

*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.*

Note the difference between the notations  $p(\mathbf{x}; \boldsymbol{\theta})$  and  $p(\mathbf{x} | \boldsymbol{\theta})$ . The former is a pdf/pmf of a random variable  $\mathbf{x}$  that is parametrised by a vector of numbers (parameters)  $\boldsymbol{\theta}$ . The latter is a *conditional* pdf/pmf of a random variable  $\mathbf{x}$  given information of another *random variable*  $\boldsymbol{\theta}$ .

**Likelihood**  $L(\boldsymbol{\theta})$  — The chance that the model generates data like the observed one when using parameter configuration  $\boldsymbol{\theta}$ . For *iid* data  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ , the likelihood of the parameters  $\boldsymbol{\theta}$  is

$$L(\boldsymbol{\theta}) = p(\mathcal{D}; \boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{x}_i; \boldsymbol{\theta}) \quad (1)$$

**Prior**  $p(\boldsymbol{\theta})$  — Beliefs about the plausibility of parameter values before we see any data.

**Posterior**  $p(\boldsymbol{\theta} | \mathcal{D})$  — Beliefs about the parameters after having seen the data. This is proportional to the likelihood function  $L(\boldsymbol{\theta})$  weighted by our prior beliefs about the parameters  $p(\boldsymbol{\theta})$

$$p(\boldsymbol{\theta} | \mathcal{D}) \propto L(\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (2)$$

**Parametric statistical model** — A set of pdfs/pmfs indexed by parameters  $\boldsymbol{\theta}$ ,

$$\{p(\mathbf{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta}} \quad (3)$$

- **Parameter estimation** Using  $\mathcal{D}$  to pick the “best” parameter value  $\hat{\boldsymbol{\theta}}$  among the possible  $\boldsymbol{\theta}$  – i.e. pick the “best” pdf/pmf  $p(\mathbf{x}; \hat{\boldsymbol{\theta}})$  from the set of pdfs/pmfs  $\{p(\mathbf{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ ,

**Bayesian model** — Considers  $p(\mathbf{x}; \boldsymbol{\theta})$  to be conditional  $p(\mathbf{x} | \boldsymbol{\theta})$ . Models the distribution of the parameters  $\boldsymbol{\theta}$ , as well as the random variable  $\mathbf{x}$

$$p(\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (4)$$

- **Bayesian inference** Determine the plausibility of all possible  $\boldsymbol{\theta}$  in light of the observed data – i.e. compute the posterior  $p(\boldsymbol{\theta} | \mathcal{D})$ .

**Maximum likelihood** — The parameters  $\hat{\boldsymbol{\theta}}$  that give the largest likelihood (or log-likelihood)

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ell(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L(\boldsymbol{\theta}) \quad (5)$$

Sometimes this can be computed directly (as in the tutorials). However, numerical methods are often needed for this optimisation problem, which leads to local optima.

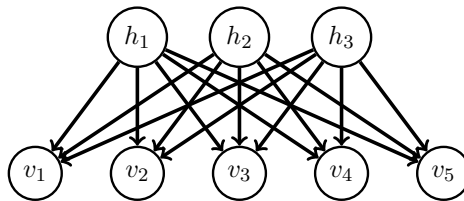
**Factor analysis** — A graphical model where statistical dependencies between the observed variables (visibles  $\mathbf{v}$ ) is modelled through unobserved variables (latents  $\mathbf{h}$ ). In factor analysis, the latents  $\mathbf{h}$  are assumed to be independent Gaussians with zero mean and unit variance.

$$p(\mathbf{h}) = \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{I})$$

$$p(\mathbf{v} | \mathbf{h}; \boldsymbol{\theta}) = \mathcal{N}(\mathbf{v}; \mathbf{F}\mathbf{h} + \mathbf{c}, \boldsymbol{\Psi})$$

$$\mathbf{v} = \mathbf{F}\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; 0, \boldsymbol{\Psi})$$



The covariance matrix  $\boldsymbol{\Psi}$  is a diagonal matrix. Probabilistic PCA is a special case of factor analysis, where  $\boldsymbol{\Psi} = \sigma^2 \mathbf{I}$ .

**Independent component analysis** — The DAG is the same as in factor analysis, but with non-Gaussian latents (one latent may be Gaussian)

$$p(\mathbf{h}) = \prod_i p(h_i)$$

$$p(\mathbf{v} | \mathbf{h}; \boldsymbol{\theta}) = \mathcal{N}(\mathbf{v}; \mathbf{A}\mathbf{h} + \mathbf{c}, \boldsymbol{\Psi})$$