UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE : **COMP0086**

ASSESSMENT PATTERN : **COMP0086A7PA**

MODULE NAME : **Probabilistic and Unsupervised Learning**

LEVEL: : **Postgraduate**

DATE : **02 May 2019**

TIME : **14:30**

TIME ALLOWED : **2 hrs 30 mins**

This paper is suitable for candidates who attended classes for this module in the following academic year(s):

Year
2018-19

EXAMINATION PAPER CANNOT BE REMOVED FROM THE EXAM HALL. PLACE EXAM PAPER AND ALL COMPLETED SCRIPTS INSIDE THE EXAMINATION ENVELOPE

Hall Instructions	.
Standard Calculators	Y
Non-Standard Calculators	N

TURN OVER

Probabilistic and Unsupervised Learning, COMP0086 (A7P)

Main Summer Examination Period, 2018/19

This examination contains SIX questions, of which you should answer FIVE.

Each question is worth 20 marks. If you have time, you may attempt all 6 questions, but only your 5 best scores will count towards your final mark. Make sure you manage your time carefully, read through ALL questions first before attempting any, and don't spend too much time on any one question. Calculators are allowed, but are unlikely to be useful.

Always provide justification and show any intermediate work for your answers. A correct but unsupported answer may not receive any marks.

In the following, bold symbols represent vectors. When asked to provide algorithms, you may provide equations or use MATLAB/Octave notation for matrix algebra. You may assume the availability of standard matrix functions such as `inv`, `pinv`, `eig` or `svd`. You may **not** assume statistical functions, including `mean` and `cov`. If you choose to use MATLAB notation, you will not lose marks for syntactic errors, as long as your intention is clear and correct.

You might find it useful to consult the following table of standard distributions:

Name	Domain	Symbol	Density or Probability fn
Gaussian (Normal)	$\mathbb{R}$	$x \sim \mathcal{N}\left(\mu, \sigma^2\right)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$
Multivariate Normal	$\mathbb{R}^D$	$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \Sigma\right)$	$\lvert 2\pi\Sigma\rvert^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$
Bernoulli	$\{0, 1\}$	$x \sim \mathsf{Bernoulli}(p)$	$p^x(1-p)^{1-x}$
Binomial	$\mathbb{Z}_{0-N}$	$x \sim \mathsf{Binom}(p)$	$\binom{N}{x} p^x (1-p)^{N-x}$
Multinomial	$[\mathbb{Z}_{0-N}]^D$	$\mathbf{x} \sim \mathsf{Multinom}(\mathbf{p})$	$\frac{N!}{x_1!\, x_2! \ldots x_D!} \prod_{d=1}^{D} p_d^{x_d}$
Poisson	$\mathbb{Z}_{0+}$	$x \sim \mathsf{Poisson}(\mu)$	$\mu^x e^{-\mu}/x!$
Beta	$[0, 1]$	$x \sim \mathsf{Beta}(\alpha, \beta)$	$\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Exponential	$\mathbb{R}_+$	$x \sim \mathsf{Exp}(\lambda)$	$\lambda e^{-\lambda x}$
Gamma	$\mathbb{R}_+$	$x \sim \mathsf{Gamma}(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
Student-t	$\mathbb{R}$	$x \sim \mathsf{Student\text{-}t}(\alpha)$	$\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\pi\alpha}\Gamma\left(\frac{\alpha}{2}\right)} \left(1 + \frac{x^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}$
Inverse-Wishart	$\mathbb{S}_+^p$	$X \sim \mathsf{InvWish}(\Psi, \nu)$	$\frac{\lvert\Psi\rvert^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}} \Gamma_p\left(\frac{\nu}{2}\right)} \lvert X\rvert^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\mathsf{Tr}\left[\Psi X^{-1}\right]}$
Dirichlet	$[0, 1]^D$	$\mathbf{x} \sim \mathsf{Dirichlet}(\boldsymbol{\alpha})$	$\frac{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} x_d^{\alpha_d - 1}$

1. **Conjugacy in regression**

Consider linear regression for input-output pairs $(\mathbf{x}_n, y_n)$, $n = 1 \dots N$ with weight vector $\mathbf{w}$ and output noise variance σ^2:

$$y_n \sim \mathcal{N}\left(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2\right)$$

(a) What is a conjugate prior? *[2 marks]*

(b) State the form of the conjugate prior for $\mathbf{w}$, assuming that σ^2 is known. You do not need to derive the result. *[2 marks]*

Now consider the case where σ^2 is unknown and is estimated along with $\mathbf{w}$. We might choose an inverse-gamma prior on σ^2:

$$p(\sigma^2) = \mathsf{InvGamma}(\sigma^2; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\beta/\sigma^2}$$

(c) Find the joint posterior on σ^2 and $\mathbf{w}$ under independent priors on $\mathbf{w}$ (of the form you identified above) and inverse-gamma on σ^2. Your notation may be simpler if you introduce a matrix $X = [\mathbf{x}_1 \dots \mathbf{x}_N]$ and row vector $Y = [y_1 \dots y_N]$. Simplify your expression (discarding constants of proportionality) to isolate the functional dependence on σ^2 and $\mathbf{w}$ as far as possible. *[6 marks]*

Your result to the previous part should have shown that the independent prior defined above is *not* conjugate. The correct conjugate prior is a joint distribution with the form:

$$p(\sigma^2, \mathbf{w}) = \mathsf{InvGamma}(\sigma^2; \alpha, \beta)\, \mathcal{N}\left(\mathbf{w}; \boldsymbol{\mu}, \sigma^2 A^{-1}\right)$$

You may take $\boldsymbol{\mu} = 0$.

(d) Show that this is indeed a conjugate prior, and derive the updates to the parameters α, β, $\boldsymbol{\mu}$ and A induced by the likelihood. *[8 marks]*

(e) In lecture we considered Bayesian learning of parameters within A (for example, the diagonal entries) by maximising the evidence or marginal likelihood obtained by integrating out $\mathbf{w}$ alone, while maximising over σ^2. Do you think the results for these parameters would have been different if we had integrated out both $\mathbf{w}$ and σ^2? You do not need to derive the result – just argue whether it will be the same or different based on your results here. *[2 marks]*

2. **A mixture of linear experts**

A mixture of experts is a latent variable model for supervised learning. The "experts" are each simple probabilistic supervised learning models. Here, we consider the case where each expert corresponds to vector-output linear regression with Gaussian noise.

We are given input-output vector pairs $(\mathbf{x}_n, \mathbf{y}_n), n = 1 \ldots N$. For each pair, we introduce a discrete latent "gating" variable $s_n \in \{1 \ldots K\}$, with the probability distribution for s_n determined by the distance of the corresponding input $\mathbf{x}_n$ from each of K points $\boldsymbol{\mu}_k$. The kth linear expert has weight matrix W_k, but all share the same diagonal output noise covariance Ψ.

The joint model can be written:

$$P(s_n = k|\mathbf{x}_n, \Theta) = \pi_k(\mathbf{x}_n) = \frac{e^{-\frac{1}{2}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2}}{\sum_{k'=1}^{K} e^{-\frac{1}{2}\|\mathbf{x}_n - \boldsymbol{\mu}_{k'}\|^2}}$$

$$P(\mathbf{y}_n|s_n = k, \mathbf{x}_n, \Theta) = \mathcal{N}(W_k\mathbf{x}_n, \Psi)$$

where the parameters are $\Theta = \{\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_K, W_1 \ldots W_K, \Psi\}$.

We will derive an Expectation Maximisation (EM) algorithm to find the maximum-likelihood parameters of this model.

(a) Show that $P(\mathbf{y}_n, s_n|\mathbf{x}_n, \Theta)$ is a joint exponential family model by expressing it in the standard exponential family form, identifying the sufficient statistic functions, and giving the natural parameters as functions of $\mathbf{x}_n$ and the conventional parameters defined above. *[6 marks]*

(b) Find the posterior distribution $P(s_n = k|\mathbf{x}_n, \mathbf{y}_n, \Theta)$. *[4 marks]*

(c) Writing r_{nk} for the posterior probability defined in part (b), derive closed-form M-step updates for:

- W_k *[4 marks]*
- $\boldsymbol{\mu}_k$ *[6 marks]*

3. **Free-energy derivatives**

Consider a general latent variable model with observed data $\mathcal{X}$, latent variables $\mathcal{Z}$ and parameters θ. The log-likelihood function is given by

$$\ell(\theta) = \log \int d\mathcal{Z}\ P(\mathcal{X}, \mathcal{Z}|\theta)$$

(a) How is the free energy defined? Show that it is a lower bound on the log likelihood. *[4 marks]*

(b) Show that the (generalised) M-step gradient evaluated after an exact E-step is (under some regularity conditions) equal to the gradient of log likelihood with respect to the parameters. *[6 marks]*

(c) Consider a mixture of spherical Gaussians:

$$p(z_n) = \mathsf{Discrete}(\boldsymbol{\pi})$$
$$p(\mathbf{x}_n|z_n) = \prod_k \left(\mathcal{N}\left(\boldsymbol{\mu}_k, I\right)\right)^{\delta_{z_n,k}}$$

where $z_n \in 1 \ldots K$ and $\delta_{z_n,k}$ is an indicator taking the value 1 if $z_n = k$ and 0 otherwise.

Compute the exact M-step update for the parameter $\boldsymbol{\mu}_k$, as well as the gradient and show that the exact step changes $\boldsymbol{\mu}_k$ along the direction of the gradient. *[6 marks]*

(d) Is this generally true? That is, is the change in parameters resulting from an exact (i.e. non-generalised) M-step always aligned with the gradient direction? Explain. *[4 marks]*

4. **Protocol sniffing.**

A secure network router handles 5 different address encryption protocols. The time taken to handle the ith incoming packet (and send it on to the next gateway) is made up of two parts:

- the routing time τ_{p_i} which depends on the protocol p_i that the current packet employs
- and a protocol-dependent startup time σ_{0,p_1} for the first packet following a reset, and a state-switching time σ_{p_{i-1},p_i} for $i \geq 2$, which depends on both the current protocol (p_i) and the protocol employed by the previous packet (p_{i-1}).

By monitoring radio-frequency emissions from the router you are able to obtain the times at which outgoing packets are sent, the differences between which (t_i) correspond to the processing delays above, plus an exponentially-distributed random wait for the next incoming packet (with parameter λ).

You wish to use the sequence $\{t_i\}_{i=1}^N$ to infer the packet protocols $\{p_i\}_{i=1}^N$. Assume that the packet protocols are *a priori* independent, with frequencies given by the probability vector $\boldsymbol{\pi}$.

(a) Draw the directed graphical model that relates the random variables $\{p_i\}$, the measured delays $\{t_i\}$, and the parameters $\boldsymbol{\pi}$, $\boldsymbol{\tau} = [\tau_k]$, $\Sigma = [\sigma_{k,k'}]$ and λ. *[3 marks]*

(b) Consider the moralised graph on the latents $\{p_i\}$ and write down expressions for the corresponding clique potentials. *[5 marks]*

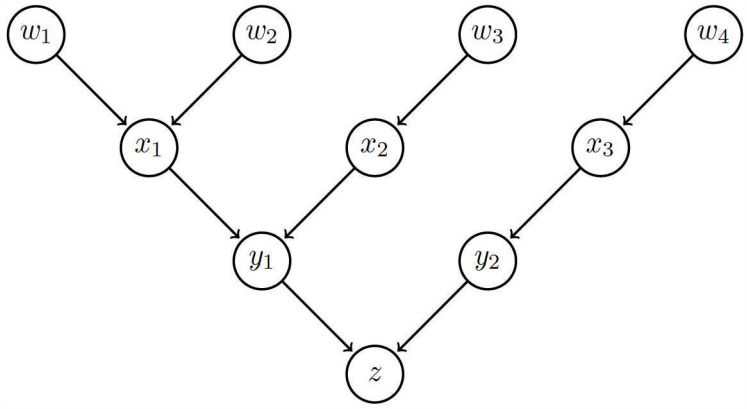
(c) This posterior factorisation is the same as would be obtained for state inference in a hidden Markov model (HMM). However, there is no equivalent standard HMM with state p_i and observations t_i that would yield the same posterior. Use the ideas of conditional independence to show why. *[2 marks]*

(d) However, it *is* possible to construct an equivalent HMM by augmenting the state. Show how, giving the corresponding transition matrix and output distribution. *[3 marks]*

(e) Derive an efficient (i.e. order N) algorithm to estimate from $\{t_i\}$ the most likely sequence of protocols that the router has handled. *[7 marks]*

5. **An inverted tree**

Consider the directed acyclic graphical (DAG) model below:



(a) List the (conditional) independence statements satisfied by the variables w_1, w_2, w_3, w_4 when:

i. no observations are made

ii. x_1 is observed

iii. y_2 is observed

iv. z is observed

[5 marks]

(b) Draw the moralised graph for the case that y_1 is observed and construct the minimal junction tree. *[5 marks]*

(c) Why is the minimal junction tree unique in this case? What is the smallest number of edges you would need to add to, or remove from, the DAG so that more than one minimal junction tree exists? [As always, justify your answer.] *[4 marks]*

(d) Describe the steps needed to obtain the posterior marginal distributions $p(w_3|y_1)$ and $p(w_4|y_1)$. What is the minimal number of messages that must be passed in each case? *[6 marks]*

6. **GPFA – Gaussian processes for unsupervised learning**

Consider a model for multivariate time series based on a set of latent functions of time $z_k(t)$, each drawn from an independent Gaussian process, which are then linearly combined with Gaussian observation noise to generate the observations:

$$z_k(\cdot) \sim \mathcal{GP}(\mu_k(\cdot), \kappa_k(\cdot,\cdot)) \qquad k = 1 \ldots K$$

$$x_d(t) \sim \mathcal{N}\left(\sum_k \Lambda_{dk} z_k(t), \psi_d\right) \qquad d = 1 \ldots D$$

This model has sometimes been called Gaussian-process factor analysis or GPFA.

Suppose we have observations of $\mathbf{x}(t_n) = [x_1(t_n) \ldots x_D(t_n)]^\mathsf{T}$ at a set of N times $t_n \in \{t_1 \ldots t_N\}$. We collect these into a $D \times N$ matrix X, with $X_{dn} = x_d(t_n)$. Define also a $K \times N$ matrix Z of latent variables with $Z_{kn} = z_k(t_n)$. In answering the questions below you may find it helpful to define further matrices assembled from the variables or parameters in the model. Make sure you define these clearly and consistently.

(a) Find the posterior distribution on Z. *[3 marks]*

(b) Give the expectation-maximisation M-step needed to learn the maximum likelihood value of Λ, assuming that all the μ_k and κ_k are known. *[3 marks]*

(c) Consider a new time t^*. Compute the predictive distribution $P(\mathbf{x}(t^*)|X)$. *[4 marks]*

(d) Suppose that we only observed a single output process at each time t_n. That is, rather than all of $\mathbf{x}(t_n)$ we have only some $x_{d_n}(t_n)$. Assume that the identity of the particular dimension d_n that is observed does not depend on any latent z. Describe an approach to estimate the posterior on Z, maximum-likelihood $\widehat{\Lambda}$ and predictive $P(\mathbf{x}(t^*)|\mathcal{X})$ (where $\mathcal{X}$ is the set of observed values) in this case. *[6 marks]*

(e) This model is related to the linear-Gaussian state-space model (LGSSM) that we considered in lecture. Under what parametric constraints (on both GPFA and the LGSSM) are the two equivalent? State conditions that are as general as possible. *[4 marks]*