Semi-Latent Linear Models

Art B. Owen

Department of Statistics
Stanford University

Based on joint work with: Stuart Kim and Jacob Zahn

Stanford CA, June 2006
Genomics of aging

In work with the Kim lab, which genes change expression:

1. as we age?
2. as worms, mice, flies, · · · age?
3. as kidney, muscle, brain, · · · age?

Microarray data

\[ Y_{ij} \] expression of gene \( j \) sample \( i \)

\[ A_i \] age of sample \( i \)

\[ i = 1, \ldots, n \quad j = 1, \ldots, p \quad n \ll p \]

NB: Here we’re consumers of matrix algorithms
Many regressions

For gene $j$

\[ Y_{ij} = \beta_{0j} + \beta_{1j} A_i + \varepsilon_{ij}, \quad \text{or,} \]
\[ Y_{ij} = \beta_{0j} + \beta_{1j} A_i + \beta_{2j} S_i \varepsilon_{ij}, \quad \text{or,} \]
\[ Y_{ij} = \beta_{0j} + \beta_{1j} A_i + \beta_{2j} S_i + \beta_{3j} T_i + \varepsilon_{ij}, \]

where

\[ A_i = \text{age}, \quad S_i = \text{sex}, \quad T_i = \text{tissue type} \quad \text{etc.} \]

Mainly interested in

\[ \hat{\beta}_{1j}, \quad j = 1, \ldots, p \]
Multivariate regression

\[ Y = X\beta \]

- \( Y \) \( n \times p \) expression
- \( X \) \( n \times r \) per tissue predictors (1, age, sex, \ldots)
- \( \beta \) \( r \times p \) coefficients (2nd row for age coefs)

\[ \hat{\beta} = (X'X)^{-1}X'Y \]

Common questions:

- which genes are age related?
- how to adjust \( p \) values for multiple tests?
- how to adjust for correlated tests?
- which gene groups are age related?
Kidney data

Patient 95 is 81 years old . . . but looks younger
Rodwell et al. (2005) P.L.O.S.
Mouse data

Courtesy of Kevin Becker, National Institute on Aging

\[ p = 8932 \text{ genes} \]

\[ n = 40 \text{ mice:} \]

- 5 male and 5 female
- ages 1, 6, 16, 24 months

16 tissues:

- Adrenal, Bone marrow, Cerebellum, . . ., Spleen, Striatum, Thymus
“Genetic” age

Minimize

\[ SS = \sum_{i=1}^{n} \sum_{j=1}^{p} (Y_{ij} - \beta_{0j} - \beta_{1j} A_i - \beta_{2j} S_i)^2 \]

over \( \beta \) and \( A_1, \ldots, A_n \)

Every mouse picks its own ‘age’ \( A_i \)

Uses it for all 8932 genes
Results

Good news: \( p > 1 \) so model does not give \( SS = 0 \)

Medium news: \( A_i \) need to be normalized \( A_i \beta_{1j} = \frac{A_i}{2} (\beta_{1j} \times 2) \)

Bad news: fitted \( A_i \) seem unrelated to age

Interpretation

\( A_i \) pick out some dominant latent structure
this need not be age

Therefore

Try

\[
\beta_{0j} + \beta_{1j} A_i + \beta_{2j} S_i + \beta_{3j} Z_i
\]

for actual age \( A_i \), latent \( Z_i \)
Model

\[ Y = X\beta + Z\gamma \]

- **Y**: \( n \times p \) Response \( n \) obs in \( \mathbb{R}^p \)
- **X**: \( n \times r \) Measured predictors \( n \) obs in \( \mathbb{R}^r \)
- **\beta**: \( r \times p \) Coefficients
- **Z**: \( n \times s \) Latent predictors \( n \) values in \( \mathbb{R}^s \)
- **\gamma**: \( s \times p \) Coefficients

Minimize \( \| Y - X\beta - Z\gamma \|_F \) over \( \beta, \gamma, Z \)
Rorschach model

\[ \text{Minimize} \quad \beta \gamma Z \quad \|Y - X\beta - Z\gamma\|_F \]

Looks like:

- Regression \( \|Y - X\beta\|_F \)
- Factor analysis \( \|Y - Z\gamma\|_F \)
- Golub Hoffman & Stewart (1987)
- Tukey’s 1 df for interaction
- Structural equation models

Extends to:

\[ \|Y - X\beta - Z\gamma - \delta W\|_F \]
- \( t \times p \) matrix \( W \) with \( t \) ’per gene’ measurements

Published in:

- Gabriel (1978) JRSS-B linear bi-linear

Special case: additive main effects plus multiplicative interaction

- Fisher and Mackenzie (1923) J Ag Sci
- popular in crop science to this day
Solution for $\beta$

$$\text{Min } ||Y - X\beta - Z\gamma||_F$$

$X$ full rank, soln still not unique

As $Z \rightarrow Z + X\theta \quad \theta \in \mathbb{R}^{r \times s}$

and $\beta \rightarrow \beta - \theta\gamma$

$X\beta + Z\gamma$ unchanged

WLOG $Z'X = 0$

or else $Z \rightarrow Z - X(X'X)^{-1}X'Z$

Given $Z\gamma$

$$\hat{\beta} = (X'X)^{-1}X'(Y - Z\gamma) = (X'X)^{-1}X'Y$$
Solution for $Z\gamma$

Minimize

$$\min \|Y - X\hat{\beta} - Z\gamma\|_F$$

over $Z \in \mathbb{R}^{n \times s}$, $\gamma \in \mathbb{R}^{s \times p}$

subject to $Z'X = 0$

The unconstrained solution . . .

Let $Y - X\hat{\beta} = U\Sigma V'$ (SVD)

$Z = \text{first } s \text{ columns of } U$

$\hat{\gamma} = \text{first } s \text{ rows of } \Sigma V'$

. . . satisfies the constraint

$$0 = (Y - X\hat{\beta})'X \implies U'X = 0 \implies Z'X = 0$$

Solution is not unique

$$\gamma \rightarrow A\gamma \text{ cancels } Z \rightarrow ZA^{-1}$$
Power iterations

WLOG $Z'Z = I$ then $Z$ unique up to rotation $Z \rightarrow ZQ$

Given $Z$:

$$\hat{\gamma} = (Z'Z)^{-1}Z'(Y - X\hat{\beta}) = (Z'Z)^{-1}Z'Y$$

Given $\gamma$:

$$\tilde{Z} = (Y - X\hat{\beta})\gamma'(\gamma\gamma')^{-1}$$

$$\tilde{Z} = QR \quad \text{(QR decompose)}$$

$$\hat{Z} = Q$$

Notes

Iteration preserves $Z'X = 0$

Often faster than svd function
Some latent variables

Latent variable by tissue

Cerebellum

Cerebrum

Gonad

Spinal

Histograms of up to 40 mice

Stanford CA, June 2006
Latent variables
Three kinds of mice?

Latent variables of mice
Outliers: not the same mouse

Latent variables of mice

Gonad

SpinalCord

Striatum
Latent var strongly influences some genes in Cerebellum

But not in Cerebrum
Inference

Regression on Const, Age and Sex

$3 \times 8932$ parameters

Regression on Const, Age, Sex and 1 Latent

$4 \times 8932 + 40$ parameters

Is it like adding $1 + \frac{40}{8932} \doteq 1.0045$ parameters per regression?

(no) mice are nearly independent but genes are strongly correlated
Permutation

Repeat many times:
  Randomly permute ages of
    20 male mice
    20 female mice
  Recompute the model
  Count significant genes
Tabulate

rationale:
The permutation world has no age related genes
yet preserves all the correlation structure among genes

Find that:
including a latent variable increases (true and) false discoveries
More aging genes

at nominal $p = 0.001$
Results at nominal $p = 0.001$

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
<th>Latent</th>
<th>Perm $\geq$ Raw</th>
<th>Perm $\geq$ Latent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adrenal</td>
<td>20</td>
<td>200</td>
<td>0.075</td>
<td>0.048</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>17</td>
<td>54</td>
<td>0.111</td>
<td>0.273</td>
</tr>
<tr>
<td>BoneMarrow</td>
<td>3</td>
<td>4</td>
<td>0.444</td>
<td>0.704</td>
</tr>
<tr>
<td>Cerebrum</td>
<td>8</td>
<td>330</td>
<td>0.190</td>
<td>0.219</td>
</tr>
<tr>
<td>Eye</td>
<td>256</td>
<td>356</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>Gonad</td>
<td>45</td>
<td>105</td>
<td>0.012</td>
<td>0.341</td>
</tr>
<tr>
<td>Heart</td>
<td>23</td>
<td>113</td>
<td>0.064</td>
<td>0.137</td>
</tr>
<tr>
<td>Hippocampus</td>
<td>2</td>
<td>9</td>
<td>0.576</td>
<td>0.554</td>
</tr>
<tr>
<td>Kidney</td>
<td>14</td>
<td>22</td>
<td>0.140</td>
<td>0.282</td>
</tr>
<tr>
<td>Liver</td>
<td>0</td>
<td>641</td>
<td>1.000</td>
<td>0.073</td>
</tr>
<tr>
<td>Lung</td>
<td>89</td>
<td>462</td>
<td>0.010</td>
<td>0.012</td>
</tr>
<tr>
<td>Muscle</td>
<td>8</td>
<td>143</td>
<td>0.179</td>
<td>0.232</td>
</tr>
<tr>
<td>Spleen</td>
<td>28</td>
<td>81</td>
<td>0.068</td>
<td>0.261</td>
</tr>
<tr>
<td>SpinalCord</td>
<td>82</td>
<td>231</td>
<td>0.007</td>
<td>0.127</td>
</tr>
<tr>
<td>Striatum</td>
<td>0</td>
<td>197</td>
<td>1.000</td>
<td>0.296</td>
</tr>
<tr>
<td>Thymus</td>
<td>346</td>
<td>1310</td>
<td>0.004</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Stanford CA, June 2006
Number of genes picked

Blue = under permutation  Red = original

Plain regression

With latent

Bone/Marrow

Aging genes p = 0.001

Adrenal

Aging genes p = 0.001
Next steps

Calibrate significance when latent variables present

Build in false discovery estimates
Thanks

Gene Golub and Lek-Heng Lim

Stuart Kim and Jacob Zahn

Patrick Perry, Jerome Friedman, Ingram Olkin, David Rogosa

Kevin Becker