HW Solutions from Mar 12 #1-7

1. Suppose we have p data points in $\mathbb{R}^n$. Show that the variance of the data, projected to a standard basis vector $\boldsymbol{e}_i$, returns the usual variance for the data in that dimension. (I want you to look back at the computations we made for this in the text, "Projections and the Variance").

   SOLUTION: Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ be our p vectors in $\mathbb{R}^n$. If we project the data so e_i, so that the scalar projections are given by

   $$\left\{\mathbf{x}_1^T \mathrm{e}_i, \ldots, \mathbf{x}_p^T \mathrm{e}_i\right\}$$

   Taking the variance of this set of scalars, and noting that the projection of the mean is the mean of the projection,

   $$\text{Var} = \frac{1}{p-1}\sum_{k=1}^{p}(\mathbf{x}_k^T \mathrm{e}_i - \bar{\mathrm{x}_k}^T \mathrm{e}_i)^2 = \frac{1}{p-1}\sum_{k=1}^{p}((\mathbf{x}_k - \bar{\mathrm{x}}_k)^T \mathrm{e}_i)^2$$

   We observe that taking the dot product with e_i extracts the i^{th} dimension.

2. Suppose we have two o.n. vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$. Given our p points in $\mathbb{R}^n$, compute the covariance between the data projected to $\boldsymbol{u}$ and the data projected to $\boldsymbol{v}$, and (i) show that the result is

   $$\boldsymbol{u}^T C \boldsymbol{v}$$

   (ii) In the special case that $\boldsymbol{u}, \boldsymbol{v}$ are eigenvectors of the covariance matrix, how does this quantity simplify?

   SOLUTION: As discussed in the video, we're taking the covariance between the sets:

   $$\left\{\mathbf{x}_1^T \mathbf{u}, \ldots, \mathbf{x}_p^T \mathbf{u}\right\}, \quad \left\{\mathbf{x}_1^T \mathbf{v}, \ldots, \mathbf{x}_p^T \mathbf{v}\right\}$$

   For simplicity, **assume that both sets have been mean-subtracted**. Then the covariance is given by

   $$\frac{1}{p-1}\sum_{k=1}^{p}(\mathbf{x}_k^T \mathbf{u})(\mathbf{x}_k^T \mathbf{v}) = \frac{1}{p-1}\sum_{k=1}^{p}\mathbf{u}^T \mathbf{x}_k \mathbf{x}_k^T \mathbf{v} = \mathbf{u}^T\left(\frac{1}{p-1}\sum_{k=1}^{p}\mathbf{x}_k \mathbf{x}_k^T\right)\mathbf{v}$$

3. Suppose we have 4 points in $\mathbb{R}^3$ as organized in the matrix X (left and below), and let $\boldsymbol{\phi}_1 = (1/\sqrt{3})[1,1,1]^T$. Use a computer (Octave/Matlab, Python or R) to compute the three quantities given in the formula to the right and below. In your script, be sure you're actually computing the covariance matrix and each quantity separately.

   $$X = \begin{bmatrix} 1 & 2 & -1 & 3 \\ 0 & 0 & 1 & 1 \\ -1 & 1 & 2 & 1 \end{bmatrix}, \qquad \frac{1}{p-1}\sum_{k=1}^{p}\|\boldsymbol{x}_k\|^2 = \boldsymbol{\phi}_1^T C \boldsymbol{\phi}_1 + \frac{1}{p-1}\sum_{k=1}^{p}\|\boldsymbol{x}_{\text{err}}^{(k)}\|^2.$$

4. Using the data (and vector ϕ_1) in the previous exercise, compuationally verify our statements: The projection of the mean is the mean of the projection, and the variance of the data projected to ϕ_1 is $\phi_1^T C \phi_1$.

5. Verify numerically that the variance of the projected data to the first best basis vector (first one) is given by the first eigenvalue of the covariance matrix. (Careful- if you use the `eig` command, the eigenvalues are not ordered).

6. Continuing with the data from Problem 3, if we retained two of the basis vectors, how much variance (as a percentage) is "explained" by them? (This refers to the discussion in the text about how to compute the rank).

7. Load the clown data, we obtain a matrix X that is 200×320. Treat this as 320 vectors in $\mathbb{R}^{200}$.

   (a) Double center the data in X (call the result X_m).

   (b) Find the best two dimensional basis for the vectors in X_m, then project the data to two dimensions and plot the result.

   (Question to think about, you don't need to answer: Did you expect a pattern or not?)

   (c) Reconstruct the data back in $\mathbb{R}^{200}$, and show the result as an image (don't add the means back in).

```
%% Homework assigned March 12:

%% Problem 3:
phi=(1/sqrt(3))*[1;1;1];
X=[1 2 -1 3;0 0 1 1; -1 1 2 1];
mx=mean(X,2);
X=X-mx;
p=4;

% The three quantities:
C=cov(X');  %Be sure this is 3 x 3, not 4 x 4
LHS=(1/(p-1))*sum(sum(X.*X));
Middle=phi'*C*phi;
Temp=X-phi*phi'*X;
RHS=(1/(p-1))*sum(sum(Temp.*Temp));
fprintf('The three quantities are %f %f %f\n',LHS,Middle,RHS);

%% Problem 4
% Projection of the mean:
```

```
fprintf('Projection of the mean is %f\n',phi'*mx)

Px=phi'*(X+mx);  %Projection of the data (it was mean subtracted, so add it back)
fprintf('Mean of the projection is %f\n',mean(Px))

%% Problem 5:
% Variance of the projected data.  I checked the eigenvalues; the biggest
% is the third one.
[V,D]=eig((1/(p-1))*(X*X'));
fprintf('The variance of the data projected to the first evec is %f \n', var(V(:,3)'*X))
fprintf('The biggest eigenvalue is %f\n',D(3,3))

%% Problem 6
% SOLUTION:  Divide the eigenvalues by the sum of all.  Sum the last two in
% this case (the largest two eigenvalues):
dd=diag(D); dsum=sum(dd);  dd=dd/dsum;
fprintf('The percentage of the variance explained is %f\n',dd(2)+dd(3))

%% Problem 7:
load clown
m1x=mean(X,1); m2x=mean(X,2); g=mean(mean(X));
Xm=X-m1x-m2x+g;  %Double center the data
[U,S,V]=svd(Xm,0);
Coords=U(:,1:2)'*Xm;
plot(Coords(1,:),Coords(2,:),'.');
Recon=U(:,1:2)*Coords;
imagesc(Recon)
```