Bayesian latent variable models for single-cell trajectory learning

Kieran Campbell & Christopher Yau
University of Oxford

May 5, 2016

Background

The transcriptomes of single cells undergoing diverse biological processes - such as differentiation or apoptosis - display remarkable heterogeneity that is averaged over in bulk sequencing. Single-cell sequencing itself offers only a snapshot of these processes by capturing cells of variable and unknown progression through them. Consequently, one outstanding problem in single-cell genomics is to find an ordering of cells (known as their pseudotime) that best reflects their progression, for which several computational methods have been proposed.

To date, the vast majority of such methods emphasise transcriptome-wide 'data-driven' approaches that assume no prior knowledge of gene dynamics along the trajectory during inference. The suitability of the inferred trajectory is typically assessed by post-hoc examination of a set of marker genes to ensure the inferred behaviour aligns with prior assumptions. Furthermore, most current methods are algorithmic and rely on heuristics as opposed to probabilistic models, which in the context of bifurcations requires the pseudotimes to be first inferred prior to the identification of any bifurcation events.

Results

Here we introduce a general probabilistic framework for single-cell trajectory learning based on Bayesian non-linear factor analysis and apply it to two outstanding problems in single-cell analysis. Firstly, we demonstrate how such a framework may be used to integrate prior knowledge of gene behaviour in trajectory inference. By assuming a parametric form of gene expression evolution across pseudotime we can place informative priors on parameters that govern gene behaviour within a Bayesian statistical framework. Consequently, we remove the need for subjective post-inference checks and simultaneously solve related problems such as trajectory orientation and setting implicit length scales. We demonstrate how using such methods only a small panel of marker genes are required to achieve comparable results to transcriptome wide 'data-driven' alternatives. We further demonstrate how such a method can be used to recover trajectories corresponding to known pathways in the presence of heavily confounding effects.

The second application of our framework is to modelling bifurcations in single-cell data. By considering a Bayesian mixture of factor analysers we simultaneously infer both the pseudotimes and branching behaviour of the cells, which is unique compared to existing methods. We derive a Gibbs sampler that allows for fast inference across hundreds of cells while accounting for the zero inflation that is pertinent to single-cell RNA-seq data. Notably, by using a Bayesian framework we can integrate prior knowledge of branch-specific gene behaviour allowing for robust inference on challenging datasets.

Conclusions

We introduce a flexible Bayesian framework that solves several outstanding issues in single-cell trajectory learning. This framework uniquely provides a principled method for integrating prior knowledge of gene behaviour along single-cell trajectories and allows for such trajectories to be learned from a

small panel of marker genes. We also introduce the first statistical method for bifurcation inference that simultaneously infers both the pseudotimes of the cells as well as the bifurcation events, providing robust trajectories as well as full uncertainty estimates. We apply our methods to a range of both synthetic and real data, and more generally discuss the challenges of single-cell latent variable modelling including the connection of principal component analysis to both pseudotime inference and dropout rate. We conclude by motivating why such methods can be applied to a wide range of 'omics' data including modelling cancer progression and patient treatment outcomes.