# Quantifying single-cell differential abundance with density difference estimation on manifolds

Sunay Joshi

July 5, 2026

## 1 Introduction

Differential abundance, or the problem identifying differences in cellular abundance between two different single-cell RNA sequencing (scRNA-seq) datasets, is a fundamental problem in single-cell analysis [2]. A popular approach for solving this problem is to first estimate the distribution of each scRNA-seq dataset on the cellular manifold and then estimate the difference between these two distributions. This is the approach taken by the MELD algorithm in [1]. However, small errors in density estimation may potentially propagate to larger errors in estimating the density difference. Separately, recent work in the machine learning literature has found that directly estimating the density difference leads to higher accuracy. For instance, the LSDD algorithm introduced in [8] fits a sum of gaussians to the density difference. However, this line of work assumes that the data lies in a Euclidean space, which is often not true for scRNA-seq data. In this report, we present OurMethod, a density difference estimation method for single-cell manifolds. We compare OurMethod to existing approaches on simulated and real data.

## 2 Methods

### 2.1 Differential abundance problem

We begin by mathematically formulating the differential abundance problem. We are given two single-cell RNA sequencing (scRNA-seq) datasets $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{X}' = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_{N'}\}$ where the vectors $\mathbf{x}_i = [x_{i,g}]$ and $\mathbf{x}'_j = [x'_{j,g}]$ contain the expression of genes $g = 1, \ldots, G$ in cells $i = 1, \ldots, N$ and $j = 1, \ldots, N'$, respectively. We model each gene expression vector $\mathbf{x}_i, \mathbf{x}'_j$ as an independent and identically distributed (IID) sample from distributions $f_0, f_1$, respectively, where $f_0, f_1$ are supported on a lower dimensional submanifold $\mathcal{M} \subseteq \mathbb{R}^G$. This manifold assumption is common in single-cell analysis; for an overview, see [5].

Given the observed data $(\mathbf{X}, \mathbf{X}')$, our goal is to partition the gene expression manifold $\mathcal{M}$ into regions $\{t \in \mathbb{R}^G : f_1(t) > f_0(t)\}$ where the density $f_1$ is larger than the density $f_0$. In practice,

since the manifold $\mathcal{M}$ is unknown, we estimate the set of datapoints $\{i \in [N] : f_1(\mathbf{x}_i) > f_0(\mathbf{x}_i)\}$, $\{j \in [N'] : f_1(\mathbf{x}'_j) > f_0(\mathbf{x}'_j)\}$ where $f_1$ is greater than $f_0$. We formalize this as the following problem.

**Differential Abundance Problem (DAP).** *Given two datasets* $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ *and* $\mathbf{X}' = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_{N'}\}$, *where* $\mathbf{x}_i \overset{i.i.d.}{\sim} f_0$ *and* $\mathbf{x}'_j \overset{i.i.d.}{\sim} f_1$, *determine* $\{i : f_1(\mathbf{x}_i) > f_0(\mathbf{x}_i)\}$ *and* $\{j : f_1(\mathbf{x}'_j) > f_0(\mathbf{x}'_j)\}$.

A schematic depiction of the DAP is given in Figure 1. A number of approaches have been developed to solve the DAP over the past few years. One popular approach is the MELD algorithm developed by [1]. The MELD algorithm uses a manifold kernel density estimator (KDE) to estimate the distributions $\widehat{f_0}$ and $\widehat{f_1}$ separately, and then identifies the region $\hat{R} \subseteq \mathcal{M}$ in the manifold where $\widehat{f_1} > \widehat{f_0}$. Another popular approach is to identify cellular neighborhoods where a certain test statistic is greater than zero. The two-sample test provided in [4] rejects the null $H_0 : f_1 = f_0$ when a weighted integrated density difference is sufficiently large on some subset of $\mathcal{M}$, but it is not immediately clear how to convert this test to an estimation procedure.

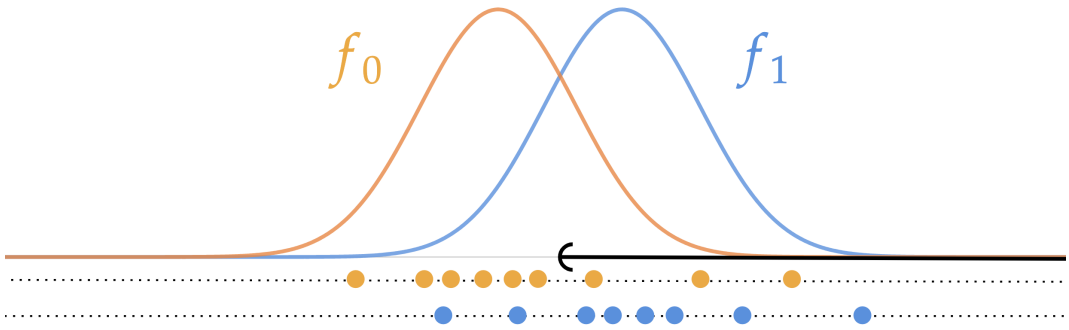



Figure 1: The DAP for a pair of densities on $\mathcal{M} = \mathbb{R}^1$. The set of points that fall on the right half-line form the desired region.

## 2.2 Density difference estimation

Recent work by [8] found that, for the Euclidean manifold $\mathcal{M} = \mathbb{R}^G$, the DAP can be solved more accurately by directly estimating the *density difference* $f_1 - f_0$. In particular, [8] developed the least squares density difference (LSDD) estimator.

The LSDD estimator $g(x)$ of the density difference $f_1 - f_0$ is given by a sum of Gaussians with fixed variance $\sigma^2 > 0$ centered on the datapoints:

$$g(\mathbf{x}) = \sum_{i=1}^{N+N'} \theta_i \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}_i\|^2}{2\sigma^2}\right) \tag{1}$$

where for notational ease, we let $\mathbf{y}_1, \ldots, \mathbf{y}_{N+N'}$ denote a given ordering of the datapoints in $\mathbf{X} \cup \mathbf{X}'$.

To define the LSDD estimator, [8] consider the regularized least squares problem

$$\theta^* = \text{argmin}_{\theta \in \mathbb{R}^{N+N'}} \int_{\mathbb{R}^G} (g(\mathbf{x}) - (f_1 - f_0)(\mathbf{x}))^2 d\mathbf{x} + \lambda \|\theta\|^2 \tag{2}$$

The solution $\theta^*$ is given by $\theta^* = (H + \lambda I)^{-1} h$, where $H$ is an $(N + N') \times (N + N')$ matrix with entries

$$H_{ij} = (\pi\sigma^2)^{G/2} \exp\left( - \frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\sigma^2} \right) \tag{3}$$

for $i, j \in [N + N']$, and $h \in \mathbb{R}^{N+N'}$ with entries

$$h_i = \left\langle \exp\left( - \frac{\|\mathbf{x} - \mathbf{y}_i\|^2}{2\sigma^2} \right), f_1(\mathbf{x}) \right\rangle_{L^2(\mathbb{R}^G)} - \left\langle \exp\left( - \frac{\|\mathbf{x} - \mathbf{y}_i\|^2}{2\sigma^2} \right), f_0(\mathbf{x}) \right\rangle_{L^2(\mathbb{R}^G)} \tag{4}$$

for $i \in [N + N']$. Since $f_0, f_1$ are unknown, $h$ cannot be computed from the data. A plug-in estimator $\hat{h} \in \mathbb{R}^{N+N'}$ of $h$ is given by

$$\hat{h}_i = \frac{1}{N'} \sum_{j=1}^{N'} \exp\left( - \frac{\|\mathbf{x}'_j - \mathbf{y}_i\|^2}{2\sigma^2} \right) - \frac{1}{N} \sum_{j=1}^{N} \exp\left( - \frac{\|\mathbf{x}_j - \mathbf{y}_i\|^2}{2\sigma^2} \right) \tag{5}$$

for $i \in [N + N']$, and the resulting approximation $\hat{\theta}$ to $\theta^*$ is given by $\hat{\theta} = (H + \lambda I)^{-1} \hat{h}$. Finally, the LSDD estimator is the function $g(\mathbf{x})$ determined by the coefficients $\theta^*$. The density difference estimator $g$ may be used to solve the DAP by returning the subset $\{i \in [N + N'] : g(\mathbf{y}_i) > 0\}$ of the dataset where $g$ is strictly positive. In their implementation of the LSDD algorithm, [8] use a 5-fold cross-validation procedure to select the variance parameter $\sigma$ and the regularization $\lambda$. In the Results section, when we run the LSDD algorithm, we will search over $\sigma \in \{0.25, 0.5, 0.75, 1, 1.2, 1.5, 2, 2.5, 2.2, 3, 5, 10, 15, 20, 25, 30\}$ and $\lambda \in \{10^{-3} \cdot 10^{4n/8} : 0 \leqslant n \leqslant 8\}$.

The main downside to the LSDD algorithm is that it only works for a Euclidean manifold $\mathcal{M} = \mathbb{R}^G$, while scRNA-seq data generally lies on a lower-dimensional non-Euclidean manifold. This motivates the following section, where we present OurMethod, a generalization to LSDD to non-Euclidean manifolds $\mathcal{M}$.

### 2.2.1 OurMethod

In OurMethod, we replace the gaussian functions in the LSDD estimator with so-called *weight vectors*. Each weight vector represents a neighborhood on the graph. In this way, the Hybrid method can be considered a synthesis of [4] and [8].

We begin by describing the construction of the weight vectors. We use the language of graph signal processing; for a reference, see [7]. Fix a positive integer $k$ (the kNN parameter), a positive integer $t$ (the scale parameter), and a regularization parameter $\lambda \geqslant 0$. Construct an unweighted $k$-nearest neighbors (kNN) graph on the complete dataset $\mathbf{X} \cup \mathbf{X}'$. Consider the heat filter $H : \mathbb{R}^{N+N'} \to \mathbb{R}^{N+N'}$ given in frequency space by the function $h(\lambda) = \exp(-\frac{t\lambda}{\lambda_{max}})$, where $\lambda_{max}$ denotes the largest eigenvalue of the Laplacian of the kNN graph. For each $i \in [N + N']$, apply the filter $H$ to the one-hot vector $e_i \in \mathbb{R}^{N+N'}$ to obtain the vector $w_i = He_i \in \mathbb{R}^{N+N'}$. We refer to $w_i$ as

the weight vector associated to the datapoint $\mathbf{y}_i$. The weight vector $w_i$ represents a neighborhood of $\mathbf{y}_i$ that becomes less concentrated as $t$ increases; thus $t$ controls the scale of each weight vector.

OurMethod constructs a density difference estimator using a linear combination of weight vectors:

$$g(\mathbf{x}) = \sum_{i=1}^{N+N'} \theta_i w_i(\mathbf{x}) \tag{6}$$

Since $w_i$ is only supported on the dataset $X \cup X'$, the resulting estimator $g$ is only supported on $X \cup X'$. We may therefore identify $w_i$, $g$, and $f_1 - f_0$ with vectors in $\mathbb{R}^{N+N'}$, having entries $w_i(\mathbf{y}_j)$, $g(\mathbf{y}_j)$, and $(f_1 - f_0)(\mathbf{y}_j)$ for all $j \in [N + N']$, respectively.

OurMethod considers the regularized least squares problem

$$\theta^* = \text{argmin}_{\theta \in \mathbb{R}^{N+N'}} \sum_{i=1}^{N+N'} (g(\mathbf{y}_i) - (f_1 - f_0)(\mathbf{y}_i))^2 + \lambda \|\theta\|^2 \tag{7}$$

Note that in vector notation, the objective can be rewritten as $\|G\theta - (f_1 - f_0)\|^2 + \lambda\|\theta\|^2$, where $G$ is the $(N + N') \times (N + N')$ matrix whose columns are $\{w_1, \ldots, w_{N+N'}\}$. Thus the infeasible solution is given by

$$\theta^* = (G^T G + \lambda I)^{-1} G^T (f_1 - f_0) \tag{8}$$

Note that $H = G^T G$ is an $(N+N') \times (N+N')$ matrix with entries $H_{ij} = \langle w_i, w_j \rangle$ for $i, j \in [N+N']$. To make this solution feasible, we use the plug-in estimator of $f_1 - f_0$ given by the empirical density difference

$$\left[ \frac{1}{N'} \mathbf{1}_{\mathbf{y}_i \in \mathbf{X}'} - \frac{1}{N} \mathbf{1}_{\mathbf{y}_i \in \mathbf{X}} \right]_{i \in [N+N']} \tag{9}$$

Therefore a plug-in estimator of $\theta^*$ is given by $\hat{\theta} = (H + \lambda I)^{-1} \hat{h}$, where $\hat{h} \in \mathbb{R}^{N+N'}$ has entries

$$\hat{h}_i = \frac{1}{N'} \sum_{j=1}^{N'} w_i(\mathbf{x}'_j) - \frac{1}{N} \sum_{j=1}^{N} w_i(\mathbf{x}_j) \tag{10}$$

Finally, OurMethod estimator is the function $g(\mathbf{x})$ determined by the coefficients $\theta^*$. The estimator $g$ may be used to solve the DAP by returning the subset $\{i \in [N + N'] : g(\mathbf{y}_i) > 0\}$ of the dataset where $g$ is strictly positive.

We now discuss parameter tuning. We assume that a good choice of $k$ is known, and we set $\lambda = 10^{-3}$. In order to select the scale parameter, we utilize the following 5-fold cross-validation procedure to select the optimal choice of $t \in \{5n : 0 \leqslant n \leqslant 10\}$. For each $i \in [5]$, randomly split $\mathbf{Y} = \mathbf{X} \cup \mathbf{X'}$ into a training set $\mathbf{Y}^{(i)}_{train}$ and a validation set $\mathbf{Y}^{(i)}_{valid}$. Fit OurMethod to the training set $\mathbf{Y}^{(i)}_{train}$. This results in a density difference estimator $g$ supported only on the training set $\mathbf{Y}^{(i)}_{train}$. In order to obtain an estimate on the validation set $\mathbf{Y}^{(i)}_{valid}$, we perform nearest-neighbor interpolation. Namely, for each datapoint $\mathbf{y}_j \in \mathbf{Y}^{(i)}_{valid}$ in the validation set, if $\mathbf{y}_\ell \in \mathbf{Y}^{(i)}_{train}$ denotes the closest training point to $\mathbf{y}_j$, then we set the value of our estimator $g(\mathbf{y}_j)$ to be $g(\mathbf{y}_\ell)$. This results in a density difference estimator $g$ supported on all of $\mathbf{Y}$.

It remains to evaluate the performance of $g$. We compare the estimate $g$ to a smoothed version of the empirical density difference. Specifically, for each datapoint $\mathbf{y}_j \in \mathbf{Y}^{(i)}_{valid}$ in the validation

set, we obtain its $k$-nearest neighbors $\mathbf{y}_{\ell_1}, \ldots, \mathbf{y}_{\ell_k}$ in the entire dataset $\mathbf{Y}$, and we compute the smoothed empirical density difference at $\mathbf{y}_j$ by the average

$$v_j = \frac{1}{k} \sum_{m=1}^{k} \left( \frac{1}{N'} \mathbf{1}_{\mathbf{y}_{\ell_m} \in \mathbf{X}'} - \frac{1}{N} \mathbf{1}_{\mathbf{y}_{\ell_m} \in \mathbf{X}} \right) \tag{11}$$

The performance of our estimator on the validation set is measured by the mean squared error

$$\frac{1}{|\mathbf{Y}^{(i)}_{valid}|} \sum_{\mathbf{y}_j \in \mathbf{Y}^{(i)}_{valid}} |g(\mathbf{y}_j) - v_j|^2 \tag{12}$$

and we select the parameter $t$ with the smallest average mean square error across all 5 folds. In practice, we observed that smoothing the empirical density difference in this manner significantly reduced the variability of the selected parameter $t$.

# 3 Results

We present the results of three simulations. In the context of the DAP, let $R = \{i : f_1(\mathbf{y}_i) > f_0(\mathbf{y}_i)\}$ denote the true region where $f_1$ dominates $f_0$, and let $\hat{R} \subseteq [N + N']$ denote an estimate of this region. We measure the performance of such an estimate $\hat{R}$ using the F-score, which is the harmonic mean of the precision and recall. In thie first two simulations, we compare the performance of the LSDD algorithm, OurMethod, and the MELD algorithm.

## 3.1 Toy example

We begin with a toy example in $\mathbb{R}^2$. For $t \in [-3, 3]$, consider the parametrized curve $\phi : \mathbb{R} \to \mathbb{R}^2$ having coordinates

$$x(t) = 2k + \frac{1}{\sin \alpha} \sin((t - 2k)\alpha) \tag{13}$$

$$y(t) = (-1)^{k+1} \frac{1}{\tan \alpha} + (-1)^k \frac{1}{\sin \alpha} \cos((t - 2k)\alpha) \tag{14}$$

where we use the shorthand $k = \lfloor \frac{t+1}{2} \rfloor$, and where we set $\alpha = \frac{7\pi}{9}$. The curve $\phi$, whose image $\mathcal{M}$ is a 1-dimensional submanifold of $\mathbb{R}^2$, is depicted in Figure 2. By construction, the intrinsic distance between two points on the manifold can be very different from the Euclidean distance between the points, particularly near the "bottleneck" at $t = \pm 1$.

We place a pair of piecewise constant densities $f_0, f_1$ on $\mathcal{M}$ as follows. Let $\tilde{f}_0$ denote the piecewise constant density $\tilde{f}_0(t) = \frac{1}{6}\mathbf{1}_{t \in [-1,0]} + \frac{3}{6}\mathbf{1}_{t \in [0,1]} + \frac{2}{6}\mathbf{1}_{t \in [1,2]}$, and let $f_0$ denote the pushforward of $\tilde{f}_0$ by $\phi$. Similarly, let $\tilde{f}_1$ denote the piecewise constant density $\tilde{f}_1(t) = \frac{1}{4}\mathbf{1}_{t \in [-2,2]}$, and let $f_1$ denote the pushforward of $\tilde{f}_1$ by $\phi$. The desired region $R$ is the set of all datapoints that land along the segment $\phi([-2, 0])$. The densities $\tilde{f}_1, \tilde{f}_0$ are depicted in Figure 3.

Our dataset is sampled as follows. Let $n = 300$ and let $p = 0.5$. For each $i \in [n]$, with probability $p$ sample datapoint $\mathbf{y}_i$ from distribution $f_1$, and with probability $1 - p$ sample $\mathbf{y}_i$ from

distribution $f_0$. Let $\mathbf{X}$ denote the set of all points $\mathbf{y}_i$ sampled from $f_0$, and let $\mathbf{X}'$ denote the set of all points $\mathbf{y}_i$ sampled from $f_1$. We seek to solve the DAP for the pair $(\mathbf{X}, \mathbf{X}')$.

We compare the performance of the LSDD algorithm, OurMethod, and the MELD algorithm. We compute the F-score for each of 100 independent datasets $(\mathbf{X}, \mathbf{X}')$, and we report the average F-score plus or minus one standard deviation. For the MELD algorithm, we set the $\beta$ parameter to $\beta = 67$ as in the MELD tutorial, and we set the kNN parameter to $k = 10$ based on a gridsearch over $k \in \{2m : 1 \leqslant m \leqslant 20\}$. Since this is a toy example, we simply set $k = 20$ in OurMethod; we do not attempt to tune this value. The average F-scores are: $0.812 \pm 0.093$ for LSDD, $0.826 \pm 0.105$ for OurMethod, and $0.925 \pm 0.073$ for MELD. In this case, OurMethod outperforms the LSDD algorithm, but underperforms the MELD algorithm. However, it is important to note that we have not tuned the parameter $k$ in OurMethod, so we cannot expect optimal results.

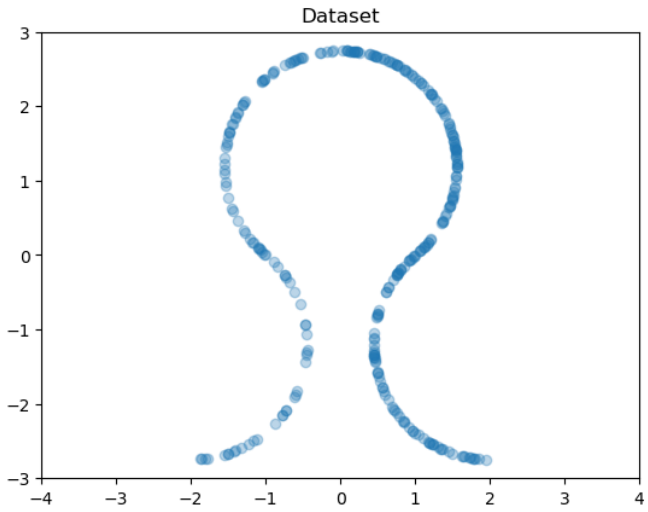



Figure 2: A dataset sampled from the densities $f_0, f_1$ supported on the image of the curve $\phi$.

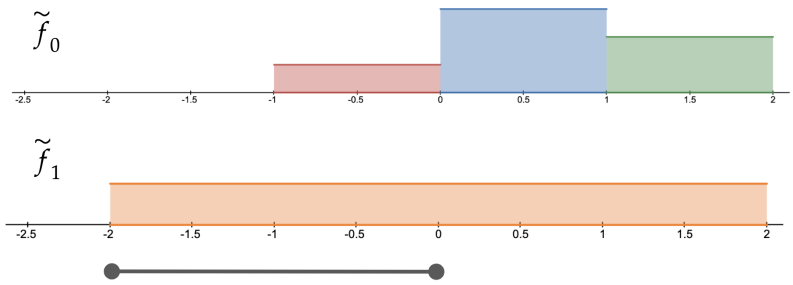



Figure 3: The densities $\tilde{f}_0, \tilde{f}_1$. The region $\{\tilde{f}_1 > \tilde{f}_0\}$ is underlined.

## 3.2 Semi-synthetic data

We now construct a semi-synthetic example. We use the zebrafish scRNA-seq dataset provided in [1]. The dataset consists of the gene expression vectors of 26785 cells projected onto the space $\mathbb{R}^{100}$ spanned by the top 100 principal components. In this experiment, the control group was injected with tyrosinase, and the treatment group was injected with chordin.

In this example, we construct two artificial discrete distributions $f_0, f_1$ on the dataset and sample datasets $\mathbf{X}, \mathbf{X}'$ from $f_0, f_1$ respectively. We construct the distributions $f_0, f_1$ as follows. Perform Leiden clustering on the entire dataset; for details on Leiden, see [9]. This yields a set of 23 distinct clusters, numbered $\{0, 1, \ldots, 21\}$ from largest to smallest. The largest cluster 0 contains 3362 cells. The top 11 clusters $\{0, 1, \ldots, 10\}$ contain 18913 cells, or 70% of the dataset. Let $f_0$ denote the uniform distribution on the set of the top 11 clusters $\{0, 1, \ldots, 10\}$. Let $f_1$ denote the piecewise constant distribution that has 50% of its mass distributed uniformly on cluster 0, and the remaining 50% of its mass distributed uniformly on clusters $\{1, \ldots, 10\}$. The desired region $R$ is simply the set of cells that belong to cluster 0. We include a schematic depiction of the distribution $f_1$ in Figure 4.

Our dataset is sampled as follows. Let $n = 2000$ and let $p = 0.5$. For each $i \in [n]$, with probability $p$ sample datapoint $\tilde{\mathbf{y}}_i$ from distribution $f_1$, and with probability $1 - p$ sample $\tilde{\mathbf{y}}_i$ from distribution $f_0$. Next, independently sample a gaussian random vector $\mathbf{u}_i \sim N(0, 0.1^2 \cdot I_{100})$ and add it to the datapoint to obtain $\mathbf{y}_i = \tilde{\mathbf{y}}_i + \mathbf{u}_i$. (We add a small amount of gaussian noise to avoid sampling the exact same datapoint twice.) Let $\mathbf{X}$ denote the set of all points $\mathbf{y}_i$ sampled from $f_0$, and let $\mathbf{X}'$ denote the set of all points $\mathbf{y}_i$ sampled from $f_1$. We seek to solve the DAP for the pair $(\mathbf{X}, \mathbf{X}')$.

We compare the performance of the LSDD algorithm, OurMethod, and the MELD algorithm. We compute the F-score for each of 100 independent datasets $(\mathbf{X}, \mathbf{X}')$, and we report the average F-score plus or minus one standard deviation. For the MELD algorithm, we set the $\beta$ parameter to $\beta = 67$ as in the MELD tutorial, and we set the kNN parameter to $k = 2$ based on a gridsearch over $k \in \{2m : 1 \leqslant m \leqslant 20\}$. We set $k = 20$ in OurMethod, as suggested in the Methods section of [1]. The average F-scores are: $0.520 \pm 0.136$ for LSDD, $0.920 \pm 0.018$ for OurMethod, and $0.878 \pm 0.021$ for MELD. In this case, OurMethod outperforms the MELD algorithm. Also, it appears that the LSDD algorithm significantly underperforms in part due to high variability in its selected variance parameter $\sigma$.

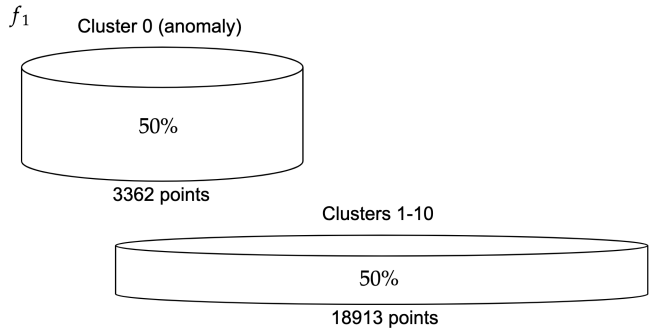



Figure 4: A schematic depiction of the distribution $f_1$ supported on the top 11 Leiden clusters.

## 3.3 Real data

For our final example, we run OurMethod on the entire zebrafish dataset provided in [1]. Figure 5 provides a visualization of the various cell types in the dataset. We let $\mathbf{X}$ consist of the cells in the control group, we let $\mathbf{X}'$ consist of the cells in the treatment group, and we run OurMethod on

$\mathbf{X} \cup \mathbf{X}'$. In other words, unlike in the previous section, we treat the zebrafish dataset as having already been sampled from unknown distributions $f_0, f_1$ on some manifold $\mathcal{M}$. Note that for the real dataset, there is no ground truth answer for the region $R$. As in the semi-synthetic example, we set the kNN parameter to $k = 20$. Further, in order to obtain a deterministic result, we set the scale parameter to $t = 45$, which appears to be the typical value of $t$ selected in semi-synthetic cross-validation.

The results are depicted in Figure 6, using the PHATE visualization algorithm of [6]. OurMethod predicts a region $\hat{R}$ containing 14338 cells, which constitutes 53.5% of the dataset $\mathbf{X} \cup \mathbf{X}'$. The top cell type in the region $\hat{R}$ by proportion is the hatching gland; 87% of its cells are present in $\hat{R}$. The top cell type in $\hat{R}$ by cell count is the tailbud, presomitic mesoderm; 2929 of its cells are present in $\hat{R}$. However, these results should be taken with a grain of salt. More testing of OurMethod is required before drawing biological insights from the predictions.

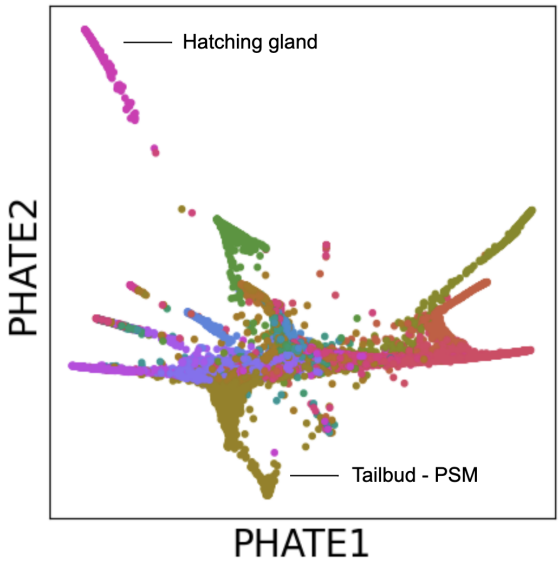



Figure 5: Cell states in the zebrafish dataset. Each color corresponds to a unique cell type. The dataset is visualized in $\mathbb{R}^2$ using the PHATE algorithm. The hatching gland and tailbud, presomitic mesoderm regions are labeled.

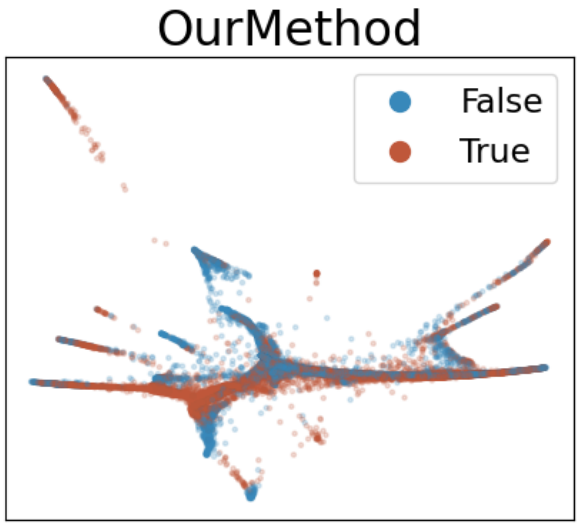



Figure 6: Results of OurMethod in PHATE space. Red points denote those for which the estimated density difference $f_1 - f_0$ is positive. The red cells in the top left of the image belong to the hatching gland.

# 4 Future work

In this report, we have presented OurMethod, a generalization of the LSDD algorithm to densities supported on manifolds. As shown in the simulations above, OurMethod outperforms the LSDD algorithm, especially in high-dimensional settings. However, OurMethod does not in general outperform the MELD algorithm presented in [1]. There are a few directions for improvement. First, one can try running OurMethod with an alternate set of weight vectors. Indeed, one can use any graph filter $H$ to define the weight vectors, and it is possible that a different set of weight vectors would improve the performance of OurMethod. Another direction for improvement is to refine the cross-validation procedure for the scale parameter $t$. At present the smoothing procedure is motivated by empirical results, but is not theoretically justified. Parameter selection for the closely related density estimation problem is typically based on an analysis of the integrated mean squared error of the estimator [3]. It would be interesting to derive equivalent criteria for density difference estimation on manifolds.

# Acknowledgements

I would like to thank my advisor, Professor Ben Raphael (COS), for his guidance throughout this project. I would also like to thank my mentors Uthsav Chitra and Julian Gold for collaborating with me and helping me prepare this report. Finally, I would like to thank Professor Jianqing Fan (ORFE) for kindly agreeing to be my PACM Reader.

# References


[1] D. B. Burkhardt, J. S. Stanley III, A. Tong, A. L. Perdigoto, S. A. Gigante, K. C. Herold, G. Wolf, A. J. Giraldez, D. van Dijk, and S. Krishnaswamy. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature biotechnology*, 39(5):619–629, 2021.

[2] S. Das, A. Rai, and S. N. Rai. Differential expression analysis of single-cell rna-seq data: current statistical approaches and outstanding challenges. *Entropy*, 24(7):995, 2022.

[3] M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American statistical association*, 91(433):401–407, 1996.

[4] B. Landa, R. Qu, J. Chang, and Y. Kluger. Local two-sample testing over graphs and point-clouds by random-walk distributions. *arXiv preprint arXiv:2011.03418*, 2020.

[5] K. R. Moon, J. S. Stanley III, D. Burkhardt, D. van Dijk, G. Wolf, and S. Krishnaswamy. Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 7:36–46, 2018.

[6] K. R. Moon, D. van Dijk, Z. Wang, W. Chen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy. Phate: a dimensionality reduction method for visualizing trajectory structures in high-dimensional biological data. *BioRxiv*, 120378, 2017.

[7] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

[8] M. Sugiyama, T. Kanamori, T. Suzuki, M. C. Du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. *Neural Computation*, 25(10):2734–2775, 2013.

[9] V. A. Traag, L. Waltman, and N. J. Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.