SCAN'2012
15'th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic and Verified Numerical Computations
Novosibirsk, September 23–29, 2012

On Methodological Foundations of Interval Analysis of Empirical Dependencies

Nikolay M. Oskorbin   Sergei I. Zhilin
Altai State University, Barnaul, Russia

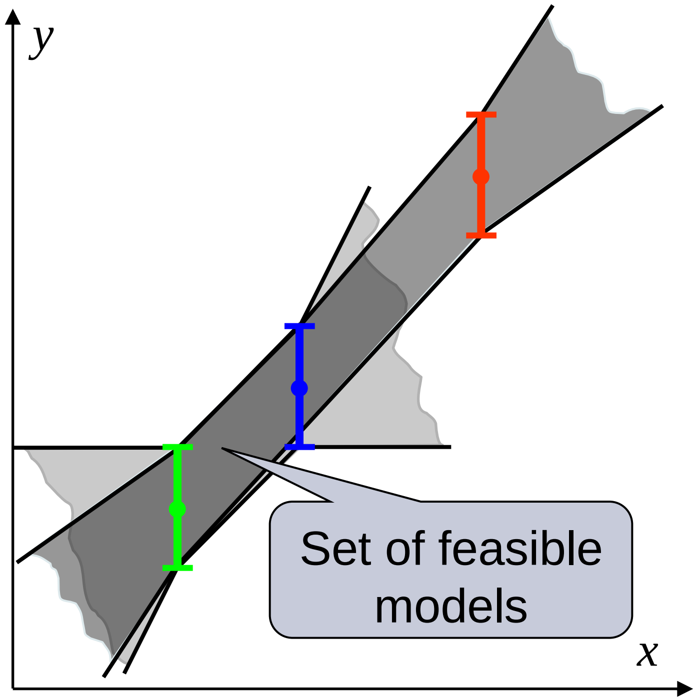
Outline

- Linear regression under interval error
- Validity of prior assumptions
- Revealing data inconsistency
- Basic principles and general scheme of empirical dependencies building
- Conclusions

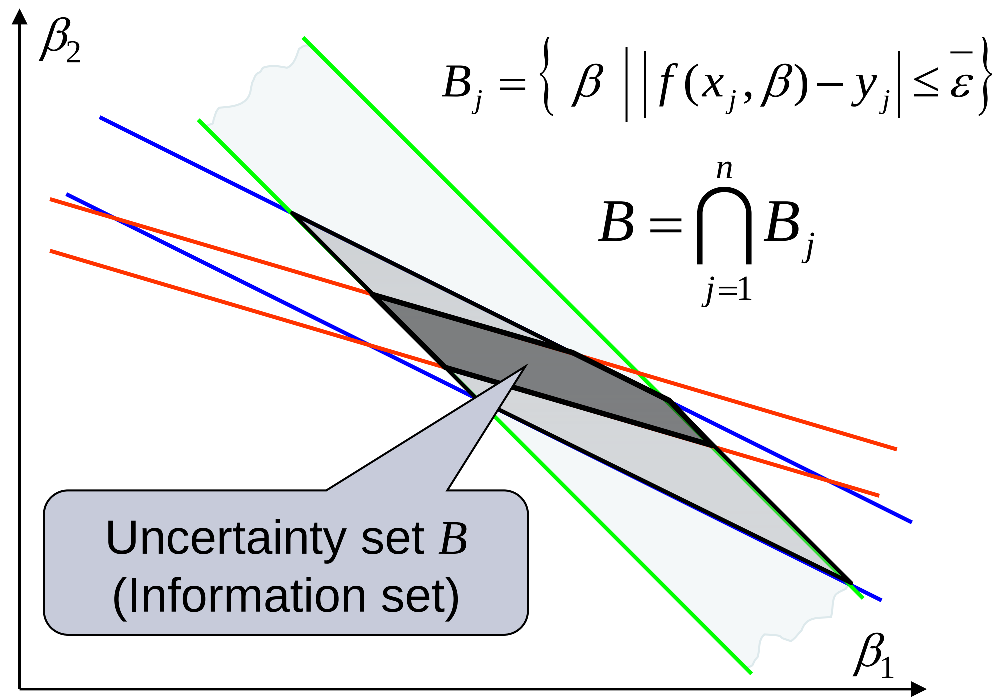
Linear Regression under Interval Error

- Building model $y = \beta_1 + \beta_2 x + \varepsilon, \quad \varepsilon \in [-\bar{\varepsilon}, \bar{\varepsilon}]$

In (x, y) domain



In (β_1, β_2) domain



Linear Regression under Interval Error

- Problems stated with respect to uncertainty set B
  - Prediction of the response value for fixed values of input variables
    - Interval estimates of y

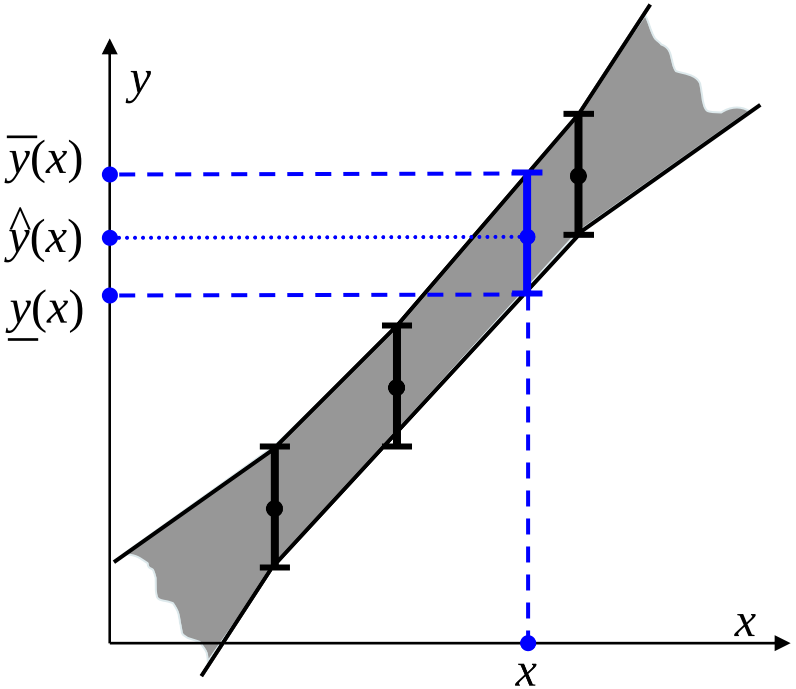
$$\mathbf{y}(x) = \left[\underline{y}(x), \overline{y}(x)\right]:$$

$$\underline{y}(x) = \min_{\beta \in B} \beta^T x,$$

$$\overline{y}(x) = \max_{\beta \in B} \beta^T x,$$

    - Point estimates of y

$$\hat{y}(x) = \tfrac{1}{2}\left(\underline{y}(x) + \overline{y}(x)\right)$$



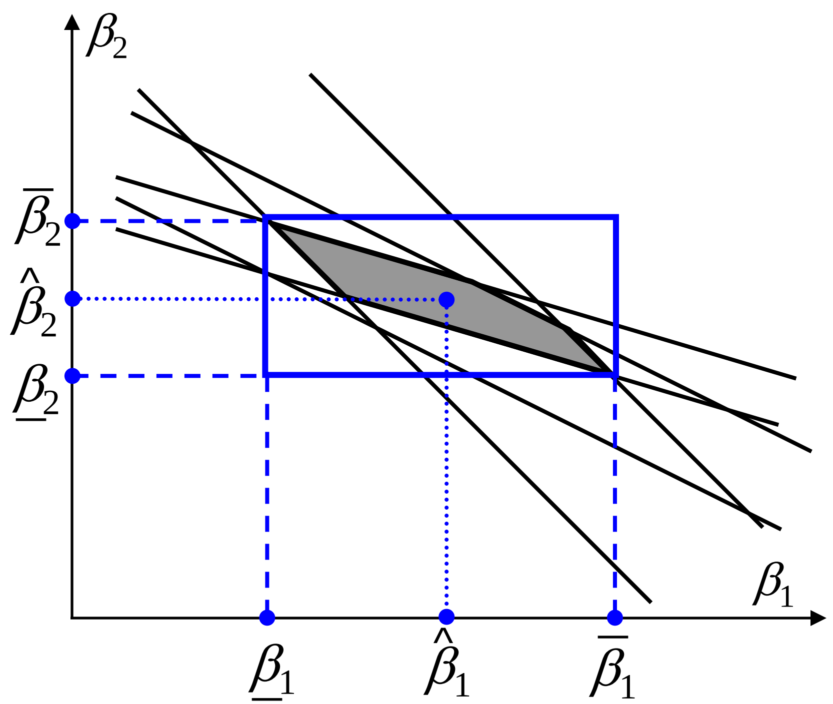
Linear Regression under Interval Error

- Problems stated with respect to uncertainty set B
  - Model parameters estimation
    - Interval estimates of β

$$\square B = \left(\left[\underline{\beta}_1, \overline{\beta}_1 \right], \ldots, \left[\underline{\beta}_p, \overline{\beta}_p \right] \right):$$

$$\underline{\beta}_i = \min_{\beta \in B} \beta_i, \quad \overline{\beta}_i = \max_{\beta \in B} \beta_i,$$

$$i = 1, \ldots, p.$$



    - Point estimates of β

$$\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p \right):$$

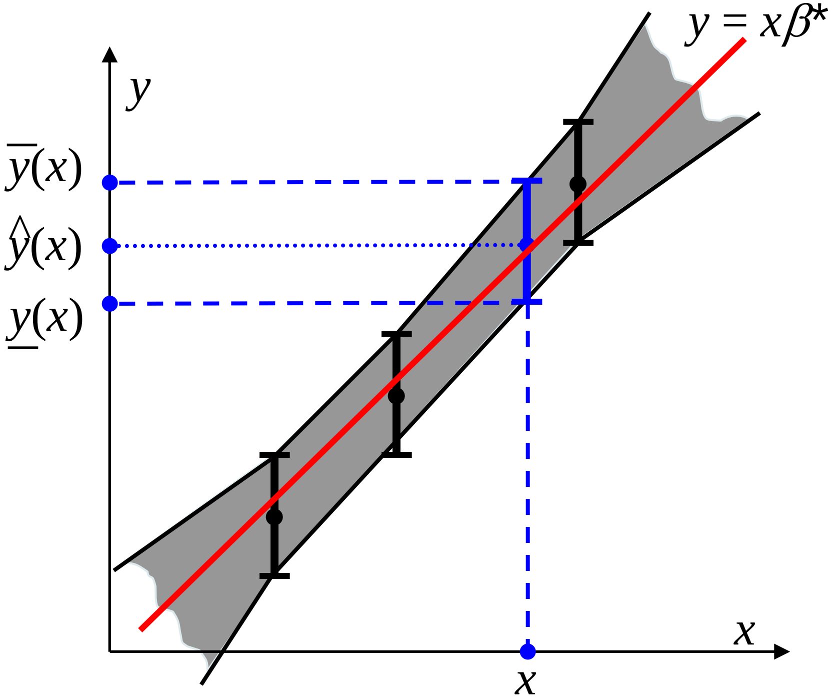
$$\hat{\beta}_i = \tfrac{1}{2} \left(\underline{\beta}_i + \overline{\beta}_i \right), \; i = 1, \ldots, p.$$

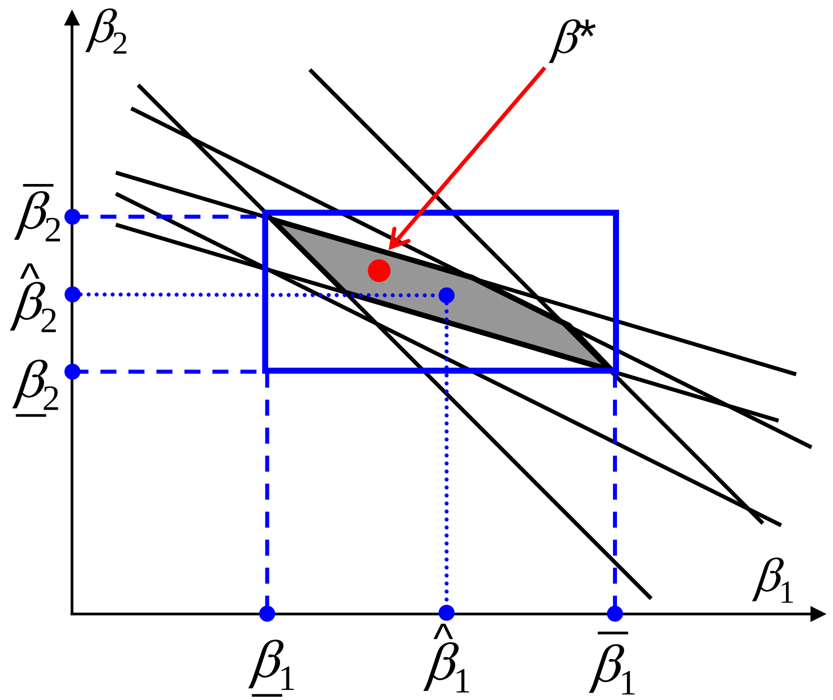
Fitting Experimental Data under Unknown-But-Bounded Error

- Years and Authors
  - 1962 L.V. Kantorovich
  - 1970 S.I. Spivak et al.
  - 1982 G. Belforte, M. Milanese et al.
  - 1983 N.M. Oskorbin et al.
  - 1986 J.P. Norton
  - 1987 S.I. Kumkov et al.
  - 1987 E. Walter, H. Piet-Lahanier
  - 1989 A.P. Voshchinin et al.
  - 1993 P.L. Combettes
  - 2000 O.E. Rodionova, A.L. Pomerantsev
  - 2003 A.A. Podruzhko, A.S. Podruzhko

Linear Regression under Interval Error

- Problems stated with respect to uncertainty set B
  - Prediction of the response for fixed values of input variables
  - Model parameters estimation





Validity of Prior Assumptions

- Interval regression assumptions
  - Structure of modeling function is fixed
  - Upper error bound is equal to $\bar{\varepsilon} \in [-\bar{\varepsilon}, \bar{\varepsilon}]$

- Milanese M., Novara C.:
  - **Definition***. Prior assumptions are considered validated if $B \neq \varnothing$.
  - The fact that the prior assumptions are validated (are consistent with the *present* data) does not exclude that they may be invalidated by *future* data**.

*Milanese M., Novara C. *Set membership identification of nonlinear systems* // Automatica 40 (2004), 957-975.

**Popper K.R. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Rontedge and Kegan Paul, 1969.

Uncertainty Center Method
(Oskorbin, 1983)

- If data are inconsistent ($B = \varnothing$) for certain model structure and upper error bound $\bar{\varepsilon}$
  - Find minimal feasible error ε_{min} (expand error bound until $B \neq \varnothing$)
  - Analyze ε_{min} and boundary samples of $B(\varepsilon_{\text{min}})$ to detect outliers or to modify model structure.

Oskorbin, N.M., Maksimov, A.V., and Zhilin, S.I., *Construction and Analysis of the Empirical Dependences Using the Uncertainty Center Method*, Izv. Alt. Gos. Univ., 1998, No. 1, pp. 35–38. (in Russian).

Simple Method for Outlier Detection
(Zhilin, 2004)

- Core idea
  - An outlier may be treated as a measurement with the underestimated error (i.e. the actual measurement error is greater than the declared error ε_j for it)
  - What are the lower bounds ε_j' for actual errors which provide non-empty uncertainty set?

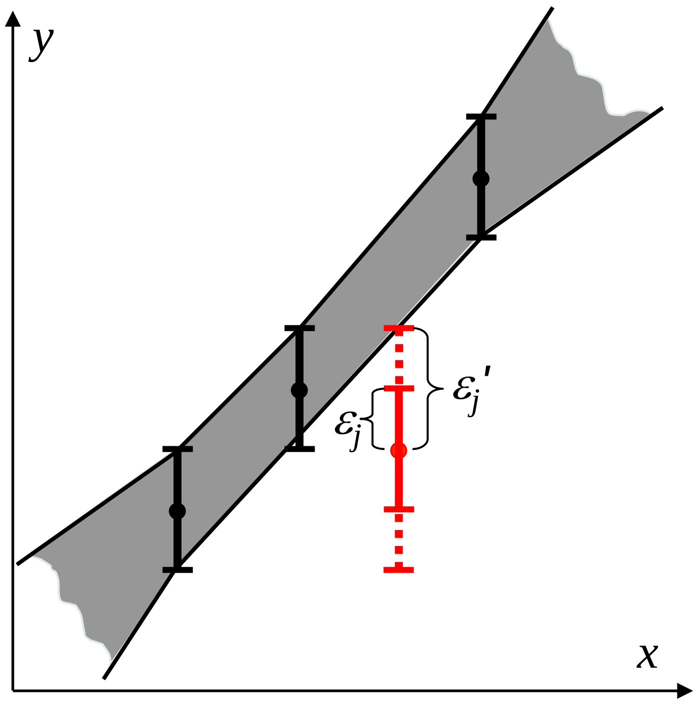
Zhilin S.I. On Fitting Empirical Data Under Interval Error // Reliable Computing (2005) 11 (5) 433-442.

Zhilin S.I. Simple Method for Outlier Detection in Fitting Experimental Data Under Interval Error // Chemometrics and Intelligent Laboratory Systems (2007) 88 (1) 60-68.

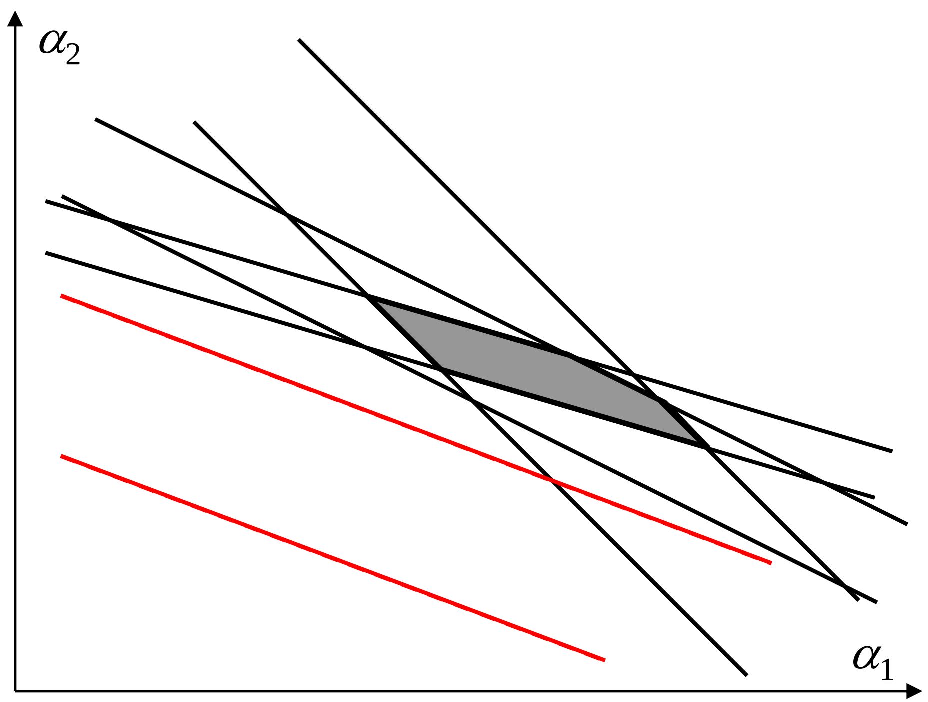
Simple Method for Outlier Detection
(Zhilin, 2004)

- How much must we stretch the declared error interval in order to «correct» an outlier?

In variables domain

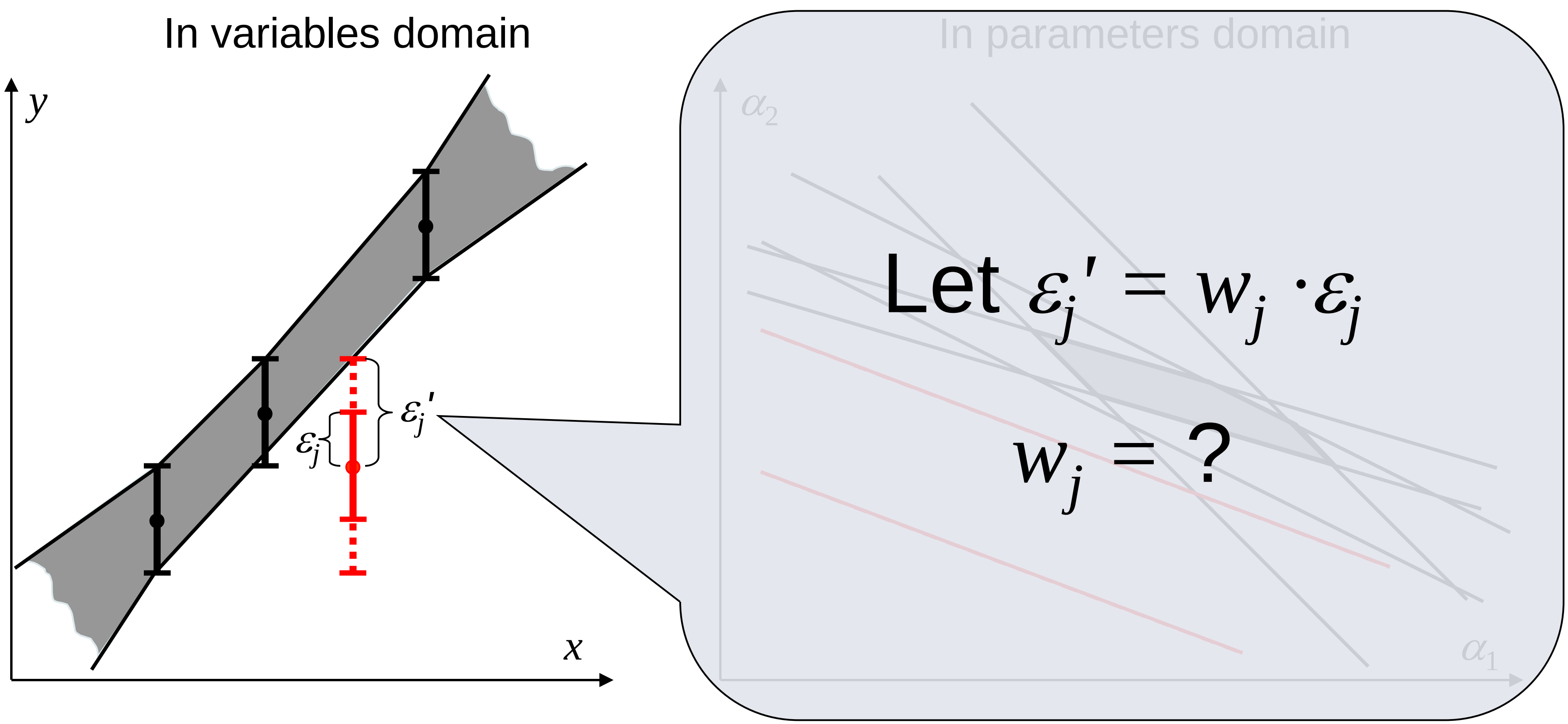


In parameters domain



Simple Method for Outlier Detection
(Zhilin, 2004)

- How much must we stretch the declared error interval in order to «correct» an outlier?



Simple Method for Outlier Detection
(Zhilin, 2004)

- Weights w_j may be found from the following optimization problem

$$\min_{\alpha,w} \sum_{j=1}^{n} w_j \qquad (1)$$

Uncertainty set constraints with movable bounds

$$y_j - w_j\varepsilon_j \le f(x_j,\beta) \le y_j + w_j\varepsilon_j, \qquad j=1,\ldots,n \qquad (2)$$

$$w_j \ge 1, \quad j=1,\ldots,k \qquad (3)$$

$$w_j = 1, \quad j=k+1,\ldots,n \qquad (4)$$

$$\begin{aligned} & w_1 = w_2 = \ldots = w_{j_1}, \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ & w_{j_m+1} = w_{j_m+2} = \ldots = w_n \end{aligned} \qquad (5)$$

Simple Method for Outlier Detection
(Zhilin, 2004)

- Weights w_j may be found from the following optimization problem

$$\min_{\alpha, w} \sum_{j=1}^{n} w_j \qquad (1)$$

$$y_j - w_j \varepsilon_j \le f(x_j, \beta) \le y_j + w_j \varepsilon_j, \qquad j = 1, \ldots, n \qquad (2)$$

$$w_j \ge 1, \quad j = 1, \ldots, k \qquad (3)$$

$$w_j = 1, \quad j = k+1, \ldots, n \qquad (4)$$

$$\begin{aligned} & w_1 = w_2 = \ldots = w_{j_1}, \\ & \ldots\ldots\ldots\ldots\ldots\ldots, \\ & w_{j_m+1} = w_{j_m+2} = \ldots = w_n \end{aligned} \qquad (5)$$

Uncertainty set constraints with movable bounds

We can only enlarge error intervals…

Simple Method for Outlier Detection
(Zhilin, 2004)

- Weights w_j may be found from the following optimization problem

$$\min_{\alpha, w} \sum_{j=1}^{n} w_j \qquad (1)$$

Uncertainty set constraints with movable bounds

$$y_j - w_j \varepsilon_j \le f(x_j, \beta) \le y_j + w_j \varepsilon_j, \qquad j = 1, \ldots, n \qquad (2)$$

$$w_j \ge 1, \quad j = 1, \ldots, k \qquad (3)$$

…or “freeze” some of error intervals

$$w_j = 1, \quad j = k+1, \ldots, n \qquad (4)$$

$$\begin{aligned} & w_1 = w_2 = \ldots = w_{j_1}, \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ & w_{j_m+1} = w_{j_m+2} = \ldots = w_n \end{aligned} \qquad (5)$$

Simple Method for Outlier Detection
(Zhilin, 2004)

- Weights w_j may be found from the following optimization problem

$$\min_{\alpha,w} \sum_{j=1}^{n} w_j \qquad (1)$$

Uncertainty set constraints with movable bounds

$$y_j - w_j \varepsilon_j \le f(x_j, \beta) \le y_j + w_j \varepsilon_j, \qquad j = 1, \ldots, n \qquad (2)$$

$$w_j \ge 1, \quad j = 1, \ldots, k \qquad (3)$$

$$w_j = 1, \quad j = k+1, \ldots, n \qquad (4)$$

Some of the measurements are obtained with equal errors

$$\begin{aligned} & w_1 = w_2 = \ldots = w_{j_1}, \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ & w_{j_m+1} = w_{j_m+2} = \ldots = w_n \end{aligned} \qquad (5)$$

Simple Method for Outlier Detection
(Zhilin, 2004)

- Relations to robust estimation

$$\min_{\alpha,w} \sum_{j=1}^{n} w_j \qquad (1)$$

$$y_j - w_j\varepsilon_j \le f(x_j, \beta) \le y_j + w_j\varepsilon_j, \qquad j = 1,\ldots,n \qquad (2)$$

$$w_j \ge 0, \quad j = 1,\ldots,n \qquad (3')$$

Uncertainty set constraints with movable bounds

~~We can only enlarge error intervals…~~

We allow free scale of error intervals (to expand and to contract)

Simple Method for Outlier Detection
(Zhilin, 2004)

- Relations to robust estimation

$$\min_{\alpha,w} \sum_{j=1}^{n} w_j \qquad (1)$$

$$y_j - w_j\varepsilon_j \le f(x_j,\beta) \le y_j + w_j\varepsilon_j, \qquad j=1,\ldots,n \qquad (2)$$

$$w_j \ge 0, \quad j=1,\ldots,n \qquad (3')$$

Uncertainty set constraints with movable bounds

W
en
in

Solution (β', w') of (1)-(3') gives
β^* is M-estimator for parameters β (known as L_1)

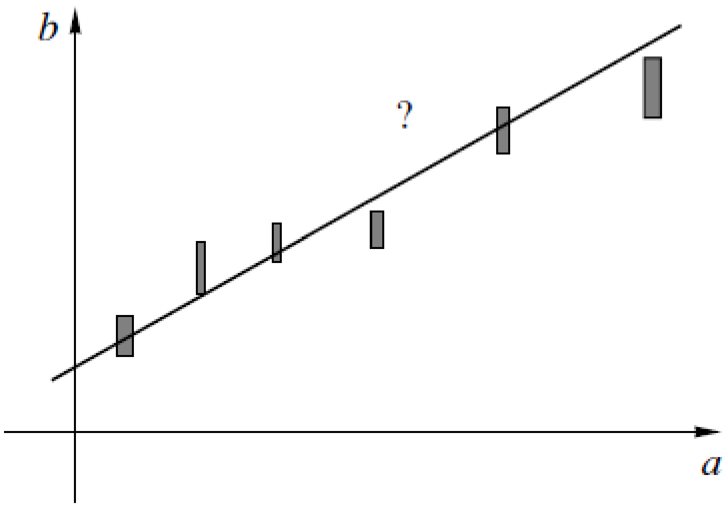
Weight function: $W(x) = 1/|x|$.

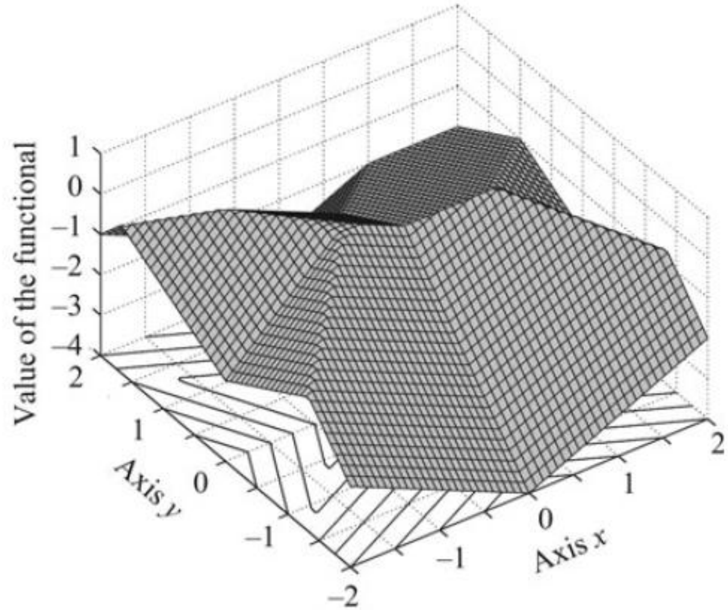
Residuals: $w_j^* \cdot \varepsilon_j$.

Method of Maximal Consistency
(Shary, 2012)

- Measure of consistency := USS recognizing functional

$$\mathrm{Uni}\,(x, \boldsymbol{A}, \boldsymbol{b}) = \min_{1\leqslant i\leqslant m} \left\{ \mathrm{rad}\, \boldsymbol{b}_i - \left\langle \mathrm{mid}\, \boldsymbol{b}_i - \sum_{j=1}^{n} \boldsymbol{a}_{ij}\, x_j \right\rangle \right\}$$

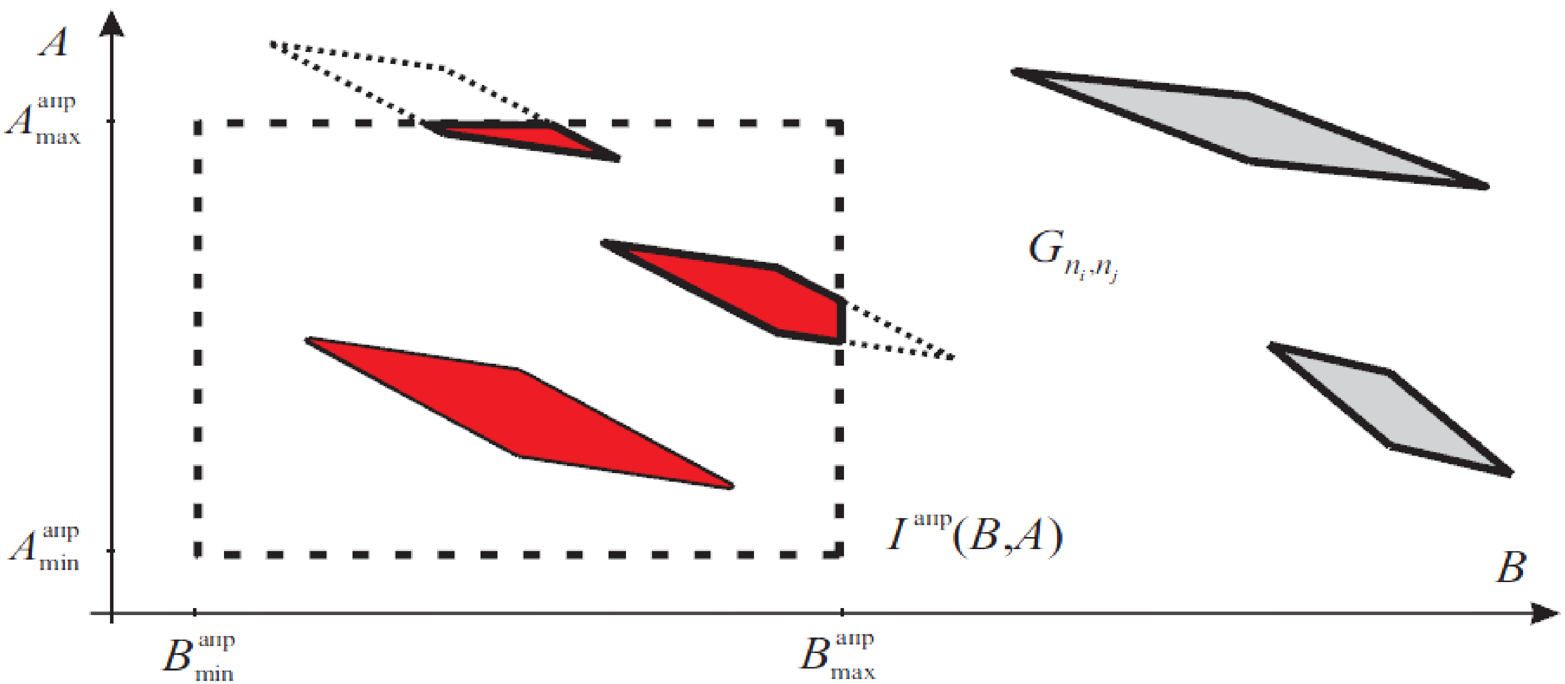




Shary S.P. *Solvability of Interval Linear Equations and Data Analysis under Uncertainty* // Automation and Remote Control (2012) 73 (2), 310–322.

Partial Information Sets
(Kumkov, 1987)

- Analysis of consistent subsamples



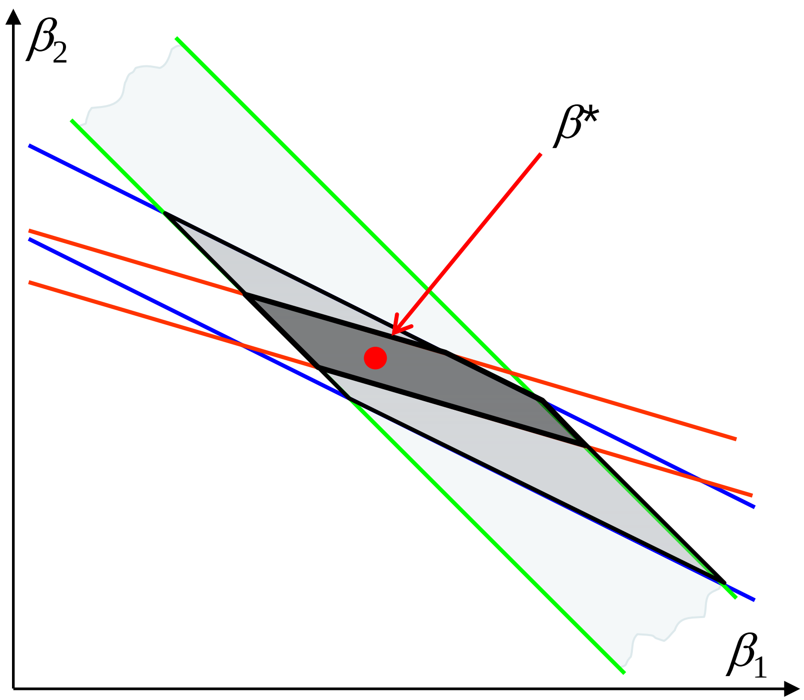
Kumkov S.I. *Processing of Experimental Data on Ionic Conductivity of Molten Electrolyte by the Interval Analysis Methods* // Rasplavy (2010), No. 3, pp. 86–96.

Potapov A.M., Kumkov S.I., and Y. Sato. *Processing of Experimental Data on Viscosity under One-Sided Character of Measuring Errors* // Rasplavy (2010), No. 3, pp. 55–70.

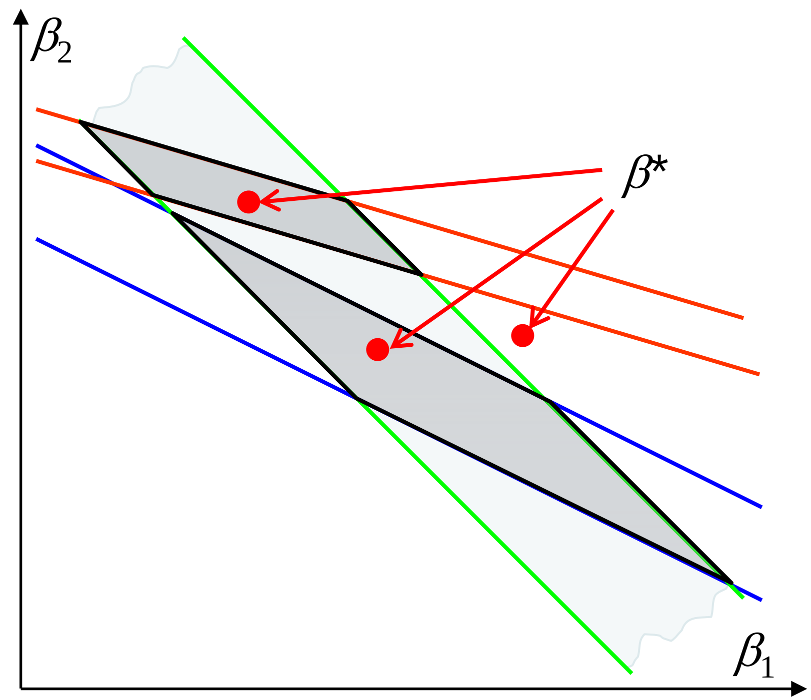
Data Consistency is Necessary But Not Sufficient

- Estimating model $y = \beta_1^* + \beta_2^* x$

$\beta^* \in B$



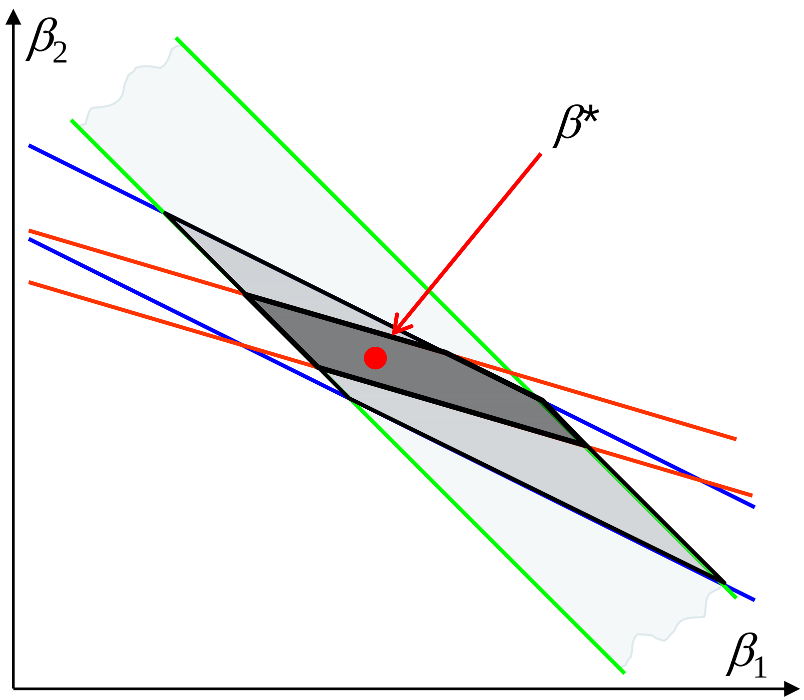
$B = \varnothing$



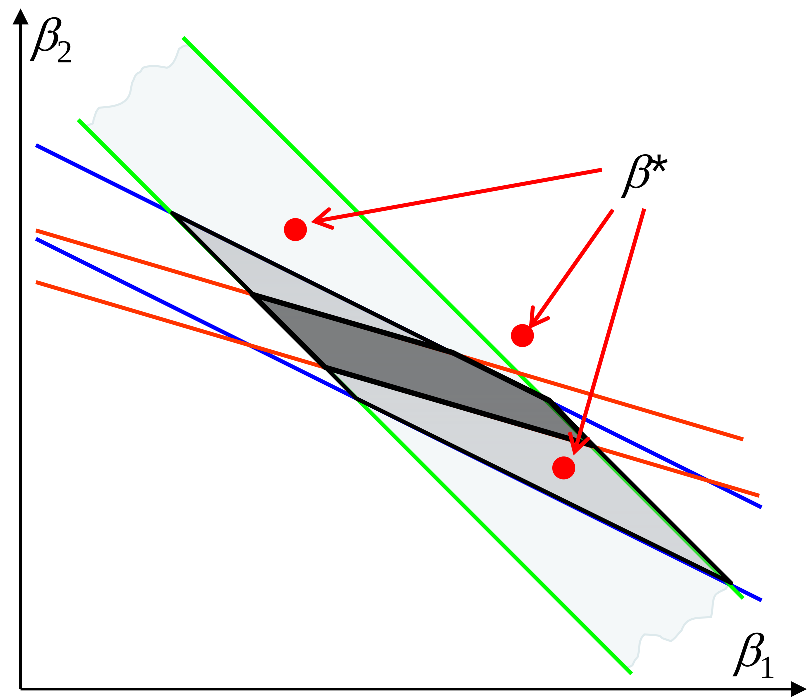
Data Consistency is Necessary But Not Sufficient

- Estimating model $y = \beta_1{}^* + \beta_2{}^* x$

$$\beta^* \in B$$



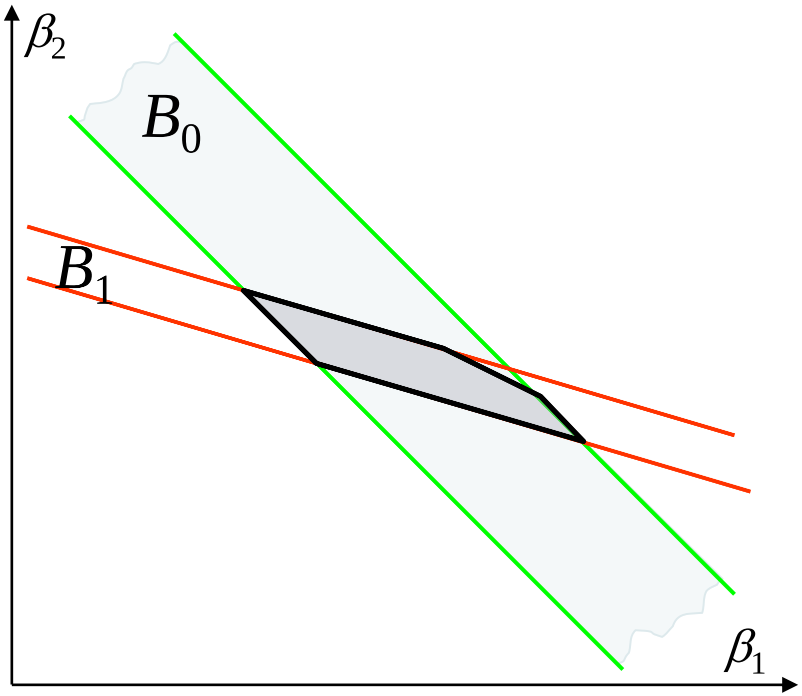
$$\beta^* \notin B \neq \emptyset$$

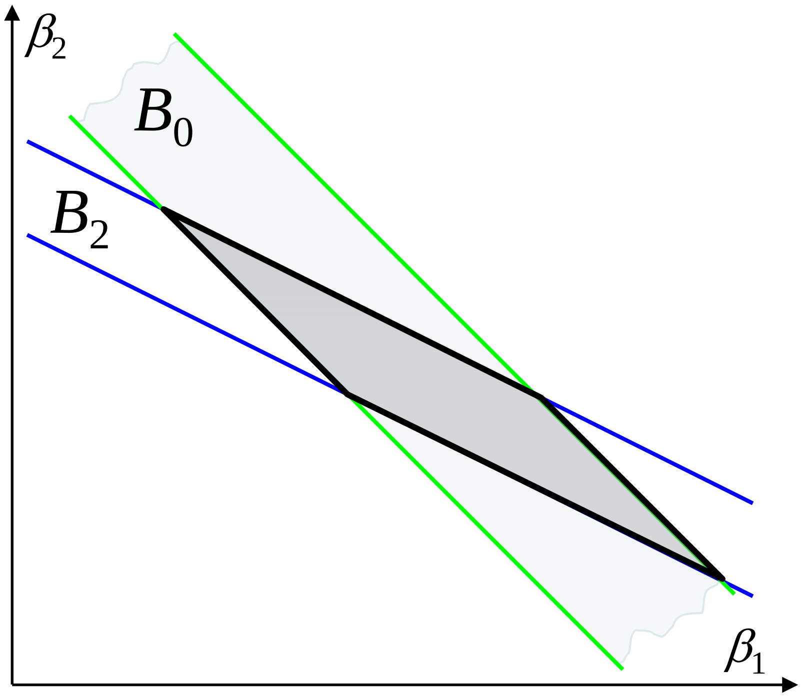


Revealing inconsistencies: possible additional indicators

- Estimate an informational value of each portion of data and knowledge with respect to the selected basic set

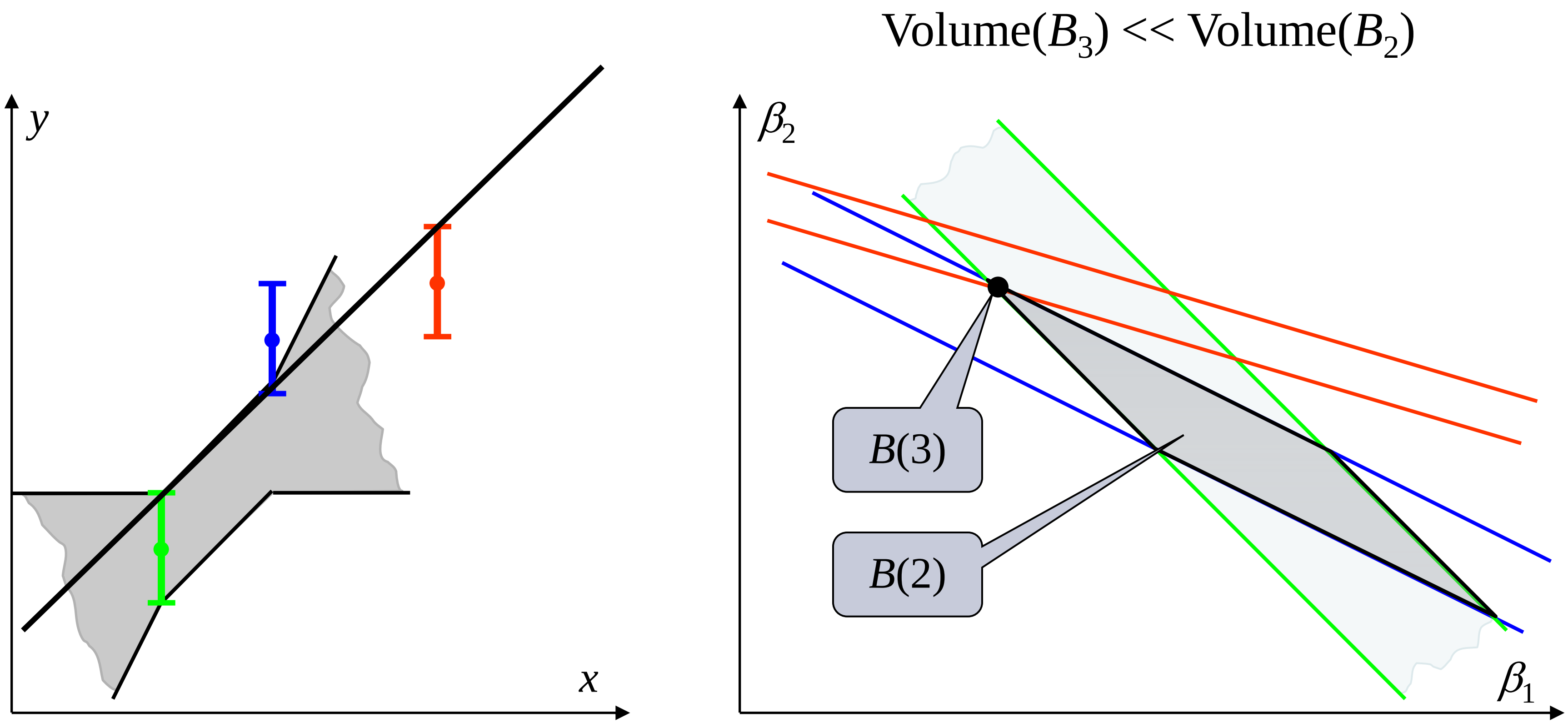
$$\text{Informativity}(B_i) = 1 - \frac{\text{Volume}(B_0 \cap B_i)}{\text{Volume}(B_0)}$$





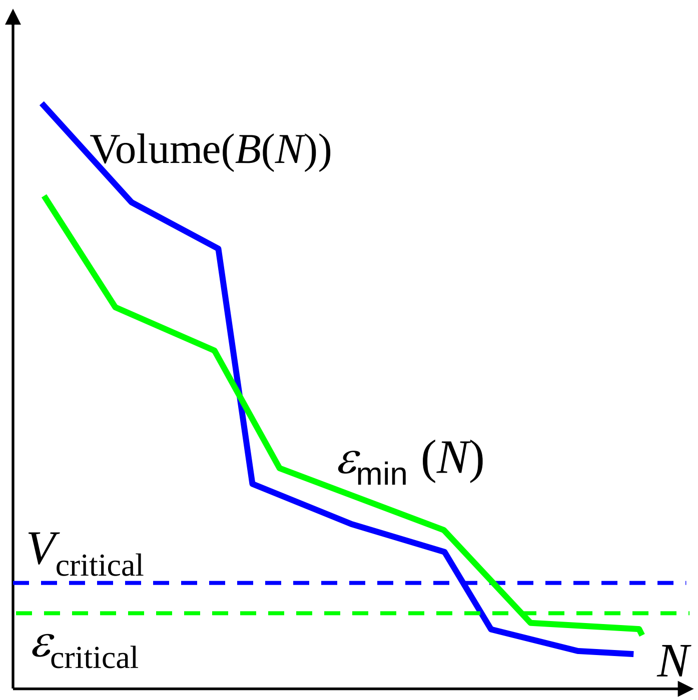
Revealing inconsistencies: possible additional indicators

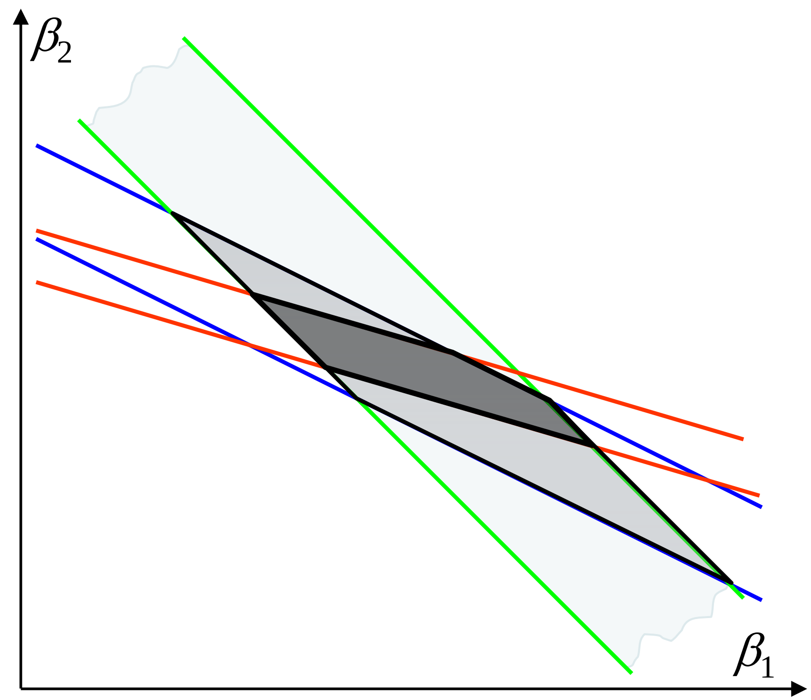
- Relate the volume of $B(N)$ to the value of N



Revealing inconsistencies: possible additional indicators

- Investigate the dynamics depending on N:
  - volume of $B(N)$
  - $\varepsilon_{\min}(N)$





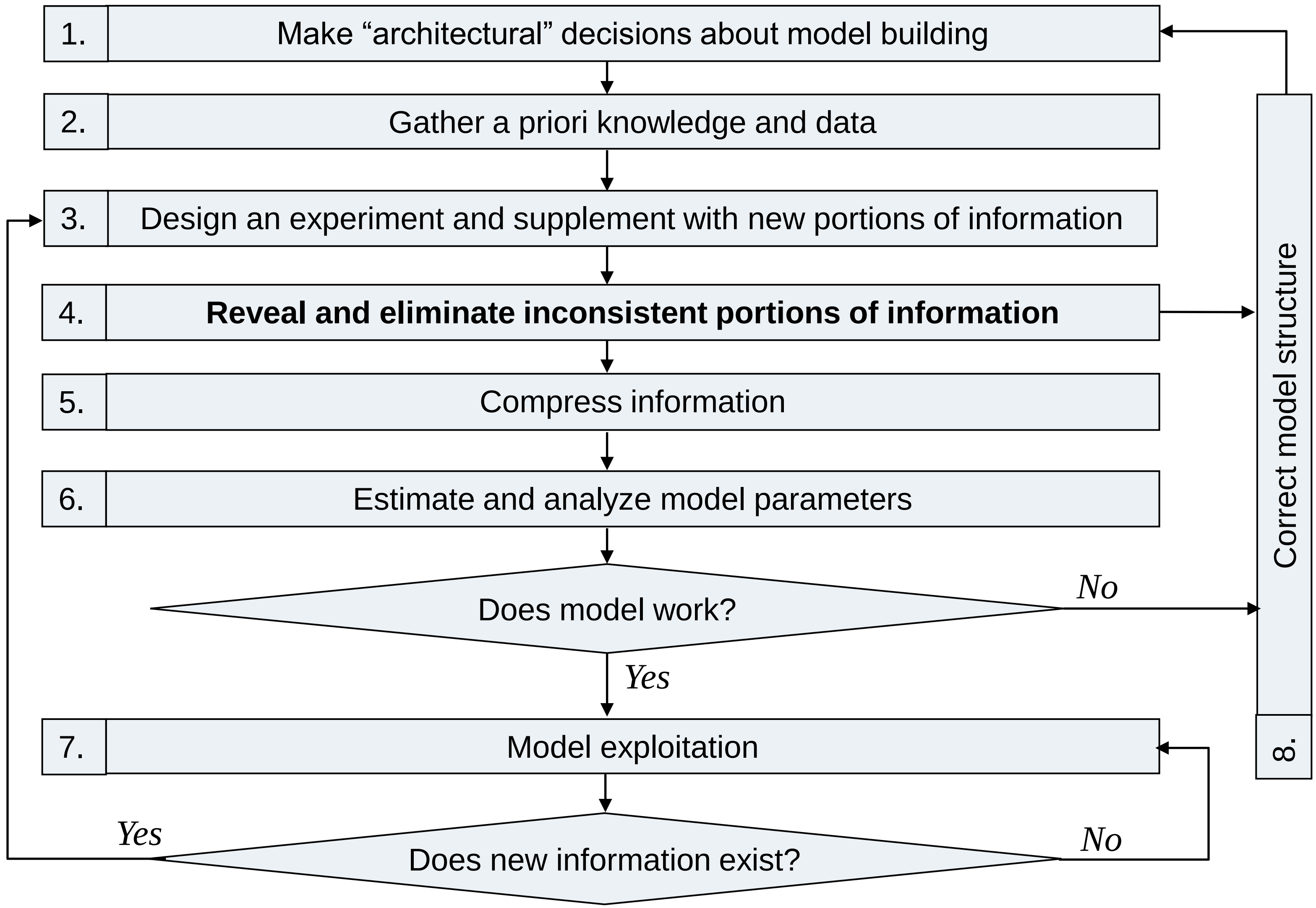
Revealing inconsistencies: possible additional indicators

- To discover inconsistencies one can
  - Estimate an informational value of each portion of data and knowledge with respect to the selected basic set
  - Relate the volume of $B(N)$ to the value of N
  - Investigate the dynamics of the volume of $B(N)$ and $\varepsilon_{min}(N)$ depending on N

Basic Principles

- It is impossible to obtain reliable estimates of process (object) parameters using an inconsistent set of data and knowledge about the process (object)
- Data consistency is necessary but not sufficient condition for reliable estimates
- None of the inner mathematical needs can be a ground for any kind of modifications of analyzed data and knowledge

General Scheme



Conclusions

- We propose
  - Basic principles and general scheme of building and analysis of empirical dependencies using interval analysis
  - Possible additional indicators of inconsistencies in data and knowledge

- Implementation of the proposed approach demands for the development of suitable mathematical tools and accumulation of experience in specific applications