Evidential likelihood flatness as a way to measure data quality: the multinomial case

Liyao Ma[1] **Sébastien Destercke**[2] **Yong Wang**[1]

[1]Department of Automation, University of Science and Technology of China, Hefei, China
[2]UMR7253 Heudiasyc, Centre de Recherches de Royallieu, Compiègne, France

Abstract

Likelihood functions, as well as the more recent concept of evidential likelihood, are essential statistical tools to perform estimation. Beyond the maximal likelihood value, the shape of the likelihood can also give interesting information about the data used to get the estimation. Indeed, it is generally acknowledged that more uncertain and scarce data will lead to flatter likelihoods. However, different aspects can affect this shape, and it may be worthwhile to separate various influences. In this paper, we discuss these influences and propose some practical ways to separate them into different measures. We demonstrate our approach on the particular case of the multinomial likelihood, which plays an important role in applications such as classification.

Keywords: data quality, evidential likelihood, consonant mass function, non-specificity

1. Introduction

The effect of data quality in statistical estimations and other tasks such as classification [1] has been acknowledged for a long time, and is typically reflected through the width of confidence intervals. A possible way to calculate such intervals is using the relative likelihood [2, 3]. The use of this likelihood can be justified in different ways, and we refer to Denoeux [4] for a recent discussion.

Clearly, the wider the intervals are, the less quality there is in the data. This means that the relative likelihood resulting from the observations is less peaked, or flatter. Hence, a measure of its flatness is a direct measurement of the data quality. In this paper, we propose a generic measure of flatness of the evidential likelihood using non-specificity, and focus on the particular case of multinomial distributions to propose practical ways to compute it. Such a particular case will be interesting to discuss the various aspects that can affect the likelihood flatness. It is also used in many applications such as classification or semi-supervised learning [5].

Indeed, it is not always clear what means data quality: Ballow and Pazer [6] consider four aspects of data quality: accuracy, timeliness, completeness and consistency, and Wang and Wang [7] divide data quality into completeness, unambiguous, meaningfulness and correctness. We will mainly focus on three aspects of data quality: uncertainty, variability and quantity, which play a major role in classification tasks. By uncertainty, we mean that the true fixed value of a given sample may be ill-known. Such uncertainty comes from a lack of knowledge and is intrinsically epistemic, hence we consider that it will be modelled under the belief function framework [8, 9] in general, or to be more specific, modelled by mass functions. By variability, we mean that the empirical distribution tends to show different values, i.e., that the entropy is likely to be high.

We first recall some basics about the relative likelihood and the theory of belief functions before proposing a general measure of flatness (Section 2). We then study in detail the multinomial evidential likelihood, showing through simple examples how different aspects of data quality can affect the likelihood (Section 3). Finally, we propose a simple, practical way to measure flatness in the case of multinomial distribution, as well as a means to separate different aspects into different measures (Section 4).

2. Data uncertainty, likelihood and non-specificity

We will use mass functions, issued from the theory of belief functions (also referred to as Dempster-Shafer theory or evidence theory), as a common tool to model data uncertainty and the relative likelihood. We will only introduce the tools we need, and the reader is referred to [8] for a full introduction. A mass function on a space Ω is simply a mapping $m : \wp(\Omega) \to [0,1]$ from the power set of Ω to the unit interval, such that $m(\emptyset) = 0$ and

$$\sum_{A \subseteq \Omega} m(A) = 1.$$

A subset A that receives a positive mass $m(A) > 0$ is called a *focal set*. The following mass functions are of particular interest:

- a *vacuous* mass is such that $m(\Omega) = 1$, and represents total ignorance;
- a *categorical* mass is such that $m(A) = 1$ for some A. It is equivalent to the set A;
- a *consonant* mass is such that its focal sets form a chain w.r.t. inclusion ordering, i.e., if A, B

are focal sets, then $A \subseteq B$ or $B \subseteq A$. It is equivalent to the fuzzy set that has its focal elements as α-cuts;

- a *Bayesian* mass is such that $m(A) > 0$ iff $|A| = 1$. It is equivalent to a probability distribution;
- a *certain* mass is such that $m(x) = 1$ for some singleton $x \in \Omega$.

2.1. Data uncertainty

Consider a quantity x that can take its values on a finite space Ω. If the true value of this quantity is ill-known, it can be modelled by a mass function. For instance, the mass $m_x(A) = 1$ models a quantity that is imprecisely observed and for which we only know that its true value lies in A. Similarly, a Bayesian mass function can model probabilistic uncertainty, and a consonant one a variable whose uncertain value is modelled by a fuzzy set.

We can associate the mass m_x with its plausibility function, such that

$$Pl_x(A) = \sum_{B \cap A \neq \emptyset} m(B),$$

for all $A \subseteq \Omega$. The function $pl_x : \Omega \to [0,1]$ such that $pl_x(w) = Pl_x(\{w\})$ is called the *contour function* associated to m_x. When m_x is consonant, the contour function contains all the information about m_x (which is not the case in general).

2.2. Evidential likelihood

Consider now a random variable X taking its values on a space Ω and following a parametric model $p_X(\cdot;\theta)$ with $\theta \in \Theta$ its parameter vector. The likelihood of θ for a perfect observation x is defined as $L(\theta;x) = p_X(x;\theta)$. If x is imprecisely observed as set-valued ($x \in A$), the imprecise likelihood of θ can be defined as

$$L(\theta;A) = p_X(A;\theta) = \sum_{x \in A} p_X(x;\theta). \quad \forall \theta \in \Theta$$

Furthermore, when the observation is both imprecise and uncertain, being represented by a mass function m_x, the evidential likelihood of θ is defined as [9]

$$\begin{aligned} L(\theta;m_x) &= \sum_{A \subseteq \Omega} L(\theta;A) m_x(A) \\ &= \sum_{\omega \in \Omega} p_X(\omega;\theta) \sum_{A \ni \omega} m_x(A) \\ &= \sum_{\omega \in \Omega} p_X(\omega;\theta) pl(\omega). \end{aligned} \tag{1}$$

As $L(\theta;m_x)$ only depends on the contour function pl induced by m_x, it can be written as $L(\theta;pl)$ instead. From (1), we have $L(\theta;pl) = \mathbb{E}_\theta[pl_x(X)]$. Now, if we consider a set $\boldsymbol{x} = (x_1,\ldots,x_n)$ of cognitively independent (see Denoeux [9] for a definition of cognitive independence) and i.i.d. uncertain observations, the evidential likelihood becomes

$$L(\theta;pl) = \prod_{i=1}^{n} \mathbb{E}_\theta[pl_{x_i}(X)] \tag{2}$$

2.3. Flatness of relative likelihood

Once we have computed the evidential likelihood by Eq. (2), the contour function (an extension of the so-called relative likelihood) pl_Θ is given by

$$pl_\Theta(\theta;m_{\boldsymbol{x}}) = \frac{L(\theta;m_{\boldsymbol{x}})}{\sup_{\theta \in \Theta} L(\theta;m_{\boldsymbol{x}})}, \tag{3}$$

which can be associated with an equivalent consonant mass function m_Θ. The flatness of $pl_\Theta(\theta;m_{\boldsymbol{x}})$, which we will denote $pl_\Theta(\theta)$ when no ambiguity can arise, can then be quantified by measuring how imprecise, or non-specific [10], is m_Θ (note that $pl_\Theta = 1$ is constant when all observations are vacuous).

This can be done by considering as a non-specificity measure of the mass function m the following extension of Hartley function [10]

$$NS(m) = \sum_{A \subseteq \Omega} m(A) \log_2 |A|$$

where Ω is the space in which is defined m. In the case of m_Θ, which is continuous, this measure would be given by

$$NS(m_\Theta) = \int_0^1 \log_2(|\{\theta | pl_\Theta(\theta) > \alpha\}|) d\alpha \tag{4}$$

where $|\{\theta|pl_\Theta(\theta) > \alpha\}|$ is an hyper-volume. In practice, computing this integral, and in particular estimating $|\{\theta|pl_\Theta(\theta) > \alpha\}|$, will be complex. This could be done through classical Monte-Carlo sampling techniques, yet it will be quickly impractical as Θ dimension grows. Also, it must be noticed that a high value of $NS(m_\Theta)$ may result from different effects: few data, numerous but uncertain data, uniformity of data, etc.

This is why, in the next sections, we study in detail the specific case of multinomial data, which are useful in practice and can give us an idea about the evidential likelihood's behaviour.

3. Non-specificity of multinomial evidential likelihood

We consider from now on that X is a multinomial variable taking its values on the space $\Omega = \{\omega_1,\ldots,\omega_\ell\}$ of ℓ outcomes, a typical situations in supervised and semi-supervised problems where X is the class variable. The parameter vector is then $\theta = (\theta_1,...,\theta_\ell)$ with $\Theta = \{\theta | \theta_i \in [0,1], \sum_{i=1}^{\ell} \theta_i = 1\}$ the (ℓ-1)-unit simplex.

3.1. Likelihood formulation

Given samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ whose uncertainty is modelled by mass functions, we will use the simplified notation $pl_i(j) := pl_{x_i}(\omega_j)$. Supposing stochastic and cognitive independence, the corresponding evidential likelihood is

$$L(\theta; pl) = \prod_{i=1}^{n} \mathbb{E}_\theta[pl_i(X)] = \prod_{i=1}^{n} \sum_{j=1}^{\ell} \theta_j pl_i(j). \quad (5)$$

The following result, whose proof can be found in Appendix A, will be useful.

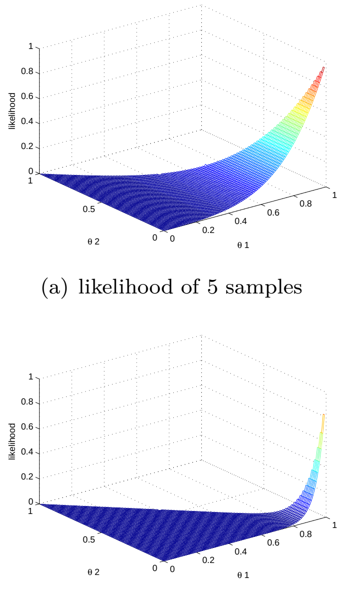
Proposition 1. *The contour function $pl_\Theta(\theta)$ corresponding to the evidential multinomial likelihood is a concave function.*

The non-specificity of $L(\theta; m_{\boldsymbol{x}})$ can then be estimated by Eq. (4), yet this can be tricky. In the next subsections we show that this flatness can be affected by various factors, and we then propose a technique to estimate the associated non-specificity.

3.2. Flatness behaviour

Before proposing various solutions to estimate or approximate $NS(m_\Theta)$, we can first wonder about what can affect the flatness, or imprecision of pl_Θ. For this reason, let us consider two examples with $\Omega = \{\omega_1, \omega_2, \omega_3\}$.

Example 1. Given data $\boldsymbol{x} = (x_1, \ldots, x_n)$, assume that every observation x_i has the same plausibility function $pl(\omega_1) = 1$, $pl(\omega_2) = 0.2$, $pl(\omega_3) = 0.2$. Calculating the evidential multinomial likelihood via a formula (5), the likelihoods for n=5 and n=24 are shown in Figure 1.

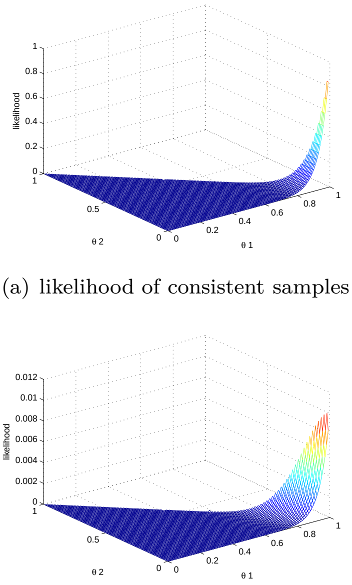


(a) likelihood of 5 samples

(b) likelihood of 24 samples

Figure 1: Effect of data quantity on likelihood

Example 2. Suppose there are 21 certain samples (i.e., modelled by certain masses). For comparison, for one case all 21 samples have the same plausibility function $pl(\omega_1) = 1$, $pl(\omega_2) = 0$, $pl(\omega_3) = 0$. For the other case, two of them are changed to $pl(\omega_1) = 0$, $pl(\omega_2) = 1$, $pl(\omega_3) = 0$. Figure 2 compares the likelihoods in these two cases.



(a) likelihood of consistent samples

(b) likelihood of inconsistent samples

Figure 2: Effect of data variability on likelihood

Example 1 shows that data quantity clearly affects pl_Θ flatness in an important way. However, when data can be uncertain (in the case of $n = 5$, we could imagine that we have 19 vacuous ones) and modelled by mass functions, both the quantity and the uncertainty have an effect, and it would be desirable to isolate this effect. Perhaps more surprisingly, Example 2 shows that data variability or empirical distribution can also have an effect on pl_Θ flatness, albeit of lesser importance. Indeed, in both cases of this example, the informational content or the uncertainty of the sample is the same (21 certain examples). It would be desirable to know when the likelihood imprecision mainly comes from this variability.

3.3. Estimation technique

Let us denote by $\hat{\theta}$ the maximal likelihood estimator (MLE), that is

$$\hat{\theta} = \arg\sup_{\theta \in \Theta} L(\theta; m_{\boldsymbol{x}}).$$

Thanks to Proposition 1, we know that the EM algorithm [9], or any efficient convex optimization technique, will provide the global maximum value.

To approximate m_Θ, let us consider a distance d between vectors of Θ and its maximal value $d_{\max}$. For example, if we select the L^2 norm distance measurement we have

$$d(\theta, \hat{\boldsymbol{\theta}}) = \|\theta - \hat{\theta}\|_2 = \sqrt{\sum_{i=1}^{\ell} (\theta_i - \hat{\theta}_i)^2}$$

and $d_{\max} = \sqrt{2}$. Let us now define a vector of values $v_1, \dots, v_D$ such that $v_i \in [0, d_{\max}]$ and $v_D = d_{\max}$. Define the region

$$\Gamma_{\boldsymbol{x}}(i) = \{\boldsymbol{\theta} \in \Theta \mid v_{i-1} \leq d(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \leq v_i\}$$

with $v_0 = 0$. Clearly, the sets $\Gamma_{\boldsymbol{x}}(i)$ form a partition of Θ. Now, consider the approximated contour function pl_{ap} such that for any $\theta \in \Gamma_{\boldsymbol{x}}(i)$, we have

$$pl_{ap}(\theta) = \max_{\theta' \in \Gamma_{\boldsymbol{x}}(i)} pl_\Theta(\theta')$$

and we denote by m_{ap} the associated consonant mass function. As pl_{ap} is constant on $\Gamma_{\boldsymbol{x}}(i)$, we will denote by $pl_{ap}(i)$ this value. Now, it is easy to see that m_{ap} value is

$$m_{ap}(\{\boldsymbol{\theta} \in \Theta \mid d(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \leq v_i\}) = pl_{ap}(i) - pl_{ap}(i+1)$$

for any $i \in 1, \dots, D$, with the convention $pl_{ap}(D + 1) = 0$. Since $pl_{ap} \geq pl_\Theta$ (with equality for some elements of $\Gamma_{\boldsymbol{x}}(i)$) by definition, m_{ap} is a consonant and conservative approximation of m_Θ, thus ensuring that the non-specificity of m_Θ will not be underestimated. Also note that as D grows, m_{ap} tends to m_Θ.

Remark 1. Note that $pl_{ap}(i) \geq pl_{ap}(i + 1)$ thanks to Proposition 1. This may not be true for non-concave functions.

Example 3. Consider $v_1 = 0.2, v_2 = 0.5, v_3 = \sqrt{2}$. They result in the sets

$$\Gamma_{\boldsymbol{x}}(1) = \{\{\boldsymbol{\theta} \in \Theta \mid d(\theta, \hat{\theta}) \in [0, 0.2]\}$$
$$\Gamma_{\boldsymbol{x}}(2) = \{\{\boldsymbol{\theta} \in \Theta \mid d(\theta, \hat{\theta}) \in (0.2, 0.5]\}$$
$$\Gamma_{\boldsymbol{x}}(3) = \{\{\boldsymbol{\theta} \in \Theta \mid d(\theta, \hat{\theta}) \in (0.5, \sqrt{2}]\}$$

and in the contour function

$$pl_{ap}(1) = max_{\{\boldsymbol{\theta} \in \Theta \mid d(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \in [0, 0.2]\}} pl_\Theta(\theta) = 1,$$
$$pl_{ap}(2) = max_{\{\boldsymbol{\theta} \in \Theta \mid d(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \in (0.2, 0.5]\}} pl_\Theta(\theta) = a,$$
$$pl_{ap}(3) = max_{\{\boldsymbol{\theta} \in \Theta \mid d(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \in (0.5, \sqrt{2}]\}} pl_\Theta(\theta) = b.$$

Then the consonant mass function is induced from the following formula:

$$m_{ap}(\{\boldsymbol{\theta} \in \Theta \mid d(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \leq 0.2\}) = 1 - a,$$
$$m_{ap}(\{\boldsymbol{\theta} \in \Theta \mid d(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \leq 0.5\}) = a - b,$$
$$m_{ap}(\{\boldsymbol{\theta} \in \Theta \mid d(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \leq \sqrt{2}\}) = b.$$

There is still one step missing to estimate NS, the cardinality of focal sets. Clearly, the hyper-volumes of $\Gamma_{\boldsymbol{x}}(i)$ will depend on the value of $\hat{\theta}$ and can be tricky to estimate, as they will correspond to hyper-spheres (in the case of L_2 norm) cut by hyper-planes. Since we are mainly interested in comparing and analyzing qualitatively non-specificities in different situations, and not in their exact values, we propose to simply consider $|\Gamma_{\boldsymbol{x}}(i)| = 1$ for any $i \in 1, \dots, D$. Unless values of NS are close to each other, this should not change too much their orders, provided we keep the same values $v_1, \dots, v_D$. The resulting value of non-specificity is then

$$NS_{ap}(m_{ap}) = \sum_{i=1}^{D} (pl_{ap}(i) - pl_{ap}(i+1)) \log_2 i$$

which is very easy to compute. For Example 3, we get

$$NS_{ap}(m_{ap}) = (a - b)log_2(2) + blog_2(3) = a + 0.58b \tag{6}$$

Note that this approach comes down to fix the sizes of the focal sets. Conversely, we could try to fix the values $1 > a_1 \geq \dots \geq a_D > 0$ and try to estimate the sets $\Psi(i)$ such that

$$\Psi(i) = \{\theta \in \Theta | pl_\Theta(\Theta) \geq a_i\},$$

that is, fixing the values of the discrete contour function approximating pl_Θ. Of course, this would not be easy in the general case, as $\Psi(i)$ will not be convex in general. However, in the case of multinomial distributions, Proposition 1 tells us that these sets will be convex, hence that clever sampling or estimation strategies can be used. We will not explore further this approach in this paper.

4. Measuring data quality via partitioning non-specificity

In this section, we propose to separate our $NS_{ap}(m_{ap})$ into three different aspects, corresponding to the three effects we have identified in Section 3.2:

- Data uncertainty;
- Data variability;
- Data quantity.

We will then provide some illustrative experiments and their associated results. Note that separating these aspects may be important in some applications, such as for example decision tree learning.

4.1. Proposed measures

Considering a sample $\boldsymbol{x} = (x_1, \dots, x_n)$ with uncertainty of each observation modelled by plausibilities pl_i on Ω, we define the following equivalences.

Definition 1 (Certain equivalent data)**.** The certain equivalent data $\boldsymbol{x^{ce}}$ of data $\boldsymbol{x}$ are defined as data having plausibilities

$$pl_i^{ce}(j) = \left\{ \begin{array}{ll} 1 & if\ \omega_j = \arg_{\omega_k \in \Omega} \max_k pl_i(\omega_k) \\ 0 & \qquad otherwise \end{array} \right. ,$$

$\forall i = 1, ..., n$.

Definition 2 (Certain consistent equivalent data). The certain consistent equivalent data $\boldsymbol{x^{ceco}}$ of data $\boldsymbol{x}$ are defined as data having plausibilities

$$pl_i^{ceco}(j) = \begin{cases} 1 & if\ \omega_j = \arg_{\omega_k \in \Omega} \max_k \hat{\theta}_k \\ 0 & otherwise \end{cases},$$

where $\hat{\theta}_k$ denotes the k-th component of optimal estimator $\hat{\boldsymbol{\theta}}$ obtained from data $\boldsymbol{x}$.

With the definitions above, the flatness of the likelihood quantitied by non-specificity can be analyzed in more detail. We propose to split the value of non-specificity into three parts NS_{unc}, NS_{inc}, and NS_{min}, which correspond to uncertainty, variability and quantity respectively:

$$NS_{ap}(m_{ap,\boldsymbol{x}}) = NS_{unc} + NS_{inc} + NS_{min}$$

where

$$NS_{unc} = NS_{ap}(m_{ap,\boldsymbol{x}}) - NS_{ap}(m_{ap,\boldsymbol{x}^{ce}}),$$

$$NS_{inc} = NS_{ap}(m_{ap,\boldsymbol{x}^{ce}}) - NS_{ap}(m_{ap,\boldsymbol{x}^{ceco}}),$$

$$NS_{min} = NS_{ap}(m_{ap,\boldsymbol{x}^{ceco}}),$$

are the difference between non-specificities calculated from $\boldsymbol{x}$ and $\boldsymbol{x^{ce}}$, the difference between non-specificities calculated from $\boldsymbol{x^{ce}}$ and $\boldsymbol{x^{ceco}}$, and the non-specificity of data $\boldsymbol{x^{ceco}}$.

It should be noted that NS_{inc} could be considered as non-reducible, while NS_{unc} and NS_{min} can be reduced by specifying uncertain data and by obtaining more samples, respectively.

4.2. Graphical and numerical illustrations

In this section, we consider the case where Θ is binary, making the multinomial distribution become a binomial distribution. This allows us to give more illustrative examples.

We assume that X is a binomial sample, each sample following a Bernoulli distribution. Original data are such that $pl_i(1) = 1, pl_i(0) = 0.2$ for $i = 1, ..., 15$ and $pl_i(1) = 0.2, pl_i(0) = 1$ for $i = 15, ..., 20$. Figure 3 shows the different contour functions obtained for data sets $\boldsymbol{x}^{ce}$ and $\boldsymbol{x}^{ceco}$. Obviously, the contour function obtained for $\boldsymbol{x}$ is flatter than the one obtained for $\boldsymbol{x}^{ce}$, which in turn is flatter than $\boldsymbol{x}^{ceco}$. The grey lines demonstrate how to achieve values a and b in formula (6) given $v_1 = 0.2, v_2 = 0.5, v_3 = \sqrt{2}$ as in Example 3.

We also varied the initial data set to show that our proposed splitting approach makes sense, as well as to show the influence of the three factors (uncertainty, variability and quantity). Figure 4 illustrates the contour functions of data with plausibility functions $pl_i(1) = 1, pl_i(0) = r, i = 1, ..., 20$, with r increasing from 0 to 1. Clearly, the higher is r, the more uncertainty there is in the data, and the flatter is the resulting contour function.

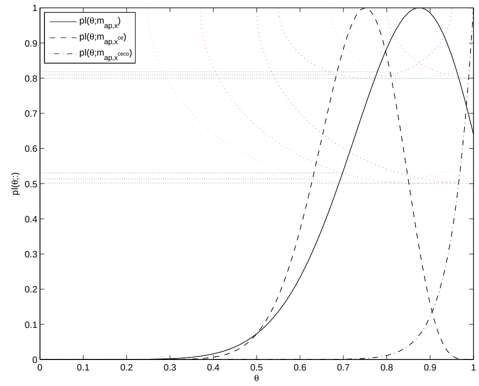


Figure 3: Comparison of contour functions obtained from original data and equivalences

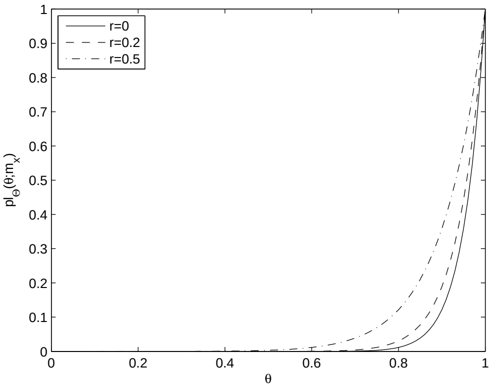


Figure 4: Contour functions for the Bernoulli distribution with consistent uncertain data

To test variability, we assumed that data $\boldsymbol{x} = (x_1, ..., x_n)$ are totally certain with plausibility functions $pl_i(1) = 1, pl_i(0) = 0$ for $i = 1, ..., 20 - INC$ and $pl_i(1) = 0, pl_i(0) = 1$ for $i = 21 - INC, ..., 20$. Figure 5 shows the evolution of the contour function as data become more inconsistent (variable), and the evolution of NS_{inc} numerical values can be seen in Table 1. As previously mentioned, we see that the effect of inconsistency on NS is minor, if not negligible.

Suppose now that every x_i is certain and has the same plausibility function $pl_i(1) = 1, pl_i(0) = 0$ for $i = 1, ..., n$. As the quantity of data n increases, the contour function becomes less flat, resulting in a smaller non-specificity. Figure 6 demonstrates this trend. Table 1 shows that our proposed splitting do identify where the flatness or imprecision of the contour function comes from.

We also performed some first experiments on data sets inspired from Examples 1 and 2. Table 2 lists the obtained results. The last two cases (lines) are certain data and the rest are uncertain. For those remaining cases, the first number is the number of uncertain samples with plausibility function $pl(\omega_1) = 1, pl(\omega_2) = 0.2, pl(\omega_3) = 0.2$, and

	NS	NS_{min}	NS_{unc}	NS_{inc}
r=0	0.2044	0.2044	0	0
r=0.2	0.2056	0.2044	0.0012	0
r=0.5	0.2088	0.2044	0.0044	0
INC=0	0.2044	0.2044	0	0
INC=5	0.2230	0.2044	0	0.0186
INC=10	0.2265	0.2044	0	0.0221
SCALE=5	0.2476	0.2476	0	0
SCALE=10	0.2144	0.2144	0	0
SCALE=20	0.2044	0.2044	0	0
$\boldsymbol{x}$	0.2311	0.2044	0.0081	0.0186
$\boldsymbol{x}^{ce}$	0.2230	0.2044	0	0.0186
$\boldsymbol{x}^{ceco}$	0.2044	0.2044	0	0

Table 1: Numerical results corresponding to data used in figures

	NS	NS_{min}	NS_{unc}	NS_{inc}
5	0.2733	0.2476	0.0257	0
4-1	0.3821	0.2476	0.0628	0.0717
24	0.2055	0.2034	0.0021	0
5	0.2733	0.2476	0.0257	0
21	0.2067	0.2044	0.0023	0
19-2	0.2120	0.2044	0.0031	0.0107
21(certain)	0.2044	0.2044	0	0
19-2(certain)	0.2151	0.2044	0	0.0107

Table 2: Numerical results under several situations

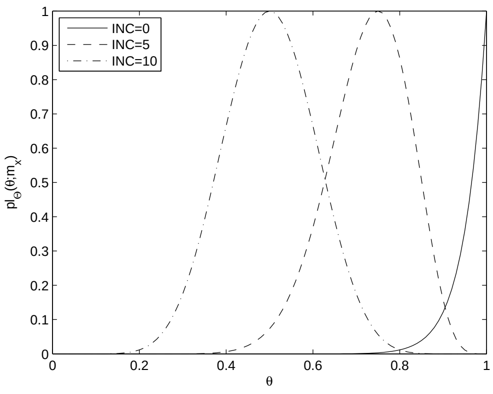


Figure 5: Contour functions for the Bernoulli distribution with increasingly inconsistent data

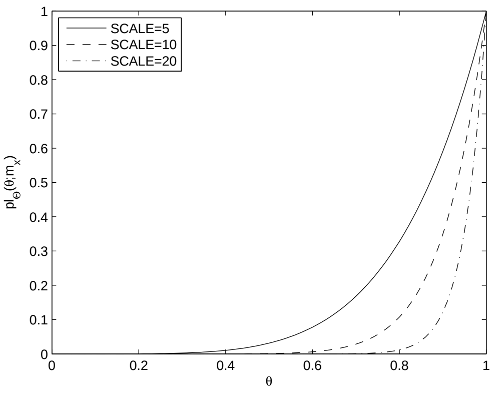


Figure 6: Contour functions for the Bernoulli distribution with data of different scales

the second number (when specified) is the number of uncertain samples with plausibility function $pl(\omega_1) = 0.2,\ pl(\omega_2) = 1,\ pl(\omega_3) = 0.2$.

The fact that no uncertainty exists in the last two cases is shown by $NS_{unc} = 0$. NS_{inc} can reflect to what extent pieces of evidence of data support the same outcome, and only inconsistent data (19-2, 4-1) will have non-zero NS_{inc} value. The value of NS_{min} shows that likelihood becomes more peaked as number of sample increases.

5. Conclusion

When using the relative likelihood to perform statistical estimation, it is commonly agreed that the flatness of the likelihood reflects the quantity of data, and when data uncertainty is modelled under the framework of belief functions, the amount of uncertainty. How to estimate this flatness in a single number, however, is still unclear. Here we have proposed to use non-specificity measures of mass functions to do so, seeing the relative likelihood as the contour function of a consonant mass function. For the particular case of multinomial likelihood, taking advantage of the function concavity, we have also proposed efficient estimation procedures.

However, the main message to recall from this paper is that there are other aspects than uncertainty or quantity, such as the distribution of the data, that can affect the flatness of the likelihood. How to isolate these various effects, as well as what really affects the evidential likelihood shape, clearly remains an open issue. Although we have proposed some preliminary answers for the case of the multinomial likelihood, this paper should mainly end with some questions: does the relative likelihood flatness really measure what we want it to measure? how can we in general separate the different factors that influence this flatness?

These are important questions, even for the multinomial case, as such distributions can be used to learn classifiers such as decision trees [11]. While the relative likelihood possesses some quite interesting properties, we should nevertheless remain cautious when using it in estimation or learning tasks.

References

[1] Carlo Batini and Monica Scannapieco. *Data quality: concepts, methodologies and techniques.* Springer, 2006.

[2] Anthony William Fairbank Edwards. *Likelihood.* CUP Archive, 1984.

[3] Allan Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

[4] Thierry Denoeux. Likelihood-based belief function: Justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547, 2014.

[5] Charu C Aggarwal. *Data Classification: Algorithms and Applications.* CRC Press, 2014.

[6] Donald P Ballou and Harold L Pazer. Modeling data and process quality in multi-input, multi-output information systems. *Management science*, 31(2):150–162, 1985.

[7] Yair Wand and Richard Y Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.

[8] Glenn Shafer et al. *A mathematical theory of evidence*, volume 1. Princeton university press Princeton, 1976.

[9] Thierry Denoeux. Maximum likelihood estimation from uncertain data in the belief function framework. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1):119–130, 2013.

[10] Didier Dubois and Henri Prade. A note on measures of specificity for fuzzy sets. *International Journal of General System*, 10(4):279–283, 1985.

[11] Nicolas Sutton-Charani, Sebastien Destercke, and Thierry Denoeux. Application of E^2M decision trees to rubber quality prediction. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 107–116. Springer, 2014.

A. Proof of Proposition 1

Proof. (i) The domain of function pl_Θ

$$\Theta = \{\theta | \theta_i \in [0,1], \sum_i \theta_i = 1, i = 1, ..., l\}$$

is a unit simplex, hence a convex set.

(ii) Given observations $\boldsymbol{x}$, take two vectors θ and θ' from Θ, define

$$g(\theta) = \log(L(\theta; m_{\boldsymbol{x}})) = \sum_{i=1}^{n} \log \sum_{j=1}^{\ell} \theta_j pl_i(j).$$

For any $\lambda \in [0,1]$, we have

$$g(\lambda\theta + (1-\lambda)\theta') = \sum_{i=1}^{n} \log \sum_{j=1}^{\ell} (\lambda\theta_j + (1-\lambda)\theta'_j) pl_i(j)$$

and

$$\begin{aligned}
&\lambda g(\theta) + (1-\lambda) g(\theta') \\
&= \sum_{i=1}^{n} [\lambda \log \sum_{j=1}^{\ell} \theta_j pl_i(j) + (1-\lambda) \log \sum_{j=1}^{\ell} \theta'_j pl_i(j)] \\
&= \sum_{i=1}^{n} \log[(\sum_{j=1}^{\ell} \theta_j pl_i(j))^{\lambda} \cdot (\sum_{j=1}^{\ell} \theta'_j pl_i(j))^{1-\lambda}].
\end{aligned}$$

Then,

$$\begin{aligned}
&g(\lambda\theta + (1-\lambda)\theta') - [\lambda g(\theta) + (1-\lambda) g(\theta')] \\
&= \sum_{i=1}^{n} \log \frac{\sum_{j=1}^{\ell} [\lambda\theta_j + (1-\lambda)\theta'_j] pl_i(j)}{[\sum_{j=1}^{\ell} \theta_j pl_i(j)]^{\lambda} \cdot [\sum_{j=1}^{\ell} \theta'_j pl_i(j)]^{1-\lambda}} \\
&= \sum_{i=1}^{n} \log \frac{\lambda \sum_{j=1}^{l} \theta_j pl_i(j) + (1-\lambda)[\sum_{j=1}^{\ell} \theta'_j pl_i(j)]}{[\sum_{j=1}^{\ell} \theta_j pl_i(j)]^{\lambda} \cdot [\sum_{j=1}^{\ell} \theta'_j pl_i(j)]^{1-\lambda}}
\end{aligned}$$

According to the weighted arithmetic mean and weighted geometric mean inequality, for any $x_1, x_2, \lambda \geq 0, \lambda x_1 + (1-\lambda) x_2 \geq x_1^{\lambda} \cdot x_2^{1-\lambda}$, therefore

$$g(\lambda\theta + (1-\lambda)\theta') - [\lambda g(\theta) + (1-\lambda) g(\theta')] \geq 0,$$

and $g(\theta)$ is a concave function. Since the exponential function is convex and increasing, the composition $e^{g(\theta)}$ remains concave. Normalization will not change concavity as well, hence pl_Θ is a concave function. □