A Semi-Supervised Dimensionality Reduction Method to Reduce Batch Effects in Genomic Data

Anusha Murali
Mentor: Dr. Mahmoud Ghandi

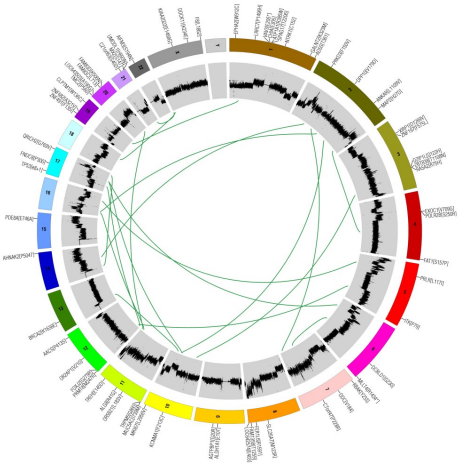
MIT PRIMES Computer Science Conference 2018

October 13, 2018

A Semi-Supervised Dimensionality Reduction Method to Reduce Batch Effects in Genomic Data

Outline

Introduction



Source: M. R. Stratton, P.J. Campbell & P.A. Futreal

Motivation for Our Research

1. Cancer classification based on molecular knowledge has gained widespread acceptance
2. Requires studying the characteristics of thousands of genes
3. Data-mining and machine learning have helped us enormously
4. Problem in Genomic Studies: Occurrence of batch effects in experimental data

Dimensionality Reduction

Dimensionality Reduction

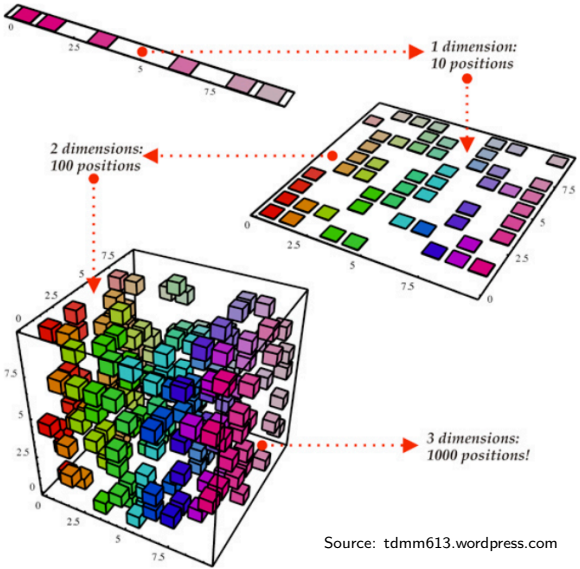
Too many variables in the study?

1. Dimensionality reduction: the process of reducing the number of random variables in the experiment
2. Higher the number of features in the results, the harder it is to visualize the problem
3. Often, most of the features are correlated, and therefore are redundant

Pros and cons

- Pros: Redundant features are removed
- Cons: May lead to some information loss

Dimensionality Reduction



Source: tdmm613.wordpress.com

Popular technique in machine learning

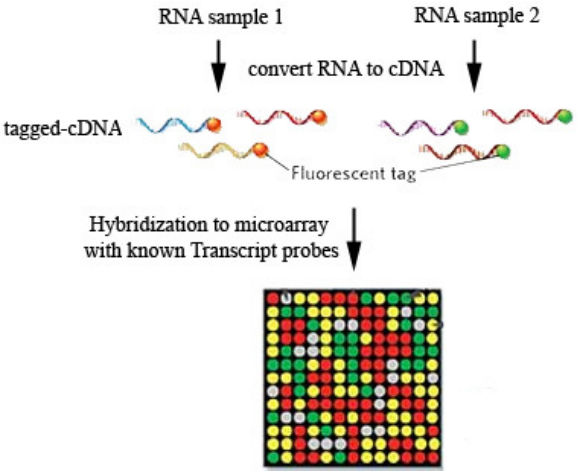
1. Reduction in memory/storage space
2. Easier to visualize and analyze

Minimizing unwanted variations in data

- Dimensionality reduction can help minimize unwanted variations in experimental data (such as batch effects)
- This can be achieved by finding a suitable projection for reducing the dimensions

Microarray & RNA-Seq

DNA Microarray

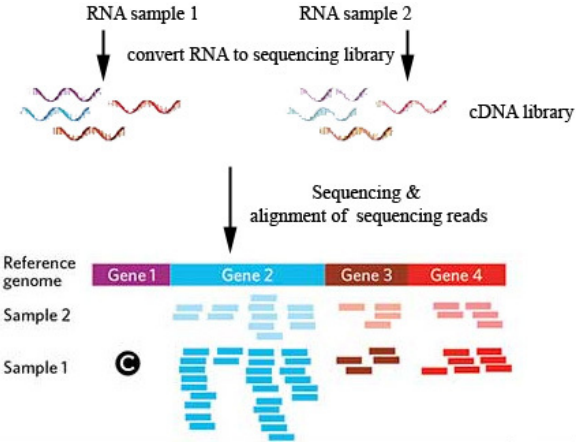


Source: https://www.otogenetics.com/rna-sequencing-vs-microarray

What is Microarray?

1. Each well in the microarray contains an isolated gene
2. Using different fluorescent labels, *c*DNA from healthy and diseased cells are added to the microarray wells
3. Fluorescence intensity level of each probe is proportional to the gene expression level

RNA Sequencing



Source: https://www.otogenetics.com/rna-sequencing-vs-microarray

What is RNA Sequencing?

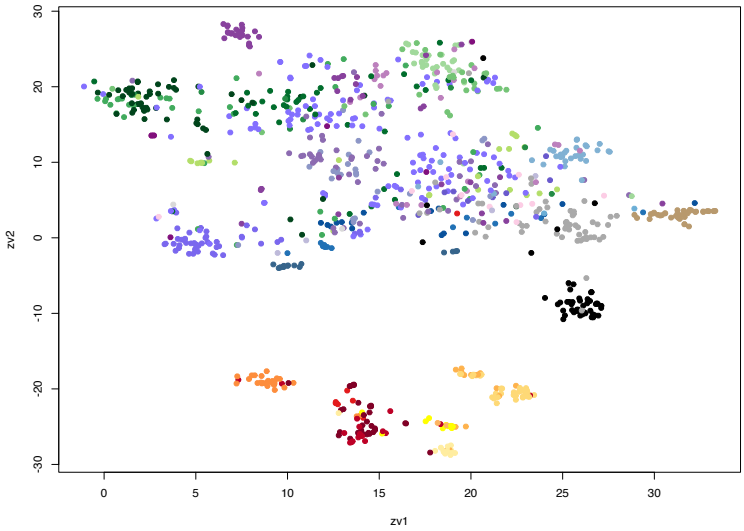
1. Enables us to discover, quantify, and profile RNAs
2. *m*RNA (and other RNAs) is first converted into *c*DNA
3. *c*DNA is then used as the input to a next-generation sequencing library, which counts the number of mappings

Applying t-SNE on Microarray and RNA-seq Data

Cancer Cell Line Encyclopedia Data for the Experiments

1. Our experiments used both Affymetrix (microarray) and RNA-seq data available from the Cancer Cell Line Encyclopedia (CCLE)
2. We pre-processed the data to identify the rows containing identical tissues appearing both in Affymetrix and RNA-seq
3. We used the t-SNE algorithm on the combined data (of $2 \times 994 = 1998$ rows) to visualize the two most important variables in the data

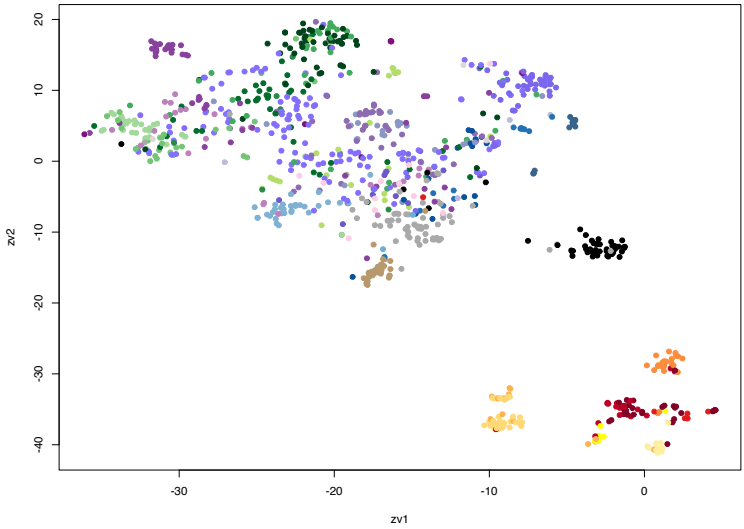
Output of t-SNE for Affymetrix Data Alone



Legend

- B–cell_ALL
- T–cell_ALL
- AML
- CML
- leukemia_other
- multiple_myeloma
- lymphoma_DLBCL
- lymphoma_Burkitt
- lymphoma_Hodgkin
- B–cell_lymphoma_ot
- T–cell_lymphoma_ot
- upper_aerodigestive
- esophagus
- stomach
- colorectal
- melanoma
- lung_NSC
- lung_small_cell
- mesothelioma
- breast
- endometrium
- ovary
- urinary_tract
- prostate
- pancreas
- bile_duct
- liver
- kidney
- thyroid
- glioma
- medulloblastoma
- neuroblastoma
- chondrosarcoma
- osteosarcoma
- soft_tissue
- Ewings_sarcoma
- giant_cell_tumour
- fibroblast_like
- other

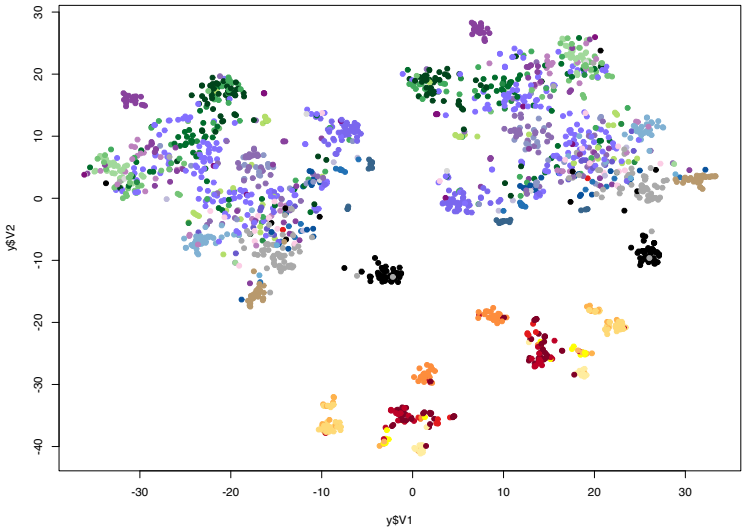
Output of t-SNE for RNA-seq Data Alone



Legend

- B–cell_ALL
- T–cell_ALL
- AML
- CML
- leukemia_other
- multiple_myeloma
- lymphoma_DLBCL
- lymphoma_Burkitt
- lymphoma_Hodgkin
- B–cell_lymphoma_ot
- T–cell_lymphoma_ot
- upper_aerodigestive
- esophagus
- stomach
- colorectal
- melanoma
- lung_NSC
- lung_small_cell
- mesothelioma
- breast
- endometrium
- ovary
- urinary_tract
- prostate
- pancreas
- bile_duct
- liver
- kidney
- thyroid
- glioma
- medulloblastoma
- neuroblastoma
- chondrosarcoma
- osteosarcoma
- soft_tissue
- Ewings_sarcoma
- giant_cell_tumour
- fibroblast_like
- other

t-SNE Output on the Combined Affymetrix and RNA-seq Data

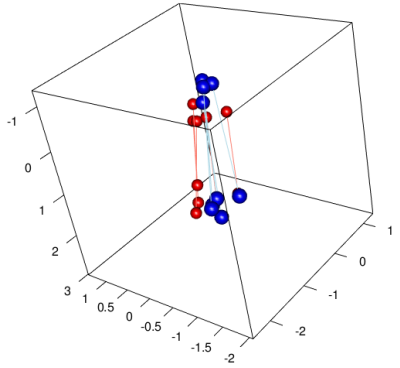


Batch Effect

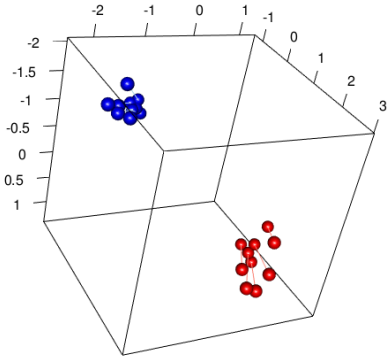
1. Affymetrix and RNA-seq represent the same genomic data
2. So, we expected the points corresponding to the same tissues to cluster together
3. Both sets of data were generated using different technologies
4. We see batch effect!

Minimizing batch effects

1. Often, batch effects are amplified or reduced depending on how the data is viewed
2. Therefore, a possible solution for reducing batch effects is to use a different approach to interpreting the data



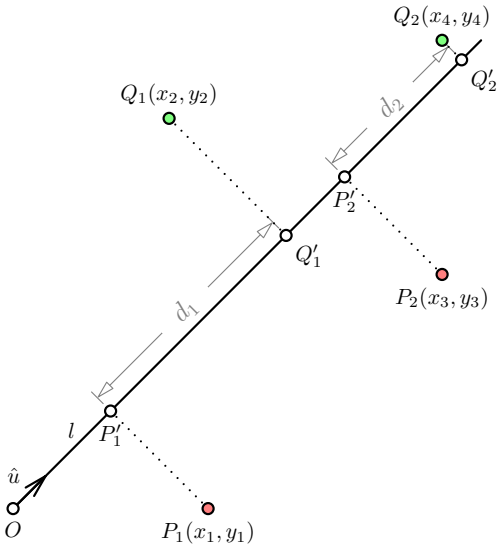
Batch effects present



No batch effects

Optimal Projection Line
for
Minimizing Batch Effects

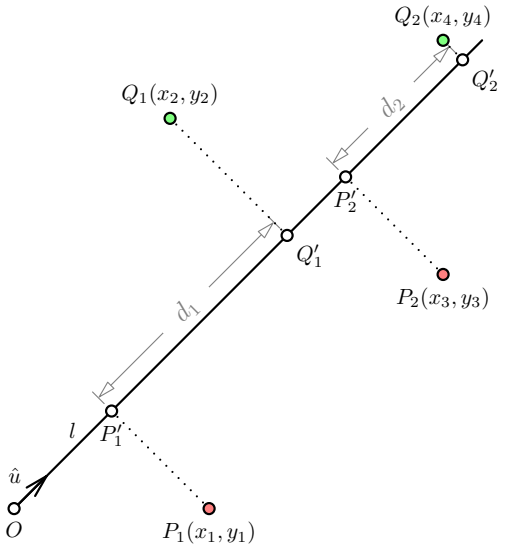
Optimal Projection Line for Minimizing Batch Effects



Finding the most optimal line

1. l: a line with unit vector $\hat{u} = (x, y)$
2. We project the points from combined data onto l
3. Let d_i be the distance between the feet of the corresponding points
4. We want to minimize $f(x, y) = \sum d_i^2$
5. Viewing this as a constrained optimization problem, we use Lagrange multipliers to find the slope m of l

Optimal Projection Line for Minimizing Batch Effects



Optimal line for two pairs of points

- We want to minimize $f(x, y) = d_1^2 + d_2^2$, using the constraint (x, y) is found on line l, which has unit vector $\hat{u} = (x, y)$
- So, our Lagrangian becomes

$$\boxed{\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda(x^2 + y^2 - 1)},$$

where $f(x, y) = d_1^2 + d_2^2$.

- We note that,
  $d_1 = (x_1 - x_2)x + (y_1 - y_2)y$,
  $d_2 = (x_3 - x_4)x + (y_3 - y_4)y$

Optimal Projection Line for Minimizing Batch Effects

Slope m of the optimal line for two pairs of points

- Solving for $\nabla \mathcal{L} = 0$, by setting the partial derivative of each of the three variables, x, y, and λ to zero, we find,

$$\begin{aligned}
\frac{\partial f}{\partial x} &= 2[(x_1 - x_2)^2 + (x_3 - x_4)^2]x + 2[(x_1 - x_2)(y_1 - y_2) + (x_3 - x_4)(y_3 - y_4)]y - 2\lambda x = 0 \\
\frac{\partial f}{\partial y} &= 2[(y_1 - y_2)^2 + (y_3 - y_4)^2]y + 2[(x_1 - x_2)(y_1 - y_2) + (x_3 - x_4)(y_3 - y_4)]x - 2\lambda y = 0 \\
\frac{\partial f}{\partial \lambda} &= -(x^2 + y^2 - 1) = 0
\end{aligned}$$

- Eliminating λ, and substituting $m = y/x$, we obtain, $\boxed{\alpha m^2 + \beta m - \alpha = 0}$, where,

$$\begin{aligned}
\alpha &= (x_1 - x_2)(y_1 - y_2) + (x_3 - x_4)(y_3 - y_4) \\
\beta &= (x_1 - x_2)^2 + (x_3 - x_4)^2 - (y_1 - y_2)^2 - (y_3 - y_4)^2
\end{aligned}$$

- One of the two solutions for m minimizes $f(x, y) = d_1^2 + d_2^2$

Optimal Projection Line for Minimizing Batch Effects

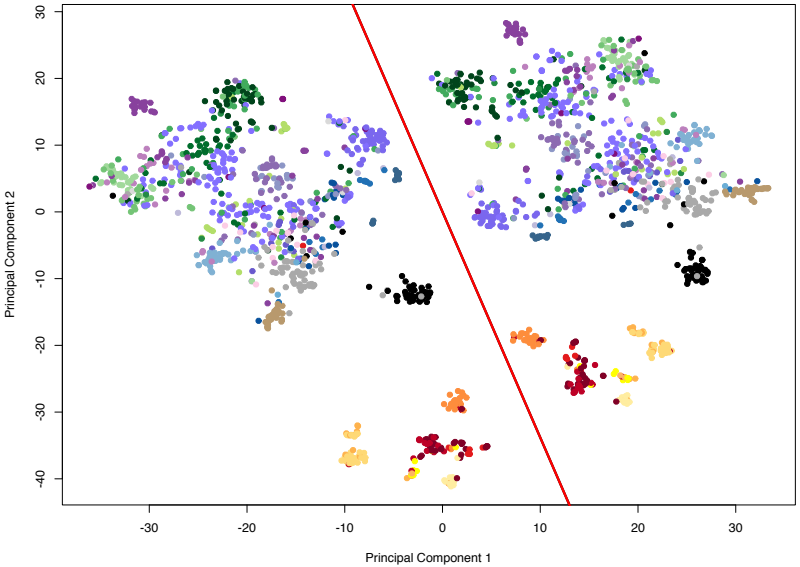
Solving for arbitrary number of n points

- Let $(x_i^{(1)}, y_i^{(1)})$ denote the i^{th} point from batch 1 (RNA-Seq), and $(x_i^{(2)}, y_i^{(2)})$ denote the i^{th} point from batch 2 (Affymetrix)
- As before, using Lagrange multipliers, we find $\boxed{\alpha m^2 + \beta m - \alpha = 0}$, where,

$$\begin{aligned} \alpha &= \sum_{i=1}^{n} \left(x_i^{(1)} - x_i^{(2)}\right)\left(y_i^{(1)} - y_i^{(2)}\right), \\ \beta &= \sum_{i=1}^{n} \left(x_i^{(1)} - x_i^{(2)}\right)^2 - \sum_{i=1}^{n} \left(y_i^{(1)} - y_i^{(2)}\right)^2 \end{aligned}$$

- Using matrix representation, $\alpha = \left(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}\right)^T \left(\boldsymbol{y}^{(1)} - \boldsymbol{y}^{(2)}\right)$ and $\beta = \left(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}\right)^T \left(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}\right) - \left(\boldsymbol{y}^{(1)} - \boldsymbol{y}^{(2)}\right)^T \left(\boldsymbol{y}^{(1)} - \boldsymbol{y}^{(2)}\right)$

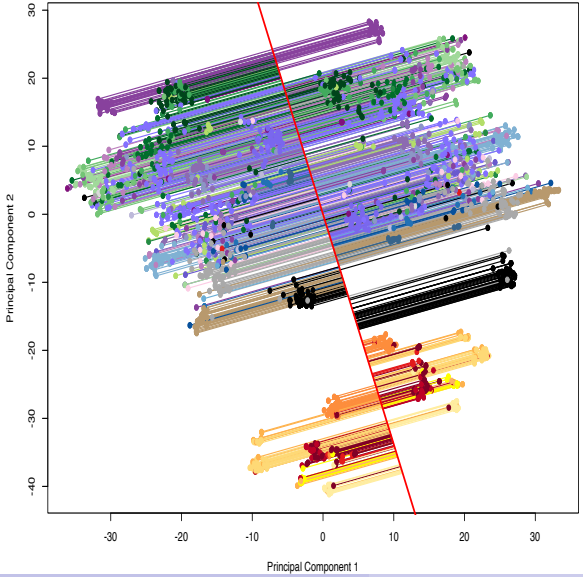
Optimal Projection Line for Minimizing Batch Effects



Projection of data on the optimal Line

- Lagrange multiplier approach found an optimal line (red) that divides the two batches cleanly
- For the given batches in the example, $m = -3.38$

Optimal Projection Line for Minimizing Batch Effects



Projection of data on the optimal Line

- Data corresponding to the same tissues tend to project close to one another!
- Some tissues are still not clustering together
- Projection of multidimensional data onto a one-dimensional line leads to loss of information
- The solution is to project data from the batches onto a higher dimension

Future Direction & Conclusion

Future Direction

Generalization to arbitrary dimensions

1. We want to generalize this approach to find an N×M dimensionality reduction matrix that maps data from N dimensions to M dimensions, minimizing the unwanted variation (e.g. batch effect) in the data
2. $C(\boldsymbol{X}) = \frac{tr(\boldsymbol{X}^T \boldsymbol{P} \boldsymbol{X})}{tr(\boldsymbol{X}^T \boldsymbol{Q} \boldsymbol{X})}$, where $\boldsymbol{X}$ is an N×M matrix of unknowns and $\boldsymbol{P}$ and $\boldsymbol{Q}$ are N×N positive definite constant matrices. We want to find $\boldsymbol{X}$ that minimizes $C(\boldsymbol{X})$ and $\boldsymbol{X}^T \boldsymbol{Q} \boldsymbol{X} = \boldsymbol{I}$
3. $\boldsymbol{P}$ and $\boldsymbol{Q}$ are N×N matrices that generate intra- and inter-cluster distances
4. It could be shown that matrix $\boldsymbol{X}$ consisting of the M eigenvectors corresponding to the smallest eigenvalues of $\boldsymbol{Q}^{-1}\boldsymbol{P}$ is a solution to the above optimization problem[a]
5. We plan to evaluate the application of this method on genomic data

[a] math.stackexchange.com/questions/2915695/minimize-tr-mathbfxt-mathbfp-mathbfx-tr-mathbfxt-mathbfq-math/2919404?noredirect=1#comment6041680_2919404

A SSDR Method to Reduce Batch Effects

Conclusion

1. Batch effects can be incorrectly attributed to outcome of interest, leading to incorrect conclusions of the investigation
2. In this investigation, we studied how to minimize batch effects arising from combining data from two different sources (Affymetrix and RNA-seq)
3. Specifically, we found an optimal straight line to project the output of t-SNE to minimize the batch effects
4. Experimental results using R showed that data from RNA-seq and Affymetrix clustered together on the straight line, eliminating some of the batch effects
5. We plan to generalize the solution to find an $N \times M$ dimensionality reduction matrix that maps data from N dimensions to M dimensions

Acknowledgments

- Dr. Mahmoud Ghandi for his tireless guidance
- The MIT PRIMES program for creating the opportunity
- My family for their unrelenting support


Q&A