Data analysis & modelling of DNA microarray data in in genetic expression profiling

Rainer Spang, Harry Zuzan, Mike West
&
Joe Nevins, Jeff Marks, Seiichi Ishida

Presentation at the

NISS Affiliates Workshop
on Gene Expression Analysis

July 13th, 2000

Project personnel:

- **Rainer Spang & Harry Zuzan**, NISS and ISDS, Duke University
- **Mike West**, ISDS, Duke University
- **Joe Nevins**, Genetics, Duke University Medical Center (DUMC)
- **Jeff Marks**, Surgery and the Cancer Center, DUMC
- **Seiichi Ishida**, Genetics, DUMC

This work is part of a project run under the auspices of two Duke Centers – the Center for Bioinformatics & Computational Biology, and the Center for Genome Technology

Web sites:

- `www.isds.duke.edu`
- `www.isds.duke.edu/bioinformatics`

(Breast Cancer) Phenotyping

Genetic features, patterns

$\updownarrow$

Physiology, Clinical outcomes

Two group problems: Binary outcomes

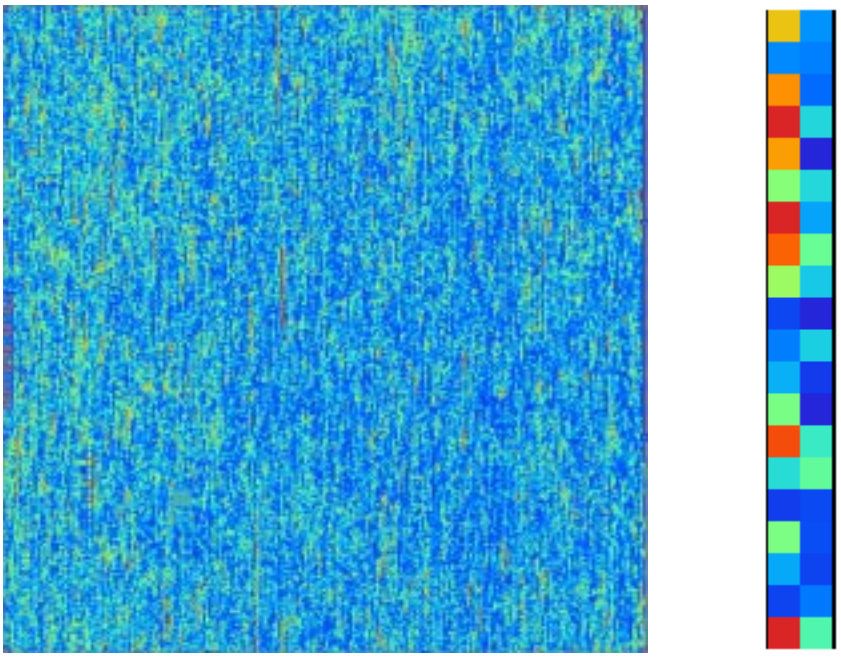
- e.g., ER+ versus ER−
- DNA microarray data: expression levels of 7129 genes (sequences) in RNA from tumour, tumour location, person, time point
- 15 ER+, 12 ER−
- Discriminatory patterns of expression?
- Predictive classification of tumours 28, 29, 30, ... ? – Decision aid
- Which genes are implicated? Surprises? What is new?

Expression array data

Microarray data: Affymetrix arrays

- 20 probe pair sets per gene: $7129 \times 20 \times 2$ intensities
- definitions of expression
- several summary measures
  - (Average) difference
  - (Average) normalised difference
  - (Average) Log Ratio – "fold" changes, level independent
- data issues: imaging, uncertain "manipulation"

One array, one probe set



Summary expression data

$p = 7126$ genes, $n = 27$ arrays

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ x_{p,1} & x_{p,2} & \cdots & x_{p,n} \end{pmatrix}$$

p (variables) $>>$ n (observations)

Understanding summary expression data

Ideas from latent factor modelling

e.g., on microarrays $j = 1, \ldots, 15$ and $k = 16, \ldots, 27$,

$$\mathbf{x}_j = \begin{pmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{p,j} \end{pmatrix} = \begin{pmatrix} 5 \\ \vdots \\ 5 \\ 50 \\ \vdots \\ 50 \end{pmatrix} + \begin{pmatrix} \epsilon_{1,j} \\ \epsilon_{2,j} \\ \vdots \\ \epsilon_{p,j} \end{pmatrix} \quad \text{and} \quad \mathbf{x}_k = \begin{pmatrix} x_{1,k} \\ x_{2,k} \\ \vdots \\ x_{p,k} \end{pmatrix} = \begin{pmatrix} -5 \\ \vdots \\ -5 \\ -50 \\ \vdots \\ -50 \end{pmatrix} + \begin{pmatrix} \epsilon_{1,k} \\ \epsilon_{2,k} \\ \vdots \\ \epsilon_{p,k} \end{pmatrix}$$

Example of factor structure (continued)

$$\mathbf{X} = \mathbf{a}\mathbf{b}' + \mathbf{E}$$

$$\mathbf{X} = \begin{pmatrix} 5 \\ \vdots \\ \\ 50 \\ \vdots \\ 50 \end{pmatrix} (1, \quad 1, \quad \cdots, \quad 1, \quad -1, \quad , -1, \quad , \ldots, \quad -1) + \mathbf{E}$$

a relates subgroups of genes; **b** relates subgroups of microarrays;
E is what's left over

Factor structure in data matrices

Reality is more complicated,

$$\mathbf{X} = \mathbf{a}_1\mathbf{b}_1' + \mathbf{a}_2\mathbf{b}_2' + \cdots \mathbf{a}_k\mathbf{b}_k' + \mathbf{E}$$

- $\mathbf{a}_1, \mathbf{a}_2, \ldots$ represent patterns/relationships among genes
- $\mathbf{b}_1, \mathbf{b}_2, \ldots$ represent patterns/relationships among arrays
- $\mathbf{E}$ is what's left over ...

If $\mathbf{E}$ is small, $\mathbf{X} \approx \mathbf{AB}$

- $\mathbf{A}$ has columns $\mathbf{a}_1, \mathbf{a}_2, \ldots$
- $\mathbf{B}$ has rows $\mathbf{b}_1', \mathbf{b}_2', \ldots$

Identifying decompositions: $\mathbf{B} = \mathbf{DF}$

– $\mathbf{F}$ orthogonal, $\mathbf{A}$ orthonormal columns, $\mathbf{D}$ diagonal –

Singular value (factor) decomposition

$$\mathbf{X} = \mathbf{ADF}$$

$$\begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ x_{2,1} & \cdots & x_{2,n} \\ \vdots & \cdots & \vdots \\ \\ \\ \\ \\ \vdots & \cdots & \vdots \\ x_{p,1} & \cdots & x_{p,n} \end{pmatrix} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ a_{2,1} & \cdots & a_{2,n} \\ \vdots & \cdots & \vdots \\ \\ \\ \\ \\ \vdots & \cdots & \vdots \\ a_{p,1} & \cdots & a_{p,n} \end{pmatrix} \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & d_n \end{pmatrix} \begin{pmatrix} f_{1,1} & \cdots & f_{1,n} \\ f_{2,1} & \cdots & f_{2,n} \\ \vdots & \cdots & \vdots \\ f_{n,1} & \cdots & f_{n,n} \end{pmatrix}$$

Factor decompositions (continued)

$$\mathbf{X} = \mathbf{ADF}$$

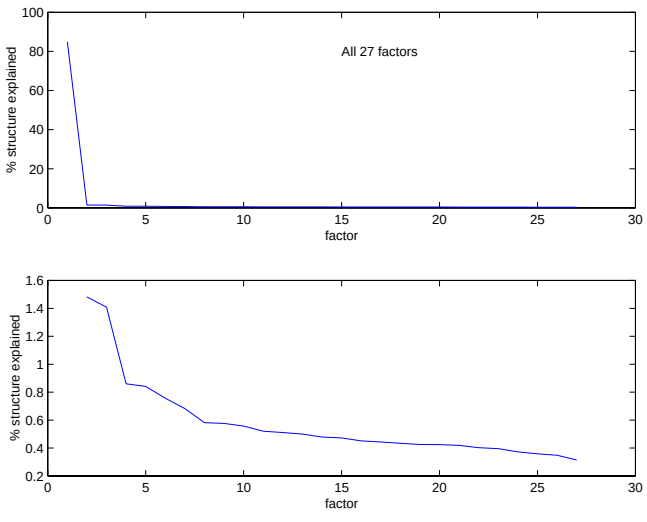
- **Factor loadings matrix** $\mathbf{A}$ has columns $\mathbf{a}_1, \mathbf{a}_2, \ldots$
  - patterns/relationships among genes
- **Latent factors** are rows of $\mathbf{F}$
  - patterns/relationships among arrays
- **Relative importance** of factors 1, 2, are $d_j > 0$
  - large $d_1, \ldots$ small d_n
- **Supergenes**=Factors are linear combinations of expression
  - microarray j: column j of $\mathbf{F}$ is $\mathbf{f}_j = \mathbf{D}^{-1}\mathbf{A}'\mathbf{x}_j$

Principal components analysis: eigenvalues/vectors of (singular) $\mathbf{XX}'$

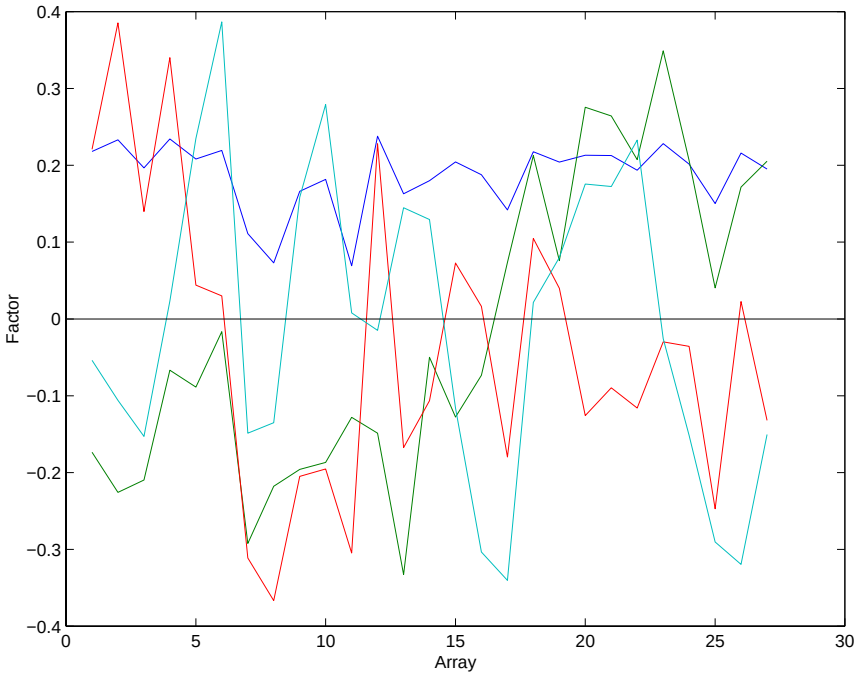
SVD of breast cancer data matrix

$\mathbf{X}$ is 7129×27, $\mathbf{A}$ is 7129×27, $\mathbf{D}$ is diagonal 27×27, $\mathbf{F}$ is 27×27

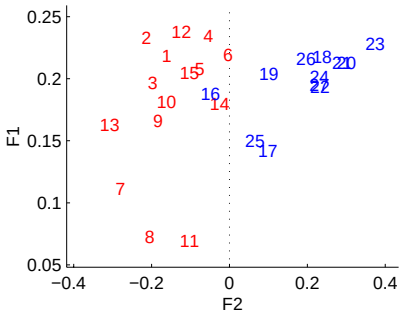
Relative contributions of factors – d_j^2 as % of total:

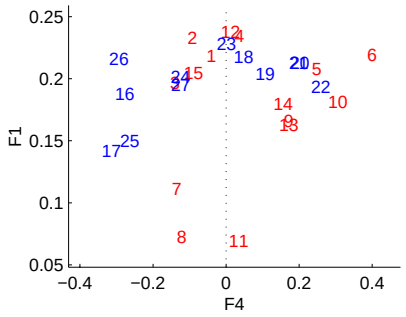


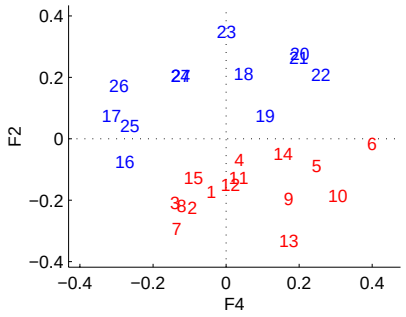
4 of the factors across arrays



Arrays plotted on pairs of 3 factors







Binary regression modelling

- Microarray j, expression profile $\mathbf{x}_j$
- Binary classification: 1 (ER+) or 0 (ER−)
- Probability array j is ER+ is $p(\mathbf{x}_j)$
- Standard "probit" model: $p(\mathbf{x}_j) = \Phi(\mathbf{x}_j'\boldsymbol{\beta})$
- Linear regression on gene expression, mapped to probability scale
  - $\mathbf{x}_j'\boldsymbol{\beta} = \sum_{i=1}^{p} \beta_i x_{i,j}$
  - β_i is regression coefficient on gene i

Binary regression modelling: $p >> n$

$n = 27$ binary observations, $p = 7129$ primary covariates

Data reduction? **Sufficient** data reduction?

$$[\mathbf{x}_1'\boldsymbol{\beta}, \mathbf{x}_2'\boldsymbol{\beta}, \ldots, \mathbf{x}_n'\boldsymbol{\beta}] = \mathbf{X}'\boldsymbol{\beta}$$

or

$$(\mathbf{DF})'\boldsymbol{\gamma}$$

with

$$\boldsymbol{\gamma} = \mathbf{A}'\boldsymbol{\beta}$$

Regression on genes (in $\mathbf{X}$) translates to regression on **supergenes** – the factor variables in $\mathbf{F}$ – weighted by elements of $\mathbf{D}$

Regression on supergenes: Priors

- Supergene (factor) model has $n = 27$ orthogonal covariates
- Precisions d_j^2 related to data on element γ_j of $\boldsymbol{\gamma}$
- Suitable priors on $\boldsymbol{\gamma}$
  - elements γ_j independent (orthogonality)
  - conditionally conjugate: $\gamma_j \sim N(0, \tau_j^2/d_j^2)$
  - differing scales τ_j
  - hyperprior on τ_j

Prior on $\boldsymbol{\gamma} = \mathbf{A}\boldsymbol{\beta}$ must be derived from prior on $\boldsymbol{\beta}$

Regression on genes: Priors

Prior: $p(\boldsymbol{\beta})$

- Generalised singular "g-prior"
- "Shape" defined by $\mathbf{A}$, "scales" to be estimated

$$p(\boldsymbol{\beta}) \propto \exp\{-\boldsymbol{\beta}'(\mathbf{A}\mathbf{D}\mathbf{T}^{-1}\mathbf{D}\mathbf{A}')\boldsymbol{\beta}/2\}$$

with

$$\mathbf{T} = \mathrm{diag}(\tau_1^2, \ldots, \tau_n^2)$$

Standard g-prior: τ_j^2 =constant, and prior non-singular

Regression on genes via supergenes

Prior for β: centred at $\mathbf{0}$

- neutral: implied classification probability of 0.5
- multiple solutions to $\boldsymbol{\gamma} = \mathbf{A}'\boldsymbol{\beta}$
- $\boldsymbol{\omega}$ in null space of $\mathbf{A}$: $\mathbf{A}'\boldsymbol{\beta} = \mathbf{A}'(\boldsymbol{\beta} + \boldsymbol{\omega})$
- same posteriors for priors on $\boldsymbol{\beta}$ with mean $\boldsymbol{\omega}$
- choose $\boldsymbol{\omega} = \mathbf{0}$ to identify unique prior, unique posterior

Posterior for β: Defined uniquely by $\boldsymbol{\beta} = \mathbf{A}\boldsymbol{\gamma}$

- 27 supergene parameters $\rightarrow$ 7129 gene parameters
- Posterior simulation samples for $\boldsymbol{\beta}$
- Posterior mean, significance of genes, effects of genes

Latent normals in probit regression

$$p(\mathbf{x}_j) = Pr(\text{array } i \text{ is } ER+) = Pr(y_i \geq 0)$$

where

$$y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, 1)$$

or, as vector for all arrays,

$$\mathbf{y} \sim N(\mathbf{X}'\boldsymbol{\beta}, \mathbf{I})$$

or

$$\mathbf{y} \sim N(\mathbf{F}'\mathbf{D}\boldsymbol{\gamma}, \mathbf{I})$$

- critical to model fitting/computation
- interpretation and implied *kernel regression structure*

Kernel regression structure under GSG prior

Marginalising over $\boldsymbol{\beta}$ implies

$$\mathbf{y} \sim N(\mathbf{0}, \mathbf{K})$$

with *kernel covariance matrix*

$$\mathbf{K} = \mathbf{F}'\mathbf{T}\mathbf{F} + \mathbf{I}$$

with

$$\mathbf{T} = \text{diag}(\tau_1^2, \ldots, \tau_n^2)$$

- correlations between arrays
- effective dependence structure with respect to classification
- key role of $\mathbf{T}$
- effective *non-linear classifier*

Regression analysis & computation

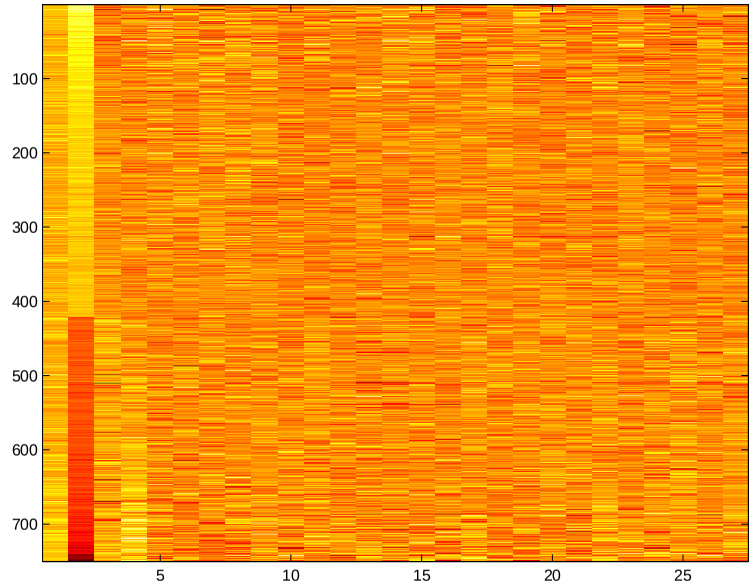
Prior: $\prod_{j=1}^{n} p(\gamma_j|\tau_j)p(\tau_j)$

- Posterior: **Easy** MCMC: iterative simulation to impute
  - latent normals $\mathbf{y}$ underlying probit
  - each γ_j, τ_j
- Hierarchical: prior on τ_j^2 – sensitivity to hyperparameters
- Variants: model selection priors, extra-probit variation, ...

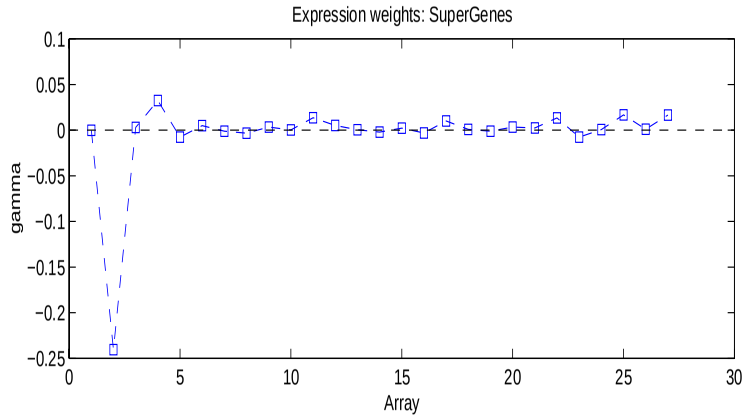
Out-of-sample or **Cross-validatory** prediction:

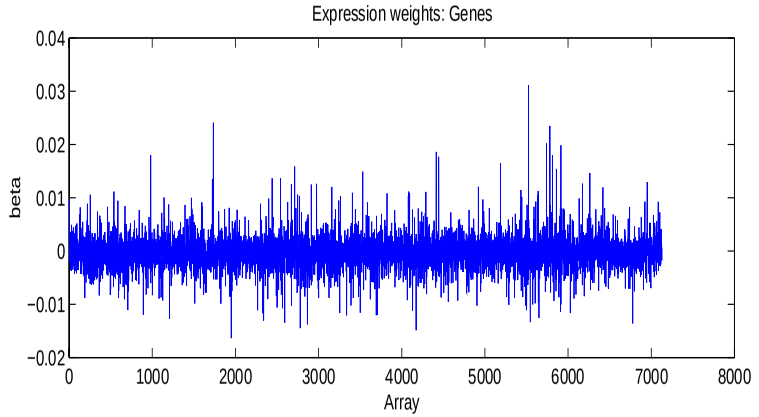
- New tumour on microarray – data $\mathbf{x}$: Posterior for $p(\mathbf{x}) = \Phi(\mathbf{x}'\boldsymbol{\beta})$
- One-at-a-time analysis: Treat array i as having "Missing" binary indicator

Factor loadings A for top 750 genes



Estimated regression coefficients





Top 10 genes with + beta coefficients

High in ER+, Low in ER−

... all almost surely > 0 ...

- **mrna for oestrogen receptor**
- **ps2 protein gene**
- intestinal trefoil factor mrna, complete cds
- nat1 gene for arylamine n-acetyltransferase
- mrna for cardiac gap junction protein
- complement component c4a gene
- **breast cancer, estrogen regulated liv-1 protein (liv-1) mrna**
- 5t4 gene for 5t4 oncofetal antigen
- nedd-4-like ubiquitin-protein ligase wwp1 mrna
- hepatocyte nuclear factor-3 alpha (hnf-3 alpha) mrna

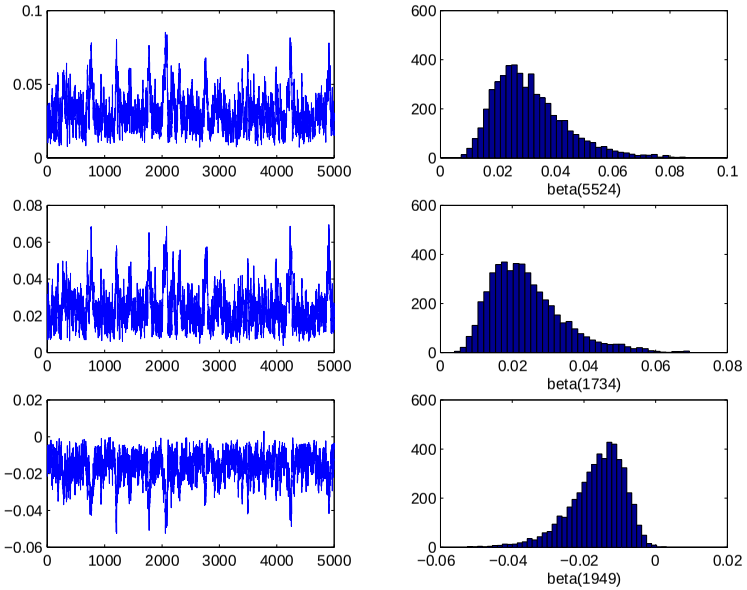
Top 8 genes with – beta coefficients

Low in ER+, High in ER−

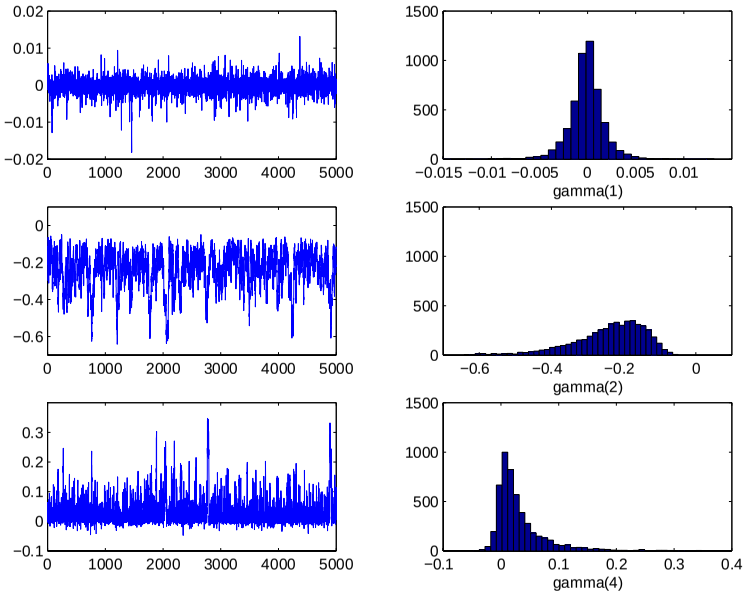
... all almost surely < 0 ...

- (hybridoma h210) anti-hepatitis a igg variable region
- rearranged immunoglobulin lambda light chain mrna
- ig alpha 2=immunoglobulin a heavy chain allotype 2
- omega light chain protein 14.1 (ig lambda chain related) gene
- sry-related hmg-box 12 protein
- matrilysin gene
- guanylate binding protein isoform i (gbp-2) mrna
- cystic fibrosis antigen mrna, complete cds

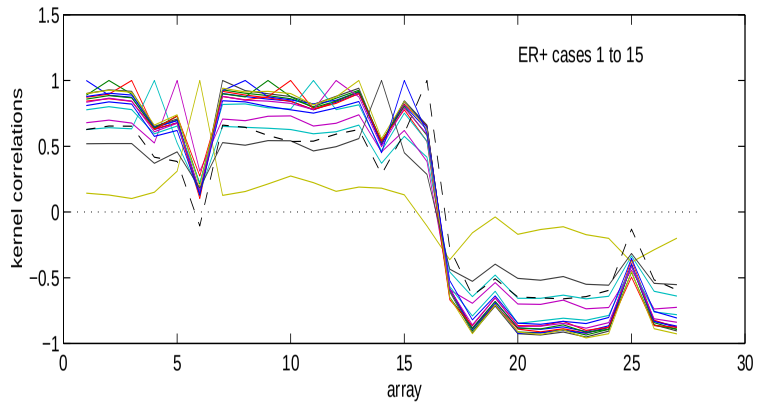
Posterior for 3 beta coefficients

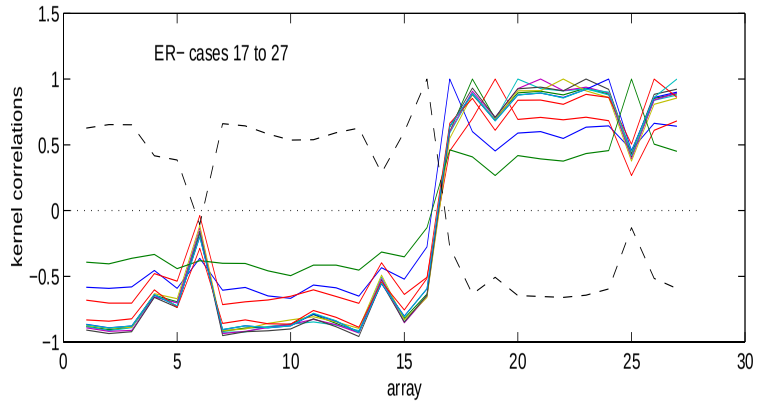


Posteriors for 3 gamma coefficients

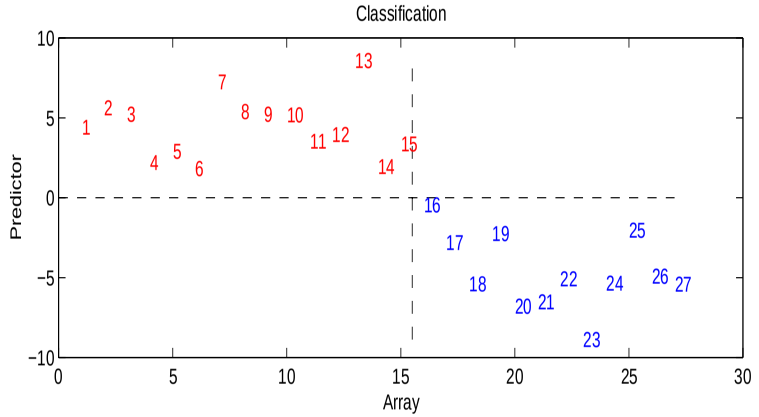


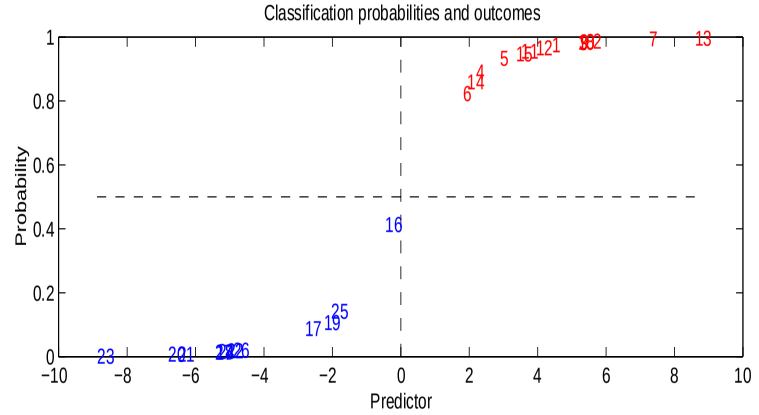
Estimated kernel correlation structure



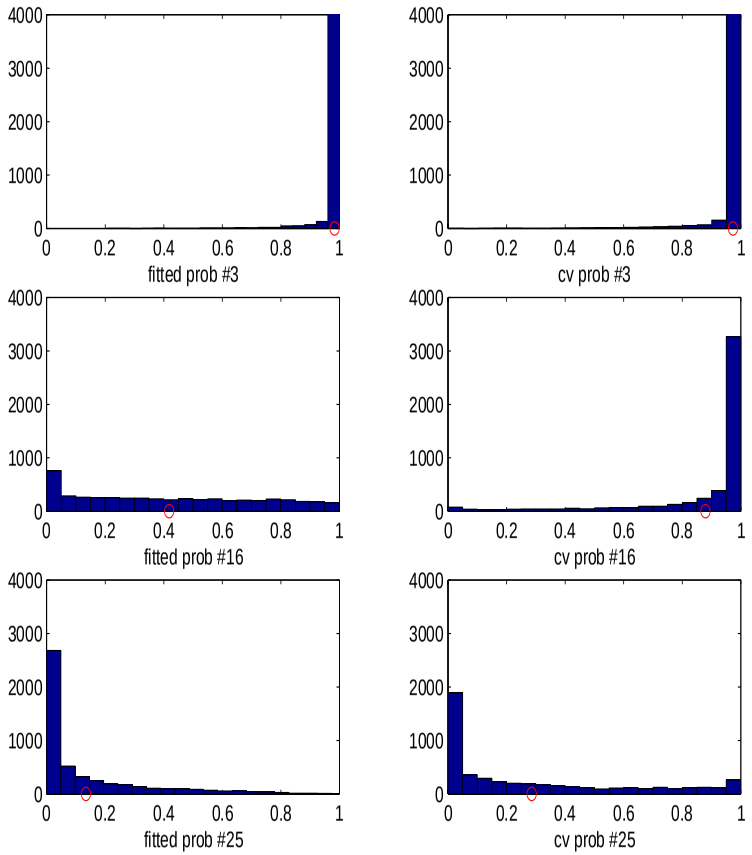


Fitted classification

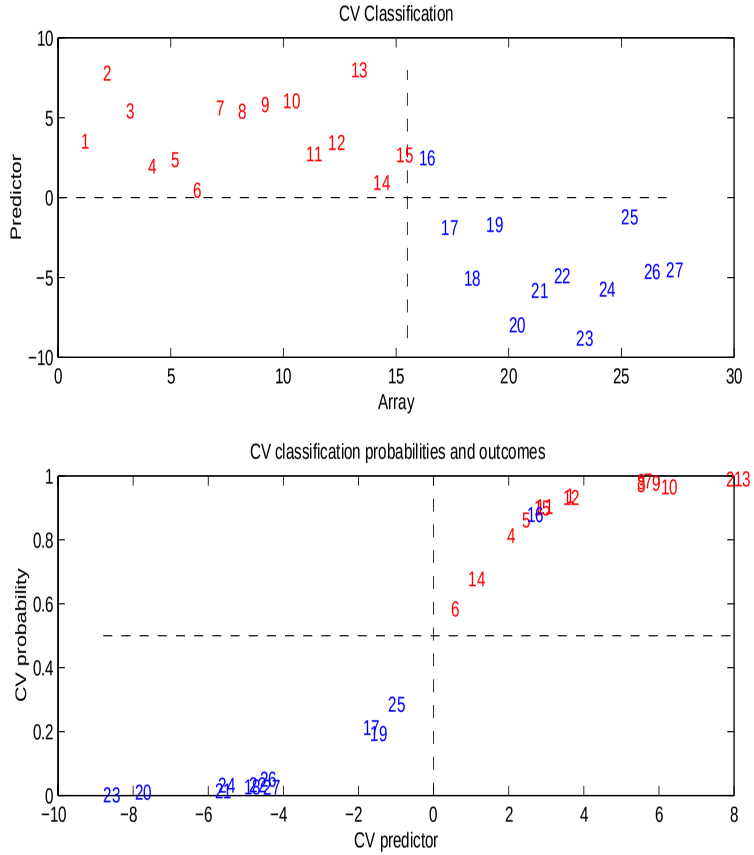




Uncertainty in classification

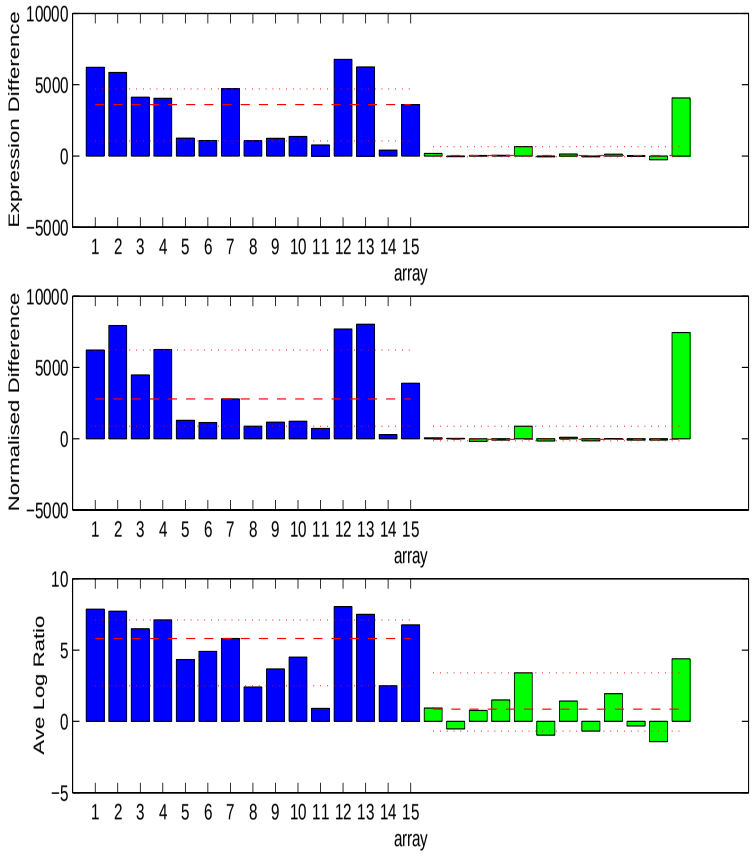


Cross-Validatory predictions

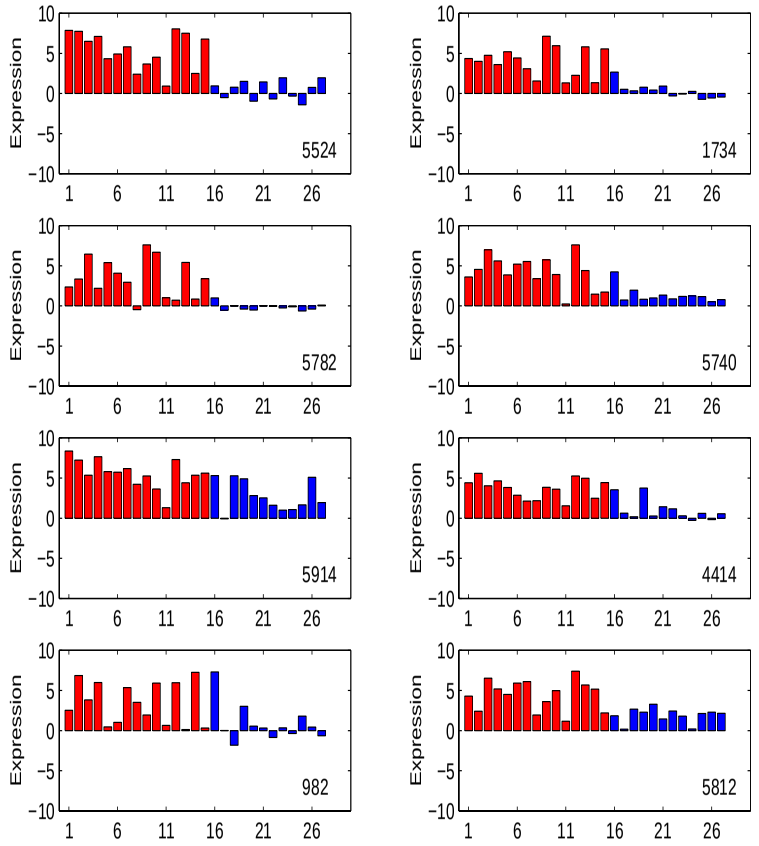


Expression measures: Top + gene

Expression Difference, Normalised difference & Log-Ratio for **oestrogen receptor gene**

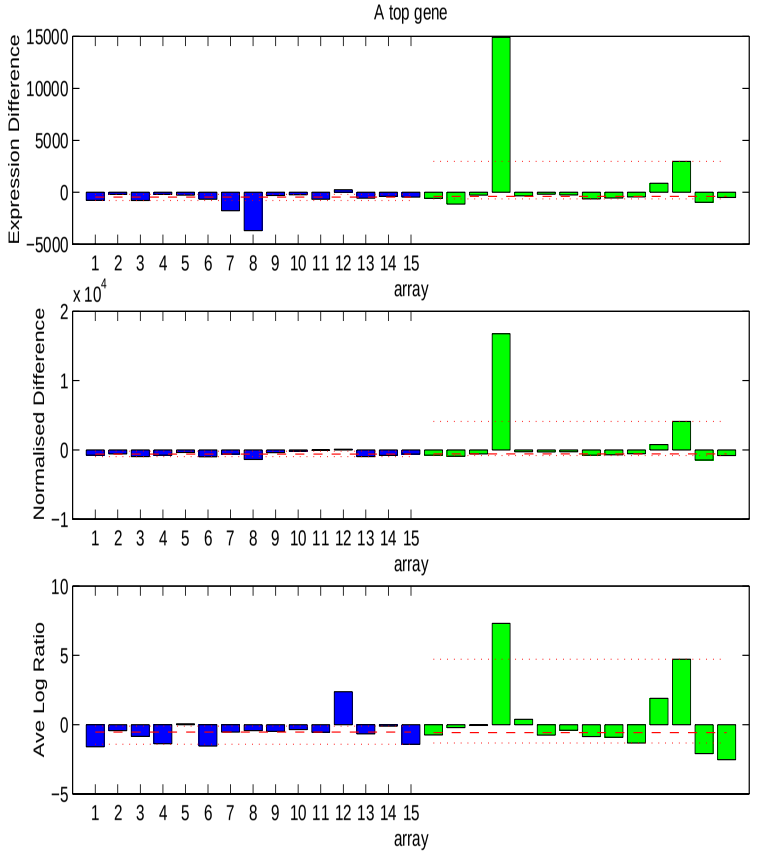


Expression by arrays: top 8(+) genes



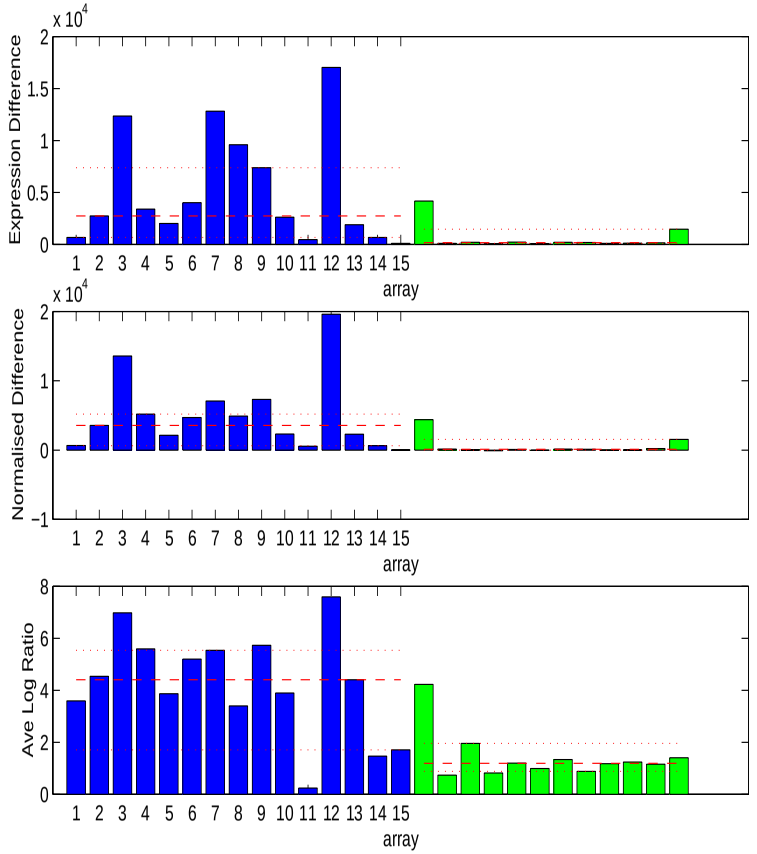
Expression measures: A – gene

Expression Difference, Normalised difference & Log-Ratio



Expression measures: A + gene

Expression Difference, Normalised difference & Log-Ratio



Screening for "Top" genes

- Screen genes by ordering $|E(\beta_i|Data)|$
- Select "top k"

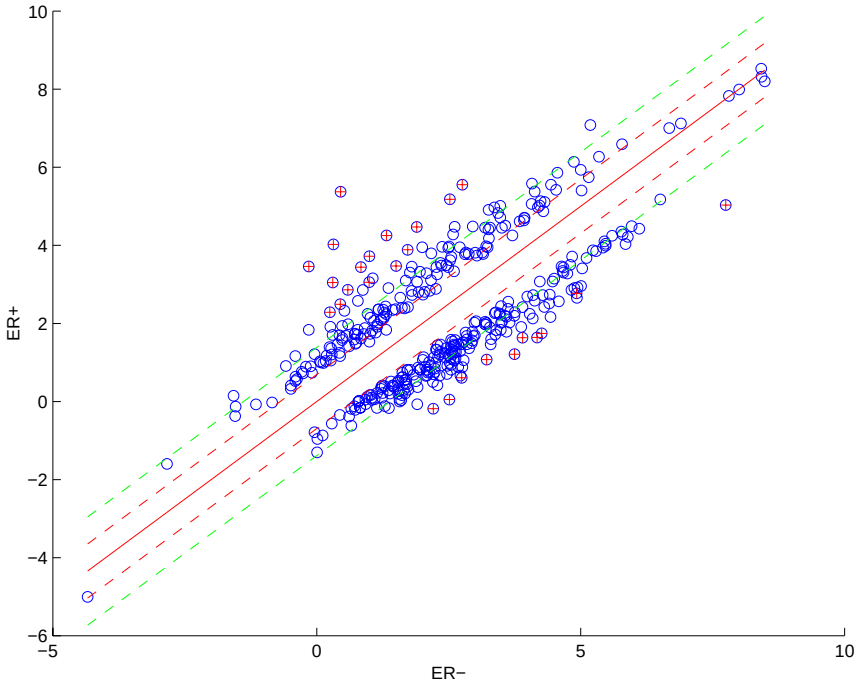
– Screen and refit model –

- Summary based on top $k = 400$ genes
- Refines classification: array 16 is "better" classified
  - "top genes" favour correct classification ER−
  - some/many genes screened favour ER+
  - learning opportunities vs. simple classification
- Warns against simple screening and use of subsets of "discriminatory" genes
- Model analysis as "filter"

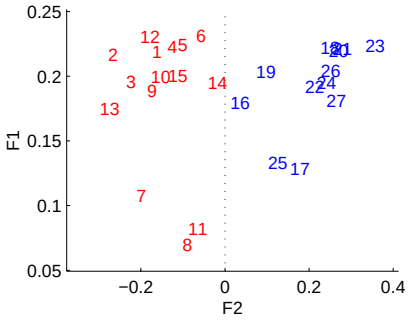
Sample 16

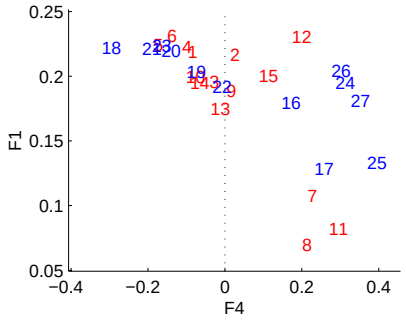
- ER+ or ER−?
- High uncertainty about $Pr(ER+)$
  – small sample size –
- Explore "top genes"
- Oestrogen gene appropriately "down"
- Others are "up"
  - PS2 protein gene: 2nd top gene
  - Liv-1 protein gene: 7th top gene
- Both regulated by oestrogen receptor
- High on array 16, as are others
- – Other regulators of PS2, Liv-1 ... ?
- – ER status determination ... ?

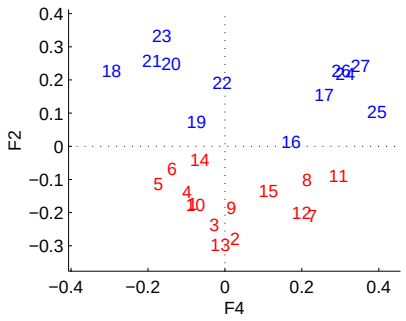
Top 400 genes: Ave(expression) on ER+ versus ER−



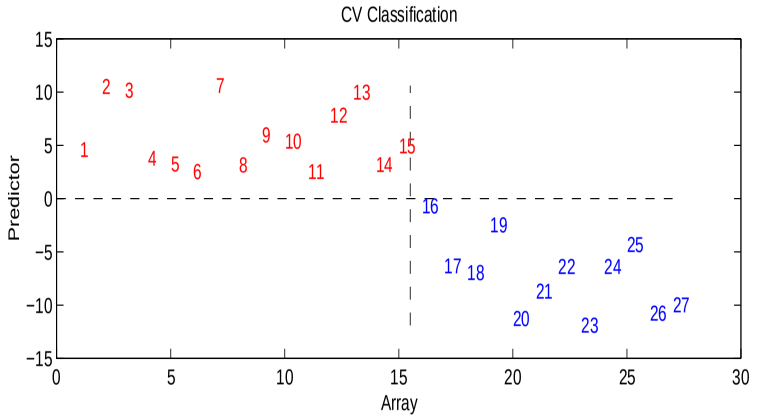
Arrays on pairs of 3 factors: Top 400 genes

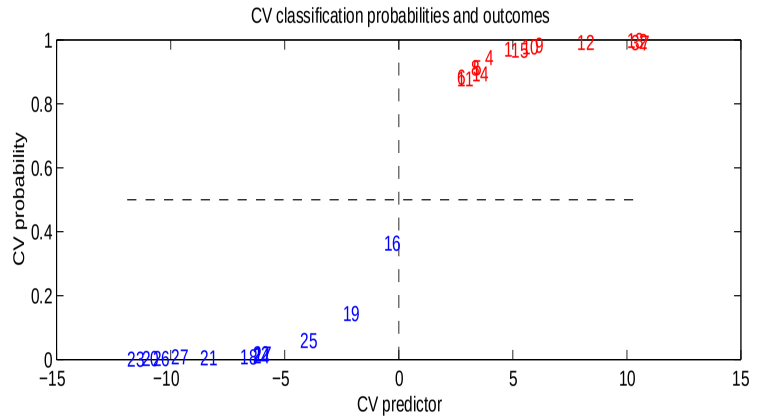




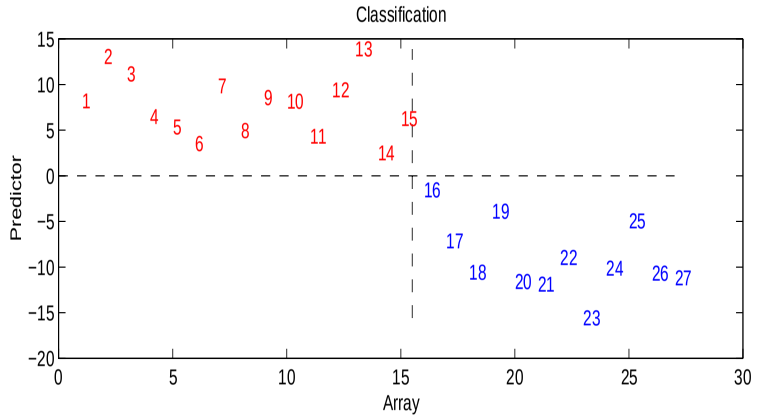


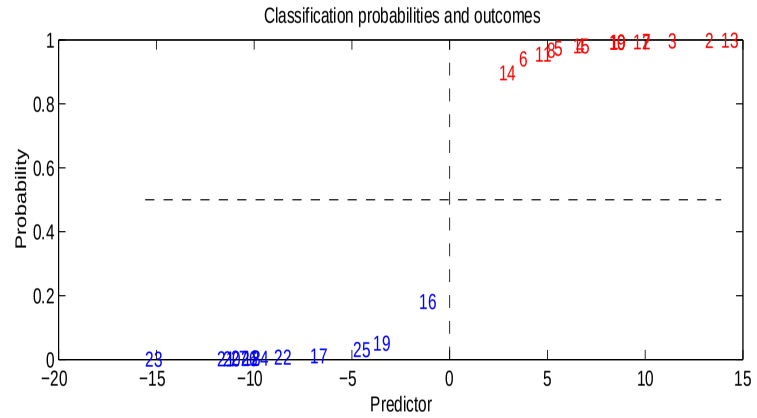
CV predictions: Top 400 genes





Fitted classification: Top 400 genes





Current and Future

Application(s)

- Other binary outcomes (remission/not, ...)
- More data
- Multiple outcomes: cancer stages/states
- Multiple expression measures
- Other biomedical contexts

Methodology

- Small samples/low information:
  – MCMC convergence, prior specifications –
- Errors in variables: uncertain expression
- Multiple outcome categories: multinomials
- Non-linear regressions (kernel methods)

Uncertainty: Top 400 genes