Visual-based Detection of Properties of Confirmation Measures

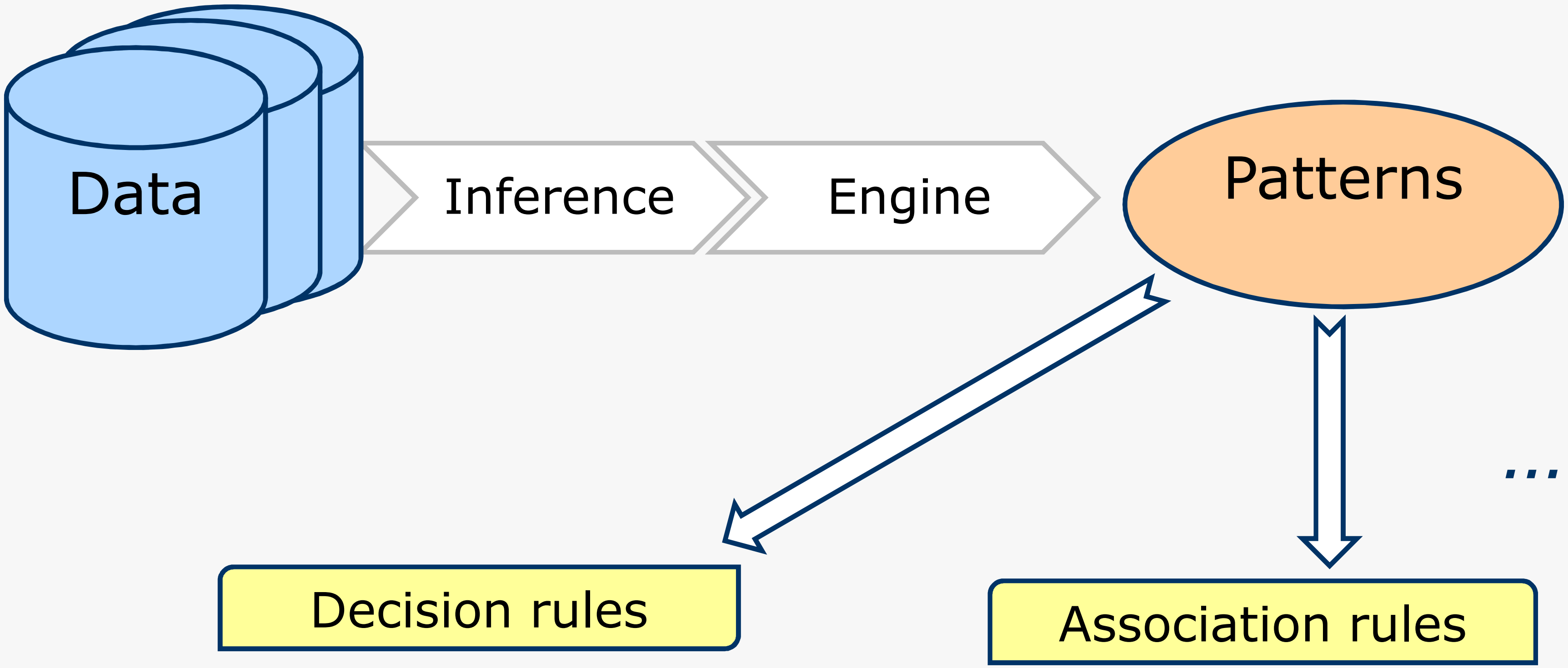
Izabela Szczęch
Robert Susmaga
Poznań University of Technology, Poland

ISMIS Roskilde, Denmark, 25-27.06.2014

Presentation plan

- Rule induction and interestingness measures
- Confirmation measures
- Properties of confirmation measures
    - Property of monotonicity M
    - Property of maximality/minimality
    - Property of hypothesis symmetry
- Visualization of measures
    - The experimental dataset
    - Visualization technique
    - Visual-based detection of properties
- Conclusions

Rule induction



If symptom s_1 is present
and symptoms s_2
and s_3 are absent
then disease d_1

If symptom s_1 is present
then symptoms s_2
and s_3 are absent

Rule induction

- Patterns in form of rules are induced from a data table
- $S=\langle U, A\rangle$ – *data table*, where U and A are finite, non-empty sets
  U – *universe of objects*; A – set of *attributes*
- $S=\langle U, C, D\rangle$ – *decision table*, where C – set of *condition attributes*,
  D – set of *decision attributes*, $C\cap D=\emptyset$

- *Rule* induced from S is a *consequence relation*:
  E → H *read as* ***if E then H***
  where

  E is condition (evidence or premise) and
  H is conclusion (hypothesis or decision)
  formula built from attribute-value pairs (q,v)

Rule induction

Characterization of nationalities

U	*Height*	*Hair*	*Eyes*	*Nationality*	Support
1	*tall*	*blond*	*blue*	*Swede*	270
2	*medium*	*dark*	*hazel*	*German*	90
3	*medium*	*blond*	*blue*	*Swede*	90
4	*tall*	*blond*	*blue*	*German*	360
5	*short*	*red*	*blue*	*German*	45
6	*medium*	*dark*	*hazel*	*Swede*	45

C (Height, Hair, Eyes) *D* (Nationality)

- E.g. decision rules induced from „characterization of nationalities”:
  1) If (*Height=tall*) then (*Nationality=Swede*)
  2) If (*Height=medium*) & (*Hair=dark*) then (*Nationality=German*)

If *Evidence* then *Hypothesis*

$$E \rightarrow H$$

Interestingness measures

The number of rules
induced from data sets is usually quite large

- overwhelming for human comprehension
- many rules are irrelevant or obvious (low practical value)

rule evaluation – interestingness (attractiveness) measures
(e.g. support, confidence, gain, rule interest, lift, measures of Bayesian confirmation)

- each measure was proposed to capture different characteristics of rules
- the number of proposed measures is very large

In this work we focus on a group of measures called
measures of confirmation

Notation

- Used notation corresponding to a 2x2 contingency table of rule's premise and conclusion. For a rule $\boldsymbol{E} \rightarrow \boldsymbol{H}$:

  $a=sup(H,E)$ is the number of objects in U satisfying both the premise E and the conclusion H of a rule $\boldsymbol{E} \rightarrow \boldsymbol{H}$,

  $b=sup(H, \neg E)$,

  $c=sup(\neg H, E)$,

  $d=sup(\neg H, \neg E)$,

  $a+c=sup(E)$,

  $a+b=sup(H)$,…

	H	$\neg H$	Σ
E	a	c	$a+c$
$\neg E$	b	d	$b+d$
Σ	$a+b$	$c+d$	$a+b+c+d=n$

$a, b, c, d \geq 0$

- a, b, c and d can also be regarded as frequencies that can be used to estimate probabilities:
  e.g., $P(E)=(a+c)/n$, $P(H)=(a+b)/n$, $P(H|E) = a/(a+c)$.

Notation

Height	**Hair**	**Eyes**	**Nationality**
tall	blond	blue	Swede
medium	dark	hazel	German
medium	blond	blue	Swede
tall	blond	blue	German
short	red	blue	German
medium	dark	hazel	Swede

$\neg E$	$\neg H$
$\neg E$	H
$\neg E$	$\neg H$
$\neg E$	H
E	H
$\neg E$	$\neg H$

	H	$\neg H$
E	1	0
$\neg E$	2	3

$a = sup(E,H)$
$b = sup(\neg E,H)$
$c = sup(E,\neg H)$
$d = sup(\neg E,\neg H)$

if (Hair = red) & (Eyes = blue) then (Nationality = German)

if Evidence then Hypothesis

- The contingency table is a form used to calculate the value of interestingness measures (e.g. confirmation measures)

Confirmation measures

- An interestingness measure *c*(*H*,*E*) has the property of confirmation (i.e. is a confirmation measure) if is satisfies the following condition:

$$c(H,E)\begin{cases} >0 & \text{if } P(H|E) > P(H) \\ =0 & \text{if } P(H|E) = P(H) \\ <0 & \text{if } P(H|E) < P(H) \end{cases} \Rightarrow c(H,E)\begin{cases} >0 & \text{if } \frac{a}{a+c} > \frac{a+b}{n} \\ =0 & \text{if } \frac{a}{a+c} = \frac{a+b}{n} \\ <0 & \text{if } \frac{a}{a+c} < \frac{a+b}{n} \end{cases}$$

- Measures of confirmation quantify the strength of confirmation that premise *E* gives to conclusion *H*
- „*H* is verified more often, when *E* is verified, rather than when *E* is not verified"

Selected confirmation measures

There are many alternative, non-equivalent measures of confirmation

$$\mathrm{M}(H,E)=\frac{a}{a+b}-\frac{a+c}{|U|}$$ (Mortimer 1988)

$$S(H,E)=\frac{a}{a+c}-\frac{b}{b+d}$$ (Christensen 1999)

$$F(H,E)=\frac{ad-bc}{ad+bc+2ac}$$ (Kemeny and Oppenheim 1952)

$$FS(H,E)=\frac{1}{2}(F(H,E)+S(H,E))$$ (Glass 2013)

$$\phi(H,E)=\frac{ad-bc}{\sqrt{(a+c)(a+b)(b+d)(c+d)}}$$ (Mortimer 1988)

$$F\phi(H,E)=\frac{1}{2}(F(H,E)+\phi(H,E))$$ (Glass 2013)

- The values of all of the above measures range from −1 to +1,
- otherwise they are undefined, e.g. when $a+b=0$ measure $M(H,E)$ is *NaN*.

Properties of confirmation measures

The choice of a confirmation measure for a certain application is a difficult problem

- the number of proposed measures is overwhelming
- there is no evidence which measure is the best
- the users’ expectations vary

properties of confirmation measures, which reflect users’ expectations towards the behaviour of measures in particular situations

- property of monotonicity M (Greco, Pawlak & Słowiński 2004)
- Ex_1 property and its generalization to weak Ex_1
- property of logicality L and its generalization to weak L (Fitelson 2006; Crupi, Tentori & Gonzalez 2007 Greco, Słowiński & Szczęch 2012)
- …

need to analyze measures with respect to their properties

Motivation for this work: **Detect properties of measures and compare measures easily through their visualizations**

Property of monotonicity M

	H	$\neg H$
E	a	c
$\neg E$	b	d

- Desirable property of $c(H,E) = f(a,b,c,d)$: monotonicity (M)*

  > f should be non-decreasing with respect to a and d and non-increasing with respect to b and c

- Interpretation of (M): ($E \rightarrow H \equiv$ *if x is a raven, then x is black*)

  a) the more black ravens we observe, the more credible becomes $E \rightarrow H$

  b) the more black non-ravens we observe, the less credible becomes $E \rightarrow H$

  c) the more non-black ravens we observe, the less credible becomes $E \rightarrow H$

  d) the more non-black non-ravens we observe, the more credible becomes $E \rightarrow H$

*S.Greco, Z.Pawlak, R.Słowiński: Can Bayesian confirmation measures be useful for rough set decision rules? *Engineering Applications of Artificial Intelligence*, 17 (2004) no.4, 345-361

Property of maximality/minimality

	H	$\neg H$
E	a	c
$\neg E$	b	d

- Desirable property of $c(H,E)$: maximality/minimality*

$c(H,E)$ is maximal if and only if $P(E,\neg H) = P(\neg E,H) = 0$ and

$c(H,E)$ is minimal if and only if $P(E,H) = P(\neg E, \neg H) = 0$.

- Interpretation of maximality/minimality:

a measure obtains its maximum iff $c=b=0$ and its minimum iff $a=d=0$.

*Glass, D.H.: Confirmation measures of association rule interestingness, Knowledge-Based Systems 44, (2013) 65–77

Property of hypothesis symmetry HS

- Desirable property of $c(H,E)$: hypothesis symmetry (HS)*

$$\boxed{c(H,E) = -c(\neg H,E)}$$

- Interpretation of (HS): ($E \rightarrow H \equiv$ *if x is a square, then x is rectangle*)

  the strength with which

  the premise (*x is a square*) confirms the conclusion (*x is rectangle*)

  is the same as the strength with which

  the premise disconfirms the negated conclsuion (*x is not a rectangle*).

*Carnap, R.: Logical Foundations of Probability, second ed. University of Chicago Press, Chicago (1962)
Eells, E., Fitelson, B.: Symmetries and asymmetries in evidential support. Philosophical Studies, 107 (2) (2002), 129-142

Visualization of measures

The experimental dataset

- Given $n > 0$ (the total number of observations), a synthetic data set is generated as the set of all possible contingency tables satisfying $a + b + c + d = n$
- The set is thus exhaustive and non-redundant (i.e. it contains exactly one copy of each contingency table satisfying the above condition)

Height	***Hair***	***Eyes***	***Nationality***
tall	*blond*	*blue*	*Swede*
medium	*dark*	*hazel*	*German*
medium	*blond*	*blue*	*Swede*
tall	*blond*	*blue*	*German*
short	*red*	*blue*	*German*
medium	*dark*	*hazel*	*Swede*

	H	$\neg H$
E	a	c
$\neg E$	b	d

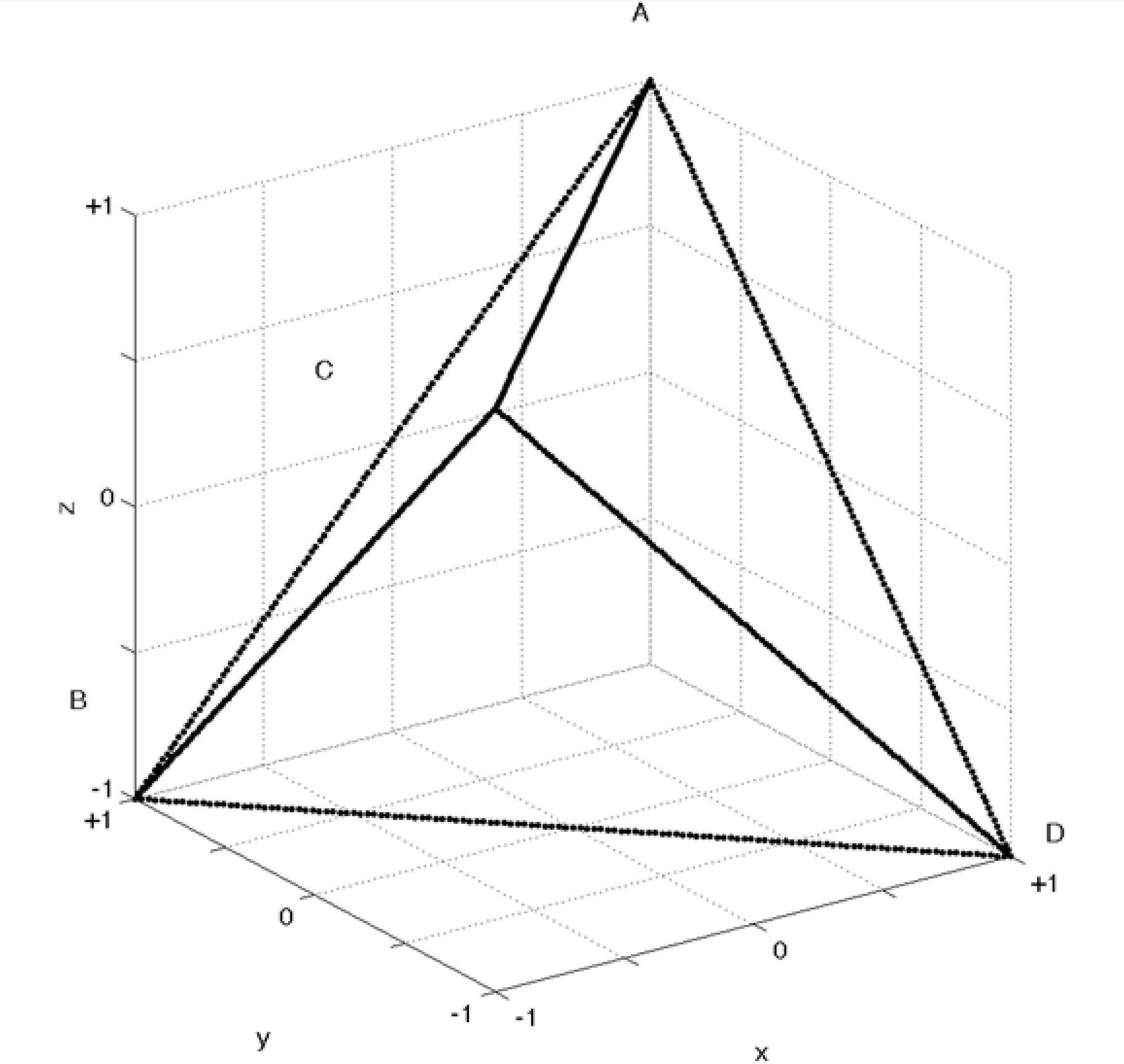
a	***b***	***c***	***d***
0	0	0	6
0	0	1	5
0	0	2	4
0	0	3	3
0	0	4	2
0	0	5	1
0	0	6	0
0	0	5	1
0	1	0	5
0	1	1	4
0	1	2	3
…	…	…	…
6	0	0	0

Visualization technique - barycentric coordinates

- Our synthetic data set comprises t rows and 4 columns: a, b, c and d; $t=(n+1)(n+2)(n+3)/6$
- In general, four independent columns correspond to four degrees of freedom, visualization of such data in the form of a scatter-plot would formally require four dimensions.

- Owing to the condition $a + b + c + d = n$ however, the number of degrees of freedom is reduced to three, so it is possible to visualize such data in three dimensions (3D) using tetrahedron-based barycentric coordinates
- The 3D view of the tetrahedron, proposed in the paper, has its four vertices A, B, C and D coinciding with points of the following $[x, y, z]$ coordinates: A: [1,1,1] C: [-1,-1,1] B: [-1,1,-1] D: [1,-1,-1]

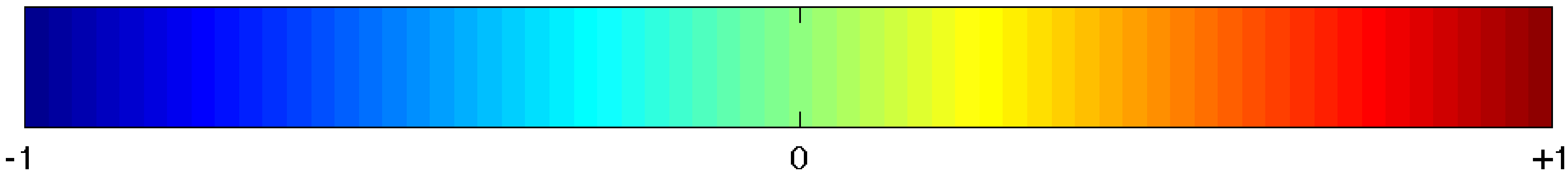
Visualization technique - barycentric coordinates

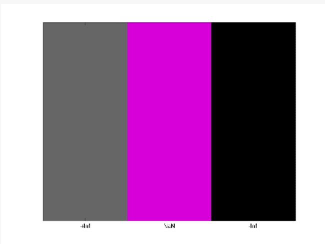


- the vertex *A* corresponds to the (single) contingency table satisfying $a=n$ and $b=c=d=0$,
- the edge *AB* corresponds to the (multiple) contingency tables satisfying $a+b=n$ and $c=d=0$,
- the face *ABC* corresponds to the (multiple) contingency tables satisfying $a+b+c=n$ and $d=0$, etc.

Visualization technique – colour map

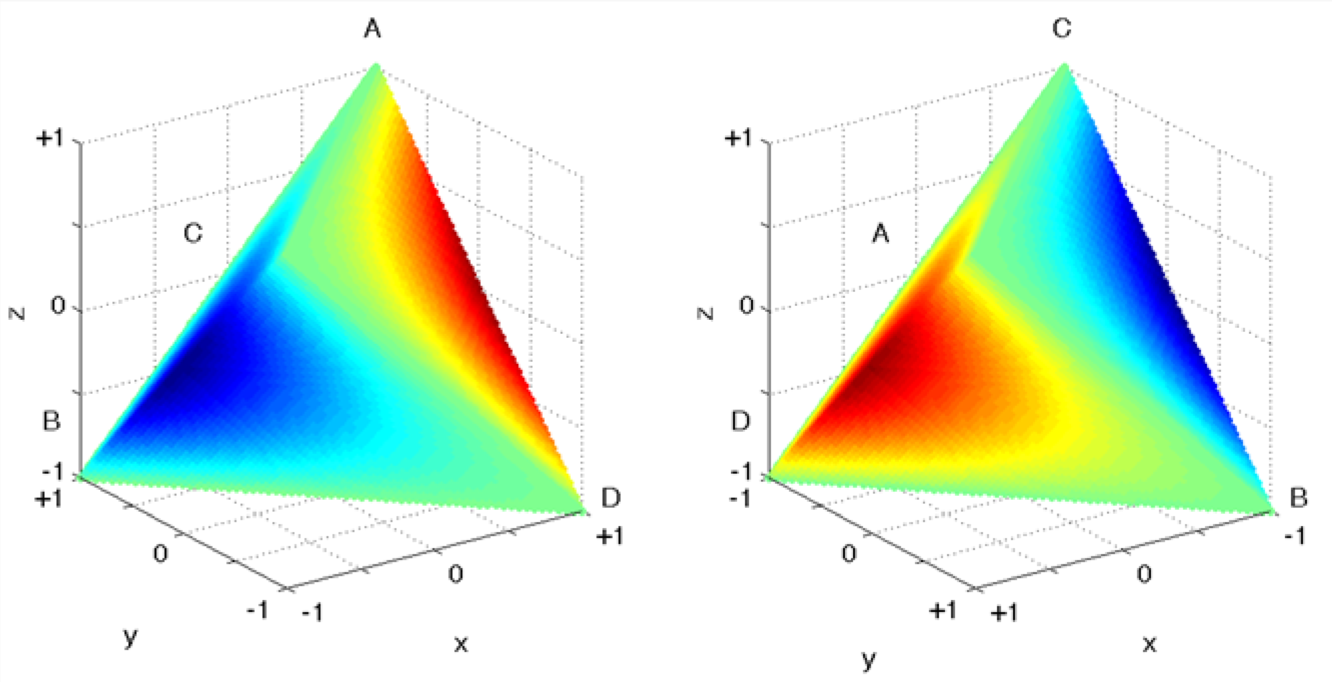
- Because the individual points of the tetrahedron may be displayed in colour, it is possible to visualize a function *f*(*a*,*b*,*c*,*d*) of the four arguments (e.g. **any interestingness measure**)
- It is assumed that the value set of this function is a real interval [*r*,*s*], with $r < s$, so that its values may be rendered using a pre-defined colour map
- For all the analysed confirmation measures the standard colour map ranges from −1 to +1:



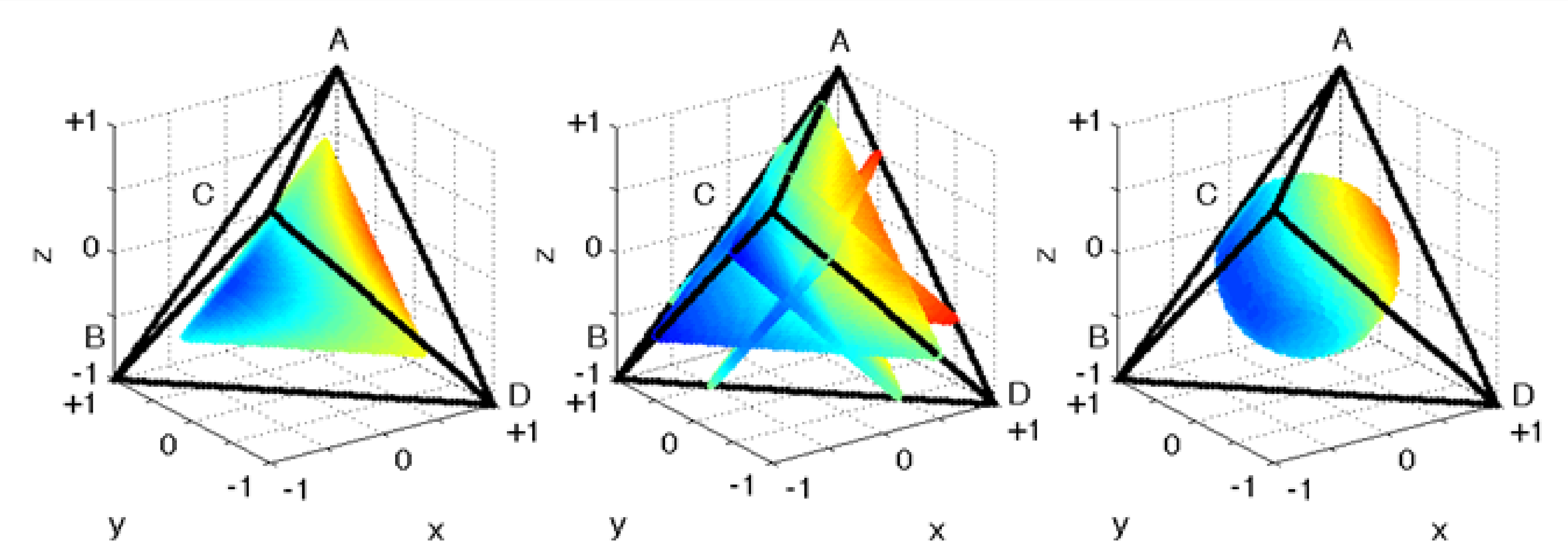


- Non-numeric values, i.e. +∞, *NaN* and - ∞, if generated by a particular function, may be rendered as colours not occurring in the map.

Visualization technique – exemplary external visualizations



Visualization technique – exemplary internal visualizations



Visual-based detection of properties

Property of monotonicity M

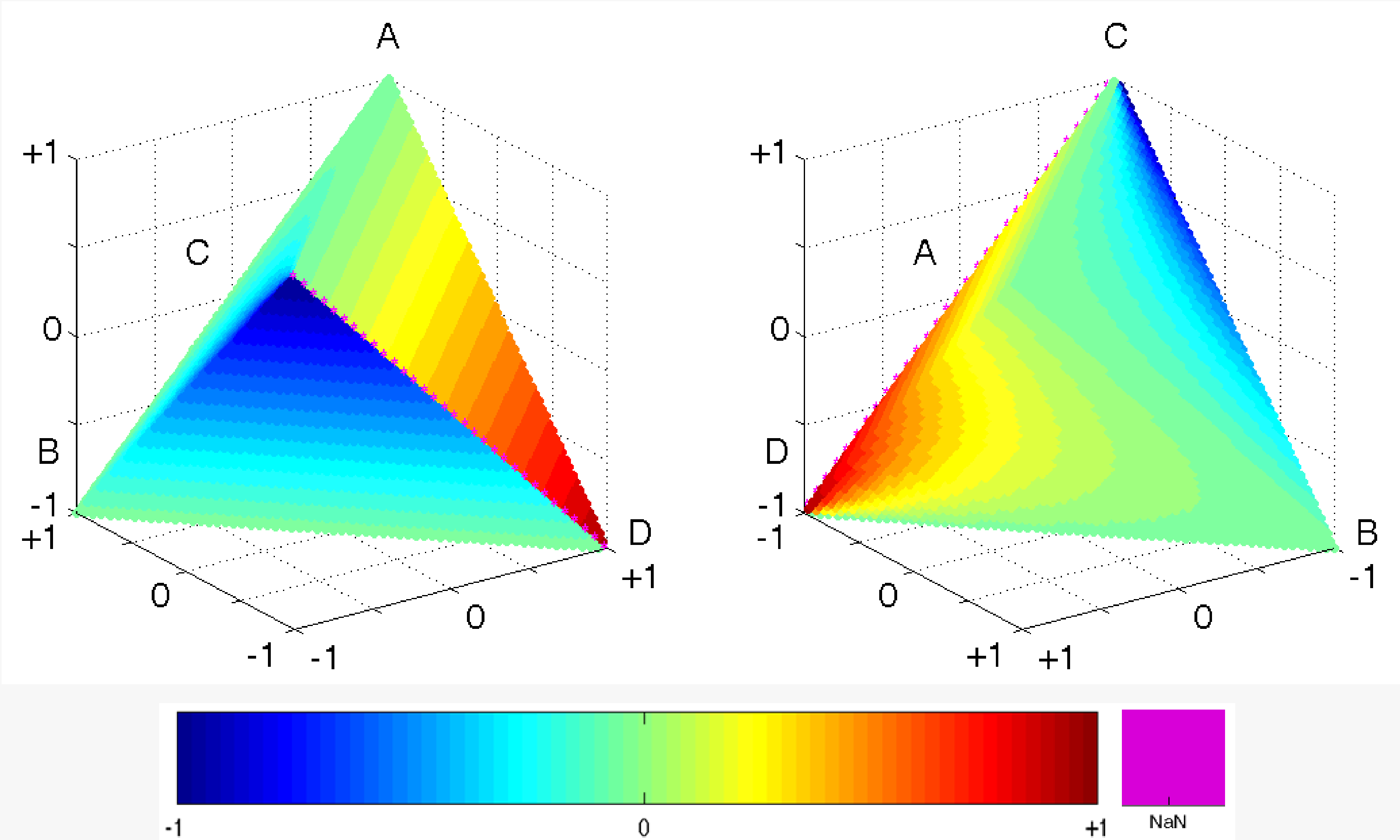
	H	$\neg H$
E	a	c
$\neg E$	b	d

- Desirable property of $c(H,E) = f(a,b,c,d)$: monotonicity (M)

  > f should be non-decreasing with respect to a and d and non-increasing with respect to b and c

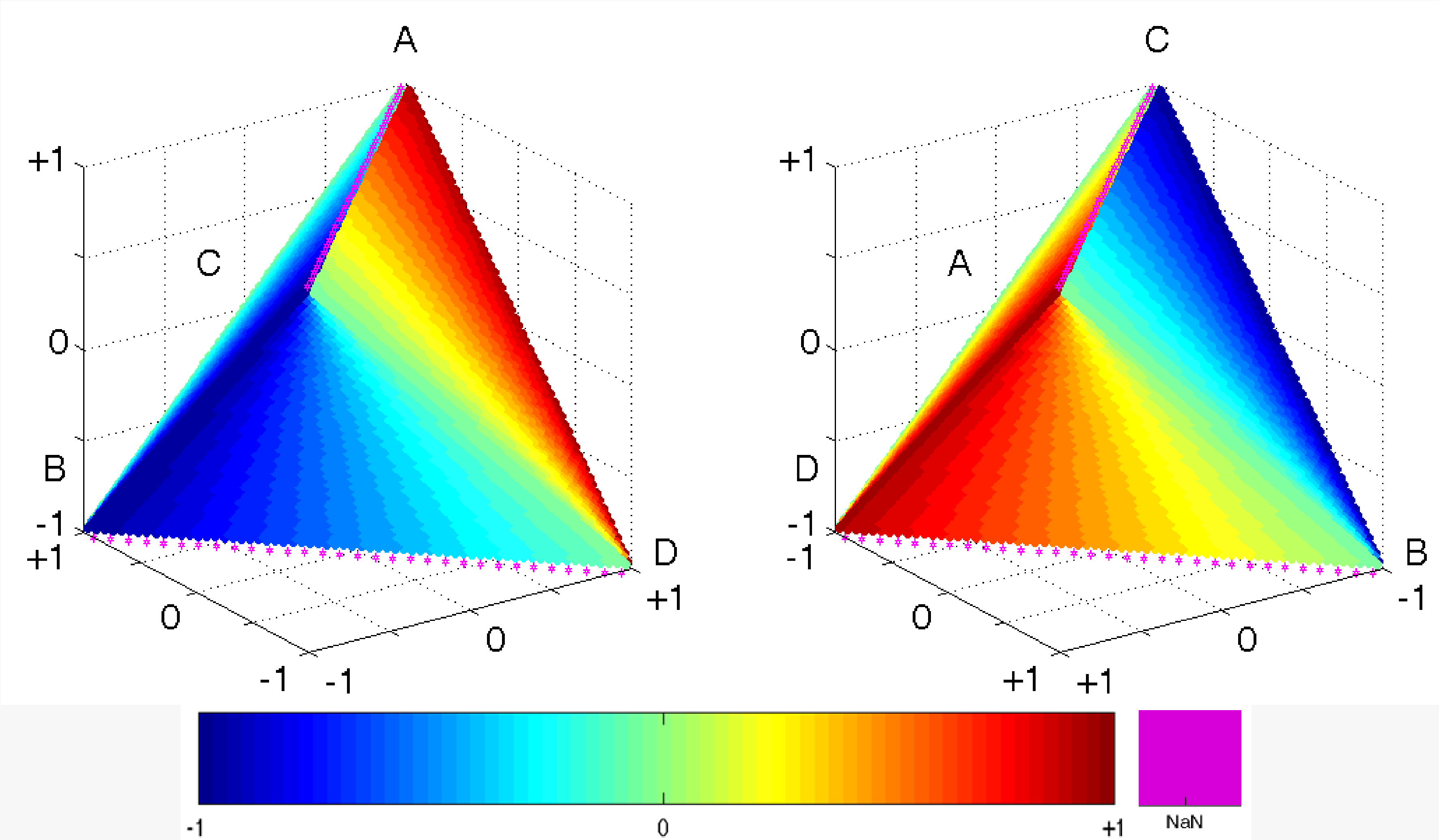
- Visual-based detection:
  - the „non-decreasing with *a* and *d*" condition should be reflected in the visualization as colours changing towards dark brown (increase of confirmation) around vertices *A* and *D* and
  - the „non-increasing with *b* and *c*" condition should be reflected in the visualization as colours changing towards dark blue (increase of disconfirmation) around vertices *B* and *C*
  - a thorough analysis with respect to property M requires an insight into the tetrahedron as potential counterexamples to this property may be located inside the shape

Measure *M*(*H*,*E*) – counterexamples to property M



Clearly, measure *M*(*H*,*E*) does not satisfy property M, as in the visualization the colour changes from dark brown at vertex *D* to pale green at vertex *A*, violating the demands the of the non-decrease with *a*.

Measure $S(H,E)$ – no observable counterexamples to M



There are no observable counterexamples to property M in the external visualizations of measure $S(H,E)$ which, together with additional analysis of the shape's inside, determines the possession of the property M by $S(H,E)$.

Property of maximality/minimality

	H	$\neg H$
E	a	c
$\neg E$	b	d

- Desirable property of $c(H,E)$: maximality/minimality*

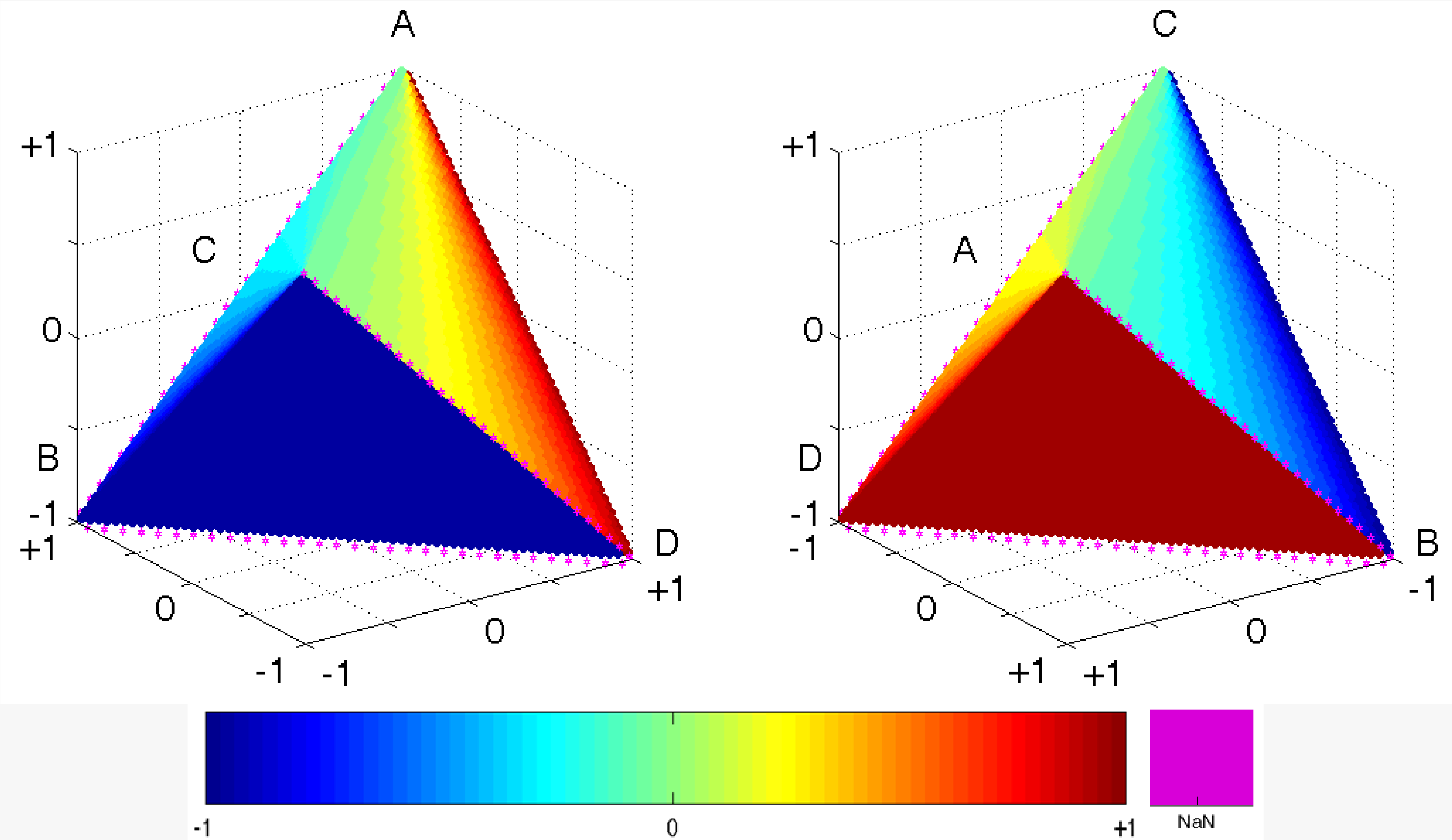
  > $c(H,E)$ is maximal if and only if $b=c=0$ and
  >
  > $c(H,E)$ is minimal if and only if $a=d=0$.

- Visual-based detection:
  - the dark brown (dark blue) colour must be found on the *AD* (*BC*) edge of the tetrahedron and cannot be found anywhere else

  Let us observe that the *AD* (*BC*) edge contains all points for which $b=c=0$ ($a=d=0$), i.e., the points most distant from the vertices *B* and *C* (*A* and *D*)

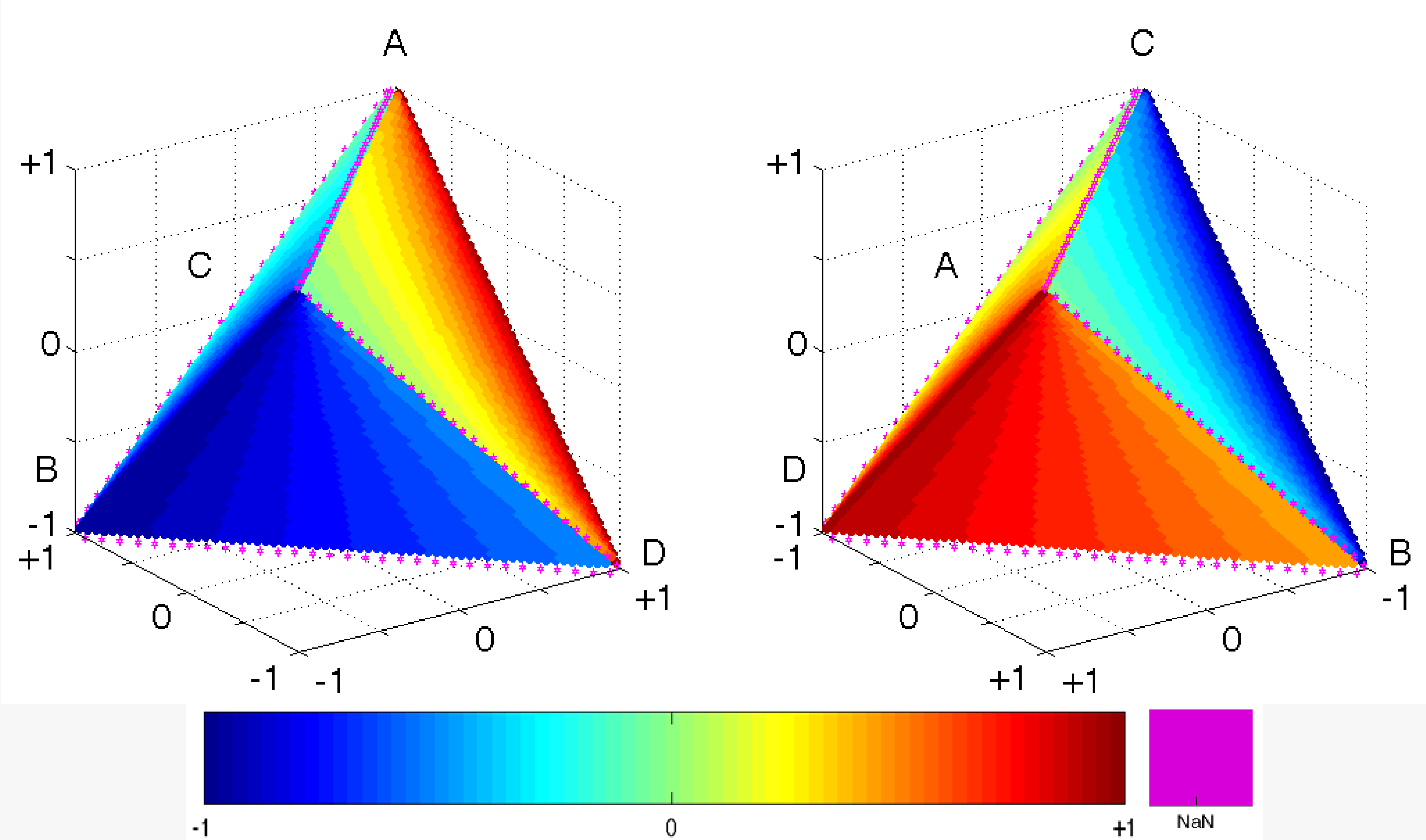
  - a thorough analysis with respect to maximality/minimality requires an insight into the tetrahedron as potential counterexamples to this property may be located inside the shape

Measure *F*(*H*,*E*) – counterexamples to property max/min



Visual-based detection of maximality/minimality property reveals that measure *F*(*H*,*E*) does not satisfy this property. It is, among others, due to the fact that the points with maximal values of *F*(*H*,*E*) cover the whole *ABD* face (i.e., there are too many of them).

Measure *FS*(*H*,*E*) – no observable counterexamples to max/min



There are no observable counterexamples to property max/min in the external visualizations of measure *FS*(*H*,*E*) which, together with additional analysis of the shape's inside, determines the possession of the maximality/minimality property by *FS*(*H*,*E*).

Property of hypothesis symmetry HS

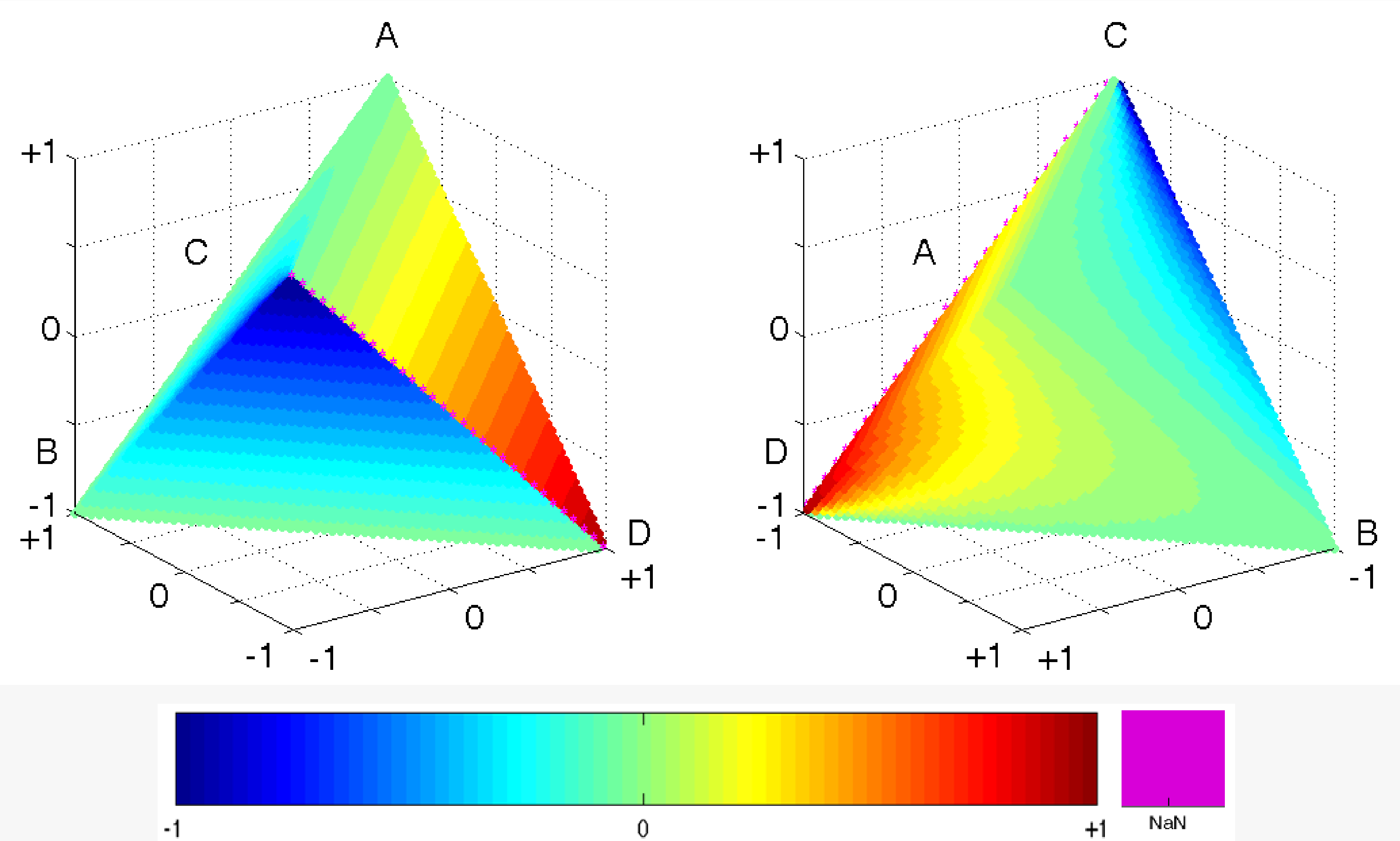
$$E \rightarrow H \qquad\qquad E \rightarrow \neg H$$

- Desirable property of $c(H,E)$: hypothesis symmetry (HS)

$$\boxed{c(H,E) = -c(\neg H,E)}$$

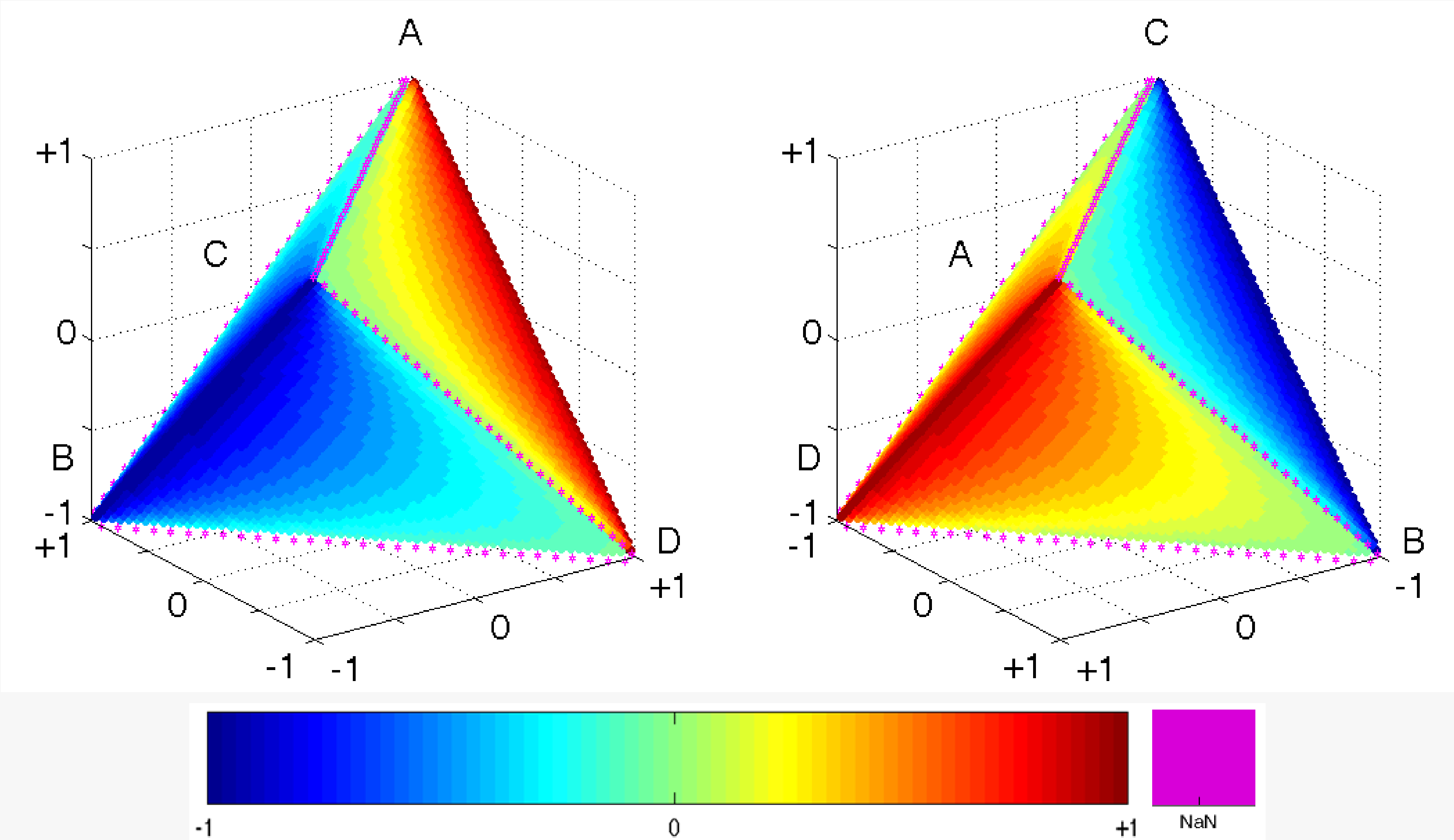
- Visual-based detection:
  - $c(H,E)=f(a,b,c,d) = -c(\neg H,E)= -f(a', b', c', d') = -f(c,d,a,b)$, reflecting the exchange of columns in the contingency tables ($a=c'$, $b=d'$, $c=a'$ $d=b'$)
  - two views must have the same gradient profile (i.e., the left view must be just like the right one, provided the colour map is reversed)
  - if the „recoloured" views are not the same, then the visualized measure does not possess the hypothesis symmetry
  - a thorough analysis with respect to HS requires an insight into the tetrahedron as potential counterexamples to this property may be located inside the shape

Measure *M*(*H*,*E*) – counterexamples to property HS



Clearly, measure *M*(*H*,*E*) does not satisfy property *HS* since e.g., the *BCD* face has a gradient profile that is characterized by straight lines, while the *DAB* face has a profile that is characterized by curved lines.

Measure $F\Phi(H,E)$ – no observable counterexamples to max/min



There are no observable counterexamples to property HS in the external visualizations of measure $F\Phi(H,E)$ which, together with additional analysis of the shape's inside, determines the possession of the property by $F\Phi(H,E)$.

Conclusions

- Determination of properties possessed by measures is an active research area

- The proposed visualization allows us to promptly detect distinct properties of the measures and compare them, increasing the general comprehension of the measures and helping the users choose one for their particular application

Conclusions

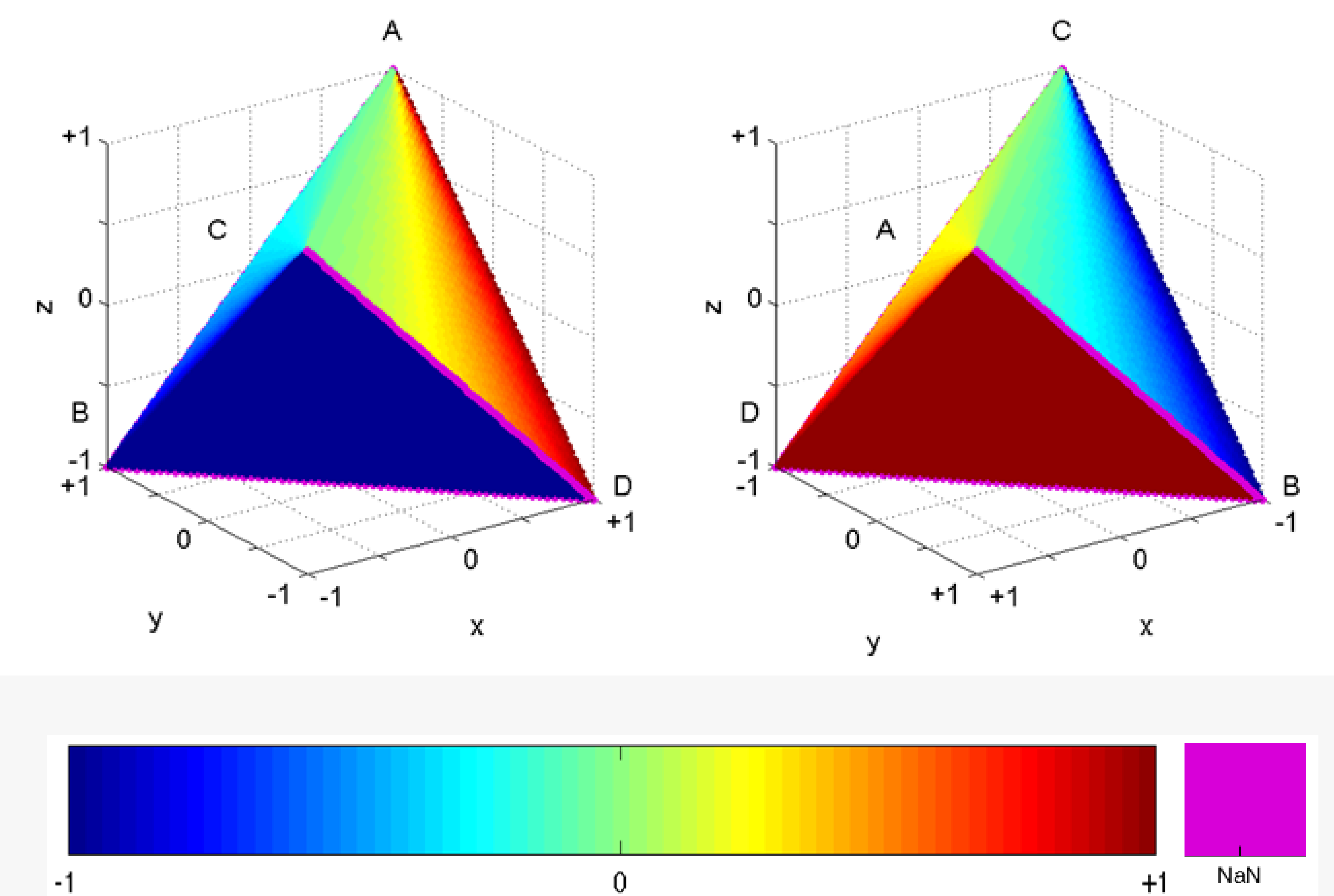
- Our proposition starts with constructing a synthetic, exhaustive and non-redundant set of contingency tables, which are commonly used to calculate the values of measures

- Using such a dataset, a 3-dimensional tetrahedron is built
- The position of points in the shape translates to corresponding contingency tables and the colour of the points represents values of the visualized measure

- Such visual-based approach is advantageous, especially when time constraints impede conducting in-depth, theoretical analyses of large numbers of such measures (e.g., generated automatically)
- Clearly, the analyses can be generalized to a wider range of measures or properties

Thank you!

Regular views of confirmation measures: $F(H,E)$

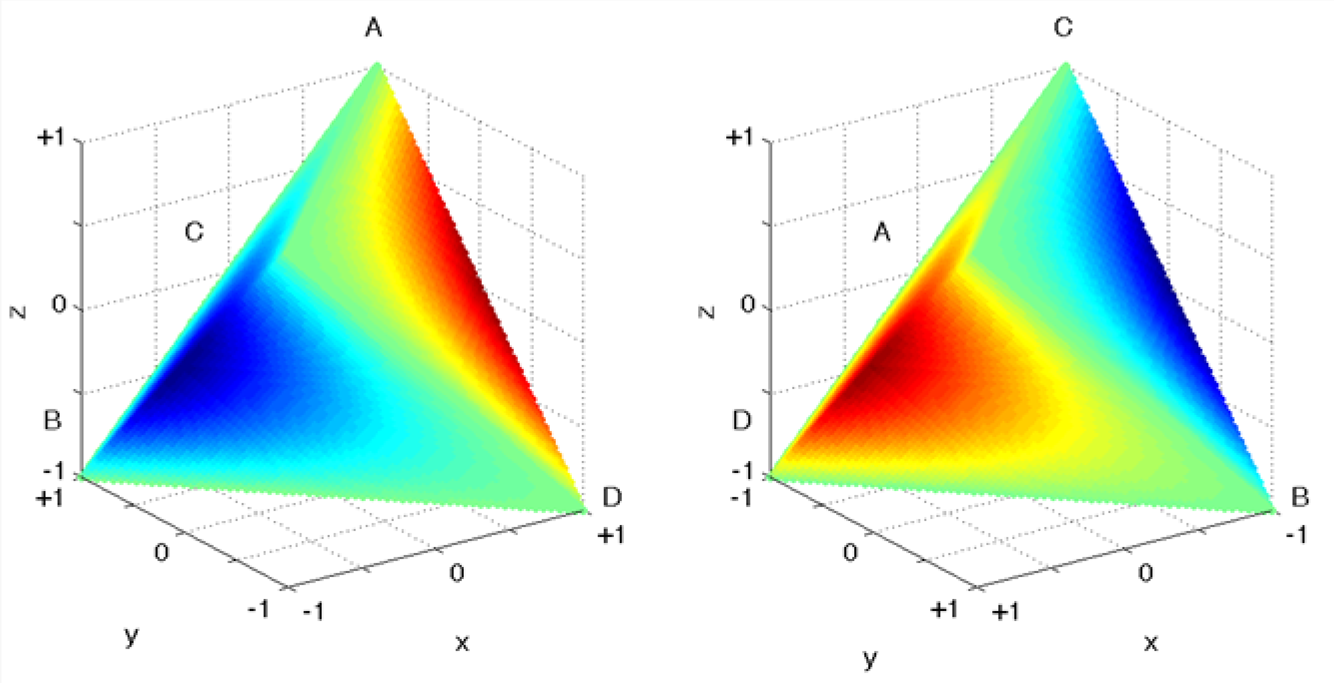


Visualization technique – exemplary external visualizations

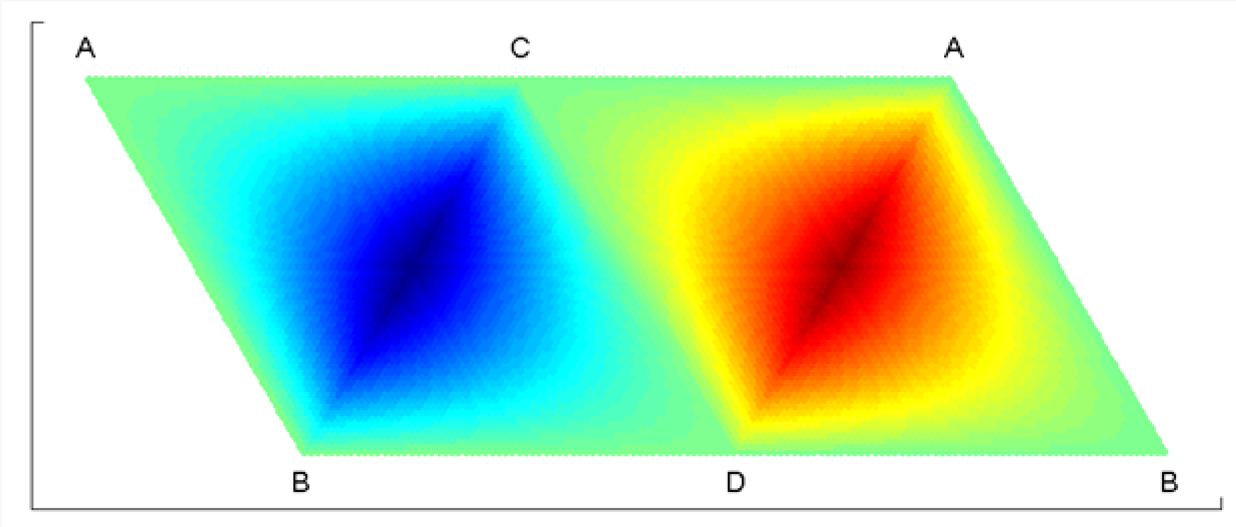
- The visualization of a 'solid' tetrahedron shows only extreme values of the arguments of the visualized function (external view):
  - 2D view
  - parallelogram – visualization of the net of the tetrahedron, i.e. a set of planar triangles, which when folded along selected edges, become the faces

Visualization technique – external visualizations for $C(H,E)$

2D view



parallelogram

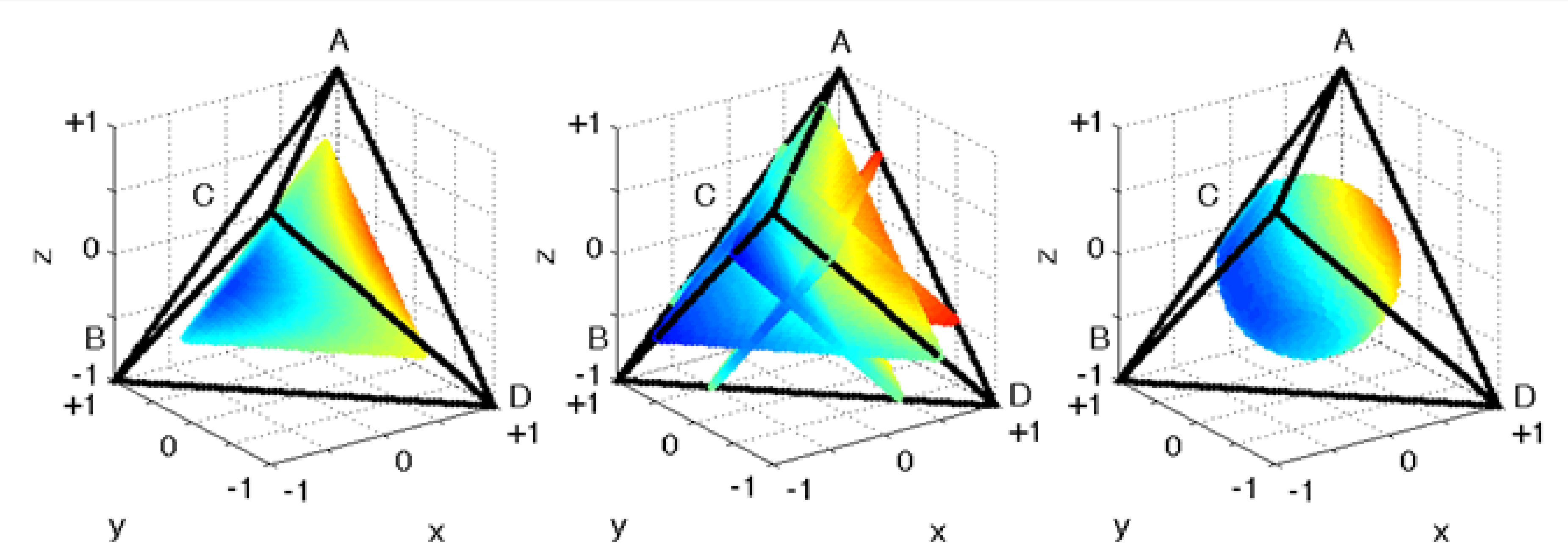


Visualization technique – exemplary internal visualizations

- If areas located strictly inside the tetrahedron have to be additionally visualized, various internal views can be generated:
  - 2D view
  - parallelogram – visualization of the net of the tetrahedron, i.e. a set of planar triangles, which when folded along selected edges, become the faces

Visualization technique – internal visualizations for *C*(*H*,*E*)

2D view



Visualization technique – internal visualizations for *C*(*H*,*E*)

parallelograms