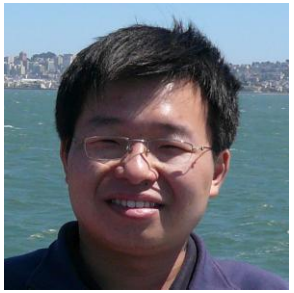
The 25th International Conference on Machine Learning (ICML) 2008, Helsinki, Finland

Dirichlet Component Analysis: Feature Extraction for Compositional Data



Hua-Yan Wang
Peking University

Qiang Yang
Hong Kong University of Science and Technology

Hong Qin
SUNY at Stony Brook

Hongbin Zha
Peking University

storyline

- **intro**
  - general concepts and background
- **a toy example**
  - how is our approach motivated
- **DCA**
  - how does it work
- **experiment results**
  - synthetic and real-world datasets

storyline

- **intro**
  - general concepts and background
- **a toy example**
  - how is our approach motivated
- **DCA**
  - how does it work
- **experiment results**
  - synthetic and real-world datasets

intro

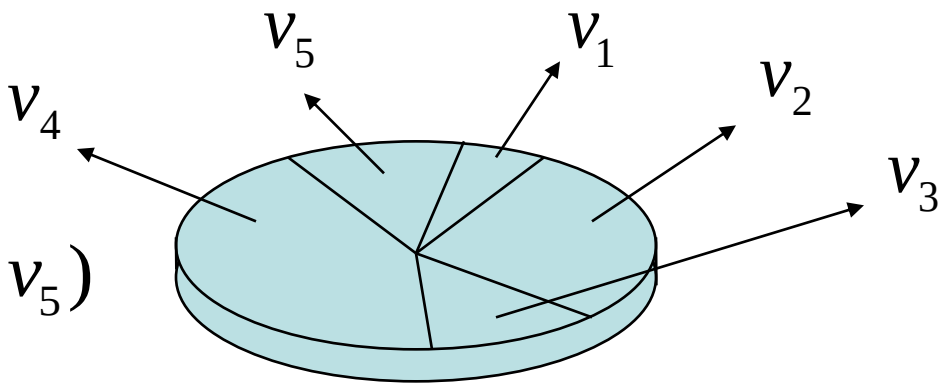
- Feature extraction (dimensionality reduction) is useful in many aspects
  - avoid *over-fitting* of classification / regression models
  - improve *domain understanding*
  - reduce *computational expense* of subsequent processing
  - facilitate *visualization* of high-D datasets

intro

- We investigate *feature extraction* for *compositional data*.

- *compositional data*: normalized histograms, representing relative proportion of different ingredients in an object

$$\vec{v} = (v_1, v_2, v_3, v_4, v_5)$$



positive, constant-sum, real vectors

points in a simplex

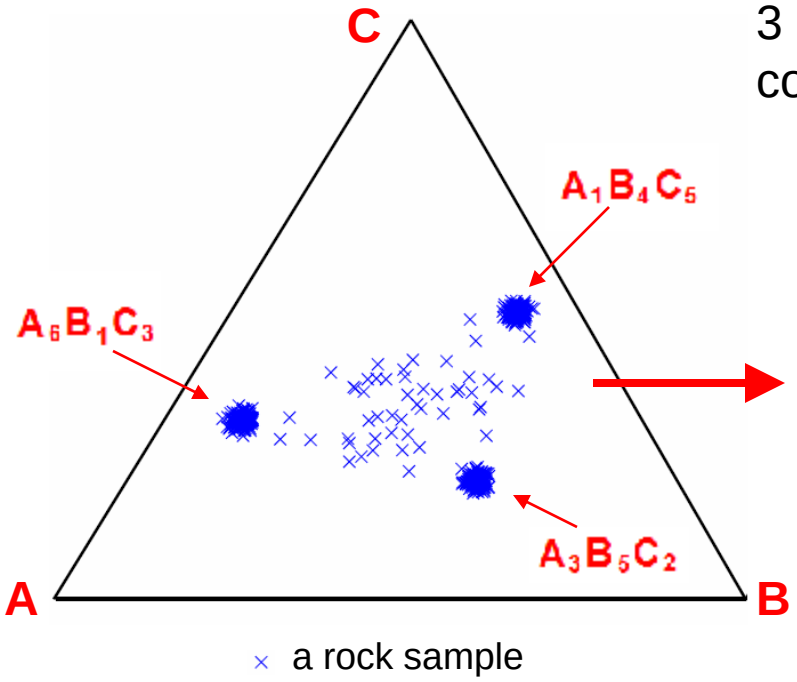
storyline

- **intro**
  - general concepts and background
- **a toy example**
  - how is our approach motivated
- **DCA**
  - how does it work
- **experiment results**
  - synthetic and real-world datasets

a toy example

- Suppose we have some rock samples collected.

- In lab, these samples are decomposed by some chemical approach, and we record *relative proportions* of 3 major elements: **A**, **B**, and **C** in each sample.

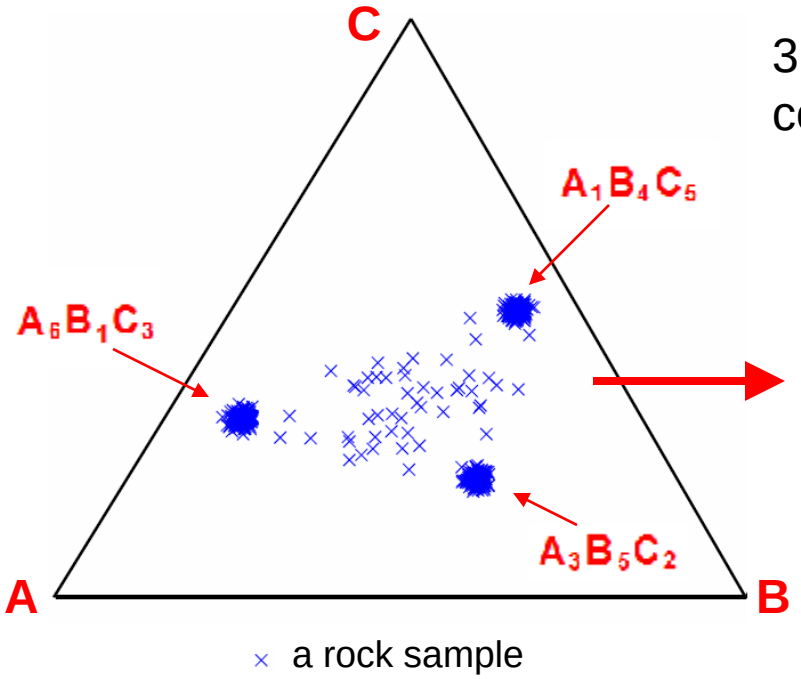


3 peaks : 3 *substances* that have fixed compositions in terms of **A**, **B**, and **C**.

1	2	3
0.0954	0.4076	0.4970
0.0983	0.4037	0.4980
0.0888	0.4130	0.4982
0.5967	0.1016	0.3017
0.5979	0.1026	0.2995
0.6288	0.0689	0.3023
0.2870	0.5009	0.2121
0.2916	0.4914	0.2170
0.2946	0.5167	0.1886
.	.	.
.	.	.
.	.	.

The major patterns (peaks) are explained by *linear combinations* of the variables (features).

a toy example



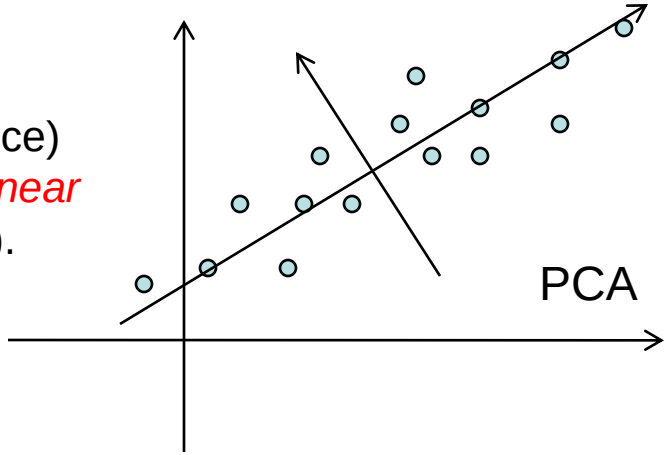
× a rock sample

3 peaks : 3 *substances* that have fixed compositions in terms of **A**, **B**, and **C**.

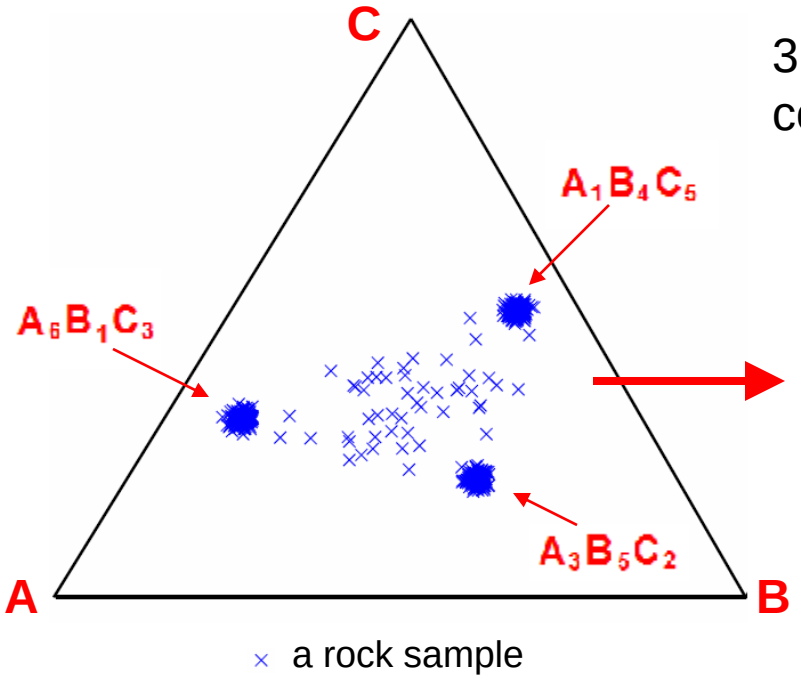
1	2	3
0.0954	0.4076	0.4970
0.0983	0.4037	0.4980
0.0888	0.4130	0.4982
0.5967	0.1016	0.3017
0.5979	0.1026	0.2995
0.6288	0.0689	0.3023
0.2870	0.5009	0.2121
0.2916	0.4914	0.2170
0.2946	0.5167	0.1886
⋮	⋮	⋮

The major patterns (peaks) are explained by *linear combinations* of the variables (features).

In PCA, we try to explain the major patterns (variance) *separately by individual variables*, instead of their *linear combinations*. (diagonalizing the covariance matrix).



a toy example



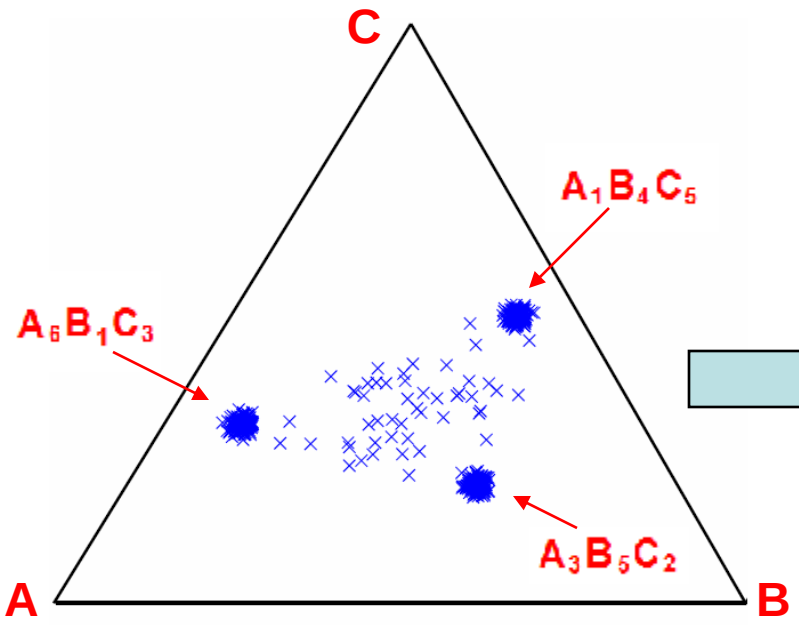
3 peaks : 3 *substances* that have fixed compositions in terms of **A**, **B**, and **C**.

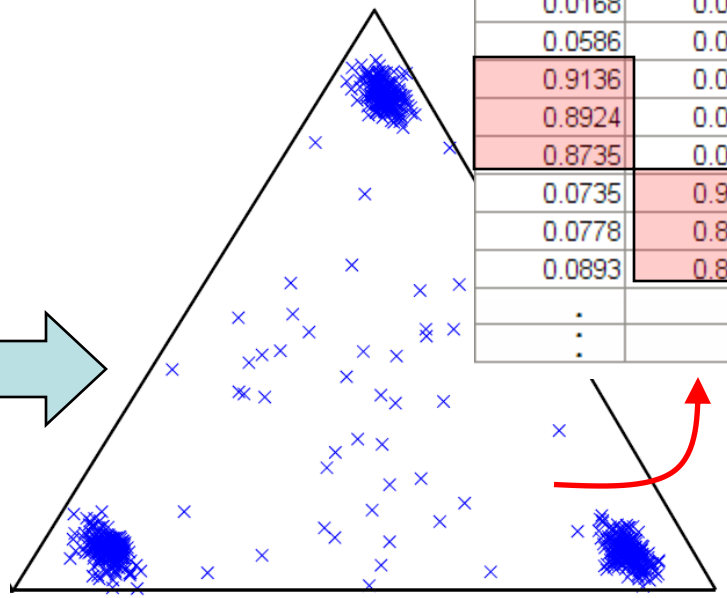
1	2	3
0.0954	0.4076	0.4970
0.0983	0.4037	0.4980
0.0888	0.4130	0.4982
0.5967	0.1016	0.3017
0.5979	0.1026	0.2995
0.6288	0.0689	0.3023
0.2870	0.5009	0.2121
0.2916	0.4914	0.2170
0.2946	0.5167	0.1886
.	.	.
.	.	.
.	.	.

The major patterns (peaks) are explained by *linear combinations* of the variables (features).

Analogously, is it possible to find a new representation for this toy example, in which the major patterns (peaks) are explained *separately by individual variables* instead of their *linear combinations* ?

a toy example



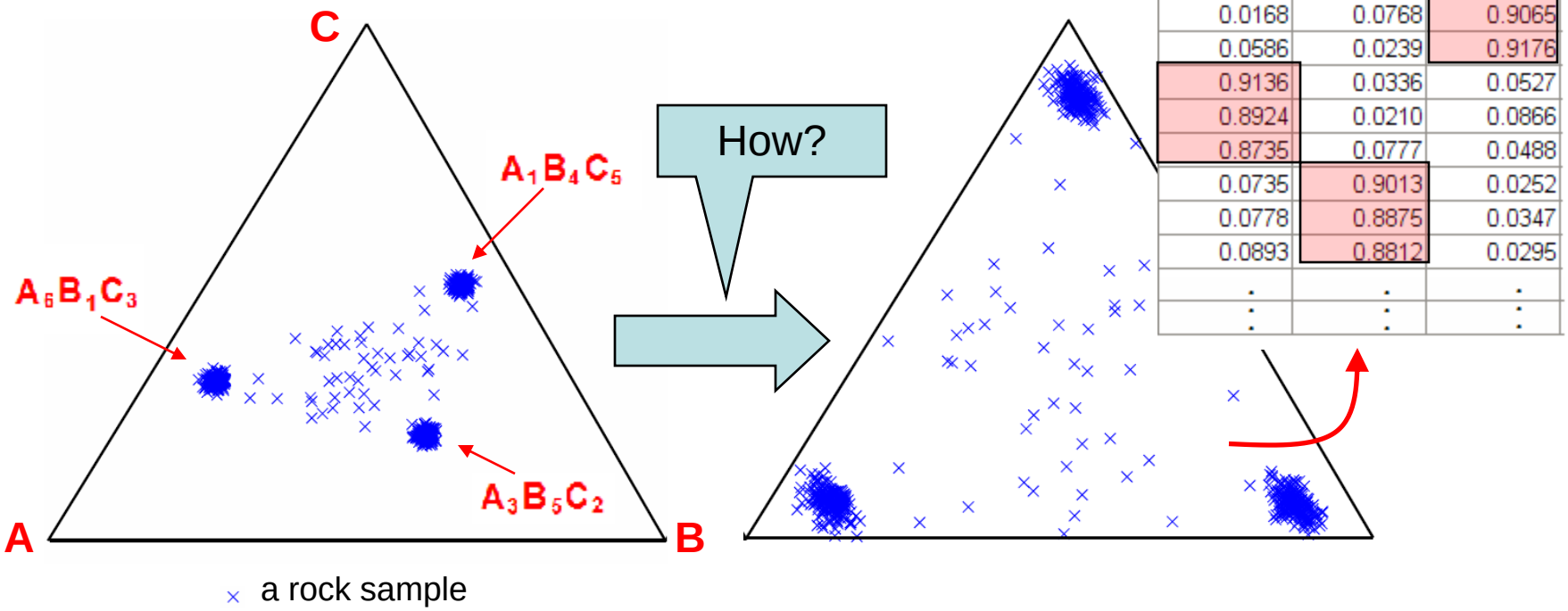


1	2	3
0.0538	0.0406	0.9056
0.0168	0.0768	0.9065
0.0586	0.0239	0.9176
0.9136	0.0336	0.0527
0.8924	0.0210	0.0866
0.8735	0.0777	0.0488
0.0735	0.9013	0.0252
0.0778	0.8875	0.0347
0.0893	0.8812	0.0295
⋮	⋮	⋮

× a rock sample

Analogously, is it possible to find a new representation for this toy example, in which the major patterns (peaks) are explained *separately by individual variables* instead of their *linear combinations* ?

a toy example



1	2	3
0.0538	0.0406	0.9056
0.0168	0.0768	0.9065
0.0586	0.0239	0.9176
0.9136	0.0336	0.0527
0.8924	0.0210	0.0866
0.8735	0.0777	0.0488
0.0735	0.9013	0.0252
0.0778	0.8875	0.0347
0.0893	0.8812	0.0295
.	.	.
.	.	.

Sometimes we need to extract features for compositions,

and the new features *also* have a natural interpretation as compositions.

That is, extract *new compositions* from *old compositions*.

storyline

- **intro**
  - general concepts and background
- **a toy example**
  - how is our approach motivated
- **DCA**
  - how does it work
- **experiment results**
  - synthetic and real-world datasets

DCA

- Such projections could be viewed as *rearranging* mass from the N original *components* to the K new *components*, while the law of conservation of mass is satisfied. Hence we refer to such linear transforms from $\mathbb{S}^N$ to $\mathbb{S}^K$ as *rearrangements.*

nents

$$i=1 \qquad (1)$$

- the family of linear projecti hat preserve the simplex constraint

Proposition 1 *For linear projections*

$$\mathbf{y} = \mathbf{R}\,\mathbf{x} \quad \text{where} \quad \mathbf{R} = (r_{ij})_{K\times N} \qquad (2)$$

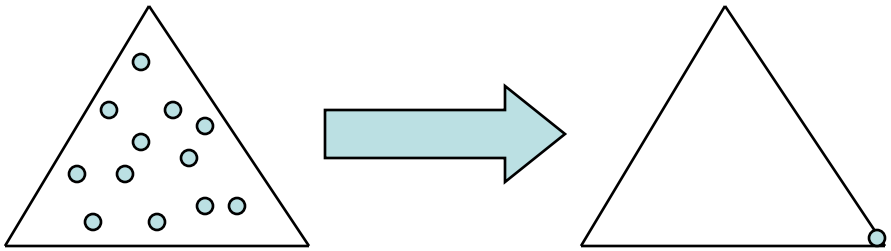
$\mathbf{y}$ *is in* $\mathbb{S}^K$ *for all* $\mathbf{x}$ *in* $\mathbb{S}^N$ ***if and only if***

1) $r_{ij} \geq 0$ *for all* i, j.

2) $\sum_{i=1}^{K} r_{ij} = 1$, *for* $j = 1, \cdots, N$.

DCA

- To avoid degenerate cases, such as:



$$R = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- we further require the rows of the projection matrix being constant-sum:

Definition 1 (Balanced Rearrangement) *A linear projection* $\mathbf{R}\,\mathbf{x} = \mathbf{y}$ *is a* ***balanced rearrangement****, if* $\mathbf{R} = (r_{ij})_{K \times N}$ *satisfies:*

1) $r_{ij} \geq 0$ *for all* i, j.

2) $\sum_{i=1}^{K} r_{ij} = 1$, *for* $j = 1, \cdots, N$.

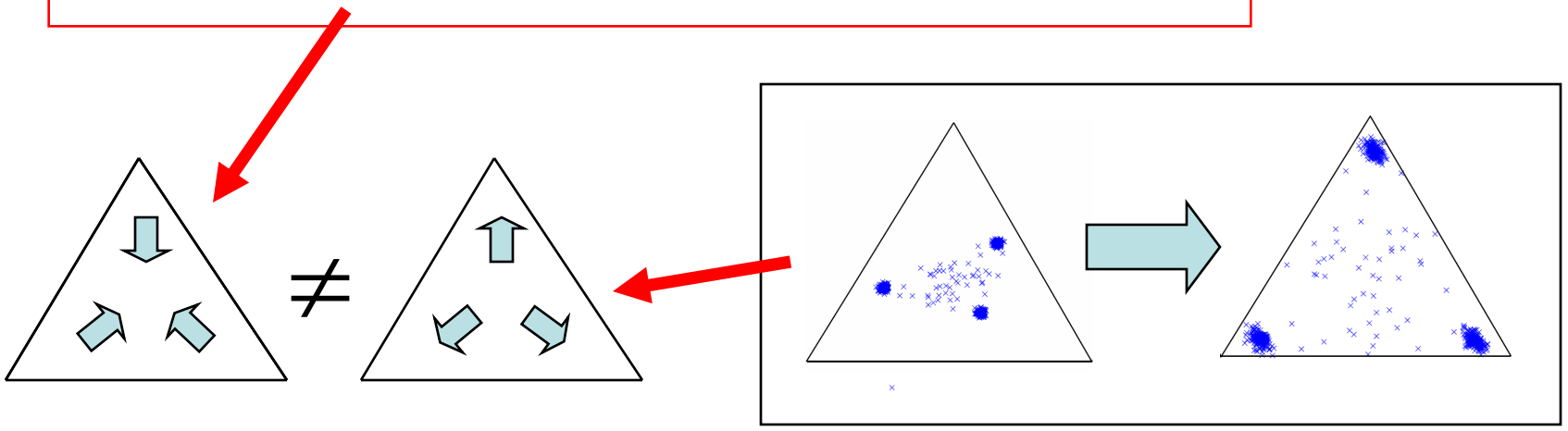
3) $\sum_{j=1}^{N} r_{ij} = N/K$, *for* $i = 1, \cdots, K$.

DCA

- So far, we've identified the family of *simplex-to-simplex non-degenerate linear* projections.

- However, such projection has an awkward property due to the simplex constraint:

Proposition 4 *Let* $\min(\mathbf{x})$ *be the minimum component of* $\mathbf{x}$. $\mathbf{R}_{K\times N}$ *is a balanced rearrangement matrix with* $K \leq N$, *then* $\min(\mathbf{Rx}) \geq \min(\mathbf{x})$ *for all* $\mathbf{x}$ *in* $\mathbb{S}^N$.



- So … r this effect:

Definition 2 (Regularization) *Given a compositional dataset* $\mathcal{X} = \{\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^M\}$, *a* ***regularization*** *on the dataset is denoted as:* $\widetilde{\mathcal{X}} = \{\widetilde{\mathbf{x}^1}, \widetilde{\mathbf{x}^2}, \cdots, \widetilde{\mathbf{x}^M}\}$, *where* $\widetilde{\mathbf{x}^i} = \frac{1}{\sum_{j=1}^{N}(x_j^i - \delta)}(x_1^i - \delta, x_2^i - \delta, \cdots, x_N^i - \delta,)$ *for* $i = 1, 2, \cdots, M$, *and the* ***regularization factor*** $\delta = \min(\min(\mathbf{x^1}), \min(\mathbf{x^2}), \cdots, \min(\mathbf{x^M}))$.

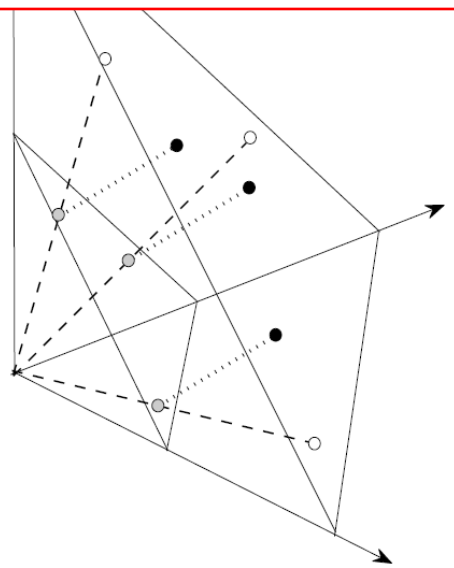
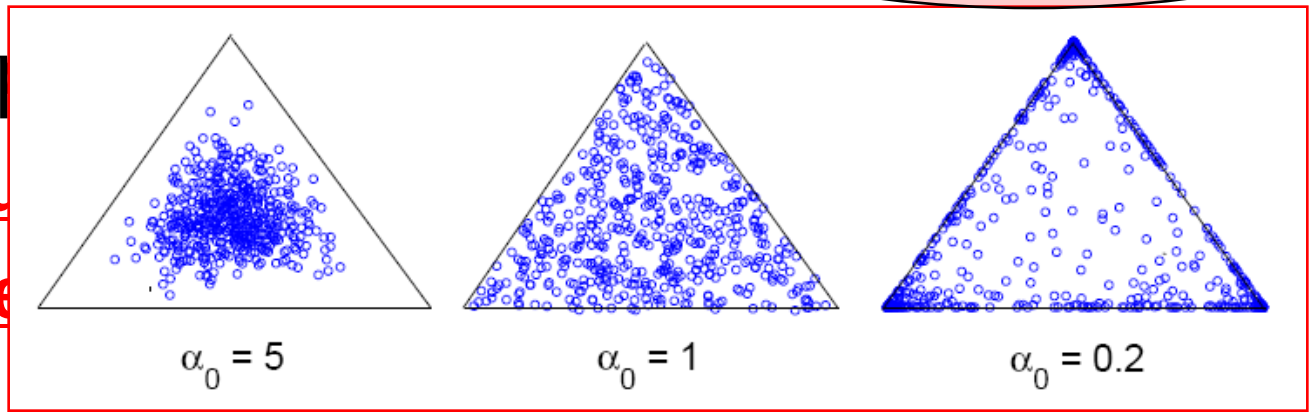


Figure 1. Regularization of compositional data points (black) is performed by parallel projection to the gray points, then radial projection to the white points.

DCA

$$\mathrm{Dir}(\mathbf{x} \mid \alpha_0) = \frac{\Gamma(N\alpha_0)}{\Gamma(\alpha_0)^N} \prod_{i=1}^{N} x_i^{\alpha_0 - 1}$$

- Princi
  - **Solu**
  - **Obje**



- Dirichlet Componen lysis (**DCA**)
  - **Solution space**: bal rearrangements
  - **Objective**: empirical Dirichlet precision

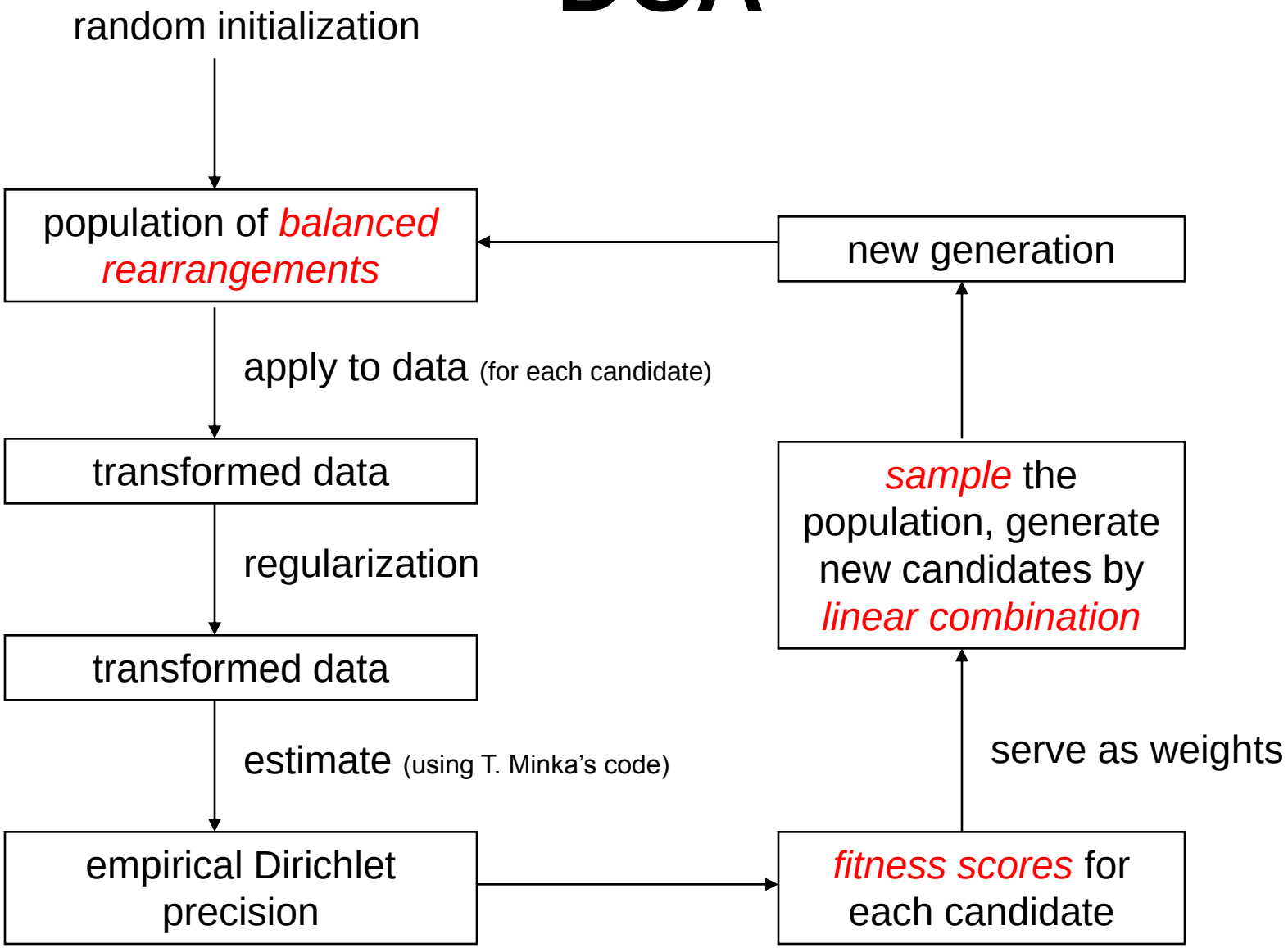
DCA

- Dirichlet component analysis:
  - Find the *balanced rearrangement*, which, when applied to data together with a *regularization operator*, minimizes the empirical *Dirichlet precision*.

- optimization: no obvious efficient solution due to…
  - the simplex constraint
  - the regularization operator

  - Our current implementation is based on the *genetic algorithm*.

DCA


random initialization
population of *balanced rearrangements*
apply to data (for each candidate)
transformed data
regularization
transformed data
estimate (using T. Minka's code)
empirical Dirichlet precision
fitness scores for each candidate
serve as weights
sample the population, generate new candidates by *linear combination*
new generation

storyline

- **intro**
  - general concepts and background
- **a toy example**
  - how is our approach motivated
- **DCA**
  - how does it work
- **experiment results**
  - synthetic and real-world datasets

experiment results (synthetic data)

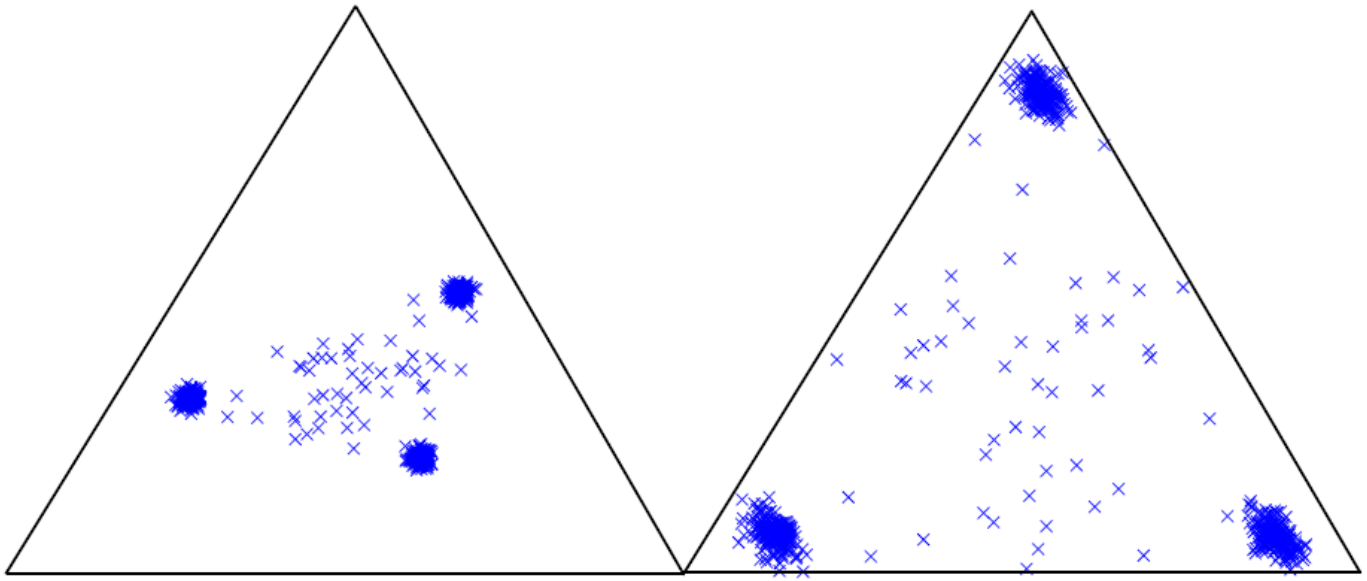


Figure 3. Left: synthetic data, composition of rock samples (small 'x') in terms of the old *components*. Right: representation in terms of new *components* (right) obtained by optimization algorithm discussed in Section 3.

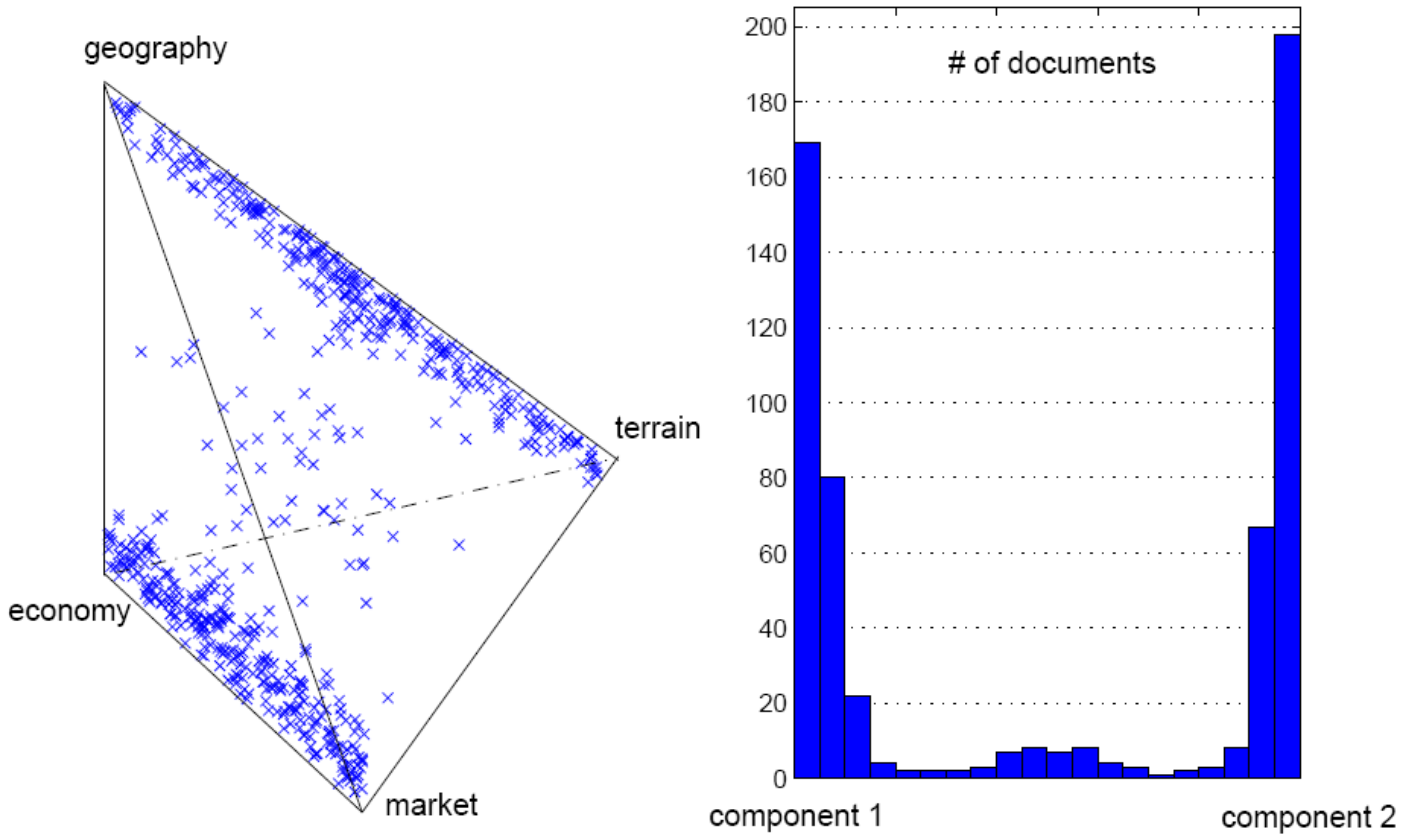
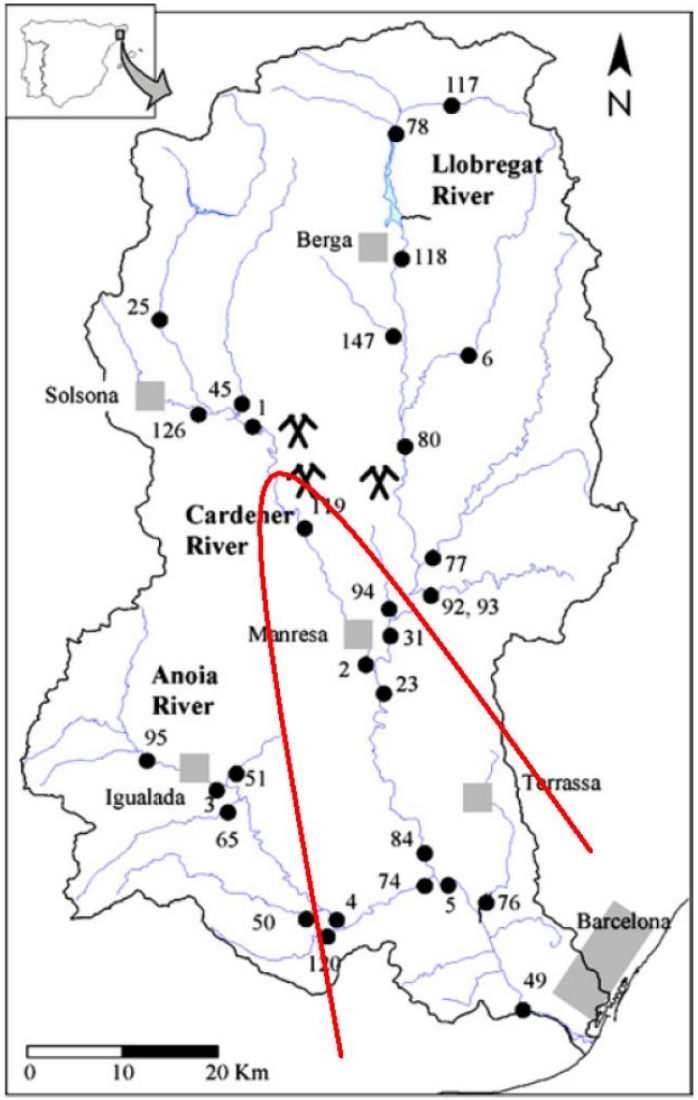


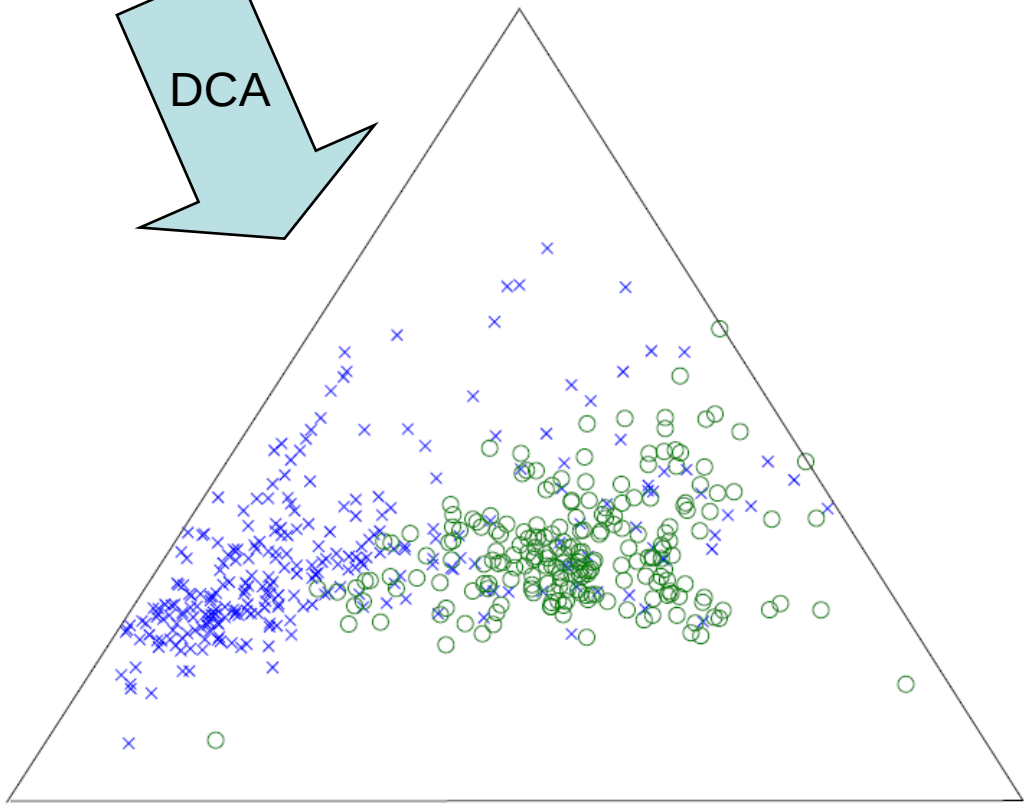
Figure 4. Left: Term frequencies of four words on $\mathbb{S}^4$, each small 'x' denotes a document. The data are synthetic. Right: new representation obtained by the optimization algorithm discussed in Section 3. The histogram illustrates the distribution of documents on $\mathbb{S}^2$.

experiment results (real-world data)



485 samples
14 dimensional
concentrations of major ions
(e.g. H^+, Na^+, NH_4^+, Cl^-, HCO_3^-, etc.)

DCA



experiment results (real-world data)

- bag-of-words data (20 newsgroup dataset)
  - validate the effect of our method in *avoiding over-fitting* of classification models (we use linear SVM), especially when the training set is extremely small

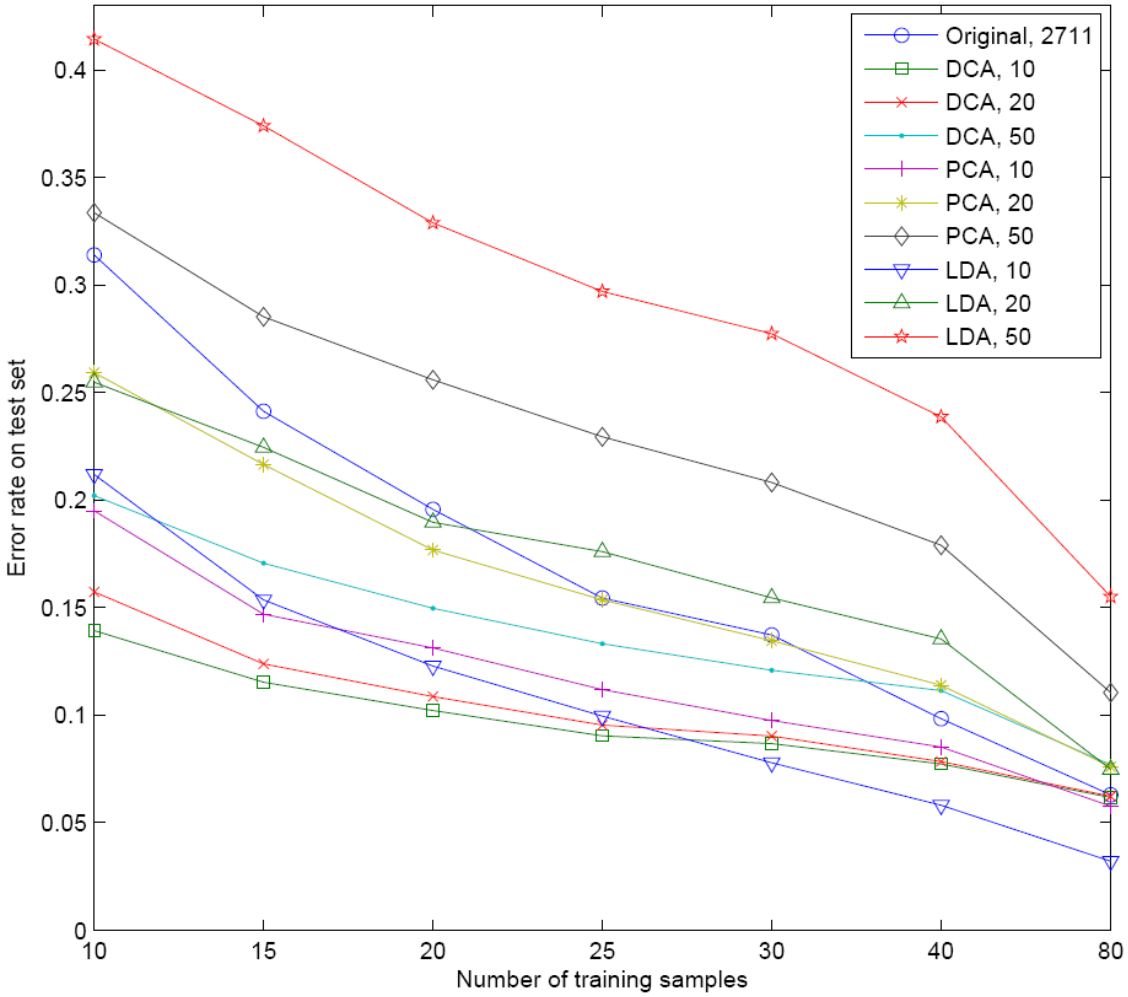


Figure 7. The "alt" versus "misc" classification performances using linear SVM. Representations with different dimensionality obtained by different methods are compared. E.g., "LDA, 10" indicates 10 dimensional representation obtained by latent Dirichlet allocation.

Thanks!

after lunch  …

S5 (3rd floor) 2:00 ~ 2:25 pm

In “multiple instance learning and learning with missing features, categorical features”

coming up…

Adaptive *p*-Posterior Mixture Model Kernels for Multiple Instance Learning

Hua-Yan Wang
Peking University

Qiang Yang
Hong Kong University of Science and Technology

Hongbin Zha
Peking University