— Global Testing —

Ulrich Mansmann, Reinhard Meister, Manuela Hummel

Practical DNA Microarray Analysis, November 2006, Heidelberg
http://compdiag.molgen.mpg.de/ngfn/pma2006nov.shtml

Abstract. *This is the tutorial for the exercises about global testing on day 4 of the course on Practical DNA Microarray Analysis. The global test introduced by Goeman et al. (2004) and the global ANCOVA approach of Mansmann and Meister (2005) are practiced.*

1 Preliminaries

Load data and packages. Have a look on the data. It stems from a study on breast cancer from van t'Veer et al. (2002). There is an expression matrix (`vantVeer`), a data frame giving phenotype information for all samples (`phenodata`) and a list of nine cancer related pathways (`pathways`) each consisting of corresponding probe set names.

```
> library(GlobalAncova)
> library(globaltest)
> data(vantVeer)
> data(phenodata)
> data(pathways)
> dim(vantVeer)
> vantVeer[1:10, 1:10]
> str(phenodata)
> str(pathways)
```

Assume we are interested in differential expression between relevant prognostic groups, defined by the development of distant metastases within five years (`metastases`). Further we have covariate information about the tumor grade (`grade`) and the Estrogen receptor status (`ERstatus`). If we attach the phenotype data frame to the R search path all its variables can be accessed by simply giving their names.

```
> attach(phenodata)
> table(metastases)
```

2 Global Testing of a Single Pathway

We start by applying global tests to all genes in the dataset so that differences in the overall gene-expression pattern can be demonstrated. Here we set the number of permutations to just 100 or 1000 so that executing the examples will not last too long. For getting more reliable results one should recompute the examples with more permutations.

```
> gt.all <- globaltest(X = vantVeer, Y = metastases)
> gt.all
```

```
> ga.all <- GlobalAncova(xx = vantVeer, group = metastases,
+     perm = 100)
> ga.all
```

GlobalAncova may also be called in a more general way by definition of two linear models that shall be compared. Hence model formulas for the full model containing all parameters and the reduced model, where the terms of interest are omitted, have to be given. An alternative is to provide the formula for the full model and a character vector naming the terms of interest. Consequently we could run the same analysis as above with two possible further function calls shown below. In both cases a data frame with information about all variables for each sample is required. (In the case of microarray data this can usually be the corresponding pData object.) Such model definitions will be useful for more complex analysis tasks (see later).

```
> GlobalAncova(xx = vantVeer, formula.full = ~metastases,
+     formula.red = ~1, model.dat = phenodata, perm = 100)
> GlobalAncova(xx = vantVeer, formula.full = ~metastases,
+     test.terms = "metastases", model.dat = phenodata,
+     perm = 100)
```

From the result we conclude that the overall gene expression profile for all 1113 genes is associated with the clinical outcome.

Now we consider a special group of genes, e.g. the p53-signalling pathway. We apply the global test to this pathway using the options genesets and test.genes, respectively.

```
> p53 <- pathways$p53_signalling
> gt.p53 <- globaltest(X = vantVeer, Y = metastases, genesets = p53)
> gt.p53
> ga.p53 <- GlobalAncova(xx = vantVeer, group = metastases,
+     test.genes = p53, perm = 1000)
> ga.p53
```

3 Adjusting for Covariates

The adjustment for covariate information is possible with both methods. For example we can adjust for the Estrogen receptor status.

```
> rownames(phenodata) <- Sample
> gt.adj <- globaltest(X = vantVeer, Y = metastases ~ ERstatus,
+     adjust = phenodata, genesets = p53)
> gt.adj
> ga.adj <- GlobalAncova(xx = vantVeer, group = metastases,
+     covars = ERstatus, test.genes = p53, perm = 1000)
> ga.adj
```

With the more general GlobalAncova function call we would simply adjust the definitions of model formulas, namely formula.full = ~ metastases + ERstatus and formula.red = ~ ERstatus.

4 Testing Several Pathways Simultaneously

The user can apply `globaltest` and `GlobalAncova` to compute p–values for a couple of pathways with one call.

```
> gt.pw <- globaltest(X = vantVeer, Y = metastases, genesets = pathways)
> gt.pw
> ga.pw <- GlobalAncova(xx = vantVeer, group = metastases,
+     test.genes = pathways, perm = 100)
> ga.pw
```

Afterwards a suitable correction for multiple testing has to be applied. Note however that due to the extremely high correlations between these tests, many procedures that correct for multiple testing here are inappropriate. An appropriate method would be for example the Holm correction.

```
> gt.pw.raw <- p.value(gt.pw)
> gt.pw.adj <- p.adjust(gt.pw.raw, "holm")
> gt.result <- cbind(raw = gt.pw.raw, Holm = gt.pw.adj)
> gt.result
> ga.pw.raw <- ga.pw[, "p.perm"]
> ga.pw.adj <- p.adjust(ga.pw.raw, "holm")
> ga.result <- cbind(raw = ga.pw.raw, Holm = ga.pw.adj)
> ga.result
```

5 Analysis of Arbitrary Clinical Variables

With `GlobalAncova` also clinical variables with more than two groups or even continuous ones can be considered. For demonstration we investigate differential expression for the three ordered levels of tumor grade and again only the p53-signalling pathway.

```
> ga.grade <- GlobalAncova(xx = vantVeer, formula.full = ~ordered(grade),
+     formula.red = ~1, model.dat = phenodata, test.genes = p53,
+     perm = 1000)
> ga.grade
```

6 Gene–Gene Interaction

Now we want to go into the matter of other interesting biological questions. For example one might ask if there exists interaction between the expression of special genes (e.g. genes from a prognosis signature) and the expression of genes in a certain pathway. This question can be answered by viewing the expression values of the signature genes as linear regressors and by testing their effects on the expression pattern of the pathway genes. For demonstration we pick the gene "AL137718". Assume that we also want to adjust for the Estrogen receptor status.

```
> signature.gene <- "AL137718"
> model <- data.frame(phenodata, signature.gene = vantVeer[signature.gene,
+     ])
> ga.signature <- GlobalAncova(xx = vantVeer, formula.full = ~signature.gene +
+     ERstatus, formula.red = ~ERstatus, model.dat = model,
```

```
+     test.genes = p53, perm = 1000)
> ga.signature
```

7 Co–Expression

Next we want to analyse co–expression regarding the clinical outcome of building distant metastases within five years. This can be done by simply adding the variable `metastases` to the full and reduced model, respectively. Such layout corresponds to testing the linear effect of the signature gene stratified not only by Estrogen receptor status but also by `metastases`.

```
> ga.coexpr <- GlobalAncova(xx = vantVeer, formula.full = ~metastases +
+     signature.gene + ERstatus, formula.red = ~metastases +
+     ERstatus, model.dat = model, test.genes = p53, perm = 1000)
> ga.coexpr
```

Supposably the most interesting question in this case concerns differential co–expression. Differential co–expression is on hand if the effect of the signature gene behaves different in both `metastases` groups. In a one dimensional context this would become manifest by different slopes of the regression lines. Hence what we have to test is the interaction between `metastases` and `signature.gene`.

```
> ga.diffcoexpr <- GlobalAncova(xx = vantVeer, formula.full = ~metastases *
+     signature.gene + ERstatus, formula.red = ~metastases +
+     signature.gene + ERstatus, model.dat = model, test.genes = p53,
+     perm = 1000)
> ga.diffcoexpr
```

(`globaltest` is not designed for the analysis of such linear models. On the other hand it is able to deal with survival times as clinical outcome.)

8 Plots

The functions `geneplot` and `Plot.genes` visualize the influence of individual genes on the test result while `sampleplot` and `Plot.subjects` visualize the influence of individual samples. If an individual does not fit into the pattern of its phenotype, negative values can occur in the sample plot. A small p–value will therefore generally coincide with many positive bars.

```
> geneplot(gt.p53)
> Plot.genes(xx = vantVeer[p53, ], group = metastases)
> sampleplot(gt.p53)
> Plot.subjects(xx = vantVeer[p53, ], group = metastases,
+     legendpos = "bottomright")
```

With `GlobalAncova` also plots for more general models are available. Additionally a variable (possibly different from the tested one) for defining the coloring can be chosen by the user and (only in the subjects plot) samples can be sorted.

```
> Plot.subjects(xx = vantVeer[p53, ], formula.full = ~ordered(grade),
+     formula.red = ~1, model.dat = phenodata, colorgroup = "grade",
+     sort = TRUE, legendpos = "topleft")
```

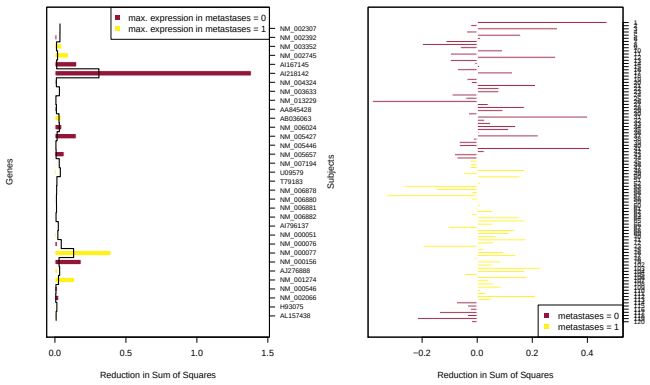

Figure 1: Gene and subjects plot (with `GlobalAncova`).

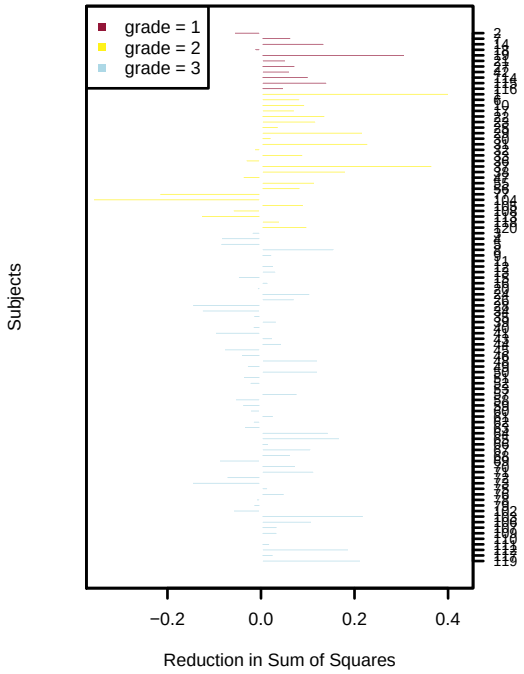

Figure 2: Subjects plot.

The package `globaltest` provides the checkerboard as another diagnostic plot. It shows similarities between samples. For high similarities the respective squares are colored white, for relatively different samples they are colored black.

```
> checkerboard(gt.p53)
```

Figure 3: Checkerboard plot.

9 References

Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 2004; 20(1): 93-9.

Holm S.A simple sequentially rejective multiple test procedure. Scand. J. Statist 1979; 6: 65-70.

Mansmann U and Meister R. Testing differential gene expression in functional groups. Methods Inf Med 2005; 44 (3): 449-53.

van t'Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415: 530-536.