Bayesian Statistics
VALENCIA 8

Modelling & Computation

Elements of Shrinkage Modelling in Regression and Multivariate Analysis

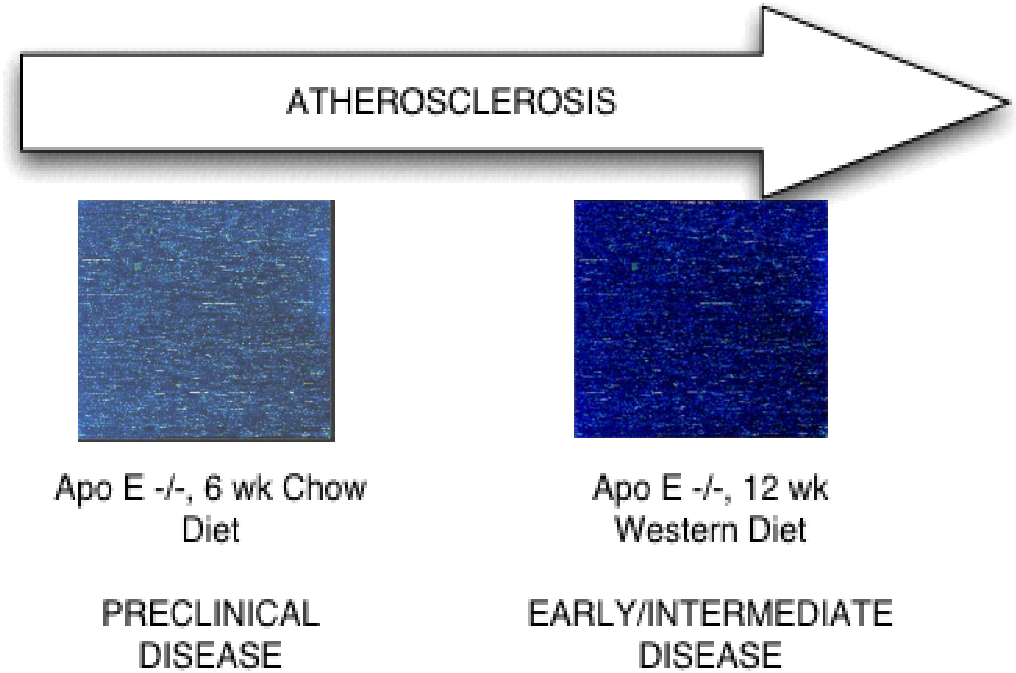
ISBA 2006

Bayesian Statistics
VALENCIA 8

ISBA 2006

Regression:
Shrinkage Prior Modelling

Example: Anova in Cardiovascular Genomics



Age, Diet, Gender, WT/ApoE

2^4x5+ balanced factorial design

Mice model aorta gene expression:
4 factors, each at 2 levels

Response: gene expression

Genes linked to design factors?

Action is interactions

One gene, one sample (p=12,500 genes in parallel)

$z = \beta$	WT, 6wk, chow, fem (baseline)
$+ \mu$	male
$+ \delta$	fat diet
$+ \alpha$	age=12wk/old
$+ \gamma$	ApoE genotype
$+ \mu\delta$	fat diet & male
$+ \mu\alpha$	12wk/old & male
$+ \mu\gamma$	ApoE & male
$+ \delta\alpha, \delta\gamma, \alpha\gamma$	
$+ \mu\delta\alpha, \mu\delta\gamma, \mu\alpha\gamma, \delta\alpha\gamma, \mu\delta\alpha\gamma$	
+ noise	

$$z_i = h_i'\beta + \epsilon_i, \qquad \epsilon_i \sim N(0, v)$$

n-vector response *z*

nxp known design matrix *H*

$$z = H\beta + \epsilon, \qquad \epsilon \sim N(0, vI)$$

(minimal) Bayes: Shrinkage priors

Relevance of zero-mean location

LSE/MLE/Reference posterior:

$$\hat{\beta} = (H'H)^{-1}H'z$$

Prior: $\beta \sim N(0, B^{-1})$

Posterior: $\beta|z \sim N(b, vB_*^{-1})$

Shrinkage:

$$b = B_*^{-1}H'z$$

$$B_* = vB + H'H$$

LSE as limiting case
- no shrinkage -

Degrees and Dimensions of Shrinkage

$$\beta \sim N(0, B^{-1})$$

$$B^{-1} = \tau I$$

$$b = (aI + H'H)^{-1}H'z$$

$$a = v/\tau$$

- act against over-fitting
- improves estimation stability
- robustness in prediction
- **key with many predictors**

Decision theory (other V8 tutorials)

Regularisation … of *H'H*
Collinearity
Numerical instabilities in inversion
Large LSE variances

Role of scale factors

Ridge regression

$$B^{-1} = \text{diag}(\tau_1, \ldots, \tau_p)$$

$$\beta' = (\beta_1, \ldots, \beta_p)$$

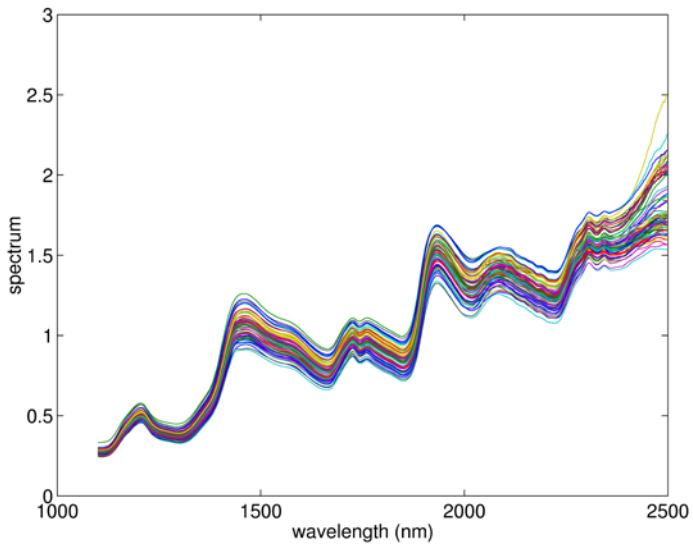
$$\beta_j \sim N(0, \tau_j)$$

Multiple shrinkage

Aim to "Shrinks out" irrelevant covariates

Other shrinkage structures:
Blocked parameters, hierarchical, time series
Substantive prior information: Non-zero prior means, prior correlations, etc

Example: Complex Patterns of Collinearity



(n=78)

Biscuit (cookie) dough spectra

(Brown, Fearn & Vannucci, 1999, Biometrika

West 2003, V7)

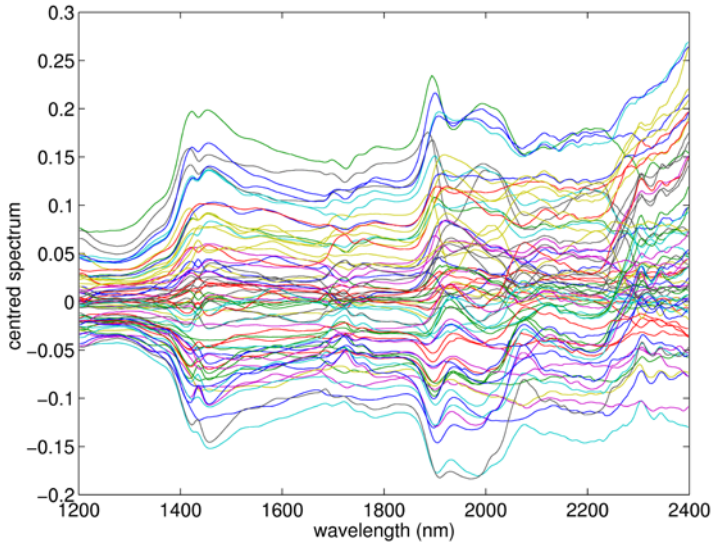
H : spectroscopic reflectance measures at 00s-000s wavelengths

z : cookie fat content

Predictors: finely discretised curves

Smoothness priors - stochastic constraints

Shrinkage priors should respect design (functional) structure …



PCA/SVD - Empirical Factor Regression

SVD: $H = \tilde{H} D E$

PCA: $H'H = E' D^2 E$

$$D = \text{diag}(d_1, \ldots, d_k),$$
$$d_1 > \cdots > d_k > 0, \quad k \leq \min(p, n)$$

Singular values in design space

Small tail values ~ collinearities

$$\tilde{\beta} = DE\beta$$

Orthogonal regression: $\tilde{H}'\tilde{H} = I$

$$z = H\beta + \epsilon$$

Regression on predictors H

$$z = \tilde{H}\tilde{\beta} + \epsilon$$

Regression on factors

Factors "underlying" structure in H are predictors

p=n or *p>n* ... *k=n*

Proper priors: Shrinkage priors key

Orthogonal regression: $\tilde{H}'\tilde{H} = I$

Concordance of independent shrinkage priors in factor regression

$$\tilde{\beta} \sim N(0, T)$$
$$T = \text{diag}(\tau_1, \ldots, \tau_k)$$

$$\begin{cases} \tilde{\beta} = DE\beta \\ \beta \leftarrow E'D^{-1}\tilde{\beta} \end{cases}$$

$$\Longrightarrow \begin{cases} \beta \sim N(0, B^{-1}) \\ B = E'D^{-1}TD^{-1}E \end{cases}$$

Bayesian coherence:
design, *n* dependence, prediction?

Single shrinkage in factor regression:

$$\tau_j = 1/g, \forall j$$
$$\beta \sim N(0, (H'H)^{-1}/g)$$

(Zellner) g-prior:
~ g/n "prior data" on same design … z=0

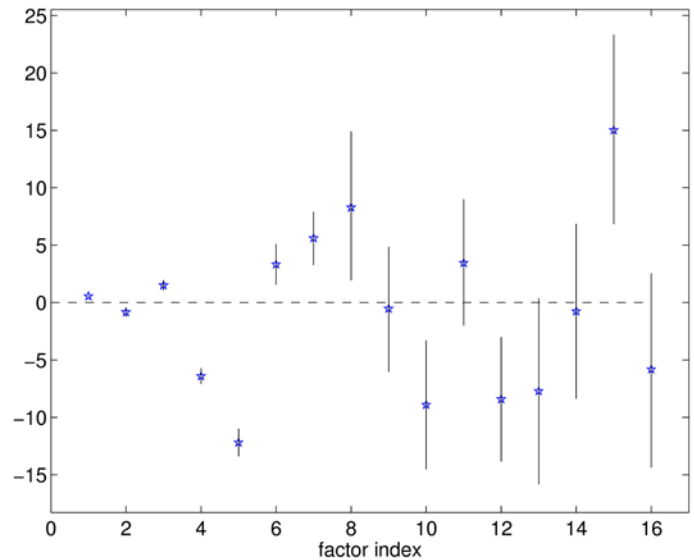
Multiple shrinkage:

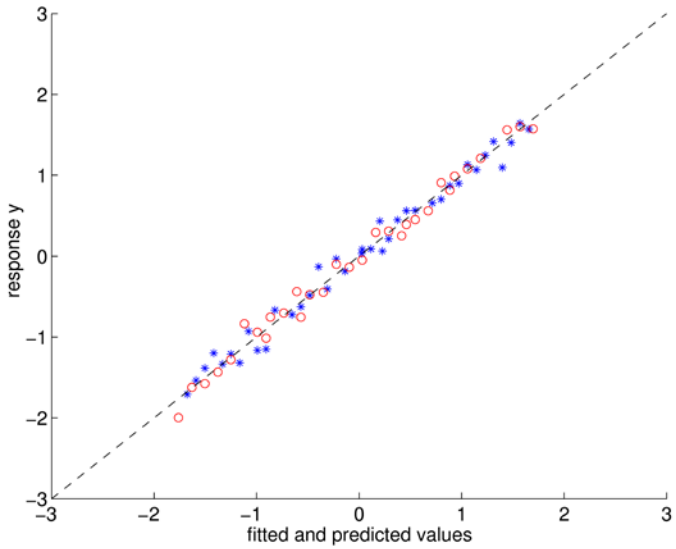
Generalised g-priors (West 2003):
Different "weights" in PCA/SV axes

Allows differing degrees of shrinkage
…. when and where it matters

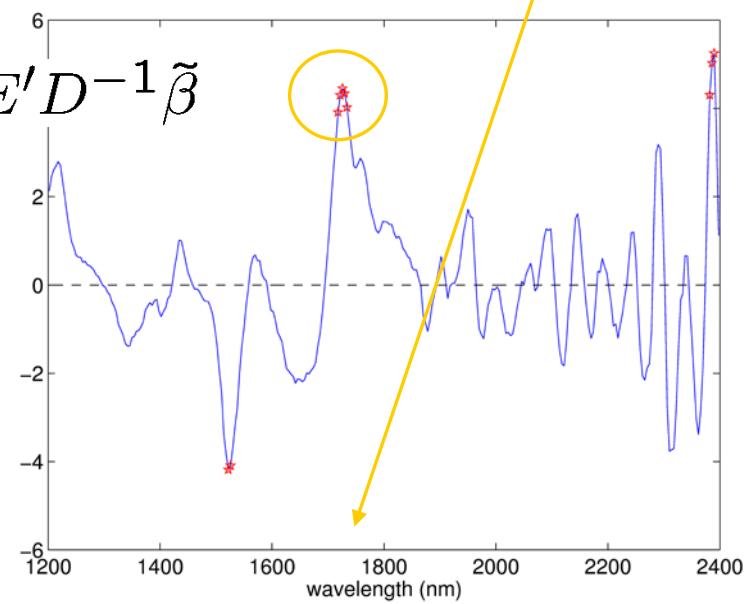
Example: SVD Regression for Cookies

$\tilde{\beta}$





$$\beta \leftarrow E' D^{-1} \tilde{\beta}$$



Spectral "wiggles" – wavelengths for fat absorbance

N=78: 39 training data

39 test data to predict

Aspects of Computation in Regression with Shrinkage Priors

e.g.: $\beta \sim N(0, T)$

$$T = \text{diag}(\tau_1, \ldots, \tau_k)$$

Priors on shrinkage parameters:

$$\tau_j \sim rs/\chi^2_r$$

IG has conditional conjugacy

s - prior estimate *r* - tail weight

Hyper-parameter specification:

Scales of predictor variables
- Standardised
- Consideration of ranges of variation
- May include weights: s/k_j in place of s

Marginal priors:
Student T on r d.o.f

Fatter tailed than *N(0,s)*

Kurtosis: Shrinkage around *0*
but sends mass out into tails

Other choices:
- (r_j, s_j), dependent, non-IG, …
- Exp: marginal priors Laplace (double Exp)
- More useful/relevant shrinkage forms below

e.g.: $\beta \sim N(0, T)$

$T = \text{diag}(\tau_1, \ldots, \tau_k)$

Priors on shrinkage parameters:

$$\tau_j \sim rs/\chi^2_r$$

Posterior computations:

- posterior modes (1970/80s)
- simulations (90s – current)

Joint posterior: $p(\beta, T|z)$

Exploit "complete" conditional posteriors

$$p(\beta|z, T)$$

$$p(T|z, \beta)$$

- EM (expectation/maximisation)
  - shrinkage parameters in *T*: "missing data"
- ICM (iterative conditional modes)
- MCMC (Markov Chain Monte Carlo)
  - canonical example of Gibbs sampling

Simulate Posterior:

Estimation/inference uses sample means, histograms - Monte Carlo approximation of posterior

Iteratively resample complete conditionals of joint posterior $p(\beta, T|z)$

$$p(\beta|z,T)$$

$$p(T|z,\beta)$$

$$\{\beta^{(i)}, T^{(i)} : i = 1, 2, \cdots\}$$

Irreducible (aperiodic) Markov chain on full (β, T) space

Limiting distribution : joint posterior

Posterior MC samples (dependent)

Computation: MCMC in Shrinkage Models

$$z|\beta \sim N(H\beta, vI)$$

e.g.: $\beta \sim N(0, T)$

$$T = \text{diag}(\tau_1, \ldots, \tau_k)$$

Complete conditionals often exploit:

- conditional conjugacy (analytic/easy to simulate)
- conditional independencies (parallel components)

$$p(\beta|z, T) = N(b, vB_*^{-1})$$

$$p(T|z, \beta) = \prod_{j=1}^{k} p(\tau_j|\beta_j)$$

$$\tau_j = (r+1)s_j/\chi^2_{r+1}, \;\; s_j = s + \beta_j^2$$

Parameter "blocking" is good
Parameter "decoupling" is good

... for MCMC convergence

Modular nature of many posterior MCMC implementations

e.g. response error variance

$$p(\beta|z,T,v) = N(b, vB_*^{-1})$$

$$p(T|z,\beta,v) = \prod_{j=1}^{k} p(\tau_j|\beta_j)$$

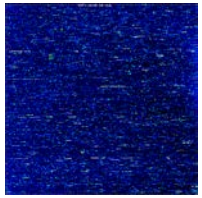
$$p(v|z,\beta,T)$$

e.g., inverse scaled chi-square for *v*

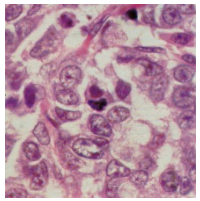
Add-on ("bolt on") components for hyperparameters

Some Responses are Binary

Example: Binary Responses



x



y

$$p(y|x)$$

Gene expression as covariates (predictors)
Molecular phenotyping e.g.:
cancer outcomes

- Predict aggressive vs. benign
- Disease susceptible vs. resistance
- Drug/treatment response

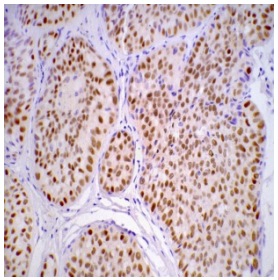
e.g.: Breast Cancer
ER – (O)Estrogen Receptor Status
Lymph node (recurrence risk) status

Clinical test – gene expression in tumour

y=0/1 (ER -/+)
(Crude) Protein assay

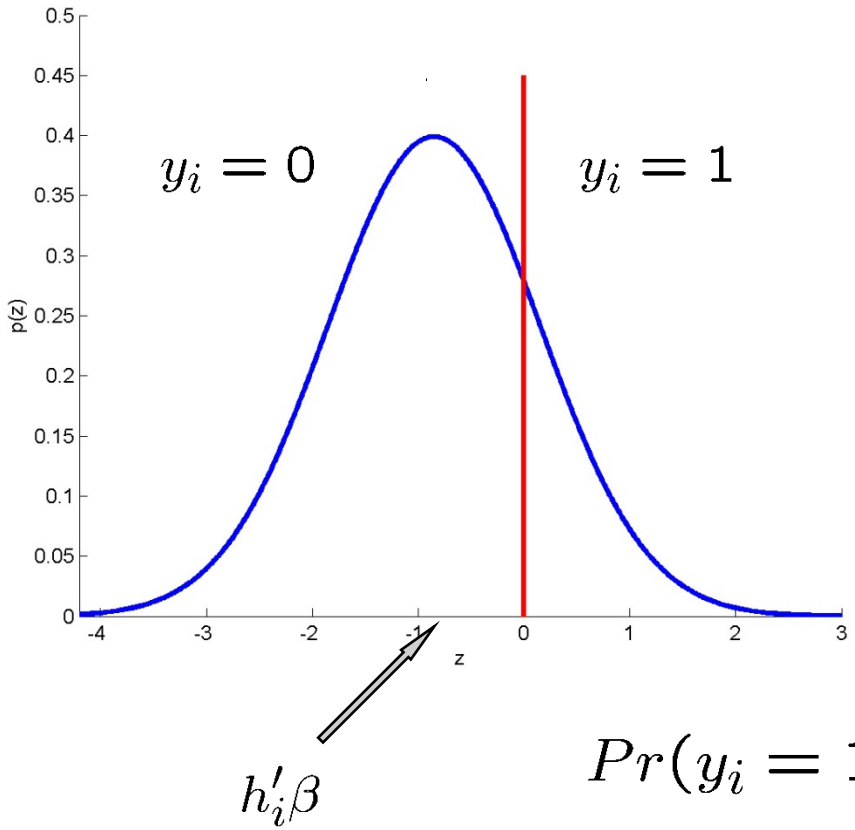
(~60x magnification)
nuclei of breast epithelial cells
cytoplasm of breast epithelial cells

brown-red & pink ~ ER+



Binary = thresholded latent continuous

probit~normal, logit~logistic, …



$$Pr(y_i = 1|\beta) = \Phi(h_i'\beta)$$

$$h_i = h_i(x_i)$$

Natural model/intepretation

Computationally nice

$$Pr(y_i = 1) = Pr(z_i > 0), \quad z_i \sim N(h_i'\beta, 1)$$

$$z = H\beta + \epsilon, \quad \epsilon \sim N(0, I)$$

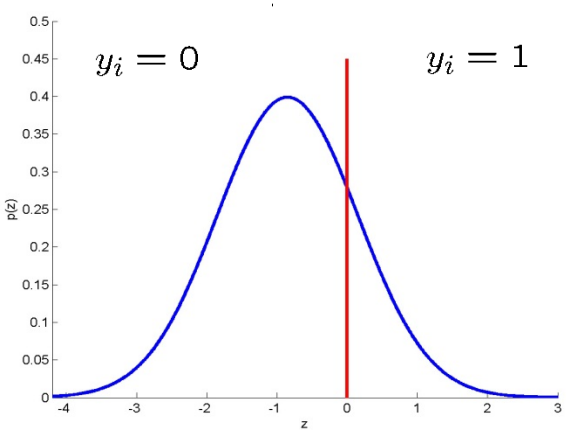
Computation: MCMC in Binary Regression

Linear regression if z known
(simpler: v=1)

$$p(\beta|z,T) = N(b, B_*^{-1})$$

$$p(T|z,\beta) = \prod_{j=1}^{k} p(\tau_j|\beta_j)$$

$$p(z|y,\beta) = \prod_{i=1}^{n} p(z_i|y_i,\beta)$$



Add module to impute latent z
MC samples for z

Prediction and {Variable, Feature} Selection

(PNAS 2001 breast cancer)

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

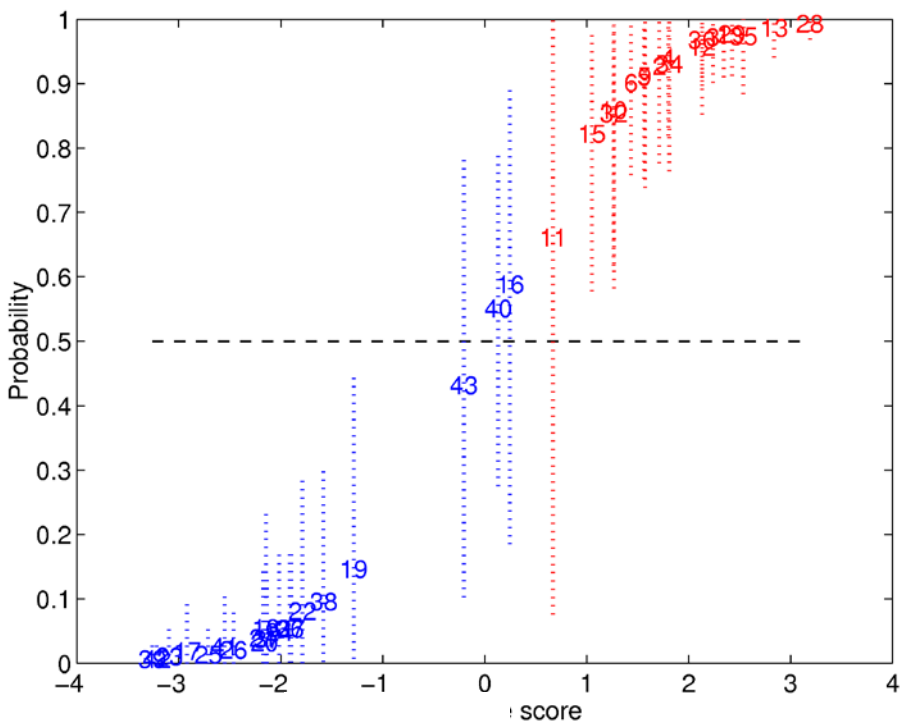
"Honest" assessment of precision

Heterogeneity, small samples

Feature/Variable selection

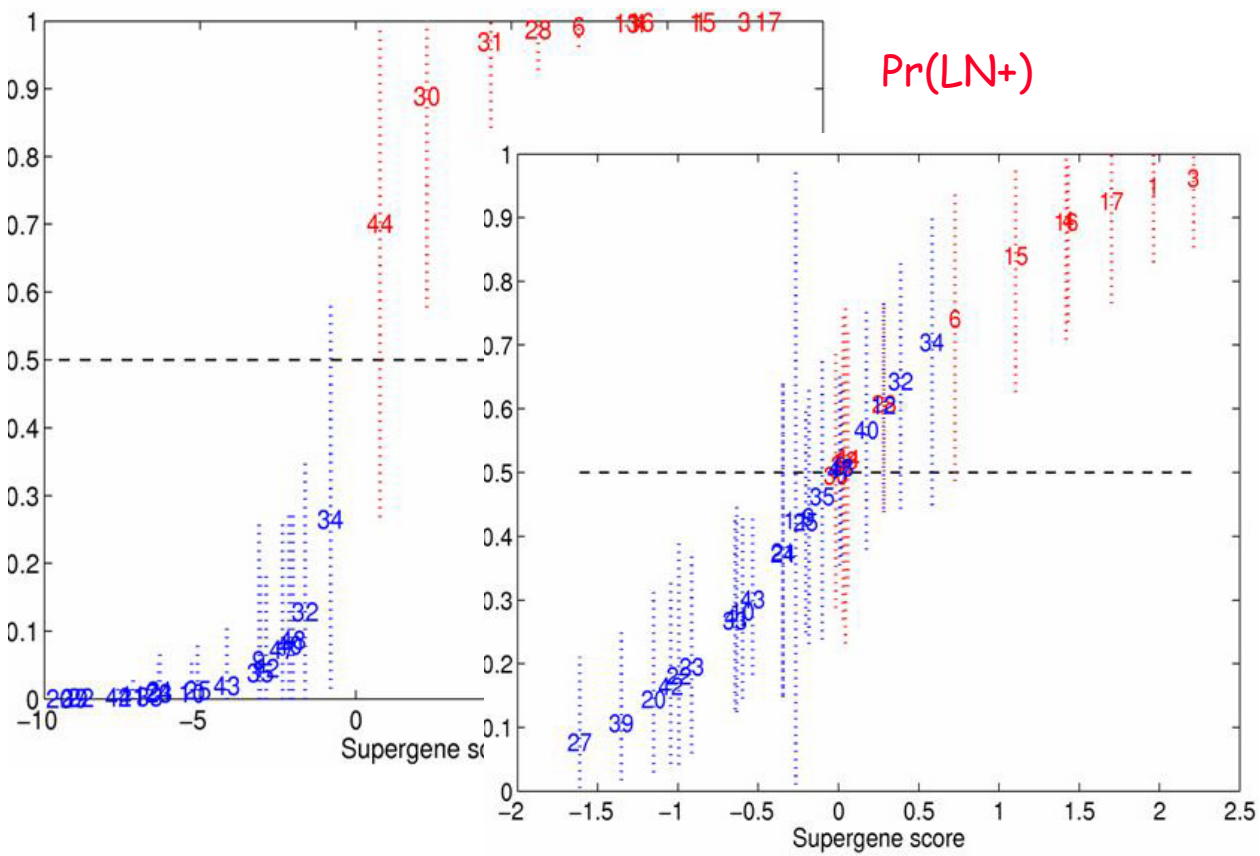
Critical component of predictive assessment with large *p*

Pr(ER+)



Predicting lymph node status:

Pre-selection of 100 genes
vs.
"Honest" CV predictions



Variable selection, Uncertainty

Complex interdependencies

Multiplicities

Shrinking Variables "Out" - Sparsity Modelling in Regression

Sparsity priors: Taking Shrinkage to the Limit

Regressors "out": $\beta_j = 0$

In/out indicators: $\gamma_j = \begin{cases} 0 & \leftrightarrow \quad \beta_j = 0 \\ 1 & \leftrightarrow \quad \beta_j \sim N(0, \tau_j) \end{cases}$

Independent Bernoulli: $Pr(\gamma_j = 1) = \pi$

$$(\beta_j | \tau_j, \pi) \sim (1-\pi)\delta_0 + \pi N(0, \tau_j), \qquad \tau_j = rs/\chi^2_r$$

Selection/sparsity shrinkage priors:

$(\beta_j | \tau_j) \sim N(0, \tau_j)$

$p(\tau_j): \qquad \tau_j = \begin{cases} 0 & \text{with probability } \pi \\ rs/\chi^2_r & \text{with probability } 1-\pi \end{cases}$

Other versions:

- constant/specified $\tau_j = s$
- non-zero but "very small" τ_j

$$Pr(\gamma_j = 1|z) = Pr(\beta_j \neq 0|z)$$
$$Pr(\gamma_j = 1, \gamma_h = 1|z)$$

Regression variable "selection"

Model uncertainty

-Berger & Bayarri V8 Tutorial -

$$\gamma = (\gamma_1, \ldots, \gamma_p)'$$

One model: M_γ $\quad Pr(M_\gamma|z)$

Subsets of regressors: Models

Simultaneous "multiple tests"

$$p(z^*|z) = \sum_\gamma p(z^*|z, M_\gamma) Pr(M_\gamma|z)$$

Model averaging for prediction

Among the issues:
- collinearities – structure among covariates
- in/out dependencies, masking effects

Computation:
Finding"interesting" models

MCMC and stochastic search

Gibbs MCMC: repeat scans through all variables - in/out?

$$p(\pi|\gamma)$$

$$p(\tau|z,\beta)$$

$$j = 1, 2, \ldots : \; p(\beta_j, \gamma_j | z, -)$$

$$= p(\beta_j | z, \gamma_j, -) p(\gamma_j | z, -)$$

$$\frac{Pr(\gamma_j = 1 | z, -)}{Pr(\gamma_j = 0 | z, -)} = \frac{\pi}{1-\pi} L_j$$

$$\begin{cases} (\beta_j | z, \gamma_j = 0, -) \rightarrow \beta_j = 0 \\ (\beta_j | z, \gamma_j = 1, -) \sim N(*, *) \end{cases}$$

Variable in/out depends on variables in/out
"Nice" in orthogonal regression

But: collinearities … masking?

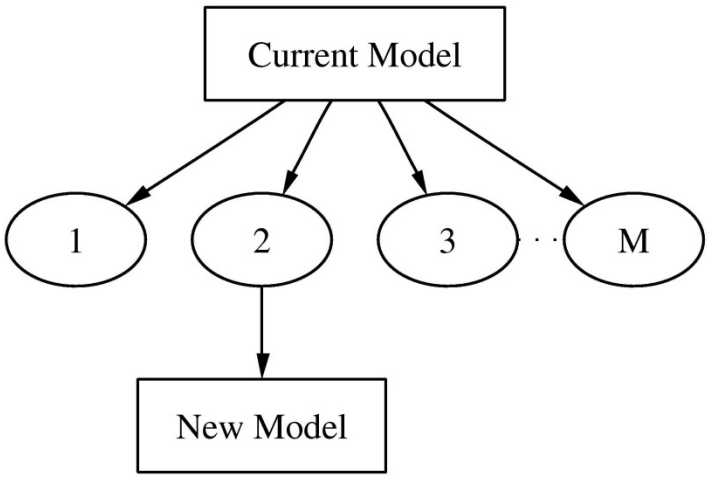
Large *p*?

LOCAL search -

"Current" model :

Add/delete sets of variables around "local" model

Large *p* - many models



MCMC "local search" inspired
Local conditional posterior proposals

Good models "near" good models

Add/drop/replace variables
... with trans-dimensional balance

Move by sampling new model

Shoot out **ALL** neighbours:
"local proposals"

Swiftly find high probability <u>regions</u>
of model space

Catalogue of many "good" models

Parallelisation

KEY: easily compute

$$\propto \quad Pr(\mathrm{M}_\gamma | Z)$$

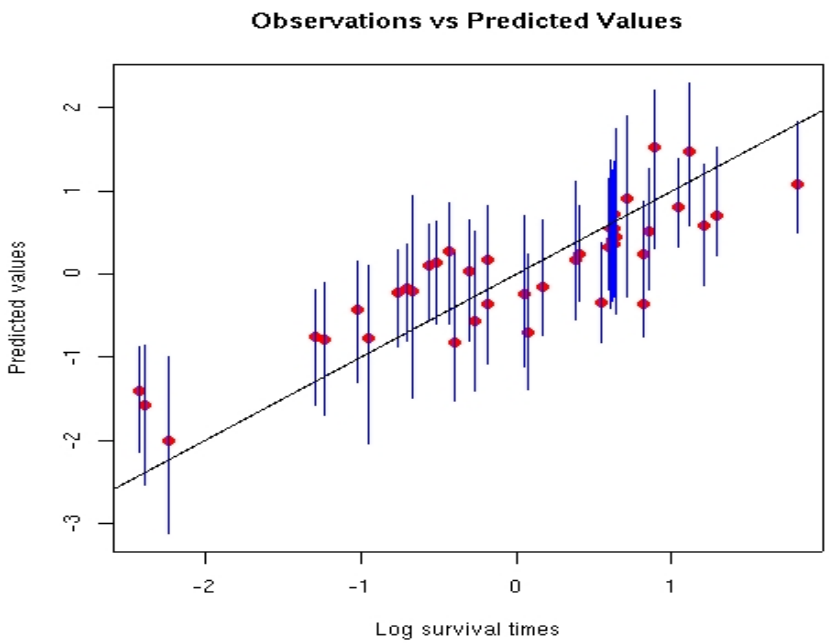
(Hans et al 05 and **V8 Poster**)

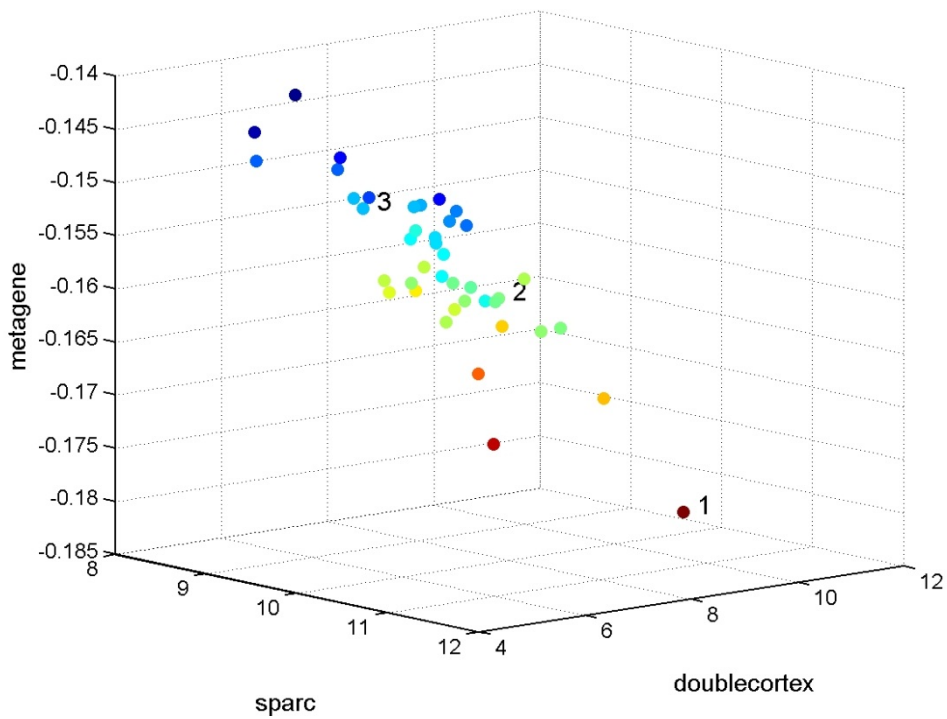
Example: Cancer Genomics – Survival Prediction

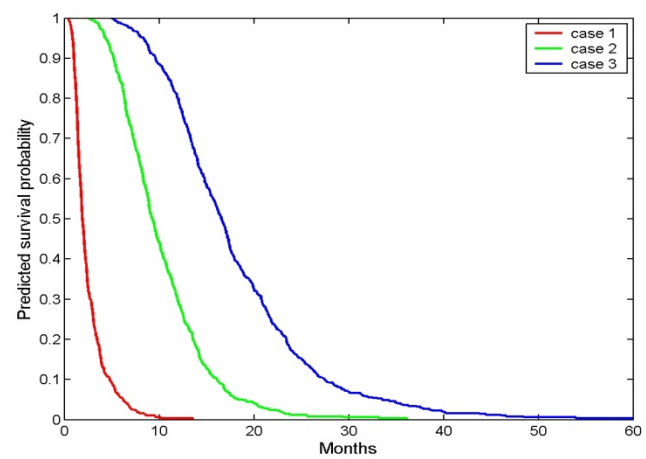
Brain cancer expression: p=8400

Survival regressions:

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







(Cancer Research, 05)

Sparsity -

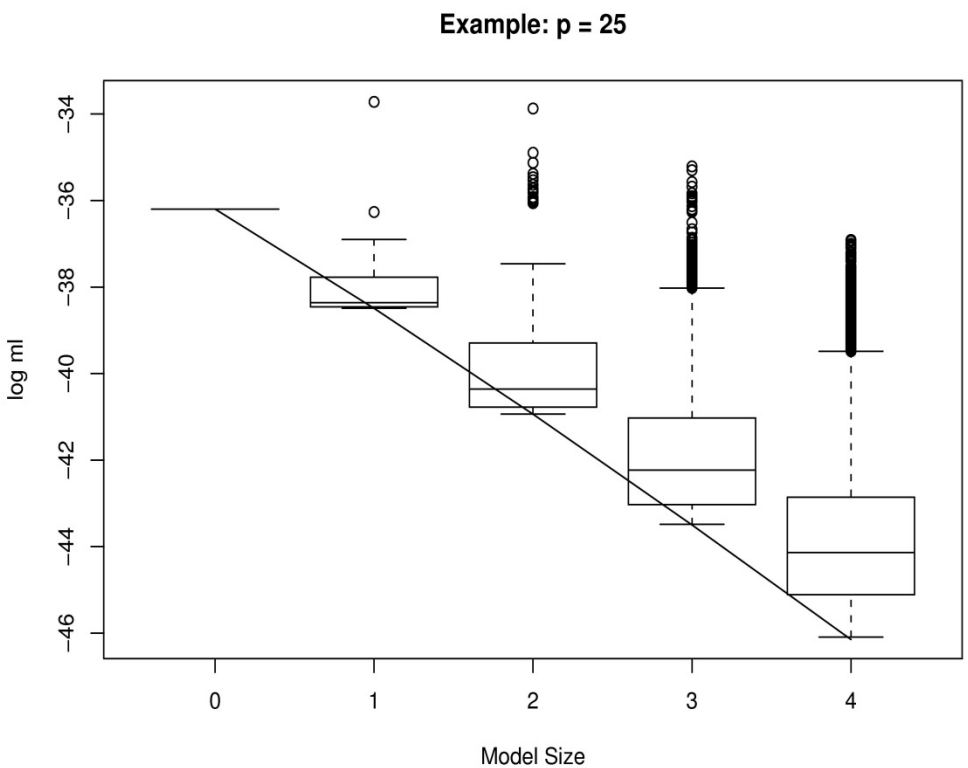
"Base rate" in/out probability π

Dimension -

Implicit in Bayesian & other likelihood-based analyses (*cf.* BIC)

Specification for base rate:

Scale with dimension to maintain parsimony, sparsity



Getting Multivariate

Example: 2^5 Anova Mice Genomics

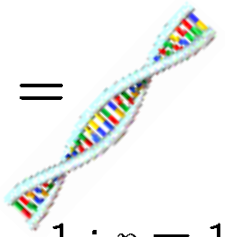
$z = x$ - gene expression

h_i fixed design (0/1)

$k = 16$ parameters

$i =$ 

$1 : n = 90$

$g =$ 

$1 : p = 12,500$

$$x_{g,i} = \beta_g' h_i + \epsilon_{g,i}, \qquad \epsilon_{g,i} \sim N(0, v_g)$$

$$x_i = B h_i + \epsilon_i, \qquad \epsilon_i \sim N(0, V), \qquad V = \text{diag}(v_1, \ldots, v_p)$$

$p \times 1$ $\qquad$ $p \times k$

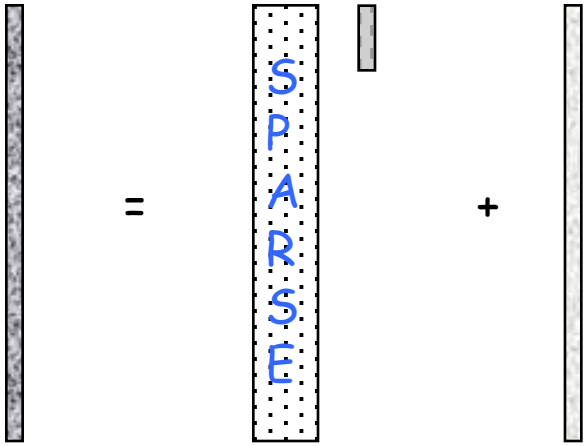
Highly multivariate anova

$$x_i = B\ h_i + \epsilon_i$$

Sample *i* $\quad x_i \quad (p \times 1)$

Variable *g*, factor *j* $\quad B = \{\beta_{g,j}\} \quad (p \times k)$

Fixed design $\quad h_i \quad (k \times 1)$



Design factor *j* : $\beta_{g,j}$ Main effect, interaction, …

Many zeros … Column/factor *j* – which are non-zero?

Full multivariate analysis – simultaneous inference – "multiple tests"

Precursor experiments

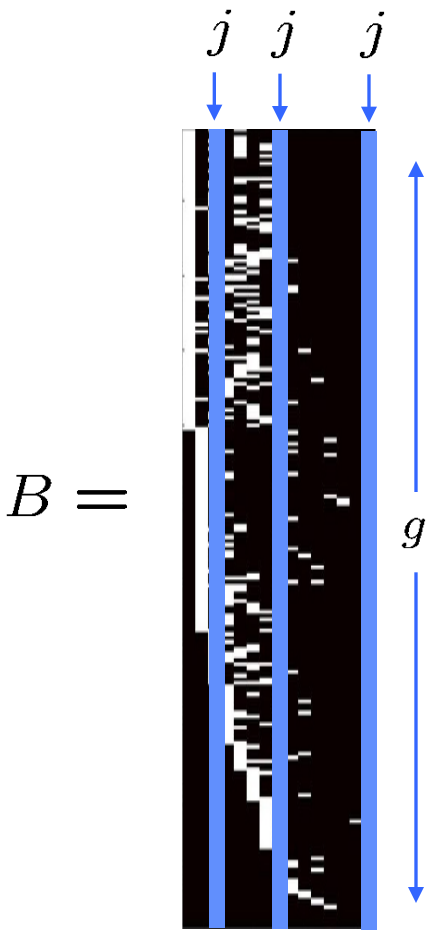
Sparsity priors: $\#\{\beta_{g,j} \neq 0\} =$ small

$$\beta_{g,j} \sim (1-\pi_j)\delta_0 + \pi_j N(0, \tau_j)$$

Variable (gene) *g*, Design factor *j* :

π_j ~ sparsity

Differing sparsity patterns



Computation: MCMC methods -

- Blocking and conditional independence of parameters within factors
- Within-factor parallelisation
- Serial Gibbs/MCMC: within iterate parallelisation

Local model stochastic search?

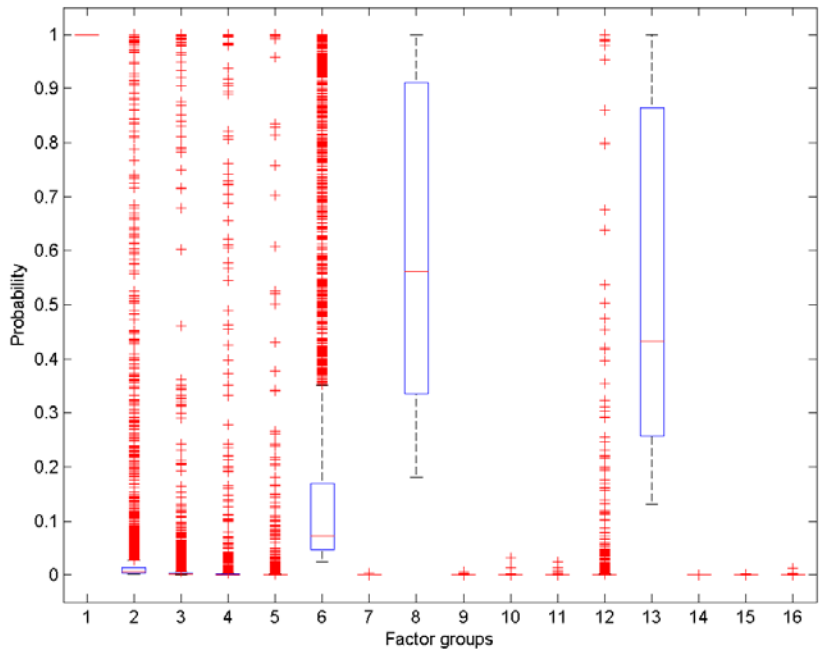
Gene-Environment Associations in Mice Experiments

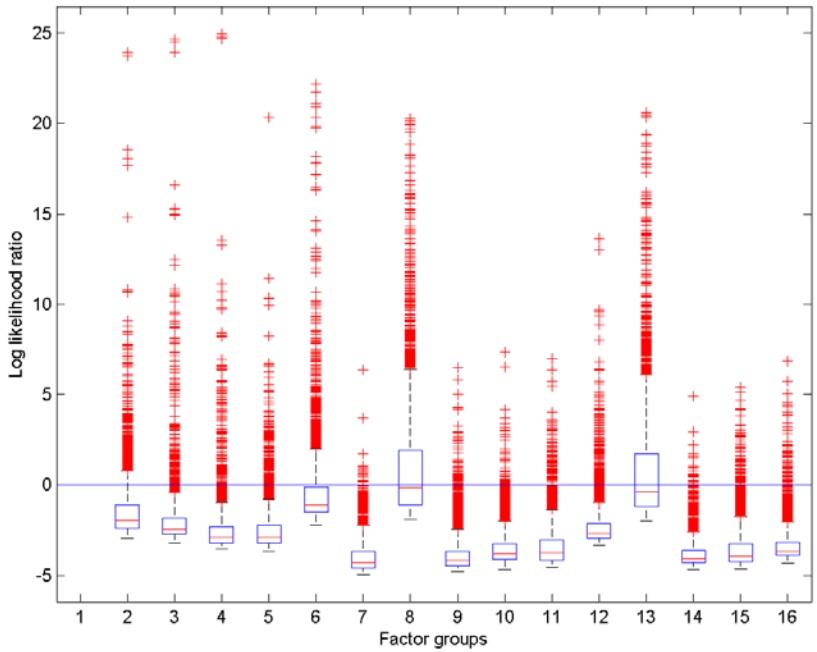
Probabilities and log-likelihood ratios

- SHRINKAGE

Variable (gene) identification within interaction effects

$$\pi^*_{g,j} = Pr(\beta_{g,j} \neq 0|X)$$

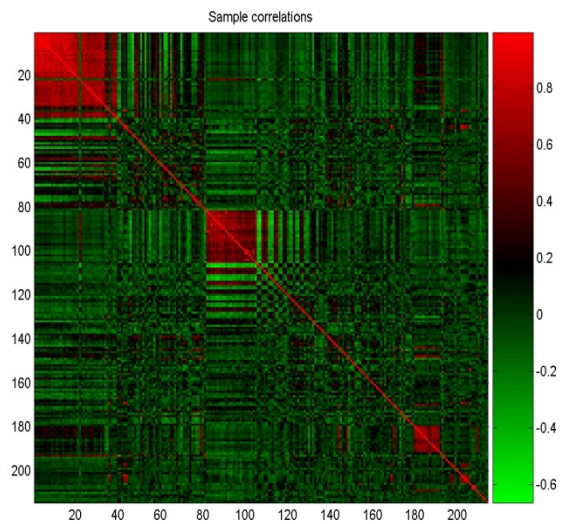




(Karra et al 04; Seo et al 05)

More Structure in Multivariate Data - Residual Correlation in High-Dimensions?

Latent Factor Models & Patterns of Association

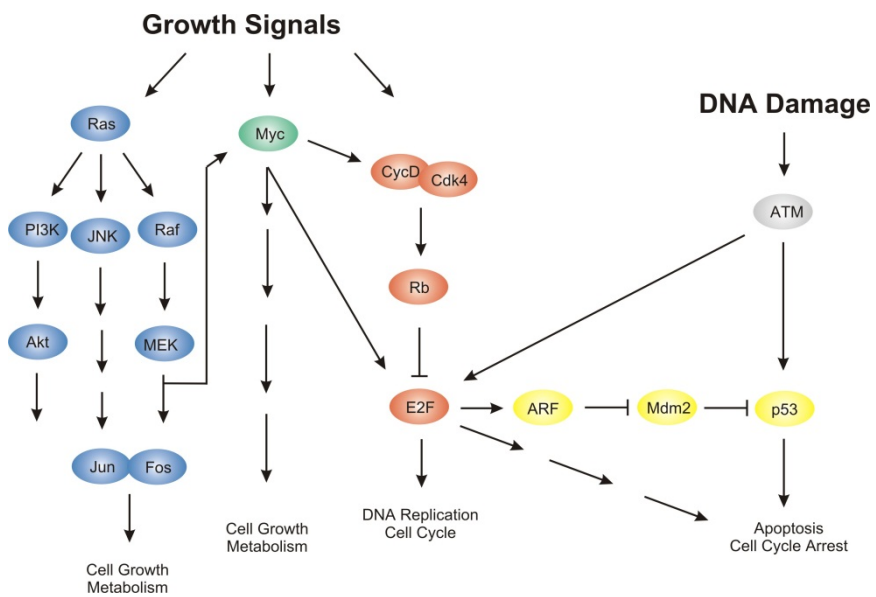


Decompositions of $p(x)$

Latent structure underlying associations

Cancer Studies: n=430 breast cancers

Multiple deregulated pathway components



Latent Factors:

intersecting sub-pathways

Latent Factor Model Structure

One sample
- column p-vector

$$x_i = Bh_i + \epsilon_i \qquad \epsilon_i \sim N(0, V)$$

Vector of k<<p latent
- underlying -
factor variables

Idiosyncratic
variation

Latent factors:

$$\lambda_i \sim N(0, T)$$

Model of covariance matrix:

$$V(x_i) = BTB' + V$$

(West 03 Valencia 7; Lopes + West 04 Stat Sinica)

Sparse Models:

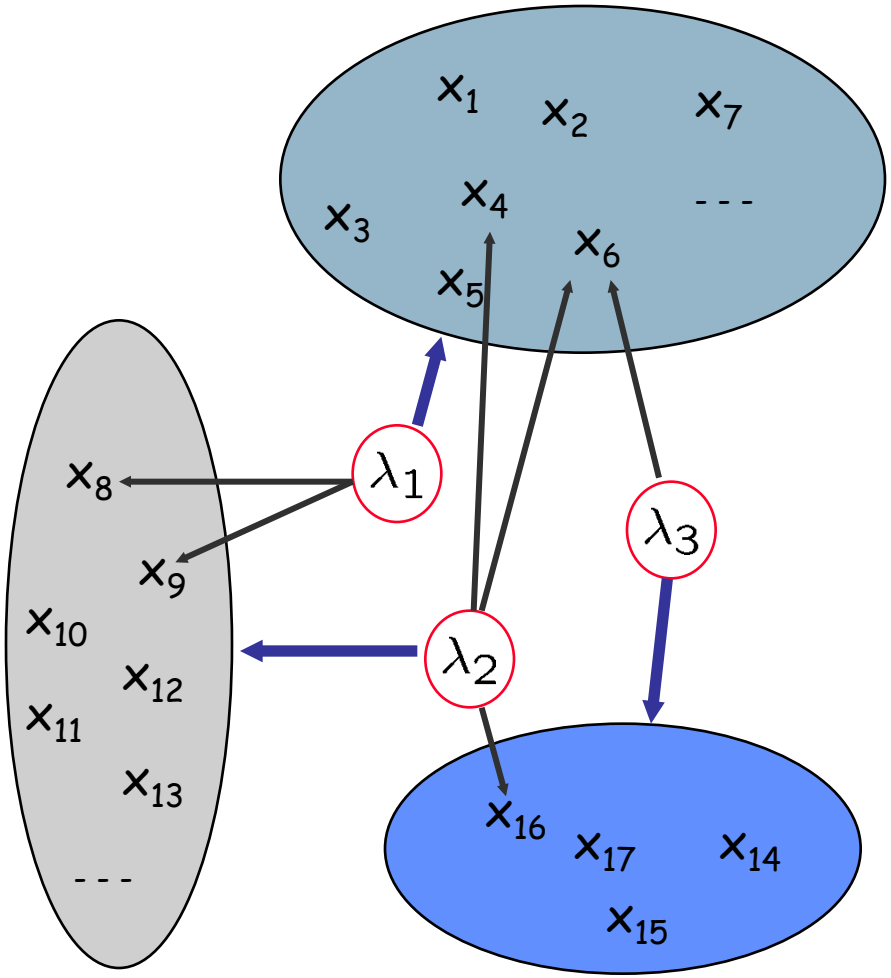
One factor – few or many variables

One variable – 0,1, or few factors

$$B = \{\beta_{g,j}\}$$

Row (variable) *g*, factor *j* :

$$\#\{\beta_{g,j} \neq 0\} = \text{ 0,1, .., small}$$

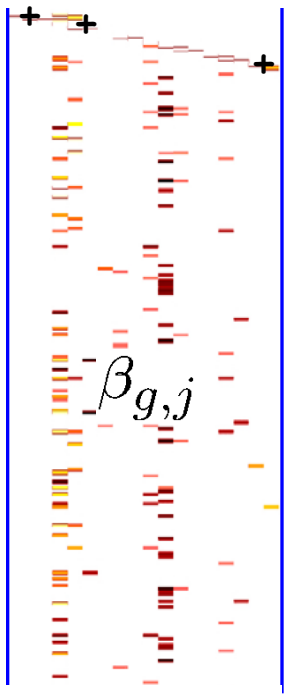


(West 2003, Valencia 7)

Uncertain sparsity patterns in latent factor models:

$$x_i = B\lambda_i + \epsilon_i$$

B =



(West 2003, Valencia 7)

Structure:

Recall multivariate regression:

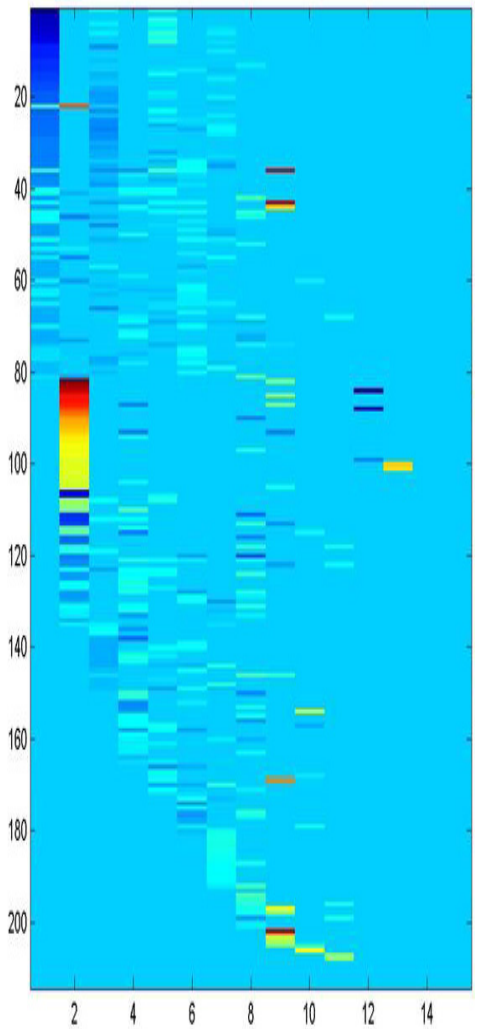
$$x_i = Bh_i + \epsilon_i$$

Same model structure: design vector becomes uncertain

Add on/bolt on module to MCMC computations

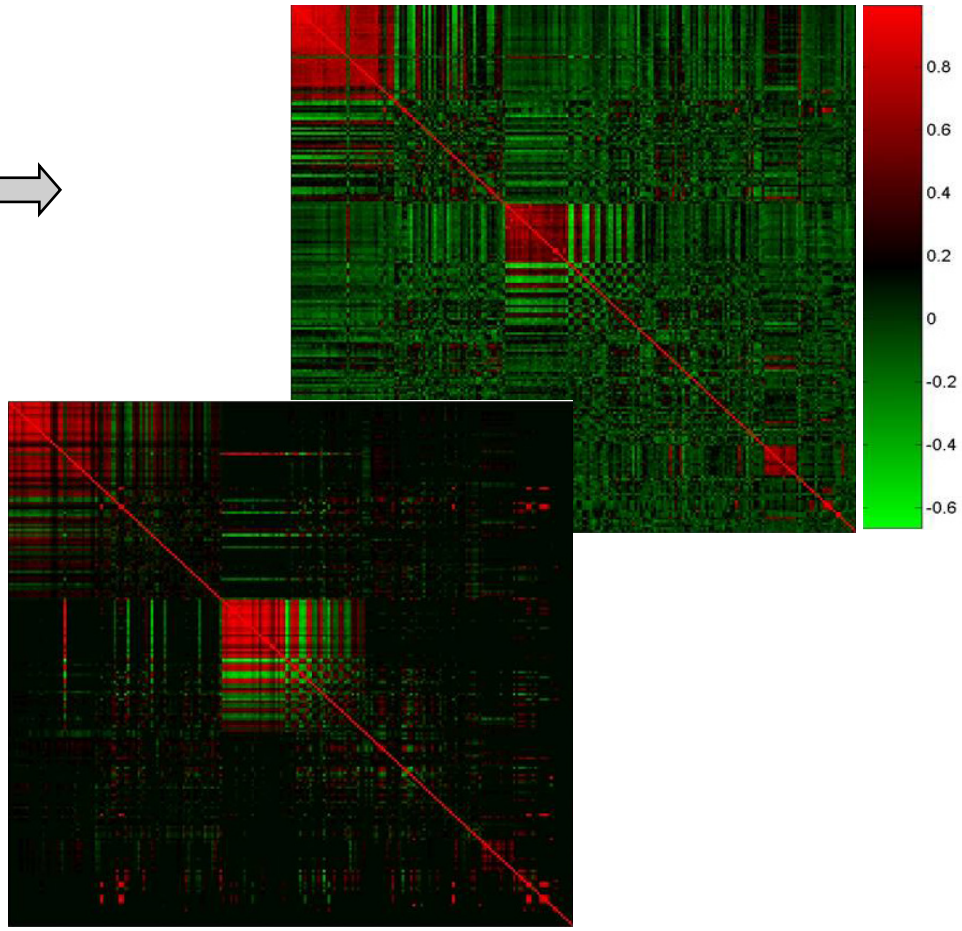
$$(\text{MC})^2 + \prod_{i=1}^{n} p(\lambda_i | X, -)$$

Inference on Sparsity and Structure



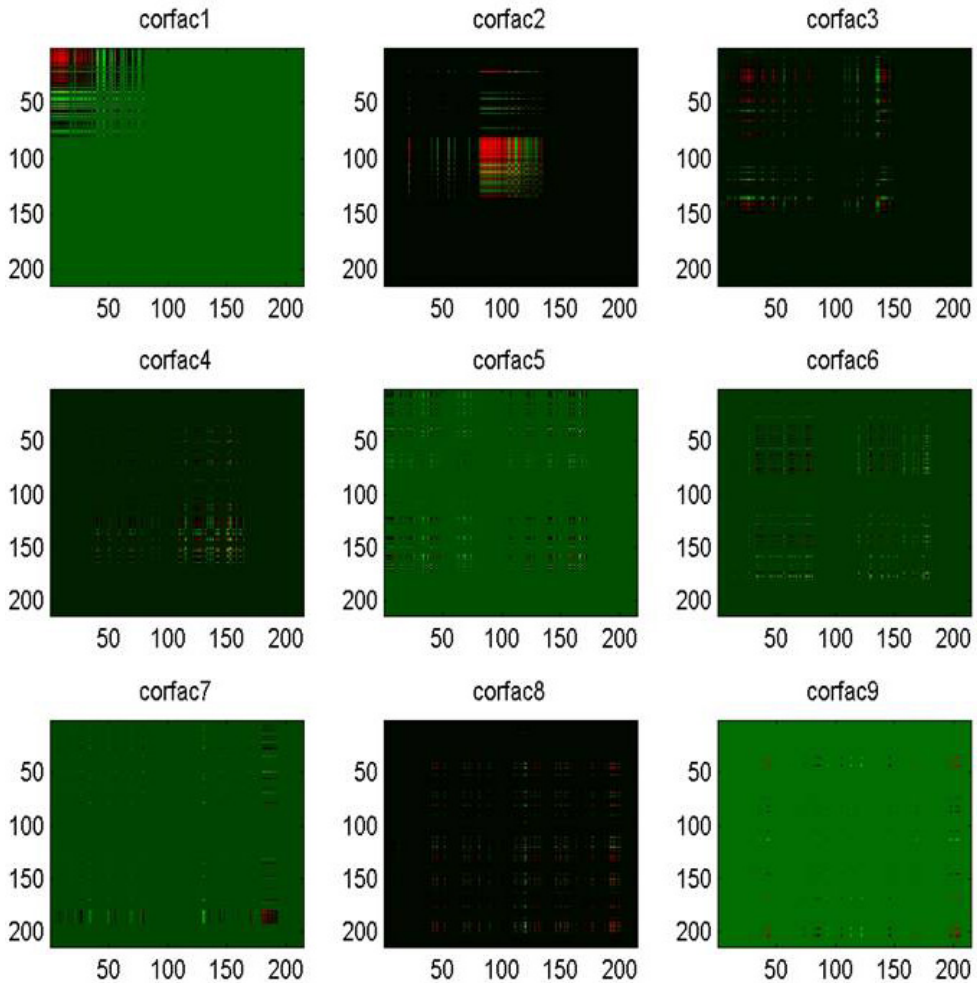
… p(B|X)

Sample correlations



Fitted correlations in B'TB+Ψ

Decomposing Associations by Factor

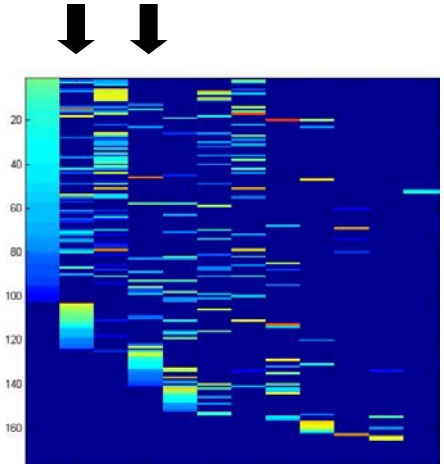


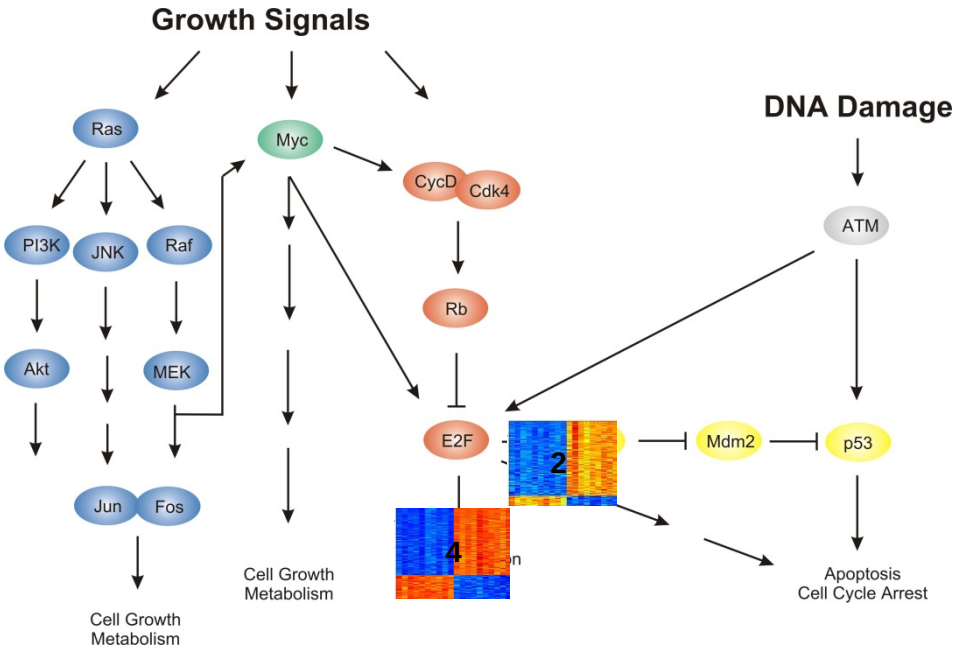
Covariance decompositions:

$$BTB' = \tau_1 b_1 b_1' + \tau_2 b_2 b_2' + \cdots$$

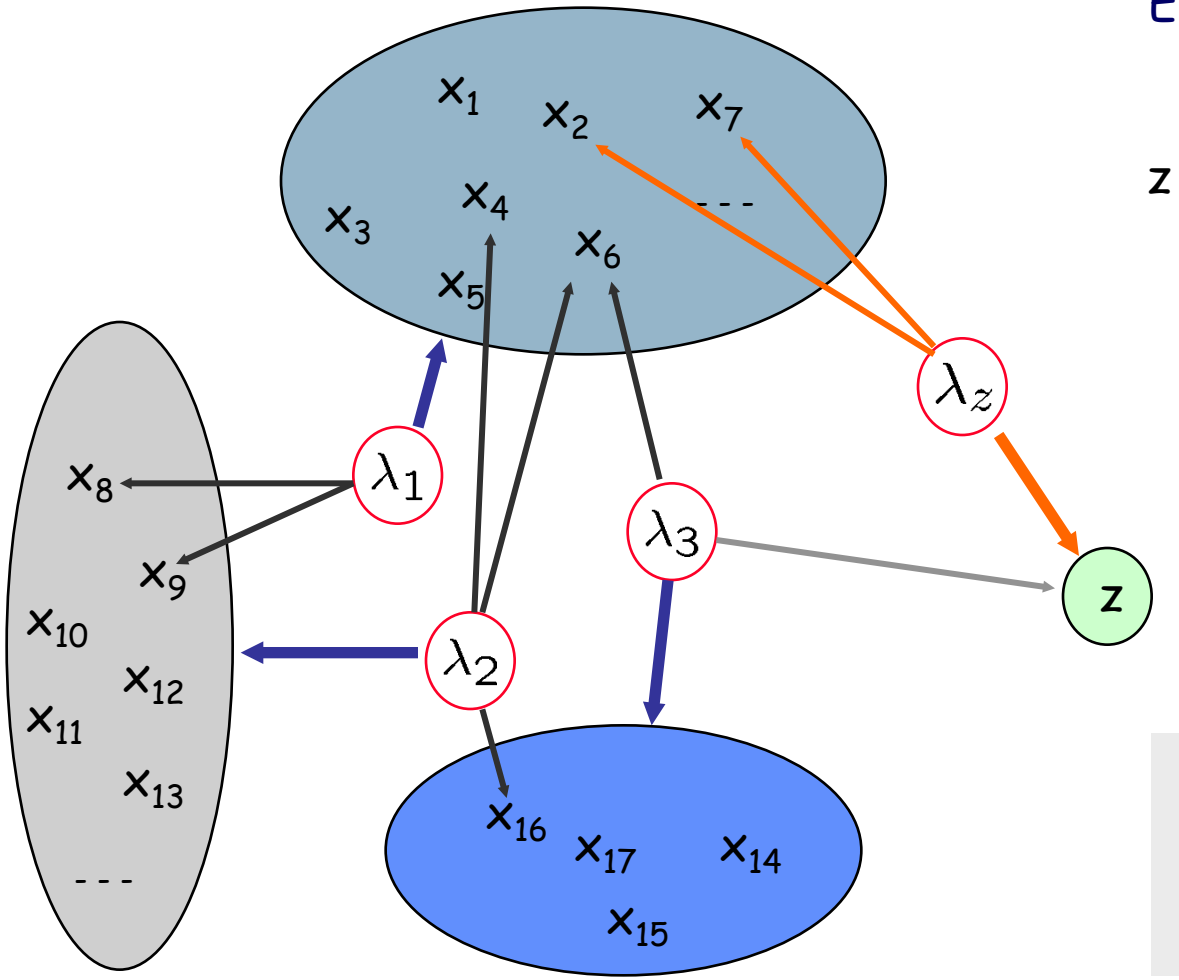
Exploratory/Discovery Variable "Selection" Analyses

2 (apoptosis)	4 (DNA replication)
MDM2	MCM6
GSN	VCAM1
AGC1	CYP2A13
RBM8A	PITX1
MTHFS	MCM2
CUGBP2	DDX39
PKN2	MCM7
SSR4	GSN
FANCG	GSTM1
NME3	CCNE1
POU4F1	MCM3
AGC1	MCM4
MDM2	MFGE8
CRHR1	ABCA3
H1FX	CDKN2A
RPS3A	MCM5
ABCB8	KIAA1026
RGS12	TOMM70A
GLG1	CDC2
DOC-1R	SAS
TNPO3	POLR2H
MDM2	CSNK1D
MAPT	NME3
LOR	CDC6
GUCA1A	CHC1
GRIA1	TNFRSF14
CDC34	BACH
COL11A2	MVD
MYC	FANCG
TBL3	LIG1
BTF3	SF3B4
UCP3	CCNE1
LBA1	ABCF3
CDKN2C	
HTR6	
CDC6	
CYP2A13	
KHDRBS1	
KIAA0284	
PEX5	
CYP2A6	
LTK	
SSTR3	
MDM2	





Sparse Latent Factor Regression



Response variables z

Evaluate p(z,x) - and - predict z

z linked to factors underlying x?

… and to individual x variables?

λ_z = response factors

Full, coherent rationale for SVD/PCA regression – but SPARSE

(West 2003, Valencia 7; Carvalho et al, 2006 & **V8 Poster**)

Shrinkage and sparsity priors:

Regression structure - variable uncertainty, "selection"

(high-dimensional) multivariate structure:

- Parsimony
- Scalability

Computation, Model Search:

MCMC and MCMC-inspired Stochastic/Evolutionary Search

Software: BFRM (C++/Java: Spring 2006 release)

(Carvalho et al 06; Lucas et al 06)

(very selective, starting points …) Links and Readings

Lindley DV (1971)
Estimation of many parameters,
Foundations of Statistical Inference
(Godambe and Sprott, eds)

Lindley DV (1972)
Bayesian Statistics: A Review, SIAM

Dobra A, Jones B, Hans C,
Nevins JR & West M (2004)
Stochastic search, regressions
and graphs, *JMVA* 90

Hans C, Dobra A and West M (2005)
Regression variable selection and
stochastic model search

West M (1985)
Shrinkage, g-/hierarchical priors, scale mixtures
Bayesian Statistics 3

Zellner A (1986) g-priors,
Bayesian Inference and Decision Techniques
(Goel and Zellner, eds)

Carvalho C, Wang Q, West M (2006)
Sparse factor models

Lucas J, Carvalho C, et al (2006)
Sparse statistical modelling,
Bayesian Bioinformatics

West M (2003) Sparse factor models
- large p, factor regression and generalized g-priors
Bayesian Statistics 7

Clyde M (1999)
Model averaging and search
Bayesian Statistics 6

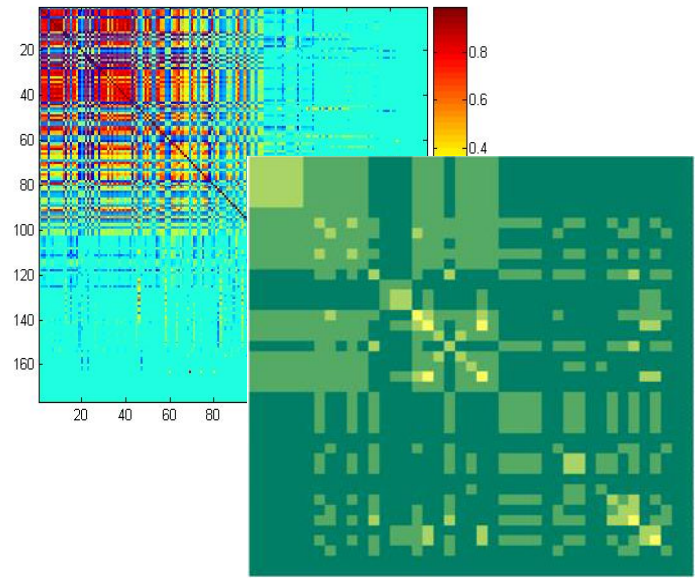
Clyde M & George EI (2004)
Model uncertainty
Statistical Science 19

George EI & McCulloch RE (1993)
Variable selection MCMC
JASA 88

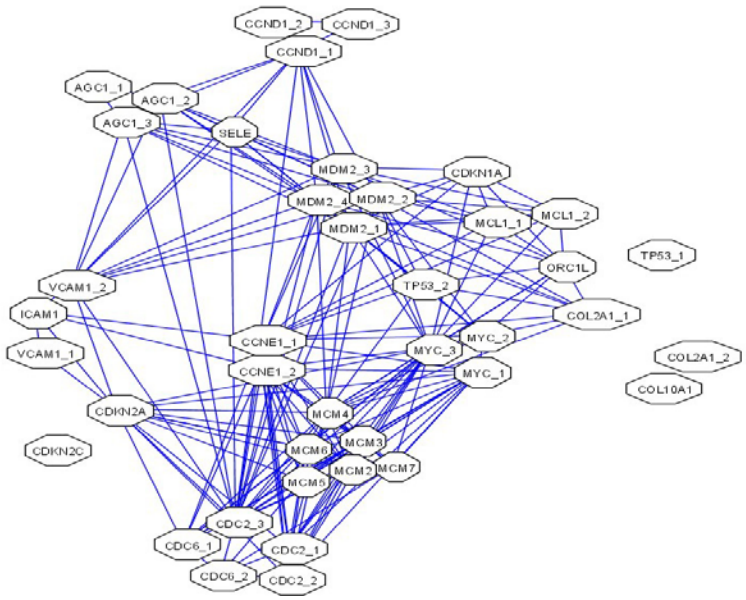
www.stat.duke.edu/~mw

Papers
Teaching (tutorials)
Software
Duke discussion papers

Covariance graphs – variance matrix of sparse factor models



Graphical models – precision matrix



(Dobra et al 04, JMVA; **V8 Poster**

Jones et al 05, Stat Sci.)

Multivariate and matrix-variate time series

(Carvalho and West 2006, **V8 Poster**)

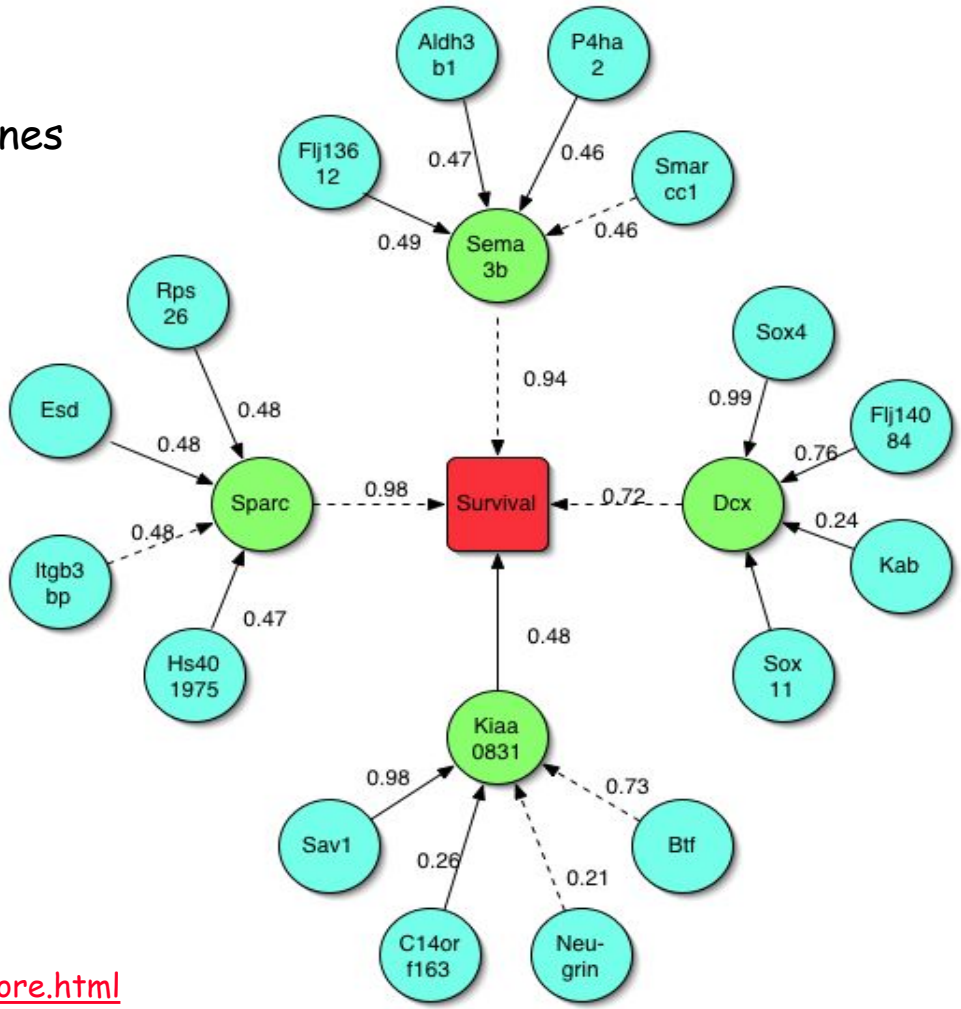
Regression Search for Multivariate Analysis

p=8400

Cascade of regression models:

- Models to predict/explain gene expression for survival predictive genes
- and so on …

Generate "graphs" of aspects of multivariate associations



Exploratory data analysis, visualization uses
http://www.stat.duke.edu/research/software/west/graphexplore.html

(Cancer Research, 05)