1. (a)  
Because \( X \) is a bivariate Gaussian random vector, we can use the conditional pdf to predict \( X_2 \) given \( X_1 \):

\[
f_{X_2|X_1=x_1} \sim N(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(x_1 - \mu_1), (1 - \rho^2)\sigma_2^2)
\]

with expectation equal to \( \mu_2 + \frac{\sigma_2}{\sigma_1}\rho(x_1 - \mu_1) \). This is our estimate.

(b)  
The log-likelihood function of the bivariate Gaussian is

\[
ln l(\theta) = \log \left( \frac{1}{(2\pi)^n|\Sigma|^{n/2}} \right) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)\Sigma^{-1}(x_i - \mu)
\]

which can also be expressed as

\[
ln l(\theta) = \log \left( \frac{1}{(2\pi)^n|\Sigma|^{n/2}} \right) - \frac{1}{2} \sum_{i=1}^{n} (x_i'\Sigma^{-1}x_i - 2\mu'\Sigma^{-1}x_i + \mu'\Sigma^{-1}\mu).
\]

Taking the derivative of \( l(\theta) \) with respect to \( \mu \) we have

\[
\frac{\partial l(x)}{\partial \mu} = -\frac{1}{2} \sum_{i=1}^{n} -2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu \\
= \sum_{i=1}^{n} \Sigma^{-1}x_i - \Sigma^{-1}\mu
\]

Setting this expression equal to 0 results in

\[
n\Sigma^{-1}\mu = \Sigma^{-1}\sum_{i=1}^{n} x_i \Rightarrow \\
\Sigma\Sigma^{-1}\mu = \frac{1}{n} \Sigma\Sigma^{-1}\sum_{i=1}^{n} x_i \Rightarrow \\
\mu = \frac{1}{n} \sum_{i=1}^{n} x_i
\]
Next we find the MLE for $\Sigma$. First let us substitute $\bar{x} = n^{-1} \sum x_i$ for $\mu$, and then rearrange $l(\theta)$ using the circular properties of trace to obtain

\[
l(\theta) = \left( \frac{1}{(2\pi)^n |\Sigma|^{n/2}} \right) \exp \left( -\frac{1}{2} \sum_{i=1}^{n} tr((x_i - \bar{x})(x_i - \bar{x})^\prime \Sigma^{-1}) \right)
\]

Let $S = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\prime$ a symmetric positive definite matrix which has a matrix square root via the Cholesky decomposition, ie $S = S^{1/2}(S^{1/2})^\prime = S^{1/2}S^{1/2}$ where $S^{1/2} = (S^{1/2})^\prime$ since $S$ is symmetric.

Substituting this in, we obtain

\[
l(\theta) = \left( \frac{1}{(2\pi)^n |\Sigma|^{n/2}} \right) \exp \left( -\frac{1}{2} tr(S^{1/2}\Sigma^{-1}S^{1/2}) \right)
\]

Now let $B = S^{1/2}\Sigma^{-1}S^{1/2}$ also a positive definite matrix and note that $|\Sigma| = |\Sigma^{-1}|^{-1} = |B|^{-1}$ so the expression becomes

\[
l(\theta) = \left( \frac{|B|^{n/2}}{(2\pi)^n} \right) \exp \left( -\frac{1}{2} tr(B) \right)
\]

Because $|B| = \prod_{i=1}^{n} \lambda_i$ and $tr(B) = \sum_{i=1}^{n} \lambda_i$ where $\lambda_i$ is the $i^{th}$ eigenvalue of $B$, we need to maximize

\[
l(\theta) = \left( \frac{(\prod_{i=1}^{n} \lambda_i)^{n/2}}{(2\pi)^n} \right) \exp \left( -\frac{1}{2} \sum_{i=1}^{n} \lambda_i \right)
\]

for each $\lambda_i$. Differentiating, we set

\[
\frac{n}{2} \lambda_i^{n/2-1} \exp\left( -\frac{1}{2} \lambda_i \right) - \frac{1}{2} \lambda_i^{n/2} \exp\left( -\frac{1}{2} \lambda_i \right) = 0
\]

giving

\[
\frac{n}{2} \lambda_i^{n/2-1} \exp\left( -\frac{1}{2} \lambda_i \right) = \frac{1}{2} \lambda_i^{n/2} \exp\left( -\frac{1}{2} \lambda_i \right) \Rightarrow \lambda_i = n
\]
So \( \lambda_i = n \) for all \( i \). Using the eigendecomposition of \( B \) with \( Q \) equal to the matrix of the eigenvectors of \( B \) we have

\[
B = Q(nI_n)Q^{-1} = nI_n
\]

So finally we have the MLE for \( \Sigma \):

\[
\Sigma = S^{1/2}B^{-1}S^{1/2} = n^{-1}S = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'.
\]

2. (a) First suppose \( \mu = 0 \) and \( \Sigma = I_n \), then we have \( Z \sim N_n(0, I_n) \). The kernel of \( Z \) is then

\[
\exp((-1/2)Z'Z) = \exp((-1/2)\sum_{i=1}^{n}Z_i^2) = \prod_{i=1}^{n}\exp((-1/2)Z_i^2)
\]

which is the product of \( n \) individual \( N(0, 1) \) random variable kernels so this product integrates to \( (2\pi)^{n/2} \). And since \( |I_n| = 1 \), the result holds in this instance.

Now let \( X = \Sigma^{1/2}Z + \mu \), where \( \Sigma^{1/2} \) is the matrix square root of \( \Sigma \) (this exists since \( \Sigma \) is positive definite. Then \( Z = h(X) = \Sigma^{-1/2}(X - \mu) \) and the Jacobian of the transformation is

\[
\det\left(\frac{\partial h(X)}{\partial X}\right) = \det(\Sigma^{-1/2})
\]

\[
= \frac{1}{\det(\Sigma^{1/2})}
\]

\[
= \frac{1}{\det(\Sigma)^{1/2}}
\]

By the multivariate change of variables theorem we have

\[
f_X(x) = f_Z(h(x))\det\left(\frac{\partial h(X)}{\partial X}\right)
\]

\[
= \frac{1}{\det(\Sigma)^{1/2}}\exp((-1/2)(\Sigma^{-1/2}x - \mu)'(\Sigma^{-1/2}x - \mu))
\]

\[
= \frac{1}{\det(\Sigma)^{1/2}}\exp((-1/2)(x - \mu)\Sigma^{-1}(x - \mu))
\]

This expression also integrates to \( (2\pi)^{n/2} \), so we have
\[
\int_{\mathbb{R}^n} \exp \left( -\frac{1}{2} (x - \mu)\Sigma^{-1} (x - \mu) \right) = (2\pi)^{n/2} \det(\Sigma)^{1/2}
\]
for all \(\mu\) and \(\Sigma > 0\).

(b) To show that \(E(X) = \mu\) and \(\text{Var}(X) = \Sigma\), we can use the same transformation as in (a), that is

\[
X = \Sigma^{1/2} Z + \mu
\]

where \(Z \sim N_n(0, I_n)\). It is trivial to show that \(E(Z) = 0\) and \(\text{Var}(Z) = I_n\).

So now

\[
EX = E(\Sigma^{1/2} Z + \mu) = \mu
\]

and

\[
\text{Var}(X) = E((\Sigma^{1/2} Z + \mu - \mu)(\Sigma^{1/2} Z + \mu - \mu)') = E(\Sigma^{1/2} ZZ' (\Sigma^{1/2})') = \Sigma^{1/2}I_n(\Sigma^{1/2})' = \Sigma
\]

3. (a) False.

Consider

\[
\Sigma = \begin{bmatrix}
1 & .5 & .5 & 0 \\
.5 & 1 & 0 & 0 \\
.5 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

So that \(\Sigma_{23} = 0\).

Then

\[
\Sigma^{-1} = \Lambda = \begin{bmatrix}
2 & -1 & -1 & 0 \\
-1 & 1.5 & 0.5 & 0 \\
-1 & 0.5 & 1.5 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

So that \(\Lambda_{23} = .5 \neq 0\)

(b) False.

Consider
\[ \Lambda = \begin{bmatrix} 1 & 1 & -1 & 0 \\ 1 & 1 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{bmatrix} \]

So that \( \Lambda_{23} = 0 \)

Then

\[ \Lambda^{-1} = \Sigma = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \]

So that \( \Sigma_{23} = 1 \neq 0 \)

\((c)\) True

In order for \(X^1, X^2, X^3\) to be mutually independent given \(X^4\), we need the term \((x - \mu)' \Lambda (x - \mu)\) to have no products involving any combination of \(X^1, X^2, X^3\).

Suppose that \(\mu = 0\) for simplicity and that

\[ \Lambda = \begin{bmatrix} a & 0 & 0 & b \\ 0 & c & 0 & d \\ 0 & 0 & e & f \\ b & d & f & g \end{bmatrix} \]

so that \(\Lambda_{12} = \Lambda_{23} = \Lambda_{31} = 0\).

Then \(x' \Lambda x = x_1(ax_1 + bx_4) + x_2(cx_2 + dx_4) + x_3(ex_3 + fx_4) + x_4(bx_1 + dx_2 + fx_3 + gx_4)\)

And so no products involve any combination of \(X^1, X^2, X^3\). Thus, when conditioning on \(X^4\), the pdf of \(X^1, X^2, X^3\) will factor into the individual marginal pdfs, implying that \(X^1, X^2, X^3\) are mutually independent.

\((d)\) False

Following the vein of part \((c)\), we need to construct a \(\Sigma\) such that \((x - \mu)' \Lambda (x - \mu)\) has products involving any combination of \(X^1, X^2, X^3\).

Let

\[ \Sigma = \begin{bmatrix} 1 & 0 & 0 & .5 \\ 0 & 1 & 0 & .5 \\ 0 & 0 & 1 & 0 \\ .5 & .5 & 0 & 1 \end{bmatrix} \]

So that \(\Sigma_{12} = \Sigma_{23} = \Sigma_{31} = 0\)
Then

\[
\Lambda = \begin{bmatrix}
1.5 & 0.5 & 0 & -1 \\
0.5 & 1.5 & 0 & -1 \\
0 & 0 & 1 & 0 \\
-1 & -1 & 0 & -2 \\
\end{bmatrix}
\]

Again, assuming \( \mu = 0 \) for simplicity, we have

\[
x' \Lambda x = \frac{3}{2} x_1^2 + \frac{3}{2} x_2^2 + x_3^2 - 2 x_1 x_4 - 2 x_2 x_4 + x_1 x_2
\]

And so the term \( x_1 x_2 \) will prevent the pdf of \( X_1, X_2, X_3 \) conditioned on \( X_4 \) from factoring into the product of the marginal pdfs. Thus \( X_1, X_2, X_3 \) are not mutually independent given \( X_4 \).

4. (a)
5. (a)

The following is the crucial steps of the EM algorithm, written in R. This part of the code updates \( \Lambda \) and \( \Psi \).

\[
\text{sigma.xy} = \text{solve}(I.p + t(\text{lambda}) \%*\% \text{psi.inv} \%*\% \text{lambda})
\]

\[
\text{Y.center} = \text{apply}(Y, 2, \text{scale}, \text{scale} = \text{F}, \text{center} = \text{T})
\]

\[
\text{multiplier} = \text{solve}(I.p + t(\text{lambda}) \%*\% \text{psi.inv} \%*\% \text{lambda}) \%*\% t(\text{lambda}) \%*\% \text{psi.inv}
\]

\[
\text{X} = \text{multiplier} \%*\% t(\text{Y.center})
\]

\[
\text{lambda.plus} = t(Y) \%*\% t(X) \%*\% \text{solve}(n*\text{sigma.xy} + X \%*\% t(X))
\]

\[
\text{s} = (t(Y) \%*\% Y - \text{lambda.plus} \%*\% X \%*\% Y)
\]

\[
\text{psi.plus} = (1/n)*\text{diag}(\text{diag(s)}, q)
\]

When running this code, I initialized \( \Lambda \) close to the given \( \lambda \), plus a small amount of gaussian noise. If the gaussian noise had small variance (less than 1) the algorithm outputted the following for \( \Lambda \) and \( \Psi \):

\[
\begin{bmatrix}
.8635 & .0278 \\
.4321 & .4689 \\
.0202 & .9206
\end{bmatrix}
\]

\[
\begin{bmatrix}
.3279 & 0 & 0 \\
0 & .1684 & 0 \\
0 & 0 & .3421
\end{bmatrix}
\]

\( \Lambda \) has entries fairly close to the truth, though somewhat off, as we might expect due to its unidentifiability.
(b)
Because $\Lambda$ is converging to have the same column space as $\Lambda^*$, we might use their projection matrices, $P_\Lambda$ and $P_{\Lambda^*}$, to compare the projections of these two matrices (assuming they exist). As the column space of $\Lambda$ goes to that of $\Lambda^*$, we would expect that the operator norm of their difference, $||P_\Lambda - P_{\Lambda^*}||_{op}$ becomes small.

This idea seems reasonable, since with the matrix found in part (a), $||P_\Lambda - P_{\Lambda^*}||_{op} \approx 0.008$, a small number.

With this idea, I initialized $\Lambda$ with $N(0,1)$ values and ran the EM algorithm until $||P_{\Lambda^t} - P_{\Lambda^{t+1}}||_{op} < \epsilon$ for varying $\epsilon > 0$, and again found $||P_\Lambda - P_{\Lambda^*}||_{op} \approx .008$

(c)
Factor analysis estimates that $Y$ mostly lives in the submanifold defined by $Y = \hat{\mu} + \Lambda X$. To project $Y$ into this submanifold, we can use the projection matrix, $P_\Lambda = \Lambda(\Lambda'\Lambda)^{-1}\Lambda'$. This projection is shown below, with a sample of size 50 from $Y$. The original $Y$‘s are in black, the projection in red.
We can better visualize the projection if we rotate the plot:

\[
\text{(d)} \quad q = 4: \quad ||P_\Lambda - P_{\Lambda^*}||_{op} \approx 0.02
\]

\[
q = 7: \quad ||P_\Lambda - P_{\Lambda^*}||_{op} \approx 0.008
\]

So we see that for \( q = 4 \), the algorithm has a harder time converging than when \( q = 7 \).

\text{(e)} Computing the principle components in R is simple:

\[
\text{Sigma} = \text{cov}(Y)
\]

\[
\text{pc} = \text{eigen}(\text{Sigma}) \text{vectors}
\]

The first two principle components account for 0.9866 of the total variance and \( ||P_{\text{pca}} - P_{\Lambda^*}||_{op} \approx .008 \), so we obtain nearly exactly the same result as from factor analysis.

Using the formula listed in the notes, we have that the variance explained by factor analysis is \( 1 - (q-p)^{-1} \sum_{i=p+1}^{q} \lambda_i \) where \( \lambda_i \) is the \( i^{th} \) leading eigenvalue of the sample covariance matrix. This value is 0.968, so PCA performs slightly better at explaining the variance of \( Y \).