Solutions to problems week 2.

1.

$$\sum_{x=0}^{1} p(x|\mu) = \sum_{x=0}^{1} \mu^x (1-\mu)^{1-x} = 1 - \mu + \mu = 1$$

$$E[x] = \sum_{x=0}^{1} x\mu^x (1-\mu)^{1-x} = \mu + 0 = \mu$$

$$var(x) = E[x^2] - (E[x])^2 = \mu - \mu^2 = \mu(1-\mu)$$

2. A student's solution: For Bayesian inference it is possible to specify any prior distribution for the parameters. But some priors result in a posterior distribution that resembles the prior in its form. Such priors are called conjugate priors. In addition to being mathematically convenient, conjugate priors most often lead to a more intuitive interpretation of the posterior. This is because the change in the shape of the posterior from the prior after incorporating the likelihood can be summarized using only the changes in the parameters governing the form of the prior(and posterior).

For example, the conjugate prior for the parameter μ for Bernoulli observations is the Beta distribution.

3. a)

$S = ULU'$. For S^{-1}, we have $S^{-1} = (ULU^{'})^{-1}$. Since U is orthogonal, we have that $U' = U^{-1}$, and $(U')^{-1} = U$. Then,

$S^{-1} = (U^{'})^{-1}L^{-1}U^{-1} = UL^{-1}U^{-1} = UL^{-1}U^{'}$.

b)

Equation 2.48 is:

$$\mathbf{\Sigma} = \sum_{\mathbf{i=1}}^{\mathbf{D}} \lambda_\mathbf{i} \mathbf{u_i} \mathbf{u_i}^\mathbf{T}$$

If we multiply by u_j on the right on both sides, we have:

$$\mathbf{\Sigma u_j} = \sum_{\mathbf{i=1}}^{\mathbf{D}} \lambda_\mathbf{i} \mathbf{u_i} \mathbf{u_i}^\mathbf{T} \mathbf{u_j} = \sum_{\mathbf{i=1}}^{\mathbf{D}} \lambda_\mathbf{i} \mathbf{u_i} \mathbf{I_{ij}} = \lambda_\mathbf{i} \mathbf{u_i} = \lambda_\mathbf{j} \mathbf{u_j},$$ where we have used equation 2.46.

4.

a) See Figure 1.

b)

The mean is [-1.9979 2.9998]. The covariance is $\begin{pmatrix} 7.8194 & -3.9136 \\ -3.9136 & 3.2636 \end{pmatrix}$

c) See Figure 2 and Figure 3.

d)

One student's answer:

The histograms in figure 2 seem to imply different underlying distri- butions for y for the two different ranges of x. So the distribution of y is most likely not independent of x. That is y is not independent of x.

Second student's answer:

The two histograms are very different. The first one has a peak between 1 and 2 while the second has peak between -1 and 0. The first has spread between -2 and 4 while the second spreads from -4 to 2. The histograms suggest that $p(Y|X \in (-2.1, -1.9)) \neq p(Y|X \in (2.9, 3.1))$ which means that X and Y are not independent.

e)

The transformation matrix ($y = U(x - \mu)$) is:

$U = \begin{pmatrix} -0.4985 & -0.8669 \\ -0.8669 & 0.4985 \end{pmatrix}$.

See Figure 4 for the results.

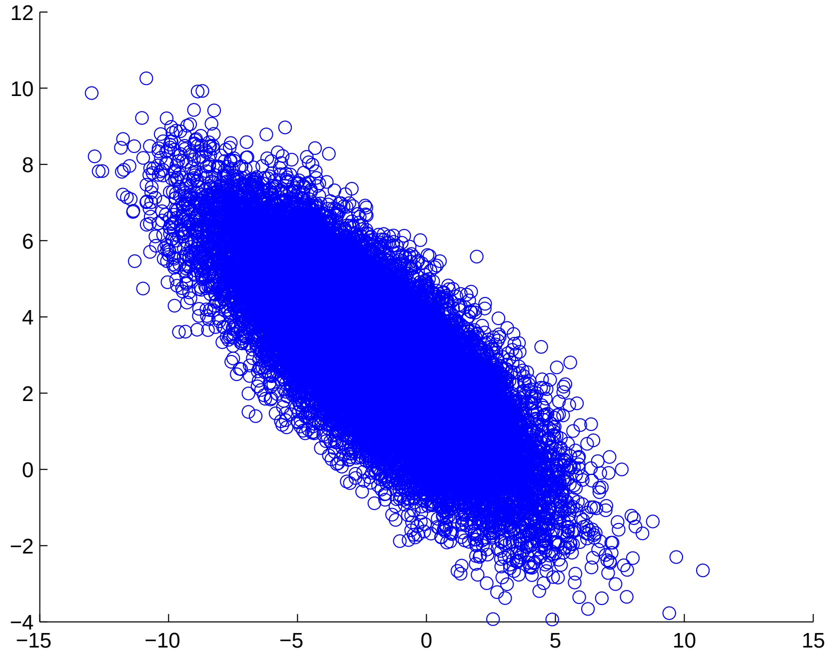

Figure 1: The data.

5. (Comments from Kobus). This problems was a bit ill-formed. As pointed out in class, if we were to use the method implied by 2.71 to prove that a formula was in fact a Gaussian, we would need to consider the interaction with any multipliers in front of the exp() that we had, because the final form of the Gaussian has a specific factor in from of the exp().

However, since any part of the constant term that is not needed to "complete the square" can be subtracted out (becoming part of the factor in front of the exp()), the exponentiation of the form 2.71 gives a form that is proportional to a Gaussian regardless of the constant term. This in turn means that either the formula will be a Gaussian (the complete expression has unit integral), or it is not a valid probability distribution.

In summary, the method outlined in page 86-87 is a good way to find the mean and variance of the Gaussian on the assumption that the function is a probability density function. However, it is not foolproof without that assumption, unless the constants are checked also.

6. See Figure 5

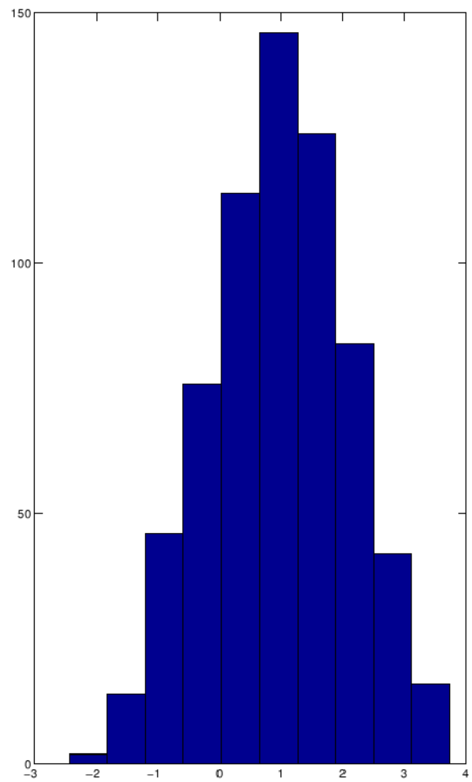

Figure 2: The data in the range (-2.1,-1.9)

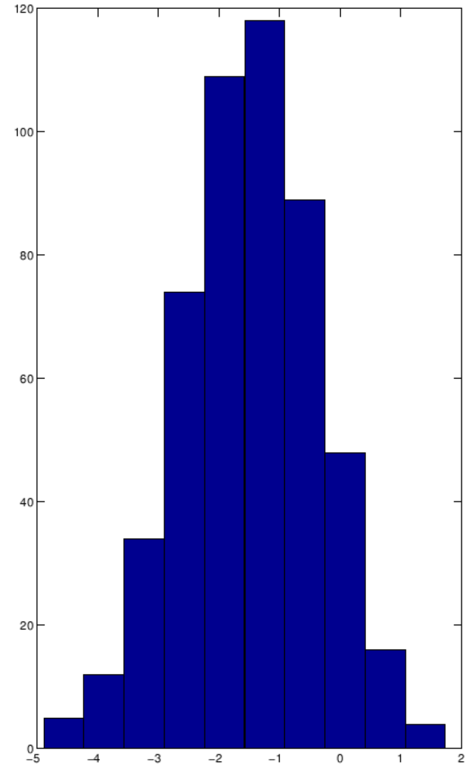

Figure 3: The data in the range (2.9,3.1)

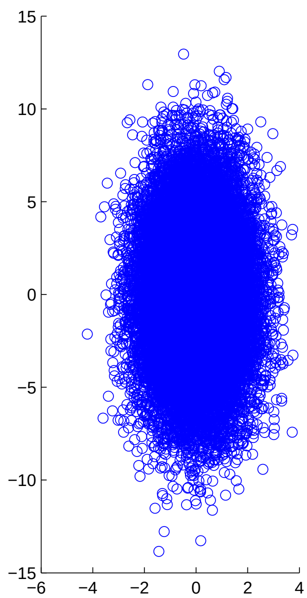

Figure 4: The data in the new coordinate system (axes scaled approximately equally).

The density function of a multivariate D dimensional Gaussian is given by

$$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) &= \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{1/2}} \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})}{2}\right] \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{1/2}} \exp\left[-\frac{\mathbf{x}^{\mathrm{T}}\Sigma^{-1}\mathbf{x} - 2\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}}{2}\right] \\
&= \frac{\exp\left[\frac{-\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}}{2}\right]}{(2\pi)^{\frac{D}{2}}|\Sigma|^{1/2}} \exp\left[-\frac{\mathbf{x}^{\mathrm{T}}\Sigma^{-1}\mathbf{x} - 2\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\mathbf{x}}{2}\right]
\end{aligned}
$$

Note that

$$
\mathbf{x}^{\mathrm{T}}\Sigma^{-1}\mathbf{x} = \sum_{i=1}^{D}\sum_{j=1}^{D} x_i \Sigma_{ij}^{-1} x_j
$$

Let $(\mathrm{A})_v$ denote a column vector formed from the matrix A by choosing the elements from A sequentially in the order left to right and top to bottom (raster scan). Comparing the above equation with the general representation of exponential family of distributions

$$
\begin{aligned}
\boldsymbol{\eta} &= \begin{pmatrix} \Sigma^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}(\Sigma^{-1})_v \end{pmatrix} \\
\mathrm{u}(\mathbf{x}) &= \begin{pmatrix} \mathbf{x} \\ (\mathbf{x}\mathbf{x}^{\mathrm{T}})_v \end{pmatrix} \\
h(\mathbf{x}) &= (2\pi)^{-D/2} \\
g(\boldsymbol{\eta}) &= \frac{1}{|\Sigma|^{1/2}} \exp\left[-\frac{\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}}{2}\right]
\end{aligned}
$$

Figure 5: Student's solution to problem 6.

Someone else simplified matters by a change of coordinates to zero mean and diagonal covariance using the eigenvectors of the covariance matrix. If one does this, then one needs to take care that the exponantial family equations are terms of the new variables (they cannot be mixed haphazardly).

The solution that was provided in class used trace identities to help manage the matrix elements. These identities are generally useful, so this solution is provided here as well:

6. Problem 2.57

ans The Exponential density has the form

$$p(x) = h(x)g(\eta)exp(\eta' u(x))$$

For D-dimensional Gaussian,

$$p(x) = (2\pi)^{-D/2}|\Sigma|^{-1/2} exp\frac{-1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)$$

The exponential part is

$$\begin{aligned}(x-\mu)'\Sigma^{-1}(x-\mu) &= x'\Sigma^{-1}x - 2x'\Sigma^{-1}\mu + \mu'\Sigma^{-1}\mu && (6.1)\\ &= Tr(\Sigma^{-1}xx') - 2(\Sigma^{-1}\mu)'x + \mu'\Sigma^{-1}\mu && (6.2)\\ &= vec(\Sigma^{-1})'vec(xx') - 2(\Sigma^{-1}\mu)'x + \mu'\Sigma^{-1}\mu && (6.3)\end{aligned}$$

In (6.2), $Tr(A) = \sum_{i=1}^{D} A_{ii}$. Then $Tr(AB) = Tr(BA)$. In (6.3), $vec(A) = (A_{11}, A_{12}, \ldots, A_{1D}, A_{21}, \ldots, A_{DD})'$. Then $vec(A)'vec(B) = Tr(AB')$. So from (6.3),

$$p(x) = (2\pi)^{-D/2}|\Sigma|^{-1/2}(exp\frac{-1}{2}\mu'\Sigma^{-1}\mu)exp\frac{-1}{2}(vec(\Sigma^{-1})'vec(xx') - 2(\Sigma^{-1}\mu)'x) \qquad (6.4)$$

Then (6.4) is the exponential form with

$$\begin{aligned} h(x) &= 1\\ g(\eta) &= (2\pi)^{-D/2}|\Sigma|^{-1/2}(exp\frac{-1}{2}\mu'\Sigma^{-1}\mu)\\ \eta &= -1/2\begin{pmatrix} vec(\Sigma^{-1})\\ -2\Sigma^{-1}\mu\end{pmatrix}\\ u(x) &= \begin{pmatrix} vec(xx')\\ x\end{pmatrix}\end{aligned}$$

7. The right hand side (RHS) is given by:

$RHS = \frac{1}{N}\sum_{i=1}^{N} u(x_i) = \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} x_i \\ x_i^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{pmatrix}$ And the left hand side (LHS) is given by:

$$= -\nabla \ln g(\eta_{ML})$$

$$= -\nabla[\tfrac{1}{2}\ln(-2\eta_2) + \tfrac{\eta_1^2}{4\eta_2}]$$

$$= -\left(\tfrac{2\eta_1}{4\eta_2}, \tfrac{1}{2}\tfrac{-2}{-2\eta_2} - \tfrac{\eta_1^2}{4\eta_2^2}\right)^T$$

$$= (\mu_{ML}, \mu_{ML}^2 + \sigma_{ML}^2)^T$$

Hence,

$$\begin{pmatrix} \mu_{ML} \\ \mu_{ML}^2 + \sigma_{ML}^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{pmatrix}$$

This is clearly the result expected, since it corresponds to the maximum likelihood estimators for the Gaussian distribution derived previously.

8. We can use Gaussian Elimination to find the inverse. Let $\mathrm{M} = (\mathbf{A} - \mathrm{BD}^{-1}\mathrm{C})^{-1}$, then

$$\left(\begin{array}{cc|cc} \mathrm{A} & \mathrm{B} & \mathrm{I} & 0 \\ \mathrm{C} & \mathrm{D} & 0 & \mathrm{I} \end{array}\right) \Rightarrow \left(\begin{array}{cc|cc} \mathrm{A}-\mathrm{CBD}^{-1} & 0 & \mathrm{I} & -\mathrm{BD}^{-1} \\ \mathrm{C} & \mathrm{D} & 0 & \mathrm{I} \end{array}\right)$$
$$\Rightarrow \left(\begin{array}{cc|cc} \mathrm{I} & 0 & \mathrm{M} & -\mathrm{MBD}^{-1} \\ \mathrm{C} & \mathrm{D} & 0 & \mathrm{I} \end{array}\right)$$
$$\Rightarrow \left(\begin{array}{cc|cc} \mathrm{I} & 0 & \mathrm{M} & -\mathrm{MBD}^{-1} \\ 0 & \mathrm{D} & -\mathrm{CM} & \mathrm{I}+\mathrm{CMBD}^{-1} \end{array}\right)$$
$$\Rightarrow \left(\begin{array}{cc|cc} \mathrm{I} & 0 & \mathrm{M} & -\mathrm{MBD}^{-1} \\ 0 & \mathrm{I} & -\mathrm{D}^{-1}\mathrm{CM} & \mathrm{D}^{-1}+\mathrm{D}^{-1}\mathrm{CMBD}^{-1} \end{array}\right)$$

Thus,

$$\begin{pmatrix} \mathrm{A} & \mathrm{B} \\ \mathrm{C} & \mathrm{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathrm{M} & -\mathrm{MBD}^{-1} \\ -\mathrm{D}^{-1}\mathrm{CM} & \mathrm{D}^{-1}+\mathrm{D}^{-1}\mathrm{CMBD}^{-1} \end{pmatrix}$$

Proof:

$$\begin{pmatrix} \mathrm{A} & \mathrm{B} \\ \mathrm{C} & \mathrm{D} \end{pmatrix}\begin{pmatrix} \mathrm{A} & \mathrm{B} \\ \mathrm{C} & \mathrm{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathrm{I} & 0 \\ 0 & \mathrm{I} \end{pmatrix}$$

9. Start to the solution is given, but not completed

$$
\begin{aligned}
p(\mathbf{X}|\mu) &= \frac{1}{(2\pi)^{ND/2}} \frac{1}{|\boldsymbol{\Sigma}|^{N/2}} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(\mathrm{x}_n-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathrm{x}_n-\mu)\right\} \\
p(\mu) &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mu-\mu_0)^T\boldsymbol{\Sigma}_0^{-1}(\mu-\mu_0)\right\}
\end{aligned}
$$

$$
\begin{aligned}
p(\mu|\mathbf{X}) &\propto p(\mathbf{X}|\mu)p(\mu) \\
&\propto \exp\left\{-\frac{1}{2}\left[\left(\sum_{n=1}^{N}(\mathrm{x}_n-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathrm{x}_n-\mu)\right)+(\mu-\mu_0)^T\boldsymbol{\Sigma}_0^{-1}(\mu-\mu_0)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-N\mu_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mu-\mu^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}N\mu_{\mathrm{ML}}-N\mu^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mu+\right.\right. \\
&\quad \left.\left.\mu^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\mu-\mu^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\mu_0-\mu_0^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\mu\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2N\mu^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mu_{\mathrm{ML}}-2\mu^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\mu_0-\mu^{\mathrm{T}}(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})\mu\right]\right\}
\end{aligned}
$$