INFERENCE ON PARAMETERS IN POSSIBILITY DISTRIBUTIONS, WITH SPECIAL REFERENCE TO POSSIBILISTIC REGRESSION

Wolfgang NÄTHER

Mining Academy of Freiberg
Department of Mathematics
9200 Freiberg, GDR

Summary: The possibilistic linear regression is used to point out the nonstatistical character of possibilistic inference. It suffices to have a sample size which ensures identifiability of the parameters. In a more general setting it is shown that Archimedean decomposable measures preserve statistical properties, but possibility is non-Archimedean.

1. Introduction

At first we recall the basic notions in possibility theory (see e.g. ZADEH /6/, DUBOIS/PRADE /1/). Let $\mathcal{B}_U$ be a σ-algebra of subsets from the universe U which contains all one-point sets $\{x\}$, $x \in U$. A set function $\Pi | \mathcal{B}_U \to [0,1]$ with $\Pi(\emptyset) = 0$, $\Pi(U) = 1$ and

$$\forall A,B \in \mathcal{B}_U: \Pi(A \cup B) = \max\{\Pi(A), \Pi(B)\} \tag{1}$$

is called possibility measure (abbr.: pm) on $(U, \mathcal{B}_U)$. Note that

$$\Pi(\{x\}) =: g(x) \; ; \; x \in U \tag{2}$$

defines the possibility distribution function (abbr.: pdf) of Π which leads to

$$\Pi(A) = \sup_{x \in A} g(x) \tag{3}$$

for a crisp subset A of U and to

$$\Pi(A) = \sup_{x \in A} \min\{\mu_A(x), g(x)\} \tag{4}$$

for a fuzzy subset A of U with membership function μ_A.

Let $(V,\mathcal{B}_V)$ be another measurable space. Any $(\mathcal{B}_U,\mathcal{B}_V)$ measurable mapping $Y\,|\,U \to V$ is called <u>possibilistic variable</u> (abbr.: pv) and the pm Π_Y on $\mathcal{B}_V$ is generated by

$$\forall A \in \mathcal{B}_V:\ \Pi_Y(A) = \Pi(Y^{-1}(A)) \tag{5}$$

As a consequence of (1) the pdf of Π_Y is given by

$$g_Y(y) = \Pi(\{x:\ Y(x)=y\}) = \Pi(\bigcup_{Y(x)=y} \{x\}) = \sup_{Y(x)=y} g(x)\ . \tag{6}$$

Similarly the pm and the pdf induced by a measurable mapping of the pv Y, say $f(Y) = Z$, is given by

$$\Pi_Z(B) = \Pi_Y(f^{-1}(B))$$
$$g_Z(z) = \Pi_Y(\{y:\ f(y)=z\}) = \sup_{f(y)=z} g_Y(y) \tag{7}$$

Note that in the frame of fuzzy set theory (7) is known as <u>extension principle</u>.

In the following we are interested in possibilistic vectors, i.e. in pv's $Y\,|\,U \to R^n$. Let $Y = (Y_1,\ldots,Y_n)^T$ a possibilistic vector. The marginal pdf is given (as a special case of (7)) by

$$g_{Y_i}(y_i) = \sup_{y_j,\,j\neq i} g_Y(y_1,\ldots,y_n)\ ;\ i=1,\ldots,n\ . \tag{8}$$

The possibilistic vector $Y = (Y_1,\ldots,Y_n)^T$ has <u>min-related</u> (noninteractive) components iff

$$g_Y(y_1,\ldots,y_n) = \min\,(g_{Y_1}(y_1),\ldots,g_{Y_n}(y_n))\ . \tag{9}$$

Our problem is the following: Assume the pdf of a pv Y is given up to an unknown parameter ϑ, say $g_Y(y,\vartheta)$; $\vartheta \in \Theta$. Assume further, there are available n observations of Y, say $y_1,\ldots,y_n$. How to "estimate" ϑ ? How to evaluate such estimates? What are main differences between possibilistic and probabilistic inference? To answer this questions we suppose:

<u>Ass.1</u>: The sample $y_1,\ldots,y_n$ is interpreted as realization of an n-dimensional possibilistic vector Y with min-related identically distributed components, the pdf of which is $g_Y(y,\vartheta)$ ∎

An **estimator** $\hat{\vartheta} = \hat{\vartheta}(Y_1,\dots,Y_n)$ of ϑ is, again, a pv generated by the mapping $\hat{\vartheta}|R^n \to \Theta$. With (7), Ass.1 and (9) the induced pdf of $\hat{\vartheta}$ for given ϑ writes

$$g_{\hat{\vartheta}/\vartheta}(t) = \sup_{\hat{\vartheta}(y_1,\dots,y_n)=t} \min\{g_Y(y_1,\vartheta),\dots,g_Y(y_n,\vartheta)\} \qquad (10)$$

To evaluate $\hat{\vartheta}$ we take a suitable loss function $L(\vartheta,\hat{\vartheta})|\Theta\times\Theta \to [0,1]$ which measures the relative loss or the degree of loss. Note that $L(\vartheta,\hat{\vartheta})$ can be interpreted as the membership function of the fuzzy set LOSS. Having in mind statistical inference, the risk of $\hat{\vartheta}$ is expressed by

$$R_P(\vartheta,\hat{\vartheta}) = \int_{\Theta} L(\vartheta,\cdot)\, dP_{\hat{\vartheta}/\vartheta} \;. \qquad (11)$$

Following ZADEH's definition of probability for fuzzy events we have the interpretation

$$R_P(\vartheta,\hat{\vartheta}) = P(\hat{\vartheta} \in \text{LOSS}) \;. \qquad (12)$$

Analogously we define the possibilistic risk of $\hat{\vartheta}$ by (see (4))

$$R_{\pi}(\vartheta,\hat{\vartheta}) = \pi(\hat{\vartheta} \in \text{LOSS}) = \sup_{t\in\Theta} \min\,(L(\vartheta,t),\, g_{\hat{\vartheta}/\vartheta}(t)) \;. \qquad (13)$$

We will use the possibilistic linear regression as an example to discuss the above questions.

2. Possibilistic linear regression

Let us consider pdf's $g_Y(y,\vartheta)$ of the following form:

$$g_Y(y,\vartheta) = s(y-f^T(z)\vartheta) \;;\; y\in R^1 \;;\; \vartheta\in R^r \;;\; z\in R^k \qquad (14)$$

where $f\,|\,R^k \to R^r$ is a known (setup-) function and s fulfils

Ass.2: The function s is continuous, symmetric about the origin and monotonically decreasing on R^+, with $s(0)\leq 1$ ∎

Note that (14) describes a linear regression model

$$Y = f^T(z)\vartheta + E \qquad (15)$$

where the "error" E is a pv which follows the pdf s.

Assume, for fixed design points $z_i \in V \subset R^k$ we observe y_i $(i=1,\dots,n)$ We want to estimate ϑ by a linear estimate

$$\hat{\vartheta} = Cy \quad ; \quad y = (y_1,\dots,y_n)^T \tag{16}$$

where C is an appropriate (r×n)-matrix. Our objective is to find the "BLUE". This requires a definition of unbiasedness for $\hat{\vartheta}$ which should be a formalization of the following demand: If all observations lie on a "straight line" $f^T(z)\vartheta$ then an unbiased $\hat{\vartheta}$ coincides with this straight line coefficient ϑ.

Def.1: The linear estimator $\hat{\vartheta}$ from (16) is called unbiased iff

$$\exists \vartheta \in R^r \ \forall i \in \{1,\dots,n\}: y_i = f^T(z_i)\vartheta \implies \hat{\vartheta} = \vartheta \ \blacksquare \tag{17}$$

It is easy to see that $\hat{\vartheta} = Cy$ is unbiased iff

$$CF = I_r \quad ; \quad F = (f(z_1),\dots,f(z_n))^T \ . \tag{18}$$

Note that (18) is the same unbiasedness condition as known from probabilistic linear regression.

The BLUE $\hat{\vartheta}^*$ minimizes, for a certain loss L, the possibilistic risk (13) w.r.t. all unbiased estimators (16).

Ass.3: The loss function $L(\vartheta,\hat{\vartheta})$ is a continuous and monotonically increasing function of $\|\vartheta - \hat{\vartheta}\|$, $\|\cdot\|$ Euklidean distance $\blacksquare$

Then the BLUE $\hat{\vartheta}^*$ is, roughly spoken, a linear unbiased estimate the pdf of which is closest to ϑ. For more detailed results let us reveal the structure of the pdf $g_{\hat{\vartheta}/\vartheta}$.

Let $e_i = y_i - f^T(z_i)\vartheta$ and $e = (e_1,\dots,e_n)^T$. Since for unbiased $\hat{\vartheta}$

$$g_{\hat{\vartheta}/\vartheta}(t) = \sup_{\vartheta + Ce = t} \min\{s(e_1),\dots,s(e_n)\} = g_{\hat{\vartheta}/\vartheta=0}(t-\vartheta) \tag{19}$$

it suffices to consider

$$g_{\hat{\vartheta}/\vartheta=0}(u) =: g_{\hat{\vartheta}}(u) = \sup_{Ce=u} \min\{s(e_1),\dots,s(e_n)\} \ , \tag{20}$$

i.e. without loss of generality we can transform the coordinates of ϑ so that $\vartheta = 0$ is the true parameter. Thus, with

$$g_E(e) = \min\{s(e_1),\dots,s(e_n)\} \tag{21}$$

for further investigations we use

$$g_{\hat{\vartheta}}(u) = \sup_{Ce=u} g_E(e) \quad . \tag{22}$$

Thus, the risk (13) writes, with Ass.3

$$R_{\pi}(\vartheta,\hat{\vartheta}) = \sup_{u\in R^+} \min(L(\|u\|), g_{\hat{\vartheta}}(u)) \quad . \tag{23}$$

Now consider the α-cuts of $g_{\hat{\vartheta}}$, i.e. $(g_{\hat{\vartheta}})_\alpha := \{u: g_{\hat{\vartheta}}(u) \geqslant \alpha\}$; $\alpha\in(0,1]$. At first we have (the proof is omitted):

Lemma 1: The α-cut of $g_{\hat{\vartheta}}$ is the image by C of the α-cut of g_E ∎

Due to the min-relatedness (Ass.1) and the unimodality of s (Ass.2) it holds:

Lemma 2: $(g_E)_\alpha$ is an n-dimensional cube with centre zero ∎

Note that the corners of $(g_E)_\alpha$ are given by the set

$$\{-q_\alpha, q_\alpha\}^n = q_\alpha \mathbb{1} \;; \quad \mathbb{1} = \{-1,1\}^n \quad ; \quad s(q_\alpha) = \alpha \quad . \tag{24}$$

Since $(g_E)_\alpha$ is a convex polyhedron from convex analysis it is known:

Lemma 3: $(g_{\hat{\vartheta}})_\alpha$ is a convex polyhedron, the corners u_{co} of which only arise as images by C of the corners from $(g_E)_\alpha$,
$u_{co} = q_\alpha Cx \;;\; x\in\mathbb{1}$ ∎

Coming back to the BLUE we now can say: Choose C so that, for every $\alpha\in(0,1]$, looked from the origin, the most distant corner of the image $(g_{\hat{\vartheta}})_\alpha$ is as near as possible. Somewhat more exactly it holds:

Theorem 1: If C^* fulfils (18) and is a solution of

$$\varphi(C) := \sup_{x\in\mathbb{1}} \|Cx\| \overset{!}{=} \min_{C:CF=I_r} \quad ; \quad \mathbb{1} = \{-1,1\}^n \tag{25}$$

then $\hat{\vartheta}^* = C^* y$ is BLUE ∎

Proof: $$R_{\pi}(\vartheta,\hat{\vartheta}) = \sup_{u\in R} \min(L(\|u\|), g_{\hat{\vartheta}}(u))$$
$$= \sup_{\alpha} \min(\alpha, \sup_{u\in(g_{\hat{\vartheta}})_\alpha} L(\|u\|)) = \alpha_0$$

where α_0 is solution of the "fix-point-equation" $\alpha_0 = \sup_{u\in(g_{\hat{\vartheta}})_{\alpha_0}} L(\|u\|)$.

Using the monotonicity of L and Lemma 3 this reduces to $\alpha_0 = L(q_{\alpha_0}\varphi(C)) = R(\vartheta,\hat{\vartheta})$. Thus, minimization of $\varphi(C)$ leads to minimization of the risk∎

Now let us present a further characterization of $\hat{\vartheta}^*$.

<u>Theorem 2</u>: Assume F is of full rank r. Then the BLUE $\hat{\vartheta}^* = C^*y$ uses only r observations out of the n observations y_i, $i=1,\ldots,n$.

<u>Proof</u>: a) Let $\hat{\vartheta}=Cy$ be unbiased, $C=(c_1,\ldots,c_n)$ with $c_i \in R^r$ and (after possible rearrangement) $C=(C_1 \vdots C_2)$ where C_1 is $r\times r$ and C_2 is $r\times(n-r)$. The unbiasedness condition $CF=I_r$ needs r degrees of freedom. Only the remaining n-r degrees we can use for optimization, i.e. the construction of the BLUE consists in determining a certain C_2.

b) The corners of the image, Cx with $x \in \mathcal{U}$, can be written, symbolically: $Cx = \sum_{i=1}^{n} \pm c_i$. According to (25) we are interested in the vector Cx with maximum length. Thus, we have to add the c_i, furnished with a certain pattern of signs, to reach a maximum length vector. At every step j of this addition procedure we have, with $c^j := \sum_{i=1}^{j} \pm c_i$, $j=1,\ldots,n-1$,

$$\|c^j\| \le \max\left(\|c^j+c_{j+1}\|, \|c^j-c_{j+1}\|\right) =: \|c^{j+1}\| \qquad (26)$$

i.e. $\{c^j\}_{j=1,\ldots,n}$ builds a sequence of vectors with nondecreasing length and with $\|c^n\| = \sup_{x\in\mathcal{U}} \|Cx\|$. In (26) the strict inequality holds iff $c_{j+1} \neq 0$. Thus, the length of c^n becomes as small as possible if n-r out of the c_i (i.e. a certain C_2) effect no increase of length , i.e. if n-r out of the c_i are zero. Hence, after rearrangement, $C^*=(C_1 \vdots 0)$, i.e. the BLUE-matrix C^* picks out only r of the n observations∎

<u>Remark 1</u>: If a linear functional $\eta = f^T\vartheta$ is to be estimated by a linear estimate $\hat{\eta} = \sum_{i=1}^{n} c_i y_i = c^T y$ the criterion (25) simplifies to $\sup_{x\in\mathcal{U}} |c^T x| = \sum_{i=1}^{n} |c_i| = \min$ and the side condition (18) changes to

$F^T c = f$. This special case is discussed in NÄTHER /2/ ∎

Remark 2: In the special case r=1 and f=1 the parameter ϑ in (14) reduces to a common location parameter. The unbiasedness condition (18) for a linear estimator $\hat{\vartheta} = \sum_{i=1}^{n} c_i y_i$ writes $\sum_{i=1}^{n} c_i = 1$. According to Theorem 2, only one observation y_i is needed for the BLUE of ϑ. Since the y_i are identically distributed, any single observation can be taken as the BLUE, i.e. $\hat{\vartheta}^* = y_i$. This example shows drastically the nonstatistical character of possibilistic inference: The estimate cannot be improved by increasing sample size. Results which take place one time are possible (but not necesserily probable) and a second observation in the same context gives, in the sense of a possibilistic evaluation, no further information. Notions as consistency of estimates, central limit theorems a.s.o. are without sense. These problems are discussed in some more detail in NÄTHER /2/ ∎

The optimization problem (25) can be solved, in principle, before experimentation. If we do so then, indeed, we have to realize only r observations. Thus, the so-called **saturated designs** play an important role in possibilistic regression. For more details on planning for possibilistic regression see NÄTHER /3/.

3. The extreme character of possibilistic inference

Let us discuss the question whether there are "intermediate" measures "between" probability and possibility which do not share the extreme nonstatistical character of possibility. We follow the WEBER-concept of decomposable measures (s. WEBER /5/).

Def.2: A set function $m | \mathfrak{B}_U \to [0,1]$ with $m(\emptyset)=0$, $m(U)=1$ and

$$m(A \cup B) = m(A) \perp m(B) \quad \text{or} \quad m(\bigcup_i A_i) = \perp_i m(A_i) \tag{27}$$

is called $\perp$- or σ-$\perp$-**decomposable**, where $\perp$ is a t-conorm i.e. a nondecreasing in each argument, commutative and associative binary operation with 0 as unit∎

Note that a possibility measure Π is $\perp$-decomposable with $\perp$=max. The crucial point is that $\perp$=max is idempotent: max(a,a)=a. This, together with min(a,a)=a, turns out to be essentially for the non-statistical character of possibilistic inference: Repeated measurements do not contribute to change of sample distribution.
Note that $\perp$=max is the only idempotent t-conorm, since with $b \leq a$

$$a = a \perp 0 \leq a \perp b \leq a \perp a = a \implies a = a \perp b = \max(a,b) \ . \qquad (28)$$

A continuous t-conorm is called <u>Archimedean</u> iff

$$\forall a \in (0,1): \quad a \perp a > a \qquad (29)$$

Thus, possibility is characterized by a non-Archimedean t-conorm and plays, accordind to (28), a singular role in the set of $\perp$-decomposable measures.
Archimedean $\perp$-decomposability is only a modification of additivity. This follows from LING's representation theorem (see WEBER /5/): There exists an increasing and continuous $g \mid [0,1] \to [0,\infty]$ with $g(0)=0$ such that

$$a \perp b = g^{(-1)}(g(a) + g(b)) \qquad (30)$$

where the pseudo-inverse $g^{(-1)}$ writes

$$g^{(-1)}(y) = \begin{cases} g^{-1}(y) & \text{if} \quad y \leq 1 \\ 1 & \text{if} \quad y > 1 \end{cases} \qquad (31)$$

The function g, the so-called <u>additive generator</u> of $\perp$, is unique up to a positive factor and is called <u>strict</u> if $g(1)=\infty$. A non-strict g with $g(1)=1$ is called <u>normed</u>.
Here we discuss the following special case: Let P be a probability measure on $(U,\mathcal{B}_U)$ and g a normed generator of a non-strict Archimedean $\perp$. Then, as a consequence of (30),

$$m := g^{-1} \circ P \qquad (32)$$

is σ-$\perp$-decomposable (A more general theorem is given in WEBER /5/). Moreover, for (membership-) functions $\mu \mid U \to [0,1]$ an <u>integral</u>

can be defined which via (30) reduces to Lebesgue-integrals (see WEBER /5/), which in case of (32) writes

$$\int_U \mu \perp m \ := \ g^{-1}\Big(\int_U \mu \, dP\Big) \ . \tag{33}$$

Note that

$$\int_U \mu_1 dP < \int_U \mu_2 dP \implies \int_U \mu_1 \perp m < \int_U \mu_2 \perp m \tag{34}$$

since with g also g^{-1} is increasing.

Consider now a $\perp$-decomposable measure space $(U, \mathcal{B}_U, m(\vartheta))$ where $m(\vartheta)$ is generated by $P(\vartheta)$ via (32) and depends on an unknown parameter $\vartheta \in \Theta$. Let $Y|U \longrightarrow R^n$ be a measurable (sample-) vector and $\hat{\vartheta}(Y)$ an estimator of ϑ. Then, the induced measure $m_{\hat{\vartheta}/\vartheta}$ is given by

$$m_{\hat{\vartheta}/\vartheta}(A) = m(\hat{\vartheta}^{-1}(A), \vartheta) = g^{-1} \ P(\hat{\vartheta}^{-1}(A), \vartheta) = g^{-1} \circ P_{\hat{\vartheta}/\vartheta}(A) \ . \tag{35}$$

Analogously to (11) we define the <u>risk</u> of $\hat{\vartheta}$ (w.r.t. loss L and evaluation with m) by

$$R_m(\vartheta, \hat{\vartheta}) = \int_{\Theta} L(\vartheta, \cdot) \perp m_{\hat{\vartheta}/\vartheta} \tag{36}$$

which with (33), (35) and (11) can be written as

$$R_m(\vartheta, \hat{\vartheta}) = g^{-1}\Big(\int_{\Theta} L(\vartheta, \cdot) dP_{\hat{\vartheta}/\vartheta}\Big) \ = \ g^{-1}(R_p(\vartheta, \hat{\vartheta})) \ . \tag{37}$$

Due to the monotonicity (34) we have the following result:

<u>Theorem 3</u>: Let $\hat{\vartheta}_1$ be better than $\hat{\vartheta}_2$ w.r.t. L and probabilistic evaluation. Then $\hat{\vartheta}_1$ is better than $\hat{\vartheta}_2$ w.r.t. L and evaluation with m where $m_{\hat{\vartheta}_i} = g^{-1} \circ P_{\hat{\vartheta}_i}$, $i=1,2$ ∎

Hence, if we accept the ordering by the risk (36), an estimator which is optimal for $(U, \mathcal{B}_U, P(\vartheta))$ is also optimal for $(U, \mathcal{B}_U, m(\vartheta))$ with $m = g^{-1} \circ P$. Note, as a special case, that an improvement of probabilistic evaluated estimators by increasing sample size leads to an improvement w.r.t. the associated m, too. From this point of view inference on Archimedean $\perp$-decomposable measures $m(\vartheta)$ preserves, qualitatively, the statistical character.

Quantitative statements on the more or less strong statistical character of inference on $m(\vartheta) = g^{-1} \circ P(\vartheta)$ are controlled by the degree of ascent of g^{-1}.

To have an **example** in mind let us mention SUGENO's λ-fuzzy-measures (see SUGENO /4/, WEBER /5/) which are $\perp$-decomposable with

$$a \perp b = \min (a+b+\lambda ab, 1) \quad ; \lambda > -1 . \tag{38}$$

Here $\perp$ is non-strict with the normed generator

$$g_\lambda(x) = \frac{\ln(1+\lambda x)}{\ln(1+x)} \quad ; \quad g_0(x) = \lim_{\lambda \to 0} g_\lambda(x) = x \tag{39}$$

and its pseudo-inverse

$$g_\lambda^{(-1)}(y) = \begin{cases} [(1+\lambda)^y - 1]/\lambda & \text{for } y \leq 1 \\ 1 & \text{for } y > 1 \end{cases} . \tag{40}$$

References

/1/ DUBOIS, D. and H. PRADE: Fuzzy Sets and Systems: Theory and Application, Academic Press New York 1980

/2/ NÄTHER, W.: On possibilistic inference, submitted to Fuzzy Sets and Systems

/3/ NÄTHER, W.: On possibilistic planning and inference, submitted to Journ. of Statist. Plann. and Inf.

/4/ SUGENO, M.: Theory of Fuzzy Integral and its Applications Ph.D.Thesis, Tokyo Inst. of Techn. 1974

/5/ WEBER, S.: Measures of fuzzy sets and measures of fuzziness, Fuzzy Sets and Systems 13 (1984) 247-271

/6/ ZADEH, L. A.: Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems 1 (1978) 3-28