

Statistical Analysis of Uncertain Data in the Belief Function Framework

Thierry Denœux¹

¹Université de Technologie de Compiègne
HEUDIASYC (UMR CNRS 6599)

NLMUA 2011
Beijing University of Technology
Beijing, September 8, 2011



Outline

- 1 Motivation and background
 - Motivation
 - Background on belief functions
- 2 Estimation from evidential data
 - Model and problem statement
 - Evidential EM algorithm
 - Example: uncertain Bernoulli sample
- 3 Partially supervised LDA
 - Model and algorithm
 - Experimental results



Introductory example

- Let us consider a population in which some disease is present in proportion θ .
- n patients have been selected **at random** from that population. Let $x_i = 1$ if patient i has the disease, $x_i = 0$ otherwise. Each x_i is a realization of $X_i \sim \mathcal{B}(\theta)$.
- We assume that the x_i 's are **not observed directly**. For each patient i , a physician gives a **degree of plausibility** $pl_i(1)$ that patient i has the disease and a **degree of plausibility** $pl_i(0)$ that patient i does not have the disease.
- The observations are **uncertain data** of the form pl_1, \dots, pl_n .
- How to estimate θ ?

Aleatory vs. epistemic uncertainty

In the previous example, uncertainty has **two distinct origins**:

- 1 **Before** a patient has been drawn at random from the population, the uncertainty is due to the **variability** of the variable of interest in the population. This **aleatory uncertainty** cannot be reduced.
- 2 **After** the random experiment has been performed, the uncertainty is due to **lack of knowledge** of the state of each particular patient. This **epistemic uncertainty** could be reduced by carrying out further investigations.

Approach

- In this talk, I will consider statistical estimation problems in which **both kinds of uncertainty are present**: it will be assumed that each data item x
 - has been generated at random from a population (aleatory uncertainty), but
 - it is ill-known because of imperfect measurement or perception (epistemic uncertainty).
- The proposed model treats these two kinds of uncertainty separately:
 - **Aleatory uncertainty** will be represented by a **parametric statistical model**;
 - **Epistemic uncertainty** will be represented using **belief functions**.



Real world applications

Uncertain data arise in many applications (but epistemic uncertainty is usually neglected). It may be due to:

- **Limitations of the underlying measuring equipment** (unreliable sensors, indirect measurements), e.g.: biological sensor for toxicity measurement in water.
- Use of **imputation, interpolation or extrapolation techniques**, e.g.: clustering of moving objects whose position is measured asynchronously by a sensor network,
- **Partial or uncertain responses in surveys or subjective data annotation**, e.g.: sensory analysis experiments, data labeling by experts, etc.

Motivation

Data labeling example

Recognition of facial expressions

joy



surprise



sadness



disgust



anger



fear



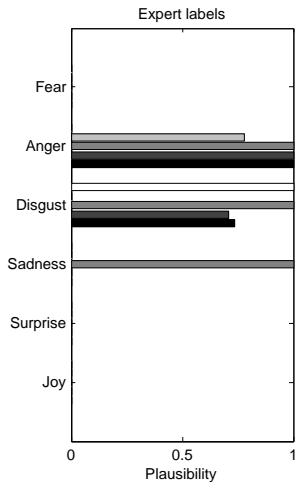
Recognition of facial expressions

Experiment

- In this kind of problem (object classification in images or videos) the **ground truth is usually unknown** or difficult to determine with high precision and reliability.
- It is then necessary to have the images subjectively annotated (labeled) by humans.
- How to **account for uncertainty** in such subjective annotations?
- Experiment:
 - 108 images were labeled by 5 subjects;
 - For each image, subjects were asked to give a **degree of plausibility** for each of the 6 basic expressions.



Example 1



Example 2



Example 3



Model

- **Complete data**: $\mathbf{x} = \{(\mathbf{w}_i, z_i)\}_{i=1}^n$ with
 - \mathbf{w}_i : feature vector for image i (pixel gray levels)
 - z_i : class of image i (one the six expressions).
- The feature vectors \mathbf{w}_i are perfectly observed but class labels are only **partially known** through subjective evaluations.
- How to **learn a decision rule** from such data?

General approach

- 1 Postulate a parametric statistical model $p_{\mathbf{x}}(\mathbf{x}; \theta)$ for the complete data;
- 2 Represent uncertain observed data using **belief functions**;
- 3 Estimate θ by **minimizing the conflict** between the model and the observed data using an extension of the **EM algorithm**: the evidential EM (E²M) algorithm.

Outline

- 1 Motivation and background
 - Motivation
 - Background on belief functions
- 2 Estimation from evidential data
 - Model and problem statement
 - Evidential EM algorithm
 - Example: uncertain Bernoulli sample
- 3 Partially supervised LDA
 - Model and algorithm
 - Experimental results



Theory of belief functions

- A formal framework for representing and reasoning with uncertain information.
- Introduced by Dempster (1968) and Shafer (1976), further developed by Smets in the 1980's and 1990's.
- Also known as **Dempster-Shafer theory**, **Evidence theory** or **Transferable Belief Model**.
- Many applications in statistics, artificial intelligence, pattern recognition, machine learning, information fusion, etc.



Mass function

- Let X be a variable taking values in a finite domain Ω , called the **frame of discernment**.
- We collect a **piece of evidence (information)** about X .
- This piece of evidence has different interpretations $\theta_1, \dots, \theta_r$ with corresponding subjective probabilities p_1, \dots, p_r .
- If interpretation θ_i holds, we only know that $X \in A_i$ for some $A_i \subseteq \Omega$, and nothing more. Let $A_i = \Gamma(\theta_i)$.
- The probability that the evidence means exactly that $X \in A$ is $m(A) = \sum_{\{i|A_i=A\}} p_i$.
- Function $m : 2^\Omega \rightarrow [0, 1]$ is called a **mass function** with **focal sets** A_1, \dots, A_r .

Example

- A murder has been committed. There are three suspects:
 $\Omega = \{Peter, John, Mary\}$.
- A witness saw the murderer going away, but he is short-sighted and he only saw that it was a man. We know that the witness is drunk 20 % of the time.
- Two interpretations:
 - ① $\theta_1 =$ the witness was not drunk, $p_1 = 0.8$;
 - ② $\theta_2 =$ the witness was drunk, $p_2 = 0.2$.
- We have $\Gamma(\theta_1) = \{Peter, John\}$ and $\Gamma(\theta_2) = \Omega$, hence

$$m(\{Peter, John\}) = 0.8, \quad m(\Omega) = 0.2$$



Belief and plausibility functions

- The total **degree of support** for A is

$$Bel(A) = P(\{\theta \in \Theta | \Gamma(\theta) \subseteq A\}) = \sum_{B \subseteq A} m(B).$$

Function $Bel : 2^\Omega \rightarrow [0, 1]$ is called a **belief function**. It is a completely monotone capacity.

- The **plausibility** of A is the degree to which the evidence does not contradict A . It is defined as

$$Pl(A) = 1 - Bel(\bar{A}) = \sum_{B \cap A \neq \emptyset} m(B)$$

- Function $pl : \omega \rightarrow Pl(\{\omega\})$ is called the **contour function**.



Special cases

- 1 If all focal elements of m are singletons, then m is said to be **Bayesian**: it is equivalent to a **probability distribution**, and $Bel = Pl$ is a probability measure.
- 2 If the focal elements of m are nested, then Pl is a **possibility measure**, i.e.,

$$Pl(A \cup B) = \max(Pl(A), Pl(B)), \quad \forall A, B \subseteq \Omega,$$

bel is the dual **necessity measure**, and the contour function pl is then the associated **possibility distribution**.



Dempster's rule

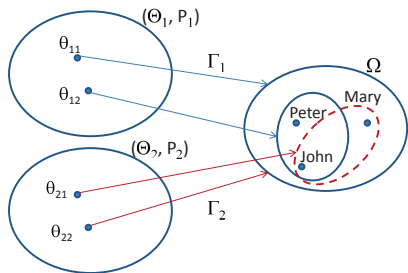
Murder example continued

- The first item of evidence gave us:
 $m_1(\{Peter, John\}) = 0.8, m_1(\Omega) = 0.2.$
- New piece of evidence: a blond hair has been found.
- There is a probability 0.6 that the room has been cleaned before the crime: $m_2(\{John, Mary\}) = 0.6, m_2(\Omega) = 0.4.$
- How to combine these two pieces of evidence?



Dempster's rule

Justification



- If $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$ both hold, then $X \in \Gamma_1(\theta_1) \cap \Gamma_2(\theta_2)$.
- If the two pieces of evidence are **independent**, then this happens with probability $P_1(\{\theta_1\})P_2(\{\theta_2\})$.
- If $\Gamma_1(\theta_1) \cap \Gamma_2(\theta_2) = \emptyset$, we know that the pair of interpretations (θ_1, θ_2) is impossible.
- The joint probability distribution on $\Theta_1 \times \Theta_2$ must be conditioned, eliminating such pairs.



Dempster's rule

Definition

$$(m_1 \oplus m_2)(A) = \frac{1}{1 - K} \sum_{B \cap C = A} m_1(B)m_2(C), \quad \forall A \neq \emptyset,$$

where

$$K = \sum_{B \cap C = \emptyset} m_1(B)m_2(C)$$

is the **degree of conflict** between m_1 and m_2 .

Dempster's rule

Combination with a Bayesian mass function

- Let m_1 be an **arbitrary mass function** and let m_2 be a **Bayesian** mass function with corresponding probability distribution p_2 .
- The combined mass function **m_{12} is Bayesian**. Its corresponding probability distribution is:

$$p_{12}(\omega) = \frac{p l_1(\omega) p_2(\omega)}{1 - K} \quad \forall \omega \in \Omega$$

with

$$K = 1 - \sum_{\omega' \in \Omega} p l_1(\omega') p_2(\omega').$$



Cognitive independence

- Let X and Y be two variables defined on Ω_X and Ω_Y , and let m^{XY} be a **joint mass function** on $\Omega_X \times \Omega_Y$.
- The **marginal** mass function on Ω_X is defined as

$$m^{XY \downarrow X}(A) = \sum_{\{C \downarrow \Omega_X = A\}} m^{XY}(C), \quad \forall A \subseteq \Omega_X,$$

where $C \downarrow \Omega_X =$ the projection of $C \subseteq \Omega_X \times \Omega_Y$ on Ω_X .

- X and Y are said to be **cognitively independent** with respect to m^{XY} if:

$$Pl^{XY}(A \times B) = Pl^X(A)Pl^Y(B), \quad \forall A \subseteq \Omega_X, \forall B \subseteq \Omega_Y.$$

- Interpretation: new evidence on one variable does not affect our beliefs in the other variable.

Outline

- 1 Motivation and background
 - Motivation
 - Background on belief functions
- 2 Estimation from evidential data
 - **Model and problem statement**
 - Evidential EM algorithm
 - Example: uncertain Bernoulli sample
- 3 Partially supervised LDA
 - Model and algorithm
 - Experimental results



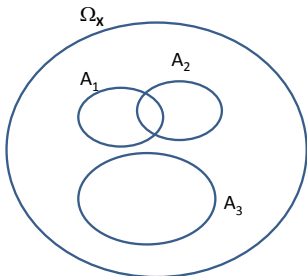
Model

- Let \mathbf{X} be a (discrete) random vector taking values in $\Omega_{\mathbf{X}}$, with probability mass function $p_{\mathbf{X}}(\cdot; \theta)$ depending on an **unknown parameter** $\theta \in \Theta$.
- Let \mathbf{x} be a realization of \mathbf{X} (**complete data**).
- We assume that \mathbf{x} is only **partially observed**, and partial knowledge of \mathbf{x} is described by a **mass function** m on $\Omega_{\mathbf{X}}$ (“observed” data).
- Problem: estimate θ .



Generalized Likelihood function

Definition



- Assume that m has focal sets A_1, \dots, A_r .
- If we new that $\mathbf{x} \in A_j$, the likelihood would be

$$L(\theta; A_j) = \mathbb{P}_{\mathbf{X}}(A_j; \theta) = \sum_{\mathbf{x} \in A_j} p_{\mathbf{X}}(\mathbf{x}; \theta).$$

- General case:

$$L(\theta; m) = \sum_{i=1}^r m(A_i) L(\theta; A_i)$$

Generalized Likelihood function

Interpretation

- It can be checked that $L(\theta; m)$ can be written as:

$$L(\theta; m) = 1 - \sum_{\mathbf{x} \notin \Omega_{\mathbf{x}}} p_{\mathbf{x}}(\mathbf{x}; \theta) p_l(\mathbf{x})$$

- $L(\theta; m)$ is equal to **one minus the degree of conflict** between $p_{\mathbf{x}}(\cdot; \theta)$ and m .
- Consequently, maximizing $L(\theta; m)$ with respect to θ amounts to **minimizing the conflict** between the parametric model and the uncertain observations.

Generalized Likelihood function

Case of fuzzy data

- Other expression for $L(\theta; m)$:

$$L(\theta; m) = \sum_{\mathbf{x} \in \Omega_{\mathbf{X}}} p_{\mathbf{X}}(\mathbf{x}; \theta) pl(\mathbf{x}) = \mathbb{E}_{\theta} [pl(\mathbf{X})]$$

- If m is **consonant**, pl may be interpreted as the membership function of a fuzzy subset of $\Omega_{\mathbf{X}}$: we have **fuzzy data**.
- $L(\theta; m)$ is then the **probability of the fuzzy data**, in the sense of Zadeh (1968).



Independence assumptions

- Let us assume that $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{np}$, where each \mathbf{x}_i is a realization from a p -dimensional random vector \mathbf{X}_i .
- Independence assumptions:

- Stochastic independence** of $\mathbf{X}_1, \dots, \mathbf{X}_n$:

$$p_{\mathbf{x}}(\mathbf{x}; \theta) = \prod_{i=1}^n p_{\mathbf{x}_i}(\mathbf{x}_i; \theta), \quad \forall \mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \Omega_{\mathbf{x}}$$

- Cognitive independence** of $\mathbf{x}_1, \dots, \mathbf{x}_n$ with respect to m :

$$p_l(\mathbf{x}) = \prod_{i=1}^n p_{l_i}(\mathbf{x}_i), \quad \forall \mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \Omega_{\mathbf{x}}$$

- Under these assumptions:

$$\log L(\theta; m) = \sum_{i=1}^n \log \mathbb{E}_{\theta} [p_{l_i}(\mathbf{X}_i)].$$

Outline

- 1 Motivation and background
 - Motivation
 - Background on belief functions
- 2 Estimation from evidential data
 - Model and problem statement
 - **Evidential EM algorithm**
 - Example: uncertain Bernoulli sample
- 3 Partially supervised LDA
 - Model and algorithm
 - Experimental results



Description

- The generalized log-likelihood function $\log L(\theta; m)$ can be maximized using an **iterative algorithm** composed of two steps:

E-step: Compute the expectation of $\log L(\theta; \mathbf{x})$ with respect to $m \oplus p_{\mathbf{X}}(\cdot; \theta^{(q)})$:

$$Q(\theta, \theta^{(q)}) = \frac{\sum_{\mathbf{x} \in \Omega_{\mathbf{X}}} \log(L(\theta; \mathbf{x})) p_{\mathbf{X}}(\mathbf{x}; \theta^{(q)}) pl(\mathbf{x})}{\sum_{\mathbf{x} \in \Omega_{\mathbf{X}}} p_{\mathbf{X}}(\mathbf{x}; \theta^{(q)}) pl(\mathbf{x})}.$$

M-step: Maximize $Q(\theta, \theta^{(q)})$ with respect to θ .

- E- and M-steps are iterated until the increase of $\log L(\theta; m)$ becomes smaller than some threshold.



Properties

- 1 When m is categorical: $m(A) = 1$ for some $A \subseteq \Omega$, then the previous algorithm reduces to the EM algorithm \rightarrow **evidential EM (E²M) algorithm.**
- 2 Monotonicity: any sequence $L(\theta^{(q)}; m)$ for $q = 0, 1, 2, \dots$ of generalized likelihood values obtained using the E²M algorithm is non decreasing, i.e., it verifies

$$L(\theta^{(q+1)}; m) \geq L(\theta^{(q)}; m), \quad \forall q.$$

- 3 The algorithm **only uses the contour function pI** , which drastically reduces the complexity of calculations.



Outline

- 1 Motivation and background
 - Motivation
 - Background on belief functions
- 2 Estimation from evidential data
 - Model and problem statement
 - Evidential EM algorithm
 - Example: uncertain Bernoulli sample
- 3 Partially supervised LDA
 - Model and algorithm
 - Experimental results



Model and data

- Let us assume that the complete data $\mathbf{x} = (x_1, \dots, x_n)$ is a realization from an i.i.d. sample X_1, \dots, X_n from $\mathcal{B}(\theta)$ with $\theta \in [0, 1]$.
- We only have **partial information** about the x_i 's in the form: pI_1, \dots, pI_n , where $pI_i(x)$ is the plausibility that $X_i = x$, $x \in \{0, 1\}$.
- Under the cognitive independence assumption:

$$\log L(\theta; pI_1, \dots, pI_n) = \sum_{i=1}^n \log [(1 - \theta)pI_i(0) + \theta pI_i(1)].$$

E- and M-steps

Complete data log-likelihood:

$$\log L(\theta, \mathbf{x}) = n \log(1 - \theta) + \log \left(\frac{\theta}{1 - \theta} \right) \sum_{i=1}^n x_i.$$

E-step: compute

$$Q(\theta, \theta^{(q)}) = n \log(1 - \theta) + \log \left(\frac{\theta}{1 - \theta} \right) \sum_{i=1}^n \xi_i^{(q)}, \text{ with}$$

$$\xi_i^{(q)} = \mathbb{E}_{\theta^{(q)}} [X_i | p_i] = \frac{\theta^{(q)} p_i(1)}{(1 - \theta^{(q)}) p_i(0) + \theta^{(q)} p_i(1)}.$$

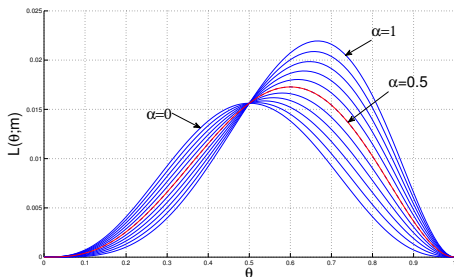
M-step:

$$\theta^{(q+1)} = \frac{1}{n} \sum_{i=1}^n \xi_i^{(q)}.$$

Example: uncertain Bernoulli sample

Numerical example

i	1	2	3	4	5	6
$pl_i(0)$	1	1	1	α	0	0
$pl_i(1)$	0	0	0	$1 - \alpha$	1	1



$$\alpha = 0.5$$

q	$\theta^{(q)}$	$L(\theta^{(q)}; pl)$
0	0.3000	6.6150
1	0.5500	16.8455
2	0.5917	17.2676
3	0.5986	17.2797
4	0.5998	17.2800
5	0.6000	17.2800

$$\hat{\theta} = 0.6$$

Outline

- 1 Motivation and background
 - Motivation
 - Background on belief functions
- 2 Estimation from evidential data
 - Model and problem statement
 - Evidential EM algorithm
 - Example: uncertain Bernoulli sample
- 3 Partially supervised LDA
 - Model and algorithm
 - Experimental results



Object classification

Problem statement

- We consider a population of **objects** partitioned in g **classes**.
- Each object is described by d **continuous features** $\mathbf{W} = (W^1, \dots, W^d)$ and a class variable Z .
- The goal of **discriminant analysis** is to learn a **decision rule** that classifies any object from its feature vector, based on a learning set.



Object classification

Learning tasks

- Classically, different learning tasks are considered:

Supervised learning: $\mathcal{L}_S = \{(\mathbf{w}_i, z_i)\}_{i=1}^n$;

Unsupervised learning: $\mathcal{L}_{ns} = \{\mathbf{w}_i\}_{i=1}^n$;

Semi-supervised learning: $\mathcal{L}_{SS} = \{(\mathbf{w}_i, z_i)\}_{i=1}^{n_s} \cup \{\mathbf{w}_i\}_{i=n_s+1}^n$

- Here, we consider **partially supervised learning**:

$$\mathcal{L}_{ps} = \{(\mathbf{w}_i, m_i)\}_{i=1}^n,$$

where m_i is a mass function representing **partial information** about the class of object i .

- This problem can be solved using the E²M algorithm using a suitable parametric model.



Linear discriminant analysis

- Generative model:
 - Complete data: $\mathbf{x} = \{(\mathbf{w}_i, z_i)\}_{i=1}^n$, assumed to be a realization of an iid random sample $\mathbf{X} = \{(\mathbf{W}_i, Z_i)\}_{i=1}^n$;
 - Given $Z_i = k$, \mathbf{W}_i is multivariate normal with mean μ_k and common variance matrix Σ .
 - The proportion of class k in the population is π_k .
 - Parameter vector: $\theta = (\{\pi_k\}_{k=1}^g, \{\mu_k\}_{k=1}^g, \Sigma)$.
- The Bayes rule is approximated by assigning each object to the class k^* that maximizes the estimated posterior probability

$$p(Z = k | \mathbf{w}; \hat{\theta}) = \frac{\phi(\mathbf{w}; \hat{\mu}_k, \hat{\Sigma}) \hat{\pi}_k}{\sum_{\ell} \phi(\mathbf{w}; \hat{\mu}_{\ell}, \hat{\Sigma}) \hat{\pi}_{\ell}},$$

where $\hat{\theta}$ is the MLE of θ .

Observed-data likelihood

- In partially supervised learning, the **observed-data log-likelihood** has the following expression:

$$\log L(\theta; \mathcal{L}_{ps}) = \sum_{i,k}^n pl_{ik} \log (\pi_k \phi(\mathbf{w}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)),$$

where pl_{ik} is the plausibility that object i belongs to class k .

- **Supervised learning** is recovered as a special case when:

$$pl_{ik} = z_{ik} = \begin{cases} 1 & \text{if object } i \text{ belongs to class } k; \\ 0 & \text{otherwise.} \end{cases}$$

- **Unsupervised learning** is recovered when $pl_{ik} = 1$ for all i and k .

E²M algorithm

E-step: Using $p_{\mathbf{x}}(\cdot; \theta^{(q)}) \oplus m$, compute

$$t_{ik}^{(q)} = \mathbb{E}(Z_{ik} | m; \theta^{(q)}) = \frac{\pi_k^{(q)} p_{l_{ik}} \phi(\mathbf{w}_i; \mu_k^{(q)}, \Sigma^{(q)})}{\sum_{\ell} \pi_{\ell}^{(q)} p_{l_{i\ell}} \phi(\mathbf{w}_i; \mu_{\ell}^{(q)}, \Sigma^{(q)})}$$

M-step: Update parameter estimates

$$\pi_k^{(q+1)} = \frac{1}{n} \sum_{i=1}^n t_{ik}^{(q)}, \quad \mu_k^{(q+1)} = \frac{\sum_{i=1}^n t_{ik}^{(q)} \mathbf{w}_i}{\sum_{i=1}^n t_{ik}^{(q)}}$$

$$\Sigma^{(q+1)} = \frac{1}{n} \sum_{i,k} t_{ik}^{(q)} (\mathbf{w}_i - \mu_k^{(q+1)})(\mathbf{w}_i - \mu_k^{(q+1)})'$$



Outline

- 1 Motivation and background
 - Motivation
 - Background on belief functions
- 2 Estimation from evidential data
 - Model and problem statement
 - Evidential EM algorithm
 - Example: uncertain Bernoulli sample
- 3 Partially supervised LDA
 - Model and algorithm
 - Experimental results



Face recognition problem

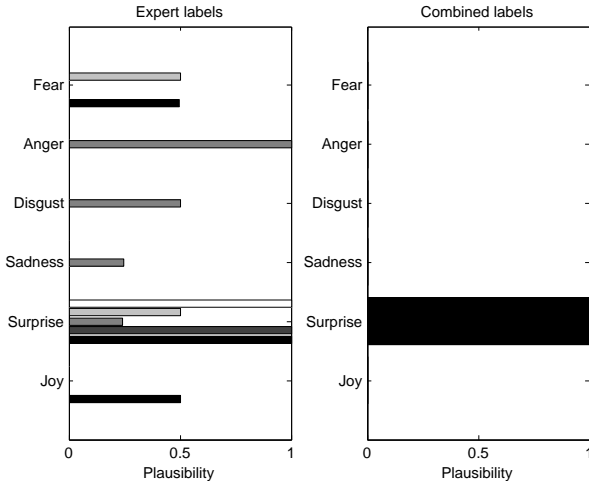
Experimental settings

- 216 images of 60×70 pixels, 36 in each class.
- One half for training, the rest for testing.
- A reduced number of features was extracted using Principal component analysis (PCA).
- Each training image was labeled by 5 subjects who gave **degrees of plausibility** for each image and each class.
- The plausibilities were combined using **Dempster's rule** (after some discounting to avoid total conflict).



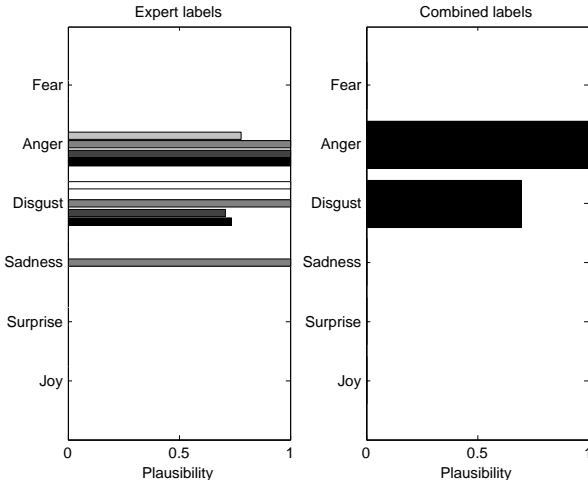
Combined labels

Example 1



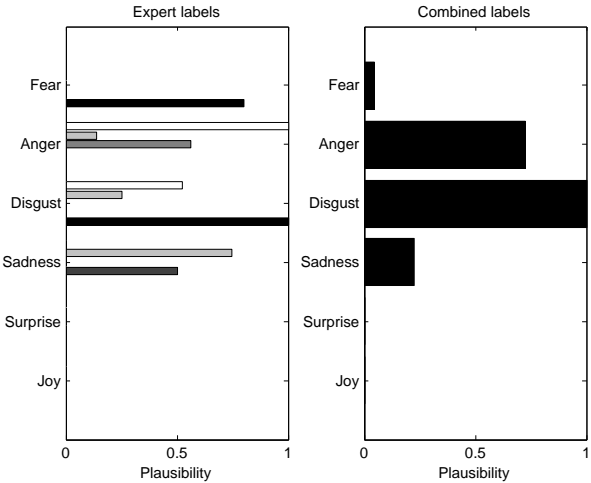
Combined labels

Example 2

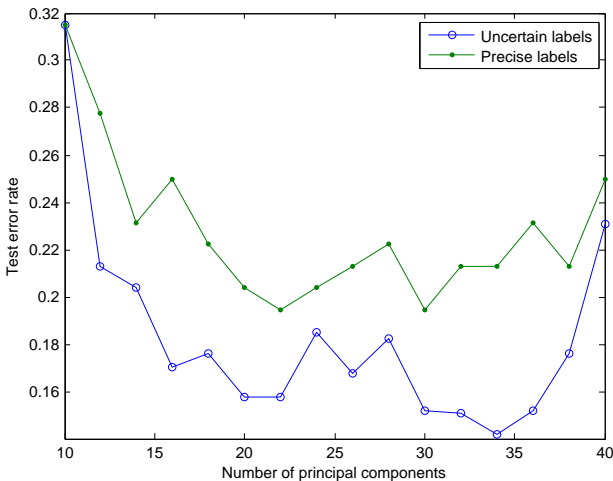


Combined labels

Example 3



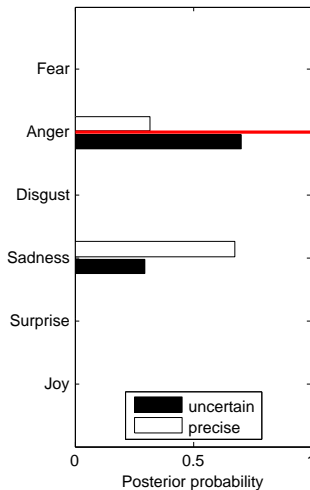
Results



Results

Example 1

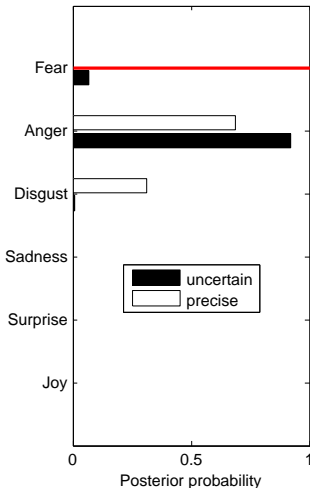
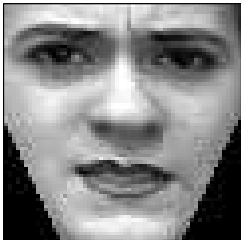
Test image 14



Results

Example 2

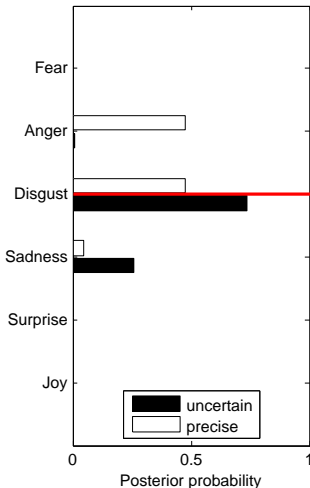
Test image 37



Results

Example 3

Test image 48



Summary

- The formalism of belief functions provides a very general setting for representing data uncertainty → **evidential data**.
- Maximizing the proposed generalized likelihood amounts to **minimizing the conflict between the uncertain data and the parametric model**.
- This can be achieved using an iterative algorithm (**evidential EM algorithm**) that reduces to the standard EM algorithm in special cases.
- In classification, the method makes it possible to handle **uncertainty on class labels** (partially supervised learning).
 Uncertainty on attributes can be handled as well.



Research challenges/Ongoing work

- The E²M algorithm can be applied to any problem involving a **parametric statistical model** and **epistemic uncertainty on observations**, e.g.:
 - Independent factor analysis (paper by Cherfi et al. in this conference);
 - Clustering of fuzzy data using Gaussian mixture models (Quost and Denœux, 2010);
 - Hidden Markov models (ongoing).
- Some open problems:
 - How to **elicit** subjective evaluations in the Dempster-Shafer framework?
 - When observations become uncertain or imprecise, this **uncertainty should be reflected in the parameter estimates**
How to do it in the proposed framework?



References



T. Denœux.

Maximum likelihood estimation from fuzzy data using the EM algorithm. *Fuzzy Sets and Systems*, 183:72-91, 2011.



T. Denœux.

Maximum likelihood estimation from Uncertain Data in the Belief Function Framework. *IEEE Trans. Knowledge and Data Engineering*, to appear, 2011.

`http://www.hds.utc.fr/~tdenoeux`

THANK YOU!