# Group Factor Analysis to disentangle common and specific sources of variation between different molecular layers

Ricard Argelaguet Calado

Copenhagen University


A thesis submitted for the degree of
*Master of Science in Bioinformatics*
Copenhagen, August 2016

Under the supervision of:
Professor Anders Krogh and Dr. Oliver Stegle

# Preface

This master thesis serves as documentation for the final assignment in the requirements to achieve the degree Master of Science in Bioinformatics by the University of Copenhagen. The work has been carried out in the period from the 1st of February 2016 to the 8th of August 2016 at the Statistical genomics and Systems genetics group in the European Bioinformatics Institute (EBI/EMBL). The work has been internally supervised by Professor Anders Krogh from the University of Copenhagen and externally supervised by Dr. Oliver Stegle from the EBI/EMBL.

Cambridge, August 2016

# Contents

# Chapter 1

# Abstract

Factor Analysis and probabilistic Principal Component Analysis are cornerstones of classical data analysis. They decompose a multivariate dataset of correlated variables in terms of a potentially smaller number of uncorrelated variables. However, both methods are not appropriate to sistematically analyse datasets consisting of multiple input matrices (views) of co-occuring samples.
In this thesis we developed single-cell Group Factor Analysis (scGFA), a fully bayesian latent variable model with an accurate noise and likelihood assumptions which is able to integrate information from different biological layers. scGFA is a multi-view extension of factor analysis that disentangles the variation unique to a single view and the variation shared between two or more views. Inference is performed using the fast variational bayes framework.
We show that scGFA outperforms current multi-view learning approaches in a simulated data set. We also applied scGFA model to a data set of 61 embryonic stem cells generated by a technology called scMT-seq, a recent method which uses single-cell genome-wide bisulfite sequencing and single-cell RNA-seq to perform a parallel profiling of the DNA methylation and the gene expression in single cells. Our results show the existence of several independent axis of variation, including a major source of covariation between the expression of pluripotency genes and the genome-wide CpG methylation rate. Furthermore, we show that using the multi-view infered latent variables we can cluster the population of embryonic stem cells into three clear subpopulations that are associated with different pluripotency potential and genome-wide methylation rate.

# Notation

## Mathematical symbols

- Matrices are denoted with bold capital letters: $\mathbf{W}$
- Vectors are denoted with bold non-capital letters. If the vector comes from a matrix, two indices separated by a comma will always be showed on the bottom: the first one corresponding to the row and the second one to the column. The symbol : denotes the entire row/column. For instance, $\mathbf{w}_{j,:}$ refers to the entire $j$th row from the $\mathbf{W}$ matrix.
- Scalars are denoted with non-bold non-capital letters. If the value comes from a matrix, two indices separated by a comma will always be showed on the bottom: the first one corresponding to the row and the second one to the column. For instance, $w_{j,k}$ refers to the value coming from the $j$th row and the $k$th column from the $\mathbf{W}$ matrix.
- $\mathbf{0}_k$ is a zero vector of length $K$.
- $\mathbf{I}_k$ is the identity matrix with rank $K$.
- $\mathbb{E}_q[x]$ denotes the expectation of $x$ under the distribution $q$. Sometimes, when the expectations are taken with respec tot the same distribution many times, to avoid cluttered notation we will use $\langle x \rangle$.
- $\mathcal{N}(x \,|\, \mu, \sigma)$: $x$ follows a univariate normal distribution with mean $\mu$ and variance $\sigma$.
- $\mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$: $x$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- $\mathcal{G}(x \,|\, a, b)$: $x$ follows a gamma distribution with parameters $a$ and $b$.
- $\mathcal{U}(x \,|\, a, b)$: $x$ follows a uniform distribution with minimum $a$ and maximum $b$.
- $diag(\mathbf{x})$ is the diagonal operator that takes as input a vector and outputs a diagonal matrix with $\mathbf{x}$ in the diagonal.

## Abbreviations

- FA: factor analysis
- vbFA: variational Bayes factor analysis
- GFA: group factor analysis
- scGFA: single-cell group factor analysis
- PCA: principal component analysis
- pPCA: probabilistic principal component analysis
- vbPCA: variational Bayes principal component analysis
- CCA: canonical correlation analysis
- vbCCA: variational Bayes canonical correlation analysis
- LVM: latent variable model
- ELBO: evidence lower bound
- scRNA-seq: single-cell RNA sequencing
- scMT: single-cell methylome and transcriptome
- ESC: embryonic stem cell
- ARD: automatic relevance determination
- VBEM: variational Bayesian expectation maximisation

## Technical terms

Throughout the report we use the terms latent variable, factor and component indistinguishably, as well as the terms loading and weight.

# Chapter 2

# Introduction

The single-cell Group Factor Analysis (scGFA) model implemented in this project results from the convergence of three areas of statistics: latent variable models, multi-view learning models and variational Bayesian inference.
This chapter provides a brief description of the background required from each of the fields. Section 2.1 describes probabilistic modelling with a main focus on variational Bayesian inference, section 2.3 describes the mathematical background of latent variable models and section 2.4 summarises the previous attempts to solve the multi-view learning problem with latent variable models.

## 2.1 Probabilistic modelling

### 2.1.1 Introduction

A scientific model is a simple theoretical representation of a complex observed phenomenon which allows the systematic study of its behaviour. The general idea is that if a model is able to explain the current observations well enough, then it might be capturing its true underlying laws and can therefore be used to make predictions.
In particular, a computational model consists of a set variables related by mathematical operators that depend on a set of parameters. The procedure of fitting the parameters to observed data is called inference or learning.
Learning is a daunting task because real datasets are noisy, so the model needs to distinguish between signal and noise. Ideally, it should only learn the information relevant for inference or prediction tasks, and disregard the noise. However, in most practical situations this is non-trivial. Highly complex models will fit the data very well but will capture larges amounts of noise, thereby overfitting the training data and leading to a bad generalisation performance. On the other hand, too simple models will fit the data poorly and will not learn its underlying pattern.

The ideas above can be formalised using the concept of probability, and can be dealt with in a rigourous manner using the Bayesian statistical framework.

### 2.1.2 Maximum likelihood inference

A common approach to statistical modelling is to define a generative model of the data $\mathbf{Y}$ with a set of parameters $\mathbf{\Theta}$ that define a probability distribution over the data $p(\mathbf{Y}|\mathbf{\Theta})$, called the likelihood function. A simple approach to fit a model is to estimate the parameters $\hat{\mathbf{\Theta}}$ that maximise the likelihood:

$$\hat{\mathbf{\Theta}} = \arg\max p(\mathbf{Y}|\mathbf{\Theta})$$

This process is called maximum likelihood learning and it generally leads to a fairly straightforward solution. However, since it does not penalise for model complexity, it is known to be prone to overfitting in cases where the data is sparse [8].

### 2.1.3 General Bayesian inference

Another approach to model learning is to use the Bayesian framework, where the parameters themselves are treated as random variables and we aim to obtain probably density functions for $\mathbf{\Theta}$, rather than a single point estimate. To do so, prior knowledge is introduced into the model by specifying a prior probabiliy distribution $p(\mathbf{\Theta})$ over the parameters. Then, using Bayes' theorem, the prior hypothesis is

updated based on the observed data $\mathbf{Y}$ by means of the likelihood $p(\mathbf{Y}|\mathbf{\Theta})$ function, which yields the posterior distribution over the parameters:

$$p(\mathbf{\Theta}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{\Theta})p(\mathbf{\Theta})}{p(\mathbf{Y})}$$

where $p(\mathbf{Y})$ is the marginal likelihood or evidence, a constant normalisation term that can usually be ignored when doing inference. However, as described in section 2.2, the marginal likelihood has a key role in the variational framework.
Again, an important benefit of Bayesian inference is that an entire posterior probability distribution is obtained for each parameter. Thus, a fully Bayesian approach attempts to integrate over all the possible settings of all uncertain quantitites. In other words, the unknown variables are averaged over and so the uncertainity is handled in a coherent way.
Nevertheless, often this calculation is intractable and one has to resort to a point estimate. The simplest approximation to the posterior distribution is to use its mode, which leads to the maximum a posteriori (MAP) estimate:

$$\hat{\mathbf{\Theta}} = \arg\max p(\mathbf{\Theta})p(\mathbf{Y}|\mathbf{\Theta})$$

Note that the only difference with the maximum likelihood objective function is the term $p(\mathbf{\Theta})$, which penalises for model complexity. Therefore, in contrast to maximum likelihood inference, Bayesian approaches naturally handle the problem of model complexity and overfitting.

### On the choice of prior distributions

The prior distribution is a key part of Bayesian inference and represents the information about an uncertain parameter before observing the data. With well-identified parameters and large sample sizes, the choice of prior has minor effects on the posterior estimates, but when data is sparse, its definition can have a major impact on the model [25].
In general, there are three commonly used approaches to guide the choice of prior distributions: subjective, empirical and objective: in the subjective approach the prior quantifies what is known before the experiment takes place. In the empirical approach the prior is estimated from the data itself, and in the objective approach the prior is chosen based on convenient mathematical properties.
Objective priors are by far the most commonly used in order to make Bayesian inference tractable [25]. If the likelihood and the prior distributions are not conjugate (they do not belong to the same family of probability distributions) then there is no closed-form solution for the posterior and one has to resort to numerical approaches. Therefore, unless absolutely required, the priors are selected in such a way that they are conjugated with the likelihood.
It is important to take into account that the term objective can be misleading, since prior information still needs to be encoded in the parameters of the prior distribution. In most scenarios, to avoid the prior from having an important effect on the posterior, the parameters are set to make the distribution have as less information as possible. For example, in case of a normal distribution, the variance is set to a large value to maximise the entropy.

### Hierarchical priors: Automatic Relevance Determination

A particularly useful modelling technique is to make priors hierarchical by introducing new random variables with its corresponding hyperprior distributions and hyper parameters.
Hierarchical priors provide a certain degree of flexibility and make the posterior less sensitive to the prior. Besides, some choices of hierarchical priors are extremely helpful to introduce sparsity and to control model complexity, such as the Automatic Relevance Determination (ARD) prior [24], which is used in this project.

The ARD prior works as follows: consider a random variable $w$ that follows a normal distribution $\mathcal{N}\left(w \,|\, 0, \alpha^{-1}\right)$. Then, we introduce a hierarchy by treating $\alpha$ also as a random variable that follows a gamma distribution $\mathcal{G}\left(\alpha \,|\, a, b\right)$ with hyperparameters $a$ and $b$.
To understand the effect of the new setting, consider the case that the posterior estimate of the precision $\alpha$ is big (or equivalently, the variance is small). Then, the posterior distribution of $w$ will be sharply peaked at zero and the variable $w$ will be effectively not used by the model. On the contrary, if the posterior estimate of $\alpha$ is small, the posterior distribution of $w$ will have significant density at non-zero values and so the variable will be used by the model (see Figure 2.1). Therefore, during inference, redundant and/or irrelevant variables can be shut down automatically by driving its corresponding precision to infinity.

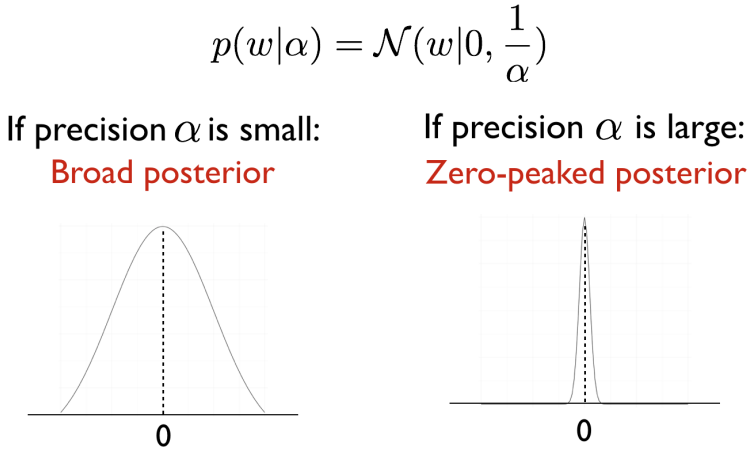



Figure 2.1: Schematic representation of the Automatic Relevance Determination prior

**On the inference algorithm**

A central task in Bayesian inference is the direct evaluation of the posterior distributions and/or the computation of expectations with respect to the posterior distributions. In most non-trivial models this is unfeasible and one has to resort to approximation schemes, which broadly fall into two classes: stochastic or deterministic [8].
The most common stochastic approximations are based on Markov Chain Monte Carlo techniques, which have the property that given infinite computational resources they would generate exact results. However, in practice, sampling approaches are computationally demanding and are difficult to use for large datasets On the other hand, deterministic approaches are based on analytical approximations to the posterior distribution by making specific asumptions regarding the nature of the approximation. As a result, deterministic approaches are usually much faster and scale to large applications, but they do not generate exact results. Well-known examples of determinisitic or analytical approaches are the Laplace approximation and the variational approximation.
In the Laplace approximation, the posterior distribution is approximated by a Gaussian at its mode. Obviously, this approach is effective in continuous unimodal distributions, but it breaks for complex multimodal posterior distributions [25]. On the other hand, the variational method replaces the true posterior by an approximating distribution which is then optimised to be as similar as possible to the true posterior distribution.

## 2.2 Variational Bayesian inference

### 2.2.1 Introduction

Broadly speaking, variational methods have their origins in the work on calculus of variations, which is concerned with the study of functionals or mappings that take one or more functions as input and return a scalar. For example, the Kullback-Leibler divergence is a statistic that takes two probability distributions as input and returns a scalar quantity, so it can be defined as a functional.
Similar to standard calculus, one can then define a functional derivative as how much the functional changes in response to infinitesimal changes to the input function. Variational inference involves the optimisation of functionals, which can be solved in a similar way than standard function optimisation problems [8]. Nevertheless, the input space of the quantity being optimised is not the set of real or complex numbers, but a set of functions.
One could solve the optimisation problem by exploring all possible functions to find the one that optimises the functional, but in order to remain in the tractable domain, one has to restrict the range of functions over which the optimisation is performed. Here is were the approximation comes up in variational methods [8].

### General variational learning

Consider a probabilistic model where all the observed variables are collectively denoted as $\mathbf{Y}$ and all the hidden variables (including parameters) are denoted by $\mathbf{X}$. Let's assume that the model is complex enough that the evaluation of the marginal log likelihood $\log p(\mathbf{Y})$ is intractable, since we have to marginalise out over all the hidden variables. However, the evaluation of the complete-data log likelihood $\log P(\mathbf{Y}, \mathbf{X})$ can be computed.
Using the sum rule of probability, we can write the marginal log likelihood as:

$$\log P(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{X}) d\mathbf{X}$$

Next, we introduce a new arbitrary distribution over the latent variables $q(\mathbf{X})$, called the variational distribution:

$$\log p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{X}) \frac{q(\mathbf{X})}{q(\mathbf{X})} d\mathbf{X}$$

One can then apply Jensen's inequality to the previous equation (because the log is a concave function):

$$\log p(\mathbf{Y}) = \log \left( \mathbb{E}_q \left[ \frac{p(\mathbf{Y}, \mathbf{X})}{q(\mathbf{X})} \right] \right) \tag{2.1}$$

$$\geq \mathbb{E}_q \left[ \log \frac{p(\mathbf{Y}, \mathbf{X})}{q(\mathbf{X})} \right] \tag{2.2}$$

$$\geq \mathcal{L}(\mathbf{X}) \tag{2.3}$$

Importantly, notice that the expectation is taken with respect to the new distribution $q(\mathbf{X})$.
Equation (2.1) is essentialy defining a functional that acts as a lower bound to the evidence or the marginal likelihood of the model, and so it is known as evidence lower bound (ELBO) in the literature. So why is the ELBO important? Recall that we are aiming at approximating the intractable posterior distribution with a tractable distribution. The link between the two is the following: notice that the inequality in Equation (2.1) turns into an equality if $q(\mathbf{X})$ is equal to the true posterior $p(\mathbf{X}|\mathbf{Y})$, so the closer the variational distribution is to the true posterior distribution, the closer the ELBO gets to the marginal likelihood. This can be clearly seen if we rewrite the ELBO as follows:

$$\mathcal{L} = \int q(\mathbf{X}) \Big( \log \frac{p(\mathbf{X}|\mathbf{Y})}{p(\mathbf{X})} + \log p(\mathbf{Y}) \Big) d\mathbf{X} \tag{2.4}$$

$$= \log p(\mathbf{Y}) - \mathrm{KL}(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})) \tag{2.5}$$

The ELBO is equal to the sum of the model evidence and the negative KL-divergence between the true posterior and the variational distribution. Therefore, increasing the ELBO is equivalent to decreasing the KL divergence between the two distributions. The following image summarises the general picture of variational learning:

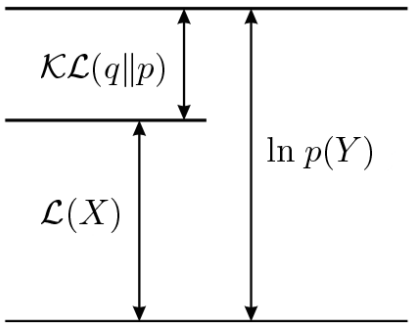



Figure 2.2: The quantity $\mathcal{L}(\mathbf{X})$ provides a lower bound on the true log marginal likelihood $\log p(\mathbf{Y})$, with the difference being given by the Kullback-Leibler divergence $\mathrm{KL}(q||p)$ between the approximating distribution $q(\mathbf{X})$ and the true posterior $p(\mathbf{X}|\mathbf{Y})$

In conclusion, variational learning involves optimising the functional $\mathcal{L}(\mathbf{X})$ with respect to the distribution $q(\mathbf{X})$. If we allow any possible choice of $q(\mathbf{X})$, then the maximum of the lower bound $\mathcal{L}(\mathbf{X})$ will occur when the KL divergence vanishes, which occurs when $q(\mathbf{X})$ equals the true posterior distribution $p(\mathbf{X}|\mathbf{Y})$.

Nevertheless, since the true posterior is intractable, this does not lead to any simplification of the problem. Instead, it is necessary to consider a restricted family of variational distributions that are tractable to compute and then seek the member of this family for which the KL divergence is minimised [6].

### Mean-field approximation

The most common type of variational Bayes, known as mean-field approach, assumes that the variational distribution factorises over $M$ disjoint groups of latent variables, that is, that all hidden variables (including parameters) are independent:

$$q(\mathbf{X}) = \prod_{i=1}^{M} q(\mathbf{x}_i)$$

Evidently, this family of distributions do not usually contain the true posterior because the hidden variables are dependent, but this is a key assumption to obtain an analytical inference scheme [6].
Importantly, notice that this is the only assumption regarding the variational distribution $q(\mathbf{X})$, no restriction is placed on its functional form.
Now, with the key assumption made, we need to find the distribution that maximises the lower bound $\mathcal{L}(\mathbf{X})$. It can be shown using calculus of variations [8], that the optimal distribution $\hat{q}_i$ for each variable $\mathbf{x}_i$, is the following:

$$\log \hat{q}_i(\mathbf{x}_i) = \mathbb{E}_{i \neq j}[\log p(\mathbf{Y}, \mathbf{X})] + \text{const} \tag{2.6}$$

where $\mathbb{E}_{i \neq j}$ denotes an expectation with respect to the $q$ distributions over all variables $\mathbf{x}_j$ except for $\mathbf{x}_i$.
In words, the log of the optimal distribution for variable $\mathbf{x}_i$ is obtained by considering the expectation of the log of the marginal likelihood with respect to all the other factors.
The additive constant is set by normalising the distribution $\hat{q}_i(\mathbf{z}_i)$:

$$\hat{q}_i(\mathbf{x}_i) = \frac{\exp(\mathbb{E}_{i \neq j}[\log p(\mathbf{Y}, \mathbf{X})])}{\int \exp(\mathbb{E}_{i \neq j}[\log p(\mathbf{Y}, \mathbf{X})]) d\mathbf{X}}$$

Nevertheless, in practice it is easier to work in terms of Equation (2.6) and infer the constant by inspection, as the rest of the expression can usually be recognised as being a known type of distribution. This becomes clear in Section 8.2 of the appendix, where the variational equations for the scGFA model are derived.

### Variational Bayesian Expectation Maximisation algorithm

In the previous section we obtained the general expression 2.6 which yields the set of variational distributions that maximise the lower bound of the log marginal likelihood (subject to the factorisation constraint). Or equivalently, the set of distributions that minimise the KL divergence between the $q(\mathbf{X})$ distribution and the true posterior $p(\mathbf{X})$.
However, notice that for a given variable $\mathbf{x}_i$, the expectation on the right-hand side is taken with respect to the other factors' variational distribution $q_j(\mathbf{x}_j)$ for $j \neq i$. Therefore, there are circular dependencies between the different equations and there is no analytical solution for Equation (2.6), which naturally suggests an iterative algorithm similar to the Expectation Maximisation (EM) algorithm [12].

To illustrate the similarity with the EM algorithm, let's split all unobserved variables $\mathbf{X}$ into the latent variables $\mathbf{Z}$ and the parameters $\boldsymbol{\theta}$. Note that this is just for comparison purposes, as the algorithm treats all variables equivalently.
In the E-step, we update the moments and parameters of the variational distribution of the latent variables $q(\mathbf{Z})$ using the current estimates of the variational distributions of the parameters $q(\boldsymbol{\theta})$. Subsequently, in the M-step, the moments and parameters of $q(\boldsymbol{\theta})$ are updated while keeping $q(\mathbf{Z})$ fixed [6]. The algorithm is stopped when the change in the ELBO is small enough.
Due to the similarity with the EM algorithm, this iterative procedure is usually called Variational Bayesian Expectation Maximisation (VBEM) algorithm (Figure 2.3).

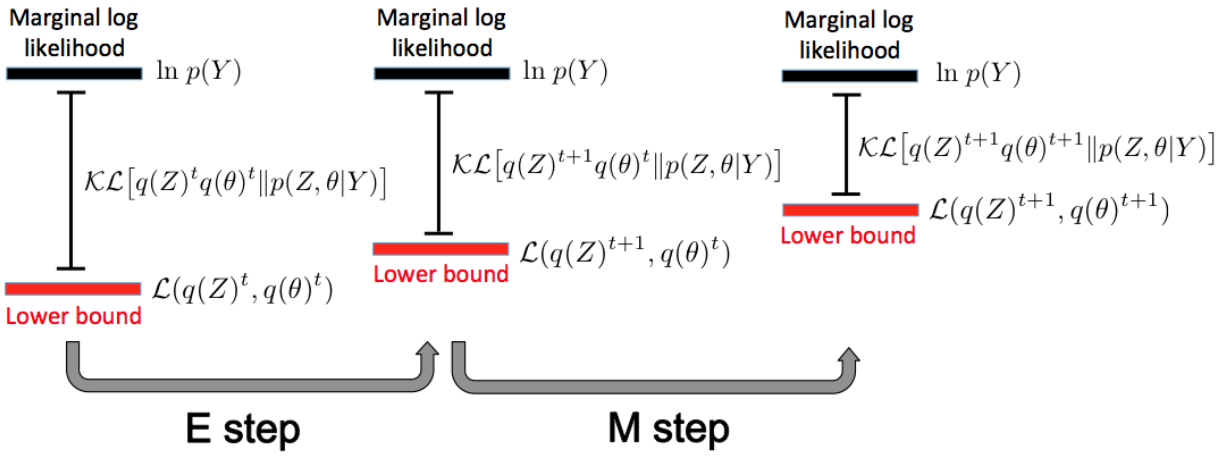



Figure 2.3: The variational Bayesian EM (VBEM) algorithm. In the VBE step, the variational posterior over hidden variables $q(\mathbf{Z})$ is updated. In the VBM step, the variational posterior over parameters $q(\boldsymbol{\theta})$ is updated. Each step is guaranteed to increase the lower bound on the marginal likelihood. Note that the exact log marginal likelihood is a fixed quantity, the algorithm only changes its corresponding lower bound. Figure adapted from [6].

It is worth discussing the main differences with respect to the EM algorithm. Both procedures are numerically similar, but the quantities being computed are very different. First, VBEM maximises a lower bound to the marginal likelihood, whereas EM maximises the likelihood function. Second, EM computes point estimates of the parameters, whereas VBEM is computing the expectations and parameters of entire distributions. Third, EM is computing parameters of the true probability distribution of the model, but VBEM does not work with the true distributions but with approximating distributions [6].

## 2.3 Latent variable models

### 2.3.1 Motivation

With the exponential growth in the use of high-throughput genomics, biology is currently facing the classical big data problem: datasets are so large and complex that traditional processing applications are adequate anymore. In particular, high dimensional biological data is a specially hard data type to analyse for multiple reasons. First, the observations are the result of complex non-linear relationships driven not only by biological processes but also by technical sources of variation. Second, the number of observations is usually much lower than the dimensionality of the datasets (large $p$ small $n$ problem). Third, many dimensions are correlated so the traditional statistical independency assumption completely breaks down. Therefore, in order to extract meaningful and rigurous patterns from biological data, the use of powerful statistical methods is mandatory. A succesful approach on traditional datasets is the use of latent variable models (LVMs), which aim to reduce the dimensionality of the dataset into a small set of latent variables which are easier to interpret and visualise [8].

### 2.3.2 General mathematical formulation

More formally, given a dataset $\mathbf{Y}$ of $N$ samples and $D$ dimensions, LVMs attempt to explain correlations between the $D$ features by means of a potentially smaller set of $K$ unobserved and less correlated variables. Usually, LVMs assume that the relationship between observed and hidden variables is linear. Thus, for a given sample $n$, the vector of observations $\mathbf{y}_n = [y_{n1}, y_{n2}, ..., y_{nD}]$ is generated by a weighted sum of the latent variables $\mathbf{z}_n = [z_{n1}, z_{n2}, ..., z_{nK}]$:

$$y_{nd} = \sum_{k=1}^{K} w_{dk} z_{n,k} + \mu_d$$

where $w_{dk}$ is the weight or loading and $\mu_d$ is the mean of feature $d$ which is sometimes omitted for mathematical simplicity. To deal with noisy data, most models also include a random term in the equation

which captures the variation not explained by the model.

$$y_{nd} = \sum_{k=1}^{K} w_{dk} z_{n,k} + \mu_d + \epsilon_{nd}$$

Notice that LVMs differ to multiple linear regression in that the explanatory variables are unobserved. To complete the specification of a LVM, one has to turn the formulation above into a probabilistic framework by deciding which variables are set to fixed parameters and which ones are set to be random. Also, in order to do probabilistic inference, assumptions must be made regarding the distributions of the random variables and also the inference framework to work with. There is a huge flexibility in LVMs, and one has to find the most appropriate design depending on the problem and the dataset in question [9]. The following sections describe some of the original and still most popular LVMs that laid the groundwork for the scGFA model.

### 2.3.3 Probabilistic Principal component Analysis

Principal Component Analysis (PCA) is one of the most widely used techniques for data analysis. It reduces a high-dimensional dataset of correlated observations into a lower dimensional dataset of uncorrelated variables called principal components, which (hopefully) reveal the internal structure of the data Figure 2.4.

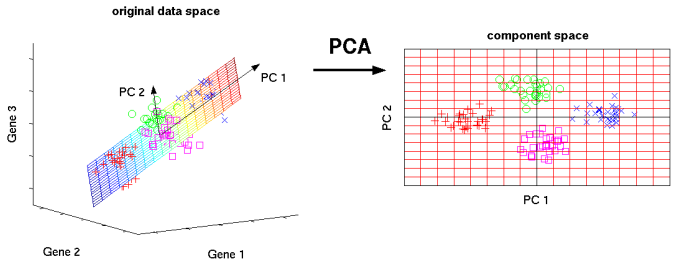



Figure 2.4: Representation of traditional principal component analysis as a dimensionality reduction technique.

However, PCA is not a true LVM, since it lacks an underlying probabilistic formulation. Instead, the most common derivation of PCA is in terms of a standarised linear projection which maximises the variance in the projected space [8].
The first probabilistic model of PCA (pPCA) was proposed by Tipping and Bishop in terms of a Gaussian latent variable model [34]. In particular, as described in Section 2.3.2 the matrix of observations $\mathbf{Y}$ is decomposed as:

$$\mathbf{Y} = \mathbf{WZ} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{2.7}$$

where the weights $\mathbf{W}$ and the means $\boldsymbol{\mu}$ are assumed to be fixed parameters whereas the noise $\boldsymbol{\epsilon}$ and the latent variables $\mathbf{Z}$ (the principal components) are assumed to be normally distributed random variables with the following characteristics:

$$\mathbf{z} \approx \mathcal{N}(\mathbf{z} \,|\, \mathbf{0}, \mathbf{I}) \tag{2.8}$$

$$\boldsymbol{\epsilon} \approx \mathcal{N}\left(\boldsymbol{\epsilon} \,|\, \mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I}\right) \tag{2.9}$$

Note that the covariance of the noise is assumed to be a diagonal matrix with the same value in all its diagonal, which implies that the noise of the features is independent but is restricted to have the same magnitude.
Since both the factors and the noise are normally distributed, the observed variables $\mathbf{Y}$ also follow a multivariate Gaussian distribution, which defines the likelihood of the model:

$$p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \sigma) = \mathcal{N}\left(\mathbf{Y} \,|\, \mathbf{WZ}, \sigma^2\mathbf{I}\right)$$

The following image shows the corresponding graphical model:

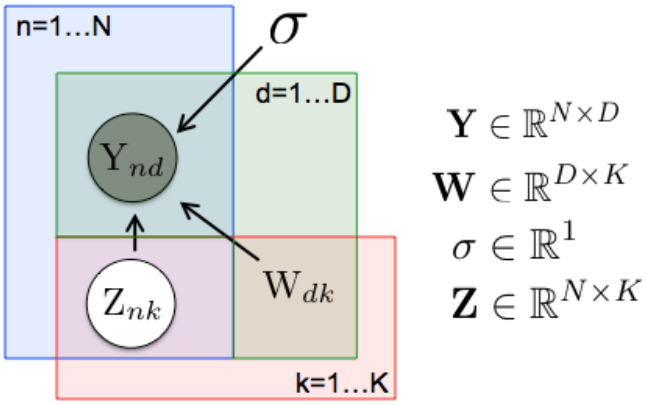



Figure 2.5: Graphical model representation of probabilistic principal component analysis. Uncircled variables are non-random parameters. Both unobserved variables and parameters are infered by the model. The mean term $\boldsymbol{\mu}$ was omitted for simplicity.

An important consideration of this model is that the number of factors $K$ has to be defined *a priori*, and it is usually optimised using cross-validation techniques. However, this is sometimes computationally expensive and is a clear drawback of the method. As described in Section 2.3.7, this problem is partially solved under the Bayesian framework.

In contrast to most latent variable models, inference on the parameters $\mathbf{W}$, $\boldsymbol{\mu}$ and $\sigma^2$ can be performed in closed form due to the normality assumptions. Interestingly, in the original report it is proved that the maximum likelihood solution lies in the same subspace that the solution of traditional PCA [34]. Thus, if the solution is the same one can what is the main advantage of a probabilistic formulation: first, it is useful for model selection as different settings can be compared under a rigurous probabilistic framework. Second, rather than its performance the main benefit is that it provides the potential to extend the model to more complex settings by adding hierarchies and relaxing assumptions. As a matter of fact, as described in section 2.4.3, the core of the first multi-view group factor analysis model is essentialy a pPCA model.

### 2.3.4 Factor Analysis

Factor analysis (FA) has historically been the most used LVM and is a generalisation of the pPCA model [17]. The main differences is that FA makes the following assumption about the noise:

$$\boldsymbol{\epsilon} \approx \mathcal{N}(\boldsymbol{\epsilon} \,|\, \mathbf{0}, \boldsymbol{\Phi})$$

where the covariance $\boldsymbol{\Phi}$ is assumed to be a diagonal matrix with potentially different elements on each entry. Therefore, this implies that the noise is allowed to have different magnitude across dimensions. The likelihood can then be writen as follows:

$$p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \sigma) = \mathcal{N}(\mathbf{Y} \,|\, \mathbf{WZ}, \boldsymbol{\Phi})$$

The graphical model associated with the FA model is shown in Figure 2.6.
The motivation and key assumption of the factor analysis model is that by constraining the error covariance $\boldsymbol{\Phi}$ to be a diagonal matrix whose elements $\phi d, d$ are estimated from the data, the observed variables $\mathbf{y}_n$ are conditionally independent given the values of the latent variables $\mathbf{z}_n$. The consequence is that the latent variables will explain covariation between observations whereas $\boldsymbol{\epsilon}_n$ will explain variation unique to a particular $\mathbf{y}_n$. Thus, in contrast to PCA, FA makes a clear distinction between variance and covariance [34].
The main drawback of FA is that the solution is not available in closed form, but one can do maximum-likelihood inference on the parameters $\mathbf{W}$, $\boldsymbol{\mu}$ and $\Phi$ using the EM algorithm [12].

### 2.3.5 Rotation Invariance

A very important consequence of the definition of most LVMs is their unidentifiability due to rotational and scaling invariance.
This problem arises from the fact that one can intercalate any full-rank matrix $\mathbf{R} \in \mathbb{R}^{K \times K}$ in the likelihood so that: $\mathbf{WX} = \mathbf{WRR^{-1}X}$, and hence the solution is defined only up to a rank-preserving linear

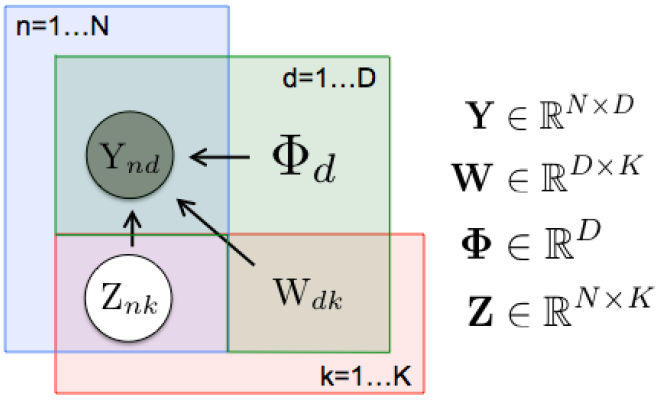



Figure 2.6: Graphical model for traditional factor analysis. Grey-circled variables are observed data whereas white-circled variables are unobserved. Uncircled variables are non-random parameters. Both unobserved variables and parameters are infered by the model. The mean term $\boldsymbol{\mu}$ was omitted for simplicity. Note that the main difference with respect to the pPCA model is that there is a different noise distribution for each feature $d$.

transformation.
This is important when analysing the model because it implies that the latent variables between different realisations of the same model are not directly comparable, as they might represent rotated and/or scaled versions.

### 2.3.6 What is a latent variable?

The concept of a latent variable and how it should be interpreted is sometimes misunderstood and difficult to grasp, in particular within the biological domain.
To illustrate the idea, consider a cellular population that is undergoing a differentiation process (Figure 2.7 top). We can define a variable called differentiation state that is going to measure, in an arbitrary scale, at which point each cell is located along the differentiation path. For example, pluripotent cells can have a score of zero and fully differentiated cells have a score of ten (in arbitrary units). Importantly, notice two things: first, we are using one single dimension to characterise the cellular state. Second, the new variable is simply a plausible artificial construction that cannot be directly measured, so we call it a hidden variable. Instead, one can obtain observations such as the gene expression profile of the cell (Figure 2.7 bottom). Nevertheless, if the hidden variable (the differentiation process) is an important driver of the gene expression, then we expect most of the genes involved in the pathway to change proportionally with the differentiation score. For instance, pluripotency genes will be highly expressed in pluripotent cells and lowly expressed in fully mature cells, whereas differentiation genes will be highly expressed in mature cells and lowly expressed in pluripotent cells. Consequently, all these group of genes should show high levels of covariation. In fact, if the correlation is large enough, one could describe the the entire dataset using a single latent variable. But in most practical cases there are multiple unknown drivers of variation, so finding them is the aim of latent variable models.

### 2.3.7 Variational Bayesian inference in latent variable models

In the previous section we described the probabilistic formulation of the two most used latent variable models for continuous variables: factor analysis and probabilistic principal components analysis. However, we have ignored the inference procedure.
The most common approach for learning in graphical models is maximum likelihood inference, but it has important caveats in LVMs. First, the choice of the right number of latent variables is usually non-trivial. Second, as there is no penalisation for model complexity, when $N$ is smaller than $D$ the solutions are usually overfitted.
These central two problems can be partially solved when reformulating the LVM in a Bayesian framework and addressing a proper choice of the prior distributions. To illustrate the point, rather than providing theoretical justification, we reproduce here the variational Bayesian treatment of the pPCA model [7], and we show how it yields a better solution than traditional PCA or pPCA.

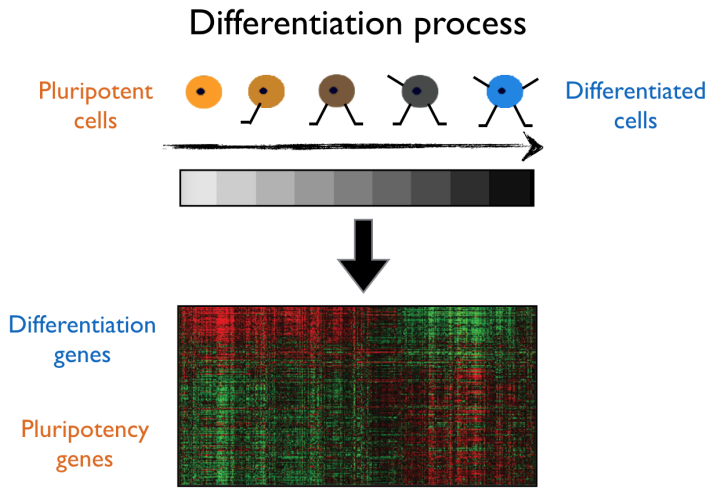



Figure 2.7: Illustration of a latent variable associated with a differentiation pathway. A single axis of variation is driving the variation of multiple observed variables (gene expression)

### Variational Bayesian probabilistic principal component analysis

We start from the probabilistic model of PCA (Figure 2.5). To do a fully Bayesian treatment of the model, we introduce prior distributions to all unobserved variables of the model, including parameters:

$$P(\mathbf{Z}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{z}_{:,n} \,|\, \mathbf{0}_k, \mathbf{I}_k\right) \qquad P(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \,|\, a_0^{\alpha}, b_0^{\alpha}\right)$$

$$P(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_{:,k} \,|\, 0, \frac{1}{\alpha_k}\mathbf{I}_D\right) \qquad P(\boldsymbol{\tau}) = \mathcal{G}\left(\tau \,|\, a_0^{\tau}, b_0^{\tau}\right)$$

where $\boldsymbol{\tau}$ is the precision (the inverse of the variance $\boldsymbol{\sigma}^2$) of the normally distributed noise term $\boldsymbol{\epsilon}$. Note the use of a hierarchical ARD prior as described in section 2.1.3. This is a crucial part of the model, since it automatically learns the appropriate dimensionality of the latent space as part of the process of Bayesian inference, and it therefore solves the latent variable selection problem mentioned above.
Following Equation (2.7), the likelihood is:

$$p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \tau) = \prod_{n=1}^{N} \prod_{d=1}^{D} \mathcal{N}\left(y_{nd} \,|\, \mathbf{w}_{d,:}\mathbf{z}_{n,:}, \tau\right) \tag{2.10}$$

The graphical model associated with the Bayesian pPCA model is represented in Figure 2.8.

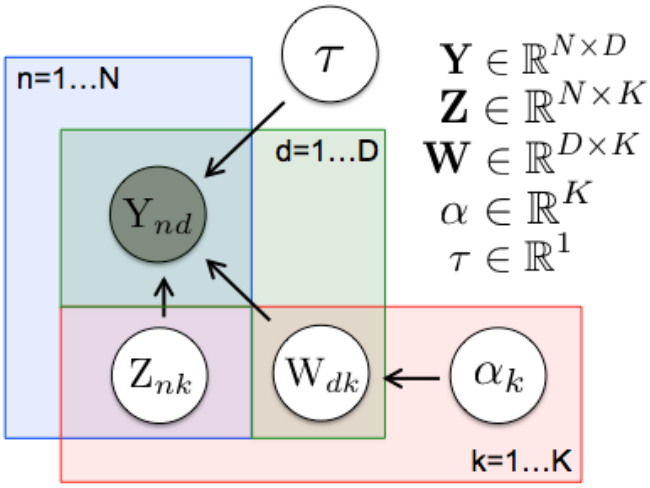



Figure 2.8: Graphical model representation of a fully Bayesian principal component analysis. Notice that in contrast to pPCA, all parameters as well as the latent variables are treated as fully Bayesian random variables.

Following Bayes formula, the corresponding posterior distributions $p(\mathbf{W}, \mathbf{Z}, \tau, \boldsymbol{\alpha}|\mathbf{Y})$ are obtained by multiplying the priors $p(\mathbf{W}, \mathbf{Z}, \tau, \boldsymbol{\alpha})$ by the likelihood $p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \tau)$. However, the solution is analytically intractable, so here the approximate mean-field variational approach (see Section 2.2) was adopted.
In short, a fully factorised variational distribution $q$ was introduced for each of the unobserved variables:

$$q(\mathbf{W}, \sigma, \mathbf{Z}|\boldsymbol{\alpha}|\mathbf{Y}) = q(\mathbf{W})q(\sigma)q(\mathbf{Z})q(\boldsymbol{\alpha})$$

Subsequently, Equation (2.6) is used to obtain the form of all the variational $q$ distributions as well as the update equations for the VBEM algorithm. The details on the derivation can be found in [8], here we only reproduce the final results:

$$q(\mathbf{Z}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{z}_n \,|\, \mathbf{m}_z^n, \Sigma_z\right) \qquad q(\mathbf{W}) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{w}_{d,:} \,|\, \mathbf{m}_w^d, \boldsymbol{\Sigma}_w\right)$$
$$q(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \mathcal{G}\left(\boldsymbol{\alpha}_k \,|\, \tilde{a}_k^\alpha, \tilde{b}_k^\alpha\right) \qquad q(\tau) = \mathcal{G}\left(\tau \,|\, \tilde{a}^\tau, \tilde{b}^\tau\right)$$

Importantly, notice that the form of the variational distributions is the the same as the corresponding prior distributions even though this was not an assumption. The parameters of the variational distributions are updated as follows:

$$\mathbf{m}_z^n = \langle\tau\rangle \boldsymbol{\Sigma}_z \langle\mathbf{W}^t\rangle(\mathbf{y}_n) \qquad \tilde{a}_k^\alpha = a_0^\alpha + D/2$$
$$\boldsymbol{\Sigma}_z = (\mathbf{I} + \langle\tau\rangle\langle\mathbf{W}^t\mathbf{W}\rangle)^{-1} \qquad \tilde{b}_k^\alpha = b_0^\alpha + \frac{1}{2}\langle\|\mathbf{w}_{:,k}\|\rangle$$
$$\mathbf{m}_w^d = \langle\tau\rangle \boldsymbol{\Sigma}_w \sum_{n=1}^{N} \langle\mathbf{x}_n\rangle y_{nd} \qquad \tilde{a}^\tau = a_0^\tau + ND/2$$
$$\boldsymbol{\Sigma}_w = (\mathrm{diag}(\langle\alpha\rangle) + \tau \sum_{n=1}^{N} \langle\mathbf{z}_n\mathbf{z}_n^t\rangle)^{-1} \qquad \tilde{b}^\tau = b_0^\tau + \frac{1}{2}\sum_{n=1}^{N} \langle\|(\mathbf{y}_n - \mathbf{W}\mathbf{z}_n)\|\rangle$$

where the expectations are taken with respect to the $q$ distributions.

To illustrate the performance of the Bayesian and maximum likelihood approach, vbPCA and pPCA were trained with a simulated dataset of $N = 100$, $d = 10$, $k = 9$ having standard deviations of $(5, 4, 3, 2)$ along four orthogonal directions and a standard deviation of 1 in the remaining six directions [7]. The comparison of fitting the two models is shown as a Hinton diagram of the weight matrix $\mathbf{W}$ in the following figure:

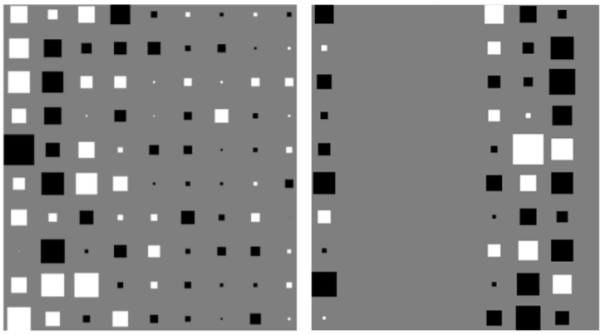


Figure 2.9: Comparison of the hintons diagram of the elements of $\mathbf{W}$ for the maximum likelihood PCA model (left) and the variational Bayes PCA model (right). The Data set of 100 samples and 10 dimensions was generated using Equation (2.10) having different and large variances in 4 independent directions and a common, smaller variance in the remaining 6. Reprinted from [7].

Both methods find the four dominant latent components, but the maximum likelihood solution is much more dense, with nonzero entries for essentialy all components, which makes interpretation harder. In contrast, the vbPCA solution finds a much more sparse solution where only the true latent variables have nonzero entries.

## 2.4 Multi-view learning

### 2.4.1 Motivation

Factor analysis and probabilistic principal component analysis are called single-view learning models because they take as input a single data matrix. However, in some occasions, data is collected from diverse domains or obtained from various feature extractors and so they can exhibit heterogeneous properties [10]. For instance, the phenotypic state of cells has usually been characterised using exploratory methods on the gene expression matrix. Nevertheless, the phenotype results from the combination of multiple layers of biological information such as the epigenetic and genetic background, protein levels and lipid composition. Each of these layers of information (or views) can be analysed separately using single-view learning methods, but the current challenge is the integration of all the different views into a single model that is able to disentangle the variation unique to a single view and the variation shared between two or more views. This is refered to as the multi-view learning problem [10, 38].

Traditionally, the multi-view problem has been approached by concatenating the different datasets into a single large matrix, which was subsequently analysed using conventional machine learning algorithms. However, this concatenation should be avoided for several reasons. First, the scale of noise is usually different between views and one view might end up overepresented in the solution. Second, this concatenation leads to a very wide matrix that causes overfitting in the case of a small number of samples. Third, it is hard to distinguish from which view the signal is coming from and whether it is shared by multiple views or it is unique to a single view.

Multi-view learning approaches can be roughly classified into three classes: co-training, multiple kernel learning and latent variable models. The scGFA model proposed here is part of the latter class, so in the next section we briefly describe some of the relevant approaches proposed before. A comprehensive review can be found in [10].

### 2.4.2 Canonical correlation analysis

The simplest multi-view learning algorithm based on a latent variable model is Canonical Correlation Analysis (CCA), which aims to find linear components that capture correlations between two datasets [16].
Mathematically it is described as follows: given two data matrices $\mathbf{Y}_1 \in \mathbb{R}^{N \times D_1}$ and $\mathbf{Y}_2 \in \mathbb{R}^{N \times D_1}$ CCA finds a set of linear combinations $\mathbf{U} \in \mathbb{R}^{D_1 \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_2 \times K}$ from $\mathbf{Y}_1$ and $\mathbf{Y}_2$ which have maximum correlation with each other. Similar to PCA and FA, the linear components are constraint to be independent. Therefore, the first pair of canonical variables $\mathbf{u}_1$ and $\mathbf{v}_1$ contain the linear combination of variables that have maximal correlation, followed by the second canonical pair $\mathbf{u}_2$ and $\mathbf{v}_2$, and so on.

Often CCA has been regarded as the two-view analogue of PCA, as they share critical features such as linearity and analytical solution. Furthermore, the solution to CCA is also found solving an eigenvalue problem, which yields a solution with a single global optimum but it is known to be overfitted for datasets where $D > N$ [16]. Likewise, CCA suffers from the same pitfalls as traditional PCA, such as the problem in selecting an appropriate dimensionality of the latent subspace, difficult interpretability and the absence of a model with probabilistic formulation. Hence, effort was made to develop probabilistic versions of CCA in terms of graphical models which are amenable to both maximum likelihood and Bayesian inference.

A probabilistic model was originally proposed by Back and Jordan [5] as an extension of the probabilistic PCA model (see Section 2.3.3). A fully Bayesian treatment was published by Wang [39] and the corresponding grahpical model is shown in Figure 2.10.

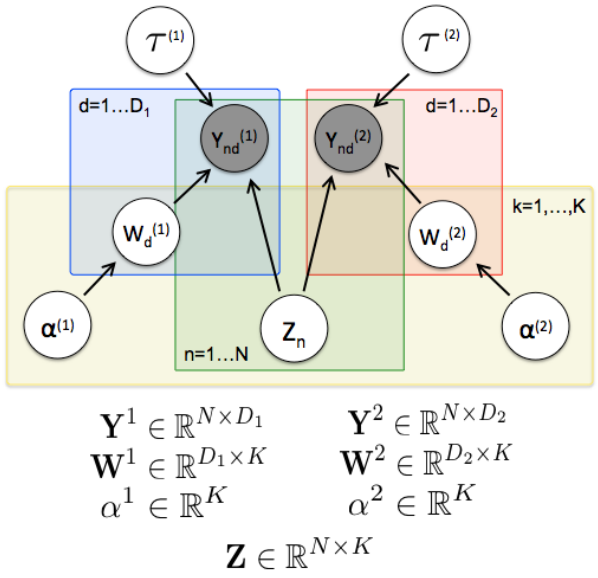



Figure 2.10: Graphical model for Bayesian Canonical Correlation Analysis. Grey-circled nodes represent observed random variables whereas white-circled nodes represent unobserved variables. The model is an extension of Bayesian PCA to two views, note that each vertical half of the model is equivalent to Figure 2.8

Notice that the observations $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ are generated from the same set of latent variables $\mathbf{Z}$, but this time the factors are shared by the two views. Therefore, they will only focus on modelling variation associated with correlated groups of variables between the two datasets.
Similarly to the Bayesian PCA model, the Automatic Relevance Determination prior (see Section 2.1.3) is able to shut off unimportant canonical variables, thereby automatically finding the correct dimension of the latent subspace. Again, this yields a much more sparse and generalisable solution than traditional CCA (Figure 2.11).

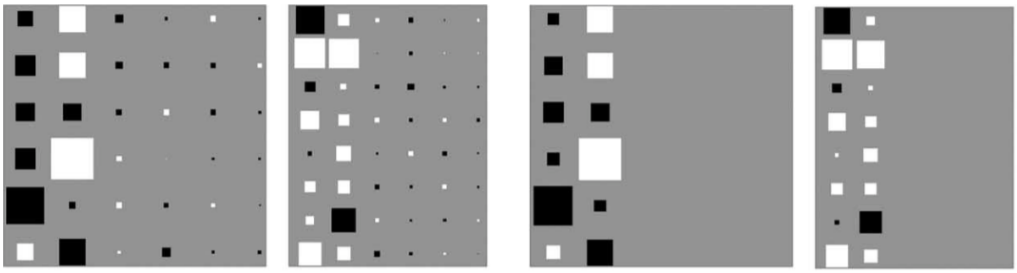


Figure 2.11: Comparison of the hintons diagram of $\mathbf{W}^1$ and $\mathbf{W}^2$ for the maximum likelihood CCA model (two left plots) and the variational bayes CCA model (two right plots). Reprinted from [39].

### 2.4.3 Group Factor Analysis

Canonical Correlation Analysis is restricted to two views. Models that can handle an arbitrary number of views have traditionally been disregarded because of their computational complexity and the inexistence of such type of data. However, with the big data revolution, multi-view learning approaches are gaining popularity and there is more demand to find approaches that deal with an arbitrary number of datasets.

Virtanen et al [38] published a very successful approach to solve the multi-view problem as a latent variable model called Group Factor Analysis (GFA). The approach is a natural generalisation of Bayesian CCA with the latent variables being able to capture variation shared either by a single view, a subset of views or all views at the same time.
More formally, the GFA solution is as follows: given a collection of $M$ views $\mathbf{Y}_1, \cdots, \mathbf{Y}_M$ with dimensionalities $D_1, \cdots D_M$, the task is to find $K$ latent factors that describe all views $\mathbf{Y}_m$ at the same time, with the condition that the relationships between the views have to be separated from the relationships within the views.

Similar to traditional latent variable models, GFA adopts the following factorisation of the data $\mathbf{Y}$:

$$\mathbf{Y} \approx \mathbf{Z}\mathbf{W}^t + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ is the column-wise concatenation of the views, $\mathbf{W} \in \mathbb{R}^{D \times K}$ $(D = D_1 + \cdots + D_m)$ is the weight matrix, $\mathbf{Z} \in \mathbb{R}^{N \times K}$ is the factor matrix and $\boldsymbol{\epsilon} \in \mathbb{R}^{D \times D}$ is a diagonal noise covariance matrix $\boldsymbol{\Sigma}$ with diagonal $[\sigma_1^2, \cdots, \sigma_M^2]$ where $\sigma_m^2$ has been repeated $D_m$ times [38]. That is, every dimension within the same view has the same variance, but the views may have different variances. Thus, in the simple case where there is just one view, the GFA model reduces to the Bayesian PCA model (see Section 2.3.7).
In principle, the model so far is equivalent to Bayesian PCA with the concatenated data set. The key difference, however, is a group-wise sparsity structure imposed on $\mathbf{W}$ using the Automatic Relevance Determination prior (see Section 2.1.3):

$$p(\mathbf{W}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{d=1}^{D_m} \mathcal{N}\left(w_{d,k}^m \,|\, 0, \frac{1}{\alpha_k^m}\right) \tag{2.11}$$

$$P(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \tag{2.12}$$

In the single-view Bayesian PCA model, the ARD prior has a factor-wise sparsity structure, so $\boldsymbol{\alpha} \in \mathbb{R}^K$ shuts down factors inactive in the single view by driving their corresponding $\alpha_k$ to infinity. In contrast, in the multi-view GFA model, The group-wies ARD prior $\boldsymbol{\alpha} \in \mathbb{R}^{K \times M}$ makes factors inactive for particular views. That is, factor $k$ might be active in view $m$ but not in another view $n$.

The generative model is completed by assuming Gaussian latent variables that are *a priori* independent and a Gamma prior for the inverse variances $\sigma_m^{-2}$. These priors are equivalent to Bayesian PCA. The full likelihood is hence given by

$$p(\mathbf{Y}, \mathbf{W}, \boldsymbol{\Sigma}, \mathbf{Z}) = p(\mathbf{W}|\boldsymbol{\alpha})p(\mathbf{Z})p(\boldsymbol{\Sigma})p(\boldsymbol{\alpha})p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \boldsymbol{\Sigma})$$

The following figure illustrates the GFA approach for three data sets:

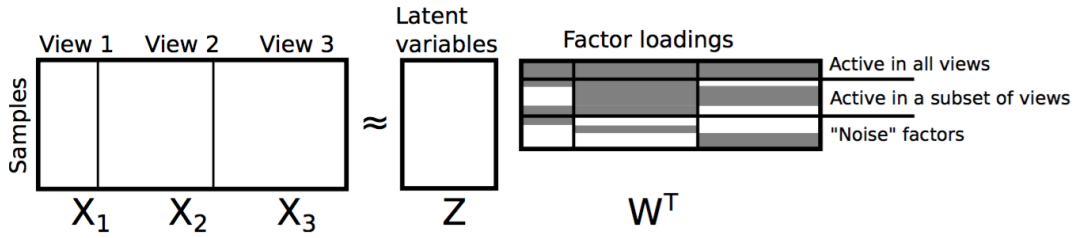



Figure 2.12: Illustration of the group factor analysis for three views. The feature-wise concatenation of the data sets $\mathbf{Y}_i$ is factorised as a product of the latent variables $\mathbf{Z}$ and factor loadings $\mathbf{W}$. The factor loadings are group-wise sparse, so that each factor is active (gray shading) only in some subset of views, or all of them. Reprinted from [38].

# Chapter 3

# Single cell Group Factor Analysis

## 3.1 Problem definition

As a more general approach to traditional single-view learning, the aim of multi-view learning is to identify patterns not only within a data set but also between different data sets (see Section 2.4.1).
To solve the multi-view learning problem, here we adopted the latent variable model approach. Hence, given a collection of $M$ views $\mathbf{Y}_1, \cdots, \mathbf{Y}_M$ with $N$ co-ocurrent samples and dimensionalities $D_1, \cdots, D_M$, the task is to find $K$ latent factors that describe the unique variability for each view as well as the shared variability between the views. Furthermore, the model has to be able to deal with the unique features of noisy single-cell sequencing data, such as highly variable variance between genes, many missing values and the zero-inflation due to the presence of dropout effects [27].

Taking the above requirements into account, an optimal solution for the single-cell multi-view learning problem has to meet the following requirements:

- The model has to be probabilistic, amenable to both maximum likelihood and Bayesian inference.
- The infered weight matrix has to be sparse, with as many zeros as possible to facilitate interpretation.
- The latent variables should be easily identified as either being unique to a single data set or shared between data sets.
- The model should be able to learn the mean and the noise values for each dimension separately.
- The model needs a rigurous way to cope with missing values and zero-inflated expression data.
- Inference has to be computationally feasible, preferibly linear with respect to the number of samples, the number of views and the number of dimensions.

Here we propose a model called single-cell Group Factor Analysis (scGFA), a Bayesian latent variable model based on the group factor analysis framework [38] that we extended by relaxing some assumptions in order to deal with the unique features of noisy single-cell sequencing data.

## 3.2 Model definition

Similar to traditional single-view latent variable models, we assume the following bilinear decomposition for the data matrix $\mathbf{Y}^m$ via a joint $K$-dimensional space:

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \mu^m + \epsilon^m$$

where the weight matrix $\mathbf{W}^m \in \mathbb{R}^{D_m \times K}$ contains the projection vectors, $\boldsymbol{\mu}^m \in \mathbb{R}^{D_m}$ the feature-wise means, $\mathbf{Z} \in \mathbb{R}^{N \times K}$ the latent variables and $\boldsymbol{\epsilon}^m \in \mathbb{R}^{D_m}$ the noise.
Importantly, notice that the same latent variable matrix is used to factorise each view (do not have index $m$), but the other variables are specific for each view. This is similar to the canonical correlation approach (Section 2.4.2), but here we define a more general case where $\mathbf{Z}$ is not only able to capture shared variation between views but also unique variability of a single view. The rationality is that we want to disentangle all sources of variation, including unique and shared ones.
The following image illustrates the factorisation:

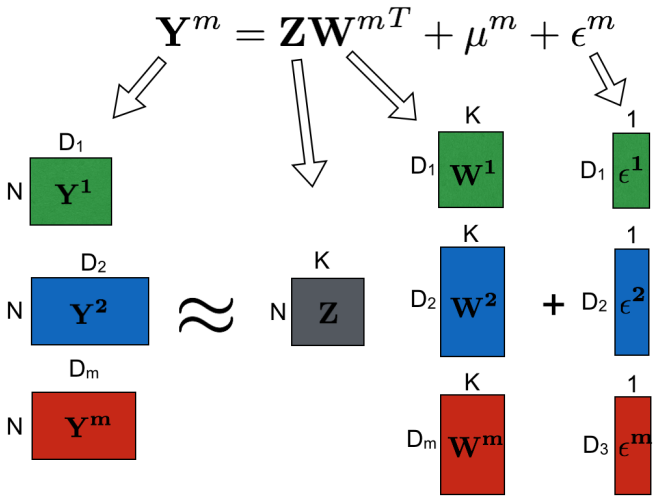



Figure 3.1: Linear decomposition adopted by the scGFA model. The mean term $\boldsymbol{\mu}$ is omited for simplicity

To do inference we adopt the variational Bayesian approach (see Section 2.2) and we perform a fully Bayesian treatment on all variables. Hence, we introduce prior distributions for each variable, including parameters, which are described in the following sections.

## Latent variables

In order to obtain non redundant factors, traditional LVMs such as FA or PCA enforce orthogonality constraints on the latent variables. In the Bayesian framework we can encode this information by setting a diagonal matrix in the prior distribution of the latent variables. Therefore, conventionally, we assume an isotropic gaussian distribution:

$$P(\mathbf{Z}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{z}_{:,n} \,|\, \mathbf{0}_k, \mathbf{I}_k\right)$$

Importantly, the posterior distributions can contain correlations, but they are minimised by setting the independency as a prior belief.

## Weight matrix

The prior distribution on the weight matrix is the most important part of the model, as we have to introduce a group-wise sparsity constraint that allows the model to shut down latent variables that are not explaining variation in particular views. Ideally, we would like to obtain a binary matrix $\mathbf{F} \in \mathbb{R}^{M \times K}$ where $f_{m,k}$ is 1 if factor $k$ is active in view $m$ and 0 otherwise. Such a discrete matrix is not easily incorporated into a continuous model, but there are approaches to approximate it.
Here we adopt the group-wise Automatic Relevance Determination prior proposed in the original group factor analysis model [38]:

$$P(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}^m_{:,k} \,|\, 0, \frac{1}{\alpha^m_k}\mathbf{I}\right) \tag{3.1}$$

$$= \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{d=1}^{D_m} \mathcal{N}\left(w^m_{d,k} \,|\, 0, \frac{1}{\alpha^m_k}\right) \tag{3.2}$$

$$P(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}\left(\alpha^m_k \,|\, a^\alpha_0, b^\alpha_0\right) \tag{3.3}$$

For simplicity and to keep the notation uncluttered, we set the hyperparameters from the Gamma distribution to be equal for all values of $\alpha$. Also, we set both hyperparameters $a^\alpha_0$ and $b^\alpha_0$ to very small values $(10^{-3})$ in order to obtain uninformative priors.

In summary, each loading follows a zero-mean gaussian distribution with precision $\alpha^m_k$. Importantly, notice that each column of $\mathbf{W}^m$ is associated with a latent variable $k$ and all values within the column share

the same precision. Therefore, if the factor $k$ is not relevant to model the variability in view $m$, the corresponding $\alpha_m^k$ is driven to infinity during inference and that direction in latent space is effectively switched off (the entire column goes to zero).
Hence, the ARD prior yields a matrix $\boldsymbol{\alpha} \in \mathbb{R}^{M \times K}$ that acts as a continuous approximation to $\mathbf{F}$ and defines three different types of factors:

- Factors that explain variation in a single data set, also called unique factors: the corresponding $\boldsymbol{\alpha}_k$ vector contains very large entries for all views except one, which has a small value.
- Factors that explain variation in a subset of data sets, also called partially shared factors: the corresponding $\boldsymbol{\alpha}_k$ vector contains very large entries for the inactive views and small entries for the active views
- Factors that explain variation in all data sets, also called fully shared factors: the corresponding $\boldsymbol{\alpha}_k$ vector contains small values for the active views

Importantly, the type of factor is learned automatically, without needing to define beforehand all combination of factor types, whose number grow exponentially in the number of views. Figure 3.2 shows an hypothetical $\boldsymbol{\alpha}$ matrix with a fully shared factor, two partially shared factors and three unique factors.

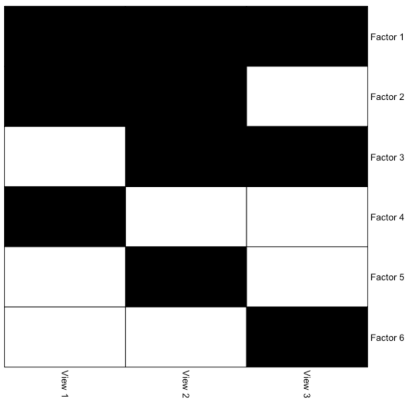



Figure 3.2: Plot of an hypothetical $\boldsymbol{\alpha}$ matrix. Each column is a view and each row is a latent variable. Black cells indicate small $\alpha$ values (active variable) and white cells indicate large $\alpha$ values (inactive variable). The first factor is shared for all three views, the second factor is shared between the first and second view, the third factor is shared between the second and third view, the fourth factor is unique to the first view, the fifth factor is unique to the second view, the sixth factor is unique to the third view.

## Mean

Most methods drop the mean for simplicity and work only with centered data. However, in RNA-seq data we are interested in mean-variance relationships, so we decided to explicitly model the mean as a Gaussian random variable with the following prior distribution:

$$P(\boldsymbol{\mu}) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \mathcal{N}\left(\mu_d^m \,|\, 0, \frac{1}{\beta_d^m}\right) \tag{3.4}$$

For simplicity and to keep the notation uncluttered, we set the hyperparameter $\beta_m^d$ to be equal for all values of $\mu_d^m$. Also, we set it to a very small value $10^{-3}$ in order to obtain an uninformative prior.

## Noise

The noise specification is a critical part of the model and depends on the type of data to handle. Here we assume that observations are real-valued so we define the noise to be normally distributed with zero mean and diagonal covariance matrix (which corresponds to independent observations). Importantly, similar to factor analysis, we allow each entry in the diagonal to have different values. That is, we introduce a different noise magnitude for each dimension:

$$P(\boldsymbol{\tau}) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \mathcal{G}\left(\tau_d^m \,|\, a_0^\tau, b_0^\tau\right)$$

For simplicity and to keep the notation uncluttered, we set the hyperparameters from the Gamma distribution to be equal for all values of $\tau$. Also, we set both hyperparameters to very small values ($10^{-3}$) in order to obtain an uninformative prior.

This is a simple but major update from the original GFA model [38], which assumes the noise to be homogeneous across all dimensions within the same view, which corresponds to a covariance matrix of the form $\mathbf{\Sigma}^m = \tau^m \mathbf{I}$. Such an assumption leads to an easier and faster implementation, but it does not hold in biological data, as the variance of genes expression values can vary even by orders of magnitude.

### Likelihood

The generative model is completed with the likelihood of the observed variables $\mathbf{Y}$, which follows from the previous equations:

$$
\begin{aligned}
P(\mathbf{Y}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau}) &= \prod_{m=1}^{M} \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{y}_{:,n}^m \,|\, \mathbf{W}^m \mathbf{z}_{:,n} + \boldsymbol{\mu}^m, \mathbf{T}^m\right) \\
&= \prod_{m=1}^{M} \prod_{n=1}^{N} \prod_{d=1}^{D_m} \mathcal{N}\left((y_{dn}^m \,|\, \mathbf{w}_{d,:}^{mT} \mathbf{z}_{:,n} + \mu_d^m, \frac{1}{\tau_d^m}\right)
\end{aligned}
$$

### Graphical model

The following graph summarises the scGFA model:

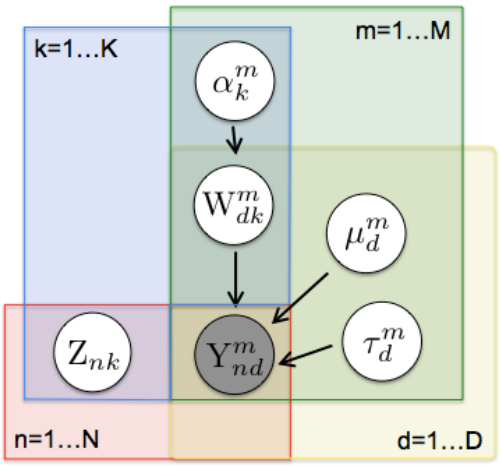



Figure 3.3: Graphical model for single-cell group factor analysis. The grey-circled variable denotes an observed variable whereas white-circled variables denote unobserved variables and have to be inferred by the model.

The explicit distributions of the different prior distributions and the likelihood can be found in Section 8.1 in the appendix.

## 3.3 Inference

As closed-form inference is not feasible, we consider the Bayesian mean-field variational framework, an efficient approximation that scales linearly in the number of cells and genes (see section **??**).
Briefly, if we denote the set of unobserved variables (parameters and factors) by $\mathbf{\Theta}$, the true posterior $p(\mathbf{\Theta})$ is approximated by a new variational distribution $q(\mathbf{\Theta})$ that has the following factorized form:

$$
\begin{aligned}
q(\mathbf{\Theta}) = q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \mathbf{T}, \boldsymbol{\mu}) &= q(\mathbf{Z}) q(\mathbf{W}) q(\boldsymbol{\alpha}) q(\mathbf{T}) q(\boldsymbol{\mu}) \\
&= \prod_{n=1}^{N} q(\mathbf{z}_{:,n}) \prod_{m=1}^{M} \prod_{k=1}^{K} q(\alpha_k^m) \prod_{d=1}^{D_m} q(\mathbf{w}_{d,:}^m) q(\tau_d^m) q(\mu_d^m)
\end{aligned}
$$

Section 2.2.1 describes in general how to do inference with the new variational distribution $q(\mathbf{\Theta})$. An example for a simpler model is also described in Section 2.3.7. The full derivation of the equations for the Variational Bayes Expectation Maximisation algorithm are presented in Section 8.2 in the appendix.

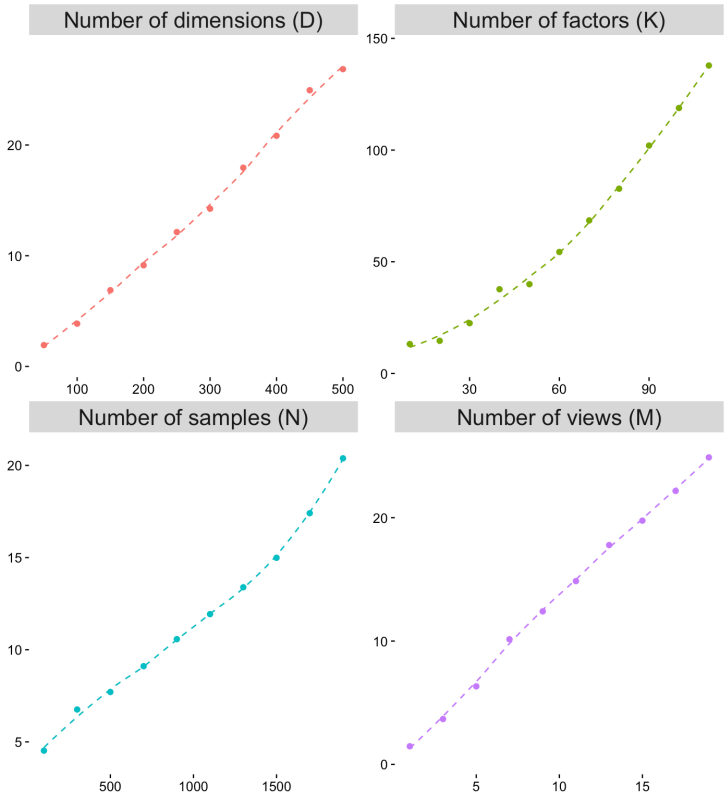
## 3.4 Special cases of the model

The scGFA model is a generalisation of simpler factor analysis models to account for multiple views and noisy data sets. As such, under some conditions it can be reduced to the simple models:

- When $\tau_d^m$ is the same for all $D_m$, then the model reduces to the original Group Factor Analysis model [38].
- When $M = 1$, all factors are active in the same view and they only describe view-specific variation. Thus the scGFA model reduces to variational bayes factor analysis [23].
- When $M = 2$ has only three types of factors: two factors associated with unique variation and one factor associated with shared variation. If we further assume that $\tau_d^m$ is the same for all $D_m$ then the model reduces to the Bayesian Canonical Correlation Analysis model formulated in [37].
- When $M = 1$ and $\tau_d^m$ is the same for all $D_m$, the model reduces to the variational probabilistic principal component analysis [7].

## 3.5 Scalability

The model scales linearly with respect to the number of samples $N$, the number of views $M$ and the number of dimensions $D$. On the other hand, it scales cubically with respect to the number of hidden variables $K$:



# Chapter 4

# Technical results: demonstration with synthetic data

## 4.1 Results: technical demonstration with synthetic data

In order to evaluate the technical capabilities of scGFA, we simulated data from the generative model in two different scenarios: in the first case we assume homogeneous noise within the same view and in the second case we allow the noise to have different magnitude for each dimension.
The results are compared against two other methods: the original GFA model as implemented in the CCAGFA package [38] and variational Bayes factor analysis (vbFA) [23] with the column-wise concatenation of the different views, which is a common approach to deal with the multi-view learning problem. vbFA is equivalent to scGFA without the group-wise sparsity structure.

The synthetic data was generated using the following settings:

1. Number of views ($M$): 3
2. Dimensionality of each view ($D_m$): sampled from a discrete uniform distribution $\mathcal{U}(D_m \,|\, 100, 1000)$
3. Number of samples ($N$): sampled from a discrete uniform distribution $\mathcal{U}(N \,|\, 50, 100)$
4. Latent factors ($\mathbf{Z}$): 6 manually constructed latent factors (Figure 4.1). The first factor is a sigmoidal function shared across all views; the second factor is also a sigmoidal function but with a phase shift, and is shared between the first two views. The third factor is a linear function that is shared between the second and third views. Finally, three random factors sampled from the latent variable prior distribution were used as view-unique factors.
5. Feature means ($\boldsymbol{\mu}$): since the traditional GFA model works with centered data, we set $\mu_d^m$ to zero for all $D_m$ to obtain a meaningful comparison.
6. Feature noise precisions ($\boldsymbol{\tau}^m$): sampled from $\mathcal{U}(\tau_d^m \,|\, 0.01, 3)$ and either set to the same value for all $D_M$ (case 1) or a different sample is used for each $D_m$ (case 2)
7. ARD prior $\boldsymbol{\alpha}^m$: factor $k$ for view $m$ is manually set to be active ($\alpha = 1$) or inactive ($\alpha = 10^6$) as specified above.
8. Weight matrices ($\mathbf{W}^m$): sampled randomly using the prior $\mathcal{N}\left(w_{d,k}^m \,|\, 0, \frac{1}{\alpha_k^m}\right)$
9. Observed data ($\mathbf{Y}^m$): sampled randomly using the likelihood $\mathcal{N}\left(\mathbf{y}_{:,n}^m \,|\, \mathbf{W}^m \mathbf{z}_{:,n}, \mathbf{T}^m\right)$

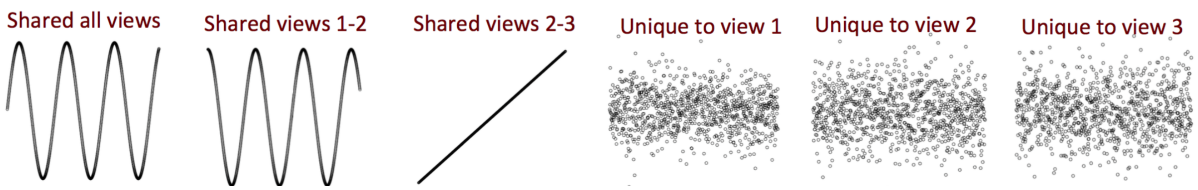



Figure 4.1: The ground truth for the latent factors used in the technical demonstration

### 4.1.1 Performance evaluation

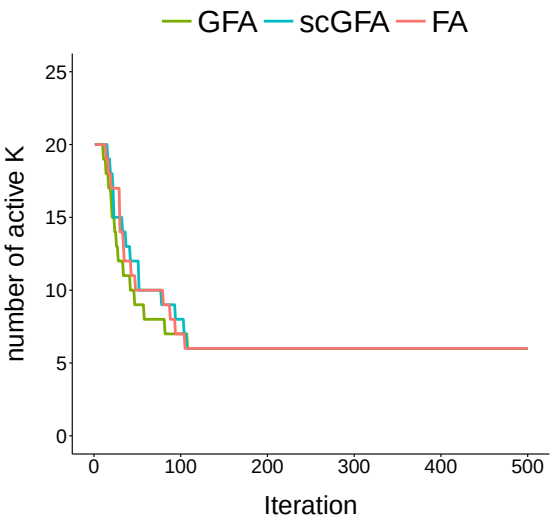
The models are evaluated using the following three different performance measurements:

- The ability to recover the true number of hidden factors: all models are started with an initial guess of $K = 20$. During training the latent variables that have an euclidean norm smaller than $10^{-7}$ (and are therefore inactive for all views) are automatically removed from the model.
- The shape of the recovered hidden factors: infered latent variables are not always directly comparable due to the rotational and scaling invariance problem (see Section 2.3.5). However, the relative distance between the points is preserved. Therefore, we computed a pairwise distance matrix $F_{ij} = \|z_i - z_j\|$ between each pair of points $i$,$j$. Then, we compared the true distance matrix with the estimated distance matrix $\hat{F}$ and we scored the correspondence between the two matrices using the sum of the element-wise absolute error: $\sum_i \sum_j F_{ij} - \hat{F}_{ij}$
- The ELBO: the variational evidence lower bound is the quantity optimised by the model and can be interpreted as a regularised version of the likelihood, so it gives the balance between model fitting and model complexity. Hence, is commonly used for model selection.

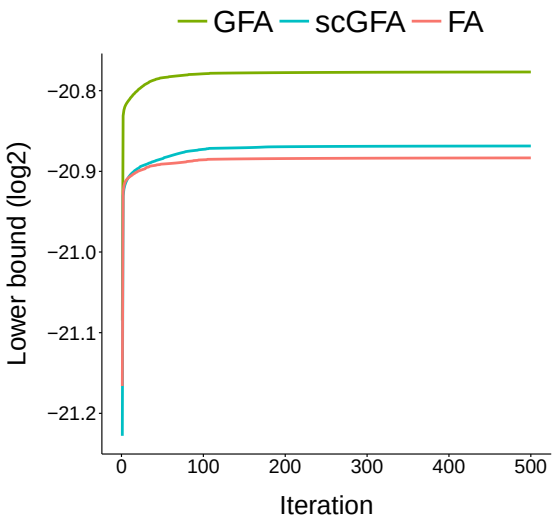
### 4.1.2 Case one: homogeneous noise

In the first scenario, all values $\boldsymbol{\tau}^m \in \mathbb{R}_+^{D_m}$ are set to the same random sample for all $D_m$. This represents the simplest (and most unrealistic) example, but it provides a good starting point for comparing the different models. The results are shown in Figures 4.2 to 4.4.

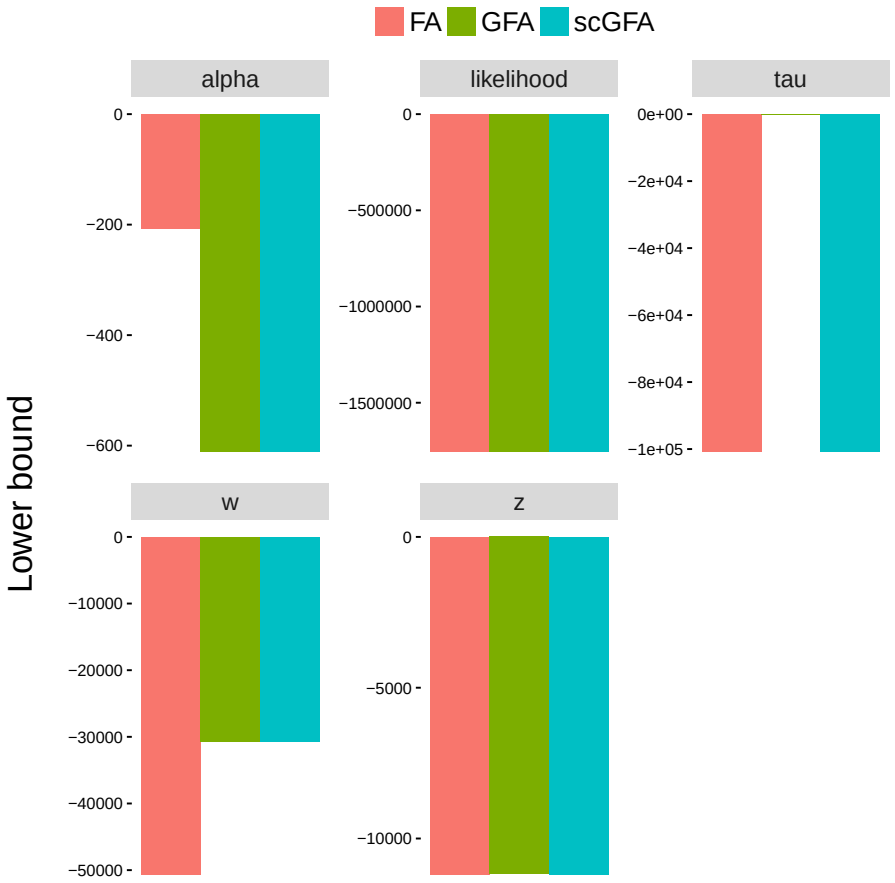
All models are able to recover the true number of latent variables (Figure 4.2(a)), although as expected, the vbFA model leads to much noisier and inaccurate estimates that can be appreciated both in quantitative and qualitative terms (Figure 4.3).
On the other hand, the GFA model reaches a better ELBO than scGFA and vbFA (Figure 4.2(b)). This result is expected, because in this scenario only one variance parameter is required for each view, and both vbFA and scGFA are estimating many more parameters than required, which leads to a high penalisation for model complexity. This can be numerically confirmed by decomposing the ELBO into its additive terms (Figure 4.2(c)): while the likelihood of all models is similar, the $\boldsymbol{\tau}$ term is much larger for the FA and scGFA models.
Interestingly, the vbFA model is also more penalised on the $\mathbf{W}$ term than the scGFA or GFA models, which can be attributed to the lack of group-wise sparsity prior on the weight matrix. In particular, vbFA is only able to determine if a component is active or not, but not for which specific views, which leads to a non-sparse weight matrix with non-zero values for entries that should have been switched off. For example, if a latent variable is active in a subgroup of views, then in both GFA and scGFA the weights for the other views are essentially zero. However, vbFA is not able to clearly distinguish which views are active and which ones are not and hence is not able to drive the weights of inactive views to zero (Figure 4.4).




(a) Training curve for the number of active latent variables.




(b) Training curve for the evidence lower bound.




(c) Decomposition of the evidence lower bound into its additive terms (see Equation (8.7))

Figure 4.2: Monitoring of training statistics for the first scenario

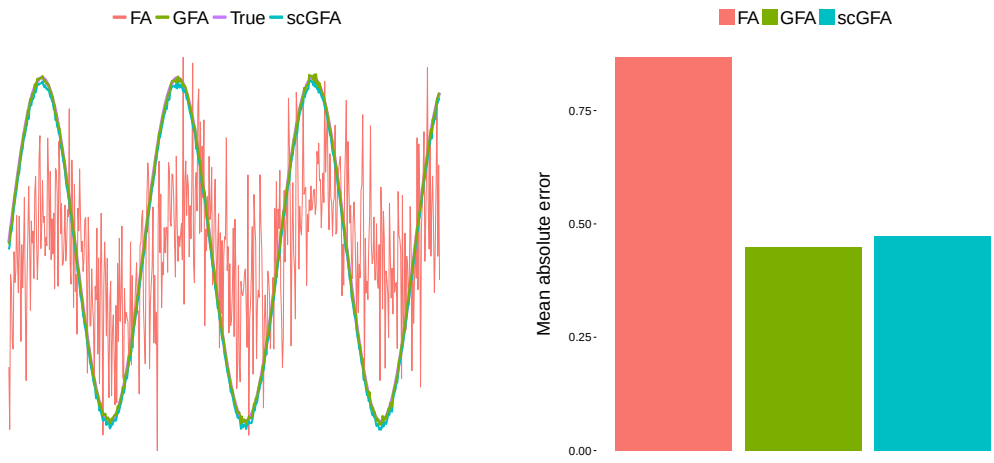



Figure 4.3: Statistics for the reconstruction of the latent variable space in the first scenario. Left: comparison of the shape of a representative predicted latent variable with the ground truth. Right: mean absolute error of the predicted and true distance matrices between the $N$ datapoints.

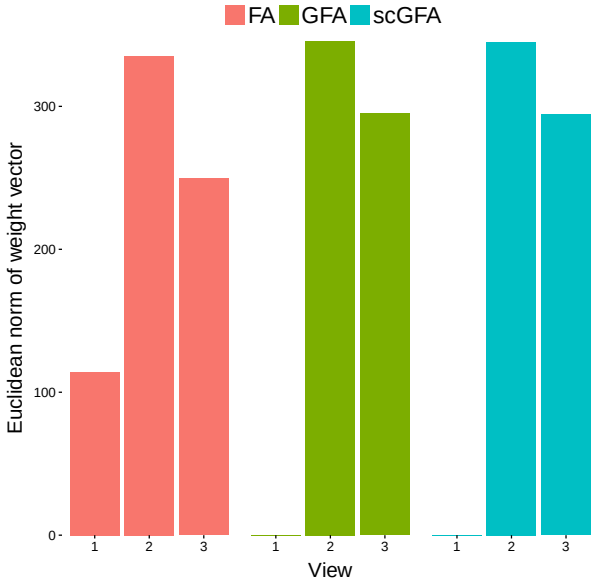



Figure 4.4: Euclidean norm of the weight vector associated to an arbitrary latent variable known to be active in the second and third views, but inactive in the first view.

### 4.1.3 Case two: heterogeneous noise

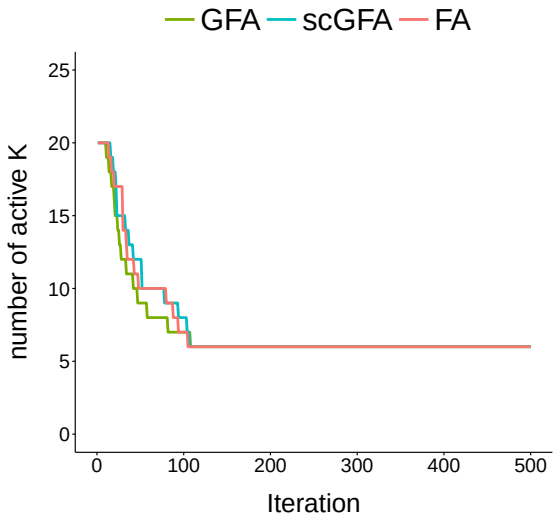
In the second scenario, $\boldsymbol{\tau}^m \in \mathbb{R}_+^{D_m}$ is not homogenous for all $D_m$, but each value is randomly sampled. This is a more realistic data set where each dimension has different variance. The results are shown in Figures 4.2 to 4.4.

In this case, the scGFA model clearly leads to the best performance: it is able to shut down unnecesary components (Figure 4.5(a)), it provides the best fit to the data (Figure 4.5(c) likelihood term) and it successfully recovers the shape of the latent variables (Figure 4.6).

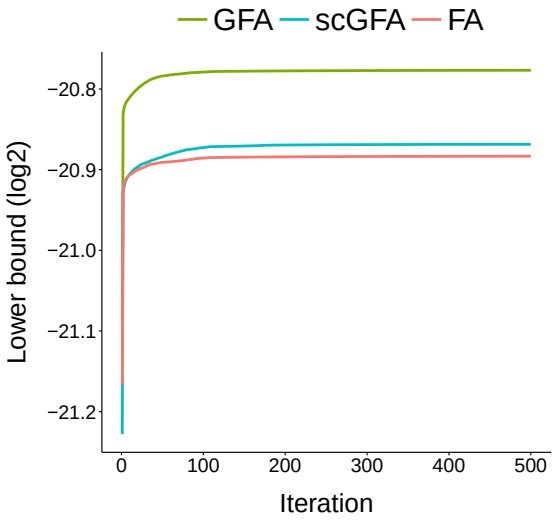
On the other hand, the GFA model recovers the shape of the latent variables, but is unable to shut down unnecesary latent variables and provides a worse fit to the data, which yields a poor ELBO mainly penalised by the worse likelihood and the excess of latent variables.

In contrast to GFA, the vbFA model does shut down unnecesary components and makes a very good fit to the data. However, it is unable to clearly separate the different latent variables, and yields a very noisy and overlapped estimate. Similarly to the first scenario, the lack of group-wise sparsity in the vbFA model

leads to a dense weight matrix and a penalisation for model complexity on the **W** term of the ELBO.




(a) Training curve for the number of active latent variables.




(b) Training curve for the evidence lower bound.


(c) Decomposition of the evidence lower bound into its additive terms (see Equation (8.7))

Figure 4.5: Monitoring of training statistics for the second scenario

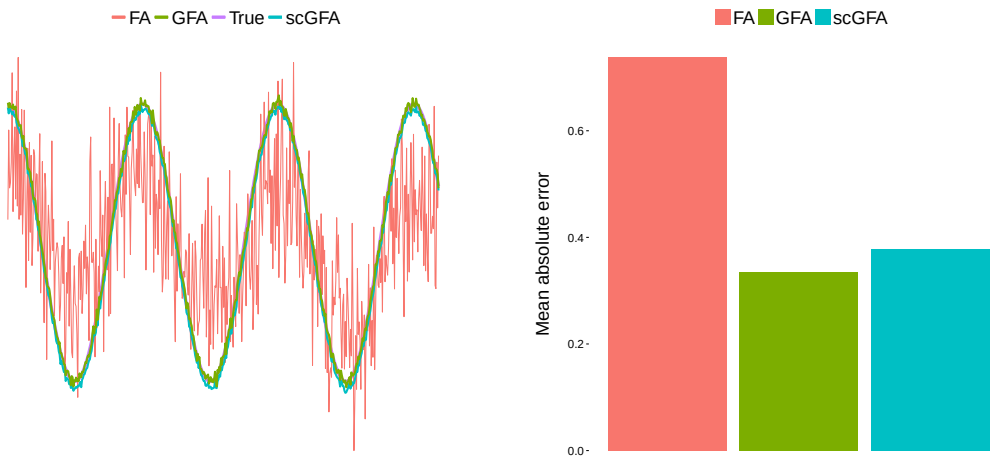



Figure 4.6: Statistics for the reconstruction of the latent variable space in the second scenario. Left: comparison of the shape of a representative predicted latent variable with the ground truth. Right: mean absolute error of the predicted and true distance matrices between the $N$ datapoints.

# Chapter 5

# Biological results: co-analysis of single-cell DNA methylation and gene expression

## 5.1 Introduction to single-cell sequencing

The use of bulk sequencing approaches have allowed the identification and characterisation of a vast amount of cellular phenotypes. As an ensemble-based approach, bulk methods estimate the average gene expression across a large population of cells. While this can be sufficient in contexts where the cells are fairly homogeneous, it is not desirable in contexts where there are larger phenotypic differences even within the same cellular population [15]. A clear example is the heterogeneity in pluripotent markers in embryonic stem cells [19] or the asynchronous differentiation processes of immunological T cells [28]. In this cases, assaying gene expression at the single-cell level provides a unique and powerful tool to unravel the complexity of cellular processes.

Single-cell studies have been done for many years with methods such as quantitative PCR or RNA fluorescence in situ hybridization, but these methods are low-throughput and not scalable to entire transcriptomes. The introduction of high-throughput single-cell RNA sequencing (scRNA-seq) have revolutionized the scope and depth of transcriptome analysis [15].
However, scRNA-seq still presents important challenges, both technical and computational [33]. To ensure that scRNA-seq data is fully exploited and interpreted correctly, it is crucial to apply appropriate statistical methods in almost every layer of the analysis. In particular, a major challenge is to disentangle technical noise and batch effects from biological variability, because if not accounted for, the effect of the most pronounced factors can mask other more subtle variation associated with signatures of primary interest [9].

## 5.2 Data set: parallel single-cell methylome and transcriptome

### 5.2.1 Protocol and data set description

To demonstrate the potential of the method, we applied scGFA to a recent data set of 61 embryonic stem cells (ESCs) under serum conditions generated by a multiparameter sequencing technology called single-cell Methylome and Transcriptome sequencing (scMT-seq) [4]. Briefly, scMT-seq performs a parallel single-cell genome-wide bisulfite sequencing and single-cell RNA sequencing to obtain a profile of the DNA methylation and the gene expression in single cells (Figure 5.1).

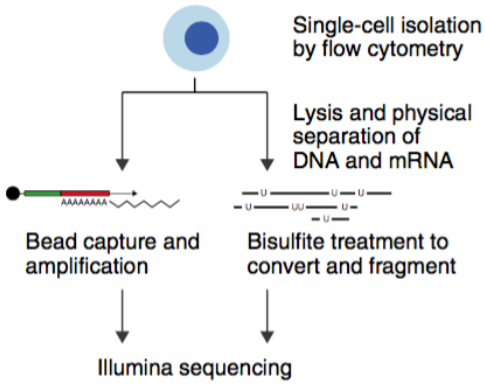



Figure 5.1: Overview of the flow chart of the scM&T-seq protocol. Single cells are collected by flow cytometry after ToPro-3 and Hoechst 33342 staining to select for live cells with low DNA content (i.e., G0 or G1 phase cells). Subsequently, cells are lysed before poly-A RNA is captured on magnetic beads and physically separated from DNA. Amplified cDNA is generated from mRNA on beads whilst DNA is bisulfite converted and Illumina sequencing libraries are prepared from both components in parallel. Reprinted from [4].

ESCs are a cellular population derived from the inner cell mass of the embryo and have two important characteristics: the capacity for differentiation into all somatic cell types and the property of unlimited self-renewal *in vitro* [3]. Given the enormous potential of these cells for medical applications, it is important to characterise their molecular machinery and to understand how to control their proliferation and differentiation into specialised cell types.
ESCs can be grown *in vitro*, but the maintenance of the pluripotency state depends also on extracellular cues. ESCs cultured in the presence of serum and leukemia inhibitory factor (LIF) are exposed to several differentiation factors and exhibit heterogeneous and dynamic expression of important pluripotency factors such as Nanog, Rex1, Dppa3 or Prdm14 [19, 30]. The fluctuations are associated with the differentiation potential of the cell, and it has been hypothesized that they define a set of coexistent metastable subpopulations [19, 30].
On the other hand, ESCs grown under a chemically defined medium without serum and the addition of two small inhibitors of the MAPK and GSK3 pathways (PD0325901 and CHIR99021) reach what is called the ground state of pluripotency, that is, cells are homogeneous and show high pluripotency potential [36]. In this project we only considered serum-grown ESCs because scGFA is focused on modelling the cell-to-cell variance of genomic features. Therefore, without heterogeneity there is no variation between cells and factor analysis models are not adequate for the analysis.

ESCs are also a model organism for studying epigenetic mechanisms such as DNA methylation, an enzyme-mediated chemical modification of DNA involving addition of a methyl group symmetrically on the cytosines of CpG dinucleotides. In general, mammalian genomes are CpG depleted, and roughly 60-80% of the CpG sites are methylated. Around 10% of CpG sites occur in CG dense regions called CpG islands, which are prevalent at transcription start sites of housekeeping and development regulator genes [3].
ESCs were shown to possess a unique DNA methylation signature when compared to differentiated cells. In particular, ESCs show widespread lack of DNA methylation that results from a global demethylation event following fertilization. On the other hand, differentiated cells show high levels of methylation throughout the entire genome, except on promoters associated with actively transcripted genes [3].
Accordingly, previous studies have shown that ESCs that lack all three DNA methyltansferases remain viable, but their differentiation capability is disrupted, apparently because of the inhability to silence the transcriptional circuitry associated with pluripotency [32]. In contrast, upon depletion of methylation levels, neither the molecular signature of pluripotency nor self-renewal capacity are significantly affected. Overall, experiments suggest that DNA methylation is critical for cellular differentiation but not for maintenance of the pluripotent state, but how is the entire process globally and locally modulated remain unanswered questions.

In conclusion, the scMT-seq ESC data set constitutes an ideal data set to test scGFA for several reasons. First, single-cell data is hard to analyse by means of traditional statistical methods. Second, there is cell-to-cell variation that is known to be associated with the pluripotency state, so this acts as a gold standard control that a factor analysis model should be able to capture. Third, the data consists of two different

types of views that are known to be associated but only a few local associations have been described [4].

### 5.2.2 Data processing

In more detail, the data generated by scMT-seq (after raw data processing) consists of a positive-real valued gene expression matrix and a binary methylation matrix of all CpG sites in the genome (Figure 5.2). Each entry in the expression matrix contains the log2 normalised read counts for each gene and cell and each entry in the methylation matrix indicates the methylation status for each CpG site and cell in the genome, 0 for unmethylated 1 for methylated and $NA$ for non-observed.
Subsequently, we used genomic annotations to obtain a set of real-valued [0-1]-restricted methylation matrices for different functional genomic contexts: coding sequences (exons), intergenic DNA, genic non-coding sequences (introns), promoters overlapping with CpG islands, promoters not overlapping with CpG islands and active enhancers (see Section 7.2 in Methods for more information about the annotations). In particular, each entry is calculated by taking the proportion of methylated CpG sites inside the annotation. Therefore, 0 indicates that all observed CpG sites are unmethylated and 1 indicates that all observed CpG sites are methylated. Figure 5.2 shows a representation of the processed data.

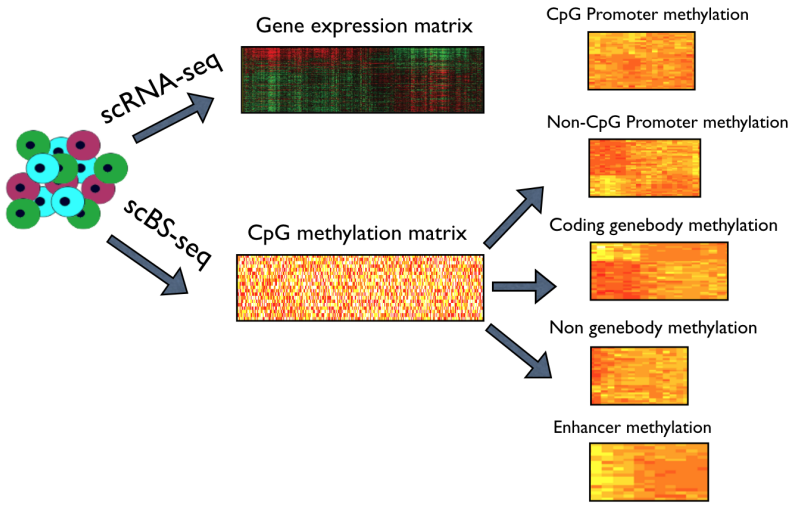



Figure 5.2: Overview of data processing for scMT-seq. A read count matrix is obtained using scRNA-seq, which is then normalised, filtered and log-transformed to yield the gene expression matrix. A binary CpG methylation matrix is obtained using scBS-seq, which is then parsed with genomic annotations to yield a set of [0-1]-restricted real valued matrices for each genomic context.

## 5.3 Quality control on expression data

To characterise and to assess the quality of the processed single-cell expression data, we performed the following set of analysis on the expression matrix.

### 5.3.1 Correlation between samples

To verify that all samples are consistent and to detect the presence of outliers we calculated the Spearman correlation coefficient between all 61 ESCs (Figure 5.3(a)) Clearly, the correlation is lower than in bulk RNA-seq experiments but it is still quite significant, with roughly $\rho = 0.65$ as average. This result is expected, since it is known that both the technical and biological component of noise is much bigger in single-cell data than in bulk data.

### 5.3.2 Coefficient of variation

It is known that genes with larger expression display a lower coefficient of variation. We verified that the same pattern is found in the scMT data set (Figure 5.3(b)).

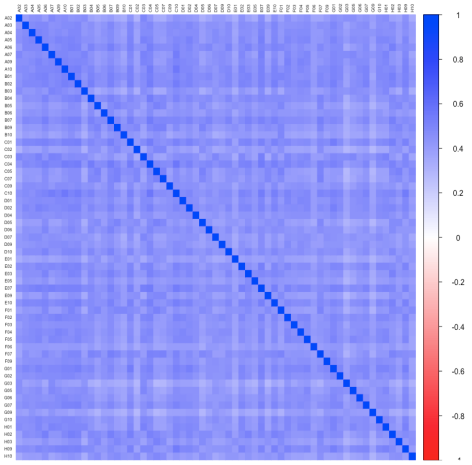
### 5.3.3 Distribution of expression levels

Next, we evaluated the overall distribution of expression levels, ignoring gene and sample labels (Figure 5.3(c)). As previously reported [13, 27], the distribution is clearly bimodal with a dense peak at zero associated with dropout events (zero measurements) and a second normally-shaped mode associated with non-zero expression levels which resembles the distribution of bulk expression data.

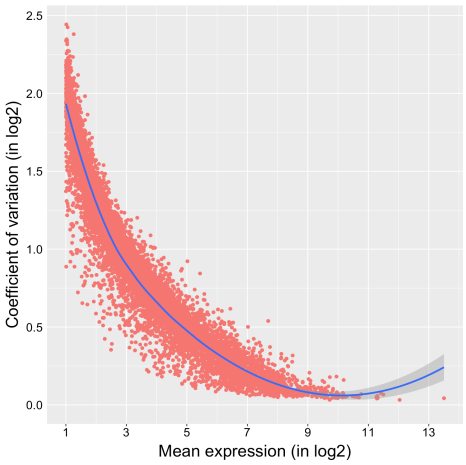
### 5.3.4 Dropout rate

scRNA-seq data contain an abundance of dropout events that lead to zero expression measurements, which are mainly the result of technical sampling effects due to low transcript numbers [13, 27]. Therefore, in single-cell data, a zero observation might be the result of a dropout event, or it might represent the scenario where the gene has no detectable expression.
To assess the effect of the dropout in the cell-to-cell variation of the scMT-seq data set, we performed traditional PCA on the expression matrix and we correlated the principal components with the cellular dropout rate (Figure 5.3(d)). The first two principal components are capturing variation that is correlated with the cellular dropout rate, which confirms it as a main source of variation.

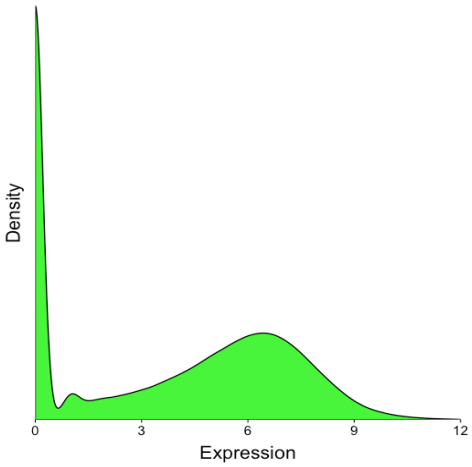
In conclusion, it is evident that the zero observations introduce an important source of variability that has to be rigurously taken into account or it may mask other more subtle biologically-relevant variation. How to deal with zero values in single-cell data is an active area of research [13, 27]. For simplicity and to avoid complicated non-gaussian noise models in scGFA, we decided to treat zero observations in the same way as bulk RNA, as a negligible expression level.



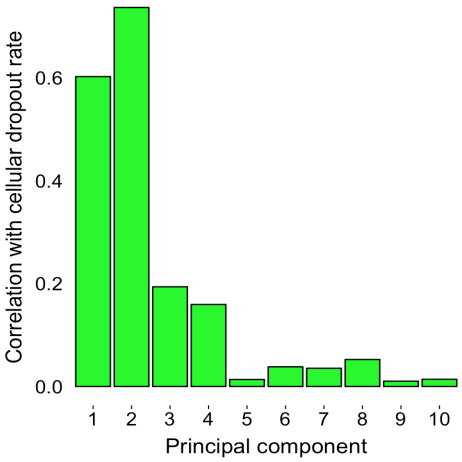
(a) Spearman correlation coefficient between the expression levels of the embryonic stem cells in the scMT data set.




(b) Biological Coefficient of Variation (CV) as a function of the average expression.




(c) Distribution of expression levels for all genes in all samples.




(d) Correlation between the principal components of the gene expression matrix and the cellular dropout rate

Figure 5.3: Quality control for the expression data

## 5.4 Quality control on methylation data

To characterise and to assess the quality of the processed single-cell methylation data, we performed the following set of analysis on the set of methylation matrices associated with the different genomic contexts.

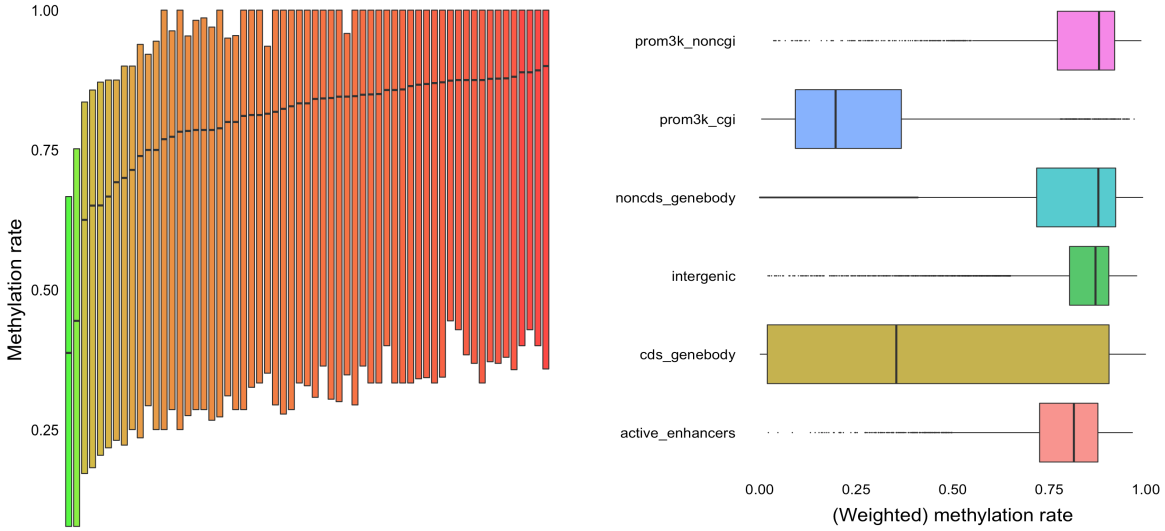
### 5.4.1 Cellular mean methylation rate

First we checked whether the ESCs show heterogeneity in the overall methylation levels. As shown in Figure 5.4(a), there is significant cell-to-cell differences in the average CpG methylation rate, ranging from a mean of 40% to 80%. As mentioned, most differentiated mammalian cells are known to be stably hypermethylated, with an average CpG methylation rate of roughly 70-80%, but cells in developmental stages are known to undergo massive changes in the methylome [32]. Therefore, even though we expected less overall methylation, the result is not surprising. In fact, genome-wide methylation heterogeneity has been previously described in ESCs grown under serum conditions [31].

### 5.4.2 Mean methylation rate for genomic contexts

It is widely known that the methylation rate is not uniformly distributed across the genome, and functional elements have particular DNA methylation signatures.
According to our results shown in Figure 5.4(b), promoters overlapping with CpG islands show by far

the lowest levels of methylation, followed by the coding part of the genes (exons), which are on average partially methylated and display a high variability in the methylation rate. The rest of genomic contexts show a consistently high methylation rate. All observations have been previously documented in both bulk and single-cell experiments [31], so it confirms that there are no major biases in our data.




(a) Boxplot of the methylation rate distribution in single cells-

(b) Boxplots of the methylation rate for different genomic contexts.

Figure 5.4: Quality control for the methylation data

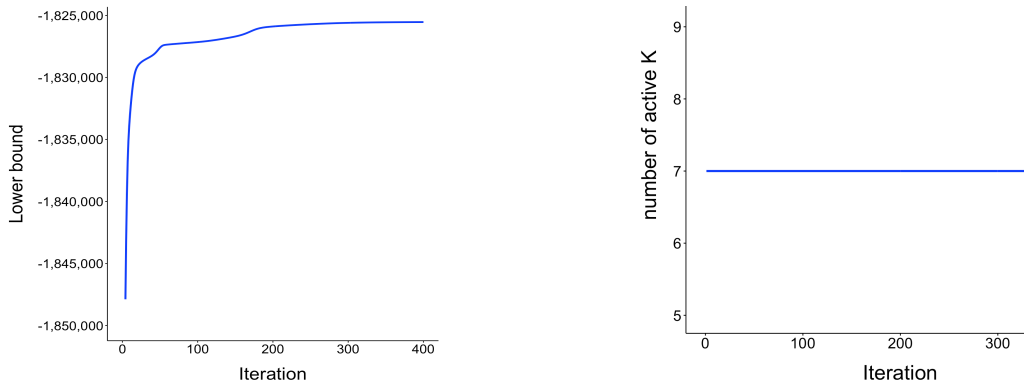
## 5.5 single-cell Group Factor Analysis

We trained the scGFA model using a total of six input views (see Figure 5.2). One view contains the gene expression levels and the other five views contain the methylation levels in different genomic contexts: promoters overlapping with CpG islands, promoters not overlapping with CpG islands, coding gene sequences, non-coding gene sequences, active enhancers and intergenic DNA.
The initialisation of the model and the selection of the hyperparameters is described in the Methods (Section 7.3).

### 5.5.1 Training

Training was monitored using the ELBO, a key statistic that in the variational Bayesian framework is optimised at each iteration (see Section 2.2 for a detailed explanation). Also, since it can be interpreted as a complexity-penalised version of the likelihood it is also useful for model selection. Hence, we trained multiple instances of the model and we calculated the lower bound at each iteration of the algorithm. We stopped training once the change in the lower bound was small enough, and we selected the instance with highest lower bound. The training curve is shown in Figure 5.5(a).

We also monitored the number of latent variables that are active in at least on view (Figure 5.5(b)). The group-wise ARD prior allows the model to shut down variables that are not explaining variability in any of the views. Thus, theoretically, one should initialise the model to a large number of latent variables and then disregard the non-active ones. However, in practice, we ended up with several moderately correlated factors capturing similar information. Therefore, in order to avoid redundancy and to obtain only a few relevant set of factors, we decided to initialise the model with a small number of latent variables (7). Then, as expected, scGFA uses them all to model variance and does not shut down any component during training.




(a) Training curve for the variational lower bound.

(b) Training curve for the number of active latent variables.

Figure 5.5: Training statistics for scGFA on the scMT dataset

#### Percentage of variance explained by the model

The introduction of a noise term is of critical importance as it allows the model to focus on modelling the structured variation and disregard the variation that does not match the patterns captured by the model (also called residual variation). This is particularly important in very noisy datasets such as scMT.
To assess the amount of detected signal and noise, we calculated the fraction of structured cell-to-cell variation $f^m$ captured by the model in data set $m$ with respect to the total variation:

$$f^m = \sum_{d=1}^{D_m} \frac{\sigma_d^m - \frac{1}{\tau_d^m}}{\sigma_d^m} \tag{5.1}$$

where $\sigma_d^m$ is the observed cell-to-cell variance for feature $d$ in view $m$ and $\tau_d^m$ is the corresponding precision of the gaussian noise learnt by the model (see Section 3.2). The results for each view are shown in Figure 5.6.

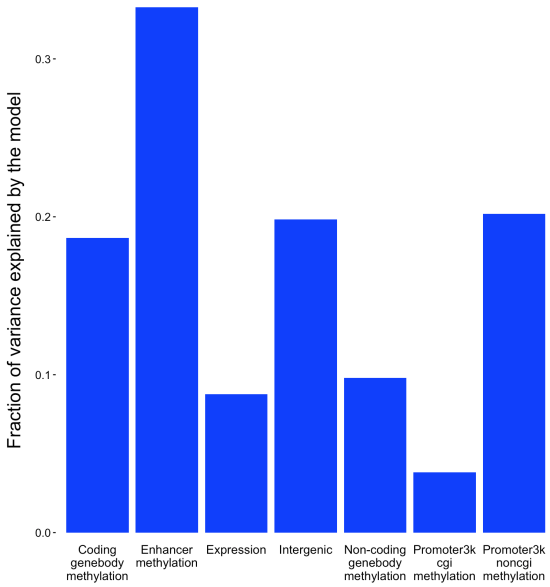



Figure 5.6: Fraction of total variance explained by the model for each view separately.

Surprisingly, the model uses a small proportion of the total variance as signal and assigns most of the variation to the noise term, which implies that the signal-to-noise ratio is remarkably small in both the expression data and the methylation data.

### 5.5.2 View versus Factor analysis

To identify which latent variables are explaining variation unique in one dataset and which ones are explaining join variation between multiple datasets, we calculated for each view the proportion of structured variance associated with each factor (Figure 5.7) as follows:

$$\sum_{d}^{D_m} \left(\frac{1}{\alpha_k^m}\right)\left(\frac{\phi_k}{\sigma_d^m - \frac{1}{\tau_d^m}}\right)$$

where $\phi_k$ is the variance of the latent factor $k$ and $\alpha_k^m$ is the precision of the weight vector $\mathbf{w}_{:,k}^m$ infered by the model. the numerator is the variance explained by factor $k$ in view $m$ and the denominator is the total variance of view $m$.
Interestingly, all methylation views share a single latent variable that explains a large proportion of the variation between cells. On the other hand, the expression view has at least two active latent variables, both of them unique.

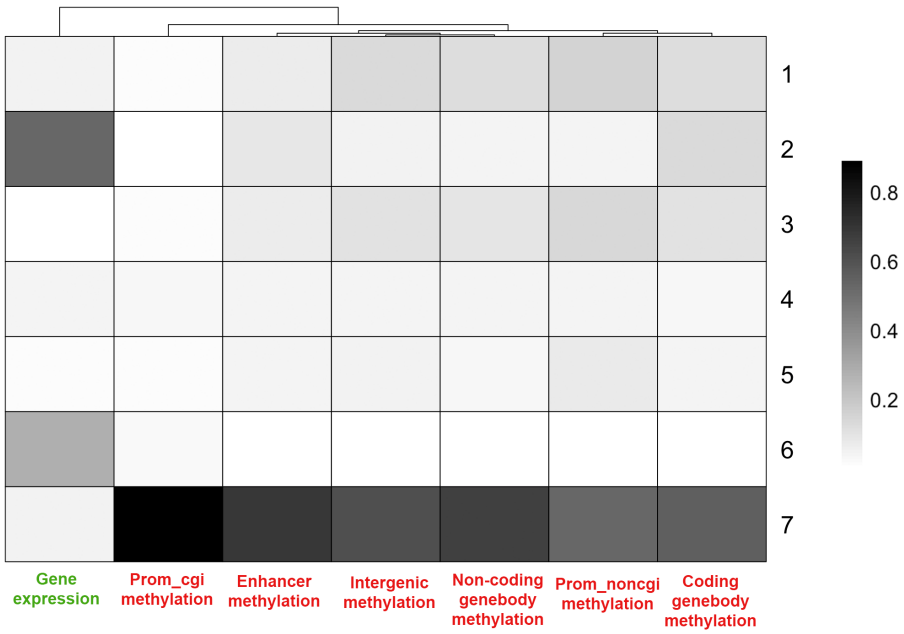



Figure 5.7: View versus Factor plot. Each cell in the matrix displays the proportion of non-residual variation aexplained by a latent variable (rows) in a given view (columns). Therefore, the darker the cell the more active the factor is.

### Correlation plot of latent variables

As explained in Section 3.2, in order to avoid redundancy between the factors, traditional LVMs usually impose the constraint of the latent variables to be orthogonal. In the Bayesian framework, one can only encourage the latent variables to be independent in the prior distributions, but one might end up with correlated latent variables in the posterior estimates.
To assess if the latent variables are redundant, we calculated the spearman correlation coefficient between all factors Figure 5.8:

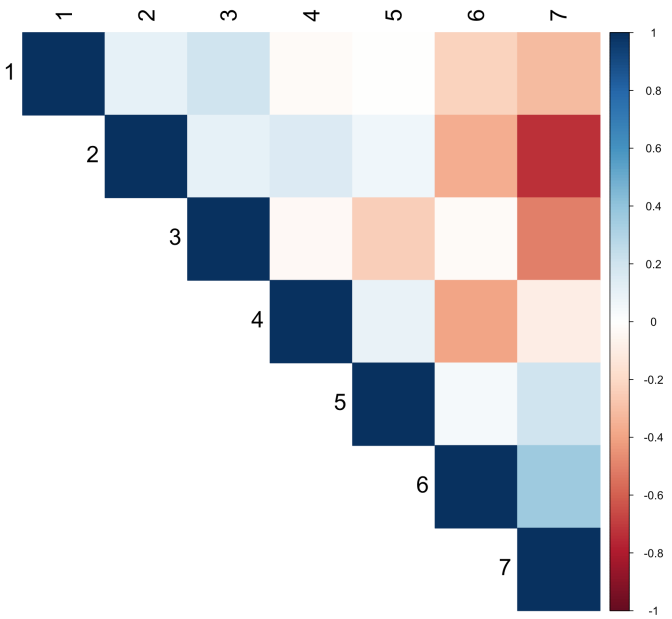



Figure 5.8: Spearman correlation coefficient between all latent variables

Except the case of the second hidden varible which is highly correlated to the seventh, there are no major correlations between the rest of factors. As a matter of fact, as described in the following section, the only significant correlation observed stems from a biologically meaningful relationship.

### Characterisation of the latent variables

To assess whether the active latent variables are associated with known biological processes, we performed a gene ontology enrichment analysis on the corresponding weight vectors for all views whose features can be associated with genes (expression, promoter methylation, cds methylation and non-cds methylation). For the views not directly associable with genes (active enhancers and intergenic DNA) no enrichment analysis was performed.

According to our results, the factors active in the methylation views do not show statistically significant enrichment for any biological pathway (data not shown), which suggests that the cell-to-cell variation is not associated to specific biological processes. This is further confirmed by the fact that the most active factor shared by all methylation views (latent variable 7) is in fact capturing differences in genome-wide mean methylation rate of cells (Figure 5.9).

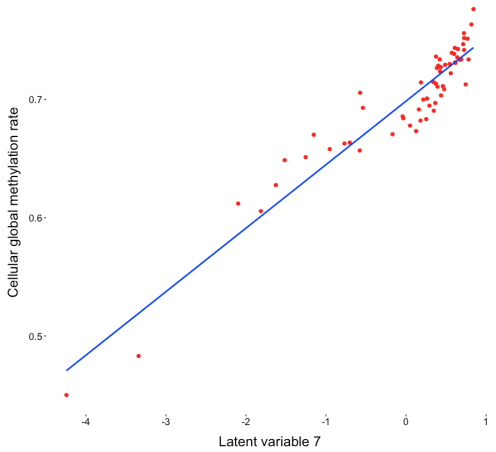



Figure 5.9: Scatterplot of the cellular mean methylation rate and the seventh latent variable infered by scGFA

Thus, we can state that the variance in the methylation rate between ESCs is attributed to broad genome-wide changes in the mean CpG methylation rate. The initial exploratory analysis on the methylation data (see Section 5.4) further supports this observation. Previous research have also reported similar observations [31].

On the other hand, the latent variables active in the expression view do show significant enrichment for some biological pathways. The most active factor (latent variable 2) is heavily enriched for pluripotency-related genes, so one can state that pluripotency is the main driver of variation in the expression data, which agrees with the well-known heterogeneity of ESCs under serum conditions [19]. In biological terms, this implies that cells have variability in their pluripotency potential.
The second strongest factor (latent variable 6) is not associated to any particular biological pathway but is capturing a known covariate in single-cell expression data: the dropout rate. As described in Section 5.3.4 the number of zero observations is a major technical driver of variability in scRNA-seq experiments.
Then, we also observe in a weak factor (latent variable 4) a significant enrichment for pathways related to DNA repair and replication stress. This is an interesting finding because it is known that the high proliferative potential of ESCs compared to differentiated cells induces a replication stress which requires the DNA damage response machinery to be much more robust [1]. Hence, there might be a biological relevance in this finding, but the associated latent variable explains only a very small proportion of the structured variation in the expression view and is not reproducible in different runs of the model (data not shown).

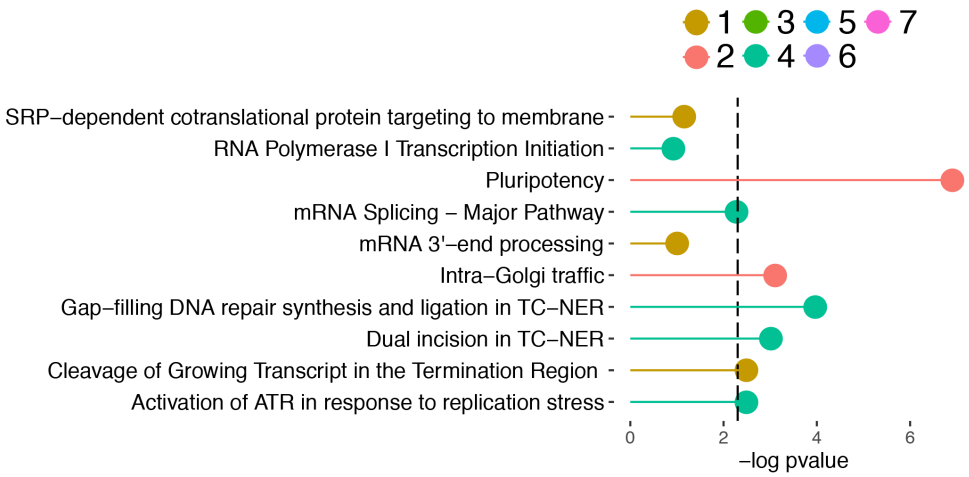



Figure 5.10: Gene ontology enrichment analysis for the latent variables and the expression view. The enrichment statistic is calculated on the weight vectors, which relate the low-dimensional latent variables with the high-dimensional views. Only the top 10 pathways with lowest p-value are shown. The dashed lines estabish the significance threshold of FDR=10%. Each colour represents a latent variable

In conclusion, we obtained two evident biologically-meaningful axis of variation: one associated with the cellular genome-wide methylation rate and another one associated with the pluripotency potential of cells.

### Characterising cellular subpopulations

Armed with the multi-view infered latent variables, we asked the question whether we can identify cellular subpopulations within the pool of ESCs.
To do so, we made a scatterplot of the two biologically-relevant latent variables, the pluripotency factor and the global genome-wide methylation rate factor (Figure 5.11). Each one of the two factors by itself is

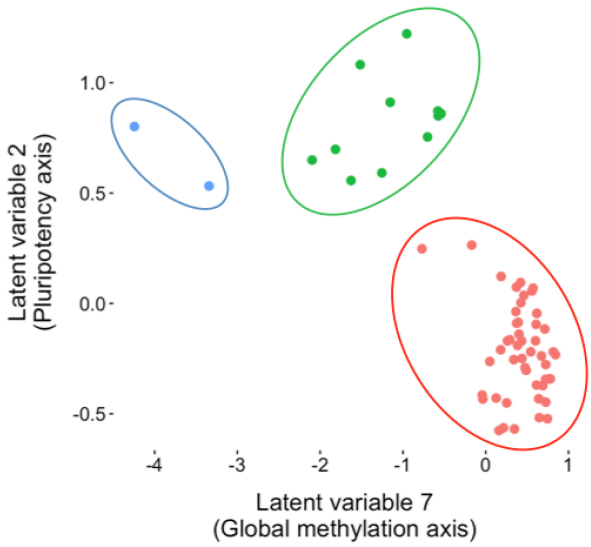



Figure 5.11: Scatterplot of the two main latent variables infered by scGFA. Each dot is a sample.

unable to define the existence of evident subpopulations, but their combination defines three clusters.
To further characterise the relevance of this finding we divided the cells into the hypothetical subpopulations (red, blue and green) and made an exploratory data analysis in the expression and the methylation data.
First, we compared the expression of genes involved in the pluripotency core network, genes associated with the three differentiated germ layers (endoderm, mesoderm, ectoderm) and housekeeping genes as a control (Section 5.5.2).

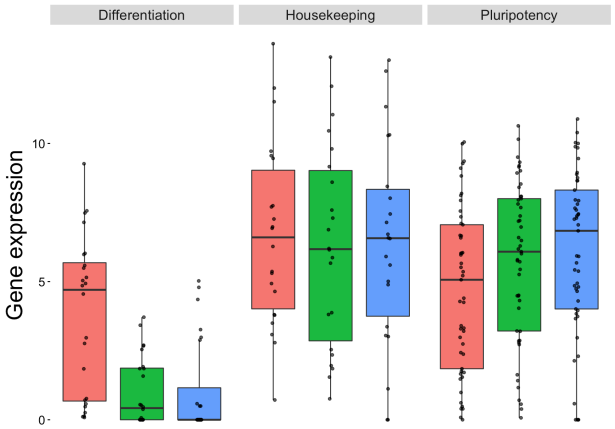



Figure 5.12: Characterisation of the expression levels of three gene sets, differentiation, housekeeping and pluripotency, for the three subpopulations. Each dot is a gene.

The red cluster shows high expression levels of differentiation genes and relatively low levels of pluripotency genes,a phenotype associated with pluripotent cells undergoing differentiation. The blue cluster shows an opposite trend, with essential no expression of differentiation genes and relatively high levels of pluripotency genes, a phenotype associated with the ground state of pluripotency. Finally, the green cluster shows intermediate levels of expression of pluripotency genes and virtually no expression of differentiation genes, which might represent a transient state between the two ends of the differentiation pathway.
Importantly, note that all clusters express significant amounts of pluripotency genes, so the results are not influenced by the presence of fully differentiated outlier cells.

Next, we performed a similar analysis for the second axis of variation associated with methylation. We grouped cells according to the cluster label and we compared the global mean methylation rate between the three clusters (Figure 5.13), which reveals that the red cluster consists of hypermethylated cells, the blue cluster of hypomethylated cells and the green cluster of partially methylated cells.

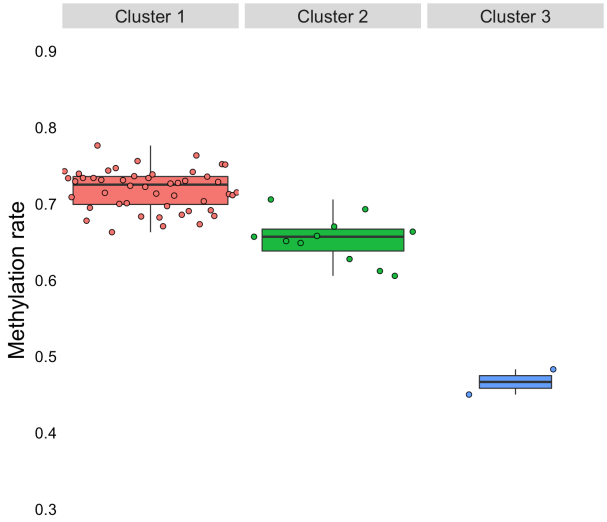



Figure 5.13: Characterisation of the genome-wide methylation rate for the three subpopulations. Each bar shows the mean methylation rate of a cell. The dashed lines define the mean over all cells in a cluster

In conclusion, we found a clear relationship between the expression levels of the pluripotency machinery and the global methylation rate: highly pluripotent cells are associated with an hypomethylated genome, and the further a cell moves along the differentiation pathway, the more methylated the genome becomes. Nevertheless, it is important to remark that with the current approach we cannot prove causality, so whether is the methylation driving the changes in pluripotency expression or is the pluripotency expression driving the changes in methylation cannot be answered with this approach.

# Chapter 6

# Discussion and conclusions

## 6.1 Model summary and technical capabilities

In this thesis we developed single-cell Group Factor Analysis (scGFA), a multi-view learning model that integrates multiple data sets and disentangles the different sources of variation, identifying which ones are unique to a single view and which ones are shared between multiple views.
The core of scGFA is based on a group factor analysis latent variable model [38] that we extended it to account for the particular features of single-cell sequencing data. In particular, we added a mean term and we generalised the noise model to have different variance for each dimension. Furthermore, we are currently working on extending the model to non-gaussian likelihoods.
The key element of the group factor analysis solution, and what makes it different from single-view factor analysis methods, is the group-wise sparsity imposed on the prior distribution of the weights. Here we adopted the Automatic Relevance Determination prior as originally proposed in [38], which allows the factors to be active or inactive in a per view basis.
Inference is performed using the variational Bayesian approximation, which scales linearly with the number of samples, views and dimensions, thereby making the model applicable to potentially very large data sets.

We showed that the model is capable of learning the hidden structure of a simulated multi-view data set, outperforming previous multi-view approaches.

## 6.2 Biological results

We applied scGFA to the scMT-ESC data set, a parallel profiling of the expression and the methylation in single ESCs. scGFA revealed four main sources of variation (or hidden variables/processes): three associated with the expression view and one shared by all methylation views. Among the expression variables, one is capturing technical variation due to dropout events and two of them are capturing biologically-meaningful variation associated with the pluripotency pathway and with the DNA repair pathway, respectively. The variable active in the methylation views is capturing variation due to global changes in the mean methylation rate.
Importantly, even though the model captures the pluripotency and the methylation factors as separate variables there is a high correlation between them, which suggests that the variation in cellular pluripotency is partially coupled to the variation in genome-wide methylation rate. This result is expected, and it has been reported previously that cells undergoing differentiation in early development suffer major epigenetic rearrangements with an increase in the methylation levels [3, 31].

An interesting question that follows from the results is: if DNA methylation is an important epigenetic mark, why did we not find more sources of variation in the methylation views? There are several reasons that can explain it. First, DNA methylation is a local epigenetic mark that normally regulates the expression of nearby genes [32]. However, we trained scGFA with genome-wide data, without any information regarding the genomic distance between features, so we are not expected to find local events but global phenomena. Second, the quality of the methylation data is poor due to low mappability, large amounts of noise, very sparse coverage of CpG sites and large amounts of missing values that have to be imputed (around 50%).

Next, we shoted that using the two dominant latent variables (pluripotency and global DNA methylation), we can identify three subpopulations within the pool of ESCs: hypomethylated cells with high pluripotency potential, partially methylated cells with slightly less pluripotency potential and hypermethylated cells with low pluripotency potential.

The existence of cell-to-cell heterogeneity in ESCs and its organization in functionally different subpopulations have been proposed before using a variety of single-cell approaches. For instance, in [19] they used single-molecule RNA-FISH to identify two coherent gene expression states with functional biases in their differentiation propensity. A Rex1+ and Nanog+ pluripotent subpopulation and a Rex- and Nanog- differentiated subpopulation. Furthermore, they also used locus-specific bisulfite sequencing to describe differential methylation patterns between the two states.
In another approach [30], they used scRNA-seq to confirm the existence of cellular heterogeneity in serum ESCs, and they proposed the existence of not two but three subpopulations: one subpopulation that express higher levels of marker of differentiation and virtually no levels of pluripotency factors, another subpopulation with low expression levels of both markers, and a third one with high levels of pluripotency factors. It is likely that these subpopulations might be the same ones that we identified, but it is hard to assess since no parallel analysis of the methylome was performed in the study.

In conclusion, whether the subpopulations that we identify are biologically relevant or not has to experimentally confirmed, but there is supporting evidence that they are not an artifact. The multi-view approach allowed us to obtain a more accurate and unbiased picture of the relationship between transcriptomics and epigenetics. To our knowledge, it is the first time that phenotypes are identified using a multi-view approach, which opens the door to a new way of characterising cellular populations.

## 6.3 Limitations and extensions

The current model has some limitations that have to be taken into account and can lead to extended versions of the scGFA model.

First, inference is performed using the mean-field variational Bayesian framework, which has the advantage of being noticeably faster than sampling approaches, thereby allowing fully bayesian inference to be performed on all variables and large data sets. However, it suffers from the important drawback that it is an analytical approximation with no asymptotical properties, so the estimates are never unbiased. If accuracy is crucial, one should switch to a Markov Chain Monte Carlo approach, at the cost of losing scalability [26].

Second, the model leads to closed-form sequential updates in the algorithm because the priors are chosen wisely so that they are conjugated with the likelihood. Nevertheless, it is common that different views have different numerical properties associated with different likelihood models that might not be conjugated anymore. A clear example is count or binary data, which is usually modelled by Poisson and Bernoulli likelihoods, respectively. In this case, the likelihood is not conjugated with the normally distributed priors, and the variational approach is not available in closed form. Thus, either numerical approximations or further analytical assumptions are necessary [29].

Third, another limitation comes from the group-wise sparsity structure on the weight matrix, the key element of the model. Here we adopted the Automatic Relevance Determination prior, which is an example of a continuous soft prior that results in elements being close to zero but not exactly so. A general property of soft priors is that they allow for efficient continuous inference, but sometimes is not trivial to separate the low-activity components from the completely inactive ones. Ideally, inactive components should have a weight of exactly zero, which could be achieved by the introduction of hard priors such as the spike-and-slab [35].

Fourth, the scGFA solution is not identifiable due to the rotation and scaling invariance property. We consider extending the model with the solution proposed by Virtanen et al [37], where they suggest to maximise the variational lower bound with respect to the linear transformation $\mathbf{R}$ (see Section 2.3.5) at each round of the VBEM updates.

Finally, an obvious pitfall is that the model is linear, and biological relationships are known to follow complex non-linear patterns. Hence, one might consider extending the model using for instance Gaussian Processes, which have been shown to be useful in the single-cell latent variable models [9].

# Chapter 7

# Methods

## 7.1 Raw data analysis of the scMT dataset

Parallel bisulfite and RNA single-cell sequencing was performed on 61 E14 mouse ESCs cultured in serum and leukemia inhibitory factor as described in [4].
scRNA-seq reads were mapped with GSNAP (version 2014.02.28) [40] in paired end mode to the GRCm38 mouse genome assembly.
BS-seq reads were mapped using Bismark (version 0.16.3) [20] in single-end mode to the GRCm38 mouse genome assembly.

In this project we were provided with the mapped reads. The quality control standards prior to mapping are detailed in [4] and for simplicity are not reproduced here.

### 7.1.1 Sequence data processing: methylation

The output of bisulfite sequencing is a set of reads that support either methylation or unmethylation in a given CpG site. First, we assigned a binary methylation status to each CpG site on the genome by sampling from a Bernoulli distribution $\mathcal{B}(\theta)$ where the parameter $\theta$ was estimated from the ratio of methylated read counts to total read counts.
Subsequently, methylation rates were computed for the different genomic contexts (see Section 7.2) by taking the mean of the binary CpG sites in the region defined by the context. Thus, a methylation rate of 1 denotes a region where all the observed CpG sites are methylated and a methylation rate of 0 denotes a region where all the observed CpG sites are unmethylated.
Finally, each genomic context was filtered by coverage by requiring at least 3 reads in 50% of the samples to be considered meaningful for downstream analysis.

### 7.1.2 Sequence data processing: expression

#### Gene abundance estimation

FeatureCounts from the Subread package (version 1.5.0) [21] was used to calculate the table of gene raw read counts with the following options: a read was assigned to a gene if it overlapped with any exonic region. Reads were not allowed to be assigned to multiple features (*allowMultiOverlap* set to False), a fragment was not allowed to map to different chromosomes (*countChimericFragments* set to False), and a fragment was considered mapped if at least one pair mapped to the genome (*requireBothEndsMapped* set to False). Reads were mapped in unstranded mode (*strandSpecific* set to 0).
The Ensembl database (version 84) in GTF format was used for gene annotations [2]. To keep only highly-confident and known genes, we filtered out all non-coding elements, resulting in a total of 19,404 genes.

#### Normalisation and transformation

The read counts were normalised by library size using the Deconvolution method [22] and they subsequently transformed to log2 scale. No gene length normalisation was applied.

#### Filtering

Genes with an average expression lower than 1 (in log2 scale) were removed.

## 7.2 Genomic annotations

In order to calculate methylation rate for functional genomic categories, we divided the genome into the following contexts using again the Ensembl database (version 84):

- Promoters overlapping with CpG islands (prom3k_cgi): 3 kb region upstream from the transcription start site and overlapping with previously documented CpG islands [18].
- Promoters not overlapping with CpG islands (prom3k_noncgi): 3 kb region located upstream from the transcription start site and not overlapping with CpG islands.
- coding gene sequences (cds_genebody): merging of all exons within the body of a gene.
- non-coding gene sequences (noncds_genebody): merging of all introns within the body of a gene.
- Active enhancers (active_enhancers): extracted from [11] by overlapping the ChIP-seq marks for H3K4me1 and H3K27ac.
- Intergenic DNA (intergenic): all DNA located between genes.

### 7.2.1 Statistics

| Genomic context | Genome-wide percentage | Total number of sites |
|---|---|---|
| prom3k_cgi | 1.30 | 9894 |
| prom3k_noncgi | 0.79 | 2187 |
| cds_genebody | 1.24 | 9518 |
| noncds_genebody | 33.22 | 49631 |
| active_enhancers | 0.74 | 3772 |
| intergenic | 63.22 | 12429 |

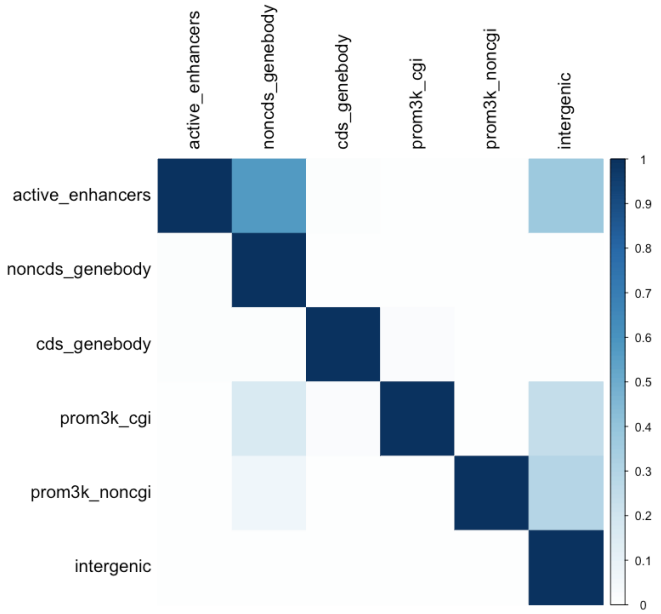



Figure 7.1: Genomic overlap between the different annotations. Each cell contains the fraction of overlap of the row with respect to the column.

## 7.3 scGFA analysis

### 7.3.1 Input data sets

The input data for the scGFA model is the expression matrix and the six methylation rate matrices. Note that they share the same $N$ samples but have different dimensions $D_m$.

**Data processing**

Factor analysis models are focused on modelling the variance between cells rather than the mean values, so sites with low variability do not provide information to the model and can cause numerical instability. Therefore, we further filtered the data by removing the 25% sites with least variance.
Missing values in the methylation data were imputed using the feature-wise mean.

### 7.3.2 Parameters

We initialised the model parameters and variables as follows:

- Latent variables ($\mathbf{Z}$): three different settings are tried, random non-orthogonal variables, maximum-likelihood pca solution and random orthogonal variables. The resulting model with the highest lower bound is selected.
- Initial number of latent variables $K$: in order to avoid redundancy and to speed up learning, we set the number of starting variables to 7, a fairly small value.
- Weights ($\mathbf{W}$): they are the first variable to be updated so no initialisation is required.
- Means ($\boldsymbol{\mu}$): initialised to the maximum likelihood estimate.
- Precision hyperparameter of the prior distribution of the means ($\beta_0$): initialised to a small value ($10^{-3}$) to make the prior uninformative
- Noise precision ($\boldsymbol{\tau}$): initialised to a large value (100) to ensure that the real structure is modelled with components instead of the noise parameters (see [23]).
- Hyperparameters of the gamma prior on $\boldsymbol{\tau}$ ($a_0^\tau$ and $b_0^\tau$): initialised to a small value $10^{-3}$ to make the prior uninformative.
- precision of the ARD prior $\boldsymbol{\alpha}$: the matrix is initialised arbitrarily to 1, so all variables are initialy active for all views.
- Hyperparameters of the gamma prior on $\boldsymbol{\alpha}$ ($a_0^\alpha$ and $b_0^\alpha$): initialised to a small value $10^{-3}$ to make the prior uninformative.

### 7.3.3 Training schedule

Training was monitored using the variational lower bound, which is guaranteed to increase at each iteration (see Section 2.2.1). The algorithm was stopped when the change in the varational lower bound was small enough.
Since the model can get stuck in local minima, we ran 30 trials of the model and selected the one with the highest lower bound.

### 7.3.4 Pathway enrichment analysis

The biological relevance of the infered latent variables was assessed by a gene ontology enrichment analysis on the corresponding weight vectors.
The approach we followed is described in more detail in [14] and was implemented as follows: first, we obtained a gene statistic that quantifies the association between a latent variable and a gene. Here, we used the absolute value of the loading in the weight matrix.
Subsequently, we computed a gene set statistic that quantifies the association between a latent variable and a gene set (a pathway) by taking the average value of the gene-level statistic across all genes annotated in the gene set.
Finally, we tested statistical significance by permutation test as follows: we generated a null distribution for the gene set statistic by permuting the gene set statistics (the weights). For each permutation, we recomputed all geneset statistics. A p-value was obtained for each geneset by calculating the proportion of permutations that obtained geneset statistics more extreme than the observed one.

## 7.4 Software

scGFA was implemented in R version 3.30 and most plots were generated with ggplot version 2.1.0.

# Bibliography


Ahuja, AK et al. (2016). "A short G1 phase imposes constitutive replication stress and fork remodelling in mouse embryonic stem cells". In: *Nat Commun* 17, p. 10660.

Aken, BL et al. (2016). "The Ensembl gene annotation system". In: Database (Oxford): baw093.

Altun, G, JF Loring, and LC Laurent (2010). "DNA methylation in embryonic stem cells". In: *J Cell Biochem* 109 (1), pp. 1–6.

Angermueller, C et al. (2016). "Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity". In: *Nat Methods* 13 (3), pp. 229–32.

Bach, RF and MI Jordan (2005). *A Probabilistic Interpretation of Canonical Correlation Analysis*. Technical Report 688. Department of Statistics, University of California, Berkeley.

Beal, JM (2003). "Variational algorithms for approximate bayesian inference". University College London.

Bishop, CM (1999). "Variational Principal Components". In: *ICAN* 1, pp. 509–14.

— (2006). *Pattern Recognition and Machine Learning*. Springer.

Buettner et al. (2015). "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells". In: *Nat Biotechnol* 33 (2), pp. 155–60.

Chang, X, T Dacheng, and X Chao (2013). "A Survey on Multi-view Learning". In: arXiv: 1304.5634.

Creyghton, MP (2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state". In: *Proc Natl Acad Sci USA* 107 (50), pp. 21931–6.

Dempster, AP, NM Laird, and DB Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society* 39, pp. 1–38.

Finak, G et al. (2015). "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data". In: *Genome Biol* 16, p. 278.

Frost, HR, Z Li, and JH Moore (2014). "Principal component gene set enrichment (PCGSE)". In: arXiv: 1403.5148.

Gawad, C, W Koh, and SR Quake (2016). "Single-cell genome sequencing: current state of the science". In: *Nat Rev Genet* 17 (3), pp. 175–88.

Hardle, W and L Simar (2007). *Applied Multivariate Statistical Analysis*. Springer, pp. 321–30. ISBN: 978-3-540-72243-4.

Harman, HH (1976). *Modern Factor Analysis*. University of Chicago Press.

Illingworth, RS et al. (2010). "Orphan CpG islands identify numerous conserved promoters in the mammalian genome". In: *PLoS Genet* 6 (9), pp. 1571–2. DOI: 10.1371/journal.pgen.1001134.

Kolodziejczyk, AA et al. (2015). "Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation". In: *Cell Stem Cell* 17 (4), pp. 471–85.

Krueger, F and SR Andrews (2011). "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". In: *Bioinformatics* 27 (11), pp. 1571–2.

Liao, Y, GK Smyth, and W Shi (2014). "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 7 (30), pp. 923–30.

Lun, AT, K Bach, and JC Marioni (2016). "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". In: *Genome Biol* 17, p. 75.

Luttinen, J and A Ilin (2010). "Transformations in variational Bayesian factor analysis to speed up learning". In: *Neurocomputing* 73, pp. 1093–102.

MacKay, DJ (1991). "Bayesian interpolation". In: *Neural Computation* 4, pp. 415–47.

Murphy, KP (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Nathoo, FS et al. (2013). "Comparing variational Bayes with Markov chain Monte Carlo for Bayesian computation in neuroimaging". In: *Stat Methods Med Res*. 4th ser. 22, pp. 398–423.

Pierson, E and C Yau (2015). "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis". In: *Genome Biol* 16, p. 241.

Proserpio, V et al. (2016). "Single-cell analysis of CD4+ T-cell differentiation reveals three major cell states and progressive acceleration of proliferation". In: *Genome Biol* 17.1, p. 103.

Seeger, M and Guillaume Bouchard (2012). "Fast variational bayesian inference for non-conjugate matrix factorization models". In: Proceedings of the 15th international conference on artificial intelligence and statistics.

Singer, ZS et al. (2014). "Dynamic heterogeneity and DNA methylation in embryonic stem cells". In: *Mol Cell* 55 (2), pp. 319–31.

Smallwood, SA et al. (2014). "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity". In: *Nat Methods* 11 (8), pp. 817–20.

Smith, ZD and A Meissner (2013). "DNA methylation: roles in mammalian development". In: *Nat Rev Genet* 14 (3), pp. 204–20.

Stegle, O, SA Teichmann, and JC Marioni (2015). "Computational and analytical challenges in single-cell transcriptomics". In: *Nat Rev Genet* 16 (3), pp. 133–45.

Tipping, ME and CM Bishop (1999). "Probabilistic Principal Component Analysis". In: *Journal of the Royal Statistical Society* 61 (3), pp. 611–22.

Titsias, MK and Guillaume M Lázaro-Gredilla (2012). "Spike and slab variational inference for multi-task and multiple kernel learning". In: Advances in neural information processing systems, pp. 2339–47.

Tosolini, M and A Jouneau (2016). "Acquiring Ground State Pluripotency: Switching Mouse Embryonic Stem Cells from Serum/LIF Medium to 2i/LIF Medium". In: *Methods Mol Biol* 1341, pp. 41–8.

Virtanen, S, A Klami, and S Kaski (2011). "Bayesian CCA via group sparsity". In: Proceedings of the 28th Inter- national Conference on Machine Learning (ICML).

Virtanen, S et al. (2012). "Bayesian group factor analysis". In: Proc. 15th Int. Conf. Artificial Intelligence and Statistics, pp. 1269–77.

Wang, C (2007). "Variational Bayesian Approach to Canonical Correlation Analysis". In: *IEEE Trans Neural Netw* 3.18.

Wu, TD and S Nacu (2010). "Fast and SNP-tolerant detection of complex variants and splicing in short reads". In: *Bioinformatics* 26 (7), pp. 873–81.

# Chapter 8

# Appendix

## 8.1 Prior distributions and likelihood

Latent variables:

$$\log P(\mathbf{Z}) = -\frac{NK}{2}\log(2\pi) - \frac{1}{2}\sum_{n=1}^{N}\|\mathbf{z}_{:,n}\|$$

Precision of the zero-centered normally-distributed noise:

$$\log P(\boldsymbol{\tau}) = \sum_{m=1}^{M}\sum_{d=1}^{D_m}\Big\{a_0^\tau \log b_0^\tau + (a_0^\tau - 1)\log \tau_d^m - b_0^\tau \tau_d^m - \log\Gamma(a_0^\tau)\Big\}$$

Weight matrix :

$$\log P(\mathbf{W}|\boldsymbol{\alpha}) = \sum_{m=1}^{M} -\frac{KD_m}{2}\log(2\pi) + \sum_{m=1}^{M}\frac{KD_m}{2}\log(\alpha_k^m) - \sum_{m=1}^{M}\sum_{k=1}^{K}\frac{\alpha_k^m}{2}(\|\mathbf{w}_{:,k}^m\|)$$

mean:

$$\frac{1}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m} -\log(2\pi) + \log(\beta) + \beta(\mu_d^m)^2$$

alpha:

$$\log P(\boldsymbol{\alpha}) = \sum_{m=1}^{M}\sum_{k=1}^{K}\Big\{a_0^\alpha \log b_0^\alpha + (a_0^\alpha - 1)\log \alpha_k^m - b_0^\alpha \alpha_k^m - \log\Gamma(a_0^\alpha)\Big\}$$

Likelihood:

$$\begin{aligned}\log P(\mathbf{Y}|\mathbf{Z},\mathbf{W},\boldsymbol{\tau}) &= -\frac{1}{2}\Big(\sum_{m=1}^{M}\sum_{n=1}^{N}\Big\{D_m\log(2\pi) - \log(|\mathbf{T}^m|) + (\mathbf{y}_{:,n}^m - \mathbf{W}^m\mathbf{z}_{:,n})^T(\mathbf{T}^m)(\mathbf{y}_{:,n}^m - \mathbf{W}^m\mathbf{z}_{:,n} - \boldsymbol{\mu}^m)\Big\}\\ &= -\frac{1}{2}\Big(\sum_{m=1}^{M} ND_m\log(2\pi) - N\sum_{m=1}^{M}\sum_{d=1}^{D_m}\log(\tau_d^m) + \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\tau_d^m\|(y_{d,n}^m - \mathbf{w}_{d,:}^m\mathbf{z}_{:,n} - \mu_d^m)\|\Big)\end{aligned}$$

## 8.2 Derivation of the variational updates

### Latent variables (Z)

To derive the variational updates for $\mathbf{Z}$, we write the expected marginal log likelihood $\mathbb{E}[\mathcal{L}(\boldsymbol{\Theta},\mathbf{Z})]$ with respect to the variational posterior distributions and neglect terms not depending on $\mathbf{Z}$. Subsequently, we obtain the distribution $Q(\mathbf{Z})$ by visual inspection.

Term from $\log p(\mathbf{Z})$:

$$-\frac{1}{2}\sum_{n=1}^{N}\|\mathbf{z}_{:,n}\| + \text{const.}$$

Term from $\log P(\mathbf{Y}|\mathbf{Z},\mathbf{W},\boldsymbol{\tau},\boldsymbol{\mu})$

$$-\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\Big\{\langle(\mathbf{y}_{:,n}^m - \mathbf{W}^m\mathbf{z}_{:,n} - \boldsymbol{\mu}^m)^T(\mathbf{T}^m)(\mathbf{y}_{:,n}^m - \mathbf{W}^m\mathbf{z}_{:,n} - \boldsymbol{\mu}^m)\rangle\Big\} + \text{const.}$$

Rewriting everything together:

$$-\frac{1}{2}\sum_{n=1}^{N}\Big\{-\sum_{m=1}^{M}\Big((\mathbf{y}_{:,n}^{mT}-\langle\boldsymbol{\mu}^m\rangle)\langle\mathbf{T}^m\rangle\langle\mathbf{W}^m\rangle\Big)\mathbf{z}_{:,n}-\sum_{m=1}^{M}\mathbf{z}_{:,n}^{T}\Big(\langle\mathbf{W}^{mT}\rangle\langle\mathbf{T}^m\rangle(\mathbf{y}_{:,n}^{m}-\langle\boldsymbol{\mu}^m\rangle)\Big) \tag{8.1}$$

$$+\mathbf{z}_{:,n}^{T}\Big(\mathbf{I}_k+\sum_{m=1}^{M}\langle\mathbf{W}^{mT}\mathbf{T}^m\mathbf{W}^m\rangle\Big)\mathbf{z}_{:,n}\Big\}+const \tag{8.2}$$

from which we can infer that $q(\mathbf{Z})$ is of the form

$$q(\mathbf{Z})=\prod_{n=1}^{N}q(\mathbf{z}_n)=\prod_{n=1}^{N}\mathcal{N}(\mathbf{z}_{:,n}|\mathbf{m}_z^n,\boldsymbol{\Sigma}_z)$$

where

$$\boldsymbol{\Sigma}_z=\left(\mathbf{I}_k+\sum_{m=1}^{M}\langle\mathbf{W}^{mT}\mathbf{T}^m\mathbf{W}^m\rangle\right)^{-1}$$

$$\mathbf{m}_z^n=\sum_{m=1}^{M}\boldsymbol{\Sigma}_z\langle\mathbf{W}^m\rangle^T\langle\mathbf{T}^m\rangle(\mathbf{y}_{:,n}^m-\boldsymbol{\mu}^m)$$

where:
$\langle\mathbf{T}_m\rangle$ is a diagonal $D_m\times D_m$ matrix where the $d$th element is equal to $\langle\tau_d^m\rangle=\frac{a_{dm}^{\tau}}{b_{dm}^{\tau}}$
$\langle\mathbf{W}^m\rangle$ is a matrix with the row-wise concatenation of $\mathbf{m}_w^{dm}$

Using the fact that $\mathbf{T}^m$ is diagonal, the expectation $\langle\mathbf{W}^{mT}\mathbf{T}^m\mathbf{W}^m\rangle$ can be calculated as follows:

$$\langle\mathbf{W}^{mT}\mathbf{T}^m\mathbf{W}^m\rangle=\sum_{d=1}^{D_m}\langle\tau_d^m\rangle\langle\mathbf{w}_{d,:}^{mT}\mathbf{w}_{d,:}^m\rangle$$

## Means (mu)

Term from $\log p(\boldsymbol{\mu})$:

$$-\frac{\beta}{2}\sum_{m=1}^{M}\|\boldsymbol{\mu}^m\|+\text{const.}$$

Term from $\log P(\mathbf{Y}|\mathbf{Z},\mathbf{W},\boldsymbol{\tau},\boldsymbol{\mu})$

$$-\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\langle(\mathbf{y}_{:,n}^m-\mathbf{W}^m\mathbf{z}_{:,n}-\boldsymbol{\mu}^m)^T(\mathbf{T}^m)(\mathbf{y}_{:,n}^m-\mathbf{W}^m\mathbf{z}_{:,n}-\boldsymbol{\mu}^m)\rangle+\text{const.}$$

Rewriting everything together:

$$-\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\Big\{\Big(-\mathbf{y}_{:,n}^{T}\langle\mathbf{T}^m\rangle\Big)\boldsymbol{\mu}^m+\Big(\langle(\mathbf{W}^m\mathbf{z}_{:,n})^T\mathbf{T}^m\rangle\Big)\boldsymbol{\mu}^m-\boldsymbol{\mu}^{mT}\Big(\langle\mathbf{T}^m\rangle\mathbf{y}_{:,n}\Big) \tag{8.3}$$

$$+\boldsymbol{\mu}^{mT}\Big(\langle\mathbf{T}^m\mathbf{W}^m\mathbf{z}_{:,n}\rangle\Big)\Big\}+\boldsymbol{\mu}^{mT}\Big(\beta\mathbf{I}_{D_m}+N\langle\mathbf{T}^m\rangle\Big)\boldsymbol{\mu}^m+\text{const.} \tag{8.4}$$

from which we can infer that $q(\boldsymbol{\mu})$ is of the form

$$q(\boldsymbol{\mu})=\prod_{m=1}^{M}\mathcal{N}\left(\boldsymbol{\mu}^m\,|\,\mathbf{m}_\mu^m,\boldsymbol{\Sigma}_\mu^m\right)$$

where

$$\boldsymbol{\Sigma}_\mu^m=\Big(\beta\mathbf{I}_{D_m}+N\langle\mathbf{T}^m\rangle\Big)^{-1}$$

$$\mathbf{m}_\mu^m=\boldsymbol{\Sigma}_\mu^m\langle\mathbf{T}^m\rangle\sum_{n=1}^{N}\langle\mathbf{y}_{:,n}-\mathbf{W}^m\mathbf{z}_{:,n})\rangle$$

## ARD precision (alpha)

To derive the variational updates for $\boldsymbol{\alpha}$, we write the expected marginal log likelihood $\mathbb{E}[\mathcal{L}(\boldsymbol{\Theta}, \mathbf{Z})]$ with respect to the variational posterior distributions and neglect terms not depending on $\boldsymbol{\alpha}$. Subsequently, we obtain the distribution $Q(\boldsymbol{\alpha})$ by inspection.

Term from the $\log P(\boldsymbol{\alpha})$:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}(a_0^\alpha - 1)\log(\alpha_k^m) - b_0^\alpha \alpha_k^m + \text{const.}$$

Term from the $\log P(\mathbf{W})$:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{D_m}{2}\log(\alpha_k^m) - \frac{\alpha_k^m}{2}(\langle\|\mathbf{w}_{:,k}^m\|\rangle) + \text{const.}$$

Writing everything together:

$$\sum_{k=1}^{K}\sum_{m=1}^{M}\left\{\left(a_0^\alpha + \frac{D_m}{2} - 1\right)\log\alpha_k^m - \left(b_0^\alpha + \frac{1}{2}\langle\|\mathbf{w}_{:,k}^m\|\rangle\right)\alpha_k^m\right\} + \text{const.}$$

from which we can infer that $q(\boldsymbol{\alpha})$ is of the form

$$q(\boldsymbol{\alpha}) = \prod_{m=1}^{M}\prod_{k=1}^{K} q(\alpha_k^m) = \prod_{m=1}^{M}\prod_{k=1}^{K}\mathcal{G}(\alpha_k^m|\hat{a}_{mk}^\alpha, \hat{b}_{mk}^\alpha)$$

where

$$\hat{a}_{mk}^\alpha = a_0^\alpha + \frac{D_m}{2}$$

$$\hat{b}_{mk}^\alpha = b_0^\alpha + \frac{\langle\|\mathbf{w}_{:,k}^m\|\rangle}{2}$$

The expectation $\langle\|\mathbf{w}_{:,k}^m\|\rangle$ can be calculated as follows:

$$\langle\|\mathbf{w}_{:,k}^m\|\rangle = \Big(\sum_{d=1}^{D_m}\boldsymbol{\Sigma}_w^m + \mathbf{m}_w^{dm}\mathbf{m}_w^{dmT}\Big)_{(k,k)}$$

## Noise precision (tau)

To derive the variational updates for $\boldsymbol{\tau}$, we write the expected complete log likelihood $\mathbb{E}[\mathcal{L}(\boldsymbol{\Theta}, \mathbf{Z})]$ with respect to the variational posterior distribution and neglect terms not depending on $\boldsymbol{\tau}$. Subsequently, we obtain the distribution $Q(\boldsymbol{\tau})$ by inspection.

Term from $\log p(\mathbf{T})$

$$\sum_{m=1}^{M}\sum_{d=1}^{D_m}(a_0^\tau - 1)\log\tau_d^m - b_0^\tau\tau_d^m + \text{const.}$$

Term from $\log P(\mathbf{Y}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau}, \boldsymbol{\mu})$

$$\frac{N}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\log(\tau_d^m) - \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{d=1}^{D_m}\frac{\tau_d^m}{2}\langle(y_{d,n}^m - \mathbf{w}_{d,:}^m\mathbf{z}_{:,n} - \mu_d^m)^2\rangle + \text{const.}$$

Rewriting everything together:

$$\sum_{m=1}^{M}\sum_{d=1}^{D_m}\left(a_0^\tau - 1 + \frac{N}{2}\right)\log\tau_d^m - \left(b_0 + \frac{1}{2}\sum_{n=1}^{N}\langle(y_{dn}^m - \mathbf{w}_{d,:}^m\mathbf{z}_{:,n} - \mu_d^m)^2\rangle\right)\tau_d^m + \text{const.}$$

from which we can infer that $q(\boldsymbol{\tau})$ is of the form

$$\prod_{m=1}^{M}\prod_{d=1}^{D_m}\mathcal{G}(\tau_d^m|\hat{a}_{dm}^\tau, \hat{b}_{dm}^\tau)$$

where

$$\hat{a}^{\tau}_{dm} = a^{\tau}_0 + \frac{N}{2}$$

$$\hat{b}^{\tau}_{dm} = b^{\tau}_0 + \frac{1}{2}\sum_{n=1}^{N}\langle (y^m_{dn} - \mathbf{w}^m_{d,:}\mathbf{z}_{:,n} - \mu^m_d)^2\rangle$$

The expectation can be calculated as follows:

$$\begin{aligned}\langle (y^m_{d,n} - \mathbf{w}^m_{d,:}\mathbf{z}_{:,n} - \mu^m_d)^2\rangle =& (\langle \mu^m_d\rangle)^2 + (y^m_{d,n})^2 - 2y^m_{d,n}\langle \mathbf{w}^m_{d,:}\rangle\langle \mathbf{z}_{:,n}\rangle - 2y^m_{d,n}\langle \mu^m_d\rangle \\ &+ 2\langle \mu^m_d\rangle\langle \mathbf{w}^m_{d,:}\rangle\langle \mathbf{z}_{:,n}\rangle + tr\Big[\langle \mathbf{w}^m_{d,:}\mathbf{w}^{mT}_{d,:}\rangle\langle \mathbf{z}_{:,n}\mathbf{z}^T_{:,n}\rangle\Big]\end{aligned}$$

### Weights (W)

To derive the variational updates for $\mathbf{W}$, we write the expected marginal log likelihood $\mathbb{E}[\mathcal{L}(\mathbf{\Theta}, \mathbf{Z})]$ with respect to the variational posterior distribution and neglect terms not depending on $\mathbf{W}$. Furthermore, we rewrite the expressions as a sum over the rows of $\mathbf{W}$:

Term from $\log p(\mathbf{W})$

$$-\frac{1}{2}\sum_{m=1}^{M}\sum_{k=1}^{K}\langle \alpha^m_k\rangle(\|\mathbf{w}^m_{:,k}\|) + \text{const.} \tag{8.5}$$

$$= -\frac{1}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\langle \mathbf{w}^m_{d,:}\hat{\boldsymbol{\alpha}}^m\mathbf{w}^{mT}_{d,:}\rangle + \text{const.} \tag{8.6}$$

Term from $\log P(\mathbf{Y}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau}, \boldsymbol{\mu})$

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\langle \tau^m_d\rangle\langle \|(y^m_{d,n} - \mathbf{w}^m_{d,:}\mathbf{z}_{:,n} - \mu^m_d)\|\rangle + const$$

$$= -\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{d}^{D_m}\langle \tau^m_d\rangle\Big\langle\Big(-\mathbf{w}^m_{d,:}\mathbf{z}_{:,n}(y^m_{d,n} - \mu^m_d) - (\mathbf{w}^m_{d,:}\mathbf{z}_{:,n})^T(y^m_{d,n} - \mu^m_d) + (\mathbf{w}^m_{d,:}\mathbf{z}_{:,n})^T(\mathbf{w}^m_{d,:}\mathbf{z}_{:,n})\Big)\Big\rangle + \text{const.}$$

Rewriting all together:

$$\sum_{m=1}^{M}\sum_{d}^{D_m}\Big\{\mathbf{w}^m_{d,:}\Big(\langle \tau^m_d\rangle\sum_{n=1}^{N}\langle \mathbf{z}_{:,n}\rangle(y^m_{dn} - \langle \mu^m_d\rangle)\Big) - \frac{1}{2}\mathbf{w}^m_{d,:}\Big(\langle \tau^m_d\rangle\sum_{n=1}^{N}\langle \mathbf{z}_{:,n}\mathbf{z}^T_{:,n}\rangle + \langle \hat{\boldsymbol{\alpha}}^m\rangle\Big)\mathbf{w}^{mT}_{d,:}\Big\} + \text{const.}$$

from which we can infer that $q(\mathbf{W})$ is of the form

$$q(\mathbf{W}) = \prod_{m=1}^{m}\prod_{d=1}^{D_m}\mathcal{N}(\mathbf{w}^m_{d,:}|\mathbf{m}^{dm}_w, \mathbf{\Sigma}^{dm}_w)$$

where

$$\mathbf{\Sigma}^{dm}_w = \Big(\langle \tau^m_d\rangle\sum_{n=1}^{N}\langle \mathbf{z}_{:n}\mathbf{z}^T_{:n}\rangle + \langle \overline{\boldsymbol{\alpha}}^m\rangle\Big)^{-1}$$

$$\mathbf{m}^{dm}_w = \mathbf{\Sigma}^{dm}_w\langle \tau^m_d\rangle\Big(\sum_{n=1}^{N}(y^m_{dn} - \langle \mu^m_d\rangle)\langle \mathbf{z}_{:,n}\rangle\Big)$$

where $\overline{\boldsymbol{\alpha}}$ is a diagonal $K \times K$ matrix with the corresponding $\alpha$ values as diagonal elements.

## 8.3 Evidence lower bound

The evidence lower bound for the scGFA model can be calculated as follows:

$$\mathbb{E}[\log p(\mathbf{Y}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau})] + \mathbb{E}[\log p(\mathbf{Z})] - \mathbb{E}[\log q(\mathbf{Z})] + \mathbb{E}[\log p(\boldsymbol{\mu})] - \mathbb{E}[\log q(\boldsymbol{\mu})] + \mathbb{E}[\log p(\boldsymbol{\alpha})] \tag{8.7}$$

$$-\mathbb{E}[\log q(\boldsymbol{\alpha})] + \mathbb{E}[\log p(\boldsymbol{\tau})] - \mathbb{E}[\log q(\boldsymbol{\tau})] + [\log p(\mathbf{W})] - [\log q(\mathbf{W})] \tag{8.8}$$

where the expectations are taken with respect to the variational distributions.
The following sectionss how how to compute each term.

### Likelihood

$$\mathbb{E}[\log p(\mathbf{Y}|\mathbf{Z},\mathbf{W},\boldsymbol{\tau})] = -\sum_{m=1}^{M}\frac{ND_m}{2}\log(2\pi)+\frac{N}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\langle\log(\tau_d^m)\rangle-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\langle\tau_d^m\rangle\langle\|(y_{d,n}^m-\mathbf{w}_{d,:}^m\mathbf{z}_{:,n}-\mu_d^m)\|\rangle$$

Using the update from $\boldsymbol{\tau}$, we can rewrite the expression as follows:

$$\mathbb{E}[\log p(\mathbf{Y}|\mathbf{Z},\mathbf{W},\boldsymbol{\tau})] = -\sum_{m=1}^{M}\frac{ND_m}{2}\log(2\pi)+\frac{N}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\langle\log(\tau_d^m)\rangle-\sum_{m=1}^{M}\sum_{d=1}^{D_m}\langle\tau_d^m\rangle(\hat{b}_{dm}^{\tau}-b_0^{\tau})$$

### Z

$$\mathbb{E}[\log P(\mathbf{Z})] = -\frac{NK}{2}\log(2\pi)-\frac{1}{2}\sum_{n=1}^{N}\langle\|\mathbf{z}_{:,n}\|\rangle$$

$$\mathbb{E}[\log Q(\mathbf{Z})] = -\frac{N}{2}\log(|\boldsymbol{\Sigma}_z|)-\frac{NK}{2}(1+\log(2\pi))$$

### mu

$$\mathbb{E}[\log P(\boldsymbol{\mu})] = \sum_{m=1}^{M}\Big\{-\frac{D_m}{2}\log(2\pi)+\frac{D_m}{2}\log(\beta)-\frac{\beta}{2}\langle\|\boldsymbol{\mu}^m\|\rangle\Big\}$$

$$\mathbb{E}[\log Q(\boldsymbol{\mu})] = -\frac{1}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\log(|\boldsymbol{\Sigma}_{\mu}^{dm}|)-\sum_{m=1}^{M}\frac{D_m}{2}(1+\log(2\pi))$$

### alpha

$$\mathbb{E}[\log P(\boldsymbol{\alpha})] = MKa_0^{\alpha}\log b_0^{\alpha}+\sum_{m=1}^{M}\sum_{k=1}^{K}(a_0^{\alpha}-1)\langle\log\alpha_k^m\rangle-\sum_{m=1}^{M}\sum_{k=1}^{K}b_0^{\alpha}\langle\alpha_k^m\rangle-MK\log\Gamma(a_0^{\alpha})$$

$$\mathbb{E}[\log Q(\boldsymbol{\alpha})] = \sum_{m=1}^{M}\sum_{k=1}^{K}\hat{a}_{mk}^{\alpha}\log\hat{b}_{mk}^{\alpha}+(\hat{a}_{mk}^{\alpha}-1)\langle\log\alpha_k^m\rangle-\hat{b}_{mk}^{\alpha}\langle\alpha_k^m\rangle-\log\Gamma(\hat{a}_{mk}^{\alpha})$$

where the expectations are calculated as follows:

$$\langle\alpha_k^m\rangle = \frac{\tilde{a}_{mk}^{\alpha}}{\tilde{b}_{mk}^{\alpha}}$$

$$\langle\log\alpha_k^m\rangle = \psi(\tilde{a}_{mk}^{\alpha})-\log\tilde{b}_{mk}^{\alpha}$$

### tau

$$\mathbb{E}[\log P(\tau)] = \sum_{m=1}^{M}D_m a_0^{\tau}\log b_0^{\tau}+\sum_{m=1}^{M}\sum_{d=1}^{Dm}(a_0^{\tau}-1)\langle\log\tau_d^m\rangle-\sum_{m=1}^{M}\sum_{d=1}^{Dm}b_0^{\tau}\langle\tau_d^m\rangle-MD\log\Gamma(a_0^{\tau})$$

$$\mathbb{E}[\log Q(\boldsymbol{\tau})] = \sum_{m=1}^{M}\sum_{d=1}^{D_m}\hat{a}_{dm}^{\tau}\log\hat{b}_{dm}^{\tau}+(\hat{a}_{dm}^{\tau}-1)\langle\log\tau_d^m\rangle-\hat{b}_{dm}^{\tau}\langle\tau_d^m\rangle-\log\Gamma(\hat{a}_{dm}^{\tau})$$

where the expectations are calculated as follows:

$$\langle\tau_d^m\rangle = \frac{\tilde{a}_{md}^{\tau}}{\tilde{b}_{md}^{\tau}}$$

$$\langle\log\tau_d^m\rangle = \psi(\tilde{a}_{md}^{\tau})-\log\tilde{b}_{md}^{\tau}$$

## W

$$\mathbb{E}[\log P(\mathbf{W})] = -\sum_{m=1}^{M} \frac{KD_m}{2} \log(2\pi) + \frac{1}{2} \sum_{m=1}^{M} \sum_{k=1}^{K} D_m \langle \log(\alpha_k^m) \rangle - \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle \mathbf{w}_{d,:}^m \hat{\boldsymbol{\alpha}}^m \mathbf{w}_{d,:}^{mT} \rangle$$

$$\mathbb{E}[\log Q(\mathbf{W})] = -\sum_{m=1}^{M} \frac{KD_m}{2} (1 + \log(2\pi)) - \frac{1}{2} \sum_{m=1}^{M} \sum_{d=1}^{D_m} \log(|\boldsymbol{\Sigma}_w^{dm}|)$$

where the expectations are calculated as follows:

$$\langle \log(\alpha_k^m) \rangle = \psi(\tilde{a}_{mk}^{\alpha}) - \log \tilde{b}_{mk}^{\alpha}$$

$$\langle \mathbf{w}_{d,:}^m \hat{\boldsymbol{\alpha}}^m \mathbf{w}_{d,:}^{mT} \rangle = \langle \hat{\boldsymbol{\alpha}}^m \rangle \langle \mathbf{w}_{d,:}^{mT} \mathbf{w}_{d,:}^m \rangle$$

# Acknowledgements

First of all, I would like to thank my two thesis advisors, Professor Anders Krogh and Dr. Oliver Stegle. Thank you Anders for the flexibility of allowing me to do my thesis in another institution and thank you Oliver for introducing me in this exciting field and steering me in the right direction whenever I needed it.

Second, my deepest and sincere gratitude to Dr. Florian Buettner for his excellent guidance during the entire internship. This thesis would have not been possible without your help. Thanks.
I also want to thank everyone in the Stegle Lab for fruitful discussions and technical help, specially Damien Arnol, Dr. Danilo Horta and Dr. Davis McCarthy.

Third, a warm acknowledgement to Obra Social La Caixa for providing funding for my entire MSc.

Finalment, donar gràcies a en Moisès Coll per la seva valuosa ajuda, i a en Pattrick Rothfuss per escriure *The Kingkiller Chronicle*.

Ricard