Likelihood in the symbolic context. Examples

Richard Emilion

[1] Denis Poisson Institute
University of Orléans

Data Science, January 11^{th}, 2019

Outline

Warmest thanks To Edwin Diday and to the organizers for this invitation.

- **I - Probabilistic setting for the symbolic paradigm**
- **II - Likelihood, examples**

I - A Probabilistic setting for the symbolic paradigm

. E. Diday (1990).
Symbolic data expresses the variability of the data within a class of data.
SDA : Considering classes of data as new statistical units.

I - A Probabilistic setting for the symbolic paradigm

. E. Diday (1990).
Symbolic data expresses the variability of the data within a class of data.
SDA : Considering classes of data as new statistical units.

. R.E. (ISI Conference, Rio 2015) Introducing a formalism: X (data r.v.), C (class variable), S symbolic variable

I - A Probabilistic setting for the symbolic paradigm

. E. Diday (1990).
Symbolic data expresses the variability of the data within a class of data.
SDA : Considering classes of data as new statistical units.

. R.E. (ISI Conference, Rio 2015) Introducing a formalism: X (data r.v.), C (class variable), S symbolic variable

.Some modifications of Rio talk: Emilion-Diday 2018, Book chapter, Eds. G. Saporta

I - A Probabilistic setting for the symbolic paradigm

. E. Diday (1990).
Symbolic data expresses the variability of the data within a class of data.
SDA : Considering classes of data as new statistical units.

. R.E. (ISI Conference, Rio 2015) Introducing a formalism: X (data r.v.), C (class variable), S symbolic variable

.Some modifications of Rio talk: Emilion-Diday 2018, Book chapter, Eds. G. Saporta

. Present talk : Examples of symbolic likelihood. Applications.

I.1 Description Variable

- Population of individuals: $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space
- $(\mathbb{V}, \mathcal{V})$ a measurable space of descriptions.
  $X : \Omega \longrightarrow \mathbb{V}, \;$ measurable w.r.t. $\mathcal{F}$ and $\mathcal{V}$, a random variable (r.v) which describes the individuals.
- Generally, $\mathbb{V}$ is a measurable subset of $\mathbb{R}^p$ and
  $X = (X_1, \ldots, X_p), \; p = 1, 2, \ldots.$
- Standard Data Analysis: $n \times p$ numerical table of a n-sample of $(X_1, \ldots, X_p)$

I.2 Class Variable

- Class variable

$$C : \Omega \longrightarrow \mathbb{C}, \quad \text{measurable w.r.t. } \mathcal{F} \text{ and } \mathcal{C}, \tag{1}$$

r.v. which assigns a class label to each individual.
$(\mathbb{C}, \mathcal{C})$ is a measurable space of class labels.

I.2 Class Variable

- Class variable

$$C : \Omega \longrightarrow \mathbb{C}, \quad \text{measurable w.r.t. } \mathcal{F} \text{ and } \mathcal{C}, \tag{1}$$

r.v. which assigns a class label to each individual.
$(\mathbb{C}, \mathcal{C})$ is a measurable space of class labels.

- X and C are correlated
- Class with label $c \in \mathbb{C}$, shortly Class c :

$$(C = c) = \{\omega \in \Omega : C(\omega) = c\} \tag{2}$$

- It is assumed that singletons $\{c\}$ belong to $\mathcal{C}$, $\forall c \in \mathbb{C}$, so that classes for $c \in range(C)$ form a measurable partition of Ω

I.5 Symbolic variable, Symbolic data

Definition : A symbolic variable S of the context (X, C) is defined as a mapping

$$S : \mathbb{C} \longrightarrow \mathbb{S}$$
$$S(c) \;=\; f(\mathbb{P}_{X|C=c}). \tag{3}$$

where

$$f : \mathcal{M}_1(\mathcal{V}) \longrightarrow \mathbb{S} \tag{4}$$

is a measurable function taking value in some measurable space of symbols $(\mathbb{S}, \mathcal{S})$.
$S(c), c \in \mathbb{C}$ is a 'symbolic data' representing the variability of the data $X(\omega)$ for $\omega \in (C = c)$.

I.6 Symbols in term of samples

- $\mathbb{P}_{X|C=c}$ is a probability distribution on $(\mathbb{V}, \mathcal{V})$, a complex object. It is generally estimated from an observed sample $(x^{(1)}, c^{(1)}), \ldots, (x^{(n)}, c^{(n)})$ of the pair (X, C) such that $c^{(j)} = c$.

I.6 Symbols in term of samples

- $\mathbb{P}_{X|C=c}$ is a probability distribution on $(\mathbb{V}, \mathcal{V})$, a complex object. It is generally estimated from an observed sample $(x^{(1)}, c^{(1)}), \ldots, (x^{(n)}, c^{(n)})$ of the pair (X, C) such that $c^{(j)} = c$.
- $S(c)$ can be estimated by an aggregating function of a sample $(x^{(1)}, c^{(1)}), \ldots, (x^{(n)}, c^{(n)})$ such that $c^{(j)} = c$.
- BLS (Beranger - Lin - Sisson) Definition

II - Likelihood in the symbolic context

II.1. Probability measures on $(\mathbb{C}, \mathcal{C})$, density

- $S : \mathbb{C} \longrightarrow \mathbb{S}$ a symbolic variable with $\mathbb{S} = \mathbb{N}^m$ or $\mathbb{S} = \mathbb{R}^m$
- Problem: density of S w.r.t. to the counting (resp. the Lebesgue) measure ? In the continuous case $\mathbb{C}$ should be uncountable and even nonatomic.

$$\begin{aligned} d_S : \mathbb{S} &\longrightarrow \quad \mathbb{R}_+ \\ \underline{s} &\longrightarrow \quad d_S(\underline{s}) \end{aligned} \tag{5}$$

Estimating d_S given a n-sample $\underline{s}_1, \ldots, \underline{s}_n$:

$$\underline{s}_i = (s_{i,1}, \ldots s_{i,m}) = S^{(i)}(c) \in \mathbb{R}^m, \quad S^{(i)} \overset{i.i.d.}{\sim} \mathbb{Q}_S, \; i = 1, \ldots, n$$

for some $c \in \mathbb{C}$: randomness of the sample of symbols.

II.2. LDA model (Blei-Ng-Jordan) Specif. a

Model in Text Mining context. Can be applied in other domains.
. *categorical* r.v. $X : \Omega \longrightarrow \mathbb{V} = \{1, \ldots, k\}$, a finite set of k *topics*,
. r.v. $N : \Omega \longrightarrow \mathbb{N} = \{0, 1, 2 \ldots, \}$,
random probability vector $\theta = (\theta_1, \ldots, \theta_k) : \Omega \longrightarrow \mathbb{T}_k$

$$\begin{cases} (N, \theta) \sim \mathit{Poisson}(\xi) \otimes \mathit{Dirichlet}(\underline{\alpha}) \\ \mathbb{P}(X = i|\theta) = \theta_i, \; i = 1, \ldots, k. \end{cases} \tag{6}$$

II.2. LDA model. Specifications b

. $\{1, \ldots, V\}$ a finite set of V *words*
$\beta = (\beta_{i,j}), i = 1, \ldots, k,\ j = 1, \ldots, V$ be a $k \times V$ Markov matrix, each of its k rows being a probability vector in dimension V. A *document*, is considered as an outcome c of our class random variable C defined as a sequence of random words:

$$\begin{cases} C(\omega) = (W^{(1)}(\omega), \ldots, W^{(N(\omega))}(\omega)),\ \omega \in \Omega \\ \text{where, given } N \text{ and } \theta \\ X^{(r)} \overset{i.i.d}{\sim} \mathbb{P}_{(X|\theta)}, \text{ for each } r = 1, \ldots N \\ W^{(r)} : \Omega \longrightarrow \{1, \ldots, V\},\ r = 1, \ldots N, \text{ are independent} \\ \mathbb{P}(W^{(r)} = v | X^{(r)}) = \beta_{(X^{(r)}, v)}, \text{ for each } r = 1, \ldots N,\ v = 1, \ldots V. \end{cases} \tag{7}$$

II.2. LDA model. Symbolic likelihood

Given a class label $c = (w_1, \ldots, w_N)$, class c is defined as

$$(C = c) = \{\omega \in \Omega : W^{(1)}(\omega) = w_1, \ldots, W^{(N)}(\omega) = w_N\}$$

Topic variability of a document c with N words $(w_1, \ldots, w_N)$ and unobserved topics $(x_1, \ldots, x_N)$ expressed by the *latent symbol*

$$s(c) = (\sum_{r=1}^{N} 1_{(x_r=1)}, \ldots, \sum_{r=1}^{N} 1_{(x_r=k)})$$

Random symbol $S = s \circ C = (\sum_{r=1}^{N} 1_{(X^{(r)}=1)}, \ldots, \sum_{r=1}^{N} 1_{(X^{(r)}=k)})$ distribution, given (N, θ), is multinomial

$$\mathbb{P}(S = (n_1, \ldots, n_k) | N = n, \theta) = \frac{n!}{n_1! \ldots n_k!} \theta_1^{n_1} \ldots \theta_k^{n_k}$$

if $n_1 + \ldots + n_k = n$

II.2. LDA. Document / Corpus Probability

We have $p(x_r|\theta) = \prod_{i=1}^{k} \theta_i^{1_{x_r=i}}$ and $p(w_r|x_r,\beta) = \prod_{j=1}^{V} \beta_{x_r,j}^{1_{w_r=j}}$

$$\begin{cases} p(x_r, w_r|\theta,\beta) = \prod_{i=1}^{k} \theta_i^{1_{x_r=i}} \prod_{j=1}^{V} \beta_{x_r,j}^{1_{w_r=j}} \\ \\ p(w_r|\theta,\beta) = \sum_{x_r} \prod_{i=1}^{k} \theta_i^{1_{x_r=i}} \prod_{j=1}^{V} \beta_{x_r,j}^{1_{w_r=j}} \end{cases} \tag{8}$$

The probability of a document is,

$$p(w_1,\ldots w_N|\theta,\beta,N) = \prod_{r=1}^{N} \sum_{x_r} \prod_{i=1}^{k} \theta_i^{1_{x_r=i}} \prod_{j=1}^{V} \beta_{x_r,j}^{1_{w_r=j}} \tag{9}$$

$$p(w_1,\ldots w_N|\underline{\alpha},\beta,N) = \int Dd(\theta|\underline{\alpha}) \prod_{r=1}^{N} \sum_{x_{d_r}} \prod_{i=1}^{k} \theta_i^{1_{x_r=i}} \prod_{j=1}^{V} \beta_{x_r,j}^{1_{w_r=j}} d\theta$$

II.3. BLS (Beranger-Lin-Sisson) method

$X : \Omega \longrightarrow \mathbb{R}^p$ r.v. with density $d_X(.|\theta)$.
$N \geq 2$ any very large integer. $(X^{(1)}, \ldots, X^{(N)}), X^{(r)} \overset{i.i.d}{\sim} \mathbb{P}_X$.
$(x^{(1)}, \ldots, x^{(N)})$ an observed large sample
$c = (B_k)_{k \in K}$ a finite partition covering the support of $\mathbb{P}_{(X^{(1)}, \ldots, X^{(N)})}$
Given partition c, the joint distribution of

$$(1_{(X^{(1)} \in B_1)}, \ldots, 1_{(X^{(N)} \in B_1)}, \ldots, 1_{(X^{(1)} \in B_K)}, \ldots, 1_{(X^{(N)} \in B_K)})$$

is captured by the symbolic variable S defined as

$$S(c) = (\sum_{r=1}^{N} 1_{(X^{(r)} \in B_1)}, \ldots, \sum_{r=1}^{N} 1_{(X^{(r)} \in B_k)}),$$

whose distribution is multinomial $(N, p_1, \ldots, p_K)$ with
$p_k = \int_{B_k} d_X(x|\theta) dx$

II.3. BLS (Beranger-Lin-Sisson) result

The main observation in BLS (Beranger B., Lin H., Scott A. S., New models for symbolic data analysis, ArXiv e-prints (2018)) is that the total probabilty formula

$$d_{S|\nu,\theta} = \int_{t\in(\mathbb{R}^p)^N} d_{S|(X^{(1)},\ldots,X^{(N)})=t,\nu} (d_X)^{\otimes N}(t|\theta) \tag{10}$$

yields an inference on parameter θ from an inference on the symbolic likelihood with parameter ν.
This considerably reduces inference complexity and seems to be a significant application of the symbolic approach

II.4 Dirichlet Process Mixture (DPM)

- h_i : symbol (e.g. histogram h_i = bins, frequencies) Bayesian parametric: $h_i|\theta \sim F(\theta)$, $\theta \sim D$: apriori D, shape

II.4 Dirichlet Process Mixture (DPM)

- h_i : symbol (e.g. histogram h_i = bins, frequencies) Bayesian parametric: $h_i|\theta \sim F(\theta)$, $\theta \sim D$: apriori D, shape
- DPM: Bayesian nonparametric, flexible, infinite mixture model

$$\begin{cases} h_i|\theta_i \overset{ind}{\sim} F(.,\theta_i), i = 1, \dots, n, \quad \theta_i \in \Theta \\ \theta_i|P = p \overset{i.i.d.}{\sim} p, i = 1, \dots, n \\ P \sim DP(c, P_0) \; a \; Dirichlet \; Process \; on \; \Theta \end{cases} \tag{11}$$

Draw p from $DP(c, P_0)$, θ_i from p and h_i from $F(.,\theta_i)$: the distribution on Θ is the mixture $\int_\Theta F(.,\theta)dp(\theta)$

II.5 Mixture of Dirichlet Processes (MDP)

- Antoniak (Ann. Stat. 1974): If the h_i's and P are as in (11) then the posterior

$$P|h_1,\ldots,h_n \sim \int DP(cP_0 + \sum_{i=1}^{n} \delta_{\theta_i}) d\mathbb{P}_{\theta_1,\ldots,\theta_n|h_1,\ldots,h_n} \qquad (12)$$

- In other words the posterior is a Mixture of Dirichlet Processes (MDP)
- The posterior MDP provides a classification of the histogram data without any apriori number of classes, mixture component = fuzzy class.
- A mixture of DD estimated from an histogram dataset converges to a MDP as the bin width goes to 0^+ (R.E. *Stat. Anal. & Data Mining* 2012)
- 'DPpackage' in R