

Enclosing vs. Non-enclosing Measurements in Interval Data Processing

Sergey P. Shary

Federal Research Center for Information and Computational Technologies,
Novosibirsk State University, Russia

January 14, 2022

I. Problem statement

**Data analysis and data processing
under interval uncertainty**

Observation and measurement errors

Data obtained from measurements and observations are almost always inexact ...

What uncertainty model do we accept?

A traditional choice is probabilistic error model

(C.F. Gauss, P.-S. Laplace, and so on):

errors of observations and measurements
are random variables of probability theory
with (more or less) known characteristics

Observation and measurement errors

Often, they are not modeled satisfactorily
by the methods of probability theory.

An alternative approach to errors analysis
is to specify two-sided bounds for the values of interest:

$$\underline{a} \leq \xi \leq \bar{a},$$

or, which is equivalent,

$$\xi \in [\underline{a}, \bar{a}].$$

Often, they are not modeled satisfactorily
by the methods of probability theory.

An alternative approach to errors analysis
is to specify two-sided bounds for the values of interest:

$$\underline{a} \leq \xi \leq \bar{a},$$

or, which is equivalent,

$$\xi \in [\underline{a}, \bar{a}].$$

L. Kantorovich — 1962

С.И. Спивак, А.П. Вошинин, Н.М. Оскорбин, С.И. Жилин, те, ...

J.P. Norton, M. Milanese, G. Belforte, L. Pronzato, E. Walter ...

The pioneering works in Interval Data Analysis

Kantorovich, L.V. (1962)

On some new approaches to numerical methods and processing observation data. *Siberian Mathematical Journal*, vol. 3, No. 5, pp. 701–709

Schweppe, F.S. (1968)

Recursive state estimation: unknown but bounded errors and system inputs. *IEEE Transactions on Automatic Control*, vol. 13 (1), pp. 22–28.

Leonid Kantorovich (1912–1986)



Outstanding mathematician,
Nobel prize winner
in economics (1975)
for linear programming, etc.

Л. В. КАНТРОВИЧ

**О НЕКОТОРЫХ НОВЫХ ПОДХОДАХ К ВЫЧИСЛИТЕЛЬНЫМ
МЕТОДАМ И ОБРАБОТКЕ НАБЛЮДЕНИЙ *****Введение**

Имевшие место сдвиги в развитии математики и вычислительных средств должны иметь следствием коренные изменения в технике, а возможно и теории численных методов и обработки наблюдений. В той или иной форме отдельные высказываемые ниже соображения встречались в литературе, но не разрабатывались систематически. В частности, мы считаем, что существенное значение имеют следующие моменты:

1. Большая ответственность за результаты расчетов, на которых сейчас нередко базируются решения, касающиеся сложных дорогостоящих объектов современной физики и техники, наличие больших не наблюдаемых этапов при машинных вычислениях повышают требования к надежности окончательных и промежуточных данных, получаемых в процессе применения численных методов и при обработке данных наблюдений. Это обуславливает систематический переход от построения приближенных значений и результатов, к получению точных двухсторонних границ для искомым величин или, если говорить о нечисловых величинах, областей расположения искомым и наблюдаемых величин; иначе говоря возникает задача возможно более точного описания расположения этих величин в соответствующих пространствах их значений. Идея рассуждений, приведенных в настоящем введении, и ставший в них

использованием вводимых нами в свое время мажорантного оператора для данного нелинейного. Мы не будем здесь подробно останавливаться на этом.

Отметим еще, что в случаях, когда обращение оператора невозможно или обратный оператор очень велик, например в окрестности собственного значения, целесообразно не строить границы областей, содержащих решения, а оценивать область расположения решения, используя технику линейного программирования.

Замечание. Во всех формулах для границ предполагалось, что действия над исходными границами производятся точно. Если эти действия производятся приближенно, например на машине, то формулы видоизменяются за счет дополнительного введения погрешностей этих действий. Не будем приводить записи формул для этого случая.

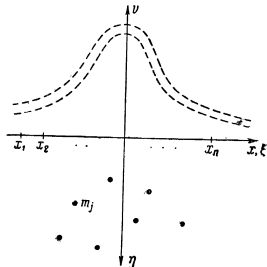
§ 3. Некоторые задачи прикладной математики

1. Задача обработки наблюдений. Обычно полученную в результате измерений избыточную систему уравнений обрабатывают по методу наименьших квадратов Гаусса. При этом происходит значительная потеря информации. По-видимому, в настоящее время более целесообразна другая техника. Уравнения, связывающие искомые величины, выписать с учетом погрешностей в форме неравенств

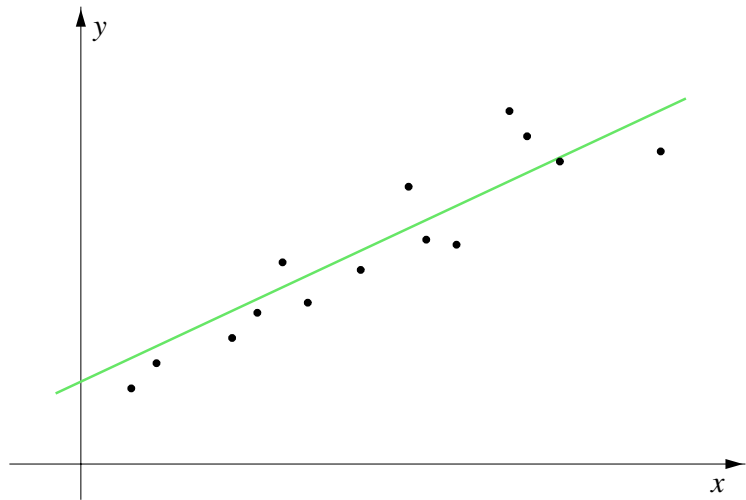
$$l_i - \delta \leq \sum_{k=1}^n c_{ik} x_k \leq l_i + \delta,$$

$$i = 1, \dots, m,$$

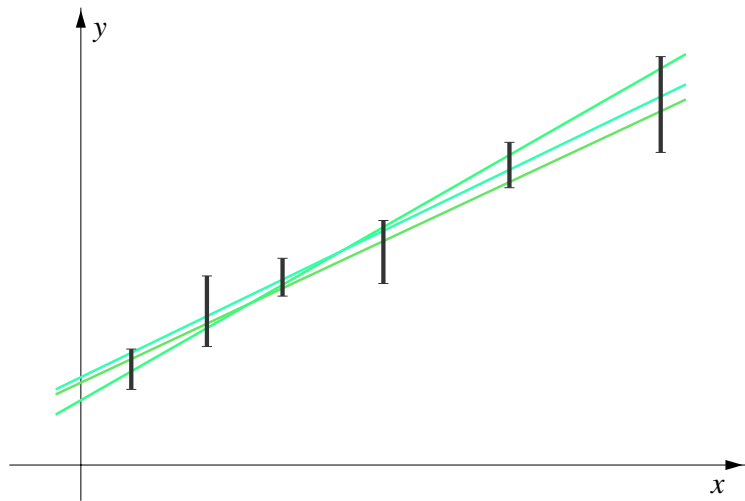
и разыскивать возможные границы для x_k методами линейного программирования.



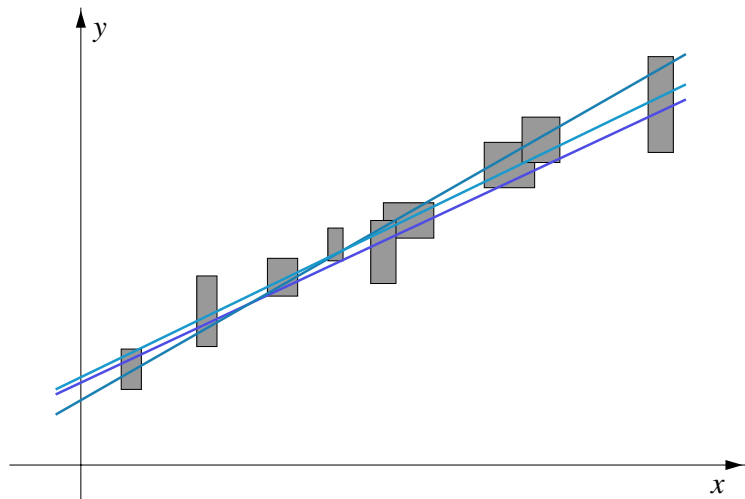
Data fitting problem



Interval data fitting problem by L. Kantorovich



Interval data fitting problem — general



Recursive State Estimation: Unknown but Bounded Errors and System Inputs

FRED C. SCHWEPPE, MEMBER, IEEE

Abstract—A method is discussed for estimating the state of a linear dynamic system using noisy observations, when the input to the dynamic system and the observation errors are completely unknown except for bounds on their magnitude or energy. The state estimate is actually a set in state space rather than a single vector. The optimum estimate is the smallest calculable set which contains the unknown system state, but it is usually impractical to calculate this set. A recursive algorithm is developed which calculates a time-varying ellipsoid in state space that always contains the system's true state. Unfortunately the algorithm is still unproven in the sense that its performance has not yet been evaluated. The algorithm is closely related in structure but not in performance to the algorithm obtained when the system inputs and observation errors are white Gaussian processes. The algorithm development is motivated by the problem of tracking an evasive target, but the results have wider applications.

I. INTRODUCTION

PROCESSING noisy observations of some function of a dynamic system's state is often necessary to provide an estimate of the system's current state. The nature of the algorithm to be used depends on the assumed structures of the dynamic system, the observa-

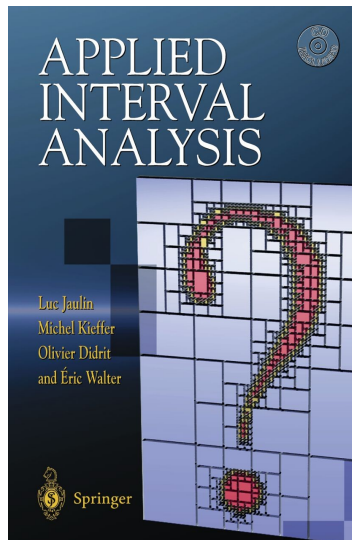
other situations. An assumption of unknown but bounded input to a dynamic system can be used simply as a way to reflect uncertainty about nature. Unknown but bounded observation errors are common; quantization errors are one example.

The basic idea of the estimation procedure is to combine knowledge of the system dynamics and bounds with the observations to specify a time-varying set in state space which always contains the true state of the system. Thus the actual estimate is a set in state space rather than a single vector. Specification of the smallest estimate set is conceptually straightforward but computationally impractical for most real problems. Therefore, an algorithm for calculating a bounding ellipsoid which always contains the state is developed. This ellipsoid is not the smallest possible estimate set, but it can be calculated recursively in real time.

The algorithm for the bounding ellipsoid estimate is computationally similar to the estimation algorithm (Kalman-Bucy) obtained when the input to the dy-

BOUNDING APPROACHES TO SYSTEM IDENTIFICATION

Edited by
MARIO MILANESE,
JOHN NORTON,
HÉLÈNE PIET-LAHANIER,
and
ÉRIC WALTER



II. A plot twist

Where to publish results on Interval Data Analysis? . . .



Advances in Data Analysis and Classification

Theory, Methods, and Applications in Data Science

Editors: M. Vichi; H.-H. Bock; W. Gaul; A. Okada; C. Weihs

- Presents research and applications on the extraction of knowable aspects from many types of data
- Topics include structural, quantitative, or statistical approaches for the analysis of data; advances in classification, clustering, and pattern recognition methods; strategies for modeling complex data and mining large data sets
- Shows how new domain-specific knowledge can be made available from data by skillful use of data analysis methods

The international journal *Advances in Data Analysis and Classification (ADAC)* is designed as a forum for high standard publications on research and applications concerning the extraction of knowable aspects from many types of data. It publishes articles on such topics as structural, quantitative, or statistical approaches for the analysis of data;

A response to my submission

Ref.: Ms. No. ADAC-D-18-00095

Weak and Strong Compatibility in Data Fitting Problems
under Interval Uncertainty

Dear Prof. Shary,

Reviewers' comments on your work have now been received.
They have advised against publication of your work.

...

Maurizio Vichi

Coordinating Editor

Advances in Data Analysis and Classification

A response to my submission

Reviewer #1:

...

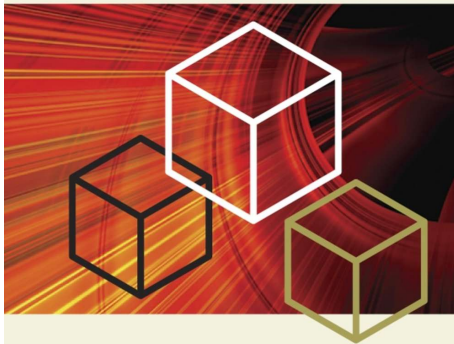
This paper faces the problem of uncertain, imprecision and variability, from a strictly numerical point of view, confuting the statistical theory of the regression model.

Moreover, the author ignores papers on regression of interval data, for instance:

- Billard L., Diday E. (2000) Regression analysis for interval-valued data. In: Kiers, Rasson, Groenen, Schader (eds) Data Analysis, Classification, and Related Methods. Springer, Berlin, Heidelberg, pp. 369-374

...

WILEY SERIES IN COMPUTATIONAL STATISTICS



Lynne Billard and Edwin Diday

SYMBOLIC DATA ANALYSIS

CONCEPTUAL STATISTICS AND DATA MINING

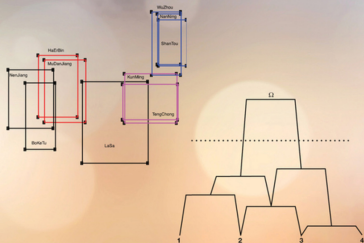
 WILEY

Contents

1	Introduction	1
	References	6
2	Symbolic Data	7
2.1	Symbolic and Classical Data	8
2.1.1	Types of data	8
2.1.2	Dependencies in the data	30
2.2	Categories, Concepts, and Symbolic Objects	34
2.2.1	Preliminaries	34
2.2.2	Descriptions, assertions, extents	35
2.2.3	Concepts of concepts	45
2.2.4	Some philosophical aspects	50
2.2.5	Fuzzy, imprecise, and conjunctive data	53
2.3	Comparison of Symbolic and Classical Analyses	56
	Exercises	66
	References	67
3	Basic Descriptive Statistics: One Variate	69
3.1	Some Preliminaries	69
3.2	Multi-Valued Variables	73
3.3	Interval-Valued Variables	77
3.4	Modal Multi-Valued Variables	92
3.5	Modal Interval-Valued Variables	96
	Exercises	105
	References	106
4	Descriptive Statistics: Two or More Variates	107
4.1	Multi-Valued Variables	107
4.2	Interval-Valued Variables	109
4.3	Modal Multi-Valued Variables	113

4.4	Modal Interval-Valued Variables	116
4.5	Baseball Interval-Valued Dataset	123
4.5.1	The data: actual and virtual datasets	123
4.5.2	Joint histograms	127
4.5.3	Guiding principles	130
4.6	Measures of Dependence	131
4.6.1	Moment dependence	131
4.6.2	Spearman's rho and copulas	138
	Exercises	143
	References	144
5	Principal Component Analysis	145
5.1	Vertices Method	145
5.2	Centers Method	172
5.3	Comparison of the Methods	180
	Exercises	185
	References	186
6	Regression Analysis	189
6.1	Classical Multiple Regression Model	189
6.2	Multi-Valued Variables	192
6.2.1	Single dependent variable	192
6.2.2	Multi-valued dependent variable	195
6.3	Interval-Valued Variables	198
6.4	Histogram-Valued Variables	202
6.5	Taxonomy Variables	204
6.6	Hierarchical Variables	214
	Exercises	227
	References	229
7	Cluster Analysis	231
7.1	Dissimilarity and Distance Measures	231
7.1.1	Basic definitions	231
7.1.2	Multi-valued variables	237
7.1.3	Interval-valued variables	241
7.1.4	Mixed-valued variables	248
7.2	Clustering Structures	249
7.2.1	Types of clusters: definitions	249
7.2.2	Construction of clusters: building algorithms	256
7.3	Partitions	257
7.4	Hierarchy–Divisive Clustering	259
7.4.1	Some basics	259
7.4.2	Multi-valued variables	262
7.4.3	Interval-valued variables	265

Wiley Series in Computational Statistics



Clustering Methodology for **Symbolic Data**

Lynne Billard | Edwin Diday

WILEY

Symbolic Data Analysis and the SODAS Software

Editors Edwin Diday and Monique Noirhomme-Fraiture

 WILEY

My article has eventually been published

Advances in Data Science and Adaptive Analysis
Vol. 12, No. 1 (2020) 2050002 (84 pages)
© World Scientific Publishing Company
DOI: [10.1142/S2424922X20500023](https://doi.org/10.1142/S2424922X20500023)



Weak and Strong Compatibility in Data Fitting Problems Under Interval Uncertainty

Sergey P. Shary

*Federal Research Center for Information and Computational Technologies
Academician M. A. Lavrentiev Avenue, 6,
630090 Novosibirsk, Russia
Novosibirsk State University,
1, Pirogova Str., 630090 Novosibirsk, Russia
shary@ict.nsc.ru*

Received 16 February 2019

Accepted 6 March 2019

Published 6 July 2020


but the questions remained ...


Symbolic Data Analysis: Conceptual Statistics and Data Mining


December 2006


DOI: [10.1002/9780470090183](https://doi.org/10.1002/9780470090183)

Source - [DAI](#)

 Edwin Diday · Lynne Billard

Citations  378

Recommendations  0 new 1

Reads  0 new 547

[See details](#)

Overview

Stats

Comments (2)

Citations (378)

...

Request full-text

Share 

More 

Comments (2)



[Sergey Shary](#)
added a comment

Apr 25, 2020 

One of the main particular types of 'symbolic data' studied in this book is interval-valued data. But the processing of interval data has been considered long before L.Billard and E.Diday, since the 60s of the last century, and a large number of studies have been devoted to it. They are based, as a rule, on the technique of interval analysis and supported by good computational methods. By the time L.Billard and E.Diday published their first works, these interval methods were summarized, for example, in the book

M.Milanese, J.Norton, H.Piet-Lahanier, E.Walter (Eds.), Bounding Approaches to System Identification, Plenum Press, New York, 1996. DOI: [10.1007/978-1-4757-9545-5](https://doi.org/10.1007/978-1-4757-9545-5)

What is the relationship between the approaches and results of the Symbolic Data Analysis by Billard & Diday (and their followers) and the results outlined in the book cited above? What methods are preferable to apply in various practical situations?

[Reply](#) [Share](#)

M.Milanesi, [J.Norton](#), H.Piet-Lahanier, [E.Walter](#) (Eds.), Bounding Approaches to System Identification, Plenum Press, New York, 1996. DOI: 10.1007/978-1-4757-9545-5

What is the relationship between the approaches and results of the Symbolic Data Analysis by Billard & Diday (and their followers) and the results outlined in the book cited above? What methods are preferable to apply in various practical situations?

[Reply](#) [Share](#)



[Sergey Shary](#)
added a comment

Apr 25, 2020 ▾

In 2001, the book

L.Jaulin, M.Kieffer, O.Didrit, [E.Walter](#), Applied Interval Analysis, Springer, London, 2001, was published, devoted to applications of interval analysis methods in various practical disciplines. Chapter 6 of this book is called "Estimation" and describes parameter estimation and state estimation in technical systems, which is, essentially, the same data fitting problem that the regression analysis deals with, although called somewhat differently.

Again, a question similar to that from my previous comment:

How do the methods of Symbolic Data Analysis elaborated by L.Billard and E.Diday relate to the methods from the book L.Jaulin, M.Kieffer, O.Didrit, and [E.Walter](#)?

[Reply](#) [Share](#)

Since my above posts with natural questions
have remained unanswered at ResearchGate for almost two years,
let us answer these questions ourselves . . .

III. A short survey of Symbolic Data Analysis in interval data fitting

Symbolic Data Analysis

From Wikipedia, the free encyclopedia

Symbolic data analysis (SDA) is an extension of standard data analysis where symbolic data tables are used as input and symbolic objects are made output as a result.

The data units are called symbolic since they are more complex than standard ones, as they not only contain values or categories, but also include internal variation and structure.

Symbolic Data Analysis

From Wikipedia, the free encyclopedia

Symbolic data analysis (SDA) is an extension of standard data analysis where symbolic data tables are used as input and symbolic objects are made output as a result.

The data units are called symbolic since they are more complex than standard ones, as they not only contain values or categories, but also include internal variation and structure.

SDA is based on four spaces: the space of individuals, the space of concepts, the space of descriptions, and the space of symbolic objects.

The space of descriptions models individuals, while the space of symbolic objects models concepts.

- BILLARD L., DIDAY E.
Symbolic regression analysis. In: Jajuga K., Sokołowski A., Bock H. (eds) Classification, Clustering, and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg, 2002, pp. 281–288.
- EUFRÁSIO DE A.LIMA NETO, FRANCISCO DE A.T.DE CARVALHO
Constrained linear regression models for symbolic interval-valued variables // Computational Statistics & Data Analysis, vol. 54 (2010), issue 2, pp. 333–347.
- YAN SUN, CHUNYANG LI
Linear regression for interval-valued data: a new and comprehensive model. — 2015, arXiv paper No. 1401.1831.
- SHAILY KABIR, CHRISTIAN WAGNER, ZACK ELLERBY
Towards handling uncertainty-at-source in AI – a review and next steps for interval regression. — 2021, arXiv paper No. 2104.07245.

Symbolic Data Analysis: Taking Variability in Data into Account

Regression for Symbolic Data

Sónia Dias

I.P. Viana do Castelo & LIAAD-INESC TEC, Univ. Porto

Paula Brito

Fac. Economia & LIAAD-INESC TEC, Univ. Porto

ECI 2015 - Buenos Aires





Paula Brito

Full Name:

Maria Paula de Pinho de Brito Duarte Silva

Associate Professor in Statistics and Data Analysis at the [Faculty of Economics \(Group of Mathematics and Information Systems\)](#) of the [University of Porto](#).
Member of the Laboratory in Artificial Intelligence and Decision Support ([LIADD - INFESC TEC](#)) of the [University of Porto](#).

[v Scientific Interests](#)

[v Contact](#)

[v Academic Degrees](#)

[v Publications](#)

[Ø Papers in international scientific periodicals](#)

[Ø Refereed chapters in books](#)

[Ø Books \(editor\)](#)

[Ø Special Issues \(editor\)](#)

[Ø Refereed papers in conference proceedings](#)

[Ø Working Papers](#)

[Ø Book Chapters](#)

[v Software](#)

Symbolic Data Analysis: Taking Variability in Data into Account

Methods for the Analysis of Symbolic Data

Regression

Linear Regression for interval-valued variables

Linear Regression for histogram-valued variables

Linear Regression for interval-valued variables

State-of-the-art

- **METHODS BASED IN SYMBOLIC COVARIANCE DEFINITIONS** (Billard and Diday, 2000;2006; Xu, 2010)
- **MINMAX METHOD** (Billard and Diday, 2002)
- **CENTER AND RANGE METHOD** (Lima Neto and De Carvalho,2008)
- **CENTER AND RANGE LEAST ABSOLUTE DEVIATION REGRESSION METHOD** (Maia and Carvalho, 2008)
- **CONSTRAINED CENTER AND RANGE METHOD** (Lima Neto and De Carvalho, 2010)
- **LASSO IR METHOD** (Giordani, 2014)
- **BIVARIANTE SYMBOLIC REGRESSION MODELS** (Lima Neto *et al*,2011)
- LINEAR REGRESSION MODELS FOR SYMBOLIC INTERVAL DATA USING PSO ALGORITHM** (Yang *et al*, 2011)
- **MONTE CARLO METHOD** (Ahn *et al*,2012)
- **RADIAL BASIS FUNCTION NETWORKS** (Su *et al*, 2012)
- **COPULA INTERVAL REGRESSION METHOD** (Neto *et al*, 2012)
- **INTERVAL DISTRIBUTIONAL MODEL** (Dias and Brito, in study)

Linear Regression for interval-valued variables

The Center Method (CM)

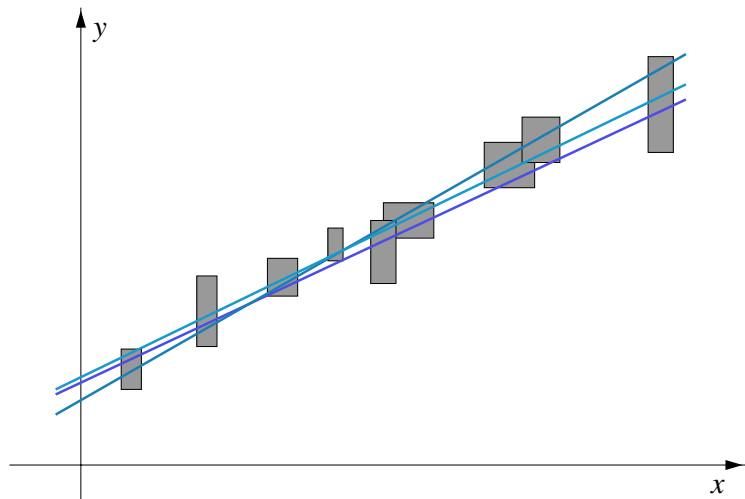
Billard and Diday, 2000. *Regression analysis for interval-valued data*. Proceedings of IFCS'00, pp.369-374. Springer.

Linear regression relation: $c_{Y(j)} = b_0 + b_1 c_{X_1(j)} + \dots + b_p c_{X_p(j)} + e^c(j)$

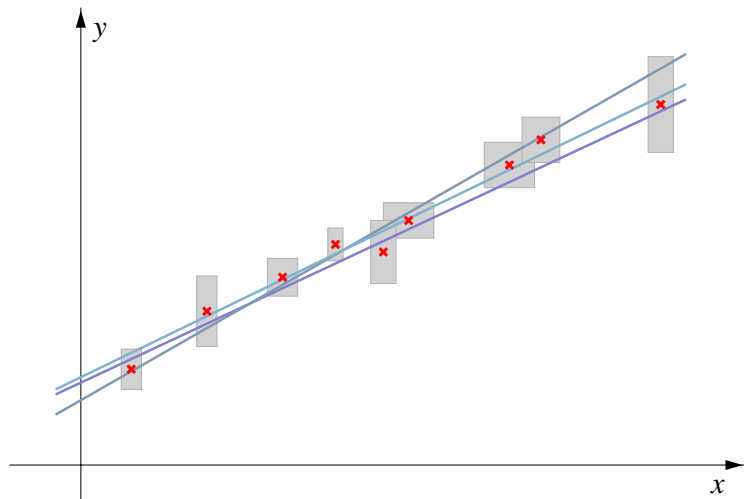
Prediction of the intervals: $I_{\hat{Y}(j)} = [L_{\hat{Y}(j)}, \bar{I}_{\hat{Y}(j)}]$ with $L_{\hat{Y}(j)} = \min \left\{ b_0 + \sum_{k=1}^p b_k L_{X_k(j)}, b_0 + \sum_{k=1}^p b_k \bar{I}_{X_k(j)} \right\}$
 $\bar{I}_{\hat{Y}(j)} = \max \left\{ b_0 + \sum_{k=1}^p b_k L_{X_k(j)}, b_0 + \sum_{k=1}^p b_k \bar{I}_{X_k(j)} \right\}$

- The coefficients of the model are estimated by applying the classical model to the mid-point of the intervals;
- Estimates separately the bounds of the interval;
- To write the estimated interval we have to consider the lower value for the lower bound and higher for the upper bound of the interval;
- The estimation of the parameters may be obtained by an adaptation of the solution obtained by the Least Square estimation method for the classical linear model, where symbolic definitions of variance and covariance are used;
- Descriptive linear regression model;
- Available in R Package: iRegression 1.2.

Interval data fitting problem



Interval data fitting problem — an SDA approach



Linear Regression for interval-valued variables

Min Max Method (MinMax)

Billard and Diday, 2002. *Symbolic regression analysis*. Proceedings of IFCS'02, pp.281-288. Springer.

$$\text{Linear regression relation : } \begin{cases} \underline{I}_{Y(j)} = b_0^L + b_1^L \underline{I}_{X_1(j)} + \dots + b_p^L \underline{I}_{X_p(j)} + \underline{e}(j) \\ \bar{I}_{Y(j)} = b_0^U + b_1^U \bar{I}_{X_1(j)} + \dots + b_p^U \bar{I}_{X_p(j)} + \bar{e}(j) \end{cases}$$

$$\text{Prediction of the intervals: } I_{\hat{Y}(j)} = [\underline{I}_{\hat{Y}(j)}, \bar{I}_{\hat{Y}(j)}] \text{ with } \begin{cases} \underline{I}_{\hat{Y}(j)} = \min \left\{ b_0^L + \sum_{k=1}^p b_k^L \underline{I}_{X_k(j)}, b_0^U + \sum_{k=1}^p b_k^U \bar{I}_{X_k(j)} \right\} \\ \bar{I}_{\hat{Y}(j)} = \max \left\{ b_0^L + \sum_{k=1}^p b_k^L \underline{I}_{X_k(j)}, b_0^U + \sum_{k=1}^p b_k^U \bar{I}_{X_k(j)} \right\} \end{cases}$$

- Requires the adjustment of two linear regression models, for the lower and upper bounds of the interval;
- The coefficients of the model are estimated by applying the classical model to the lower and upper bounds of the interval;
- The estimated value for the upper bound of the interval may be smaller than the lower. This can happen if there are negative coefficients in the model;
- Descriptive linear regression model;
- Available in R Package: iRegression 1.2.

Linear Regression for interval-valued variables

The Center and Range Method (CRM)

Lima Neto and de Carvalho, 2008. *Center and Range method for fitting a linear regression model to symbolic interval data*. Computational Statistics & Data Analysis 52 (3), 1500-1515.

Linear regression relation :
$$\begin{cases} c_{Y(j)} = b_0^c + b_1^c c_{X_1(j)} + \dots + b_p^c c_{X_p(j)} + e^c(j) \\ r_{Y(j)} = b_0^r + b_1^r r_{X_1(j)} + \dots + b_p^r r_{X_p(j)} + e^r(j) \end{cases}$$

Prediction of the intervals:
$$I_{\hat{Y}(j)} = [L_{\hat{Y}(j)}, \bar{I}_{\hat{Y}(j)}]$$
 with
$$L_{\hat{Y}(j)} = \min\{c_{\hat{Y}(j)} - r_{\hat{Y}(j)}, c_{\hat{Y}(j)} + r_{\hat{Y}(j)}\}$$
$$\bar{I}_{\hat{Y}(j)} = \max\{c_{\hat{Y}(j)} - r_{\hat{Y}(j)}, c_{\hat{Y}(j)} + r_{\hat{Y}(j)}\}$$

- Requires the adjustment of two linear regression models, for the mid-point and half range of the interval;
- The coefficients of the model are estimated by applying the classical model to the mid-point and half range of the interval;
- The estimated value for the range of the interval may be negative. This can happen if there are negative coefficients in the model that estimates the half range;
- Descriptive linear regression model;
- Available in R Package: iRegression 1.2.

Linear Regression for interval-valued variables

The Constrained Centre and Range Method (CCRM)

Lima Neto and de Carvalho, 2010. *Constrained linear regression models for symbolic interval-valued variables*. Computational Statistics & Data Analysis 54 (2), 333-347.

Linear regression relation :

$$\begin{cases} c_{Y(j)} = b_0^c + b_1^c c_{X_1(j)} + \dots + b_p^c c_{X_p(j)} + e^c(j) \\ r_{Y(j)} = b_0^r + b_1^r r_{X_1(j)} + \dots + b_p^r r_{X_p(j)} + e^r(j) \end{cases}$$

with $b_k^r \geq 0$

Prediction of the intervals: $I_{\hat{Y}(j)} = [c_{\hat{Y}(j)} - r_{\hat{Y}(j)} \cdot c_{\hat{Y}(j)} + r_{\hat{Y}(j)}]$

- The mid-points and half ranges of the intervals are estimated independently;
- The coefficients of the centers model are estimated by applying the classical model to the mid-points of the intervals;
- The coefficients of the half ranges model are estimated using the Lawson and Hanson's algorithm (Lawson and Hanson, 1995).
- Because of the restriction imposed, the linear relation between the half range of the intervals has to be always positive;
- Descriptive linear regression model;
- Available in R Package: iRegression 1.2.

III. Understanding and analysis

- its approach to processing interval data is “perpendicular”
to that of the classical Interval Data Analysis.

Does this even make sense? . . .

- its approach to processing interval data is “perpendicular”
to that of the classical Interval Data Analysis.

Does this even make sense? ...

My answer is “YES” ...

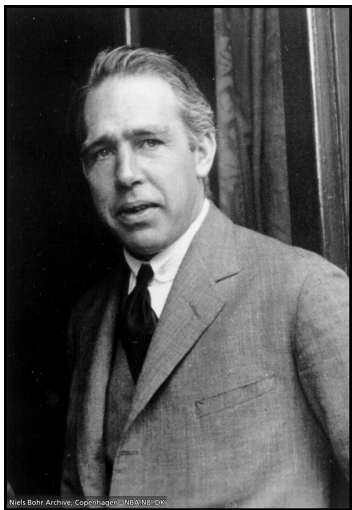
It makes, to some extent

Correspondence principle

The correspondence principle in the methodology of science is the statement that *any new scientific theory must include the old theory and its results as a particular limiting case.*

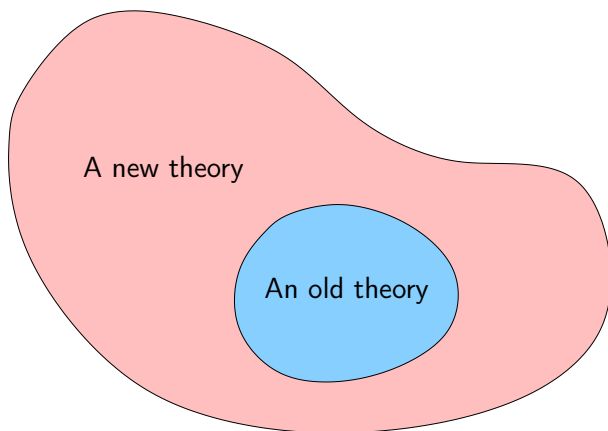
The conclusions of the new theories in the area
where the old “classical” theory was valid,
pass into the conclusions of the classical theory.

The correspondence principle was proposed by N. Bohr in 1913–1918 to explain the mutual relationship between the emerging quantum mechanics and the traditional physics of macroobjects.



Niels Bohr (1885–1962)

Correspondence principle



The correspondence principle can serve as a means of constructing and correcting new theoretical systems, new systems of concepts.

Correspondence principle

In the XX century, it was necessary to ensure the “continuous gluing” of the new physics of micro-objects with classical physics.

The latter perfectly describes and explains a huge number of phenomena around us and cannot be simply rejected on the grounds that it is just “old”, “unfashionable”, etc.

Correspondence principle

In the XX century, it was necessary to ensure the “continuous gluing” of the new physics of micro-objects with classical physics.

The latter perfectly describes and explains a huge number of phenomena around us and cannot be simply rejected on the grounds that it is just “old”, “unfashionable”, etc.

Another popular example

In the special theory of relativity for small velocities, we obtain the same equations of motion as in Newton’s classical mechanics.

Correspondence principle in Data Analysis for interval data

The correspondence principle has a broader methodological meaning, being applicable not only to physics and natural sciences, but also to mathematics, computer science, etc.

As applied to our situation, when processing interval data, the correspondence principle requires that

*good and reasonable interval methods in the limit,
when the width of the data intervals tends to zero,
switch to some methods for processing point data,*

since real numbers are the limiting case of intervals.

Correspondence principle in Data Analysis for interval data

We can use the correspondence principle as a tool for checking the adequacy of the concepts and methods of interval data analysis, which allows us to cut off the “unreasonable” methods.

Correspondence principle in Data Analysis for interval data

We can use the correspondence principle as a tool for checking the adequacy of the concepts and methods of interval data analysis, which allows us to cut off the “unreasonable” methods.

- “Symbolic data analysis” for interval data processing is in agreement with Correspondence principle.
- Not all interval data processing methods proposed so far are consistent with Correspondence principle.

IV. A comparison

Symbolic Data Analysis versus Interval Data Analysis

What is the difference between

Symbolic Data Analysis applied to intervals

and classical Interval Data Analysis ?

Symbolic Data Analysis vs. Interval Data Analysis

The basis of symbolic data analysis applied to intervals is
the traditional techniques of probabilistic mathematical statistics,
the least squares method, etc.

The basis of the classical interval approach is
the idea of approximating interval data
taking into account specific nature of intervals.

Usually, no probability at all.

Symbolic Data Analysis vs. Interval Data Analysis

The basis of symbolic data analysis applied to intervals is
the traditional techniques of probabilistic mathematical statistics,
the least squares method, etc.

The basis of the classical interval approach is
the idea of approximating interval data
taking into account specific nature of intervals.

Usually, no probability at all.

But this is not the characteristic distinction!

Symbolic Data Analysis vs. Interval Data Analysis

In Symbolic Data Analysis, the main assumption

$$\underline{a} \leq \xi \leq \bar{a},$$

or, which is equivalent,

$$\xi \in [\underline{a}, \bar{a}]$$

— is not valid

Symbolic Data Analysis vs. Interval Data Analysis

In Symbolic Data Analysis, the main assumption

$$\underline{a} \leq \xi \leq \bar{a},$$

or, which is equivalent,

$$\xi \in [\underline{a}, \bar{a}]$$

— is not valid

Do data intervals make sense then?

Symbolic Data Analysis vs. Interval Data Analysis

In Symbolic Data Analysis, the main assumption

$$\underline{a} \leq \xi \leq \bar{a},$$

or, which is equivalent,

$$\xi \in [\underline{a}, \bar{a}]$$

— is not valid

Do data intervals make sense then?

Yes, of course, if nothing else is given to us.

Symbolic Data Analysis vs. Interval Data Analysis

If true values are not included in measurement uncertainty boxes,
then what is left for us?

The only criterion remains to determine how good or bad our result is.

Symbolic Data Analysis vs. Interval Data Analysis

If true values are not included in measurement uncertainty boxes,
then what is left for us?

The only criterion remains to determine how good or bad our result is.

This is its distance to these boxes,
deviation from the uncertainty boxes,
which we perceive as some kind of integral objects.

Enclosing vs. non-enclosing interval measurements

Definition

The true value of a physical quantity is the value that ideally reflects the considered quantity or phenomenon within the framework of the model (theory) we have adopted to describe it.

Enclosing vs. non-enclosing interval measurements

Definition

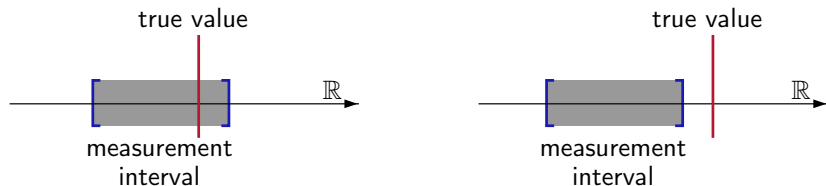
The true value of a physical quantity is the value that ideally reflects the considered quantity or phenomenon within the framework of the model (theory) we have adopted to describe it.

Definition

Enclosing measurement (covering measurement) is an interval measurement result that is guaranteed to contain the true value of the measured quantity.

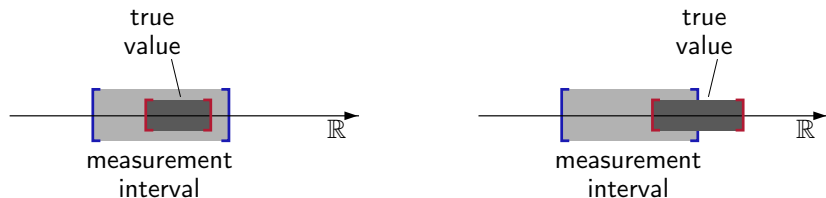
A measurement for which it cannot be claimed that it contains the true value of the measured quantity will be called *non-enclosing* (non-covering).

Enclosing vs. non-enclosing interval measurements



Enclosing (left) and non-enclosing (right) measurements
of a point true value for a physical quantity.

Enclosing vs. non-enclosing interval measurements



Enclosing (left) and non-enclosing (right) measurements
of an interval true value for a physical quantity.

Enclosing vs. non-enclosing interval samples

Definition

A sample of interval measurements will be called *enclosing sample* (covering sample) if the dominant part (majority, etc.) of its measurements are enclosing (covering). A sample, is called *non-enclosing* (non-covering) if the dominant portion of its measurements are non-enclosing (non-covering).

Enclosing vs. non-enclosing interval samples

Definition

A sample of interval measurements will be called *enclosing sample* (covering sample) if the dominant part (majority, etc.) of its measurements are enclosing (covering). A sample, is called *non-enclosing* (non-covering) if the dominant portion of its measurements are non-enclosing (non-covering).

The simplest way:

- a set of enclosing measurement is an “enclosing sample”,
- a “non-enclosing sample” has at least one non-enclosing measurement.

That would be purely theoretical, not taking into account the actual practice of measurements, where errors and outliers are inherent in data.

Interval Data Analysis

is the analysis of enclosing (covering) interval data.

Symbolic Data Analysis

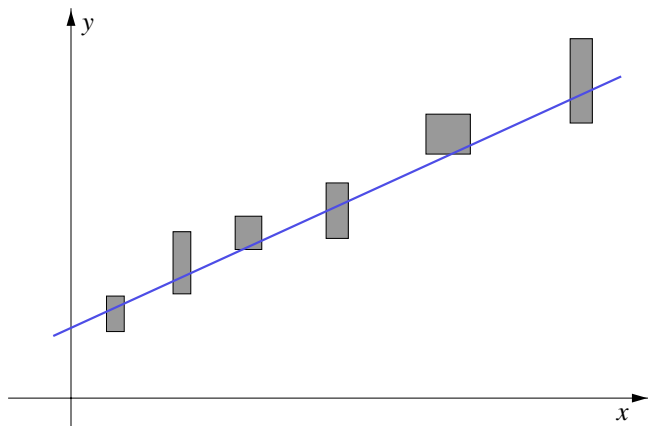
is the analysis of non-enclosing (non-covering) interval data.

They complement each other.

V. Further development of Symbolic Data Analysis for intervals

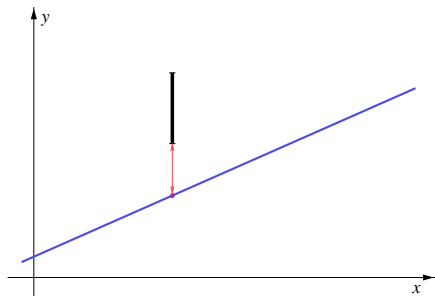
(jointly with Maxim Zvyagin)

Our contribution to Symbolic Data Analysis



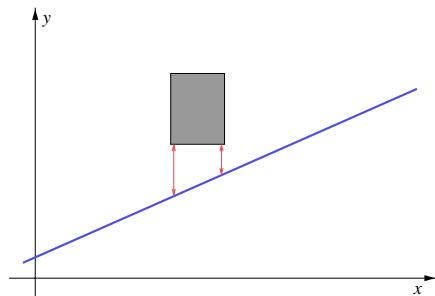
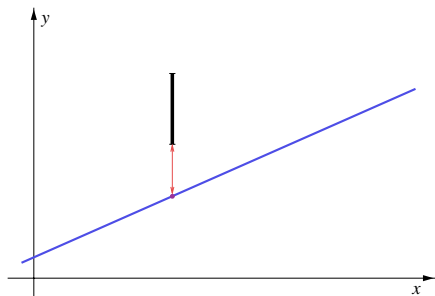
- 1) Assign a distance between the uncertainty boxes and the line.
- 2) Aggregate the individual distances into a total one.
- 3) Minimize the total distance.

Our contribution to Symbolic Data Analysis



Assign a distance from the uncertainty interval to the regression line

Our contribution to Symbolic Data Analysis



Assign a distance from the uncertainty interval to the regression line

Assign a distance from the uncertainty box to the regression line

Our contribution to Symbolic Data Analysis

Deviation of the regression line from interval data

= a norm of the vector made up
of separate deviations for uncertainty boxes

Our contribution to Symbolic Data Analysis

Deviation of the regression line from interval data

= a norm of the vector made up
of separate deviations for uncertainty boxes

We take Chebyshev norm (maximum norm) in \mathbb{R}^n

$$\|z\| = \max_{1 \leq i \leq n} |z_i|$$

Then we minimize the total deviation

by developed non-smooth optimization methods ...

We call the technique described above Simple Interval Approximation (SIA)

An example

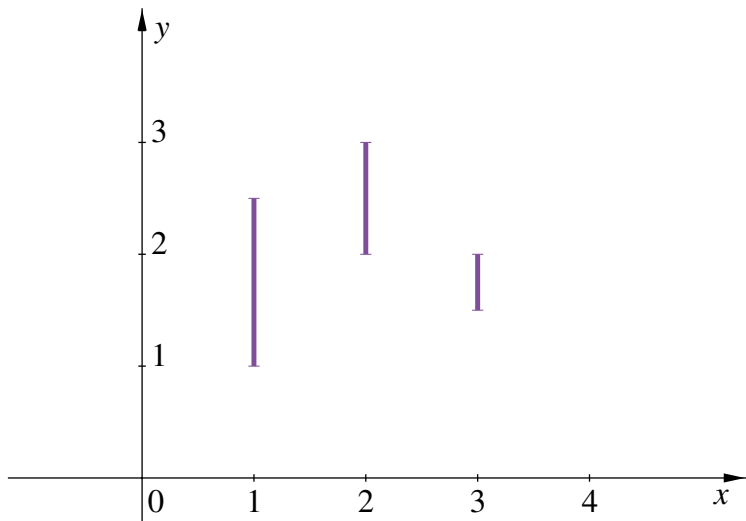
Let us construct a linear function

$$y = \beta_1 x_1 + \beta_0$$

from the data

x	1	2	3
y	[1, 2.5]	[2, 3]	[1.5, 2]

An example



An example: the case of enclosing data

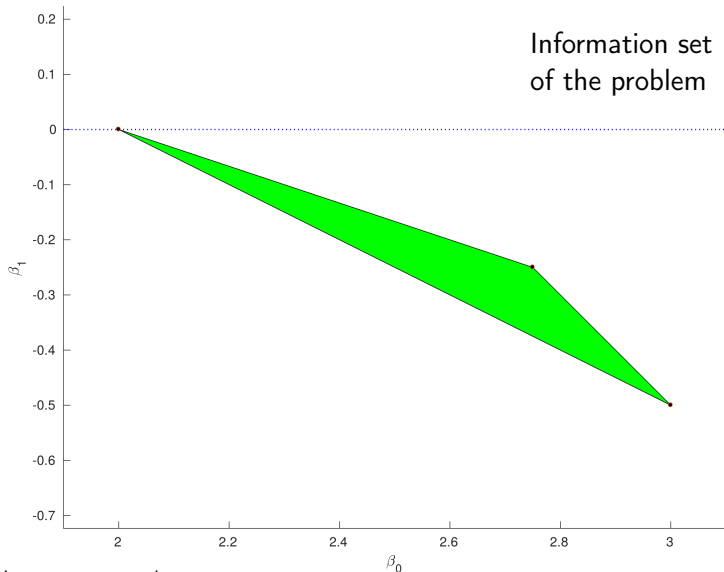
Substituting the data

into the expression of the function,

we get the interval equation system

$$\begin{cases} \beta_0 + \beta_1 = [1, 2.5], \\ \beta_0 + 2\beta_1 = [2, 3], \\ \beta_0 + 3\beta_1 = [1.5, 2] \end{cases}$$

An example: the case of enclosing data



Plotted by IntLinInc2D package

An example: the case of enclosing data

Using e. g. the Maximum Compatibility Method,

we get “the best fit” linear function

$$y = -0.25x + 2.625$$

Sergey P. Shary, Maximum consistency method for data fitting under interval uncertainty // Journal of Global Optimization, vol. 66 (2016), pp. 111–126

Sergey P. Shary, Weak and strong compatibility in data fitting problems under interval uncertainty // Advances in Data Science and Adaptive Analysis, vol. 12 (2020), No. 1, 2050002

An example: the case of non-enclosing data

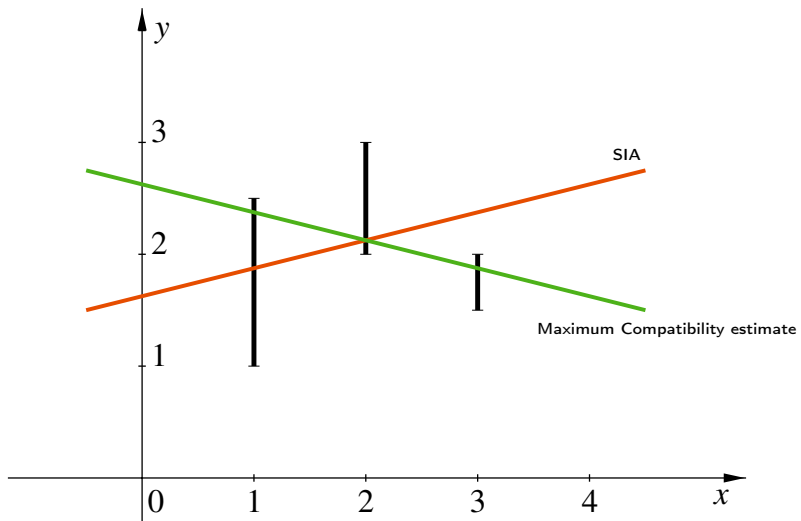
But the SIA method with the Chebyshev norm gives as the best,
from his point of view, the linear function

$$y = 0.25x + 1.625$$

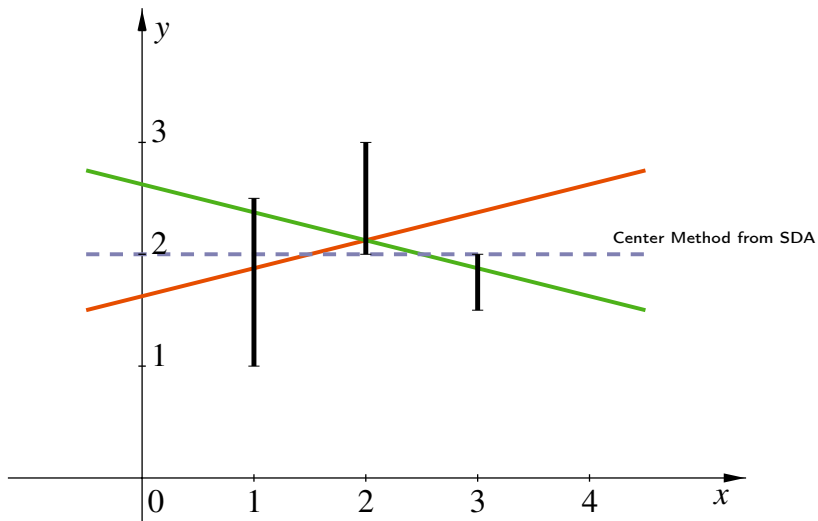
This differs sharply from the result of the Maximum Compatibility Method.

Its parameters do not lie in the information set ...

An example



An example



I appreciate your attention