Aspects of Statistical Modelling & Data Analysis in Gene Expression Genomics

Mike West
Duke University
These slides:
www.isds.duke.edu/~mw/downloads/SemStat05

Papers, software, many links:
www.isds.duke.edu/~mw

ABS04 web site: Lecture slides, stats notes, papers, data, links:
www.isds.duke.edu/~mw/ABS04

Integrated Cancer Biology Program
icbp.genome.duke.edu

Genome Institute @ Duke
www.genome.duke.edu
#1  
*Genomics, Microarrays, Data:*

Big picture

#2  
*Bayesics - Regression and Shrinkage:*

Gene expression as predictors

#3  
*Patterns and Factors:*

Prediction via pattern profiling

#4  
*Sparse Modelling:*

Regression subset-structure uncertainty

#5  
*Sparse Models and Profiling:*

Gene expression as response: Designed experiments

#6  
*Sparse Models and Profiling:*

Gene expression as response: Latent factor models
(Yet another) History of Life as we know it...

Homo Apriorius
Homo Pragmaticus
Homo Frequentistus
Homo Sapiens
Homo Bayesianis
Genomics, Microarrays, Data: 
Big picture

#1

#2
Bayesics - Regression and Shrinkage: 
Gene expression as predictors

#3
Patterns and Factors: 
Prediction via pattern profiling

#4
Sparse Modelling: 
Regression subset-structure uncertainty

#5
Sparse Models and Profiling: 
Gene expression as response: Designed experiments

#6
Sparse Models and Profiling: 
Gene expression as response: Latent factor models
Transitions in Biology: Data and Observation

Observational science

Molecular science

Genomic science

Data: Scale, Complexity -

Computational & Statistical Science
Low resolution phenotypes
“Small worlds”, small data

Breast cancer:
- Lymph node involvement
- Hormone receptor status
- Tumor size
- Visual assessment

Phenotypes & data

SemStat05 - Warwick - Sept 11 & 12th 2005
Phenotypes & DATA

Higher resolution
Genome scale, big data

Increased understanding
\[ p(X) \]

Improved prediction
\[ p(Y|X) \]
Genomic & Biologic Data

Biological/disease state ...

Gene expression

DNA copy (CGH)

DNA methylation

Proteomics

Metabolics

Clinical markers

Environmental factors

Serum-proteomics

Serum-metabolics

Genotypes

Haplotypes

SNPs

Serum - gene expression

‘host’ ...

SemStat05 - Warwick - Sept 11 & 12th 2005
Data from Experiment and Observation: Challenges

Translation of inferences
Gene expression profiles: Signatures of states
Laboratory/In vitro
Laboratory/Animal models
Human Observational Studies
Human Clinical Studies
**Affymetrix DNA Microarray Data**

*Gene probesets*

*Imaging/Scanning*

*100Mb raw data*

---

Expression intensity estimates \( X \pm S \)

\( p \) genes, \( n \) samples

---

Background, noise, gross defects, ...

Cross-hybridization

Sample-sample normalisation

'Low level" data processing, analysis

---

West et al 2001

Wong & Li (dChip) 2001

Bolstad, Irizarry, Speed et al 2003a,b

RMA estimates - www.bioconductor.org
Multiple expression data sets

Multiple array technologies

Multiple species: genome A - B mappings

Same array platform: sample/lab/study/gene effects

Assay/batch/reagent/hybridisation sensitivities

Sporadic - Sparse
Genomics, Microarrays, Data:
Big picture

Bayesics - Regression and Shrinkage:
Gene expression as predictors

Patterns and Factors:
Prediction via pattern profiling

Sparse Modelling:
Regression subset-structure uncertainty

Sparse Models and Profiling:
Gene expression as response: Designed experiments

Sparse Models and Profiling:
Gene expression as response: Latent factor models
Gene expression as covariates (predictors)

Molecular phenotyping:

- Predict aggressive vs. benign
- Disease susceptible vs. resistance
- Drug/treatment response
- Finding genes linked to response
- Patterns of association among genes
- Signatures of effect - multiple genes
\[
z = H\beta + \nu, \quad \nu \sim N(0, \sigma^2 I)
\]

**Prior:** \( \beta | C \sim N(0, C^{-1}) \)

**Posterior:** \( \beta | z, C \sim N(b, \sigma^2 B^{-1}) \)

**LSE:**
\[
\hat{\beta} = B_*^{-1} H'z \\
B_* = H'H
\]

**Shrinkage:**
\[
b = B^{-1} H'z \\
B = \sigma^2 C + H'H
\]

**Phenotype** \( z \)

**H** \( \sim \) subsets of genes

**Decision theory**

**Regularisation** - Ridge regression

**Key with many predictors**

**Relevance of zero-mean location**

(minimal) Bayes: Shrinkage priors
Degrees and Dimensions of Shrinkage

\[ \beta \sim N(0, C^{-1}) \]

\[ C^{-1} = \tau I, \quad \tau \sim \text{InvGamma} \]

LSE as limiting case - no shrinkage: \( \tau^{-1} \to 0 \)

Shrinks when it matters - weak/no association

Acts against over-fitting, improves stability and robustness in prediction

\[ C^{-1} = \text{diag}(\tau_1, \ldots, \tau_k), \quad \tau_j \sim \text{InvGamma} \]

Multiple shrinkage

\( \beta' = (\beta_1, \ldots, \beta_k) \)

\[ \beta_j \sim N(0, \tau_j) \]

“Shrinks out” irrelevant covariates
Computation: MCMC in Regression

Simulate Posterior:
Iteratively resample conditional posteriors

Sample means, histograms
MC approximation of posterior

\[ p(\beta | z, C) = N(b, \sigma^2 B^{-1}) \]
\[ p(C | z, \beta) = \prod_{j=1}^{k} p(\tau_j | \beta_j) \]
Computation: MCMC in Regression

Modules in MCMC
e.g. response error variance

\begin{align*}
p(\beta | z, C, \sigma^2) &= N(b, \sigma^2 B^{-1}) \\
p(C | z, \beta, \sigma^2) &= \prod_{j=1}^{k} p(\tau_j | \beta_j) \\
p(\sigma^2 | z, \beta, C) &= \text{InvGamma}
\end{align*}
Binary Regression

Binary = thresholded latent continuous
probit~normal, logit~logistic, ...

\[ P_r(y_i = 1) = \Phi(h_i'\beta) \]

Natural model/interpretation
Computationally nice

\[ P_r(y_i = 1) = P_r(z_i > 0), \quad z_i \sim N(h_i'\beta, 1) \]

\[ z = H\beta + \nu, \quad \nu \sim N(0, I) \]
Computation: MCMC in Binary Regression

Linear model if $z$ known

Add module to impute latent $z$
MC samples for $z$
Easy summary, prediction

\[
p(\beta | z, C) = N(b, B^{-1})
\]
\[
p(C | z, \beta) = \prod_{j=1}^{k} p(\tau_j | \beta_j)
\]
\[
p(z | y, \beta) = \prod_{i=1}^{n} p(z_i | y_i, \beta)
\]
Basic Examples: Breast Cancer Data

- ER - (O) Estrogen Receptor Status
- HER2 hormone status
- Lymph node (recurrence risk) status

Frozen tumour: Gene expression
- Higher resolution
- Future clinical tests: Pr(ER+)

y=0/1 (ER -/+)
Protein assay
Immunohistochemical staining
0/1 (0-3)

ER positive tumour
IHC for Estrogen Receptor
(≈60x magnification)

- Brown-red & pink ~ ER+
- Nuclei of breast epithelial cells
- Cytoplasm of breast epithelial cells
- Collagen
- Nuclei of stromal cells
Prediction and \{Gene, Variable, Feature\} Selection

Leave-one-out Cross-Validation (CV) analysis:

"Honest" assessment of precision

Heterogeneity, small samples

Feature/Variable selection

Critical (dominant) component of predictive assessment

(又名 2001 breast cancer)
Predicting lymph node status

Pre-selection of 100 genes

"Honest" CV predictions

Large p:
Small models-
Sparsity

Variable selection,
Uncertainty

Complex
interdependencies

Multiplicities
#1 **Genomics, Microarrays, Data:**
Big picture

#2 **Bayesics - Regression and Shrinkage:**
Gene expression as predictors

#3 **Patterns and Factors:**
Prediction via pattern profiling

#4 **Sparse Modelling:**
Regression subset-structure uncertainty

#5 **Sparse Models and Profiling:**
Gene expression as response: Designed experiments

#6 **Sparse Models and Profiling:**
Gene expression as response: Latent factor models
Patterns of coordinately expressed genes:
  - Signatures

Metagenes
PCA, SVD of expression data

- Biologically selected gene subsets
- Trained subset selection
- Clusters

Cardiovascular disease: High/Low
(Seo, West et al 2004)
Empirical Factor Regression

SVD: $X = ADF$

PCA: $XX' = AD^2A'$

$\beta_f = DA'\beta_x$

$\dim(\beta_f) = n << p = \dim(\beta_x)$

$z = X'\beta_x + \nu$

Genes X

$z = F'\beta_f + \nu$

Metagene factors F

Patterns: Factors "underlying" X are predictors

X variable set selection

$p=n$: Shrinkage priors key

F variable selection
Expression Profiles: Signatures of States

Metagene factor regression: characterising genomic patterns

Predictive + Translational profiling: oncogenic pathway deregulation

(Huang et al 03, Black et al 03)
Out-of-sample prediction

Cell line derived signatures predict differences in oncogenic activity in mouse tumours

\textbf{c-Myc up-expression}

Metagene: gene subset & pattern as a predictor

(Huang et al 03, Black et al 03)
Oncogene Sub-Pathway Profiles: Translation

Single Oncogenes
- pathway characterisation
- potential targets

Cell lines signatures
Human lung cancers (ovarian, breast)

Clinical prognostic
- clinical evaluation
- therapeutic evaluation

(Bild et al 05)
Metagenes in Clinico-Genomic Prognostic Models

Genomic Medicine
Personalised Prognostics

Gene expression clustering
Metagene factors

Non-linear regressions - CART models

Integration:
non-genomic predictors

(Breast cancer - Pittman et al PNAS 04)
Flightplan

#1 Genomics, Microarrays, Data: Big picture

#2 Bayesics - Regression and Shrinkage: Gene expression as predictors

#3 Patterns and Factors: Prediction via pattern profiling

#4 Sparse Modelling: Regression subset-structure uncertainty

#5 Sparse Models and Profiling: Gene expression as response: Designed experiments

#6 Sparse Models and Profiling: Gene expression as response: Latent factor models
Standard Sparsity Priors in Regression

$$(z_i | \beta) \sim N(h_i^T \beta, \sigma^2)$$

$$\beta = \begin{pmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_p
\end{pmatrix}$$

Variable inclusion uncertainty

Large p: parsimony
sparsity

$$\#\{\beta_j \neq 0\} = \text{small}$$

Sparsity priors: $$\beta_j \sim (1 - \pi)\delta_0(\beta_j) + \pi N(\beta_j | 0, \tau)$$

Augment: $$\gamma_j \sim Ber(\pi)$$

$$\gamma = \begin{pmatrix}
\gamma_1 \\
\gamma_2 \\
\vdots \\
\gamma_p
\end{pmatrix}$$

MCMC computation: $$p(\beta, \sigma, \gamma, \tau, \pi | Z)$$
Large p - Shrinkage and Sparsity

Model-based, automatic shrinkage - Simultaneous "multiple tests"

Multiple shrinkage: conservative, parsimonious
Decision theory/false discovery?
Estimation versus Decision?

\[ \pi_j^* = Pr(\gamma_j = 1|Z) = Pr(\beta_j \neq 0|Z) \]
\[ p(\beta_j | \beta_j \neq 0, Z) \]

Model/subset probabilities:
\[ Pr(M_\gamma | Z) \]

Issues:
Collinearity
Multiple related models
Computation with very large p

(Clyde & George StatSci 04)
Stochastic Search Methods

**MCMC “local search” inspired**

Good models “near” good models

Add/drop/replace variables

Move by sampling new model

**KEY:** easily compute

\[ \propto Pr(M_\gamma | Z) \]

Shoot out ALL neighbours: “local proposals”

Swiftly find high probability regions of model space

Catalogue of many “good” models

Parallelisation

(Hans, Dobra, West 05; Rich et al 2005 – p=8400)
Brain cancer expression: \( p=8400 \)

Survival regressions:
- multiple related 3-5 gene subsets
- key cellular motility/infiltration genes
- regression model uncertainty in prediction

(Cancer Research, 05)
Sparsity -
Regression variable in/out probabilities

Dimension -
Implicit in Bayesian & other likelihood-based analyses
(cf. BIC)
Cascade of regression models:
- Models to predict/explain gene expression for survival predictive genes
- and so on ...

Generate Directed Acyclic Graphical models (DAGs) of association patterns in gene expression

Exploratory data analysis, visualization uses

http://graphexplore.cgt.duke.edu

(Cancer Research, 05)
Sparse Graphs from Sparse Regressions

EGFR
Brain cancer gene expression
Duke Keck Center for Neurooncogenomics

(Dobra et al JMVA, 04; Jones et al Stat Sci 05)
These slides:
www.isds.duke.edu/~mw/downloads/SemStat05

Papers, software, many links:
www.isds.duke.edu/~mw

ABS04 web site: Lecture slides, stats notes, papers, data, links:
www.isds.duke.edu/~mw/ABS04

Integrated Cancer Biology Program
icbp.genome.duke.edu

Genome Institute @ Duke
www.genome.duke.edu