Apprentissage de classifieurs naïfs de Bayes à partir de données soumises à un bruit de classification conditionnel à chaque classe

François Denis, Christophe Magnan, Liva Ralaivola

Laboratoire d'Informatique Fondamentale de Marseille, UMR CNRS 6166

CAp 2006, ICML 2006

Ce travail est partiellement financé
par l'ACI masses de données GENOTO3D

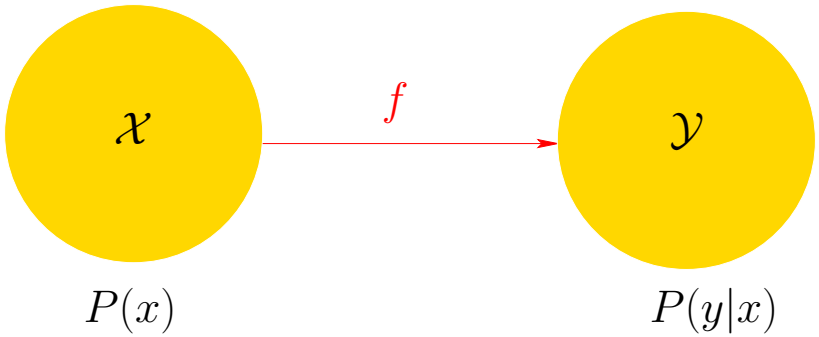
Cadre de l'apprentissage statistique

$\mathcal{X}$: espace discret de représentation
$\mathcal{Y}$: ensemble de classes ($\mathcal{Y} = \{0, 1\}$)

Soit $S = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. selon $P(x, y) = P(x) \cdot P(y|x)$

Cadre de l'apprentissage statistique

$\mathcal{X}$: espace discret de représentation
$\mathcal{Y}$: ensemble de classes ($\mathcal{Y} = \{0, 1\}$)

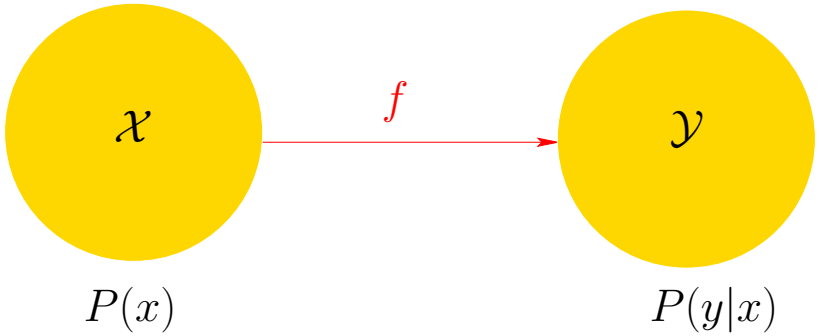


Soit $S = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. selon $P(x, y) = P(x) \cdot P(y|x)$

Objectif: calculer à partir de S un classifieur $f : \mathcal{X} \to \mathcal{Y}$ qui minimise $R(f) = P(f(x) \neq y)$.

Cadre de l'apprentissage statistique

$\mathcal{X}$: espace discret de représentation
$\mathcal{Y}$: ensemble de classes ($\mathcal{Y} = \{0, 1\}$)



Soit $S = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. selon $P(x, y) = P(x) \cdot P(y|x)$

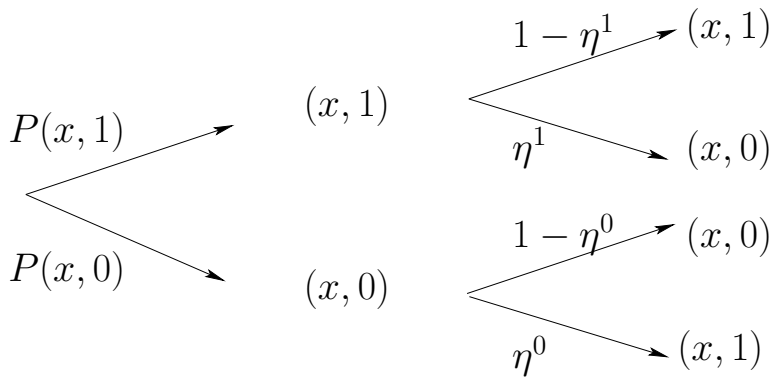
Classifieur optimal: classifieur de Bayes f_{Bayes}, qui
$\forall x \in \mathcal{X}, f_{Bayes}(x) = argmax_y P(y|x)$

Apprentissage à partir de données bruitées conditionnellement à chaque classe

Bruit CCCN

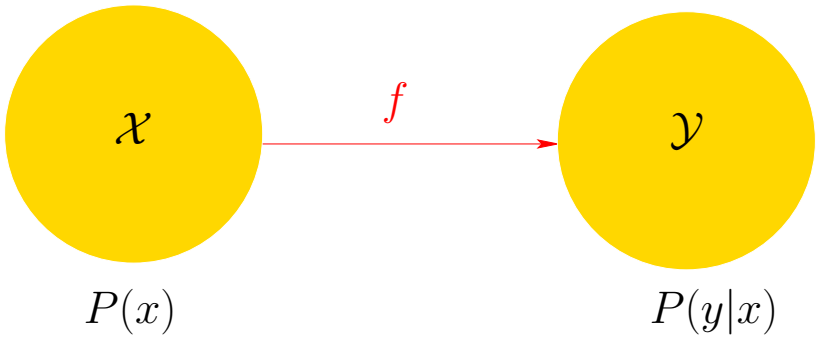
Bruit de classification conditionnel à chaque classe (CCCN)

Soit $S^{\overrightarrow{\eta}} = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. selon une distribution $P^{\overrightarrow{\eta}}(x, y)$ avec $\overrightarrow{\eta} = [\eta^0\ \eta^1]$ et $\eta^0, \eta^1 \in [0, 1]$



$$\begin{cases} P^{\overrightarrow{\eta}}(x,1) = (1-\eta^1)\cdot P(x,1) + \eta^0 \cdot P(x,0) \\ P^{\overrightarrow{\eta}}(x,0) = \eta^1 \cdot P(x,1) + (1-\eta^0)\cdot P(x,0) \end{cases}$$

Bruit de classification conditionnel à chaque classe (CCCN)



Objectif: calculer à partir de $S^{\overrightarrow{\eta}}$ un classifieur $f : \mathcal{X} \to \mathcal{Y}$ qui minimise $R(f) = P(f(x) \neq y)$ pour la distribution originale P.

Remarque

$$\begin{cases} P^{\overrightarrow{\eta}}(x,1) = (1-\eta^1)\cdot P(x,1) + \eta^0 \cdot P(x,0) \\ P^{\overrightarrow{\eta}}(x,0) = \eta^1 \cdot P(x,1) + (1-\eta^0)\cdot P(x,0) \end{cases}$$

Soient:

- $P'(x,y) = P(x,1-y)$
- $\eta'^0 = 1-\eta^1$
- $\eta'^1 = 1-\eta^0$

alors $P^{\overrightarrow{\eta}} = P'^{\overrightarrow{\eta}'}$ et les classifieurs de Bayes associés sont complémentaires.

Pour lever l'ambiguité, considérons $\eta^0 + \eta^1 < 1$

Apprenabilité avec CCCN

Peut-on apprendre à partir de données soumises à du bruit CCCN?

- Cas triviaux
- Cas général: problème mal posé
- Hypothèses sur les distributions P

Cas triviaux

P et $P^{\overrightarrow{\eta}}$ définissent le même classifieur de Bayes

$$P(1|x) > P(0|x) \Leftrightarrow P^{\overrightarrow{\eta}}(1|x) > P^{\overrightarrow{\eta}}(0|x)$$

Cas triviaux

Par définition de $P^{\overrightarrow{\eta}}$:

$$P^{\overrightarrow{\eta}}(1|x) \geq P^{\overrightarrow{\eta}}(0|x)$$
$$\Leftrightarrow$$
$$(1 - 2\eta^1) \cdot P(1|x) \geq (1 - 2\eta^0) \cdot P(0|x)$$

Cas triviaux

Par définition de $P^{\overrightarrow{\eta}}$:

$$P^{\overrightarrow{\eta}}(1|x) \geq P^{\overrightarrow{\eta}}(0|x)$$
$$\Leftrightarrow$$
$$(1 - 2\eta^1) \cdot P(1|x) \geq (1 - 2\eta^0) \cdot P(0|x)$$

- si $\eta^0 = \eta^1$ et $< \frac{1}{2}$, alors c'est le cas de P et $P^{\overrightarrow{\eta}}$

Cas triviaux

Par définition de $P^{\overrightarrow{\eta}}$:

$$P^{\overrightarrow{\eta}}(1|x) \geq P^{\overrightarrow{\eta}}(0|x)$$
$$\Leftrightarrow$$
$$(1 - 2\eta^1) \cdot P(1|x) \geq (1 - 2\eta^0) \cdot P(0|x)$$

- si $\eta^0 = \eta^1$ et $< \frac{1}{2}$, alors c'est le cas de P et $P^{\overrightarrow{\eta}}$
- si $\forall x \in \mathcal{X}, P(1|x) = 0$ ou $P(0|x) = 0$ et $\eta^0, \eta^1 < \frac{1}{2}$

Cas général: un problème mal posé

Exemple

Soit $\mathcal{X} = \{a\}$ et soient:

- P_1 telle que $P_1(0|a) = \frac{1}{3}$, $(f_{Bayes}(a) = 1)$
- $\overrightarrow{\eta}_1 = (0,0) \Rightarrow P_1^{\overrightarrow{\eta}}(0|a) = \frac{1}{3}$

- P_2 telle que $P_2(0|a) = \frac{2}{3}$, $(f_{Bayes}(a) = 0)$
- $\overrightarrow{\eta}_2 = (\frac{1}{2},0) \Rightarrow P_2^{\overrightarrow{\eta}}(0|a) = \frac{1}{3}$

- $P_1^{\overrightarrow{\eta}_1} = P_2^{\overrightarrow{\eta}_2}$
- les classifieurs de Bayes associés à P_1 et P_2 sont complémentaires

Restrictions à des ensembles de distributions

Définition

$\mathcal{P}$: ensemble de distributions sur $\mathcal{X} \times \mathcal{Y}$.

$\mathcal{P}$ est *identifiable* sur $\mathcal{X} \times \mathcal{Y}$ avec bruit CCCN ssi $\forall P \in \mathcal{P}, \forall \eta^0, \eta^1 / \eta^0 + \eta^1 < 1$, $P^{\overrightarrow{\eta}}$ détermine P.

- $P_1^{\overrightarrow{\eta}_1} = P_2^{\overrightarrow{\eta}_2} \Rightarrow P_1 = P_2$ et $\overrightarrow{\eta}_1 = \overrightarrow{\eta}_2$.

Distribution $P^{\overrightarrow{\eta}}$

$$\begin{cases} P_{+}^{\overrightarrow{\eta}}(x) = \alpha P_{+}(x) + (1-\alpha)P_{-}(x) \\ P_{-}^{\overrightarrow{\eta}}(x) = \beta P_{+}(x) + (1-\beta)P_{-}(x) \end{cases}$$

$$P_{+}(x) = P(x|y=1),\ P_{-}(x) = P(x|y=0)$$

Distribution $P^{\overrightarrow{\eta}}$

$$\begin{cases} P_+^{\overrightarrow{\eta}}(x) = \alpha P_+(x) + (1-\alpha)P_-(x) \\ P_-^{\overrightarrow{\eta}}(x) = \beta P_+(x) + (1-\beta)P_-(x) \end{cases}$$

$$\alpha = \frac{p\cdot(1-\eta^1)}{p\cdot(1-\eta^1)+(1-p)\eta^0},\ \beta = \frac{p\cdot\eta^1}{p\cdot\eta^1+(1-p)\cdot(1-\eta^0)}$$

$$(p = P(y=1))$$

Distribution $P^{\overrightarrow{\eta}}$

$$\begin{cases} P_+^{\overrightarrow{\eta}}(x) = \alpha P_+(x) + (1-\alpha)P_-(x) \\ P_-^{\overrightarrow{\eta}}(x) = \beta P_+(x) + (1-\beta)P_-(x) \end{cases}$$

$$\alpha = \frac{p\cdot(1-\eta^1)}{p\cdot(1-\eta^1)+(1-p)\eta^0} \text{ et } \beta = \frac{p\cdot\eta^1}{p\cdot\eta^1+(1-p)\cdot(1-\eta^0)}$$

- $\alpha, \beta, P_+^{\overrightarrow{\eta}}, P_-^{\overrightarrow{\eta}} \not\Rightarrow p, \eta^0, \eta^1$

Distribution $P^{\overrightarrow{\eta}}$

$$\begin{cases} P_+^{\overrightarrow{\eta}}(x) = \alpha P_+(x) + (1-\alpha)P_-(x) \\ P_-^{\overrightarrow{\eta}}(x) = \beta P_+(x) + (1-\beta)P_-(x) \end{cases}$$

$$\alpha = \frac{p\cdot(1-\eta^1)}{p\cdot(1-\eta^1)+(1-p)\eta^0} \text{ et } \beta = \frac{p\cdot\eta^1}{p\cdot\eta^1+(1-p)\cdot(1-\eta^0)}$$

- $\alpha, \beta, P_+^{\overrightarrow{\eta}}, P_-^{\overrightarrow{\eta}} \not\Rightarrow p, \eta^0, \eta^1$
- Proposition: $\alpha, \beta, P^{\overrightarrow{\eta}} \Rightarrow p, \eta^0, \eta^1$

Distribution $P^{\overrightarrow{\eta}}$

$$\begin{cases} P_+^{\overrightarrow{\eta}}(x) = \alpha P_+(x) + (1-\alpha)P_-(x) \\ P_-^{\overrightarrow{\eta}}(x) = \beta P_+(x) + (1-\beta)P_-(x) \end{cases}$$

$$\alpha = \frac{p\cdot(1-\eta^1)}{p\cdot(1-\eta^1)+(1-p)\eta^0} \text{ et } \beta = \frac{p\cdot\eta^1}{p\cdot\eta^1+(1-p)\cdot(1-\eta^0)}$$

- $\alpha, \beta, P_+^{\overrightarrow{\eta}}, P_-^{\overrightarrow{\eta}} \not\Rightarrow p, \eta^0, \eta^1$
- Proposition: $\alpha, \beta, P^{\overrightarrow{\eta}} \Rightarrow p, \eta^0, \eta^1$
- Corollaire: Soit P tel que les 2-mélanges de P_+ et P_- sont identifiables, alors $P^{\overrightarrow{\eta}} \Rightarrow \alpha, \beta \Rightarrow p, \eta^0, \eta^1$

Distributions produits

Définition - distributions produits

Soit $\mathcal{X} = \prod_{i=1}^{m} \mathcal{X}^i$ un espace défini par m attributs symboliques et soit $\mathcal{Y}$ un ensemble de classes. Si les attributs sont indépendants conditionnellement à chaque classe, alors

$\forall y \in \mathcal{Y}, P(x|y) = \prod_{i=1}^{m} P(x^i|y)$ est une distribution produit.

[Geiger et al, 2001, Whiley & Titterington, 2002, Freund & Mansour, 1999, Feldman et al, 2005]

Théorème: les mélanges de distributions produits sont identifiables

Application du corollaire

L'ensemble des distributions P telles que P_+ et P_- soient des distributions produits est identifiable avec bruit de classification conditionnel à chaque classe.

Classifieur naïf de Bayes

Si P est une telle distribution, alors le classifieur de Bayes f_{Bayes} devient le classifieur naïf de Bayes f_{NB} définit par:

$$f_{NB}(x) = argmax_y P(y) \prod_{i=1}^{m} P(x^i|y)$$

Classifieur naïf de Bayes

Si P est une telle distribution, alors le classifieur de Bayes f_{Bayes} devient le classifieur naïf de Bayes f_{NB} définit par:

$$f_{NB}(x) = argmax_y P(y) \prod_{i=1}^{m} P(x^i|y)$$

Les classifieurs naïfs de Bayes sont identifiables avec bruit de classification conditionnel à chaque classe ($P^{\overrightarrow{\eta}} \rightarrow \alpha, \beta \rightarrow P$)

Apprentissage à partir de données positives et non étiquetées

[Magnan, CAp 2005] Les classifieurs naïfs de Bayes sont identifiables à partir de données positives et non étiquetées.

Ce résultat est un corollaire de la proposition précédente.

Apprendre efficacement les classifieurs naïfs de Bayes

Il doit être possible d'identifier efficacement les classifieurs de Bayes à partir de bruit CCCN

Notre apport

- formules analytiques pour calculer les classifieurs naïfs de Bayes avec CCCN
- convergence rapide (résultats expérimentaux)

Etude expérimentale

Algorithmes

- `NB` l'algorithme naïf de Bayes classique

Algorithmes

- `NB` l'algorithme naïf de Bayes classique
- `NB-CCCN` induit par les formules analytiques

Algorithmes

- `NB` l'algorithme naïf de Bayes classique
- `NB-CCCN` induit par les formules analytiques
- `NB-UNL` induit par les formules fournies dans [Geiger et al, 2001]. Apprentissage à partir de données non-étiquetées

Algorithmes

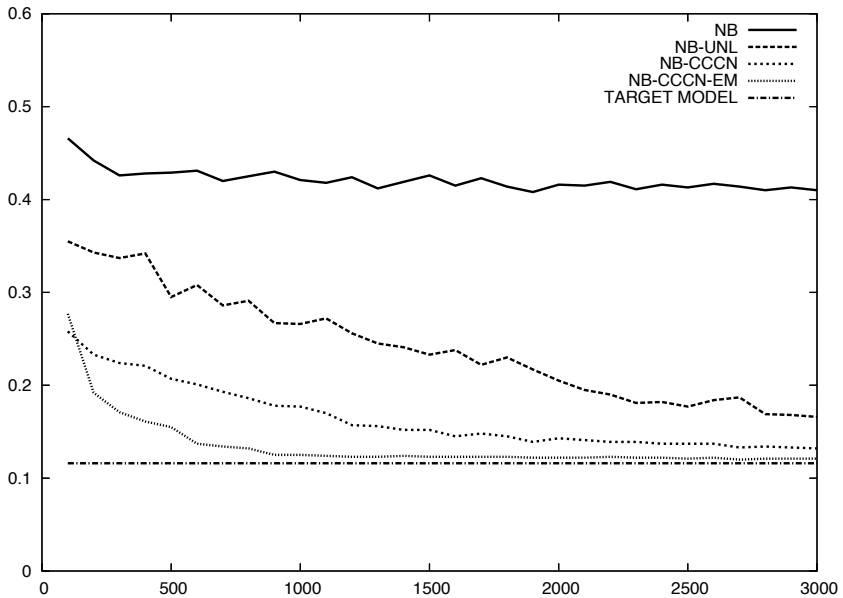
- `NB` l'algorithme naïf de Bayes classique
- `NB-CCCN` induit par les formules analytiques
- `NB-UNL` induit par les formules fournies dans [Geiger et al, 2001]. Apprentissage à partir de données non-étiquetées
- `NB-CCCN-EM`: à partir du classifieur apprit par `NB-CCCN`, appliquer la méthode E.M.. Itération des phases:
  - Estimation: re-étiquetage des données par le classifieur courant
  - Maximisation: calcul d'un nouveau classifieur sur les nouvelles données

Expériences sur données artificielles

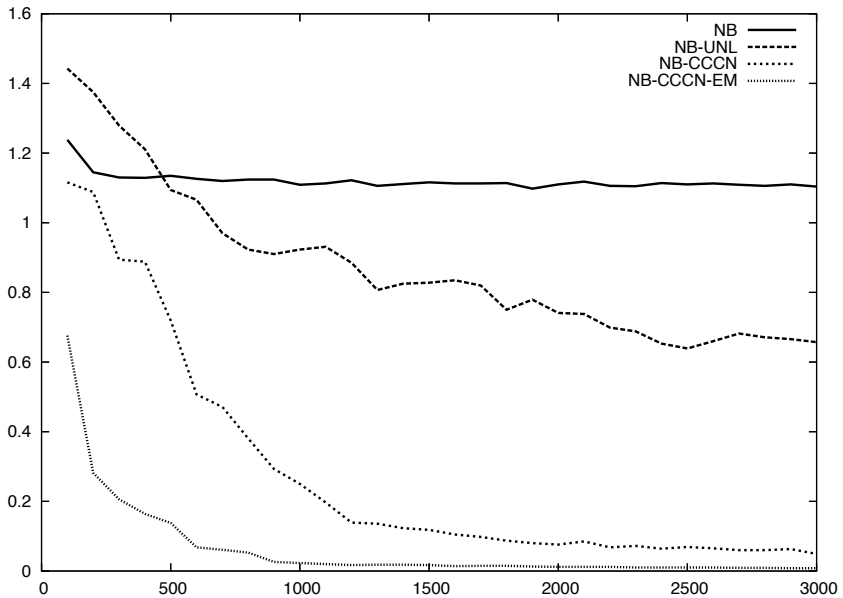
Génération des distributions cibles

Distributions cibles P générées aléatoirement (loi uniforme).
P_+ et P_- sont des distributions produit sur $\{0,1\}^{10}$.
Les ensembles d'apprentissage sont tirés à partir de cette distribution. Les classes des données d'apprentissage sont bruitées avec des probabilités $\eta^0 = 0.2$ et $\eta^1 = 0.5$.
Les ensembles test contiennent 10000 exemples générés à partir des distributions cibles. Leur classe n'est pas bruitée.

Résultats



$\hat{R}(f)$



Vraisemblance

Expériences sur jeux de données issus de l'UCI

Les jeux de données utilisés

Nom	$\|S\|$	$NbAtt$	$\|\mathcal{X}^i\|$
House Votes	433	16	2
Tic Tac Toe	958	9	3
Hepatitis	155	19	2-10
Breast Cancer	286	9	2-11
B. C. Wisc.	699	9	10
Bal. Scale	576	4	5

Résultats

Dataset		MC	NB	NB-CCCN	NB-CCCN-EM
H.Votes	ac	0.62	0.904	**0.916**	0.882
no noise	lk	-	-3134	-3035	**-2915**
	$\overrightarrow{\hat{\eta}}$	-	-	(.02,.08)	(.04,.20)
H.Votes	ac	0.38	0.866	**0.900**	0.873
$\overrightarrow{\eta}$ noise	lk	-	-4130	**-3037**	-3041
	$\overrightarrow{\hat{\eta}}$	-	-	(.33,.58)	(.20,.56)
T.T.T.	ac	0.65	**0.697**	0.682	**0.697**
no noise	lk	-	**-8726**	-8854	**-8726**
	$\overrightarrow{\hat{\eta}}$	-	-	(.09,.19)	(.00,.00)
T.T.T.	ac	0.35	0.562	**0.664**	0.587
$\overrightarrow{\eta}$ noise	lk	-	-8828	-8818	**-8815**
	$\overrightarrow{\hat{\eta}}$	-	-	(.24,.62)	(.21,.56)

Résultats

Dataset		MC	NB	NB-CCCN	NB-CCCN-EM
Hepat.	ac	0.79	0.827	**0.850**	0.770
no noise	lk	-	-1982	-2416	**-1902**
	$\overrightarrow{\hat{\eta}}$	-	-	(.31,.03)	(.50,.03)
Hepat.	ac	0.21	0.590	**0.811**	0.758
$\overrightarrow{\eta}$ noise	lk	-	-2095	-2273	**-1946**
	$\overrightarrow{\hat{\eta}}$	-	-	(.25,.55)	(.29,.45)
Br.Can.	ac	0.70	0.730	**0.760**	0.718
no noise	lk	-	-2520	-2682	**-2448**
	$\overrightarrow{\hat{\eta}}$	-	-	(.06,.20)	(.13,.27)
Br.Can.	ac	0.30	0.581	**0.732**	0.722
$\overrightarrow{\eta}$ noise	lk	-	-2573	-2623	**-2479**
	$\overrightarrow{\hat{\eta}}$	-	-	(.19,.59)	(.33,.56)

Conclusions

- Poser un cadre formel de l'apprentissage avec CCCN
- En général, c'est un problème mal posé
- Mais il y a des cas intéressants
- En particulier, les classifieurs naïfs de Bayes sont identifiables avec bruit CCCN
- Une méthode analytique efficace pour l'identification de P à partir de $P^{\overrightarrow{\eta}}$

Perspectives

- Comment valider les résultats si les données test sont également bruitées

$$R(f) = \frac{R^{\overrightarrow{\eta}}(f) - \eta^1 \cdot p_f - \eta^0 \cdot (1 - p_f)}{1 - \eta^0 - \eta^1}$$

- D'autres classes de distributions identifiables à partir de données bruitées?