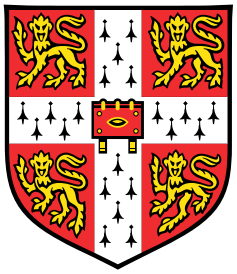
Statistical Methods for the Analysis of Contextual Gene Expression Data



Damien Arnol

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Queens' College

September 2018

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words exclusive of tables, footnotes, bibliography, and appendices.

Damien Arnol
September 2018

Acknowledgements

First and foremost, I would like to thank my co-supervisors Oliver Stegle and Julio Saez-Rodriguez for their guidance and support throughout my PhD. I am also very grateful to my Thesis Advisory Committee: Sarah Teichmann, Lorenz Wernisch and Martin Jechlinger, for their support and their valuable feedback during our meetings. I thank the whole Stegle and Saez-Rodriguez groups, and more specially the people I have worked most closely with during the past four years: Ricard Argelaguet in particular, as well as Danila Bredikhin and Yonatan Deloro. Additionally and in no particular order, I wish to mention a few people with whom insightful discussions inspired my work: Danilo Horta, Paolo Casale, Amélie Baud, Bogdan Mirauta, Attila Gabor, Aurélie Dugourd, Verena Zuber, Florian Buettner, Na Cai, Valentine Svensson, Mi Yang, Marc Jan Bonder, Emanuel Gonçalves, Roser Vento and Britta Velten. I am grateful to Daniel Seaton, Jordan Billiald, Alec Greaves-Tunnell and Jovan Tanevski for their valuable feedback on this thesis, and for Krishna Kumar who provided the LaTeX template. Last but not least, thank you to Amal and to my parents, friends and family for your unconditional support.

Abstract

Technological advances have enabled profiling gene expression variability, both at the RNA and the protein level, with ever increasing throughput. In addition, miniaturisation has enabled quantifying gene expression from small volumes of the input material and most recently at the level of single cells. Increasingly these technologies also preserve context information, such as assaying tissues with high spatial resolution. A second example of contextual information is multi-omics protocols, for example to assay gene expression and DNA methylation from the same cells or samples.

Although such contextual gene expression datasets are increasingly available for both population and single-cell variation studies, methods for their analysis are not established. In this thesis, we propose two modelling approaches for the analysis of gene expression variation in specific biological contexts.

The first contribution of this thesis is a statistical method for analysing single cell expression data in a spatial context. Our method identifies the sources of gene expression variability by decomposing it into different components, each attributable to a different source. These sources include aspects of spatial variation such as cell-cell interactions. In applications to data across different technologies, we show that cell-cell interactions are indeed a major determinant of the expression level of specific genes with a relevant link to their function.

The second contribution is a latent variable model for the unsupervised analysis of gene expression data, while accounting for structured prior knowledge on experimental context. The proposed method enables the joint analysis of gene expression data and other omics data profiled in the same samples, and the model can be used to account for the grouping structure of samples, e.g. samples from individuals with different clinical covariates or from distinct experimental batches. Our model constitutes a principled framework to compare the molecular identities of these distinct groups.

Table of contents

List of figures

Chapter 1

Introduction

Analysing patterns of gene-expression variation across and within individuals is key to understand the molecular basis of phenotypic diversity and diseases. Microarrays were the first technology to provide genome-wide gene expression profiles, offering insight into gene expression variation across individuals (Taub et al., 1983; Tarca et al., 2006). Their use for differential expression analysis in case control studies has identified candidate molecular determinants of human diseases (Frolov et al., 2003; Ritchie et al., 2015b).

Microarray protocols are targeted approaches and offer a limited detection range, due to probe saturation as well as high background signal owing to cross-hybridisation. In contrast, the rise of whole transcriptome sequencing provided higher resolution measurements (Wang et al., 2009; Kukurba and Montgomery, 2015; Soneson and Delorenzi, 2013; Chu and Corey, 2012). The non-targeted nature of the approach, providing full transcript sequences, also allowed the additional quantification of single nucleotide mutations (Quinn et al., 2013; Kang et al., 2016), as well as post-transcriptional modifications such as alternative splicing events (Marguerat and Bähler, 2010; Nachtergaele and He, 2017).

More recently, miniaturisation of the protocols and the ability to operate with low volumes of input material has enabled the profiling of gene expression in single-cells (Tang et al., 2009; Svensson et al., 2018b). The analysis of gene expression variation at this level has tremendous implications in multiple fields of biology (Macaulay and Voet, 2014; Kolodziejczyk et al., 2015; Wang and Navin, 2015). Among many other things, it provides a new understanding of the composition of tissues in terms of cell types (Schelker et al., 2017; Hu et al., 2017; Chen et al., 2017), provides insights in developmental and differentiation processes (Macaulay et al., 2016a; Kumar et al., 2017; Griffiths et al., 2018) and enables us to study the dynamics

of transcription (Rafalska-Metcalf et al., 2010; Skinner et al., 2016).

A myriad of new experimental techniques are being developed to measure gene expression profiles in an ever-increasing diversity of contexts. Gene expression is increasingly measured at multiple time points (Androulakis et al., 2007), with spatial resolution (Strell et al., 2018), in the context of multi-omics studies (Hasin et al., 2017) and also across multiple organs for the same samples (GTEx Consortium, 2013). Such contextual experiments provide data with known categorical, hierarchical or continuous dependencies between individual measurements. In spatial assays or time series for example, expression profiles are related continuously by their spatial distance or by the time lapse between their measurement. In multi-omics experiments, measurements fall into categorical data sources. Standard methods for the analysis of gene expression variation are not designed to account for the new dimensions created by this contextual information.

This PhD thesis is concerned with the development of adapted statistical tools to explicitly model this contextual gene expression data, focussing on probabilistic generative models. Generative models are an ideal framework to encode our assumptions about the relationship between observations, including those due to the experimental context, such as the measurement time or spatial location. The probabilistic formulation provides a rigorous framework for inference, where objective functions such as the data likelihood have a well-defined mathematical interpretation, and accounts for uncertainty about the model parameters in a principled manner.

The key contributions of this thesis are twofold. First, we present Spatial Variance Component Analysis (SVCA), a model for the analysis of spatial expression data. Most of the current single-cell technologies require to first isolate the cells from their native context and subsequent analysis therefore ignores their spatial arrangement in the tissue of origin (Hu et al., 2016). This is an important limitation, as tissue function relies on interacting cells rather than isolated components. Technological advances for single-cell expression profiling in tissue context (Strell et al., 2018) are now bringing an opportunity to relate single-cell expression variation to the spatial structure of tissues, and in particular to model cell-cell interactions. At the same time this creates a need for principled statistical tools to account for spatial information in the analysis.

SVCA is a statistical model that decomposes gene expression variation in relation to the cells' spatial context. Specifically, it is based on an additive Gaussian process to model gene expression variation as resulting from multiple drivers, some of which dependent on the spatial context (e.g. cell-cell interactions). The generative modelling approach enables us to encode flexible hypotheses about the effect of the spatial context on gene expression level, as well as other sources of variations. SVCA infers the relative importance of the spatial context as a determinant of gene expression variation, and in particular quantifies the effect of cell-cell interactions.

The second contribution of this thesis is BIO-Factor Analysis Model (biofam), a dimensionality reduction method for the exploratory analysis of gene expression and other biological layers measured in a multi-omics context (Hasin et al., 2017), or in multiple tissues or sample types (GTEx Consortium, 2013). Biofam is a latent variable model of the Factor Analysis family. High dimensional gene expression data is modelled as arising from the additive effects of a small number of latent factors. These factors are often interpretable as representing the activity of known biological processes. This approach is motivated by the observation that gene expression variation is highly structured (low-rank) due to genes acting in a coordinated manner (Szklarczyk et al., 2017).

In biofam, the generative probabilistic modelling approach enables us to model explicitly that the drivers of gene expression variation (i.e. the latent factors) may themselves be context-dependent (e.g. tissue-specific pathways). This is achieved by using hierarchical Bayesian priors that are structured to reflect prior knowledge about the data context. A Bayesian inference scheme then discriminates automatically between context-dependent and context-independent factors. Biofam unifies existing Factor Analysis models within a coherent inference and software framework, with the addition of new model features. It is powered by an efficient implementation and inference method and comes with user-friendly tools for downstream analysis and visualisation of the results.

In Chapter 2, we lay the theoretical foundations of the modelling approaches forming the core of SVCA and biofam models.

In Chapter 3, we present SVCA. We first introduce the biological context with a short review of the experimental and computational state of the art in spatial gene expression analysis. We then describe the SVCA model in detail, highlight its differences with related models,

and validate it on simulated data. Finally, we present applications of SVCA to two different biological systems and data from different technologies and discuss the biological relevance of the results.

In Chapter 4, we present biofam, a modular software that allows to fit existing Factor Analysis models as well as new extensions of those. We validate our model using simulated data, highlighting strengths and weaknesses of the method in different simulation settings. We then demonstrate the efficiency of our inference scheme and introduce an extension using stochastic inference for future applications of biofam to even larger datasets.

Chapter 5 illustrates two use cases of biofam, first on a multi-omics dataset, second on data consisting of distinct sample groups corresponding to different biological contexts. We demonstrate that biofam is a powerful tool for the exploratory analysis of data in a structured context, capable of providing a comprehensive overview of interpretable drivers of variation in a single unsupervised analysis. We reproduce multiple results from the literature and show that biofam can be used to unveil novel molecular drivers of heterogeneity.

Finally, Chapter 6 gives a summary of this thesis and an outlook of future research.

Chapter 2

Theoretical foundations

This chapter introduces the two main classes of Machine Learning models this thesis is based on: Gaussian Processes and Factor Analysis, which we first introduce independently, respectively in Section 2.1 and Section 2.2, before addressing a connection between these models in Section 2.3. The aim is to lay out the conceptual foundations of the specific modelling approaches developed in later chapters.

2.1 Gaussian Processes

The first part of this chapter introduces Gaussian Processes (GPs), a class of kernel methods of broad applicability in Machine Learning. I first introduce GPs from the perspective of linear regression, which can be regarded as a special case of the more general class of Gaussian Process regression models. This introduction is largely inspired from the Gaussian Process textbook of Rasmussen and Williams (2006), which may be consulted for further explanations or references.

2.1.1 Linear least squares regression

Let us consider the task of predicting the value of a one dimensional output variable y from a D-dimensional input x. In linear regression, we model the output variable as a linear function f of the input variable x, assuming additive Gaussian noise ε:

$$y = f(x) + \varepsilon = \sum_{d=1}^{D} x_d w_d + \varepsilon = x^T w + \varepsilon \tag{2.1}$$

Here, $x = \{x_d\}_{d\in[\![1;D]\!]}$ is the vector of input variables, $w = \{w_d\}_{d\in[\![1;D]\!]}$ is the vector of the model parameters, controlling the effect size of each input variable. $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is a Gaussian observational noise, with residual variance σ_ε^2, which results in the following likelihood for every data point: $p(y|x,w) = \mathcal{N}\left(y|\sum_{d=1}^{D} x_d w_d, \sigma_\varepsilon^2\right)$

Given N observed samples, we note Y the vector of length N of their output values and X the matrix of dimensions $N \times D$ of their input features. Fitting a linear regression consists in determining the value of the vector w which maximises the likelihood of the data given the independent identically distributed Gaussian noise:

$$P(Y|X,w) = \prod_{i=1}^{N} P(y_i|X_{i,:},w) = \mathcal{N}\left(Y|Xw, \sigma_\varepsilon^2 I\right) \tag{2.2}$$

Writing down the log of this likelihood, we find that maximising the data likelihood is equivalent to minimising the sum of the squared differences between the true values of the target variable y and the values predicted by the linear model:

$$\ln P(Y|X,w) \propto -\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{N} (y_i - X_{i,:}w)^2 \tag{2.3}$$

This is the quadratic loss function used in least squares regression. The fitted weight vector $\hat{w}$ that maximises the data likelihood can be used to predict the output value y^* given the input x^* of a test datapoint.

2.1.2 Bayesian linear regression

As seen before, the solution to the least squares regression problem corresponds to the maximum likelihood estimate of the regression weights assuming a Gaussian likelihood. Alternatively, one may consider a Bayesian analysis of the linear regression, where the weight vector w is modelled as a random variable with a Gaussian prior distribution $P(w) \sim \mathcal{N}(0, \Sigma_w)$. The Bayes rule gives rise to the posterior distribution of w given the data and the prior distribution (Eq. 2.4). The computation is analytically tractable here because a product of two Gaussian probability density functions (pdfs) is proportional to a Gaussian pdf, and the denominator can be rewritten as a convolution between two Gaussian pdfs which is also proportional to a Gaussian pdf (Rasmussen and Williams, 2006, Chap. 2). The resulting posterior distribution is the following Normal distribution[1]:

[1] As the posterior distribution is from the same type as the prior, we say that the Gaussian prior distribution on the weights is conjugate for the Gaussian likelihood.

$$\begin{aligned} P(w|Y,X) &= \frac{P(Y|X,w)P(w)}{\int_w P(Y|X,w)P(w)\mathrm{d}w} \\ &= \mathcal{N}\left(w\middle|\frac{1}{\sigma_\varepsilon^2}\left(\sigma_\varepsilon^{-2}X^TX+\Sigma_w^{-1}\right)^{-1}X^TY, \sigma_\varepsilon^{-2}X^TX+\Sigma_w^{-1}\right) \end{aligned} \tag{2.4}$$

The hyperparameters σ_ε and Σ_w are fixed and chosen a priori. The problem of their optimisation is addressed in Section 2.1.8.

Note: Maximum A Posteriori and Ridge regression

Note that in the special case of an independent prior on the regression weights, $P(w) \sim \mathcal{N}(0, \sigma_w^2 I)$, the Maximum A Posteriori (MAP) solution of the Bayesian linear regression corresponds to the solution of the linear regression using Ridge regularisation. This can be shown by taking the log of the Bayes rule for $P(w|Y,X)$, which gives the objective function of Ridge regression, where the penalisation term $\frac{1}{2\sigma_w^2}\sum_{d=1}^{D} w_d^2$ comes from the Gaussian prior on w:

$$\ln(P(w|Y,X)) \propto -\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}(y_i - X_{i,:}w)^2 - \frac{1}{2\sigma_w^2}\sum_{d=1}^{D} w_d^2 \tag{2.5}$$

This highlights the link between standard regularisation methods in linear regression and Bayesian analysis.

2.1.3 Predictive distribution in Bayesian Linear regression

Given test data points with input matrix X^*, the Bayesian treatment of linear regression gives rise to a predictive distribution for the output vector y^* by averaging the distribution $P(y^*|X^*,w)$ across all possible values of w given the posterior distribution of Equation 2.4:

$$\begin{aligned} P(y^*|X^*,X,Y) &= \int_w P(y^*| \quad X^*,w)P(w|X,Y)\mathrm{d}w \\ &= \mathcal{N}\left(y^*\middle| \quad \frac{1}{\sigma_\varepsilon^2}X^{*T}\left(\frac{1}{\sigma_\varepsilon^2}X^TX+\Sigma_w^{-1}\right)^{-1}X^TY, \right. \\ &\qquad \left. X^{*T}\left(\frac{1}{\sigma_\varepsilon^2}X^TX+\Sigma_w^{-1}\right)^{-1}X^*\right) \end{aligned} \tag{2.6}$$

Using matrix inversion lemmas and linear algebra, this predictive distribution may be rewritten as on Equation 2.7 (Rasmussen and Williams, 2006, Chap. 2), where, the input X and X^* only appear in the inner products $X\Sigma_w X^T$, $X^*\Sigma_w X^{*T}$ and $X^*\Sigma_w X^T$.

$$P(y^*|X^*,X,Y) = \mathcal{N}\left(y^*\middle|X^*\Sigma_w X^T\left(X\Sigma_w X^T+\sigma_\varepsilon^2 I\right)^{-1}Y,\right.$$
$$\left.X^*\Sigma_w X^{*T}-X^*\Sigma_w X\left(X\Sigma_w X^T+\sigma_\varepsilon^2 I\right)^{-1}X\Sigma_w X^{*T}\right) \tag{2.7}$$

The importance of this identity will become clear in Section 2.1.5 when we show how this linear modelling framework can be extended into Gaussian Process regression.

2.1.4 Feature map

A major limitation of linear regression is the restrictive assumption of a linear relationship between the input and the output variables.

One way to overcome this limitation is to transform the D dimensional input into a set of features of higher dimension $P > D$. For example, if the relationship between an input x and an output y resembles a polynomial function, one can fit a linear regression on the powers of x: $\phi(x) = \{1, x^1, x^2, x^3, ...\}$. The function ϕ is commonly called a feature map and its components ϕ: $\{\phi_0(x) = x^0, \phi_1(x) = x, \phi_2(x) = x^2, \phi_3(x) = x^3\}$ basis functions.

One can then model non-linear functions of the input using standard Bayesian linear regression on the features defined by the mapping function ϕ. The predictive distribution of Bayesian linear regression in this feature space becomes:

$$P(y^*|X^*,X,Y) = \mathcal{N}(y^*|\phi(X^*)\Sigma_w\phi(X)^T\left(\phi(X)\Sigma_w\phi(X)^T+\sigma_\varepsilon^2 I\right)^{-1}Y,$$
$$\phi(X^*)\Sigma_w\phi(X^*)^T-\phi(X^*)\Sigma_w\phi(X)\left(\phi(X)\Sigma_w\phi(X)^T+\sigma_\varepsilon^2 I\right)^{-1}\phi(X)\Sigma_w\phi(X^*)^T) \tag{2.8}$$

where the features appear in the equations in the same inner products as the input in Equation 2.7.

2.1.5 The kernel trick

In all quantities that need to be computed to evaluate the marginal likelihood or calculate the predictive distribution, the dependency to the data can be expressed in terms of the inner product $\phi(X)\Sigma_w\phi(X)^T$. This observation is key, as for some complex transformations ϕ, this inner product will be easier to compute than the explicit features.

Illustration with a feature space of infinite dimension

In order to illustrate the importance of the observation thereof, let us consider the following basis functions (Xing, 2015; MacKay, 1998):

$$\phi_i(x) = \exp\left(-\frac{(x-c_i)^2}{2l^2}\right) \tag{2.9}$$

We consider a linear regression with J basis functions: $y = \sum_{i=1}^{J} \phi_i(x) w_i + \varepsilon$, and put the following priors on the weights: $w_i \sim \mathcal{N}(0, \frac{\sigma_w^2}{J} I)$. In the predictive distribution, the input appears in the following inner product $\phi(X)\Sigma_w \phi(X)^T$, with:

$$\left(\phi(X)\Sigma_w \phi(X)^T\right)_{k,l} = \frac{\sigma_w^2}{J} \sum_{i=1}^{J} \phi_i(x_k)\phi_i(x_l), \forall (k,l) \tag{2.10}$$

Let us now define an infinite set of basis functions ϕ_i, by letting $c_{i+1} - c_i = 1/J$ taking $J \to \infty$, with the limits $c_0 = -\infty$ and $c_\infty = \infty$. The inner product defined above becomes:

$$\begin{aligned}\left(\phi(X)\Sigma_w \phi(X)^T\right)_{k,l} &= \lim_{J\to\infty} \frac{\sigma_w^2}{J} \sum_{i=1}^{J} \phi_i(x_k)\phi_i(x_l) \\ &= \int_{-\infty}^{\infty} \exp\left(-\frac{(x_k-c)^2}{2l^2}\right) \exp\left(-\frac{(x_l-c)^2}{2l^2}\right) \mathrm{d}c \\ &= \sqrt{\pi} l \sigma_w^2 \exp\left(-\frac{(x_k-x_l)^2}{4l^2}\right)\end{aligned} \tag{2.11}$$

Once simplified, this inner product is easily computable, despite the use of a feature space of infinite dimension. This is a first example of what we will call the *kernel trick* (Michael I. Jordan, 2004), where a simple inner product is computed and used for prediction in Bayesian linear regression, instead of the explicit features $\phi(x)$ themselves. In general we will use the notation $\langle \phi(x); \phi(x') \rangle$ for inner products in the feature space.

We will now formally introduce the concept of covariance functions and their link to this inner product, and show how the kernel trick permits the extension of linear regression to Gaussian Process regression.

Duality between feature maps and covariance functions

Given a linear model $Y = \phi(X)w + \varepsilon$, with Gaussian likelihood $P(Y|w) = \mathcal{N}(Y|Xw, \sigma_\varepsilon^2 I)$ and the prior distribution $P(w) \sim \mathcal{N}(0, \Sigma_w)$, marginalising over the weights gives rise to the following marginal likelihood:

$$\begin{aligned} P(Y) &= \int_w P(Y|w)P(w)\mathrm{d}w \\ &= \mathcal{N}\left(0, \phi(X)\Sigma_w\phi(X)^T + \sigma_\varepsilon^2 I\right) \end{aligned} \tag{2.12}$$

It appears from this marginalisation that the covariance of Y is a function of the input, parametrised by the inner product $\phi(X)\Sigma_w\phi(X)^T$ introduced in Equation 2.7. We call this inner product $\langle\phi(X);\phi(X')\rangle = \phi(X)\Sigma_w\phi(X')^T$ a covariance function or kernel.

Noting χ the definition domain of the input space, any semi-positive definite function $k : \chi^2 \to \mathbb{R}$ gives rise to a valid covariance. One can show that for any such k, there exists a Hilbert space $\mathscr{H}$ and a (generally non-unique) feature map $\phi : \chi \to \mathscr{H}$ such that $\forall(x,x') \in \chi^2, k(x,x') = \langle\phi(x);\phi(x')\rangle_{\mathscr{H}}$ (see the Reproducing Kernel Hilbert Space property (Carmeli et al., 2005; Gretton, 2017)). The Mercer Theorem (Mercer, 1909; Minh et al., 2006) enables us to reformulate explicitly any covariance function as an inner product, thereby exhibiting a feature map ϕ.

In practise, this means that the choice of a covariance function k defined in the input space circumvents the need to define a feature map ϕ and enables us to work implicitly in very rich and complex feature spaces. Using this duality is often referred to as the *kernel trick* (Michael I. Jordan, 2004) and it is at the core of the Gaussian Process framework (Section 2.1.6).

2.1.6 Gaussian Process formalism

Definition

A Gaussian Process is a stochastic process (an infinite collection of random variables), of which any finite subset has a joint Normal distribution. Each random variable i in the collection is typically associated to an input x_i, and the Gaussian Process is entirely characterised by the mean function $m(x)$ and the covariance function or kernel $k(x,x')$ a semi-positive definite function defined for any pair of inputs x and x'. We will write the Gaussian Process f as:

$$f(x) \sim GP(m(x), k(x,x')) \tag{2.13}$$

For N data points with an input matrix X of dimension $N \times D$, we write $m(X)$ the vector of the values taken by the mean function, and $K(X,X)$ the covariance matrix of dimension $N \times N$ made of the values of $k(x,x')$ for all pairs of points. We note $\mathrm{f} = \{f(x_i)\}_{i\in[\![1;N]\!]}$ the

subset of the Gaussian Process f for these inputs. The random vector f follows the following Normal distribution:

$$\mathrm{f} \sim \mathcal{N}(m(X), K(X,X)) \tag{2.14}$$

In this thesis, we will always consider noisy observations: $Y = \mathrm{f} + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$. We will marginalise over the noise-free observations f and directly consider the marginal distribution:

$$Y \sim \mathcal{N}\left(m(X), K(X,X) + \sigma_\varepsilon^2 I\right) \tag{2.15}$$

In the rest of this thesis, we assume $m(x) = 0$.

Weight-space view: generalisation of Bayesian linear regression

In Section 2.1.5, we have derived the duality between semi-positive definite covariance functions and feature maps. Based on this duality, Gaussian Processes can essentially be regarded as a generalisation of Bayesian linear regression, where the feature space is implicit and the covariance function becomes the main object of interest. Different choices of the covariance function model our assumptions about how the similarity between inputs relates to how similar the output variables are. The predictive distribution can be expressed as a function of this covariance function like in Bayesian Linear regression:

$$\begin{aligned} P(y^*|x^*, X, Y) = \mathcal{N}(y^*|&K(x^*,X)\left(K(X,X) + \sigma_\varepsilon^2 I\right)^{-1} Y, \\ &K(x^*,x^*) - K(x^*,X)\left(K(X,X) + \sigma_\varepsilon^2 I\right)^{-1} K(X,x^*)) \end{aligned} \tag{2.16}$$

Function-space view

Alternatively, the so-called function space view interprets Gaussian Processes as a distribution over functions.

In the input space, we model the target variable y as linked to the input x by a function f and the observational noise: $y = f(x) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. The aim of a supervised machine learning task is to infer from the data the form of this unknown function f. Defining a Gaussian Process can be regarded as defining a prior distribution over regression functions f: $f(x) \sim GP(m(x), K(x,x'))$, so that the marginal distribution of Y over all possible functions f given this GP prior is $Y \sim \mathcal{N}(m(X), K(X,X) + \sigma_\varepsilon^2 I)$.

Given a training set, the predictive posterior $P(y^*|x^*,X,Y)$, can be interpreted as the posterior distribution of f given the observed data: $f^*(x^*|X,Y)$. In the case of noise-free observations ($\sigma_\varepsilon^2 = 0$), this posterior distribution would restrict the space of possible functions f to functions going through the observed data points, while modelling observational noise relaxes this constraint (Fig. 2.1).

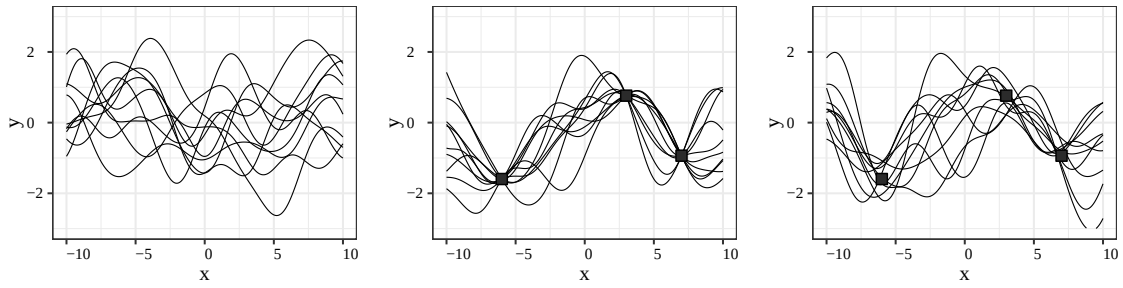


Fig. 2.1 Function space view of Gaussian Processes. Left: functions $f : x \mapsto y$ drawn from a Gaussian Process prior using a squared-exponential kernel with length scale 2 and scaling hyperparameter 1 (see Section 2.1.7). Middle: noise-free posterior over functions after the observation of three data points. Right: posterior over functions when modelling observational noise with a noise scaling hyperparameter of 0.1.

Example: modelling spatial or temporal context

Given a time series dataset or a spatially resolved dataset such as a biological image, it is often difficult to know in advance if and how space or time specifically affects the data, or to encode these assumptions in conventional linear covariates.

The GP framework allows for making flexible assumptions about how the covariance of the output variables may relate to their relative measurement time or spatial location, by choosing a smooth and flexible covariance function. A common choice of covariance function for this purpose is the squared exponential kernel (see Section 2.1.7): $\mathrm{cov}(y_i, y_j) = \exp\frac{-d_{i,j}^2}{2l^2}$, where $d_{i,j}$ represents the distance between data points in time or space (Rizvi et al., 2017; Hensman et al., 2015, 2013b; Groot et al., 2011; Roberts et al., 2012; Niu et al., 2016).

2.1.7 Kernel design

In Gaussian Processes, the covariance function (or kernel) is the object of primary interest and its choice is the cornerstone of the design of a Gaussian Process model. As we have seen,

it corresponds to the choice of a feature map in the weight space view, while in the function space view, it defines our prior over the function f we are trying to infer.

The choice of the covariance function encodes our intuition about how correlated a pair of output variable is as a function of their similarity in the input space. For example, in a time series application, one might assume that the similarity between two output variables y_i and y_j will only depend on how far apart in time they were measured, irrespective of the specific measurement time. This means that the covariance between y_i and y_j should only depend on the difference between the two time points t_i and t_j: $\text{cov}(y_i, y_j) = k(t_j - t_i)$. This assumption will suggest the use of a type of covariance functions called stationary, which fulfils this translation invariance requirement. In contrast, other covariance functions will be more suited to periodic signals.

Extensive lists of usual covariance functions, as well as general recipes for designing and combining covariance functions can be found in Rasmussen and Williams (2006), Chap. 4, and Duvenaud (2014). In this thesis, we will only present the two covariance functions of interest for the model of Chapter 3.

Linear covariance functions

Linear covariance functions derive directly from the marginalisation of the weights in Bayesian linear regression and are of the form: $k(x,x') = \sigma^2 x^T x'$, where σ^2 is a scaling hyperparameter, corresponding to the following independent prior on the Bayesian regression weights: $w \sim \mathcal{N}(0, \sigma^2 I)$.

The corresponding feature space is the non-transformed input space (or a feature space of finite dimension with explicit basis functions), and in the function-space view, this corresponds to a prior over all linear functions of the input x (Fig. 2.2).

One example for the use of linear covariance functions in bioinformatics, is to model population structure based on genotype in linear mixed models for Genome Wide Association Studies (Casale, 2016; Widmer et al., 2014; Li and Zhu, 2013).

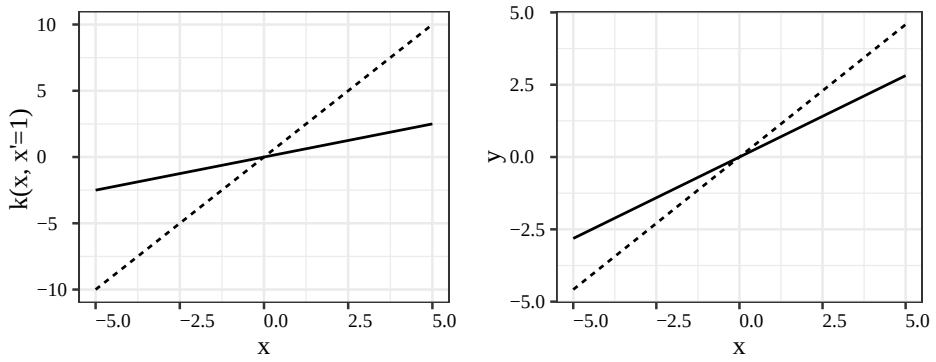


Fig. 2.2 Linear covariance function: $k(x,x') = \sigma^2 x^T x'$. Left panel: values of the covariance function $k(x,x'=1)$. Right panel: Samples from the Gaussian Process (i.e. samples of the mapping function f between x and y in the function-space view). Solid line: $\sigma = 0.5$. Dashed line: $\sigma = 2$

Squared Exponential covariance functions

Squared exponential covariance functions are of the form $k(x,x') = \sigma^2 \exp(-d^2/2l^2)$, where d is the Euclidean distance between x and x', σ is a scaling hyperparameter, and l a length scale parameter which controls the smoothness of the Gaussian Process (fig. 2.3).

There are multiple ways to exhibit non-unique feature spaces and maps corresponding to a squared-exponential covariance function. Section 2.1.5 gives a possible set of basis functions and the Bochner theorem offers a principled way to exhibit another possible feature space of infinite dimension using Fourier transforms (Bochner, 1959; Samo and Roberts, 2015; Oliva et al., 2016). Squared exponential covariance functions are infinitely differentiable and correspond to very smooth functions in the function space view. This kernel can be used for the approximation of a wide variety of non-linear functions as shown in Micchelli et al. (2006) (Fig. 2.3). Another important property of squared exponential covariance functions is that they are stationary: $k(x,x') = k(x-x')$.

In bioinformatics, squared-exponential covariance functions have been used to analyse gene expression patterns in time (Kalaitzis and Lawrence, 2011; Lawrence et al., 2007; McDowell et al., 2018), sometimes in combination with other covariance functions to detect periodicity (Durrande et al., 2016), or in GP mixtures to model perturbations or branching processes (Yang et al., 2016; Lönnberg et al., 2017; Boukouvalas et al., 2018). More recently,

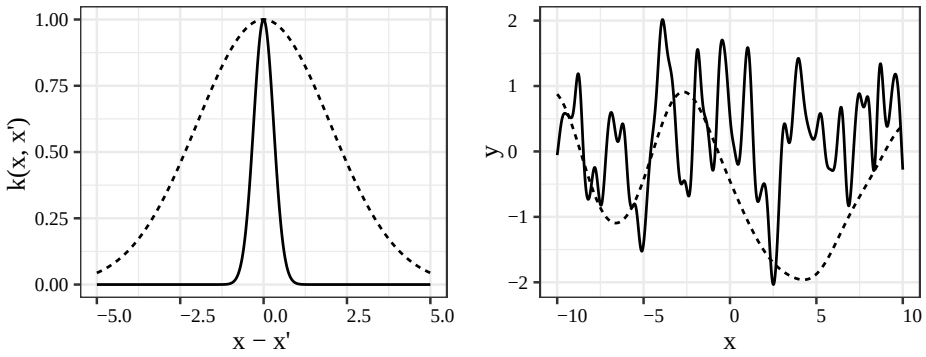


Fig. 2.3 Squared Exponential covariance function: $k(x,x') = \sigma^2 \exp(-(x-x')^2/l^2)$. Left panel: Covariance value as a function of the distance between x and x'. Right panel: samples from the Gaussian Process (i.e. samples of the mapping function f between x and y in the function-space view). Solid line: $l = 0.3$. Dashed line: $l = 2$.

Svensson et al. (2018a) used a squared-exponential kernel to model gene expression patterns in space.

Combining covariance functions

Two important properties of covariance functions is that sums and products of covariance functions are covariance functions. We can therefore build more complex covariance functions by combining existing ones.

Interpreted in feature space, the sum of two covariance functions corresponds to the concatenation of all the features associated to each covariance function, modelling our assumption of the additive effect of all features. This can be shown by writing the covariance functions in terms of inner products in the feature space. In function-space view, the sum of two covariance functions k_A and k_B corresponds to summing the functions with respective priors $GP(0,k_A)$ and $GP(0,k_B)$, as illustrated in Figure 2.4 with the sum of a linear and a squared exponential covariance function. This is because the sum of two independent variables A and B following multivariate Normal distributions $\mathcal{N}(0,k_A)$ and $\mathcal{N}(0,k_B)$ follows a multivariate Normal distribution with covariance $k_A + k_B$.

If we assume that the output variable y is affected by two different inputs x_A and x_B with independent additive effects, we can use two covariance functions k_A and k_B, and sum their contributions: $\text{cov}(y_i, y_j) = k_A((x_A)_i, (x_A)_j) + k_B((x_B)_i, (x_B)_j)$. This is useful when using

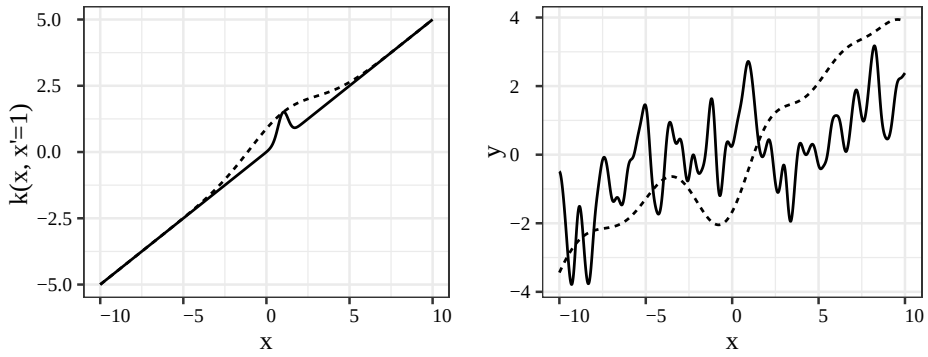


Fig. 2.4 Sum of a linear covariance function with scaling hyperparameter $\sigma = 0.5$ and a squared exponential with scaling hyperparameter $\sigma = 1$ and length scale l. Left panel: value of the resulting covariance function $k(x, x' = 1)$. Right Panel: Samples from the Gaussian Process. Solid line: $l = 0.3$ Dashed line: $l = 2$

Gaussian Processes to disentangle the contributions from multiple sources of variation (Hoffman and Schadt, 2016), as we will do in Chapter 3 (see also Section 2.1.9).

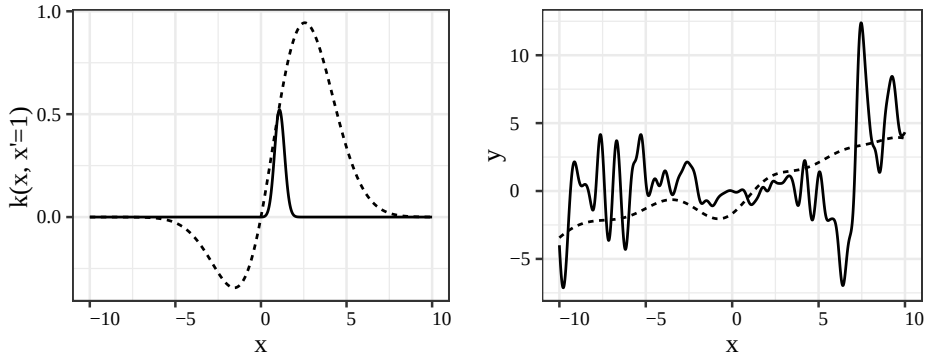


Fig. 2.5 Product of a linear covariance function with scaling hyperparameter $\sigma = 0.5$ and a squared exponential with scaling hyperparameter $\sigma = 1$ and length scale l. Left panel: value of the resulting covariance function $k(x, x' = 1)$. Right Panel: Samples from the Gaussian Process. Solid line: $l = 0.3$. Dashed line: $l = 2$

The multiplication of two covariance functions corresponds to the concatenation of the products of all pairs of features from the two feature spaces, which models interactions between them. In the function-space view, the product of two covariance functions, illustrated in Figure 2.5 with the example of the product of a linear and a squared exponential covariance function, does not have an analogous interpretation as a product of function. It is not

surprising, as the product of two Normally distributed variables does not in general follow a Normal distribution.

2.1.8 Hyperparameters optimisation

Flexible covariance functions like the squared exponential kernel constitute a great modelling tool by enabling us to work in a rich feature space of infinite dimension with only simple computations in the input space. However, these functions also depend on hyperparameters such as the scaling hyperparameter and length scales seen before. The choice of these hyperparameters, as well as the noise hyperparameter σ_ε, have pronounced effects on model predictions (Fig. 2.6), hence principled strategies are needed for their choice or optimisation.

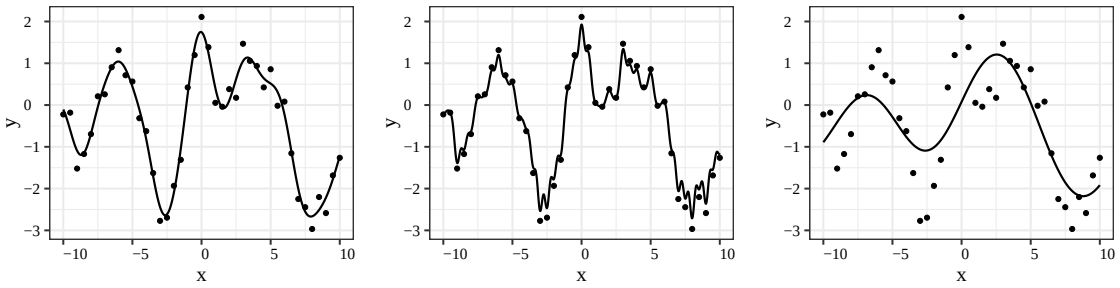


Fig. 2.6 Gaussian Process predictions for different hyperparameters. The data points are generated from a joint Normal distribution with covariance $\mathrm{cov}(y_i, y_j) = \exp(-(x_i - x_j)^2/2l^2) + 0.1\delta_{i,j}$. The lines show the mean of the GP predictive distribution with covariance functions $k(x_i, x_j) = \exp(-(x_i - x_j)^2/2l^2) + 0.1\delta_{i,j}$. Left panel: $l = 1$ (ground truth), middle panel: $l = 0.2$ (overfitting), right panel: $l = 4$ (under fitting)

Type-II maximum likelihood using gradient ascent

In the following, we denote θ the vector of all hyperparameters for a given covariance function k, as well as the noise hyperparameter σ_ε. In a full Bayesian treatment, we would use another level of prior distribution $P(\theta)$ on the elements of θ and use the Bayes rule to compute the posterior distribution:

$$P(\theta|Y,X) = \frac{P(Y|X,\theta)P(\theta)}{\int_\theta P(Y|X,\theta)P(\theta)\mathrm{d}\theta} \tag{2.17}$$

However, a simpler alternative which is widely used in Gaussian Process regression (Pedregosa et al., 2011; GPy, 2012) is to simply maximise the marginal likelihood $P(Y|X,\theta)$ with respect to θ. Note here that $P(Y|X,\theta)$ is a marginal likelihood because, in the weight

space view, the weights of the linear mapping in the feature space have been marginalised over, while in the function space view, the noise free variables f have been marginalised over (see Section 2.1.6). We call this optimisation procedure type II maximum likelihood (Rasmussen and Williams, 2006, Chap. 5). In Bayesian terms, this also corresponds to a Maximum A Posteriori (MAP) estimate using a uniform prior on θ. MAP estimates with different choices of priors can also be derived and will only add a regularisation term to the log likelihood (e.g. a quadratic regularisation for a Normal prior).

An appealing property of Gaussian Processes is that the log marginal likelihood and its gradient can be computed in closed form (Rasmussen and Williams, 2006, Chap. 5):

$$\begin{aligned} \ln P(Y|X,\theta) &= -\frac{1}{2}Y^T (K)^{-1} Y - \frac{1}{2}\ln|K| - \frac{n}{2}\ln 2\pi \\ \frac{\partial}{\partial\theta_i}\ln P(Y|X,\theta) &= \frac{1}{2}Y^T K^{-1}\frac{\partial K}{\partial\theta_i}K^{-1}Y - \frac{1}{2}\mathrm{tr}\left(K^{-1}\frac{\partial K}{\partial\theta_i}\right) \end{aligned} \tag{2.18}$$

using the short notation $K = K(X,X) + \sigma_\varepsilon^2 I$.

The hyperparameters θ are then typically optimised using standard gradient ascent techniques such as lbfgs (Bonnans et al., 2006).

Computational complexity

Evaluating the log marginal likelihood and its gradient (Eq. 2.18) involves computing the inverse of the covariance matrix $K(X,X) + \sigma_\varepsilon^2 I$ of dimensions $N \times N$. This is the bottleneck of GP regression and scales typically in N^3.

For large datasets (large N), approximate methods have been proposed to make this inference faster. For example, Sparse Gaussian Processes (Hensman et al., 2013a; Snelson and Ghahramani, 2006; Quiñonero-Candela and Rasmussen, 2005) approximate the full dataset with inducing variables, which consists in a small number $M < N$ of well chosen pseudo input points which best represent the data. The position of this pseudo input is typically optimised in a probabilistic manner and the complexity is typically reduced to M^3. Alternatively, random feature methods (Rahimi and Recht, 2008; Oliva et al., 2016) use the duality between covariance function and feature space, to approximate a Gaussian Process by a linear regression with a relatively small number of features, which are chosen in a

probabilistic manner so as to provide the best kernel approximation. These methods typically scale linearly in the number of data points N.

2.1.9 Variance decomposition

In many biological applications, supervised machine learning is not used for the main purpose of predictions, but rather in order to find out what input variables are predictive of the output variable, so as to unveil the drivers of variation of the output variable Y (Hoffman and Schadt, 2016). Gaussian Process regression is well suited for that purpose, and provides a principled way to assess the variance explained by multiple groups of input features.

Let us consider an additive model: $Y \sim \mathcal{N}(0, \sum_i \sigma_i^2 K_i(X_i, X_i) + \sigma_\varepsilon^2 I)$, where multiple groups of input variables X_i are linked to the output Y via the respective covariance functions $\sigma_i^2 K_i$ We explicitly represented the scaling hyperparameters $\sigma_{i,\forall i}^2$ and σ_ε^2 but each kernel might have additional hyperparameters such as a length scale for a squared exponential covariance function. These hyperparameters are optimised by maximising the marginal likelihood seen before for a given training set. Note that the use of a scaling hyperparameter ensures that if a group of input X_i does not explain any variance of Y, the corresponding covariance function may take a constant value of zero.

The variance explained by each covariance term K_i can then be estimated using Gower factors (Searle, 1982; Kostem and Eskin, 2013), which are defined as follow:

$$G(K) = \frac{\mathrm{tr}\,(PKP)}{N-1}, \text{ with } P = I_N - J_N \tag{2.19}$$

where I_N is the identity matrix of dimensions $N \times N$ and J_N is a matrix of ones of dimensions $N \times N$. The Gower factor of a covariance term computes the expected variance of a random variable which is Normally distributed with the considered covariance. In other words, the Gower factor of each covariance term computes the amount of variance of Y across samples explained by the corresponding group of features X_i: for $Y \sim \mathcal{N}(0, K), G(K) = \mathbb{E}[\mathrm{var}(Y)]$.

2.2 Factor Analysis

The second part of this chapter introduces linear latent variable models for dimensionality reduction, and more specifically Bayesian Factor Analysis. We show how extending this modelling approach using hierarchical priors is a natural way to encode prior knowledge

about data context and we introduce Variational Bayesian inference, an approximate inference method which is commonly used when the posterior distribution over parameters is intractable.

In this section and in Chapter 4, probabilistic models are illustrated with Bayesian networks which express all conditional dependencies between the parameters of the models and the observed variables.

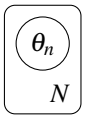
2.2.1 Graphical notations for probabilistic models

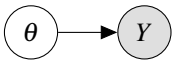
Notations were adapted from Dietz (2010). The different types of model variables are represented with different types of nodes in the network; nodes repetitions are expressed with plate notations and the conditional dependency between nodes is expressed with directed arrows:

Observed variables 

Probabilistic parameters 

Deterministic hyperparameters 

Repetition of node θ_n for $n \in [\![1;N]\!]$ 

Conditional dependency between nodes: $P(Y,\theta) = P(Y|\theta)P(\theta)$ 

For simplicity, deterministic hyperparameters are only represented on graphical models when, although not modelled as random variables, they are optimised using Maximum Likelihood, typically in an Expectation-Maximisation scheme.

A more complete factor graph notation (Loeliger, 2004; Kschischang et al., 2001) would represent all components of the model including the prior distributions used and any function that couples the random variables. In this thesis, we prefer to give those informations in separate equations to avoid cluttering the graphical notations.

2.2.2 Dimensionality reduction and latent variable models

Many biological datasets, such as transcriptomics and proteomics are high dimensional (large number of features), but their variability is very structured: for example a given biological

process typically involves multiple interacting genes which vary in a coordinated manner (Ma and Gao, 2012; Margolin et al., 2006; Markowetz and Spang, 2007; Komili and Silver, 2008; Van Dam et al., 2018; Markowetz, 2010). This redundancy in the data motivates and permits the use of dimensionality reduction techniques (Van der Maaten et al., 2008) for the unsupervised analysis of high dimensional biological data.

The aim of dimensionality reduction is to explain a dataset of dimensionality D (number of features) with a small number K of latent (hidden) variables. We assume that there is a mapping function Φ between the low dimensional space and the high-dimensional space, which typically depends on some parameters Θ (Eq. 2.20). The inference process consists in learning the values of the latent variables and the values of the mapping function parameters Θ which best explain the data.

$$Y = \Phi_{\Theta}(Z) \tag{2.20}$$

Y is the dataset matrix with dimensionality $N \times D$, N is the number of samples and D the number of features. Z is the latent variable matrix of dimension $N \times K$.

Dimensionality reduction techniques discover the hidden structure of a high dimensional dataset. In a biological context, latent variables can reflect biological processes or technical factors (Stegle et al., 2012; Leek and Storey, 2007). The value of a latent variable $z_{n,k}$ quantifies the strength of the effect of latent variable k on sample n, while the mapping function models how it affects the different data features.

2.2.3 Principal Component Analysis (PCA)

The most widely used dimensionality reduction technique is Principal Component Analysis (Hotelling, 1933; Ringnér, 2008; Pearson, 1901). PCA assumes that there exists a linear mapping between the latent variables Z (the *principal components*) and the data Y (Eq. 2.21): $Y = ZW$, where W is a weight matrix of dimensions $K \times D$. Each parameter $w_{k,d}$ of this linear mapping corresponds to the *loading* of the principal component k on the feature d.

$$y_{n,d} = \sum_{k=1}^{K} z_{n,k} w_{k,d} \tag{2.21}$$

In its minimum error formulation, Principal Component Analysis aims at finding the values of Z and W which best explain the data in terms of mean squared error between the observations

and the model predictions (Eq. 2.22). It can be shown that a solution for $\hat{Z}$ is the eigenvectors of the data covariance matrix (samples times samples) for the K biggest eigenvalues and the corresponding weight matrix W is the pseudo-inverse of the projection matrix on these eigenvectors (Richardson, 2009; Bishop, 2006, Chap. 12). This provides a deterministic inference method used in PCA, which adds an orthogonality constraint on the latent variables.

$$\hat{Z}, \hat{W} = \underset{Z,W}{\arg\max} \frac{1}{N} \sum_{n=1}^{N} \left(y_{n,d} - \sum_{k=1}^{K} z_{n,k} \times w_{k,d} \right)^2 \tag{2.22}$$

A linear mapping between the latent space and the data space has the advantage of yielding results which are in principle interpretable. Inspection of $W_{k,:}$ will reveal families of genes which are jointly affected by principal component k. Providing that the data is already normalised for technical confounders, these families of genes will hopefully correspond to gene sets linked to known biological processes, enabling a biological interpretation of the corresponding latent variable. This inspection step can be further automatised with standard tools such as Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005).

2.2.4 Probabilistic frameworks for linear dimensionality reduction

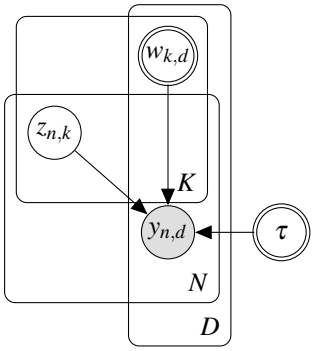
As such, PCA does not offer a principled way of modelling prior information about the data and the data context, and has additional limitations such as the inability to cope with missing values in the data. To address this, multiple probabilistic models for dimensionality reduction have been proposed based on the linear mapping of PCA (Cunningham and Ghahramani, 2015), which differ by the specific choice of prior distributions on the model random parameters. The posterior distribution of the model parameters can be computed using the Bayes rule (Eq. 2.23), either in closed form or using approximate inference methods.

$$P(\Theta|Y) = \frac{P(Y|\theta)P(\Theta)}{\int_{\Theta} P(Y|\theta)P(\Theta)\mathrm{d}\Theta} \tag{2.23}$$

In this section we will present a probabilistic formulation of PCA and its extension to Factor Analysis, which the modelling approach of Chapter 4 builds on. In Section 2.2.5, we will then explain how this probabilistic framework allows accounting for prior knowledge about data context using hierarchical priors.

Probabilistic PCA (pPCA)

First probabilistic formulations of PCA were proposed by Tipping and Bishop (1999) and Roweis (1998). Like in PCA, a linear mapping between the latent variables and the data is assumed, and a Gaussian observational noise ε is explicitly modelled: $y_{n,d} = \sum_{k=1}^{K} z_{n,k} w_{k,d} + \varepsilon_{n,d}$. Assuming $\varepsilon_{n,d} \sim \mathcal{N}(0, \tau)$, where τ is a precision hyperparameter, the distribution of the data given the model parameters is $P(y_{k,d}|Z, W, \tau) \sim \mathcal{N}(\sum_{k=1}^{K} z_{n,k} w_{k,d}, \tau)$[2]. Latent variables are modelled as random with the following prior distribution: $z_{n,k} \sim \mathcal{N}(0, 1)$ for all n and k, while the precision τ and the weights W are deterministic (Fig. 2.7).



$$y_{n,d} \sim \mathcal{N}\left(\sum_{k=1}^{K} z_{n,k} w_{k,d}, \tau\right)$$

$$P(z_{n,k}) \sim \mathcal{N}(0, 1)$$

Fig. 2.7 Left: Graphical model of probabilistic Principal Component Analysis (pPCA). The only parameters to be modelled as random variables are the latent variables $z_{n,k}$, but deterministic hyperparameters $w_{k,d}$ and τ are also inferred from the data. Right: specification of the model likelihood and the parameters prior distributions.

For a given value of the weight and precision hyperparameters, the Normal priors on the latent variables are conjugate for the Gaussian likelihood (Murphy, 2007). The posterior distribution of the latent variables can therefore be computed in closed form using the Bayes rule (see Section 2.1.2).

The weight and precision hyperparameters are optimised using an Expectation-Maximisation (EM) algorithm (Rubin and Thayer, 1982). In the Expectation step, the posterior distribution $P(Z|Y, W, \tau)$ is computed given the current values of the hyperparameters W and τ. Then, the expected value of the joint log likelihood $\ln(P(Y, Z))$ is computed given the posterior distribution of Z: $\mathbb{E}_{P(Z|Y,W,\tau)} \ln(P(Y, Z))$. In the Maximisation step this quantity is maximised with respect to the hyperparameters W and τ: $\hat{W}, \hat{\tau} = \arg\max_{W,\tau} \mathbb{E}_{P(Z|Y,W,\tau)} \ln(P(Y, Z))$. Tipping

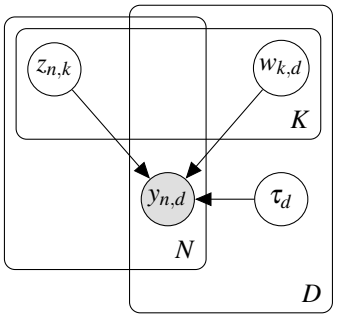
[2]Note that PCA (Section 2.2.3) provides maximum likelihood estimates for W and Z. This can be verified by taking the log of the Gaussian likelihood which results in the quadratic loss function of Equation 2.22.

and Bishop (1999) show that the solution of the maximisation step can also be computed in closed form.

Whereas PCA and its implementation relying on the eigenvalue decomposition of the data covariance matrix requires full data matrices, a first advantage of the probabilistic formulation and its EM resolution is its applicability to datasets with missing observations. In addition, we will show in the next paragraph and in Section 2.2.5 that the probabilistic framework forms the basis of a Bayesian treatment of linear dimensionality reduction models and permits a broad range of principled extensions.

Bayesian Factor Analysis

pPCA assumes an isotropic noise model (hyperparameter τ shared for all d). Factor Analysis (FA) (McDowell et al., 2018; Bartholomew, 1985) relaxes this assumption by using a feature specific precision τ_d. Unlike pPCA, this more flexible assumption allows for the analysis of data with features on different scales. In the full Bayesian treatment of FA, we also model the weights $w_{k,d}$ as random variables with isotropic Normal prior distributions as well as the precision parameters τ_d with gamma prior distributions (Fig. 2.8). To avoid any ambiguity, we will call weights the $w_{k,d}$ parameters and factors the $z_{n,k}$ parameters.



$$y_{n,d} \sim \mathcal{N}\left(\sum_{k=1}^{K} z_{n,k} w_{k,d}, \tau_d\right)$$
$$P(z_{n,k}) \sim \mathcal{N}(0,1)$$
$$P(w_{k,d}) \sim \mathcal{N}(0,1)$$
$$P(\tau_d) \sim \Gamma(1,1)$$

Fig. 2.8 Left: graphical model for Bayesian Factor Analysis. The precision τ_d is defined for each feature independently and all model parameters are treated as random variables. Right: specification of the model likelihood and the prior distributions of the parameters.

The choice of conjugate priors makes full Bayesian inference easier, although analytically intractable in this specific case. Section 2.2.6 will present one way to approximate the posterior distribution of the model parameters for the Gaussian noise model presented here. Inference methods for an extension of Factor Analysis with non-Gaussian likelihoods are

presented in Appendix C.

In the next section, we will show that this full Bayesian model allows for principled extensions to model prior knowledge about data context in a principled manner, using hierarchical priors.

2.2.5 Hierarchical priors to model prior knowledge about the data

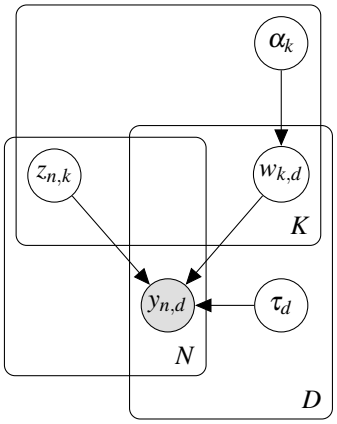
Using explicit probabilistic models and Bayesian inference for dimensionality reduction is a powerful approach as it provides a principled way to account for prior knowledge about the data, including the context in which it was collected, or sensible assumptions about its hidden structure, such as sparsity assumptions. This is done using hierarchical priors on the model parameters. Prior distributions over the model parameters are themselves parametrised (eg by their mean and their variance for Normal priors). Building hierarchical priors in a model consists in modelling those parameters as random variables too, which may be done in successive layers.

Automatic Relevance Determination

In Factor Analysis, a common assumption about the hidden structure of the data is that only a small number of factors are relevant to explain the observed variance. This exact dimensionality is however unknown, while the number of factors used in the model of Figure 2.8 needs to be determined a priori. A solution to infer the dimensionality of the latent space from the data is to initially assume a high number of factors, while making sure that the posterior distribution prunes irrelevant ones.

This behaviour can be achieved using Automatic Relevance Determination (ARD) (David J. C. MacKay, 1994; Neal, 1995; Wipf and Nagarajan, 2008), which consists in optimising the precision of the prior distributions of the model weights, adding a level of hierarchy in the model parameters. In a Bayesian framework, this is done by modelling this precision as a random variable with a Gamma prior (Fig. 2.9).

If a factor is irrelevant, the posterior distribution $P(\alpha_k|Y)$ will take a high mean, enabling, in turn, the posterior $P(W_{k,:}|Y)$ to be very peaked at zero (Fig. 2.10). The resulting Factor Analysis model will as a result learn automatically an appropriate number of factors for a given dataset. This is another improvement allowed by probabilistic models compared to less flexible frameworks like PCA.



$$y_{n,d} \sim \mathcal{N}\left(\sum_{k=1}^{K} z_{n,k} w_{k,d}, \tau_d\right)$$
$$P(z_{n,k}) \sim \mathcal{N}(0,1)$$
$$w_{k,d} \sim \mathcal{N}(0, \alpha_k)$$
$$\alpha_k \sim \Gamma(a_0, b_0)$$

Fig. 2.9 Left: graphical model for Bayesian Factor Analysis with ARD priors. A factor-specific gamma prior is used for the precision of the weights. Right: specification of the model likelihood and the parameter prior distributions. Typical hyperparameters for the prior distribution of α_k are $a_0 = b_0 = 10^{-14}$ (Virtanen et al., 2012).

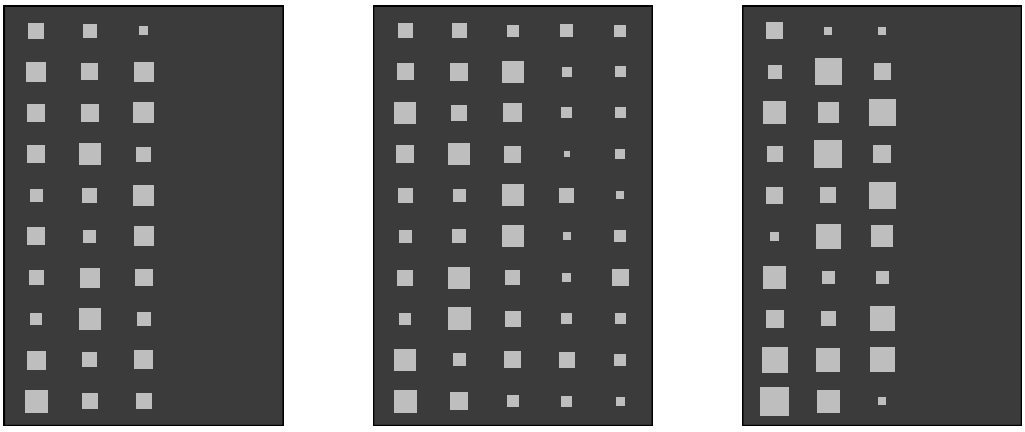
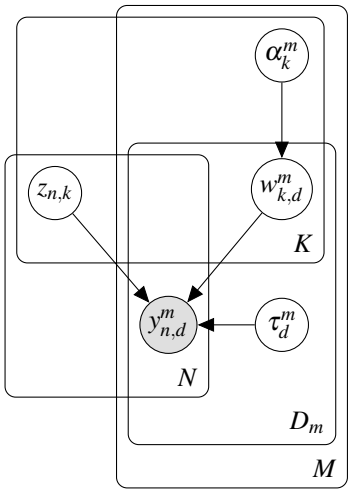


Fig. 2.10 Hinton plot of the weight matrix for the true weights (left), weights inferred by a FA model with no ARD (middle), weights inferred by a FA model with ARD priors (right). Rows represent the features $d \in [\![1; D]\!]$, columns represent the factors $k \in [\![1; K]\!]$, the absolute value of a weight $w_{k,d}$ is represented by the size of the corresponding square. For this figure, synthetic data was generated for 100 samples and 10 features with three generative factors. FA models with and without ARD were fitted with 5 factors using the bofam package (see Chapter 4).

Modelling the data context: Group Factor Analysis

Hierarchical priors might also be used to model contextual knowledge about the data. For example, we may know that certain groups of features have similar properties. In bioinformatics, an obvious case is when multiple omics types (eg transcriptomics and methylation data) are measured for the same samples (*multi-omics data*, Dihazi et al. (2018); Huang et al. (2017)).

In such a case, some factors may only be relevant for a specific group of features, and modelling these groups specifically may help to uncover this hidden structure. This is addressed by a class of models called Group Factor Analysis (GFA) (Virtanen et al., 2012). For M groups of features, a specific ARD prior $\alpha_{k,m}, \forall m \in [\![1;M]\!]$ is used for every factor (Fig. 2.11), enabling the deactivation of factors in specific groups only (Fig. 2.12).



$$y_{n,d}^m \sim \mathcal{N}\left(\sum_{k=1}^{K} z_{n,k} w_{k,d}^m, \tau_d\right)$$
$$P(z_{n,k}) \sim \mathcal{N}(0,1)$$
$$w_{k,d}^m \sim \mathcal{N}(0, \alpha_k)$$
$$\alpha_k^m \sim \Gamma(a_0, b_0)$$

Fig. 2.11 Left: graphical model for Group Factor Analysis (GFA). A factor and group specific gamma prior is used for the precision of the weights. Right: specification of the model prior distributions. Typical hyperparameters for the prior distribution of α_k^m are $a_0 = b_0 = 10^{-14}$.

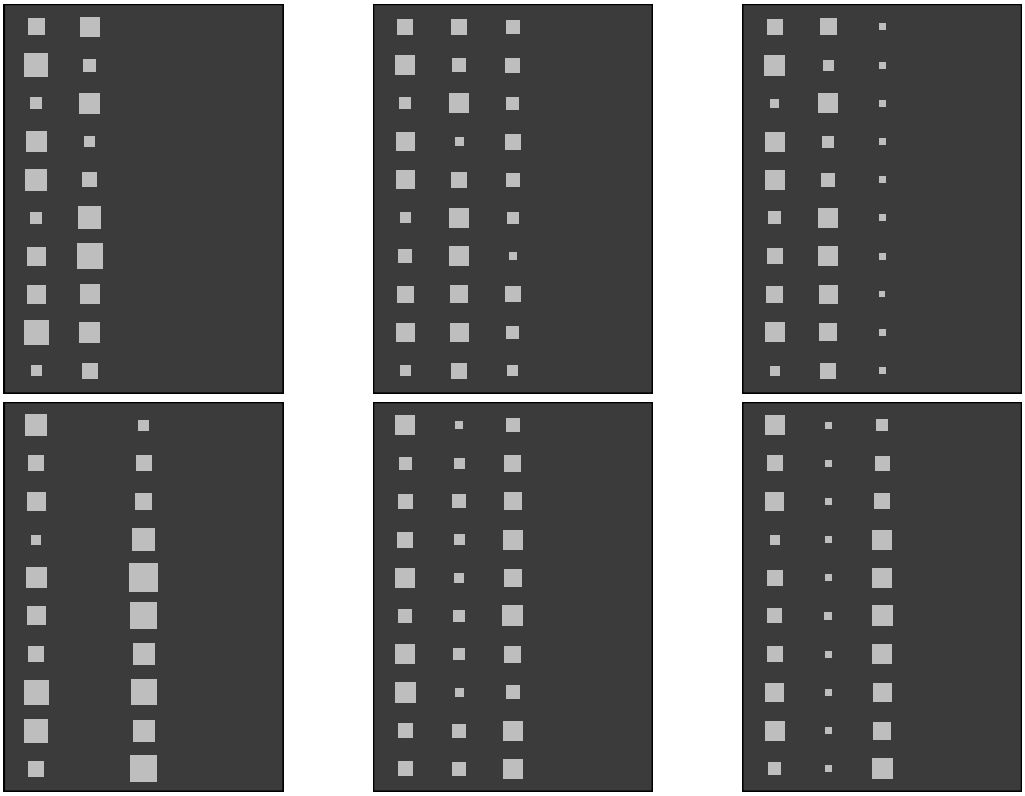


Fig. 2.12 Hinton plot of the weight matrix for the true weights (left), weights inferred by a FA model with an ARD per factor (middle), weights inferred by a Group Factor Analysis model (right). Rows represent the features $d \in [\![1; D]\!]$ (group 1 on top, group 2 at the bottom), columns represent the factors $k \in [\![1; K]\!]$, the absolute value of a weight $w_{k,d}$ is represented by the size of the corresponding square. For this figure, synthetic data was generated for 100 samples and two groups of 10 features with three generative factors relevant to specific groups. The FA model with ARD and the GFA model were fitted with 5 factors using the biofam software (see Chapter 4).

2.2.6 Approximate inference: variational methods

A major challenge posed by the use of hierarchical priors is the inference of the posterior distribution of the model parameters. The resulting complexity of the probabilistic model renders the exact posterior distribution given by the Bayes rule intractable. It becomes necessary to use approximate inference methods.

One approach is to use sampling based methods such as Gibbs sampling, which has the advantage of converging to the exact posterior distributions given infinite computational resources (Bishop, 2006, Chap. 11). Alternatively, deterministic methods such as variational

inference (Jordan et al., 1999; Beal and Others, 2003; Wainwright and Jordan, 2008) or expectation propagation (Minka, 2013) are less precise, but typically converge much faster and are also widely used in Bayesian modelling (Blei et al., 2016) for larger datasets. In this thesis, we made the choice of using variational inference, which we introduce here. Textbook references include Murphy (2012) (Chap. 21) and Bishop (2006) (Chap. 10).

Mean field approximation and Kullback-Leibler divergence

The posterior distribution of a hierarchical Bayesian model parameters may follow a complex joint distribution. In variational inference, we introduce a simpler distribution $q(\Theta)$, which aims at approximating the true posterior distributions $p(\Theta|Y)$ of the model parameters, and we minimise the Kullback-Leibler (KL) divergence between this q distribution and the true posterior distribution (Eq. 2.24).

$$\mathrm{KL}\left(q(\Theta)||P(\Theta|Y)\right) = -\int_{\Theta} q(\Theta) \ln \frac{P(\Theta|Y)}{q(\Theta)} \mathrm{d}\Theta \tag{2.24}$$

To make inference tractable, we need to restrict the form of this q distribution, although not necessarily in a parametric manner. The most common restriction used for q is called the mean field approximation (Parisi, 1988; Tanaka, 1999), in which q is factorised over all or some model parameters (Eq. 2.25). This family of distributions does not usually contain the true posterior distribution, in which model parameters would generally covary, but we will see that this assumption allows for the derivation of an analytical inference scheme with no further assumption on the q_i distributions.

$$q(\Theta) = \prod_i q_i(\theta_i) \tag{2.25}$$

The KL divergence of equation 2.24 is itself intractable as it requires to compute the intractable true posterior. However, this expression can be transformed as on equation 2.26.

$$\begin{aligned}
\mathrm{KL}\left(q(\Theta)||P(\Theta|Y)\right) &= -\int_{\Theta} q(\Theta) \ln \frac{P(\Theta|Y)}{q(\Theta)} \mathrm{d}\Theta \\
&= -\int_{\Theta} q(\Theta) \ln \frac{P(Y,\Theta)}{q(\Theta)P(Y)} \mathrm{d}\Theta \\
&= -\int_{\Theta} q(\Theta) \ln \frac{P(Y,\Theta)}{q(\Theta)} \mathrm{d}\Theta + P(Y)
\end{aligned} \tag{2.26}$$

As $P(Y)$ does not depend on the parameters Θ, we deduce that minimising the KL divergence of equation 2.26 is equivalent to maximising the term of equation 2.27.

$$\mathscr{L} = \int_{\Theta} q(\Theta) \ln \frac{P(Y,\Theta)}{q(\Theta)} \mathrm{d}\Theta \tag{2.27}$$

We first give an interpretation of $\mathscr{L}$ and then show how the mean field approximation makes the maximisation of this term analytically tractable.

Evidence Lower Bound (ELBO)

We have $P(Y) = \mathscr{L} + \mathrm{KL}\left(q(\Theta)||P(\Theta|Y)\right)$, and know that KL divergences are greater than zero. This means that $\mathscr{L}$ defines a lower bound to the evidence $P(Y)$. We will call this term the Evidence Lower BOund (ELBO).

The ELBO can be decomposed into two contributions shown on Equation 2.28. The first contribution, coming from the data likelihood under the approximate posterior distribution, is a measure of the goodness of fit. The second contribution is the opposite of the KL divergence between the prior and the posterior distribution of the parameters. It acts as a regularising term which prevents overfitting by accounting for prior assumptions on the parameter distributions.

$$\begin{aligned} \mathscr{L} &= \int_{\theta} q(\Theta) \ln \frac{P(Y|\Theta)P(\Theta)}{q(\Theta)} \mathrm{d}\Theta \\ &= \mathop{\mathbb{E}}_{q(\theta)} \ln P(Y|\Theta) - \mathrm{KL}\left(q(\Theta)||p(\Theta)\right) \end{aligned} \tag{2.28}$$

Variational Bayes (VB) alogrithm

The aim is now to maximise the ELBO $\mathscr{L}$ with respect to every q_i distribution coming from the mean-field approximation. This ELBO may be rewritten as on equation 2.29, where the constant cst does not depend on θ_i. Note that $\mathbb{E}_{\theta_{j\neq i}}$ designates an expectation with respect to the parameters $\theta_j, \forall j \neq i$ in their q distributions.

$$\begin{aligned} \mathscr{L} &= \int_{\Theta} q(\Theta) \ln \frac{P(Y,\Theta)}{q(\Theta)} \mathrm{d}\Theta \\ &= \int_{\theta_i} q(\theta_i) \left[\int_{\theta_{j\neq i}} \prod_{j\neq i} q(\theta_j) \ln P(Y,\Theta) \mathrm{d}\theta_j + cst \right] \mathrm{d}\theta_i - \int_{\theta_i} q(\theta_i) \ln q(\theta_i) \mathrm{d}\theta_i \\ &= -KL\left[q_i(\theta_i) \middle|\middle| \exp \mathop{\mathbb{E}}_{\theta_{j\neq i}} \left(\ln P(Y,\Theta)\right) + cst \right] \end{aligned} \tag{2.29}$$

It follows that maximising the ELBO with respect to q_i, with all other $q_{j,j\neq i}$ fixed, is equivalent to minimise the KL divergence shown on equation 2.29, which is achieved when:

$$\ln q_i(\theta_i) = \mathop{\mathbb{E}}_{\theta_{j\neq i}} (\ln P(Y,\Theta)) + cst \tag{2.30}$$

Equation 2.30 consists in a set of one equation per model variable θ_i, with dependencies in the first and second moment of the other distributions $q_{j\neq i}$ via the expectation $\mathbb{E}_{\theta_{j\neq i}}$. In practice, this suggests an algorithm where the q_i distributions are updated iteratively while keeping the $q_{i\neq j}$ distributions fixed, until convergence of the ELBO $\mathscr{L}$. When conjugate priors are used, the q_i distributions have the same functional forms as the priors $P(\theta_i)$, and their parameters can easily be identified from the functional form of $\mathbb{E}_{\theta_{j\neq i}} (\ln P(Y,\Theta))$.

Extension beyond the fully-factorised approximation

We have seen how to use variational methods to compute an approximation to the posterior distribution of the parameters of a probabilistic model, in the context of a fully factorised approximation to the posterior. However, we will see in Chapter 4 that there exists cases where it is necessary to model jointly some parameters in the q distribution: $q(\Theta) = q(\theta_k|\theta_l)q(\theta_l)\prod_{i,\notin\{k,l\}} q(\theta_i)$ with $q(\theta_k|\theta_l) \neq q(\theta_k)$. Appendix B describes the derivation of variational inference for this partially factorised case.

2.3 A note on the connection between Gaussian Processes and Factor Analysis

In the first section of this chapter, we have seen how Gaussian Processes enable us to infer complex functions f linking an input x to an output variable y, in the context of supervised Machine Learning tasks. The high capacity of Gaussian Processes rely on the implicit transformation of the input data X into a set of high dimensional features using the kernel trick. A Gaussian Process can also be seen as a prior over the function f. Data observation allows the computation of a posterior for f.

In the second section, we presented Factor Analysis and some extensions. The aim was to infer the low dimensional latent representation of a high-dimensional dataset in an unsupervised manner. The algorithm learns a mapping function between factors Z in a latent space and observed data y. This time, the latent variables Z are unobserved and their distribution is

also inferred probabilistically from the data.

In Factor Analysis, the mapping function between the latent variables and the observed data is linear, which has the advantage of being easily interpretable but has little capacity. GP-LVMs (Lawrence, 2005) bridge the gap between Gaussian Processes and Latent Variable models by putting a GP prior on the mapping function f between the latent and the observed space. The posterior f is inferred from the data jointly with the posterior distribution of the latent variables.

In bioinformatics, such models have been successful for pseudo-time ordering of cells, where the unobserved time point of a temporal process such as cell differentiation is treated as a latent variable, and the expression level of specific genes vary smoothly along this pseudo time space (Campbell and Yau, 2016; Ahmed et al., 2017; Macaulay et al., 2016b).

2.3.1 Model

For a sample i, a noise free observation f_i is modelled as being generated from the associated latent variables z_i via a mapping function f: $\mathrm{f}_i = f(z_i)$. f has a GP prior with a covariance function k which takes as input the latent variables such that $\mathrm{cov}(\mathrm{f}_i, \mathrm{f}_j) = k(z_i, z_j)$. Assuming an independent identically distributed Gaussian noise model, the noisy observations are $y_i = f(z_i) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

The log marginal likelihood of the data (integrating over the noise-free variables f, or the weights in the input space in the weight-space view) is as in Section 2.1.8:

$$\ln P(Y|Z,\theta) = -\frac{1}{2} Y^T \left(K(Z,Z) + \sigma_\varepsilon^2 I\right)^{-1} - \frac{1}{2} \ln \left|K(Z,Z) + \sigma_\varepsilon^2 I\right| - \frac{n}{2} \ln 2\pi \tag{2.31}$$

2.3.2 Inference

In GP-LVMs, as in other Latent Variable models, the main aim is to infer the posterior distribution (or point estimates) of the latent variables. In Equation 2.31, they appear in the inverse of the covariance matrix, parametrised by a possibly non-linear covariance function, which renders Bayesian inference intractable.

Maximum likelihood solution

A first inference method in GP-LVMs computes maximum likelihood point estimates of the latent variables using gradient ascent. Like in Section 2.1.8, the gradient of the log marginal likelihood with respect to any latent variable $z_{n,k}$ can be computed in closed form:

$$\frac{\partial}{\partial z_{n,k}} \ln P(Y|Z,\theta) = \frac{1}{2} Y^T K^{-1} \frac{\partial K}{\partial z_{n,k}} K^{-1} Y - \frac{1}{2} \mathrm{tr}\left(K^{-1} \frac{\partial K}{\partial z_{n,k}} \right) \tag{2.32}$$

using the short notation $K = \left(K(Z,Z) + \sigma_\varepsilon^2 I\right)$.

Variational inference solution

Variational methods can also be applied to compute a factorised approximation to the posterior distribution of the latent variables.

As seen in Section 2.2.6, the idea is to approximate the true posterior distribution over the latent variables $p(Z|Y)$ with a simpler distribution $q(Z)$, which in the mean-field approximation is of the fully factorised form $q(Z) = \prod_{n,k} q_{n,k}(z_{n,k})$. The inference consists in minimising the KL divergence between the true posterior and the q approximation, which is equivalent to maximising the Evidence Lower Bond:

$$\mathscr{L} = \int_Z q(Z) \ln P(Y|Z) \mathrm{d}Z - \mathrm{KL}\left(q(Z)||p(Z)\right) \tag{2.33}$$

In the case of GP-LVMs however, the first term of this lower bond is in general intractable due to the potentially non-linear dependency in the latent variables, which makes the recipe of Section 2.2.6 unusable. Methods relying on auxiliary variational variables exist to circumvent this problem (Damianou et al., 2014; Titsias and Lawrence, 2010) but are beyond the scope of this thesis.

Optimisation of kernel hyperparameters

Kernel hyperparameters can be optimised as with standard GPs using gradient ascent to maximise the log marginal likelihood (Section 2.1.8)

Chapter 3

Modelling cell-cell interactions from spatial gene expression data with spatial variance component analysis

In this Chapter, we present Spatial Variance Component Analysis (SVCA), a statistical framework that aims to account for the spatial context of single cell expression data in a principled manner. SVCA accounts for multiple types of spatial and non-spatial effects on single cell expression variation, and in particular measures the effect of cell-cell interactions.

This work was co-supervised by Oliver Stegle and Julio Saez-Rodriguez. I developed and implemented the statistical method under the supervision of Oliver Stegle and I performed the simulations and the two real data analysis presented in this chapter, under the supervision of Oliver Stegle and Julio Saez-Rodriguez. Denis Schapiro and Bernd Bodenmiller from the University of Zurich contributed to the interpretation of the results in the first application (Imaging Mass Cytometry dataset). The SVCA software is open source and freely accessible here: http://github.com/damienArnol/svca. The paper is currently in revision and a preprint can be found on biorxiv: https://www.biorxiv.org/content/early/2018/03/27/265256.

3.1 Introduction

3.1.1 Spatial gene expression data

Experimental advances have enabled assaying RNA and protein abundances of single cells in spatial contexts, thereby allowing the study of single cell variation in tissues. Already, these

technologies have delivered new insights into tissue systems and the sources of transcriptional variation (Bodenmiller, 2016; Battich et al., 2013), with a potential use as biomarkers for human health (Bodenmiller, 2016).

Different technologies allow for generating spatially resolved expression profiles. Imaging Mass Cytometry (IMC) (Giesen et al., 2014; Chang et al., 2017) and Multiplexed Ion Beam Imaging (MIBI) (Angelo et al., 2014) rely on protein labelling with antibodies coupled to metal isotopes of specific masses followed by high-resolution tissue ablation and ionisation. IMC currently allows for the profiling of up to 37 targeted proteins with subcellular resolution. Other methods such as MxIF and CycIF use immunofluorescence for protein quantification of dozens of markers in single cells (Gerdes et al., 2013; Lin et al., 2015). Increasingly, there also exist fluorescence-based assays to measure single cell RNA levels in spatial context (Strell et al., 2018). Mer-FISH and seqFISH use a combinatorial approach of fluorescence-labeled small RNA probes to identify and localise single RNA molecules (Shah et al., 2016; Chen et al., 2015; Gerdes et al., 2013; Lin et al., 2015), which allows for a larger number of readouts (currently between 130 and 250). Even higher-dimensional expression profiles can be obtained from spatial expression profiling techniques such as Spatial Transcriptomics (Ståhl et al., 2016). However, they currently do not offer single cell resolution and are therefore not sufficient to study cell-to-cell variation.

3.1.2 Modelling the spatial context

The availability of spatially resolved expression profiles from a population of cells provides new opportunities to disentangle the sources of gene expression variation. Spatial context can for example be utilised to distinguish intrinsic sources of variation due to differences in cell types or states (Buettner et al., 2015), e.g. cell cycle stage (Scialdone et al., 2015), from sources of variation which relate to the spatial structure of the tissue, such as microenvironmental effects linked to the cell position (Fukumura, 2005), access to glucose or other metabolites (Meugnier et al., 2007; Lyssiotis and Kimmelman, 2017), or cell-cell interactions. To perform their function, proximal cells may interact via direct molecular signals (Sieck, 2014), adhesion proteins (Franke, 2009), or other types of physical contacts (Varol et al., 2015). In addition, certain cell types such as immune cells may migrate to specific locations in a tissue to perform their function in interaction with local cells (Moreau et al., 2018). In this thesis, *cell-cell interactions* is used as a general term to designate any of these phenomena.

More specific biological interpretations are discussed in Section 3.5.3 and Section 3.6.3.

While intrinsic sources of variation have been extensively studied, the cell-cell interaction component is arguably less understood and yet one of the most important, as it holds the promise to understand how genes are expressed in cells that participate in different tissue level functions. Yet, although experimentally spatial omics profiles can already be generated with high throughput, the required computational strategies for interpreting the resulting data are only beginning to emerge. Only a few methods quantify the impact of spatial features on the variance of individual genes, and even fewer methods specifically measure the effect of cell-cell interactions.

On the one hand, there exist methods to link the spatial position of cells to their expression profile. For example, there exist clustering methods that infer groups of cells from the same spatial location, solely based on their expression profiles (Achim et al., 2015). Other methods implement statistical tests of differential expression in space, which provide an overall assessment of the effect of the spatial topology on gene expression (Svensson et al., 2018a). However, none of these two approaches allow for directly quantifying cell-cell interactions.

On the other hand there exist methods which study cell-cell interactions, but only qualitatively or relying on discretisation steps which limit their interpretability or applicability. For example, some methods study tissue organisation by looking at the spatial cooccurrence of discrete cell types in predefined cellular neighbourhoods (Schapiro et al., 2017; Schulz et al., 2018). These approaches provide qualitative insights into interactions between cell types but they do not allow for quantifying their impact on individual genes. In contrast, some regression-based models assess interaction effects on individual gene expression levels, based on predefined features of cell neighbourhood (Goltsev et al., 2018; Battich et al., 2015). However the prior engineering of microenvironmental features relies on discretisation steps which are arbitrary and not always directly interpretable (see Section 3.2.7).

3.1.3 SVCA: Spatial Variance Component Analysis

Here, we present Spatial Variance Component Analysis (SVCA), a computational framework to model spatial sources of variation of individual genes. SVCA allows for decomposing

gene expression variation into intrinsic effects, environmental effects and, most importantly, an explicit cell-cell interaction component. In contrast to previous modelling approaches, the model uses the spatial coordinates directly and the continuous expression profiles of individual cells as inputs, thereby avoiding the need to define discrete cell types and microenvironmental variables.

We validate our model using simulated data, by showing that SVCA yields more accurate estimates of cell-cell interactions than alternative methods. We also illustrate the flexibility of SVCA by showing that it is more robust to confounding factors such as cell mis-segmentation.

We then illustrate SVCA using two real datasets from different technologies and biological domains: IMC proteomics profiles data from human breast cancer tissue (Schapiro et al., 2017) and spatial single-cell RNA profiles from the mouse hippocampus generated using seqFISH (Shah et al., 2017). Across these applications, we find that our model, and in particular the cell-cell interaction component, explains a major share of expression variability and facilitates the identification of biologically relevant genes and gene families participating in cell-cell interactions, such as glutamate receptors or cell junction genes in the brain.

3.2 SVCA: A spatial Gaussian Process model of gene expression variation

3.2.1 Overview of the model

SVCA builds upon the random effect framework to model variations in gene expression in terms of additive components from an intrinsic effect of the cell state, U_{int}, an environmental effect linked to the cell position, U_{env}, and an effect due to cell-cell interactions, U_{c-c}:

$$Y = U_{int} + U_{env} + U_{c-c} + \varepsilon \tag{3.1}$$

where Y is the vector of the expression level of a target gene across all cells of an image and $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$ is additive Gaussian noise.

These random effects are assumed to follow Normal distributions, defined by specific covariances which are functions of the cell spatial positions and expression profiles: $U_{int} \sim \mathcal{N}(0, K_{int})$, where K_{int} is a covariance matrix that determines the pairwise similarity between

cells based on intrinsic features; $U_{env} \sim \mathcal{N}(0, K_{env})$ where K_{env} is a measure of the similarity between the cell microenvironments, based on their spatial proximity; $U_{cc} \sim \mathcal{N}(0, K_{c-c})$ where K_{c-c} measures the similarity between cellular neighbourhoods, in order to account for cell-cell interactions. The marginalisation over the additive effects in Equation 3.1 results in the following Gaussian Process: $Y = \mathcal{N}(0, K_{int} + K_{env} + K_{c-c} + \sigma_\varepsilon^2 I)$. (Fig. 3.1)

For every individual gene, we use Maximum Likelihood to assess the relative importance of the four additive terms of the covariance in explaining the observed distribution of the gene across cells. The fitted model can be used to estimate the fraction of variance explained by each term using Gower factors. This results in a breakdown of the variance of every gene into spatial and non-spatial components, which gives a compact overview of the effect of the spatial structure of the tissue on gene expression levels (Fig. 3.1). We call this representation a spatial variance signature.

In the next sections, we give the exact definition of the covariance terms of the SVCA model (Fig. 3.2), explain how they are parametrised and how the resulting model is fitted to the data.

3.2.2 Nomenclature and notation of the SVCA model

To describe the SVCA model, we will use the following nomenclature and notation.

Nomenclature

Gene of interest	Individual molecule, typically a gene or protein, whose expression profile the SVCA model is fitted on.
Cell state	Intrinsic characteristic of a cell. Here, we consider the overall expression profile excluding the gene of interest as a multidimensional and continuous measure of cell state. Other possibilities include the discrete classification of cells into cell types.
Cellular neighbourhood matrix	Continuous measure of the cellular neighbourhood, which aggregates, for each cell, the molecular composition of the neighbouring cells. This is achieved by weighting the molecular profiles of all neighbouring cells by a squared exponential function of the distance.

Spatial variance signature	Concatenation of all variance estimates (intrinsic effect, environmental effect, cell-cell interaction effect and residual noise) across every molecule for a given image.

Mathematical Notation

N	Number of cells in a given image
D	Number of molecules (e.g. genes or proteins) in a given image
Y	Expression level of the gene of interest in all cells. Dimensions: $N \times 1$
X	Cell state matrix made of the expression profile of all genes excluding the gene of interest. Dimensions: $N \times (D-1)$.
$d_{i,j}$	Euclidean distance between cells i and j
K_{int}	Cell-cell covariance for the intrinsic effect. Dimensions: $N \times N$
K_{c-c}	Cell-cell covariance for the cell-cell interaction effect. Dimensions: $N \times N$
K_{env}	Cell-cell covariance for the environmental effect. Dimensions: $N \times N$

3.2.3 Definition of the covariance terms

Intrinsic term

The covariance term K_{int} uses as input the expression profile X of all genes excluding the gene of interest (dimension $N \times (D-1)$), which represents a continuous measure of intrinsic cell states (Fig. 3.2). The covariance function used is linear:

$$K_{int} = \sigma^2_{int} XX^T \tag{3.2}$$

As seen in Section 2.1.7, this covariance term corresponds to a Bayesian linear regression that models the effect of the cell expression profile on the expression of the gene of interest: $Y_i = \sum_{d \neq c} X_{i,d} \beta^{int}_d$, where c denotes the index of the gene of interest, with the following Normal prior on the effect sizes: $\beta^{int} \sim \mathcal{N}(0, \sigma^2_{int} I)$.

This covariance function has a scaling hyperparameter σ^2_{int}, which is proportional to the variance explained by this component.

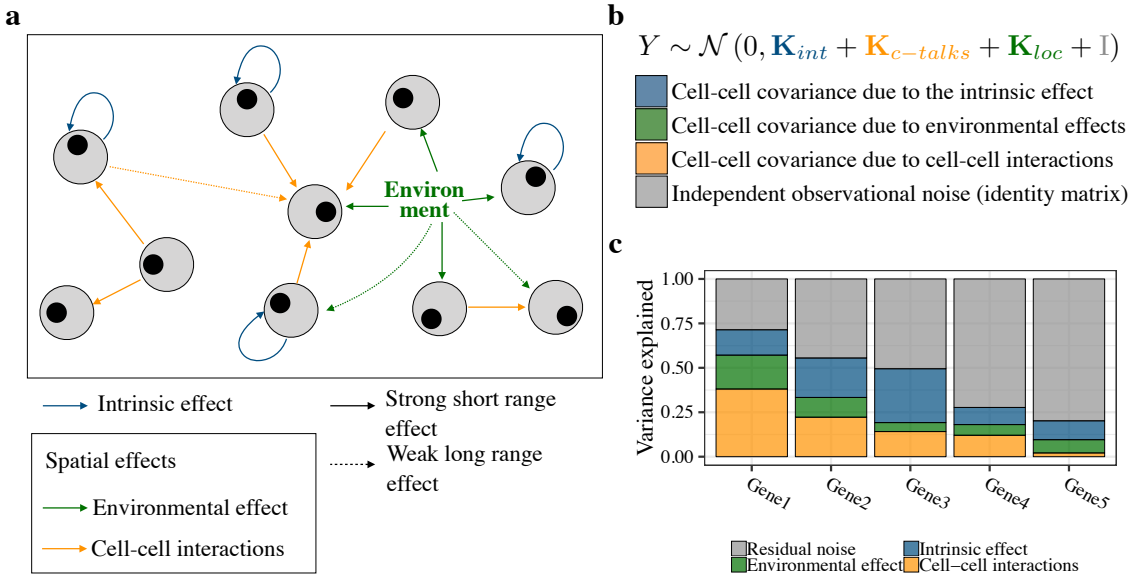


Fig. 3.1 Spatial Variance Component Analysis (SVCA). (a) SVCA decomposes the variability of individual genes into i) an intrinsic component capturing the effect of the intrinsic cell type or state (blue), ii) an environmental component capturing the effect of non-specific local factors dependent on the cell position (green) and iii) a cell-cell interaction component capturing the effect of the cellular composition of the cell neighbourhood (yellow), thereby accounting for interactions between neighbouring cells. The strength of a spatial effect on a given cell depends on its distance to the source, as symbolised by the solid and dotted lines. (b) SVCA builds on a Gaussian Process framework, defining additive covariance components to explain the different effects. Details on the definition of the corresponding covariance terms are given in Section 3.2.3 and Figure 3.2 (c) SVCA spatial variance signature: gene-level break down of the proportion of variance attributable to the different components.

Environmental term

The second term, K_{env}, aims at modelling the positional effect explained by the cell location in the tissue on the expression profile of the gene of interest. This variance component can for example model the effect of the local microenvironment (such as oxygen access). These microenvironmental factors are not measured explicitly, however the cell position can be used as a proxy, as nearby cells will be exposed to the same environmental factors. We use the squared exponential covariance defined on the relative spatial position of the cells (see Section 2.1.7), where $d_{i,j}$ is the distance between the centroids of cells i and j. (Fig. 3.2):

$$(K_{env})_{i,j} = \sigma^2_{env} \frac{-d^2_{i,j}}{2l^2} \tag{3.3}$$

The squared exponential covariance function is widely used for modelling non-linear spatial and temporal dependencies and is highly flexible. This covariance function has a scaling hyperparameter σ^2_{env} and a length scale hyperparameter l (see Section 2.1.7).

Defined as such, the environmental covariance may capture any spatially correlated residuals with no specific biological interpretation. For example in Section 3.4, we will discuss its role in capturing segmentation errors between proximal cells or cells in physical contact. If there is no spatial structure, and the length scale tends to zero, the environmental covariance may also become unidentifiable with noise, as we will discuss in Section 3.5.2. For these reasons, we will only use the environmental covariance as a normalisation term and we will not be concerned with its interpretation.

Cell-cell interaction term

The cell-cell interaction covariance term K_{c-c} explains the effect of the cellular composition of the neighbourhood of a given cell on gene expression. This component can in particular explain interactions between cells. Borrowing ideas from social genetic effect studies (Baud et al., 2017), we define a covariance function which measures the similarity between cells based on their cellular neighbourhood, which we compute by aggregating, for each cell, the molecular composition of all other cells, while down-weighting interactions between cells which are further apart, using a squared exponential function of the distance: ZX, where $Z_{i,j} = \exp\left(-d^2_{i,j}/2l^2\right)$ (Fig. 3.2).

$$K_{c-c} = \sigma^2_{c-c} ZXX^T Z^T \tag{3.4}$$

This covariance function has a scaling hyperparameter σ^2_{c-c} and a length scale hyperparameter l.

Residual noise

The residual noise term is a diagonal matrix with a scaling hyperparameter σ^2_ε. The assumption is that, for a given gene, the residual noise is independently identically distributed across cells. For these assumptions to be met, appropriate data normalisation and variance stabilisation are necessary processing steps. Additionally, the independence assumption may not hold in the case of imaging batch effects or cell mis-segmentation, which may result in correlated errors between neighbouring cells. In SVCA, this problem is addressed by the environmental component, which captures spatially correlated residuals, as we further

discuss in Section 3.4.3.

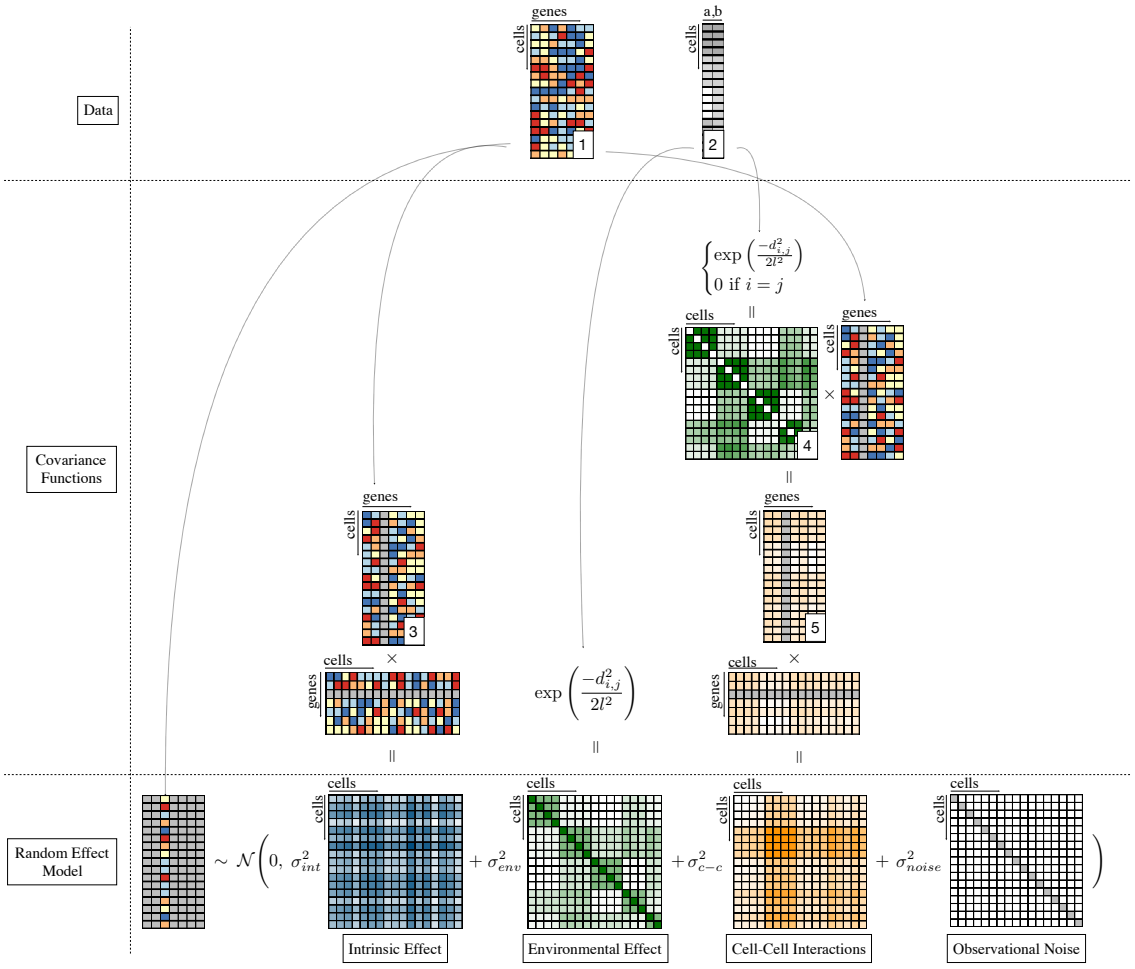


Fig. 3.2 SVCA model definition. SVCA takes as input a single cell expression data as a cell times gene matrix (1) and a matrix of the cells spatial coordinates (a,b) (2). The expression profile of individual genes is modelled as Normally distributed, with additive covariance components that account for intrinsic, environmental and cell-cell interaction effects. The intrinsic effect covariance is computed as the empirical covariance of the expression profiles between cells, using all genes except the gene of interest (3). The environmental effect is modelled using a Squared Exponential covariance defined on the relative distance between cells. The cell-cell interaction covariance measures the similarity between cell cellular neighbourhoods that aggregate, for each cell, the molecular composition of the neighbouring cells (5). This aggregation step can be written as a product between a squared exponential covariance matrix whose diagonal elements were removed (4) and the expression matrix (3).

3.2.4 Model fitting - optimisation of hyperparameters

The hyperparameters of SVCA are optimised by maximising the log likelihood of the data (type-II maximum likelihood (Rasmussen and Williams, 2006, Chap. 5), see Section 2.1.8).

The scaling hyperparameters $\{\sigma^2_{int}, \sigma^2_{env}, \sigma^2_{c-c}, \sigma^2_{\varepsilon}\}$ are optimised with gradient ascent using an lbfgs optimiser (Bonnans et al., 2006), while the common length scale hyperparameter[1] l of the environmental and the local terms is optimised with a grid search strategy, to avoid possible local optima. Specifically, the scaling hyperparameters are refitted for every length scale of the grid independently, and the set of hyperparameters (scaling and length scale) providing the best log-marginal likelihood is selected at the end of the training.

SVCA is implemented in the *limix* framework (Lippert et al., 2014), a software which is mostly used for linear mixed models in genetics, but also has more general Gaussian Process capabilities. The derivation of the gradient of the cell-cell interaction covariance is given in Appendix A, while the remaining components are based on linear and squared exponential covariance functions which are standard in any Gaussian Process framework.

3.2.5 Variance estimates

We used Gower factors (see Section 2.1.9) to estimate the variance explained by each term of the model:

$$\text{var}_{\text{effect}} = \frac{G(K_{\text{effect}})}{\sum_{\text{eff} \in \text{ other effects}} G(K_{\text{eff}})} \tag{3.5}$$

Computing these variance estimates for every gene and effect enables us to compute the spatial variance signatures schematised in Figure 3.1, which gives a compact representation of the effect of spatial and non-spatial drivers of variation.

3.2.6 Significance of the variance components

As the individual components have additive contributions, the significance of any term can be assessed using the Log Likelihood Ratio (LLR) between the SVCA model and a reduced model omitting the tested covariance term (in fact a log ratio between the marginal likelihoods

[1] Choosing a common length scale for the environmental and the cell-cell interaction covariance functions was necessary to reduce the size of the grid-search

of the two models, see Section 2.1).

Given that the reduced model is nested in the full SVCA model, we rely on Wilks theorem (Wilks, 1938), which states that under the null hypothesis, the LLR statistics should follow a χ^2 distribution. To calibrate this statistic, we fit the degrees of freedom of the χ^2 distribution to an empirical null distribution of LLRs. To obtain this null distribution for a given gene, we fit the null model to the data and simulate 100 expression profiles from the fitted Normal distribution under the null hypothesis. By fitting SVCA and the alternative model to these 100 data points, we obtain a null distribution of LLRs for the considered gene on which we fit a χ^2 distribution using an off-the-shelf non-linear optimisation method (*nloptr* in R). The significance of a covariance term is then assessed by comparing LLRs to the empirical χ^2 distribution (Casale et al., 2017; Bůžková et al., 2011).

Unless stated otherwise, we use the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to adjust for multiple testing across genes and images.

3.2.7 Comparison with related models

In comparison to alternative methods, SVCA has a number of unique features. First, it does not require assigning cells to discrete types, but instead is based on a continuous measure of cell-cell similarities that are directly estimated from cell expression profiles. The model also circumvents the need to define discrete neighbourhoods but instead weights interactions between any pair of cells as a function of their distance. Both advantages are tackling the problem of cellular classification and neighbourhood definition by providing a continuous representation of space and cellular identity in a unique modelling framework (Wagner et al., 2016). Additionally, SVCA includes a non-linear environmental component which captures non-specific spatial effects, ensuring robustness to confounders in the identification of cell-cell interactions, as we will show with simulations.

This section provides a more detailed comparison of SVCA to related models which have been proposed for spatial expression analysis.

Schapiro et al 2017 - HistoCAT

HistoCAT (Schapiro et al., 2017) aims at measuring spatial cooccurrence of different cell types. Briefly, cells of one or multiple images are classified into discrete cell-types based

on their expression profile using a clustering algorithm. For every cell, a neighbourhood is defined as containing all cells within a fixed distance threshold (measured from membrane to membrane). Using this fixed neighbourhood definition, histoCAT counts the number of occurrences of a given pair of cell types, in the same neighbourhood. This number is then compared to a null distribution obtained from permuting the cell positions, which gives a P value for positive and negative cell types interactions.

HistoCAT therefore provides an understanding of tissue structure in terms of cell types co-occurrence. While these co-occurrences are probably related to functional interactions between cells, histoCAT, unlike SVCA does not quantify the effect of these interactions on the expression profile of individual genes.

Battich et al 2015

Battich et al. (2015) uses a regression approach to measure the effect of the cell microenvironment on individual expression levels. Briefly, 183 features are collected, quantifying intrinsic cell properties and microenvironmental properties. Microenvironmental features namely account for local cell crowding, number of adjacent neighbours, intercellular space around the cell, as well as the molecular profile of the neighbours, based on a fixed distance threshold. The dimensionality of this feature set is then reduced using principal component analysis (PCA), and single cell expression profiles are modelled with a fixed effect linear model with the first 20 PCs as covariates. The PCs are then a posteriori linked to the microenvironmental features of interest. Biological replicates are used to quantify the amount of variance explained by each covariate using out of sample prediction.

This method therefore quantifies directly the effect of microenvironmental features including cell-cell interactions. Unlike SVCA however, it relies on a definition of discrete microenvironmental features and the definition of fixed parameters such as a distance threshold to define a cell neighbourhood, which limits the applicability of the method to general spatial data. In addition, standard linear regressions have other limitations like the inability to capture spatially correlated measurement errors, as we further discuss in Section 3.4.3.

Goltsev et al 2018

The approach from Goltsev et al. (2018) also relies on the definition of discrete microenvironmental variables, used in a fixed effect linear model to predict the expression level of

individual markers out of sample. In contrast to Battich et al. (2015), microenvironmental variables are not defined directly based on the molecular profile of neighbouring cells, but based on the cell-type composition of the neighbourhood. The different neighbourhood cell-type compositions are clustered into discrete i-niches, used as an input for the linear model.

This method therefore enables us to quantify directly the effect of cell-cell interactions on individual molecular profiles of single cells. However, it again relies on a priori definition of microenvironmental variables, this time based on discrete cell-type assignments.

3.3 Validation of SVCA using simulations from the generative model

In this section, we consider simulated data to validate SVCA, and the identification of cell-cell interactions in particular.

3.3.1 Simulation procedure

We used empirical parameters derived from 11 real datasets (see Section 3.5), including their gene expression profiles and cell positions, as well as ranges of fractions of variance explained by different components that reflect those observed in real data.

The experimental workflow was as follows:

i) Fitting the SVCA model to a real dataset

ii) Simulating data from a multivariate Normal distribution, with a covariance made of:

- the intrinsic covariance from the fitted model
- the environmental covariance from the fitted model
- the noise covariance from the fitted model
- a cell-cell interaction covariance which is a rescaled version of the one fitted to the data: $Y = \mathcal{N}(0, \hat{K}_{int} + k_{sim} \times \hat{K}_{c-c} + \hat{K}_{env} + \hat{\sigma}^2_{\varepsilon} I_n)$, where $\hat{K}$ represents a fitted covariance term and k_{sim} is the rescaling term for cell-cell interactions.

iii) Refitting SVCA to the simulated data

The rescaling of the cell-cell interaction covariance term provides a ground truth value for the fraction of variance explained by cell-cell interactions. For a cell-cell interaction effect explaining a fraction $\eta \in [0.1;0.9]$ of variance in the simulated data, it can be shown that the rescaling factor is:

$$k_{sim} = \frac{\eta}{1-\eta} \times \frac{G(\hat{K}_{int} + \hat{K}_{c-c} + \hat{K}_{env} + \hat{\sigma}_{\varepsilon}^2 I_n)}{G(\hat{K}_{c-c})} \tag{3.6}$$

3.3.2 Accuracy of cell-cell interaction estimates

We simulated expression profiles with a cell-cell interaction component that explains a variance $\eta \in [0.1;0.9]$ in the generated data. 10 repeat experiments were performed for 11 images and 26 proteins, and for every value of η. Variance estimates were then averaged across these 110 data points for every protein and every value of η. We then compared variance estimates for cell-cell interactions obtained with SVCA to the ground truth values.

We found that SVCA yielded accurate cell-cell interaction estimates, although slightly conservative, especially for high simulated cell-cell interactions (Fig. 3.3).

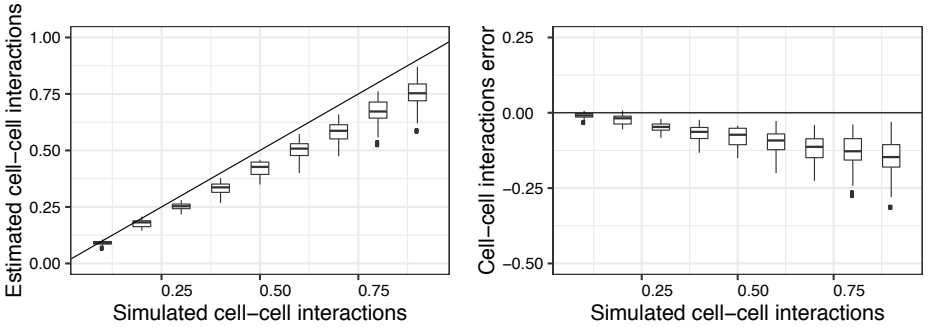


Fig. 3.3 Left: Fraction of variance due to cell-cell interactions estimated by SVCA when varying the simulated fraction of variance explained by cell-cell interactions. Right: error in cell-cell interaction estimates as a function of the simulated cell-cell interactions.

3.3.3 Statistical calibration

Then, we simulated expression profiles with no cell-cell interactions ($k_{sim} = 0$) to assess the calibration of the corresponding Log Likelihood Ratio test. Again, we used 10 repeat experiments for 11 real images and 26 proteins used for simulations. For varying P value thresholds, we computed a False Discovery Rate of the statistical test across the resulting

110 data points for the 26 proteins independently.

Results show that the statistical test for cell-cell interactions significance is conservative (Fig. 3.4).

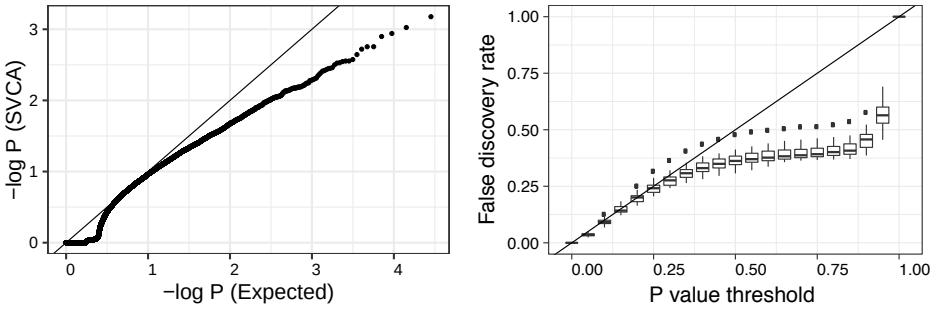


Fig. 3.4 We performed 10 repeat experiments for 11 images and 26 proteins with no cell-cell interactions and computed P values for the cell-cell interaction test. Left: log P values as a function of the expected value under the null hypothesis Right: empirical false discovery rate for the cell-cell interaction test (FDR) as a function of the P value threshold.

3.3.4 Statistical power

Finally, we assessed the detection power of the cell-cell interaction statistical test, when varying the variance explained by this effect, and when varying the number of cells in the dataset based on subsampling. We performed 10 repeat experiments for every value of the simulated cell-cell interactions, for every fraction of cells used and each time considering the same set of 11 images and 26 proteins used for simulations. True Positive Rates were computed across the 110 data points for the 26 proteins independently (Fig. 3.5).

We found that SVCA Log Likelihood Ratio test had little power, with, on average, a true positive rate of 50% and below for cell-cell interactions below 20%. This limitation is discussed in Section 3.7.1.

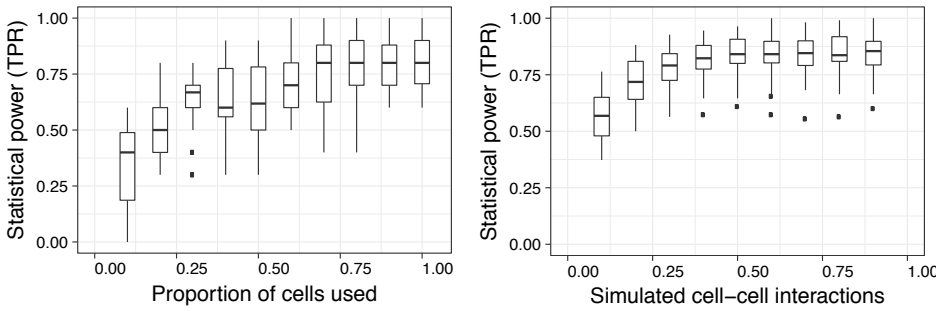


Fig. 3.5 Left panel: True positive rate as a function of the variance explained by cell-cell interactions using all cells of the images. Right panel: True positive rate as a function of the proportion of cells used in the images (downsampling) for a cell-cell interaction terms explaining 30% of variance ($\eta = 0.3$).

3.4 Benchmarking of SVCA in comparison to alternative linear regressions

3.4.1 Simulation setting

In order to compare SVCA to baseline regression models, we considered simulating expression profiles from a linear model accounting for an intrinsic effect, a cell-cell interaction effect variable in size and Gaussian noise. In addition, we simulated a confounding effect due to cell mis-segmentation (Fig. 3.6).

Analogously to the approach from Section 3.3, we used empirical parameters derived from 11 real datasets (see Section 3.5), including their gene expression profiles and cell positions to generate in silico target genes, using a linear model accounting for an intrinsic effect and a cell-cell interaction effect variable in size.

The expression profile of an in-silico target gene Y is generated using the true expression profile across all observed genes X and the following linear model:

$$Y = \underbrace{\sqrt{\eta_{c-c}} Z X \beta_{c-c}}_{\text{Cell-cell interactions}} + \underbrace{\sqrt{1-\eta_{c-c}} (X \beta_{int} + \varepsilon)}_{\text{Intrinsic effect}} \tag{3.7}$$

where X is a matrix of dimension $N \times D$ which corresponds to the empirical expression profiles of the real data considered, β_{c-c} and β_{int} are effect sizes drawn from standard Normal distributions and $\eta_{c-c} \in [0;1]$ corresponds to the variance explained by cell-cell interactions

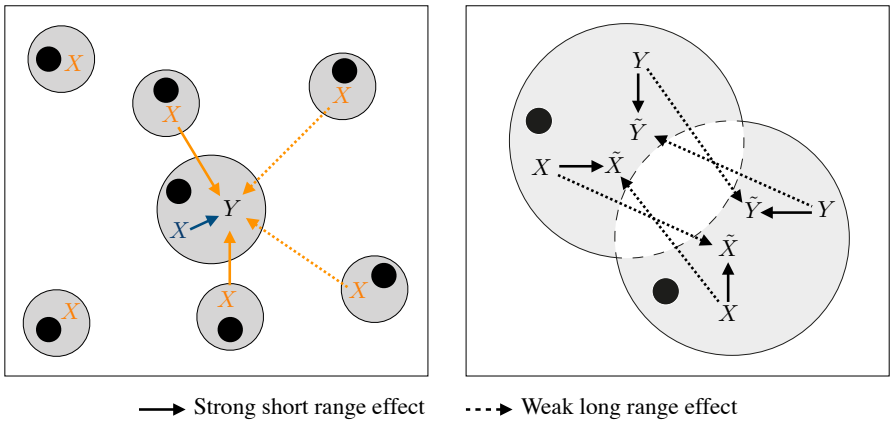


Fig. 3.6 Simulation approach for comparing SVCA with alternative regression models. Left: Expression levels are generated from a linear model including an intrinsic term and a cell-cell interaction component to which participate only the the first N_{nn} neighbours (here $N_{nn} = 4$). Right: Modelling approach for cell mis-segmentation: expression profiles of two mis-segmented cells are perturbed by receiving a fraction of each other's expression profile: $\tilde{Y} = \mu Y + (1-\mu)Y_{neighbour}$, with $\mu \in [0;1]$

in the simulated data. ε is a standard Gaussian noise, $\varepsilon \sim \mathcal{N}(0,I)$

Z is a function which weights the contribution of the expression profiles of the N_{nn} nearest neighbours of each cell to the expression level of Y:

$$Z_{i,j} = f(d_{i,j}) = \begin{cases} 1/d_{i,j}^2 & \text{if the cell } j \text{ is one of the } N_{nn} \text{ nearest neighbours of } i \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

$d_{i,j}$ is the distance between cells i and j. N_{nn} is the number of neighbours considered for cell-cell interactions. This parameter was also varied across simulations to assess the robustness of the results to multiple underlying cell-cell interaction structures.

Simulation of cell mis-segmentation

To simulate mis-segmentation, the expression profiles generated from the linear model were further perturbed by receiving a share of the expression profiles of mis-segmented neighbouring cells.

For every cell in the image, an arbitrary number of two cells were randomly drawn as mis-segmented with the focal cell. To model our assumption that the closer the cells, the more likely they were to be mis-segmented, the probability for a cell j to be mis-segmented with the focal cell i was taken from the probability vector:

$$p_{i,j} = \frac{1/d_{i,j}^2}{\sum_j 1/d_{i,j}^2} \tag{3.9}$$

To model mis-segmentation, the expression level Y_i in a given cell i is then transformed into a weighted average between itself and the average signal from the cells which are mis-segmented with cell i, which we call signal spillover:

$$\tilde{Y}_i = \sqrt{1-\eta_{mis}} Y_i + \underbrace{\sqrt{\eta_{mis}} \text{mean}_j(Y_j)}_{\text{signal spillover}} \tag{3.10}$$

where $\text{mean}_j(Y_j)$ is the mean of the expression profile of Y in the cells which are mis-segmented with the cell i. η_{mis} controls the effect size of the mis-segmentation.

Assuming that all genes are affected in the same way by mis-segmentation, which is reasonable but does not account for different sub-cellular localisation of genes, the same perturbation was applied to the expression level matrix X:

$$\tilde{X}_{i,:} = \sqrt{1-\eta_{mis}} X_{i,:} + \underbrace{\sqrt{\eta_{mis}} \text{mean}_j(X_{j,:})}_{\text{signal spillover}} \tag{3.11}$$

The size of this mis-segmentation effect was also varied through multiple simulations and the perturbed profiles $\tilde{X}$ and $\tilde{Y}$ were used to fit SVCA and alternative models.

3.4.2 Alternative models

We compared SVCA to four alternative regression models, which included an intrinsic component and a cell-cell interaction component.

Three of the four models were based on linear regressions with Ridge regularisation. The intrinsic effect was modelled as a linear combination of the expression profile of all genes measured in the cell, excluding the gene of interest. Cell-cell interactions were accounted for in three alternative ways:

i) Using all the cells in the images and weighting their impact by a function of the distance $f(d_{i,j}) = 1/d_{i,j}^2$.

ii) Using an average of the expression profiles of the first five neighbouring cells.

iii) Using a weighted average of the expression profiles of the first five neighbouring cells, with the same weighting function $f(d_{i,j}) = 1/d_{i,j}^2$.

The coefficient of the Ridge regularisation was determined using cross-validation using the *RidgeCV* function from the *scikit-learn* package (Pedregosa et al., 2011) with default parameters.

Additionally, we considered a reduced Gaussian Process model containing all the covariance terms of SVCA apart from the local effect.

3.4.3 Results

Cell-cell interaction accuracy

The simulation setting described before was used to generate data with cell-cell interaction effects η_{c-c} ranging from 0 to 0.25, which reflected the range of values observed on real data applications (see Section 3.5 and Section 3.6). The mis-segmentation effect was fixed to $\eta_{mis} = 0.2$ and the number of neighbours used for simulating cell-cell interactions was $N_{nn} = 5$

The variance estimates obtained by SVCA and the alternative models were compared to the simulated ground truth (Fig. 3.7), using an in sample coefficient of determination r^2 for the linear regressions. Results showed that SVCA was more accurate than alternative models.

Spurious cell-cell interaction effects from cell mis-segmentation

We also assessed cell-cell interaction variance estimates when simulating data with $\eta_{c-c} = 0$ (no cell-cell interactions), and mis-segmentation effects η_{mis} ranging from 0 to 0.25. We observed that mis-segmentation yielded the inference of spurious cell-cell interaction effects, and that SVCA was the most robust model in the presence of this confounder. The comparison with the reduced Gaussian Process with no local term showed that this term is critical for capturing the mis-segmentation effect.

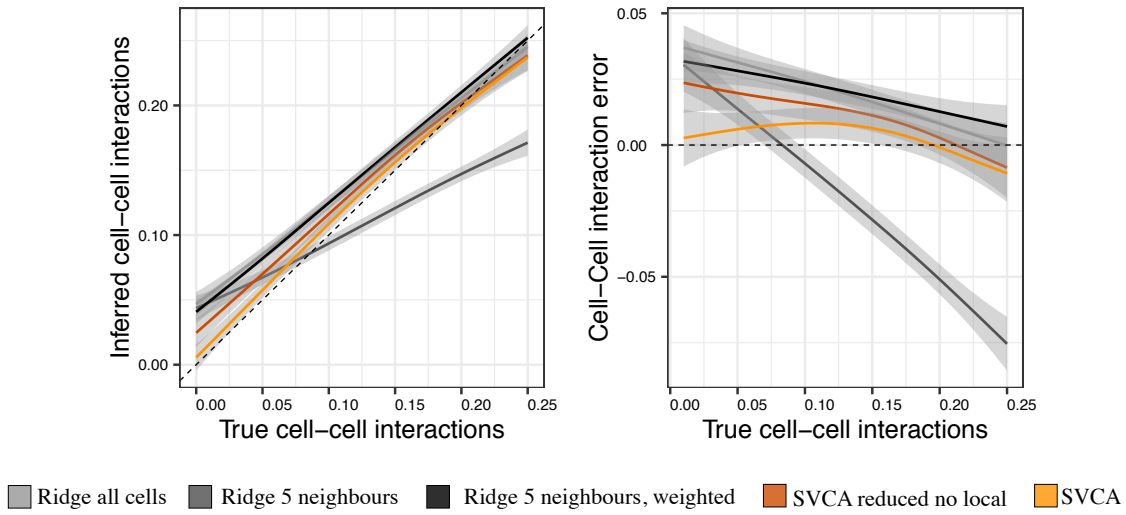


Fig. 3.7 Accuracy of SVCA cell-cell interaction estimates compared to alternative regression models. Left panel: inferred cell-cell interaction variance estimates as a function of the simulated cell-cell interactions. Values correspond to proportions of variance explained. Right panel: error in the variance estimates as a function of the simulated cell-cell interactions. Values on the x-axis correspond to proportions of variance explained, values on the y-axis correspond to the absolute error between the simulated proportions and the estimations from the different models.

Robustness to different simulation settings

Finally, we considered the robustness of the results across multiple cell-cell interaction ranges, varying the simulated number of neighbours N_{nn} involved in cell-cell interaction effects. We computed the average error in the inferred variance estimates across all simulations for $\eta_{c-c} \in [0, 0.25]$ and for $\eta_{mis} = 0.2$. Results show that SVCA was the method with the lower bias (error more centred on zero) for all simulation settings (Fig. 3.9).

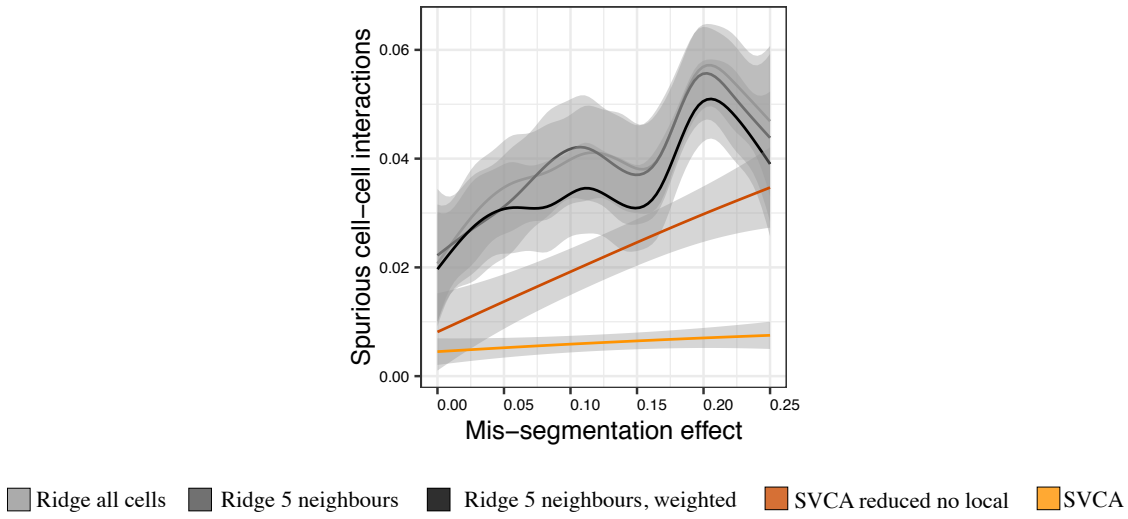


Fig. 3.8 Spurious cell-cell interaction variance estimate as a function of the mis-segmentation effect size. Values on the x-axis correspond to the proportion of variance explained by mis-segmentation in the simulated data, values on the y-axis correspond the cell-cell interaction variance estimates yielded by the different models. The data is generated with no cell-cell interactions.

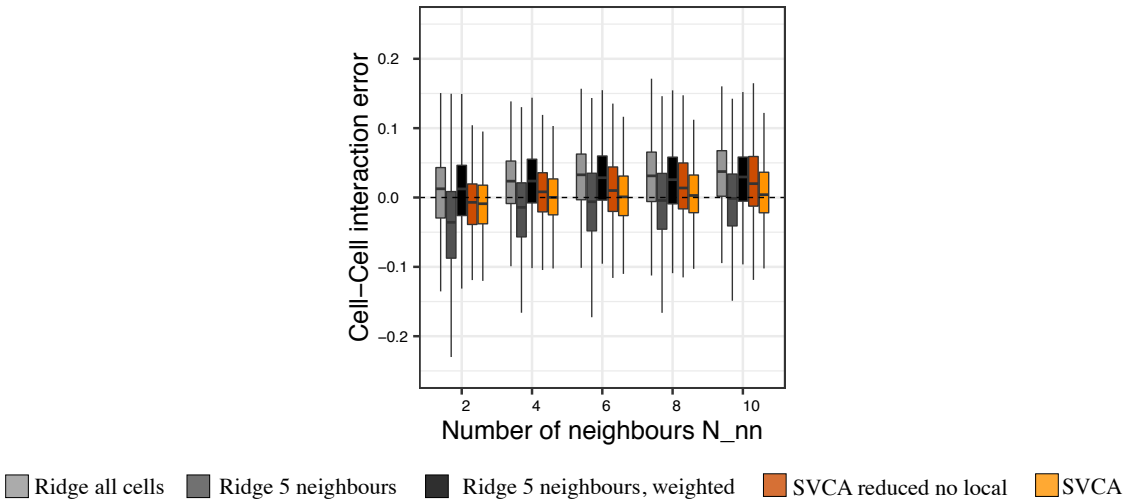


Fig. 3.9 Absolute difference between inferred cell-cell interaction variance estimates and simulated values as a function of the numbers of neighbours N_{nn} used in the simulations. Values are averaged across the whole range of simulated variance components.

3.5 Application of SVCA to Imaging Mass Cytometry breast cancer data

Here, we present results obtained from applying SVCA to an Imaging Mass Cytometry (IMC) dataset from human breast cancer (Schapiro et al., 2017). We first give a brief overview of the experimental method and the processing steps performed to normalise the raw data. Then we present the results obtained with SVCA and discuss their biological interpretation.

3.5.1 Experimental method and data processing

Imaging Mass Cytometry (IMC) allows sub-cellular resolution measurements of the abundance of up to 37 proteins (Giesen et al., 2014). In a paraffin embedded tissue, targeted proteins are labelled with specific antibodies coupled with metal isotopes of distinct masses. The tissue is then laser ablated into a sub-cellular resolution grid of so-called voxels of dimension $1\mu m \times 1\mu m$, and subsequently injected into a CyTOF (Kay et al., 2016), which measures the protein abundances based on the detection of the metal isotopes. This results in protein counts in each voxel, which are aggregated into single cell expression levels after cell segmentation, for example using cell profiler (Sommer et al., 2011; Schüffler et al., 2015; Carpenter et al., 2006).

Raw data description

We analysed a dataset of 46 breast cancer biopsies from 23 patients (Schapiro et al., 2017) (6 images were removed from the original dataset as they exhibited one or multiple markers with no variance across pixels, indicating failure in abundance measurements). Clinical data was available for 38 of these images: (ER status, PR status, Her2 status, Grade). The images contained between 267 and 1455 cells, with an average of 900 cells and 26 proteins were targeted. An example of an IMC image, visualised for three proteins, is shown in Figure 3.10

Preprocessing

We first quantified the expression level of every given cell using the median of the signal across all voxels assigned to it. The single cell counts obtained exhibited over-dispersion (Fig. 3.11), which motivated using the Anscombe transformation for variance stabilisation

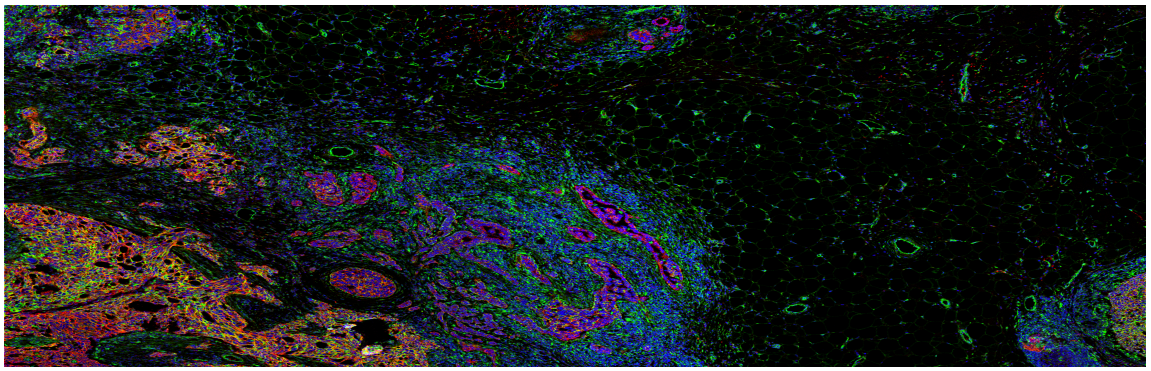


Fig. 3.10 Example of IMC image for a breast cancer sample. Proteins shown are E-cadherin in Red, Vimentin in Green and Histone H3 in blue.

of Negative Binomial data (Anscombe, 1948). Specifically, the dispersion parameter ϕ in the negative binomial mean-variance equation $\sigma^2 = \mu + \phi\mu^2$ was optimised using gradient descent and the following log transformation was applied to the data: $y = \log(x + 1/2\phi)$. The resulting expression profiles were then normalised by regressing out the log of the total signal in the cell to remove possible batch effects resulting in a higher detection rate across all proteins for some cells.

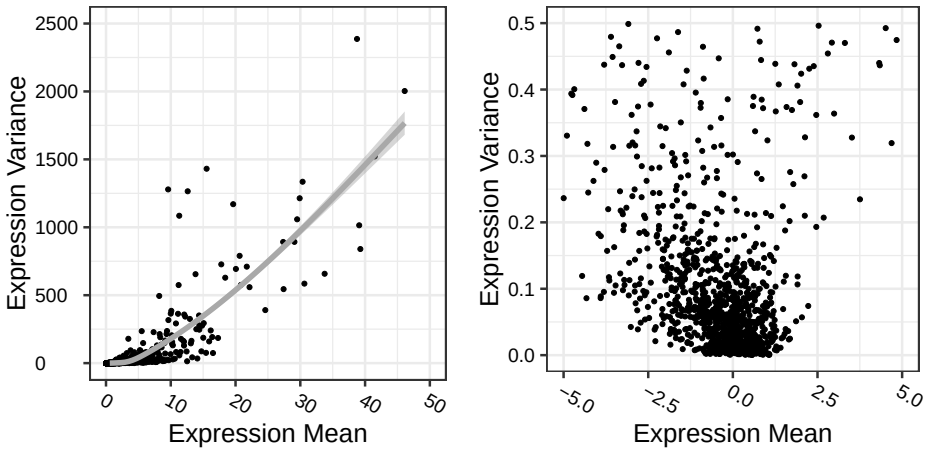


Fig. 3.11 Mean Variance relationship of single cell expression levels for the IMC data. Left: Before data processing. Right: after data processing. Means and variances are computed across cells, for every image and every protein independently (one dot per protein and image). A few outlying plots were removed for clearer visualisation of the relationship.

Figure 3.12 shows the impact of the two processing steps on the distribution of the protein expression levels across cells and images, while Figure 3.13 shows that the data processing

reduces the correlation between proteins across cells.

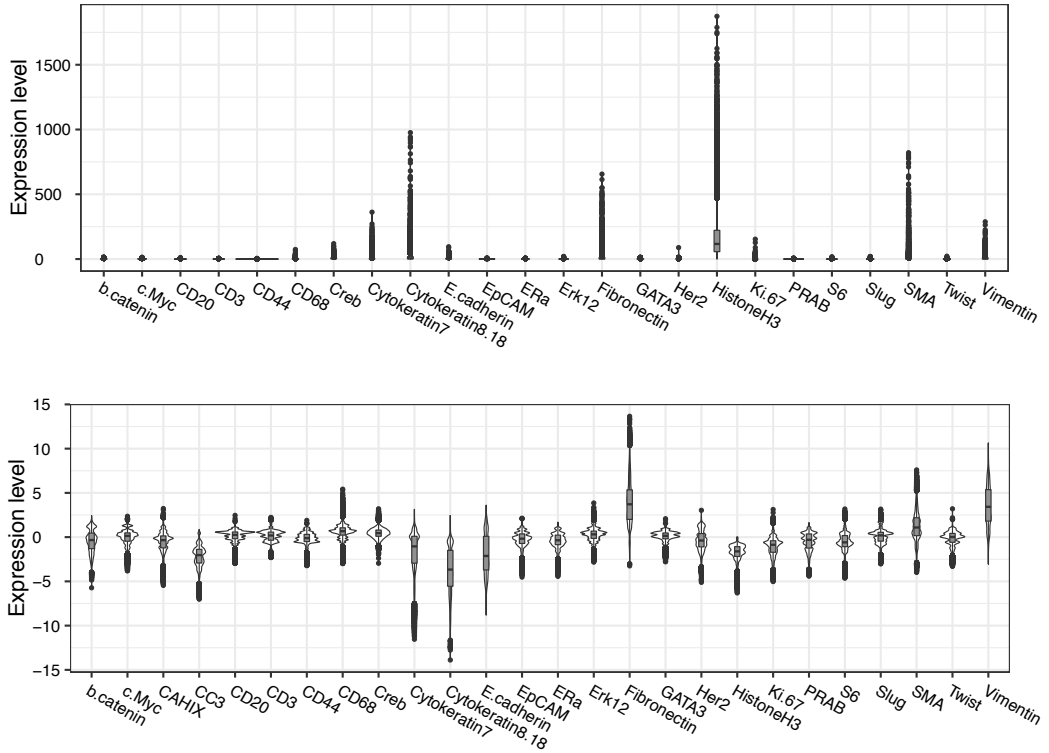


Fig. 3.12 Protein expression profiles across cells and images Top: before processing. Bottom: after processing.

Before fitting SVCA, the stabilised expression profile of the target gene Y was subsequently ranked standardised and transformed into Normally distributed data using the probit function, in order to reduce sensitivity to outliers. Note that this last step was performed for the target gene Y only, so that the true scale and distribution of the input genes are preserved.

3.5.2 SVCA variance signatures

We computed the SVCA spatial variance signatures for every image independently. Figure 3.14 shows the average variance components across all 46 images. We also computed the statistical significance of the cell-cell interaction component, corrected for multiple testing across proteins and images using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) (Fig. 3.14).

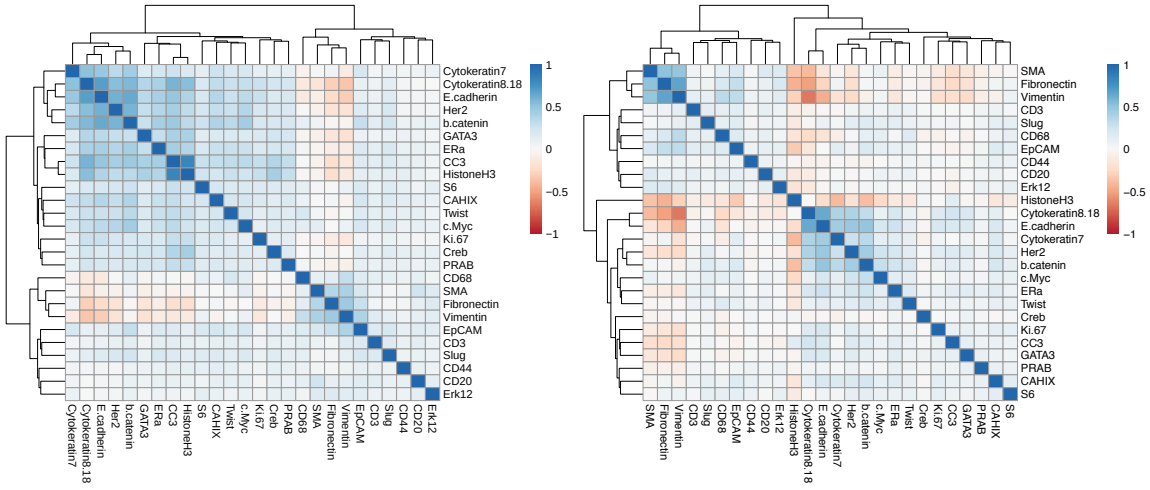


Fig. 3.13 Matrix of the Pearson correlation between protein expression levels. Left: raw data. Right: processed data. Correlations between proteins are computed within each image independently. Shown is the average value across images for each protein pair

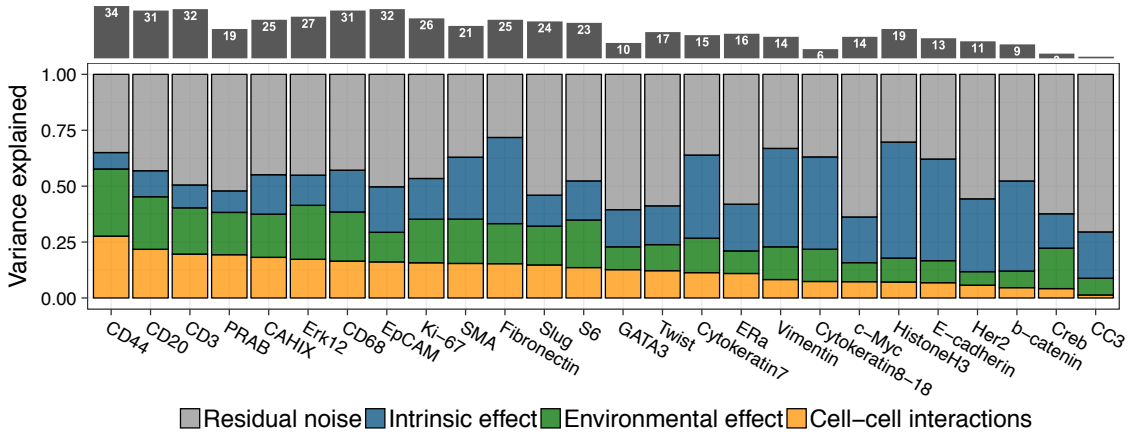


Fig. 3.14 Top panel (grey): number of images with a significant cell-cell interaction component for the corresponding proteins (out of 46 images; FDR < 1%). Bottom panel: SVCA spatial variance signatures averaged across images for the IMC breast cancer dataset. Proteins are ordered by the magnitude of the cell-cell interaction component.

Out of sample predictions

We performed out of sample predictions of gene expression profiles in test cells sampled at random (details in Appendix A, Section A.1.2[2]), to validate the inferred variance estimates, finding that SVCA yielded more accurate gene expression profile imputations than the alternative linear regressions discussed in the previous section, as well as reduced Gaussian Process

[2]Appendix A, Section A.1.2 explains the marginalisation property of Gaussian Processes, and how out of sample prediction is implemented in the context of SVCA so as to preserve it

models ignoring cell-cell interactions (Fig. 3.15). The only small difference between SVCA and a reduced Gaussian Process accounting for a local effect but no cell-cell interactions may arguably be due to the capacity of the local term to capture non-specific spatial effects, including cell-cell interactions when they are not modelled explicitly (see Appendix A).

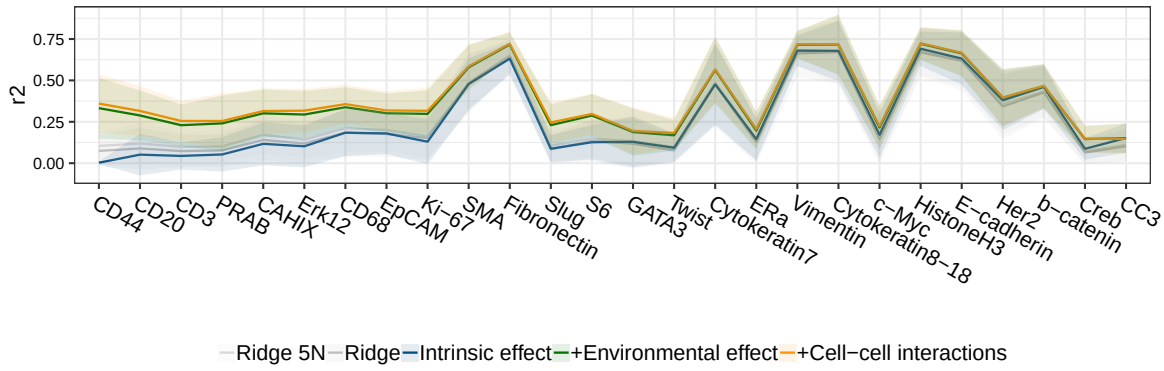


Fig. 3.15 Prediction accuracy for SVCA and alternative models using 5-fold cross-validation. The blue and green lines correspond to two reduced Gaussian Processes including respectively an intrinsic component, and an intrinsic component plus a local component. The two grey lines correspond to alternative linear regressions (see Section 3.4). The solid lines correspond to the coefficient of determination between the predicted and the observed values. Shaded areas correspond to plus and minus one standard deviation across images.

Validation with cell permutations

As a sanity check, we also computed the spatial variance signatures after permuting the cell positions and found that no cell-cell interactions were inferred when permuting the cells (Fig. 3.16).

The remaining variance component for the environmental effect can arguably be due to the capacity of the squared-exponential covariance function to capture uncorrelated noise for very small length scales. To validate this hypothesis, we plotted the variance estimate for the environmental effect as a function of the fitted length scale of the covariance function. We observed that strong environmental effects, in the case of permuted cell positions, corresponded indeed to very small length scales (Fig. 3.17).

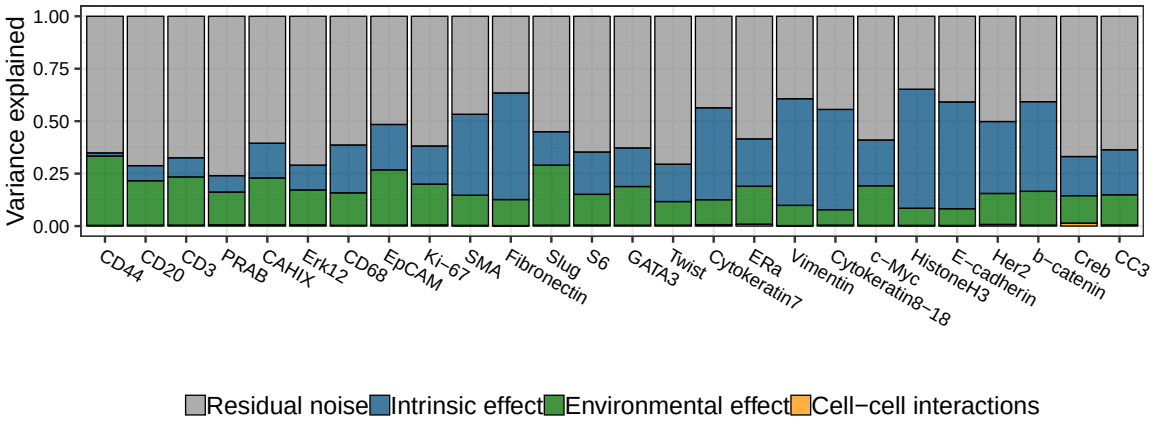


Fig. 3.16 SVCA signatures for permuted cell positions for the IMC Breast cancer data.

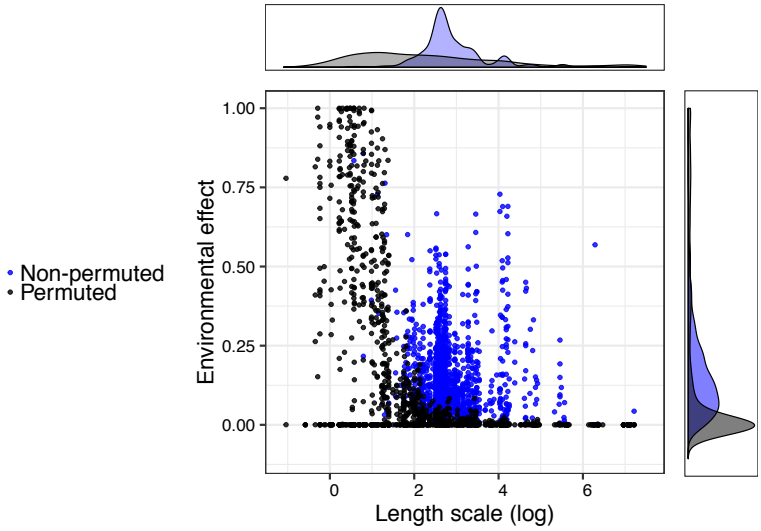


Fig. 3.17 Variance estimates for the environmental effect as a function of the fitted length scale, for both permuted and non-permuted cell positions. Shown are the log of the length scale on the x-axis, and the proportion of variance explained by the environmental effect on the y-axis. Each point corresponds to one image and one protein.

3.5.3 Biological interpretation

Effect of cell-cell interactions on expression variability

SVCA revealed substantial differences of the overall importance of cell-cell interaction components across proteins, explaining up to 25% of the total expression variance averaged across images. Immune cell markers in particular were found among the proteins with the largest cell-cell interaction component: for CD44, CD20, CD3 and CD68, we detected

significant cell-cell interaction effects in 34, 31, 32 and 31 out of the 46 images respectively (FDR<1%, Benjamini-Hochberg adjusted, Fig. 3a). We hypothesise that this effect could reflect the recruitment of immune cells by specific cellular environments (Moreau et al., 2018; "Chlon and Markowetz, 2017). CAHIX, a marker of hypoxia, was also found among the top markers linked to cell-cell interaction effects.

Signatures variability across images

We also observed substantial differences in the estimated spatial variance signatures across images (Fig. 3.18), motivating the investigation of the relationship between these variations and clinical covariates, including tumour grade.

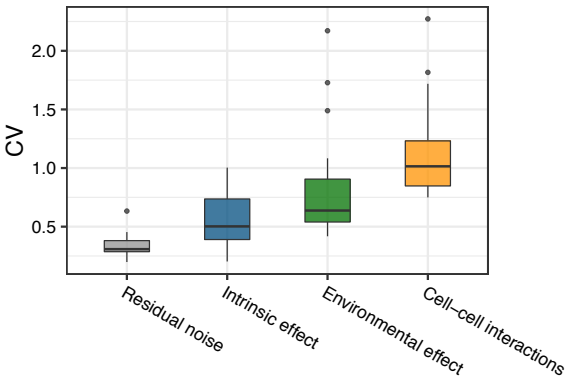


Fig. 3.18 Coefficient of variation (CV) across images of the SVCA variance components, computed independently for every protein. Boxplots are made using one point per protein.

A projection of the SVCA spatial variance signatures for every image using principal component analysis (PCA) identified tumour grade as an important source of variation in spatial variance components (Fig. 3.19) ($P = 3.8 \times 10^{-3}$, using the *ClusterSignificance* package in R (Serviss et al., 2017)). Inspection of the PCA loadings identified the cell-cell interaction component and the environmental component for a subset of proteins (including CD20 and CD44) as the most informative SVCA features related to tumour grade (Fig. 3.19).

Tumour progression is characterised by disorganisation and irregular cellular architecture which is associated with increased cell sizes, increased proliferation and thus higher cell density in comparison to healthy breast tissue (Elston and Ellis, 1991). We investigated how SVCA signatures were related to these environmental features and discovered a significant

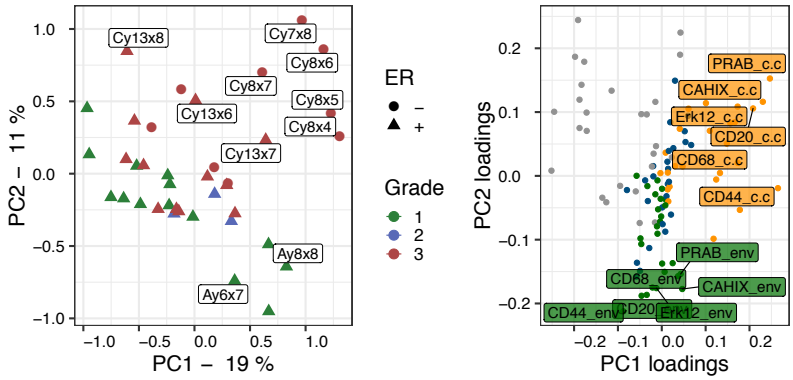


Fig. 3.19 Left: First two principal components for the 38 clinically annotated images, calculated based on the spatial variance signature, with individual images coloured by tumour grade. The shape corresponds to the ER status of the sample Right: Loadings of the principal components, depicting the relevance of individual proteins and types of variance components.

correlation ($P = 3 \times 10^{-3}$) between the average number of neighbours per cell and the average cell-cell interaction components across proteins (using cellProfiler (Carpenter et al., 2006) to compute the number of neighbours per cell and the *lm* function from R to compute significance). This difference in tissue composition could contribute to the separation of grades observed on the PCA. It is not surprising that cell-cell interactions would be higher in tissues with higher cell densities, compared to adipose tissue with only sparse cells. However, this effect was quite small (Fig 3.20), suggesting the existence of other drivers of variation across spatial variance signatures.

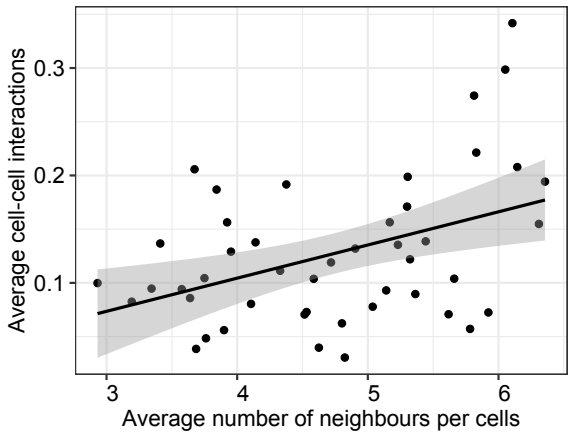


Fig. 3.20 Average cell-cell interaction component across proteins as function of the average number of neighbours per cells, as defined by cellProfiler (Carpenter et al., 2006). One point per IMC image. The line corresponds to the fitted linear regression model using the *lm* function in R ($P = 3 \times 10^{-3}$), and the shaded areas to 95% confidence intervals.

Relationship to histoCAT results

We noticed that the images with the strongest separation on the PCA (image names highlighted in Fig. 3.19) were already identified in the primary analysis (Schapiro et al., 2017) for the specificity of their tissue organisation compared to other images.

Schapiro et al. (2017) used the HistoCAT software to assess the significance of the spatial proximity of pre-annotated cell types, based on cell permutations. This analysis resulted in a signature of the cellular neighbourhood structure of every tissue, made of repulsion and attraction scores between each pair of cell types. Schapiro et al. (2017) clustered these signatures to detect images with similar tissue structures. As a result of this procedure, the highlighted images were separated in a grade 1 enriched cluster containing the images Ay6x7 and Ay8x8 and a grade 3 enriched cluster containing the images Cy7x8, Cy8x4, Cy8x5, Cy8x6, Cy8x7, Cy13x6, Cy13x7 and Cy13x8 (Schapiro et al., 2017, Fig. 3c). This suggests a relationship between SVCA signatures and signatures obtained with a permutation based neighbourhood analysis, although SVCA does not rely on cell type classification and arbitrary neighbourhood definitions.

3.6 Application of SVCA to a mouse hippocampus seqFISH data

SVCA can in principle be applied to data from any spatially resolved technology, including optical imaging-based assays. To explore this, we considered a mouse hippocampus RNA dataset imaged using the seqFISH technique (Shah et al., 2016). We start with a brief overview of the experimental method and the data processing steps. Then, we present the spatial variance signatures obtained with SVCA for this dataset, as well as validation steps using out of sample prediction and cell permutations. The section is concluded with a discussion on the biological relevance of the results.

3.6.1 Experimental method and data processing

Sequential FISH (seqFISH) is a targeted approach for in-situ RNA quantification, building upon single molecule Fluorescence In Situ Hybridisation (sm-FISH). In sm-FISH, cells are fixed and messenger RNAs are typically targeted in-situ by fluorescently labeled complementary sequences, whose positions are then detected with high resolution imaging (Raj et al.,

2008; AM et al., 2003).

In order to quantify a high number of transcripts in a single experiment, seqFISH uses a multiplexing approach that relies on the sequential application of multiple hybridisation and stripping rounds with fluorescent RNA-probes of multiple colours. A given gene is targeted by a unique combination of probes, each hybridising on a small portion of the RNA molecule. For F distinct colours and N hybridisation rounds, this barcoding approach can target up to F^N different transcripts. In contrast, a non-combinatorial approach where each gene is targeted by only one probe only allows the quantification up to $F \times N$ transcripts. The latter sequential approach is also used in seqFISH for some of the genes. (Shah et al., 2016; Lubeck et al., 2014; Shah et al., 2017).

To improve the method performance, seqFISH also borrows from information theory the use of error correcting barcodes (Shah et al., 2016; Biswas, 2008). Barcodes are chosen with a sufficient Hamming distance between them to ensure that the mis-detection of one or more of its element still allows for the non-ambiguous assignment to the right transcript. Increasing the Hamming distance between barcodes however reduces the number of possible target transcripts. Finally, seqFISH uses hybridisation chain reaction to amplify the brightness of every spot (Shah et al., 2016; Strell et al., 2018).

Raw data description

We analysed a dataset of 20 images from the hippocampus of a single mouse, with 249 genes quantified. A barcoding approach with 5 rounds of hybridisation and error correction was used for 214 of them. The remaining genes were imaged in 7 rounds of non-barcoding serial hybridisation. The authors also quantified the efficiency of the experiment which was found to be 71% (Shah et al., 2016).

Each image contained between 97 and 362 cells (average of 140), pre-segmented by the authors. The data consisted in a list of every single molecule detected within each cell, and its precise location, which was aggregated into single cell-counts. The location of the biopsies within the hippocampus spatial organisation were also provided by the authors.

Data processing

The single-cell counts were normalised following the same procedure as for the IMC data in Section 3.5.1, resulting in variance stabilisation of the expression profiles (Fig. 3.21) and removal of spurious correlations between genes (Fig. 3.22). As in the IMC application, the stabilised expression profile of the target gene Y was subsequently ranked standardised and transformed into Normally distributed data using the probit function, before fitting SVCA.

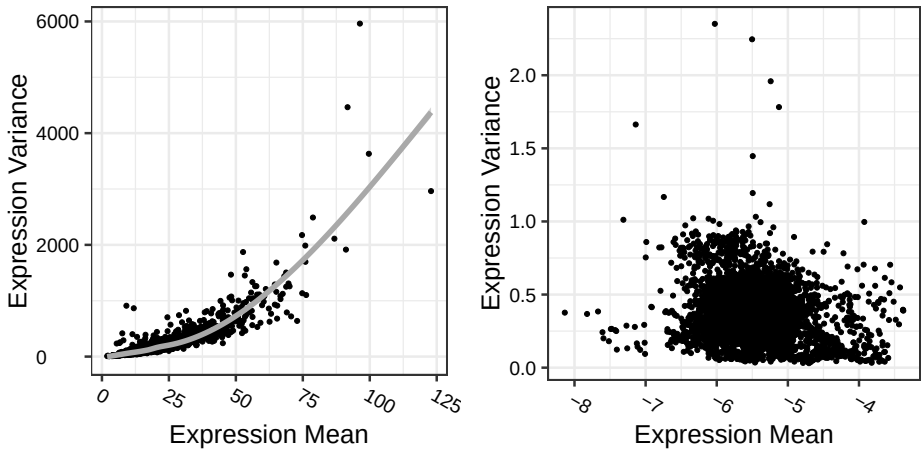


Fig. 3.21 Mean Variance relationship of single cell expression levels for the seqFISH data. Left: Before data processing. Right: after data processing. Means and variances are computed across cells, for every image and every protein independently (one dot per protein and image).

3.6.2 SVCA variance signatures

We computed the SVCA spatial variance signatures for every image independently. Figure 3.23 shows the average variance components across images. Because of the higher number of genes measured, we only represent the twenty genes with the highest cell-cell interaction component (Fig. 3.23).

Out of sample predictions

Again, we used 5-fold cross-validation to validate the variance estimates (Fig. 3.24), confirming at the same time that SVCA yields more accurate out of sample predictions than alternative models.

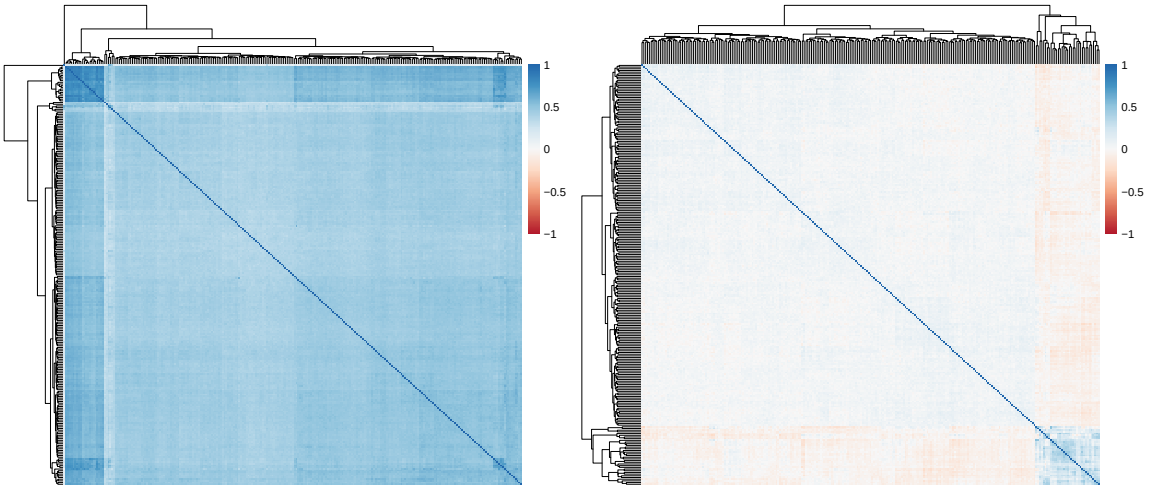


Fig. 3.22 Matrix of Pearson correlation between gene expression levels. Left: raw data. Right: processed data. Correlations between genes are computed within each image independently. Shown is the average value across images for each gene pair.

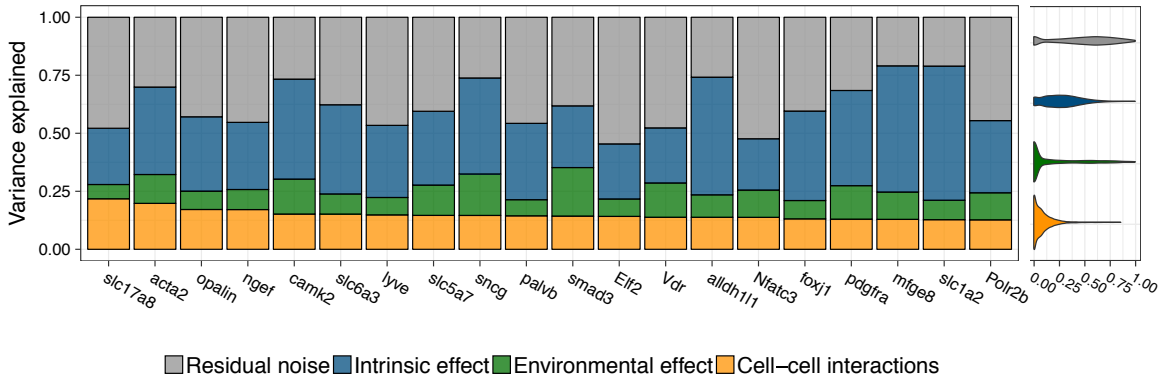


Fig. 3.23 SVCA Spatial Variance signatures averaged across images for the seqFISH hippocampus dataset. Genes are ordered by the magnitude of the cell-cell interaction component. Violin plot: variance estimates distribution across images and genes for all 249 genes.

Validation with cell permutations

We performed the same sanity check as in the IMC application, fitting the SVCA model on cell-permuted data (Fig. 3.25). Cell-cell interactions were greatly reduced by the procedure. However, unlike in the IMC application, there was a residual spurious cell-cell interaction effect, which could arguably be due to the smaller number of cells in each image, due to which, randomisation may not completely eliminate the spatial structure in the data. We investigated this issue further in Appendix A.6

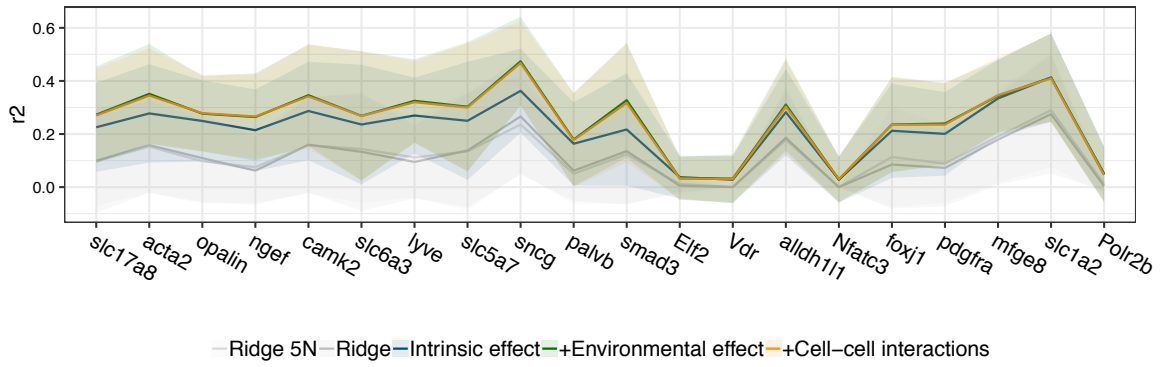


Fig. 3.24 Prediction accuracy for SVCA and alternative models using 5-fold cross-validation. The blue and green lines correspond to two reduced Gaussian Processes including respectively an intrinsic component only, and both an intrinsic and a local component. The two grey lines correspond to alternative linear regressions (see Section 3.4). Results are shown for the 20 genes with highest cell-cell interactions. The solid lines correspond to the coefficients of determination between predicted gene expression and observed values. The shaded areas correspond to plus and minus one standard deviation across images.

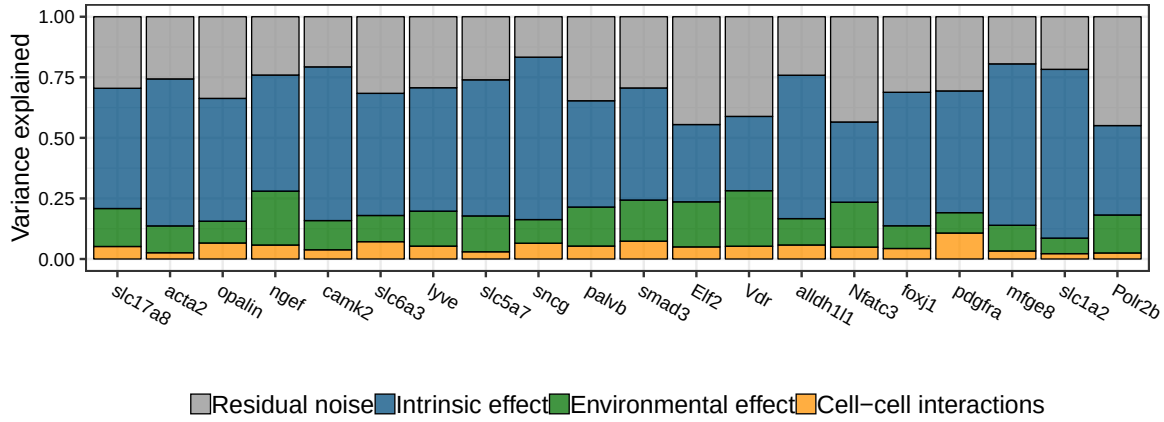


Fig. 3.25 SVCA signatures for permuted cell positions for the seqFISH hippocampus data for the genes of Figure 3.23.

3.6.3 Biological interpretation

Identification of relevant gene families involved in cell-cell interactions

Making use of the higher dimensionality of the data, we sought to identify gene families that participate in cell-cell interactions. First, we manually classified genes in non-overlapping categories based on prior annotations (see table in Appendix A.7), and considered categories with more than five genes, including cell cycle, cell junctions, immune system, neurotransmitter transporters and transcription factors. The neurotransmitter transporter category was

made of six glutamate transporters of the Solute Carrier family. The immune system category was made of six genes with various functions, all associated to immune response, such as MFGE8 involved in phagocytosis or the interferon regulatory factor IRF2. The eight cell junction genes included ACTA2 (Actin), Opalin (Yoshikawa et al., 2008) and MOG. The largest group was the transcription factors group with 166 genes.

We assessed the enrichment of these gene categories for the cell-cell interaction and the intrinsic components, using a permutation strategy similar to GSEA (Subramanian et al., 2005; Mootha et al., 2003):

i) Genes were ranked based on the size of the variance component of interest (cell-cell interaction or intrinsic effect)

ii) A GSEA-like trace was computed for each gene category and the height of this trace was considered as a test statistic.

iii) Gene labels were permuted 10,000 times in order to estimate an empirical P value for the statistic described above.

iv) P values were adjusted for multiple testing using a Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

We found that cell junction genes and neurotransmitter transporters were the most enriched groups for cell-cell interactions ($q = 6 \times 10^{-4}$ and $q = 1 \times 10^{-3}$, Benjamini-Hochberg adjusted) (Fig. 3.26).

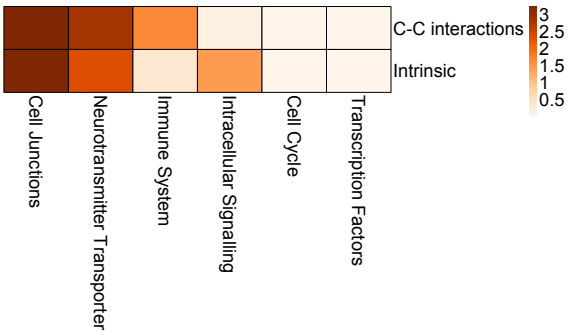


Fig. 3.26 Enrichment of the annotated gene families for cell-cell interaction and intrinsic effect, using a permutation based strategy. Values correspond to $-\log q$, Benjamini-Hochberg corrected.

Some cell junction genes such as GJA1 (connexin) are involved in gap junction intercellular communication (Cheng et al., 2015) while, for example, the actin skeleton has a known role in the adaptation of tissue structure and geometry to external stimulus (Carpenter, 2000; Brakebusch and Fässler, 2003), which may explain the enrichment of this category for cell-cell interactions. The enrichment of glutamate transporters is also consistent with their involvement in the transport and (re)uptake of neurotransmitter at the neuronal synapses, a critical cell-cell interaction in the brain (Masson et al., 1999; Iversen, 2009; Angulo et al., 2004; Mason, 2017). In addition Slc5a7 (CHT) was found to be preferentially expressed in specific interneurones which are linked to the spatial organisation of the tissue (Yi et al., 2015). To a smaller extent, genes related to the immune system were also significantly enriched for cell-cell interactions ($q = 2 \times 10^{-2}$). Genes such as CTSS (Cathepsin) and MFGE8 (Lactadherin), which play a role in phagocytosis in the brain (Fricker et al., 2012; Neher et al., 2013; Vitner et al., 2010), were amongst the top cell-cell interaction related genes. Notably however, Cell junction genes and Neurotransmitter transporters were also enriched amongst genes with a high intrinsic component, suggesting that the expression level of these genes also relate to intracellular processes (Fig. 3.26). This observation also raised the question of whether cell-cell interaction and intrinsic components were globally correlated. In Appendix A, Section A.3, we compared the cell-cell interaction variance component with the three other model components for both the seqFISH data application and the IMC application of Section 3.5. We did not observe any strong global dependency between intrinsic effect and cell-cell interactions.

Five out of the ten genes with the highest cell-cell interaction variance components did not fall into any of the annotated gene sets and we therefore analysed them individually. NGEF (Ephexin) is an exchange factor that plays a role in axon guidance (Shamah et al., 2001; O'Donnell et al., 2009); CAMK2 is a kinase known to play a role in long-term potentiation and neurotransmitter release (Wang, 2008; Lisman et al., 2012); LYVE is a membrane receptor (Banerji et al., 1999) and SNCG (Synuclein Gamma) is involved in axonal architecture (Surguchov et al., 2001; Vargas et al., 2017). Taken together, this shows that genes with a high cell-cell interaction component are reported in the literature as being involved in important cell-cell communications in neurones, or having a function in the spatial architecture of the tissue.

Signature variability across samples

Similarly to results obtained on the IMC datasets, we observed high variation in the spatial variance signatures across images, which were sampled from functionally distinct regions of the hippocampus (Shah et al., 2017) (Fig. 3.27).

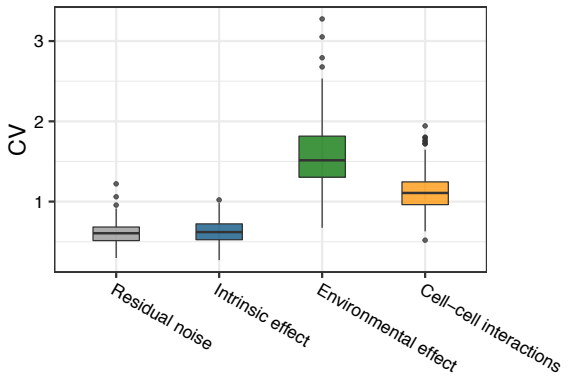


Fig. 3.27 Coefficient of variation across images of the SVCA variance components, computed independently for every gene (one point per gene).

We used principal component analysis, to see if this variability could be linked to prior knowledge about the structure of the hippocampus. We used the compartmentations of the hippocampus into dorsal, ventral and intermediate regions, as well as CA1, CA3 and DG regions as provided by Shah et al. (2017) (Fig. 3.28). We found that the first two principal components of the spatial variance signatures in the dorsal region clustered together, irrespective of their CA1/CA3 location. Similarly, images from the Dentate Gyrus (DG) also clustered together, and there was some proximity between signatures from the ventral region, although with more variation between them. This is consistent with the observation from Shah et al. (2017) that the ventral and dorsal regions of the CA1 and the CA3 mirror each other with respect to their cellular compositions, and that ventral regions are more heterogeneous in their cellular composition. Spatial variance signatures for intermediate regions, however, did not show much resemblance.

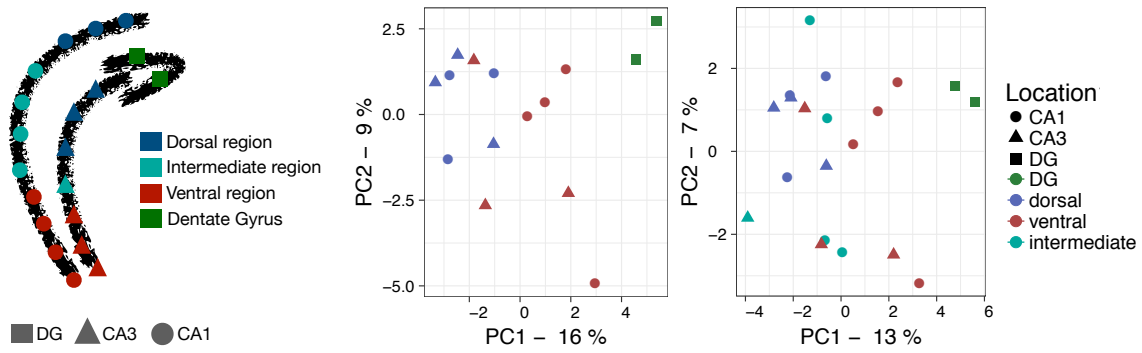


Fig. 3.28 Left: spatial organisation of the mouse hippocampus with each dot corresponding to an individual image. Colours and shapes denote regions using the classification as in Shah et al. (2017). Middle: first two principal components of the spatial variance signatures for individual images from the DG, the dorsal region and the ventral region. Colour and shape represent the location of the biopsy in the hippocampus. Right: first two principal components of the spatial variance signatures for all 20 images.

3.7 Discussion

We presented Spatial Variance Component Analysis (SVCA), a regression-based framework for the analysis of spatially resolved molecular expression data. Our model uses a Gaussian Process with specific covariance terms to compute a spatial variance signature for individual mRNA or protein levels, which decomposes their sources of variation into spatial and non-spatial components. Most prominently, SVCA provides a quantitative assessment of the effect of cell-cell interactions on the expression profile of individual molecules. The model avoids the explicit definition of cell types and neighbourhoods, instead using a continuous measure of cell state and Euclidean distances between cells.

3.7.1 Technical limitations

Although we have tested the calibration and robustness of SVCA, the model is not free of limitations.

Noise model

At present, the model does not account for technology-specific noise and instead assumes Gaussian distributed residuals, requiring suitable data processing. Further development could consider a generalised random effects model, for example to couple the random effect component with a negative-binomial likelihood.

Univariate model

A second limitation of SVCA is that the model is univariate, meaning that individual genes or proteins are modelled independently from each other. Multivariate extensions could account for relationships between genes involved in the same pathways, either in an unsupervised manner or using prior knowledge (Buettner et al., 2017). Such approaches could give a more comprehensive understanding of how biological processes are affected by tissue structure.

Scalability

As the size of spatial expression datasets increases with the development of higher-throughput technologies, scalability will also become an important challenge for SVCA. The model is linear in the number of genes, and massive parallelisation can be obtained with adequate computational infrastructure. Also, the Gaussian Process approach typically scales cubically in the number of cells, which can be circumvented by splitting bigger images into multiple patches analysed independently. Additionally, we could explore using faster inference methods relying on sparse approximations (Hensman et al., 2013a; Snelson and Ghahramani, 2006; Quiñonero-Candela and Rasmussen, 2005) or random feature selection (Rahimi and Recht, 2008; Oliva et al., 2016).

Significance testing

We have seen in Section 3.3.4 that the SVCA significance testing procedure is quite underpowered. For the applications presented in this thesis, we believe that the output of interest was the variance signatures rather than any significance measurement. However, if SVCA is to be used for the purpose of assessing the significance of the covariance terms, a more calibrated testing procedure needs to be implemented.

Miscellaneous

In addition to those main points, there are multiple minor issues and open discussion topics for the SVCA model and software. Among other things, the model could use decoupled length scales for the cell-cell interaction and the environmental covariance functions. Different covariance functions could also be designed and compared to the current choices. Finally, the current software uses command line executions, meaning that user experience could be improved by the development of a graphical interface and SVCA could be integrated in existing platforms such as the histoCAT software (Schapiro et al., 2017).

3.7.2 Biological applications

Broad applicability of SVCA

We have applied SVCA to datasets generated using alternative technologies, probing either RNA transcripts or proteins, demonstrating the broad applicability of the approach. Across these applications, we observed that cell-cell interactions can substantially contribute to gene expression variation, which is consistent with previous reports (Battich et al., 2013; Goltsev et al., 2018; Kamińska et al., 2015; Nasra Naeim Ayuob and Soad Shaker Ali, 2012) and supports the concept that studying single cell expression in the native context is important for understanding their molecular differences.

We noticed a relatively high variability of the SVCA signatures across samples and investigated the possible causes of this variability. We provided evidence that differences in SVCA signatures could result from differences in the spatial structure of tissues, as well as different clinical and biological contexts. For the IMC data, we also noticed that this variability reflected previous findings about different tissue organisations between samples.

Interpretation of the variance signatures remains challenging

We used gene annotation and enrichments to interpret the spatial variance signatures. In the seqFISH application in particular, we found genes known to be involved in cellular interactions, such as SLCs, to be predominantly enriched in the corresponding term of our model.

Further interpretation of these signatures remains however challenging. This could arguably be due to our limited knowledge of such multi-cellular processes in comparison to intracellular pathways. In addition, cell-cell interactions may be caused by a diversity of biological contexts and processes: for example, it is intrinsically challenging to tell apart cell-type co-occurrences from more specific molecular interactions. As emerging technologies provide data sets rich in this type of information, and methods such as SVCA are developed to analyse it, our knowledge and understanding should increase, and thereby the ability to interpret the signatures of cell-cell interactions. In particular, more hypothesis-driven research, possibly with simpler biological systems with clear positive and negative controls, can be instrumental towards this goal.

3.7.3 Conclusion

There is a growing appreciation of the role of spatial distribution of proteins, RNA transcripts and other molecules in determining tissues functioning and its deregulation in disease, with potential value as predictors of clinical outcomes (Bodenmiller, 2016). This is largely driven by vigorous development of novel technologies that enable us to capture such data (Bodenmiller, 2016; Aichler and Walch, 2015; Lin et al., 2017; Schulz et al., 2018). We believe that the SVCA framework and extensions thereof could be of broad use to analyse this burgeoning spatially resolved molecular data to advance our understanding of the pathophysiology of multiple diseases.

Chapter 4

Biofam: a flexible framework for Factor Analysis models in biology

In this Chapter, we present biofam, a flexible Factor Analysis framework which can be applied to gene expression data and other biological layers, while accounting for structured data context such as the existence of multiple sample groups or the combined analysis of multiple omics.

The development of biofam was motivated by the prior implementation of the MOFA package (Argelaguet et al., 2018) a more limited factor analysis model for the integration of multiple data types (see Section 5.2). Biofam provides a more efficient inference scheme, additional sparsity-inducing priors allowing the user to address new biological questions and is based on a more modular implementation enabling the selection of optional model features in any combination.

This is joint work with Oliver Stegle, Ricard Argelaguet, Danila Bredikin and Yonatan Deloro. I led the development of the package. I designed the model and software in collaboration with Ricard Argelaguet. Ricard Argelaguet and I derived all variational updates and implemented most of the core software, which includes the probabilistic model, the inference routine and the python user interface. Danila Bredikin and Yonatan Deloro joined the project later and contributed to the implementation of some core software components and implemented most of the R package for downstream analysis presented in Section 5.1, based on previous work from Britta Velten and Ricard Argelaguet. I designed and implemented the stochastic inference extension, and tested it with the help of Yonatan Deloro, an intern I supervised.

All analysis presented in this chapter is the result of my own work, except the simulations from the non-Gaussian data which were performed by Ricard Argelaguet (Fig. 4.14 and 4.15), In Chapter 5, we illustrate use cases of biofam on real data by others and in collaboration with me. The software is open source and freely available here: https://github.com/biofam/biofam.

4.1 Introduction

Unsupervised models such as clustering or dimensionality reduction are widely used approaches for the exploratory first-pass analysis of high-dimensional data. For example, Principal Component Analysis (Hotelling, 1933) (see Section 2.2) computes a low dimensional representation of the data which can be easily visualised and highlights the main dependencies between samples, such as subpopulations (Novembre et al., 2008; Pollen et al., 2014; Islam et al., 2011; Shalek et al., 2014; Tang et al., 2010), technical batch effects (Luo et al., 2010; Holmes et al., 2011; Yang et al., 2008; Lazar et al., 2013) or continuous relationships (Wang et al., 2013; Haghverdi et al., 2015; Guo et al., 2010).

With recent advances in experimental techniques, gene expression datasets are measured in an increasing number of different contexts, including multiple tissues (GTEx Consortium, 2013; Nica et al., 2011), and in combination with other biological layers such as methylation (Hasin et al., 2017). This gives rise to structured datasets, where samples belong to distinct groups, and features to distinct views, with sometimes different data modalities. These new datasets pose additional requirements to perform dimensionality reduction while explicitly accounting for this contextual information.

We will here build on the probabilistic Factor Analysis framework introduced in Chapter 2. This formulation allows for the definition of hierarchical priors which reflect prior knowledge about the data context and structure, or prior assumptions about the distribution of the model parameters. In addition, Factor Analysis has other advantages such as the ability to handle datasets with missing values.

Probabilistic methods for linear dimensionality reduction

Multiple variants of Factor Analysis have been developed and applied to the analysis of gene expression datasets. They differ in the choice of the prior distributions used on the model parameters, which are tailored to a given context or application. For example, while most

models typically assume Normally distributed weights and factors (see Section 2.2), Independent Component Analysis assumes a non-Normal distribution of the weights (Lawrence and Bishop, 2000), which induces statistical independence across latent factors.

In addition, Factor Analysis extensions differ in their choice of hierarchical priors. Group Factor Analysis (Virtanen et al., 2012; Klami et al., 2015) models prior knowledge about existing groups of features, which allows for the combined analysis of data from multiple sources. As conventional (Group) Factor Analysis tends to infer dense weight matrices, so that the latent factors capture a maximum variability in the data (minimising the reconstruction error), relating these latent factors to known biological processes involving a small subset of genes is challenging. To address this issue, sparse extensions of Factor Analysis (Zhao et al., 2016; Khan et al., 2014; Gao et al., 2013) model the assumption that meaningful drivers of variation affect a limited number of features, resulting in the inference of sparse weight matrices which are more easily interpretable as biological processes. Bi-clustering models extend this sparsity assumption to the factors (Hochreiter et al., 2010; Suvitaival et al., 2014; Bunte et al., 2016; Leppäaho et al., 2017)

While Group Factor Analysis models are useful for the purpose of integrating multiple types of data sources for matching samples, existing models do not account for group structures on the sample axis. Thus, they do not offer a principled framework for the unsupervised analysis of multidimensional data across biological contexts, a problem which remains less studied. A notable exception is the GFA mixture model of Remes et al. (2015), which however offers only limited options and does not provide a runnable software (see Section 4.7.1).

In addition, current GFA implementations each have a number of specific limitations. Commonly employed inference using Gibbs sampling does not scale to large datasets (Khan et al., 2014; Bunte et al., 2016; Leppäaho et al., 2017). Other implementations do not handle missing values (Khan et al., 2014; Zhao et al., 2016; Remes et al., 2015; Klami et al., 2015; Virtanen et al., 2012). These models rely solely on a Gaussian likelihood, which is not adapted for the analysis of data with different modalities such as binary mutation or methylation data. Finally, most of the tools available for dimensionality reduction provide only sparse functionalities for downstream analysis of the latent factors and weights inferred. At the end of this Chapter (Section 4.7.1), Table 4.3 gives a summary of the characteristics of published Group Factor Analysis models.

Biofam: a unified framework for Factor Analysis in biology

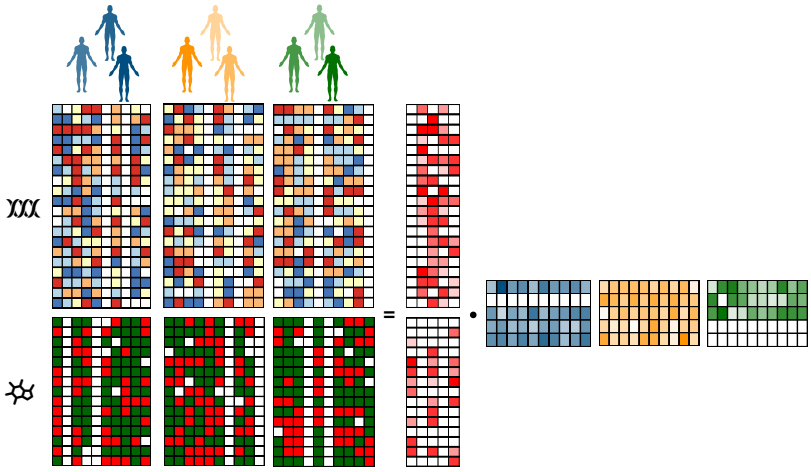


Fig. 4.1 Biofam enables joint Factor Analysis across multiple omics and multiple sample groups. Weight matrices are represented in red, and factor matrices have the colour of the sample groups they correspond to. Latent factors may be relevant to specific omics and specific sample groups only, as illustrated by the null column in the weight matrices and the null rows in the factor matrices (white cells). Biofam also supports datasets with missing values, including samples missing an entire assay, as represented by empty columns in the data matrix (and likewise, features missing across an entire sample group).

Here, we propose biofam (Bio - Factor Analysis models), a unified Factor Analysis modelling framework which brings together the individual strengths of the multiple separate GFA models and implementations mentioned before. Biofam is implemented in a modular manner, so that grouping structures and sparsity-inducing priors can be encoded flexibly on the feature axis, on the sample axis or on both. Thus, it is able to fit existing FA, GFA, ICA and bi-clustering models, with or without sparsity-inducing priors.

In addition to unifying existing GFA modelling approaches, biofam offers a number of unique features. Firstly, it models grouping structures on the sample axis, facilitating the analysis of new types of gene expression analysis problems such as the combined analysis of datasets across different biological contexts (Fig. 4.1). Biofam also extends GFA to support Poisson and Bernoulli data likelihoods, which are particularly suited to the analysis of multi-omics data where different data sources exhibit different modalities, such as binary data in the case of somatic mutations. Biofam is implemented using an efficient inference scheme using variational methods and GPU optimisation, which makes it scalable to large datasets. We also present ongoing work regarding an extension to stochastic variational inference, which holds

the promise to make biofam scalable to datasets which do not fit on the computer memory (see Section 4.6), although this still requires the implementation of additional functionalities. Finally, biofam provides an R package for results visualisation and downstream analysis (Fig. 4.2), which will be presented in Chapter 5.

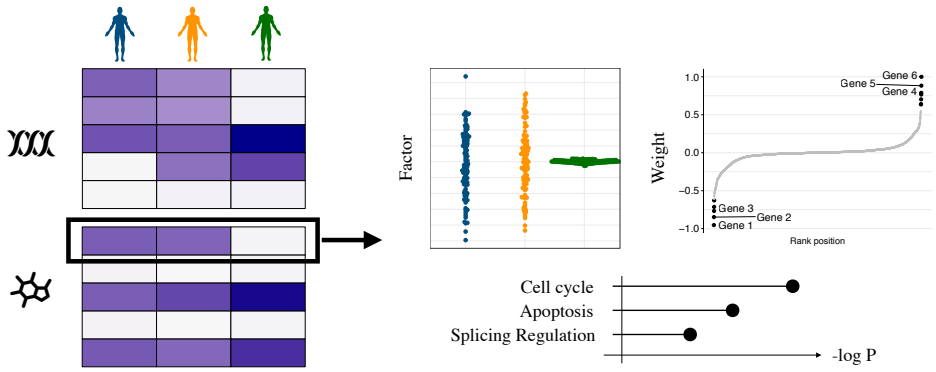


Fig. 4.2 Biofam downstream analysis packages provides a compact representation of factors relevance in multiple sample groups and omics. It also provides a visual way to inspect the weights of the Factor Analysis model and relate them to biological processes using Gene Set Enrichment Analysis

4.2 Model

Biofam builds upon the modelling principles and techniques introduced in Section 2.2, thereby extending Factor Analysis using hierarchical Bayesian priors, whose architecture reflects the data context. After introducing our notation and naming convention, this Section presents the Bayesian architecture and main characteristics of the biofam model. Unless stated otherwise, graphical models use the notations of Section 2.2

4.2.1 Mathematical notation and naming convention

In general, we consider structured datasets where multiple sample groups are analysed jointly, and the features consist of multiple distinct groups or views. These datasets are described with the following notation:

G	Number of sample groups
M	Number of feature groups or *views*
N_g	Number of samples in group $g \in [\![1;G]\!]$
D_m	Number of features in view $m \in [\![1;M]\!]$

$y_{n,d}^{m,g}$ Data point for sample n of group g and feature d of view m
Y Full data matrix concatenated over samples, features, groups and views

Note that $y_{n,d}^{m,g}$ does not designate a tensor, as views have in general different number of non-matching features, and groups different numbers of non-matching samples. The model however relies on the assumption that samples are matching across views, and features are matching across sample groups (Fig. 4.1).

Biofam builds upon Factor Analysis, which assumes that the observed dataset is generated by a (small) number of latent factors, through a linear function (Section 2.2.4). In this chapter, the model parameters are described with the following notation:

K Number of latent factors
$z_{n,k}^{g}$ Value of factor k associated to sample n of group g
$w_{k,d}^{m}$ Weight associated to feature d of view m and factor k
τ_{d}^{m} Precision of the Gaussian likelihood for feature d of view m
Z Matrix of all factors concatenated over samples and groups
W Matrix of all weights concatenated over features and views
Θ Ensemble of all model parameters, including W, Z, $\{\tau_{d}^{m}\}_{\forall m,\forall d}$ and any additional optional parameters introduced in this Section.

Unless stated otherwise, we assume a Gaussian likelihood model with feature and view specific precision; using the notation just introduced: $y_{n,d}^{m,g} \sim \mathcal{N}\left(\sum_{k=1}^{K} z_{n,k}^{g} w_{k,d}^{m}, \tau_{d}^{m}\right)$. In the interest of uncluttered notations, view and group indices m and g are typically dropped when the dataset consists in only one view or group, or when these grouping structures are not needed for clarity.

This section introduces optional hierarchical priors which encode prior knowledge about the data structure and prior assumptions about the latent space. The mathematical notation for the corresponding parameters are introduced in due time.

4.2.2 Structured Sparsity

Known feature groups

When the data is structured into multiple feature groups, called views, such as multiple omics, some factors may affect specific views only, resulting in the structured sparsity of the

weight matrix W illustrated in Figure 4.1 and in Figure 2.12 of Section 2.2.5. Borrowing prior distributions from Group Factor Analysis (Section 2.2.5), biofam models this structured sparsity assumption with view-specific ARD priors on the weights:

$$\begin{aligned} w^m_{k,d} &\sim \mathcal{N}(0, \alpha^m_k) \\ \alpha^m_k &\sim \Gamma(a_0, b_0) \end{aligned} \tag{4.1}$$

As in Virtanen et al. (2012), we use the hyperparameters $a_0 = b_0 = 10^{-14}$.

Known sample groups

Similarly, when the data is structured into multiple sample groups reflecting different biological contexts, some factors may be relevant to a subset of groups only, resulting in the structured sparsity of the factor matrix Z, illustrated in Figure 4.1. Structured sparsity of the factors is also modelled with group specific ARD priors, with the same hyperparameters $a_0 = b_0 = 10^{-14}$:

$$\begin{aligned} z^g_{k,d} &\sim \mathcal{N}(0, \alpha^g_k) \\ \alpha^g_k &\sim \Gamma(a_0, b_0) \end{aligned} \tag{4.2}$$

4.2.3 Element-wise sparsity

Weight sparsity

Individual biological processes typically only affect a small fraction of observed features (Gao et al., 2013). However, commonly used models, including PCA and conventional Factor Analysis, do not exploit this sparsity assumption and instead assume that the factors are in general dense. As a result, interpreting the inferred factors in relation to biological processes can be challenging.

In contrast, we here use a spike-and-slab sparsity-inducing prior (Mitchell and Beauchamp, 1988) on the weights of the Factor Analysis model, corresponding to a zero-inflated Normal distribution. Specifically, weights are modelled as drawn from a product between a Normally distributed random variable $\hat{w}$ and a Bernoulli distributed variable s:

$$
\begin{aligned}
w^m_{k,d} &= \hat{w}^m_{k,d} \times s^m_{k,d} \\
\hat{w}^m_{k,d} &\sim \mathcal{N}(0, \alpha^m_k) \\
s^m_{k,d} &\sim \text{Ber}(\theta^m_k) \\
\theta^m_k &\sim \beta(a_0, b_0)
\end{aligned}
\tag{4.3}
$$

The parameter θ^m_k of the Bernoulli variables $s^m_{k,d}$ corresponds to the fraction of features affected by factor k in view m. We model θ^m_k as a random parameter with a conjugate beta prior with hyperparameters $a_0 = b_0 = 1$, corresponding to a uniform distribution.

Factor sparsity

In biofam, we extend element-wise sparsity to the factor distributions using an analogous model, and the same hyperparameters. We use the same notations for the Bernoulli variables s and their beta-distributed parameter θ, but the parameters can be distinguished by the indices used (m for indices of the views and g for the indices of the sample groups):

$$
\begin{aligned}
z^g_{k,n} &= \hat{w}^g_{k,n} \times s^g_{k,n} \\
\hat{w}^g_{k,n} &\sim \mathcal{N}(0, \alpha^g_k) \\
s^g_{k,n} &\sim \text{Ber}(\theta^g_k) \\
\theta^g_k &\sim \beta(a_0, b_0)
\end{aligned}
\tag{4.4}
$$

Rotational invariance and model identifiability

Conventional Factor Analysis is invariant to rotation in the latent space. To demonstrate this property, let us consider a rotation matrix R of dimension $K \times K$ and note $\tilde{W} = RW$, the model weights rotated by R, and $\tilde{Z} = ZR^{-1}$ the rotated factors. The model likelihood is unchanged by this rotation, $P(Y|\tilde{Z}\tilde{W}, \tau) = P(Y|ZR^{-1}RW, \tau) = P(Y|ZW, \tau)$, and isotropic Normal priors on the weights and factors are also unaffected. For the weights for example, $\ln P(\tilde{W}) \propto \sum_{k,d} \tilde{w}^2_{k,d} = \text{Tr}(\tilde{W}^T\tilde{W})$, and $\tilde{W}^T\tilde{W} = W^T R^{-1} RW = W^T W$, using the fact that the inverse of a rotation matrix is its transposed. This property obviously renders conventional Factor Analysis unidentifiable.

Sparsity assumptions, however, partially address the rotational invariance problem. Independent identically distributed spike-and-slab priors on the weights or factors of the model is not rotationally invariant and will encourage sparser solutions. When the true generative

factors are dense however, rotational invariance remains to some extent an issue, which we will discuss in Section 4.4.2. Additionally, the model remains evidently invariant to factor permutations, which should be remembered when comparing the results given by two independent fittings of a Factor Analysis model.

4.2.4 Multiple data likelihoods

Standard Factor Analysis typically assumes Normally distributed data. In biofam, we extend this model to binary and count data, using respectively the Poisson likelihood of Equation 4.5 and the Bernoulli likelihood of Equation 4.6.

Bernoulli Likelihood for binary data

$$\begin{aligned} P\left(y_{n,d}^{m,g}\middle|\Theta\right) &= \text{Ber}\left(y_{n,d}^{m,g}\middle|\sigma\left(\sum_k z_{n,k}w_{k,d}\right)\right) \\ &= \frac{1}{1+\exp\left[-\left(2y_{n,d}^{m,g}-1\right)\sum_k z_{n,k}w_{k,d}\right]} \end{aligned} \tag{4.5}$$

Poisson likelihood for count data

$$\begin{aligned} &P\left(y_{n,d}^{m,g}\middle|\Theta\right) \propto \lambda\left(\sum_k z_{n,k}^g w_{k,d}^m\right)^{y_{n,d}^{m,g}} \exp\left(-\lambda\left(\sum_k z_{n,k}^g w_{k,d}^m\right)\right) \\ &\text{with } \lambda\left(\sum_k z_{n,k}w_{k,d}\right) = \ln\left[1+\exp\left(\sum_k z_{n,k}^g w_{k,d}^m\right)\right] \end{aligned} \tag{4.6}$$

In Section 4.3.2 and Appendix C, variational updates for these data likelihoods are defined.

4.2.5 Handling missing values

Nothing in the Factor Analysis formalism requires the completeness of the data matrices, meaning that these models naturally handle missing values. In a vectorised implementation however, care needs to be taken, so that we keep track of the indices of non-observed data in matrix operations, and remove their contribution to the variational updates and Evidence Lower Bound terms. Biofam encodes the position of the missing values in memory efficient Boolean masks, which are built based on the presence of invalid values such as *NA* in the input matrices, and propagates this information to all update operations.

4.2.6 Modular implementation

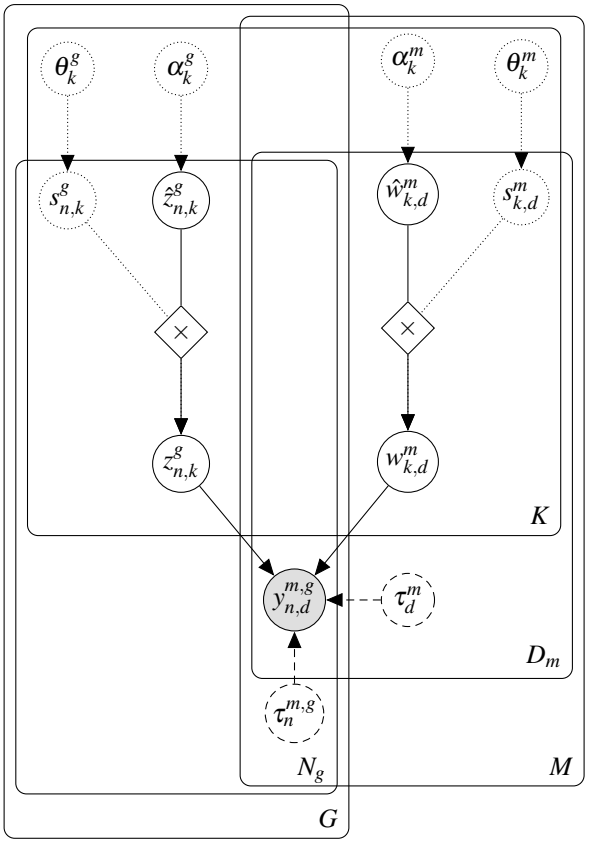


Fig. 4.3 Full graphical model for biofam with modularity representation. Full nodes correspond to the core Factor Analysis model implemented in biofam. Dotted nodes correspond to optional parameters which can be used in any combination. Dashed nodes correspond to the two mutually exclusive noise parameters to choose from for the Gaussian likelihood.

Biofam is implemented in a modular fashion which enables the user to choose any combination of sparsity-inducing priors that suits their assumptions, enabling data integration across feature groups (such as multi-omics), sample groups (such as multiple tissues or experimental conditions) or both in a flexible manner. This modularity also allows for using different likelihoods for multiple groups of features, which is particularly useful for applications to multi-omics data. Finally, for the sake of model symmetry, we allow the Gamma distributed precision τ to be defined on a per sample basis. This full modular model is shown in Figure 4.3.

The design of the software allows combining assumptions and functionalities from models that have been considered individually, such as sparse and non-sparse Group Factor Analysis,

or an implementation of Independent Component Analysis, using spike-and-slab priors on the factors. Comparison between different models can therefore be performed with the same software, aiding the objective comparison of alternative models and assumptions. In Section 4.4, we make use of this flexibility in order to investigate the effect of different choices of priors.

4.3 Inference

The ultimate goal of Bayesian inference is to compute the posterior distribution of all parameters given the observed data using the Bayes rule (Eq. 2.23, Section 2.2). However, as seen in Section 2.2, this computation is generally intractable and the inference must be addressed by approximate methods. Biofam uses variational inference (Section 2.2.6) to approximate the posterior distribution of the latent parameters.

In this section, we outline the variational inference scheme of biofam.

4.3.1 Posterior factorisation

The mean field approximation is widely used in variational inference to approximate the posterior distribution of the parameters using a fully factorised distribution $q(\Theta) = \prod_i q_i(\theta_i)$. Although this assumption is common and most convenient (see Section 2.2.6), it may be advantageous to retain dependencies for selected subsets of model parameters.

Specifically, in the case of the spike-and-slab prior, the fully factorised mean-field approximation has limitations because of the strong connection between the Normally distributed parameter $\hat{w}^m_{k,d}$ (e.g. for the model weights) and the Bernoulli parameter $s^m_{k,d}$. For example, notice that if $s^m_{k,d} = 0$, $\hat{w}^m_{k,d}$ becomes unconnected to the data. Therefore $P(\hat{w}^m_{k,d}|s = 0)$ should be equal to the prior $P(\hat{w}^m_{k,d})$ and only $P(\hat{w}^m_{k,d}|s = 1)$ should be influenced by the data. Titsias and Lázaro-Gredilla (2011) show that using a joint q distribution for the parameters $\hat{w}^m_{k,d}$ and $s^m_{k,d}$ such that $q(\hat{w}^m_{k,d}, s^m_{k,d}) = q(\hat{w}^m_{k,d}|s^m_{k,d})q(s^m_{k,d})$ yields a more accurate approximation to the posterior distribution of these parameters than the fully factorised approximation used in alternative studies (Yoshida and West, 2010).

In biofam, we therefore use the partially factorised approximation of Equation 4.7 when using spike-and-slab priors on the model weights and factors.

$$
\begin{aligned}
q(\Theta) = &\prod_{k,m,d} q(\hat{w}^m_{d,k}|s^m_{d,k})q(s^m_{d,k}) \prod_{n,k,g} q(\hat{z}^g_{k,n}|s^g_{k,n})q(s^g_{k,n}) \\
&\prod_{m,k} q(\alpha^m_k)q(\theta^m_k) \prod_{g,k} q(\alpha^g_k)q(\theta^g_k) \prod_{d,m} q(\tau^m_d)
\end{aligned}
\tag{4.7}
$$

In Appendix B, we derive an analytical inference scheme for this partial factorisation. The update rule of Section 2.2.6, $\ln q_i(\theta_i) = \mathbb{E}_{\theta_{j\neq i}} \ln(P(Y,\Theta)) + cst$ remains unchanged for all parameters of the model except $s^m_{k,d}$ and $\hat{w}^m_{k,d}$, for which it becomes:

$$
\begin{aligned}
\ln q(\hat{w}^m_{k,d}|s^m_{k,d}) &= \mathop{\mathbb{E}}_{\Theta'|s^m_{k,d}} \ln P(Y,\Theta) + cst \\
\ln q(s^m_{k,d}) &= \mathop{\mathbb{E}}_{\Theta',\hat{w}^m_{k,d}|s^m_{k,d}} \left[\ln \frac{P(Y,\Theta)}{q(\hat{w}^m_{k,d}|s^m_{k,d})} \right] + cst
\end{aligned}
\tag{4.8}
$$

where Θ' includes all model parameters except $\hat{w}^m_{k,d}$ and $s^m_{k,d}$. The same update rules can be derived for spike-and-slab priors on the factors.

As biofam uses only conjugate priors, the derivation of the update equations for each term of the model is straightforward and the approximate posterior distributions q belong to the same family as the prior distributions. Update rules for all model parameters are provided in Appendix D.

4.3.2 Non-Gaussian likelihoods

Prior conjugacy does not hold for the non-Gaussian likelihoods of Section 4.2.4, which renders the derivation of the variational updates more challenging. To address this, we adapt prior work from Seeger and Bouchard (2012) and Jaakkola and Jordan (2000), who derive Gaussian lower bounds to the Poisson and Bernoulli likelihoods. This section outlines the general principle of these approximations and provides the resulting update rules. Detailed derivations can be found in Appendix C.

General principle

Recall that variational inference can be regarded as an optimisation problem where the aim is to find the distribution $q(\Theta)$ that maximises $\mathscr{L} = \mathbb{E}_{q(\Theta)} \ln P(Y|\Theta) - \mathrm{KL}(q(\Theta)||p(\Theta))$, where $\mathscr{L} < P(Y)$ is an evidence lower bound. In the case of an independent data likelihood, $\ln P(Y|\Theta)$ can be rewritten as a sum over samples and features $\ln P(Y|\Theta) = \sum_{n,d} \ln P(y_{n,d}|\Theta)$.

When the model likelihood is Gaussian, each term $\ln P(y_{n,d}|\Theta)$ is quadratic in $\sum_{k=1}^{K} z_{n,k}w_{k,d}$.

For non-Gaussian likelihoods, Seeger and Bouchard (2012) and Jaakkola and Jordan (2000) use Taylor expansions to derive a quadratic lower bound to each log likelihood term such that $\ln P(y_{n,d}|\Theta) > q_{n,d}(y_{n,d},\Theta)$, with $q_{n,d}(y_{n,d},\Theta)$ quadratic in $\sum_{k=1}^{K} z_{n,k}w_{k,d}$. The quadratic property allows rewriting this function as a Gaussian log likelihood: $q_{n,d}(y_{n,d},\Theta) = \ln \tilde{P}(\tilde{y}_{n,d}|\Theta)$ where $\tilde{P}(\tilde{y}_{n,d}|\Theta) = \mathcal{N}(\tilde{y}_{n,d}|\sum_{k=1}^{K} z_{n,k}w_{k,d}, \tilde{\tau}_{n,d})$. We can then define a new evidence lower bound $\mathscr{L}_2 = \sum_{n,d} \ln \tilde{P}(\tilde{y}_{n,d}|\Theta) - \mathrm{KL}(q(\Theta)||p(\Theta))$ such that $\mathscr{L}_2 < \mathscr{L} < P(Y)$, which can be optimised using variational updates for the Gaussian likelihood $\tilde{P}$, with pseudo-data $\tilde{y}_{n,d}$. This pseudo-data is a transformed version of the data which depends on the $q(\Theta)$ distribution and is dynamically updated throughout the optimisation process.

Approach from Seeger and Bouchard (2012)

The second derivative of the Bernoulli and Poisson log likelihoods has a negative lower bound κ, meaning that their second order Taylor expansion provides a quadratic lower bound $\mathscr{L}_2$. Using this expansion, one can show that the variational updates of the model parameters can be approximated by those of a Factor Analysis model with the following Normal likelihood:

$$\tilde{P}(\tilde{y}_{n,d}|Z,W) \sim \mathcal{N}(\sum_{k=1}^{K} z_{n,k}w_{k,d}, -\kappa) \tag{4.9}$$

with pseudo data $\tilde{y}_{n,d} = \xi_{n,d} - f'(\xi_{n,d})/\kappa$, where

$$f'(\xi_{n,d}) = \frac{\mathrm{d}\ln P(y_{n,d}|\sum_{k=1}^{K} z_{n,k}w_{k,d})}{\mathrm{d}\sum_{k=1}^{K} z_{n,k}w_{k,d}} \tag{4.10}$$

This introduces a new variational parameter $\xi_{n,d}$ corresponding to the location of the Taylor expansion. This parameter is updated using the following rule, which maximises the evidence lower bound $\mathscr{L}_2$:

$$\xi_{n,d} = \mathop{\mathbb{E}}_{q(\Theta)}\left[\sum_{k=1}^{K} z_{n,k}w_{k,d}\right] \tag{4.11}$$

In each iteration, the variational inference algorithm first updates the new parameters $\xi_{n,d}$ to compute the pseudo data $\tilde{Y}$, and uses it to update every model parameter using the update rules for Gaussian likelihood.

Approach from Jaakkola and Jordan (2000) for the Bernoulli likelihood

A major drawback of the lower bound provided by Seeger and Bouchard (2012) is that it provides a Gaussian likelihood approximation which is homoscedastic (its variance $-\kappa$ is the same for all sample and features). In biofam, we adapt the work from Jaakkola and Jordan (2000) who introduce the following heteroscedastic Gaussian lower bound to the Bernoulli likelihood:

$$\tilde{P}(\tilde{Y}_{n,d}|Z,W) \sim \mathcal{N}(\sum_{k=1}^{K} z_{n,k} w_{k,d}, \tau_{n,d}) \tag{4.12}$$

with pseudo data $\tilde{y}_{n,d} = (2y_{n,d}-1)\xi_{n,d}/\tanh(\xi_{n,d}/2)$, and precision $\tau_{n,d} = \tanh(\xi_{n,d}/2)/(2\xi_{n,d})$.

The variational parameter $\xi_{n,d}$ is updated with the following rule:

$$\xi_{n,d}^2 = \mathop{\mathbb{E}}_{q(\Theta)} \left[\left(\sum_{k=1}^{K} z_{n,k} w_{k,d} \right)^2 \right] \tag{4.13}$$

While the unique variance of the lower bound from Seeger and Bouchard (2012) needs to be suited for all values of ξ, the variance of the lower bound from Jaakkola and Jordan (2000) is adjusted to the location of the Taylor expansion, allowing a more flexible and therefore tighter approximation to the Bernoulli likelihood for all values of ξ, as illustrated in Figure 4.4.

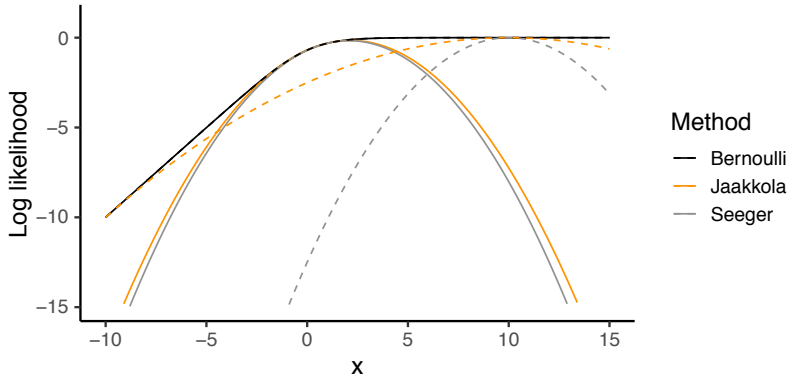


Fig. 4.4 Comparison of the lower bounds given by the Seeger and the Jaakkola approaches. The real Bernoulli likelihood as a function of $x = \sum_{k=1}^{K} z_{n,k} w_{k,d}$ is shown in black. The approximation to this likelihood for the Jaakkola and the Seeger approach are drawn for $\xi = 1$ (full line) and $\xi = 10$ (dashed line) as representative examples. It is apparent that the variance of the Jaakola lower bound is adjusted to the location of the approximation, providing a tighter fit.

4.4 Model Validation

In this section, we use simulated data in various settings to validate the different aspects of the biofam framework. We also investigate the identifiability of Factor Analysis in general. Unless stated otherwise, downstream analysis of the biofam results consider point estimates of the inferred variational distributions[1].

4.4.1 Structured Sparsity

The structured sparsity-inducing priors introduced in 4.2.2 can be used to encode our prior knowledge about existing groups of samples and features, for example in the joint analysis of samples from multiple tissues or cell types, with different feature views corresponding to multiple omics. Biofam uses group-specific ARD priors on the factors α_k^g, and view-specific ARD priors on the weights, α_k^m, which enable us to automatically detect to which sample groups and feature views a given factor is relevant. For example, a factor capturing cell-cycle state may be driving gene expression variation in liver tissue and irrelevant to samples from some brain regions.

[1]Note that variational inference is known to underestimate the parameter variances due to its objective function

Simulation setting

We simulated data for 800 samples and 1600 features drawn from the following generative model: $y_{n,d}^{m,g} = \sum_{k=1}^{K} z_{n,k}^{g} w_{k,d}^{m} + \varepsilon_{n,d}^{m,g}$, using 8 latent factors.

We defined two groups of 400 samples each (labelled group 0 and group 1), and two views, each consisting of 800 features, (labelled view 0 and view 1). The weights and factors were drawn from standard Normal distributions, and subsequently masked to obtain a desired sparsity pattern. Specifically, chosen weights and factors were set to zero so that factors 1 and 2 were relevant to all samples and all features; factors 3 and 4 were relevant to both views but only in group 0 and group 1 respectively; factors 5 and 6 were relevant to view 0 and view 1 respectively; factor 7 was relevant to group 0 and view 0 only and Factor 8 was relevant to group 1 and view 1 only. This structured sparsity pattern is illustrated in Figure 4.5.

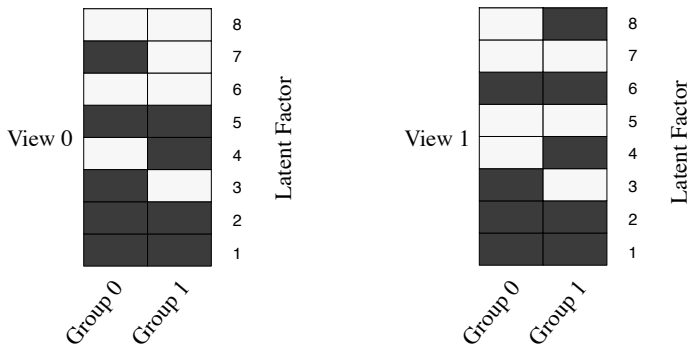


Fig. 4.5 Simulated sparsity structure. The binary matrices represent the relevance of every factor in each view and sample group. Left: factors relevance in view 0. Right: factor relevance in view 1. In each matrix, rows correspond to factors and columns to sample groups. A grey cell means that a given factor is relevant for a given sample group in the considered view.

The $\varepsilon_{n,d}^{m,g}$ noise terms were drawn from a Normal distribution $\varepsilon \sim \mathcal{N}(0, \sigma_{\varepsilon}^{2})$, where σ_{ε}^{2} was set so that the factors explained 30% of the total data variance (70% of noise).

Results

To assess the value added by the ARD priors for structured sparsity, we compared four models of increasing complexity (Fig. 4.6). The first model is a conventional Factor Analysis model, the second model uses per factor ARD priors on the weights, the third model employs

feature-group-specific ARD priors (Group factor Analysis) and the fourth model uses both ARD priors per sample group (on the factors) and ARD priors per feature group (on the samples). All models were implemented in the biofam framework, exploiting the modularity of the software implementation (Section 4.2.6).

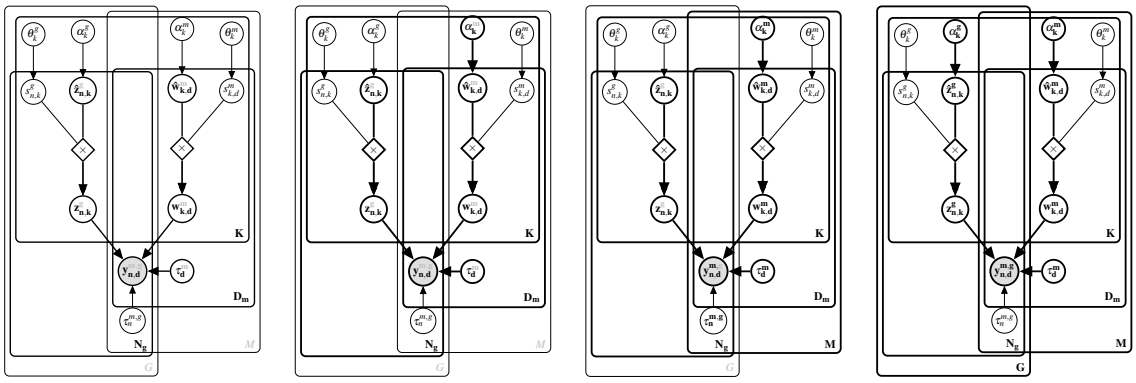


Fig. 4.6 Models compared for structured sparsity tests. From left to right: conventional Factor Analysis model; Factor analysis with ARD priors on the weights; Factor Analysis with view specific ARD priors on the weights; Factor Analysis with view-specific ARD priors on the weights and group-specific ARD priors on the factors. The specific priors used in each model are highlighted from the full graphical model of Figure 4.3.

Each model was fitted setting the number of latent factors to $K = 10$. We measured the relevance of the factor in each view and sample group using the coefficient of determination r^2 between data predicted using a given factor and the observed data (Fig. 4.7).

All models except for standard Factor Analysis identified the true number of factors, deactivating factors 9 and 10. This illustrates the sparsity-inducing ARD prior as a regulariser of model complexity, as already demonstrated in Section 2.2.5 (Fig. 2.10). In addition, the Group Factor Analysis model, which included view-specific ARD priors on the weights, was able to infer the view specific factor relevance, deactivating factor 5 and 6 respectively in view 1 and view 0. It was however unable to detect that factors 3 and 4 were specific to group 0 and group 1 respectively. Finally, the most complex model with both view-specific and group-specific ARD priors correctly inferred the simulated structured sparsity. These results demonstrate that models using structured sparsity-inducing priors provide a more accurate assessment of the relevance of individual factors in specific sample groups and views.

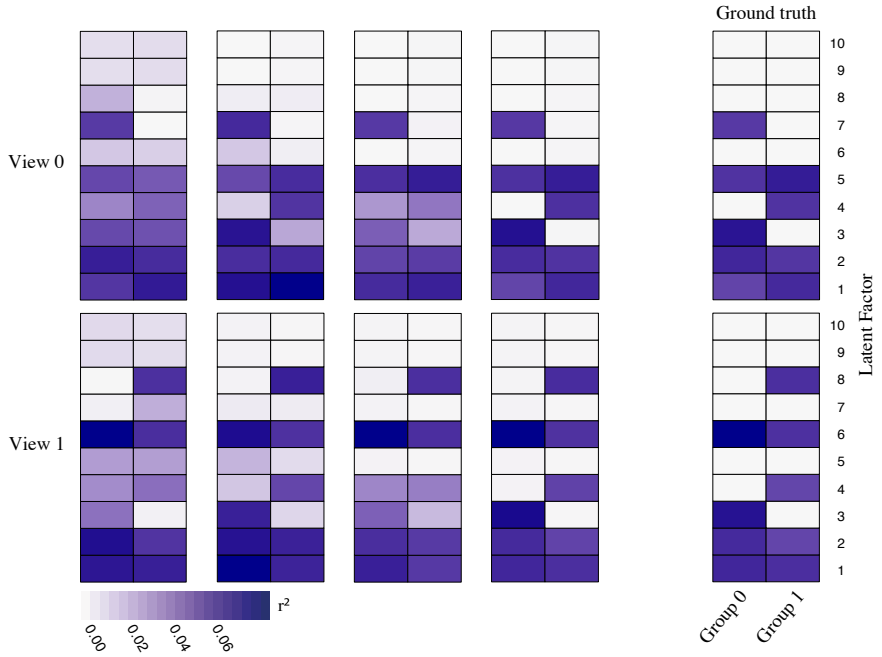


Fig. 4.7 Structured Sparsity inference for the models of Figure 4.6. From left to right: standard Factor Analysis; Factor Analysis with an ARD prior per factor on the weights; Factor Analysis with view-specific ARD-priors; Factor Analysis with both view-specific and group-specific ARD priors; ground truth. The first row of matrices correspond to view 0 and the second row to view 1. In each matrix, the first column corresponds to group 0 and the second column to group 1, rows correspond to the inferred factors. The colour scale corresponds to the fraction of variance explained by a given factor in a given view and group.

We then computed the correlation between the simulated weights and factors and the values inferred by all four models (Fig. 4.8) and found that models which explicitly accounted for the data context yielded more accurate results.

4.4.2 Element-wise Sparsity

Biological factors tend to affect a small subset of the features. In Section 4.2.3, we introduced spike-and-slab priors on the weights to encode this prior belief. In the following, we explore the effect of these priors on the inferred weight and factor matrices.

Simulation settings

First, to illustrate the effect of the spike-and-slab prior, we used a small number of factors so as to easily visualise the posterior distribution of the weights. We simulated data for 500

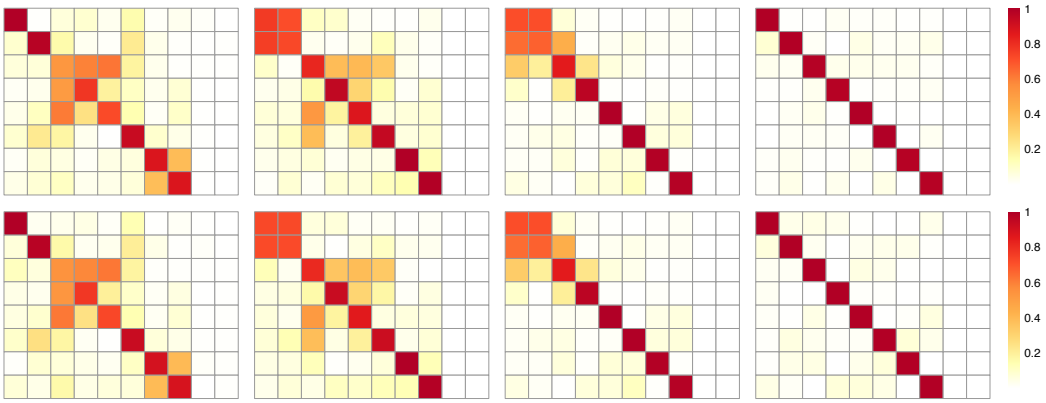


Fig. 4.8 Top: Pearson correlation between the 8 simulated factors and the 10 inferred factors for the 4 models compared in Figure 4.7. From left to right: conventional Factor Analysis model; Factor analysis with ARD priors on the weights; Factor Analysis for feature groups; Factor Analysis with view-specific ARD priors on the weights and group-specific ARD priors on the factors. Bottom: Analogous figure for the factor values.

samples and 20,000 features using 3 latent factors and a generative model as in Section 4.4.1: $y_{n,d}^{m,g} = \sum_{k=1}^{K} z_{n,k}^{g} w_{k,d}^{m} + \varepsilon$. The first factor was simulated to exhibit dense effects ($w_{k,d}^{m}$ drawn from a Normal distribution with no added sparsity); for the second factor 40% of the weights were set to zero and for the third factor, 95% of the weights were set to zero.

Additionally, we considered a second simulation setting, using 30 factors covering the entire range of sparsity levels: from 0% to 97% of weights set to zero. We simulated data for 800 samples and 1600 features and the factors explained 60% of the total variability.

Spike-and-slab priors enable inference of zero-inflated weight distributions

In both simulation settings, we first fitted a model with no spike-and-slab priors and a model with spike-and-slab priors on the weights. Both models had an ARD prior per factor on the weights. (Fig. 4.9).

In the first simulation setting, the two models were fitted with 5 factors. Both of them recovered the true number of factors and the two remaining factors were completely pruned. For both models, inferred weights were then compared to the ground truth, as shown in Figure 4.10. The sparse model inferred more accurate estimates of both sparse and dense weights.

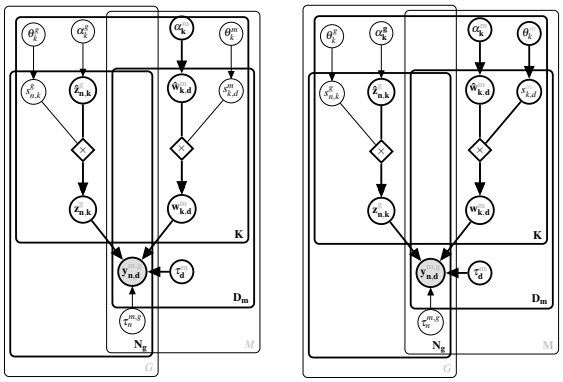


Fig. 4.9 Models compared for element-wise sparsity tests. Both models had ARD priors on the weights. In addition, the model on the right had spike-and-slab priors on the weights. The specific priors used in each model are highlighted from the full graphical model of Figure 4.3.

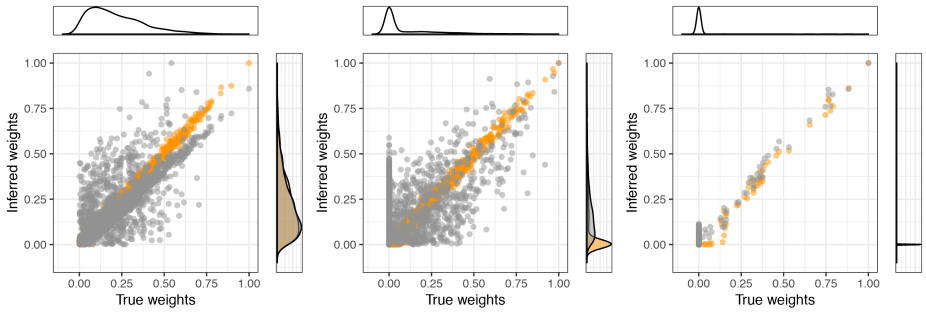


Fig. 4.10 Comparison of the absolute values of the weights inferred by the two Factor Analysis models of Figure 4.9 with the true simulated weights. From left to right factor 1 (0% sparse); factor 2 (40% sparse); factor 3 (95% sparse). In Yellow, model with spike-and-slab priors, in grey model without.

In the second simulation setting, we fitted both models with 30 factors and compared the inferred weight distributions for the five factors with the lowest degree of sparsity and the five factors with the highest degree of sparsity (Fig. 4.11). The model with spike-and-slab priors inferred both dense and sparse weight distributions, whereas a model without could not infer zero-inflated distributions. For the sparse model, we also compared the value of the inferred sparsity level per factor, $1 - \mathbb{E}_q(\theta_k)$ (where θ_k is the hyperparameter of the Bernoulli variables $s_{k,d}$), with the simulated sparsity level, and found that they were in accordance for the 11 sparsest weights. We will see in the next paragraph that denser weights are not identifiable by the model, which could explain the mismatch between simulated and inferred sparsity levels.

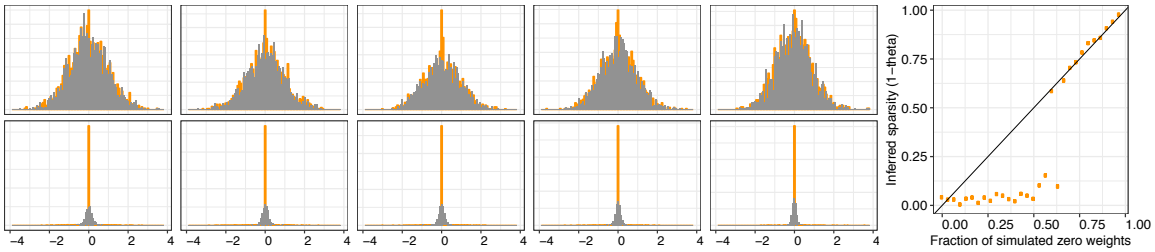


Fig. 4.11 Histograms of the weight distributions inferred by the two Factor Analysis models of Figure 4.9 for a simulation with 30 factors of different sparsity levels. Top: 5 denser factors. Bottom: 5 sparser factors. Yellow: model with spike-and-slab. Grey: model without spike-and-slab. The scatter plot shows the comparison across factors between simulated sparsity levels and the value of $1 - \mathbb{E}_q(\theta_k)$, where θ_k is the hyperparamter of $s_{k,d}$.

Taken together, these results illustrate the utility of the spike-and-slab prior to infer sparse weight matrices, which in turn may facilitate the interpretation of the resulting factors.

Spike-and-slab priors improve model identifiability

In the second simulation setting, we compared the simulated weights and factors with the inferred values using the two models of Figure 4.9, as well as a model which uses all biofam sparsity-inducing priors, including ARD and spike-and-slab priors on the factors (Fig. 4.12).

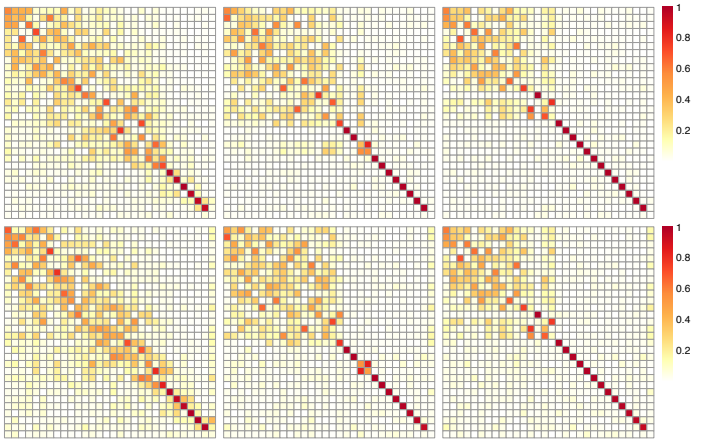


Fig. 4.12 Top row: Pearson correlation of ground truth weights with weights inferred with a model without spike-and-slab priors (left), with spike-and-slab priors on the weights (middle) and with spike-and-slab priors both on weights and factors (right). Factors are ordered by sparsity level: from the densest weights (left) to the sparsest weights (right). Bottom: Analogous figures for factor values.

For all models, we found that factors exhibiting sparse effects were inferred more accurately. These are the factors of main interest for biological applications as they are more likely to relate to meaningful and interpretable biological processes. We also found that the models using spike-and-slab priors yielded more accurate results than models with no element-wise sparstiy priors.

We then investigated the robustness of these results across three trials using different random initialisations of the latent factors. We found that, for all models, factors exhibiting sparse effects were more robustly estimated, and that models using spike-and-slab priors yielded more robust results than models with no element-wise sparsity-inducing priors. (Fig. 4.13).

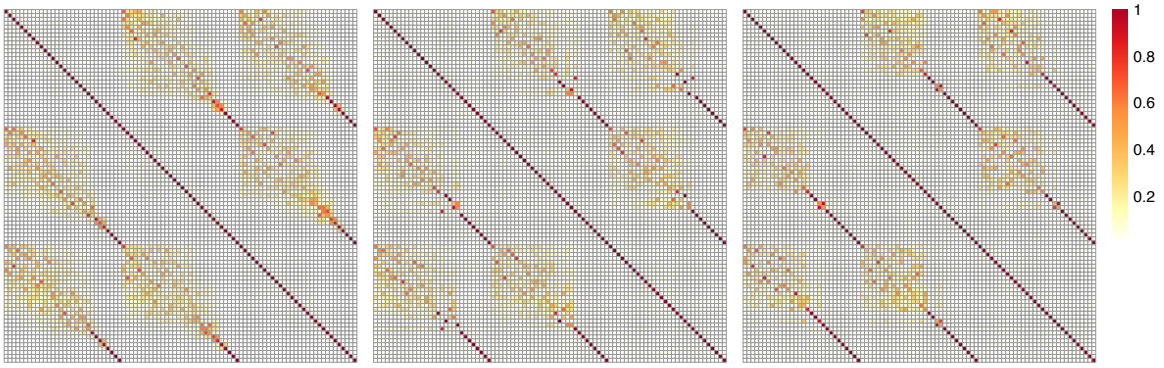


Fig. 4.13 Robustness of weights inference for a model without spike-and-slab priors (left), with spike-and-slab priors on the weights (middle) and with spike-and-slab priors both on weights and factors (right). Heat maps show the Pearson correlation matrices for all inferred weights for all three separate runs, so that off-diagonal blocks correspond to correlation plots between runs. Weight vectors are ordered by sparsity level as in Figure 4.12 From left to right: 5, 11 and 12 factors were robustly inferred across runs.

Taken together, these results show that sparser latent structures are more easily identifiable by Factor Analysis models, which is further illustrated in Appendix E, and that the use of element-wise sparsity-inducing priors further improves this identifiability. Interestingly, the use of additional spike-and-slab priors on the factors also yielded a marginal increase in identifiability, despite the fact that the factor matrix was not simulated as sparse.

As discussed in Section 4.2.3, the spike-and-slab priors breaks the rotational invariance property of Factor Analysis by encouraging the inference of sparser weights or factors. True sparse latent factors then become identifiable as shown here and in Appendix E. When

generative weights are truly dense however, sparsity-inducing priors do not solve the rotation invariance problem.

4.4.3 Multiple data modalities

Simulation setting

In order to validate the non-Gaussian likelihoods implemented in biofam, we simulated binary and count data for 100 samples and 3,000 features using 10 generative factors. As before, the factors and weights were drawn from standard Normal distributions. To simulate binary data from the generative model of Section 4.2.4, we used the rule of Equation 4.14, while we generated count data by rounding the value of the poisson rate $\lambda\left(\sum_k z_{n,k} w_{k,d}\right)$ of Section 4.2.4. In both cases, 25 repeat experiments were performed.

$$y_{n,d} = \begin{cases} 0 \text{ if } \sigma\left(\sum_k z_{n,k} w_{k,d}\right) < 0.5 \\ 1 \text{ otherwise} \end{cases} \tag{4.14}$$

Results

Biofam, with ARD and spike-and-slab priors on the weights, was fitted on both in silico datasets using a Gaussian likelihood and a Bernoulli and a Poisson likelihood respectively. The performance of the alternative models were compared using the evidence lower bound and the data reconstruction error between the ground truth and values generated from the fitted biofam model. We also reported the distributions of the reconstructed data using the biofam model.

Results show that both the Bernoulli likelihood (Fig. 4.14) and the Poisson likelihood (Fig. 4.15) yielded higher evidence lower bounds and lower reconstruction errors than the Gaussian likelihood for these in silico datasets. By looking at the distributions of the reconstructed data, one can also qualitatively appreciate the value of modelling these likelihoods explicitly.

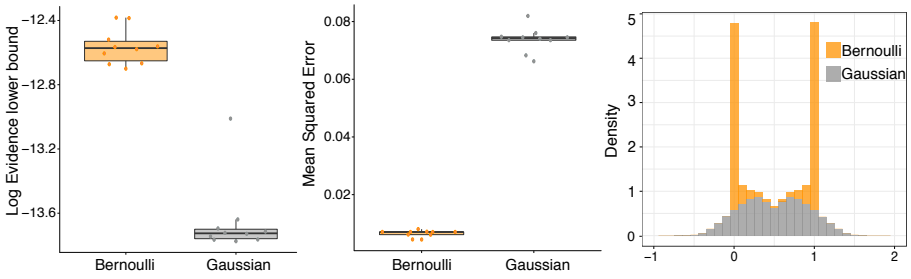


Fig. 4.14 Comparison of biofam results with a Bernoulli and a Gaussian likelihood on simulated binary data. Left: Evidence lower bound compared between the Gaussian and the Bernoulli likelihood. Middle: Difference in reconstruction errors from the fitted biofam model. Right: Distribution of the reconstructed data. For the Bernoulli likelihood, we show the values of Bernoulli parameter $\sigma(\sum_k z_{n,k} w_{k,d})$ (Section. 4.2.4).

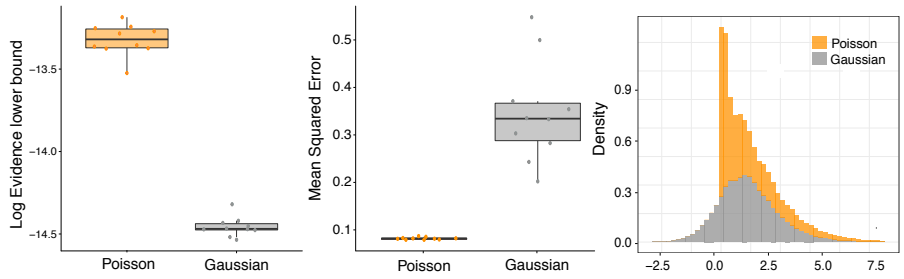


Fig. 4.15 Comparison of biofam results with a Poisson and a Gaussian likelihood on simulated count data. Left: Evidence lower bound compared between the Gaussian and the Poisson likelihood. Middle: Difference in reconstruction errors from the fitted biofam model. Right: Distribution of the reconstructed data. For the Poisson likelihood, we show the values of the rate $\lambda(\sum_k z_{n,k} w_{k,d})$ (Section. 4.2.4).

4.5 Computational cost and scalability

4.5.1 Standard inference

We compared the biofam performance to the main Group Factor Analysis competitor from Leppäaho et al. (2017), which uses an inference scheme based on Gibbs sampling. We found that biofam was faster and scaled linearly in the number of features, the number of samples and the number of latent factors used in our simulations (Fig 4.16).

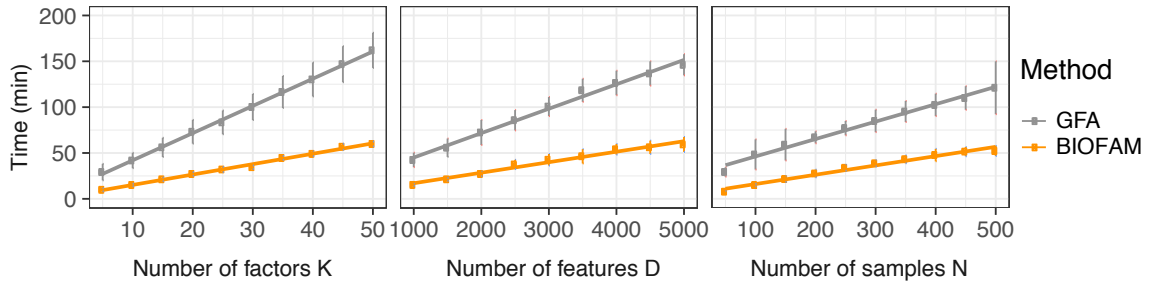


Fig. 4.16 Time required for model training for biofam and GFA (Leppäaho et al., 2017) as a function of the number of factors K, the number of features D and the number of samples N. Baseline parameters were $K = 10$, $D = 1,000$ and $N = 100$. Shown are average time across 10 trials, and error bars denote standard deviation.

4.5.2 GPU optimisation

The biofam computational bottleneck involves several operations involving large matrices. Performance may therefore be further improved by the execution of these operations on GPU using CUDA-implemented libraries. We use *cupy* (Okuta et al., 2017), a python library which provides an implementation of most standard matrix operations on GPU with an API based on *numpy*. For most sizes of datasets, GPU computations provided a three-fold performance increase (Fig. 4.17).

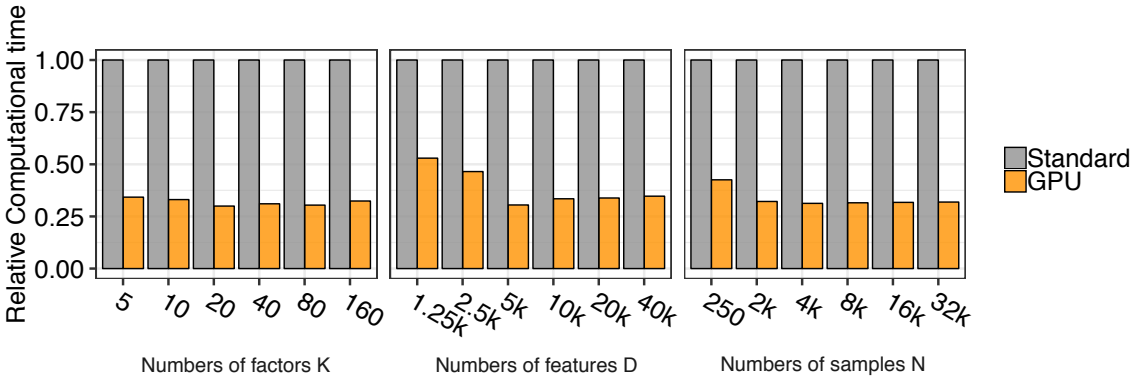


Fig. 4.17 Effect of GPU optimisation on relative computational time. The baseline dimensions are 1,000 samples, 5,000 features and 10 factors. GPU operations are run on a NVIDIA® Quadro® M6000 GM200GL GPU with 3,072 CUDA™ cores. Both the standard and GPU-optimised optimisations use an Intel® Xeon® CPU E5-2660 v3, 2.60GHz.

4.6 Extension: Stochastic variational inference

Variational inference is a fast method which typically outperforms sampling-based inference schemes such as Gibbs sampling (Blei et al., 2016). However, the size of biological datasets is rapidly increasing, for example in the field of single cell sequencing (Cao et al., 2017; Rosenberg et al., 2017; Dixit et al., 2016), hence motivating the development of an even more efficient inference framework to make biofam scalable to these latest datasets. Specifically, we adapt work from Hoffman et al. (2013) to implement a stochastic version of the variational inference algorithm using stochastic natural gradient ascent. Section 4.6.1 and 4.6.2 introduce the theoretical concepts supporting the method, Section 4.6.4 presents the specific algorithm implemented in biofam and Section 4.6.4 provides preliminary application results.

4.6.1 Natural gradient ascent

Variational inference can be regarded as an optimisation problem where the Evidence Lower Bound is maximised with respect to the variational distribution $q(\Theta)$ of the model parameters, which provides a tractable approximation to their true posterior distribution. A widely used approach in optimisation is gradient ascent, which is an iterative method where small steps are taken in the direction of the gradient until convergence towards a local maximum.

The gradient of a function f with respect to an input vector x, noted $\nabla_x f$, points in the direction of steepest ascent, i.e. the direction for which the smallest step dx in the input space results in the highest increase in the value of $f(x)$. In most gradient ascent methods, such as lbfgs (Bonnans et al., 2006), the size of the step dx is measured using a Euclidean distance $||\mathrm{d}x||$, meaning that the gradient is proportional to the solution of Equation 4.15.

$$\lim_{||\mathrm{d}x||\to 0}\left[\arg\max_{\mathrm{d}x} f(x+\mathrm{d}x)\right] \tag{4.15}$$

Let us define λ, the parameters of the variational distribution of the model parameters: $q(\Theta) = q(\Theta|\lambda)$. In variational inference, a small step dλ in the input space would ideally ensure that the *distributions* $q(\Theta|\lambda)$ and $q(\Theta|\lambda+\mathrm{d}\lambda)$ are close to each other. However, the Euclidean distance $||\mathrm{d}\lambda||$ between distribution parameters often poorly measures the similarity between the corresponding distributions[2]. The natural gradient (Amari, 1998; Asa-aki Sato, 2001; Honkela et al., 2010) addresses this issue by relying on the symmetric

[2]For example, consider the two Normal distributions $\mathcal{N}(0,1000)$ and $\mathcal{N}(10,1000)$. Their Euclidean distance would be 10 although the high variance makes them indistinguishable. In contrast , the distributions $\mathcal{N}(0,0.001)$ and $\mathcal{N}(0.1,0.001)$ would show very little overlap, as they are very peaked, but have a Euclidean distance of 0.1.

KL divergence (Eq. 4.16) as a measure of the distance between the distributions $q(\Theta|\lambda)$ and $q(\Theta|\lambda + \mathrm{d}\lambda)$.

$$\mathrm{KL}^{\mathrm{sym}}(p_1, p_2) = \mathrm{KL}(p_1||p_2) + \mathrm{KL}(p_2||p_1) \tag{4.16}$$

Formally, the natural gradient is proportional to the solution of Equation 4.17.

$$\lim_{\mathrm{KL}^{\mathrm{sym}}(q(\Theta|\lambda), q(\Theta|\lambda+\mathrm{d}\lambda)\to 0} \left[\arg\max_{\mathrm{d}\lambda} f(\lambda + \mathrm{d}\lambda) \right] \tag{4.17}$$

Hoffman et al. (2013) show that the natural gradient may be computed with a linear transformation of the Euclidean gradient using a Riemannian metric. The details of this computation for the Evidence Lower Bound is beyond the scope of this thesis and we will only give the final result demonstrated in Hoffman et al. (2013). As seen in Section 2.2.6, if the prior on the parameter Θ is conjugate for the considered likelihood, then $\exp\left[\mathbb{E}_{\theta_{j\neq i}}(\ln P(Y,\Theta))\right]$ remains in the same distributional family as $P(\theta_i)$, and therefore also the variational distribution $q_i(\theta_i)$. Assuming in addition that the prior distributions of these parameters are members of the exponential family, Hoffman et al. (2013) show that the natural gradient of the ELBO with respect to the parameters λ_i of $q_i(\theta_i)$ is:

$$\nabla_{\lambda_i}\mathscr{L} = \tilde{\lambda}_i - \lambda_i \tag{4.18}$$

where $\tilde{\lambda}_i$ represents the parameters of the $\exp\left[\mathbb{E}_{\theta_{j\neq i}}(\ln P(Y,\Theta))\right]$ distribution.

Natural gradient ascent then consists in optimising all λ_i iteratively using the rule of Equation 4.19, until convergence of the evidence lower bound.

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \rho \nabla_{\lambda_i}\mathscr{L}(\lambda_i^{(t)}) \tag{4.19}$$

Note that the update rule of the standard VB algorithm, $q(\theta_i) = \exp\left[\mathbb{E}_{\theta_{j\neq i}}(\ln P(Y,\Theta))\right]$, therefore corresponds to a natural gradient ascent algorithm with a step size $\rho = 1$.

4.6.2 Stochastic Gradient ascent

Stochastic gradient ascent (Robbins and Monro, 1951; Bottou, 2011; Spall, 2003) is a variation of gradient ascent, in which the gradient is approximated by noisy and unbiased estimates, which are cheaply computed using only a randomly sampled subset of the data. A

function f is maximised with respect to an input x using the rule of Equation 4.20, which is applied iteratively until f converges to a local maximum.

$$x^{(t+1)} = x^{(t)} + \rho^{(t)} b^{(t)}(x^{(t)}) \tag{4.20}$$

$b^{(t)}(x^{(t)})$ is the realisation at iteration t of a random variable B whose expectation is equal to the gradient of f: $\mathbb{E}(B(x)) = \nabla_x f(x)$. In stochastic gradient ascent, the step size $\rho^{(t)}$ is also adjusted at each iteration t. When this series satisfies $\sum_t \rho^{(t)} = \infty$ and $\sum_t (\rho^{(t)})^2 < \infty$, f is guaranteed to converge to a local maximum (Robbins and Monro, 1951). This result also applies to natural gradient ascent (Hoffman et al., 2013).

Stochastic gradient ascent is typically useful when the gradient of the objective function can be written as a sum over all observations: $\nabla_x f = \sum_{n=1}^{N} (\nabla_x f)_n$, where $n \in [\![1;N]\!]$ designates the observation index. When the size of the dataset is prohibitively high to compute the full gradient for inference, a common practice is then to sample uniformly at each iteration a random subset of the data called a mini-batch *mb*, and compute the following gradient noisy estimate: $b^{(t)}(f) = N/\text{size}(mb) \sum_{n \in mb} (\nabla_x f)_n$

4.6.3 Stochastic VB algorithm

Biofam implements a stochastic VB algorithm based on stochastic gradient ascent and adapted from Hoffman et al. (2013). The algorithm is as follows:

i) Initialise the weights $w^m_{k,d}$ randomly before the first iteration

ii) Sample uniformly a data mini-batch *mb* of size S. Mini-batches are sampled without replacement across each epoch[3]

iii) Update the parameters of $z_{n,k}$ $\forall k$ and $\forall n \in mb$ using standard variational methods

iv) Update all other parameters using stochastic natural gradient ascent as on Equation 4.24

v) Iterate steps 2 to 4 until ELBO convergence

Note that the factors are not updated stochastically but instead using standard variational updates. This is possible because factors are *local* parameters, meaning that the update of $z_{n,k}$ depends on observation n only, and does not depend on other factors $z_{n' \neq n,k}$. In the gradient

[3] an epoch contains the number of iterations necessary for the algorithm to see the entire dataset. For a batch-size of 20% of the data, an epoch contains 5 iterations.

ascent perspective, this means that the gradient of the elbo with respect to the parameters of $q(z_{n,k})$ can be computed deterministically from the sampled mini-batch, thus not requiring a resort to a noisy estimate.

Let us now derive the stochastic natural gradient ascent step for biofam. The Evidence Lower Bound can be decomposed as follows:

$$\begin{aligned} \mathscr{L} &= \mathop{\mathbb{E}}_{q(\Theta)} \ln \frac{P(Y,\Theta)}{q(\Theta)} \\ &= \mathop{\mathbb{E}}_{q(\Theta)} \ln P(\Theta) - \mathop{\mathbb{E}}_{q(\Theta)} \ln q(\Theta) + \sum_n \mathop{\mathbb{E}}_{q(\Theta)} \ln P(y_{n,:}|\Theta) \end{aligned} \tag{4.21}$$

A noisy estimates of $\mathscr{L}$ for the selected mini-batch can therefore be written as on Equation 4.22, where N is the number of observations and S the size of the mini-batch. The factor N/S ensures that the estimate is unbiased[4].

$$\mathscr{L}^S = \mathop{\mathbb{E}}_{q(\Theta)} \ln P(\Theta) - \mathop{\mathbb{E}}_{q(\Theta)} \ln q(\Theta) + \frac{N}{S} \sum_{n \in mb} \mathop{\mathbb{E}}_{q(\Theta)} \ln P(y_{n,:}|\Theta) \tag{4.22}$$

The natural gradient of $\mathscr{L}^S$ is a noisy unbiased estimate of the natural gradient of $\mathscr{L}$. Similarly to 4.18, it is given by:

$$\nabla_{\lambda_i} \mathscr{L}^S = \tilde{\lambda}_i^S - \lambda_i \tag{4.23}$$

where $\tilde{\lambda}_i^S$ represents the parameters of the distribution $\exp\left[\mathbb{E}_{\theta_{j\neq i}}\left(N/S \sum_{n\in mb} \ln P(y_{n,:},\Theta)\right)\right]$. This natural gradient can be easily computed by recycling the standard variational updates on the data mini-batch and their associated factors $z_{n\in mb,:}$ and the update rules for all other parameters become:

$$\begin{aligned} \lambda_i^{(t+1)} &= \lambda_i^{(t)} + \rho^{(t)} \left(\tilde{\lambda}_i^S - \lambda_i^{(t)}\right) \\ &= \left(1 - \rho^{(t)}\right) \lambda_i^{(t)} + \rho^{(t)} \tilde{\lambda}_i^S \end{aligned} \tag{4.24}$$

In their derivation of the ELBO natural gradient, Hoffman et al. (2013) make use of the fact that all prior distributions are of the exponential family. It is not the case for the spike-and-slab prior used in biofam for element-wise sparsity. At present, variational inference is therefore only implemented in biofam for non-sparse factors and weights. A similar inference

[4]One can prove that $\mathscr{L}^S$ is an unbiased estimate of $\mathscr{L}$, meaning that $\mathbb{E}\left(\mathscr{L}^S\right) = \mathscr{L}$, by using the fact that the mini-batches are sampled uniformly

scheme could easily be extended to sparse parameters, but with no theoretical guarantees of correctness.

4.6.4 Application

Simulation setting

We tested the stochastic inference algorithm using data simulated from the same generative model as in Section 4.4.1 and 4.4.2. We simulated 1,000 features for 30,000 and 50,000 samples alternatively, and using 10 and 20 factors alternatively. Weights were simulated as sparse, with only 15% of non-zero values.

Results

We fitted a Factor Analysis model with ARD priors on the weights and the factors, but no spike-and-slab prior. Stochastic inference was used with mini-batches made of 20% of the dataset, and a step size function of the form $\rho(t) = 1/(1+\nu t)^{3/4}$, where ν is a forgetting rate, which was set to a value of 0.9 in our applications. Figure 4.18 shows the evolution of biofam evidence lower bound as a function of the training time.

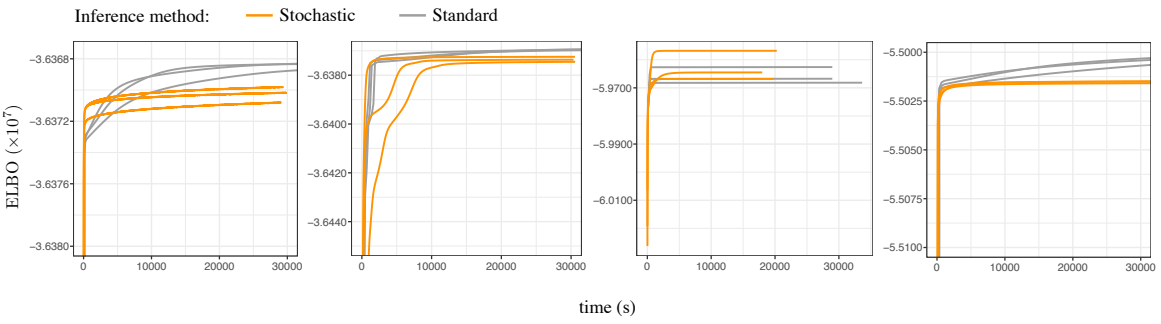


Fig. 4.18 Biofam performance with stochastic VB inference in simulated data of different dimensions. Shown is the Evidence Lower Bound as a function of the time. The number of simulated features is 1,000 for all simulation settings. The number of samples and factors is, from left to right: 30,000 samples and 10 factors; 30,000 samples and 20 factors; 50,000 samples and 20 factors; 50,000 samples and 20 factors.

For some tests, the first iterations of stochastic inference provided a fast convergence towards an approximate result, a illustrated in Figure 4.19 for an example with 30,000 samples and 10 factors, although after a longer time, the non-stochastic inference method converged towards

a higher evidence lower bound. In other cases, performances were similar between the two methods, or worse using stochastic inference.

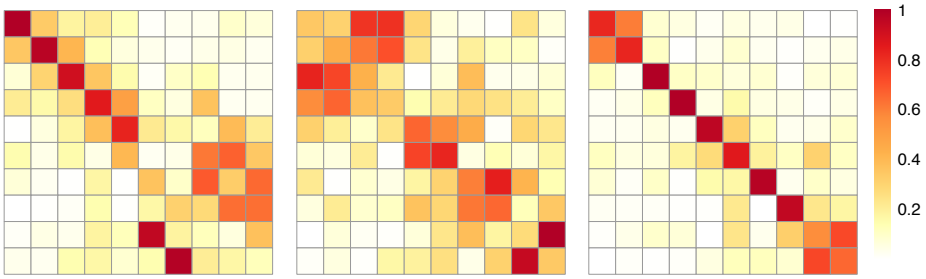


Fig. 4.19 Correlation across samples of the inferred factor values with the simulated values, for 50,000 samples and 10 latent factors. Left: after 2,475s of stochastic training. Middle: after 2,475s of non-stochastic training. Right: after 11,000s of non-stochastic training. In the stochastic case. 7 out of the 10 true factors are well recovered, while the non-stochastic version recovers only 2 of the 10 factors. After 11,000s of training, non-stochastic inference recovers 8 out of the 10 factors. In each case, shown are results with the highest ELBO across three random trials.

These mixed results could be due to a poor choice of step function hyperparameters, but at present we were unable to find any setting providing better performances. Attempts to implement more sophisticated algorithms for stochastic gradient ascent such as algorithms using momentum (Ruder, 2016) were also unfruitful. Section 4.7.3 discusses a possible direction of future work to make this inference scheme more beneficial.

4.7 Discussion

4.7.1 Comparison with other GFA implementations

Other implementations of Group Factor Analysis have been proposed and are related to the models presented in this Chapter. Table 4.3 compares the main characteristics of these implementations. The implementation which is the closest to biofam in terms of flexibility of the model options is the implementation of Leppäaho et al. (2017). However, Leppäaho et al. (2017) do not model group structure for samples. Additionally biofam variational inference is more scalable, as demonstrated in Figure 4.16, which is further improved by the GPU optimisation. Biofam also provides a comprehensive R package for the visualisation and downstream analysis of the results, which will be briefly presented in Chapter 5.

	Factor Sparsity		Weight Sparsity					
Model/Software	Group	Elem.	Group	Elem.	Likelihoods	Missing values	inference	GPU support
Virtanen et al. (2012)	✗	✗	ARD	✗	Gaussian	✗	VB	✗
Klami et al. (2015)	✗	✗	ARD	✗	Gaussian	✗	VB	✗
Remes et al. (2015)	GFA Mix.	✗	ARD	✗	Gaussian	✗	VB	✗
Zhao et al. (2016)	✗	✗	ARD	β-mix.	Gaussian	✗	VB	✗
Khan et al. (2014)	✗	✗	S&S	ARD	Gaussian	✗	Gibbs	✗
Bunte et al. (2016)	✗	S&S	ARD	S&S	Gaussian	✓	Gibbs	✗
Leppäaho et al. (2017)	✗	S&S	ARD	S&S	Gaussian	✓	Gibbs	✗
Biofam	ARD	S&S	ARD	S&S	Gaussian Bernoulli Poisson	✓	VB (+ Stochastic)	✓

Table 4.3 Comparison of the main characteristics of GFA implementations. Abbreviations: S&S: spike-and-slab; β-mix.: Beta mixture; VB: Variational Bayes; Elem.: Element-wise; GFA Mix.: Mixture of GFA models. Bunte et al. (2016) and Khan et al. (2014) link to the same software repository but describe two different models, probably reflecting the evolution of the software. The implementation of Leppäaho et al. (2017) is also derived from the work of Bunte et al. (2016) but provides a more user friendly R package.

The only model which offers a way of accounting for group structure on the sample axis was the model proposed by Remes et al. (2015). Briefly, the model relies on a mixture of factors, some of which are shared across all samples while some others are specific to sample groups. These groups do not need to be defined a priori. Instead, sample assignment to an arbitrary number of discrete clusters is modelled as a random parameter with a multinomial prior, and a conjugate Dirichlet prior for the parameters of the bionomial distribution. This is an appealing property with a slightly different aim from the biofam approach. Unfortunately, we were unable to find an implementation of the model online.

4.7.2 Comparison with alternative approaches

Alternative approaches have been proposed for the unsupervised analysis of structured datasets, in particular in the context of multi-omics analysis.

iCluster

iCluster+ (Mo and Shen, 2013; Mo et al., 2013), an extension of iCluster (Shen et al., 2009; Curtis et al., 2012) is a latent variable model, primarily used to perform clustering in multi-omics data. The underlying model is similar to the one of probabilistic PCA (Section 2.2.4), extended to multiple data modalities. Additionally, a lasso regularisation (Tibshirani, 1996) term is used on the weight matrices of each omic. Unlike the ARD prior, however, this regularisation technique does not enable selection of the model complexity as it is not factor specific. Instead, iCluster relies on multiple model fittings with different numbers of latent factors and post-hoc selection of the best fit/complexity tradeoff, which is computationally expensive and lacks a probabilistic interpretation. The lack of structured sparsity-inducing priors also precludes the model from distinguishing sources of variability which are shared across omics from the ones that are unique to specific omics, as we demonstrate in Argelaguet et al. (2018).

mixOmics

The mixOmics R package (Rohart et al., 2017) proposes multiple deterministic projection methods for the combined unsupervised analysis of multi-omics (DIABLO, Singh et al. (2016)) or multiple sample groups (MINT). Like iCluster, mixOmics rely on Lasso regularisation methods, whose parameters are fitted using cross validation, which can be computationally expensive. Although mixOmics offers functions for the analysis of data with grouping structures on the features or on the samples exclusively, these deterministic

projections do not rely on generative probabilistic models, and cannot model structured sparsity like we do with hierarchical priors. Like iCluster, they are tailored mostly for the purpose of clustering and two dimensional visualisation rather than disentangling and interpreting sources of variations across and between groups.

Network based methods

Kernel or graph-based methods have been proposed to combine different data types into a common similarity network between samples (Wang et al., 2014; Lanckriet et al., 2004). Briefly, these methods measure sample similarities in each data type independently and aggregate the results into a consensus network. As such, these methods provide a statistical framework for the integration of multiple datatypes to infer global relationships between samples. However, they do not pinpoint the molecular determinants of the resulting graph structure, and they do not aim to disentangle the sources of variability within each data type or within and between sample groups.

Tensor decomposition

When the same set of features (i.e. gene expression levels) is measured in the same samples in multiple environments, such as multiple tissues or experimental conditions, the resulting structured dataset can be represented as a tensor, a generalisation of matrices to higher order arrays (a matrix being a second order tensor). Latent variable models have been proposed to decompose the sources of variation in such structured datasets, both using deterministic methods (Carroll and Chang, 1970; Harshman and Lundy, 1994) and probabilistic methods (Zhao et al., 2014; Kolda and Bader, 2009; Hoff, 2010). In particular Hore et al. (2016) apply tensor decomposition for the joint dimensionality reduction of gene expression assayed in multiple tissues for the same samples.

The advantage of the tensor representation is that it enables the user to encode structured contextual information of any order. For example, a 4^{th} order tensor representation could be used to represent gene expression measured in multiple tissues and at multiple time points, and such structure would be accounted for when learning the distribution of the latent variables in tensor decomposition. However, the assumption made of matching samples and features across the multiple arrays of the tensor (e.g. tissues or time points) is very restrictive and only met by few datasets such as the Gtex dataset (GTEx Consortium, 2013). In contrast, the grouping structure of biofam enables us to jointly analyse multiple feature

sets of different dimensionality, or non-matching samples coming from different tissues or experimental contexts.

An extension of biofam to handle sample and feature grouping structures in higher order tensors could however be a direction of future work which would bridge the gap between the two modelling approaches.

Other Non-probabilistic latent variable models

Although this thesis focusses mostly on probabilistic models for dimensionality reduction, non-probabilistic methods are also widely used for the purpose of contextual data analysis such as multi-omics integration (Meng et al., 2016). Examples include Co Inertia Analysis (Fagan et al., 2007; Doledec and Chessel, 1994) and the MCIA (Meng et al., 2014) extension tailored to multi-omics or Generalised CCA (Tenenhaus and Tenenhaus, 2011, 2014; Tenenhaus et al., 2014). They often rely on linear generative models, but do not make use of prior architecture to model the data context and the resulting structured sparsity, using instead deterministic regularisation methods which require expensive cross-validation for hyperparameter fitting, and do not offer the modularity of probabilistic models.

4.7.3 Technical limitations and directions for future work

Informative priors

In biofam, we use a Bayesian framework to account for prior knowledge about sample and feature group structures in a principled manner. This Bayesian framework could be extended to account for other types of prior knowledge about the data features. For example, we could encode known pathways with binary informative priors on the model weights as in Buettner et al. (2017). Alternatively, weights could be modelled with a multivariate Normal prior distribution where the covariance encodes known continuous relationships between genes such as regulatory networks (Türei et al., 2016; Snel et al., 2000; Szklarczyk et al., 2017). Such approaches could improve factors interpretability and make the model more identifiable.

Similarly, relatedness between samples could be encoded using a multivariate prior on the factors. The covariance matrix could encode genetic relatedness (Speed and Balding, 2015) or other similarity measures between samples using a flexible covariance function. In an application to spatial expression data for example (see Section 2.1 and Chapter 3), the spatial relatedness could be encoded with a squared exponential covariance prior on the factors,

whose length scale could be jointly optimised with the model in an Expectation-Maximisation scheme.

Views coupling

Another type of prior knowledge that is not accounted for in the current biofam implementation is the coupling between different views. For example, if two views correspond respectively to DNA methylation and RNA expression, it would be desirable to encode explicitly the correspondence between methylation sites and the expressed genes, as the two biological variables are known to be strongly coupled (Clark et al., 2018). Biofam does not offer this flexibility, which is a major limitation of the current modelling approach. A solution to this problem could be based on an informative covariance prior on the model weight, as explained before, but this would hardly be combinable with the view-specific ARD priors which is at the core of the biofam approach. More adapted solutions to this pitfall would therefore likely require a probabilistic model which goes beyond the scope of Factor Analysis but should definitely be considered as a direction of future work.

Noise models

Although biofam differentiates itself from other Group Factor Analysis software by the greater diversity of likelihoods it offers, they are not ideal for some data modalities. For example, the Poisson likelihood does not account for overdispersion, and is therefore unadapted to the analysis of single cell expression data. Single cell RNA-seq techniques also provide zero-inflated count data for which specific noise models have been proposed (Pierson and Yau, 2015; Risso et al., 2018; Perraudeau et al., 2017) which are not implemented in the current version of biofam. RNA splicing yields binomial distributed data for which biofam does not offer a specific likelihood (Huang and Sanguinetti, 2017). Future work could include the implementation of those additional noise models in the variational framework of biofam.

Stochastic inference

In Section 4.6.4, we presented preliminary results from a stochastic extension of biofam variational inference scheme. These results do not represent a substantial improvement as they stand, but the implementation of stochastic inference in biofam is still the subject of ongoing research. A possible extension of the current framework could include the implementation of lazy IO functions, so that mini-batches are only loaded in memory when required for an iteration. Indeed, in stochastic variational inference, a single iteration of the algorithm does

not require the use of the entire dataset. In principle, this means that stochastic inference combined with lazy IO functions would make biofam applicable to datasets which do not fit in the computer memory.

Additionally, future work will consider the extension of the stochastic inference framework to sparse factors and weights.

Statistical testing

Finally, future work could also take advantage of the probabilistic interpretation and the modular implementation of the biofam software to perform hypothesis testing. For example, one could use the model evidence lower bound to compute a Bayes factor between a model using an ARD prior per factor and group and a model using an ARD per factor only, and thereby quantify the statistical significance of the differential relevance of factors between groups. To be truly insightful however, such tests should be performed on a per factor basis, so as to measure differential relevance of specific biological processes between groups, as opposed to a global effect of the grouping structure. This requires improving further the modularity of the software, and finding an efficient way to perform the test for every factor without refitting the model entirely each time. Another concern is the validity and interpretation of a Bayes factor computed using evidence lower bond instead of proper model likelihoods.

Chapter 5

Biofam applications

This Chapter exemplifies how bioFAM has been applied to different biological systems. The biological interpretation of the results presented in this Chapter is mainly the work of others. My individual contributions are detailed at the beginning of each section.

In Section 5.1, we introduce the biofamtools R package, which we use for the visualisation and downstream analysis of the biofam results. Section 5.2 illustrates the use of biofam for the joint analysis of data from multiple omics (group structure on the feature axis), while Section 5.3 showcases the application of biofam for the joint analysis of samples from multiple biological contexts (group structure on the sample axis). Finally, we give an outlook on ongoing applications in Section 5.4.

5.1 Biofamtools: visualisation and downstream analysis of the biofam results

The biofam core software is implemented in python and provides a basic interface which allows the user to define a specific biofam model, including the definition of known feature and group structures, specific sparsity-inducing priors (see Fig. 4.3) and data likelihoods. After training, the first moment of every model parameter is saved in an hdf5 file.

We developed an R package, biofamtools, to visualise and analyse these results. Biofamtools is an extension of the MOFAtools package, developed for the analysis of the MOFA results (Argelaguet et al., 2018). I designed the package with Danila Bredikin and Ricard

Argelaguet. Danila Bredikin and Ricard Argelaguet did most of the implementation, with the help of Yonatan Deloro, based on the MOFAtools package, implemented by Ricard Argelaguet and Britta Velten. I implemented the python IO functions to ensure compatibility between the core software and the R package.

Biofamtools provides the following main functionalities for the downstream analysis of biofam results:

- Calculation and visualisation of the fraction of variance explained by each factor in every sample group and view (see Figure 5.1).
- Visualisation of the samples in factor space, which provides, like in a standard usage of Principal Component Analysis, a compact overview of some of the main dependencies between samples (see Figure 5.4).
- Comparison of the weights and factors of different models (see Figure 5.2). This functionality is particularly convenient to check the robustness of a fitting process across multiple random starting points.
- Visualisation of the weights distribution and inspection of the top features with largest weights (see Figure 5.5). The loadings can give insights into the biological process underlying the heterogeneity captured by a latent factor.
- Feature Set Enrichment Analysis and visualisation of the enrichment results. This feature enables the automatic biological interpretation of factors in terms of Gene Set or more generally feature sets for other types of omics (see Section 5.2.4).
- Clustering of samples in discrete groups based on factors values.
- Imputation of missing data (see Section 5.2.3).
- Miscellaneous visualisation tools such as visualisation of data heat-maps ordered by factors values, giving insights into the effect of a given factor on features profiles; bees-warm plot visualisation of a given factor coloured by covariate values such as clinical covariates; correlation between factors, factors sparsity levels, training curve (elbo as a function of time).

In addition, biofamtools provides functions to run the model directly from R.

The package is open source and available at https://github.com/bioFAM/biofam/tree/master/BioFAMtools, where all functions are documented in detail. In this Chapter, most downstream analysis is based on biofamtools.

5.2 Application of biofam to multi-omics data

In this Section, we illustrate the use of biofam on multi-omics data, with selected results from Argelaguet et al. (2018). Note that all analysis were performed with the original MOFA and MOFAtools software (https://github.com/biofam/mofa), before they were extended into the biofam framework, but could be reproduced identically with this new software version[1]. In this section, we will therefore always refer to the *biofam* and *biofamtools* names.

The work presented here was supervised by Florian Buettner, Oliver Stegle, Wolfgang Huber and John C. Marioni. I conceived the model with Oliver Stegle and Florian Buettner and implemented the software with Ricard Argelaguet and Britta Velten. Ricard Argelaguet and Britta Velten led the work on the application presented in this section and I helped interpret the results with all other contributors.

5.2.1 Introduction

Experimental techniques increasingly enable gene expression profiles to be assayed in combination with multiple other omics, including genome, epigenome, proteome and metabolome (Hasin et al., 2017), and are applied across an increasing number of biological domains, including cancer biology (Gerstung et al., 2015; Iorio et al., 2016; Cancer Genome Atlas Research Network, 2017; Mertins et al., 2016), regulatory genomics (Chen et al., 2016), microbiology (Kim et al., 2016) or host-pathogen biology (Söderholm et al., 2016). The most recent technological advances also enable performing multi-omics analyses at the single-cell level (Colomé-Tatché and Theis, 2018; Guo et al., 2017; Clark et al., 2018; Angermueller et al., 2016; Macaulay et al., 2015). A common aim of such applications is to characterise heterogeneity between samples, as manifested in one or several of the data modalities (Ritchie et al., 2015a).

A basic strategy for the integration of omics data is testing for marginal associations between different data modalities. A prominent example is molecular quantitative trait locus map-

[1] MOFA can be regarded as a specific case of biofam

ping, where large numbers of association tests are performed between individual genetic variants and gene expression levels (GTEx Consortium, 2013) or epigenetic marks (Chen et al., 2016). While eminently useful for variant annotation, such association studies are inherently local and do not provide a coherent global map of the molecular differences between samples. In Section 4.7.2, we have also seen a number of unsupervised methods based on network analysis or clustering which have been applied to the analysis of multi-omics data. Although these methods can be used as rigorous tools for the integration of multi-omics data, and provide insight into relationships between samples, they do not reconstruct the underlying factors that drive the observed variation in an interpretable manner.

Group Factor Analysis provides a rigorous probabilistic framework for the reconstruction of interpretable drivers of variation in multi-omics data set. Here, we illustrate this by an application of biofam to a Chronic Lymphocytic Leukaemia (CLL) study, which combined gene expression data (transcriptome) with ex vivo drug response measurements, somatic mutation status and DNA methylation assays (Dietrich et al., 2018).

5.2.2 Data description and processing

The dataset consisted of RNA expression (RNA-Seq), somatic mutations (combination of targeted and whole exome sequencing), DNA methylation (Illumina arrays) and ex-vivo drug response screens (ATPbased CellTiter-Glo assay) (Dietrich et al., 2018). We selected the 200 samples for which at least two omics were measured. Yet, nearly 40% of these samples were profiled with some but not all omics types, highlighting the importance of a software which properly handles missing values (Fig. 5.1).

The drug response view included 62 drug response measurements at five concentrations each, making a total of 310 features. Somatic mutations which were present in at least three samples were included (69 in total). Low counts from RNA-Seq data were filtered out and the data was normalised using the *estimateSizeFactors* and *varianceStabilizingTransformation* function of the DESeq2 software (Love et al., 2014). We considered the top 5,000 most variable mRNAs after exclusion of genes from the Y chromosome. Methylation data were transformed to M-values (Du et al., 2010), and we extracted the top 1% most variable CpG sites excluding sex chromosomes, which resulted in 4,248 features.

More details on the data generation and processing can be found in the primary analysis paper (Dietrich et al., 2018).

5.2.3 Biofam results

We ran biofam with an ARD prior per view and per factor and spike-and-slab priors on the weights. Normal priors were used for the latent variables. The somatic mutation view was modelled with a Bernoulli likelihood, while other views were modelled with Gaussian likelihoods.

Structured variance overview

We initially fitted biofam with 25 factors and selected factors explaining 2% of variation or more in at least one view. This resulted in 10 factors which cumulatively explained 41% of variation in the drug response data, 38% in the mRNA data, 24% in the DNA methylation data and 24% in the mutation data (Fig. 5.1).

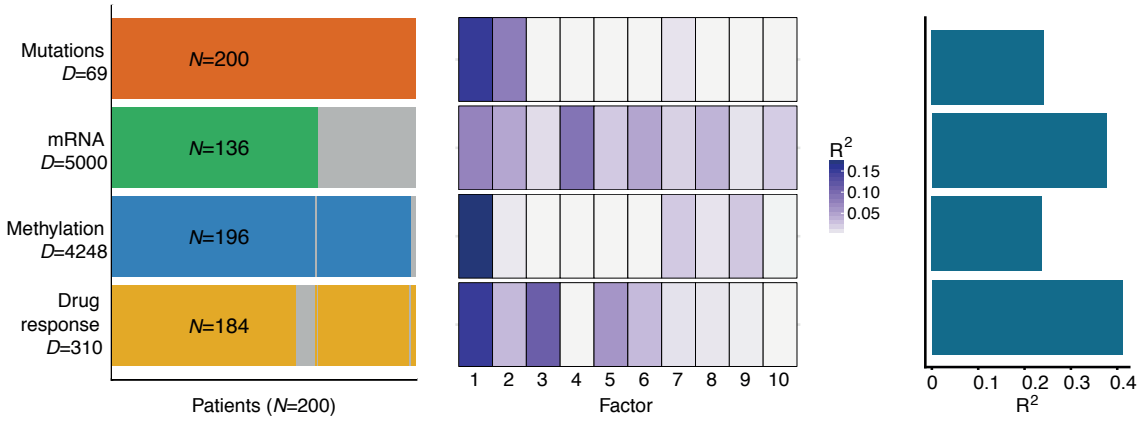


Fig. 5.1 Overview of the biofam results. Left: study overview and data types. Data modalities are shown in different rows (D = number of features) and samples (N) in columns, with missing samples shown using grey bars. Middle: proportion of total variance explained (r^2) by individual factors for each assay. Right: cumulative proportion of total variance explained.

Robustness

We assessed the robustness of factors and weights using 25 random initialisations of the latent variables, and comparing biofam results across trials. We found that most factors and weights were unchanged across trials (Fig. 5.2)

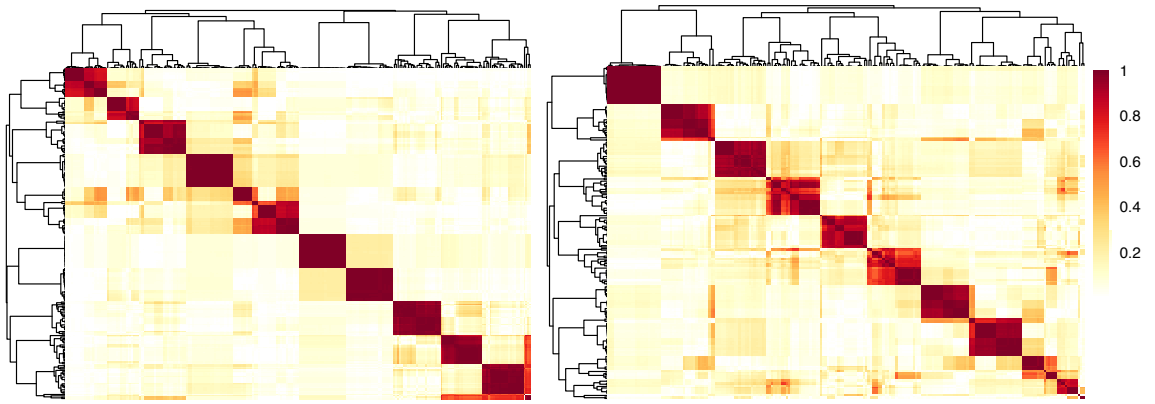


Fig. 5.2 Robustness of the biofam results across multiple initialisations. Left: absolute value of the Pearson correlation coefficient between the weights of the mRNA data. Right: absolute value of the Pearson correlation coefficient between the factors. Rows and Columns are clustered so that each block in the diagonal captures a weight (resp. latent factor) consistently learnt across multiple trials.

Missing values imputation

Incomplete data is a common problem in studies that combine multiple high-throughput assays, we assessed the ability of biofam to impute missing values within assays as well as when entire data modalities were missing for some of the samples. Non-missing data points were held out during training For both imputation tasks, biofam yielded more accurate predictions than other established imputation strategies, including imputation by feature-wise mean, SoftImpute (Mazumder et al., 2010) and a k-nearest neighbour method (Troyanskaya et al., 2001) (Fig 5.3).

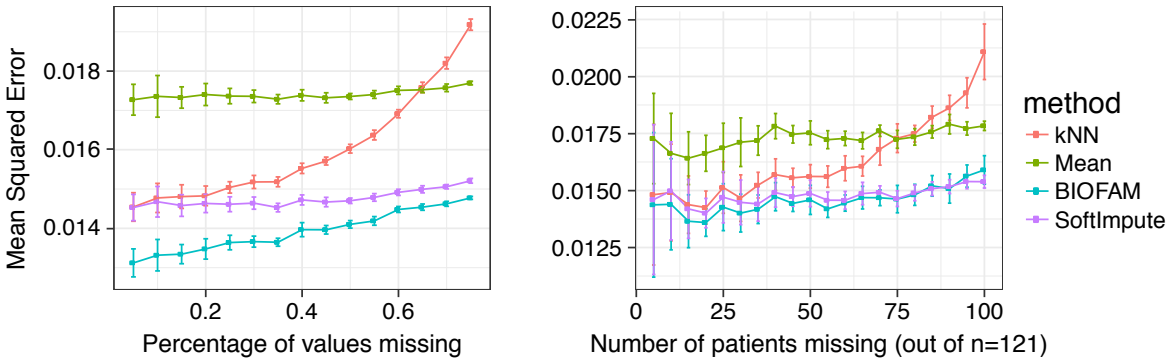


Fig. 5.3 Missing values imputation using biofam and alternative methods. Values are Mean squared error between ground truth and values imputed using the first moment of the weights and latent variables.

5.2.4 Factor interpretation

Sample visualisation in the latent space

We first plotted the samples in the two dimensional space made of the two factors explaining the most variance across all views. This identified four distinct subgroups, separated by their trisomy 12 status and somatic mutations on the immunoglobulin heavy-chain variable region gene (IGHV) (Fig. 5.4), two of the most important clinical markers and drivers of molecular disease heterogeneity in CLL (Zenz et al., 2010; Fabbri and Dalla-Favera, 2016).

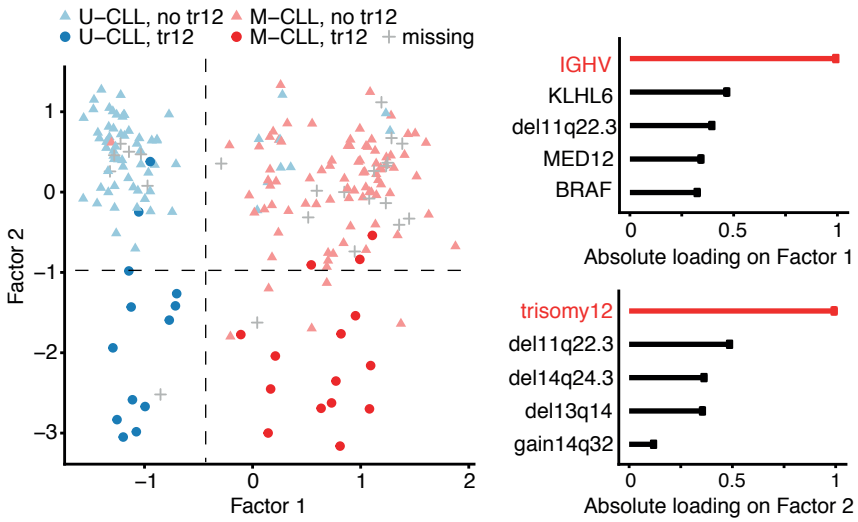


Fig. 5.4 Left: representation of samples using Factors 1 and 2. The colours denote the IGHV status of the tumours; symbol shape and colour tone indicate chromosome 12 trisomy status. Right: absolute weights of the top features of Factors 1 and 2 in the somatic mutations view.

This is in accordance with the high proportion of variance explained by Factors 1 and 2 in the somatic mutation view, which is largely attributable to these two mutations as can be seen by inspecting the weights of the biofam model (Fig. 5.4).

Characterisation of the IGHV factor across omics

IGHV status, the marker associated with Factor 1, is a surrogate of the differentiation state of the tumour's cell of origin and the level of activation of the B-cell receptor. While in clinical practice this axis of variation is generally considered binary (Fabbri and Dalla-Favera, 2016), our results indicate a more complex substructure (Fig. 5.5). At the current resolution, this factor was consistent with three subgroup models such as proposed by Oakes et al. (2016)

and Queirós et al. (2015), although there is suggestive evidence for an underlying continuum.

The variance breakdown of Figure 5.1 connected this factor to multiple molecular layers, including gene expression profile and drug response, which motivated the inspection of the associated weights to investigate the molecular determinants of this connection. We found that factor 1 affected genes previously linked to IGHV status (Plesingerova et al., 2017; Morabito et al., 2015; Trojani et al., 2011; Maloum et al., 2009; Vasconcelos et al., 2005) and responses to drugs targeting kinases in or downstream of the B-cell receptor pathway, in accordance with the reported association of the IGHV status with the level of activation of the B-cell receptor.

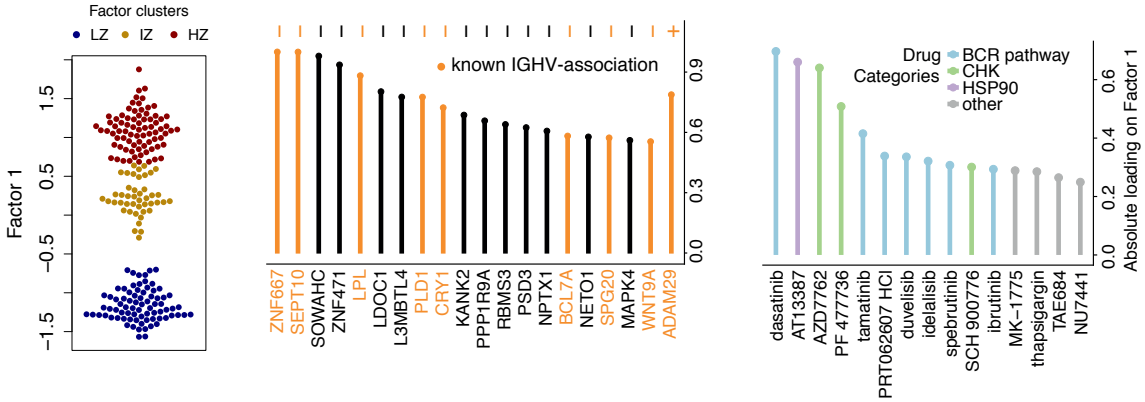


Fig. 5.5 Characterisation of Factor 1. Left: beeswarm plot with Factor 1 values for each sample with colours corresponding to three groups found by 3-means clustering with low factor values (LZ), intermediate factor values (IZ) and high factor values (HZ). Middle: highest absolute weights in the mRNA view. Plus or minus symbols on the right indicate the sign of the loading. Genes highlighted in orange were previously described as prognostic markers in CLL and associated with IGHV status. Right: absolute loadings of the drugs with the largest weights, annotated by target category.

Biofam therefore unveiled, in a single unsupervised analysis, multiple known associations between punctual mutations and molecular features, without the need of performing multiple individual tests as in eQTL analysis, demonstrating its utility for exploratory analysis.

Factors characterisation based on Gene Set Enrichment Analysis

Despite the clinical importance of the first two factors studied before, the other 8 factors accounted for more than 80% of the explained variance and the manual inspection of the

weight vectors for every factor and in every view can prove tedious. Gene Set enrichment analysis (Subramanian et al., 2005), which can be run directly from the biofamtools package, provides a compact overview of possible biological interpretations of every factor (Fig. 5.6). Weight inspection can be performed as a second step to refine GSEA results or better understand the link between multiple omics.

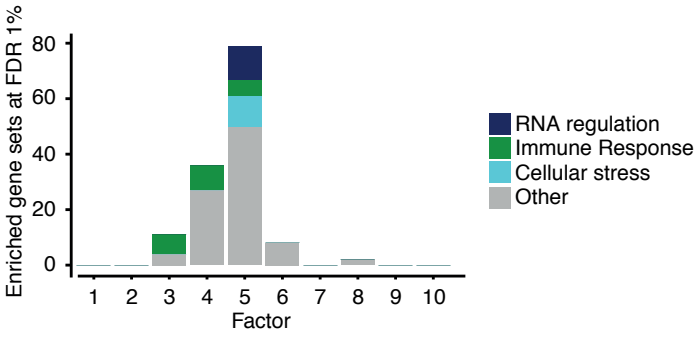


Fig. 5.6 GSEA results for biofam factors interpretation. Shown are numbers of enriched Reactome gene (Croft et al., 2011; Haw and Stein, 2012) sets per factor based on the gene expression data (FDR < 1%). The colours denote categories of related pathways.

For example, inspection of the top weights in the gene expression view for Factor 5, enriched for cellular stress, revealed genes coding for heat-shock proteins (HSP), which are essential for protein folding and are up-regulated upon stress conditions (Srivastava, 2002; Akerfelt et al., 2010). In accordance with this, we found that the drugs with the strongest weights on Factor 5 were associated with response to oxidative stress, such as target reactive oxygen species (ROS), DNA damage response and apoptosis (Fig. 5.7).

Although genes in HSP pathways are up-regulated in some cancers and have known roles in tumour cell survival (Trachootham et al., 2009), thus far this gene family has received little attention in the context of CLL, suggesting that unsupervised exploratory analysis of multi-omics data may be used to build hypothesis and direct future research.

Factors utility as predictors of clinical outcome

Finally, we explored the utility of the latent factors inferred by biofam as predictors in models of clinical outcomes. We used a Cox model (Spruance et al., 2004) to measure the significance of associations between biofam factors and patient survival. We identified three factors which were significantly associated with time to next treatment (FDR < 1%,

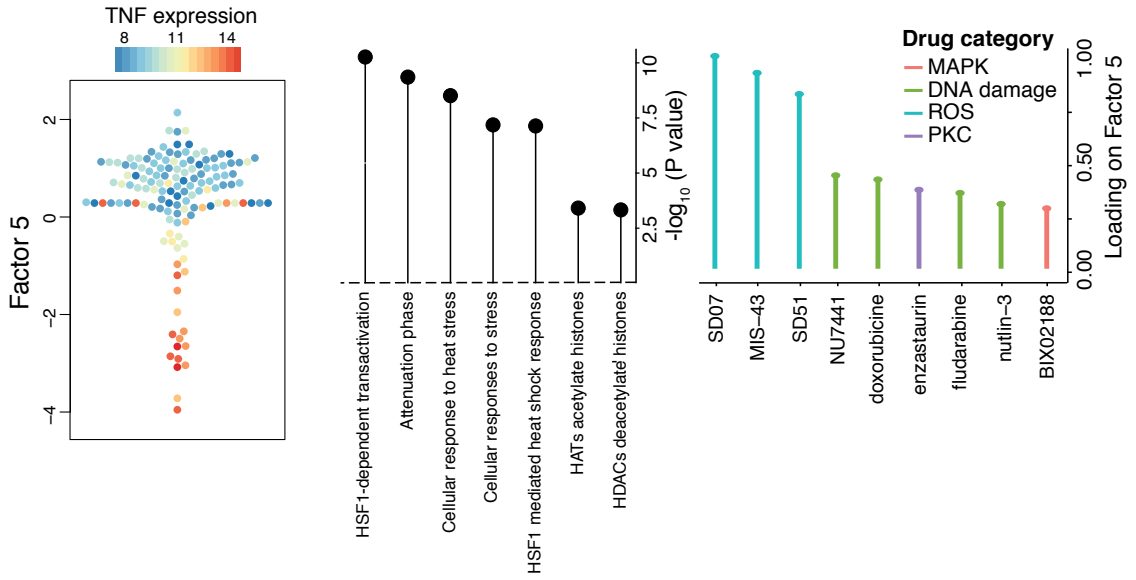


Fig. 5.7 Left: beeswarm plot of Factor 5. Colours denote the expression of TNF, an inflammatory stress marker. Middle: gene set enrichment analysis for the top Reactome pathways in the mRNA data using a t-test. Right: highest weights in the drug response view coloured by category.

Fig 5.8). Factor 1, characterised before, related to the B-cell of origin and Factors 7 and 8, were associated with chemo-immunotherapy treatment prior to sample collection ($P < 0.01$, t-test). In particular, inspection of the weights revealed that Factor 7 captured del17p and TP53 somatic mutations, as well as differences in methylation patterns of oncogenes such as Protein Kinase-C (Garg et al., 2014) and Crebb-P (Fluhr et al., 2016) while Factor 8 was associated with WNT signalling, a causative factor for several human cancers (Komiya and Habas, 2008).

Using 5-fold cross-validation in a multivariate Cox regression model, we also assessed the overall performance in predicting the time to next treatment when combining the 10 biofam factors. Notably, this model yielded higher prediction accuracy than models using individual molecular features (Fig 5.9). The predictive value of biofam factors was similar to clinical covariates (such as lymphocyte doubling time) that are used to guide treatment decisions.

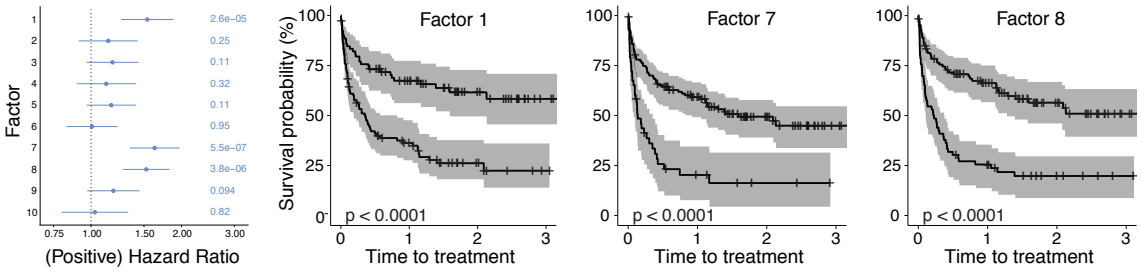


Fig. 5.8 Left panel: association of biofam factors to time to next treatment using a univariate Cox regression (Spruance et al., 2004) and P values based on the Wald statistic (Bangdiwala, 1989). Error bars denote 95% confidence intervals. Numbers on the right denote p-values for each predictor. Other panels: Kaplan–Meier plots measuring time to next treatment for the individual biofam factors. The cut-points on each factor were chosen using maximally selected rank statistics (Hothorn and Lausen, 2003), and P values were calculated using a log-rank test on the resulting groups.

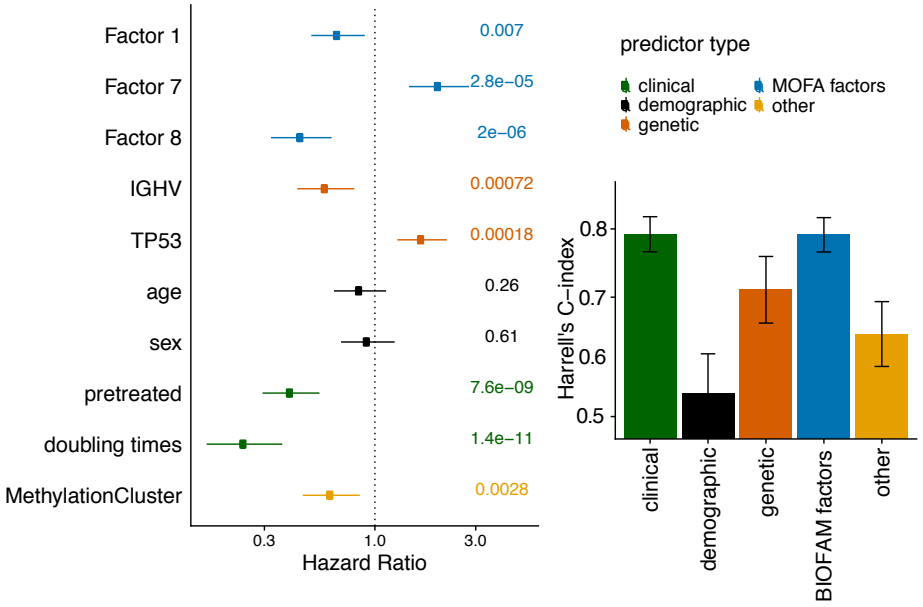


Fig. 5.9 Left: association of biofam factors and clinical covariates with time to next treatment using a univariate Cox models for 76 samples, for which the clinical information was available. Error bars denote 95% confidence intervals. Numbers on the right denote P values for each predictor. Right: prediction accuracy of time to treatment using multivariate Cox regression trained using the 10 factors derived using biofam as well as the selected clinical predictors. Shown are average values of Harrell C index from 5-fold cross-validation. Error bars denote standard error of the mean.

5.3 Joint analysis of multiple development stages of the mouse embryo

In a second application, we illustrate how biofam can be used for the joint analysis of multiple sample groups with an analysis of single cell RNA expression profiles measured in the mouse embryo at different stages of the gastrulation process. This work is an ongoing collaboration with Ricard Argelaguet, and builds on unpublished data generated by Hisham Mohammed and Stephen Clark, from Wolf Reik's group at the Babraham institute. Ricard Argelaguet processed the data, ran biofam, and interpreted most of the results. I organised the figures, wrote the text, and provided additional interpretation of the biofam weights based on literature research.

5.3.1 Introduction

Gastrulation is a phase in mouse embryonic development during which a single-layered blastula is reorganised to give rise, at embryonic day 7.5 (E7.5) to the three primordial germ layers: ectoderm, mesoderm and endoderm.

The onset of gastrulation is determined by the formation of the primitive streak at E6.5, a structure that emerges from the epiblast and establishes the initial bilateral symmetry of the body. Subsequently, involution of cells through the primitive streak gives rise to the mesoderm and endoderm, whereas epiblast cells establish the ectoderm (Solnica-Krezel and Sepich, 2012; Tam and Loebel, 2007; Tam and Behringer, 1997).

The gene expression dynamics at the different embryonic time points have been well characterised (Ibarra-Soria et al., 2018; Arnold and Robertson, 2009; Scialdone et al., 2016; Mohammed et al., 2017; Peng and Jing, 2017; Wen et al., 2017; Chan et al., 2018), and it hence provides an ideal system to showcase the application of biofam to multiple sample groups

5.3.2 Data description and processing

Here we used an unpublished data set where single-cell Nucleosome Methylation and Transcriptome (scNMT-seq) (Clark et al., 2018) was used to jointly profile chromatin accessibility, DNA methylation and gene expression from 743 single cells isolated from mouse embryos at

three developmental stages (E5.5, E6.5 and E7.5).

The aim of the study is to give a proof of concept for the utilisation of the biofam software in the comparative analysis of samples from different biological contexts. Thus, we only considered the RNA expression data, for which the gastrulation dynamics are well characterised in the literature. Yet, the final aim of biofam is to quantify the contributions from all three molecular layers to cellular diversity during germ layer formation, which is the subject of ongoing work.

RNA-seq libraries were aligned to the GRCm38 mouse genome build using HiSat2 (v2.1.0) (Kim et al., 2015) using options *–dta –sp 1000,1000 –no-mixed –no-discordant*, which yielded a mean of 611,000 aligned reads per cell. We discarded cells that had less than 100,000 reads mapped and less than 2,500 genes expressed. Gene expression counts were quantified from the mapped reads using featureCounts (Liao et al., 2014) with the Ensembl gene annotation (version 87) (Yates et al., 2016). Only protein-coding genes matching canonical chromosomes were considered. The read counts were log-transformed and size-factor adjusted (L. Lun et al., 2016).

Lineages were annotated using Single-Cell Consensus Clustering (Kiselev et al., 2017) to provide orthogonal cell labels to interpret the biofam results.

5.3.3 Biofam results

We built a biofam model with a Gaussian likelihood, ARD priors per gastrulation stage on the factors, and spike-and-slab priors on both the weights and the factors. For the purpose of this demonstration, we fixed the number of factors to 10. Features were centered within each group to avoid capturing trivial differential expression between stages, and focus instead on the structured variability within each stage.

Structured variance overview

First, we calculated the proportion of variance explained (r^2), at each developmental stage, by each factor individually and by the combination thereof (Fig. 5.10).

We observed factors capturing shared variability across all sample groups (i.e. Factors 1,3,7), and factors capturing variability unique to E6.5 (Factors 5 and 6) and E7.5 (Factors 2,4,8,9

and 10). No factor unique to E5.5 was observed. Notably, the total r^2, as well as the number of relevant factors inferred increased steadily across the gastrulation stages. This reflects an increase in the structured transcriptional variability of the gene expression profiles as cells transition from a relatively homogeneous blastula (E5.5) to a heterogeneous gastrula (E7.5).

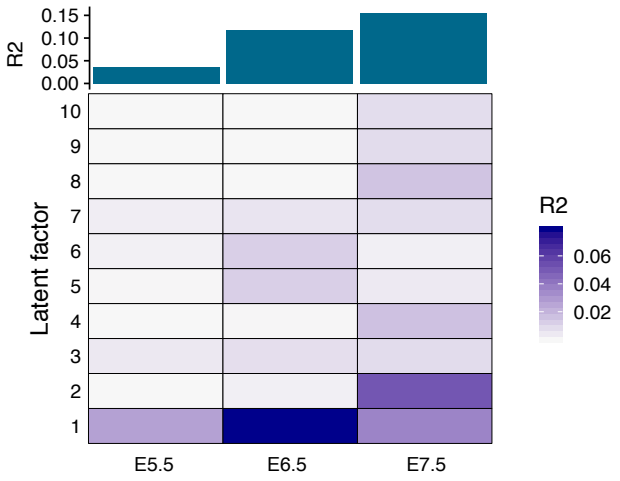


Fig. 5.10 Top: cumulative proportion of total variance explained for each stage. Bottom: proportion of total variance explained (r^2) by individual factors at each measurement stage. Factors were ordered by the total fraction of variance they captured across the three stages.

Results robustness

We then assessed the robustness of the biofam factors and weights using 5 random initialisations of the latent variables, and comparing results across trials. We found that all factors and weights were consistent across trials (Fig. 5.11)

5.3.4 Factor interpretation

Subsequently we characterised each one of the inferred factors by three complementary approaches: visualisation of the cells in the factor space, coloured by lineage identity; inspection of the genes with top weights and gene set enrichment analysis on the weights.

Factors capturing shared variability across stages

Factor 1 and 3 captured an important proportion of the gene expression variability shared across all stages (Fig. 5.10). Ordination of cells in the factor space, shows that these factors

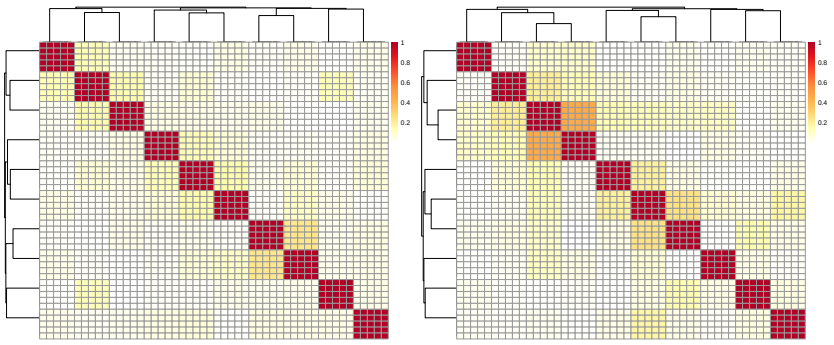


Fig. 5.11 Robustness of the biofam results across multiple initialisations. Left: absolute value of the Pearson correlation coefficient between the biofam weights. Right: absolute value of the Pearson correlation coefficient between the factors. Rows and Columns are clustered so that each block in the diagonal captures a weight (resp. latent factor) consistently learnt across multiple trials.

do not capture heterogeneity linked to lineage commitment (Fig. 5.12).

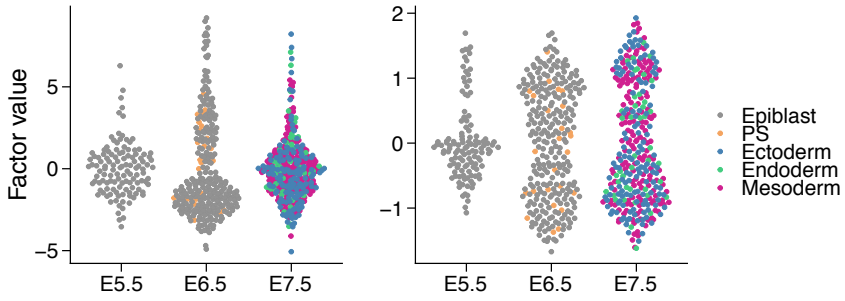


Fig. 5.12 Ordination of the cells along Factor 1 (left) and Factor 3 (right), coloured by lineage.

To investigate the molecular determinants of these factors, we inspected the corresponding biofam weights, revealing that Factor 1 shows a dense distribution of weights affecting most genes (Fig. 5.10). Previous studies have shown that dense factors tend to capture technical effects, as opposed to sparse factors that capture local gene coordinated variations (Gao et al., 2013). Accordingly, we find that Factor 1 is related to the cellular detection rate (i.e. number of expressed genes), a known technical artefact in scRNA-seq data (Finak et al., 2015).

In contrast, we observed that the weights of Factor 3 were enriched with genes involved in the cell cycle (Fig. 5.10), which is coherent with the high division rate observed in the considered stages of development (Solnica-Krezel and Sepich, 2012; Tam and Loebel, 2007; Tam and Behringer, 1997).

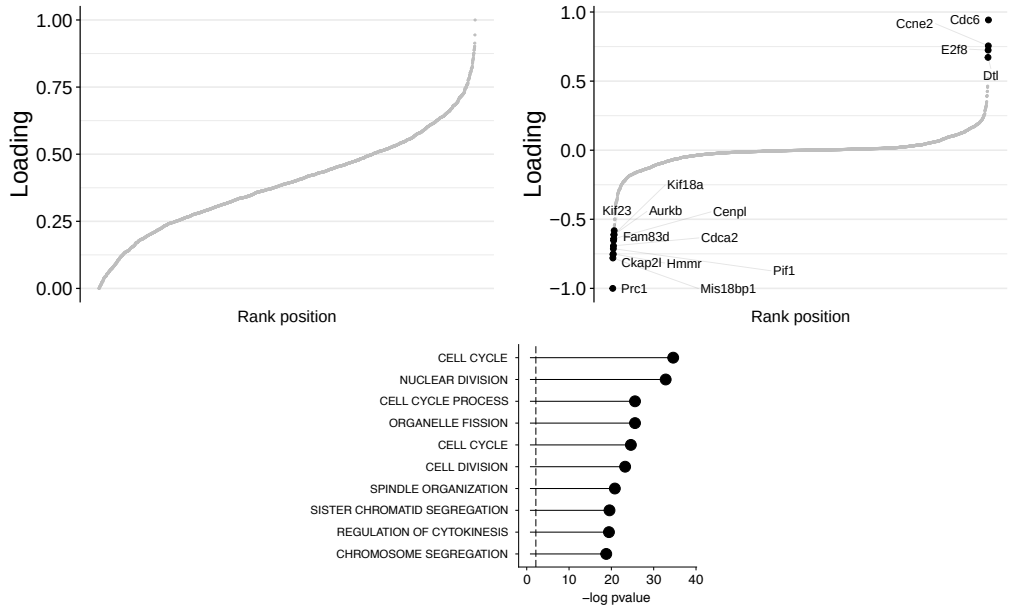


Fig. 5.13 Top: Ordination of the weights associated to Factor 1 (left) and Factor 3 (middle). Shown are the names of the genes with top weights in absolute value. Bottom: Gene ontologies enriched in the top weights associated to Factor 3.

Factors capturing variability unique to E6.5

Next, we considered Factor 5, which was specifically relevant to the E6.5 stage. Ordination of samples in factor space revealed that this factor captured the segregation of cells from the epiblast to the primitive streak (PS) state, a process which is known to determine the onset of gastrulation (Tam and Loebel, 2007) (Fig. 5.14).

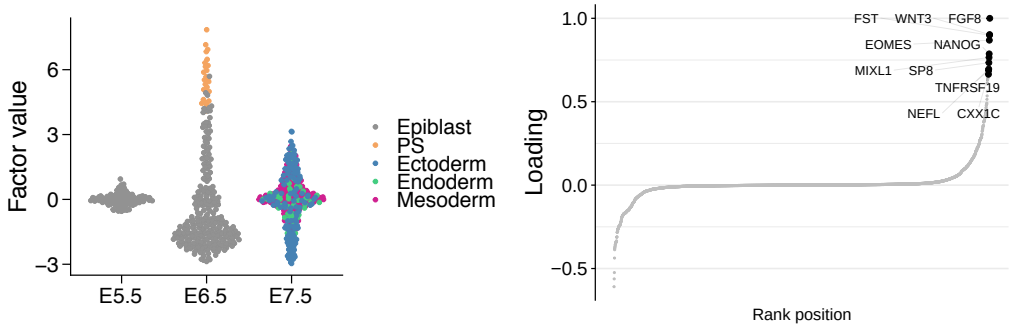


Fig. 5.14 Left: ordination of the cells along Factor 5 coloured by lineage. Right: ordination of the weights associated to Factor 5. Shown are the names of the genes with top weights in absolute value

Consistently with this observation, the genes with the top weights contained well-studied markers of PS formation, including Fgf8 (Sun et al., 1999), Wnt3 (Yoon et al., 2015) and Mixl1 (Ng et al., 2005) (Fig. 5.14).

Factors capturing variability unique to E7.5

As illustrated in Figure 5.10, the majority of factors were relevant to the E7.5 stage. Factor 2 separated the mesoderm cells from the ectoderm cells. Coherently, inspection of the corresponding weights revealed genes involved in the formation of mesoderm and ectoderm amongst the most determinant features, yet with opposite signs (Fig 5.16). Prominent examples are Lefty2 (Dai et al., 2016), Mesp1 (Chan et al., 2013), and Dll3 (Takahashi et al., 2003) for the Mesoderm; and Utf1 (Van den Boom et al., 2007) for the Ectoderm.

Factor 4 captured variation within the mesoderm lineage. While some of the top genes associated to this Factor were related to the mesoderm formation, other genes such as Mixl1 (Hart et al., 2002) and Sp5 (Weidinger et al., 2005) were coherently reported as specifically involved in the Mesoderm patterning, a process during which the mesoderm is further subdivided into organ domains (Fig 5.15).

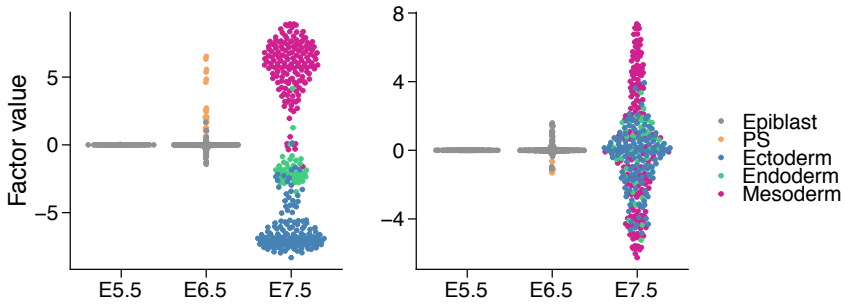


Fig. 5.15 Ordination of the cells along Factor 2 (left) and Factor 4 (right), coloured by lineage.

Analogously, Factor 8 prominently segregated endoderm cells, while Factor 9 captured patterning within this lineage (Fig. 5.17). Consistently, the genes most affected by these factors had functions related to the endoderm formation and subsequent patterning including Cer1 (Iwashita et al., 2013), Dkk1 (Ou et al., 2016), Trh (McKnight et al., 2007) (Fig. 5.18) and APLNR (Deshwar et al., 2016).

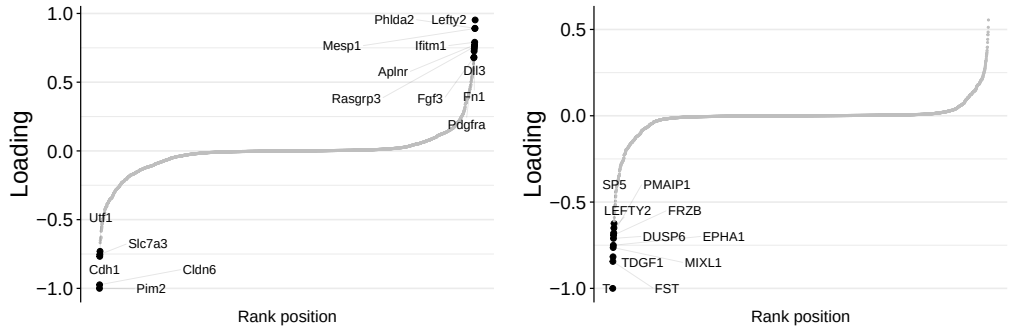


Fig. 5.16 Ordination of the weights associated to Factor 2 (left) and Factor 4 (right). Shown are the names of the genes with top weights in absolute value

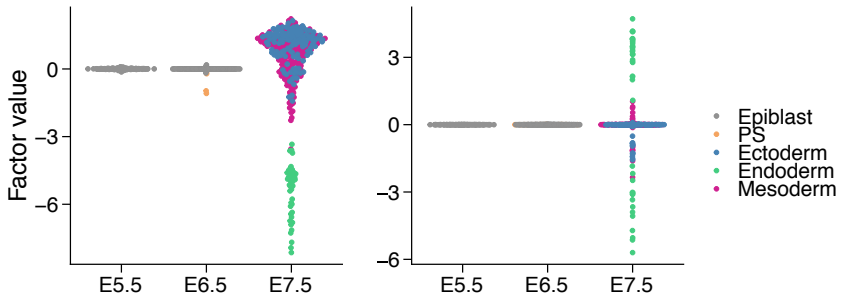


Fig. 5.17 Ordination of the cells along Factor 8 (left) and Factor 9 (right), coloured by lineage.

5.3.5 Conclusion

In a single analysis, biofam captured the main molecular determinant of the germ layer commitment from early E6.5 cells, which can be identified using only two dimensions of the biofam latent space (Fig. 5.19).

In addition, we illustrated how modelling explicitly the sample grouping structure formed by the distinct development stages provided a principled framework to compare the biological determinants of gene expression variation across biological contexts, thereby providing a unique perspective on the sequential nature of the gastrulation process. Although we restricted our analysis to the well-characterised RNA expression view, future work will include the generalisation of this analysis to the three biological layers measured by the sc-NMT technique (see Section 5.3.2).

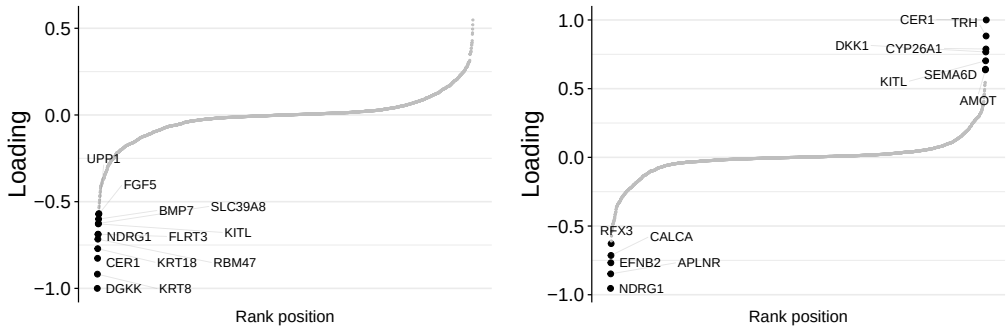


Fig. 5.18 Ordination of the weights associated to Factor 8 (left) and Factor 9 (right). Shown are the names of the genes with top weights in absolute value

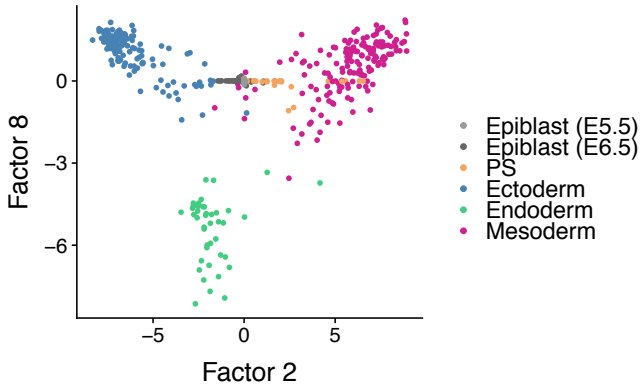


Fig. 5.19 Representation of the cells in factor space using Factor 2 (mesoderm formation) and Factor 8 (endoderm formation).

5.4 Future applications outlook

A key interest of the biofam framework is that it provides a novel perspective on the analysis of molecular differences between distinct sample groups. This is allowed by the explicit modelling of group structures on the sample axis, combined with an ergonomic visualisation package to summarise molecular differences between those groups.

To further explore this direction, we are currently analysing a dataset of 768 cells of the Hematopoietic Stem Cell Compartment in mice, across 2 age groups and 2 mutant groups (Kirschner et al., 2017). Here, biofam can provide a global map of the molecular differences between young and old mice, and investigate how those are affected by differences in genotype. We are also analysing the Tabula Muris dataset (The Tabula Muris Consortium et al., 2017) which consists in more than 100,000 cells from 20 mouse organs.

Here, the scalability of the biofam software is key to integrate so many cells in a single analysis, thereby comparing the biological determinants of multiple organ functions. Other applications could consider the application of Factor Analysis to standard case control studies, or the comparative study of molecular phenotypes between multiple environments.

Chapter 6

Concluding remarks

Recent technological advances in gene expression profiling have resulted in a variety of contextual gene expression datasets. This thesis aimed at developing statistical approaches to explicitly account for contextual information when modelling gene expression. The presented methods build on two distinct fields of Machine Learning research, Gaussian Processes and Factor Analysis. We gave a theoretical perspective on these approaches in Chapter 2.

In Chapter 3, we presented Spatial Variance Component Analysis (SVCA), a probabilistic model based on Gaussian Processes for the analysis of spatial gene expression data. Most prominently, SVCA assesses the effect of cell-cell interactions on gene expression profiles.

Using simulated data, we showed that SVCA yielded more accurate estimates of cell-cell interactions than alternative regression models and was more robust to different simulation settings. This was enabled by the flexibility of the Gaussian Process framework which allows modelling of non-linear positional effects with little prior knowledge about their functional form. We applied SVCA to a protein expression dataset assayed in human breast cancer biopsies, and an RNA expression dataset assayed in the mouse hippocampus. In these applications, we showed that cell-cell interactions are a major driver of gene expression variation at the single cell level, underlying the importance of developing models of gene expression variation which account for the spatial relationship between cells. We also discussed the biological interpretation of the SVCA cell-cell interaction estimates and found that they were largely in accordance with the function of the genes concerned, even if precise mechanistic interpretation remains difficult.

In Chapter 4, we presented biofam, a software for the unsupervised analysis of gene expression data in the context of multi-omics experiments and for the combined analysis of multiple sample batches, such as samples from different biological contexts, experimental conditions or tissues.

Biofam extends the framework of Group Factor Analysis. It combines the strengths of published implementations of Group Factor Analysis methods, and adds novel extensions to these models, such as the implementation of non-Gaussian likelihoods and the modelling of a sample group structure. It is implemented in a modular software which enables the selection of different types of sparsity-inducing priors in any combination, to best reflect assumptions about the data and to enable the comparison of different models implemented within the same framework. We showed that biofam variational inference scheme was performant and we presented ongoing work on a stochastic extension to this inference scheme. In applications to simulated data, we validated the impact of sparsity-inducing priors, and investigated their effect on model identifiability in different simulation settings. This showed that element-wise sparsity-inducing priors helped identify the true latent structure of the data.

In Chapter 5, we illustrated two use cases of the biofam software, with the help of the biofamtools R package for the visualisation and downstream analysis of biofam results.

In an application to a multi-omics dataset of chronic lymphocytic leukaemia, we showed that biofam was able to identify major drivers of variation in a clinically and biologically heterogeneous disease. Most notably, biofam identified previously known clinical markers as well as novel putative molecular drivers of heterogeneity, some of which were predictive of clinical outcome. In a second application we analysed single-cell RNA expression data assayed across multiple stages of the mouse embryo development. We illustrated how biofam can be used in the context of definite sample groups to provide a compact map of their molecular differences. In the gastrulation context, we showed that the biofam approach provided a coherent way to dissect the major processes involved in germ layer commitment. Biofam is still an ongoing project with more real data applications in preparation, in the context of ageing and cross-species comparisons.

Although the SVCA and biofam modelling approaches extend the state of the art in their respective fields, we have shown in Section 3.7.1 and Section 4.7.3 that both models have several limitations and still offer a lot of room for improvement and extensions. Most

importantly, SVCA would benefit from a finer understanding of the biological meaning of the cell-cell interactions measured, for which we could use hypothesis-driven research with simpler biological systems showing clear positive and negative controls. For biofam, a pressing direction of improvement is the extension of the current stochastic inference framework with lazy IO functions in order to permit the analysis of datasets which do not fit on the computer memory.

SVCA and biofam were developed as disconnected models and are tailored for different types of data contexts. However, future work may try to bring these approaches together in a unified framework where spatial context is modelled jointly with latent factors of variations. A first approach could be to encode the spatial relatedness of cells with a squared exponential covariance prior on the latent variables of the biofam model, whose length scale could be jointly optimised with the model in an Expectation-Maximisation scheme. Although we originally implemented such a spatial covariance feature in the biofam package, it was subsequently dropped for two reasons. First, experiments on simulated data, as well as spatial transcriptomics (Ståhl et al., 2016) and seqFISH (Shah et al., 2017) data, yielded undistinguishable results from standard biofam. Second, the dependency across samples introduced in the prior distribution of the latent variables had a cost in terms of computational complexity, as also demonstrated in Hore (2015). In the case of a univariate prior on the latent variables, and an approximation to the posterior distribution which is factorised over samples, updates of a given latent variable are independent across samples (see Appendix D). This enables a fast vectorised implementation which is no longer possible with a multivariate prior. Future work should address these problems or explore other alternatives.

Building generative models for bioinformatics requires a thought process which goes in round trips between three components: the model, the data and the biological question or purpose. New perspectives on the studied question or data may emerge from this thought process, which go beyond technical developments alone. For example, Group Factor Analysis did not bring any technical novelties from the already widely used ARD prior. It did however provide a new outlook or perspective on the analysis of data from different sources, and, with it, a principled framework to explore novel questions. Likewise, the models presented in this thesis combine existing models and methods to approach specific datasets and biological questions from a new angle. For example, biofam offers a new perspective on the analysis of global molecular differences between experimental conditions, and provides a principled

statistical framework to study them. We hope that this work will inspire users to address new biological questions in this direction.

Appendix A

Supplementary materials for SVCA

A.1 Methodological notes

A.1.1 Gradient derivation for the cell-cell interaction term

As seen in Section 2.1.8, Gaussian Process hyperparameters are optimised by maximising the following Log marginal likelihood using gradient ascent:

$$\ln P(Y|X,\theta) = -\frac{1}{2}Y^T (K)^{-1} - \frac{1}{2}\ln|K| - \frac{n}{2}\ln 2\pi \tag{A.1}$$

The gradient with respect to the hyperparameter θ_i can be computed in closed form, and depends on the derivative of the kernel matrix $\partial K/\partial\theta_i$

$$\frac{\partial}{\partial\theta_i}\ln P(Y|X,\theta) = \frac{1}{2}Y^T K^{-1}\frac{\partial K}{\partial\theta_i}K^{-1}Y - \frac{1}{2}\mathrm{tr}\left(K^{-1}\frac{\partial K}{\partial\theta_i}\right) \tag{A.2}$$

For an additive kernel, as in SVCA, the derivative $\partial K/\partial\theta_i$ is equal to the sum of the derivatives for each kernel term which can be computed independently. Generally speaking, this property enables a very modular implementation of Gaussian Processes. In SVCA, we reused the gradient derivations of the squared exponential kernel and the linear kernel, as implemented in the limix package (Lippert et al., 2014).

For the cell-cell interaction term with scaling hyperparameters σ_{c-c} and length scale l, we implemented the following gradient:

$$\begin{aligned}\frac{\partial}{\partial \sigma_{c-c}} K_{c-c} =& \frac{\partial}{\partial \sigma_{c-c}} \sigma^2_{c-c} ZXX^T Z^T \\ =& 2 \times \sigma_{c-c} ZXX^T Z^T\end{aligned} \tag{A.3}$$

$$\begin{aligned}\frac{\partial}{\partial l} K_{c-c} =& \frac{\partial}{\partial l} \sigma^2_{c-c} Z(l) XX^T Z(l)^T \\ =& \sigma^2_{c-c} \frac{\partial Z(l)}{\partial l} \times XX^T Z(l)^T + \sigma^2_{c-c} Z(l) XX^T \times \frac{\partial Z(l)^T}{\partial l}\end{aligned} \tag{A.4}$$

With $Z_{i,j} = \exp(-d^2_{i,j}/2l^2)$, we find $\partial Z_{i,j}/\partial l = -(d^2_{i,j}/l^3)\exp(-d^2_{i,j}/2l^2)$ and:

$$\frac{\partial}{\partial l} K_{c-c} = -(d^2_{i,j}/l^3) \times 2 \times \sigma^2_{c-c} Z(l) XX^T Z(l)^T \tag{A.5}$$

We implemented the cell-cell interaction covariance term and gradient derivation in the limix package, although the applications of this thesis used a grid search strategy for the length scale hyperparameter to avoid local optima.

A.1.2 Note on the marginalisation property and out of sample predictions with SVCA

Out of sample predictions are performed using the predictive distribution for Gaussian Processes and taking its mean as a point estimate:

$$\begin{aligned}P(y^*|X^*,X,Y) = \mathcal{N}(y^*|& K(X^*,X)\left(K(X,X)+\sigma^2_\varepsilon I\right)^{-1} Y, \\ & K(X^*,X^*) - K(X^*,X)\left(K(X,X)+\sigma^2_\varepsilon I\right)^{-1} K(X,X^*))\end{aligned} \tag{A.6}$$

This requires to compute the covariance between training samples $K(X,X)$ and the cross-covariance between test samples and training samples $K(X^*,X)$. A fundamental property of Gaussian Processes is that the covariance between two samples i and j is only a function of their input x_i and x_j: $\text{cov}(y_i,y_j) = k(x_i,x_j)$. As a consequence, the covariance between training samples $K(X,X)$ should not be affected by the observation of new samples with input x^*. This is known as the consistency requirement or marginalisation property of Gaussian Processes (Rasmussen and Williams, 2006).

In SVCA, the cell-cell interaction term renders fulfilling this requirement challenging, because the cell-cell interaction covariance term between two cells involves the expression profile of all cells in the tissue. To address that, we consider that the test set's expression

profiles x^* are observed during training when computing the cell-cell interaction covariance. The output value y^*, in contrast, is not observed and not used for training. This ensures that the value of the training set's covariance $K(X,X)$ is not affected by the observation of a new sample y^*.

When comparing out of sample predictions results with simpler regressions, we make sure that the comparison is fair by also accounting for the expression profiles of all cells (including test sets) in the computation of the cell-cell interaction term.

A.2 Signature robustness on real data using bootstrapping

We used a bootstrapping strategy, combined with t-SNE dimensionality reduction (Maaten and Hinton, 2008) to visualise the robustness of the SVCA variance estimates. For a given image and a given protein, the model was fitted 4 times on a different randomly drawn subset of 80% of the cells and we represented the signatures obtained for all bootstraps and all images in a low dimensional latent space using t-SNE (Fig. A.1 for IMC, Fig. A.2 for seqFISH). Every point corresponds to a SVCA signatures, and signatures for multiple bootstraps of the same image are linked with a line. The closer the bootstraps in the latent space the most robust the signatures were. Intrinsic, environmental and cell-cell interaction components were also analysed independently with the same procedure.

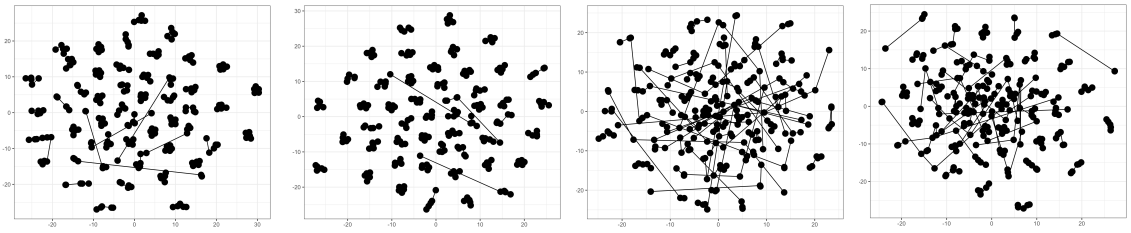


Fig. A.1 Robustness of the IMC variance signatures using bootstrapping and t-SNE visualisation. From left to right: full signatures, intrinsic effect, environmental effect, cell-cell interactions effect.

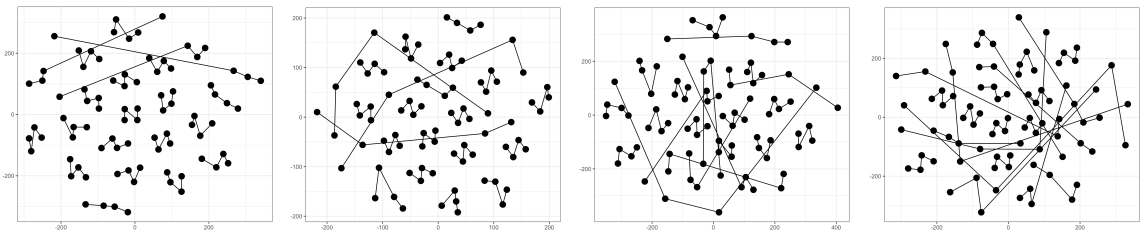


Fig. A.2 Robustness of the seqFISH variance signatures using bootstrapping and t-SNE visualisation. From left to right: full signatures, intrinsic effect, environmental effect, cell-cell interactions effect.

A.3 Comparison between variance components for both real data applications

Motivated by the similarity of the results obtained in the gene set enrichment analysis of the intrinsic component and in the gene set enrichment analysis of the cell-cell interaction component in the application to the seqFISH hippocampus data (Section 3.6.3), we compared the cell-cell interaction component with the three other model components for both the IMC application of Section 3.5 (Fig. A.3) and the seqFISH application of Section 3.6 (Fig. A.4).

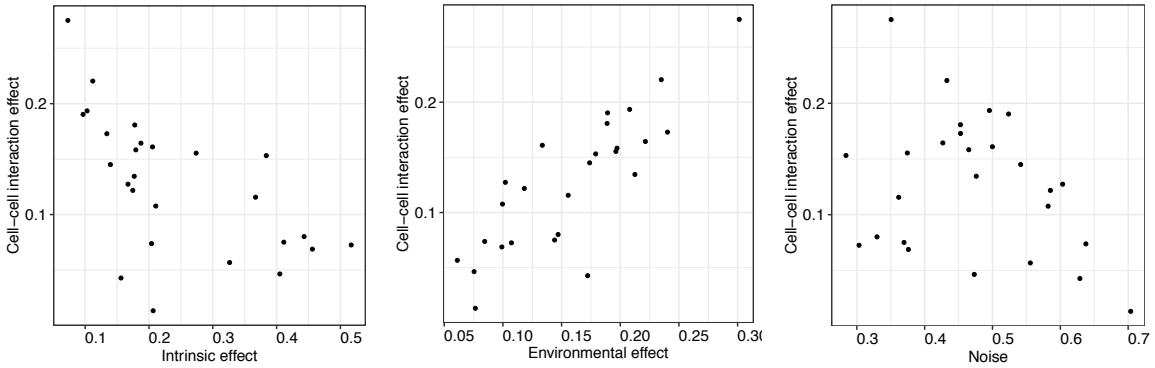


Fig. A.3 Comparison of SVCA cell-cell interaction estimate with the three other model's components in the application to the IMC breast cancer data. Values are proportion of variance explained, averaged across images. One data point on each panel corresponds to one protein.

We did not observe any strong dependency between the variance components, expect a strong correlation between the environmental and the cell-cell interaction term in the application to the IMC data, which can arguably be due to the fact that the two components model spatial

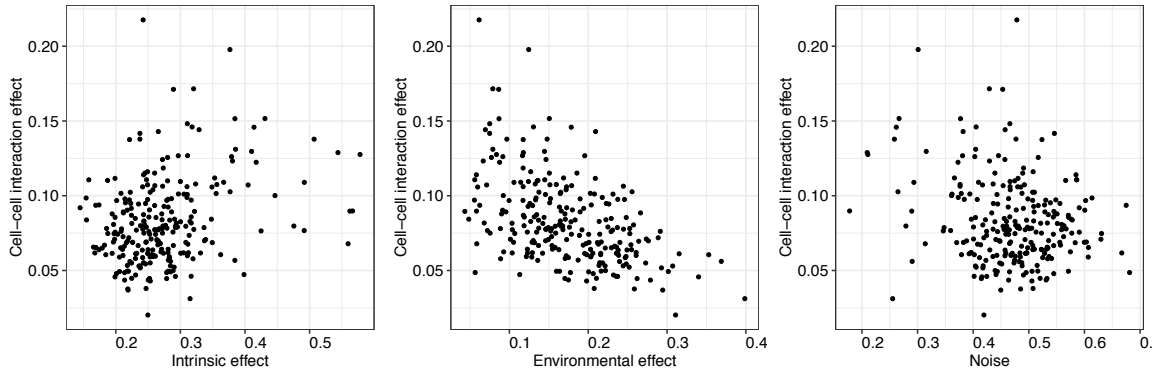


Fig. A.4 Comparison of SVCA cell-cell interaction estimate with the three other model's components in the application to the seqFISH hippocampus data. Values are proportion of variance explained, averaged across images. One data point on each panel corresponds to one protein.

effects with some non-identifiability between them, as discussed in Section A.4. Cell-cell interaction effects could be captured by both terms jointly, and the signal split between them.

A.4 Note on the environmental term

We have seen in section 3.4.3 that the environmental component of SVCA captures spatial effects with no specificity to cell-cell interactions. This does mean however that if cell-cell interactions are not explicitly modelled, the environmental component has the capacity to capture part of it.

To test this hypothesis, we fitted a reduced Gaussian Process with no cell-cell interaction component to the IMC data, and assessed how the residual variance due to cell-cell interactions was captured by the other components of the model (Fig A.5). We found that the environmental term absorbed a major share of this residual variance, which may also explain the small difference in predictive power between this reduced model and the model with explicit cell-cell interactions.

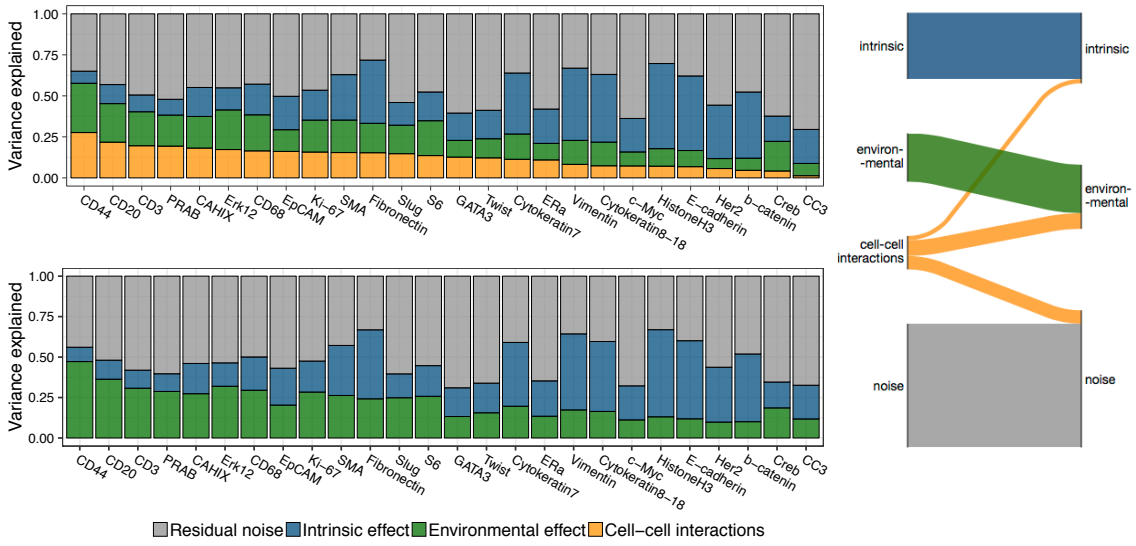


Fig. A.5 Effect of the environmental term on SVCA variance signature when cell-cell interactions are not explicitly modelled. Left: comparison of the variance signatures with and without cell-cell interactions. Right: Visualisation of the capture of non-modelled cell-cell interactions by other terms

A.5 Variability of the variance signatures

A.5.1 Clinical covariates in the IMC application

Figure A.6 shows the first two PCs of the full SVCA variance signatures, as well as the individual variance components considered separately, overlaid with all the breast cancer clinical covariates available. The strongest separation between grades is for the full signatures, followed by the environmental and cell-cell interaction terms. Other clinical covariates did not show a strong relationship with the variance signature terms.

A.5.2 Relationship to gene mean expression and variance for the IMC application

For every gene, we also compared the cell-cell interaction component, in a given image, with the mean expression level of the gene across cells (Fig. A.7), as well as its standard deviation (Fig. A.8). We fond no obvious relationship with the mean, and found that for some genes the cell-cell interaction terms was stronger in images where they had a relatively low variance.

A.5.3 Relationship to gene mean expression and variance for the seqFISH application

For the seqFISH signature, we did the same comparison but pooling all the genes together because of their higher number. No relationship was observed between cell-cell interactions and genes' mean expression or standard deviations (Fig. A.9)

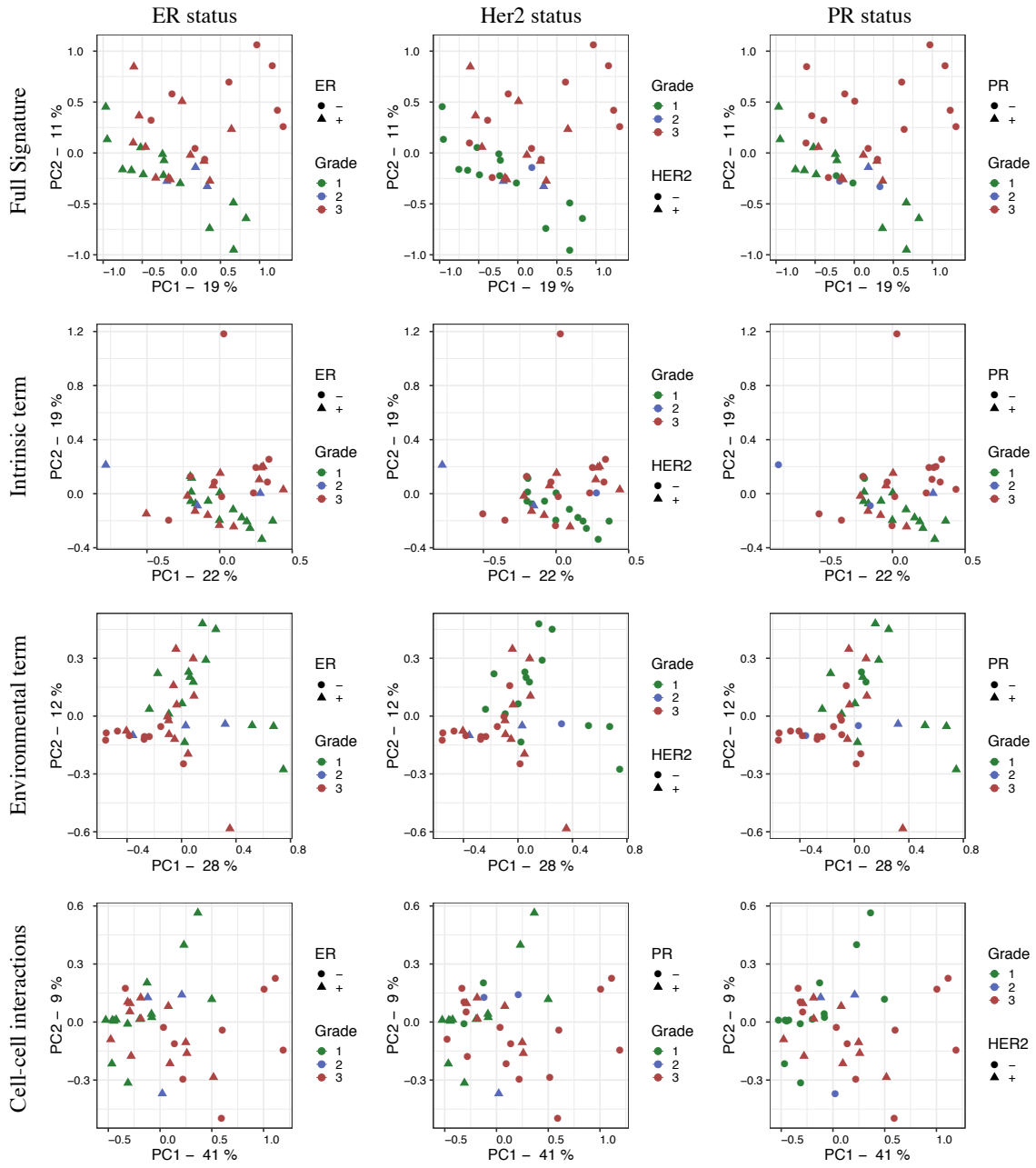


Fig. A.6 PCAs of individual SVCA variance components in IMC and comparison with clinical covariates. Variance components analysed are listed no the rows. Covariates are listed on the columns

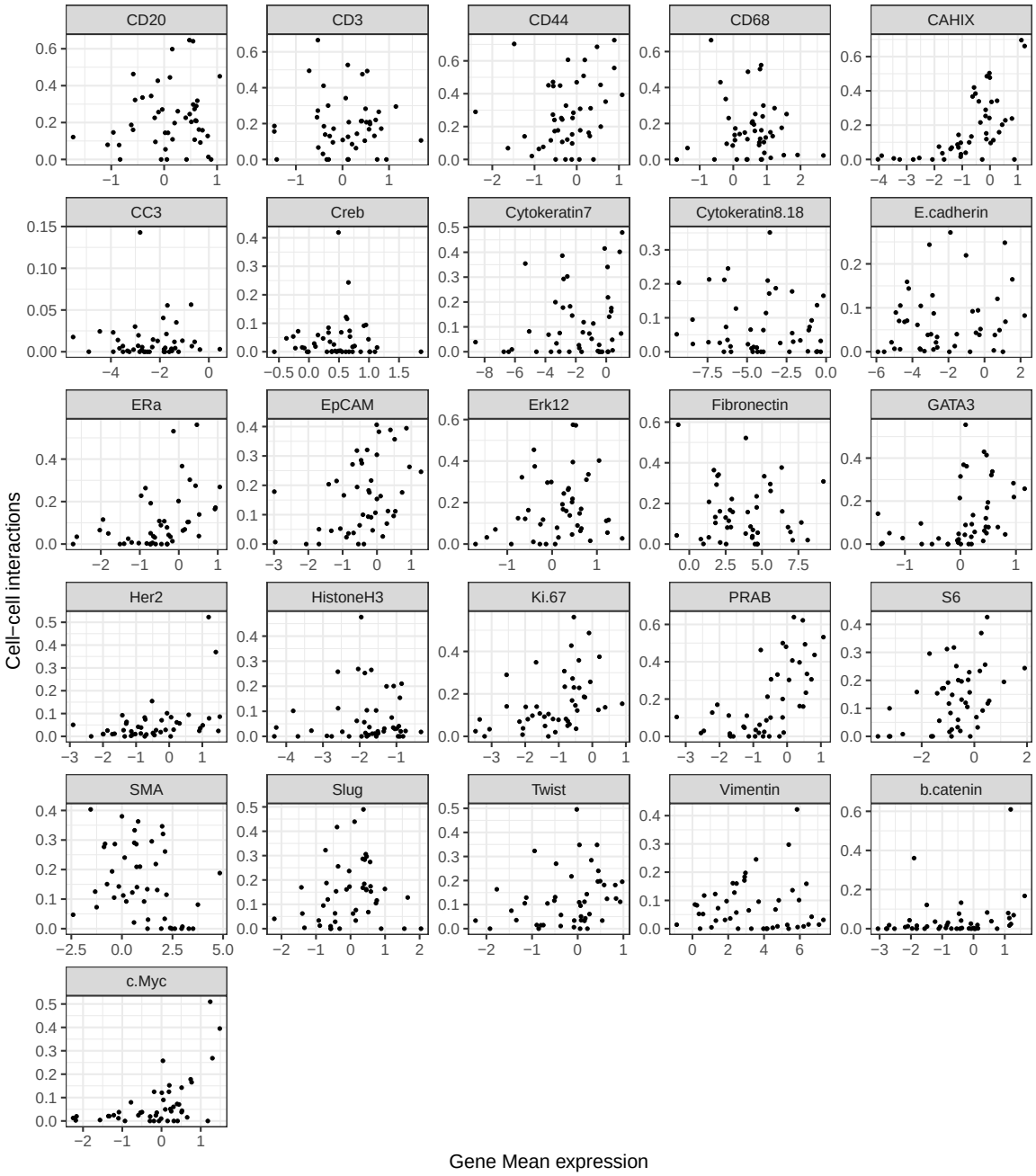


Fig. A.7 Comparison between cell-cell interaction components and mean expression levels in the application to the IMC breast cancer data

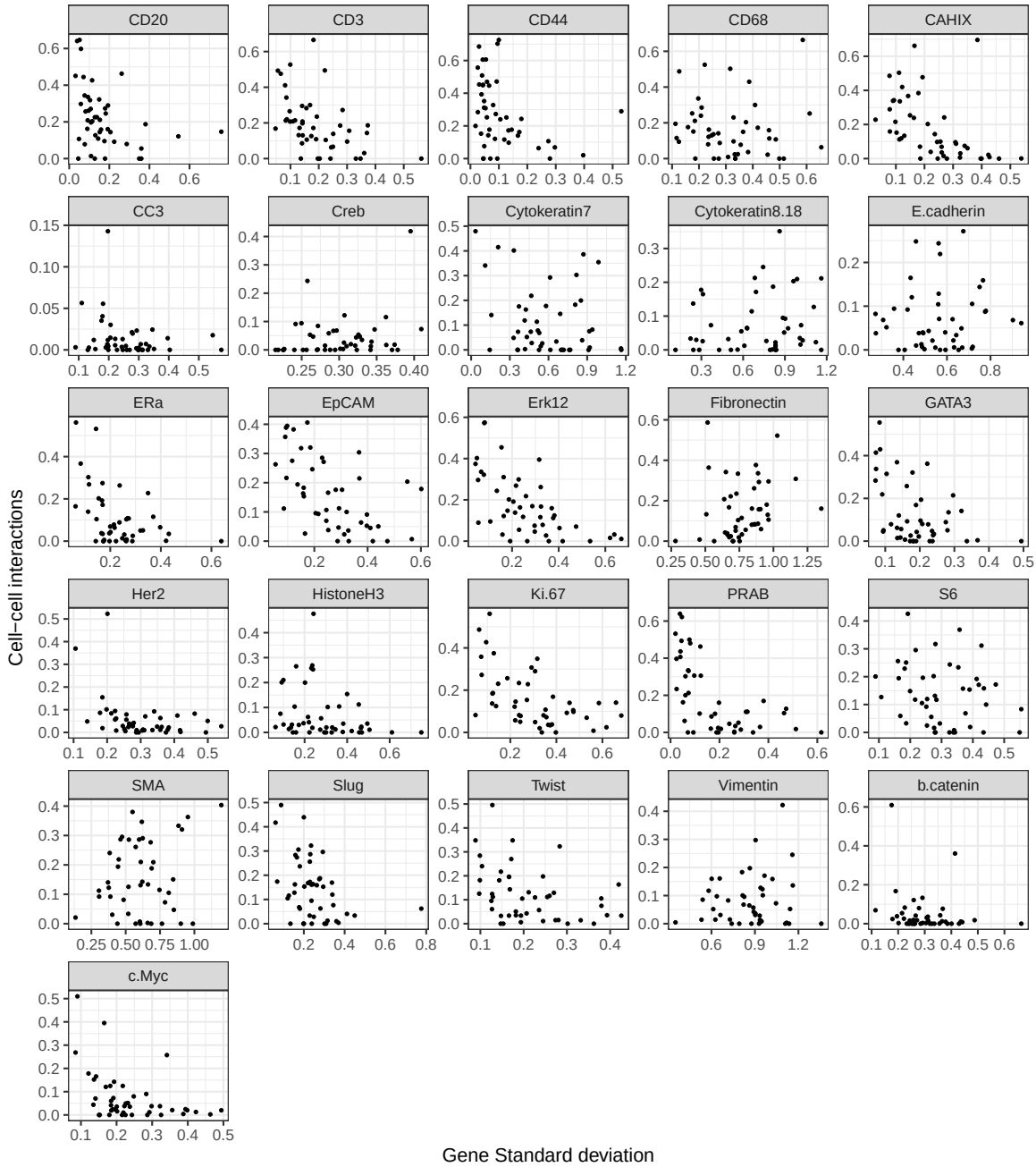


Fig. A.8 Comparison between cell-cell interaction components and the standard deviation of gene expression levels across cells in the application to the IMC breast cancer data

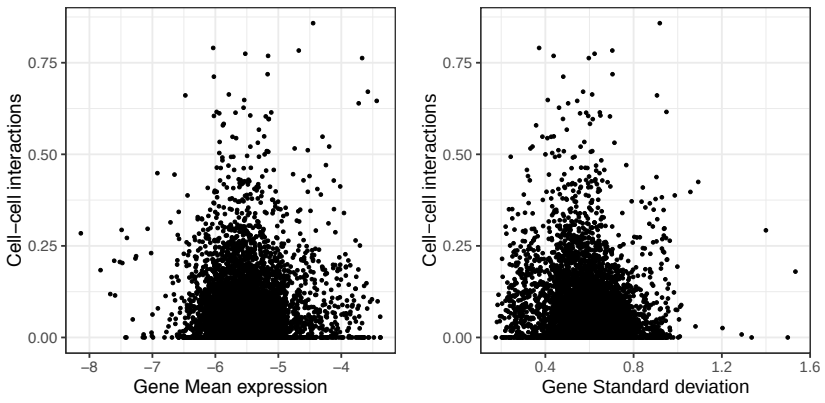


Fig. A.9 Comparison between cell-cell interaction components and mean expression levels (left) and standard deviation of gene expression levels (right) across cells in the application to the seqFISH hippocampus data

A.6 Cell permutation results in seqFISH

To validate the spatial variance signatures obtained in the seqFISH application, we investigated further the results obtained with cell permutations (Section 3.6.2).

First, we focussed our analysis on the top 20 genes, ranked based on their cell-cell interaction estimates obtained with the true cell positions. After permuting the cells, these genes were found to have some residual cell-cell interaction terms (Fig. 3.25). We checked whether these residual variance component had any predictive power out of sample, and found that out of sample predictions were worsened by the use of the spatial covariance terms in SVCA (Fig. A.10).

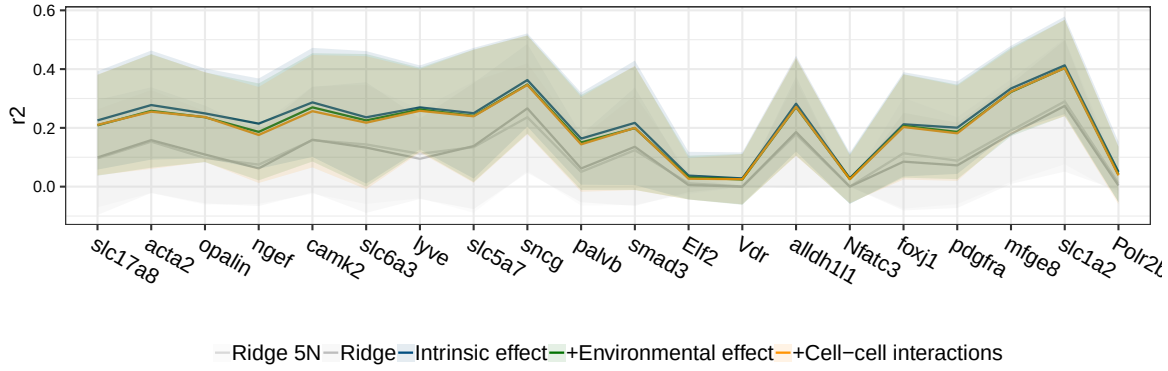


Fig. A.10 Prediction accuracy for SVCA and simpler models using 5-fold cross-validation, with permuted cells. The blue and green lines correspond to two reduced Gaussian Processes including respectively an intrinsic component only, and both an intrinsic and a local component. The two grey lines correspond to simpler linear regressions (see Section 3.4). Results shown for the top 20 cell-cell interaction genes, ranked based on the spatial variance signatures obtained with the true cell positions. The solid lines correspond to the coefficients of determination between predicted gene expression and observed values. The shaded areas correspond to plus and minus one standard deviation across images.

This confirms that the spatial variance components measured on the seqFISH data are indeed dependent on the cell positions.

We then considered the 20 genes showing the highest cell-cell interaction component in the permuted case. We compared the strength of these cell-cell interactions with the top 20 genes from the analysis with true cell positions (Fig A.11) and found a two-fold decrease in the median of these variance components when permuting the cells.

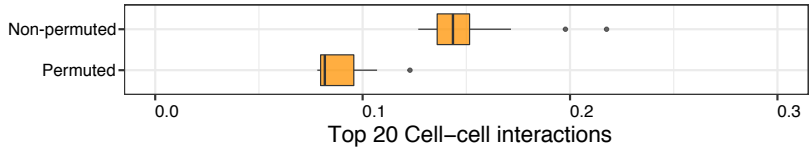


Fig. A.11 Comparison between top 20 cell-cell interactions with and without cell permutations

Looking at out of sample prediction for these 20 genes (Fig. A.12), we also confirmed that the spurious cell-cell interactions detected had no predictive power out of sample, and found that the 20 genes corresponded to very noisy genes.

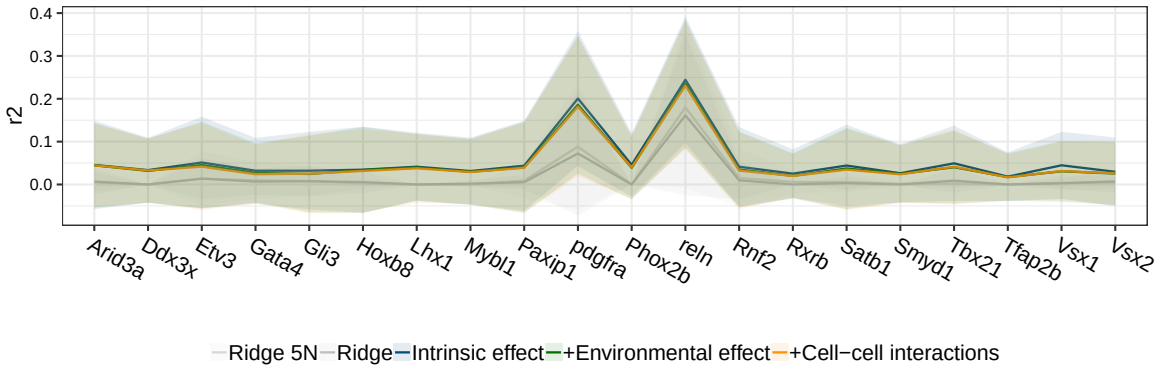


Fig. A.12 Prediction accuracy for SVCA and simpler models using 5-fold cross-validation, with permuted cells. The blue and green lines correspond to two reduced Gaussian Processes including respectively an intrinsic component only, and both an intrinsic and a local component. The two grey lines correspond to simpler linear regressions (see Section 3.4). Results shown for the top 20 cell-cell interaction genes, ranked based on the spatial variance signatures obtained with the permuted cell positions. The solid lines correspond to the coefficients of determination between predicted gene expression and observed values. The shaded areas correspond to plus and minus one standard deviation across images.

Finally, Figure A.13 shows the comparison between i) out of sample predictions using true cells positions for the top 20 genes for cell-cell interactions, ii) out of sample predictions for the same genes, but using permuted cell positions, iii) out of sample predictions for the 20 genes with the highest cell-cell interaction components based on the signatures obtained for the permuted positions. This confirms that spatial components' prediction power indeed relies on the spatial structure of the tissue.

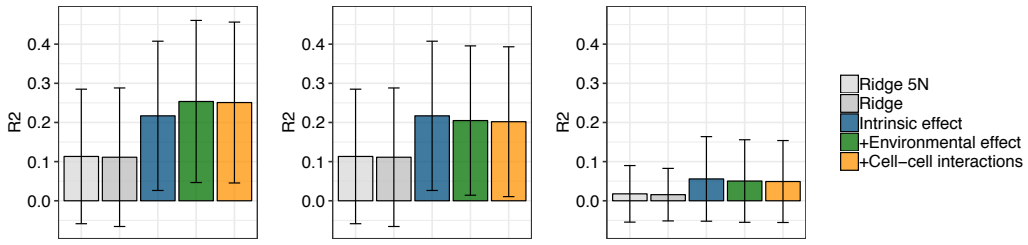


Fig. A.13 Out of sample predictions for (left) true cell positions, top 20 genes for cell-cell interactions; (middle) permuted cell positions, same 20 genes; (right) permuted cell positions, top 20 genes for cell-cell interactions in the permuted case. The ridge models are always fitted on the true cell positions and are shown to give a standard benchmark

A.7 Manual gene annotation for the seqFISH dataset

Gene	Category
ACTA2	Cell Junctions
AHR	Transcription Factor
ALLDH1L1	Metabolism
ALX1	Transcription Factor
AR	Intracellular Signalling
ARID2	Transcription Factor
ARID3A	Transcription Factor
ATM	Cell Cycle
ATR	Cell Cycle
BACH1	Transcription Factor
BACH2	Transcription Factor
BARHL1	Transcription Factor
BHLHE41	Transcription Factor
BLZF1	transporter
CAMK2	kinase
CBFA2T3	Transcription Factor
CDC5L	Cell Cycle
CDC6	Cell Cycle
CEBPG	Transcription Factor
CHAT	neurotransmission
CIITA	Immune System
CLDN5	Cell Junctions
CLOCK	Transcription Factor
CREB1	Transcription Factor
CTNNB1	Cell Junctions
CTSS	Immune System
DDX3X	Transcription Factor
DLX2	Transcription Factor
DMBX1	Transcription Factor
E2F2	Transcription Factor
E2F7	Transcription Factor
EGF	Growth Factor
EHF	Transcription Factor
ELF1	Transcription Factor
ELF2	Transcription Factor
ELF4	Transcription Factor
ELK4	Transcription Factor
EMX2	Transcription Factor
EN1	Development
EN2	Development
EOMES	Transcription Factor
ESR2	Intracellular Signalling
ESRRB	Intracellular Signalling
ESRRG	Intracellular Signalling
ETS1	Transcription Factor
ETV3	Transcription Factor
FBLL1	na
FOXA1	Transcription Factor
FOXB1	Transcription Factor
FOXC1	Transcription Factor
FOXD3	Transcription Factor
FOXD4	Transcription Factor
FOXJ1	Transcription Factor
FOXN1	Transcription Factor
FOXN4	Transcription Factor
FOXO1	Transcription Factor
FOXO4	Transcription Factor
FOXP3	Transcription Factor
GABPA	Transcription Factor
GAD1	neurotransmission
GATA4	Transcription Factor
GATA5	Transcription Factor
GATA6	Transcription Factor
GFI1	Transcription Factor
GJA1	Cell Junctions
GLI1	Transcription Factor
GLI2	Transcription Factor
GLI3	Transcription Factor
GMEB2	Transcription Factor
GRHL1	Transcription Factor
HIC1	Transcription Factor
HLTF	Transcription Factor
HNF1A	Transcription Factor
HOXA1	Transcription Factor
HOXB3	Transcription Factor
HOXB8	Transcription Factor
HOXB9	Transcription Factor
HOXD12	Transcription Factor
HOXD13	Transcription Factor
HTR3A	Intracellular Signalling
IGTP	Immune System
IKZF1	Transcription Factor
IRF2	Immune System

IRX4	Transcription Factor
IRX5	Transcription Factor
KLF1	Transcription Factor
LHX1	Transcription Factor
LHX3	Transcription Factor
LHX5	Transcription Factor
LHX6	Transcription Factor
LMX1A	Transcription Factor
LPP	Cell Junctions
LYVE	Intracellular Signalling
MAFK	Transcription Factor
MAML3	Transcription Factor
MED14	Intracellular Signalling
MFGE8	Immune System
MITF	Transcription Factor
MN1	Transcription Factor
MNAT1	Transcription Factor
MOG	Cell Junctions
MTF2	Transcription Factor
MXD1	Transcription Factor
MYB	Transcription Factor
MYBL1	Transcription Factor
MYBL2	Transcription Factor
MYCN	Transcription Factor
MYL14	na
MZF1	Transcription Factor
NDNF	Cell Junctions
NEUROD4	Transcription Factor
NFATC3	Transcription Factor
NFATC4	Transcription Factor
NFE2L2	Transcription Factor
NFE2L3	Transcription Factor
NFIA	Transcription Factor
NFIL3	Transcription Factor
NFKB2	Transcription Factor
NFKBIZ	Transcription Factor
NFYA	Transcription Factor
NGEF	Exchange Factor
NHLH1	Transcription Factor
NKX3-1	Transcription Factor
NOTCH3	Intracellular Signalling
NPAS3	neurogenesis
NR2E1	Intracellular Signalling
NR2F2	Intracellular Signalling
NR3C2	Intracellular Signalling
NR4A3	Intracellular Signalling
NR5A1	Intracellular Signalling
ONECUT2	Transcription Factor
OPALIN	Cell Junctions
PALVB	na
PAX1	Transcription Factor
PAX2	Transcription Factor
PAX3	Transcription Factor
PAX6	Transcription Factor
PAX7	Transcription Factor
PAX9	Transcription Factor
PAXIP1	Cell Cycle
PBX3	Transcription Factor
PDGFRA	Intracellular Signalling
PGR	Intracellular Signalling
PHOX2B	Transcription Factor
PIAS3	Transcription Factor
PIN1	isomerase
PKNOX1	Transcription Factor
PKNOX2	Transcription Factor
PLAG1	Transcription Factor
PML	Cell Cycle
POLR2B	Polymerase Catalyser
POU3F2	Transcription Factor
POU4F1	Transcription Factor
PPARA	Transcription Factor
PPARGC1B	Transcription Factor
PRDM1	Immune System
PTCH1	Intracellular Signalling
RBAK	Zinc Finger
RBPJ	Transcription Factor
RBPJL	Transcription Factor
RELB	Transcription Factor
RELN	Transcription Factor
REST	Transcription Factor
RFX2	Transcription Factor
RFX4	Transcription Factor
RNF2	Cell Proliferation
RORC	Transcription Factor
RUNX1	Transcription Factor
RUNX3	Transcription Factor

RXRA	Intracellular Signalling
RXRB	Intracellular Signalling
RYBP	Transcription Factor
SALL1	Zinc Finger
SALL3	Zinc Finger
SALL4	Zinc Finger
SATB1	Transcription Factor
SCML2	Transcription Factor
SIN3A	Transcription Factor
SIX4	Transcription Factor
SLC17A7	Neurotransmitter Transporter
SLC17A8	Neurotransmitter Transporter
SLC1A2	Neurotransmitter Transporter
SLC5A7	Neurotransmitter Transporter
SLC6A3	Neurotransmitter Transporter
SLC6A8	Neurotransmitter Transporter
SMAD3	Intracellular Signalling
SMAD5	Intracellular Signalling
SMAD9	Intracellular Signalling
SMARCA4	Transcription Factor
SMYD1	Transcription Factor
SNCG	Axonal Architecture
SOX11	Transcription Factor
SOX13	Transcription Factor
SOX17	Transcription Factor
SOX5	Transcription Factor
SOX6	Transcription Factor
SOX9	Transcription Factor
SP1	Transcription Factor
SP7	Transcription Factor
SP8	Transcription Factor
SREBF1	Transcription Factor
SST	Endocrine System
TAF2	Transcription Factor
TAF4B	Transcription Factor
TAF6L	Transcription Factor
TAL1	Transcription Factor
TBR1	Transcription Factor
TBX15	Transcription Factor
TBX2	Transcription Factor
TBX21	Transcription Factor
TBX4	Transcription Factor
TCF23	Transcription Factor
TFAP2B	Transcription Factor
TFAP2E	Transcription Factor
TFDP2	Transcription Factor
TH	na
TIAM1	Exchange Factor
TMF1	Intracellular Signalling
TOPORS	ligase
TRIM33	Transcription Factor
TRP73	Ion Channels
TRPS1	Ion Channels
TSC2	Growth Factor
TTF1	Transcription Factor
UACA	Apoptosis
UNCX	Transcription Factor
VAV1	Exchange Factor
VDR	Intracellular Signalling
VEZF1	Transcription Factor
VIP	Intracellular Signalling
VSX1	Development
VSX2	Development
WT1	Transcription Factor
WWTR1	Transcription Factor
XDH	Metabolism
ZBTB33	Transcription Factor
ZFP128	Transcription Factor
ZFP263	Transcription Factor
ZFP287	Transcription Factor
ZFP354A	Transcription Factor
ZFP422	Transcription Factor
ZFP423	Transcription Factor
ZFP64	Transcription Factor
ZIC2	Transcription Factor
ZIC3	Transcription Factor
ZIC4	Transcription Factor
ZIC5	Transcription Factor
ZKSCAN17	Transcription Factor
ZSCAN21	Transcription Factor

Appendix B

Variational Inference beyond the mean field approximation

In this Appendix, we discuss the case of variational inference with partially factorised approximations to the posterior distribution of the parameters: $q(\Theta) = q(\theta_k|\theta_l)q(\theta_l)\prod_{i\notin\{k,l\}} q(\theta_i)$ where $q(\theta_k|\theta_l) \neq q(\theta_k)$.

Y denotes the observed data, Θ denotes the parameters (unobserved random variables). For simpler notations, we will write $\Theta' = \{\theta_i\}_{\forall i\notin\{l,k\}}$, so that $q(\Theta) = q(\theta_k|\theta_l)q(\theta_l)q(\Theta')$.

We have shown in 2.2.6 the relation of equation B.1.

$$\ln P(Y) = \mathscr{L} + \mathrm{KL}(q(\Theta)||P(\Theta|Y)) \tag{B.1}$$

with,

$$\begin{aligned} \mathscr{L} &= \int_\Theta q(\Theta)\ln\left[\frac{p(Y,\Theta)}{q(\Theta)}\right]\mathrm{d}\Theta \\ \mathrm{KL}(q||p) &= -\int_\Theta q(\Theta)\ln\left[\frac{p(\Theta|Y)}{q(\Theta)}\right]\mathrm{d}\Theta \end{aligned} \tag{B.2}$$

The aim is to find the variational distribution $q(\Theta)$ which minimises $\mathrm{KL}(q(\Theta)||p(\Theta|Y))$, or equivalently which maximises $\mathscr{L}(q)$. For $q(\theta_{i\notin\{l,k\}})$, the derivations of 2.2.6 apply with no differences. We here derive the update rules for the conditional distributions $q(\theta_k|\theta_l)$ and for

$q(\theta_l)$

B.1 Optimisation of $q(\theta_l)$

The ELBO $\mathscr{L}$ can be rewritten as on equation B.3.

$$
\begin{aligned}
\mathscr{L}(q) &= \int_{\Theta} q(\Theta')q(\theta_k|\theta_l)q(\theta_l)\ln\left[\frac{p(Y,\Theta)}{q(\Theta')q(\theta_k|\theta_l)q(\theta_l)}\right]\mathrm{d}\Theta \\
&= \int_{\theta_l} q(\theta_l)\left[\int_{\Theta',\theta_k} q(\Theta')q(\theta_k|\theta_l)\ln\left[\frac{p(Y,\Theta)}{q(\Theta')q(\theta_k|\theta_l)q(\theta_l)}\right]\mathrm{d}\Theta'\mathrm{d}\theta_k\right]\mathrm{d}\theta_l \\
&= \int_{\theta_l} q(\theta_l)\left[\int_{\Theta',\theta_k} q(\Theta')q(\theta_k|\theta_l)\ln\left[\frac{p(Y,\Theta)}{q(\theta_k|\theta_l)}\right]\mathrm{d}\Theta'\mathrm{d}\theta_k\right]\mathrm{d}\theta_l \\
&\quad - \int_{\theta_l} q(\theta_l)\left[\int_{\Theta',\theta_k} q(\Theta')q(\theta_k|\theta_l)\ln q(\theta_l)\mathrm{d}\Theta'\mathrm{d}\theta_k\right]\mathrm{d}\theta_l \\
&\quad - \int_{\theta_l} q(\theta_l)\left[\int_{\Theta',\theta_k} q(\Theta')q(\theta_k|\theta_l)\ln q(\Theta')\mathrm{d}\Theta'\mathrm{d}\theta_k\right]\mathrm{d}\theta_l \\
&= \int_{\theta_l} q(\theta_l)\mathbb{E}_{\Theta',\theta_k|\theta_l}\left[\ln\frac{p(Y,\Theta)}{q(\theta_k|\theta_l)}\right]\mathrm{d}\theta_l - \int_{\theta_l} q(\theta_l)\ln q(\theta_l)\mathrm{d}\theta_l - \int_{\Theta'} q(\Theta')\ln q(\Theta')\mathrm{d}\Theta'
\end{aligned}
\tag{B.3}
$$

We write $\tilde{P}(Y,\Theta)$ the distribution such that:

$$
\ln\tilde{P}(Y,\theta) = \mathbb{E}_{\Theta',\theta_k|\theta_l}\left[\ln\frac{p(Y,\Theta)}{q(\theta_k|\theta_l)}\right] \tag{B.4}
$$

We then recognise that $\mathscr{L}(q) = -\mathrm{KL}\left(q(\theta_l)||\tilde{P}(Y,\theta)\right) + cst$ and deduce that $\mathscr{L}(q)$ is maximised when:

$$
\ln q(\theta_l) = \mathbb{E}_{\Theta',\theta_k|\theta_l}\left[\ln\frac{p(Y,\Theta)}{q(\theta_k|\theta_l)}\right] \tag{B.5}
$$

We recognise a slightly altered version of the update rules of 2.2.6, which additionally involves an expectation of $\ln q(\theta_k|\theta_l)$. The iterative algorithm used in variational inference can be adapted using equation B.5

B.2 Optimisation of the conditional distribution $q(\theta_k|\theta_l)$

As $q(\theta_k|\theta_l)$ is conditioned on θ_l, the update rule for $q(\theta_k|\theta_l)$ can be obtained easily by fixing θ_l in Equation B.2. The derivations of 2.2.6 remain then unchanged and the update rule is the same as in 2.2.6 but with a conditional expectation on θ_l, as on equation B.6

$$\ln q(\theta_k|\theta_l) = \mathbb{E}_{\Theta'|\theta_l}[\ln p(Y,\Theta)] \tag{B.6}$$

Appendix C

Variational inference with non-Gaussian likelihoods

Bayesian models such as Factor Analysis in conjunction with non-Gaussian likelihoods are useful when dealing with different data modalities, such as binary values in the case of genotyping data or counts in the case of RNA-seq data. However, this poses an additional challenge in the inference of the posterior distribution of the parameters, which is that the prior distributions of the parameters of the model are no longer conjugate.

In variational inference, one way to address this challenge is to find a lower bound to the model log likelihood which is quadratic in the model parameters. The quadratic form enables to approximate the model log likelihood with a Gaussian, and the log likelihood lower bound can then be re-injected in the standard evidence lower bound of the model without affecting the inequality.

C.1 Approach from Seeger and Bouchard (2012)

If the second derivative of the model log likelihood is lower bounded, a natural quadratic lower bound of the log likelihood is its second order Taylor expansion. This is the approach used by Seeger and Bouchard (2012) and it applies to Poisson, Bernoulli and Binomial data.

Recall that in the Bayesian framework, we approximate the true posterior distribution of the parameters, $P(\Theta|Y)$, with a variational distribution $q(\Theta)$ which is typically factorised

over some or all of the parameters. Computing the best approximation $q(\Theta)$ is equivalent to maximising the evidence lower bound $\mathscr{L}$ of Equation C.1.

$$\begin{aligned}\mathscr{L} &= \int_{\Theta} q(\Theta) \ln \frac{P(Y,\Theta)}{q(\Theta)} \mathrm{d}\Theta \\ &= \int_{\Theta} q(\Theta) \ln P(Y|\Theta)\Theta + \mathrm{KL}\left[q(\Theta)||p(\Theta)\right]\end{aligned} \tag{C.1}$$

We now assume a log likelihood which can be written as in Equation C.2

$$\ln P(Y|\Theta) = \sum_{n=1}^{N}\sum_{d=1}^{D} f_{n,d}(c_{n,d}) \tag{C.2}$$

In the case of Factor Analysis, $c_{n,d} = \sum_k z_{n,k} w_{k,d}$ and the specific form of $f_{n,d}(c_{n,d}) = \ln P(y_{n,d}|c_{n,d})$ depends on the likelihood used for the data. The sum over n and d comes from the fact that observations are treated as independent.

If $f_{n,d}$ is twice differentiable and $f''_{n,d} \geq \kappa_d$, we can write the lower bound of Equation C.3, which comes from the second order Taylor expansion in $\xi_{n,d}$ of the individual terms of the model log likelihood for every observation.

$$f_{n,d}(c_{n,d}) \geq f_{n,d}(\xi_{n,d}) + f'_{n,d}(\xi_{n,d})(c_{n,d} - \xi_{n,d}) + \frac{\kappa_d}{2}(c_{n,d} - \xi_{n,d})^2 := q_{n,d}(c_{n,d}, \xi_{n,d}) \tag{C.3}$$

Therefore, the optimisation problem of maximising the ELBO $\mathscr{L}$ of Equation C.1, may be approximated by the maximisation of the new Evidence Lower Bound of Equation C.4.

$$\begin{aligned}\mathscr{L}_2 &= \sum_{n,d} \mathop{\mathbb{E}}_{q}\left(q_{n,d}(c_{n,d}, \xi_{n,d})\right) + \mathrm{KL}\left[q(\Theta)||p(\Theta)\right] \\ &= \int_{\Theta} q(\Theta) \sum_{n,d} \ln \tilde{P}(\tilde{y}_{n,d}|\xi_{n,d}, c_{n,d}, -\kappa_d)\mathrm{d}\Theta + \mathrm{KL}\left[q(\Theta)||p(\Theta)\right]\end{aligned} \tag{C.4}$$

where $\tilde{P}(\tilde{Y}_{n,d}|\xi_{n,d}, c_{n,d}, -\kappa_d)$ corresponds to a Normal distribution with pseudo-data $\tilde{Y}_{n,d} = \xi_{n,d} - f'(\xi_{n,d})/\kappa_d$, mean $c_{n,d}$ and precision $-\kappa_d$, which is obtained by completing the square in Equation C.3.

We are now reduced to the Gaussian case where usual variational update rules may be used with the pseudo data defined before. However, it also becomes necessary to maximise the

lower bound defined above with respect to the location of the Taylor expansions $\xi_{n,d}$ for all n and d. Taking the first derivative of $q_{n,d}(c_{n,d}, \xi_{n,d})$ with respect to $\xi_{n,d}$, we find that $\mathscr{L}_2$ is maximised when $\xi_{n,d} = \mathbb{E}_q(c_{n,d})$.

C.1.1 Poisson likelihood

A standard way of modelling count data is to use a Poisson likelihood as in Equation C.5

$$P(y_{n,d}|c_{n,d}) = \lambda(c_{n,d})^{y_{n,d}} \exp\left(-\lambda(c_{n,d})\right) \tag{C.5}$$

where $\lambda(c_{n,d})$ is a log-concave rate function such as $\lambda(c_{n,d}) = \ln\left(1+\exp(c_{n,d})\right)$.
It can be shown that in that case:

$$f''_{n,d}(c_{n,d}) \geq \kappa_d = -\left(1/4 + 0.17 * \max(Y_{:,d})\right) \tag{C.6}$$

which enables to use the method described above for inference.

C.1.2 Bernoulli likelihood

A standard way to model binary data in Factor Analysis models is to use the Bernoulli likelihood of Equation C.7

$$\begin{aligned} P(y_{n,d}|c_{n,d}) &= \text{Ber}\left(y_{n,d}|\sigma(c_{n,d})\right) \\ &= \frac{1}{1+\exp\left[-(2y_{n,d}-1)c_{n,d}\right]} \end{aligned} \tag{C.7}$$

It can then be shown that

$$f''_{n,d}(c_{n,d}) \geq \kappa_d = -1/4 \tag{C.8}$$

C.2 Bernoulli case with the approach from Jaakkola and Jordan (2000)

For the Bernoulli case, it is possible to derive a tighter lower bound to the log likelihood. Let's rewrite the Bernoulli likelihood as in Equation C.9

$$\begin{aligned} f_{n,d}(c_{n,d}) &= -\ln\left[1+\exp-(2y_{n,d}-1)c_{n,d}\right] \\ &= \frac{(2y_{n,d}-1)c_{n,d}}{2} - \ln\left[\exp\frac{(2y_{n,d}-1)c_{n,d}}{2} + \exp\frac{-(2y_{n,d}-1)c_{n,d}}{2}\right] \\ &= x/2+g(x) \text{ with } x=(2y_{n,d}-1)c_{n,d} \text{ and } g(x) = -\ln\left[\exp\frac{x}{2}+\exp\frac{-x}{2}\right] \end{aligned} \tag{C.9}$$

It can be shown that the function g is convex in x^2, which is used by Jaakkola and Jordan (2000) in order to write the lower bound of Equation C.10 which comes from the first order Taylor expansion of g with respect to x^2.

$$\begin{aligned} f_{n,d}(x) &\geq x/2+g(\xi)+\frac{\mathrm{d}g}{\mathrm{d}x^2}(\xi)(x^2-\xi^2) \\ &= x/2-\xi/2-\ln(1+\exp-\xi)-\frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right)(x^2-\xi^2) \\ &= \frac{(2y_{n,d}-1)c_{n,d}-\xi_{n,d}}{2} - \ln(1+\exp-\xi_{n,d}) - \frac{1}{4\xi_{n,d}}\tanh\left(\frac{\xi_{n,d}}{2}\right)(c_{n,d}^2-\xi_{n,d}^2) \end{aligned} \tag{C.10}$$

By completing the square in Equation C.10 and re-injecting the likelihood lower bound into the evidence lower bound of the model, we observe that maximising the ELBO $\mathscr{L}$ of Equation C.1 may be reduced to the simpler problem of maximising the new ELBO of Equation C.11.

$$\mathscr{L}_2 = \int_\Theta q(\Theta)\sum_{n,d}\ln\tilde{P}(\tilde{y}_{n,d}|\xi_{n,d},c_{n,d},\tau_{n,d})\mathrm{d}\Theta + \mathrm{KL}\left[q(\Theta)||p(\Theta)\right] \tag{C.11}$$

Where $\tilde{P}$ defines a Normal distribution with pseudo data $\tilde{y}_{n,d} = (2y_{n,d}-1)\xi_{n,d}/\tanh(\xi_{\mathrm{n,d}}/2)$ mean $c_{n,d}$ and precision $\tau_{n,d} = \tanh(\xi_{n,d}/2)/2\xi_{n,d}$

Like in the approach by Seeger et al., we are reduced to the Gaussian case where usual variational update rules can be used with the pseudo data defined above. Like in the Seeger et al. approach, we also maximise $\mathscr{L}_2$ with respect to the location of the Taylor expansion and find that the update for this additional parameter is given by $\xi_{n,d}^2 = \mathbb{E}(c_{n,d}^2)$.

The lower bound provided by the Jaakkola method for the Bernoulli likelihood is tighter than the one provided by the Seeger method, as illustrated in Figure 4.4. This is because in the

Jaakkola method, the precision of the approximate Gaussian $\tilde{P}$ is dependent on the location of the Taylor expension $\xi_{n,d}$ (heteroscedasticity)

Appendix D

BIOFAM variational updates

In this section, we give the analytical solution for the approximate posterior distribution of BIOFAM parameters in the variational inference framework. Additionally, we write down the analytical form of the Evidence Lower Bound (ELBO) of Equation 2.27. Although computing the ELBO is not necessary in order to estimate the posterior distribution of the parameters, it is used to monitor the convergence of the iterative algorithm.

D.1 Variational Updates

D.1.1 Latent variables

Variational distribution:

$$q(z^m_{n,k}) = \mathcal{N}\left(z_{nk} | \mu_{z_{nk}}, \sigma_{z_{nk}}\right) \tag{D.1}$$

where

$$\begin{aligned} \sigma^2_{z_{nk}} &= \left(\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle \tau^m_d \rangle \left\langle (s^m_{dk} \hat{w}^m_{dk})^2 \right\rangle + 1 \right)^{-1} \\ \mu_{z_{nk}} &= \sigma^2_{z_{nk}} \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle \tau^m_d \rangle \langle s^m_{dk} \hat{w}^m_{dk} \rangle \left(y^m_{nd} - \sum_{j \neq k} \left\langle s^m_{dj} \hat{w}^m_{dj} \right\rangle \langle z_{nj} \rangle \right) \end{aligned} \tag{D.2}$$

D.1.2 Spike-and-slab weights

Update for $q(s^m_{dk})$:

$$q(s^m_{d,k}) = \text{Ber}(s^m_{d,k}|\gamma^m_{d,k}) \tag{D.3}$$

with

$$\begin{aligned}
\gamma^m_{d,k} &= \frac{1}{1+\exp(-\lambda^m_{d,k})} \\
\lambda^m_{dk} &= \left\langle \ln \frac{\theta}{1-\theta} \right\rangle + 0.5 \ln \frac{\left\langle \alpha^m_k \right\rangle}{\left\langle \tau^m_d \right\rangle} - 0.5 \ln \left(\sum_{n=1}^{N} \left\langle z^2_{nk} \right\rangle + \frac{\left\langle \alpha^m_k \right\rangle}{\left\langle \tau^m_d \right\rangle} \right) \\
&+ \frac{\left\langle \tau^m_d \right\rangle}{2} \frac{\left(\sum_{n=1}^{N} y^m_{nd} \left\langle z_{nk} \right\rangle - \sum_{j \neq k} \left\langle s^m_{dj} \hat{w}^m_{dj} \right\rangle \sum_{n=1}^{N} \left\langle z_{nk} \right\rangle \left\langle z_{nj} \right\rangle \right)^2}{\sum_{n=1}^{N} \left\langle z^2_{nk} \right\rangle + \frac{\left\langle \alpha^m_k \right\rangle}{\left\langle \tau^m_d \right\rangle}}
\end{aligned} \tag{D.4}$$

Update for $q(\hat{w}^m_{dk})$:

$$\begin{aligned}
q(\hat{w}^m_{dk}|s^m_{dk} = 0) &= \mathcal{N}\left(\hat{w}^m_{dk}|0, 1/\alpha^m_k\right) \\
q(\hat{w}^m_{dk}|s^m_{dk} = 1) &= \mathcal{N}\left(\hat{w}^m_{dk} \middle| \mu_{w^m_{dk}}, \sigma^2_{w^m_{dk}}\right)
\end{aligned} \tag{D.5}$$

with

$$\begin{aligned}
\mu_{w^m_{dk}} &= \frac{\sum_{n=1}^{N} y^m_{nd} \left\langle z_{nk} \right\rangle - \sum_{j \neq k} \left\langle s^m_{dj} \hat{w}^m_{dj} \right\rangle \sum_{n=1}^{N} \left\langle z_{nk} \right\rangle \left\langle z_{nj} \right\rangle}{\sum_{n=1}^{N} \left\langle z^2_{nk} \right\rangle + \frac{\left\langle \alpha^m_k \right\rangle}{\left\langle \tau^m_d \right\rangle}} \\
\sigma_{w^m_{dk}} &= \frac{\left\langle \tau^m_d \right\rangle^{-1}}{\sum_{n=1}^{N} \left\langle z^2_{nk} \right\rangle + \frac{\left\langle \alpha^m_k \right\rangle}{\left\langle \tau^m_d \right\rangle}}
\end{aligned} \tag{D.6}$$

D.1.3 ARD precision (alpha)

Feature-wise

Variational distribution:

$$q(\alpha^m_k) = \Gamma\left(\alpha^m_k \middle| \hat{a}^{\alpha}_{m,k}, \hat{b}^{\alpha}_{m,k}\right) \tag{D.7}$$

where

$$\begin{aligned}\hat{a}^{\alpha}_{m,k} &= a^{\alpha}_{0} + \frac{D_m}{2} \\ \hat{b}^{\alpha}_{m,k} &= b^{\alpha}_{0} + \frac{\sum_{d=1}^{D_m} \left\langle (\hat{w}^{m}_{d,k})^2 \right\rangle}{2}\end{aligned} \tag{D.8}$$

Sample-wise

Variational distribution:

$$q(\alpha^{g}_{k}) = \Gamma\left(\alpha^{g}_{k} \middle| \hat{a}^{\alpha}_{g,k}, \hat{b}^{\alpha}_{g,k}\right) \tag{D.9}$$

where

$$\begin{aligned}\hat{a}^{\alpha}_{g,k} &= a^{\alpha}_{0} + \frac{N_g}{2} \\ \hat{b}^{\alpha}_{g,k} &= b^{\alpha}_{0} + \frac{\sum_{n=1}^{N_g} \left\langle (z^{g}_{n,k})^2 \right\rangle}{2}\end{aligned} \tag{D.10}$$

D.1.4 Noise precision (tau)

Variational distribution:

$$q(\tau^{m}_{d}) = \Gamma\left(\tau^{m}_{d} \middle| \hat{a}^{\tau}_{m,d}, \hat{b}^{\tau}_{m,d}\right) \tag{D.11}$$

where

$$\begin{aligned}\hat{a}^{\tau}_{md} &= a^{\tau}_{0} + \frac{N}{2} \\ \hat{b}^{\tau}_{md} &= b^{\tau}_{0} + \frac{1}{2}\sum_{n=1}^{N} \left\langle \left(y^{m}_{nd} - \sum_{k}^{K} \hat{w}^{m}_{dk} s^{m}_{dk} z_{n,k} \right)^2 \right\rangle\end{aligned} \tag{D.12}$$

D.1.5 Spike-and-slab sparsity parameter (theta)

Variational distribution:

$$q(\theta_k^m) = \beta\left(\theta_k^m \middle| \hat{a}_{mk}^\theta, \hat{b}_{mk}^\theta\right) \tag{D.13}$$

where

$$\begin{aligned} \hat{a}_{mk}^\theta &= \sum_{d=1}^{D_m} \langle s_{dk}^m \rangle + a_0^\theta \\ \hat{b}_{mk}^\theta &= b_0^\theta - \sum_{d=1}^{D_m} \langle s_{dk}^m \rangle + D_m \end{aligned} \tag{D.14}$$

D.2 Evidence Lower Bound

As seen in 2.2.6, the ELBO can be decomposed into a contribution coming from the data likelihood under the current estimate of the posterior distribution of the parameters and a contribution accounting for the KL divergence between the prior and the posterior distributions of the parameters:

$$\mathscr{L} = \mathop{\mathbb{E}}_{q(\theta)} \ln P(Y|\Theta) - \mathrm{KL}\left(q(\Theta)||p(\Theta)\right) \tag{D.15}$$

D.2.1 Contribution from the data likelihood (Gaussian case)

$$\begin{aligned} \mathop{\mathbb{E}}_{q(\theta)} \ln P(Y|\Theta) = &- \sum_{m=1}^{M} \frac{ND_m}{2} \ln(2\pi) + \frac{N}{2} \sum_{m=1}^{M} \sum_{d=1}^{D_m} \ln(\langle \tau_d^m \rangle) \\ &- \sum_{m=1}^{M} \sum_{d=1}^{D_m} \frac{\langle \tau_d^m \rangle}{2} \sum_{n=1}^{N} \left(y_{nd}^m - \sum_{k=1}^{K} \langle s_{dk}^m \hat{w}_{dk}^m \rangle \langle z_{nk} \rangle\right)^2 \end{aligned} \tag{D.16}$$

D.2.2 Contribution from the KL divergence regulariser

Note that $\mathrm{KL}\left(q(\Theta)||P(\Theta)\right) = \mathbb{E}_q(q(\Theta)) - \mathbb{E}_q(P(\Theta))$. Below, we will write the analytical form for these two expectations.

W and S terms

$$\begin{aligned}\mathop{\mathbb{E}}_{q}[\ln p(\hat{W},S)] = &-\sum_{m=1}^{M}\frac{KD_m}{2}\ln(2\pi)+\sum_{m=1}^{M}\frac{D_m}{2}\sum_{k=1}^{K}\ln(\alpha_k^m)-\sum_{m=1}^{M}\frac{\alpha_k^m}{2}\sum_{d=1}^{D_m}\sum_{k=1}^{K}\left\langle(\hat{w}_{dk}^m)^2\right\rangle \\ &+\langle\ln(\theta)\rangle\sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K}\langle s_{dk}^m\rangle+\langle\ln(1-\theta)\rangle\sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K}(1-\langle s_{dk}^m\rangle)\end{aligned} \tag{D.17}$$

$$\begin{aligned}\mathop{\mathbb{E}}_{q}[\ln q(\hat{W},S)] = &-\sum_{m=1}^{M}\frac{KD_m}{2}\ln(2\pi)+\frac{1}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K}\ln(\langle s_{dk}^m\rangle\,\sigma_{w_{dk}^m}^2+(1-\langle s_{dk}^m\rangle)/\alpha_k^m) \\ &+\sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K}(1-\langle s_{dk}^m\rangle)\ln(1-\langle s_{dk}^m\rangle)-\langle s_{dk}^m\rangle\ln\langle s_{dk}^m\rangle\end{aligned} \tag{D.18}$$

Z term

$$\begin{aligned}\mathop{\mathbb{E}}_{q}[\ln p(Z)] &= -\frac{NK}{2}\ln(2\pi)-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\left\langle z_{nk}^2\right\rangle \\ \mathop{\mathbb{E}}_{q}[\ln q(Z)] &= -\frac{NK}{2}(1+\ln(2\pi))-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\ln(\sigma_{z_{nk}}^2)\end{aligned} \tag{D.19}$$

ARD terms (applies to both sample and feature-wise ARD)

$$\begin{aligned}\mathop{\mathbb{E}}_{q}[\ln p(\alpha)] &= \sum_{m=1}^{M}\sum_{k=1}^{K}\left(a_0^\alpha\ln b_0^\alpha+(a_0^\alpha-1)\langle\ln\alpha_k\rangle-b_0^\alpha\langle\alpha_k\rangle-\ln\Gamma(a_0^\alpha)\right) \\ \mathop{\mathbb{E}}_{q}[\ln q(\alpha)] &= \sum_{m=1}^{M}\sum_{k=1}^{K}\left(\hat{a}_k^\alpha\ln\hat{b}_k^\alpha+(\hat{a}_k^\alpha-1)\langle\ln\alpha_k\rangle-\hat{b}_k^\alpha\langle\alpha_k\rangle-\ln\Gamma(\hat{a}_k^\alpha)\right)\end{aligned} \tag{D.20}$$

Tau term

$$\begin{aligned}\mathop{\mathbb{E}}_{q}[\ln p(\tau)] &= \sum_{m=1}^{M}D_m a_0^\tau\ln b_0^\tau+\sum_{m=1}^{M}\sum_{d=1}^{Dm}(a_0^\tau-1)\langle\ln\tau_d^m\rangle-\sum_{m=1}^{M}\sum_{d=1}^{Dm}b_0^\tau\langle\tau_d^m\rangle-\sum_{m=1}^{M}D_m\ln\Gamma(a_0^\tau) \\ \mathop{\mathbb{E}}_{q}[\ln q(\tau)] &= \sum_{m=1}^{M}\sum_{d=1}^{D_m}\left(\hat{a}_{dm}^\tau\ln\hat{b}_{dm}^\tau+(\hat{a}_{dm}^\tau-1)\langle\ln\tau_d^m\rangle-\hat{b}_{dm}^\tau\langle\tau_d^m\rangle-\ln\Gamma(\hat{a}_{dm}^\tau)\right)\end{aligned} \tag{D.21}$$

theta term

$$\mathop{\mathbb{E}}_{q}[\ln p(\theta)] = \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{d=1}^{D_m}\left((a_0-1)\times\left\langle\ln(\pi_{d,k}^m)\right\rangle + (b_0-1)\left\langle\ln(1-\pi_{d,k}^m)\right\rangle - \ln(\mathrm{B}(a_0,b_0))\right)$$

$$\mathop{\mathbb{E}}_{q}[\ln q(\theta)] = \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{d=1}^{D_m}\left((a_{k,d}^m-1)\times\left\langle\ln(\pi_{d,k}^m)\right\rangle + (b_{k,d}^m-1)\left\langle\ln(1-\pi_{d,k}^m)\right\rangle - \ln(\mathrm{B}(a_{k,d}^m,b_{k,d}^m))\right) \tag{D.22}$$

Appendix E

Supplementary Analysis of the BIOFAM software

E.1 Identifiability of the latent structure for sparse and dense factors

In this section, we analyse further the question of the identifiability of the latent structure raised in Section 4.4.2, in scenarios where the latent structure is made of factors with dense effects only, or sparse effects only. We show that dense latent structures are mostly non-identifiable by any of the biofam models. In the case of latent factors with sparse effects, the use of spike-and-slab priors on the weights renders factors readily identifiable. Spike-and-slab priors on the factors provides a marginal identifiability improvement, even when the factors are simulated as dense with only sparse weights.

E.1.1 Simulations with sparse factors only

We generated data using 30 latent variables with 10 % of active weights for each one of them. Data was simulated for 800 samples and 1600 features and the latent variables explained 25% of the total variability.

We compared three models. The first model had no spike-and-slab prior, the second model had spike-and-slab priors on the weights and the third model had both spike-and-slab priors on the weights and on the factors. All three models were fitted three times with random initialisations of the latent variables in order to assess the robustness of the inference process.

Correlating the inferred weights with the simulated weights for the three model showed that only models with spike-and-slab priors identified the true latent weights (Fig. E.1). In addition, only those models inferred reproducible weights across multiple trials (Fig. E.2).

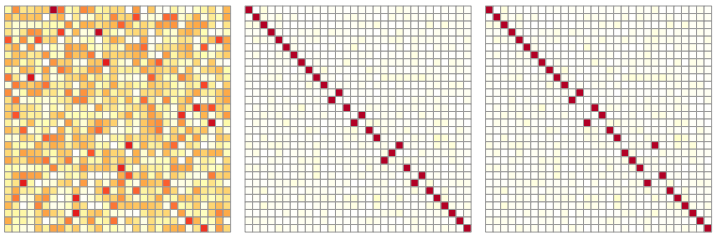


Fig. E.1 Correlation of the simulated weights with weights inferred with a model without any spike-and-slab prior (left), with spike-and-slab priors on the weights (middle) and with both spike-and-slab priors on the weights and on the factors (right). The generative latent factors all exhibited sparse effects (10% of active weights). Note that off-diagonal elements for the two sparse models are only due to factors permutation.

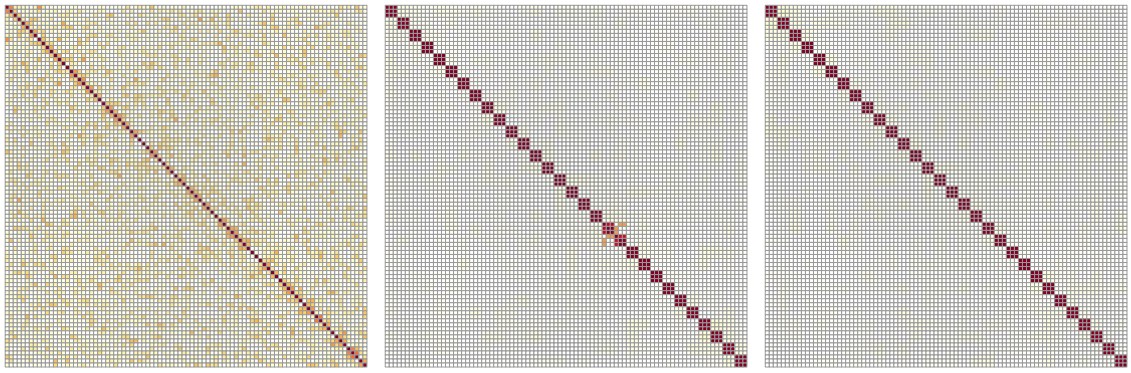


Fig. E.2 Robustness of weights inference for a model without any spike-and-slab prior (left), with spike-and-slab priors on the weights (middle) and with both spike-and-slab priors on the weights and on the factors (right). Each heat map shows the correlation matrix between all inferred weights for all three separate runs, which is then further hierarchically clustered so that reproduced weights between runs appear as blocks of three on the diagonal. The generative latent factors all exhibited sparse effects (10% of active weights).

E.1.2 Simulations with dense factors only

We then generated data using 30 latent variables with 100 % of active weights. Data was simulated for 800 samples and 1600 features and the latent variables explained 75% of the

total variability (due to the higher density of the weights).

Again, we compared three models: a first model with no spike-and-slab prior; a second model with spike-and-slab priors on the weights and a third model with both spike-and-slab priors on the weights and on the factors. Like in Section E.1.1, the three models were fitted three times with random initialisations of the latent variables in order to assess the robustness of the inference process. Correlation of the inferred weights with the simulated weights showed that none of the models could identify the true latent weights (Fig. E.4). Across our three random trials, inference was only reproducible for the model with spike-and-slab priors on both weights and factors (Fig. E.4), although the inferred weights were in any case rotated compared to the true simulated weights.

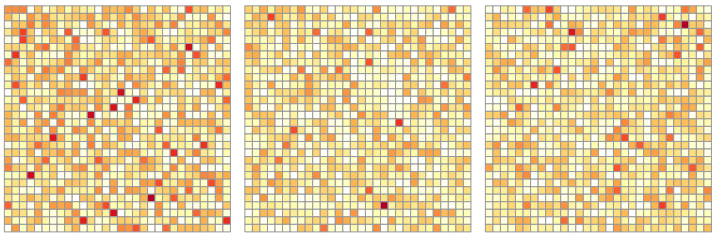


Fig. E.3 Correlation of the simulated weights with weights inferred with a model without any spike-and-slab prior (left), with spike-and-slab priors on the weights (middle) and with both spike-and-slab priors on the weights and on the factors (right). The generative latent factors were all dense (100% of active weights).

Taken together, these results confirm the usefulness of sparsity-inducing priors for the identification of sparse latent factors, while showing that dense latent factors pose in any case an identifiability problem.

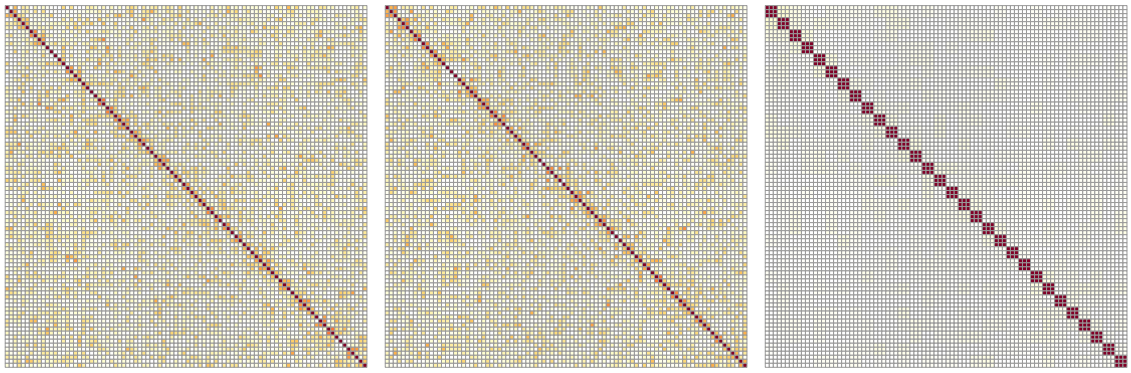


Fig. E.4 Robustness of weights inference for a model without any spike-and-slab prior (left), with spike-and-slab priors on the weights (middle) and with both spike-and-slab priors on the weights and on the factors (right). Each heat map shows the correlation matrix between all inferred weights for all three separate runs, so that off-diagonal blocks correspond to correlation plots between runs. The generative latent factors were all dense (100% of active weights).

References

Achim, K., Pettit, J.-B., Saraiva, L. R., Gavriouchkina, D., Larsson, T., Arendt, D., and Marioni, J. C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.*, 33(5):503–509.

Ahmed, S., Rattray, M., and Boukouvalas, A. (2017). GrandPrix: Scaling up the Bayesian GPLVM for single-cell data. *bioRxiv*, page 227843.

Aichler, M. and Walch, A. (2015). MALDI imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Lab. Invest.*, 95(4):422–431.

Akerfelt, M., Morimoto, R. I., and Sistonen, L. (2010). Heat shock factors: integrators of cell stress, development and lifespan. *Nat. Rev. Mol. Cell Biol.*, 11(8):545–555.

AM, F., K, F., LM, L., W, C., and RH, S. (2003). Visualization of single molecules of mrna in situ. *Nat. Methods*, 361:245–304.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276.

Androulakis, I. P., Yang, E., and Almon, R. R. (2007). Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annu. Rev. Biomed. Eng.*, 9:205–228.

Angelo, M., Bendall, S. C., Finck, R., Hale, M. B., Hitzman, C., Borowsky, A. D., Levenson, R. M., Lowe, J. B., Liu, S. D., Zhao, S., Natkunam, Y., and Nolan, G. P. (2014). Multiplexed ion beam imaging of human breast tumors. *Nat. Med.*, 20(4):436–442.

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O., and Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*, 13:229.

Angulo, M. C., Kozlov, A. S., Charpak, S., and Audinat, E. (2004). Glutamate released from glial cells synchronizes neuronal activity in the hippocampus. *J. Neurosci.*, 24(31):6920–6927.

Anscombe, F. J. (1948). The transformation of poisson, binomial and Negative-Binomial data. *Biometrika*, 35(3-4):246–254.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, 14(6):e8124.

Arnold, S. J. and Robertson, E. J. (2009). Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat. Rev. Mol. Cell Biol.*, 10(2):91–103.

Asa-aki Sato, M. (2001). Online model selection based on the Variational. *Neural Comput.*, 13:1649–1681.

Banerji, S., Ni, J., Wang, S. X., Clasper, S., Su, J., Tammi, R., Jones, M., and Jackson, D. G. (1999). LYVE-1, a new homologue of the CD44 glycoprotein, is a lymph-specific receptor for hyaluronan. *J. Cell Biol.*, 144(4):789–801.

Bangdiwala, S. I. (1989). The wald statistic in proportional hazards hypothesis testing. *Biom. J.*, 31(2):203–211.

Bartholomew, D. J. (1985). Foundations of factor analysis: Some practical implications. *Br. J. Math. Stat. Psychol.*, 38(1):1–10.

Battich, N., Stoeger, T., and Pelkmans, L. (2013). Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat. Methods*, 10(11):1127–1133.

Battich, N., Stoeger, T., and Pelkmans, L. (2015). Control of transcript variability in single mammalian cells. *Cell*, 163(7):1596–1610.

Baud, A., Mulligan, M. K., Casale, F. P., Ingels, J. F., Bohl, C. J., Callebert, J., Launay, J.-M., Krohn, J., Legarra, A., Williams, R. W., and Stegle, O. (2017). Genetic variation in the social environment contributes to health and disease. *PLoS Genet.*, 13(1):e1006498.

Beal, M. J. and Others (2003). *Variational algorithms for approximate Bayesian inference*. university of London London.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Biswas, S. (2008). Introduction to coding theory: basic codes and shannon's theorem. [online] http://www.math.uchicago.edu/~may/VIGRE/VIGRE2008/REUPapers/Biswas.pdf.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2016). Variational inference: A review for statisticians.

Bochner, S. (1959). *Lectures on Fourier integrals*. Princeton University Press.

Bodenmiller, B. (2016). Multiplexed Epitope-Based tissue imaging for discovery and healthcare applications. *Cell Syst*, 2(4):225–238.

Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., and Sagastizábal, C. A. (2006). *Numerical Optimization: Theoretical and Practical Aspects (Universitext)*. Springer-Verlag, Berlin, Heidelberg.

Bottou, L. (2011). From machine learning to machine reasoning.

Boukouvalas, A., Hensman, J., and Rattray, M. (2018). BGP: identifying gene-specific branching dynamics from single-cell data with a branching Gaussian process. *Genome Biol.*, 19(1):65.

Brakebusch, C. and Fässler, R. (2003). The integrin-actin connection, an eternal love affair. *EMBO J.*, 22(10):2324–2333.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, 33(2):155–160.

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. (2017). f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.*, 18(1):212.

Bunte, K., Leppäaho, E., Saarinen, I., and Kaski, S. (2016). Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics*, 32(16):2457–2463.

Bůžková, P., Lumley, T., and Rice, K. (2011). Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Ann. Hum. Genet.*, 75(1):36–45.

Campbell, K. R. and Yau, C. (2016). Order under uncertainty: Robust differential expression analysis using probabilistic models for pseudotime inference. *PLoS Comput. Biol.*, 12(11):e1005212.

Cancer Genome Atlas Research Network (2017). Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, 169(7):1327–1341.e23. Electronic address: wheeler@bcm.edu.

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., and Shendure, J. (2017). Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. *bioRxiv*.

Carmeli, C., De Vito, E., and Toigo, A. (2005). Reproducing kernel hilbert spaces and mercer theorem.

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. A., Chang, J. H., Lindquist, R. A., Moffat, J., Golland, P., and Sabatini, D. M. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, 7(10):R100.

Carpenter, C. L. (2000). Actin cytoskeleton and cell signaling. *Crit. Care Med.*, 28(4 Suppl):N94–9.

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319.

Casale, F. P. (2016). *Multivariate linear mixed models for statistical genetics*. PhD thesis, University of Cambridge. [online] https://doi.org/10.17863/CAM.13422.

Casale, F. P., Horta, D., Rakitsch, B., and Stegle, O. (2017). Joint genetic analysis using variant sets reveals polygenic gene-context interactions. *PLoS Genet.*, 13(4):e1006693.

Chan, M., Smith, Z. D., Grosswendt, S., Kretzmer, H., Norman, T., Adamson, B., Jost, M., Quinn, J. J., Yang, D., Meissner, A., and Weissman, J. S. (2018). Molecular recording of mammalian embryogenesis. *bioRxiv*.

Chan, S. S.-K., Shi, X., Toyama, A., Arpke, R. W., Dandapat, A., Iacovino, M., Kang, J., Le, G., Hagen, H. R., Garry, D. J., and Kyba, M. (2013). Mesp1 patterns mesoderm into cardiac, hematopoietic, or skeletal myogenic progenitors in a context-dependent manner. *Cell Stem Cell*, 12(5):587–601.

Chang, Q., Ornatsky, O. I., Siddiqui, I., Loboda, A., Baranov, V. I., and Hedley, D. W. (2017). Imaging mass cytometry. *Cytometry Part A*, 91(2):160–169.

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). RNA imaging. spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090.

Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., Datta, A., Richardson, D., Burden, F., Mead, D., Mann, A. L., Fernandez, J. M., Rowlston, S., Wilder, S. P., Farrow, S., Shao, X., Lambourne, J. J., Redensek, A., Albers, C. A., Amstislavskiy, V., Ashford, S., Berentsen, K., Bomba, L., Bourque, G., Bujold, D., Busche, S., Caron, M., Chen, S.-H., Cheung, W., Delaneau, O., Dermitzakis, E. T., Elding, H., Colgiu, I., Bagger, F. O., Flicek, P., Habibi, E., Iotchkova, V., Janssen-Megens, E., Kim, B., Lehrach, H., Lowy, E., Mandoli, A., Matarese, F., Maurano, M. T., Morris, J. A., Pancaldi, V., Pourfarzad, F., Rehnstrom, K., Rendon, A., Risch, T., Sharifi, N., Simon, M.-M., Sultan, M., Valencia, A., Walter, K., Wang, S.-Y., Frontini, M., Antonarakis, S. E., Clarke, L., Yaspo, M.-L., Beck, S., Guigo, R., Rico, D., Martens, J. H. A., Ouwehand, W. H., Kuijpers, T. W., Paul, D. S., Stunnenberg, H. G., Stegle, O., Downes, K., Pastinen, T., and Soranzo, N. (2016). Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, 167(5):1398–1414.e24.

Chen, R., Wu, X., Jiang, L., and Zhang, Y. (2017). Single-Cell RNA-Seq reveals hypothalamic cell diversity. *Cell Rep.*, 18(13):3227–3241.

Cheng, J.-C., Chang, H.-M., Fang, L., Sun, Y.-P., and Leung, P. C. K. (2015). TGF-β1 up-regulates connexin43 expression: a potential mechanism for human trophoblast cell differentiation. *J. Cell. Physiol.*, 230(7):1558–1566.

"Chlon, L. and Markowetz, F. (2017). Causal modeling dissects Tumour–Microenvironment interactions in breast cancer. *bioRxiv*, page 144832.

Chu, Y. and Corey, D. R. (2012). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther.*, 22(4):271–274.

Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., and Reik, W. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.*, 9(1):781.

Colomé-Tatché, M. and Theis, F. J. (2018). Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, 7:54–59.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., and Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39(Database issue):D691–7.

Cunningham, J. P. and Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *J. Mach. Learn. Res.*, 16:2859–2900.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., METABRIC Group, Caldas, C., Aparicio, S., Curtis†, C., Shah, S. P., Caldas, C., Aparicio, S., Brenton, J. D., Ellis, I., Huntsman, D., Pinder, S., Purushotham, A., Murphy, L., Caldas, C., Aparicio, S., Caldas, C., Bardwell, H., Chin, S.-F., Curtis, C., Ding, Z., Gräf, S., Jones, L., Liu, B., Lynch, A. G., Papatheodorou, I., Sammut, S. J., Wishart, G., Aparicio, S., Chia, S., Gelmon, K., Huntsman, D., McKinney, S., Speers, C., Turashvili, G., Watson, P., Ellis, I., Blamey, R., Green, A., Macmillan, D., Rakha, E., Purushotham, A., Gillett, C., Grigoriadis, A., Pinder, S., de Rinaldis, E., Tutt, A., Murphy, L., Parisien, M., Troup, S., Caldas, C., Chin, S.-F., Chan, D., Fielding, C., Maia, A.-T., McGuire, S., Osborne, M., Sayalero, S. M., Spiteri, I., Hadfield, J., Aparicio, S., Turashvili, G., Bell, L., Chow, K., Gale, N., Huntsman, D., Kovalik, M., Ng, Y., Prentice, L., Caldas, C., Tavaré, S., Curtis, C., Dunning, M. J., Gräf, S., Lynch, A. G., Rueda, O. M., Russell, R., Samarajiwa, S., Speed, D., Markowetz, F., Yuan, Y., Brenton, J. D., Aparicio, S., Shah, S. P., Bashashati, A., Ha, G., Haffari, G., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C., and Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486:346.

Dai, H.-Q., Wang, B.-A., Yang, L., Chen, J.-J., Zhu, G.-C., Sun, M.-L., Ge, H., Wang, R., Chapman, D. L., Tang, F., Sun, X., and Xu, G.-L. (2016). TET-mediated DNA demethylation controls gastrulation by regulating Lefty-Nodal signalling. *Nature*, 538(7626):528–532.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. (2014). Variational inference for uncertainty on the inputs of Gaussian process models.

David J. C. MacKay (1994). Bayesian non-linear modelling for the prediction competition. In *In ASHRAE Transactions, V.100, Pt.2*.

Deshwar, A. R., Chng, S. C., Ho, L., Reversade, B., and Scott, I. C. (2016). The apelin receptor enhances Nodal/TGFβ signaling to ensure proper cardiac development. *Elife*, 5.

Dietrich, S., Oleś, M., Lu, J., Sellner, L., Anders, S., Velten, B., Wu, B., Hüllein, J., da Silva Liberio, M., Walther, T., Wagner, L., Rabe, S., Ghidelli-Disse, S., Bantscheff, M., Oleś, A. K., Słabicki, M., Mock, A., Oakes, C. C., Wang, S., Oppermann, S., Lukas, M., Kim, V., Sill, M., Benner, A., Jauch, A., Sutton, L. A., Young, E., Rosenquist, R., Liu, X.,

Jethwa, A., Lee, K. S., Lewis, J., Putzker, K., Lutz, C., Rossi, D., Mokhir, A., Oellerich, T., Zirlik, K., Herling, M., Nguyen-Khac, F., Plass, C., Andersson, E., Mustjoki, S., von Kalle, C., Ho, A. D., Hensel, M., Dürig, J., Ringshausen, I., Zapatka, M., Huber, W., and Zenz, T. (2018). Drug-perturbation-based stratification of blood cancer. *J. Clin. Invest.*, 128(1):427–445.

Dietz, L. (2010). Directed factor graph notation for generative models. *Technical Report*.

Dihazi, H., Asif, A. R., Beißbarth, T., Bohrer, R., Feussner, K., Feussner, I., Jahn, O., Lenz, C., Majcherczyk, A., Schmidt, B., Schmitt, K., Urlaub, H., and Valerius, O. (2018). Integrative omics - from data to biology. *Expert Rev. Proteomics*, 15(6):463–466.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., and Regev, A. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.

Doledec, S. and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshw. Biol.*, 31(3):277–294.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11:587.

Durrande, N., Hensman, J., Rattray, M., and Lawrence, N. D. (2016). Detecting periodicities with Gaussian Processes. *PeerJ Comput. Sci.*, 2:e50.

Duvenaud, D. (2014). *Automatic model construction with Gaussian Processes*. PhD thesis, University of Cambridge.

Elston, C. W. and Ellis, I. O. (1991). Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410.

Fabbri, G. and Dalla-Favera, R. (2016). The molecular pathogenesis of chronic lymphocytic leukaemia. *Nat. Rev. Cancer*, 16(3):145–162.

Fagan, A., Culhane, A. C., and Higgins, D. G. (2007). A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics*, 7(13):2162–2171.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., and Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, 16:278.

Fluhr, S., Boerries, M., Busch, H., Symeonidi, A., Witte, T., Lipka, D. B., Mücke, O., Nöllke, P., Krombholz, C. F., Niemeyer, C. M., Plass, C., and Flotho, C. (2016). CREBBP is a target of epigenetic, but not genetic, modification in juvenile myelomonocytic leukemia. *Clin. Epigenetics*, 8(1):50.

Franke, W. W. (2009). Discovering the molecular components of intercellular junctions—a historical view. *Cold Spring Harbor Perspectives in Biology*, 1(3):160–169.

Fricker, M., Neher, J. J., Zhao, J.-W., Théry, C., Tolkovsky, A. M., and Brown, G. C. (2012). MFG-E8 mediates primary phagocytosis of viable neurons during neuroinflammation. *J. Neurosci.*, 32(8):2657–2666.

Frolov, A. E., Godwin, A. K., and Favorova, O. O. (2003). [differential gene expression analysis by DNA microarrays technology and its application in molecular oncology]. *Mol. Biol.*, 37(4):573–584.

Fukumura, D. (2005). Role of microenvironment on gene expression, angiogenesis and microvascular function in tumors. In Meadows, G. G., editor, *Integration/Interaction of Oncologic Growth*, pages 23–36. Springer Netherlands, Dordrecht.

Gao, C., Brown, C. D., and Engelhardt, B. E. (2013). A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects.

Garg, R., Benedetti, L. G., Abera, M. B., Wang, H., Abba, M., and Kazanietz, M. G. (2014). Protein kinase C and cancer: what we know and what we do not. *Oncogene*, 33(45):5225–5237.

Gerdes, M. J., Sevinsky, C. J., Sood, A., Adak, S., Bello, M. O., Bordwell, A., Can, A., Corwin, A., Dinn, S., Filkins, R. J., Hollman, D., Kamath, V., Kaanumalle, S., Kenny, K., Larsen, M., Lazare, M., Li, Q., Lowes, C., McCulloch, C. C., McDonough, E., Montalto, M. C., Pang, Z., Rittscher, J., Santamaria-Pang, A., Sarachan, B. D., Seel, M. L., Seppo, A., Shaikh, K., Sui, Y., Zhang, J., and Ginty, F. (2013). Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl. Acad. Sci. U. S. A.*, 110(29):11982–11987.

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Porta, M. G. D., Jädersten, M., Dolatshad, H., Verma, A., Cross, N. C. P., Vyas, P., Killick, S., Hellström-Lindberg, E., Cazzola, M., Papaemmanuil, E., Campbell, P. J., and Boultwood, J. (2015). Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat. Commun.*, 6:5901.

Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., Varga, Z., Wild, P. J., Günther, D., and Bodenmiller, B. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods*, 11:417.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G. P. (2018). Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell*, 174(4):968–981.e15.

GPy (since 2012). GPy: A Gaussian process framework in python. http://github.com/SheffieldML/GPy.

Gretton, A. (2017). Introduction to rkhs, and some simple kernel algorithms. [online] http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture4_introToRKHS.pdf.

Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.*, 14(4):e8046.

Groot, P., Lucas, P., and Bosch, P. (2011). Multiple-step time series forecasting with sparse Gaussian Processes.

GTEx Consortium (2013). The Genotype-Tissue expression (GTEx) project. *Nat. Genet.*, 45(6):580–585.

Guo, F., Li, L., Li, J., Wu, X., Hu, B., Zhu, P., Wen, L., and Tang, F. (2017). Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.*, 27(8):967–988.

Guo, G., Huss, M., Tong, G. Q., Wang, C., Sun, L. L., Clarke, N. D., and Robson, P. (2010). Resolution of cell fate decisions revealed by Single-Cell gene expression analysis from zygote to blastocyst. *Dev. Cell*, 18(4):675–685.

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998.

Harshman, R. A. and Lundy, M. E. (1994). PARAFAC: Parallel factor analysis. *Comput. Stat. Data Anal.*, 18(1):39–72.

Hart, A. H., Hartley, L., Sourris, K., Stadler, E. S., Li, R., Stanley, E. G., Tam, P. P. L., Elefanty, A. G., and Robb, L. (2002). Mixl1 is required for axial mesendoderm morphogenesis and patterning in the murine embryo. *Development*, 129(15):3597–3608.

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.*, 18(1):83.

Haw, R. and Stein, L. (2012). Using the reactome database. *Curr. Protoc. Bioinformatics*, Chapter 8:Unit8.7.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013a). Gaussian processes for big data.

Hensman, J., Lawrence, N. D., and Rattray, M. (2013b). Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics*, 14(1):252.

Hensman, J., Rattray, M., and Lawrence, N. D. (2015). Fast nonparametric clustering of structured Time-Series. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):383–393.

Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Van Sanden, S., Lin, D., Talloen, W., Bijnens, L., Göhlmann, H. W. H., Shkedy, Z., and Clevert, D.-A. (2010). FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527.

Hoff, P. (2010). Hierarchical multilinear models for multiway data. *arXiv*.

Hoffman, G. E. and Schadt, E. E. (2016). variancepartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics*, 17(1):483.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine*.

Holmes, S., Alekseyenko, A., Timme, A., Nelson, T., Pasricha, P. J., and Spormann, A. (2011). Visualization and statistical comparisons of microbial communities using R packages on phylochip data. *Pac. Symp. Biocomput.*, pages 142–153.

Honkela, A., Raiko, T., Kuusela, M., Tornio, M., and Karhunen, J. (2010). Approximate riemannian conjugate gradient learning for Fixed-Form Variational bayes. *J. Mach. Learn. Res.*, 11(Nov):3235–3268.

Hore, V. (2015). *Latent Variable Models for Analysing Multidimensional Gene Expression Data*. PhD thesis, University of Oxford.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.*, 48(9):1094–1100.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Hothorn, T. and Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Comput. Stat. Data Anal.*, 43(2):121–137.

Hu, P., Fabyanic, E., Kwon, D. Y., Tang, S., Zhou, Z., and Wu, H. (2017). Dissecting Cell-Type composition and Activity-Dependent transcriptional state in mammalian brains by massively parallel Single-Nucleus RNA-Seq. *Mol. Cell*, 68(5):1006–1015.e7.

Hu, P., Zhang, W., Xin, H., and Deng, G. (2016). Single cell isolation and analysis. *Front Cell Dev Biol*, 4:116.

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: Recent progress in Multi-Omics data integration methods. *Front. Genet.*, 8:1005.

Huang, Y. and Sanguinetti, G. (2017). BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol.*, 18(1):123.

Ibarra-Soria, X., Jawaid, W., Pijuan-Sala, B., Ladopoulos, V., Scialdone, A., Jörg, D. J., Tyser, R. C. V., Calero-Nieto, F. J., Mulas, C., Nichols, J., Vallier, L., Srinivas, S., Simons, B. D., Göttgens, B., and Marioni, J. C. (2018). Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat. Cell Biol.*, 20(2):127–134.

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., Van Dyk, E., Chang, H., de Silva, H., Heyn, H., Deng, X., Egan, R. K., Liu, Q., Mironenko, T., Mitropoulos, X., Richardson, L., Wang, J., Zhang, T., Moran, S., Sayols, S., Soleimani, M., Tamborero, D., Lopez-Bigas, N., Ross-Macdonald, P., Esteller, M., Gray, N. S., Haber, D. A., Stratton, M. R., Benes, C. H., Wessels, L. F. A., Saez-Rodriguez, J., McDermott, U., and Garnett, M. J. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, 21(7):1160–1167.

Iversen, L. (2009). Neurotransmitter transporters and their impact on the development of psychopharmacology. *Br. J. Pharmacol.*, 147(S1):S82–S88.

Iwashita, H., Shiraki, N., Sakano, D., Ikegami, T., Shiga, M., Kume, K., and Kume, S. (2013). Secreted cerberus1 as a marker for quantification of definitive endoderm differentiation of the pluripotent stem cells. *PLoS One*, 8(5):e64291.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Stat. Comput.*, 10(1):25–37.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to Variational methods for graphical models. *Mach. Learn.*, 37(2):183–233.

Kalaitzis, A. A. and Lawrence, N. D. (2011). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics*, 12(1):180.

Kamińska, K., Szczylik, C., Bielecka, Z. F., Bartnik, E., Porta, C., Lian, F., and Czarnecka, A. M. (2015). The role of the cell-cell interactions in cancer progression. *J. Cell. Mol. Med.*, 19(2):283–296.

Kang, E. Y., Martin, L. J., Mangul, S., Isvilanonda, W., Zou, J., Ben-David, E., Han, B., Lusis, A. J., Shifman, S., and Eskin, E. (2016). Discovering single nucleotide polymorphisms regulating human gene expression using allele specific expression from RNA-seq data. *Genetics*, 204(3):1057–1064.

Kay, A. W., Strauss-Albee, D. M., and Blish, C. A. (2016). Application of mass cytometry (CyTOF) for functional and phenotypic analysis of natural killer cells. *Methods Mol. Biol.*, 1441:13–26.

Khan, S., Virtanen, S., Kallioniemi, O., Wennerberg, Poso, A., and Kaski, S. (2014). Identification of structural features in chemicals associated with cancer drug response: A systematic data-driven analysis. *Bioinformatics*, 30(17):497–504.

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, 12:357.

Kim, M., Rai, N., Zorraquino, V., and Tagkopoulos, I. (2016). Multi-omics integration accurately predicts cellular state in unexplored conditions for escherichia coli. *Nat. Commun.*, 7:13090.

Kirschner, K., Chandra, T., Kiselev, V., Flores-Santa Cruz, D., Macaulay, I. C., Park, H. J., Li, J., Kent, D. G., Kumar, R., Pask, D. C., Hamilton, T. L., Hemberg, M., Reik, W., and Green, A. R. (2017). Proliferation drives Aging-Related functional decline in a subpopulation of the hematopoietic stem cell compartment. *Cell Rep.*, 19(8):1503–1511.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., and Hemberg, M. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, 14(5):483–486.

Klami, A., Virtanen, S., Leppäaho, E., and Kaski, S. (2015). Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2136–2147.

Kolda, T. and Bader, B. (2009). Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell*, 58(4):610–620.

Komili, S. and Silver, P. A. (2008). Coupling and coordination in gene expression processes: a systems biology view. *Nat. Rev. Genet.*, 9:38.

Komiya, Y. and Habas, R. (2008). Wnt signal transduction pathways. *Organogenesis*, 4(2):68–75.

Kostem, E. and Eskin, E. (2013). Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *Am. J. Hum. Genet.*, 92(4):558–564.

Kschischang, F. R., Frey, B. J., and . Loeliger, H. (2001). Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory*, 47(2):498–519.

Kukurba, K. R. and Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harb. Protoc.*, 2015(11):951–969.

Kumar, P., Tan, Y., and Cahan, P. (2017). Understanding development and stem cells using single cell-based analyses of gene expression. *Development*, 144(1):17–32.

L. Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17(1):75.

Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J. Mach. Learn. Res.*, 6(Nov):1783–1816.

Lawrence, N. D. and Bishop, C. M. (2000). Variational Bayesian independent component analysis. *Univ of Cambridge Tech Report*.

Lawrence, N. D., Sanguinetti, G., and Rattray, M. (2007). Modelling transcriptional regulation using Gaussian processes. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 785–792. MIT Press.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D. Y., Duque, R., Bersini, H., and Nowé, A. (2013). Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.*, 14(4):469–490.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, 3(9):e161.

Leppäaho, E., Ammad-ud din, M., and Kaski, S. (2017). GFA: Exploratory analysis of multiple data sources with group factor analysis. *J. Mach. Learn. Res.*, 18(39):1–5.

Li, G. and Zhu, H. (2013). Genetic studies: The linear mixed models in genome-wide association studies. *The open bioinformatics journal*.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Lin, J.-R., Fallahi-Sichani, M., and Sorger, P. K. (2015). Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nat. Commun.*, 6:8390.

Lin, J.-R., Izar, B., Mei, S., Wang, S., Shah, P., and Sorger, P. (2017). A simple open-source method for highly multiplexed imaging of single cells in tissues and tumours. *bioRxiv*, page 151738.

Lippert, C., Casale, F. P., Rakitsch, B., and Stegle, O. (2014). LIMIX: genetic analysis of multiple traits. *bioRxiv*, page 003905.

Lisman, J., Yasuda, R., and Raghavachari, S. (2012). Mechanisms of CaMKII action in long-term potentiation. *Nat. Rev. Neurosci.*, 13(3):169–182.

Loeliger, H.-A. (2004). An introduction to factor graphs. *IEEE Sig. Proc. Mag.*, page 28–41.

Lönnberg, T., Svensson, V., James, K. R., Fernandez-Ruiz, D., Sebina, I., Montandon, R., Soon, M. S. F., Fogg, L. G., Nair, A. S., Liligeto, U., Stubbington, M. J. T., Ly, L.-H., Bagger, F. O., Zwiessele, M., Lawrence, N. D., Souza-Fonseca-Guimaraes, F., Bunn, P. T., Engwerda, C. R., Heath, W. R., Billker, O., Stegle, O., Haque, A., and Teichmann, S. A. (2017). Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria. *Sci Immunol*, 2(9).

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550.

Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods*, 11:360.

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., Shi, T., Tong, W., Shi, L., Hong, H., Zhao, C., Elloumi, F., Shi, W., Thomas, R., Lin, S., Tillinghast, G., Liu, G., Zhou, Y., Herman, D., Li, Y., Deng, Y., Fang, H., Bushel, P., Woods, M., and Zhang, J. (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.*, 10(4):278–291.

Lyssiotis, C. A. and Kimmelman, A. C. (2017). Metabolic interactions in the tumor microenvironment. *Trends Cell Biol.*, 27(11):863–875.

Ma, X. and Gao, L. (2012). Biological network analysis: insights into structure and functions. *Brief. Funct. Genomics*, 11(6):434–442.

Maaten, L. V. d. and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9(Nov):2579–2605.

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., and Voet, T. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods*, 12:519.

Macaulay, I. C., Svensson, V., Labalette, C., Ferreira, L., Hamey, F., Voet, T., Teichmann, S. A., and Cvejic, A. (2016a). Single-Cell RNA-Sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep.*, 14(4):966–977.

Macaulay, I. C., Svensson, V., Labalette, C., Ferreira, L., Hamey, F., Voet, T., Teichmann, S. A., and Cvejic, A. (2016b). Single-Cell RNA-Sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep.*, 14(4):966–977.

Macaulay, I. C. and Voet, T. (2014). Single cell genomics: Advances and future perspectives. *PLoS Genet.*, 10(1):e1004126.

MacKay, D. J. C. (1998). Introduction to Gaussian processes. *Neural Networks and Machine Learning*.

Maloum, K., Settegrana, C., Chapiro, E., Cazin, B., Leprêtre, S., Delmer, A., Leporrier, M., Dreyfus, B., Tournilhac, O., Mahe, B., Nguyen-Khac, F., Lesty, C., Davi, F., and Merle-Béral, H. (2009). IGHV gene mutational status and LPL/ADAM29 gene expression as clinical outcome predictors in CLL patients in remission following treatment with oral fludarabine plus cyclophosphamide. *Ann. Hematol.*, 88(12):1215–1221.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(1):S7.

Marguerat, S. and Bähler, J. (2010). RNA-seq: from technology to biology. *Cell. Mol. Life Sci.*, 67(4):569–579.

Markowetz, F. (2010). How to understand the cell by breaking it: Network analysis of gene perturbation screens. *PLoS Comput. Biol.*, 6(2):e1000655.

Markowetz, F. and Spang, R. (2007). Inferring cellular networks – a review. *BMC Bioinformatics*, 8(6):S5.

Mason, S. (2017). Lactate shuttles in Neuroenergetics-Homeostasis, allostasis and beyond. *Front. Neurosci.*, 11:43.

Masson, J., Sagné, C., Hamon, M., and El Mestikawy, S. (1999). Neurotransmitter transporters in the central nervous system. *Pharmacol. Rev.*, 51(3):439–464.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 11(Aug):2287–2322.

McDowell, I. C., Manandhar, D., Vockley, C. M., Schmid, A. K., Reddy, T. E., and Engelhardt, B. E. (2018). Clustering gene expression time series data using an infinite Gaussian process mixture model. *PLoS Comput. Biol.*, 14(1):e1005896.

McKnight, K. D., Hou, J., and Hoodless, P. A. (2007). Dynamic expression of thyrotropin-releasing hormone in the mouse definitive endoderm. *Dev. Dyn.*, 236(10):2909–2917.

Meng, C., Kuster, B., Culhane, A. C., and Gholami, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, 15:162.

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.*, 17(4):628–641.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character (1905-1934)*, 83(March):69–70.

Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J. T., Gatza, M. L., Wilkerson, M., Perou, C. M., Yellapantula, V., Huang, K.-L., Lin, C., McLellan, M. D., Yan, P., Davies, S. R., Townsend, R. R., Skates, S. J., Wang, J., Zhang, B., Kinsinger, C. R., Mesri, M., Rodriguez, H., Ding, L., Paulovich, A. G., Fenyö, D., Ellis, M. J., Carr, S. A., and NCI CPTAC (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605):55–62.

Meugnier, E., Rome, S., and Vidal, H. (2007). Regulation of gene expression by glucose. *Curr. Opin. Clin. Nutr. Metab. Care*, 10(4):518–522.

Micchelli, C. A., Xu, Y., and Zhang, H. (2006). Universal kernels. *J. Mach. Learn. Res.*, 7:2651–2667.

Michael I. Jordan, R. T. (2004). The kernel trick. [online] https://people.eecs.berkeley.edu/~jordan/courses/281B-spring04/lectures/lec3.pdf.

Minh, H. Q., Niyogi, P., and Yao, Y. (2006). Mercer's theorem, feature maps, and smoothing. *on Computational Learning Theory*.

Minka, T. P. (2013). Expectation propagation for approximate Bayesian inference.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.*, 83(404):1023–1032.

Mo, Q. and Shen, R. (2013). iClusterPlus: integrative clustering of multiple genomic data sets.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. U. S. A.*, 110(11):4245–4250.

Mohammed, H., Hernando-Herraez, I., Savino, A., Scialdone, A., Macaulay, I., Mulas, C., Chandra, T., Voet, T., Dean, W., Nichols, J., Marioni, J. C., and Reik, W. (2017). Single-Cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep.*, 20(5):1215–1228.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, 34(3):267–273.

Morabito, F., Cutrona, G., Mosca, L., D'Anca, M., Matis, S., Gentile, M., Vigna, E., Colombo, M., Recchia, A. G., Bossio, S., De Stefano, L., Maura, F., Manzoni, M., Ilariucci, F., Consoli, U., Vincelli, I., Musolino, C., Cortelezzi, A., Molica, S., Ferrarini, M., and Neri, A. (2015). Surrogate molecular markers for IGHV mutational status in chronic lymphocytic leukemia for predicting time to first treatment. *Leuk. Res.*, 39(8):840–845.

Moreau, H. D., Piel, M., Voituriez, R., and Lennon-Duménil, A.-M. (2018). Integrating physical and molecular insights on immune cell migration. *Trends Immunol.*, 39(8):632–643.

Murphy, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution. *DEF*, 1($2\sigma 2$):16.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Nachtergaele, S. and He, C. (2017). The emerging biology of RNA post-transcriptional modifications. *RNA Biol.*, 14(2):156–163.

Nasra Naeim Ayuob and Soad Shaker Ali (2012). Cell-Cell interactions and cross talk described in normal and disease conditions: Morphological approach. In Gowder, S., editor, *Cell Interaction*. InTech.

Neal, R. M. (1995). Bayesian learning for neural networks.

Neher, J. J., Emmrich, J. V., Fricker, M., Mander, P. K., Théry, C., and Brown, G. C. (2013). Phagocytosis executes delayed neuronal death after focal brain ischemia. *Proc. Natl. Acad. Sci. U. S. A.*, 110(43):E4098–E4107.

Ng, E. S., Azzola, L., Sourris, K., Robb, L., Stanley, E. G., and Elefanty, A. G. (2005). The primitive streak gene mixl1 is required for efficient haematopoiesis and BMP4-induced ventral mesoderm patterning in differentiating ES cells. *Development*, 132(5):873–884.

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., Hedman, Å. K., Bataille, V., Bell, J. T., Surdulescu, G., Dimas, A. S., Ingle, C., Nestle, F. O., di Meglio, P., Min, J. L., Wilk, A., Hammond, C. J., Hassanali, N., Yang, T.-P., Montgomery, S. B., O'Rahilly, S., Lindgren, C. M., Zondervan, K. T., Soranzo, N., Barroso, I., Durbin, R., Ahmadi, K., Deloukas, P., McCarthy, M. I., Dermitzakis, E. T., Spector, T. D., and The MuTHER Consortium (2011). The architecture of gene regulatory variation across multiple human tissues: The MuTHER study. *PLoS Genet.*, 7(2):e1002003.

Niu, M., Dai, Z., Lawrence, N., and Becker, K. (2016). Spatio-temporal Gaussian Processes modeling of dynamical systems in systems biology.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. (2008). Genes mirror geography within europe. *Nature*, 456:98.

Oakes, C. C., Seifert, M., Assenov, Y., Gu, L., Przekopowitz, M., Ruppert, A. S., Wang, Q., Imbusch, C. D., Serva, A., Koser, S. D., Brocks, D., Lipka, D. B., Bogatyrova, O., Weichenhan, D., Brors, B., Rassenti, L., Kipps, T. J., Mertens, D., Zapatka, M., Lichter, P., Döhner, H., Küppers, R., Zenz, T., Stilgenbauer, S., Byrd, J. C., and Plass, C. (2016). DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.*, 48(3):253–264.

O'Donnell, M., Chance, R. K., and Bashaw, G. J. (2009). Axon growth and guidance: receptor regulation and signal transduction. *Annu. Rev. Neurosci.*, 32:383–412.

Okuta, R., Unno, Y., Nishino, D., Hido, S., and Loomis, C. (2017). Cupy: A numpy-compatible library for nvidia gpu calculations. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*.

Oliva, J. B., Dubey, A., Wilson, A. G., Poczos, B., Schneider, J., and Xing, E. P. (2016). Bayesian nonparametric Kernel-Learning. In *Artificial Intelligence and Statistics*, pages 1078–1086.

Ou, L., Fang, L., Tang, H., Qiao, H., Zhang, X., and Wang, Z. (2016). Dickkopf wnt signaling pathway inhibitor 1 regulates the differentiation of mouse embryonic stem cells in vitro and in vivo. *Mol. Med. Rep.*, 13(1):720–730.

Parisi, G. (1988). *Statistical Field Theory*. Addison-Wesley.

Pearson, K. (1901). LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(Oct):2825–2830.

Peng, G. and Jing, N. (2017). The genome-wide molecular regulation of mouse gastrulation embryo. *Sci. China Life Sci.*, 60(4):363–369.

Perraudeau, F., Risso, D., Street, K., Purdom, E., and Dudoit, S. (2017). Bioconductor workflow for single-cell RNA sequencing: Normalization, dimensionality reduction, clustering, and lineage inference. *F1000Res.*, 6.

Pierson, E. and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, 16(1):241.

Plesingerova, H., Librova, Z., Plevova, K., Libra, A., Tichy, B., Skuhrova Francova, H., Vrbacky, F., Smolej, L., Mayer, J., Bryja, V., Doubek, M., and Pospisilova, S. (2017). COBLL1, LPL and ZAP70 expression defines prognostic subgroups of chronic lymphocytic leukemia patients with high accuracy and correlates with IGHV mutational status. *Leuk. Lymphoma*, 58(1):70–79.

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., Li, N., Szpankowski, L., Fowler, B., Chen, P., Ramalingam, N., Sun, G., Thu, M., Norris, M., Lebofsky, R., Toppani, D., Kemp, 2nd, D. W., Wong, M., Clerkson, B., Jones, B. N., Wu, S., Knutsson, L., Alvarado, B., Wang, J., Weaver, L. S., May, A. P., Jones, R. C., Unger, M. A., Kriegstein, A. R., and West, J. A. A. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, 32(10):1053–1058.

Queirós, A. C., Villamor, N., Clot, G., Martinez-Trillos, A., Kulis, M., Navarro, A., Penas, E. M. M., Jayne, S., Majid, A., Richter, J., Bergmann, A. K., Kolarova, J., Royo, C., Russiñol, N., Castellano, G., Pinyol, M., Bea, S., Salaverria, I., López-Guerra, M., Colomer, D., Aymerich, M., Rozman, M., Delgado, J., Giné, E., González-Díaz, M., Puente, X. S., Siebert, R., Dyer, M. J. S., López-Otín, C., Rozman, C., Campo, E., López-Guillermo, A., and Martín-Subero, J. I. (2015). A b-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia*, 29(3):598–605.

Quinn, E. M., Cormican, P., Kenny, E. M., Hill, M., Anney, R., Gill, M., Corvin, A. P., and Morris, D. W. (2013). Development of strategies for SNP detection in RNA-Seq data: Application to lymphoblastoid cell lines and evaluation using 1000 genomes data. *PLoS One*, 8(3):e58815.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.*, 6(Dec):1939–1959.

Rafalska-Metcalf, I. U., Powers, S. L., Joo, L. M., LeRoy, G., and Janicki, S. M. (2010). Single cell analysis of transcriptional activation dynamics. *PLoS One*, 5(4):e10272.

Rahimi, A. and Recht, B. (2008). Random features for Large-Scale kernel machines. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

Raj, A., Van den Bogaard, P., Rifkin, S. A., Van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, 5(10):877–879.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Remes, S., Mononen, T., and Kaski, S. (2015). Classification of weak multi-view signals by sharing factors in a mixture of Bayesian group factor analyzers.

Richardson, M. (2009). Principal component analysis. *URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3. 5. 2013). Aleš Hladnik Dr. , Ass. Prof. , Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si*, 6:16.

Ringnér, M. (2008). What is principal component analysis? *Nat. Biotechnol.*, 26:303.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, 9(1):284.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015a). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.*, 16(2):85–97.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015b). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47–e47.

Rizvi, S. A. A., Roberts, S. J., Osborne, M. A., and Nyikosa, F. (2017). A novel approach to forecasting financial volatility with Gaussian process envelopes.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2012). Gaussian processes for timeseries modelling.

Rohart, F., Gautier, B., Singh, A., and Cao, K.-A. L. (2017). mixomics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.*, 13(11):e1005752.

Rosenberg, A. B., Roco, C., Muscat, R. A., Kuchina, A., Mukherjee, S., Chen, W., Peeler, D. J., Yao, Z., Tasic, B., Sellers, D. L., Pun, S. H., and Seelig, G. (2017). Scaling single cell transcriptomics through split pool barcoding. *bioRxiv*.

Roweis, S. T. (1998). EM algorithms for PCA and SPCA. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10*, pages 626–632. MIT Press.

Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76.

Ruder, S. (2016). An overview of gradient descent optimization algorithms.

Samo, Y.-L. K. and Roberts, S. (2015). Generalized spectral kernels. *arXiv*.

Schapiro, D., Jackson, H. W., Raghuraman, S., Fischer, J. R., Zanotelli, V. R. T., Schulz, D., Giesen, C., Catena, R., Varga, Z., and Bodenmiller, B. (2017). histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat. Methods*, 14(9):873–876.

Schelker, M., Feau, S., Du, J., Ranu, N., Klipp, E., MacBeath, G., Schoeberl, B., and Raue, A. (2017). Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.*, 8(1):2032.

Schüffler, P. J., Schapiro, D., Giesen, C., Wang, H. A. O., Bodenmiller, B., and Buhmann, J. M. (2015). Automatic single cell segmentation on highly multiplexed tissue images. *Cytometry A*, 87(10):936–942.

Schulz, D., Zanotelli, V. R. T., Fischer, J. R., Schapiro, D., Engler, S., Lun, X.-K., Jackson, H. W., and Bodenmiller, B. (2018). Simultaneous multiplexed imaging of mRNA and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry. *Cell Syst*, 6(1):25–36.e5.

Scialdone, A., Natarajan, K. N., Saraiva, L. R., Proserpio, V., Teichmann, S. A., Stegle, O., Marioni, J. C., and Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61.

Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N. K., Macaulay, I. C., Marioni, J. C., and Göttgens, B. (2016). Resolving early mesoderm diversification through single-cell expression profiling. *Nature*, 535(7611):289–293.

Searle, S. R. (1982). In *Matrix algebra useful for statistics (wiley series in probability and statistics)*, page 67.

Seeger, M. and Bouchard, G. (2012). Fast variational Bayesian inference for non-conjugate matrix factorization models. *Artificial Intelligence and Statistics*.

Serviss, J. T., Gådin, J. R., Eriksson, P., Folkersen, L., and Grandér, D. (2017). ClusterSignificance: a bioconductor package facilitating statistical analysis of class cluster separations in dimensionality reduced data. *Bioinformatics*, 33(19):3126–3128.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2):342–357.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2017). seqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. *Neuron*, 94(4):752–758.e1.

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., Schwartz, S., Fowler, B., Weaver, S., Wang, J., Wang, X., Ding, R., Raychowdhury, R., Friedman, N., Hacohen, N., Park, H., May, A. P., and Regev, A. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510:363.

Shamah, S. M., Lin, M. Z., Goldberg, J. L., Estrach, S., Sahin, M., Hu, L., Bazalakova, M., Neve, R. L., Corfas, G., Debant, A., and Greenberg, M. E. (2001). EphA receptors regulate growth cone dynamics through the novel guanine nucleotide exchange factor ephexin. *Cell*, 105(2):233–244.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

Sieck, G. (2014). Physiology in perspective: Cell-Cell interactions: The physiological basis of communication. *Physiology*, 29(4):220–221.

Singh, A., Gautier, B., Shannon, C. P., Vacher, M., Rohart, F., Tebutt, S. J., and Le Cao, K.-A. (2016). DIABLO - an integrative, multi-omics, multivariate method for multi-group classification. *bioRxiv*, page 067611.

Skinner, S. O., Xu, H., Nagarkar-Jaiswal, S., Freire, P. R., Zwaka, T. P., and Golding, I. (2016). Single-cell analysis of transcription kinetics across the cell cycle. *Elife*, 5:e12175.

Snel, B., Lehmann, G., Bork, P., and Huynen, M. A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.*, 28(18):3442–3444.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In Weiss, Y., Schölkopf, B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT Press.

Söderholm, S., Fu, Y., Gaelings, L., Belanov, S., Yetukuri, L., Berlinkov, M., Cheltsov, A. V., Anders, S., Aittokallio, T., Nyman, T. A., Matikainen, S., and Kainov, D. E. (2016). Multi-Omics studies towards novel modulators of influenza a Virus-Host interaction. *Viruses*, 8(10).

Solnica-Krezel, L. and Sepich, D. S. (2012). Gastrulation: making and shaping germ layers. *Annu. Rev. Cell Dev. Biol.*, 28:687–717.

Sommer, C., Straehle, C., Köthe, U., and Hamprecht, F. A. (2011). Ilastik: Interactive learning and segmentation toolkit. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 230–233.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):91.

Spall, J. C. (2003). *Introduction to stochastic search and optimization: estimation, simulation, and control*. J. Wiley, Hoboken, N.J.

Speed, D. and Balding, D. J. (2015). Relatedness in the post-genomic era: is it still useful? *Nat. Rev. Genet.*, 16(1):33–44.

Spruance, S. L., Reid, J. E., Grace, M., and Samore, M. (2004). Hazard ratio in clinical trials. *Antimicrob. Agents Chemother.*, 48(8):2787–2792.

Srivastava, P. (2002). Roles of heat-shock proteins in innate and adaptive immunity. *Nat. Rev. Immunol.*, 2(3):185–194.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., and Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, 7(3):500–507.

Strell, C., Hilscher, M. M., Laxman, N., Svedlund, J., Wu, C., Yokota, C., and Nilsson, M. (2018). Placing RNA in context and space - methods for spatially resolved transcriptomics. *FEBS J.*

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550.

Sun, X., Meyers, E. N., Lewandoski, M., and Martin, G. R. (1999). Targeted disruption of fgf8 causes failure of cell migration in the gastrulating mouse embryo. *Genes Dev.*, 13(14):1834–1846.

Surguchov, A., Palazzo, R. E., and Surgucheva, I. (2001). Gamma synuclein: subcellular localization in neuronal and non-neuronal cells and effect on signal transduction. *Cell Motil. Cytoskeleton*, 49(4):218–228.

Suvitaival, T., Parkkinen, J. A., Virtanen, S., and Kaski, S. (2014). Cross-organism toxicogenomics with group factor analysis. *Systems Biomedicine*, 2(4):71–80.

Svensson, V., Teichmann, S. A., and Stegle, O. (2018a). SpatialDE: identification of spatially variable genes. *Nat. Methods*, 15(5):343–346.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018b). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.*, 13:599.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., and von Mering, C. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.*, 45(D1):D362–D368.

Takahashi, Y., Inoue, T., Gossler, A., and Saga, Y. (2003). Feedback loops comprising dll1, dll3 and mesp2, and differential involvement of psen1 are essential for rostrocaudal patterning of somites. *Development*, 130(18):4259–4268.

Tam, P. P. and Behringer, R. R. (1997). Mouse gastrulation: the formation of a mammalian body plan. *Mech. Dev.*, 68(1-2):3–25.

Tam, P. P. L. and Loebel, D. A. F. (2007). Gene function in mouse embryogenesis: get set for gastrulation. *Nat. Rev. Genet.*, 8(5):368–381.

Tanaka, T. (1999). A theory of mean field approximation. In Kearns, M. J., Solla, S. A., and Cohn, D. A., editors, *Advances in Neural Information Processing Systems 11*, pages 351–360. MIT Press.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., and Surani, M. A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell*, 6(5):468–478.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, 6(5):377–382.

Tarca, A. L., Romero, R., and Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. *Am. J. Obstet. Gynecol.*, 195(2):373–388.

Taub, F. E., DeLeo, J. M., and Thompson, E. B. (1983). Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs. *DNA*, 2(4):309–327.

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569–583.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257.

Tenenhaus, A. and Tenenhaus, M. (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *Eur. J. Oper. Res.*, 238(2):391–403.

The Tabula Muris Consortium, Quake, S. R., Wyss-Coray, T., and Darmanis, S. (2017). Transcriptomic characterization of 20 organs and tissues from mouse at single cell resolution creates a tabula muris. *bioRxiv*, page 237446.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 58(1):267–288.

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, 61(3):611–22.

Titsias, M. and Lawrence, N. D. (2010). Bayesian Gaussian process latent variable model. *Journal of Machine Learning Research*.

Titsias, M. K. and Lázaro-Gredilla, M. (2011). Spike and slab Variational inference for Multi-Task and multiple kernel learning. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2339–2347. Curran Associates, Inc.

Trachootham, D., Alexandre, J., and Huang, P. (2009). Targeting cancer cells by ROS-mediated mechanisms: a radical therapeutic approach? *Nat. Rev. Drug Discov.*, 8:579.

Trojani, A., Di Camillo, B., Tedeschi, A., Lodola, M., Montesano, S., Ricci, F., Vismara, E., Greco, A., Veronese, S., Orlacchio, A., Martino, S., Colombo, C., Mura, M., Nichelatti, M., Colosimo, A., Scarpati, B., Montillo, M., and Morra, E. (2011). Gene expression profiling identifies ARSD as a new marker of disease progression and the sphingolipid metabolism as a potential novel metabolism in chronic lymphocytic leukemia. *Cancer Biomark.*, 11(1):15–28.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods*, 13(12):966–967.

Van Dam, S., Võsa, U., Van der Graaf, A., Franke, L., and de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.*, 19(4):575–592.

Van den Boom, V., Kooistra, S. M., Boesjes, M., Geverts, B., Houtsmuller, A. B., Monzen, K., Komuro, I., Essers, J., Drenth-Diephuis, L. J., and Eggen, B. J. L. (2007). UTF1 is a chromatin-associated protein involved in ES cell differentiation. *J. Cell Biol.*, 178(6):913–924.

Van der Maaten, L. J. P., Postma, E. O., and Van den Herik, H. J. (2008). Dimensionality reduction: A comparative review. *citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.112.5472*.

Vargas, K. J., Schrod, N., Davis, T., Fernandez-Busnadiego, R., Taguchi, Y. V., Laugks, U., Lucic, V., and Chandra, S. S. (2017). Synucleins have multiple effects on presynaptic architecture. *Cell Rep.*, 18(1):161–173.

Varol, C., Mildner, A., and Jung, S. (2015). Macrophages: Development and tissue specialization. *Annual Review of Immunology*, 33(1):643–675. PMID: 25861979.

Vasconcelos, Y., De Vos, J., Vallat, L., Rème, T., Lalanne, A. I., Wanherdrick, K., Michel, A., Nguyen-Khac, F., Oppezzo, P., Magnac, C., Maloum, K., Ajchenbaum-Cymbalista, F., Troussard, X., Leporrier, M., Klein, B., Dighiero, G., Davi, F., and French Cooperative Group on CLL (2005). Gene expression profiling of chronic lymphocytic leukemia can discriminate cases with stable disease and mutated ig genes from those with progressive disease and unmutated ig genes. *Leukemia*, 19(11):2002–2005.

Virtanen, S., Klami, A., Khan, S., and Kaski, S. (2012). Bayesian group factor analysis. *Artificial Intelligence and Statistics*.

Vitner, E. B., Dekel, H., Zigdon, H., Shachar, T., Farfel-Becker, T., Eilam, R., Karlsson, S., and Futerman, A. H. (2010). Altered expression and distribution of cathepsins in neuronopathic forms of gaucher disease and in other sphingolipidoses. *Hum. Mol. Genet.*, 19(18):3583–3590.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, 34(11):1145–1160.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and Variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, 11(3):333–337.

Wang, M., Yang, X., Wang, F., Li, R., Ning, H., Na, L., Huang, Y., Song, Y., Liu, L., Pan, H., Zhang, Q., Fan, L., Li, Y., and Sun, C. (2013). Calcium-deficiency assessment and biomarker identification by an integrated urinary metabonomics analysis. *BMC Med.*, 11:86.

Wang, Y. and Navin, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Mol. Cell*, 58(4):598–609.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63.

Wang, Z.-W. (2008). Regulation of synaptic transmission by presynaptic CaMKII and BK channels. *Mol. Neurobiol.*, 38(2):153–166.

Weidinger, G., Thorpe, C. J., Wuennenberg-Stapleton, K., Ngai, J., and Moon, R. T. (2005). The sp1-related transcription factors sp5 and sp5-like act downstream of wnt/beta-catenin signaling in mesoderm and neuroectoderm patterning. *Curr. Biol.*, 15(6):489–500.

Wen, J., Zeng, Y., Fang, Z., Gu, J., Ge, L., Tang, F., Qu, Z., Hu, J., Cui, Y., Zhang, K., Wang, J., Li, S., Sun, Y., and Jin, Y. (2017). Single-cell analysis reveals lineage segregation in early post-implantation mouse embryos. *J. Biol. Chem.*, 292(23):9840–9854.

Widmer, C., Lippert, C., Weissbrod, O., Fusi, N., Kadie, C., Davidson, R., Listgarten, J., and Heckerman, D. (2014). Further improvements to linear mixed models for Genome-Wide association studies. *Sci. Rep.*, 4:6874.

Wilks, S. S. (1938). The Large-Sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, 9(1):60–62.

Wipf, D. P. and Nagarajan, S. S. (2008). A new view of automatic relevance determination. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1625–1632. Curran Associates, Inc.

Xing, E. P. (2015). Advanced Gaussian processes. [online] https://www.cs.cmu.edu/~epxing/Class/10708-15/notes/10708_scribe_lecture21.pdf.

Yang, H., Harrington, C. A., Vartanian, K., Coldren, C. D., Hall, R., and Churchill, G. A. (2008). Randomization in laboratory procedure is key to obtaining reproducible microarray results. *PLoS One*, 3(11):e3724.

Yang, J., Penfold, C. A., Grant, M. R., and Rattray, M. (2016). Inferring the perturbation time from biological time course data. *Bioinformatics*, 32(19):2956–2964.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Res.*, 44(D1):D710–6.

Yi, F., Catudio-Garrett, E., Gábriel, R., Wilhelm, M., Erdelyi, F., Szabo, G., Deisseroth, K., and Lawrence, J. (2015). Hippocampal "cholinergic interneurons" visualized with the choline acetyltransferase promoter: anatomical distribution, intrinsic membrane properties, neurochemical characteristics, and capacity for cholinergic modulation. *Front. Synaptic Neurosci.*, 7:4.

Yoon, Y., Huang, T., Tortelote, G. G., Wakamiya, M., Hadjantonakis, A.-K., Behringer, R. R., and Rivera-Pérez, J. A. (2015). Extra-embryonic wnt3 regulates the establishment of the primitive streak in mice. *Dev. Biol.*, 403(1):80–88.

Yoshida, R. and West, M. (2010). Bayesian learning in sparse graphical factor models via Variational Mean-Field annealing. *J. Mach. Learn. Res.*, 11:1771–1798.

Yoshikawa, F., Sato, Y., Tohyama, K., Akagi, T., Hashikawa, T., Nagakura-Takagi, Y., Sekine, Y., Morita, N., Baba, H., Suzuki, Y., Sugano, S., Sato, A., and Furuichi, T. (2008). Opalin, a transmembrane sialylglycoprotein located in the central nervous system myelin paranodal loop membrane. *J. Biol. Chem.*, 283(30):20830–20840.

Zenz, T., Mertens, D., Küppers, R., Döhner, H., and Stilgenbauer, S. (2010). From pathogenesis to treatment of chronic lymphocytic leukaemia. *Nat. Rev. Cancer*, 10(1):37–50.

Zhao, Q., Zhang, L., and Cichocki, A. (2014). Bayesian CP factorization of incomplete tensors with automatic rank determination.

Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. (2016). Bayesian group factor analysis with structured sparsity. *J. Mach. Learn. Res.*, 17(196):1–47.