Dealing with missing data through mixture models

Vincent Vandewalle[1,2,3]
Joint work with Christophe Biernacki[3,4,5]

154th ICB Seminar on "Statistics and clinical practice"
Warsaw
11 may 2017

1 Université de Lille 2 DROIT ET SANTÉ
2 équipe d'accueil EA 2694
3 Inria informatics mathematics
4 cnrs dépasser les frontières
5 Université de Lille 1 SCIENCES ET TECHNOLOGIES

Overview

Data (1/2)

Many variables collected :

- Heterogeneous data (of various natures)
- Missing data
- Incertain data (by interval)

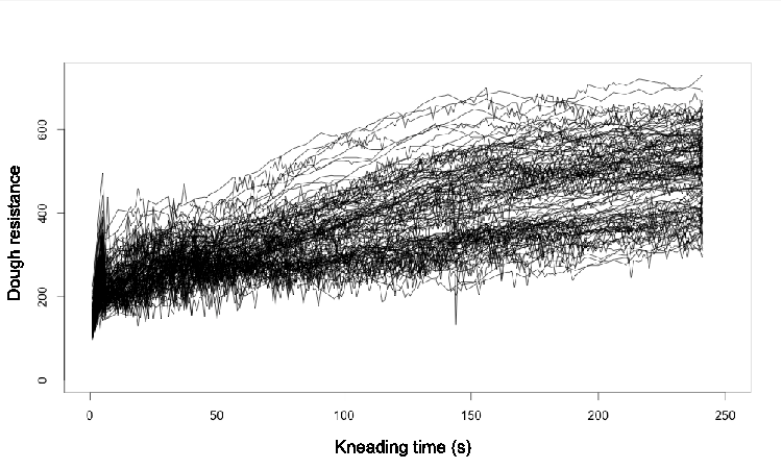
Heterogeneous, missing, incertain

Units $\mathbf{x}^O \in \mathcal{X}$			
?	0.5	?	5
0.3	0.1	green	3
0.3	0.6	{red, green}	3
0.9	[0.25 0.45]	red	?
↓	↓	↓	↓
continuous	continuous	categorical	integer

Data (2/2)

And also :

- Rank data
- Directional data
- Ordinal data
- Functional data
- Network data
- ...



Imputation of missing data

Motivation for imputing missing values

Most of standard statistical methods cannot cope with missing data $\Rightarrow$ need to complete the dataset

Single imputation

- $E[\mathbf{x}^M]$: average
- $\arg\max p(\mathbf{x}^M)$: mode
- $E[\mathbf{x}^M|\mathbf{x}^O]$: conditional expectation (regression, NPALS, ...)
- Underestimation of the variability related to missing values

Multiple imputation

- Replace missing value by many possibles values ideally coming from $p(\mathbf{x}^M|\mathbf{x}^O)$
- Variability inherent to missing data taken into account
- Necessity to consider many version of the dataset when performing the learning

The fully conditional specification approach (Van Buuren & *al.* (2006)

Presentation

- Specify a conditional model $p(x_j|\mathbf{x}_{-j}, \theta_j)$ for each j (linear regression, logistic regression, ...)
- At each iteration
  1. Draw for each variable j the missing value given the observed values, the simulated missing values and the current parameter
  2. Update the parameters of the conditional models

Advantage

- Missing data variability taken into account
- Simulated missing values expected to come from $p(\mathbf{x}^M|\mathbf{x}^O)$
- Good performance in practice : implemented in the R package `mice`.

Supervised classification (1/3)

- Data : learning data $\mathcal{D} = (\mathbf{x}^O, \mathbf{z})$
  - n units : $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) = (\mathbf{x}^O, \mathbf{x}^M)$ belonging to the space $\mathcal{X}$
  - Observed variables $\boldsymbol{x}^O$
  - Missing variables $\boldsymbol{x}^M$
  - Partition in K clusters $G_1, \ldots, G_K$: $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$

$$\mathbf{x}_i \in G_k \quad \Leftrightarrow \quad z_{ik} = 1 \text{ et } \forall h \neq k, z_{ih} = 0$$

- Goal : learn a classification rule r $\mathcal{D}$

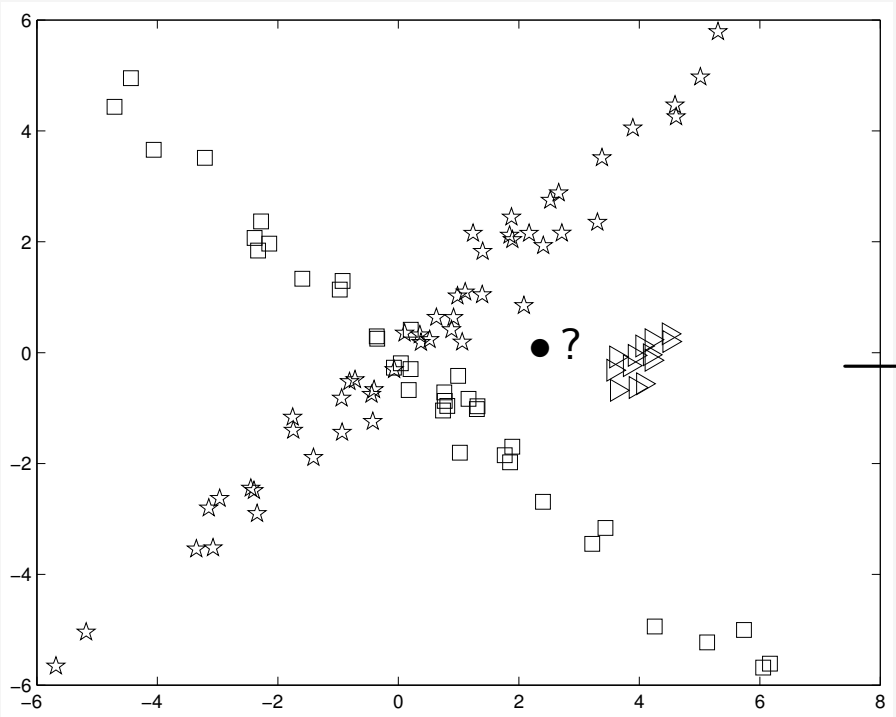
$$\begin{array}{rccc} r : & \mathcal{X} & \longrightarrow & \{1, \ldots, K\} \\ & \mathbf{x}^O_{n+1} & \longmapsto & r(\mathbf{x}^O_{n+1}). \end{array}$$

Supervised classification (2/3)

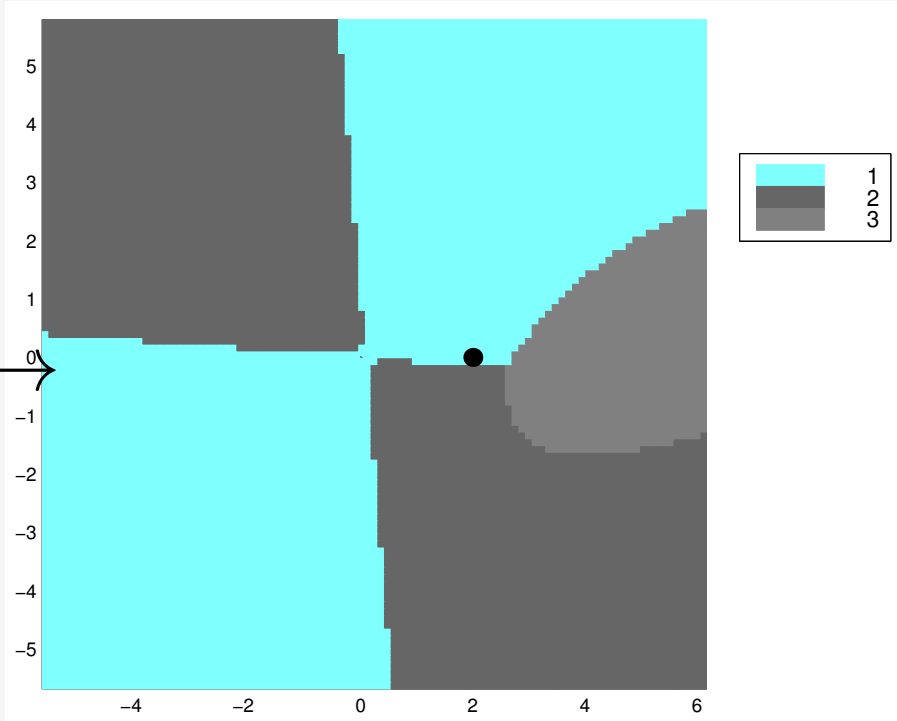
Heterogeneous, missing, uncertain

Units $\mathbf{x}^O$				Partition $\mathbf{z}$			$\Leftrightarrow$	Class
?	0.5	red	5	0	1	0	$\Leftrightarrow$	G_2
0.3	0.1	vert	3	1	0	0	$\Leftrightarrow$	G_1
0.3	0.6	{red, green}	3	1	0	0	$\Leftrightarrow$	G_1
0.9	[0.25 0.45]	red	?	0	0	1	$\Leftrightarrow$	G_3
↓	↓	↓	↓					
continuous	continuous	categorical	integer					

Supervised classification (3/3)



$(\mathbf{x}^O, \mathbf{z}^O)$ et $\mathbf{x}^O_{n+1}$



$\hat{r}$ et $\hat{\mathbf{z}}^0_{n+1}$

Semi-supervised classification (1/3)

- Data : learning data $\mathcal{D} = (\mathbf{x}^O, \mathbf{z}^O)$
  - n units : $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) = (\mathbf{x}^O, \mathbf{x}^M)$ belonging to the space $\mathcal{X}$
  - Observed varibles $\boldsymbol{x}^O$
  - Missing variables $\boldsymbol{x}^M$
  - Partition : $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \mathbf{z}_n) = (\mathbf{z}^O, \mathbf{z}^M)$
  - Observed class $\boldsymbol{z}^O$
  - Missing class $\boldsymbol{z}^M$
- Goal : estimate the classification rule r from $\mathcal{D}$

$$\begin{array}{rccc} r: & \mathcal{X} & \longrightarrow & \{1, \ldots, K\} \\ & \mathbf{x}^O_{n+1} & \longmapsto & r(\mathbf{x}^O_{n+1}). \end{array}$$

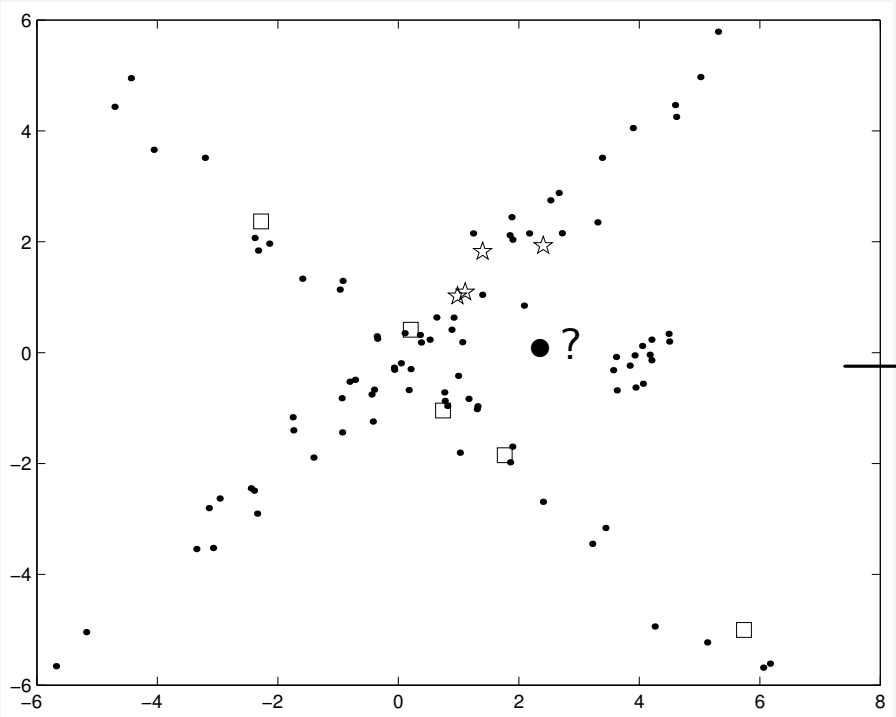
- Idea : $\mathbf{x}$ is cheaper than $\mathbf{z}$ thus $n \gg n_\ell$, n_ℓ number of labeled units.

Semi-supervised classification (2/3)

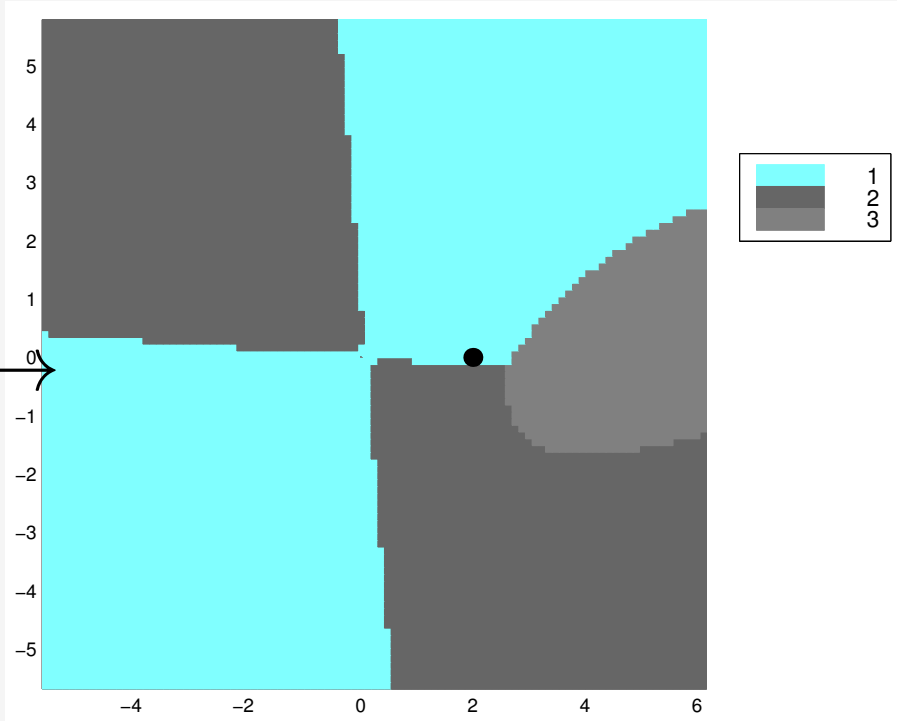
heterogeneous, missing, uncertain

Units $\mathbf{x}^O$				Partition $\mathbf{z}^O$			$\Leftrightarrow$	Class
?	0.5	red	5	0	?	?	$\Leftrightarrow$	G_2 ou G_3
0.3	0.1	vert	3	1	0	0	$\Leftrightarrow$	G_1
0.3	0.6	{red,green}	3	?	?	?	$\Leftrightarrow$	? ? ?
0.9	[0.25 0.45]	red	?	0	0	1	$\Leftrightarrow$	G_3
↓	↓	↓	↓					
continue	continuous	categorical	integer					

Semi-supervised classification (3/3)



$(\mathbf{x}^O, \mathbf{z}^O)$ et $\mathbf{x}^O_{n+1}$



$\hat{r}$ et $\hat{\mathbf{z}}^O_{n+1}$

Unsupervised classification (1/3)

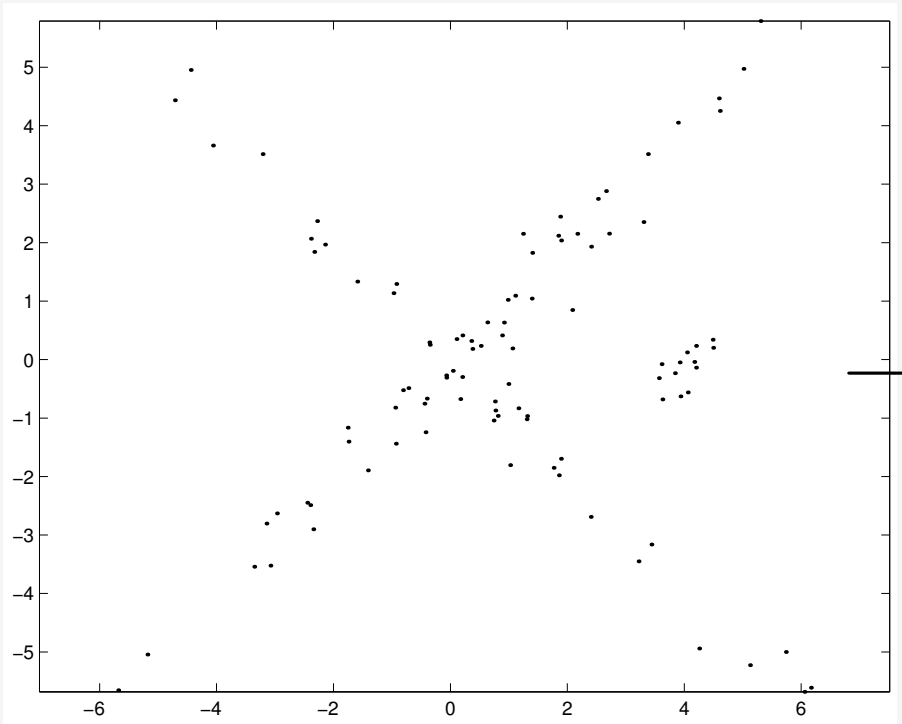
- Data : learning data $\mathcal{D} = \mathbf{x}^O$, with $\mathbf{z}^O = \emptyset$
- Goal : estimate the partition $\mathbf{z}$ and the number of clusters K
- Usually usually known as : clustering

Unsupervised classification (2/3)

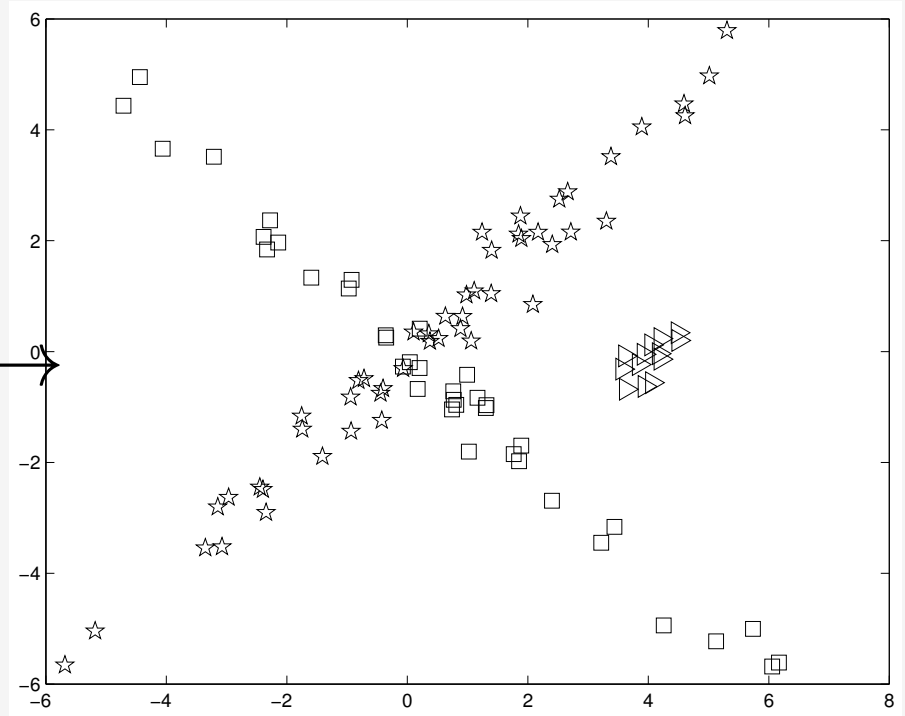
heterogneous, missing, uncertain

Units $\mathbf{x}^O$				Partition $\mathbf{z}^O$			$\Leftrightarrow$	Cluster
?	0.5	red	5	?	?	?	$\Leftrightarrow$	? ? ?
0.3	0.1	green	3	?	?	?	$\Leftrightarrow$	? ? ?
0.3	0.6	{red, green}	3	?	?	?	$\Leftrightarrow$	? ? ?
0.9	[0.25 0.45]	red	?	?	?	?	$\Leftrightarrow$	? ? ?
↓	↓	↓	↓					
continuous	continuous	categorical	interger					

Unsupervised classification (3/3)



$\mathbf{x}^O$



$(\mathbf{x}^O, \hat{\mathbf{z}})$

Usuals solutions (1/2)

Two types of models

- Generatives models
  - Modeling of $p(\boldsymbol{x}, \boldsymbol{z})$
  - Which implies a direct modeling of $p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z})$
  - Missing data easily taken into account on $\boldsymbol{z}$ and $\boldsymbol{x}$
- Predictive models
  - Modeling of $p(\boldsymbol{z}|\boldsymbol{x})$ eventually only $\mathbf{1}_{\{p(\boldsymbol{z}|\boldsymbol{x})>1/2\}}$, or only the rank on $p(\boldsymbol{z}|\boldsymbol{x})$
  - Avoids assumptions on $p(\boldsymbol{x})$, and thus limits the modeling errors
  - Difficulty to take into account missing data on $\boldsymbol{z}$ and/or $\boldsymbol{x}$

Solutions usuelles (2/2)

- Missing / uncertain data : multiple inputation
- Mixed data : empirical recoding

How to deal with heterogenous, missing, uncertain data in an integrated way ?

Solutions usuelles (2/2)

- Missing / uncertain data : multiple inputation
- Mixed data : empirical recoding

How to deal with heterogenous, missing, uncertain data in an integrated way ?
By using the mixture models.

Mixture models (1/2)

- Rigorous definition of a cluster :

$$\mathbf{x}_1 \in G_k \quad \Leftrightarrow \quad \mathbf{x}_1 \text{ is a realization of } \mathbf{X}_1 \sim \mathrm{p}_k(\mathbf{x}_1)$$

- Mixture formulation :

$$\begin{aligned} \mathbf{Z}_1 &\sim \mathrm{Mult}_K(1, \pi_1, \ldots, \pi_K) \\ \mathbf{X}_{1|Z_{1k}=1} &\sim \mathrm{p}_k(\mathbf{x}_1) \end{aligned}$$

- Joint and marginal distribution (mixture) :

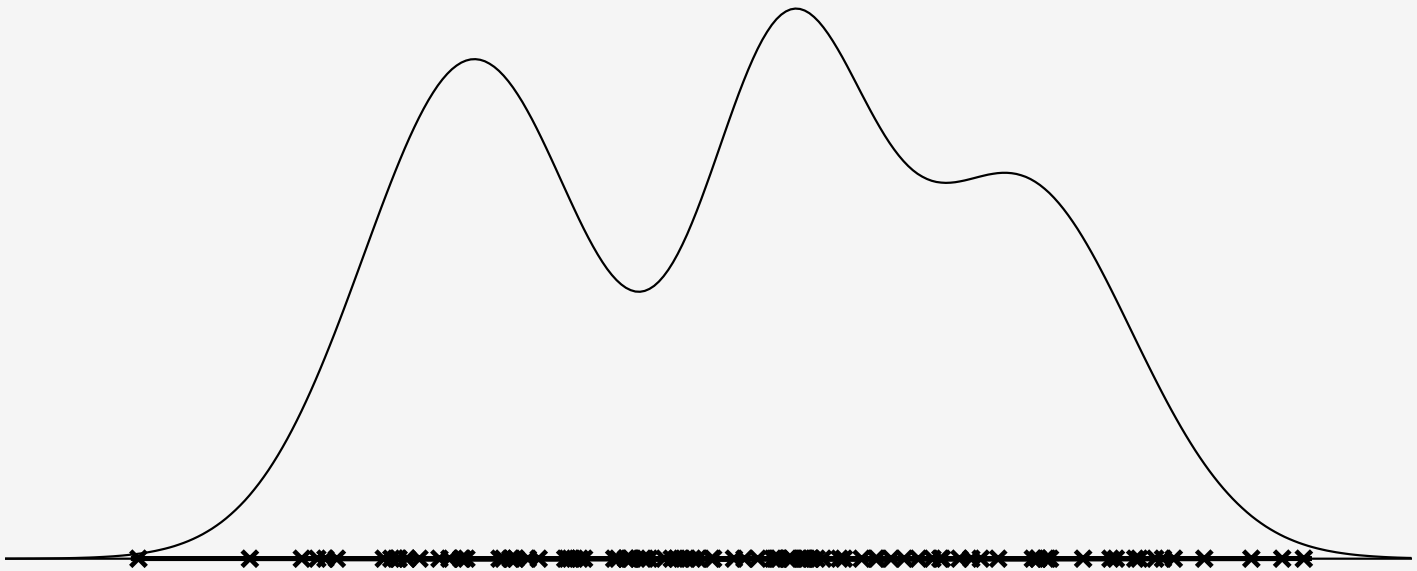
$$\begin{aligned} (\mathbf{X}_1, \mathbf{Z}_1) &\sim \prod_{k=1}^{K} [\pi_k \mathrm{p}_k(\mathbf{x}_1)]^{z_{1k}} \\ \mathbf{X}_1 &\sim \mathrm{p}(\mathbf{x}_1) = \sum_{k=1}^{K} \pi_k \mathrm{p}_k(\mathbf{x}_1) \end{aligned}$$

Mixture models (2/2)

$$p(\mathbf{x}_1) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_1).$$

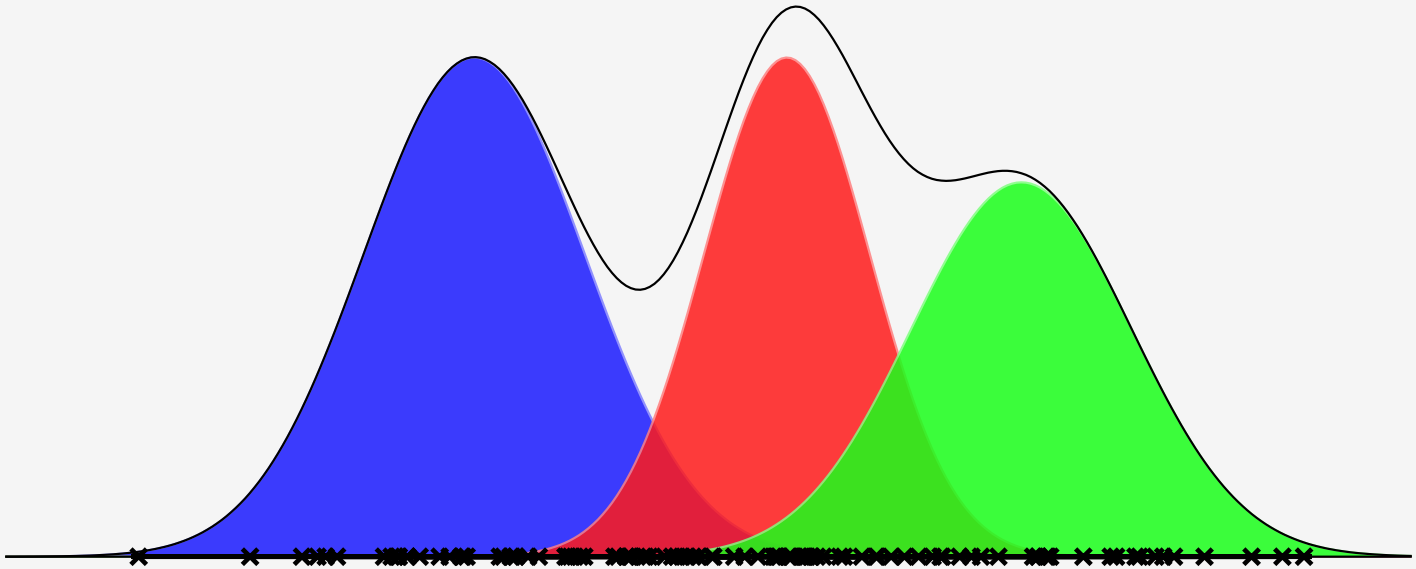
Mixture models (2/2)

$$p(\mathbf{x}_1) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_1).$$



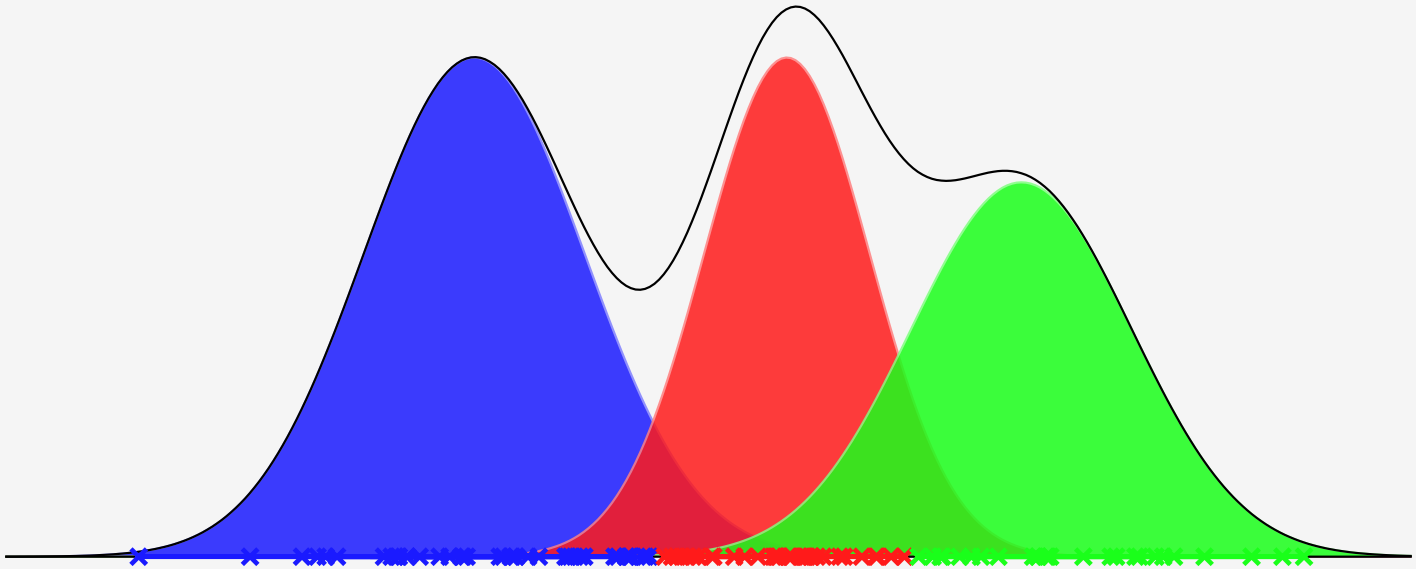
Mixture models (2/2)

$$p(\mathbf{x}_1) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_1)$$



Mixture models (2/2)

$$p(\mathbf{x}_1) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_1)$$



Prostate cancer data (1/5)

- Units : patients with prostate cancer classified in grade 3 or 4 of cancer based on advanced study.
- Variables : $d = 12$ preliminary measures on each patient :
  - 8 continuous variables : age, weight, systolic blood pressure, diastolic blood pressure, serum hemoglobin, size of primary tumor, index of tumor stage and histologic grade, serum prostatic acid phosphatase
  - 4 categorical variables : performance rating, cardiovascular disease history, electrocardiogram code, bone metastases
  - Missing data
- Goal : predict the grade of the cancer based on preliminary analysis.

Covariate	*Abbreviation*	*Number of Levels (if categorical)*
Age	Age	
Weight	Wt	
Performance rating	PF	4
Cardiovascular disease history	HX	2
Systolic Blood pressure	SBP	
Diastolic blood pressure	DBP	
Electrocardiogram code	EKG	7
Serum haemoglobin	HG	
Size of primary tumour	SZ	
Index of tumour stage and histolic grade	SG	
Serum prostatic acid phosphatase	AP	
Bone metastases	BM	2