Application to Drug Consumption data

Mark de Rooij *Methodology and Statistics Unit, Leiden University*
Patrick Groenen *Econometric Institute, Erasmus University*

This is a document describing the MLD model and a MM algorithm.

Introduction

Fehrman E., Muhammad A.K., Mirkes E.M., Egan V., Gorban A.N. (2017) The Five Factor Model of Personality and Evaluation of Drug Consumption Risk. In: Palumbo F., Montanari A., Vichi M. (eds) Data Science. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham

The drug consumption data has records for 1885 respondents. For each respondent 9 attributes are measured: Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), age, gender. [1]

In addition, participants were questioned concerning their use of 18 legal and illegal drugs and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug patricipants had to indicate whether they never used the drug, used it over a decade ago, in the last decade, in the last year, month, week, or day. In our analysis we coded whether participants used the particular drug in the last year. Furthermore, we focussed on the drugs that had a minimal percentage of 10% and a maximum of 90%, which are Amphetamine, Benzodiazepine, Cannabis, Cocaine, Extasy, Ketamine, legal highs consumption, LSD, Methadone, Mushrooms, and Nicotine.

[1] Also level of education, ethnicity, and country of origin are available in the original data base.

Analysis

```
setwd("~/surfdrive/MLDM/mldm2")
source("melodic.R")

drugdat <- read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/00373/d
    sep = ",")
for (v in 14:32) {
    drugdat[, v] = ifelse(drugdat[, v] == "CL3", 1, ifelse(drugdat[, v] == "CL4",
        1, ifelse(drugdat[, v] == "CL5", 1, ifelse(drugdat[, v] == "CL6", 1, 0))))
}

# add variable names
colnames(drugdat) = c("id", "age", "gender", "educ", "country", "ethnic", "N", "E",
    "O", "A", "C", "impulse", "SS", "Alcohol", "Am", "Amyl", "Be", "Caff", "Ca",
    "Choc", "Co", "Crack", "Ex", "Heroin", "Ke", "Le", "LSD", "Me", "Mu", "Ni", "Semer",
    "VSA")

X = as.matrix(drugdat[, c(2, 3, 7:13)])
Y = as.matrix(drugdat[, 14:32])
idx = which(colMeans(Y) > 0.1 & colMeans(Y) < 0.9)
Y = Y[, idx]
```

The first step in the analysis is to select the dimensionality. We fit models in one till 7 dimensions and compute the AIC statistic for comparison.

```
AIC.dim
```

```
##      Dimensionality Deviance #param      AIC      BIC
## [1,]              1 18311.76     30 18371.76 18538.01
## [2,]              2 18117.49     48 18213.49 18479.49
## [3,]              3 18030.36     65 18160.36 18520.57
## [4,]              4 17998.55     81 18160.55 18609.43
## [5,]              5 17987.31     96 18179.31 18711.31
## [6,]              6 17980.86    110 18200.86 18810.45
## [7,]              7 17975.78    123 18221.78 18903.41
```

We can see that either the two- or three dimensional solution is optimal. Further let us check the influence of the predictor variables in the two dimensional solution.

```
AIC.pred
```

```
##       Left Out Deviance #param      AIC      BIC
##  [1,]        1 19303.36     46 19395.36 19650.28
##  [2,]        2 18417.77     46 18509.77 18764.69
##  [3,]        3 18181.24     46 18273.24 18528.16
##  [4,]        4 18134.47     46 18226.47 18481.39
##  [5,]        5 18449.94     46 18541.94 18796.86
##  [6,]        6 18137.69     46 18229.69 18484.61
##  [7,]        7 18172.36     46 18264.36 18519.28
##  [8,]        8 18121.73     46 18213.73 18468.65
##  [9,]        9 18409.45     46 18501.45 18756.37
```

Interpretation

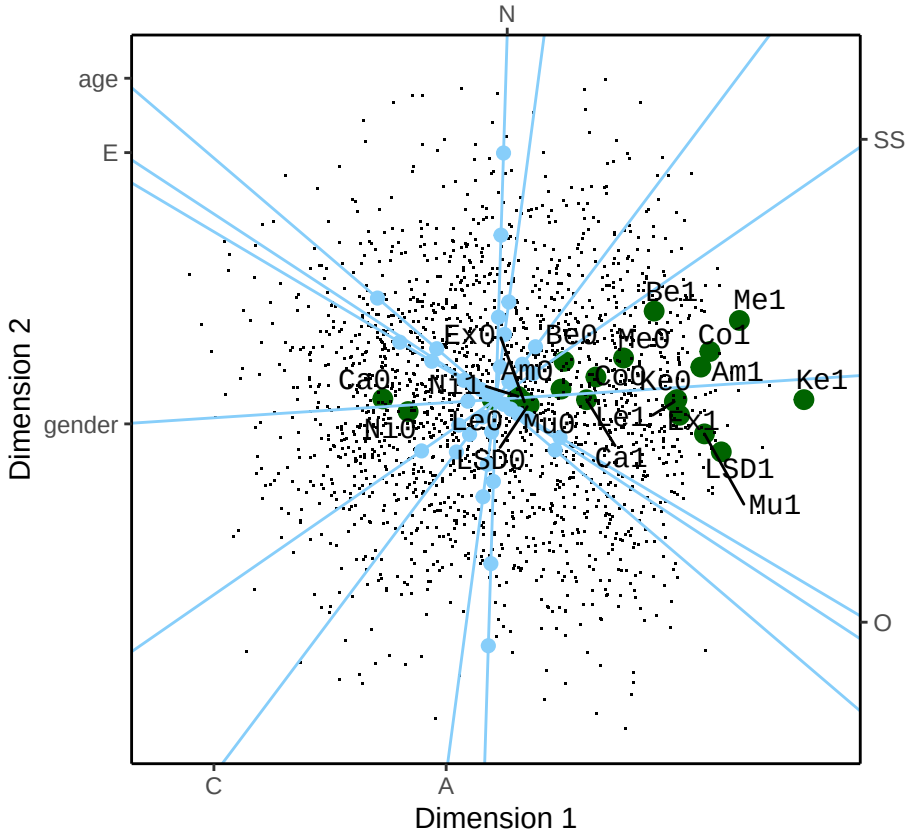
```
p = plot.mldm(out2h, dec.lines = FALSE)

p
```

```
## Warning in min(x): no non-missing arguments to min; returning Inf

## Warning in max(x): no non-missing arguments to max; returning -Inf

## Warning: Removed 12 rows containing missing values (geom_point).
```




```
ggsave(filename = "~/surfdrive/MLDM/mldm2/Figures/drug_mldm.pdf", plot = p, width = 11.7
    height = 8.3, units = "in", limitsize = FALSE)
```

```
## Warning in min(x): no non-missing arguments to min; returning Inf
```

```
## Warning in max(x): no non-missing arguments to max; returning -Inf
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
```

Further let us look at the logistic regression coefficients. Since we standardized the preidtcor variables these are changes in log odds for one standard deviations increases in the predictors.

```
out2h$LRcoef
```

```
##                  Am         Be          Ca         Co          Ex          Ke
## age    -0.58785391 -0.2288500 -0.99274792 -0.4476326 -0.80964120 -0.61737083
## gender -0.32583391 -0.2153792 -0.46959733 -0.2666443 -0.35586438 -0.29281717
## N       0.17749989  0.3574361  0.03706066  0.1955156 -0.05809466  0.02559945
## E      -0.07634901 -0.0347515 -0.12435383 -0.0591901 -0.09987704 -0.07737777
## O       0.33334615  0.1583974  0.53686278  0.2598250  0.42907290  0.33411802
## A      -0.10803750 -0.1656826 -0.06981966 -0.1081444 -0.01907479 -0.04451373
## C      -0.18292350 -0.1697827 -0.21910954 -0.1599237 -0.14850288 -0.13713249
## SS      0.47343435  0.3813101  0.62003549  0.4017427  0.44532969  0.38733187
##                  Le         LSD         Me          Mu          Ni
## age    -0.89503203 -1.13045408 -0.4063956 -0.97190354 -0.47039694
## gender -0.42002972 -0.44060708 -0.2725631 -0.39975566 -0.25562120
## N       0.02251780 -0.26436008  0.2767808 -0.15906709  0.12539513
## E      -0.11192370 -0.13625631 -0.0554689 -0.11833562 -0.06080378
## O       0.48293785  0.58089601  0.2457457  0.50619570  0.26508948
## A      -0.05827611  0.05193361 -0.1407466  0.01540237 -0.07931689
## C      -0.19381807 -0.14470095 -0.1791300 -0.14772637 -0.14068592
## SS      0.55156165  0.49658606  0.4325632  0.47354566  0.36747057
```

We can verify how well each response variable is represented in the low dimensional space. Therefore we define a measure called Quality of Representation, Q_r which is defined by

$$Q_r = (L_{(0,r)} - L_r)/(L_{(0,r)} - L_{lr}),$$

where $L_{(0,r)}$ is the deviance of the intercept only logistic regression model for response variable r, L_r is the part of the loss function for our new model, and L_{lr} is the deviance from the logistic regression with the same predictor variables $\mathbf{X}$.

```
mldm.diag(out2h)
```

```
##          [,1]
## Am  0.9958998
## Be  0.9648054
## Ca  0.9700373
## Co  0.8983786
## Ex  0.9555837
## Ke  0.9547956
## Le  0.9957183
## LSD 0.9823361
## Me  0.9429159
## Mu  0.9929817
## Ni  0.9811682
```

We see that most response variables are well represented. The response variable 'cocaine' is worst rerpesented with only 89.8% recovered deviance.