Exploring complex data sets with generative models

Fabrice Rossi

SAMM
Université Paris 1 Panthéon Sorbonne

December 1st, 2017

Fabrice.Rossi@apiacoa.org
http://apiacoa.org/

Exploratory data analysis

Data

- a set of N objects/instances described by P variables
- $(X_i)_{1 \leq i \leq N}$, $X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{iP})$
- no particular structure (e.g., no target variable)

Exploratory data analysis

Data

- a set of N objects/instances described by P variables
- $(X_i)_{1 \leq i \leq N}$, $X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{iP})$
- no particular structure (e.g., no target variable)

Goals

- to “make sense” of the data
- to find “interesting aspects” in the data
- and many more imprecise, ill-defined, ambiguous and vague goals

Exploratory data analysis

Data

- a set of N objects/instances described by P variables
- $(X_i)_{1 \leq i \leq N}$, $X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{iP})$
- no particular structure (e.g., no target variable)

Goals

- to “make sense” of the data
- to find “interesting aspects” in the data
- and many more imprecise, ill-defined, ambiguous and vague goals

Rationalization

- clusters
- frequent patterns
- outliers

A standard recipe

for data exploration

Summarizing the data set

Clustering

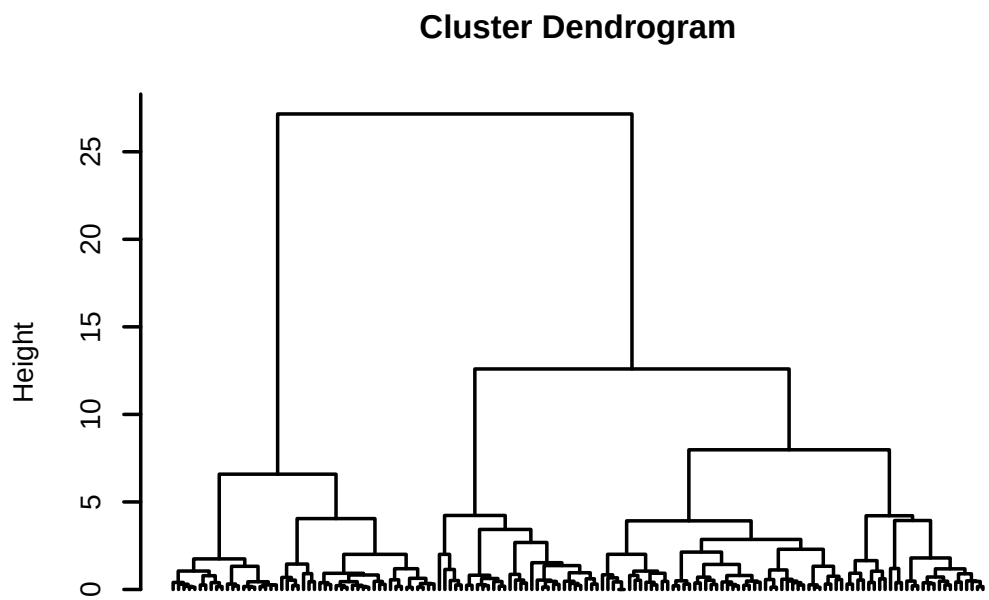
- building groups of similar objects (possibly with prototypes)
- coping with data size (object wise)
- e.g. hierarchical clustering, k-means

Dimensionality reduction

- building or selecting a small number of variables/features
- enabling scatter plot like visualization
- coping with data size (variable wise)
- e.g. Principal Component Analysis

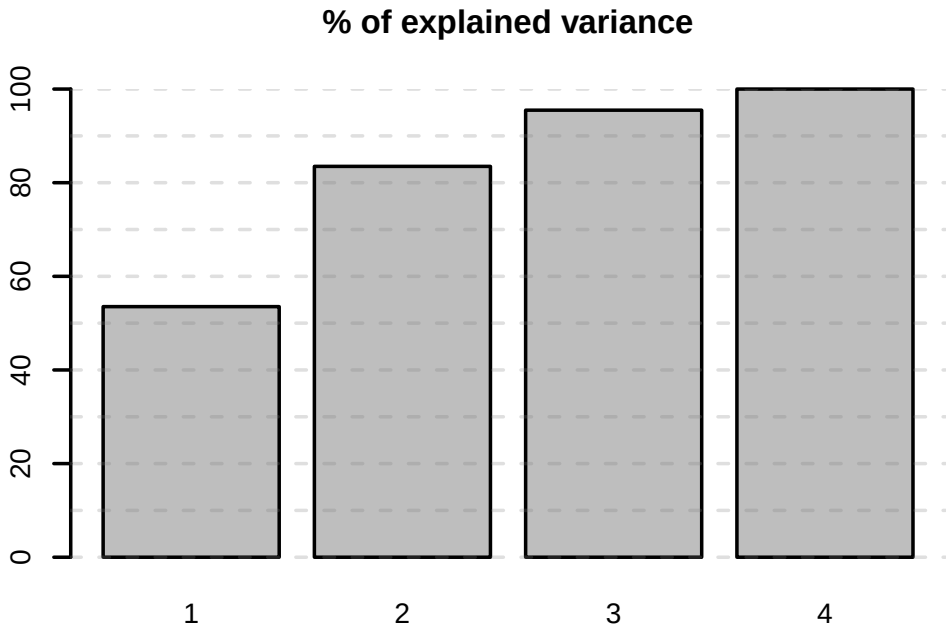
Basic example

- a basic 4 dimensional data set with only 150 observations (iris data set)
- Ward hierarchical clustering
- 2 or 3 classes
- possibly some substructures

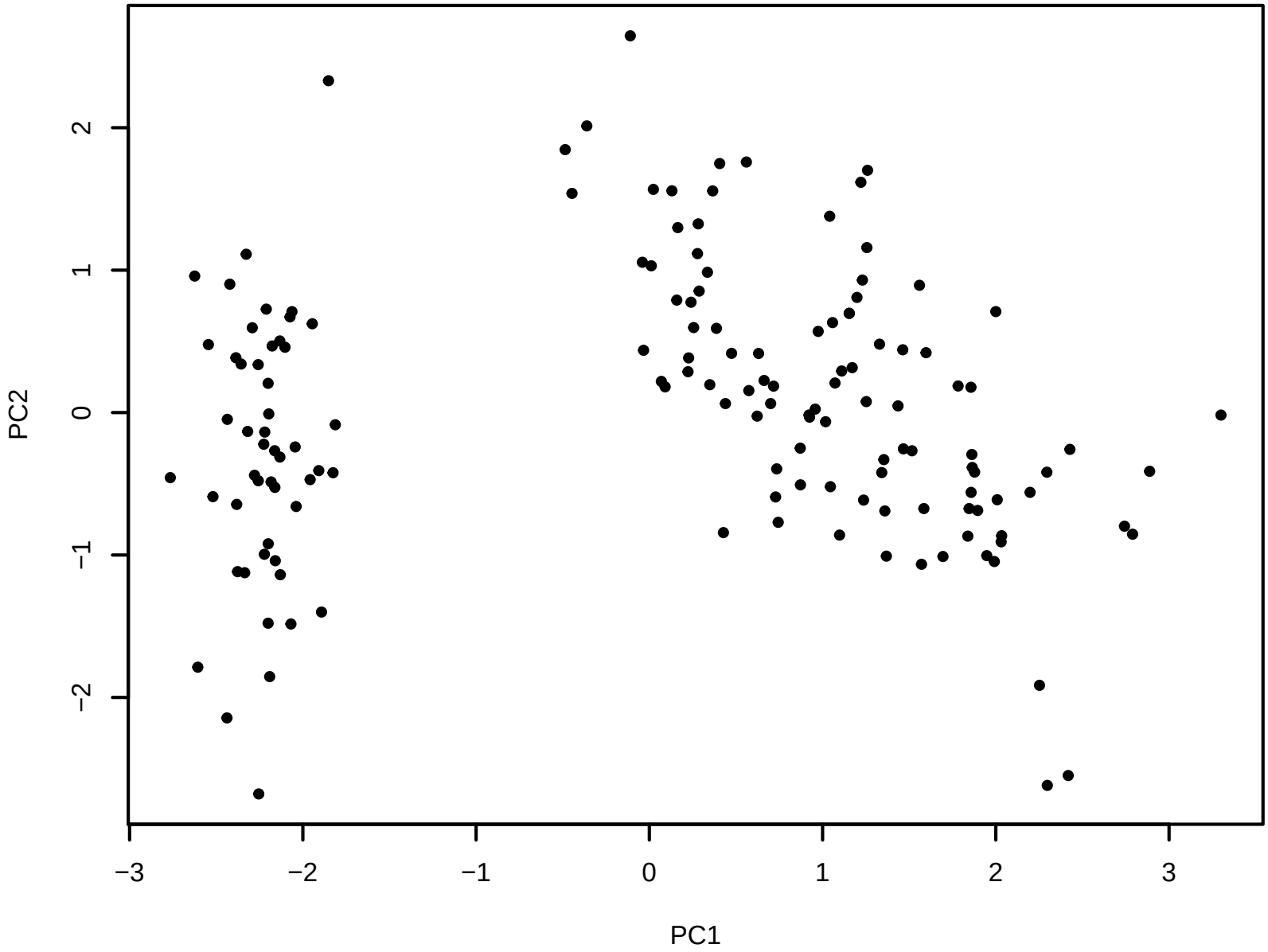


Basic example

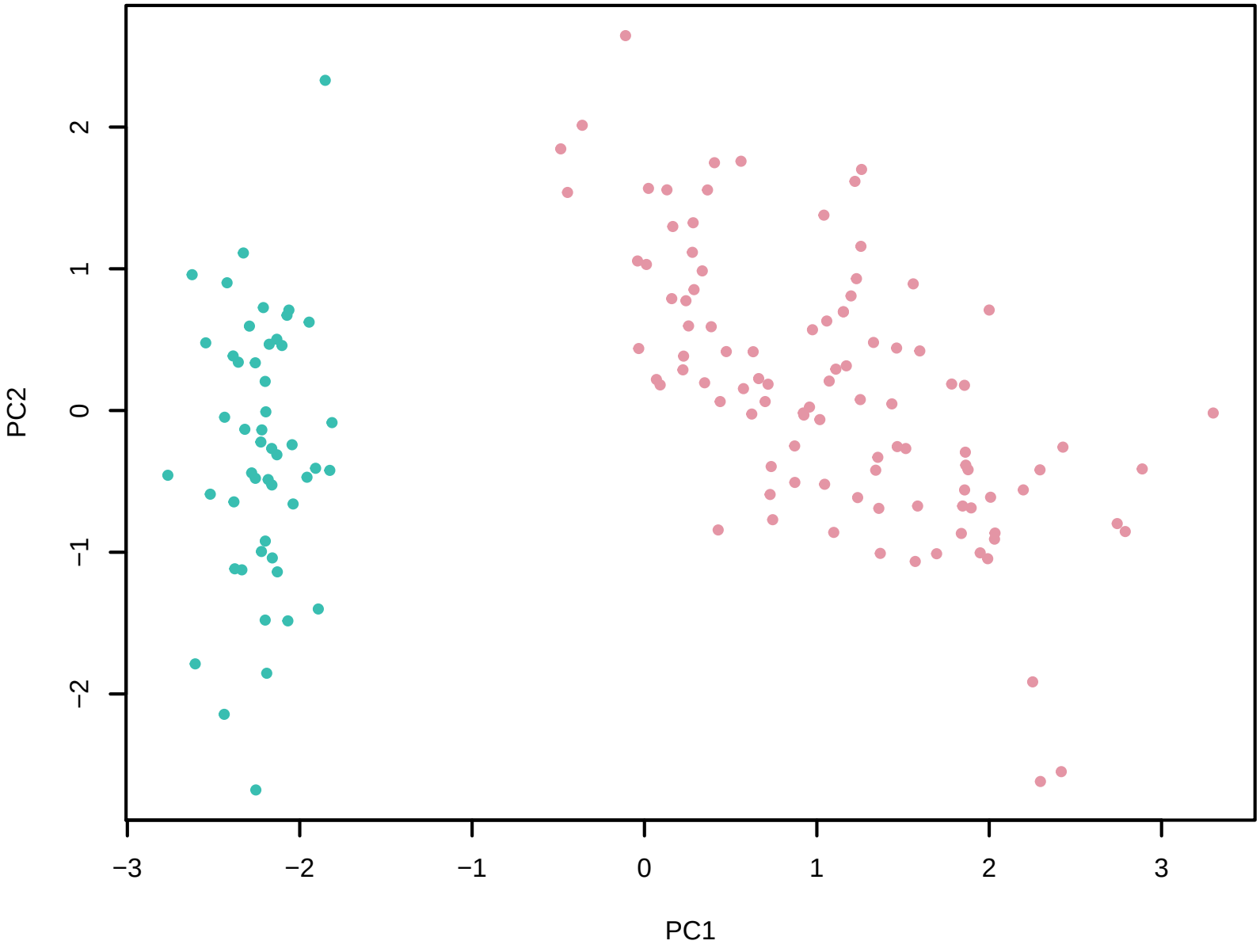
- a basic 4 dimensional data set with only 150 observations (iris data set)
- principal component analysis
- 2 or 3 components



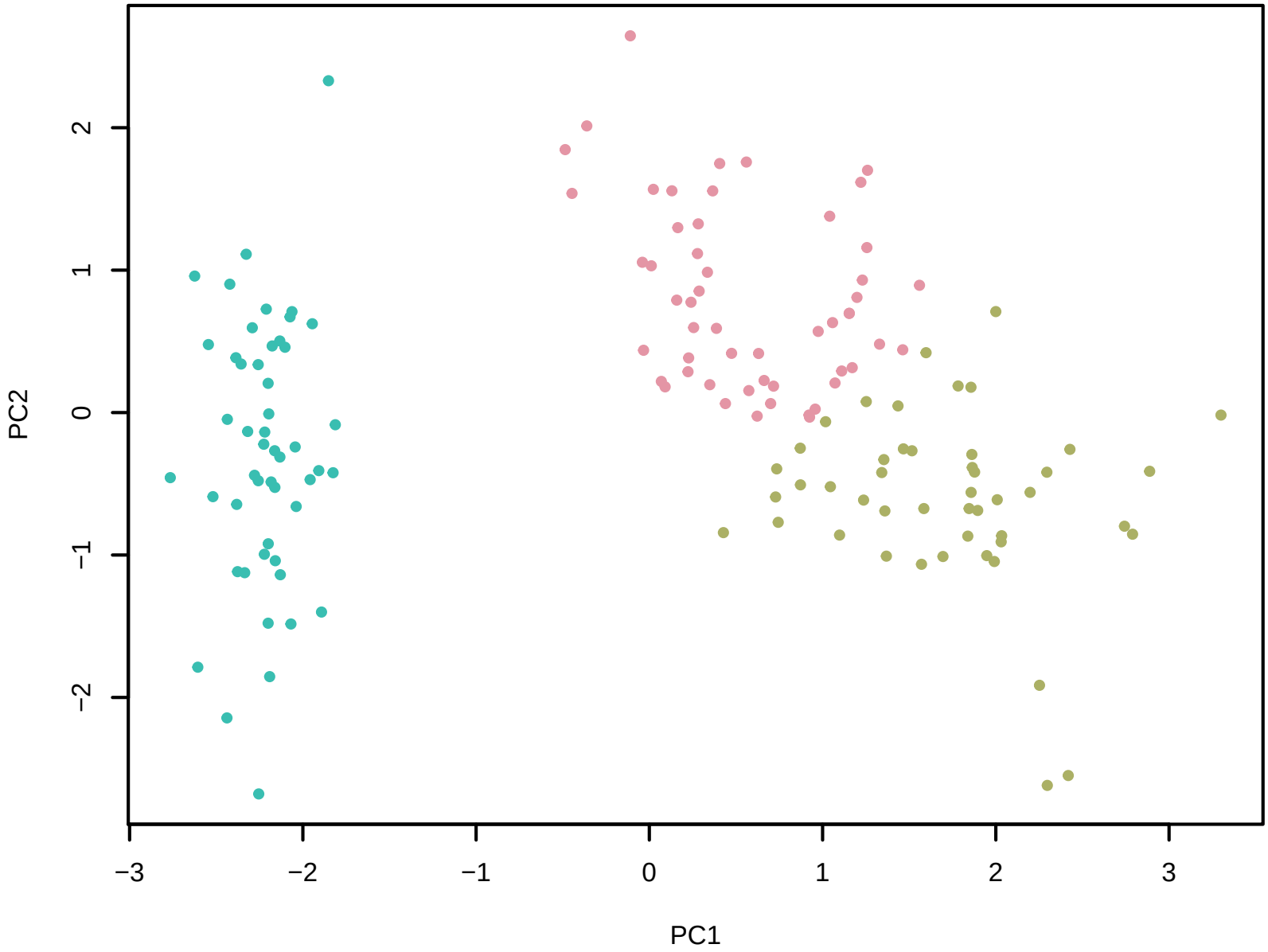
Basic example



Basic example



Basic example



Limitations

Computational scalability

- PCA: $O(NP^2 + P^3)$
- K means: $O(KNP)$
- hierarchical clustering: $O(N^2 P)$ (in general)

Human scalability

- a small number of variables (2 or 3, slightly more with brush and link techniques)
- a reasonable number of clusters (less than e.g. 100)

Limitations

Computational scalability

- PCA: $O(NP^2 + P^3)$
- K means: $O(KNP)$
- hierarchical clustering: $O(N^2 P)$ (in general)

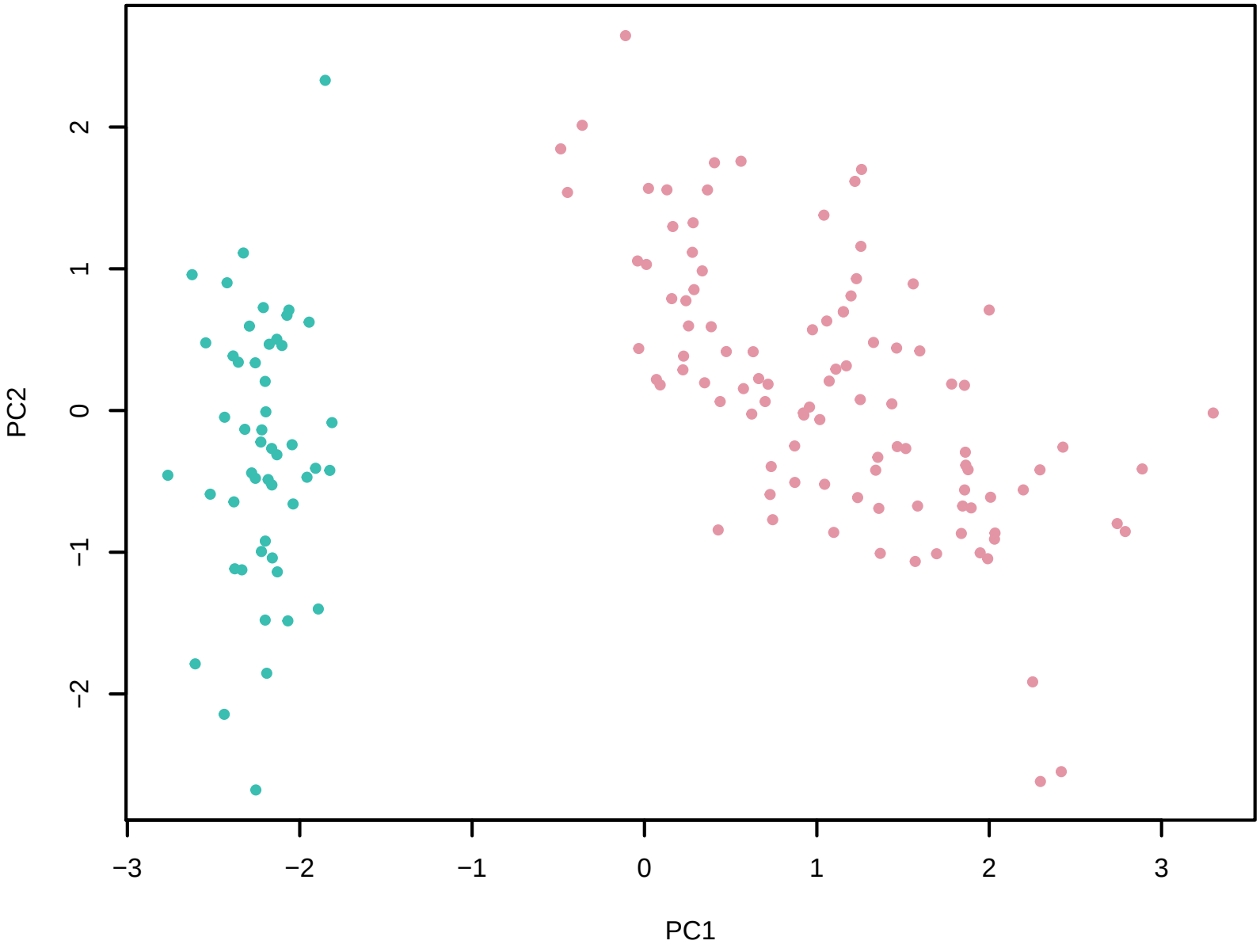
Human scalability

- a small number of variables (2 or 3, slightly more with brush and link techniques)
- a reasonable number of clusters (less than e.g. 100)

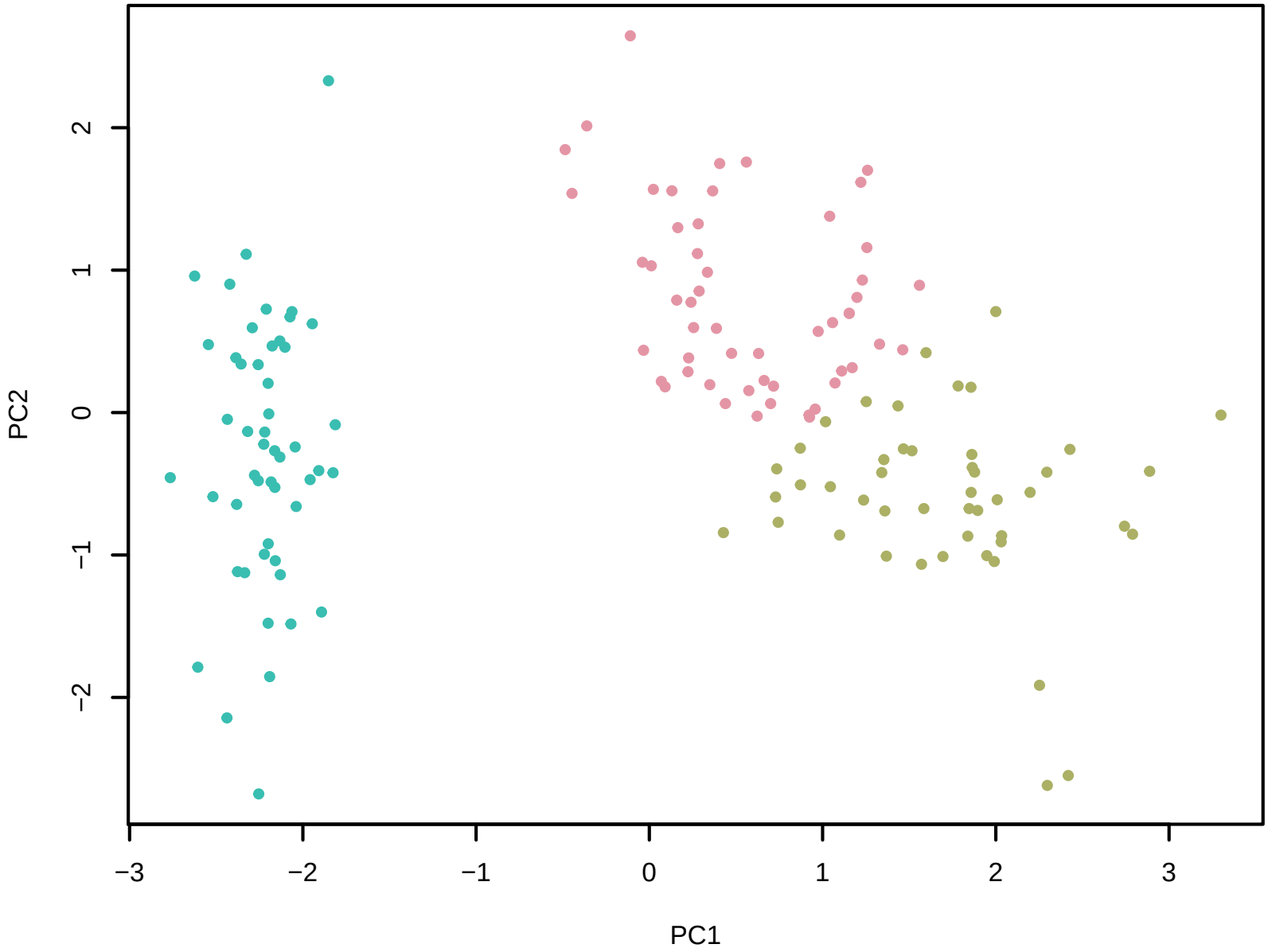
Additional problems

- distorsion
- non significant patterns

Basic example



Basic example



Ideal tool set

Scalable

- ideally linear with NP
- sublinear multiplicative terms are acceptable e.g. $\sqrt{NP}$
- a quadratic complexity strongly limits the scalability:
  - $NP \sim 10^7$: 10 millions, e.g. 100,000 objects with 100 variables
  - $N^2P^2 \sim 10^{14}$: 100 Tera operations
  - current best GPU (Intel Xeon Phi and Nvidia Tesla): 4 Tera flops per second

Usable

- mixed data processing (numerical and categorical)
- complex data processing (graphs and texts)
- automated models: baseline results should not need user steering
- interpretable models: provided with tools that empower the user

Outline

Introduction

Generative models

Co-clustering

Probabilistic modeling

Back to the data

- observation: $\mathcal{D} = (X_i)_{1 \leq i \leq N}$
- probabilistic point of view:
  - $\mathcal{D}$ is a random variable
  - the best way to characterize it is to compute its distribution $\mathbb{P}_{\mathcal{D}}$

Pros and cons

+ very rich setting: mixed data, missing data, graphs, texts, etc.
+ large tool set: tests, simulations, imputation, etc.
+ automatic model selection (mostly Bayesian)
- complex on a mathematical point of view
- somewhat costly on a computational point of view

Probabilistic modeling

Back to the data

- observation: $\mathcal{D} = (X_i)_{1 \leq i \leq N}$
- probabilistic point of view:
  - $\mathcal{D}$ is a random variable
  - the best way to characterize it is to compute its distribution $\mathbb{P}_{\mathcal{D}}$

Pros and cons

+ very rich setting: mixed data, missing data, graphs, texts, etc.
+ large tool set: tests, simulations, imputation, etc.
+ automatic model selection (mostly Bayesian)
- complex on a mathematical point of view
- somewhat costly on a computational point of view

How to estimate $\mathbb{P}_{\mathcal{D}}$?

Independence

Too many parameters

- simple case: $X_{ij} \in \{0, 1\}$
- full specification of $\mathbb{P}_{\mathcal{D}}$: $2^{NP} - 1$ parameters (!)
- simplification hypotheses are needed

Independence

Too many parameters

- simple case: $X_{ij} \in \{0, 1\}$
- full specification of $\mathbb{P}_{\mathcal{D}}$: $2^{NP} - 1$ parameters (!)
- simplification hypotheses are needed

Independence

- main tool: regularity assumption
- strong regularity: independence and identity (structural assumptions on $\mathbb{P}_{\mathcal{D}}$)
- very common at the object level: $X_i \perp\!\!\!\perp X_l$

Independence

Too many parameters

- simple case: $X_{ij} \in \{0, 1\}$
- full specification of $\mathbb{P}_{\mathcal{D}}$: $2^{NP} - 1$ parameters (!)
- simplification hypotheses are needed

Independence

- main tool: regularity assumption
- strong regularity: independence and identity (structural assumptions on $\mathbb{P}_{\mathcal{D}}$)
- very common at the object level: $X_i \perp\!\!\!\perp X_l$

Back to the example

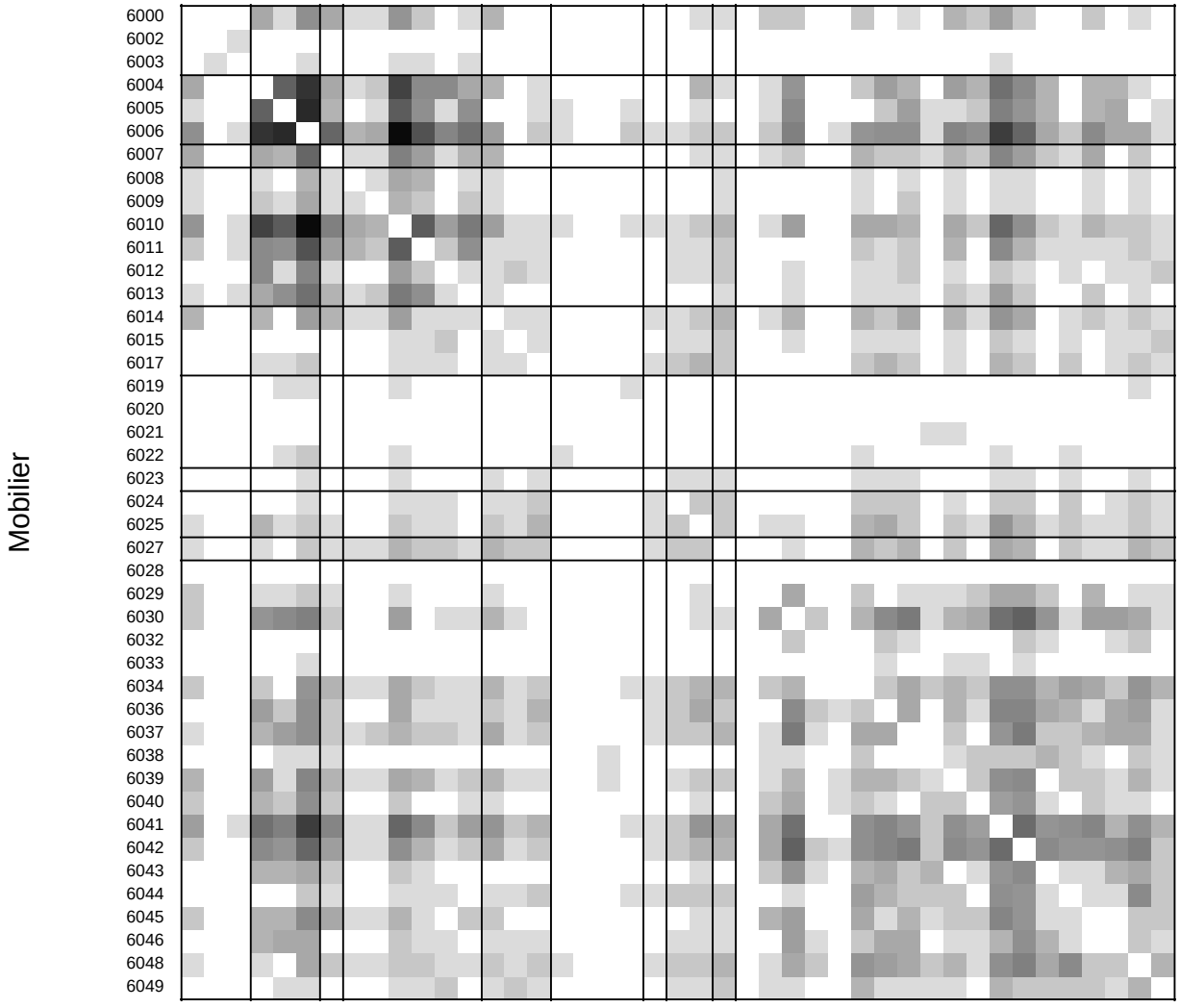
- $(X_i)_{1 \leq i \leq N}$ are independent and identically distributed
- only $\mathbb{P}_{X_1}$ is needed: $2^P - 1$ parameters

Independence

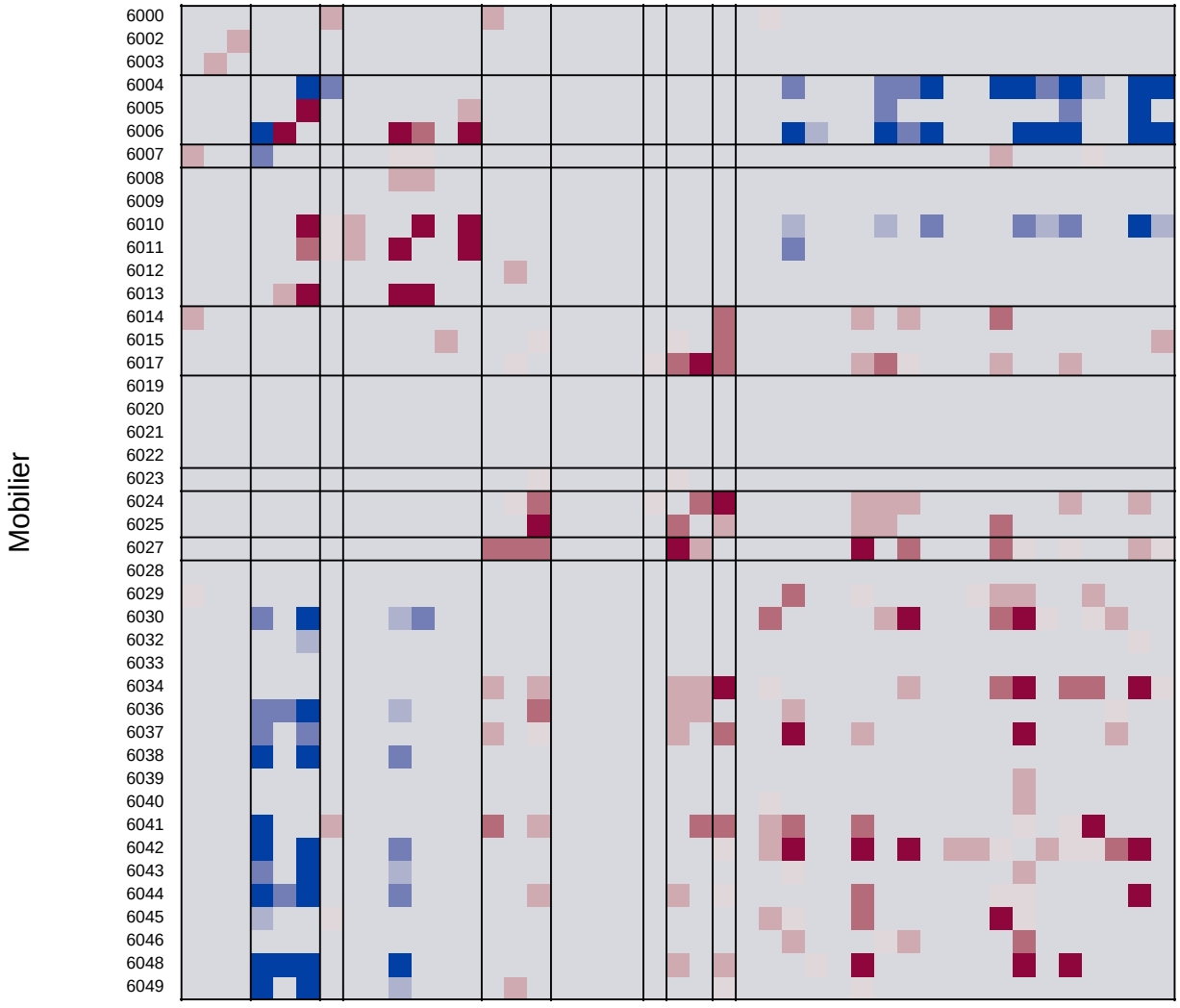
Limitations

- independence is a very strong assumption: does not make sense in many settings
  - $(X_{ij})_{1 \leq j \leq P}$ independent: only P parameters
  - but no links between the variables...
- binary point of view: independent or not, no middle ground

Application



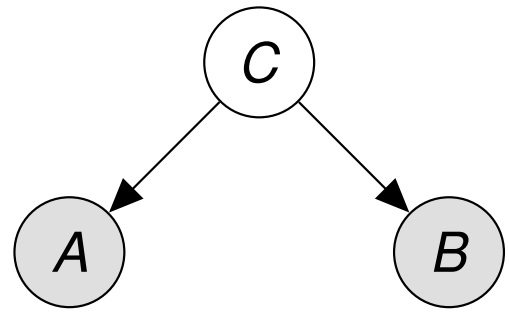
Application



Conditional Independence

Conditional independence

- $A \perp\!\!\!\perp B | C$:
  - A and B are conditionally independent given C
  - C captures all the dependencies between A and B
- C can be hidden: latent variable
- different conditional distributions: mixture models



Mixture models

Binary example

- $\perp\!\!\!\perp_{1 \leq j \leq P} X_{ij} | Z_i$ with $Z_i \in \{1, \ldots, K\}$
- $X_{ij} | Z_i = k \sim \mathcal{B}(\gamma_{kj})$ i.e. $\mathbb{P}(X_{ij} = 1 | Z_i = k) = \gamma_{kj}$
  - X_{ij} is still a Bernoulli distribution with $\gamma_j = \sum_k \gamma_{kj} \mathbb{P}(Z_i = k)$
  - but $\mathbb{P}(X_{ij} = 1, X_{il} = 1) \neq \mathbb{P}(X_{ij} = 1)\mathbb{P}(X_{il} = 1)$!
- KP parameters with K arbitrary

Mixture models

Binary example

- $\perp\!\!\!\perp_{1 \leq j \leq P} X_{ij} | Z_i$ with $Z_i \in \{1, \ldots, K\}$
- $X_{ij} | Z_i = k \sim \mathcal{B}(\gamma_{kj})$ i.e. $\mathbb{P}(X_{ij} = 1 | Z_i = k) = \gamma_{kj}$
  - X_{ij} is still a Bernoulli distribution with $\gamma_j = \sum_k \gamma_{kj} \mathbb{P}(Z_i = k)$
  - but $\mathbb{P}(X_{ij} = 1, X_{il} = 1) \neq \mathbb{P}(X_{ij} = 1)\mathbb{P}(X_{il} = 1)$!
- KP parameters with K arbitrary

Classical mixture model

- $\perp\!\!\!\perp_{1 \leq i \leq N} Z_i$ and $\perp\!\!\!\perp_{1 \leq i \leq N} X_i | (Z_l)_{1 \leq l \leq N}$
- $X_i | Z_i = k \sim f_k$
- $X_i \sim \sum_k \mathbb{P}(Z_i = k) f_k$
- notice that “guessing” Z_i is not needed (but useful!)

Mixture models

Binary example

- $\perp\!\!\!\perp_{1 \leq j \leq P} X_{ij} | Z_i$ with $Z_i \in \{1, \ldots, K\}$
- $X_{ij} | Z_i = k \sim \mathcal{B}(\gamma_{kj})$ i.e. $\mathbb{P}(X_{ij} = 1 | Z_i = k) = \gamma_{kj}$
  - X_{ij} is still a Bernoulli distribution with $\gamma_j = \sum_k \gamma_{kj} \mathbb{P}(Z_i = k)$
  - but $\mathbb{P}(X_{ij} = 1, X_{il} = 1) \neq \mathbb{P}(X_{ij} = 1)\mathbb{P}(X_{il} = 1)$!
- KP parameters with K arbitrary

Classical mixture model

- $\perp\!\!\!\perp_{1 \leq i \leq N} Z_i$ and $\perp\!\!\!\perp_{1 \leq i \leq N} X_i | (Z_l)_{1 \leq l \leq N}$
- $X_i | Z_i = k \sim f_k$
- $X_i \sim \sum_k \mathbb{P}(Z_i = k) f_k$
- notice that “guessing” Z_i is not needed (but useful!)
- EM algorithm

Applications

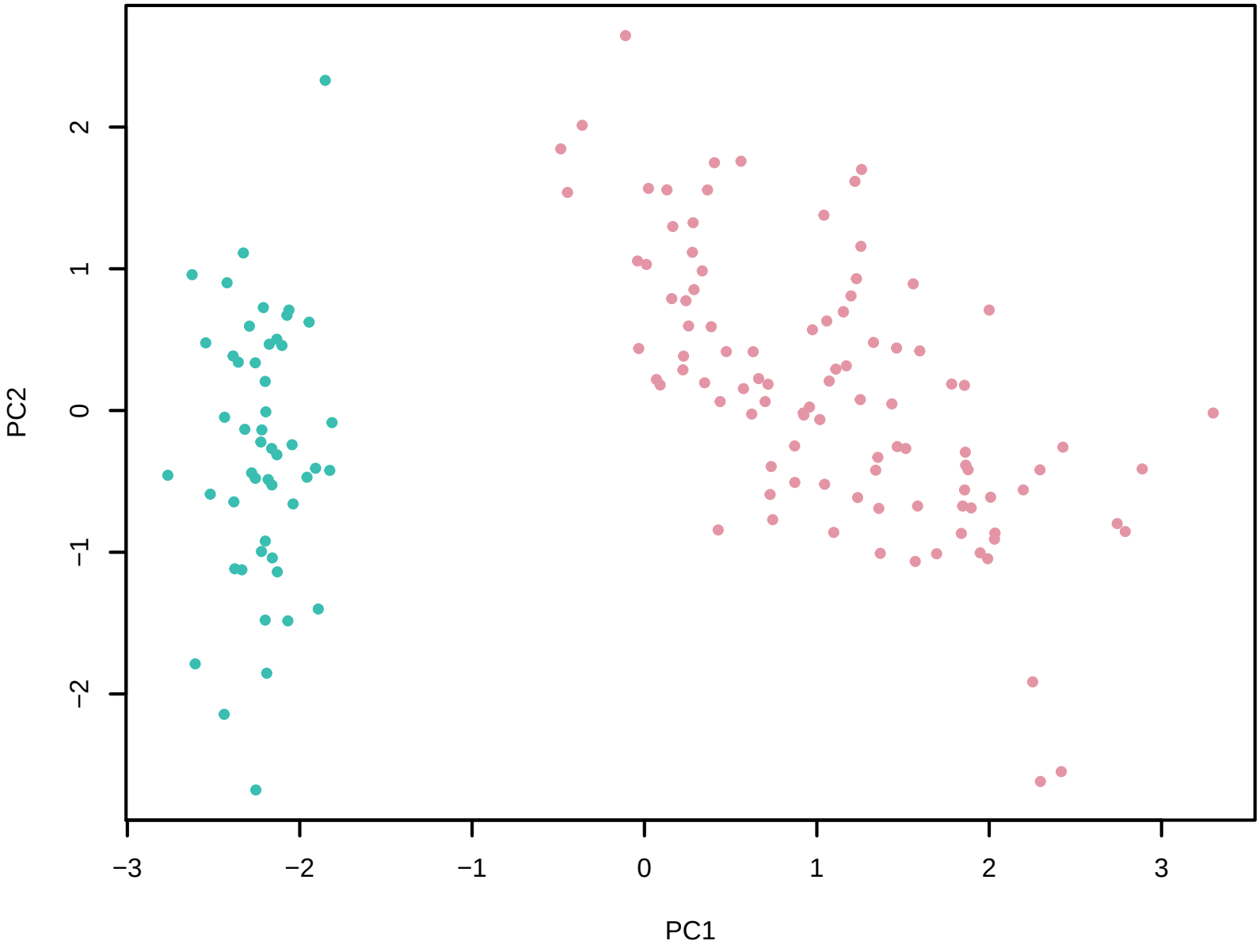
Clustering

- Z_i can be seen as cluster index
- f_k gives the shape of cluster number k
- extends the k-means approach: f_k is a multivariate Gaussian distribution
- many more possibilities especially for non numerical data

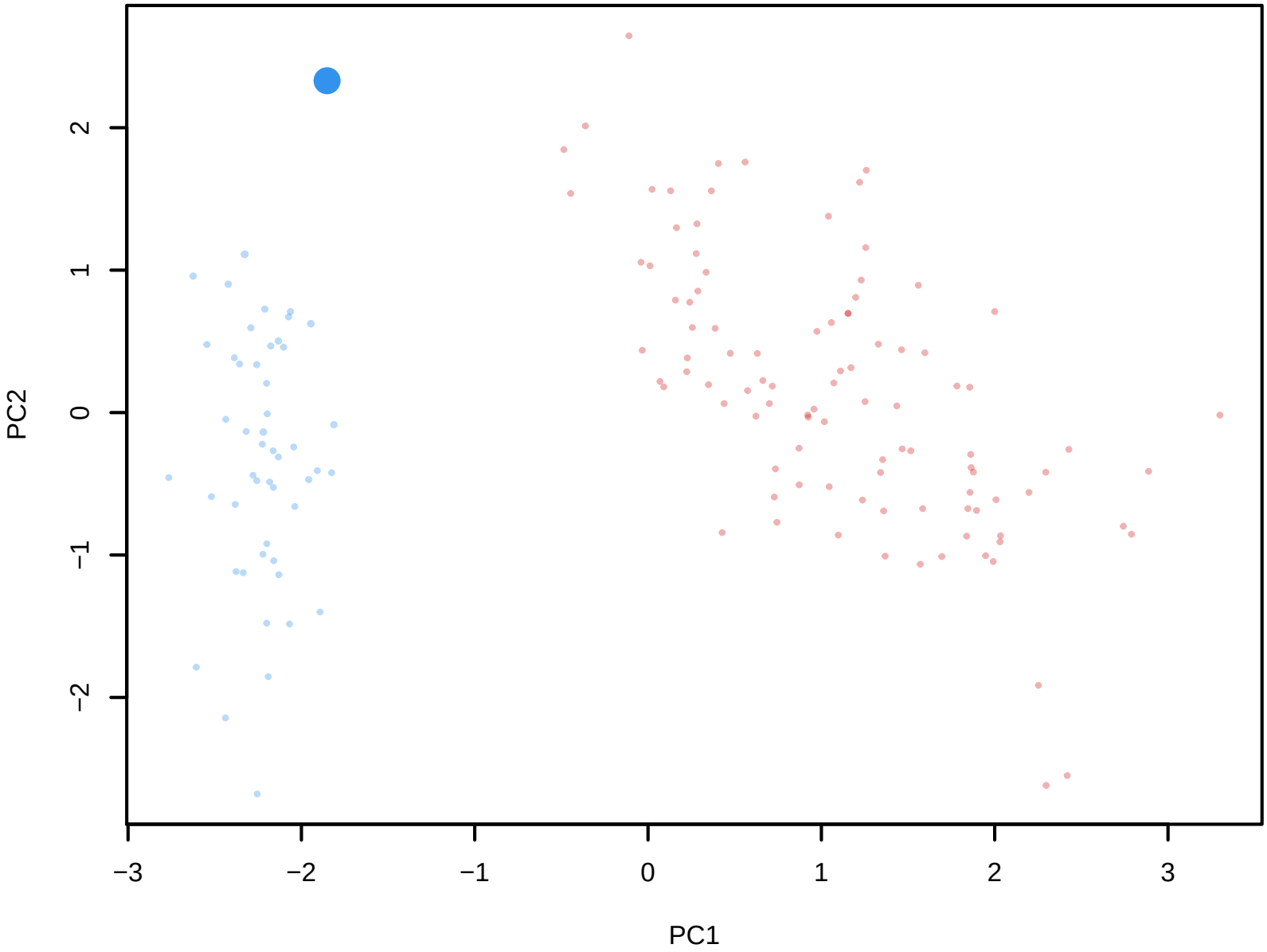
Added value

- principled model selection (e.g. BIC criterion)
- principled missing data handling
- uncertainty measure
- data generation

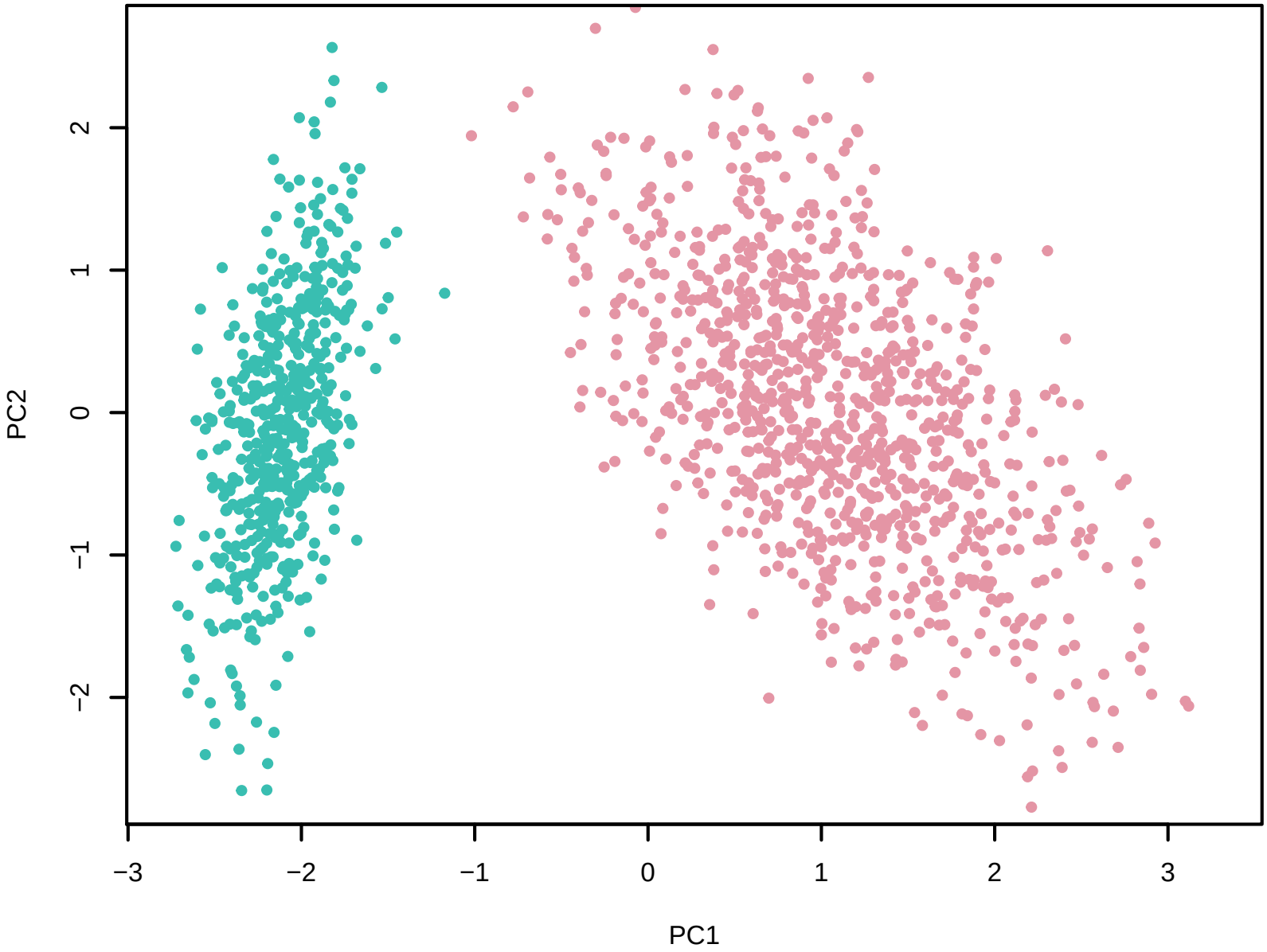
Basic example



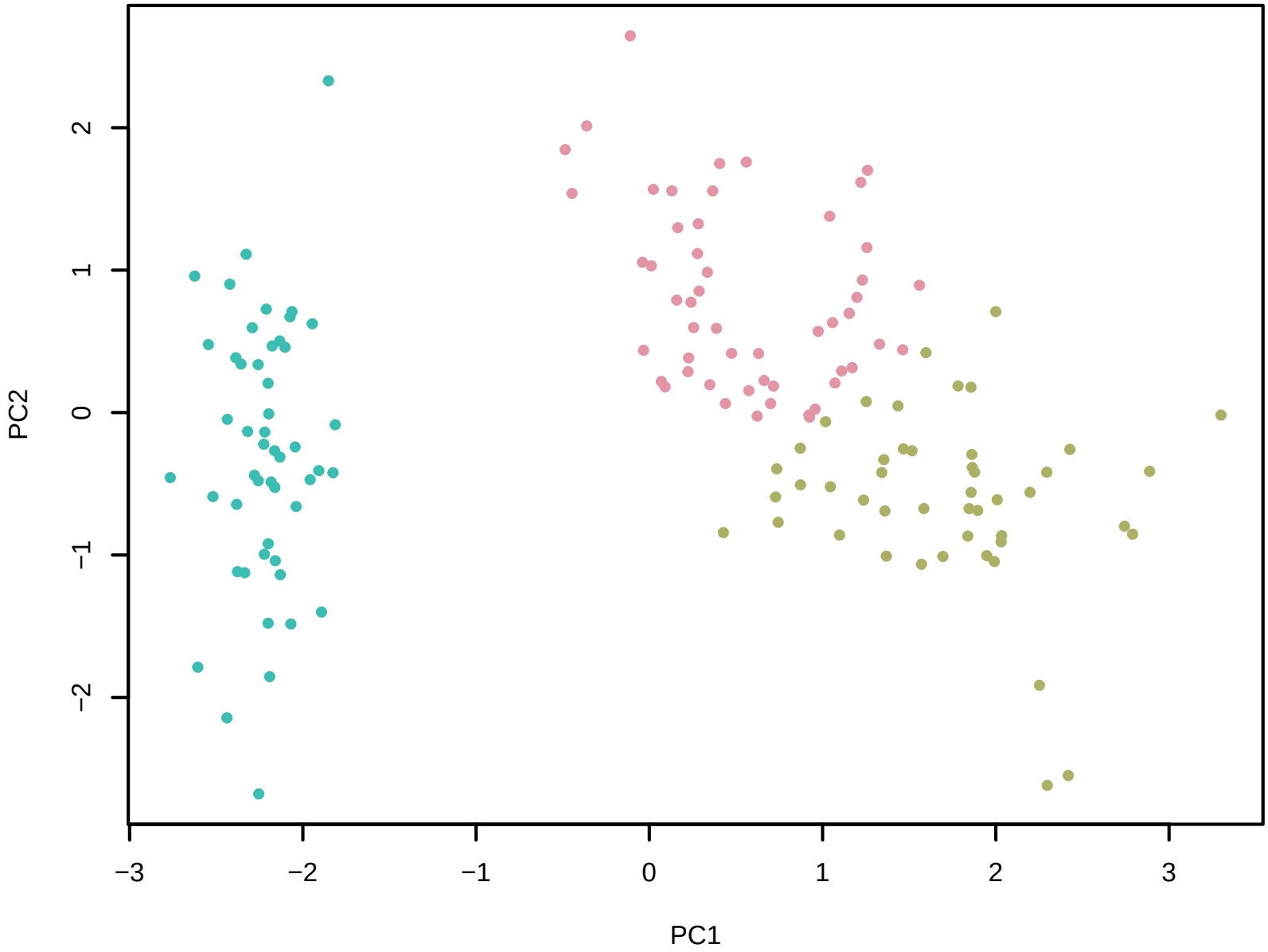
Basic example



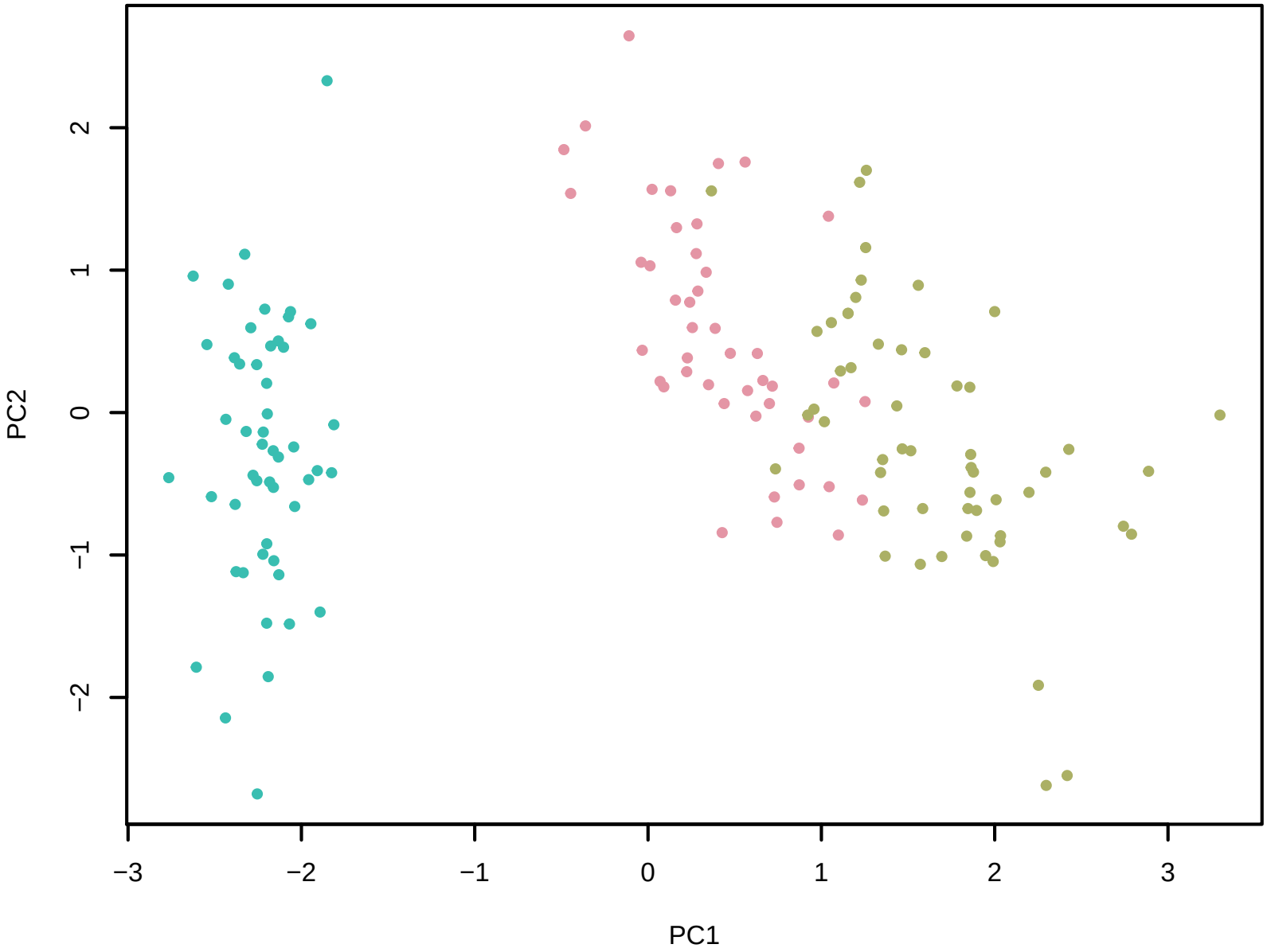
Basic example



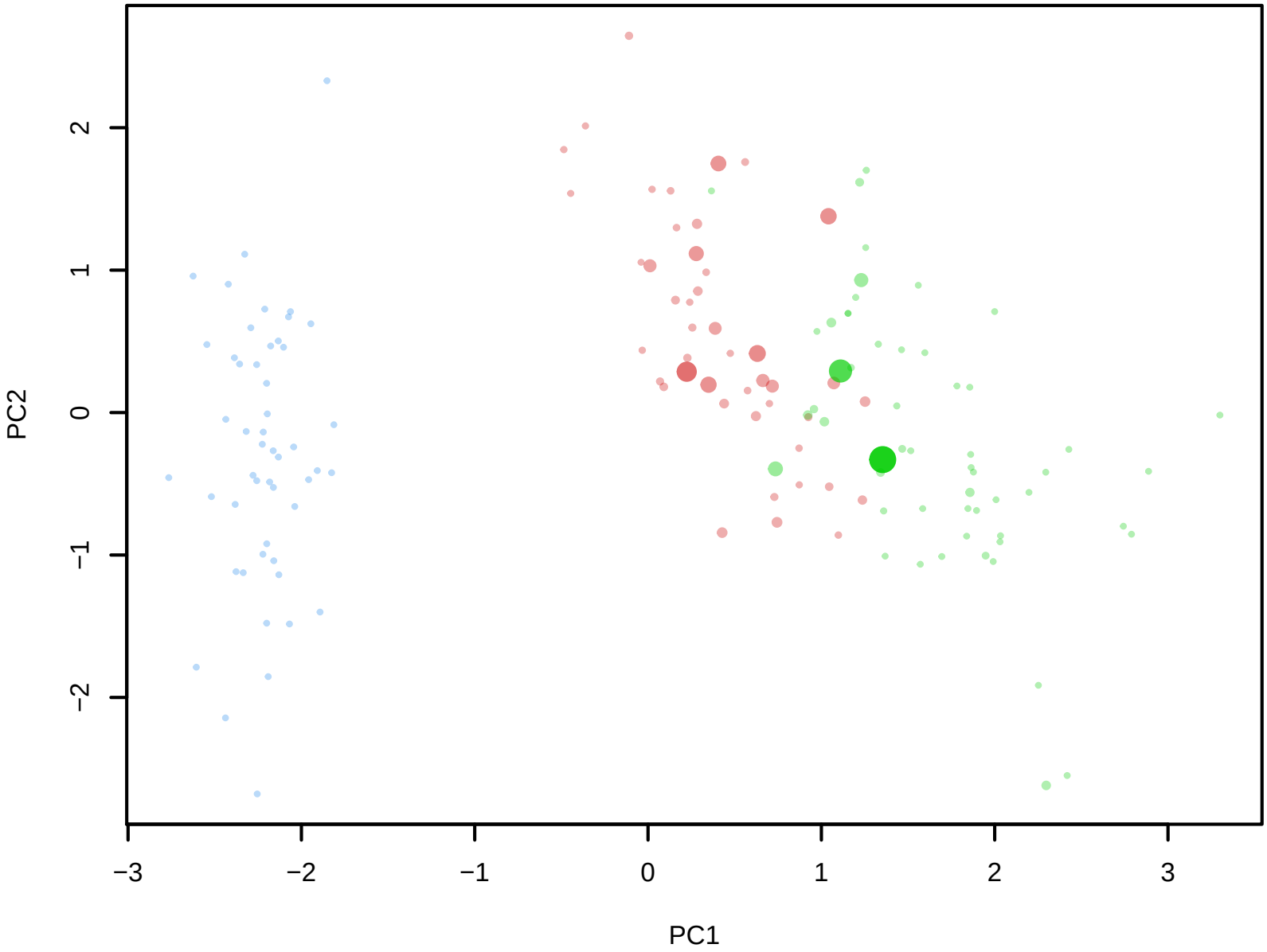
Basic example



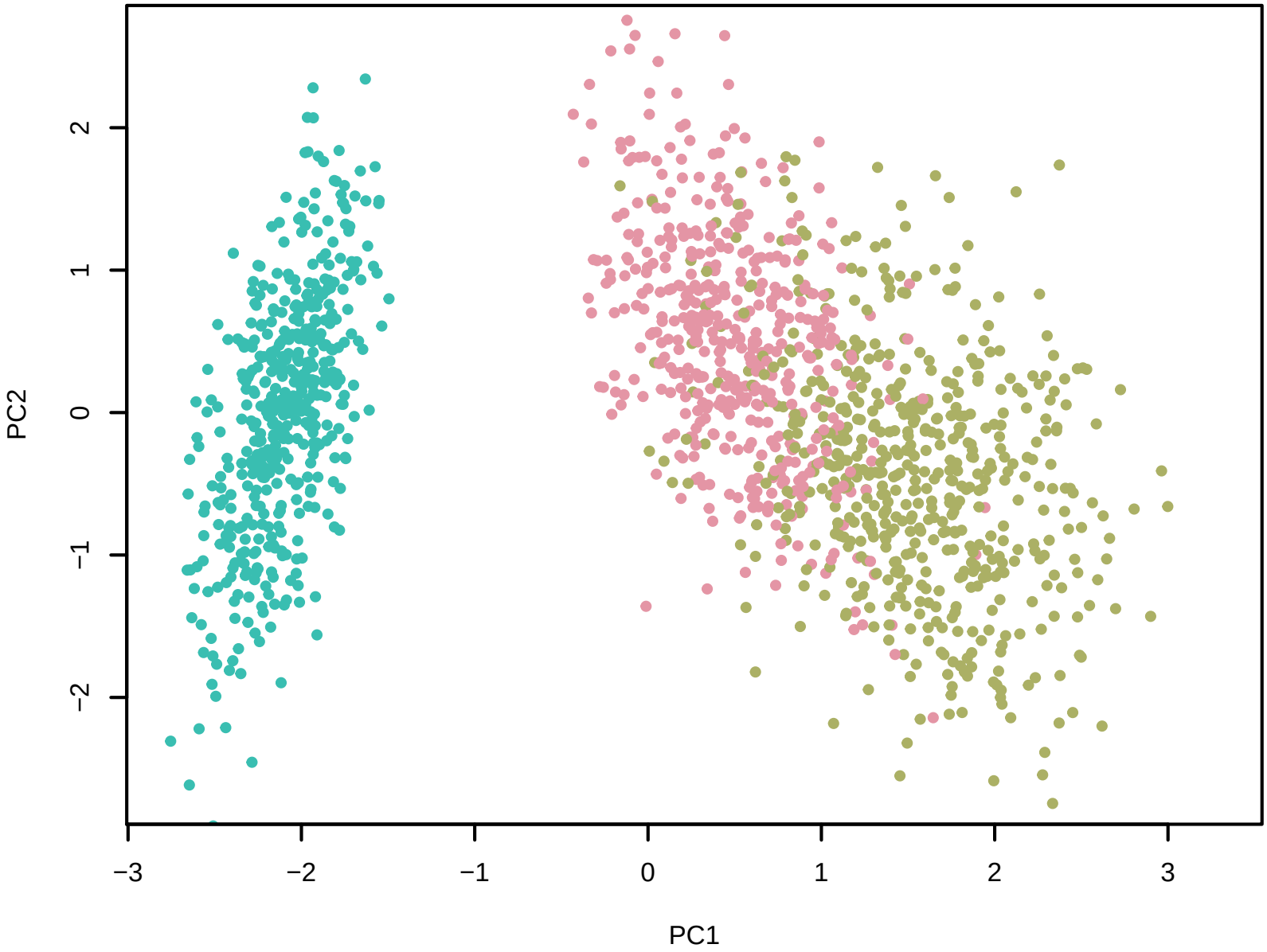
Basic example



Basic example



Basic example



Outline

Introduction

Generative models

Co-clustering

Variable level summary

Dimensionality reduction methods

- probabilistic PCA
- mixture of pPCA
- etc.

Variable level summary

Dimensionality reduction methods

- probabilistic PCA
- mixture of pPCA
- etc.

Co-clustering

- dimensionality reduction by variable clustering
- combined variable clustering and object clustering
- bloc structure extraction

Generative co-clustering

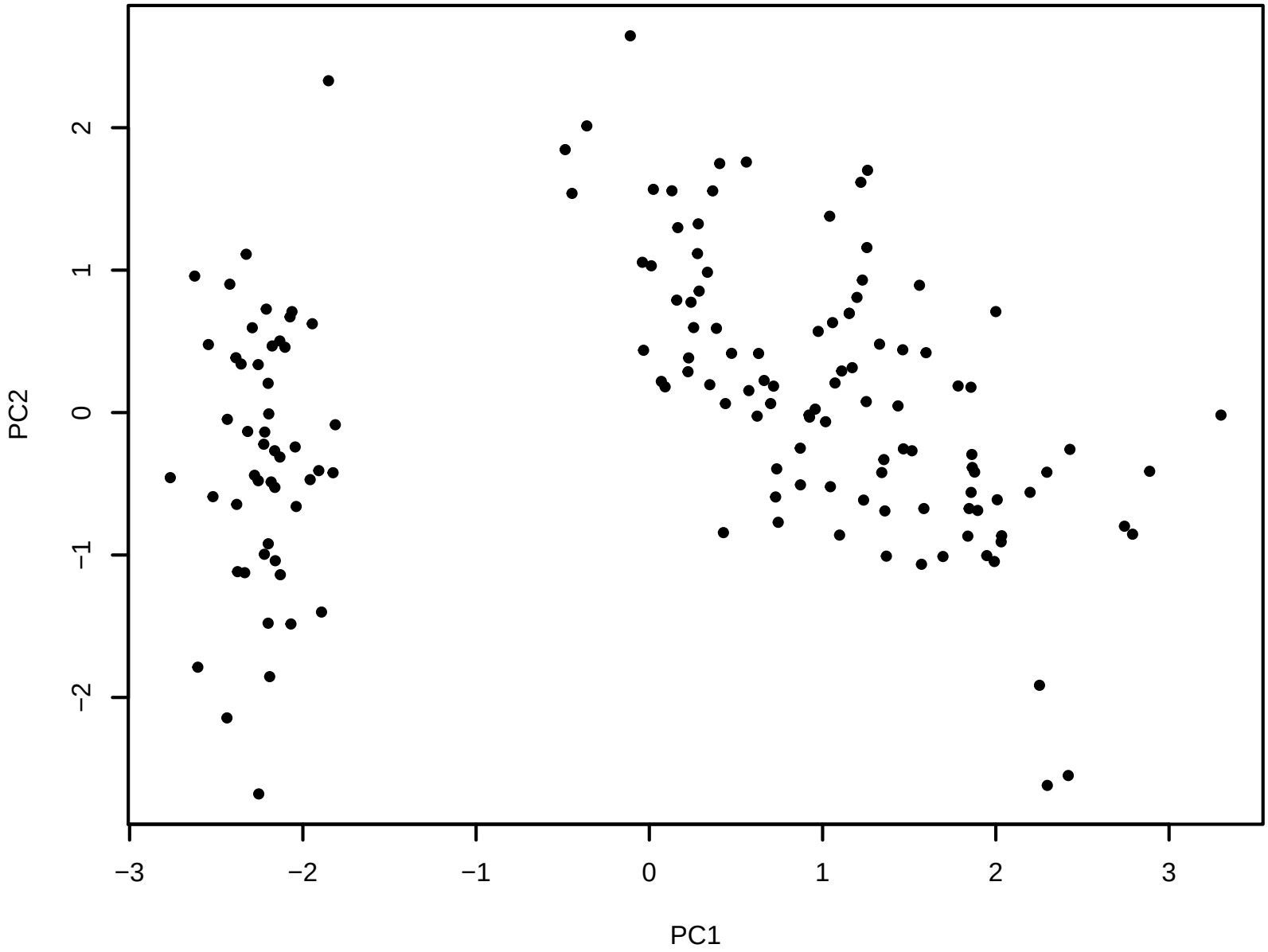
Latent variables

- each object is associated to a hidden variable Z_i
- each variable is associated to a hidden variable W_j
- $\perp\!\!\!\perp_{1 \leq i \leq N, 1 \leq j \leq P} X_{ij} | (Z_i)_{1 \leq i \leq N}, (W_j)_{1 \leq j \leq P}$
- each block $(Z_i = k, W_j = l)$ has a single generative model (e.g. Gaussian)

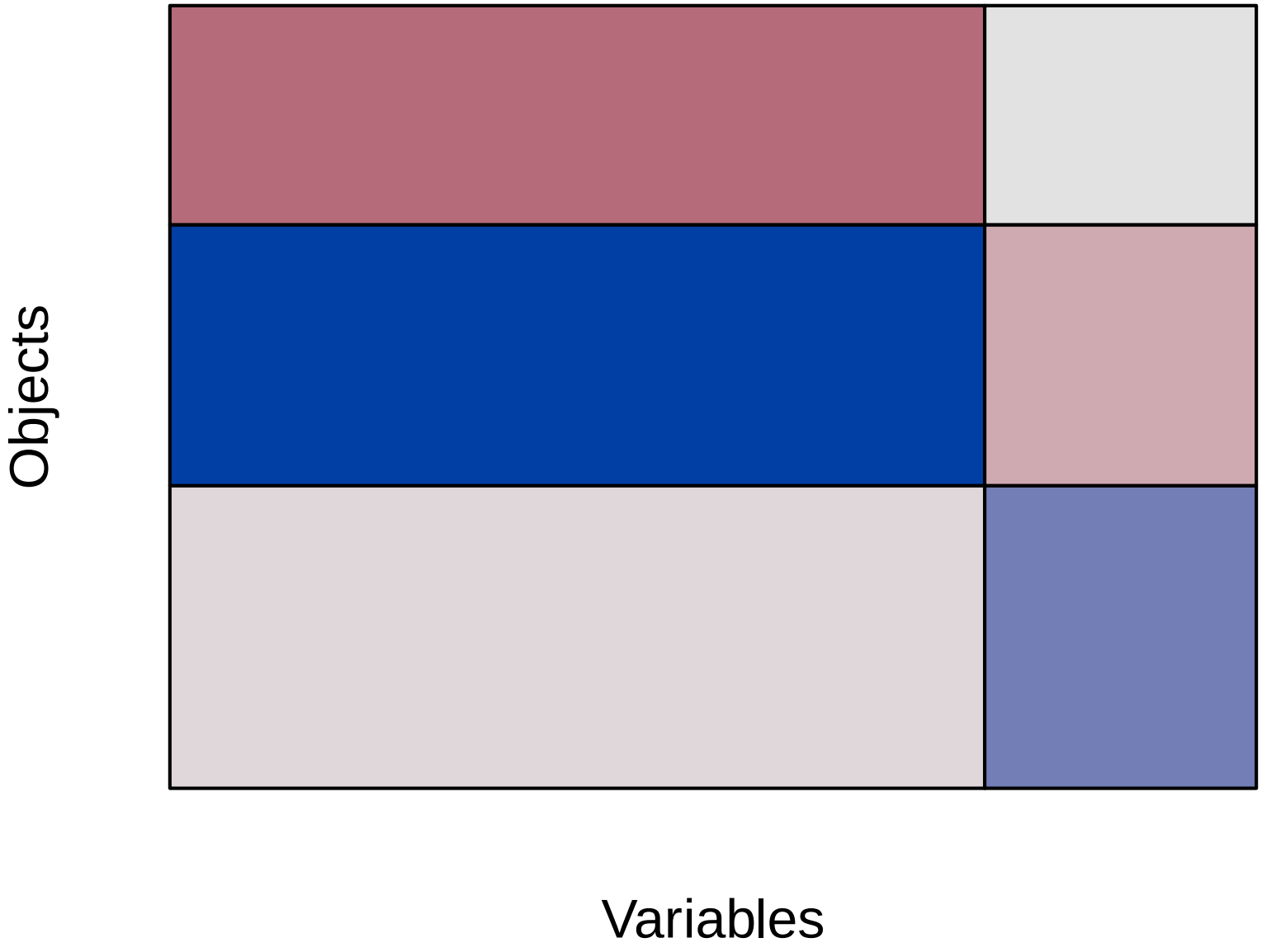
As a mixture model

- $X_{ij} \sim \sum_k \sum_l \mathbb{P}(Z_i = k, W_j = l) f_{kl}$
- variables can be of different nature
- but each variable cluster has a specific type

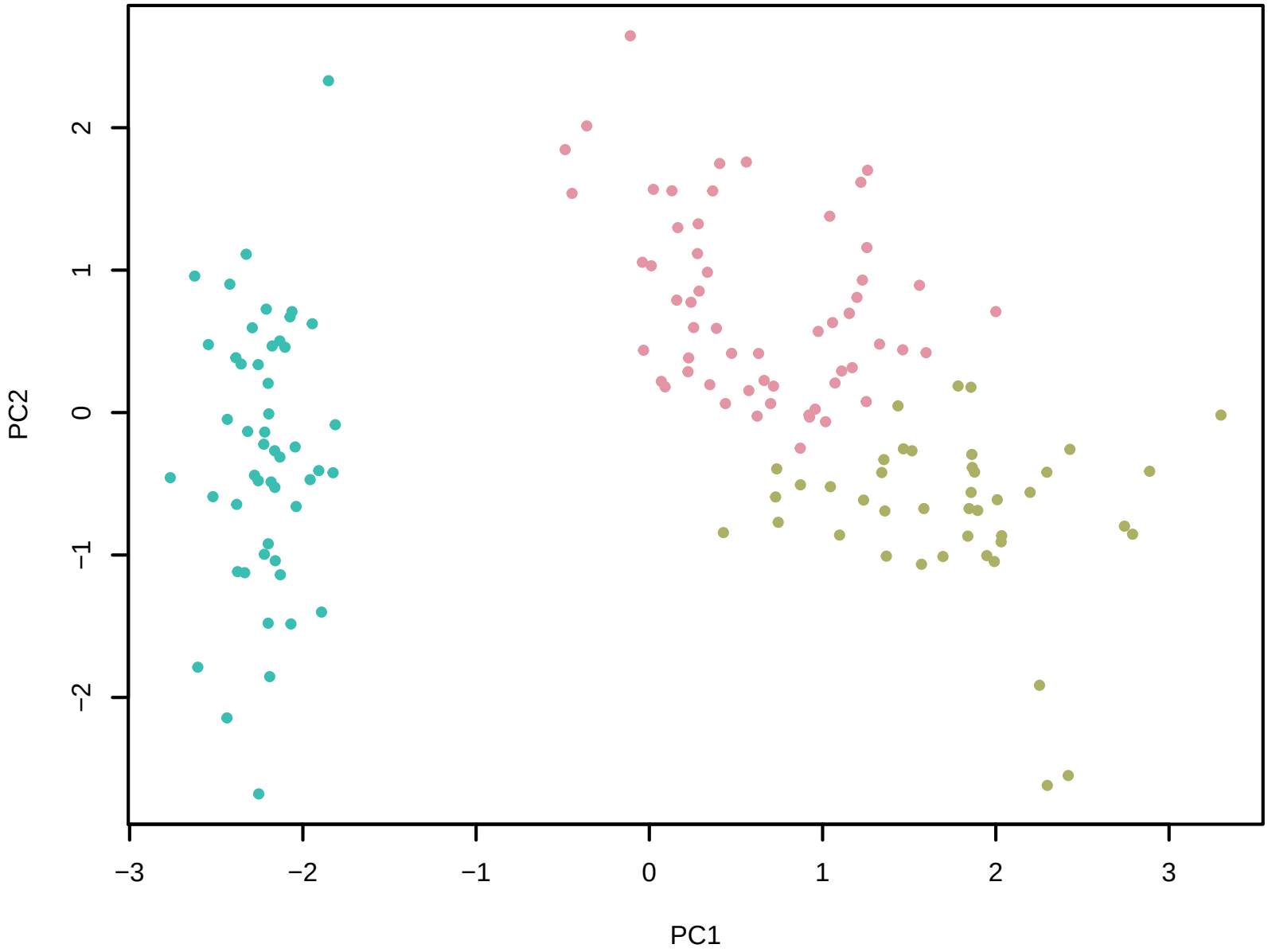
Basic example



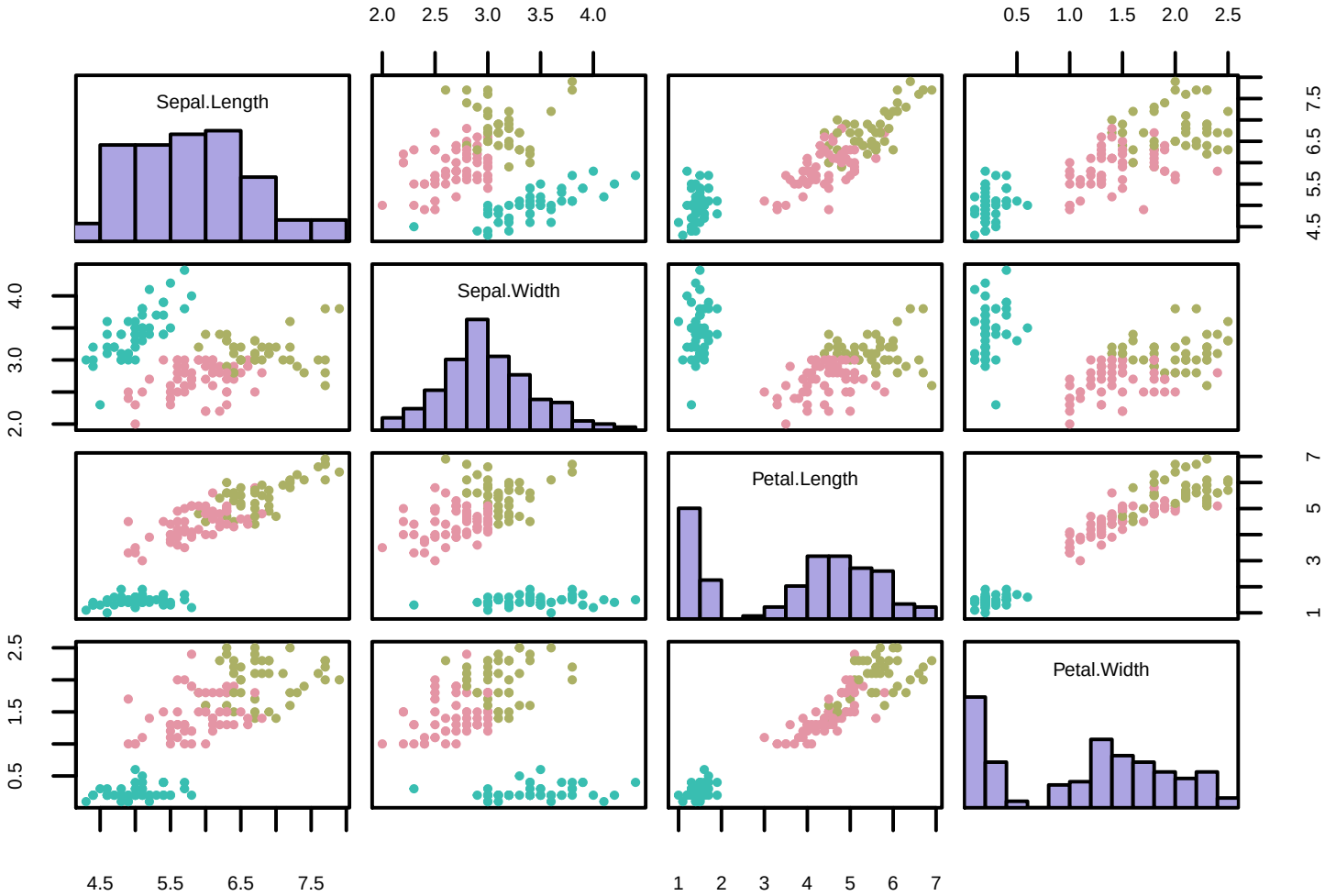
Basic example



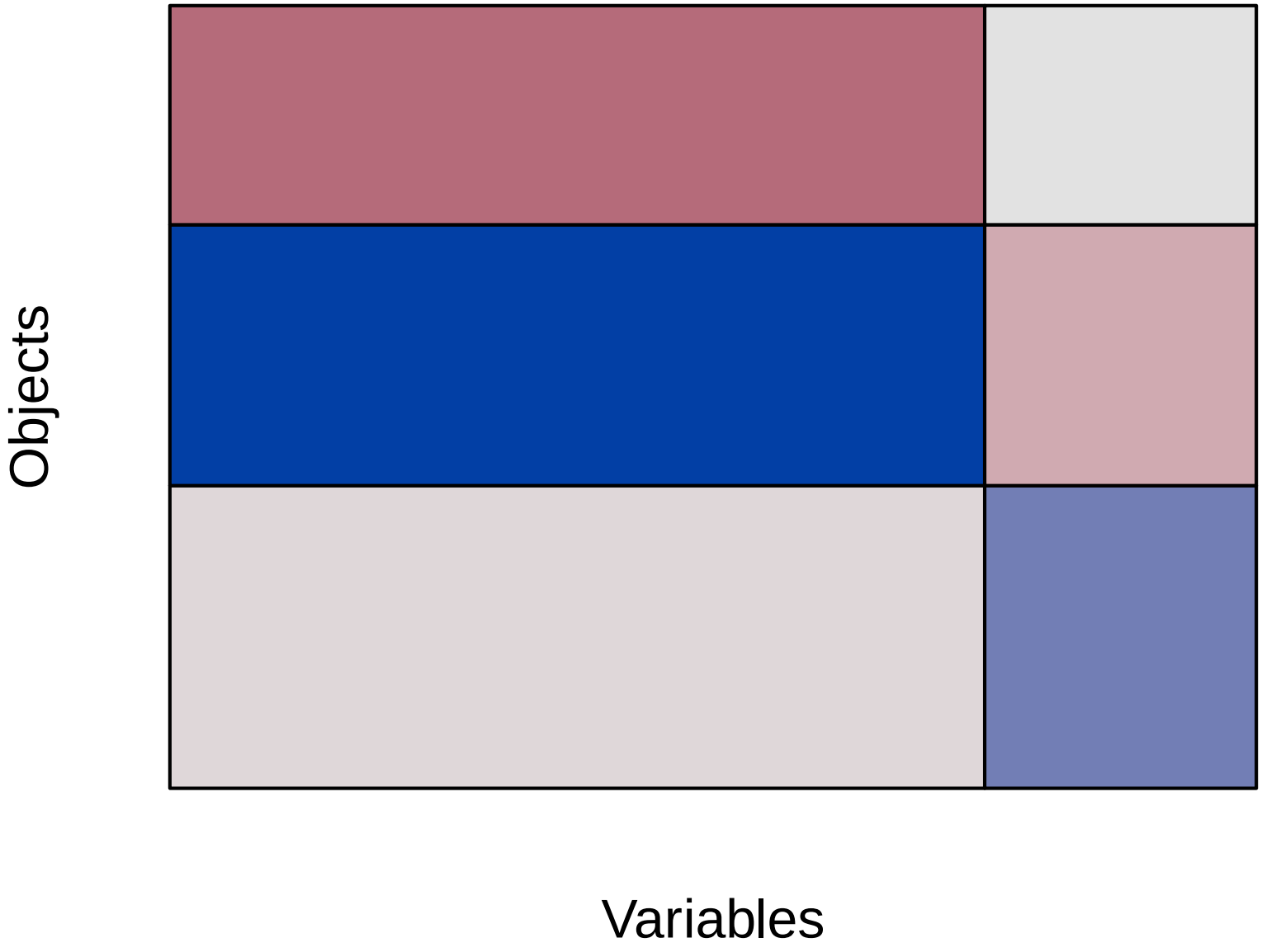
Basic example



Basic example



Basic example



A more complicated example

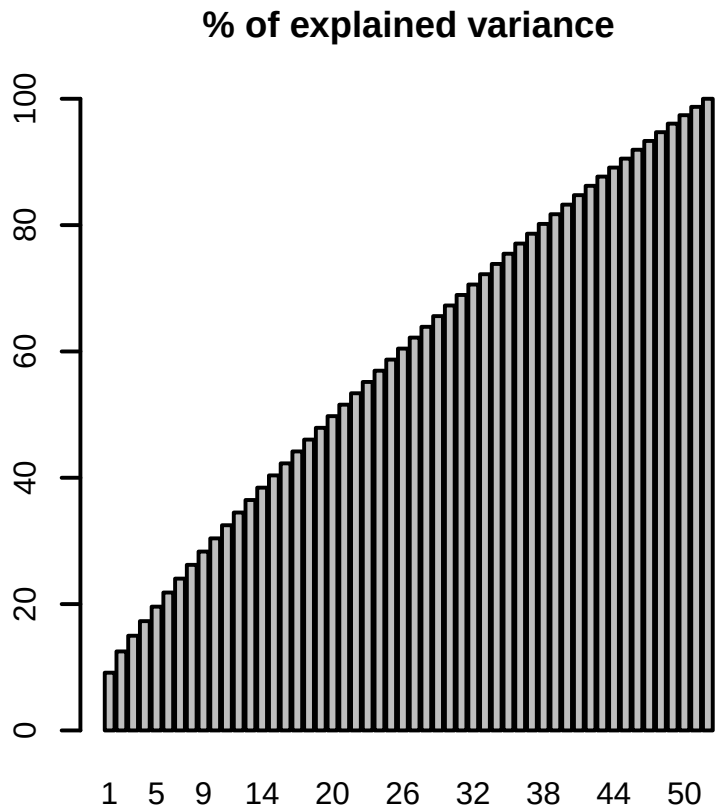
Data

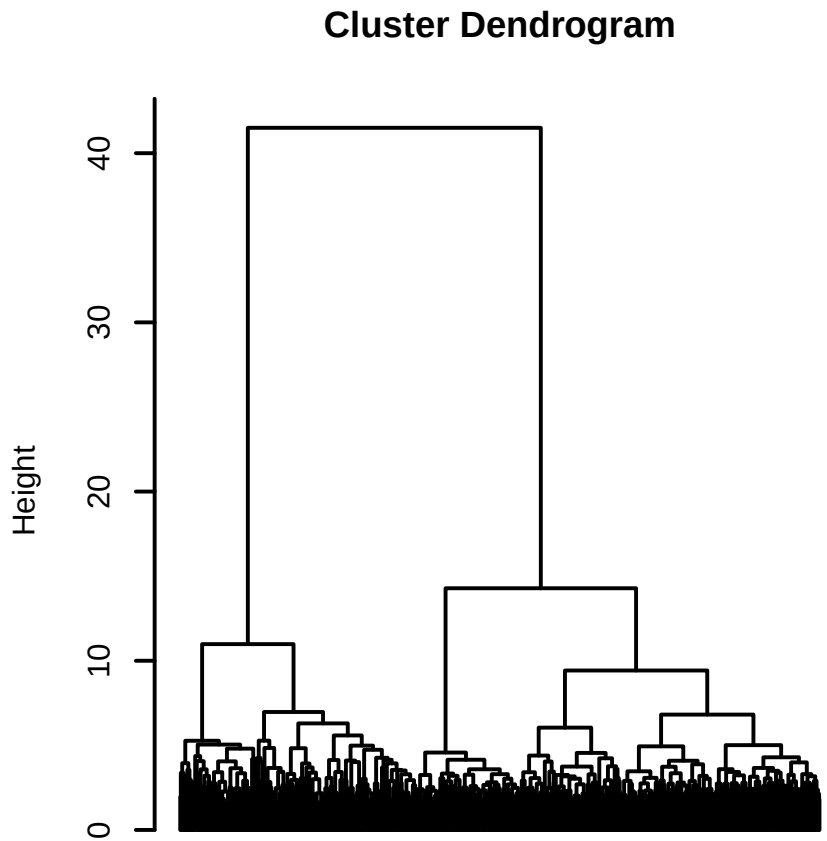
- weekly sales of 811 products over one year
- 52 variables

A more complicated example

Data

- weekly sales of 811 products over one year
- 52 variables

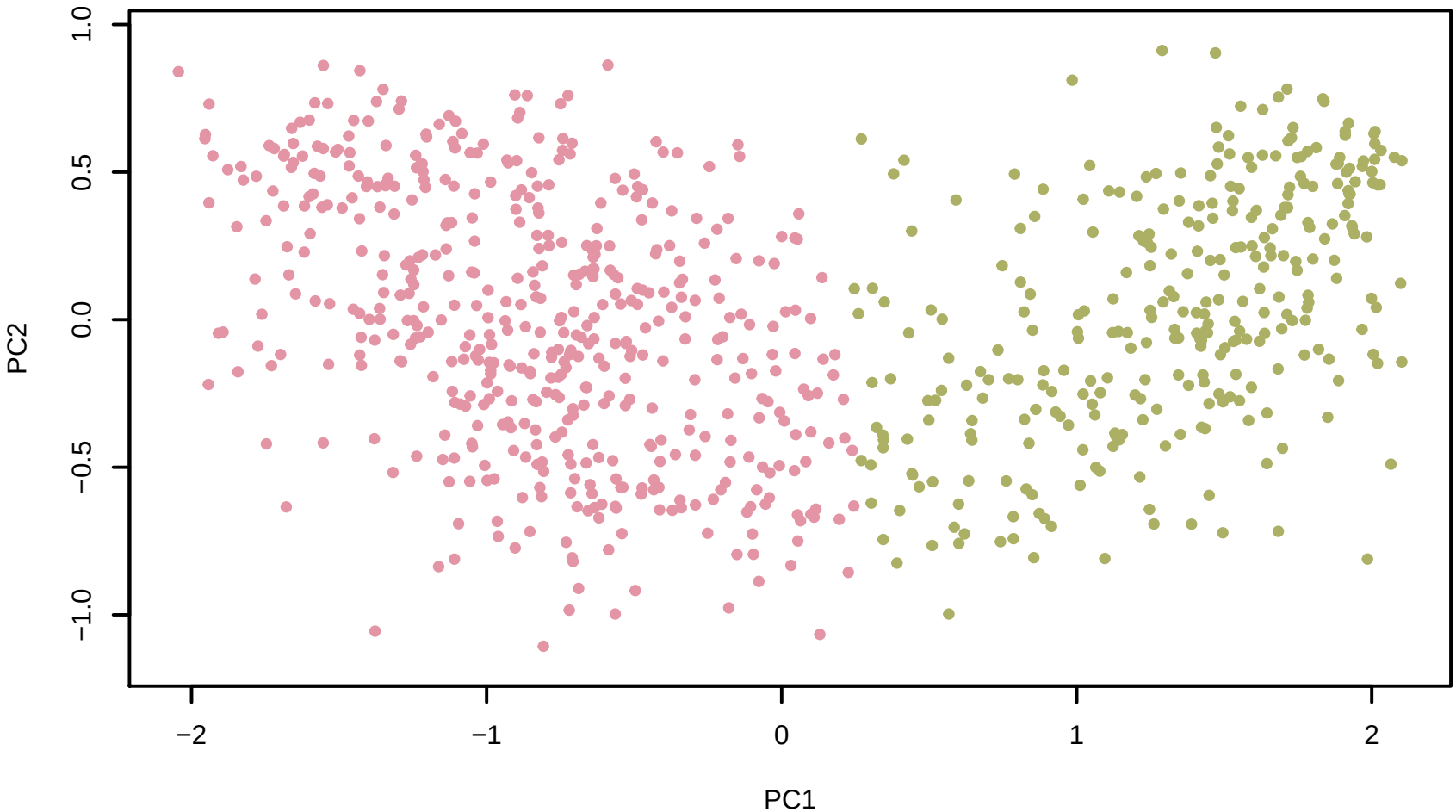




A more complicated example

Data

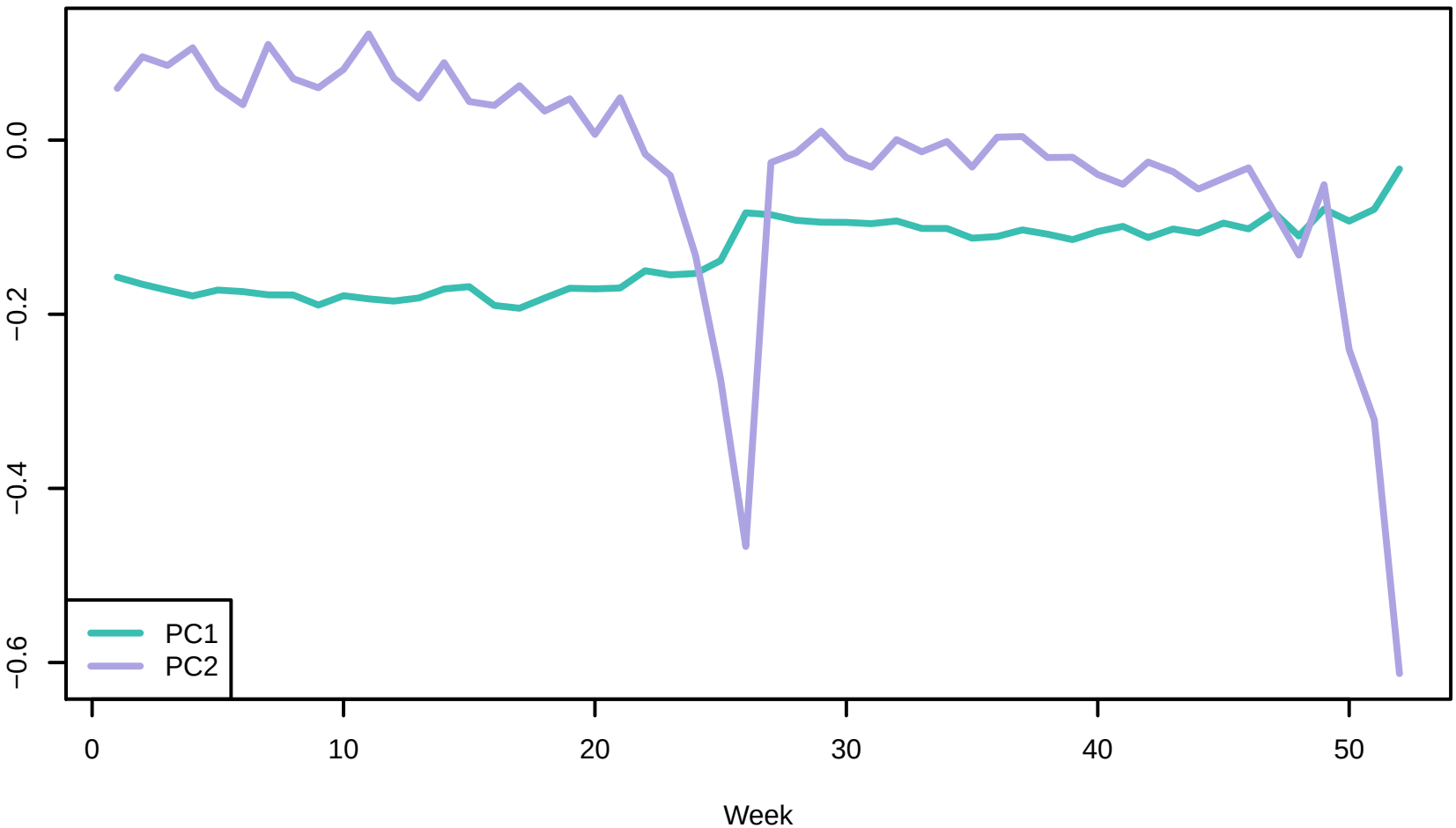
- weekly sales of 811 products over one year
- 52 variables



A more complicated example

Data

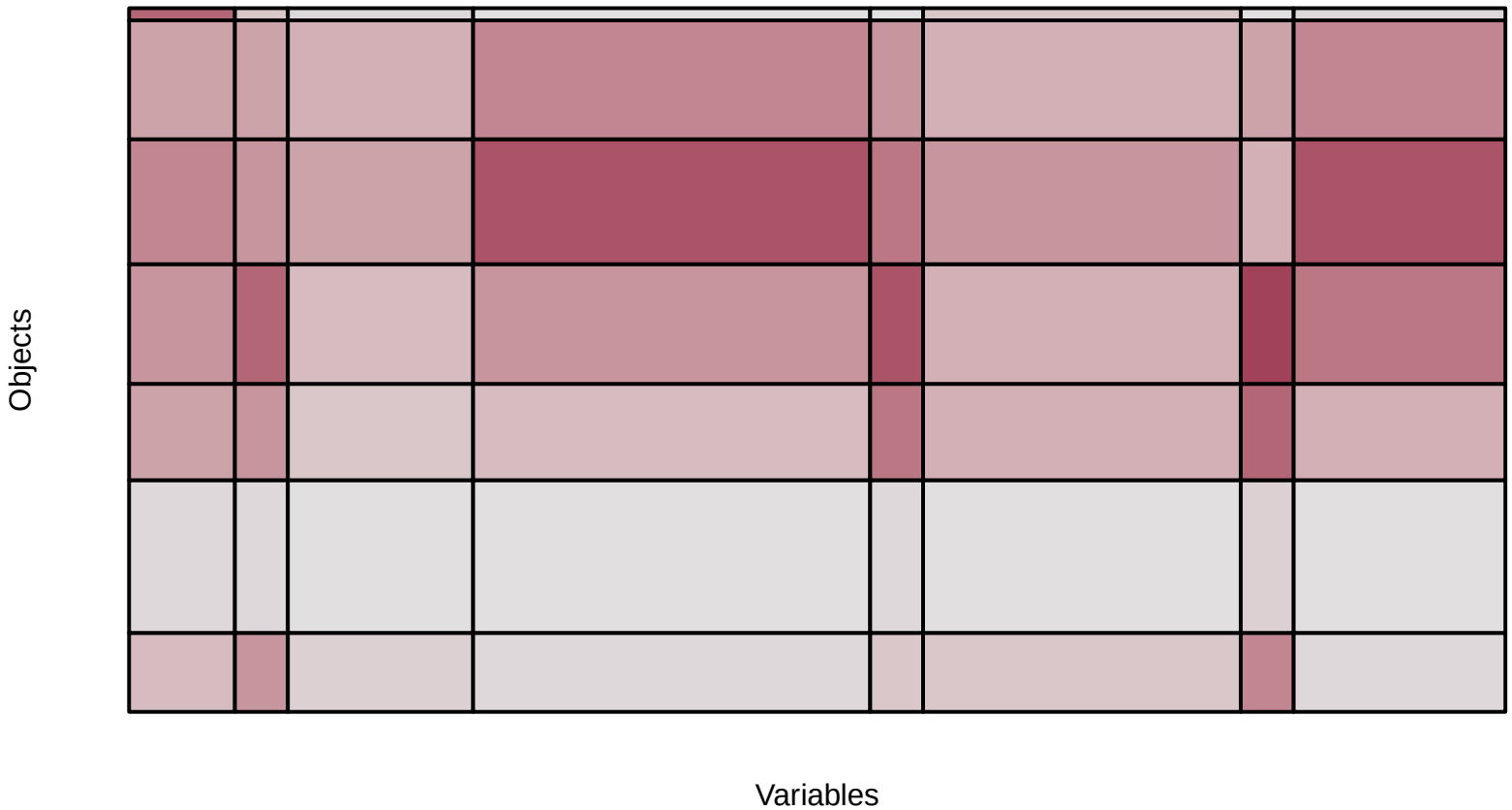
- weekly sales of 811 products over one year
- 52 variables



A more complicated example

Data

- weekly sales of 811 products over one year
- 52 variables



Extensions

Text mining

- objects: texts
- variables: words
- values: number of occurrences of a word in a text

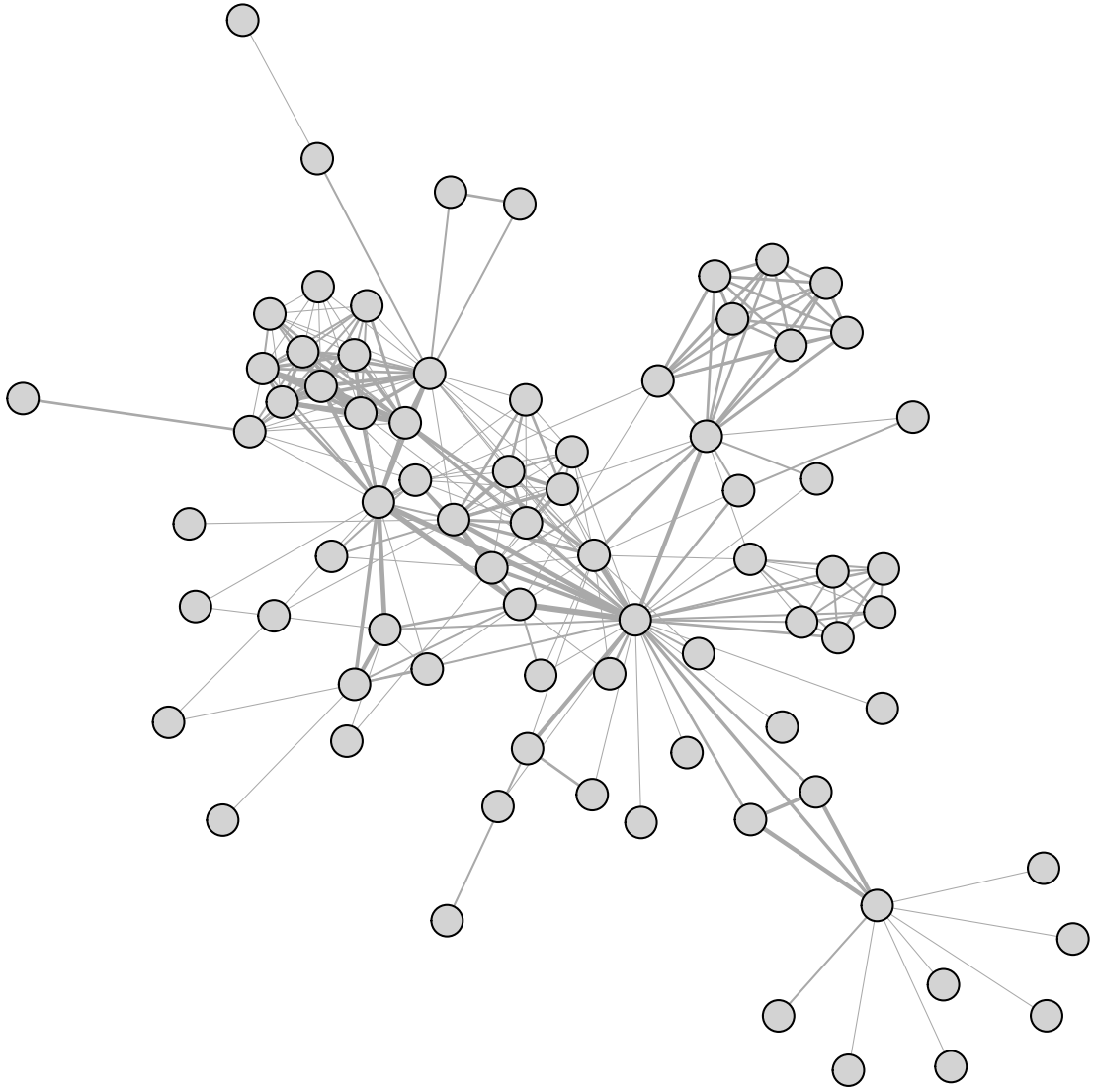
Graph mining

- objects: nodes
- variables: nodes
- value: number of arcs from one “object” node to one “variable” node

Generic framework

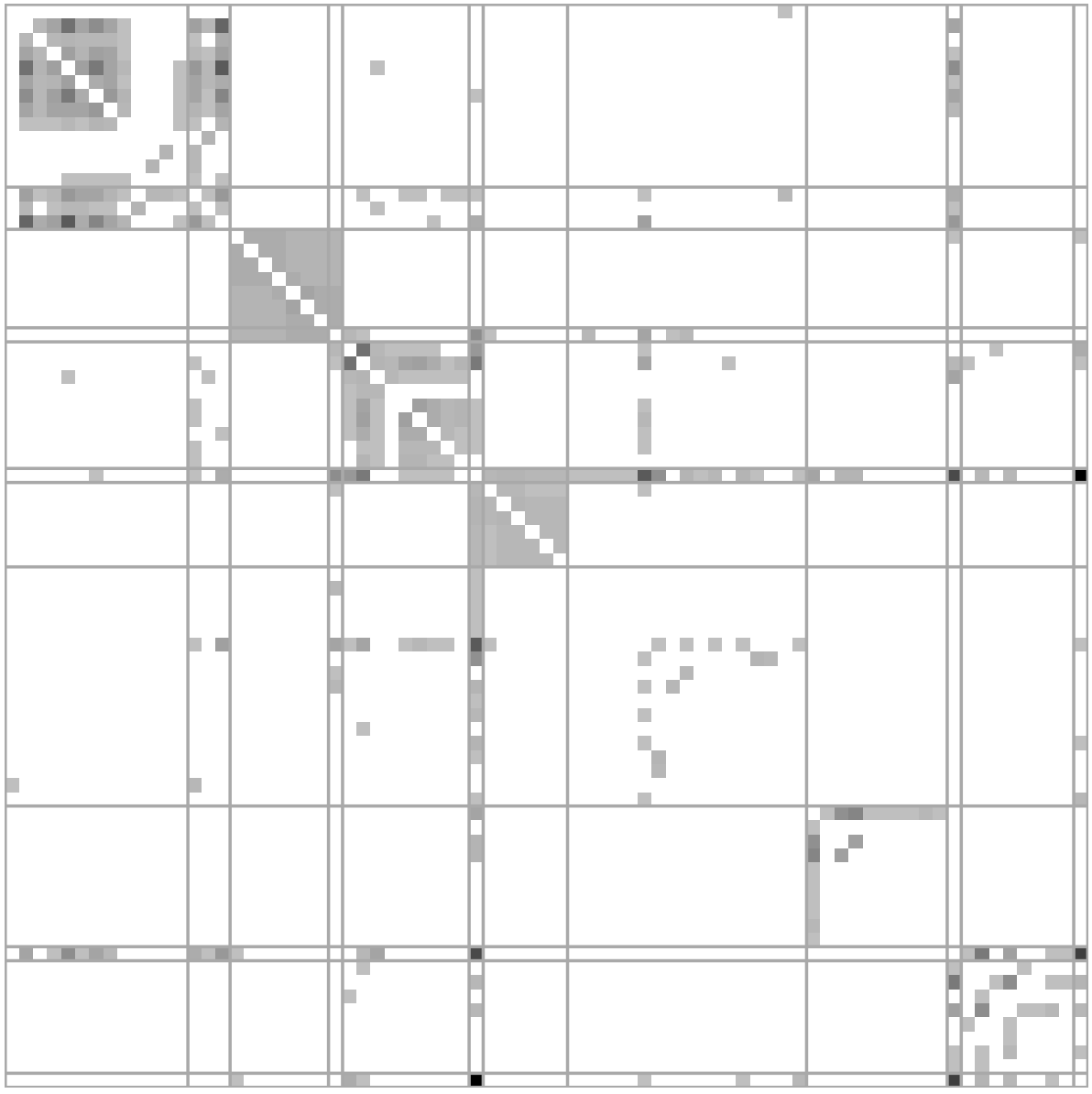
- “true” variables
- indicator or derived variables
- n-mode clustering

Example

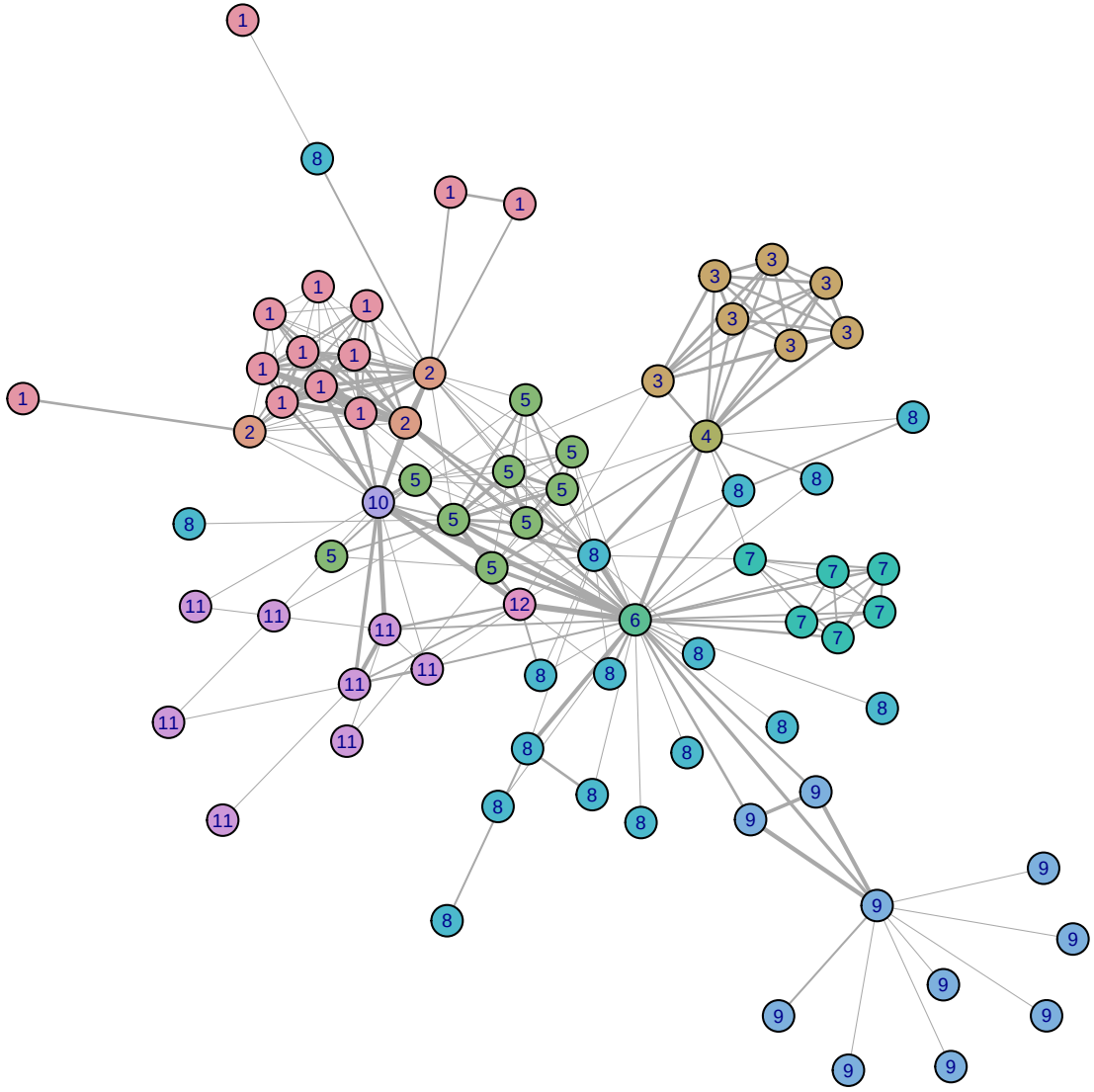


Example

Example



Example



Real Data

Phone Calls in Ivory Coast

- Cellular phone calls to Ivory Coast from other countries
- Emitters: countries ($\sim$ 190)
- Receivers: cellular antenna (1216 antennas)
- minute level timestamps
- two months of communication: roughly 13 millions of incoming calls

Raw results

- very fine clustering: 286 clusters of antennas, 33 clusters of countries and 10 temporal intervals
- greedy simplification: 12 clusters of antennas, 11 clusters of countries and 6 temporal intervals

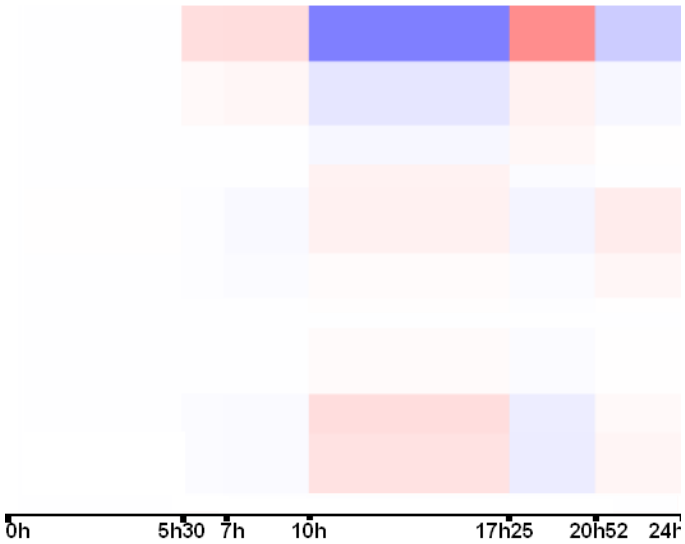
Burkina Faso

Burkina Faso

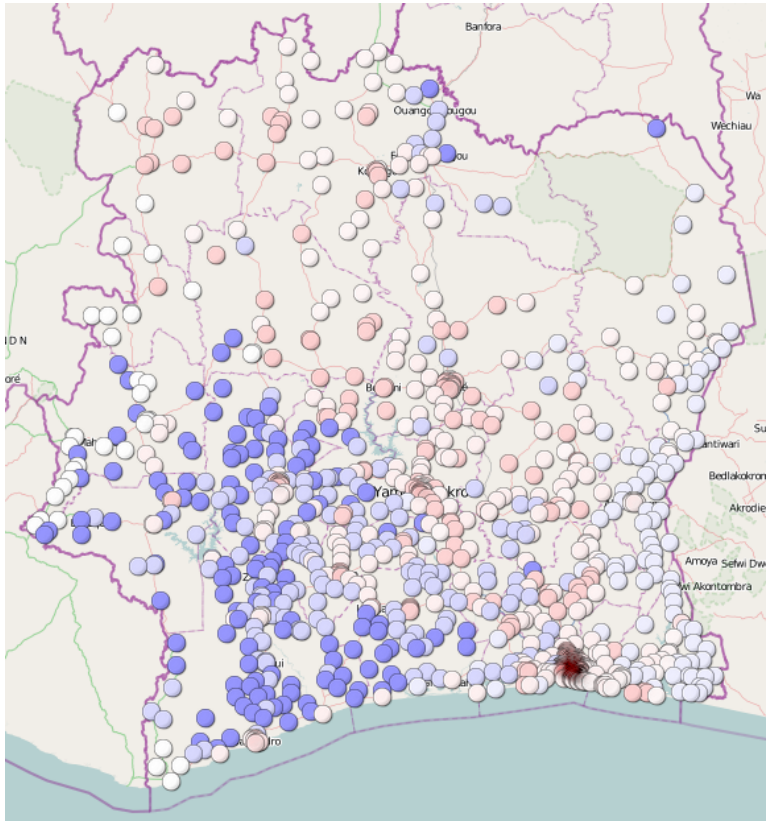
- neighbor of Ivory Coast
- provider of the first group of non Ivorian inhabitants of the Ivory Coast (roughly 15 % of the population)
- largest emitter of phone calls to Ivory Coast
- found isolated in a cluster of countries (even after simplification)

A typical result

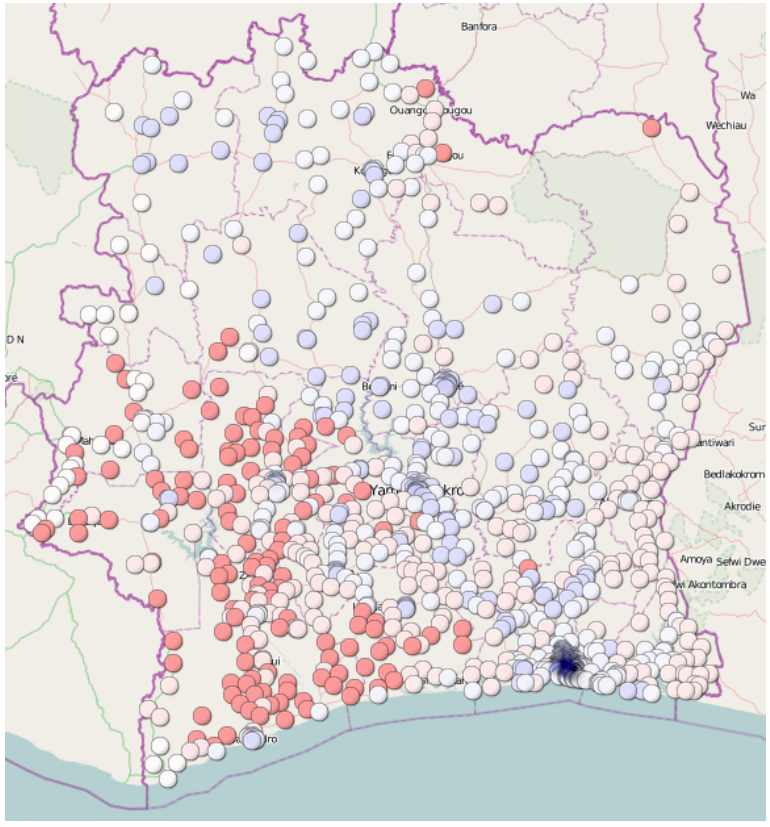
Mutual information between antenna clusters and time interval in the Burkina's cluster



Geographical view



[10*h*; 17*h*25]



[17*h*25; 20*h*52[

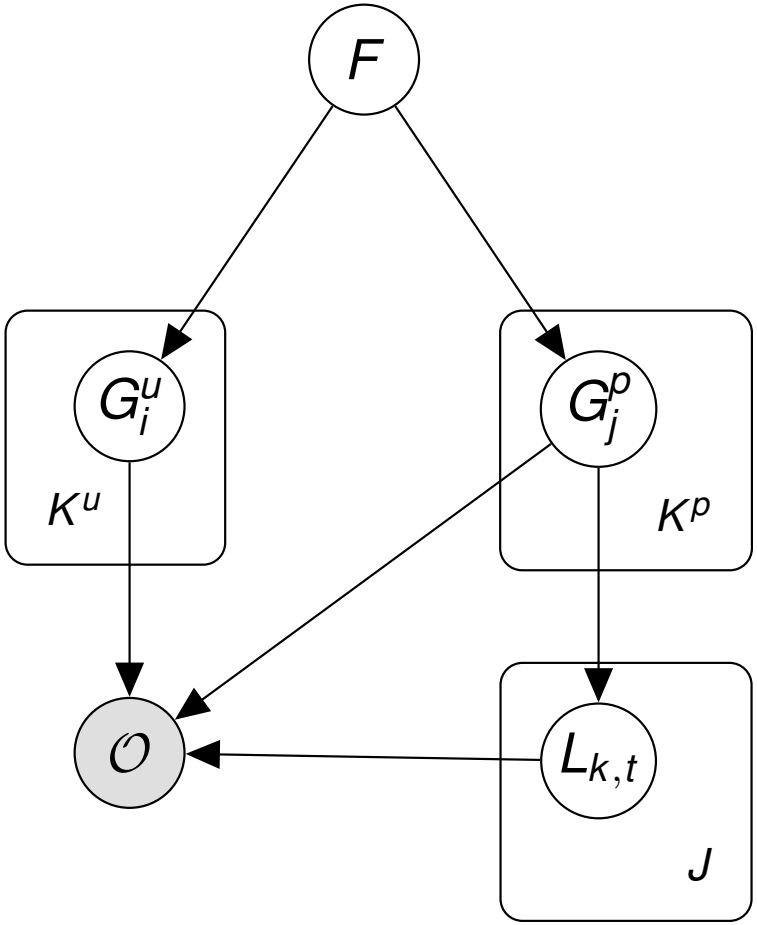
More extensions

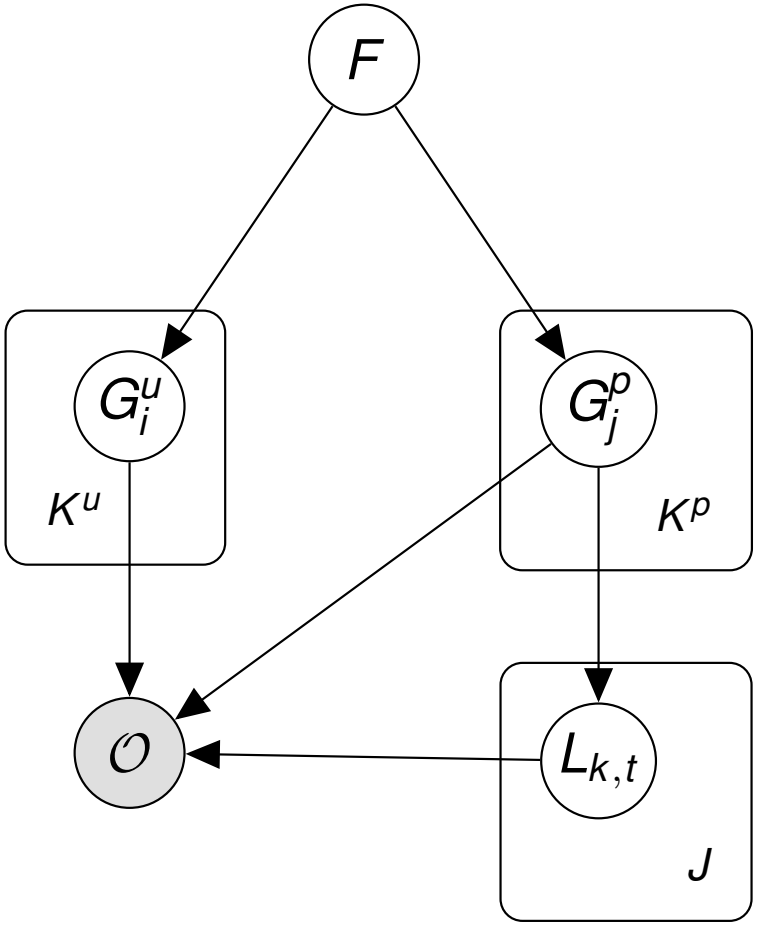
Current mixed data processng

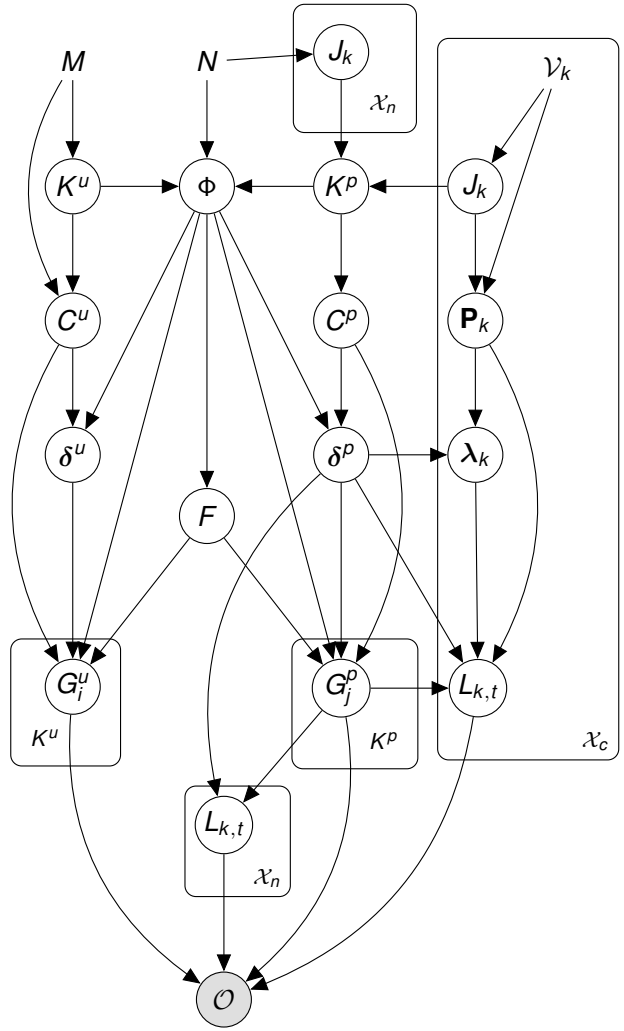
- a variable cluster gather a single type of variables
- coupling done at the object level

Proposed approach

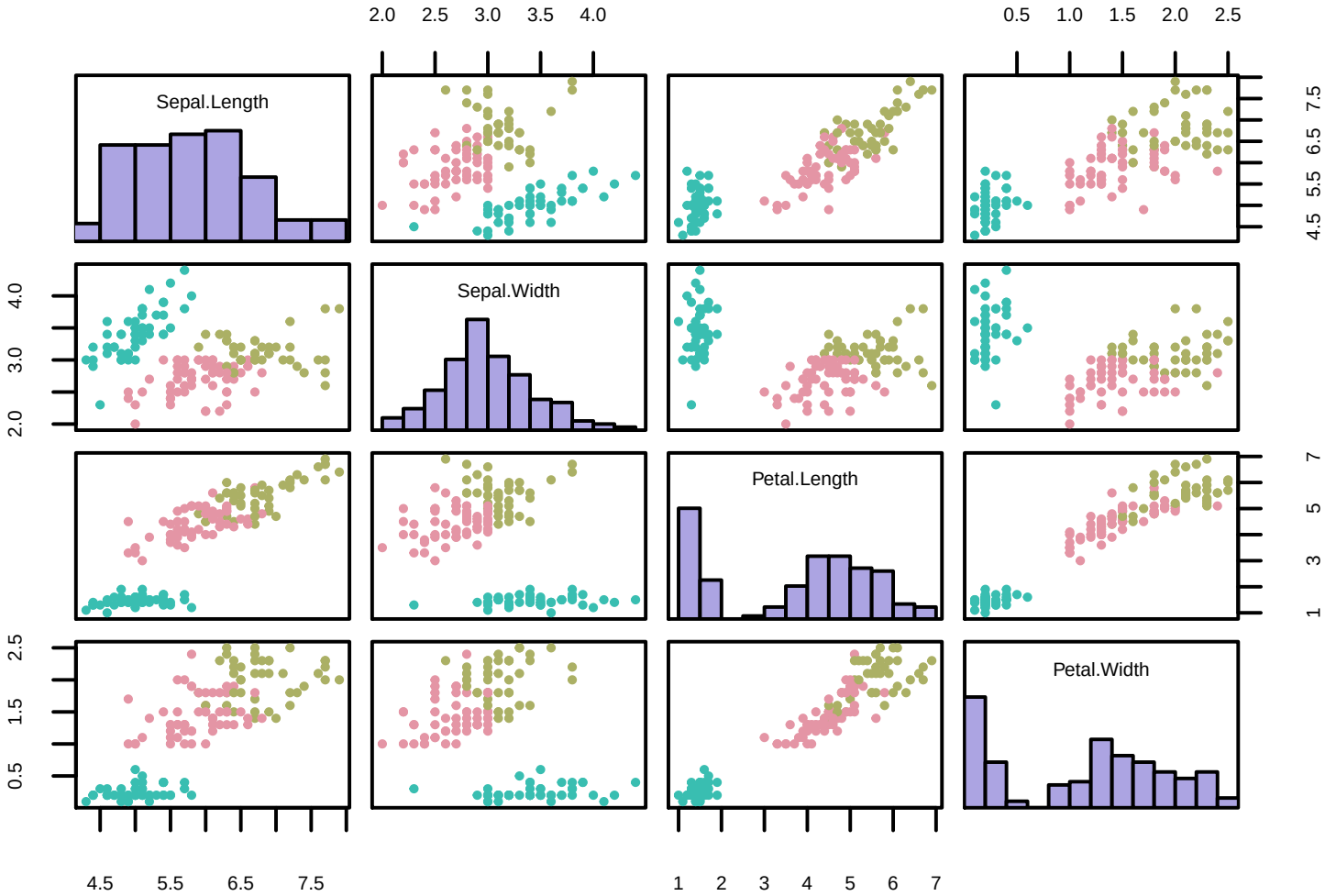
- true mixed data processing
- embedded data encoding:
  - each variable is replaced by a set of binary indicators $X_j \in C_l$
  - co-clustering on binary indicators
- full optimization: co-clustering and binarization
- ongoing work


F
G_i^u
K^u
G_j^p
K^p
$\mathcal{O}$
$L_{k,t}$
J

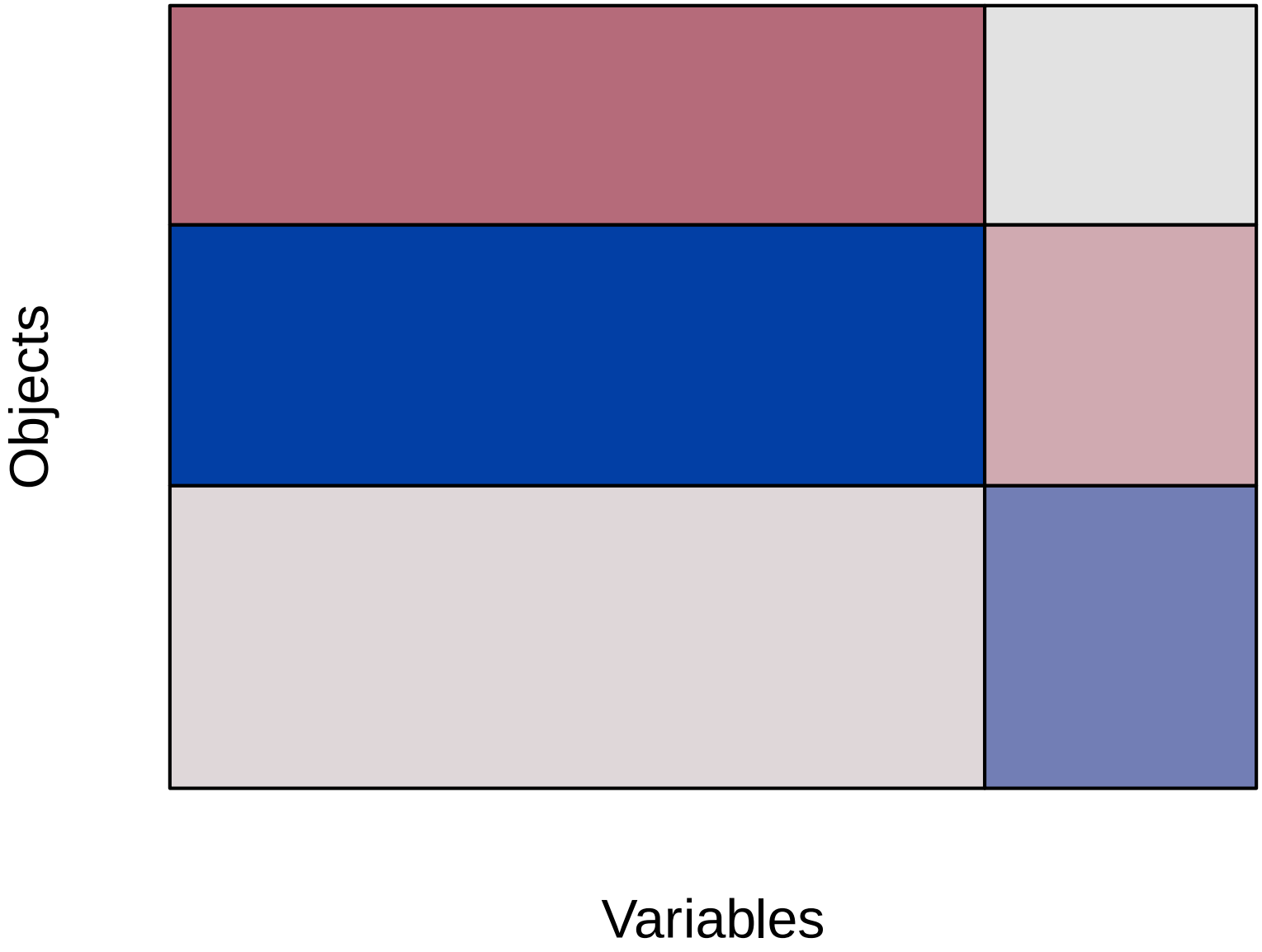

F
G_i^u
K^u
G_j^p
K^p
$\mathcal{O}$
$L_{k,t}$
J


M
N
J_k
$\mathcal{X}_n$
$\mathcal{V}_k$
K^u
Φ
K^p
J_k
C^u
C^p
$\mathbf{P}_k$
δ^u
δ^p
λ_k
F
G_i^u
K^u
G_j^p
K^p
$L_{k,t}$
$\mathcal{X}_c$
$L_{k,t}$
$\mathcal{X}_n$
$\mathcal{O}$

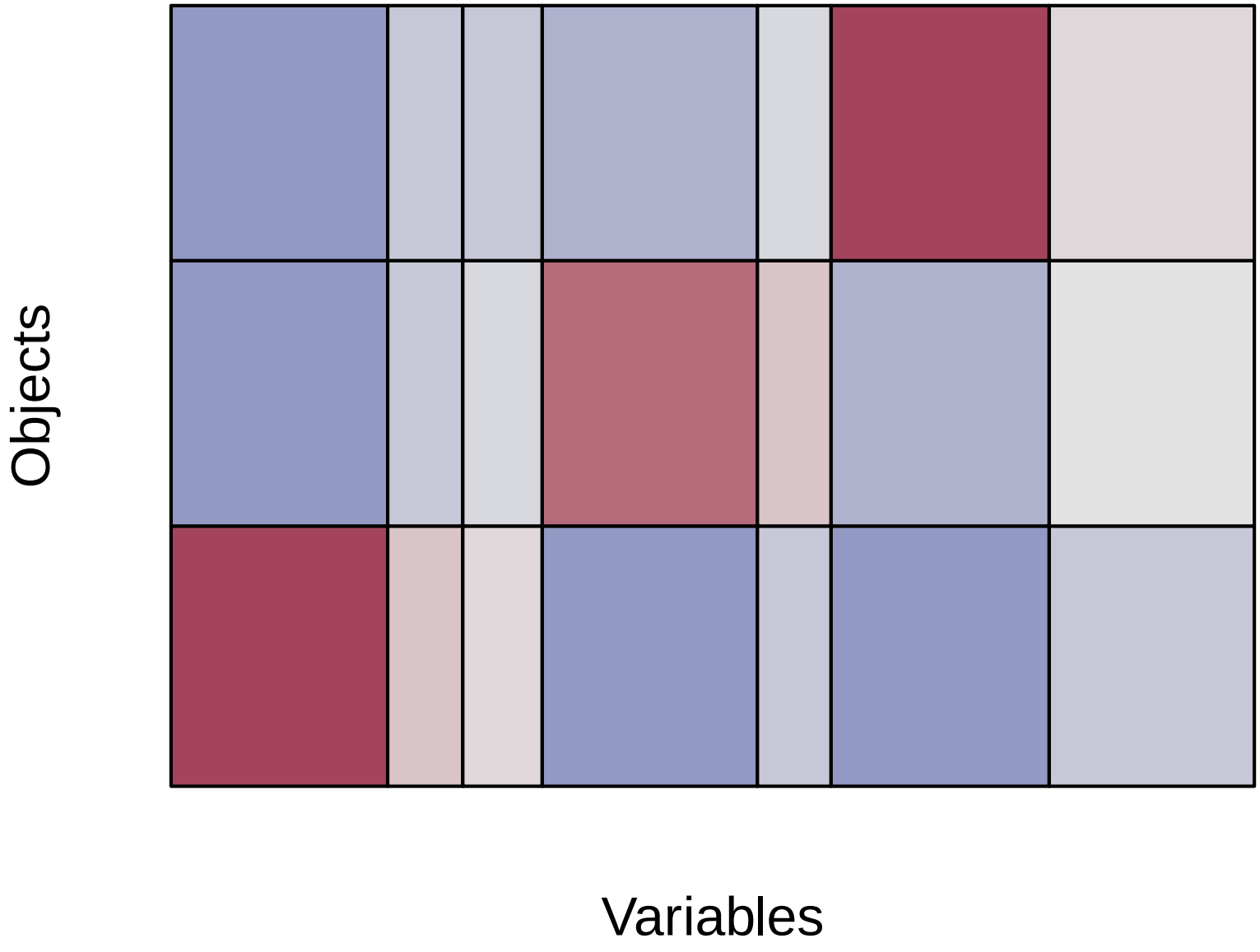
Basic example



Basic example



Basic example



Basic example

Data

- 4 numerical variables (as always)
- 1 class variable (specie of the Iris)

Results

- each value of the class variable is clustered with specific intervals of the petal variables:
  - Setosa with length ≤ 2.4 and width ≤ 0.8
  - etc.
- sepal variables are mostly left alone
- object clusters are strongly associated to the mixed variable clusters

Yet another model

Timestamped text based interaction

- fixed topic model
- role based use of the topics
- change detection

Yet another model

Timestamped text based interaction

- fixed topic model
- role based use of the topics
- change detection

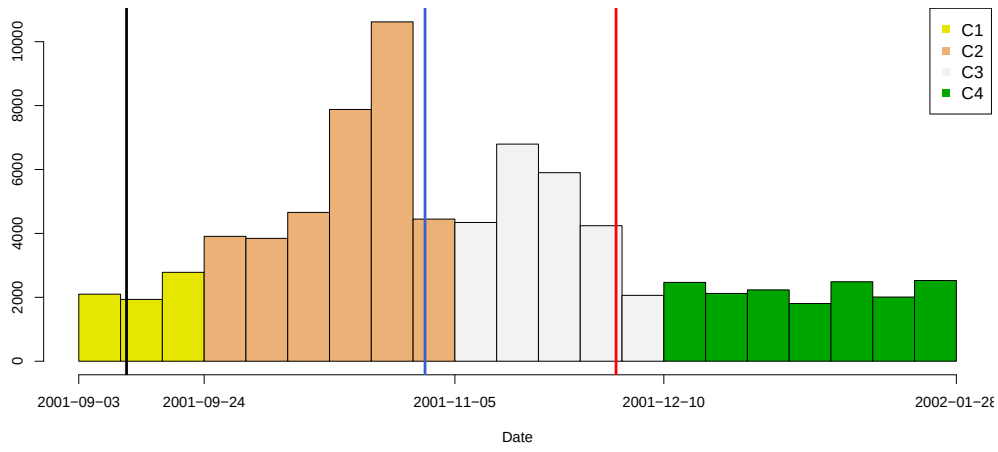
Enron data set

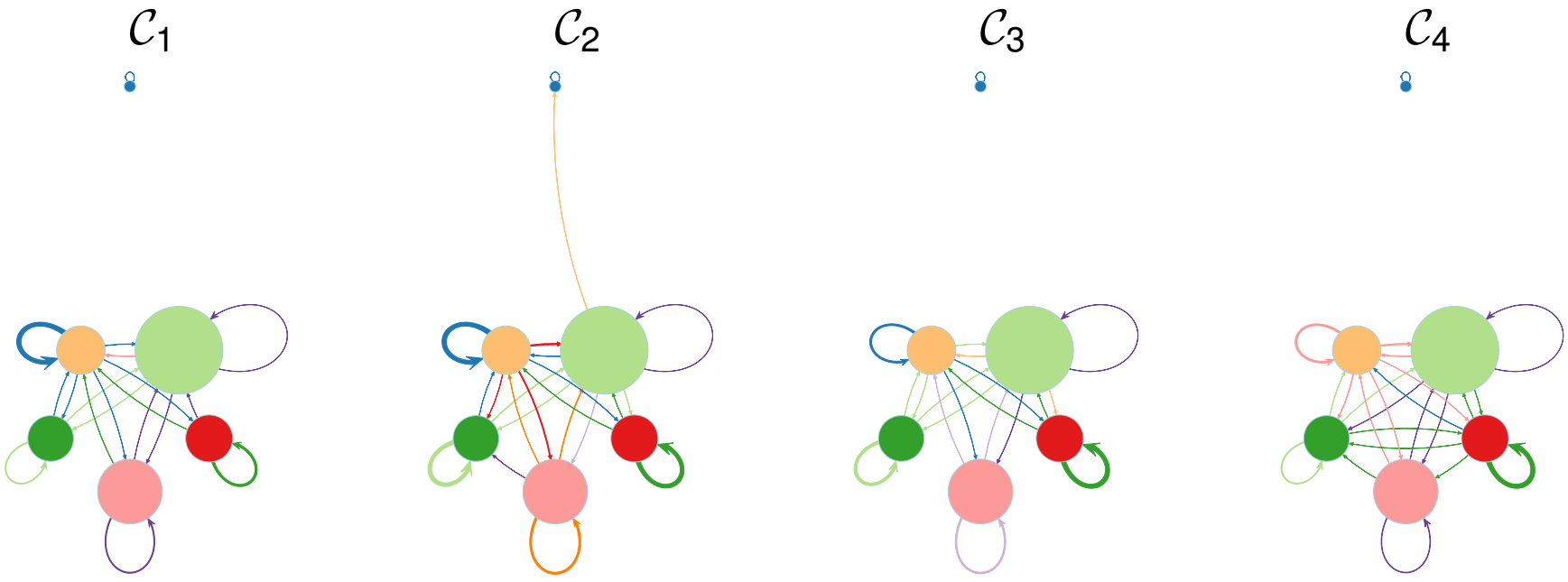
E-mail exchanges between 149 Enron employees.

4 time clusters

Key dates (vertical lines):

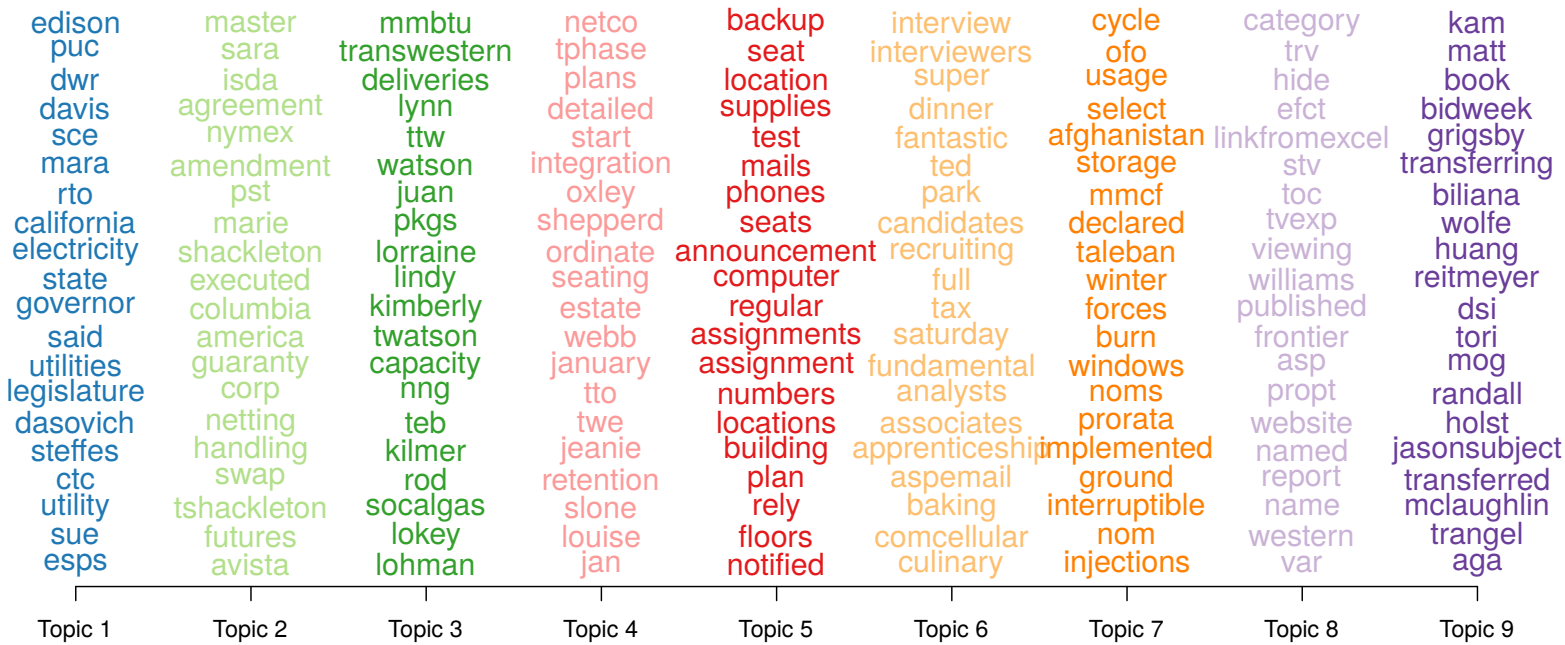
- September, 11th, 2001
- October, 31st, 2001: Enron investigated for fraud
- December, 2nd, 2001: Enron failed bankruptcy





- 6 Groups of nodes
- 9 Topics

More representative words for each topic



Summary

Generative models

- very flexible probabilistic modeling
- models for mixed data, graphs, texts, etc.
- model selection, missing data, simulation, etc.

Personal work

- n mode clustering from graph data
- dynamic graph block models
- dynamic text based interaction
- mixed data co-clustering

Future work

More models

- more complex text models (beyond topic models)
- ad hoc models for specific applications:
  - homonyms in old texts
  - import structure for cultural diffusion analysis
- null models

Better models and tools

- fast models:
  - approximation
  - online approaches
- interactive visualization

References I

M. Boullé, R. Guigourès, and F. Rossi.

Nonparametric hierarchical clustering of functional data.
In F. Guillet, B. Pinaud, G. Venturini, and D. A. Zighed, editors, *Advances in Knowledge Discovery and Management*, volume 527 of *Studies in Computational Intelligence*, pages 15–35. Springer International Publishing, 2 2014.
doi: 10.1007/978-3-319-02999-3_2.
URL http://hal.archives-ouvertes.fr/hal-01017553.

M. Corneli, P. Latouche, and F. Rossi.

Modelling time evolving interactions in networks through a non stationary extension of stochastic block models.
In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2015)*, pages 1590–1591, Paris, France, 8 2015a. ACM.
doi: 10.1145/2808797.2809348.
URL http://hal.archives-ouvertes.fr/hal-01263540.

M. Corneli, P. Latouche, and F. Rossi.

Exact icl maximization in a non-stationary time extension of the latent block model for dynamic networks.
In *Proceedings of the 23-th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015)*, pages 225–230, Bruges, Belgique, 4 2015b.
URL http://hal.archives-ouvertes.fr/hal-01163367.

References II

M. Corneli, P. Latouche, and F. Rossi.

Modelling time evolving interactions in networks through a non stationary extension of stochastic block models.
In *Actes des 47èmes Journées de Statistique de la SFdS*, Lille, France, 6 2015c. Société Française de Statistique.
URL http://hal.archives-ouvertes.fr/hal-01195688.

M. Corneli, P. Latouche, and F. Rossi.

Block modelling in dynamic networks with non-homogeneous poisson processes and exact ICL.
Social Network Analysis and Mining, 6(1):55, 2016a.
ISSN 1869-5469.
doi: 10.1007/s13278-016-0368-3.
URL http://hal.archives-ouvertes.fr/hal-01468548.

M. Corneli, P. Latouche, and F. Rossi.

Exact ICL maximization in a non-stationary temporal extension of the stochastic block model for dynamic networks.
Neurocomputing, 192:81 – 91, 6 2016b.
ISSN 0925-2312.
doi: 10.1016/j.neucom.2016.02.031.
URL http://hal.archives-ouvertes.fr/hal-01312596.

References III

M. K. El Mahrsi, R. Guigourès, F. Rossi, and M. Boullé.

Co-clustering network-constrained trajectory data.
In F. Guillet, B. Pinaud, G. Venturini, and D. A. Zighed, editors, *Advances in Knowledge Discovery and Management*, volume 615 of *Studies in Computational Intelligence*, pages 19–32. Springer International Publishing, 9 2015.
doi: 10.1007/978-3-319-23751-0_2.
URL http://hal.archives-ouvertes.fr/hal-01222649.

R. Guigourès, M. Boullé, and F. Rossi.

A triclustering approach for time evolving graphs.
In *Co-clustering and Applications, IEEE 12th International Conference on Data Mining Workshops (ICDMW 2012)*, pages 115–122, Brussels, Belgium, décembre 2012.
doi: 10.1109/ICDMW.2012.61.
URL http://hal.archives-ouvertes.fr/hal-00773133.

R. Guigourès, M. Boullé, and F. Rossi.

A study of the spatio-temporal correlations in mobile calls networks.
In F. Guillet, B. Pinaud, G. Venturini, and D. A. Zighed, editors, *Advances in Knowledge Discovery and Management*, volume 615 of *Studies in Computational Intelligence*, pages 3–17. Springer International Publishing, 9 2015a.
doi: 10.1007/978-3-319-23751-0_1.
URL http://hal.archives-ouvertes.fr/hal-01222355.

References IV

R. Guigourès, D. Gay, M. Boullé, F. Clérot, and F. Rossi.

Country-scale exploratory analysis of call detail records through the lens of data grid models.
In A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, J. Cardoso, and M. Spiliopoulou, editors, *Machine Learning and Knowledge Discovery in Databases (Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD 2015)*, volume 9286 of *Lecture Notes in Computer Science*, pages 37–52, Porto, Portugal, 9 2015b. Springer International Publishing.

doi: 10.1007/978-3-319-23461-8_3.
URL http://hal.archives-ouvertes.fr/hal-01261122.

R. Guigourès, M. Boullé, and F. Rossi.

Discovering patterns in time-varying graphs: a triclustering approach.
Advances in Data Analysis and Classification, pages 1–28, 2016.
ISSN 1862-5347.
doi: 10.1007/s11634-015-0218-6.
URL http://hal.archives-ouvertes.fr/hal-01356993.