Semi-Latent Linear Models

Art B. Owen

Department of Statistics

Stanford University

Based on joint work with: Stuart Kim and Jacob Zahn

Genomics of aging

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

1. as we age?
2. as worms, mice, flies, $\cdots$ age?
3. as kidney, muscle, brain, $\cdots$ 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 = age, S_i = sex, T_i = tissue type etc.

Mainly interested in

$$\hat{\beta}_{1j}, \quad j = 1, \ldots, p$$

Multivariate regression

$$Y \doteq X\beta$$

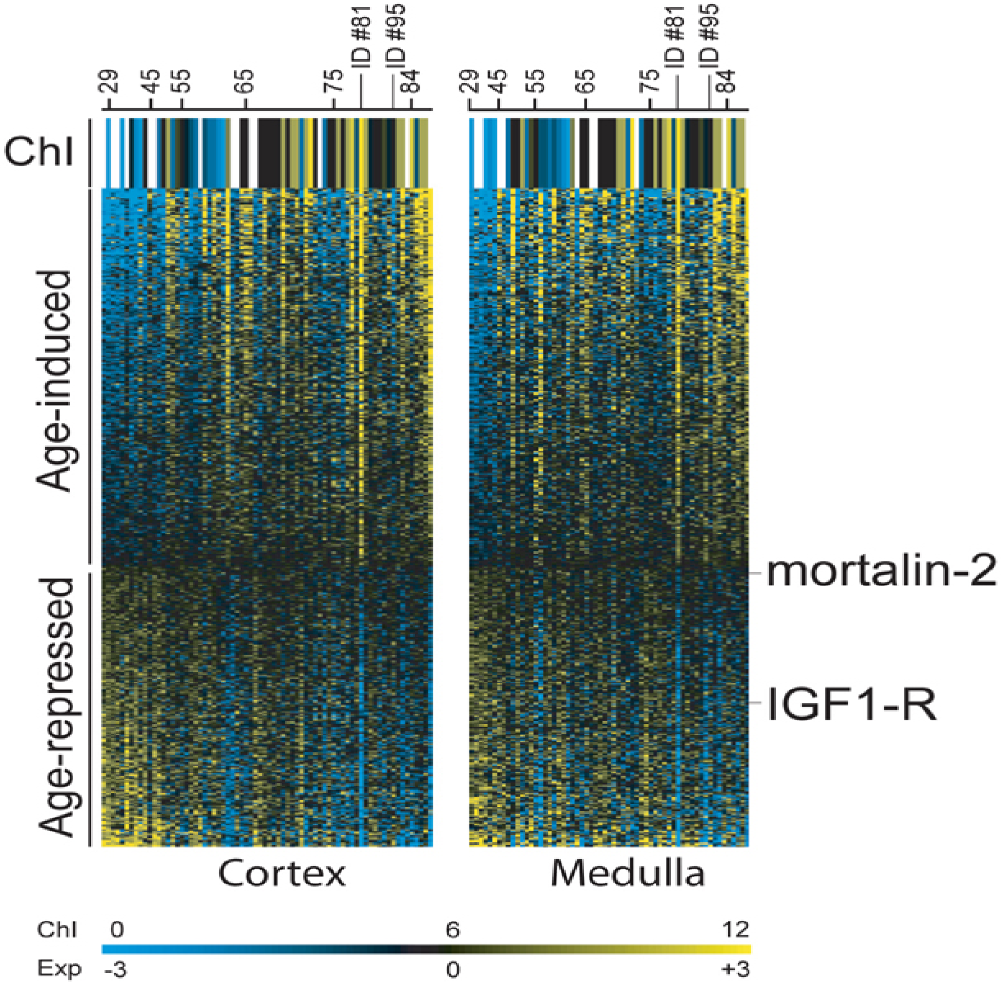
Y	$n \times p$	expression
X	$n \times r$	per tissue predictors (1, age, sex, . . .)
β	$r \times p$	coefficients (2nd row for age coefs)

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad r \times p$$

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

$p = 8932$ genes

$n = 40$ 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 β and $A_1, \ldots, A_n$

Every mouse picks it’s 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

$$Y \doteq X\beta + Z\gamma$$

Y	$n \times p$	Response	n obs in $\mathbb{R}^p$
X	$n \times r$	Measured predictors	n obs in $\mathbb{R}^r$
β	$r \times p$	Coefficients	
Z	$n \times s$	Latent predictors	n values in $\mathbb{R}^s$
γ	$s \times p$	Coefficients	

Minimize $\|Y - X\beta - Z\gamma\|_F$ over β, γ, Z

Rorschach model

$$\underset{\beta\ \gamma\ Z}{\text{Minimize}} \quad \|Y - X\beta - Z\gamma\|_F$$

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 β

Min $\|Y - X\beta - Z\gamma\|_F$

X full rank, soln still not unique

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

and $\quad \beta \longrightarrow \beta - \theta\gamma$

$X\beta + Z\gamma$ unchanged

WLOG $Z'X = 0$

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

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

subject to $\quad Z'X = 0$

The unconstrained solution . . .

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

$Z =$ first s columns of U

$\hat{\gamma} =$ first s 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

$$\gamma \to A\gamma \quad \text{cancels} \quad Z \to ZA^{-1}$$

Power iterations

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

Given Z:

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

Given γ:

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

$\widetilde{Z} = QR$ (QR decomp)

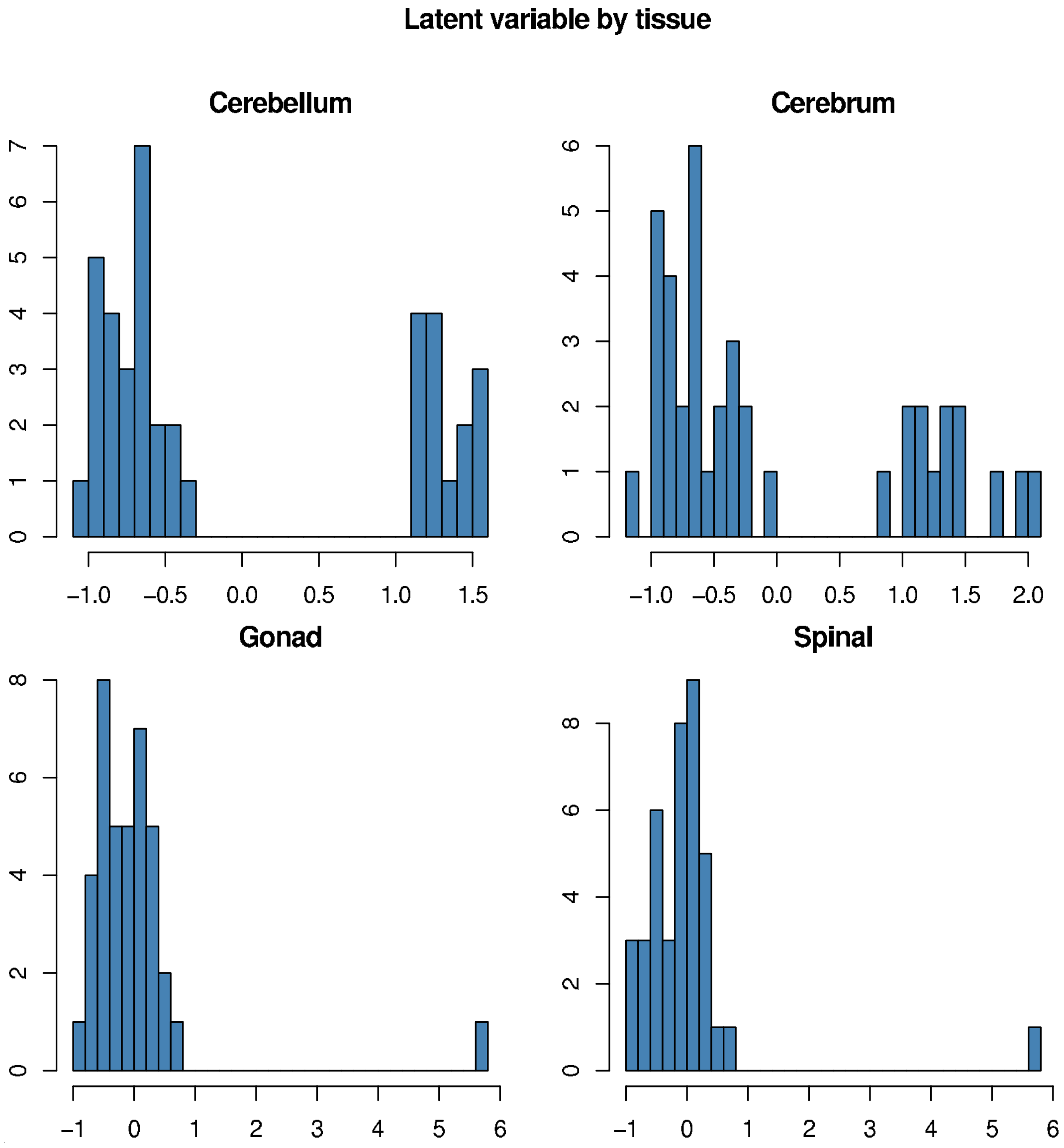
$$\hat{Z} = Q$$

Notes

Iteration preserves $Z'X = 0$

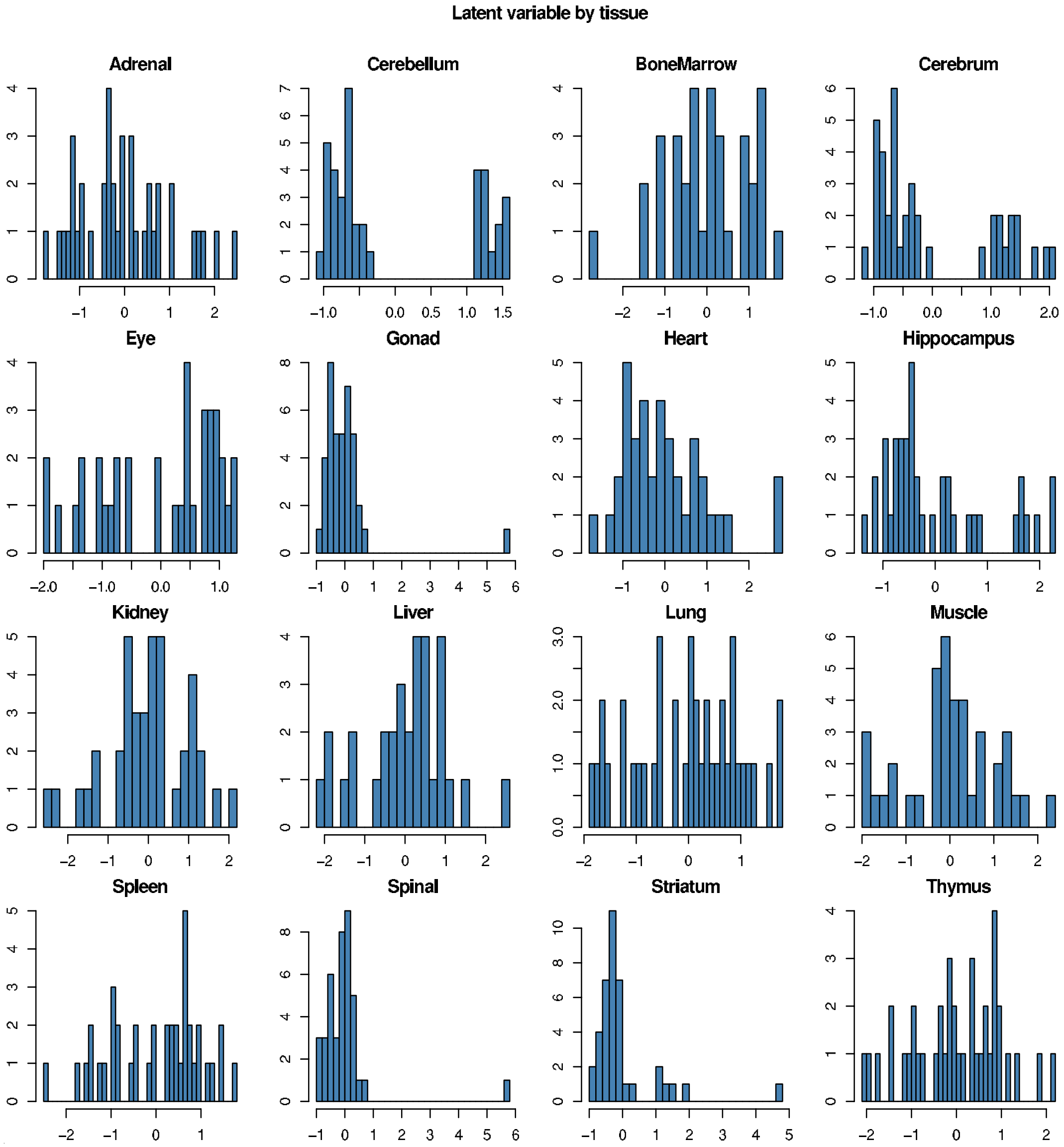
Often faster than svd function

Some latent variables

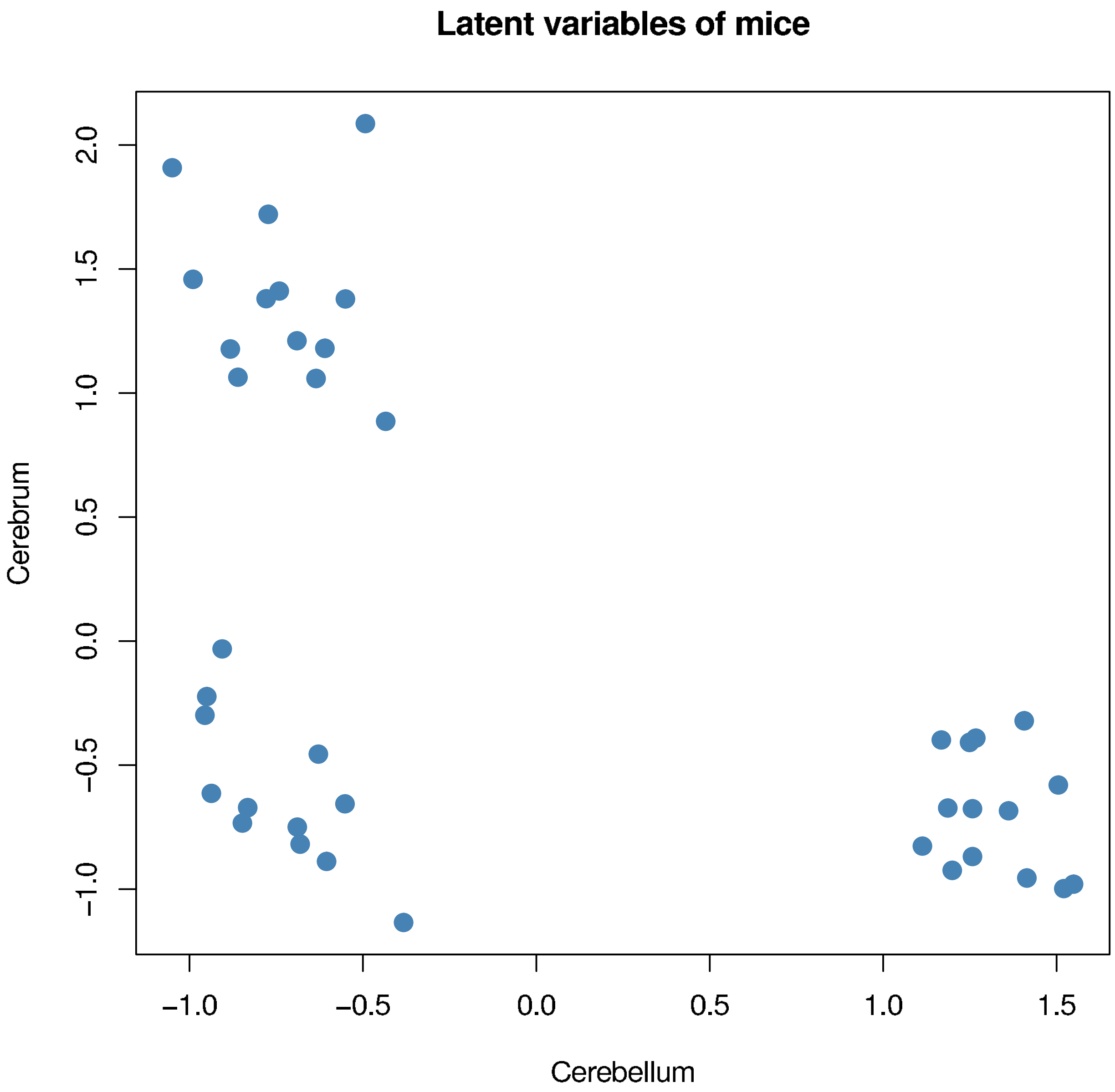


Histograms of up to 40 mice

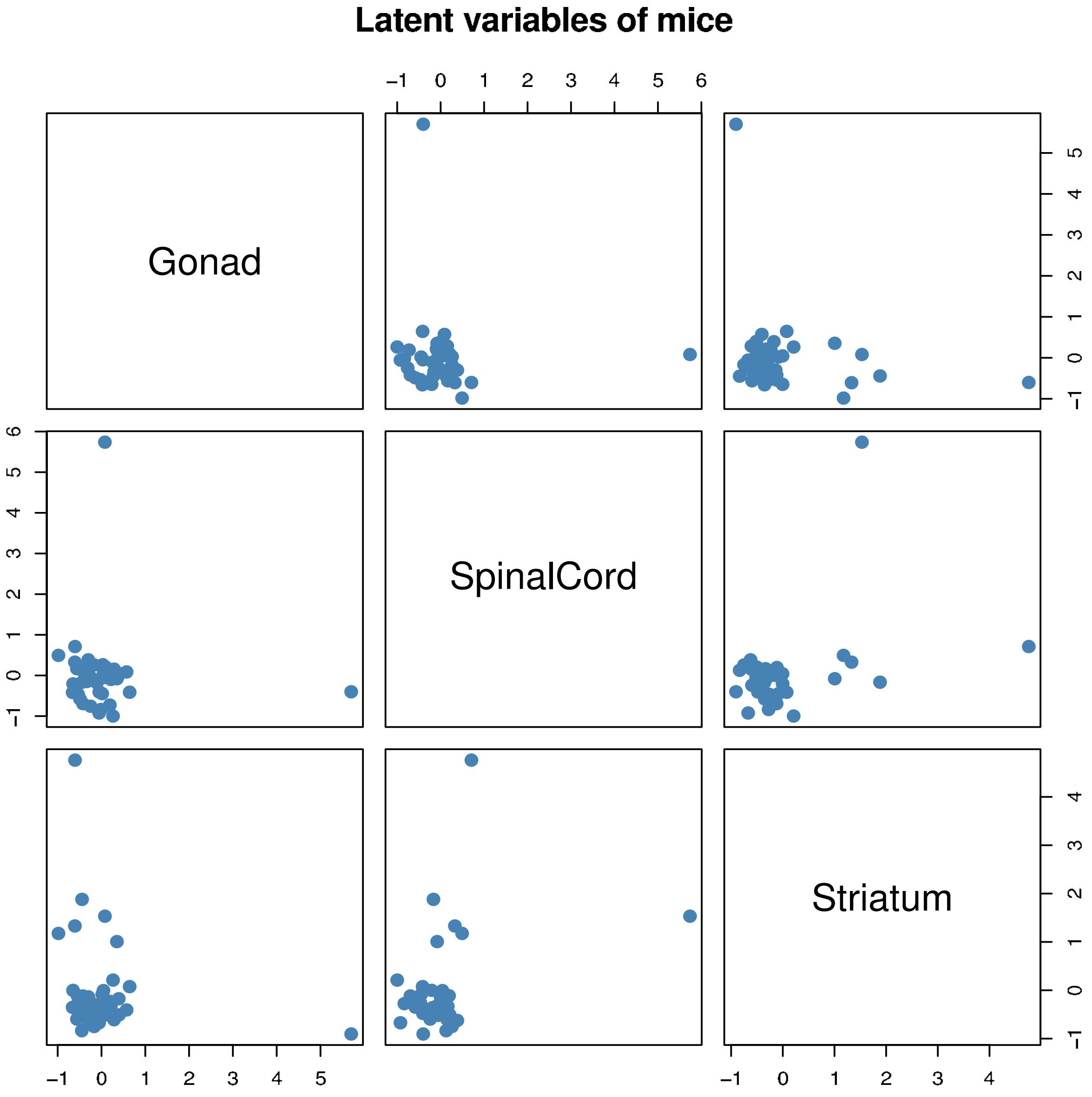
Latent variables



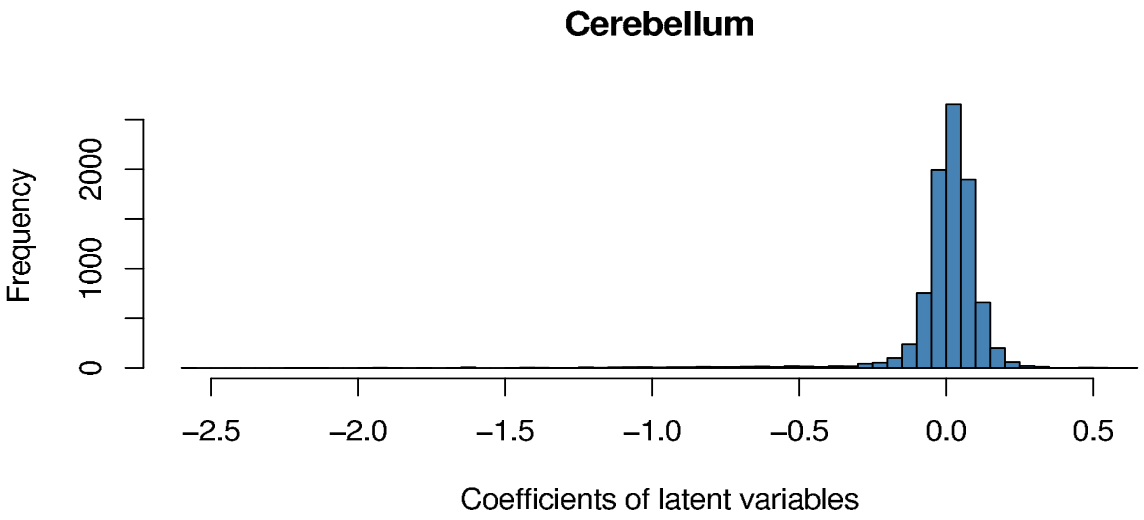
Three kinds of mice?

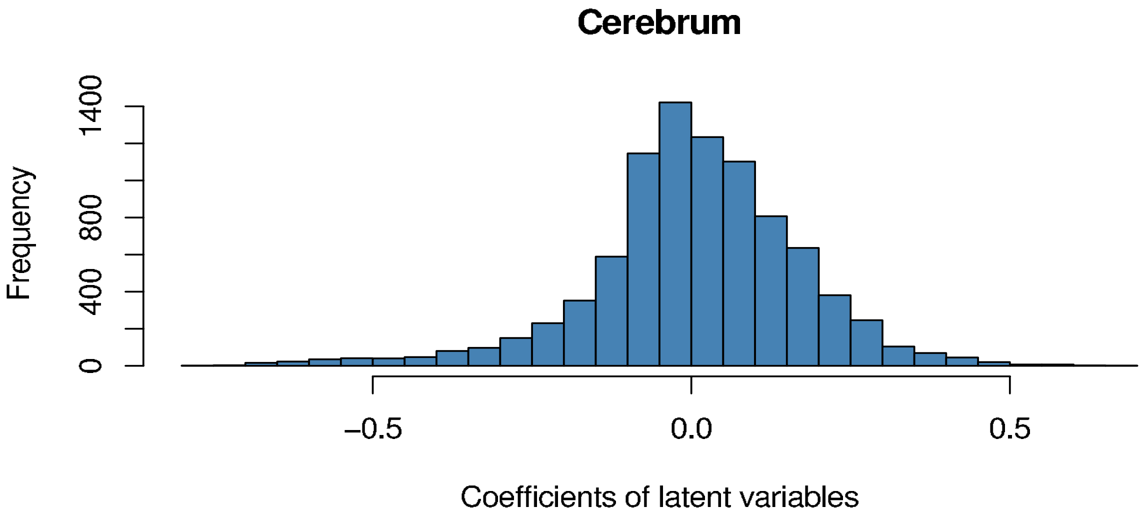


Outliers: not the same mouse



Latent var strongly influences some genes in Cerebellum





But not in Cerebrum

Inference

Regression on Const, Age and Sex

3×8932 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

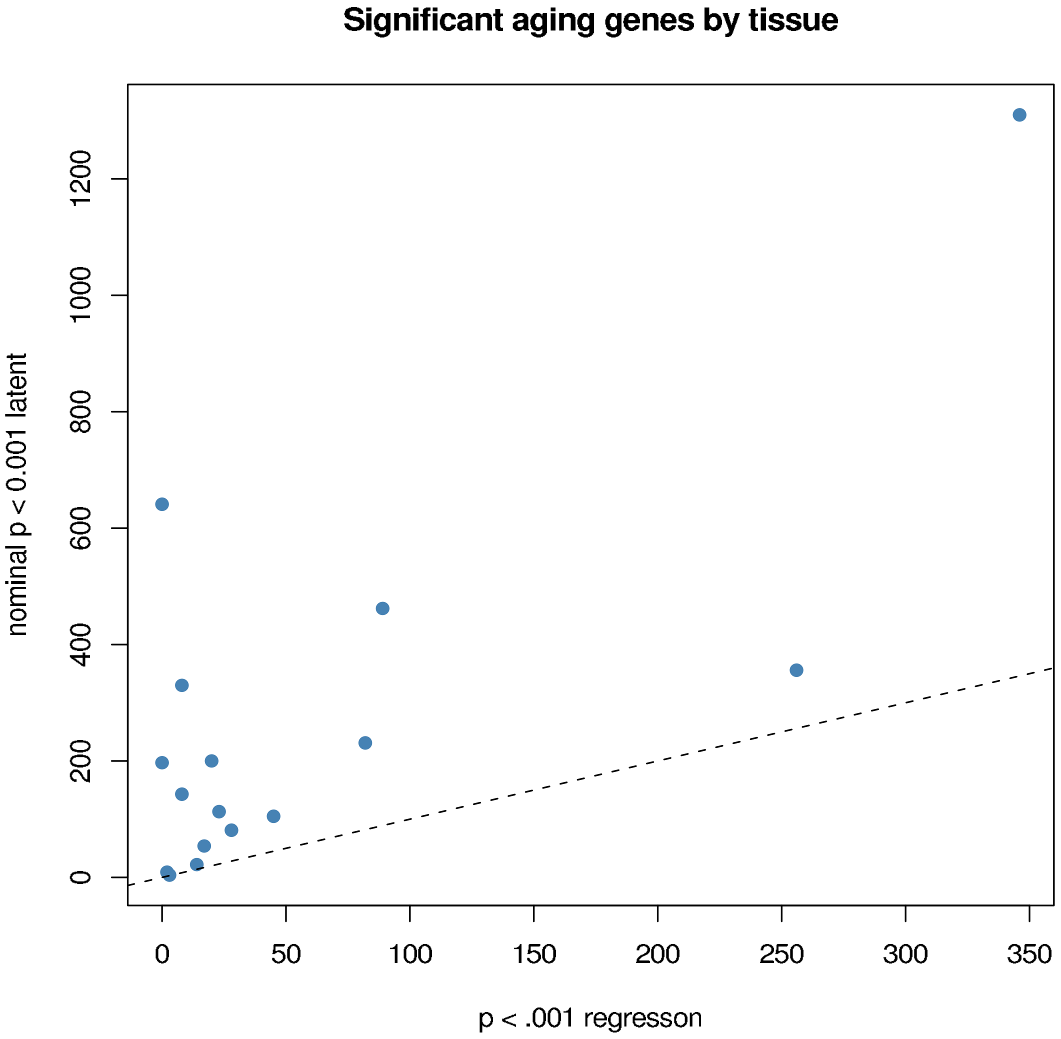
yet preserves all the correlation structure among genes

Find that:

including a latent variable increases (true and) false discoveries

More aging genes

at nominal $p = 0.001$

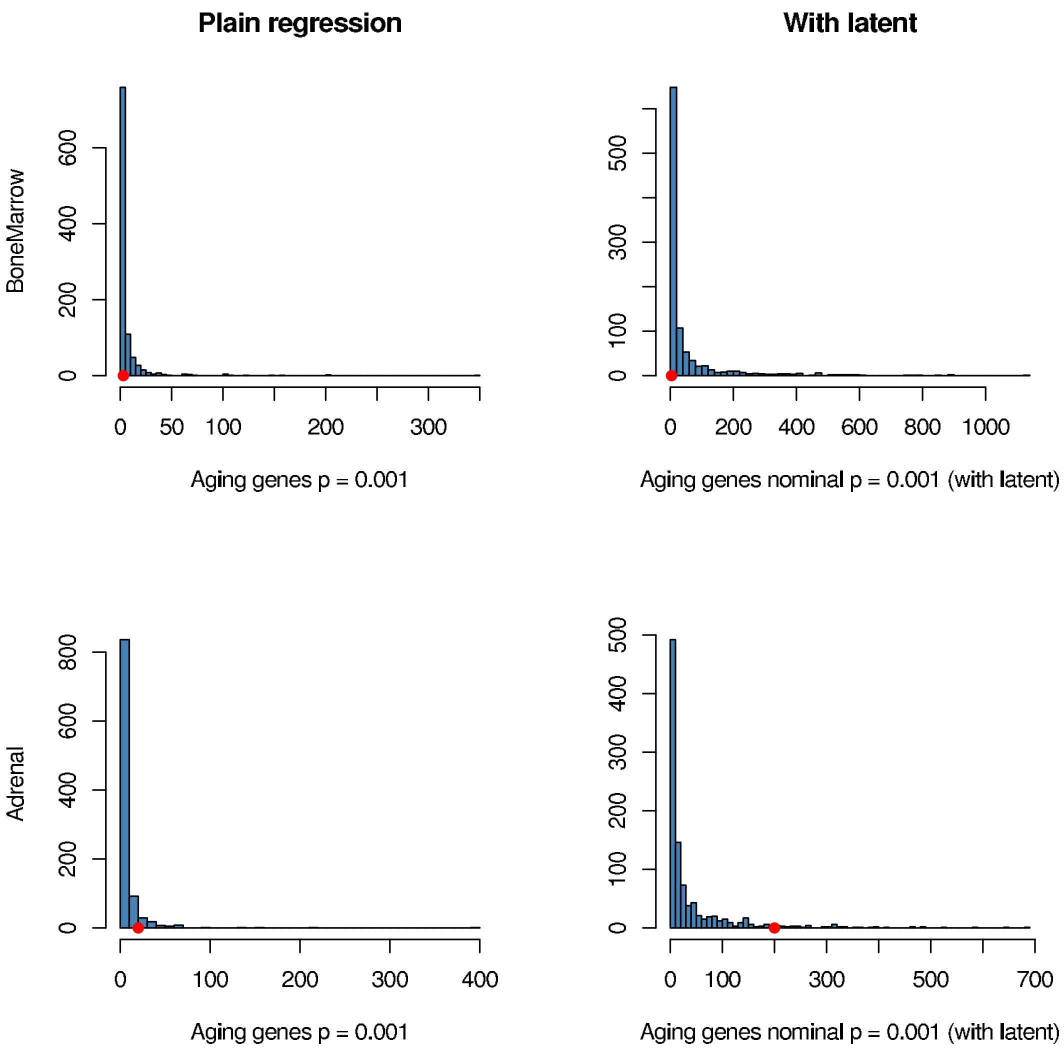


Results at nominal p $= 0.001$

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

Number of genes picked

Blue = under permutation Red = original



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