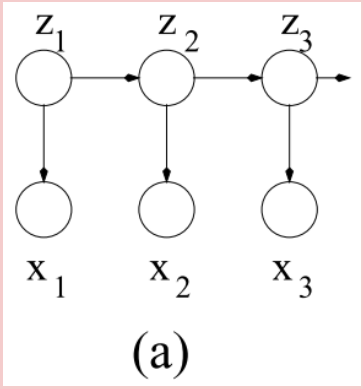
[PMR] exam 2011

---

1 .

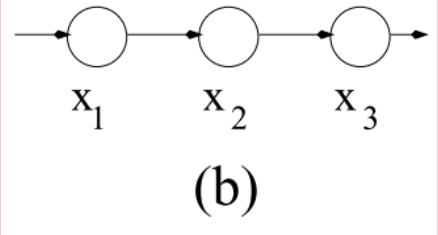
(a)



1. (a) Figure (a) above illustrates a hidden Markov model (HMM) drawn as a graphical model for 3 timeslices. [*6 marks*]

   Describe the structure of the HMM and the parameters that define it. Draw and label a diagram of a HMM as a finite state automaton.

(b)



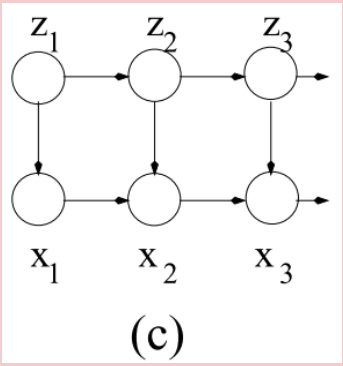
(b) In Figure (b) above, each of the variables $x_1,\ x_2, \ldots$ is a real-valued scalar variable. Together they form a first-order auto-regressive or AR(1) process, so that [*7 marks*]

$$x_n = cx_{n-1} + w_n \qquad \text{for integer } n$$

where c is a scalar constant, and w_n is drawn from a zero-mean Gaussian with variance σ^2, i.e. $w_n \sim N(0, \sigma^2)$. The w's at different timesteps are independent.

Assuming that the sequence of x's is stationary, compute the mean, variance and covariance function for this AR(1) process in terms of c and σ^2. (One way to compute the covariance function is via the Yule-Walker equations.) Also state and explain the condition on c for the process to be stable.

(c)



(c) Figure (c) illustrates a more complex HMM, known as an auto-regressive HMM or AR-HMM. The state z_n $(n = 1,\ 2, \ldots)$ is discrete, taking on M possible values. Under this model $p(x_n|x_{n-1}, z_n)$ is M different AR(1) models (with different parameters), indexed by z_n. [*6 marks*]

Let $\alpha(z_n = i) = p(x_1, x_2, \ldots, x_n, z_n = i)$, as for the standard HMM forward recursion. Derive an expression for $\alpha(z_{n+1} = j)$ in terms of $\alpha(z_n)$ and the parameters of the model for the AR-HMM, justifying any conditional independences used in terms of the graphical model.

(d)

(d) The AR(1) and AR-HMM models used above are both models for the x-sequence. (For the AR-HMM we simply marginalize out the hidden z variables to obtain a model on the x's.) [*6 marks*]

State the number of parameters used to define the AR(1) model and the AR-HMM model. Explain how the models could be compared using maximum likelihood, and the potential problems of this method.

Describe one way by which the problems of using the maximum likelihood method for model selection might be addressed in this case.

2 .

(a)

2. (a) Explain what is meant by a latent variable model (using a diagram if you wish), and why it might be used. [*3 marks*]

(b)

(b) Independent Components Analysis (ICA) is a latent variable model for data. Here we consider the noiseless square case so that $\mathbf{x} = A\mathbf{z}$. Consider the case where $\mathbf{z}$ and $\mathbf{x}$ are two-dimensional, and $p(\mathbf{z}) = \frac{1}{4}\prod_{i=1}^{2}\exp(-|z_i|)$. Here each z_i has zero mean and variance 2. Let the mixing matrix A be given by [*8 marks*]

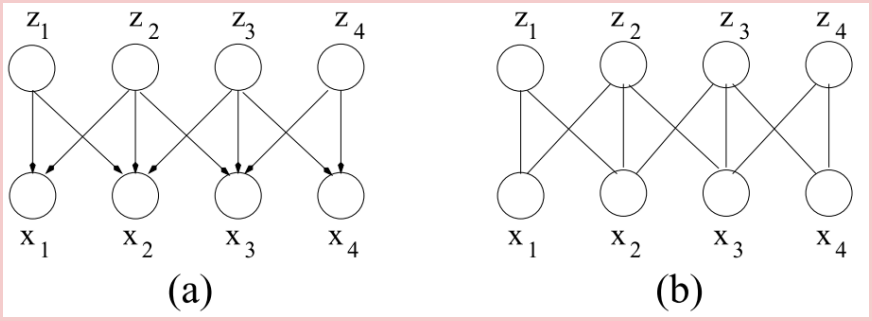
$$A = \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix}.$$

i. Sketch the distribution in $\mathbf{x}$-space, showing the directions of the independent components and the contours of $p(\mathbf{x})$. Explain your reasoning.

ii. Compute the covariance matrix of $\mathbf{x}$ under this model.

iii. List three similarities or differences between ICA and factor analysis.

(c)



(c) The figure above shows (a) a directed and (b) an undirected graphical model with similar structure. For the *directed* model (figure (a)) [*10 marks*]

i. Write down the factorization of the joint distribution implied by the graphical model.

ii. Carry out moralization and (if necessary) triangulation operations on the graph; explain your reasoning. Identify the cliques.

iii. Draw a valid junction tree for the above graph, and explain why it is valid. Define the separators.

iv. Give a suitable initialization of the clique and separator potentials.

(d)

(d) For the *undirected* model (figure (b) above) [*4 marks*]

i. Write down the factorization of the joint distribution implied by the graphical model.

ii. If necessary, carry out triangulation operations on the graph; explain your reasoning. Identify the cliques.

3 .

(a)

3. (a) i. Derive the maximum likelihood estimator for the probabilities $\theta_1, \ldots, \theta_r$ of an r-state multinomial distribution with $\sum_{i=1}^{r} \theta_i = 1$, given observed counts $n_1, \ldots, n_r$ for each of the states. [*8 marks*]

ii. Under the Bayesian methodology, the conjugate prior for the multinomial likelihood is the Dirichlet distribution (as given in the preamble) with parameters $\alpha_1, \ldots, \alpha_r$. Show that the posterior distribution is also a Dirichlet distribution and state its parameters. (You do not need to worry about the normalization constants for the Dirichlet distribution in this question.)

iii. Now consider a non-conjugate prior which is a mixture of three Dirichlet distributions

$$p(\boldsymbol{\theta}) = \pi_1 \text{Dir}(\boldsymbol{\alpha}^{(1)}) + \pi_2 \text{Dir}(\boldsymbol{\alpha}^{(2)}) + \pi_3 \text{Dir}(\boldsymbol{\alpha}^{(3)}),$$

where $\pi_1 + \pi_2 + \pi_3 = 1$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ and $\boldsymbol{\alpha}^{(j)}$ denotes the parameters of the jth Dirichlet distribution. Derive the posterior in this case.

(b)

(b) A Boltzmann machine is an undirected graphical model. Consider a Boltzmann machine where there are m units $x_1, \ldots, x_m$ taking on values ± 1. The energy $E(\mathbf{x})$ of a configuration is given by [*7 marks*]

$$E(\mathbf{x}) = -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} x_i x_j$$

where $w_{ij} = w_{ji}$, and $w_{ii} = 0$ for all $i = 1, \ldots, m$. The probability of a configuration $\mathbf{x}$ is given by

$$p(\mathbf{x}) = \frac{1}{Z} \exp -E(\mathbf{x}),$$

where $Z = \sum_{\mathbf{x}} \exp -E(\mathbf{x})$ is the partition function, and the sum is over all possible states of $\mathbf{x}$.

Given the distribution $p(\mathbf{x})$ as defined above, show that

$$p(x_k = 1 | \mathbf{x}_{-k}) = \frac{1}{1 + e^{-2h_k}}$$

where $\mathbf{x}_{-k}$ denotes the state of all other variables except x_k, and $h_k = \sum_{i=1}^{m} w_{ik} x_i$.

(c)

(c) Consider a multi-class classification problem. You construct a classifier that outputs predictions of $p(C_j|\mathbf{x})$ for all classes $j = 1, \ldots, M$. In some situations it is not necessary to classify all input examples; some can be rejected. Explain how a reject option is implemented, and describe (with the aid of a sketch graph) what an error-reject curve is and how it is computed. [*5 marks*]

(d)

(d) Explain how probability theory can be combined with utility theory to make optimal decisions. Give an example where the loss matrix is asymmetric. [*5 marks*]