

GlobalAncova with Special Sum of Squares Decompositions

Ramona Scheufele Manuela Hummel Reinhard Meister
Ulrich Mansmann

October 3, 2007

Contents

1	Abstract	1
2	Sequential and Type III Decomposition	2
3	Testing Groups of Genes	4
4	Diagnostic Plots	5
5	Pairwise Comparisons of Factor Levels	7
6	Adjusting for a Global Covariate	8
7	Acknowledgements	10

1 Abstract

This vignette shows the enhancements made for *GlobalAncova*. Basically, there are four ideas implemented:

- decomposition of the sum of squares of a linear model ([2])
- a plotting function for the sequential decomposition
- pairwise comparison for factor levels
- adjustment for global covariates

The decomposition of the model sum of squares results in an ANOVA table, which shows the sum of squares due to each term of a linear model. It can be used either on a global basis or for a small group of genes on a gene-wise basis. Pairwise comparisons allow conclusions about whether or not the gene expressions of two levels of a factor are significantly different. An adjustment for global covariates is possible in cases where not only one but two expression sets per subject exist. If the second set describes the subject's 'normal' status, it can be used to reduce the variance between subjects.

Since a permutation approach is not (yet) implemented as it is in the basic `GlobalAncova` function, the functionalities described in this vignette may be seen like rather descriptive tools. Large p-values indicate that there is no significant effect, whereas small p-values have to be interpreted with caution. Appropriate p-values for testing any linear hypothesis about phenotype effects can be derived using the basic function `GlobalAncova`.

This document was created using R version 2.6.0 and version 3.4.0 of the *GlobalAncova* package.

2 Sequential and Type III Decomposition

Decomposition of sum of squares (SSQ) can be used so as to show the effect of each factor of a linear model on the gene expression. Two types of decomposition of the model SSQ are presented here: the sequential and the type III decomposition. Both are computed using the extra sum of squares principle repeatedly, yet they have different approaches with regard to which factors are to be kept or to be left out. For both decompositions an F-test can be performed. The two methods are implemented in the function `GlobalAncova.decomp` and can be selected by specifying the argument *method*, which can be either one of "sequential" (default), "type3" or "all". The latest yields a list of both. Other arguments to `GlobalAncova.decomp` are *xx* specifying the expression matrix, *formula* describing the model to be decomposed and *model.dat* determining phenotype data. The names or indices of the group of genes to be analysed can be set in the *test.genes* parameter.

Sequential Decomposition

In the sequential or hierarchical decomposition, the sum of squares of each term of the model are calculated by adding to the model term after term and obtaining the increase in model SSQ due the addition. Thus, each term is adjusted for all preceding terms but not for the succeeding ones. Consequently, the order of terms in the model is important as the following examples show. Applying an F-test to the sequential SSQ is only meaningful if a logical or hierarchical order of factors exists (e.g. main effects first followed by interaction effects). Only in this case, it can be tested when the factors start to be insignificant. The following examples are based on the van't Veer breast cancer data set [3], which is included in the *GlobalAncova* package. A subset of the data consisting of the expression values for 96 patients without *BRCA1* or *BRCA2* mutations is available. The dataset (`vantVeer`) is restricted to 1113 genes associated with 9 cancer related pathways that are provided as a list named (`pathways`). The phenotype data of this study (`phenodata`) include the grade of the tumour (`grade`), whether or not metastases were developed (`metastases`) and the estrogen receptor status (`ERstatus`). Giving the assumption that all three factors measured have an effect on gene expression, a possible linear model for the gene expressions could be

```
> library(GlobalAncova)
> data(vantVeer)
> data(phenodata)
> data(pathways)
```

```
> formula <- ~grade + metastases + ERstatus
```

We will now investigate which of the terms of the former model have effects on the overall gene expression.

```
> GlobalAncova.decomp(xx = vantVeer, formula = formula,
+   model.dat = phenodata, method = "sequential")
```

	SSQ	df	MS	F	p
Intercept	443.62541	1113	0.39858527	9.776182	0.000000e+00
grade	194.50524	2226	0.08737881	2.143158	4.095934e-180
metastases	53.97303	1113	0.04849329	1.189405	1.325949e-05
ERstatus	209.86860	1113	0.18856118	4.624878	0.000000e+00
error	4129.41514	101283	0.04077106	NA	NA

Apparently, all factors of the model are significant. Note however, that since expressions are correlated and have non-normal distribution, the p-values derived from the F-distribution lead to alpha-inflation.

In order to demonstrate the impact of the order of terms we will repeat the analysis using the same model but in reverse order.

```
> formula2 <- ~ERstatus + metastases + grade
> GlobalAncova.decomp(xx = vantVeer, formula = formula2,
+   model.dat = phenodata, method = "sequential")
```

	SSQ	df	MS	F	p
Intercept	443.62541	1113	0.39858527	9.776182	0.000000e+00
ERstatus	269.99053	1113	0.24257909	5.949786	0.000000e+00
metastases	63.67287	1113	0.05720833	1.403160	1.822444e-17
grade	124.68347	2226	0.05601234	1.373826	1.131387e-28
error	4129.41514	101283	0.04077106	NA	NA

Differences in sum of squares due to different ordering of terms indicate a non-orthogonal design matrix. One major property of the sequential method is that the sums of squares due to the single terms add up to the full model SSQ. This is not generally the case in the type III decomposition.

Type III Decomposition

In contrast to the sequential decomposition, the type III decomposition is calculated by removing only a single term at a time from the full model. That is, every term is adjusted for every other term in the model and is thus treated as if ordered last. To select the type III sum of squares the option *method = "type3"* has to be set. The terms of the model derived from the van't Veer dataset can now be tested by using the following call

```
> GlobalAncova.decomp(xx = vantVeer, formula = formula,
+   model.dat = phenodata, method = "type3")
```

	SSQ	df	MS	F	p
Intercept	186.06742	1113	0.16717648	4.100371	0.000000e+00
grade	124.68347	2226	0.05601234	1.373826	1.131387e-28
metastases	47.70751	1113	0.04286389	1.051331	1.154727e-01
ERstatus	209.86860	1113	0.18856118	4.624878	0.000000e+00
error	4129.41514	101283	0.04077106	NA	NA

The SSQ due to the last term is the same in both decompositions. All terms but the last have generally smaller type III than sequential sums of squares. Only in case of an orthogonal design matrix they have equal sums of squares in both decompositions. This examples demonstrates how important the adjustment for more than one phenotype-characterizing covariate can be. Besides the occurrence of metastases, grade and estrogen receptor status contribute substantially to the observed differential gene expressions.

3 Testing Groups of Genes

In some studies, the interest lies not with the entire expression matrix but only with certain groups of genes, e.g. pathways. The names or indices of these genes can be specified in the option *test.genes*. As example, we use the first three cancer relevant pathways given in the object *pathways*.

```
> GlobalAncova.decomp(xx = vantVeer, formula = formula,
+   model.dat = phenodata, method = "type3", test.genes = pathways[1:3])
```

\$androgen_receptor_signaling

	SSQ	df	MS	F	p
Intercept	31.070643	72	0.43153671	12.4661789	2.060236e-132
grade	9.448638	144	0.06561554	1.8954937	9.546578e-10
metastases	1.568993	72	0.02179157	0.6295123	9.939806e-01
ERstatus	29.125115	72	0.40451549	11.6855933	1.671092e-122
error	226.807952	6552	0.03461660	NA	NA

\$apoptosis

	SSQ	df	MS	F	p
Intercept	33.073033	187	0.17686114	5.350719	3.067227e-107
grade	16.433991	374	0.04394115	1.329386	2.488033e-05
metastases	6.267846	187	0.03351789	1.014043	4.331773e-01
ERstatus	45.666917	187	0.24420811	7.388220	1.799784e-172
error	562.475084	17017	0.03305372	NA	NA

\$cell_cycle_control

	SSQ	df	MS	F	p
Intercept	4.173999	31	0.13464512	3.268404	3.399027e-09
grade	8.024304	62	0.12942426	3.141672	5.765491e-15
metastases	1.397575	31	0.04508307	1.094356	3.295908e-01
ERstatus	10.863558	31	0.35043735	8.506590	9.911064e-37
error	116.213873	2821	0.04119598	NA	NA

Estrogen receptor status and tumour grade are significant in all three pathways, whereas the indicator of the development of metastases is not significant in any of them.

Gene-wise Analysis

For a more detailed analysis, it is furthermore possible to display the sequential decomposition for each gene separately. Due to the large numerical output, this option will only be interesting for a limited number of genes. It can be chosen

by setting `genewise=TRUE`. This option also serves the purpose of providing a numerical view of what is shown by the function `Plot.sequential`, see section 4. Just for demonstration we will show the gene-wise result of the sequential decomposition for the first three genes of the first pathway.

```
> GlobalAncova.decomp(xx = vantVeer, formula = formula,
+   model.dat = phenodata, test.genes = pathways[[1]][1:3],
+   genewise = TRUE)
```

\$terms

```
[1] "Intercept" "grade" "metastases" "ERstatus" "error"
```

\$SSQ

	Intercept	grade	metastases	ERstatus	error
AW025529	0.0047040	0.12624647	0.0639334130	0.19956603	2.116086
NM_001648	0.1137127	0.04911051	0.0016380351	0.03591376	2.068737
NM_001753	0.1944000	0.47020168	0.0001075029	0.07604684	4.033718
all	0.3128167	0.64555866	0.0656789509	0.31152663	8.218541

\$df

	Intercept	grade	metastases	ERstatus	error
gene	1	2	1	1	91
all	3	6	3	3	273

\$MS

	Intercept	grade	metastases	ERstatus	error
AW025529	0.0047040	0.06312324	0.0639334130	0.19956603	0.02325369
NM_001648	0.1137127	0.02455526	0.0016380351	0.03591376	0.02273337
NM_001753	0.1944000	0.23510084	0.0001075029	0.07604684	0.04432657
all	0.1042722	0.10759311	0.0218929836	0.10384221	0.03010455

\$F

	Intercept	grade	metastases	ERstatus
AW025529	0.2022904	2.714547	2.749387475	8.582122
NM_001648	5.0020145	1.080141	0.072054202	1.579781
NM_001753	4.3856313	5.303835	0.002425247	1.715604
all	3.4636703	3.573982	0.727231813	3.449386

\$p

	Intercept	grade	metastases	ERstatus
AW025529	0.7309261	0.4132602	0.3922692	0.2555246
NM_001648	0.2676727	0.5930650	0.9349027	0.5434297
NM_001753	0.2836113	0.3065339	0.9975835	0.5258760
all	0.3138884	0.3665289	0.6382954	0.3903414

4 Diagnostic Plots

In this section a graphical method will be introduced to visualize the result of genewise sequential decomposition. This graphic can either be plotted on its own or in combination with the function `Plot.genes`, see *GlobalAncova* vignette.

The function `Plot.sequential` yields a bar plot with bars for the single genes and one for the over all result. The segments of the bars indicate the extra sum of squares due to each factor relative to the model SSQ of the corresponding gene. In order to enable an easier comparison between the genes, the model SSQ of each gene is set to 1. The arguments to this function are the expression matrix *xx*, the *formula* of the model to be decomposed, the phenotype data *model.dat* and optionally a vector of gene names or indices *test.genes* specifying the gene set and the name of the gene set *name.geneset* (for the title of the plot). As an example we will investigate the structure of the cell cycle pathway genes from the van't Veer study, see figure 1.

```
> Plot.sequential(vantVeer, formula = ~ERstatus + metastases +
+ grade, model.dat = phenodata, test.genes = pathways[[3]],
+ name.geneset = "cell cycle pathway")
```

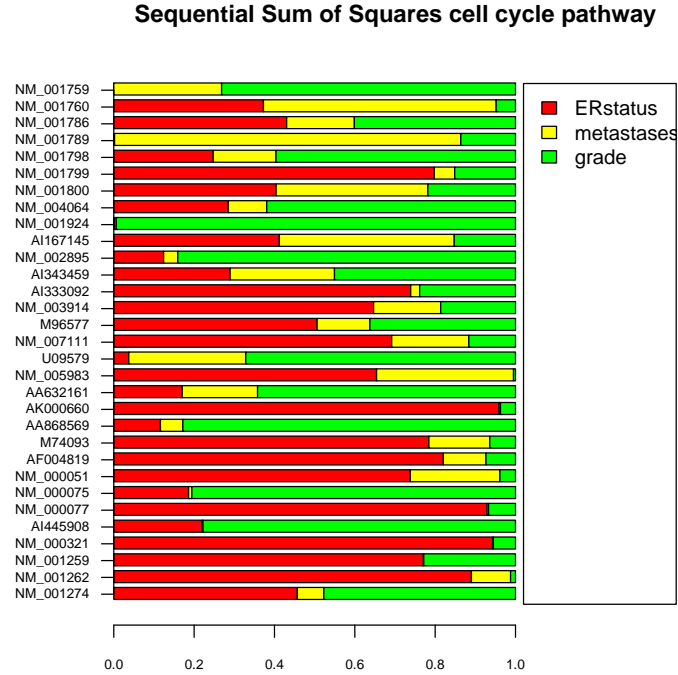


Figure 1: Sequential decomposition of model sum of squares for the cell cycle genes in the van't Veer data.

In order to gain information about the significance of the factors, it is also possible to combine the former plot with the *GlobalAncova* gene plot. The gene plot shows for two given models the gene-wise extra sum of squares and the mean square error in a barplot. In the combined plot, the gene plot is used to show the extra sum of squares due to all factors of the model altogether. Hence, not only the decomposition can be visualized but also the significance of the full

model for each gene. For displaying again the effects on the cell cycle genes in the van't Veer data with the combined plot (figure 2) we use

```
> Plot.all(vantVeer, formula = ~ERstatus + metastases +
+         grade, model.dat = phenodata, test.genes = pathways[[3]],
+         name.geneset = "cell cycle pathway")
```

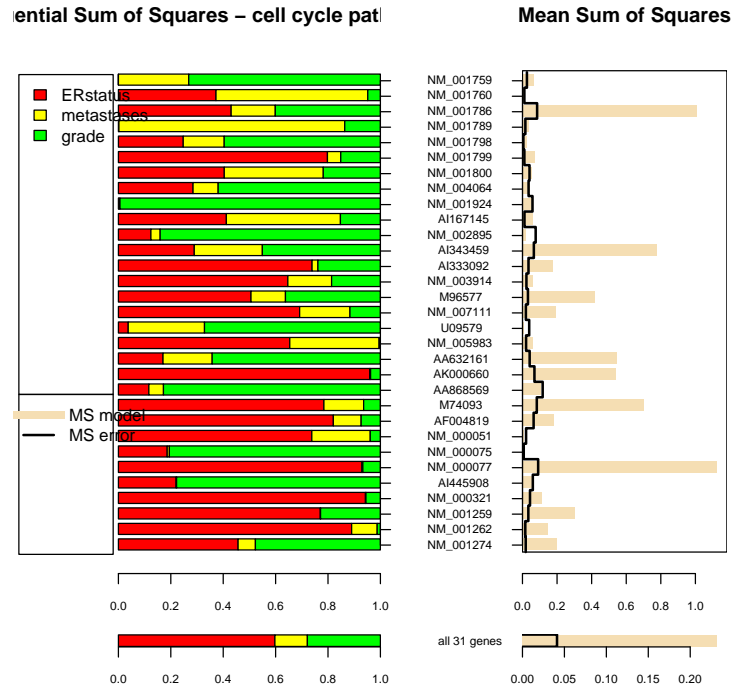


Figure 2: Sequential decomposition of model sum of squares and model mean squares for the cell cycle genes in the van't Veer data.

From figure 2 we can conclude, that effects of the model terms vary from gene to gene. **ERstatus** seems to play a major role, however one has to remember, that a sequential decomposition is displayed and **ERstatus** was ordered first. This question could be answered by using a different order for the model formula. Most importantly, the graph on the left can be used for understanding the significance of the model displayed on the right hand graph. This graphical method helps to investigate a lot of different and complex models, e.g. differential co-expression.

5 Pairwise Comparisons of Factor Levels

Some studies focus on the connection between a single factor and gene expression. This is, for instance, the case when the factor in question provides some kind of classification. To investigate how a classification is reflected in the gene

expression, pair-wise comparisons of factor levels can be used. The method presented here analyses for a pair of factor levels whether the genes under observation are differently expressed, i.e. it can detect whether the difference between two levels is reflected in the gene expression. As measurement for the extra sum of squares due to this difference, the method uses the decrease in residual sum of squares obtained when the two are set to the same level. Pairwise comparisons can be performed by using the function `pair.compare`. Similarly to the function `GlobalAncova.decomp`, `pair.compare` requires the arguments `xx` (expression matrix), `formula` (model) and `model.dat` (phenotype data). Furthermore, the argument `group` has to be set to the name of the factor which is to be analysed. The function yields an ANOVA table with a row for each possible combination of the factor levels. Since pair-wise comparison is a multiple testing problem, not only the p-values based on the permutation test are returned but also the Bonferroni-Holm adjusted p-values.

As an example we will perform pair-wise comparisons of the three levels of tumour grade in the van't Veer data.

```
> pair.compare(xx = vantVeer, formula = ~grade, model.dat = phenodata,
+             group = "grade", perm = 100)
```

	SSQ	df	MS	F	p.perm
1 : 2	43.80796	1113	0.03936025	0.9273622	0.56
1 : 3	103.10757	1113	0.09263932	2.1826641	0.00
2 : 3	114.99741	1113	0.10332202	2.4343578	0.00
error	4393.25676	103509	0.04244323	NA	NA

There are essential differences between stages 1 and 3 and 2 and 3, respectively, whereas stages 1 and 2 seem to be rather similar.

6 Adjusting for a Global Covariate

Considerable variability between patients is the rule in medical statistics. Therefore, adjustment for covariates is often applied so as to reduce variance. Natural candidates for adjustment are covariates representing the baseline status of patients. For microarray data, gene expressions from normal (non-cancer) tissue of a patient might play this role of giving baseline status, which can then be compared to the tumour probe of the same patient. Analogously, tumour probes of a former biopsy might be used to trace changes.

The adjustment for covariates is also implemented in the `GlobalAncova.decomp` call. It uses the arguments `zz` and `zz.per.gene`. The former defines the expression matrix of the global covariate, and the latter is a logical value specifying whether or not different parameters should be used for each gene.

The data for the following examples are taken from the colon cancer study of [1]. In this study, tumour as well as normal tissue expressions were measured. Just for demonstration we picked a set of 1747 genes which are associated with cell proliferation and which are measured in tumour and normal tissue of 12 colon cancer patients. The most important factors considered are the carcinoma stage (`UICC.stage`), the gender (`sex`) and the location of the tumour (distal/proximal). At first a type III decomposition will be performed only for the tumour data in order to show potential differences made by the later adjustment:


```

> data(colon.tumour)
> data(colon.normal)
> data(colon.pheno)
> formula <- ~UICC.stage + sex + location
> GlobalAncova.decomp(xx = colon.tumour, formula = formula,
+   model.dat = colon.pheno, method = "type3")

```

	SSQ	df	MS	F	p
Intercept	742527.9761	1747	425.0303241	2345.1917477	0.00000000
UICC.stage	242.3041	1747	0.1386973	0.7652905	1.00000000
sex	269.6725	1747	0.1543632	0.8517306	0.99999380
location	339.0278	1747	0.1940628	1.0707814	0.02669185
error	2532.9374	13976	0.1812348	NA	NA

In this analysis neither UICC stage nor sex seem to have an influence on gene expression. We will see whether this conclusion remains true after adjustment. In order to adjust for the expression in normal state the of cells, intuitively we could use the difference between tumour and normal tissue expression for analysis:

```

> GlobalAncova.decomp(xx = colon.tumour - colon.normal,
+   formula = formula, model.dat = colon.pheno, method = "type3")

```

	SSQ	df	MS	F	p
Intercept	2062.5365	1747	1.1806162	4.5956873	0.00000000
UICC.stage	459.3349	1747	0.2629278	1.0234775	0.25531679
sex	428.9081	1747	0.2455112	0.9556814	0.89368340
location	482.5561	1747	0.2762199	1.0752183	0.02022423
error	3590.3861	13976	0.2568965	NA	NA

In this example the adjustment does not yield a significant difference to the previous result. Note however, that for the UICC status a considerable reduction in p-value is achieved (see *type3* results).

Adjustment by using one parameter for all genes

A more general way is to adjust by $xx - \beta \cdot zz$ allowing β to be different from 1. The parameter β is estimated as the least square estimate of the model $xx = \beta \cdot zz$. Thus, we account for a relatively different ground level in the two expression sets. This kind of adjustment can be chosen by specifying *zz* as the gene expression matrix of the global covariate:

```

> GlobalAncova.decomp(xx = colon.tumour, formula = formula,
+   model.dat = colon.pheno, method = "all", zz = colon.normal)

```

\$adjustment

	ssq	df
adjustment	2017200	1

\$sequential

	SSQ	df	MS	F	p
Intercept	4779.2569	1747	2.7356937	10.6058621	0.00000000

UICC.stage	374.1633	1747	0.2141747	0.8303224	0.99999978
sex	431.1914	1747	0.2468182	0.9568760	0.88723374
location	484.2787	1747	0.2772059	1.0746844	0.02092398
error	3604.7348	13975	0.2579417	NA	NA

\$typeIII

	SSQ	df	MS	F	p
Intercept	2026.3657	1747	1.1599117	4.4967985	0.00000000
UICC.stage	461.6535	1747	0.2642550	1.0244760	0.24661362
sex	431.2039	1747	0.2468253	0.9569038	0.88708056
location	484.2787	1747	0.2772059	1.0746844	0.02092398
error	3604.7348	13975	0.2579417	NA	NA

Adjustment by using a different parameter for every gene

In the last approach using the model $xx = zz\beta$ for adjustment, the parameter vector β is estimated on a gene-wise basis. This adjustment can be selected by additionally setting the option `zz.per.gene = TRUE`. However, the biological meaning of this model seems unclear.

```
> GlobalAncova.decomp(xx = colon.tumour, formula = formula,
+   model.dat = colon.pheno, method = "all", zz = colon.normal,
+   zz.per.gene = TRUE)
```

\$adjustment

	ssq	df
adjustment	2021963	1747

\$sequential

	SSQ	df	MS	F	p
Intercept	5.552879	1747	0.003178522	0.01076723	1.00000000
UICC.stage	370.891238	1747	0.212301796	0.71917107	1.00000000
sex	437.010911	1747	0.250149348	0.84737942	0.9999962
location	486.723097	1747	0.278605093	0.94377309	0.9428552
error	3610.043251	12229	0.295203471	NA	NA

\$typeIII

	SSQ	df	MS	F	p
Intercept	222.9053	1747	0.1275932	0.4322213	1.00000000
UICC.stage	462.8026	1747	0.2649128	0.8973905	0.9983780
sex	436.9670	1747	0.2501242	0.8472943	0.9999962
location	486.7231	1747	0.2786051	0.9437731	0.9428552
error	3610.0433	12229	0.2952035	NA	NA

7 Acknowledgements

This work was supported by the NGFN project 01 GR 0459, BMBF, Germany.

References

- [1] J. Groene, U. Mansmann, R. Meister, E. Staub, S. Roepcke, M. Heinze, I. Klamann, T. Brummendorf, K. Hermann, C. Loddenkemper, C. Pilarsky, B. Mann, H.P. Adams, H.J. Buhr, and A. Rosenthal. Transcriptional census of 36 microdissected colorectal cancers yields a gene signature to distinguish uicc ii and iii. *Int J Cancer*, pages 1829–1836, 2006.
- [2] S.R. Searle. *Linear Models*. Wiley, 1971.
- [3] L. J. van't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.