Identification of branching using pseudotime estimation

Alexis Boukouvalas

alexis.boukouvalas@manchester.ac.uk

MANCHESTER
1824

The University of Manchester

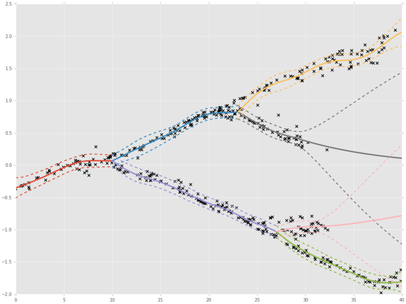
http://personalpages.manchester.ac.uk/staff/alexis.boukouvalas/

October 3, 2016

Joint work with James Hensman, Magnus Rattray

Single cell pseudotime inference

- High throughput single cell protocols typically provide snapshot view of gene expression.
- Pseudotime methods place cells on a continuous path reflecting the similarity and rate of change of their gene expression.
- Biological processes such as differentiation also exhibit distinct cell fates across a common lineage.



Single cell non-linear branching models

- Using probabilistic models, ensures a logical and consistent way of including relevant prior information such as cell capture times in synchronised populations.
- Using the GPLVM framework allows us to infer pseudotime including such prior information [Reid].
- A non-linear mapping increases the accuracy of pseudotime estimation [TopSlam, DPT].
- Missing information (dropout) is relatively straightforward to handle in a probabilistic model [ZIFA].

Branching approaches

- Diffusion pseudotime → transition matrix between cells, then branching.
- Wishbone → k-nn from root cell, then identify branching.
- Slicer → LLE+entropy criterion.
- Scuba → if capture times available, models non-linear dynamics.

Our two stage approach

- Infer pseudotime
    - Use capture time if available.
- Initialisation: Use overlapping mixture of GPs (OMGP) to infer K trajectories. No branching.
- Create branching model using OMGP to initialise allocation probabilities Φ and kernel hyperparameters θ.
- Use Bayesian optimisation (GPyOpt) to learn θ and branching locations B.
    - Local optimization for allocation probabilities Φ.

Running example

Mouse early embryonic development

We apply our approach on a published dataset of mouse development[a]

[a] Guo, Guoji, et al. 'Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst.'Developmental cell, 2010.

- RT-PCR to quantify expression of 48 genes including 27 development transcription factors.
- 438 cells extracted at seven time points corresponding to cell-doubling events.
- Two branching points at 32-cell stage, differentiation to trophectoderm (TE) and inner cell mass (ICM) and at 64-cell state, ICM branch differentiates to primitive endoderm (PE) and epiblast (EPI).

Using capture times

DeLorean

- In Reid et al, 2016[a] a Gaussian process latent variable model was used to infer pseudotime when an informative experimental capture time is available.
- Consistent way of incorporating prior information.
- Generic approach leveraging the STAN probabilistic language.

[a] John E. Reid and Lorenz Wernisch 'Pseudotime estimation: deconfounding single cell time series', Bioinformatics (2016) 32 (19): 2973-2980

Using capture times

Improvements

- Extend the variational Bayesian GPLVM[a] for a non-standard Gaussian.
- Allows for an analytic lower bound calculation whereas the generic STAN bound is numerically estimated.
- Bound is calculated via a reduced set of auxiliary variables that allows the method to scale up as the number of cells is increased.
- In Reid et al, a second approximation is performed to reduce the computational complexity (FITC kernel).
- Implemented in scalable architecture[b]

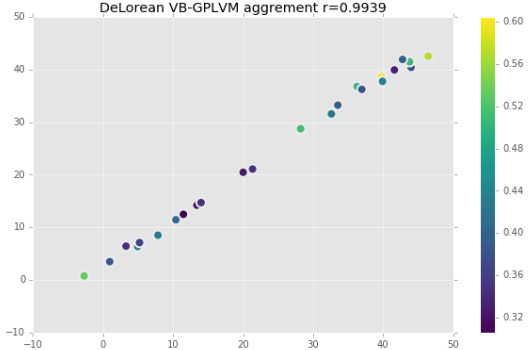
[a] Bayesian Gaussian Process Latent Variable Model. MK Titsias, ND Lawrence, AISTATS, 2010

[b] GPflow by James Hensman, Alex G Matthews, ...

Computational results

Elapsed time in seconds.

Data	Genes	Cell	DeLorean	Bayesian GPLVM
Windram	100	24	120	14
Guo[1]	48	440	1191∼20 mins	74



[1] Using 40 inducing points.

Learning pseudotime without capture times via Topslam (Max Zwiessele)

Density in Bayesian GPLVM

Probabilistic dimensionality reduction technique allows for estimating the density of the landscape, depicted in gray shading of the background of the two dimensional image of the landscape. Light areas are preferred by the algorithm, whereas dark areas increase the between cells distance in the landscape.

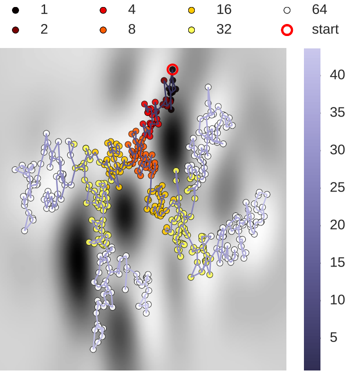
Minimum spanning tree

The extraction of time is done by shortest paths along the extracted graph, depicted in the blue shading of edges, starting from the red circled starting cell.

Learning pseudotime via Topslam (Max Zwiessele)

Key assumptions

- Bayesian Gaussian process latent variable model to reduce dimensionality.
- Minimum spanning tree to infer pseudotime: temporal order from snapshot view of gene expression.



Our two stage approach

- Infer pseudotime
    - Use capture time if available.
- $\boxed{\text{Initialisation}}$: Use overlapping mixture of GPs (OMGP) to infer K trajectories. No branching.
- Create branching model using OMGP to initialise.
- Use Bayesian optimisation (GPyOpt) to learn θ and branching locations B.
    - Local optimization for allocation probabilities Φ.

Initialisation in latent space: Overlapping mixture of GPs

Data association

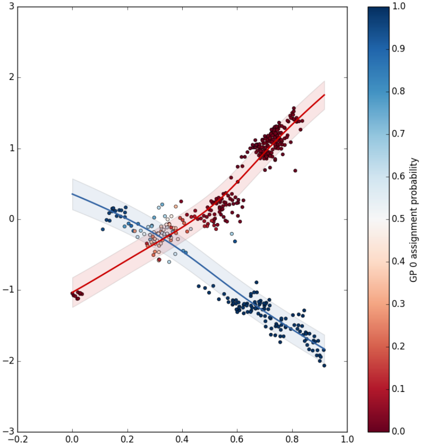
The OMGP[a] seeks to label observations according to the sources that generated them.

[a] Lázaro-Gredilla, Miguel, Steven Van Vaerenbergh, and Neil D. Lawrence. 'Overlapping mixtures of Gaussian processes for the data association problem.'Pattern Recognition 45.4, 2012.

- Flexible non-linear model based on a set of independent GPs.
- The number of sources can be specified or inferred. The latter tends to overestimate the number of functions.
- Fast variational approach using natural gradients available.

Initialisation in latent space: Overlapping mixture of GPs

Working on the pseudotime and principal GPLVM direction.



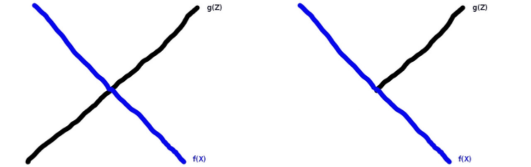
Our two stage approach

- Infer pseudotime
  - Use capture time if available.
- Initialisation: Use overlapping mixture of GPs (OMGP) to infer K trajectories. No branching.
- Create branching model using OMGP to initialise.
- Use Bayesian optimisation (GPyOpt) to learn θ and branching locations B.
  - Local optimization for allocation probabilities Φ.

Inferring perturbation time

Perturbation time

- Yang et al[a] developed a tractable GP model for the identification of a single perturbation point.
- Define a novel kernel that constrains two functions f and g to cross at a single point.
- Bifurcation point is identified by numerically approximating the posterior and selecting a point estimate. This is a model selection approach.



[a] Yang, et al. 'Inferring the perturbation time from biological time course data.'Bioinformatics, 2016

Inferring perturbation time II

- Hyperparameters are estimated by assuming the two functions are independent, that is they do not cross.
- Model used 'to identify at which time point a gene becomes differentially expressed in time course gene expression data under two various conditions.'
- Both GPy and R implementations are available

Key assumptions

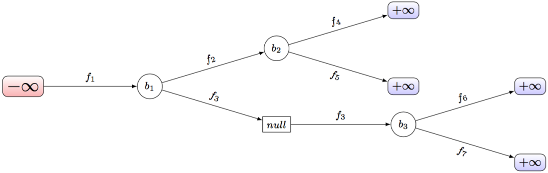
- All data points have been labelled as to which function (f/g) they belong to.
- The ordering of time points is assumed known and fixed.

Branching Gaussian processes

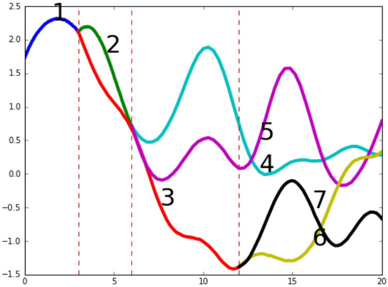
- Extend the kernel to multiple branching points assuming a tree structure.
- Infer the function labels.
- Perform efficient inference via optimisation of a variational lower bound.
- Efficient implementation using GPflow.

Tree prior

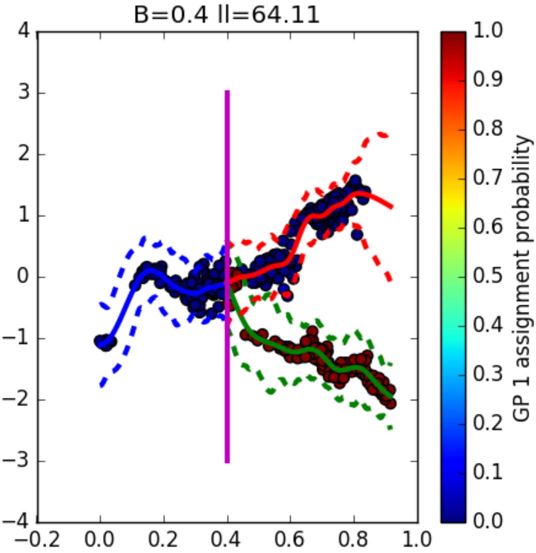
Notional prior



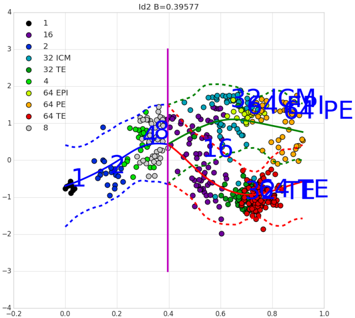
Sample from the model

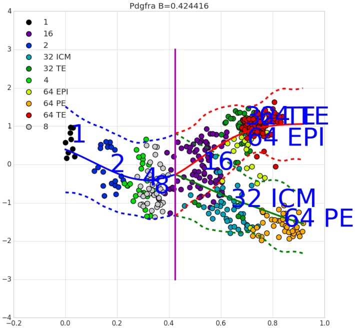


Learning the branching structure



Individual gene branch reconstruction





In SCUBA also found to be have high relative weight for branching.

Summary

- Infer pseudotime which allows an unsynchronised or partly synchronised cell population to be placed on a developmental continuum.
- Branching model to identify the earliest developmental point where cell fate decisions are evident and rank genes in terms of earliest divergence.
- Quantification of uncertainty of branching location.
- Easy to extend to multiple branching points: harder optimization problem using same objective function on higher dimensions.
- Sparse GP approach (Sparse GP) to improve performance, e.g. Drop-seq=50k cells.

Extensions

- Compare to Diffusion pseudotime, Waterfall and Slicer and other approaches.
- Pseudotime inference; jointly identify labels and time order.
- Constrain derivatives to be the same at crossing points so transitions are smooth at branching points.
- Different kernels in tree structure via model selection; e.g. periodic vs non-periodic kernels.
- Stochastic process prior on trees. Place non-parametric prior on tree structure and perform inference on tree structure as well as branching GP.