Gaussian processes for identifying branching dynamics in single cell data

Alexis Boukouvalas, James Hensman and Magnus Rattray

Abstract

Single cell gene quantification allows for the analysis of heterogeneous cell populations and the analysis of the whole transcriptome without the need for a priori gene target selection. Identifying branching dynamics in cell populations undergoing differentiation is computationally challenging due to lack of time course data and high technical and biological noise. We develop the branching Gaussian process (BGP), a non-parametric flexible model that is able to robustly identify branching dynamics on an individual gene level whilst also providing an uncertainty estimate of the branching times. The code is open-source and available at https://github.com/alexisboukouvalas/BranchedGP.

1 Introduction

Single-cell gene expression data can be used to uncover cellular progression through different states of a temporal transformation, e.g. during development, differentiation or disease. As single cell protocols improve, a flurry of methods have been recently proposed to model branching behaviour in transcriptional cell fates (Haghverdi *et al.*, 2016; Setty *et al.*, 2016; Qiu *et al.*, 2017). In these and similar methods, pseudotime is estimated and a global branching structure is inferred. Our focus in this paper is to propose a downstream analysis method that can be subsequently used to identify the individual gene branching time. Recently Qiu *et al.* (2017) have proposed the branch expression analysis modelling (BEAM) approach that uses penalised splines to infer the individual gene branching time. Here we propose an alternative non-parametric method based on Gaussian process inference to estimate the branching tree structure. The model provides a log likelihood ratio estimate of the evidence for branching and a posterior estimate of the most likely branching location as well as a confidence interval.

2 Methods

Our approach builds on Gaussian processes (GP), a flexible on-parametric probabilistic model. A GP can be defined as a probability distribution over functions $f(t) \sim \mathcal{GP}\left(\text{mean}(t), \text{cov}(t, t')\right)$ (Rasmussen and Williams, 2006). GPs have been applied to single-cell data for pseudotime estimation where the effect of of uncertainty can be quantified (Campbell and Yau, 2016) and capture time included as prior information (Reid and Wernisch, 2016).

Yang *et al.* (2016) developed a GP model for the identification of a single perturbation time point from two-sample time series data. They define a novel kernel that constrains two functions to intersect at a single point. The bifurcation point is identified by numerically approximating the posterior and selecting a point estimate. The model is used to identify when a gene becomes differentially expressed in time course gene expression data under control and perturbed conditions. In their approach all data points have been labelled with the branch that generated them and the ordering of time points is assumed known.

Conversely, the overlapping mixture of GPs (Lázaro-Gredilla *et al.*, 2012) assumes the functions are independent (no branching) but estimates the generating labels using a mixture model. Our method combines these two approaches resulting in a probabilistic model where the branching posterior is available whilst the data labels are estimated. More generally, the BGP method is a generalisation of the overlapping mixture of GPs approach to correlated latent functions.

For pseudotime estimation we use the reversed graph embedding approach of Qiu *et al.* (2017), termed DDRTree, which we have found to be effective in recovering pseudotime in the presence of branching, although other approaches may be selected. Qiu *et al.* (2017) have shown the reverse graph embedding approach to outperform DPT (Haghverdi *et al.*, 2016), Wishbone (Setty *et al.*, 2016) and other methods in their analysis.

A naive implementation of GP models scales cubically with the size of the data. As increasing number of cells may be measured in new single cell protocols, we ensure the scalability of our approach by employing two complementary approaches. Firstly we use sparse inference (Quiñonero-Candela and Rasmussen, 2005) that allows model fitting to scale with the number of inducing points. The latter is a user-defined value that trades off model accuracy and training time. Specifically for N training points, covariance inversion scales as $O(N^3)$

whereas sparse inference with k inducing points is $O(k^2 N)$. Secondly, we provide an open-source implementation that leverages the GPflow library (Matthews *et al.*, 2017), which both simplifies the implementation due to automatic symbolic differentiation and allows for the necessary matrix operations to be computed in parallel across many CPU nodes or GPUs.

Let F a branching Gaussian Process evaluated for N data points with M branches and $Z \in \{0,1\}^{N \times M}$ indicates which branch each cell comes from. The likelihood is $p(Y|F,Z) = \mathcal{N}\left(Y|ZF, \sigma^2 I\right)$ and as in Lázaro-Gredilla *et al.* (2012) we place a categorical prior on the indicator matrix $p(Z) = \prod_{n=1}^{N} \prod_{m=1}^{M} [\Pi]_{n,m}^{[Z]_{nm}}$. We place a GP prior on the latent functions $p(F|\, t_b) = \mathcal{GP}(0, K|\, t_b)$ which constrains the latent functions to branch at pseudotime t_b. Note that the latter does not factorize as in Lázaro-Gredilla *et al.* (2012) as the latent functions are dependent.

Global branching labels such as those provided by DDRTree can provide an informative prior $p(Z)$ for all genes. The prior preceding to the global branching point is uninformative as no global assignment is available. After the global branching point, the prior favours increased assignment probability to the globally assigned branch. However as the prior we use places non-zero mass on the alternative assignment, the resulting assignment may differ from the global allocation given enough evidence from the likelihood term.

The log likelihood is not analytically tractable as it involves integrating out the indicator matrix Z. We proceed to compute a lower bound using Jensen's inequality

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

where $q(Z,F) = q(Z)\,q(F)$ as by mean-field assumption the latent functions F are independent of the association indicators Z and $q(Z) = \prod_{nm} \phi_{nm}$. The ϕ_{nm} approximates the posterior probability of which branch cell n belongs to. The latter is either the trunk state or one of the two branches in the case of a single branching considered here. Then F can be integrated out to get marginal likelihood $p(Y)$.

The branching time posterior probability is calculated using the approximate marginal likelihood $p(t_b|Y) = \frac{p(Y|t_b)}{\sum_{t_b} p(Y|t_b)}$. We can also calculate a likelihood ratio of branching versus not branching to rank genes by how likely their expression exhibits branching.

3 Synthetic study

We evaluate three methods, the mixture of factors analysers (Campbell and Yau, 2017), the BEAM approach (Qiu *et al.*, 2017) and the branching GP model on synthetically generated data. The synthetic scenarios are summarized in Table 1.

Table 1: Synthetic scenarios settings.. The branching location and number of outputs is provided. A branching location of 1.1 refers to a non-branching output. All scenarios use $N = 200$ points.

Scenario	Branching	Description
1	[0.2, 15], [0.6, 15], [1.1, 10]	Multiple branching points
2	[0.2, 15], [0.6, 15], [1.1, 10]	Short lengthscale
3	[0.1, 3], [0.7, 27], [1.1, 10]	Majority of late branching genes
4	[0.2, 20], [1.1, 20]	High branching variance
5	[0.2, 20], [1.1, 20]	Linear model branching
6	[0.2, 10], [0.8, 20], [1.1, 10]	High noise

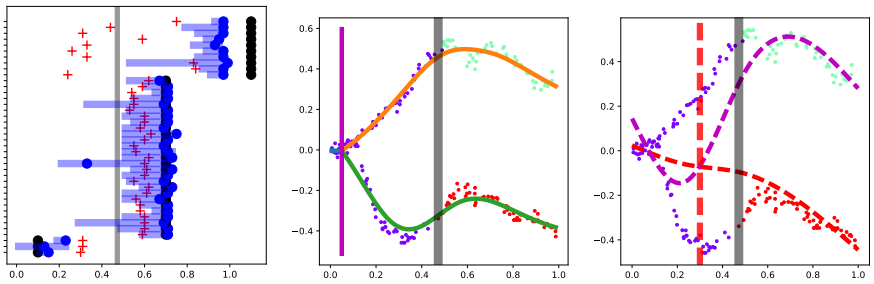
The log likelihood ratio of the branching GP can be used to rank the evidence of branching for each gene. Similar measures exist for the MFA and BEAM method. We first compare the three methods on their ability to discriminate branching genes from non-branching genes. The metric we use is the area under the curve which has been classically used to evaluate classification models. The branching GP and BEAM methods achieve consistently good performance (Table 2) whilst MFA performance varies significantly. The MFA does not perform well due to poor estimation of the branching pseudotime.

The error in estimating branching time for the BEAM and BGP methods is given in Table 2 (b). As the BGP methods allows for the inclusion of a prior assignment probability, we report the error of an uninformative prior (50% prior probability of assignment to each branch) and a prior derived from the global DDRTree assignment (90% prior probability). The error for the BGP method is consistently lower than the BEAM method. As can be seen in the Figure 1(a), the estimates for the BEAM method tend to biased towards the global branching time. The underestimation of branching times in BEAM for genes that branch later than the global branching time is most likely due to the spline regularisation employed by the Monocle software that tends to over-smooth the spline fit. The overestimation of branching times for genes branching prior to the global branching time is due to the arbitrary assignment of cells prior the global branching time as no labels are provided by the global algorithm and no estimation is performed by the spline-fitting algorithm; see Figure 1 for an illustrative example. The former could possibly be rectified by tuning the regularisation approach employed but the latter is a fundamental restriction of the BEAM approach that does not directly estimate branching assignments but only

uses the globally derived labels estimates. The BGP approach does not suffer from this deficiency as the branch assignment is performed on a gene by gene basis at the cost of increased computation time. We note however the problem is parallelisable as each gene is treated independently. The prior has little effect on estimation accuracy for low noise-levels (Table 2 (b)). However in the more realistic higher-noise scenario (6), the inclusion of an informative prior results in significantly lower estimation errors.

Table 2: Synthetic study: (a) area under the curve for detecting branching genes and (b) root mean squared error of estimated branching time.

Scenario	(a) Branching Score (AUC)				(b) Branching Time (RMSE)		
	MFA	BEAM	BGP No prior	BGP Weak prior	BEAM	BGP No prior	BGP Weak prior
1	0.19	1.00	0.99	1.00	0.39	0.32	0.31
2	0.66	1.00	1.00	1.00	0.35	0.24	0.24
3	0.80	1.00	1.00	1.00	0.34	0.09	0.09
4	0.90	1.00	1.00	1.00	0.53	0.19	0.22
5	1.00	1.00	1.00	1.00	0.58	0.30	0.30
6	0.73	0.90	0.82	0.95	0.34	0.38	0.28



(a) Posterior (b) BGP (c) BEAM

Figure 1: Synthetic data: fitting BGP and BEAM on an early branching gene. (a) The true branching times (black dots), BEAM times (red crosses) and BGP mean (blue dots) and 95% credible regions are shown. (b)-(c)The vertical grey bar is the global branching time. The vertical coloured bar is the model branching time estimate and cells have been coloured by the global assignment.

4 Hematopoiesis single cell RNA-seq

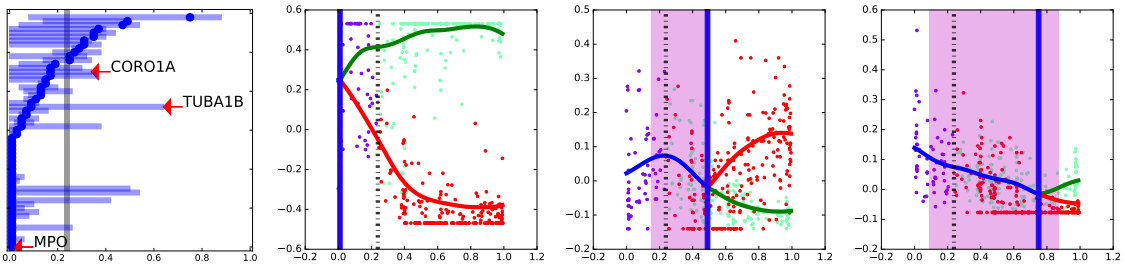
We apply the BGP model on single-cell RNA-seq of haematopoietic stem cells (HSC) differentiating into myeloid and erythroid precursors (Paul *et al.*, 2015). The data consists of 4423 cells and we retain the top 500 genes as ranked by the difference in median between the DDRTree end-states at each branch. The root state was selected using marker genes for the common myeloid progenitors, erythrocytes and granulocyte-macrophage progenitors (GMP).

A probabilistic model is an appropriate choice for early haematopoiesis which has been described as a cellular continuum of low-primed HSCs (Velten *et al.*, 2017). The continuum contains transitory states rather than discrete progenitor cell types with some cell state transitions and lineage combinations more likely to occur than others. (Velten *et al.*, 2017). A probabilistic model such as BGP better reflects the probabilistic nature of lineage selection highlighted in Velten *et al.* (2017). In the BGP model in particular, each cell is associated with an allocation probability for each branch. The branching point can be interpreted as the earliest pseudotime from which probabilistic biases in lineage selection can be detected.

In Figure 2(a) we show the posterior summary for the 65 genes with highest branching probability. The uncertainty of the branching posterior increases for later branching locations as identification of later branching points is more challenging as the amount of informative cells decreases. Three examples are shown in more detail in Figure 2(a). The MPO gene is a well known GMP marker and exhibits early branching behaviour prior to the global branching time. The TUBA1B and CORO1A genes exhibit later branching times with increased posterior uncertainty.

5 Discussion

We have presented a flexible non-parametric probabilistic approach to robustly identify individual gene branching times. For scalability our model uses sparse variational inference implemented in a scalable computing



(a) Posterior (b) MPO (c) TUBA1B (d) CORO1A

Figure 2: Hematopoiesis gene expression: (a) Posterior summary of 65 genes with highest branching probability. The genes are ordered by the branching location. (b)-(d) BGP fits for three different genes. The global branching time (dashed black vertical line), BGP branching point mode (blue vertical line) and 99% posterior intervals (magenta vertical span) are also shown. The cells are marked according to the global allocation estimated by the DDRTree algorithm.

architecture. Previous methods such as BEAM are unable to accurately identify branching times earlier than the global branching time whereas the BGP method can robustly do so as it estimates cell branch association for each gene independently. Further, the BGP approach provides an uncertainty estimate that can be used in downstream analysis of the individual gene branching times.

We have also included in our comparison a probabilistic linear method (Campbell and Yau, 2017). The linearity allows for an efficient joint estimation of both the pseudotime and global branching structure. Although this method does not estimate gene bifurcation times, a probabilistic estimate of an individual gene exhibiting branching behaviour is available. However in our synthetic study we have found the pseudotime estimation not to be robust and this reduces the effectiveness of the method.

A related approach to the BGP model also uses a Gaussian process (GP) mixture model to model branching (Lönnberg *et al.*, 2017). However the mixture model used assumes the latent functions are independent without any branching and a heuristic is then proposed to identify the most likely branching time. We were unable to obtain meaningful results with this method.

Acknowledgements We wish to thank Xiaojie Qiu for his help with the BEAM Monocle 2 code.

References

Campbell, K. and Yau, C. (2017). Probabilistic inference of bifurcations in single-cell data using a hierarchical mixture of factor analysers. *Wellcome Open Res*.

Campbell, K. R. and Yau, C. (2016). Order under uncertainty: Robust differential expression analysis using probabilistic models for pseudotime inference. *PLOS Computational Biology*, **12**(11), 1–20.

Haghverdi, L., Buettner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, **13**(10), 845–848.

Lázaro-Gredilla, M., Van Vaerenbergh, S., and Lawrence, N. D. (2012). Overlapping Mixtures of Gaussian Processes for the data association problem. *Pattern Recognition*, **45**(4), 1386–1395.

Lönnberg, T., Svensson, V., James, K. R., Fernandez-Ruiz, D., Sebina, I., Montandon, R., Soon, M. S., Fogg, L. G., Nair, A. S., Liligeto, U., *et al.* (2017). Single-cell rna-seq and computational analysis using temporal mixture modelling resolves th1/tfh fate bifurcation in malaria. *Science immunology*, **2**(9).

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. (2017). GPflow: A Gaussian process library using Tensorflow. *Journal of Machine Learning Research*, **18**(40), 1–6.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., *et al.* (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, **163**(7), 1663–1677.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell developmental trajectories. *bioRxiv*, page 110668.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, **6**(Dec), 1939–1959.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Reid, J. E. and Wernisch, L. (2016). Pseudotime estimation: deconfounding single cell time series. *Bioinformatics*, page btw372.

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature biotechnology*, **34**(6), 637–645.

Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., Hirche, C., Lutz, C., Buss, E. C., Nowak, D., *et al.* (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology*, **19**(4), 271–281.

Yang, J., Penfold, C. A., Grant, M. R., and Rattray, M. (2016). Inferring the perturbation time from biological time course data. *Bioinformatics*, page btw329.