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 identification 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 development.

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 from various stages of human development over the next decade. [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 often unobserved from single cell capture technologies, these analysis questions motivate the use of unsupervised learning approaches to approximate latent variables in high-dimensional data.

1.1 Pseudotime ordering of single cells

The establishment of biomarkers related to early versus late development in hematopoesis acute myeloid leukemia (AML). [2]

Recently reviewed in [3]

single cell heme atac [4]

In brief, the inference of cellular pseudotime as a latent variable can be represented as a specific application of unsupervised learning.

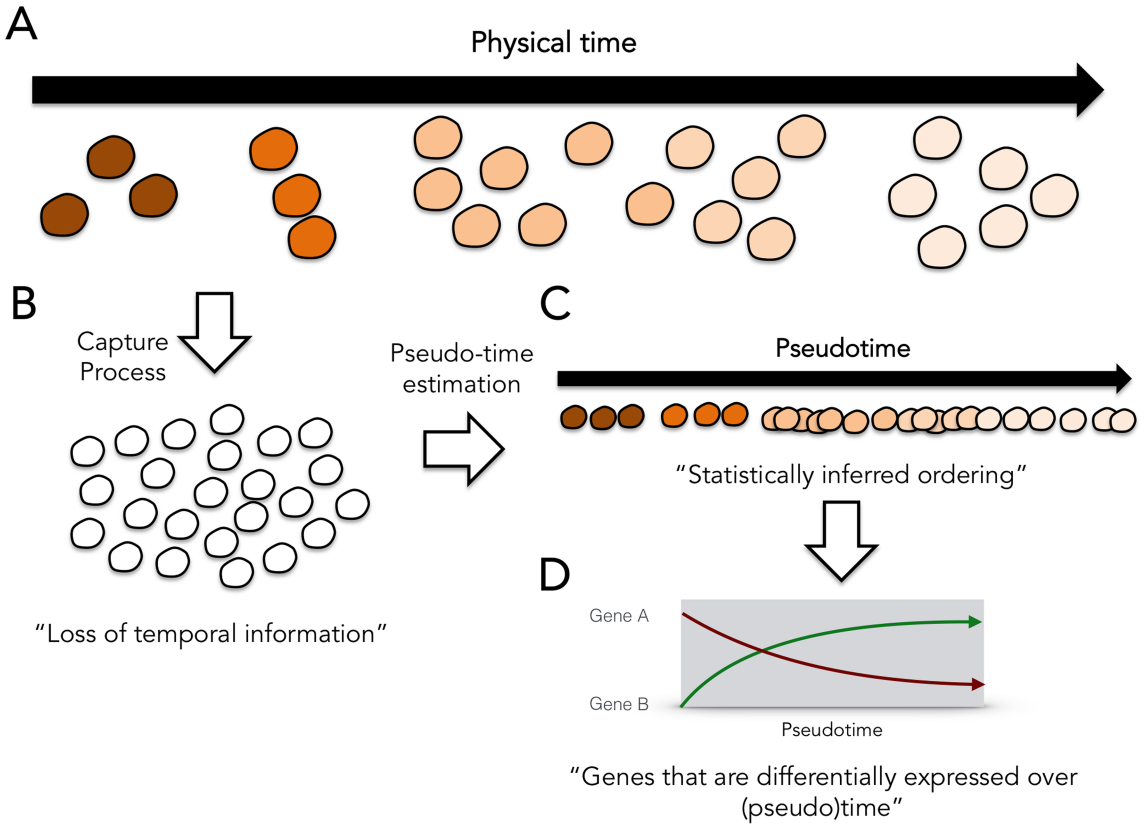

Figure 1: **Overview of single-cell sample collection and developmental ordering.** (A) For a set of cells (each single cell is indicated by a circle), a true developmental time associated with its development is a latent variable (indicated by the color gradient) that accounts for a large proportion 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 (except for 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, significant genetic factors responsible 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. In traditional inferential statistics, a popular form of unsupervised learning is method of moments where unknown parameters are related to the moments of random variables [6]. Using moment-based estimators and the form of the relationship between the random variables and the parameters, one can estimate the values of these unknown parameters.

PCA [7]

pPCA [8]

1.3 Uncertainty in latent variables

In the 1940s and 1950s, Whittle[9] and Young [10] 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.

2 Setting

Introduce your notation and clearly define your assumptions.

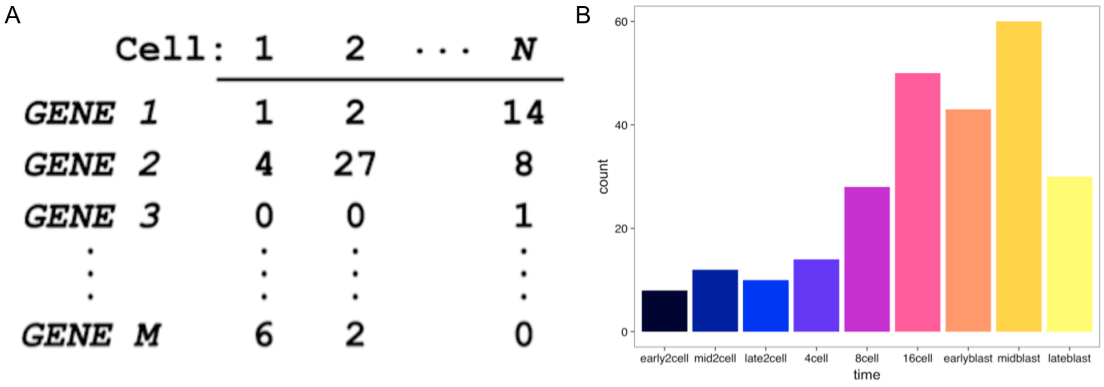

Figure 2: **Data summary for single cell analysis setting.** (A) An example of a "perfect" latent dimension (x-axis) where the true ordering of the nine embryonic states (y-axis) is inferred. (B) The comparison of the true developmental ordering (y-axis) against a linear unsupervised learning dimension (principal component 1), indicating a modest

3 Methods

Over a dozen computational approaches have been proposed to infer pseudotime from scRNA-Seq data (recently reviewed in [3]) though nearly all of these methods 1) do not provide an explicit form of pseudotime inference or 2) allow for uncertainty in the resulting pseudotime estimations.

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 2(A)**), the covariance matrix can be computed–

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

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 $j \in (1, ..., D)$. Then a_j represent the eigenvectors of the data matrix $\mathbf{Y}$. Note the ordering of j is meaningful–

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D \geq 0$$

3.2 Factor analysis

$\mathbf{Y}$ 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

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

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

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 analysis, the derivation of probablistic PCA enables a Bayesian framework for performing linear dimensionality reduction and putting a prior on these reduced dimensions. [11] 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

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

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

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

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

4 Data Application

From mice [12] A small simulation study or application to existing dataset.

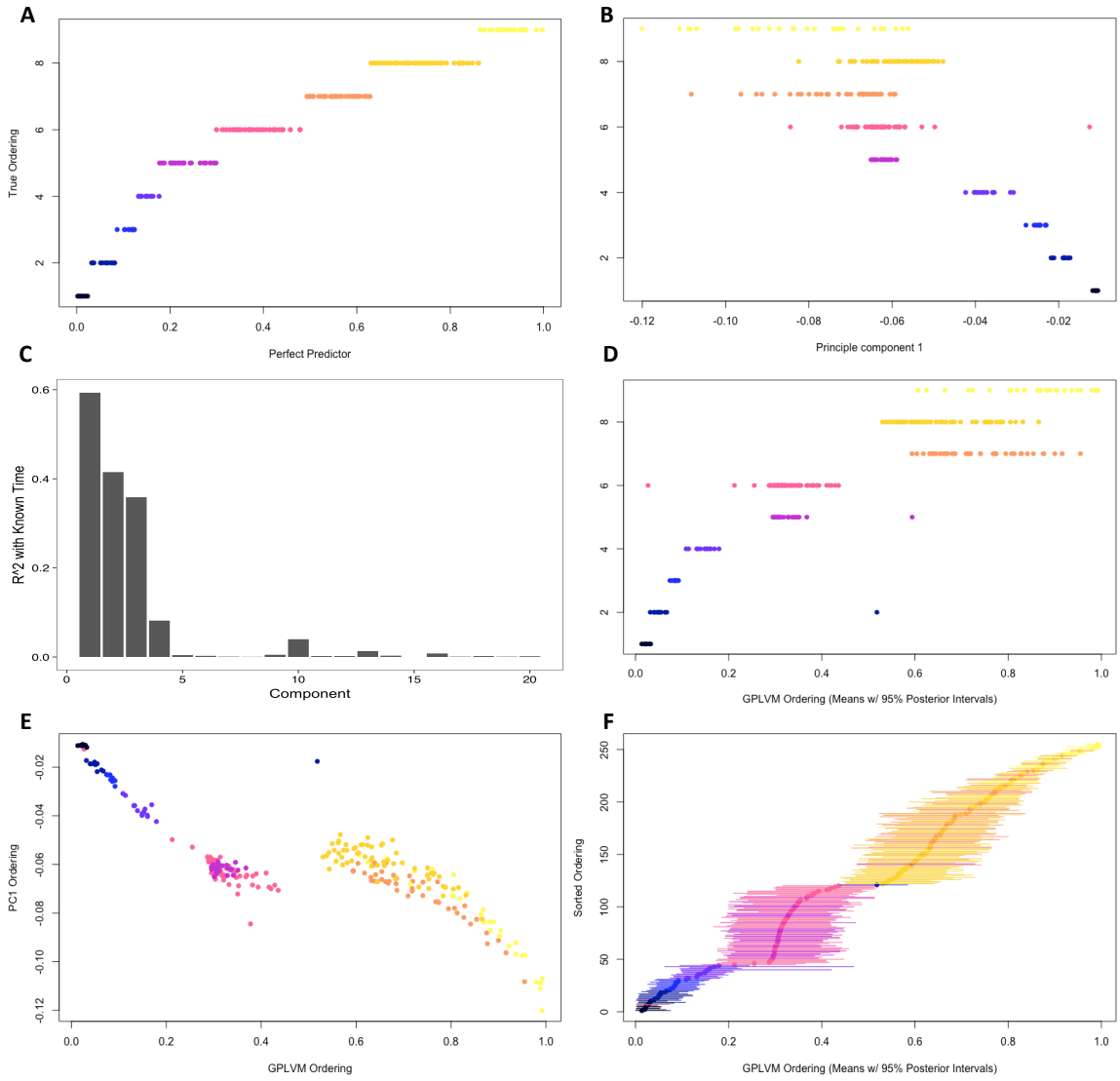

Figure 3: **Summary of unsupervised learning results applied to murine embryonic stem cell development.** (A) An example of a "perfect" latent dimension (x-axis) where the true ordering of the nine embryonic states (y-axis) is inferred ($r^2 = 0.88$). (B) The comparison of the true developmental ordering (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.

GPLVM [13]

5 Conclusion

GPLVM have recently been applied to resolve Th1/Tfh fate bifurcation in mice exposed to malaria, providing novel insights into the genetic perturbations associated with T-cell response and memory to foreign antigens. [14] 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. [15]

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 (specifically 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 made available through `Bioconductor` repository 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] R. Cannoodt, W. Saelens, and S. Yvan, "Computational methods for trajectory inference from single-cell transcriptomics," *European Journal of Immunology*, 2016.

[4] 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.

[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] 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.

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

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.