DIETARY HABITS AND BREAST CANCER

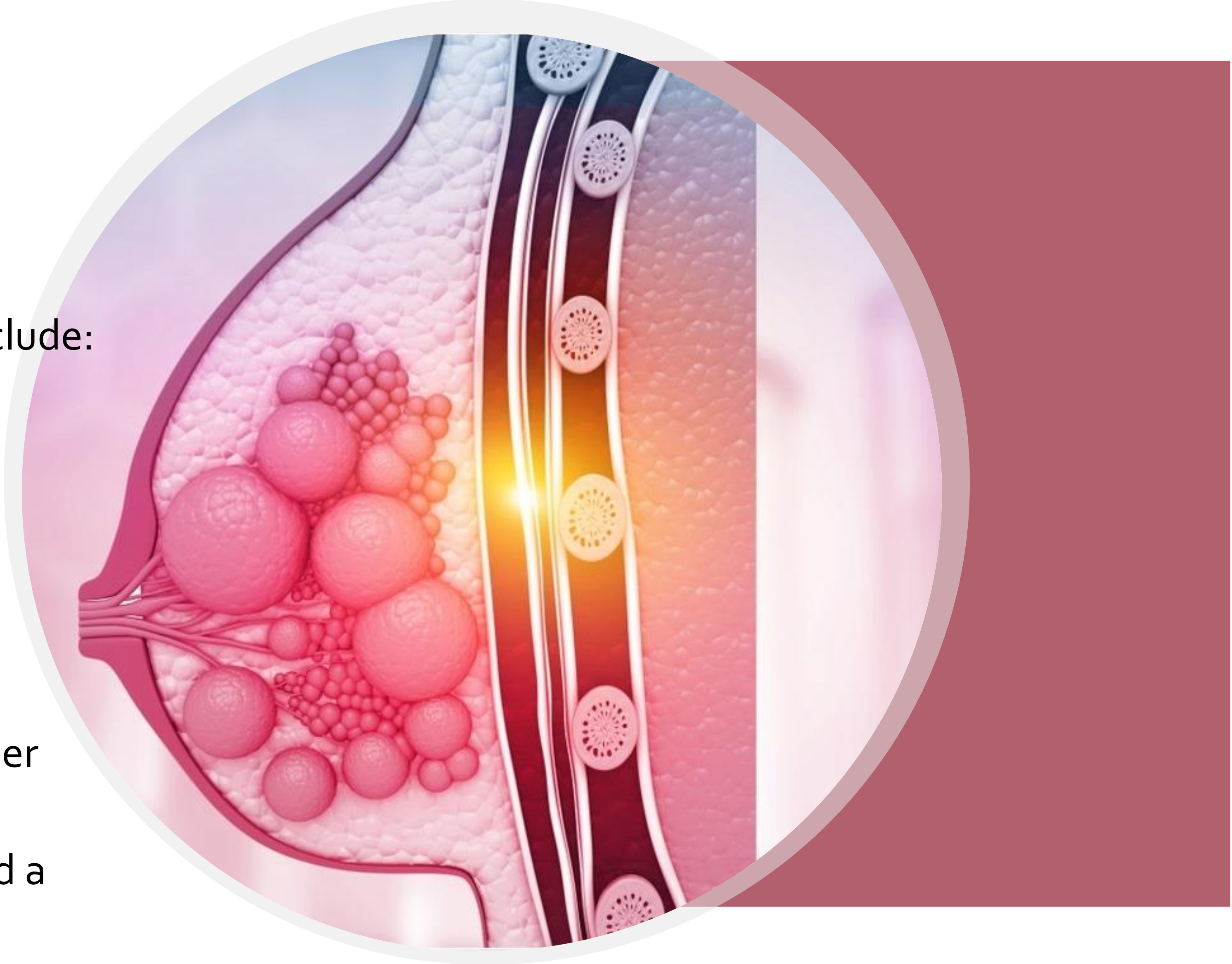
The Desease

What do we know?

Breast cancer is cancer that develops
from breast tissue.

Risk factors for developing breast cancer include:

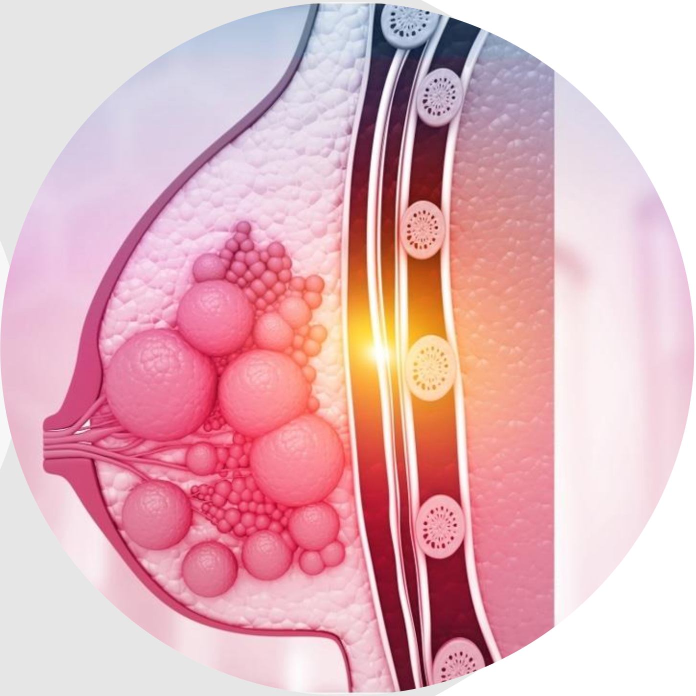
- being female
- obesity
- a lack of physical exercise
- alcoholism
- an early age at first menstruation
- having children late in life or not at all, older age
- having a prior history of breast cancer, and a family history of breast cancer.



Is diet associated to cancer risk?

So far...

A priori and *a posteriori* dietary patterns allowed to consider diet as an overall exposure and were weakly associated with breast cancer risk



Our goal:

Find new methods to consider diet as an overall exposure and its relationship with the breast cancer risk

Strategy

STEP 1

Extract *a posteriori* dietary patterns from the available data

STEP 2

Identify an association between breast cancer and dietary patterns

Traditional approach

STEP 1

Extract *a posteriori* dietary patterns from the available data

Dietary patterns identified using data driven statistical methods as for example:

- Principal Component Analysis
- Factor Analysis
- Cluster Analysis

Our approach

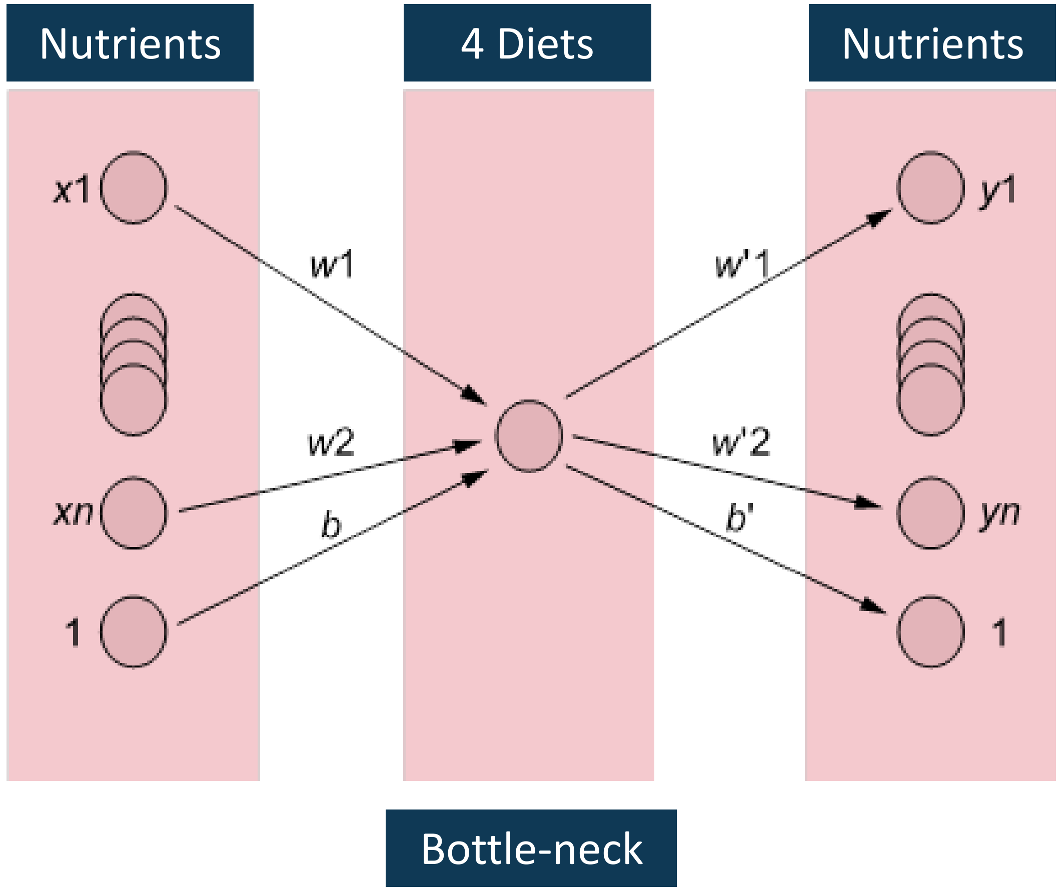
STEP 1

Extract *a posteriori* dietary patterns from the available data

Dietary patterns identified using a neural network typically used for dimensionality reduction:
Autoencoder

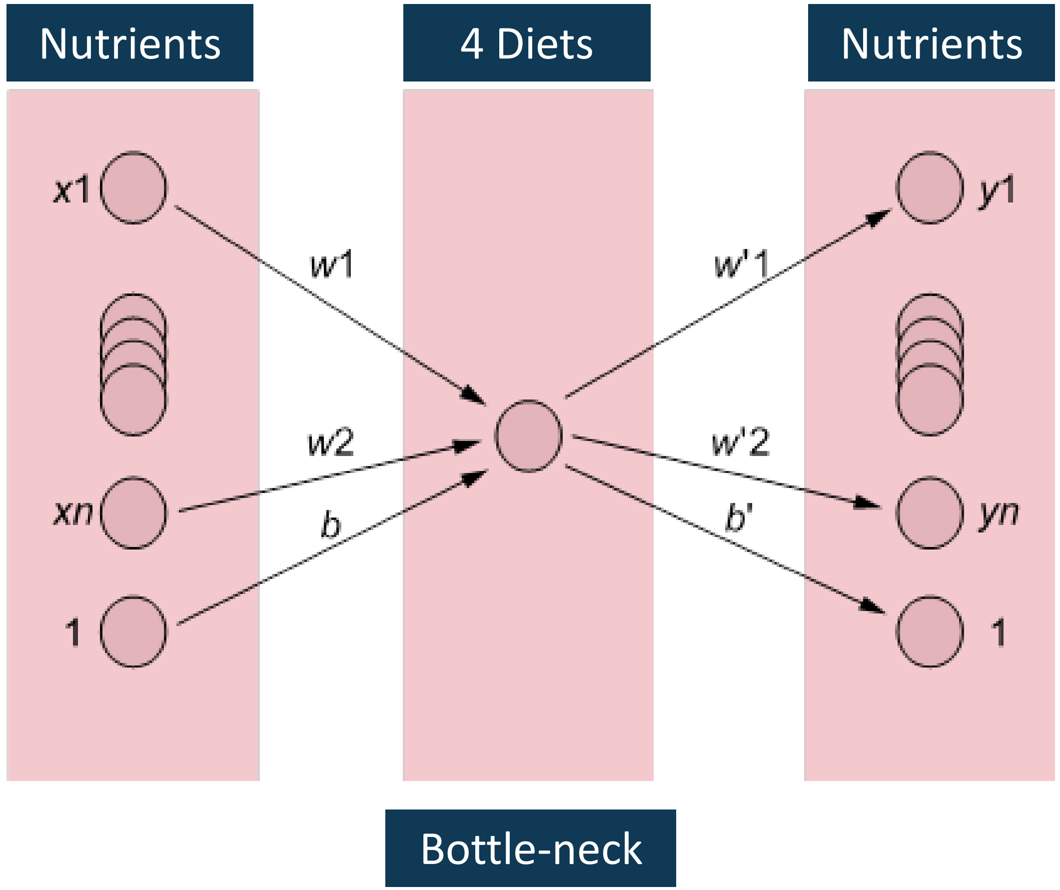
Autoencoder

It is a neural network based model to compress the data. Therefore, it has the ability to learn the compressed representation of our input data

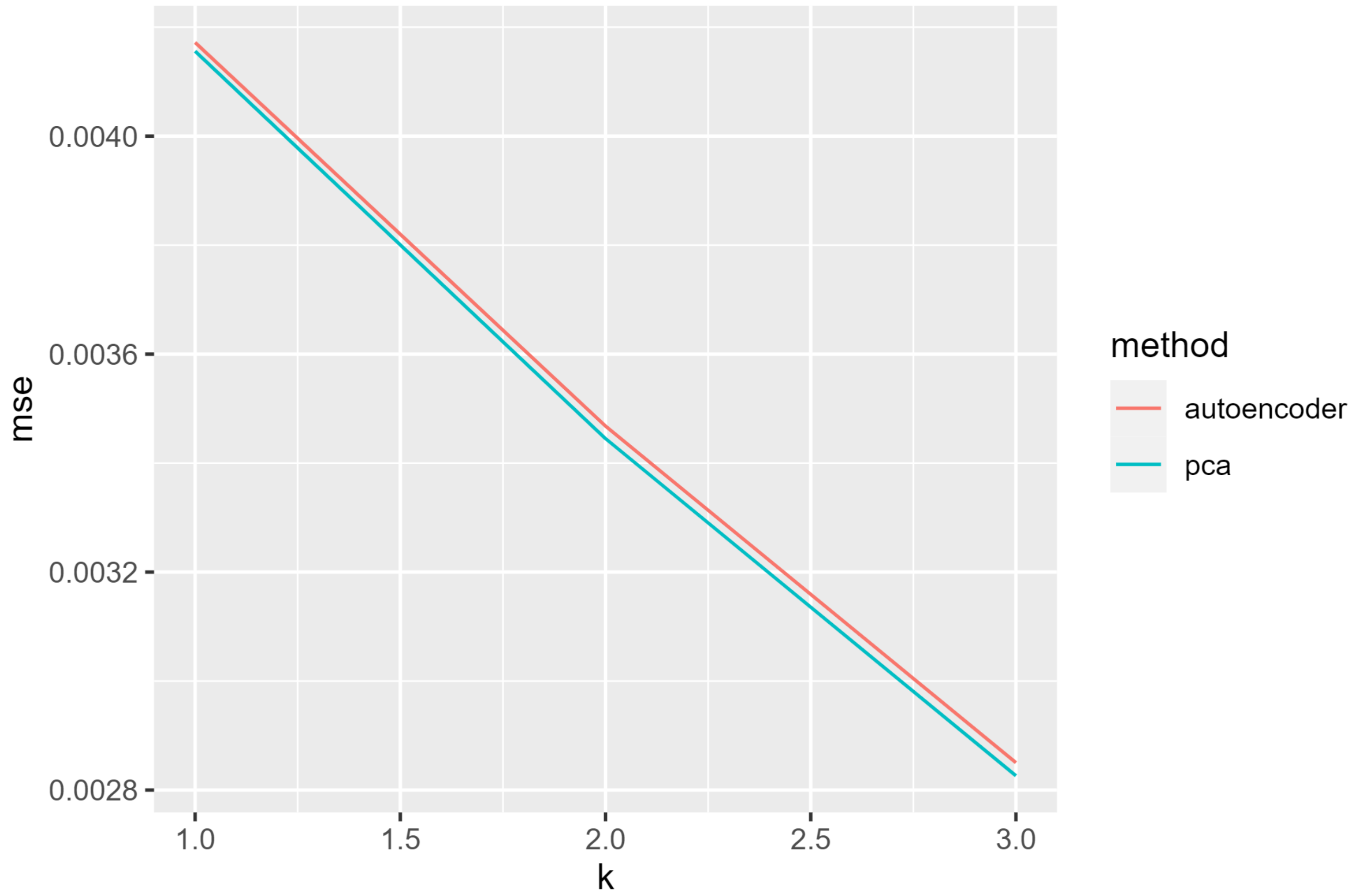


Autoencoder

In our application the first layer was populated with 27 nutrients, then reduced into 4 dimensions, which correspond to the 4 dietary patterns extracted



Autoencoder vs PCA performance



Autoencoder vs PCA

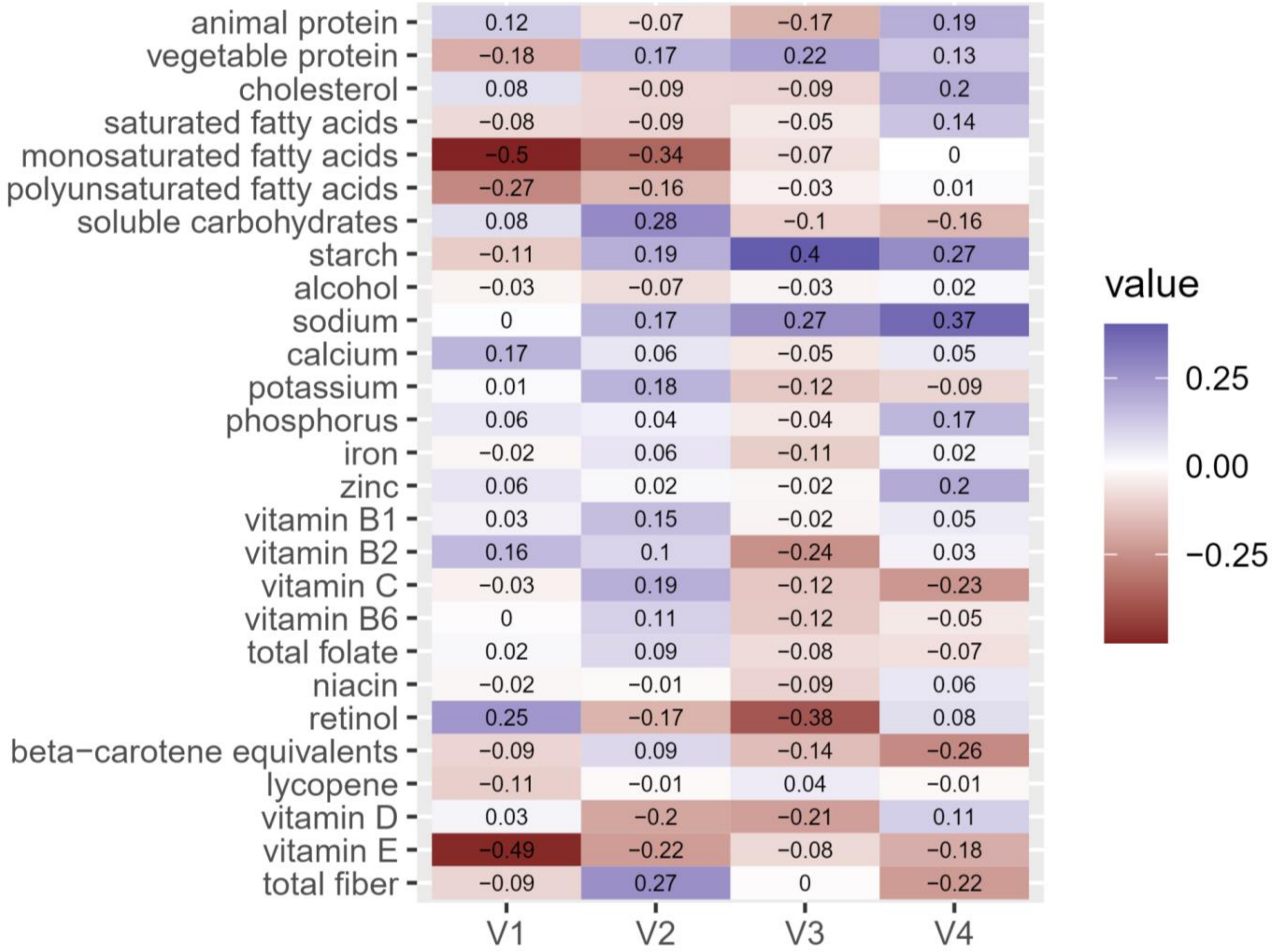
- PCA is essentially a linear transformation but Auto-encoders are capable of modelling complex non-linear functions

- PCA features are totally linearly uncorrelated with each other since features are projections onto the orthogonal basis. But auto-encoded features might have correlations, thus allowing to reconstruct more realistic dietary patterns

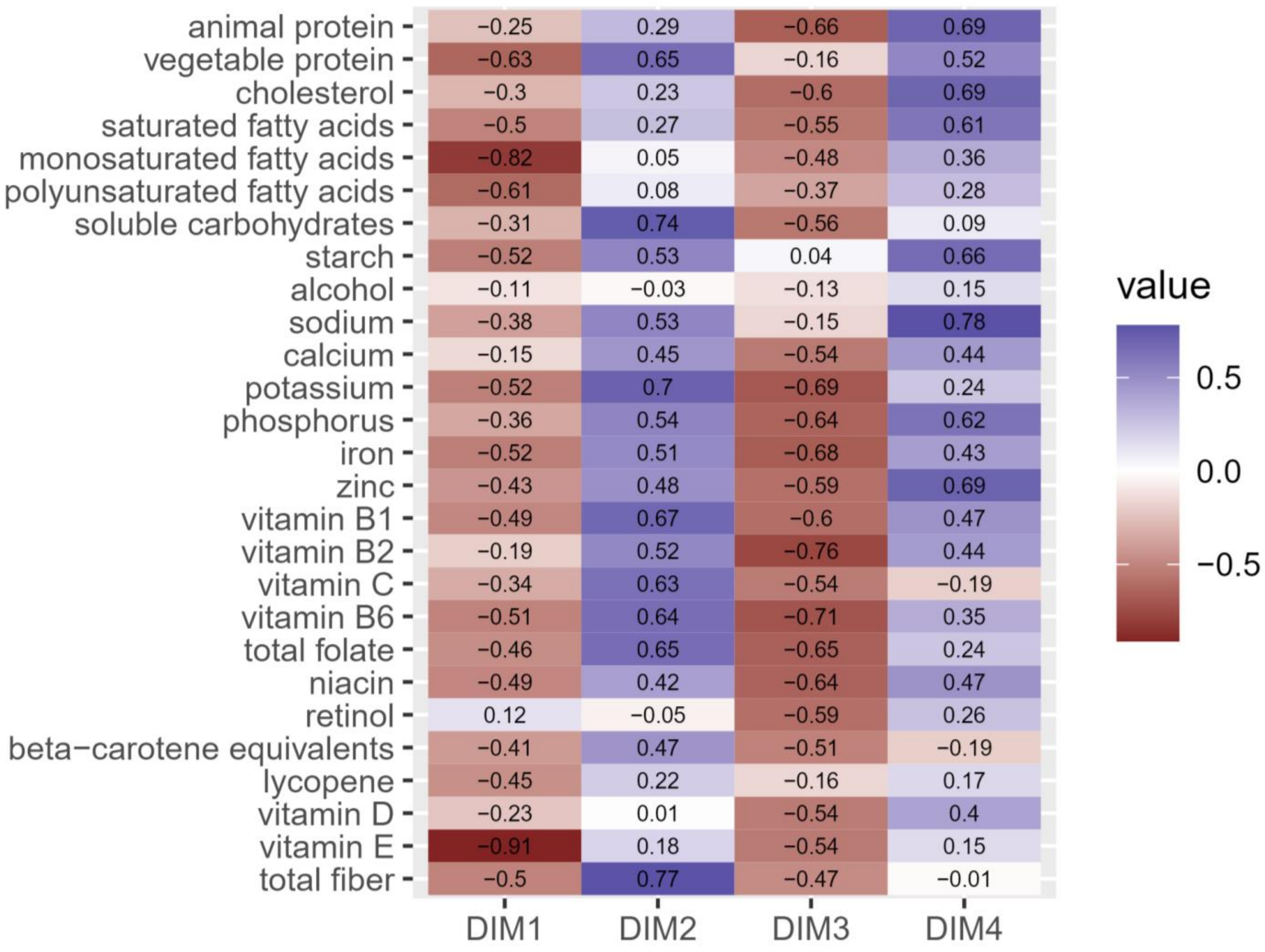
- A single layered autoencoder, like the one used in this case, is very similar to PCA, but with some advantages

- Autoencoders results are harder to interpret

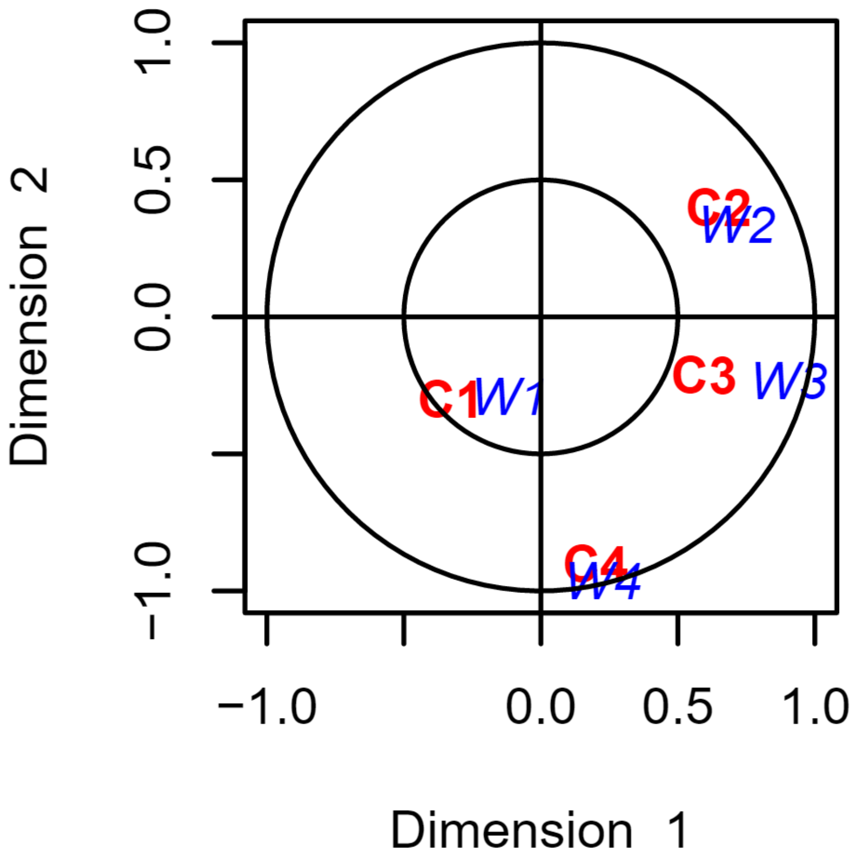
Autoencoder Results: layer weights



Autoencoder Results: correlations



Autoencoder Results: correlations vs weights



Autoencoder Results: interpretation

Weights	Correlation	Dietary Pattern Name
V1	DIM1	Animal products vs vegetable fats
V2	DIM2	Sugar vs vegetable fats
V3	DIM3	Starch vs animal products
V4	DIM4	Sodium vs beta-carotene

Traditional approach

Analyze the relationship between risk and disease with a logistic regression model and interpret the odds ratio adjusted for counfounding variables

STEP 2

Identify an association between breast cancer and dietary patterns

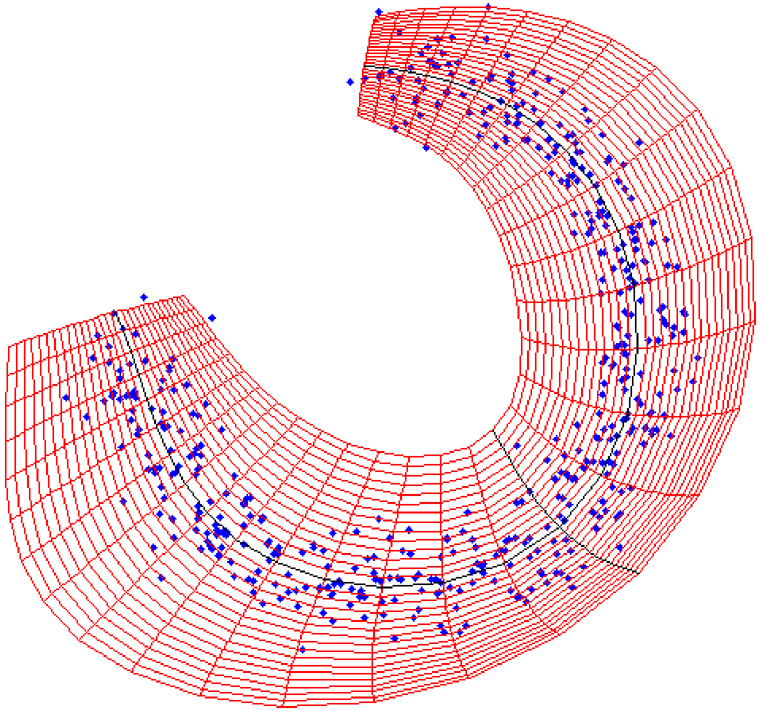
Our approach

Apply a Bayesian Network and describe the links among variables using the resulting conditional dependencies

STEP 2

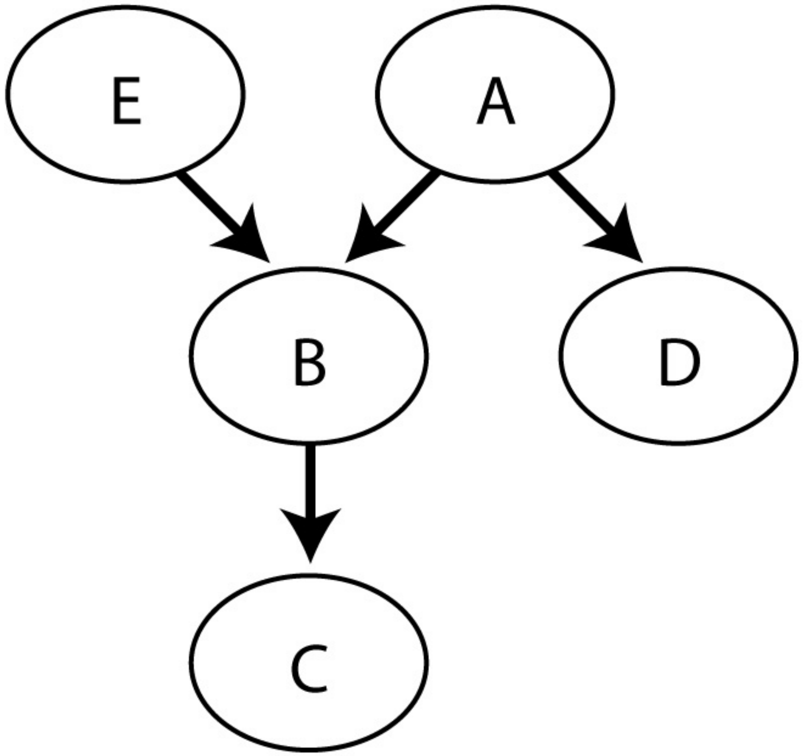
Identify an association between breast cancer and dietary patterns

Forward Imputation



It is an algorithm which applies non-linear PCA in order to impute missing values in matrices with ordinal data

Bayesian Network



It is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph

Logistic Regression Results

List of confounding factors: Education, Menopausal status, Number of Children, Smoking Status, Alcohol Status, Physical activity in 15-19 years old.

Unadjusted Model:

```
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.007242   0.027923  -0.259 0.795365    
X1          -0.018565   0.028802  -0.645 0.519214    
X2           0.072941   0.028833   2.530 0.011414 *  
X3           0.096806   0.029129   3.323 0.000889 ***
X4           0.095634   0.028484   3.358 0.000787 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Adjusted Model:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.656214   0.162649  -4.035 5.47e-05 ***
X1          -0.026122   0.029550  -0.884 0.376705    
X2           0.065373   0.029227   2.237 0.025302 *  
X3           0.132125   0.029752   4.441 8.96e-06 ***
X4           0.090980   0.028981   3.139 0.001693 ** 
V12          0.072488   0.008208   8.831  < 2e-16 ***
GIN4        -0.034134   0.033738  -1.012 0.311661    
V11         -0.064463   0.021488  -3.000 0.002700 ** 
FUM1         0.023659   0.042001   0.563 0.573230    
ALC1         0.169106   0.050267   3.364 0.000768 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Unadjusted vs Adjusted odds ratio:

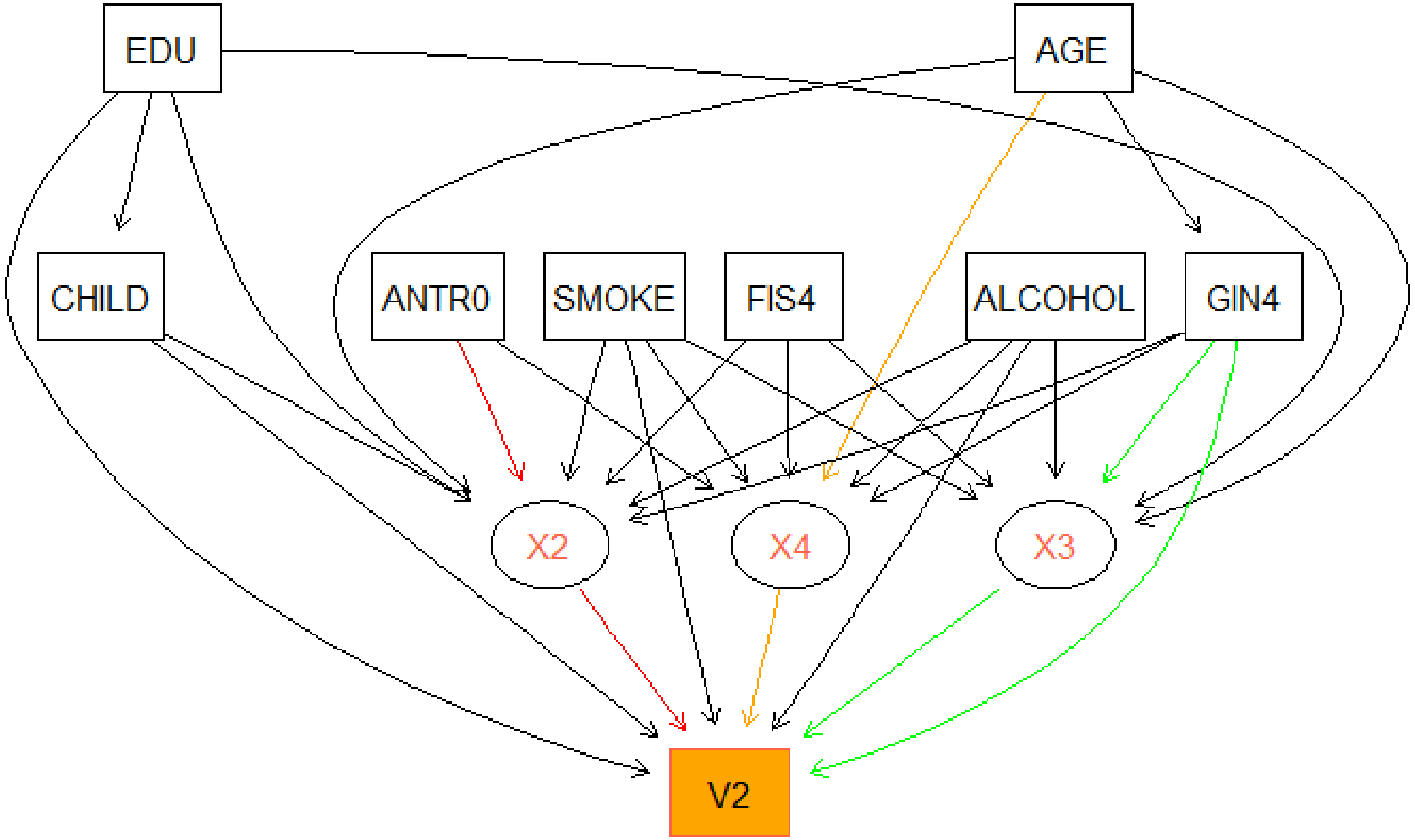
```
(Intercept)          X1          X2          X3          X4 
  0.9927843   0.9816066   1.0756674   1.1016468   1.1003560 
```

```
(Intercept)          X1          X2          X3          X4         V12        GIN4         V11 
  0.5188119   0.9742165   1.0675567   1.1412513   1.0952470   1.0751798   0.9664423   0.9375706 
       FUM1        ALC1 
  1.0239415   1.1842451 
```

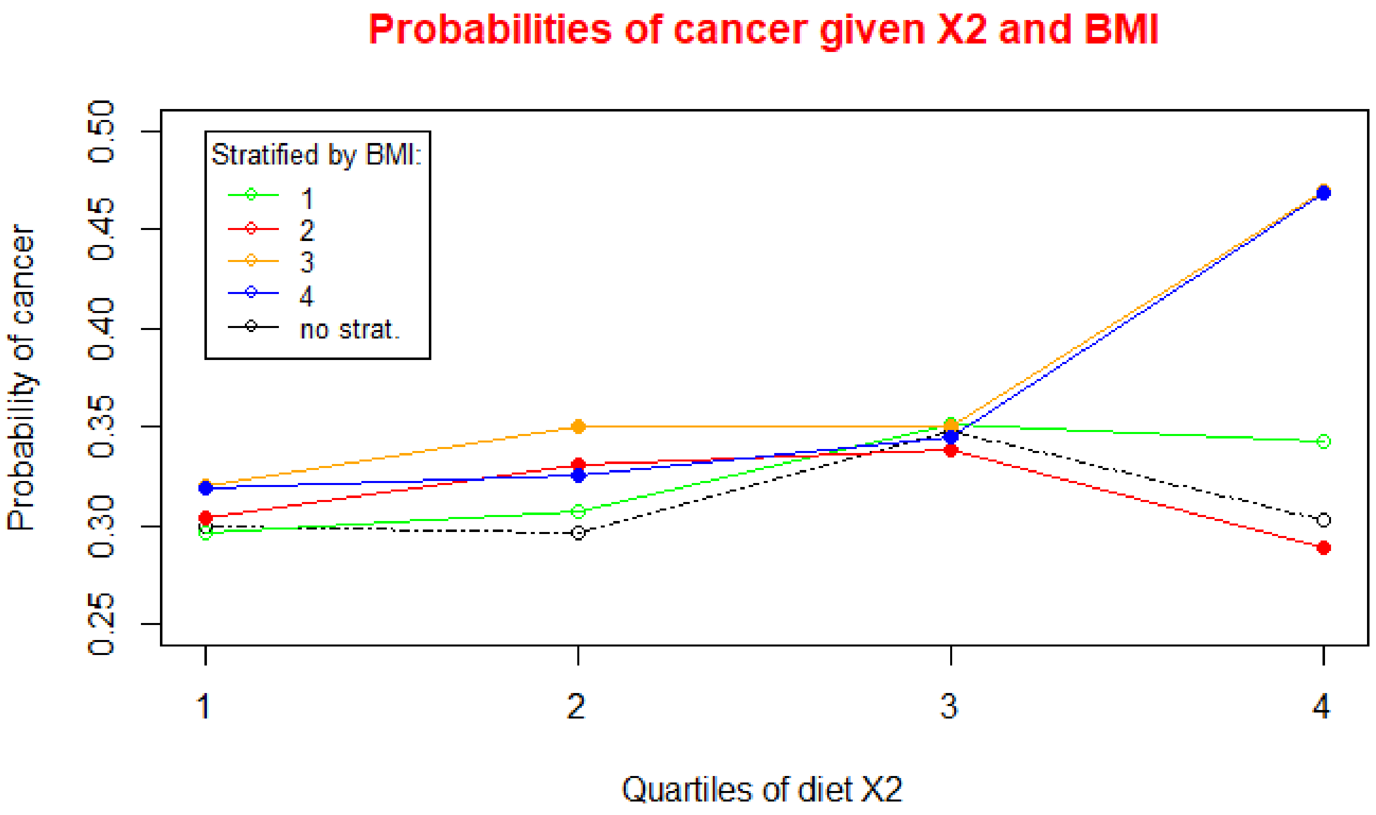
Autoencoder Results: interpretation

Weights	Correlation	Dietary Pattern Name	Impact on breast cancer
V1	DIM1	Animal products vs vegetable fats	-
V2	DIM2	Sugar vs vegetable fats	↑ Sugar ↓ Vegetable fats
V3	DIM3	Starch vs animal products	↑ Starch ↓ Animal Products
V4	DIM4	Sodium vs beta-carotene	↑ Sodium ↓ Beta-carotene

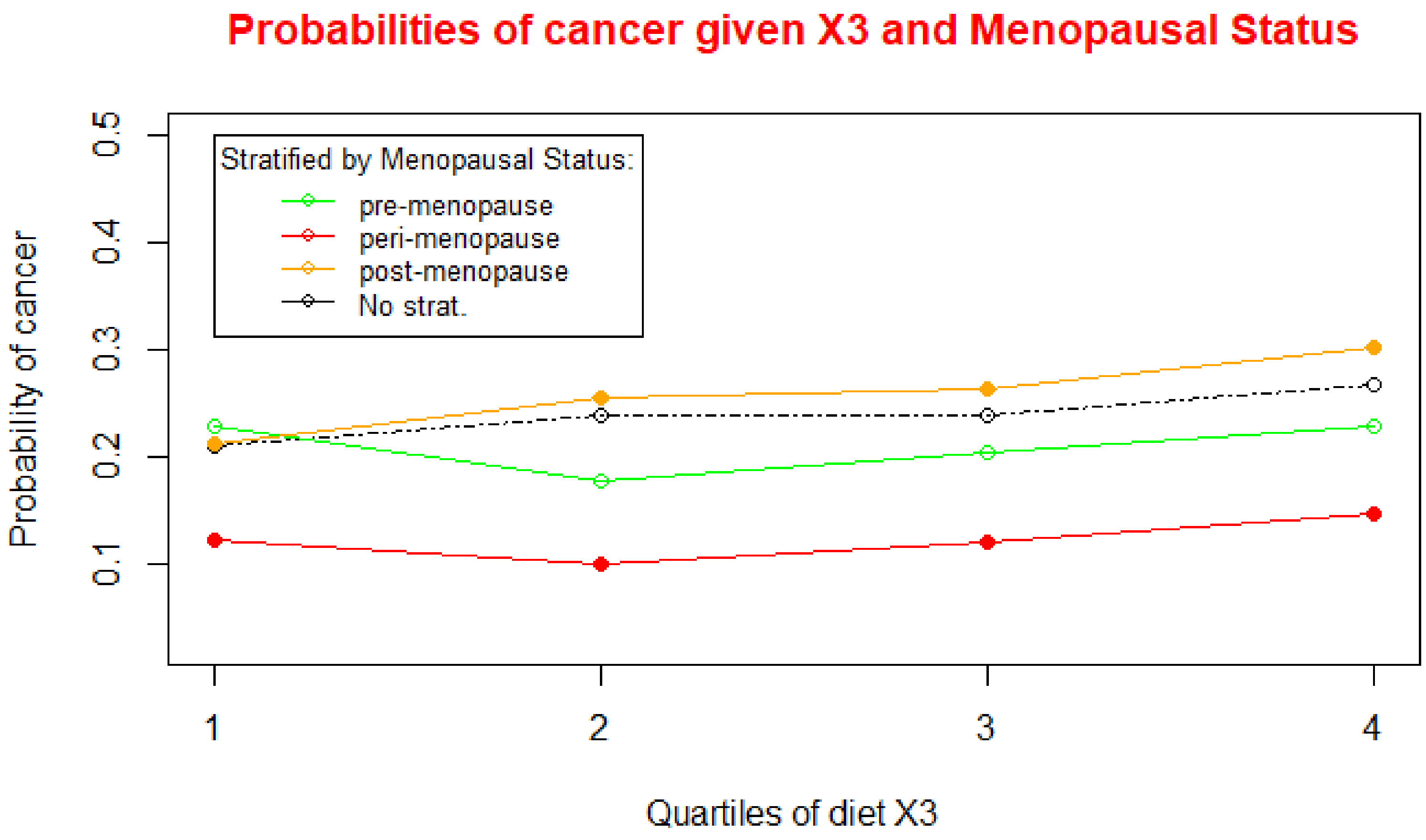
Bayesian Network Results



Bayesian Network – Some examples



Bayesian Network – Some examples



Bayesian Network – Some examples