Modelling & Computation

Elements of Shrinkage Modelling in Regression and Multivariate Analysis
Regression: Shrinkage Prior Modelling
**Example: Anova in Cardiovascular Genomics**

Mice model aorta gene expression:
- 4 factors, each at 2 levels
- Response: gene expression
- Genes linked to design factors?
- Action is interactions

Apo E -/-, 6 wk Chow Diet
- PRECLINICAL DISEASE

Apo E -/-, 12 wk Western Diet
- EARLY/INTERMEDIATE DISEASE

Age, Diet, Gender, WT/ApoE
- $2^4 \times 5^+$ balanced factorial design
One gene, one sample

\[ z = \beta + \mu + \delta + \alpha + \gamma + \mu\delta + \mu\alpha + \mu\gamma + \delta\alpha + \delta\gamma + \alpha\gamma + \mu\delta\alpha + \mu\delta\gamma + \mu\alpha\gamma + \mu\delta\alpha\gamma + \text{noise} \]

\( p=12,500 \) genes in parallel

**Multivariate ANOVA**

- WT, 6wk, chow, fem (baseline)
- male
- fat diet
- age=12wk/old
- ApoE genotype
- fat diet & male
- 12wk/old & male
- ApoE & male

\[ z = \beta + \mu + \delta + \alpha + \gamma + \mu\delta + \mu\alpha + \mu\gamma + \delta\alpha + \delta\gamma + \alpha\gamma + \mu\delta\alpha + \mu\delta\gamma + \mu\alpha\gamma + \mu\delta\alpha\gamma + \text{noise} \]
Regression, Anova and Shrinkage Modelling

\[ z_i = h_i^T \beta + \epsilon_i, \quad \epsilon_i \sim N(0, \nu) \]
\[ z = H \beta + \epsilon, \quad \epsilon \sim N(0, \nu I) \]

LSE/MLE/Reference posterior:
\[ \hat{\beta} = (H'H)^{-1} H'z \]

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

Posterior:
\[ \beta | z \sim N(b, \nu B_*^{-1}) \]

Shrinkage:
\[ b = B_*^{-1} H'z \]
\[ B_* = \nu B + H'H \]

(n-vector response $z$

$n \times p$ known design matrix $H$

(minimal) Bayes: Shrinkage priors

Relevance of zero-mean location

LSE as limiting case
- no shrinkage -
Degrees and Dimensions of Shrinkage

\[ B^{-1} = \tau I \]
\[ b = (aI + H' H)^{-1} H' z \]
\[ a = v / \tau \]

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

\[ B^{-1} = \text{diag}(\tau_1, \ldots, \tau_p) \]
\[ \beta' = (\beta_1, \ldots, \beta_p) \]
\[ \beta_j \sim N(0, \tau_j) \]

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

Decision theory (other V8 tutorials)

Regularisation ... of \( H'H \)
Collinearity
Numerical instabilities in inversion
Large LSE variances

Role of scale factors
Ridge regression

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

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:**

\[ H' H = E' D^2 E \]

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

Orthogonal regression:

\[ \tilde{H}' \tilde{H} = I \]

Factors “underlying” structure in \( H \) are predictors

Proper priors: Shrinkage priors key

Regression on predictors \( H \)

Regression on factors

Small tail values ~ collinearities
Factor Regressions, Multiple Shrinkage and g-Priors

Orthogonal regression:
\[ \tilde{H}' \tilde{H} = I \]

Concordance of independent shrinkage priors in factor regression

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

\[ \begin{aligned}
\tilde{\beta} &= D E \beta \\
\beta &\leftarrow E' D^{-1} \tilde{\beta}
\end{aligned} \]

Single shrinkage in factor regression:
\[ \begin{aligned}
\tau_j &= 1/g, \forall j \\
\beta &\sim N(0, (H'H)^{-1}/g)
\end{aligned} \]

(Zellner) g-prior:
\[ \sim g/n \text{ "prior data" on same design ... } z=0 \]

Multiple shrinkage:

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

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

Bayesian coherence: design, n dependence, prediction?
Example: SVD Regression for Cookies

N=78: 39 training data
39 test data to predict

Spectral “wiggles” - wavelengths for fat absorbance

\[ \beta \leftarrow E' D^{-1} \tilde{\beta} \]
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:**

\[ \tau_j \sim rs / \chi_r^2 \]

IG has conditional conjugacy

$s$ - prior estimate \hspace{1cm} $r$ - tail weight

**Hyper-parameter specification:**

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

Other choices:
- $(r_j, s_j)$, dependent, non-IG, ...
- Exp: marginal priors Laplace (doubleExp)
- More useful/relevant shrinkage forms below
Iterative Posterior Computations

Prior on shrinkage parameters:
\[ \tau_j \sim r s / \chi_r^2 \]

Posterior computations:
- posterior modes (1970/80s)
- simulations (90s - current)

Exploit “complete” conditional posteriors

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

\[ \beta \sim N(0, T) \]
\[ T = \text{diag}(\tau_1, \ldots, \tau_k) \]
Simulate Posterior:
Estimation/inference uses sample means, histograms - Monte Carlo approximation of posterior

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

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

Irreducible (aperiodic) Markov chain on full ($\beta, T$) space

Limiting distribution: joint posterior

Posterior MC samples (dependent)
\[ z | \beta \sim N(H \beta, \nu I) \]

**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)

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

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

\[
\tau_j = (r + 1) s_j / \chi_r^2, \quad s_j = s + \beta_j^2
\]

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

... for MCMC convergence
Modular nature of many posterior MCMC implementations
e.g. response error variance

\[
p(\beta | z, T, v) = N(b, vB_\star^{-1})
\]

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

\[
p(v | z, \beta, T)
\]
e.g., inverse scaled chi-square for \(v\)

Add-on ("bolt on") components for hyperparameters
Some Responses are Binary
Gene expression as covariates (predictors)
Molecular phenotyping e.g.:
cancer outcomes

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

**Example: Binary Responses**
e.g.: Breast Cancer
ER - (O)Estrogen Receptor Status
Lymph node (recurrence risk) status

Clinical test - gene expression in tumour

\[ p(y|x) \]

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

(~60x magnification)
nuclei of breast epithelial cells
cytoplasm of breast epithelial cells
brown-red & pink ~ ER+
Binary Regression

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

\[ y_i = 0 \quad y_i = 1 \]

\[ Pr(y_i = 1 | \beta) = \Phi(h_i^T \beta) \]

\[ h_i = h_i(x_i) \]

Natural model/interpretation
Computationally nice

\[ Pr(y_i = 1) = Pr(z_i > 0), \quad z_i \sim N(h_i^T \beta, 1) \]

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

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

“Honest” assessment of precision

Heterogeneity, small samples

Feature/Variable selection

Critical component of predictive assessment with large $p$
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: $\gamma_j = \begin{cases} 0 & \iff \beta_j = 0 \\ 1 & \iff \beta_j \sim N(0, \tau_j) \end{cases}$

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

\[
(\beta_j | \tau_j, \pi) \sim (1-\pi)\delta_0 + \pi N(0, \tau_j), \quad \tau_j = rs/\chi_r^2
\]

Selection/sparsity shrinkage priors:

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

\[
p(\tau_j) : \quad \tau_j = \begin{cases} 0 & \text{with probability } \pi \\ rs/\chi_r^2 & \text{with probability } 1 - \pi \end{cases}
\]

Other versions:

- constant/specified $\tau_j = s$
- non-zero but “very small” $\tau_j$
$\Pr(\gamma_j = 1 | z) = \Pr(\beta_j \neq 0 | z)$

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

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

One model: $M_\gamma$ \hspace{2cm} $\Pr(M_\gamma | z)$

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

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

Regression variable “selection”
Model uncertainty

- Berger & Bayarri V8 Tutorial -

Subsets of regressors: Models
Simultaneous “multiple tests”

Model averaging for prediction

Computation:
Finding “interesting” models

MCMC and stochastic search
**Gibbs MCMC**: repeat scans through all variables - in/out?

\[ p(\pi | \gamma) \]
\[ p(\tau | z, \beta) \]

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

\[ = p(\beta_j | z, \gamma_j, -) p(\gamma_j | z, -) \]

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

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

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
Stochastic Search Methods

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 regions of model space

Catalogue of many “good” models

Parallelisation

KEY: easily compute

\[ \propto Pr(M_γ | Z) \]

(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 $\pi$

Dimension -

Implicit in Bayesian & other likelihood-based analyses

(cf. BIC)

Specification for base rate:

Scale with dimension to maintain parsimony, sparsity

Example: $p = 25$
Getting Multivariate
Example: $2^5$ Anova Mice Genomics

$z = x$ - gene expression

$h_i$ fixed design (0/1)

$k = 16$ parameters

$g =
\begin{align*}
1 : p &= 12,500 \\
\end{align*}$

$x_{g,i} = \beta'_g h_i + \epsilon_{g,i}, \quad \epsilon_{g,i} \sim N(0, \upsilon_g)$

$x_i = B h_i + \epsilon_i, \quad \epsilon_i \sim N(0, V), \quad V = \text{diag}(\upsilon_1, \ldots, \upsilon_p)$

Highly multivariate anova
Multivariate Anova

\[ x_i = B h_i + \epsilon_i \]

Sample \( i \)

\( x_i \) \((p \times 1)\)

Variable \( g \), factor \( j \)

\( B = \{\beta_{g,j}\} \) \((p \times k)\)

Fixed design

\( h_i \) \((k \times 1)\)

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

$\pi_j \sim \text{sparsity}$

Differing sparsity patterns

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

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

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

Local model stochastic search?
\[ \pi^*_g,j = Pr(\beta_{g,j} \neq 0|X) \]

Probabilities and log-likelihood ratios

- SHRINKAGE

Variable (gene) identification within interaction effects

(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

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

Latent factors:

Model of covariance matrix:

\[ x_i = Bh_i + \epsilon_i \]

\[ \epsilon_i \sim N(0, V) \]

Idiosyncratic variation

\[ \lambda_i \sim N(0, T) \]

\[ V(x_i) = BTB' + V \]

(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 \( j \):

\[ \# \{ \beta_{g,j} \neq 0 \} = 0, 1, \ldots, \text{small} \]

(West 2003, Valencia 7)
Uncertain sparsity patterns in latent factor models:

\[ x_i = B\lambda_i + \epsilon_i \]

Same model structure: design vector becomes uncertain

Add on/bolt on module to MCMC computations

Structure:
Recall multivariate regression:

\[ x_i = Bh_i + \epsilon_i \]

\[(MC)^2 + \prod_{i=1}^{n} p(\lambda_i | X, -)\]
Inference on Sparsity and Structure

\[ \cdots p(B|X) \]

Sample correlations

Fitted correlations in \( B'TB+\Psi \)
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)  (DNA replication)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDM2</td>
<td>MDM2</td>
</tr>
<tr>
<td>GSN</td>
<td>VCAM1</td>
</tr>
<tr>
<td>AGC1</td>
<td>CYP2A13</td>
</tr>
<tr>
<td>RBM8A</td>
<td>P1X1</td>
</tr>
<tr>
<td>MTHFS</td>
<td>MCM2</td>
</tr>
<tr>
<td>CUGBP2</td>
<td>DDX39</td>
</tr>
<tr>
<td>PKN2</td>
<td>MCM7</td>
</tr>
<tr>
<td>SSR4</td>
<td>GSN</td>
</tr>
<tr>
<td>FANCG</td>
<td>GSTM1</td>
</tr>
<tr>
<td>NME3</td>
<td>CCNE1</td>
</tr>
<tr>
<td>POU4F1</td>
<td>MCM3</td>
</tr>
<tr>
<td>AGC1</td>
<td>MCM4</td>
</tr>
<tr>
<td>MDM2</td>
<td>MFGES7</td>
</tr>
<tr>
<td>CRHR1</td>
<td>ABCA3</td>
</tr>
<tr>
<td>H1FX</td>
<td>CDKN2A</td>
</tr>
<tr>
<td>RPS3A</td>
<td>MCM5</td>
</tr>
<tr>
<td>ABCB8</td>
<td>KIAA1026</td>
</tr>
<tr>
<td>RGS12</td>
<td>TOMM70A</td>
</tr>
<tr>
<td>GLG1</td>
<td>CDC2</td>
</tr>
<tr>
<td>DOC-1R</td>
<td>SAS</td>
</tr>
<tr>
<td>TNPO3</td>
<td>POLR2H</td>
</tr>
<tr>
<td>MDM2</td>
<td>CSNK1D</td>
</tr>
<tr>
<td>MAPT</td>
<td>NME3</td>
</tr>
<tr>
<td>LOR</td>
<td>CDC6</td>
</tr>
<tr>
<td>GUCA1A</td>
<td>CHC1</td>
</tr>
<tr>
<td>GRIA1</td>
<td>TNFRSF14</td>
</tr>
<tr>
<td>CDC34</td>
<td>BACH</td>
</tr>
<tr>
<td>COL11A2</td>
<td>MVD</td>
</tr>
<tr>
<td>MYC</td>
<td>FANCG</td>
</tr>
<tr>
<td>TBL3</td>
<td>LIG1</td>
</tr>
<tr>
<td>BTF3</td>
<td>SF3B4</td>
</tr>
<tr>
<td>UCP3</td>
<td>CCNE1</td>
</tr>
<tr>
<td>LBA1</td>
<td>ABCF3</td>
</tr>
<tr>
<td>CDKN2C</td>
<td></td>
</tr>
<tr>
<td>HTR6</td>
<td></td>
</tr>
<tr>
<td>CDC6</td>
<td></td>
</tr>
<tr>
<td>CYP2A13</td>
<td></td>
</tr>
<tr>
<td>KHDRBS1</td>
<td></td>
</tr>
<tr>
<td>KIAA0284</td>
<td></td>
</tr>
<tr>
<td>PEX5</td>
<td></td>
</tr>
<tr>
<td>CYP2A6</td>
<td></td>
</tr>
<tr>
<td>LTK</td>
<td></td>
</tr>
<tr>
<td>SSTR3</td>
<td></td>
</tr>
<tr>
<td>MDM2</td>
<td></td>
</tr>
</tbody>
</table>

Cell Growth Metabolism

DNA Damage

2  DNA replication

4  DNA replication

Valencia 8: Modelling & Computation Tutorial - Mike West - June 1st 2006
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?

$\lambda_z = \text{response factors}$

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)
Stochastic search, regressions and graphs, *JMLR* 90

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

Sparse factor models

Sparse statistical modelling, *Bayesian Bioinformatics*

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

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

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

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)

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
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.)
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)