Maximum likelihood with coarse data based on robust optimisation

Romain Guillaume, Inés Couso, Didier Dubois

IRIT, Université Toulouse, France, Universidad de Oviedo, Gijon (Spain)

July 10, 2017

Contents

Introduction

- Maximum likelihood is a standard approach to finding a probabilistic model based on data.
- It maximizes the "probability" of obtaining the observations (supposed to belong to a set of mutually exclusive outcomes)
- When observations are coarse and may overlap, it is not clear how to define the likelihood function: several options
- *Here we try to optimize the likelihood function that we should have observed, had observations been precise, using a robust optimization approach.*

The random phenomenon and its measurement process

- $X : \Omega \to \mathcal{X}$ represent the outcome of a certain random experiment. $\mathcal{X} = \{a_1, \ldots, a_m\}$.
- $Y : \Omega \to \wp(\mathcal{X})$ that models the reports of a measurement device, where $\wp(\mathcal{X})$ is the set of subsets of $\mathcal{X}$. $\mathcal{Y} = \{A_1, \ldots, A_r\}$.

The information about the joint distribution of the random vector (X, Y) modeling the random variable X $(P(X))$ and its measurement process $(P(Y|X))$ can be represented by a joint probability on $\mathcal{X} \times \mathcal{Y}$:

$$p_{ij} = P(Y = A_j | X = a_i) \cdot P(X = a_i) = \begin{cases} P(X = a_i) \text{ if } a_i \in A_j \\ 0 \text{ otherwise.} \end{cases}$$

We shall just ignore the measurement process.

Formalisation in the finite case

Let n_k be the number of appearances of a_k in the virtual sample $\mathbf{x}$, we have that any $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$ satisfies:

$$\begin{cases} \sum_{k=1,...,r} n_k = \sum_{j=1,...,r} q_j = N \\ n_k = \sum_{j=1}^{r} n_{kj}, \forall k = 1, ..., m \\ q_j = \sum_{k=1}^{m} n_{kj} \forall j = 1, ..., r. \\ n_{kj} = 0 \text{ if } a_k \notin A_j, \forall k, j. \end{cases} \tag{1}$$

For a complete sample $\mathbf{z}$ to be compatible with the observation $\mathbf{y}$, we have that any $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ satisfies:

$$\begin{cases} \sum_{k=1,...,r} \sum_{j=1,...,r} n_{kj} = N \\ q_j = \sum_{k=1}^{m} n_{kj}, \forall j = 1, ..., r. \\ n_{kj} = 0 \text{ if } a_k \notin A_j, \forall k, j. \end{cases} \tag{2}$$

Different likelihood functions

We may consider three different log-likelihood functions depending on whether we refer to

1. the observed sample in $\mathcal{Y}$: $L^{\mathbf{y}}(\theta) = \log \prod_{i=1}^{N} p(y_i; \theta)$.
2. the hidden sample in $\mathcal{X}$: $L^{\mathbf{x}}(\theta) = \log \prod_{i=1}^{N} p(x_i; \theta)$.
3. the complete sample in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$: $L^{\mathbf{z}}(\theta) = \log \prod_{i=1}^{N} p(z_i; \theta)$.

We focus on the hidden sample log-likelihood functions.

Imprecise likelihood evaluation

There are two strategies of likelihood maximization, based on a sequence of imprecise observations $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$:

1. The maximax strategy : $(\mathbf{x}^*, \theta^*) = \arg\max_{\mathbf{x} \in \mathcal{X}_\mathbf{y}, \theta \in \Theta} L^\mathbf{x}(\theta)$.
2. The maximin strategy : it aims at finding $\theta_* \in \Theta$ that maximizes $\underline{L}(\theta) = \min_{\mathbf{x} \in \mathcal{X}_\mathbf{y}} L^\mathbf{x}(\theta)$.

We study the maximin strategy in the scope of robust optimisation.

The robust approach to discrete probability estimation with coarse data

$$\max_{\mathbf{p}} \min_{\mathbf{n}} \sum_{k=1,\ldots,m} n_k \cdot \log p_k \tag{3}$$

$s.t.$

$$(a) \quad \sum_{k=1,\ldots,m} n_k = \sum_{j=1,\ldots,r} q_j = N$$

$$(b) \quad n_k = \sum_{j:(j,k)\in\mathbb{E}^{\mathbf{y}}} n_{j,k}, \qquad \forall k = 1, \ldots, m$$

$$(c) \quad q_j = \sum_{k:(j,k)\in\mathbb{E}^{\mathbf{y}}} n_{j,k}, \qquad \forall j = 1, \ldots, r$$

$$(d) \quad \sum_{k=1,\ldots,m} p_k = 1$$

$$(e) \quad n_k, n_{j,k} \in \mathbb{N}, p_k > 0, \qquad \forall k = 1, \ldots, m,$$

The maxmin strategy maximizes entropy

Proposition

Assuming $\boldsymbol{n}$ *is not restricted to being integer-valued, the equality* $\max_{\boldsymbol{p}} \min_{\boldsymbol{n}} \sum_{k=1,\ldots,m} n_k \cdot \log p_k = \min_{\boldsymbol{n}} \max_{\boldsymbol{p}} \sum_{k=1,\ldots,m} n_k \cdot \log p_k$ *holds.*

(from a standard result in game theory)

Corollary

The optimal solution to the maxmin likelihood problem (3) is the solution with maximal entropy, namely the solution to: $\max_{\boldsymbol{n}} -\sum_{k=1,\ldots,m} \frac{n_k}{N} \cdot \log \frac{n_k}{N}$ *under conditions (3(a, b, c)), and* $n_k \in \mathbb{R}^+$, *i.e.* $\boldsymbol{n}$ *in the convex hull of* $\mathcal{N}^{\mathbf{y}}$.

This is a min cost max-flow problem

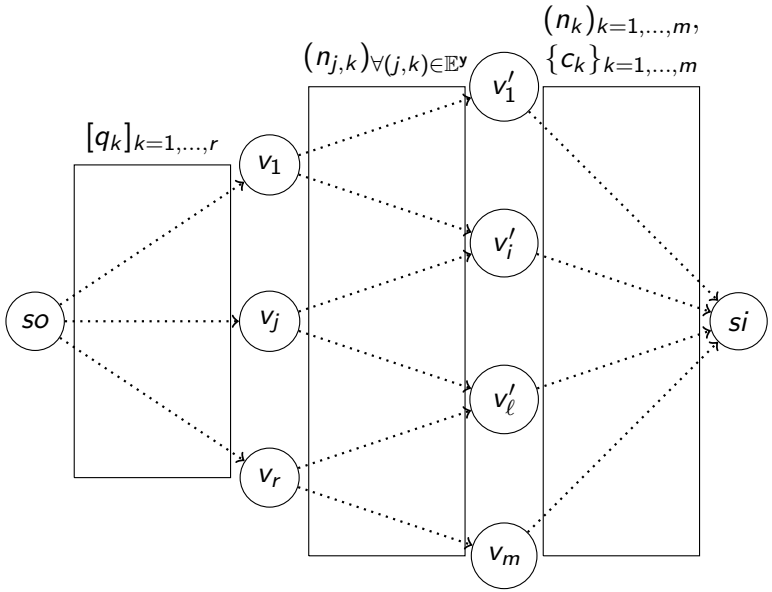

Figure: Graph representation of the problem

Hence the maximal entropy solution is indeed of the form $(n_1/N, \dots, n_m/N)$ for integer values of n_k.

What the two strategies mean

The above results shed light on the significance of the maximin and the maximax strategies and are useful to understand when to apply one or the other.

- The maximin strategy:
  - the process generating the variable X is genuinely non-deterministic
  - the imprecision of the observation may hide some variability
- The maximax strategy:
  - the underlying phenomenon is deterministic but the observations are noisy and coarse.
  - If we try to learn a best model taken from a class of models and we have some good reason to think that the phenomenon under study can be represented by one of these models

The dual problem

For fixed probability vectors $\mathbf{p}$ and using duality theorem, the minimization parts of problem (3) is equivalent to:

$$\max_{\alpha,\beta,\gamma} -(\alpha \sum_{k=1}^{r} q_k + \sum_{k=1}^{r} \gamma_k q_k) \tag{4}$$

$$\begin{aligned} s.t. \quad & \alpha + \beta_k \geq -\log(p_k), && \forall k = 1, ..., m \\ & -\beta_j + \gamma_k \geq 0, && \forall (k,j) \in \mathbb{E}^{\mathbf{y}} \\ & \alpha, \beta_j, \gamma_k \in \mathbb{R}, && \forall j = 1, ..., r, k = 1, ..., m \end{aligned}$$

New formulation of robust model

The problem (3) can be now written as follows :

$$\min_{\mathbf{p},\alpha,\beta,\gamma} \alpha \sum_{k=1}^{r} q_k + \sum_{k=1}^{r} \gamma_k q_k$$

$s.t.$

$$(a) \quad \alpha + \beta_k \geq -\log(p_k), \qquad \forall k = 1, ..., m$$

$$(b) \quad -\beta_j + \gamma_k \geq 0, \qquad \forall (j,k) \in \mathbb{E}^{\mathbf{y}}$$

$$(c) \quad \sum_{k=1,...,m} p_k = 1$$

$$(d) \quad p_k + \epsilon \geq 0, \qquad \forall k = 1, ..., m$$

$$(e) \quad p_k, \alpha, \beta_j, \gamma_k \in \mathbb{R}, \qquad \forall j = 1, ..., r, k = 1, ..., m$$

It can be solved using standard optimisation software.

Example: cars in a parking lot

- Three colors: red (r), blue(b),grey (g)
- Two situations for doors: 3 doors (3) and 5 doors (5)

The information reported by the custodian can be, for each car:

- both the color and the number of doors,
- or only the color,
- or only the number of doors.

$\mathcal{Y}$	$\{r3\}$	$\{r5\}$	$\{b3\}$	$\{b5\}$	$\{g3\}$	$\{g5\}$
q	9	167	120	199	164	188
$\mathcal{Y}$	$\{r3, b3, g3\}$	$\{r5, b5, g5\}$	$\{r3, r5\}$	$\{b3, b5\}$	$\{g3, g5\}$	
q	80	80	18	107	100	

Table: Distribution of Coarse Observations

Example: cars in a parking lot

We compare maximin probability distribution (noted p^{Mm}) with the probability distribution obtained using a maximax approach (noted p^{MM})

$\mathcal{X}$	$\{r3\}$	$\{r5\}$	$\{b3\}$	$\{b5\}$	$\{g3\}$	$\{g5\}$
$p^{Mm}(X = a_i) \approx$	0.141	0.171	**0.171**	**0.171**	0.173	0.173
$p^{MM}(X = a_i) \approx$	0.007	0.150	**0.098**	**0.313**	0.279	0.153

Table: Estimations of probability distributions on the latent variable

- Around half of observations concerning $\{b3\}$ or $\{b5\}$ are imprecise.
- The maximax solution makes precise predictions, while the maximin approach is cautious and uncertain.

Example: cars in a parking lot

Let us swap observations $\{b5\}$ and $\{b3\}$:

$\mathcal{X}$	$\{r3\}$	$\{r5\}$	$\{b3\}$	$\{b5\}$	$\{g3\}$	$\{g5\}$
$p^{Mm}(X = a_i) = \text{New}p^{Mm}(X = a_i)$	0.141	0.171	0.171	0.171	0.173	0.173
$p^{MM}(X = a_i)$	0.007	0.150	0.098	0.313	0.279	0.153
$\text{New}p^{MM}(X = a_i)$	0.008	0.150	*0.313*	*0.097*	**0.133**	**0.299**

Table: Estimations of probability distributions on the latent variable

- The result of the maximin strategy remains the same despite the swapping as the data about $\{b5\}$ and $\{b3\}$ is quite imprecise.
- The maximax strategy seems to be sensitive to change in data.

Conclusions and perspectives

- The close connections between maximax and maximin strategies with entropy optimization shed light on the significance of each approach
- We have proposed an efficient solving technique that can use existing non-linear optimization software.
- Further work is needed to test the approach on real data,
- Compare obtained results with other approaches
  - that use belief functions,
  - and also recent possibilistic maximum likelihood methods.