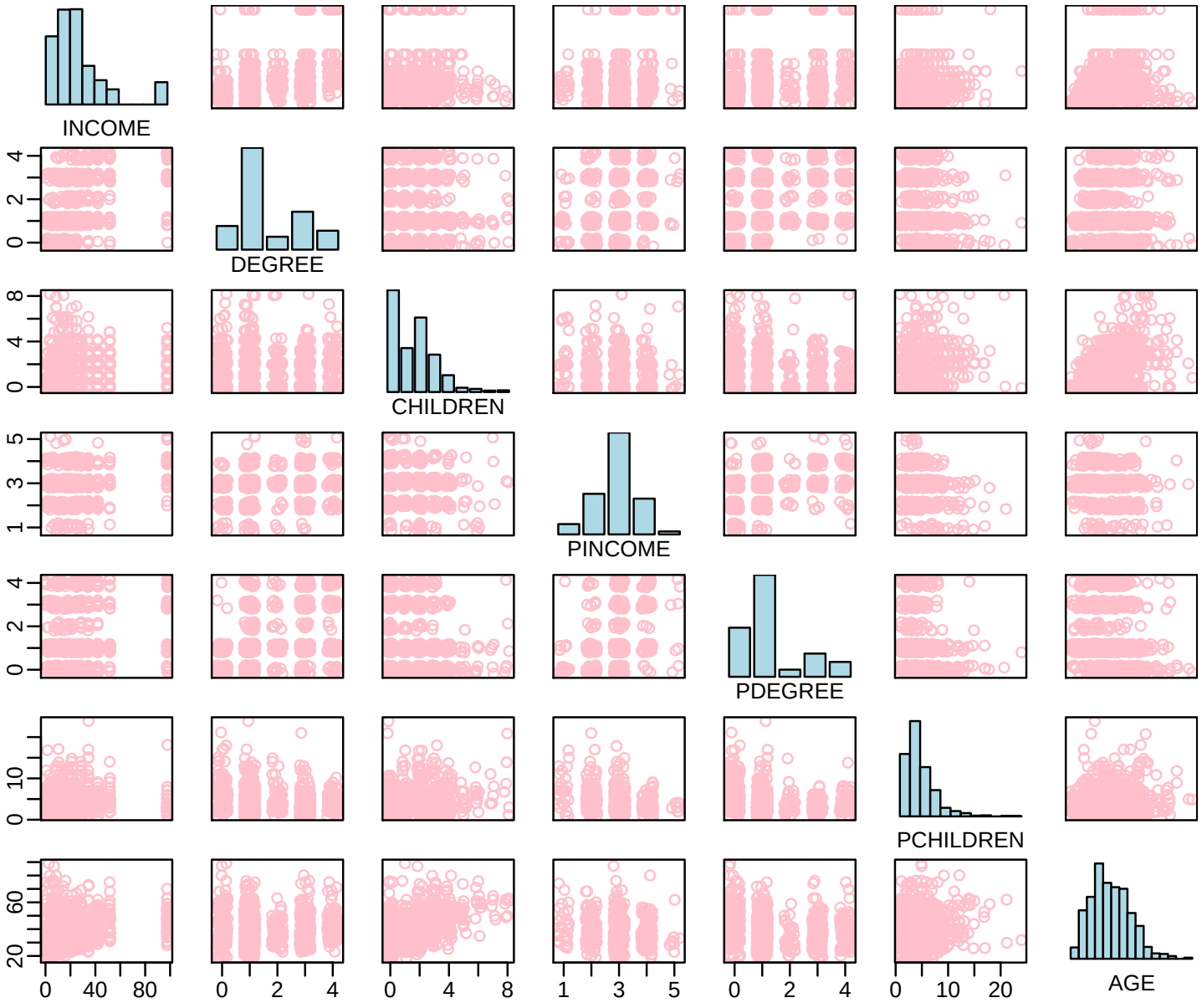
Some research in multivariate analysis
Stat 542

Peter Hoff

University of Washington

Outline

GSS data


INCOME
DEGREE
CHILDREN
PINCOME
PDEGREE
PCHILDREN
AGE
0 2 4
0 4 8
1 3 5
0 10 20
0 40 80

Conditional models

Interest is typically in the conditional relationship between pairs of variables, accounting for heterogeneity in other variables of less interest. Standard bivariate rank-based methods are inappropriate.

Model 1

$$\mathrm{INC}_i = \beta_0 + \beta_1 \mathrm{CHILD}_i + \beta_2 \mathrm{DEG}_i + \beta_3 \mathrm{AGE}_i + \beta_4 \mathrm{PCHILD}_i + \beta_5 \mathrm{PINC}_i + \beta_6 \mathrm{PDEG}_i + \epsilon_i$$

p-value for β_1 is 0.11: "little evidence" that $\beta_1 \neq 0$

Model 2

$$\mathrm{CHILD}_i \sim \mathrm{Pois}(\exp\{\beta_0 + \beta_1 \mathrm{INC}_i + \beta_2 \mathrm{DEG}_i + \beta_3 \mathrm{AGE}_i + \beta_4 \mathrm{PCHILD}_i + \beta_5 \mathrm{PINC}_i + \beta_6 \mathrm{PDEG}_i\})$$

p-value for β_1 is 0.01: "strong evidence" that $\beta_1 \neq 0$.

	Predictor						
Response	INC	CHILD	DEG	AGE	PCHILD	PINC	PDEG
INC	NA	**1.10 (.11)**	7.03 (<.01)	.34 (<.01)	4.07 (<.01)	.28 (.41)	1.40 (.12)
CHILD	**.01 (.01)**	NA	-.07 (.06)	.04 (<.01)	-.06 (.20)	.02 (.08)	-.05 (.20)

Multivariate normal copula model

This idea motivates the following "latent variable" model:

$$\begin{aligned}(z_1, \ldots, z_p) &\sim \text{mvn}(\mathbf{0}, \Sigma) \\ (y_1, \ldots, y_p) &= (F_1^{-1}(z_1), \ldots, F_p^{-1}(z_p))\end{aligned}$$

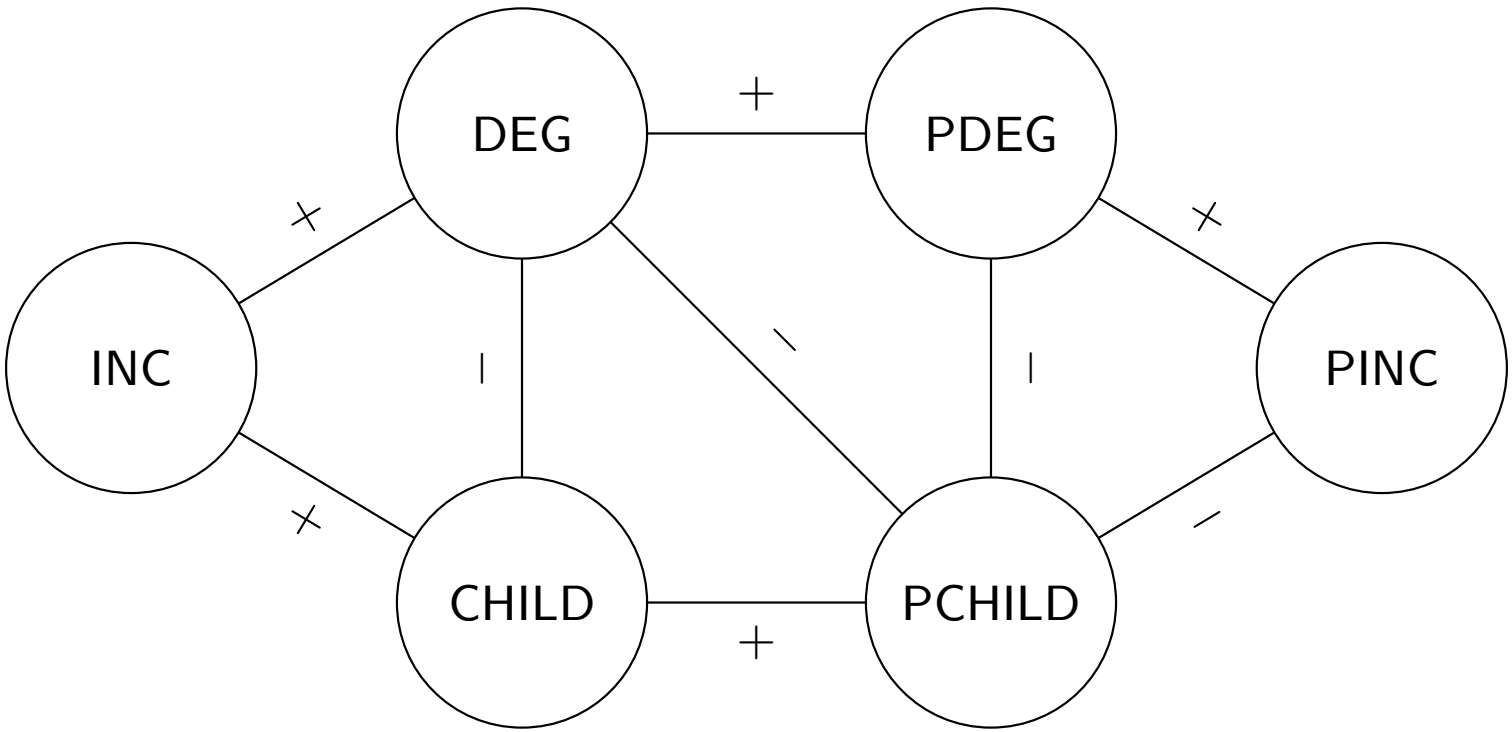
Σ parameterizes the dependence, $F_1^{-1}, \ldots, F_p^{-1}$ the marginal distributions.

This model

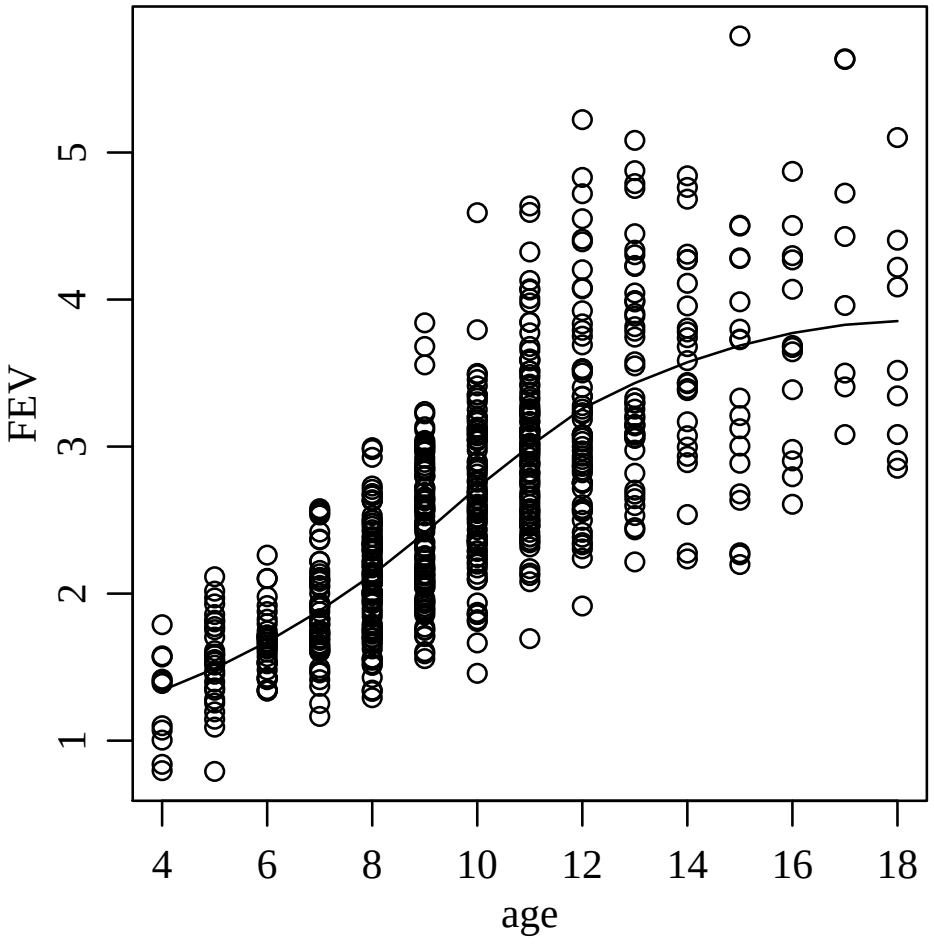
- is scale free
- is appropriate for discrete and continuous data
- gives compatible full conditional distributions

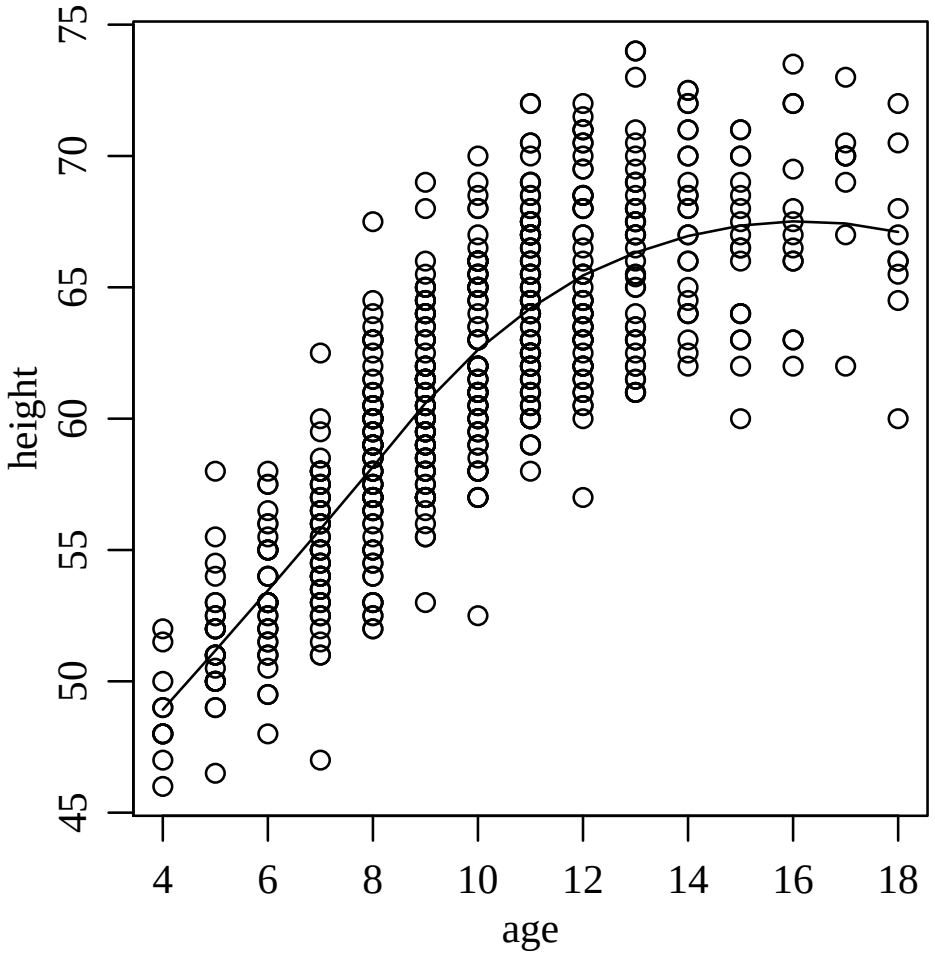
$$\text{E}[z_j | z_{-j}] = \Sigma_{[j,-j]} \Sigma_{[-j,-j]}^{-1} z_{-j}$$

Compatible regression coefficients

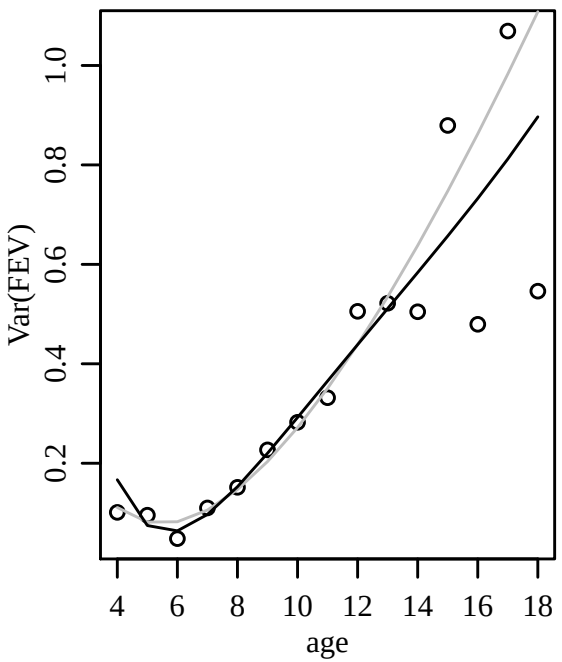


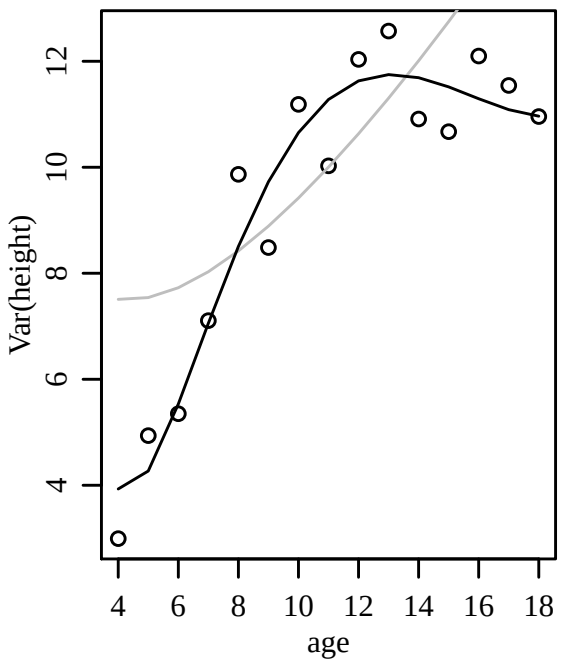
FEV data

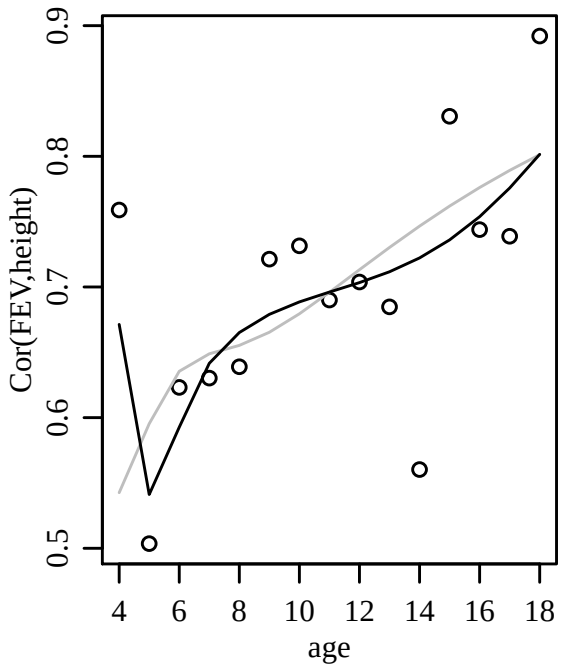

FEV
1
2
3
4
5
4
6
8
10
12
14
16
18
age


height
45
50
55
60
65
70
75
4
6
8
10
12
14
16
18
age

FEV data

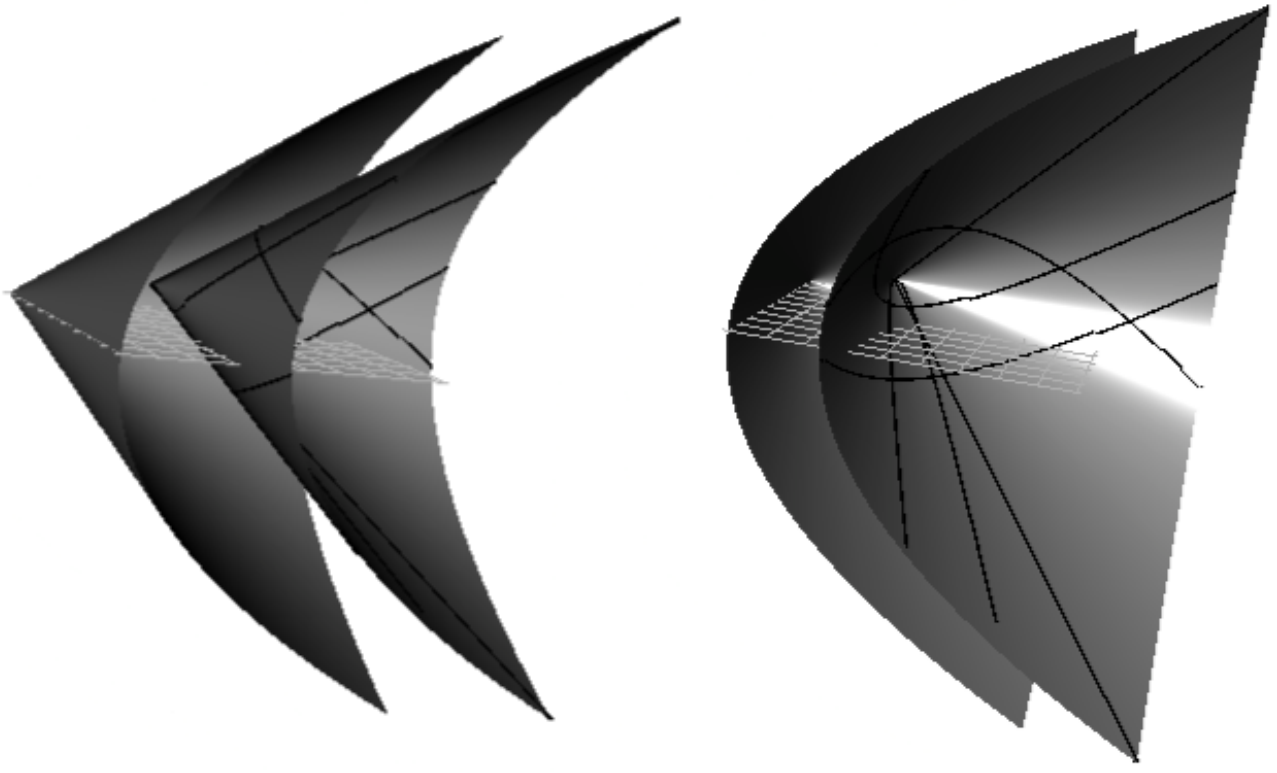

Var(FEV)
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
14
16
18
age


Var(height)
4
6
8
10
12
4
6
8
10
12
14
16
18
age

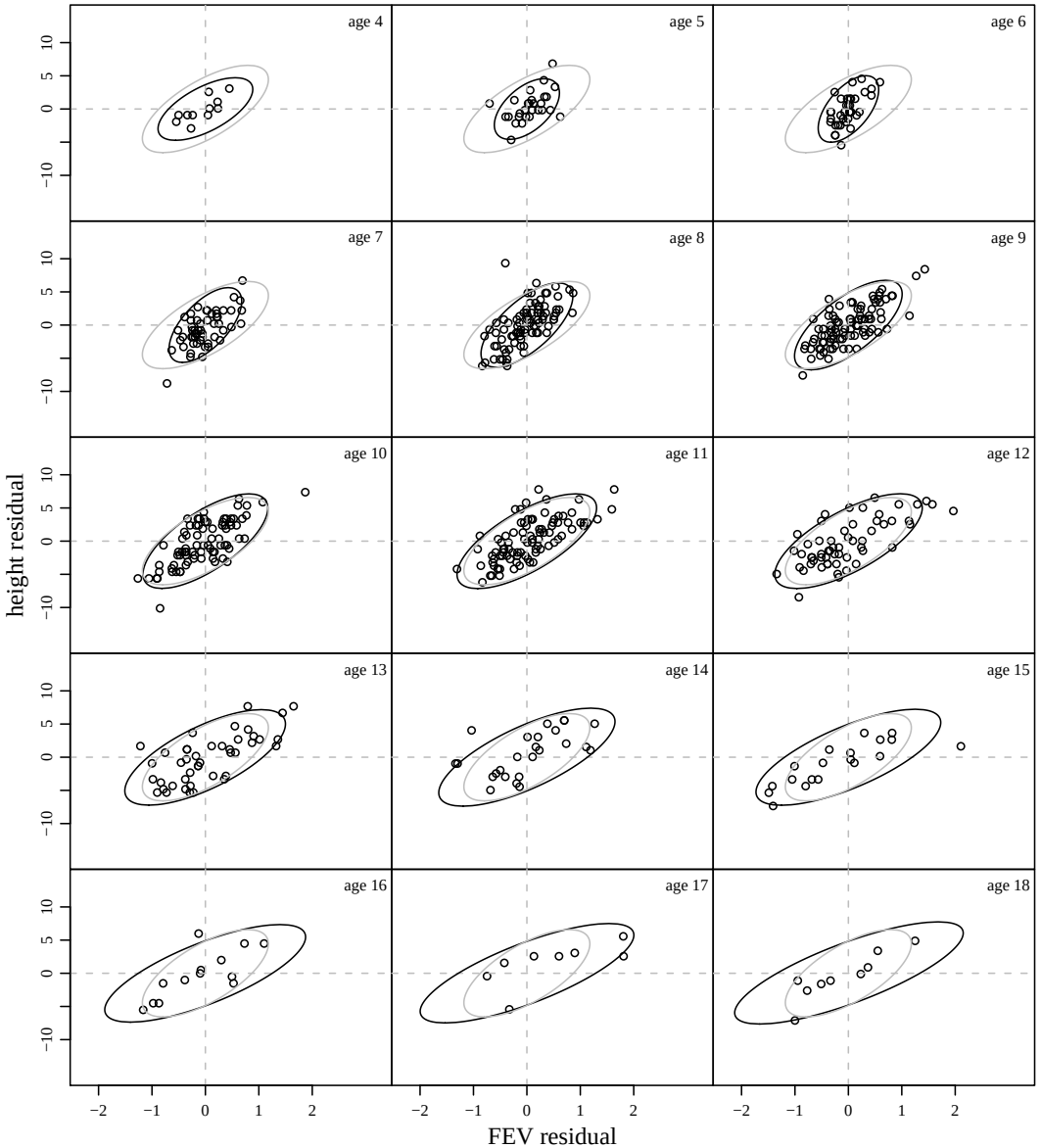

Cor(FEV,height)
0.5
0.6
0.7
0.8
0.9
4
6
8
10
12
14
16
18
age

A covariance regression model

$$\mathbf{\Sigma}_x = \mathbf{A} + \mathbf{B}\mathbf{x}\mathbf{x}^T\mathbf{B}^T \quad (1)$$



FEV data by age

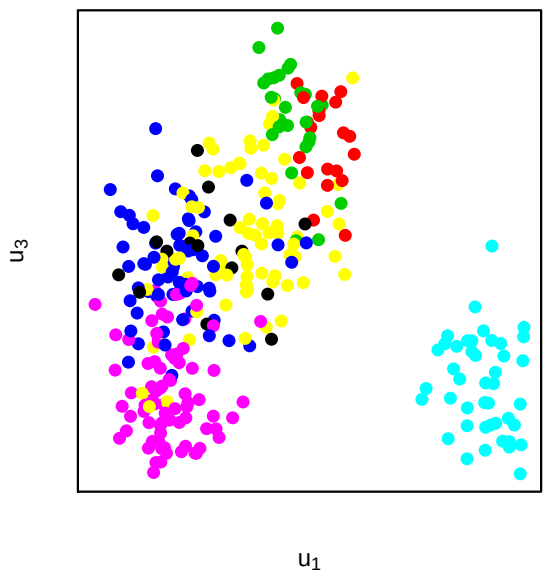


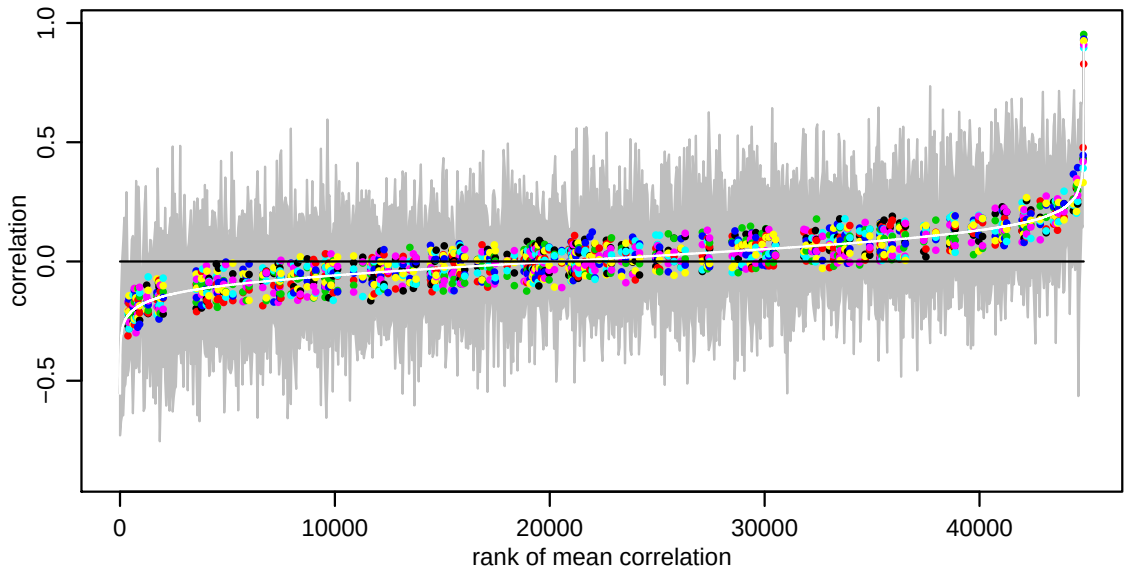
Leukemia data

Gene expression data on 327 cancer patients, each in one of seven groups:

group	BCR	E2A	Hyperdip50	MLL	T	TEL	other
sample size	15	27	64	20	43	79	79

We look at the 300 genes with highest rank variation across subjects.





$\mathbf{Y} = \mathbf{UDV}^T$ Left-singular vectors of $\mathbf{U}$ separate the groups.

$\mathbf{Y}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^T$ How do covariances $\mathbf{V}_k\mathbf{D}_k^2\mathbf{V}_k^T$ differ across groups?

Reduced rank matrix approximation

For high-dimensional data, low rank approximations are useful for describing the main patterns in row/column variability:

Symmetric matrices: $\mathbf{Y} = \mathbf{U}\Lambda\mathbf{U}^T + \mathbf{E}$, $y_{i,j} = \mathbf{u}_i^T \Lambda \mathbf{u}_j + e_{i,j}$

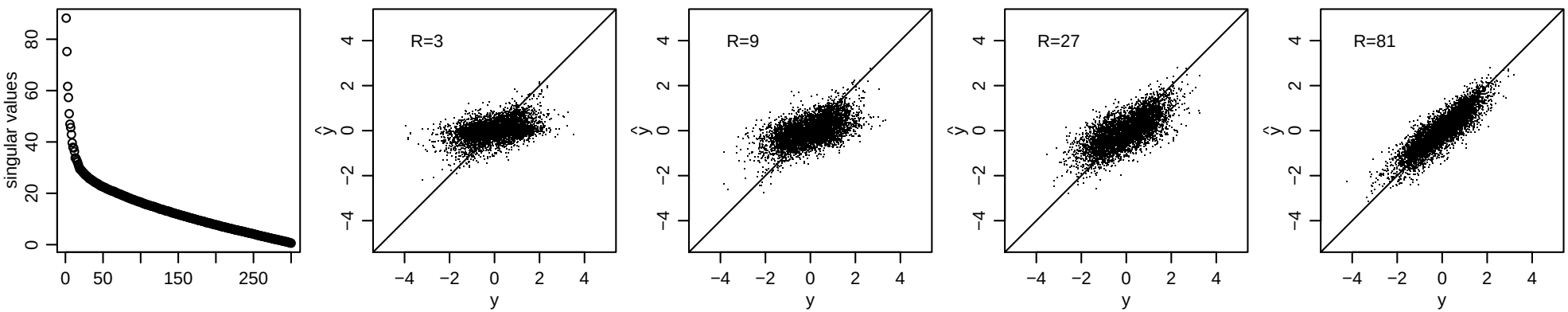
Rectangular matrices: $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^T + \mathbf{E}$, $y_{i,j} = \mathbf{u}_i^T \mathbf{D} \mathbf{v}_j + e_{i,j}$

The column dimension R of $\mathbf{U}$ is generally much smaller than that of $\mathbf{Y}$,

$$R << \min(m, n)$$

so that $\mathbf{U}\Lambda\mathbf{U}^T$, $\mathbf{U}\mathbf{D}\mathbf{V}^T$ provide low-rank approximations to $\mathbf{Y}$.

$$\min_{\mathbf{M}:\mathrm{rank}(\mathbf{M})=R} ||\mathbf{Y} - \mathbf{M}||^2 = ||\mathbf{Y} - \hat{\mathbf{U}}_{[,1:R]}\hat{\mathbf{D}}_{[1:R,1:R]}\hat{\mathbf{V}}^T_{[,1:R]}||^2$$



A hierarchical eigenmodel

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{U}_1\mathbf{D}_1\mathbf{V}_1^T + \mathbf{E}_1 \\ \vdots \quad &\vdots \quad \vdots \\ \mathbf{Y}_K &= \mathbf{U}_K\mathbf{D}_K\mathbf{V}_K^T + \mathbf{E}_K \end{aligned}$$

$$\begin{array}{lll} \mathbf{U}_1 \sim \text{uniform}(\mathcal{V}_{n_1,R}) & \text{diag}(\mathbf{D}_1) \sim \text{normal}(\mathbf{0}, \tau^2 I) & \mathbf{V}_1 \sim \text{Bingham}(\mathbf{A}, \mathbf{B}, \mathbf{V}) \\ & \vdots & \\ \mathbf{U}_K \sim \text{uniform}(\mathcal{V}_{n_K,R}) & \text{diag}(\mathbf{D}_K) \sim \text{normal}(\mathbf{0}, \tau^2 I) & \mathbf{V}_K \sim \text{Bingham}(\mathbf{A}, \mathbf{B}, \mathbf{V}) \end{array}$$

$$\mathbf{V} \sim \text{uniform}(\mathcal{O}_R)$$