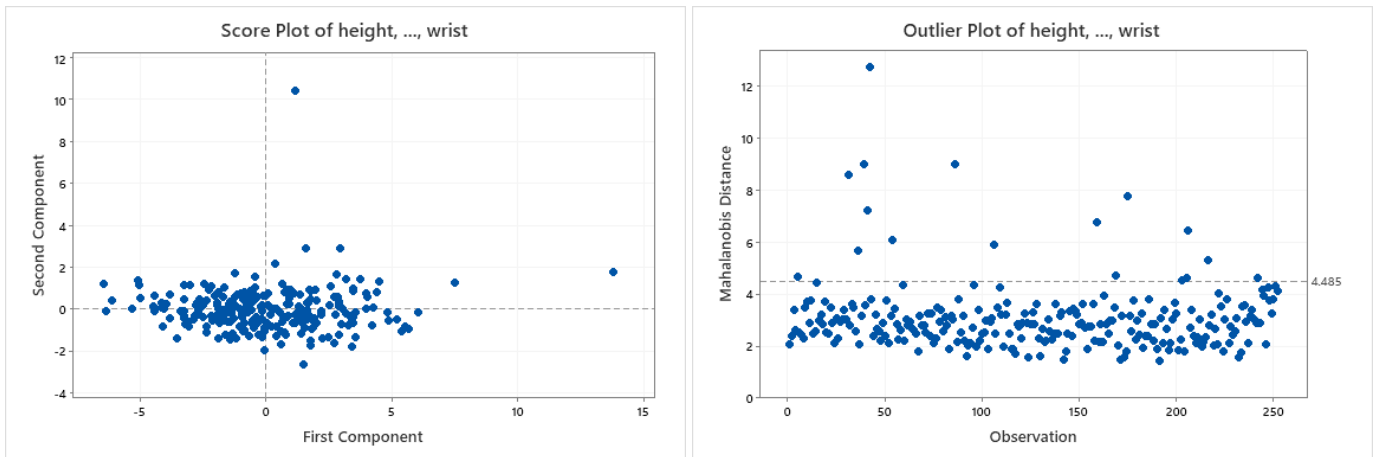


Additional Multivariate Exercise 11

Data: A description of the dataset with some basic statistics (including a full correlation matrix between the variables) were given in the solution for additional exercise 2.9. It is not surprising that most of the variables are fairly strongly, positively correlated.

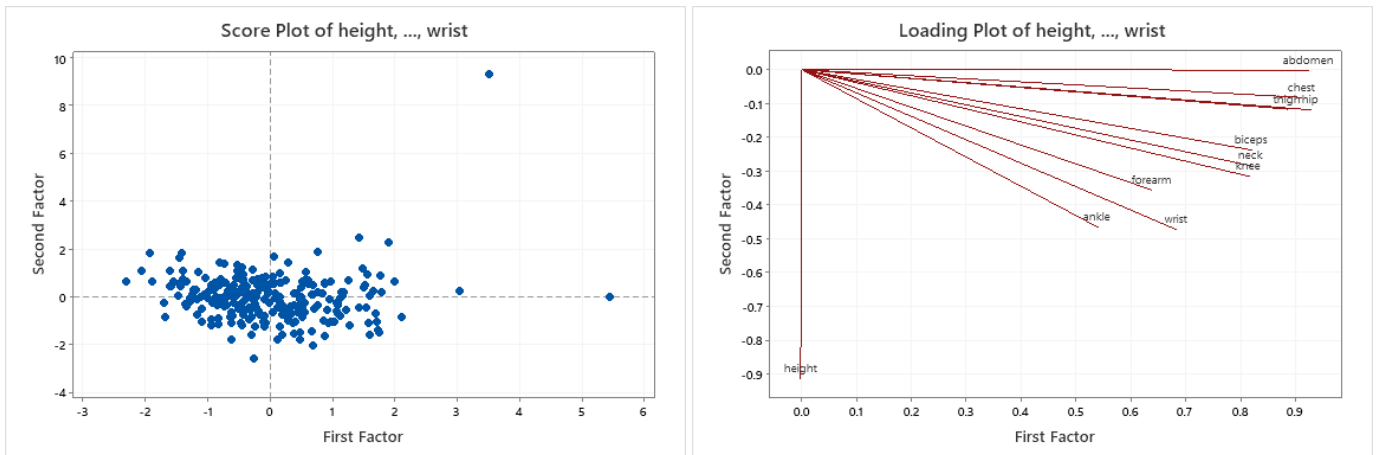
Part 1)

A principal components analysis (PCA) for the correlations among 11 physical body measurements (height – wrist) shows one dominant eigenvalue (7.09) and only one other eigenvalue above 1 (barely, at 1.10); together the two first components explain 74% of the variation. Because we will look at refinements of this analysis in subsequent questions, we focus here on only these two components. The loadings for the first component are all positive, and all in the range 0.25–0.35, except for height (0.11). This could be interpreted to represent the person's size. The second component loads strongly (-0.83) on height, whereas all other loadings are of small/moderate size around zero (within ± 0.27). This component very clearly represents height, with no obvious interpretation of the loadings for the other variables. The most interesting plots are perhaps the score plot and outlier plot (from Minitab).



The visual impression of the score plot is dominated by two extreme points: #39 with a high score on the first component and #42 with a very high score on the second component. These also stand out in the outlier plot, as the observations with the two highest values of the Mahalanobis distance, although other observations have almost as high a distance as #39.

For the factor analysis (FA), we follow up the PCA by a varimax rotation (in Minitab) of the two first components, the loading and score plots are shown on the next page. After the rotation, height has a zero loading on the first component, whereas the other variables kept their positive loadings. On the second component, all loadings have been shifted to have the same sign, still with the by far strongest loading for height, followed by moderate loadings for ankle and knee (without any obvious interpretation). The interpretation of the rotated factors is therefore quite similar as for the original PCA eigenvectors. The communalities range are all in the 0.7–0.9 range, except for ankle and forearm (both around 0.50). The FA score plot (not shown) is still dominated by the two observations discussed previously.

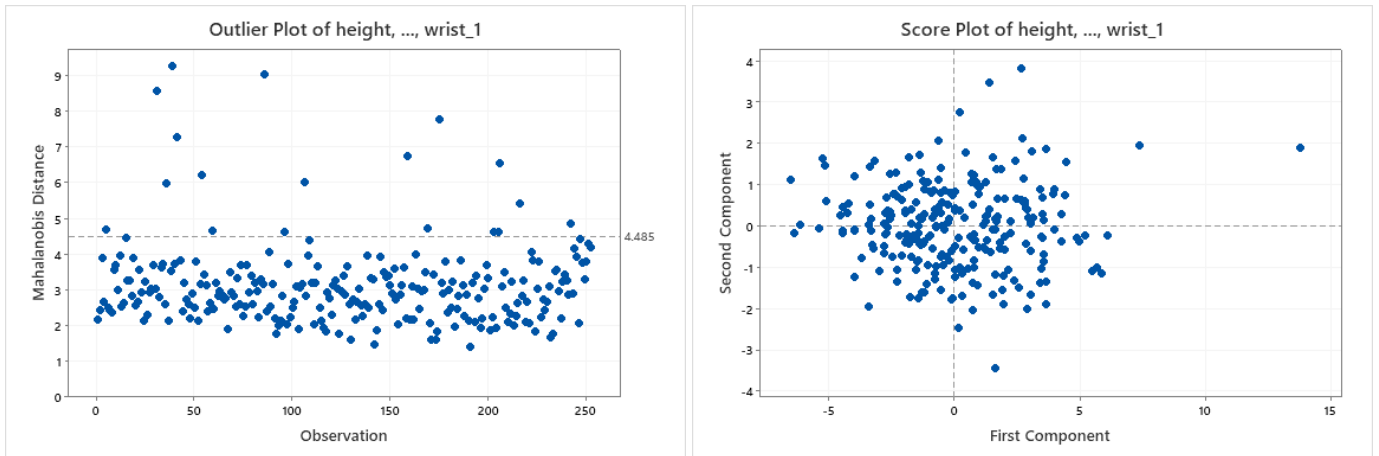


Part 2)

The two extreme observations identified (# 39 and 42) were already discussed in the additional exercise 2.9. The observation for height for #42 clearly does not match the other observations and must reasonably be considered as an error (it seems impossible to imagine a person with such body dimensions). Observation #39 has large values for all variables, in fact the largest value in the dataset for all variables except height, ankle and forearm. This is conceivable as corresponding to simply a very large person. This person may therefore not be an error, but could still be an outlier in terms of deviating from the other individuals. We try PCAs with and without observation 39, after dropping observation 42 as an error. We show a table of loadings for the first two components, before and after varimax rotation (in Minitab), with and without observation 39, as well as (on the next page) PCA outlier and score plots for the analysis with observation 39.

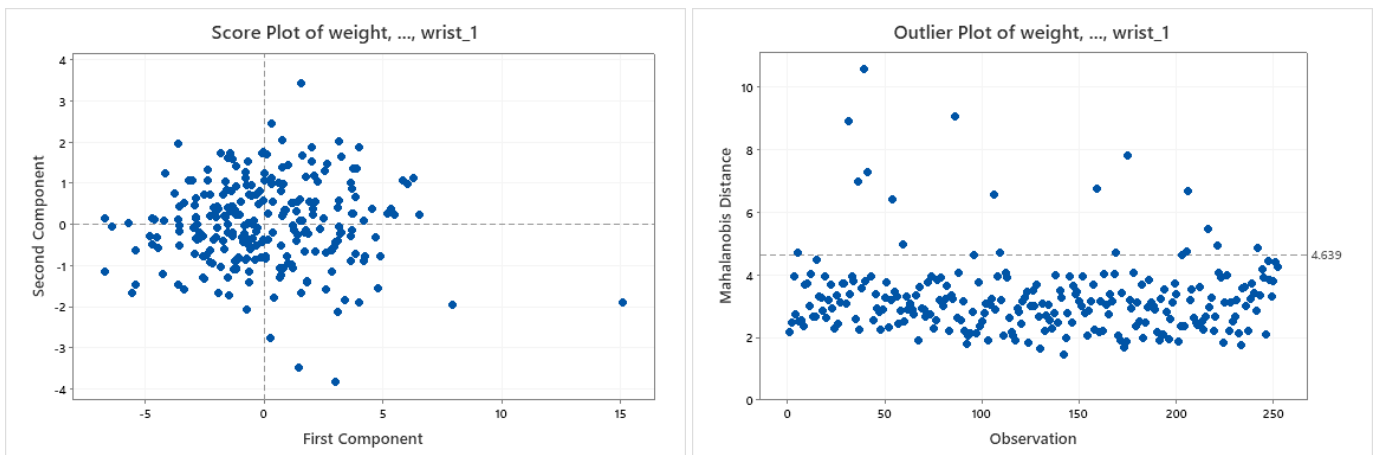
Variable /Components	With observation 39				Without observation 39			
	PCA loadings		FA loadings		PCA loadings		FA loadings	
	1	2	1	2	1	2	1	2
height	0.17	-0.77	0.05	-0.91	0.17	-0.74	0.06	-0.89
neck	0.32	0.12	0.83	-0.28	0.32	0.12	0.81	-0.28
chest	0.33	0.28	0.92	-0.14	0.33	0.30	0.92	-0.13
abdomen	0.32	0.33	0.93	-0.09	0.32	0.34	0.93	-0.08
hip	0.34	0.10	0.87	-0.32	0.35	0.09	0.86	-0.34
thigh	0.33	0.08	0.83	-0.32	0.33	0.07	0.80	-0.34
knee	0.33	-0.15	0.72	-0.53	0.33	-0.17	0.69	-0.55
ankle	0.25	-0.36	0.43	-0.63	0.24	-0.41	0.37	-0.67
biceps	0.32	0.08	0.80	-0.31	0.32	0.09	0.79	-0.30
forearm	0.27	-0.06	0.62	-0.38	0.28	-0.02	0.66	-0.36
wrist	0.30	-0.13	0.66	-0.48	0.30	-0.13	0.64	-0.49
variance	7.23	1.05	6.00	2.27	7.10	1.07	5.83	2.33
prop. variance	0.66	0.10	0.55	0.21	0.65	0.10	0.53	0.21

The results are seen to be very similar between analyses with and without observation 39, both for PCA and FA. The outlier plot still shows #39 as the most extreme in terms of Mahalanobis distance, but # 86 and 31 have only slightly smaller distances. The score plot still shows #39 as extreme on the first component, which as before can be interpreted to represent size. The PCA and FA loadings are actually also similar to those with observation 42 included, telling us (among other things) that the high loading for height on the second component was not due to the influence of this observation. It seems reasonable to continue analysis with the dataset with only the suspected error #42 excluded.



Part 3)

Including weight among the variables does not change the results too much. The first two eigenvalues still represent around 76% of the variation, and the next eigenvalues are still well below 1 (starting at 0.68). It is not obvious that the analysis should include further components. In the PCA, the variable weight loads as strongly as most of the other variable on the first component, therefore apparently just adding to the information about the person's size. It does not load on the second component, and actually not on any of the subsequent ones either. So it seems that weight did not add anything unique to the data. The FA further substantiates this interpretation by the very high value (0.96) of the communality for weight, whereas the communalities for the other variables show only very changes compared to the previous analysis. The factor loadings in Minitab also show a sign change for the second component, but this is unimportant for the interpretation. The variable weight has loadings of 0.88 and 0.43 on the two components, not contradicting our previous interpretations. The score and outlier plots for the PCA show observation 39 a bit more extreme than previously, because this person also had the highest weight. It does however not change our decision to include it in the analysis.



Part 4)

The interest is in whether the two extra variables, age and bodyfat, add anything substantial to the previous analyses. For this reason we shift to looking at the first three components, even if it means that the factor loadings are less comparable to those from previous analyses with only two factors. As will be seen, adding bodyfat to the variables does not alter the results substantially, so the 3-factor model with bodyfat included also (roughly) represents the results without this variable.

Variable /Components	With age variable						With bodyfat variable					
	PCA loadings			FA loadings			PCA loadings			FA loadings		
	1	2	3	1	2	3	1	2	3	1	2	3
age	0.01	0.73	0.44	0.02	-0.13	0.96	—	—	—	—	—	—
bodyfat	—	—	—	—	—	—	0.23	0.52	0.19	0.89	-0.14	0.10
weight	0.34	-0.02	-0.04	0.90	0.40	0.02	0.33	-0.03	0.11	0.70	-0.48	-0.50
height	0.16	-0.48	0.56	0.11	0.87	-0.18	0.15	-0.64	0.30	-0.06	-0.16	-0.87
neck	0.30	0.13	0.02	0.80	0.31	0.19	0.29	0.01	-0.27	0.53	-0.67	-0.27
chest	0.31	0.23	-0.14	0.91	0.14	0.21	0.31	0.22	-0.02	0.78	-0.48	-0.13
abdomen	0.31	0.27	-0.18	0.92	0.08	0.23	0.31	0.30	0.15	0.89	-0.35	-0.18
hip	0.33	-0.01	-0.19	0.90	0.27	-0.06	0.32	0.06	0.22	0.76	-0.36	-0.44
thigh	0.31	-0.13	-0.32	0.88	0.23	-0.24	0.30	0.02	0.14	0.68	-0.39	-0.42
knee	0.31	-0.07	0.12	0.74	0.50	0.03	0.30	-0.14	0.23	0.58	-0.36	-0.60
ankle	0.23	-0.21	0.31	0.44	0.62	-0.02	0.22	-0.32	0.23	0.30	-0.27	-0.65
biceps	0.30	-0.02	-0.16	0.82	0.26	-0.05	0.29	-0.01	-0.30	0.49	-0.68	-0.27
forearm	0.25	-0.09	-0.01	0.62	0.35	-0.06	0.24	-0.13	-0.66	0.20	-0.87	-0.18
wrist	0.28	0.11	0.41	0.60	0.56	0.36	0.27	-0.20	-0.27	0.33	-0.66	-0.42
variance	8.18	1.43	0.86	6.87	2.32	1.29	8.58	1.32	0.70	4.82	3.19	2.60
prop. variance	0.63	0.11	0.07	0.53	0.18	0.10	0.66	0.10	0.05	0.37	0.25	0.20

As indicated already, the results for bodyfat are similar to those previously seen. The third PCA component has an eigenvalue of only 0.70, and is in fact similar to those of previous PCAs (not shown). The bodyfat variable has a (weak) positive loading on the first component, and thus contributes (weakly) to its size. It loads positively on the second component, which for the other coefficients still contrast height to other variables. Thus this second component moves towards a contrast between height and bodyfat. The FA rotation changes the interpretations of the second and third factors considerably, but these factors do not substantially involve bodyfat, which is therefore essentially expressed on the first factor, with its usual size interpretation.

Turning next to the analysis with age, its impact shows up on the second and third components of the PCA. The third eigenvalue is also here below 1 (at 0.86) but large enough that a third component/factor could reasonably be included. The factor rotations show a nice separation of the age variable onto its own (third) factor, whereas the two first factors are similar to what we have seen before. The communality for age is very high (at 0.94) as one would expect with its large loading for the third factor. As this analysis was the best example of a substantial change in the factors with the rotation, we also include loading plots for the three factors before and after rotation (from Stata, on the next page).

