Additional Multivariate Exercise 9

Data: The `beef_ultra` dataset is described in the Datasets chapter of the VER2 textbook. Briefly, the data collected on 487 cattle, as they entered a feedlot for fattening prior to slaughter, consist of demographic information plus readings obtained from an ultrasonic evaluation of the animal.

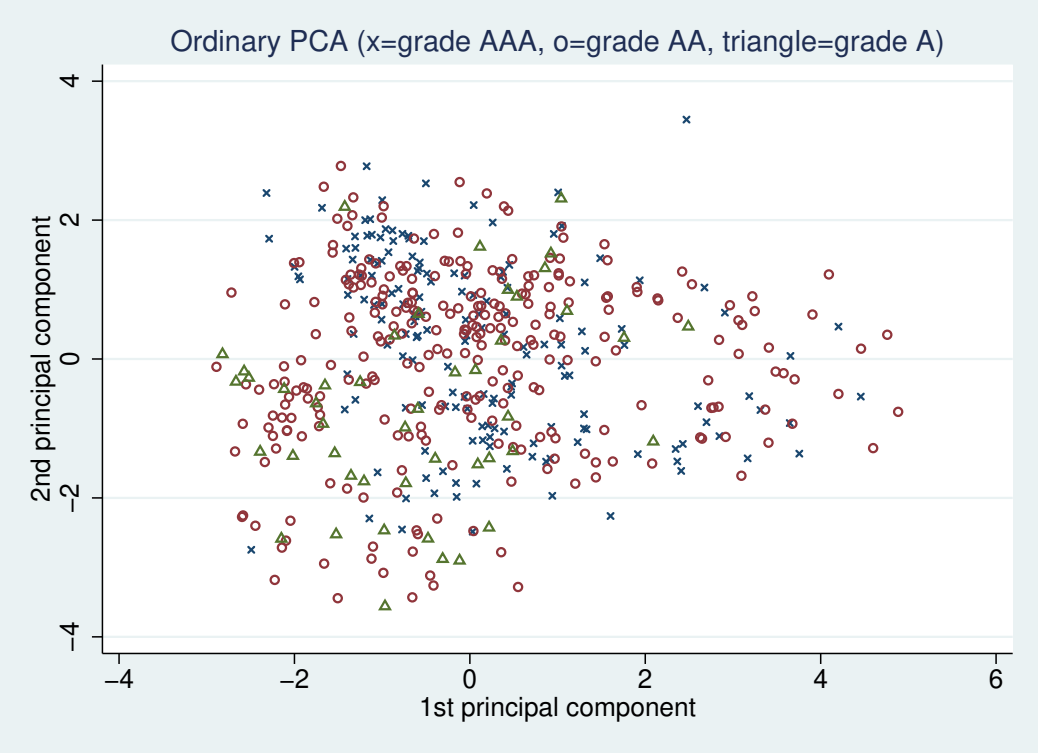
Descriptive analysis: Three of the variables are nominal categorical, and hence not suitable for multivariate analysis, namely animal and farm id's, as well as the breed. Three variables are binary, namely whether the animals were backgrounded (75% were, versus 25% weaned) and whether a hormone implant was used (for 26% of animals, versus 74% not), as well as the sex (64% steer versus 36% female). These variables can be used for multivariate analysis. Only the grade is ordinal categorical, and the intent was to use this variable for display purposes only. The 5 quantitative variables are on totally different scales, so analysis without standardizing these variables will make little sense. In other words, the dimension-reduction techniques should be applied to the correlation matrix. Except for the carcass weight, the quantitative variables all have a right-skewed distribution, but none of them show any obvious extreme outliers.

Principal component analysis: The eigenvalues decline slowly, and it takes five components to explain more than 80% of the variation. By the rule to focus mainly on components with eigenvalues greater than one, we would arrive at three components of main interest; together they explain about 65% of the variation. The Stata listing below shows the eigenvalues and eigenvectors (component loadings).

```
Principal components/correlation                  Number of obs    =        487
                                                  Number of comp.  =          8
                                                  Trace            =          8
    Rotation: (unrotated = principal)             Rho              =     1.0000
    --------------------------------------------------------------------------
       Component |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
           Comp1 |      2.36475      .616135             0.2956       0.2956
           Comp2 |      1.74861      .618665             0.2186       0.5142
           Comp3 |      1.12995      .228449             0.1412       0.6554
           Comp4 |      .901497      .281181             0.1127       0.7681
           Comp5 |      .620316     .0858525             0.0775       0.8456
           Comp6 |      .534464      .106948             0.0668       0.9124
           Comp7 |      .427516      .154612             0.0534       0.9659
           Comp8 |      .272904            .             0.0341       1.0000
    --------------------------------------------------------------------------
```

```
Principal components (eigenvectors)
-------------------------------------------------------------------------------------------
    Variable |    Comp1     Comp2     Comp3     Comp4     Comp5     Comp6     Comp7     Comp8
-------------+-----------------------------------------------------------------------------
         sex |  -0.1921    0.5408    0.3335    0.2182    0.0633   -0.4429    0.5107    0.2255
     bckgrnd |   0.3197    0.3905   -0.3465    0.1828   -0.5199    0.4770    0.2652    0.1537
     implant |  -0.3461    0.1085    0.6121    0.2026    0.0230    0.6330   -0.2161    0.0694
     backfat |   0.5109   -0.0879    0.2438    0.0026    0.4369    0.2533    0.4808   -0.4309
      ribeye |   0.4776    0.2788    0.1015   -0.3217    0.3599    0.0306   -0.3223    0.5883
       imfat |   0.2857   -0.1953   -0.0385    0.8711    0.1346   -0.1549   -0.2437    0.1357
        days |  -0.4073   -0.0665   -0.5111    0.0917    0.5644    0.2936    0.2695    0.2876
     carc_wt |  -0.0666    0.6444   -0.2481    0.0849    0.2610   -0.0194   -0.3940   -0.5365
-------------------------------------------------------------------------------------------
```

Several of the components have high loadings on multiple variables. The loading plot for the first two components show an almost circular distribution of the variables. High loadings are found for `sex` and `carc_wt` on component 2, for `implant` and `days` on component 3, and for `imfat` on component 4 (if included). Descriptive statistics for the scores for the different grades show grade A with different means for more or less all four two components, whereas grades AA and AAA are only separated by components 3 and 4. A scatter plot of scores with symbols corresponding to grades, shown below for the first two scores, does not very clearly show a separation between grades (i.e., A versus the others).



Choric correlations, and corresponding principal component analysis: We start by listing the two correlation matrices; the matrix below has the Pearson correlations below the diagonal and the -choric correlations above the diagonal (with correlations between quantitative variables omitted, because they are the same as the Pearson correlations).

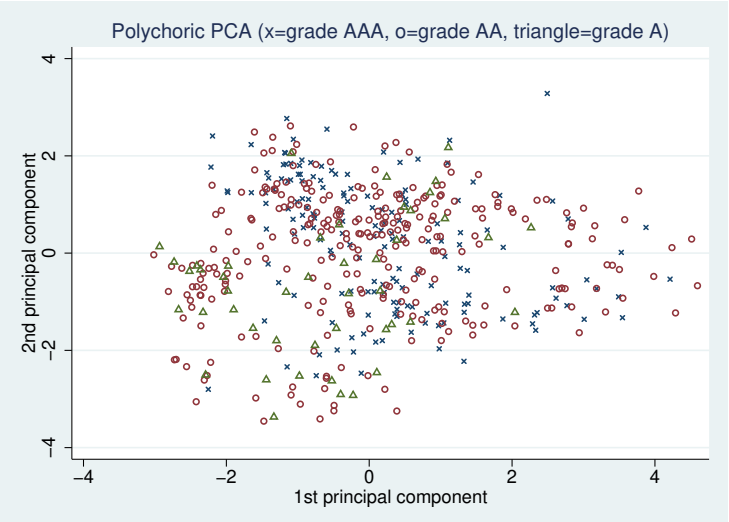
	sex	bckgrnd	implant	backfat	ribeye	imfat	days	carc_wt
sex	1.0000	0.1097	0.6247	-0.2269	-0.0070	-0.1996	-0.0290	0.5991
bckgrnd	0.0634	1.0000	-0.4322	0.3334	0.5572	0.1963	-0.2817	0.4602
implant	0.3386	-0.2615	1.0000	-0.4009	-0.4267	-0.2335	0.1052	0.0574
backfat	-0.1872	0.1913	-0.2264	1.0000				
ribeye	-0.0056	0.3386	-0.2702	0.5277	1.0000			
imfat	-0.1606	0.1364	-0.1638	0.3161	0.0534	1.0000		
days	-0.0232	-0.2026	0.0719	-0.4081	-0.4378	-0.1529	1.0000	
carc_wt	0.4586	0.3445	0.0442	-0.1974	0.2119	-0.1430	0.1404	1.0000

As expected, the -choric correlations are generally in the same direction as the Pearson correlations but numerically larger; in some cases, the difference is quite substantial (e.g. between `sex` and `implant`). PCA results for the -choric correlation matrix are shown on the next page.

```
Principal components/correlation                  Number of obs    =        487
...
    ------------------------------------------------------------------------------
       Component |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+----------------------------------------------------------------
           Comp1 |      2.77499      .781521             0.3469       0.3469
           Comp2 |      1.99347      .876001             0.2492       0.5961
           Comp3 |      1.11747      .218417             0.1397       0.7357
           Comp4 |       .89905      .382725             0.1124       0.8481
           Comp5 |      .516325      .193137             0.0645       0.9127
           Comp6 |      .323188     .0777155             0.0404       0.9531
           Comp7 |      .245473      .115435             0.0307       0.9837
           Comp8 |      .130038            .             0.0163       1.0000
    ------------------------------------------------------------------------------
```

```
Principal components (eigenvectors)
------------------------------------------------------------------------------------------------
    Variable |    Comp1     Comp2     Comp3     Comp4     Comp5     Comp6     Comp7     Comp8
-------------+----------------------------------------------------------------------------------
         sex |  -0.2269    0.5735   -0.3001    0.1614    0.1722   -0.0231   -0.0496   -0.6862
     bckgrnd |   0.3957    0.3896    0.2328    0.1280   -0.2300   -0.5201    0.5438    0.0436
     implant |  -0.4440    0.1722   -0.5033    0.1258    0.1753   -0.0100    0.4058    0.5555
     backfat |   0.4631   -0.0708   -0.2595   -0.0011    0.7052   -0.3782   -0.2471    0.1085
      ribeye |   0.4544    0.2627   -0.0516   -0.3070    0.1922    0.6974    0.3222    0.0210
       imfat |   0.2547   -0.1633   -0.0680    0.9026   -0.0450    0.2950    0.0115   -0.0007
        days |  -0.3284   -0.0832    0.6456    0.1344    0.5920    0.0536    0.2991   -0.0866
     carc_wt |  -0.0321    0.6181    0.3331    0.1214    0.0033    0.0975   -0.5317    0.4461
------------------------------------------------------------------------------------------------
```

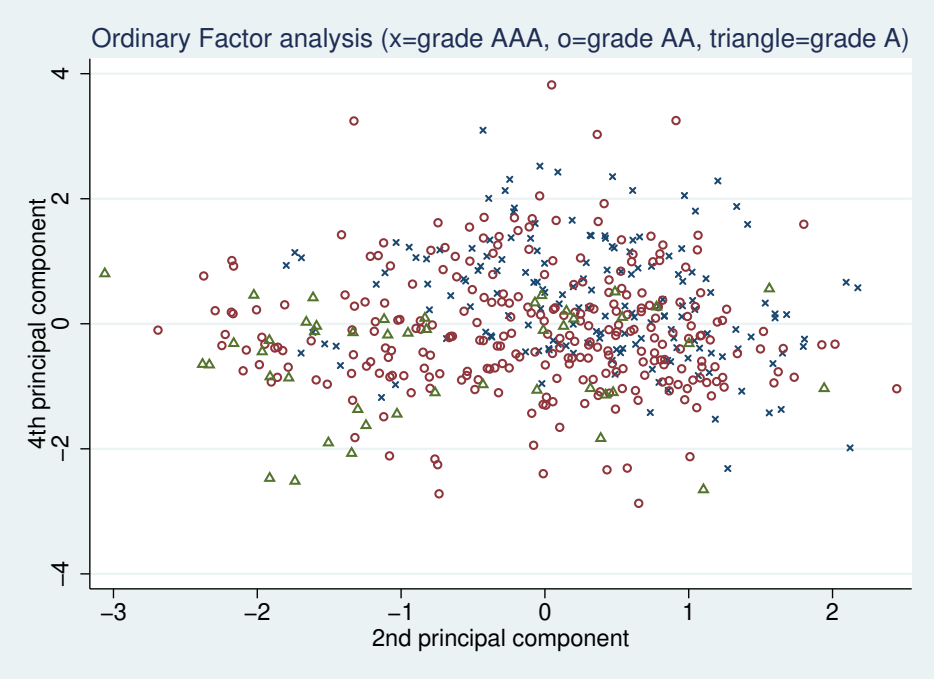
It is seen that the first eigenvalues are larger than for the Pearson correlations, leading to about 75% of the variation being explained by the first three components, and 85% when including the fourth component as well. The loadings show quite similar patterns as above, except for a sign-reversal of the third component (which does not make any major difference). Because the analysis was based on the computed correlation matrix (and not the variables themselves), a manual calculation of the scores is necessary (as shown in the do-file). The relation of the scores to the grades show similar patterns as above; we show again a scatter plot of the first two scores (below). The plot looks remarkably similar to the one previously shown; one needs to look carefully to detect the differences! Considering the strong similarity between the two plots, one could be tempted to conclude that use of the -choric correlations was hardly worth the trouble, despite that the first eigenvalues were somewhat higher (and therefore a higher proportion of the variation was explained).



Factor analysis: For simplicity, the solution here will focus on the analysis based on the Pearson correlation matrix (but the Stata do-file includes commands for analysis based on the -choric correlation matrix as well, in particular, for the generation of the scores). We extract four factors, thereby including one component with an eigenvalue less than 1; with only 3 factors, the uniqueness for `imfat` would be quite high (or the communality be low). We show only the factor loadings and uniqueness (1 minus the communality), because the eigenvalues are the same as for the PCA, followed by the results from a varimax rotation. We use the Stata default, i.e. no Kaiser normalization (contrary to the Minitab default, so results will differ slightly).

```
Factor loadings (pattern matrix) and unique variances
    ---------------------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3   Factor4 |   Uniqueness
    -------------+-------------------------------------+--------------
             sex |  -0.2955    0.7151    0.3545    0.2072 |      0.2327
         bckgrnd |   0.4916    0.5164   -0.3683    0.1736 |      0.3258
         implant |  -0.5323    0.1435    0.6507    0.1924 |      0.2357
         backfat |   0.7856   -0.1162    0.2591    0.0025 |      0.3022
          ribeye |   0.7345    0.3687    0.1079   -0.3054 |      0.2197
           imfat |   0.4394   -0.2583   -0.0409    0.8271 |      0.0545
            days |  -0.6263   -0.0880   -0.5433    0.0870 |      0.2973
         carc_wt |  -0.1025    0.8521   -0.2637    0.0807 |      0.1874
    ---------------------------------------------------------------------

. rotate, varimax

Factor analysis/correlation                       Number of obs    =        487
    Method: principal-component factors           Retained factors =          4
    Rotation: orthogonal varimax (Kaiser off)     Number of params =         26

    --------------------------------------------------------------------------
         Factor  |     Variance   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      1.94661      0.30486            0.2433       0.2433
        Factor2  |      1.64175      0.21708            0.2052       0.4485
        Factor3  |      1.42468      0.29292            0.1781       0.6266
        Factor4  |      1.13176            .            0.1415       0.7681
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(28) =  877.29 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    ---------------------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3   Factor4 |   Uniqueness
    -------------+-------------------------------------+--------------
             sex |   0.0143    0.5340    0.6881   -0.0919 |      0.2327
         bckgrnd |   0.2659    0.6735   -0.3089    0.2335 |      0.3258
         implant |  -0.1312   -0.1156    0.8533   -0.0750 |      0.2357
         backfat |   0.7433   -0.1409   -0.1900    0.2990 |      0.3022
          ribeye |   0.7933    0.2974   -0.2292   -0.0999 |      0.2197
           imfat |   0.0791   -0.0402   -0.0706    0.9657 |      0.0545
            days |  -0.8150    0.0879   -0.1414   -0.1038 |      0.2973
         carc_wt |  -0.0789    0.8786    0.1187   -0.1426 |      0.1874
    ---------------------------------------------------------------------
```

As we have seen before, the variance becomes more equally distributed on the four factors, and the factor loadings become more extreme for a few variables and get closer to zero for the others. The rotated third factor differs more substantially from the unrotated factor, with large changes for the `sex` and `carc_wt` variables. Except for `sex`, each variable only loads strongly on one of the factors. Perhaps not too surprisingly, the fourth factor is relatively unchanged, because it was already determined essentially by a single variable only. The grade differences in the scores however changed remarkably. Only scores 2 and 4 showed significant differences between grades. We therefore show instead the score plot for those two factors; the separation of grades is perhaps more clearly visible in the plot now.



Without going into details, it is perhaps interesting to note that the factor analysis based on the -choric correlation matrix does not at all end up with the same factors (and scores) after the varimax rotation, see the do-file for details. This illustrates one marked difference between PCA and factor analysis: the former is relatively robust to minor changes in the matrix (and not affected by the choice of components), whereas results of a factor analysis after rotation can change a lot between similar starting points.