

Problem Set 4 Solutions

Calvin Walker

1.

(a) Letting $\theta = (\mu, \sigma^2)$

$$\begin{aligned}\theta^* &= \arg \max_{\theta} L(\theta : D) \\ &= \arg \max_{\theta} \prod_m^M \mathcal{N}(x^{(m)}; \theta) \\ &= \arg \min_{\theta} \sum_m -\log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(x^{(m)} - \mu)^2}{2\sigma^2} \right) \right)\end{aligned}$$

Minimizing with respect to μ and σ^2 :

$$\frac{\partial}{\partial \mu} = 0 = M\mu - \sum_m x^{(m)} \qquad \frac{\partial}{\partial \sigma^2} = 0 = \frac{-M}{\sigma^2} + \frac{1}{\sigma^4} \sum_m (x^{(m)} - \mu)^2$$

So we have

$$\mu_{MLE} = \frac{1}{M} \sum_m x^{(m)} \qquad \sigma_{MLE}^2 = \frac{1}{M} \sum_m (x^{(m)} - \mu_{MLE})^2$$

(b)

$$\begin{aligned}P(\mu_x | D) &\propto P(\mu_x) P(D | \mu_x) \\ &\propto \exp \left(\frac{-\lambda_{\mu_x} (\mu_x - \mu_{\mu_x})^2}{2} \right) \exp \left(\frac{-\lambda_x}{2} \sum_m (x^{(m)} - \mu_x)^2 \right) \\ &= \exp \left(-\frac{\lambda_{\mu_x}}{2} (\mu_x^2 - 2\mu_x \mu_{\mu_x} + \mu_{\mu_x}^2) - \frac{\lambda_x}{2} \sum_m (x^{(m)2} - 2x^{(m)} \mu_x + \mu_x^2) \right) \\ &= \exp \left(-\frac{\mu_x^2}{2} (\lambda_{\mu_x} + M\lambda_x) + \mu_x (\lambda_{\mu_x} \mu_{\mu_x} + \lambda_x \sum_m x^{(m)}) - \frac{1}{2} (\lambda_{\mu_x} \mu_{\mu_x}^2 + \lambda_x \sum_m x^{(m)2}) \right) \\ &= \exp \left(\frac{-\lambda'_{\mu_x}}{2} (\mu_x^2 - 2\mu_x \mu'_{\mu_x} + \mu_{\mu_x}'^2) \right)\end{aligned}$$

Which is of the form $\mathcal{N}(\mu_x; \mu'_{\mu_x}, (\lambda'_{\mu_x})^{-1})$, where:

$$\begin{aligned}-\frac{1}{2} \lambda'_{\mu_x} \mu_x^2 &= -\frac{\mu_x^2}{2} (\lambda_{\mu_x} + M\lambda_x) \\ \lambda'_{\mu_x} &= \lambda_{\mu_x} + M\lambda_x\end{aligned}$$

and

$$\begin{aligned}-\lambda'_{\mu_x} \mu_x \mu'_{\mu_x} &= \mu_x (\mu_{\mu_x} \lambda_{\mu_x} + \lambda_x \sum_m x^{(m)}) \\ \mu'_{\mu_x} &= \frac{\lambda_{\mu_x}}{\lambda'_{\mu_x}} \mu_{\mu_x} + \frac{M\lambda_x}{\lambda'_{\mu_x}} \mathbb{E}_D[x]\end{aligned}$$

(c) (i)

$$\begin{aligned}P(\lambda_x | D, \mu) &\propto P(\lambda_x | \mu) P(D | \lambda_x, \mu) \\ &\propto \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_x^\alpha \exp(-\beta \lambda_x) \prod_m \left(\lambda_x^{1/2} \exp \left(-\frac{\lambda_x}{2} (x^{(m)} - \mu)^2 \right) \right) \\ &\propto \lambda_x^{\alpha + \frac{1}{2}M} \exp \left(-\beta \lambda_x - \frac{1}{2} \lambda_x \sum_m (x^{(m)} - \mu)^2 \right) \\ &= \lambda_x^{\alpha + \frac{1}{2}M} \exp \left(-\left(\beta + \frac{1}{2} \sum_m (x^{(m)} - \mu)^2 \right) \lambda_x \right)\end{aligned}$$

Which is of the form $\text{Gamma}(\alpha', \beta')$, with $\alpha' = \alpha + \frac{1}{2}M$ and $\beta' = \beta + \frac{1}{2} \sum_m (x^{(m)} - \mu)^2$

(ii) Since $P(\lambda_x|D, \mu) \sim \text{Gamma}(\alpha', \beta')$, we have

$$\mathbb{E}[\lambda_x] = \frac{\alpha + \frac{1}{2}M}{\beta + \frac{1}{2} \sum_m (x^m - \mu)^2} \quad \text{Var}[\lambda_x] = \frac{\alpha + \frac{1}{2}M}{(\beta + \frac{1}{2} \sum_m (x^m - \mu)^2)^2}$$

...

(d) (i)

$$\mathcal{N}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; 0, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \propto \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$$

Which we can express in terms of the Schur complement as:

$$\begin{aligned} &= \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) \\ &= \exp\left(-\frac{1}{2}(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2)^T(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2)\right) \exp\left(-\frac{1}{2}x_2^T\Sigma_{22}^{-1}x_2\right) \end{aligned}$$

We can see that this takes the form of the product of the conditional $P(x_1|x_2)$ and the marginal $P(x_2)$, so

$$\begin{aligned} \exp\left(-\frac{1}{2}(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2)^T(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2)\right) &= \exp\left(-\frac{1}{2}(x_1 - \mu_{1|2})^T\Sigma_{1|2}^{-1}(x_1 - \mu_{1|2})\right) \\ &= P(x_1|x_2) \end{aligned}$$

Where $\mu_{1|2} = \Sigma_{12}\Sigma_{22}^{-1}x_2$ and $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

(ii) We can see from the Schur complement that:

$$\Lambda = \Sigma^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & \dots \\ \dots & \dots \end{bmatrix} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

and from (i) we know that $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Thus,

$$\text{Var}(x_1|x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1}$$

(iii) We can see in part (i) that if $\Lambda_{12} = 0$, in the canonical parameterization, the first term becomes only a function of x_1 , so the distribution decomposes as $P(x_1)P(x_2)$.

(e)

$$\begin{aligned} L(\Lambda : D) &= \prod_m 2\pi^{n/2} |\Lambda|^{1/2} \exp\left(-\frac{1}{2}x^{(m),T}\Lambda x^{(m)}\right) \\ &= 2\pi^{Mn/2} |\Lambda|^{M/2} \prod_m \exp\left(-\frac{1}{2}x^{(m),T}\Lambda x^{(m)}\right) \end{aligned}$$

Since $x^T A x = \text{tr}(x x^T A)$,

$$L(\Lambda : D) = 2\pi^{Mn/2} |\Lambda|^{M/2} \exp\left(\text{tr}\left(-\frac{1}{2} \sum_m x^{(m)} x^{(m),T} \Lambda\right)\right)$$

Taking the log,

$$\begin{aligned} \ell(\Lambda : D) &= -\frac{Mn}{2} \log 2\pi + \frac{M}{2} \log |\Lambda| - \frac{M}{2} \text{tr}\left(\frac{1}{M} D^T D \Lambda\right) \\ &\propto \log \det(\Lambda) - \text{tr}(S\Lambda) \end{aligned}$$

(f) The maximum likelihood estimate for Λ is given by:

$$\Lambda^* = \arg \min_{\Lambda} \text{tr}(S\Lambda) - \log \det(\Lambda)$$

So,

$$\begin{aligned} \frac{\partial \ell}{\partial \Lambda} &= 0 = S - \Lambda^{-1} \\ \Lambda_{MLE} &= S^{-1} \end{aligned}$$

2.

(a)

(b)

$$\begin{aligned} H(X | \mathbf{X} - X) &= - \sum_{X, \mathbf{X} - X} P(X, \mathbf{X} - X) \log P(X | \mathbf{X} - X) \\ &= - \sum_{X, \mathbf{X} - X - \text{MB}(X), \text{MB}(X)} P(X, \mathbf{X} - X - \text{MB}(X), \text{MB}(X)) \log P(X | \mathbf{X} - X - \text{MB}(X), \text{MB}(X)) \end{aligned}$$

Where $P(X | \mathbf{X} - X - \text{MB}(X), \text{MB}(X)) = P(X | \text{MB}(X))$ and (by chain rule),

$$\begin{aligned} P(X, \mathbf{X} - X - \text{MB}(X), \text{MB}(X)) &= P(X | \mathbf{X} - X - \text{MB}(X), \text{MB}(X)) P(\mathbf{X} - X - \text{MB}(X) | \text{MB}(X)) P(\text{MB}(X)) \\ &= P(X | \text{MB}(X)) P(\mathbf{X} - X - \text{MB}(X) | \text{MB}(X)) P(\text{MB}(X)) \end{aligned}$$

So,

$$\begin{aligned} H(X | \mathbf{X} - X) &= - \sum_{X, \mathbf{X} - X - \text{MB}(X), \text{MB}(X)} P(X | \text{MB}(X)) P(\mathbf{X} - X - \text{MB}(X) | \text{MB}(X)) P(\text{MB}(X)) \log P(X | \text{MB}(X)) \\ &= - \sum_{X, \text{MB}(X)} P(X, \text{MB}(X)) \log P(X | \text{MB}(X)) \\ &= H(X | \text{MB}(X)) \end{aligned}$$

- (c) Since $H(A | B, C) \leq H(A | B)$ for disjoint sets A, B, C , we can minimize the conditional entropy $H(X | Y)$ by conditioning on the rest of the nodes. So $H(X | \mathbf{X} - X) = H(X | \mathbf{X} - X - \text{MB}(X) \text{MB}(X)) = H(X | \text{MB}(X))$ minimizes the conditional entropy. Thus, $\text{MB}(X) = \arg \min_Y H(X | Y)$.
- (d) For each node X , compute its Markov Blanket as $\text{MB}(X) = \arg \min_Y H(X | Y)$. We have n nodes, and there are $\binom{n}{d}$ possible Markov Blankets for each node, the conditional entropy of which takes c complexity to compute. So the algorithm runs in $O(n \binom{n}{d} c)$.

3.

(a)