R SubtypeDiscovery: a data mining scenario for the inference of subtypes by cluster analysis

F. Colas

July 24, 2009

Introduction

In the study of phenomena characterized by heterogeneity, an important and general data analysis problem is the search for more homogeneous subtypes in the data distribution. In clinical research on Osteoarthritis, Parkinson's disease, major depressive and anxiety disorders, or glioblastoma and metastasis discrimination, the identification of more homogeneous patient subtypes may contribute to understand more specifically the underlying mechanisms of the these pathologies, and thus help to develop tailored prevention strategies and treatments.

To advance research on these phenomena, we developed a data-mining scenario designed to infer subtypes by cluster analysis [Col09b, Col09a]. This scenario is referred to as the SubtypeDiscovery and it was implemented as an R package. As a result, other research teams can benefit of this tested scenario to perform their subtyping analyses. With this package, analyses are straightforward and therefore accessible to many investigators. Furthermore, the well-defined data structures and the public availability of the package greatly enhance reproducibility.

The scenario features various data preparation techniques, an approach that repeats data modeling in order to select for the number of subtypes and/or the type of model, with a selection of methods to characterize, compare and evaluate the most likely mixture models. The combination of steps of a typical subtype discovery analysis outlines as follows:

1. data configuration, processing and exploratory data analysis (EDA),

2. model based clustering [Fra02, Fra06] repeated from different initialization points,

3. selection of the top models based on a Bayesian Information Criterion ranking,

4. cross-comparison of those mixture models and characterization of their subtypes,

5. relevance evaluation of each subtype.

Prepare the data, EDA

Before loading the SubtypeDiscovery package with the usual `library` command, package dependencies must be resolved. We install the `mclust`, `RColorBrewer`, `abind`, `e1071`, `class` and `R2HTML` package and then, load the library, using the following commands:

```
> install.packages(c("mclust", "RColorBrewer", "abind", "e1071",
+     "class"))

> library(SubtypeDiscovery)

by using mclust, you accept the license agreement in the LICENSE file
and at http://www.stat.washington.edu/mclust/license.txt
```

The SubtypeDiscovery package includes a public chemoinformatics dataset made available by Ed O Cannon, for which more information can be obtained from the man page (see `?wada2008`). To search for subtypes in this dataset, we start by loading it into the R environment. Then, we generate automatically a pre-filled `settings` file, which defines how SubtypeDiscovery must interpret the data. Yet, because variables may have an ordering, or because some variables should be left out of the clustering, or used in the validation stage, the file needs to be edited. For this reason, we save the `settings` into a `csv` file, and we will further edit it with a spreadsheet like Excel or Openoffice.

```
> data(wada2008)
> settings <- generate_cdata_settings(wada2008)
> write.csv(settings, file = "settings.csv")
```

Edit the settings file in a spreadsheet, save it, and read it back into the R environment.

```
> settings <- read.csv("settings.csv", row.names = 1)
```

Along with the dataset, we stored a pre-edited settings file in SubtypeDiscovery and it is saved as wada2008_settings. We retrieve it and show a few of its lines.

```
> colnames(settings)

[1] "group"          "in_canalysis"   "fun_transform"  "visu_groups"
[5] "visu_ycoord"    "heatmap_ycoord"

> data(wada2008_settings)
> wada2008_settings[c(1:3, 53:54), ]

            group                                                in_canalysis
balabanJ    "Adjency and distance matrix descriptors"            "TRUE"
diameter    "Adjency and distance matrix descriptors"            "TRUE"
KierFlex    "Kier and Hall connectivity and Kappa shape indices" "TRUE"
Data.Source NA                                                   "FALSE"
Labels      NA                                                   "FALSE"
            fun_transform
balabanJ    "transform_AVG transform_SIGMA"
diameter    "transform_AVG transform_SIGMA"
KierFlex    "transform_AVG transform_SIGMA"
```

```
Data.Source NA
Labels      NA
            visu_groups
balabanJ    "Adjency, distance matrix, Kier and Hall\nconnectivity, Kappa shape indices"
diameter    "Adjency, distance matrix, Kier and Hall\nconnectivity, Kappa shape indices"
KierFlex    "Adjency, distance matrix, Kier and Hall\nconnectivity, Kappa shape indices"
Data.Source NA
Labels      NA
            visu_ycoord heatmap_ycoord
balabanJ    "10.00"     "1"
diameter    " 9.50"     "2"
KierFlex    " 9.00"     "3"
Data.Source NA          NA
Labels      NA          NA
```

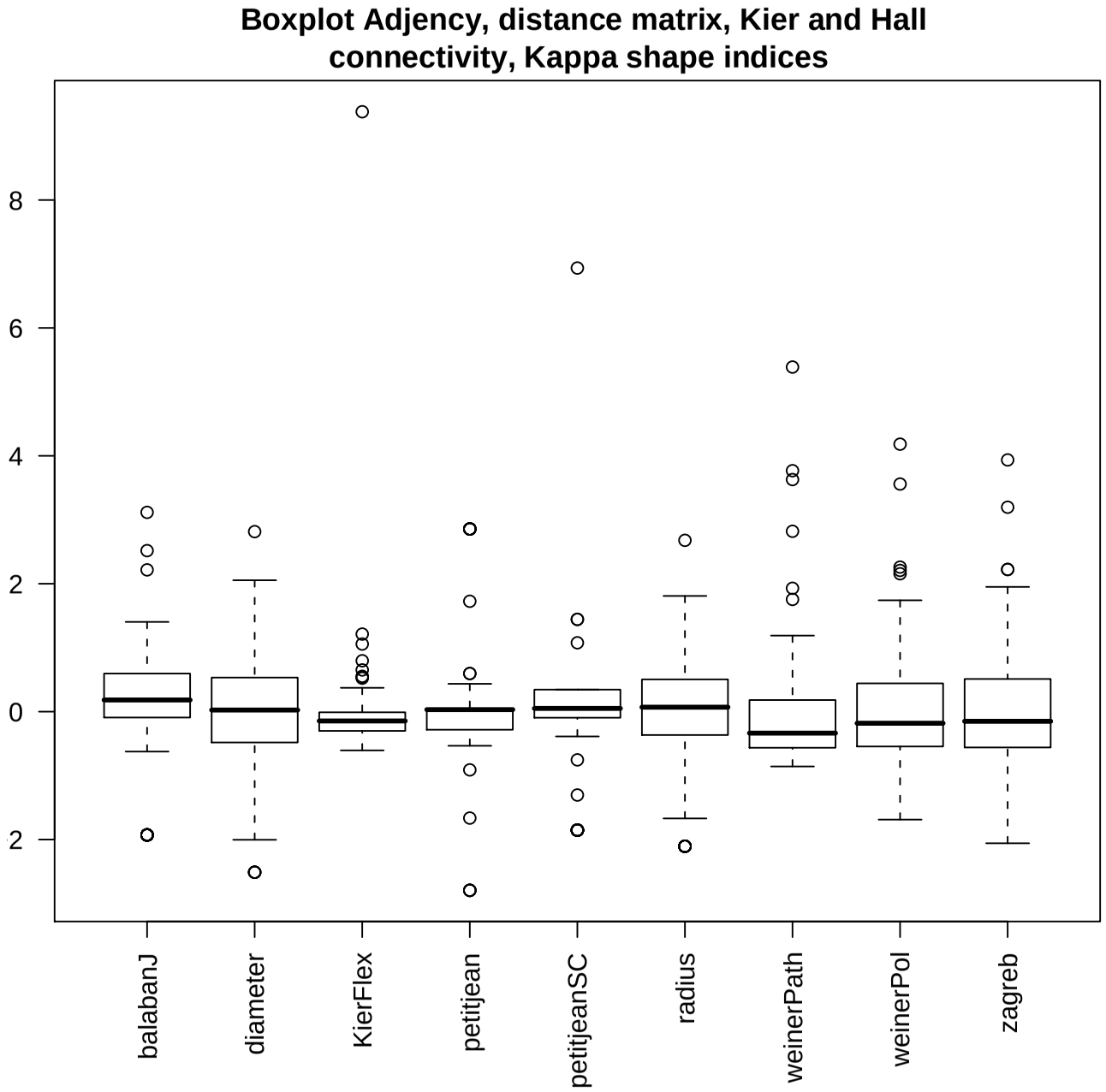
We then proceed to the creation of the data container instantiation (`cdata`):

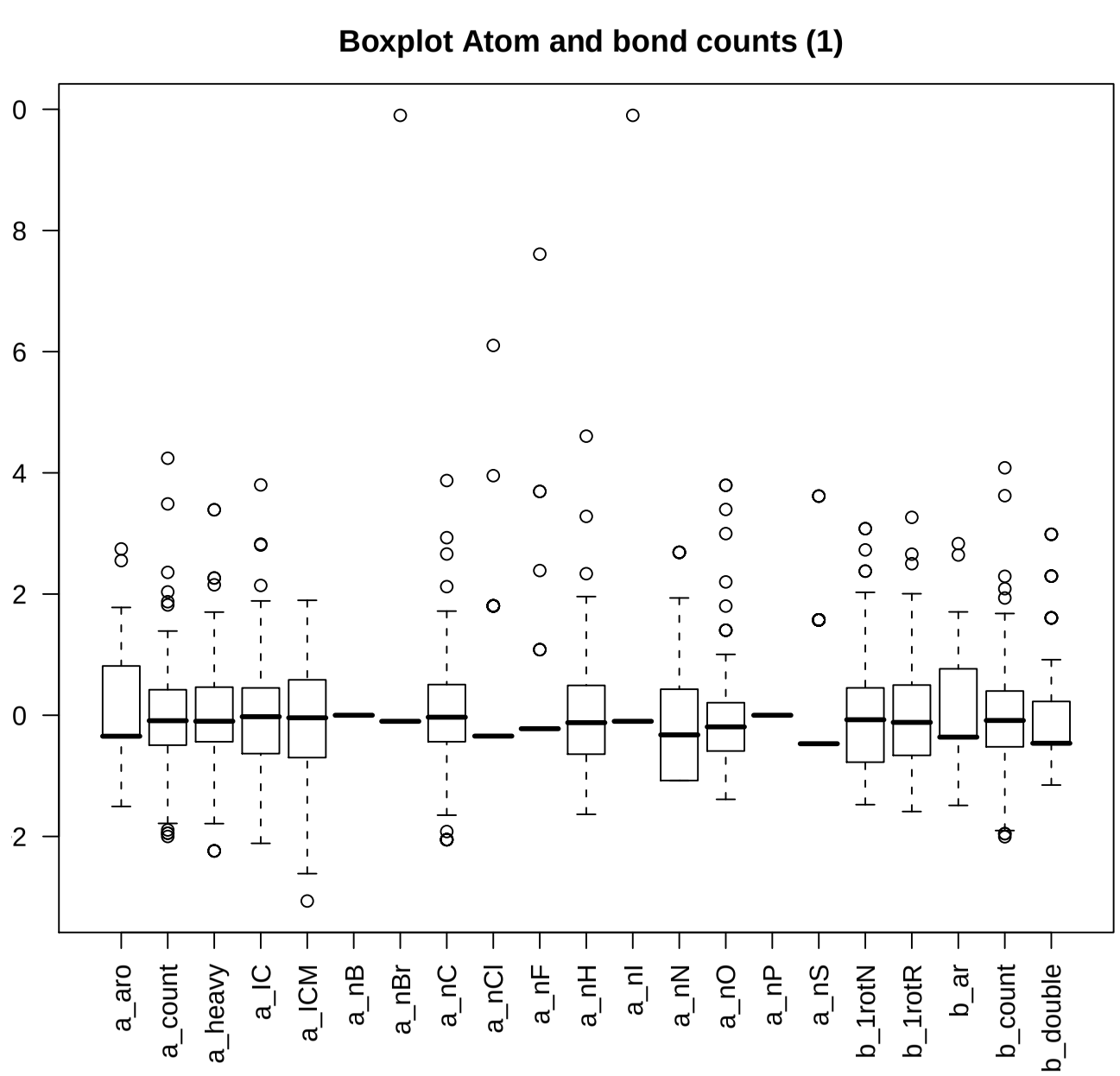
```
> wada2008_cdata <- set_cdata(data = wada2008[1:100, ], settings = wada2008_settings,
+     prefix = "Wada2008")
```

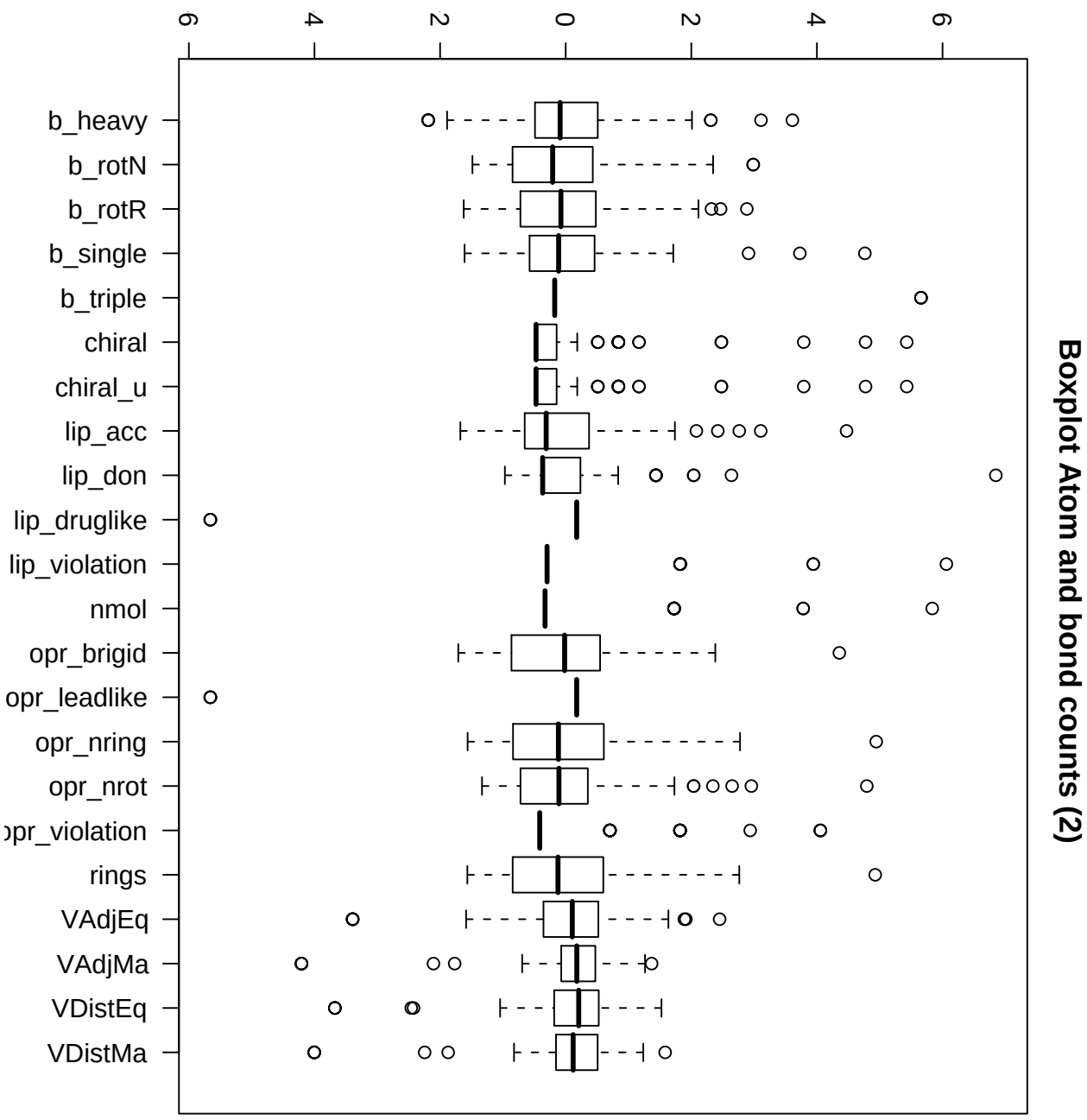
An exploratory data analysis (EDA) can be made from the plot function of a `cdata` dataset. Because there may be a large number of boxplots, histograms and statistics, the output is redirected into a postscript file `YYYY-MM-DD_PREFIX_CDATA.ps`.

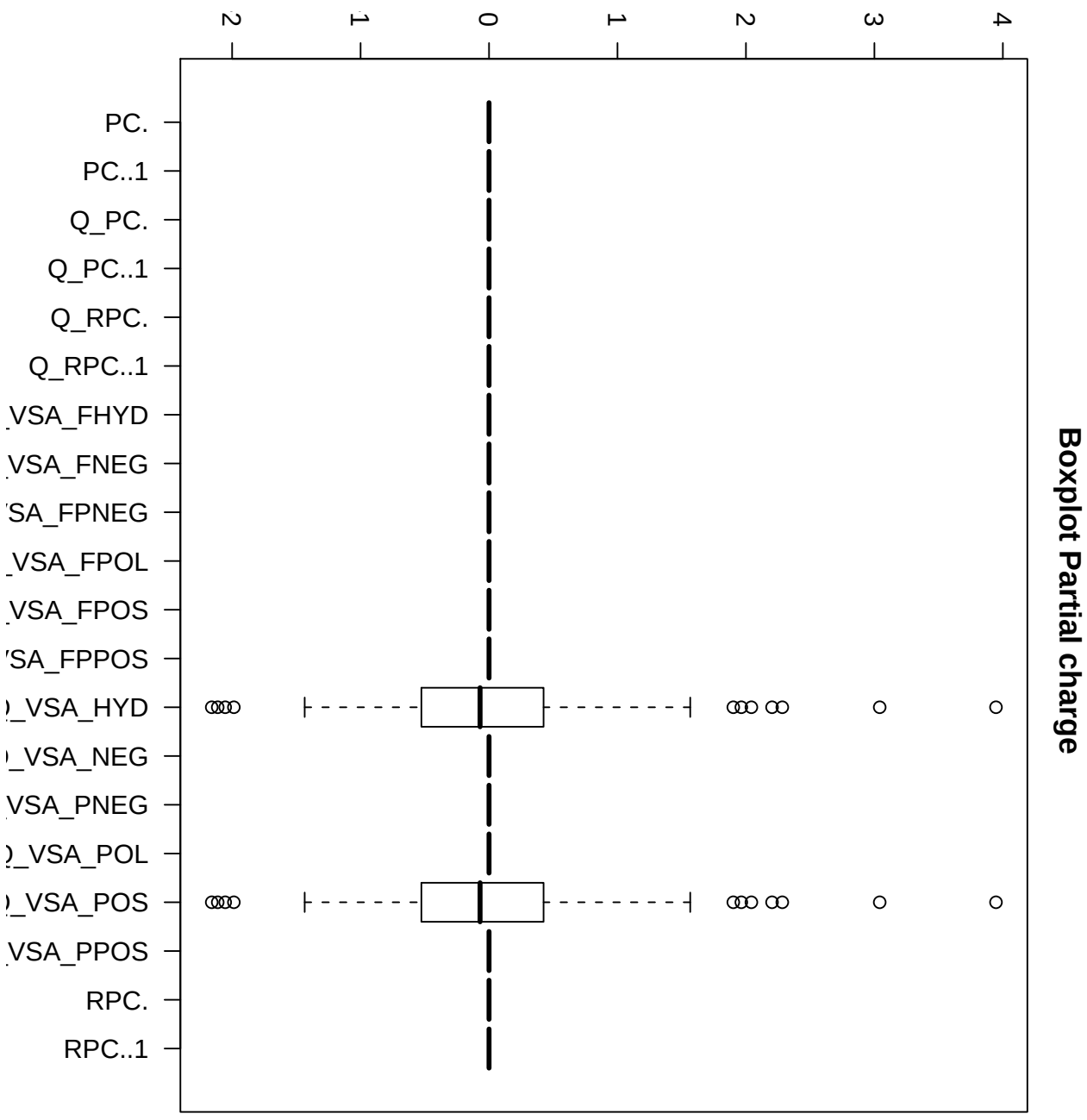
```
> plot(wada2008_cdata)
```

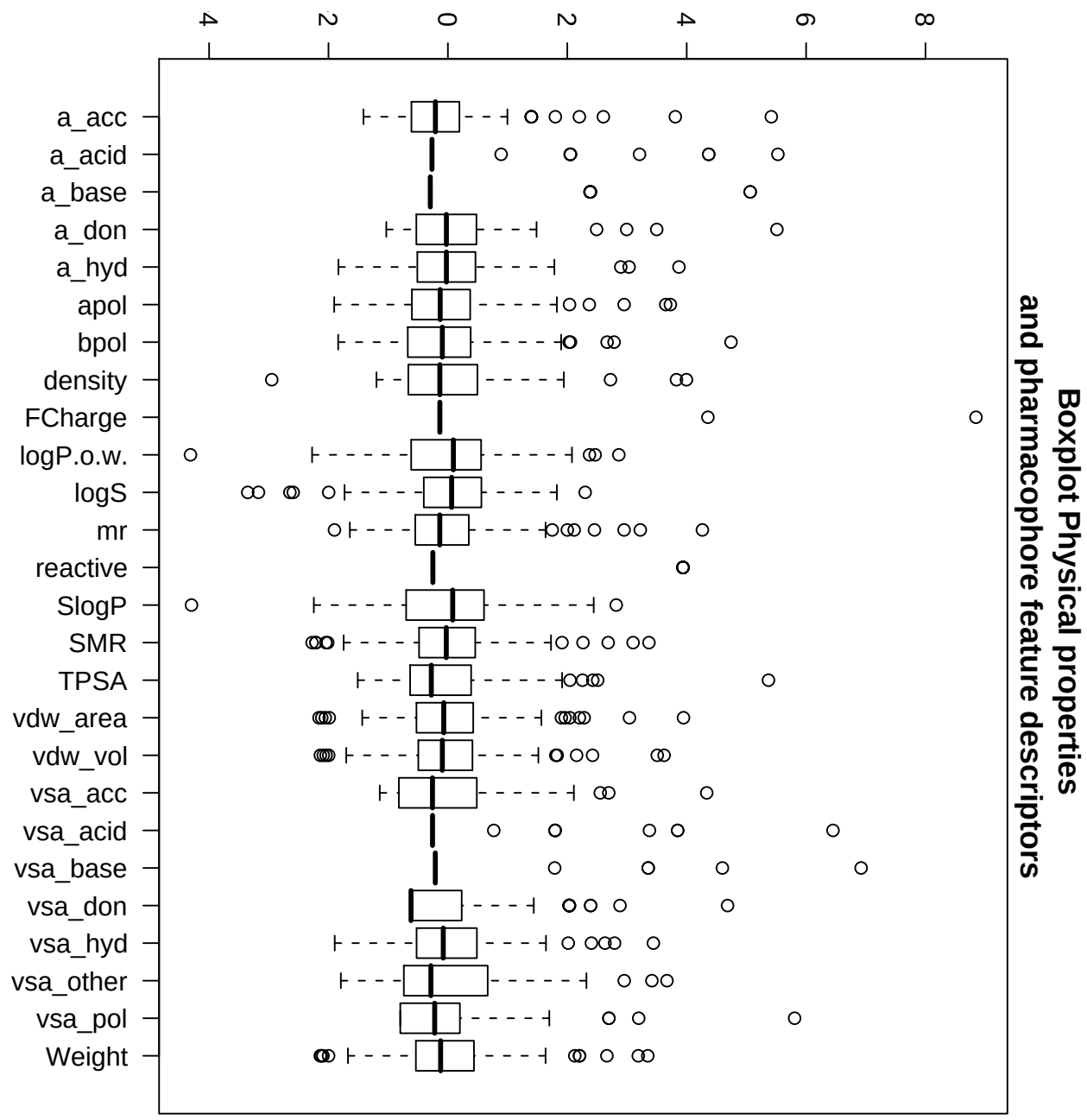
```
[1] "figures/2009-07-24_Wada2008_001-BB.pdf"
[1] "figures/2009-07-24_Wada2008_001-H.pdf"
[1] "figures/2009-07-24_Wada2008_002-BB.pdf"
[1] "figures/2009-07-24_Wada2008_002-H.pdf"
[1] "figures/2009-07-24_Wada2008_003-BB.pdf"
[1] "figures/2009-07-24_Wada2008_003-H.pdf"
[1] "figures/2009-07-24_Wada2008_004-BB.pdf"
[1] "figures/2009-07-24_Wada2008_004-H.pdf"
[1] "figures/2009-07-24_Wada2008_005-BB.pdf"
[1] "figures/2009-07-24_Wada2008_005-H.pdf"
```

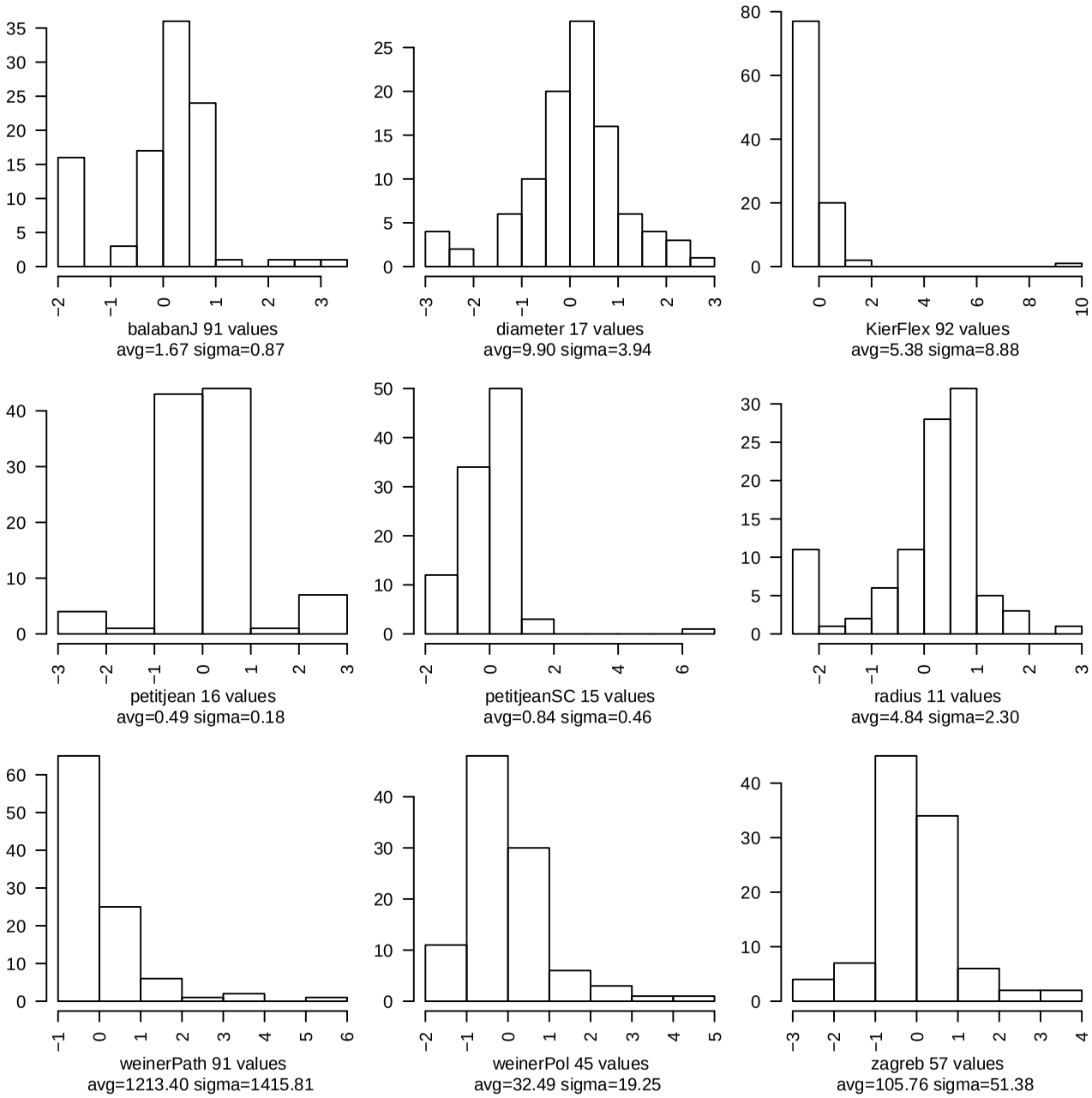

Boxplot Adjency, distance matrix, Kier and Hall connectivity, Kappa shape indices
8
6
4
2
0
2
balabanJ
diameter
KierFlex
petitjean
petitjeanSC
radius
weinerPath
weinerPol
zagreb

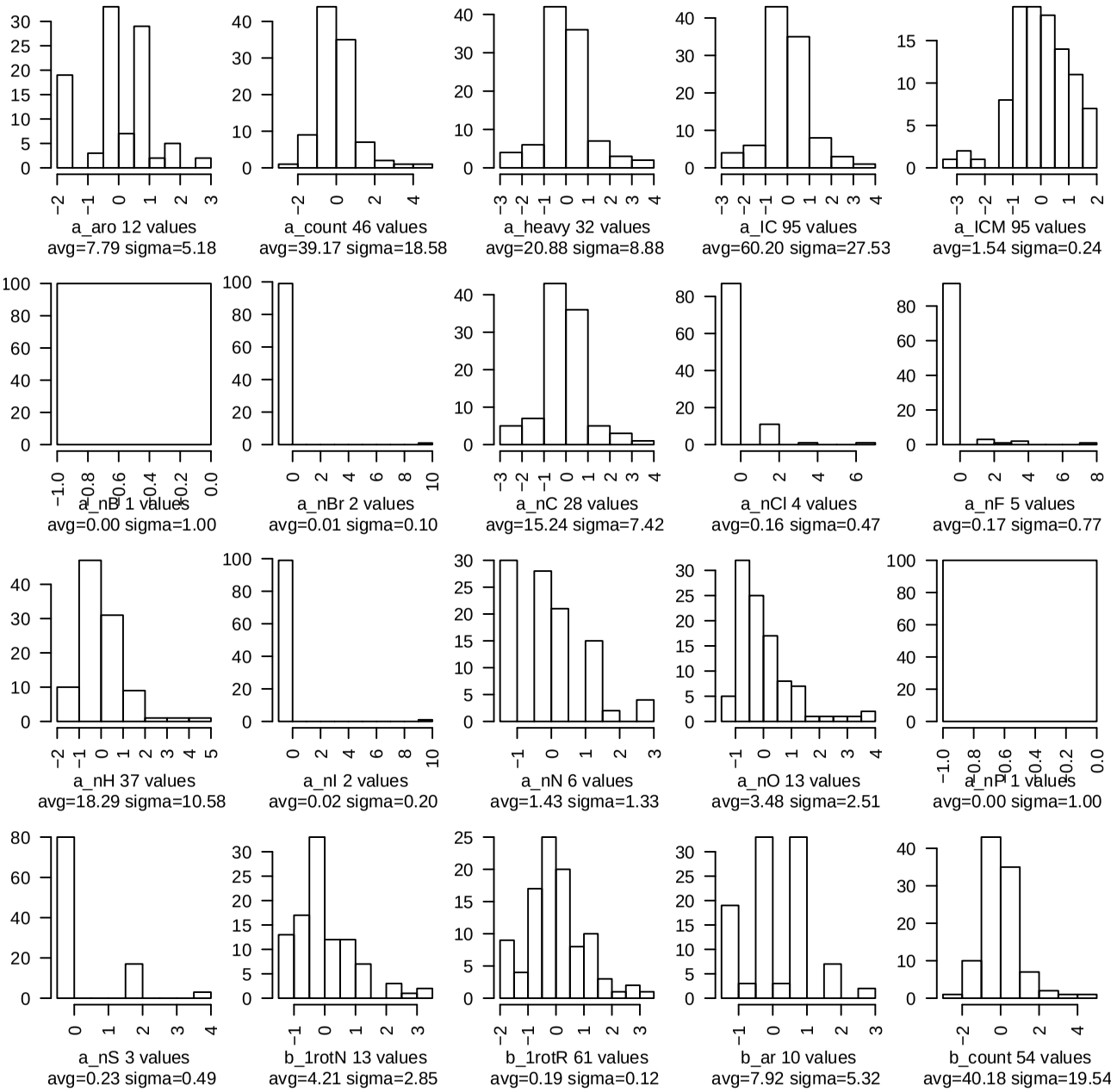

Boxplot Atom and bond counts (1)
0
8
6
4
2
0
2
a_aro
a_count
a_heavy
a_IC
a_ICM
a_nB
a_nBr
a_nC
a_nCl
a_nF
a_nH
a_nI
a_nN
a_nO
a_nP
a_nS
b_1rotN
b_1rotR
b_ar
b_count
b_double

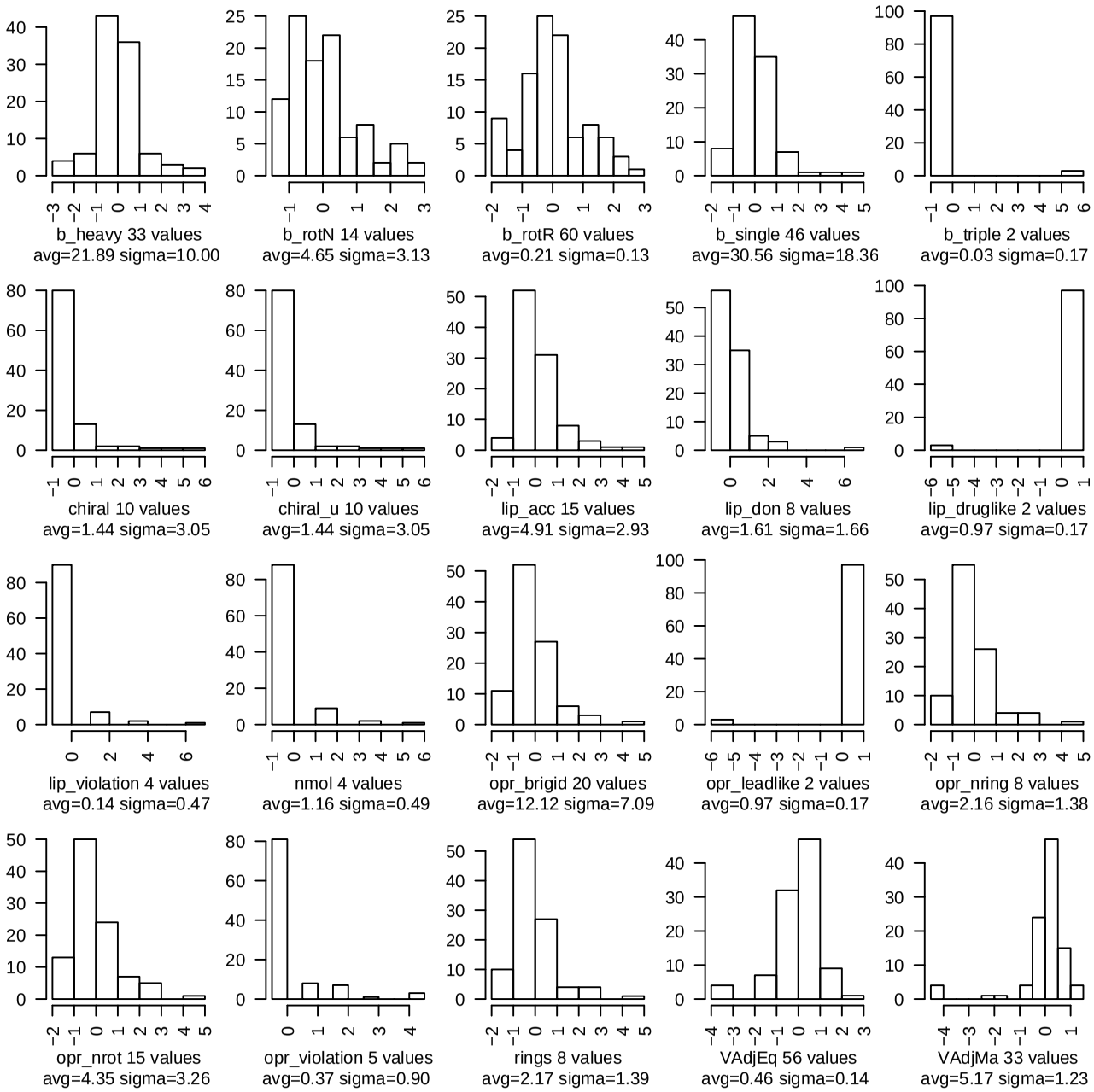

Boxplot Atom and bond counts (2)
b_heavy
b_rotN
b_rotR
b_single
b_triple
chiral
chiral_u
lip_acc
lip_don
lip_druglike
lip_violation
nmol
opr_brigid
opr_leadlike
opr_nring
opr_nrot
opr_violation
rings
VAdjEq
VAdjMa
VDistEq
VDistMa
6
4
2
0
2
4
6

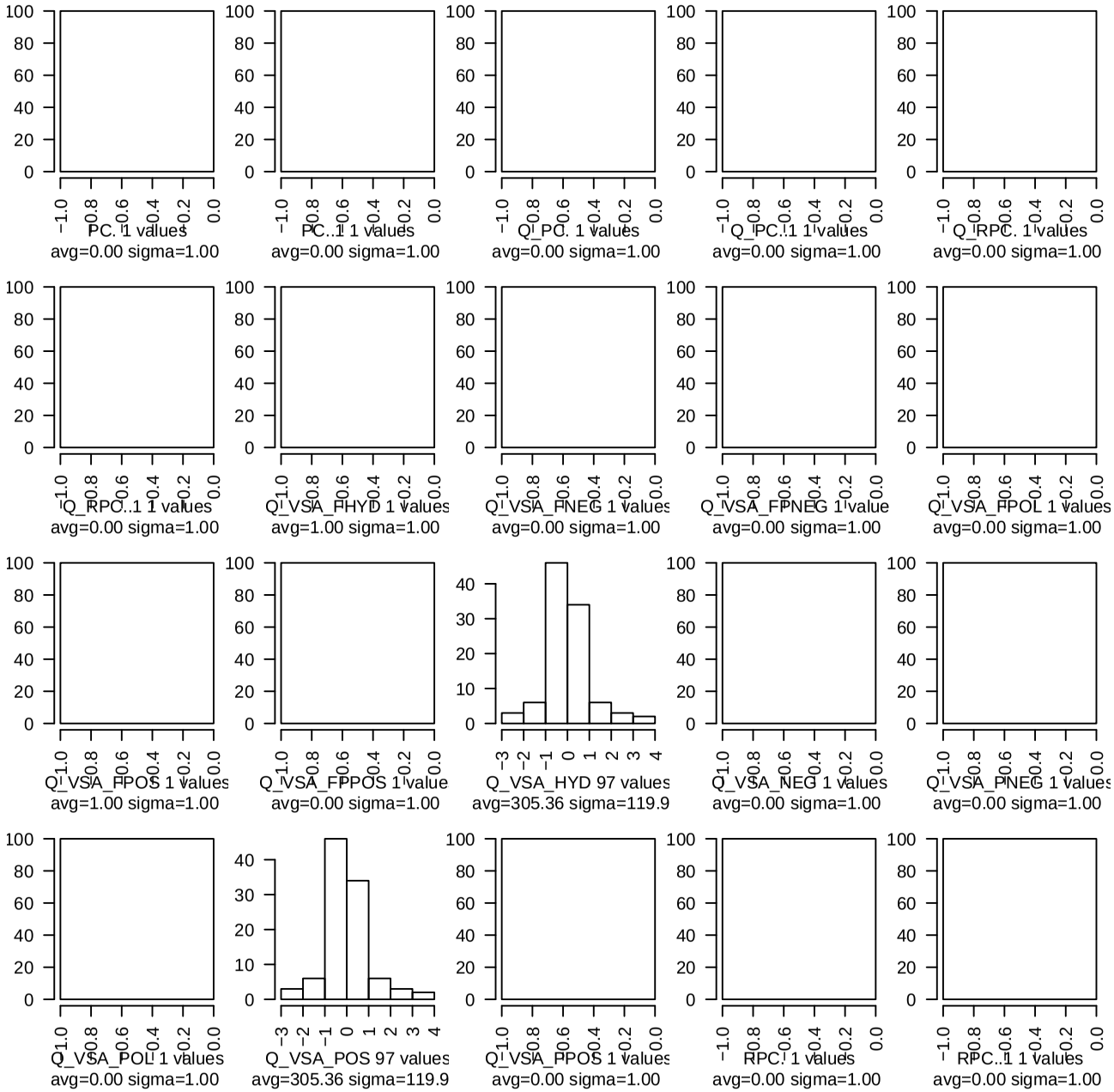

Boxplot Partial charge
2
1
0
1
2
3
4
PC.
PC..1
Q_PC.
Q_PC..1
Q_RPC.
Q_RPC..1
_VSA_FHYD
_VSA_FNEG
'SA_FPNEG
_VSA_FPOL
_VSA_FPOS
'SA_FPPOS
)_VSA_HYD
)_VSA_NEG
VSA_PNEG
)_VSA_POL
)_VSA_POS
_VSA_PPOS
RPC.
RPC..1

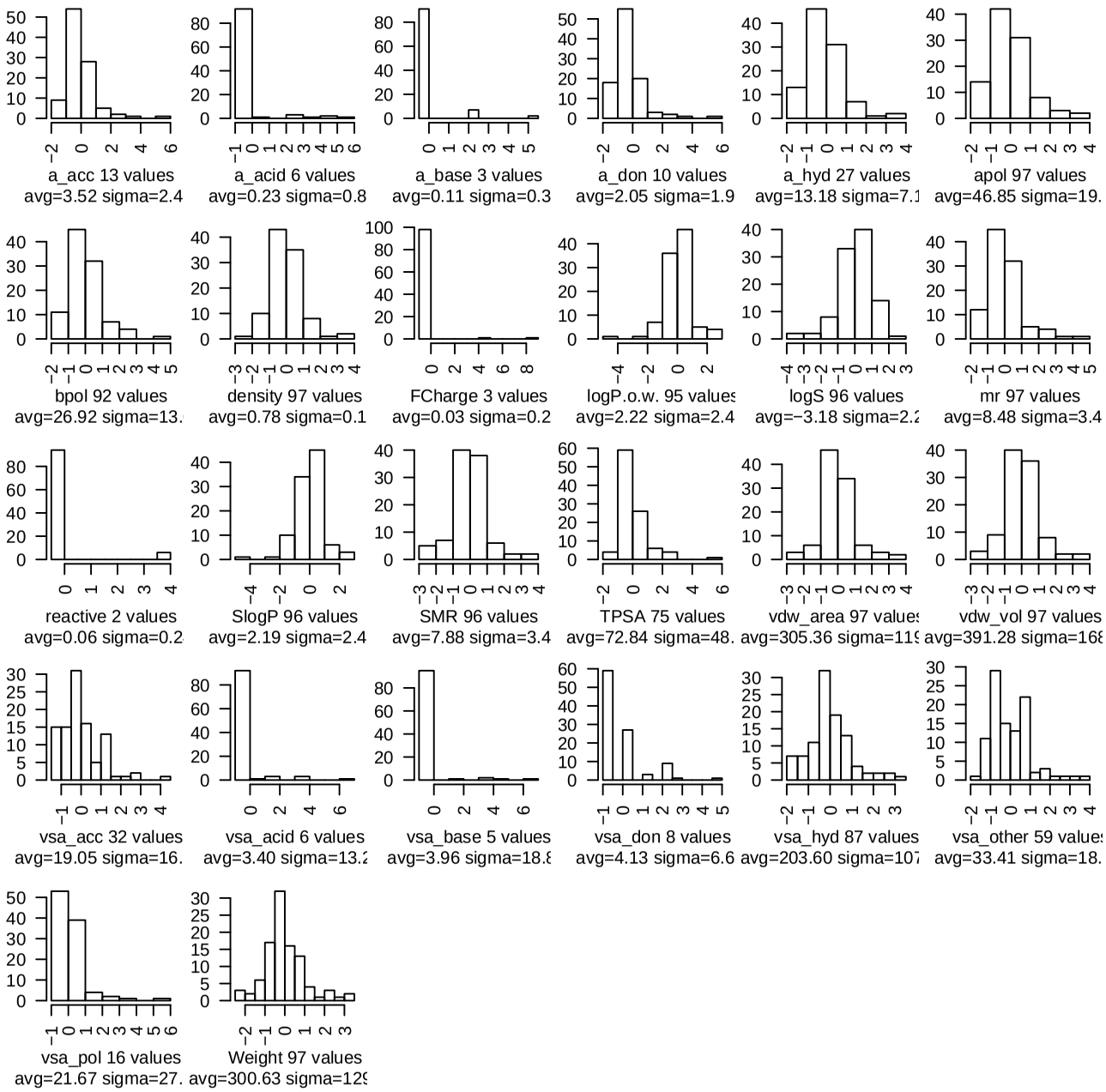

Boxplot Physical properties
and pharmacophore feature descriptors
a_acc
a_acid
a_base
a_don
a_hyd
apol
bpol
density
FCharge
logP.o.w.
logS
mr
reactive
SlogP
SMR
TPSA
vdw_area
vdw_vol
vsa_acc
vsa_acid
vsa_base
vsa_don
vsa_hyd
vsa_other
vsa_pol
Weight
4
2
0
2
4
6
8


balabanJ 91 values
avg=1.67 sigma=0.87
diameter 17 values
avg=9.90 sigma=3.94
KierFlex 92 values
avg=5.38 sigma=8.88
petitjean 16 values
avg=0.49 sigma=0.18
petitjeanSC 15 values
avg=0.84 sigma=0.46
radius 11 values
avg=4.84 sigma=2.30
weinerPath 91 values
avg=1213.40 sigma=1415.81
weinerPol 45 values
avg=32.49 sigma=19.25
zagreb 57 values
avg=105.76 sigma=51.38


a_aro 12 values
avg=7.79 sigma=5.18
a_count 46 values
avg=39.17 sigma=18.58
a_heavy 32 values
avg=20.88 sigma=8.88
a_IC 95 values
avg=60.20 sigma=27.53
a_ICM 95 values
avg=1.54 sigma=0.24
a_nB 1 values
avg=0.00 sigma=1.00
a_nBr 2 values
avg=0.01 sigma=0.10
a_nC 28 values
avg=15.24 sigma=7.42
a_nCl 4 values
avg=0.16 sigma=0.47
a_nF 5 values
avg=0.17 sigma=0.77
a_nH 37 values
avg=18.29 sigma=10.58
a_nI 2 values
avg=0.02 sigma=0.20
a_nN 6 values
avg=1.43 sigma=1.33
a_nO 13 values
avg=3.48 sigma=2.51
a_nP 1 values
avg=0.00 sigma=1.00
a_nS 3 values
avg=0.23 sigma=0.49
b_1rotN 13 values
avg=4.21 sigma=2.85
b_1rotR 61 values
avg=0.19 sigma=0.12
b_ar 10 values
avg=7.92 sigma=5.32
b_count 54 values
avg=40.18 sigma=19.54


b_heavy 33 values
avg=21.89 sigma=10.00
b_rotN 14 values
avg=4.65 sigma=3.13
b_rotR 60 values
avg=0.21 sigma=0.13
b_single 46 values
avg=30.56 sigma=18.36
b_triple 2 values
avg=0.03 sigma=0.17
chiral 10 values
avg=1.44 sigma=3.05
chiral_u 10 values
avg=1.44 sigma=3.05
lip_acc 15 values
avg=4.91 sigma=2.93
lip_don 8 values
avg=1.61 sigma=1.66
lip_druglike 2 values
avg=0.97 sigma=0.17
lip_violation 4 values
avg=0.14 sigma=0.47
nmol 4 values
avg=1.16 sigma=0.49
opr_brigid 20 values
avg=12.12 sigma=7.09
opr_leadlike 2 values
avg=0.97 sigma=0.17
opr_nring 8 values
avg=2.16 sigma=1.38
opr_nrot 15 values
avg=4.35 sigma=3.26
opr_violation 5 values
avg=0.37 sigma=0.90
rings 8 values
avg=2.17 sigma=1.39
VAdjEq 56 values
avg=0.46 sigma=0.14
VAdjMa 33 values
avg=5.17 sigma=1.23


PC. 1 values
avg=0.00 sigma=1.00
PC..1 1 values
avg=0.00 sigma=1.00
Q_PC. 1 values
avg=0.00 sigma=1.00
Q_PC..1 1 values
avg=0.00 sigma=1.00
Q_RPC. 1 values
avg=0.00 sigma=1.00
Q_RPC..1 1 values
avg=0.00 sigma=1.00
Q_VSA_FHYD 1 values
avg=1.00 sigma=1.00
Q_VSA_FNEG 1 values
avg=0.00 sigma=1.00
Q_VSA_FPNEG 1 value
avg=0.00 sigma=1.00
Q_VSA_FPOL 1 values
avg=0.00 sigma=1.00
Q_VSA_FPOS 1 values
avg=1.00 sigma=1.00
Q_VSA_FPPOS 1 value
avg=0.00 sigma=1.00
Q_VSA_HYD 97 values
avg=305.36 sigma=119.9
Q_VSA_NEG 1 values
avg=0.00 sigma=1.00
Q_VSA_PNEG 1 values
avg=0.00 sigma=1.00
Q_VSA_POL 1 values
avg=0.00 sigma=1.00
Q_VSA_POS 97 values
avg=305.36 sigma=119.9
Q_VSA_PPOS 1 values
avg=0.00 sigma=1.00
RPC. 1 values
avg=0.00 sigma=1.00
RPC..1 1 values
avg=0.00 sigma=1.00


a_acc 13 values
avg=3.52 sigma=2.4
a_acid 6 values
avg=0.23 sigma=0.8
a_base 3 values
avg=0.11 sigma=0.3
a_don 10 values
avg=2.05 sigma=1.9
a_hyd 27 values
avg=13.18 sigma=7.1
apol 97 values
avg=46.85 sigma=19.
bpol 92 values
avg=26.92 sigma=13.
density 97 values
avg=0.78 sigma=0.1
FCharge 3 values
avg=0.03 sigma=0.2
logP.o.w. 95 values
avg=2.22 sigma=2.4
logS 96 values
avg=−3.18 sigma=2.2
mr 97 values
avg=8.48 sigma=3.4
reactive 2 values
avg=0.06 sigma=0.2
SlogP 96 values
avg=2.19 sigma=2.4
SMR 96 values
avg=7.88 sigma=3.4
TPSA 75 values
avg=72.84 sigma=48.
vdw_area 97 values
avg=305.36 sigma=11
vdw_vol 97 values
avg=391.28 sigma=16
vsa_acc 32 values
avg=19.05 sigma=16.
vsa_acid 6 values
avg=3.40 sigma=13.2
vsa_base 5 values
avg=3.96 sigma=18.
vsa_don 8 values
avg=4.13 sigma=6.6
vsa_hyd 87 values
avg=203.60 sigma=107
vsa_other 59 values
avg=33.41 sigma=18.
vsa_pol 16 values
avg=21.67 sigma=27.
Weight 97 values
avg=300.63 sigma=12

Prepare the analysis and start the calculations

The dataset `cdata` is the only mandatory parameters of `set_cresult`. Thus, with the other parameters set as default, it is possible to perform a sample subtype discovery analysis. In the case of the `wada2008` data, commands are as follows:

```
> wada2008_cresult <- set_cresult(cdata = wada2008_cdata)
> analysis(wada2008_cresult)
```

The output of `analysis` describes the sequence of calculations performed. First, there are the mixture modelling (`EII,VII`), the number of mixture components (`3,4,5`), and the random initialization integer (`6013,6014,6015`). Then, statistical patterns such as the empirical mean, the standard deviation, or other quantile statistics are estimated (`Patterns`). For each mixture model, by default, a euclidean distance based hierarchical clustering is performed on both the variables and the observations (`Dendros`), the `Patterns` are re-ordered accordingly (`Ordering`). Finally, statistics for each subtype are calculated during the `Stats` step, e.g. the odds ratios.

Once calculations are completed, the subtype discovery analysis (`cresult`) is stored on the hard drive into an `RData` file (`YYYY-MM-DD_PREFIX_IMAGE.RData`). Further, to enable post-hoc analysis of the discovered subtypes, the set of best mixture models is stored into `csv` files. These files report the likelihood of every element to belong to each mixture component, along with the component affectation.

Next, a graphical characterization of the different mixture models is realised. For the same reasons than before, because the number of graphics is pretty large, the result of the characterization is stored into an additional postscript file (`YYY-MM-DD_PREFIX_CRESULT.ps`). Last, to perform the statistical inference of the subtypes, a number of summary measures are calculated on the set of BIC scores, cross-comprisons are performed between the most likely mixture models, showing in particular the joint distribution between these models, and thus, their consistency, etc. A report is assembled from these calculations and will assist the investigators in their inference (`Generate HTML report...`).

Optional parameters `set_cresult` does accept a number of additional parameters to adapt the calculations to the application area. Among these parameters, there is first the cluster modelling method (`cfun`), whose parameters are provided in `cfun_params`. There is, too, the graphic characterization of the mixture models, which is defined by a `fun_plot` parameter expecting a list of functions provided by `get_plot_fun`. With a similar mechanism, a number of statistical methods to characterize or evaluate the results of an analysis can be defined into the `fun stats` parameter, which expects a list of function, result of `get_fun_stats`. The parameter `nbr_top_models` specifies how many top-ranking models will be selected as likely and, thus, be cross-compared for consistency assessment. Finally, a number of statistics to summary the set of BIC scores and the subtypes, may be defined as list of functions into the `fun_bic_pattern` and `fun_pattern` parameters. More details in the man page of `?set_cresult`.

How to carry a statistical inference for subtypes?

Implementation

This scenario is implemented as the R SubtypeDiscovyer package. To implement these different steps, our package relies on three important containers:

1. `cdata` takes as input the data and its `settings`,

2. `cmodel` stores the cluster models,

3. `cresult` stores the whole subtype discovery analysis.

Concluding remarks

References

[Col09b] R SubtypeDiscovery to assist subtyping analyses in clinical research. F Colas, S van Rooden, I Meulenbelt, J J Houwing-Duistermaat, T van Veen and J N Kok in the Bioinformatics journal, Vol X, Number X, XXX 2009.

[Col09a] A Scenario Implementation in R for Subtype Discovery Examplified on Chemoinformatics Data. F Colas and I Meulenbelt and J J Houwing-Duistermaat and M Kloppenburg and I Watt and S M van Rooden and M Visser and H Marinus and Edward O Cannon and Andreas Bender and J J van Hilten and P E Slagboom and J N Kok, 3rd International Symposium on Leveraging Applications of Formal Methods, Verification and Validation (ISOLA'08), Greece.

[Col08] Stability of Clusters for Different Time Adjustments in Complex Disease Research. F Colas and I Meulenbelt and J J Houwing-Duistermaat and M Kloppenburg and I Watt and S M van Rooden and M Visser and H Marinus and J J van Hilten and P E Slagboom and J N Kok, 30th Annual International IEEE EMBS Conference (EMBC'08), Vancouver, Canada

[Fra02] Model-Based Clustering, Discriminant Analysis and Density Estimation. C Fraley and A E Raftery. In Journal of the American Statistical Association, Vol 97, 611-631, 2002.

[Fra06] MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. C Fraley and A E Raftery. In technical report 504, Department of Statistics, University of Washington, September, 2006.