Antonella D'Amico


# Breast cancer analysis

Academic year 2020/2021

# Contents

**1 Dataset description** **2**
1.1 Variables description . . . . . 2

**2 Methodology** **4**
2.1 Mixed Interaction Model . . . . . 4
2.2 Model Selection . . . . . 5

**3 Anlaysis** **6**

**4 Conclusions** **9**

**5 Appendix: R Code** **10**

**Abstract.** The aim of this project is to analyze the dependencies among a set of variables, which can be gathered in routine blood analysis, through the use of Mixed Interaction models. The dataset contains records related to women patients located in Coimbra (Portugal) who have been diagnosed or not breast cancer. The result of the analysis is consistent with those found in the literature: BMI, Age, Glucose and Resistin are good candidates to predict breast cancer.

# 1 Dataset description

The dataset analyzed is **Breast Cancer Coimbra Data Set** stored in UCI Machine Learning repository [1]. The dataset contains records of routine blood tests of 116 patients, living in Coimbra (Portugal). These analysis were used to check whether patients were suffering from breast cancer or not.
The dataset contains ten continuous variables and one categorical.

| Name | Type | Domain |
|---|---|---|
| Classification | discrete | {0,1} |
| Age | continuous | {24, 89} |
| BMI | continuous | {18, ..., 39} |
| Glucose | continuous | {60, ..., 201} |
| Resistin | continuous | {3, ..., 83} |
| HOMA | continuous | {0, ..., 26} |
| Insulin | continuous | {2, ..., 59} |
| Leptin | continuous | {4, ..., 91} |
| Adiponectin | continuous | {1, ..., 39} |
| MCP.1 | continuous | {45, ..., 1700} |

## 1.1 Variables description

- The variable *Classification* describes whether a patient has or not breast cancer. It contains two levels:

  1. Healthy
  2. Patients

  The dataset contains 64 records associated to individual with breast cancer and 52 records for healthy individual.

- The variable *BMI* stands for Body Mass Index. According to some studies, obesity is associated with increased cancer risk.

- *Glucose* is the level of glucose resulted in the routine blood analysis. It is considered as a simple sugar that is transported from the blood to the cells and is able to provide energy to them. A normal blood glucose level for adults, without diabetes, two hours after eating is 90 to 110 mg/dL.

- *Resistin* is an important hormone protein that is mainly produced by adipose tissue, and has been found to be actively involved in the regulation of inflammation and insulin resistance.
- *HOMA*, Homeostatic Model Assessment is a method used to quantify insulin resistance, a pathological condition in which cells fail to respond normally to the hormone insulin, and beta-cell function.
- *Insulin* is a hormone made in the pancreas. It allows body to use glucose for energy. Insulin can also help to balance blood glucose levels. It is highly correlated to diabetes since it occurs when the body does not use insulin properly or does not make enough of it.
- *Leptin* is one of the hormones directly connected to body fat and obesity. Because it comes from fat cells, leptin amounts are directly connected to an individual's amount of body fat; it increases or decreases, according to body weight.
- *Adiponectin* is a protein hormone, which is involved in regulating glucose levels.

After preprocessing phase only 5 variables were realised and considered for the analysis: *Classification, Age, BMI, Glucose and Resistin.*

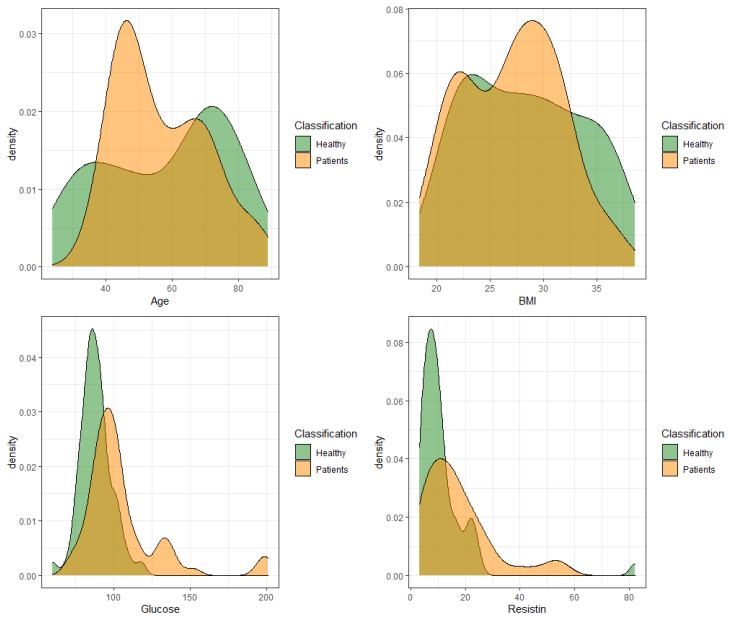



Figure 1: Conditional Probability density functions

# 2 Methodology

## 2.1 Mixed Interaction Model

Since the dataset contains both discrete and continuos variables, this analysis was performed by using the **Mixed Interaction Model**. These models were developed to describe associations between variables that can be discrete or continuous. Discrete variables are described through Log-linear models, while continuous variables are describer by using Gaussian graphical models.
Graphical models are advanced tools used to described the relationship among a set of random variables through a probability distribution. They are used to compactly represent complex, real-world phenomena.
A graph is an ordered pair $\mathcal{G} = \{\boldsymbol{V},\boldsymbol{E}\}$ where $\mathbf{V}$ is the set of vertices and $\mathbf{E}$ is the set of edges. Graph can be divided into *Undirected Graphs* where vertices are connected to each others by segments, and *Directed Acyclic Graphs* where edges can be only arrows.
As said before, Graphical models are powerful tools which can be used with both continuous or discrete variables. In these two situations we need to face different domains and due to this it is necessary to use different parametrization strategies. In the discrete case the distribution of an undirected graph is factorized according to the *Gibbs Representation* as following:

$$P(X) = \frac{1}{Z}\prod_{i=1}^{m} g_i(\mathbf{X}_{C_i})$$

where $g_i$ are distinct functions and $C_i$ are the cliques of the graph.
In the continuous case, the set of variables follows the multinomial Gaussian distribution, $Y \sim N_{\mathrm{d}}(\mu,\Sigma)$, where $\mu$ is the vector of means and $\Sigma$ is the covariance matrix. The inverse of the covariance matrix, $\mathrm{K} = \Sigma^{-1}$, is the concentration matrix. The latter can be useful in understanding the conditional independence between two variables. This is true if the entry in correspondence of two independent variables is null.
In general, given a vector of random variables which follows the multinomial Gaussian distribution, $\boldsymbol{Y} = \{\mathrm{Y}_1,...,\mathrm{Y}_{\mathrm{d}}\} \sim \mathrm{N}_{\mathrm{d}}(\mu,\Sigma)$. Two variables $Y_i$ and $Y_j$ are said to be conditional independent if the correlation between them is equal to zero, that is:

$$Y_i \perp\!\!\!\perp Y_j \leftrightarrow corr(Y_i, Y_j) = 0$$

According to the structure of many real-world datasets, Mixed Interaction models are one of the possible solution to deal with domains problem. Specifically, by setting the strong assumption of homogeneous conditional Gaussian density, it is possible to define *Homogeneous Mixed Interacion model* (HMI-models) as a triplet of parameters, $\{g, h, K\}$ where $K$ is the **concentration matrix** and $g$ and $h$ are the log-linear expansion of the probability $P$, the so-called **canonical parameters**.

## 2.2 Model Selection

The model selection is not an easy task, even if in presence of a small number of variables the number of possible models could be very high. To fit all possible models could result computationally expensive. According to this, three main categories exist to select the appropriate graphical model:

1. **Constraints-based algorithms**. This use low-order conditional independence tests to infer the structure of the joint model;

2. **Score-based algorithms**. The optimal model is selected by minimizing the *score function*:

$$score(j) = -2logL(j) + k\ df(j)$$

   where $L(j)$ is the likelihood of the model $\mathcal{M}(j)$, $df(j)$ is the number of free (unconstrained) parameters of the model and $k$ is a penalty parameter.

3. **Bayesian Methods**. This category often involve the use of Markov Chain Monte Carlo methods.

By focusing on score-based algorithms, two metrics are commonly used to evaluate a model $\mathcal{M}(j)$: Akaike's Information Criterion (AIC), when $k = 2$

$$AIC(j) = -2[logL(j) - df(j)]$$

or Bayesian Information Criterion (BIC), when $k = \log(n)$

$$BIC(j) = -2logL(j) + log(n)df(j)$$

The *stepwise* learing procedure can work in two modalities *backward* or *forward*. In backward modality it uses the saturated model, the one with all possible arcs, and at each step, deletes the arcs that induces the maximum decrease in the selected metric; on the other hand, the forward modality starts by a null model, the one without arcs, and at each step, adds the possible arc accordingly to the chosen metric. The algorithm ends when there are no any model that improves the *score*. The metric-minimizing models is selected.

# 3 Anlaysis

Five Mixed Interaction model were investigated in order to find the best relationship among the chosen set of variables; three of them work in backward modality, while the others in forward mode.

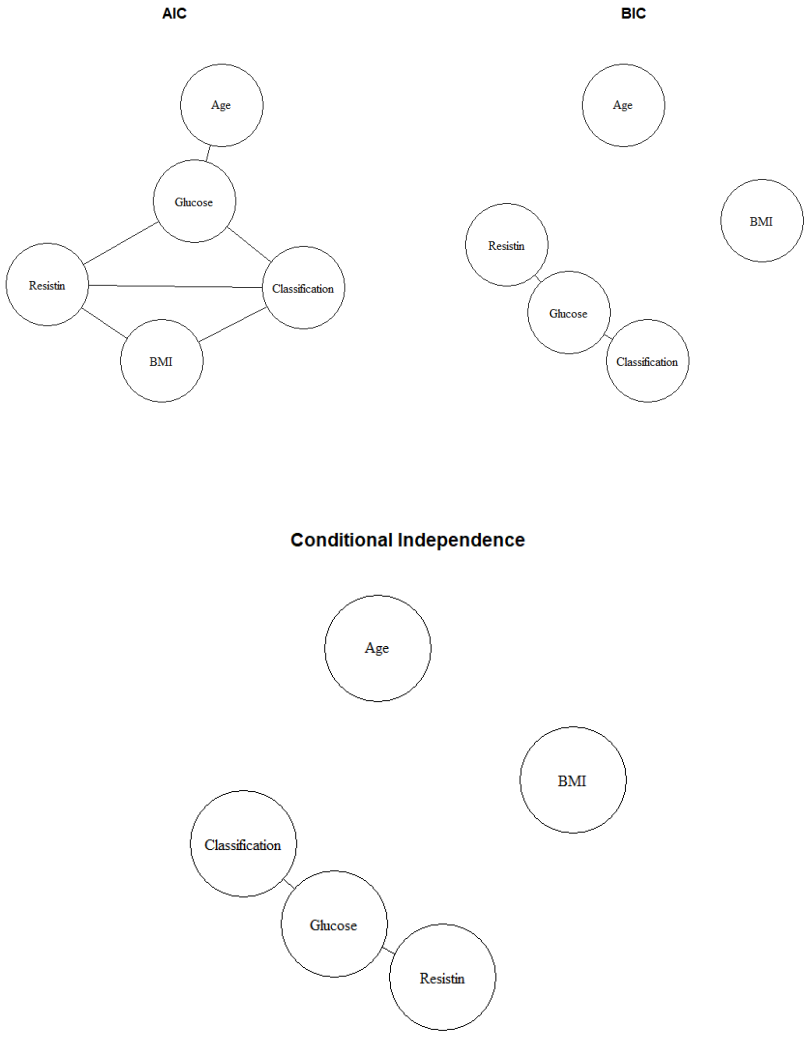



Figure 2: Backward modality

The model obtained using the BIC score function and the Conditional indepen-

dence test are more parsimonious in terms of edges with respect to the AIC model. Both BIC and Conditional independence test model have 8 degree of freedom, which corresponds to the constrained parameter; the AIC model has only 4 degree of freedom.

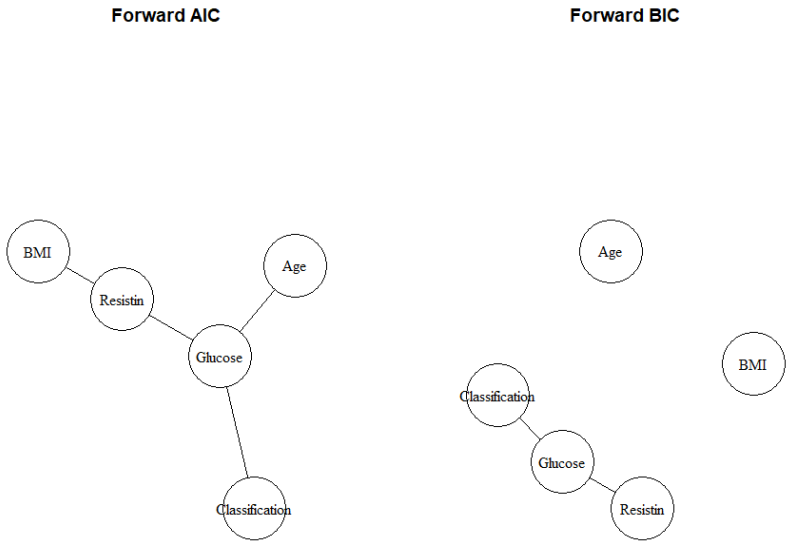


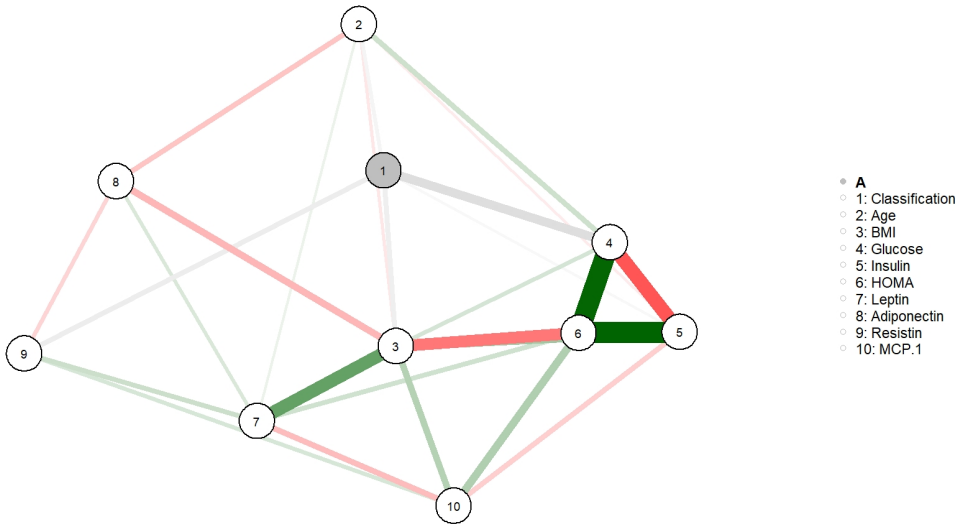

Figure 3: Forward modality

The other two models obtained with forward modality show that classification is linked only with glucose; the number of constrained parameters in the forward AIC increased to 6, while in the case forward BIC they remain the same.
In order to select the best model between those above, a possible solution is to focus the attention on the *deviance* $\mathcal{D}$. The deviance is the twice log-likelihood ratio between the saturated and the testing model and has asymptotic Chi-square $\mathcal{X}^2$ distribution with degree of freedom equal to the number of constrained parameters: $\mathcal{D} = 2(l_0 - l(\mathcal{M}))$.
According to the results shown in the table below the best model to be selected is the AIC implemented with backward modality, since it has the highest p-value with respect to the others models.

| Modality | Tested model | p-value |
|---|---|---|
| *Backward* | BIC | 0.1727523 |
| | AIC | 0.5401098 |
| | Conditional independence | 0.1727523 |
| *Forward* | BIC | 0.1727523 |
| | AIC | 0.4086012 |

Table 1: P-value results

In addition to this analysis a predictive one is conducted. All the variables contained in the dataset were now considered in order to confirm whether the dependence within a given set of variables is consistent with the objective.
The R function *mgm* was used to perform the prediction and to inspect which set of given variables could be considered meaningful for the analysis.
In principle the dataset is randomly split into traing and test dataset. By executing the *mgm* function, it returns a list of entries: fit_mgm$pairwise contains the weighted adjacency matrix and the signs (if defined) of the parameters in the weighted adjacency matrix; by using both of them is possible to generate a plot like the one below.

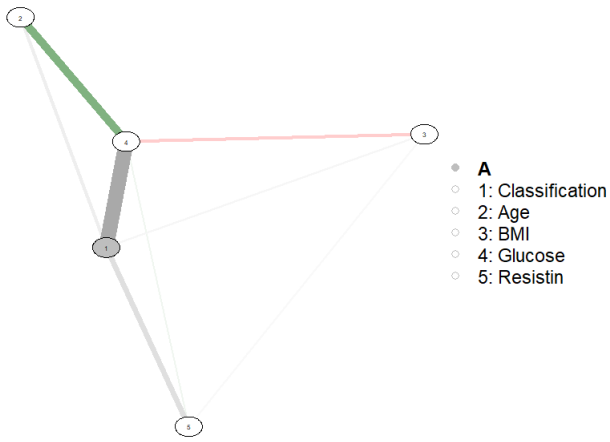



The categorical variable (i.e. *Classification*) is the one in gray and it is linked to continuous variables by gray arcs; green arcs connect two continuous variables in case of positive correlation while red edges in case of negative correlation. The

positive and negative correlations between continuous variables are consistent with studies made in this field. By using the function predict() it was possible to compute predictions and nodewise errors from the model estimated previously. Four different errors were investigated: **Root Mean Squared Error**, RMSE and **R2** as error functions for the continuous variables and the proportion of **correct classification**, CC and the **normalized proportion of correct classification**, nCC for categorical variables.

# 4 Conclusions

Overall, combining BMI and Age together with Glucose and Resistin can be considered a powerful tool in order to detect whether a patient is affect or not to breast cancer with an accuracy $> 85\%$.




This result is consistent to those in the literature, which suggest that Resistin and Glucose, taken together with Age and BMI, may be considered a good set of candidates for breast cancer biomarkers to implement into screening tests.[2]

# 5 Appendix: R Code

```
# Import libraries
library(gRbase)
library(gRim)
library(ggplot2)
library(gridExtra)
library(stats)
library(igraph)
library(mgm)
library(arm)

# Read the dataset
setwd("C:/Users/Utente/Desktop/UniMi/Probabilistic Modeling/Project")
breast_cancer <- read.csv(file = 'breast_cancer.csv')
breast_cancer <- breast_cancer[c(10,1,2,3,8)]

breast_cancer$Classification[breast_cancer$Classification == 1] <- "Healthy"
breast_cancer$Classification[breast_cancer$Classification == 2] <- "Patients"
breast_cancer$Classification <- factor(breast_cancer$Classification)

# Density distribution of continuous variables with respect to Classification
d1 <- ggplot(breast_cancer, aes(x = Age)) + geom_density(aes(fill = Classification),
     alpha= 0.5) +  scale_fill_manual(values = c("#228b22","#ff8c00")) + theme_bw()
d2 <- ggplot(breast_cancer, aes(x = BMI)) + geom_density(aes(fill = Classification),
     alpha= 0.5) +  scale_fill_manual(values = c("#228b22","#ff8c00")) + theme_bw()
d3 <- ggplot(breast_cancer, aes(x = Glucose)) + geom_density(aes(fill = Classification),
     alpha= 0.5) +  scale_fill_manual(values = c("#228b22","#ff8c00")) + theme_bw()
d4 <- ggplot(breast_cancer, aes(x = Resistin)) + geom_density(aes(fill = Classification),
      alpha= 0.5) + scale_fill_manual(values = c("#228b22","#ff8c00")) + theme_bw()
x11()
grid.arrange(d1,d2,d3,d4, nrow = 2, ncol=2)

SS <- CGstats(breast_cancer)
SS

apply(SS$center,1,sd)/apply(SS$center,1,mean)

# Saturated Model
ms <- mmod(~Classification*Glucose*Resistin, data = breast_cancer)
can.parms <- ms$fitinfo$parms
can.parms

apply(can.parms$h,1,sd) / apply(can.parms$h,1,mean)
```

```
# Partial correlation matrix
pc <- cov2pcor(solve(can.parms$K))
pc

# Score-based selection
sat <- mmod(~.^., data = breast_cancer)
forw <- mmod (~.^1 , data = breast_cancer)

m_aic <- stepwise(sat, data = breast_cancer)
m_bic <- stepwise(sat, k=log(nrow(breast_cancer)), details = 0)
m_ci <- stepwise(sat, "test",alpha=0.05, details = 0, headlong = TRUE)
f_bic <- stepwise (forw ,k=log ( nrow (breast_cancer)), direction ="forward", details = 0)
f_aic <- stepwise (forw, direction ="forward", detail = 0)

x11()
par(mfrow=c(3,2))
plot(as(m_aic, "igraph"), main = "AIC",
     vertex.shape = "circle", vertex.label.font = 50, vertex.color = "white",
     vertex.label.color = "black", vertex.size = 65, edge.color = "black")

plot(as(m_bic, "igraph"), main = "BIC",
     vertex.shape = "circle", vertex.label.font = 65, vertex.color = "white",
     vertex.label.color = "black", vertex.size = 65, edge.color = "black")

plot(as(m_ci, "igraph"), main = "Conditional Independence",
     vertex.shape = "circle", vertex.label.font = 65, vertex.color = "white",
     vertex.label.color = "black", vertex.size = 65, edge.color = "black")

plot(as(f_aic, "igraph"), main = "Forward AIC",
     vertex.shape = "circle", vertex.label.font = 60, vertex.color = "white",
     vertex.label.color =  "black", vertex.size = 50, edge.color = "black")

plot(as(f_bic, "igraph"), main = "Forward BIC",
     vertex.shape = "circle", vertex.label.font = 60, vertex.color = "white",
     vertex.label.color = "black", vertex.size = 50, edge.color = "black")

# Deviance comparison
pchisq(m_bic$fitinfo$dev, m_bic$fitinfo$dimension["df"], lower.tail = FALSE)
pchisq(m_aic$fitinfo$dev, m_aic$fitinfo$dimension["df"], lower.tail = FALSE)
pchisq(m_ci$fitinfo$dev, m_ci$fitinfo$dimension["df"], lower.tail=FALSE)
pchisq(f_bic$fitinfo$dev , f_bic$fitinfo$dimension["df"], lower.tail = FALSE)
pchisq(f_aic$fitinfo$dev , f_aic$fitinfo$dimension["df"], lower.tail = FALSE)


# Prediction
breast_pred <- read.csv(file = 'breast_cancer.csv')
```

```
breast_pred <- breast_pred[c(10,c(1,2,3,4,5,6,7,8,9))]

breast_pred$Classification[breast_pred$Classification == 1] <- "Healthy"
breast_pred$Classification[breast_pred$Classification == 2] <- "Patients"
breast_pred$Classification <- factor(breast_pred$Classification)

log_reg <- bayesglm(Classification~., data = breast_pred, family = "binomial")
summary(log_reg)

smp <- floor(0.75 * nrow(breast_pred))
set.seed(123)
train_ind <- sample(seq_len(nrow(breast_pred)), size = smp)
train <- breast_pred[train_ind, ]
test <- breast_pred[-train_ind, ]

train$Classification <- ifelse(train$Classification == "Healthy" ,0 ,1)
test$Classification <- ifelse(test$Classification == "Healthy" ,0 ,1)

fit_mgm<- mgm ( data =train ,
              type = c("c", rep("g" ,9)) ,
              level = c(2, rep (1 ,9)) ,
              k=2,
              lambdaSel = "CV",
              lambdaFolds = 10,
              ruleReg = "OR")

f <- (fit_mgm$pairwise$wadj)
rownames(f) <- colnames(f) <- names(breast_pred)
f[1 ,]
x11()
qgraph :: qgraph (fit_mgm$pairwise$wadj , layout ="spring", repulsion = 1.2 ,
                  edge.color =fit_mgm$pairwise$edgecolor ,
                  nodeNames = colnames(train), legend = TRUE )



p_mgm <- predict ( fit_mgm , data = test ,
                     errorCon = c("RMSE", "R2"),
                     errorCat = c("CC", "nCC"))
p_mgm$errors
```

# References


[1] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

[2] Patrício, M., Pereira, J., Crisóstomo, J. et al. *Using Resistin, glucose, age and BMI to predict the presence of breast cancer*. BMC Cancer 18, 29 (2018). https://doi.org/10.1186/s12885-017-3877-1

[3] Feng Z., Zhang H., *Resistin and Cancer Risk: A Mini-Review*. (2011)