FACTOR REGRESSION MODELS

Mike West

Institute of Statistics & Decision Sciences

Duke University

June 6th 2002, Valencia VII - Tenerife

Scope

Regression with $p >> n$ — Factor regressions

- *Empirical factor models:*
  - SVD (PCA) regression variables
  - Multiple shrinkage priors – Generalised ridge regression
- *General class of latent factor models:*
  - Regression on latent factors
  - SVD (PCA) regression as special case
    - * resolves questions/issues in Bayesian SVD regression
- *Sparse latent factor models*
- *Examples*

SVD/PCA regression with $p >> n$

- $\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \nu$
- $p > n :$ $\mathbf{X}$ is "tall and skinny"
- SVD: $\mathbf{X} = \mathbf{AF}$ transforms model to $\mathbf{y} = \mathbf{F}'\boldsymbol{\theta} + \nu$ with $\boldsymbol{\theta} = \mathbf{A}'\boldsymbol{\beta}$
- Dimension reduction from p to n – many one
- priors: $\theta_i \sim T_k(0, 1)$
  or $N(0, \tau_i^2)$ with inverse gamma prior on τ_i^2
  - different "weights" in PCA/SV axes
  - conditionally conjugate, generalised "ridge regression" prior
- MCMC for inference on $\boldsymbol{\theta}, \tau_1, \ldots, \tau_k$
- binary regression: observe indicators of $y_i \geq 0$ for probit (or other)

Using SVD/PCA regression

Inference required for $\boldsymbol{\beta}$ where $\boldsymbol{\theta} = \mathbf{A}'\boldsymbol{\beta}$

- Multiple generalised inverses $\boldsymbol{\beta} = \mathbf{A}^{-}\boldsymbol{\theta}$
- Implicit prior(s) on $\boldsymbol{\beta}$
  - Generalised $g-$priors, generalised shrinkage

Issues:

- Choice of inverse transformation? Special choice $\boldsymbol{\beta} = \mathbf{A}\boldsymbol{\theta}$?
- Design-data dependent priors
  - *Prediction* at new design points
  - New design points, new parameters, new priors!
  - Fit model, define prior on *all* design points

SVD regression: Example

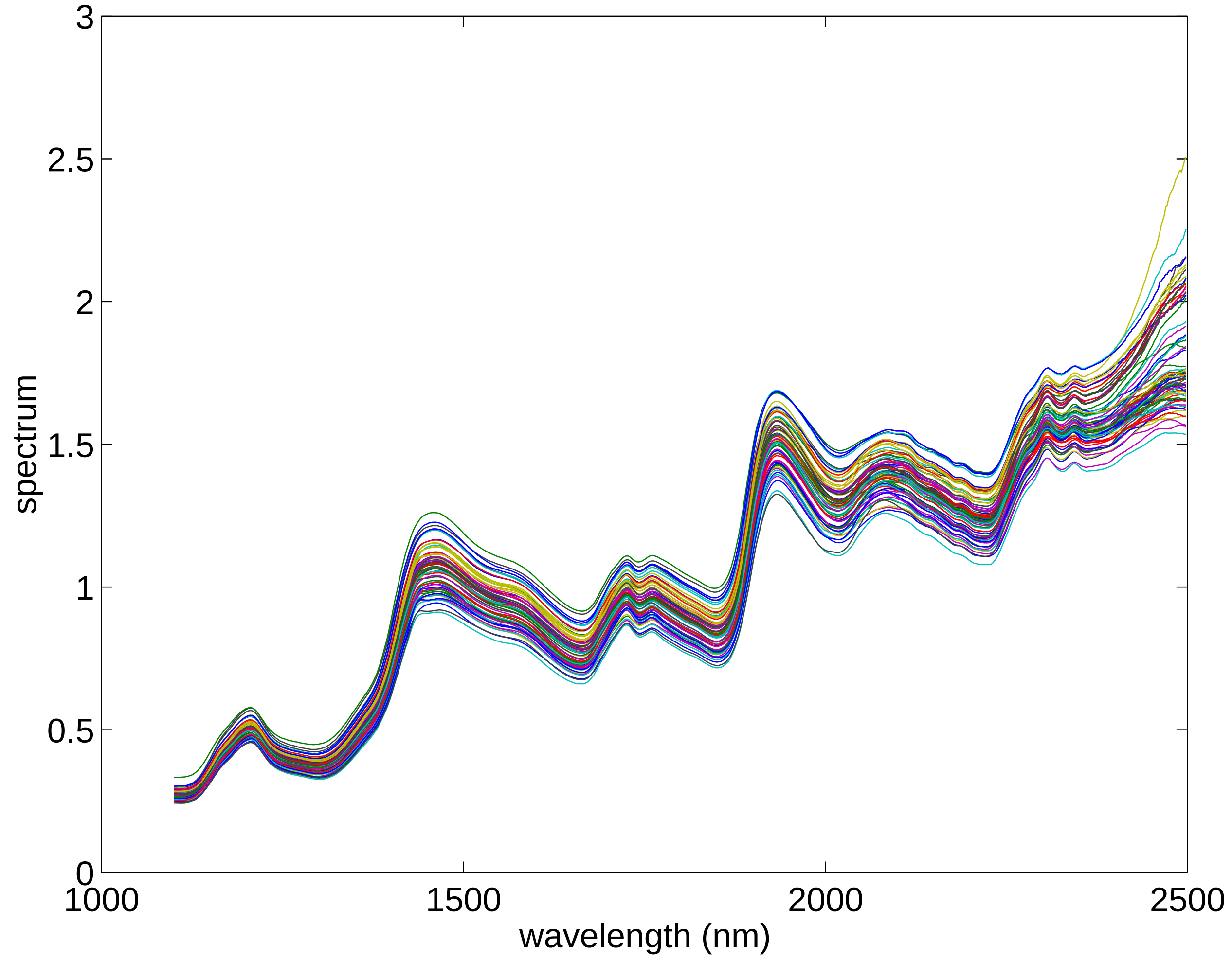
Cookie dough spectra - Brown, Fearn & Vanucci (1999 Biometrika)

- **Predictors**: Spectra: near-infrared spectroscopy of cookie dough
  NIR reflectance measures: spectrum over 300 wavelengths
- **Response**: fat content of cookies

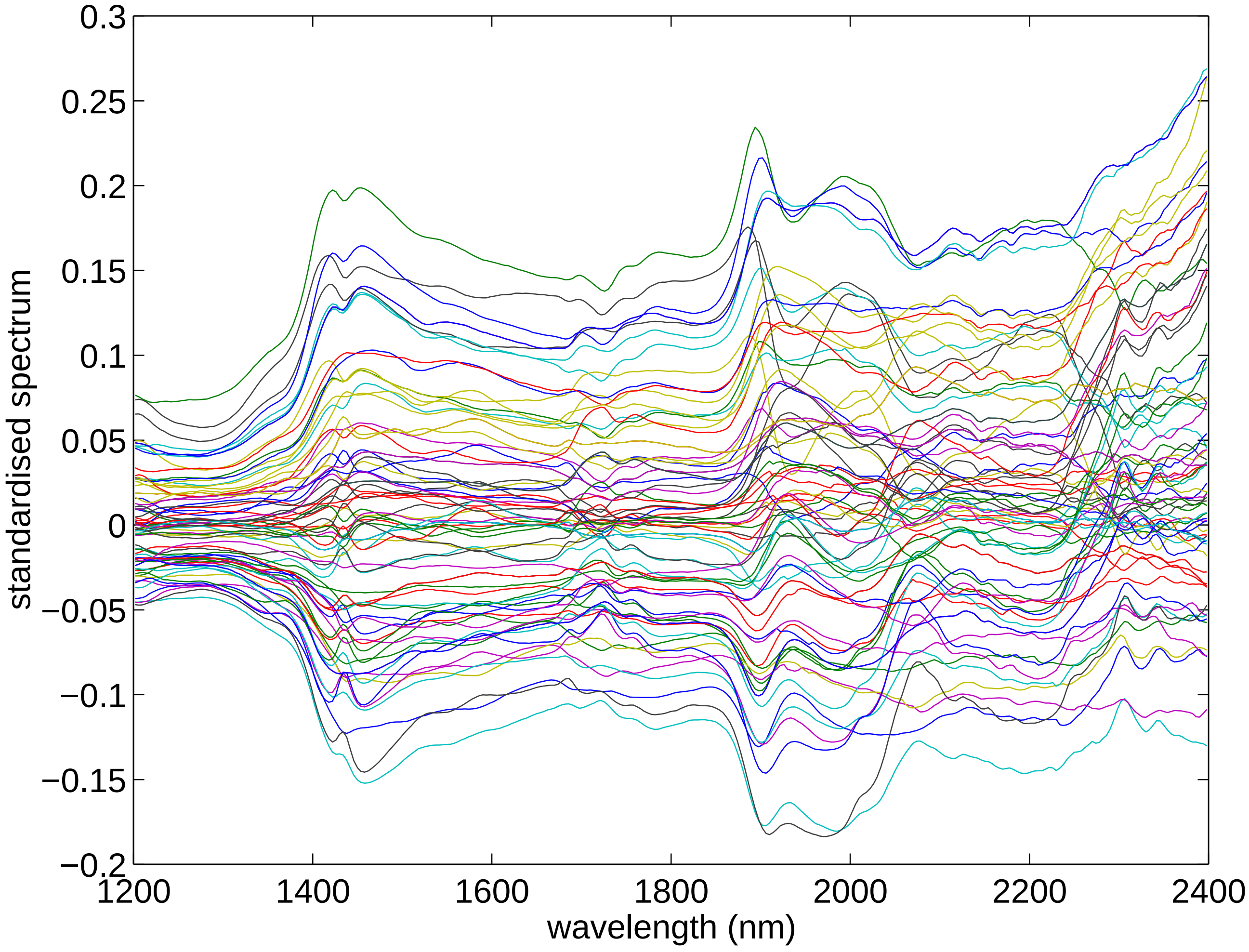
Aim: Predict fat content of cookies using features of spectrum

Analysis: 39 training cases, 39 validation cases

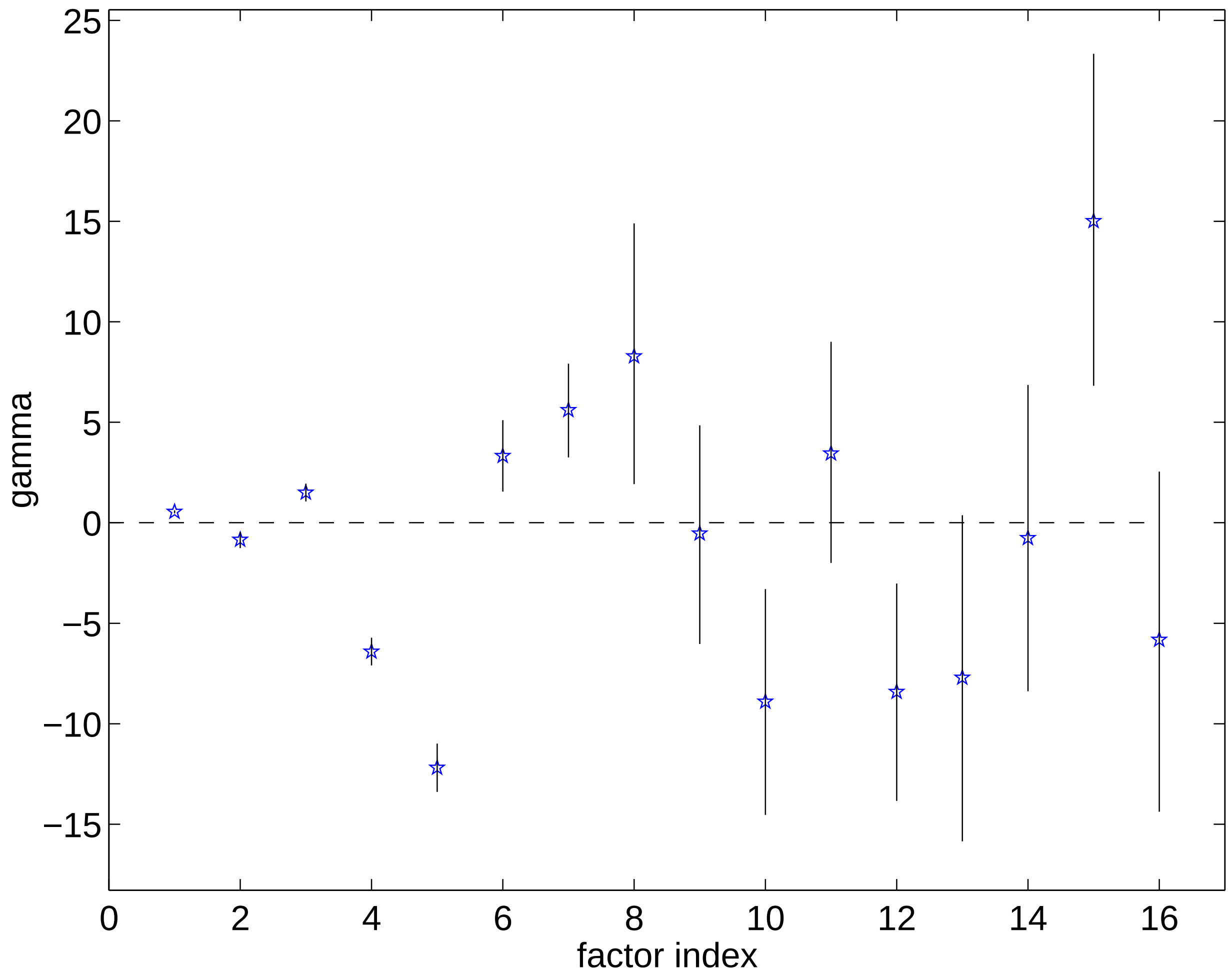
Spectra of 78 biscuit dough samples



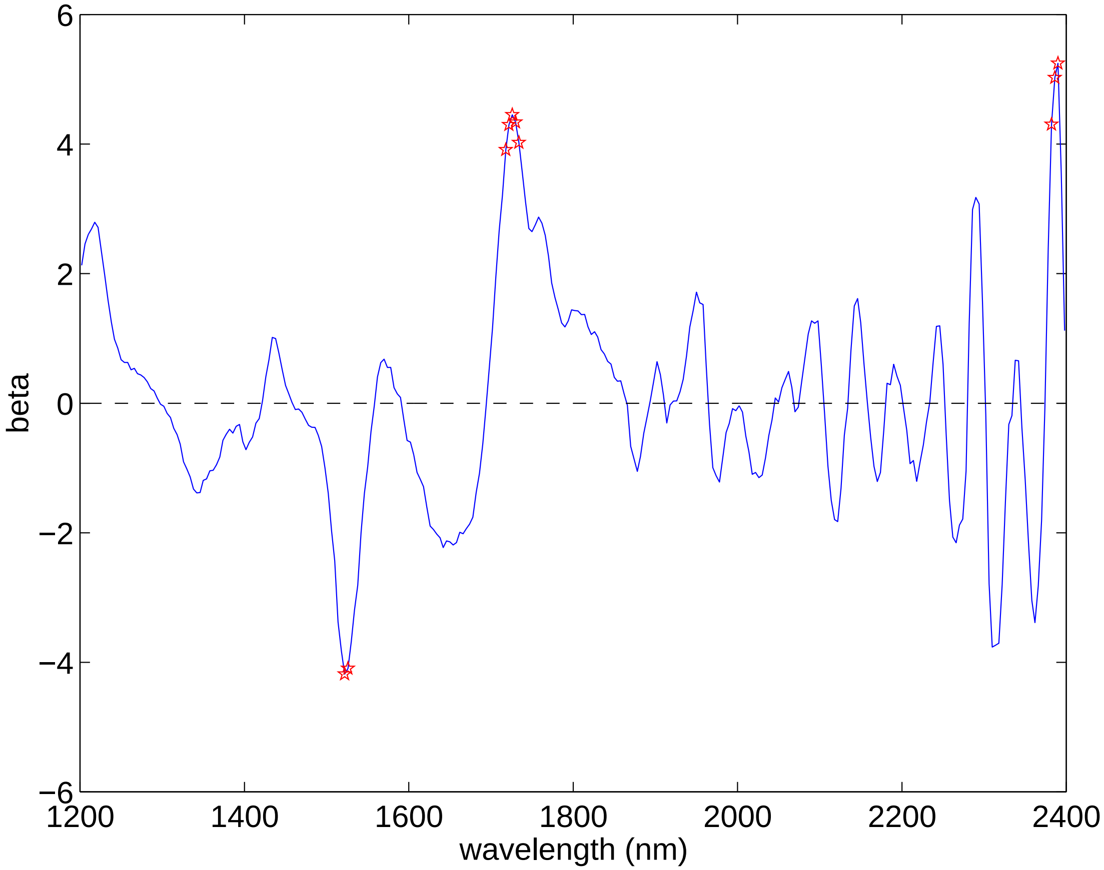
Standardised spectra of 78 biscuit dough samples



Factor coefficients θ



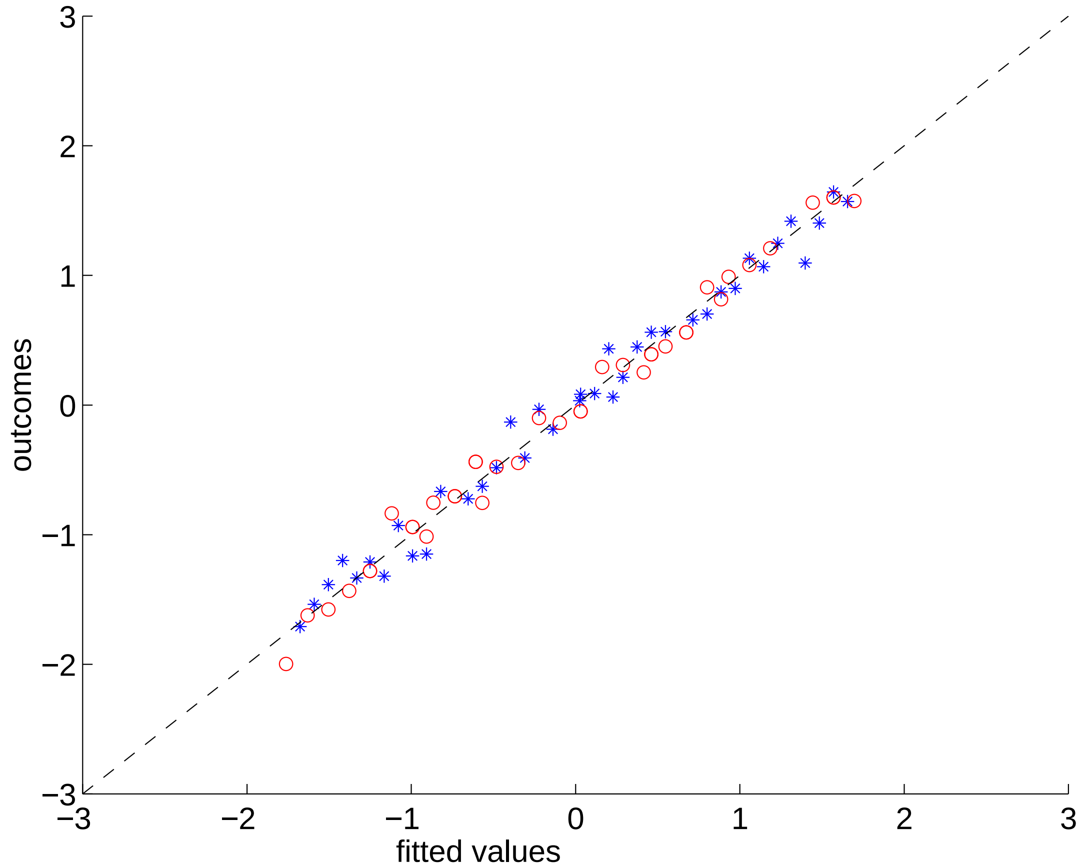
Regression coefficients β for wavelengths



* 10 largest post means

Cluster near 1720 is characteristic of fat absorbance (tiny wiggles in spectra)

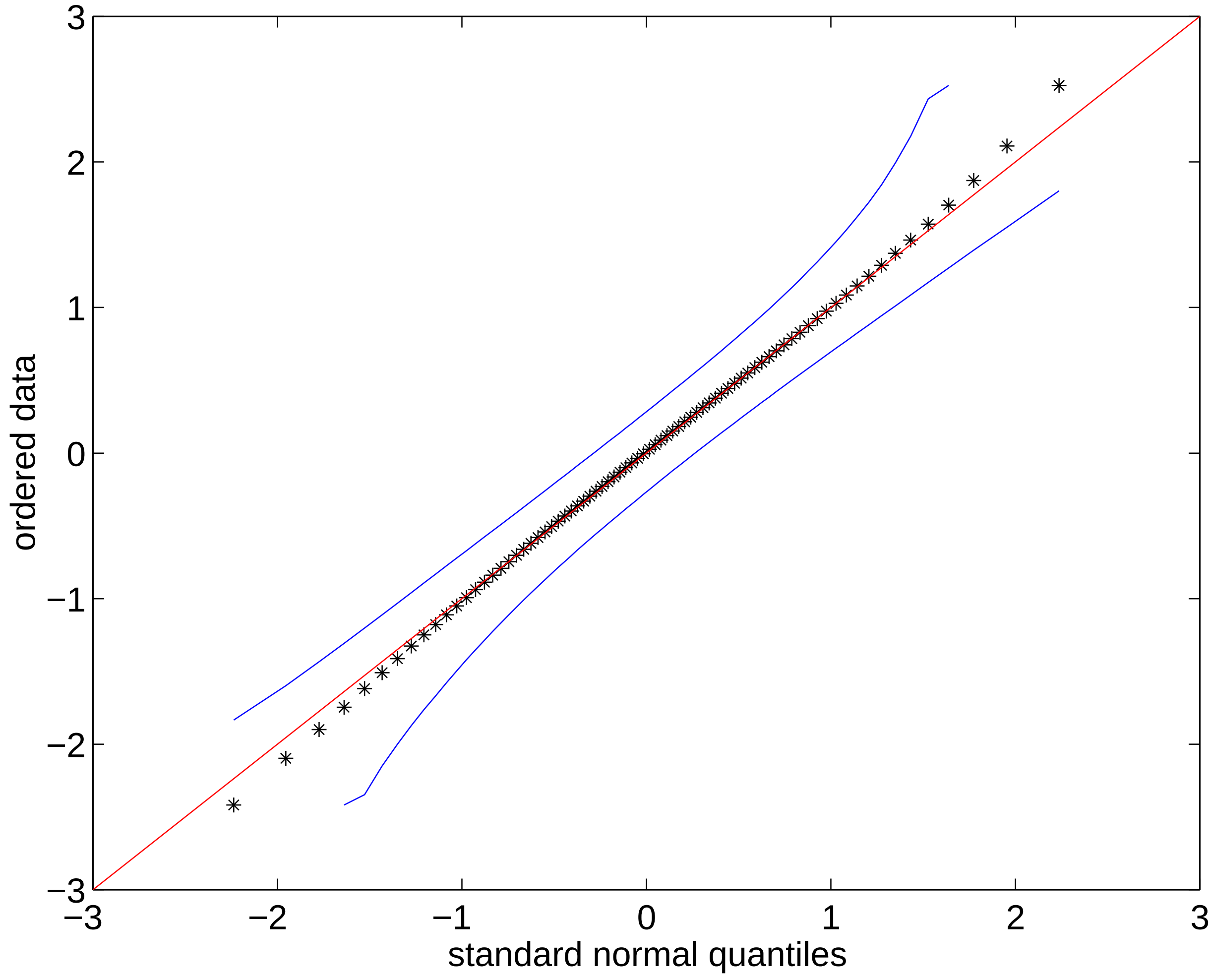
Response versus fitted values



* training data   o Validation cases

Better out-of-sample prediction MSE than original authors

Normal qqplot of fitted residuals



Latent factor regression models

Sample i: column i of $\mathbf{X}$ is

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

- $\boldsymbol{\lambda}_i \sim N(\mathbf{0}, \boldsymbol{\Delta})$ and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$
- diagonal variance matrices
- common patterns: (few) latent factors: $k = \mathbf{dim}(\boldsymbol{\lambda}_i)$

Regression:

$$y_i \sim N(\boldsymbol{\lambda}_i'\boldsymbol{\theta}, v)$$

- outcomes regress on latent factors in $\mathbf{x}_i$ – indirect regression on $\mathbf{x}_i$
- different outcomes relate to different latent factors

Latent factor models: SVD regression case

- Latent factor model defines $p(y_i, \mathbf{x}_i, \boldsymbol{\lambda}_i)$
- Implied $p(y_i|\mathbf{x}_i)$**:** linear regression of y_i on $\mathbf{x}_i$
- Linear regression coefficient $\boldsymbol{\beta} = \mathbf{H}\boldsymbol{\theta}$
- $\mathbf{H}$ depends on $\mathbf{B}, \boldsymbol{\Delta}, \boldsymbol{\Psi}$

Some implications:

- Prior on $\boldsymbol{\theta}$ implies (unique) prior on $\boldsymbol{\beta}$
- Limiting case: $\boldsymbol{\Psi} \rightarrow \mathbf{0}$ *(plus orthogonal columns)* leads to SVD regression
- Complete theoretical rationale for Bayesian SVD analysis

Analysis of latent factor models

- SVD is empirical, "noisy" estimates of factors, loadings
- Artificial orthogonality constraints
- Sample size dependence of number of factors

Fitting latent factor models:

- MCMC
- Coupled factor model with regression model
- Identification questions – constraints on loadings matrix **B**
- Informative priors

Sparse factor models

Interest in high-dimensional $\mathbf{x}$: $p >> n$

Example: Gene expression and molecular characterisation problems

- Samples are gene expression from tumour, cell line experiment, etc
- Regression to characterise subgroups or physiological/clinical outcome

Sparse factor concept:

- Many genes purely ideosyncratic
- Factors represent biology: Gene networks/pathways
  - A factor involves relatively small number of genes
  - A genes is involved in 1 factor, or a very few factors

Model sparsity: many zeros in $\mathbf{B}$

Fitting sparse factor models

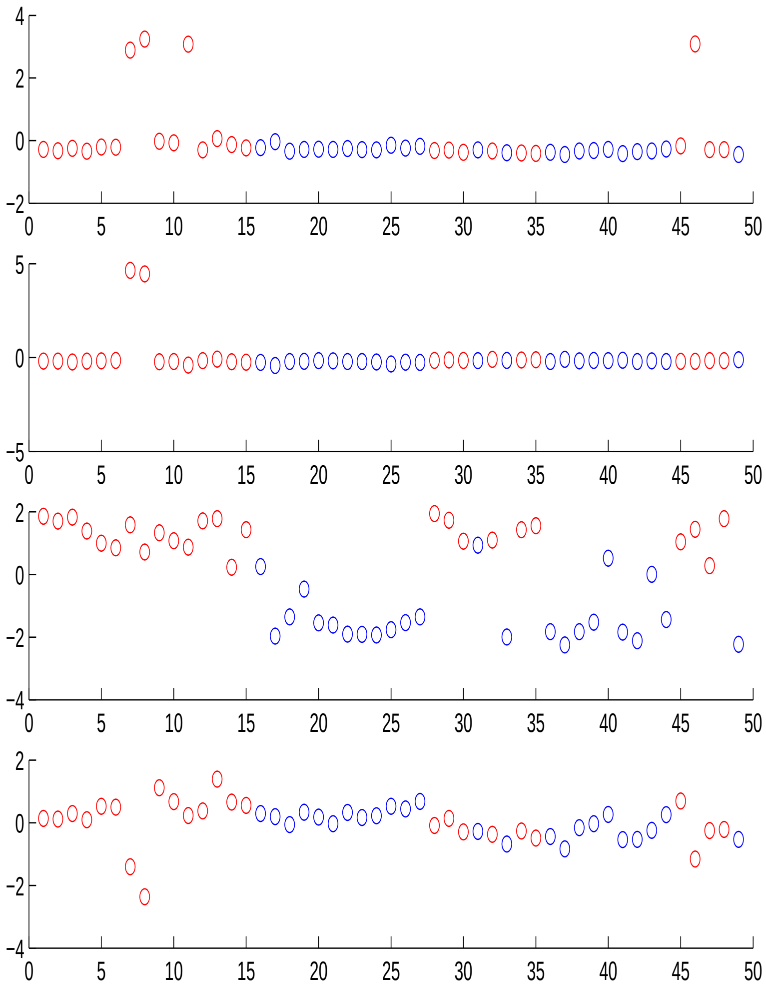
- Priors on factor loadings **B** induce many zeros
- Column (factor) j: elements independent, high probability of zero
- "Strong" prior to focus on dimension reduction within factors
- Computationally (very) intensive
- Confounding difficulties with too few factors in model
- Next stages of development will involve more elaborate prior structure

Breast cancer example:

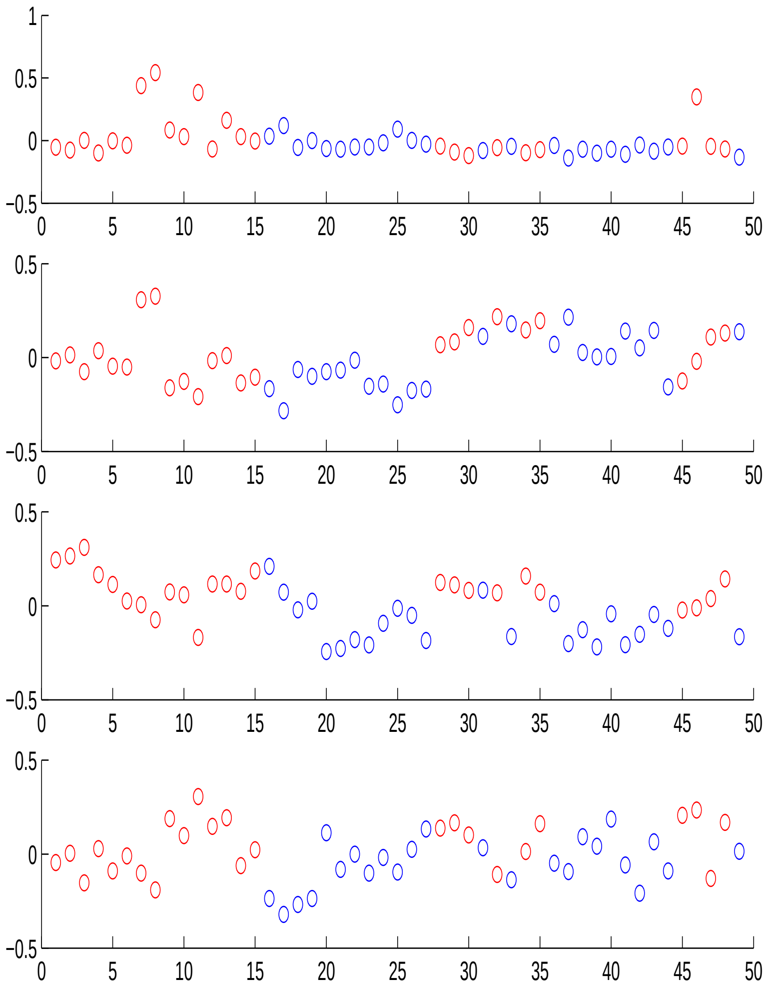
- Sparse factor analysis of 7000 genes, 25 factors
- Improves out-of-sample (CV) predictions in binary regressions
- ER factors - Oestrogen receptor status

Duke BC factor analysis: $p = 7000, \ k = 25$

Factor model



SVD



Current developments

- Computation: Cluster-based implementations
- Mixture models for latent factors
- Gaussian process/SVM regressions with kernels in factor space
- Hierarchical factor models – multiple layers
- Domain knowledge structuring priors

Co-conspirators:

Ming Liao (Duke)

Hedibert Lopes (UFRJ)

Rainer Spang (MPI, Berlin)