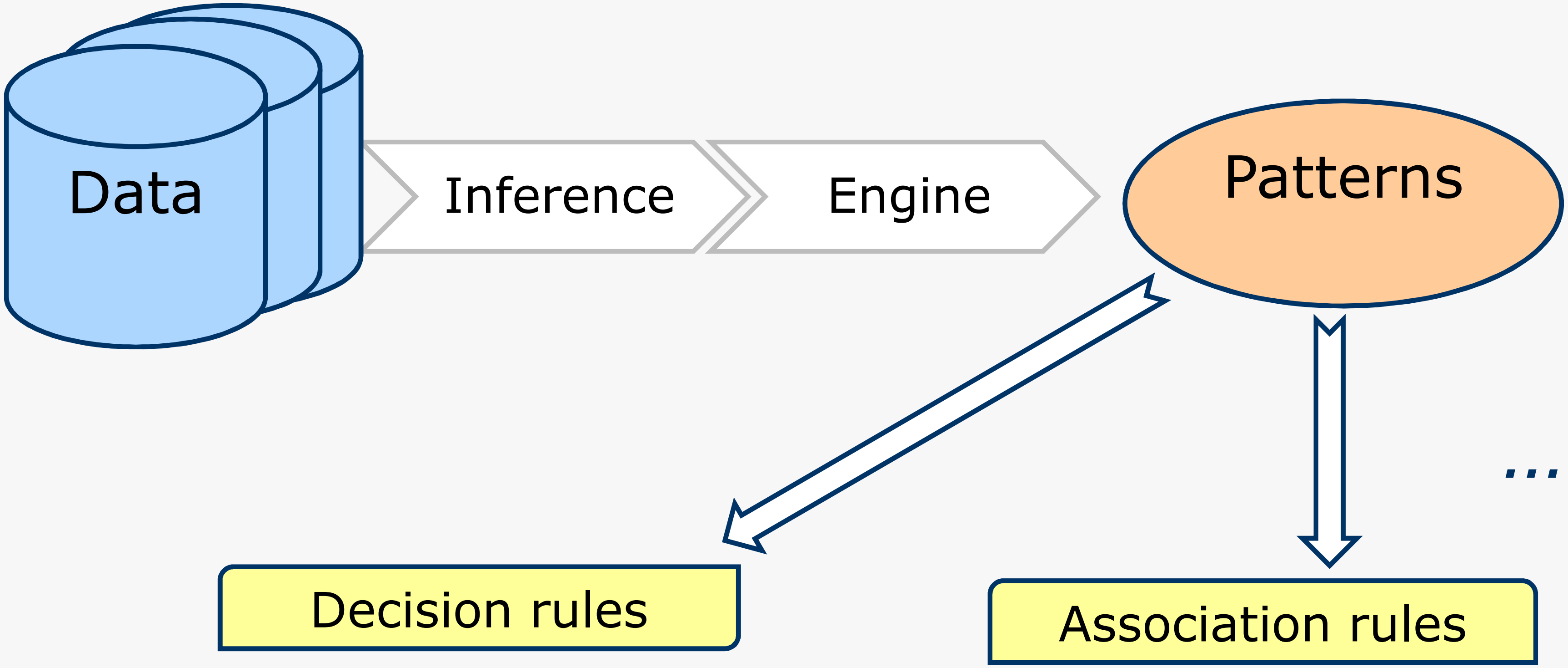
Visualization of Interestingness Measures

Izabela Szczęch
Robert Susmaga
Poznań University of Technology, Poland

Presentation plan

- Rule induction and interestingness measures
- The property of confirmation and popular confirmation measures
- Visualization of measures
    - The experimental dataset
    - Visualization techniques
    - Application of the visualization techniques
- 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 *D*

- 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*

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)

The property of confirmation

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

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

- 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"

Popular confirmation measures

There are many alternative, non-equivalent measures of confirmation

$$D(H,E)=\frac{a}{a+c}-\frac{a+b}{|U|}$$ (Carnap 1950/1962)

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

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

$$N(H,E)=\frac{a}{a+b}-\frac{c}{c+d}$$ (Nozick 1981)

$$\mathrm{C}(H,E)=4\left[\frac{a}{|U|}-\frac{(\mathrm{a}+\mathrm{c})(a+b)}{|U|^2}\right]$$ (Carnap 1950/1962)

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

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

Properties of confirmation measures

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

- there is no evidence which measure(s) is the best
- the users' expectations vary
- the number of proposed measures is overwhelming

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: **Discover properties of measures and compare measures easily through their visualizations**

The experimental data set

- 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

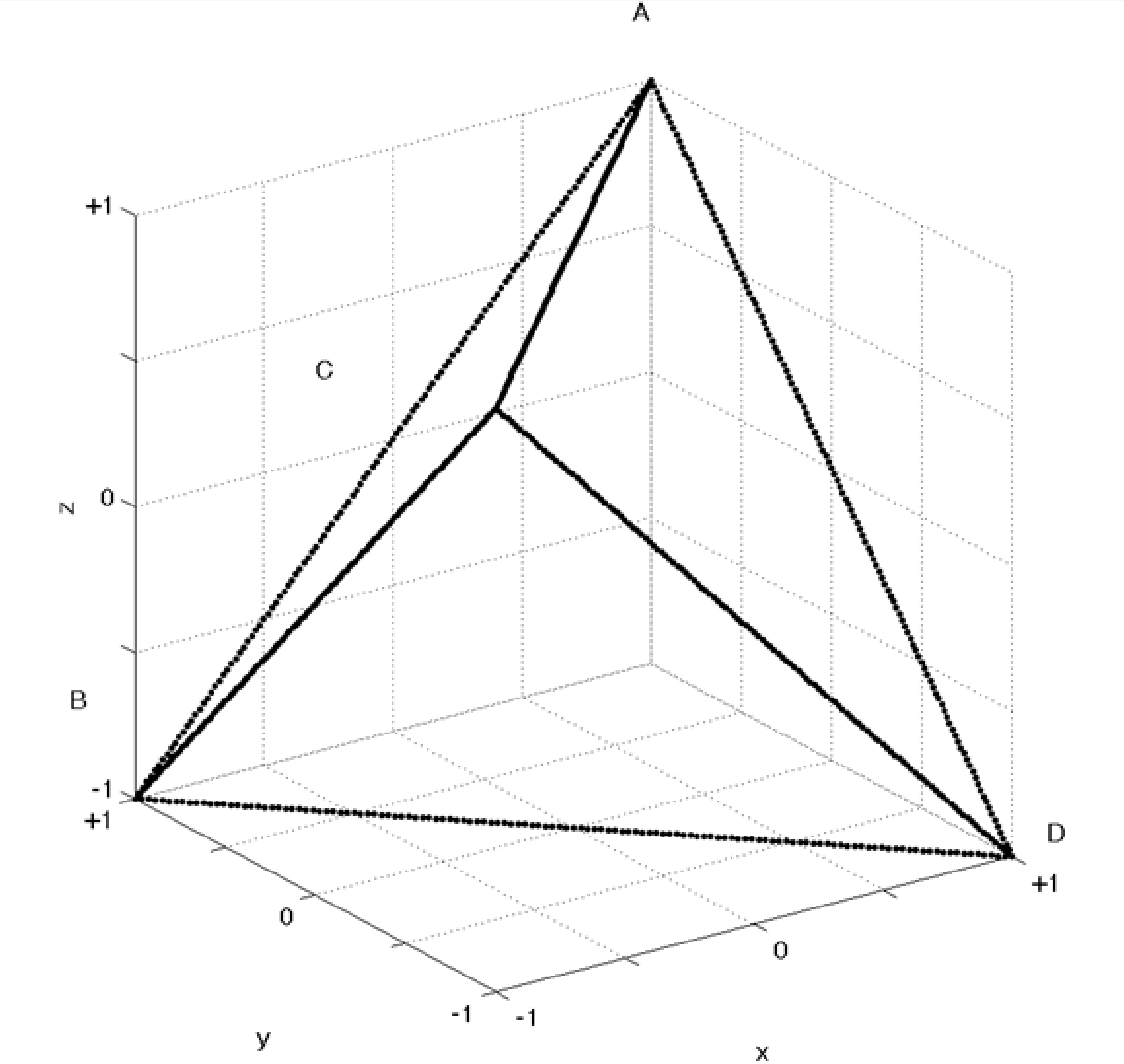
Visualization techniques - barycentric coordinates

- 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]

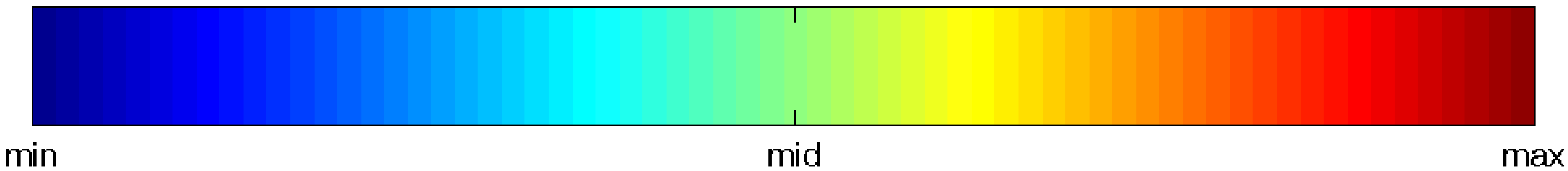
Visualization techniques - barycentric coordinates

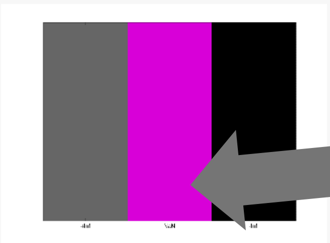


- the vertex *A* corresponds to the (single) contingency table satisfying $a=n$ and $b=c=d=0$,
- the edge *AB* corresponds to the (multiple) contingency tables satisfying $a+b=n$ and $c=d=0$,
- the face *ABC* corresponds to the (multiple) contingency tables satisfying $a+b+c=n$ and $d=0$, etc.

Visualization techniques – 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, further referred to as the operational function (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
- The standard colour map:

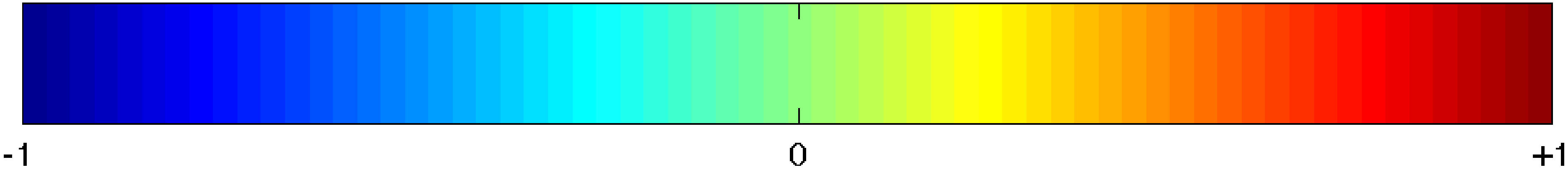




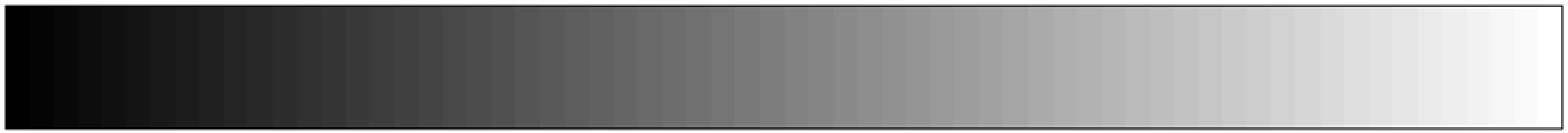
- Non-numeric values, i.e. [illegible] *NaN* and - ∞, if [illegible] by a particular function, may be rendered as colours not occurring in the map.

Visualization techniques – colour map

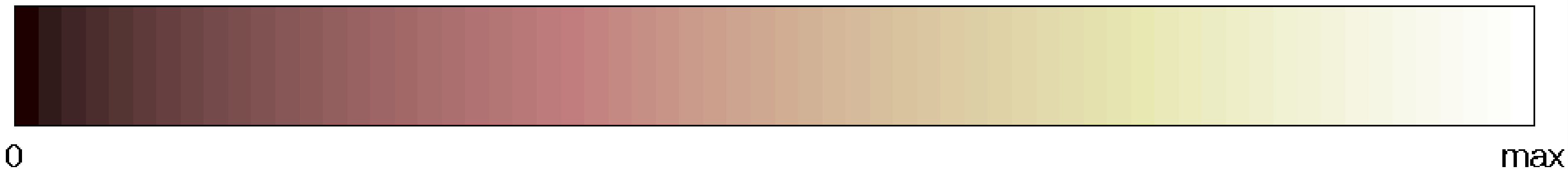
- For all the analysed confirmation measures the standard colour map ranges from -1 to +1



- The grey colour map is used only to provide the necessary perspective; the colours do not translate to values of the measure



- The pink colour map is used for presenting functions ranging from 0 to some positive value (e.g. variances of groups of measures)

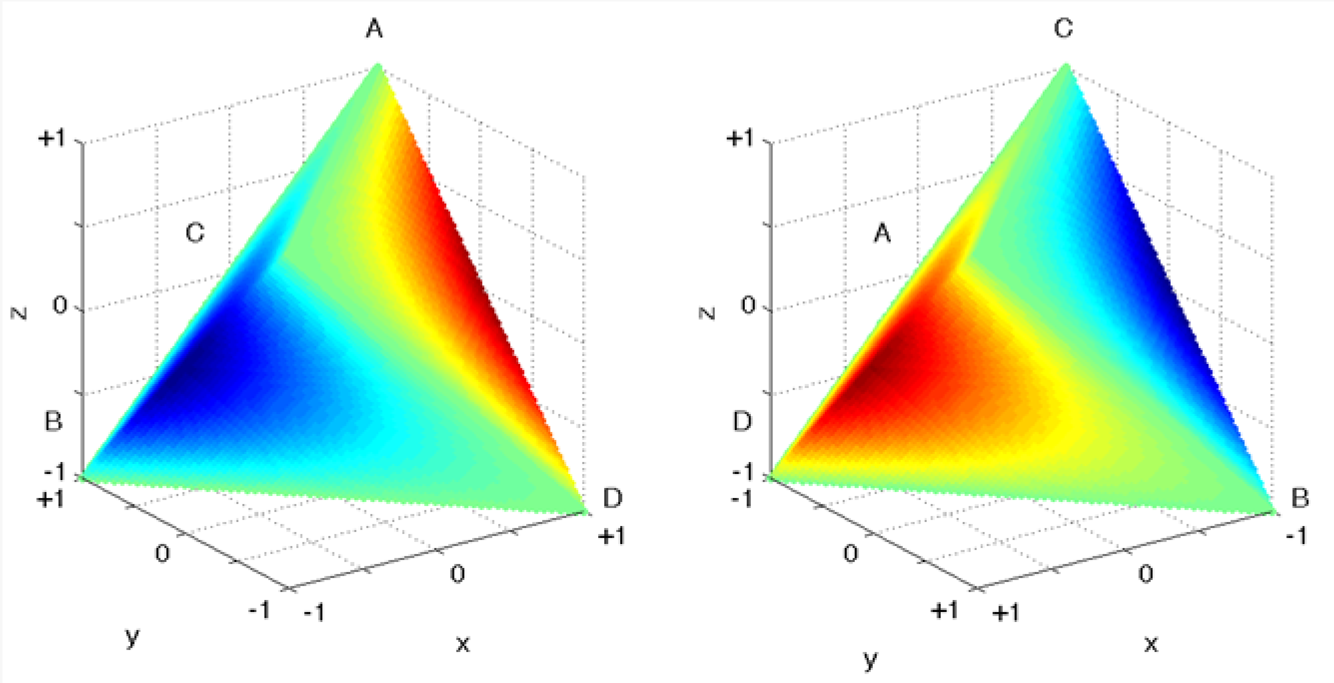


Visualization techniques – exemplary external visualizations

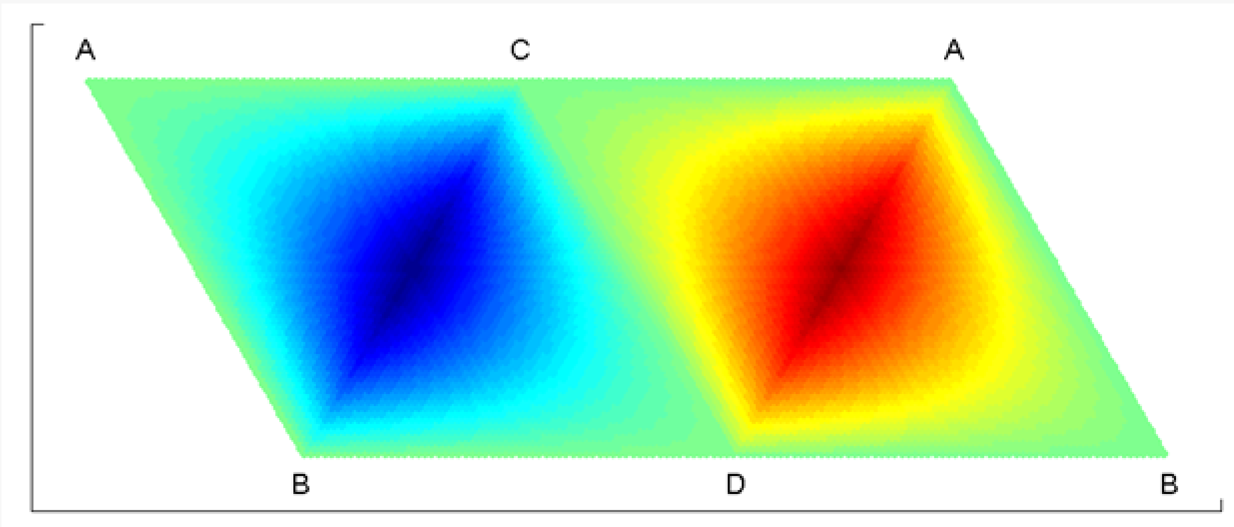
- The visualization of a 'solid' tetrahedron shows only extreme values of the arguments of the visualized function (external view):
  - 2D view
  - parallelogram – visualization of the net of the tetrahedron, i.e. a set of planar triangles, which when folded along selected edges, become the faces

Visualization techniques – external visualizations for *C*(*H*,*E*)

2D view



parallelogram

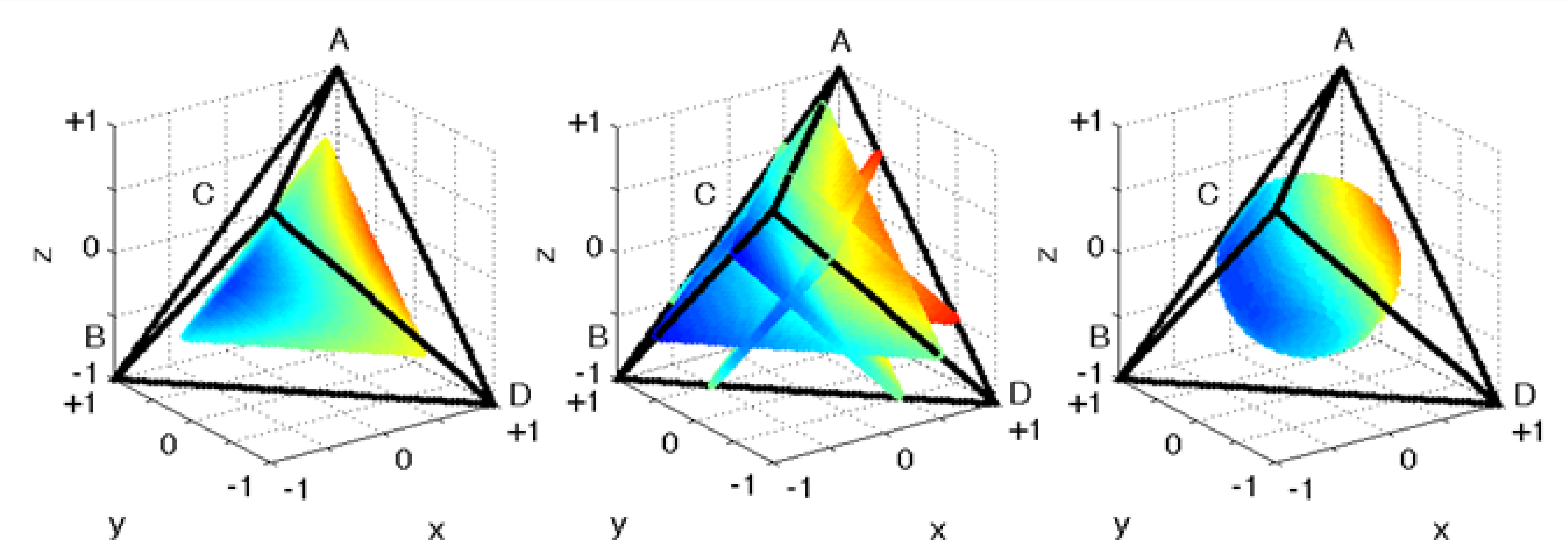


Visualization techniques – exemplary internal visualizations

- 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

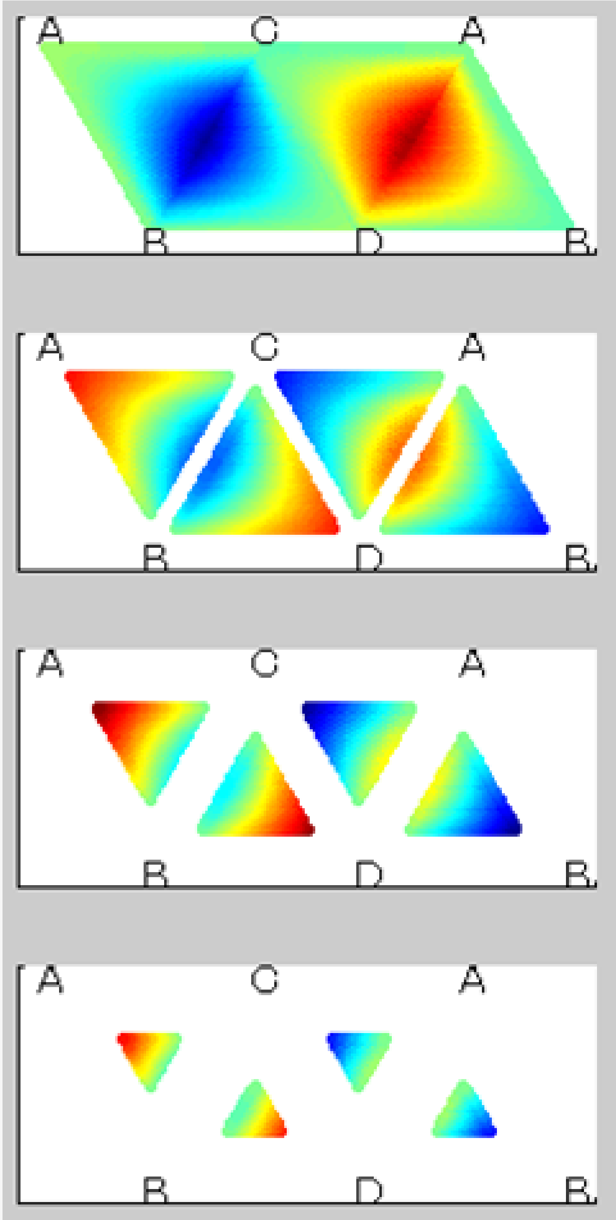
Visualization techniques – internal visualizations for $C(H,E)$

2D view



Visualization techniques – internal visualizations for $C(H,E)$

parallelograms



Visualization techniques – summary of the capabilities

- The capabilities of the visualization techniques include:
  - regular views of any operational function
  - specialized views of a region of interest, i.e. only points satisfying pre-defined conditions, e.g. $f(a,b,c,d)=0$, of any operational function
  - specialized views of any number of operational functions
    - differences between two operational functions
    - variances/means of a number of operational functions

Application of the visualization techniques to confirmation measures

Visualization techniques – summary of the capabilities

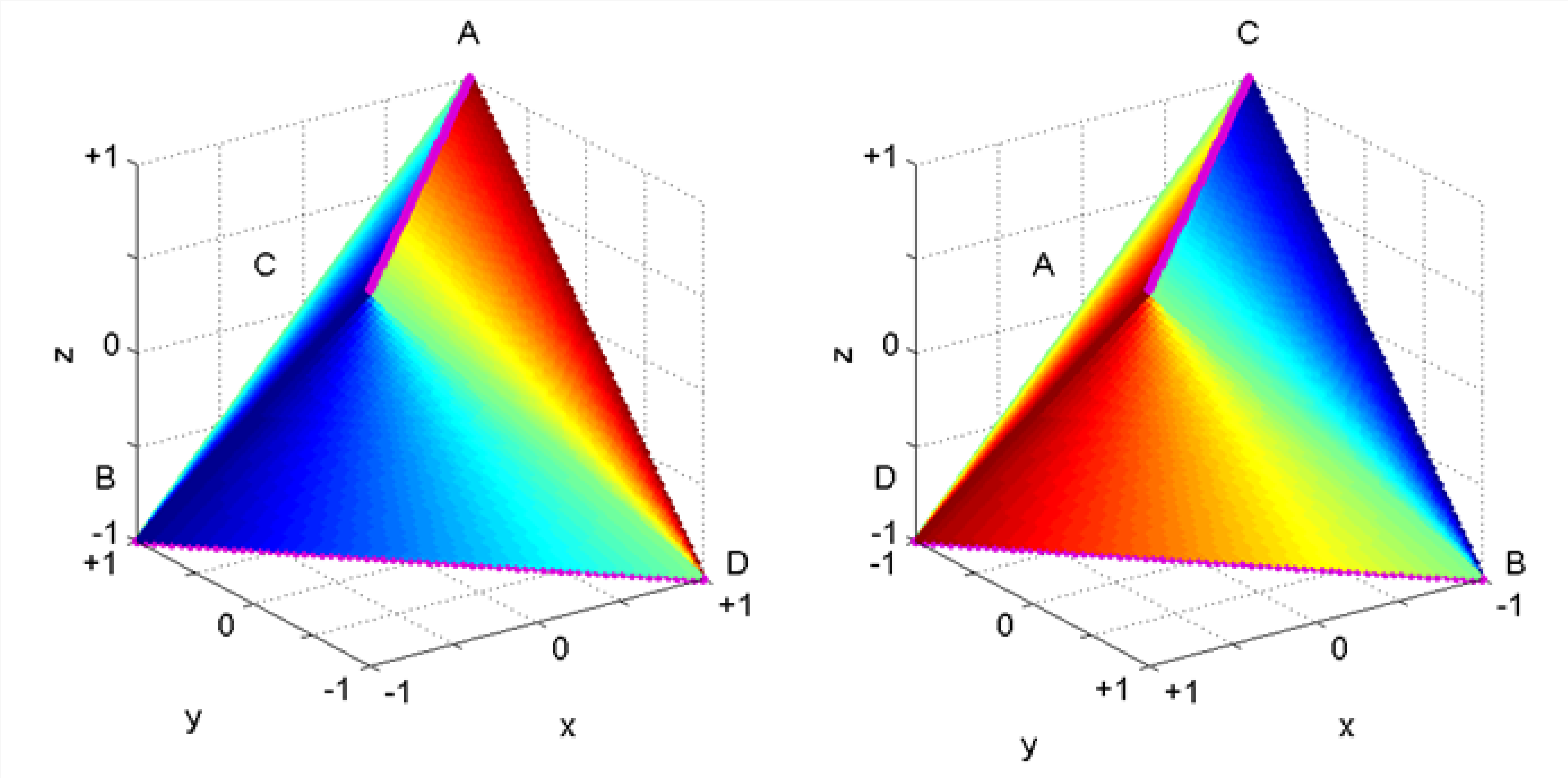
- The capabilities of the visualization techniques include:
  - regular views of any operational function
  - specialized views of a region of interest
  - specialized views of any number of operational functions
    - differences between two operational functions
    - variances/means of a number of operational functions

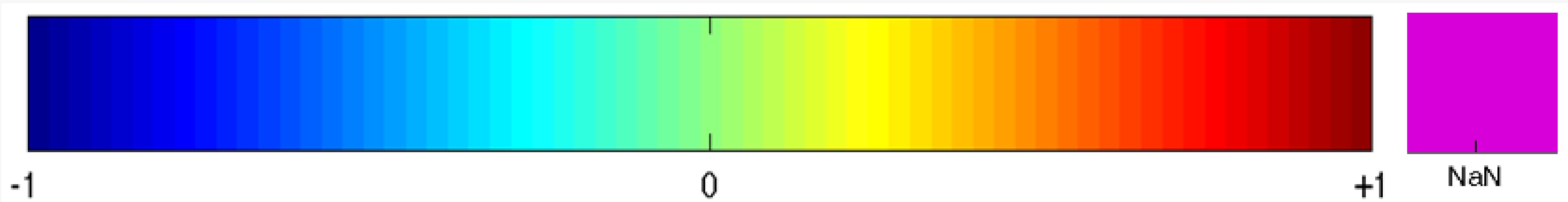
Regular views of confirmation measures

- The regular views of the measures may be used to practically compare their general configurations of values and gradient profiles

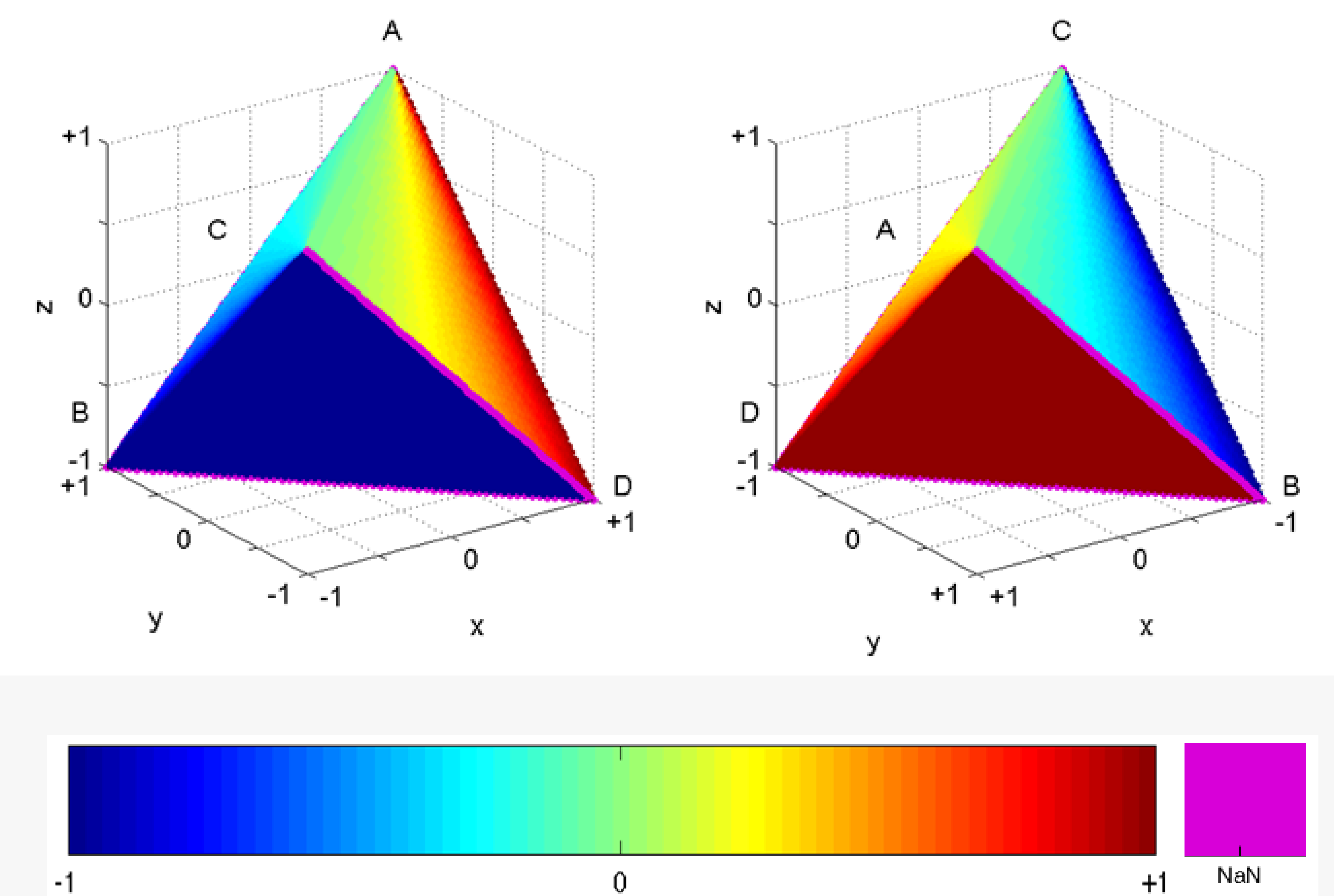
- Such visual analyses allow to tentatively conclude about the ordinal equivalence of the visualized measures, an especially important issue in evaluating rules with multiple measures
- In general, this kind of equivalence analysis may require an insight into the interior of the tetrahedron

Regular views of confirmation measures: *S*(*H*,*E*)





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



Regular views of confirmation measures

- In all their faces measures $S(H,E)$ manifest 'radial' gradients, while measure $F(H,E)$ is characterized by constant values (no gradient) in two faces (*ABD* and *BCD*) and a 'radial' gradient in the other two

- In the case of $S(H,E)$ and $F(H,E)$ the different gradient profiles in the external areas of the corresponding tetrahedrons constitute conclusive counterexamples to the ordinal equivalence of those measures

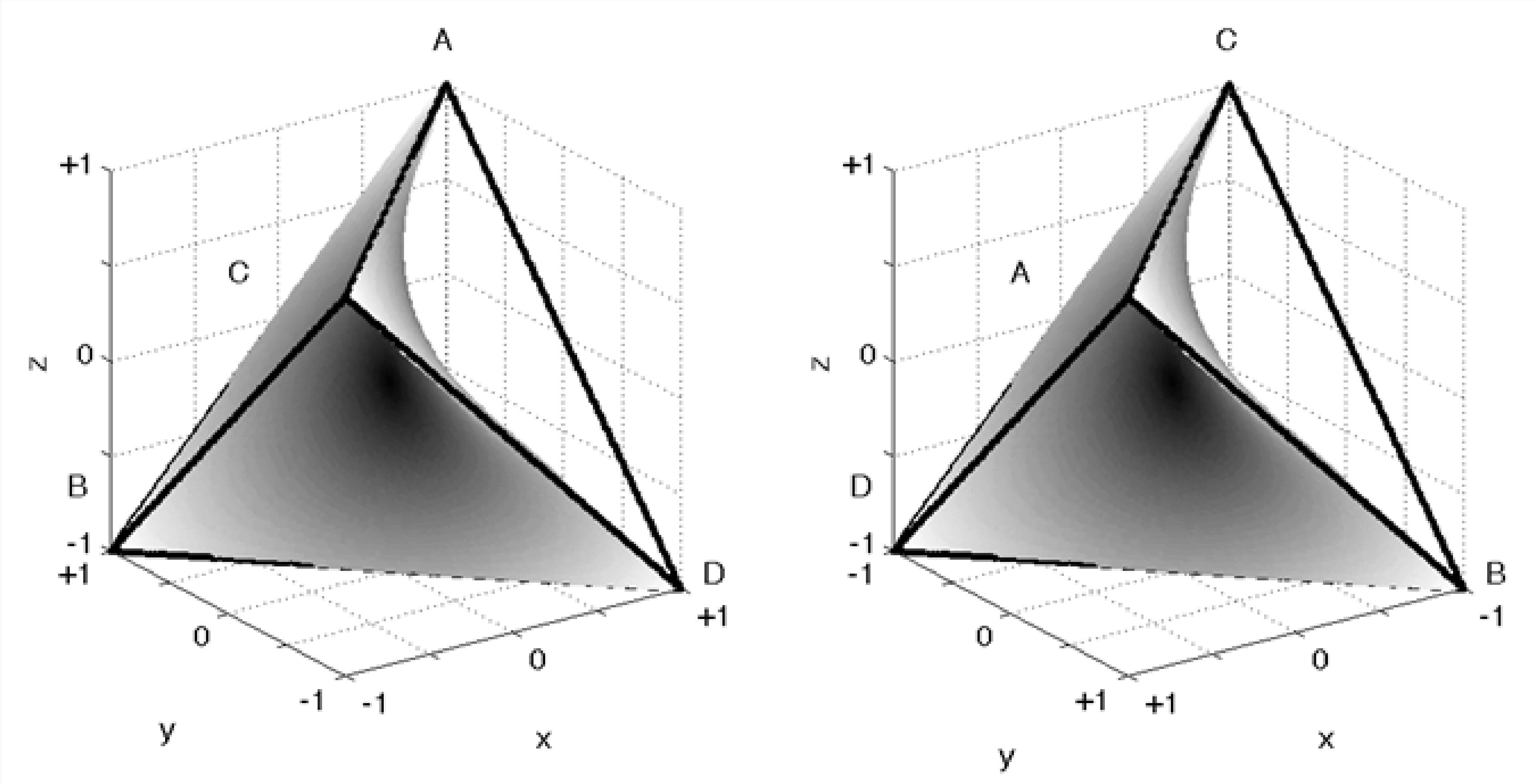
Visualization techniques – summary of the capabilities

- The capabilities of the visualization techniques include:
  - regular views of any operational function
  - specialized views of a region of interest
  - specialized views of any number of operational functions
    - differences between two operational functions
    - variances/means of a number of operational functions

Specialized views of regions of interest

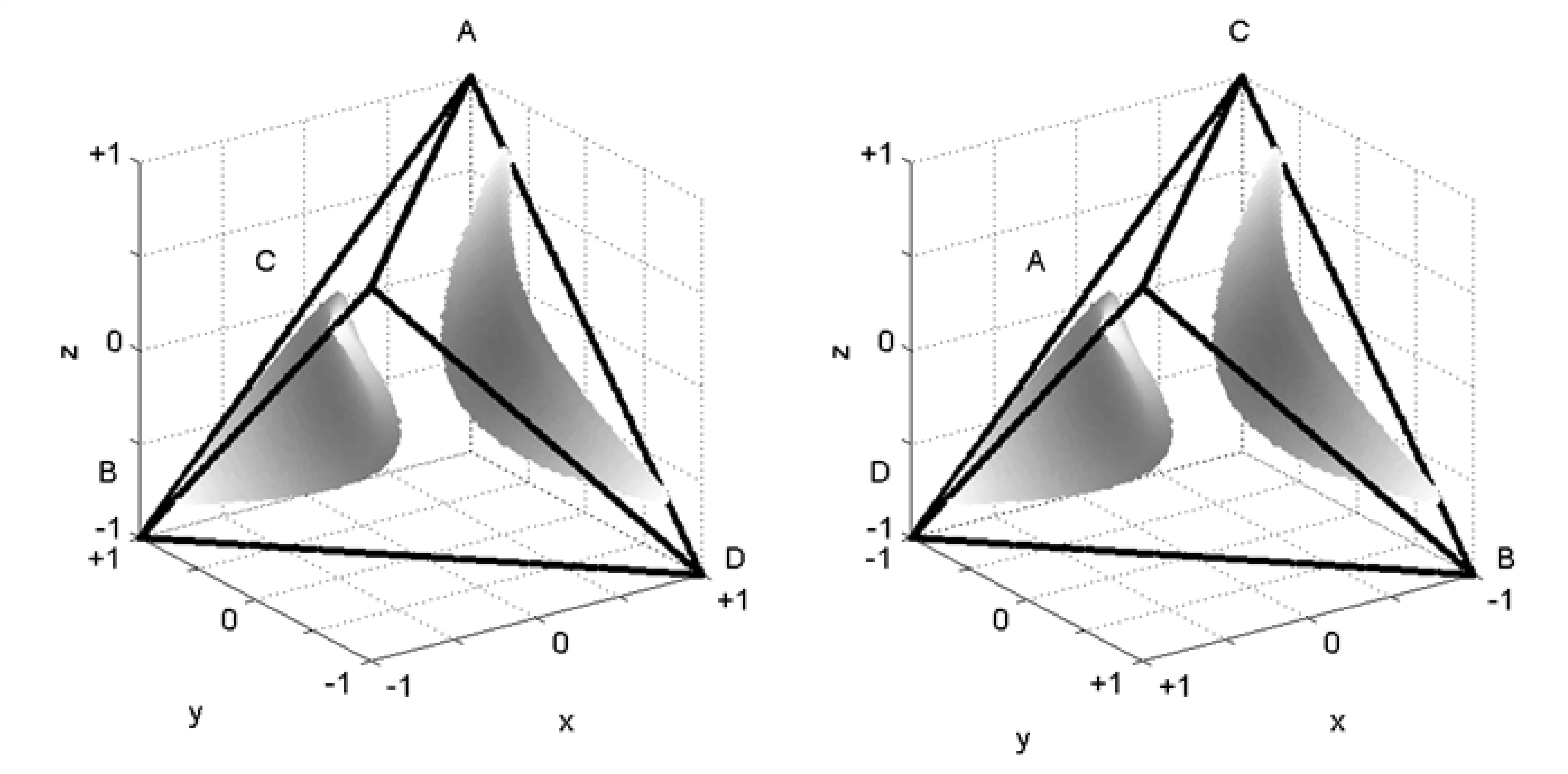
- The specialized views of regions of interest are useful since they allow to instantly detect and localize interesting characteristics of the measures (extreme values, zeros, etc.), which would otherwise have to be laboriously derived from the analytic definitions of the measures

Specialized views of regions of interest: $c(H,E)=0$



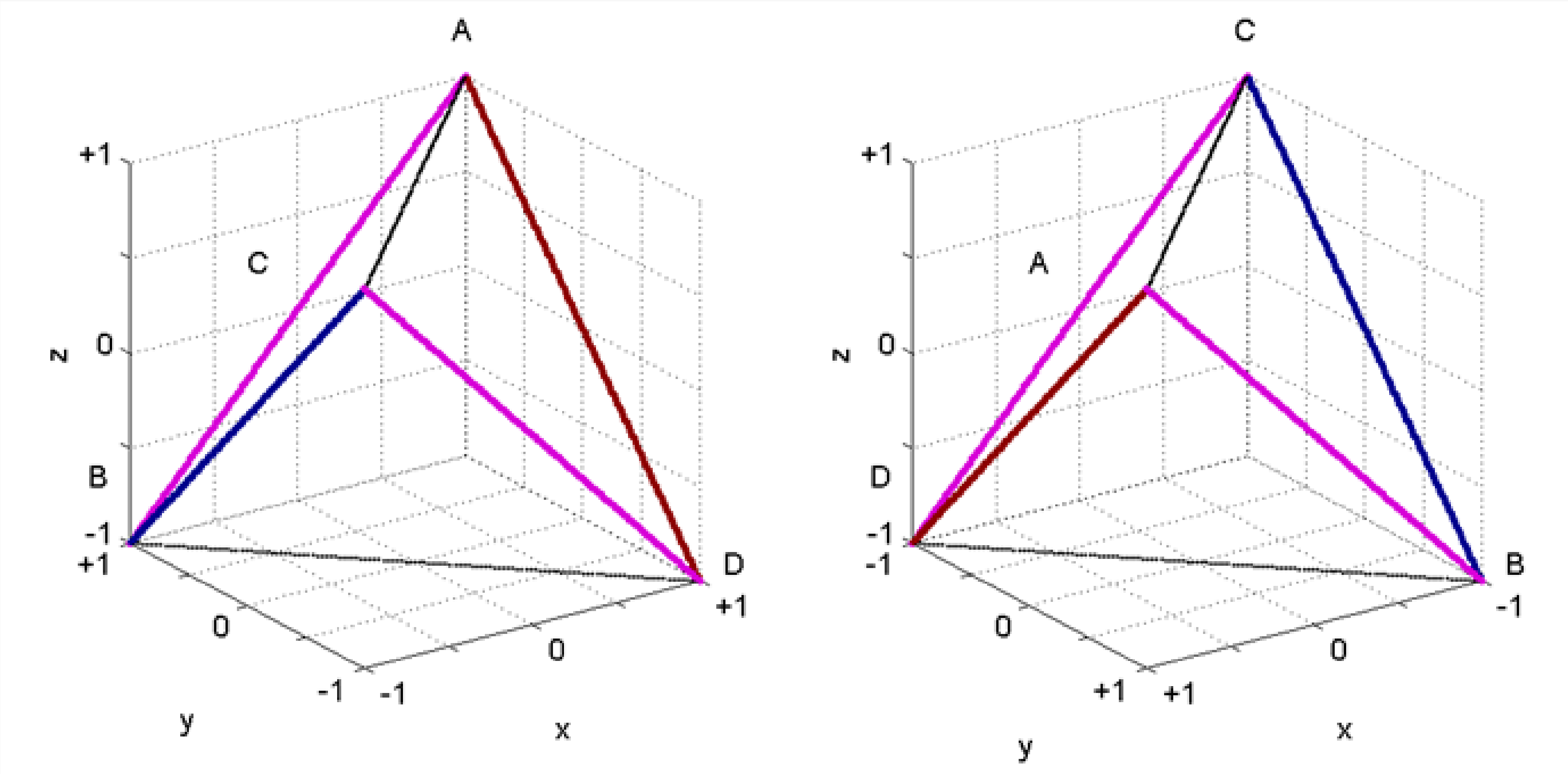
- Regions with neutral values of confirmation measures
- The grey colour map is used only to provide the necessary perspective; the colours do not translate to values of the measure (which are constant in this case)

Specialized views of regions of interest: $C(H,E)=0.5$

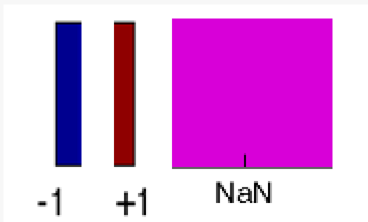


- Regions for which $|C(H,E)|=0.5$; notice their full symmetry
- The grey colour map is used only to provide the necessary perspective; the colours do not translate to values of the measure (which are constant in this case)

Specialized views of regions of interest: $N(H,E)$=min/NaN/max



- Regions of extreme (-1 and +1) and non-numeric values (*NaN*) of measure *N*(*H*,*E*)



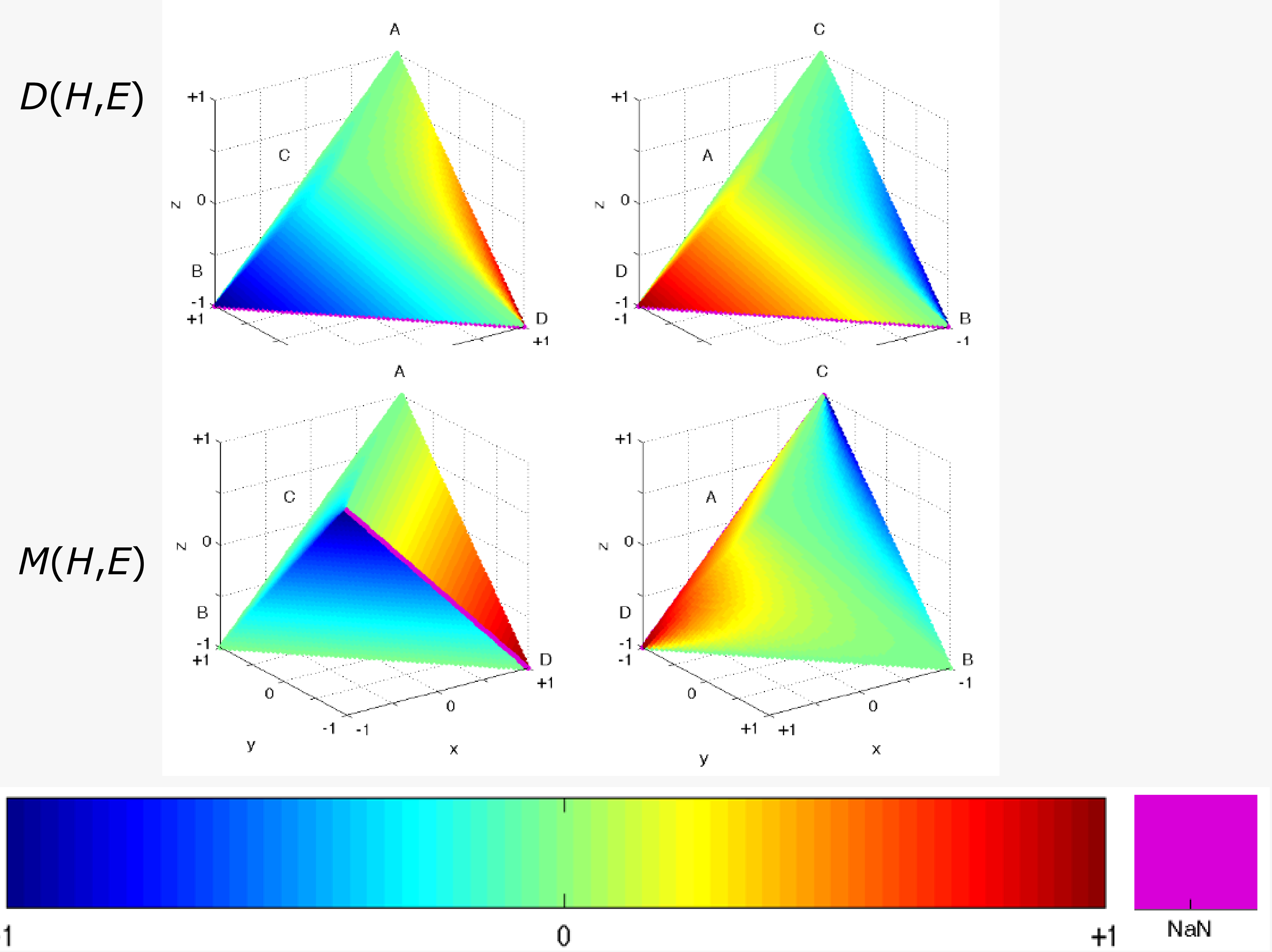
Visualization techniques – summary of the capabilities

- The capabilities of the visualization techniques include:
  - regular views of any operational function
  - specialized views of a region of interest
  - specialized views of any number of operational functions
    - differences between two operational functions
    - variances/means of a number of operational functions

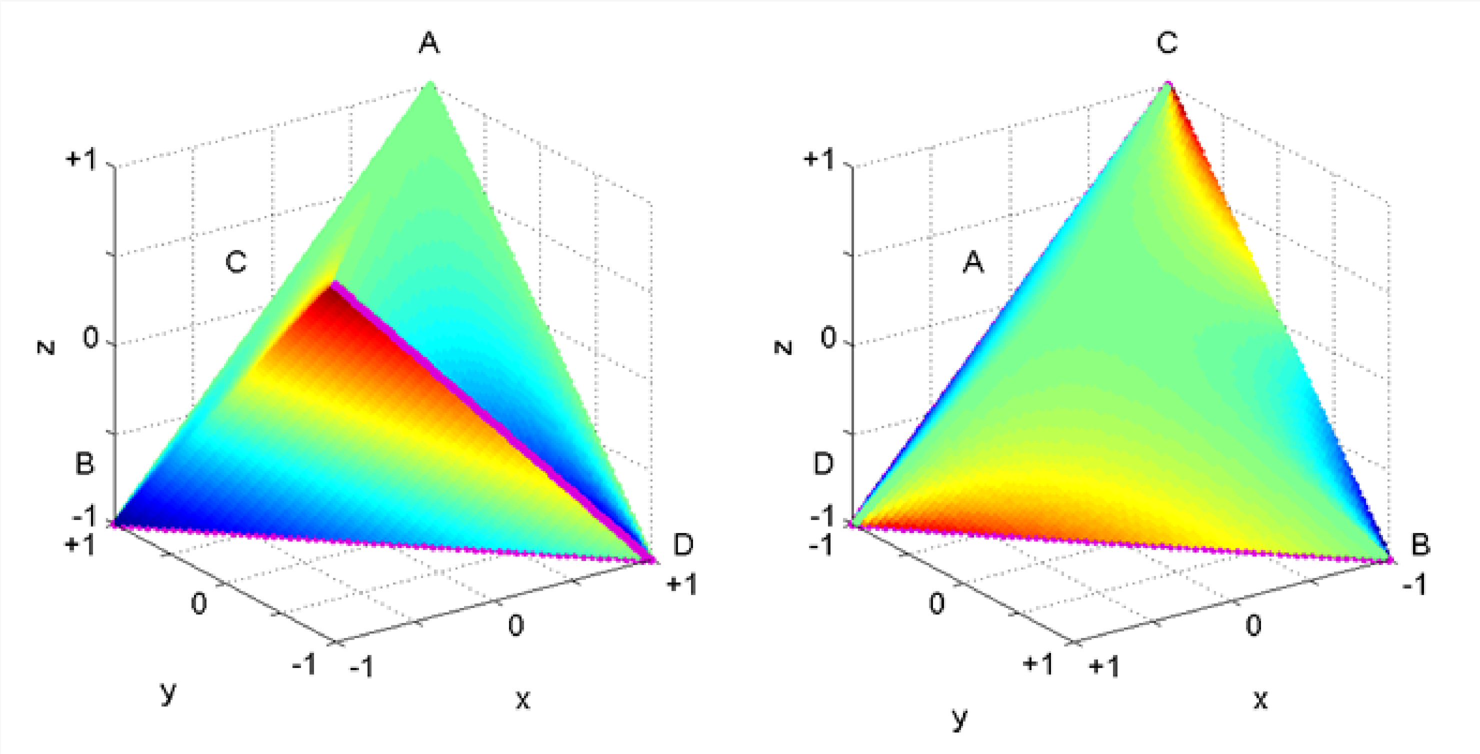
Specialized views – differences /variance among measures

- Visualization of differences between measures or variances among groups of measures allows to identify those arguments (i.e. values of *a*, *b*, *c* and *d*) for which two given measures differ only insignificantly (similarity of the measures) or differ considerably (dissimilarity of the measures)
- Thus, it guides practitioners towards measures that suit them most

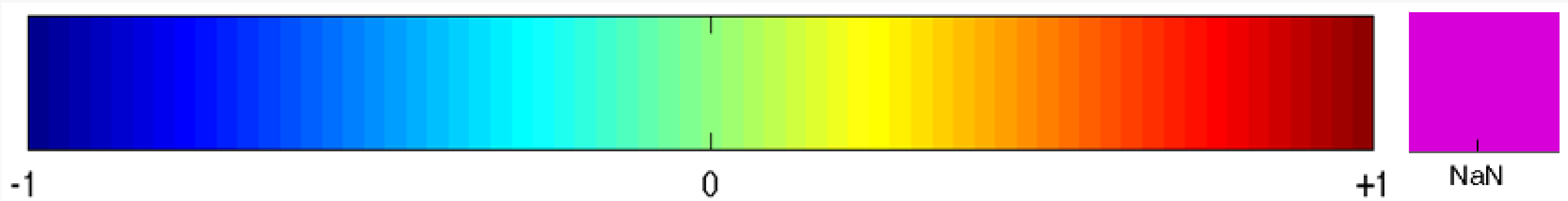
Specialized views-differences between measures: $D(H,E)$-$M(H,E)$



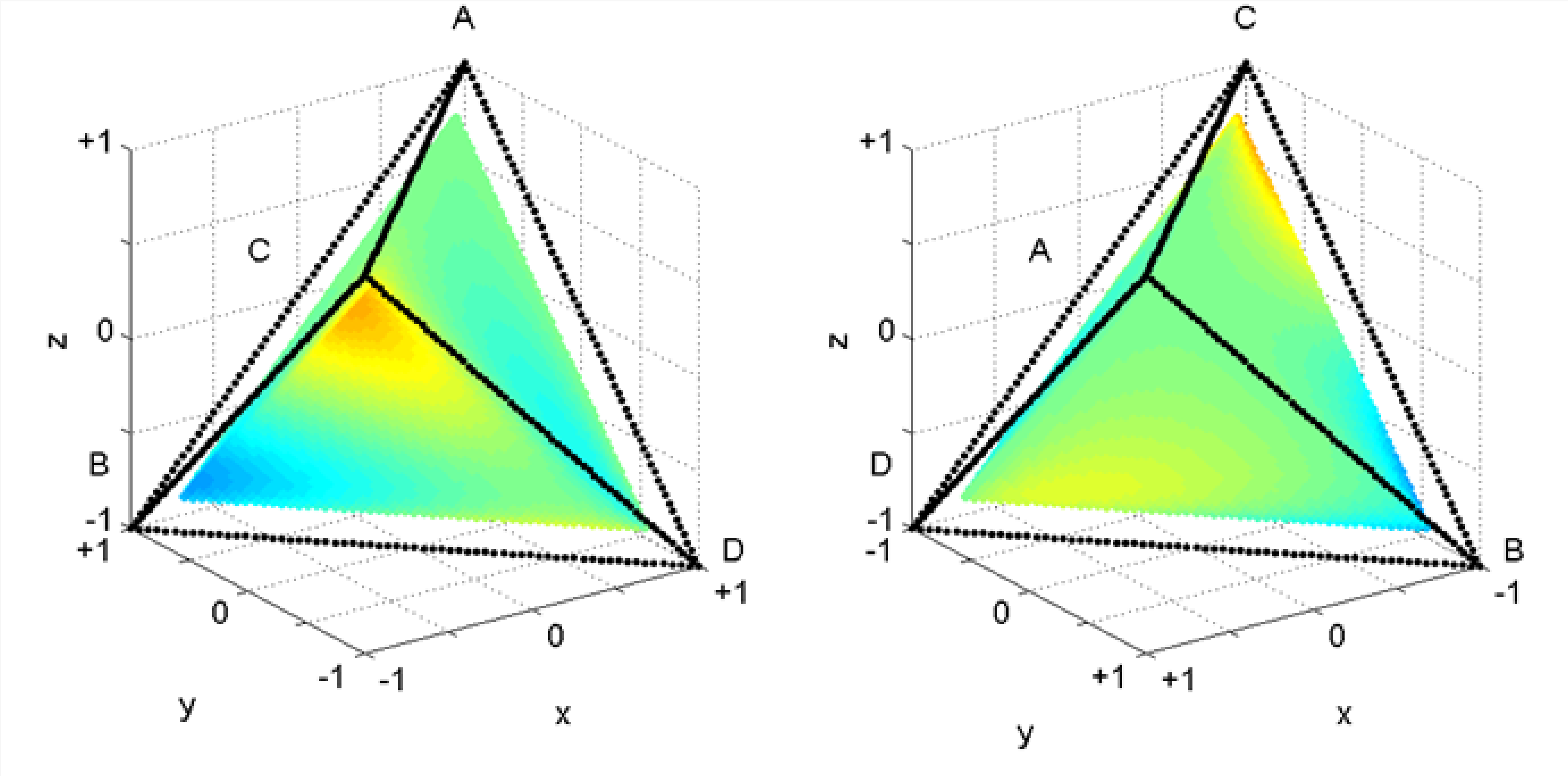
Specialized views-differences between measures: *D*(*H*,*E*)-*M*(*H*,*E*)



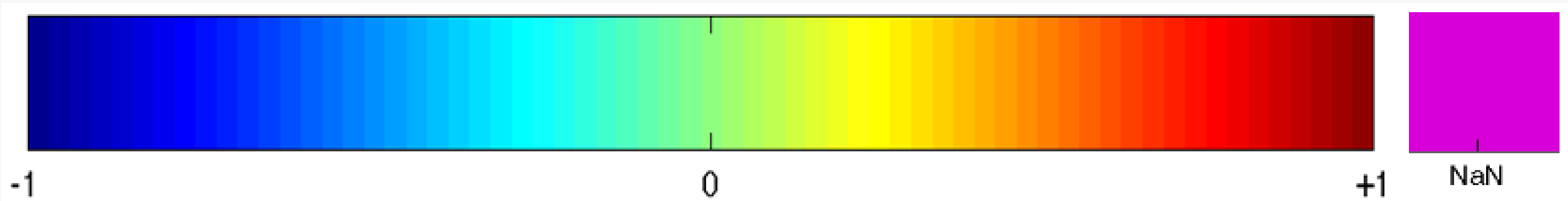
- The exterior view of *D*(*H*,*E*) - *M*(*H*,*E*)
- *D*(*H*,*E*) exceeds *M*(*H*,*E*) most in the vicinity of the *C* vertex, while *M*(*H*,*E*) exceeds *D*(*H*,*E*) most in the vicinity of the *B* vertex



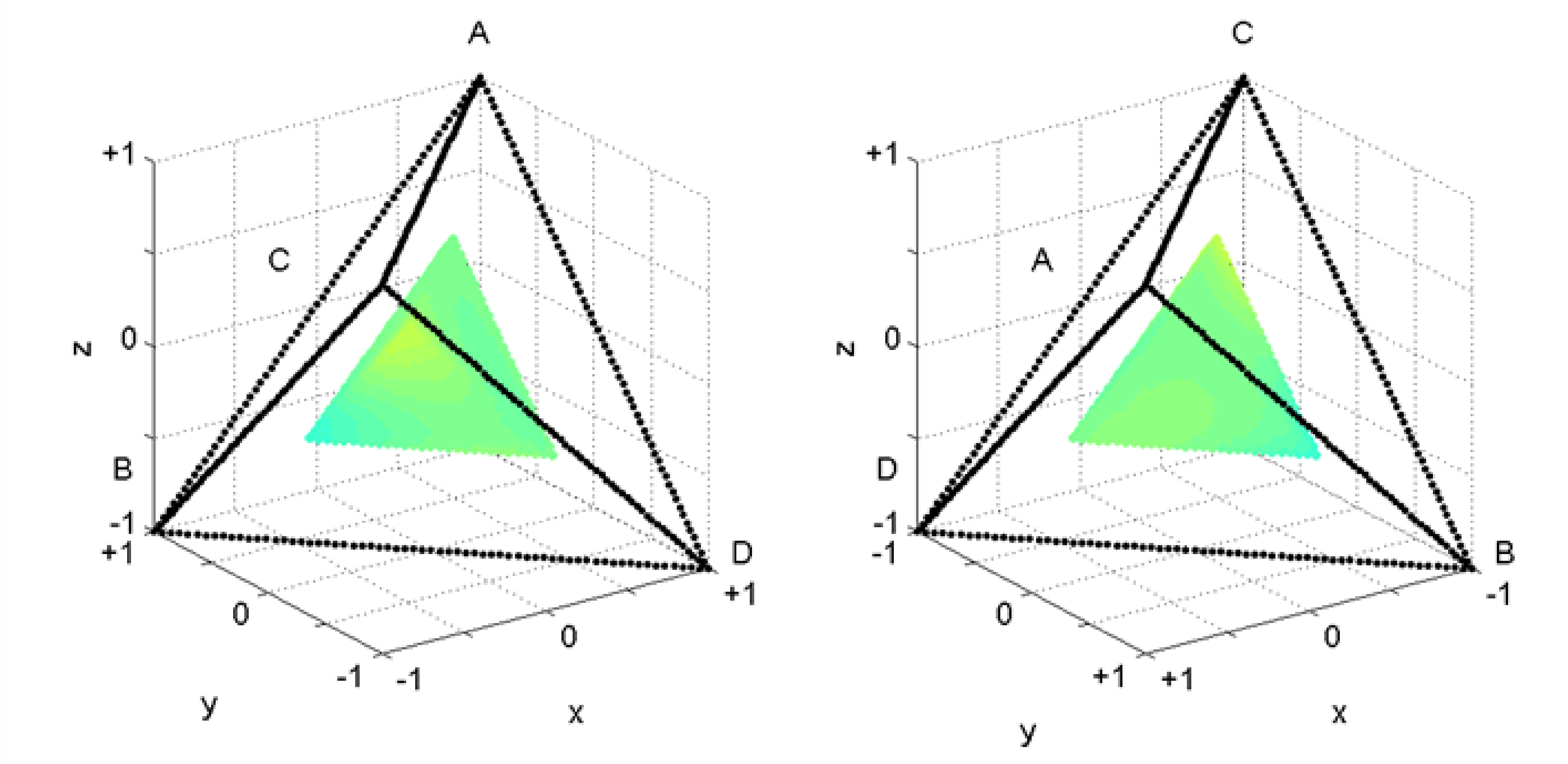
Specialized views-differences between measures: *D*(*H*,*E*)-*M*(*H*,*E*)



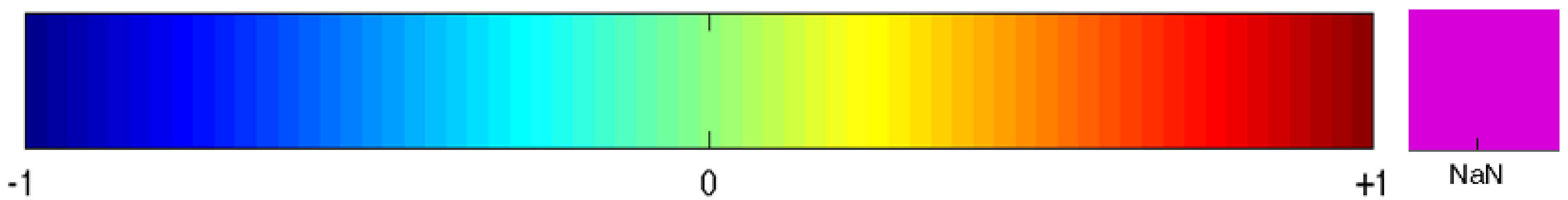
- The inner view of *D*(*H*,*E*) - *M*(*H*,*E*)



Specialized views-differences between measures: $D(H,E)$-$M(H,E)$



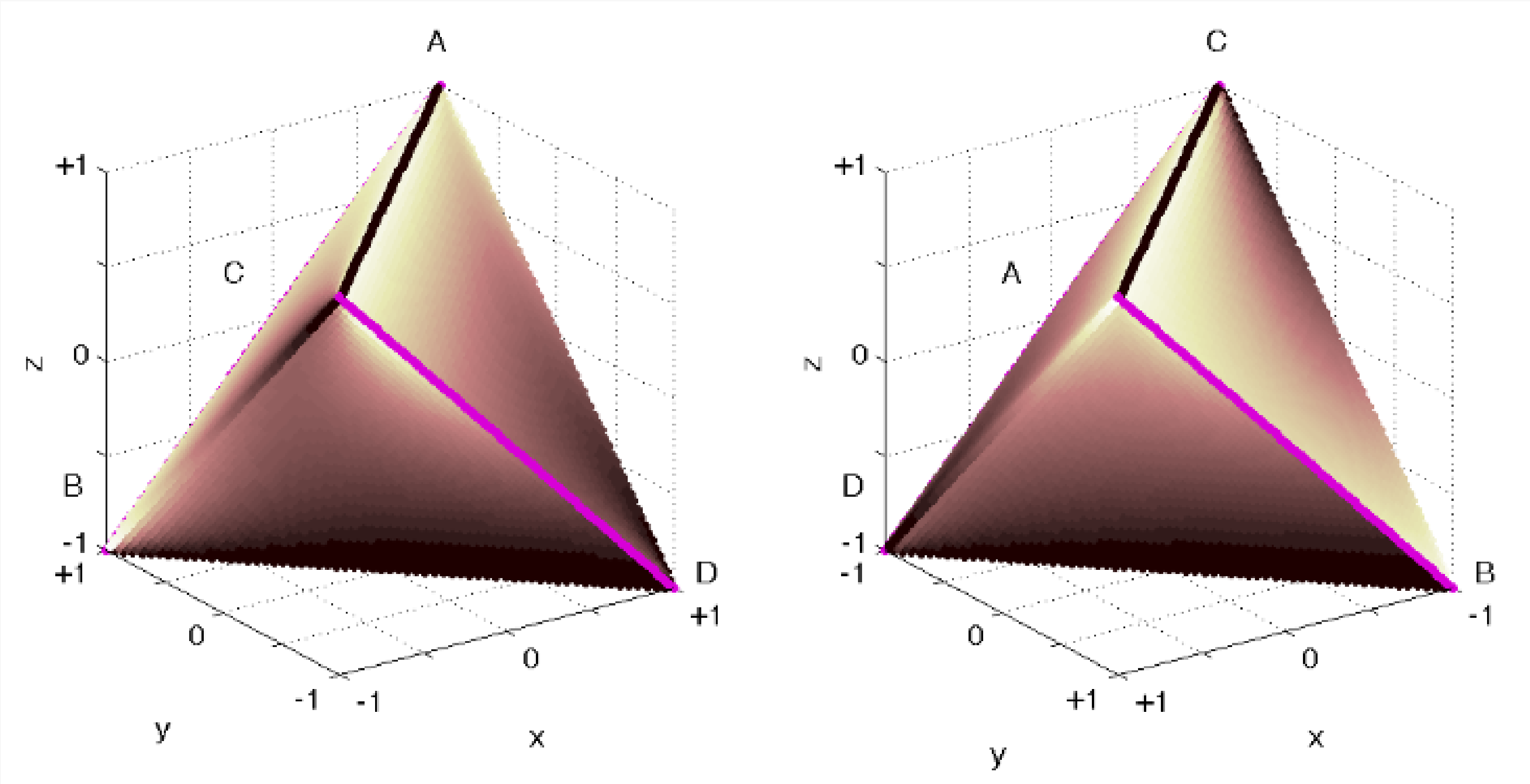
- The inner view of $D(H,E)$ - $M(H,E)$



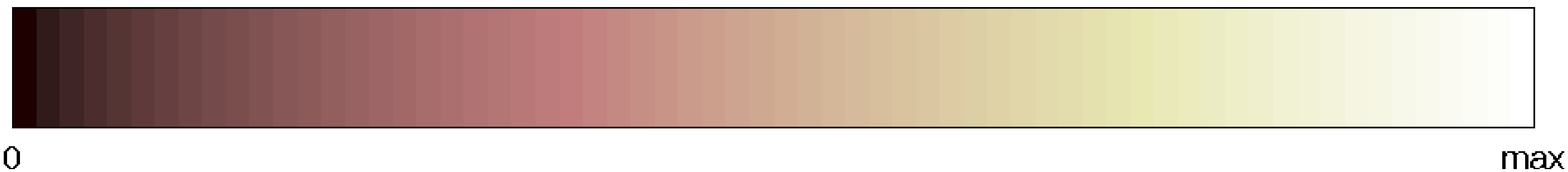
Visualization techniques – summary of the capabilities

- The capabilities of the visualization techniques include:
  - regular views of any operational function
  - specialized views of a region of interest
  - specialized views of any number of operational functions
    - differences between two operational functions
    - variances/means of a number of operational functions

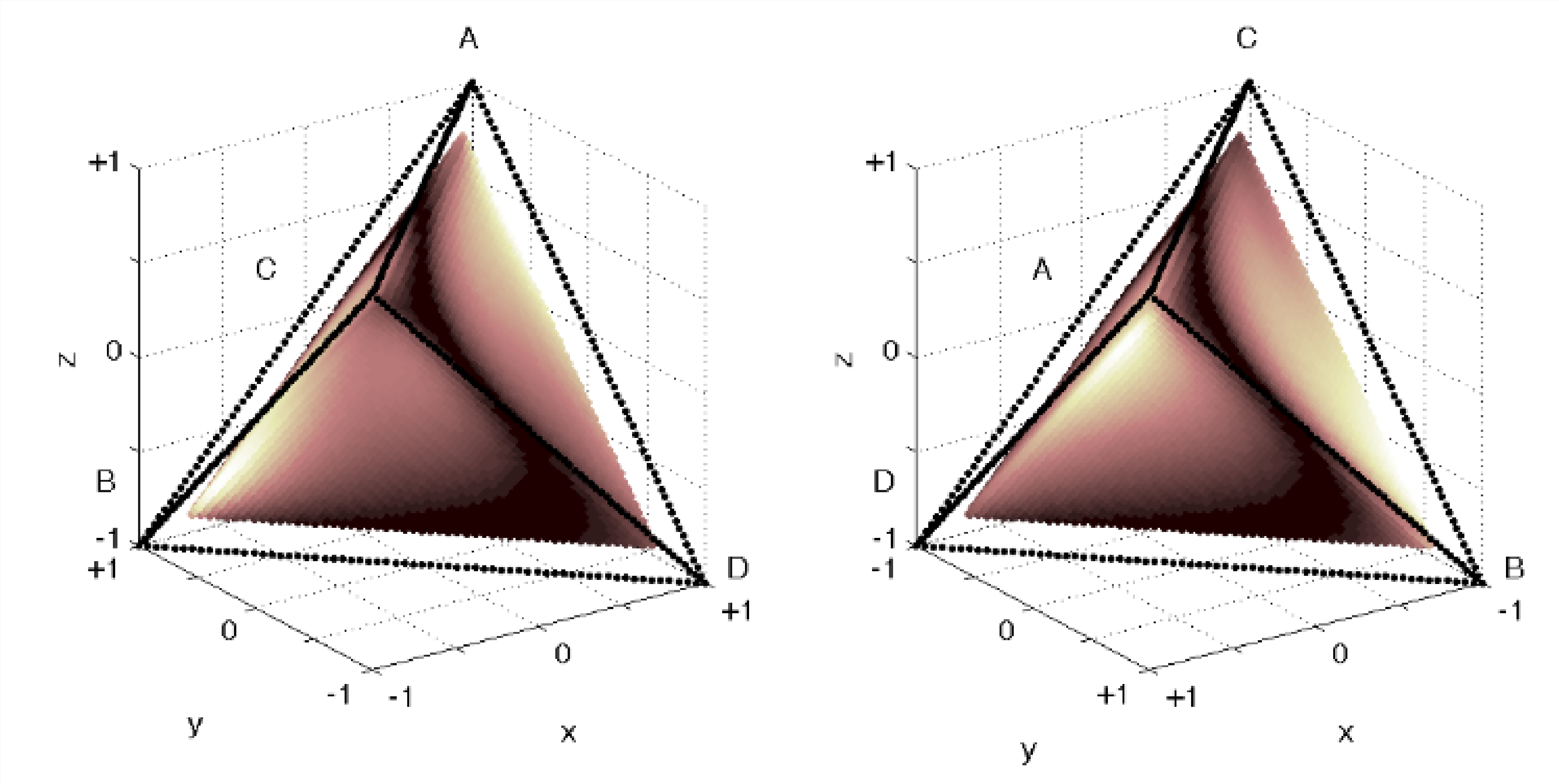
Specialized views - variance among likelihoodist measures



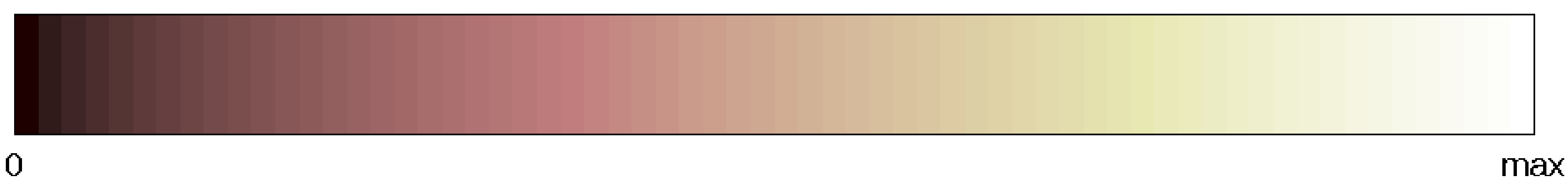
- The variance among measures: $M(H,E)$, $N(H,E)$, $A(H,E)$, $c_2(H,E)$



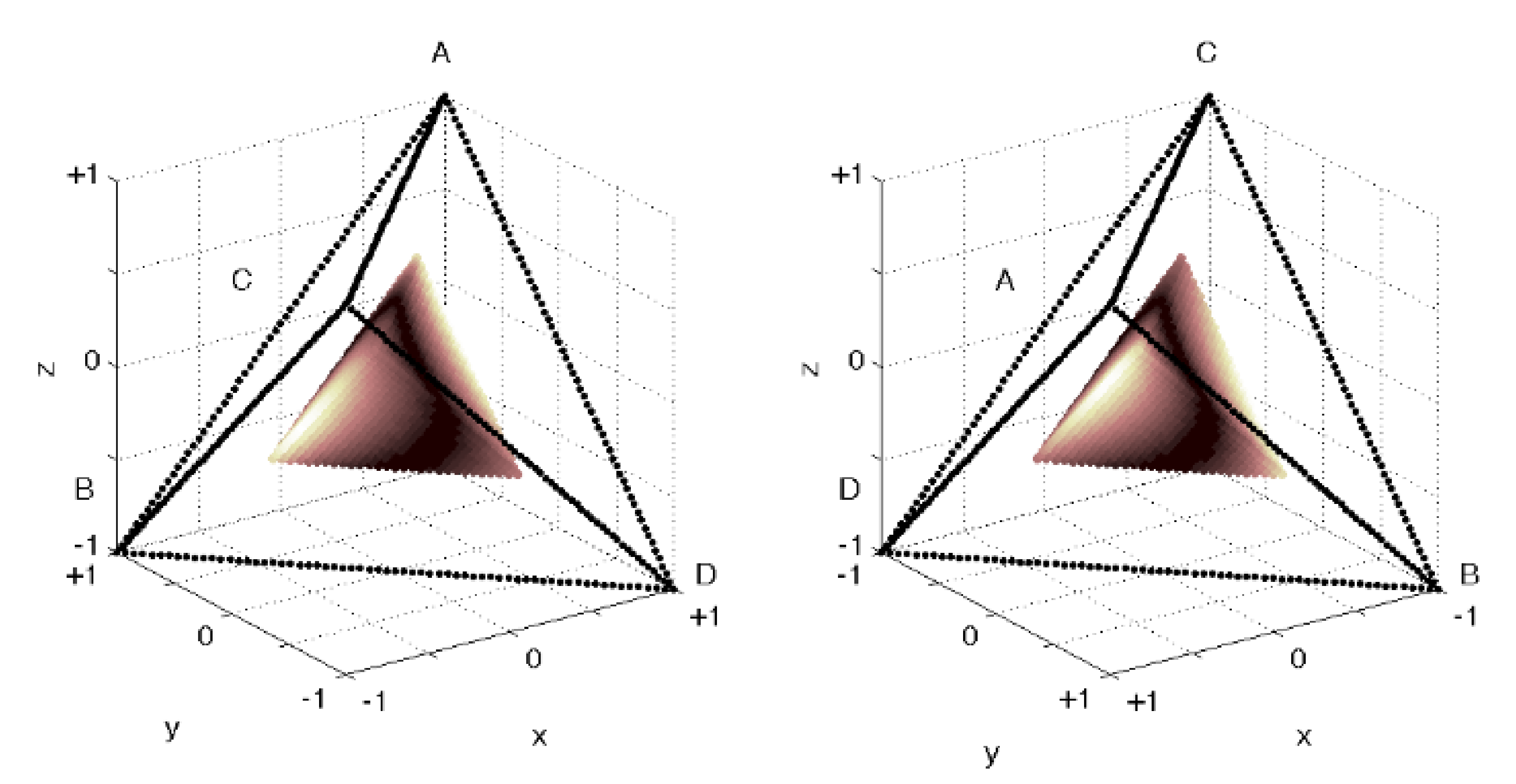
Specialized views - variance among likelihoodist measures



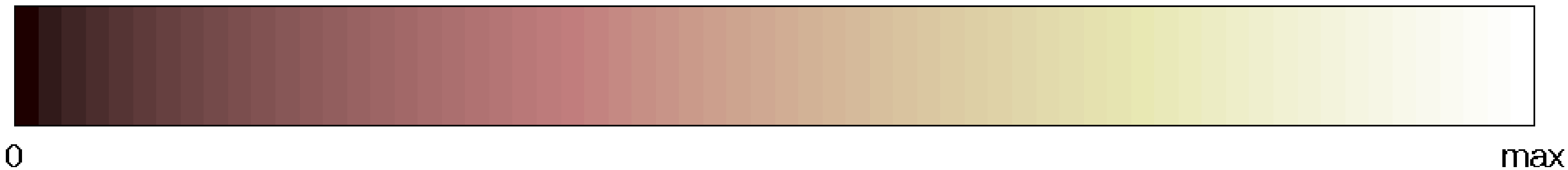
- The inner view of the variance among: $M(H,E)$, $N(H,E)$, $A(H,E)$, $c_2(H,E)$



Specialized views - variance among likelihoodist measures



- The inner view of the variance among: $M(H,E)$, $N(H,E)$, $A(H,E)$, $c_2(H,E)$



Conclusions

- Our visualization tool for interestingness measures provides practical insights into different details of the analysed measures

- The originally 4-dimensional arguments of the measures are effectively represented in three dimensions using a tetrahedron-based barycentric coordinate system, with values of any operational function, e.g. an interestingness measure, rendered as colour

Conclusions

- The visual analyses are especially useful since they allow to instantly detect and localize interesting characteristics of the measures (extreme values, zeros, etc.), which would otherwise have to be laboriously derived from the analytic definitions of the measures
- Our visualization helps to determine e.g. if the visualized measures are identical or similar in particular domain regions, or if they are ordinally equivalent
- The gained insights swiftly guides the practitioner towards interestingness measures that best reflect his/her expectations

Thank you!