AI 501 Machine Learning for AI, Spring 2021

Final Exam

1. (60 points) Consider the Probabilistic Principal Component Analysis (PPCA) model with generative process defined as

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}_h, \boldsymbol{I}_h), \quad \boldsymbol{x}|\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d), \tag{1}$$

where $\boldsymbol{z} \in \mathbb{R}^h$, $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{W} \in \mathbb{R}^{d\times h}$, and $\boldsymbol{\mu} \in \mathbb{R}^d$. Let $\boldsymbol{\theta} := \{\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2\}$ be the collection of parameters.

(a) (20 points) Show that the conditional distribution $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})$ and the marginal distribution $p(\boldsymbol{x};\boldsymbol{\theta})$ are given as follows:

$$p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{V}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}-\boldsymbol{\mu}), \sigma^2\boldsymbol{V}^{-1}), \quad p(\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}_d), \tag{2}$$

where

$$\boldsymbol{V} = \boldsymbol{W}^\top\boldsymbol{W} + \sigma^2\boldsymbol{I}_h. \tag{3}$$

Hint: you may use the following matrix inverse identity:

$$(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{A})^{-1}\boldsymbol{B}. \tag{4}$$

(b) (10 points) Let $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a training set. Using the expression of $p(\boldsymbol{x};\boldsymbol{\theta})$ obtained above, compute the maximum-likelihood estimator of the parameter $\boldsymbol{\mu}$.

(c) (30 points) Now consider the Expectation-Maximization (EM) procedure to learn parameters $(\boldsymbol{W}, \sigma^2)$ via maximum likelihood while holding $\boldsymbol{\mu}$ fixed. Let $\boldsymbol{\theta}_t = (\boldsymbol{W}_t, \boldsymbol{\mu}, \sigma_t^2)$ be the parameter estimate at step t. Show that the M-step update equations are given as follows:

$$\boldsymbol{W}_{t+1} = \boldsymbol{S}\boldsymbol{W}_t(\sigma_t^2\boldsymbol{I}_h + \boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\boldsymbol{W}_t)^{-1}, \tag{5}$$

$$\sigma_{t+1}^2 = \frac{1}{d}\mathrm{Tr}(\boldsymbol{S} - \boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}_{t+1}^\top), \tag{6}$$

where

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top, \quad \boldsymbol{V}_t = \boldsymbol{W}_t^\top\boldsymbol{W}_t + \sigma_t^2\boldsymbol{I}_h. \tag{7}$$

Hint: you may use the following identity,

$$\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{y} = \mathrm{Tr}(\boldsymbol{A}\boldsymbol{y}\boldsymbol{x}^\top). \tag{8}$$

You can also refer to the matrix cookbook for the computation of derivatives.

2. (40 points) Consider the Mixture of Probabilistic Principal Component Analysers (MPPCA) defined as follows:

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}_h, \boldsymbol{I}_h), \quad c \sim \mathrm{Cat}(\boldsymbol{\pi}), \quad \boldsymbol{x}|\boldsymbol{z}, c \sim \mathcal{N}(\boldsymbol{W}_c\boldsymbol{z} + \boldsymbol{\mu}_c, \sigma_c^2\boldsymbol{I}_d), \tag{9}$$

where $\boldsymbol{z} \in \mathbb{R}^h$, $\boldsymbol{x} \in \mathbb{R}^d$, $c \in \{1, \ldots, k\}$. Let $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \{\boldsymbol{W}_j, \boldsymbol{\mu}_j, \sigma_j^2\}_{j=1}^k\}$ be the parameters with $\boldsymbol{W}_j \in \mathbb{R}^{d\times h}$, $\boldsymbol{\mu}_j \in \mathbb{R}^h$, and $\boldsymbol{\pi} \in [0, 1]^k$.

(a) (10 points) Compute the distributions $p(c|\boldsymbol{x};\boldsymbol{\theta})$, $p(\boldsymbol{z}|c,\boldsymbol{x};\boldsymbol{\theta})$ and $p(\boldsymbol{x};\boldsymbol{\theta})$.

(b) (15 points) Let $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a training set to fit MPPCA. Let us first consider the parameters $\boldsymbol{\pi}$ and $\{\boldsymbol{\mu}_j\}_{j=1}^k$. Given a current parameter estimate $\boldsymbol{\theta}_t = \{\{\boldsymbol{W}_{t,j}, \boldsymbol{\mu}_{t,j}, \sigma_{t,j}^2\}_{j=1}^k, \boldsymbol{\pi}_j\}$, we first compute the expected complete-data log-likelihood *only* for c as follows:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_{i=1}^n \mathbb{E}_q[\log p(\boldsymbol{x}_i, c_i; \boldsymbol{\theta})] \\
&= \sum_{i=1}^n \mathbb{E}_q\left[\log \prod_{j=1}^k \Big(\pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{W}_j\boldsymbol{W}_j^\top + \sigma_j^2 \boldsymbol{I}_d)\Big)^{\mathbb{1}\{c_i=j\}}\right] \\
&= \sum_{i=1}^n \sum_{j=1}^k r_{t,i,j}\Big(\log \pi_j + \log \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{W}_j\boldsymbol{W}_j^\top + \sigma_j^2 \boldsymbol{I}_d)\Big), \qquad (10)
\end{aligned}$$

where

$$r_{t,i,j} := \frac{\pi_{t,j}\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{W}_{t,j}\boldsymbol{W}_{t,j}^\top + \sigma_{t,j}^2\boldsymbol{I}_d)}{\sum_{\ell=1}^k \pi_{t,\ell}\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_\ell, \boldsymbol{W}_{t,\ell}\boldsymbol{W}_{t,\ell}^\top + \sigma_{t,\ell}^2\boldsymbol{I}_d)}. \qquad (11)$$

Show that the M-step update for $\boldsymbol{\pi}$ and $\{\mu_j\}_{j=1}^k$ are given as

$$\pi_{t+1,j} = \frac{\sum_{i=1}^n r_{t,i,j}}{n}, \quad \boldsymbol{\mu}_{t+1,j} = \frac{\sum_{i=1}^n r_{t,i,j}\boldsymbol{x}_i}{\sum_{i=1}^n r_{t,i,j}}. \qquad (12)$$

(c) (15 points) Consider the parameters $\{\boldsymbol{W}_j, \sigma_j\}_{j=1}^k$. While holding $\boldsymbol{\pi}, \{\boldsymbol{\mu}_j\}_{j=1}^k$ fixed, we would like to apply M-step with the expected complete-data log-likelihood with both $\boldsymbol{z}_i$ and c_i,

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbb{E}_q[\log p(\boldsymbol{x}_i, \boldsymbol{z}_i, c_i; \boldsymbol{\theta})], \qquad (15)$$

where

$$q(\boldsymbol{z}_i, c_i; \boldsymbol{\theta}_t) = p(\boldsymbol{z}_i|c_i, \boldsymbol{x}_i; \boldsymbol{\theta}_t)p(c_i|\boldsymbol{x}_i; \boldsymbol{\theta}_t). \qquad (16)$$

Derive the M-step update equations for the parameters $\{\boldsymbol{W}_j, \sigma_j\}_{j=1}^k$.