrix} \\
  (PC 4) & \quad Z_4 = \begin{bmatrix} 0.64X_1 - 0.77X_2 + 0X_3 + 0X_4 + 0X_5 + 0X_6 \end{bmatrix} \\
  (PC 5) & \quad Z_5 = \begin{bmatrix} 0X_1 + 0X_2 + 0X_3 - 0.74X_4 - 0.67X_5 + 0X_6 \end{bmatrix} \\
  (PC 6) & \quad Z_6 = \begin{bmatrix} 0X_1 + 0X_2 + 0X_3 + 0X_4 + 0X_5 + 1X_6 \end{bmatrix}
  \end{align*}

- **Introduce sparseness** to the loadings through adjusting tuning parameters

Extensive Simulation

- Consider two datasets with p and q variables, obtained on n observations.

  \[ \begin{bmatrix} X \\ Y \end{bmatrix} \]
CCA: Cardiac Surgery Data

• Need to find relationship between risk factors and various outcomes in cardiac surgery
• n = 2605 patients who underwent cardiac surgery
• p = 74 potential risk factors
• q = 12 different outcome measures

Example-1

CCA: Cardiac Surgery Data

\[ X = \begin{pmatrix} x_1 & x_2 & \ldots & x_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 & y_2 & \ldots & y_q \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \ldots & y_{nq} \end{pmatrix} \]

• Canonical correlation analysis (CCA) is a classical multivariate method used for finding correlations between two sets of multi-dimensional variables.
  - CCA can be used for dimension reduction and data visualization.
  - \( \text{maximize } a'X'Yb \text{ subject to } a'X'Xa = 1, \ b'Y'Yb = 1 \)
  - CCA gives a linear combination of X that is highly associated with a linear combination of Y measurements

CCA ...

• Need samples at least 20 x number of variables to avoid computational problems and to estimate parameters accurately
• Solutions are linear combinations of entire sets of variables under consideration
• In high-dimensional data, sample size is very small compared to number of variables

Sparse canonical correlation analysis

• Automated selection of variables based on mathematical objective function
• Developed fast-converging computer algorithm


**SCCA – a comparison of methods**

(Sathish Pichika’s MSc project)

- We compared three SCCA methods: Parkhomenko et al. (2009), Witten et al. (2009), and Lee et al. (2011)
- In SCCA, only a sparse set of variables will be included in the solution from each set.

\[
\text{maximize}_{a,b} \quad a'X'Yb \\
\text{subject to} \quad a'a \leq 1, \quad b'b \leq 1, \quad P_1(a) \leq c_1, \quad P_1(b) \leq c_2
\]

- CCA/SCCA seeks weights \(a, b\) such that \(\text{Cor}(Xa, Yb)\) is large but in SCCA most of the weights are 0. i.e., \(a_i's, b_i's\) are 0.
- The penalty functions vary and tuning parameters estimated using cross-validation

**Simulation**

- Let \(X\) contain \(p\) variables and \(Y\) contains \(q\) variables and sample sizes be \(n\).
- Suppose only a subset of variables in \(X\) is correlated with a subset of variables in \(Y\)
  - first few variables in \(X\) is highly correlated with the first few variables in \(Y\)

\[
X = \begin{bmatrix}
x_{11} & \cdots & x_{1r} & \cdots & x_{1p} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{n1} & \cdots & x_{nr} & \cdots & x_{np}
\end{bmatrix} \quad Y = \begin{bmatrix}
y_{11} & \cdots & y_{1r} & \cdots & y_{1q} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
y_{n1} & \cdots & y_{nr} & \cdots & y_{nq}
\end{bmatrix}
\]

**Parameters Varied in Simulation**

<table>
<thead>
<tr>
<th>Varied Parameter</th>
<th>Description</th>
<th>Values between</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)</td>
<td>Observations</td>
<td>30 and 500</td>
</tr>
<tr>
<td>(p)</td>
<td>Number of variables in (X)</td>
<td>50 and 2000</td>
</tr>
<tr>
<td>(q)</td>
<td>Number of variables in (X)</td>
<td>30 and 1500</td>
</tr>
<tr>
<td>(r)</td>
<td>Number of Correlated variables</td>
<td>5 and 50</td>
</tr>
<tr>
<td>(\sigma_e)</td>
<td>Std. Dev. of Latent variable ((\mu))</td>
<td>1.8 and 4</td>
</tr>
<tr>
<td>(\sigma_\nu)</td>
<td>Std. Dev. of nuisance variable</td>
<td>0.1 and 0.5</td>
</tr>
</tbody>
</table>
Using the angle between the true canonical variates and their estimates as the measure of closeness (Johnstone and Lu (2009)) given by

\[ \text{dist}(\hat{a}_1, \hat{a}_1) = \sin(\gamma) = \sqrt{1 - (\hat{a}_1^T \hat{a}_1)^2} \]

Discordance measures
- number of false negatives (FNN) and false positives (FPN)
- FP = number of nuisance variables with non-zero loadings in the resulting vector
- FN = number of correlated variables with zero loadings in the resulting vector

For each scenario, measures are averaged over 1000 simulated datasets.

Table 1: n = 100, p = 300, q = 200, r = 15, \( \sigma_\mu = 2 \), \( \sigma_e = 0.01 \)

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Corr.</th>
<th>Dist (a)</th>
<th>Dist (b)</th>
<th>FPN (a)</th>
<th>FPN (b)</th>
<th>FNN (a)</th>
<th>FNN (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>0.9738</td>
<td>0.17</td>
<td>0.19</td>
<td>285</td>
<td>185</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PT</td>
<td>0.9975</td>
<td>0.09</td>
<td>0.09</td>
<td>2.75</td>
<td>2.55</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>WT</td>
<td>0.9935</td>
<td>0.21</td>
<td>0.20</td>
<td>3</td>
<td>0.36</td>
<td>6.91</td>
<td>8.82</td>
</tr>
<tr>
<td>LT</td>
<td>0.9975</td>
<td>0.02</td>
<td>0.02</td>
<td>0.63</td>
<td>0.91</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CCA</td>
<td>0.9741</td>
<td>0.17</td>
<td>0.17</td>
<td>285</td>
<td>185</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PT</td>
<td>0.9975</td>
<td>0.09</td>
<td>0.09</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>WT</td>
<td>0.9930</td>
<td>0.19</td>
<td>0.18</td>
<td>0</td>
<td>9</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>LT</td>
<td>0.9975</td>
<td>0.02</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: n = 50, p = 100, q = 80, r = (5, 15), \( \sigma_\mu = 2 \), \( \sigma_e = 0.5 \)

<table>
<thead>
<tr>
<th>Method</th>
<th>r</th>
<th>Test Corr.</th>
<th>Dist (a)</th>
<th>Dist (b)</th>
<th>FPN (a)</th>
<th>FPN (b)</th>
<th>FNN (a)</th>
<th>FNN (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>5</td>
<td>0.7519</td>
<td>0.31</td>
<td>0.31</td>
<td>11.52</td>
<td>9.16</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>WT</td>
<td>5</td>
<td>0.75543</td>
<td>0.47</td>
<td>0.49</td>
<td>0.83</td>
<td>0.14</td>
<td>2.72</td>
<td>2.84</td>
</tr>
<tr>
<td>LT</td>
<td>5</td>
<td>0.7737</td>
<td>0.16</td>
<td>0.15</td>
<td>17</td>
<td>14</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>PT</td>
<td>15</td>
<td>0.7541</td>
<td>0.24</td>
<td>0.23</td>
<td>12.84</td>
<td>9.97</td>
<td>0.77</td>
<td>0.71</td>
</tr>
<tr>
<td>WT</td>
<td>15</td>
<td>0.7492</td>
<td>0.53</td>
<td>0.57</td>
<td>0.21</td>
<td>0</td>
<td>9.52</td>
<td>10.39</td>
</tr>
<tr>
<td>LT</td>
<td>15</td>
<td>0.7743</td>
<td>0.18</td>
<td>0.16</td>
<td>15.9</td>
<td>14</td>
<td>0.76</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Gene and protein expression data were obtained from the National Cancer Institute http://discover.nci.nih.gov/cellminer/

- The data contains 60 humans cancer cell lines that include a variety of cancer tissues of origins such as leukemias, lymphomas, and carcinomas of ovarian, renal, breast, prostate, colon, lung, and CNS origin

Pre-processing (normalization, filtering) prior to applying SCCA methods
- n = 59
- X (gene expression data): p = 10,123
- Y (protein data): q = 89
### Summary of results of three different SCCA methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Non-Zeros in Gene Expression Data</th>
<th>Non-Zero in Protein Data</th>
<th>Canonical Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCCA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LT</td>
<td>3232</td>
<td>7</td>
<td>0.8579</td>
</tr>
<tr>
<td>WT</td>
<td>3418</td>
<td>24</td>
<td>0.9516</td>
</tr>
<tr>
<td>PT</td>
<td>310</td>
<td>34</td>
<td>0.9559</td>
</tr>
</tbody>
</table>

Nov 14, 2014

### Part-III
**Class prediction**

Nov 14, 2014

### Kernel-based statistical methods

- Reduce data to the same dimension and common format
  - Each data source is represented as a kernel matrix $K_i$
  - Kernels are similarity measures e.g., Gaussian and polynomial kernels
- Let $K = \{K_1, K_2, ..., K_m\}$ we can define a combined kernel as $\sum \mu_i K_i$
- We proposed weights based on classification accuracy
- wKFD analysis is performed on the combined kernel

Nov 14, 2014

### Kernels

- Kernels will be of size $n$ by $n$
- Any symmetric, positive semi-definite function is a valid kernel
- Linear, polynomial, Gaussian etc.

Weighted kernels

\[ K = \sum_{i=1}^{m} w_i K_i, \]

where \( w_i \) are given by

\[ w_i = \begin{cases} 1 & \text{if } e_i \leq 0.5 \\ 0 & \text{otherwise} \end{cases} \]


Simulation I

We generated two data sets with similar information in predicting outcome (blue and black curves).

Integration provided improved accuracy

Naïve (green) and weighted (red) integration provided similar performances (as expected)

Simulation II

One of the data sets is generated to have more information (black) than the other (blue)

Integration (both naïve and weighted) provided improved accuracy

Weighted integration (red) performed better than naïve integration (green)

We also showed that the kernel weights can be interpreted as relative importance of the data sets

Illustrative example – integration of clinical and gene expression data in cancer prediction

- We used a publicly available breast cancer data set
- 295 breast cancer patients of whom 180 had poor prognosis (distant metastases) and 115 had good prognosis (free of distant metastases).
- Aim: Predict disease outcome (good or poor prognosis)
Illustrative example ...

- Clinical data consists of 12 variables - age, number of positive nodes, tumor diameter, histologic grade, mastectomy, chemotherapy, hormonal therapy, NIH risk, Estrogen Receptor status, St Gallen recommendation, NIH recommendation, Tumor stage
- Gene expression data consists of 24,479 genes
- We used IQR to filter gene expression data – 214 genes were included in the analysis
- wkFD analysis was performed to estimate the relative importance of clinical and gene expression data in predicting disease outcome

<table>
<thead>
<tr>
<th>Method</th>
<th>Weight</th>
<th>Error (SE)</th>
<th>AUC (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>KFD</td>
<td>1</td>
<td>0.68 (0.18)</td>
</tr>
<tr>
<td>Clinical</td>
<td>KFD (naive)</td>
<td>0</td>
<td>0.60 (0.08)</td>
</tr>
<tr>
<td>Gen expression</td>
<td>wkFD</td>
<td>0.05</td>
<td>0.53 (0.04)</td>
</tr>
<tr>
<td>Gen expression</td>
<td>wkFD (naive)</td>
<td>0.498 (0.019)</td>
<td>0.531 (0.049)</td>
</tr>
</tbody>
</table>

Standard errors for the weights
Equal weights are assigned to the clinical and gene expression data.

- Clinical and gene expression data provided similar predictive accuracy
- Integration of the two data sets provided little improvement, this may be due to data redundancy
- Both naive and weighted integration provided similar performance

Challenge

- Needle in the haystack problem
- Over fitting is a huge issue
- Biological validation is critical
Challenges ...

- The “curse” of technology
  - Capacity for collecting data has surpassed the data analysis techniques, and it is only getting worse with newer data types (e.g. whole genome sequence)
- Interdisciplinary collaboration is crucial for success
  - Basic biologists; clinicians; statisticians; computer scientists; mathematicians

Conclusions ...

- With the availability of many large-scale ‘omics’ data along with clinical and environmental data, integrative analysis is becoming crucial
  - can help unravel relationships between different biological functional levels
  - May lead to improved accuracy in the context of prediction
  - Allows detection of small effects sizes
  - Explore heterogeneity
- There are major computational and statistical challenges due to high-dimensional nature of data (e.g., large number of variables but small sample size)

Conclusions ...

- New and efficient statistical methodologies need to be developed and validated
- Appropriate pre-processing of data, quality assessment and adjustment, biological validation etc. are crucial.