Results & Discussion/Conclusion

Kasper Notebomer

10/2/2021

Short introduction

This report is to describe the results of the exploratory data analysis and data cleaning done on a dataset of membrane composition and characteristics. The data was provided by Tsjerk Wassenaar of the RuG. The data set contains data on membrane composition and characteristics. The data set is not publicly available. The data set also does not have a publication linked to it.

Seeing as there is no paper linked to the data set I can only guess using the information that I did get that the data was gathered by making a membrane using specific variables and the measuring the resulting membranes to get variables like the thickness and compressibility. The data set contains 14 variables and 2843 different measurement. The goal is to use measurement to predict the composition of the membrane. Typically you would use independent variables to predict one dependent variable. In this data set this doesn't seem to be the case, seeing as the variables that are given as parameters are seen as class variables, this means that these would be considered the labels. The parameters are: Temperature, Sterol type, Sterol concentration, Other (phospho)lipids in membrane, Aliphatic tails, Saturation index, Phosphatidyl choline concentration and Ethanol concentration. So the biggest question that the EDA should answer is which of these variables is most interesting to use as the label, or could we even predict multiple of them using the given variables.

Results

The results are divided into two different subsection: data exploration and data cleaning. This is because both of these sections have different goals and are both major subject of discussion. The code for both the data exploration steps and the data cleaning can be found in appendix A. All of the abbreviation are explained in the codebook, shown in appendix B.

Data exploration

The biggest goal of the exploration phase was to give insight into the data, like relations between variables. Using this insight, the research question could be tweaked to have a higher likelihood of being achievable. Another question the exploration phase was looking to answer was whether or not the chosen dataset was suitable for machine learning.

Table 1: Five number summary of the original dataset

	temperature	sterol.conc	satur.index	PC.conc	ethanol.conc	APL	thickness	bending	tilt	zorder	compress
Minimum	298.0	0.00	0.0000	0.00	0	0.4514	3.003	0.2025	6.774	0.02736	1.286
Q1	298.0	10.00	0.0000	25.00	5	0.6350	3.627	9.8579	14.490	0.18604	23.097
Median	298.0	20.00	0.5000	50.00	15	0.6885	3.812	12.1000	18.157	0.25860	31.346
Mean	305.2	16.55	0.6208	50.02	15	0.6807	3.934	22.0256	35.732	0.33269	48.714
Q3	298.0	30.00	1.0000	75.00	25	0.7394	4.238	25.5712	35.933	0.45559	43.348
Maximum	328.0	30.00	2.0000	100.00	30	0.8924	5.101	125.3820	227.813	0.93013	538.446
Number of NA's	0	0	0	0	0	104	103	104	104	153	103

Figure 1:

Density plots of multiple variables coloured on membrane components

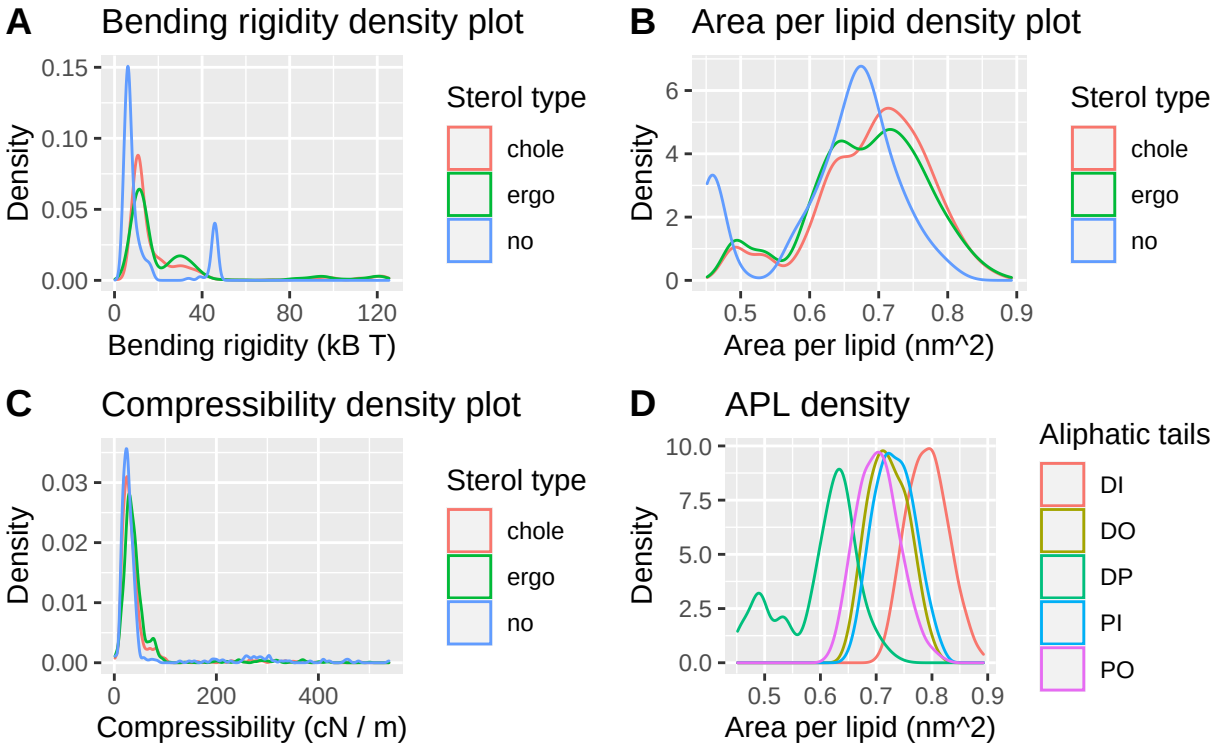


Table 1 shows the five number summary of all of the numerical columns that are present in the dataset. Inspecting this table closely, a deviation in the mean and the median of some of the columns can be ob-

served. This is an indication that our data might be skewed. This is an important observation, seeing as some algorithms expect the data to have a Gaussian(normal) distribution. Any type of normalization or transformation was dealt with in the data cleaning. The NA's that are present were also be removed in the data cleaning.

The results of plotting some variables and coloring them based on certain class variables can be seen in figure 1. These density plots are made to look whether or not it might by possible to distinguish between class variables based on one specific variable.

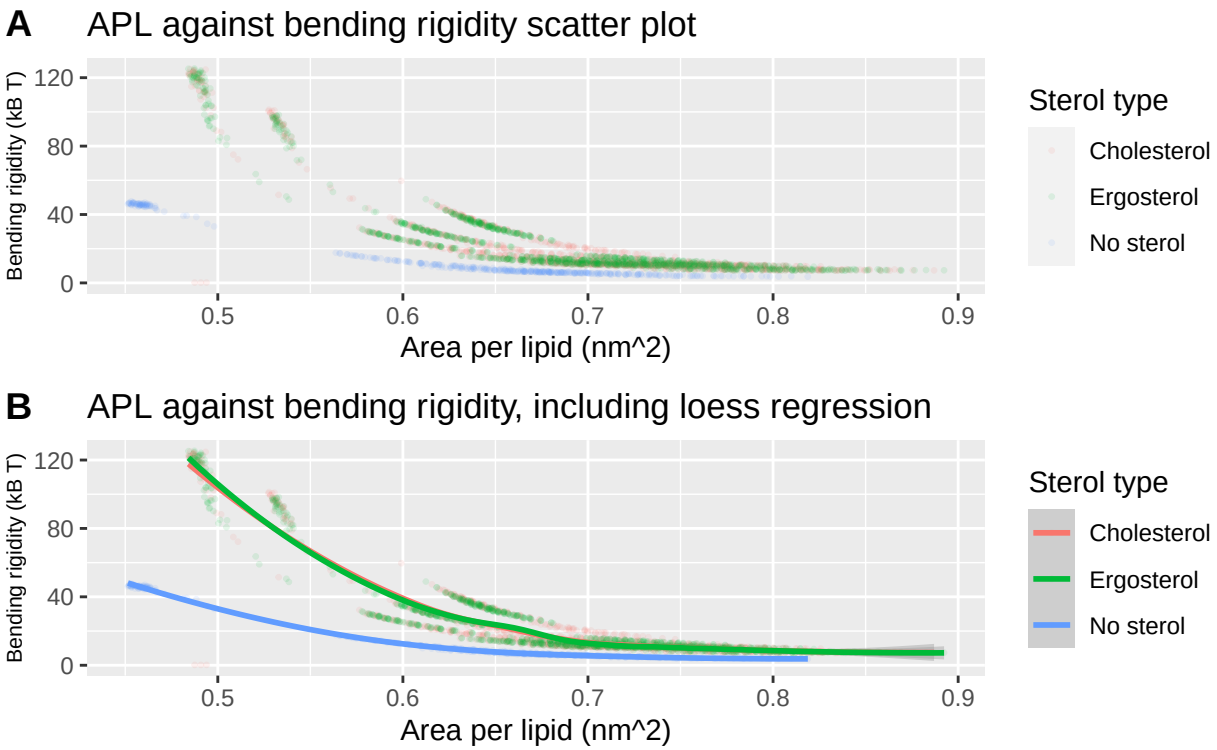
Plot A in figure 1 show that there is basically no distinction between cholesterol and ergosterol based on the bending rigidity. However, distinction between no sterol or a sterol does seem to be possible based on the bending rigidity seeing as the peaks of there classes only overlaps a small bit. There does seem to be an odd peak in the no sterol class at around 50 kB/t bending rigidity.

In plot B it looks like there is absolutely no way of distinguishing different classes based on the area per lipid, seeing as the peaks are basically in the same place. But just to be certain we'll still plot in against some other variables to make sure, seeing as it might be a very useful variable when paired with something like bending rigidity. The results depicted in plot C also don't look very promising seeing as every peak is around the same place again. Plot D on the other hand seems a lot more interesting seeing as all of the peaks are in slightly different locations, the DO, PO and the PI seem to overlap a lot but the DP and DI tails seem to have little overlap with the rest. It seems like APL could possibly be used as a variable to distinguish between aliphatic tails when paired with another variable.

When looking at all of the plots it looks like most of the data is skewed in one way or another, confirming the numbers given by the five number summary.

Figure 2:

Relation between APL and bending rigidity of three different sterol types, with and without loess regression line



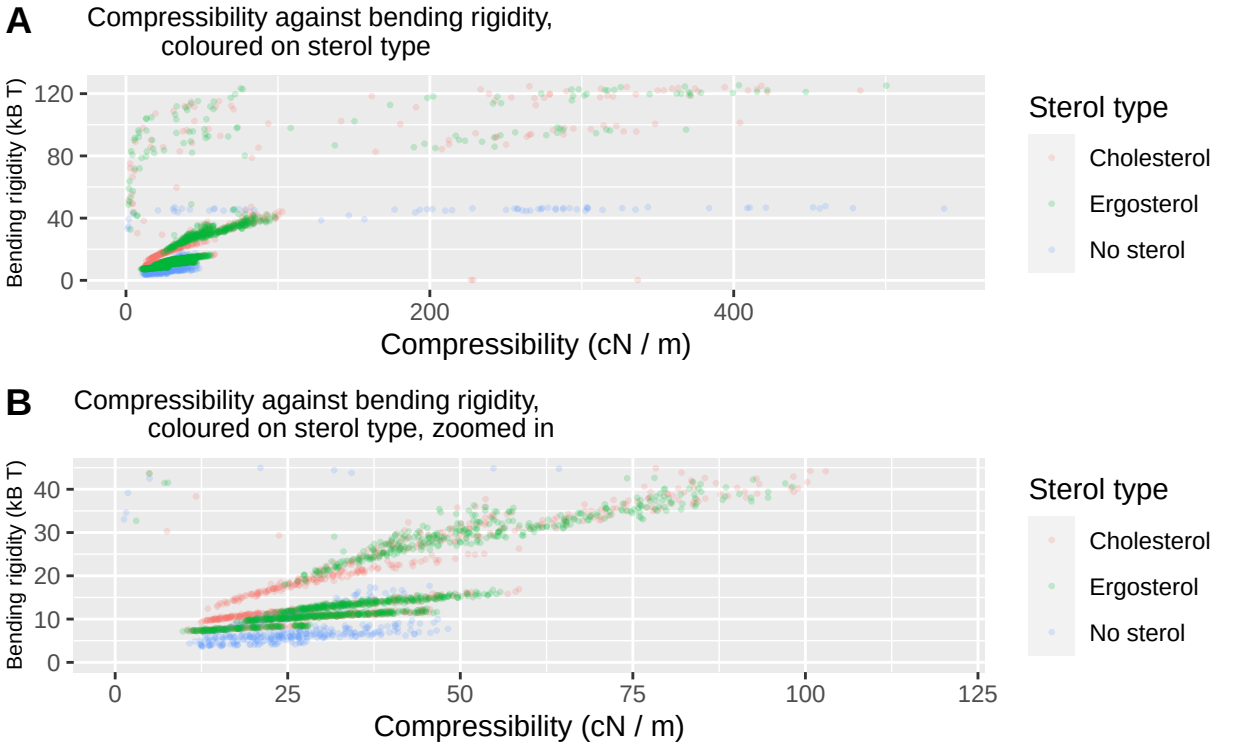
An interesting correlation that was found in the data exploration is shown in figure 2. It shows that there is an inverse logarithmic correlation between the area per lipid of a membrane and the bending rigidity. An even more interesting observation is the fact that, when colored on sterol type, it shows that there is a big

difference between the no sterol group and the other two groups. This means that this ratio could be very important in the prediction of sterol type based on the membrane characteristics. It does however seem like these variables aren't useful when it comes to discerning between ergosterol and cholesterol.

The results of plotting the compressibility of the membrane against the bending rigidity can be found in figure 3. In plot A of figure 3 we can't really make out much of a pattern in the dense clout to the left, but when looking at the rest of the data points there does seem to be some grouping based on whether or not no sterol or a sterol is present. When zooming in in plot B we can now see the same kind of grouping happen again based on either a sterol present or no sterol present. Something else that figure 3 shows it that a log transformation on the compressibility variable might be in order seeing as there are a lot of point in the magnitude of tens but also a lot in the hundreds.

Figure 3:

Relation between compressibility and bending rigidity of three different sterol types, full and zoomed

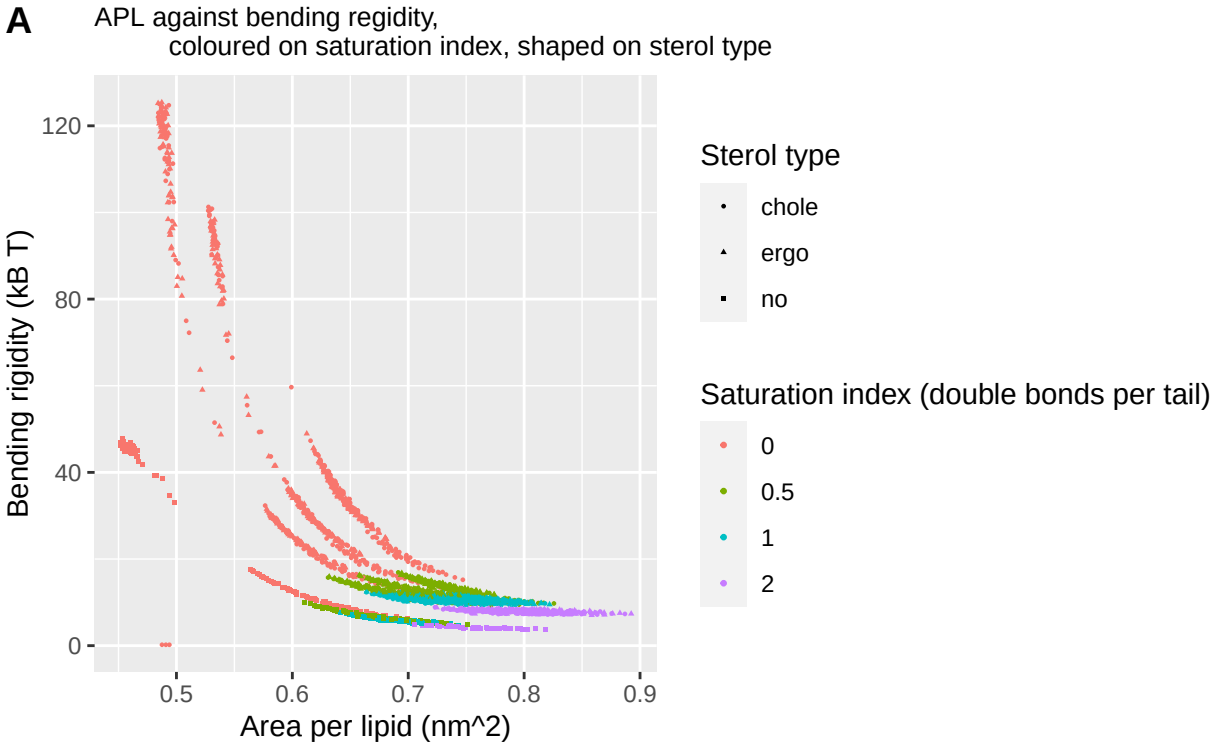


In figure 4 the area per lipid is again plotted against the bending rigidity, only this time, the plot is colored based on the saturation index.

The results shown in figure 4 show an interesting pattern. It seems that a specific saturation index is only found in membranes within a specific range of area per lipid and bending rigidity.

Figure 3:

Relation between APL and bending rigidity of three different sterol types with different saturation levels

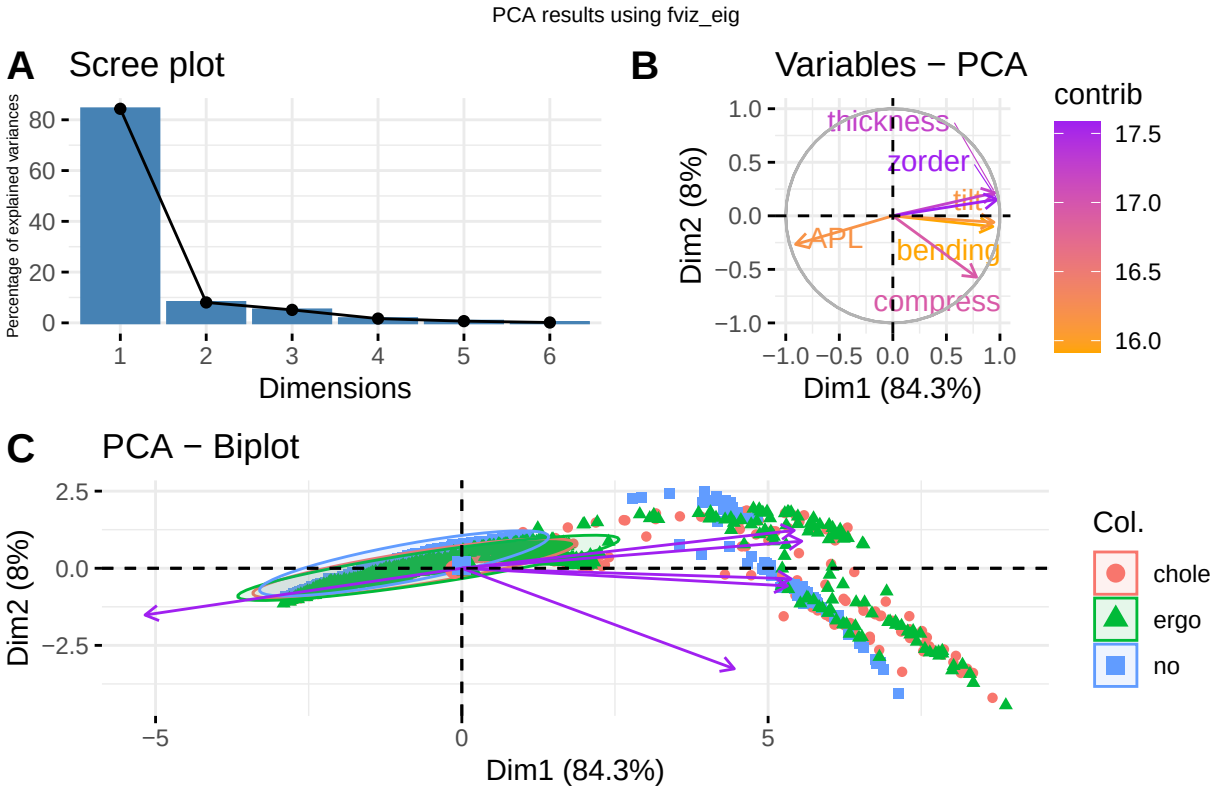


To look for correlation between the variables and clustering a PCA plot was made. This plot is shown in figure 4.

Plot A in figure 4 shows something that is quite strange. It states that over 80% of the variance can be explained by the first principle component, this is quite odd. Plot B shows that some of the variables seem to be highly correlated like z order and thickness just like tilt angle and bending rigidity seeing as their arrows are pointing in the same direction. Area per lipid(APL) on the other hand seems to be negatively correlated to z order and thickness. When looking at plot C there doesn't seem to be any obvious form of clustering, if any, for that matter.

All of the results from the data exploration led to the following research question: Is it possible to use machine learning to reliably predicts the sterol type in a membrane with a higher than 80% accuracy, given certain measurements like the bending rigidity, the are per lipid and the compressibility?

Figure 4:



Data cleaning

To make the dataset usable for machine learning it had to be cleaned up using various of methods. The difference between the cleaned dataset and the original dataset can be seen when comparing table 2 and table 3. Unused columns like temperature and ethanol concentration where removed. All of the data in the columns was scaled using min max normalization. This is visible in table 3, seeing as the values now range from zero to one. NA's where also omitted. It was decided not to remove any variables based on a high degree of correlation. Table 3 doesn't have all of the variables that table 2 has. This is because the variables that where removed would be seen as labels from a machine learning point of view. Seeing as this research is interested in only one of the labels, sterol type, the other where removed.

Table 2: Top of original data

temperature	sterol.type	sterol.conc	other.phosph	tails	satur.index	PC.conc	ethanol.conc	APL	thickness	bending	tilt	zorder	compress
298	chole	20	PE	PI	1	33	20	0.736911	3.628973	10.08020	14.7578	0.175499	24.18218
298	chole	20	PE	PI	1	50	20	0.746226	3.596991	10.14030	14.3075	0.172084	24.32495
298	chole	20	PE	PI	1	25	20	0.731945	3.652637	10.25470	14.9777	0.178494	23.76607
298	chole	20	PE	PI	1	100	20	0.769630	3.496052	10.02320	13.2029	0.163349	22.96077
298	chole	20	PE	PI	1	0	20	0.719509	3.710018	10.46540	15.7530	0.182191	23.10503
298	chole	20	PE	PI	1	75	20	0.760091	3.538908	9.86163	13.7148	0.166807	23.86050

In table 4, the five number summary of the cleaned dataset can be found. It shows that all of the data is now on a common scale of zero to one, and that all of the NA's from the original dataset where omitted. This can be deduced from the fact that there is no line present in the table with the number of NA's, seeing as when there are no NA's present, this line won't be returned.

Table 3: Top of cleaned data

APL	thickness	bending	tilt	zorder	compress	sterol.type
0.6474103	0.2983812	0.0789085	0.0361181	0.1640910	0.0426240	chole
0.6685354	0.2831408	0.0793886	0.0340809	0.1603082	0.0428898	chole
0.6361482	0.3096582	0.0803025	0.0371130	0.1674086	0.0418494	chole
0.7216122	0.2350388	0.0784531	0.0290836	0.1506323	0.0403502	chole
0.6079452	0.3370028	0.0819857	0.0406205	0.1715037	0.0406188	chole
0.6999791	0.2554618	0.0771624	0.0313995	0.1544628	0.0420252	chole

Table 4: Five number summary of the cleaned dataset

	APL	thickness	bending	tilt	zorder
Minimum	0.0000	0.0000	0.00000	0.00000	0.0000
Q1	0.4163	0.2954	0.07817	0.03477	0.1758
Median	0.5412	0.3865	0.09643	0.05161	0.2561
Mean	0.5211	0.4453	0.17643	0.13251	0.3382
Q3	0.6560	0.5915	0.20684	0.13454	0.4743
Maximum	1.0000	1.0000	1.00000	1.00000	1.0000

Discussion and Conclusion

Conclusion

To conclude, based on the finding of the data exploration it is safe to assume that the sterol type of a given membrane can be predicted by specific measurements of the membrane. This can be said because a multitude of plots states that there is a difference between ratios of variables, that is caused by a specific sterol in the membrane. An example of this is the ratio of area per lipid to bending rigidity, which shows a clear difference between no sterol and a sterol present in the membrane. It is also safe to assume that the dataset is suitable for machine learning experiments after some small adjustments and modifications.

Discussion

Exploration

One of the most important things to note is that there was probably an subconscious bias in the exploration of the data. This is because most of the steps where taken while there was already a research question. Because of this the exploration was probablymore focused on refining the research question then on looking for unexpected patterns.

The PCA plot was made before any of the data cleaning step. This means that the data wasn't transformed or normalized. The function used to plot the PCA plot did have the option to scale the data enabled, which should have taken care of any scaling issues. Except that this function might make use of z score scaling, in which case the results might not be totally accurate seeing as there is a slight skew in some of our variables. Z score scaling assumes that our data has a normal distribution, which is why it might have caused problems. Even after taking this into account and trying to reduce the skew by using log and square transformations, we were still left with a slight skew. Even reducing the skew, the scree plot still showed around the same amount of variance explained by each dimension and PCA plot still showed around the same amount of correlation. The main difference being that the plot is now flipped upside down.

An important thing to address is the fact that outliers in the dataset aren't removed. This is for the sole reason that there is no justification to do so, there also wasn't much information given on how the data was gathered. Seeing as the data was most likely gathered from a controlled experiment, it should be

assumed that these data points aren't incorrect inputs or caused by environmental errors, but actual correct measurement. This means that these outliers are actual possibilities that could be found in nature, which is why they shouldn't be removed. Some of the outliers could also be a result of a bias in the making of the samples. There were also no obvious erroneous instances that needed to be removed.

Data cleaning

The only data cleaning that was done was the removal of columns and min max normalization. The reason behind the removal of these columns was that these columns where technically other labels that could have been used in machine learning. Some of them could have been left in seeing as they could be interesting in future research, but this would just clutter up the dataset. Even though some of the variables seem to be highly correlated, it was decided not to remove some of them. This is mostly because they could still hold valuable information seeing as they don't have a correlation of one. Another reasons is because of the possible problems that where raised about the PCA plot. These variables could still be removed in the future to check if this further improves the machine learning algorithm. The biggest reason why they weren't removed was because the PCA plot is based on the assumption that the correlation between variables is linear. In some comparisons variables will indeed have a strong correlation, but this might change when another dimension like the sterol type is added. An example of this can be found in the results chapter in figure 2. When trying to fit a linear regression line through all of the data point, you would probably get a pretty strong correlation. But once colored on sterol type, it becomes very clear that the correlation isn't linear but inverse logarithmic.

In the cleaning of the data, min max normalization was used. This leads to a suppression of the effect of outliers. To combat this, z score normalization could have been used. The only problem with this is that z score normalization is based on a normal distribution, while all of the data in the dataset is heavily skewed. Something to combat this would be using transformation. It was however decided not to make use of this, seeing as this lessens the skew in most data, but not by a big margin. The choice not to transform the data could lead to problems in the future if we where to use an algorithm that assumes the data has a Gaussian distribution. However, as previously stated, even transforming the data didn't reduce skewness by a large margin.

Future research

The research done can be used to show more insight into sterols effect on characteristics of a membrane. It also gives insight into the dataset, and shows how it can be transformed for use in machine learning. Future research done should focus on the actual usage of machine learning algorithms on the dataset.

Apendix

A, The code

```
# Libraries
library("ggplot2")
library("kableExtra")
library("factoextra")
library("cowplot")
library("gridExtra")
library("dplyr")
library("scales")
library(ggpubr)
library(ggfortify)
library(tidyr)
library(purrr)
library(ggcorrplot)
library(reshape2)

# Read the data and transform it to a tibble
data <- read.csv("data/dataFrame_all_sims.csv",
                 sep = ",", header=TRUE,
                 stringsAsFactors=FALSE)
data <- as_tibble(data)

# Remove NA's
nona_data <- na.omit(data)
# Scaling
scale_min_max <- function(x) (x - min(x)) / (max(x) - min(x))
nona_copy <- nona_data[c("APL", "thickness", "bending",
                                              "tilt", "zorder","compress")]

# Min max normalized data
scaled_data <- as.data.frame(lapply(nona_copy, FUN=scale_min_max))
scaled_data$sterol.type <- nona_data$sterol.type

# Dataframe used for PCA plot
correlation <- nona_data[,9:14]

#Read the codebook
codebook <- read.csv("data/codebook.csv", sep = ",", header=TRUE)

# Get the description by the abbreviation from the codebook
get_des_by_ab <- function(df, abbreviation){
  description <- df[df$Abbreviation == abbreviation,]$Desciption
  return(description)
}

# Get the five number summary of the selected columns
sum <- summary(data[c("temperature","sterol.conc","satur.index",
                      "PC.conc","ethanol.conc","APL", "thickness", "bending",
                      "tilt", "zorder","compress")])
```

```r
sum <- sub(".*:", "", sum)
sum[is.na(sum)] <- 0
rownames(sum) <- c("Minimum", "Q1", "Median", "Mean", "Q3", "Maximum", "Number of NA's")

# Print five number summary using kable
kable(sum, caption="Five number summary of the original dataset") %>%
  kable_styling(latex_options = c("scale_down", "hold_position"))

# Create density plots
p <- ggplot(nona_data, aes(x=bending, colour = factor(sterol.type))) +
  geom_density() +
  xlab(get_des_by_ab(codebook, "bending")) +
  ylab("Density") +
  ggtitle("Bending rigidity density plot") +
  labs(colour = get_des_by_ab(codebook, "sterol.type"))

p2 <- ggplot(nona_data, aes(x=APL, colour = factor(sterol.type))) +
  geom_density() +
  xlab(get_des_by_ab(codebook, "APL")) +
  ylab("Density") +
  ggtitle("Area per lipid density plot") +
  labs(colour = get_des_by_ab(codebook, "sterol.type"))

p3 <- ggplot(nona_data, aes(x=compress, colour = factor(sterol.type))) +
  geom_density() +
  xlab(get_des_by_ab(codebook, "compress")) +
  ylab("Density") +
  ggtitle("Compressibility density plot") +
  labs(colour = get_des_by_ab(codebook, "sterol.type"))

p4 <- ggplot(nona_data, aes(x=APL, colour = factor(tails))) +
  geom_density() +
  xlab(get_des_by_ab(codebook, "APL")) +
  ylab("Density") +
  ggtitle("APL density") +
  labs(colour = get_des_by_ab(codebook, "tails"))
# Plot density plots
plots <- ggarrange(p, p2, p3, p4,
                   labels = c("A", "B", "C", "D"),
                   ncol = 2, nrow = 2)
title <- expression(atop(bold("Figure 1:"),
                         scriptstyle(paste("Density plots of multiple",
                                           " variables coloured on membrane components"))))
annotate_figure(plots,
                top=text_grob(title))

#Plot APL against bending rigidity
chart <- ggplot(nona_data, aes(APL, bending, colour = factor(sterol.type) )) +
  geom_point(alpha=1/10, size=0.5) +
  xlab(get_des_by_ab(codebook, "APL")) +
  ylab(get_des_by_ab(codebook, "bending")) +
  theme(axis.title.y = element_text(size=8)) +
  ggtitle("APL against bending rigidity scatter plot") +
```

```
  scale_color_manual(name=get_des_by_ab(codebook, "sterol.type"),
                     labels = c("Cholesterol",
                                "Ergosterol",
                                "No sterol"),
                     values = hue_pal()(3))

plots <- ggarrange(chart,
                   chart +
                     geom_smooth(method = "loess") + ggtitle("APL against bending rigidity, including lo
                   labels = c("A", "B"),
                   ncol = 1, nrow = 2)

title <- expression(atop(bold("Figure 2:"),
                         scriptstyle(paste("Relation between APL and bending ",
                                           "rigidity of three different sterol ",
                                           "types, with and without loess regression line"))))
# Give the plot a number
annotate_figure(plots,
                top=text_grob(title))

# Create plot
chart <- ggplot(nona_data, aes(compress, bending, colour = factor(sterol.type) )) +
  geom_point(alpha=2/10, size=0.5) +
  xlab(get_des_by_ab(codebook, "compress")) +
  ylab(get_des_by_ab(codebook, "bending")) +
  theme(axis.title.y = element_text(size=8)) +
  ggtitle("Compressibility against bending rigidity,
          coloured on sterol type") +
  scale_color_manual(name=get_des_by_ab(codebook, "sterol.type"),
                     labels = c("Cholesterol",
                                "Ergosterol",
                                "No sterol"),
                     values = hue_pal()(3)) +
  theme(plot.title = element_text(size=10))

chart2 <- ggplot(nona_data, aes(compress, bending, colour = factor(sterol.type) )) +
  geom_point(alpha=2/10, size=0.5) +
  xlab(get_des_by_ab(codebook, "compress")) +
  ylab(get_des_by_ab(codebook, "bending")) +
  theme(axis.title.y = element_text(size=8)) +
  ggtitle("Compressibility against bending rigidity,
          coloured on sterol type, zoomed in") +
  scale_color_manual(name=get_des_by_ab(codebook, "sterol.type"),
                     labels = c("Cholesterol",
                                "Ergosterol",
                                "No sterol"),
                     values = hue_pal()(3)) +
  scale_x_continuous(limits=c(0,120)) +
  scale_y_continuous(limits=c(0, 45))  +
  theme(plot.title = element_text(size=10))

# Arrange plots
plots <- ggarrange(chart,
```

```
                  chart2,
                  labels = c("A", "B"),
                  ncol = 1, nrow = 2)

title <- expression(atop(bold("Figure 3:"),
                         scriptstyle(paste("Relation between compressibility and bending ",
                                           "rigidity of three different sterol ",
                                           "types, full and zoomed"))))
annotate_figure(plots,
                top=text_grob(title))

# Create plots
chart2 <- ggplot(nona_data, aes(APL, bending, colour = factor(satur.index), shape = factor(sterol.type)
  geom_jitter(alpha=1, size=0.5) +
  xlab(get_des_by_ab(codebook, "APL")) +
  ylab(get_des_by_ab(codebook, "bending"))  +
  ggtitle("APL against bending regidity,
          coloured on saturation index, shaped on sterol type") +
  theme(plot.title = element_text(size=10))
chart <- chart2 + labs(colour = get_des_by_ab(codebook, "satur.index"), shape = get_des_by_ab(codebook,

# Arrange plots
plots <- ggarrange(chart,
                  labels = c("A"),
                  ncol = 1)

title <- expression(atop(bold("Figure 3:"),
                         scriptstyle(paste("Relation between APL and bending ",
                                           "rigidity of three different sterol ",
                                           "types with different saturation levels"))))
annotate_figure(plots,
                top=text_grob(title))

# Create PCA plot
res.pca <- prcomp(correlation, scale = TRUE)
plot1 <- fviz_eig(res.pca) +
  theme(axis.title.y = element_text(size=6))
plot2 <- fviz_pca_var(res.pca,
                      col.var = "contrib", # Colour by contributions to the PC
                      gradient.cols = c("orange", "purple"),
                      repel = TRUE     # Avoid text overlapping
                      )
plot3 <- fviz_pca_biplot(res.pca, repel = FALSE,
                         col.var = "purple", # Variables colour
                         col.ind = nona_data$sterol.type,  # Individuals colour
                         label = FALSE,
                         addEllipses =  TRUE,
                         ellipse.type = "t"
                         )
# Arrange plots
plot <- arrangeGrob(plot1,  plot2,
                    plot3,
                    ncol = 2, nrow = 2,
```

```
                    layout_matrix = rbind(c(1,2), c(3,3)))

plot <- as_ggplot(plot) +
  draw_plot_label(label = c("A", "B", "C"), size = 15,
                  x = c(0, 0.5, 0), y = c(1, 1, 0.5))
title <- expression(atop(bold("Figure 4:"),
                         scriptstyle("PCA results using fviz_eig")))
annotate_figure(plot,
                top=text_grob(title))

# Get the five number summary of the selected columns
sum <- summary(scaled_data[1:5])

sum <- sub(".*:", "", sum)
sum[is.na(sum)] <- 0
rownames(sum) <- c("Minimum", "Q1", "Median", "Mean", "Q3", "Maximum")

# Print five number summary using kable
kable(sum, caption="Five number summary of the cleaned dataset") %>%
  kable_styling(latex_options = c("hold_position"), full_width = TRUE)
```

B, The codebook

```
codebook <- read.csv("data/codebook.csv", sep = ",", header=TRUE)
# Turn dataframe into a tibble and print
codebook <- as_tibble(codebook)
kable(codebook, caption = "Codebook") %>%
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

Table 5: Codebook

Abbreviation	Type	Class	Desciption	Units
temperature	integer	integer	Temperature (Kelvin)	Kelvin
sterol.type	character	character	Sterol type	
sterol.conc	integer	integer	Sterol concentration (%)	%
other.phosph	character	character	Other (phospho)lipids in membrane (headgroup)	
tails	character	character	Aliphatic tails	
satur.index	double	numeric	Saturation index (double bonds per tail)	
PC.conc	integer	integer	Phosphatidyl choline concentration (% of non-sterol lipids)	
ethanol.conc	integer	integer	Ethanol concentration (% of solvent)	
APL	double	numeric	Area per lipid (nm^2)	nm^2
thickness	double	numeric	Thickness (nm)	nm
bending	double	numeric	Bending rigidity (kB T)	kB T
tilt	double	numeric	Tilt angle (degrees)	degrees
zorder	double	numeric	Z-order	
compress	double	numeric	Compressibility (cN / m)	cN / m