Mixture of contaminated normal distributions with different variables inflation factor within classes

Jorge Sanchez

2024-02-01

Introduction

The traditional contaminated mixture model assumed that the contamination is the same for all variables within groups. The contaminated mixture model each group with a mixture of two normal distributions with two components. The first normal distribution models the non-contaminated samples while the second component models the contaminated samples. The contamination is control by two parameters which are the proportion of non-contaminated samples in each group α_g and the inflation factor η_g that is the same for all variables within group.

$$f(\mathbf{x}|\vartheta) = \sum_{g=1}^{G} \pi_g \left[\alpha_g \mathcal{N}(x|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + (1-\alpha_g)\mathcal{N}(x|\boldsymbol{\mu}_g, \eta_g \boldsymbol{\Sigma}_g)\right]$$

There are cases where the assumption that the inflation factor is the same for all variables measured in an observation within groups might be unrealistic. It is possible that the characteristics or variables being contaminated are a few instead of a all variables. To model this scenario the previous equation can be modified by replacing the scalar η_g that is the inflation factor for all variables within group g by a matrix N_g which is a diagonal matrix where each element of the diagonal η_{gj} for $j = 1, \ldots, p$ represent the inflation factor for the corresponding variable.

$$f(\mathbf{x}|\vartheta) = \sum_{g=1}^{G} \pi_g \left[\alpha_g \mathcal{N}(x|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + (1-\alpha_g)\mathcal{N}(x|\boldsymbol{\mu}_g, \dot{N}_g \boldsymbol{\Sigma}_g \dot{N}_g^T)\right]$$

$$\dot{N}_g = \begin{bmatrix} \sqrt{\eta_{g1}} & 0 & \ldots & 0 \\ 0 & \sqrt{\eta_{g2}} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sqrt{\eta_{gp}} \end{bmatrix}$$

Simulation study: Comparison between mixtures with equal and different inflation factors within group

Overview

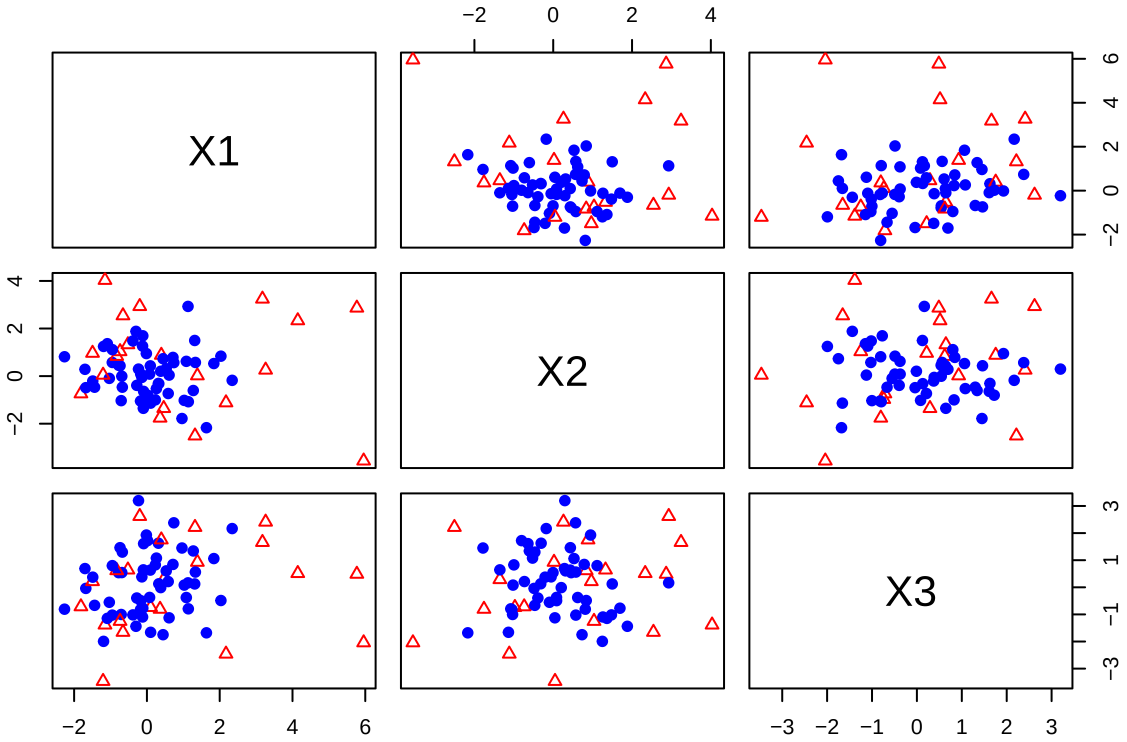
In this section, the behavior of contaminated mixture normal models assuming equal and different variables inflation factor within group is investigated. To generate the data the following process is conducted to generate three datasets where 75% of the samples composed the training subset and the remaining are part of the test subset.

a) A contaminated dataset of 100 samples in $p = 3$ dimensions and same variable inflation factor within group and $G = 1$;
b) A contaminated dataset of 100 samples in $p = 3$ dimensions with different variable inflation factor within group and $G = 1$;
c) A contaminated balanced dataset of 100 samples in $p = 3$ dimensions and $G = 2$ with equal proportions for both groups and one group with same variable inflation factor and another group with different variable inflation factor;

The parameters for the first dataset are

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma = I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \alpha = 0.8, \eta = 5$$

```
pairs(DatasetA$Xtrain, col = c("red","blue")[DatasetA$vtrain + 1],
      pch = ifelse(DatasetA$vtrain == 0,24,19), cex = 1)
```



Looking at the training data above, it is clear that the inflation factor is the same across all variables with a few contaminated samples further than the contaminated samples cloud. This dataset is fitted firstly with the contaminated mixture of normals model assuming equal variable inflation factor within group and secondly with the model assuming different variable inflation factor within group.

The parameter estimates were obtained by using the command Cnmixt until covergence from the library ContaminatedMixt. Next, the self-coded E-step was used to obtain the labels whether a sample is non-contaminated or contaminated. The function Cnmixt assumes equal variable inflation factor within group and their parameters are shown below.

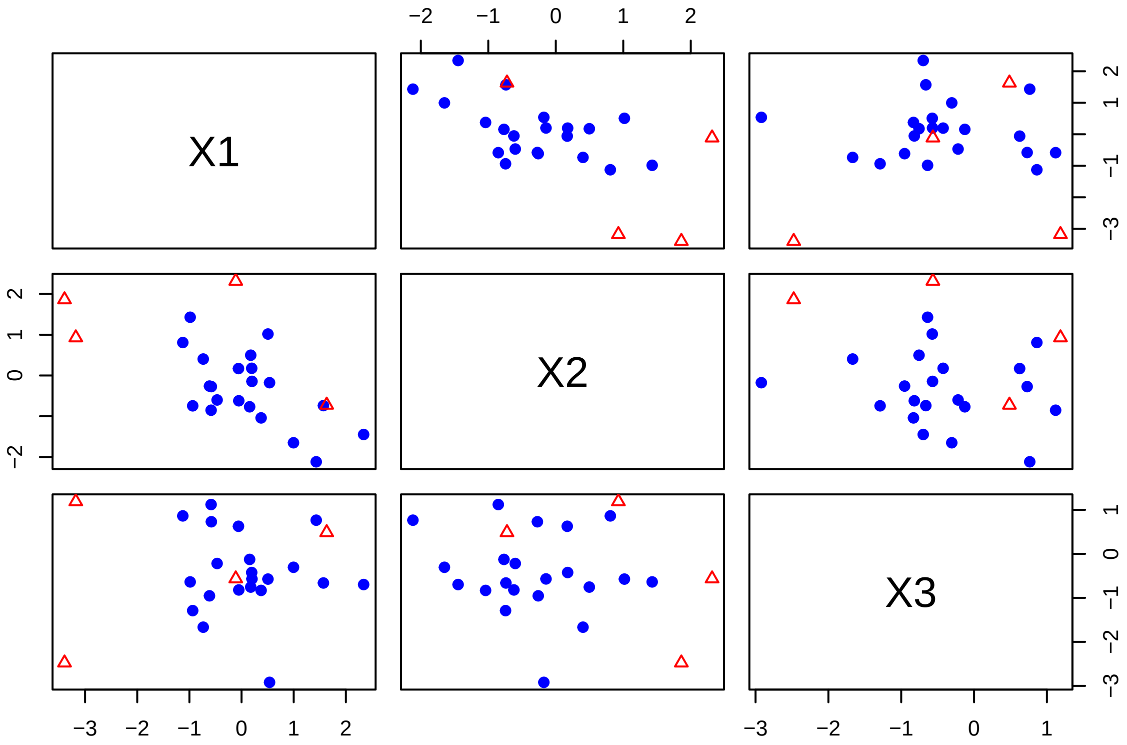
$$\mu = \begin{bmatrix} 0.07 \\ 0.15 \\ 0.13 \end{bmatrix}, \Sigma = \begin{bmatrix} 1.05 & 0 & 0 \\ 0 & 1.05 & 0 \\ 0 & 0 & 1.05 \end{bmatrix}, \alpha = 0.78, \eta = 4.95$$

The initial values for the parameters of the model were taken from the contaminated mixture model produced by the function Cnmixt from the ContaminatedMixt package that assumes same variable inflation factor within group. Next, an iteration of a self-coded m-step that assumes different variable inflation factor within group for EII model was run to obtain the estimated parameters shown below. It is possible to see that there are some differences between the estimates obtained for both models. The main difference between the estimates is observed in the values that the model with different variables inflation factor take for α which suggest a 51% of non-contaminated samples and implies a greater contamination that the other model. Also, the inflation factors in the diagonal of the matrix $\dot{N}$ are different and much smaller than 4.95.

$$\mu = \begin{bmatrix} 0.34 \\ 0.16 \\ 0.1 \end{bmatrix}, \Sigma = \begin{bmatrix} 1.14 & 0 & 0 \\ 0 & 1.14 & 0 \\ 0 & 0 & 1.14 \end{bmatrix}, \alpha = 0.51, \dot{N} * \dot{N}^T = N = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1.83 & 0 \\ 0 & 0 & 1.68 \end{bmatrix}$$

Next, a self-coded function E-step was used to obtain the estimates of labels corresponding whether the observations are non-contaminated or contaminated in the test set. It was observed that the estimates for α and N were affected for the choose of the initial values of ν_i which denotes if the i^{th} sample is non-contaminated if $\nu_i = 1$ otherwise is 0. The pairs plot of the samples in the test subset shows a little more dispersion for the pair X_1 and X_2 than for the pair X_1 and X_3.

```
pairs(DatasetA$Xtest, col = c("red","blue")[DatasetA$vtest + 1],
      pch = ifelse(DatasetA$vtest == 0,24,19), cex = 1)
```



The confusion matrix where the rows are the actual values and the columns the predicted values for the model assuming same variable inflation factor within group , shows that it was possible to identify half of the contaminated observations and all the non-contaminated observations correctly.

```
knitr::kable(t_AEq, format = "markdown")
```

	0	1
0	2	2
1	0	21

However, taking a look of the performance of the model assuming different variables inflation factor within group is also able to identify half of the non-contaminated samples and misclassified 3 non-contaminated samples as contaminated. As the data is generated for a model with same variable inflation factor within group, there is not surprise that it overcame their counterpart that assume different variable inflation factors within group. It is possible to confirm this result taking a look to the other metrics specially F1_Score.

```
knitr::kable(t_ADif, format = "markdown")
```

	0	1
0	2	2
1	3	18

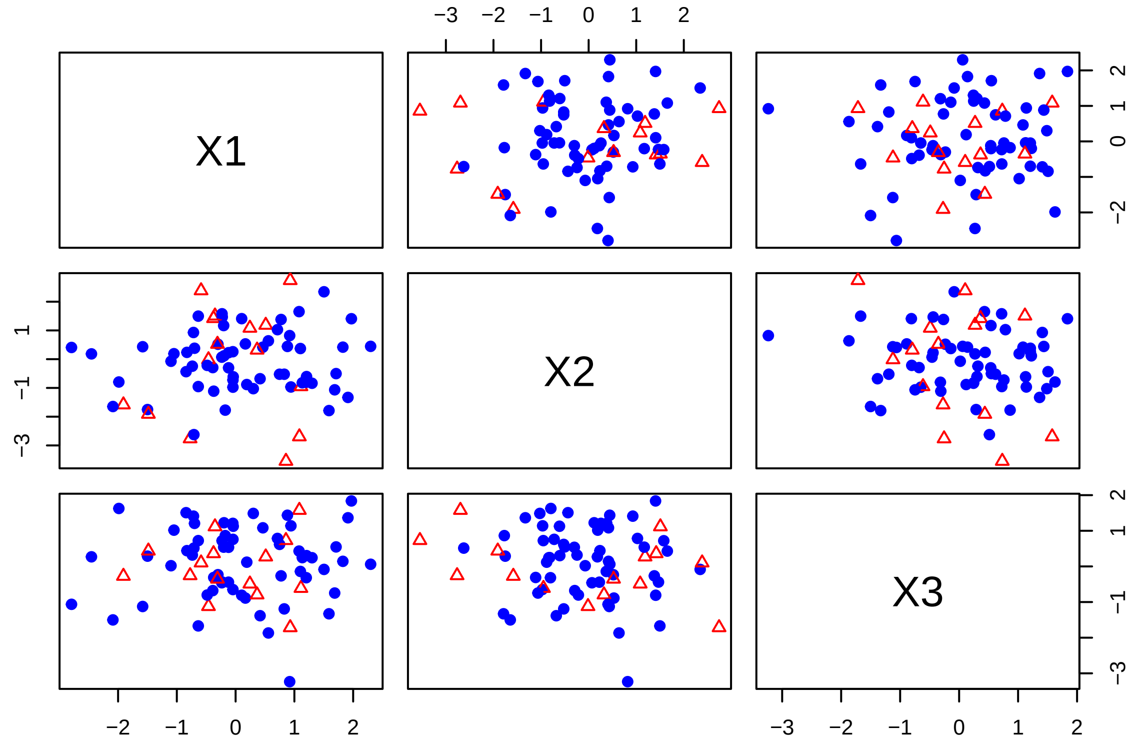
```
#knitr::kable(df_A, format = "markdown")
format_table(df_A)
```

```
##        Metric Equal_IF Different_IF
## 1    Accuracy     0.92         0.80
## 2   Precision     1.00         0.40
## 3      Recall     0.50         0.50
## 4 Sensitivity     0.50         0.50
## 5 Specificity     1.00         0.86
## 6    F1 score     0.67         0.44
```

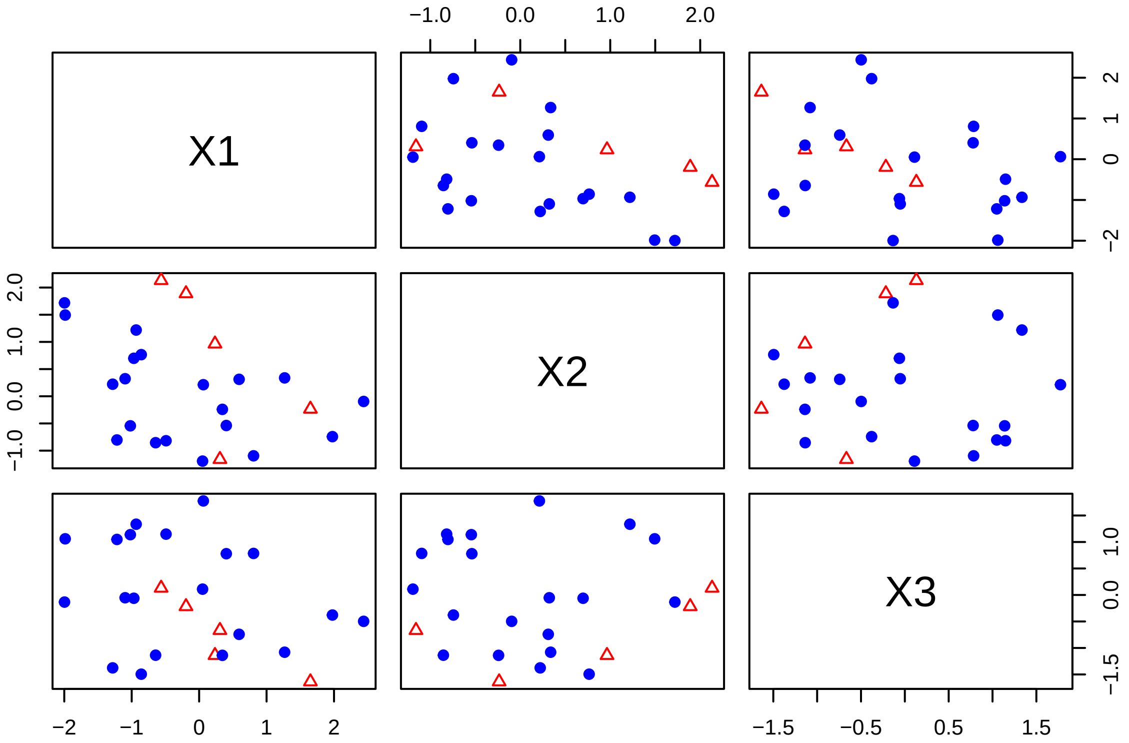
The parameters for the second dataset are

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma = I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \alpha = 0.8, N_g = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

```
pairs(DatasetB$Xtrain, col = c("red","blue")[DatasetB$vtrain + 1],
      pch = ifelse(DatasetB$vtrain == 0,24,19), cex = 1)
```



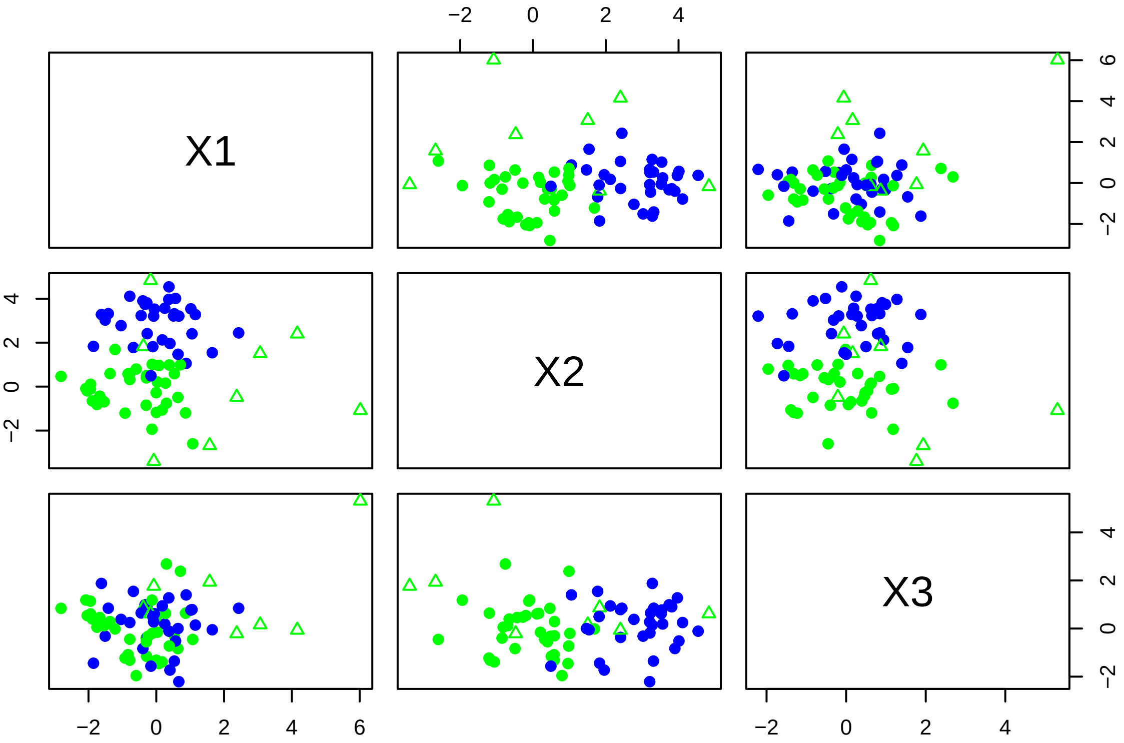
```
pairs(DatasetB$Xtest, col = c("red","blue")[DatasetB$vtest + 1],
      pch = ifelse(DatasetB$vtest == 0,24,19), cex = 1)
```



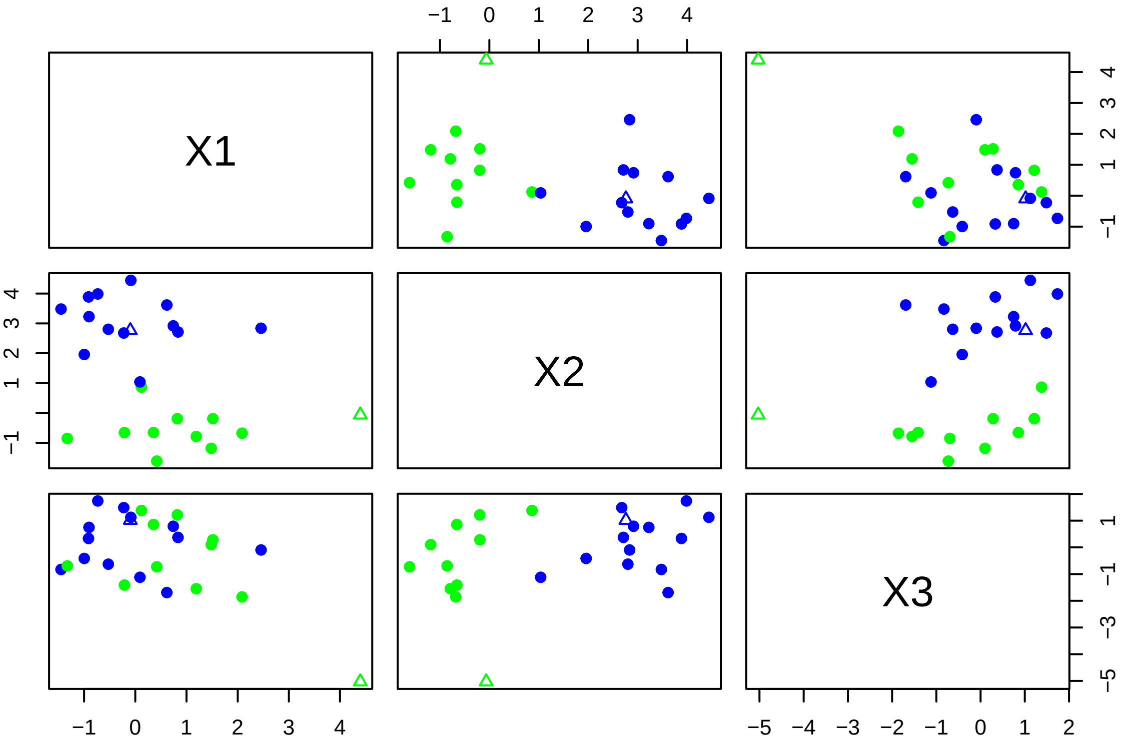
The parameters for the third dataset are

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0 \\ 5 \\ 0 \end{pmatrix} \Sigma_1 = I_3, \Sigma_2 = I_3, \alpha_1 = 0.8, \alpha_2 = 0.9, \eta_1 = 5, N_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{10} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

```
pairs(DatasetC$Xtrain, col = c("green","blue")[DatasetC$ltrain],
      pch = ifelse(DatasetC$vtrain == 0,24,19), cex = 1)
```



```
pairs(DatasetC$Xtest, col = c("green","blue")[DatasetC$ltest],
      pch = ifelse(DatasetC$vtest == 0,24,19), cex = 1)
```


X1
X2
X3
−1
0
1
2
3
4
−5
−4
−3
−2
−1
0
1
2