Statistical methods for the integrative analysis of single-cell multi-omics data

Ricard Argelaguet

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Robinson College

March 2020

Chapter 1

MOFA+: an improved framework for the comprehensive integration of structured single-cell data

In Chapter 2 we developed Multi-Omics Factor Analysis (MOFA), a statistical framework for the unsupervised integration of multi-omics data.
MOFA addresses key challenges in multi-omics data integration, including overfitting, noise reduction, handling of missing values and improved interpretation of the model output. Hower, when applied to increasingly-large (single-cel) data sets, the variational inference scheme implemented in MOFA has limited scalability.

In addition, the increased experimental throughput has facilitated the study of larger numbers of experimental conditions. MOFA makes strong assumptions about the dependencies across samples and it hence has no principled way of modelling data sets where the samples are structured into multiple groups, where groups can be defined as batches, donors or different experiments. By pooling and contrasting information across studies or experimental conditions, it would be possible to obtain more comprehensive insights into the complexity underlying biological systems.

In this new study we improve the first model formulation with the aim of performing integrative analysis of large-scale datasets simultaneously across multiple data modalities and across multiple groups.

1.1 Theoretical fundations

1.1.1 Stochastic gradient ascent

Gradient ascent is a common first-order optimization algorithm for finding the maximum of a function [**Bishop2006**, **Murphy**]. It works iteratively by taking steps proportional to the gradient of the function evaluated at each iteration. Formally, for a differentiable function $F(x)$, the iterative

scheme of gradient ascent is:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \rho^{(t)} \nabla F(\mathbf{x}^{(t)}) \tag{1.1}$$

At each iteration, the gradient ∇F is re-evaluated and a step is performed towards its direction. The step size is controlled by $\rho^{(t)}$, a parameter called the learning rate, which is typically adjusted at each iteration t [**Ranganath2013**].

Gradient ascent is appealing because of its simplicity, but it becomes prohibitively slow with large datasets, mainly because of the computational cost (both in terms of time and memory) associated with the iterative calculation of gradients [**Spall2003**].

Assuming the existence of redundancy in the dataset, a fast approximation of the gradient $\hat{\nabla} F$ can be calculated using a random subset of the data (minibatch). Formally, as in standard gradient ascent, the iterative training schedule proceeds by taking steps of size ρ in the direction of the approximate gradient $\hat{\nabla} F$:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \rho^{(t)} \hat{\nabla} F(\mathbf{x}^{(t)}) \tag{1.2}$$

The step size $\rho^{(t)}$ can also be adjusted at each iteration t. When the series $\rho(t)$ satisfies the Robbins-Monro conditions: $\sum_t \rho^{(t)} = \infty$ and $\sum_t (\rho^{(t)})^2 < \infty$, F is guaranteed to converge to a local maximum [**Robbins-Monro1951**]. If F is not convex, the algorithm is sensible to the initialisation $\mathbf{x}^{t=0}$.

1.1.2 Natural gradient ascent

Let us consider a model with a single hidden variable x and corresponding variational parameter θ. The objective function is the ELBO $\mathcal{L}(\theta)$. From the definition of the gradient:

$$\nabla \mathcal{L}(\theta) = \lim_{||h|| \to 0} \frac{\mathcal{L}(\theta + h) - \mathcal{L}(\theta)}{||h||}$$

where h represents an infinitesimally small positive step in the space of θ.
To find the steepest gradient, one would need to search over all possible directions d in an infinitely small distance h, and select the $\hat{d}$ with the largest gradient:

$$\nabla \mathcal{L}(\theta) = \lim_{h \to 0} \frac{1}{h} \arg\max_{d \, s.t. \|d\| = h} \mathcal{L}(\theta + d) - \mathcal{L}(\theta)$$

Notice that the neighborhood of θ is measured in terms of its Euclidean norm, and the direction of steepest ascent is hence dependent on the Euclidean geometry of the θ space. This is the key problematic issue when working with probability distributions. A small step from $\theta^{(t)}$ to $\theta^{(t+1)}$ does not guarantee an equivalently small change from $\mathcal{L}(\theta^{(t)})$ to $\mathcal{L}(\theta^{(t+1)})$. As an example, consider four random variables:

$$\begin{aligned} \psi_1 &\sim \mathcal{N}(0 \,|\, 5) \qquad \psi_3 \sim \mathcal{N}(0 \,|\, 1) \\ \psi_2 &\sim \mathcal{N}(10 \,|\, 5) \qquad \psi_4 \sim \mathcal{N}(10 \,|\, 1) \end{aligned} \tag{1.3}$$

Using the Euclidean metric, the distance between ψ_1 and ψ_2 is the same as the distance between ψ_3 and ψ_4. However, the distance in distribution space (measured for example by the KL divergence) is clearly much larger between ψ_1 and ψ_2 than between ψ_3 and ψ_4 (**??**).

Hence, rather than using the Euclidean distance, it is more appropriate to use the KL divergence as a distance metric:

$$\nabla_{KL}\mathcal{L}(\theta) = \lim_{h\to 0} \frac{1}{h} \underset{d\, s.t. KL[p_\theta || p_{\theta+d}]=h}{\arg\max} \mathcal{L}(\theta + d) - \mathcal{L}(\theta)$$

The direction of steepest ascent measured by the KL divergence is called the natural gradient [**Amari1998**, **Martens2014**].
By introducing Lagrange multipliers and Taylor expansions, one can solve the optimisation problem to obtain the direction of the steepest natural gradient (see [**Amari1998**, **Kristiadi2019**]). The solution corresponds to the standard (Euclidean) gradient pre-multiplied by the inverse of the Fisher Information Matrix of $q(x|\theta)$:

$$\hat{d}_{KL} \propto \mathbf{F}^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta) \tag{1.4}$$

where $\mathbf{F}(\theta)$ is defined as

$$\mathbf{F}(\theta) = \mathbb{E}_{q(x|\theta)}[(\nabla_\theta \log q(x|\theta))(\nabla_\theta \log q(x|\theta))^T]$$

Effectively, the premultiplication by $\mathbf{F}^{-1}$ takes into account the local curvate of $q(\theta)$ in distribution space.
Importantly, when $q(x|\theta)$ belongs to the exponential family, the Fisher Information matrix is simply the Hessian of the log normalizer.

In conclusion, while the standard gradient points to the direction of steepest ascent in Euclidean space, the natural gradient points to the direction of steepest ascent in a space where distances are defined by the KL divergence [**Kristiadi2019**, **Amari1998**, **Hoffman2012**].

1.2 Model description

This section is reproduced from [**Argelaguet2018**] with some modifications.

1.2.1 Multi-group inference

1.2.2 GPU-accelerated stochastic variational inference

1.3 Model validation

1.4 Applications

1.4.1 Integration of a heterogeneous time-course single-cell RNA-seq dataset

1.4.2 Identification of context-dependent methylation signatures associated with cellular diversity in the mammalian cortex

1.4.3 Identification of molecular signatures of lineage commitment during mammalian embryogenesis