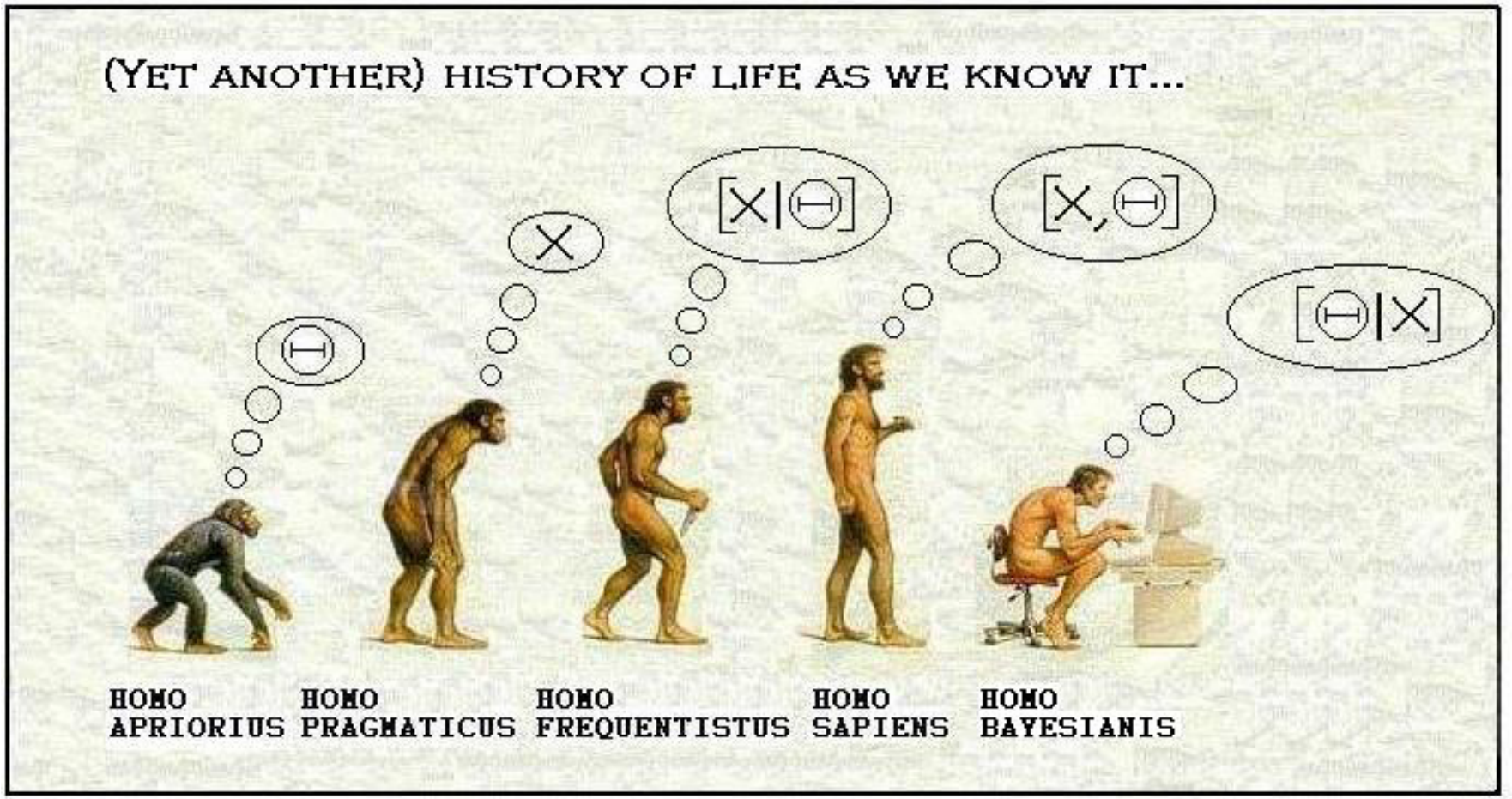
Expression Data Exploration:

Association, Patterns, Factors & Regression Modelling


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
Θ
X
$[X|\Theta]$
$[X,\Theta]$
$[\Theta|X]$
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

Exploring gene expression data

Scale factors, *median chip* correlation on gene subsets for crude 'data quality' investigation

Selecting genes correlated with an outcome, or a gene

Image displays: Expression intensity images

Common structure in gene subsets – Underlying Factors

- Principal components/Singular value decomposition

Data sets: exploration in Matlab
Some data, papers linked to MW info@ABS2004 web site
More data in papers on Duke genomics web site

Some data sets linked to MW info@ABS2004 web site

- 2001 PNAS paper: Duke breast cancer genomics (49 tumours)
- 2003 Lancet, 2004 PNAS papers: Duke-KFSYS breast cancer - clinico-genomics (158 tumours, clinical data)
- 2003 Nature Genetics paper: Mice cell lines and tumours in controlled experiments on several key oncogenes
- MIT/Whitehead 1999 Leukemia data

Regression models

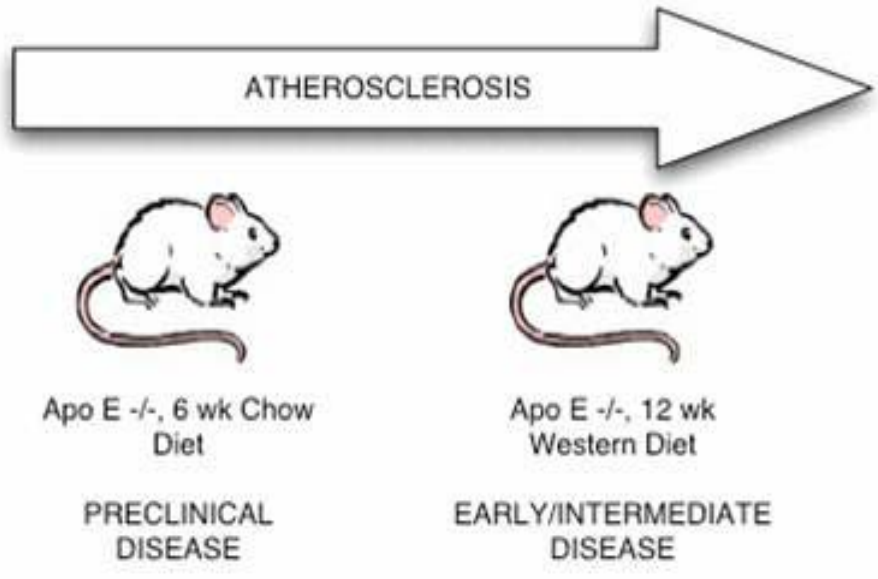
Gene expression as response

Designed experiments: e.g. Mice models: age, sex, diet, genotype

- Finding genes, effects of covariates
- Grouping genes by effects
- Patterns of coordinately expressed genes
- Signature of effect – multiple genes

Predictive interest: Human disease

p genes: Multiple models in parallel



Regression models

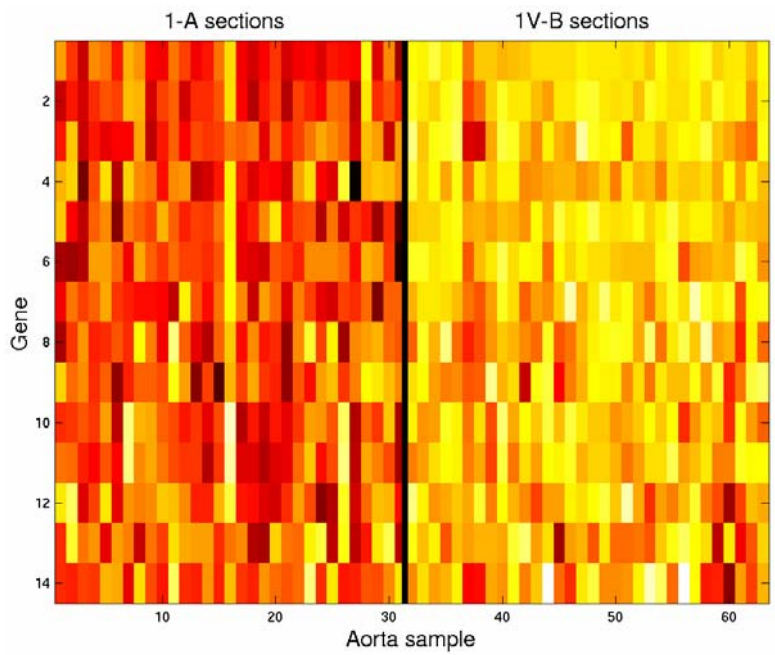
Gene expression as covariates (predictors)

Molecular phenotyping: e.g. Predict aggressive vs. benign tumour
Disease susceptible vs. resistance

- Finding genes linked to response
- Grouping genes
- Patterns of coordinately expressed genes
- Signature of effect – multiple genes

Predictive interest: Human disease

Multiple genes as predictors



Empirical Factors:

Principal Components ~ Singular Value Decompositions

SVD of Expression Data Matrices

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{2,1} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p,1} & x_{p,2} & \cdots & x_{p,n} \end{pmatrix}$$

n arrays

p genes

"Tall & Skinny" - *p>>n*

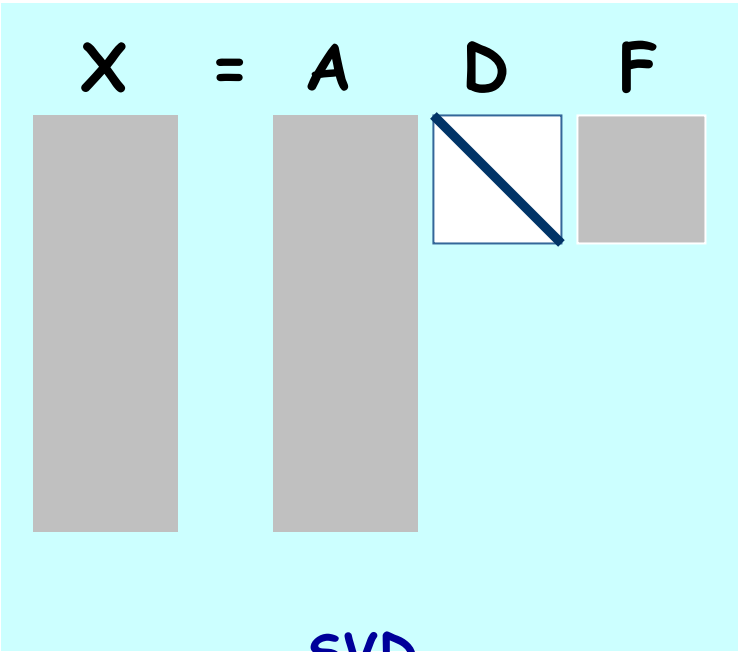
Gene-gene associations:
sample variances/covariances
- centered data (row means 0)

$$(n-1)^{-1}\mathbf{X}\mathbf{X}'$$

Collinearities
Co-dependencies
Co-regulation

Transcription factors
Cascading expression
Pathways, interactions

SVD of Data Expression Matrices



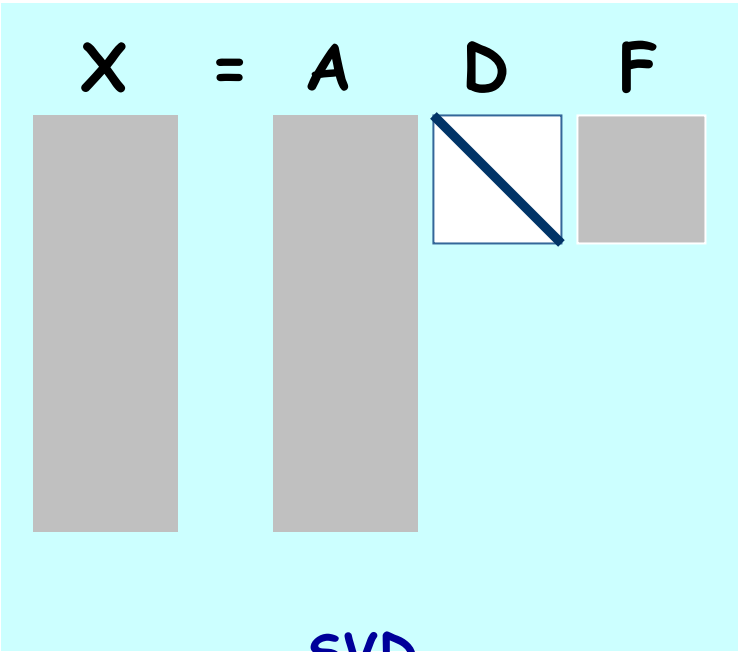
A *pxn* orthonormal columns: $\mathbf{A'A=I_n}$

D *nxn* diagonal, non-negative

F *nxn* orthogonal: $\mathbf{F'F=FF'=I_n}$

Principal Components (PCA): $\mathbf{XX' = AD^2A'}$

SVD of Expression Matrices



F - factors:

rows are *n* factors

columns are *n* samples

A – loadings:

A_{ij} loads gene i on factor j

D – singular values

relative importance of factors

Patterns of covariation observed among genes is "driven" by underlying factors

- mediated by loadings -

Exploring Expression Data

Data sets: exploration in Matlab

Gene subset selection
SVD of smaller subsets "of interest"
Loadings to order genes

Links to clustering – group/cluster by loadings

Scatter plots, Image displays

Signatures: Dominant pattern – Metagene pattern

Elements of Regression

Regression models : ideas & theory (see notes)

Linear model $\mathbf{y} = \mathbf{H}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}/\phi)$

Least Squares

$$\mathbf{V}^{-1} = \mathbf{H}'\mathbf{H}$$

$$\hat{\boldsymbol{\beta}} = \mathbf{V}\mathbf{H}'\mathbf{y}$$

(minimal) Bayes: Shrinkage priors

Decision theory
Regularisation - Ridge regression
Key with many predictors
Relevance of zero-mean location

Prior $\boldsymbol{\beta} \mid \phi \sim N(\mathbf{0}, \mathbf{C}^{-1}/\phi)$

Posterior $\boldsymbol{\beta} \mid \mathbf{y}, \phi \sim N(\mathbf{b}, \mathbf{B}^{-1}/\phi)$

$$\mathbf{B} = \mathbf{C} + \mathbf{H}'\mathbf{H}$$

$$\mathbf{b} = \mathbf{B}^{-1}\mathbf{H}'\mathbf{y}$$

Intercept term: 'Large' $C_{1,1}^{-1}$
other $C_{i,i}^{-1} = \tau$ *(covariates scaled)*
Common shrinkage towards 0 - 'controlled' by τ

Regression models : full theory

Gamma prior/posterior on precision
Normal posterior for β becomes T

Prior $\phi \sim Ga(a/2, b/2)$

Posterior

$$\phi \mid \mathbf{y} \sim Ga(a'/2, b'/2)$$

$a' = a + n$ $b' = b + \mathbf{y}'(\mathbf{y} - \mathbf{Hb})$

$$\boldsymbol{\beta} \mid \mathbf{y}, \phi \sim N(\mathbf{b}, \mathbf{B}^{-1}/\phi)$$

$$\boldsymbol{\beta} \mid \mathbf{y} \sim T_{a+n}(\mathbf{b}, \mathbf{B}^{-1}s)$$

Residual variance estimate $s = (b + q)/(a + n)$

Marginal likelihood (to assess shrinkage) $p(\mathbf{y})$

Prediction of new samples: $p(y_{new}|\mathbf{y})$

Bayesian/Shrinkage elements:

- LSE as limiting case (no shrinkage) $\tau \to \infty$
- Shrinks when it matters – weak/no association $\tau \to 0$
- Marginal likelihood to assess shrinkage degree
- Theoretical dominance of shrinkage estimation
- Regularisation: acts against over-fitting, improves stability and robustness in prediction

Examples (see Matlab code explorations)

- Explore genes predictive of others – collinear expression patterns, gene co-regulation (e.g., ER in breast data)

- Design effects: finding genes related to genetic interventions, environmental factors (e.g., mice cell lines and tumours, time course experiments)

Simple Designed Experiment Example

Gene pathway studies -

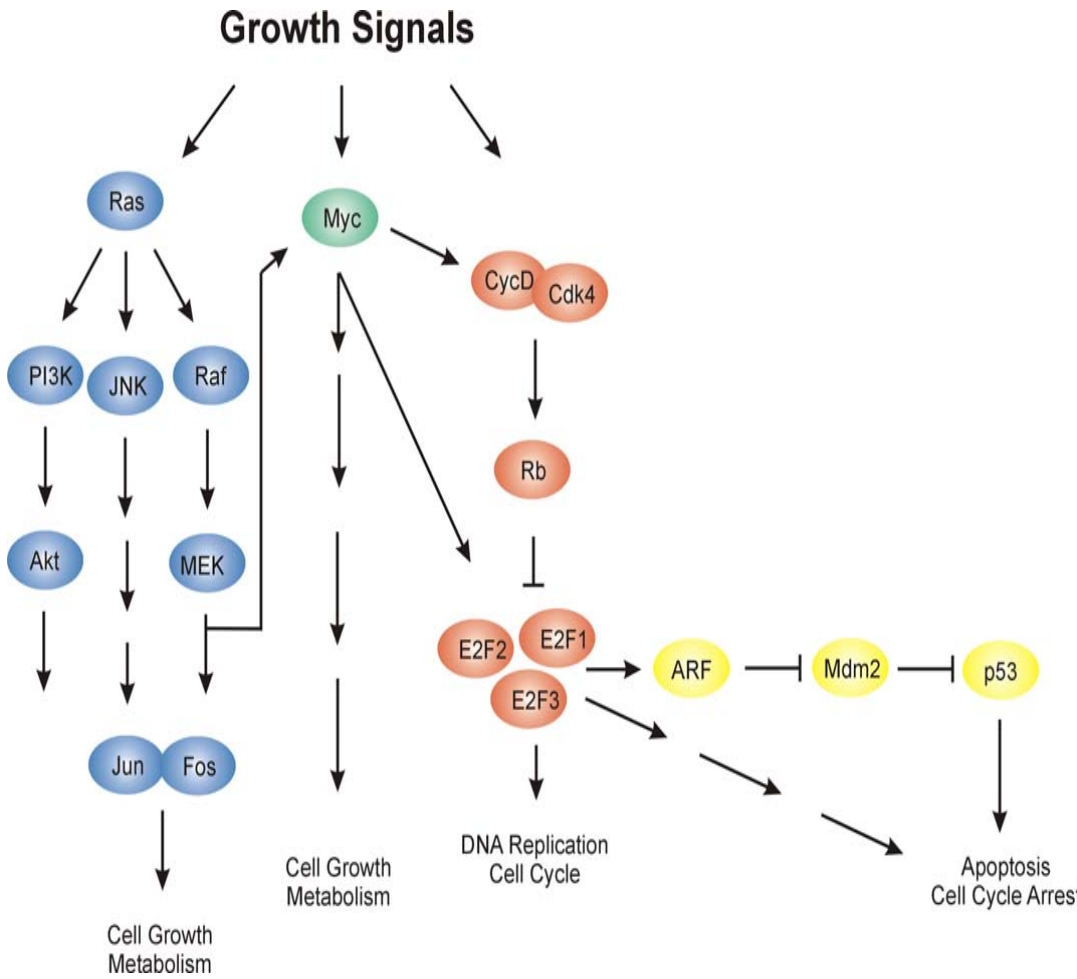
Myc & Ras oncogenes
E2F transcription factors

(Nat Gene 2003, Huang et al)

Genes and gene subsets (signatures) related to up-regulation of oncogenes.

Compare & characterise oncogenic states

Why?



Simple Myc/Ras/E2F Example

y = a	Wild type mice (baseline)
+m	if Myc up-expressed
+r	if Ras up-expressed
+mr	if Myc & Ras up
+e1, or +e2, or +e3	if E2F1, 2 or 3 up
+ε	

Design matrix H 0/1 entries

Same for each gene

Parallel processing to compute posterior, predictive summaries

Examples (see Matlab code explorations)

A more elaborate example: Mice & Heart Disease

Factor Regression

Linear Regression on SVD Factors

Linear model with genes as predictors

$$\mathbf{y} = \mathbf{H}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \mathbf{H} = \mathbf{X}'$$

SVD implies

$$\mathbf{y} = \mathbf{F}'\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\theta} = \mathbf{D}\mathbf{A}'\boldsymbol{\beta}$$

$$\boldsymbol{\beta} = \mathbf{A}\mathbf{D}^{-1}\boldsymbol{\theta}$$

Regression on *n*<<*p* factors

Massive dimension reduction
Many-one: No unique inverse

Common "optimal" least-norm inverse

Priors: Normal priors are consistent/coherent
Generalised g-priors (West, Bayesian Statistics 7, 2003)

Linear Regression on SVD Factors

Prior

$$\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{C}^{-1})$$

$$N(\mathbf{0}, \mathbf{C}^{-1} / \phi)$$

Posterior

$$\boldsymbol{\theta} \mid \mathbf{y}, \phi \sim N(\mathbf{g}, \mathbf{G}^{-1})$$

$$\mathbf{G} = \mathbf{C} + \mathbf{I}\phi$$

$$\mathbf{g} = \phi \mathbf{G}^{-1} \mathbf{F}\mathbf{y}$$

Implied posterior

$$\boldsymbol{\beta} \mid \mathbf{y}, \phi \sim N(\mathbf{A}\mathbf{D}^{-1}\mathbf{g}, \mathbf{A}(\mathbf{D}\mathbf{G}\mathbf{D})^{-1}\mathbf{A}')$$

Extension of shrinkage prior:
Multiple shrinkage

Different degrees of shrinkage for different factor dimensions

Estimation of multiple shrinkage factors?

$$\mathbf{C}^{-1} = diag(\tau_1, \ldots, \tau_n)$$

MCMC – Gibbs Sampling in Factor Regression

Multiple shrinkage prior model

Simulate: Iteratively resample conditional posteriors

Sample means, histograms MC approximation of posterior quantities of interest

$$\boldsymbol{\theta} \mid \mathbf{y}, \phi \sim N(\mathbf{g}, \mathbf{G}^{-1})$$

$$\phi \mid \boldsymbol{\theta}, y \sim Ga((a+n)/2, (b+\varepsilon'\varepsilon)/2)$$

$$\varepsilon = \mathbf{y} - \mathbf{F}'\boldsymbol{\theta}$$

.. and for shrinkage variances:

$$\tau_i^{-1} \sim Ga(k/2, h/2)$$

$$\tau_i^{-1} \mid \boldsymbol{\theta} \sim Ga((k+1)/2, (h+\theta_i^2)/2)$$

MCMC – Gibbs Sampling in Factor Regression

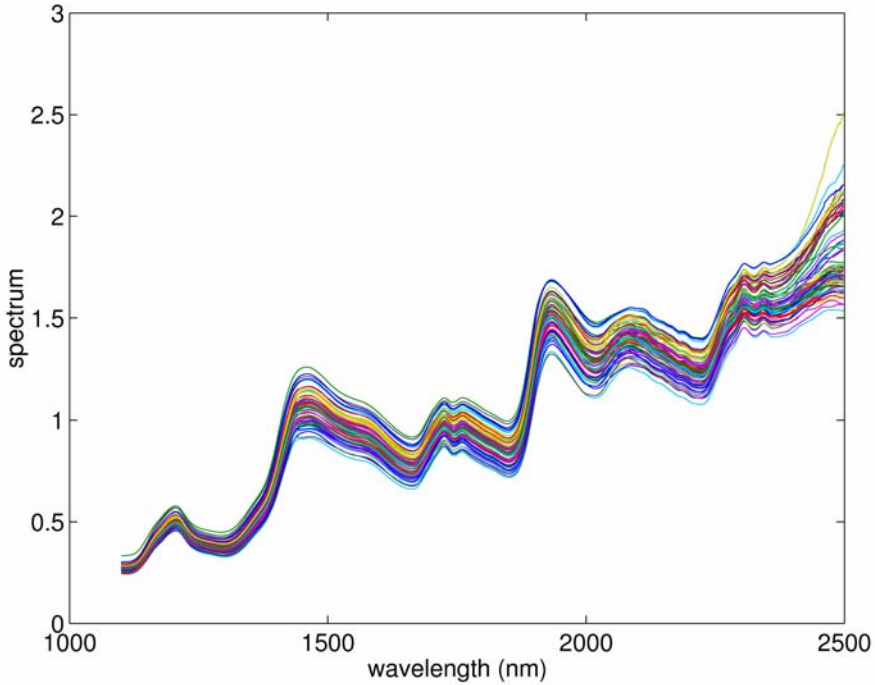
Illustrative/exploratory example:
Breast cancer predicting ER gene RNA levels to identify
Metagene predictor of ER
-pathway related genes –

PNAS 2004 data and code

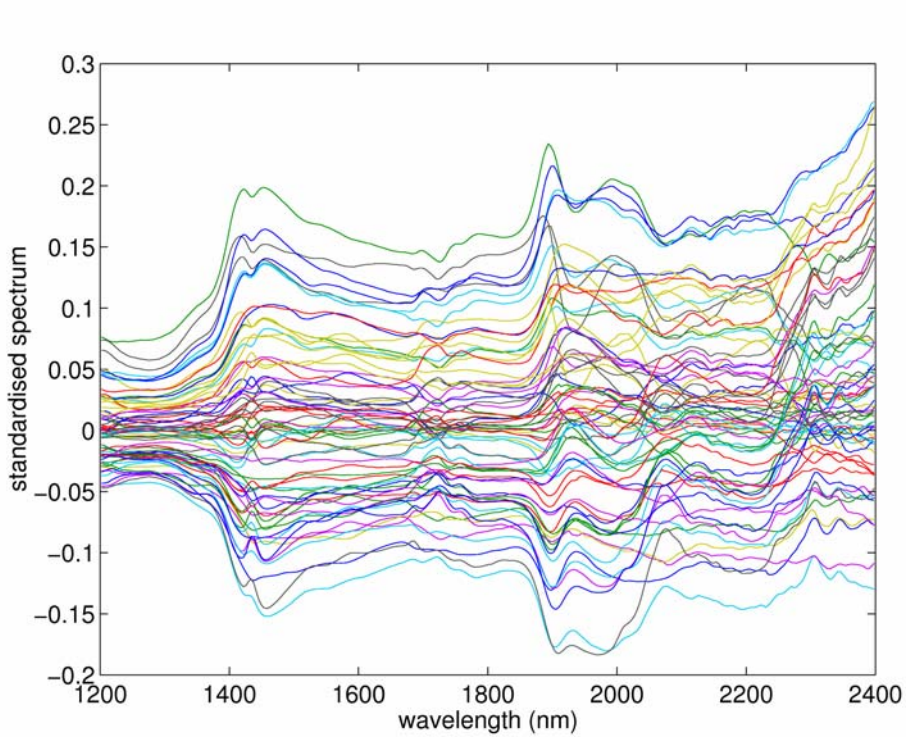
Second, non-genomic example:
Predicting fat content of cookies from infrared mass spectroscopy
- covariates are spectra/finely sampled curves -
(Bayesian Statistics 7 paper/M West)

39 training samples, 39 to predict
p=300 wavelengths on NIR reflectance spectra:
Aim to produce model to predict fat of future cookies

Cookie Data

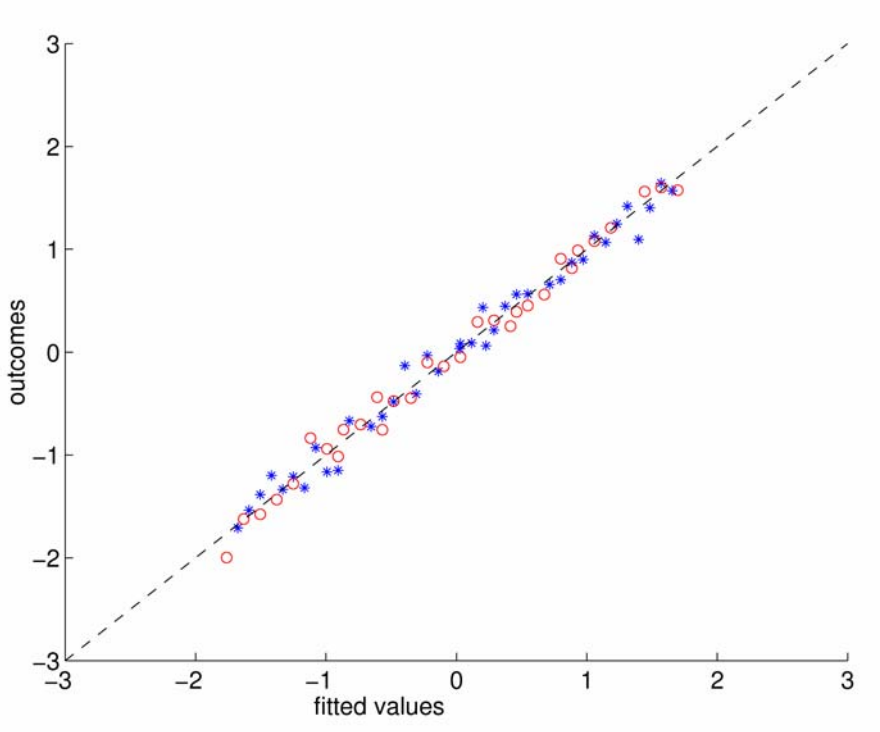


78 spectra



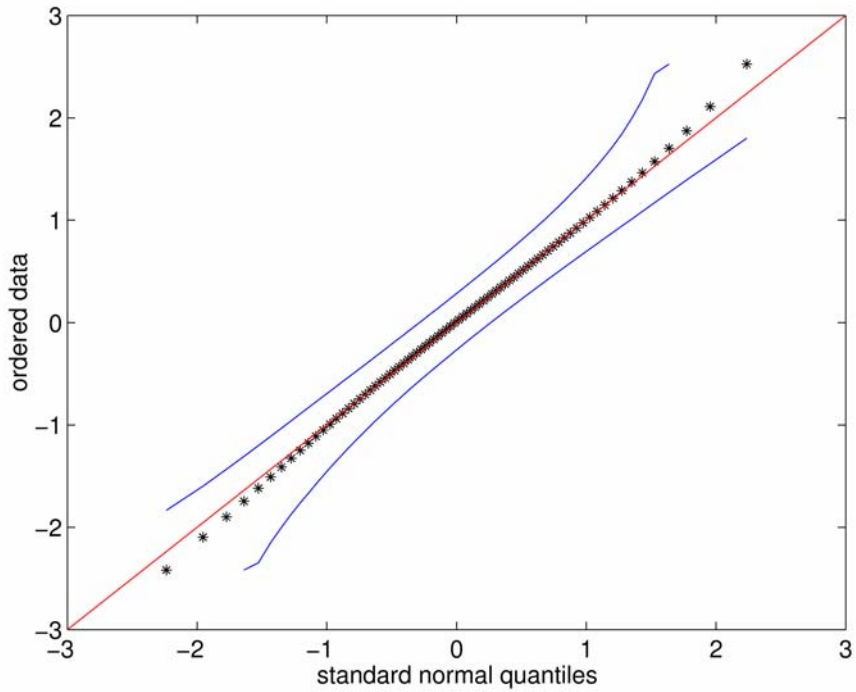
standardised spectra

Cookie Data



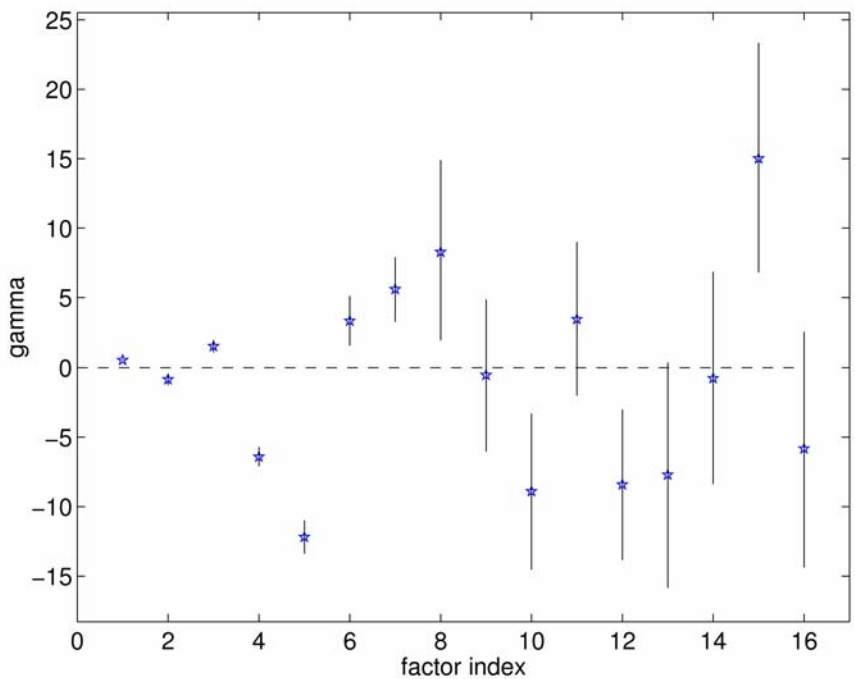
Data and fitted/predicted values

* training, o predicted/validation

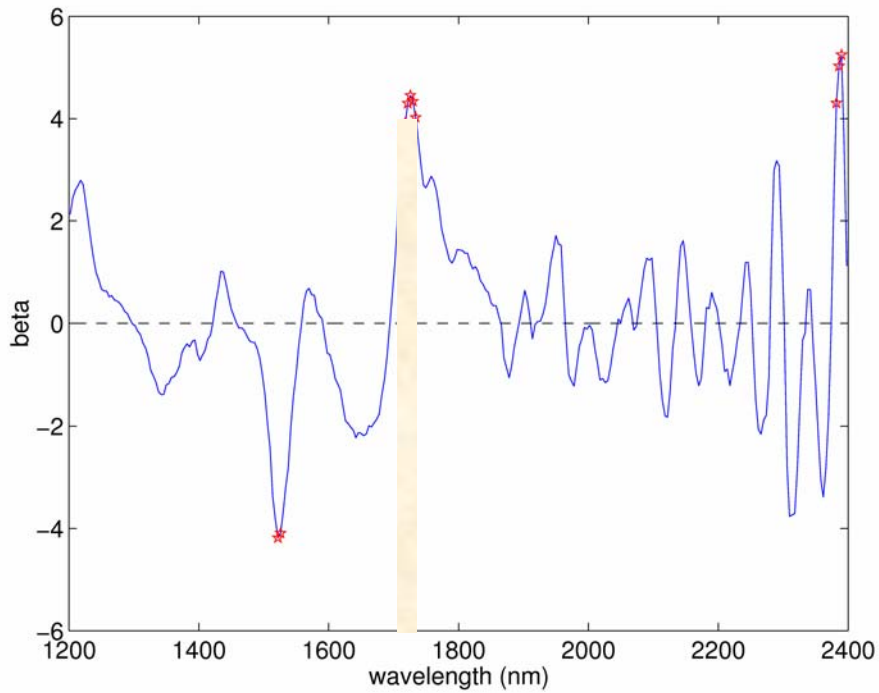


Bayesian qqplot of residuals

Cookie Data



Posterior summaries: θ



Posterior means: β

1710-1730: charactistic fat absorbance range

Coherent Theoretical Basis for SVD Regression?

Model & design data

$$\mathbf{y} = \mathbf{H}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

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

Transform

$$\mathbf{y} = \mathbf{F}'\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\theta} = \mathbf{DA}'\boldsymbol{\beta}$$

Prior ?

$$\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{C}^{-1})$$

?

$$\boldsymbol{\beta} = \mathbf{AD}^{-1}\boldsymbol{\theta}$$

Conceptual, technical questions:

- *n* parameters θ - sample size dependent?
- prior on θ - design-dependency
- non-unique reverse map to β

e.g., predict new y_{n+1} at new design point …

… which prior to use?

Theoretical Foundation: Latent Factor Models

Sample *i* - Column *i* of **X**

$$\mathbf{x}_i = \mathbf{B}\boldsymbol{\lambda}_i + \mathbf{v}_i \qquad \mathbf{v}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$$

Vector of *k<<p* latent factors on sample *i*

Ideosyncratic variation on sample *i*

Independent latent factors:

$$\boldsymbol{\lambda}_i \sim N(\mathbf{0}, \mathbf{I}) \qquad \boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n]$$

Ψ *diagonal*

B *constrained*

Identifiability

Limiting case: **Ψ≈0**, **B** orthogonal …
SVD: **B≈AD**, **Λ≈F**

Latent Factor Regression Models
- & SVD Regression Limiting Case -

$$\mathbf{x}_i = \mathbf{B}\boldsymbol{\lambda}_i + \mathbf{v}_i$$

$$y_i \sim N(\boldsymbol{\lambda}_i'\boldsymbol{\theta}, 1/\phi)$$

Latent/common structure in gene expression patterns …

… is predictive of outcome, phenotype

Everything is normal, so: $E(y|x) = x'\beta$

$$\beta = M\theta$$

Coherent model:

prior on θ transfers uniquely to β

Limiting case: SVD …

$$\beta \approx AD^{-1}\theta$$

Latent Factor Models

- many open research questions -

Model fitting:

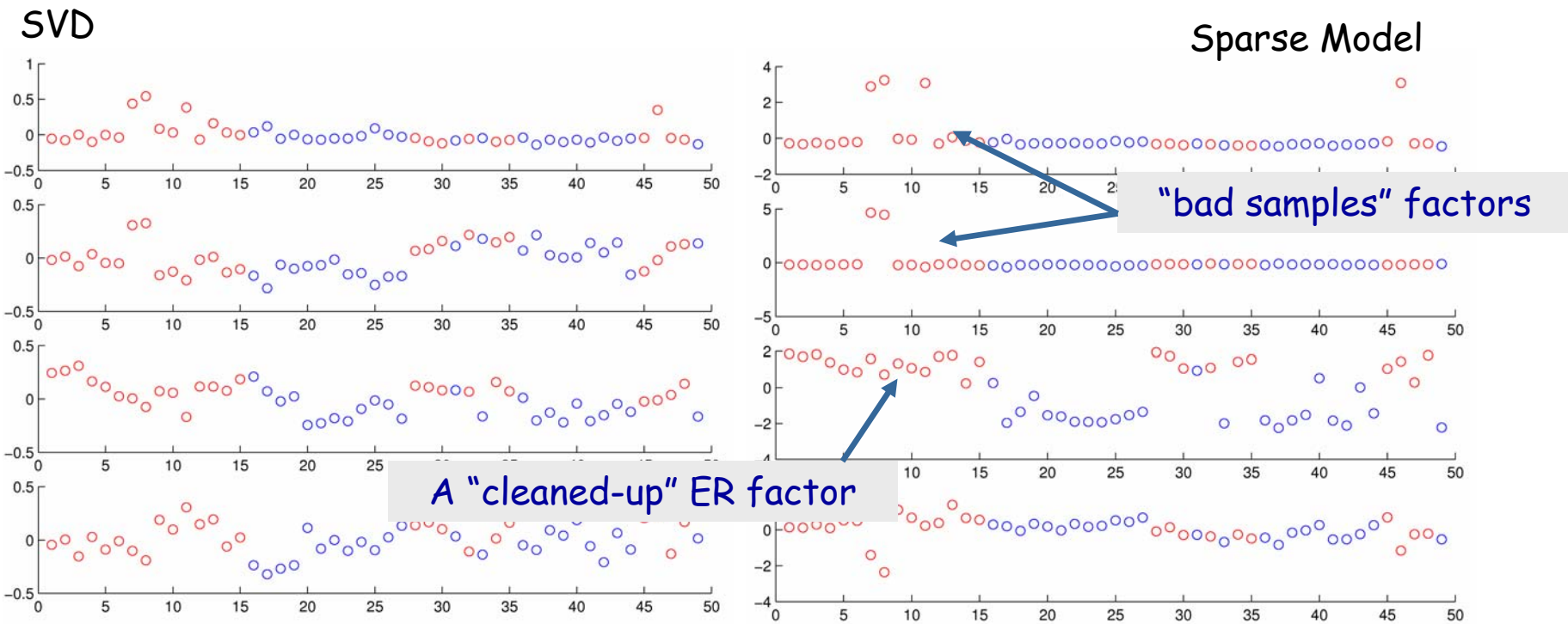
- Parametrisation of **B**
- Identification
- MCMC

Sparse Models:

- **B sparse**
  - One gene – one or a few "pathways"
  - One "pathway" – few or many genes, but not all
- Mixture priors on B_{ij} …

$$B_{ij} \sim pI(B_{ij} = 0) + (1-p)N(0, \omega)$$

Factors in Breast Cancer Data
SVD ~ Sparse latent factor model



Factor models "clean up" SVD

SVD "noisy" factors

West 2003, Valencia 7

Next up:

Binary Regressions & Molecular Phenotyping