


Integrative 'biostatistics':
Concepts, Methods and Challenges



Nov 14, 2014 Joseph Beyene McMaster University 1

Acknowledgements


Collaborators:

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

Funding:

- CIHR
- NSERC
- Genome Canada
- MITACS

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



Nov 14, 2014 Joseph Beyene McMaster University 2

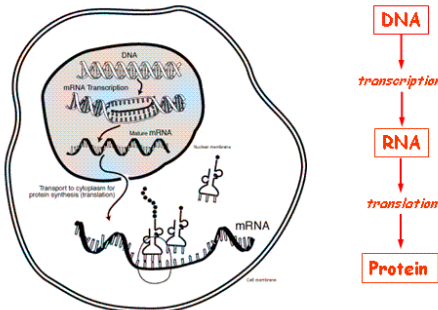
Motivation

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

Nov 14, 2014 Joseph Beyene McMaster University 3

Central Dogma of Biology: Classic View



Nov 14, 2014 McMaster University 4

Microarrays

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



Nov 14, 2014

Joseph Beyene McMaster University

5

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.)?

Nov 14, 2014

Joseph Beyene McMaster University

6

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?

Nov 14, 2014

Joseph Beyene McMaster University

7

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

Nov 14, 2014

Joseph Beyene McMaster University

8

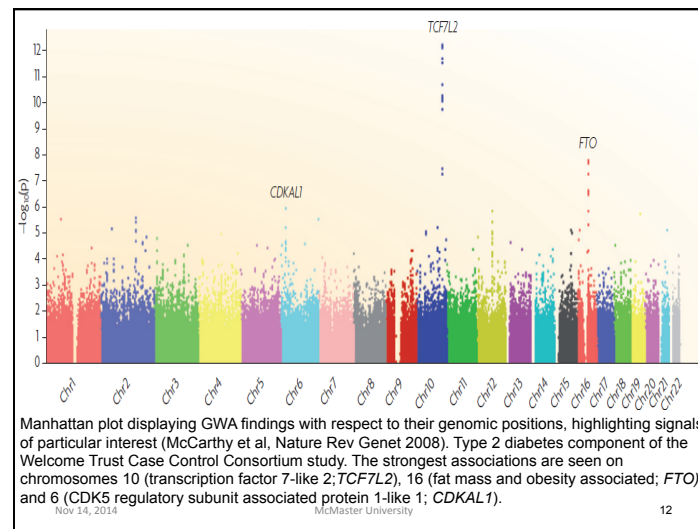
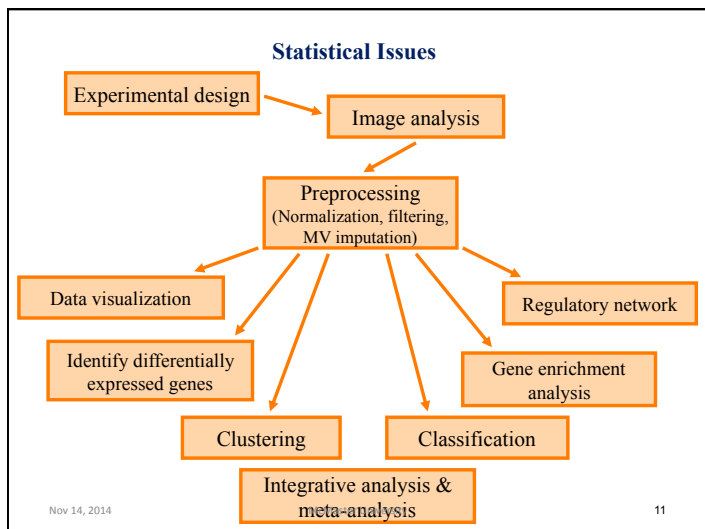
The “Omics” era

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

Nov 14, 2014 Joseph Beyene McMaster University 9

	Sample 1	Sample 2	...	Sample n
Gene 1				
Gene 2				
...				
Gene p				

Nov 14, 2014 McMaster University 10



Data visualization - heatmap

Golub et al.,
Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999.

ALL – acute lymphoblastic leukemia
AML – acute myeloid leukemia

Nov 14, 2014 Joseph Beyene McMaster University 13

fashionomics The art of systemizing your wardrobe

about Debby
services
contact
faq
tips & trends
testimonials
press
links

Debby Jett Allbright
Wardrobe Consultant

"so many clothes, but nothing to wear"

214-675-0121 | debby@fashionomics.net

http://fashionomics.net/

Nov 14, 2014 McMaster University 14

“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*

Nov 14, 2014 Joseph Beyene McMaster University 15

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

Nov 14, 2014 Joseph Beyene McMaster University 16

Data integration

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

Nov 14, 2014

Joseph Beyene McMaster University

17

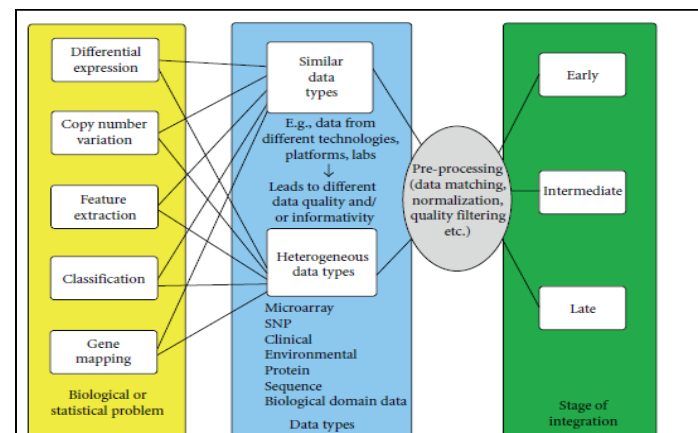


FIGURE 1: Conceptual framework for data integration in genetics and genomics.

Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CM, Beyene J. Data integration in genetics and genomics. Methods and challenges. *Human Genomics and Proteomics*, 2009

Part-1: Class comparison / univariate association

Nov 14, 2014

Joseph Beyene McMaster University

19

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

Nov 14, 2014

Joseph Beyene McMaster University

20

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

Hu, P., Greenwood, C.M.T., Beyene, J. (2005). Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics*, vol. 6, 128, pp. 1–11.

Hu P, Greenwood CM, Beyene J. Using the ratio of means as the effect size measure in combining results of microarray experiments. *BMC Syst Biol*. 2009, 5;3:106.

Neupane B, Loeb M, Anand SS, Beyene J. Meta-analysis of genetic association studies under heterogeneity. *Eur J Hum Genet*. 2012, 11:1174-81.

Friedrich JO, Adhikari NKJ, Beyene J. Ratio of geometric means to analyze continuous outcomes in meta-analysis: comparison to mean differences and ratio of arithmetic means using empiric data and simulation. *Stat Med*. 2012

Nov 14, 2014

Joseph Beyene McMaster University

21

Hypotheses

- ❖ Suppose there are k cohorts (studies/populations)
- ❖ let the effect in i^{th} cohort = β_i

❖ Fixed effect model

- ❖ **(Traditional) FE:** $H_0: \beta_1 = \dots = \beta_k = \beta = 0$ vs. $H_1: \beta_1 = \dots = \beta_k = \beta \neq 0$
- ❖ **New FE:** $H_0: \beta_1 = \dots = \beta_k = 0$ vs. $H_1: \beta_i \neq 0$ for at least one cohort

❖ Random effect model:

Let $\beta_1, \beta_2, \dots, \beta_k \sim \text{iid } N(\mu, \tau^2)$; μ = overall mean, τ^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 $\tau^2 > 0$

Nov 14, 2014

McMaster University

22

Test statistics

- ❖ Let $\hat{\beta}_i$ and s_i be the estimate of β_i and its SE in the i^{th} cohort

❖ Fixed effect model

- ❖ **FE:** $T = (\sum_i w_i \hat{\beta}_i)^2 / \sum_i w_i \sim \chi_1^2$; where $w_i = 1/s_i^2$
- ❖ **New FE:** $T = \sum_i (\hat{\beta}_i/s_i)^2 \sim \chi_k^2$

❖ Random effect model:

Obtain $\hat{\tau}^2$, $\hat{\mu} = \sum_i w_i \hat{\beta}_i / \sum_i w_i$; $\text{var}(\hat{\mu}) = 1/\sum_i w_i$; where $w_i = 1/(s_i^2 + \hat{\tau}^2)$

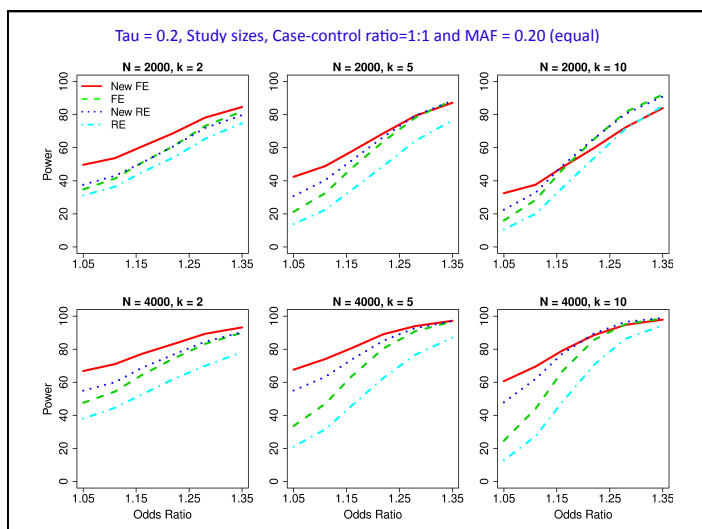
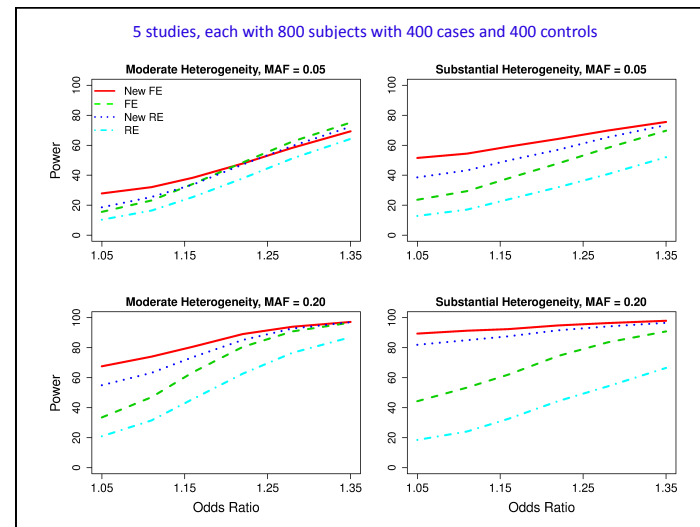
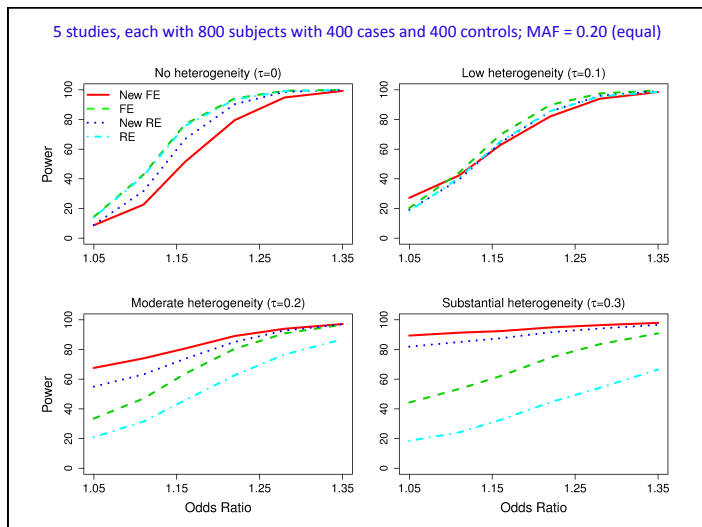
- ❖ **RE:** Wald test, $T = \hat{\mu}^2 / \text{var}(\hat{\mu}) \sim \chi_1^2$
- ❖ **New RE:** LR test, $T = 2(l(\hat{\mu}, \hat{\tau}^2) - l(0, 0)) \sim (\chi_1^2 + \chi_2^2)/2$

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
- ❖ $\beta_1, \beta_2, \dots, \beta_k$ were simulated from $N(\mu, \tau^2)$
- ❖ 10,000 simulations for each combination of (μ, τ) , 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

Neupane B, Loeb M, Anand SS, Beyene J. Meta-analysis of genetic association studies under heterogeneity. *Eur J Hum Genet*. 2012, 11:1174-81.

24



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 (wKFDA)

Nov 14, 2014 Joseph Beyene McMaster University 28

Part-II Correlation across data types

Nov 14, 2014
Joseph Beyene
McMaster University
29

Classical PCA Review

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

X_1	X_2	...	X_p

→

Z_1	Z_2	...	Z_p

Nov 14, 2014
Joseph Beyene
McMaster University
30

Classical PCA Review

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

(PC 1) $Z_1 = v_{11}X_1 + v_{12}X_2 + \dots + v_{1p}X_p = \mathbf{v}'_1\mathbf{X}$

(PC 2) $Z_2 = v_{21}X_1 + v_{22}X_2 + \dots + v_{2p}X_p = \mathbf{v}'_2\mathbf{X}$

⋮

(PC p) $Z_p = v_{p1}X_1 + v_{p2}X_2 + \dots + v_{pp}X_p = \mathbf{v}'_p\mathbf{X}$

- Loading values: $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p)$

Nov 14, 2014
Joseph Beyene
McMaster University
31

Classical PCA Review

- Three identifying properties:
 1. **Maximized Variances:**
 $Var(Z_1) \geq Var(Z_2) \geq \dots \geq Var(Z_p) \geq 0.$
 2. **Orthonormal Loading Vectors, Uncorrelated PCs:**
 $\|\mathbf{v}_j\|_2 = 1, \mathbf{v}'_j\mathbf{v}_m = 0$ and $Cov(Z_j, Z_m) = 0$ for $j \neq m.$
 3. **Total Variance preserved:**
 $\sum_{j=1}^p Var(Z_j) = \sum_{j=1}^p Var(X_j).$
- **Variance-covariance matrix** dictates solutions.
- Easily found with **SVD**: $\mathbf{X} = \mathbf{UDV}'$

Nov 14, 2014
Joseph Beyene
McMaster University
32

Issues in High-dimensional data

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

(PC 1) $Z_1 = -0.56X_1 - 0.59X_2 - 0.57X_3 + 0.07X_4 + 0.02X_5 - 0.01X_6 + \dots$ →

(PC 2) $Z_2 = 0.01X_1 + 0.10X_2 - 0.04X_3 + 0.77X_4 - 0.63X_5 + 0.01X_6 + \dots$ →

(PC 3) $Z_3 = 0.61X_1 - 0.14X_2 - 0.51X_3 - 0.38X_4 - 0.44X_5 - 0.01X_6 + \dots$ →

(PC 4) $Z_4 = 0.47X_1 - 0.72X_2 + 0.32X_3 + 0.31X_4 + 0.25X_5 - 0.02X_6 + \dots$ →

(PC 5) $Z_5 = 0.29X_1 + 0.32X_2 - 0.56X_3 + 0.40X_4 + 0.58X_5 + 0.06X_6 + \dots$ →

(PC 6) $Z_6 = 0.05X_1 - 0.05X_2 - 0.01X_3 - 0.07X_4 - 0.07X_5 + 0.99X_6 + \dots$ →

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

Nov 14, 2014
Joseph Beyene McMaster University
33

Sparse PCA

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

(PC 1) $Z_1 = -0.72X_1 - 0.60X_2 - 0.35X_3 + 0X_4 + 0X_5 - 0X_6$

(PC 2) $Z_2 = 0X_1 + 0X_2 + 0X_3 - 0.67X_4 + 0.74X_5 + 0X_6$

(PC 3) $Z_3 = -0.27X_1 - 0.22X_2 + 0.94X_3 + 0X_4 + 0X_5 + 0X_6$

(PC 4) $Z_4 = 0.64X_1 - 0.77X_2 + 0X_3 + 0X_4 + 0X_5 + 0X_6$

(PC 5) $Z_5 = 0X_1 + 0X_2 + 0X_3 - 0.74X_4 - 0.67X_5 + 0X_6$

(PC 6) $Z_6 = 0X_1 + 0X_2 + 0X_3 + 0X_4 + 0X_5 + 1X_6$

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

Nov 14, 2014
Joseph Beyene McMaster University
34

Extensive Simulation

McMaster University
DigitalCommons@McMaster

Open Access Dissertations and Theses Open Dissertations and Theses

10-1-2012

Sparse Principal Component Analysis for High-Dimensional Data: A Comparative Study

Ashley J. Bonner
McMaster University, ashleybonner@gmail.com

<http://digitalcommons.mcmaster.ca/cgi/viewcontent.cgi?article=8155&context=opendissertations>

Nov 14, 2014
Joseph Beyene McMaster University
35

Canonical Correlation

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

Nov 14, 2014
Joseph Beyene McMaster University
36

Example-1 CCA: Cardiac Surgery Data

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_p \\ x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 & y_2 & \dots & y_q \\ y_{11} & y_{12} & \dots & y_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nq} \end{pmatrix}$$

- 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

Ridderstople et al. Canonical correlation analysis of risk factors and clinical outcomes in cardiac surgery, *Journal of Medical Systems*, 2005 37

Canonical Correlation Analysis

- 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.
- maximize_{a,b} $\mathbf{a}'\mathbf{X}'\mathbf{Y}\mathbf{b}$ subject to $\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a} = 1$, $\mathbf{b}'\mathbf{Y}'\mathbf{Y}\mathbf{b} = 1$
- CCA gives a linear combination of X that is highly associated with a linear combination of Y measurements

Nov 14, 2014 Joseph Beyene McMaster University 38

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

Barcikowski and Stevens. A Monte Carlo study of stability of canonical correlations, canonical weights and canonical variate-variable correlations. *Multivariate Behavioral Research*, 1975

Nov 14, 2014 Joseph Beyene McMaster University 39

Sparse canonical correlation analysis

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

E Parkhomenko, D. Tritchler, J. Beyene .Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. *Statistical Applications in Genetics and Molecular Biology*, 2009

Nov 14, 2014 Joseph Beyene McMaster University 40

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.

$$\text{maximize}_{\mathbf{a}, \mathbf{b}} \mathbf{a}'\mathbf{X}'\mathbf{Y}\mathbf{b}$$

subject to $\mathbf{a}'\mathbf{a} \leq 1, \mathbf{b}'\mathbf{b} \leq 1, P_1(\mathbf{a}) \leq c_1, P_2(\mathbf{b}) \leq c_2$

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

Nov 14, 2014 Joseph Beyene McMaster University 41

McMaster University
DigitalCommons@McMaster

Open Access Dissertations and Theses Open Dissertations and Theses

4-1-2012

Sparse Canonical Correlation Analysis (SCCA): A Comparative Study

Sathish chandra Pichika
McMaster University, pichiksc@math.mcmaster.ca

<http://digitalcommons.mcmaster.ca/cgi/viewcontent.cgi?article=7714&context=opendissertations>

Nov 14, 2014 McMaster University 42

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

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

Nov 14, 2014 Joseph Beyene McMaster University 43

Parameters Varied in Simulation

Varied Parameter	Description	Values between
n	Observations	30 and 500
p	Number of variables in X	50 and 2000
q	Number of variables in Y	30 and 1500
r	Number of Correlated variables	5 and 50
σ_e	Std. Dev. of Latent variable (μ)	1.8 and 4
σ_μ	Std. Dev. of nuisance variable	0.1 and 0.5

Nov 14, 2014 Joseph Beyene McMaster University 44

Performance evaluation

- Using the angle between the true canonical variates and their estimates as the measure of closeness (Johnstone and Lu (2009)) given by

$$dist(a_1, \hat{a}_1) = \sin \angle(a_1, \hat{a}_1) = \sqrt{1 - (a_1^T \hat{a}_1)^2}$$
- 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

Nov 14, 2014 Joseph Beyene McMaster University 45

Table 1: $n = 100, p = 300, q = 200, r = 15, \sigma_\mu = 2, \sigma_\epsilon = 0.01$
 WT = Witten et al. (2009), LT = Lee et al. (2011), PT = Parkhomenko et al. (2009), CCA = Classical Canonical Correlation Analysis.

Method		Test Corr.	Dist (a)	Dist (b)	FPN (a)	FPN (b)	FNN (a)	FNN (b)
CCA	M E A N	0.9738	0.17	0.19	285	185	0	0
PT		0.9975	0.09	0.09	2.75	2.55	0	0
WT		0.9935	0.21	0.20	3	0.36	6.91	8.82
LT		0.9975	0.02	0.02	0.63	0.91	0	0
CCA	M E D I A N	0.9741	0.17	0.17	285	185	0	0
PT		0.9975	0.09	0.09	0	0	0	0
WT		0.9930	0.19	0.18	0	0	9	11
LT		0.9975	0.02	0.02	0	0	0	0

Nov 14, 2014 McMaster University 46

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

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

Nov 14, 2014 McMaster University 47

Application

- 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): $q = 89$

Nov 14, 2014 Joseph Beyene McMaster University 48

Summary of results of three different SCCA methods

Method	Non-Zeros in Gene Expression Data	Non-Zero in Protein Data	Canonical Correlation Coefficient
SCCA			
LT	3232	7	0.8579
WT	3418	24	0.9516
PT	310	34	0.9559

Nov 14, 2014

Joseph Beyene McMaster University

49

Part-III Class prediction

Nov 14, 2014

Joseph Beyene

McMaster University

50

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 $K = \{K_1, K_2, \dots, K_m\}$ 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

Joseph Beyene McMaster University

51

Kernels

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

Hamid JS, Greenwood CMT, Beyene J. Weighted kernel Fisher discriminant analysis for integrating heterogeneous data. *Comput Stat Data Anal.*, 2012, 56:2031–40.

Nov 14, 2014

Joseph Beyene

McMaster University

52

Weighted kernels

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

where w_i are given by

$$w_i = \begin{cases} \frac{1}{e_i} - 2, & \text{if } e_i \leq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

Hamid JS, Greenwood CMT, Beyene J. Weighted kernel Fisher discriminant analysis for integrating heterogeneous data. *Comput Stat Data Anal.*, 2012, 56:2031–40.

Nov 14, 2014 Joseph Beyene McMaster University 53

Simulation I

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

We performed naïve (equal weight) and weighted integration using wKFD.

Integration provided improved accuracy

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

Nov 14, 2014 McMaster University 54

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

Nov 14, 2014 McMaster University 55

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)

Nov 14, 2014 Joseph Beyene McMaster University 56

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

Nov 14, 2014

Joseph Beyene McMaster University

57

Average standardized weights, classification error, and area under the ROC curve and their corresponding standard errors (se) for the breast cancer data.

Method	Weight		Error (se)	AUC (se)
	Clinical	Gene expression		
KFD	1	0	0.300 (0.038)	0.613 (0.054)
KFD	0	1	0.299 (0.038)	0.593 (0.049)
wKFD (naïve) [†]	0.5	0.5	0.293 (0.042)	0.646 (0.051)
wKFD	0.4995 (0.015) [‡]	0.5005 (0.015) [‡]	0.275 (0.039)	0.644 (0.053)

[†] 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 naïve and weighted integration provided similar performance

Nov 14, 2014 McMaster University 58

Challenge

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


Nov 14, 2014

Joseph Beyene McMaster University

59

The NEW ENGLAND JOURNAL of MEDICINE

CORRESPONDENCE



Retraction: A Genomic Strategy to Refine Prognosis in Early-Stage Non–Small-Cell Lung Cancer. N Engl J Med 2006;355:570-80.

TO THE EDITOR: We would like to retract our article, "A Genomic Strategy to Refine Prognosis in Early-Stage Non–Small-Cell Lung Cancer,"¹ which was published in the *Journal* on August 10, 2006. Using a sample set from a study by the American College of Surgeons Oncology Group (ACOSOG) and a collection of samples from a study by the Cancer and Leukemia Group B (CALGB), we have tried and failed to reproduce results supporting the validation of the lung metagene model described in the article. We deeply regret the effect of this action on the work of other investigators.

Jason Koontz, M.D.
Duke University Medical Center
Durham, NC

Robert Kratzke, M.D.
University of Minnesota
Minneapolis, MN

Mark A. Watson, M.D., Ph.D.
Washington University School of Medicine
St. Louis, MO

Michael Kelley, M.D.
Geoffrey S. Ginsburg, M.D., Ph.D.
Mike West, Ph.D.
David H. Harpole, Jr., M.D.

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

Nov 14, 2014

Joseph Beyene McMaster University

61

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)

Nov 14, 2014

Joseph Beyene McMaster University

62

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.

Nov 14, 2014

Joseph Beyene McMaster University

63