IE 529 - Homework 2: Estimators. Matrix analysis.

Due Monday, October 3rd

Grading note: This homework was graded with 5 points for the three graded problems and 10 points for completing the rest of the assignment.

I. Estimators:

a. Determine the maximum likelihood estimator of θ when $X_1, X_2, \ldots, X_n$ is a random sample (i.i.d.) with density function

$$f(x) = \frac{1}{2} e^{-|x-\theta|}, \quad -\infty < x < \infty.$$

Proof: The log-likelihood function is given by

$$l(\theta) = -nlog(2) - \sum_{i=1}^{n} |x_i - \theta|.$$

Therefore

$$\frac{dl(\theta)}{d\theta} = \sum_{i=1}^{n} sgn(x_i - \theta),$$

where $sgn(\cdot)$ is the sign function equaling 1, 0, -1 for positive, zero, and negative arguments, respectively. Setting this derivative equal to zero gives that $\theta^\star$ is the median of the $\{x_1, \ldots, x_n\}$. ■

b. Sketch of proof for Bayes' estimation:

Suppose $\overline{x}$ is the sample mean for a random sample of size n taken from a normal distribution with unknown mean (denoted μ) and known variance σ^2 (i.e., $x_1, x_2, \ldots, x_n \in \mathcal{N}(\mu, \sigma^2)$); further make the *prior* assumption that the distribution for the mean is also normal, i.e., $\mu \in \mathcal{N}(\nu, \rho^2)$.

Show that the *posterior* distribution for the population mean μ is also normal, with mean $\mu*$ and standard deviation σ^* given by

$$\mu^* = \left(\frac{\rho^2}{\rho^2 + \frac{\sigma^2}{n}} \right) \overline{x} + \left(\frac{\frac{\sigma^2}{n}}{\rho^2 + \frac{\sigma^2}{n}} \right) \nu; \text{ and } \sigma^* = \sqrt{\frac{\rho^2 \sigma^2}{n\rho^2 + \sigma^2}}.$$

Hints: recall we know the following density functions apply:

$$f(x_1, x_2, \ldots, x_n | \mu) = \left(\frac{1}{2\pi} \right)^{\frac{n}{2}} \left(\frac{1}{\sigma} \right)^{n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\};$$

$$f(\mu) = \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \left(\frac{1}{\rho}\right) \exp\left\{-\frac{1}{2\rho^2}(\mu-\nu)^2\right\}.$$

The posterior distribution is $f(\mu|x_1, x_2, \ldots, x_n)$; show that this has a normal form, i.e., is $\mathcal{N}(\mu^*, \sigma^{*2})$.

Proof: Note that the marginal joint pdf is

$$\begin{aligned}
f(x_1, x_2, \ldots, x_n) &= \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n|\mu) f(\mu) d\mu \\
&= \int_{-\infty}^{\infty} \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \left(\frac{1}{\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \left(\frac{1}{\rho}\right) \exp\left\{-\frac{1}{2\rho^2}(\mu-\nu)^2\right\} d\mu \\
&= \left(\frac{1}{2\pi}\right)^{\frac{n+1}{2}} \frac{1}{\rho\sigma^n} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\left(\frac{n}{\sigma^2}+\frac{1}{\rho^2}\right)\mu^2 - 2\left(\frac{n\bar{x}}{\sigma^2}+\frac{\gamma}{\rho^2}\right)\mu + \frac{\sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\gamma^2}{\rho^2}\right)\right\} d\mu \\
&= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\rho\sigma^n} \exp\left\{-\frac{1}{2}\left(\frac{\sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\gamma^2}{\rho^2} - \left(\frac{n\bar{x}}{\sigma^2}+\frac{\gamma}{\rho^2}\right)^2 \left(\frac{n}{\sigma^2}+\frac{1}{\rho^2}\right)^{-1}\right)\right\} \left(\frac{n}{\sigma^2}+\frac{1}{\rho^2}\right)^{-\frac{1}{2}}.
\end{aligned}$$

Therefore using this and Bayes' rule

$$\begin{aligned}
f(\mu|x_1, x_2, \ldots, x_n) &= \frac{f(x_1, x_2, \ldots, x_n|\mu) f(\mu)}{f(x_1, x_2, \ldots, x_n)} \\
&= \frac{\left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \left(\frac{1}{\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \left(\frac{1}{\rho}\right) \exp\left\{-\frac{1}{2\rho^2}(\mu-\nu)^2\right\}}{\left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\rho\sigma^n} \exp\left\{-\frac{1}{2}\left(\frac{\sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\gamma^2}{\rho^2} - \left(\frac{n\bar{x}}{\sigma^2}+\frac{\gamma}{\rho^2}\right)^2 \left(\frac{n}{\sigma^2}+\frac{1}{\rho^2}\right)^{-1}\right)\right\} \left(\frac{n}{\sigma^2}+\frac{1}{\rho^2}\right)^{-\frac{1}{2}}} \\
&= \left(\frac{n\rho^2+\sigma^2}{2\pi\rho^2\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{n\rho^2+\sigma^2}{2\rho^2\sigma^2}\left(x - \left(\frac{\rho^2}{\rho^2+\frac{\sigma^2}{n}}\right)\overline{x} + \left(\frac{\frac{\sigma^2}{n}}{\rho^2+\frac{\sigma^2}{n}}\right)\nu\right)^2\right\} \\
&= \left(\frac{1}{2\pi\frac{\rho^2\sigma^2}{n\rho^2+\sigma^2}}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\frac{\rho^2\sigma^2}{n\rho^2+\sigma^2}}\left(x - \left(\frac{\rho^2}{\rho^2+\frac{\sigma^2}{n}}\right)\overline{x} + \left(\frac{\frac{\sigma^2}{n}}{\rho^2+\frac{\sigma^2}{n}}\right)\nu\right)^2\right\} \\
&= \left(\frac{1}{2\pi(\sigma^*)^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2(\sigma^*)^2}(x-\mu^*)^2\right\}.
\end{aligned}$$

■

II. Linear Algebra:

a. Show that the following statements are true:
For $x,\ y \in \mathbf{R}^n$,

$$\text{(i)}\ \|y\|_\infty = \max_{\|x\|_1 \neq 0} \left(\frac{|y^* x|}{\|x\|_1} \right), \qquad \text{(ii)}\ \|y\|_1 = \max_{\|x\|_\infty \neq 0} \left(\frac{|y^* x|}{\|x\|_\infty} \right).$$

Proof of (i): Notice that since $\frac{x}{\|x\|_1}$ has norm $= 1$, we can write the RHS as

$$\begin{aligned}
\max_{\|u\|_1=1} (|y^* u|) &\leq \max_{\|u\|_1=1} \sum_i |y_i u_i| \\
&= \max_{\|u\|_1=1} \sum_i |y_i||u_i| \\
&\leq \max_{\|u\|_1=1} \|y\|_\infty \sum_i |u_i| \\
&\leq \|y\|_\infty \max_{\|u\|_1=1} \|u\|_1 \\
&= \|y\|_\infty.
\end{aligned}$$

To show that this upper bound is achieved, suppose $\|y\|_\infty = |y_i|$. Choose $u_i = sgn(y_i), u_j = 0\ \forall\ j \neq i$. This implies $\|u\|_1 = 1$ and $|y^* u| = |y_i| = \|y\|_\infty$. ■

Proof of (ii): Using similar ideas as (i), we can write the RHS as

$$\begin{aligned}
\max_{\|u\|_\infty=1} (|y^* u|) &\leq \max_{\|u\|_1=\infty} \sum_i |y_i u_i| \\
&= \max_{\|u\|_1=\infty} \sum_i |y_i||u_i| \\
&\leq \max_{\|u\|_1=\infty} \|u\|_\infty \sum_i |y_i| \\
&\leq \max_{\|u\|_1=\infty} \|u\|_\infty \|y\|_1 \\
&= \|y\|_1.
\end{aligned}$$

To show that this upper bound is achieved, choose $u_i = sgn(y_i)\ \forall\ i$. This implies that $\|u\|_\infty = 1$ and $|y^* u| = \sum_i |y_i| = \|y\|_1$. ■

b. Suppose $a_1, a_2, \ldots, a_n$ are fixed positive real numbers. Determine which of the following are proper vector norms on $\mathbf{R}^n$ (i.e., which of the following satisfy the four conditions required of functions to be vector norms).

1. $\|x\| := \max_i\{a_i|x_i|\}$
2. $\|x\| := \sum_{i=1}^n a_i|x_i|$.

The four conditions required to be a norm are as follows:
(i) $\|x\| = 0 \Leftrightarrow x = 0$,
(ii) $\|cx\| = |c|\|x\|$,
(iii) $\|x + y\| \leq \|x\| + \|y\|$,
(iv) $\|x\| \geq 0$.

We will discuss these conditions for 1 and 2:

1. *Proof:* (i) $\|x\| = 0 \Leftrightarrow a_i|x_i| = 0 \;\forall\; i \Leftrightarrow |x_i| = 0 \;\forall\; i$ since $a_i > 0 \Leftrightarrow x = 0$.
(ii) $\|cx\| = \max_i\{a_i|cx_i|\} = \max_i\{a_i|c||x_i|\} = |c| \max_i\{a_i|x_i|\} = |c|\|x\|$.
(iii) Note $|x_i + y_i| \leq |x_i| + |y_i|$ and $a_i|x_i + y_i| \leq a_i|x_i| + a_i|y_i|$ imply
$\max_i\{a_i|x_i + y_i|\} \leq \max_i\{a_i|x_i| + a_i|y_i|\} \leq \max_i\{a_i|x_i|\} + \max_i\{a_i|y_i|\}$.
(iv) Since $a_i > 0$ and $|x_i| \geq 0$ for all $i \implies a_i|x_i| \geq 0$ for all i. Therefore $\|x\| = \max_i\{a_i|x_i|\} \geq 0$. ■

2. *Proof:* (i) $\|x\| = 0 \Leftrightarrow \sum_{i=1}^n a_i|x_i| = 0 \Leftrightarrow |x_i| = 0 \;\forall\; i$ since a_i's strictly positive $\Leftrightarrow x = 0$.
(ii) $\|cx\| = \sum_{i=1}^n a_i|cx_i| = \sum_{i=1}^n a_i|c||x_i| = |c| \sum_{i=1}^n a_i|x_i| = |c|\|x\|$.
(iii) Note $|x_i + y_i| \leq |x_i| + |y_i|$ and $a_i|x_i + y_i| \leq a_i|x_i| + a_i|y_i|$ imply
$\sum_{i=1}^n a_i|x_i + y_i| \leq \sum_{i=1}^n (a_i|x_i| + a_i|y_i|) = \sum_{i=1}^n a_i|x_i| + \sum_{i=1}^n a_i|y_i|$.
(iv) Since $a_i > 0$ and $|x_i| \geq 0$ for all $i \implies a_i|x_i| \geq 0$ for all i. Therefore $\|x\| = \sum_{i=1}^n a_i|x_i| \geq 0$. ■

c. For this problem we will prove that the induced matrix 2-norm for a matrix $A \in \mathbf{R}^{m \times n}$ is given by the maximum singular value, $\sigma_1(A)$.

In particular, show that

$$\begin{aligned} \max f(x) &= \|Ax\|_2^2 = x^T A^T A x \\ \text{subj. to } x^T x &= 1 \end{aligned},$$

is given by σ_1^2. Hint: consider using the SVD of the matrix A.

Proof: Let the svd of A be

$$A = \sum_i \sigma_i u_i v_i^T.$$

Note that

$$A^T A = (\sum_i v_i u_i^T \sigma_i)(\sum_i \sigma_i u_i v_i^T) = \sum_i \sigma_i^2 v_i v_i^T.$$

Any x can be written in the form

$$x = \sum_i \alpha_i v_i,$$

where the α_i are scalars, since $v_1, ..., v_n$ form a basis for $\mathbb{R}^n$. Therefore

$$x^T A^T A = (\sum_i \alpha_i v_i^T)(\sum_i \sigma_i^2 v_i v_i^T) = \sum_i \alpha_i \sigma_i^2 v_i^T,$$

which implies

$$x^T A^T A x = (\sum_i \alpha_i \sigma_i^2 v_i^T)(\sum_i \alpha_i v_i) = \sum_i \alpha_i^2 \sigma_i^2.$$

We need to maximize this over x such that $x^T x = 1$, which is equivalent to maximizing over $\alpha_1, ..., \alpha_n$ subject to $\sum_i \alpha_i^2 = 1$. This implies the optimal choice is $\sigma_1^2 = 1$, with $\sigma_2^2 = \cdots = \sigma_n^2 = 0$, since the singular values are in descending order. ■

III. Please see the posted Matlab file, PCAdata.mat, or the csv version for Python. Compute the SVD of the original matrix, and using the SVD discuss what you think a proper PCA should reveal.

Write code to compute a PCA. Namely,

1. compute and subtract off the mean from the data,
2. compute the covariance matrix (i.e., A*A'), including scaling by 1/(n-1), and then
3. compute an eigenvalue decomposition, and sort both the eigenvalues and associated eigenvectors in descending order.

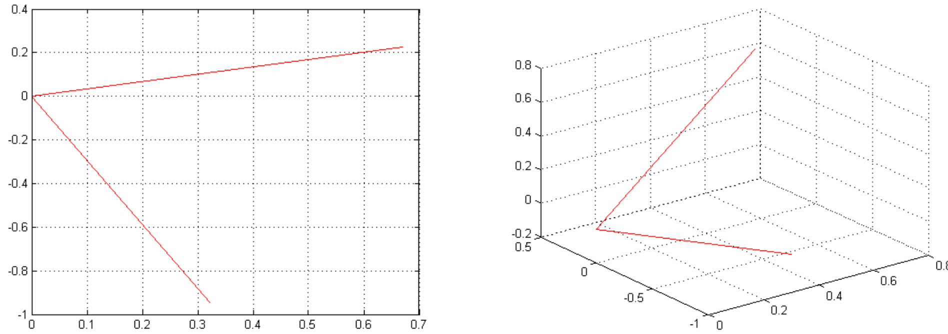
Plot and discuss (i) the principal components; and (ii) how this process may be done more directly using a SVD of the (de-biased, scaled) data.

(i) Various plots of the principal components are provided. Figures 1a-1b plot just the principal components in two- and three-dimensions, respectively.

Figures 2a-2b plot the de-biased, scaled data in three-dimensions with the principal components, unscaled and scaled by the corresponding eigenvalues, respectively. Any of those four were acceptable answers. If the third eigenvector was plotted no points were subtracted.

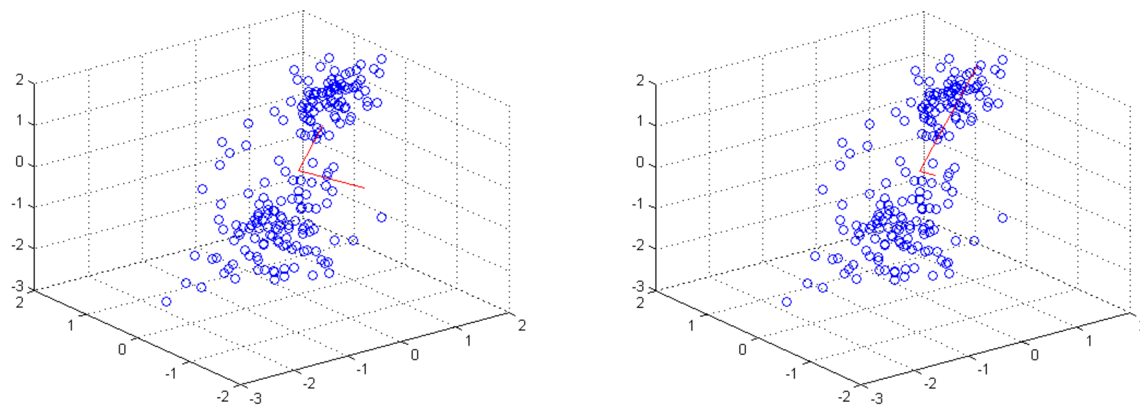
The plot in Figure 3 was not required but could also have been plotted.

For completeness we include the Matlab code, if you did it in Python and are not sure what was wrong, please come to office hours and we can walk through it:



(a) Two Dimensions (b) Three Dimensions

Figure 1: Principal Components



(a) Principal components unscaled (b) Principal components scaled by the corresponding eigenvalues

Figure 2: Principal Components in three dimensions with de-biased, scaled data

```
% hw2prob3.m
% Philip Pare
% 10/7/2016

[m, n] = size(A);

% 1. Compute and subtract mean
meanA = mean(A,2);
Amean0 = A - repmat(meanA,1,n);

% 2. Compute the covariance matrix
Cova = Amean0*Amean0'/(n-1);
```

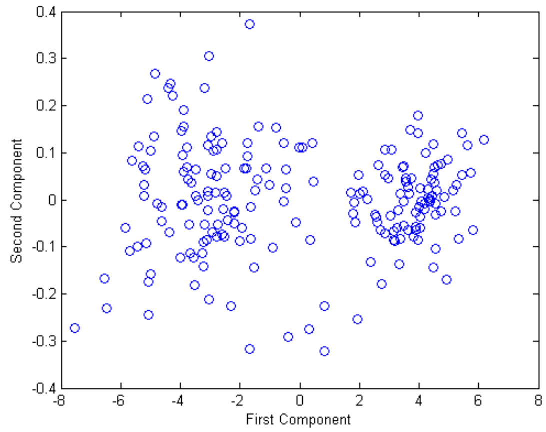


Figure 3: Two dimensional de-biased, scaled, and transformed data plotted against the principal components

```
% 3. Compute an eigenvalue decomposition and sort both sort both the
%    eigenvalues and associated eigenvectors in descending order
[U, D] = eig(Cova);
[eigvals,Ind] = sort(diag(D),'descend');
Us = U(:,Ind);

% (i) Possible plots for principal components
%      Note: we only need to plot the two eigenvectors corresponding to
%      nonzero eigenvalues

% 2-dimensional
figure(1)
clf
plot([0; Us(1,1)],[0;  Us(1,2)],'r')
hold on
plot([0; Us(2,1)],[0;  Us(2,2)],'r')
grid on

% 3-dimensional
figure(2)
clf
plot3([0; Us(1,1)],[0;  Us(1,2)],[0;  Us(1,3)],'r')
hold on
plot3([0; Us(2,1)],[0;  Us(2,2)],[0;  Us(2,3)],'r')
grid on

% if you want to be fancier you can also plot the data
```

```
figure(3)
clf
plot3([0; Us(1,1)],[0;  Us(1,2)],[0;  Us(1,3)],'r')
hold on
plot3([0; Us(2,1)],[0;  Us(2,2)],[0;  Us(2,3)],'r')
scatter3(Amean0(Ind(1),:),Amean0(Ind(2),:),Amean0(Ind(3),:))
grid on

% and we can scale the principal components by the eigenvalues
figure(4)
clf
plot3([0; eigvals(1)*Us(1,1)],[0;  eigvals(1)*Us(1,2)],[0;  eigvals(1)*Us(1,3)],'r
hold on
plot3([0; eigvals(2)*Us(2,1)],[0;  eigvals(2)*Us(2,2)],[0;  eigvals(2)*Us(2,3)],'r
scatter3(Amean0(Ind(1),:),Amean0(Ind(2),:),Amean0(Ind(3),:))
grid on

% This was not required but you could also plot the reduced, un-biased, and
% scaled data
data2d=D*U'*Amean0;
data2d = data2d(Ind,:);
data2d(3,:) = [];

figure(5)
clf
plot(data2d(1,:),data2d(2,:),'o')
hold on
xlabel('First Component')
ylabel('Second Component')
```

(ii) Let A be the data and A_0 be the mean zero data. Let the sorted eigenvalue decomposition of $\frac{A_0A_0'}{n-1}$, where n is the number of data points, be $\frac{A_0A_0'}{n-1} = U_sD_sU_s'$. Let the svd of $\frac{A_0'}{(n-1)^{\frac{1}{2}}}$ be $\frac{A_0'}{(n-1)^{\frac{1}{2}}} = UVD'$. Note that

$$U_sD_sU_s' = \frac{A_0A_0'}{n-1} = \left(\frac{A_0'}{(n-1)^{\frac{1}{2}}}\right)' \frac{A_0'}{(n-1)^{\frac{1}{2}}} = (UVD')'UVD' = DV'U'UVD' = DV'VD'.$$

Therefore D is the sorted principal components of the data ($D = U_s$) and the diagonal of V is the square root of the sorted eigenvalues ($V'V = D_s$).

IV. Seminar summaries:

a. Provide a summary of each talk attended (or viewed) from the *S*ymposium on Frontiers of Big Data. You should watch a minimum of 45 minutes of talks, and submit 3 paragraphs (min) summarizing the presentations.

b. Provide a summary of the session/talks attended from the *A*llerton Conference. You should either attend one of the tutorial sessions on Tuesday, or at least one full session (approx. 6 talks) at the conference (held at Allerton House, in Monticello). Please submit a 1/2 - 1 page summary of the presentation(s) you attended.

Please provide the titles and speaker names for all talks summarized.