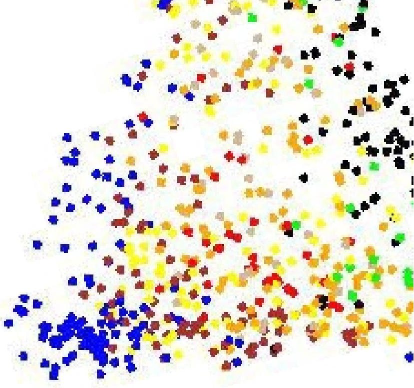


Pattern Change Discovery between High Dimensional Data Sets

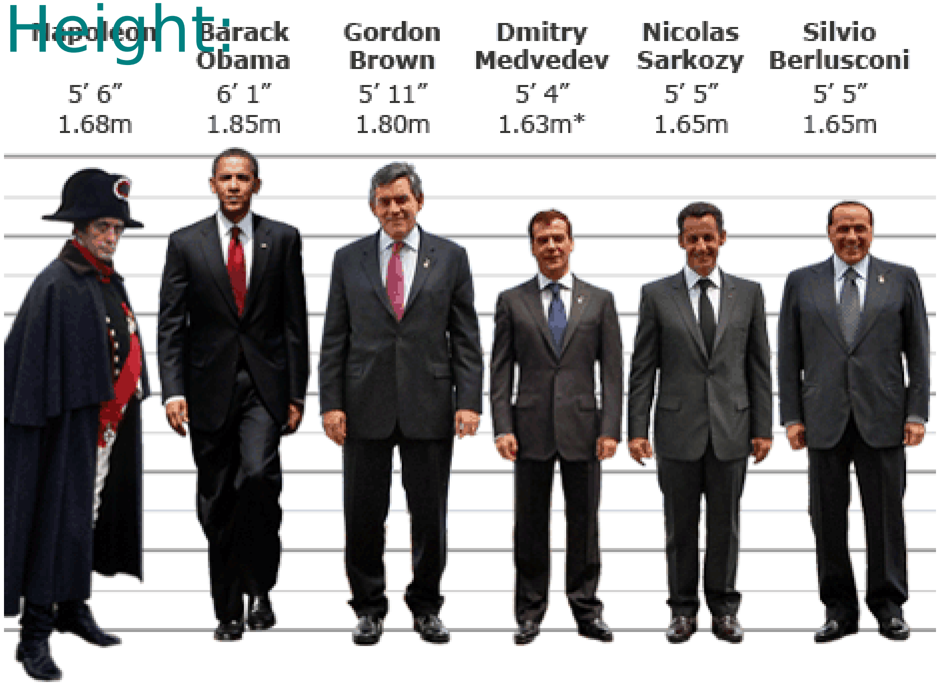
Yi Xu
yxu@cs.binghamton.edu
Computer Science
Binghamton University

BINGHAMTON
UNIVERSITY
STATE UNIVERSITY OF NEW YORK

What is High Dimensional data?

Low Dimensional Data

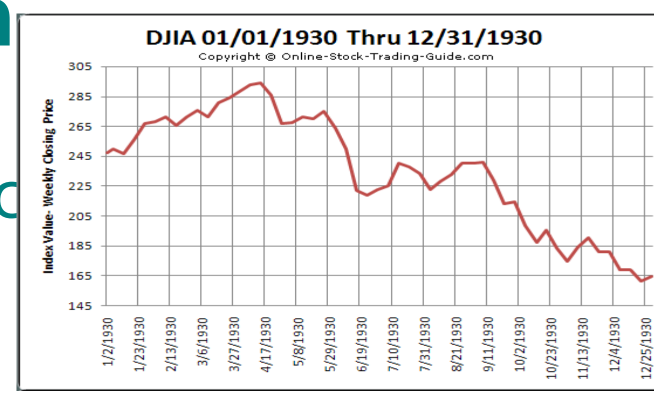
People's Height:



1 feature: meter

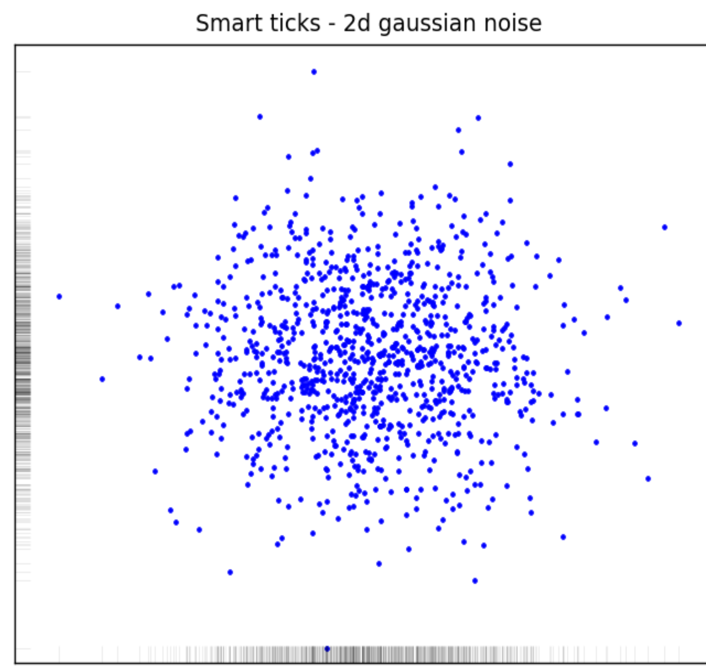
Magnitude Matters!

Stock Price:



1 feature: price

Gaussian Noise:



2 features: x and y

To view a data sample as a vector

Vector Space

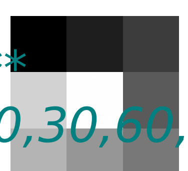
Constructed by

Ordered Features

Text*	
To	2
Be	2
Or	1
Not	1
That	1
Is	1
A	1
Question	1

Image**	
(1,1)	0
(1,2)	30
(1,3)	60
(2,1)	210
(2,2)	255
(2,3)	90
(3,1)	180
(3,2)	150
(3,3)	120

$*Hamlet = (2,2,1,1,1,1,1,1)^T$



$** = (0,30,60,210,255,90,180,150,120)^T$

Some High Dimensional Feature Space:

- Vocabulary of one month news from New York Times/Politics:

  >8000

- An image data set containing images with 600x800 resolution:

  =480,000

- Algae genome data set acquired using DNA microarray

  $>10^7$

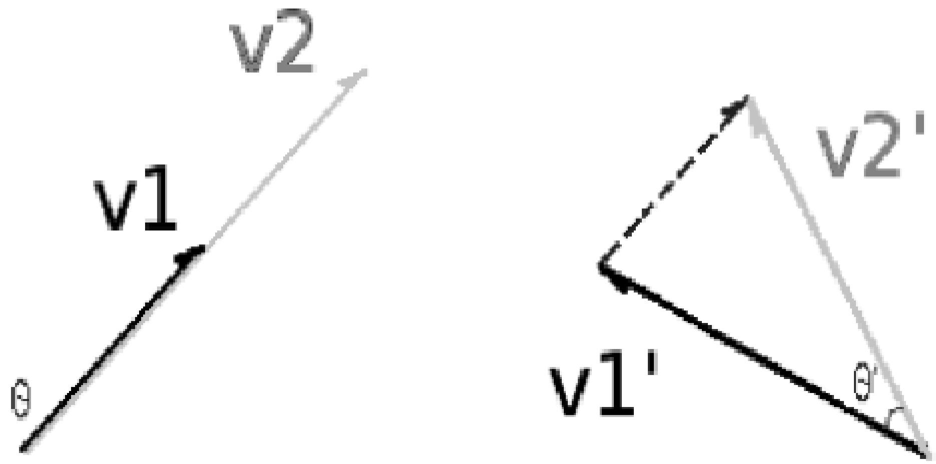
Magnitude matters?

Or something else?

Difference between two scalars is the difference of their *magnitude*:

$$|x_1 - x_2|$$

Difference between two vectors involves both the *magnitude* and the *direction*:



The Euclidean metric fails to differentiate the length difference from the direction difference

For high dimensional data, the rotation of a subspace is different from and usually more informative than the magnitude difference.

--- Breaking news – new combination of key words

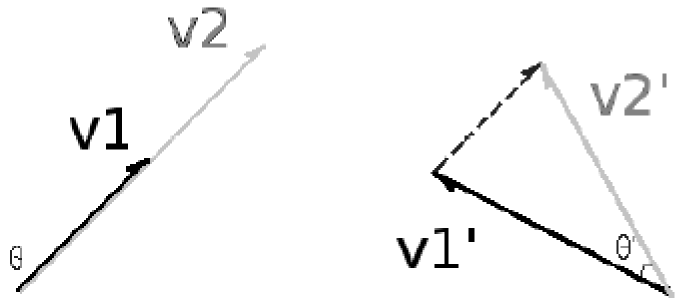
--- A baby v.s. An adult v.s. A monkey

What measurement is invariant under data’s magnitude change and
only characterize the rotation introduced by dimensionality?

~~Euclidean Distance~~

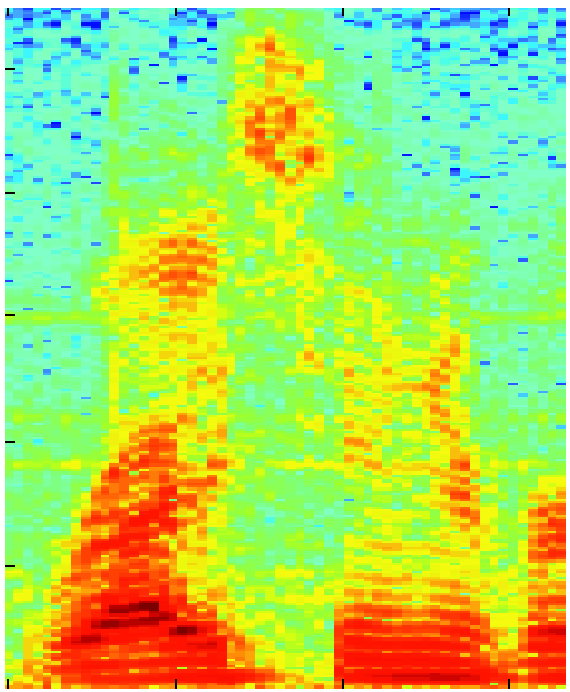
~~L-Norms~~

~~Bregman Divergenc~~

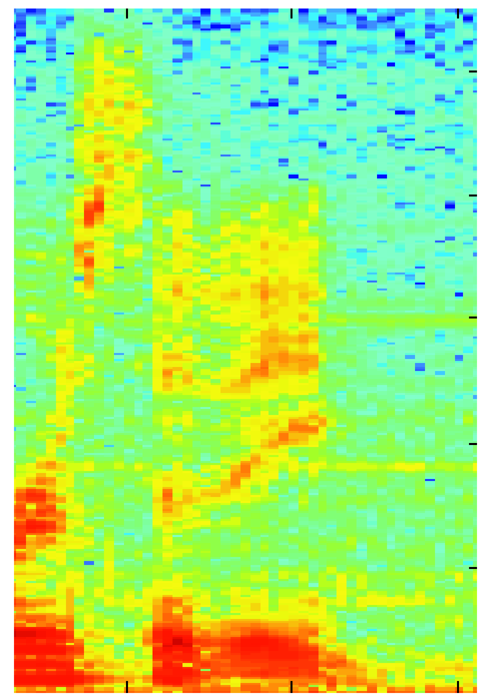


θ

How to compute subspace rotation (pattern change) between two vector sets?



$\{\mathbf{x}_i\}_{i=1}^{70}$



$\{\mathbf{y}_i\}_{i=1}^{50}$

Principal angles between two subspaces (Golub and Loan)

DEFINITION 1. *Let S_1 and S_2 be subspaces in $\mathbb{R}^n$ whose dimensions satisfy*

$$p = dim(S_1) \geq dim(S_2) = q \geq 1$$

The principal angles $\theta_k \in [0, \pi/2]$, $k = 1, \ldots, q$, *between* S_1 *and* S_2 *are defined recursively as*

$$\cos(\theta_k) = \max_{\mathbf{u} \in S_1, \mathbf{v} \in S_2} \mathbf{u}^T \mathbf{v} = \mathbf{u}_k^T \mathbf{v}_k$$

when $k = 1$, $\|\mathbf{u}_1\| = \|\mathbf{v}_1\| = 1$; *when* $k \geq 2$, $\|\mathbf{u}_k\| = \|\mathbf{v}_k\| = 1$; $\mathbf{u}_k^T \mathbf{u}_i = 0$; $\mathbf{v}_k^T \mathbf{v}_i = 0$ *where* $i = 1, \ldots, k-1$.

Principal angles between two subspaces (cont.)

- A generalization of the angle between two vectors
- A unique set of angles $\{\cos^{-1}\theta_i\}_{i=1}^{k}$ defined using orthonormal basis of two subspaces
- Invariant under an isomorphism and thus independent of the magnitude change.
- The largest angles reflect the pattern change between two data sets

- No statistical significance
- Sensitive to noise

Not practical to directly compute the largest principal angles.

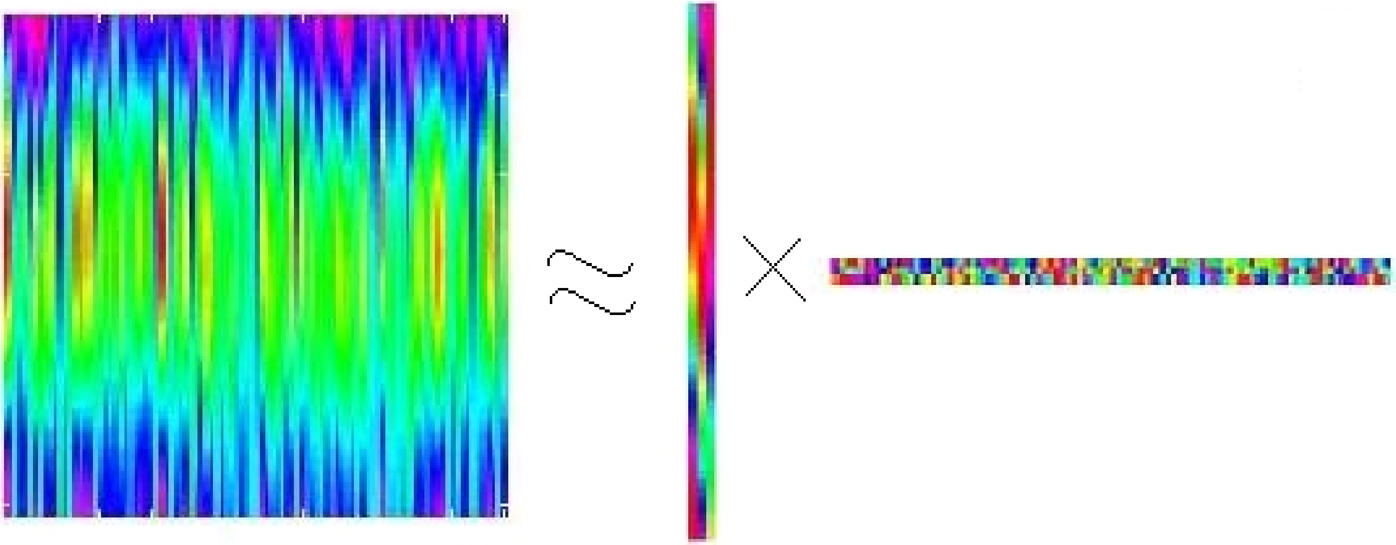
What we really want?

Principal angles between subspaces of high like

Prior information:

- Data
- The number of patterns k that we recognize from the data,usually based on common sense and widely acknowledged facts.
- ~~Label of the data (classification problem)~~

Pattern Summary via Matrix Factorization



$$X \approx PS^T$$

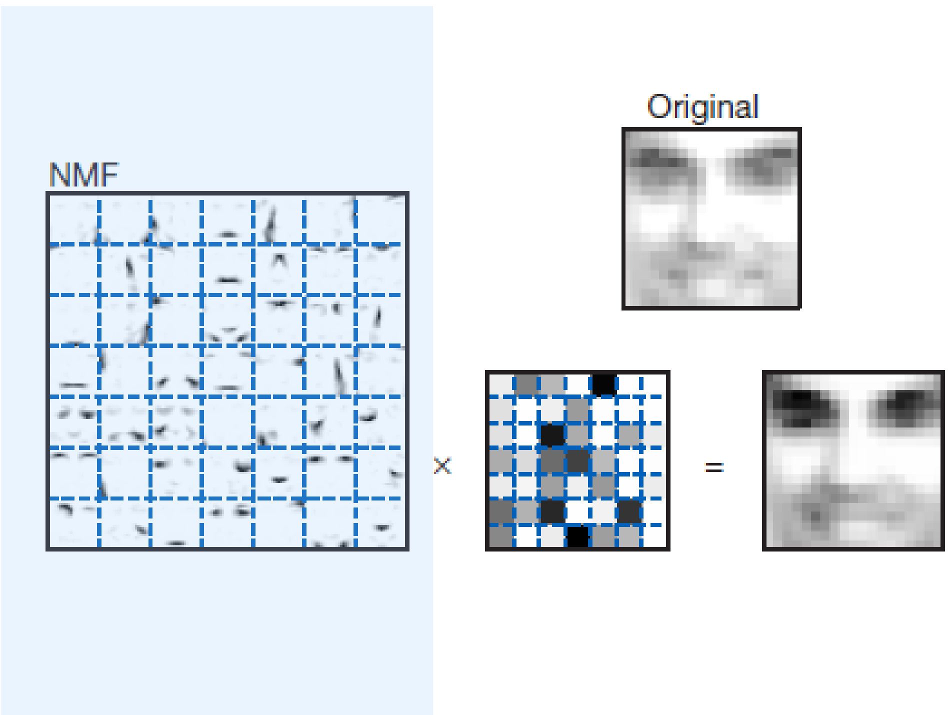
P --- Each column vector represents a summarized pattern from X

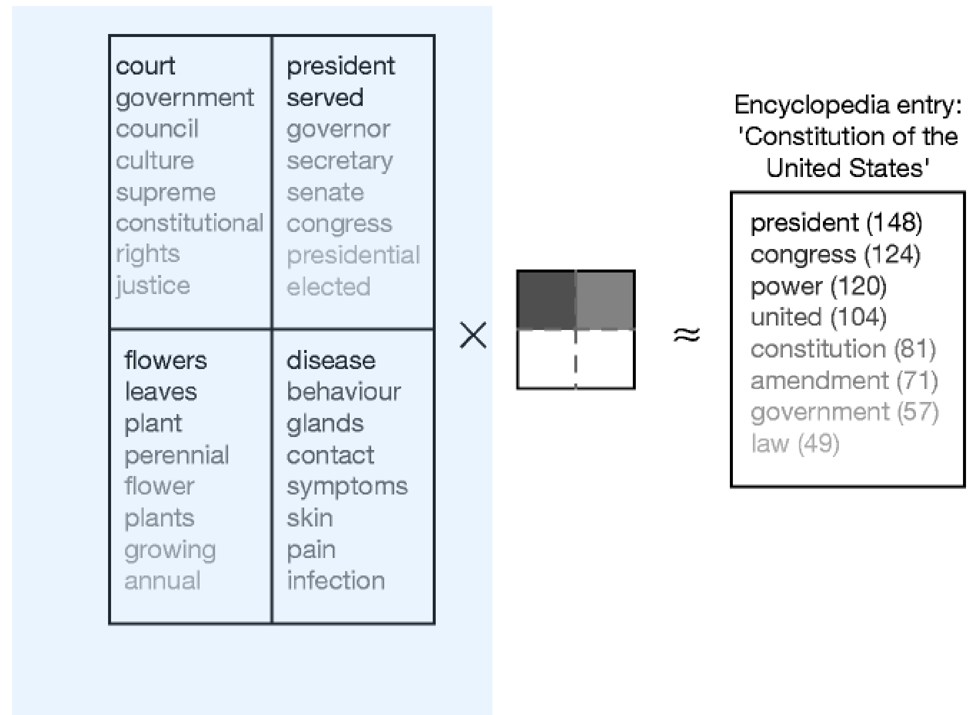
S --- Each row vector represents the weight on each pattern to

restore the corresponding sample.

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

Pattern Summary via Matrix Factorization (cont.)





P P

The subspace $span(P)$ is where we find principal an

Given two data sets X_1 and X_2

First compute $X_1 \approx P_1 S_1^T \quad X_2 \approx P_2 S_2^T$

Then compute principal angles between $span(P_1)$ and $span(P_2)$

Still does not know what is new in subspace $span($

Given two data sets X_1 and X_2 , how do we find $span(P_2)$, which has the largest principal angles from $span(P_1)$

Construct hypothesis test on principal angles betw

and $span\,(P_2)$.

Pattern Change Detection Based on Hypothesis Test

- Construct null-hypothesis using principal angles between $span(P_1)$

  and $span(P_2)$:

  $$H_o : \|diag(\{\cos \theta_i\}_{i=1}^{k})\| = 0$$

  to assume the largest principal angles in between.

- Ho has an equivalent form using only P_1 and P_2:

  $$H_o : P_1^T P_2 = 0$$

Pattern Change Detection Based on Hypothesis Test *(cont.)*

- Maximum likelihood estimation with hypothesis *Ho*: $P_1^T P_2 = 0$

$$\mathcal{L}(\mathbf{P}_2, \mathbf{S}_2) = \|\mathbf{X}_2 - \mathbf{P}_2\mathbf{S}_2^T\|^2 + \lambda\|\mathbf{P}_2^T\mathbf{P}_1\|^2$$

- Maximum likelihood estimation without hypothesis *Ho*:

$$\mathcal{L}(\mathbf{P}_2, \mathbf{S}_2) = \|\mathbf{X}_2 - \mathbf{P}_2\mathbf{S}_2^T\|^2$$

- Likelihood ratio test:

$$\Lambda = \frac{\|\mathbf{X}_2 - \hat{\mathbf{P}}\hat{\mathbf{S}}^T\|^2}{\|\mathbf{X}_2 - \hat{\mathbf{P}}_H\hat{\mathbf{S}}_H^T\|^2}$$

Reject *Ho* when $\Lambda < h$

Experiment summary

- Text data sets

  New combination of words – New topic

- Face image data sets

  New combination of pixels – New object

- Surveillance videos

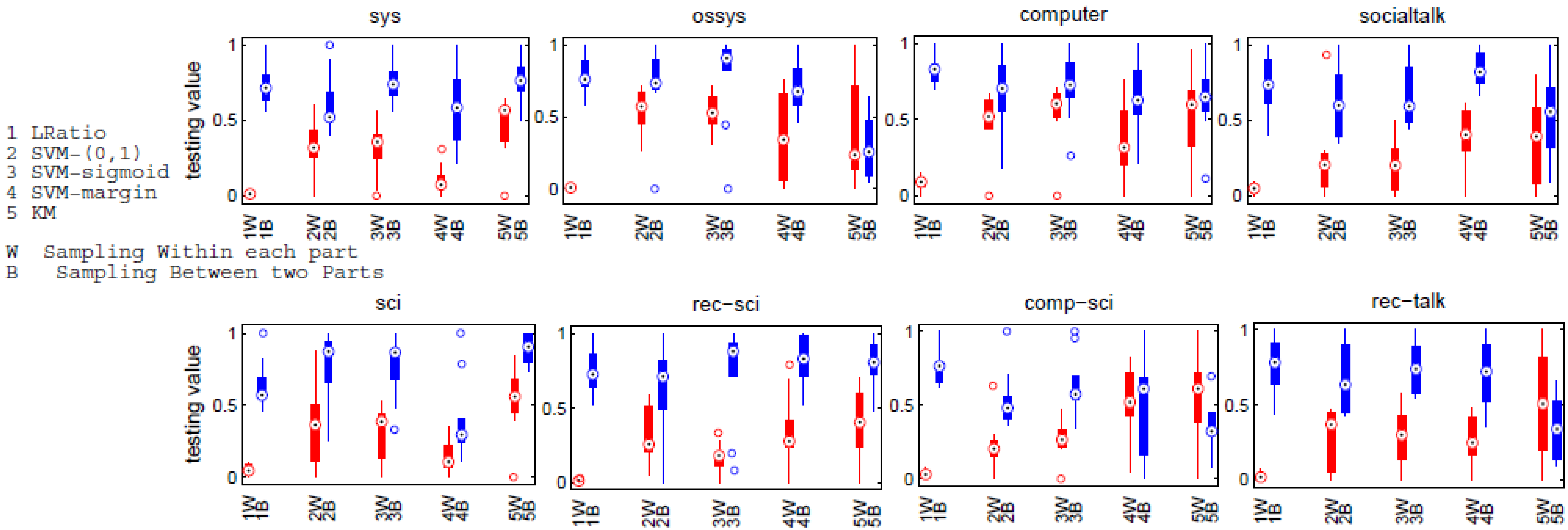
  New combination of motion vectors – New event

Experimental results: synthetic change

Name	Part 1	Part 2	Sample no.	Dim.
sys	comp.sys.ibm.pc	comp.sys.mac	400 × 2	1558
ossys	comp.os.ms-windows.misc, comp.sys.ibm.pc	comp.sys.mac comp.windows.x	200 × 4	2261
computer	comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc	comp.sys.ibm.pc, comp.sys.mac, sci.electronics	100 × 6	1606
socialtalk	talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc	alt.atheism, soc.religion.christian, talk.politics.misc, talk.religion.misc,	100 × 8	3312
sci	sci.crypt	sci.med	400 × 2	2870
rec-sci	rec.sport.baseball, rec.sport.hockey	sci.electronics, sci.space	200 × 4	2800
comp-sci	comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc	sci.electronics, sci.med, sci.space	100 × 6	1864
rec-talk	rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey	talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc	100 × 8	2992

The configuration of the pattern change data sets by using 20-news-group data sets.

Experimental results: synthetic change (*cont.*)



For each pair of red and blue bars, a smaller overlap in between indicates a better performance.

Experimental results: event detection from video

http://cs.binghamton.edu/~mrldata/Report

The Summary

- *Principal angles* to measure pattern change between high-dimensional data sets.
- *Matrix factorization* to summarize pattern space of high likelihood
- *Likelihood ratio test* based on linear model to unify the two tasks
- Experiments on text, images, and videos for justification.

Thank You!