FEM21045-20 & FEM31002-20

Machine Learning (in Finance)

Unsupervised Learning - part 1

Dick van Dijk

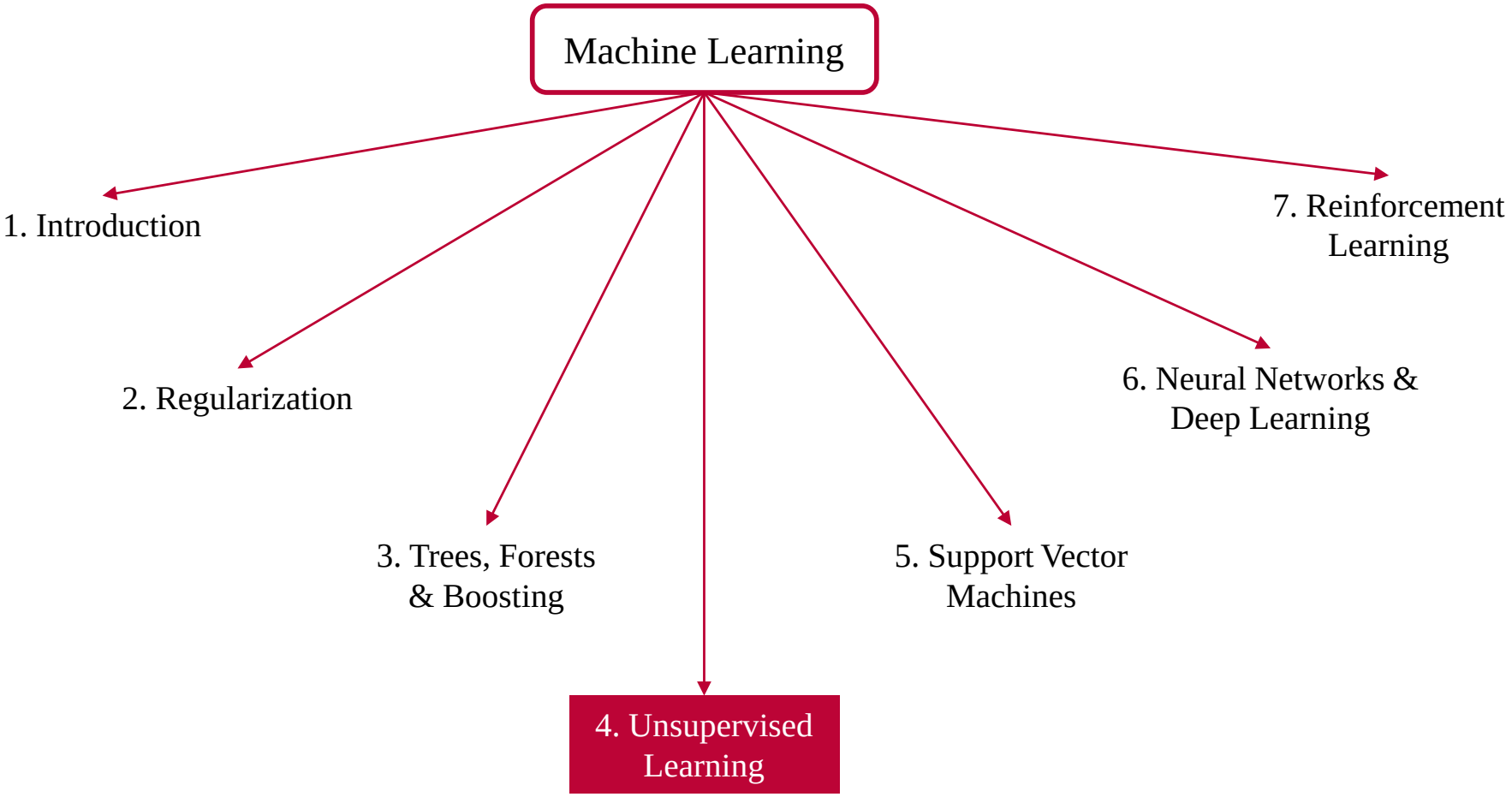
Econometric Institute

Erasmus University Rotterdam

e-mail: djvandijk@ese.eur.nl

Block 1 (Sep-Oct 2020)

Unsupervised Learning



Unsupervised Learning

Outline

- ★ Unsupervised Learning: What and why?
- ★ Principal Components Analysis
- ★ Non-negative Matrix Factorization
  (or: The secret behind online recommendation systems)
- ★ K-means clustering
- ★ Hierarchical clustering
- ★ Gaussian mixture models and the EM algorithm
- ★ Google's PageRank Algorithm

Unsupervised Learning: What and why?

★ Supervised learning: predict output Y using inputs X
We learn about the relationship between Y and X using training sample $\{(x_i, y_i); i = 1, \ldots, N\}$

$\Rightarrow$ Key aspect: both inputs and outputs are observed.

★ Unsupervised learning: we only observe X
$\Rightarrow$ Thus, training sample now is $\{x_i; i = 1, \ldots, N\}$
But still, we want to learn something...

★ Consider X and Y as random variables with joint probability density

$$\Pr(X, Y) = \Pr(Y|X)\Pr(X).$$

Supervised learning focuses on $\Pr(Y|X)$, taking $\Pr(X)$ as given.
Unsupervised learning is all about $\Pr(X)$.

Unsupervised Learning: What and why?

★ $X^T = (X_1, X_2, \ldots, X_p)$ is p-dimensional, and p can be large.

Unsupervised learning aims to get insight into characteristics of the joint density Pr(X), based on N observations $(x_1, x_2, \ldots, x_N)$. In particular, characterize X-values where Pr(X) is large.

Two main types of techniques:

1. Dimension reduction
2. Cluster analysis

Dimension reduction

★ $X^T = (X_1, X_2, \ldots, X_p)$ is p-dimensional, and p can be large.

★ Probably, the X_j's are not (all) independent.

For example, suppose we have information on 'features' of our customers such as age, years of schooling, income, level of education, ZIP code, brand/type of car, holiday expenses, etc, etc

Each feature possibly has some unique information, but they will probably also have some part (or in fact quite a lot) in common.

Hence, in order to analyze and exploit the information in the p variables, it might suffice to consider a lower-dimensional set of q 'driving forces'.

Principal Components Analysis

★ Is it possible to reduce the dimension (p) of the X variables, while still describing a large fraction of their variation?

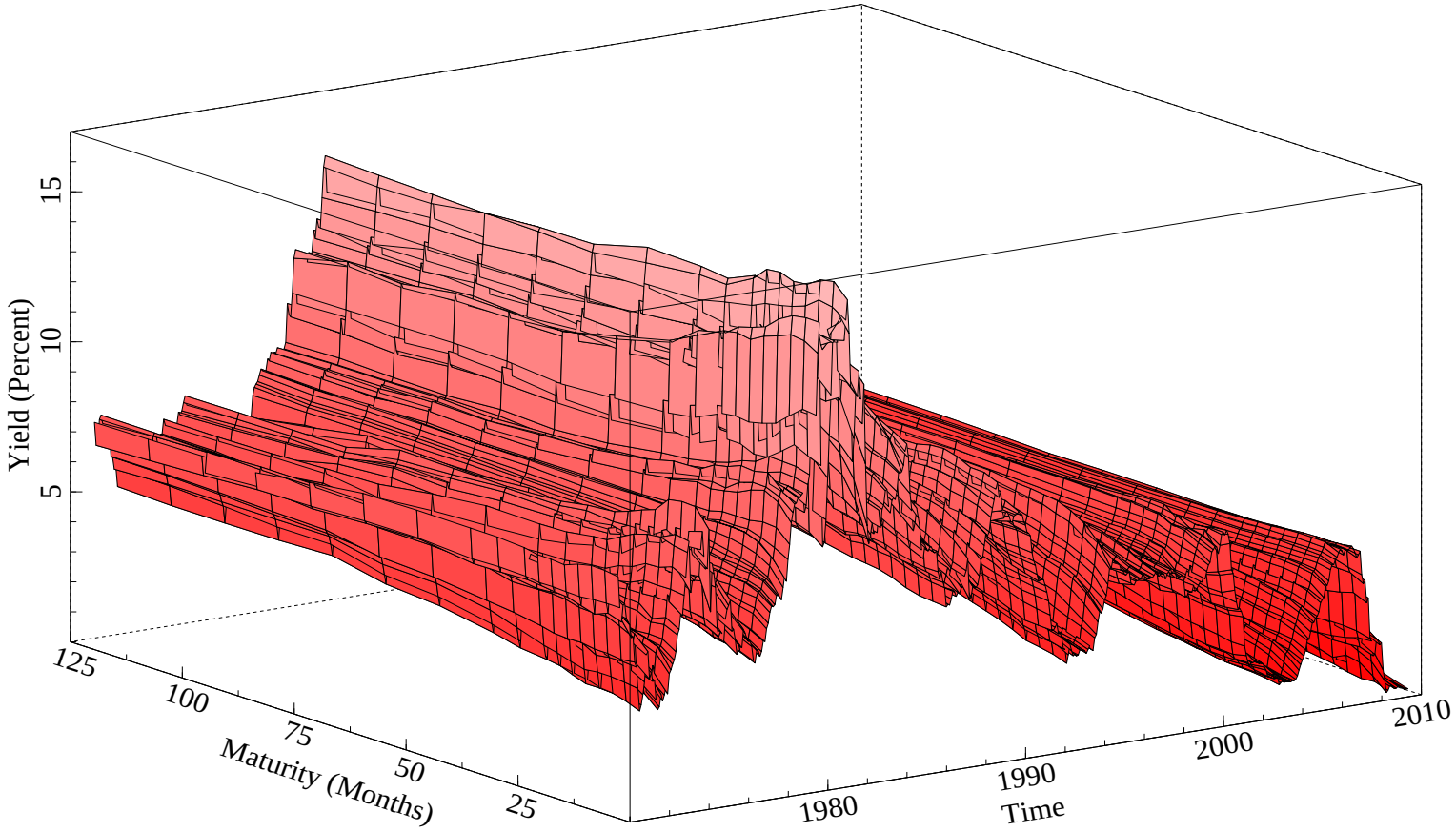
$$x_{ij} = \phi_{j1} z_{i1} + \phi_{j2} z_{i2} + \ldots + \phi_{jq} z_{iq} + e_{ij}, \qquad i = 1, \ldots, N; j = 1, \ldots, p.$$

★ Essential idea: we attempt to describe the (co-)variation in the p-dimensional variable $X^T = (X_1, \ldots, X_p)$ with a limited number of q factors $Z^T = (Z_1, \ldots, Z_q)$, where the 'factors' (directions) are **linear combinations** of $X_1, \ldots, X_p$.

Of course, in general we would look for those factors $Z_1, \ldots, Z_q$ that can **best** describe the (co-)variation in $X_1, \ldots, X_p$.

$\Rightarrow$ This leads to **principal components analysis** [PCA]

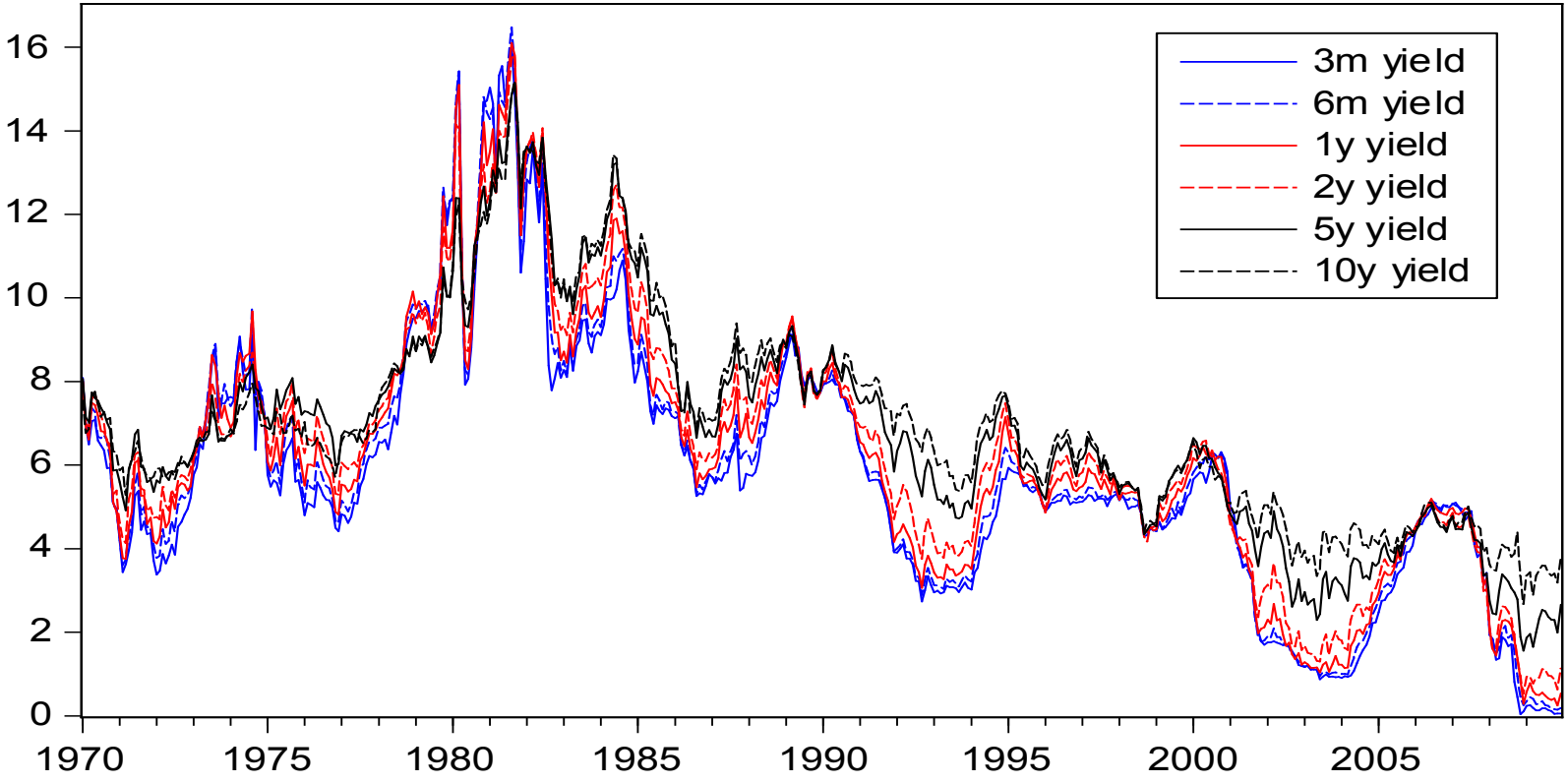
Example: US Treasury yields



US Treasury zero-coupon yields ($p = 17$)

Monthly observations, 1970:1-2009:12 ($N = 480$)

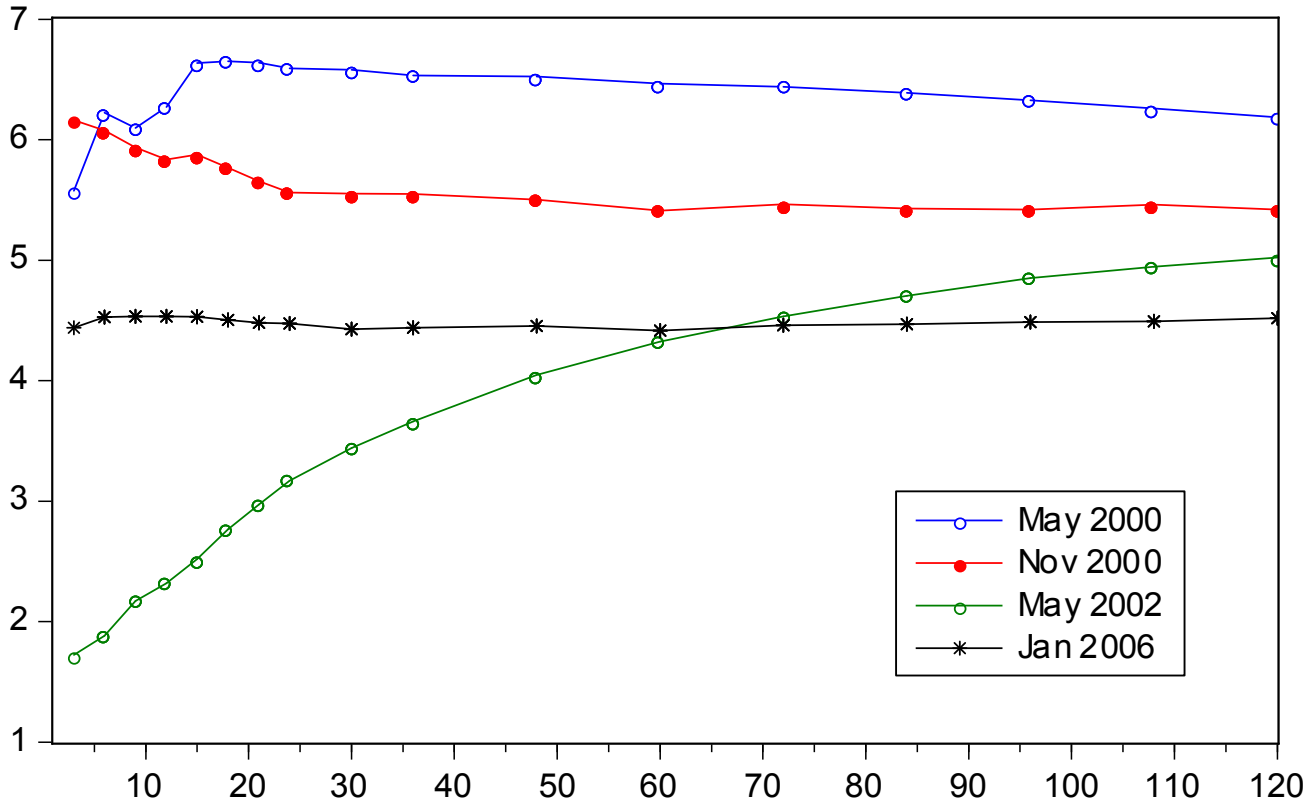
Example: US Treasury yields



US Treasury zero-coupon yields

Monthly observations, 1970:1-2009:12

Example: US Treasury yields



US Treasury zero-coupon yields

Principal Components Analysis

★ Principal components:

$$Z_j = \phi_{1j} X_1 + \phi_{2j} X_2 + \ldots + \phi_{pj} X_p$$

or, in terms of realizations:

$$z_{ij} = \phi_{1j} x_{i1} + \phi_{2j} x_{i2} + \ldots + \phi_{pj} x_{ip}$$

★ z_{ij} $(i = 1, \ldots, N)$ are the *scores* of the jth principal component

★ $\phi_{1j}, \ldots, \phi_{pj}$ are the *loadings* of the jth principal component

★ Assuming that variables are standardized to have mean zero the variance of z_j is equal to

$$\frac{1}{N} \sum_{i=1}^{N} (\phi_j^T x_i)^2 = \frac{1}{N} \phi_j^T \mathbf{X}^T \mathbf{X} \phi_j = \phi_j^T \widehat{\mathbf{\Sigma}}_{\mathbf{x}} \phi_j.$$

Principal Components Analysis

★ Principal components of $\mathbf{X}$ are those linear combinations (directions) that have **maximum variance** and are **uncorrelated**:

1. The first PC is the linear combination $z_{i1} = \phi_1^T x_i$ that maximizes $\text{Var}(z_1) = \phi_1^T \widehat{\Sigma}_{\mathbf{x}} \phi_1$ subject to the constraint $\phi_1^T \phi_1 = \sum_{j=1}^p \phi_{j1}^2 = 1$.
2. The k-th PC is the linear combination $z_{ik} = \phi_k^T x_i$ that maximizes $\text{Var}(z_k) = \phi_k^T \widehat{\Sigma}_{\mathbf{x}} \phi_k$ subject to the constraints $\phi_k^T \phi_k = 1$ and $\phi_k^T \phi_j = 0$ for $j = 1, \ldots, k-1$.

★ The principal component directions $\phi_1, \phi_2, \ldots$ are the ordered sequence of eigenvectors of $\widehat{\Sigma}_{\mathbf{x}}$. The variances of the PCs are the corresponding eigenvalues $\lambda_1, \lambda_2, \ldots$ (with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$).

Principal Components Analysis

★ $\frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}$ is the fraction of the total variance in $\mathbf{X}$ 'explained' by the k-th principal component.

★ With PCA, we hope to find some $q \ll p$, such that

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

is reasonably large.

★ Different approaches to determine an 'appropriate' value of q.

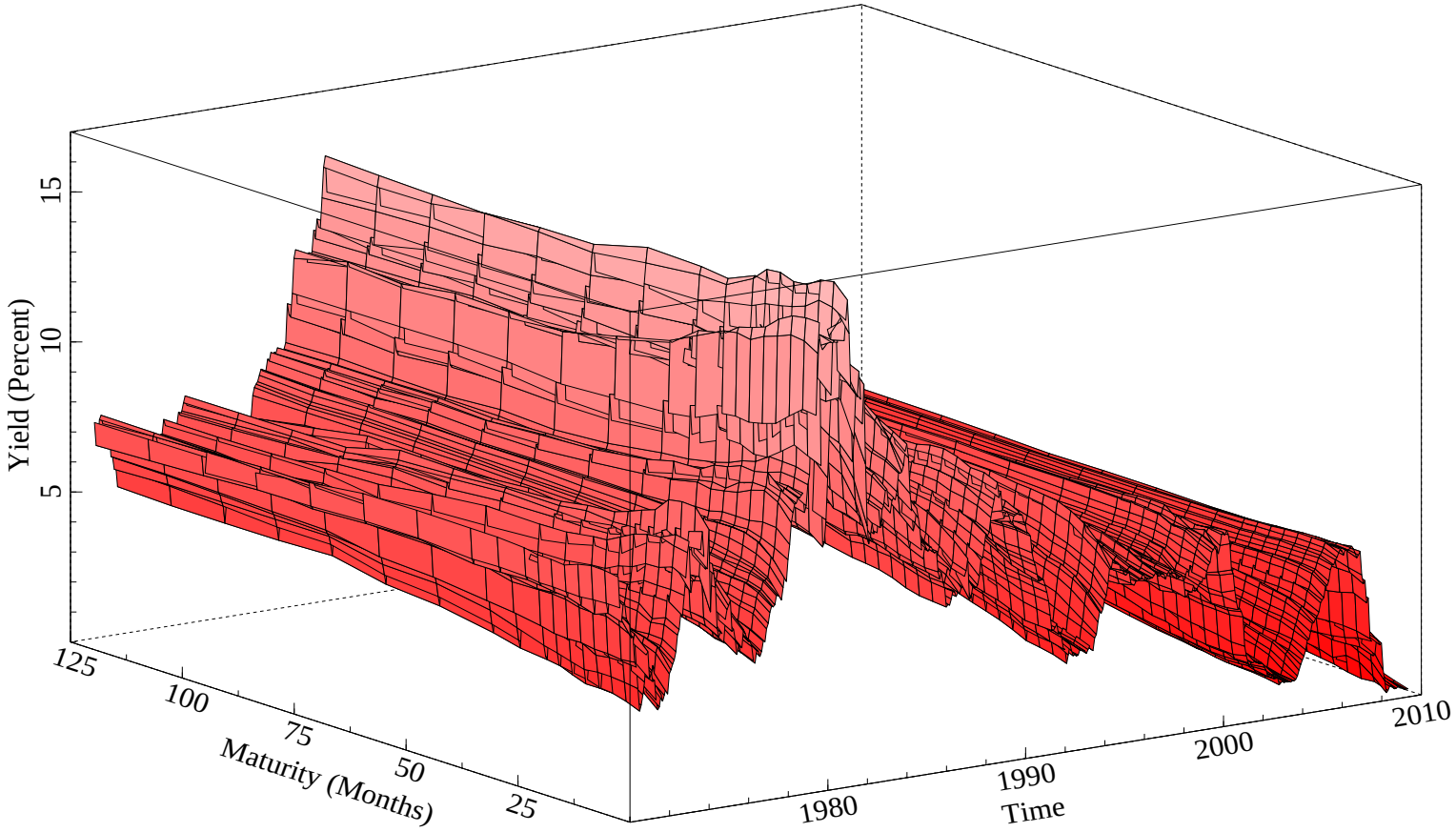
Principal Components Analysis

★ What is an appropriate value of q?

$\Rightarrow$ Different criteria may be used:

- Choose q as small as possible, but such that the first q PCs explain at least a fraction δ of the total variance, for some threshold $0 < \delta < 1$
- Choose q such that $\lambda_q > 1$ but $\lambda_{q+1} \leq 1$ [when X is standardized to have unit variances]
- Scree plot
- Choose q based on the interpretation of the PCs

Example: US Treasury yields



US Treasury zero-coupon yields

Monthly observations, 1970:1-2009:12.

Example: US macro data

• 116 monthly US macro-economic variables for the period 1970:1-2003:12.

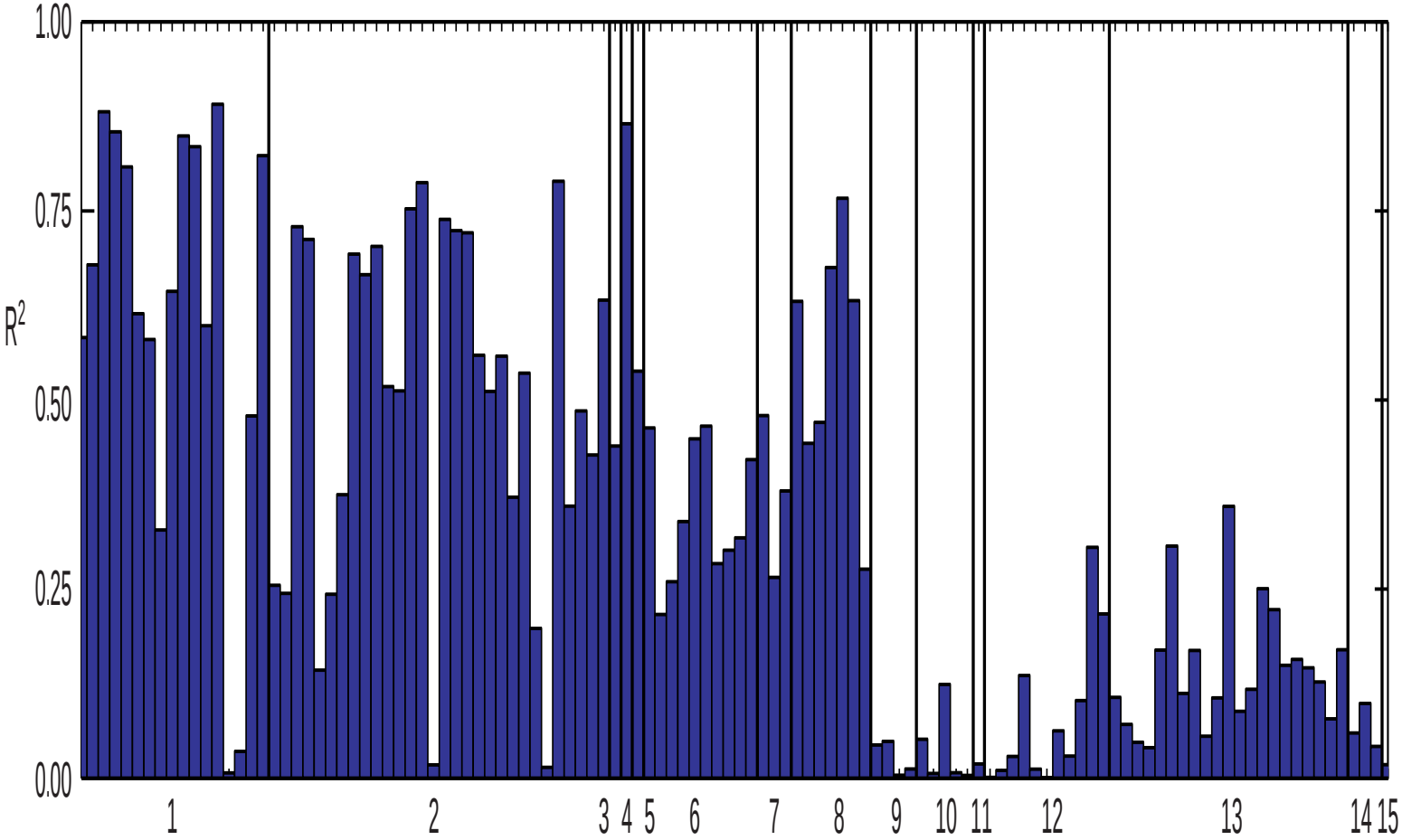
15 different categories [including real output, employment, housing starts, price indexes,...].

$\Rightarrow$ First 3 PCs explain roughly 60% of the variation.

'Tentative' interpretation:

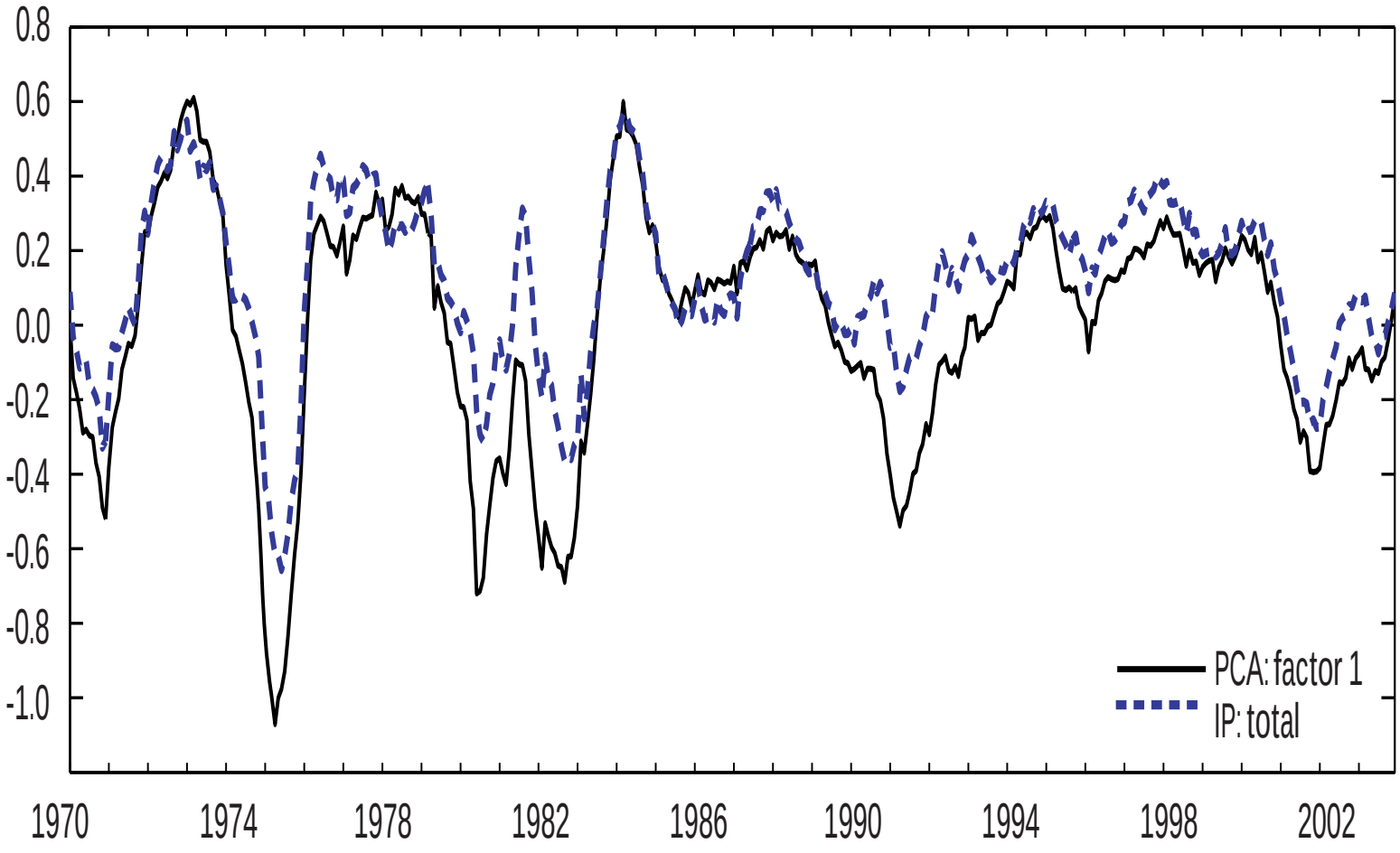
- 1st PC: **business cycle / real activity**
- 2nd: **inflation**
- 3rd: **monetary aggregates**.

US macro data – 1st PC



R^2 of regression of individual variables on PC #1.

US macro data – 1st PC



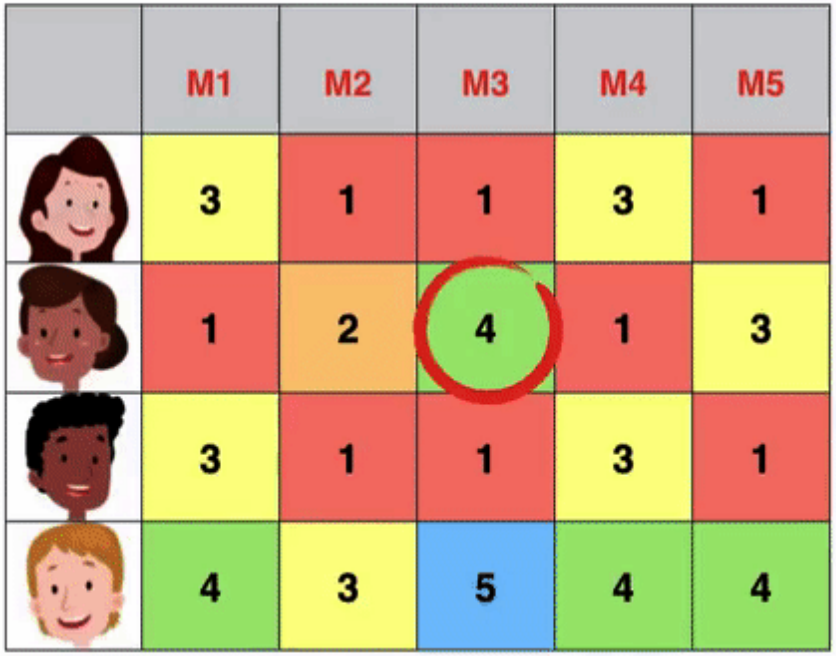
PC #1 & IP growth.

Non-negative Matrix Factorization

Approximate the $N \times p$ data matrix $\mathbf{X}$ by

$$\mathbf{X} \approx \mathbf{W}\mathbf{H}$$

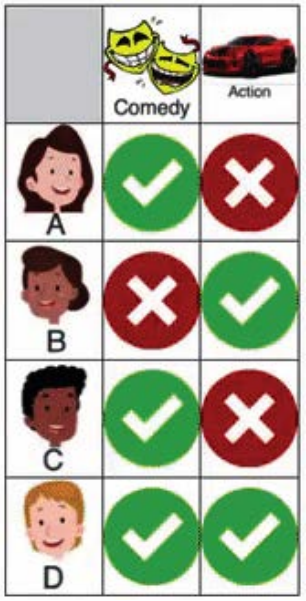
where $\mathbf{W}$ is $N \times r$ and $\mathbf{H}$ is $r \times p$, with r (hopefully) much smaller than N and p.



	M1	M2	M3	M4	M5
	3	1	1	3	1
	1	2	4	1	3
	3	1	1	3	1
	4	3	5	4	4

Non-negative Matrix Factorization

	M1	M2	M3	M4	M5
Comedy	3	1	1	3	1
Action	1	2	4	1	3



	M1	M2	M3	M4	M5
A	3	1	1	3	1
B	1	2	4	1	3
C	3	1	1	3	1
D	4	3	5	4	4

Non-negative Matrix Factorization

$\mathbf{W}$ and $\mathbf{H}$ can be found by minimizing

$$||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 = \sum_{(i,j)} (x_{ij} - w_i h_j)^2$$

- ★ Estimation can be done using Alternating Least Squares (but only gives local optimum)
- ★ Estimation is done for fixed r
- ★ Regularization terms (Ridge, LASSO, etc) can be added
- ★ $\mathbf{X}$ may have some empty cells, or may in fact be rather sparse (we only need to estimate $(N + p)r$ unknown elements in $\mathbf{W}$ and $\mathbf{H}$, which is substantially smaller than the Np cells in $\mathbf{X}$ for typical values of N, p and r.)
  $\Rightarrow$ In fact, empty cells are the most interesting ones, because they can be predicted once $\mathbf{W}$ and $\mathbf{H}$ have been estimated, and can then be used to make recommendations!