Solution to Project 4

In this project we are examining the metabolome of ovarian cancer cells. We compare the tumor cells (OCC) with ovarian cancer initiating cells (OCIC) - cancer cells which are hard to treat. The cells are examined over time after culturing in a medium. The data was provided by Mark Styczynski, a former graduate of this department, and currently a professor at Georgia Tech. The procedure is to first import the data, subtract the mean value, use SVD to determine the principal basis set, and finally determine the magnitude of the two principal components for each sample. The different principal components are hypothesized to correspond to different metabolic pathways.

Contents

Part 1: Importing the data

We have previously loaded the data portion of the excel file and saved the variable under the name "data" in the file metabcombdata.mat. We thus load it back in and subtract off the mean:

```
clear
clf
load('metabcombdata.mat')
meandata=mean(data')';
[n m]=size(data);
data=data-meandata*ones(1,m);
```

Part 2: Determining the principal components

We use SVD to determine the basis set, and then project the sample data back onto this basis set to determine the first two principal components for each sample. We can also determine the fraction of the information contained in these two principal components by examining the first two singular values:

```
[u s v]=svd(data);
pc1=data'*u(:,1);
pc2=data'*u(:,2);
frac1=s(1,1)/sum(diag(s))
frac2=s(2,2)/sum(diag(s))
```

```
frac1 =
    0.1147
frac2 =
    0.0757
```

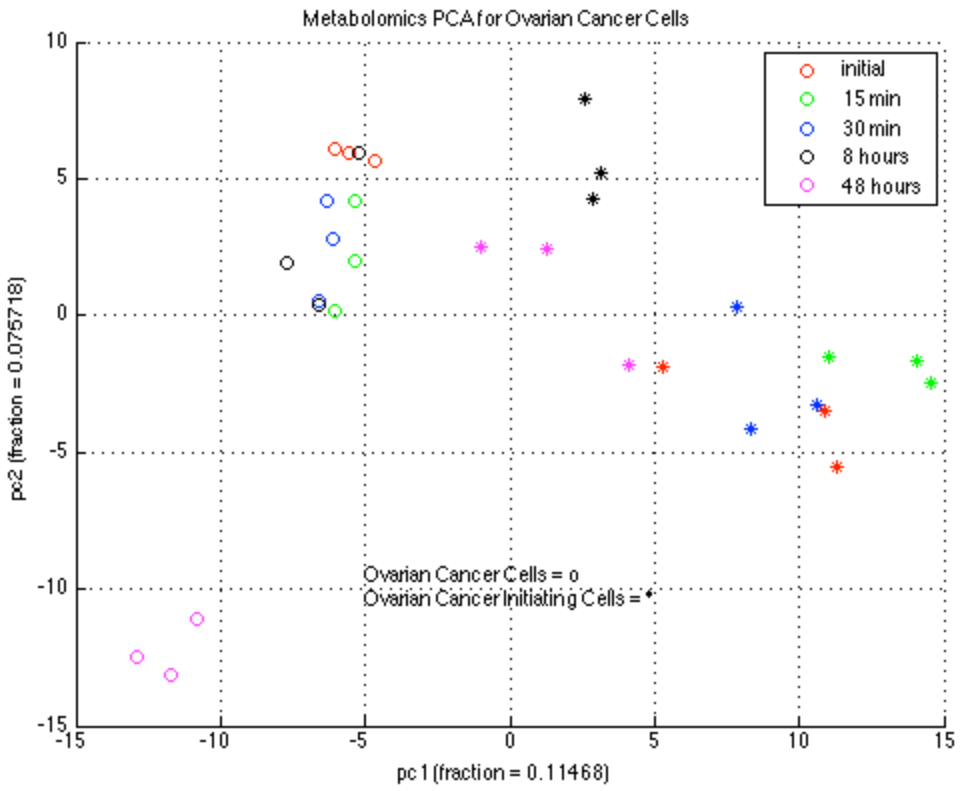
Part 3: Plotting up the principal components for each subset

In order to use a more compact way of plotting the first two principal components, we set up

a vector of five colors (one for each time), and two symbols (one for each type of cell). We then plot the results for each sample in groups with the appropriate color and symbol type:

```
colors='rgbkm';
colors=[colors,colors];
syms='o*';

for i=1:m/3
    figure(1)
    hold on
    b=(i-1)*3;
    plot(pc1(b+1:b+3),pc2(b+1:b+3),[syms(1+floor(i/5.01)),colors(i)])
end
hold off
legend('initial','15 min','30 min','8 hours','48 hours')
xlabel(['pc1 (fraction = ',num2str(frac1),')'])
ylabel(['pc2 (fraction = ',num2str(frac2),')'])
grid on
title('Metabolomics PCA for Ovarian Cancer Cells')
stuff=str2mat('Ovarian Cancer Cells = o','Ovarian Cancer Initiating Cells = *');
text(-5,-10,stuff)
```



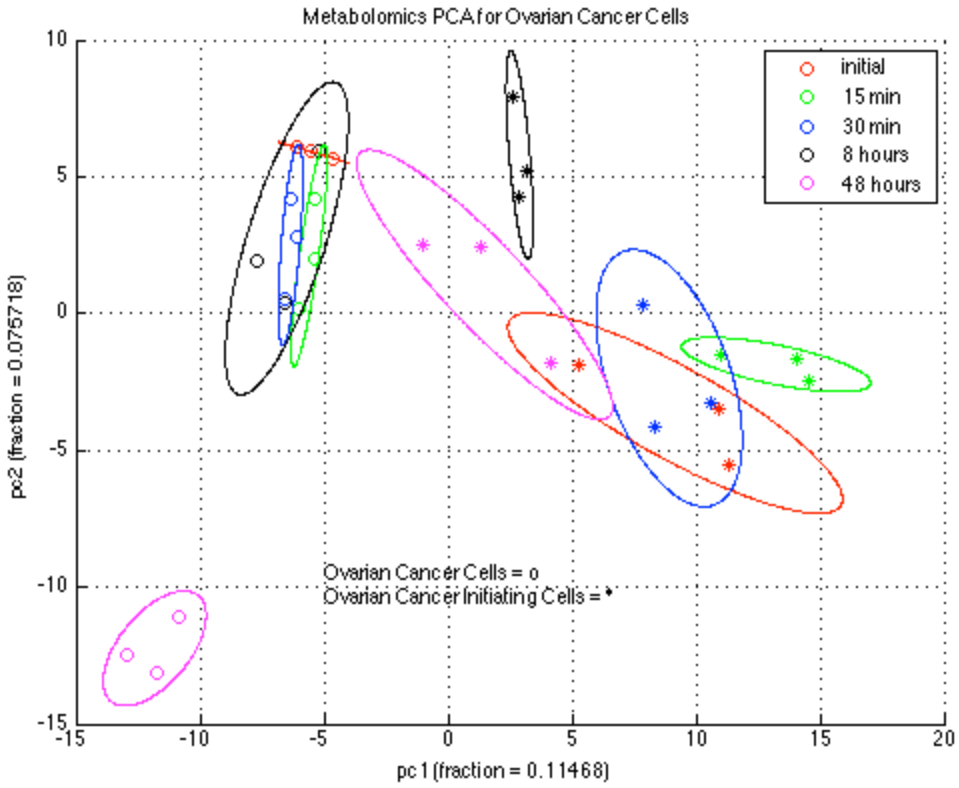
Part 4: Adding the 2-sigma ellipse

This is done by first computing the mean principal components for each set, then determining the matrix of covariance. We then discretize a domain which would encompass the ellipse (faster than using the entire domain of the graph, although that would work too) and calculate an expression which will be equal to 4 (e.g., 2^2) when we are on the 2-sigma ellipse. This

ellipse is then plotted using the appropriate color.

```
for i=1:m/3
    b=(i-1)*3;
    xdat=pc1(b+1:b+3);
    ydat=pc2(b+1:b+3);
    xmean=mean(xdat);
    ymean=mean(ydat);
    %Now we get the matrix of covariance of these points
    n=length(xdat);
    var=[mean((xdat-xmean).^2),mean((xdat-xmean).*(ydat-ymean)); ...
        mean((xdat-xmean).*(ydat-ymean)),mean((ydat-ymean).^2)]*n/(n-1);
    %Next we plot an ellipse.  We want it to be a 2-sigma ellipse, so we
    %choose a range that would contain it.  We get this from the variance
    %in x and y:
    xrange=linspace(xmean-var(1,1)^.5*2.1,xmean+var(1,1)^.5*2.1,200);
    yrange=linspace(ymean-var(2,2)^.5*2.1,ymean+var(2,2)^.5*2.1,200);
    zsq=zeros(length(xrange),length(yrange));
    varinv=inv(var);
    for j=1:length(xrange)
        for k=1:length(yrange)
            zsq(j,k)=[xrange(j)-xmean,yrange(k)-ymean]*varinv*[xrange(j)-xmean;yrang
        end
    end

    figure(1)
    hold on
    contour(xrange,yrange,zsq',[4 4],colors(i))
    drawnow
end
hold off
```



Part 5: Interpretation of the graph

As you can see from the figure, the metabolome of the cancer cells and cancer initiating cells are quite different. The PC1 and PC2 of the OCC cells are tightly clustered initially and after 48 hours, with a significant shift in PC2 over time. The initiating cells (OCIC) are less tightly clustered and evolve in a different direction. In fact, the final principal components PC1 and PC2 of the OCIC cells are both close to zero, while the final PC1 and PC2 of the OCC cells are of large magnitude. This means that the metabolic pathways associated with these principal components play a large role in OCC cells, but are less active in OCIC cells. Further examination of the PC's of the OCIC cells may show what metabolic pathways are more active, and may provide targets for attack on these dangerous cells.

Part 6: Extra Credit!

We are asked to determine the probablility that an additional sample would fall within the relevant ellipses. If a large number of samples had been used, this would have been about 95%. Since there are only three, however, we only have 2 degrees of freedom, and thus we are governed by the t-distribution. We can look it up, or use matlab's tcdf function:

```
probability = tcdf(2,2)-tcdf(-2,2)
```

```
probability =
    0.8165
```

Published with MATLAB® 7.12