Unsupervised Learning of Cellular Development Time

Caleb Lareau

11 May 2017

Abstract

Recent advances in microfluidic technologies have enabled an unprecedented characterization of epigenomic and transcriptomic profiles of single cells. Though the dimension of the single cell feature space often exceeds 20,000, a commonly desired yet unobserved variable associated with cellular development time must be estimated using computational methods. Herein, I motivate the approximation of this latent dimension termed *pseudotime* and contextualize its inference as an application of unsupervised learning. Moreover, I present the statistical basis of both linear (principal component analysis; PCA) and non-linear (Gaussian process latent variable modeling; GPLVM) dimension reduction from a probabilistic perspective. To evaluate the performance of the methods of unsupervised learning in this context, I examine the inferred pseudotime against the true developmental ordering of murine embryonic stem cell maturation.

1 Introduction

A fundamental question in developmental biology is how organisms such as humans and mice that originate as single zygotes mature into complex entities composed of trillions of cells in adulthood. Moreover, identifying the longitudinal genetic and epigenetic factors that control stages of development is critical for understanding disease presentation and manifestation. Recent technological advances have enabled a high-throughput characterization of single cells, providing a formidable means of identifying factors critical in complex organism development and disease. In particular, the breakthroughs in microfluidic capture and sequencing of RNA from single cells (scRNA-Seq) has incited the development of a "Human Cell Atlas," which expects to profile an estimated 35 trillion cells over the next decade from various stages of human development. [1]

With this unprecendented source of human cellular data, robust statistical and computational frameworks are needed to extract meaningful structure from the Human Cell Atlas and similar single cell profiles to answer these fundamental questions in developmental biology. In particular, low-dimension latent features like cell-cycle stage and developmental time are often desired to be inferred from a higher dimension feature set (*i.e.* $> 20,000$ gene expression values per cell). As these latent characteristics are almost always unobserved from single cell capture technologies, inference of unobserved features motivate the use of unsupervised learning approaches to approximate latent variables in high-dimensional data. Here, I examine the utility and statistical framework for unsupervised learning through a probablistic lense.

1.1 Pseudotime ordering of single cells

Though the technical advances in profiling single cell transcriptomes has become markedly higher-throughput over the past decade, vital characteristics that define a cell's phenotype is often unprofiled through these techniques. **Figure 1** provides a graphical overview of a typical experimental design, capture process, and analysis framework for single cell data where a latent variable (color gradient/development time) is left unobserved and must be approximated algorithmically. Aside from specific exceptional experimental settings, both the cell cycle and developmental time of a given cell is left unprofiled though these variables are often of distinct interest to investigators. For example, a computational annotation of developmental time across myeloid differentiation enabled the establishment a "stemness" program in acute myeloid leukemia (AML), which was shown to be significantly associated with patient surivival and relapse-free survival. [2] Other applications of these methdologies have been used to infer novel cell types in blood [3] and the genetic markers that distinguish these cellular processes. Consequently, the availability of scRNA-Seq data coupled with methods of pseudotime inference provide a mechanism for investigating outstanding questions in disease, developmental, and regulatory biology.

As shown in **Figure 1C**, statistical methods are needed to approximate the unobserved latent feature of the data. As the observed features (genes) reflect variation in the unobserved variables, > 15 computational approaches (recently reviewed in [4]) have been developed for inferring the color gradient via methods of unsupervised learning. While many of the proposed approaches perform fairly well in estimating developmental time, few provide uncertainty in the calculations of per-cell pseudotime. As insights from this analysis framework may have clinical [2] implications, reliable methods should seemingly quantify uncertainty in the estimation of pseudotime. As such, I will examine **probablistic** unsupervised learning techniques and their efficacy in estimation of pseudotime for cellular developmental systems.

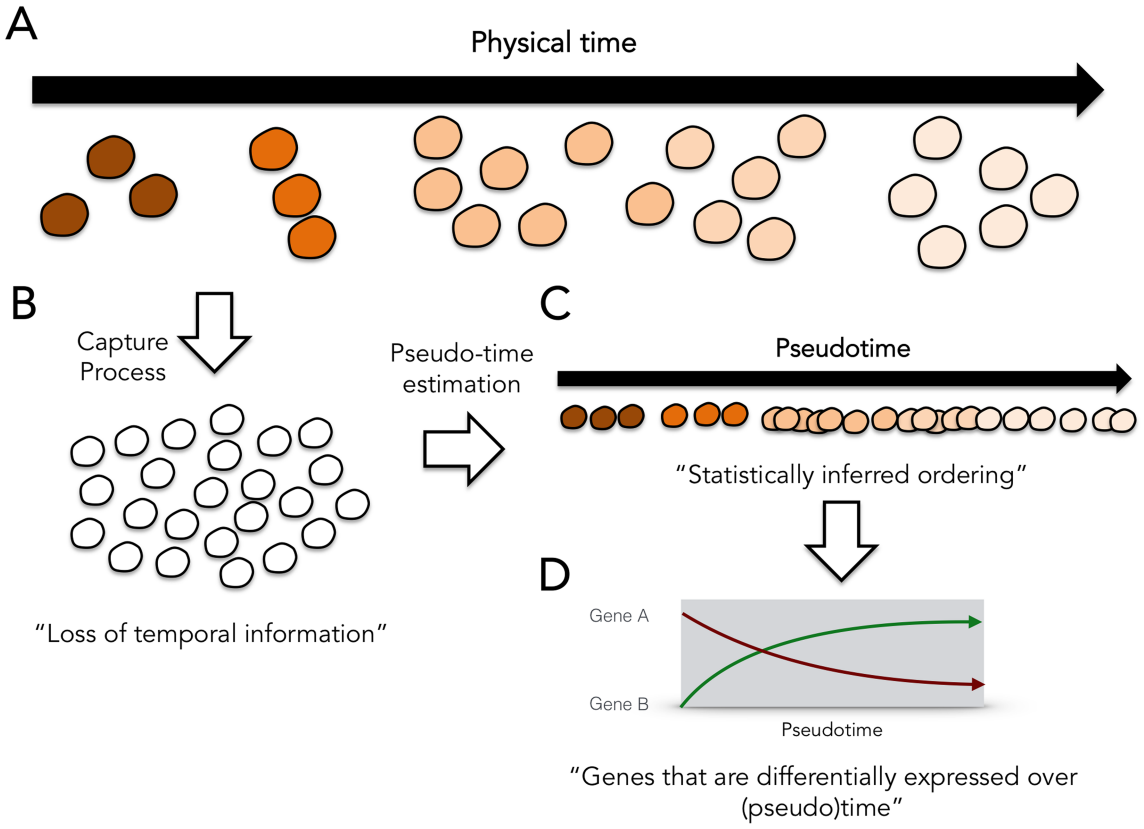

Figure 1: **Overview of single-cell sample collection and developmental ordering.** (A) For a set of cells (each single cell is indicated by a circle), a true developmental time associated with its development is a latent variable (indicated by the color gradient) that accounts for a large proportion of the variance in the transcriptomic profiles. (B) While this developmental time is a true feature of these cells, current technologies do not allow for the capture of this physical/developmental time ordering (aside from rare exceptions). (C) Thus, computational and statistical approaches are required to approximate the developmental time. This approximation is referred to as *pseudotime*. (D) Once ordering of cells can be approximated, genetic factors associated for a developmental process can be inferred. Image reproduced without permission from [5].

1.2 Unsupervised learning

Unsupervised learning is a term used to describe the inference of hidden structure from "unlabeled" data. [6] In contrast to supervised or reinforcement learning, the accuracy of the structure that is inferred by the algorithm cannot be obtained directly in unsupervised learning. Conversely, supervised learning approaches use an outcome of interest (often accuracy) to help guide feature selection and parameter approximations in model building. As the true developmental ordering "label" is typically absent from single cell datasets, unsupervised approaches are required for both feature (gene) selection and parameter estimation in inferring the developmental gradient from the observed feature space.

In classical statistics, a popular form of unsupervised learning is **method of moments** where unknown parameters are related to the moments of random variables. [6] From values of moment-

based estimators and the form of the relationship between random variables and parameters, a straightforward application of unsupervised learning leads to the estimation unknown parameters. In high-dimensional data settings, a canonical technique for unsupervised learning is **principal component analysis** (PCA), which infers a set of orthogonal dimensions that maximize the variance of the samples through a linear transformation of a feature set. [7] Though PCA is adept at inferring structure that accounts for variance in observations (single cells) even with covariance in the feature set (gene expression values), its specification as a linear dimensionality reduction technique has limited its utility in situations where non-linear representations of the observed data may best achieve the learning task at hand (as has been hypothesized with cellular developmental time). [4] To this end, a class of methods called **Gaussian process latent variable models** (GPLVM) have been proven useful to derive structure from high-dimensional data settings in especially in machine-learning contexts. [8] Here, I'll examine the formulation of PCA and GPLVM in a likelihood-based framework, which enables a straightforward approximation of uncertainty in this application of unsupervised learning.

2 Setting

A typical representation of single cell data is embedded in a matrix $\mathbf{Y}$ of dimension M genes x N cells. Elements of this matrix $\mathbf{Y}$ are populated by the observed gene counts that result from a scRNA-Seq experiment. **Figure 2A** provides an example of such a matrix, which will serve as the data basis for pseudotime estimation approaches. While $M > 20,000$ in most scRNA-Seq experiments, considerable covariation is observed in the gene expression values. Like many high-dimensional settings, some smaller number d ($d < M$) "true signals" are parameters of the data-generating mechanisms associated with the M genes. In scRNA-Seq, two of thse d "signals" or latent variables may be the developmental ordering of the cell and the cell's stage in the cell cycle.

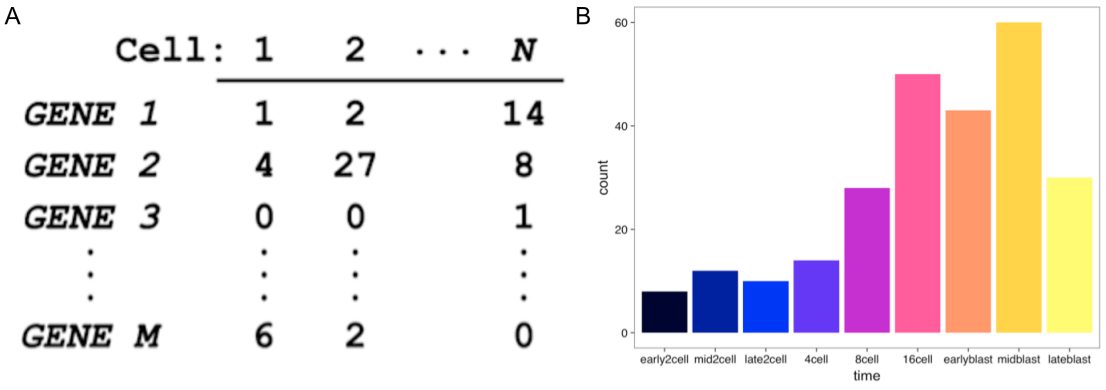

Figure 2: **Data summary for single cell analysis setting.** (A) An example of a data matrix $\mathbf{Y}$ of dimension M genes x N cells. This matrix (which may be normalized by a variety of techniques) serves as the input for unsupervised learning methods discussed herein. Image reproduced without permission from [9] (B) Annotation of the N single cells used in the sample data analysis. Embryonic stem cells were derived from *in vitro* fertilized mice at various stages of embryonic development and M gene transcripts were quantified using the SMART-Seq2 protocol as previously described. [10] The color gradient from dark to light reflects the known developmental ordering (uncommon in most scRNA-Seq experiments) and is consistent throughout the figures presented in this document.

As an example, cells cycling in G2 that are early in development may be highly expressing genes *GENE 3, GENE M-4,* and *GENE M-2* and lowly expressing *GENE 2* and *GENE 15,000* whereas cells in G1 that are also early in development may be highly expressing *GENE 3, GENE M-4,* and *GENE M*. Thus, we seek to compute a matrix $\mathbf{L}$ of d latent variables x N cells that gives an approximation of the reduced dimensionality features responsible for the observed gene expression counts. It has been hypothesized [4] and experimentally verified [11] that the latent variable that explains the most variance in the cells is associated with developmental time in scRNA-Seq datasets (when the cells come from a single, sufficiently dynamic biological system). Thus, we will specify a vector P of dimension 1 x N that is a column vector of $\mathbf{L}$ that explains the most variance in the data to be our estimates of pseudotime for these N cells. P can be thought of as principal component 1 (PC1) or GPLVM1, depending on the method of unsupervised learning. Note that while the cell cycle latent variable has previously been estimated using GPLVM [12] and may be of biological interest, this latent variable is not explicitly considered in this manuscript.

Finally, to evaluate the efficacy of these unsupervised learning approaches, I will examine the the squared correlation (r^2) between the vector P estimated by different approaches and the true developmental ordering of a set of 255 murine embryonic stem cells. **Figure 2B** shows the 9 true developmental timepoints that were available through a carefully controlled experiment as previously described. [10] Notably, this scRNA-Seq dataset is exceptional in that these developmental annotations are available and a feature of the dataset whereas this annotation is typically unobserved in most scRNA-Seq settings. The true developmental ordering will be represented by an integer vector of dimension 1 x N with elements $\{1, 2, ..., 9\}$ (1 = early2cell; 9 = lateblast; see x-axis of **Figure 2B**), which will be correlated with P.

3 Methods

Over a dozen computational approaches have been proposed to infer pseudotime from scRNA-Seq data (recently reviewed in [4]) though nearly all of these methods 1) do not provide an explicit form of pseudotime inference (and rely on algorithmic approaches) or 2) fail to provide uncertainty in the resulting pseudotime estimations. Thus, I will examine a set of methodologies, specifically PCA and GPLVM, that represent pseudotime calculation in a likelihood-based framework to enable a probablistic interpretation of this computed dimension.

3.1 Principal component analysis

PCA is a form of unsupervised learning through linear dimensionality reduction. Though many different methods have been proposed to compute the principal components of a matrix, the following approach is most common. For a matrix $\mathbf{Y}$ with dimension M x N (as shown in **Figure 2A**), the cell-cell covariance matrix can be computed–

$$\mathbf{\Sigma} = E(\mathbf{Y^T Y}) - \mu^\mathbf{T}\mu$$

where $\mu = E(\mathbf{Y})$. Next, the spectral decomposition of the covariance matrix $\mathbf{\Sigma}$ can be computed using the following form–

$$\mathbf{\Sigma} a_j = \lambda_j a_j$$

for $j \in (1, ..., N)$. Then a_j represent the eigenvectors of the data matrix $\mathbf{Y}$. Note the ordering of j

is meaningful–

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d \geq ... \geq \lambda_N \geq 0$$

Thus, the reduced dimensionality matrix $\mathbf{L}$ will be synthesized of the eigen vectors a_j, and d (the number of statistically significant principal components) can be determined using a variety of approaches, including the jackstraw. [13] Notably, our pseudotime vector P for our PCA computation will simply be the first PC (specifically a_j^T).

3.2 Factor analysis

Notably, the common form of performing PCA presented in **3.1** does not present a likelihood framework, hindering a probablistic interpretation of the components, specifically, P. However, a probablistic framework for PCA follows from a different route of inferring a linear, reduced-dimensionality set of variables.

In the 1940s and 1950s, Whittle [14] and Young [15] provided a likelihood-based framework that enabled the derivations of principal components using a factor analysis framework. Notably, this likelihood-based framework enabled a probablistic interpretation of PCA, enabling uncertainty in both the components and the loadings to be estimated from a posterior distribution.

$\mathbf{Y}$ is a D x n matrix. $\mathbf{x}$ is a d x n matrix ($d < D$). Want to relate $\mathbf{x}$ and $\mathbf{Y}$ and assume that the relationship is linear–

$$\mathbf{Y} = \mathbf{W}\mathbf{x} + \epsilon$$

 where W is a D x d matrix.

By convention (after locating the matrix),

$$\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$$

 where $\mathbf{I}$ is the identity matrix of dimension d x d.

3.3 Probablistic PCA

pPCA [16]

ψ_i element of the diagonal of Ψ; add constraint $\psi_i = \sigma^2$

 Assume σ^2 known, MLE yields same W as PCA

$$\mathcal{L} = \frac{-N}{2}\{D\log(2\pi) + \log|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{\Sigma})\}$$

$$\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathbf{T}} + \mathbf{\Psi}$$

$$\mathbf{x} = \mathbf{C}^{-\mathbf{1}}\mathbf{W}^{\mathbf{T}}\mathbf{Y}$$

Probablistic PCA (Tipping and Bishop) - Assuming σ^2 may not be reasonable; want to estimate it from the data (keep in likelihood) - Can estimate $\mathbf{W}, \sigma^2$ using EM and with a prior over $\mathbf{x}$

For $\mathbf{W}_{\mathrm{MLE}}$,

$$\sigma^2_{MLE} = \frac{1}{D-d}\sum_{j=d+1}^{D}\lambda_j$$

- Similarly, we can integrate over $\mathbf{W}$ given a prior, yielding

$$\mathbf{x} \sim \mathcal{N}(\mathbf{C^{-1}W^T Y}, \sigma^2 C^{-1})$$

aaa

Computational methods for inferring principal components under this likelihood framework are availabel through the `pcaMethods` `R` package.

3.4 Bayesian PCA

Though not utilized in this particular data application to remain purely unsupervised, a corollary of the pPCA derivation is Bayesian PCA. In brief, the statistical framework for performing linear dimensionality reduction in the likelihood framework above can be modified by specifying the distribution on these reduced dimensions *a priori*. [17] Methods for computing principal components with Bayesian priors and interpretations are also accessible through the `pcaMethods` `R` package.

3.5 Gaussian Process Latent Variable Modeling

$$\mathbf{Y} \sim \mathcal{N}(0, \mathbf{C}), \mathbf{C} = \mathbf{WW^T} + \mathbf{\Psi} \tag{1}$$

$$\mathbf{C}(\mathbf{W_i}, \mathbf{W_j}) = \mathbf{W_i^T W_j} + \sigma^2 \delta_{ij} \tag{2}$$

$$\mathbf{C}(\mathbf{W_i}, \mathbf{W_j}) = \theta_{rbf} \exp\Big(\frac{-\gamma}{2}(\mathbf{W_i} - \mathbf{W_j})^T(\mathbf{W_i} - \mathbf{W_j})\Big) + ... \tag{3}$$

- (2) being a special case (linear, iid) of (3) - Computationally more challenging, but there are fast algorithms out there

4 Data Application

From mice [10] A small simulation study or application to existing dataset.

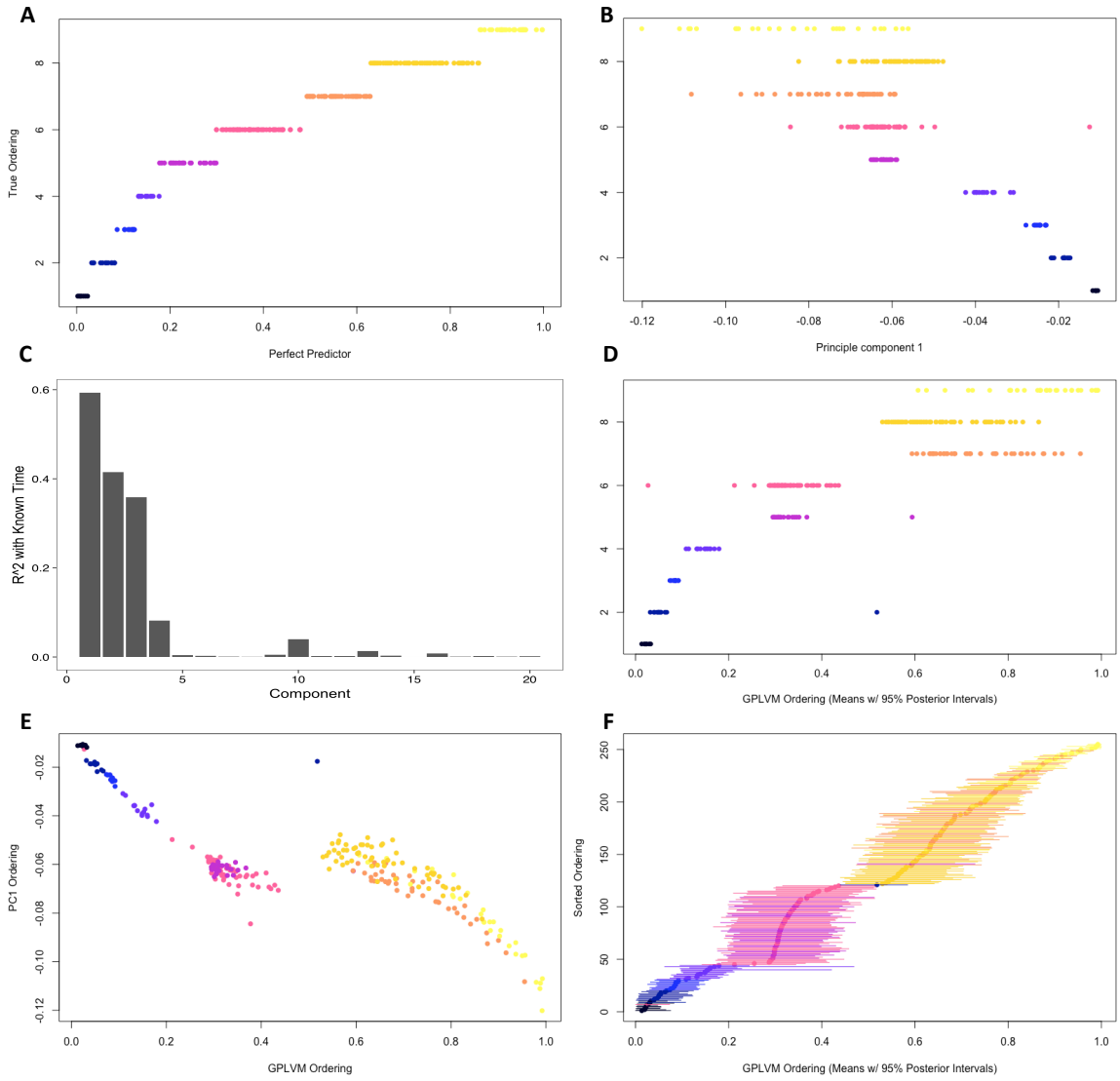

Figure 3: **Summary of unsupervised learning results applied to murine embryonic stem cell development.** (A) An example of a "perfect" latent dimension (x-axis) where the true ordering of the nine embryonic states (y-axis) is inferred ($r^2 = 0.88$). (B) The comparison of the true developmental ordering (y-axis) against a linear unsupervised learning dimension (principal component 1), indicating a decent approximation of true developmental ordering ($r^2 = 0.59$). (C) Barplot depicting the r^2 for the first 20 principal components similar to what was shown for PC1 in (B). This plot suggests that a non-linear dimension reduction may best approximate the true developmental ordering. (D) The comparison of the true developmental ordering (y-axis) against a non-linear unsupervised learning dimension (Gaussian Process latent variable 1), indicating a better approximation of true developmental ordering ($r^2 = 0.78$). (E) Comparison of linear (y-axis; PC1) and non-linear (x-axis; GP latent variable 1) reduced dimensions. The non-linear unsupervised learning technique separates day 8 and day 16 development from the blastocyst samples. (F) Uncertainty associated with GP latent variable 1 depicted with the 95% posterior confidence interval. The uncertainty computed in the posterior distribution provides a probablistic measure for pseudotime ordering of these single cell samples.

5 Discussion / Conclusion

In brief, the application of GPLVM provided a good approximation of the cellular developmental time from the specificed sample of murine embryonic stem cells. Notably, the extra flexibility afforded through non-linear feature reduction lead to a better squared correlation ($r^2 = 0.78$) with the true developmental ordering than the linear dimension reduction through PCA. ($r^2 = 0.59$) Moreover, as PCA is a special case of GPLVM, [8] this approach is a flexible, encompassing approach for unsupervised learning that has gained considerable popularity in a variety of data applications outside of single-cell biology.

As previously mentioned, the utility of GPLVM in this data setting extends beyond the estimation of cellular development time but has been shown to approxmiate other latent variabels in single cell gene expression values, including the cell cycle stage. [12] In specific biological applications, GPLVM have been applied to resolve Th1/Tfh cell-fate decisions in mice exposed to malaria, providing novel insights into the genetic perturbations associated with T-cell response and memory to foreign antigens. [11] More generally, Welch *et al*. showed that GPLVM provide a flexible means to infer development ordering in single cells using not just RNA profiles but also chromatin accessibility, methylation, and histone modifications derived from single-cell experiments. [18] Thus, for unsupervised learning in single-cell biology, GPLVM provides a generalized, flexible tool for handling a variety of data types and modeling a variety of desired structures with statistical ease and confidence.

An important feature of GPLVM as shown in the **Methods** section of this document is the formal probalistic interpretation of the latent dimensions inferred from this class of unsupervised learning. As such, uncertainty in the estimates of these variables can be quantified and visualized as shown in **Figure 3F**. These uncertainty estimates are a vital component of uncovering true variation associated with dynamic processes in developmental biology, such as cancer progression [2] and hematopoiesis. [3] The application of GPLVM to our murine embryonic scRNA-Seq data suggests that resolving meaningful gene expression differences within *a priori* labeled populations may not be sensible. However, differences between classes of labels (*i.e.* differences in gene expression between early2cell, 8/16cell, and blastocyst) may evaluated with statistical confidence.

As the richness and depth of large datasets become more prevalent in the 21st century (*e.g.* transcript profiles of > 35 trillion cells), sophisticated statistical theory and methodologies must be developed to extract meaningful structure from these data. While carefully considered and motivated through an application to single-cell biology, methods of unsupervised learning have become a key component of many data analysis tasks for a large variety of data types. The form of pPCA and GPLVM provides a facile integration into well-established likelihood-based methods for high-dimensional data settings, making these techniques a worthwhile addition to the biostatistician's toolkit.

Accessibility

All code and data used to generate the figures, slides, and writeup are made publicly available at `https://github.com/caleblareau/BST245_Final`. An `R` package for running GPLVM (tailored to single-cell data) is distributed through `Github` and can be installed typing the following command into an `R` console: `devtools::install_github("kieranrcampbell/pseudogp")`. The `pcaMethods` `R` package for running PCA, Bayesian PCA, and Probablistic PCA is distributed via `Bioconductor` and is available at `http://bioconductor.org/packages/release/bioc/html/pcaMethods.html`.

References

[1] I. Sample, "Human cell atlas project aims to map the human body's 35 trillion cells.," *The Guardian*, p. Published 14 Oct. 2016; Accessed 02 May 2017.

[2] M. R. Corces, J. D. Buenrostro, B. Wu, P. G. Greenside, S. M. Chan, J. L. Koenig, M. P. Snyder, J. K. Pritchard, A. Kundaje, W. J. Greenleaf, *et al.*, "Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution," *Nature Genetics*, 2016.

[3] J. D. Buenrostro, W. Greenleaf, R. Corces, B. Wu, A. N. Schep, C. Lareau, R. Majeti, and H. Chang, "Single-cell epigenomics maps the continuous regulatory landscape of human hematopoietic differentiation," *bioRxiv*, p. 109843, 2017.

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

[5] K. R. Campbell and C. Yau, "Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference," *PLOS Computational Biology*, vol. 12, no. 11, p. e1005212, 2016.

[6] T. Hastie, R. Tibshirani, and J. Friedman, "Springer series in statistics," *The elements of statistical learning: data mining, inference, and prediction*, 2009.

[7] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[8] N. D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," in *Advances in neural information processing systems*, pp. 329–336, 2004.

[9] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, *et al.*, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015.

[10] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, "Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells," *Science*, vol. 343, no. 6167, pp. 193–196, 2014.

[11] T. Lönnberg, V. Svensson, K. R. James, D. Fernandez-Ruiz, I. Sebina, R. Montandon, M. S. Soon, L. G. Fogg, A. S. Nair, U. Liligeto, *et al.*, "Single-cell rna-seq and computational analysis using temporal mixture modelling resolves th1/tfh fate bifurcation in malaria," *Science immunology*, vol. 2, no. 9, 2017.

[12] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells," *Nature biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.

[13] N. C. Chung and J. D. Storey, "Statistical significance of variables driving systematic variation in high-dimensional data," *Bioinformatics*, vol. 31, no. 4, pp. 545–554, 2015.

[14] P. Whittle, "On principal components and least square methods of factor analysis," *Scandinavian Actuarial Journal*, vol. 1952, no. 3-4, pp. 223–239, 1952.

[15] G. Young, "Maximum likelihood estimation and factor analysis," *Psychometrika*, vol. 6, no. 1, pp. 49–53, 1941.

[16] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[17] C. M. Bishop, "Bayesian pca," in *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pp. 382–388, 1999.

[18] J. D. Welch, A. J. Hartemink, and J. F. Prins, "Manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics," *bioRxiv*, p. 130336, 2017.