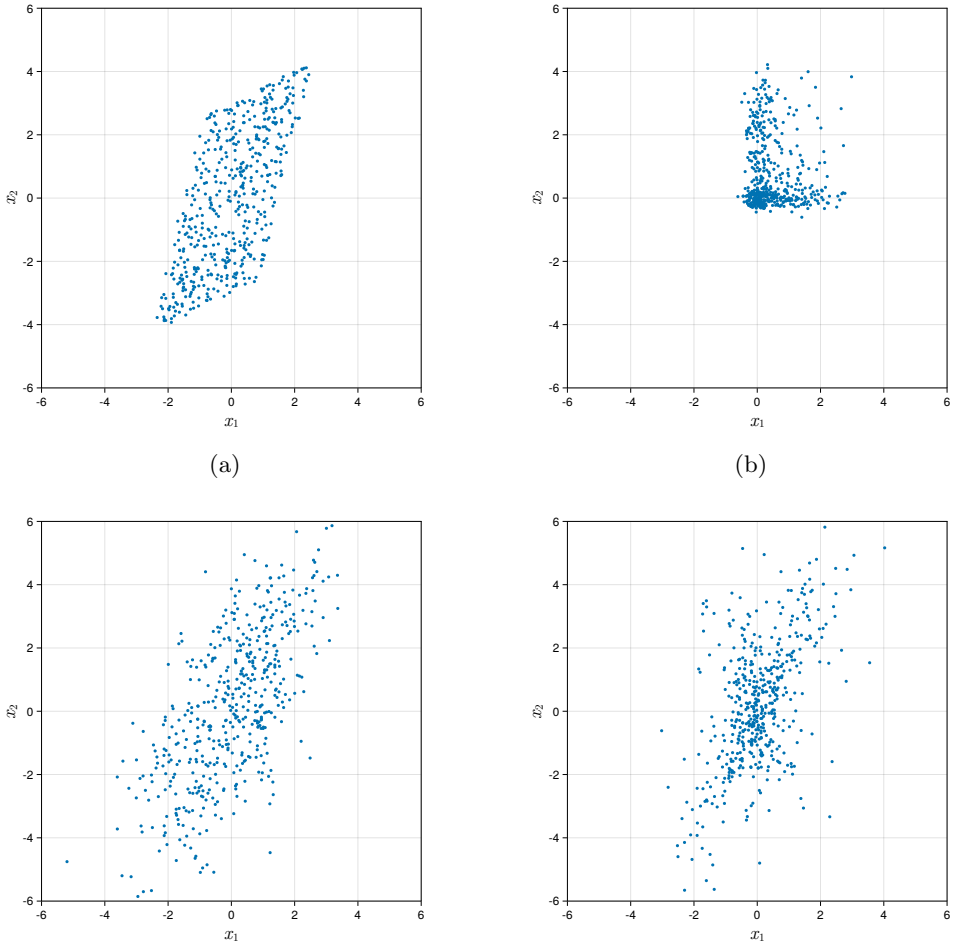
These are exercises for self-study and exam preparation. All material is examinable unless otherwise mentioned.

Exercise 1. *Generative model of variational autoencoders*

Factor analysis, (noisy) independent component analysis, and variational auto-encoders have the same directed graphical model. However, the three models make different distributional assumptions which leads to models with markedly different properties which we here investigate.

(a) Consider the four scatter plots below that show observed two-dimensional data (x_1, x_2). For each plot indicate the simplest model, from the four listed below, that may have generated the data. Each of the four models should be used once:

- *a factor analysis model*
- *a Gaussian VAE*
- *a noise-free square ICA model with super-Gaussian sources*
- *a noise-free square ICA model with sub-Gaussian sources*



(a) (b) (c) (d)

Solution. The mapping from graphs to models is as follows:

- Graph (i): ICA model with sub-Gaussian sources
- Graph (ii): VAE model
- Graph (iii): FA model
- Graph (iv): ICA model with super-Gaussian sources

(b) Briefly explain under which assumptions a Gaussian VAE model as discussed in the course becomes a factor analysis model.

Solution. In a Gaussian VAE, the generative model $p(\mathbf{v}|\mathbf{h})$ is Gaussian with a nonlinear dependency of the mean and variance on $\mathbf{h}$. We obtain the FA model when the mean function is linear in $\mathbf{h}$, the covariance matrix is diagonal, and the variances do not depend on $\mathbf{h}$.

Exercise 2. *ELBO of variational autoencoders*

The general expression for the ELBO for parameter estimation in presence of unobserved variables is

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \mathbb{E}_{q(\mathbf{h}|\mathcal{D})}\left[\log \frac{p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})}{q(\mathbf{h}|\mathcal{D})}\right] \tag{1}$$

where $\mathcal{D}$ are the observed data, $\mathbf{h}$ the unobserved variables, $\boldsymbol{\theta}$ the parameters of the model for the observed and unobserved variables, and $q(\mathbf{h}|\mathcal{D})$ is the variational distribution.

In the case of some VAEs, the above ELBO becomes

$$\begin{aligned} J(\boldsymbol{\theta}, \boldsymbol{\phi}) = &\sum_{i=1}^{n}\sum_{k=1}^{d} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{v}_i)}\left[v_{ik}\log p_k(\mathbf{h}) + (1 - v_{ik})\log(1 - p_k(\mathbf{h}))\right] + \\ &+ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{H}\left(1 + \log(\sigma_j^2(\mathbf{v}_i)) - \sigma_j^2(\mathbf{v}_i) - \mu_j^2(\mathbf{v}_i)\right) \end{aligned} \tag{2}$$

where the $p_k(\mathbf{h})$ are some functions parameterised by $\boldsymbol{\theta}$ that map $\mathbf{h}$ to $[0, 1]$, $\sigma_j(\mathbf{v})$ and $\mu_j(\mathbf{v})$ are some functions parameterised by $\boldsymbol{\phi}$, and the $\mathbf{v}_i = (v_{i1}, \ldots, v_{id})$ are the observed data points. Variable $\mathbf{h}$ is H dimensional.

(a) State two independence assumptions that we made for $p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})$ when deriving Equation (2) from Equation (1).

Solution. The two independence assumptions are:

- The data points $\mathbf{v}_i$ are independent. (In more detail, they are iid, and we have a latent variable $\mathbf{h}_i$ per observed data point $\mathbf{v}_i$, and the $\mathbf{h}_i$ are independent too.)
- With $\mathbf{v} = (v_1, \ldots, v_d)$, the v_k are conditionally independent given the latents.

(b) Would you use a VAE trained with the ELBO in Equation (2) for real-valued or binary data? Justify your answer.

Solution. For binary data. This is because the term $v_{ik} \log p_k(\mathbf{h}) + (1 - v_{ik}) \log(1 - p_k(\mathbf{h}))$ is the log pmf of a Bernoulli distribution with success probability $p_k(\mathbf{h})$.

(c) Equation (2) uses amortised variational inference. Briefly explain what amortisation refers to in the context of VAEs and state a reason why we may use it.

Solution. Amortisation in the context of VAEs means that we do not aim to estimate $q(\mathbf{h}|\mathbf{v}_i)$ for each $\mathbf{v}_i$ separately but that we consider parametric distributions $q(\mathbf{h}; \boldsymbol{\eta}_{\boldsymbol{\phi}}(\mathbf{v}_i))$ where the values of the parameters $\boldsymbol{\eta}$ are the output of a nonlinear function (neural network) that takes $\mathbf{v}_i$ as input. The nonlinear function is shared among all $\mathbf{v}_i$ and has tunable parameters $\boldsymbol{\phi}$ that are learned by maximising the ELBO.

The main reason for using amortisation is computational: optimising each $q(\mathbf{h}|\mathbf{v}_i)$ separately is often too costly. Another reason is that we can use the learned $\boldsymbol{\eta}_{\boldsymbol{\phi}}$ to cheaply approximate the variational distribution of a test point. The caveat here is that the approximation may not be very accurate due to the amortisation gap.