Homework 9 Least squares and PCA answers

Least squares

1

(a) Show that the best least-squares fit to a set of measurements $y_1, \ldots, y_m$ by a *horizontal line* (a constant function $y = C$) is their average

$$C = \frac{y_1 + \cdots + y_m}{m}.$$

(b) Show that the slope of the line that passes through the origin in $\Re^2$ and comes closest in the least squares sense to passing through the points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is given by $m = \sum_i x_i y_i / \sum_i x_i^2$.

(a) The sum-of-squares error for the horizontal line $y=c$ is

$$\sum_{j=1}^{m} (y_j - c)^2 = m \cdot c^2 - 2c \sum_j y_j + \sum_j y_j^2$$

This is a quadratic function where the coefficient of c is >0, so the minimum occurs where the derivative with respect to c is 0:

$$0 = 2mc - 2\sum_j y_j$$

$$\Rightarrow c = \frac{1}{m} \sum_j y_j \text{, as claimed } \checkmark$$

(b) The sum-of-squares error for the line $y = mx$ is

$$\sum_{j=1}^{n} (y_j - m x_j)^2$$

Just as above, setting the derivative with respect to the free parameter m to zero gives:

$$0 = \sum_j 2(y_j - m x_j) \cdot (-x_j)$$

$$\Rightarrow 0 = \sum_j (-x_j y_j + x_j^2 m)$$

$$\Rightarrow m = \frac{1}{\sum_j x_j^2} \sum_j x_j y_j \text{, as claimed } \checkmark$$

- Both these problems can also be solved more geometrically. In (a), we want to find the C that minimizes

$$\left\| \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} C - \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \right\|.$$

By the theory of least-squares, the best C is given by

By the theory of least-squares, the best c is given by

$$\underbrace{\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}^{+}}_{\text{pseudoinverse}}\begin{pmatrix}y_1\\y_2\\\vdots\\y_m\end{pmatrix} = \frac{1}{m}(1\ 1\ \cdots\ 1)\begin{pmatrix}y_1\\y_2\\\vdots\\y_m\end{pmatrix} = \frac{1}{m}\sum_j y_j,$$

since it is easy to see (e.g., by symmetry) that the pseudoinverse of $\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}$ is $\frac{1}{m}(1\ 1\ \cdots\ 1)$.

Similarly, in b, we want to minimize

$$\left\| \underbrace{\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}}_{A} \cdot m - \begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix} \right\|$$

An easy way of finding the best m is to solve the normal-form equations

$$A^TAm = A^Ty.$$

$$\left(\sum_j x_j^2\right)\cdot m \quad \sum_j x_j y_j \qquad \Rightarrow m = \frac{\sum_j x_j y_j}{\sum_j x_j^2} \ \checkmark$$

(2)

Find the best straight-line fit (least squares) to the measurements

$$b=4 \text{ at } t=-2, \qquad b=3 \text{ at } t=-1,$$
$$b=1 \text{ at } t=0, \qquad b=0 \text{ at } t=2.$$

Then find the projection of $b=(4,3,1,0)$ onto the column space of

$$A=\begin{bmatrix}1 & -2\\1 & -1\\1 & 0\\1 & 2\end{bmatrix}.$$

Answer: The x that minimizes $\|Ax-b\|$ is given by

$$x = \underbrace{A^+}_{\text{pseudoinverse}}b = (A^TA)^{-1}A^Tb$$

Here it is in Matlab, computed both ways:

```
>> format rat;
A = [1 -2; 1 -1; 1 0; 1 2];
b = [4; 3; 1; 0];
pinv(A) * b          % solving for the least-squares fit using the pseudoinverse...
(A' * A) \ (A' * b)  % is equivalent to solving the normal equations A'Ax = A'b

ans =

      61/35
     -36/35


ans =

      61/35
     -36/35
```

To find the projection of b onto the column space of A, recall that the best least-squares x (found above) solves

$$Ax = P_{R(A)}b$$

So to compute $P_{R(A)}b$, just multiply x by A!

```
>> A * pinv(A) * b

ans =

      19/5
      97/35
      61/35
     -11/35
```

We have also shown in class that $AA^+ = P_{R(A)}$, so that's another way of seeing this.

Here are the same calculations in Mathematica. (Unlike Matlab, Mathematica uses exact arithmetic, so we can be confident that the answers are exact, not just close to the above fractions.)

```
A = {{1, -2}, {1, -1}, {1, 0}, {1, 2}};
b = {4, 3, 1, 0};
x = LinearSolve[Transpose[A].A, Transpose[A].b]
```

$$\left\{\frac{61}{35}, -\frac{36}{35}\right\}$$

```
A.x
```

$$\left\{\frac{19}{5}, \frac{97}{35}, \frac{61}{35}, -\frac{11}{35}\right\}$$

(3) Find the best least-squares error **parabola** to the four points $(0,0), (1,8), (3,8), (4,20)$. (Your answer should minimize the summed squared error in the y-coordinates.) What is the R^2 value of your fit?

Answer:

```
>> X = [
1 0 0;
1 1 1;
1 3 9;
1 4 16
];
b = [0 8 8 20]';
```

each row of X has the form $(x^0=1, x, x^2)$ for one of the data points

```
1 3 9;
1 4 16
];
b = [0 8 8 20]';
p = pinv(X)*b
```

```
p =

     2
   4/3
   2/3
```

$\Rightarrow$ best fit polynomial is

$$p(x) = 2 + \frac{4}{3}x + \frac{2}{3}x^2$$

$$R^2 = 1 - \frac{\sum_{j=1}^{4}(p(x_j)-y_j)^2}{\sum_{j=1}^{4}(y_j-\bar{y})^2} \qquad \bar{y} = \frac{1}{4}\sum_{j=1}^{4} y_j = \frac{36}{4} = 9$$

```
>> errorsfromp = X*p-b;
errorsfrommean = b - 9;
R2 = 1 - norm(errorsfromp)^2 / norm(errorsfrommean)^2
```

```
R2 =

     41/51
```

Here it is in Mathematica:

```
X = {{1, 0, 0}, {1, 1, 1}, {1, 3, 9}, {1, 4, 16}};
b = {0, 8, 8, 20};
p = PseudoInverse[X].b
```

$$1 - \frac{\text{Norm}[b - X.p]^2}{\text{Norm}\left[b - \frac{1}{\text{Length}[b]}\,\text{Plus @@ } b\right]^2}$$

$p = \left\{2, \frac{4}{3}, \frac{2}{3}\right\}$

$R^2 = \frac{41}{51}$

Economics

(4) An economist hypothesizes that the change (in dollars) in the price of a loaf of bread is primarily a linear combination of the change in the price of a bushel of wheat and the change in the minimum wage. That is, if B is the change in bread prices, W is the change in wheat prices, and M is the change in the minimum wage, then $B = \alpha W + \beta M$. Suppose that for three consecutive years the change in bread prices, wheat prices, and the minimum wage are as shown below.

	Year 1	Year 2	Year 3
B	+\$1	+\$1	+\$1
W	+\$1	+\$2	0\$
M	+\$1	0\$	−\$1

M	+\$1	0\$	−\$1

Use the theory of least squares to estimate the change in the price of bread in Year 4 if wheat prices and the minimum wage each fall by \$1.

Answer:

```
X = [ 1 1; 2 0; 0 -1 ];
b = [1 ; 1 ; 1];
x = pinv(X) * b

x =

    0.6667
   -0.3333

>> % this is the predicted change of price in year 4 if wheat prices and the minimum wage each fall by $1
x' * [-1; -1]

ans =

   -0.3333
```

This is also easy to do by hand. The normal equations are

$$X^T X \vec{x} = X^T \vec{b} \qquad X = \begin{pmatrix} 1 & 1 \\ 2 & 0 \\ 0 & -1 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix} \vec{x} = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$$

$$\Rightarrow \vec{x} = \frac{1}{3}\begin{pmatrix} 2 \\ -1 \end{pmatrix} \Rightarrow \boxed{\alpha = \frac{2}{3},\ \beta = -\frac{1}{3}}$$

$$\Rightarrow \alpha(-1) + \beta(-1) = \boxed{-\frac{1}{3}} \checkmark$$

(5)

Okuns "law", in economics, states that the annual change in gross domestic product (GDP) should relate to the annual change in the unemployment rate via an equation

$$\underbrace{\Delta G}_{\text{change in GDP}} = k - c \cdot \underbrace{\Delta U}_{\text{change in unemployment rate}}$$

I have uploaded two datasets, giving US GDP growth rates and unemployment rates. You can read them into Matlab using the csvread() function.

Find the least-squares best values for k and c.

Plot the data and your best-fit line.

Explain briefly what this means (one or two sentences).

(Note: The data gives the unemployment rates. You'll need to compute the changes in unemployment rates.)

Answer: Here is the transcript of a Matlab session:

```
>> gdp = csvread('HW 8 GDP change.csv');
unemp = csvread('HW 8 Annual unemployment rates.csv');

'Notice that the data sets start in different years, GDP in 1930 and unemployment rates in 1947';
gdp(1,:)
unemp(1,:)

ans =
```

1930

```
   1.9300e+03  -8.5000e+00


ans =
```

1947

```
   1.9470e+03   3.9000e+00

>> 'Cut start of the data to start GDP in 1948';
gdp = gdp(19:length(gdp),:);
gdp(1,:)
gdp = gdp(:,2);

ans =
```

1948

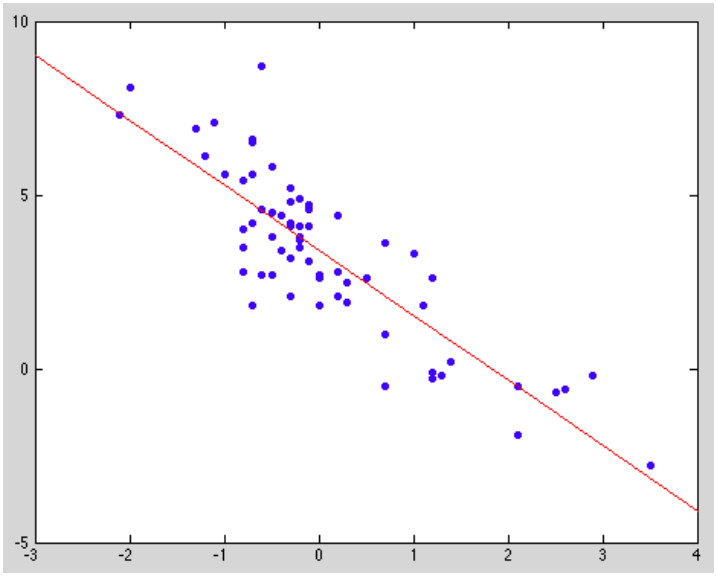
```
   1.9480e+03   4.1000e+00



>> m = length(gdp)
X = [ones(m,1), unempchanges];
b = gdp;

fit = pinv(X)*b

m =

    65

fit =

   3.3979e+00
  -1.8729e+00
```

k

-c

Now let's plot it:

```
>> k = fit(1);
c = -fit(2);
plot(unempchanges, gdp, '.') % '.' => scatterplot
hold on
x = -3:4;
y = k - c .* x;
plot(x, y, 'r')  % r => red line
```



If GDP grows faster than its mean growth rate k, then the unemployment rate falls. When GDP grows more slowly, unemployment rises.

Least squares and PCA in astronomy

(6)

Henrietta Leavitt discovered thousands of "variable stars" (stars with varying brightness) in the Magellanic Clouds (dwarf galaxies orbiting our Milky Way galaxy). Among them were 25 stars with regular periods now known as "Cepheid variable stars." https://en.wikipedia.org/wiki/Henrietta_Swan_Leavitt

In 1912, she published a relationship between the period and the luminosity of Cepheids. (Since they were in the same cluster, the stars she studied were all roughly at the same distance from the Earth, so the measured luminosities were comparable.)

Cepheids became the first "standard candle" in astronomy, allowing astronomers to calculate distances to other galaxies. https://en.wikipedia.org/wiki/Cosmic_distance_ladder

In fact, the Cepheids provided strong evidence that there **were** other galaxies. Using a telescope on Mt Wilson (here in LA!), Edwin Hubble discovered Cepheids, measured their brightness, and extrapolated their distances---far outside the Milky Way. https://timesmachine.nytimes.com/timesmachine/1924/11/23/issue.html p.6

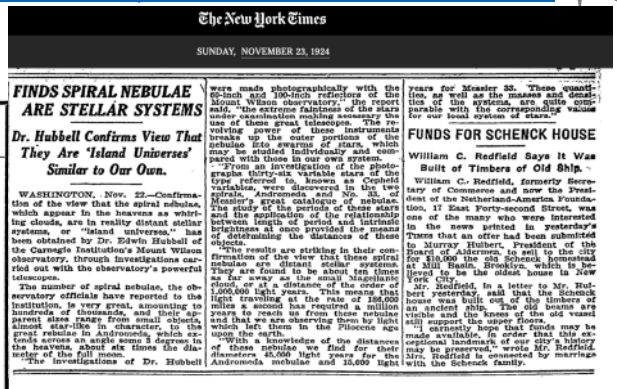
Here is her data:

TABLE I.

PERIODS OF VARIABLE STARS IN THE SMALL MAGELLANIC CLOUD.

H.	Max.	Min.	Epoch.	Period.	Res. *M.*	Res. *m.*	H.	Max.	Min.	Epoch.	Period.	Res. *M.*	Res. *m.*
			d.	d.						d.	d.		
1505	14.8	16.1	0.02	1.25336	−0.6	−0.5	1400	14.1	14.8	4.0	6.650	+0.2	−0.3
1436	14.8	16.4	0.02	1.6637	−0.3	+0.1	*1355*	14.0	14.8	4.8	7.483	+0.2	−0.2
1446	14.8	16.4	1.38	1.7620	−0.3	+0.1	1374	13.9	15.2	6.0	8.397	+0.2	−0.3
1506	15.1	16.3	1.08	1.87502	+0.1	+0.1	818	13.6	14.7	4.0	10.336	0.0	0.0
1413	14.7	15.6	0.35	2.17352	−0.2	−0.5	*1610*	13.4	14.6	11.0	11.645	0.0	0.0
1460	14.4	15.7	0.00	2.913	−0.3	−0.1	*1365*	13.8	14.8	9.6	12.417	+0.4	+0.2
1422	14.7	15.9	0.6	3.501	+0.2	+0.2	*1351*	13.4	14.4	4.0	13.08	+0.1	−0.1
842	14.6	16.1	2.61	4.2897	+0.3	+0.6	827	13.4	14.3	11.6	13.47	+0.1	−0.2
1425	14.3	15.3	2.8	4.547	0.0	−0.1	*822*	13.0	14.6	13.0	16.75	−0.1	+0.3
1742	14.3	15.5	0.95	4.9866	+0.1	+0.2	823	12.2	14.1	2.9	31.94	−0.3	+0.4
1646	14.4	15.4	4.30	5.311	+0.3	+0.1	824	11.4	12.8	4.	65.8	−0.4	−0.2
1649	14.3	15.2	5.05	5.323	+0.2	−0.1	821	11.2	12.1	97.	127.0	−0.1	−0.4
1492	13.8	14.8	0.6	6.2926	−0.2	−0.4							

http://adsbit.harvard.edu/cgi-bin/nph-iarticle_query?1912HarCi.173....1L&defaultprint=YES&filetype=.pdf



The New York Times

SUNDAY, NOVEMBER 23, 1924

FINDS SPIRAL NEBULAE ARE STELLAR SYSTEMS

Dr. Hubbell Confirms View That They Are 'Island Universes' Similar to Our Own.

WASHINGTON, Nov. 22.—Confirmation of the view that the spiral nebulae, which appear in the heavens as whirling clouds, are in reality distant stellar systems, or "island universes," has been obtained by Dr. Edwin Hubbell of the Carnegie Institution's Mount Wilson observatory, through investigations carried out with the observatory's powerful telescopes.

The number of spiral nebulae, the observatory officials have reported to the Institution, is very great, amounting to hundreds of thousands, and their apparent sizes ranges from small objects, almost star-like in character, to the great nebulae in Andromeda, which extends across an angle some 3 degrees in the heavens, about six times the diameter of the full moon.

"The investigations of Dr. Hubbell were made photographically with the 60-inch and 100-inch reflectors of the Mount Wilson observatory," the report said, "the extreme faintness of the stars under examination making necessary the use of these great telescopes. The revolving power of these instruments breaks up the outer portions of the nebulae into swarms of stars, which may be studied individually and compared with those in our own system.

"From an investigation of the photographs thirty-six variable stars of the type referred to, known as Cepheid variables, were discovered in the two spirals, Andromeda and No. 33, of Messier's great catalogue of nebulae. The study of the periods of these stars and the application of the relationship between length of period and intrinsic brightness at once provided the means of determining the distances of these objects.

"The results are striking in their confirmation of the view that these spiral nebulae are distant stellar systems. They are found to be about ten times as far away as the small Magellanic cloud, or at a distance of the order of 1,000,000 light years. This means that light traveling at the rate of 186,000 miles a second has required a million years to reach us from these nebulae and that we are observing them by light which left them in the Pliocene age upon the earth.

"With a knowledge of the distances of these nebulae we find for their diameters 45,000 light years for the Andromeda nebulae and 15,000 light years for Messier 33. These quantities, as well as the masses and densities of the systems, are quite comparable with the corresponding values for our local system of stars."

FUNDS FOR SCHENCK HOUSE

William C. Redfield Says It Was Built of Timbers of Old Ship.

William C. Redfield, formerly Secretary of Commerce and now the President of the Netherland-America Foundation, 17 East Forty-second Street, was one of the many who were interested in the news printed in yesterday's Times that an offer had been submitted to Murray Hulbert, President of the Board of Aldermen, to sell to the city for $10,000 the old Schenck homestead at Mill Basin, Brooklyn, which is believed to be the oldest house in New York City.

Mr. Redfield, in a letter to Mr. Hulbert yesterday, said that the Schenck house was built out of the timbers of an ancient ship. The old beams are visible and the knees of the old vessel still support the upper floors.

"I earnestly hope that funds may be made available, in order that this exceptional landmark of our city's history may be preserved," wrote Mr. Redfield. Mrs. Redfield is connected by marriage with the Schenck family.

Find the best-fitting lines for maximum luminosity as a function of log(period), and for minimum luminosity as a function of log(period). Plot your results. Also give the R^2 values.

Find the best-fitting lines for maximum luminosity as a function of log(period), and for minimum luminosity as a function of log(period). Plot your results. Also give the R^2 values.

Answer:

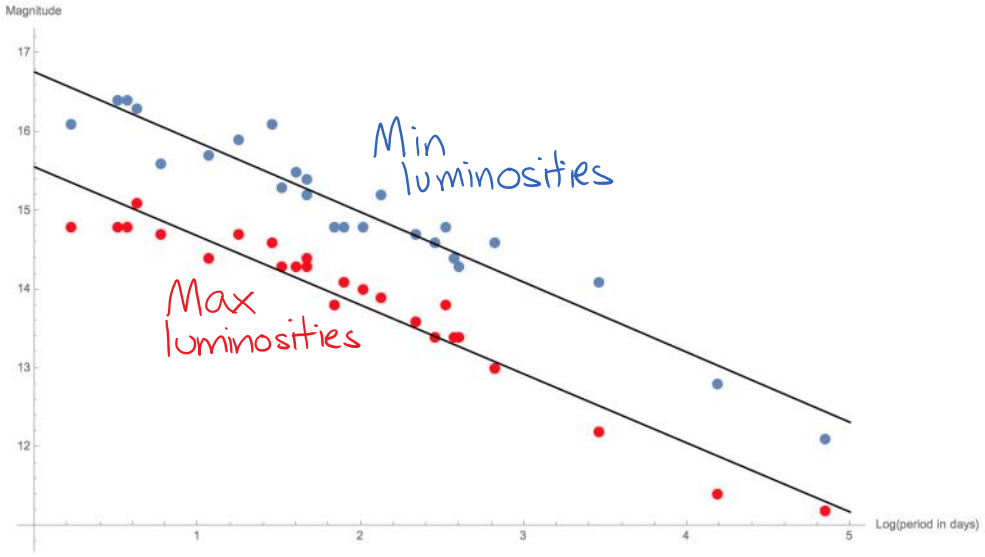
y-intercept slope

{15.56, -0.87626} ← max line

0.933845 ← R^2

{16.7663, -0.889748} ← min line

0.925526 ← R^2



https://en.wikipedia.org/wiki/Hubble%27s_law

(7) In 1929, Edwin Hubble famously showed that the universe is expanding. Specifically, he showed a roughly linear relationship between the distances of other galaxies and their velocities away from us.

Here is the data he used:

NGC #	Distance (x10^6 parsecs)	Radial velocity (km/sec)	Right ascension	Declination	Adjusted radial velocity
NA	0.032	170	NA	NA	170
NA	0.034	290	NA	NA	290
6822	0.214	-130	{19, 44, 57.8}	{-14, 48, 11}	60
598	0.263	-70	{1, 33, 51.}	{30, 39, 37}	15
221	0.275	-185	{0, 42, 41.9}	{40, 51, 57}	-30
224	0.275	-220	{0, 42, 44.3}	{41, 16, 9}	-65
5457	0.45	200	{14, 3, 12.5}	{54, 20, 53}	395
4736	0.5	290	{12, 50, 52.6}	{41, 7, 9}	405
5194	0.5	270	{13, 29, 52.4}	{47, 11, 41}	430
4449	0.63	200	{12, 28, 11.}	{44, 5, 33.4}	305
4214	0.8	300	{12, 15, 39.2}	{36, 19, 41}	370
3031	0.9	-30	{9, 55, 33.2}	{69, 3, 55}	90
3627	0.9	650	{11, 20, 15.1}	{12, 59, 22}	580
4826	0.9	150	{12, 56, 43.9}	{21, 41, 0}	205
5236	0.9	500	{13, 37, 0.8}	{-29, 51, 59}	425
1068	1	920	{2, 42, 40.8}	{0, 0, 48}	830
5055	1.1	450	{13, 15, 49.3}	{42, 1, 47}	585
7331	1.1	500	{22, 37, 4.3}	{34, 24, 59}	740
4258	1.4	500	{12, 18, 57.5}	{47, 18, 14}	610
4151	1.7	960	{12, 10, 32.7}	{39, 24, 20}	1035
4382	2	500	{12, 25, 24.2}	{18, 11, 27}	515
4472	2	850	NA	NA	850
4486	2	800	{12, 30, 49.4}	{12, 23, 28}	800
4649	2	1090	{12, 43, 40.2}	{11, 33, 9}	1100

The second column gives the distance to each galaxy, and the last column gives the velocity away from us. (Hubble actually started with the data in column 3, but I have adjusted these velocities for the

motion of our Sun.)

(a) Run linear regression of distance versus velocity to find the best-fitting line. Make sure your line goes through (0,0)!!

I put the relevant data into two columns of a matrix:

```
>> format long e
>> data = csvread('data.csv')

data =

   3.200000000000000e-02   1.700000000000000e+02
   3.400000000000000e-02   2.900000000000000e+02
   2.140000000000000e-01   6.000000000000000e+01
   2.630000000000000e-01   1.500000000000000e+01
   2.750000000000000e-01  -3.000000000000000e+01
   2.750000000000000e-01  -6.500000000000000e+01
   4.500000000000000e-01   3.950000000000000e+02
   5.000000000000000e-01   4.050000000000000e+02
   5.000000000000000e-01   4.300000000000000e+02
   6.300000000000000e-01   3.050000000000000e+02
   8.000000000000000e-01   3.700000000000000e+02
   9.000000000000000e-01   9.000000000000000e+01
   9.000000000000000e-01   5.800000000000000e+02
   9.000000000000000e-01   2.050000000000000e+02
   9.000000000000000e-01   4.250000000000000e+02
   1.000000000000000e+00   8.300000000000000e+02
   1.100000000000000e+00   5.850000000000000e+02
   1.100000000000000e+00   7.400000000000000e+02
   1.400000000000000e+00   6.100000000000000e+02
   1.700000000000000e+00   1.035000000000000e+03
   2.000000000000000e+00   5.150000000000000e+02
   2.000000000000000e+00   8.500000000000000e+02
   2.000000000000000e+00   8.000000000000000e+02
   2.000000000000000e+00   1.100000000000000e+03
```

Normally, with linear regression you add a column of 1's for the line's y-intercept. But since we want the line to go through 0 that's not needed:

```
>> slopeA = pinv(data(:,1)) * data(:,2)

slopeA =

   4.638002262702888e+02
```

← slope of best-fit line
the units are $\frac{\text{km/sec}}{10^6 \text{ parsecs}}$.

(b) Now run linear regression of velocity versus distance. Why does this give a different answer than part (a)?

```
>> slopeB = 1 / ( pinv(data(:,2)) * data(:,1) )

slopeB =

   5.320349501400984e+02
```

← the new best-fit line.
I have taken the inverse to compare to part (a): $\frac{\text{velocity}}{\text{distance}}$.

In part (a), we tried to minimize errors only in the velocity,

while here we are trying to minimize errors only in distance, so of course they are different.

© Now use PCA to find the best-fitting line.
Plot all three lines, and the data, on one labeled graph.
Why is PCA more appropriate for analyzing this data than either linear regression?

PCA is more appropriate because there are errors in both components of the data, distance and velocity.

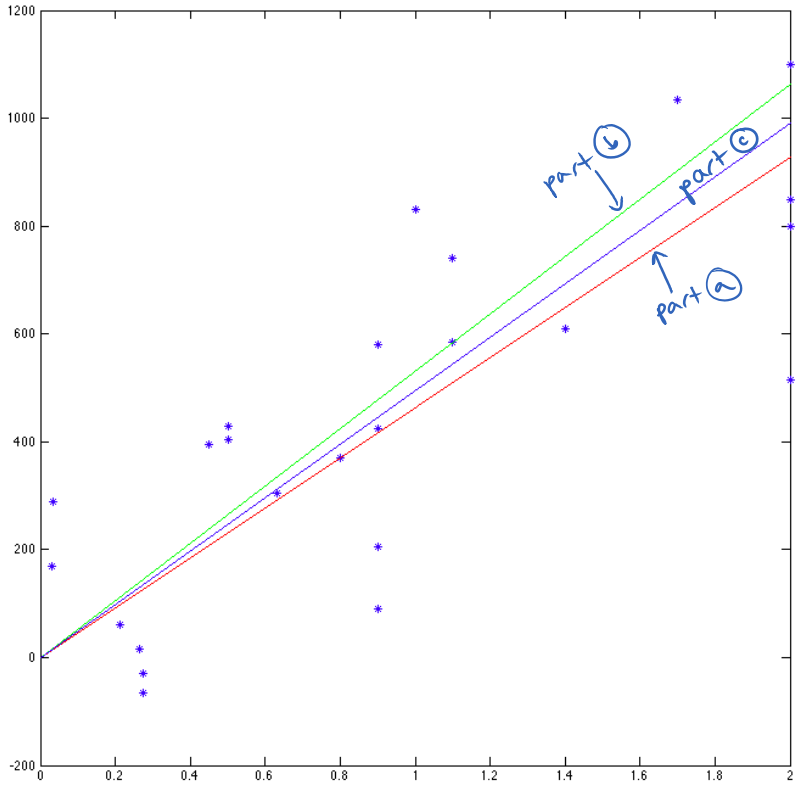
Normally in PCA you center each component of the data by subtracting off its mean. Then to make errors in the different components comparable, you divide each component by its observed standard deviation.

In this case, though, we want a line that goes through 0. So we only divide by the observed stddev, and do not subtract off the mean.

```
>> stddevs = sqrt(var(data))
normalizedData = data * diag(stddevs)^-1;

[U,S,V] = svd(normalizedData);
S(1:2,:)

projectedNormalizedData = S(1,1) * U(:,1) * V(:,1)';
% now rescale each component back, so it can be plotted with the original scales
projectedRescaledData = projectedNormalizedData * diag(stddevs);
slopeC = projectedRescaledData(1,2) / projectedRescaledData(1,1)

stddevs =

     6.454957523519018e-01     3.299810600625436e+02


ans =

     1.154000239668598e+01                         0
                         0     2.136307712571286e+00


slopeC =

     4.957362749141023e+02
```

keep just the first left and right singular vectors

```
>> % now let us plot the results:
plot(data(:,1), data(:,2), 'LineStyle', 'none', 'Marker', '*')
hold on
plot([0 2], slopeA * [0,2], 'r')
plot([0 2], slopeB * [0,2], 'g')
plot([0 2], slopeC * [0,2], 'b')
hold off
```



(2) Some of these data points are more precise than others. For example, they may have been collected by different telescopes.

In class, we saw how to get the k-dimensional subspace S that minimizes

$$\sum_{j=1}^{m} \| \vec{x}_j - P_S \vec{x}_j \|^2,$$

the sum of the squared distances from the data points to their projections on S. (The answer was to set S = Span{k largest e-value e-vectors of $\sum_j x_j x_j^T$.)

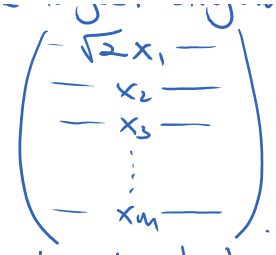
Extend this analysis to show how to get the k-dim! subspace S that minimizes

$$2 \cdot \|\vec{x}_1 - P_S \vec{x}_1\|^2 + \sum_{j=2}^{m} \| \vec{x}_j - P_S \vec{x}_j \|^2.$$

(This situation would arise if data point $\vec{x}_1$ was more precise than the others.)

If you multiply $\vec{x}_1$ by $\sqrt{2}$ then for any subspace S, $\|x_1 - P_S x_1\|^2$ will be multiplied by 2. Thus we can simply keep the k largest singular values of the matrix

$$\begin{pmatrix} — \sqrt{2} x_1 — \\ — x_2 — \end{pmatrix}$$



In general, to put a higher weight on data point j (because it is more reliable), just multiply $\vec{x}_j$ by a weight before taking the singular-value decomposition (SVD).

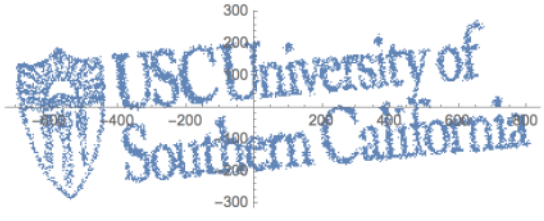
Principle component analysis

(8) The file "data.mat" contains roughly 15,000 data points in 32 dimensions. Use the singular-value decomposition (SVD) to project the data onto the best two-dimensional affine plane. Plot the projected data set. x-axis = principal component, y-axis = second component
(If you are using Matlab, the "scatter" function might be helpful for plotting.)

Answers:

Mathematica answers

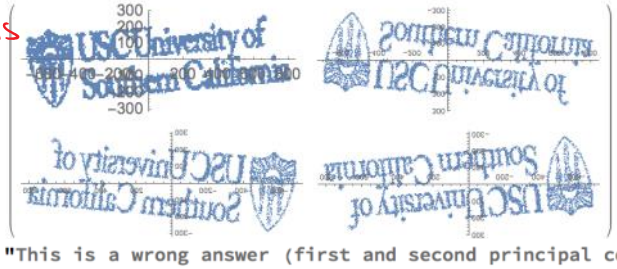
```
In[1114]:= data = embeddeddata;

In[1139]:= data = Import["data.mat"][[1]];
           data // Dimensions

Out[1140]= {15418, 32}

In[1141]:= means = Mean[data]
           data = (# - means) & /@ data;

Out[1141]= {-213.499, -153.992, 352.13, -27.4404, 71.5703, 102.55,
            64.4337, 75.5283, -294.045, -295.394, -33.3214, -15.5504,
            126.736, 88.3582, 6.5469, 27.9305, 44.4251, -73.6134, 48.0674,
            48.3388, -34.1894, 81.6966, 179.795, 26.4465, -5.4802, -12.2393,
            -53.8381, -134.369, 129.906, 20.5587, 198.028, 78.2909}

svd = SingularValueDecomposition[data, 2];   <- we only need the first 2 dimensions of the SVD
Dimensions /@ svd
projecteddata = svd[[1]].svd[[2]];

projecteddata // Dimensions
{projectedwidth, projectedheight} =
  (Max[#] - Min[#]) & /@ Transpose[projecteddata];
answer = ListPlot[projecteddata, AspectRatio → projectedheight / projectedwidth]

{{15418, 2}, {2, 2}, {32, 2}}

{15418, 2}
```



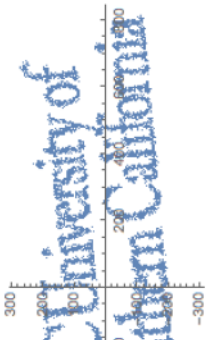
Matlab answer

```
load 'data.mat'
size(data)

data = data - mean(data);  % center it to find the best **affine** plane
[U,S,V] = svds(data,2);
projecteddata = U*S;
scatter(projecteddata(:,1),projecteddata(:,2))

"These are all correct answers (the +/- sign of a singular
   vector doesn't matter):";
{{answer, ImageReflect[answer]},
 {ImageReflect[answer, Left],
  ImageReflect[ImageReflect@answer, Left]}} // MatrixForm
```

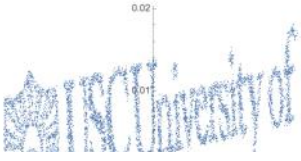
MatrixForm=

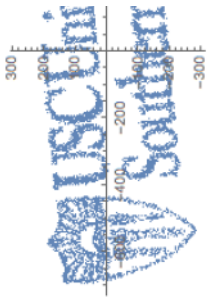


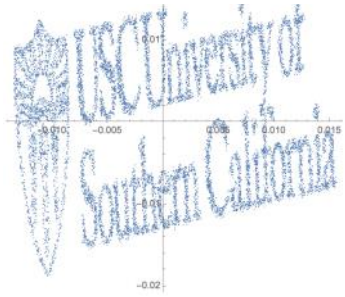
```
"This is a wrong answer (first and second principal components
   are switched):";
ImageRotate[answer, π/2]
```



```
"This is a wrong answer, because the x and y components aren't
   scaled properly:"
projecteddata = svd[[1]];
{projectedwidth, projectedheight} =
  (Max[#] - Min[#]) & /@ Transpose[projecteddata];
answer = ListPlot[projecteddata, AspectRatio → 1]

This is a wrong answer, because
   the x and y components aren't scaled properly:
```




300
200
100
-200
-300
400


USC University of Southern California
0.010
0.005
-0.005
-0.02