5MB20 - Exercises Part 1: Linear Gaussian Models and EM

course year 2010

Note: For some of these exercises you will be able to find solutions quickly on the internet. Try to resist this route to solving the problems. You will not be graded for these exercises and solutions will be made available through the class web page. **Your ability to solve these exercises without external help (apart from Sam Roweis' cheat sheets) provides an excellent indicator of your readiness to pass the exam**. Finally, the provided solutions come without guarantee and cannot be used as 'evidence' in case you choose to argue about your exam grade.

Ex. 1 — (lesson 2: Probability Theory). Box 1 contains 8 apples and 4 oranges. Box 2 contains 10 apples and 2 oranges. Boxes are chosen with equal probability.
(a) What is the probability of choosing an apple?
(b) If an apple is chosen, what is the probability that it came from box 1?

Answer (ex. 1) — The following probabilities are given in the problem statement,

$$p(b_1) = p(b_2) = 1/2$$
$$p(a|b_1) = 8/12 \qquad p(a|b_2) = 10/12$$
$$p(o|b_1) = 4/12 \qquad p(o|b_2) = 2/12$$

(a) $p(a) = \sum_i p(a, b_i) = \sum_i p(a|b_i)p(b_i) = \frac{8}{12} \cdot \frac{1}{2} + \frac{10}{12} \cdot \frac{1}{2} = \frac{3}{4}$
(b) $p(b_1|a) = \frac{p(a,b_1)}{p(a)} = \frac{p(a|b_1)p(b_1)}{p(a)} = \frac{\frac{8}{12} \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{4}{9}$

Ex. 2 — (lesson 2: Probability Theory). Derive the 'generalized sum rule',

$$p(\mathcal{A} + \mathcal{B}) = p(\mathcal{A}) + p(\mathcal{B}) - p(\mathcal{A}, \mathcal{B})$$

from the 'elementary sum rule' $p(\mathcal{A}) + p(\bar{\mathcal{A}}) = 1$ and the product rule. Use the fact that $\mathcal{A} + \mathcal{B} = \overline{\bar{\mathcal{A}}\bar{\mathcal{B}}}$ (from Boolean logic).

Answer (ex. 2) —

$$
\begin{aligned}
&p\left(\mathcal{A}+\mathcal{B}\right) \underset{bool}{=} p\left(\overline{\bar{\mathcal{A}}\bar{\mathcal{B}}}\right) \underset{sum}{=} 1-p\left(\bar{\mathcal{A}}\bar{\mathcal{B}}\right) \underset{prod}{=} 1-p\left(\bar{\mathcal{A}}|\bar{\mathcal{B}}\right)p\left(\bar{\mathcal{B}}\right) \\
&\underset{sum}{=} 1-\left(1-p\left(\mathcal{A}|\bar{\mathcal{B}}\right)\right)\left(1-p\left(\mathcal{B}\right)\right) = p(\mathcal{B})+\left(1-p\left(\mathcal{B}\right)\right)p\left(\mathcal{A}|\bar{\mathcal{B}}\right) \\
&\underset{prod}{=} p(\mathcal{B})+\left(1-p\left(\mathcal{B}\right)\right)p\left(\bar{\mathcal{B}}|\mathcal{A}\right)\frac{p\left(\mathcal{A}\right)}{p\left(\bar{\mathcal{B}}\right)} \underset{sum}{=} p(\mathcal{B})+p\left(\bar{\mathcal{B}}|\mathcal{A}\right)p\left(\mathcal{A}\right) \\
&\underset{sum}{=} p(\mathcal{B})+\left(1-p\left(\mathcal{B}|\mathcal{A}\right)\right)p\left(\mathcal{A}\right) \underset{prod}{=} p\left(\mathcal{A}\right)+p(\mathcal{B})-p\left(\mathcal{A},\mathcal{B}\right)
\end{aligned}
$$

If $\mathcal{A}$ and $\mathcal{B}$ can not be true at the same time (we say: $\mathcal{A}$ and $\mathcal{B}$ are *mutually exclusive*), it follows that $p(\mathcal{A},\mathcal{B})=0$ and consequently in this case $p(\mathcal{A}+\mathcal{B}) = p(\mathcal{A})+p(\mathcal{B})$.

Ex. 3 — (*DM03, ex.2.36*)[1]. (lesson 2: Probability Theory). The inhabitants of an island tell the truth one third of the time. They lie with probability 2/3. On an occasion, after one of them made a statement, you ask another 'was that statement true?' and he says 'yes'. What is the probability that the statement was indeed true?

Answer (ex. 3) — We use variables S_1 and S_2 for statements 1 and 2 and shorthand 'y','n','t' and 'f' for 'yes', 'no', 'true' and 'false', respectively. The problem statement provides us with the following probabilities,

$$
\begin{aligned}
p(S_1 = \text{'t'}) &= 1/3 \\
p(S_1 = \text{'f'}) &= 1 - p(S_1 = \text{'t'}) = 2/3 \\
p(S_2 = \text{'y'}|S_1 = \text{'t'}) &= 1/3 \\
p(S_2 = \text{'y'}|S_1 = \text{'f'}) &= 1 - p(S_2 = \text{'y'}|S_1 = \text{'t'}) = 2/3
\end{aligned}
$$

We are asked to compute $p(S_1 = \text{'t'}|S_2 = \text{'y'})$. Use Bayes rule,

$$
\begin{aligned}
p(S_1 = \text{'t'}|S_2 = \text{'y'}) &= \frac{p(S_1 = \text{'t'}, S_2 = \text{'y'})}{p(S_2 = \text{'y'})} \\
&= \frac{p(S_2 = \text{'y'}|S_1 = \text{'t'})p(S_1 = \text{'t'})}{p(S_2 = \text{'y'}|S_1 = \text{'t'})p(S_1 = \text{'t'}) + p(S_2 = \text{'y'}|S_1 = \text{'f'})p(S_1 = \text{'f'})} \\
&= \frac{\frac{1}{3}\cdot\frac{1}{3}}{\frac{1}{3}\cdot\frac{1}{3}+\frac{2}{3}\cdot\frac{2}{3}} = \frac{1}{5}
\end{aligned}
$$

Ex. 4 — (*DM03, ex.3.12*). (lesson 2: Probability Theory). A bag contains one ball, known to be either white or black. A white ball is put in, the bag is shaken, and a ball is drawn out, which proves to be white. What is now the chance of drawing a white ball? [Notice that the state of the bag, after the operations, is exactly identical to its state before.]

Answer (ex. 4) — There are two hypotheses: let $H = 0$ mean that the original ball in the bag was white and $H = 1$ that is was black. Assume the

[1] David Mackay, Information Theory, Inference, and Learning algorithms, Cambridge Press, 2003 (accessible at http://www.inference.org.uk/itila/book.html)

prior probabilities are equal. The data is that when a randomly selected ball was drawn from the bag, which contained a white one and the unknown one, it turned out to be white. The probability of this result according to each hypothesis is:

$$P(D|H=0)=1, \quad P(D|H=1)=1/2$$

So by Bayes theorem, the posterior probability of H is

$$P(H=0|D)=2/3, \quad P(H=1|D)=1/3$$

Ex. 5 — (lesson 2: Probability Theory). A dark bag contains five red balls and seven green ones.
(a) What is the probability of drawing a red ball on the first draw? Balls are not returned to the bag after each draw.
(b) If you know that on the second draw the ball was a green one, what is now the probability of drawing a red ball on the first draw?

Answer (ex. 5) — (a) $p(S_1=R)=\frac{N_R}{N_R+N_G}=\frac{5}{12}$
(b) The outcome of the nth draw is referred to by variable S_n. Use Bayes rule to get

$$\begin{aligned} p(S_1=\mathrm{R}|S_2=\mathrm{G}) &= \frac{p(S_2=\mathrm{G}|S_1=\mathrm{R})p(S_1=\mathrm{R})}{p(S_2=\mathrm{G}|S_1=\mathrm{R})p(S_1=\mathrm{R})+p(S_2=\mathrm{G}|S_1=\mathrm{G})p(S_1=\mathrm{G})} \\ &= \frac{\frac{7}{11}\cdot\frac{5}{12}}{\frac{7}{11}\cdot\frac{5}{12}+\frac{6}{11}\cdot\frac{7}{12}}=\frac{5}{11} \end{aligned}$$

Ex. 6 — (lesson 2: Probability Theory). Consider two jointly distributed variables x and y, where x is an input and y is a response variable. As usual, the (joint) expectation is defined as $\mathbb{E}[g(x,y)]=\int_x\int_y g(x,y)p(x,y)\,\mathrm{d}x\,\mathrm{d}y$ and the *conditional expectation* as $\mathbb{E}[g(y)|x]=\int_y g(y)p(y|x)\,\mathrm{d}y$.
(a) Is $\mathbb{E}[y|x]$ a function of x only, of y only or is it a function of both x and y?
(b) Let's interpret $\hat{y}=\mathbb{E}[y|x]$ as an estimator for the outcomes y. Prove that for any function $f(x)$ of the input variables,

$$\mathbb{E}[yf^T(x)]=\mathbb{E}[\hat{y}f^T(x)]$$

(c) Prove that the conditional mean estimator $\mathbb{E}[y|x]$ minimizes the mean squared error loss function over all regression functions $f(x)$, i.e., prove that

$$\mathbb{E}\left[(y-f(x))^2\right]\geq\mathbb{E}\left[(y-\mathbb{E}[y|x])^2\right]$$

(d) Interpret this result.

Answer (ex. 6) — (a) $\mathbb{E}[y|x]$ a function of x only.

(b)

$$\begin{aligned}
\mathbb{E}[yf^T(x)] &= \int_x \int_y yf^T(x)p(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_x \int_y yf^T(x)p(y|x)p(x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_x \left[\int_y yp(y|x)\,\mathrm{d}y\right] f^T(x)p(x)\,\mathrm{d}x \\
&= \int_x \hat{y}f^T(x)p(x)\,\mathrm{d}x \\
&= \mathbb{E}[\hat{y}f^T(x)]
\end{aligned}$$

(c) If we decompose the error between y and $f(x)$ into

$$y - f(x) = (y - \hat{y}) + (\hat{y} - f(x))$$

then the MSE between y and $f(x)$ is

$$\begin{aligned}
\mathbb{E}[(y - f(x))^T(y - f(x))] &= \mathbb{E}[(y - \hat{y})^T(y - \hat{y})] + \mathbb{E}[(\hat{y} - f(x))^T(\hat{y} - f(x))] \\
&\geq \mathbb{E}[(y - \hat{y})^T(y - \hat{y})]
\end{aligned}$$

with equality only if $f(x) = \hat{y}$.

Ex. 7 — (lesson 3: Bayesian Machine Learning). (a) Explain shortly the relation between machine learning and Bayes rule.
(b) How are Maximum a Posteriori (MAP) and Maximum Likelihood (ML) estimation related to Bayes rule and machine learning?

Answer (ex. 7) — (a) Machine learning is inference over models (hypotheses, parameters, etc.) from a given data set. Bayes rule makes this statement precise. Let $\theta \in \Theta$ and D represent a model parameter vector and the given data set, respectively. Then, Bayes rule,

$$p(\theta|D) = \frac{p(D|\theta)}{p(D)}p(\theta)$$

relates the information that we have about θ before we saw the data (i.e., the distribution $p(\theta)$) to what we know after having seen the data, $p(\theta|D)$.
(b) The *Maximum a Posteriori* (MAP) estimate picks a value $\hat{\theta}$ for which the posterior distribution $p(\theta|D)$ is maximal, i.e.,

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D)$$

In a sense, MAP estimation approximates Bayesian learning, since we approximated $p(\theta|D)$ by $\delta(\theta - \hat{\theta}_{MAP})$. Note that, by Bayes rule,

$$\arg\max_{\theta} p(\theta|D) = \arg\max_{\theta} p(D|\theta)p(\theta)$$

If we further assume that prior to seeing the data all values for θ are equally likely (i.e., $p(\theta) = \text{const.}$), then the MAP estimate reduces to the *Maximum Likelihood* estimate,

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(D|\theta)$$

Ex. 8 — (lesson 5: density estimation).
We are given an IID data set $D = \{x_1, x_2, \ldots, x_N\}$, where $x_n \in \mathcal{R}^M$. Let's assume that the data were drawn from a multivariate Gaussian (MVG),

$$\begin{aligned} p(x_n|\theta) &= \mathcal{N}(x_n|\,\mu, \Sigma) \\ &= |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)\right\} \end{aligned}$$

(a) Derive the log-likelihood of the parameters for these data.

(b) Derive the maximum likelihood estimates for μ and Σ.

Answer (ex. 8) — See lecture notes (on class homepage).

Ex. 9 — (lesson 5: density estimation).
Now we consider IID data $D = \{x_1, x_2, \ldots, x_N\}$ obtained from tossing a K-sided die. We use a *binary selection variable*

$$x_{nk} \equiv \begin{cases} 1 & \text{if } x_n \text{ lands on } k\text{th face} \\ 0 & \text{otherwise} \end{cases}$$

with probabilities $p(x_{nk} = 1) = \theta_k$.
(a) Write down the probability for the nth observation $p(x_n|\theta)$ and derive the log-likelihood $\ell(\theta; D) \equiv \log p(D|\,\theta)$.

(b) Derive the maximum likelihood estimate for θ.

Answer (ex. 9) — See lecture notes (on class homepage).

(a) $p(x_n|\theta) = \prod_k \theta_k^{x_{nk}}$ subject to $\sum_k \theta_k = 1$.

$$\ell(\theta) = \sum_k m_k \log \theta_k$$

where $m_k = \sum_k x_{nk}$.
(b) $\hat{\theta} = \frac{m_k}{N}$, the *sample proportion*.

Ex. 10 — (lesson 7: Generative Classification). You have a machine that measures property x, the 'orangeness' of liquids. You wish to discriminate between C_1 = 'Fanta' and C_2 = 'Orangina'. It is known that

$$p(x|C_1) = \begin{cases} 10 & 1.0 \le x \le 1.1 \\ 0 & \text{otherwise} \end{cases}$$

$$p(x|C_2) = \begin{cases} 200(x-1) & 1.0 \le x \le 1.1 \\ 0 & \text{otherwise} \end{cases}$$

The prior probabilities $p(C_1) = 0.6$ and $p(C_2) = 0.4$ are also known from experience.

(a) A 'Bayes Classifier' is given by

$$\text{Decide } C_1 \text{ if } p(C_1|x) > p(C_2|x); \text{ otherwise decide } C_2$$

Calculate the optimal Bayes classifier.
(b) The probability of making the wrong decision, given x, is

$$p(\text{error}|x) = \begin{cases} p(C_1|x) & \text{if we decide } C_2 \\ p(C_2|x) & \text{if we decide } C_1 \end{cases} \tag{1}$$

Compute the *total* error probability $p(\text{error})$ for the Bayes classifier in this example.

Answer (ex. 10) — (a) We choose C_1 if $p(C_1|x)/p(C_2|x) > 1$. This condition can be worked out as

$$\frac{p(C_1|x)}{p(C_2|x)} = \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} = \frac{10 \times 0.6}{200(x-1) \times 0.4} > 1$$

which evaluates to choosing

$$\begin{aligned} C_1 \quad & \text{if } 1.0 \leq x < 1.075 \\ C_2 \quad & \text{if } 1.075 \leq x \leq 1.1 \end{aligned}$$

The probability that x falls outside the interval $[1.0, 1.1]$ is zero.

(b) The total probability of error $p(\text{error}) = \int_x p(\text{error}|x)p(x)\,\mathrm{d}x$. We can work this out as

$$\begin{aligned} p(\text{error}) &= \int_x p(\text{error}|x)p(x)\,\mathrm{d}x \\ &= \int_{1.0}^{1.075} p(C_2|x)p(x)\,\mathrm{d}x + \int_{1.075}^{1.1} p(C_1|x)p(x)\,\mathrm{d}x \\ &= \int_{1.0}^{1.075} p(x|C_2)p(C_2)\,\mathrm{d}x + \int_{1.075}^{1.1} p(x|C_1)p(C_1)\,\mathrm{d}x \\ &= \int_{1.0}^{1.075} 0.4 \cdot 200(x-1)\,\mathrm{d}x + \int_{1.075}^{1.1} 0.6 \cdot 10\,\mathrm{d}x \\ &= 80 \cdot [x^2/2 - x]_{1.0}^{1.075} + 6 \cdot [x]_{1.075}^{1.1} \\ &= 0.225 + 0.15 \\ &= 0.375 \end{aligned}$$

Ex. 11 — (lesson 7: Generative Classification and lesson 8: Discriminative Classification). Describe shortly in your own words the similarities and differences between the discriminative and generative approach to classification.

Ex. 12 — (Logistic regression). (lesson 8: Discriminative Classification). Given a data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_n \in \mathcal{R}^M$ and $y_n \in \{0, 1\}$. The probabilistic classification method known as *logistic regression* attempts to model these data as

$$p(y_n = 1|x_n) = \sigma(\theta^T x_n + b)$$

where $\sigma(x) = 1/(1+e^{-x})$ is the *logistic function*. Let's introduce shorthand notation $\mu_n = \sigma(\theta^T x_n + b)$. So, for every input x_n, we have a model output μ_n and an actual data output y_n.

(a) Express $p(y_n|x_n)$ as a Bernoulli distribution in terms of μ_n and y_n.
(b) If furthermore is given that the data set is IID, show that the log-likelihood is given by

$$\ell(\theta; D) = \sum_n \{y_n \log \mu_n + (1-y_n)\log(1-\mu_n)\}$$

(c) Prove that the derivative of the logistic function is given by

$$\sigma'(\xi) = \sigma(\xi)\cdot[1-\sigma(\xi)]$$

(d) Show that the derivative of the log-likelihood is

$$\nabla_\theta \ell = \sum_{n=1}^{N} \left(y_n - \sigma(\theta^T x_n + b)\right) x_n$$

(e) Design a gradient-ascent algorithm for maximizing $\ell(\theta; D)$ with respect to θ.
(f) Interpret this result.

Answer (ex. 12) — -

(a) $p(y_n|x_n) = p(y_n=1|x_n)^{y_n} p(y_n=0|x_n)^{1-y_n} = \mu_n^{y_n}(1-\mu_n)^{1-y_n}$
(b) The log-likelihood is given by

$$\begin{aligned}\ell(\theta; D) &= \log p(D|\theta) = \sum_n \log p(y_n|x_n, \theta) \\ &= \sum_n \{y_n \log \mu + (1-y_n)\log(1-\mu_n)\}\end{aligned}$$

(c)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\xi}\left(\frac{1}{1+e^{-\xi}}\right) &= \frac{(1+e^{-\xi})\cdot 0 - (-e^{-\xi}\cdot 1)}{(1+e^{-\xi})^2} \\ &= \frac{e^{-\xi}}{(1+e^{-\xi})^2} = \frac{1}{1+e^{-\xi}}\cdot\frac{e^{-\xi}}{1+e^{-\xi}} \\ &= \sigma(\xi)\,[1-\sigma(\xi)]\end{aligned}$$

(d)

$$\begin{aligned}\nabla_\theta \ell(\theta) &= \sum_n \left(\frac{y_n}{\mu_n} - \frac{1-y_n}{1-\mu_n}\right)\cdot\frac{\partial \mu_n}{\partial(\theta^T x_n + b)}\cdot\frac{\partial(\theta^T x_n + b)}{\partial\theta} \\ &= \sum_n \frac{y_n - \mu_n}{\mu_n(1-\mu_n)}\cdot \mu_n(1-\mu_n)\cdot x_n \\ &= \sum_n (y_n - \mu_n) x_n\end{aligned}$$

(e)

$$\theta^{(t+1)} = \theta^{(t)} + \rho \sum_n (y_n - \mu_n^{(t)}) x_n$$

Ex. 13 — (lesson 9: Clustering). Consider a data set $D = \{x_1, x_2, \ldots, x_N\}$ and a MVG generative model for these data. The maximum likelihood estimate (MLE) of the mean value of this MVG distribution is given by

$$\hat{\mu} = \frac{1}{N} \sum_n x_n \tag{2}$$

In the lecture notes on linear generative classification, we derived the following expression for the MLE of the *class-conditional* mean,

$$\hat{\mu}_k = \frac{\sum_n t_{nk} x_n}{\sum_n t_{nk}} \tag{3}$$

(a) Explain this formula. What does t_{nk} represent? Relate this formula to the expression for the MLE of the mean for a MVG.

We then discussed MLE for a *clustering* problem and derived

$$\hat{\mu}_k^{(t)} = \frac{\sum_n r_n^{k(t)} x_n}{\sum_n r_n^{k(t)}} \tag{4}$$

(b) Again, explain this latter formula. What does r_n^k represent? Why the superscript (t)?

Let z_n^k be the usual binary selection variable, corresponding to the kth cluster. (c) Express $r_n^{k(t)}$ as a conditional probability distribution (that involves both x_n and z_n).

Answer (ex. 13) — (a) For a classification problem, we use t_{nk} as a binary indicator variable, i.e.,

$$t_{nk} = \begin{cases} 1 & \text{if } k\text{th class} \\ 0 & \text{else} \end{cases}$$

Equation 3 computes the sample proportion, just like eq. 2 , but now only for samples from class k.

(b) $0 \leq r_n^k \leq 1$ is a *soft* class indicator. It is our best estimate of the binary class indicator t_{nk}, given the input x_n. The superscript (t) is the iteration index in an iterative update algorithm such as EM; we need this because in clustering we don't have a one-step solution to the maximum likelihood estimation problem.

(c) The E-step (in the EM algo) for clustering:

$$r_n^{k(t+1)} = p(z_n^k | x_n, \theta^{(t)})$$

Ex. 14 — (lesson 11: Continuous Latent Variable models). Again we consider an observed data set $D = \{x_1, x_2, \ldots, x_N\}$ where $x_n \in \mathcal{R}^M$. This time we assume that the data were generated according to a model $x_n = \Lambda z_n + v_n$ where $z_n \in \mathcal{R}^K$, $K < M$. The samples z_n are IID drawn from a distribution $\mathcal{N}(0, I)$ and $v_n \sim \mathcal{N}(0, \Psi)$. Furthermore, we assume that z_n and v_n are uncorrelated, i.e., $\epsilon[z_n v_n^T] = 0$.

(a) Write this model in terms of $p(x_n|z_n)$ and a prior $p(z_n)$.

Note that $p(x_n)$ is a Gaussian distribution since products and marginals of Gaussian distributions yield again Gaussian distributions.
(b) Given that $p(x_n)$ is Gaussian, proof that

$$p(x_n) = \mathcal{N}(0, \Lambda\Lambda^T + \Psi)$$

(hint: workout the expectation and variance of x_n).
(c) Why is this model not very interesting if the only constraint for Ψ is that it's a symmetric positive definite matrix? What's so interesting about this model if Ψ is constrained to be a diagonal matrix?

In the E-step of an EM-algorithm for estimating Ψ, we compute the posterior distribution of hidden factors z_n, given the observed data, as

$$\begin{aligned} q_n^{(t+1)}(z) &= p(z|x_n, \theta^{(t)}) = \mathcal{N}(z|\, m_n, V) \\ V &= (I + \Lambda^T \Psi^{-1} \Lambda)^{-1} \\ m_n &= V \Lambda^T \Psi^{-1} x_n \end{aligned}$$

In the M-step, we then maximize the free energy function w.r.t Ψ and obtain

$$\hat{\Psi}^{(t+1)} = \text{diag}\left\{ \Lambda^{(t+1)} V (\Lambda^{(t+1)})^T + \frac{1}{N} \sum_n (x_n - \Lambda m_n)(x_n - \Lambda m_n)^T \right\}$$

(d) Use the expression for $\hat{\Psi}^{(t+1)}$ to explain differences (and similarities) between this model and the linear regression model.

Answer (ex. 14) — (a)

$$p(x_n|z_n) = \mathcal{N}(\Lambda z_n, \Psi), \quad p(z_n) = \mathcal{N}(0, I)$$

(b)

$$\begin{aligned} \epsilon[x_n] &= \epsilon[\Lambda z_n + v_n] = \Lambda \epsilon[z_n] + \epsilon[v_n] = 0 \\ \text{var}[x_n] &= \epsilon[(x_n - \epsilon[x_n])(x_n - \epsilon[x_n])^T] \\ &= \epsilon[(\Lambda z_n + v_n)(\Lambda z_n + v_n)^T] \\ &= \Lambda \epsilon[z_n z_n^T] \Lambda^T + \epsilon[v_n v_n^T] \\ &= \Lambda\Lambda^T + \Psi \end{aligned}$$

(c) Because setting $\Lambda = 0$ would result in a 'regular' gaussian model; i.o.w. it's no more interesting than any other gaussian model. If Ψ is diagonal, then all

correlations between the components in x *must* be absorbed ('explained') by the rank-K matrix $\Lambda\Lambda^T$. This is interesting because $K < M$ and hence this model attempts to achieve dimensionality reduction.

(d) Factor Analysis is much like unsupervised linear regression (plus a few constraints on the noise covariance matrix Ψ). The MLE for Ψ looks the same for both FA and LR, apart from an extra term $\Lambda V \Lambda^T$ in the FA case. This latter term reflects the uncertainty about the inputs (z_n). In FA, our knowledge about the inputs is expressed as the posterior $p(z|x_n, \theta^{(t)}) = \mathcal{N}(z|\, m_n, V)$. I.o.w., the uncertainty about the inputs is reflected by the covariance matrix V (and it shows up as $\Lambda V \Lambda^T$ in the M-step for the MLE of Ψ. In LR, the inputs are exactly known, i.e. $V = 0$ and in the MLE expression the term $\Lambda V \Lambda^T$ vanishes. Lastly, since FA contains hidden inputs, we cannot solve MLE in one step, but need to resort to an iterative algorithm (such as EM).

Ex. 15 — (Lesson 13: Dynamic latent variable models).
(a) What is the 1st-order Markov assumption?
(b) Derive the joint probability distribution $p(x_{1:T}, z_{0:T})$ (where x_t and z_t are observed and latent variables respectively) for the state-space model with ftransition and observation models ($p(z_t|z_{t-1})$ and $p(x_t|z_t)$).
(c) What is a Hidden Markov Model (HMM)?
(d) What is a Linear Dynamical System (LDS)?
(e) What is a Kalman Filter?
(f) How does the Kalman Filter relate to the LDS? And to Factor Analysis (FA)?
(g) Explain the popularity of Kalman filtering and HMMs?
(h) How relates a HMM to a GMM?

Answer (ex. 15) — (a) An auto-regressive model is first-order Markov if $p(x_t|x_{t-1}, x_{t-1}, \ldots, x_1) = p(x_t|x_{t-1})$.
(b)

$$p(x_{1:T}, z_{0:T}) = p(z_0) \prod_{t=1}^{T} p(z_t|z_{t-1}) \prod_{t=1}^{T} p(x_t|z_t)$$

(c) An HMM is a state-space model (as described in (b)) where the latent variable z_t is discretely valued. Iow, the HMM has hidden clusters.
(d) An LDS is a state-space model (also described by the eq in (b)), but now the latent variable z_t is continuously valued.
(e) A Kalman filter is a recursive solution to the inference problem $p(z_t|x_t, x_{t-1}, \ldots, x_1)$, based on an estimate at the previous time step $p(z_{t-1}|x_{t-1}, x_{t-2}, \ldots, x_1)$ and a new observation x_t. Basically, it's a filter that updates the optimal Bayesian estimate of the current state z_t based on all past observations $x_t, x_{t-1}, \ldots, x_1$.
(f) The LDS describes a (generative) *model*. The Kalman filter does not describe a model, but rather describes an *inference task* on the LDS model. The Kalman filter can also be understood as factor-analysis-over-time. In both cases, there is a latent continuously-valued (set of) variables. In the case of the Kalman filter, the corresponding model (LDS model) introduces 1st order Markov temporal dependencies between the states by the transition probabilities $p(z_t|z_{t-1})$. These temporal dependencies are absent for factor analysis models.

(g) The LDS and HMM models are both quite general and flexible generative probabilistic models for time series. There exists very efficient algorithms for executing the latent state inference tasks (Kalman filter for LDS and there is a similar algorithm for the HMM). That makes these models flexible and practical. Hence the popularity of these models.
(h) An HMM can be interpreted as a Gaussian-Mixture-model-over-time, in the same way as an LDS can be seen as factor-analysis-over-time.

Ex. 16 — Work through previous exams!!