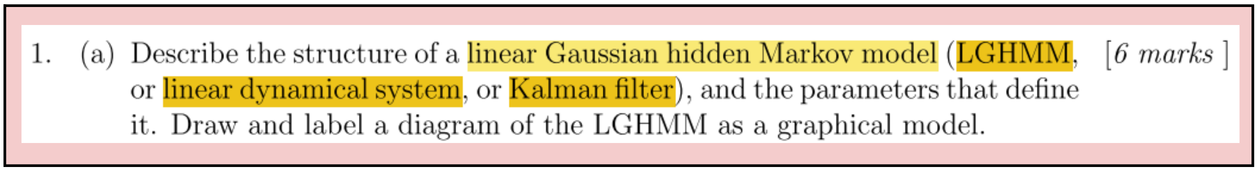
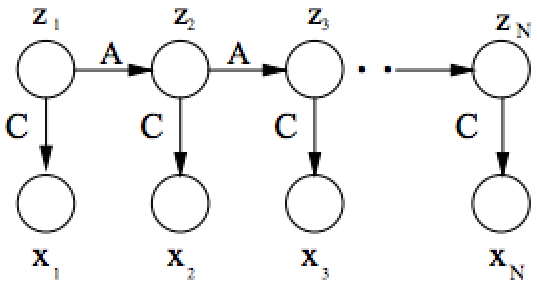
t[PMR] exam 2012 may

1 .

(a) LGHMM / linear dynamic system / Kalman filter



1. (a) Describe the structure of a linear Gaussian hidden Markov model (LGHMM, or linear dynamical system, or Kalman filter), and the parameters that define it. Draw and label a diagram of the LGHMM as a graphical model. [*6 marks*]



- Dynamical model

$$\mathbf{z}_{n+1} = A\mathbf{z}_n + \mathbf{w}_{n+1}$$

  where $\mathbf{w}_{n+1} \sim N(\mathbf{0}, \Gamma)$ is Gaussian noise, i.e.

$$p(\mathbf{z}_{n+1}|\mathbf{z}_n) \sim N(A\mathbf{z}_n, \Gamma)$$

- Observation model

$$\mathbf{x}_n = C\mathbf{z}_n + \mathbf{v}_n$$

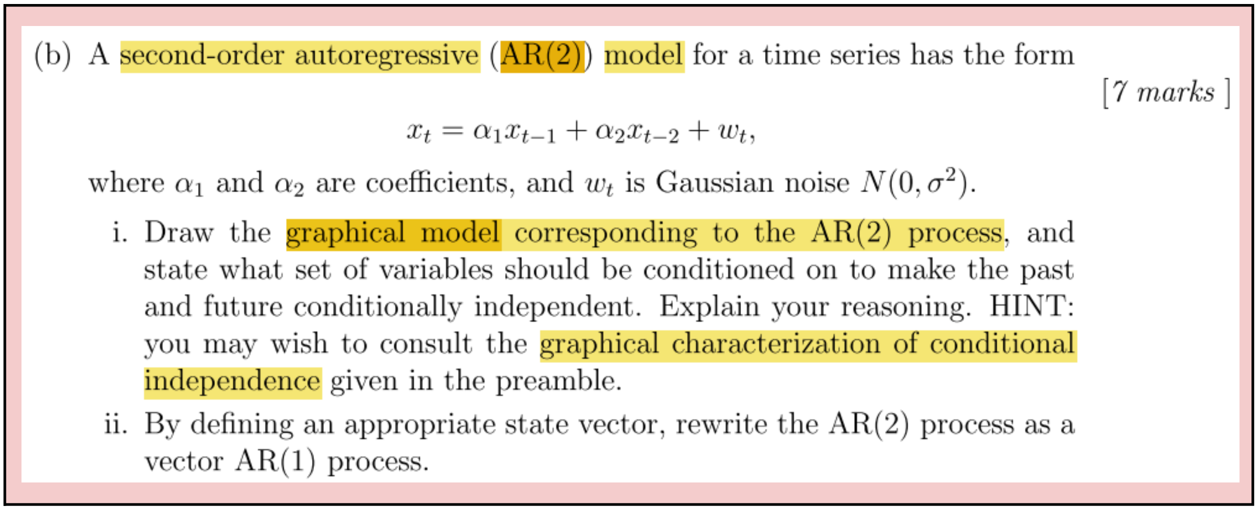
  where $\mathbf{v}_n \sim N(\mathbf{0}, \Sigma)$ is Gaussian noise, i.e.

$$p(\mathbf{x}_n|\mathbf{z}_n) \sim N(C\mathbf{z}_n, \Sigma)$$

- Initialization

$$p(\mathbf{z}_1) \sim N(\mu_0, V_0)$$

(b) Second-order autoregressive (AR(2)) model



(b) A second-order autoregressive (AR(2)) model for a time series has the form

[7 marks]

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + w_t,$$

where α_1 and α_2 are coefficients, and w_t is Gaussian noise $N(0, \sigma^2)$.

i. Draw the graphical model corresponding to the AR(2) process, and state what set of variables should be conditioned on to make the past and future conditionally independent. Explain your reasoning. HINT: you may wish to consult the graphical characterization of conditional independence given in the preamble.

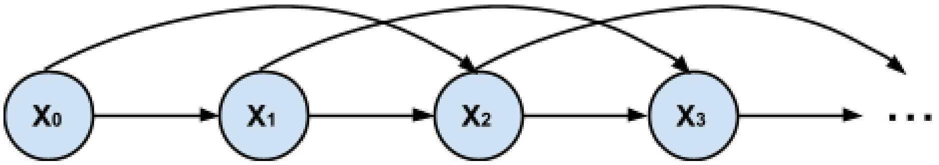
ii. By defining an appropriate state vector, rewrite the AR(2) process as a vector AR(1) process.

Kshitij, Frederico, Tim:

(i)

note:

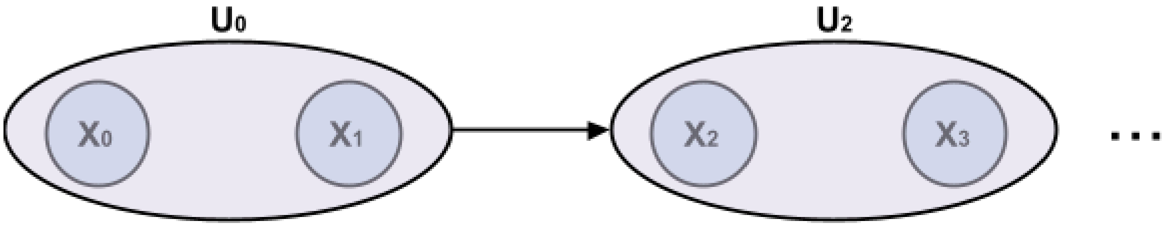
*we are dealing with a **second-order markov chain** because there are no hidden variables*



$$I(X_t, X_{t-3:0} \mid X_{t-2}, X_{t-1})$$

...read $I(A, B \mid C)$: A independent of B given C

(ii)

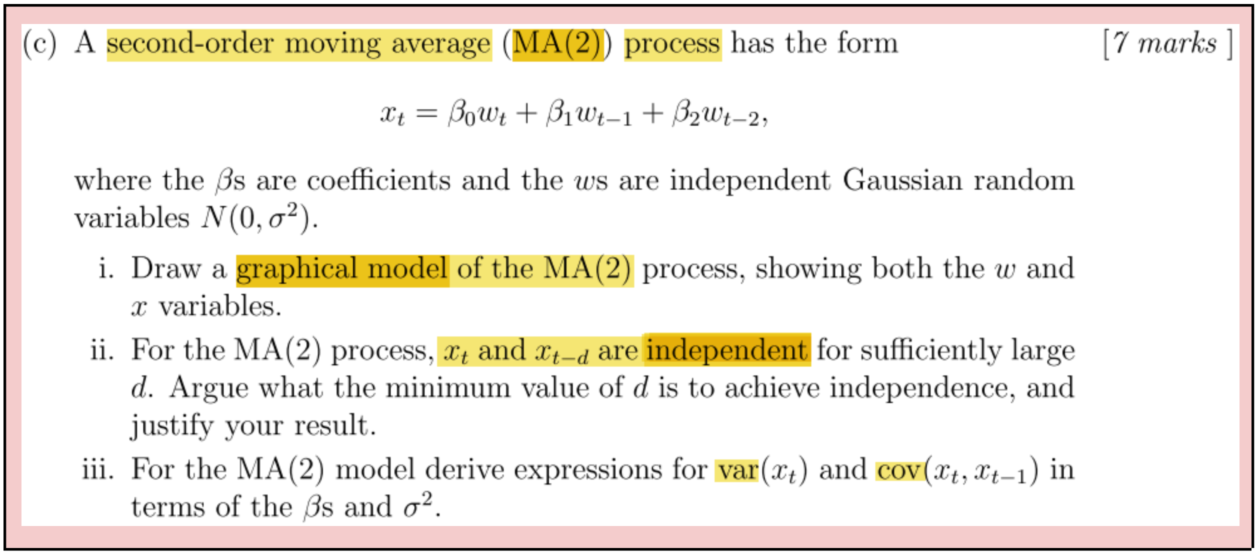


$$U_t = \begin{pmatrix} X_{2t} \\ X_{2t+1} \end{pmatrix} = \underbrace{\begin{pmatrix} \alpha_1 X_{2t-1} + \alpha_2 X_{2t-2} + W_t \\ \alpha_1 X_{2t} + \alpha_2 X_{2t-1} + W_t \end{pmatrix}}_{\text{dependencies to } X_{2t-1} \text{ and } X_{2t-2}} \Rightarrow U_{t-1} = \begin{pmatrix} X_{2t-2} \\ X_{2t-1} \end{pmatrix}$$

- An AR(2) process can be written as a vector AR(1) process:

$$\begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ x_{t-2} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} w_t \\ w_{t-1} \end{pmatrix}$$

(c) Second-order moving average (MA(2)) process



(c) A second-order moving average (MA(2)) process has the form [*7 marks*]

$$x_t = \beta_0 w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2},$$

where the βs are coefficients and the ws are independent Gaussian random variables $N(0, \sigma^2)$.

i. Draw a graphical model of the MA(2) process, showing both the w and x variables.

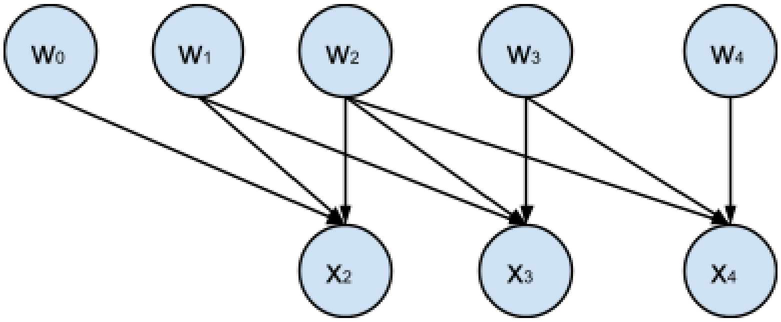
ii. For the MA(2) process, x_t and x_{t-d} are independent for sufficiently large d. Argue what the minimum value of d is to achieve independence, and justify your result.

iii. For the MA(2) model derive expressions for $\text{var}(x_t)$ and $\text{cov}(x_t, x_{t-1})$ in terms of the βs and σ^2.

Kshitij, Frederico, Tim:

(i)

$$x_t = \sum_{k=0}^{2} \beta_k w_{t-k}$$



(ii)

$p(x_t, x_{t-3}) = p(\beta_0 w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2},\ \beta_0 w_{t-3} + \beta_1 w_{t-4} + \beta_2 w_{t-5})$

… no "overlap" of the w parameters

$= p(\beta_0 w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2}) p(\beta_0 w_{t-3} + \beta_1 w_{t-4} + \beta_2 w_{t-5})$

$= p(x_t) p(x_{t-3})$

$I(x_{t,} x_{t-3:0} \mid \phi = \{\})$

ϕbeing the empty set

(iii)

given:

$w \sim N(0, \sigma^2)$ IID

$var(x_t) = var(\beta_0 w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2})$

$= var(\beta_0 w_t) + var(\beta_1 w_{t-1}) + var(\beta_2 w_{t-2})$

$= \beta_0^{\ 2} var(w_t) + \beta_1^{\ 2} var(\ w_{t-1}) + \beta_2^{\ 2} var(w_{t-2})$

remeber: $var(w) = \sigma^2$

$= \beta_0^{\ 2}\sigma^2 + \beta_1^{\ 2}\sigma^2 + \beta_2^{\ 2}\sigma^2$

$= \sigma^2(\beta_0^{\ 2} + \beta_1^{\ 2} + \beta_2^{\ 2})$

$cov(x_t, x_{t-1}) = E[x_t x_{t-1}] - E[x_t]E[x_{t-1}]$

split the problem up…

$E[x_t x_{t-1}]$

$= E\Big[(\beta_0 w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2})(\beta_0 w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2})\Big]$

*note for w with **different indexes**:*

$E[\beta_0^2 w_t w_{t-1}] = \beta_0^2 E[w_t w_{t-1}]$

(w_t and w_{t-1} are independent, b/c no connections within hidden layer)

(and expectation/mean of w is 0)

$= \beta_0^2 E[w_t]E[w_{t-1}] = \beta_0^2 \cdot 0 \cdot 0$

thus, only need to take into account multiplications with w of same index

$= E[\beta_1\beta_0 w_{t-1}^2 + \beta_2\beta_1 w_{t-2}^2]$

$= \beta_1\beta_0 E[w_{t-1}^2] + \beta_2\beta_1 E[w_{t-2}^2]$

note that $var(w) = E[w]^2 - E[w^2]$

thus $E[w^2] = E[w]^2 + var(w)$

where $E[w]^2 = 0$ *still*

$= \beta_1\beta_0 var(w_{t-1}) + \beta_2\beta_1 var(w_{t-2})$

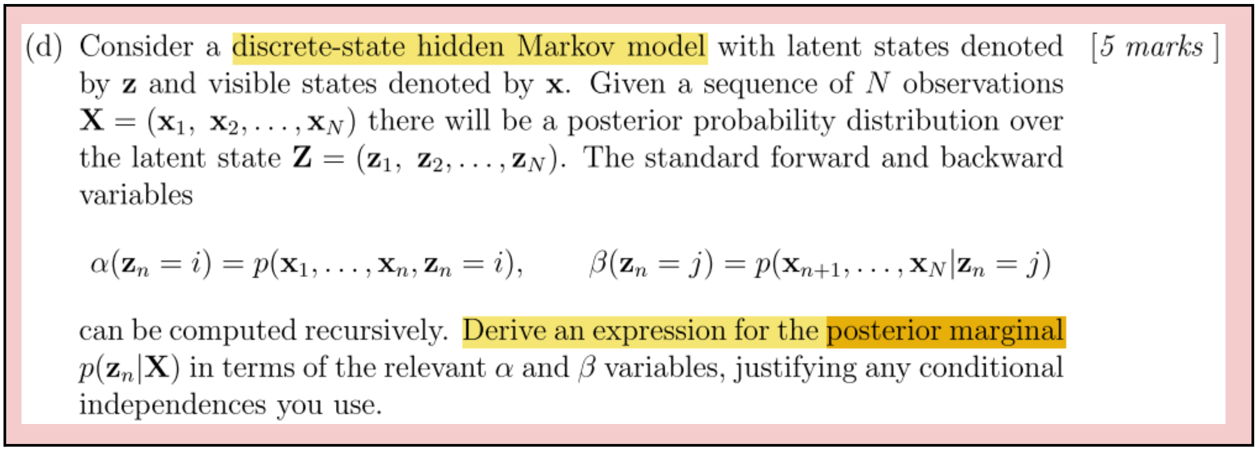
$= \beta_1\beta_0\sigma^2 + \beta_2\beta_1\sigma^2$

$E[x_t]E[x_{t-1}] = 0$

*because x is a **linear combination of w** which are **zero mean**, thus x is also zero mean*

$\Rightarrow cov(x_t, x_{t-1}) = E[x_t x_{t-1}] = \beta_1\beta_0\sigma^2 + \beta_2\beta_1\sigma^2$

(d) Discrete-state hidden Markov model (posterior marginal)



(d) Consider a discrete-state hidden Markov model with latent states denoted by $\mathbf{z}$ and visible states denoted by $\mathbf{x}$. Given a sequence of N observations $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ there will be a posterior probability distribution over the latent state $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N)$. The standard forward and backward variables [*5 marks*]

$$\alpha(\mathbf{z}_n = i) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n = i), \qquad \beta(\mathbf{z}_n = j) = p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n = j)$$

can be computed recursively. Derive an expression for the posterior marginal $p(\mathbf{z}_n|\mathbf{X})$ in terms of the relevant α and β variables, justifying any conditional independences you use.

$p(z_n | x_{1:N})$

$= p(z_n, x_{1:N}) / p(x_{1:N})$

$= p(z_n, x_{1:n}, x_{n+1:N}) / p(x_{1:N})$

$= p(x_{n+1:N} \mid z_n, x_{1:n})\ p(z_n, x_{1:n}) / \text{Sum\_}z_n\ p(x_{n+1:N} \mid z_n, x_{1:n})\ p(z_n, x_{1:n})$

$= p(x_{n+1:N} \mid z_n)\ p(z_n, x_{1:n}) / \text{Sum\_}z_n\ p(z_n, x_{1:N})$

= $a(z_n)$ $b(z_n)$ / Sum_z_n $a(z_n)$ $b(z_n)$

Define $a(z_n) = p(z_n, x_{1:n})$

= Sum_z_{n-1} $p(z_n, z_{n-1}, x_{1:n-1}, x_n)$

= Sum_z_{n-1} $p(x_n|z_n, z_{n-1}, x_{1:n-1})$ $p(z_n|z_{n-1}, x_{1:n-1})$ $p(z_{n-1}, x_{1:n-1})$

= Sum_z_{n-1} $p(x_n|z_n)$ $p(z_n|z_{n-1})$ $a(z_{n-1})$

= $p(x_n|z_n)$ S_z_{n-1} $p(z_n|z_{n-1})$ $a(z_{n-1})$

Define $b(z_n) = p(x_{n+1:N}|z_n)$

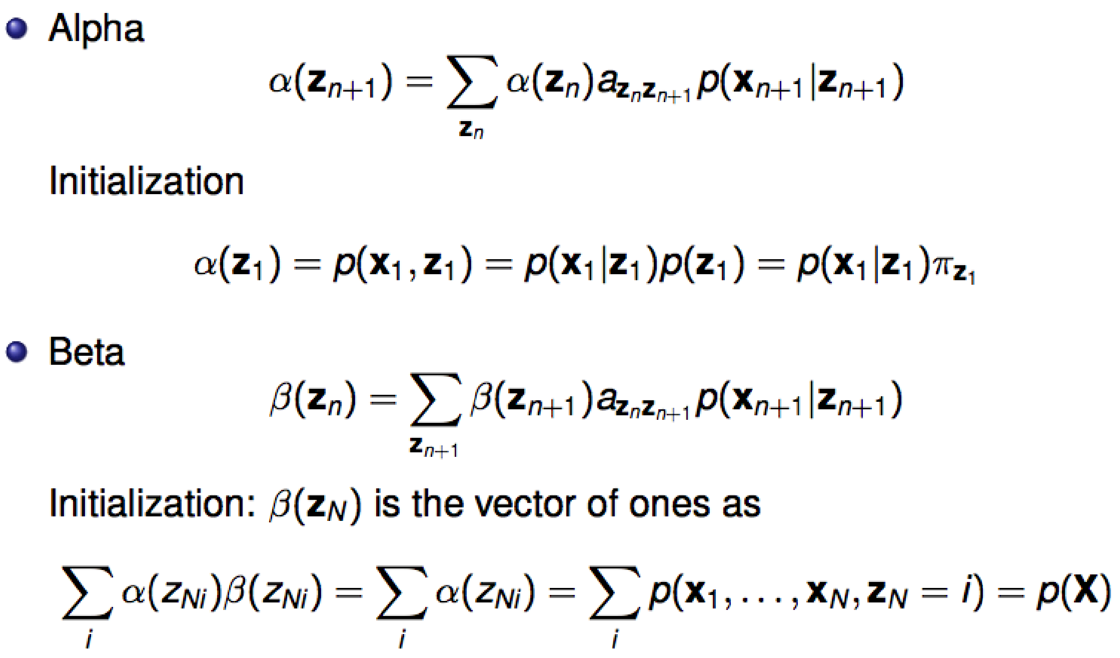
= Sum_z_{n+1} $p(z_{n+1}, x_{n+1}, x_{n+2:N}|z_n)$

= Sum_z_{n+1} $p(x_{n+1}|z_{n+1}, x_{n+2:N}, z_n)$ $p(x_{n+2:N}|z_{n+1}, z_n)$ $p(z_{n+1}|z_n)$

= Sum_ z_{n+1} $p(x_{n+1}|z_{n+1})$ $p(x_{n+2:N}|z_{n+1})$ $p(z_{n+1}|z_n)$

= Sum_z_{n+1} $p(x_{n+1}|z_{n+1})$ $b(z_{n+1})$ $p(z_{n+1}|z_n)$

Green: conditional independence from red



- Alpha

$$\alpha(\mathbf{z}_{n+1}) = \sum_{\mathbf{z}_n} \alpha(\mathbf{z}_n) a_{\mathbf{z}_n \mathbf{z}_{n+1}} p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})$$

Initialization

$$\alpha(\mathbf{z}_1) = p(\mathbf{x}_1, \mathbf{z}_1) = p(\mathbf{x}_1|\mathbf{z}_1)p(\mathbf{z}_1) = p(\mathbf{x}_1|\mathbf{z}_1)\pi_{\mathbf{z}_1}$$

- Beta

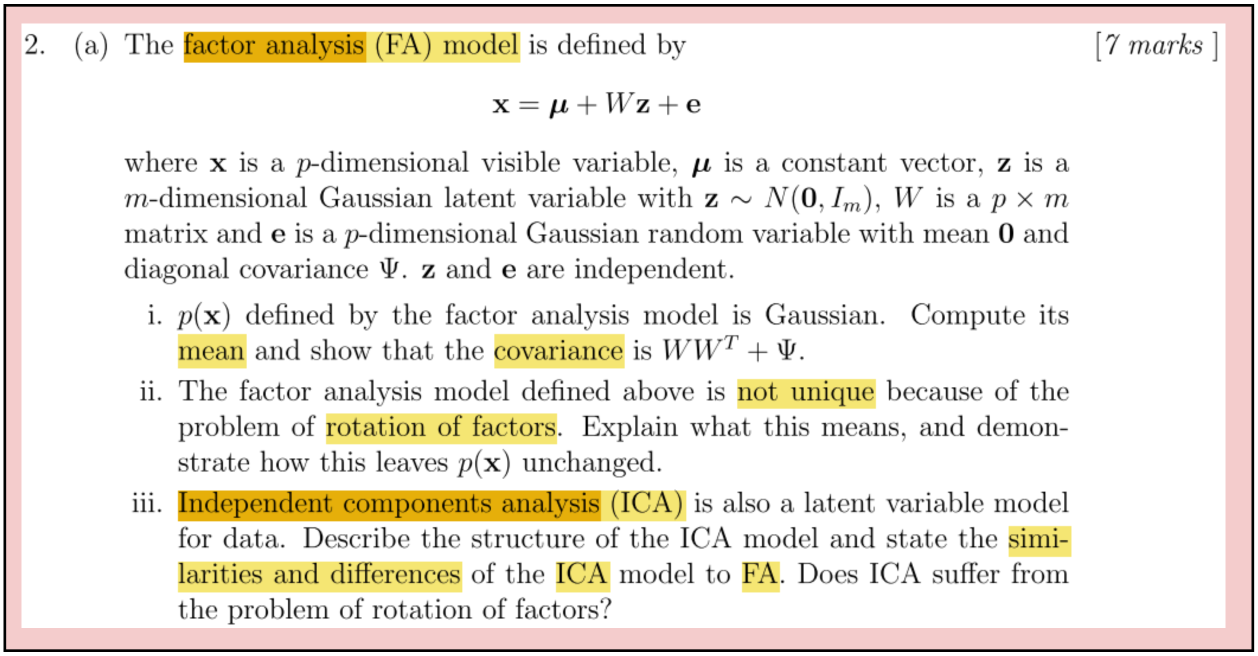
$$\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1}) a_{\mathbf{z}_n \mathbf{z}_{n+1}} p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})$$

Initialization: $\beta(\mathbf{z}_N)$ is the vector of ones as

$$\sum_i \alpha(z_{Ni})\beta(z_{Ni}) = \sum_i \alpha(z_{Ni}) = \sum_i p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_N = i) = p(\mathbf{X})$$

2 .

(a) Factor analysis (FA) model (mean, covariance, rotation of factors, ICA vs. FA)



2. (a) The factor analysis (FA) model is defined by [7 marks]

$$\mathbf{x} = \boldsymbol{\mu} + W\mathbf{z} + \mathbf{e}$$

where $\mathbf{x}$ is a p-dimensional visible variable, $\boldsymbol{\mu}$ is a constant vector, $\mathbf{z}$ is a m-dimensional Gaussian latent variable with $\mathbf{z} \sim N(\mathbf{0}, I_m)$, W is a $p \times m$ matrix and $\mathbf{e}$ is a p-dimensional Gaussian random variable with mean $\mathbf{0}$ and diagonal covariance Ψ. $\mathbf{z}$ and $\mathbf{e}$ are independent.

i. $p(\mathbf{x})$ defined by the factor analysis model is Gaussian. Compute its mean and show that the covariance is $WW^T + \Psi$.

ii. The factor analysis model defined above is not unique because of the problem of rotation of factors. Explain what this means, and demonstrate how this leaves $p(\mathbf{x})$ unchanged.

iii. Independent components analysis (ICA) is also a latent variable model for data. Describe the structure of the ICA model and state the similarities and differences of the ICA model to FA. Does ICA suffer from the problem of rotation of factors?

(i)

given:

$x = \mu + Wz + e$

with

μ constant vector

$z \sim N(0, I_m)$

$W \in R^{p\,x\,m}$ (constant)

$e \sim N(0, \psi)$

mean:

$E[x] = E[\mu] + E[Wz] + E[e]$

note: $E[Wz] = WE[z]$ where E[z] = 0

= ~~$\mu + 0 + \psi$~~ |← ψ should be 0 right? I think so as well. my bad!

$= \mu + 0 + 0$

variance

$cov(x) = cov(\mu) + cov(Wz) + cov(e)$

note $\mu$is constant, thus cov(μ)=0

$= 0 + cov(Wz) + \psi$

$$= 0 + WIW^T + \psi$$
$$= WW^T + \psi$$

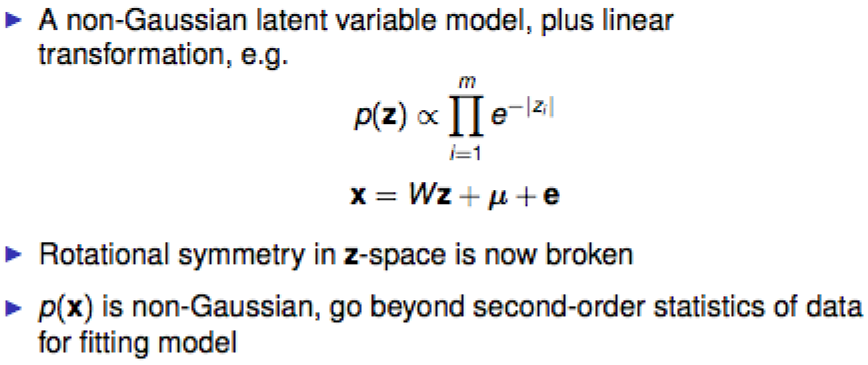
(ii)

Wibi: p(x) = N(μ, WW^T + Ψ) depends on W only in the form WW^T. So, given any rotation matrix R, we can rotate W with R yielding W' = WR. W'W'^{T} = WR(WR)^T = WRR^{T}W^T = WW^T.

iii. ICA does not suffer that problem, it find the actual latent variables, not a linear subspace of the data.

Yes it does, with a Gaussian prior on the hidden sources, which is why you have to use a non-Gaussian prior - see Barber p.475.

ICA structure:

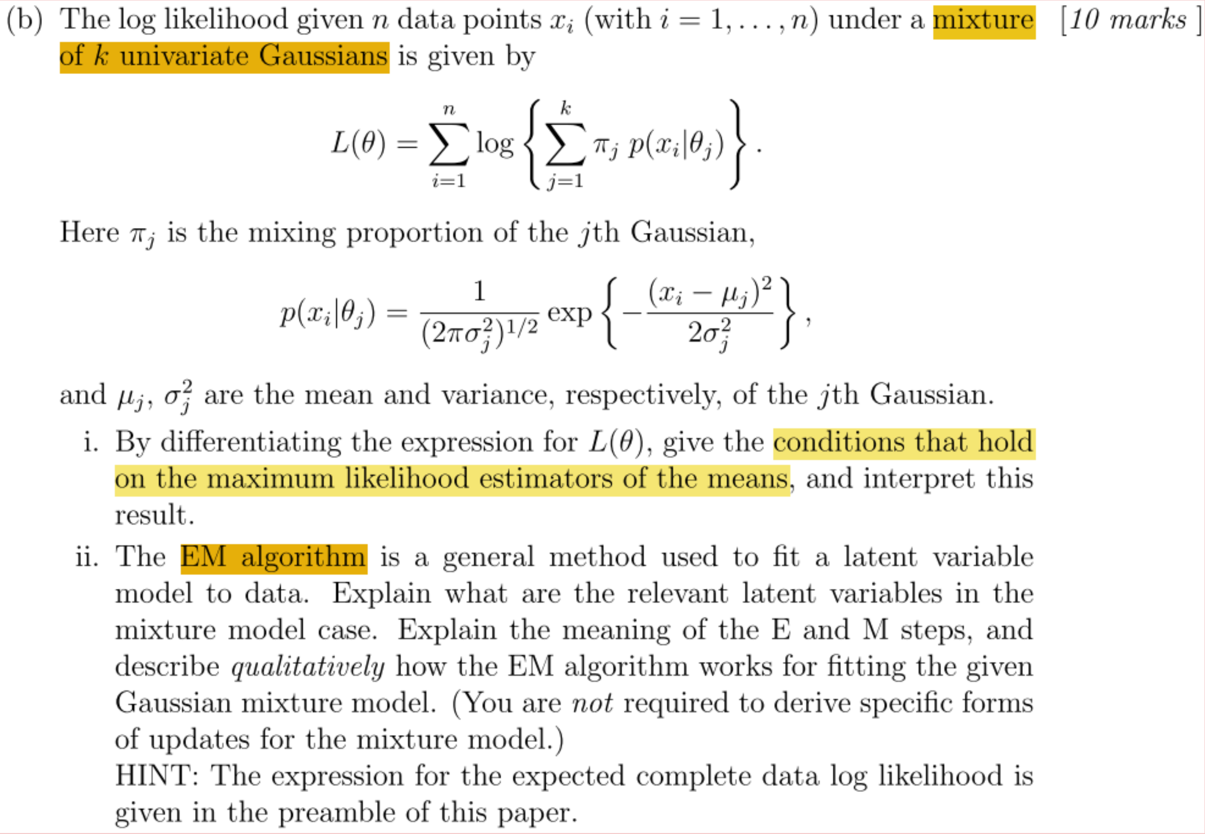


- A non-Gaussian latent variable model, plus linear transformation, e.g.

$$p(\mathbf{z}) \propto \prod_{i=1}^{m} e^{-|z_i|}$$
$$\mathbf{x} = W\mathbf{z} + \mu + \mathbf{e}$$

- Rotational symmetry in **z**-space is now broken
- $p(\mathbf{x})$ is non-Gaussian, go beyond second-order statistics of data for fitting model

(b) Mixture of Gaussians (EM algorithm)



(b) The log likelihood given n data points x_i (with $i = 1, \ldots, n$) under a mixture of k univariate Gaussians is given by [*10 marks*]

$$L(\theta) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \pi_j \, p(x_i|\theta_j) \right\}.$$

Here π_j is the mixing proportion of the jth Gaussian,

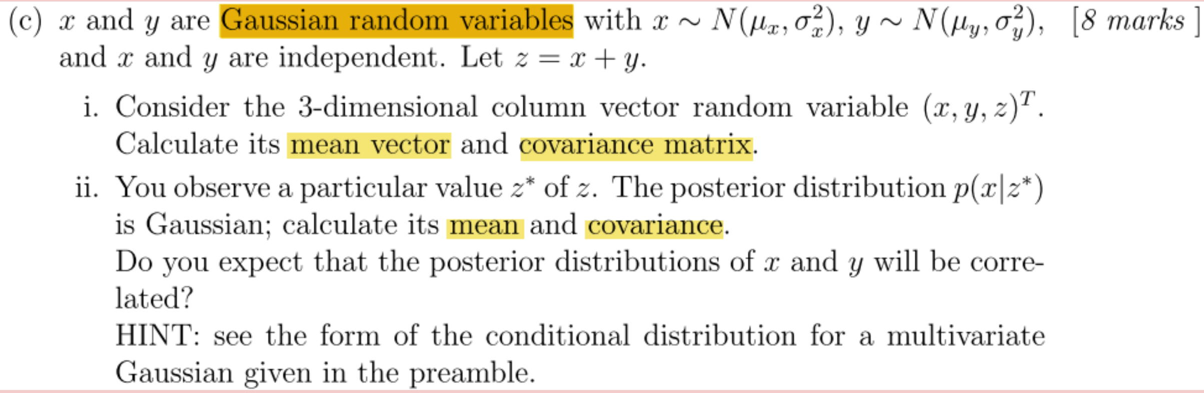
$$p(x_i|\theta_j) = \frac{1}{(2\pi\sigma_j^2)^{1/2}} \exp\left\{ -\frac{(x_i - \mu_j)^2}{2\sigma_j^2} \right\},$$

and μ_j, σ_j^2 are the mean and variance, respectively, of the jth Gaussian.

i. By differentiating the expression for $L(\theta)$, give the conditions that hold on the maximum likelihood estimators of the means, and interpret this result.

ii. The EM algorithm is a general method used to fit a latent variable model to data. Explain what are the relevant latent variables in the mixture model case. Explain the meaning of the E and M steps, and describe *qualitatively* how the EM algorithm works for fitting the given Gaussian mixture model. (You are *not* required to derive specific forms of updates for the mixture model.)
HINT: The expression for the expected complete data log likelihood is given in the preamble of this paper.

(c) Gaussian random variables (addition, mean, covariance)

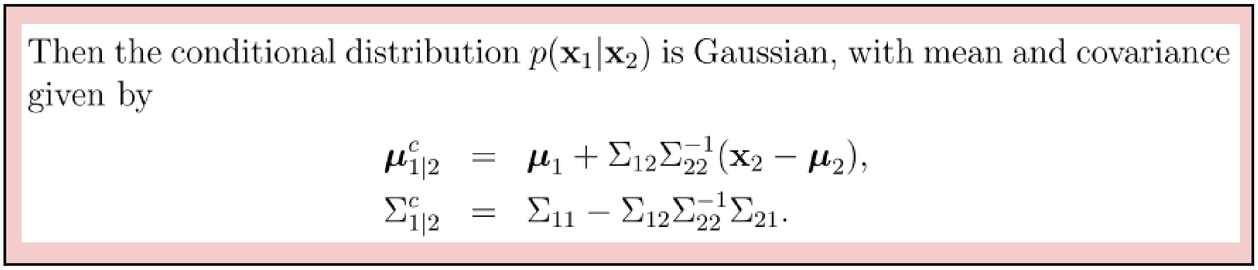


(c) x and y are Gaussian random variables with $x \sim N(\mu_x, \sigma_x^2)$, $y \sim N(\mu_y, \sigma_y^2)$, and x and y are independent. Let $z = x + y$. [*8 marks*]

i. Consider the 3-dimensional column vector random variable $(x, y, z)^T$. Calculate its mean vector and covariance matrix.

ii. You observe a particular value z^* of z. The posterior distribution $p(x|z^*)$ is Gaussian; calculate its mean and covariance.
Do you expect that the posterior distributions of x and y will be correlated?
HINT: see the form of the conditional distribution for a multivariate Gaussian given in the preamble.

HINT:

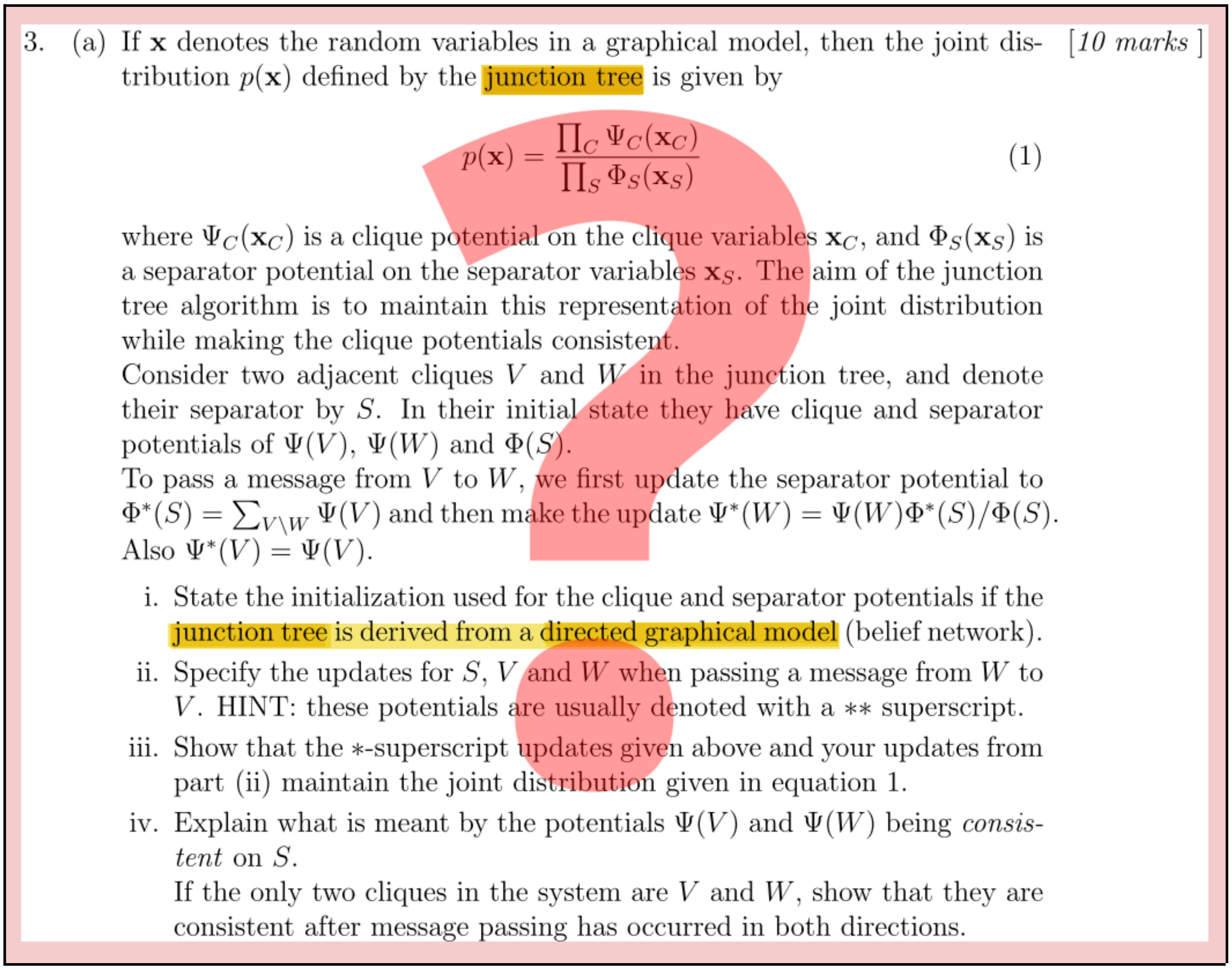


Then the conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ is Gaussian, with mean and covariance given by

$$\begin{aligned}\boldsymbol{\mu}^c_{1|2} &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2),\\ \Sigma^c_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.\end{aligned}$$

3 .

(a) (?) Junction Trees



3. (a) If $\mathbf{x}$ denotes the random variables in a graphical model, then the joint distribution $p(\mathbf{x})$ defined by the junction tree is given by [*10 marks*]

$$p(\mathbf{x}) = \frac{\prod_C \Psi_C(\mathbf{x}_C)}{\prod_S \Phi_S(\mathbf{x}_S)} \qquad (1)$$

where $\Psi_C(\mathbf{x}_C)$ is a clique potential on the clique variables $\mathbf{x}_C$, and $\Phi_S(\mathbf{x}_S)$ is a separator potential on the separator variables $\mathbf{x}_S$. The aim of the junction tree algorithm is to maintain this representation of the joint distribution while making the clique potentials consistent.
Consider two adjacent cliques V and W in the junction tree, and denote their separator by S. In their initial state they have clique and separator potentials of $\Psi(V)$, $\Psi(W)$ and $\Phi(S)$.
To pass a message from V to W, we first update the separator potential to $\Phi^*(S) = \sum_{V\backslash W}\Psi(V)$ and then make the update $\Psi^*(W) = \Psi(W)\Phi^*(S)/\Phi(S)$.
Also $\Psi^*(V) = \Psi(V)$.

i. State the initialization used for the clique and separator potentials if the junction tree is derived from a directed graphical model (belief network).

ii. Specify the updates for S, V and W when passing a message from W to V. HINT: these potentials are usually denoted with a $**$ superscript.

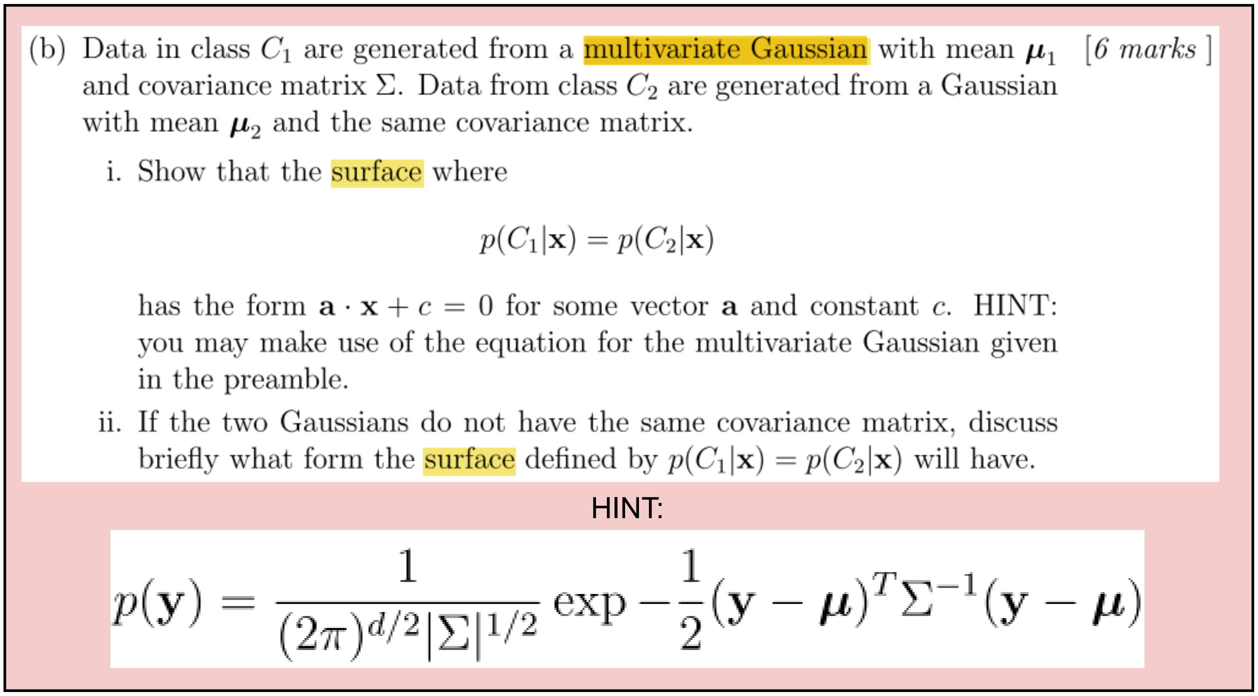
iii. Show that the $*$-superscript updates given above and your updates from part (ii) maintain the joint distribution given in equation 1.

iv. Explain what is meant by the potentials $\Psi(V)$ and $\Psi(W)$ being *consistent* on S.
If the only two cliques in the system are V and W, show that they are consistent after message passing has occurred in both directions.

Tim: not examinable to this extent this year?
Tadas: No. But we are expected to know what are junction trees in general and how they help us computing probabilities in trees (source Amos).
Tim: does anybody know know “what are junction trees in general and how they help us computing probabilities in trees”? :)

Tadas: They help in with complicated graphical networks where usual inference is hard (contains loops is the usual case). It replaces connected nodes in a specific way with cliques that can be represented later as a single node. In the end, you get a tree on which you can do (easier) inference.

(b) Multivariate Gaussians



(b) Data in class C_1 are generated from a multivariate Gaussian with mean $\boldsymbol{\mu}_1$ and covariance matrix Σ. Data from class C_2 are generated from a Gaussian with mean $\boldsymbol{\mu}_2$ and the same covariance matrix. [*6 marks*]

i. Show that the surface where

$$p(C_1|\mathbf{x}) = p(C_2|\mathbf{x})$$

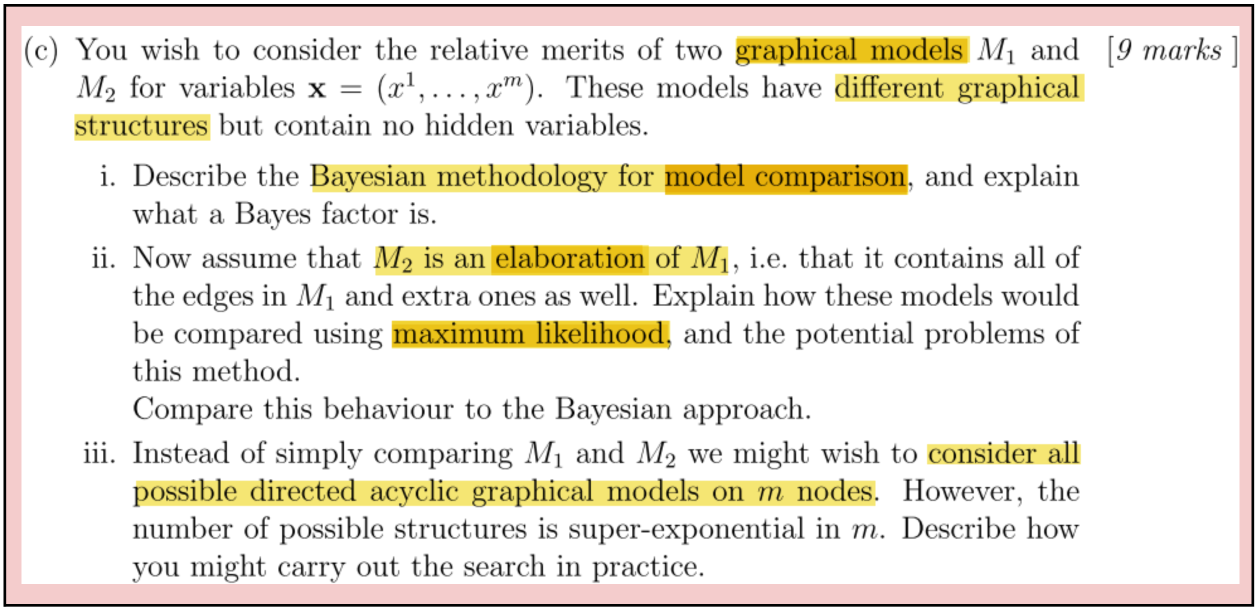
has the form $\mathbf{a} \cdot \mathbf{x} + c = 0$ for some vector $\mathbf{a}$ and constant c. HINT: you may make use of the equation for the multivariate Gaussian given in the preamble.

ii. If the two Gaussians do not have the same covariance matrix, discuss briefly what form the surface defined by $p(C_1|\mathbf{x}) = p(C_2|\mathbf{x})$ will have.

HINT:

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})$$

(c) Model comparison (Bayesian approach, MLE approach)



(c) You wish to consider the relative merits of two graphical models M_1 and M_2 for variables $\mathbf{x} = (x^1, \ldots, x^m)$. These models have different graphical structures but contain no hidden variables. [*9 marks*]

i. Describe the Bayesian methodology for model comparison, and explain what a Bayes factor is.

ii. Now assume that M_2 is an elaboration of M_1, i.e. that it contains all of the edges in M_1 and extra ones as well. Explain how these models would be compared using maximum likelihood, and the potential problems of this method.
Compare this behaviour to the Bayesian approach.

iii. Instead of simply comparing M_1 and M_2 we might wish to consider all possible directed acyclic graphical models on m nodes. However, the number of possible structures is super-exponential in m. Describe how you might carry out the search in practice.