Statistical framework for the integration of single-cell multi-omics data sets

Ricard Argelaguet

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Robinson College

March 2019

Abstract

Single-cell profiling techniques have provided an unprecented opportunity to study cellular heterogeneity at multiple molecular levels. The maturation of single-cell RNA-sequencing technologies has enabled the identification of transcriptional profiles associated with lineage diversification and cell fate commitment. This represents a remarkable advance over traditional bulk sequencing methods, particularly for the study of complex and heterogeneous biological processes, including the immune system, embryonic development and cancer. However, the accompanying epigenetic changes and the role of other molecular layers in driving cell fate decisions remains poorly understood. The profiling the epigenome at the single-cell level is receiving increasing attention. However, without associated transcriptomic readouts, the conclusions that can be extracted from epigenetic measurements are limited.

More recently, technological advances enabled multiple biological layers to be probed in parallel at the single-cell level, unveling a powerful approach for investigating regulatory relationships. Such single-cell multi-modal technologies can reveal multiple dimensions of cellular heterogeneity and uncover how this variation is coupled between the different molecular layers, hence enabling a more profound mechanistic insight than can be inferred by analysing a single data modality in separate cells. Yet, multi-modal sequencing protocols face multiple challenges, both from the experimental and the computational front.

In this thesis we propose an experimental methodology and a computational framework for the integrative study of multiple omics in single cells.

The first contribution of this thesis is Nucleosome, Methylome and Transcriptome sequencing (scNMT-seq), a multi-modal single-cell sequencing protocol for profiling RNA expression, DNA methylation and chromatin accessibility in single cells. scNMT-seq provides genome-wide epigenetic readouts at a base-pair resolution, hence expanding our ability to investigate the dynamics of the epigenome across cell fate transitions.

The second contribution of this thesis is Multi-Omics Factor Analysis (MOFA), a statistical framework for the unsupervised integration of large-scale multi-omics data sets. MOFA aims at discovering the principal sources of variation while disentangling the axes of heterogeneity that are shared across multiple modalities from those specific to individual data modalities. This framework enables the unbiased interrogation of large (single-cell) data sets simultaneously across multiple data modalities and across different experiments or conditions.

The third contribution of this thesis is generation of an epigenetic roadmap of mouse gastrulation, resulting from the combined use of scNMT-seq and MOFA. Notably, we show that regulatory elements associated with the formation of the three germ layers are either epigenetically primed or epigenetically remodelled prior to overt cell fate decisions, providing the molecular logic for a hierarchical emergence of the primary germ layers.

0.1 Probabilistic modelling

A scientific model is a simple theoretical representation of a complex natural phenomenon to allow the systematic study of its behaviour. The general idea is that if a model is able to explain the current observations, it might be capturing its true underlying laws and can therefore be used to make future predictions.

Statistical models are a powerful abstraction of nature. They consist on a set of observed variables and a set of (hidden) parameters. The procedure of fitting the parameters to explain the observations is called inference or learning.

The major challenge of inference in dealing with real data stes is the disentanglement of signal and noise. An ideal model should learn only the information relevant to gain explanatory power while disregarding the noise. However, this is non-trivial in most practical situations. Very complex models will tend to overfit the training data, thereby capturing large amounts of noise and leading to a bad generalisation performance to independent data sets. On the other hand, simplistic models will fit the data poorly, leading to poor explanatory power.

The ideas above can be formalised using the framework of probability and statistics.

0.1.1 Maximum likelihood inference

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

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

This process is called maximum likelihood learning. However, in this setting there is no penalisation for model complexity, and maximum likelihood solutions are prone to overfit in cases where the data is relatively sparse **?**. Generalisations that account for model complexity have been proposed and include regularising terms that shrink parameters to small values. However, these are often particular cases of the more general framework of Bayesian statistics **?**.

0.1.2 Bayesian inference

In the Bayesian framework, the parameters themselves are treated as random variables and we aim to obtain probably distributions for Θ, rather than a single point estimate. To do so, prior beliefs are introduced into the model by specifying a prior probability distribution $p(\Theta)$. Then, using Bayes' theorem **??**, the prior hypothesis is updated based on the observed data $\mathbf{Y}$ by means of the likelihood $p(\mathbf{Y}|\Theta)$ function, which yields a posterior distribution over the parameters:

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

where $p(\mathbf{Y})$ is a constant term called the marginal likelihood, or model evidence.

The choice of the prior distribution is a key part of Bayesian inference and captures beliefs about the distribution of a variable before observing the data. With asymptotically large sample sizes, the choice of prior has negligible effects on the posterior estimates, but it becomes critical with sparse data **??**. There are two common considerations to define the prior distributions. The first concerns the incorporation of subjective information, or predefined assumptions, into the model. An example is the use of sparsity assumptions on the loadings, as in [[SECTION X]]. The second one is based on convenient mathematical properties to make inference tractable. In particular, if the likelihood and the prior distributions do not belong to the same family of probability distributions (they are not conjugate) then inference becomes more problematic and one has to resort to numerical approaches **?**. An example is the Automatic Relevance Determination prior as discussed in [[SECTION X]].

Again, the milestone of Bayesian inference is that an entire posterior probability distribution is obtained for each parameter. This has the clear advantage of naturally handling uncertainity in the estimation of parameters. Hence, when making predictions, a fully Bayesian approach attempts to integrate over all the possible values of all uncertain quantitites. Nevertheless, this calculation is often intractable and one has to resort to point estimates **??**. The simplest approximation to the posterior distribution is to use its mode, which leads to the maximum a posteriori (MAP) estimate:

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

This is similar to the maximum likelihood objective function, but with the addition of a term $p(\Theta)$, that penalises for model complexity. Therefore, in contrast to maximum likelihood inference, Bayesian approaches naturally handle the problem of model complexity and overfitting**??**.

0.1.3 Sampling approaches: Markov Chain Monte Carlo and Gibbs Sampling

The central task in Bayesian inference is the direct evaluation of the posterior distributions and/or the computation of expectations with respect to the posterior distributions. In sufficiently complex models, closed-form solutions are not available and one has to resort to approximation schemes, which broadly fall into two classes: stochastic or deterministic **?**. Stochastic approaches hinge on the generation of samples from the posterior distribution via a Markov Chain Monte Carlo (MCMC) framework. Such techniques have the appealing property of exact results at the asymptotic limit of infinite computational resources. However, in practice, sampling approaches are computationally demanding and suffer from limited scalability to large data sets **?**.

0.1.4 Deterministic approaches: Variational Bayesian inference

Deterministic approaches are based on analytical simplified approximations to the posterior distribution. As a result, deterministic approaches are usually much faster and scale to large applications, at the expense of biased results **??**. Variational inference is a deterministic technique that has been receiving widespread attention due to a positive balance between accuracy, speed, and ease of use **??**.

The key idea behind variational inference is to approximate a posterior distribution using a (variational) distribution with free (variational) parameters by minimising their Kullback-Leibler divergence. Different types of variational inference exist depending on how the family of variational distributions is chosen, and how the optimisation is performed **?**. The basic framework is derived below.

Consider a probabilistic model where all the observed variables are collectively denoted as **Y** and all the hidden variables are denoted by **X**. In variatonal inference, the true (but complex) posterior distribution

$p(\mathbf{X}|\mathbf{Y})$ is approximated via a simpler (variational) distribution $q(\Theta)$, parameterized by some variational parameters $\boldsymbol{\theta}$. The variational optimisation finds the settings of parameters that minimise the Kullback-Leibler divergence with the true posterior:

$$\mathrm{KL}q(\mathbf{X}|\theta)p(\mathbf{Y}|\mathbf{X}) = \int q(\mathbf{X}|\theta)\log\frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X}|\theta)}d\mathbf{X}$$

This integral is intractable as it requires knowing the true posterior. However, this expression can be rearranged to:

$$\mathrm{KL}q(\mathbf{X}|\theta)p(\mathbf{Y}|\mathbf{X}) = -\int q(\mathbf{X}|\theta)\log\frac{p(\mathbf{Y},\mathbf{X})}{q(\mathbf{X}|\theta)}d\mathbf{X} + \log p(\mathbf{Y})$$

The KL divergence decomposes into two terms. The second term is the marginal log likelihood $\log p(\mathbf{Y})$, which does not depend on the parameters. The first term is an expectation with respect to the variational distributions $q(\mathbf{X}|\theta)$ that involves the complete-data log likelihood $\log P(\mathbf{Y},\mathbf{X})$, which is tractable, given the right choice of priors and variational distributions, as discussed below.

As the first term is the only one that depends on the parameters θ, we infer that increasing the first term (by changing the parameters accordingly) is an approach to minimise the KL divergence. By rearranging we can see that this term acts as a lower bound to the log marginal likelihood [[FIGURE X]] and is consequently called the Evidence Lower Bound (ELBO) $\mathscr{L}(\mathbf{X})$:

$$\mathscr{L}(\mathbf{X}|\theta) = \log p(\mathbf{Y}) - \mathrm{KL}(q(\mathbf{X}|\theta)||p(\mathbf{X}|\mathbf{Y}))$$

The ELBO, which acts as the cost function to be maximised, can be further decomposed into two terms [[EQXX]]:

$$\mathscr{L}(\mathbf{X}|\theta) = E_{q(\mathbf{X}|\theta)}[\log p(\mathbf{Y}|\mathbf{X})] - \mathrm{KL}(q(\mathbf{X}|\theta)||p(\mathbf{X}))$$

The first term corresponds to the expectation of the log likelihood under the variational distribution, and it measures the goodness of fit . The second contribution is the KL divergence between the prior and the posterior distribution of the parameters. If priors are chosen to promote sparsity, this term acts as a regularisarer to prevent overfitting.

To conclude the derivation above, variational inference involves optimising $\mathscr{L}(\mathbf{X}|\theta)$. If we allow any possible choice of $q(\mathbf{X})$, then the maximum of the lower bound $\mathscr{L}(\mathbf{X})$ occurs when the KL divergence vanishes and $q(\mathbf{X})$ equals the true posterior distribution $p(\mathbf{X}|\mathbf{Y})$. As the posterior is intractable and this does not lead to any simplification of the problem, it is necessary to consider a restricted and tractable family of variational distributions and seek the member of this family for which the KL divergence is minimised **?**.

There are a myriad of assumptions and simplifications placed on the form of the $q(\mathbf{X}|\theta)$ distribution, for example by assuming a parametric form **?** and/or by factorising over the hidden variables **?**. When the the variational distribution of the hidden variables is in the same family as the prior, given the markov blanket (i.e. conditional conjugacy), then a closed form coordinate-ascent algorithm can be derived **??**, as we show below.

In this thesis we considered the mean-field approximation **???**, a factorisation over M disjoint groups of hidden variables:

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

Evidently, this family of distributions do not usually contain the true posterior, but this is a key assumption to obtain an analytical inference scheme with no further assumptions on the $q(\mathbf{x}_i)$ distributions. Extensions where dependencies between variables are taken into account have been considered **???**, which improve the fidelity of the approximation at the cost of a harder optimisation problem.
Applying the mean-field assumption to [[EQ XX]]:
With the key assumption made, we need to find the specific distributions within the family that maximise the lower bound $\mathscr{L}(\mathbf{X}|\theta)$. It can be shown using calculus of variations **?**, that the optimal distribution $\hat{q}_i$ for each variable $\mathbf{x}_i$, is the following:

$$\log \hat{q}_i(\mathbf{x}_i) = \mathbb{E}_{i \neq j}[\log p(\mathbf{Y}, \mathbf{X})] + \text{const} \tag{1}$$

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

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

In practice it is easier to work in terms of **??** and infer the constant by inspection, as the rest of the expression can usually be recognised as being a known type of distribution. This becomes clear in **??**.