BANKNOTES IDENTIFICATION

Alberto Negri - 880254
Gaia Ghidoni - 890957

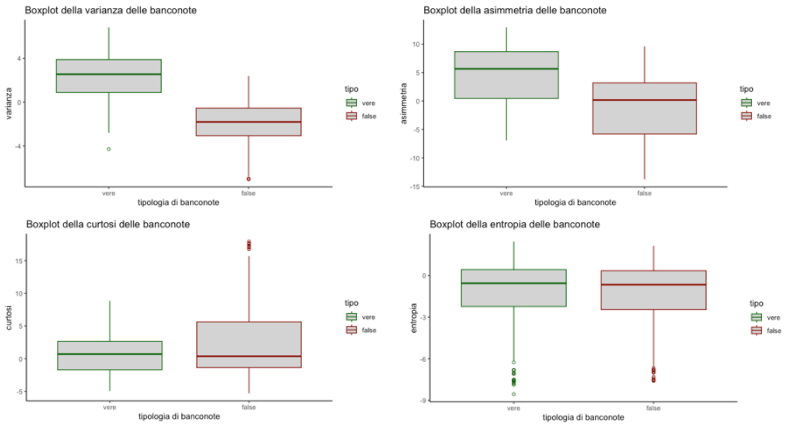
La presente analisi si propone di applicare metodi di clustering e classificazione al fine di distinguere tra banconote autentiche e contraffatte sulla base dei pixel di immagini delle stesse. Si tratta di un'applicazione di rilevante importanza, data l'entità della problematica delle banconote falsificate nel contesto economico. Questi approcci possono infatti essere utilizzati come strumento di previsione della veridicità delle banconote.

I dati utilizzati provengono da un dataset creato dal Professore Volker Lohweg per il medesimo scopo. I dati sono stati estratti da immagini digitalizzate, ottenute in scala di grigi, di banconote autentiche e contraffatte. Queste ultime sono di valute diverse (USD, euro serie 1, altri campioni): questo è causa di ulteriore discriminazione tra i valori delle banconote.
Il dataset contiene 1371 osservazioni e in nessuna di esse sono presenti dati mancanti. Le variabili sono 5: 4 numeriche e una qualitativa. Si inizi con una breve analisi descrittiva delle variabili del dataset.

```
    varianza        asimmetria        curtosi          entropia       tipo
Min.   :-7.0421   Min.   :-13.773   Min.   :-5.2861   Min.   :-8.5482   0:761
1st Qu.:-1.7747   1st Qu.: -1.711   1st Qu.:-1.5534   1st Qu.:-2.4170   1:610
Median : 0.4957   Median :  2.313   Median : 0.6166   Median :-0.5867
Mean   : 0.4314   Mean   :  1.917   Mean   : 1.4007   Mean   :-1.1922
3rd Qu.: 2.8146   3rd Qu.:  6.813   3rd Qu.: 3.1816   3rd Qu.: 0.3948
Max.   : 6.8248   Max.   : 12.952   Max.   :17.9274   Max.   : 2.4495
```

Le variabili quantitative denotano caratteristiche strutturali della banconota, riferendosi infatti ai pixel della stessa. Grazie alla rilevazione di esse è possibile discriminare tra una banconota reale e una contraffatta.
Nello specifico: la varianza è una misura del contrasto di colore (non si tratta dell'indice di variabilità in senso statistico, come si nota anche dalla presenza di valori negativi); l'asimmetria rileva il grado di asimmetria della distribuzione dei pixel; la curtosi dà indicazioni sulla forma della distribuzione delle intensità dei pixel; l'entropia fornisce una misura della casualità della distribuzione dei pixel.
Osservando gli indici descrittivi si noti come i range di variazione dei valori di ciascuna variabile siano abbastanza simili tra loro e di ampiezza contenuta.
Inoltre, non c'è evidenza di dati sbilanciati: la variabile qualitativa si distribuisce equamente tra le due categorie.

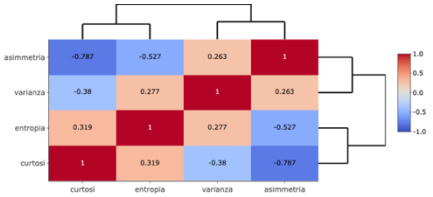
Per meglio descrivere le variabili numeriche ed evidenziare la presenza di eventuali outlier, si riportano i boxplot relativi a ciascuna di esse differenziando per tipo di banconote.



Si notano differenze tra le due tipologie di banconote soprattutto sui valori di varianza e asimmetria, mentre sull'entropia i valori sono molto simili. I pochi outlier presenti non destano preoccupazioni.

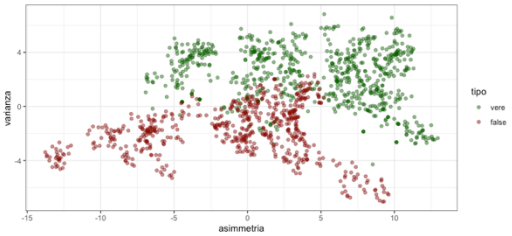
Da una prima analisi sono emerse 24 unità statistiche duplicate, tutte non contraffatte. Non necessariamente si tratta di un errore di imputazione dei dati, ma potrebbe derivare dal fatto che queste banconote provengano dagli stessi lotti. Venendo stampate in serie, infatti, appare sensato che assumano gli stessi valori sulle varie variabili. Per questo motivo non verranno rimosse dal dataset.

Si vada ora ad analizzare la struttura di correlazione tra le variabili quantitative.



Asimmetria e curtosi sono le variabili più correlate, con un valore di -0.787; anche asimmetria ed entropia hanno un valore discretamente alto di correlazione, di -0.526.

Si effettua ora l'analisi delle componenti principali, dalla quale emerge che le prime due componenti spiegano cumulativamente il 90% della variabilità totale. Le variabili ad esse associate sono varianza e asimmetria.
Osservando la distribuzione delle unità rispetto a quest'ultime si notano i due gruppi ma non in maniera distinta: ciò è probabilmente dovuto al fatto che nel dataset sono contenute banconote di valute differenti.



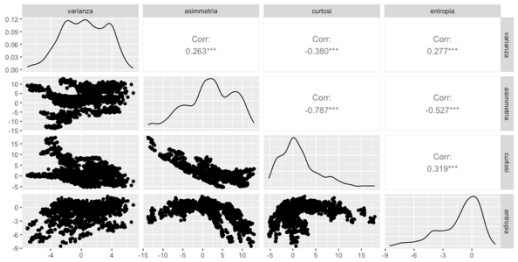
Innanzitutto, si implementa un modello di regressione logistica per valutare l'effetto delle variabili sulla classificazione binaria.

```
Call:
glm(formula = tipo ~ ., family = binomial, data = banconoteglm)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7000   0.0000   0.0000   0.0003   2.2461

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.3218     1.5589   4.697 2.64e-06 ***
varianza     -7.8593     1.7383  -4.521 6.15e-06 ***
asimmetria   -4.1910     0.9041  -4.635 3.56e-06 ***
curtosi      -5.2874     1.1612  -4.553 5.28e-06 ***
entropia     -0.6053     0.3307  -1.830   0.0672 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

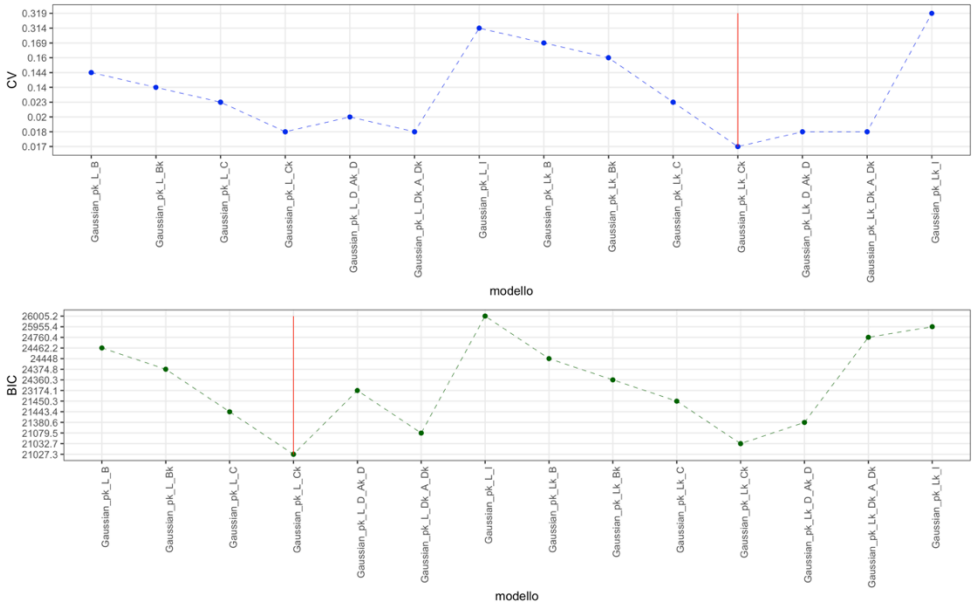
I coefficienti associati alle esplicative sono tutti altamente significativi, eccetto quello dell'entropia: quest'ultima non aiuta nella discriminazione tra banconote vere e false.
Qualora avessimo a disposizione dati di banconote non ancora classificate, questo potrebbe essere un primo metodo per fare previsione e classificarle.

MODEL BASED CLASSIFICATION

Avendo a disposizione la corretta classificazione di ciascuna banconota la prima tecnica presa in considerazione è la model based classification. Confrontando le vere etichette con quelle previste dal modello scelto in fase di allenamento verrà valutata la correttezza della classificazione fatta.
Potenzialmente, qualora l'errore di classificazione fosse trascurabile, questo approccio potrebbe essere usato per valutare la veridicità di nuove banconote di cui si sono rilevate le variabili numeriche ma non la tipologia.
Si osservi la densità delle variabili per avere un'indicazione approssimativa di quale possa essere il miglior metodo di classificazione.



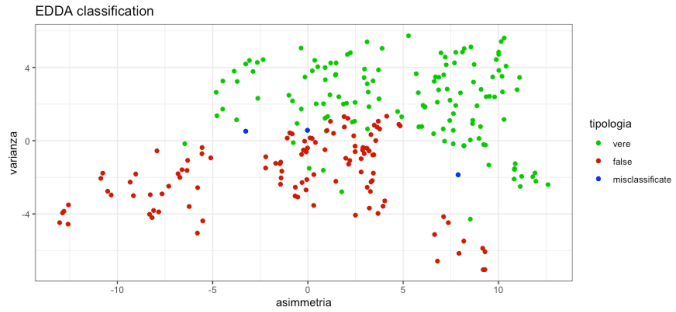
L'andamento della densità di curtosi ed entropia ricorda quello di una normale asimmetrica: proviamo a partire da un'analisi EDDA, che ipotizza come struttura della popolazione una mistura di normali d-dimensionali.
Questa viene realizzata allenando il classificatore su un training set di 1122 osservazioni (80%), che viene poi applicato alla restante parte di osservazioni, costituenti il test set.
Stimando più volte il modello, il migliore risulta essere Gaussian_pk_Lk_Ck (VVV). La scelta è stata fatta considerando come criteri sia il Cross Validation che il BIC e, per vedere come essi variano a seconda dei modelli, si riporta il grafico sottostante.



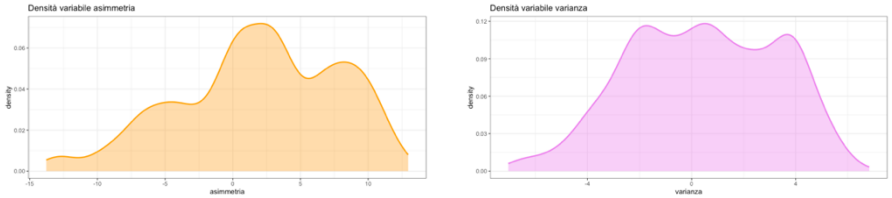
Si noti come il modello scelto sia il migliore in termini di CV ma non secondo il BIC. Il BIC è infatti leggermente inferiore per il modello Pk_L_Ck (EVV), più vincolato di quello considerato. Per entrambi i criteri le differenze sono però minime: scegliere un modello piuttosto che l'altro non dovrebbe essere eccessivamente discriminante.
Si consideri quindi il caso meno vincolato (VVV) e si proceda alla seconda fase della classificazione, assegnando le osservazioni del test set ai due gruppi.

```
Confusion Matrix and Statistics

          Reference
Prediction vere false
     vere   135     0
     false    3   112

               Accuracy : 0.988
                 95% CI : (0.9653, 0.9975)
    No Information Rate : 0.552
    P-Value [Acc > NIR] : <2e-16

                  Sensitivity : 0.9783
                  Specificity : 1.0000
               Pos Pred Value : 1.0000
               Neg Pred Value : 0.9739
                   Prevalence : 0.5520
               Detection Rate : 0.5400
         Detection Prevalence : 0.5400
            Balanced Accuracy : 0.9891
```

Confrontando le etichette emerge che, delle 138 banconote vere, 3 sono state classificate come false: l'accuracy è infatti del 98.8% (il classificatore sbaglia del 1.2%). Il dato più importante riguarda però le banconote false, infatti tutte le 112 banconote contraffatte sono riconosciute come tali (specificity = 1). In ottica di previsione questo errore provocherebbe danni più ingenti di quanto possano fare delle banconote vere scambiate per contraffatte.
Nel grafico sottostante viene riportata la classificazione delle unità del test set mettendo in evidenza i due gruppi e le 3 unità erroneamente classificate.



La classificazione attraverso EDDA potrebbe però non essere la scelta più opportuna. MDA, infatti, è basata su una mistura di misture di normali e, osservando la densità delle due variabili più rilevanti, sembra essere la scelta più appropriata.



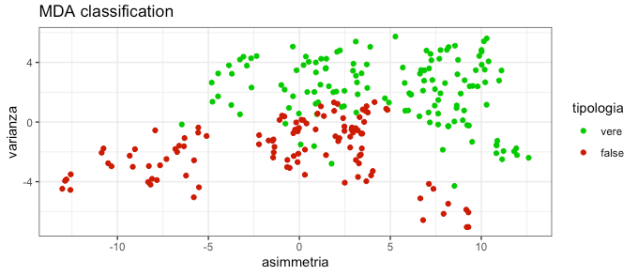
Utilizzando tutte le variabili e non imponendo vincoli sul numero di mixture components, i modelli migliori risultano essere VEV e VVV con 5 componenti ciascuno. La scelta del modello è stata effettuata guardando al BIC migliore (più basso).
Il numero di parametri da stimare è abbastanza elevato ma non desta preoccupazioni, essendo le osservazioni in misura molto maggiore. Inoltre, si nota che il training del classificatore avviene senza errori.
Classificando le unità del test set e confrontando le etichette assegnate con quelle reali, si noti come la classificazione sia riuscita perfettamente.

```
MclustDA model summary:

 log-likelihood    n  df       BIC
      -8671.641 1121 136 -18298.27

Classes   n     % Model G
  vere  623 55.58   VEV 5
  false 498 44.42   VVV 5

Training confusion matrix:
       Predicted
Class   vere false
  vere   623     0
  false    0   498
Classification error = 0
Brier score          = 0
```

```
Confusion Matrix and Statistics

          Reference
Prediction vere false
     vere   138     0
     false    0   112

               Accuracy : 1
                 95% CI : (0.9854, 1)
    No Information Rate : 0.552
    P-Value [Acc > NIR] : < 2.2e-16

                  Sensitivity : 1.000
                  Specificity : 1.000
               Pos Pred Value : 1.000
               Neg Pred Value : 1.000
                   Prevalence : 0.552
               Detection Rate : 0.552
         Detection Prevalence : 0.552
            Balanced Accuracy : 1.000
```

Un'ulteriore conferma viene fornita applicando il metodo v-fold cross validation. L'intero set di osservazioni viene diviso in v=10 sottoinsiemi, 9 vengono usati come training set e uno come test. L'operazione viene ripetuta 10 volte: ciò permette di ridurre la distorsione dovuta alla scelta di un campione casuale piuttosto che un altro.
Essendo il dataset composto da 1371 osservazioni, la divisione per 10 lascia fuori un'unità. Questo non desta preoccupazioni per due motivi: si tratta di una sola unità; se si volesse considerare un numero di fold per cui il dataset è divisibile, si dovrebbe impostare v=4, che semplificherebbe eccessivamente la procedura.
Anche in questo caso l'errore di classificazione è nullo: questo metodo è quindi da prediligere rispetto ad EDDA.

Il grafico sottostante rappresenta la classificazione delle unità del test set.



Provando ad aumentare il numero delle componenti (ad esempio impostando G=1:10) il modello si complica inutilmente: aumenta il numero di parametri da stimare e non c'è alcun'utilità: già con G=5 l'errore è nullo.
Per avere un numero di parametri da stimare meno oneroso si può adottare una riduzione della dimensionalità.
Considerando infatti solo le variabili varianza e asimmetria e rifacendo i calcoli su esse vengono scelti modelli molto più semplici (VII e VVE). La riduzione dei parametri naturalmente è rilevante, se ne stimano infatti solo 44.
Queste semplificazioni vengono pagate in termini di errore nella classificazione. Infatti, considerando il test set, compaiono delle unità misclassificate. La prima classificazione MDA eseguita risulta quindi essere la migliore.

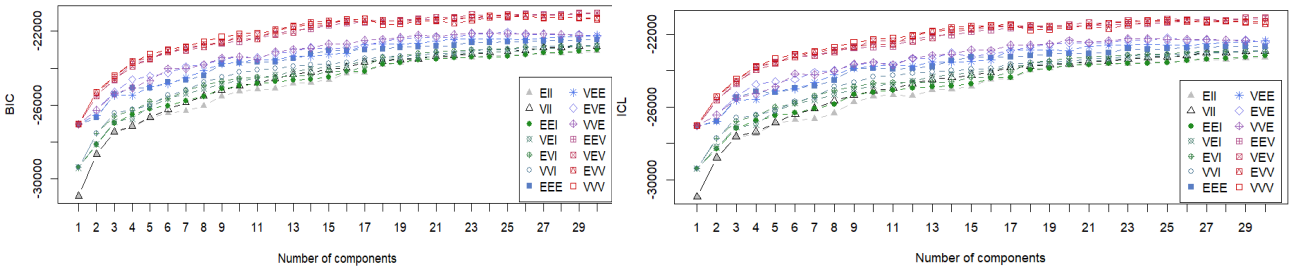
MODEL BASED CLUSTERING

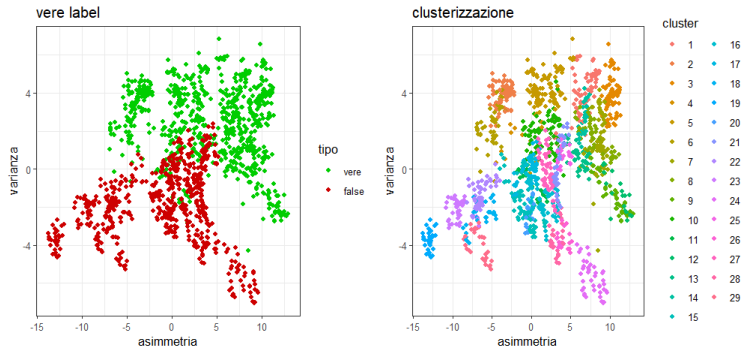
Si ipotizzi ora che non siano disponibili le label dei diversi di gruppi: si usi quindi la tecnica del model clustering con modelli mistura di normali. Questo è molto interessante in quanto fornisce un risultato di aggregazione dei dati più naturale e basato sui valori assunti dalle variabili piuttosto che da un modello allenato.
Inizialmente si provi a creare un cluster con tutte le variabili disponibili (asimmetria, varianza, curtosi ed entropia): si ottengono 29 cluster (VEV) con volume e orientamento variabili e forma uguale, che implicano la stima di 350 parametri.
La scelta del miglior modello (VEV) è stata fatta guardando sia al BIC che all'ICL (BIC corretto con l'entropia).
Si è scelto di modificare il numero massimo di cluster in quanto lasciando il default (ovvero 9) il risultato ottimale era in funzione di 9 cluster: non era chiaro se fosse veramente il numero di cluster ottimali o la scelta migliore nel range 1: 9.
Il modello risulta molto complesso, ma il miglioramento dell'ICL, che passa da -22464.4 (9 cluster - VVV) a -21093.1, giustifica in parte lo sforzo computazionale dovuto ai parametri da stimare. Ciò va infatti a favore di un migliore fitting dei dati.

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEV (ellipsoidal, equal shape) model with 29 components:

 log-likelihood    n  df       BIC      ICL
      -9242.002 1372 350 -21012.41 -21093.1

Clustering table:
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22
 69  69  77  29 107  45  68  53  42  81  23  28  40  31  72  71  38  32  30  20  50  56
 23  24  25  26  27  28  29
 38  40  30  20  56  27  30
```

```
Top 3 models based on the ICL criterion:
   VEV,29    VEV,30    EEV,30
-21093.10 -21093.65 -21131.03
```

Si grafichi l'andamento di BIC e ICL per il numero crescente di G e per i 14 modelli disponibili: i valori massimi si ottengono in corrispondenza di 29 cluster e un modello VEV, anche se si nota una certa "stazionarietà" con l'aumentare del numero di gruppi. Inoltre, con ulteriori prove, si è riscontrato che è possibile individuare modelli con più cluster e ICL migliori, nonostante questo complichi ancora di più l'analisi.





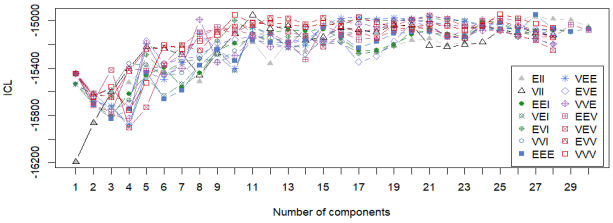
Anche il plot dei cluster individuati risulta poco leggibile per via dell'alto numero di cluster. Sono state scelte come variabili per graficare le due variabili selezionate dall'analisi delle componenti principali.

Si ripete l'analisi riducendo la dimensionalità delle variabili (si selezionino solo asimmetria e varianza). Con il criterio ICL si evidenziano i seguenti modelli:

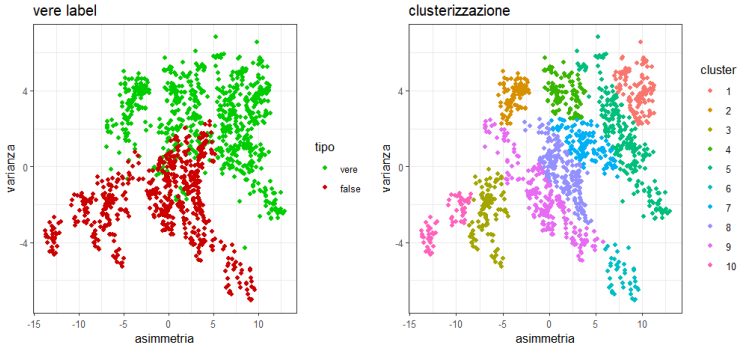
```
Best ICL values:
              VVV,25        EEE,27        VVV,10
ICL      -14945.04 -14948.45311 -14950.663877
ICL diff      0.00     -3.41301     -5.623778
```

Senza perdere di generalità del clustering si opta per il modello VVV con 10 cluster per via dello scarto quasi esiguo in termini relativi di ICL (confrontando il BIC risulta meglio il modello con 10 cluster). Vengono stimati 59 parametri: rispetto alla clusterizzazione con tutte le variabili è presente una semplificazione. È interessante vedere nel grafico dell'ICL come il modello migliore cambi molto frequentemente rispetto alla situazione precedente in cui vi era una ripetizione dei modelli migliori.

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 10 components:

 log-likelihood    n df       BIC       ICL
      -7105.912 1372 59 -14638.04 -14950.66

Clustering table:
  1   2   3   4   5   6   7   8   9  10
123  89 116 109 252  40 140 211 227  65
```



Il grafico sottostante confronta i cluster veri con quelli stimati dal modello.

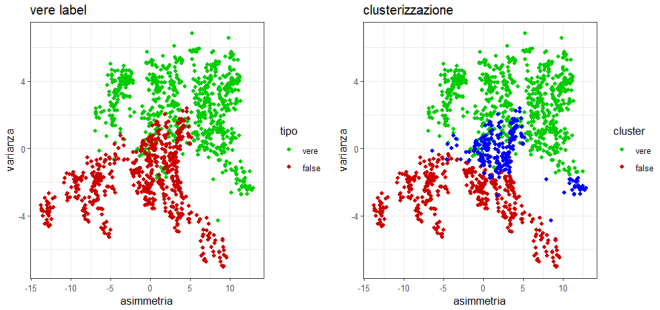


In entrambi i casi i modelli scelti non selezionano i due cluster definiti all'interno del dataset. Questo perché nella costruzione di esso sono state scannerizzate diverse tipologie di banconote e di diverso taglio, come riportato dell'autore del dataset stesso, senza però registrarla come variabile. Si può perciò ipotizzare che il tentativo di clustering sia effettuato discriminando non solo per le banconote vere o false, ma anche per ciascuna valuta. Questo spiega il motivo per cui, utilizzando tutte le variabili, si sono trovati 29 (con ulteriori analisi anche 35) cluster. Sarebbe perciò interessante avere a disposizione anche questa variabile latente per riuscire a valutare la precisione di raggruppamento.
Si provi ora a forzare un numero di cluster uguale a 2 per fare un confronto con le label esistenti, tenendo solamente le variabili selezionate dalla pca. Il modello stimato è stato selezionato sempre tramite ICL e risulta essere un modello EVV:

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust EVV (ellipsoidal, equal volume) model with 2 components:

 log-likelihood    n df       BIC       ICL
      -7621.642 1372 10 -15315.52 -15613.58

Clustering table:
  1   2
919 453
```

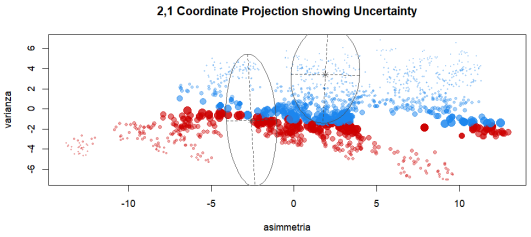
```
Confusion Matrix and Statistics

          Reference
Prediction vere false
     vere   717   202
     false   45   408
```

A destra è presente la confusion matrix, che confronta le etichette predette dal modello con quelle reali. Si evidenzia la grossa lacuna di questo modello, ovvero un valore di Specificity = 0.6689, che attesta la scarsissima capacità di identificare le banconote false. Difatti è molto più grave identificare una banconota falsa come vera, piuttosto che il contrario. Per questo motivo la specificità risulta essere una misura molto più adeguata rispetto al CER (che comunque risulta essere del 18%, quindi abbastanza alto).



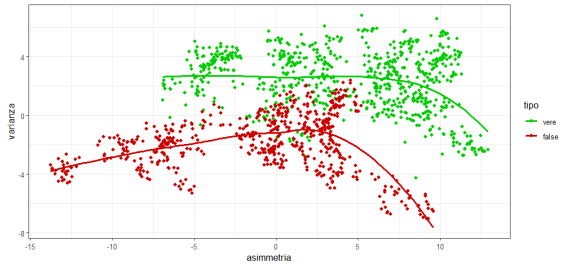
In questo confronto sono evidenziate in blu tutte le unità classificate erroneamente dal modello.
Si nota come il modello non riconosce soprattutto le osservazioni che si trovano in nell'area di piano in cui le banconote false si sovrappongono a quelle vere (probabilmente a causa delle considerazioni precedentemente fatte sulle valute diverse).

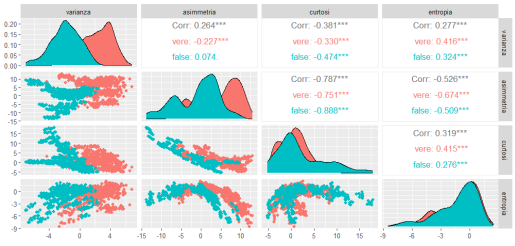
Nel grafico dell'incertezza vengono rappresentate le probabilità a posteriori con un puntino tanto più grande quanto maggiore è l'incertezza dell'assegnazione a quel determinato cluster. Anche qui si evidenzia come l'incertezza maggiore sia sui punti in cui i due cluster si intersecano e nella parte in cui l'asimmetria assume valori massimi: in entrambi i casi ci sono banconote vere clusterizzate come false.



MODEL BASED CLUSTERS WITH COVARIATES

Poiché il model based clustering ha prodotto risultati che sono migliorabili, si prova a valutare una clusterizzazione basata sulla regressione. Un modello errato sarebbe quello che considera le etichette come risposta e le altre variabili come esplicative le altre: i cluster dovrebbero infatti essere individuati tramite l'uso delle covariate e non attraverso le label.
Si grafichi quindi la relazione tra le due variabili usate per l'analisi precedente, asimmetria e varianza.





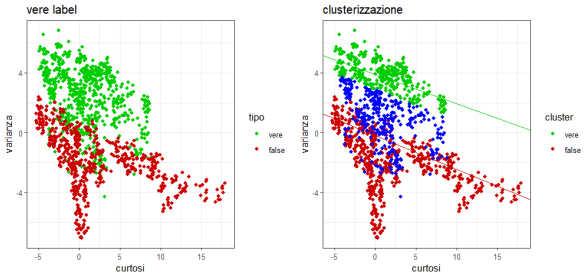
Le regressioni che fittano meglio i dati risultano essere due curve loess, quindi non parametriche.
Per questo motivo si sceglie di analizzare nuovamente il pairs delle variabili per vedere se si individuano dei modelli più facili.
Da questi grafici sembrerebbe che un modello semplice come due regressioni lineari che hanno come risposta la varianza e come covariata la curtosi (sembra avere una distribuzione normale nonostante abbia una coda di destra rilevante; non può essere una gamma per via del suo supporto), possa avere un fitting accettabile.
Siccome non è possibile graficare una relazione tra tre o più variabili non si esclude la possibilità di inserire anche altre esplicative nel modello.
Usando perciò come esplicativa solamente la curtosi e come risposta la varianza, ipotizzando due cluster basati su due regressioni lineari con distribuzione della esplicativa normale e usando una full mixture of expert model, si ottengono i seguenti risultati:

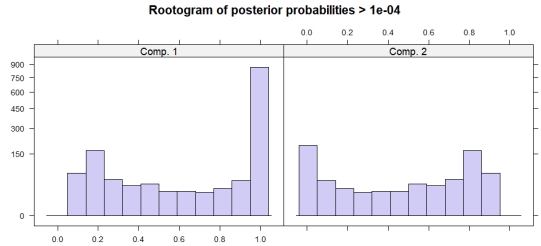
```
Call:
stepFlexmix(varianza ~ curtosi, data = banconote, concomitant = FLXMCmvnorm(),
    model = FLXMRglm(family = "gaussian"), k = 2)

      prior size post>0 ratio
Comp.1  0.77 1010   1372 0.736
Comp.2  0.23  362    698 0.519

'log Lik.' -3199.743 (df=7)
AIC: 6413.486   BIC: 6450.055
```

```
$Comp.1
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -0.192759   0.127145  -1.5161   0.1295
curtosi     -0.226132   0.016786 -13.4715   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Comp.2
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  3.939817   0.090080  43.737 < 2.2e-16 ***
curtosi     -0.196917   0.019318 -10.194 < 2.2e-16 ***
---
```

```
         vere
predette vere false
   vere   362     0
   false  400   610
```

È evidente che il modello sovrastimi il numero di banconote false nel dataset; infatti, si ha un CER di circa il 30%. È però rilevante che il valore della specificità sia 1: questo modello è ottimo per evitare che una banconota falsa rimanga in circolazione, al contrario è pessimo nella clusterizzazione delle banconote vere, in quanto sono più quelle erroneamente raggruppate che quelle identificate come vere (sensibilità del circa 47,5%).
Osservando i coefficienti si evidenzia come quello riferito alla curtosi sia significativo per entrambi i gruppi, mentre l'intercetta sia significativa solo per il gruppo delle vere.
La ratio indica che l'incertezza sulle banconote allocate come vere è maggiore (0.519<0.736=ratio false). Questo potrebbe giustificare la traslazione della retta di regressione riferite alle banconote false.
Anche il rootogram evidenzia un cluster con molta incertezza (l'area al centro è non trascurabile per entrambi i cluster).





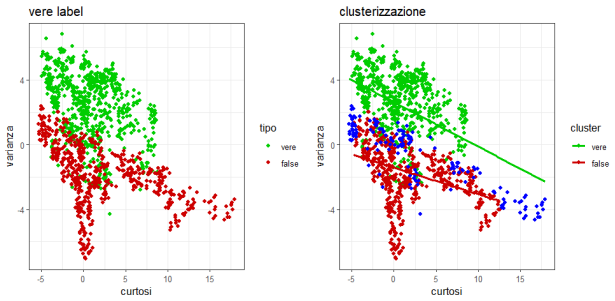
Si provi a inserire come esplicativa anche l'entropia, anch'essa con distribuzione simile alla curtosi, per vedere se sono presenti miglioramenti nel modello.

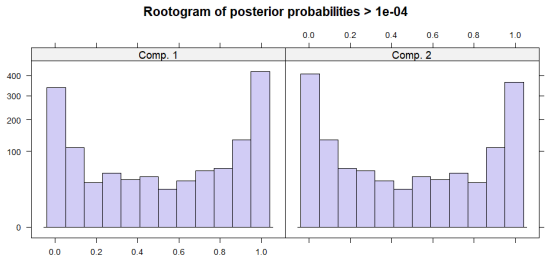
```
Call:
stepFlexmix(varianza ~ curtosi + entropia, data = banconote, concomitant = FLXMCmvnorm(varianza ~
    curtosi + entropia), model = FLXMRglm(family = "gaussian"), k = 2)

      prior size post>0 ratio
Comp.1 0.533  731   1347 0.543
Comp.2 0.467  641   1360 0.471

'log Lik.' -3046.282 (df=9)
AIC: 6110.564   BIC: 6157.58
```

```
$Comp.1
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  3.432691   0.140421  24.446 < 2.2e-16 ***
curtosi     -0.383645   0.015208 -25.227 < 2.2e-16 ***
entropia     0.647060   0.027396  23.619 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Comp.2
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -0.540198   0.163247  -3.3091  0.000936 ***
curtosi     -0.253958   0.020903 -12.1492 < 2.2e-16 ***
entropia     0.597232   0.043711  13.6631 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
          Reference
Prediction vere false
     vere   659    72
     false  103   538
```

L'errore di classificazione è diminuito (12,8%) e la clusterizzazione segue un andamento più conforme ai dati, a discapito di una diminuzione della specificità, che resta comunque sopra l'88%. Inoltre, c'è un netto miglioramento netto della sensivity, che passa dal pessimo 45% dello scorso modello all'86%. La significatività dei coefficienti conferma l'utilità di avere inserito una nuova covariata (l'entropia), in quanto tutte e sei le stime sono significative. Si grafichi il rootogram e la clusterizzazione.





La nuvola di punti classificati in maniera erronea (colore blu) è diminuito notevolmente. Il rootogram risulta migliore del modello precedente in quanto ha delle frequenze più elevate sugli estremi 0 e 1. La parte centrale non è nulla, il che evidenzia comunque un certo grado di incertezza nell'assegnare le unità (indicato anche dai ratio abbastanza bassi).
In definitiva quindi questo modello risulta più adatto alle vere labels nonostante ci sia più incertezza.

CONCLUSIONI

Al termine di quest'analisi si può affermare come il miglior metodo risulti la classificazione MDA. Il modello stimato permette infatti di identificare in maniera precisa la tipologia di banconote avendo potenzialmente solo i dati quantitativi a disposizione.
È stato comunque interessante scoprire strutture latenti del dataset attraverso la cluster analysis.
Si cita inoltre il professore Volker Lohweg, che si è reso disponibile a fornire delucidazioni sul dataset.

Fonte dei dati: https://archive.ics.uci.edu/dataset/267/banknote+authentication