Bayesian Gaussian Process Latent Variable Models for pseudotime inference in single-cell RNA-seq data

Kieran Campbell[1] and Christopher Yau[2,3]

[1]MRC Functional Genomics Unit, University of Oxford, UK
[2]Wellcome Trust Centre for Human Genetics, University of Oxford, UK
[3]Department of Statistics, University of Oxford, UK

September 15, 2015

Abstract

Single-cell genomics has revolutionised modern biology while requiring the development of advanced computational and statistical methods. Advances have been made in uncovering gene expression heterogeneity, discovering new cell types and novel identification of genes and transcription factors involved in cellular processes. One such approach to the analysis is to construct pseudotime orderings of cells as they progress through a particular biological process, such as cell-cycle or differentiation. These methods assign a score - known as the pseudotime - to each cell as a surrogate measure of progression. However, all published methods to date are purely algorithmic and lack any way to give uncertainty to the pseudotime assigned to a cell. Here we present a method that combines Gaussian Process Latent Variable Models (GP-LVM) with a recently published electroGP prior to perform Bayesian inference on the pseudotimes. We go on to show that the posterior variability in these pseudotimes leads to nontrivial uncertainty in the pseudo-temporal ordering of the cells and that pseudotimes should not be thought of as point estimates.

1 Introduction

Single-cell RNA-seq (scRNA-seq) has emerged as a powerful method for the quantification of gene and transcript abundance in individual cells. In only a few years it has uncovered exciting new biology including the identification of novel cell types [1], hidden heterogeneity in gene expression [2] and regulatory networks operating at the single-cell level [3]. It has particular advantages over bulk RNA sequencing such as the ability to identify rare cell types and cell-to-cell variability [4, 5], which are typically hidden in bulk analyses.

Despite being simultaneously sequenced individual cells may be of variable progression through a variety of cellular processes due to heterogeneous responses to stimuli and inherent transcriptomic variability. This has consequently lead to the idea of pseudotime as an artificial measure of a cell's progression through a process such as differentiation or apoptosis [6]. The statistical problem is to assign a pseudotime label between 0 and 1 to each observation where values near 0 indicates that the cell is in a state near the start of the biological process and values near 1 denote cells that are toward the end of the process.

Early attempts and pseudotime ordering algorithms include *Monocle*[6], which uses Independent Component Analysis (ICA) with Minimum Spanning Trees (MST) applied to scRNA-seq data, *Wanderlust* [7], which uses ensembles of k-nearest-neighbour graphs applied to mass spectrometry data and *embeddr* [8], which uses Laplacian Eigenmaps and principal curves to identify pseudotime trajectory using nonparametric curve fitting. However, a common feature of all these algorithms is that they provide point estimates of the pseudotime for each cell and are unable to quantify the uncertainty in each estimate[1]. As a result, the differential expression of particular genes over pseudotime may simply be an artefact of that ordering and disappear when the uncertainty is taken into account.

Here we present an innovative approach that uses full Bayesian inference for Gaussian Process Latent Variable Models [10] combined with a repulsive prior [11] to infer posterior distributions of pseudotimes. First, a reduced dimension representation of the cells from a dimensionality reduction algorithm such as principal components analysis (PCA) or Laplacian Eigenmaps [12] is created. Subsequently, a probabilistic curve is fitted through the cells that jointly fits posterior distributions of pseudotimes over all cells. As a result the uncertainty in each pseudotime assignment can be accurately assessed as well as the extent to which the ordering of any two cells is robust against the inherent noise. Our method does not require cell capture times as a prerequisite. An implementation of our inference method is available in the `Julia` programming language at `http://github.com/kieranrcampbell/gpseudotime`.

2 Gaussian Processes for pseudo-time assignment

2.1 Gaussian Process Latent Variable Models

A real-valued stochastic process $\{\mu_t, t \in T\}$, where T is an index set, is a Gaussian Process (GP), if all its finite dimensional marginal distributions are multivariate Gaussian distributions. That is, for any given distinct values $t_1, \ldots, t_n$, the random vector $\boldsymbol{\mu} = (\mu_{t_1}, \ldots, \mu_{t_n}) \sim N(\mathbf{m}, \mathbf{K})$ where $\mathbf{m} \equiv \mathbb{E}[\boldsymbol{\mu}]$ and $\mathbf{K} \equiv \mathrm{cov}(\boldsymbol{\mu}, \boldsymbol{\mu})$.

Gaussian Processes allow us to define prior probability distributions over real-valued functions in Bayesian nonparametric modelling. A popular model exploiting this property is the Gaussian Process Latent Variable Model (GPLVM) [13]. GPLVMs are a family of non-parametric methods that define a distribution over functions μ linking a latent variable t to an output variable x, for example, through the relationship $x = \mu(t) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. If the prior distribution over the latent function μ is given by a Gaussian Process then, given a set of N observations $\mathbf{x} = \{x_{t_1}, \ldots, x_{t_N}\}$ and corresponding latent points $\mathbf{t} = \{t_1, \ldots, t_N\}$, the marginal distribution $p(\mathbf{x}|\mathbf{t})$ is multivariate Gaussian with mean vector $\mathbf{m}(\mathbf{t})$ and covariance matrix given by $\mathbf{K} + \sigma^2 \boldsymbol{I}_N$ where $K_{ij} = \kappa(t_i, t_j)$ for a kernel function κ and $\boldsymbol{I}_N$ is an $N \times N$ identity matrix [10]. The specification of the kernel function κ allows us to control the favourable features of the latent functions, e.g. smoothness.

2.2 Model Specification

In our problem, the data consists of N P-dimensional observation vectors $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ which are assumed to be conditionally independent given the latent, unobserved pseudo-times $\mathbf{t} = \{t_1, \ldots t_n\}$, $t_i \in (0, 1]$, a mean function $\boldsymbol{\mu}(t)$ and a observation covariance matrix $\boldsymbol{\Sigma}$. The data we begin with are 'features' measured across N cells that we wish to order. These features can either be genes of particular interest or co-ordinates of some previously applied manifold learning algorithm, such as Laplacian Eigenmaps [12]. Each dimension of the mean function μ_j is given an independent Gaussian Process prior with covariance function $K^{(j)}$.

[1] In theory a bootstrap measure of uncertainty could be assigned, though this is computationally expensive.

Our model is described succinctly in the following hierarchical representation:

$$\begin{aligned}
\mathbf{x}_i &\sim \mathrm{N}(\boldsymbol{\mu}(t_i), \boldsymbol{\Sigma}),\ i = 1, \ldots, N, \\
\mu_j &\sim \mathrm{GP}(0, K^{(j)}),\ j = 1, \ldots, P, \\
K^{(j)}(t, t') &= \exp(-\lambda_j (t - t')^2),\ j = 1, \ldots, P, \\
\boldsymbol{\Sigma} &= \mathrm{diag}(\sigma_1^2, \ldots, \sigma_P^2) \\
\lambda_j &\sim \mathrm{Exp}(\gamma),\ j = 1, \ldots, P, \\
\sigma_j &\sim \mathrm{InvGamma}(\alpha, \beta) \\
\mathbf{t} &\sim \mathrm{Corp}(r),
\end{aligned} \tag{1}$$

where in the last line we define a repulsive prior on the pseudo-times $\mathbf{t}$ based on the Coulomb repulsive process [11].

The Coulomb repulsive process models repulsions between adjacent points using a process inspired by physical models of electrostatic potentials. In our model, this has the effect of preferentially favouring pseudo-time configurations that 'fill out' the interval $(0, 1]$. For pseudo-times $t_1, \ldots, t_N$ the prior is given by

$$p(\mathbf{t}) = p(t_1, \ldots, t_n) \propto \prod_{m=1}^{N} \prod_{n=m+1}^{N} \sin^{2r} \left[\pi(t_m - t_n)\right] \tag{2}$$

where r is the repulsion parameter. Under this prior, as $|t_i - t_j|$ gets smaller, the probability decreases and there is zero probability that two pseudo-times will coincide. [11] demonstrated that embedding this process within a GPLVM gave superior results to standard alternatives such as uniform $t_i \sim U(0, 1)$ or normal priors $t_i \sim N(0, 1)$ by avoiding identifiability issues due to scale- and translational-invariance under these standard priors.

The likelihood of $\mathbf{X}$ given the latent pseudotimes $\mathbf{t}$ is conditionally independent across features [10] so we can write it as

$$p(\mathbf{X}|\mathbf{t}, \boldsymbol{\sigma}^2, \boldsymbol{\lambda}) = \prod_{j=1}^{P} p(\mathbf{x}_j|\mathbf{t}, \sigma_j^2, \lambda_j)$$

where

$$p(\mathbf{x}_j|\mathbf{t}, \sigma_j^2, \lambda_j) = \mathcal{N}\left(\mathbf{x}_j|\mathbf{0}, K^{(j)}(\lambda_j, \mathbf{t}) + \sigma_j^2 \mathbf{I}\right).$$

Therefore, we can write the joint posterior as

$$\begin{aligned}
p(\mathbf{t}, \boldsymbol{\lambda}, \boldsymbol{\sigma}^2|\mathbf{X}) \propto &\prod_{j=1}^{P} \mathcal{N}\left(\mathbf{x}_j|\mathbf{0}, K^{(j)}(\lambda_j, \mathbf{t}) + \sigma_j^2 \mathbf{I}\right) \\
&\times \prod_{m=1}^{N} \prod_{n=m+1}^{N} \sin^{2r}\left(\pi|t_m - t_n|\right) \pi(\boldsymbol{\lambda}) \pi(\boldsymbol{\sigma}).
\end{aligned} \tag{3}$$

Note that for maximum likelihood inference $\mathbf{t}$ will never leave its initial ordering as the prior goes to zero probability when $t_m = t_n$.

Interestingly, the parameter $\boldsymbol{\lambda}$ has an intuitive interpretation in the context of curve fitting. In a one-dimensional Gaussian Process regression setting, λ corresponds to the 'horizontal' length scale over which the function varies. Therefore, in the two-dimensional plane $|\boldsymbol{\lambda}|$ loosely corresponds to arc-length, with larger $|\boldsymbol{\lambda}|$ generating longer curves. We effectively regularize $\boldsymbol{\lambda}$ by placing an exponential prior on it, with larger γ corresponding to shorter curves passing through the manifold.

2.3 Statistical Inference

Statistical inference for GPLVMs is typically performed using approximate maximum *a posteriori* [11] or variational methods [10] but Markov Chain Monte Carlo (MCMC) approaches are also possible [14]. As our primary objective is to characterise full posterior uncertainty measures of pseudo-time, MCMC-based inference was a necessity.

We therefore adopted used a random-walk Metropolis-Hastings (MH) with normal proposals for the unknown parameters of the model $(\mathbf{t}, \boldsymbol{\lambda}, \boldsymbol{\sigma}^2)$:

$$\begin{aligned}\mathbf{t}_{i+1} &\sim \mathcal{N}_{[0,1]}(\mathbf{t}_i, \sigma_t^2 \boldsymbol{I})\\ \boldsymbol{\lambda}_{i+1} &\sim \mathcal{N}_{(0,\infty)}(\boldsymbol{\lambda}_i, \sigma_\lambda^2 \boldsymbol{I})\\ \boldsymbol{\sigma}_{i+1} &\sim \mathcal{N}_{(0,\infty)}(\boldsymbol{\sigma}_i, \sigma_\sigma^2 \boldsymbol{I})\end{aligned}$$

where $\mathcal{N}_{[a,b]}$ is the multivariate Gaussian distribution truncated on $[a, b]$. The marginal properties of the Gaussian Process meant that the latent functions μ could be integrated out as therefore we did not need to impute this infinite-dimensional quantity.

One difficulty with the proposed model is the extreme multi-modal nature of the posterior. Many local maxima exist in the prior alone due to the $N(N-1)/2$ values for which it vanishes whenever $t_i = t_j$. As such, setting initial values to avoid becoming stuck in local maxima is challenging[2]. However, we can exploit the *repulsive* nature of the prior, by initialising $t_i \sim U(\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon)$ for arbitrarily small ϵ. As the prior heavily discourages pseudo-times that are in close proximity, the prior pushes the pseudotimes apart like charged particles being repelled while the likelihood biases this in an order that is consistent with the data reflecting the *a posteriori* more likely pseudotime orderings. We call this a “Big Bang” initialisation using an analogy to the famous cosmological phenomenon. Note that in practice it is really the variance σ_t^2 in the proposal distribution for $\mathbf{t}$ that provides the spread of points at the first iteration rather than the value of ϵ itself, so any sensible $\epsilon < \sigma_t$ will work.

3 Results

We applied our method to two datasets: (i) a synthetic dataset generated from the model, and (ii) a Laplacian Eigenmaps representation of the *Monocle* [6] dataset of differentiating myoblasts.

3.1 Synthetic Data

We generated data from the Gaussian Process described in Equation 1. Specifically, we set $\boldsymbol{\lambda} = [1, 2]$, $\boldsymbol{\sigma} = [2 \times 10^{-3}, 2 \times 10^{-2}]$ and sampled $n = 100$ pseudotimes from U(0, 1). We performed MH inference as described in Section 2.3 using 2×10^5 iterations thinned by 100. We set $\epsilon = 10^{-6}$, $\sigma_t = 9 \times 10^{-3}$, $\sigma_\lambda = 5 \times 10^{-1}$ and $\sigma_\sigma = 5 \times 10^{-3}$. The hyper-parameters were set to $r = 10^{-3}$, $\alpha = \beta = \gamma = 1$. The results of inference on the synthetic data can be seen in Figure 1. Pseudotime estimates clearly converge to stable values that are close to the ‘true’ values and are almost always within the 95% highest probability density (HPD) credible interval.

[2] The local maxima of the posterior generally have geometrically intuitive interpretations such as the posterior mean curve folding back on itself one or more times.

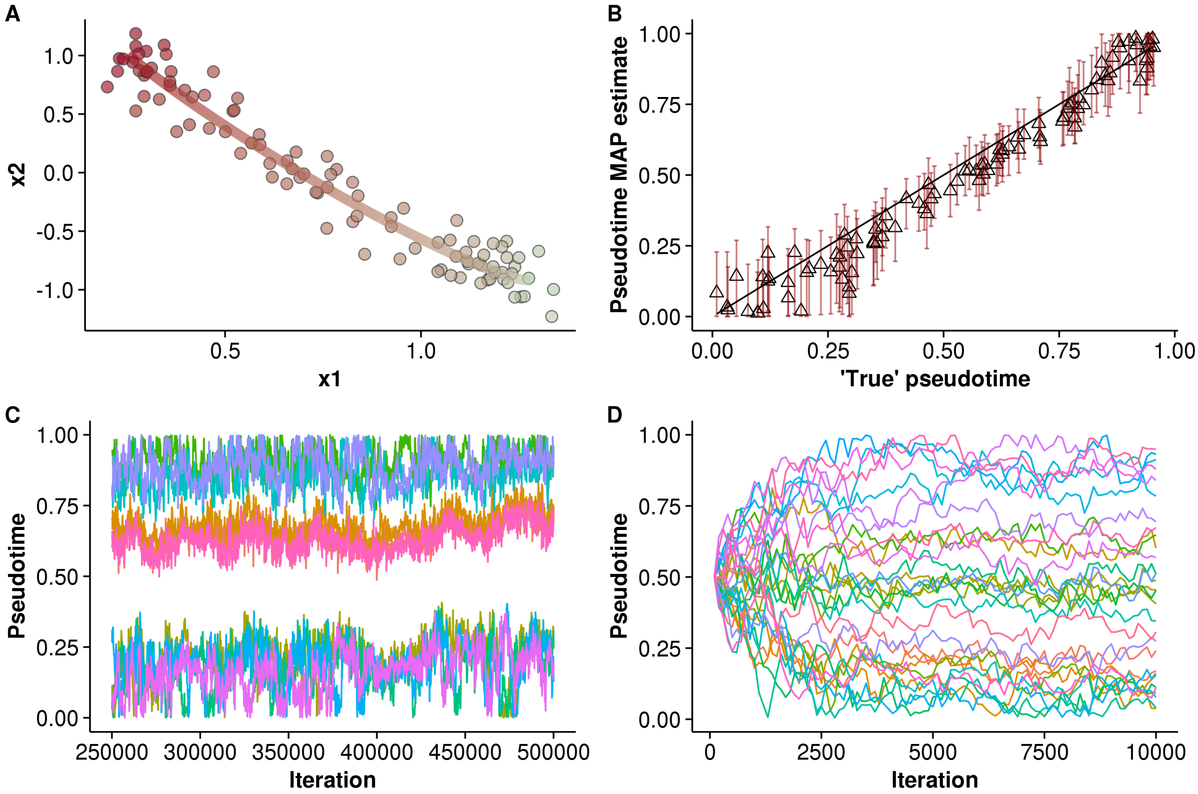

Figure 1: **Pseudotime inference on synthetic data. A.** The synthetic data points along with the MAP estimate of the mean function $\mu(t)$. Points are coloured by 'true' pseudotime while the curve is coloured by the MAP estimate. **B.** 'True' pseudotimes plotted against the MAP pseudotime estimates. Error bars report the 95% HPD credible interval; the solid line corresponds to $y = x$. **C.** MCMC traces for ten randomly chosen cells after the burn-in period. **D.** The 'big bang' intialisation for thirty randomly chosen cells up to 10^4 iterations.

3.2 Single-cell RNA-seq dataset

We next applied our method to a dataset of differentiating myoblasts [6]. A Laplacian Eigenmaps reduced-dimensionality representation of the data was used, as described at `https://github.com/kieranrcampbell/embeddr/blob/master/vignettes/vignette.Rmd`. Four outlier cells were discarded and the points were centre scaled to have mean 0 and standard deviation 1 in each dimension. We performed MH inference as described in Section 2.3 using 5×10^5 iterations thinned by 500. We set $\epsilon = 10^{-6}$, $\sigma_t = 6.5 \times 10^{-3}$, $\sigma_\lambda = 9 \times 10^{-13}$ and $\sigma_\sigma = 8 \times 10^{-3}$. The hyper-parameters were set to $r = 10^{-3}$, $\alpha = \beta = 1$, $\gamma = 100$.

The results of the inference can be seen in Figure 2. Clearly the GP-LVM curve traces through the centre of the manifold in a manner similar to the principal curve fit obtained using the `embeddr` [8] (Figure 2A). The MAP pseudotime values are very similar to the principal curve values with almost all falling within the 95% CI (Figure 2B).[3]

The variation in pseudotime estimate can be seen in Figure 2C/D. The 95% CI for some cells is as large as 0.5 - half the overall pseudotime window. This suggest that point estimation of pseudotimes could severely underestimate the potential variability in the estimates (which in the point estimate case is assumed to be negligible). Our model predicts uncertainty in such

[3] Although it looks like the pseudotimes are reversed, since they are 'pseudo' by nature they're entirely equivalent up to a parity transformation.

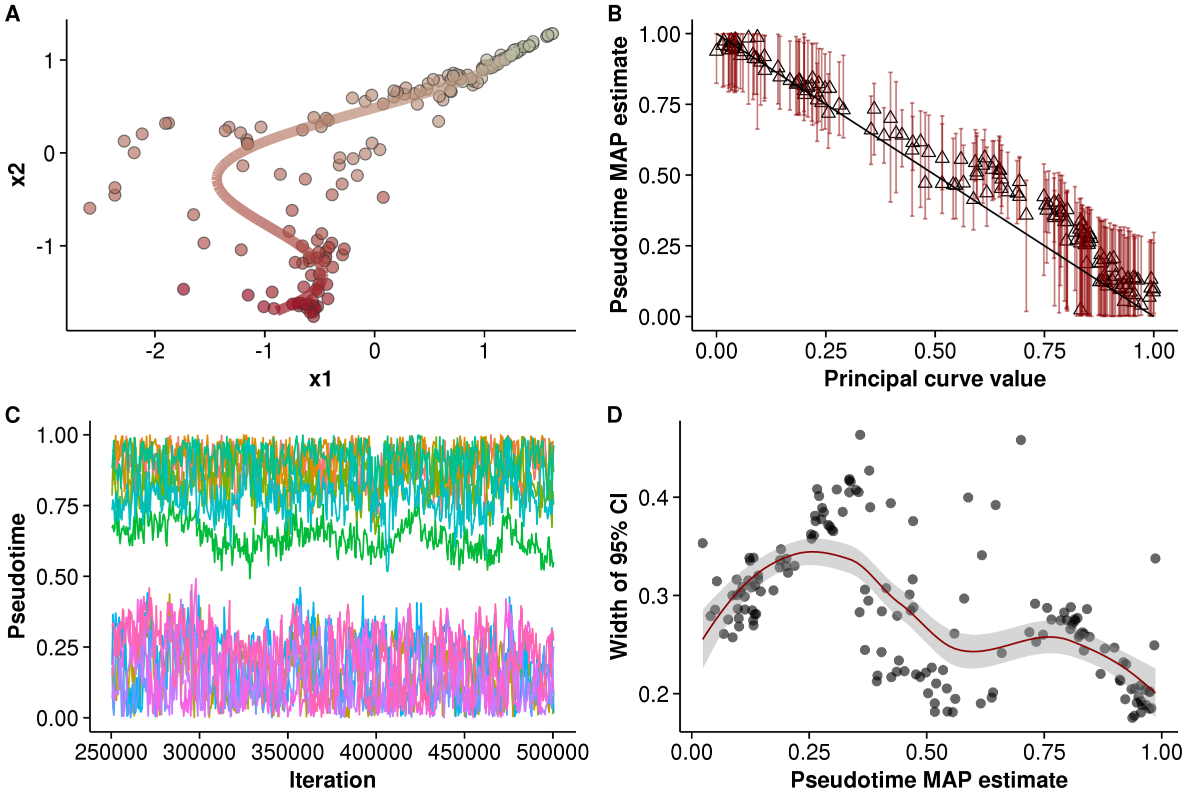

Figure 2: **Pseudotime inference on scRNA-seq data of differentiating myoblasts. A.** The Laplacian Eigenmaps representation of the dataset, with each point coloured by the MAP pseudotime estimate, along with the posterior mean curve of the GPLVM. **B.** The MAP pseudotime estimates compared to the principal curve fit, along with the 95% HPD credible interval. **C.** The thinned MCMC traces after burn-in for 12 randomly chosen cells. **D.** The width of the 95% HPD CI as a function of pseudotime.

assignments can extend to almost half of the entire biological process of interest. Visually, this is reasonable since the pseudotime assignment depends on the relative positioning along a curve drawn through the data points. There are clearly a range of possible curves that could be compatible with this data and integrating out the latent function allows us to characterise this uncertainty.

An alternative way to summarise the variation is using a non-parametric measure of correlation between subsequent (thinned) MCMC pseudotime traces. We computed the mean Kendall-Tau correlation - a non-parametric measure of rank correlation - along the entire chain and found it to be 0.84. In other words, the ordering of cells consistently changes meaning the idea of a well defined ordering of cells is meaningless.

4 Discussion

We have presented a novel probabilistic method for inferring pseudotimes from single-cell RNA-seq data by applying Bayesian Gaussian Process Latent Variable modelling. We have shown that the use of the Coulomb repulsive prior is appropriate to provide a well defined posterior over pseudotimes. Furthermore, the use of this prior combined with the MCMC sampling can decide the ordering of points using no prior knowledge, as opposed to the method suggested in

[11] where the structure of the prior combined with maximum likelihood inference means the points will never leave their initial ordering. Finally, because our method works in a reduced-dimension space an immediate visual check of plotting the MAP mean curve exists to ensure the inference method is not stuck in a local maximum.

By applying our method to both synthetic and real data we have shown that a pseudotime ordering of cells can be recovered and the inherent uncertainty in it characterised. We also uncovered a surprisingly large posterior uncertainty in pseudotimes and variability in the cell ordering, suggesting the ideas of 'fixed' pseudotimes should be reconsidered.

Previous approaches to pseudotime assignment have given point estimates of temporal ordering [6–8] but we have been able to infer the uncertainty associated with the pseudotime assignment to each cell with our model. Recently [9] have also used Bayesian Gaussian Process Latent Variable Models for pseudo-time assignment allowing uncertainty to be quantified. However, their approach requires the measurement of cell capture times upon which prior distributions can be centred solving identifiability issues by providing a physical calibrated temporal scaling. This information is often not available in single cell experiments. Our approach is more general providing support for data sets where actual temporal information cannot be attained. Prior capture time information can be included in our model, this would modifying the repulsive prior to be conditional on cells which have temporal information.

In future extensions of our work we will consider closer integration with the initial dimensionality reduction problem. Our simulations have assumed that the high-dimensional gene expression measurements have already been preprocessed and reduced to a low-dimensional representation (in our case using Laplacian Eigenmaps). The GPLVM framework provides a natural extension to avoid the need for this preprocessing step but inference could be challenging in the Monte Carlo sampling framework we desire. We would also like to consider improved Monte Carlo inference approaches as early experiments using standard MCMC techniques, such as parallel tempering, have yielded no significant sampling efficiency due to the extreme multi-modal nature of the Coulomb repulsion prior. We are developing novel sampling techniques to address the unique properties of this prior but will also consider alternative repulsive processes. Finally, we are also examining downstream applications of our techniques for quantifying temporal gene expression behaviour.

5 Acknowledgements

We thank Michalis Titsias for comments and advice regarding the manuscript. K.C. is supported by the MRC. C.Y. is supported by a UK Medical Research Council New Investigator Research Grant (Ref. No. MR/L001411/1), the Wellcome Trust Core Award Grant Number 090532/Z/09/Z, the John Fell Oxford University Press (OUP) Research Fund and the Li Ka Shing Foundation via a Oxford-Stanford Big Data in Human Health Seed Grant.

References

1. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–5. ISSN: 1476-4687 (May 2014).

2. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology.* ISSN: 1087-0156. doi:10.1038/nbt.3102. <http://www.nature.com/doifinder/10.1038/nbt.3102> (Jan. 2015).

3. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology* **33.** ISSN: 1087-0156. doi:10.1038/nbt.3154. <http://www.nature.com/doifinder/10.1038/nbt.3154> (Feb. 2015).

4. Macaulay, I. C. & Voet, T. Single cell genomics: advances and future perspectives. *PLoS genetics* **10,** e1004126. ISSN: 1553-7404 (Jan. 2014).

5. Stegle, O., Teichmann, S. a. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16,** 133–145. ISSN: 1471-0056 (Jan. 2015).

6. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32,** 381–6. ISSN: 1546-1696 (Apr. 2014).

7. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157,** 714–25. ISSN: 1097-4172 (Apr. 2014).

8. Campbell, K., Ponting, C. & Webber, C. *Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles* (submitted). 2015.

9. Reid, J. E. & Wernisch, L. Pseudotime estimation: deconfounding single cell time series. *bioRxiv,* 019588 (2015).

10. Titsias, M. & Lawrence, N. Bayesian Gaussian Process Latent Variable Model. *Artificial Intelligence* **9,** 844–851 (2010).

11. Wang, Y. & Dunson, D. B. *Probabilistic Curve Learning: Coulomb Repulsion and the Electrostatic Gaussian Process* in *Advances in Neural Information Processing Systems* (2015).

12. Belkin, M. & Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data. **1396,** 1373–1396 (2003).

13. Lawrence, N. D. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems* **16,** 329–336 (2004).

14. Titsias, M. K., Lawrence, N. & Rattray, M. Markov chain Monte Carlo algorithms for Gaussian processes. *Inference and Estimation in Probabilistic Time-Series Models,* 9 (2008).