Bayesian Gaussian Process Latent Variable Model for Pseudotime Inference of Cancer Progression

Melissa Slaughter and Larry Zhang
October 20, 2016

1. Introduction

Cancer is a worldwide problem with deadly consequences. Approximately 10 million people are diagnosed with cancer yearly, of which 6 million people die [1]. In many cancer-related cases, the disease is poorly prognosed, and so treatment cannot occur until the cancer is already at a lethal stage. Unfortunately, current cancer stage classifications are coarsely differentiated and thus do not accurately reflect the continuous and variable progression of cancer. Overall, this makes treatment selection very difficult. Thus, it is important to accurately determine a finer level of cancer progression, and even to predict treatment responses.

One possible way to artificially measure cancer progression is the idea of pseudotime. Pseudotime ordering assigns a score between 0 and 1 to each observation, where values near 0 indicate the start of a biological process and values near 1 indicate the end of a biological process. For example, pseudotime ordering has been used fairly extensively in determining a cell's progression through differentiation or apoptosis [2].

Current attempts at pseudotime ordering algorithms take very different approaches: *Monocle* [2] and *TSCAN* [3] use Independent Component Analysis with Minimum Spanning Trees; *Wanderlust* [4] uses *k*-nearest-neighbors graphs; *embeddr* [5] uses Laplacian Eigenmaps and non-parametric curve fitting. While all of these methods have been shown to provide pseudotime estimates, they cannot evaluate the uncertainty of their estimates. One proposed method for determining pseudotime estimates with their corresponding uncertainty is by using a Gaussian Process Latent Variable Model (GPLVM) [6] to perform Bayesian inference on the pseudotimes [7].

Thus, we propose developing a GPLVM to predict cancer progression from various cancer data sets. We would like to expand our model to extend beyond just RNA-seq analysis, incorporating features from clinical, miR-seq, and mutation profile data. We

would also like to attempt applying our GPLVM to other biological properties that can be modeled along a pseudotime axis, such as treatment outcome.

2. Specific Aims

Aim 1: Develop a GPLVM to make pseudotime predictions of cancer progression in cancer data sets.
We will base our GPLVM on the pseudotime ordering algorithm described in [7]. We will train and test our predictive model using data from three cohorts of cancer: breast invasive carcinoma, lung squamous cell carcinoma, and skin cutaneous melanoma. Initially, we will only focus on RNA-seq data to develop our GPLVM.

Aim 2: Expand our pseudotime ordering algorithm to incorporate other cancer features.
As cancers have more than just RNA-seq data, we aim to expand the feature space of our GPLVM. Currently, we plan to also include clinical, miR-seq, and mutation profile data in order to improve the applicability of our algorithm, as well as the overall accuracy of our predictions. The specific features that we ultimately use depend on how effective those features are in predicting cancer progression.

Aim 3: Apply our pseudotime ordering algorithm to predicting treatment outcomes.
In addition to progression, there are many other aspects of cancer that would be useful if modeled along a pseudotime axis. One such aspect that we are interested in modeling is treatment outcomes. This refers to how well a patient responds to a particular treatment, as well as the overall health of the patient following the treatment. Note that data concerning treatment outcomes may be difficult to acquire.

Aim 4: If time permits, compare our GPLVM model to non probabilistic pseudotime ordering algorithms.
As mentioned above, current pseudotime ordering strategies do not provide uncertainties in their predictions. Thus, it would be interesting to compare our Bayesian model with current algorithms and investigate any differences that arise in the resulting pseudotime predictions.

3. Research Strategy

3.1 Significance

Current measures of cancer progression are extremely coarsely defined (i.e. cancer stages). Thus, a more fine-tuned evaluation of a cancer's progression level can be invaluable for both doctors and researchers. While an actual physical measurement of cancer progression may not be feasible, the idea of pseudotime as an artificial measure of cancer progression may serve as an accurate representation. Thus, our pseudotime predictions of cancer progression can help evaluate the types of treatment and likely responses for cancer patients.

Interestingly, likely treatment responses can also be represented on a pseudotime axis. If our model is capable of accurately predicting these treatment outcomes, it can serve as a very valuable tool in immunotherapy research and in clinical settings.

3.2 Innovation

Currently, most pseudotime inference methods simply assign estimates, but cannot evaluate the uncertainty of their estimates. Furthermore, the use of pseudotime ordering algorithms is very limited to modelling single-cell progression, such as investigating cell fate decisions [2]. Thus, we aim to implement a novel pseudotime ordering algorithm using Bayesian inference (GPLVM), and expand its use to predicting non-single-cell properties, such as cancer progression and treatment outcome. To date, GPLVM has not been used for inferring either of these properties. We also plan to utilize multiple sources of patient profiling (tumor gene expression patterns, mutational profiles, clinical data) for our predictions, while many current methodologies focus on only one source (commonly RNA-seq).

3.3 Approach

First, we will develop a GPLVM to predict cancer progression for three types of cancer based on their RNA-seq data (breast invasive carcinoma, lung squamous cell carcinoma, and skin cutaneous melanoma). We will evaluate the effectiveness of our algorithm by comparing the resulting pseudotime ordering with current coarse stage classifications.

It is possible that RNA-seq data alone will not serve as a good indicator for predicting cancer progression. Thus, we will expand our feature space to include additional data from the three types of cancer including clinical, miR-seq, and mutation profile data. We hope that by doing so, we can improve the relevance and accuracy of our model. Current GPLVM implementations are limited to a single set of features (i.e. only RNA-seq), and so we will have to modify the existing implementation to expand our feature space. We can again evaluate the effectiveness of our algorithm by comparing the resulting pseudotime orderings with both the current coarse stage classifications and the results from our previous implementation.

In addition to inferring cancer progression, we can attempt to apply our model to other properties that can be represented with pseudotime, such as patient responses to treatments. We can do this by mapping the above feature space to a new pseudotime axis, where the assigned pseudotime represents the probability of a treatment's success. We will need to acquire patient response data for specific treatments, which may be difficult. We will use the collected patient response data to evaluate our model by comparing the pseudotime ordering resulting from our model with the actual patient response data.

Finally, time permitting, we will compare our GPLVM to other non probabilistic pseudotime ordering algorithms, such as those described in the introduction.

4. Resources

4.1 Datasets and Computational Resources

We will be using Kieran Campbell's gpseudotime repository on Github as a basis for our algorithm. To implement our model, we will use Edward, a probabilistic programming language framework written in Python. Both expression and clinical datasets for various cancers will be obtained from TCGA.

As we do not expect our project to be computationally heavy, we plan on performing all computational analysis on our personal laptops.

4.2 Lectures

Our research is relevant to a few lectures taught this semester. Primarily, we will be focusing on expression analysis and Bayesian classification (Lecture 6), specifically in

cancer genomics and single-cell sequencing (Lecture 23). We will hopefully also employ some Bayesian inference methodologies that will be taught in Lecture 12.

4.3 Advisors

We are thankful to Yue Li and Alvin Shi, both from the Kellis Lab in CSAIL at MIT, for their guidance in statistical machine learning models and cancer genomics.

5. Timeline

Date	**Goal**
10/20 (Thu)	First draft of project proposal due
10/27 (Thu)	Finish code and literature review of existing methods
11/3 (Thu)	Begin initial implementation of GPLVM and pseudotime ordering algorithm to predict cancer progression
11/10 (Thu)	Final draft of project proposal due
11/17 (Thu)	Complete initial implementation of pseudotime ordering algorithm
11/23 (Wed)	Midcourse report due
12/1 (Thu)	Expand pseudotime ordering algorithm to additional feature sets
12/8 (Thu)	Apply pseudotime ordering algorithm to predict treatment outcomes
12/11 (Sun)	Final report due
12/13 (Tue)	Final presentation

6. Collaboration Plan

We intend to meet at least twice a week to share and discuss progress toward weekly goals. Our project is fairly linear in work progression, and so it requires that we work together closely to develop our algorithm. We plan to work in parallel as much as possible, and thus our frequent meetings will allows us to address blockages and other issues as they arise. We also plan to consult Yue and Alvin regularly for overall advice and feedback on our implementation and analysis.

7. References

1. Steward, B. W. and Kleihues, P. (Eds): World Cancer Report. *IARCPress*. Lyon 2003.
2. 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).
3. Ji, Z. and Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Research 44 (13), e117. ISSN: 1362-4962 (May 2016).
4. 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).
5. Campbell, K., Ponting, C. and Webber, C. Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles (submitted). 2015.
6. Lawrence, N. D. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems* **16**, 329–336 (2004).
7. Campbell, K., Yau, C. Bayesian Gaussian Process Latent Variable Models for pseudotime inference in single-cell RNA-seq data (submitted). 2015.