General Global Ancova - Model, Algorithm, Examples

Reinhard Meister TFH Berlin

Why testing for global differences in gene expression?

Global tests are useful for

- evaluating treatment effects
- checking signatures used for classification
- comparing results between different studies
- exploring the rôle of pathways and other functional groups

Differential gene expression: different views

Notation: Y: phenotype data X: gene-expression data.

- H_0 : $P(Y|X) = P(Y)$ no expression structure in phenotype
  Goeman et al (2004) globaltest
- H_0' : $P(X|Y) = P(X)$: no phenotype structure in expression
  Mansmann, Meister (2005) GlobalAncova

Both hypotheses are equivalent, the tests derived are not.

General GlobalAncova: the model

Decomposition of the $p \times n$ data matrix $\mathbf{X}$ into a systematic part $\mathcal{M}$ and an error term $\mathcal{E}$: $\mathbf{X} = \mathcal{M} + \mathcal{E}$, assuming $E\mathcal{E} = \mathbf{0}$.
gene-wise view: modeling phenotype structure by $\mathbf{D}$ and $\boldsymbol{\theta}$:

$$\mathcal{M} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \vdots \\ \boldsymbol{\mu}^{(p)} \end{pmatrix} \qquad E \begin{pmatrix} \mathbf{x}^{(1)\,\prime} \\ \mathbf{x}^{(2)\,\prime} \\ \vdots \\ \mathbf{x}^{(p)\,\prime} \end{pmatrix} = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & \dots & \mathbf{0} \\ \vdots & \dots & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \mathbf{D} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \\ \vdots \\ \boldsymbol{\theta}^{(p)} \end{pmatrix}$$

subject wise view: equal covariance structure among genes:

$$Cov \begin{pmatrix} \boldsymbol{\varepsilon}_{(1)} \\ \boldsymbol{\varepsilon}_{(2)} \\ \vdots \\ \boldsymbol{\varepsilon}_{(n)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \dots & \mathbf{0} \\ \vdots & \dots & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \boldsymbol{\Sigma} \end{pmatrix}$$

General GlobalAncova: extra-sum-of-squares principle

Univariate considerations:

Decomposition of a linear model $\mathbf{x} = \mathbf{D}\boldsymbol{\theta} + \mathbf{e}$ into two parts:

$\mathbf{D} = (\mathbf{D}_1, \mathbf{D}_2)$ and $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}$ $\quad \mathbf{x} \in \mathbb{R}^n, \quad \boldsymbol{\theta} \in \mathbb{R}^q, \quad \boldsymbol{\theta}_2 \in \mathbb{R}^f$

full model: $\mathbf{x} = \mathbf{D}_1\boldsymbol{\theta}_1 + \mathbf{D}_2\boldsymbol{\theta}_2 + \boldsymbol{\varepsilon}_{\text{full}}$

reduced model: $\mathbf{x} = \mathbf{D}_1\boldsymbol{\theta}_1 + \boldsymbol{\varepsilon}_{\text{red}}$

Computation sums of squares of residuals: SSR_{full} and SSR_{red}

Test of $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$ using $F = \frac{(SSR_{\text{red}} - SSR_{\text{full}})/f}{SSR_{\text{full}}/(n-q)}$

$F \sim F_{f,n-q}$ holds under H_0 and $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$

General GlobalAncova: global test

Multivariate considerations:

$$F_{\text{global}} = \frac{(1/p)\ \sum_{i=1}^{p}(SSR_{\text{red}}^{(i)} - SSR_{\text{full}}^{(i)})/f}{(1/p)\ \sum_{i=1}^{p} SSR_{\text{full}}^{(i)}/(n-q)}$$

Assumption: homoskedastic uncorrelated normal errors:

$$\boldsymbol{\varepsilon}_{(j)} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \quad j = 1, \ldots, n$$

Distribution: global tests-statistic under H_0:

$$H_0 : \bigcap_{i=1}^{p} \boldsymbol{\theta}_2^{(i)} = \mathbf{0} \quad \Rightarrow \quad F_{\text{global}} \sim F_{p\times f, p\times(n-q)}$$

General GlobalAncova: approximative p-values

Permutation test approach

Basic assumption: residuals from reduced model interchangeable between subjects under H_0. (correlation structure is preserved)

- **loop**

  generate permutation of subject numbers
  recompute statistic $\widetilde{F}$
  using permutation of terms tested in design matrix

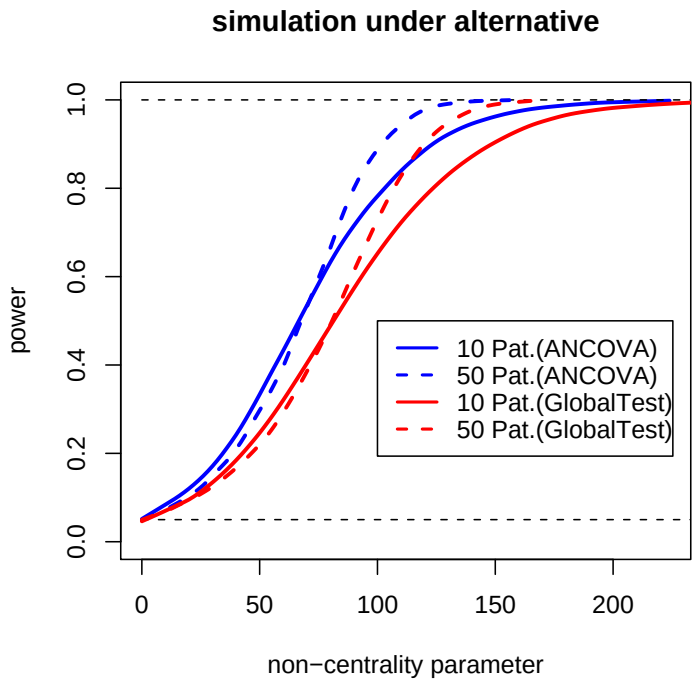
  end loop

- **result** $\widehat{p}_{\text{value}} = \#\{\widetilde{F} \geq F_{\text{orig}}\}/\#\{\text{permutations}\}$

Comparing Power: GlobalAncova vs GlobalTest

Niebank (2004) Results of simulations using `snow` (**Rossini**)

- uncorrelated homoscedastic normal data: theoretical power (noncentral F), Goeman's test and GlobalAncova identical
- correlated data: permutations correct for α-inflation but loss in power
- power simulated using 100 genes containing two blocks of 20 genes, being correlated $\rho = 0.9$ and differentially expressed Advantage GlobalAncova.



General GlobalAncova: algorithm R – Code

```
# Generate Designmatrix
D <- model.matrix(model, phenodata)

# Define projection matrix
hat.matrix <- function(D) { D %*% solve(t(D) %*% D) %*% t(D) }

# Compute projection matrix
H <- hat.matrix(D)
I <- diag(N.subjects)

# Compute residuals
R <- X %*% (I - H)
```

Applications of Generalized Concept

Types of models and symbolic description in R syntax

Design	Model notation
ANOVA, many groups	~ `group`
dose–response	~ `dose`, ~ `group*dose`
time trends	~ `time`, ~ `group*time`
complex phenotypes	~ `type+grade+enzyme`
gene - gene interaction	~ `gene`, ~ `poly(gene,2)`
co–expression	~ `group+gene`
differential co–expression	~ `group*gene`

Tests of specific hypotheses provided by specifying reduced model or collection of single terms.

General GlobalAncova: work in progress

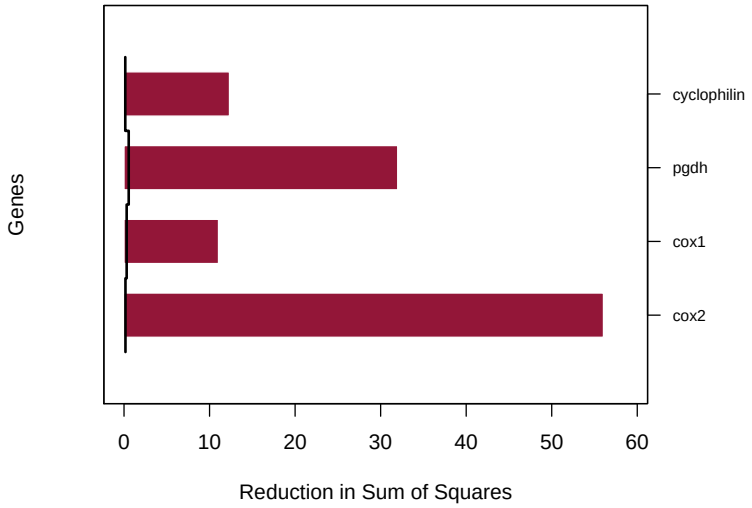
Extensions

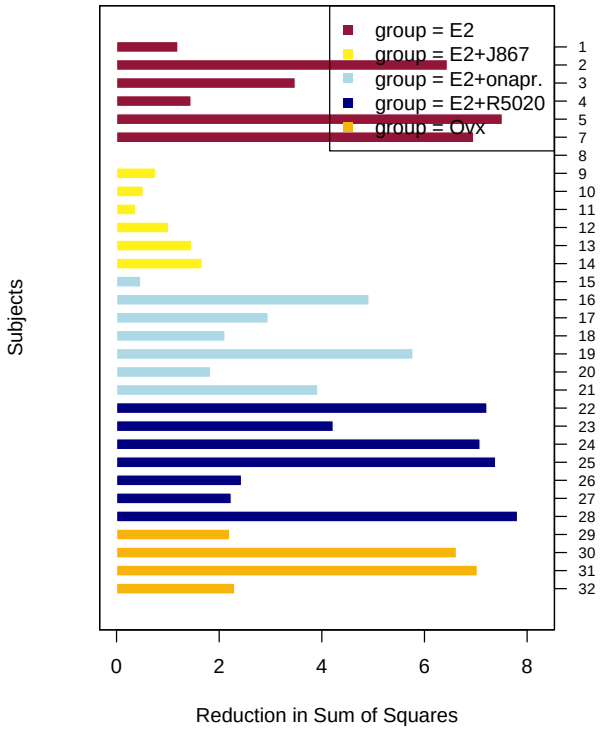
- provide full AOV table (sequential) for all terms in the model
- enhance plots
- include gene-wise covariates $\mathbf{Z}$ with $\dim(\mathbf{Z}) = \dim(\mathbf{X})$
  (e.g. corresponding gene-expressions in normal tissue)
- provide parametric approximation of p-values
  (see Ulrich Mansmann next talk)

Example 1: Treatment effects in guinea pigs

Elger (pers. comm.) progesterone regulation

	SSQ	DF	MS	F.value
Effect	111.4	16	6.96	23.7
Error	30.5	104	0.29	





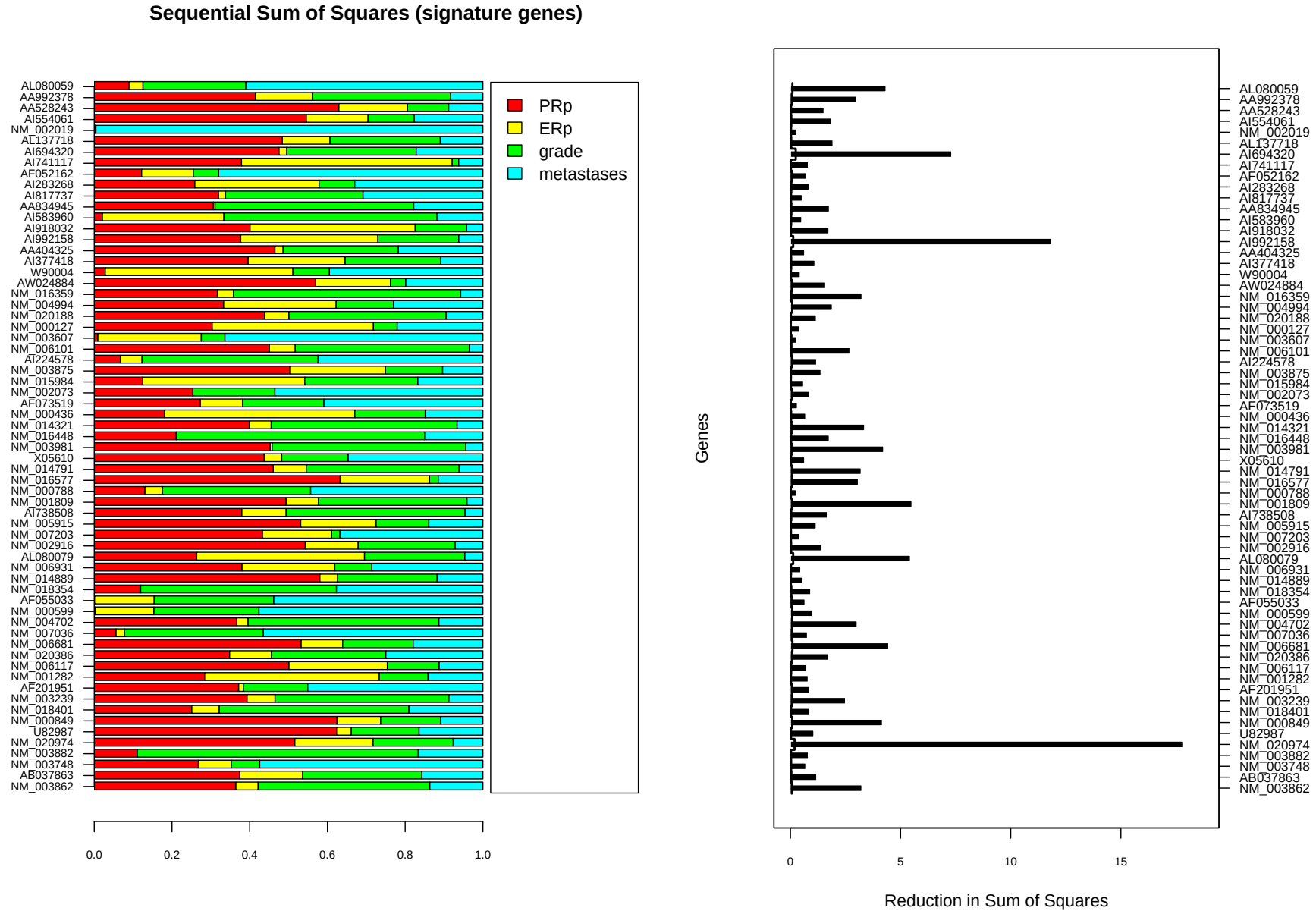
Example 2: Structure in gene signature

van't Veer breast cancer: 65 signature genes

Model formula: $\sim$ `PRp+ERp+grade+metastases`

Sequential decomposition

	SSQ	DF	MS	F
Intercept	56.4	65	0.867	17.46
PRp	56.0	65	0.861	17.34
ERp	21.7	65	0.334	6.72
grade	39.8	130	0.306	6.16
metastases	20.7	65	0.319	6.42
error	290.6	5850	0.050	


Sequential Sum of Squares (signature genes)
PRp
ERp
grade
metastases
0.0
0.2
0.4
0.6
0.8
1.0
Genes
Reduction in Sum of Squares
0
5
10
15

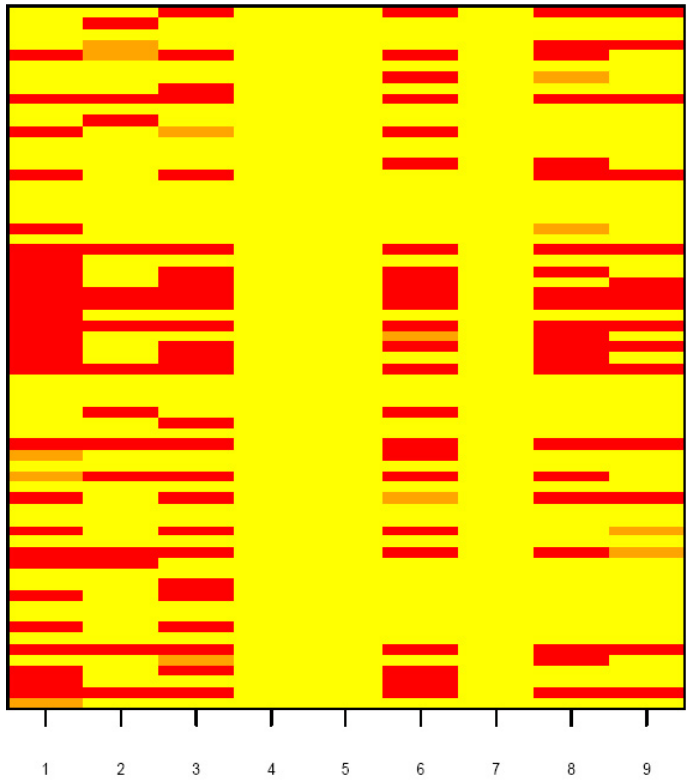
Example 3: Coexpression in pathways

van't Veer breast cancer:
116 patients,65 signature genes, 9 pathways:
1:androgen, 2:apoptosis, 3:cell-cycle, 4:notch-delta, 5:p53, 6:ras, 7:tgf-beta, 8:tightjunction, 9:wnt

`expr~metastases+ER+sign.gene`

$$H_0 : \bigcap_{i \in pathway} \beta^{(i)}_{signature} = 0$$

p-values Holm-adjusted ($p < .01$)



Summary

Basics and Applications of GlobalAncova

- univariate methods are inappropriate for micro-array data
- GlobalAncova provides permutation based inference for multivariate linear modeling of expression data
- extension to even gene-specific covariates possible
- adjustment for covariates a *must* for observational studies
- broad range of possible applications

Thank you
for
your
attention!