require(mclust)

##### summaryMclust: function which returns information about model fit and classification with gaussian mixture models

summaryMclust <- function(mdl, data, G) {

mod <- Mclust(data, modelNames = mdl, G = G)

return(data.frame(Variance = rep(mdl, G), G = 1:G, log.likelihood = rep(mod$loglik, G), BIC = rep(-1.75\*mod$loglik+log(length(data))\*mod$df, G),

prop.classes = mod$parameters$pro, mean = mod$parameters$mean, sdv = sqrt(mod$parameters$variance$sigmasq), stringsAsFactors = FALSE))

#return(list(mdl = mod,

# Assessment = data.frame(Variance = mdl, G = G, log.likelihood = round(mod$loglik, 2), BIC = round(-2\*mod$loglik+log(length(data))\*mod$df, 2), prop.classes = paste(round(mod$parameters$pro, 3), collapse = "-"), mean = paste(round(mod$parameters$mean, 3), collapse = "-"), sdv = paste(round(sqrt(mod$parameters$variance$sigmasq), 3), collapse = "-"), stringsAsFactors = FALSE),

# density = densityMclust(data, modelNames = mdl, G = G)))

}

summaryFitThresholds <- function(data, thresh, leg) {

# Manual split with fixed thresholds on V

data <- na.omit(data)

ind.1 <- which(data <= thresh)

ind.2 <- which(data > thresh)

mean.1 <- mean(data[ind.1])

mean.2 <- mean(data[ind.2])

sd.1 <- sd(data[ind.1])

sd.2 <- sd(data[ind.2])

n <- length(data)

res <- data.frame(legend = rep(leg, 2), group = c(paste("<=", thresh, sep = ""), paste(">", thresh, sep = "")),

nb = c(length(ind.1), length(ind.2)), percent = 100\*c(length(ind.1), length(ind.2))/n, mean = c(mean.1, mean.2), sd = c(sd.1, sd.2))

return(res)

}

##### Read S452D data #####

test <- read.table("S452D.csv", sep = ",", h = T, dec = ".")

test\_keep <- test[] # keep unproblematic dataset

S452D <- na.omit(unlist(test\_keep))

n\_S452D <- length(S452D)

# Histogram of each replicate and per pool

hist(S452D, breaks = 200)

par(mfrow = c(3,3))

for (i in 1:dim(test\_keep)[2]) {

hist(test\_keep[,i], main = i, breaks = 80)

}

hist(S452D, main = paste("Pool (n=", n\_S452D, ")", sep = ""), breaks = 80)

# Summary with thresholds

fit.thresholds.pool <- summaryFitThresholds(data = S452D, thresh = -1.67, leg = "Pool S452D")

fit.thresholds.replicate <- sapply(1:dim(test\_keep)[2], function(x) summaryFitThresholds(data = test\_keep[,x],

thresh = -1.67, leg = paste("S452D ", x, sep = "")), simplify = F)

fit.thresholds <- rbind(fit.thresholds.pool, do.call(rbind, fit.thresholds.replicate))

# Summary with mixture model

fit.varequal <- summaryMclust(mdl = "E", data = S452D, G = 2)

mod.varequal <- Mclust(S452D, modelNames = "E", G = 2)

boot.varequal <- MclustBootstrap(mod.varequal, nboot = 999, type = "bs")

summary(boot.varequal, what = "ci")

#plot(mod.varequal, what = "classification")

#plot(mod.varequal, what = "density")

percent\_correct\_classif\_varequal <- 100\*(length(which(S452D<=1.67 & mod.varequal$classification == 1)) +

length(which(S452D>1.67 & mod.varequal$classification == 2)))/length(S452D)

plot(mod.varequal, what = "density")

fit.vardiff <- summaryMclust(mdl = "V", data = S452D, G = 2)

mod.vardiff <- Mclust(S452D, modelNames = "V", G = 2)

boot.vardiff <- MclustBootstrap(mod.vardiff, nboot = 999, type = "bs")

summary(boot.vardiff, what = "ci")

#plot(mod.vardiff, what = "classification")

#plot(mod.vardiff, what = "density")

percent\_correct\_classif\_vardiff <- 100\*(length(which(S452D<=1.67 & mod.vardiff$classification == 1)) +

length(which(S452D>1.67 & mod.vardiff$classification == 2)))/length(S452D)

plot(mod.vardiff, what = "density")

write.table(fit.thresholds, "fit.thresholds\_S452D.xls", col=NA, sep="\t",dec=".")

write.table(fit.vardiff, "fit.vardiff\_S452D.xls", col=NA, sep="\t",dec=".")

write.table(fit.varequal, "fit.varequal\_S452D.xls", col=NA, sep="\t",dec=".")