

# Extracting Features from Nonnegative Data with topical application to Molecular Biology

Olly Downs

Hopfield Group, Princeton University

<http://olly.downs.net>



*Extracting Features from Nonnegative Data*

1

Enabling the *future*  
through innovation.  
IBM Research

## Overview

- Feature extraction - What do we mean?
- Classical approaches to feature extraction
- Principal Components Analysis
- Novel ideas about bounded data
- Nonnegative Features for faces and the yeast cell cycle
- Summary



*Extracting Features from Nonnegative Data*

2

Enabling the *future*  
through innovation.  
IBM Research

## Feature Extraction

- What do we mean by feature extraction?
- Reconstruct a complicated set of data using only mixtures of a small number of its *key features*
- Consequences
  - for compressing the data
  - for understanding the main factors controlling the process generating it
- To perform this, we must define
  - a criterion (empirical or mathematical) for assessing the 'importance' of components of the data to the whole
  - how to build these 'important' components of the data into meaningful features

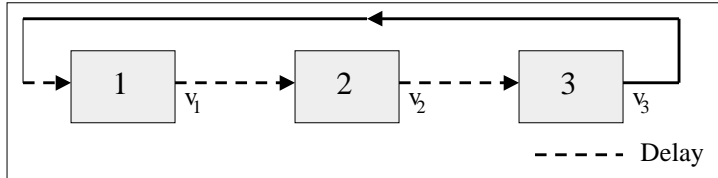


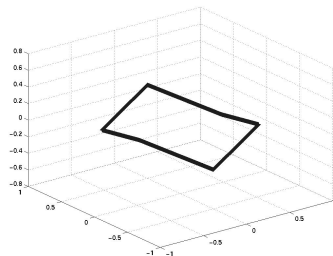
Extracting Features from Nonnegative Data

3

Enabling the *future*  
through innovation.  
Research

## Spatial view of feature extraction

- Consider a simple system which has delayed feedback
- 
- The states of the system do not span the whole of the 'available space'



Extracting Features from Nonnegative Data

4

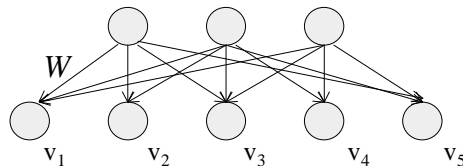
Enabling the *future*  
through innovation.  
Research

## Neural Network Model

- We view extracting a set of features in terms of learning the connections in a two-layer neural network

Hidden Layer

Visible layer



- Mixed activations of the hidden units reconstruct the data
- Assumptions about the hidden units and connections to the visible layer embody how we expect the features to relate to the data.



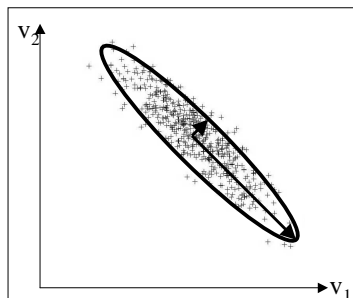
Extracting Features from Nonnegative Data

5

Enabling the *future*  
through innovation.  
Microsoft Research

## Principal Components Analysis

- Mathematically well-understood
  - Hidden layer units are continuous, and Gaussian-distributed
  - Visible layer units are Gaussian distributed
- Finds orthogonal directions of maximal variance in the data
- These directions are the *eigenvectors* of the covariance matrix of the data



Extracting Features from Nonnegative Data

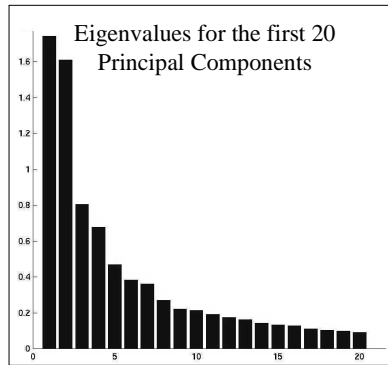
6



Enabling the *future*  
through innovation.  
Microsoft Research

## Eigenfaces and Eigengenes

- Two recent applications of PCA to high-dimensional systems
  - Eigenfaces (Atick et. al 1996)



Extracting Features from Nonnegative Data

7

Enabling the *future*  
through innovation.  
Research

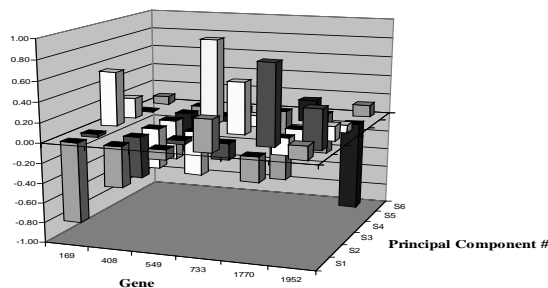
## Eigenfaces and Eigengenes

- Eigengenes (Alter et. al. 2000)
- The analysis tends to capture the smoothly-varying components of the data
- Travelling waves of activation of the eigengenes throughout the cycle

$$\frac{\max - \min}{\max} > \text{Threshold}$$

169	YCK3	CELL PROLIFERATION	PLASMA MEMBRANE-BOUND CASEIN KINASE I
408	PCL6	CELL CYCLE	CYCLIN (PHO85P)
549	VPS15	VACUOLAR PROTEIN TARGET I	SER/THR PROTEIN KINASE
733	MSR1	PROTEIN SYNTHESIS	ARGINYL-TRNA SYNTHETASE
1770	FIG1	MATING	EXTRACELLULAR INTEGRAL MEMBRANE PROTEIN
1952	NUP1	NUCLEAR PROTEIN TARGETIN	NUCLEAR PORE PROTEIN

Eigengenes for the Yeast Cell-Cycle



Extracting Features from Nonnegative Data

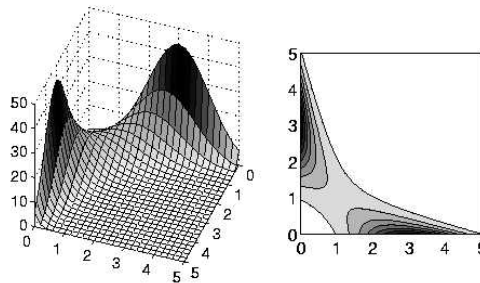
8

Enabling the *future*  
through innovation.  
Research

## Novel ideas about bounded data

(Downs, Lee & MacKay 2000)

- Data which is continuous, but bounded does not conform well to the Gaussian PCA model of  $\text{Mean} \pm \text{Eigenvectors of Covariance Matrix}$
- The *natural* distribution for nonnegative data can have multiple peaks



Extracting Features from Nonnegative Data

9

Enabling the *future*  
through innovation.  
Microsoft Research

## The Nonnegative Matrix Factorisation

(Lee & Seung 1999)

- Learns a set of nonnegative features which *additively* reconstruct the data



Extracting Features from Nonnegative Data

10

Enabling the *future*  
through innovation.  
Microsoft Research

## The Nonnegative Matrix Factorisation

(Lee & Seung 1999)



Extracting Features from Nonnegative Data

11

Enabling the *future*  
through innovation.  
Research

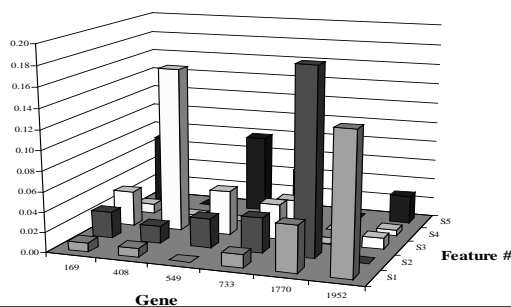
## Nonnegativity in DNA Microarray Analysis

- Since the system (levels of gene expression) is inherently *nonnegative* we expect the NMF to be a more informative model
- We only look at genes exhibiting 'switching' behaviour

$$\frac{\max - \min}{\max} > \text{Threshold}$$

169	YCK3	CELL PROLIFERATION	PLASMA MEMBRANE-BOUND CASEIN KINASE I
408	PCL6	CELL CYCLE	CYCLIN (PHO85P)
549	VPS15	VACUOLAR PROTEIN TARGET I	SER/THR PROTEIN KINASE
733	MSR1	PROTEIN SYNTHESIS	ARGINYL-TRNA SYNTHETASE
1770	FIG1	MATING	EXTRACELLULAR INTEGRAL MEMBRANE PROTEIN
1952	NUP1	NUCLEAR PROTEIN TARGET I	NUCLEAR PORE PROTEIN

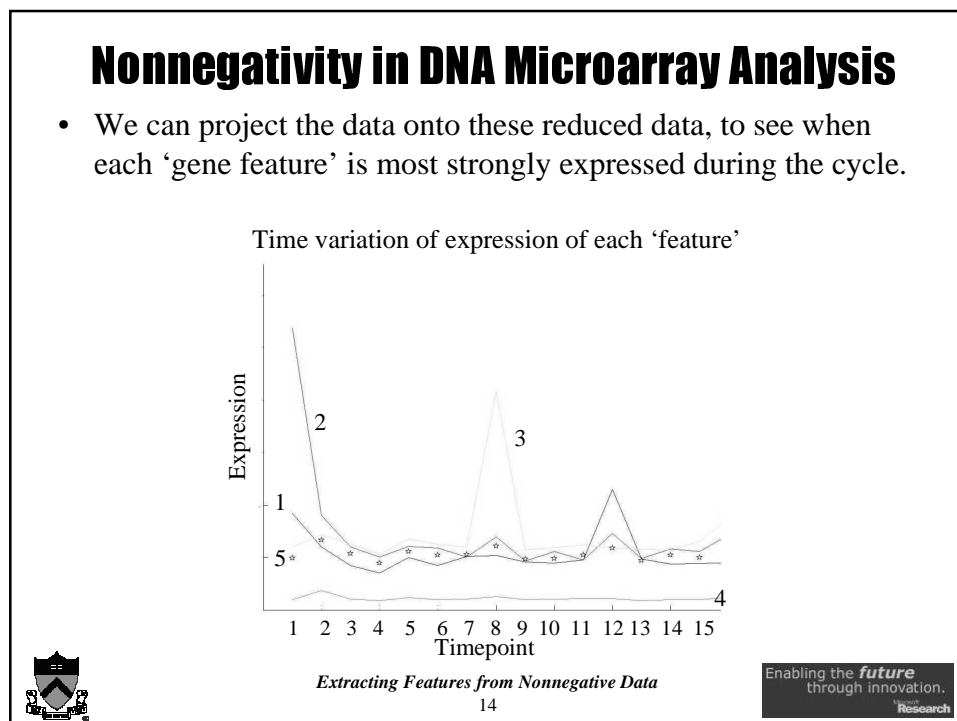
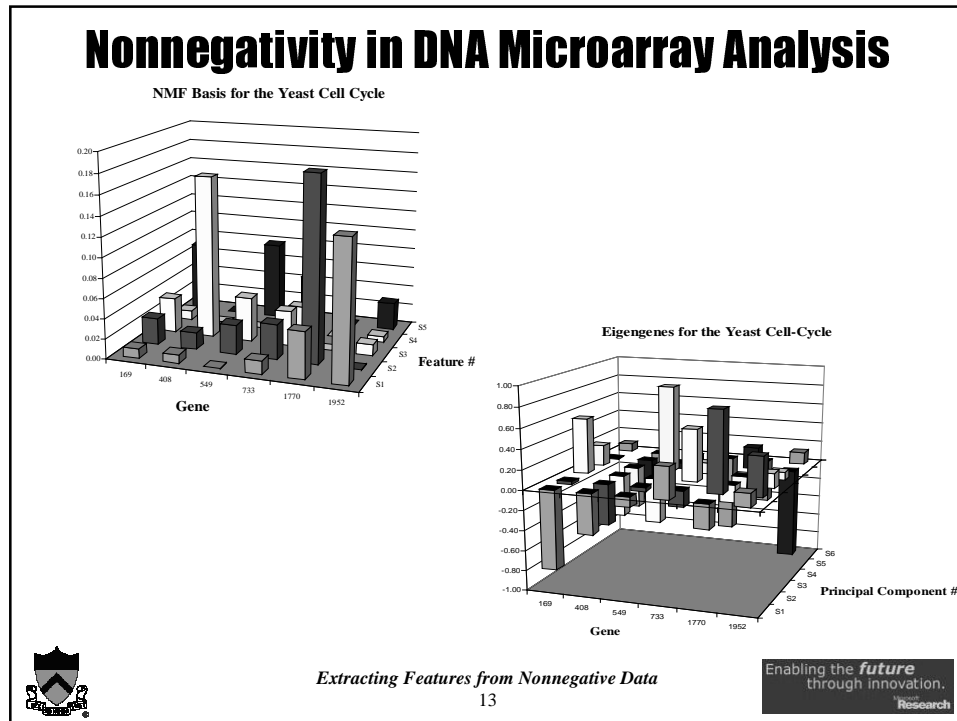
Cell-Cycle Movers and Shakers



Extracting Features from Nonnegative Data

12

Enabling the *future*  
through innovation.  
Research



## Summary

- Brief indication of what feature-extraction is all about
- Common models can be viewed in terms of (artificial) neural networks
- Described PCA in terms of a Gaussian model for the data
- Seen recent work showing Eigenfaces and Eigengenes
- Important distinction between bounded and unbounded (nonnegative) data
- Demonstrated the NMF on faces and on feature-finding in the yeast cell-cycle
- Work continues in trying to describe nonnegative feature extraction in terms of an interpretable probabilistic model (potentially the NNBM).

