Computational and Theoretical Analysis of Novel Dimensionality Reduction Algorithms in Data Mining

BRANDON GUO, MONTA VISTA HIGH SCHOOL

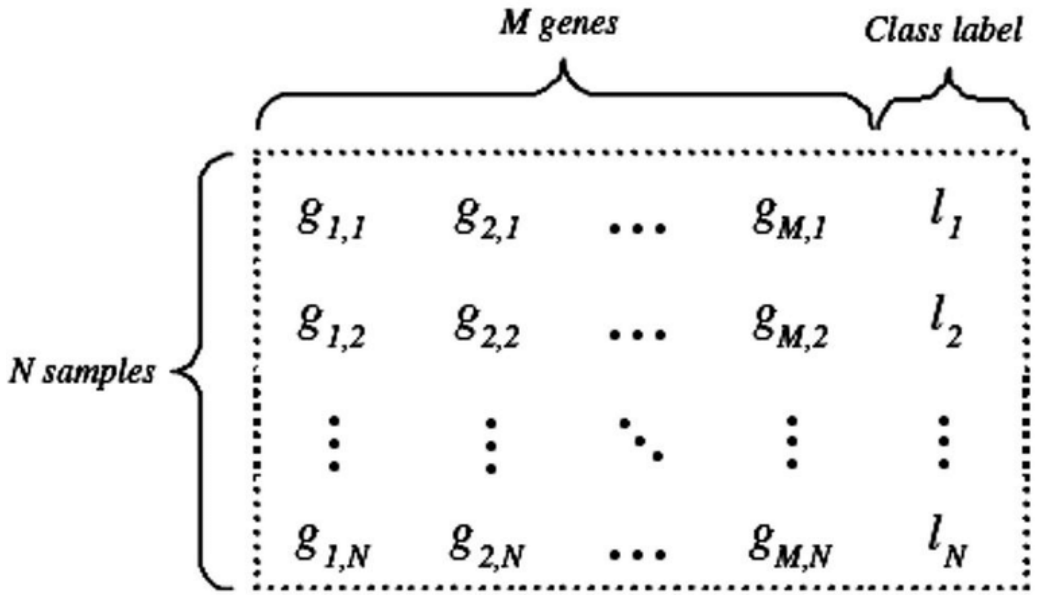
CUPERTINO, CALIFORNIA

Modern Data Mining

- Given: Data with many attributes to explain one result
- Goal: **Understand** the data
- Past developments: Very fast linear regression, neural networks

Case Study: Disease Prediction

Using matrix defined by N samples of M genes and whether they carry disease l, optimize a function f to predict future instances of the array:



The problem

- The value of M is too big!
- How quickly the computer runs an algorithm τ:
  - $\tau \propto n^{-\frac{x}{2x+d}}$
- where d is the number of dimensions (Stone 1982).
- Overfitting is also a problem

The problem (cont.)

We want to reduce the number d enough to retain the important information and but still saving computation time.

Why this problem?

- Implementation of predictive algorithm is trivial
- Data > Algorithm
- How to make data “better”

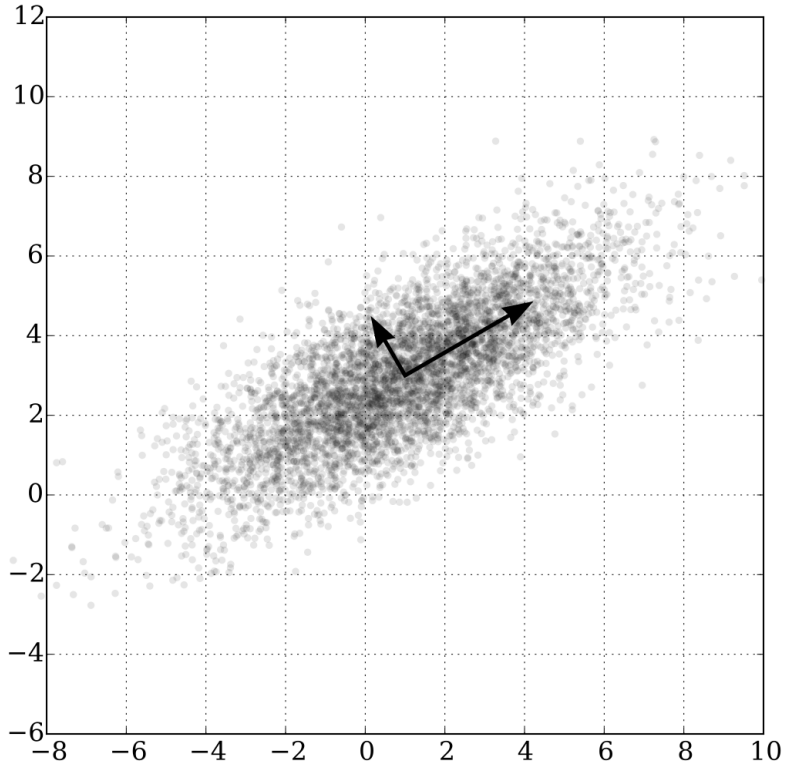
The solution – reduction algorithms

Dimensionality Reduction Algorithm – given a dataset with d predictor variables, construct a new dataset with $k < d$ dimensions that retains as much variance a as possible

Four Algorithms

- Principal Component Analysis
- Kernel Component Analysis
- Nonnegative Matrix Factorization
- Independent Component Analysis

Principal Component Analysis



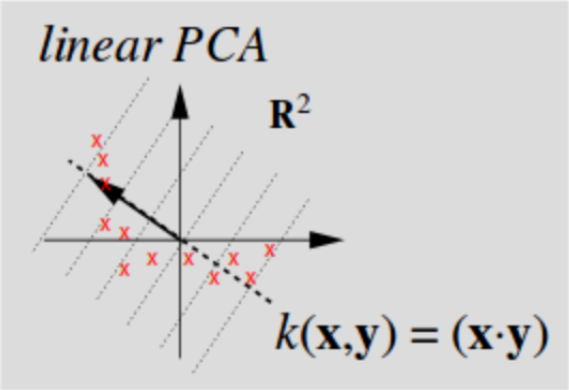
Given a dataset X with covariance matrix C_x, we claim that the maximal eigenvalues of the eigenproblem

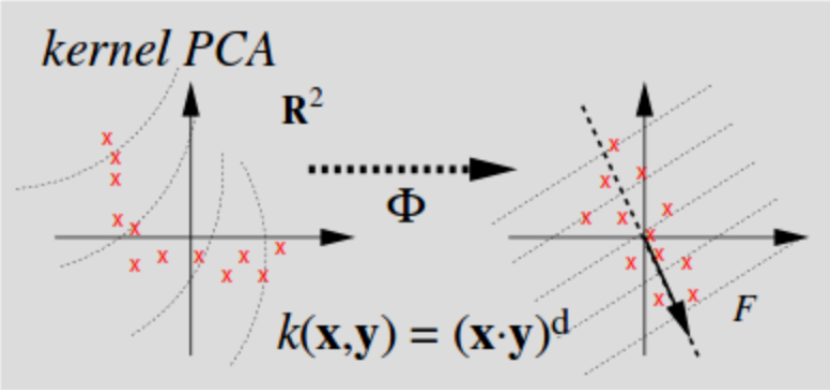
$$C_x W = \lambda W$$

produce the vectors W that capture the most variance in the dataset.

The data is then rotated with k^{th}-best W_i as the axes.

Kernel Component Analysis

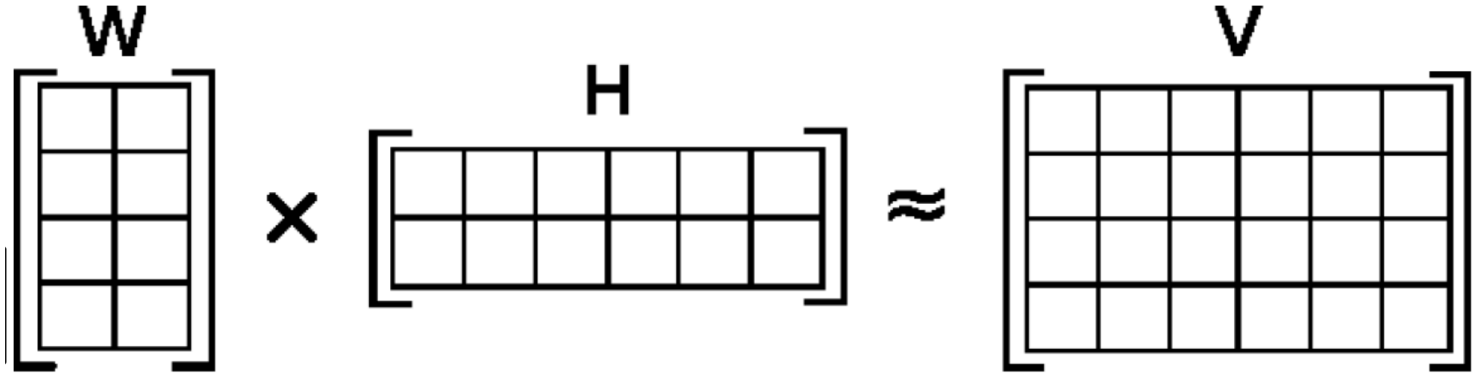




Like PCA, except the eigenvalues are given by $\Phi W = \lambda W$ and Φ is no longer linearly defined like C_x, but rather nonlinearly defined.* These nonlinear models are projected to be straight axes.

*Can be Gaussian, Poisson, exponential, etc.

Nonnegative Matrix Factorization

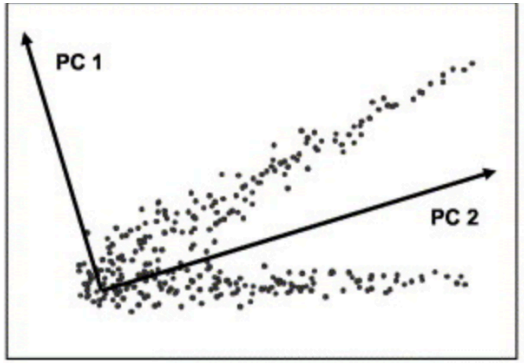


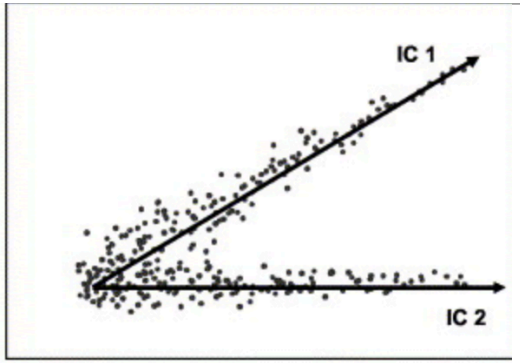
Approximate two matrices W, H that multiply to the original data matrix.

V is a linear combination of W with weights H.

Prediction on W, H allows for dimensions to be omitted (simply by assigning 0 weight)

Independent Component Analysis



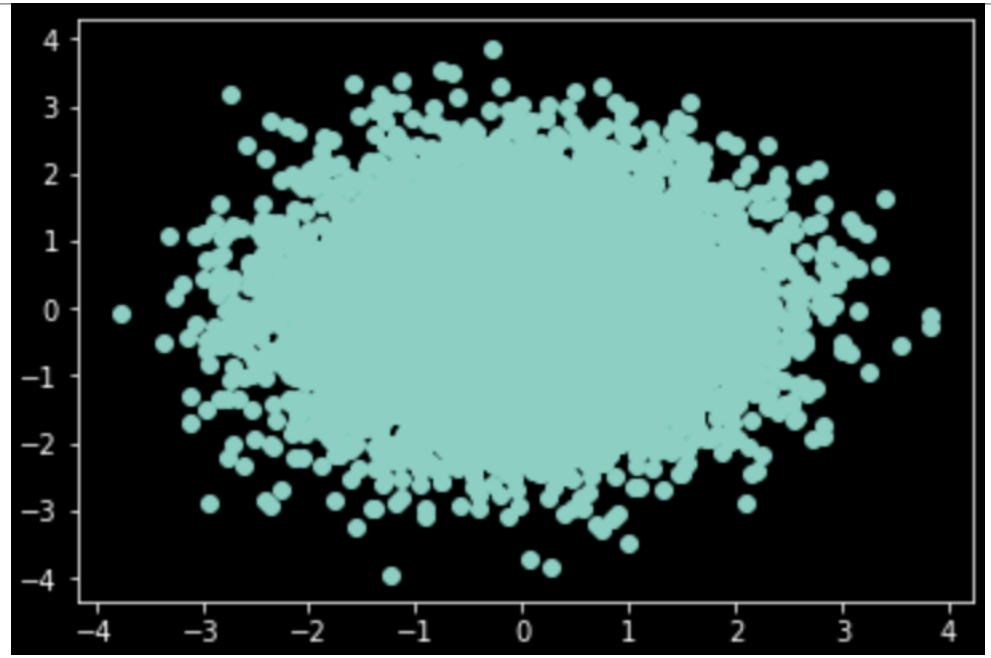


Like PCA, but instead of describing the most variant components, describe the most independent.

Maximize the independence of a set of vectors chosen to describe data

ICA (cont.)

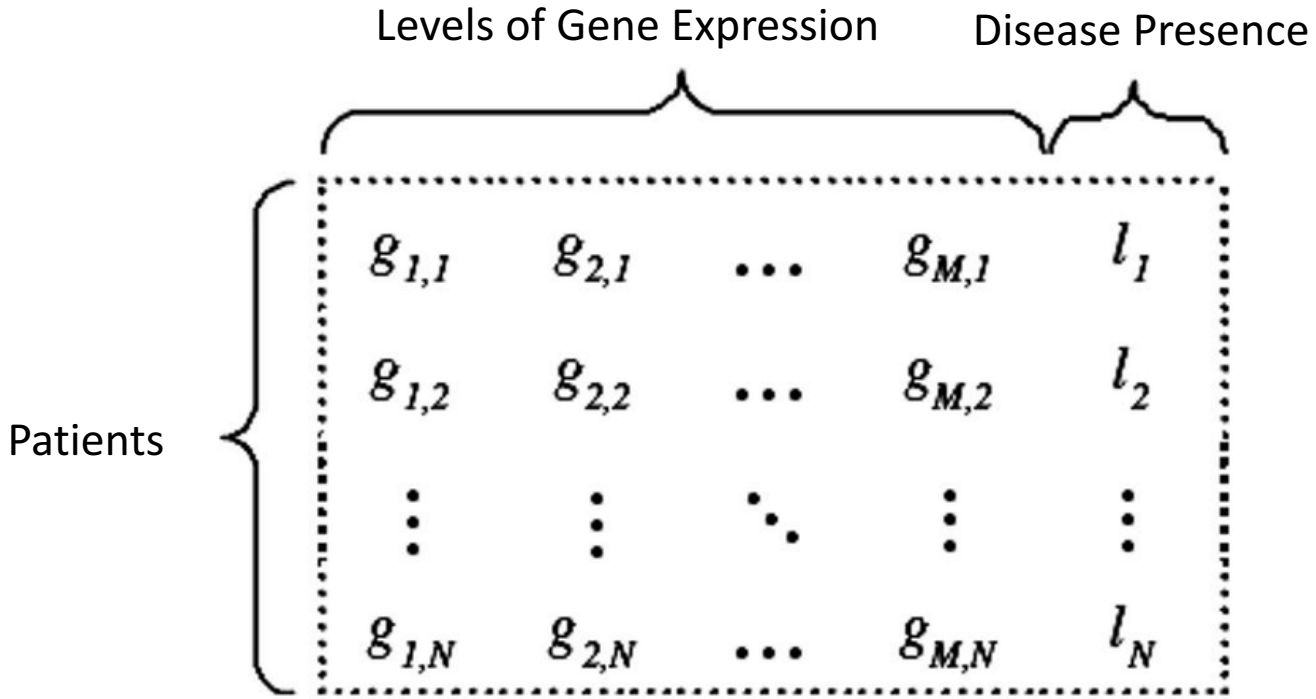
- We don’t want Gaussian data
  - Gaussian distribution is completely symmetric
- Uniquely interested in non-Gaussianity
  - Non-Gaussian when **kurtosis** is nonzero
  - $E(x^4) - 3$
- Converge at a maximal value of kurtosis
  - Each component is the most effective



Plot of 10000 normally distributed numbers $\mathrm{N}(0,1)$

Experimentation

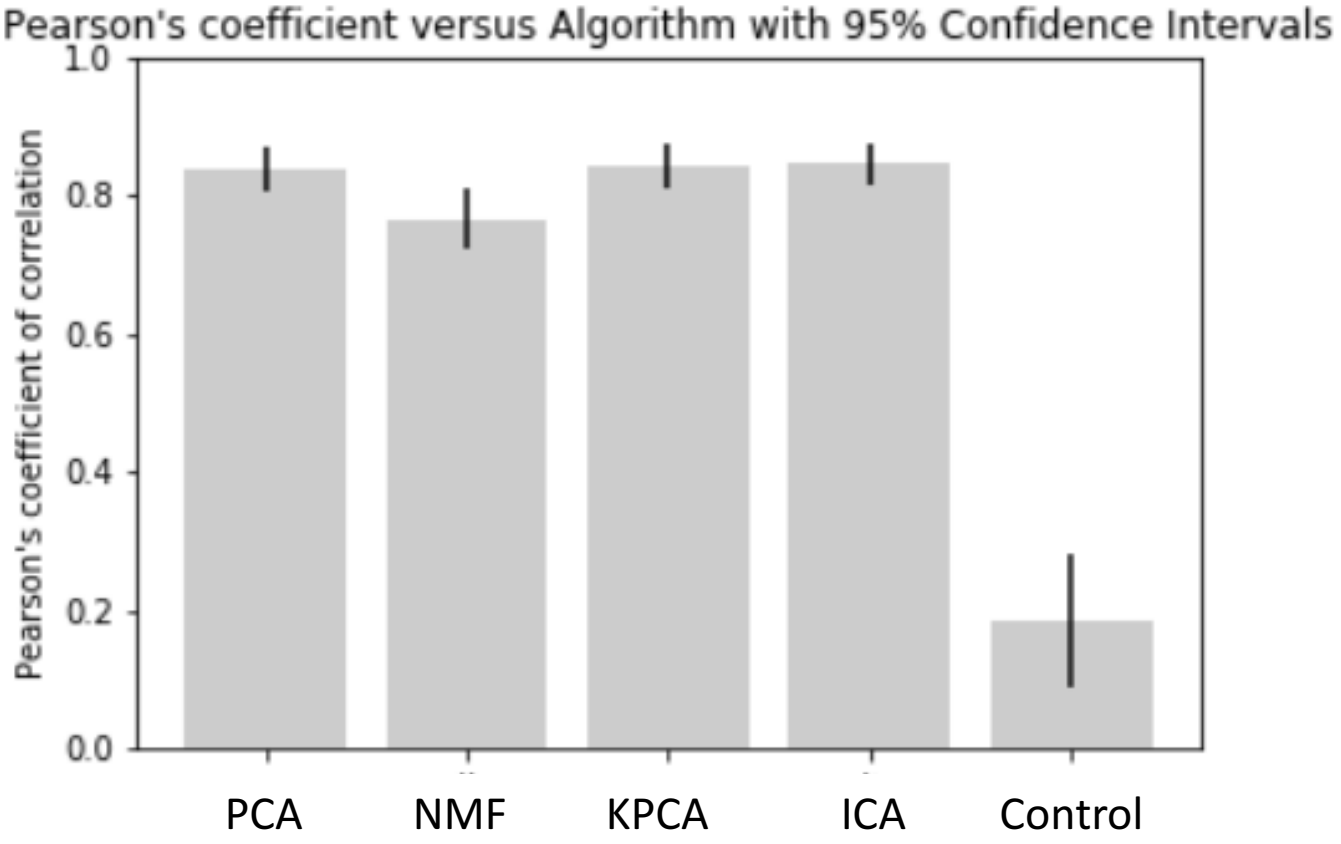
Data from LGRC – Lung Disease: 16721 total dimensions!



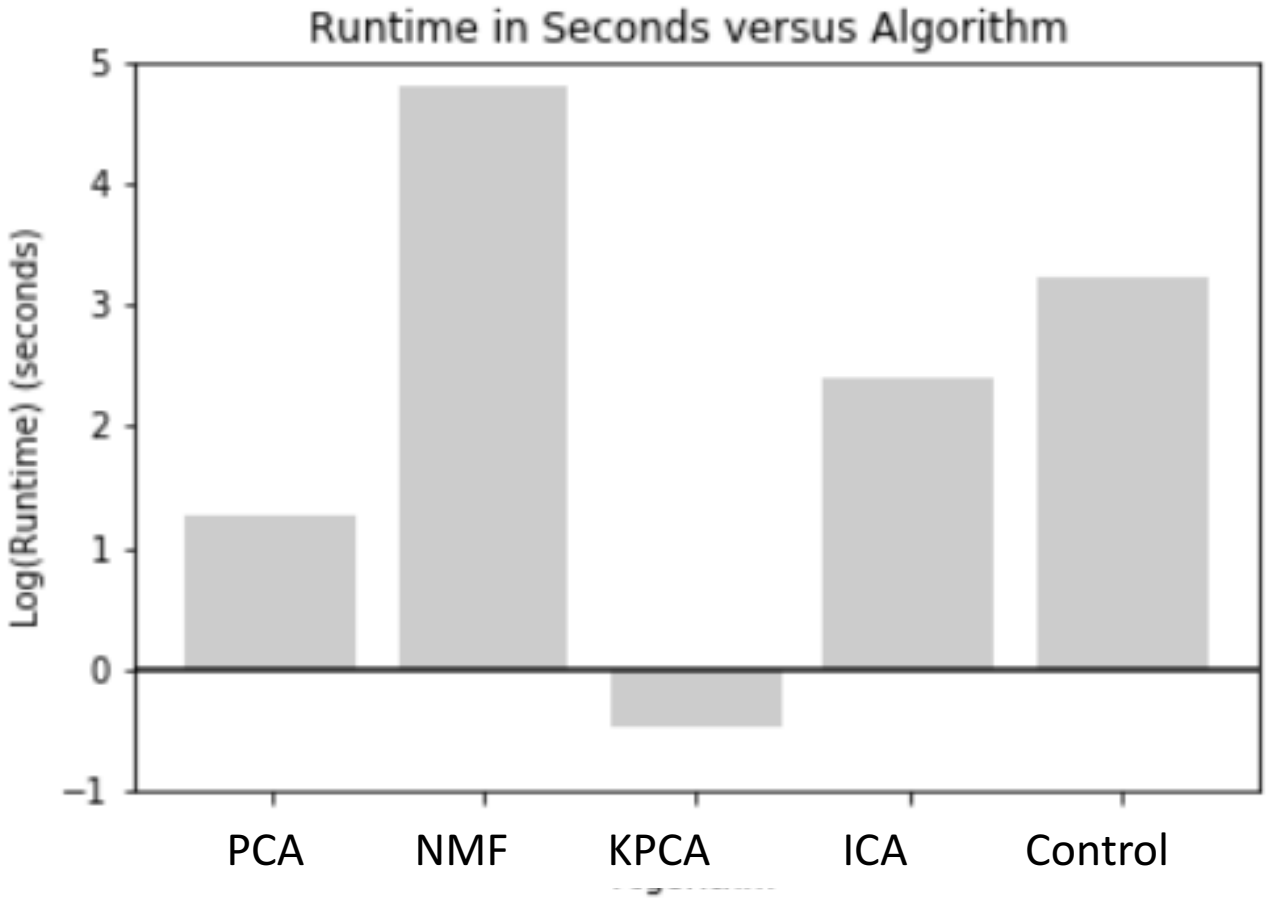
Procedure

- Using each algorithm, reduce dimensions to $600 \ll 16721$
- Use linear regression model on reduced set
- Accuracy by Mean-Squared Error
- Tabulated Runtime
- Calculated R-value as well its 95% confidence interval

Results


Pearson's coefficient versus Algorithm with 95% Confidence Intervals
Pearson's coefficient of correlation
1.0
0.8
0.6
0.4
0.2
0.0
PCA
NMF
KPCA
ICA
Control

Results (cont.)



Conclusions

- On Efficiency
  - Why “less efficient” algorithms exist
- Acknowledgment

References Used

An and Chen 2006: Jiyuan An, Yi-Ping Phoebe Chen, Finding rule groups to classify high dimensional gene expression datasets, Computational Biology and Chemistry, Volume 33, Issue 1,2009, Pages 108-113,ISSN 1476-9271.

Blumer, Anselm $\&$ Ehrenfeucht, Andrzej $\&$ Haussler, David $\&$ K. Warmuth, Manfred. (1989). Learnability and the Vapnik-Chervonenkis Dimension. J. ACM. 36. 929-965. 10.1145/76359.76371.

Burges, C. J. C. (2009). Dimension Reduction: A Guided Tour. Foundations and Trends® in Machine Learning, 2(4), 275–364. https://doi.org/10.1561/2200000002

Foldiak, P $\&$ Young, M (1995). Sparse coding in the primate cortex. The Handbook of Brain Theory and Neural Networks, 895- 898. (MIT Press, Cambridge, MA).

Hyvarinen: A.Hyvärinen and E.Oja. A Fast Fixed-Point Algorithm for Independent Component Analysis.Neural Computation, 9 (7): 1483-1492, October 1997.

Lee and Seung 2001: Lee, Daniel $\&$ Seung, Hyunjune. (2001). Algorithms for Non-negative Matrix Factorization. Adv. Neural Inform. Process. Syst.. 13.

Lung Genomics Research Consortium. Lung Genomics.

Keogh, E., $\&$ Mueen, A. (2017). Curse of Dimensionality. In Encyclopedia of Machine Learning and Data Mining (pp. 314–315). Springer US. https://doi.org/10.1007/978-1-4899-7687-1192

Maaten, L.V., $\&$ Hinton, G.E. (2008). Visualizing Data using t-SNE.Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.

References Used (cont.)

Python Software Foundation. Python Language Reference, version 3.0. Available at http://www.python.org.

Srivastava, Nitish $\&$ Hinton, Geoffrey $\&$ Krizhevsky, Alex $\&$ Sutskever, Ilya $\&$ Salakhutdinov, Ruslan. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 15. 1929-1958.

Stones 1982: Stone, C. (1982). Optimal Global Rates of Convergence for Nonparametric Regression.The Annals of Statistics,10(4), 1040-1053. Retrieved from http://www.jstor.org/stable/2240707

Subramanian, J., $\&$ Simon, R. (2013). Overfitting in prediction models – Is it a problem only in high dimensions? Contemporary Clinical Trials, 36(2), 636–641. https://doi.org/10.1016/j.cct.2013.06.011

Wang, Weiran $\&$ Á. Carreira-Perpiñán, Miguel. (2014). The role of dimensionality reduction in classification. Proceedings of the National Conference on Artificial Intelligence.

Thank you for listening!