Большие проблемы с маленькими данными

Сергей Жилин
CSort, Барнаул
szhilin@gmail.com

Small Data

Big Data

Michael Frei
KIDS Film
http://playkids.ch

Маленькие данные — это сколько?

Маленькие данные: это сколько?

https://www.kaggle.com/rafjaa/dealing-with-very-small-datasets

Маленькие данные: это сколько?

Small Data problems

Problems of small-data are numerous, but mainly revolve around high variance:

- **Over-fitting** becomes much harder to avoid
- You don't only over-fit to your training data, but sometimes you over-fit to your validation set as well.
- **Outliers** become much more dangerous.
- Noise in general becomes a real issue, be it in your target variable or in some of the features.

Summary

This could be a somewhat long list of things to do or try, but they all revolve around three main themes: constrained modeling, smoothing and quantification of uncertainty.

Most figures used in this post were taken from the book "Pattern Recognition and Machine Learning" by Christopher Bishop.

So what to do in these situation?

1- Hire a statistician

2- Stick to simple models

3- Pool data when possible

4- Limit Experimentation

5- Do clean up your data

6- Do perform feature selection

7- Do use Regularization

8- Do use Model Averaging

9- Try Bayesian Modeling and Model Averaging

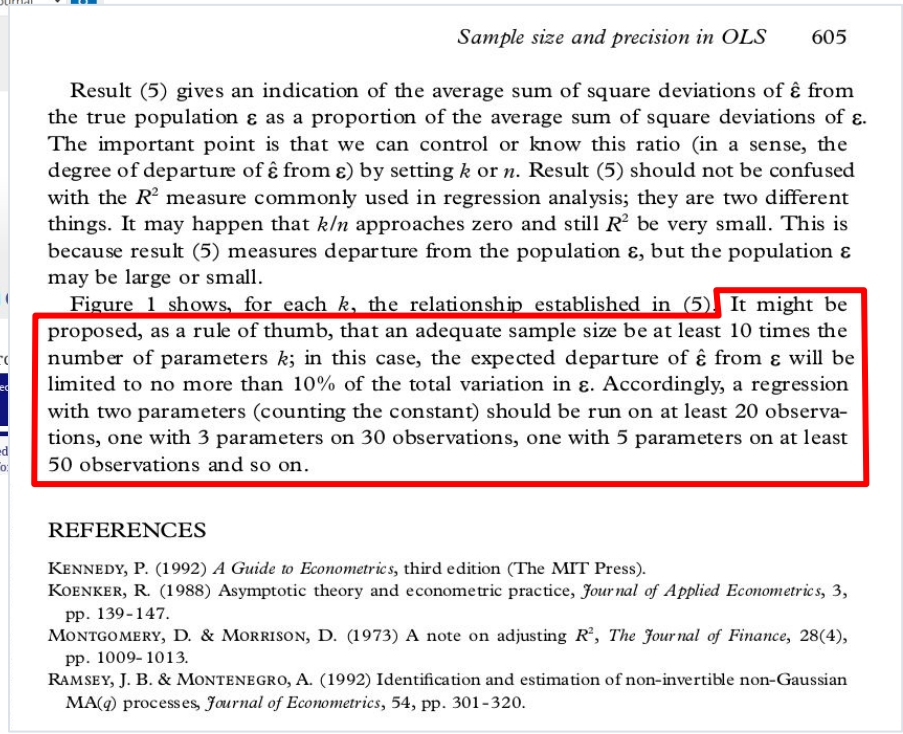
10- Prefer Confidence Intervals to Point Estimates

It is usually a good idea to get an estimate of confidence in your prediction in addition to producing the prediction itself. For regression analysis this usually takes the form of predicting a range of values that is calibrated to cover the true value 95% of the time or in the case of classification it could be just a matter of producing class probabilities. This becomes more crucial with small data sets as it becomes more likely that certain regions in your feature space are less represented than others. Model averaging as referred

https://medium.com/rants-on-machine-learning/what-to-do-with-small-data-d253254d1a89

Маленькие данные: это сколько?

МНК: наблюдений нужно в 10 раз больше, чем параметров



Sample size and precision in OLS 605

Result (5) gives an indication of the average sum of square deviations of $\hat{\varepsilon}$ from the true population ε as a proportion of the average sum of square deviations of ε. The important point is that we can control or know this ratio (in a sense, the degree of departure of $\hat{\varepsilon}$ from ε) by setting k or n. Result (5) should not be confused with the R^2 measure commonly used in regression analysis; they are two different things. It may happen that k/n approaches zero and still R^2 be very small. This is because result (5) measures departure from the population ε, but the population ε may be large or small.

Figure 1 shows, for each k, the relationship established in (5). It might be proposed, as a rule of thumb, that an adequate sample size be at least 10 times the number of parameters k; in this case, the expected departure of $\hat{\varepsilon}$ from ε will be limited to no more than 10% of the total variation in ε. Accordingly, a regression with two parameters (counting the constant) should be run on at least 20 observations, one with 3 parameters on 30 observations, one with 5 parameters on at least 50 observations and so on.

REFERENCES

KENNEDY, P. (1992) *A Guide to Econometrics*, third edition (The MIT Press).
KOENKER, R. (1988) Asymptotic theory and econometric practice, *Journal of Applied Econometrics*, 3, pp. 139-147.
MONTGOMERY, D. & MORRISON, D. (1973) A note on adjusting R^2, *The Journal of Finance*, 28(4), pp. 1009-1013.
RAMSEY, J. B. & MONTENEGRO, A. (1992) Identification and estimation of non-invertible non-Gaussian MA(q) processes, *Journal of Econometrics*, 54, pp. 301-320.

Montenegro A. / , J. Appl. Stat. 28(5) (2001) 603–605. DOI: 10.1080/02664760120047933

Маленькие данные: это сколько?
ГОСТ Р 8.736-2011: Оценка границ погрешности

7 Доверительные границы случайной погрешности

7.1 Доверительные границы случайной погрешности оценки измеряемой величины в соответствии с настоящим стандартом устанавливают для результатов измерений, принадлежащих нормальному распределению.

При невыполнении этого условия методы вычисления доверительных границ случайной погрешности должны быть указаны в методике измерений.

7.2 При числе результатов измерений $n \le 15$ принадлежность их к нормальному распределению не проверяют. При этом вычисление доверительных границ случайной погрешности оценки измеряемой величины по методике, предусмотренной настоящим стандартом, допускается только в том случае, если заранее известно, что результаты измерений принадлежат нормальному распределению.

Примечание — Если не известно распределение погрешностей оценки искомой величины, способы нахождения доверительных границ случайной погрешности могут быть указаны в методике измерений с учетом того, что подобные измерения повторяют.

7.3 При числе результатов измерений $15 < n \le 50$ для проверки принадлежности их к нормальному распределению предпочтителен составной критерий, приведенный в приложении Б.

7.4 При числе результатов измерений $n > 50$ для проверки принадлежности их к нормальному распределению предпочтителен один из критериев: χ^2 К. Пирсона или ω^2 Мизеса—Смирнова. Критерий К. Пирсона χ^2 приведен в приложении В, критерий ω^2 Мизеса—Смирнова — в приложении Г.

7.5 Доверительные границы ε (без учета знака) случайной погрешности оценки измеряемой величины вычисляют по формуле

$$\varepsilon = t S_{\bar{x}}, \qquad (6)$$

где t — коэффициент Стьюдента, который в зависимости от доверительной вероятности P и числа результатов измерений n находят по таблице, приведенной в приложении Д.

ГОСТ Р 8.736-2011. ГСИ. ИЗМЕРЕНИЯ ПРЯМЫЕ МНОГОКРАТНЫЕ. Методы обработки результатов измерений. Основные положения https://docs.cntd.ru/document/1200089016

Маленькие данные: это сколько?

ГОСТ Р 8.736-2011: Исключение выбросов

ГОСТ Р 8.736—2011

6 Исключение грубых погрешностей

6.1 Для исключения грубых погрешностей используют критерий Граббса. Статистический критерий Граббса исключения грубых погрешностей основан на предположении о том, что группа результатов измерений принадлежит нормальному распределению. Для этого вычисляют критерии Граббса G_1 и G_2, предполагая, что наибольший x_{max} или наименьший x_{min} результат измерений вызван грубыми погрешностями:

$$G_1 = \frac{|x_{max} - \overline{x}|}{S}, \qquad G_2 = \frac{|\overline{x} - x_{min}|}{S}. \tag{5}$$

Сравнивают G_1 и G_2 с теоретическим значением G_T критерия Граббса при выбранном уровне значимости q. Таблица критических значений критерия Граббса приведена в приложении А.

Если $G_1 > G_T$, то x_{max} исключают как маловероятное значение. Если $G_2 > G_T$, то x_{min} исключают как маловероятное значение. Далее вновь вычисляют среднее арифметическое и среднее квадратическое отклонения ряда результатов измерений и процедуру проверки наличия грубых погрешностей повторяют.

Если $G_1 \le G_T$, то x_{max} не считают промахом и его сохраняют в ряду результатов измерений. Если $G_2 \le G_T$, то x_{min} не считают промахом и его сохраняют в ряду результатов измерений.

Таблица А.1 — Критические значения G_T для критерия Граббса

n	Одно наибольшее или одно наименьшее значение при уровне значимости q	
	Свыше 1 %	Свыше 5 %
3	1,155	1,155
4	1,496	1,481
5	1,764	1,715

ГОСТ Р 8.736-2011. ГСИ. ИЗМЕРЕНИЯ ПРЯМЫЕ МНОГОКРАТНЫЕ. Методы обработки результатов измерений. Основные положения https://docs.cntd.ru/document/1200089016

Логические основания

Логические основания

Логические основания

“Закат” научного метода

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

CHRIS ANDERSON SCIENCE 06.23.2008 12:00 PM

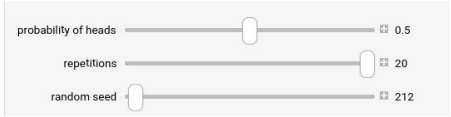
The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

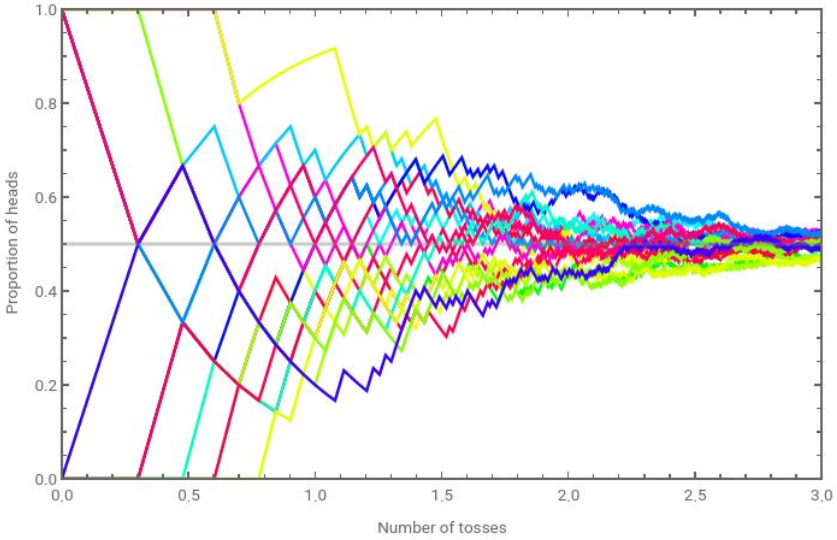
Illustration: Marian Bantjes “All models are wrong, but some are useful.” So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies [...]

https://www.wired.com/2008/06/pb-theory/

- Massimo Pigliucci. The end of theory in science? / doi: 10.1038/embor.2009.111
- The Paradox of Deductive Reason in Data Science, Featuring Donald Trump's Twitter Account https://towardsdatascience.com/the-paradox-of-deductive-reason-in-data-science-featuring-donald-trumps-twitter-account-43839d4dda82
- Rothchild Irving. Induction, Deduction, And The Scientific Method. An Eclectic Overview Of The Practice Of Science. https://higherlogicdownload.s3.amazonaws.com/SSR/fbd87d69-d53f-458a-8220-829febdf990b/UploadedImages/Documents/rothchild_scimethod.pdf
- Mazzocchi Fulvio. Could Big Data be the end of theory in science? A few remarks on the epistemology of data- driven science https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4766450/
- Big data: farewell to Cartesian thinking? http://parisinnovationreview.com/articles-en/big-data-farewell-to-cartesian-thinking
- Gelman Andrew. Induction and Deduction in Bayesian Data Analysis* http://www.stat.columbia.edu/~gelman/research/published/philosophy_online4.pdf

Вероятностная модель случайных явлений





- Случайное явление — **математический** объект (событие, величина, функция), который исчерпывающе характеризуется законом распределения вероятностей
- Вероятность $P(A)$ случайного события A — предел частоты $p_N(A)$ его наблюдения в одинаковых условиях

$$P(A) = \lim_{N \to \infty} p_N(A)$$

- При малых N частота $p_N(A)$ может сильно флуктуировать

https://demonstrations.wolfram.com/SimulatedCoinTossingExperimentsAndTheLawOfLargeNumbers/

Статистическая (не)устойчивость

- Корректность применения теории вероятностей обеспечивается принятием следующих гипотез
  - **(A) гипотезы идеальной статистической устойчивости** (статистической прогнозируемости) — наличие сходимости статистик к постоянным величинам
  - **(B) гипотезы адекватного описания реальных физических явлений случайными моделями**
- Экспериментальные исследования физических явлений на больших интервалах наблюдений показывают нарушения гипотезы **(A)**
- **Вероятность не имеет физической интерпретации**
- **Вероятность — математическая абстракция**

УФН

RSS-ленты

Выпуск 6, 2021

Русский English

Вход →

Получить статью →

Выпуски Авторы PACS Подписчикам Для авторов

Поиск →

Выпуски / 1992 / Июль

Методические заметки

← →

Является ли вероятность «нормальной» физической величиной?

Ю.И. Алимов, Ю.А. Кравцов[a]

[a]Институт космических исследований РАН, ул. Профсоюзная 84/32, Москва, 117997, Российская Федерация

Обсуждаются неформальные аспекты теории вероятности и математической статистики, возникающие при интерпретации физических экспериментов. Изложены требования к верифицирующему эксперименту и на примере математического ожидания проанализирована роль эвристических (внеопытных) утверждений. Перечислены главные гипотезы, скрывающиеся в тени экспериментов: принцип многократного воспроизведения ("как раньше бывало, так, видимо, и будет"); принцип разумной достаточности; статистический принцип («лучше прогнозировать что-нибудь, чем ничего'). Значительное внимание уделено фишеровским и многовыборочным доверительным интервалам. Отмечена несостоятельность фишеровских доверительных интервалов. Перечислены поводы для вовлечения домысливания в практическое исчисление вероятностей: неполнота любой системы гипотез; субъективные оценки вероятностей; примысливание статистического ансамбля; нестационарность и неустойчивость; редкие явления; использование классических вероятностей и закона больших чисел. Сделан вывод, что относительная частота (эмпирическая вероятность) является «нормальной» физической величиной в том смысле, что она допускает физическое измерение. Ее «ненормальность» выражается в том, что она больше других физических величин нагружена условностями и гипотезами, которые требуют специальной проверки (верификации).

Текст: pdf

Войдите или зарегистрируйтесь чтобы получить доступ к полным текстам статей.

PACS: 02.50.Cw, 06.20.Dk, 05.20.Gg, 02.70.Rr (все)

DOI: 10.3367/UFNr.0162.199207d.0149

URL: https://ufn.ru/ru/articles/1992/7/d/

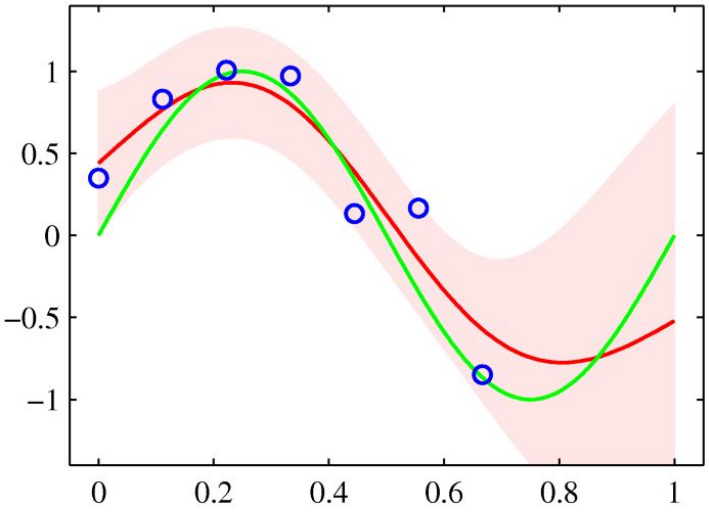
Алимов Ю И, Кравцов Ю А / УФН 162 (7) 149–182 (1992). DOI: 10.3367/UFNr.0162.199207d.0149

Горбань И.И. Феномен стаистической устойчивости. — Киев: Наукова думка, 2014. https://www.researchgate.net/publication/311653342_Fenomen_statisticeskoj_ustojcivosti

Нужно “хорошее” описание неопределённости

10- Prefer Confidence Intervals to Point Estimates

It is usually a good idea to get an estimate of confidence in your prediction in addition to producing the prediction itself. For regression analysis this usually takes the form of predicting a range of values that is calibrated to cover the true value 95% of the time or in the case of classification it could be just a matter of producing class probabilities. This becomes more crucial with small data sets as it becomes more likely that certain regions in your feature space are less represented than others. Model averaging as referred to in the previous two points allows us to do that pretty easily in a generic way for regression, classification and density estimation. It is also useful to do that when evaluating your models. Producing confidence intervals on the metrics you are using to compare model performance is likely to save you from jumping to many wrong conclusions.

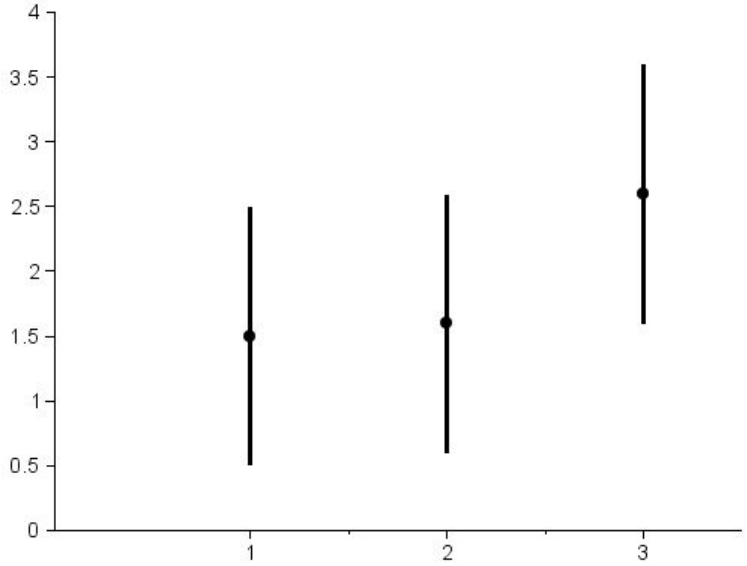


Некоторые части признакового пространства хуже покрыты данными и доверительные интервалы предсказаний должны отражать этот факт

https://medium.com/rants-on-machine-learning/what-to-do-with-small-data-d253254d1a89

Нужно “хорошее” описание неопределённости

ГОСТ Р 8.736-2011: Исключение выбросов

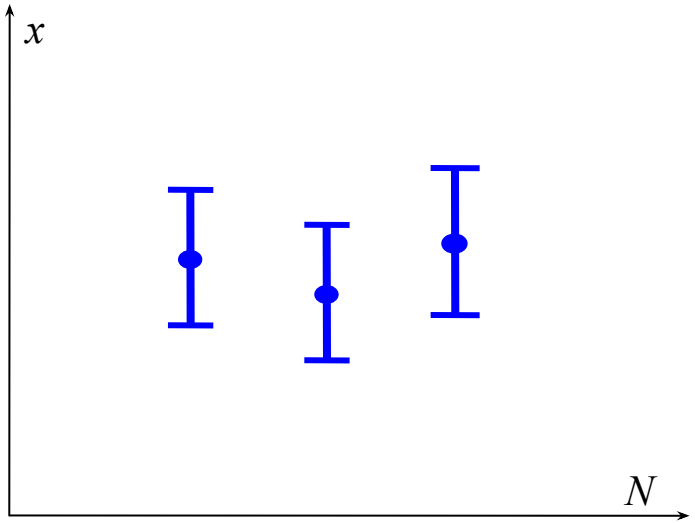


Выброс по критерию Граббса

Что делать?

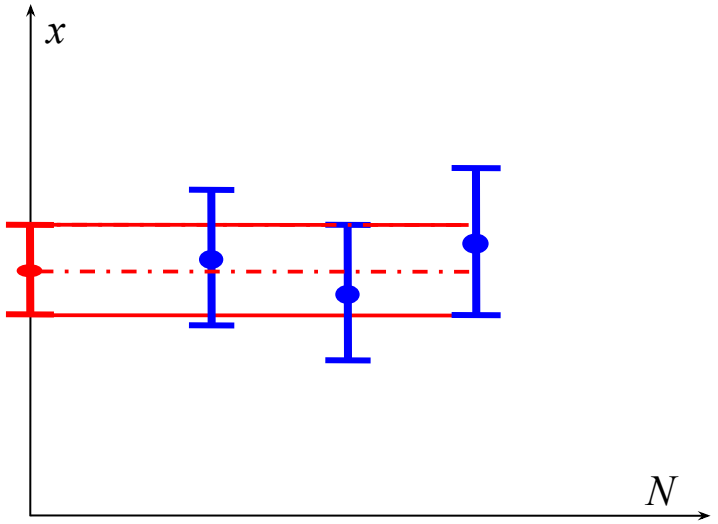
Интервальный подход: основные гипотезы

- Неопределённость измерения полагается ограниченной, т.е. принадлежащей интервалу
- Никаких других предположений о неопределённости нет



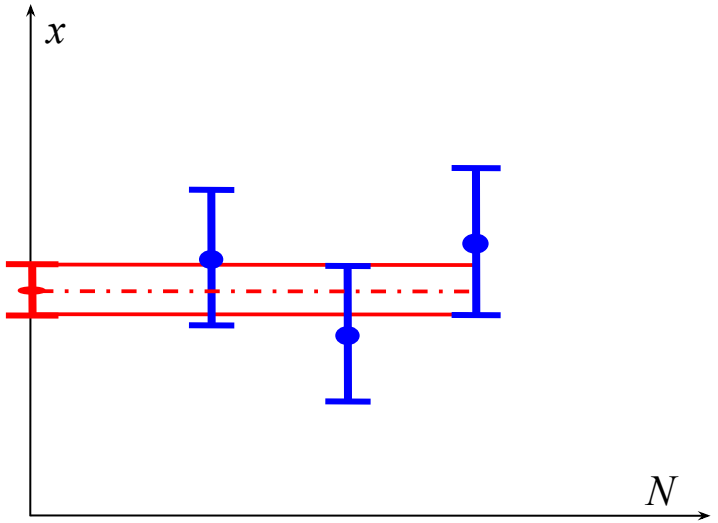
Интервальный подход: оценивание истинного значения

- Интервальная оценка истинного значения – пересечение интервалов наблюдений
- Точечной оценкой может служить центр интервальной оценки



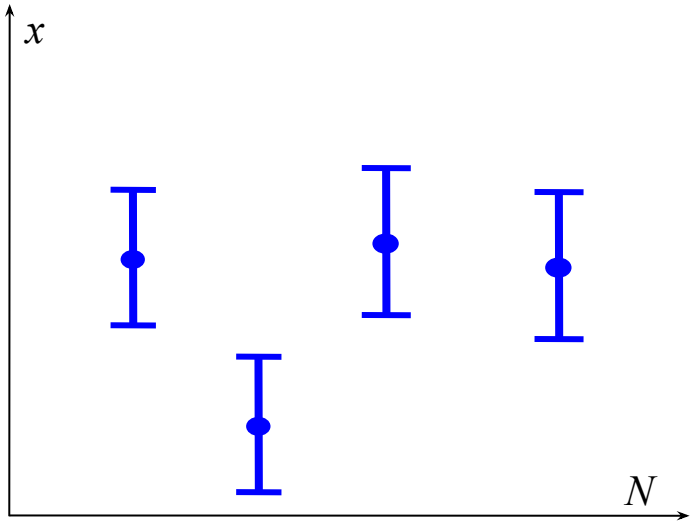
Интервальный подход: оценивание истинного значения

- Интервальная оценка истинного значения – пересечение интервалов наблюдений
- Точечной оценкой может служить центр интервальной оценки



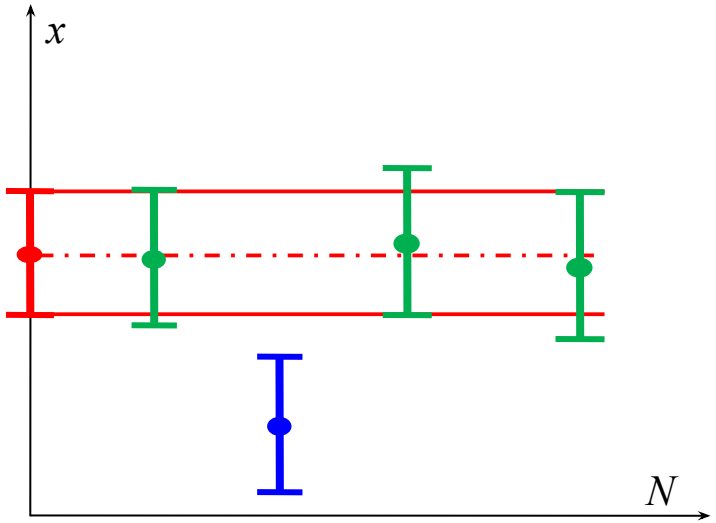
Интервальный подход: выявление выбросов

- Индикатор наличия выбросов – пустота интервальной оценки



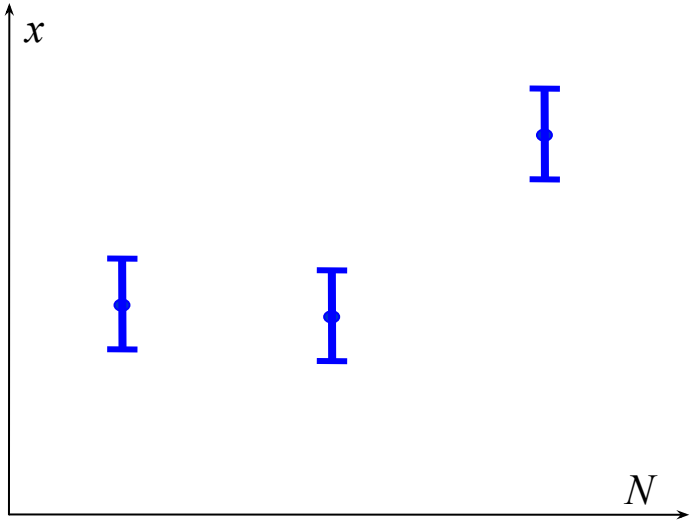
Интервальный подход: выявление выбросов

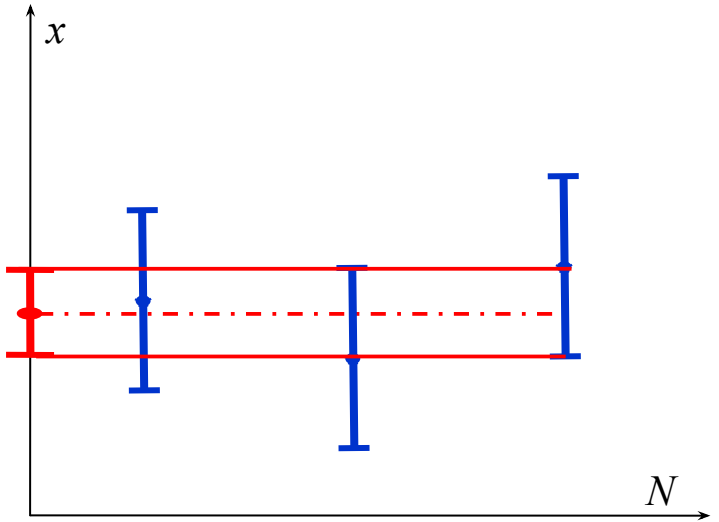
- Индикатор наличия выбросов – пустота интервальной оценки
- Выделение совместных подвыборок – например, поиск полных подграфов в графе попарной совместности измерений



Интервальный подход: мера совместности выборки

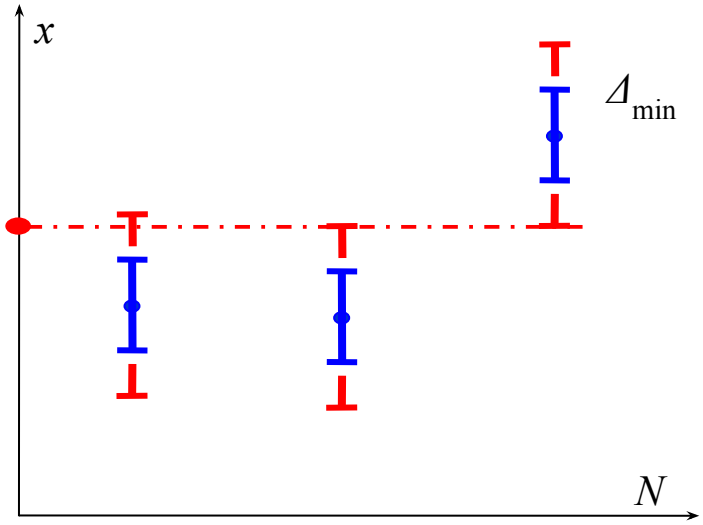
- Мерой совместности выборки служит предельный минимальный уровень погрешности $\Delta_{\min}$

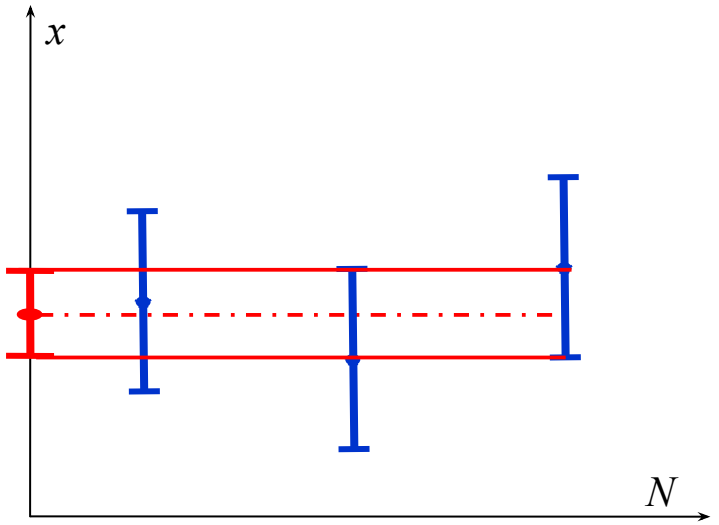




Интервальный подход: мера совместности выборки

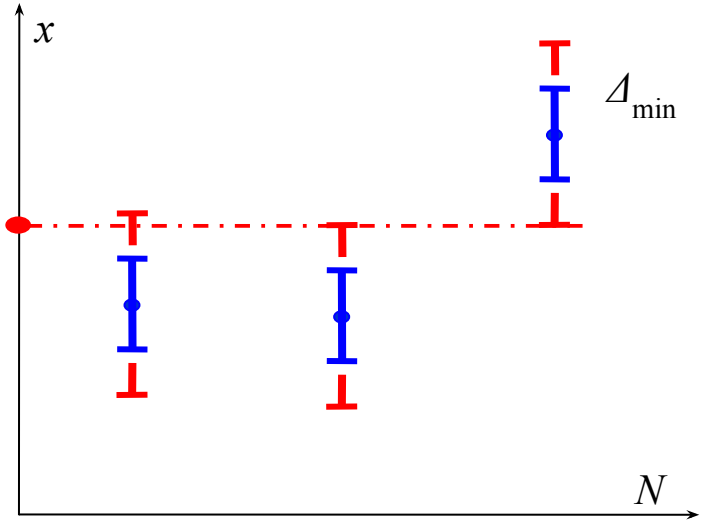
- При несовместности выборки интервалы измерений растягиваются до появления общей точки

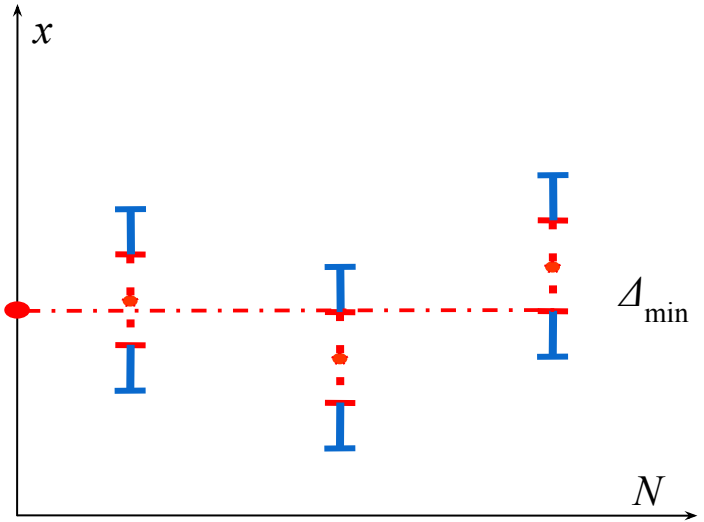




Интервальный подход: мера совместности выборки

- При несовместности выборки интервалы измерений растягиваются до появления общей точки
- Для совместной выборки интервалы измерений сжимаются до единственной общей точки

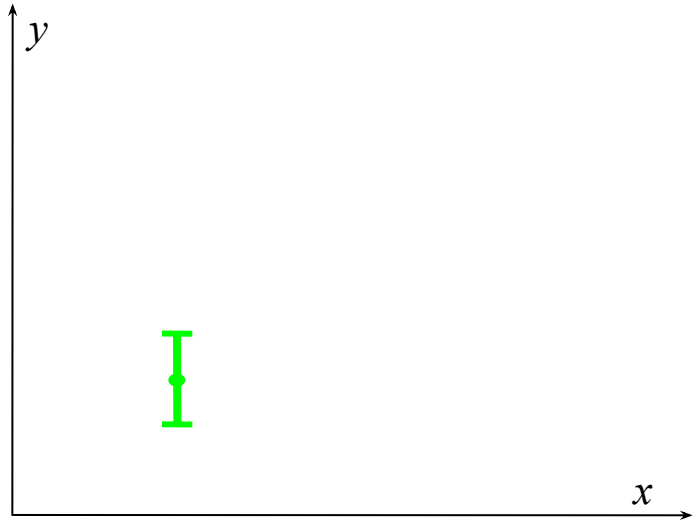




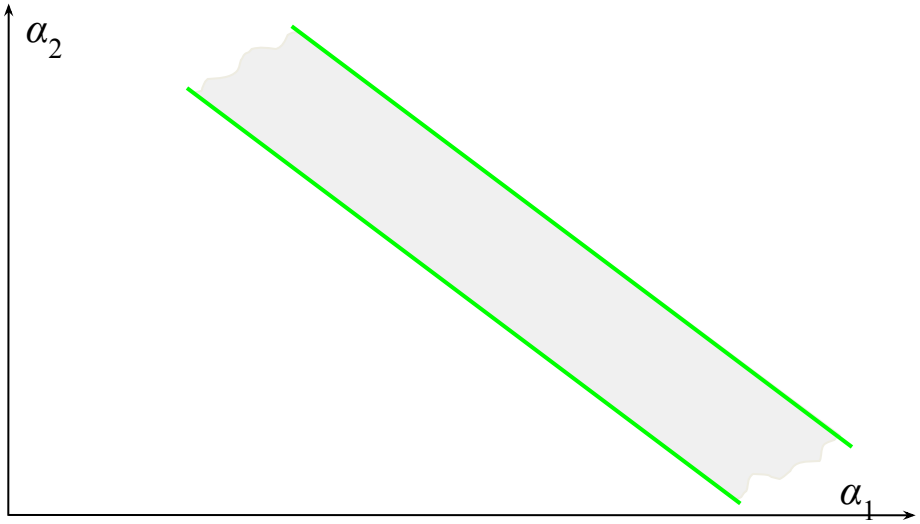
Интервальный подход: построение зависимости

- Построение модели $y = \alpha_1 + \alpha_2 x$

Пространство (x, y)



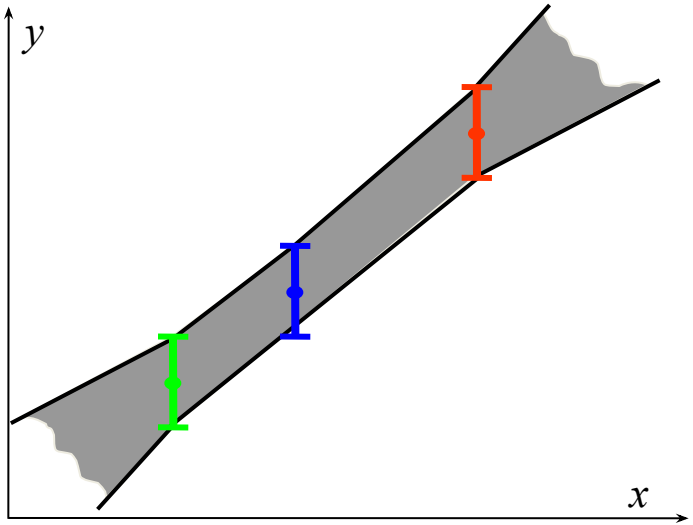
Пространство (α_1, α_2)



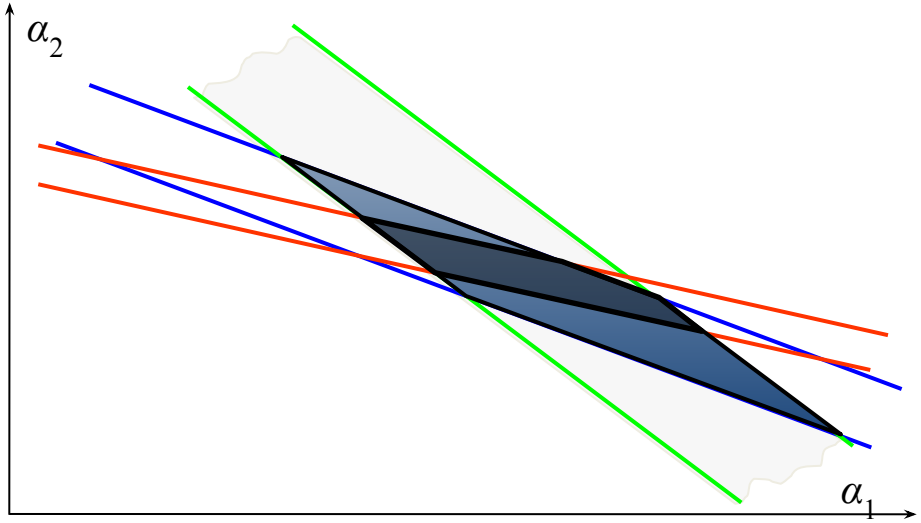
Интервальный подход: построение зависимости

- Построение модели $y = \alpha_1 + \alpha_2 x$

Пространство (x, y)



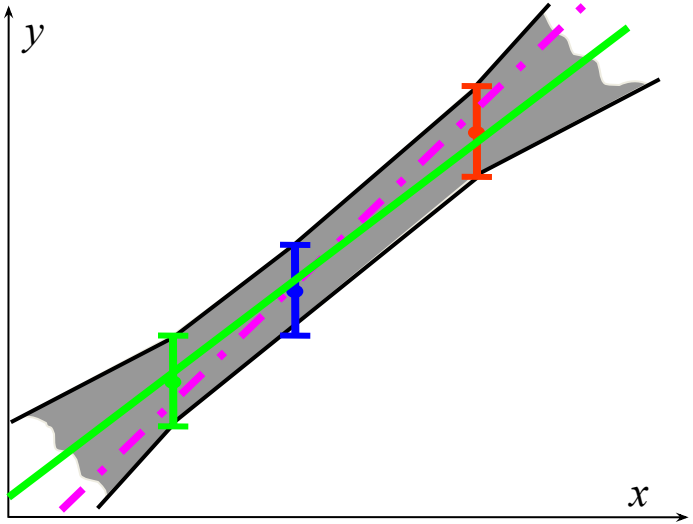
Пространство (α_1, α_2)



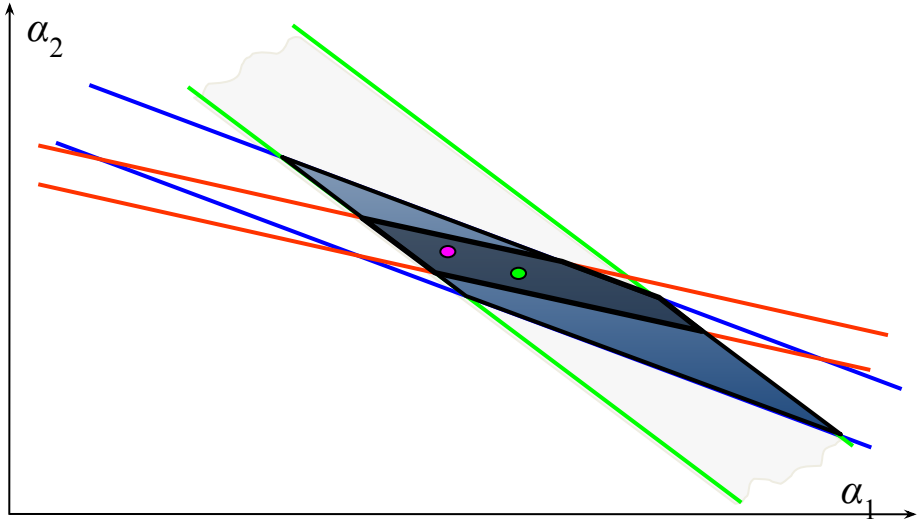
Интервальный подход: построение зависимости

- Построение модели $y = \alpha_1 + \alpha_2 x$

Пространство (x, y)



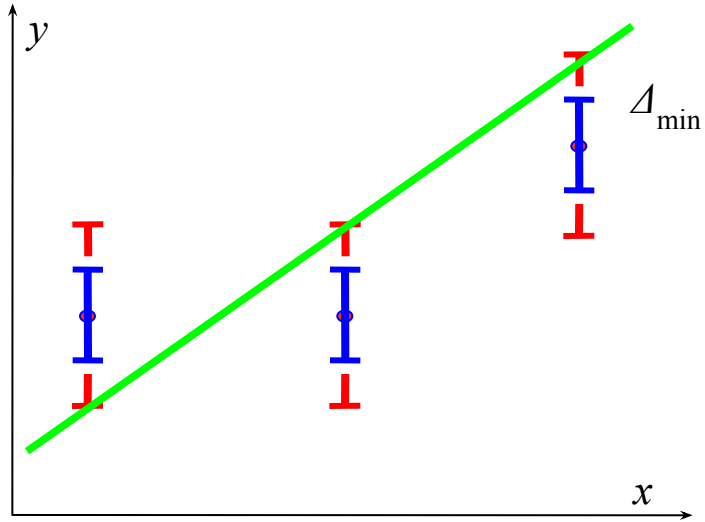
Пространство (α_1, α_2)



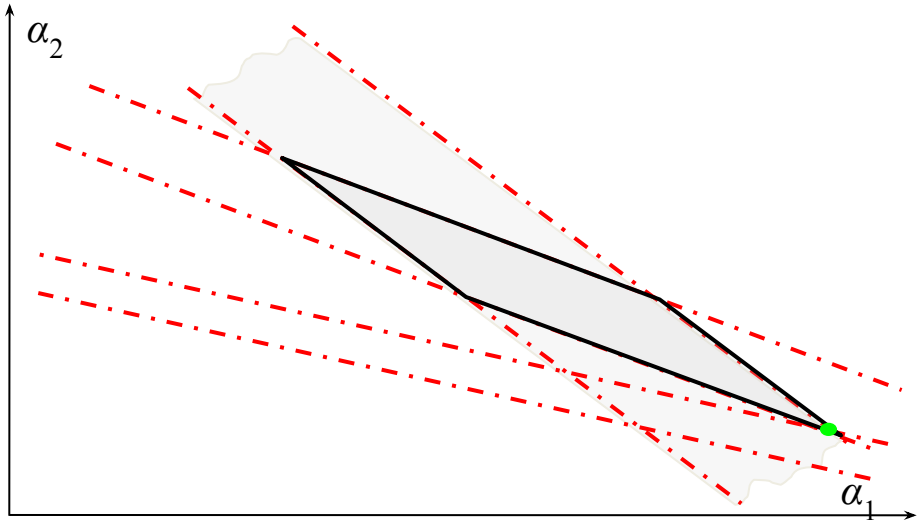
Интервальный подход: построение зависимости

- Мера совместности выборки – минимальный предельный уровень погрешности $\Delta_{\min}$

Пространство (x, y)



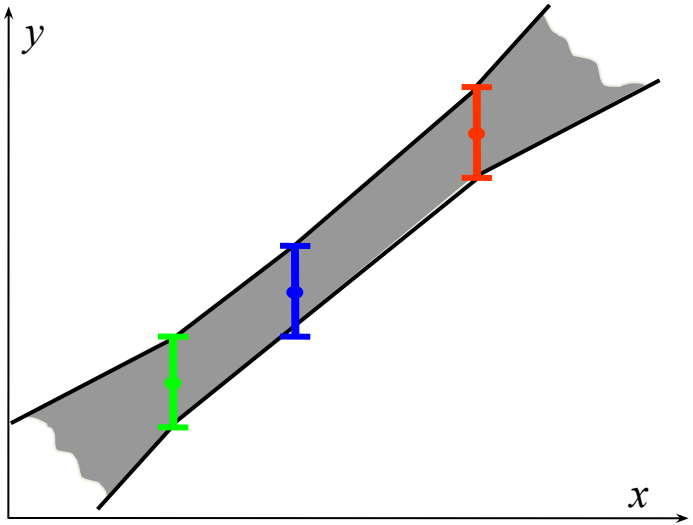
Пространство (α_1, α_2)



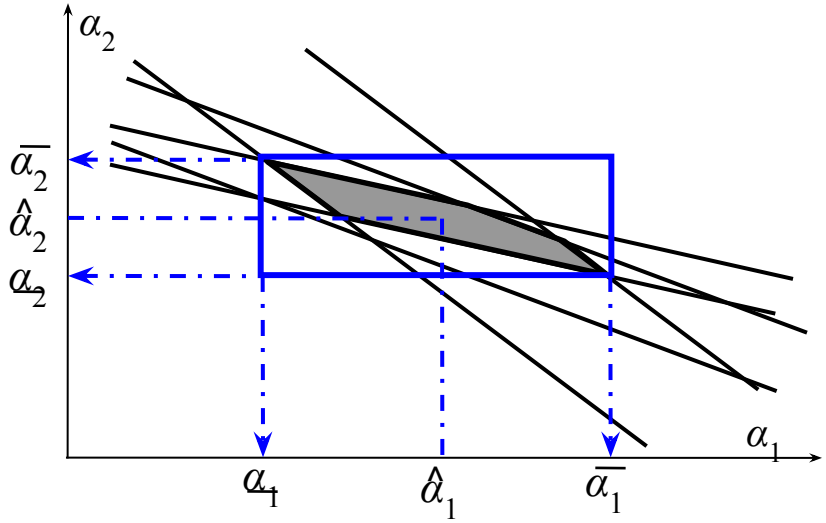
Интервальный подход: построение зависимости

- Оценка параметров модели $y = \alpha_1 + \alpha_2 x$

Пространство (x, y)



Пространство (α_1, α_2)



Интервальный подход: идея

СИБИРСКИЙ МАТЕМАТИЧЕСКИЙ ЖУРНАЛ
Том III, № 5 Сентябрь — Октябрь 1962 г.

Л. В. КАНТОРОВИЧ

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. Бóльшая 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 искомых величин или, если говорить о нечисловых величинах, областей расположения искомых и наблюдаемых величин; иначе говоря возникает задача возможно более точного описания расположения этих величин в соответствующих пространствах их зна-

Канторович Л.В. // Сиб. мат. журнал. – 1962 3(5) , 701-709. http://www.nsc.ru/interval/Library/Thematic/DataProcs/Kantorovich62.pdf

Интервальный подход: имена и годы

Интервальный анализ, оценивание при ограниченной неопределённости, bounding approach, unknown-but-bounded errors, set membership estimation

- 1962 Л.В. Канторович
- 1970 С.И. Спивак и др.
- 1982 G. Belforte, M. Milanese et al.
- 1983 Н.М. Оскорбин и др.
- 1986 J.P. Norton
- 1987 С.И. Кумков и др.
- 1987 E. Walter, H. Piet-Lahanier
- 1989 А.П. Вощини и др.
- 1993 P.L. Combettes
- 1995 С.П. Шарый
- 2000 О.Е. Родионова, А.Л. Померанцев

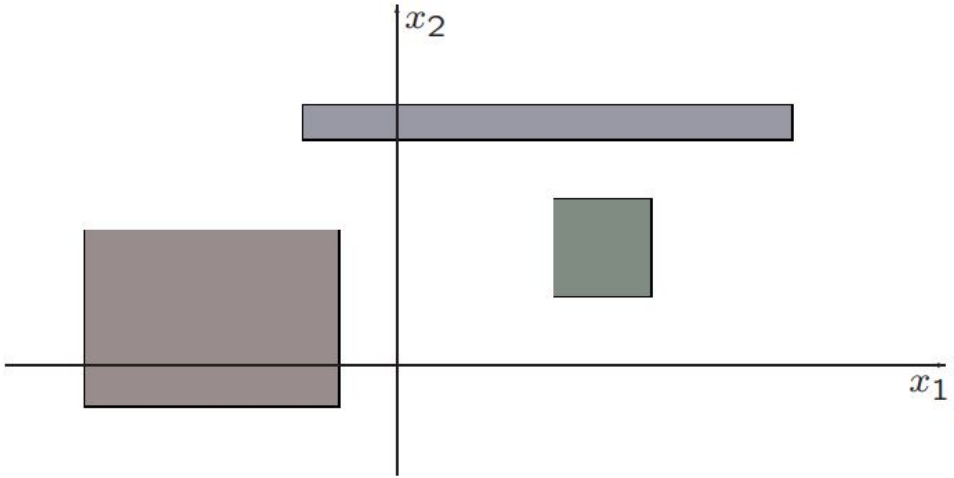
Интервальная математика

Интервалы и интервальные векторы

$[1, 2], \quad [1000, 1003], \quad \ldots$

$$\big([1, 2],\ [1000, 1003]\big)$$

$$\begin{pmatrix} [1, 2] \\ [1000, 1003] \end{pmatrix}$$



Интервальный анализ и его приложения http://www.nsc.ru/interval/

Шарый С.П. Интервальный анализ: современное состояние и перспективы http://www.nsc.ru/interval/Education/Manuals/SShary-RASession.pdf

Шарый С.П. Конечномерный интервальный анализ. – Новосибирск: XYZ, 2021. http://www.nsc.ru/interval/Library/InteBooks/SharyBook.pdf

Характеристики интервалов и расстояние между интервалами

$\underline{x}, \quad \overline{x}$ — нижний и верхний концы

$\operatorname{mid} x = \frac{1}{2}(\overline{x} + \underline{x})$ — середина

$\operatorname{wid} x = \overline{x} - \underline{x}$ — ширина

$\operatorname{rad} x = \frac{1}{2}(\overline{x} - \underline{x})$ — радиус

$|x| = \max\{|\underline{x}|, |\overline{x}|\}$ — абсолютное значение (модуль)

$$\operatorname{dist}(x, y) = \max\left\{ |\underline{x} - \underline{y}|, \ |\overline{x} - \overline{y}| \right\}$$

Операции над интервалами

$$x \in [1, 2] \qquad y \in [3, 7]$$

$$x + y \in ?$$

$$1 \leq x \leq 2 \qquad 3 \leq y \leq 7$$

$$x + y \in [4, 9] = [1 + 3, 2 + 7]$$

Классическая интервальная арифметика

— алгебраическая система, образованная интервалами $\boldsymbol{x} = [\underline{\boldsymbol{x}}, \overline{\boldsymbol{x}}] \subset \mathbb{R}$ так, что

$$\boldsymbol{x} \star \boldsymbol{y} = \left\{ x \star y \mid x \in \boldsymbol{x},\, y \in \boldsymbol{y} \right\} \qquad \text{для} \qquad \star \in \{+, -, \cdot, /\}$$

$$\begin{aligned}
\boldsymbol{x} + \boldsymbol{y} &= \left[\underline{\boldsymbol{x}} + \underline{\boldsymbol{y}},\ \overline{\boldsymbol{x}} + \overline{\boldsymbol{y}}\right] \\
\boldsymbol{x} - \boldsymbol{y} &= \left[\underline{\boldsymbol{x}} - \overline{\boldsymbol{y}},\ \overline{\boldsymbol{x}} - \underline{\boldsymbol{y}}\right] \\
\boldsymbol{x} \cdot \boldsymbol{y} &= \left[\min\{\underline{\boldsymbol{x}}\,\underline{\boldsymbol{y}}, \underline{\boldsymbol{x}}\,\overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\,\underline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\,\overline{\boldsymbol{y}}\},\ \max\{\underline{\boldsymbol{x}}\,\underline{\boldsymbol{y}}, \underline{\boldsymbol{x}}\,\overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\,\underline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\,\overline{\boldsymbol{y}}\}\right] \\
\boldsymbol{x} / \boldsymbol{y} &= \boldsymbol{x} \cdot \left[1/\overline{\boldsymbol{y}},\ 1/\underline{\boldsymbol{y}}\right] \qquad \text{для} \quad \boldsymbol{y} \not\ni 0
\end{aligned}$$

IEEE 1788-2015 - IEEE Standard for Interval Arithmetic https://standards.ieee.org/standard/1788-2015.html
IEEE 1788.1-2017 - IEEE Standard for Interval Arithmetic (Simplified) https://standards.ieee.org/standard/1788_1-2017.html

Интервалы как средство учета погрешностей вычислений

$$x \in [1.1, 1.2] \qquad y \in [5.3, 5.4]$$

$$x + y \in [6.4, 6.6] = [1.1 + 5.3, 1.2 + 5.4]$$

Аналогично и с другими арифметическими операциями . . .

Результаты такого вычисления погрешностей
можно использовать далее
в цепочке вычислений!

Интервалы как средство для работы с неопределенностями

«Неопределённость» — состояние частичного знания о рассматриваемой величине

Модели неопределённости

- вероятностная (стохастическая)
- интервальная (ограниченная по величине)
- нечёткая (размытая)

Интервальное описание неопределённости — наиболее «скупое», но математический аппарат для его обработки наиболее развит.

Интервальные линейные системы уравнений

$$\begin{cases} \boldsymbol{a}_{11}x_1 + \boldsymbol{a}_{12}x_2 + \ldots + \boldsymbol{a}_{1n}x_n = \boldsymbol{b}_1, \\ \boldsymbol{a}_{21}x_1 + \boldsymbol{a}_{22}x_2 + \ldots + \boldsymbol{a}_{2n}x_n = \boldsymbol{b}_2, \\ \quad\vdots \qquad\qquad \ddots \qquad\qquad \vdots \\ \boldsymbol{a}_{n1}x_1 + \boldsymbol{a}_{n2}x_2 + \ldots + \boldsymbol{a}_{nn}x_n = \boldsymbol{b}_n, \end{cases}$$

или, кратко,

$$\boldsymbol{A}x = \boldsymbol{b}$$

с интервальными матрицей $\boldsymbol{A} = (\boldsymbol{a}_{ij})$ и вектором $\boldsymbol{b} = (\boldsymbol{b}_i)$.

Интервальные линейные системы уравнений

$$\boldsymbol{A}x = \boldsymbol{b}$$

— семейство точечных линейных систем $Ax = b$ с $A \in \boldsymbol{A}$ и $b \in \boldsymbol{b}$.

Множество решений
интервальной линейной системы уравнений —

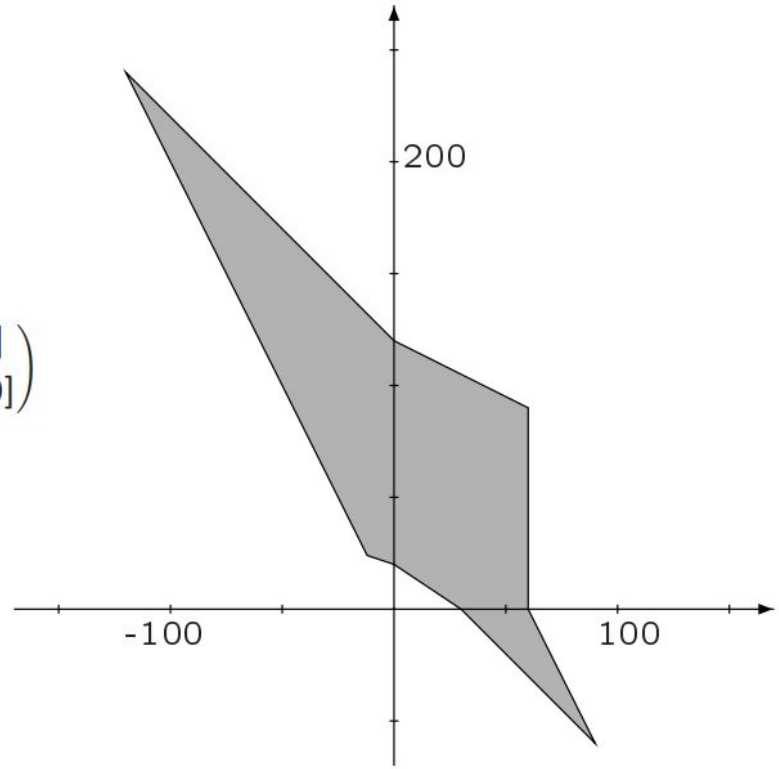
$$\Xi(\boldsymbol{A}, \boldsymbol{b}) = \left\{ x \in \mathbb{R}^n \mid (\exists\, A \in \boldsymbol{A})(\exists b \in \boldsymbol{b})(\, Ax = b\,) \right\}$$

Также *объединённое множество решений* . . .

Интервальные линейные системы уравнений.

Пример — система Хансена

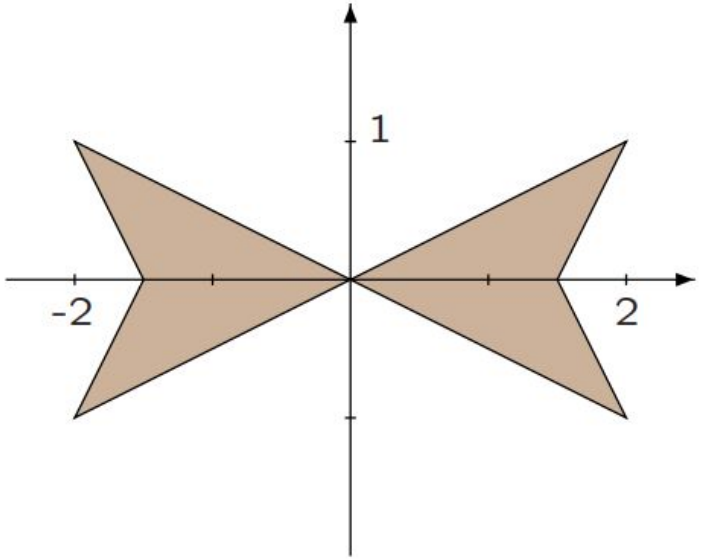
$$\begin{pmatrix} [2,3] & [0,1] \\ [1,2] & [2,3] \end{pmatrix} x = \begin{pmatrix} [0,120] \\ [60,240] \end{pmatrix}$$



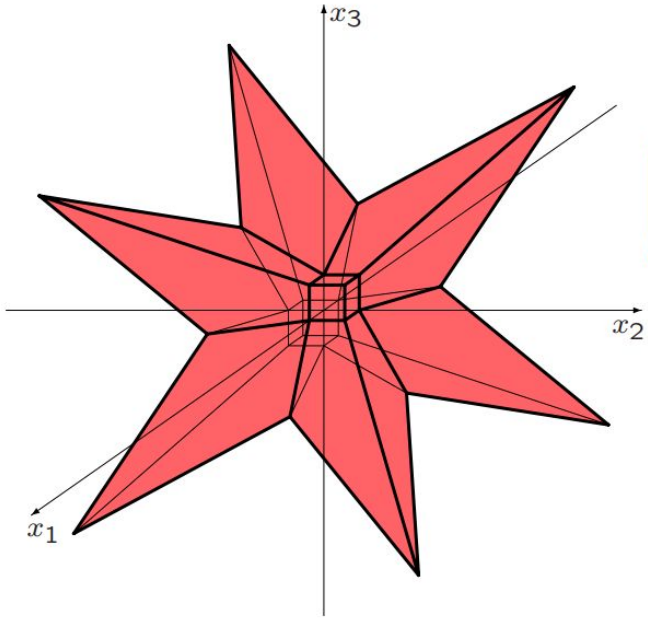
Интервальные линейные системы уравнений.

Пример — почти несвязное множество решений

$$\begin{pmatrix} [2,4] & [-1,1] \\ [-1,1] & [2,4] \end{pmatrix} x = \begin{pmatrix} [-3,3] \\ 0 \end{pmatrix}$$

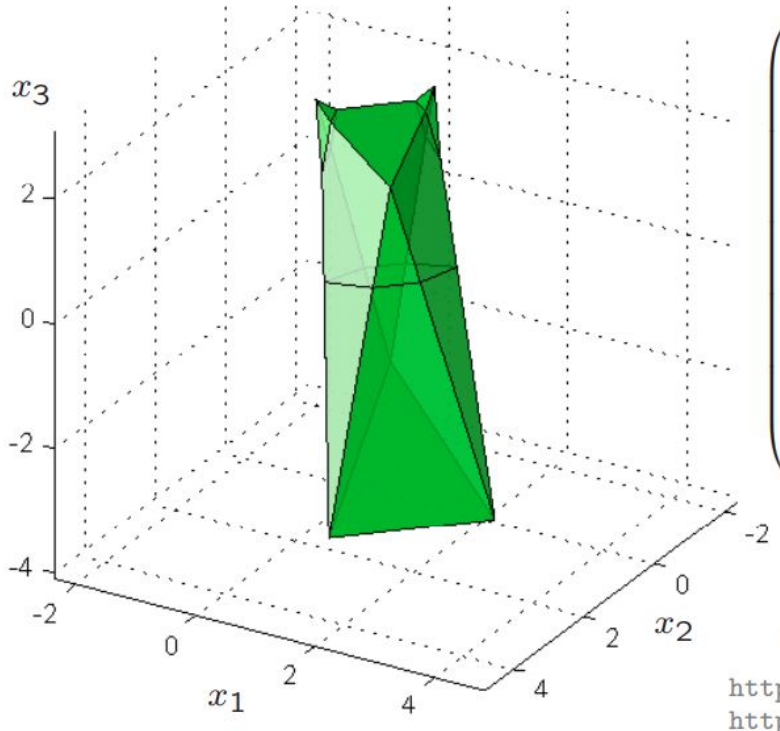


Интервальные линейные системы уравнений.
Пример — система Ноймайера



$$\begin{pmatrix} 3.5 & [0,2] & [0,2] \\ [0,2] & 3.5 & [0,2] \\ [0,2] & [0,2] & 3.5 \end{pmatrix} x = \begin{pmatrix} [-1,1] \\ [-1,1] \\ [-1,1] \end{pmatrix}$$

Интервальные линейные системы уравнений.
Пример — “бесхвостый котик”



$$\begin{pmatrix} [0.8,1.2] & [0.8,1.2] & 1 \\ [0.8,1.2] & [1.8,2.2] & 1 \\ [0.8,1.2] & [2.8,3.2] & 1 \\ [1.8,2.2] & [0.8,1.2] & 1 \\ [1.8,2.2] & [1.8,2.2] & 1 \\ [1.8,2.2] & [2.8,3.2] & 1 \\ [2.8,3.2] & [0.8,1.2] & 1 \\ [2.8,3.2] & [1.8,2.2] & 1 \\ [2.8,3.2] & [2.8,3.2] & 1 \end{pmatrix} x = \begin{pmatrix} [1,3] \\ [2,4] \\ [3,5] \\ [2,4] \\ [3,5] \\ [4,6] \\ [3,5] \\ [4,6] \\ [5,7] \end{pmatrix}$$

Пакет IntLinIncR3, автор Ирина Шарая

http://www.nsc.ru/interval/Programing

http://www.nsc.ru/interval/sharaya/irash.html

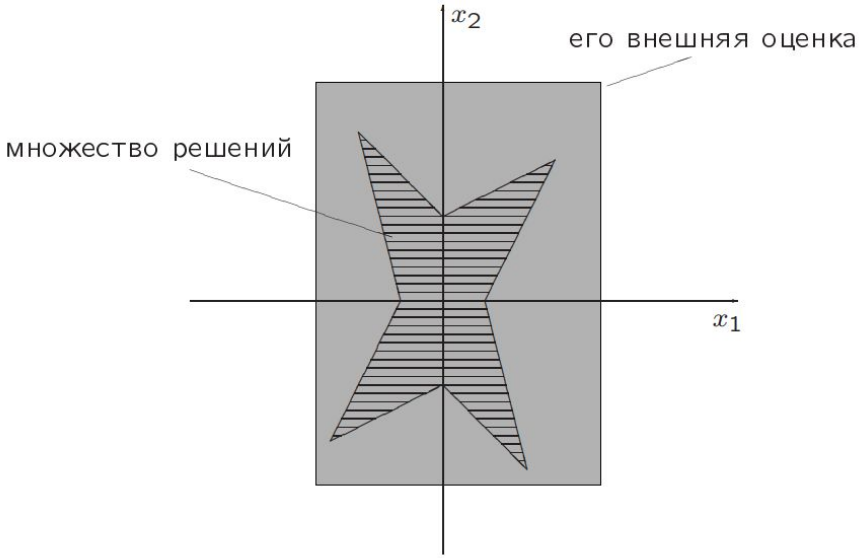
Интервальные линейные системы уравнений.

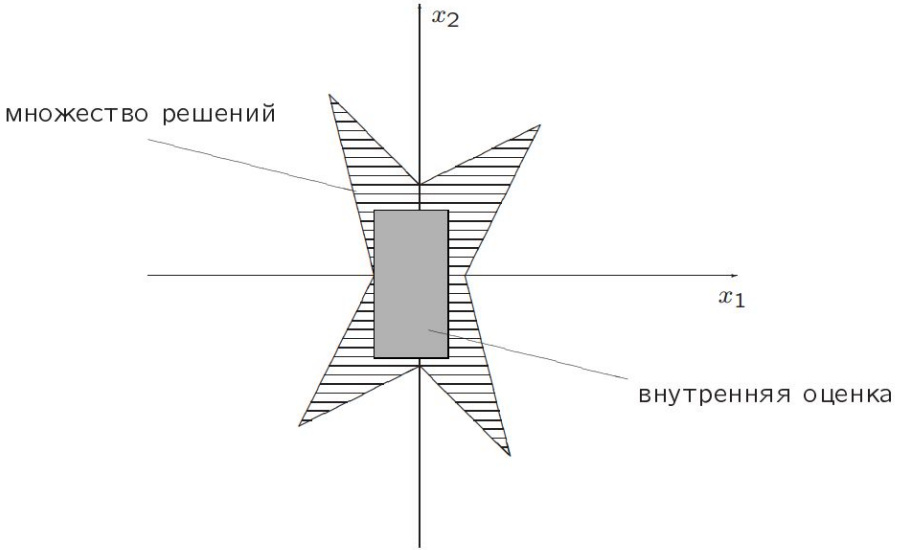
Точное и полное описание множества решений

- практически невозможно в силу огромной сложности,
- реально не нужно.

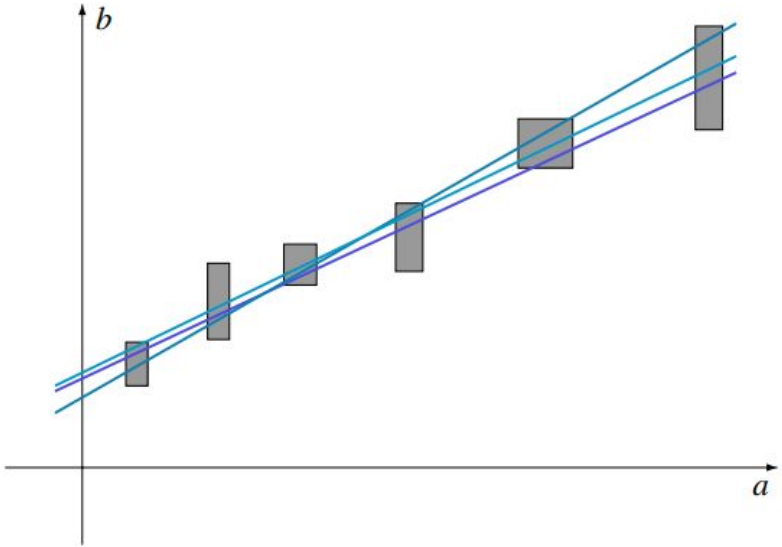
В большинстве случаев достаточно знать *приближённое описание*, или *оценку* множества решений более простыми множествами (т.е. имеющими меньшую конструктивную сложность).

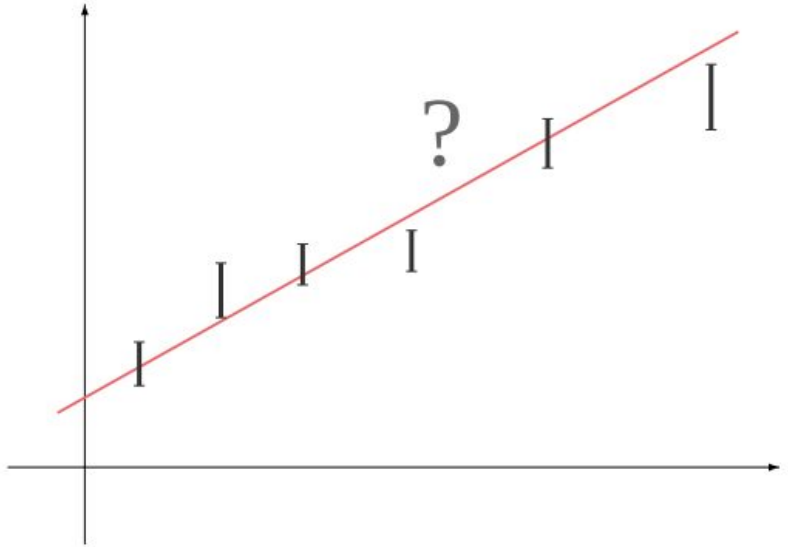
Внешняя и внутренняя задачи





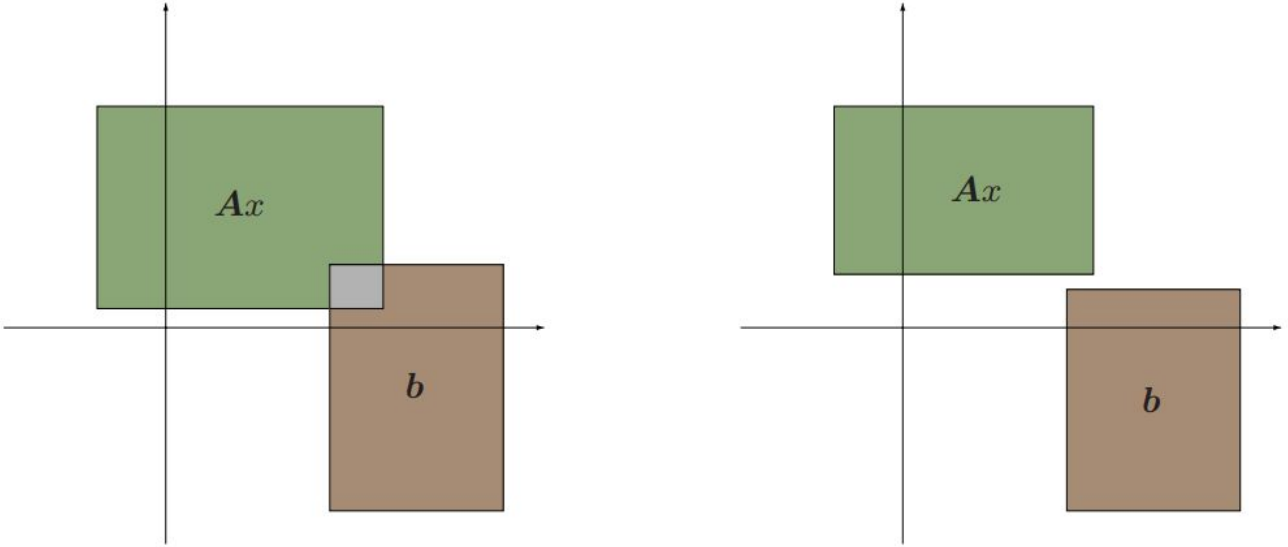
Восстановление зависимостей по неточным данным





Характеризация точек множества решений

«Мерой согласования» параметров x и данных $\boldsymbol{A}$, $\boldsymbol{b}$ может служить функционал, характеризующий взаимное расположение $\boldsymbol{A}x$ и $\boldsymbol{b}$, меру их пересечения или непересечения друг с другом:

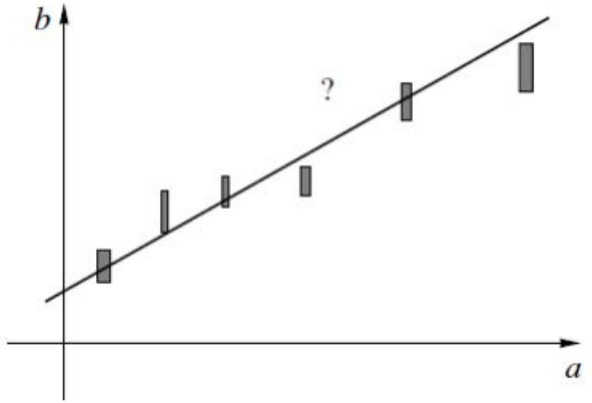


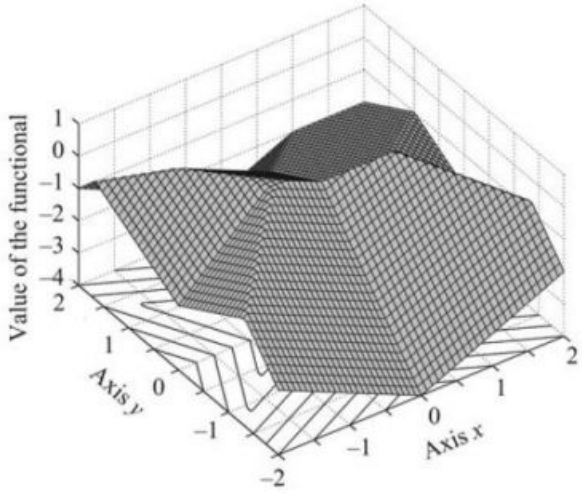
— *аналог невязки*

Метод максимума согласования

- Мера совместности := распознающий функционал Uss

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



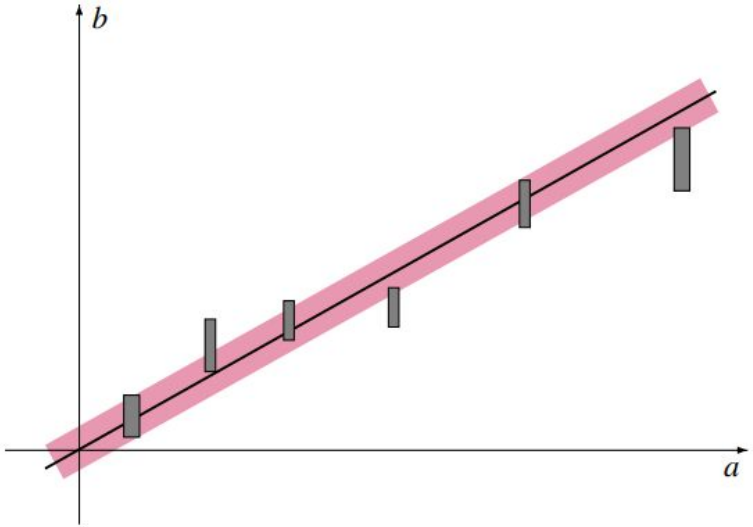


Shary S.P. *Solvability of Interval Linear Equations and Data Analysis under Uncertainty* // Automation and Remote Control (2012) 73 (2), 310–322.

Метод максимума согласования

Ещё одна практическая интерпретация:

arg max Uss даёт параметры такой регрессионной линии, которую следует наименьшим образом расширить до «регрессионной полосы», уже пересекающей все брусы данных



Основная теорема интервальной арифметики

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — рациональная функция от аргументов $x_1, x_2, \ldots, x_n$.

Если для некоторого бруса $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ определён результат $\boldsymbol{f}_\natural(\boldsymbol{x})$ подстановки вместо аргументов функции $f(x)$ интервалов $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ и выполнения всех действий над ними по правилам интервальной арифметики, то

$$\{ f(x_1, \ldots, x_n) \mid x_1 \in \boldsymbol{x}_1, \ldots, x_n \in \boldsymbol{x}_n \} \subseteq \boldsymbol{f}_\natural(\boldsymbol{x}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n),$$

т. е. результат интервального оценивания $\boldsymbol{f}_\natural(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ содержит множество значений функции $f(x_1, x_2, \ldots, x_n)$ на $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$.

— естественное интервальное расширение

Особенности интервальной арифметики

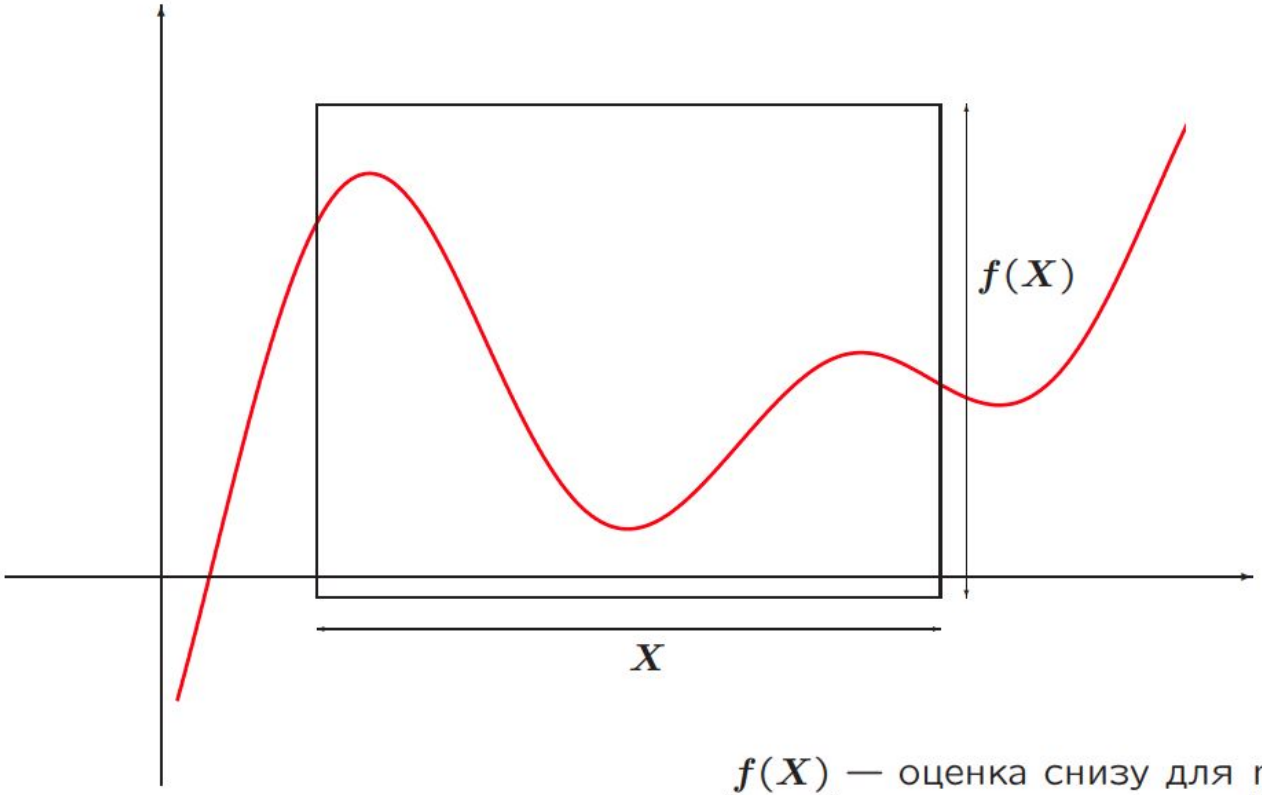
$$f(x) = \frac{x}{x+y} \quad \text{для } x \in [1,2],\ y \in [3,4]$$

$$\frac{[1,2]}{[1,2]+[3,4]} = \frac{[1,2]}{[4,6]} = \left[\frac{1}{6}, \frac{2}{4}\right] = [0.166\ldots, 0.5]$$

$$g(x) = \frac{1}{1+y/x} \quad \text{для } x \in [1,2],\ y \in [3,4]$$

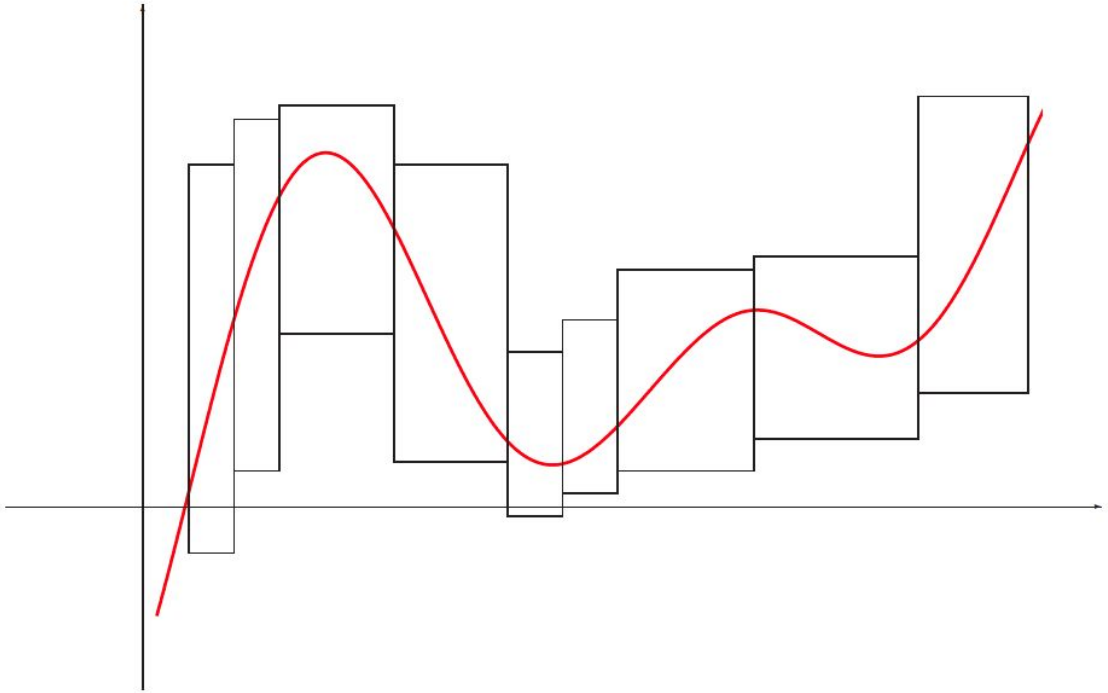
$$\frac{1}{1+\frac{[3,4]}{[1,2]}} = \frac{1}{1+\left[\frac{3}{2}, 4\right]} = \frac{1}{\left[\frac{5}{2}, 5\right]} = \left[\frac{1}{5}, \frac{2}{5}\right] = [0.2, 0.4]$$

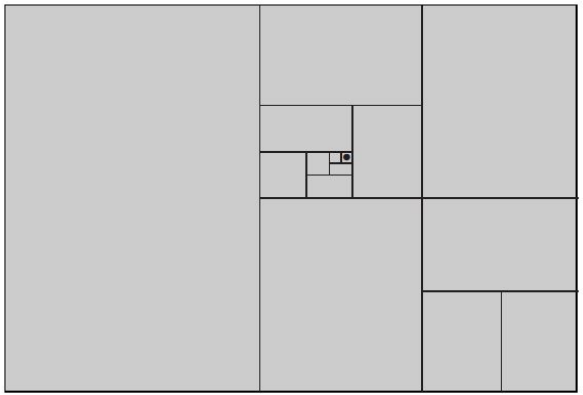
Интервальное расширение функции



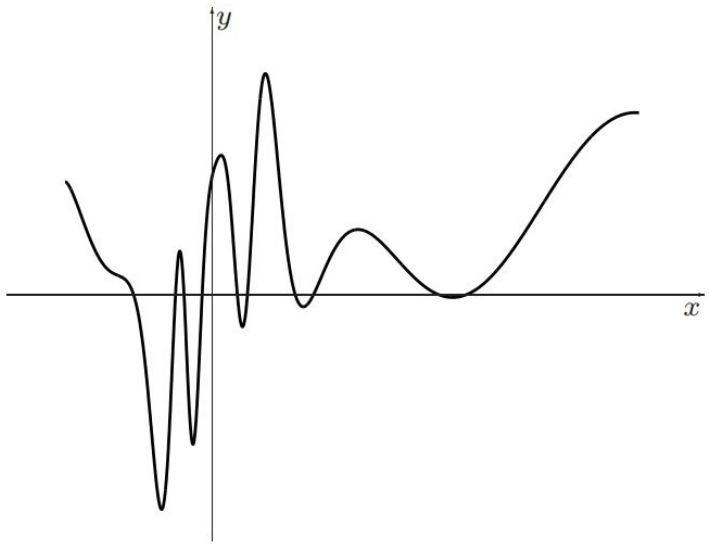
$\underline{f(X)}$ — оценка снизу для $\min\limits_{x \in X} f(x)$

Интервальный подход к глобальной оптимизации





Глобальное решение систем уравнений



Найти решения системы уравнений

$$\begin{cases} F_1(x_1, x_2, \ldots, x_n) = 0, \\ F_2(x_1, x_2, \ldots, x_n) = 0, \\ \vdots \quad \ddots \quad \vdots \quad \vdots \\ F_n(x_1, x_2, \ldots, x_n) = 0, \end{cases}$$

или, кратко,

$$F(x) = 0,$$

где $F(x) = \big(F_1(x), F_2(x), \ldots, F_n(x)\big)^\top$, $x = (x_1, x_2, \ldots, x_n)^\top$.

Найти *все* решения системы уравнений

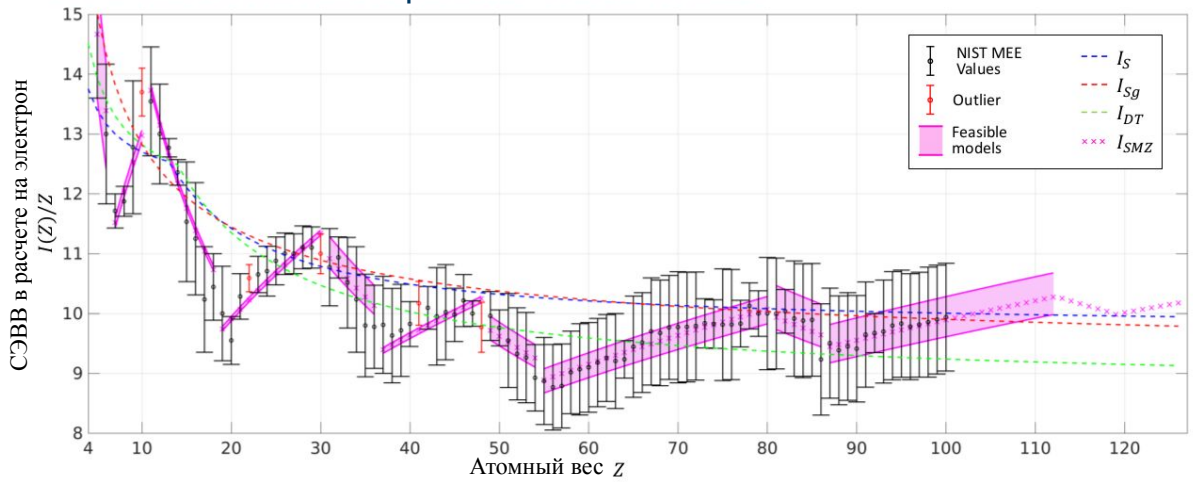
$$F(x) = 0$$

на данном множестве $D \subseteq \mathbb{R}^n$, определив для них гарантированные двусторонние границы.

Пример

Полу-эмпирическая формула средней энергии возбуждения

Экспериментальные значения СЭВВ



Известные формулы для СЭВВ

- Формула Штернхеймера[4] (1963):

$$\frac{I_S}{Z} = \begin{cases} 12 + 7/Z \text{ eV}, & Z < 13; \\ 9.76 + 58.8Z^{-1.19} \text{ eV}, & Z \geq 13, \end{cases}$$

- Формула Далтона – Тёрнера[5] (1968):

$$I_{DT} = \begin{cases} 11.2 + 11.7Z \text{ eV}, & Z \leq 13; \\ 52.8 + 8.71Z \text{ eV}, & Z > 13. \end{cases}$$

- Формула Сегрэ[6] (1977):

$$I_{Sg} = 9.1\left(Z + 1.9Z^{1/3}\right) \text{ eV}, Z \geq 4$$

Предлагаемая модель СЭВВ

- Сохраняется общая структура модели Томаса – Ферми

$$I = C_n^{\alpha} Z^{\alpha},$$

но

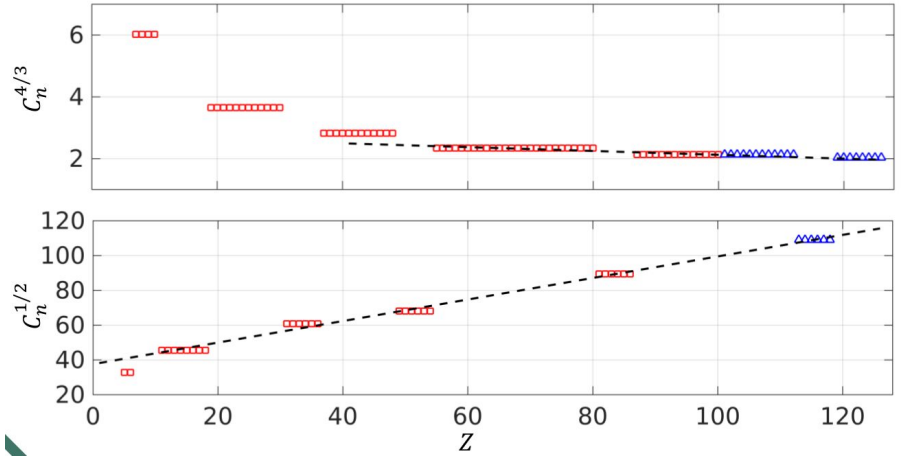
$$\alpha = \begin{cases} {}^4/_3 & \text{для } s-, d-, f- \text{ и } g-\text{блоков}; \\ {}^1/_2 & \text{для } p-\text{блоков}. \end{cases}$$

Smolar V.A., Maglevanny I.I., Zhilin S.I. / https://www.researchgate.net/publication/342106802_Semi-empirical_formula_for_mean_excitation_energy_interval_approach

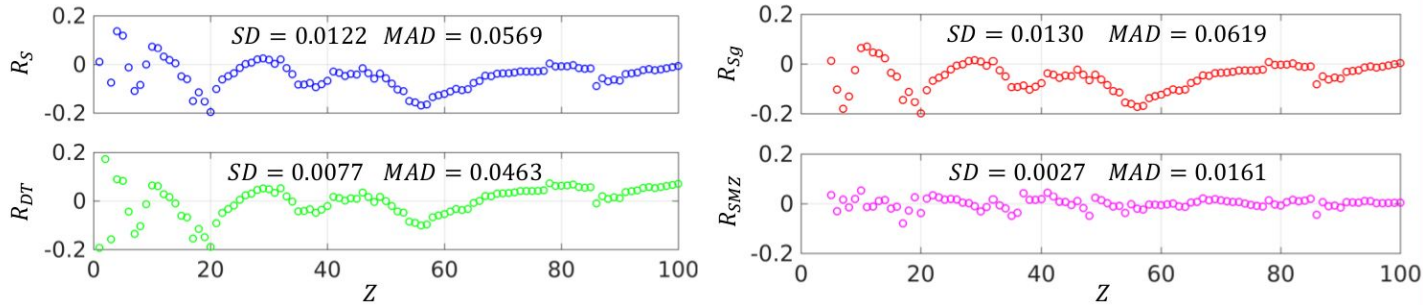
Полу-эмпирическая формула средней энергии возбуждения

Оценки коэффициентов

Пе-риод	Блок	Электронная конфигурация	Интервалы Z	α	Оценки $\left[\underline{C_n^\alpha}, \overline{C_n^\alpha}\right]$	Оценки C_n^α
1	s	H, He	1(H) – 2(He)			
2	s	[He]$2s^1 - 2s^2$	3(Li) – 4(Be)			
	p	[He]$2s^2 2p^{1-2}$	5(B) – 6(C)	1/2	[30.4, 34.7]	32.8
	p	[He]$2s^2 2p^{3-6}$	7(N) – 10(Ne)	4/3	[5.97, 6.06]	6.02
3	s-p	[Ne]$3s^{1-2}3p^{1-6}$	11(Na) – 18(Ar)	1/2	[45.5, 45.8]	45.5
4	s-d	[Ar]$3d^{1-10}4s^{1-2}$	19(K) – 30(Zn)	4/3	[3.64, 3.67]	3.65
	p	[Zn]$4p^{1-6}$	31(Ga) – 36(Kr)	1/2	[59.9, 62.8]	60.8
5	s-d	[Kr]$4d^{1-10}5s^{1-2}$	37(K) – 48(Zn)	4/3	[2.80, 2.83]	2.82
	p	[Cd]$5p^{1-6}$	49(Ga) – 54(Kr)	1/2	[66.9, 69.7]	68.0
6	s-f-d	[Xe]$4f^{1-14}5d^{1-10}6s$	55(Cs) – 80(Hg)	4/3	[2.28, 2.39]	2.34
	p	[Hg]$6p^{1-6}$	81(Tl) – 86(Rn)	1/2	[87.6, 94.2]	89.4
7	s-f-d	[Rn]$5f^{1-14}6d^{1-10}6s$	87(Fr) – 112(Cn)	4/3	[2.07, 2.22]	2.13
	p	[Cn]$7p^{1-6}$	113(Uut) – 118(Uuo)	1/2		≈109
8	s-g-f-d	[Uuo]$5g^{1-18}6f^{1-14}7d^{1-14}8s$	119(Uue) – 126(Ubh)	4/3		≈2.03



Остаточные отклонения



Софт

- Интервальная арифметика на C++
  https://github.com/nehmeier/libieeep1788
- PyInterval — интервальная арифметика на Python
  https://pypi.org/project/pyinterval/
- Python-обертка на Ibex (контракторы и т.п.)
  https://github.com/codac-team/pyIbex
- IntervalArithmetic.jl —интервальная арифметика на Julia
  https://github.com/JuliaIntervals/IntervalArithmetic.jl
- Интервальная арифметика на Octave
  https://octave.sourceforge.io/interval/
- JInterval — интервальная арифметика на Java
  https://github.com/jinterval/jinterval
- ПО для интервального анализа С.П. Шарого
  http://www.nsc.ru/interval/Programing/
- Примеры анализа интервальных данных в Octave (сборник jupyter-ноутбуков)
  https://github.com/szhilin/octave-interval-examples/

Примеры анализа интервальных данных в Octave

master · octave-interval-examples / SteamGenerator.ipynb · Go to file

szhilin Fix convexhull error in old versions of Octave — Latest commit 8a7407f on Apr 14 · History

1 contributor

765 lines (765 sloc) | 128 KB · Raw · Blame

Производительность парогенератора

Этот простой пример построения модели, описывающей изменение количества вырабатываемого парогенератором пара в зависимости от подаваемого на вход топлива, приведён в книге [1]. Он иллюстрирует основные шаги построения и анализа интервальных регрессионных моделей при точно задаваемых входных переменных и наличии интервальной неопредлённости в выходной переменной.

Набор данных

Рассмотрим процедуру построения и анализа зависимости по интервальным данным на примере исследования парогенератора типа БК-3-210, широко используемого на ТЭЦ для выработки пара высокого давления.

Основной эксплуатационной характеристикой агрегата является зависимость его производительности y, измеряемой как расход пара на выходе агрегата, от расхода топлива x на входе агрегата.

В результате испытаний парогенератора после проведения регламентных работ были получены данные трёх измерений, причём расход пара измерялся с 5%-й относительной ошибкой. Полученные данные сведены в таблицу.

i	x	y	e	$y^- = y - e$	$y^+ = y + e$
1	12	128	6	122	134
2	16	180	9	171	189
3	20	230	11	119	241

Коридор совместных зависимостей

Информационное множество задачи определяется в пространстве параметров. Каждая его точка (β_1, β_2) задаёт зависимость в пространстве переменных (x, y). Множество всех таких моделей именуется **коридором совместных зависимостей**.

In [9]:
```
## Графическое представление коридора совместных зависимостей для модели y = beta1 + beta2 * x
figure(1, 'position',[0, 0, 800, 600]);
xlimits = [0 25];
ir_plotmodelset(irp_steam, xlimits)     # коридор совместных зависимостей
hold on
ir_scatter(irp_steam,'bo')              # интервальные измерения
ir_plotline(b_maxdiag, xlimits, 'r-')   # зависимость с параметрами, оцененными как центр наибольшей диагонали ИМ
#ir_plotline(b_gravity, xlim, 'b--')     # зависимость с параметрами, оцененными как центр тяжести ИМ
#ir_plotline(b_lsm, xlim, 'b--')         # зависимость с параметрами, оцененными МНК
#ir_scatter(ir_problem(Xp,ypmid,yprad),'ro')

grid on
set(gca, 'fontsize', 12)
xlabel('Fuel consumtion')
ylabel('Steam quantity')
title('Set of models compatible with data and constraints')
```