Advanced Data Analysis for Industrial Applications

Zdeněk Wagner

INSTITUTE OF CHEMICAL PROCESS FUNDAMENTALS OF THE CAS

Pavel Kovanic

retired from

INSTITUTE OF INFORMATION THEORY AND AUTOMATION OF THE CAS

Modelling Smart Grids 2015, Prague, September 10–11, 2015
http://www.smartgrids2015.eu/

Introductory note

This is an extended version of the slides to the lecture. During the lecture several pieces of information were only said without being displayed. For this reason a few explanatory slides and remarks have been added. If you have any questions, you can contact the authors via e-mail:

wagner@icpf.cas.cz

kovanic@email.cz

Typical tasks

Marketing – analysis of *big data* available from eShops, social media, internet of things etc. Extremal data may be present, they sometimes disturb analysis, sometimes supply the most valuable information.

Quality control – detection of defects, preferably when the quality of the product is still within acceptable limits.

Process control – analysis of real time data, early detection of departure from optimum conditions.

Safety – real-time analysis of concentration of hazardous waste, early detection of dangerous concentration.

Demand for robust methods of data analysis!

Statistical paradigm of uncertainty

- Distribution of errors known *a priori*, normal distribution often silently assumed in textbooks of statistics for engineers (ANOVA, F-test, χ^2 test)
- Robust statistical methods require additional assumptions on the distribution function of outliers
- Continuous distribution function derived for an infinite data set
- Properties of data obtained by extrapolation from an infinite to a finite data sample

Questions

- Do we know the distribution of data? *(quality of products, concentration of poisonous waste, flow rate of leakage, heterogeneities in the raw material, power consumption of home and industrial consumers)*
- Is the data sample large enough to make the extrapolation to the finite data sample valid?
- Are the outliers rare?
- Can the outliers be discarded without loss of important information?
- Is the data analysis algorithm robust so that it can run unattended and produce reliable results?

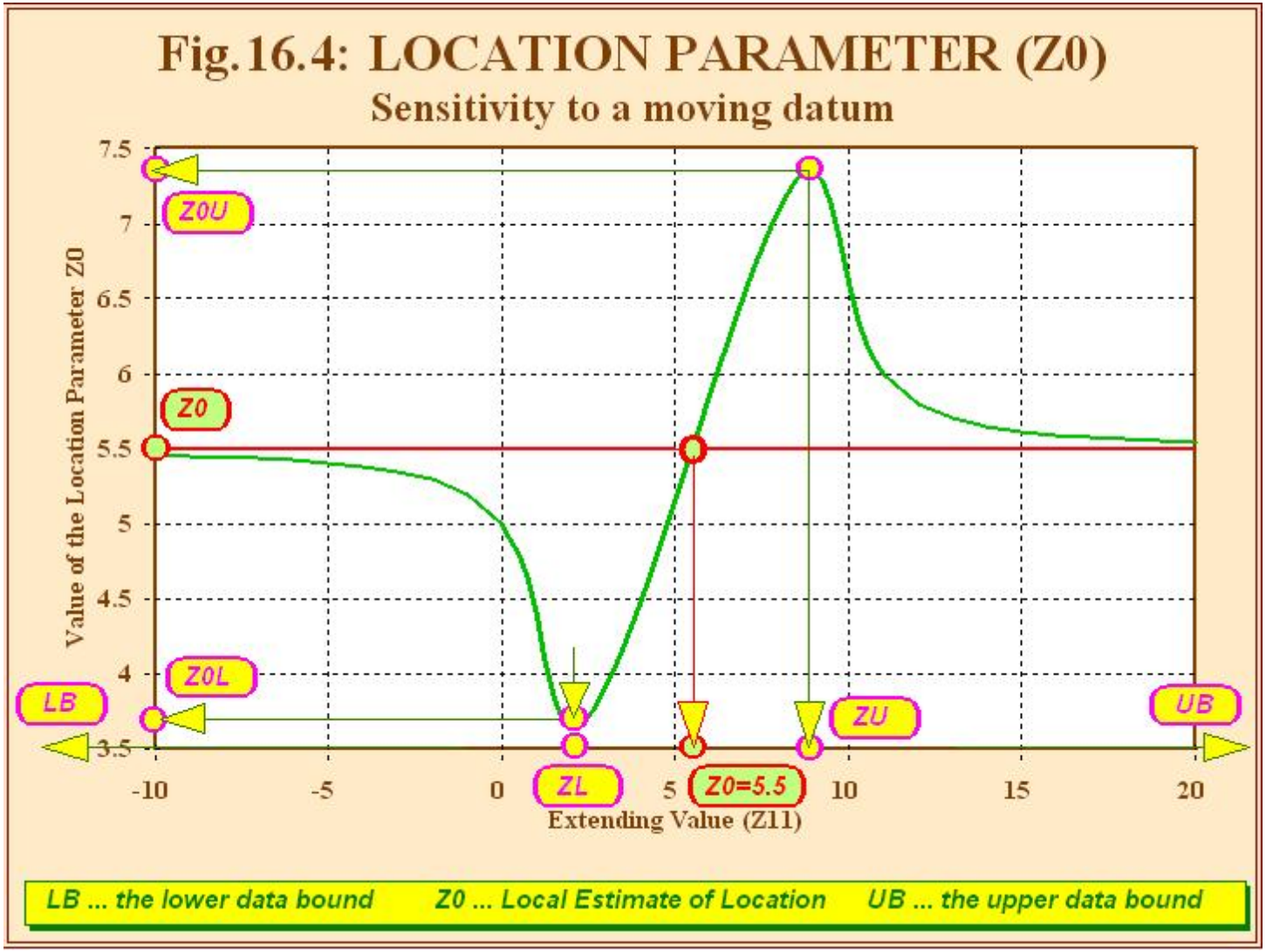
Principles of mathematical gnostics

- Derived from the fundamental laws of nature
- Based on the properties of each individual measurement
- Properties of a data sample obtained by aggregation of properties of individual data, hence the results are valid also for small data samples
- The distribution function as well as the metrics of the space estimated during data analysis: *Let the data speak for themselves!*
- Robustness is the inherent property

P. Kovanic, M. B. Humber: *The Economics of Information (Mathematical Gnostics for Data Analysis).* 717 pages. Updated in September 2013.

http://www.math-gnostics.com/index.php?a=books

Properties of the local estimate of location



Example 1, marginal analysis

Data from NIST Webbook Chemistry, http://webbook.nist.gov

Normal boiling temperature of 1,4-dichlorobutane (CAS 110-56-5)

Available data: 12 measured values

Value reported by NIST: 410 ± 80 K

Results obtained by mathematical gnostics

Parameter	Certifying Bound	Cum. Probability
LB	426.187	0
LSB	426.250	0.071
ZL	426.938	0.411
Z0L	427.057	0.457
Z0	427.097	0.472
Z0U	427.130	0.484
ZU	427.261	0.533
USB	428.150	0.929
UB	428.216	1

Explanation of symbols

Dx	additional (moving) datum
LB, UB	bounds of the data support
LSB, USB	bounds of domain of sample's homogeneity
ZL, UL	bounds of the interval of typical data
Z0L, Z0U	bounds of the tolerance interval
Z0	local estimate of location

Notes:

1. If the moving datum falls outside the interval delimited by (*LSB*, *USB*), the extended data set in *not* homogeneous.
2. Notation Z_* is used for a multiplicative model, A_* for an additive model. Both models are mathematically equivalent and one model can be transformed to the other one by $Z_x = \exp(A_x)$.

Data classification

Class No.	Condition	Data class
1	$Dx \leq LB$	L-outlier
2	$LB < Dx \leq LSB$	L-dubious
3	$LSB < Dx \leq ZL$	L-subtypical
4	$ZL < Dx \leq Z0L$	L-typical
5	$Z0L < Dx < Z0$	L-tolerated
6	$Dx = Z0$	Max. density
7	$Z0 < Dx \leq Z0U$	U-tolerated
8	$Z0U < Dx \leq ZL$	U-typical
9	$ZL < Dx \leq USB$	U-overtypical
10	$USB < Dx < UB$	U-dubious
11	$UB \leq Dx$	U-outlier

Results of data certification

Standard data			
Data No.	Value	Cum. Prob.	Class No.
8	426.25	0.071	2
12	426.65	0.293	3
10	427.05	0.454	4
3	427.1	0.473	6
6	427.15	0.492	7
11	428	0.830	8
4	428.15	0.929	8

Nonstandard data (outliers)					
Data No.	9	5	2	7	1
Data value	308.15	322	433	434.65	435.2

Example 2, marginal analysis

Data from NIST Webbook Chemistry, http://webbook.nist.gov

Normal boiling temperature of chloroform (CAS 67-66-3)

Available data: 37 measured values

Value reported by NIST: 334.3 ± 0.2 K

Results obtained by mathematical gnostics

Data split to 7 subsamples, 5 with 5 items each, 2 with 6 items each.

Parameter	Median	MAD	%
LB	334.199	0.104	0.031
LSB	334.240	0.059	0.018
ZL	334.328	0.040	0.012
Z0L	334.331	0.033	0.010
Z0	334.334	0.041	0.012
Z0U	334.340	0.043	0.013
ZU	334.339	0.043	0.013
USB	334.450	0.044	0.013
UB	334.451	0.071	0.021

MAD = *mean absolute deviation from the median*

Example 3, particle size distribution

- Particle size distribution in atmospheric aerosol measured by an SMPS (scanning mobility particle sizer) and the data transfered via internet once per hour
- Time series filtered in order to remove disturbances caused by instrument malfunction and local pollution events
- Distribution function estimated, number of modes estimated using a condition of equality of entropy of the data and the distribution function
- The results graphically displayed in near real time on the web – http://hroch486.icpf.cas.cz/Kosetice/

The procedure runs reliably since May 1, 2008. The graphical display offers early detection of instrument malfunction and usually even diagnostics on distance.

Example 4, energetics

- Real time measurement of transfered power plant output
- Real time measurement of the electrical network frequency
- Measurement of frequency/power sensitivity
  (failure of 1000 MW block in Germany not detected in Prague but the quasiperiodic response to switching the Vltava cascade on/off for 2 minutes repeated four times can be detected)

Kovanic P., Votlučka J., Blecha K.: *Experimental determination of the frequency/power coefficients of an electricity distributing system by means of periodical impulses of power* (in Russian), Elektrotechnický obzor (Review of Electrical Engineering) 68 (1979), 3, 133–139.

Development of an experimental technique

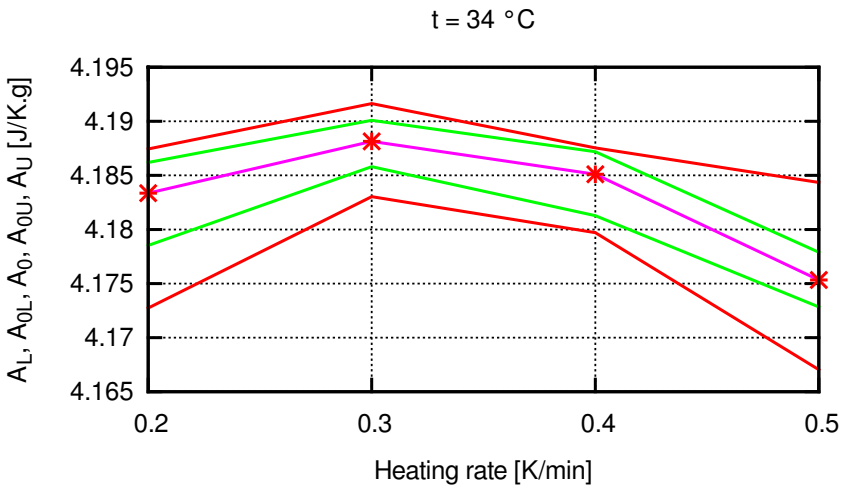
Measurement of heat capacity (C_p) by a continuous method by using a Setaram DSC3EVO calorimeter

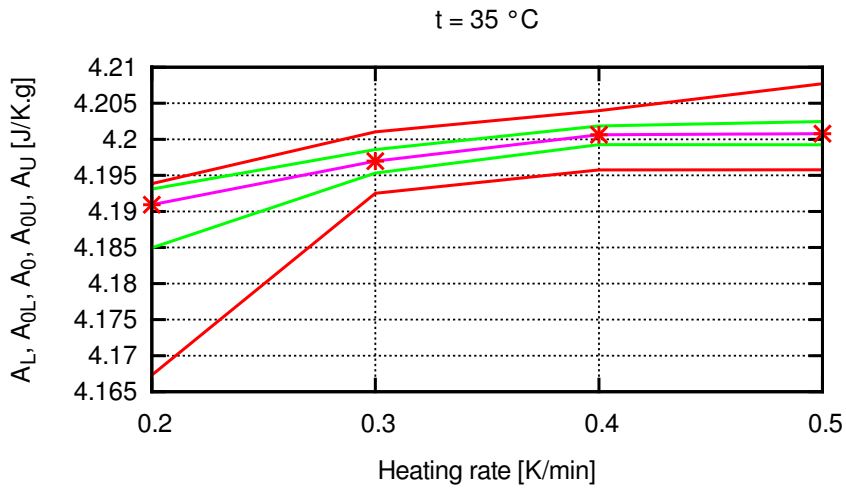
Task: find the heating rate ensuring the best repeatibility
(n_{min} = minimum sample size for 10% error in deviation)

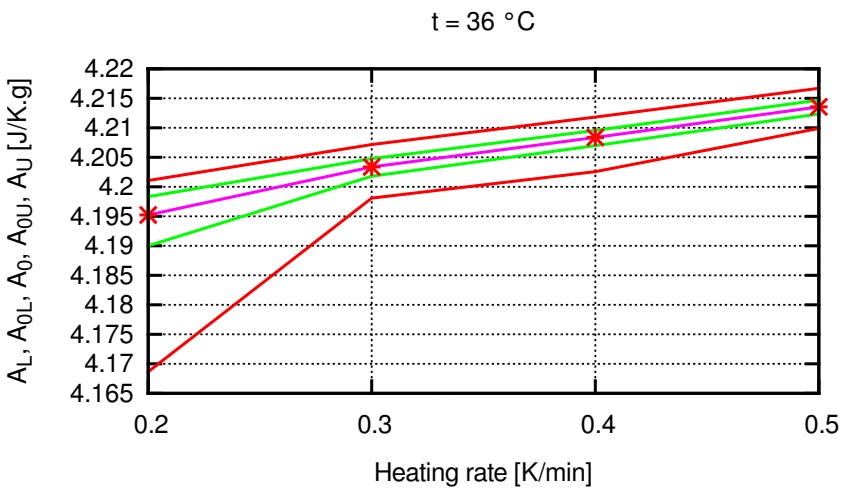
Distribution	**Kurtosis**	n_{min}	time [weeks]
Uniform	1.8	21	4
Normal	3.0	51	10
Exponential	6.0	126	26
Laplace	9.0	201	41
Lognormal	15.0	351	72

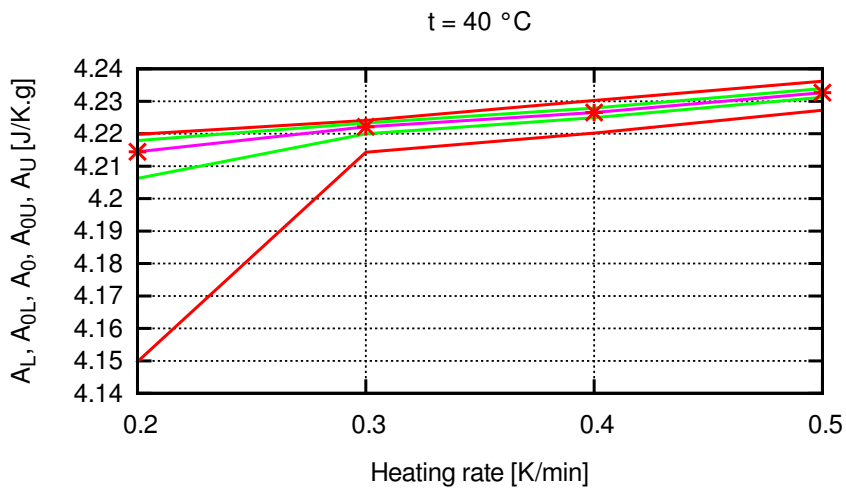
Time needed for reliable determination of tolerance interval and interval of typical data by mathematical gnostics: *less than 1 week*

Analysis of results of C_p measurement

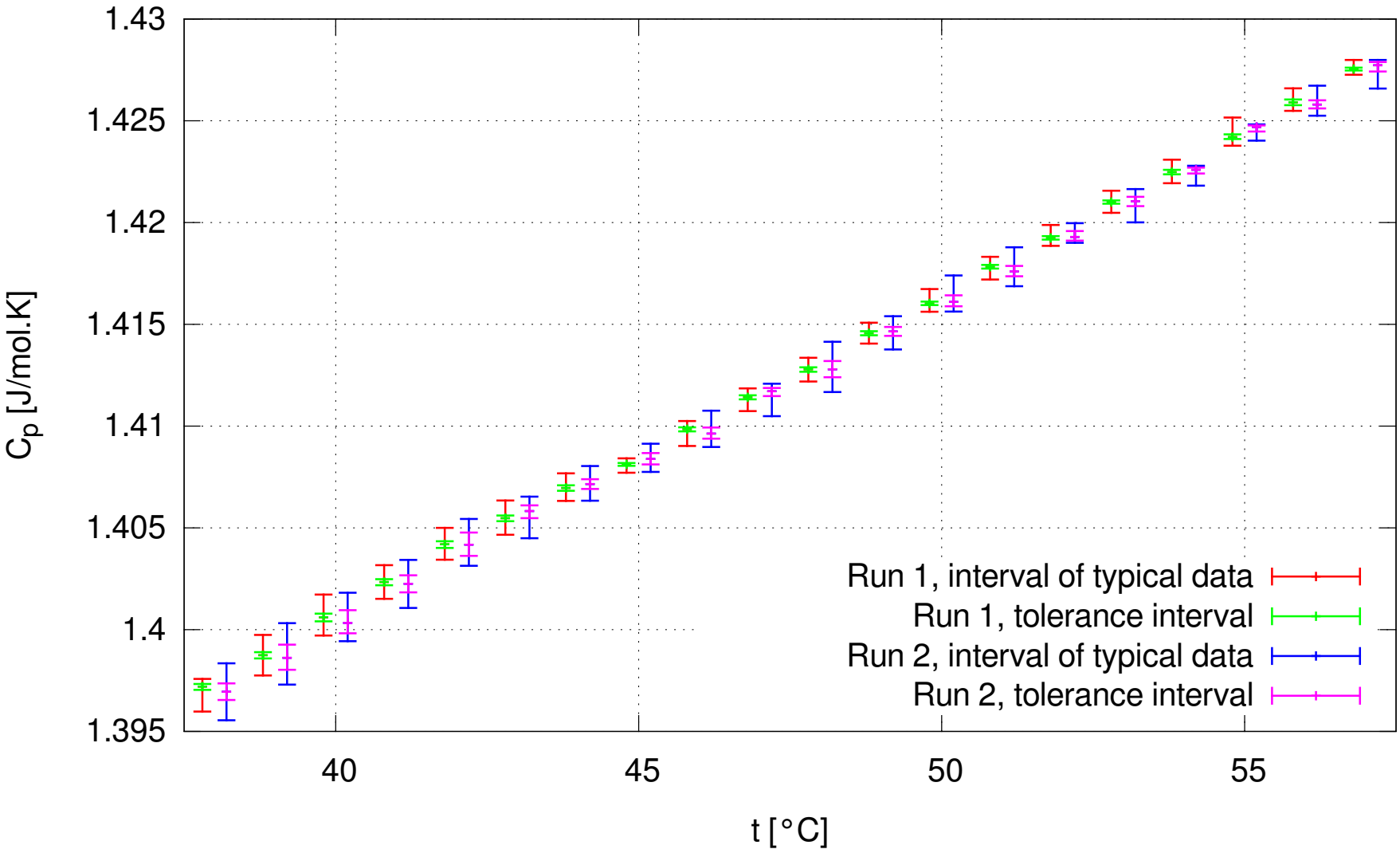








Comparison of two series of C_p measurement



8 values in each run, by mistake heating rate 0.2 K/min used

Conclusion

- Methods of data analysis by mathematical gnostics do not impose any kind of a distribution function *a priori*.
- Robustness is the inherent property of mathematical gnostics.
- The algorithms of mathematical gnostics are robust, can run *unattended* so that large number of data samples can be analyzed automatically.
- In many cases mathematical gnostics can extract additional information that is not obtainable by statistical methods.
- *It is important to understand that mathematics provides us with tools that can only extract information from data, nothing less, nothing more. The information must be interpreted in order to be useful.*
  See also – Nassim Taleb: *The Black Swan*.

विद्यैव सर्वधनम्

KNOWLEDGE IS THE GREATEST WEALTH

http://ttsm.icpf.cas.cz/team/wagner.shtml