CS-6190: Homework 1

James Brissette

September 24, 2019

1 Warm Up

1. To get $p(x_1)$ from $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ we need to use a fair bit of wizardry, and assume $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, so $\Sigma^{-1} = V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$. Furthermore, we know that $x^T Ax = \sum_i \sum_j x_i A_{ij} x_j$, so the term in our multivariate gaussian exponent becomes:

$$\begin{aligned} &\left(\frac{1}{2\pi}\right)^{\frac{k}{2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} exp\left[\frac{1}{2} x^T \Sigma^{-1} x\right] \\ &= x_1^T V_{11} x_1 + x_1^T V_{12} x_2 + x_2^T V_{21} x_1 + x_2^T V_{22} x_2 \end{aligned}$$

   When we go through and tediously complete the square we get something that looks like the following:

$$x^T \Sigma^{-1} x = (x_2 + V_{22}^{-1} V_{21} x_1)^T V_{22} (X_2 + V_{22}^{-1} V_{21} x_1) + x_1^T (V_{11} - V_{21}^T V_{22}^{-1} V_{21}) x_1$$

   and with some guidance from various sources that illustrate that $f(x) = f(x_2|x_1) f(x_1)$, we can simplify this and recover the marginal distribution of x_1 as follows:

$$\begin{aligned} f(x_2|x_1) &\propto exp(x_2 + V_{22}^{-1} V_{21} x_1)^T V_{22} (X_2 + V_{22}^{-1} V_{21} x_1) \\ f(x_1) &\propto exp((x_1^T - \mu)^T (V_{11} - V_{21}^T V_{22}^{-1} V_{21})(x_1 - \mu)) \\ X_1 &\sim \mathcal{N}(\mu, (V_{11} - V_{21}^T V_{22}^{-1} V_{21})) \end{aligned}$$

   Which turns out to be exactly $X_1 \sim \mathcal{N}(\mu, \Sigma_{11})$

2.

3. We can simplify this equation by using the definition of Expected value and some Matrix

Cookbook trace tricks:

$$
\begin{aligned}
H(x) &= -\int p(x)log(p(x))dx \\
&= -\int \mathcal{N}(x|\mu,\Sigma) * log\mathcal{N}(x|\mu,\Sigma) \\
&= -\mathbb{E}\Big(log\mathcal{N}(x|\mu,\Sigma)\Big) \\
&= -\mathbb{E}\Big(log[2\pi^{\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}exp\big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\big)]\Big) \\
&= -\mathbb{E}(-\frac{d}{2}log(2\pi)) - \mathbb{E}(-\frac{1}{2}log|\Sigma|) - \mathbb{E}(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \\
&= \frac{d}{2}log(2\pi) + \frac{1}{2}log|\Sigma| + \frac{1}{2}\mathbb{E}((x-\mu)^T\Sigma^{-1}(x-\mu)) \\
&= \frac{d}{2}log(2\pi) + \frac{1}{2}log|\Sigma| + \frac{1}{2}\mathbb{E}(tr((x-\mu)^T\Sigma^{-1}(x-\mu))) \\
&= \frac{d}{2}log(2\pi) + \frac{1}{2}log|\Sigma| + \frac{1}{2}\mathbb{E}(tr(\Sigma^{-1}(x-\mu)(x-\mu)^T)) \\
&= \frac{d}{2}log(2\pi) + \frac{1}{2}log|\Sigma| + \frac{1}{2}tr(\mathbb{E}(\Sigma^{-1}(x-\mu)(x-\mu)^T)) \\
&= \frac{d}{2}log(2\pi) + \frac{1}{2}log|\Sigma| + \frac{1}{2}tr(\Sigma^{-1}\mathbb{E}((x-\mu)(x-\mu)^T)) \\
&= \frac{d}{2}log(2\pi) + \frac{1}{2}log|\Sigma| + \frac{1}{2}tr(\Sigma^{-1}\Sigma) \\
&= \frac{d}{2}log(2\pi) + \frac{1}{2}log|\Sigma| + \frac{1}{2}tr(I_d) \\
&= \frac{d}{2}log(2\pi) + \frac{1}{2}log|\Sigma| + \frac{d}{2} \\
&= \frac{d}{2}(1+log(2\pi)) + \frac{1}{2}log|\Sigma|
\end{aligned}
$$

4. We know that $KL(q||p) = \int \big[log(q(x)) - log(p(x))\big]p(x)dx$. If we expand out each term we get:

$$
\begin{aligned}
&\int log\Big[\frac{1}{2|\Lambda|}exp(\frac{1}{2}(x-m)^T\Lambda^{-1}(x-m))\Big] - log\Big[\frac{1}{2|\Sigma|}exp(\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))\Big]p(x)dx \\
&= \int \Big[\frac{1}{2}log\frac{|\Lambda|}{|\Sigma|} - \frac{1}{2}(x-m)^T\Lambda^{-1}(x-m) + \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big]p(x)dx \\
&= \frac{1}{2}log\frac{|\Lambda|}{|\Sigma|} - tr\big[E[(x-\mu)(x-\mu)^T]\Sigma^{-1}\big] + \frac{1}{2}E[(x-m)^T\Lambda^{-1}(x-m)]
\end{aligned}
$$

and we know from the matrix cookbook that $E(x^TAx) = tr(A\Sigma) + m^TAm$:

$$
\begin{aligned}
&= \frac{1}{2}log\frac{|\Lambda|}{|\Sigma|} - tr\big[I_d\big] + \frac{1}{2}(m-\mu) + \frac{1}{2}tr(^{-1}\Sigma)) \\
&= \frac{1}{2}\big[log\frac{|\Lambda|}{|\Sigma|} - d + tr(\Lambda^{-1}\Sigma) + (m-\mu)^T\Lambda^{-1}(m-\mu)\big]
\end{aligned}
$$

5. So if we use a fair bit of what is in the lecture notes, this isn't too bad. We know that $\nabla log(Z(\eta)) = \frac{\nabla Z(\eta)}{Z(\eta)}$, that $Z(\eta) = \int h(x)exp(u(x)^T\eta)dx$ and that $\nabla Z(\eta) = \int h(x)exp(u(x)^T\eta)u(x)dx$.

   If we combine terms to simplify the gradient by taking $\frac{1}{Z(\eta)}$ to be $exp(-log(Z(\eta)))$ we can combine terms and calculate the second derivative as follows:

$$\begin{aligned}
\nabla log(Z(\eta)) &= \int u(x)h(x)exp(u(x)^T\eta - log(Z(\eta))) \\
\nabla^2 \frac{log(Z(\eta))}{d\eta} &= \int u(x)h(x)exp(u(x)^T\eta - log(Z(\eta)))(u(x) - \nabla log(Z(\eta))
\end{aligned}$$

And we can simplify $u(x)exp(u(x)^T\eta - log(Z(\eta))) = \mathbb{E}[u(x)]$ which gives us:

$$\begin{aligned}
&= \int u(x)^2 exp(u(x)^T\eta - log(Z(\eta)) - u(x)\mathbb{E}[u(x)]exp(u(x)^T\eta - log(Z(\eta))) \\
&= \mathbb{E}[u(x)^2] - \mathbb{E}[u(x)]^2 \\
&= \mathbb{V}[u(x)]
\end{aligned}$$

6. I've never heard of negative Variance, so the covariance matrix must all be positive, meaning it's positive semi-definite, meaning it's convex.

7. In order to calculate the mutual information in terms of the entropy of two random variables, we'll start with the definition for Mutual Information:

$$\begin{aligned}
I &= -\int\int p(x,y) log\frac{p(x)p(y)}{p(x,y)} \\
&= -\int\int p(x,y)\Big[log\frac{p(y)}{p(x,y)} + log(p(x))\Big] \\
&= -\int\int p(x,y) * log\frac{p(y)}{p(x,y)} + \int\int p(x,y)log(p(x)) \\
&= -\int\int p(y)p(x|y) * log(p(x|y)) + \int\int p(x,y)log(p(x)) \\
&= -\int p(y)\int p(x|y) * log(p(x|y)) + \int log(p(x))\int p(x,y) \\
&= -\int p(y) * H(x|y) + \int log(p(x)) * p(x) \\
&= -\int p(y) * H(x|y) \quad + H(x) \\
&= -H(x|y) + H(x) \\
&= H(x) - H(x|y)
\end{aligned}$$

8. (a) Dirichlet:

$$\begin{aligned}
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_k)}\prod_{k=1}^{K}\mu_k^{a_k-1} \\
&= log(\Gamma(\alpha_0)) - \dots + \sum_{k=1}^{K}(a_k - 1)(log(\mu_k)) \\
&= log(\Gamma(\alpha_0)) - \dots + (a_k - 1)^T(log(\mu_k)) \\
&= exp(\eta^T T(x) + log(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_k)}))
\end{aligned}$$

where $\eta = (\vec{a} + \vec{1})$, $T(x) = log(\vec{\mu})$ and $log(Z(\eta)) = log(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_k)})$

(b) Gamma:

$$\begin{aligned} Gam(\lambda|a,b) &= \frac{1}{\Gamma(a)} b^a \lambda^{a-1} exp(-b\lambda) \\ &= exp(log(\frac{1}{\Gamma(a)} b^a \lambda^{a-1} exp(-b\lambda))) \end{aligned}$$

After taking the log, we have an idea that based on the function values, we have $\eta(\Theta) = \frac{1}{\Gamma(a)}$, $T(x) = b^a \lambda^{a-1}$. After simplifying and rearranging terms we get:

$$exp(-b\lambda + a * log(b) + (a-1)log(\lambda) - log(\Gamma(a)))$$

from which we can deduce that $\eta(a, \lambda) = (-b, a)^T$ and $T(b) = (b, log(b))^T$

(c)

9.

10. Yes.

11. (a)

 (b)

 (c)

 (d)