Multimodal data visualization, denoising and clustering with integrated diffusion

Manik Kuchroo [*1] **Abhinav Godavarthi** [*2] **Guy Wolf** [†3 4] **Smita Krishnaswamy** [†5 6]

Abstract

We propose a method called integrated diffusion for combining multimodal datasets, or data gathered via several different measurements on the same system, to create a joint data diffusion operator. As real world data suffers from both local and global noise, we introduce mechanisms to optimally calculate a diffusion operator that reflects the combined information from both modalities. We show the utility of this joint operator in data denoising, visualization and clustering, performing better than other methods when applied to multi-omic data generated from peripheral blood mononuclear cells. Our approach better visualizes the geometry of the joint data, captures known cross-modality associations and identifies known cellular populations. More generally, integrated diffusion is broadly applicable to multimodal datasets generated in many medical and biological systems.

1. Introduction

Recently there has been a profusion of multimodal data measured in parallel on the same system. Some examples include multiple modalities of data collected on biological specimens, such a single cell RNA-sequencing or single cell ATAC-sequencing, or multiple measurements collected on hospitalized patients, such as lab tests and continuous monitoring systems. There is a dire need for integration of this data in order to perform a wide variety of downstream tasks such as clustering, differential or comparative analysis, denoising and cross-modality correlations between features. We believe that the key to integrating data is to discover which entities are similar to each other across modalities by creating a data affinity graph on the basis of information from all modalities available. However, it is not immediately clear how to determine distances or similarities between entities on the basis of multiple modalities of data, which could be measured on entirely differently scales and suffer from different amounts of noise and sparsity. This is particularly problematic in the biomedical domain, where issues of 'drop out' or under-sampling make correlation analysis in single cell technologies extremely difficult. In order to address this, we turn to the manifold learning framework of data diffusion that was developed by Coifman & Lafon (2006).

Although measurement strategies create high dimensional observations, the intrinsic dimensionality, or degrees of freedom within the data, is relatively low. This manifold assumption is at the core of the data diffusion framework, which learns the intrinsic manifold space of the data by powering a Markov transition matrix to a power t, implicitly calculating a t-step random walk on the data graph. This process accumulates probabilities in paths that traverse through relatively dense regions of the data and diminish in sparse outlier regions. In Coifman & Lafon (2006), the powered diffusion operator is eigendecomposed to uncover intrinsic data dimensions called a diffusion map however since that seminal work, data diffusion has been shown to be useful in a myriad of data processing tasks (Moon et al., 2018), including clustering (Burkhardt et al., 2020), denoising (Van Dijk et al., 2018) and dimensionality reduction (Moon et al., 2019). Recently, a multimodal diffusion approach named alternating diffusion (Katz et al., 2019) generalized the random walk to "hop" between different metric spaces by taking a matrix product of the markov transition matrices. As explained by Katz et al. (2019), the diffusion distances in this joint space constitute the joint diffusion map embedding, which captures information shared between modalities but removes modality-specific information. While this approach creates a joint manifold, it does not generalize well to noisey biological datasets which contain modality specific sources of noise.

Here, we define an *integrated diffusion operator* for multiple

[*]Equal contribution [1]Department of Neuroscience, Yale University, New Haven, CT [2]Department of Applied Mathematics, Yale University, New Haven, CT, USA [3]Department of Mathematics & Statistics, Université de Montréal, Montréal, QC, Canada [4]Mila – Quebec AI Institute, Montréal, QC, Canada [5]Department of Genetics, Yale University, New Haven, CT, USA [6]Department of Computer Science, Yale University, New Haven, CT, USA. Correspondence to: Smita Krishnaswamy <smita.krishnaswamy@yale.edu>.

The 2021 ICML Workshop on Computational Biology.

data modalities which accounts for local noise and intrinsic dimensionality of each modality. Conceptually diffusion probabilities in our integrated operator are computed by taking several steps in the data graph from one modality, and several steps on the data graph defined by the other modality. The number of steps is carefully chosen based on the *spectral entropy* of each operator. Furthermore, we emphasize dominant directions in the diffusion operator by locally denoising using PCA-based low-rank approximations.

2. Method

2.1. Problem Formulation

Let $\mathbf{X} \subseteq \mathbb{R}^{D_X}$ and $\mathbf{Y} \subseteq \mathbb{R}^{D_Y}$ be two sets of data, perhaps with different dimensionalities, capturing two modalities gathered (e.g., via different measurement techniques) from the same underlying system. We consider a setting where the underlying system of the data can be modeled via a d-dimensional manifold (with $d \ll \min\{D_X, D_Y\}$) that is embedded in a high dimensional ambient space given by both modalities, but is only partially captured by each individual dataset. Here, we describe an unsupervised approach to integrate information from such multimodal settings based on the principles of data diffusion in order to recover the underlying joint manifold. By utilizing methods that capture both local and global manifold geometric information, our method is robust to vastly differing quantities of noise. Our method allows for amenable visualization, data denoising, and clustering of this jointly recovered manifold.

Neighborhood low rank approximation for local noise correction We begin by estimating a measure of local signal in various neighborhoods in the dataset. To do this, we first run spectral clustering on each modality to obtain N partitions $X_1, \ldots, X_N$, written as submatrices of $\mathbf{X}$ (and accordingly for $\mathbf{Y}$). Next, we compute SVD on the centered points in the partition $X_i - \bar{X}_i = USV^T$ where $\bar{X}_i$ is a matrix with all rows containing the partition center, U consist of left singular vectors, V right singular vectors and S contains singular values. In order to denoise a neighborhood, we estimate the intrinsic dimensionality using an eigengap heuristic, counting the first *k+1* singular values. Finally, we obtain a low rank approximation of the data in each local partition by using a truncated SVD, i.e., $\widetilde{X}_i = U'S'(V')^T + \bar{X}_i$ where S' only takes the first (most significant) $k+1$ singular values, and vectors U', V' consist of the first $k+1$ columns of U, V (correspondingly). It is important that this method be highly local so as not to destroy the manifold structure via elimination of linear dimensions in the data.

Modality specific diffusion time scale calculation via spectral entropy In addition to correcting for varying local noise within a single modality, it is crucial to estimate the intrinsic dimensionality of each modality to understand how much information each contains. Previous implementations of data diffusion methods, such as alternating diffusion and diffusion maps, provide no means of calculating a correct timescale. Here, we apply *spectral entropy*, computed on the diffusion operator, to estimate ideal number of t-steps to take in each modality. This refers to the theory of graph signal processing (Shuman et al., 2013) where the eigenvectors of the diffusion operator form frequency harmonics on a data graph. The spectral entropy of the operator is then the amount of variability explained by each frequency in the graph spectrum, i.e., the diffusion dimension.

To quantify the significance of each diffusion dimension in describing the data geometry, we can observe the corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. Quantitatively, this is given by the spectral entropy defined on probability distribution of eigenvalues normalized by their sum, $\eta(t)$. This is parameterized by the diffusion timescale t as this spectrum changes with the powering of the diffusion operator $\lambda_1^t, \lambda_2^t, \ldots, \lambda_k^t$.

$$H(t) = -\sum_{i=1}^{N} \eta(t)_i \log[\eta(t)]_i, \tag{1}$$

When the diffusion operator is powered to a value t, there is an application of a low-pass filter to the eigenspectrum of the operator such that the eigenvalues corresponding to higher frequencies are diminished. Thus the spectral entropy decreases with subsequent powering of the operator t—but not steadily. For low values of t the spectral entropy rapidly decreases and then stabilizes to create an elbow. We believe this elbow refers to the elimination of noise, with further powering removing signal. We find the elbow of this operator for the modality-specific operators. In this manner, the higher frequency components of the data graph, corresponding to noise dimensions will be eliminated in a frequency-specific manner globally on the graph, as opposed to locally in a vertex-specific manner using local PCA. We note that a similar heuristic is used in Moon et al. (2019) where any value beyond an elbow is chosen for visualization using PHATE.

Fusion of operators We compute t using the spectral entropy heuristic for each modality taken independently, giving us an estimate for the relative quantities of information present between modalities. While the absolute degree of information within each view is informative, a ratio of information is perhaps more meaningful. We raise each modalities diffusion operator to the lowest possible multiple of the ideal view specific t computed via spectral entropy. For example, if we obtain time values of 2 and 8 for two individual modes, then we will assume a ratio of 1:4 of information. Intuitively, this ratio indicates that for every diffusion "step" taken in modality 1, four diffusion steps will be taken in modality 2. More generally, we can write our joint diffusion operator, $\mathbf{J}$, to reflect the differing levels

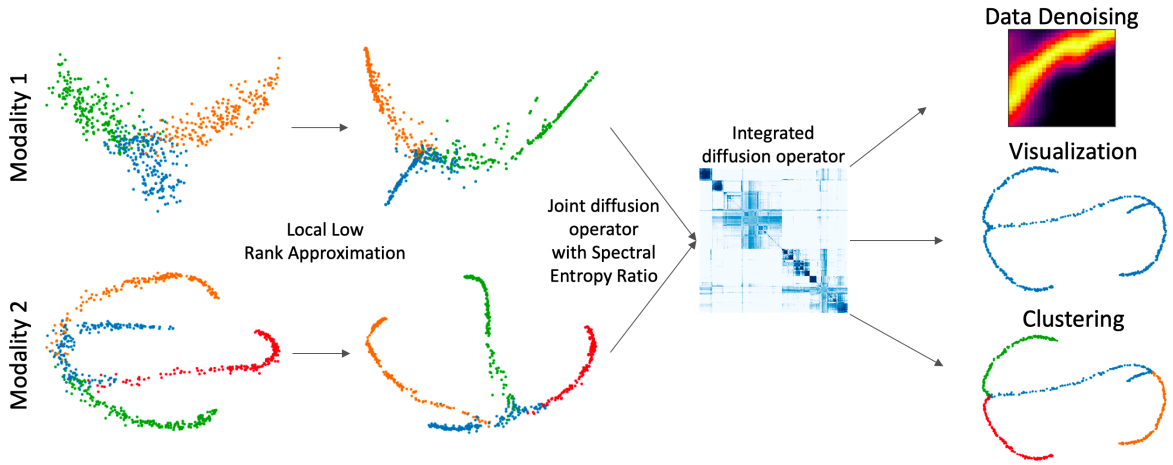

Figure 1. Overall workflow of integrated diffusion. First, local low rank approximation is performed on each dataset independently, removing local noise. Next, the intrinsic dimensionality of each modality is calculated by the spectral entropy of each operator to determine the ideal number of t-steps to place in each modality. Each operator is then powered to a ratio of the calculated t-steps and finally multiplied together to simulate steps in a random walk. The resulting diffusion operator can help denoise, visualize and cluster.

of global information between views as follows:

$$\mathbf{J} = \mathbf{P_1^{t_1}} * \mathbf{P_2^{t_2}}, \tag{2}$$

where $\mathbf{t_1}$ and $\mathbf{t_2}$ are integer values obtained from the reduced ratio as described above, and $\mathbf{P_1}$ and $\mathbf{P_2}$ are modality specific diffusion operators. We use the reduced ratio instead of directly applying the values of t obtained from the spectral entropy heuristic, as this joint operator is then powered once more to correct for spurious noise generated when integrating the datasets (i.e., noise present from one measurement modality affecting signal present from another measurement modality). Powering directly by $\mathbf{t_1}$ and $\mathbf{t_2}$ would lead to an oversmoothing effect in the final computed manifold which would collapse independent clusters together. We determine the adequate timescale for powering this joint diffusion via the same spectral entropy approach and calculate an embedding using the method of Moon et al. (2019).

3. Biological Applications

New methods allow for the measurement tens to hundreds of thousands of features in single cells, allowing for unprecedented insight into biological and cell type specific processes. Until recently, only a single modality could be measured in each cell, be it expression of genes through RNA sequencing or the accessibility of chromatin regions through ATAC sequencing. Now novel techniques allow for the measurement of different modalities at single-cell resolution. Increasingly commonly, individual cells are measured with a combination of chromatin accessibility, RNA expression, protein expression and spatial location (Ma et al., 2020; Cao et al., 2018; Liu et al., 2020). This new type of data is powerful, as it not only allows for the study of each modality independently, but also allows for the discovery of regulatory mechanisms between modalities. Currently, no computational techniques are capable of modelling and predicting these dynamics as there are no strategies that integrate different modalities of data to jointly visualize, cluster and denoise multimodal single-cell data.

We apply integrated diffusion to multimodal single cell data of 11,909 blood cells, visualizing the joint manifold, identifying known cell types and uncovering key cross modalities interactions. Visualizing each modality, gene expression and chromatin accessibility, independently reveals similar overall structure however different resolutions. Chromatin accessibility data, when compared to gene expression data, is incredibly sparse and generally considered to be far less informative. When computing the spectral entropy of each modality, we can clearly see that the chromatin accessibility diffusion operator has a far fewer informative dimensions than the gene expression operator. The alternating diffusion approach, which does not take into account the information present within each modality, creates an embedding that blends the distinct structure of gene expression data with the less informative structure of chromatin accessibility data. Integrated diffusion, however, appears to better resolve differences in information across dataset, producing a visualization that contains sharper borders between populations and displays clear structure when visualized with PHATE (Figure 2A).

These more clearly resolved populations also correspond with more biologically relevant clusters. Using cellular annotations of this dataset which predict celltype based on the expression of known marker genes and accessibility regions

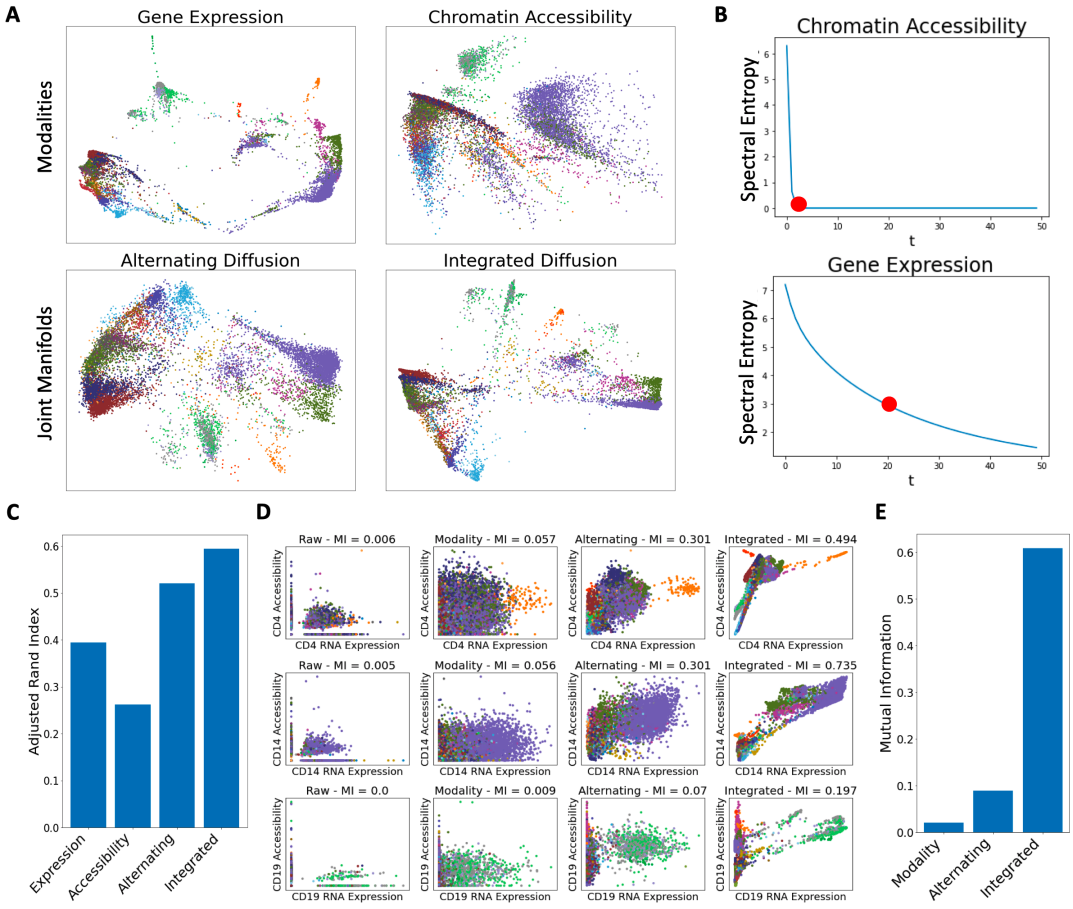

Figure 2. Application of integrated diffusion to multimodal single cell data. A) Visualization of gene expression and chromatin accessibility manifolds as well as alternating and integrated diffusion operators via PHATE. Points colored by annotated cell type. B) Visualization of spectral entropy of each modality. C) Adjusted Rand Index between spectrally computed clusters on each diffusion operator and known annotated cell types. D) Mutual information between the expression of a gene and it's accessibility across differing denoising strategies: modality specific denoising, denoising both modalities with alternating diffusion operator or integrated diffusion operator. E) Average mutual information for differing denoising strategies across all gene expression-gene accessibility pairs.

(Hao et al., 2020), we computed clusters from the diffusion operator of each modality as well as alternating and integrated operators. Clusters from the integrated operator best overlapped with annotated cell types (Figure 2C).

A major issue in single cell data is sparsity due to under sampling which makes it very difficult to measure and model cross modality interactions. Theoretically, if a gene is expressed, then the chromatin encoding that gene must be accessible. With this understanding of the data, we try to recover these known associations between gene expression and chromatin accessibility (Figure 2D). Due to sparsity, there is no association as computed by mutual information between these variables without denoising. There are several strategies to recovering these cross modality interactions: denoising with modality specific diffusion operators, denoising with a single alternating diffusion operator or denoising with a single integrated diffusion operator. Using the integrated diffusion operator appears to best recover known gene expression and chromatin accessibility associations as shown in genes CD19, CD14 and CD4 (Figure 2D). We then computed these associations across all genes with each of our denoising strategies. Across 18,659 genes, integrated diffusion recovered significantly more information between a gene's accessibility and its expression than alternating diffusion and modality-specific diffusion (Figure 2E).

4. Conclusion

We introduce the integrated diffusion operator, a method for learning the joint data geometry as described by multiple data measurement modalities applied to a single system. We show its improvement over alternating diffusion in integrating multimodal information. We apply our method in the biomedical setting to a multi-omics dataset, where we generated rich joint manifolds, compute cell popula-

tions with increased accuracy and recover cross modality gene-chromatin associations. Our flexible framework is extendable to multiple modalities and will allow for the successful integration and analysis of massive multi-omic datasets from a wide variety of fields. Future work will involve multiscale diffusion operators designed to integrate data at many levels of granularity.

References

Burkhardt, D. B., Stanley, J. S., Tong, A., Perdigoto, A. L., Gigante, S. A., Herold, K. C., Wolf, G., Giraldez, A. J., van Dijk, D., and Krishnaswamy, S. Quantifying the effect of experimental perturbations in single-cell rna-sequencing data using graph signal processing. *bioRxiv*, 2020. doi: 10.1101/532846. URL https://www.biorxiv.org/content/early/2020/08/01/532846.

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., and Shendure, J. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, August 2018. doi: 10.1126/science.aau0730. URL https://doi.org/10.1126/science.aau0730.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. doi: 10.1016/j.acha.2006.04.006.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zagar, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. B., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. Integrated analysis of multimodal single-cell data. *bioRxiv*, 2020. doi: 10.1101/2020.10.12.335331. URL https://www.biorxiv.org/content/early/2020/10/12/2020.10.12.335331.

Katz, O., Talmon, R., Lo, Y.-L., and Wu, H.-T. Alternating diffusion maps for multimodal data fusion. *Information Fusion*, 45:346–360, January 2019. doi: 10.1016/j.inffus.2018.01.007. URL https://doi.org/10.1016/j.inffus.2018.01.007.

Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C. C., Tebaldi, T., Zhang, D., Kim, D., Bai, Z., Norris, E., Pan, A., Li, J., Xiao, Y., Halene, S., and Fan, R. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell*, 183(6):1665–1681.e18, December 2020. doi: 10.1016/j.cell.2020.10.026. URL https://doi.org/10.1016/j.cell.2020.10.026.

Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., Law, T., Lareau, C., Hsu, Y.-C., Regev, A., and Buenrostro, J. D. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, 183(4):1103–1116.e20, November 2020. doi: 10.1016/j.cell.2020.09.056. URL https://doi.org/10.1016/j.cell.2020.09.056.

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

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. v. d., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., and Krishnaswamy, S. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37 (12):1482–1492, 2019.

Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., and Vandergheynst, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3): 716–729, 2018.