kimod A K-tables approach to integrate multiple Omics-Data in R

M L Zingaretti[1], J A Demey Zambrano[2], J L Vicente Villardón[2], and J R Demey[3]

[1]IAPCBA-IAPCH, Universidad Nacional de Villa María
[3]Departamento de Estadística, Universidad de Salamanca
[3]Fellow Prometeo Senescyt, Escuela Superior Politécnica del Litoral (ESPOL)

October 30, 2017

Abstract

kimod is to do multivariate data analysis of k-tables, in particular it makes STATIS methodology, designed to handle multiple data tables that quantity sets of variables collected on the same observations. This package allows to work with mixed data, with the introduction of the following improvements: distance options (for numeric and/or categorical variables) for each of the tables, bootstrap resampling techniques on the residual matrix of STATIS- compromise, that enable perform confidence ellipses for the projection of observations, and regressions Biplot to project all variables on the compromise matrix. In this way, goodness of fit criteria are used for variables selection and building relationships between observations and variables. Moreover, this allows generating clustering of variables which are powerfully related to each other and consequently get the same information. Since the main purpose of the package is to use these techniques to omic data analysis, it includes an example data from four different microarray platforms of the NCI-60 cell lines.

1 Introduction

In the last years, the data of microarrays has not only gained a great importance but also it is availability for the public has increase. The "omics" technologies allow quantitative knowledge of hundreds of biological data of complex nature and have enabled the opportunity of study simultaneously, based on multiple datasets, the expression levels of thousands of genes over the effects of certain treatments or diseases. However, the joint analysis of the different subspaces that generate these technologies and their relations is not simple. Several statistical methods have been developed to handle these problems and to calculate a consensus from data matrices. STATIS-ACT (des Plantes, 1976),(Escoufier et al., 1976) is one of the families of methods that are concerned with analysis of data arising from several configurations and is a powerful technique to

compare subspaces. The aim of this package is to combine STATIS, Biplot (Gabriel, 1971), (Demey et al., 2008) and Cluster methodologies to study the relationships between genes expressions of multiple omics datasets measuring the same biological samples or the expression of the same genes over different experimental conditions.

2 STATIS methodology

The STATIS methodology is an family of exploratory technique of multivariate data analysis based on linear algebra and especially Euclidean vector spaces (ACT stands for Analyse Conjointe de Tableaux, STATIS stands for Structuration des Tableaux A Trois Indices de la Statistique). It has been devised for multiway data situations on the basic idea of computing Euclidean distances between configurations of points (Escoufier, 1973).

In studies of genetic diversity the STATIS is a technique that it allows determine contribution of each observation to the Euclidean distance between the subspaces defined by the molecular markers and morphological traits.

Formally, the central idea of the technique is to compare configurations of the same observations obtained in different circumstances. Thus we need to introduce a measure of similarity between two configurations. This is equivalent to define a distance between the corresponding scalar product matrices. These matrices are:

$$W_k = XX^T \tag{1}$$

We can use the classic Euclidean norm

$$\|W_1 - W_2\|^2 = \sum_k \sum_{k^T} [(W_1 - W_2)_{kk^T}]^2 = Tr[(W_1 - W_2)^2] \tag{2}$$

On some cases when the variables are not all continuous, the scalar product can not compute. DISTATIS approach, we compute K distance matrices instead of Scalar Product (See 1) between observations, further we transform these matrices into cross-product matrices and then use the cross-product approach to STATIS (See (Abdi et al., 2007), (Abdi et al., 2012)). In these works, the autors only proposed the euclidean metrics, however in this package, we extend this approach and incorporating different metrics, extending the use of STATIS-ACT to other types of variables. Three aspects are considered in the application of the method, the study of Interstructure, the boundary of the Compromise space and the Graphical representation of the trajectories.

2.1 STEPS of STATIS

1. Interstructure: Define a distance between $W_k's$ configurations matrix and generate a matrix of scalar product W_{kxk}, later, use the spectral decompositon of W to projection of all studies in a space of low dimension.

2. Compromise: Define a matrix W_{nxn} that $\sum_{k=1}^{n} \alpha_k W_k$ with the property that is the linear combination of the $W_k's$ the most related to each W_k. Finally, use the singular value decomposition for plotting all observations on consensus espace.

3. Trajectories: These gives a idea of the importance and the direction of the change of position of all observations between the stages k and k'.

2.2 Sampling Variability and Biplot Analysis

Following (Demey, 2008), the results of any data analysis are not thorough if they do not offer information about the stability of the solution that show whether the structure detected by the analysis is not random. There are several ways to accomplish this purpose, including the introduction of small perturbations in the data, resampling techniques or applying permutations.

As for other sorting techniques, the sensitivity study of solutions in methods K- tables has hardly received any attention. Therefore, as part of this work, intents to study the stability sample of the average projections of individuals/variables or individuals-variables on parent commitment of the various methods.

Specifically, the use of bootstrap (Efron and Tibshirani, 1993) is proposes for the building confidence regions on the projection of the individual on the compromise matrix (W).

In order to acquire the sampling variability, B configurations must be generated of matrix W, for an algorithm, which on the matrix of residuals is used as detailed below.

The eigen-decomposition of W matrix is:

$$\hat{W} = U_q D_q V_q \tag{3}$$

The objective is then to find a configuration P in a lower dimensional Euclidean space. A lower dimensional approximation can be obtained projecting using the equation 3 (usually $q = 2$). (W), can be break down as $W = \hat{W} + \epsilon$, making ϵ a matrix of residual with the same properties as W and $\hat{W}$, it is the low range estimation ($q < r$) of W. Resampling B times on $n(n-1)/2$ different elements outside the diagonal matrix ϵ, B replicates are generated so $W_i^* = \hat{W} + \epsilon_i^*, i = 1..B$. Using the new matrices W_i^* , it is possible to generate (from eigen-decomposition) new B matrices P_i^* that can compared to the original configuration (P) and create the desired sampling variability.

2.3 Using Biplot to project variables on compromise

In general, from the use of methods STATIS is only possible to show in a graph of individual or variables and no individual and variables. Often, researchers do not only want to know the relationships established between observations, but also between these and variables involved or between the variables themselves. In case of omics data where K-tables can be different platforms or technologies, study relationship between genes and observations is appropiate and necessary and enables gene selection. The classical definition of these methods is that enables the graphical approximation of the multivariable data matrixes -of order (nxp) and range $r-$, using columns and lines makers to study the relationships between individuals and variables from the singular values decomposition. In this case,Y be the matrix obtained by interactively coding all the matrices $Y = [X_1|X_2|...X_K]$:

$$Y = UDV^T \tag{4}$$

The approximation of equation can also be performed through a general multiplicative bilinear model (Vicente-Villardón et al., 2006), (Demey et al., 2008), (Sánchez and Vicente-Villardón, 2013):

$$Y = P\beta^T + \epsilon \tag{5}$$

Which can be understood as a multivariate regression of Y on the coordinates of individuals P, when they are fixed, or they are multivariate regression Y^T on the coordinates of the variables β, if they are who are fixed.

We add here a biplot interpretation of the method, based on the projection of the all variables onto the compromise space. Let Y be the matrix obtained by interactively coding all the matrices (X_k: $Y = [X_1|X_2|...X_K]$, of order nxJ, where J is $J = \sum_{i=1}^{K} j_i$, and P is an fixed matrix of the projection of all observations on the compromise, obtained from the eigen-decomposition of W_c, such as stated in section 1.
Given that P coordinates are known, obtaining the $\beta's$ is equivalent to performing a linear regression using the $j-th$ column of Y as a response variable and the columns of P as regressors. Thus, the projection of a individuals on the direction of an variable predicts the level of expression of such variable on this observations. With this biplot approximation on the compromise in the classic STATIS method, it is possible to project the variables of different data tables (in other words, all the genes involved in the study) and determine relationships between tissues and genes, genes with each other.

In classical STATIS, it is not possible to interpret the relationship between variables or variables and observations. With biplot aproximation, obtained by fitting linear regressions to that configuration as described in this section, itÂ´s posible obtain these relationships.
Due to the high number of variables usually studied, it is convenient to situate on the graph only those that are related to the configuration, i.e. those that have an adequate goodness of fit after adjusting the regression model (Demey, 2008). In addition, the technique can be used to select candidate genes, representatives of the data structure, using measures of goodness of fit, among this can be the adjusted R-squared, p value, p-value corrected by Bonferroni or criteria AIC (Akaike Information Criterious) and BIC (Bayesian Information Criterious).
Besides, these method allows generate groups of variables using a clustering algorithm.

3 Examples

In this section we provide an overview of the kimod package. The example consist in the analysis of four different microarrays platforms (i.e., Agilent, Afymetrix HGU 95, Afymetrix HGU 133 and Afymetrix HGU 133plus 2.0) on the NCI-60 cell lines (Shankavaram et al., 2009),(Reinhold et al., 2012). These datasets are illustrative and they have only a subset of microarray gene expression of the NCI 60 cell lines from four different platforms.

3.1 Package overview

The original data Files are available at Cell-Miner WebSite (Shankavaram et al., 2009),(Reinhold et al., 2012). In this dataset, the 60 human tumour cell lines are

derived from patients with leukaemia, melanoma, lung, colon, central nervous system, ovarian, renal, breast and prostate cancers. The cell line panel is widely used in anti-cancer drug screen. In this dataset, a subset of microarray gene expression of the NCI 60 cell lines from four different platforms are combined in a list.

```
> library("kimod")
```

```
> data(NCI60Selec_ESet)
```

Once we call the datasets, we ckeck your class using the class() comand:

```
> class(NCI60Selec_ESet)
```

```
[1] "list"
```

Then, check the dimensions of datasets.

```
> lapply(NCI60Selec_ESet,dim)
```

```
[[1]]
Features  Samples
      60      300

[[2]]
Features  Samples
      60      298

[[3]]
Features  Samples
      60      268

[[4]]
Features  Samples
      60      288
```

Finally, we ckeck if all tables have the same observations:

```
> Tissues<-c(rep("Breast",5),rep("CNS",6),rep("Colon",7),
+ rep("Leukemia",6),rep("Melanoma",10),rep("Lung",9),
+ rep("Ovarian",7),rep("Prostate",2),rep("Renal",8))
```

Next command returns an array with the rownames of all tables

```
> Names<-sapply(NCI60Selec_ESet,rownames)
```

And if the following command is TRUE, it means of all matrix have the same observations:

```
> unique(apply(Names[,-1],2,function(y)identical(y,Names[,1])))
```

```
[1] TRUE
```

Once the preprocessing of the experiment data is completed, the STATIS method can be carried out using by calling DiStatis function of **kimod** package:

```
> Z1<-DiStatis(NCI60Selec_ESet)

> class(Z1)

[1] "DiStatis"
attr(,"package")
[1] "kimod"
```

Z1 if an object of DiStatis S4-class, if is printing the main slots of Z1 are: distance. methods (that indicates the kind of distance (or scalar product) that is calculated in each study, Inertia of Vectorial Correlation, Euclidean image of studies, compromise matrix, P matrix for projection all observations in consensus- space, representation quality of observations and trajectories(i.e, the rows of the initial tables are projected in the the compromise-strucucture). To obtain the euclidean image of studies, runs:

```
> RVPlot(Z1)
```

The figure 1 shows the relative contributions of each of the tables to Components 1 and 2. Thus, we can see that Study 1 (correspondent to Agilent platform it has the lowest contribute to the compromise.

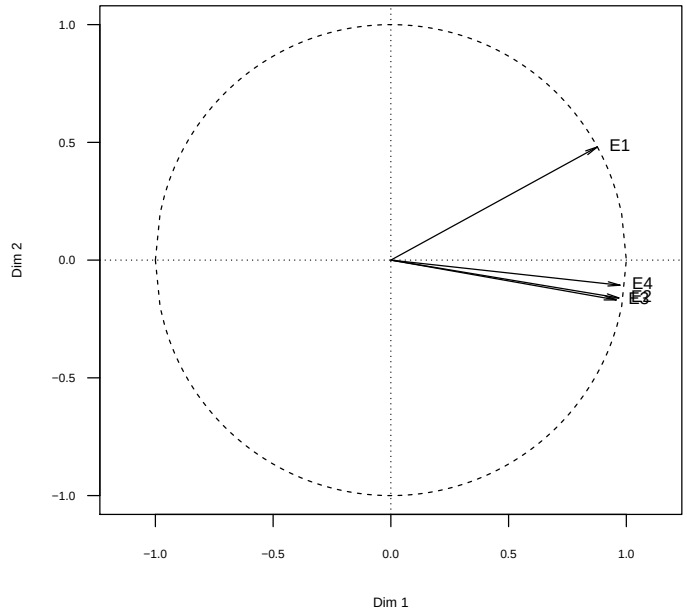

Figure 1: Contribution of all tables to the compromise.

To obtain the projection of observations on compromise, runs:

```
> Tissues<-c(rep("Breast",5),rep("CNS",6),rep("Colon",7),
+ rep("Leukemia",6),rep("Melanoma",10),rep("Lung",9),
+ rep("Ovarian",7),rep("Prostate",2),rep("Renal",8))

> Colours<-c(rep(colors()[657],5),rep(colors()[637],6),
+  rep(colors()[537],7),rep(colors()[552],6),rep(colors()[57],10),
+  rep(colors()[300],9),rep(colors()[461],7),rep(colors()[450],2),
+  rep(colors()[432],8))

> CompPlot(Z1,xlabBar="",colObs=Colours,pch=15,las=1,
+  cex=2,legend=FALSE,barPlot=FALSE,cex.main=0.6,cex.lab=0.6,
+  cex.axis=0.6,las=1)
>  legend("topleft",unique(Tissues),col=unique(Colours),
+  bty="n",pch=16,cex=1)
```

The figure 2 shows the projection of s cell lines onto the first two principal components of Compromise-structure. Cell lines of leukemia, melanoma and colon are clearly distinguished from the others. However, a melanoma cell line has similar profiles to carcinomas (CNS, renal ovarian, lung). Furthermore, the breast cancer varies widely, grouping itself some samples with colon tissues and others with CNS.

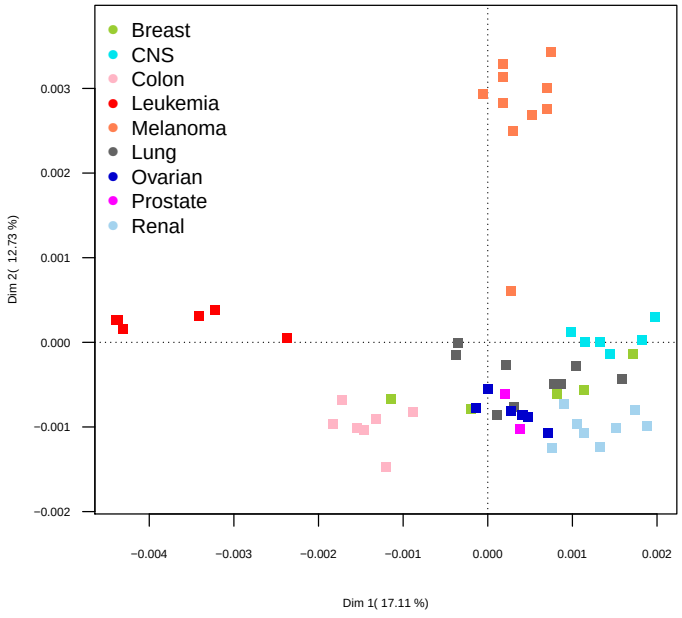

Figure 2: Compromise Plot. Projection of all tumoral tissues in the consensus space.

The Sample Variability is obtained by using Bootstrap and BootPlot functions. Bootstrap receives as argument an object of DiStatis Class and Boot-

Plot performs the Sample-Variaibility-Plot (see figure **??**). The Slot "Comparision.Boot" show difference between observations using the Bonferroni Correction for all dimensions.

```
> B<-Bootstrap(Z1)
> BootPlot(B,Points=FALSE,cex.lab=0.7,cex.axis=0.7,
+  las=1,xlimi=c(-0.003,0.002),ylimi=c(-0.005,0.007)
+  ,legend=FALSE,col=Colours)
>  legend("topleft",unique(Tissues),col=unique(Colours),
+  bty="n",pch=16,cex=1)
> Comparisions.Boot(B)

[[1]]
[[1]]$Pos
 [1] "BR.BT_549"     "BR.HS578T"     "BR.MDA_MB_231" "CNS.SF_268"
 [5] "CNS.SF_295"    "CNS.SF_539"    "CNS.SNB_19"    "CNS.SNB_75"
 [9] "CNS.U251"      "LC.EKVX"       "LC.HOP_62"     "LC.HOP_92"
[13] "LC.NCI_H226"   "ME.M14"        "ME.MALME_3M"   "ME.SK_MEL_28"
[17] "ME.UACC_62"    "OV.OVCAR_4"    "OV.OVCAR_5"    "OV.OVCAR_8"
[21] "OV.SK_OV_3"    "PR.DU_145"     "RE.786_0"      "RE.A498"
[25] "RE.ACHN"       "RE.CAKI_1"     "RE.RXF_393"    "RE.SN12C"
[29] "RE.TK_10"      "RE.UO_31"

[[1]]$Neg
 [1] "BR.MCF7"      "CO.COLO205"   "CO.HCC_2998"  "CO.HCT_116"   "CO.HCT_15"
 [6] "CO.HT29"      "CO.KM12"      "CO.SW_620"    "LC.NCI_H522"  "LE.CCRF_CEM"
[11] "LE.HL_60"     "LE.K_562"     "LE.MOLT_4"    "LE.RPMI_8226" "LE.SR"


[[2]]
[[2]]$Pos
 [1] "BR.BT_549"     "BR.HS578T"     "BR.MDA_MB_231" "CNS.SF_268"
 [5] "CNS.SF_295"    "CNS.SF_539"    "CNS.SNB_19"    "CNS.SNB_75"
 [9] "CNS.U251"      "LC.EKVX"       "LC.HOP_62"     "LC.HOP_92"
[13] "LC.NCI_H226"   "ME.M14"        "ME.MALME_3M"   "ME.SK_MEL_28"
[17] "ME.UACC_62"    "OV.OVCAR_4"    "OV.OVCAR_5"    "OV.OVCAR_8"
[21] "OV.SK_OV_3"    "PR.DU_145"     "RE.786_0"      "RE.A498"
[25] "RE.ACHN"       "RE.CAKI_1"     "RE.RXF_393"    "RE.SN12C"
[29] "RE.TK_10"      "RE.UO_31"

[[2]]$Neg
 [1] "BR.MCF7"      "CO.COLO205"   "CO.HCC_2998"  "CO.HCT_116"   "CO.HCT_15"
 [6] "CO.HT29"      "CO.KM12"      "CO.SW_620"    "LC.NCI_H522"  "LE.CCRF_CEM"
[11] "LE.HL_60"     "LE.K_562"     "LE.MOLT_4"    "LE.RPMI_8226" "LE.SR"

>
```

On figure 3 can be seen than then melanoma tissues have high internal variability. Moreover, from slot(B,"Comparisions.Boot"), we can see that Colon, Leukemia, $BR.MCF$ and LC_NCI_H522 tissues separates from others.

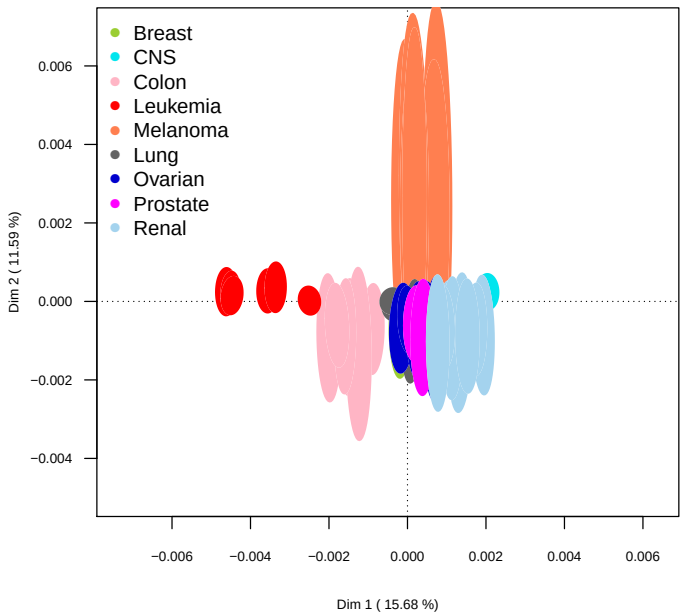

Figure 3: Sample-Variability Plot

For performs gene selection, responsibles of the tissues projections, and explore gene expression profiles, we can use the SelectVar function, that receives an main argument of DiStatis class. This function allows to build the biplot for continuous response, using an external procedure to obtained the regresors in the linear model (see section 4). Furthermore, allows select genes using measures of goodness of fit of the Models Biplot: adjusted R^2, P-value with bonferroni correction, AIC or BIC. The percentage of selected variables is an user input (See figure 4).

```
> M1<-SelectVar(Z1,Crit="R2-Adj",perc=0.95)
> layout(matrix(c(1,1,1,1,1,1,2,2),c(1,1,1,1,1,1,2,2),byrow=TRUE))
> Biplot(M1,labelObs = FALSE,labelVars=FALSE,
+        colObs=Colours,Type="SQRT",las=1,cex.axis=0.8,
+        cex.lab=0.8,xlimi=c(-3,3),ylimi=c(-3,3))
> plot(0,type='n',axes=FALSE,ann=FALSE)
> legend("topright",unique(Tissues),col=unique(Colours),
+        bty="n",pch=15,cex=1)
```

Besides,f Groups argument in this function is TRUE, the variables will be clustered using Euclidean distance and Ward algorithm (see figure 5).

Finally, to see relationships between gene clusters and tissues, may be used the GroupProj function, that receives an main argument of SelectVar class. This

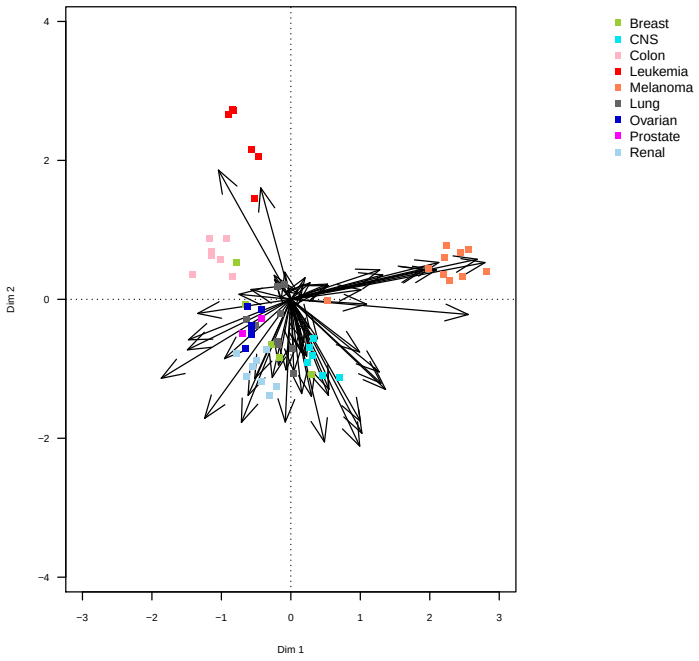

Figure 4: Biplot. Projection of gene-selected on Compromise

function use the cluster package (Maechler et al., 2015) which is automatically called in our package.

```
> A1<-GroupProj(M1,method="ward",metric="euclidean",NGroups=4)
> head(SortList(A1)[[1]])

               (+1 over-exp) (-1 under-exp)
BR.MCF7                                -1
BR.MDA_MB_231                          -1
BR.HS578T                               1
BR.BT_549                              -1
BR.T47D                                -1
CNS.SF_268                              1

>
```

The list shows that genes of cluster 1 are over-expressed in melanoma and CSN tissues and under-expressed in colon and leukemia (black in figure 5).
The gene on cluster 2 are over-expressed in Breast, CSN, Lung, Renal, Ovarion and Colon and under-expressed in melanoma and leukemia (red in figure 5).
The cluster 3 is related to under-expression in colon and leukemia tissues and over-expression on CSN and melanoma, mainly (green in in figure 5) .
Finally, the cluster 4 is associated to high expression in Colon and leukemia

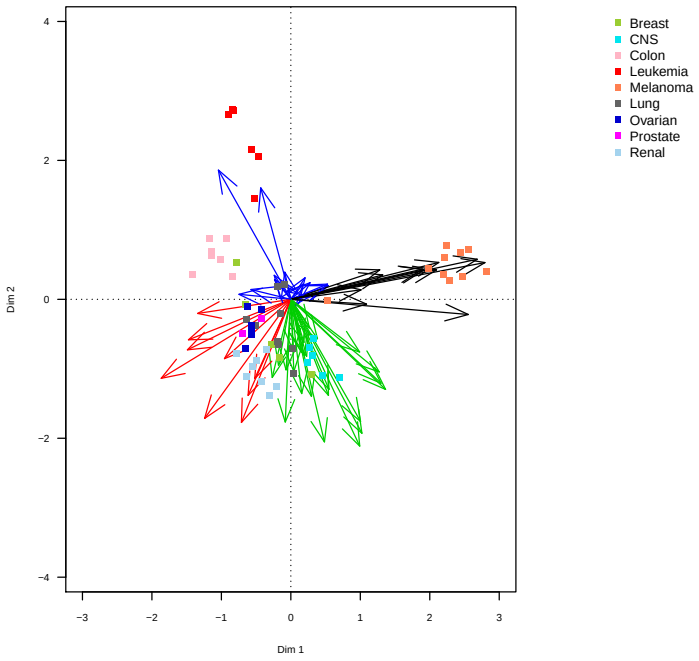

Figure 5: Biplot. Projection of gene-selected on Compromise

tissues and breast: BR.MCF7 and BR.T47D. The list of all cluster gene is obtained:

```
> A1<-GroupProj(M1,method="ward",metric="euclidean",NGroups=4)
> Groups(A1)

[[1]]
 [1] "ACP5"     "C10orf90" "C6orf218" "CA14"     "DUSP4"    "GPNMB"
 [7] "KAT2B"    "PLP1"     "S100A1"   "S100B"    "SOX10"

[[2]]
[1] "ANXA3"   "CA12"    "ECHDC2"  "F3"      "FERMT1"  "KRT8P23" "OCLN"
[8] "PTGES"   "TBC1D2"

[[3]]
 [1] "ARAP3"     "BACE1"     "BACE2"     "C3orf59"   "C9orf30"   "FAM57A"
 [7] "FKBP10"    "GPC6"      "LEPREL1"   "LHFP"      "LOC344978" "MLF1"
[13] "PDLIM2"    "PPIC"      "PRKD1"     "PTPN21"    "PVR"       "RAB32"
[19] "SC65"      "SMAD3"     "SPEG"      "SYNM"      "VAMP3"     "WWC2"

[[4]]
 [1] "ATP1A3"   "C19orf39" "C1orf131" "CPNE7"    "ETV3"     "FGD3"
 [7] "GMFG"     "GNPTAB"   "HELZ"     "HERC1"    "HSPBAP1"  "TOR2A"
```

```
[13] "ZFP36"
```

References

Abdi, H., Valentin, D., Chollet, S., and Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality and Preference*, 18(4):627–640.

Abdi, H., Williams, L. J., Valentin, D., and Bennani-Dosse, M. (2012). STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):124–167.

Demey, J. R. (2008). *DIVERSIDAD GENETICA EN BANCOS DE GERMOPLASMA : UN ENFOQUE BIPLOT*. PhD thesis, Universidad de Salamanca.

Demey, J. R., Vicente-Villardón, J. L., Galindo-Villardón, M. P., and Zambrano, a. Y. (2008). Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. *Bioinformatics (Oxford, England)*, 24(24):2832–8.

des Plantes, H. L. (1976). *Structuration des tableaux à trois indices de la statistique: théorie et application d'une méthode d'analyse conjointe*. PhD thesis, Université des sciences et techniques du Languedoc.

Efron, B. and Tibshirani, R. J. (1993). An introduction to the bootstrapchapman & hall. *New York*, 436.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, pages 751–760.

Escoufier, Y., Cazes, P., et al. (1976). Opérateurs et analyse des tableaux à plus de deux dimensions. *Cahiers du bureau universitaire de recherche opérationnelle*, 25:61–89.

Gabriel, K. R. (1971). The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika*, 58(3):453.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2015). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.3 — For new features, see the 'Changelog' file (in the package source).

Reinhold, W. C., Sunshine, M., Liu, H., Varma, S., Kohn, K. W., Morris, J., Doroshow, J., and Pommier, Y. (2012). CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer research*, 72(14):3499–511.

Sánchez, J. C. H. and Vicente-Villardón, J. L. (2013). Logistic biplot for nominal data. *arXiv preprint arXiv:1309.5486*.

Shankavaram, U. T., Varma, S., Kane, D., Sunshine, M., Chary, K. K., Reinhold, W. C., Pommier, Y., and Weinstein, J. N. (2009). CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC genomics*, 10:277.

Vicente-Villardón, J. L., Galindo Villardón, M. P., Blázquez Zaballos, A., Greenacre, M., and Blasisus, J. (2006). Logistic biplots. *Multiple correspondence analysis and related methods. London: Chapman & Hall*, pages 503–521.

Session Info

```
> sessionInfo()

R version 3.4.2 Patched (2017-10-07 r73498)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows Server 2012 R2 x64 (build 9600)

Matrix products: default

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] kimod_1.6.0

loaded via a namespace (and not attached):
[1] compiler_3.4.2      parallel_3.4.2      tools_3.4.2
[4] Biobase_2.38.0      BiocGenerics_0.24.0 cluster_2.0.6
```