Assignment 02

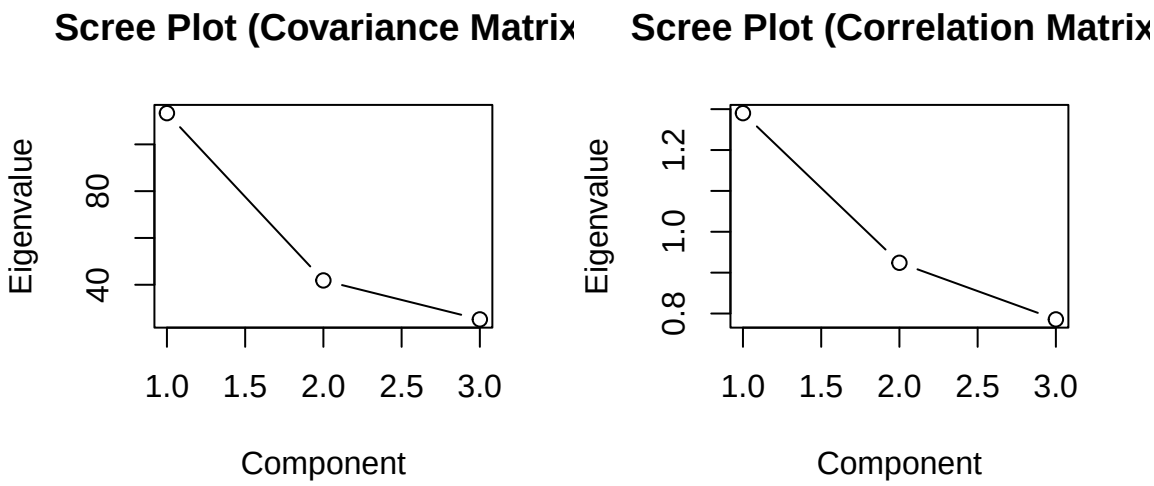
Mauricio Vazquez & Mariana Luna

2024-11-08

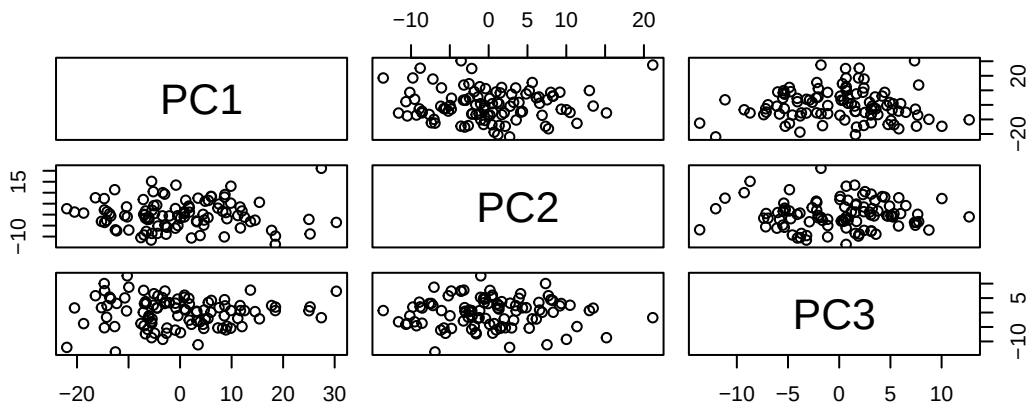
Repository link: https://github.com/MauricioVazquezM/Multivariate_Statistical_Course_Assignments_Fall2024

(Izenman) Ex. 7.1: Modern Multivariate Statistical

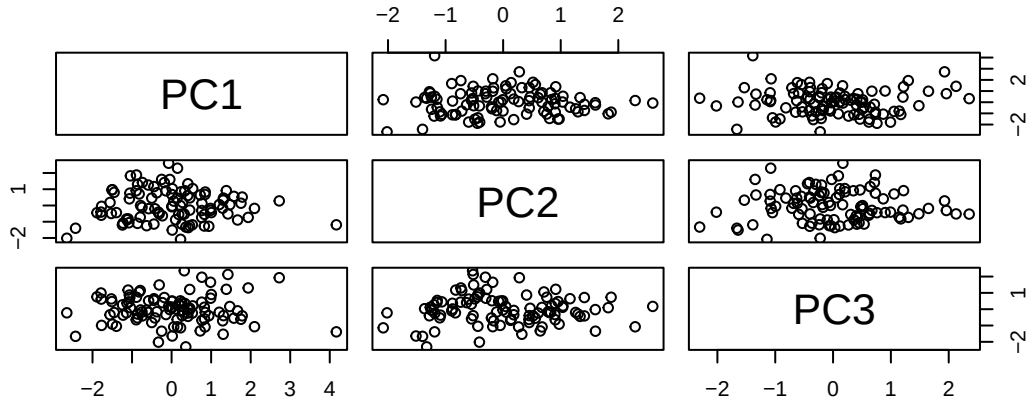
- Generate a random sample of size $n = 100$ from a three-dimensional ($r = 3$) Gaussian distribution, where one of the variables has very high variance (relative to the other two). Carry out PCA on these data using the covariance matrix and the correlation matrix. In each case, find the eigenvalues and eigenvectors, draw the scree plot, compute the PC scores, and plot all pairwise PC scores in a matrix plot. Compare results.



Pairwise PC Scores (Covariance Matrix)



Pairwise PC Scores (Correlation Matrix)




```
## Eigenvalues (Covariance Matrix): 113.3416 41.83933 25.19447

## Eigenvectors (Covariance Matrix):

##             PC1        PC2        PC3
## X1 -0.97443115  0.2230967  0.0266800
## X2 -0.21468549 -0.9595039  0.1823797
## X3 -0.06628786 -0.1719886 -0.9828661

## Eigenvalues (Correlation Matrix): 1.290194 0.9241341 0.785672

## Eigenvectors (Correlation Matrix):

##            PC1        PC2         PC3
## X1 -0.6175791  0.3977132 -0.67854273
## X2 -0.6384475  0.2503372  0.72781599
## X3 -0.4593266 -0.8826978 -0.09931593
```

(Izenman) Ex. 7.3: Modern Multivariate Statistical Techniques

- In the file `turtles.txt`, there are three variables, length, width, and height, of the carapaces of 48 painted turtles, 24 female and 24 male. Take logarithms of all three variables. Estimate the mean vector and covariance matrix of the male turtles and of the female turtles separately. Find the eigenvalues and eigenvectors of each estimated covariance matrix and carry out a PCA of each data set. Find an expression for the volume of a turtle carapace for males and for females. (Hint: use the fact that the variables are logarithms of the original measurements.) Compare volumes of male and female carapaces.

```
## Vector de medias (machos): 4.725444 4.477574 3.703186

## Vector de medias (hembras): 4.900356 4.623264 3.938253

## Matriz de covarianza (machos):

##             log_length   log_width  log_height
## log_length 0.011072004 0.008019142 0.008159648
## log_width  0.008019142 0.006416726 0.006005271
## log_height 0.008159648 0.006005271 0.006772758

## Matriz de covarianza (hembras):

##            log_length  log_width log_height
## log_length 0.02639101 0.02019836 0.02544294
## log_width  0.02019836 0.01629653 0.01987634
## log_height 0.02544294 0.01987634 0.02589861

## Valores propios (machos): 0.02330335 0.0005983049 0.000359836

## Vectores propios (machos):

##                   PC1        PC2        PC3
## log_length 0.6831023  0.1594791 -0.7126974
## log_width  0.5102195  0.5940118  0.6219534
## log_height 0.5225392 -0.7884900  0.3244015

## Valores propios (hembras): 0.06732574 0.0007362018 0.0005242089

## Vectores propios (hembras):

##                   PC1        PC2         PC3
## log_length 0.6216304 -0.5003288 -0.60269952
## log_width  0.4854905 -0.3577251  0.79770403
## log_height 0.6147151  0.7884820 -0.02053215

## Volumen promedio del caparazón (machos): 402803.1

## Volumen promedio del caparazón (hembras): 702129.2

## Diferencia de volumen entre machos y hembras: 299326
```

(Izenman) Ex. 7.10: Modern Multivariate Statistical Techniques

- Consider an $(r \times r)$ correlation matrix with the same correlation, ρ, say, in the off-diagonal entries. Find the eigenvalues and eigenvectors of this matrix when $r = 2, 3, 4$. Generalize your results to any r variables. As examples, set $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$.

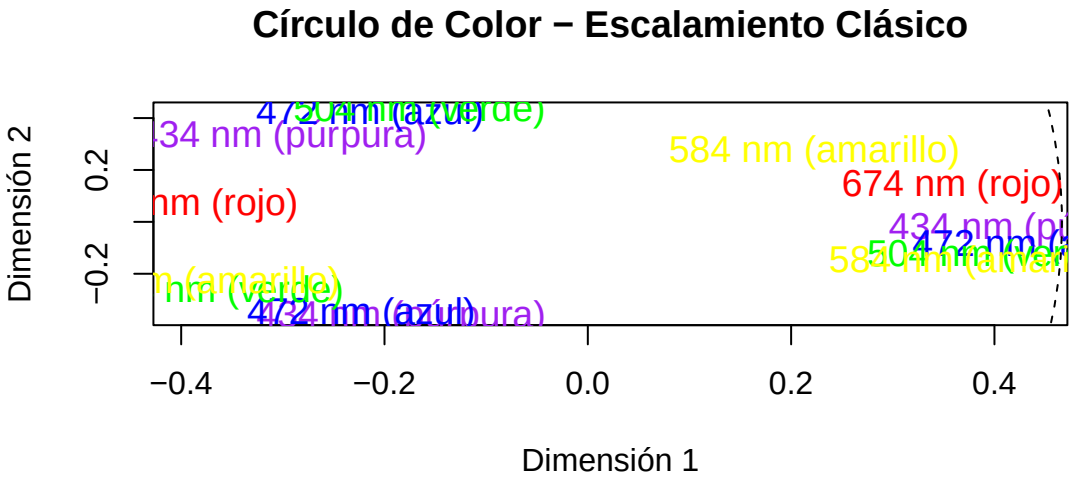
```
## r = 2
## p = 0.1
## Eigenvalues:
## [1] 1.1 0.9
## Eigenvectors:
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
##
## p = 0.3
## Eigenvalues:
## [1] 1.3 0.7
## Eigenvectors:
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
##
## p = 0.5
## Eigenvalues:
## [1] 1.5 0.5
## Eigenvectors:
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
##
## p = 0.7
## Eigenvalues:
## [1] 1.7 0.3
## Eigenvectors:
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
##
## p = 0.9
## Eigenvalues:
## [1] 1.9 0.1
## Eigenvectors:
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
##
## r = 3
## p = 0.1
## Eigenvalues:
## [1] 1.2 0.9 0.9
## Eigenvectors:
##           [,1]       [,2]       [,3]
## [1,] 0.5773503  0.0000000  0.8164966
```

```
## [2,] 0.5773503 -0.7071068 -0.4082483
## [3,] 0.5773503  0.7071068 -0.4082483
##
## p = 0.3
## Eigenvalues:
## [1] 1.6 0.7 0.7
## Eigenvectors:
##            [,1]       [,2]       [,3]
## [1,] -0.5773503  0.8164966  0.0000000
## [2,] -0.5773503 -0.4082483 -0.7071068
## [3,] -0.5773503 -0.4082483  0.7071068
##
## p = 0.5
## Eigenvalues:
## [1] 2.0 0.5 0.5
## Eigenvectors:
##            [,1]       [,2]       [,3]
## [1,] -0.5773503  0.0000000  0.8164966
## [2,] -0.5773503 -0.7071068 -0.4082483
## [3,] -0.5773503  0.7071068 -0.4082483
##
## p = 0.7
## Eigenvalues:
## [1] 2.4 0.3 0.3
## Eigenvectors:
##           [,1]       [,2]        [,3]
## [1,] 0.5773503  0.3555207  0.73503175
## [2,] 0.5773503 -0.8143165 -0.05962589
## [3,] 0.5773503  0.4587958 -0.67540586
##
## p = 0.9
## Eigenvalues:
## [1] 2.8 0.1 0.1
## Eigenvectors:
##           [,1]       [,2]        [,3]
## [1,] 0.5773503  0.4586617  0.67549693
## [2,] 0.5773503 -0.8143284  0.05946423
## [3,] 0.5773503  0.3556666 -0.73496116
##
## r = 4
## p = 0.1
## Eigenvalues:
## [1] 1.3 0.9 0.9 0.9
## Eigenvectors:
##      [,1]       [,2]          [,3]         [,4]
## [1,] -0.5  0.5099852  0.000000e+00  0.699939347
## [2,] -0.5  0.4899141  8.756053e-17 -0.714131779
## [3,] -0.5 -0.4999496 -7.071068e-01  0.007096216
## [4,] -0.5 -0.4999496  7.071068e-01  0.007096216
##
## p = 0.3
## Eigenvalues:
## [1] 1.9 0.7 0.7 0.7
## Eigenvectors:
```

```
##      [,1]       [,2]       [,3]       [,4]
## [1,] -0.5  0.8660254  0.0000000  0.0000000
## [2,] -0.5 -0.2886751  0.0000000  0.8164966
## [3,] -0.5 -0.2886751 -0.7071068 -0.4082483
## [4,] -0.5 -0.2886751  0.7071068 -0.4082483
##
## p = 0.5
## Eigenvalues:
## [1] 2.5 0.5 0.5 0.5
## Eigenvectors:
##      [,1]       [,2]       [,3]       [,4]
## [1,] -0.5  0.8660254  0.0000000  0.0000000
## [2,] -0.5 -0.2886751 -0.5773503 -0.5773503
## [3,] -0.5 -0.2886751 -0.2113249  0.7886751
## [4,] -0.5 -0.2886751  0.7886751 -0.2113249
##
## p = 0.7
## Eigenvalues:
## [1] 3.1 0.3 0.3 0.3
## Eigenvectors:
##      [,1]       [,2]       [,3]       [,4]
## [1,] -0.5  0.0000000  0.0000000  0.8660254
## [2,] -0.5 -0.5773503 -0.5773503 -0.2886751
## [3,] -0.5 -0.2113249  0.7886751 -0.2886751
## [4,] -0.5  0.7886751 -0.2113249 -0.2886751
##
## p = 0.9
## Eigenvalues:
## [1] 3.7 0.1 0.1 0.1
## Eigenvectors:
##      [,1]       [,2]       [,3]       [,4]
## [1,] -0.5  0.8519866  0.0000000  0.1553024
## [2,] -0.5 -0.4304160  0.0000000  0.7514932
## [3,] -0.5 -0.2107853 -0.7071068 -0.4533978
## [4,] -0.5 -0.2107853  0.7071068 -0.4533978

## Resultado general:

## Un valor propio es 1 + (r-1)*p, con el vector propio (1, 1, ..., 1) / sqrt(r).

## Los otros (r-1) valores propios son 1 - p, con vectores propios que suman cero.
```

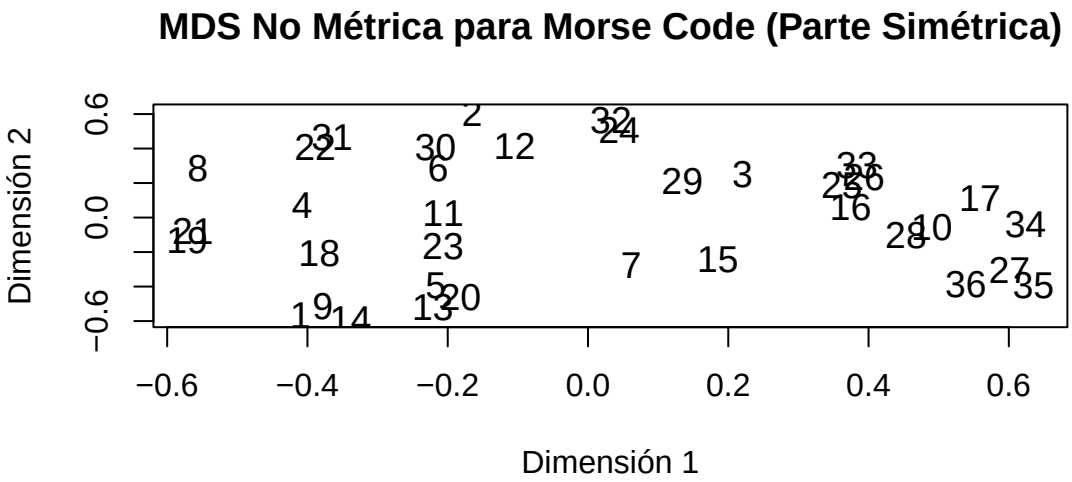
(Izenman) Ex. 13.1: Modern Multivariate Statistical Techniques

- Consider the color-stimuli experiment outlined in Section 13.2.1. The similarity ratings are given in the file `color-stimuli` on the book's website. Carry out a classical scaling of the data and show that the solution is a "color circle" ranging from violet (434 mmu) to blue (472 mmu) to green (504 mmu) to yellow (584 mmu) to red (674 mmu). Compare your solution to the nonmetric scaling solution given in Figure 13.3.



(Izenman) Ex. 13.2: Modern Multivariate Statistical Techniques

- Consider the Morse-code experiment outlined in Section 13.2.2. The file `Morse-code` on the book's website gives a table of the percentages of times that a signal corresponding to the row label was identified as being the same as the signal corresponding to the column label. A row of this table shows the confusion rate for that particular Morse-code signal when presented *before* each of the column signals, whereas a column of the table shows the confusion rate for that particular signal when presented *after* each of the row signals. This table of confusion rates is not symmetric and the diagonal elements are not each 100%. Now, every square matrix M can be decomposed uniquely into the sum of two orthogonal matrices, $M = A + B$, where $A = \frac{1}{2}(M + M^T)$ is symmetric ($A^T = A$), and $B = \frac{1}{2}(M - M^T)$ is skew-symmetric ($B^T = -B$) with zero diagonal entries.Find the decomposition for the Mose-code data. Ignore that part of the Morse-code data provided by **B** and carry out a nonmetric scaling only of the symmetric part **A**. Decide how many dimensions you think are appropriate for representing the data.



```
## Eigenvalores de la MDS no métrica:

##  [1]  5.205365e+00  4.353316e+00  3.487860e+00  2.845465e+00  2.458606e+00
##  [6]  2.362910e+00  2.167358e+00  1.843835e+00  1.802066e+00  1.766812e+00
## [11]  1.667725e+00  1.503634e+00  1.380941e+00  1.321786e+00  1.222535e+00
## [16]  1.102229e+00  1.006210e+00  8.711539e-01  8.458618e-01  7.976113e-01
## [21]  7.654828e-01  7.215666e-01  6.957785e-01  6.102155e-01  5.832254e-01
## [26]  5.520200e-01  5.133764e-01  4.738729e-01  3.979680e-01  3.202436e-01
## [31]  2.751521e-01  2.286207e-01  1.479234e-01  2.536853e-02  5.237705e-15
## [36] -2.652066e-16
```

(Izenman) Ex. 13.4: Modern Multivariate Statistical Techniques

- Show that the dissimilarities in the matrix Δ are Euclidean distances if and only if the doubly centered matrix $\mathbf{B} = \mathbf{HAH}$ is nonnegative definite, where $\mathbf{A}$ is given in the classical scaling algorithm of Table 13.5.

```
## Valores propios de B:

## [1]  0  0 -2 -2

## ¿B es semidefinida positiva?: FALSE
```

(Johnson & Wichern) Ex. 8.9: Applied Multivariate Statistical Analysis

- **Check book**

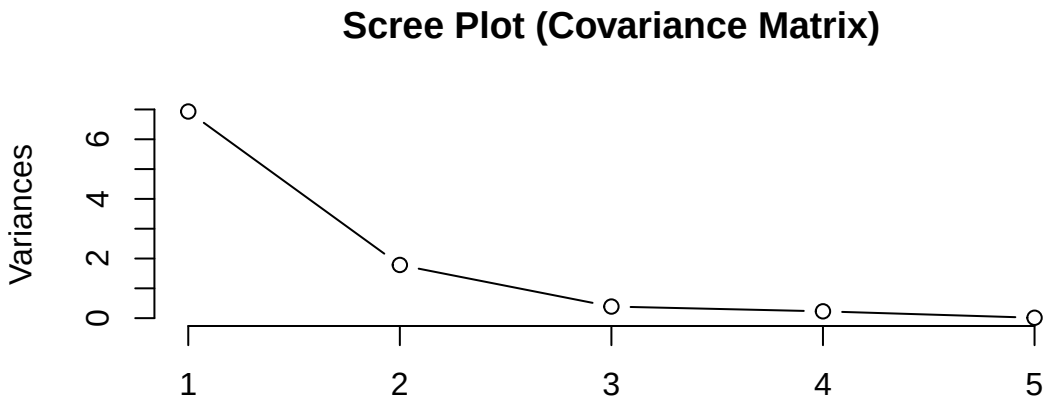
Respuesta en pdf anexo

(Johnson & Wichern) Ex. 8.12: Applied Multivariate Statistical Analysis

- Consider the air-pollution data listed in Table 1.5. Your job is to summarize these data in fewer than $p = 7$ dimensions if possible. Conduct a principal component analysis of the data using both the covariance matrix $\mathbf{S}$ and the correlation matrix $\mathbf{R}$. What have you learned? Does it make any difference which matrix is chosen for analysis? Can the data be summarized in three or fewer dimensions? Can you interpret the principal components?

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.7404 1.1363 0.7566 0.30887 0.11005
## Proportion of Variance 0.6058 0.2582 0.1145 0.01908 0.00242
## Cumulative Proportion  0.6058 0.8640 0.9785 0.99758 1.00000

## [1] 0.605779211 0.258227592 0.114491133 0.019079697 0.002422368

## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5
## Standard deviation     2.6327 1.3361 0.62422 0.47909 0.11897
## Proportion of Variance 0.7413 0.1909 0.04168 0.02455 0.00151
## Cumulative Proportion  0.7413 0.9323 0.97394 0.99849 1.00000
```

```
## [1] 0.741326833 0.190933682 0.041675786 0.024549724 0.001513975
```




```
## [1] "Cargas usando matriz de correlación:"
```

```
##                                    PC1          PC2          PC3         PC4
## Total_population           -0.5583589  0.131392987  0.007945807 -0.55055321
## Median_school_years        -0.3132830  0.628872546 -0.549030533  0.45265380
## Total_employment           -0.5682577  0.004262264  0.117280380 -0.26811649
## Health_services_employment -0.4866246 -0.309560576  0.454923806  0.64798227
## Median_value_home           0.1742664  0.701005911  0.691224986 -0.01510711
##                                    PC5
## Total_population            0.606464575
## Median_school_years        -0.006564747
## Total_employment           -0.769040874
## Health_services_employment  0.201325679
## Median_value_home          -0.014203097
```

```
## [1] "Cargas usando matriz de covarianza:"
```

```
##                                    PC1         PC2          PC3         PC4
## Total_population           -0.78120807 -0.07087183 -0.003656607  0.54171007
## Median_school_years        -0.30564856 -0.76387277  0.161817438 -0.54479937
## Total_employment           -0.33444840  0.08290788 -0.014841008  0.05101636
## Health_services_employment -0.42600795  0.57945799 -0.220453468 -0.63601254
## Median_value_home           0.05435431 -0.26235528 -0.961759720  0.05127599
##                                    PC5
## Total_population           -0.302039670
## Median_school_years        -0.009279632
## Total_employment            0.937255367
## Health_services_employment -0.172145212
## Median_value_home           0.024583093
```

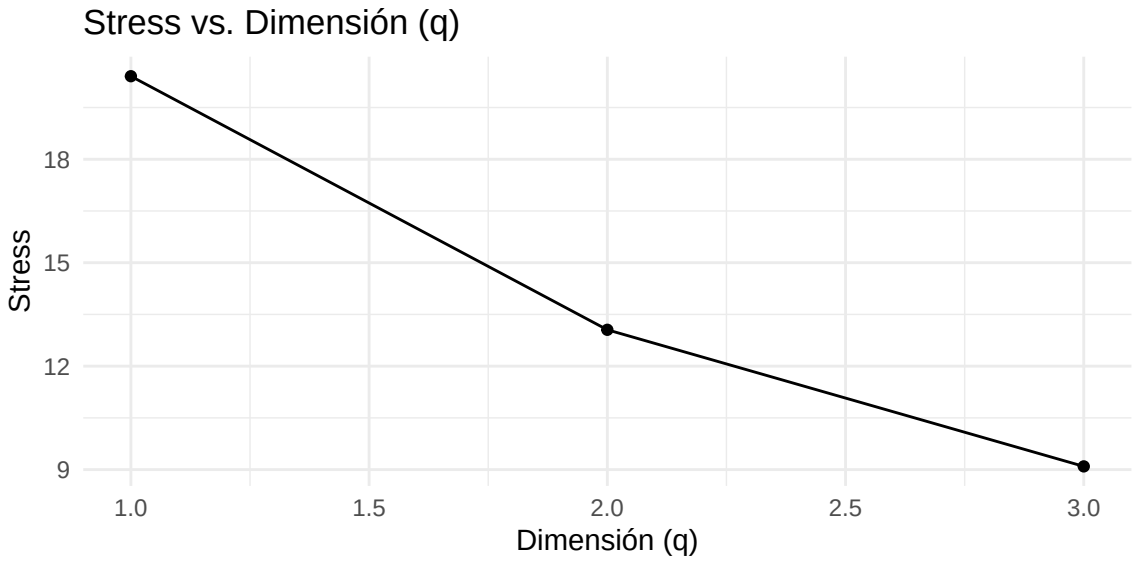
(Johnson & Wichern) Ex. 12.18: Applied Multivariate Statistical Analysis

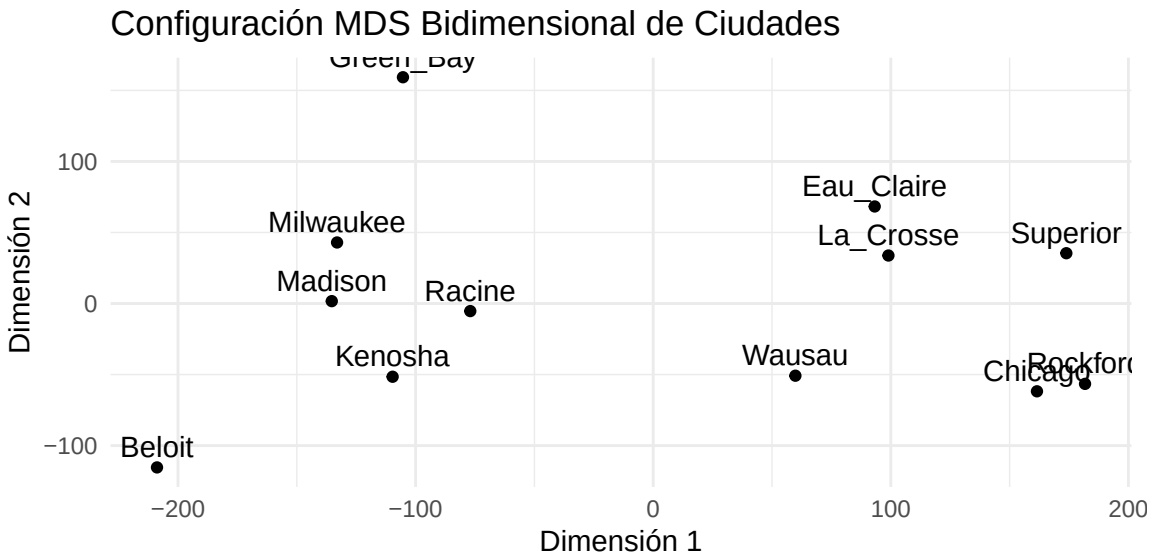
- Table 12.12 gives the road distances between 12 Wisconsin cities and cities in neighboring states. Locate the cities in $q = 1, 2$, and 3 dimensions using multidimensional scaling. Plot the minimum stress (q) versus q and interpret the graph. Compare the two-dimensional multidimensional scaling configuration with the locations of the cities on a map from an atlas.

```
## initial  value 20.411904
## final  value 20.408388
## converged
```

```
## initial  value 13.060463
## final  value 13.054540
## converged
```

```
## initial  value 9.098001
## final  value 9.093091
## converged
```



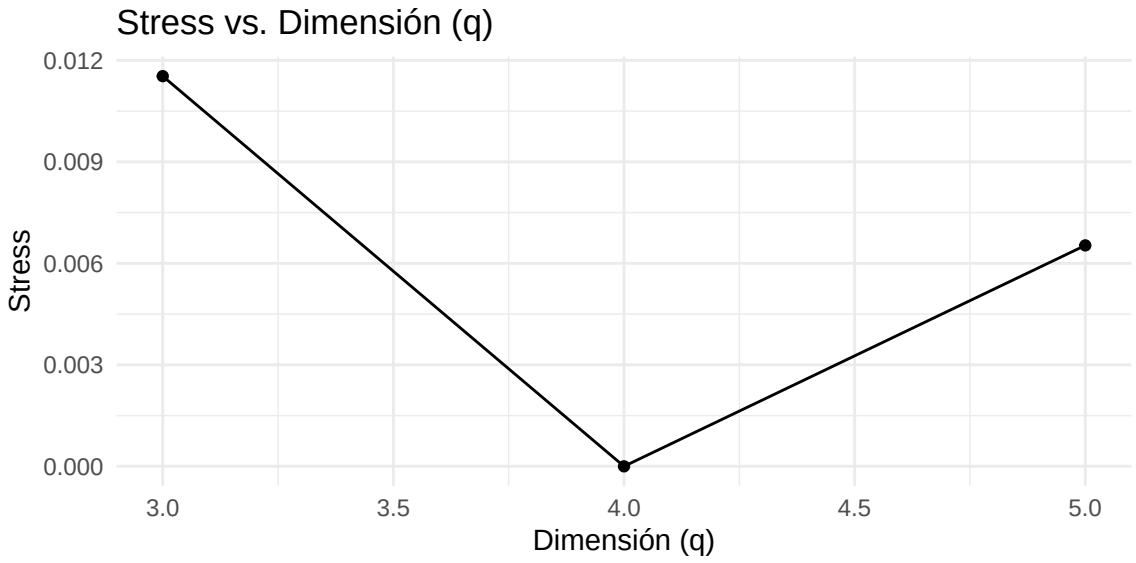


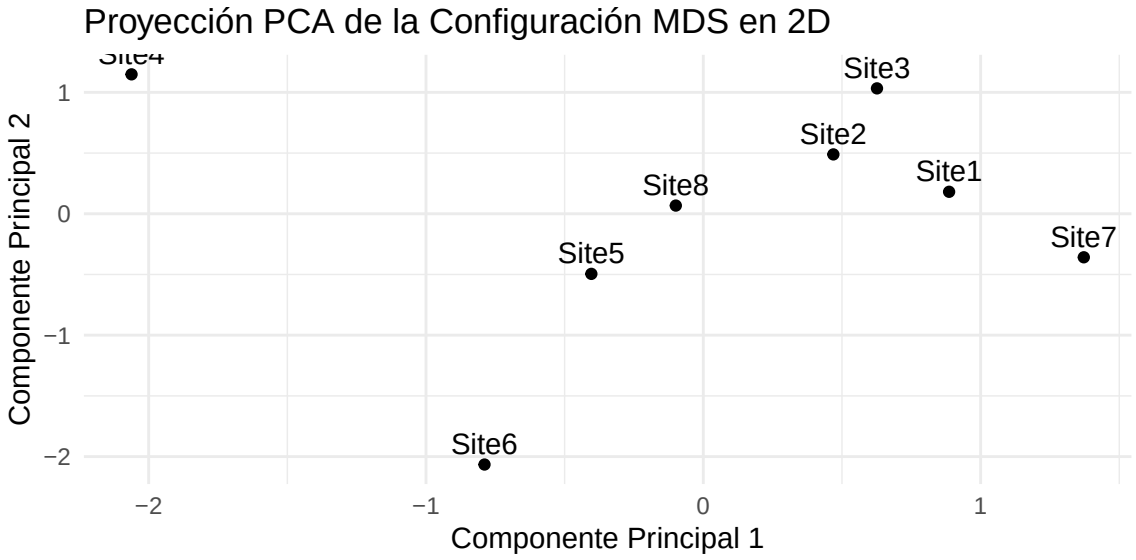
(Johnson & Wichern) Ex. 12.19: Applied Multivariate Statistical Analysis

- Table 12.13 on page 744 gives the "distances" between certain archaeological sites from different periods, based upon the frequencies of different types of potsherds found at the sites. Given these distances, determine the coordinates of the sites in $q = 3, 4$, and 5 dimensions using multidimensional scaling. Plot the minimum stress (q) versus q and interpret the graph. If possible, locate the sites in two dimensions (the first two principal components) using the coordinates for the $q = 5$-dimensional solution. (Treat the sites as variables.) Noting the periods associated with the sites, interpret the two-dimensional configuration.

```
## initial  value 7.177846
## iter   5 value 2.996872
## iter  10 value 1.036919
## iter  15 value 0.674844
## iter  20 value 0.464397
## iter  25 value 0.424302
## iter  30 value 0.403276
## iter  35 value 0.314097
## iter  40 value 0.140983
## iter  45 value 0.051686
## iter  50 value 0.011533
## final  value 0.011533
## stopped after 50 iterations

## initial  value 3.819719
## iter   5 value 1.285489
## iter  10 value 0.651872
## iter  15 value 0.156843
## iter  20 value 0.037688
## iter  20 value 0.000000
## iter  20 value 0.000000
## final  value 0.000000
## converged
```

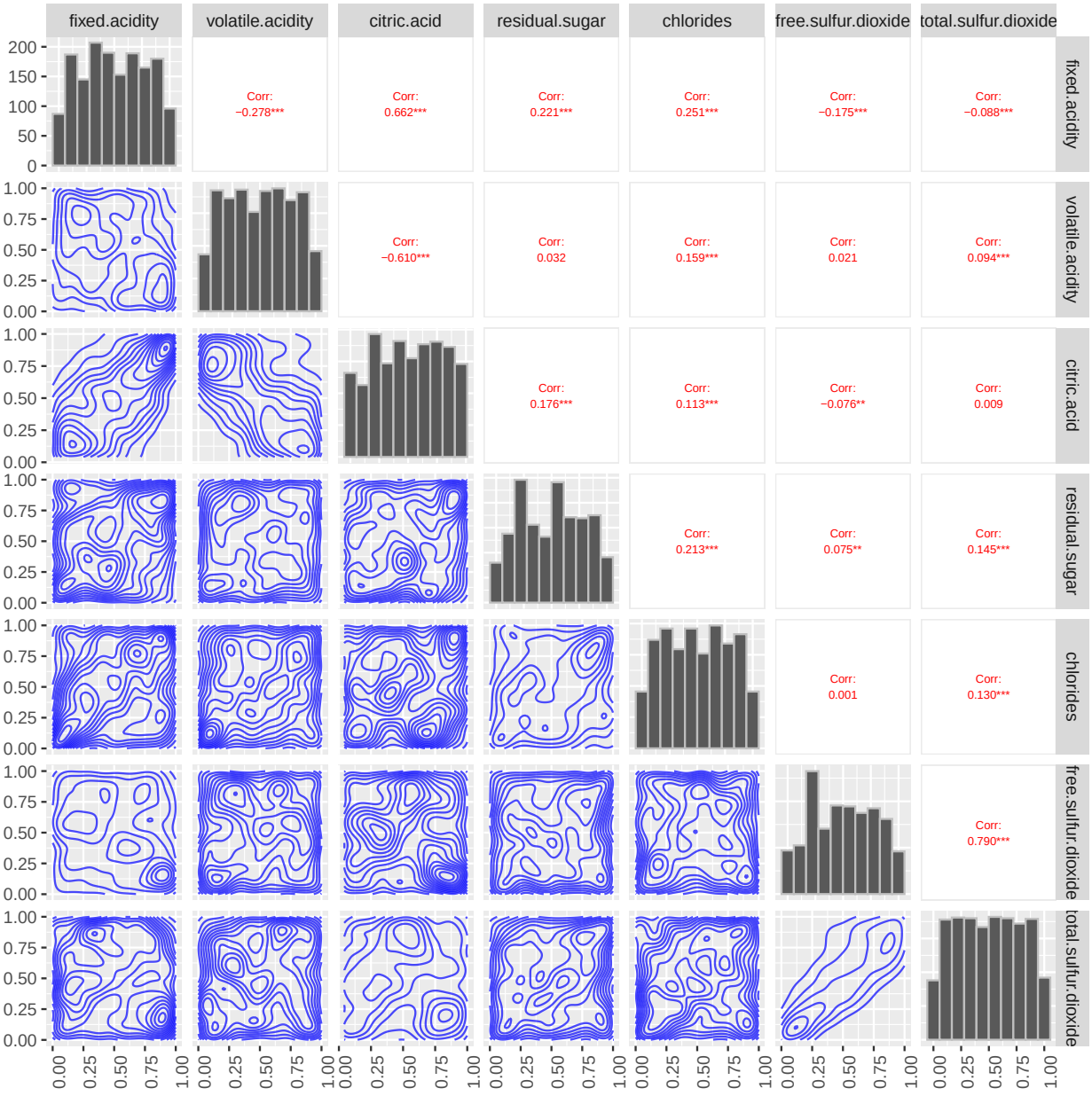
```
## initial  value 2.236951
## iter   5 value 0.217803
## iter  10 value 0.011232
## iter  10 value 0.006531
## iter  10 value 0.006531
## final  value 0.006531
## converged
```





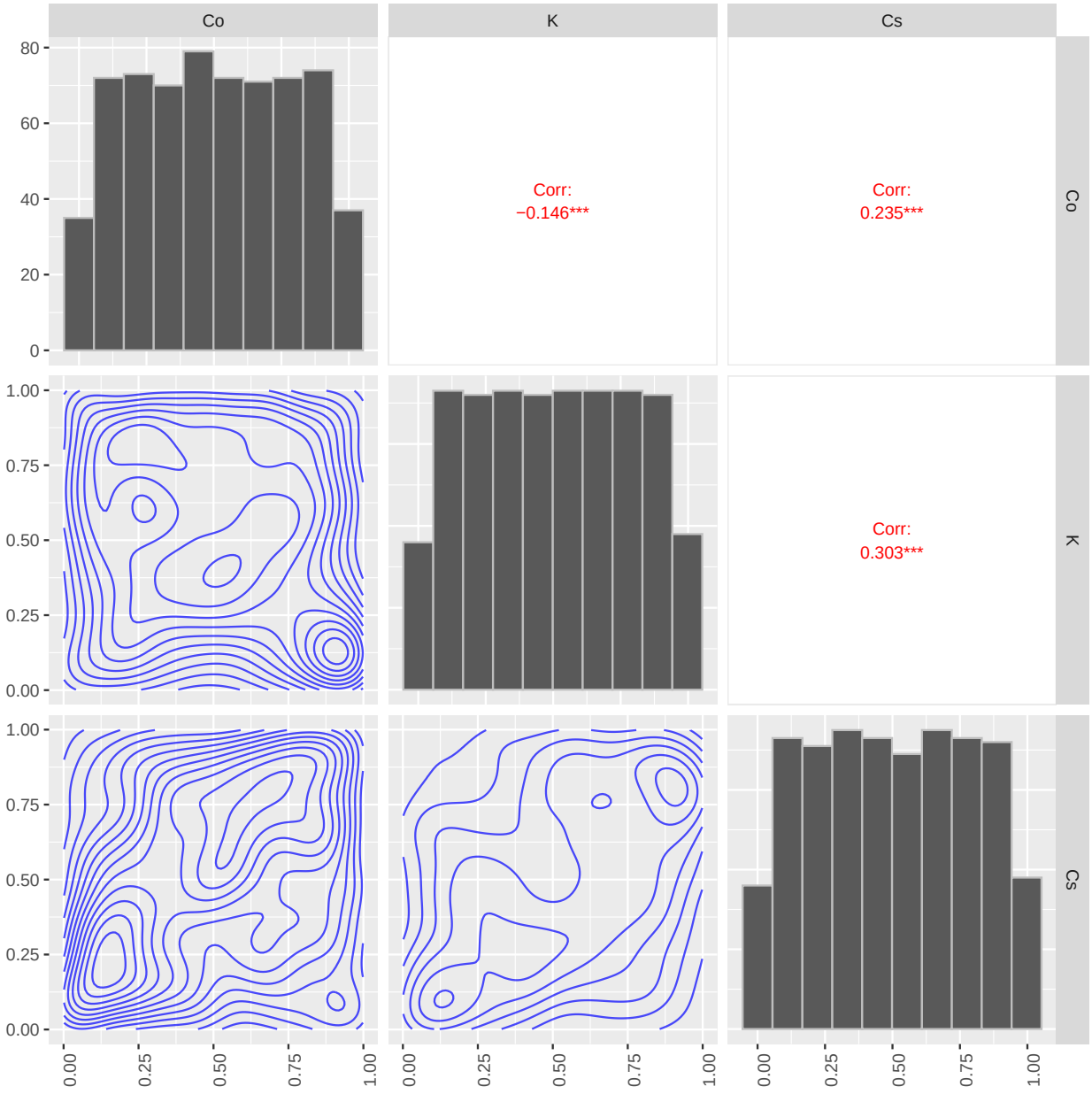
(Czado) Ex. 3.3: Analyzing Dependent Data with Vine Copulas

- *Exploratory bivariate copula choices for the seven-dimensional red wine data:* For the data set considered in Exercise 1.7 the pairs plot of the associated pseudo-copula data is given in Fig. 3.15. For each pair of variables propose a pair copula family.



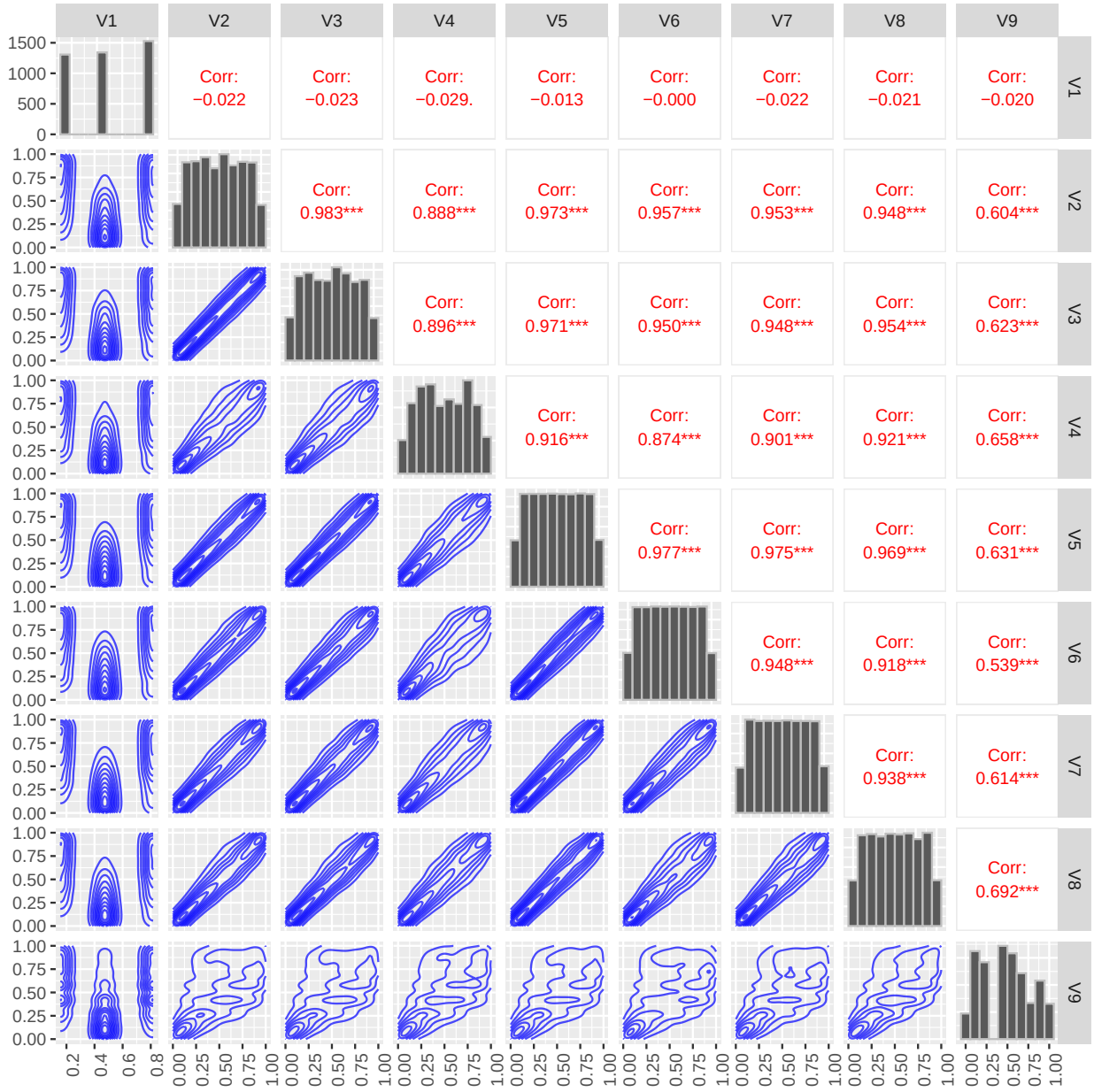
(Czado) Ex. 3.4: Analyzing Dependent Data with Vine Copulas

- *URAN3: Exploratory copula choices for the three-dimensional uranium data:* Consider as in Example 2.2 the three-dimensional subset of the *uranium* data set contained in the R package `copula` with variables Cobalt (Co), Titanium (Ti) and Scandium (Sc). As in Example 3.4 transform the original data to the copula scale using marginal empirical distributions. Then explore the empirical normalized contour plots for all pairs of variables and suggest appropriate parametric pair copula families. Check your choices by comparing the fitted to the empirical normalized contour plots.



(Czado) Ex. 3.5: Analyzing Dependent Data with Vine Copulas

- *ABALONE3: Exploratory copula choices for the three-dimensional abalone data:* Consider as in Example 2.3 the three-dimensional subset of the *abalone* data set contained in the R package `PivotalR` with variables *shucked*, *viscera*, and *shell*. As in Example 3.4 transform the original data to the copula scale using marginal empirical distributions. Then explore the empirical normalized contour plots for all pairs of variables and suggest appropriate parametric pair copula families. Check your choices by comparing the fitted to the empirical normalized contour plots.

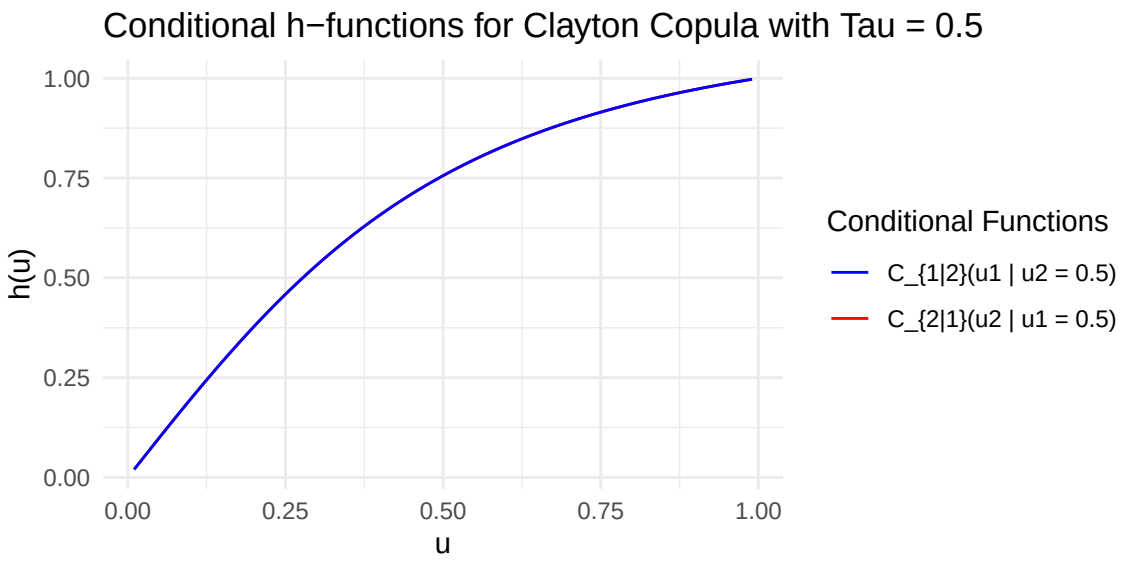


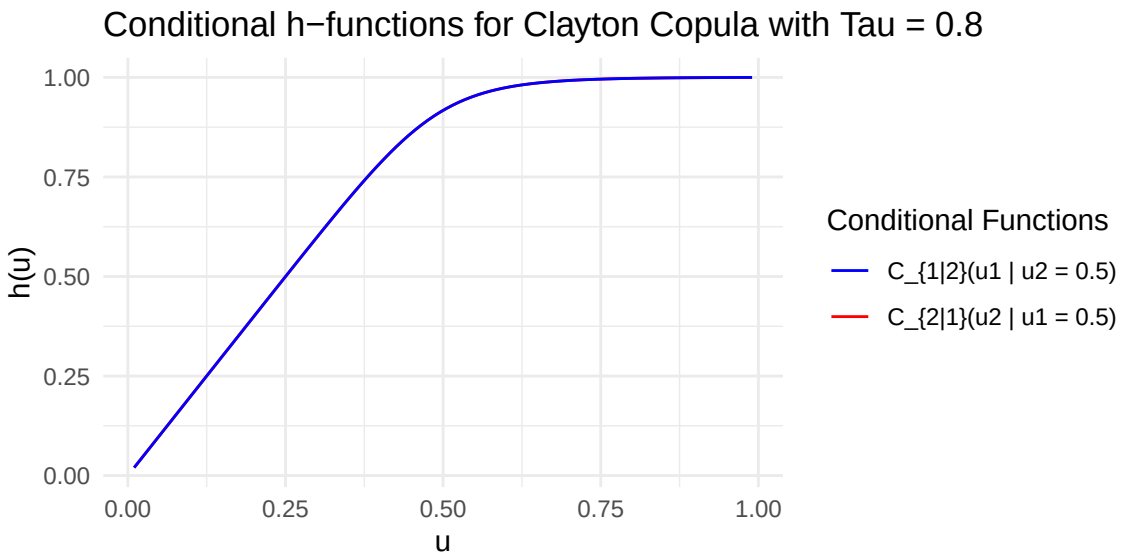
(Czado) Ex. 3.6: Analyzing Dependent Data with Vine Copulas

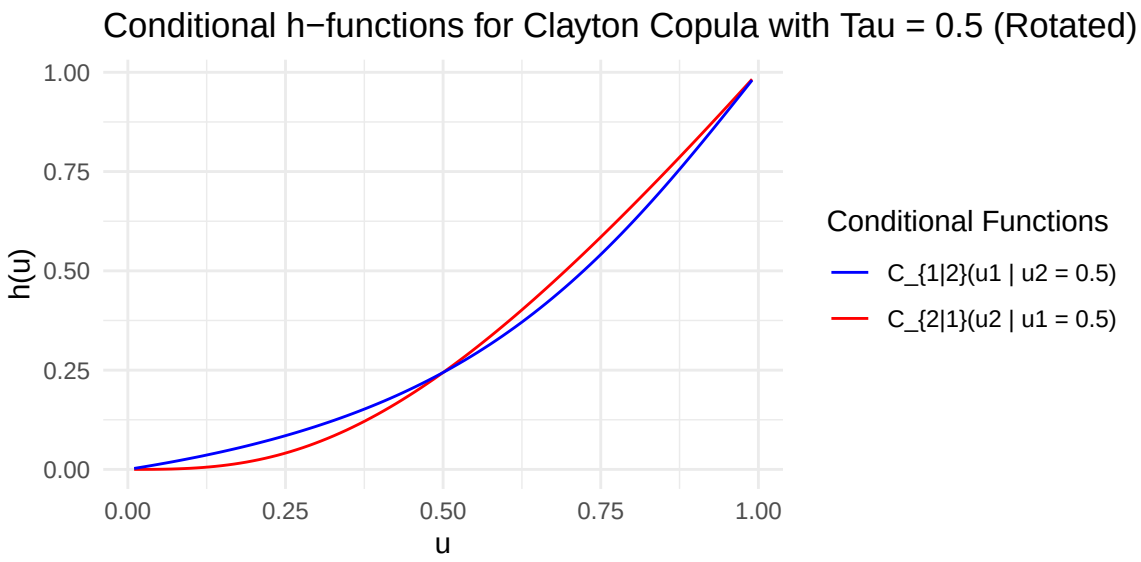
- *The effect of the degree of freedom in a bivariate Student's t copula on the contour shapes:* For $df = 2, \ldots, 30$ draw the normalized contour plots, when the association parameter is $\rho = .7$. Do the same for $\rho = -.2$. How do these plots change when you fix $\tau = .7$ and $\tau = -.2$, respectively.

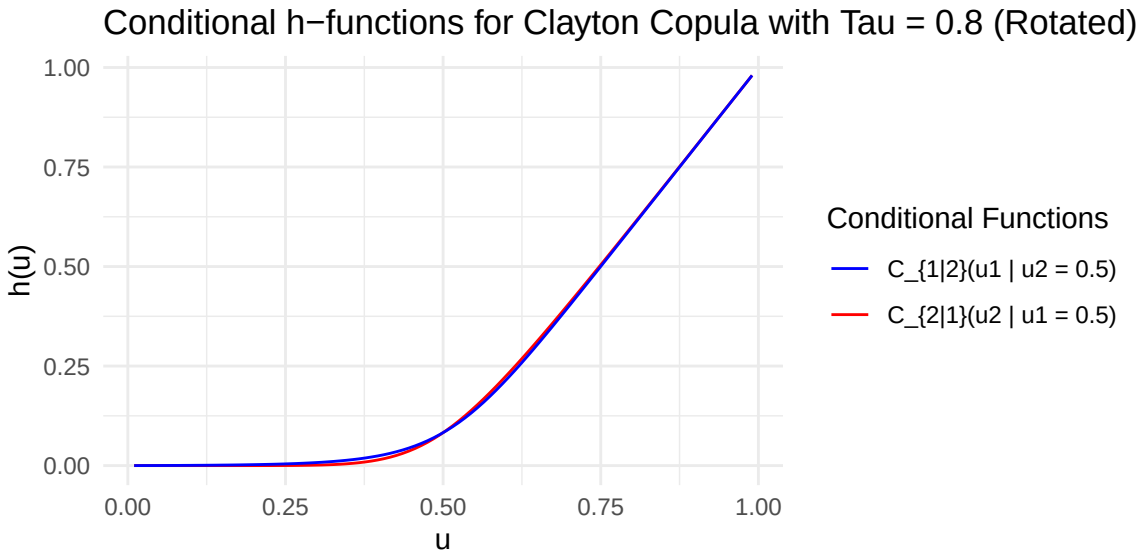
(Czado) Ex. 3.10: Analyzing Dependent Data with Vine Copulas

- *Conditional distribution of the Clayton copula:* Derive and visualize the h functions $C_{2|1}(u_2|u_1 = .5)$ and $C_{1|2}(u_1|u_2 = .5)$ of a bivariate Clayton copula with a Kendall's $\tau = .5$ and $\tau = .8$, respectively. Compare the two functions. Do the same for a 90° rotated Clayton copula.


Conditional h−functions for Clayton Copula with Tau = 0.5
1.00
0.75
0.50
0.25
0.00
h(u)
0.00
0.25
0.50
0.75
1.00
u
Conditional Functions
C_{1|2}(u1 | u2 = 0.5)
C_{2|1}(u2 | u1 = 0.5)


Conditional h−functions for Clayton Copula with Tau = 0.8
1.00
0.75
0.50
0.25
0.00
h(u)
0.00
0.25
0.50
0.75
1.00
u
Conditional Functions
C_{1|2}(u1 | u2 = 0.5)
C_{2|1}(u2 | u1 = 0.5)





Cuando tao es más alta, la dependencia es más fuerte, lo que se refleja en una curvatura más pronunciada en h.