Machine Learning Analysis of Pulmonary Artery Metafeatures

Roshan Lodha

18 February, 2023

Introduction

```
knitr::opts_chunk$set(warning = FALSE, # turn off warnings
                      message = FALSE,
                      results = 'hide') # hide console output
knitr::opts_chunk$set(fig.width = 10, fig.height = 7) # set figure height and width

pkgs <- c("umap", "readxl", "emulator", "Hmisc", "tidyverse", "glmnet",
    "ggprism", "ggfortify", "RColorBrewer", "plotly", "viridis", "GGally",
    "boot", "table1", "papaja",
    "corrplot")

invisible(lapply(pkgs,
    function(x) suppressMessages(library(x, character.only = TRUE))))

# git token: ghp_pzSZZdurZOrhz6vq78eOKPpcp7t0SM10AHY1
```

Loading Demographic Data

```
ct <- read.csv("./data/vanderbilt/vanderbilt_ct_phenotype_2-14-23.csv",
               header = TRUE,
               row.names = 1)
mri <- read.csv("./data/vanderbilt/vanderbilt_mri_phenotype_2-14-23.csv",
                header = TRUE,
                row.names = 1)

ct
mri
```

```
ct$scan <- "CT"
mri$scan <- "MRI"

ct <- ct %>% dplyr::select(colnames(mri))
dem <- rbind(ct, mri)
dem
```

```
dictionary <- read_csv("./data/vanderbilt/radiogenomics_data_dictionary.csv")
dictionary
```

Table 1

```
dem$gender <-
  factor(dem$gender,
         levels = c(0, 1, 2),
         labels = c("Female",
                    "Male",
                    "Unknown"))
dem$race <-
  factor(dem$race,
         levels = c(0, 1, 2, 3, 4, 5, 6),
         labels = c("American Indian/Alaska Native",
                    "Asian",
                    "Black or African American",
                    "Native Hawaiian or Other Pacific Islander",
                    "White",
                    "Other",
                    "Declined/Prefer not to answer"))

dem$ethnicity <-
  factor(dem$ethnicity,
         levels = c(0, 1, 2),
         labels = c("Hispanic",
                    "Not Hispanic",
                    "Declined/Prefer not to answer"))
```

```
dem$htn <-
  factor(dem$htn,
         levels = c(0, 1),
         labels = c("No Hx of Hypertension",
                    "Hx of Hypertension"))

dem$diabetes <-
  factor(dem$diabetes,
         levels = c(0, 1),
         labels = c("No Hx of Diabetes",
                    "Hx of Diabetes"))

dem$chf <-
  factor(dem$chf,
         levels = c(0, 1),
         labels = c("No Hx of CHF",
                    "Hx of CHF"))

dem$cad <-
  factor(dem$cad,
         levels = c(0, 1),
         labels = c("No Hx of CAD",
```

```
                    "Hx of CAD"))
#dem$pad <-
#  factor(dem$pad,
#         levels = c(0, 1),
#         labels = c("No Hx of PAD",
#                    "Hx of PAD"))

#dem$mi <-
#  factor(dem$mi,
#         levels = c(0, 1),
#         labels = c("No Hx of MI",
#                    "Hx of MI"))

dem$stroke_tia <-
  factor(dem$stroke_tia,
         levels = c(0, 1),
         labels = c("No Hx of Stroke or TIA",
                    "Hx of Stroke or TIA"))

dem$osa <-
  factor(dem$osa,
         levels = c(0, 1),
         labels = c("No Hx of OSA",
                    "Hx of OSA"))
```

```
label(dem$scan) <- "Imaging Modality"
label(dem$gender) <- "Gender"
label(dem$race) <- "Race"
label(dem$ethnicity) <- "Ethnicity"
label(dem$age_ablation) <- "Age at Ablation"
label(dem$pt_height) <- "Height"
label(dem$weight) <- "Weight"
label(dem$htn) <- "History of Hypertension"
label(dem$diabetes) <- "History of Diabetes"
label(dem$chf) <- "History of Congestive Heart Failure"

table1 <- table1(~ gender + race + ethnicity + age_ablation + pt_height + weight + htn + diabetes + chf

table1
```

	CT	MRI	Overall
	(N=725)	(N=618)	(N=1343)
Gender			
Female	262 (36.1%)	221 (35.8%)	483 (36.0%)
Male	463 (63.9%)	397 (64.2%)	860 (64.0%)
Unknown	0 (0%)	0 (0%)	0 (0%)
Race			
American Indian/Alaska Native	1 (0.1%)	1 (0.2%)	2 (0.1%)
Asian	2 (0.3%)	6 (1.0%)	8 (0.6%)
Black or African American	16 (2.2%)	9 (1.5%)	25 (1.9%)
Native Hawaiian or Other Pacific Islander	0 (0%)	1 (0.2%)	1 (0.1%)
White	697 (96.1%)	591 (95.6%)	1288 (95.9%)

	CT	MRI	Overall
Other	0 (0%)	2 (0.3%)	2 (0.1%)
Declined/Prefer not to answer	9 (1.2%)	8 (1.3%)	17 (1.3%)
Ethnicity			
Hispanic	4 (0.6%)	5 (0.8%)	9 (0.7%)
Not Hispanic	709 (97.8%)	605 (97.9%)	1314 (97.8%)
Declined/Prefer not to answer	12 (1.7%)	8 (1.3%)	20 (1.5%)
Age at Ablation			
Mean (SD)	66.3 (10.6)	63.8 (10.9)	65.2 (10.8)
Median [Min, Max]	67.4 [25.7, 87.4]	64.9 [22.9, 87.1]	66.2 [22.9, 87.4]
Height			
Mean (SD)	175 (10.6)	176 (10.5)	176 (10.5)
Median [Min, Max]	178 [147, 206]	177 [145, 201]	177 [145, 206]
Weight			
Mean (SD)	98.1 (24.5)	95.1 (20.6)	96.7 (22.8)
Median [Min, Max]	95.3 [42.4, 199]	93.0 [43.0, 180]	95.0 [42.4, 199]
History of Hypertension			
No Hx of Hypertension	189 (26.1%)	201 (32.5%)	390 (29.0%)
Hx of Hypertension	536 (73.9%)	417 (67.5%)	953 (71.0%)
History of Diabetes			
No Hx of Diabetes	572 (78.9%)	517 (83.7%)	1089 (81.1%)
Hx of Diabetes	153 (21.1%)	101 (16.3%)	254 (18.9%)
History of Congestive Heart Failure			
No Hx of CHF	594 (81.9%)	524 (84.8%)	1118 (83.2%)
Hx of CHF	131 (18.1%)	94 (15.2%)	225 (16.8%)

Metafeatures Analysis

Loading Metafeatures

```
vandy <- read.csv("./data/vanderbilt/primary_vanderbilt_filtered.csv",
                          header = TRUE,
                          row.names = 1)
vandy
```

Dimensionality Analysis

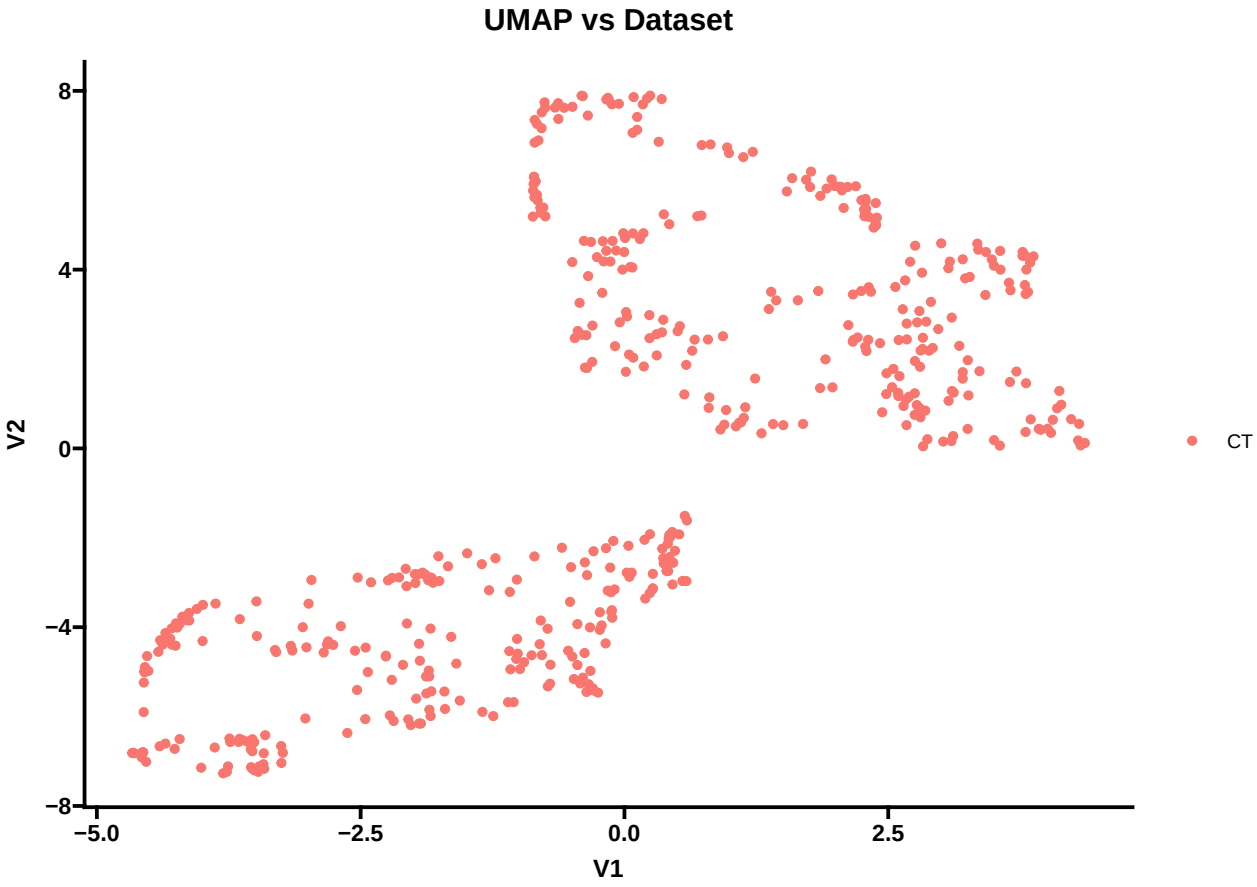
```
vandy <- vandy[, which(apply(vandy, 2, var) != 0)]
vandy <- as.matrix(vandy)
vandy[!is.finite(vandy)] <- NA

vandy <- scale(vandy, center = TRUE, scale = colSums(vandy))
```

UMAP Plot

```
vandy.umap <- umap(na.omit(vandy))
vandy.umap.df <- merge(as.data.frame(dem), as.data.frame(vandy.umap$layout), by = 0)

vandy.umap.plot <- ggplot(data = vandy.umap.df, aes(x = V1, y = V2, color = factor(scan))) +
  geom_point(size = 2) +
  ggtitle("UMAP vs Dataset") +
  theme_prism()
#ggsave("./plots/LA.umap.png", plot = LA.umap.plot, height = 6, width = 6)
vandy.umap.plot
```

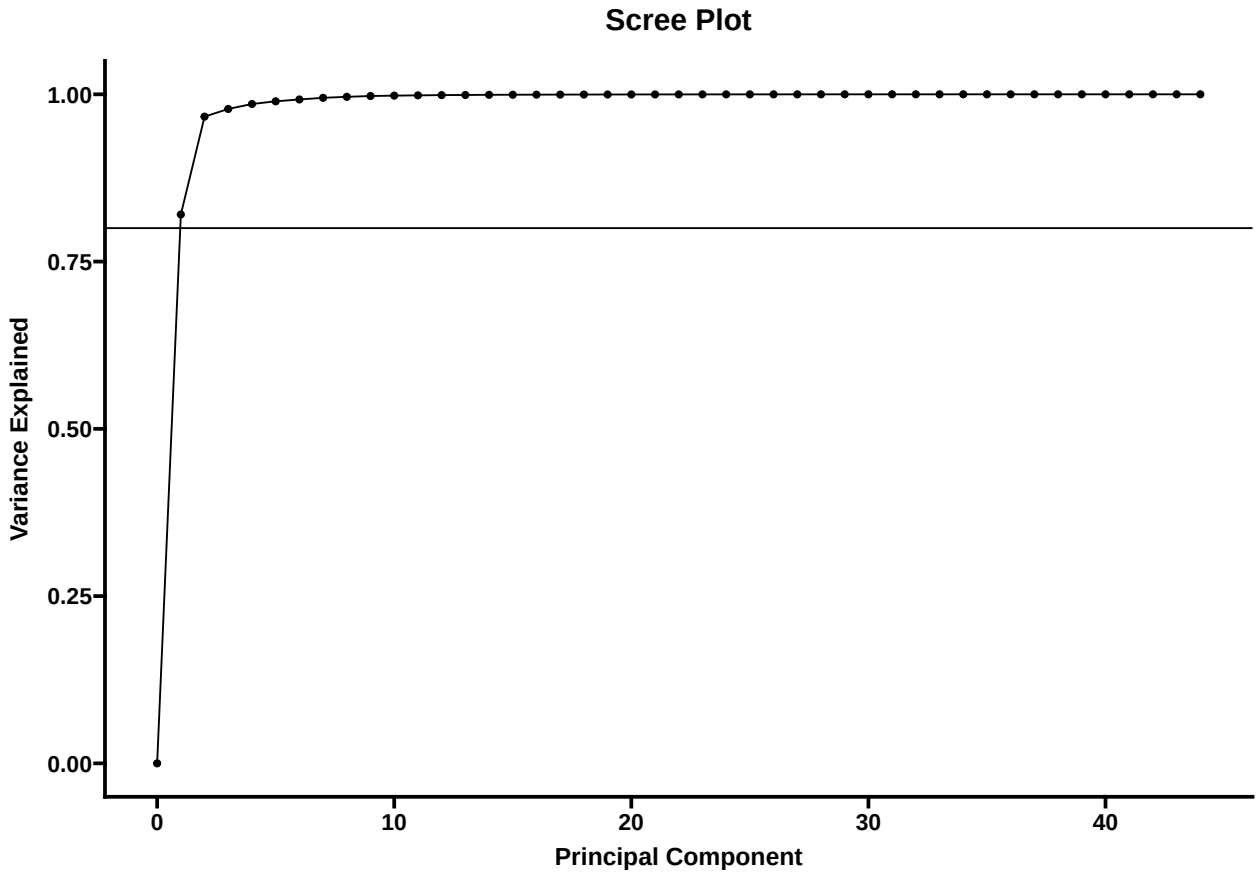


PCA Analysis

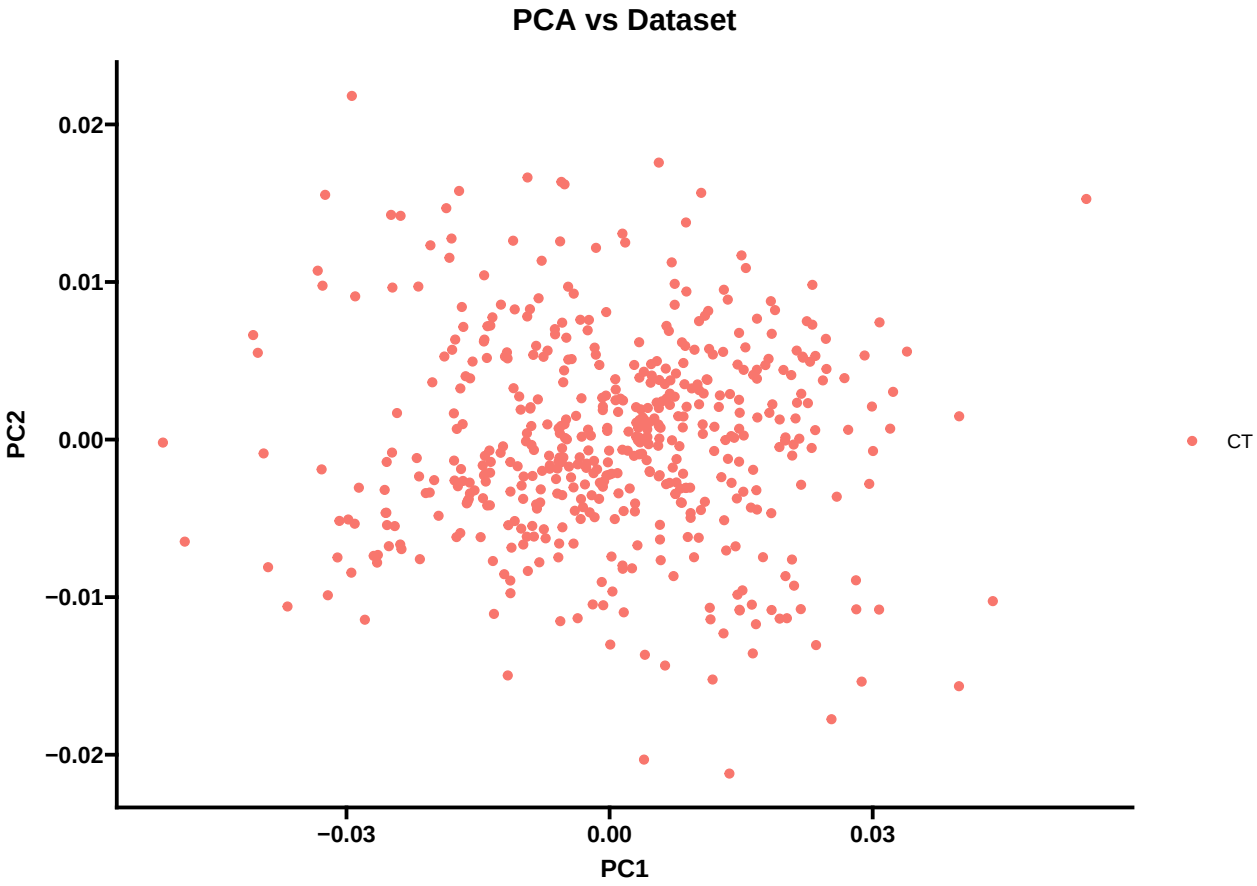
```
vandy.pca <- prcomp(na.omit(vandy))
pca.var.explained <- cumsum(vandy.pca$sdev^2 / sum(vandy.pca$sdev^2))
```

```
scree.plot <- qplot(c(0:length(pca.var.explained)), c(0, pca.var.explained)) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  geom_hline(yintercept = 0.8) +
  ylim(0, 1) +
```

```
  theme_prism()
#ggsave("./plots/LA.scree.png", plot = scree.plot, height = 6, width = 6) #12
scree.plot
```

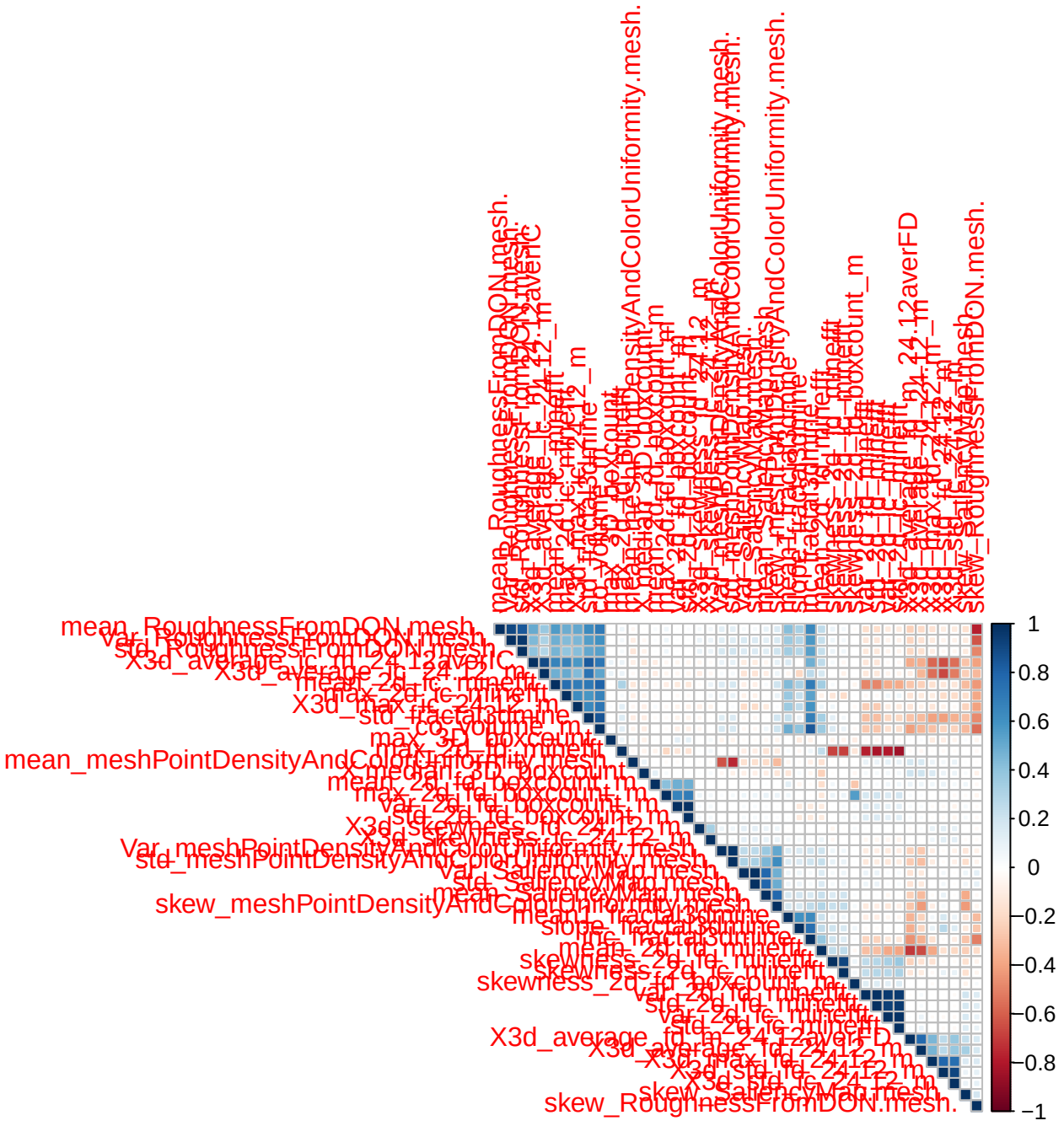



```
vandy.pca.df <- merge(as.data.frame(dem), as.data.frame(vandy.pca$x[,1:44]), by = 0)
vandy.pca.plot <- ggplot(data = vandy.pca.df, aes(x = PC1, y = PC2, color = factor(scan))) +
  geom_point(size = 2) +
  ggtitle("PCA vs Dataset") +
  theme_prism()
#ggsave("./plots/LA.pca.png", plot = LA.pca.plot, height = 6, width = 6)
vandy.pca.plot
```



Correlation Between Metafeatures

```
vandy.corr <- cor(vandy)
#corrplot.mixed(vandy.corr, lower = "number", upper = "circle", order = "hclust", type = "full")
corrplot(vandy.corr, method = "square", order = "hclust", type = "upper")
```


mean_RoughnessFromDON.mesh.
Var_RoughnessFromDON.mesh.
std_RoughnessFromDON.mesh.
X3d_average_fc_fd_24_12_m
mean_meshPointDensityAndColorUniformity.mesh.
Var_meshPointDensityAndColorUniformity.mesh.
std_meshPointDensityAndColorUniformity.mesh.
skew_meshPointDensityAndColorUniformity.mesh.
skew_RoughnessFromDON.mesh.
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1