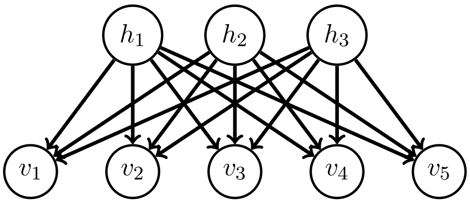
These notes are intended to give a summary of relevant concepts from the lectures which are helpful to complete the tutorial sheet. It is not intended to cover the lectures thoroughly. Learning this content is not a replacement for working through the lecture material and the tutorial sheet.

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.

$$
\begin{aligned}
p(\mathbf{h}) &= \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{I}) \\
p(\mathbf{v} \mid \mathbf{h}; \boldsymbol{\theta}) &= \mathcal{N}(\mathbf{v}; \mathbf{F}\mathbf{h} + \mathbf{c}, \boldsymbol{\Psi})
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{v} = \mathbf{F}\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon}& \\
\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; 0, \boldsymbol{\Psi})&
\end{aligned}
$$



Where the covariance $\boldsymbol{\Psi}$ 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

$$
\begin{aligned}
p(\mathbf{h}) &= \prod_i p(h_i) \\
p(\mathbf{v} \mid \mathbf{h}; \boldsymbol{\theta}) &= \mathcal{N}(\mathbf{v}; \mathbf{A}\mathbf{h} + \mathbf{c}, \boldsymbol{\Psi})
\end{aligned}
$$

Score matching — A parameter estimation method for models over continuous random variables when the partition function is intractable. The score matching cost function $J_{\mathrm{sm}}(\boldsymbol{\theta})$ is the expectation under the data distribution $p_*(\mathbf{x})$ of the squared difference between the model score function $\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta})$ and the data score function $\boldsymbol{\psi}_*(\mathbf{x})$

$$
\begin{aligned}
\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}) &= \nabla_{\mathbf{x}} \log p(\mathbf{x}; \boldsymbol{\theta}) = \nabla_{\mathbf{x}} \log \tilde{p}(\mathbf{x}; \boldsymbol{\theta}) \\
\boldsymbol{\psi}_*(\mathbf{x}) &= \nabla_{\mathbf{x}} \log p_*(\mathbf{x}) \\
J_{\mathrm{sm}}(\boldsymbol{\theta}) &= \frac{1}{2}\mathbb{E}_{p_*(\mathbf{x})} \|\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}) - \boldsymbol{\psi}_*(\mathbf{x})\|^2 \qquad (1)
\end{aligned}
$$

Working with gradients removes the intractable partition function. We cannot compute the data score function $\boldsymbol{\psi}_*(\mathbf{x})$ directly. However, we do not need to, under mild conditions, the optimisation problem can be written as:

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, J(\boldsymbol{\theta}) \\
J(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\left[\partial_j \psi_j(\mathbf{x}_i; \boldsymbol{\theta}) + \frac{1}{2}\psi_j(\mathbf{x}_i; \boldsymbol{\theta})^2\right] \qquad (2)
\end{aligned}
$$