Scaling up probabilistic pseudotime estimation with the GPLVM

Sumon Ahmed, Alexis Boukouvalas and Magnus Rattray

May 18, 2017

Abstract

The analysis of single cell genomics data promises to reveal novel states of complex biological processes, but is challenging due to inherent biological and technical noise. We propose a probabilistic approach based on sparse variational Bayesian Gaussian process latent variable model (GPLVM) to perform robust pseudotime estimation whilst allowing for the incorporation of prior information such as cell capture times. The model converges an order of magniture faster compared to existing methods whilst achieving similar levels of estimation accuracy. We demonstrate the flexibility of our approach by extending the model to higher-dimensional latent spaces that can be used to simulteneously infer pseudotime and branching structure.

1 Introduction

The analysis of cell differentiation and maturation at single-cell level has been shown to be promising, and offers numerous advantages over bulk analysis. The average transcriptomes across a cell population fails to capture the crucial transcriptomic signal in individual cells and recent studies have shown that many questions of cellular developments can be answered in a more refined way in single-cell level (Cannoodt *et al.*, 2016).

During single cell sampling process, the actual temporal label that identifies the cells' position in the differentiation trajectory is lost and these parameters become unobserved, latent quantities known as pseudotime (Trapnell *et al.*, 2014). The initial single cell data may contain a mixture of cells of different cell cycle stages (McDavid *et al.*, 2014) or a set of cells sampled at distinct time points (Windram *et al.*, 2012). Starting with the high dimensional data, the inferrence algorithms first apply dimensionality reduction techniques to get a compressed lower (usually two or three) dimensional representation. A number of dimensionality reduction techniques have been adopted in single cell transcriptomics studies such as Principal and Independent Components Analysis (P/ICA) (Trapnell *et al.*, 2014; Ji and Ji, 2016); non-linear techniques such as t-stochastic neighbourhood embedding (t-SNE) (Becher *et al.*, 2014) and diffusion map (Haghverdi *et al.*, 2015).

The low-dimensional embedding is used to characterize the pseudotime trajectory. Different formalisms are used to represent a pseudotime trajectory. In graph based methods such as Monocle (Trapnell *et al.*, 2014), Wanderlust (Bendall *et al.*, 2014), Waterfall (Shin *et al.*, 2015) and TSCAN (Ji and Ji, 2016), a simplified graph or tree is used as input. By using different path-finding algorithms, these methods try to find a path through a series of nodes. These nodes can correspond to individual cells (Trapnell *et al.*, 2014; Bendall *et al.*, 2014) or group of cells (Shin *et al.*, 2015; Ji and Ji, 2016) in the graph. SCUBA (Marco *et al.*, 2014) uses curve fitting to characterize the pseudotime trajectory. Principal curves are used to model trajectory and each cell is assigned a pseudotime according to its low-dimensional projection on the principal curves.

One major drawback of these methods is the absence of a probabilistic framework. They only provide a single point estimate of pseudotimes concealing the impact of biological and technical noises.Campbell and Yau (2016) have used the GPLVM where pseudotime trajectories have been modelled by the latent variables. They have used Markov Chain Monte Carlo (MCMC) simulation to draw samples from the posterior pseudotime distribution where each sample corresponds to one possible pseudotime ordering for the cells with associated uncertainties.

The pseudotime estimation in the above methods is unstructured lacking any physical or biological interpretation of the space. These methods do not incorporate experimental covariates such as cell type or true capture time, and may fail to uncover a specific structure of interest. As an example, in immune response, after the combat with the infection is finished the natural course is to go back to a healthy state. Thus, the expression profiles show a cyclic behaviour where it is challenging to estimate a single pseudotime. Reid and Wernisch (2016) have developed a Bayesian approach that uses GPLVM and impose a prior structure on the latent dimension. The latent dimension in their model is one dimensional pseudotime and the imposed structure relates it to the cell capture time. This helps the model not only to maximise some relevant statistics but also to identify sample specific features such as cyclic behaviour of cell cycle data. The pseudotime points estimated by their model are in proximity to the actual capture time and in the same scale.

2 Methods

The primary latent variables in the proposed method are the pseudotimes. The expression profile of each gene is modelled by a Gaussian process

$$y_g = GP(0, k(t, t^*))$$

where y_g is the expression profile of gene g; and $k(t, t^*)$ is the covariance function between two distinct pseudotime points t and t^*. Thus, the expression profiles are the functions of pseudotime and the covariance function imposes the smoothness constraint that is shared by all genes.

The Bayesian GPLVM has the computational complexity of $O(GC^3)$, where G is the number of genes and C is the number of cells. To make the model computationally tractable for large datasets, a sparse approximation has been incorporated in different models (Reid and Wernisch, 2016). Sparse GP approximation has the complexity $O(GCM^2)$ where $M << C$ is the number of auxiliary or inducing points. These inducing points may or may not coincide with the actual points. As M is chosen much smaller than C, sparse approximation provides a great efficiency in terms of computational time.

The computation of the log marginal likelihood is mathematically intractable and MCMC methods (Campbell and Yau, 2016; Reid and Wernisch, 2016) have been employed for inference. However their computational complexitity has motivated variational approaches (Damianou *et al.*, 2015) that provide a lower bound on the likelihood at a fraction of the computational cost. Reid and Wernisch (2016) also use black box variational approaches that rely on sampling to increase inference efficiency. However for the Bayesian GPLVM an analytic exact bound exists (Damianou *et al.*, 2015) but the original derivation and all currently available packages such as GPy (2012) assume an uninformative prior. We modify the exact bound to allow for informative priors

$$\log p\left(Y\right) \geq \mathbb{E}_{q(t)}\left[\log p\left(Y|t\right)\right] - KL\left[q\left(t\right)||p\left(t\right)\right]$$

where $q\left(t\right)$ the variational distribution and $p\left(t\right) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|\tau_n, \sigma_t^2\right)$ the modified prior centered at the capture time τ_n of cell n with prior variance σ_t^2

We have implemented our approach in the GPflow package (Matthews *et al.*, 2016) whose flexible architecture allows to perform the computation across multiple CPU cores and GPUs. The source of the scalability of our approach is therefore two-fold: model estimation using an exact variational bound and implementation on a scalable software architecture.

3 Results and Discussion

The performance of the proposed framework has been investigated by analysing three different data sets from three different organisms: whole leaf microarrays of *Arabidopsis thaliana* (Windram *et al.*, 2012); single cell expression profiles of a human prostate cancer cell line (McDavid *et al.*, 2014) and single cell RNA-Seq libraries of mouse dendritic cells (Shalek *et al.*, 2014).

The proposed model has been compared with the work of Reid and Wernisch (2016) in terms of model fitting as well as the time required to fit the model. As GPflow provides exact bound for variational approximation, the parameter estimation of the proposed model shows robustness with respect to the validation methods used to campare the models. Moreover, the proposed method converges quickly by using a small number of inducing points even for large data. The proposed model outperforms the DeLorean approach (Reid and Wernisch, 2016) in all aspects[1,2,3]. We also demonstrate the flexibility of the model to infer higher-dimensional latent spaces (Guo *et al.*, 2010).

3.1 Infering withheld time points and smooth pseudotime trajectories

Windram *et al.* (2012) examined the effects of *Botrytis cinera* infection on *Arabidopis thaliana*, and time series contains 24 distinct capture time points. These 24 times points have been grouped into 4 separate groups, each consisting of 6 consecutive time points, which have been fed to the model for prior initialization. Figure 1 depicts the comparison of the proposed method to the Delorean approach (Reid and Wernisch, 2016) for *Arabidopsis thaliana* data. All the experiments have been carried out by using the same experimental setup. The Figure 1(**Top**) shows the best and average spearman correlation between the actual capture time and the estimated pseudotime for different number of inducing points used. Both the best and average correlation values show that the proposed method has better convergence for relatively smaller number of inducing points than Delorean method. Figure 1(**Bottom**) depicts the fitting time required by both models for different number of inducing points.

To verify smoothness of the predicted trajectory, roughness statistics R_g (Reid and Wernisch, 2016) has been calculated. The average R_g for different prior initialization is 0.71 which is smaller than R_g 0.72 calculated by Reid and Wernisch (2016).

3.2 Recovering cell cycle peak times

McDavid *et al.* (2014) examined the effect of cell cycle on single cell gene expression and this work uses the expression data from the PC3 human prostate cancer cell line. The inference has been carried out by using the top 56 differentially expressed genes in 361 cells. To identify the cyclic nature of the cell cycle, the method uses a periodic kernel function (MacKay, 1998). The expression profiles of some selected genes over the estimated pseudotime are shown in Fig. 2. DeLorean approach requires 7h 31m to fit the model for this data while the proposed method uses 20 inducing points and takes only 4m 45s to converge.

To evaluate the model's performance, estimated peak times from the expression profiles fit by the proposed model have been compared with the peaks times defined by the CycleBase database (Santos *et al.*,

[1] Time indicated for DeLorean method is for 40 initializations while the mentioned for the proposed model is for 1 initialization

[2] DeLorean Machine: DELL R815 server having 4 x AMD 6174 2.2 Ghz processors (12 cores each), with 128 Gb of 1333 MHz RAM memory

[3] Our Machine: Intel(R) Core(TM) i5-3570 CPU @ 3.40GHz, Ram: 16 GB

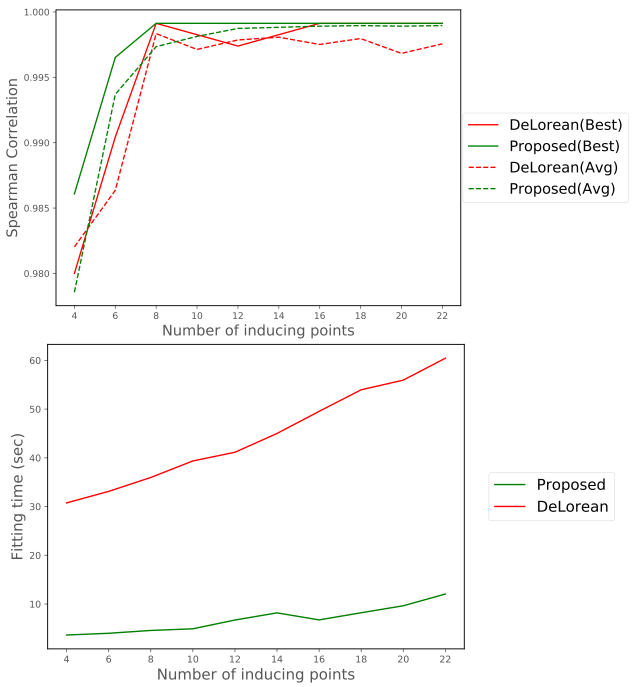


Figure 1: A comparision of performance and fitting time between the proposed method and Delorean method for Windram *et al.* (2012) microarray data. **(Top)** Spearman correlation between the actual capture time and the estimated pseudotime for different number of inducing points. **(Bottom)** Fitting time required by the models for the same experimental setups

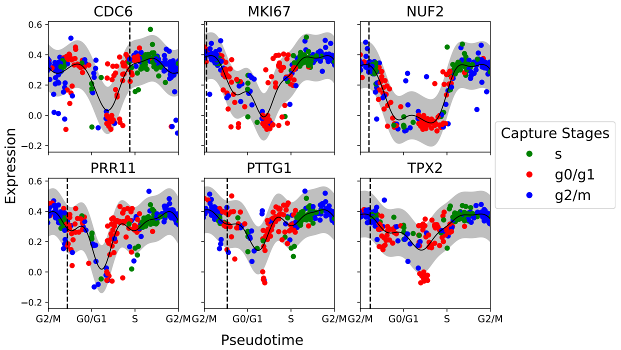


Figure 2: Expression profiles over estimated pseudotime for some selected genes from PC3 human prostate cancer cell line. Each point corresponds to a particular gene expression in a cell. The points are colored based on cell cycle stages according to McDavid *et al.* (2014). The circular horizontal axis (where both first and last lables are G2/M) represents the periodicity realized by the method in pseudotime inference. The solid black line is the posterior predicted mean of expression profiles while the grey ribbon depicts the 95% confidence interval. The vertical dotted lines are the CyclyBase peak times for the selected genes.

2014). The root mean square error (RMSE) between the estimated peaks and the CycleBased defined peaks is 13.6 ± 0.4 which is smaller than the RMSE 14.5 calculated between the same quantities in Reid and Wernisch (2016).

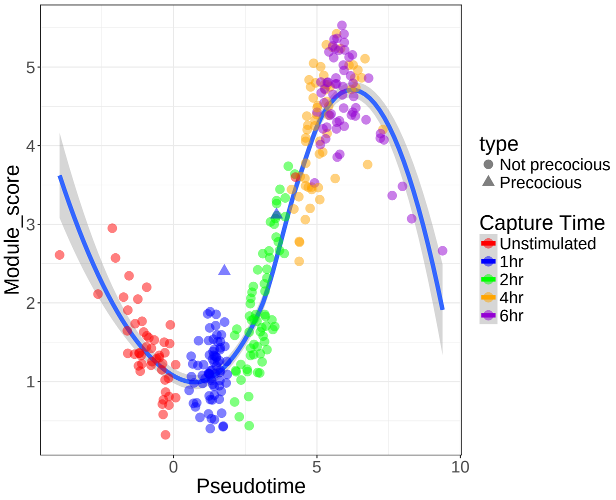


Figure 3: The module score (Shalek *et al.*, 2014) of core antiviral cells over pseudotime. The two precocious cells (plotted as triangles) have been placed in later pseudotimes than the other cells of captured at 1 h. A loess curve (solid blue line) has been plotted thorough the data.

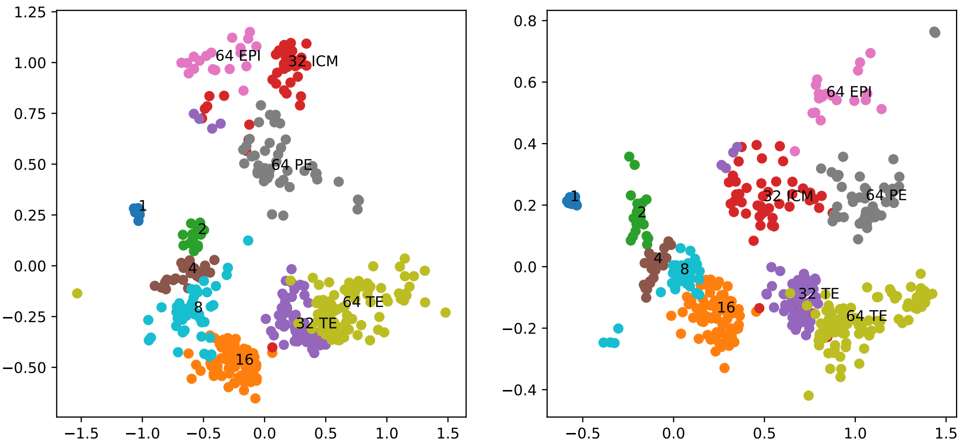
3.3 Correctly identify precocious cells

Shalek *et al.* (2014) identified a core antiviral gene module expressed in LPS after 2-4 hours. They defined two cells captured at 1 h which switched to this group precociously. The inference process uses top 74 variationaly expressed genes from the clusters Id, IIIb, IIIc, IIId. The time used by DeLorean method is 20m while the proposed method takes 3m for the same number of inducing points. Fig. 3 shows the module score (Shalek *et al.*, 2014) of core antiviral genes over the estimated pseudotime. Two precocious cells have been assigned pseudotimes in the middle of 2 h group. Thus, the model successfully simulates the concept that some cells can progress across the differentiation (pseudotime trajectories) faster than others.

3.4 Pseudotime-branching inference

The model has been extended for 2-D latent spaces and has been applied on the single cell qPCR data of early developmental stages from multicellular organisms (Guo *et al.*, 2010). The gene expression profiles of 48 genes was measured across 437 cells. Cells differentiate from the single cell stage into three different cell states in the 64 cell stage: trophectoderm (TE), epiblast (EPI), and primitive endoderm (PE).

Both models with informative and non-informative priors were examined. The informative prior (Figure 4) on capture time helps with the identifiability of the model as it aligns the first latent dimension (horizontal axis) with pseudotime and the second latent dimension (vertical axis) with the branching structure. The rank correlation is also higher (0.95 vs 0.79) as well as the log likelihood.



(a) No prior

(b) Informative prior

Figure 4: Latent space reconstruction without and with prior. The latter captures both developmental time and branching structure (Guo *et al.*, 2010). The cell stage and type labels are also shown.

4 Conclusion

Pseudotime estimation on single cell genomics faces a number of challenges as many sources of variability introduce a significant amount of statistical uncertainty in the inference process. The proposed method uses a sparse variational Bayesian GPLVM with an informative prior on the latent space. Four different datasets have been used to examine the model's suitability to estimate pseudotimes. The model can be extended to higher dimensional latent spaces where the interaction of pseudotime with other factors can be captured. We have demonstrated this capability on a two-dimensional latent space where pseudotime is estimated jointly with the developmental branching structure. In all cases, the model requires a small number of inducing points a and less computation to generate biologically plausible estimates compared to existing approaches. Thus the scalability and flexibility of the proposed method ensures its utility for analysing larger datasets such as those generated from droplet-based techniques.

Acknowledgements We wish to thank John Reid for his help with the DeLorean code and data. We also thank James Hensman for his help with extending the GPflow package.

References

Becher, B., Schlitzer, A., Chen, J., Mair, F., Sumatoh, H. R., Teng, K. W. W., Low, D., Ruedl, C., Riccardi-Castagnoli, P., Poidinger, M., *et al.* (2014). High-dimensional analysis of the murine myeloid cell system. *Nature immunology*, **15**(12), 1181–1189.

Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Peer, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, **157**(3), 714–725.

Campbell, K. and Yau, C. (2016). Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference. *PLoS Computational Biology*, **12**(11).

Cannoodt, R., Saelens, W., and Yvan, S. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *European Journal of Immunology*.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. (2015). Variational inference for latent variables and uncertain inputs in gaussian processes. *Journal of Machine Learning Research (JMLR)*, **2**.

GPy (since 2012). GPy: A gaussian process framework in python. http://github.com/SheffieldML/GPy.

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. *Developmental 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.

Ji, Z. and Ji, H. (2016). Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, **44**(13).

MacKay, D. J. (1998). Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, **168**, 133–166.

Marco, E., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., and Yuan, G.-C. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, **111**(52), E5643–E5650.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. (2016). GPflow: A Gaussian process library using TensorFlow. *arXiv preprint 1610.08733*.

McDavid, A., Dennis, L., Danaher, P., Finak, G., Krouse, M., Wang, A., Webster, P., Beechem, J., and Gottardo, R. (2014). Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLoS Comput Biol*, **10**(7), e1003696.

Reid, J. E. and Wernisch, L. (2016). Pseudotime estimation: deconfounding single cell time series. *Bioinformatics*, **32**(19), 2973–2980.

Santos, A., Wernersson, R., and Jensen, L. J. (2014). Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic acids research*, page gku1092.

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., *et al.* (2014). Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, **510**(7505), 363–369.

Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., Bonaguidi, M. A., Enikolopov, G., Nauen, D. W., Christian, K. M., Ming, G.-l., *et al.* (2015). Single-cell rna-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell*, **17**(3), 360–372.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, **32**(4), 381–386.

Windram, O., Madhou, P., McHattie, S., Hill, C., Hickman, R., Cooke, E., Jenkins, D. J., Penfold, C. A., Baxter, L., Breeze, E., *et al.* (2012). Arabidopsis defense against botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *The Plant Cell*, **24**(9), 3530–3557.