Recovering Developmental Dynamics from Single-Cell Data via Penalized Principal Curves

Slav Kirov

Monday 14th May, 2018

Background. Modern single-cell technologies offer a detailed view of the conditions and states of thousands of cells at the individual level. Often the full spectrum of developmental stages can be captured in a single snapshot, and a common objective is to temporally order cells according to their progression. Despite the number of existing methods, we are not aware of any that directly recover the typical development along a lineage as a curve in the original (or used) feature space.

Aim. For given single-cell data representing a single lineage process, our aim is to obtain a curve in the space of used features that represents the underlying development.

Data. We analyze a publicly available single-cell mass cytometry dataset consisting of cells undergoing B cell development from 5 human bone marrow samples. Each contains between approximately $2K$ and $20K$ cells and 41 markers of protein expression levels.

Methods. We compute regularized principal curves using a sequential solve of the algorithm of Kirov and Slepčev (2017) for multiple penalized principal curves. For comparison, we also use the results of a pioneering method, Wanderlust, that analyzed the same dataset.

Results. Our results closely resemble those of Wanderlust, and computed trajectories from all samples exhibit canonical marker expression patterns that are consistent with domain knowledge. The results of a feature analysis that follows from our approach also largely agree with literature on the markers' significance to B cell development.

Conclusions. The results suggest that our approach does effectively recover curves representing B cell development. They also suggest potential added benefits, including an extended ability of feature analysis and trajectory comparison across samples.

Broader impacts. With added comparison capabilities, curves representative of development may offer additional insight into how standard cellular processes differ from those that are abnormal, such as cancer.

Keywords: single-cell data, curve reconstruction, nonlinear dimensionality reduction

DAP Committee members:
Ziv Bar-Joseph (Machine Learning)
Dejan Slepčev (Mathematical Sciences)

1 Introduction

Advances in single-cell technologies can provide a view of the states and conditions of thousands of cells at unprecedented resolution. Unlike bulk analyses that average out cell features over large pools, single-cell data provides simultaneous measurements of many cellular features at the individual (single-cell) level. The ability to finely resolve heterogeneities in cell populations presents an opportunity to advance the current understanding of a variety of cellular processes, and many discoveries have already been made. Among these are the identification of new cell populations (Bendall et al. (2014)), refined characterization of cell types (Jaitin et al. (2014)), and discovery of unexpectedly early cell fate commitment of hematopoetic progenitors (Paul et al. (2015)).

When the sampled cells are assumed to be undergoing a common process, a frequent objective is to temporally order cells according to their development. This is the case for asynchronous processes that recur continuously, such as hematopoeisis (blood cell development), for which it is possible to effectively sample the entire process with one single-cell experiment. The subsequent temporal ordering or *trajectory inference* problem, as it is often referred, has lead to the development of a variety of computational methods targeted to both single-lineage and branching processes. Methods are centered around data simplification and dimensionality reduction, as in the case of a single lineage the problem consists of assigning a pseudotime to each cell (i.e., a one-dimensional reduction of the data). To the extent of our knowledge, no existing methods concurrently obtain with the pseudotimes a curve in the original feature space that represents the typical dynamics of a lineage. Having such a representative curve could be desirable for many reasons, especially since it allows for more direct comparisons between developmental trajectories residing in the same space, otherwise unavailable with mere pseudotimes.

In this report we investigate the applicability of an approach for finding *penalized principal curves* for approximating the dynamics of single-lineage cell processes. The curves, regularized via a penalty on their length, are sought to pass close to the "mean" of the data distribution. Our approach to computing them is based on the work of Kirov and Slepčev (2017). We focus our attention to a single-cell mass cytometry dataset from human bone marrow samples in order to recover trajectories for B cell development. Our evaluation will largely consist of comparing the found trajectories with prior knowledge, and with results of Wanderlust, an algorithm for pseudotime inference introduced together with the dataset by Bendall et al. (2014). We then investigate the impact of features on the trajectory, and we compare the computed trajectories from different samples. In doing so, we take advantage of additional analyses allowed by our approach in order to further examine the role of features and appropriately compare the found curves.

1.1 Background and related work

The problem of recovering the temporal order of data with respect to an underlying process has been studied in different contexts dating to the pre single-cell era. Dealing with data from microarray experiments, Gupta and Bar-Joseph (2008) investigated a traveling salesman approach for inferring their temporal order. Another approach for microarray data by Magwene et al. (2003) is based on constructing a minimum spanning tree and using its diameter path to induce an ordering on the data. The method was later extended by Trapnell et al. (2014) to allow for branching processes in

what became the first and one of the most well known trajectory inference methods (Monocle) for single-cell RNA-seq data. Wanderlust, another pioneering method for single-lineage processes, was introduced around the same time, and uses a graph representation of the data and shortest path distances to obtain pseudotimes. Wanderlust was introduced by Bendall et al. (2014) to analyze the mass cytometry dataset we study here, and we provide a brief summary of the algorithm in Section 2.2. Setty et al. (2016) later extended Wanderlust to allow for processes with bifurcating branches, and additionally applied it to single-cell RNA-seq data.

Following Monocle and Wanderlust, a multitude of methods for trajectory inference have been proposed for both single-lineage and more complex branching topologies. A recent article by Saelens et al. (2018) found a remarkable 59 methods (and counting) since 2014. Prior to assigning pseudotimes, the majority of approaches obtain a simplified representation of the data via graph construction, clustering, or other dimensionality reduction (e.g. PCA, ICA, t-SNE). Many approaches combine multiple simplification steps (e.g. clustering after dimensionality reduction), while others use a direct reduction to one dimension to obtain pseudotimes (e.g. via diffusion maps: Haghverdi et al. (2016)). Review articles and comparison studies have been provided by Cannoodt et al. (2016a) and Saelens et al. (2018).

Despite the number and variety of methods for single-cell trajectory inference, to our knowledge none have yet approached the problem by searching for representative curves in the original feature space. This is a more difficult problem (pseudotimes can be obtained from a curve via projection), and high-dimensional noisy single-cell data does not make it any easier. One of the main origins of curve reconstruction is due to Hastie and Stuetzle (1989), who introduced a *principal curve* as a nonlinear generalization of the first principal component, and more specifically, as a curve for which every point is the mean of data that project there. The original formulation can lead to problems of overfitting, and subsequent variants addressing it (for example with regularization: Kegl et al. (2000); Smola et al. (2001); Tibshirani (1992), or reformulation: Gerber and Whitaker (2013)) still faced the obstacle of obtaining good initialization to eventually compute desirable curves. We note that a handful of methods for single-cell trajectory inference do attempt to find principal curves after an initial dimensionality reduction step (Campbell et al. (2015); Cannoodt et al. (2016b); Guo et al. (2017); Marco et al. (2014); Street et al. (2017)), and all do so via the original algorithm of Hastie and Stuetzle (1989).

Our approach is based on a regularization of principal curves through a penalty on their length, studied by Lu and Slepčev (2015), and a relaxed problem that allows for multiple curves, introduced in Kirov and Slepčev (2017). The relaxed setting allows energy-descent algorithms to evade undesirable (high-energy) local minima, by guiding curves to disconnect in low-density regions, and reconnect in higher density regions. The larger configuration space also permits initialization with singletons (that can be set as cluster centers), which can then gradually connect to form a single curve, all in one unified approach that we further describe in Section 2.1.

1.2 Data

We focus our attention on single-cell mass cytometry data that were collected from human bone marrow samples. The data, which were introduced and studied by Bendall et al. (2014), were previously

found online[1], and include measurements of 41 cellular features undergoing B cell development in five healthy human patients. The features consist of marker expression levels of phenotypic surface molecules and internal functional proteins. The five samples corresponding to each patient, labelled A, B, C, D, and G, contain 1,900, 3,436, 10,864, 5,955, and 19,486 cells respectively. The public dataset also contains expression levels of the markers measured after multiple cellular signaling perturbations, but here we restrict our attention to these basal samples.

The algorithm of Bendall et al. (2014), Wanderlust, is our main basis for comparison, and we therefore attempt to replicate some of their tests to the best of our abilities. To that end, we predominantly apply our method to the same 18-feature set that Wanderlust was tested on. These 18 markers are listed in Table 1, and the remaining 23 features, listed in Table 2, were only added to the feature set for fit analysis in Section 3.5.2. We focus our attention to dataset G for much of the analysis because it contains the most cells, and the public datasets A, B, C, D do not contain three of the markers (CD117, CD179b, IgMi) in the 18-feature set used by Bendall et al. (2014). We later omit these three markers from G when comparing results from all datasets in Section 3.6.

2 Methods

We start by first describing penalized principal curves, the algorithm for finding them, and how we use it for the problem here. We then provide a brief description of the Wanderlust algorithm of Bendall et al. (2014).

2.1 Penalized principal curves

In the general case, the multiple penalized principal curve approach of Kirov and Slepčev (2017) seeks a set of curves that minimize a functional consisting of an approximation error term along with regularization terms on the number of curves used, and their length. For given point cloud data $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^d$ with equal weight and parameters $\lambda_1, \lambda_2 > 0$ the functional (or energy) to be minimized is

$$E^{\lambda_1,\lambda_2}(\gamma) := \frac{1}{n}\sum_{i=1}^{n} d(x_i,\Gamma)^2 + \lambda_1\left(\mathrm{L}(\gamma) + \lambda_2\left(\mathrm{k}(\gamma)-1\right)\right), \tag{MPPC}$$

where γ represents a set of curves, $d(x_i,\Gamma)$ represents the Euclidean distance from the x_i to the image of the curves, $\mathrm{L}(\gamma)$ denotes the total length of the curves, and $\mathrm{k}(\gamma)$ is the number of curves in the set γ (cardinality). [Formally, one considers γ belonging to the admissible set

$$\mathcal{A} := \left\{\gamma = \{\gamma^i\}_{i=1}^k : k \in \mathbb{N},\, \gamma^i \in \mathcal{C},\;\; i = 1,...,k\right\},$$

where the class of curves $\mathcal{C}$ is defined

$$\mathcal{C} := \{\gamma : [0,a] \to \mathbb{R}^d : a \geq 0, \gamma \text{ is Lipschitz with } |\gamma'| \leq 1,\ \mathcal{L}^1 - \text{ a.e.}\}.]$$

[1]The data were found at `https://www.c2b2.columbia.edu/danapeerlab/html/wanderlust-data.html`. It seems the website was updated, and we cannot find an alternate public source of the data.

The first term of the functional measures how well the curves approximate the data, while λ_1 controls the complexity of the curves through their length, and together with λ_2, their number.

For our application we are interested in obtaining a single curve to describe the underlying developmental process, so it may seem unnecessary to allow for multiple curves, and one may instead seek a single curve $\gamma \in \mathcal{C}$ minimizing

$$E^{\lambda}(\gamma) := \frac{1}{n}\sum_{i=1}^{n} d(x_i, \Gamma)^2 + \lambda \, \mathrm{L}(\gamma). \tag{PPC}$$

The utility in allowing for multiple curves comes from the non-convexity of the single curve functional (PPC). Computed local minimizers of (PPC) can be very sensitive to the initial curves, and Kirov and Slepčev (2017) show that expanding the configuration space allows curves to disconnect in low density regions of the data, and reconnect in higher density regions. These topological changes are the means by which configurations following energy descent evade high-energy local minima of (PPC).

For further improvement, Kirov and Slepčev (2017) propose initializing with singletons (curves whose image is a single point) found by running an instance of the k-means problem. For a large enough value of λ_2, singletons between regions with high enough density will connect to decrease the (MPPC) energy. We extend this idea here in such a way that we can virtually eliminate the parameter λ_2 and end up with a single curve. That is, we solve a sequence of (MPPC) problems, corresponding to an increasing sequence of λ_2 values, and with λ_1 fixed throughout. We start with a low enough value of λ_2 such that the minimizer of (MPPC) consists of at least ~ 100 singletons. Once the local minimizer is computed, we increase the value of λ_2 just enough so that at least one energy-decreasing connection between curves exists, and is added. We repeat these steps until we obtain a local minimizer consisting of only one curve, which we take as our local minimizer to (PPC). We will refer to this algorithm as sequential solve penalized principal curve (SPPC).

We now provide a very brief description of the algorithm for multiple penalized principal curves of Kirov and Slepčev (2017). The algorithm alternates geometric updates (local curve fitting) with topological updates (connecting and disconnecting curves) that decrease the energy of the configuration at every step. In doing so, one considers sets of k polygonal curves

$$\left\{ \{y^c\}_{c=1}^{k} : y^c = (y_{c,1}, ..., y_{c,m_c}) \in \mathbb{R}^{m_c \times d}, m_c \in \mathbb{N}^+ \text{ for } c = 1, ..., k, \text{ and } k \in \mathbb{N}^+ \right\}$$

and the discrete functional

$$\sum_{c=1}^{k} \left(\frac{1}{n} \sum_{j=1}^{m_c} \sum_{i \in I_{c,j}} ||x_i - y_{c,j}||^2 + \lambda_1 \sum_{j=1}^{m_c - 1} ||y_{c,j+1} - y_{c,j}|| \right) + \lambda_1 \lambda_2 (k-1), \tag{1}$$

where the set $I_{c,j}$ includes all data x_i that project to $y_{c,j}$. Specifically,

$$I_{c,j} := \{ i \ : \ (\forall \tilde{c} = 1, ..., k \, \forall \tilde{j} = 1, ..., m_c) \ ||x_i - y_{c,j}|| \leq ||x_i - y_{\tilde{c},\tilde{j}}|| \},$$

and in case the closest point is not unique an arbitrary assignment is made so that the sets $I_{c,j}$ partition $\{1, ..., n\}$, by for example setting $\tilde{I}_{c,j} = I_{c,j} \setminus \left(\bigcup_{s=1}^{c-1} \bigcup_{i=1}^{j-1} I_i \right)$. With the discrete functional,

the distance to the curves $d(x_i, \Gamma)$ is approximated by $\min\{||x_i - y_{c,j}|| : c = 1, ..., k, j = 1, ..., m_c\}$, and therefore the fidelity term depends on how finely each curve y^c is discretized, given by m_c. In the algorithm, the numbers m_c are chosen large enough so that the curves satisfy a specified limit on the average turning angle (5° for experiments here), and so that topological updates can be appropriately performed. We also note that the computational complexity of the algorithm (per iteration) is $\mathcal{O}(mnd)$, where $m = \sum_{c=1}^{k} m_c$ is the total number of points on the curves. For further details we refer to Kirov and Slepčev (2017).

Once we have a curve, we obtain pseudotimes for the data via projection. That is, for an arc-length parametrized curve $\gamma : [0, L] \to \mathbb{R}^d$, we take as the pseudotime $s_i = \frac{1}{L}\gamma^{-1}(\Pi_\gamma(x_i))$ for data point x_i, where $\Pi_\gamma(x) := \arg\min_{y \in \gamma([0,L])} |y - x|$. In general the endpoints corresponding to $s = 0, 1$ are arbitrary, and for the data we study here, we designate the endpoints based on expected marker expression levels. Specifically, we assign $s = 0$ to the endpoint that has higher expression of the marker CD34 (in accordance with domain knowledge, see Figure 1(a)).

2.2 Wanderlust background

We now provide a very brief description of the Wanderlust algorithm of Bendall et al. (2014). The algorithm constructs nearest-neighbor graphs, on which shortest-path distances are then calculated. In order to align the cells and obtain pseudotimes, an initiator cell is defined by the user, and several "waypoint" cells are randomly chosen that serve as anchors. The waypoints are ordered according to their distance from the initiator cell, and pseudotimes for all cells are then assigned according to weighted distances to the ordered waypoints. This process is repeated until convergence for each graph in an ensemble of nearest-neighbor graphs, and the algorithm outputs final pseudotimes $\in [0, 1]$ as the average of those obtained from each graph.

We note that Wanderlust uses cosine distance ($d_{cos}(x_1, x_2) = 1 - \frac{x_1 \cdot x_2}{||x_1|| ||x_2||}$) in both constructing the nearest neighbor graphs and computing shortest path distances. The stated motivation for using cosine distance was that it is "scale-invariant" (it is invariant with respect to scaling of each argument) while Euclidean and other distances are sensitive to scale. We return to this point this in the following sections.

3 Procedures and Results

3.1 Preprocessing and parameter selection

Before applying the SPPC algorithm, we first transformed the data. We note that this is an essential step, as the original distribution of values for each feature is highly skewed and with long tails. We follow the transformation as described in Bendall et al. (2011), where they apply the function $\text{arcsinh}(\frac{\cdot}{5})$ for mass cytometry data. Although the denominator of 5 seems somewhat arbitrary, we apply this exact transformation since, to the extent of our knowledge, it is also what was used by Wanderlust in Bendall et al. (2014).

Before feeding in the transformed data to SPPC, we also normalize it so that it has standard deviation 1. This relieves us from taking scaling effects into consideration when choosing the pa-

rameter λ, which then has a fairly standard range of $[0.001, 0.1]$. After some experimental runs with values in this range, we settled on the value of $\lambda = 0.02$, and use it for all tests and results below.

3.2 Outline of evaluation and analysis methods

Here we briefly outline the evaluation and analysis methods that we apply in the following sections. In the first section, we compare our results on the dataset G to prior knowledge on markers, and to the results of Wanderlust in Bendall et al. (2014). For the latter, we make qualitative comparisons based on marker trajectories, and we compute correlations between found pseudotimes. Since we are mostly interested in the ranked pseudotimes, we predominantly use the Spearman correlation r_S, and in places also report the Pearson correlation r_P. For brevity, we sometimes use WL and PC to denote results from Wanderlust and our (penalized) principal curve approach.

In the second section, we investigate the impact of the markers on the found developmental trajectory. We do this first through removing features from the trajectory, and computing correlations of the resulting set of pseudotimes to the original. We then investigate the impact of markers through their contribution to the squared error term in the functional, after appropriate normalization.

In the third section, we compare results across all five datasets A, B, C, D, and G. In addition to qualitative comparisons, we also test how consistently each curve orders the cells of different datasets, and perform a similar test of how well the trajectories themselves correspond to each other. We then use the latter to reparametrize and align three of the curves that have near perfect correspondence.

We begin by first summarizing prior knowledge on the used markers.

3.3 Marker prior knowledge summary

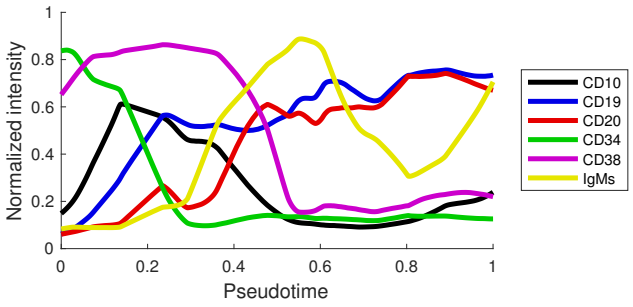
We provide a very brief and basic review of B cell development and domain knowledge on the role of relevant markers. As sources may vary in the number and definition of cell subtypes present at different stages development, we report simplified information on markers that appeared most consistent across references we examined, including Bendall et al. (2014); Coico and Sunshine (2015); Wickramasinghe et al. (2011).

Of the features included in the dataset, there are six canonical markers – CD34, CD38, CD10, CD19, CD20, IgM – whose expression behavior throughout B cell development is somewhat coarsely established. The common understanding of how their expression patterns relate to the development is indicated in Figure 1(a) and is as follows. B cells originate from hematopoetic stem cells in the bone marrow, and differentiate to common lymphoid progenitor (CLP) cells, where they express CD34. Once committed to the B-lineage, they develop into pro-B cells, where CD10, CD19, and CD38 are expressed. In the next pre-B stage, CD34 is no long expressed, and in the final stage before leaving the bone marrow (immature B cell), IgM and CD20 expression occurs, and CD10 expression vanishes.

Other markers not mentioned above that have been closely related to B cell development include CD24, CD40, CD45, CD79a, CD79b, CD179a, IgD, IgMi, Kappa, Lambda, TdT.

	CLP	Pro-B	Pre-B	Imm. B
CD10	−	+		−
CD19		+	+	+
CD20	−	−		+
CD34	+	+	−	−
CD38		+	+	+
IgMs				+

(a)



(b)

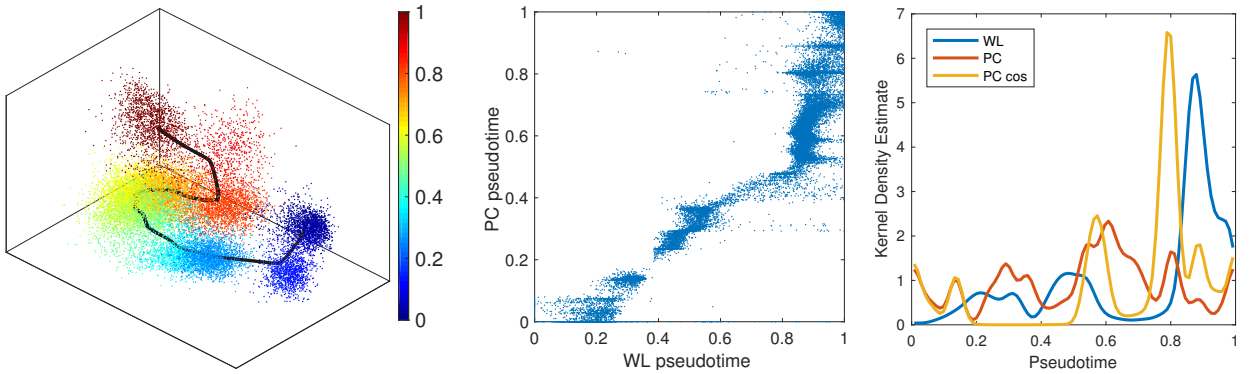
Figure 1: (a) Summary of prior knowledge on marker expression (+ for present, − for absent, blank for mixed or unkown) throughout development, starting from common lymphoid progenitor cells and ending with immature B cells. (b) Normalized marker expression levels of the developmental curve found using our approach. The pseudotimes are proportional to arc length along the curve.

3.4 Results on G and comparisons to Wanderlust

We first applied our approach on dataset G, using the same 18 markers used by Wanderlust in Bendall et al. (2014). A visualization of the data, computed curve, and induced ordering is shown in Figure 2(a), using the first 3 principal components of the data. In Figure 1(b) we plot the found curve in the dimensions corresponding to each of the six canonical markers earlier reviewed. [The levels of expression of each marker were normalized on a $[0, 1]$ scale that includes 90% of the values (around the mean), and the pseudotimes represent proportion of length along the curve.] The plot approximately coincides with the coarse prior knowledge from Figure 1(a), with expression of CD34 followed by CD38, CD10, CD19, and finally IgMs and CD20. Roughly, it appears CLP's correspond to pseudotimes in $[0, 0.1]$, pro-B cells to $[0.1, 0.2]$, pre-B cells to $[0.2, 0.4]$, while immature B cells occupy the majority of the curve in the interval $[0.4, 1]$.

The pseudotimes we found have a solid correlation with those of Wanderlust ($r_S = 0.90$, $r_P = 0.92$), and in Figure 2(b) we provide a scatterplot of both. One may observe though that the distributions of the found pseudotimes differ signifcantly. Many WL pseudotimes lie between 0.8 and 1, while the density of PC peudotimes also varies noticeably throughout. The discrepancy in distributions is likely due to the use of cosine distance in Wanderlust, as after reparametrizing the pseudotimes using cosine distance along the curve, we obtain a distribution that better resembles that of Wanderlust. Kernel density estimates of these distributions are shown in Figure 2(c).

To allow for a more direct comparison between the WL and PC trajectories, we reparametrized both sets of pseudotimes to have uniform distribution. That is, we ranked the values and normalized them to be between 0 and 1. We plotted the PC trajectory where available, and the median values in a sliding window for WL and PC for the six canonical markers along with CD24 and TdT in Figure 3. Since TdT was not included in the construction of the curve, only the median curves are shown. The PC and associated median trajectory are close for the most part, and seem to differ most at points where the data distribution is skewed. (This is expected since the PC curve should



(a) (b) (c)

Figure 2: (a) A visualization of the computed curve using the first 3 principal components of the data, which are colored based on pseudotimes s_i. (b) A correlation plot of the pseudotimes from WL and our computed curve. (c) Comparison of kernel density estimates of pseudotimes from WL, and from our curve with respect to arc length using both euclidean and cosine distances.

be close to the mean curve when there is not much curvature). The median curves for WL and PC are very similar, and nearly identical in some cases. We did not find any significant differences in the trajectories for markers not included in Figure 3.

The scattered marker values over the PC reparameterization are also shown in Figure 3. One may notice vertical streaks in the scattered feature values, where several cells have the same rank. They are most apparent shortly after 0.2 and 0.8, and correspond to sharp turns, or kinks, in the curve where several points project (they are also visible in the PCA visualization in Figure 2(a)). Similar streaks appear at the endpoints 0 and 1 as well, due to the presence of the length term in the functional, which leads to shorter curves and consequently more data project to the endpoints. Both the endpoint effect and kinks are unfavorable byproducts our approach. The latter simply may exist in penalized principal curves beyond the influence of discretization [Slepčev (2014)]. On the other hand, the endpoint shortcomings are likely corrigible, and we leave carefully addressing them as future improvements outside the scope of our main goals here.

We end this section with a note on the initialization of the algorithm. Since the algorithm uses random seeds in its initialization through k-means, in the presence of noisy data it sometimes does lead to different results. To investigate, we ran the algorithm with 10 different random seeds. Eight of the 10 runs were nearly identical to the results presented here, with both $r_S, r_P > 0.99$. The two other runs, nearly identical to each other, both had correlations $r_P = 0.70$ and $r_S = 0.55$ with the result here, and had notably higher energy (objective functional value) than the eight good runs, which were very close in energy. Despite our efforts to avoid local minima of the (PPC) functional, the noisy dataset here sometimes does pose obstacles. As one can run the algorithm multiple times using different initializations though, it is encouraging that the lowest energy curves we found have good properties (in the context of prior knowledge and comparison to WL).

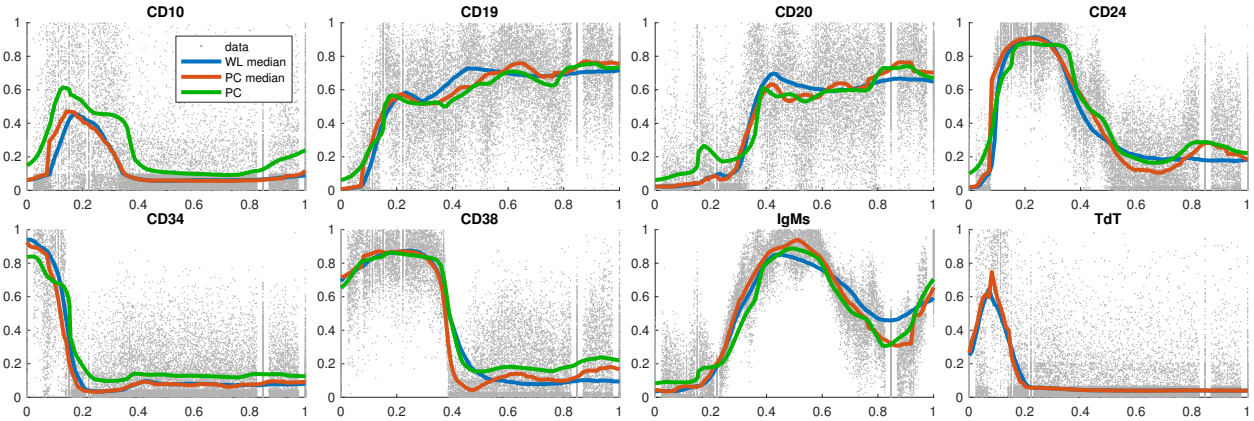


Figure 3: Normalized expression values of six canonical markers, along with CD24 and TdT, over ranked pseudotimes. The plots include scattered marker values as ordered by PC, the computed curve where available, and the median values in a sliding window for both WL and PC. The normalization is such that the range $[0, 1]$ includes 90% of the values around the mean.

3.5 Impact of features: consistency and analysis

Here we investigate the impact of the individual features on the curve that we compute. We test the consistency of the outputted orderings with respect to removal of features, and we also analyze their role by how well they are approximated when they take on equal variance.

3.5.1 Effects of feature removal

We first describe our findings on the impact of features through their removal. One can observe that the variances of the features (if unnormalized) may have a large impact on the curve that minimizes (PPC). For these experiments we therefore report results for both normalized and unnormalized feature variance settings, and we note that the curve found on the normalized 18-feature set gave pseudotimes with high correlation to those found on the unnormalized set ($r_S = 0.95$, $r_P = 0.97$) and to those of WL ($r_s = 0.90$, $r_P = 0.93$) [curves were nearly identical for 9/10 runs with different initial seeds giving lowest energy].

In the first experiment, we computed penalized principal curves by leaving out one feature at a time. For each left out feature, we found a curve and its associated pseudotimes, for which we computed Pearson and Spearman correlations to those of the full (normalized or unnormalized) 18-feature set. The variances of the (unnormalized) features, along with the correlations for both normalized and unnormalized settings are reported in Table 1. The correlations correspond to the lowest-energy curves found using at most three different initialization seeds. With unnormalized variances, the correlations for 15/18 of the markers were very high ($r_S \geq 0.97$). Marker Lambda had a slightly lower rank correlation (0.94), and the two lowest correlations were for IgMs (0.59) and Kappa (0.37) – both of which had higher than average variance (0.13 and 0.10 respectively compared to $\frac{1}{18} \approx 0.056$). On the normalized feature set, Lambda is the only marker for which we

obtained a rank correlation below 0.9. [2]

We also tested the effects of leaving multiple markers out simultaneously. In the first test, we omitted six key markers CD10, CD19, CD20, CD79b, IgMi, and IgMs, following an experiment of Bendall et al. (2014). The found pseudotimes had good correlation ($r_S = 0.87$, $r_P = 0.92$ on the unnormalized features, $r_S = 0.88$, $r_P = 0.94$ on normalized features) to those found on the 18-feature set, while $r_P = 0.95$ was reported for WL. [3]

Since the four other samples A, B, C, and D do not contain three of the 18 features used in sample G (CD117, CD179b, and IgMi), our second test was to check the impact their removal. We found that the resulting pseudotimes had strong correlation to the original ($r_S, r_P \geq 0.97$ for both normalized and unnormalized features), indicating that the exclusion of these markers in the four other samples should have little effect.

[We briefly comment on the role of local minima in these experiments. In the leave-one-out tests for features with reported low correlations (IgMs, Kappa, Lambda for unnormalized, Lambda for normalized), we ran the algorithm with at least 5 initial seeds. One would expect that at least IgMs should not have such a large effect on the trajectory, since when removed together with five other markers the resulting pseudotimes still had good correlation to the original. And we did, for all three markers, obtain curves with better correlations and lower energy by initializing with the respective curve found on the (normalized or unnormalized) 18-feature set. The difficulty of the algorithm to locate these lower energy minima with automatic initialization suggests that the difference in the energy between the minima has decreased, and there should be a point when the lowest energy minima no longer correspond to curves that correlate well with those found on the 18-feature set. Indeed, using the original curve as initialization on the 'leave-6-out' test returned a curve with higher energy (and higher correlations) than that of what was earlier reported.]

3.5.2 Marker fit analysis

We additionally analyzed the role of the features by inspecting how well each was fit by the computed curve. To motivate this, note that the fidelity term of the functional decomposes into mean-squared errors in each dimension. For the discrete functional and a single curve $y = (y_1, ..., y_m) \in \mathbb{R}^{m \times d}$ we have

$$\frac{1}{n} \sum_{j=1}^{m} \sum_{i \in I_j} ||x_i - y_j||^2 = \sum_{p=1}^{d} MSE_p \qquad \text{where} \qquad MSE_p := \frac{1}{n} \sum_{j=1}^{m} \sum_{i \in I_j} (x_{i,p} - y_{j,p})^2.$$

We can then view the problem as finding a partial ordering of the data $\{I_1, I_2, ..., I_m\}$ and an associated curve y, where the ordering seeks to efficiently (and jointly) align the values in each dimension such that they can be approximated well by a curve without much length. Thus, if the feature variances are normalized so that they are not a factor in this tug-of-war, then one may measure the degree to which a given feature aligns with a penalized principal curve and therefore

[2] We note that these correlations differ from the results of Bendall et al. (2014) for the same test. There they only found one marker, HLA-DR, whose corresponding correlation ($r_P = 0.79$) upon removal was below $r_P = 0.93$. This difference is somewhat surprising, and could be due to their exclusive use of cosine distance in Wanderlust.

[3] It seems that in their experiment they used a significantly larger dataset that is not publicly available – it may be that this also applies to their 'leave-one-out' test.

also (to some extent) with the rest of the features based on its mean-squared error. Note also that for any minimizing curve y the mean-squared error in any dimension will be bounded by data's variance in that dimension: $MSE_p \leq \text{var}(x_{:,p})$ (a curve with constant value in that dimension equal to the mean of the data achieves this). When features have equal variance, we have that $MSE_p \leq \frac{1}{d}$ for all features p, which also gives us a uniform basis for comparison. The mean-squared errors might then be able to help identify which features are noisy to the developmental trajectory, and which play an important role.

We report the mean-squared errors MSE_p for each normalized feature in Table 1 column 6, and for convenience multiplied them by $d = 18$ so that they are on a $[0, 1]$ scale. We do not expect there to be a strong relation to the leave-one-out correlations (in the normalized setting), given our far-from-exhaustive search for global minimizers and high correlations for most features, but it is slightly encouraging that there is some positive correlation between the mean-squared errors and the Spearman and Pearson correlations reported there ($r_S = 0.34, 0.52$ respectively). We also note that there are no features with mean-squared errors ≥ 0.7, and this perhaps indicates that the features were well chosen. To investigate this, we did additional tests.

In the first, we compute a surrogate for the mean-squared error for the omitted features. For this we use (one-dimensional) trend filtering (Kim et al. (2009); Tibshirani (2014)), a nonparametric regression method, to fit curves over the same ordering for each feature separately. We use the code of Kim et al. (2009) [`https://web.stanford.edu/~boyd/l1_tf/`], and then compute the mean-squared errors using the trend filtering curves for each of the 41 features. These are reported in Table 1 column 7, and we note that on the 18 used features, while the values are on average slightly lower than the MSE_p values, they are representative ($r_S, r_P = 0.99$). Of the 23 omitted features, a few have notably low mean-squared errors: CD40 (0.25), CD22 (0.44), and TdT (0.46). That they align relatively well with the found developmental trajectory is consistent with their active role in B cell development, as investigated by Bendall et al. (2014). On the other hand, there are several features with high mean-squared errors – namely 15 of them are at least 0.9.

To check the degree to which their high values could be due to their omission from the feature set, we repeated the analysis after running our algorithm using all 41 normalized features. We display both sets of mean-squared errors with a scatterplot in Figure 4, where we marked the 18 initially included features in blue, and those initially omitted in red. Satisfyingly, the MSE's for all initially omitted features decreased while the opposite held for all those initially included. More remarkable is that the MSE's for most initially used features remained low (the highest of the six canonical markers is CD10 with MSE 0.63), and the MSE's for 13 of the features remained high (≥ 0.9). This broadly reinforces the selection of the 18 markers used by Bendall et al. (2014), with some markers believed to be relevant left out for testing or exploratory purposes (in particular CD40, CD22, TdT). On the other hand, that 13 of the markers maintained a high MSE indicates that their values have little impact on the found developmental trajectory. Only a few of these markers (cPARP, pPLCg, pSTAT5) were further investigated in Bendall et al. (2014), and from our understanding they did so primarily for data collected following cellular signaling perturbations (i.e. not the basal dataset that we analyze here) [4].

[4] After contacting a few of the authors of Bendall et al. (2014), we understood that spikes were present for cPARP

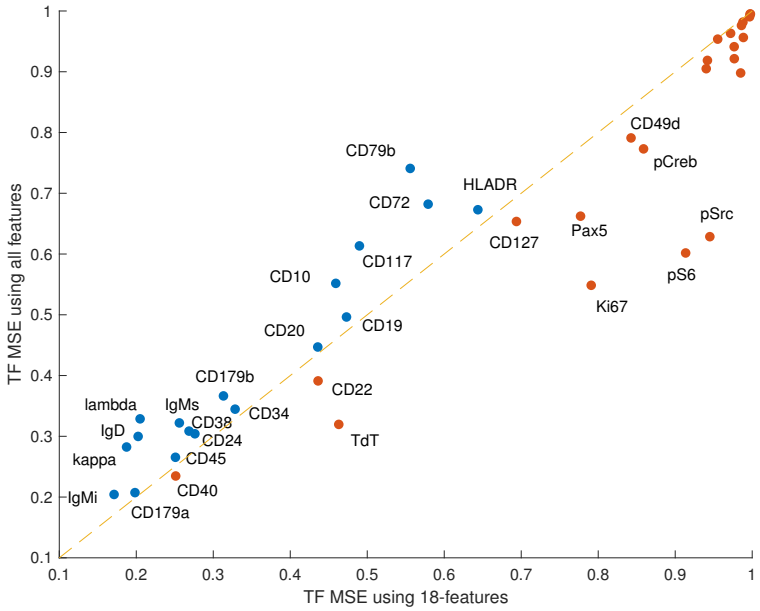


Figure 4: The mean-squared errors of each feature as computed by trend filtering using the ordering obtained with 18-features (x-axis), marked in blue, and with all features (y-axis). The dashed line corresponds to $y = x$.

3.6 Comparison across other samples

Here we describe the results obtained on the four other datasets A, B, C, and D. As stated earlier, these datasets do not contain three of the markers that we used with dataset G, so we applied the algorithm on the 15 common features, unnormalized, and we report the curves with lowest energy over 5 random seeds. As on set G, the pseudotimes we obtained for sets A, B, C, D had high correlations with those of WL, with $r_S = 0.88, 0.97, 0.91, 0.94$ ($r_P = 0.93, 0.97, 0.92, 0.97$) respectively. In addition, they agreed with prior knowledge on the rise and fall of canonical markers, which followed the same basic pattern across all datasets. They did however, notably differ from each other in the pseudotimes at which these events ocurred, and this discrepancy held for ranked pseudotimes as well (Figure 5). This presented us with interesting questions on how the developmental trajectories from different datasets should be compared.

One of our advantages is that we have curves lying in the same space that represent the pseudotimes, which allows for a number of direct comparisons. In the first, we investigated how consistently each of the five curves would order the cells from the different samples. For every curve-sample pair, we obtained induced pseudotimes (by projection), and computed correlations to the original pseudotimes for the sample. The resulting Spearman correlations are reported in Table 3. The sets B, C,

and pSTAT5 in some of the basal data they analyzed. Our PC and Wanderlust plots of the trajectory for cPARP in set G though appeared constant, while for pSTAT5 they did consistently feature a relatively small spike.

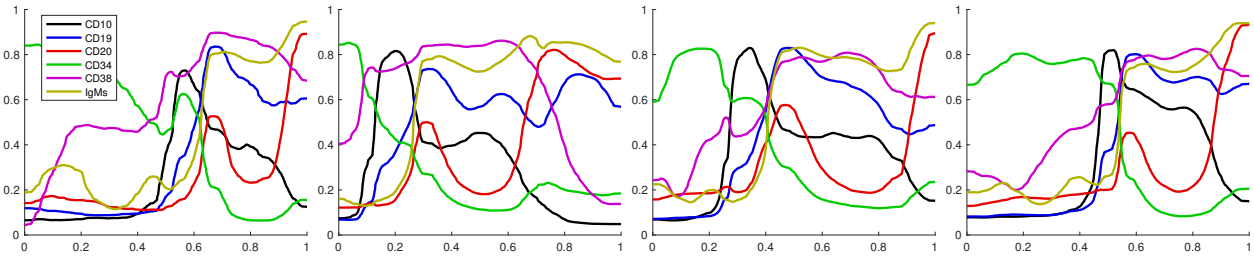


Figure 5: Normalized expression values of six canonical markers over the normalized (ranked) pseudotimes for the datasets A, B, C, D (left to right).

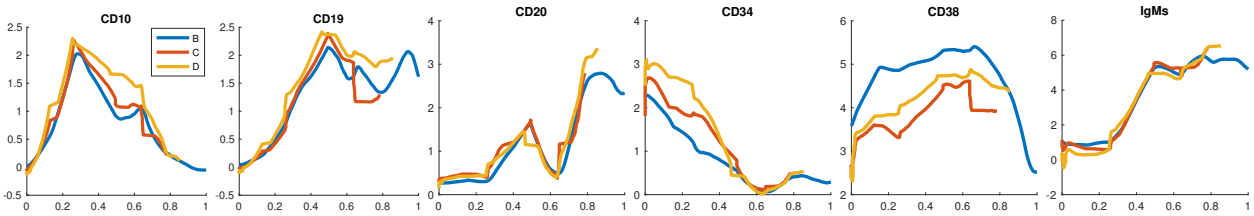


Figure 6: Expression values (unnormalized) of six canonical markers over reparametrized pseudotimes for the found curves from datasets B, C, and D.

and D appear to be represented well by all other curves, with the lowest correlation being $r_S = 0.86$ (for set C projected onto the curve from G). Notably set G gives most inconsistent pseudotimes when projected onto the other sets ($r_S = 0.43, 0.64, 0.34, 0.39$), and yet the curve from G appears to represent the other sets well ($r_S = 0.87, 0.93, 0.86, 0.94$).

The extent of this asymmetry may initially seem odd, and to further investigate we performed a similar test using just the curves. Since the curves are polygonal, we proceeded as before using their vertices and associated pseudotimes (or ranks) in place of the data samples. The resulting Spearman correlations between vertex ranks and their projected ranks (reported in Table 4) give a simpler representation of the earlier Table 3, and the correlations are expectedly higher (since curves are denoised representations of data). In particular, all correspondences between the curves B, C, and D are nearly perfect in this sense. This suggests that using the projected pseudotimes (after sorting as the correspondence is not perfect) for these datasets could allow their developmental trajectories to be more easily compared. In Figure 6, we plot curves B, C and D with respect to their projected (and sorted) pseudotimes onto B for (the usual) six canonical markers. We note that in some markers there are vertical discrepancies, as we did not normalize the expression values. There are also some undesirable properties of the reparametrizations: namely the projected pseudotimes from C and D do not map onto the entire curve B, and at places both extreme steepness and flatness appear for some markers (due to jumps and non-uniqueness of the projected pseudotimes). Aside from this, the trajectories do align fairly well horizontally, indicating some level of consistency in the developmental process across the datasets. More care would be needed to obtain smoother

parametrizations for this purpose, and especially in cases where the correspondence between curves is far from perfect (e.g. sets A and G). We end by remarking that the Fréchet distance between curves, and the parametrizations it is associated with, could be useful for further and more exact comparisons of the developmental trajectories.

4 Discussion and Conclusion

In this report, we applied a method for computing regularized principal curves to recover trajectories representing B cell development in human bone marrow samples. We obtained trajectories that agree with prior knowledge on marker behavior and closely resemble those of Wanderlust in Bendall et al. (2014). We then investigated the impact of the features by measuring how much the trajectory changes when they are excluded, and favorably found that nearly all have little impact in this sense.

Unlike most methods for trajectory inference, with our method we directly obtain a curve in the original feature space without first reducing the dimensionality of the data or altering its representation (e.g. via a graph). This allows for more direct analyses of the trajectories and their comparison across different datasets. In particular, we found that the trajectories of all five human samples exhibited the same pattern of canonical marker expression, and we were able to closely align three of the five that had very good correspondence via projection. Moreover, as our approach aims to find a curve that minimizes a combination of mean-squared error and length, comparing the mean-squared errors of features obtained after normalization could help identify important markers and those that are noisy. We found that markers known to be indicative of cell development stage had low mean-squared error, while a group of markers with high mean-squared error largely appear (in literature) less characteristic of development. A more thorough analysis would be needed to determine the strength of the implications, but our experiments suggest such comparison could have potential in identifying informative features.

There are of course drawbacks to our approach and some aspects that can be improved. One of the challenges is that we seek minimizers of a functional that is non-convex, and although our algorithm has mechanisms that help avoid high energy local minima, for the noisy datasets here it is usually necessary to check the curves over multiple initialization seeds. Another drawback of the functional is the possibility of kinks in penalized principal curves that do not get mitigated by finer discretization, leading to multiple cells being assigned the same pseudotimes. We did not find this to have a significant impact on our analysis of the data, though, and domain experts and practitioners should be able to better judge the importance of unique pseudotimes. Our approach is also restricted to using Euclidean distance (in the sense that the constructed curves approximate the data), which is in contrast to graph-based approaches that can rely on other notions of similarity. The effect of this limitation depends on the exact nature of the data, and the results suggest that it was not a major factor for the mass cytometry data here, but it could be a consideration for other types of data. In particular, it would be interesting to see if our approach could be extended to handle single-cell RNA-seq data, which has much higher dimensionality and a majority of missing values, as well as data representing more topologically complex (branching) processes.

5 Acknowledgements

I would like to thank Ziv Bar-Joseph and Dejan Slepčev for suggestions and valuable discussions. I am also grateful to Sean C. Bendall, El-ad David Amir, and Kara L. Davis for their help in answering questions regarding the dataset, and details on how the Wanderlust algorithm was tested. This research was supported by the National Science Foundation under grants CIF 1421502 and DMS 1516677.

References

Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725.

Bendall, S. C., Simonds, E. F., Qiu, P., El-ad, D. A., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., et al. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696.

Campbell, K., Ponting, C. P., and Webber, C. (2015). Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell rna-seq profiles. *bioRxiv*, page 027219.

Cannoodt, R., Saelens, W., and Saeys, Y. (2016a). Computational methods for trajectory inference from single-cell transcriptomics. *European journal of immunology*, 46(11):2496–2506.

Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B. N., De Preter, K., and Saeys, Y. (2016b). Scorpius improves trajectory inference and identifies novel modules in dendritic cell development. *bioRxiv*, page 079509.

Coico, R. and Sunshine, G. (2015). *Immunology: a short course*. John Wiley & Sons.

Gerber, S. and Whitaker, R. (2013). Regularization-free principal curve estimation. *The Journal of Machine Learning Research*, 14(1):1285–1302.

Guo, M., Bao, E. L., Wagner, M., Whitsett, J. A., and Xu, Y. (2017). Slice: determining cell differentiation and lineage based on single cell entropy. *Nucleic acids research*, 45(7):e54–e54.

Gupta, A. and Bar-Joseph, Z. (2008). Extracting dynamics from static cancer expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(2):172–182.

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.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84(406):502–516.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779.

Kegl, B., Krzyzak, A., Linder, T., and Zeger, K. (2000). Learning and design of principal curves. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(3):281–297.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM review*, 51(2):339–360.

Kirov, S. and Slepčev, D. (2017). Multiple penalized principal curves: Analysis and computation. *Journal of Mathematical Imaging and Vision*, 59(2):234–256.

Lu, X. Y. and Slepčev, D. (2015). Average-distance problem for parameterized curves. *ESAIM: COCV*.

Magwene, P. M., Lizardi, P., and Kim, J. (2003). Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics*, 19(7):842.

Marco, E., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., and Yuan, G.-C. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, 111(52):E5643–E5650.

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.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2018). A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv*, page 276907.

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.

Slepčev, D. (2014). Counterexample to regularity in average-distance problem. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 31(1):169 – 184.

Smola, A. J., Mika, S., Schölkopf, B., and Williamson, R. C. (2001). Regularized principal manifolds. *J. Mach. Learn. Res.*, 1(3):179–209.

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2017). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *bioRxiv*, page 128843.

Tibshirani, R. (1992). Principal curves revisited. *Stat. Comput.*, 2:182–190.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386.

Wickramasinghe, S., Porwit, A., and Erber, W. (2011). Normal bone marrow cells: development and cytology. In *Blood and Bone Marrow Pathology (Second Edition)*, pages 19–44. Elsevier.

A Tables

18-feature set	Leave-one-out analysis					Mean-squared errors			
	Unnormalized			Normalized		Using 18 features		Using all features	
	Var	r_S	r_P	r_S	r_P	MSE_p	using TF	using TF	MSE_p
CD10	0.01	1.00	1.00	0.99	1.00	0.55	0.46	0.55	0.63
CD117	0.01	1.00	1.00	0.99	0.99	0.60	0.49	0.61	0.72
CD179a	0.05	0.99	1.00	0.98	0.99	0.23	0.20	0.21	0.26
CD179b	0.03	1.00	1.00	0.99	0.99	0.36	0.31	0.37	0.42
CD19	0.03	0.99	0.99	0.98	0.99	0.50	0.47	0.50	0.53
CD20	0.04	0.99	0.99	0.98	0.99	0.46	0.44	0.45	0.47
CD24	0.09	0.97	0.98	0.97	0.98	0.29	0.28	0.30	0.35
CD34	0.04	0.99	0.99	0.98	0.98	0.34	0.33	0.34	0.35
CD38	0.09	0.97	0.98	0.98	0.99	0.28	0.27	0.31	0.32
CD45	0.04	0.99	1.00	0.99	0.99	0.26	0.25	0.26	0.27
CD72	0.02	1.00	1.00	0.98	0.99	0.63	0.58	0.68	0.69
CD79b	0.03	0.99	1.00	0.95	0.98	0.63	0.56	0.74	0.83
HLA-DR	0.04	0.99	0.99	0.98	0.99	0.69	0.64	0.67	0.71
IgD	0.05	0.99	0.99	0.93	0.97	0.25	0.20	0.30	0.31
IgMi	0.12	0.98	0.99	0.91	0.96	0.19	0.17	0.20	0.23
IgMs	0.13	**0.59**	**0.73**	0.97	0.99	0.28	0.26	0.32	0.36
Kappa	0.10	**0.37**	**0.31**	0.98	0.99	0.22	0.19	0.28	0.29
Lambda	0.06	0.94	0.93	**0.70**	**0.85**	0.23	0.21	0.33	0.38

Table 1: Markers included in the 18-feature set, along with values from leave-one-out and mean-squared-error analysis. From the leave-one-out analysis, values include variances (Var) of feature values, and Spearman (r_S) and Pearson (r_P) correlations in both normalized and unnormalized variance settings. The mean-squared errors include values obtained using trend filtering (TF) and the found curve (MSE_p) on both the 18-feature set and full feature set.

Other features	Mean-squared errors		
	Using 18 features	Using all features	
	using TF	using TF	MSE_p
CD3-1*	0.99	0.96	0.97
CD3-2*	0.98	0.94	0.96
CD3-3*	0.98	0.92	0.95
CD22	**0.44**	**0.39**	**0.42**
CD33*	0.94	0.90	0.92
CD40	**0.25**	**0.23**	**0.25**
CD49d	0.84	0.79	0.81
CD127	0.69	0.65	0.66
CD235*	0.99	0.97	0.98
Ki67	0.79	0.55	0.69
Pax5	0.78	0.66	0.72
RAG1*	0.97	0.96	0.96
TdT	**0.46**	**0.32**	**0.41**
cPARP*	1.00	0.99	0.99
pAKT*	0.99	0.98	0.98
pCreb	0.86	0.77	0.81
pErk12*	1.00	0.99	0.99
pP38*	0.96	0.95	0.95
pPLCg*	1.00	0.99	0.99
pS6	0.91	0.60	0.73
pSTAT5*	0.94	0.92	0.93
pSrc	0.95	0.63	0.77
pSyk*	0.99	0.90	0.94

Table 2: Mean-squared errors for features not included in the 18-feature set. Trend filtering estimates (TF) were used both when these features were omitted (column 1), and when all were included (column 2), for comparison. Features marked with an asterisk have MSE ≥ 0.9 and correspond to the unlabeled cluster in Figure 4.

	A	B	C	D	G
A	1.00	0.97	0.96	0.90	0.43
B	0.87	1.00	0.96	0.97	0.64
C	0.81	0.96	1.00	0.98	0.34
D	0.84	0.97	0.96	1.00	0.39
G	0.87	0.93	0.86	0.94	1.00

Table 3: r_S for induced pseudotimes obtained by projecting the datasets (columns) onto the different curves (rows).

	A	B	C	D	G
A	1.00	0.99	0.99	0.96	0.76
B	0.92	1.00	0.99	1.00	0.84
C	0.87	0.99	1.00	1.00	0.48
D	0.89	1.00	0.99	1.00	0.60
G	0.93	0.99	0.96	0.97	1.00

Table 4: r_S for induced pseudotimes obtained by projecting the curves (columns) onto different curves (rows).