These notes are intended to give a summary of relevant concepts from the lectures which are helpful to complete the exercises. It is not intended to cover the lectures thoroughly. Learning this content is not a replacement for working through the lecture material and the exercises.

KL divergence — The Kullback-Leibler divergence measures the "distance" between p and q:

$$\mathrm{KL}(p||q) = \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right] \tag{1}$$

It satisfies: $\mathrm{KL}(p||q) = 0 \Leftrightarrow p = q$, $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$, $\mathrm{KL}(p||q) \geq 0$. Optimising with respect to the first argument when the second is fixed leads to mode seeking. Optimising with respect to the second argument when the first is fixed produces global fits (moment-matching).

ELBO — For a joint model $p(\mathbf{x}, \mathbf{y})$, the evidence lower bound (ELBO) is

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right] \tag{2}$$

where $q(\mathbf{y}|\mathbf{x})$ is the variational distribution. It can be rewritten as

$$\log p(\mathbf{x}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x})) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{y}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y})) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log p(\mathbf{x}, \mathbf{y}) + \mathcal{H}(q)$$

where $\mathcal{H}(q) = -\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log q(\mathbf{y}|\mathbf{x})\right]$ is the entropy of q. The ELBO is a lower bound on $\log p(\mathbf{x})$. It is maximised when $q(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})$ which makes the bound tight.

EM algorithm — The expectation maximisation (EM) algorithm can be used to learn the parameters $\boldsymbol{\theta}$ of a statistical model $p(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})$ with latent (unobserved) variables $\mathbf{h}$ and visible (observed) variables $\mathbf{v}$ for which we have data $\mathcal{D}$. It updates the parameters $\boldsymbol{\theta}$ by iterating between the expectation (E) and the maximisation (M) step:

$$\text{E-step: compute } J(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_{\text{old}})}[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})] \qquad \text{M-step: } \boldsymbol{\theta}_{\text{new}} \leftarrow \operatorname*{argmax}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{3}$$

The update rule produces a sequence of parameters for which the log-likelihood is guaranteed to never decrease, i.e. $\ell(\boldsymbol{\theta}_{\text{new}}) \geq \ell(\boldsymbol{\theta}_{\text{old}})$.