Biologically-aware Latent Dirichlet Allocation (BaLDA) for the classification of expression microarray data

A. Perina[1], P. Lovato[1], V. Murino[1,3], M. Bicego[1,3]

[1] *University of Verona (Italy)*
[2] *Istituto Italiano di Tecnologia (Italy)*

5th IAPR Int. Conf. on Pattern Recognition in Bioinformatics, 22-24 September 2010

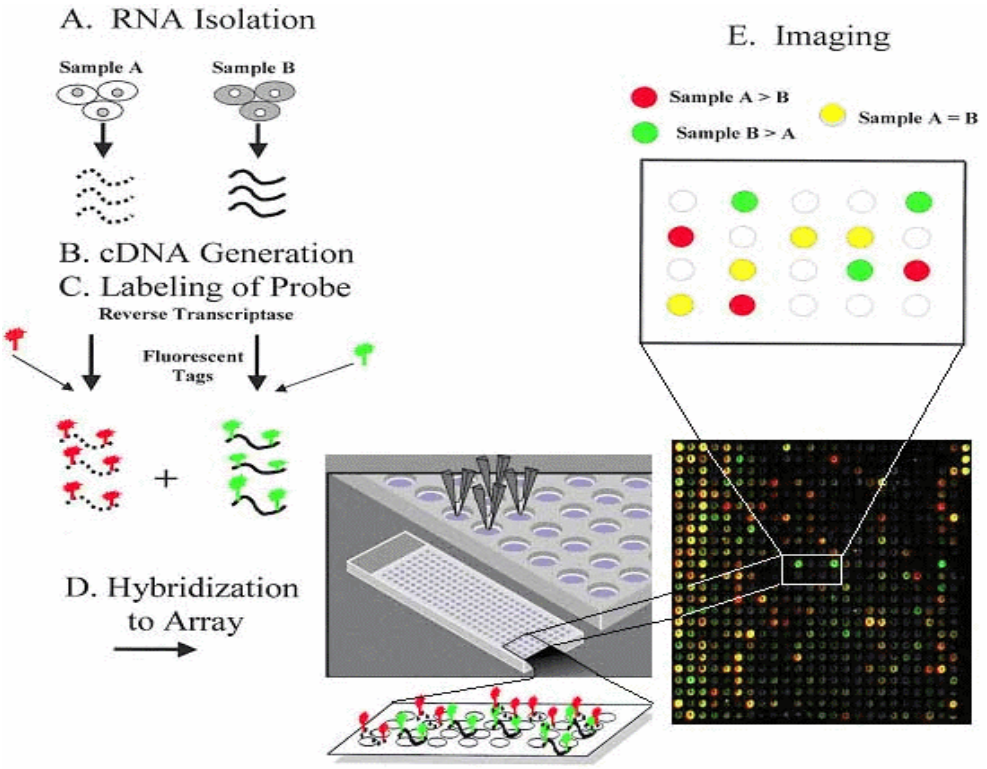
Summary

- Introduction: expression microarray data analysis
- Background: topic models
  - topic models & microarray
- The BaLDA model
- Some preliminary results
- Conclusions

Introduction: microarrays

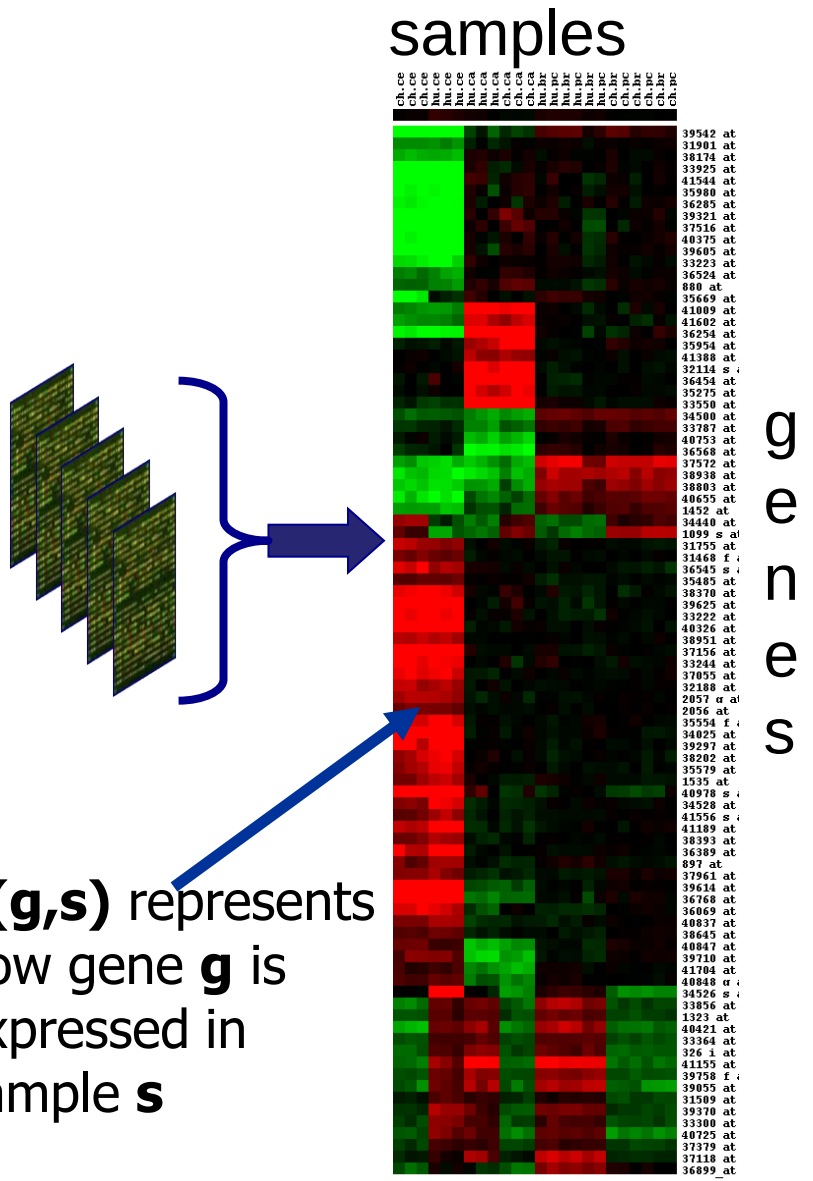
- Microarray: technology able to simultaneously analyze thousands of genes

Expression microarrays: they measure the expression levels of the different genes



Introduction: microarrays

- Different samples are analysed:
  - different subjects (healthy/diseased people)
  - different growth conditions
  - different development steps
- Result:
  - expression matrix **e(g,s)**



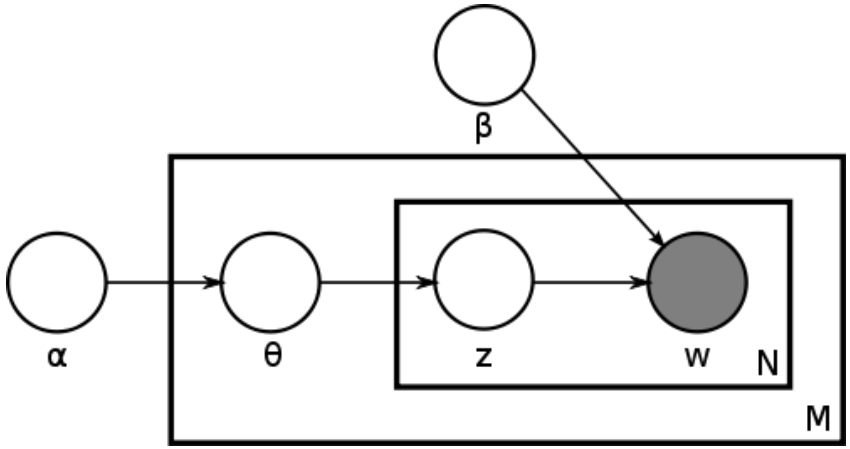
Microarray data analysis

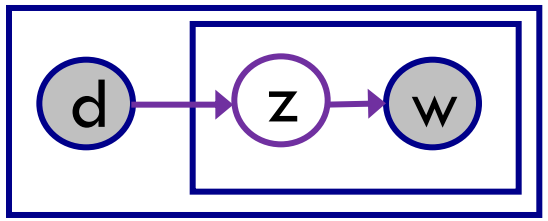
- Many Pattern Recognition problems in microarray data analysis
- Starting from the expression matrix
  - Gene selection
  - Clustering (biclustering)
  - Classification of samples

The problem is faced with a novel topic model

Background: topic models

- Probabilistic tools widely used in text analysis and computer vision communities





- They can model a dataset in terms of hidden topics (processes)

  ... which are highly interpretable

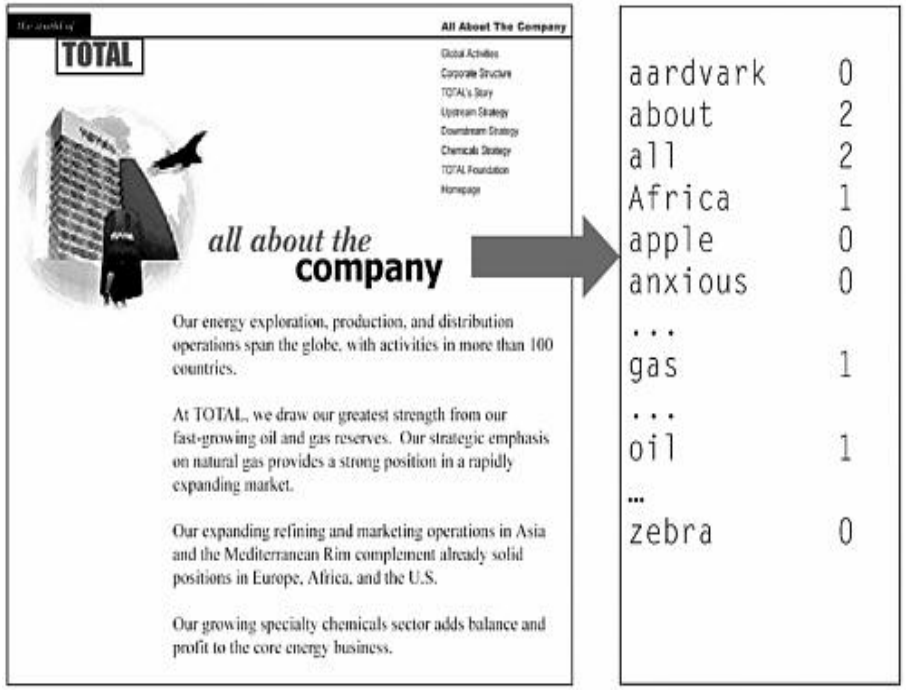
Background: topic models

(From the text analysis point of view)

- In topic models a document is seen as an unordered collection of words

A document is characterized by a "word counting vector"

(how many times every word is found in the document)



Bacground: topic models

- Observation: a single word can have different meanings depending on the context

"Home"	"sports"	"space"	"computers"	"weather"
Kitchen	Team	Space	Drive	Rain
Door	Game	Sun	Windows	Snow
Garden	Play	Research	Card	Sun
Windows	Year	Center	DOS	Season
Bedroom	Games	Earth	SCSI	Weekend
Space	Season	NASA	Sun	Cloudy

Topic models

- Topic models solve this problem
  - words can be disambiguated by looking at the context

- Topic models introduce an intermediate level, based on the concept of **topic**
  - it represents “what we are talking about”
  - the topics are extracted looking at co-occurrence of words in documents

Topic models

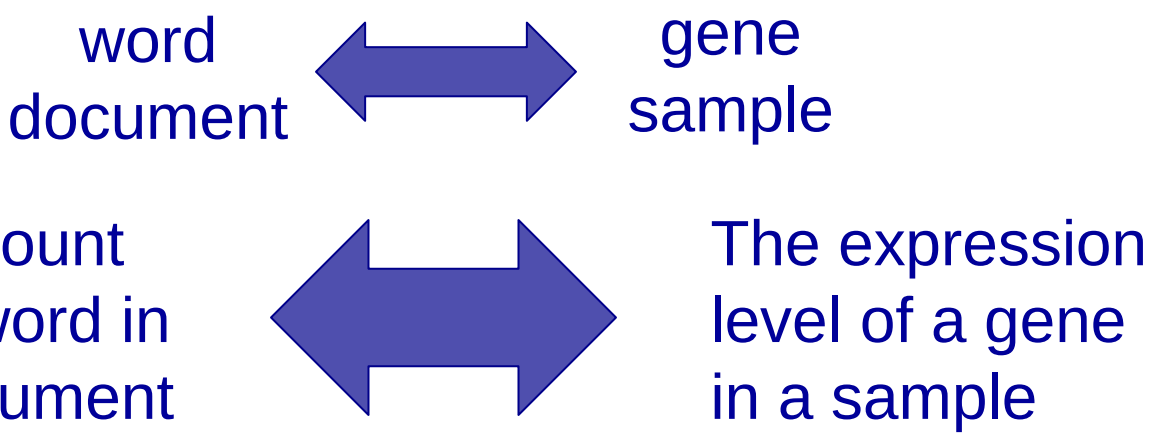
Summarizing:

- Every document is characterized by the presence of one or more topics (e.g. sport, finance, politics)

  .... which may induce the presence of some words

Topic models and Microarray

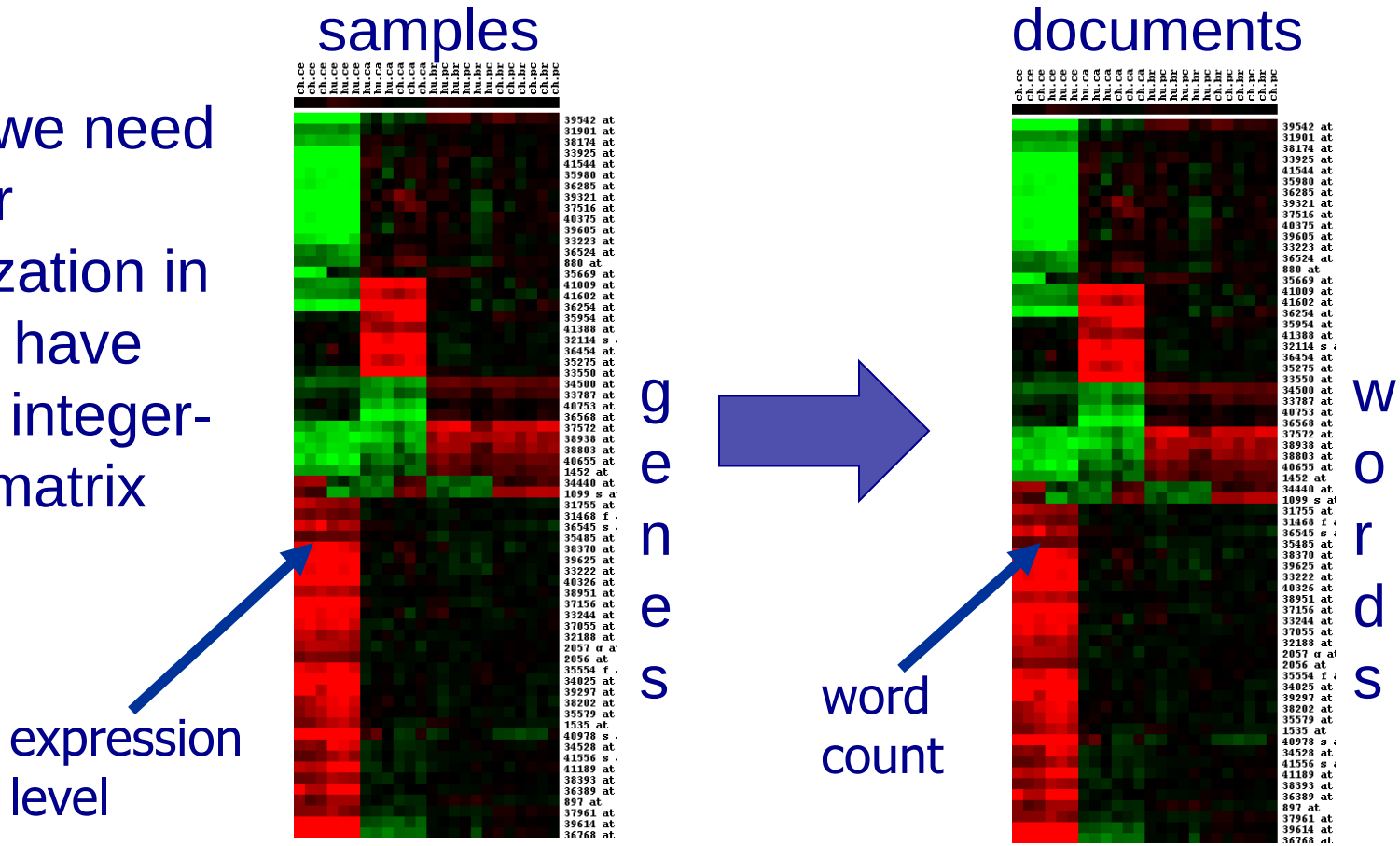
- We set an analogy between the text analysis and the microarray scenarios:
  - a document is characterized by the different presence of the words
  - a sample is characterized by the different expression level of the genes



Topic models and Microarray

We can apply TM to microarray considering the expression matrix as the counting matrix

NOTE: we need a proper normalization in order to have positive integer-valued matrix



Topic models and Microarray

- Main feature: interpretability!
  - a topic may be easily associated to a biological process
  - which may be active only in some samples
  - which may involve only some genes
- Already used with microarray data:
  - clustering and biclustering [Rogers et al. 2005] [Ying et al. 2008] [Masada et al. 2009] [Bicego et al. 2010]
  - classification [Bicego et al. 2010]

The proposed approach

- Problem of topic models with microarrays:
  - Each gene expression is *independently generated* given its corresponding topic
  - *“genes are independent”*: not true in biology!
- Here we propose a new topic model (BaLDA) which can integrate in the model relations between genes
  - the dependencies may be extracted from a priori information or different sources

BaLDA

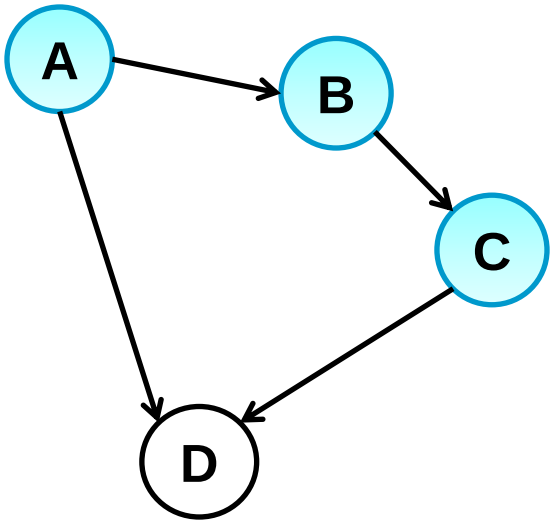
- The name BaLDA:
  - **LDA**: *Latent Dirichlet Allocation*: a topic model already used in the microarray scenario [Blei et al., 2003]
    - another version: *Latent Process Decomposition* [Rogers et al. 2005]
  - ***Ba**: Biologically aware:* (possibly biological) a priori information is taken into account in the model construction

BaLDA

- The form of the a priori information we are exploiting
  - subdivision of genes in groups (i.e. clustering of genes)
    - known relations between genes
  - it may be also computed, on the basis of different a priori information
    - spatial proximity, sequence similarity, ...

BaLDA: notation

- BaLDA is a graphical model



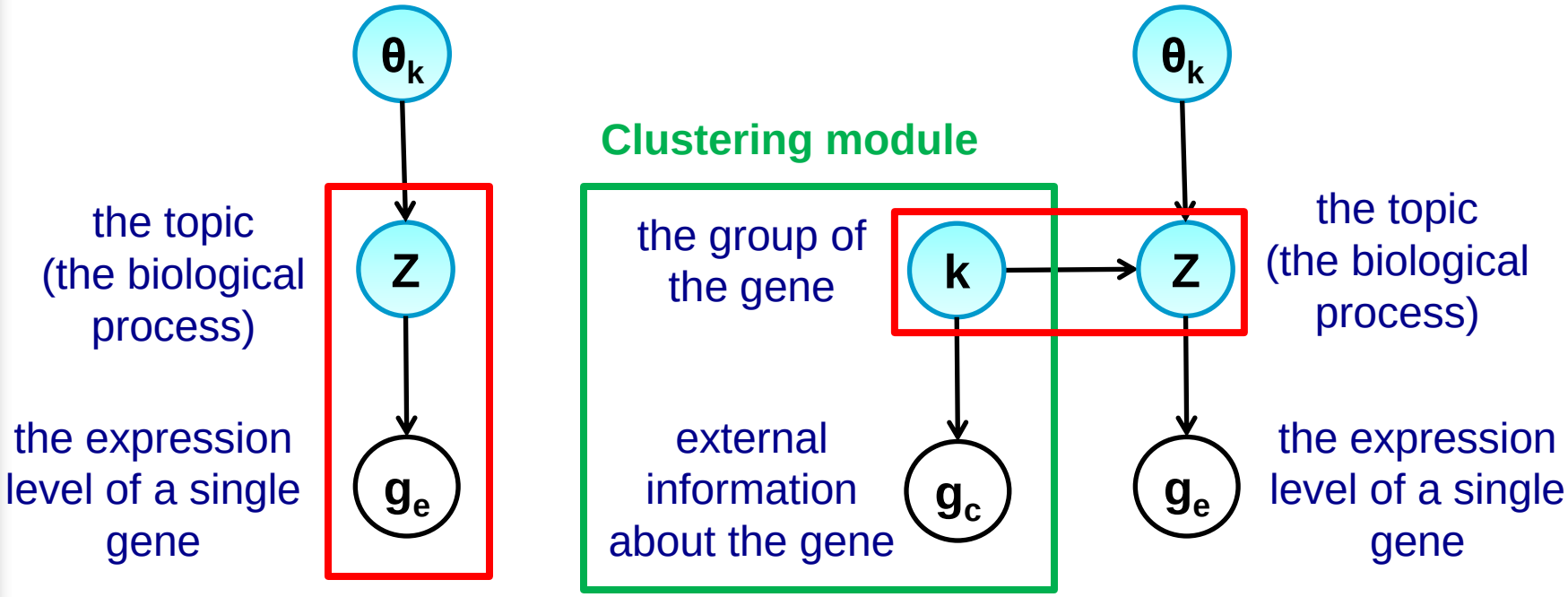
A graph where:

- Nodes represent variables
  - Visible (Effects) D
  - Hidden (Causes) B
- Arcs represent probabilistic dependencies

BaLDA: the main idea

LDA (simplified)

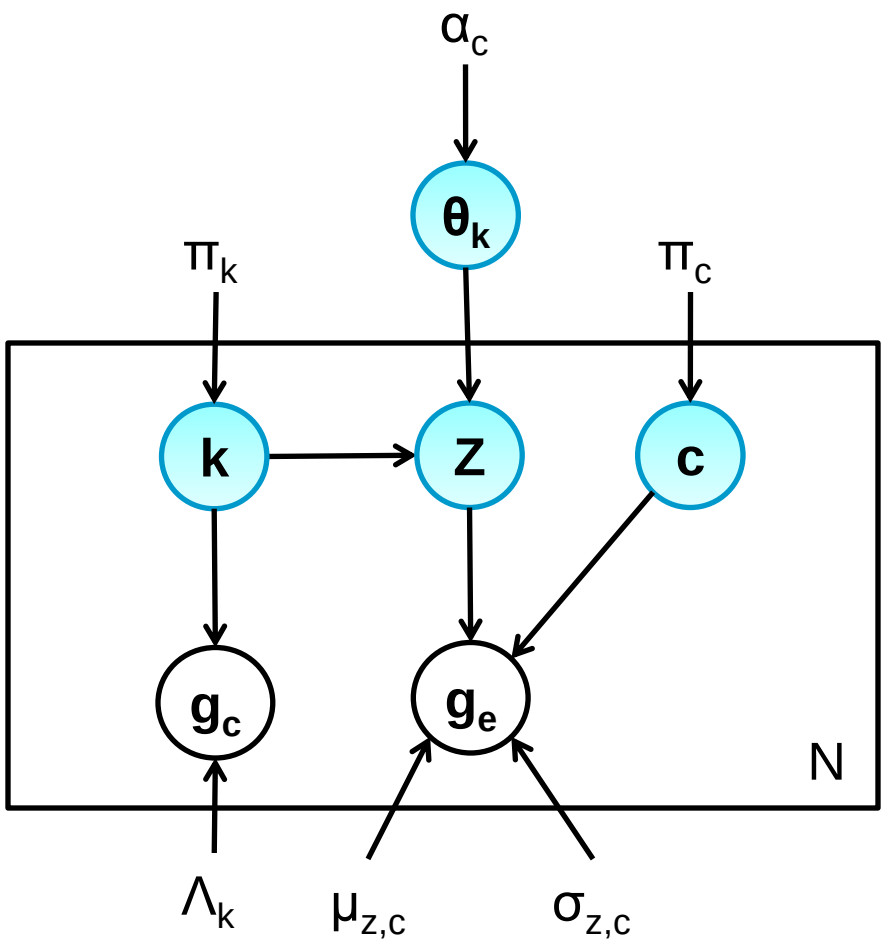
BaLDA (simplified)



the expression level depends from the topic (no relation with other genes)

the topic is influenced by the group the gene belongs to

BaLDA: the full model



In the paper you can find more details on:

- some different versions
- parameters
- learning
- inference
- interpretability

or come to the poster session and ask directly!

Preliminary evaluation

- Question: is it useful for classification?
- Classification with topic models:
  - every sample is characterized by its mixture of topics
- classification is performed with SVM and KNN
  - different kernels: linear and Information Theoretic kernels
- Cross validation classification accuracies

Preliminary evaluation

- Tests on two datasets, comparing BaLDA with LDA
  - Prostate cancer : 54 samples with 9984 features, divided in 3 classes.
  - Brain tumor : 90 samples with 5920 features, divided in 5 classes.
- A priori information used for clustering: *again the expression levels*
- (all the details are in the paper)

Classification accuracies

Prostate Cancer Dataset

	SVM – linear	SVM – best ITK	KNN
LDA (3 topics)	0.651	0.685	0.777
LDA (12 topics)	0.862	0.855	0.822
BaLDA (3 topics, 4 clusters)	**0.899**	**0.912**	**0.852**

Brain Tumor Dataset

	SVM – linear	SVM – best ITK	KNN
LDA (15 topics)	0.833	0.841	0,786
LDA (90 topics)	0.667	0.667	**0.821**
BaLDA (15 topics, 6 clusters)	**0.852**	**0.889**	0.811

In all cases, standard errors of the mean are all less than 0.03

Conclusions

- A novel topic model for the classification of expression microarray data has been introduced
  - Able to take into account known relations between genes
- Preliminary results are promising

- Next: exploit different a priori information (e.g. spatial proximity, sequence similarities,...)

THANK YOU!

QUESTIONS?