Max Kramer

I affirm that I have adhered to the honor code on this assignment

Hello again, scientist! I'll do all my writing in italics, and problems for you will be in **bold.** *Comment your code, and* explain your ideas in plaintext. *As a general rule, I expect you to do at least as much writing as I do. Code should be part of your solution, but I expect variables to be clear and explanation to involve complete sentences. Cite your sources; if you work with someone in the class on a problem, that's an extremely important source. Don't work alone.*

Problem F.08: Interpret the grades.

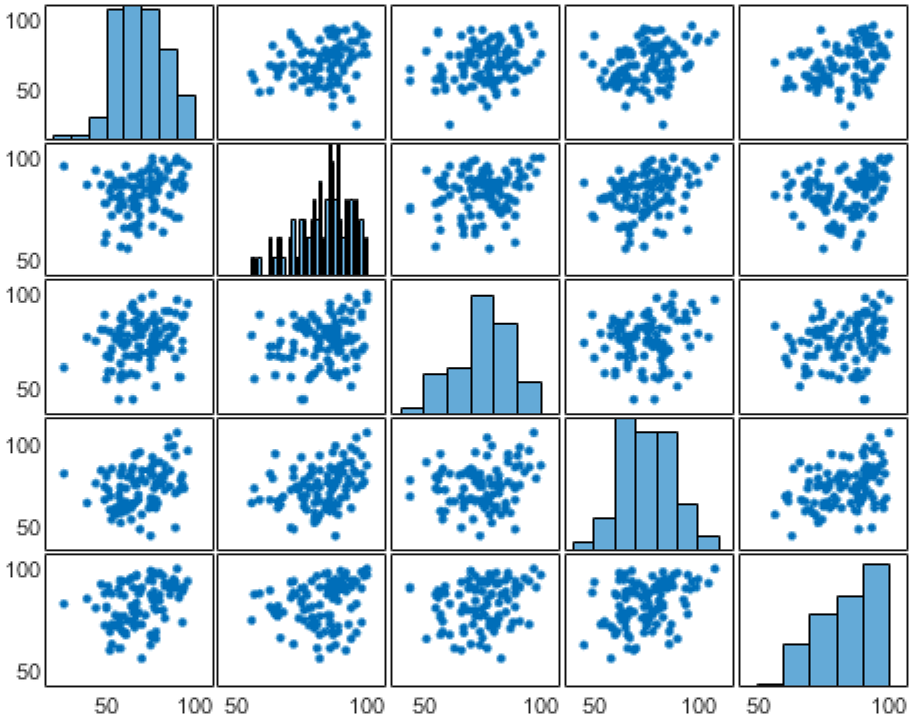
This problem assumes that you have mostly completed F.07.

```
grades = csvread('grades2017.csv');
figure;
plotmatrix(grades)

m = mean(grades)
```

```
m =

   69.0900   83.0000   75.2800   75.2000   83.4250
```



What you're looking at is the performance of the top 100 students in a mainstream Calculus I class on five exams: Exam 1, Exam 2, Exam 3, Exam 4, and Final. They covered limits, derivatives, optimization, integrals, and everything. While it's clear that each pair of exams is generally positively correlated, it's also clear that scores are all over the place.

```
[U,S,V] = svd(grades-m,'econ'); % this is the PCA
```

Real-world social science data is messy.

```
v=diag(S).^2; v=(v/sum(v))'*100


v =

   42.2139   18.6031   17.2256   11.7164   10.2410
```

Unlike the flower problem, we really do want to keep and interpret multiple principal components this time; the first principal component explains less than half of the variance. There seems to be a large drop in explanatory power after the third principal component, so let's try to do a three-dimensional analysis.

```
V(:,1:3)


ans =

    0.6165    0.5885   -0.4786
    0.3024   -0.2404    0.1795
    0.3474    0.2946    0.8533
    0.5026   -0.7073   -0.1023
    0.3939   -0.0938   -0.0107
```

The first principal component is positively correlated with all five exams; not surprisingly, a one-dimensional interpretation of the data set is that some students did well all semester and other students didn't. The second and third are more interesting. The second column of V has a large positive loading for exam 1 and a very large negative loading for exam 4, and the third column has a negative loading for exam 1 and a large positive loading for exam 3. Principal components are orthogonal bases, so they're only determined up to a plus or minus sign; I think it makes sense to flip the sign of the second component. That way positive factor scores are all "good," though for different reasons. A positive first score means that a student did well overall, while positive second and third scores reflect improvement over the course of the semester. (We could add the second and third scores to create a single "overall improvement" score, but that would no longer be a PCA because a linear combination of eigenvectors is almost never an eigenvector. More on this in the next problem.)

```
V(:,2)=-V(:,2); U(:,2)=-U(:,2);
V(:,1:3)
```

```
grades(60,:) % I chose this student pretty much at random


ans =

    0.6165   -0.5885   -0.4786
    0.3024    0.2404    0.1795
    0.3474   -0.2946    0.8533
    0.5026    0.7073   -0.1023
    0.3939    0.0938   -0.0107


ans =

   77.0000   79.0000   89.0000   74.0000   81.5000
```

At a glance, it looks to me like Student 60 did pretty well overall but dropped off towards the end of the semester. Let's see what our model thinks.

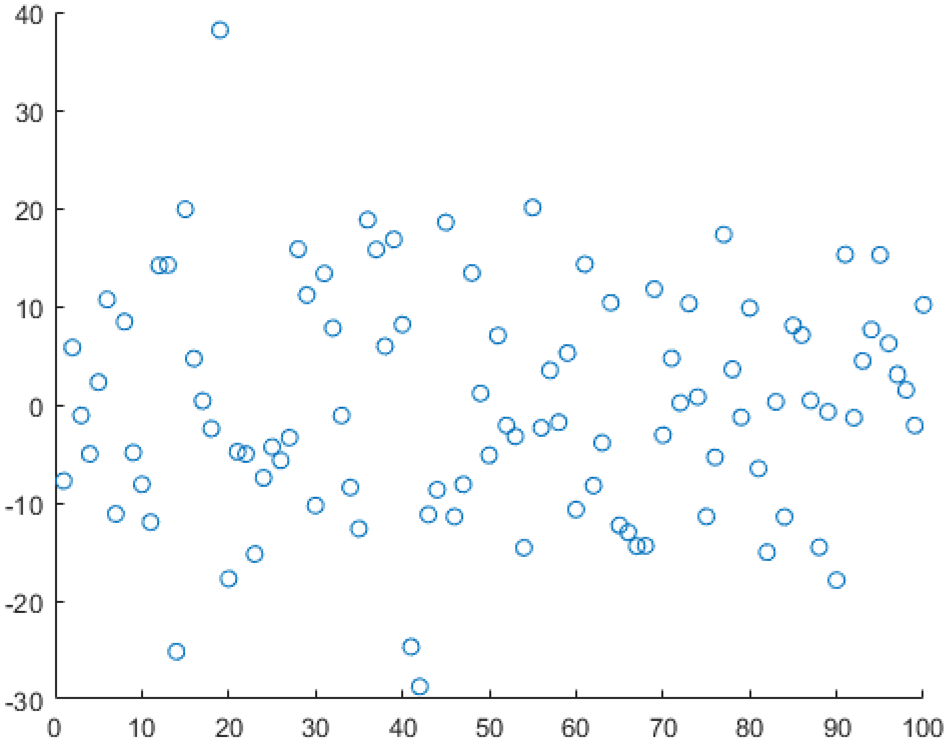
```
rot = (grades-m)*V; % this is the rotated data
rot(60,:)
```

```
ans =

    7.0720  -10.6881    7.3470   -0.8404    7.2814
```

Looks good to me. Their score on the first principal component is positive, the second is negative, and the third is positive. Let's take a deeper look at the data by plotting the factor scores for the second component. (If you run this section instead of publishing it, you can click on individual data points.)

```
figure;
scatter(1:100,rot(:,2))
```



Find the students with the largest negative, and largest positive, second component scores. Show me their raw grades. *Both of these students have a low-C average, but to me it feels like there's a huge qualitative difference that overall average alone doesn't see.*

```
largest_neg = min(rot(:,2)) % student 42
largest_pos = max(rot(:,2)) % student 19

grades(42,:)
grades(19,:)
```

```
largest_neg =

  -28.7470


largest_pos =

   38.1671


ans =

   89.0000   71.0000   66.0000   50.0000   93.5000


ans =

    26    96    61    83    83
```

Okay, now let's use this data to make predictions. Check out the last principal component.

```
P=V(:,5)
```

```
P =

    0.1262
   -0.5923
    0.2464
    0.4859
   -0.5801
```

*By definition, this vector is orthogonal to the first four principal components. That means that you can describe the first four components using the normal equation P'*x = 0, which lets you solve for a missing exam score.* ***Write down a "master equation" that predicts the final exam score as a linear combination of the first four.*** *The coefficients might seem weird. If you do it right, it should say that the coefficient on the second exam will be quite negative. But this is still the best that a linear analysis can do, and that makes sense if you think about it -- the second exam was on derivatives and had the highest average of the four midterms, so it's not surprising that students who do well on the comprehensive final have a disproportional mastery of the rest of the content in Calc 1.*

```
P'
```

```
ans =

    0.1262   -0.5923    0.2464    0.4859   -0.5801
```

The master equation is formed by multiplying the components of P' by the exam scores. The master equation is (0.1262*e1)-(0.5923*e2)+(0.2464*e3)+(0.4859*e4)-(0.5801*e5) = 0. This can be rewritten as ((0.1262*e1)-(0.5923*e2)+(0.2464*e3)+(0.4859*e4))/(0.5801) = e5.

One of my colleagues at Maryland offered his students the following "deal": if you got a B average on the midterms, you could take a B on the final without sitting for it. He was lazy and this is an absolutely terrible deal. ***Show that if you earn an 85 on each of Exams 1-4, then your predicted score on the final is actually a 97.*** *Once again, if you're getting weird and broken numbers, remember that PCA only applies to mean-centered data.*

```
scores = [85;85;85;85];
scores_mc = scores' - m(1:4);


Es = P(1:4)';

deviation = (Es * scores_mc')/(-P(5)) % master equation finds
 deviation on e5 from mean

m(5) + deviation
```

```
deviation =

   13.7554


ans =

   97.1804
```

The first step was to mean center the data. The four exam scores had the means on exams 1-4 subtracted and saved as scores_mc. Then the scores and values from P' (called Es) were input into the master equation found above and solved for a deviation from e5. That deviation added to the mean on the final exam resulted in a predicted exam score of 97.1804%.

What I just showed you was an ethical use of PCA. Here's an unethical one. ***Predict the final exam score of a student who earned midterm scores of 53, 86, 59, 60.*** *(It's not a passing grade.)*

```
scores2 = [53;86;59;60];
scores_mc2 = scores2' - m(1:4);


deviation2 = (Es * scores_mc2')/(-P(5))

m(5) + deviation2
```

```
deviation2 =

  -26.2094


ans =

   57.2156
```

The procedure was the exact same as the previous question with a different set of scores. This time, the deviation applied to the mean of the final exam resulted in a predicted exam score of 57.2156%.

Look at the grades of Student 16. *Real-world data is messy, and trying to use a classifier like this to predict the future in specific cases requires thought and care. I ran the numbers: the "master equation" you came up with correctly predicts a student's score on the final within one letter grade 50% of the time. (But this is still much better than just averaging their previous grades.)*

```
grades(16,:)
```

```
ans =

    53    86    59    60    88
```

The actual final exam score for student 16 was 88. This is approximately 31% better than the predicted result from the analysis.

Published with MATLAB® R2019b