

Tesina di Inferenza Statistica

Dario Macii

Indice

1	Creazione del dataset	2
1.1	Scelta delle variabili	2
2	Analisi descrittiva	6
2.1	Analisi marginale	6
2.1.1	Boxplot	7
2.1.2	Iistogrammi	8
2.2	Analisi bivariata	9
2.2.1	Boxplot condizionati	9
2.2.2	Diagrammi a dispersione	13
2.3	Analisi delle componenti principali	15
2.4	Stimatore di nucleo	17
2.5	Stimatore di nucleo bivariato	20

Capitolo 1

Creazione del dataset

Il dataset è stato costruito utilizzando le immagini contenute nel dataset LC25000 (disponibile al link: https://github.com/tampapath/lung_colon_image_set), selezionando esclusivamente quelle relative ai tumori polmonari. Sono presenti 15000 immagini, 5000 per ogni tipo di tumore. Sono presenti tre tipi di tumori: **benigno**, **adenocarcinoma** e **carcinoma a cellule squamose**. Questi ultimi due sono tipologie di tumore maligno.

1.1 Scelta delle variabili

Per la costruzione del dataset sono state estratte le intesità dei pixel relative ai tre canali colore: rosso, verde e blu (RGB).

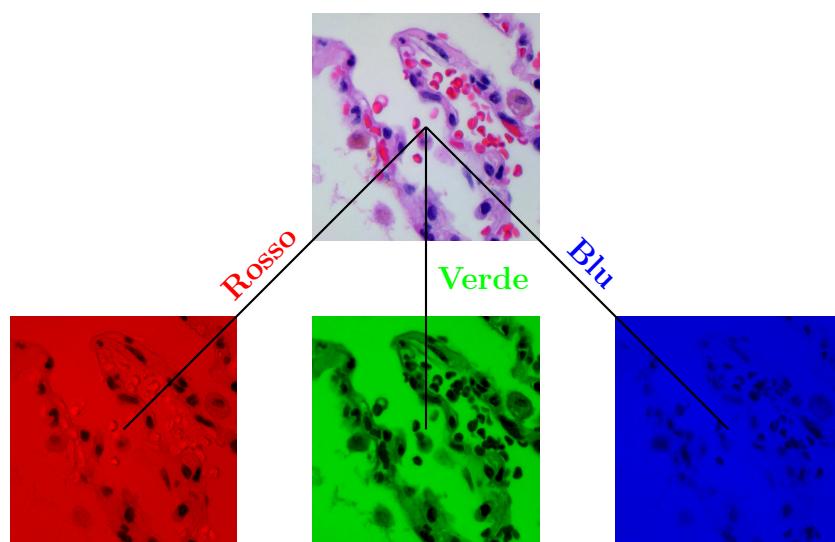


Figura 1.1.1: Scomposizione di un’immagine nei tre canali RGB

Sono state successivamente calcolate quattro statistiche del primo ordine relative ai ad ogni canale colore. Le statistiche utilizzate sono: **media**, **varianza**, **coefficiente di asimmetria** e **curtosi**.

Il codice R utilizzato per la costruzione del dataset è il seguente:

```
1 library(magick)
2 library(moments)
3 library(dplyr)
4
5 # crea una lista con i percorsi alle cartelle
6 cartelle_immagini <- list(
7   "benigno" = "lung_colon_image_set/lung_image_sets/lung_n",
8   "adenocarcinoma" = "lung_colon_image_set/lung_image_sets/lung_aca",
9   "carcinoma a cellule squamose" =
10  ↪ "lung_colon_image_set/lung_image_sets/lung_scc"
11 )
12
13 df_stat <- function(cartelle_immagini){
14   #crea una lista vuota
15   lista_immagini <- list()
16
17   #cicli for annidati per selezionare tutte le immagini
18   for (etichetta in names(cartelle_immagini)){
19     file_immagini <- list.files(cartelle_immagini[[etichetta]], full.names
20      ↪ = TRUE)
21
22     for(file_immagine in file_immagini){
23       #lettura dell'immagine
24       img <- image_read(file_immagine)
25       img_rgb <- image_data(img)
26
27       # Estraie singolarmente i canali
28       canale_rosso <- as.numeric(img_rgb[1, , ])
29       canale_verde <- as.numeric(img_rgb[2, , ])
30       canale_blu <- as.numeric(img_rgb[3, , ])
31
32       #calcola le statistiche
33       media_rosso <- mean(canale_rosso)
34       media_verde <- mean(canale_verde)
35       media_blu <- mean(canale_blu)
36
37       varianza_rosso <- var(canale_rosso)
38       varianza_verde <- var(canale_verde)
39       varianza_blu <- var(canale_blu)
40
41       skewness_rosso <- skewness(canale_rosso)
```

```

39     skewness_verde <- skewness(canale_verde)
40     skewness_blu <- skewness(canale_blu)
41
42     curtosi_rosso <- kurtosis(canale_rosso)
43     curtosi_verde <- kurtosis(canale_verde)
44     curtosi_blu <- kurtosis(canale_blu)
45
46     #inserisce le variabili estratte nella lista
47     stat_immagine <- c(media_rosso, media_verde, media_blu,
48     ↪ varianza_rosso, varianza_verde, varianza_blu, skewness_rosso,
49     ↪ skewness_verde, skewness_blu, curtosi_rosso, curtosi_verde,
50     ↪ curtosi_blu)
51     lista_immagini <- append(lista_immagini, list(c(stat_immagine,
52     ↪ etichetta)))
53   }
54 }
55
56 # Converte la lista in dataframe
57 df <- as.data.frame(do.call(rbind, lista_immagini))
58
59 # Aggiunge nomi alle colonne
60 colnames(df) <- c("media_rosso", "media_verde", "media_blu",
61   ↪ "varianza_rosso", "varianza_verde", "varianza_blu", "skewness_rosso",
62   ↪ "skewness_verde", "skewness_blu", "curtosi_rosso", "curtosi_verde",
63   ↪ "curtosi_blu", "classe")
64
65 # converte le varibili in valori numerici
66 df <- df %>%
67   mutate(across(1:12, as.numeric))
68   return(df)
69 }
70
71
72 #richiamo della funzione
73 df <- df_stat(cartelle_immagini)

```

Il dataset risultante è costituito da 13 variabili: 12 variabili quantitative composte da media, varianza, coefficiente di asimmetria e curtosi per ciascun canale RGB, e una variabile qualitativa che indica la tipologia di tumore. Le prime righe del dataframe ottenuto sono le seguenti.

	media_rosso	media_verde	media_blu	varianza_rosso	varianza_verde	varianza_blu	skewness_rosso	skewness_verde	skewness_blu	curtosi_rosso	curtosi_verde	curtosi_blu	classe
1	211.68	164.53	198.14	909.82	2441.28	244.46	-3.03	-1.25	-2.43	13.55	3.68	10.94	benigno
2	215.20	161.58	201.68	997.52	2524.44	259.12	-2.71	-1.06	-2.65	11.35	3.17	13.44	benigno
3	202.59	165.32	201.50	1196.10	2916.13	447.44	-2.83	-1.19	-2.34	11.82	3.52	9.46	benigno
4	218.42	165.68	169.14	690.39	2017.02	153.36	-3.18	-1.17	-2.91	15.55	3.69	16.34	benigno
5	201.21	164.37	189.75	966.28	2364.02	297.15	-2.91	-1.38	-2.63	12.89	4.23	11.73	benigno
6	201.21	164.37	189.75	966.28	2364.02	297.15	-2.91	-1.38	-2.63	12.89	4.23	11.73	benigno

Tabella 1.1.1: Prime righe del dataframe

Capitolo 2

Analisi descrittiva

In questo capitolo verrà svolta l'analisi descrittiva del dataset, prima marginale e poi bivariata. Successivamente verrà svolta l'analisi delle componenti principali e infine verrà svolta la stima di nucleo.

2.1 Analisi marginale

Come punto di partenza dell'analisi marginale sono stati calcolati alcuni indici di sintesi, tra cui i quantili, la media e la deviazione standard per ciascuna variabile quantitativa. Non sono presenti valori mancanti.

Tipo Variabile	Nome Variabile	Valori Mancanti	Media	Deviazione Standard	Minimo	p25	Mediana	p75	Massimo
character	classe	0							
numeric	media_rosso	0	170.78	26.72	107.70	148.31	168.61	194.39	231.84
numeric	media_verde	0	136.20	22.88	71.86	119.56	137.76	154.72	197.25
numeric	media_blu	0	217.06	19.79	157.78	198.83	222.08	232.34	252.08
numeric	varianza_rosso	0	1117.17	410.22	126.19	842.18	1039.08	1326.31	3070.85
numeric	varianza_verde	0	2073.44	849.92	283.20	1413.18	2126.66	2681.80	4832.58
numeric	varianza_blu	0	381.73	191.72	11.99	246.81	345.21	488.54	1562.03
numeric	skewness_rosso	0	-1.67	0.95	-5.14	-2.60	-1.33	-0.90	0.28
numeric	skewness_verde	0	-0.63	0.46	-3.39	-0.95	-0.62	-0.29	0.83
numeric	skewness_blu	0	-1.83	0.70	-6.70	-2.31	-1.82	-1.30	-0.13
numeric	curtosi_rosso	0	6.91	4.73	1.72	3.19	4.55	10.76	43.54
numeric	curtosi_verde	0	3.16	0.89	1.52	2.57	3.01	3.57	14.26
numeric	curtosi_blu	0	8.00	4.79	2.15	5.03	7.17	9.74	106.12

Tabella 2.1.1: Indici di sintesi

La sola variabile qualitativa presente è la tipologia di tumore, suddivisa equamente in tre categorie da 5.000 osservazioni ciascuna:

adenocarcinoma	5000
benigno	5000
carcinoma a cellule squamose	5000

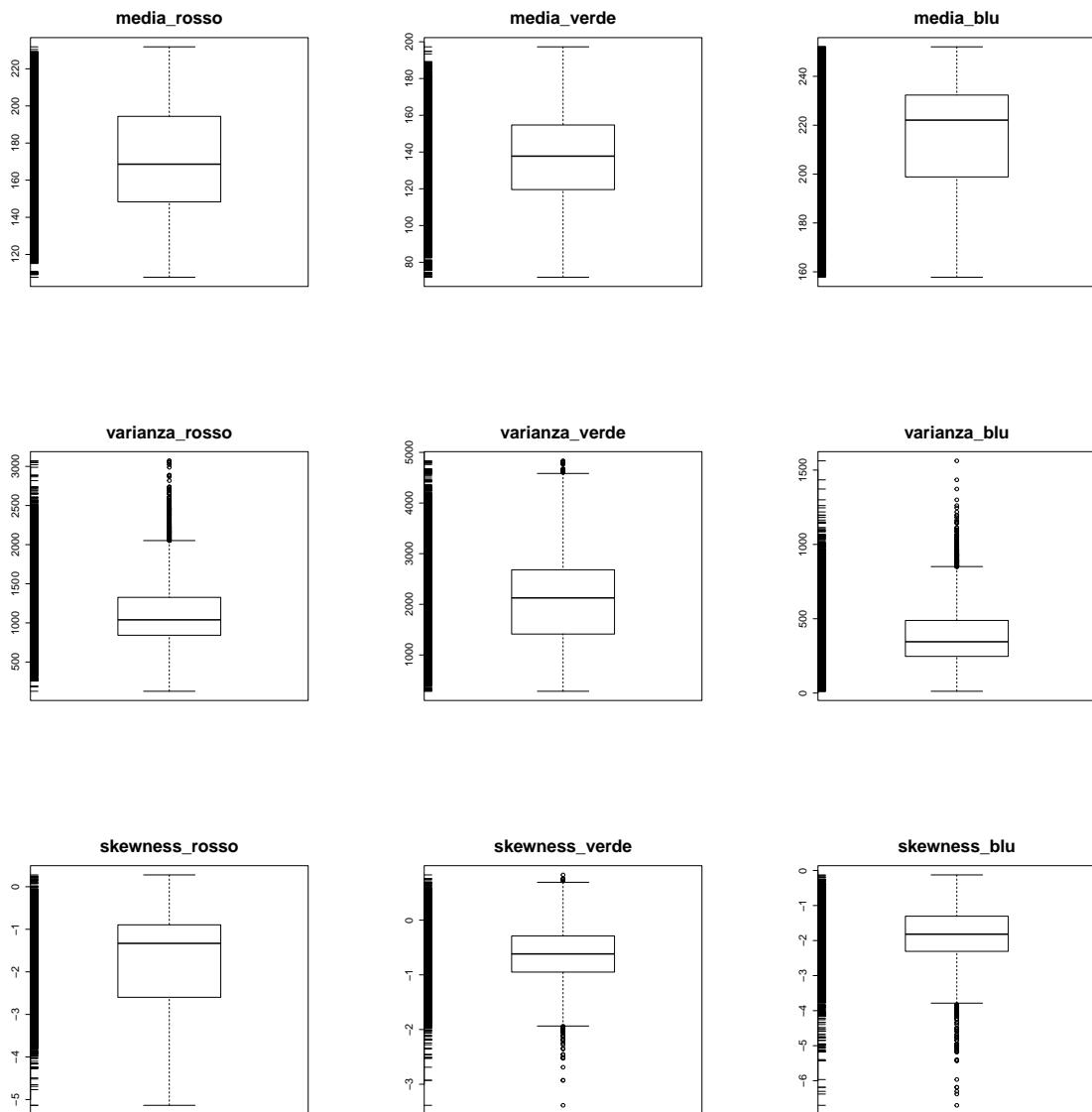
Tabella 2.1.2: Modalità della variabile classe

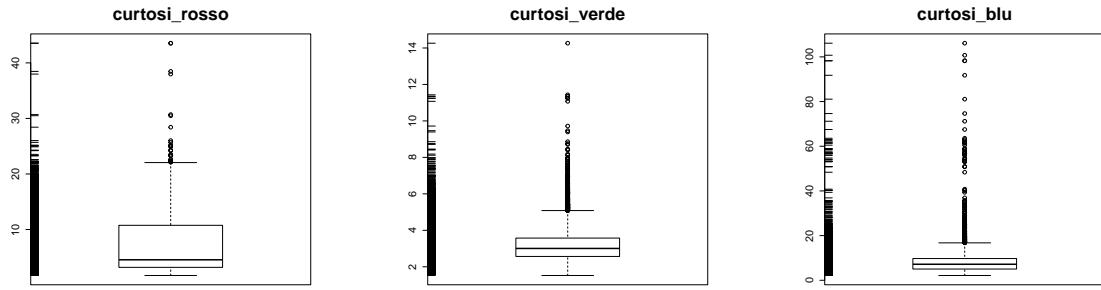
2.1.1 Boxplot

I boxplot la presenza di eventuali asimmetrie nella distribuzione e di identificare valori anomali.

Il codice utilizzato per costruire i boxplot è il seguente:

```
1 for (i in 1:12) {  
2   boxplot(df[, i], main = colnames(df)[i], col = "white")  
3   rug(df[, i], side = 2, lwd = 0.1)  
4 }
```





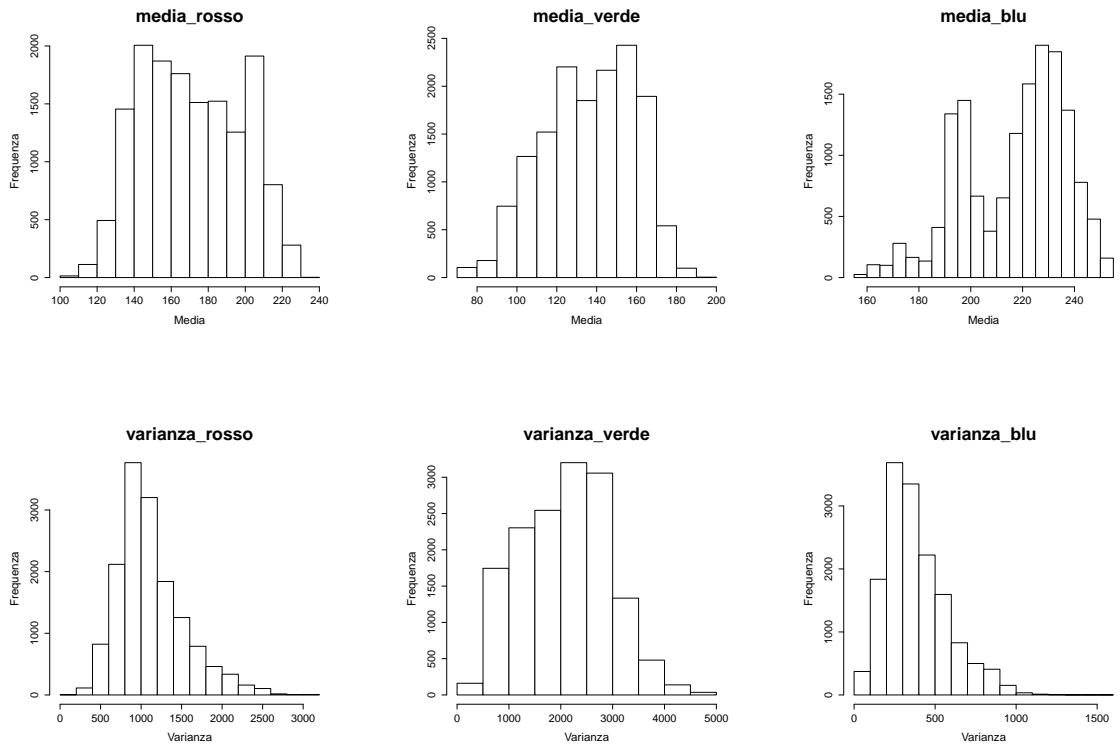
Si osserva la presenza di variabili con un numero elevato di valori anomali, come varianza_rosso, varianza_blu, skewness_verde, skewness_blu, curtosi_verde e curtosi_blu. Siamo in presenza di distribuzioni piuttosto asimmetriche, ad eccezione della variabile media_verde.

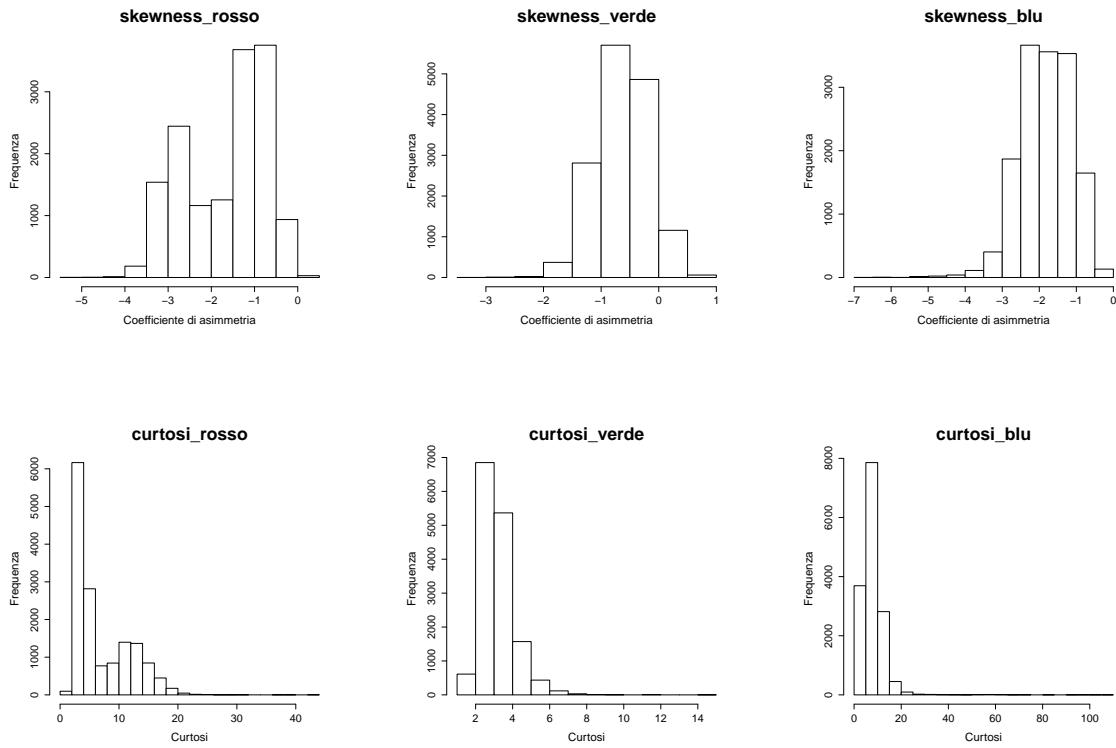
2.1.2 Iistogrammi

Il codice utilizzato per costruire gli istogrammi è il seguente:

```

1 for (i in 1:12) {
2   hist(df[, i], main = colnames(df)[i], col = "white")
3 }
```





Come evidenziato dai boxplot, le distribuzioni delle statistiche derivate dalle intensità dei pixel mostrano una marcata asimmetria. In particolare, le misure di varianza e curtosi relative ai tre canali colore presentano una asimmetria positiva. Al contrario, il coefficiente di asimmetria delle intensità dei pixel calcolato per ciascun canale mostra una asimmetria negativa.

Si osserva inoltre una distribuzione trimodale nel caso della variabile media_blu e una bimodale nel caso delle variabili skewness_rosso e curtosi_rosso.

2.2 Analisi bivariata

Per svolgere l'analisi bivariata verranno utilizzati i boxplot condizionati al tipo di tumore e i diagrammi a dispersione tra ogni coppia di variabili quantitative con i relativi coefficienti di correlazione.

2.2.1 Boxplot condizionati

A titolo di esempio è stato riportato il codice utilizzato per costruire i boxplot condizionati relativi alle variabili media_rosso, media_verde e media_blu.

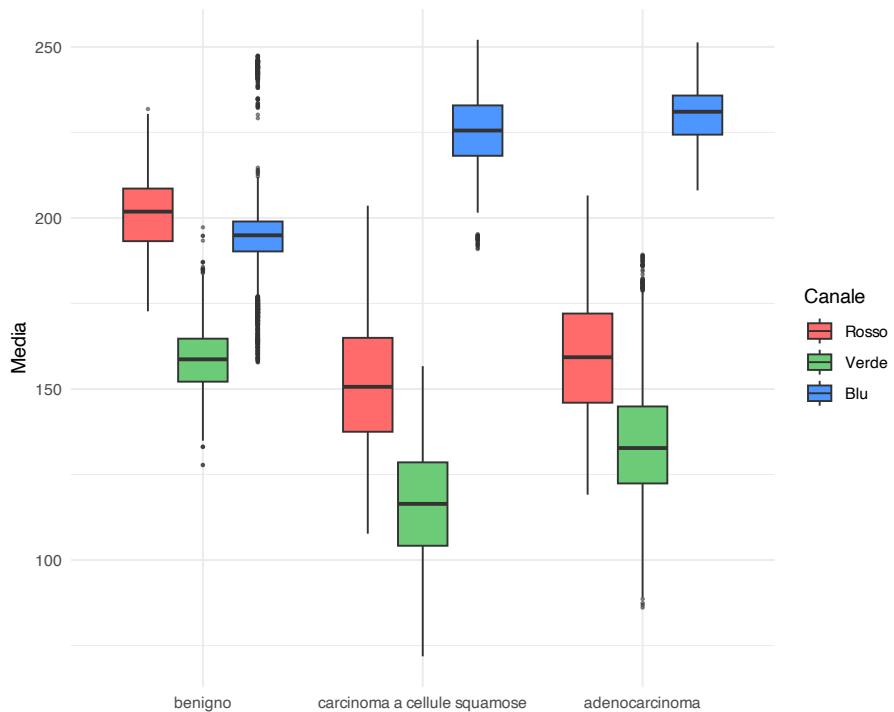
```

1 library(dplyr)
2 library(ggplot2)
3 library(reshape2)
```

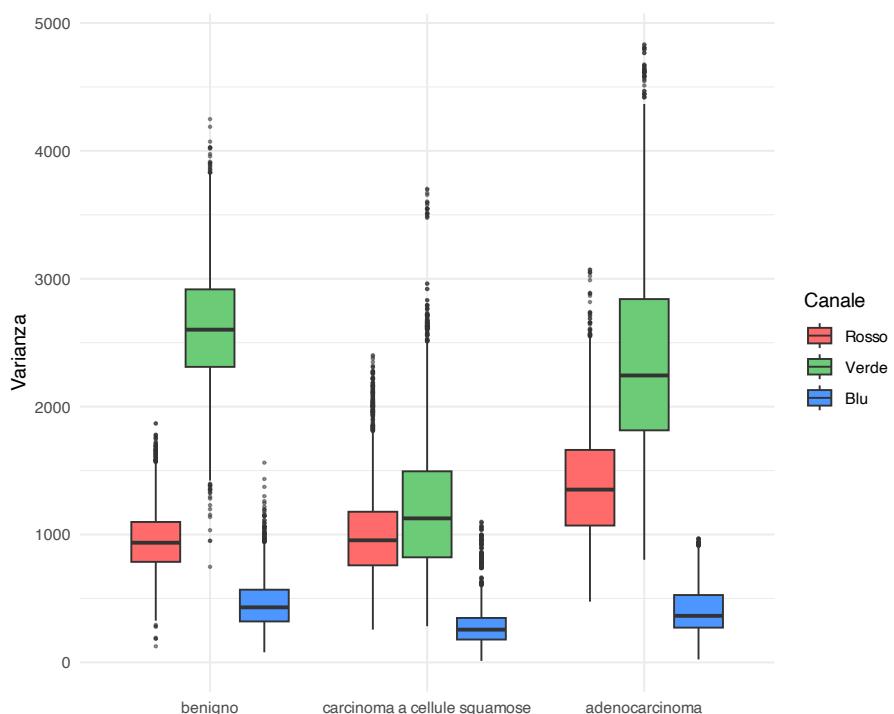
```

4 df_long <- df %>%
5   pivot_longer(cols = c(media_rosso, media_verde, media_blu), names_to =
6     "Canale", values_to = "Media")
7
8 df_long <- melt(df, id.vars = "classe", measure.vars = c("media_rosso",
9   "media_verde", "media_blu")) %>%
10  mutate(classe = factor(classe, levels = c("benigno", "carcinoma a cellule
11    squamose", "adenocarcinoma")))
12
13 ggplot(df_long, aes(x = factor(classe), y = value, fill = variable)) +
14   geom_boxplot(outlier.size = 0.5, outlier.alpha = 0.5) +
15   labs(y = "Media") +
16   scale_fill_manual(
17     name = "Canale",
18     values = c("media_rosso" = "#FF6B6B", "media_verde" = "#6BCB77",
19       "media_blu" = "#4D96FF"),
20     labels = c("media_rosso" = "Rosso", "media_verde" = "Verde",
21       "media_blu" = "Blu"))
22 ) + theme_minimal()

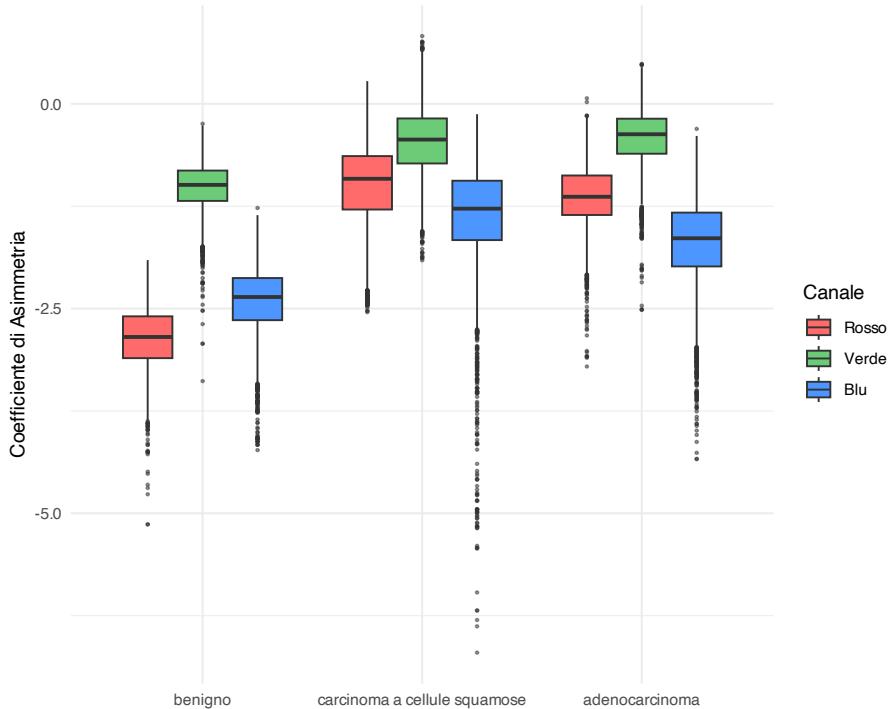
```



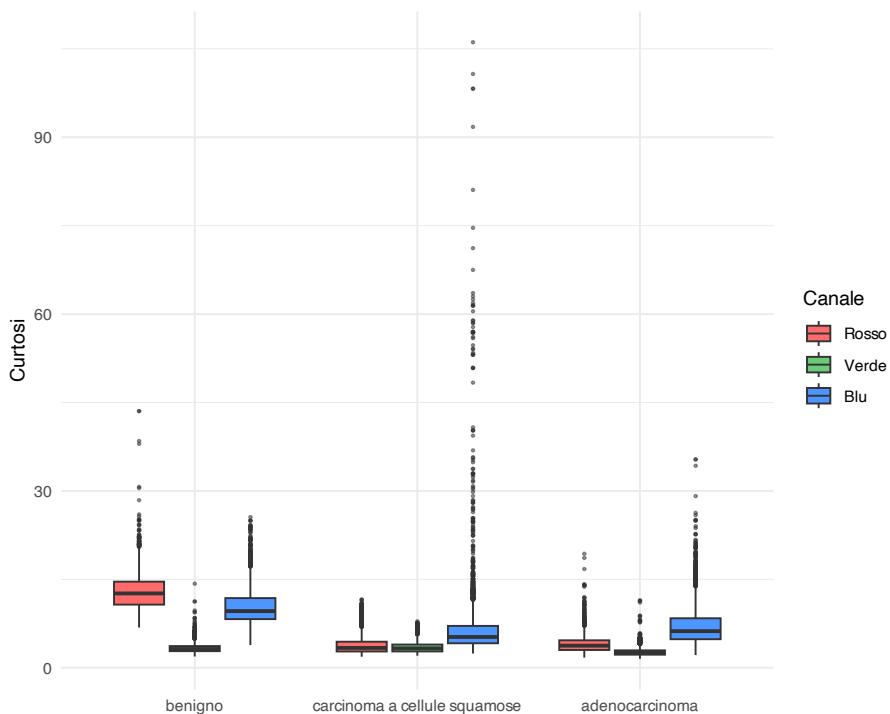
Boxplot della media dell'intensità dei pixel dei tre canali RGB condizionati al tipo di tumore



Boxplot della varianza dell'intensità dei pixel dei tre canali RGB condizionati al tipo di tumore



Boxplot del coefficiente di asimmetria dell'intensità dei pixel dei tre canali RGB condizionati al tipo di tumore



Boxplot della curtosi dell'intensità dei pixel dei tre canali RGB condizionati al tipo di tumore

Osservando i grafici si nota che le immagini relative al carcinoma a cellule squamose e all'adenocarcinoma seguono distribuzioni molto simili. Le variabili che si discostano maggiormente sono varianza_verde e varianza_rosso.

2.2.2 Diagrammi a dispersione

Per osservare la relazione tra le coppie di variabili sono stati utilizzati dei grafici a dispersione. Sono inoltre stati colorati i punti del grafico in base al tipo di tumore: giallo per il tumore benigno, verde per l'adenocarcinoma e viola per il carcinoma a cellule squamose.

Il codice utilizzato è il seguente:

```
1 colori <- c("benigno" = "#E7B800",
2           "adenocarcinoma"= "#1b9E77",
3           "carcinoma a cellule squamose" = "#7570B3")
4 classe_colori <- colori[df$classe]
5 pairs(df[,1:12], col = classe_colori, pch = 16, cex = 0.4, upper.panel =
   ↪ NULL)
```

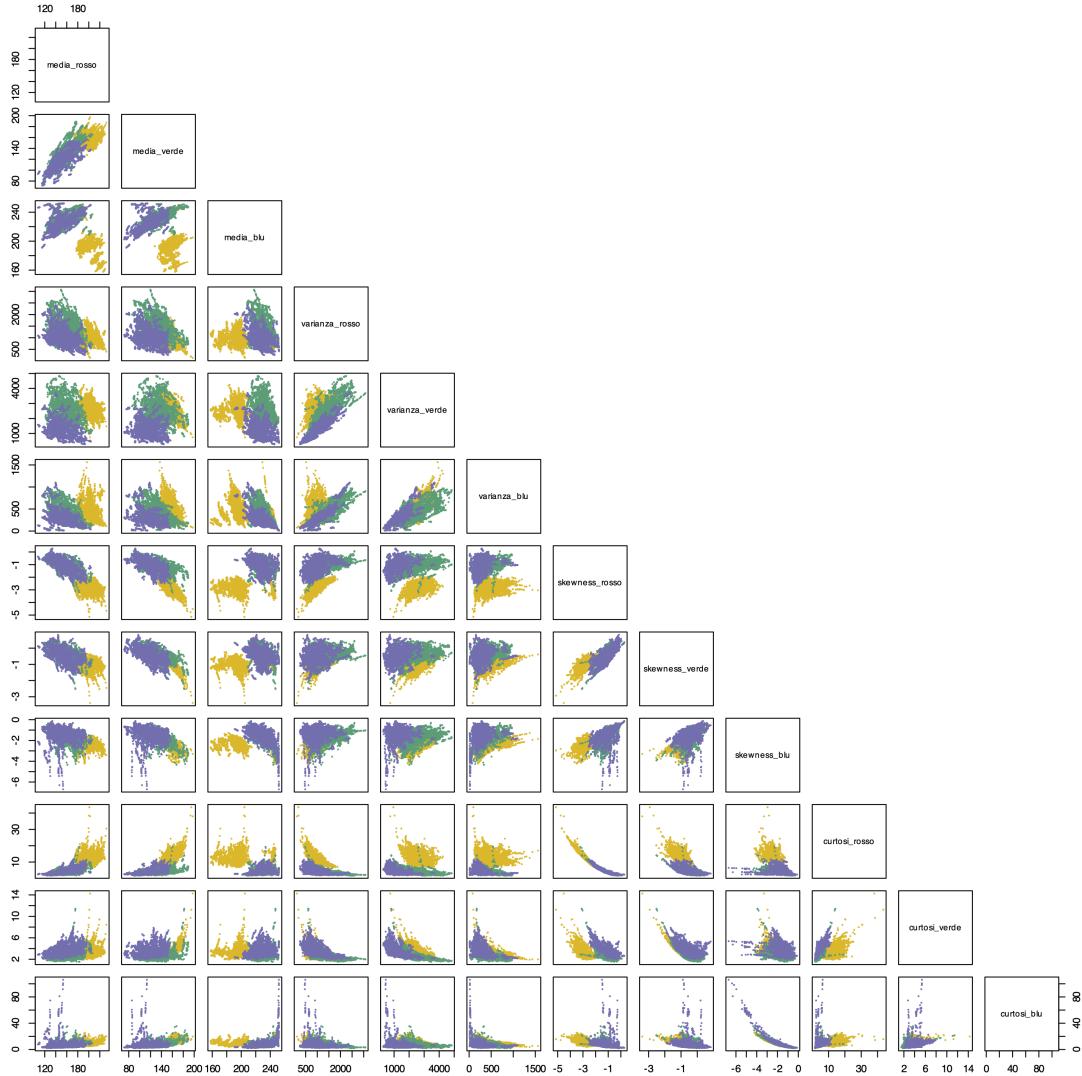


Figura 2.2.3: Matrice dei diagrammi di dispersione (giallo: benigno, verde: adenocarcinoma, viola: carcinoma a cellule squamose)

È stata inoltre utilizzata la matrice di correlazione per visualizzare il coefficiente di correlazione tra le variabili, implementato con il seguente codice:

```

1 library(corrplot)
2 corrplot(cor(df[,1:12]), method="color", type="lower", addCoef.col =
  ↴ "black", tl.col="black", tl.srt=45)

```

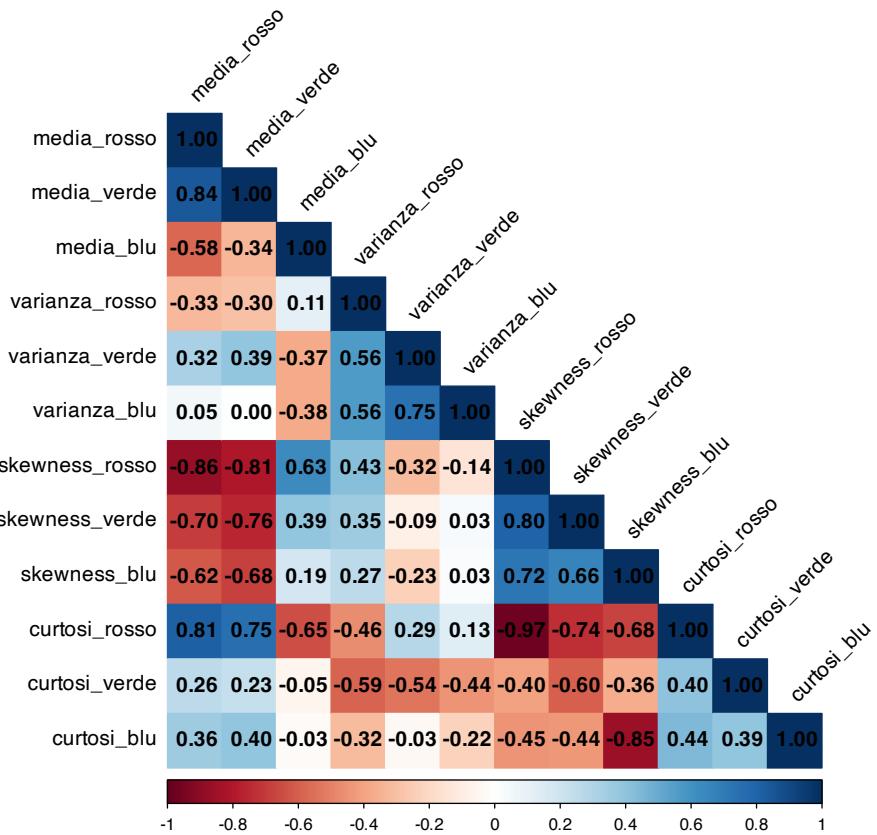


Figura 2.2.4: Matrice di correlazione

Il coefficiente di correlazione più elevato si riscontra tra le variabili `media_rosso` e `media_verde`, pari a 0.84, mentre le variabili maggiormente correlate negativamente sono `skewness_rosso` e `curtosi_rosso`. È presente inoltre un coefficiente di correlazione pari a 0 tra le variabili `media_verde` e `varianza_blu`.

È interessante notare che in alcuni casi si verifica il paradosso di Simpson: ad esempio il coefficiente di correlazione tra le variabili `media_verde` e `media_blu` è negativo e pari a -0.34, ma se si osserva il diagramma a dispersione suddiviso per classi di tumore si nota che c'è una correlazione positiva se si considerano le unità suddivise in gruppi.

2.3 Analisi delle componenti principali

Il codice utilizzato per implementare l'analisi delle componenti principali è il seguente:

```

1 library(FactoMineR)
2 library(factoextra)
3

```

```

4 df_numeric <- df[, sapply(df, is.numeric)]
5 pca_result <- PCA(df[,1:12], scale.unit = TRUE, graph = FALSE)
6
7 fviz_pca_biplot(pca_result,
8     col.ind = df$classe,
9     col.var = "black",
10    geom.ind = "point",
11    palette = c("benigno" = "#E7B800", "adenocarcinoma" =
12      "#1b9E77", "carcinoma a cellule squamose" = "#7570B3"),
13    pointsize = 1,
14    repel = TRUE)

```

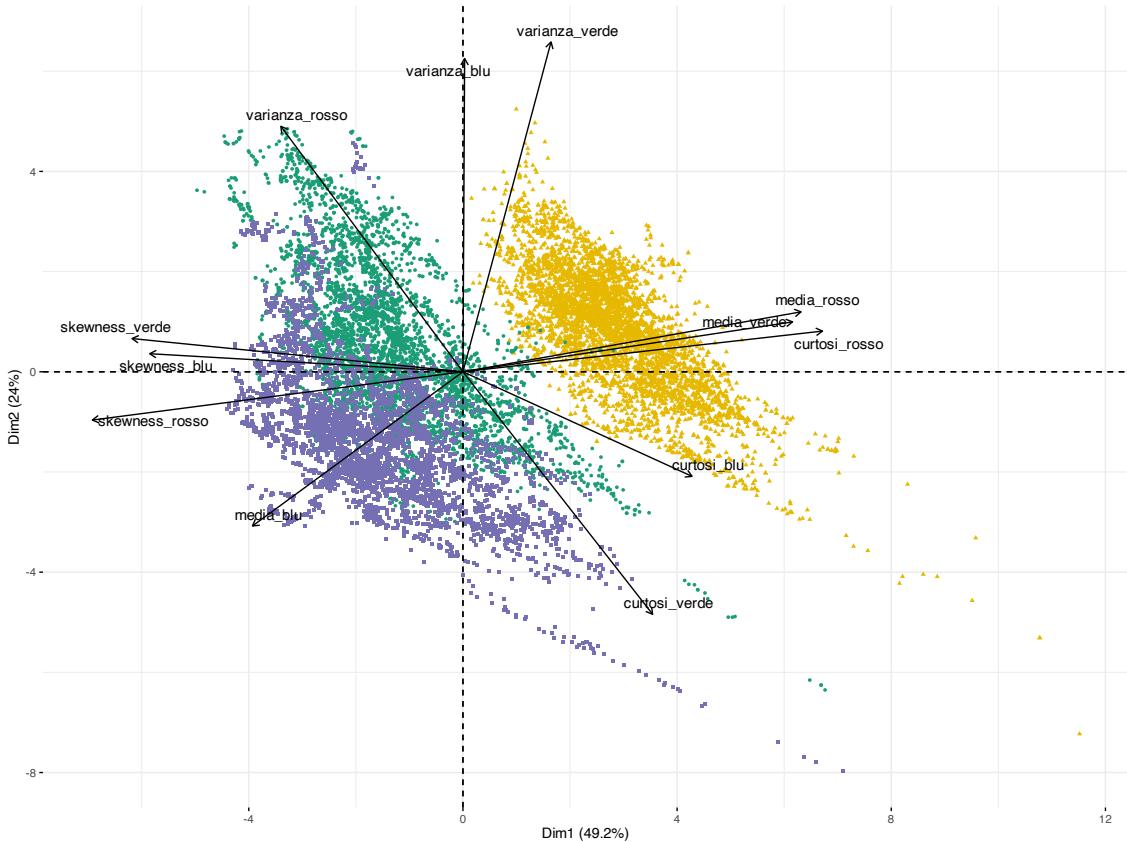


Figura 2.3.1: Biplot ricavato dall’analisi delle componenti principali (giallo: benigno, verde: adenocarcinoma, viola: carcinoma a cellule squamose)

Le prime due componenti principali preservano il 73.2% della varianza totale. Dal biplot si osserva una netta separazione del tumore benigno dalle altre due tipologie, mentre c’è una separazione meno evidente tra l’adenocarcinoma e il carcinoma a cellule squamose. Le immagini di queste due ultime tipologie di tumore sono caratterizzate principalmente dalle variabili `skewness_rosso`, `skewness_verde` e `skewness_blu`, oltre che dalla variabile `media_blu`. Le immagini del tumore benigno sono

invece caratterizzate da valori più alti delle variabili media_rosso, media_verde e curtosi_rosso.

Sono stati inoltre calcolati i coefficienti di correlazione tra le variabili e le componenti principali:

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
media_rosso	0.88	0.17	-0.12	0.26	-0.04
media_verde	0.86	0.14	0.10	0.43	0.13
media_blu	-0.55	-0.43	0.53	0.28	0.24
varianza_rosso	-0.47	0.68	0.32	-0.10	0.31
varianza_verde	0.23	0.92	0.21	0.10	0.01
varianza_blu	0.00	0.87	-0.03	-0.34	0.10
skewness_rosso	-0.96	-0.13	0.13	0.02	0.02
skewness_verde	-0.86	0.09	0.03	0.04	-0.42
skewness_blu	-0.82	0.05	-0.53	0.10	0.08
curtosi_rosso	0.94	0.11	-0.17	-0.07	-0.11
curtosi_verde	0.49	-0.67	-0.14	-0.32	0.36
curtosi_blu	0.60	-0.29	0.63	-0.26	-0.26

Tabella 2.3.1: Coefficienti di correlazione tra le variabili e le prime cinque componenti principali

La prima componente principale contrappone principalmente le variabili media_rosso, media_verde e curtosi_rosso alle variabili skewness_rosso, skewness_verde e skewness_blu. La prima componente principale potrebbe essere interpretata come un indice sulla tipologia di tumore: valori elevati indicano che l'immagine contiene un tumore benigno, mentre valori più bassi indicano che siamo in presenza di un tumore maligno.

Elevati valori del coefficiente di asimmetria dei pixel dell'immagine di tutti e tre i canali colore sembrano quindi associati a tumori maligni.

2.4 Stimatore di nucleo

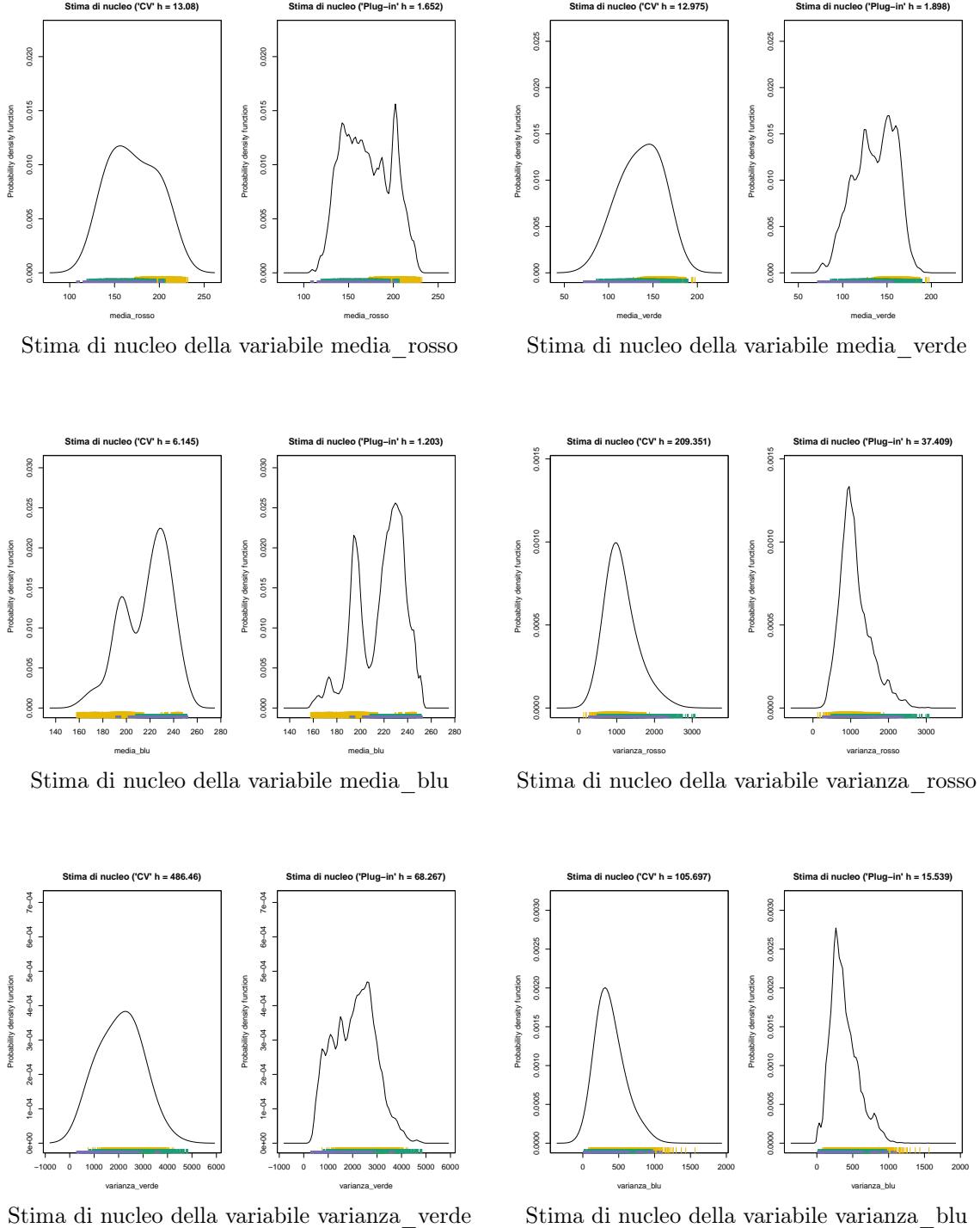
Un metodo alternativo e più raffinato rispetto agli istogrammi per rappresentare la forma della distribuzione di una variabile quantitativa è la stima di nucleo.

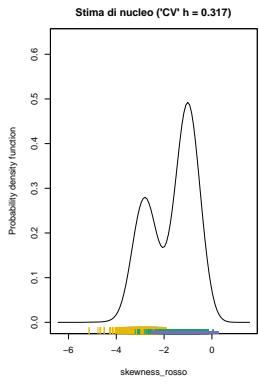
Il codice utilizzato è il seguente. Sono stati utilizzati dei rug colorati per tipologia di tumore: giallo per il tumore benigno, verde per l'adenocarcinoma e viola per il carcinoma a cellule squaumose.

¹ `library(sm)`
² `attach(df)`

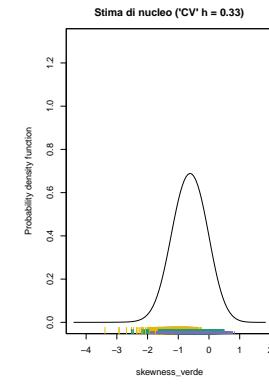
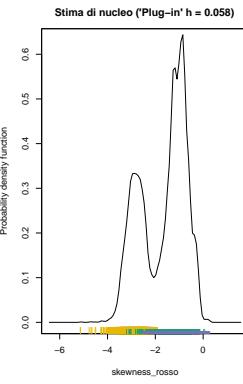
```

3
4  for (variabile in names(df[,1:12])){
5    var <- get(variabile)
6    h_cv <- h.select(var, method = "cv")
7    h_pi <- hsj(var)
8
9    par(mfrow = c(1, 2))
10
11   # con metodo cross validation
12   sm.density(
13     var,
14     h = h_cv,
15     xlab = variabile,
16     rugplot = FALSE
17   )
18
19   rug(var[classe == "benigno"], col = "#E7B800", side = 1, ticksize = 0.03,
20     ↪ lwd = 0.3)
21   rug(var[classe == "adenocarcinoma"], col = "#1b9E77", side = 1, ticksize
22     ↪ = 0.02, lwd = 0.3)
23   rug(var[classe == "carcinoma a cellule squamose"], col = "#7570B3", side
24     ↪ = 1, ticksize = 0.01, lwd = 0.3)
25
26   title(main = paste0("Stima di nucleo ('CV' h = ", round(h_cv, 3), ")"))
27   # con metodo plug-in
28   sm.density(
29     var,
30     h = h_pi,
31     xlab = variabile,
32     rugplot = FALSE
33   )
34
35   rug(var[classe == "benigno"], col = "#E7B800", side = 1, ticksize = 0.03,
36     ↪ lwd = 0.3)
37   rug(var[classe == "adenocarcinoma"], col = "#1b9E77", side = 1, ticksize
38     ↪ = 0.02, lwd = 0.3)
39   rug(var[classe == "carcinoma a cellule squamose"], col = "#7570B3", side
40     ↪ = 1, ticksize = 0.01, lwd = 0.3)
41
42   title(main = paste0("Stima di nucleo ('Plug-in' h = ", round(h_pi,
43     ↪ 3), ")"))
44 }
```

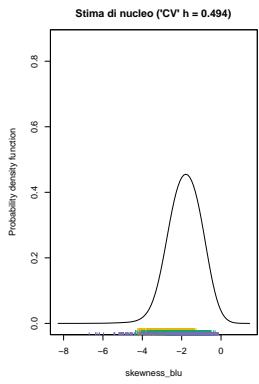
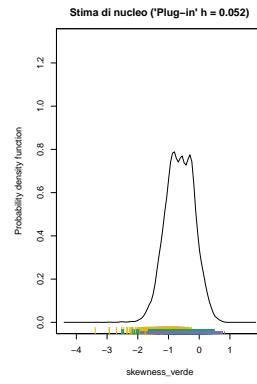




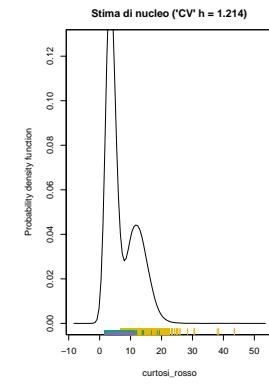
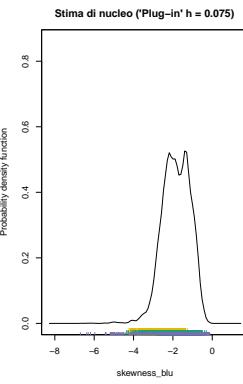
Stima di nucleo della variabile skewness_rosso



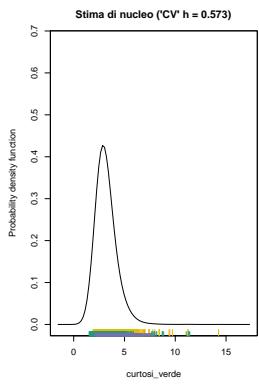
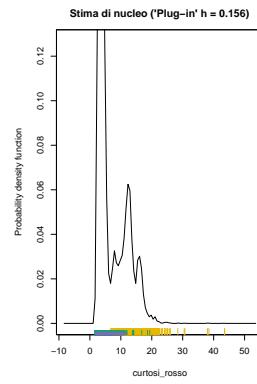
Stima di nucleo della variabile skewness_verde



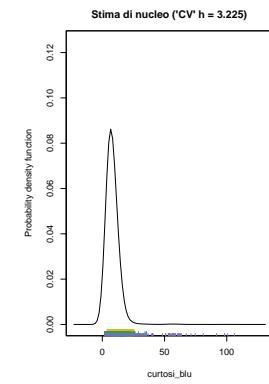
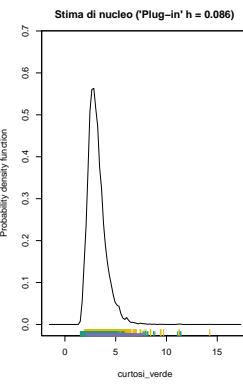
Stima di nucleo della variabile skewness_blu



Stima di nucleo della variabile curtosi_verde



Stima di nucleo della variabile curtosi_verde



Stima di nucleo della variabile curtosi_blu

2.5 Stimatore di nucleo bivariato

La stima di nucleo si può utilizzare anche per distribuzioni bivariate.

A titolo di esempio è stato inserito il codice utilizzato per creare i grafici della prima coppia di variabili:

```
1 library(sm)
```

```

2
3 var_1 <- df$media_rosso
4 var_2 <- df$media_blu
5
6 h_cv_1 <- h.select(var_1, method = "cv")
7 h_cv_2 <- h.select(var_2, method = "cv")
8
9 sm.density(df[, c("media_rosso", "media_blu")], h = c(h_cv_1, h_cv_2))
10
11 title(main = paste0("Stima di nucleo bivariato ('CV' h1 = ", round(h_cv_1,
12   ↵ 3), ", h2 = ", round(h_cv_2, 3), ")"))
13
14 sm.density(df[, c("media_rosso", "media_blu")], h = c(h_cv_1, h_cv_2),
15   ↵ display = "image",)
16 title(main = paste0("Stima di nucleo bivariato ('CV' h1 = ", round(h_cv_1,
17   ↵ 3), ", h2 = ", round(h_cv_2, 3), ")"))

```

- **media_rosso e media_blu**

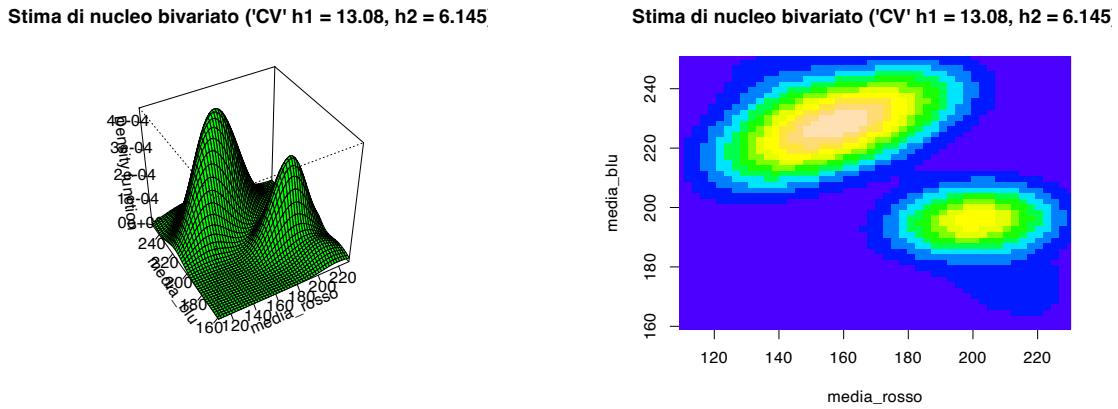


Figura 2.5.1: Stima di nucleo bivariato delle variabili media_rosso e media_blu

- **curtosi_rosso e skewness_rosso**

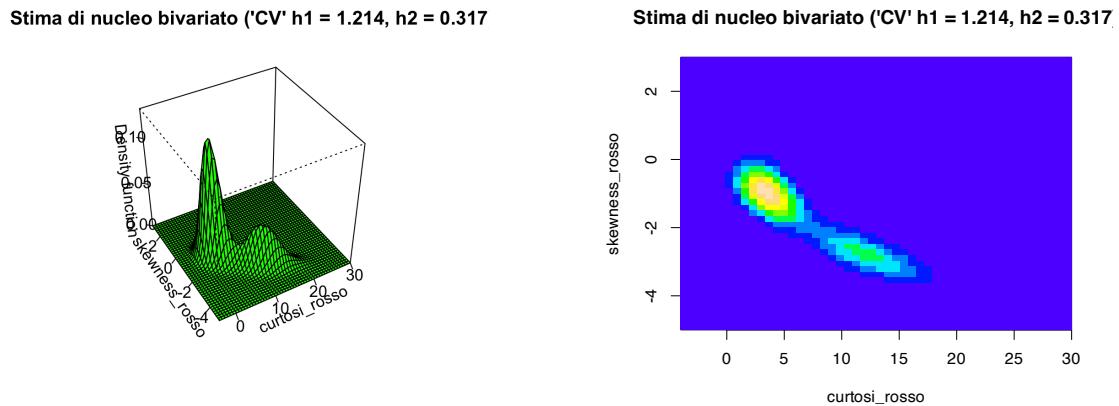


Figura 2.5.2: Stima di nucleo bivariato delle variabili `curtosi_rosso` e `skewness_rosso`

- `media_blu` e `skewness_blu`

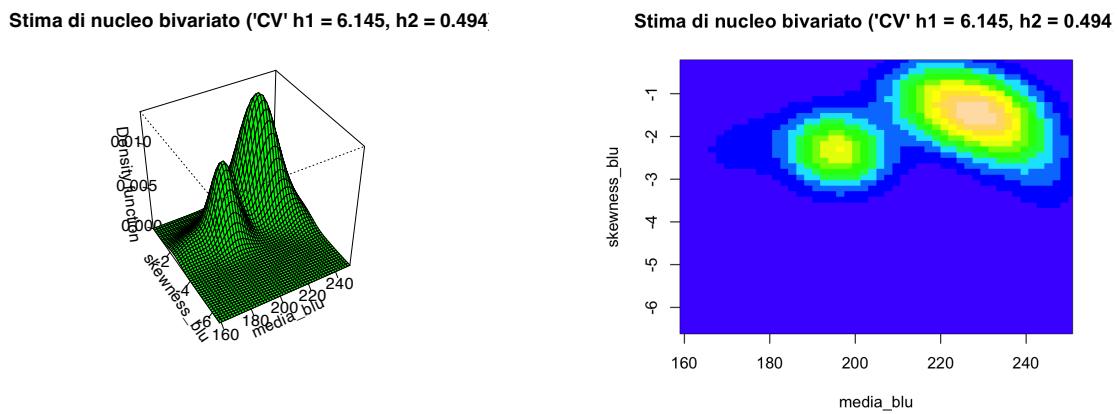
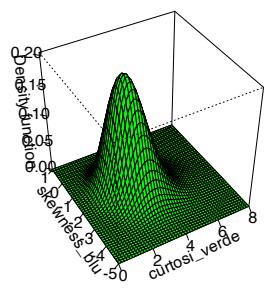


Figura 2.5.3: Stima di nucleo bivariato delle variabili `media_blu` e `skewness_blu`

- `curtosi_verde` e `skewness_blu`

Stima di nucleo bivariato ('CV' h1 = 0.573, h2 = 0.494)



Stima di nucleo bivariato ('CV' h1 = 0.573, h2 = 0.494)

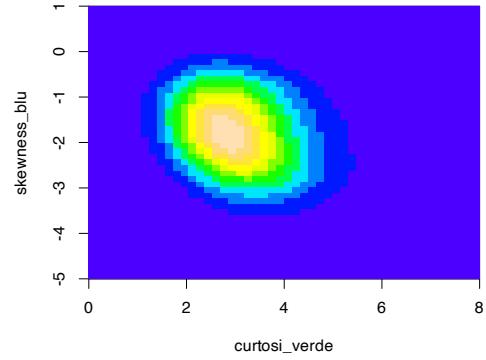


Figura 2.5.4: Stima di nucleo bivariato delle variabili curtosi_verde e skewness_blu

Nella quasi totalità delle distribuzioni bivariate stimate si osservano due modalità distinte: una caratterizzata da picchi di densità più elevati, probabilmente riconducibili alla presenza sia di adenocarcinoma sia di carcinoma a cellule squamose, e l'altra associata alle osservazioni del tumore benigno. L'unica eccezione è rappresentata dalla coppia di variabili curtosi_verde e skewness_blu, per la quale la stima di densità mostra una distribuzione monomodale.