Code for chapter: Benefit of Bayesian Clustering of Longitudinal Data: Study of Cognitive Decline for Precision Medicine

Anais Rouanet

02/10/2020

This file provides the code used in the chapter:

Rouanet, A., Richardson, S., & Tom, B. D. M. (2020). Benefit of Bayesian clustering of longitudinal data: study of cognitive decline for precision medicine. In Bayesian Methods in Pharmaceutical Research (pp. 223-242).

Load the following libraries:

```r
library("NormPsy")
library("lcmm")
library("remotes")
# remotes::install_github('anarouanet/PReMiuMar')
# #install this one from github
library(PReMiuMar)
library(ggplot2)
library(patchwork)
```

Data were leveraged from the ADNI study. We selected the following variables:

- ID: subject identifier
- Ventricles.bl: ventricule volumes at baseline
- Hippocampus.bl: hippocampus volume at baseline
- WholeBrain.bl: whole brain volume at baseline
- Entorhinal.bl: entorhinal volume at baseline
- Fusiform.bl: volume of fusiform girus at baseline
- MidTemp.bl: volume of mid-temporal girus at baseline
- ICV.bl: Intracranial volume at baseline
- Age: Age at baseline
- PTEDUCAT: years of education
- APOE4: APOE4 status (0, 1 or 2 alleles)
- DX.bl: dementia status at baseline
- PTGENDER: gender
- time: age in decades, centered in 55 years old
- MMSE: Mini Mental State Examination score

We randomly selected 199 subjects who are representative of the overall sample with respect to gender and baseline disease state (cognitively normal - CN, early mild cognitive impairment - EMCI, late mild cognitive impairment - LMCI, subjective memory complaints - SMC).

```r
head(covariables_select)
```

```
##     ID Ventricles.bl Hippocampus.bl WholeBrain.bl Entorhinal.bl Fusiform.bl
## 12   3         84599           5319       1129834          1791       15506
## 23   5         34062           7075       1116633          4433       24788
## 66  15         33420          6732        942730          4307       14953
## 135 31         25669          7206        921781          3227       13595
## 188 42         48933          4087        952780          2784       16454
## 300 58         23647          7987       1014209          3489       17461
##     MidTemp.bl  ICV.bl  AGE PTEDUCAT APOE4 DX.bl PTGENDER
## 12       18422 1920691 81.3       18     1    AD        0
## 23       21614 1640766 73.7       16     0    CN        0
## 66       17273 1500995 80.8       18     1    CN        0
## 135      20044 1341605 77.7       18     0    CN        1
## 188      16009 1519691 72.8       18     0  LMCI        0
## 300      21620 1432548 70.1       16     1    CN        0
```

```
head(ydata_select)
```

```
##    ID     time MMSE
## 12  3 2.630000   20
## 13  3 2.679829   24
## 14  3 2.729932   17
## 16  3 2.829863   19
## 23  5 1.870000   29
## 24  5 1.920103   29
```

We transformed the MMSE outcome using the normalizing function proposed by Philipps et al. (2014), categorized the Education variable and defined standardized volumetric variables by dividing the 6 baseline imaging variables (Ventricles.bl, WholeBrain.bl, Entorhinal.bl, Fusiform.bl, MidTemp.bl) by the intracranial volume. Finally, we center and reduce these 6 variables.

```
ydata_select$outcome <- normMMSE(ydata_select$MMSE)

covariables_select$Educ <- as.factor(sapply(covariables_select$PTEDUCAT,
    function(x) ifelse(x < 16, 0, 1)))
covariables_select$Educ <- as.factor(covariables_select$Educ)

covariables_select$APOE4 <- as.factor(covariables_select$APOE4)
covariables_select$PTGENDER <- as.factor(covariables_select$PTGENDER)

covariables_select$Ventricles_ICV.bl <- covariables_select$Ventricles.bl/covariables_select$ICV.bl
covariables_select$Hippocampus_ICV.bl <- covariables_select$Hippocampus.bl/covariables_select$ICV.bl
covariables_select$Entorhinal_ICV.bl <- covariables_select$Entorhinal.bl/covariables_select$ICV.bl
covariables_select$Fusiform_ICV.bl <- covariables_select$Fusiform.bl/covariables_select$ICV.bl
covariables_select$MidTemp_ICV.bl <- covariables_select$MidTemp.bl/covariables_select$ICV.bl
covariables_select$WholeBrain_ICV.bl <- covariables_select$WholeBrain.bl/covariables_select$ICV.bl

covv <- covariables_select[c("Ventricles_ICV.bl", "Hippocampus_ICV.bl",
    "Entorhinal_ICV.bl", "Fusiform_ICV.bl", "MidTemp_ICV.bl", "WholeBrain_ICV.bl")]
covariables_select[c("Ventricles_ICV.bl", "Hippocampus_ICV.bl", "Entorhinal_ICV.bl",
    "Fusiform_ICV.bl", "MidTemp_ICV.bl", "WholeBrain_ICV.bl")] <- apply(covv,
    2, function(x) (x - mean(x))/sqrt(var(x)))
```

The normalising function can be represented as follows:

```
x <- seq(0, 30, 1)
y <- normMMSE(x)
ggplot(data.frame(MMSE = x, normMMSE = y)) + geom_line(aes(x = MMSE,
  y = normMMSE), size = 2) + labs(y = "Normalised MMSE") +
  theme(axis.title.y = element_text(size = 20, angle = 90)) +
  labs(x = "MMSE") + theme(axis.title.x = element_text(size = 20)) +
  scale_x_continuous(breaks = seq(0, 30, 5)) + scale_y_continuous(breaks = seq(0,
  100, 10))
```

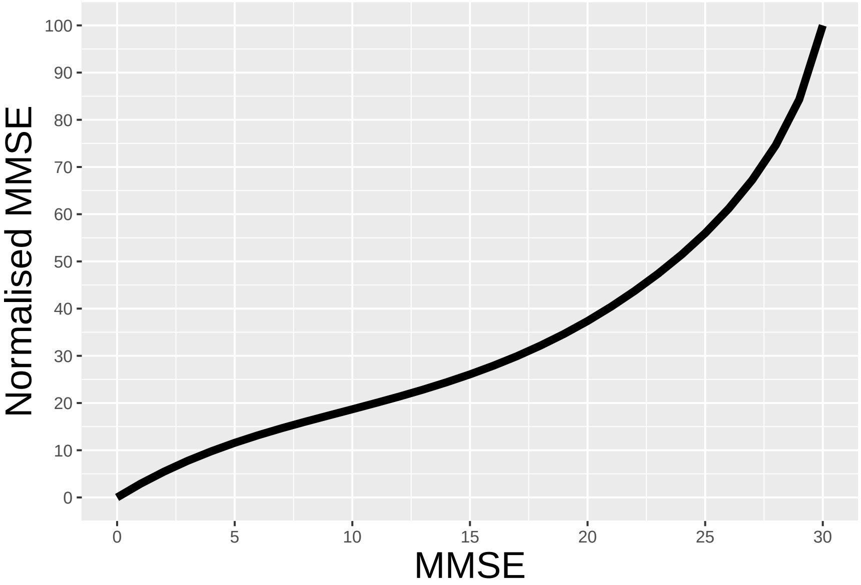

FIGURE 11.1: Normalising transformation for MMSE

The longitudinal data are displayed below in the MMSE and normalized MMSE scores:

```
age <- ydata_select$time * 10 + 55
data_plot <- cbind.data.frame(normMMSE = ydata_select$outcome,
  MMSE = ydata_select$MMSE, subjects = ydata_select$ID,
  Age = age, Delay = ydata_select$time)
plot_normMMSE <- ggplot(data = data_plot) + ylab("Normalised MMSE") +
  geom_line(aes(y = normMMSE, x = Age, group = subjects)) +
  theme_bw() + theme(axis.text = element_text(size = 12),
  axis.title = element_text(size = 16))

plot_MMSE <- ggplot(data = data_plot) + ylab("MMSE") +
  geom_line(aes(y = MMSE, x = Age, group = subjects)) +
  theme_bw() + theme(axis.text = element_text(size = 12),
  axis.title = element_text(size = 16))

plot_MMSE + plot_normMMSE
```

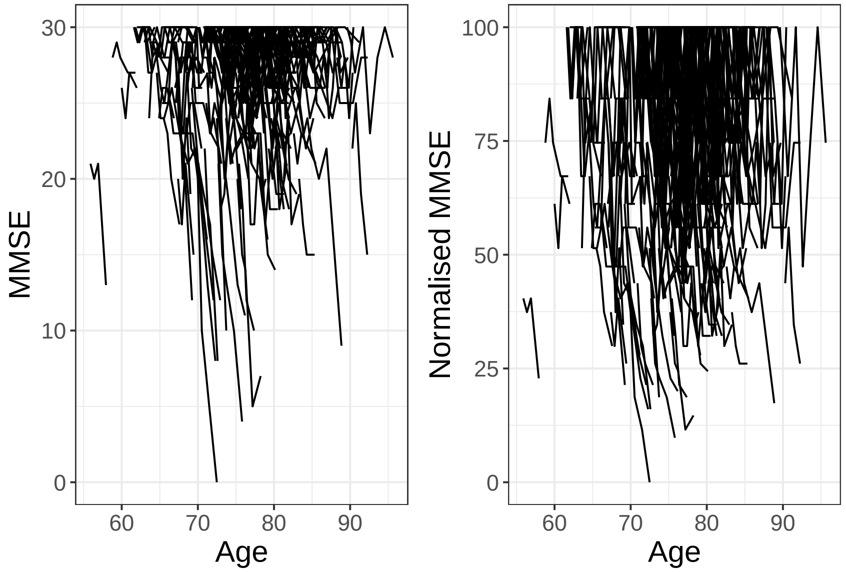

FIGURE 11.2: ADNI cohort: Observed cognitive trajectories of the 199 selected subjects on both the original and normalized MMSE scales.

11.4 Standard frequentist analysis: Latent class mixed models

We create ydata_lcmm dataset for the latent class mixed model, which has a long format (one line per observation). We specify a quadratic trend for the normalized MMSE trajectories, and the model is adjusted on a practice effect (learn, equal to 1 at the first visit and 0 otherwise), education (Educ), gender (PTGENDER) and APOE4 status (APOE4).

```
ydata_lcmm <- merge(ydata_select, covariables_select[,
  c("ID", "Educ", "APOE4", "PTGENDER")], by = "ID")
ydata_lcmm$Educ <- ifelse(as.numeric(as.character(ydata_lcmm$Educ)) >
  0, 1, 0)
ydata_lcmm$APOE4 <- ifelse(as.numeric(as.character(ydata_lcmm$APOE4)) >
  0, 1, 0)

ydata_lcmm$learn <- rep(0, dim(ydata_lcmm)[1])
unique_ID_lcmm_select <- sapply(unique(ydata_lcmm$ID),
  function(x) min(which(ydata_lcmm$ID == x)))
ydata_lcmm$learn[unique_ID_lcmm_select] <- rep(1, length(unique_ID_lcmm_select))
head(ydata_lcmm)
```

```
##   ID     time MMSE outcome Educ APOE4 PTGENDER learn
## 1  3 2.630000   20   37.37    1     1        0     1
## 2  3 2.679829   24   51.44    1     1        0     0
## 3  3 2.729932   17   29.93    1     1        0     0
## 4  3 2.829863   19   34.64    1     1        0     0
## 5  5 1.870000   29   84.32    1     0        0     1
## 6  5 1.920103   29   84.32    1     0        0     0
```

We then run the latent class mixed models for 1 to 4 classes:

```
M1 <- lcmm(fixed = outcome ~ time + I(time^2) + Educ +
  APOE4 + PTGENDER + learn, random = ~time + I(time^2),
  subject = "ID", ng = 1, idiag = F, link = "linear",
  data = ydata_lcmm)

M2 <- lcmm(fixed = outcome ~ time + I(time^2) + Educ +
  APOE4 + PTGENDER + learn, mixture = ~time + I(time^2) +
  Educ + learn, random = ~time + I(time^2), subject = "ID",
  ng = 2, idiag = F, nwg = T, link = "linear", data = ydata_lcmm,
  maxiter = 300)

M3 <- lcmm(fixed = outcome ~ time + I(time^2) + Educ +
  APOE4 + PTGENDER + learn, mixture = ~time + I(time^2) +
  Educ + learn, random = ~time + I(time^2), subject = "ID",
  ng = 3, idiag = F, nwg = T, link = "linear", data = ydata_lcmm,
  maxiter = 300)

M4_1 <- lcmm(fixed = outcome ~ time + I(time^2) + Educ +
  APOE4 + PTGENDER + learn, mixture = ~time + I(time^2) +
  learn, random = ~time + I(time^2), subject = "ID",
  ng = 4, idiag = F, nwg = T, link = "linear", data = ydata_lcmm,
  maxiter = 500)

B <- rep(0, length(M4_1$best) + 3)
B[-c(15:18)] <- M4_1$best[-which(names(M4_1$best) ==
  "Educ")]

M4_2 <- lcmm(fixed = outcome ~ time + I(time^2) + Educ +
  APOE4 + PTGENDER + learn, mixture = ~time + I(time^2) +
  Educ + learn, random = ~time + I(time^2), subject = "ID",
  ng = 4, idiag = F, nwg = T, link = "linear", data = ydata_lcmm,
  B = B, maxiter = 500)
```

The BIC for the four models are:

```
knitr::kable(data.frame(M1 = M1$BIC, M2 = M2$BIC, M3 = M3$BIC,
  M4 = M4_2$BIC))
```

M1	M2	M3	M4
9197.371	9171.646	9184.643	9209.561

We choose the two-class model with the lowest BIC value and display the estimated normalized MMSE trajectories. Note that we switch class labels (1 and 2) to have class 1 with the steepest cognitive decline.

```
time <- seq(min(ydata_lcmm$time), max(ydata_lcmm$time),
  by = 0.05)
profile <- data.frame(time = time, learn = c(1, rep(0,
  length(time) - 1)), Educ = rep(0, length(time)),
  APOE4 = rep(0, length(time)), PTGENDER = as.factor(rep(0,
    length(time))))

pred_m2 <- predictY(M2, profile, var.time = "time",
  draws = TRUE)
age <- time * 10 + 55
```

```
prediction <- data.frame(pred_m2$pred)
names(prediction)
prediction$age <- age

prediction2 <- cbind.data.frame(normMMSE = c(prediction$Ypred_50_class1,
  prediction$Ypred_50_class2), Class = c(rep("2",
  nrow(prediction)), rep("1", nrow(prediction))),
  Age = c(prediction$age, prediction$age), CI_inf = c(prediction$Ypred_2.5_class1,
    prediction$Ypred_2.5_class2), CI_sup = c(prediction$Ypred_97.5_class1,
    prediction$Ypred_97.5_class2))
prediction2$Class <- factor(ifelse(prediction2$Class ==
  1, 2, 1))

ggplot(data = prediction2) + geom_line(aes(y = normMMSE,
  x = Age, color = Class, linetype = Class), size = 1.3) +
  ylab("Normalised MMSE") + geom_ribbon(aes(ymin = CI_inf,
  ymax = CI_sup, x = Age, fill = Class), alpha = 0.2) +
  theme_bw() + theme(axis.text = element_text(size = 12),
  axis.title = element_text(size = 16))
```

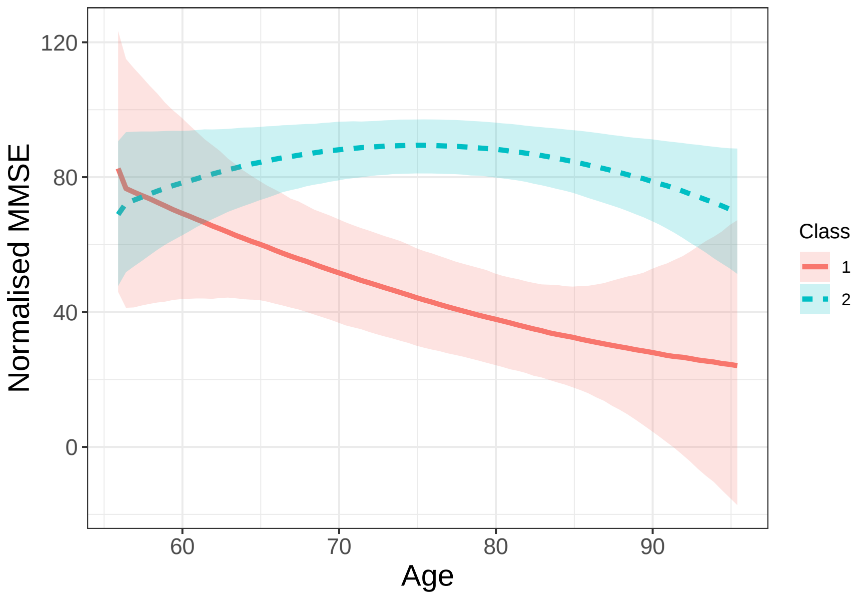

Figure 11.5 : Class-specific trajectories estimated by the two-latent class mixed model, on the normalized MMSE scale, as a function of age for a man with no APOE4 alleles and fewer than 16 years of education. The shaded regions represent 95% confidence bands.

The estimates are displayed below:

```
summary(M2)
```

	coef	Se	Wald	p-value
intercept class1 (not estimated)	0.00000	NA	NA	NA
intercept class2	0.98891	2.52321	0.392	0.69511
time10 class1	1.84891	0.92611	1.996	0.04589
time10 class2	-2.06167	1.90557	-1.082	0.27929
I(time10ˆ2) class1	-0.45742	0.21192	-2.158	0.03089
I(time10ˆ2) class2	0.19274	0.47119	0.409	0.68250

	coef	Se	Wald	p-value
Educ class1	0.17924	0.40941	0.438	0.66153
Educ class2	1.73367	0.67291	2.576	0.00998
APOE4	-0.55587	0.17771	-3.128	0.00176
PTGENDER1	0.30744	0.17098	1.798	0.07215
learn class1	-0.23557	0.13454	-1.751	0.07996
learn class2	0.45668	0.13968	3.269	0.00108

Table 11.1 : Regression parameter estimates, standard errors and P-values from the two- latent-class mixed model.

The imaging variables are compared across classes:

```
postclas <- M2$pprob$class
cov_X_ICV <- c("Ventricles_ICV.bl", "Hippocampus_ICV.bl",
  "Entorhinal_ICV.bl", "Fusiform_ICV.bl", "MidTemp_ICV.bl",
  "WholeBrain_ICV.bl")

pval <- apply(covariables_select[, cov_X_ICV], 2, function(x) {
  t.test(x ~ postclas, var.equal = FALSE, mu = 0,
    alternative = "two.sided")$p.value
})

Table11_2 <- data.frame(cbind(apply(covariables_select[,
  cov_X_ICV][postclas == 2, ], 2, mean), apply(covariables_select[,
  cov_X_ICV][postclas == 2, ], 2, function(x) sqrt(var(x))),
  apply(covariables_select[, cov_X_ICV][postclas ==
    1, ], 2, mean), apply(covariables_select[,
    cov_X_ICV][postclas == 1, ], 2, function(x) sqrt(var(x))),
  pval))
names(Table11_2) <- c("mean Class 1", "sd Class 1",
  "mean Class 2", "sd Class 2", "P-value")
knitr::kable(Table11_2)
```

	mean Class 1	sd Class 1	mean Class 2	sd Class 2	P-value
Ventricles_ICV.bl	0.2020882	1.1099154	-0.1883541	0.8482599	0.0061388
Hippocampus_ICV.bl	-0.4694194	0.9361804	0.4375171	0.8509992	0.0000000
Entorhinal_ICV.bl	-0.4688260	0.9758921	0.4369640	0.8090065	0.0000000
Fusiform_ICV.bl	-0.3296411	1.0148782	0.3072383	0.8856023	0.0000049
MidTemp_ICV.bl	-0.3308281	1.0656000	0.3083446	0.8273948	0.0000051
WholeBrain_ICV.bl	-0.2625470	1.0281456	0.2447040	0.9117537	0.0003126

Table 11.2: Description of the standardized imaging markers using their means (standard deviations) by the two latent classes, with associated Student's 2-sample t-test P-values for comparing between classes.

11.5.1 Profile regression analysis: Integrative analysis of summarized cognitive and imaging data —-

We obtain individual random effects from a linear mixed model

```
mod_lme_age10 <- lcmm(outcome ~ I(time * 100), random = ~I(time *
  100), ng = 1, subject = "ID", data = ydata_lcmm,
  maxiter = 200)
```

```
cov_X_ICV <- c("Ventricles_ICV.bl", "Hippocampus_ICV.bl",
  "Entorhinal_ICV.bl", "Fusiform_ICV.bl", "MidTemp_ICV.bl",
  "WholeBrain_ICV.bl")
covariables_select$Gender <- as.numeric(covariables_select$PTGENDER)
covariables_select$Educ <- covariables_select$Educ_2
covariables_select$APOE4 <- ifelse(covariables_select$APOE4 ==
  0, 0, 1)

data_combine_RE_age10_obs <- cbind(covariables_select,
  mod_lme_age10$predRE[, 2:3] * mod_lme_age10$best["Linear 2 (std err)"])
names(data_combine_RE_age10_obs)[(dim(data_combine_RE_age10_obs)[2] -
  1):dim(data_combine_RE_age10_obs)[2]] <- c("outcome1",
  "outcome2")
head(data_combine_RE_age10_obs)
```

```
##     ID Ventricles.bl Hippocampus.bl WholeBrain.bl Entorhinal.bl Fusiform.bl
## 12   3         84599           5319       1129834          1791       15506
## 23   5         34062           7075       1116633          4433       24788
## 66  15         33420           6732        942730          4307       14953
## 135 31         25669           7206        921781          3227       13595
## 188 42         48933           4087        952780          2784       16454
## 300 58         23647           7987       1014209          3489       17461
##     MidTemp.bl  ICV.bl  AGE PTEDUCAT APOE4 DX.bl PTGENDER Ventricles_ICV.bl
## 12       18422 1920691 81.3       18     1    AD        0         1.3137949
## 23       21614 1640766 73.7       16     0    CN        0        -0.4876686
## 66       17273 1500995 80.8       18     1    CN        0        -0.3712077
## 135      20044 1341605 77.7       18     0    CN        1        -0.6135170
## 188      16009 1519691 72.8       18     0  LMCI        0         0.3973075
## 300      21620 1432548 70.1       16     1    CN        0        -0.8166759
##     Hippocampus_ICV.bl Entorhinal_ICV.bl Fusiform_ICV.bl MidTemp_ICV.bl
## 12         -2.00330195        -2.5645218      -1.7841331     -1.5918542
## 23         -0.07756219         0.8902099       2.3988001      0.3413549
## 66          0.13841097         1.2175473      -0.6609127     -0.5575288
## 135         1.24459538         0.3113491      -0.5590341      1.2951758
## 188        -2.10310772        -0.8082193      -0.1464664     -1.0828635
## 300         1.49950119         0.3703035       0.6631748      1.3770421
##     WholeBrain_ICV.bl Gender   outcome1   outcome2
## 12         -1.4728650      1   9.491943 -13.235750
## 23          0.4201089      1  -5.762515  10.798060
## 66         -0.6561773      1  -9.777402  11.141363
## 135         0.5537592      2 -14.820652  16.907649
## 188        -0.6790116      1  22.786303 -19.884424
## 300         0.9823878      1   5.646065   3.599809
```

The MVN profile regression is run as follows, specifying the profile variables (covNames, including the discrete ones-discreteCovs, and the continuous ones-continuousCovs) as well as the outcomes, provided in the data dataframe (one line per subject).

```
runInfoObj_combine_RE_age10 <- PReMiuMar::profRegr(yModel = "MVN",
  xModel = "Mixed", nSweeps = 50000, nBurn = 1000,
  data = data_combine_RE_age10_obs, discreteCovs = c("Gender",
    "Educ", "APOE4"), continuousCovs = cov_X_ICV,
  outcome = c("outcome1", "outcome2"), output = "output/output",
  covNames = c(cov_X_ICV, "Gender", "Educ", "APOE4"),
  outcomeT = NA, nClusInit = 30, run = TRUE, nProgress = 500,
  seed = 1567)
```

```
runInfoObj <- runInfoObj_combine_RE_age10
runInfoObj$directoryPath
runInfoObj$fileStem <- "output_obs"
dissimObj <- calcDissimilarityMatrix(runInfoObj)
clusObj <- calcOptimalClustering(dissimObj)
clusObjMVN <- clusObj
table(clusObj$clustering)
myheatDissMat(dissimObj)
```

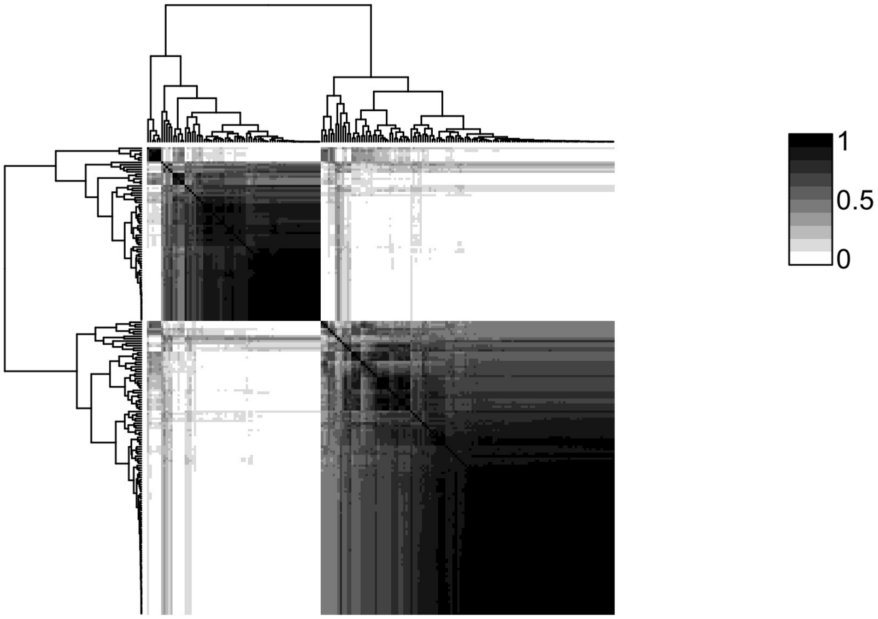

FIGURE 11.6: Posterior similarity matrix obtained by profile regression on random intercepts and slopes and profile variables. This identified 2 clusters comprising 70 (35.2%) and 129 subjects (64.8%) respectively.

The estimated cluster-specific parameters are computed by averaging over the clusterings sampled at each iteration, and allow to plot the outcome and variable profiles using the following functions:

```
clusObj$nOutcomes <- 2
riskProfileObj <- calcAvgRiskAndProfile_AR(clusObj,
  nSweeps1 = nSweeps)
# Figure 11.7
clusterOrderObj_chapter <- plotRiskProfile_AR_chapter(riskProfileObj,
  "Figure11_7.pdf", nSweeps1 = nSweeps)
# Figure 11.8
clusterOrderObj_chapter2 <- plot_trajectories_AR_chapter(riskProfileObj,
  "Figure11_8.png", nSweeps1 = nSweeps)
```

11.5.2 Integrative analysis of longitudinal cognitive and imaging data —-

The GP profile regression is run as follows, specifying the profile variables (covNames, including the discrete ones-discreteCovs, and the continuous ones-continuousCovs) provided in the data dataframe (one line per subject), the outcome (outcome) provided in the longData dataframe (one line per observation).

```
head(ydata_select)
```

```
##    ID     time MMSE outcome
## 12  3 2.630000   20   37.37
## 13  3 2.679829   24   51.44
## 14  3 2.729932   17   29.93
## 16  3 2.829863   19   34.64
## 23  5 1.870000   29   84.32
## 24  5 1.920103   29   84.32
```

```
runInfoObj_200s_1000s__cov_X_ICV <- PReMiuMar::profRegr(yModel = "Longitudinal",
  xModel = "Mixed", nSweeps = 10000, nBurn = 5000,
  data = data_combine_RE_age10_obs, longData = ydata_select,
  discreteCovs = c("Gender", "Educ", "APOE4"), continuousCovs = cov_X_ICV,
  outcome = c("outcome"), output = "output_long/output",
  covNames = c(cov_X_ICV, "Gender", "Educ", "APOE4"),
  outcomeT = NA, nClusInit = 30, run = TRUE, nProgress = 500,
  seed = 1567)
```

The following functions provide the dissimilarity matrix and the output plots:

```
runInfoObj <- runInfoObj_200s_1000s__cov_X_ICV

dissimObj <- calcDissimilarityMatrix(runInfoObj)
clusObj <- calcOptimalClustering(dissimObj)
clusObjGP <- clusObj
myheatDissMat(dissimObj)
```

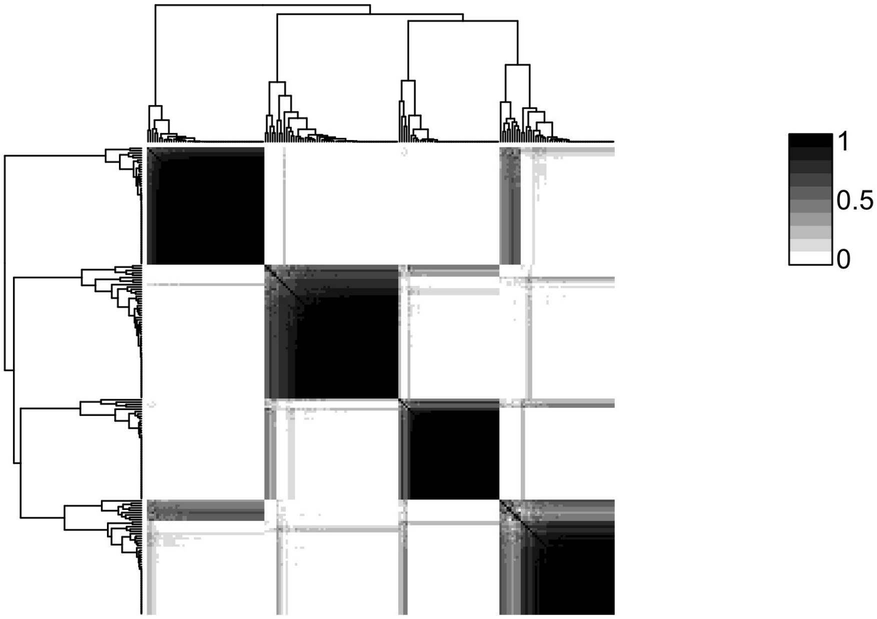

FIGURE 11.9: Posterior similarity matrix obtained by profile regression on repeated normal- ized MMSE scores and volumetric imaging biomarkers, identifying 4 clusters of 57 (28.6%), 55 (27.6%), 44 (22.1%) and

43 (21.6%) subjects, respectively.

The estimated cluster-specific trajectories and variable patterns are obtained using:

```r
clusObj$nOutcomes <- 2
riskProfileObj <- calcAvgRiskAndProfile_AR(clusObj,
  nSweeps1 = nSweeps)
# Figure 11.10
clusterOrderObj_chapter2 <- plot_trajectories_AR_chapter(riskProfileObj,
  "Figure11_10.png", nSweeps1 = nSweeps)
# Figure 11.11
clusterOrderObj_chapter <- plotRiskProfile_AR_chapter(riskProfileObj,
  "Figure11_11.pdf", nSweeps1 = nSweeps)
```

Finally, we compare the two clusterings, considering the individual random intercepts and slopes as outcomes (MVN) or the repeated normalized MMSE scores (GP):

```r
post_class_MVN <- clusObjMVN$clustering
post_class_GP <- clusObjGP$clustering
Table <- table(post_class_GP, post_class_MVN)
Table <- as.matrix(Table)
Table11_4 <- data.frame(Cluster1 = Table[, 1], Cluster2 = Table[,
  2], Total = apply(Table, 1, sum))
Table11_4 <- rbind(Table11_4, apply(Table11_4, 2, sum))
row.names(Table11_4) <- c("Cluster1", "Cluster2", "Cluster3",
  "Cluster4", "Total")
knitr::kable(Table11_4)
```

	Cluster1	Cluster2	Total
Cluster1	33	24	57
Cluster2	2	53	55
Cluster3	6	38	44
Cluster4	29	14	43
Total	70	129	199

TABLE 11.4: Cross-tabulation of the clustering structures obtained by profile regression based on a GP model (rows: Clusters 1 to 4) and the two-stage profile regression approach (columns: Clusters 1 and 2).