Learning, Probabilities and Causality - Part 1

Xavier Alameda-Pineda and Thomas Hueber

MSIAM 2022-2023

Contents

Chapter 1

Introduction to Probabilistic Learning

1.1 Introduction

Probabilistic learning means we **model** our data using probabilities. For example in classification, we aim to estimate the posterior probability of an item x: $p(c|x)$ for every possible class c. To develop complex probabilistic models, we will use basic concepts in probabilities such as the Bayes rule:

$$p(x, c) = p(x|c)p(c) \quad \Rightarrow \quad p(c|x) = \frac{p(x|c)p(c)}{\sum\limits_{k} p(x|k)p(k)}. \tag{1.1}$$

But before that, we should quickly recall what do these "p" mean.

- For **discrete variables** (i.e. measurable events are discrete), the $p(c)$ is the probability that the random variable takes the value c, and we write:
$$p(c) = P(C = c). \tag{1.2}$$
- For **continuous variables** (i.e. measurable events are continuous), $p(x)$ denotes the probability density function f_X at x:
$$p(x) = f_X(x) \tag{1.3}$$
and that probabilities can be computed over measurable sets, as the integral over the set:
$$p(\mathcal{X}) = P(x \in \mathcal{X}) = \int_{\mathcal{X}} f_X(x)\mathrm{d}x. \tag{1.4}$$
As a consequence, the probability of a single value is always zero $P(\{x\}) = 0$.

We will be discussing probabilistic models with **hidden** or **latent** random variables. This means that some variables of the model will be observed, while other will not be observed (thus latent/hidden). This is useful for a wide variety of applications.

Example 1: Clustering Given a set of data points $(x_j)_{j=1,\dots,n}$ in $\mathbb{R}^d$, we aim to find (& predict) clusters (groups of points). In a probabilistic model-based approach, we will have one latent variables assigned to each point, z_j. The realisation of the latent variable will index the cluster the associated observation belongs to. The model states that two observations belonging to the same cluster should have the same conditional distribution. We will discuss this model in detail in the Gaussian mixture model (GMM) chapter.

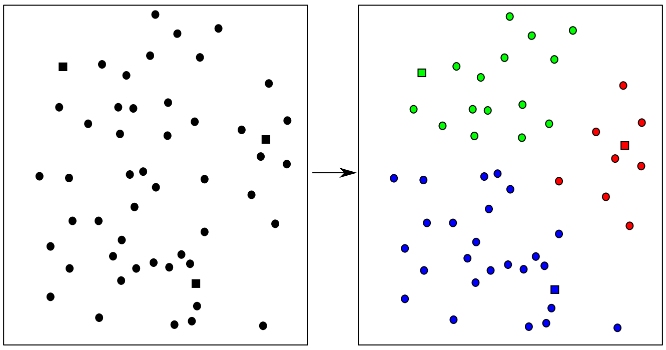


Figure 1.1: Example of a clustering result in 2D with 3 clusters.

Example 2: Dimensionality Reduction It is difficult to work with high-dimensional data (visualisation, pattern identification, etc). Ideally we would like to use a lower-dimensional representation. One possibility is to project in the subspace spanning the largest variance/energy. We will discuss more about this in the probabilistic principal component analysis (PPCA) chapter.

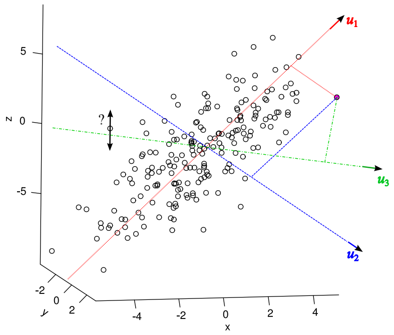


Figure 1.2: Example of dimensionality reduction result from 3 to 2 dimensions (spanned by the blue and red vectors. The original points are projected in the sub-space with lower dimension (e.g. green arrow).

Example 3: Analysis of Sequential Data Temporal data segmentation can be seen as a special case of clustering, where temporal dependencies are important. One can hypothesize that the observation/latent variables at time t depend only on a few of the past variables. In that case we say that the model is Markovian. A very important example are hidden Markov models (HMM), that will we widely discussed.

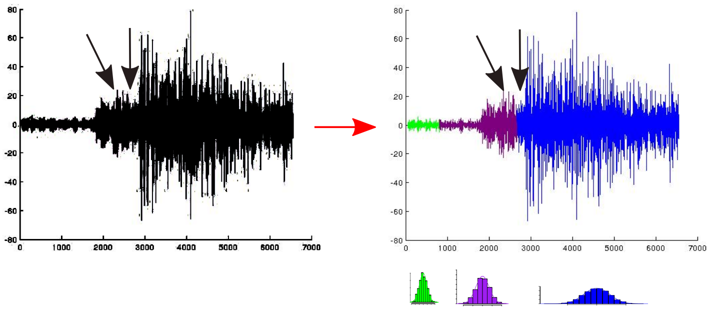


Figure 1.3: Example of temporal segmentation of a time-series into three color-coded different segments.

In addition, we will also discuss approximate inference tools, mainly variational inference. In particular the family of variational autoencoders. We will also discuss other non-linear models such as normalising flows and diffusion models.

1.2 Multivariate Gaussians

1.2.1 Recalling the 1D Gaussian Distribution

Let's recall the definition of the probability density function of **univariate Gaussian distribution**:

$$p(x) = \mathcal{N}(x; \mu, \nu) = \frac{1}{\sqrt{2\pi\nu}} \exp\Big(-\frac{(x-\mu)^2}{2\nu}\Big), \tag{1.5}$$

with $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ and $\nu \in \mathbb{R}^+$. The parameters μ and ν are called respectively the mean and the variance, and satisfy:

$$\mu = \mathbb{E}_{p(x)}\{x\} \qquad \nu = \mathbb{E}_{p(x)}\{(x-\mu)^2\}. \tag{1.6}$$

Remark 1.1: We will often use the **expectation** of a function f of a random variable x w.r.t. the probability density function $p(x)$, and denote it by:

$$\mathbb{E}_{p(x)}\{f(x)\} = \int_{\mathcal{X}} f(x)p(x)\mathrm{d}x, \tag{1.7}$$

where $\mathcal{X}$ is the domain of the random variable x.

One very common and useful concept is the maximum-likelihood estimators. To derive them, we assume the existence of N observations $X = \{x_1, \dots, x_N\}$ following a univariate Gaussian distribution, and the aim is to find the values of the parameters maximising the log-likelihood:

$$\mathcal{L}(\mu, \nu | X) = \sum_{n=1}^{N} \log \mathcal{N}(x_n; \mu, \nu) \tag{1.8}$$

Exercise 1.1: Prove that the expressions for the maximum likelihood estimators μ^* and ν^* are:

$$\mu^* = \frac{1}{N} \sum_{n=1}^{N} x_n \qquad \nu^* = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu^*)^2. \tag{1.9}$$

1.2.2 Constructing Multivariate Gaussian Distributions

As Product of Independent Gaussians But what happens now if we want to define a Gaussian on $\mathbb{R}^D$? One trivial extension of the 1D Gaussian is to consider each dimension as an independent Gaussian, with a different mean and variance: $\mu_1, \dots, \mu_D$ and $\nu_1, \dots, \nu_D$.
In that case the density writes:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\nu}) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\nu_d}} \exp\Big(-\frac{(x_d - \mu_d)^2}{2\nu_d}\Big), \tag{1.10}$$

where we have defined $\mathbf{x} = (x_1, \dots, x_D) \in \mathbb{R}^D$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_D) \in \mathbb{R}^D$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_D), \nu_d \in \mathbb{R}^+$.

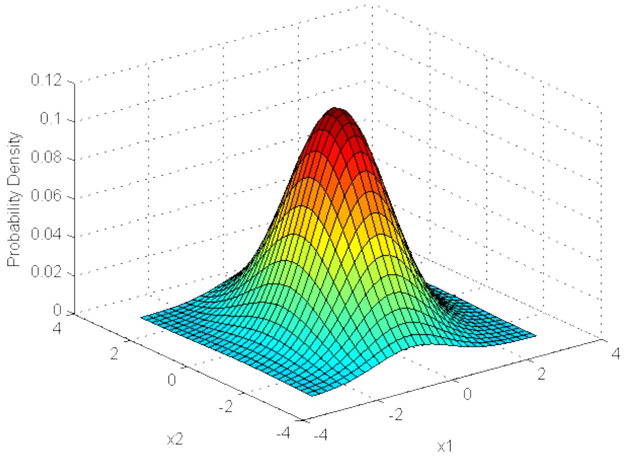


Figure 1.4: Probability density function of a 2D Gaussian distribution.

Exercise 1.2: Derive the maximum likelihood estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.

This is a special case of a multivariate Gaussian distribution. To present the more general case, we first remark that the above expression can be rewritten using matrix notation. Indeed, by defining:

$$\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\nu}) = \begin{pmatrix} \nu_1 & 0 & \dots & 0 \\ 0 & \nu_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \nu_D \end{pmatrix} \tag{1.11}$$

the density rewrites as:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\Big). \tag{1.12}$$

Exercise 1.3: Prove it!

Generalisation to "any" Covariance Matrix Naturally, the question of the constraints/conditions on the matrix $\mathbf{\Sigma}$ arises. In other words, one wonders whether or not a multivariate Gaussian distribution can be defined for any matrix $\mathbf{\Sigma}$. The answer is simply no. The so-called **covariance matrix** must be *symmetric* and *positive definite* matrices.

Remark 1.2: A $D \times D$ symmetric matrix $\mathbf{\Sigma}$ is **positive definite** if and only if $\mathbf{v}^\top \mathbf{\Sigma} \mathbf{v} > 0, \forall \mathbf{v} \neq \mathbf{0}$.

For the Gaussian distribution, this is intuitive, since the variance should be strictly positive in any direction, see Figure 1.4. Let $\mathbf{\Sigma}$ be a symmetric and positive definite matrix, the $\mathbf{\Sigma}$ has the following properties:

- All eigenvalues of $\mathbf{\Sigma}$ are real and strictly positive.
- The inverse of $\mathbf{\Sigma}$ is symmetric and positive definite.
- We can write $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with $\mathbf{\Lambda}$ diagonal and $\mathbf{U}$ orthogonal, with $\mathbf{\Lambda}$ containing the eigenvalues and $\mathbf{U}$ the eigenvectors (as columns).
- With this notation the inverse writes: $\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top$.

Exercise 1.4: Prove that the inverse of a (symmetric) positive definite matrix always exists.

If $\mathbf{\Sigma}$ is a covariance matrix we can define the *Mahalanobis* distance:

$$\mathcal{M}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{1.13}$$

(which is the Euclidean distance for $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{\Sigma} = \mathbf{I}$).

Remark 1.3: Given a vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and symmetric and positive definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{D\times D}$, we can define the probability density function of a **multivariate Gaussian** distribution as:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}|}} \exp\Big(-\frac{1}{2}\mathcal{M}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})\Big) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}|}} \exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\Big) \tag{1.14}$$

The vector $\boldsymbol{\mu}$ and the matrix $\mathbf{\Sigma}$ are usually referred to as the *mean vector* and the *covariance matrix*. The covariance matrix corresponds to the variance (and not to the standard deviation) of a univariate Gaussian. Formally, the mean vector and covariance matrix are defined as:

$$\boldsymbol{\mu} = \mathbb{E}_{\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\mathbf{\Sigma})}\{\mathbf{x}\} \qquad \mathbf{\Sigma} = \mathbb{E}_{\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\mathbf{\Sigma})}\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\}. \tag{1.15}$$

Exercise 1.5: Prove that the normalisation constant of a multivariate Gaussian distribution with covariance matrix $\mathbf{\Sigma}$ is $\sqrt{|2\pi\mathbf{\Sigma}|}$.

As Linear Transformations of the Standard Gaussian Multivariate Gaussian distributions can be constructed from the so-called standard Gaussian by using linear transformations.

Remark 1.4: The standard multivariate Gaussian is defined as the zero-mean and unit-variance Gaussian distribution, or equivalently, as the cartesian product of D univariate standard Gaussian distributions. The associated probability density function writes:

$$\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}) = \frac{1}{(2\pi)^{D/2}} \exp\Big(-\frac{1}{2}\|\mathbf{z}\|^2\Big). \tag{1.16}$$

Exercise 1.6: Let us consider the case where $\mathbf{z}$ follows a standard multivariate Gaussian distribution, and we define $\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$ with $\mathbf{A} \in \mathbb{R}^{D\times D}$ being an invertible matrix ($|\mathbf{A}| \neq 0$). Prove that:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}), \quad \text{with} \quad \mathbf{\Sigma} = \mathbf{A}\mathbf{A}^\top. \tag{1.17}$$

Implicitly, we are saying that for any invertible matrix $\mathbf{A}$, $\mathbf{A}\mathbf{A}^\top$ is symmetric (obvious) and positive definite. We will use this construction using linear transformations to better understand multivariate Gaussians. We will take several particular cases, and plot the level curves and associated probability density functions. For plotting purposes we will restrict to $D = 2$. For simplicity and without loss of generality, we will ignore the mean vector. We will consider four cases:

$$\mathbf{x} = \mathbf{z} \qquad \mathbf{x} = \underbrace{\begin{pmatrix} \sqrt{2} & 0 \\ 0 & 1/\sqrt{3} \end{pmatrix}}_{\mathbf{A}_1} \mathbf{z} \qquad \mathbf{x} = \underbrace{\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}}_{\mathbf{A}_2(\theta),\ \theta=30^\circ} \mathbf{z} \qquad \mathbf{x} = \mathbf{A}_2\mathbf{A}_1\mathbf{z}. \tag{1.18}$$

The first case corresponds to plot the standard Gaussian:

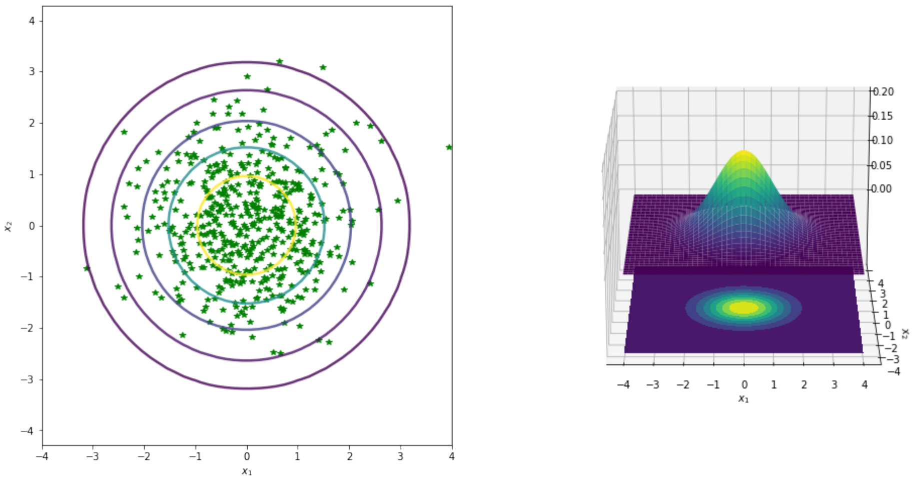


Figure 1.5: Plots corresponding to the standard multivariate Gaussian. (left) Level curves and sample data points in green. (right) 3D plot of the probability density function, and associated level curves.

The second case corresponds to scaling the points with different values on the first and second dimension.

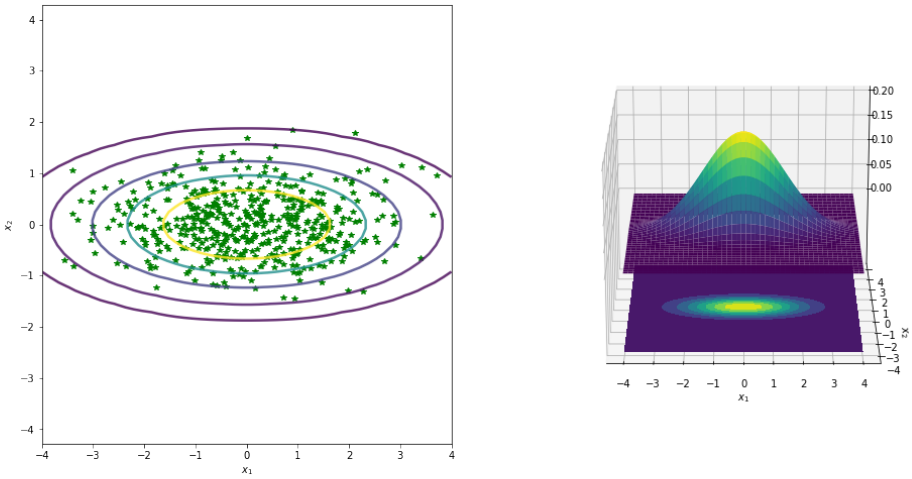


Figure 1.6: Plots corresponding to the standard multivariate Gaussian scaled using $\mathbf{A}_1$.

The third case will have the same plot as the first one, except that the samples will be rotated. The fourth case corresponds to composing scaling the rotation.

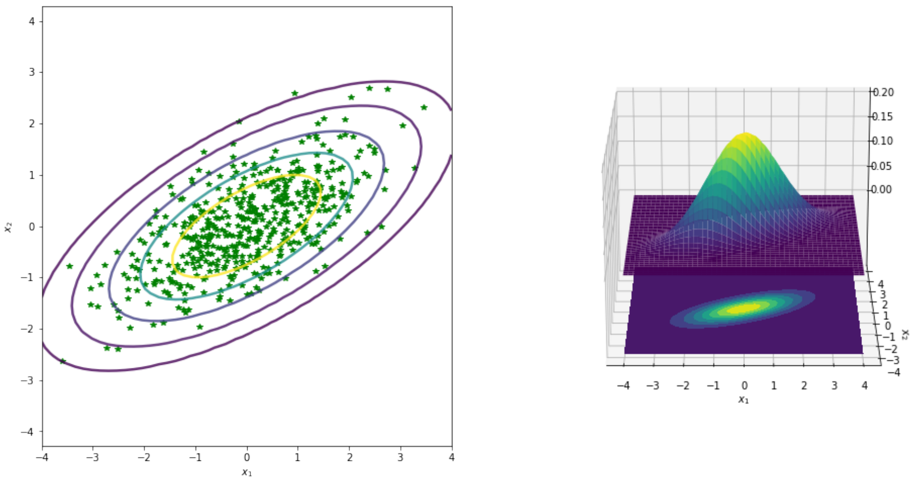


Figure 1.7: Plots corresponding to the standard multivariate Gaussian scaled using $\mathbf{A}_2\mathbf{A}_1$.

It is clear now that for every invertible matrix we can obtain a different Gaussian distribution. Because of the decomposition of positive definite matrices into $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ discussed before, we can also see that every Gaussian distribution corresponds to a linear transformation of the standard Gaussian distribution. Indeed, we can simply consider the entry-wise squared-root of $\mathbf{\Lambda}$, $\mathbf{L} = \mathbf{\Lambda}^{1/2}$ and define $\mathbf{A} = \mathbf{U}\mathbf{L}$, then $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^\top$, and $\mathbf{A}$ is the linear (invertible) transformation applied to the standard normal.
We can state further properties of the multivariate Gaussian distribution regarding its level curves, defined as:

$$\mathcal{C}_\lambda = \{\mathbf{x} | \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = \lambda\}. \tag{1.19}$$

These level curves are:

- For $\lambda < 0, \mathcal{C}_\lambda = \emptyset$.
- For $\lambda = 0, \mathcal{C}_0 = \{\boldsymbol{\mu}\}$.
- For $\lambda > 0, \mathcal{C}_\lambda$? is an ellipsoid with center $\boldsymbol{\mu}$, axis given by the columns of $\mathbf{U}$ and axis length given by the elements in $\mathbf{\Lambda}$, where $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$. See the figures above.

1.2.3 Properties of Multivariate Gaussians

Deriving Maximum Likelihood Estimators for Multivariate Gaussians In order to derive maximum likelihood estimators for the parameters of the multivariate Gaussian distribution, we need to know how to take derivatives w.r.t. vectors and matrices. We will use some results from the matrix cookbook [1]. Useful formulae:

$$\frac{\partial \mathrm{Tr}(\mathbf{M}\mathbf{A})}{\partial \mathbf{M}} = \mathbf{A}^\top \qquad \frac{\partial \mathrm{Tr}(\mathbf{B}^\top \mathbf{M}^\top \mathbf{C}\mathbf{M}\mathbf{B})}{\partial \mathbf{M}} = \mathbf{C}^\top \mathbf{M}\mathbf{B}\mathbf{B}^\top + \mathbf{C}\mathbf{M}\mathbf{B}\mathbf{B}^\top \tag{1.20}$$

$$\frac{\partial \log |\mathbf{M}|}{\partial \mathbf{M}} = (\mathbf{M}^{-1})^\top \qquad [\mathrm{Tr}(\mathbf{A}\mathbf{B}\mathbf{C}) = \mathrm{Tr}(\mathbf{B}\mathbf{C}\mathbf{A}) = \mathrm{Tr}(\mathbf{C}\mathbf{A}\mathbf{B})] \tag{1.21}$$

Exercise 1.7: Prove that the ML estimators of the multivariate Gaussian are:

$$\boldsymbol{\mu}^* = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \qquad \mathbf{\Sigma}^* = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}^*)(\mathbf{x}_n - \boldsymbol{\mu}^*)^\top. \tag{1.22}$$

The shape is all you need

Remark 1.5: By developing the expression of the multivariate Gaussian distribution:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}|}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big) \overset{\mathbf{x}}{\propto} \exp\Big(-\frac{1}{2}\mathbf{x}^\top \mathbf{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\top \mathbf{\Sigma}^{-1}\boldsymbol{\mu}\Big), \tag{1.23}$$

we observe that only two terms depend on $\mathbf{x}$. Both are in the exponential: one is quadratic and the other is linear.

The notation $\overset{\mathbf{x}}{\propto}$ means that is proportional up to a constant that does NOT depend on $\mathbf{x}$. This notation will be very useful along these notes. In other words, as long as the shape of the distribution is exponential with a quadratic and a linear term on the random variable, then the distribution is Gaussian. The only constraint is that the quadratic term must have negative signe and the matrix should be symmetric and positive definite.

Exercise 1.8: Prove that given a symmetric and positive definite matrix $\mathbf{\Omega}$ and a vector $\mathbf{m}$, a probability distribution of the form:

$$p(\mathbf{x}) \overset{\mathbf{x}}{\propto} \exp\Big(-\frac{1}{2}\mathbf{x}^\top \mathbf{\Omega}\mathbf{x} + \mathbf{x}^\top \mathbf{m}\Big) \tag{1.24}$$

corresponds to a multivariate Gaussian distribution with the following covariance matrix and mean vector:

$$\mathbf{\Sigma} = \mathbf{\Omega}^{-1} \qquad \boldsymbol{\mu} = \mathbf{\Sigma}\mathbf{m} = \mathbf{\Omega}^{-1}\mathbf{m}. \tag{1.25}$$

More on multivariate Gaussians in Chapter 4.

1.3 Latent Variables and Conditional Independence

1.3.1 Introduction

We will be using *models* between variables. In our case this model describes the relationships between two or more variables, see Figure 1.8. In practice this means that we choose:

- the nature of $\mathbf{z}$ & $\mathbf{x}$: cont./discrete, bounded, ...
- the dependencies, i.e. $p(\mathbf{x},\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$.
- the prior distribution $p(\mathbf{z})$.
- the likelihood distribution $p(\mathbf{x}|\mathbf{z})$.

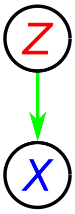


Figure 1.8: Simple model with two variables.

Remark 1.6: For the rest of the courses, we will make the difference between **observed** and **latent** or hidden variables. Observed variables are the samples obtained via measuring a system, and latent variables are quantities that cannot be measured directly.

For instance, in the clustering task, the observations are the positions of the data points in the $\mathbb{R}^D$ space (measured), and the latent variables are the class index of each of the samples (not observed). Usually, $\mathbf{x}$ will denote the observation variable and $\mathbf{z}$ the hidden or latent variable.

Remark 1.7: The two classical questions in probabilistic models with latent variables, once the prior and likelihood distributions are defined are the **marginal distribution** of $\mathbf{x}$ (left) and the **posterior distribution** of $\mathbf{z}$ given $\mathbf{x}$ (right):

$$p(\mathbf{x}) = \int_{\mathcal{Z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \qquad p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} \tag{1.26}$$

Formally, both $p(\mathbf{z})$ and $p(\mathbf{x})$ are marginal distributions. Because of the sense of the dependency, the former is referred to as the marginal, while the latter as the prior. The same happens with $p(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{z})$: both are conditional or posterior distributions, but their name is different because of the sense of the dependency.

Example: Gaussian Mixture Model The one-dimensional Gaussian mixture model (GMM) can be defined as follows:

- The nature: z is discrete & bounded, x is 1D & continuous.
- The dependencies are: $p(x,z) = p(x|z)p(z)$.
- The prior distribution $p(z)$, $z \in \{1,\ldots,K\}$ is categorical:

$$p(z=k) = \pi_k, \qquad \pi_k \geq 0, \ \sum_{k=1}^{K} \pi_k = 1. \tag{1.27}$$

- The likelihood distribution $p(x|z)$ is Gaussian:

$$p(x|z=k) = \mathcal{N}(x;\mu_k,\nu_k) = \frac{1}{\sqrt{2\pi\nu_k}} \exp\Big(-\frac{(x-\mu_k)^2}{2\nu_k}\Big) \tag{1.28}$$

with $\mu_k \in \mathbb{R}$ and $\nu_k > 0, \forall k$.

Exercise 1.9: Prove that given the prior and likelihood distributions of the GMM, the marginal writes:

$$p(x) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\nu_k}} \exp\Big(-\frac{(x-\mu_k)^2}{2\nu_k}\Big). \tag{1.29}$$

Exercise 1.10: Prove that given the prior and likelihood distribution of the GMM, the posterior writes:

$$p(z=k|x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\nu_k}} \exp\Big(-\frac{(x-\mu_k)^2}{2\nu_k}\Big)}{\sum_{m=1}^{K} \pi_m \frac{1}{\sqrt{2\pi\nu_m}} \exp\Big(-\frac{(x-\mu_m)^2}{2\nu_m}\Big)}. \tag{1.30}$$

More on this on the Chapter 2.

1.3.2 Conditional Independence

We will start study conditional independence with the smallest model that can be used: 3 variables. There are four types of 3-variable models, as depicted in Figure 1.9:

Full all dependencies are set:

$$p(x,y,z) = p(z|x,y)p(y|x)p(x) \tag{1.31}$$

Two-kids $y \to z$ dependency missing:

$$p(x,y,z) = p(z|x,y)p(y|x)p(x) \tag{1.32}$$

Two-parents $x \to y$ dependency missing:

$$p(x,y,z) = p(z|x,y)p(y|x)p(x) \tag{1.33}$$

Cascaded $x \to z$ dependency missing:

$$p(x,y,z) = p(z|x,y)p(y|x)p(x) \tag{1.34}$$

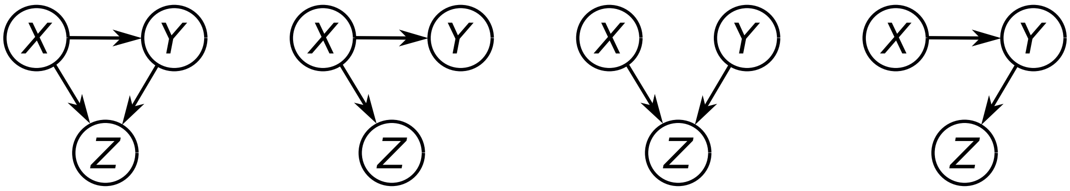


Figure 1.9: Three-variable models, from left to right: Full (all dependencies are used), Two-kids ($y \to z$ dependency missing), Two-parents ($x \to y$ dependency missing) and Cascaded ($x \to z$ dependency missing).

We will let aside the Full model, since we are interested in understanding the imapct of the missing dependencies. Let's analyse the Two-kids model in depth.

Exercise 1.11: Prove that in the Two-kids model:

$$p(y|z) \neq p(y) \qquad \text{and} \qquad p(y|z,x) = p(y|x) \tag{1.35}$$

The first statement is equivalent to say that y and z are not independent. The second statement, however, says that y and z are **conditionally independent** w.r.t. x. Let us repeat a similar analysis with the Two-parents model.

Exercise 1.12: Prove that in the Two-parents model:

$$p(y|x) = p(x) \qquad \text{and} \qquad p(x|y,z) \neq p(x|z). \tag{1.36}$$

We are in the opposite case. The first statement says that y and x are independent, while the second statement says that y and x are conditionally dependent w.r.t. z. We will repeat this analysis for the Cascaded model.

Exercise 1.13: Prove that in the Cascaded model:

$$p(x,z) \neq p(x)p(z) \qquad \text{and} \qquad p(x,z|y) = p(x|y)p(z|y). \tag{1.37}$$

In this case, we obtain similar results than with the Two-kids model.

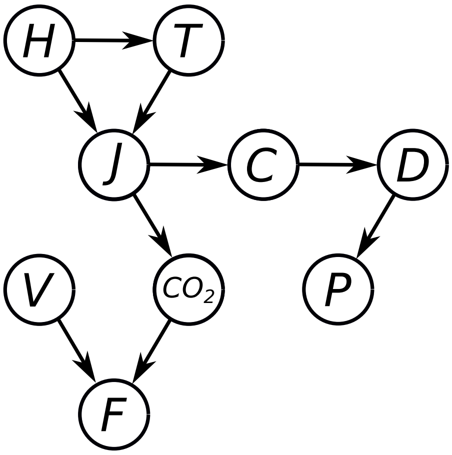
Remark 1.8: At this point it should be clear that independence and conditional independence are **two very different properties** of random variables.

Remark 1.9: Let x, y, and z be random variables, we say that x and y are **conditionally independent** given z, and write $x \perp\!\!\!\perp y \mid z$, iff one of the following equivalent expressions holds:

- $p(x, y|z) = p(x|z)p(y|z)$
- $p(x|y, z) = p(x|z)$
- $p(y|x, z) = p(y|z)$

1.3.3 D-separation

Let's consider the following variables and dependencies.



Is $P \perp\!\!\!\perp V \mid T$? How would you do it ? Is the previous strategy scalable ? Let us recall the 3-variable models:

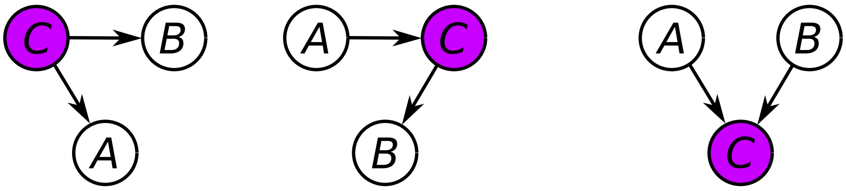


Figure 1.10: Recalling the 3-variable models with new variable names. From left to right: two-kids, cascaded and two-parents.

Two-kids The path from a to b is called "tail-to-tail."

$$p(a, b|c) = p(a|c)p(b|c) \tag{1.38}$$

Cascaded The path from a to b is called "head-to-tail."

$$p(a, b|c) = p(a|c)p(b|c) \tag{1.39}$$

Two-parents The path from a to b is called "head-to-head."

$$p(a, b|c) \neq p(a|c)p(b|c) \tag{1.40}$$

⇒ If $A \perp\!\!\!\perp B \,|\, C$, nodes within tail-to-tail or head-to-tail can be in C and nodes within head-to-head or any of their descendents must not be in C.

Remark 1.10: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A path from A to B is said to be **blocked** by C if it includes a node that either:

- the path meets tail-to-tail or head-to-tail at the node and the node is in C;
- the path meets head-to-head at the node and neither the node nor any of its descendants are in C.

Exercise 1.14: Regarding the graphical model in Figure 1.11, please answer the following questions:

1. Is the path from $\{x\}$ to $\{v\}$ blocked by $\{u\}$?
2. Is the path from $\{x\}$ to $\{v\}$ blocked by $\{y\}$?
3. Is the path from $\{x\}$ to $\{v\}$ blocked by $\{z\}$?
4. Is the path from $\{y\}$ to $\{v\}$ blocked by $\{u\}$?

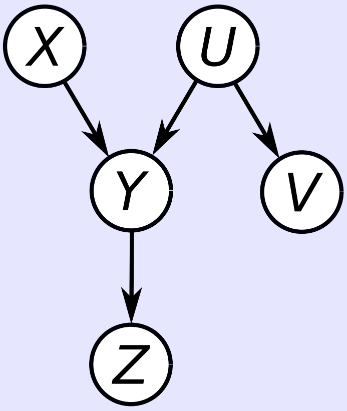


Figure 1.11: Example of graphical model to discuss path blocking.

Remark 1.11: Let A, B and C be three non-intersecting sets of nodes of a directed acyclic graph. A and B are **D-separated** by C, if all paths from any node from A to B are blocked by C.

Exercise 1.15: Regarding the graphical model in Figure 1.11, please answer the following questions:

- Is $\{x\}$ D-separated from $\{v\}$ by $\{u\}$?
- Is $\{x\}$ D-separated from $\{v\}$ by $\{y\}$?
- Is $\{x\}$ D-separated from $\{v\}$ by $\{y, u\}$?

Remark 1.12: A and B are D-separated by C if and only if $A \perp\!\!\!\perp B|C$.

1.3.4 Markovian Structures

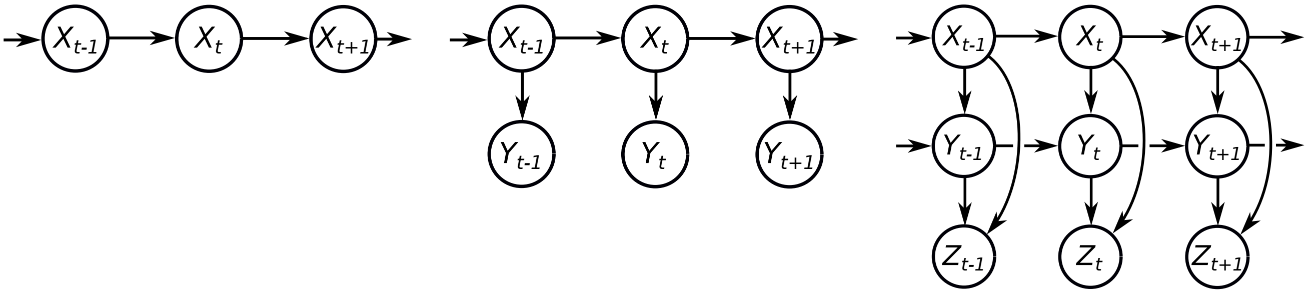
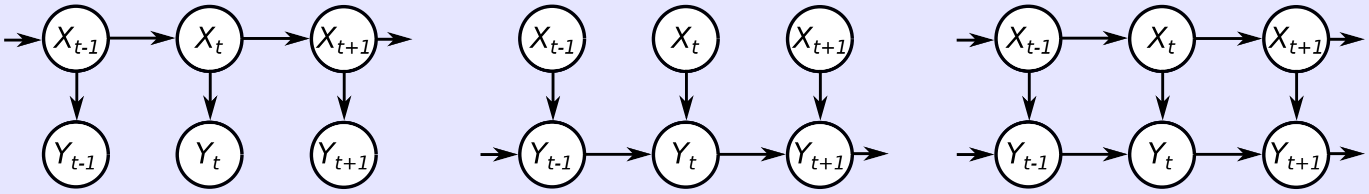


Figure 1.12: Examples of models with Markovian dependencies, from left to right: Markov chain, hidden Markov model, double hidden Markov model.

Exercise 1.16: For each of the three Markov models shown below, is $\{x_{t-1}\}$ D-separated from $\{y_{t+1}\}$ by:

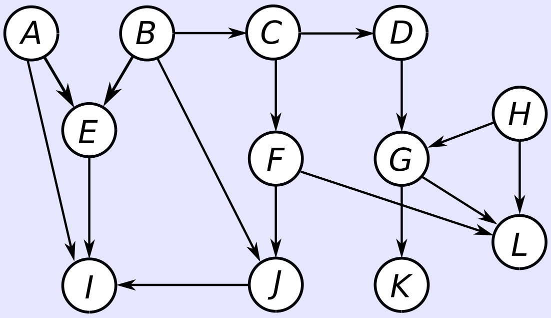
1. $\{x_t\}$?
2. $\{y_t\}$?
3. $\{x_t, y_t\}$?



Remark 1.13: Given a directed acyclic graph representing a probabilitic model, with random variable set Ω, for any given random variable $x \in \Omega$, its **Markov blanket (or boundary)** is defined as the smallest subset $\mathcal{B}_x$ of Ω such that:

$$p(x|\Omega/x) = p(x|\mathcal{B}_x) \tag{1.41}$$

Exercise 1.17: Given the probabilistic graphical model shown below, identify $\mathcal{B}_k$, $\mathcal{B}_l$, $\mathcal{B}_c$, $\mathcal{B}_e$.



Remark 1.14: Construction of the Markov blanket. Given a directed acyclic graph, and a node x on that graph, the Markov blanket of x, $\mathcal{B}_x$ is the set of all parents, children and co-parents of x.

Solutions to the Exercises

Solution to Exercise 1.1: In order to find the optimal value of the parameters in the maximum likelihood, we first develop the expression of $\mathcal{L}$:

$$\mathcal{L}(\mu,\nu|X) = \sum_{n=1}^{N} \log \mathcal{N}(x_n;\mu,\nu) = \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\nu}} - \frac{(x_n-\mu)^2}{2\nu} = -\frac{N}{2}\log(2\pi\nu) - \frac{1}{2\nu}\sum_{n=1}^{N}(x_n-\mu)^2 \tag{1.42}$$

And then compute $\frac{\partial \mathcal{L}}{\partial \mu}$ and $\frac{\partial \mathcal{L}}{\partial \nu}$:

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{\nu}\sum_{n=1}^{N}(x_n-\mu) \qquad \frac{\partial \mathcal{L}}{\partial \nu} = -\frac{N}{2\nu} + \frac{1}{2\nu^2}\sum_{n=1}^{N}(x_n-\mu)^2 \tag{1.43}$$

By setting the derivatives to zero:

$$\mu^* = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \nu^* = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu^*)^2. \tag{1.44}$$

Solution to Exercise 1.2: By developping the log-likelihood, we can easily observe that the total log-likelihood is the sum of D terms, one corresponding to each dimension of the observations. Therefore the ML values of $\mathbf{mu}$ and $\boldsymbol{\nu}$ can be computed per-dimension, and the univariate ML estimator formulae can be used.

Solution to Exercise 1.3: Just write things down.

Solution to Exercise 1.4: As explained, all eigenvalues of a symmetric and positive definite matrix are positive. Therefore, the determinant is strictly positive, since it is the product of the eigenvalues. Since the determinant is not zero, the matrix is invertible.

In addition, the eigenvalues of the inverse matrix correspond to the inverse of the eigenvalues of $\boldsymbol{\Sigma}$, and they will all be positive. Since the inverse will also be symmetric, then the inverse of a symmetric and positive definite matrix is also symmetric and positive definite.

Solution to Exercise 1.5: The trick is to start by using the decomposition $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$, where $\boldsymbol{\Lambda}$ is diagonal with strictly positive entries, and $\mathbf{U}$ is orthogonal. We then consider $\mathbf{L} = \boldsymbol{\Lambda}^{1/2}$, meaning the entry-wise squared root, and realise that: $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{L}\mathbf{L}^\top\mathbf{U}^\top$. With this notation we can rewrite the inside of the exponential to compute the normalisation constant:

$$\begin{aligned}
\mathcal{C}(\boldsymbol{\Sigma}) &= \int_{\mathbb{R}^D} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\mathrm{d}\mathbf{x} && (1.45)\\
&= \int_{\mathbb{R}^D} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{U}\mathbf{L}^{-1}\mathbf{L}^{-1}\mathbf{U}^\top(\mathbf{x}-\boldsymbol{\mu})\Big)\mathrm{d}\mathbf{x} && (1.46)\\
&= |\mathbf{L}|\int_{\mathbb{R}^D} \exp\Big(-\frac{1}{2}\mathbf{y}^\top\mathbf{y}\Big)\mathrm{d}\mathbf{y} && (1.47)\\
&= |\boldsymbol{\Sigma}|^{1/2}(2\pi)^{D/2} && (1.48)\\
&= |2\pi\boldsymbol{\Sigma}|^{1/2}, && (1.49)
\end{aligned}$$

where we used the change of variables $\mathbf{y} = \mathbf{L}^{-1}\mathbf{U}^\top(\mathbf{x}-\boldsymbol{\mu})$, leading to $\mathrm{d}\mathbf{y} = |\mathbf{L}|^{-1}\mathrm{d}\mathbf{x}$ (where we used that $\mathbf{U}$ is orthogonal).

Solution to Exercise 1.6: By definition, we have that $\mathbf{z} = \mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. By taking the definition of the standard normal distribution, we write:

$$\exp\Big(-\frac{1}{2}\mathbf{z}^\top\mathbf{z}\Big) = \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{A}^{-\top}\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big) = \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top(\mathbf{A}\mathbf{A}^\top)^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big). \qquad (1.50)$$

By setting $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$ we conclude the proof (notice that the normalisation coefficient must be adapted following the variable change, as in Exercise 1.5.

Solution to Exercise 1.7: By considering a set of i.i.d. observations $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, we can develop the log-likelihood:

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = \sum_{n=1}^{N} \log \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1.51}$$

$$= \sum_{n=1}^{N} -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \tag{1.52}$$

$$= -\frac{1}{2}\Big(DN\log(2\pi) + N\log|\boldsymbol{\Sigma}| + \sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\Big). \tag{1.53}$$

We now focus on the last term, and redevelop it as follows:

$$(\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = \mathrm{Tr}\big((\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\big) = \mathrm{Tr}\big(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top\big), \tag{1.54}$$

where we used the circular property of the trace. Regarding the optimal value of $\boldsymbol{\mu}$, we can use the other derivative formulae shown in the main text and write:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}\big(\mathbf{x}_n - \boldsymbol{\mu}\big) = 0 \Leftrightarrow \boldsymbol{\mu}^* = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n. \tag{1.55}$$

We are now ready to take the derivative w.r.t. $\boldsymbol{\Sigma}$. We will actually take the derivative w.r.t. $\boldsymbol{\Sigma}^{-1}$:

$$\frac{\partial \mathcal{L}}{\partial(\boldsymbol{\Sigma}^{-1})} = N\boldsymbol{\Sigma} - \sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top = 0 \Leftrightarrow \boldsymbol{\Sigma}^* = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}^*)(\mathbf{x}_n - \boldsymbol{\mu}^*)^\top. \tag{1.56}$$

Solution to Exercise 1.8: In order to prove this, we will first notice that the quadratic term has to be associated to the covariance distribution, hence: $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$, leading to:

$$p(\mathbf{x}) \overset{\mathbf{x}}{\propto} \exp\Big(-\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\top \mathbf{m}\Big). \tag{1.57}$$

We can now multiply the linear term by $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}$:

$$p(\mathbf{x}) \overset{\mathbf{x}}{\propto} \exp\Big(-\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\mathbf{m}\Big). \tag{1.58}$$

By setting $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{m}$ we obtain:

$$p(\mathbf{x}) \overset{\mathbf{x}}{\propto} \exp\Big(-\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big). \tag{1.59}$$

By multiplying the right-hand side of the above expression by $|2\pi\boldsymbol{\Sigma}|^{1/2}\exp(-\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/2)$, which is a constant that does not depend on $\mathbf{x}$ we end up with:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{1.60}$$

Solution to Exercise 1.9: The definition of the marginal distribution writes:

$$p(x) = \int_{\mathcal{Z}} p(x|z)p(z)\mathrm{d}z = \sum_{k=1}^{K} p(x|z=k)p(z=k). \tag{1.61}$$

And then we replace:

$$p(x) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\nu_k}} \exp\Big(-\frac{(x-\mu_k)^2}{2\nu_k}\Big). \tag{1.62}$$

Solution to Exercise 1.10: Use the Bayes rule, and write down the definitions and the previously computed marginal distribution.

Solution to Exercise 1.11: Let us write what the equality would imply:

$$p(y|z) = \frac{p(y,z)}{p(z)} = \frac{\int_{\mathcal{X}} p(x,y,z)\mathrm{d}x}{p(z)} = \frac{\int_{\mathcal{X}} p(z|x)p(y|x)p(x)\mathrm{d}x}{p(z)} \overset{?}{=} \int_{\mathcal{X}} p(y|x)p(x)\mathrm{d}x = p(y). \tag{1.63}$$

In other words, it would imply that $p(z|x) = p(z)$, $\forall x$, meaning that z is independent of x, which is against the model definition. Regarding the second statement:

$$p(y|z,x) = \frac{p(y,z,x)}{p(z,x)} = \frac{p(y|x)p(z|x)p(x)}{p(z|x)p(x)} = p(y|x). \tag{1.64}$$

Solution to Exercise 1.12: Regarding the first statement:

$$p(y|x)p(x) = p(y,x) = \int_{\mathcal{Z}} p(z,x,y)\mathrm{d}z = p(x)p(y)\int_{\mathcal{Z}} p(z|x,y)\mathrm{d}z = p(x)p(y) \Rightarrow p(y|x) = p(y). \tag{1.65}$$

Regarding the second statement, we impose the equality:

$$p(x|y,z) = \frac{p(x,y,z)}{p(y,z)} = \frac{p(z|y,x)p(x)p(y)}{p(y)\int_{\mathcal{X}} p(z|y,x)p(x)\mathrm{d}x} \overset{?}{=} p(x|z), \tag{1.66}$$

which implies that $p(z|y,x)$ does not depend on y and is contrary to the model.

Solution to Exercise 1.13: Regarding the first statement:

$$p(x,z) = \int_{\mathcal{Y}} p(x,y,z)\mathrm{d}y = \int_{\mathcal{Y}} p(z|y)p(y|x)p(x)\mathrm{d}x = p(x)\int_{\mathcal{Y}} p(z|y)p(y|x)\mathrm{d}y \overset{?}{=} p(x)p(y), \tag{1.67}$$

which implies that $p(y|x)$ does not depend on x. Regarding the second statement:

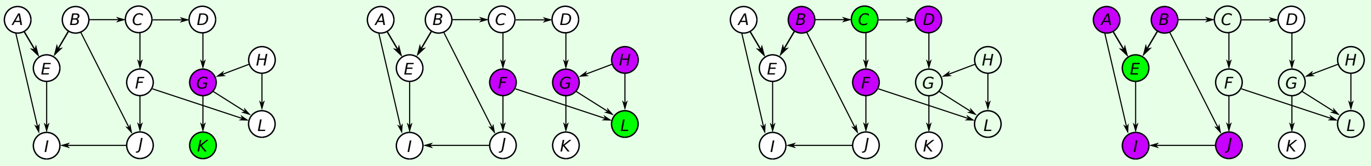
$$p(x,z|y) = \frac{p(x,z,y)}{p(y)} = \frac{p(z|y)p(y|x)p(x)}{p(y)} = p(x|y)p(z|y). \tag{1.68}$$

Solution to Exercise 1.14: 1. Yes, 2. No, 3. No, 4. Yes.

Solution to Exercise 1.15: 1. Yes, 2. No, 3. Yes.

Solution to Exercise 1.16: With the left model: 1.Yes, 2. No, 3. Yes. With the center model: 1. No, 2. Yes, 3. Yes. With the right model: 1. No, 2. No, 3. Yes.

Solution to Exercise 1.17: The isolated variables are shown and their associated Markov blanket in purple.



Chapter 2

Gaussian Mixture Models

2.1 Motivation: Clustering

Suppose you are given the set of points in the left hand side of Figure 2.1, and you are asked to devise an algorithm to automatically find K clusters. One option is the following:

1. Initialise randomly K centroids.
2. Assign each data point to the closes centroid.
3. Recompute centroids from the assignments.
4. Iterate the past two steps.

This is called the K-means algorithm, and it results on the color-coded assignment shown in the center of Figure 2.1.

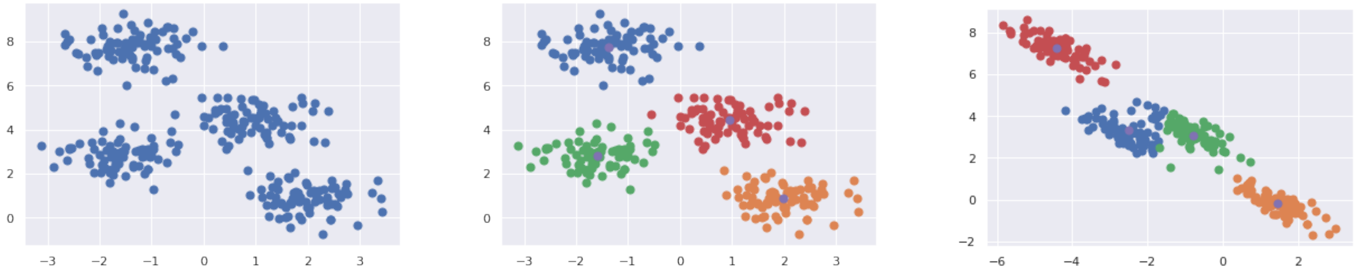


Figure 2.1: Clustering example.

Important properties of the K-means algorithm:

- Automatic inference of latent variables. The point-to-cluster assignment variable is **unknown/latent/hidden**, and must be infered together with the parameters.
- Limited to spherical and equally populated clusters. Because the assignment criterion is the Euclidean distances, see the right hand side of Figure 2.1.

Remark 2.1: Gaussian mixture model (GMM):

- For each data point $\mathbf{x}_n$ there is a hidden variable z_n taking discrete values from 1 to K: $z_n \in \{1, \ldots, K\}$.
- Its prior probability is defined as: $p(z_n = k) = \pi_k$, $\sum_{k=1}^{K} \pi_k = 1$.
- Given z_n, the data point is modeled as a multivariate Gaussian:

$$p(\mathbf{x}_n | z_n = k) = \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.1}$$

Advantages:

1. Having $\pi_1, \ldots, \pi_K$ means that groups can be differently populated.
2. The shape of the groups is modeled by $\boldsymbol{\Sigma}_k$.

2.2 Maximum Likelihood for GMM: the Expectation-Maximisation algorithm

Let's compute $p(\mathbf{x}_n)$:

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} p(\mathbf{x}_n, z_n = k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2.2}$$

The log-likelihood writes:

$$\mathcal{L}(\boldsymbol{\Theta}|\mathbf{X}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.3}$$

with $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$. One can easily check by computing $\frac{\partial \mathcal{L}}{\partial \pi_k}$, $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k}$ or $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k}$ that direct optimisation is very complicated.

Notice that $\log p(\mathbf{x})$ is not a nice function to take the derivative of, but that $\log p(\mathbf{x}, z)$ is. However, z is not observed and we need to take some sort of expectation. We will do that w.r.t. the posterior distribution, and this choice will be justified later on.

Remark 2.2: The expected complete-data log-likelihood writes:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \mathbb{E}_{p(z|\mathbf{x};\boldsymbol{\Theta}^0)} \log p(\mathbf{x}, z; \boldsymbol{\Theta}) \tag{2.4}$$

Remark 2.3: We can now propose the EM algorithnm for GMM, with some notation first: observations $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$, latent variables $\mathbf{Z} = \{z_n\}_{n=1}^{N}$ and parameters $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$. Given $\boldsymbol{\Theta}^0$, we use the expected complete-data log-likelihood $\mathcal{Q}$:

1. Expectation:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X};\boldsymbol{\Theta}^0)} \log p(\mathbf{Z}, \mathbf{X}; \boldsymbol{\Theta}) \tag{2.5}$$

2. Maximisation:

$$\boldsymbol{\Theta}^1 = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) \tag{2.6}$$

We can look back to K-means:

1. Infer latent variables (assignment) given the parameters (centroids).
2. Estimate the parameters (centroids) given the assignments.

Algorithm	K-means	EM for GMM
Criterion	Sum Euclidean distance to assigned cluster center	Expected complete-data log-likelihood
Inference	Optimal hard-assignment	Posterior distribution of z_n
Param. Est.	Cluster centroids	Optimise for $\boldsymbol{\Theta}$

To derive the expectation (E) and maximisation (M) steps, of the EM algorithm, we recall: $p(z_n = k) = \pi_k$ and $p(\mathbf{x}_n|z_n = k) = \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

2.2.1 The Expectation Step

Compute $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X};\boldsymbol{\Theta}^0)} \log p(\mathbf{Z}, \mathbf{X}; \boldsymbol{\Theta})$.

- Start with $p(z_n = k|\mathbf{x}_n; \boldsymbol{\Theta}^0)$. And name it $\eta_{nk} = p(z_n = k|\mathbf{x}_n; \boldsymbol{\Theta}^0)$. η_{nk} is the posterior probability that $\mathbf{x}_n$ belongs to group k.
- Then $p(\mathbf{x}_n, z_n|\boldsymbol{\Theta}) = \pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.
- Also $\mathbb{E}_{p(z_n|\mathbf{Z}_n;\boldsymbol{\Theta}^0)} \log p(z_n, \mathbf{x}_n; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \eta_{nk} \log \pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Remark 2.4: In the case of Gaussian mixture models, the expected complete-data log-likelihood writes:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk} \log \pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2.7}$$

2.2.2 The Maximisation Step

The exected complete-data log-likelihood splits:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk} \log \pi_k + \eta_{nk} \log \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2.8}$$

We consider now the Lagrangian for $\pi_1, \ldots, \pi_K$:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk} \log \pi_k + \beta\Big(1 - \sum_{k=1}^{K} \pi_k\Big). \tag{2.9}$$

Exercise 2.1: By computing the derivatives of the previous function, prove that:

$$\pi_k^* = \frac{1}{N} S_k, \qquad S_k = \sum_{n=1}^{N} \eta_{nk}, \tag{2.10}$$

and

$$\boldsymbol{\mu}_k^* = \frac{1}{S_k}\sum_{n=1}^{N} \eta_{nk}\mathbf{x}_n \qquad \boldsymbol{\Sigma}_k^* = \frac{1}{S_k}\sum_{n=1}^{N} \eta_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k^*)(\mathbf{x}_n - \boldsymbol{\mu}_k^*)^\top. \tag{2.11}$$

2.3 The EM algorithm in general

Let us assume a probabilistic graphical model, with observed variables $\mathbf{X}$, hidden variables $\mathbf{z}$ and parameters $\boldsymbol{\Theta}$.

Remark 2.5: Initialise the parameters $\boldsymbol{\Theta}^0$. For iteration $r = 1, \ldots, R$:

E-step Compute $p(\mathbf{z}|\mathbf{X}; \boldsymbol{\Theta}^{r-1})$ and $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{r-1})$.

M-step Compute $\boldsymbol{\Theta}^r = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{r-1})$.

Comments

- EM is sensible to initialisation.
- It may converge to a local maxima or saddle point.
- We still need to compute and optimise $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{r-1})$.

Why does the EM work? The main mathematical object in EM is $\mathcal{Q}$ (the expected complete-data log-likelihood). In order to study the relationship with the log-likelihood, we consider any distribution of $\mathbf{z}$: $q(\mathbf{z})$ and ignore $\boldsymbol{\Theta}$ for the time being.

$$\begin{aligned}
\log p(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{z})}\Big[\log p(\mathbf{x})\Big] && (2.12)\\
&= \mathbb{E}_{q(\mathbf{z})}\Big[\log p(\mathbf{x}) \frac{p(\mathbf{z}|\mathbf{x})q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})q(\mathbf{z})}\Big] && (2.13)\\
&= \mathbb{E}_{q(\mathbf{z})}\Big[\log \frac{p(\mathbf{x})p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})}\Big] + D_{\mathsf{KL}}\Big(q(\mathbf{z})\Big\|p(\mathbf{z}|\mathbf{x})\Big) && (2.14)
\end{aligned}$$

We therefore obtain:

$$\log p(\mathbf{x}; \boldsymbol{\Theta}) = \underbrace{\mathbb{E}_{q(\mathbf{z})}\Big[\log \frac{p(\mathbf{x})p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})}\Big]}_{\text{M-step}} + \underbrace{D_{\mathsf{KL}}\Big(q(\mathbf{z})\Big\|p(\mathbf{z}|\mathbf{x})\Big)}_{\text{E-step}}, \tag{2.15}$$

where D_{KL} stands for the Kullback-Leibler divergence, and is always positive. In these terms, the EM can be interpreted as follows, for a given $\boldsymbol{\Theta}^0$:

1. Set $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}; \boldsymbol{\Theta}^0)$.
2. Optimise $\mathbf{E}_{q(\mathbf{z})}\Big[\log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\Theta})}{q(\mathbf{z})}\Big]$ w.r.t. $\boldsymbol{\Theta}$.

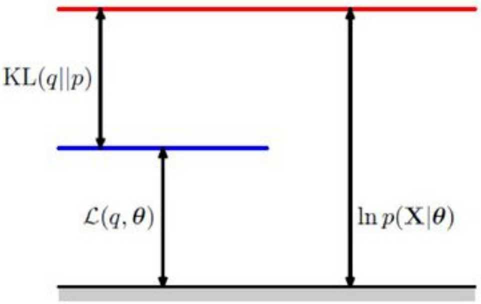
By running the E-step (step 1), we make $D_{\text{KL}} = 0$, and therefore the expected complete-data log-likelihood becomes a tight lower bound of the log-likelihood. The M-step (step 2), thus optimises directly the log-likelihood. In other words:

E-step reduce the distance between log-likelihood and $\mathcal{Q}$.

M-step push $\mathcal{Q}$ and therefore push the log-likelihood.

A graphical representation of this phenomenon can be found in the following figure with the following notations:

$$\log p(\mathbf{x};\boldsymbol{\Theta}) = \underbrace{\mathbb{E}_{q(\mathbf{z})}\left[\log\frac{p(\mathbf{x})p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})}\right]}_{\mathcal{L}(q,\boldsymbol{\Theta})} + \underbrace{D_{\text{KL}}\Big(q(\mathbf{z})\Big\|p(\mathbf{z}|\mathbf{x})\Big)}_{\text{KL}(q\|p)}$$


$\mathrm{KL}(q||p)$
$\mathcal{L}(q,\boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$

Solutions to the Exercises

Solution to Exercise 2.1: For $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ re-use the same strategy used for the ML estimators of a multivariate Gaussian distribution in Chapter 1, see Exercise 1.7. Regarding the π_k, we need to compute the derivative:

$$\frac{\partial Q}{\partial \pi_k} = \sum_{n=1}^{N} \eta_{nk}\frac{1}{\pi_k} - \beta. \tag{2.16}$$

By setting the derivative to zero, we obtain:

$$\pi_k = \frac{1}{\beta}\sum_{n=1}^{N}\eta_{nk}, \tag{2.17}$$

and we recall that $\sum_k \pi_k = 1$, thus:

$$1 = \sum_{k=1}^{K}\pi_k = \frac{1}{\beta}\sum_{k,n=1}^{K,N}\eta_{nk} = \frac{N}{\beta} \Rightarrow \pi_k^* = \frac{1}{N}\sum_{k=1}^{N}\eta_{nk}. \tag{2.18}$$

Chapter 3

Hidden Markov Models

(Empty)

Chapter 4

Probabilistic Principal Component Analysis

4.1 From Discrete to Continuous Latent Variables

So far we have studied probabilistic models with discrete latent variables, that is $z \in \{1, \ldots, K\}$. In particular we have discussed the Gaussian mixture model and the hidden Markov model, depicted in Figure 4.1.

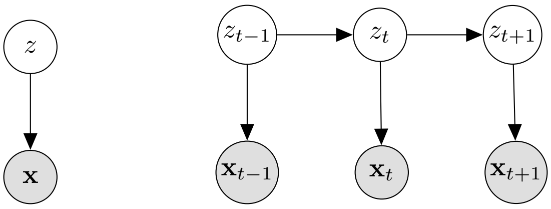


Figure 4.1: Graphical representation of a GMM (left) and an HMM (right).

In the following courses we will discuss models with continuous z. This means that $\mathbf{z} \in \mathbb{R}$. The continuous equivalent of GMM is called probabilistic principal component analysis (PPCA), and the one corresponding to HMM is called Linear Dynamical System (LDS) and will be discussed in Chapter 5.

4.2 The PPCA Model

The PPCA model uses some of the concepts and intuitions around the multivariate Gaussian distribution discussed in Chapter 1. From a graphical perspective it is the same as GMM (left of Figure 4.1), but the nature of the random variables and their interdependencies change.

Remark 4.1: The PPCA model is characterised by:

- Two continuous variables, one hidden one observed. Traditionally, $\mathbf{z} \in \mathbb{R}^{d_\mathsf{z}}$ and $\mathbf{x} \in \mathbb{R}^{d_\mathsf{x}}$ denote the hidden and observed dimensions, and $d_\mathsf{z} \ll d_\mathsf{x}$.
- The prior on the latent variable is a standard Gaussian:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}). \tag{4.1}$$

- The conditional likelihood is a multivariate Gaussian:

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mathbf{A}\mathbf{z} + \mathbf{b}, \nu\mathbf{I}). \tag{4.2}$$

Exercise 4.1: What is the number of free parameters of the PPCA model?

Alternatively, one can write $\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{b} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \nu\mathbf{I})$, being $\boldsymbol{\epsilon}$ and $\mathbf{z}$ independent. An example of the generative process of PPCA can be found in Figure 4.2.

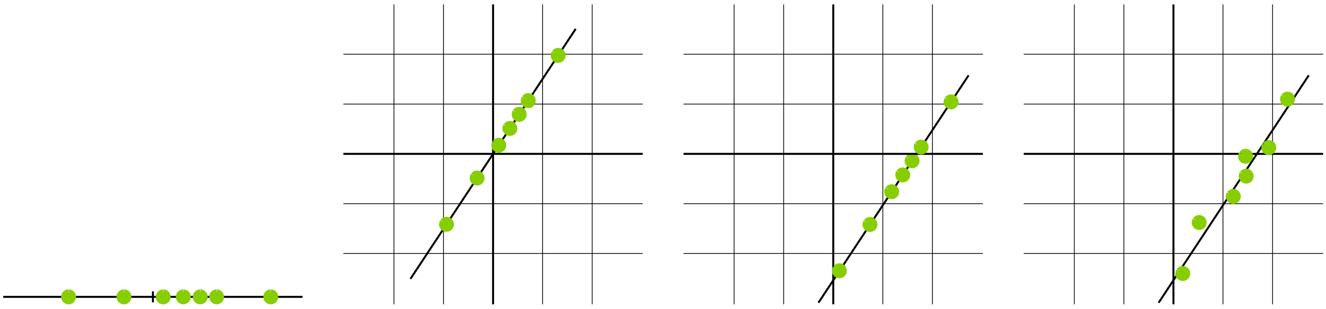


Figure 4.2: Example of the generative process of PPCA. From left to right. We first generate the low-dimensional latent variable $\mathbf{z}$, which gets multiplied by the linear operator $\mathbf{A}$, and added the bias $\mathbf{b}$, before adding the high-dimensional noise $\boldsymbol{\epsilon}$, to obtain the observations $\mathbf{x}$.

Exercise 4.2: Prove that, given the PPCA model, the marginal of $\mathbf{x}$ has mean vector $\mathbf{b}$ and covariance matrix $\mathbf{A}\mathbf{A}^\top + \nu\mathbf{I}$.

Exercise 4.3: Prove that the shape of the marginal distribution $p(\mathbf{x})$ corresponds to the one of a multivariate Gaussian. To do so, first prove that the joint distribution in $\mathbf{x}$ and $\mathbf{z}$ can be expressed as:

$$p(\mathbf{x},\mathbf{z}) \overset{\mathbf{x},\mathbf{z}}{\propto} \exp\left(-\frac{1}{2}\left[\frac{\|\mathbf{x}-\mathbf{b}\|^2}{\nu} - \mathbf{m}^\top\boldsymbol{\Omega}^{-1}\mathbf{m}\right]\right)\mathcal{N}(\mathbf{z};\mathbf{m},\boldsymbol{\Omega}), \tag{4.3}$$

with

$$\boldsymbol{\Omega} = \left(\frac{\mathbf{A}^\top\mathbf{A}}{\nu} + \mathbf{I}\right)^{-1} \qquad \mathbf{m} = \boldsymbol{\Omega}\left(\frac{\mathbf{A}^\top(\mathbf{x}-\mathbf{b})}{\nu}\right). \tag{4.4}$$

While it is clear from the previous equation that the mean of the marginal distribution is $\mathbf{b}$, understanding that the corresponding covariance matrix is what was found in Exercise 4.2 is not trvial. You might use the **Woodbury matrix inversion lemma**:

Remark 4.2: For matrices $\mathbf{D} \in \mathbb{R}^{n\times n}$, $\mathbf{U} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{m\times n}$, the following identity holds:

$$(\mathbf{D}+\mathbf{U}\mathbf{C}\mathbf{V})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{U}(\mathbf{C}^{-1}+\mathbf{V}\mathbf{D}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{D}^{-1}. \tag{4.5}$$

Exercise 4.4: Prove the Woodbury matrix inversion lemma.

4.3 Expectation-Maximisation for PCCA

We will now derive the EM algorithm for PPCA. The parameters of the model are $\boldsymbol{\Theta} = \{\mathbf{A},\mathbf{b},\nu\}$. As usual, the EM will provide a way to estimate these parameters provided a set of data $\mathbf{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$. For each of these data points, we assume the existence of a low-dimensional hidden random variable $\mathbf{z}_n$, and write $\mathbf{Z} = \{\mathbf{z}_1,\ldots,\mathbf{z}_N\}$.

Given an initialisation of the parameter set, $\bar{\boldsymbol{\Theta}} = \{\bar{\mathbf{A}},\bar{\mathbf{b}},\bar{\nu}\}$, the EM algorithm considers the expecteded complete-data log-likelihood:

$$\mathcal{Q}(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X};\bar{\boldsymbol{\Theta}})}\log p(\mathbf{X},\mathbf{Z};\boldsymbol{\Theta}) = \sum_{n=1}^{N}\mathbb{E}_{p(\mathbf{z}_n|\mathbf{x}_n;\bar{\boldsymbol{\Theta}})}\log p(\mathbf{x}_n,\mathbf{z}_n;\boldsymbol{\Theta}). \tag{4.6}$$

We already know that the posterior distribution is a Gaussian distribution that writes:

$$\mathcal{N}(\mathbf{z}_n;\bar{\mathbf{m}}_n,\bar{\boldsymbol{\Omega}}), \quad \text{with} \quad \bar{\boldsymbol{\Omega}} = \left(\frac{\bar{\mathbf{A}}^\top\bar{\mathbf{A}}}{\bar{\nu}} + \mathbf{I}\right)^{-1} \quad \text{and} \quad \bar{\mathbf{m}}_n = \bar{\boldsymbol{\Omega}}\left(\frac{\bar{\mathbf{A}}^\top(\mathbf{x}_n-\bar{\mathbf{b}})}{\bar{\nu}}\right). \tag{4.7}$$

In order to compute the expecation in $\mathcal{Q}$, we need the following result.

Exercise 4.5: The following two formulae hold:

$$\mathbb{E}_{\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})}\{\mathbf{A}\mathbf{z}\} = \mathbf{A}\boldsymbol{\mu} \quad \text{and} \quad \mathbb{E}_{\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})}\{\mathbf{z}^\top\boldsymbol{\Lambda}\mathbf{z}\} = \boldsymbol{\mu}^\top\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathrm{Tr}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}). \tag{4.8}$$

4.3.1 E-step

We are now ready to derive the E-step of the EM algorithm for PPCA:

Exercise 4.6: Show that the n-th term of the sum of the $\mathcal{Q}$ function for PPCA writes:

$$\mathbb{E}_{\mathcal{N}(\mathbf{z}_n;\bar{\mathbf{m}}_n,\bar{\Omega})}\left\{\log\mathcal{N}(\mathbf{x}_n;\mathbf{A}\mathbf{z}_n+\mathbf{b},\nu\mathbf{I})\mathcal{N}(\mathbf{z}_n;\mathbf{0},\mathbf{I})\right\} \overset{\boldsymbol{\Theta}}{=} -\frac{1}{2}\left(d_{\mathsf{x}}\log\nu+\frac{1}{\nu}\|\mathbf{x}_n-\mathbf{A}\bar{\mathbf{m}}_n-\mathbf{b}\|^2+\frac{1}{\nu}\mathrm{Tr}(\mathbf{A}^\top\mathbf{A}\bar{\boldsymbol{\Omega}})\right), \quad (4.9)$$

where the notation $\overset{\boldsymbol{\Theta}}{=}$ means equality up to an additive constant that does not depend on $\boldsymbol{\Theta}$.

Therefore the $\mathcal{Q}$ function for PPCA writes:

$$\mathcal{Q}(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}) \overset{\boldsymbol{\Theta}}{=} -\frac{N}{2}\left(d_{\mathsf{x}}\log\nu+\frac{1}{\nu}\mathrm{Tr}(\mathbf{A}^\top\mathbf{A}\bar{\boldsymbol{\Omega}})+\frac{1}{N\nu}\sum_{n=1}^{N}\|\mathbf{x}_n-\mathbf{A}\bar{\mathbf{m}}_n-\mathbf{b}\|^2\right), \quad (4.10)$$

4.3.2 M-step

By taking the derivative of $\mathcal{Q}$ w.r.t. the parameters in $\boldsymbol{\Theta}$, obtain the following results.

Exercise 4.7: The optimal value for ν writes:

$$\nu^* = \frac{1}{d_{\mathsf{x}}}\left(\mathrm{Tr}(\mathbf{A}^\top\mathbf{A}\bar{\boldsymbol{\Omega}})+\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{x}_n-\mathbf{A}\bar{\mathbf{m}}_n-\mathbf{b}\|^2\right). \quad (4.11)$$

Regarding the other two parameters, nulling out their respective derivatives yields:

$$\mathbf{A} = (\mathbf{S}_{\mathsf{XM}}-\mathbf{b}\mathbf{S}_{\mathsf{M}}^\top)\mathbf{S}_2^{-1} \quad \text{and} \quad \mathbf{b} = \mathbf{S}_{\mathsf{X}}-\mathbf{A}\mathbf{S}_{\mathsf{M}}, \quad (4.12)$$

where $\mathbf{S}_{\mathsf{X}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \in \mathbb{R}^{d_{\mathsf{x}}}$, $\mathbf{S}_{\mathsf{M}} = \frac{1}{N}\sum_{n=1}^{N}\bar{\mathbf{m}}_n \in \mathbb{R}^{d_{\mathsf{z}}}$, $\mathbf{S}_{\mathsf{XM}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\bar{\mathbf{m}}_n^\top \in \mathbb{R}^{d_{\mathsf{x}}\times d_{\mathsf{z}}}$, $\mathbf{S}_2 = \frac{1}{N}\sum_{n=1}^{N}\bar{\mathbf{m}}_n\bar{\mathbf{m}}_n^\top+\bar{\Omega} \in \mathbb{R}^{d_{\mathsf{z}}\times d_{\mathsf{z}}}$. This means that the optimal values of $\mathbf{A}$ and $\mathbf{b}$ must be found by solving a linear system of equations, yielding:

$$\mathbf{A}^* = (\mathbf{S}_{\mathsf{XM}}-\mathbf{S}_{\mathsf{X}}\mathbf{S}_{\mathsf{M}}^\top)\mathbf{S}_2^{-1}(\mathbf{I}-\mathbf{S}_m\mathbf{S}_m^\top\mathbf{S}_2^{-1})^{-1}, \quad \text{and} \quad \mathbf{b}^* = \mathbf{S}_{\mathsf{X}}-\mathbf{A}^*\mathbf{S}_{\mathsf{M}}. \quad (4.13)$$

4.4 More on Multivariate Gaussian Distributions

We perform a deeper analysis on the multivariate Gaussians. Indeed, in Exercise 4.3 we have shown that for the PPCA model, the marginal distribution on $\mathbf{x}$ is also a multivariate Gaussian. Another immendiate consequence of this exercise is that the posterior distribution $p(\mathbf{z}|\mathbf{x})$ is also Gaussian.

Remark 4.3: Under the notations and assumptions of the PPCA model, the posterior distribution writes:

$$p(\mathbf{z}|\mathbf{x}) \overset{\mathbf{z}}{\propto} p(\mathbf{z},\mathbf{x}) \overset{\mathbf{z}}{\propto} \mathcal{N}(\mathbf{z};\mathbf{m},\boldsymbol{\Omega}). \quad (4.14)$$

While we have proven this for the PPCA model, the formulae can be easily extended to the more general case.

Remark 4.4: The so-called **linear Gaussian** model, which is actually afine, is defined as:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma}) \qquad p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x};\mathbf{A}\mathbf{z}+\mathbf{b},\mathbf{L}), \quad (4.15)$$

$\boldsymbol{\mu} \in \mathbb{R}^{d_{\mathsf{z}}}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d_{\mathsf{z}}\times d_{\mathsf{z}}}$, $\mathbf{A} \in \mathbb{R}^{d_{\mathsf{x}}\times d_{\mathsf{z}}}$, $\mathbf{b} \in \mathbb{R}^{d_{\mathsf{x}}}$ and $\mathbf{L} \in \mathbb{R}^{d_{\mathsf{x}}\times d_{\mathsf{x}}}$.

Exercise 4.8: What are the number of free parameters of this model?

Under these assumptions, $p(\mathbf{x},\mathbf{z})$, $p(\mathbf{x})$ and $p(\mathbf{z}|\mathbf{x})$ are all multivariate Gaussian distributions.

Exercise 4.9: Using the Gaussian completion, prove the following results. For the posterior distribution:

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mathbf{m}, \boldsymbol{\Omega}) \quad \text{with} \quad \boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}^{-1} + \mathbf{A}^\top \mathbf{L}^{-1}\mathbf{A} \quad \text{and} \quad \mathbf{m} = \boldsymbol{\Omega}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{A}^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{b})). \tag{4.16}$$

For the marginal distribution:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top + \mathbf{L}). \tag{4.17}$$

For the joint distribution:

$$p\left(\left(\begin{array}{c}\mathbf{x}\\ \mathbf{z}\end{array}\right)\right) = \mathcal{N}\left(\left(\begin{array}{c}\mathbf{x}\\ \mathbf{z}\end{array}\right); \left(\begin{array}{c}\mathbf{L}^{-1}\mathbf{b}\\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{A}^\top\mathbf{L}^{-1}\mathbf{b}\end{array}\right), \left(\begin{array}{cc}\mathbf{L}^{-1} & \mathbf{L}^{-1}\mathbf{A}\\ \mathbf{A}^\top\mathbf{L}^{-1} & \boldsymbol{\Sigma}^{-1} + \mathbf{A}^\top\mathbf{L}^{-1}\mathbf{A}\end{array}\right)\right) \tag{4.18}$$

Solutions to Exercises

Solution to Exercise 4.1: The model has $d_\mathsf{z} d_\mathsf{x}$ free parameters for $\mathbf{A}$, d_x free parameters for $\mathbf{b}$ and one for ν, so a total of $(d_\mathsf{z} + 1)d_\mathsf{x} + 1$.

Solution to Exercise 4.2: By direct computation:

$$\mathbb{E}\{\mathbf{x}\} = \mathbb{E}\{\mathbf{A}\mathbf{z} + \mathbf{b} + \boldsymbol{\epsilon}\} = \mathbf{b}, \tag{4.19}$$

since the two Gaussians are independent with null mean vector. Regarding the covariance matrix:

$$\mathbb{E}\{(\mathbf{x} - \mathbf{b})(\mathbf{x} - \mathbf{b})^\top\} = \mathbb{E}\{(\mathbf{A}\mathbf{z} + \boldsymbol{\epsilon})(\mathbf{A}\mathbf{z} + \boldsymbol{\epsilon})^\top\} = \mathbf{A}\mathbf{A}^\top + \nu\mathbf{I}, \tag{4.20}$$

where we have used the independence properties and the covariances of $\mathbf{z}$ and $\boldsymbol{\epsilon}$.

Solution to Exercise 4.3: In order to do that we write:

$$p(\mathbf{x}) = \int_{\mathcal{Z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}. \tag{4.21}$$

By writing the joint, we observe:

$$\begin{aligned}
p(\mathbf{x}, \mathbf{z}) &\overset{\mathbf{x}}{\propto} \exp\Big(-\frac{1}{2\nu}\|\mathbf{x} - \mathbf{A}\mathbf{z} - \mathbf{b}\|^2\Big)\exp\Big(-\frac{1}{2}\|\mathbf{z}\|^2\Big) && (4.22)\\
&\overset{\mathbf{x}}{\propto} \exp\left(-\frac{1}{2}\left[\frac{\|\mathbf{x} - \mathbf{b}\|^2}{\nu} + \frac{\|\mathbf{A}\mathbf{z}\|^2}{\nu} - \frac{2\mathbf{z}^\top\mathbf{A}^\top(\mathbf{x} - \mathbf{b})}{\nu} + \|\mathbf{z}\|^2\right]\right) && (4.23)\\
&\overset{\mathbf{x}}{\propto} \exp\Big(-\frac{1}{2\nu}\|\mathbf{x} - \mathbf{b}\|^2\Big)\exp\left(-\frac{1}{2}\left[\mathbf{z}^\top\left(\frac{\mathbf{A}^\top\mathbf{A}}{\nu} + \mathbf{I}\right)\mathbf{z} - 2\mathbf{z}^\top\frac{\mathbf{A}^\top(\mathbf{x} - \mathbf{b})}{\nu}\right]\right) && (4.24)
\end{aligned}$$

The Gaussian on $\mathbf{z}$ can be completed with:

$$\boldsymbol{\Omega} = \left(\frac{\mathbf{A}^\top\mathbf{A}}{\nu} + \mathbf{I}\right)^{-1} \qquad \mathbf{m} = \boldsymbol{\Omega}\left(\frac{\mathbf{A}^\top(\mathbf{x} - \mathbf{b})}{\nu}\right) \tag{4.25}$$

thus obtaining:

$$p(\mathbf{x}, \mathbf{z}) \overset{\mathbf{x}}{\propto} \exp\left(-\frac{1}{2}\left[\frac{\|\mathbf{x} - \mathbf{b}\|^2}{\nu} - \mathbf{m}^\top\boldsymbol{\Omega}^{-1}\mathbf{m}\right]\right)\mathcal{N}(\mathbf{z}; \mathbf{m}, \boldsymbol{\Omega}), \tag{4.26}$$

which concludes the proof (why?).

Solution to Exercise 4.4: Just multipy the right hand side expression by the left hand side expression (without the inverse) and simplify to obtain the identity.

Solution to Exercise 4.5: The key property is that expectations commute with linear operators such as matrix multiplication and the trace:

$$\mathbb{E}_{\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})}\left\{\mathbf{A}\mathbf{z}\right\} = \mathbf{A}\mathbb{E}_{\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})}\left\{\mathbf{z}\right\} = \mathbf{A}\boldsymbol{\mu} \tag{4.27}$$

and:

$$\begin{aligned}
\mathbb{E}_{\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})}\left\{\mathbf{z}^\top\boldsymbol{\Lambda}\mathbf{z}\right\} &= \mathbb{E}_{\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})}\left\{\mathrm{Tr}(\mathbf{z}^\top\boldsymbol{\Lambda}\mathbf{z})\right\} && (4.28)\\
&= \mathbb{E}_{\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})}\left\{\mathrm{Tr}(\boldsymbol{\Lambda}\mathbf{z}\mathbf{z}^\top)\right\} && (4.29)\\
&= \mathrm{Tr}(\boldsymbol{\Lambda}\mathbb{E}_{\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})}\left\{\mathbf{z}\mathbf{z}^\top\right\}) && (4.30)\\
&= \mathrm{Tr}(\boldsymbol{\Lambda}(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma})) && (4.31)\\
&= \boldsymbol{\mu}^\top\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathrm{Tr}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}). && (4.32)
\end{aligned}$$

Solution to Exercise 4.6: Use the formulae before, and get rid of all terms that do not depend on $\boldsymbol{\Theta}$.

Solution to Exercise 4.7: The optimal value for ν can be obtained by taking the derivative:

$$\frac{\partial\mathcal{Q}}{\partial\nu} = -\frac{1}{2}\left(\frac{Nd_{\mathsf{x}}}{\nu} - \frac{1}{\nu^2}\left(N\mathrm{Tr}(\mathbf{A}^\top\mathbf{A}\bar{\boldsymbol{\Omega}}) + \sum_{n=1}^{N}\|\mathbf{x}_n - \mathbf{A}\bar{\mathbf{m}}_n - \mathbf{b}\|^2\right)\right), \tag{4.33}$$

from what it follows that:

$$\frac{\partial\mathcal{Q}}{\partial\nu} = 0 \quad\Rightarrow\quad \nu^* = \frac{1}{d_{\mathsf{x}}}\left(\mathrm{Tr}(\mathbf{A}^\top\mathbf{A}\bar{\boldsymbol{\Omega}}) + \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{x}_n - \mathbf{A}\bar{\mathbf{m}}_n - \mathbf{b}\|^2\right). \tag{4.34}$$

Regarding the other two parameters, you need to use the derivatives w.r.t. matrices. Nulling out these derivatives yields:

$$\mathbf{A} = (\mathbf{S}_{\mathsf{XM}} - \mathbf{b}\mathbf{S}_{\mathsf{M}}^\top)\mathbf{S}_2^{-1} \quad \text{and} \quad \mathbf{b} = \mathbf{S}_{\mathsf{X}} - \mathbf{A}\mathbf{S}_{\mathsf{M}}, \tag{4.35}$$

where $\mathbf{S}_{\mathsf{X}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \in \mathbb{R}^{d_{\mathsf{x}}}$, $\mathbf{S}_{\mathsf{M}} = \frac{1}{N}\sum_{n=1}^{N}\bar{\mathbf{m}}_n \in \mathbb{R}^{d_{\mathsf{z}}}$, $\mathbf{S}_{\mathsf{XM}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\bar{\mathbf{m}}_n^\top \in \mathbb{R}^{d_{\mathsf{x}}\times d_{\mathsf{z}}}$, $\mathbf{S}_2 = \frac{1}{N}\sum_{n=1}^{N}\bar{\mathbf{m}}_n\bar{\mathbf{m}}_n^\top + \bar{\Omega} \in \mathbb{R}^{d_{\mathsf{z}}\times d_{\mathsf{z}}}$. This means that the optimal values of $\mathbf{A}$ and $\mathbf{b}$ must be found by solving a linear system of equations, yielding:

$$\mathbf{A}^* = (\mathbf{S}_{\mathsf{XM}} - \mathbf{S}_{\mathsf{X}}\mathbf{S}_{\mathsf{M}}^\top)\mathbf{S}_2^{-1}(\mathbf{I} - \mathbf{S}_m\mathbf{S}_m^\top\mathbf{S}_2^{-1})^{-1}, \quad \text{and} \quad \mathbf{b}^* = \mathbf{S}_{\mathsf{X}} - \mathbf{A}^*\mathbf{S}_{\mathsf{M}}. \tag{4.36}$$

Solution to Exercise 4.8: It's d_{z} for $\boldsymbol{\mu}$, d_{x} for $\mathbf{b}$, $d_{\mathsf{z}} \times d_{\mathsf{x}}$ for $\mathbf{A}$. For the covariance matrices we have $d_{\mathsf{z}}(d_{\mathsf{z}}+1)/2$ for $\boldsymbol{\Sigma}$ and $d_{\mathsf{x}}(d_{\mathsf{x}}+1)/2$ for $\mathbf{L}$, since they are symmetric.

Solution to Exercise 4.9: It's somewhat long to write, but it *only* needs to be written.

Chapter 5

Linear dynamical systems

Linear dynamical systems are composed of a hidden variable with a temporal structure, and an observation. They are very similar to hidden Markov models, with the very important difference that the hidden variable is continuous. **LDS are to PPCA what HMM are to GMM**.

	Discrete	Continuous
No Temporal	GMM $p(z = k) = \pi_k$	PPCA $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, I)$
Temporal	HMM $p(z_t = k \vert z_{t-1} = j) = \tau_{jk}$	LDS $p(\mathbf{z}_t \vert \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \mathbf{A}\mathbf{z}_{t-1}, \mathbf{\Gamma})$

In more detail the equations defining the model are:

$$p(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1; \mathbf{d}, \mathbf{\Omega}), \qquad p(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \mathbf{A}\mathbf{z}_{t-1}, \mathbf{\Gamma}), \qquad p(\mathbf{x}_t|\mathbf{z}_t) = \mathcal{N}(\mathbf{x}_t; \mathbf{C}\mathbf{z}_t, \mathbf{\Sigma}), \tag{5.1}$$

where $\mathbf{\Omega}$, $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ are covariance matrices.

Exercise 5.1: What is the dimension of the free parameters of the model?

5.1 Structured Multivariate Gaussians

Before going through the EM, we will be discussing properties of the multivariate Gaussian distribution. We **forget for a while about the sequences** and consider now the concatenation of $\mathbf{x}$ and $\mathbf{z}$: $\mathbf{y} = [\mathbf{x}^\top, \mathbf{z}^\top]^\top$, $\mathbf{y} \in \mathbb{R}^{d_x+d_z}$. We have seen that if $p(\mathbf{z})$ is a Gaussian, and $p(\mathbf{x}|\mathbf{z})$ is a linear-Gaussian, then the joint distribution on $\mathbf{y} = [\mathbf{x}^\top, \mathbf{z}^\top]^\top$ is a Gaussian as well. But is the opposite true?

If we write $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_y, \mathbf{\Sigma}_{yy})$ then we can write:

$$\boldsymbol{\mu}_y = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_z \end{pmatrix} \qquad \mathbf{\Sigma}_{yy} = \begin{pmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xz} \\ \mathbf{\Sigma}_{zx} & \mathbf{\Sigma}_{zz} \end{pmatrix} \tag{5.2}$$

We would like to derive properties of $p(\mathbf{z})$ and $p(\mathbf{x}|\mathbf{z})$ from $p(\mathbf{y})$. In order to evaluate $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_y, \mathbf{\Sigma}_{yy})$, we need to compute:

$$\mathbf{\Sigma}_{yy}^{-1} = \begin{pmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xz} \\ \mathbf{\Sigma}_{zx} & \mathbf{\Sigma}_{zz} \end{pmatrix}^{-1} \neq \begin{pmatrix} \mathbf{\Sigma}_{xx}^{-1} & \mathbf{\Sigma}_{xz}^{-1} \\ \mathbf{\Sigma}_{zx}^{-1} & \mathbf{\Sigma}_{zz}^{-1} \end{pmatrix} \tag{5.3}$$

You cannot invert a matrix block-by-block. Think about how do you invert a 2×2 matrix, or what if $\mathbf{\Sigma}_{xy}$ is not square. We have the help of the following results.

Remark 5.1: There are two version of the block-wise inversion lemma. The first:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix} \tag{5.4}$$

with $\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$.
And the second:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{N}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{C}\mathbf{N} \\ -\mathbf{N}\mathbf{B}\mathbf{A}^{-1} & \mathbf{N} \end{pmatrix} \tag{5.5}$$

with $\mathbf{N} = (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}$.

Exercise 5.2: Prove that both are indeed inverses of the original matrix. As a consequence they have to be equal. You will need the Woodbory lemma.

If we apply this to $p(\mathbf{y})$, the precision matrix $\mathbf{\Lambda}_{yy}$ (the inverse of $\mathbf{\Sigma}_{yy}$) is:

$$\mathbf{\Sigma}_{yy}^{-1} = \mathbf{\Lambda}_{yy} = \begin{pmatrix} \mathbf{\Lambda}_{xx} & \mathbf{\Lambda}_{xz} \\ \mathbf{\Lambda}_{zx} & \mathbf{\Lambda}_{zz} \end{pmatrix} = \begin{pmatrix} \mathbf{\Lambda}_{xx} & -\mathbf{\Lambda}_{xx}\mathbf{\Sigma}_{xz}\mathbf{\Sigma}_{zz}^{-1} \\ -\mathbf{\Sigma}_{zz}^{-1}\mathbf{\Sigma}_{zx}\mathbf{\Lambda}_{xx} & \mathbf{\Sigma}_{zz}^{-1} + \mathbf{\Sigma}_{zz}^{-1}\mathbf{\Sigma}_{zx}\mathbf{\Lambda}_{xx}\mathbf{\Sigma}_{xz}\mathbf{\Sigma}_{zz}^{-1} \end{pmatrix} \tag{5.6}$$

with $\mathbf{\Lambda}_{xx} = (\mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xz}\mathbf{\Sigma}_{zz}^{-1}\mathbf{\Sigma}_{zx})^{-1}$.

Exercise 5.3: Use the previous expression to prove that the conditional distribution writes:

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{x|z}, \mathbf{\Sigma}_{x|z}) \quad \text{with} \quad \mathbf{\Sigma}_{x|z} = \mathbf{\Lambda}_{xx}^{-1}, \quad \text{and} \quad \boldsymbol{\mu}_{x|z} = \boldsymbol{\mu}_x - \mathbf{\Lambda}_{xx}^{-1}\mathbf{\Lambda}_{xz}(\mathbf{z} - \boldsymbol{\mu}_z) \tag{5.7}$$

Exercise 5.4: Use the previous result to prove that the marginal distribution writes:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_z, \mathbf{\Sigma}_{zz}). \tag{5.8}$$

The **linear-Gaussian** model showed that two independent random variables with both marginal and conditional distributions being Gaussian, then the joint is Gaussian. We now showed that if the joint is Gaussian, then both the marginal and the conditional are Gaussians. This completes our discussion on multivariate Gaussian distributions that we will use for deriving the EM algorithm.

5.2 Expectation Maximization for LDS

As usual we have:

- Observations: $\mathbf{x}_{1:T} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$, with $\mathbf{x}_t \in \mathbb{R}^{d_x}$.
- Latent variables: $\mathbf{z}_{1:T} = (\mathbf{z}_1, \dots, \mathbf{z}_T)$, with $\mathbf{z}_t \in \mathbb{R}^{d_z}$.
- Model parameters: $\mathbf{\Theta} = \{\boldsymbol{\mu}_0, \mathbf{\Omega}, \mathbf{A}, \mathbf{\Gamma}, \mathbf{C}, \mathbf{\Sigma}\}$.

Given an estimate of the parameters $\bar{\mathbf{\Theta}}$, the expectation-maximisation algorithm has two steps:

- **Expectation**: compute the auxiliar function:

$$Q(\mathbf{\Theta}, \bar{\mathbf{\Theta}}) = \mathbb{E}_{p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T};\bar{\mathbf{\Theta}})}\{\log p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}; \mathbf{\Theta})\}. \tag{5.9}$$

- **Maximisation**: of the auxiliary function Q w.r.t. $\mathbf{\Theta}$:

$$\mathbf{\Theta}^* = \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}, \bar{\mathbf{\Theta}}), \tag{5.10}$$

  then set $\bar{\mathbf{\Theta}} = \mathbf{\Theta}^*$ and go back to the E-step.

5.2.1 E step

The E-step starts by deriving further the $\mathcal{Q}$ function. To this aim we first get the logarithm of the joint distribution (see Figure 5.1 as well):

$$\log\, p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}; \mathbf{\Theta}) = \log p(\mathbf{x}_{1:T}|\mathbf{z}_{1:T}; \mathbf{\Theta}) + \log p(\mathbf{z}_{1:T}; \mathbf{\Theta}) \tag{5.11}$$

$$= \sum_{t=1}^{T} \log p(\mathbf{x}_t|\mathbf{z}_t; \mathbf{\Theta}) + \sum_{t=2}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1}; \mathbf{\Theta}) + \log p(\mathbf{z}_1; \mathbf{\Theta}) \tag{5.12}$$

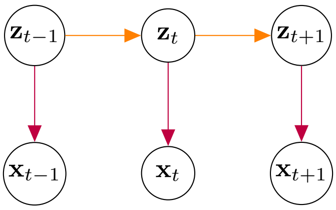


Figure 5.1: Graphical model of a first order hidden Markov chain (common to HMM and LDS models), with the transition and observation models highlighted in blue and green respectively.

If we develop $\mathcal{Q}$:

$$\begin{aligned}
\mathcal{Q} &= \mathbb{E}_{p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T};\bar{\boldsymbol{\Theta}})}\left\{\sum_{t=1}^{T}\log p(\mathbf{x}_t|\mathbf{z}_t;\boldsymbol{\Theta}) + \sum_{t=2}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1};\boldsymbol{\Theta}) + \log p(\mathbf{z}_1;\boldsymbol{\Theta})\right\} \qquad (5.13)\\
&= \sum_{t=1}^{T}\mathbb{E}_{p(\mathbf{z}_t|\mathbf{x}_{1:T};\bar{\boldsymbol{\Theta}})}\{\log p(\mathbf{x}_t|\mathbf{z}_t;\boldsymbol{\Theta})\} \qquad (5.14)\\
&+ \sum_{t=2}^{T}\mathbb{E}_{p(\mathbf{z}_{t-1},\mathbf{z}_t|\mathbf{x}_{1:T};\bar{\boldsymbol{\Theta}})}\{\log p(\mathbf{z}_t|\mathbf{z}_{t-1};\boldsymbol{\Theta})\} + \mathbb{E}_{p(\mathbf{z}_1|\mathbf{x}_{1:T};\bar{\boldsymbol{\Theta}})}\{\log p(\mathbf{z}_1;\boldsymbol{\Theta})\}. \qquad (5.15)
\end{aligned}$$

In order to compute $\mathcal{Q}$, we need to first compute:

$$p(\mathbf{z}_t|\mathbf{x}_{1:T};\bar{\boldsymbol{\Theta}}) \quad \text{and} \quad p(\mathbf{z}_{t-1},\mathbf{z}_t|\mathbf{x}_{1:T};\bar{\boldsymbol{\Theta}}), \quad \forall t. \tag{5.16}$$

If we start from scratch, we write:

$$\begin{aligned}
p(\mathbf{z}_t|\mathbf{x}_{1:T}) &= \int\dots\int p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\mathrm{d}\mathbf{z}_{1:t-1}\mathrm{d}\mathbf{z}_{t+1:T} \qquad (5.17)\\
&= \frac{1}{p(\mathbf{x}_{1:T})}\int\dots\int p(\mathbf{x}_{1:T},\mathbf{z}_{1:T})\mathrm{d}\mathbf{z}_{1:t-1}\mathrm{d}\mathbf{z}_{t+1:T} \qquad (5.18)\\
&= \frac{1}{p(\mathbf{x}_{1:T})}\int\dots\int \prod_{\tau=1}^{T} p(\mathbf{x}_\tau|\mathbf{z}_\tau)\prod_{\tau=2}^{T} p(\mathbf{z}_\tau|\mathbf{z}_{\tau-1})p(\mathbf{z}_1)\mathrm{d}\mathbf{z}_{1:t-1}\mathrm{d}\mathbf{z}_{t+1:T} \qquad (5.19)
\end{aligned}$$

This requires $T-1$ integrals for every $\mathbf{z}_t$, so $T(T-1)$ integrals. We need a **smarter strategy**!
As in the case of HMM, we will use the **forward-backward** algorithm.

Main idea: decompose as a product of a forward and backward *distributions* that can be computed recursively and re-used at every t:

$$p(\mathbf{z}_t|\mathbf{x}_{1:T}) \overset{(\mathbf{z}_t)}{\propto} p(\mathbf{z}_t,\mathbf{x}_{1:T}) = p(\mathbf{z}_t,\mathbf{x}_{1:t})p(\mathbf{x}_{t+1:T}|\mathbf{z}_t) \tag{5.20}$$

Exercise 5.5: Prove that the past dependencies drop from the backward term, or equivalently that $\mathbf{x}_{1:t}$ and $\mathbf{x}_{t+1:T}$ are D-separated by $\mathbf{z}_t$.

We refer to $p(\mathbf{z}_t,\mathbf{x}_{1:t})$ and $p(\mathbf{x}_{t+1:T}|\mathbf{z}_t)$ as the forward and backward *distributions*, but this is a language abuse. They are denoted respectively by $\alpha_t(\mathbf{z}_t) = p(\mathbf{z}_t,\mathbf{x}_{1:t})$ and $\beta_t(\mathbf{z}_t) = p(\mathbf{x}_{t+1:T}|\mathbf{z}_t)$. We recall that $p(\mathbf{z}_t|\mathbf{x}_{1:T}) \overset{(\mathbf{z}_{1:T})}{\propto} p(\mathbf{z}_t,\mathbf{x}_{1:t})p(\mathbf{x}_{t+1:T}|\mathbf{z}_t)$, but also:

$$p(\mathbf{z}_t|\mathbf{x}_{1:T}) = \frac{1}{p(\mathbf{x}_{1:T})}\int\dots\int\prod_{\tau=1}^{t}p(\mathbf{x}_\tau|\mathbf{z}_\tau)\prod_{\tau=2}^{t}p(\mathbf{z}_\tau|\mathbf{z}_{\tau-1})p(\mathbf{z}_1)\mathrm{d}\mathbf{z}_{1:t-1}\int\dots\int\prod_{\tau=t+1}^{T}p(\mathbf{x}_\tau|\mathbf{z}_\tau)\prod_{\tau=t+1}^{T}p(\mathbf{z}_\tau|\mathbf{z}_{\tau-1})\mathrm{d}\mathbf{z}_{t+1:T}. \tag{5.21}$$

Exercise 5.6: Prove the general forward-backward recursions:

$$\alpha_t(\mathbf{z}_t) \triangleq p(\mathbf{z}_t,\mathbf{x}_{1:t}) = p(\mathbf{x}_t|\mathbf{z}_t)\int p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{z}_{t-1},\mathbf{x}_{1:t-1})\mathrm{d}\mathbf{z}_{t-1} = p(\mathbf{x}_t|\mathbf{z}_t)\int p(\mathbf{z}_t|\mathbf{z}_{t-1})\alpha_{t-1}(\mathbf{z}_{t-1})\mathrm{d}\mathbf{z}_{t-1}, \tag{5.22}$$

$$\beta_t(\mathbf{z}_t) \triangleq p(\mathbf{x}_{t+1:T}|\mathbf{z}_t) = \int p(\mathbf{x}_{t+1}|\mathbf{z}_{t+1})p(\mathbf{z}_{t+1}|\mathbf{z}_t)p(\mathbf{x}_{t+2:T}|\mathbf{z}_{t+1})\mathrm{d}\mathbf{z}_{t+1} = \int p(\mathbf{x}_{t+1}|\mathbf{z}_{t+1})p(\mathbf{z}_{t+1}|\mathbf{z}_t)\beta_{t+1}(\mathbf{z}_{t+1})\mathrm{d}\mathbf{z}_{t+1}. \tag{5.23}$$

So we have:

- **Forward** Start with $\alpha_1(\mathbf{z}_1) = p(\mathbf{z}_1|\mathbf{x}_1)$ and compute:

$$\alpha_t(\mathbf{z}_t) = p(\mathbf{x}_t|\mathbf{z}_t)\int p(\mathbf{z}_t|\mathbf{z}_{t-1})\alpha_{t-1}(\mathbf{z}_{t-1})\mathrm{d}\mathbf{z}_{t-1}, \quad \forall t. \tag{5.24}$$

- **Backward** Start with $\beta_{T-1}(\mathbf{z}_{T-1}) = p(\mathbf{x}_T|\mathbf{z}_{T-1})$ and compute:

$$\beta_t(\mathbf{z}_t) = \int p(\mathbf{x}_{t+1}|\mathbf{z}_{t+1})p(\mathbf{z}_{t+1}|\mathbf{z}_t)\beta_{t+1}(\mathbf{z}_{t+1})\mathrm{d}\mathbf{z}_{t+1}, \forall t. \tag{5.25}$$

- **Posterior** Compute $p(\mathbf{z}_t|\mathbf{x}_{1:T}) \propto \alpha_t(\mathbf{z}_t)\beta_t(\mathbf{z}_t)$, $\forall t$.

This is the exact same formulation as for HMM (we did not use the distributions of $\mathbf{z}_t$ and $\mathbf{x}_t$), so this is general. With this strategy, we need $2(T-1)$ integrals, instead of $T(T-1)$. Differently from HMM, we need to see how these recursions look like for LDS.

Forward & backward recursions In order to compute the forward recursion we assume that the forward distribution at the previous time step is Gaussian (see below for initialisation).

Exercise 5.7: Let us now assume that: $\alpha_{t-1}(\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\mu}_{t-1}, \mathbf{V}_{t-1})$, prove that the recursion in 5.24 leads to a Gaussian as follows:

$$\alpha_t(\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_t, \mathbf{V}_t), \quad \text{with} \quad \begin{cases} \mathbf{V}_t = \left(\mathbf{C}^\top\boldsymbol{\Sigma}^{-1}\mathbf{C} + (\boldsymbol{\Gamma} + \mathbf{A}\mathbf{V}_{t-1}\mathbf{A}^\top)^{-1}\right)^{-1}, \\ \boldsymbol{\mu}_t = \mathbf{V}_t\left((\boldsymbol{\Gamma} + \mathbf{A}\mathbf{V}_{t-1}\mathbf{A}^\top)^{-1}\mathbf{A}\boldsymbol{\mu}_{t-1} + \mathbf{C}\boldsymbol{\Sigma}^{-1}\mathbf{x}_t\right). \end{cases} \tag{5.26}$$

Exercise 5.8: Using similar tools than in the forward pass, and when necessary completing the quadratic form, prove that if $\beta_{t+1}(\mathbf{z}_{t+1}) = \mathcal{N}(\mathbf{z}_{t+1}; \boldsymbol{\nu}_{t+1}, \mathbf{W}_{t+1})$, then:

$$\beta_t(\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_t; \boldsymbol{\nu}_t, \mathbf{W}_t) \quad \text{with} \quad \mathbf{W}_t^{-1} = \mathbf{A}^\top(\boldsymbol{\Gamma} + (\mathbf{W}_{t+1}^{-1} + \mathbf{C}^\top\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1})^{-1}\mathbf{A} \tag{5.27}$$

and

$$\boldsymbol{\nu}_t = \mathbf{W}_t\mathbf{A}^\top\boldsymbol{\Gamma}^{-1}(\mathbf{C}^\top\boldsymbol{\Sigma}^{-1}\mathbf{C} + \mathbf{W}_{t+1}^{-1} + \boldsymbol{\Gamma}^{-1})^{-1}(\mathbf{C}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}_{t+1} + \mathbf{W}_{t+1}^{-1}\boldsymbol{\nu}_{t+1}) \tag{5.28}$$

Marginal a posteriori All the previous computations were targetted to obtain the marginal a posteriori distribution:

$$p(\mathbf{z}_t|\mathbf{x}_{1:T}) \overset{(\mathbf{z}_t)}{\propto} p(\mathbf{z}_t, \mathbf{x}_{1:t})p(\mathbf{x}_{t+1:T}|\mathbf{z}_t) \overset{(\mathbf{z}_t)}{\propto} \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_t, \mathbf{V}_t)\mathcal{N}(\mathbf{z}_t; \boldsymbol{\nu}_t, \mathbf{W}_t) \tag{5.29}$$

$$\overset{(\mathbf{z}_t)}{\propto} \exp\left(-\frac{1}{2}\left[\mathbf{z}_t^\top(\mathbf{V}_t^{-1} + \mathbf{W}_t^{-1})\mathbf{z}_t - 2\mathbf{z}_t^\top(\mathbf{V}_t^{-1}\boldsymbol{\mu}_t + \mathbf{W}_t^{-1}\boldsymbol{\nu}_t)\right]\right) \tag{5.30}$$

$$\overset{(\mathbf{z}_t)}{\propto} \mathcal{N}(\mathbf{z}_t; \mathbf{m}_t, \boldsymbol{\Lambda}_t) \tag{5.31}$$

$$\text{with} \qquad \boldsymbol{\Lambda}_t = (\mathbf{V}_t^{-1} + \mathbf{W}_t^{-1})^{-1} \qquad \mathbf{m}_t = \boldsymbol{\Lambda}_t(\mathbf{V}_t^{-1}\boldsymbol{\mu}_t + \mathbf{W}_t^{-1}\boldsymbol{\nu}_t) \tag{5.32}$$

But how to initialise the recursions?

Initialisation of the recursions If we take a look to the forward at time $t = 1$:

$$\alpha_1(\mathbf{z}_1) = p(\mathbf{z}_1, \mathbf{x}_1) = p(\mathbf{x}_1|\mathbf{z}_1)p(\mathbf{z}_1) = \mathcal{N}(\mathbf{x}_1; \mathbf{C}\mathbf{z}_1, \boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}_1; \mathbf{d}, \boldsymbol{\Omega}). \tag{5.33}$$

Therefore:

$$\alpha_1(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1; \boldsymbol{\mu}_1, \mathbf{V}_1) \text{ with } \begin{cases} \mathbf{V}_1 = (\boldsymbol{\Omega}^{-1} + \mathbf{C}\boldsymbol{\Sigma}^{-1}\mathbf{C}^\top)^{-1}, \\ \boldsymbol{\mu}_1 = \mathbf{V}_1(\boldsymbol{\Omega}^{-1}\mathbf{d} + \mathbf{C}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}_1) \end{cases} \tag{5.34}$$

Regading the backward, let's first reason with the following equation:

$$p(\mathbf{z}_T|\mathbf{x}_{1:T}) \overset{(\mathbf{z}_T)}{\propto} \underbrace{p(\mathbf{z}_T, \mathbf{x}_{1:T})}_{\alpha_T(\mathbf{z}_T)}, \tag{5.35}$$

meaning that actually $\beta_T(\mathbf{z}_T) \overset{(\mathbf{z}_T)}{\propto} 1$, and thus β_T should be constant. This would be achieved if we initialise β_T as a normal with infinite covariance or zero precision $\mathbf{W}_T^{-1} \to 0$, but in that case it will not propertly a normal. However, we can compute the parameters of β_{T-1} that correspond to this choice:

$$\mathbf{W}_T^{-1} \to 0 \Rightarrow \begin{cases} \mathbf{W}_{T-1}^{-1} = \mathbf{A}^\top(\boldsymbol{\Gamma} + (\mathbf{C}^\top\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1})^{-1}\mathbf{A} \\ \boldsymbol{\nu}_{T-1} = \mathbf{W}_{T-1}\mathbf{A}^\top\boldsymbol{\Gamma}^{-1}(\mathbf{C}^\top\boldsymbol{\Sigma}^{-1}\mathbf{C} + \boldsymbol{\Gamma}^{-1})^{-1}(\mathbf{C}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}_{t+1}). \end{cases}$$

This is equivalent to start from $\beta_{T-1}(\mathbf{z}_{T-1}) = \int p(\mathbf{x}_T, \mathbf{z}_T|\mathbf{z}_{T-1})\mathrm{d}\mathbf{z}_{T-1}$. We can now compute $p(\mathbf{z}_t|\mathbf{x}_{1:T}; \bar{\boldsymbol{\Theta}})$ (previous formulae):

- Initialise and compute the parameters of the forward distribution.
- Initialise and compute the parameters of the backward distribution.
- Compute the parameters of the posterior.

To finish the E-step, we need to compute the joint distribution $p(\mathbf{z}_t, \mathbf{z}_{t+1}|\mathbf{x}_{1:T})$, and compute the $\mathcal{Q}$ function.

Joint a posteriori In order to compute the joint a posteriori distribution of two consecutive hidden variables, we start by marginalising, as we did for the posterior of a single latent variable:

$$p(\mathbf{z}_t, \mathbf{z}_{t+1}|\mathbf{x}_{1:T}) = \int \dots \int p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\mathrm{d}\mathbf{z}_{1:t-1}\mathrm{d}\mathbf{z}_{t+2:T} \tag{5.36}$$

$$\overset{(\mathbf{z}_t, \mathbf{z}_{t+1})}{\propto} \alpha_t(\mathbf{z}_t)p(\mathbf{x}_{t+1}|\mathbf{z}_{t+1})p(\mathbf{z}_{t+1}|\mathbf{z}_t)\beta_{t+1}(\mathbf{z}_{t+1}) \tag{5.37}$$

$$\overset{(\mathbf{z}_t, \mathbf{z}_{t+1})}{\propto} \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_t, \mathbf{V}_t)\mathcal{N}(\mathbf{x}_{t+1}; \mathbf{C}\mathbf{z}_{t+1}, \boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}_{t+1}; \mathbf{A}\mathbf{z}_t, \boldsymbol{\Gamma})\mathcal{N}(\mathbf{z}_{t+1}; \boldsymbol{\nu}_{t+1}, \mathbf{W}_{t+1}) \tag{5.38}$$

$$\begin{aligned} \overset{(\mathbf{z}_t, \mathbf{z}_{t+1})}{\propto} \exp\Bigg(-\frac{1}{2}\Big[&\mathbf{z}_t^\top(\mathbf{V}_t^{-1} + \mathbf{A}^\top\boldsymbol{\Gamma}^{-1}\mathbf{A})\mathbf{z}_t + \mathbf{z}_{t+1}^\top(\mathbf{C}^\top\boldsymbol{\Sigma}^{-1}\mathbf{C} + \boldsymbol{\Gamma}^{-1} + \mathbf{W}_{t+1}^{-1})\mathbf{z}_{t+1} \\ &- 2\mathbf{z}_{t+1}^\top\boldsymbol{\Gamma}^{-1}\mathbf{A}\mathbf{z}_t - 2\mathbf{z}_t^\top\mathbf{V}_t^{-1}\boldsymbol{\mu}_t - 2\mathbf{z}_{t+1}^\top(\mathbf{C}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}_{t+1} + \mathbf{W}_{t+1}\boldsymbol{\nu}_{t+1})\Big]\Bigg). \end{aligned} \tag{5.39}$$

This means that we will re-use the forward and backward distribution. Another consequence of the previous equation is that the joint posterior distribution is also Gaussian.

Exercise 5.9: From the expression above, prove that $p(\mathbf{z}_t, \mathbf{z}_{t+1}|\mathbf{x}_{1:T}) = \mathcal{N}([\mathbf{z}_t; \mathbf{z}_{t+1}]; \mathbf{r}_t, \mathbf{\Xi}_t)$ with:

$$\mathbf{\Xi}_t = \begin{pmatrix} \mathbf{V}_t{}^{-1} + \mathbf{A}^\top\mathbf{\Gamma}^{-1}\mathbf{A} & -\mathbf{A}^\top\mathbf{\Gamma}^{-1} \\ -\mathbf{\Gamma}^{-1}\mathbf{A} & \mathbf{W}_{t+1}{}^{-1} + \mathbf{\Gamma}^{-1} + \mathbf{C}^\top\mathbf{\Sigma}^{-1}\mathbf{C} \end{pmatrix}^{-1} \quad \text{and} \quad \mathbf{r}_t = \mathbf{\Xi}_t \begin{pmatrix} \mathbf{V}_t{}^{-1}\boldsymbol{\mu}_t \\ \mathbf{C}^\top\mathbf{\Sigma}^{-1}\mathbf{x}_{t+1} + \mathbf{W}_{t+1}{}^{-1}\boldsymbol{\nu}_{t+1} \end{pmatrix}. \tag{5.40}$$

In the previous equations we droped the bar from $\bar{\mathbf{\Theta}}$ for simplicity. For the E-step ALWAYS use the parameters of the previous iteration! Thus, all equations above should use $\bar{\mathbf{A}}$, $\bar{\mathbf{\Gamma}}$, $\bar{\mathbf{C}}$, $\bar{\mathbf{\Sigma}}$, $\bar{\mathbf{d}}$ and $\bar{\mathbf{\Omega}}$.

5.2.2 M-step

In order to derive the M-step we need some tools. We recall this important result.

$$\mathbb{E}_{\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\mathbf{\Sigma})}\{\mathbf{A}\mathbf{z}\} = \mathbf{A}\boldsymbol{\mu} \quad \text{and} \quad \mathbb{E}_{\mathcal{N}(\mathbf{z};\boldsymbol{\mu},\mathbf{\Sigma})}\left\{\mathbf{z}^\top\mathbf{\Lambda}\mathbf{z}\right\} = \boldsymbol{\mu}^\top\mathbf{\Lambda}\boldsymbol{\mu} + \mathrm{Tr}(\mathbf{\Lambda}\mathbf{\Sigma}).$$

and certain properties:

$$(i)\ \mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{BCA}) \qquad (ii)\ \frac{\partial}{\partial\mathbf{A}}\mathrm{Tr}(\mathbf{A}^\top\mathbf{B}) = \mathbf{B} \tag{5.41}$$

$$(iii)\ \frac{\partial}{\partial\mathbf{A}}\mathrm{Tr}(\mathbf{ABA}^\top\mathbf{C}) = \mathbf{CAB} + \mathbf{C}^\top\mathbf{AB}^\top \qquad (iv)\ \frac{\partial}{\partial\mathbf{A}}\log|\mathbf{A}| = (\mathbf{A}^{-1})^\top \tag{5.42}$$

Let's recall that the posterior distributions:

$$p(\mathbf{z}_t|\mathbf{x}_{1:T}) = \mathcal{N}(\mathbf{z}_t; \mathbf{m}_t, \mathbf{\Lambda}_t) \quad \text{and} \quad p(\mathbf{z}_t, \mathbf{z}_{t+1}|\mathbf{x}_{1:T}) = \mathcal{N}([\mathbf{z}_t; \mathbf{z}_{t+1}]; \boldsymbol{\xi}_t, \mathbf{\Xi}_t) \tag{5.43}$$

($\mathbf{m}_t \in \mathbb{R}^{d_z}$, $\mathbf{\Lambda}_t \in \mathbb{R}^{d_z \times d_z}$ and $\boldsymbol{\xi}_t \in \mathbb{R}^{2d_z}$, $\mathbf{\Xi}_t \in \mathbb{R}^{2d_z \times 2d_z}$)

M-step: $\mathbf{C}$ and $\mathbf{\Sigma}$ The expectation of the t-th term writes:

$$\mathbb{E}_{p(\mathbf{z}_t|\mathbf{x}_{1:T};\bar{\mathbf{\Theta}})}\{\log p(\mathbf{x}_t|\mathbf{z}_t;\mathbf{\Theta})\} \overset{(\mathbf{\Theta})}{=} -\frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\left(\mathbf{x}_t\mathbf{x}_t^\top - 2\mathbf{C}\mathbf{m}_t\mathbf{x}_t^\top + \mathbf{C}(\mathbf{m}_t\mathbf{m}_t^\top + \mathbf{\Lambda}_t)\mathbf{C}^\top\right)\right\} \tag{5.44}$$

Sum over t:

$$\mathcal{Q}_{\mathbf{C},\mathbf{\Sigma}} = -\frac{T}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\sum_{t=1}^{T}\left(\mathbf{x}_t\mathbf{x}_t^\top - 2\mathbf{C}\mathbf{m}_t\mathbf{x}_t^\top + \mathbf{C}(\mathbf{m}_t\mathbf{m}_t^\top + \mathbf{\Lambda}_t)\mathbf{C}^\top\right)\right\} \tag{5.45}$$

By definining: $\mathbf{S}_{xx} = \sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t^\top$ $\quad \mathbf{S}_{zx} = \sum_{t=1}^T \mathbf{m}\mathbf{x}^\top$ $\quad \mathbf{S}_{zz} = \sum_{t=1}^T \mathbf{m}_t\mathbf{m}_t^\top + \mathbf{\Lambda}_t$:

$$\mathcal{Q}_{\mathbf{C},\mathbf{\Sigma}} = -\frac{T}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\left(\mathbf{S}_{xx} - 2\mathbf{C}\mathbf{S}_{zx} + \mathbf{C}\mathbf{S}_{zz}\mathbf{C}^\top\right)\right\} \tag{5.46}$$

By taking derivatives and nulling them out we obtain:

$$\frac{\partial\mathcal{Q}}{\partial\mathbf{C}} = 0 \Rightarrow \mathbf{C}^* = \mathbf{S}_{xz}(\mathbf{S}_{zz})^{-1} \quad \text{and} \quad \frac{\partial\mathcal{Q}}{\partial\mathbf{\Sigma}^{-1}} = 0 \Rightarrow \mathbf{\Sigma}^* = \frac{1}{T}\left(\mathbf{S}_{xx} - \mathbf{S}_{xz}(\mathbf{S}_{zz})^{-1}\mathbf{S}_{zx}\right). \tag{5.47}$$

M-step: $\mathbf{A}$ and $\mathbf{\Gamma}$ The t-th term of the sum writes:

$$-\frac{1}{2}\log|\mathbf{\Gamma}| - \frac{1}{2}\mathrm{Tr}\left\{\mathbb{E}_{p(\mathbf{z}_{t+1},\mathbf{z}_t|\mathbf{x}_{1:T};\bar{\mathbf{\Theta}})}\left\{\mathbf{\Gamma}^{-1}\left(\mathbf{z}_{t+1}\mathbf{z}_{t+1}^\top - 2\mathbf{A}\mathbf{z}_t\mathbf{z}_{t+1}^\top + \mathbf{A}\mathbf{z}_t\mathbf{z}_t^\top\mathbf{A}^\top\right)\right\}\right\} \tag{5.48}$$

Remember $p(\mathbf{z}_t, \mathbf{z}_{t+1}|\mathbf{x}_{1:T}) = \mathcal{N}([\mathbf{z}_t; \mathbf{z}_{t+1}]; \boldsymbol{\xi}_t, \mathbf{\Xi}_t)$:

$$\boldsymbol{\xi}_t =: \begin{pmatrix} \boldsymbol{\xi}_{t-} \\ \boldsymbol{\xi}_{t+} \end{pmatrix} \qquad \mathbf{\Xi}_t =: \begin{pmatrix} \mathbf{\Xi}_{t--} & \mathbf{\Xi}_{t-+} \\ \mathbf{\Xi}_{t+-} & \mathbf{\Xi}_{t++} \end{pmatrix}, \tag{5.49}$$

We need to take the expectation, and the sum over t:

$$\mathbf{S}_{++} = \sum_{t=1}^{T}\boldsymbol{\xi}_{t+}\boldsymbol{\xi}_{t+}^\top + \mathbf{\Xi}_{t++}, \quad \mathbf{S}_{--} = \sum_{t=1}^{T}\boldsymbol{\xi}_{t-}\boldsymbol{\xi}_{t-}^\top + \mathbf{\Xi}_{t--}, \quad \mathbf{S}_{-+} = \sum_{t=1}^{T}\boldsymbol{\xi}_{t-}\boldsymbol{\xi}_{t+}^\top + \mathbf{\Xi}_{t-+}, \tag{5.50}$$

As a result (very similar formula to the previous one):

$$\mathcal{Q}_{\mathbf{A},\mathbf{\Gamma}} = -\frac{T-1}{2}\log|\mathbf{\Gamma}| - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{\Gamma}^{-1}\left(\mathbf{S}_{++} - 2\mathbf{A}\mathbf{S}_{-+} + \mathbf{A}\mathbf{S}_{--}\mathbf{A}^\top\right)\right\} \tag{5.51}$$

The optimal values can be derived as in the previous case.

M-step: $\mathbf{d}$ and $\mathbf{\Omega}$ It is quite easy to see that:

$$\mathbf{d}^* = \mathbf{m}_1 \qquad \mathbf{\Omega}^* = \mathbf{\Lambda}_1. \tag{5.52}$$

5.2.3 Summary

Start with initial parameters $\bar{\mathbf{\Theta}}$. Iterate the following:

E-step Initialise and compute the forward recursion: $\boldsymbol{\mu}_t, \mathbf{V}_t, \forall t$.

Initialise and compute the backward recursion: $\boldsymbol{\nu}_t, \mathbf{W}_t, \forall t$.

Compute the parameters of the posterior, $\mathbf{m}_t, \mathbf{\Lambda}_t$, and joint posterior $\boldsymbol{\xi}_t, \mathbf{\Xi}_t, \forall t$.

M-step Compute the summary statistics $\mathbf{S}_{xx}$, $\mathbf{S}_{zx}$, and $\mathbf{S}_{zz}$, and the optimal values $\mathbf{C}^*$ and $\mathbf{\Sigma}^*$.

Compute the summary statistics $\mathbf{S}_{++}$, $\mathbf{S}_{-+}$, and $\mathbf{S}_{--}$ and the optimal values $\mathbf{A}^*$ and $\mathbf{\Gamma}^*$.

Compute the optimal initial values $\mathbf{d}^*$ and $\mathbf{\Omega}^*$.

Until some convergence criterion is met.

Solutions to (some) Exercises

Solution to Exercise 5.3: To do so, we will consider $\mathbf{z}$ fixed in the joint distribution, and complete the Gaussian on $\mathbf{x}$. Developing the joint distribution with the blocks of the precision matrix of $\mathbf{y}$:

$$\log \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_y, \mathbf{\Sigma}_{yy}) \overset{\mathbf{y}}{=} -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_y)^\top \mathbf{\Sigma}_{yy}^{-1}(\mathbf{y}-\boldsymbol{\mu}_y) \tag{5.53}$$

$$\overset{(\mathbf{x},\mathbf{z})}{=} -\frac{1}{2}\Big((\mathbf{x}-\boldsymbol{\mu}_x)^\top \mathbf{\Lambda}_{xx}(\mathbf{x}-\boldsymbol{\mu}_x) + 2(\mathbf{x}-\boldsymbol{\mu}_x)^\top \mathbf{\Lambda}_{xz}(\mathbf{z}-\boldsymbol{\mu}_z) \tag{5.54}$$

$$+ (\mathbf{z}-\boldsymbol{\mu}_z)^\top \mathbf{\Lambda}_{zz}(\mathbf{z}-\boldsymbol{\mu}_z)\Big). \tag{5.55}$$

Now considering a fixed value of $\mathbf{z}$ we write:

$$\log p(\mathbf{x}|\mathbf{z}) \overset{(\mathbf{x})}{=} -\frac{1}{2}\Big(\mathbf{x}^\top \underbrace{\mathbf{\Lambda}_{xx}}_{\mathbf{\Sigma}_{x|z}^{-1}} \mathbf{x} - 2\mathbf{x}^\top \mathbf{\Sigma}_{x|z}^{-1} \underbrace{\mathbf{\Sigma}_{x|z}\Big(\mathbf{\Lambda}_{xx}\boldsymbol{\mu}_x - \mathbf{\Lambda}_{xz}(\mathbf{z}-\boldsymbol{\mu}_z)\Big)}_{\boldsymbol{\mu}_{x|z}}\Big). \tag{5.56}$$

Therefore, $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{x|z}, \mathbf{\Sigma}_{x|z})$ with:

$$\mathbf{\Sigma}_{x|z} = \mathbf{\Lambda}_{xx}^{-1}, \qquad \boldsymbol{\mu}_{x|z} = \mathbf{\Sigma}_{x|z}\Big(\mathbf{\Lambda}_{xx}\boldsymbol{\mu}_x - \mathbf{\Lambda}_{xz}(\mathbf{z}-\boldsymbol{\mu}_z)\Big) \tag{5.57}$$

$$= \boldsymbol{\mu}_x - \mathbf{\Lambda}_{xx}^{-1}\mathbf{\Lambda}_{xz}(\mathbf{z}-\boldsymbol{\mu}_z) \tag{5.58}$$

Solution to Exercise 5.4: We will start by considering the Bayes theorem in logarithm form:

$$\log p(\mathbf{z}) = \log p(\mathbf{x}, \mathbf{z}) - \log p(\mathbf{x}|\mathbf{z}) \tag{5.59}$$

$$\overset{(\mathbf{x},\mathbf{z})}{=} -\frac{1}{2}\Big((\mathbf{x}-\boldsymbol{\mu}_x)^\top \mathbf{\Lambda}_{xx}(\mathbf{x}-\boldsymbol{\mu}_x) + 2(\mathbf{x}-\boldsymbol{\mu}_x)^\top \mathbf{\Lambda}_{xz}(\mathbf{z}-\boldsymbol{\mu}_z) \tag{5.60}$$

$$(\mathbf{z}-\boldsymbol{\mu}_z)^\top \mathbf{\Lambda}_{zz}(\mathbf{z}-\boldsymbol{\mu}_z)\Big) + \frac{1}{2}\Big((\mathbf{x}-\boldsymbol{\mu}_{x|z})^\top \mathbf{\Sigma}_{x|z}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{x|z})\Big) \tag{5.61}$$

$$\overset{(\mathbf{x},\mathbf{z})}{=} -\frac{1}{2}\Big((\mathbf{x}-\boldsymbol{\mu}_x)^\top \mathbf{\Lambda}_{xx}(\mathbf{x}-\boldsymbol{\mu}_x) + 2(\mathbf{x}-\boldsymbol{\mu}_x)^\top \mathbf{\Lambda}_{xz}(\mathbf{z}-\boldsymbol{\mu}_z) \tag{5.62}$$

$$+ (\mathbf{z}-\boldsymbol{\mu}_z)^\top \mathbf{\Lambda}_{zz}(\mathbf{z}-\boldsymbol{\mu}_z)\Big) + \frac{1}{2}\Big((\mathbf{x}-\boldsymbol{\mu}_x)^\top \mathbf{\Lambda}_{xx}(\mathbf{x}-\boldsymbol{\mu}_x) \tag{5.63}$$

$$+ 2(\mathbf{x}-\boldsymbol{\mu}_x)^\top \mathbf{\Lambda}_{xx}\mathbf{\Lambda}_{xx}^{-1}\mathbf{\Lambda}_{xz}(\mathbf{z}-\boldsymbol{\mu}_z) + (\mathbf{z}-\boldsymbol{\mu}_z)^\top \mathbf{\Lambda}_{xz}^\top \mathbf{\Lambda}_{xx}^{-1}\mathbf{\Lambda}_{xz}(\mathbf{z}-\boldsymbol{\mu}_z)\Big) \tag{5.64}$$

$$\overset{\mathbf{z}}{=} -\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_z)^\top \Big(\mathbf{\Lambda}_{zz} - \mathbf{\Lambda}_{xz}^\top \mathbf{\Lambda}_{xx}^{-1}\mathbf{\Lambda}_{xz}\Big)(\mathbf{z}-\boldsymbol{\mu}_z), \tag{5.65}$$

which leads to the desired result after applying the Woodbury lemma to see that $\mathbf{\Sigma}_{zz} = \Big(\mathbf{\Lambda}_{zz} - \mathbf{\Lambda}_{xz}^\top \mathbf{\Lambda}_{xx}^{-1}\mathbf{\Lambda}_{xz}\Big)^{-1}$.

Bibliography

[1] K. B. Petersen and M. S. Pedersen. The Matrix Cookbook, Nov. 2012.