COMP0085: Formative Assignments

Probabilistic and Unsupervised Learning / Approximate Inference

Maneesh Sahani

1. **Gaussian graphical models.** Consider a multivariate Gaussian variable $\mathbf{x} = (x_1, \ldots, x_n)$ with given mean vector μ and covariance matrix Σ.

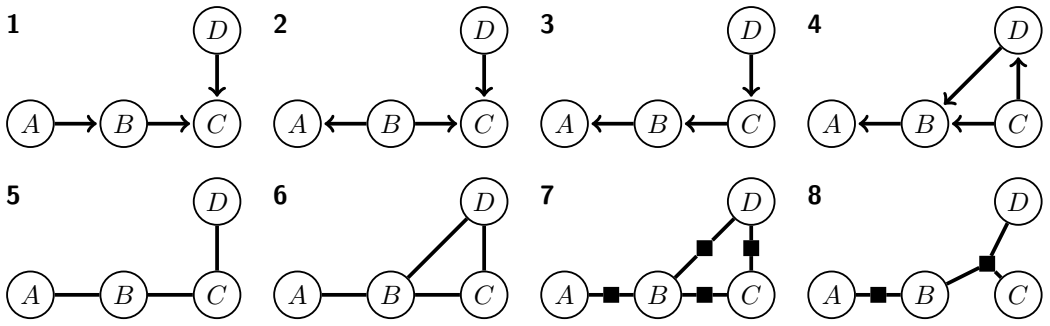
   (a) Write out the probability density function for $\mathbf{x}$.

   (b) Let $n = 4$, $\mu = (0, 1, 1, 0)$ and

   $$\Sigma = \frac{1}{6}\begin{pmatrix} 7 & -2 & -2 & 1 \\ -2 & 7 & 1 & -2 \\ -2 & 1 & 7 & -2 \\ 1 & -2 & -2 & 7 \end{pmatrix},$$

   draw the corresponding undirected graph and define clique potentials consistent with the above Gaussian. [Hint: multiply out the terms that appear in the exponent.]

2. **Conditional independencies and expressiveness of graphical models.** Consider the following graphical models:



   (a) For graphs 2, 4, 6 and 8, write down **all** the conditional independence relationships that hold for variable C of the form $C \perp\!\!\!\perp \mathcal{X} | \mathcal{V}$, where $\mathcal{X}$ and $\mathcal{V}$ are sets of other variables.

   (b) Two graphs are **equivalent** if they express *all* the same marginal and conditional independence relationships between their variables. A graph G is **subsumed** by graph H if all conditional independence relationships in H are exhibited in G. Divide the above 8 graphs into the smallest number of non-overlapping sets of equivalent graphs, and state which of these sets of equivalent graphs are subsumed by one of other sets.

3. **Noisy ICA** Consider a noisy independent factor model for data $\mathbf{x} \in \mathbb{R}^D$:

$$\begin{aligned} \mathbf{x} &\sim \mathcal{N}(A\mathbf{z}, \Psi) && A \in \mathbb{R}^{D\times K}; \quad \Psi \in \mathbb{R}^{D\times D}, \text{ diagonal.} \\ z_k &\sim \mathcal{N}\left(0, u_k^{-1}\right) && k = 1 \ldots K \\ u_k &\sim \mathsf{Gamma}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right) && k = 1 \ldots K \end{aligned}$$

(a) Derive the resulting marginal $P(z_k)$ (integrating out u_k). What distribution is this? Sketch the density function and explain the relationship between this model and ICA.

This two-stage representation of the prior on z_k is valuable because it allows us to design approximate methods for inference in noisy ICA. For example, consider variational Bayes (VB).

(b) Write down a *minimal* VB factorisation of the joint posterior on latents and parameters that would make approximate learning tractable. By minimal we mean that you should not assume any factorisations beyond those necessary to obtain a tractable algorithm.

(c) Derive the update for the term $q(\Psi, \dots)$ (where the dots stand for any other variables that may be included in the factor according to your factorisation). You may assume conjugate priors wherever needed.

(d) Derive hyperparameter optimisation rules to:

(i) learn the shape of the marginal distributions $P(z_k)$;

(ii) learn the number of factors.

You may need to alter the model parameterisation in the second case. Explain the reasoning behind your choice of hyperparametrisation, and derive the corresponding update rules.

4. **EP for sign constraints**

Consider a linear dynamical system:

$$y_1 \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

$$y_i | y_{i-1} \sim \mathcal{N}(y_{i-1}, \sigma^2) \qquad \text{for } i = 2, 3, \dots \tag{2}$$

$$x_i | y_i \sim \mathcal{N}(y_i, \tau^2) \qquad \text{for } i = 1, 2, \dots \tag{3}$$

with each random variable being scalar. Suppose we observe only the *signs* $s_i = \pm 1$ of the outputs x_i, rather than their magnitudes. Derive two different expectation propagation algorithms to approximate the resulting posterior over the y_is.

(a) To incorporate the sign observations, we could include additional factors of the form:

$$f_i(x_i) = \begin{cases} 1 & \text{if } s_i x_i > 0, \\ 0 & \text{otherwise} \end{cases}$$

Derive an expectation propagation algorithm to estimate the marginal distributions over all x_i and y_i in the joint distribution given by the (normalized) product of these factors with the distribution of equations (1-3). Approximate each factor with a Gaussian. You may assume access to a function which can compute the mean $E(m, v^2)$ and variance $V(m, v^2)$ of the truncated Gaussian:

$$P(z | m, v) \propto \begin{cases} e^{-\frac{(z-m)^2}{2v^2}} & \text{if } z > 0; \\ 0 & \text{otherwise} \end{cases}$$

(b) An alternative approach would be to first compute the probabilities:

$$g_i(y_i) = P(\text{sign}(x_i) = s_i | y_i),$$

and then use expectation propagation to estimate the marginals of y_i's in the joint distribution given by the product of the g_i factors with the prior $P(y_1, \ldots, y_t)$ given in equations (1-2). Show that both EP algorithms are equivalent in that they should have the same fixed points.

5. **Inconsistency of Local Marginals** Loopy belief propagation approximates the distribution over a pairwise MRF using a set of locally consistent beliefs $\{b_i(x_i), b_{ij}(x_i, x_j)\}$:

$$\sum_{x_i} b_i(x_i) = 1 \qquad \text{for all } i;$$

$$\sum_{x_i} b_{ij}(x_i, x_j) = b_j(x_j) \qquad \text{for all } i, j \text{ and } x_j.$$

(a) Give an example set of beliefs that are locally consistent but not globally consistent. That is, there is no distribution $p(\mathbf{X})$ over all variables such that

$$p(X_i = x_i) = b_i(x_i) \qquad \text{for all } i, x_i;$$

$$p(X_i = x_i, X_j = x_j) = b_{ij}(x_i, x_j) \qquad \text{for all } i, j, x_i, x_j.$$

Explain why this set of beliefs is not globally consistent.

(b) Construct a graphical model with specific parameter settings, such that the local marginals you came up with in the previous question form a fixed point of the loopy belief propagation algorithm run on this model.