Representation learning on gene expression data

CSIC 5011 Final project

Xinwei Shen and Yunfei Yang

The Hong Kong University of Science and Technology

May 19, 2019

Representation learning

- Data representation plays an essential role in machine learning.
- High-dimensional data often concentrate around a low dimensional manifold.
- Data representation methods can help us reduce the dimension and extract useful information when building classifiers or other predictors.

Methodology

Three levels of representation learning methods:

- linear dimension reduction methods: PCA and robust PCA
- Non-linear dimensionality techniques: manifold learning
  - Isomap, Locally linear embedding (LLE) and t-distributed stochastic neighbor embedding (tSNE)
- Deep representation methods: variational autoencoder (VAE)

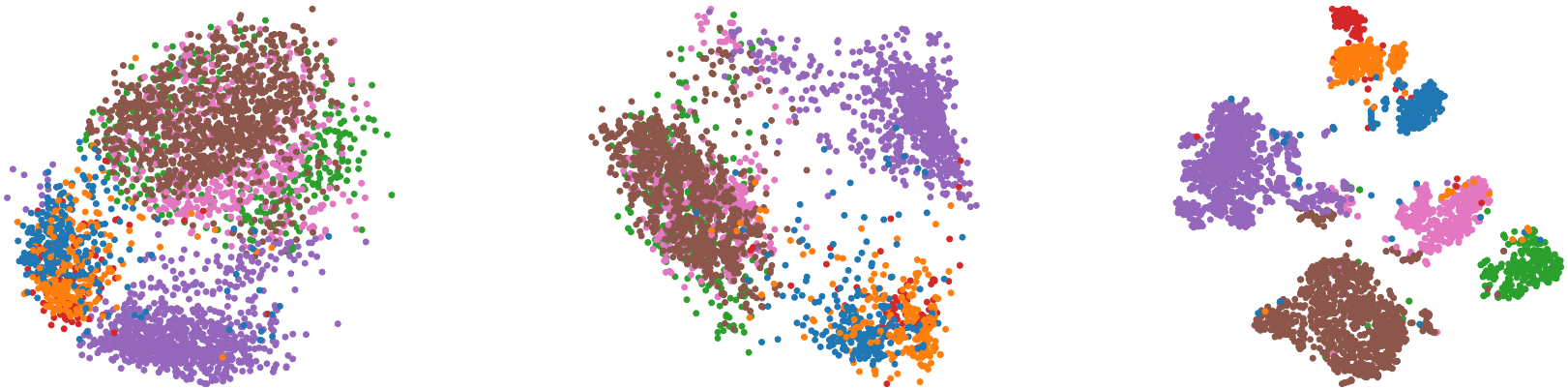
Our work

- We explore three levels of representation learning methods:
  - linear: PCA and robust PCA
  - non-linear: manifold learning
  - deep: VAE
- We apply these methods to a real single-cell expression data set.
- We also give detailed analysis and comparison of different approaches.

Gene expression analysis

- The Mouse Cortex Cells dataset contains
  - 3,005 mouse cortex cells
  - 558 genes
  - labels for 7 distinct cell types
- Cell clustering: to identify the distinct cellular subtypes or states.
  - (1) Dimension reduction
  - (2) Clustering (k-means, Ward's hierarchical clustering method)

Data reduction and visualization



(a) MDS (b) Isomap (c) tSNE

Figure 1: Visualization of the dataset

Data reduction and visualization

PCA and robust PCA

- Top ten PCs and robust PCs

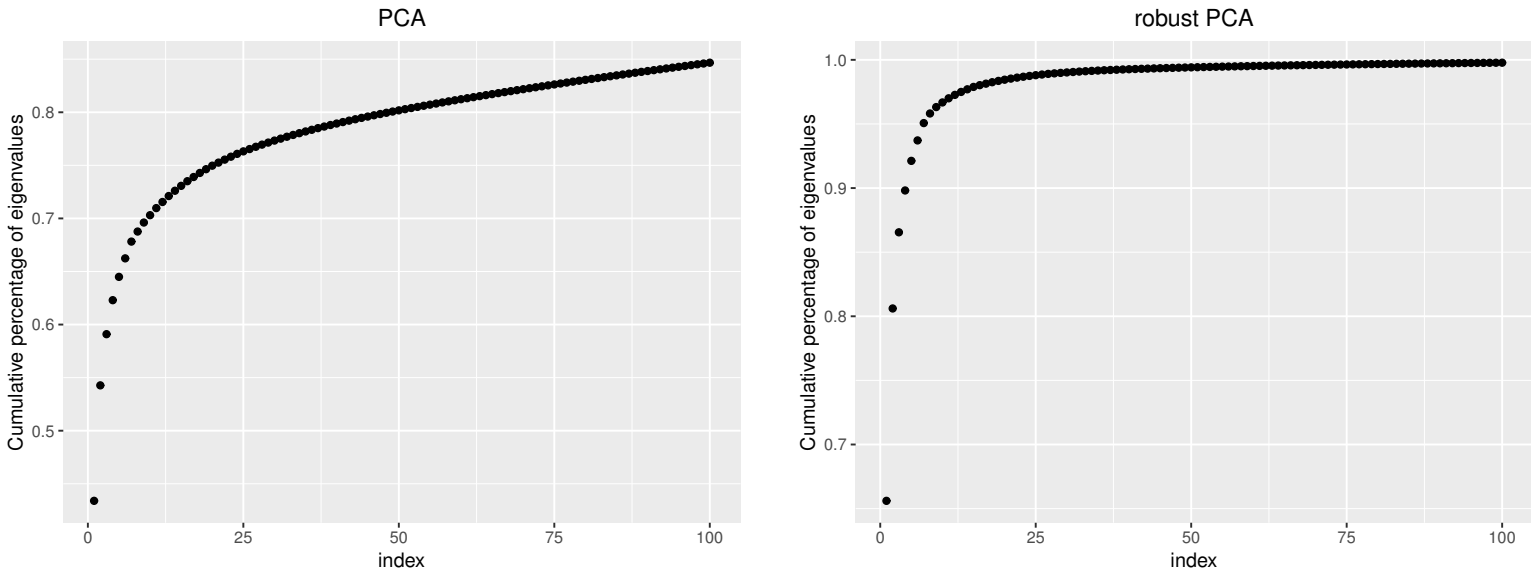


Figure 2: Cumulative percentage of eigenvalues

Clustering

- Performance measure: adjusted Rand index (ARI)
- PCA + kmeans: 0.4030
- RPCA + kmeans: 0.3817
- This dataset is well preprocessed and is of high quality, the outliers may be removed beforehand.

Clustering

- Manifold learning (Isomap, LLE and tSNE) and VAE + k-means and Ward's hierarchical clustering

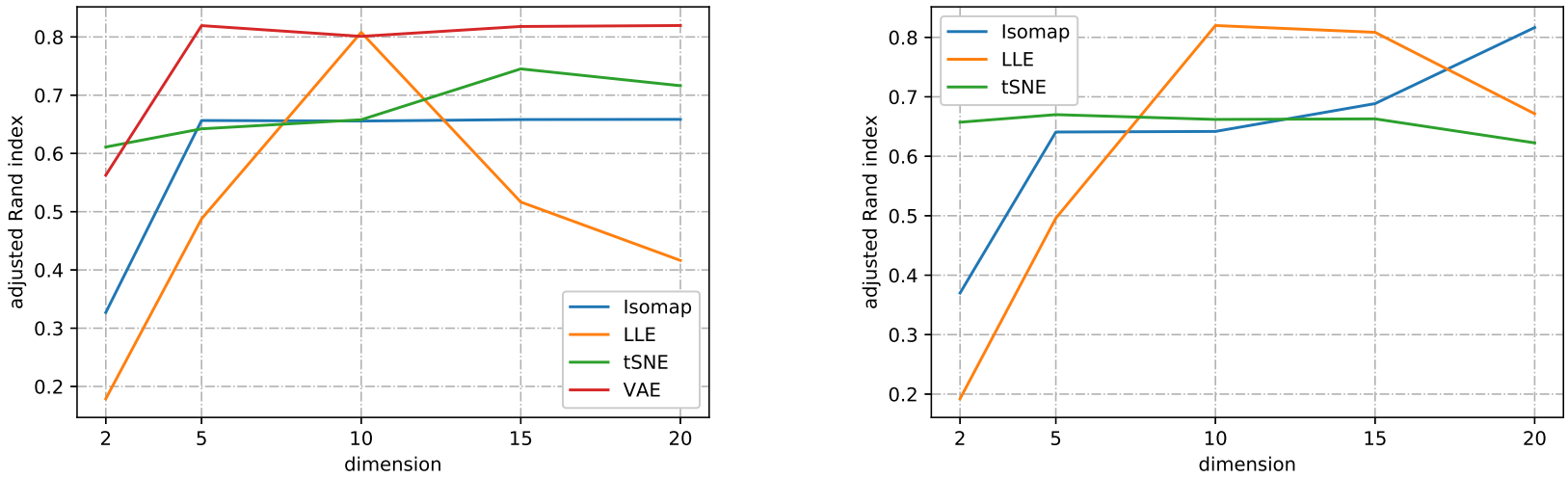
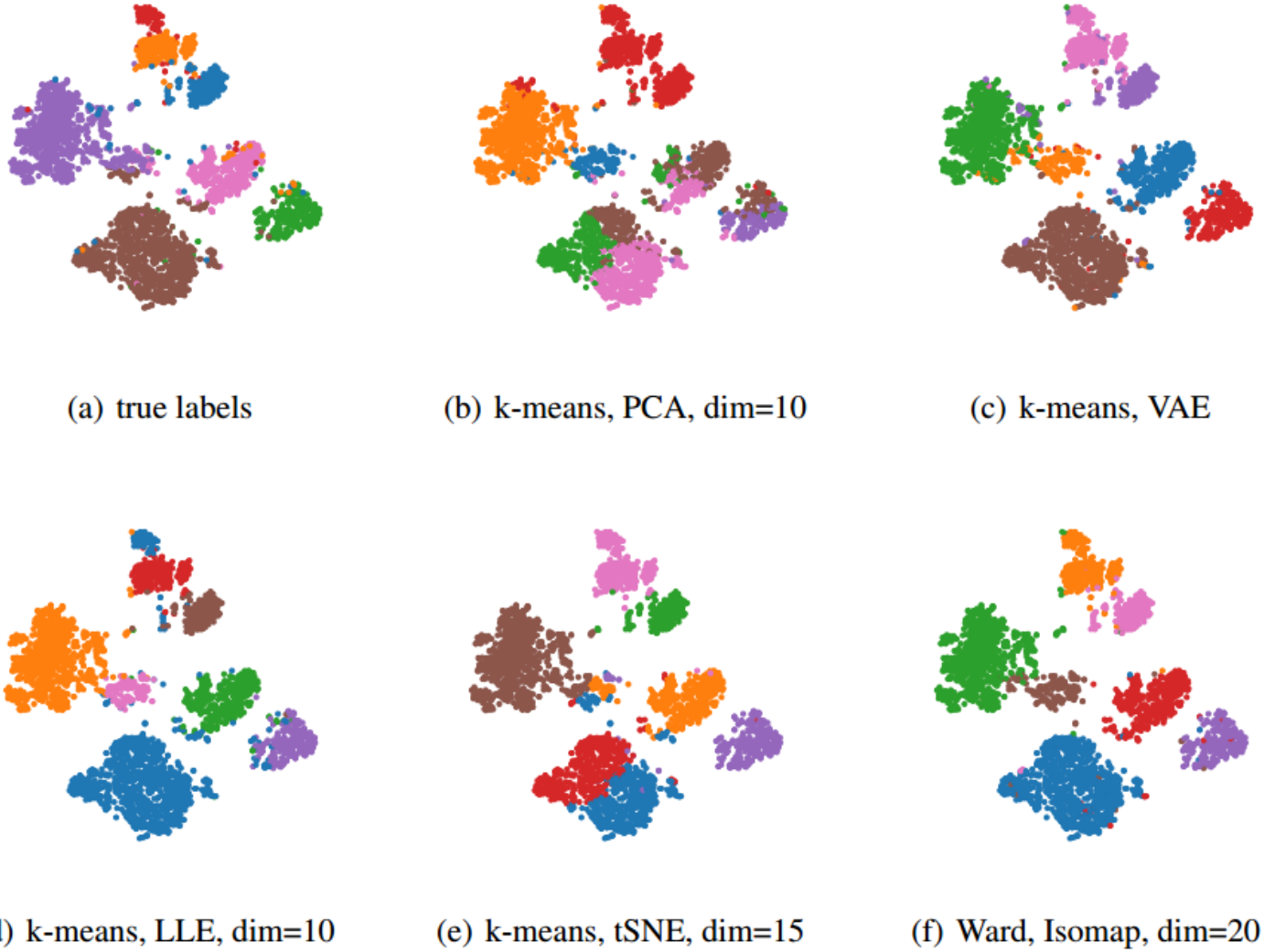


Figure 3: ARI of k-means and Ward's hierarchical clustering with respect to different latent dimensions

Clustering visulization



(a) true labels (b) k-means, PCA, dim=10 (c) k-means, VAE

(d) k-means, LLE, dim=10 (e) k-means, tSNE, dim=15 (f) Ward, Isomap, dim=20

Figure 4: Visualization of some clustering results

Conclusion

- We explore three levels of representation learning methods.
- We apply various methods to single-cell expression analysis.
- We also give detailed analysis and comparison of different approaches as well as different settings of hyperparameters.
- In conclusion, manifold learning and VAE outperform linear dimensionality reduction approaches on this dataset.

Thank you.

References I


Y. Bengio, A. C. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, 2013.


R. A. Fisher, "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 1922.


E. Pierson and C. Yau, "Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis," in *Genome Biology*, 2015.


C. Doersch, "Tutorial on variational autoencoders," *CoRR*, vol. abs/1606.05908, 2016.


D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2014.


A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. L. Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson, "Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq," *Science*, vol. 347, pp. 1138–1142, 2015.

References II


S. Prabhakaran, E. Azizi, A. Carr, and D. Pe'er, "Dirichlet process mixture model for correcting technical variation in single-cell gene expression data," *JMLR workshop and conference proceedings*, vol. 48, pp. 1070–1079, 2016.


R. R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines," in *AISTATS*, 2009.


E. J. Candès, X. Li, Y. Ma, and J. N. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, pp. 11:1–11:37, 2011.


C. Gao, Y. Yao, and W. Zhu, "Generative adversarial nets for robust scatter estimation: A proper scoring rule perspective," *CoRR*, vol. abs/1903.01944, 2019.


Y. Yao, *A mathematical introduction to data science*.
2019.


I. Borg and P. Groenen, "Modern multidimensional scaling: Theory and applications," *Journal of Educational Measurement*, vol. 40, no. 3, pp. 277–280, 2003.


J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.


S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

References III

L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.