API

Granulometric profiles of 10 batches of an API described by 25 class diameters (D. Copelli, A. Cavecchi, C. Merusi, R. Leardi, “Multivariate evaluation of the effect of the particle size distribution of an Active Pharmaceutical Ingredient on the performance of a pharmaceutical drug product: a real-case study”, *Chemometrics and Intelligent Laboratory Systems*, 178, 1-10 (2018))
The first row contains the class diameter.
The first column contains the name of the batch.

Classification

Artificial data set for classification purposes, made by 40 samples and 10 variables (source: R. Brereton, used in the book "Basic Chemometrics for Analytical Chemists").

Cognac

Sensory data, with five different cognacs tasted by six assessors, each scoring on 12 characteristics (source: Pernod Ricard Research Center).
The first row contains the label of the columns.
The first column contains the code of the product.
The second column contains the name of the assessor (fantasy names).

Colorants

Samples containing mixtures of two colorants (E-102 and E-110), described by 47 absorbances measured every 5 nm in the range 340-570 nm. Samples 1-16 contain known amount of colorants, while rows 17-22 are two samples of different food colorants purchased in a local market (source: L. Sarabia, used in the book "Basic Chemometrics for Analytical Chemists").
The first row contains the name of the variable.
The first column contains the name of the sample.
The second and the third column contain the concentration of the colorants.

DoE

A collection of data from different experimental designs.

Economic data

35 countries described by 75 economic descriptors (source: OECD)
The first row (not to be imported) contains the full name of the descriptor.
The second row contains a shortened name of the descriptor (can be used as header).
The first column contains the name of the country.

Forensic

44 human beings described by 15 variables (the height and 14 measurements made on humerus and femur). (Buikstra J. y Ubelaker H., 1994. Standards for Data Collection from Human Skeletal Remains. *Arkansas Archaeological Survey Research Series* No. 44, Arkansas, used in the book "Basic Chemometrics for Analytical Chemists".)
The first row contains the name of the variable.

Four whiskeys

93 samples from four commercial brands of Irish whiskeys described by 57 chemico-physical variables (K. MacNamara, Irish Distillers, personal communication)
The first row contains the name of the variable.
The first column contains the name of the sample.
The second column contains the full name of the brand.
The third column contains the shortened name of the brand (to be used as label in plots).

Note: inside each brand the samples are reported according to the production/analysis sequence.

Fraud
17 samples of lard, pure or adulterated with tallow, described by 5 fatty acids (percentage areas of the chromatographic peaks) (Sarabia, L. A., Ortiz, M. C., and Checa, M. A. (1989). Pattern recognition for detection of tallow in lard. In *Agriculture, Food Chemistry and the Consumer*, Vol. 2. Paris: l'Institute Nationale de la Recherche Agronomique, pp. 602–606, used in the book "Basic Chemometrics for Analytical Chemists").
The first row contains the name of the variable.
The first column contains the label of the chromatographic column used for the analysis
The second column contains the percentage of tallow.

Milk
250 samples of sheep milk, described by their fatty acid composition (71 variables) (M. Caredda, M. Addis, I. Ibba, R. Leardi, M.F. Scintu, G. Piredda, G. Sanna, "Building of prediction models by using Mid-Infrared spectroscopy and fatty acid profile to discriminate the geographical origin of sheep milk", *LWT- Food Science and Technology*, 75, 131-136 (2017))
Sheet train: 150 samples of the training set.
Sheet test: 100 samples of the test set.
The first row contains the name of the variable.
The first column contains the code of the geographic origin.
The second column contains the code of the season.
The third column contains the code of the year.
The fourth column contains the code of the category obtained combined the three previous codings.

Moisture
54 samples of soy wheat (R. Leardi, A. Lupiáñez González, "Genetic Algorithms applied to feature selection in PLS regression: how and when to use them", *Chemometrics and Intelligent Laboratory Systems*, 41, 195-207 (1998))
moisture content: column 1; NIR spectra: columns 2-176 (1104-2496 nm, step 8 nm);
Rows 1-40: training set; rows 41:54 test set.

Two whiskeys
43 samples from two types of Irish whiskeys described by the concentration of 12 compounds from a GC analysis (K. MacNamara, Irish Distillers, personal communication)
The first row contains the name of the variables.
The first column contains the name of the sample.
The second column contains the type of the sample.
Note: inside each type the samples are reported according to the production/analysis sequence.

Venice
Pollution data from the Venice lagoon, from 16 sampling site on each of which 13 variables have been measured once per month during one year (R. Leardi, M.L. Tercier-Waeber, B. Gianni, G. Ferrari, "Application of 3-way Principal Component Analysis to water quality assessment of the Venice lagoon", *Colloquium Chemiometricum Mediterraneum V,* Ustica, 25-27 June 2003, Book of Abstracts, O23).
Sheet data: the original data.
Sheet datalog: the data after logarithmic transformation.
The first row contains the label of the columns.
The first column contains the code of the sampling station.
The second column contains the month.

The third column contains the type of the sampling station.

Vinegars
84 samples of Spanish vinegars from 4 different types, described by 20 compositional variables (Benito, M. J., Ortiz, M. C., S´anchez, M. S., Sarabia, L. A., and ´Iniguez, M. (1999). Typification of vinegars from Jerez and Rioja using classic chemometric techniques and neural network methods, used in the book "Basic Chemometrics for Analytical Chemists").
The first row contains the label of the columns.
The first column contains the type of the sample.
Note: the data set has already been autoscaled.

Washing
Quality control data of washing machines. The 46 variables (different characteristics at different wavelengths) describe the noise of the engine (not allowed to disclose data origin).
No header, no row names.
Sheet train: 356 samples inside specifications according to the univariate approach (all the variables inside the acceptability interval).
Sheet test: 50 samples outside specifications according to the univariate approach (at least one variable outside the acceptability interval).

Wheat
Granulometric profiles of 41 samples of wheat described by 38 class diameters, from three different producers (not allowed to disclose data origin)
The first row contains the class diameter.
The first column contains the coded name of the producer.

Wines
178 samples from three types of wines from Piedmont described by 13 chemico-physical analyses (M. Forina, C. Armanino, M. Castino, M. Ubigli, “Multivariate data analysis as a discriminating method of the origin of wines”, *Vitis*, 25, 189-201 (1986))
The first row contains the name of the variable.
The first column contains the name of the sample.
The second column contains the category of the sample (OLO = Barolo, GR = Grignolino, ERA = Barbera).

Winesclass
The same data set as Wines, with samples divided into training set and test set
Sheet TRAIN: 90 samples (samples 1-30 of each category); same structure as in Wines.
Sheet TEST: the remaining 88 samples; no first row with the name of the variable.

Whisky
29 mixtures prepared with an high quality whisky, a poor quality whisky and water, described by 91 absorbances recorded in the UV range between 400 and 220 nm, every 2 nm (source: L. Sarabia, used in the book "Basic Chemometrics for Analytical Chemists").
The first row contains the name of the variable.
The first three columns contain the percentage of the three components.

Many more data sets are “embedded” in the packages used by the software.
The list of all of them can be obtained by typing

```
data(package = .packages(all.available = TRUE))
```

In order to “extract” a data set into the workspace type

```
data(dataname,package=”packagename”)
```

As an example, in order to extract the data set environmental, present in the package lattice, type

```
data(environmental,package="lattice")
```

By selecting “Data Handling”, “Workspace Management”, “Tell Me” it can be checked that the data set has been correctly extracted and can be used as such by CAT:

```
[1] ***************************************************************************
[1] Name              :  environmental
[1] Type              :  data.frame
[1] Row               :  111
[1] Column            :  4
[1] NA                :  0
[1] %NA               :  0
[1] Column Names         :
[1] ozone       radiation   temperature wind
[1] Row Names            :
  [1] 1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27
28
 [29] 29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53
54  55  56
 [57] 57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81
82  83  84
 [85] 85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100 101 102 103 104 105 106 107
108 109 110 111
[1] ***************************************************************************
```

In some cases, the structure of these data sets is not directly compatible with CAT.
For instance, the data set gasoline, present in the package pls, apparently has only two variables:

```
[1] ***************************************************************************
[1] Name              :  gasoline
[1] Type              :  data.frame
[1] Row               :  60
[1] Column            :  2
[1] NA                :  0
[1] %NA               :  0
[1] Column Names         :
[1] octane NIR
[1] Row Names            :
 [1] 1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
33 34 35 36 37 38
[39] 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
[1] ***************************************************************************
```

while, when typing gasoline on the console, the whole data set is displayed.
To solve this problem it is enough to export the data set as .txt and then load it again, after which a structure compatible with CAT is obtained.

```
 [1] ******************************************************************************
[1] Name                :  gasoline
[1] Type               :  data.frame
[1] Row                :  60
[1] Column              :  402
[1] NA                :  0
[1] %NA                :  0
[1] Column Names          :
  [1] V1   V2   V3   V4   V5   V6   V7   V8   V9   V10  V11  V12  V13  V14  V15  V16  V17  V18
V19  V20  V21  V22  V23 
 [24] V24  V25  V26  V27  V28  V29  V30  V31  V32  V33  V34  V35  V36  V37  V38  V39  V40
V41  V42  V43  V44  V45  V46 
 [47] V47  V48  V49  V50  V51  V52  V53  V54  V55  V56  V57  V58  V59  V60  V61  V62  V63
V64  V65  V66  V67  V68  V69 
 [70] V70  V71  V72  V73  V74  V75  V76  V77  V78  V79  V80  V81  V82  V83  V84  V85  V86
V87  V88  V89  V90  V91  V92 
 [93] V93  V94  V95  V96  V97  V98  V99  V100 V101 V102 V103 V104 V105 V106 V107 V108
V109 V110 V111 V112 V113 V114 V115
[116] V116 V117 V118 V119 V120 V121 V122 V123 V124 V125 V126 V127 V128 V129 V130
V131 V132 V133 V134 V135 V136 V137 V138
[139] V139 V140 V141 V142 V143 V144 V145 V146 V147 V148 V149 V150 V151 V152 V153
V154 V155 V156 V157 V158 V159 V160 V161
[162] V162 V163 V164 V165 V166 V167 V168 V169 V170 V171 V172 V173 V174 V175 V176
V177 V178 V179 V180 V181 V182 V183 V184
[185] V185 V186 V187 V188 V189 V190 V191 V192 V193 V194 V195 V196 V197 V198 V199
V200 V201 V202 V203 V204 V205 V206 V207
[208] V208 V209 V210 V211 V212 V213 V214 V215 V216 V217 V218 V219 V220 V221 V222
V223 V224 V225 V226 V227 V228 V229 V230
[231] V231 V232 V233 V234 V235 V236 V237 V238 V239 V240 V241 V242 V243 V244 V245
V246 V247 V248 V249 V250 V251 V252 V253
[254] V254 V255 V256 V257 V258 V259 V260 V261 V262 V263 V264 V265 V266 V267 V268
V269 V270 V271 V272 V273 V274 V275 V276
[277] V277 V278 V279 V280 V281 V282 V283 V284 V285 V286 V287 V288 V289 V290 V291
V292 V293 V294 V295 V296 V297 V298 V299
[300] V300 V301 V302 V303 V304 V305 V306 V307 V308 V309 V310 V311 V312 V313 V314
V315 V316 V317 V318 V319 V320 V321 V322
[323] V323 V324 V325 V326 V327 V328 V329 V330 V331 V332 V333 V334 V335 V336 V337
V338 V339 V340 V341 V342 V343 V344 V345
[346] V346 V347 V348 V349 V350 V351 V352 V353 V354 V355 V356 V357 V358 V359 V360
V361 V362 V363 V364 V365 V366 V367 V368
[369] V369 V370 V371 V372 V373 V374 V375 V376 V377 V378 V379 V380 V381 V382 V383
V384 V385 V386 V387 V388 V389 V390 V391
[392] V392 V393 V394 V395 V396 V397 V398 V399 V400 V401 V402
[1] Row Names            :
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
33 34 35 36 37 38
[39] 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
[1] ******************************************************************************
```