Extending the Bayesian Deep Learning Method MultiSWAG

Scott Brownlie

Master of Science
School of Informatics
University of Edinburgh
2021

Abstract

This skeleton demonstrates how to use the `infthesis` style for MSc dissertations in Artificial Intelligence, Cognitive Science, Computer Science, Data Science, and Informatics. It also emphasises the page limit, and that you must not deviate from the required style. The file `skeleton.tex` generates this document and can be used as a starting point for your thesis. The abstract should summarise your report and fit in the space on the first page.

Acknowledgements

Any acknowledgements go here.

Table of Contents

Chapter 1

Introduction

The preliminary material of your report should contain:

- The title page.
- An abstract page.
- Optionally an acknowledgements page.
- The table of contents.

As in this example `skeleton.tex`, the above material should be included between:

```
\begin{preliminary}
    ...
\end{preliminary}
```

This style file uses roman numeral page numbers for the preliminary material.

The main content of the dissertation, starting with the first chapter, starts with page 1. ***The main content must not go beyond page 40.***

The report then contains a bibliography and any appendices, which may go beyond page 40. The appendices are only for any supporting material that's important to go on record. However, you cannot assume markers of dissertations will read them.

You may not change the dissertation format (e.g., reduce the font size, change the margins, or reduce the line spacing from the default 1.5 spacing). Over length or incorrectly-formatted dissertations will not be accepted and you would have to modify your dissertation and resubmit. You cannot assume we will check your submission before the final deadline and if it requires resubmission after the deadline to conform to the page and style requirements you will be subject to the usual late penalties based on your final submission time.

1.1 Using Sections

Divide your chapters into sub-parts as appropriate.

1.2 Citations

Citations (such as [**?**] or [**?**]) can be generated using `BibTeX`. For more advanced usage, the `natbib` package is recommended. You could also consider the newer `biblatex` system.

These examples use a numerical citation style. You may also use (Author, Date) format if you prefer.

Chapter 2

Your next chapter

A dissertation usually contains several chapters.

Chapter 3

Factor Analysis

FA is a latent variable model which generates observations $\theta \in \mathbb{R}^d$ as follows. First, a latent vector $\mathbf{h} \in \mathbb{R}^K$, for some $K < d$, is sampled from $p(\mathbf{h}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Next, $\mathbf{h}$ is transformed onto a K-dimensional linear subspace of $\mathbb{R}^d$ by left-multiplying it by a *factor loading* matrix $\mathbf{F} \in \mathbb{R}^{d \times K}$. The origin of this subspace is then shifted by adding a bias term $\mathbf{c} \in \mathbb{R}^d$. Finally, the data is perturbed by adding some zero mean Gaussian noise $\varepsilon \in \mathbb{R}^d$ sampled from $\mathcal{N}(\mathbf{0}, \Psi)$, where Ψ is a $d \times d$ diagonal matrix [1]. Putting all this together, an observation $\theta \in \mathbb{R}^d$ is generated according to

$$\theta = \mathbf{F}\mathbf{h} + \mathbf{c} + \varepsilon. \tag{3.1}$$

In this context, an observation θ is the parameter vector of a neural network.

It follows that, given $\mathbf{h}$, the observations θ are Gaussian distributed with mean $\mathbf{F}\mathbf{h} + \mathbf{c}$ and covariance Ψ [1]. Formally,

$$p(\theta|\mathbf{h}) = \mathcal{N}\Big(\mathbf{F}\mathbf{h} + \mathbf{c}, \Psi\Big) = \frac{1}{\sqrt{(2\pi)^d |\Psi|}} \exp\Big(-\frac{1}{2}(\theta - \mathbf{F}\mathbf{h} - \mathbf{c})^\intercal \Psi^{-1} (\theta - \mathbf{F}\mathbf{h} - \mathbf{c})\Big), \tag{3.2}$$

where $|\Psi|$ is the *determinant* of Ψ. From [1], integrating $p(\theta|\mathbf{h})$ over $\mathbf{h}$ gives the marginal distribution

$$p(\theta) = \mathcal{N}\big(\mathbf{c}, \mathbf{F}\mathbf{F}^\intercal + \Psi\big). \tag{3.3}$$

The parameters of the model are $\mathbf{c}, \mathbf{F}$ and Ψ. The value of $\mathbf{c}$ which maximises the likelihood of the observed data is the empirical mean of the observations [1], which in this case is θ_{SWA}. Having set the bias term, an expectation maximisation (EM) or singular value decomposition (SVD) algorithm can find the maximum likelihood estimates of $\mathbf{F}$ and Ψ [1]. However, both methods require storing all the observations in memory, making them impractical for high-dimensional data with lots of observations. Two alternative online algorithms are presented in Sections 3.1 and 3.2.

3.1 Online Stochastic Gradient Algorithm

In situations where learning latent variable models with the classic EM algorithm is slow, [1] suggests optimising the log-likelihood of the model parameters via gradient methods. Since FA is a latent variable model, this approach can be applied here. In this case, the log-likelihood of the parameters $\mathbf{F}$ and Ψ is

$$L(\mathbf{F},\Psi) = \frac{1}{T}\sum_{t=1}^{T} \log p(\theta_t|\mathbf{F},\Psi). \tag{3.4}$$

The partial derivatives of the log-likelihood with respect to $\mathbf{F}$ and Ψ are therefore

$$\nabla_{\mathbf{F},\Psi} L(\mathbf{F},\Psi) = \frac{1}{T}\sum_{t=1}^{T} \nabla_{\mathbf{F},\Psi} \log p(\theta_t|\mathbf{F},\Psi). \tag{3.5}$$

Computing $\nabla_{\mathbf{F},\Psi} L(\mathbf{F},\Psi)$ in full would require a pass over all observations, $\theta_1,\ldots,\theta_T$. However, a stochastic gradient algorithm can be used instead, as long as the expectation of the sample derivatives is proportional to $\nabla_{\mathbf{F},\Psi} L(\mathbf{F},\Psi)$. By using sample derivatives $\nabla_{\mathbf{F},\Psi} \log p(\theta_t|\mathbf{F},\Psi)$ for $t = 1,\ldots,T$, this condition clearly holds. Hence, as long as the partial derivatives $\nabla_{\mathbf{F},\Psi} \log p(\theta_t|\mathbf{F},\Psi)$ can be computed efficiently, they can be used in conjunction with SGD or any of its variants to optimise $\mathbf{F}$ and Ψ online. After each θ_t is sampled and used to perform a gradient step, it can immediately be discarded.

By adapting the argument for general latent variable models in [1] to FA, the required sample derivatives can be written as

$$\begin{aligned}
\nabla_{\mathbf{F},\Psi} \log p(\theta_t|\mathbf{F},\Psi) &= \frac{1}{p(\theta_t|\mathbf{F},\Psi)} \nabla_{\mathbf{F},\Psi} p(\theta_t|\mathbf{F},\Psi) \\
&= \frac{1}{p(\theta_t|\mathbf{F},\Psi)} \nabla_{\mathbf{F},\Psi} \int_{\mathbf{h}_t} p(\theta_t,\mathbf{h}_t|\mathbf{F},\Psi) \\
&= \frac{1}{p(\theta_t|\mathbf{F},\Psi)} \int_{\mathbf{h}_t} \nabla_{\mathbf{F},\Psi} p(\theta_t,\mathbf{h}_t|\mathbf{F},\Psi) \\
&= \frac{1}{p(\theta_t|\mathbf{F},\Psi)} \int_{\mathbf{h}_t} p(\theta_t,\mathbf{h}_t|\mathbf{F},\Psi) \nabla_{\mathbf{F},\Psi} \log p(\theta_t,\mathbf{h}_t|\mathbf{F},\Psi) \\
&= \int_{\mathbf{h}_t} \frac{p(\theta_t,\mathbf{h}_t|\mathbf{F},\Psi)}{p(\theta_t|\mathbf{F},\Psi)} \nabla_{\mathbf{F},\Psi} \log p(\theta_t,\mathbf{h}_t|\mathbf{F},\Psi) \\
&= \int_{\mathbf{h}_t} p(\mathbf{h}_t|\theta_t,\mathbf{F},\Psi) \nabla_{\mathbf{F},\Psi} \log p(\theta_t,\mathbf{h}_t|\mathbf{F},\Psi) \\
&= \mathbb{E}_{p(\mathbf{h}_t|\theta_t,\mathbf{F},\Psi)} \left[\nabla_{\mathbf{F},\Psi} \log p(\theta_t,\mathbf{h}_t|\mathbf{F},\Psi) \right].
\end{aligned} \tag{3.6}$$

This is as far as the derivation in [1] goes. However, given the form of the FA model, it is possible to manipulate the sample derivatives further. In particular, using the fact

that $\mathbf{h}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is independent from $\mathbf{F}$ and Ψ,

$$\begin{aligned}\nabla_{\mathbf{F},\Psi} \log p(\theta_t, \mathbf{h}_t|\mathbf{F},\Psi) &= \nabla_{\mathbf{F},\Psi} \log \big(p(\theta_t|\mathbf{h}_t,\mathbf{F},\Psi)p(\mathbf{h}_t|\mathbf{F},\Psi)\big) \\ &= \nabla_{\mathbf{F},\Psi} \log \big(p(\theta_t|\mathbf{h}_t,\mathbf{F},\Psi)p(\mathbf{h}_t)\big) \\ &= \nabla_{\mathbf{F},\Psi} \big(\log p(\theta_t|\mathbf{h}_t,\mathbf{F},\Psi) + \log p(\mathbf{h}_t)\big) \\ &= \nabla_{\mathbf{F},\Psi} \log p(\theta_t|\mathbf{h}_t,\mathbf{F},\Psi).\end{aligned} \tag{3.7}$$

Substituting Equation (3.7) into Equation (3.6),

$$\nabla_{\mathbf{F},\Psi} \log p(\theta_t|\mathbf{F},\Psi) = \mathbb{E}_{p(\mathbf{h}_t|\theta_t,\mathbf{F},\Psi)}\big[\nabla_{\mathbf{F},\Psi} \log p(\theta_t|\mathbf{h}_t,\mathbf{F},\Psi)\big]. \tag{3.8}$$

Note that $p(\theta_t|\mathbf{h}_t,\mathbf{F},\Psi)$ is just the Gaussian distribution in Equation (3.2), given $\mathbf{F}$ and Ψ. Hence, substituting $\mathbf{c} = \theta_{\text{SWA}}$ into Equation (3.2) and applying the logarithm,

$$\begin{aligned}\log p(\theta_t|\mathbf{h}_t,\mathbf{F},\Psi) &= -\frac{1}{2}(\theta_t - \mathbf{F}\mathbf{h}_t - \theta_{\text{SWA}})^\intercal \Psi^{-1}(\theta_t - \mathbf{F}\mathbf{h}_t - \theta_{\text{SWA}}) - \frac{1}{2}\log|\Psi| - \frac{d}{2}\log 2\pi \\ &= -\frac{1}{2}(\mathbf{d}_t - \mathbf{F}\mathbf{h}_t)^\intercal \Psi^{-1}(\mathbf{d}_t - \mathbf{F}\mathbf{h}_t) - \frac{1}{2}\log|\Psi| - \frac{d}{2}\log 2\pi,\end{aligned} \tag{3.9}$$

where $\mathbf{d}_t = \theta_t - \theta_{\text{SWA}}$. This is convenient, since Equation (3.9) can be differentiated with respect to both $\mathbf{F}$ and Ψ. Of course, this requires access to θ_{SWA}, which is not available during training. As a compromise - and following the approach of the SWAG covariance approximation - θ_{SWA} can be replaced by the running average of the neural network's parameter vectors.

3.1.1 Partial derivatives with respect to F

From [2], for any symmetric matrix $\mathbf{W}$,

$$\nabla_{\mathbf{A}}(\mathbf{x} - \mathbf{A}\mathbf{s})^\intercal \mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s}) = -2\mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s})\mathbf{s}^\intercal. \tag{3.10}$$

Hence, differentiating Equation (3.9) with respect to $\mathbf{F}$ gives

$$\nabla_{\mathbf{F}} \log p(\theta_t|\mathbf{h}_t,\mathbf{F},\Psi) = \Psi^{-1}(\mathbf{d}_t - \mathbf{F}\mathbf{h}_t)\mathbf{h}_t^\intercal. \tag{3.11}$$

It then follows from Equation (3.8) that $\nabla_{\mathbf{F}} \log p(\theta_t|\mathbf{F},\Psi)$ is the expected value of $\Psi^{-1}(\mathbf{d}_t - \mathbf{F}\mathbf{h}_t)\mathbf{h}_t^\intercal$ over the distribution $p(\mathbf{h}_t|\theta_t,\mathbf{F},\Psi)$. Letting $\mathbb{E}[\cdot]$ denote $\mathbb{E}_{p(\mathbf{h}_t|\theta_t,\mathbf{F},\Psi)}[\cdot]$ to simplify the notation,

$$\begin{aligned}\nabla_{\mathbf{F}} \log p(\theta_t|\mathbf{F},\Psi) &= \mathbb{E}\big[\Psi^{-1}(\mathbf{d}_t - \mathbf{F}\mathbf{h}_t)\mathbf{h}_t^\intercal\big] \\ &= \mathbb{E}\big[\Psi^{-1}\mathbf{d}_t\mathbf{h}_t^\intercal\big] - \mathbb{E}\big[\Psi^{-1}\mathbf{F}\mathbf{h}_t\mathbf{h}_t^\intercal\big] \\ &= \Psi^{-1}\mathbf{d}_t\mathbb{E}\big[\mathbf{h}_t^\intercal\big] - \Psi^{-1}\mathbf{F}\mathbb{E}\big[\mathbf{h}_t\mathbf{h}_t^\intercal\big].\end{aligned} \tag{3.12}$$

From the E-step of the EM algorithm in [1], $p(\mathbf{h}_t|\theta_t,\mathbf{F},\Psi) \propto \mathcal{N}(\mathbf{m}_t,\Sigma)$, where

$$\mathbf{m}_t = (\mathbf{I}+\mathbf{F}^\intercal\Psi^{-1}\mathbf{F})^{-1}\mathbf{F}^\intercal\Psi^{-1}\mathbf{d}_t \quad \text{and} \quad \Sigma = (\mathbf{I}+\mathbf{F}^\intercal\Psi^{-1}\mathbf{F})^{-1}. \tag{3.13}$$

Hence, using identities from [2],

$$\mathbb{E}\left[\mathbf{h}_t^\intercal\right] = \mathbf{m}_t^\intercal \quad \text{and} \quad \mathbb{E}\left[\mathbf{h}_t\mathbf{h}_t^\intercal\right] = \Sigma+\mathbf{m}_t\mathbf{m}_t^\intercal. \tag{3.14}$$

Substituting Equation (3.14) into Equation (3.12),

$$\nabla_{\mathbf{F}} \log p(\theta_t|\mathbf{F},\Psi) = \Psi^{-1}\mathbf{d}_t\mathbf{m}_t^\intercal - \Psi^{-1}\mathbf{F}(\Sigma+\mathbf{m}_t\mathbf{m}_t^\intercal). \tag{3.15}$$

3.1.2 Partial derivatives with respect to Ψ

In order to differentiate Equation (3.9) with respect to Ψ, it helps to use the fact that Ψ is a diagonal matrix. First consider $\mathbf{X}^{-1} = \operatorname{diag}(\frac{1}{x_1},\ldots,\frac{1}{x_d})$ and $\mathbf{a} = (a_1,\ldots,a_d)^\intercal$. Then

$$\mathbf{a}^\intercal\mathbf{X}^{-1}\mathbf{a} = \sum_{i=1}^{d}\frac{a_i^2}{x_i}, \tag{3.16}$$

and so

$$\frac{\partial}{\partial x_i}\mathbf{a}^\intercal\mathbf{X}^{-1}\mathbf{a} = \frac{-a_i^2}{x_i^2} \tag{3.17}$$

for $i = 1,\ldots,d$. Since the partial derivatives of Equation (3.16) with respect to the off-diagonal entries of $\mathbf{X}$ are zero,

$$\begin{aligned}\nabla_{\mathbf{X}}(\mathbf{a}^\intercal\mathbf{X}^{-1}\mathbf{a}) &= \operatorname{diag}\left(\frac{-a_1^2}{x_1^2},\ldots,\frac{-a_d^2}{x_d^2}\right)\\ &= -\operatorname{diag}\Big(\operatorname{diag}(\mathbf{X}^{-2})\odot(\mathbf{a}\odot\mathbf{a})\Big),\end{aligned} \tag{3.18}$$

where $\odot$ denotes the element-wise matrix product. Note that when applied to a d-length vector, $\operatorname{diag}(\cdot)$ represents the $d\times d$ diagonal matrix with the vector on its diagonal, and when applied to a $d\times d$ matrix, $\operatorname{diag}(\cdot)$ represents the d-length vector consisting of the diagonal entries of the matrix.

Substituting $\mathbf{X} = \Psi$ and $\mathbf{a} = \mathbf{d}_t - \mathbf{F}\mathbf{h}_t$, into Equation (3.18),

$$\nabla_\Psi(\mathbf{d}_t-\mathbf{F}\mathbf{h}_t)^\intercal\Psi^{-1}(\mathbf{d}_t-\mathbf{F}\mathbf{h}_t) = -\operatorname{diag}\Big(\operatorname{diag}(\Psi^{-2})\odot\big((\mathbf{d}_t-\mathbf{F}\mathbf{h}_t)\odot(\mathbf{d}_t-\mathbf{F}\mathbf{h}_t)\big)\Big). \tag{3.19}$$

Also, using the identity $\nabla_{\mathbf{X}}\log|\mathbf{X}| = \mathbf{X}^{-\intercal}$ from [2] and the fact that $\Psi^{-\intercal} = \Psi^{-1}$,

$$\nabla_\Psi \log|\Psi| = \Psi^{-1}. \tag{3.20}$$

Hence, using Equation (3.19) and Equation (3.20), the partial derivatives of (3.9) with respect to Ψ are

$$\nabla_\Psi \log p(\theta_t|\mathbf{h}_t,\mathbf{F},\Psi) = \frac{1}{2}\operatorname{diag}\Big(\operatorname{diag}(\Psi^{-2})\odot\big((\mathbf{d}_t-\mathbf{F}\mathbf{h})\odot(\mathbf{d}_t-\mathbf{F}\mathbf{h})\big)\Big)-\frac{1}{2}\Psi^{-1}. \tag{3.21}$$

Again, letting $\mathbb{E}[\cdot]$ denote $\mathbb{E}_{p(\mathbf{h}_t|\theta_t,\mathbf{F},\Psi)}[\cdot]$, it follows from Equation (3.8) that

$$\begin{aligned}
2\cdot\nabla_\Psi \log p(\theta_t|\mathbf{F},\Psi) &= \mathbb{E}\Big[\operatorname{diag}\Big(\operatorname{diag}(\Psi^{-2})\odot\big((\mathbf{d}_t-\mathbf{F}\mathbf{h}_t)\odot(\mathbf{d}_t-\mathbf{F}\mathbf{h}_t)\big)\Big)-\Psi^{-1}\Big]\\
&= \operatorname{diag}\Big(\mathbb{E}\big[\operatorname{diag}(\Psi^{-2})\odot\big((\mathbf{d}_t-\mathbf{F}\mathbf{h}_t)\odot(\mathbf{d}_t-\mathbf{F}\mathbf{h}_t)\big]\Big)-\mathbb{E}\big[\Psi^{-1}\big]\\
&= \operatorname{diag}\Big(\operatorname{diag}(\Psi^{-2})\odot\mathbb{E}\big[\big((\mathbf{d}_t-\mathbf{F}\mathbf{h}_t)\odot(\mathbf{d}_t-\mathbf{F}\mathbf{h}_t)\big]\Big)-\Psi^{-1}.
\end{aligned} \tag{3.22}$$

Now, expanding the expectation inside Equation (3.22) and using Equation (3.14),

$$\begin{aligned}
\mathbb{E}\big[\big((\mathbf{d}_t-\mathbf{F}\mathbf{h})\odot(\mathbf{d}_t-\mathbf{F}\mathbf{h}_t)\big] &= \mathbb{E}\big[\mathbf{d}_t\odot\mathbf{d}_t\big]-2\mathbb{E}\big[\mathbf{d}_t\odot\mathbf{F}\mathbf{h}_t\big]+\mathbb{E}\big[\mathbf{F}\mathbf{h}_t\odot\mathbf{F}\mathbf{h}_t\big]\\
&= \mathbf{d}_t\odot\mathbf{d}_t-2\mathbf{d}_t\odot\mathbf{F}\mathbb{E}\big[\mathbf{h}_t\big]+\mathbb{E}\big[\operatorname{diag}(\mathbf{F}\mathbf{h}_t\mathbf{h}_t^\intercal\mathbf{F}^\intercal)\big]\\
&= \mathbf{d}_t\odot\mathbf{d}_t-2\mathbf{d}_t\odot\mathbf{F}\mathbf{m}_t+\operatorname{diag}\big(\mathbb{E}\big[\mathbf{F}\mathbf{h}_t\mathbf{h}_t^\intercal\mathbf{F}^\intercal\big]\big)\\
&= \mathbf{d}_t\odot\mathbf{d}_t-2\mathbf{d}_t\odot\mathbf{F}\mathbf{m}_t+\operatorname{diag}\big(\mathbf{F}\mathbb{E}\big[\mathbf{h}_t\mathbf{h}_t^\intercal\big]\mathbf{F}^\intercal\big)\\
&= \mathbf{d}_t\odot\mathbf{d}_t-2\mathbf{d}_t\odot\mathbf{F}\mathbf{m}_t+\operatorname{diag}\big(\mathbf{F}(\Sigma+\mathbf{m}_t\mathbf{m}_t^\intercal)\mathbf{F}^\intercal\big)\\
&= \mathbf{d}_t\odot\mathbf{d}_t-2\mathbf{d}_t\odot\mathbf{F}\mathbf{m}_t+\operatorname{sum}\big(\mathbf{F}(\Sigma+\mathbf{m}_t\mathbf{m}_t^\intercal)\odot\mathbf{F},\ \dim=1\big),
\end{aligned} \tag{3.23}$$

where $\operatorname{sum}(\cdot,\ \dim=1)$ denotes the operation of summing along the rows of a matrix. Finally, substituting Equation (3.23) into Equation (3.22) and rearranging,

$$\begin{aligned}
\nabla_\Psi \log p(\theta_t|\mathbf{F},\Psi) = \frac{1}{2}\operatorname{diag}\Big(\operatorname{diag}(\Psi^{-2})\odot\big(&\mathbf{d}_t\odot\mathbf{d}_t-2\mathbf{d}_t\odot\mathbf{F}\mathbf{m}_t\\
&+\operatorname{sum}\big(\mathbf{F}(\Sigma+\mathbf{m}_t\mathbf{m}_t^\intercal)\odot\mathbf{F},\ \dim=1\big)\Big)-\frac{1}{2}\Psi^{-1}.
\end{aligned} \tag{3.24}$$

3.2 Online EM

The classic EM algorithm for FA iteratively optimises the log-likelihood of $\mathbf{F}$ and Ψ by alternating "E" and "M" steps until convergence. Using properties of the Kullback-Leibler divergence, it can be shown that

$$L(\mathbf{F},\Psi)\geq -\sum_{t=1}^{T}\mathbb{E}_{q(\mathbf{h}_t|\theta_t)}\big[\log q(\mathbf{h}_t|\theta_t)\big]+\sum_{t=1}^{T}\mathbb{E}_{q(\mathbf{h}_t|\theta_t)}\big[\log p(\mathbf{h}_t,\theta_t|\mathbf{F},\Psi)\big], \tag{3.25}$$

where the first and second terms on the right-hand side are called the *entropy* and *energy*, respectively, and $q(\mathbf{h}_t|\theta_t)$, $t=1,\ldots,T$, are known as the *variational* distributions

[1]. The EM algorithm optimises this lower bound on the log-likelihood with respect to $\mathbf{F}$, Ψ and also $q(\mathbf{h}_t|\theta_t)$, hence the name "variational distributions". The idea is that, by pushing up the lower bound, the log-likelihood $L(\mathbf{F},\Psi)$ will hopefully increase as well. In fact, it is guaranteed that each iteration of EM does not decrease $L(\mathbf{F},\Psi)$, which again follows from the properties of the Kullback-Leibler divergence [1].

3.2.1 E-step

In the batch E-step, $\mathbf{F}$ and Ψ are fixed and Equation (3.25) is maximised with respect to $q(\mathbf{h}_t|\theta_t)$, $t = 1,\ldots,T$. From [1], the optimal variational distributions are

$$q(\mathbf{h}_t|\theta_t) = p(\mathbf{h}_t|\theta_t,\mathbf{F},\Psi) \propto \mathcal{N}(\mathbf{m}_t,\Sigma), \tag{3.26}$$

where $\mathbf{m}_t$ and Σ are given in Equation (3.13). Note that this optimisation can be performed separately for each θ_t as it is sampled, using the estimates of $\mathbf{F}$ and Ψ on iteration t. However, in batch EM all $q(\mathbf{h}_t|\theta_t)$ are update on every iteration. This is clearly not possible in an online algorithm which discards θ_t before sampling the next observation. Therefore, as a compromise, in the online version each $q(\mathbf{h}_t|\theta_t)$ will be computed once only, on iteration t, and held fixed thereafter. The only other detail is that the batch algorithm uses θ_{SWA} to compute $\mathbf{m}_t$. As θ_{SWA} is not available during training, it will be replaced by the running average of the neural network's parameter vectors, as in the online gradient-based FA algorithm from Section 3.1.

3.2.2 M-step

In the batch M-step, $q(\mathbf{h}_t|\theta_t)$, $t = 1,\ldots,T$, are fixed and Equation (3.25) is maximised with respect to $\mathbf{F}$ and Ψ. From [1], the optimal values are

$$\mathbf{F} = \mathbf{A}\mathbf{H}^{-1}, \tag{3.27}$$

where

$$\mathbf{A} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{d}_t\mathbf{m}_t^\intercal \quad \text{and} \quad \mathbf{H} = \Sigma + \frac{1}{T}\sum_{t=1}^{T}\mathbf{m}_t\mathbf{m}_t^\intercal, \tag{3.28}$$

and

$$\Psi = \operatorname{diag}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{d}_t\mathbf{d}_t^\intercal - 2\mathbf{F}\mathbf{A}^\intercal + \mathbf{F}\mathbf{H}\mathbf{F}^\intercal\right). \tag{3.29}$$

Note that this optimisation involves summing over $t = 1,\ldots,T$. Moreover, on each iteration all components of the sums in Equation (3.28) and Equation (3.29) are updated.

As in the E-step, updating all components is not possible in an online algorithm. Therefore, in the online version the sums will be updated incrementally on each iteration t, with $\mathbf{d}_t$ and $\mathbf{m}_t$ derived from θ_t and the current estimates of $\mathbf{F}$ and Ψ. As in the E-step, θ_{SWA} will be replaced by the running average of the neural network's parameter vectors when computing $\mathbf{d}_t$ and $\mathbf{m}_t$.

Chapter 4

Conclusions

4.1 Final Reminder

The body of your dissertation, before the references and any appendices, *must* finish by page 40. The introduction, after preliminary material, should have started on page 1.

You may not change the dissertation format (e.g., reduce the font size, change the margins, or reduce the line spacing from the default 1.5 spacing). Over length or incorrectly-formatted dissertations will not be accepted and you would have to modify your dissertation and resubmit. You cannot assume we will check your submission before the final deadline and if it requires resubmission after the deadline to conform to the page and style requirements you will be subject to the usual late penalties based on your final submission time.

Bibliography

[1] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[2] Kaare B Petersen and Michael S Pedersen. *The Matrix Cookbook*. Technical University of Denmark, 2012.