

Тема X-1. Обработка и анализ данных с интервальной неопределённостью.

А.Н. Баженов

Санкт-Петербургский политехнический университет Петра Великого

a_bazhenov@inbox.ru

21.09.2021

Интервальный анализ и его методы

Интервал — замкнутый отрезок вещественной оси, а **интервальная неопределенность** — состояние неполного знания об интересующей нас величине, когда известна лишь ее принадлежность **некоторому интервалу**.

Интервальный анализ — отрасль математического знания, исследующая задачи с интервальными неопределенностями и методы их решения.

Поиск множества, удовлетворяющего **постановке задачи**.

Понятие интервала

Интервалом $[a, b]$ вещественной оси R называется множество всех чисел, расположенных между заданными числами включая их самих,
т.е.

$$[a, b] := \{x \in \mathbb{R} \mid a \leq x \leq b\}$$

При этом a и b называются концами интервала.

Интервальный анализ и его методы

«... В большинстве случаев некорректно говорить о «решении интервальных уравнений» (систем уравнений, неравенств и т. п.) вообще.

Правильнее вести речь о решении тех или иных **постановок задач**, связанных с интервальными уравнениями (системами уравнений, неравенств и т. п.). В свою очередь, формулировка постановки интервальной задачи подразумевает указание, по крайней мере, **множества решений задачи и способа его оценивания**.

С.П.Шарый. Конечномерный интервальный анализ, 2021

Обработка и анализ данных с интервальной неопределённостью.

ПЛАН

Общий план

- Общие понятия
- Обработка константы (физической величины)
- Задача восстановления зависимостей

Теория:

А.Н. Баженов, С.И. Жилин, С.И. Кумков, С.П. Шарый.
Обработка и анализ данных с интервальной неопределённостью. РХД.
Серия «Интервальный анализ и его приложения». Ижевск. 2021. с.200.

Общие понятия.

Отношения между интервалами.

Интервалы являются множествами, составленными из вещественных чисел, и неудивительно, что большую роль для них играют теоретико-множественные отношения и операции (объединение, пересечение и др.). Особенно важно отношение включения одного интервала в другой:

$$\mathbf{a} \subseteq \mathbf{b} \text{ равносильно тому, что } \underline{\mathbf{a}} \geq \underline{\mathbf{b}} \text{ и } \bar{\mathbf{a}} \leq \bar{\mathbf{b}}. \quad (1)$$

Отношение включения является частичным порядком и превращает множество интервалов в частично упорядоченное множество, важную и хорошо изученную математическую структуру.

Отношения между интервалами.

Помимо порядка по включению на множестве интервалов огромную роль играют также другие отношения, которые обобщают хорошо известный порядок \leq на вещественной оси \mathbb{R} .

Фундаментальным фактом является то, что порядок \leq между вещественными числами может быть обобщен на интервалы многими осмысленными способами (и даже бесконечно большим числом способов). Значительная часть получающихся при этом отношений на \mathbb{IR} не являются полноценными порядками.

Отношения между интервалами.

Помимо порядка по включению на множестве интервалов огромную роль играют также другие отношения, которые обобщают хорошо известный порядок \leq на вещественной оси \mathbb{R} .

Фундаментальным фактом является то, что порядок \leq между вещественными числами может быть обобщен на интервалы многими осмысленными способами (и даже бесконечно большим числом способов). Значительная часть получающихся при этом отношений на \mathbb{IR} не являются полноценными порядками.

Отношения между интервалами.

Важную роль играет следующее упорядочение

Definition

Для интервалов $\mathbf{a}, \mathbf{b} \in \mathbb{IR}$ условимся считать, что \mathbf{a} не превосходит \mathbf{b} и писать « $\mathbf{a} \leq \mathbf{b}$ » тогда и только тогда, когда $\underline{\mathbf{a}} \leq \underline{\mathbf{b}}$ и $\bar{\mathbf{a}} \leq \bar{\mathbf{b}}$.

Интервал называется *неотрицательным*, т. е. « ≥ 0 », если неотрицательны оба его конца. Интервал называется *неположительным*, т. е. « ≤ 0 », если неположительны оба его конца.

Теоретико-множественные операции между интервалами.

Если интервалы \mathbf{a} и \mathbf{b} имеют непустое пересечение, т. е. $\mathbf{a} \cap \mathbf{b} \neq \emptyset$, то можно дать простые выражения для результатов теоретико-множественных операций пересечения и объединения через концы этих интервалов

$$\mathbf{a} \cap \mathbf{b} = [\max\{\underline{\mathbf{a}}, \underline{\mathbf{b}}\}, \min\{\bar{\mathbf{a}}, \bar{\mathbf{b}}\}], \quad \mathbf{a} \cup \mathbf{b} = [\min\{\underline{\mathbf{a}}, \underline{\mathbf{b}}\}, \max\{\bar{\mathbf{a}}, \bar{\mathbf{b}}\}]. \quad (2)$$

Если же $\mathbf{a} \cap \mathbf{b} = \emptyset$, т. е. интервалы \mathbf{a} и \mathbf{b} не имеют общих точек, то эти равенства уже неверны.

Теоретико-множественные операции между интервалами.

Обобщением операций пересечения и объединения являются операции взятия минимума и максимума относительно включения « \subseteq »:

$$\mathbf{a} \wedge \mathbf{b} = [\max\{\underline{a}, \underline{b}\}, \min\{\bar{a}, \bar{b}\}], \quad \mathbf{a} \vee \mathbf{b} = [\min\{\underline{a}, \underline{b}\}, \max\{\bar{a}, \bar{b}\}]. \quad (3)$$

Они также понадобятся нам при обработке интервальных измерений.

Первая из этих операций, « \wedge », не всегда выполнима во множестве обычных интервалов, но это затруднение преодолевается путём расширения множества интервалов специальными элементами — неправильными интервалами.

Измерения

Definition

Измерением (замером, наблюдением) будем называть измеренное значение величины.

По способу получения результата измерения все процессы измерения разделяются на *прямые, косвенные и совокупные*.

Измерения и их результаты

- Погрешности квантования
- Неопределённость измерения нуля
- Агрегирование результатов многократных наблюдений

Агрегирование результатов многократных наблюдений.

Во многих практических ситуациях измерение интересующей нас величины выполняется для надёжности многократно. Тем не менее, повторные измерения над одними и теми же явлениями не показывают разумное (в пределах точности измерений) совпадение результатов.

Приняв все необходимы меры предосторожности, обеспечив постоянные условия измерения, мы всё равно не получаем разумно согласующихся друг с другом результатов.

Скажем, в промышленности, как бы тщательно ни был отрегулирован измерительный прибор, колебания в его показаниях не могут быть уменьшены ниже некоторого предела.

Агрегирование результатов многократных наблюдений.

В этих условиях результатом серии повторяющихся измерений можно взять интервал от минимального до максимального из полученных результатов, т. е. агрегировать (объединить) результаты отдельных измерений.

Математически, если результаты повторных измерений величины равны x_1, x_2, \dots, x_n , то интервальным результатом следует взять

$$x = \left[\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i \right].$$

Будем называть этот способ получения интервального результата измерения *агрегированием*.

Агрегирование результатов многократных наблюдений.

Используя операции взятия *интервальной оболочки* множества и *максимума по включению* этот результат можно записать следующим равносильным образом:

$$x = \square\{x_1, x_2, \dots, x_n\}$$

или

$$x = \bigvee_{1 \leq i \leq n} x_i.$$

Эти представления хороши тем, что могут быть обобщены на более сложные случаи.

Модель погрешности наблюдения.

Интервалы в результатах измерений могут возникать различным способом. Они могут получаться сразу, в виде готовых интервалов, но могут возникать в результате коррекции точечных результатов.

Один из распространённых способов получения интервальных результатов в первичных измерениях — это «обинтерваливание» точечных значений, когда к точечному базовому значению \hat{x} прибавляется *интервал погрешности* ϵ :

$$x = \hat{x} + \epsilon \quad (4)$$

Модель погрешности наблюдения.

Интервал погрешности, вообще говоря, может быть произвольным, но если он уравновешен, то есть

$$\epsilon = [-\epsilon, \epsilon],$$

то это можно трактовать, как отсутствие систематических погрешностей в прямом измерении.

Твины.

На практике концы интервалов, представляющие результаты измерений, сами могут быть известны неточно, так что возникает необходимость работы с интервалами, имеющими интервальные концы.

Такие объекты известны в интервальном анализе и называются *твинами* (по английски twin, как сокращение фразы twice interval, «двойной интервал»).

Твины были введены в научный оборот в начале 80-х годов XX века в работах испанских исследователей.

Развёрнутый анализ дан в диссертации В.М.Нестерова, 1999. Твинные арифметики и их применение в методах и алгоритмах двустороннего интервального оценивания. – Санкт-Петербург, 1999.

<http://www.nsc.ru/interval/Library/InteDiss/Nesterov-disser-1999.pdf>

Твины.

Твин, как «интервал интервалов» или интервал с интервальными концами, можно представить как

$$X = [a, b] = [[\underline{a}, \bar{a}], [\underline{b}, \bar{b}]]. \quad (5)$$

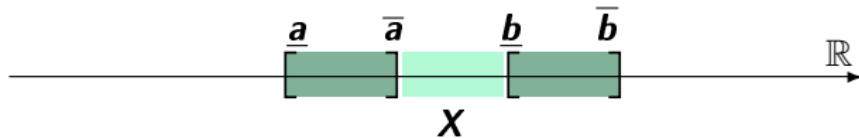


Рис.: Твины на вещественной оси.

На рисунке твин X представлен в графической форме. Концы твина, т. е. интервалы a и b , даны более тёмной заливкой, чем остальная часть твина.

Твины.

Твин является множеством всех интервалов, больших или равных $[\underline{a}, \bar{a}]$ и меньших или равных $[\underline{b}, \bar{b}]$, и точное определение зависит от смысла, который вкладывается в понятия «больше или равно», «меньше или равно».

Поскольку интервалы могут быть упорядочены различными способами, то существуют различные виды твинов. Двум основным частичным порядкам на \mathbb{IR} и \mathbb{KR} , « \subseteq » и « \leq », соответствуют два основных типа твинов. Разработаны различные операции с твинами, а также способы оценок значений функций от них.

Пример

Измерение температуры термометром сопротивления.

В повседневной лабораторной и промышленной практике широко применяются термометры сопротивления.

Один из типов таких датчиков, платиновый термометр Pt100, имеет номинальное сопротивление 100 Ом при температуре 0°C и систематическую погрешность

$$\Delta t = \pm 0.35 \text{ } ^\circ\text{C}.$$

Пример

Пусть измеряемая температура находится в диапазоне $[19.5, 20.5]$ $^{\circ}\text{C}$, которую представим как интервал t :

$$t = [19.5, 20.5] \text{ } ^{\circ}\text{C}. \quad (6)$$

Аналогично рассмотренному выше примеру, представим границы \underline{t} , \bar{t} интервала t как интервалы. С учётом систематической погрешности твин температур T , даваемый датчиком, составит

$$T = [[19.15, 19.85], [20.15, 20.85]] \text{ } ^{\circ}\text{C}. \quad (7)$$

Графическое представление твина T (7) дано на рисунке 2.

Пример

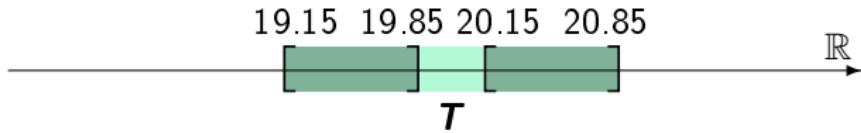


Рис.: Температура как твин.

Форма записи температуры в виде твина T (7) выразительно и полно представляет информацию об измеряемых данных. В случае, если концы интервала в выражении (6) могут меняться независимо, возможны различные ситуации. В частности, может реализоваться ситуация, подобная рассмотренной выше для твина R_2 . Также может оказаться, что значения температур для левого конца будут выше, чем для правого.

Мультиинтревалы.

В ряде разделов науки и техники имеют место ситуации, когда исследуемая величина содержится в неодносвязной области.

Мультиинтервал — это объединение конечного числа несвязных интервалов числовой оси (Рис. 3).



Рис.: Мультиинтервал в \mathbb{R} .

Мультиинтервалы.

Между мультиинтервалами также могут быть определены арифметические операции «по представителям», аналогично тому, как это делается на множестве интервалов.

Мультиинтервальная арифметика применяется редко ввиду серьёзных ограничений, которые возникают при алгебраических операциях с мультиинтервальными величинами и вычислительных сложностей. Тем не менее, сама по себе идея мультиинтервалов содержательна и полностью отметать её не стоит.

Ряд научных и технических примеров возникновения мультиинтервалов приводится в материале А.Н.Баженов.

Естественнонаучные и технические применения интервального анализа: учебное пособие.

<https://elib.spbstu.ru/dl/5/tr/2021/tr21-169.pdf/info>.

Пример

Рассмотрим задачу калибровки временной шкалы прибора. Для этого на прибор подаётся гармонический сигнал. В силу того, что на промышленно выпускаемых генераторах положительный и отрицательный фронт имеет разную длительность, необходимо различать эти части временной шкалы.

На рисунке 4 чёрным цветом показан гармонический сигнал и выделены соответственно красным и синим цветом области положительной и отрицательной производной сигнала. Эти области образуют мультиинтревалы. Они преобразуются при изменении калибровочного сигнала.

При изменении частоты составляющие мультиинтревалов расширяются или сужаются. При изменении фазы происходит их сдвиг.

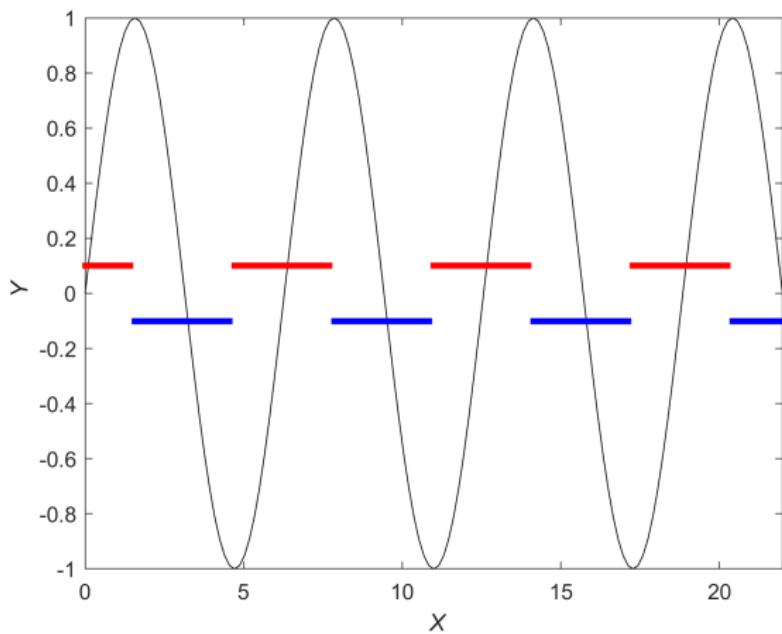


Рис.: Мультиинтревалы фаз гармонического сигнала.

Погрешность измерений

На практике измерения и наблюдения, как правило, подвержены неизбежным внешним влияниям, выполняющие их средства измерений и приборы не вполне точны и т. п., что в целом приводит к отличию измеренного значения от истинного (идеального) значения физической величины.

По отношению к неточным измерениям иногда используют термин «зашумлённые» (зашумлённые данные и т. п.), особенно, когда проводится целая серия таких измерений или наблюдений. Чтобы количественно охарактеризовать неточности измерений, вводится понятие *погрешности*.

Погрешность измерений — вещественная арифметика

Погрешность измерения — это отклонение результата измерения от истинного значения измеряемой величины. Математически погрешность равна алгебраической разности измеренного значения и истинного значения величины.

Если это истинное значение x^* и результат измерения \tilde{x} — вещественные числа, то погрешностью является разность $\tilde{x} - x^*$.

Погрешность измерений — интервальная арифметика

Если истинное значение и результат измерения — интервалы x^* и \tilde{x} соответственно, то погрешность Δ определяется как алгебраическая разность

$$\Delta = \tilde{x} \ominus x^* \quad (8)$$

в полной интервальной арифметике Каухера.

Напомним, что обычное интервальное вычитание, которое обозначается традиционным знаком « $-$ » и является интервальным расширением вычитания, не является операцией, алгебраически обратной сложению и для нашей цели непригодно.

Формула (8) справедлива и в том случае, когда истинное значение величины x^* — точечное, а результат её измерения \tilde{x} интервальный. При этом в (8) полагаем $x^* = [x^*, x^*]$.

Расстояние на множестве интервалов.

Расстояние между интервалами \mathbf{a} и \mathbf{b} из \mathbb{IR} или \mathbb{KR} определяется как

$$\text{dist}(\mathbf{a}, \mathbf{b}) = \max\{|\underline{\mathbf{a}} - \underline{\mathbf{b}}|, |\bar{\mathbf{a}} - \bar{\mathbf{b}}|\}. \quad (9)$$

Оно обладает всеми свойствами абстрактного расстояния (метрики) и
ещё некоторыми хорошими свойствами в связи с интервальными
арифметическими операциями. Кроме того, легко убедиться, что

$$\text{dist}(\mathbf{a}, \mathbf{b}) = |\mathbf{a} \ominus \mathbf{b}|.$$

Эта формула является полным аналогом расстояния между точками
вещественной оси, как модуля их разности, т. е. $|a - b|$.

Расстояние на множестве интервалов.

Рассмотрим интервал $[3, 5]$ и точку 3.6 внутри него. Расстояние от этой точки, отождествляемой с вырожденным интервалом $[3.6, 3.6]$, до данного интервала равно

$$\text{dist} (3.6, [3, 5]) = \max\{|3.6 - 3|, |3.6 - 5|\} = 1.4.$$

Рассмотрим дуальный интервал к интервалу $[3, 5]$. Это интервал $\text{dual}[3, 5] = [5, 3]$. Расстояние его до исходного интервала равно

$$\text{dist} ([3, 5], [5, 3]) = 2.$$

Расстояние важно для определения отклонения интервалов друг от друга и, как следствие, для определения погрешности интервальных измерений.

Погрешность измерений

Абсолютной погрешностью измерения назовём модуль (абсолютное значение) погрешности.

Для интервальных измерений абсолютная погрешность равна модулю интервала разности $\tilde{x} \ominus x$, и, как легко видеть, она равна расстоянию (9) между измеренным и истинным значениями величины.

Пример

Рассмотрим для примера ситуацию, когда истинное значение измеряемой величины, скажем, массы какого-то груза, является интервалом $[3, 4]$ кг, а её измерение дало интервал $[3, 5]$ кг. Тогда его погрешность равна

$$[3, 5] \text{ кг} \ominus [3, 4] \text{ кг} = [0, 1] \text{ кг}. \quad (10)$$

Пример

Если в результате измерения мы получим вещественное значение 3.8 кг, которое отождествляется с интервалом $[3.8, 3.8]$ кг, то его погрешность

$$[3.8, 3.8] \text{ кг} \ominus [3, 4] \text{ кг} = [0.8, -0.2] \text{ кг} \quad (11)$$

— неправильный интервал.

Может показаться, что он бессмыслен с физической точки зрения, но это поспешный вывод. Ситуация здесь совершенно аналогична, например, тому, как при измерении положительных физических величин (массы, плотности, давления и т. п.) мы получаем отрицательную погрешность, если измеренное значение приближает истинное значение снизу.

Абсолютная погрешность измерения равна 1 в случае (10) и 0.8 в случае (11).

Накрывающие и ненакрывающие измерения

Если результат измерения — точечная величина, то для неё возможны только два исхода проведения измерения: либо она получается равной истинному значению интересующей нас физической величины, либо не равной ей. Как говорят математики и программисты, исход измерения является «булевозначным», «да» или «нет».

При этом ясно, что в случае измерения непрерывных физических величин равенство является исключительным событием и почти никогда не достигается. Если же оно по каким-то причинам произошло, то является неустойчивым к сколь угодно малым возмущениям или же погрешностям в вычислительных алгоритмах.

Накрывающие и ненакрывающие измерения

Принципиально другая ситуация возникает, если результат измерения может быть интервалом.

Интервал по своей сути является двусторонней «вилкой» значений, и принадлежность ей истинного значения — это уже не исключительное событие. Оно, как правило, устойчиво к возмущениям и погрешностям обработки. Как следствие, для теории обработки интервальных данных фундаментальный характер имеют следующие определения:

Накрывающие и ненакрывающие измерения

Definition

Накрывающее измерение (накрывающий замер) — это интервальная оценка неизвестной истинной величины, гарантированно ее содержащая.

Измерение, не являющееся накрывающим, будем называть *ненакрывающим* (Рис. 5 и Рис. 6).

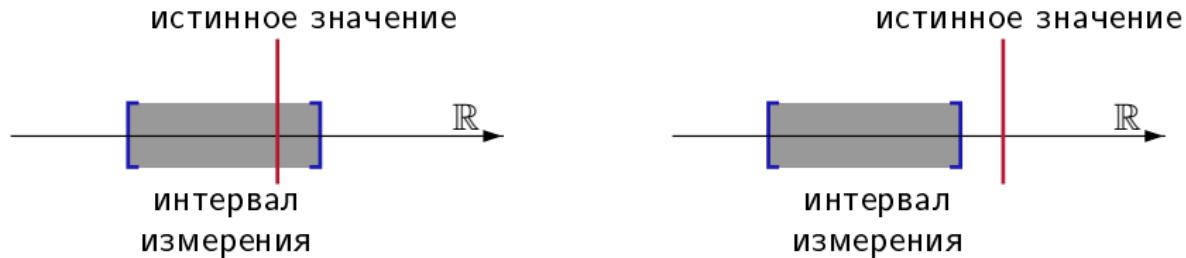


Рис.: Накрывающее (слева) и ненакрывающее (справа)

измерения точечного истинного значения некоторой физической величины,

Накрывающие и ненакрывающие выборки

Definition

Накрывающая выборка — совокупность накрывающих измерений, т. е. выборка, в которой все измерения (наблюдения) являются накрывающими. Напротив, выборка называется *ненакрывающей*, если хотя бы одно из входящих в неё измерений — ненакрывающее.

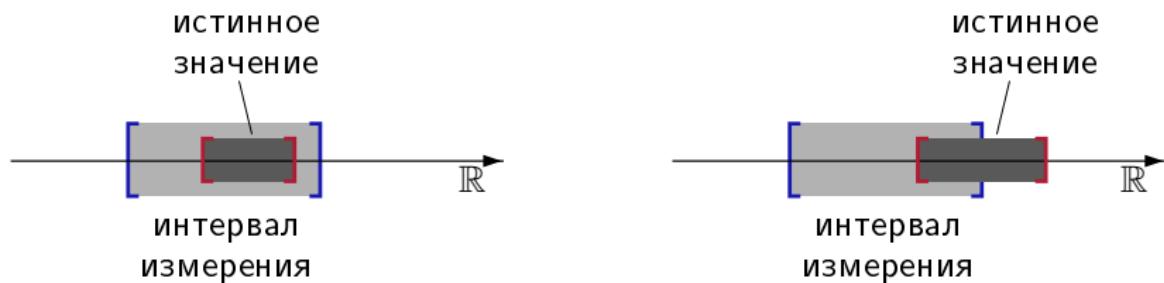


Рис.: Накрывающее (слева) и ненакрывающее (справа)

измерения интервального истинного значения некоторой физической величины.

Информационное множество

Неформально говоря, *информационное множество* — это множество параметров задачи, которые совместны с данными измерений в рамках выбранной модели их обработки.

Информационное множество

Аналогом «информационного множества» может отчасти служить понятие *доверительного интервала* оцениваемой случайной величины в традиционной вероятностной статистике.

В определение доверительного интервала входит дополнительный параметр — *уровень статистической значимости*, без которого понятие становится бессодержательным из-за неограниченности носителей большинства вероятностных распределений, но смысл доверительного интервала примерно соответствует «информационному множеству».

Информационное множество

Далее для обозначения различных информационных множеств мы будем использовать прописную греческую букву

Ω

(«омега»), добавляя к ней при необходимости параметры, обозначающие контекст задачи.

Так как информационное множество может быть достаточно произвольным множеством в пространстве параметров и не обязательно является интервалом, интервальным вектором или интервальной матрицей, мы не выделяем его символ жирным шрифтом.

Принцип соответствия

Принцип соответствия в методологии науки — это утверждение, что любая новая научная теория должна включать старую теорию и её результаты как частный предельный случай.

Мы будем использовать принцип соответствия, как инструмент проверки «разумности» и адекватности наших конструкций, понятий и методов обработки данных с интервальными неопределённостями, который позволяет отсекать заведомо «неразумные».

Выбросы и промахи

Выбросами или промахами в метрологии называются такие измерения, результаты которых не приносят информацию об исследуемом объекте в рамках его принятой модели.

Другое популярное определение выбросов или промахов состоит в том, что это результаты измерений, которые для данных условий резко отличаются от остальных результатов общей выборки.

Выбросы и промахи

Что считать выбросом (промахом) в случае интервальных результатов измерений? Прежде всего, не стоит связывать выбросы со свойством измерений быть накрывающими или ненакрывающими.

Более точно, из того, что интервальное измерение не является накрывающим, не следует, что оно представляет выброс или промах.

Отождествление выбросов (промахов) со свойством ненакрывания противоречит принципу соответствия, сформулированному в предыдущем параграфе.

В самом деле, при стремлении ширины интервальных измерений к нулю они переходят в точечные измерения, которые, как правило, всегда ненакрывающие. Тем не менее, различие для них выбросов (промахов) от этого не исчезает.

Измерение физической величины (константы).

Физическая величина взята в качестве примера. Данные могут быть любой природы: из наук о Земле, биологии, науках об обществе, экономики, etc.

Измерение физической величины — пример.

Проведём рассмотрение обработки данных физического эксперимента по измерению константы. В качестве источника данных будем использовать публикацию [2], представляющую результаты измерения циркулярной поляризации гамма-кванта в реакции захвата поляризованного нейтрона протоном.

Приведём часть данных таблицы 1 из публикации [2].
В таблице 1 основные данные измерения содержатся в столбцах Peak — средние значения и std Peak — оценки ошибки. В столбцах BG и std BG приведены данные, которые можно использовать для коррекции систематических ошибок. В первом столбце дан условный номер эксперимента.

Исходные данные. Величина $\delta \times 10^5$.

Номер замера	Peak	std Peak	BG	std BG
1	-4.4	2.7	4.2	6.7
2	-3.4	1.9	-3.2	4.8
3	-6.9	2.4	12.1	9
4	-1.2	2.4	12.4	7.2
5	-1.0	2.7	9.4	5.1
6	-10.8	3.5	1	12.4
7	-10.2	2.8	-0.6	6.1
8	-6.3	2	3.9	4.3
9	-10.4	4.1	10.3	10
10	0.6	3.4	-4.8	10.6
11	-1.8	2	4.6	4.2
12	-6.6	2.1	-5.7	4.6
13	-4.9	2.1	13	3
14	-6.0	2.4	8.4	4.6
15	-4.0	2.7	10.6	5.5

Таблица: Данные таблицы 1 для величины $\delta \times 10^5$ [2].

Представление данных.

В первую очередь представим данные таким образом, чтобы применить понятия статистики данных с интервальнойной неопределённостью.

Согласно терминологии интервального анализа, рассматриваемая выборка — это вектор интервалов, или интервальный вектор $\mathbf{x} = (x_1, x_2, \dots, x_n)$.

Для того, чтобы придать данным таблицы 1 необходимую форму, примем, что в качестве элементов \mathbf{x} будут выступать данные

$$\text{mid } x_k = \text{Peak}(k), \quad \text{rad } x_k = \text{std Peak}(k), \quad k = 1, 2, \dots, 15.$$

Для наглядного представления выборки часто рисуют образующие её интервалы в виде графика, изображённого на Рис. 10, который по статистической традиции мы будем называть *диаграммой рассеяния*.

Диаграмма рассеяния интервальных измерений.

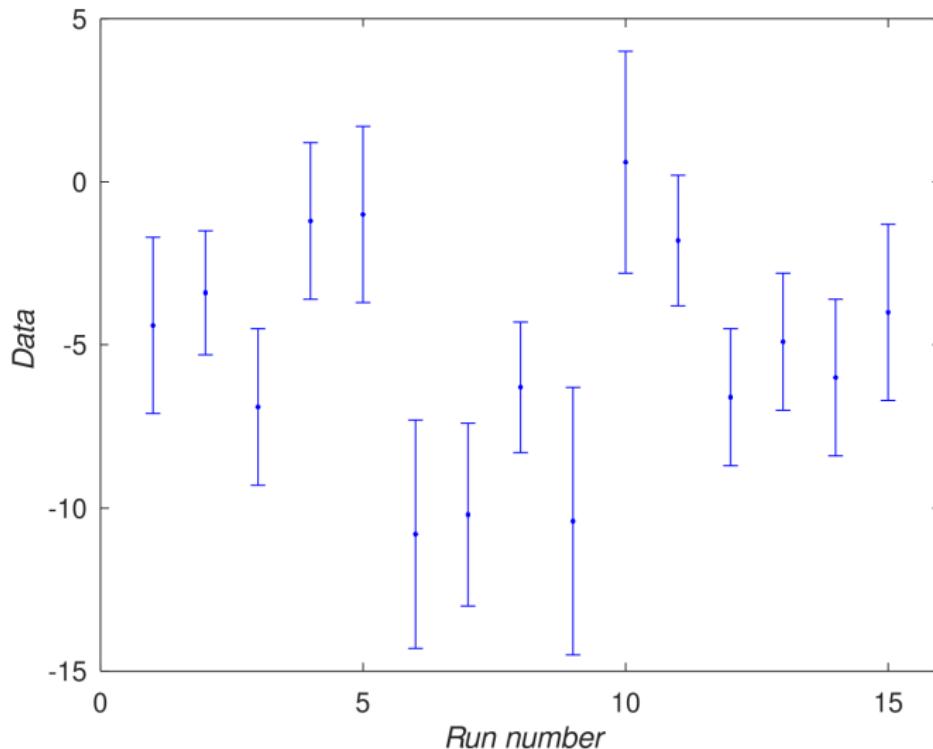


Рис.: Диаграмма рассеяния интервальных измерений [2]

Диаграмма рассеяния интервальных измерений.

Из таблицы 1 и Рис. 10 видно, что элементы выборки *неравноширичные*, поскольку величина неопределённости $\text{rad}x_k$ меняется в зависимости от измерения выборки, $k = 1, \dots, n$.

Информационное множество.

Информационным множеством в случае оценивания единичной физической величины по выборке интервальных данных будет также интервал, который называют *информационным интервалом*.

Неформально говоря, это интервал, содержащий значения оцениваемой величины, которые «совместны» с измерениями выборки («согласуются» с данными этих измерений).

Конкретный смысл, вкладываемый в понятия «совместные» или «согласующиеся», будет различен для разных ситуаций. В частности, он зависит от того, является ли выборка интервальных данных накрывающей или нет.

Совместность выборки

Важным внутренним свойством интервальной выборки, характеризующим согласование её данных между собой, является понятие совместности.

Definition

Выборка $\{x_k\}_{k=1}^n$ называется *совместной*, если пересечение всех интервалов составляющих её измерений непусто, т. е.

$$\bigcap_{1 \leq k \leq n} x_k \neq \emptyset.$$

В противном случае, если пересечение всех интервалов x_k , $k = 1, \dots, n$, является пустым, то выборка называется *несовместной*.

Совместность выборки

Свойство совместности характеризует саму выборку и, строго говоря, не связано напрямую с её свойством быть накрывающей выборкой, т. е. с включением в неё истинного значения измеряемой величины.

Выборка может быть совместной, но ненакрывающей. Но если выборка накрывающая, то она обязана быть совместной.

Эквивалентная формулировка этого свойства: если выборка несовместна, то она и ненакрывающая.

Основываясь на этих соображениях, в практической обработке результатов измерений трудный анализ накрытия выборкой истинного значения часто заменяют анализом её совместности, так как это удобнее и нагляднее (хотя и не вполне строго).

Совместность выборки

Если обрабатываемая выборка несовместна, то это может вызываться следующими причинами:

- (а) неверно заданным значением неопределённости измерений radx_k для каких-то $k \in \{1, 2, \dots, n\}$, которое занижено в сравнении с фактическим значением неопределённости;
- (б) наличием в этой выборке выбросов (промахов), т. е. сбойных измерений;
- (в) невыполнением условий на измеряемую физическую величину (её непостоянство и т. п.).

Обработка накрывающей выборки

Если истинное значение величины содержится во всех интервалах измерений выборки $\{x_k\}_{k=1}^n$, то оно должно принадлежать также пересечению этих интервалов. Следовательно, уточнённым интервалом принадлежности истинного значения можно взять

$$I = \bigcap_{1 \leq k \leq n} x_k. \quad (12)$$

Это и будет информационный интервал I оценки измеряемой физической величины (см. Рис. 8). Явные выражения для его левой (нижней) и правой (верхней) границ даются следующими формулами:

$$\underline{I} = \max_{k=1, \dots, n} \underline{x}_k, \quad \bar{I} = \min_{k=1, \dots, n} \bar{x}_k. \quad (13)$$

Обработка накрывающей выборки

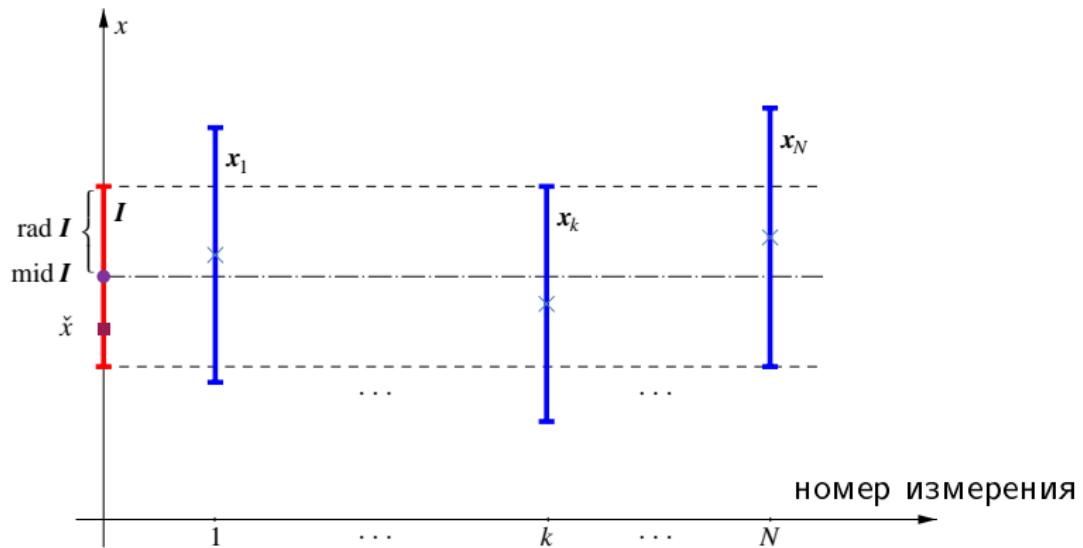


Рис.: Обработка накрывающей выборки интервальных измерений величины.

Предел совместности выборки

В силу сделанного допущения о том, что выборка накрывает истинное значение величины, имеем $\underline{I} \leq \bar{I}$.

При этом интересен предельный случай совместной выборки, когда

$$\underline{I} = \bar{I} = x^*.$$

Тогда выборка совместна, но мы, образно говоря, находимся на пределе её совместности, и информационный интервал I вырождается при этом в точку.

Уточнение априорным интервалом

Если известен некоторый априорный интервал возможных значений оцениваемой физической величины $I_{\text{апр}} = [l_{\text{апр}}, \bar{l}_{\text{апр}}]$, который должен гарантированно содержать её, то границы результирующего интервала (12) могут быть уточнены пересечением

$$I = I \cap I_{\text{апр}}. \quad (14)$$

Отметим, что априорный интервал $I_{\text{апр}}$ может задавать одностороннее ограничение, если он имеет вид $[l_{\text{апр}}, +\infty]$ или $[-\infty, \bar{l}_{\text{апр}}]$, т. е. является полубесконечным интервалом из арифметики Кахана.

Центральная оценка

На практике часто необходимо работать не с интервалами интересующей нас величины — (12) или (14), а с некоторой точечной оценкой \hat{x} . Все точки информационного интервала вполне равноценны друг другу, так что эту точечную оценку \hat{x} можно выбирать достаточно произвольно (см. Рис. 8). Тем не менее, имеет смысл взять из интервала некоторое точечное значение, которое представляет его наилучшим образом.

В качестве такой величины можно использовать, к примеру, его центральную оценку x_c ,

$$x_c = \text{mid } I = \frac{1}{2} (\underline{I} + \bar{I}). \quad (15)$$

Напомним, что середина интервала обладает определённой оптимальностью, являясь точкой, которая наименее удалёна от других точек этого интервала.

Обработка ненакрывающей выборки

Если выборка — ненакрывающая, так что некоторые из её измерений не содержат истинного значения измеряемой величины, то приведённые в предыдущем параграфе рассуждения и приёмы частично теряют свой смысл.

Поскольку кроме информации, представленной выборкой, в нашем распоряжении ничего нет, то следует бережно отнестись ко всем измерениям и считать, что каждое из них несёт существенную информацию об измеряемой величине, которая не должна быть потеряна.

Уточнение пересечением здесь уже неуместно, и информационное множество для истинного значения величины имеет смысл взять в виде объединения всех интервалов выборки, т. е. как

$$\bigcup_{1 \leq k \leq n} x_k. \quad (16)$$

Обработка ненакрывающей выборки

Это множество может не быть единым интервалом на вещественной оси (подобное часто случается, к примеру, если выборка несовместна). Разумно тогда воспользоваться вместо объединения обобщающей его операцией « \vee » (см. (3)), т. е. взятием максимума по включению, и вместо (16) взять информационный интервал в виде

$$J = \bigvee_{1 \leq k \leq n} x_k = \left[\min_{1 \leq k \leq n} x_k, \max_{1 \leq k \leq n} \bar{x}_k \right]. \quad (17)$$

Точечной оценкой измеряемой величины может служить середина полученного интервала, т. е.

$$x_c = \text{mid } J = \frac{1}{2} \left(\min_{1 \leq k \leq n} x_k + \max_{1 \leq k \leq n} \bar{x}_k \right). \quad (18)$$

Обработка ненакрывающей выборки

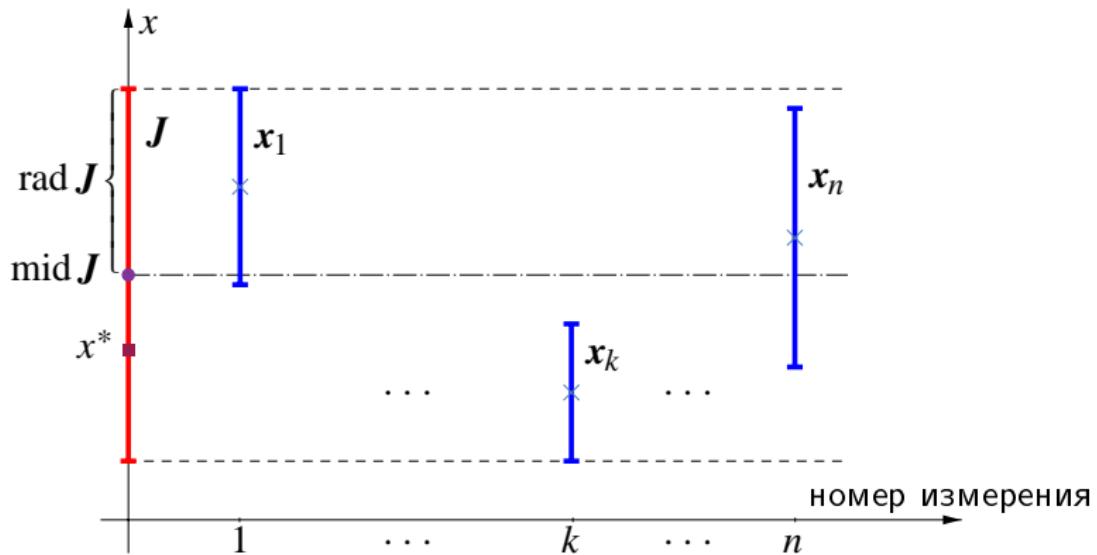


Рис.: Обработка ненакрывающей выборки интервальных измерений величины.

Уточнение априорным интервалом

Как и ранее, нам может быть известен некоторый априорный интервал возможных значений оцениваемой физической величины

$J_{\text{апр}} = [\underline{J}_{\text{апр}}, \bar{J}_{\text{апр}}]$, который должен гарантированно содержать её. Его могут задавать внешние физические (химические, биологические, экономические и т. п.) условия или ограничения.

Тогда границы результирующего интервала (17) могут быть уточнены пересечением

$$J = J \cap J_{\text{апр}}. \quad (19)$$

В данной ситуации это уточнение имеет даже больший смысл, чем в случае накрывающей выборки.

Взятие минимума по включению

Другой возможный сценарий обработки данных ненакрывающей выборки может состоять в том, что вместо пересечения интервальных измерений мы используем обобщающую её операцию « \wedge », т. е. взятие минимума всех интервальных результатов измерений относительно упорядочения по включению:

$$I = \bigwedge_{1 \leq k \leq n} x_k = \left[\max_{1 \leq k \leq n} \underline{x}_k, \min_{1 \leq k \leq n} \bar{x}_k \right]. \quad (20)$$

Здесь по существу требуется использование полной интервальной арифметики Каухера, так как интервал (20) может оказаться неправильным.

Точечная оценка ненакрывающей выборки

Соответственно, точечной оценкой измеряемой величины целесообразно взять

$$x_c = \text{mid } I = \frac{1}{2} \left(\max_{1 \leq k \leq n} \underline{x}_k + \min_{1 \leq k \leq n} \bar{x}_k \right), \quad (21)$$

т. е. середину интервала, который получается как минимум по включению всех интервалов выборки (см. (3)).

Если выборка совместна, то (21) совпадает с (15). Если же выборка несовместна, то результатом (20) является неправильный интервал I , $\text{rad } I < 0$. Соответственно, информационное множество результатов измерений по обрабатываемой выборке пусто.

Оптимальность точечной оценки

Но даже когда интервал (20) неправилен, его середина (21) — это точка, обладающая определёнными условиями оптимальности. Она первой появляется в непустом пересечении интервалов выборки, если мы станем равномерно уширять их, увеличивая неопределённость измерений.

В самом деле, пусть радиусы всех интервалов выборки увеличились на s , $s \geq 0$, тогда как середины остались неизменными. Вместо радиусов $\text{rad}x_k$ мы получили $\text{rad}x_k + s$, $k = 1, 2, \dots, n$. Кроме того, все нижние концы интервальных измерений стали теперь $\underline{x}_k - s$, а верхние концы — $\bar{x}_k + s$, $k = 1, 2, \dots, n$.

Как следствие, $\max_{1 \leq k \leq n} \underline{x}_k$ уменьшается на s , а $\min_{1 \leq k \leq n} \bar{x}_k$ увеличивается на s , а радиус получающегося интервала (20) теперь равен $\text{rad}I + s$.

Оптимальность точечной оценки

Как следствие, $\max_{1 \leq k \leq n} x_k$ уменьшается на s , а $\min_{1 \leq k \leq n} \bar{x}_k$ увеличивается на s , а радиус получающегося интервала (20) теперь равен $\text{rad}I + s$.

Поэтому, если взять s таким, чтобы $s \geq |\text{rad}I|$, то получившийся интервал станет правильным, и точка x_c будет лежать в нём.

Можно также сказать, что в точке (21) минимизируется равномерное уширение интервалов данных рассматриваемой выборки, необходимое для достижения её совместности.

«Средняя» оценка ненакрывающей выборки

Наконец, если выборка интервальных измерений — ненакрывающая, то иногда имеет смысл взять среднее арифметическое образующих её интервалов, т. е.

$$K = \frac{1}{n} \sum_{k=1}^n x_k.$$

Его середина может служить точечной оценкой измеряемой величины.

Принцип соответствия

Нетрудно убедиться в том, что все три рассмотренных выше приёма обработки ненакрывающей выборки при стремлении ширины интервальных данных к нулю переходят в осмысленные методы оценивания физической величины по точечным данным.

В частности, она полагается равной среднему арифметическому измерений выборки в третьем случае. То есть, эти методы удовлетворяют «принципу соответствия».

Пример выборки данных [2].

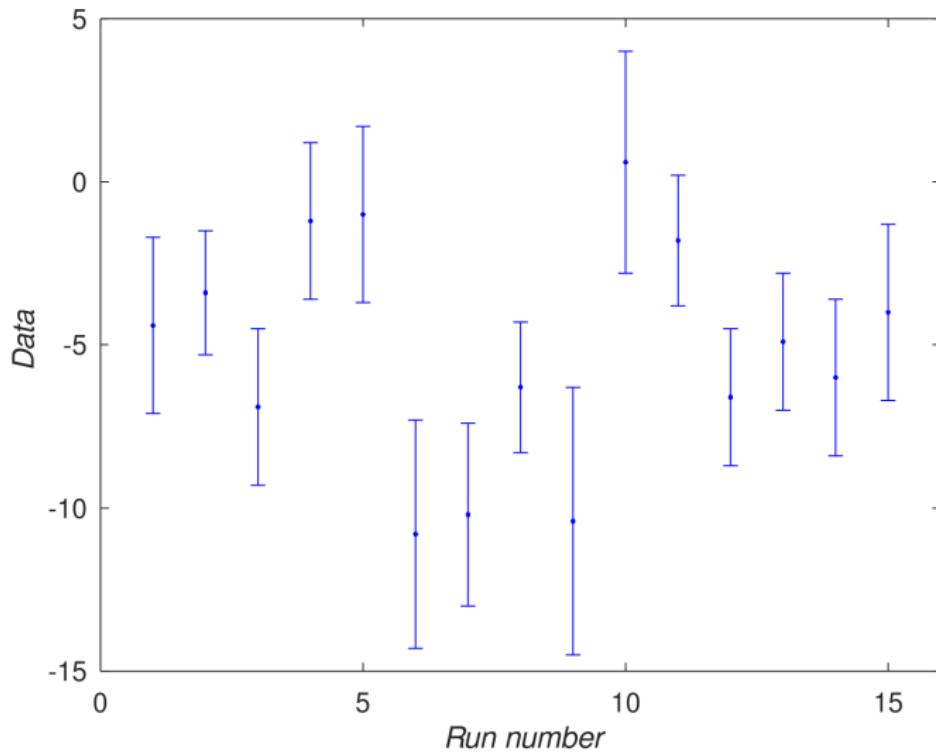


Рис.: Диаграмма рассеяния интервальных измерений [2]

Пример данных [2].

Информация, представленная выборкой Табл. 1, уникальна, так что следует бережно отнестись ко всем измерениям и считать, что каждое из них несёт существенную информацию об измеряемой величине, которая не должна быть потеряна.

Попробуем взять в качестве информационного множества для истинного значения величины объединение всех интервалов выборки, т. е.

$$I_{Uni} = \bigcup_{1 \leq k \leq n} x_k = [-14.5, 4.0]. \quad (22)$$

Пример данных [2].

По существу измеряемая величина является константой неизвестного, но определённого знака. Оценка (16) в данном случае имеет разные знаки концов интервалов и противоречит постановке задачи.

Можно было бы отбросить элементов выборки, имеющие «неправильный» знак, но это представляется недопустимым произволом.

Вместе с тем, середина интервала (16)

$$\text{mid } I_{U_{hi}} = -5.25$$

может быть разумной точечной оценкой, и её будет полезно сравнить с оценками, полученными на основе других подходов.

Пример данных [2].

Продемонстрируем наглядно, что получается в конкретном случае. Будем представлять теперь данные в несколько ином виде, чем на рисунке 10, откладывая номер измерения по вертикальной шкале.

При этом мы будем действовать согласовано с представлением подобных результатов при обработке данных на ресурсе С.И.Жилина [3].

Вычисления проводились в среде Octave в классической интервальной арифметике с использованием стандартной библиотеки `interval` и полной интервальной арифметики с использованием библиотеки `kinterval` [4].

Пример данных [2].

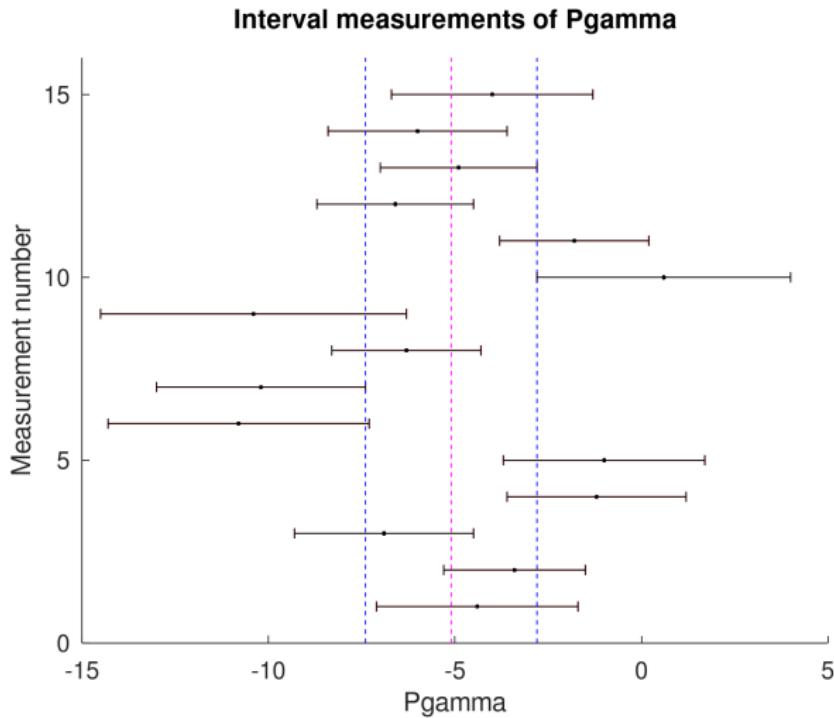


Рис.: Диаграмма рассеяния интервальных измерений величины, полоса минимума по включению (20) и точечная оценка (21).

Пример данных [2].

На Рис. 11 синими вертикальными линиями показаны границы информационного множества, полученные по формуле (20)

$$I = \bigwedge_{1 \leq k \leq n} x_k = \left[\max_{1 \leq k \leq n} \underline{x}_k, \min_{1 \leq k \leq n} \bar{x}_k \right] = [-2.8, -7.4].$$

Также вычислим точечную оценку измеряемой величины по формуле (21)

$$x_c = \text{mid } I = \frac{1}{2} \left(\max_{1 \leq k \leq n} \underline{x}_k + \min_{1 \leq k \leq n} \bar{x}_k \right) = -5.1.$$

На Рис. 11 эта величина показана вертикальной линией цветом magenta. Интервал I — неправильный. Смысл значения x_c прояснён в комментарии после формулы (21) как точки, которая первой появляется в непустом пересечении интервалов выборки, если мы станем равномерно уширять их.

Пример данных [2].

Наконец, если выборка интервальных измерений — ненакрывающая, то иногда имеет смысл взять среднее арифметическое образующих её интервалов, т. е.

$$K = \frac{1}{n} \sum_{k=1}^n x_k = [-7.77, -2.54]. \quad (23)$$

Середина этого интервала

$$\text{mid } K = -5.15$$

также может служить точечной оценкой измеряемой величины.

Вариабельность оценки — радиус

Рассмотрим теперь характеристики разброса оценок физической величины, полученных по интервальной выборке. Её наиболее естественной мерой, если информационный интервал непуст, является его *радиус* ϱ , т. е.

$$\varrho = \text{rad}I = \frac{1}{2}(\bar{I} - I).$$

Фактически, это максимальное отклонение границ информационного интервала от центральной оценки.

Вариабельность оценки — отклонения

При анализе данных имеет также смысл знать отклонения точечных или интервальных измерений выборки от итоговой точечной оценки. Они дают возможность судить о степени разброса измерений относительно полученной оценки, что помогает при анализе «качества» выборки и выявлении выбросов.

Отклонения Δ_k для первичных интервальных измерений рассчитываются как

$$\Delta_k = \text{dist}(\mathbf{x}_k, x_c), \quad k = 1, \dots, n. \quad (24)$$

Вариабельность оценки

В некоторых случаях имеет смысл отсчитывать отклонения от базовых точечных измерений, вокруг которых строятся далее интервальные результаты, т. е. рассматривать в качестве отклонений результатов отдельных измерений величины

$$\Delta_k = |\hat{x}_k - x_c|, \quad k = 1, \dots, n. \quad (25)$$

Норма вектора $\Delta = (\Delta_1, \dots, \Delta_n)$ может служить аналогом выборочной дисперсии оценки из традиционной вероятностной статистики.

Приём варьирования неопределённости

Выше мы видели, что величина реальной неопределенности измерения, т. е. радиуса интервала измерения, определяется непросто и подчас неоднозначно. С другой стороны, он сильно влияет на свойства как отдельного измерения, так и выборки интервальных измерений. Совместность выборки и свойство накрытия истинного значения существенно зависят от правильно назначенной величины неопределенности — радиуса интервальных измерений. Наконец, если некоторое Δ является величиной неопределенности интервального измерения или выборки, то и любое Δ' , удовлетворяющее $\Delta' \geq \Delta$, также может служить величиной неопределенности.

Сказанное выше приводит к мысли о том, что при обработке интервальных данных величиной неопределенности можно управлять, виртуально варьируя её, с целью исследования интервальных измерений, их выборок и построения оценок с нужными свойствами.

Приём варьирования неопределённости

Если выборка интервальных измерений несовместна, то, увеличивая одновременно величину неопределённости всех измерений, мы всегда сможем добиться того, чтобы выборка сделалась совместной, т. е. чтобы пересечение интервалов стало непустым, а интервал минимума по включению (20) — правильным.

Кроме того, точка (или точки), которая первой появляется в непустом пересечении интервалов при расширении интервальных измерений, и тем самым требует наименьшего увеличения неопределённости измерений для достижения совместности выборки, является «наименее несовместной». Её разумно брать в качестве оценки величины (или оценки параметров зависимости).

Приём варьирования неопределённости

В конкретной ситуации данных [2], измерения выборки являются существенно неравноширичными. Одновременное изменение величины неопределенности для всех измерений на одно и то же значение может оказаться неразумным.

Пусть задан некоторый положительный весовой вектор

$w = (w_1, w_2, \dots, w_n)$, $w_k > 0$, размерность которого равна длине исследуемой выборки, причём изменение величины неопределенности k -го измерения — $\text{rad}x_k$, должно быть пропорциональным w_k , т. е. для любых k и l справедливо

$$\frac{\text{изменение } \text{rad}x_k}{\text{изменение } \text{rad}x_l} = \frac{w_k}{w_l}.$$

Приём варьирования неопределённости

Идея варьирования величины неопределенности интервальных измерений оформилась в 80-е годы XX века (Н.М. Оскорбин [5] и др.), и далее неоднократно переоткрывалась различными исследователями.

Применительно к данным таблицы 1, применение методики приведено на Рис. 12.

Красным цветом даны исходные данные таблицы 1, а чёрным цветом — «расширенные» интервалы данных при выбранном коэффициенте расширения.

Приём варьирования неопределённости

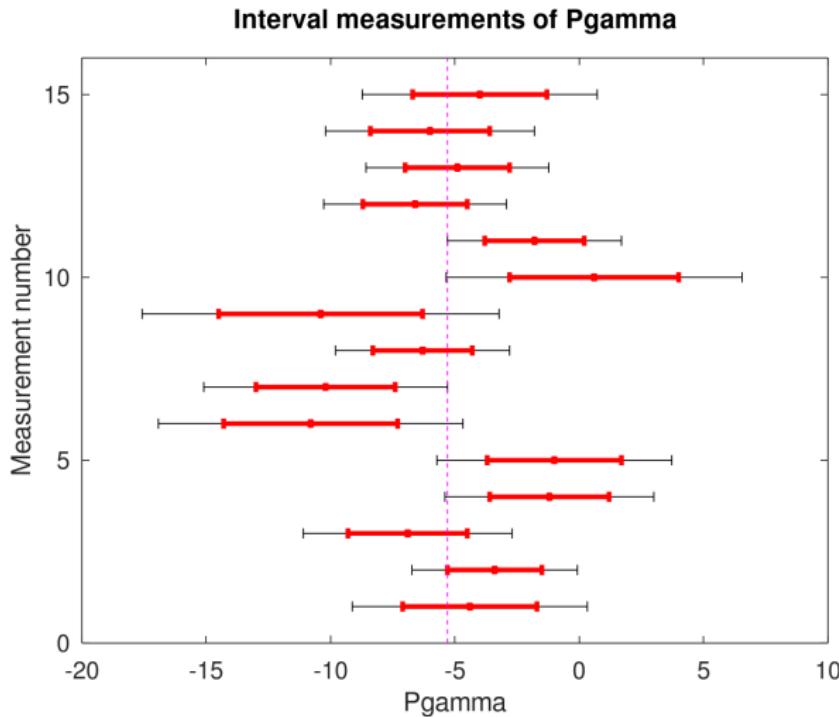


Рис.: Графическое представление интервальных данных и результаты обработки по методике [5].

Приём варьирования неопределённости

Вычисления проведены по методике [5] и с использованием кода С.И.Жилина [3]. При этом решается задача линейного программирования, в ходе которой вычисляются 2 параметра: оптимальное положение «центра неопределенности» `oskorbin_center` и коэффициент расширения радиусов замеров.

$$x_{MM} = \text{oskorbin_center} = -5.30, \quad k = 1.75.$$

Здесь в индексе x_{MM} , MM соответствует Minimal Module, функции оптимизации задачи линейного программирования.

Информационное множество представляет точку

$$I_{MM} = \bigcap_{1 \leq k \leq n} x_k = x_{MM}.$$

Приём варьирования неопределённости

Содержательным результатом вычислений является уточнение положения наиболее вероятной точечной оценки физической величины [2] и вычисление дополнительной погрешности для каждого элемента выборки, необходимой для достижения совместности данных.

Следует заметить, что значение x_{MM} , полученное варьированием неопределённости, ненамного отличается от полученных ранее оценок.

Это свидетельствует в пользу того, что выборка данных таблицы 1 не обладает какими-то патологическими свойствами. При этом для данных требуется увеличение неопределённости. Таким образом, можно говорить о наличии систематических погрешностей.

Литература

-  А.Н. Баженов, С.И. Жилин, С.И. Кумков, С.П. Шарый. Обработка и анализ данных с интервальной неопределенностью. РХД. Серия «Интервальный анализ и его приложения». Ижевск. 2021. с.200.
-  V.M.LOBASHEV ET AL, Circular polarization of γ -quanta in the $pr \rightarrow d\gamma$ reactions with polarized neutrons. Physics Letters B, Volume 289, Issues 1–2, 3 September 1992, Pages 17-21.
-  С.И.Жилин. Примеры анализа интервальных данных в Octave <https://github.com/szhilin/octave-interval-examples>
-  С.И.Жилин. Библиотека полной интервальной арифметики kinterval в среде Octave. Частное сообщение.
-  Оскорбин Н.М. Некоторые задачи обработки информации в управляемых системах // Синтез и проектирование многоуровневых иерархических систем. Материалы конференции. – Барнаул: Алтайский государственный университет, 1983.

Тема X2. Обработка и анализ данных с интервальной неопределенностью.

А.Н. Баженов

Санкт-Петербургский политехнический университет Петра Великого

a_bazhenov@inbox.ru

28.09.2021

Обработка и анализ данных с интервальной неопределённостью.

ПЛАН

ПЛАН

- Общие понятия
- Обработка константы
- Задача восстановления зависимостей

Теория:

А.Н. Баженов, С.И. Жилин, С.И. Кумков, С.П. Шарый.
Обработка и анализ данных с интервальной неопределенностью. РХД.
Серия «Интервальный анализ и его приложения». Ижевск. 2021. с.200.

ПЛАН

Задача восстановления зависимостей. Часть 1.

Задача восстановления зависимостей

Даются определения новых терминов и понятий, которые возникают в связи с восстановлением функциональных зависимостей по данным их измерений и наблюдений, имеющих интервальную неопределённость.

Мы рассмотрим основные идеи и типичные приёмы восстановления зависимостей по интервальным данным, а также возникающие при этом проблемы.

Подробно исследуется случай простейшей линейной зависимости, но большинство построений и рассуждений легко переносятся на общий нелинейный случай.

Постановка задачи

Предположим, что величина y является функцией некоторого заданного вида от независимых аргументов x_1, x_2, \dots, x_m , т. е.

$$y = f(x, \beta), \quad (1)$$

где $x = (x_1, \dots, x_m)$ — вектор независимых переменных,
 $\beta = (\beta_1, \dots, \beta_l)$ — вектор параметров функции. Имея набор значений переменных x и y , нам нужно найти β_1, \dots, β_l , которые соответствуют конкретной функции f из параметрического семейства (1).

Мы будем называть эту задачу *задачей восстановления зависимости*.

Постановка задачи

Важнейший частный случай поставленной задачи — определение параметров линейной функциональной зависимости вида

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m, \quad (2)$$

в которой x_1, x_2, \dots, x_m — независимые переменные (которые называются также *экзогенными*, *предикторными* или просто *входными* переменными), y — это зависимая переменная (которая называется также *эндогенной*, *критериальной* или *выходной* переменной), а $\beta_0, \beta_1, \dots, \beta_m$ — некоторые коэффициенты.

Эти неизвестные коэффициенты должны быть определены из ряда измерений значений x_1, x_2, \dots, x_m и y .

Постановка задачи

Результаты измерений неточны, и мы предполагаем что они имеют ограниченную неопределенность, когда нам известны лишь некоторые интервалы, дающие двусторонние границы измеренных значений.

Таким образом, результатом i -го измерения являются такие интервалы $x_1^{(i)}, x_2^{(i)}, \dots, x_m^{(i)}, y^{(i)}$, относительно которых мы предполагаем, что истинное значение x_1 лежит в пределах $x_1^{(i)}$, истинное значение x_2 лежит в $x_2^{(i)}$ и т.д. вплоть до y , истинное значение которого находится в интервале $y^{(i)}$.

В целом имеется n измерений, так что индекс i может принимать значения из множества натуральных чисел $\{1, 2, \dots, n\}$.

Постановка задачи

Далее для удобства построений и выкладок обозначим номер измерения i не верхним, а нижним индексом, который мы поставим первым при обозначении входов. Таким образом, полный набор данных будет иметь вид

$$\begin{aligned} & x_{11}, \quad x_{12}, \quad \dots \quad x_{1m}, \quad y_1, \\ & x_{21}, \quad x_{22}, \quad \dots \quad x_{2m}, \quad y_2, \\ & \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ & x_{n1}, \quad x_{n2}, \quad \dots \quad x_{nm}, \quad y_n. \end{aligned} \tag{3}$$

Нам необходимо найти или как-то оценить коэффициенты β_j , $j = 0, 1, \dots, m$, для которых линейная функция (2) «наилучшим образом» приближала бы интервальные данные измерений (3).

Постановка задачи

Для обозначения $n \times m$ -матрицы, составленной из данных (3) для независимых переменных часто используют термины **матрица плана эксперимента** или просто **матрица плана**, которые возникли в теории планирования эксперимента .

Интервалы $x_{i1}, x_{i2}, \dots, x_{im}, y_i$ мы называем, как и раньше, **интервалами неопределённости i -го измерения**.

Но кроме них нам также потребуется обращаться ко всему множеству, ограничеваемому в многомерном пространстве \mathbb{R}^{m+1} этими интервалами по отдельным координатным осям.

Брус неопределённости

Definition

Брусом неопределённости i -го измерения рассматриваемой зависимости будем называть интервальный вектор-брус $(x_{i1}, x_{i2}, \dots, x_{im}, y_i) \subset \mathbb{R}^{m+1}$, $i = 1, 2, \dots, n$.

Таким образом, каждый брус неопределённости измерения зависимости является прямым декартовым произведением интервалов неопределённости независимых переменных и зависимой переменной. На Рис. 1 на плоскости Oxy наглядно показаны брусы неопределённости измерений и график линейной функции, которую мы восстанавливаем.

Далее мы рассматриваем данные (3) как «спущенные свыше» и никак не обсуждаем их выбор, коррекцию или оптимизацию.

Пример

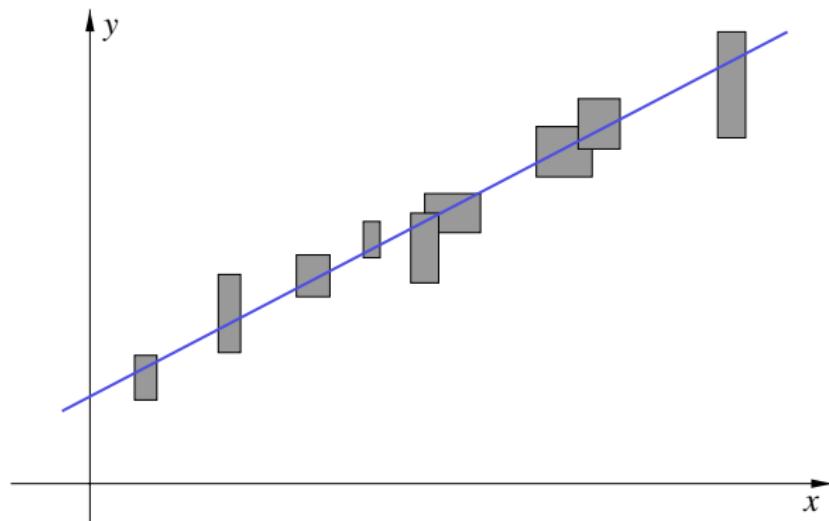


Рис.: Наглядная иллюстрация задачи восстановления линейной зависимости по данным с интервальной неопределённостью.

Накрывающие и ненакрывающие брусы

Definition

Будем называть брус неопределённости измерения зависимости **накрывающим**, если он гарантированно содержит истинные значения измеряемых величин входных и выходных переменных зависимости.

Брус неопределённости измерения зависимости, который не является накрывающим, будем называть **ненакрывающим**.

Возможные альтернативные термины — «включающий брус неопределённости», «охватывающий брус неопределённости» (их отрицание — «невключающий», «неохватывающий»).

Диаграммы рассеяния

Для визуализации интервальных данных, аналогично традиционному точечному случаю, используют *диаграммы рассеяния*.

В традиционном понимании диаграмма рассеяния используется в статистике и анализе данных для визуализации значений двух переменных в виде «облака» точек на декартовой плоскости и позволяет оценить наличие или отсутствие корреляции и других взаимосвязей между двумя переменными.

На диаграмме рассеяния для интервальных данных каждое интервальное наблюдение отображается в виде бруса (брюса неопределённости). При отсутствии неопределенности по одной из переменных, брусы наблюдений могут «схлопываться» в одномерные вертикальные или горизонтальные отрезки («ворота»).

Примерами диаграмм рассеяния могут служить Рис. 1 и Рис. 3.

Накрывающая и ненакрывающая выборка

Definition

Накрывающая выборка — совокупность накрывающих измерений, т. е. выборка, в которой все измерения (наблюдения) являются накрывающими.

Напротив, выборка называется *ненакрывающей*, если хотя бы одно из входящих в неё измерений — ненакрывающее.

Решение задачи восстановления зависимостей для обычных точечных данных

Существует большое количество более или менее стандартных подходов к решению задачи восстановления зависимостей для обычных точечных данных.

Наиболее популярные из них — это метод наименьших квадратов, метод наименьших модулей и метод максимальной энтропии. Часто используется чебышёвское (минимаксное) сглаживание.

Все эти методы основаны на нахождении глобального (абсолютного) минимума определённым образом подобранный целевой функции. Мы пытаемся найти наиболее набор параметров, который доставляет минимум этому функционалу. Очевидно, что конечный результат будет существенно отличаться в зависимости от формы этого целевого функционала.

В любом случае, «идеальным решением» задачи можно считать ту функциональная зависимость вида (если она существует), линия графика которой проходит через все точки данных.

Что следует считать решением?

Что следует считать решением задачи восстановления зависимости по интервальным данным (3)?

Очевидно, что функцию, вида (1) или (2), нужно считать точным решением задачи восстановления искомой зависимости, если её график проходит через все брусы неопределённости данных.

В случае точечных данных эта идеальная ситуация почти никогда не реализуется и неустойчива к малым возмущениям в данных. Но в случае данных с существенной интервальной неопределенностью прохождение графика функции через брусы данных (3) может реализовываться, и оно устойчиво к возмущениям в данных.

Кроме того, дополнительную специфику задаче придаёт то новое обстоятельство, что брусы неопределённости данных (3), в отличие от бесконечно малых и бесструктурных точек, получают структуру и потому нужно различать, как именно проходит график функции через эти брусы.

Информационное множество

В соответствии с терминологией, намеченной для нахождения констант, будем называть *информационным множеством* задачи восстановления зависимости множество значений параметров зависимости, совместных с данными в каком-то определённом смысле.

Информационное множество

В традиционном «точечном» случае, когда данные неинтервальны, решение задачи восстановления зависимостей получается по следующей общей схеме. Мы подставляем данные в формулу для зависимости (2) и получаем для каждого отдельного измерения одно уравнение. В целом в результате этой процедуры возникает система уравнений, решив которую, в обычном или обобщённом смысле, мы найдём параметры зависимости.

В интервальном случае, действуя аналогичным образом, мы получим уже интервальную систему уравнений, которую также можно решать. Её решением, обычным или в некотором обобщённом смысле, будет вектор оценки параметров восстанавливаемой зависимости (2).

Информационное множество задачи получается при этом как множество решений этой интервальной системы уравнений, построенной на основе формулы (2) и данных (3).

Коридор совместных зависимостей

Определение параметров функциональной зависимости производится, как правило, для того, чтобы затем найденную формулу использовать для предсказания значений зависимости в других интересующих нас точках её области определения.

Ясно, что такое предсказание будет осуществляться с некоторой погрешностью, вызванной неопределённостями данных, неоднозначностью самой процедуры восстановления и т. п. Эту неопределённость предсказания также необходимо знать и учитывать в нашей деятельности.

Коридор совместных зависимостей и его сечение

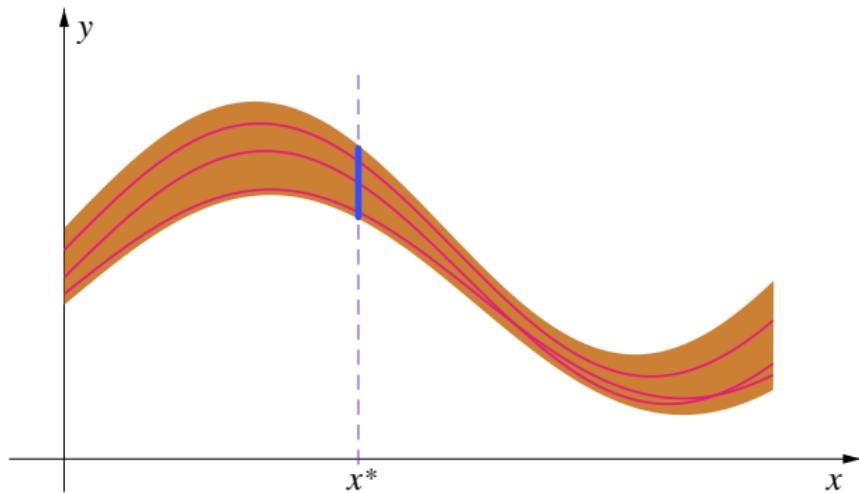


Рис.: Коридор совместных зависимостей и его сечение
для какого-то значения аргумента x^* .

Коридор совместных зависимостей

Если информационное множество задачи восстановления зависимостей непусто, то обычно оно задаёт целое семейство зависимостей, совместных с данными задачи, которое имеет смысл рассматривать вместе, как единое целое.

Это необходимо делать в вопросах, касающихся оценивания неопределённости предсказания, учёта всех возможных сценариев развития и т. п. Как следствие, возникает необходимость рассматривать вместе, единым целым, множество всех функций, совместных с интервальными данными задачи восстановления зависимости. Мы будем называть его *коридором совместных зависимостей* (см. Рис. 2).

Многозначные отображения

В литературе использовались также другие термины для обозначения этого объекта — «трубка» совместных зависимостей (имеет происхождение в теории управления), «полоса» или даже «слой неопределённости», «коридор неопределённости» и т. п.

Строгое определение коридора совместных зависимостей может быть дано на основе математического понятия многозначного отображения. Напомним, что для произвольных множеств X и Y **многозначным отображением F из X в Y** называется соответствие (правило), сопоставляющее каждой точке $x \in X$ непустое подмножество $F(x) \subset Y$, называемое **значением** или **образом x** .

Definition

Пусть в задаче восстановления зависимостей информационное множество Ω параметров зависимостей $y = f(x, \beta)$, совместных с данными, является непустым. *Коридором совместных зависимостей* рассматриваемой задачи называется многозначное отображение Υ , сопоставляющее каждому значению аргумента x множество

$$\Upsilon(x) = \bigcup_{\beta \in \Omega} f(x, \beta).$$

Сечение коридора совместных зависимостей

Значение $\Upsilon(\tilde{x})$ коридора совместных зависимостей при каком-то определённом аргументе \tilde{x} («сечение коридора») — это множество $\cup_{\beta \in \Omega} f(\tilde{x}, \beta)$, образованное всевозможными значениями, которые принимают на этом аргументе функциональные зависимости, совместные с интервальными данными измерений.

Рис. 2 изображает коридор совместных зависимостей в задаче восстановления нелинейной зависимости, но для рассматриваемого нами линейного случая коридор совместных значений имеет существенно более специальный вид .

Нетрудно показать, что границы коридора совместных зависимостей в этом случае являются кусочно-линейными.

Случай точных измерений входных переменных

Важнейшим и часто встречающимся частным случаем рассмотренной задачи является ситуация, когда независимые (экзогенные, предикторные, входные) переменные x_1, x_2, \dots, x_m измеряются точно, и вместо телесных брусов неопределённости измерений (как на Рис. 1) мы имеем отрезки прямых $(x_{i1}, x_{i2}, \dots, x_{im}, y_i)$, $i = 1, 2, \dots, n$, параллельные оси зависимой (эндогенной, критериальной, выходной) переменной (см. Рис. 3).

Именно такая постановка задачи была рассмотрена в пионерской работе Л.В. Канторовича.

Случай точных измерений входных переменных

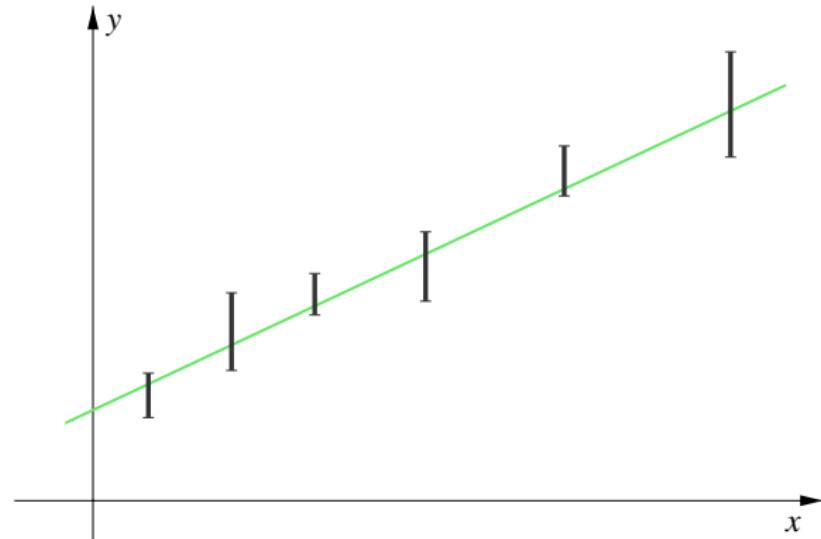


Рис.: Частный случай задачи восстановления линейной зависимости по неточным данным, когда входные переменные измеряются точно.

Постановка задачи

Отсутствие неопределённости значений независимых переменных приводит к кардинальному упрощению математической модели. Брусы неопределённости измерений зависимости, введённые ранее, схлопываясь по независимым переменным, превращаются в *отрезки неопределённости*.

Как следствие, для решения и полного исследования этого частного случая предложено большое количество эффективных вычислительных методов. Рассмотрим эти математические вопросы более детально.

Совместность зависимости с данными

Линейная зависимость (2) совместна (согласуется) с интервальными данными измерений, если её график проходит через все отрезки неопределённости, задаваемые интервалами измерений выходной переменной y , как это изображено на Рис. 3).

Подобное понимание совместности (согласования) является прямым обобщением того понимания «совместности», которое традиционно для неинтервального случая и используется, к примеру в постановке задачи интерполяции.

Совместность зависимости с данными

Подставляя в зависимость (2) данные для входных переменных x_1, x_2, \dots, x_m в i -ом измерении и требуя включения полученного значения в интервалы y_i , получим

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in y_i, \quad i = 1, 2, \dots, n. \quad (4)$$

Фактически, это интервальная система линейных алгебраических уравнений

$$\left\{ \begin{array}{l} \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1m}\beta_m = y_1, \\ \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \dots + x_{2m}\beta_m = y_2, \\ \vdots \qquad \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ \beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \dots + x_{nm}\beta_m = y_n, \end{array} \right.$$

у которой интервальность присутствует только в правой части.

Совместность зависимости с данными

С другой стороны, (4) равносильно системе

$$\left\{ \begin{array}{l} \underline{\mathbf{y}}_1 \leq \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_m x_{1m} \leq \bar{\mathbf{y}}_1, \\ \underline{\mathbf{y}}_2 \leq \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_m x_{2m} \leq \bar{\mathbf{y}}_2, \\ \vdots \quad \vdots \quad \ddots \quad \vdots \quad \vdots \\ \underline{\mathbf{y}}_n \leq \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_m x_{nm} \leq \bar{\mathbf{y}}_n. \end{array} \right. \quad (5)$$

Система двусторонних линейных неравенств

Это система двусторонних линейных неравенств относительно неизвестных параметров $\beta_0, \beta_1, \beta_2, \dots, \beta_m$, решив которую, мы можем найти искомую линейную зависимость. Множество решений системы неравенств (5) естественно считать информационным множеством параметров восстанавливаемой зависимости для рассматриваемого случая.

Для i -го двустороннего неравенства из системы (5) множество решений — это полоса в пространстве \mathbb{R}^{m+1} параметров $(\beta_0, \beta_1, \dots, \beta_m)$, ограниченная с двух сторон гиперплоскостями с уравнениями

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} = \underline{y}_i,$$

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} = \bar{y}_i.$$

Система двусторонних линейных неравенств

Множество решений системы неравенств (5) является пересечением n штук таких полос, отвечающих отдельным измерениям. Можно рассматривать эти полосы как информационные множества отдельных измерений.

На Рис. 4 изображено формирование множества решений системы неравенств (5) для случая двух параметров (т. е. $m = 1$) и $n = 3$.

Образование информационного множества параметров

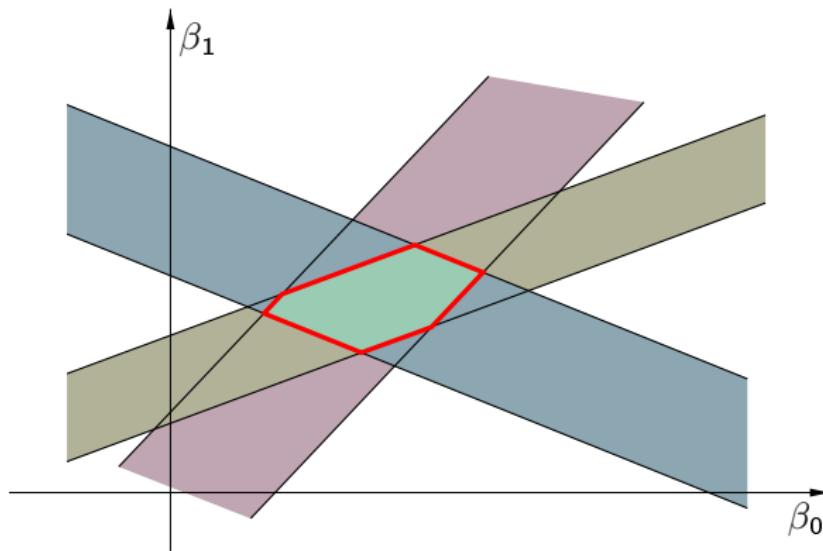


Рис.: Образование информационного множества параметров
линейной зависимости (ограничено красной линией)
для случая точных входных переменных.

Информационное множество — трудоёмкость распознавания

В целом множество решений системы линейных алгебраических неравенств (5) является *выпуклым многогранным множеством в пространстве \mathbb{R}^{m+1}* .

Распознавание того, пусто оно или непусто, а также нахождение какой-либо точки из него, являются задачами, сложность которых ограничена полиномом от их размера. Существуют эффективные и хорошо разработанные вычислительные методы для решения этих вопросов и для нахождения оценок множества решений, например, основанные на сведении рассматриваемой задачи к задаче линейного программирования.

Информационное множество — трудоёмкость распознавания

В общем случае, когда входные (экзогенные, предикторные) переменные известны неточно, ситуация существенно усложняется и множество параметров, совместных (согласующихся) с интервальными данными не может быть описано так же просто, с помощью системы линейных неравенств (5).

Трудоёмкость распознавания его пустоты или непустоты также становится экспоненциальной в зависимости от количества переменных [2].

Пример

Случай точных измерений входных переменных

Общий случай задачи восстановления зависимостей

Рассмотрим теперь случай, когда неопределённость присутствует как в измерениях значений зависимой переменной, так и в измерениях значений аргументов.

Это может быть вызвано различными причинами. Например, существенно неточное измерение входных переменных происходит в ситуациях, когда они должны устанавливаться в течение значительного времени.

Тогда их уместно выразить какими-то интервалами, а не точечными значениями.

Пример

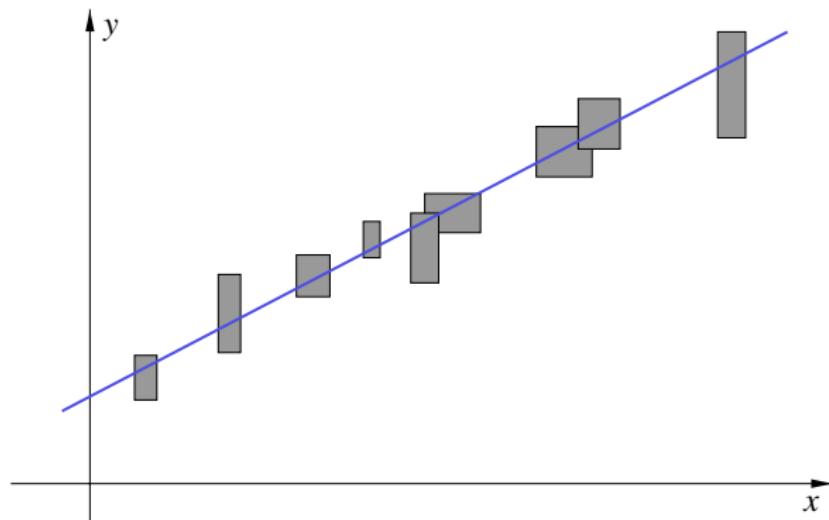


Рис.: Наглядная иллюстрация задачи восстановления линейной зависимости по данным с интервальной неопределенностью.

Пример

[https://github.com/szhilin/octave-interval-examples/blob/
master/SteamGenerator.ipynb.](https://github.com/szhilin/octave-interval-examples/blob/master/SteamGenerator.ipynb)

Общий случай задачи восстановления зависимостей

Если выборка измерений независимых переменных и зависимой переменной — накрывающая, то

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n,$$

где все x_{ij} могут принимать значения из соответствующих интервалов $x_{ij}, i = 1, 2, \dots, n, j = 1, 2, \dots, m$. Как следствие, получаем интервальную систему линейных алгебраических уравнений

$$\left\{ \begin{array}{l} \beta_0 + \mathbf{x}_{11}\beta_1 + \mathbf{x}_{12}\beta_2 + \dots + \mathbf{x}_{1m}\beta_m = \mathbf{y}_1, \\ \beta_0 + \mathbf{x}_{21}\beta_1 + \mathbf{x}_{22}\beta_2 + \dots + \mathbf{x}_{2m}\beta_m = \mathbf{y}_2, \\ \vdots \qquad \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ \beta_0 + \mathbf{x}_{n1}\beta_1 + \mathbf{x}_{n2}\beta_2 + \dots + \mathbf{x}_{nm}\beta_m = \mathbf{y}_n. \end{array} \right. \quad (6)$$

Общий случай задачи восстановления зависимостей

Это формальная запись, означающая совокупность обычных (точечных) систем линейных алгебраических уравнений того же размера и с теми же неизвестными переменными, у которых коэффициенты и правые части лежат в предписанных им интервалах (см. [2]).

Восстановление параметров линейной зависимости можно рассматривать как «решение», в том или ином смысле, выписанной интервальной системы уравнений.

Общий случай задачи восстановления зависимостей

В случае присутствия погрешностей как в измерениях аргумента, так и в измерениях зависимости множество параметров зависимостей, совместных (согласующихся) с данными, характеризуются новыми свойствами, которыми не обладают задачи с точными измерениями входных переменных.

Прежде всего, множества решений отдельных интервальных уравнений уже *не являются полосами в пространстве \mathbb{R}^n* , вроде тех, что изображены на Рис. 4. Они выглядят существенно иначе, и их конкретный вид зависит от того, какой смысл вкладывается в понятие совместности (согласования) параметров и данных, т. е. от того, *какое множество решений ИСЛАУ взято в качестве информационного множества* (см. Рис. 6).

Пример

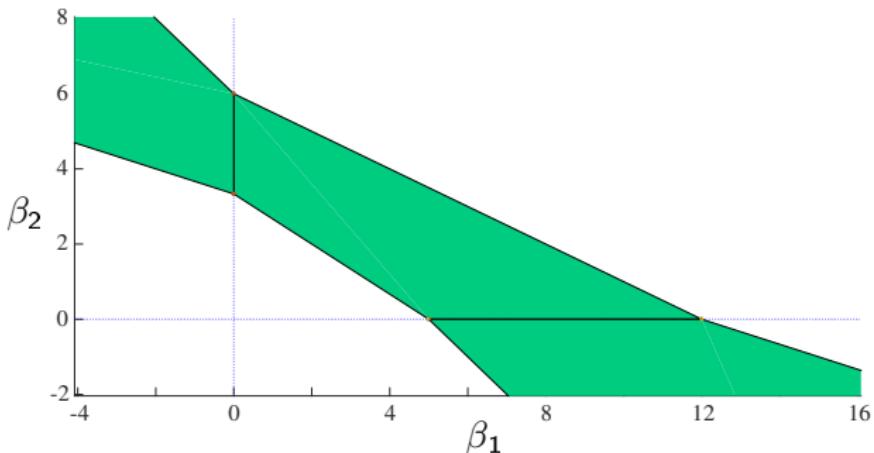


Рис.: Объединённое множество решений интервального линейного уравнения $[1, 2]\beta_1 + [2, 3]\beta_2 = [10, 12]$.

Общий случай задачи восстановления зависимостей

Само понятие согласования (совместности) параметров и данных должно быть расширено и переосмыслено.

В обычном неинтервальном случае результаты измерений — это бесконечно малые точки, и прохождение через них графика функциональной зависимости адекватно описывается двумя значениями — «да» или «нет», т. е. имеет булевский (логический) тип данных.

Общий случай задачи восстановления зависимостей

Если мы переходим от точек к брусьям неопределённости, то прохождение графика зависимости через них можно понимать по-разному.

Брусы неопределённости измерений являются прямыми декартовыми произведениями интервалов по различным осям координат, и эти оси имеют разный смысл:

интервалы $x_{i1}, x_{i2}, \dots, x_{im}$ соответствуют входным (экзогенным, предикторным) переменным,
а интервал y ; соответствует выходной (эндогенной, критериальной) переменной.

По этой причине становится важным, как именно проходит график восстанавливаемой зависимости через брусы неопределённости измерений (см. Рис. 7).

Общий случай задачи восстановления зависимостей

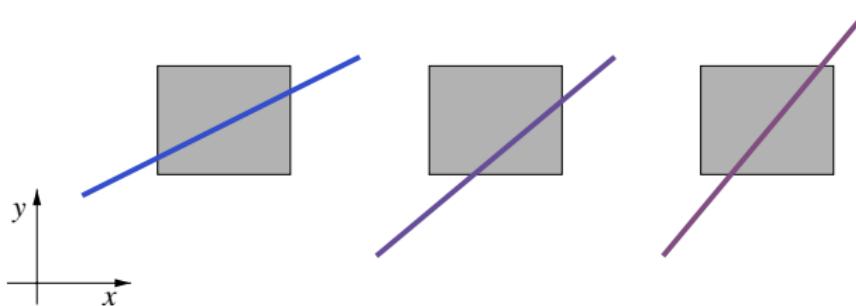


Рис.: Различные способы пересечения линии с бруском
неопределённости измерения зависимости.

Слабо совместная зависимость

Функциональную зависимость назовём *слабо совместной* с интервальными данными, если её график проходит через каждый брус неопределённости измерений хотя бы для одного значения аргумента.

Наглядно это означает, что график зависимости пересекает брусы неопределённости, но как именно — неважно (средний чертёж на Рис. 7),

достаточно лишь одной точки пересечения.

достаточно лишь одной точки пересечения.

Слабо совместная зависимость

Для случая линейной зависимости это условие наиболее удобно выразить с помощью формального языка логического исчисления предикатов:

$$(\exists x_{i1} \in \mathbf{x}_{i1}) \cdots (\exists x_{im} \in \mathbf{x}_{im}))(\exists y_i \in \mathbf{y}_i) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Равносильная упрощённая формулировка этого свойства выглядит следующим образом:

$$(\exists x_{i1} \in \mathbf{x}_{i1}) \cdots (\exists x_{im} \in \mathbf{x}_{im}) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Сильно совместная зависимость

Функциональную зависимость назовём *сильно совместной* с интервальными данными, если её график проходит через каждый брус неопределённости измерений для любого значения аргумента из интервалов неопределённости входных переменных.

Наглядно это означает, что график зависимости

целиком содержится в коридорах,

задаваемых интервалами выходной переменной при всех значениях входных переменных из соответствующих им интервалов

(левый чертёж на Рис. 7).

Сильно совместная зависимость

Для случая линейной зависимости это условие может быть формально записано в следующем виде:

$$(\forall x_{i1} \in \mathbf{x}_{i1}) \cdots (\forall x_{im} \in \mathbf{x}_{im})(\exists y_i \in \mathbf{y}_i) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Равносильная упрощённая формулировка этого свойства выглядит следующим образом:

$$(\forall x_{i1} \in \mathbf{x}_{i1}) \cdots (\forall x_{im} \in \mathbf{x}_{im}) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Сильно и слабо совместные зависимости

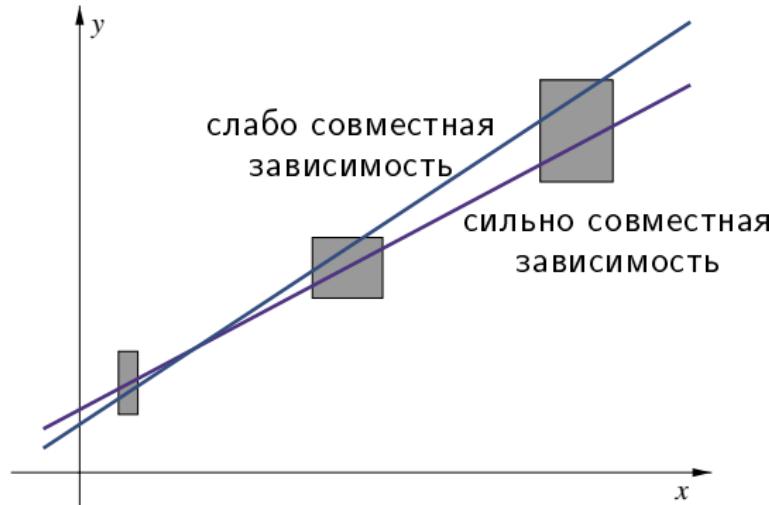


Рис.: Линейные зависимости с разными типами
согласования с данными.

Сильно совместная зависимость

В чём содержательный смысл сильной совместности?

На практике измерения на входах и выходах системы осуществляются, как правило, разными способами и даже в разное время.

Мы измеряем выход (зависимую переменную) уже тогда, когда входные значения (независимых переменных) зафиксированы, и мы их измерили. Получив при этом какие-то интервалы.

Сильная совместность функциональной зависимости с интервальными данными означает тогда, что выходная величина остаётся в пределах измеренного для неё интервала вне зависимости от того, какими конкретно в своих интервалах являются значения входных переменных.

Сильно совместная зависимость

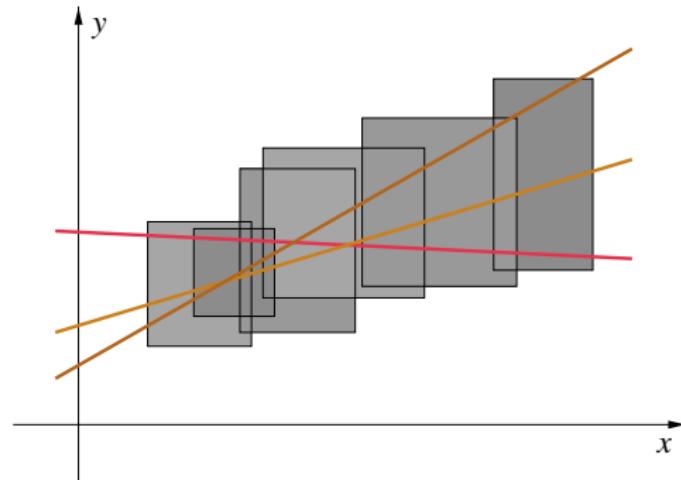


Рис.: Сложный случай восстановления зависимости
по широким перекрывающимся интервальным данным.

Множества решений

Если матрица системы (6) уравнений — точечная, т. е. коэффициенты при неизвестных β_i являются обычными вещественными числами, то объединённое множество решений в целом является выпуклым.

Но в общем случае, когда матрица интервальной системы линейных алгебраических уравнений существенно интервальна, то объединённое множество решений может быть невыпуклым.

Допусковое множество решений всегда выпукло. В целом, количество гиперплоскостей, ограничивающих множества решений, может быть очень большим.

Приближённое описание информационного множества

Возвращаясь к решению задачи восстановления зависимостей, следует отметить, что непростое строение множеств решений интервальных систем уравнений делает очень трудоёмким и малополезным их точное и полное описание.

Имеет смысл найти какое-нибудь приближённое описание информационного множества.

Здесь могут встретиться различные ситуации.

Приближённое описание информационного множества

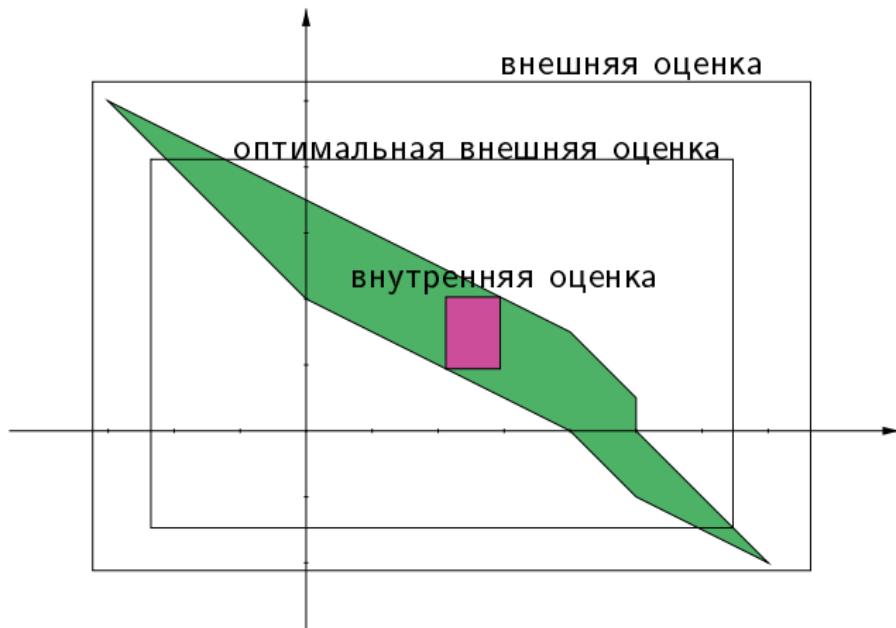


Рис.: Различные способы оценивания

информационного множества.

Оценки информационного множества

Часто бывает необходимо оценить разброс точек из информационного множества, то есть определить, насколько сильно оно «растекается» в пространстве параметров.

Часто это делается для его отдельных компонент, так что в целом нам требуется интервальный брус, содержащий множество решений. Это *внешняя оценка информационного множества*

Среди всех внешних оценок наилучшей служит минимальная по размерам внешняя оценка, которую также называют *оптимальной внешней оценкой*. Она единственна и является интервальной оболочкой информационного множества задачи.

Внешняя оценка информационного множества необходима, к примеру, при построении внешней оценки коридора совместных зависимостей, когда мы хотим просчитать гарантированный эффект от реализации всех сценариев, могущих встретиться по восстановленным зависимостям.

Оценки информационного множества

Во многих задачах требуется оценивание информационного множества с помощью какого-то несложно описываемого подмножества — *внутреннее оценивание*. Такая оценка будет содержать только точки из информационного множества и ничего лишнего.

Внешняя оценка информационного множества в этом смысле плоха тем, что включает в себя точки, не принадлежащие информационному множеству.

Если в качестве подмножества информационного множества берётся вписанный брус, то он называется *внутренней интервальной оценкой* множества решений. Среди двух внутренних оценок лучшей является та, которая целиком содержит другую, но максимальных по включению внутренних оценок, которые несравнимы друг с другом, может быть много.

Оценки информационного множества

Английские термины для обозначения внешней и внутренней оценки — outer estimate и inner estimate соответственно. Внешнюю оценку часто называют также термином «closure».

Кроме внешнего и внутреннего оценивания информационных множеств могут встретиться и другие, которые требуются по смыслу задачи.

Например, «слабое внешнее» оценивание , оценивание вдоль какого-то специального выделенного направления, исчерпывающее оценивание с помощью набора брусов и т.п.

Варианты точечной оценки информационного множества

Помимо оценивания информационного множества «целиком», во многих ситуациях достаточно найти какую-либо точку из него (здесь мы имеем аналогию с оцениванием «точечным» и «интервальным» в традиционной статистике). Естественно выбирать такую одну точку удовлетворяющей некоторым условиям оптимальности.

Варианты точечной оценки информационного множества

- центр интервального бруса, который является минимальной по включению внешней оценкой информационного множества,
- центр Оскорбина,
- чебышёвский центр,
- центр тяжести,
- точка максимума совместности (аргумент максимума распознающего функционала, который является точкой максимума совместности соответствующей интервальной системы уравнений).

Пример обработки накрывающей выборки

Пример обработки накрывающей выборки.

Пример обработки накрывающей выборки

Пример иллюстрирует практическое применение методики главы «Задача восстановления зависимостей» книги «Обработка и анализ данных с интервальной неопределённостью» [1].

Технологически изложение следует канве, представленной в виде блокнота на ресурсе С.Жилина [3].

Набор данных.

При измерении параметров шагового двигателя была получена зависимость положения вала от времени.

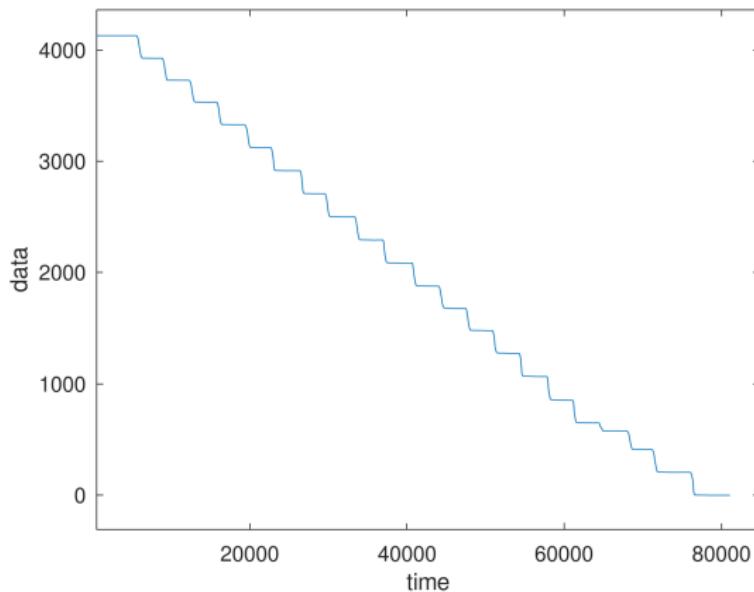
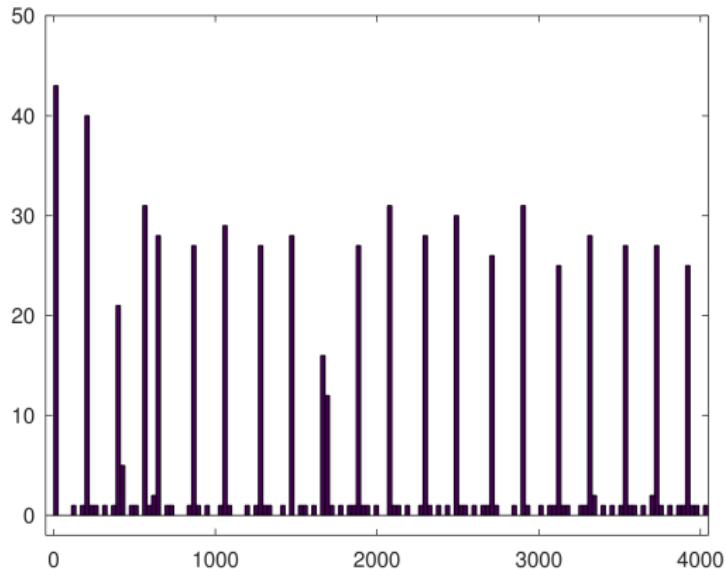


Рис.: Положение вала от времени. Данные энкодера углового перемещения.

Гистограмма данных.

На Рис. 11 горизонтальные участки соответствуют устойчивым положениям вала, а вертикальные — его повороту. Для выделения устойчивых положений, построим гистограмму



Устойчивые положения.

На основании данных гистограммы Рис. 12 можно выделить устойчивые положения как те, в которых двигатель находился больше какого-то времени. Таким образом приходим к зависимости положения вала от номера шага.

Рис. 13 подобен Рис. 11 с заменой горизонтальных участков данных на одиночные значения. Сдвинуто начало отсчета энкодера, чтобы работать с более удобными для визуальной оценки числами. Также для удобства график показан возрастающим по коду энкодера.

Устойчивые положения.

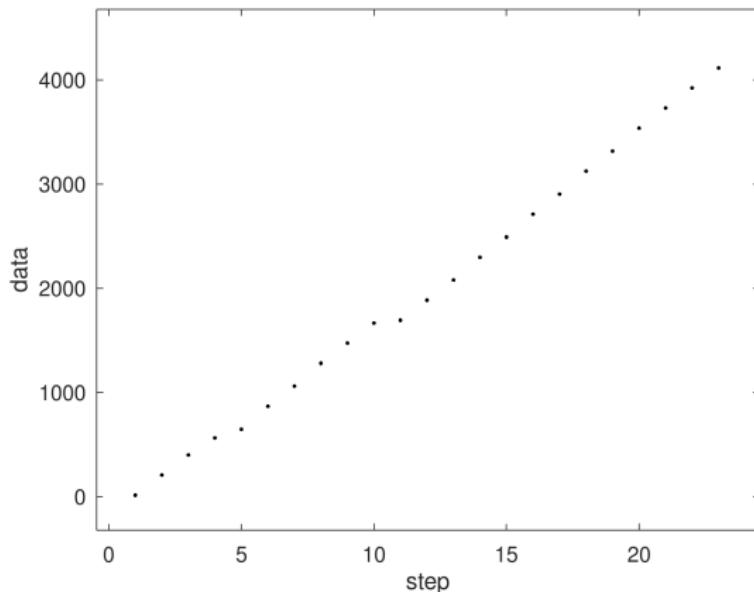


Рис.: Зависимость положения вала от номера шага.

Рабочая выборка.

Для удобства восприятия, выберем 10 значений замеров из числа данных, представленных на Рис. 13. Конкретно выбрано 10 первых нечётных значений для статических положений вала двигателя.

Номер измерения	Данные энкодера
1	399
2	646
3	1059
4	1472
5	1692
6	2078
7	2491
8	2904
9	3316
10	3729

Таблица: Частичная выборка данных.

Точечная оценка параметров регрессии.

Данные энкодера выдаются в виде целых десятичных значений, так что неопределённость представления — младший десятичный разряд. Реально погрешность, как мы увидим, существенно выше, и включает много факторов, о части которых неизвестно ничего.

В качестве первого подхода к проблеме, проведем точечную оценку параметров регрессии. Пусть модель задаётся в классе линейных функций

$$y = \beta_1 + \beta_2 x, \tag{7}$$

x — номер измерения в выборке Табл. 1,

y — угол поворота вала двигателя.

Точечная оценка параметров регрессии.

Для согласования с данными поставим задачу оптимизации и решим её методами линейного программирования [1].

$$\text{mid } \mathbf{y}_i - w_i \cdot \text{rad } \mathbf{y}_i \leq X\beta \leq \text{mid } \mathbf{y}_i + w_i \cdot \text{rad } \mathbf{y}_i, \quad i = 1, m, \quad (8)$$

$$\sum_{i=1}^m w_i \longrightarrow \min \quad (9)$$

$$w_i \geq 0, \quad i = 1, m, \quad (10)$$

$$w, \beta = ? \quad (11)$$

Здесь X — матрица $m \times 2$, в первом столбце которой элементы, равные 1, во втором — значения x_i .

В качестве значений середины и радиуса возьмём $\text{mid } \mathbf{y}_i = y_i$ и $\text{rad } \mathbf{y}_i = 1$.

Уравнение регрессионной прямой получилось

$$y = 0.0 + 363.13 \cdot x.$$

Точечная оценка параметров регрессии.

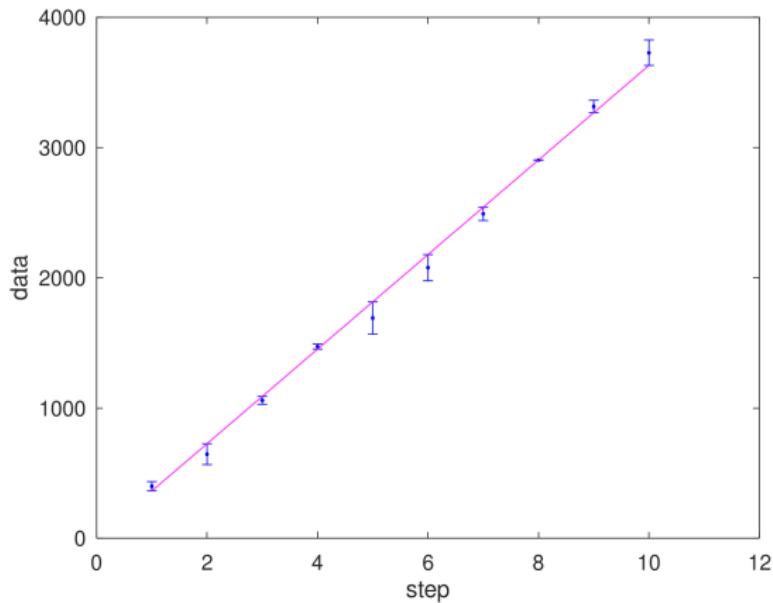


Рис.: Регрессия с оценкой по норме L_1 .

Вектор весов достижения совместности.

Вектор весов w радиусов отдельных замеров изображен на Рис. 15.

Вместе с Рис. 14, высокая неоднородность значений w свидетельствует о разной по величине степени отклонении данных от регрессионной прямой на разных участках оси абсцисс.

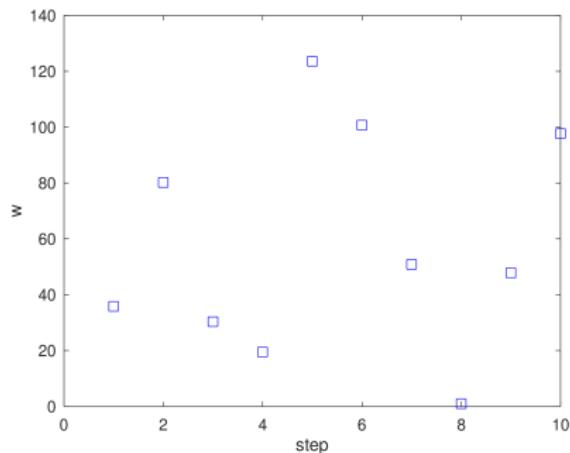


Рис.: Значения весов в задаче оптимизации.

Модель погрешности данных.

Приступим к интервальной оценке параметров регрессии. Ясно, при достаточно высокой погрешности данных выборка станет накрывающей или, по крайней мере совместной, согласно терминологии [1].

Для этого необходимо приписать данным какие-то погрешности. Значения компонент вектора w несут индивидуальную информацию о каждом измерении. Такая информация обладает высокой степенью избыточности, и желательно её заменить на более экономное представление.

Как видно из Рис. 15, имеет смысл в качестве первой оценки реалистичной погрешности данных взять близкой к максимальному значению w . Итак, пусть значение

$$\text{rad } \mathbf{y}_i := \varepsilon = \max_i w_i \simeq 150.$$

Диаграмма рассеяния данных.

Приведём диаграмму рассеяния данных для конкретного $\varepsilon = 150$ — Рис. 16.

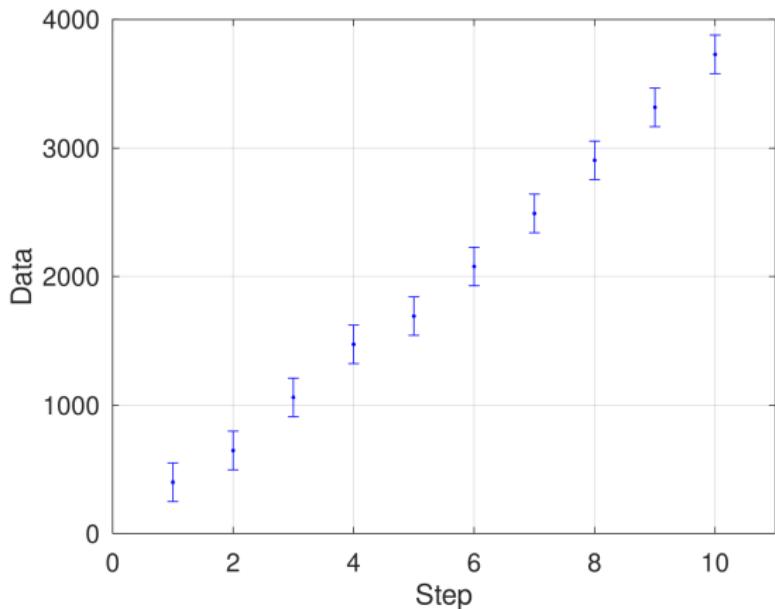
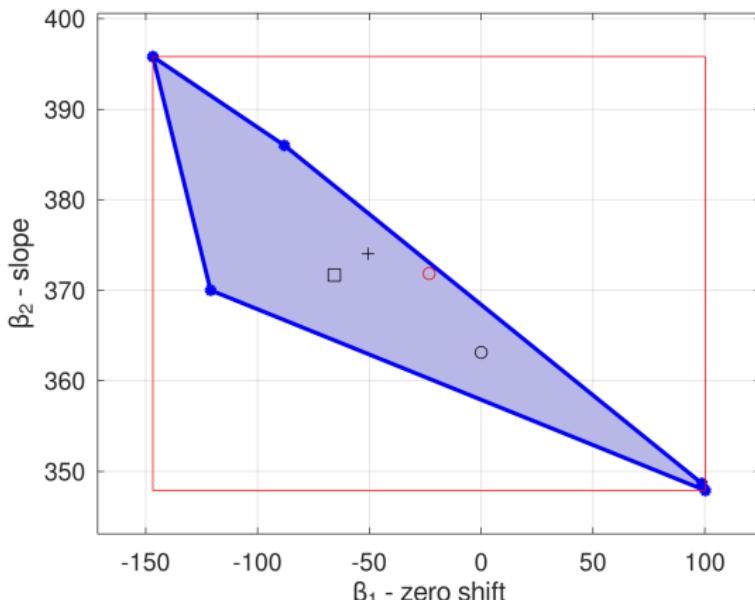


Рис.: Диаграмма рассеяния, погрешность данных $\varepsilon = 150$.

Информационное множество параметров I.

Определим теперь интервальные параметры регрессии по методике [3]. На Рис. 17 приведено информационное множество сдвигов и наклонов регрессионной прямой. Оно ограничено многоугольником и дано заливкой.



Информационное множество параметров I.

Также на Рис. 17 приведены различные точечные оценки.

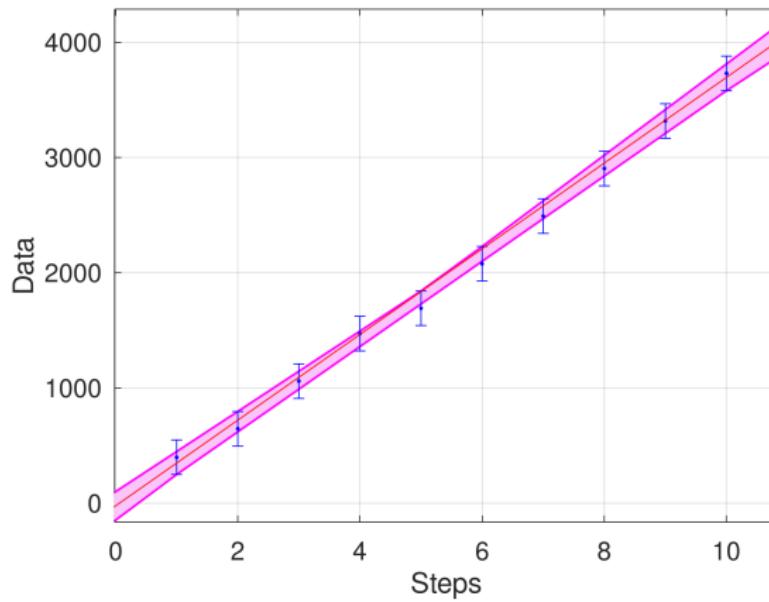
Они достигнуты вычислением

- максимальной диагонали,
- центра тяжести,
- методом наименьших квадратов,
- точечной регрессией.

Для заданного значения погрешности данных все точечные оценки содержатся в информационном множестве.

Коридор совместности Υ .

На Рис. 18 приведены диаграмма рассеяния данных и коридор совместности параметров модели регрессии для заданной погрешности данных.



Коридор совместности Υ .

Также дана прямая регрессии по параметрам, соответствующим центру тяжести множества, показанного на Рис. 17.

Видно, что для значения независимой переменной, равному 5, эта прямая касается границ коридора совместности.

То есть, в этом месте имеется «излом» множества Υ .

Прогноз значений выходной переменной.

Важнейшим назначением регрессионной модели является предсказание значений выходной переменной для заданных значений входной.

С помощью построенной модели — Рис. 18

$$y(x) = [-150, 100] + [348, 395] \cdot x \quad (12)$$

можно получить прогнозные значения выходной переменной в точках эксперимента.

Прогноз значений выходной переменной.

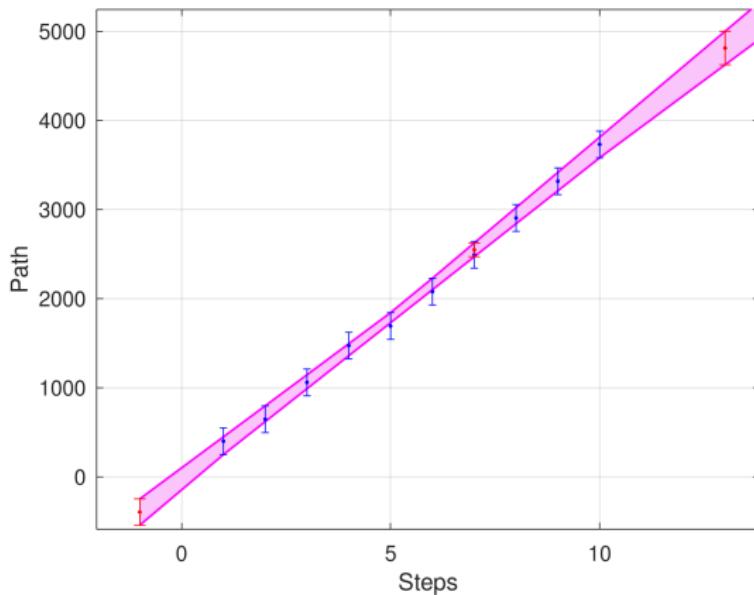


Рис.: Прогноз значений внутри и вне интервала имеющихся данных, погрешность данных $\varepsilon = 150$.

Прогноз значений выходной переменной.

Ценность модели заключается в возможности её употребления для предсказания выходной переменной в точках, где измерения не производились. Для иллюстрации приведём прогнозы в одной точке внутри диапазона $x = 7$ и двух точках за его границами $x = -1, 13$. Результаты расчётов представлены в Табл. 2.

i	x_i	mid \mathbf{y}	rad \mathbf{y}_i	$\underline{\mathbf{y}}_i$	$\bar{\mathbf{y}}_i$
1	-1	-395.11	147.487	-542.60	-247.62
2	7	2546.40	77.400	2469.00	2623.80
3	13	4810.61	187.987	4622.62	4998.60

Таблица: Прогноз измерений по модели (12).

Прогноз значений выходной переменной.

Погрешность прогноза для «внутренней» точки $x = 7$ составляет $\simeq 77$ кодов энкодера и меньше назначеннной погрешности 150.

При выборе точек прогноза со значениями -1 и 13 за пределами диапазона данных, даёт соответственно погрешность прогноза $\simeq 147$ и $\simeq 188$.

Чем более удалена точка прогноза от области данных, тем больше предсказываемая погрешность.

Уточнение модели погрешности данных.

Итак, при значении погрешности данных, равной $\varepsilon = 150$, получены согласованные оценки параметров линейной модели данных (12).

Напомним, что величина ε выбрана «с запасом» из соображений обеспечения заведомого согласования данных и линейной модели.

Посмотрим, что произойдёт при попытке уменьшить эту неопределённость. Пусть $\varepsilon = 100$.

Определим интервальные параметры регрессии. На Рис. 20 приведено новое информационное множество сдвигов и наклонов регрессионной прямой.

Информационное множество.

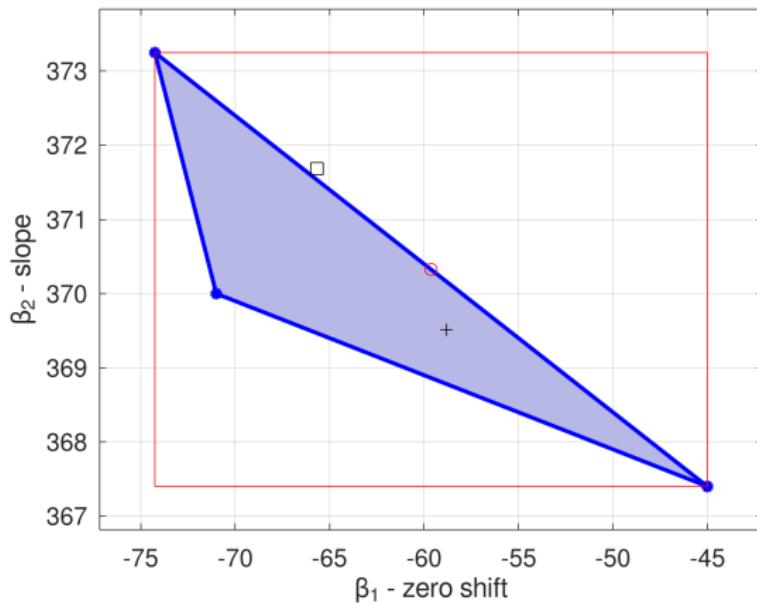


Рис.: Информационное множество, погрешность данных $\varepsilon = 100$.

Множество параметров линейной модели.

Множество параметров линейной модели на Рис. 20 существенно меньше аналогичного множества Рис. 17. Конкретные значения ширин параметров β приведены в Табл. 3.

ε	wid β_1	wid β_2
100	$\simeq 29$	$\simeq 4$
150	$\simeq 250$	$\simeq 46$

Таблица: Размеры множества параметров линейной модели.

информационное множество.

Таким образом, информационное множество очень уменьшилось в размерах: примерно на десятичный порядок по каждой компоненте.

Согласование становится в таких условиях весьма проблематичным. В частности, оценка точечных параметров модели методом наименьших квадратов (черный квадратик на Рис. 20) находится за пределами I .

Уменьшение информационного множества приводит к сужению коридора совместности параметров модели. На Рис. 21 приведены диаграмма рассеяния данных и коридор совместности параметров модели регрессии Υ для заданной погрешности данных.

Коридор совместности.

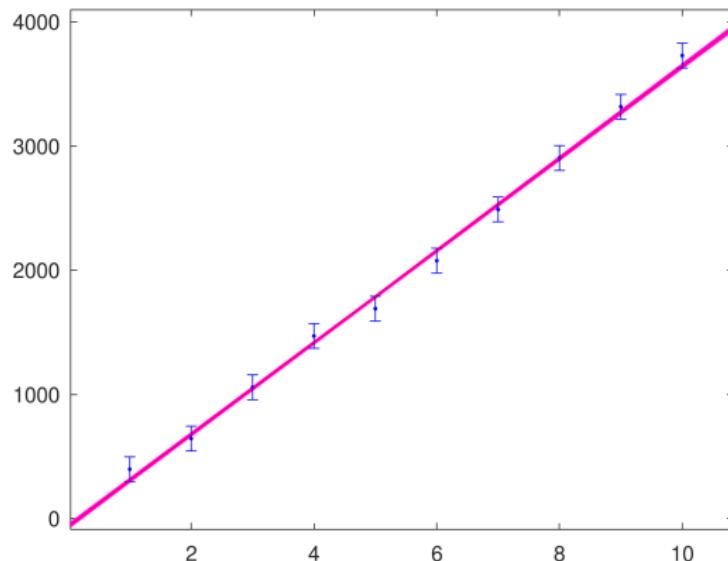


Рис.: Диаграмма рассеяния и коридор совместности \mathcal{Y} , погрешность данных $\varepsilon = 100$.

Коридор совместности.

Коридор совместности Υ представляет собой узкую полосу, проходящую через крайние значения нескольких брусов.

Именно, коридор совместности касается вершин брусов

- \underline{y}_1 ,
- \bar{y}_5, \bar{y}_6 ,
- \underline{y}_{10} .

Как уже было замечено ранее, в середине графика имеется «излом».

Дальнейшее уменьшение ε приводит к пустоте множества параметров. Выборка становится ненакрывающей.

Границные измерения — Лекция 4.

Границными называют измерения, определяющие какой-либо фрагмент границы информационного множества. Очевидно, это свойство имеет смысл рассматривать для наблюдений, принадлежащих выборке S_n , по которой сконструирована модель и её информационное множество $\Omega(S_n)$.

Подмножество всех границных наблюдений в S_n играет особую роль, поскольку оно является

минимальной подвыборкой, полностью определяющей модель.

Удаление неграницных наблюдений из выборки не изменяет модель.

Пример обработки накрывающей выборки — заключение.

В приведённом примере была продемонстрирована технология обработки выборки с *неизвестной заранее погрешностью данных*.

Выбором модели погрешностей выборка была сделана *накрывающей*.

Далее было показано, что при занижении погрешности данных происходит уменьшение информационного множества вплоть до его пустоты.

Пример обработки ненакрывающей выборки

Пример обработки ненакрывающей выборки.

Литература

-  А.Н. Баженов, С.И. Жилин, С.И. Кумков, С.П. Шарый.
Обработка и анализ данных с интервальной неопределенностью.
РХД. Серия «Интервальный анализ и его приложения». Ижевск.
2021. с.200.
-  С.П. Шарый. Конечномерный интервальный анализ. —
Новосибирск: XYZ, 2021. — Электронная книга, доступная на
<http://interval.ict.nsc.ru/Library/InteBooks/SharyBook.pdf>
-  С.И.Жилин. Примеры анализа интервальных данных в Octave
<https://github.com/szhilin/octave-interval-examples>
-  С.И.Жилин. Библиотека полной интервальной арифметики
kinterval в среде Octave. Частное сообщение.

Тема X3. Обработка и анализ данных с интервальной неопределенностью.

А.Н. Баженов

Санкт-Петербургский политехнический университет Петра Великого

a_bazhenov@inbox.ru

05.10.2021

Обработка и анализ данных с интервальной неопределённостью.

ПЛАН

ПЛАН

- Общие понятия
- Обработка константы
- Задача восстановления зависимостей

Теория:

А.Н. Баженов, С.И. Жилин, С.И. Кумков, С.П. Шарый.
Обработка и анализ данных с интервальной неопределенностью. РХД.
Серия «Интервальный анализ и его приложения». Ижевск. 2021. с.200.

ПЛАН

Задача восстановления зависимостей. Часть 2.

Задача восстановления зависимостей

Даются определения новых терминов и понятий, которые возникают в связи с восстановлением функциональных зависимостей по данным их измерений и наблюдений, имеющих интервальную неопределённость.

Мы рассмотрим основные идеи и типичные приёмы восстановления зависимостей по интервальным данным, а также возникающие при этом проблемы.

Подробно исследуется случай простейшей линейной зависимости, но большинство построений и рассуждений легко переносятся на общий нелинейный случай.

Постановка задачи

Предположим, что величина y является функцией некоторого заданного вида от независимых аргументов x_1, x_2, \dots, x_m , т. е.

$$y = f(x, \beta), \quad (1)$$

где $x = (x_1, \dots, x_m)$ — вектор независимых переменных,
 $\beta = (\beta_1, \dots, \beta_l)$ — вектор параметров функции. Имея набор значений переменных x и y , нам нужно найти β_1, \dots, β_l , которые соответствуют конкретной функции f из параметрического семейства (1).

Мы будем называть эту задачу *задачей восстановления зависимости*.

Постановка задачи

Важнейший частный случай поставленной задачи — определение параметров линейной функциональной зависимости вида

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m, \quad (2)$$

в которой x_1, x_2, \dots, x_m — независимые переменные (которые называются также *экзогенными*, *предикторными* или просто *входными* переменными), y — это зависимая переменная (которая называется также *эндогенной*, *критериальной* или *выходной* переменной), а $\beta_0, \beta_1, \dots, \beta_m$ — некоторые коэффициенты.

Эти неизвестные коэффициенты должны быть определены из ряда измерений значений x_1, x_2, \dots, x_m и y .

Постановка задачи

Результаты измерений неточны, и мы предполагаем что они имеют ограниченную неопределенность, когда нам известны лишь некоторые интервалы, дающие двусторонние границы измеренных значений.

Таким образом, результатом i -го измерения являются такие интервалы $x_1^{(i)}, x_2^{(i)}, \dots, x_m^{(i)}, y^{(i)}$, относительно которых мы предполагаем, что истинное значение x_1 лежит в пределах $x_1^{(i)}$, истинное значение x_2 лежит в $x_2^{(i)}$ и т.д. вплоть до y , истинное значение которого находится в интервале $y^{(i)}$.

В целом имеется n измерений, так что индекс i может принимать значения из множества натуральных чисел $\{1, 2, \dots, n\}$.

Постановка задачи

Далее для удобства построений и выкладок обозначим номер измерения i не верхним, а нижним индексом, который мы поставим первым при обозначении входов. Таким образом, полный набор данных будет иметь вид

$$\begin{aligned} & x_{11}, \quad x_{12}, \quad \dots \quad x_{1m}, \quad y_1, \\ & x_{21}, \quad x_{22}, \quad \dots \quad x_{2m}, \quad y_2, \\ & \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ & x_{n1}, \quad x_{n2}, \quad \dots \quad x_{nm}, \quad y_n. \end{aligned} \tag{3}$$

Нам необходимо найти или как-то оценить коэффициенты β_j , $j = 0, 1, \dots, m$, для которых линейная функция (2) «наилучшим образом» приближала бы интервальные данные измерений (3).

Постановка задачи

Для обозначения $n \times m$ -матрицы, составленной из данных (3) для независимых переменных часто используют термины **матрица плана эксперимента** или просто **матрица плана**, которые возникли в теории планирования эксперимента .

Интервалы $x_{i1}, x_{i2}, \dots, x_{im}, y_i$ мы называем, как и раньше, **интервалами неопределённости i -го измерения**.

Но кроме них нам также потребуется обращаться ко всему множеству, ограничеваемому в многомерном пространстве \mathbb{R}^{m+1} этими интервалами по отдельным координатным осям.

Брус неопределённости

Definition

Брусом неопределённости i -го измерения рассматриваемой зависимости будем называть интервальный вектор-брус $(x_{i1}, x_{i2}, \dots, x_{im}, y_i) \subset \mathbb{R}^{m+1}$, $i = 1, 2, \dots, n$.

Таким образом, каждый брус неопределённости измерения зависимости является прямым декартовым произведением интервалов неопределённости независимых переменных и зависимой переменной. На Рис. 1 на плоскости Oxy наглядно показаны брусы неопределённости измерений и график линейной функции, которую мы восстанавливаем.

Далее мы рассматриваем данные (3) как «спущенные свыше» и никак не обсуждаем их выбор, коррекцию или оптимизацию.

Пример

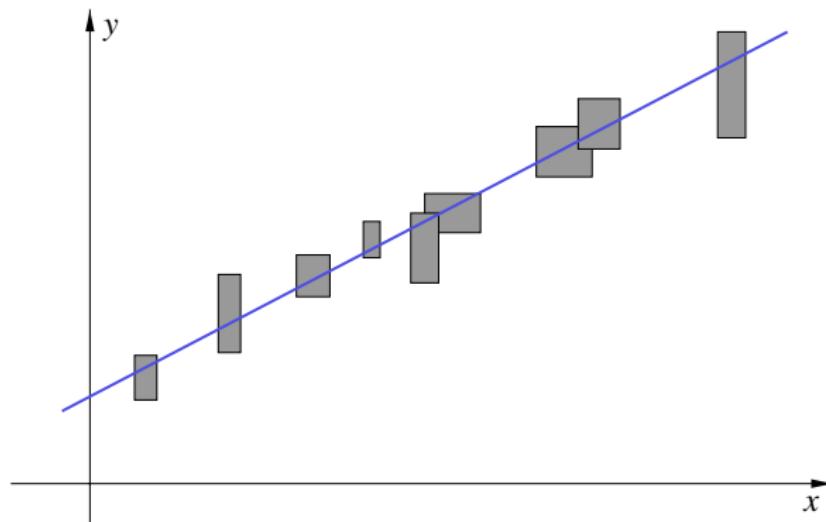


Рис.: Наглядная иллюстрация задачи восстановления линейной зависимости по данным с интервальной неопределённостью.

Накрывающие и ненакрывающие брусы

Definition

Будем называть брус неопределённости измерения зависимости **накрывающим**, если он гарантированно содержит истинные значения измеряемых величин входных и выходных переменных зависимости.

Брус неопределённости измерения зависимости, который не является накрывающим, будем называть **ненакрывающим**.

Возможные альтернативные термины — «включающий брус неопределённости», «охватывающий брус неопределённости» (их отрицание — «невключающий», «неохватывающий»).

Диаграммы рассеяния

Для визуализации интервальных данных, аналогично традиционному точечному случаю, используют *диаграммы рассеяния*.

В традиционном понимании диаграмма рассеяния используется в статистике и анализе данных для визуализации значений двух переменных в виде «облака» точек на декартовой плоскости и позволяет оценить наличие или отсутствие корреляции и других взаимосвязей между двумя переменными.

На диаграмме рассеяния для интервальных данных каждое интервальное наблюдение отображается в виде бруса (брюса неопределённости). При отсутствии неопределенности по одной из переменных, брусы наблюдений могут «схлопываться» в одномерные вертикальные или горизонтальные отрезки («ворота»).

Примерами диаграмм рассеяния могут служить Рис. 1 и Рис. 3.

Накрывающая и ненакрывающая выборка

Definition

Накрывающая выборка — совокупность накрывающих измерений, т. е. выборка, в которой все измерения (наблюдения) являются накрывающими.

Напротив, выборка называется *ненакрывающей*, если хотя бы одно из входящих в неё измерений — ненакрывающее.

Решение задачи восстановления зависимостей для обычных точечных данных

Существует большое количество более или менее стандартных подходов к решению задачи восстановления зависимостей для обычных точечных данных.

Наиболее популярные из них — это метод наименьших квадратов, метод наименьших модулей и метод максимальной энтропии. Часто используется чебышёвское (минимаксное) сглаживание.

Все эти методы основаны на нахождении глобального (абсолютного) минимума определённым образом подобранный целевой функции. Мы пытаемся найти наиболее набор параметров, который доставляет минимум этому функционалу. Очевидно, что конечный результат будет существенно отличаться в зависимости от формы этого целевого функционала.

В любом случае, «идеальным решением» задачи можно считать ту функциональная зависимость вида (если она существует), линия графика которой проходит через все точки данных.

Что следует считать решением?

Что следует считать решением задачи восстановления зависимости по интервальным данным (3)?

Очевидно, что функцию, вида (1) или (2), нужно считать точным решением задачи восстановления искомой зависимости, если её график проходит через все брусы неопределённости данных.

В случае точечных данных эта идеальная ситуация почти никогда не реализуется и неустойчива к малым возмущениям в данных. Но в случае данных с существенной интервальной неопределённостью прохождение графика функции через брусы данных (3) может реализовываться, и оно устойчиво к возмущениям в данных.

Кроме того, дополнительную специфику задаче придаёт то новое обстоятельство, что брусы неопределённости данных (3), в отличие от бесконечно малых и бесструктурных точек, получают структуру и потому нужно различать, как именно проходит график функции через эти брусы.

Информационное множество

В соответствии с терминологией, намеченной для нахождения констант, будем называть *информационным множеством* задачи восстановления зависимости множество значений параметров зависимости, совместных с данными в каком-то определённом смысле.

Информационное множество

В традиционном «точечном» случае, когда данные неинтервальны, решение задачи восстановления зависимостей получается по следующей общей схеме. Мы подставляем данные в формулу для зависимости (2) и получаем для каждого отдельного измерения одно уравнение. В целом в результате этой процедуры возникает система уравнений, решив которую, в обычном или обобщённом смысле, мы найдём параметры зависимости.

В интервальном случае, действуя аналогичным образом, мы получим уже интервальную систему уравнений, которую также можно решать. Её решением, обычным или в некотором обобщённом смысле, будет вектор оценки параметров восстанавливаемой зависимости (2).

Информационное множество задачи получается при этом как множество решений этой интервальной системы уравнений, построенной на основе формулы (2) и данных (3).

Коридор совместных зависимостей

Определение параметров функциональной зависимости производится, как правило, для того, чтобы затем найденную формулу использовать для предсказания значений зависимости в других интересующих нас точках её области определения.

Ясно, что такое предсказание будет осуществляться с некоторой погрешностью, вызванной неопределённостями данных, неоднозначностью самой процедуры восстановления и т. п. Эту неопределённость предсказания также необходимо знать и учитывать в нашей деятельности.

Коридор совместных зависимостей и его сечение

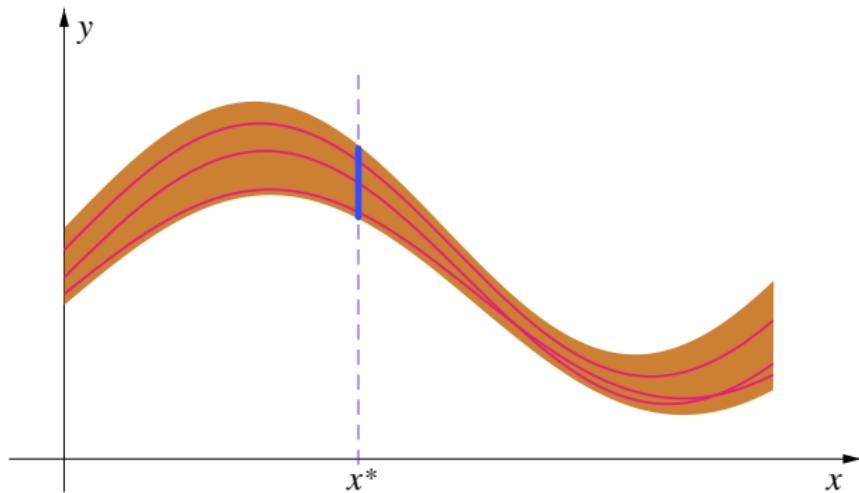


Рис.: Коридор совместных зависимостей и его сечение
для какого-то значения аргумента x^* .

Коридор совместных зависимостей

Если информационное множество задачи восстановления зависимостей непусто, то обычно оно задаёт целое семейство зависимостей, совместных с данными задачи, которое имеет смысл рассматривать вместе, как единое целое.

Это необходимо делать в вопросах, касающихся оценивания неопределённости предсказания, учёта всех возможных сценариев развития и т. п. Как следствие, возникает необходимость рассматривать вместе, единым целым, множество всех функций, совместных с интервальными данными задачи восстановления зависимости. Мы будем называть его *коридором совместных зависимостей* (см. Рис. 2).

Многозначные отображения

В литературе использовались также другие термины для обозначения этого объекта — «трубка» совместных зависимостей (имеет происхождение в теории управления), «полоса» или даже «слой неопределённости», «коридор неопределённости» и т. п.

Строгое определение коридора совместных зависимостей может быть дано на основе математического понятия многозначного отображения. Напомним, что для произвольных множеств X и Y **многозначным отображением F из X в Y** называется соответствие (правило), сопоставляющее каждой точке $x \in X$ непустое подмножество $F(x) \subset Y$, называемое **значением** или **образом x** .

Definition

Пусть в задаче восстановления зависимостей информационное множество Ω параметров зависимостей $y = f(x, \beta)$, совместных с данными, является непустым. *Коридором совместных зависимостей* рассматриваемой задачи называется многозначное отображение Υ , сопоставляющее каждому значению аргумента x множество

$$\Upsilon(x) = \bigcup_{\beta \in \Omega} f(x, \beta).$$

Сечение коридора совместных зависимостей

Значение $\Upsilon(\tilde{x})$ коридора совместных зависимостей при каком-то определённом аргументе \tilde{x} («сечение коридора») — это множество $\cup_{\beta \in \Omega} f(\tilde{x}, \beta)$, образованное всевозможными значениями, которые принимают на этом аргументе функциональные зависимости, совместные с интервальными данными измерений.

Рис. 2 изображает коридор совместных зависимостей в задаче восстановления нелинейной зависимости, но для рассматриваемого нами линейного случая коридор совместных значений имеет существенно более специальный вид .

Нетрудно показать, что границы коридора совместных зависимостей в этом случае являются кусочно-линейными.

Случай точных измерений входных переменных

Важнейшим и часто встречающимся частным случаем рассмотренной задачи является ситуация, когда независимые (экзогенные, предикторные, входные) переменные x_1, x_2, \dots, x_m измеряются точно, и вместо телесных брусов неопределённости измерений (как на Рис. 1) мы имеем отрезки прямых $(x_{i1}, x_{i2}, \dots, x_{im}, y_i)$, $i = 1, 2, \dots, n$, параллельные оси зависимой (эндогенной, критериальной, выходной) переменной (см. Рис. 3).

Именно такая постановка задачи была рассмотрена в пионерской работе Л.В. Канторовича.

Случай точных измерений входных переменных

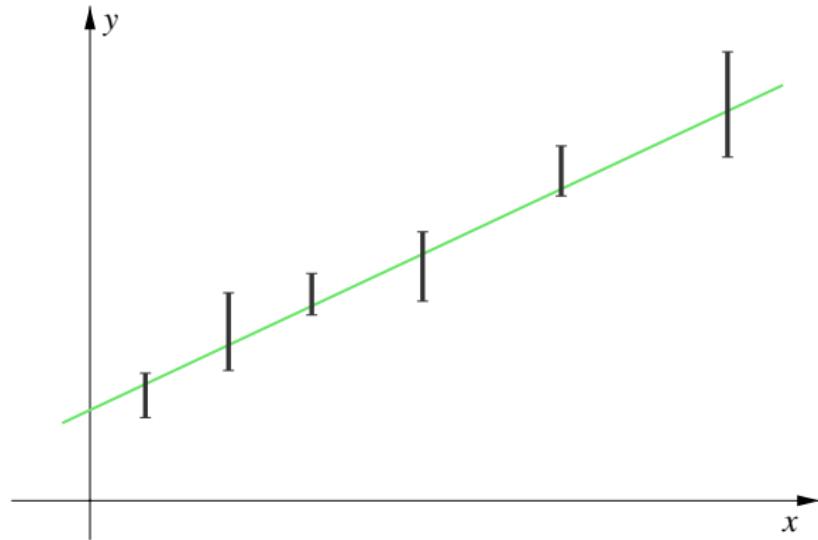


Рис.: Частный случай задачи восстановления линейной зависимости по неточным данным, когда входные переменные измеряются точно.

Постановка задачи

Отсутствие неопределённости значений независимых переменных приводит к кардинальному упрощению математической модели. Брусы неопределённости измерений зависимости, введённые ранее, схлопываясь по независимым переменным, превращаются в *отрезки неопределённости*.

Как следствие, для решения и полного исследования этого частного случая предложено большое количество эффективных вычислительных методов. Рассмотрим эти математические вопросы более детально.

Совместность зависимости с данными

Линейная зависимость (2) совместна (согласуется) с интервальными данными измерений, если её график проходит через все отрезки неопределённости, задаваемые интервалами измерений выходной переменной y , как это изображено на Рис. 3).

Подобное понимание совместности (согласования) является прямым обобщением того понимания «совместности», которое традиционно для неинтервального случая и используется, к примеру в постановке задачи интерполяции.

Совместность зависимости с данными

Подставляя в зависимость (2) данные для входных переменных x_1, x_2, \dots, x_m в i -ом измерении и требуя включения полученного значения в интервалы y_i , получим

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in y_i, \quad i = 1, 2, \dots, n. \quad (4)$$

Фактически, это интервальная система линейных алгебраических уравнений

$$\left\{ \begin{array}{l} \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1m}\beta_m = y_1, \\ \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \dots + x_{2m}\beta_m = y_2, \\ \vdots \qquad \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ \beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \dots + x_{nm}\beta_m = y_n, \end{array} \right.$$

у которой интервальность присутствует только в правой части.

Совместность зависимости с данными

С другой стороны, (4) равносильно системе

$$\left\{ \begin{array}{l} \underline{\mathbf{y}}_1 \leq \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_m x_{1m} \leq \bar{\mathbf{y}}_1, \\ \underline{\mathbf{y}}_2 \leq \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_m x_{2m} \leq \bar{\mathbf{y}}_2, \\ \vdots \quad \vdots \quad \ddots \quad \vdots \quad \vdots \\ \underline{\mathbf{y}}_n \leq \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_m x_{nm} \leq \bar{\mathbf{y}}_n. \end{array} \right. \quad (5)$$

Система двусторонних линейных неравенств

Это система двусторонних линейных неравенств относительно неизвестных параметров $\beta_0, \beta_1, \beta_2, \dots, \beta_m$, решив которую, мы можем найти искомую линейную зависимость. Множество решений системы неравенств (5) естественно считать информационным множеством параметров восстанавливаемой зависимости для рассматриваемого случая.

Для i -го двустороннего неравенства из системы (5) множество решений — это полоса в пространстве \mathbb{R}^{m+1} параметров $(\beta_0, \beta_1, \dots, \beta_m)$, ограниченная с двух сторон гиперплоскостями с уравнениями

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} = \underline{y}_i,$$

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} = \bar{y}_i.$$

Система двусторонних линейных неравенств

Множество решений системы неравенств (5) является пересечением n штук таких полос, отвечающих отдельным измерениям. Можно рассматривать эти полосы как информационные множества отдельных измерений.

На Рис. 4 изображено формирование множества решений системы неравенств (5) для случая двух параметров (т. е. $m = 1$) и $n = 3$.

Образование информационного множества параметров

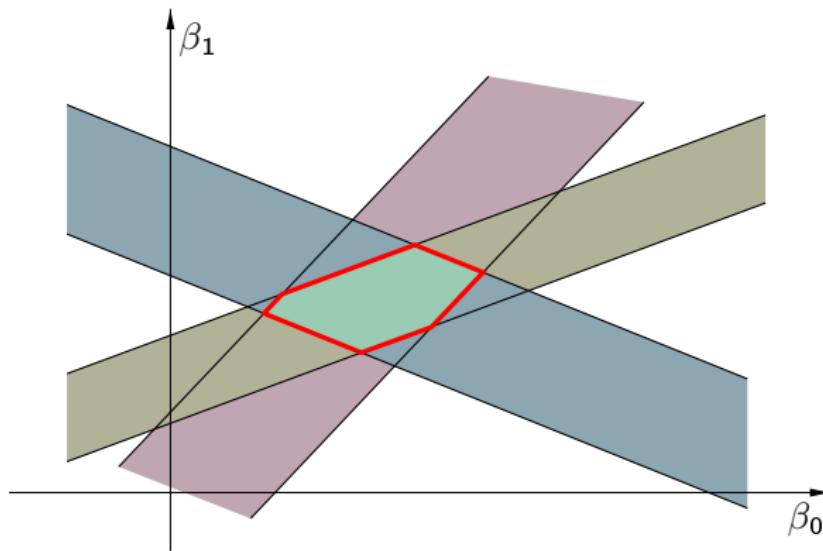


Рис.: Образование информационного множества параметров
линейной зависимости (ограничено красной линией)
для случая точных входных переменных.

Информационное множество — трудоёмкость распознавания

В целом множество решений системы линейных алгебраических неравенств (5) является *выпуклым многогранным множеством в пространстве \mathbb{R}^{m+1}* .

Распознавание того, пусто оно или непусто, а также нахождение какой-либо точки из него, являются задачами, сложность которых ограничена полиномом от их размера. Существуют эффективные и хорошо разработанные вычислительные методы для решения этих вопросов и для нахождения оценок множества решений, например, основанные на сведении рассматриваемой задачи к задаче линейного программирования.

Информационное множество — трудоёмкость распознавания

В общем случае, когда входные (экзогенные, предикторные) переменные известны неточно, ситуация существенно усложняется и множество параметров, совместных (согласующихся) с интервальными данными не может быть описано так же просто, с помощью системы линейных неравенств (5).

Трудоёмкость распознавания его пустоты или непустоты также становится экспоненциальной в зависимости от количества переменных [2].

Пример

Случай точных измерений входных переменных

Общий случай задачи восстановления зависимостей

Рассмотрим теперь случай, когда неопределённость присутствует как в измерениях значений зависимой переменной, так и в измерениях значений аргументов.

Это может быть вызвано различными причинами. Например, существенно неточное измерение входных переменных происходит в ситуациях, когда они должны устанавливаться в течение значительного времени.

Тогда их уместно выразить какими-то интервалами, а не точечными значениями.

Пример

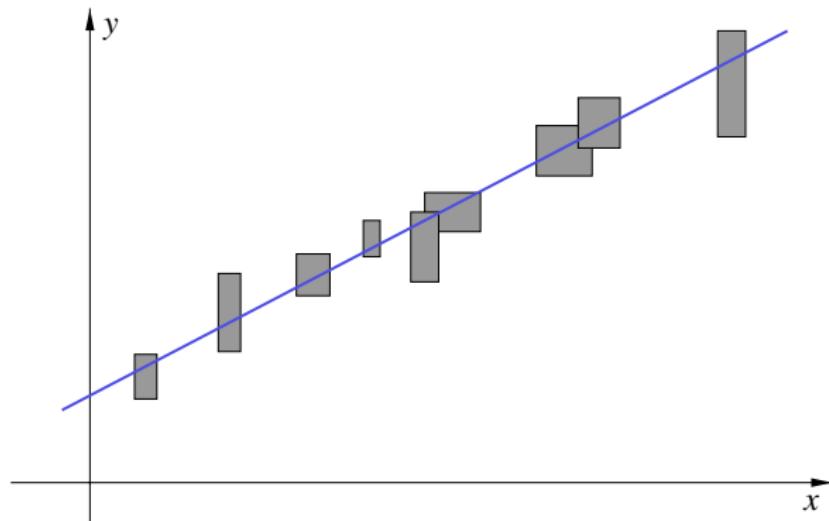


Рис.: Наглядная иллюстрация задачи восстановления линейной зависимости по данным с интервальной неопределённостью.

Пример

[https://github.com/szhilin/octave-interval-examples/blob/
master/SteamGenerator.ipynb.](https://github.com/szhilin/octave-interval-examples/blob/master/SteamGenerator.ipynb)

Общий случай задачи восстановления зависимостей

Если выборка измерений независимых переменных и зависимой переменной — накрывающая, то

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n,$$

где все x_{ij} могут принимать значения из соответствующих интервалов $x_{ij}, i = 1, 2, \dots, n, j = 1, 2, \dots, m$. Как следствие, получаем интервальную систему линейных алгебраических уравнений

$$\left\{ \begin{array}{l} \beta_0 + \mathbf{x}_{11}\beta_1 + \mathbf{x}_{12}\beta_2 + \dots + \mathbf{x}_{1m}\beta_m = \mathbf{y}_1, \\ \beta_0 + \mathbf{x}_{21}\beta_1 + \mathbf{x}_{22}\beta_2 + \dots + \mathbf{x}_{2m}\beta_m = \mathbf{y}_2, \\ \vdots \qquad \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ \beta_0 + \mathbf{x}_{n1}\beta_1 + \mathbf{x}_{n2}\beta_2 + \dots + \mathbf{x}_{nm}\beta_m = \mathbf{y}_n. \end{array} \right. \quad (6)$$

Общий случай задачи восстановления зависимостей

Это формальная запись, означающая совокупность обычных (точечных) систем линейных алгебраических уравнений того же размера и с теми же неизвестными переменными, у которых коэффициенты и правые части лежат в предписанных им интервалах (см. [2]).

Восстановление параметров линейной зависимости можно рассматривать как «решение», в том или ином смысле, выписанной интервальной системы уравнений.

Общий случай задачи восстановления зависимостей

В случае присутствия погрешностей как в измерениях аргумента, так и в измерениях зависимости множество параметров зависимостей, совместных (согласующихся) с данными, характеризуются новыми свойствами, которыми не обладают задачи с точными измерениями входных переменных.

Прежде всего, множества решений отдельных интервальных уравнений уже *не являются полосами в пространстве \mathbb{R}^n* , вроде тех, что изображены на Рис. 4. Они выглядят существенно иначе, и их конкретный вид зависит от того, какой смысл вкладывается в понятие совместности (согласования) параметров и данных, т. е. от того, *какое множество решений ИСЛАУ взято в качестве информационного множества* (см. Рис. 6).

Пример

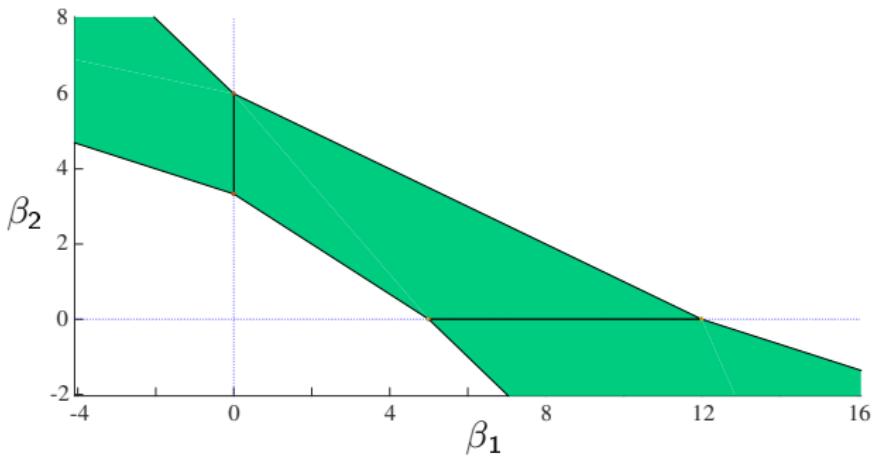


Рис.: Объединённое множество решений интервального линейного уравнения $[1, 2]\beta_1 + [2, 3]\beta_2 = [10, 12]$.

Общий случай задачи восстановления зависимостей

Само понятие согласования (совместности) параметров и данных должно быть расширено и переосмыслено.

В обычном неинтервальном случае результаты измерений — это бесконечно малые точки, и прохождение через них графика функциональной зависимости адекватно описывается двумя значениями — «да» или «нет», т. е. имеет булевский (логический) тип данных.

Общий случай задачи восстановления зависимостей

Если мы переходим от точек к брусьям неопределённости, то прохождение графика зависимости через них можно понимать по-разному.

Брусы неопределённости измерений являются прямыми декартовыми произведениями интервалов по различным осям координат, и эти оси имеют разный смысл:

интервалы $x_{i1}, x_{i2}, \dots, x_{im}$ соответствуют входным (экзогенным, предикторным) переменным,
а интервал y ; соответствует выходной (эндогенной, критериальной) переменной.

По этой причине становится важным, как именно проходит график восстанавливаемой зависимости через брусы неопределённости измерений (см. Рис. 7).

Общий случай задачи восстановления зависимостей

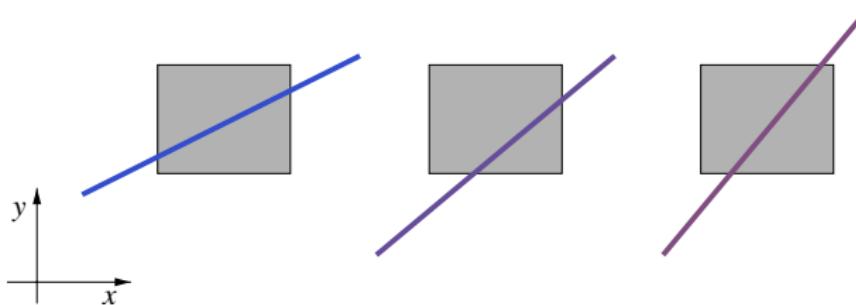


Рис.: Различные способы пересечения линии с бруском
неопределённости измерения зависимости.

Слабо совместная зависимость

Функциональную зависимость назовём *слабо совместной* с интервальными данными, если её график проходит через каждый брус неопределённости измерений хотя бы для одного значения аргумента.

Наглядно это означает, что график зависимости пересекает брусы неопределённости, но как именно — неважно (средний чертёж на Рис. 7),

достаточно лишь одной точки пересечения.

достаточно лишь одной точки пересечения.

Слабо совместная зависимость

Для случая линейной зависимости это условие наиболее удобно выразить с помощью формального языка логического исчисления предикатов:

$$(\exists x_{i1} \in \mathbf{x}_{i1}) \cdots (\exists x_{im} \in \mathbf{x}_{im}))(\exists y_i \in \mathbf{y}_i) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Равносильная упрощённая формулировка этого свойства выглядит следующим образом:

$$(\exists x_{i1} \in \mathbf{x}_{i1}) \cdots (\exists x_{im} \in \mathbf{x}_{im}) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Сильно совместная зависимость

Функциональную зависимость назовём *сильно совместной* с интервальными данными, если её график проходит через каждый брус неопределённости измерений для любого значения аргумента из интервалов неопределённости входных переменных.

Наглядно это означает, что график зависимости

целиком содержится в коридорах,

задаваемых интервалами выходной переменной при всех значениях входных переменных из соответствующих им интервалов

(левый чертёж на Рис. 7).

Сильно совместная зависимость

Для случая линейной зависимости это условие может быть формально записано в следующем виде:

$$(\forall x_{i1} \in \mathbf{x}_{i1}) \cdots (\forall x_{im} \in \mathbf{x}_{im})(\exists y_i \in \mathbf{y}_i) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Равносильная упрощённая формулировка этого свойства выглядит следующим образом:

$$(\forall x_{i1} \in \mathbf{x}_{i1}) \cdots (\forall x_{im} \in \mathbf{x}_{im}) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Сильно и слабо совместные зависимости

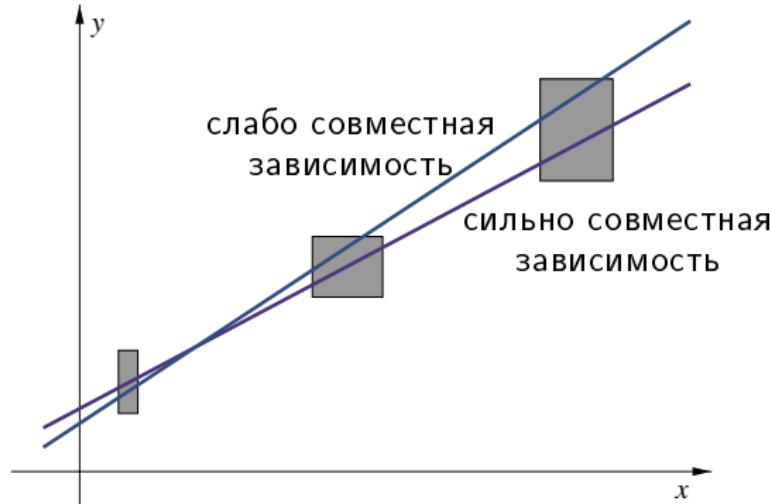


Рис.: Линейные зависимости с разными типами
согласования с данными.

Сильно совместная зависимость

В чём содержательный смысл сильной совместности?

На практике измерения на входах и выходах системы осуществляются, как правило, разными способами и даже в разное время.

Мы измеряем выход (зависимую переменную) уже тогда, когда входные значения (независимых переменных) зафиксированы, и мы их измерили. Получив при этом какие-то интервалы.

Сильная совместность функциональной зависимости с интервальными данными означает тогда, что выходная величина остаётся в пределах измеренного для неё интервала вне зависимости от того, какими конкретно в своих интервалах являются значения входных переменных.

Сильно совместная зависимость

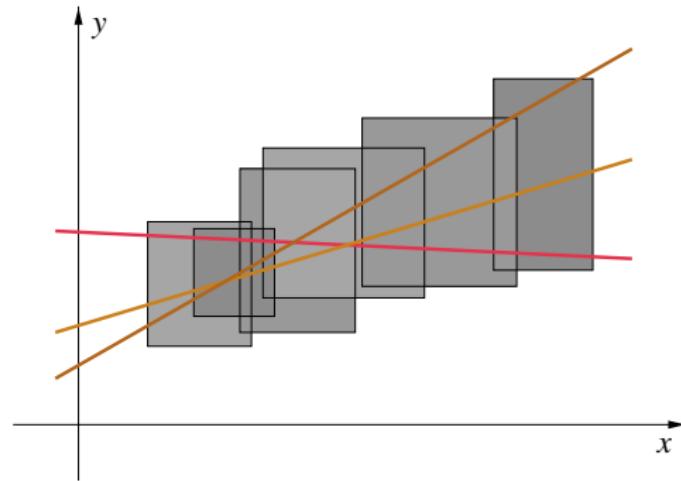


Рис.: Сложный случай восстановления зависимости
по широким перекрывающимся интервальным данным.

Множества решений

Если матрица системы (6) уравнений — точечная, т. е. коэффициенты при неизвестных β_i являются обычными вещественными числами, то объединённое множество решений в целом является выпуклым.

Но в общем случае, когда матрица интервальной системы линейных алгебраических уравнений существенно интервальна, то объединённое множество решений может быть невыпуклым.

Допусковое множество решений всегда выпукло. В целом, количество гиперплоскостей, ограничивающих множества решений, может быть очень большим.

Приближённое описание информационного множества

Возвращаясь к решению задачи восстановления зависимостей, следует отметить, что непростое строение множеств решений интервальных систем уравнений делает очень трудоёмким и малополезным их точное и полное описание.

Имеет смысл найти какое-нибудь приближённое описание информационного множества.

Здесь могут встретиться различные ситуации.

Приближённое описание информационного множества

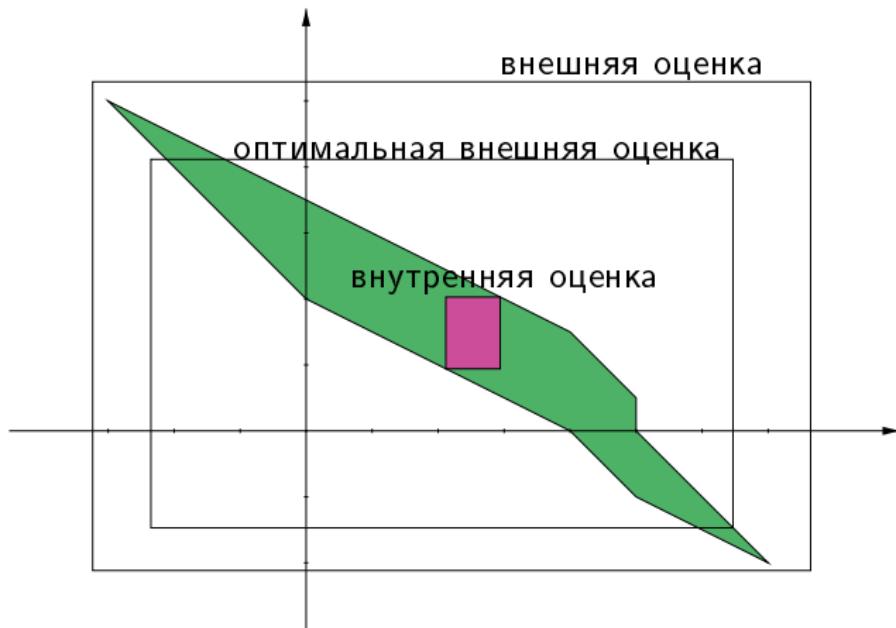


Рис.: Различные способы оценивания

информационного множества.

Оценки информационного множества

Часто бывает необходимо оценить разброс точек из информационного множества, то есть определить, насколько сильно оно «растекается» в пространстве параметров.

Часто это делается для его отдельных компонент, так что в целом нам требуется интервальный брус, содержащий множество решений. Это *внешняя оценка информационного множества*

Среди всех внешних оценок наилучшей служит минимальная по размерам внешняя оценка, которую также называют *оптимальной внешней оценкой*. Она единственна и является интервальной оболочкой информационного множества задачи.

Внешняя оценка информационного множества необходима, к примеру, при построении внешней оценки коридора совместных зависимостей, когда мы хотим просчитать гарантированный эффект от реализации всех сценариев, могущих встретиться по восстановленным зависимостям.

Оценки информационного множества

Во многих задачах требуется оценивание информационного множества с помощью какого-то несложно описываемого подмножества — *внутреннее оценивание*. Такая оценка будет содержать только точки из информационного множества и ничего лишнего.

Внешняя оценка информационного множества в этом смысле плоха тем, что включает в себя точки, не принадлежащие информационному множеству.

Если в качестве подмножества информационного множества берётся вписанный брус, то он называется *внутренней интервальной оценкой* множества решений. Среди двух внутренних оценок лучшей является та, которая целиком содержит другую, но максимальных по включению внутренних оценок, которые несравнимы друг с другом, может быть много.

Оценки информационного множества

Английские термины для обозначения внешней и внутренней оценки — outer estimate и inner estimate соответственно. Внешнюю оценку часто называют также термином «closure».

Кроме внешнего и внутреннего оценивания информационных множеств могут встретиться и другие, которые требуются по смыслу задачи.

Например, «слабое внешнее» оценивание , оценивание вдоль какого-то специального выделенного направления, исчерпывающее оценивание с помощью набора брусов и т.п.

Варианты точечной оценки информационного множества

Помимо оценивания информационного множества «целиком», во многих ситуациях достаточно найти какую-либо точку из него (здесь мы имеем аналогию с оцениванием «точечным» и «интервальным» в традиционной статистике). Естественно выбирать такую одну точку удовлетворяющей некоторым условиям оптимальности.

Варианты точечной оценки информационного множества

- центр интервального бруса, который является минимальной по включению внешней оценкой информационного множества,
- центр Оскорбина,
- чебышёвский центр,
- центр тяжести,
- точка максимума совместности (аргумент максимума распознающего функционала, который является точкой максимума совместности соответствующей интервальной системы уравнений).

Пример обработки ненакрывающей выборки

Пример обработки ненакрывающей выборки.

Набор данных.

Рассмотрим другой пример данных, полученных при измерении параметров шагового двигателя.

Изучалась зависимость положения вала от управляющего воздействия. Из одного устойчивого равновесия был проведён цикл вращений «вперёд-назад» с возвращением в начальное положение.

При этом было подано 7 одинаковых команд с шагом +64 и затем столько же с шагом -64 в единицах контроллера управления. Данные контроллера и энкодера собраны в Табл. 1.

Набор данных.

Код управления	Данные энкодера
0	30
64	30
128	26
192	24
256	17
320	11
384	7
448	0
384	6
320	7
256	11
192	14
128	20
64	25
0	29

Таблица: Выборка данных движения «вперёд-назад». Точка останова соответствует коду управления 448.

Раздельная обработка данных для каждой ветви.

Раздельная обработка данных
для каждой ветви.

Диаграмма рассеяния данных с двумя ветвями.

Диаграмма рассеяния данных имеет две ветви, выделенными синим и красным цветом. Точка останова перед возвратным движением показана черным цветом. В силу дискретности данных энкодера им приписана погрешность, равная младшему значащему разряду.

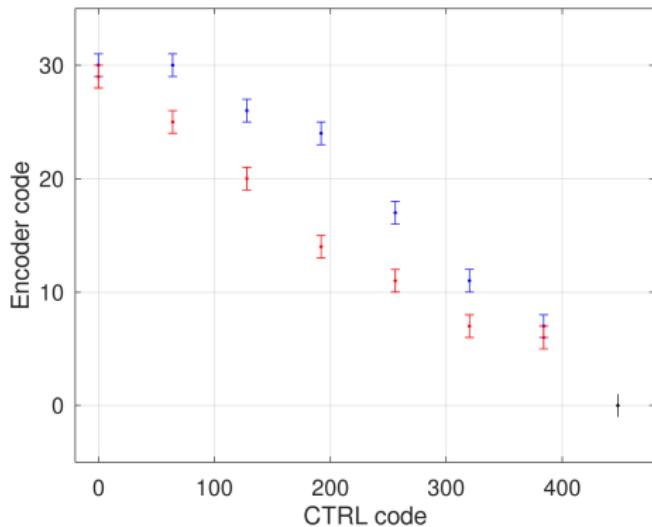


Рис.: Диаграмма рассеяния движения «вперёд-назад».

Набор данных.

Характер данных Табл. 1 и Рис. 11 совершенно типичен и является нормой для подобных измерений.

Управление происходило в так называемом режиме дробления шага. Величина кода управления ± 64 отвечает одной четверти полного шага. При меньших кодах управления траектории движения зачастую приобретают ещё более сложный вид.

Выборка из Табл. 1 несовместна. Интересно попробовать эти данные для апробирования различных математических приёмов.

Линейная регрессия на отдельные ветви зависимости.

Начнём с отдельной обработки ветвей движения. Как и в предыдущем примере сделаем данные возрастающими.

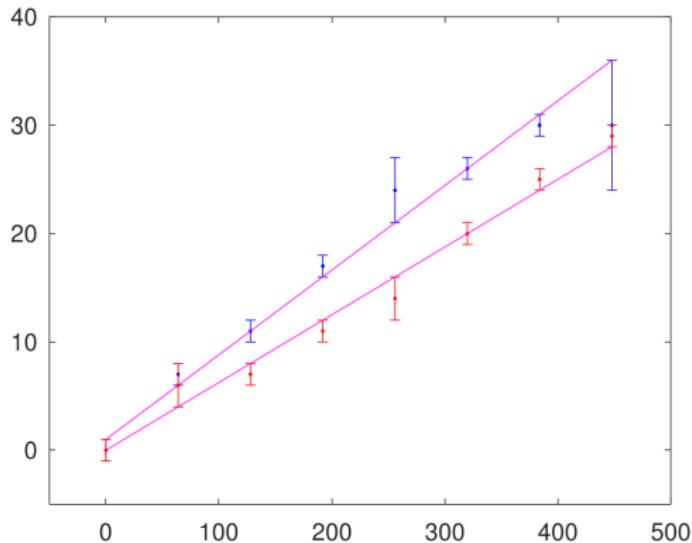


Рис.: Регрессии на разные ветви данных для движения «вперёд-назад» с оценкой по норме L_1 .

Векторы весов для отдельных ветвей зависимости.

Рис. 12 иллюстрирует несовместность данных как внутри отдельных ветвей, так и между ними. Свидетельством внутренней несогласованности служит большой разброс значений весов w_i .

$$w_{fw} = (1, 1, 1, 1, 3, 1, 1, 6)^\top, \quad (7)$$

$$w_{bk} = (1, 2, 1, 1, 2, 1, 1, 1)^\top. \quad (8)$$

Разница между ветвями проявляется в величинах коэффициентов регрессии:

$$\beta_1^{fw} = 1.00, \quad \beta_2^{fw} = 0.078, \quad (9)$$

$$\beta_1^{bk} = 0.00, \quad \beta_2^{bk} = 0.063. \quad (10)$$

Информационные множества для ветвей данных $I_{1,2}$.

Определим теперь интервальные параметры регрессии [3].

При малых оценках погрешности данных первая («синяя») ветвь несовместна даже внутренне. Непустое информационное множество I_1 возникает при $\varepsilon = 4.5$

При этом значении пересечение информационных множеств ветвей данных пусто:

$$I_1 \cap I_2 = \emptyset.$$

Интервальные оценки для разных ветвей зависимости.

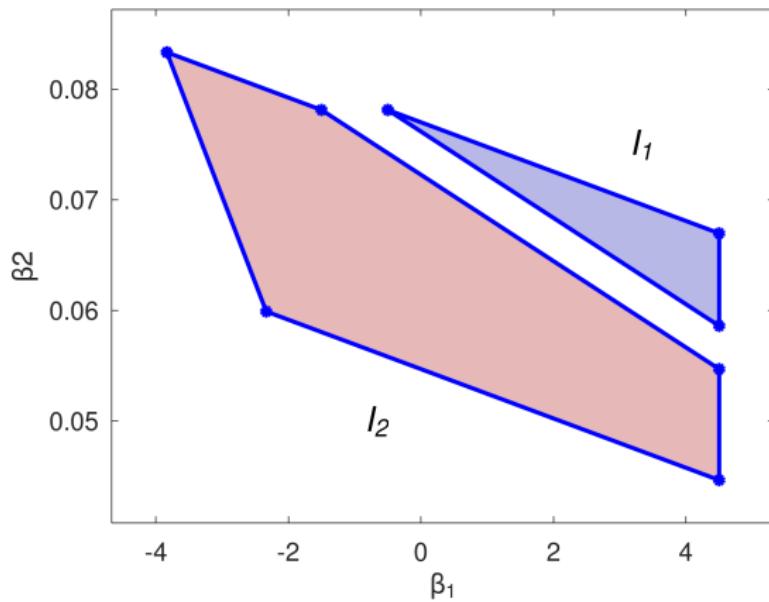


Рис.: Интервальные оценки для разных ветвей зависимости. Назначенное значение погрешности данных $\varepsilon=4.5$.

Достижение совместности.

Для достижения совместности между ветвями данными зададимся оценкой погрешности данных будем увеличивать ε пока не будет достигнуто условие

$$I = I_1 \cap I_2 \neq \emptyset.$$

Иначе, ищем

$$\arg \varepsilon = \min_{\varepsilon} \{ I_1(\varepsilon) \cap I_2(\varepsilon) \neq \emptyset \}. \quad (11)$$

Информационные множества.

На Рис. 14 приведены информационные множества сдвигов и наклонов регрессионных прямых для обеих ветвей данных. Они ограничены многоугольниками и даны заливкой того же цвета, что и данные на Рис. 12.

Их пересечение — сторона многоугольника, отрезок с вершинами

$$I(\beta_1, \beta_2) = I_1 \cap I_2 = (-1.00, 0.078) - (5.00, 0.055), \quad (12)$$

показан красным цветом.

Красным прямоугольником дана внешняя оценка параметров регрессионных прямых для обеих ветвей данных.

Информационные множества. Интервальные оценки.

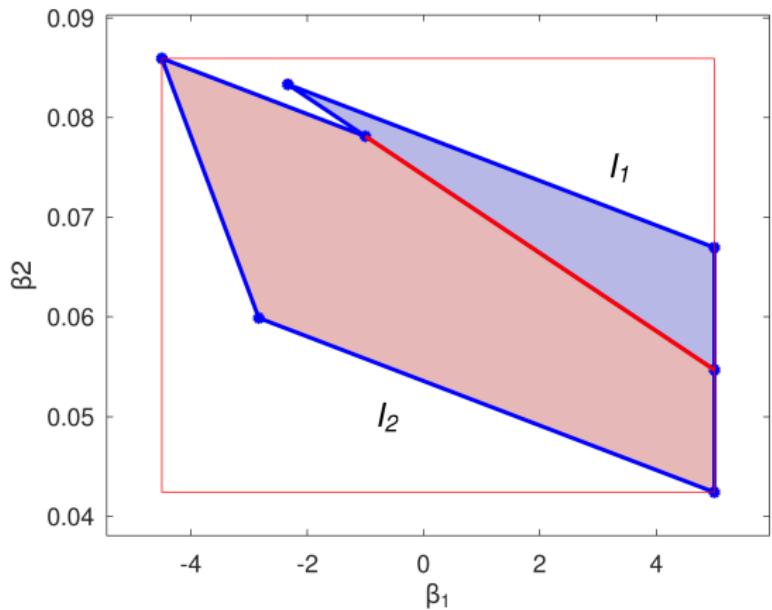
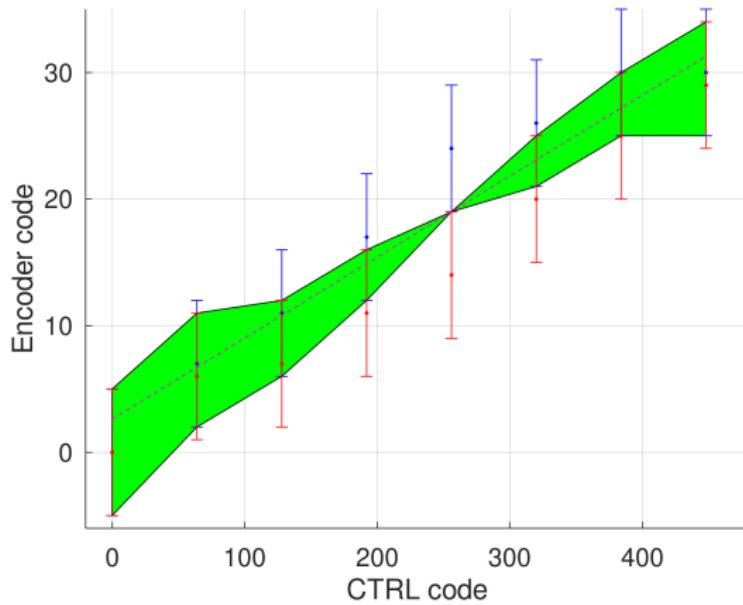


Рис.: Интервальные оценки для разных ветвей зависимости и множество $I(\beta_1, \beta_2)$. Назначенное значение погрешности данных (11) $\varepsilon=5$.

Коридор совместности Υ .

На Рис. 15 приведены диаграмма рассеяния данных и коридор совместности параметров модели регрессии Υ для погрешности данных согласно (11).



Сечение коридора совместности.

Также дана прямая регрессии по параметрам, соответствующим середине информационного множества

$$\text{mid } I(\beta_1, \beta_2) = [2.616, 0.064].$$

При значении $x^* = 256$, сечение коридора совместности $\Upsilon(x^*)$ состоит из одной точки.

Линейные регрессии.

Построим линейные регрессии с параметрами из крайних точек отрезка (12) и его середины.

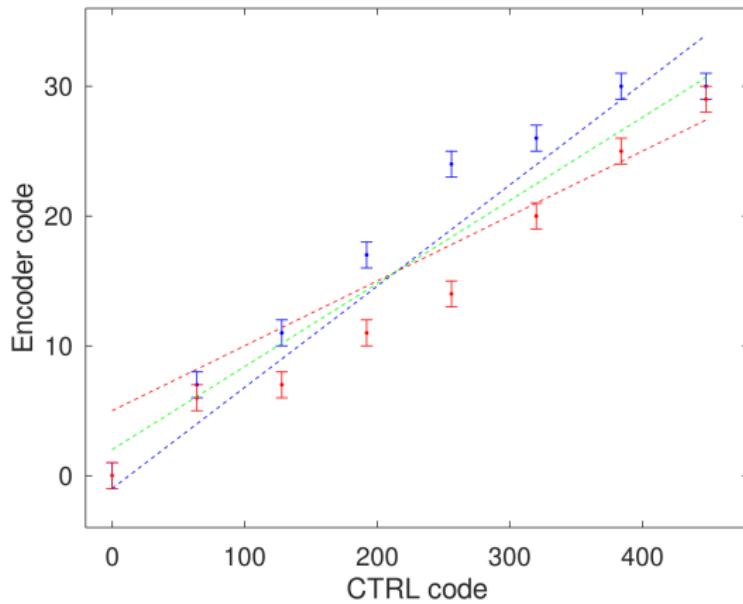


Рис.: Набор линейных регрессий.

Линейные регрессии.

Из Рис. 16 ясно, что прямые, определяемые множеством (12), заполняют два открытых угла и дают внутреннюю оценку коридора совместности.

Брусы совместности данных.

Посмотрим на вопрос с другой точки зрения. Пусть погрешность измерений находится не в выходных данных, которые весьма точны, а во входных.

Будем считать, что данные

$$\mathbf{y}_i = \mathbf{y}_i^1 \cup \mathbf{y}_i^2,$$

где 1, 2 — разные ветви данных. В общем случае, \mathbf{y}_i — неодносвязный интервал.

Для работы с обычными интервалами \mathbb{IR} , возьмём внешнюю оценку выходных данных

$$\mathbf{y}_i = \left[\min\{\underline{\mathbf{y}}_i^1, \underline{\mathbf{y}}_i^2\}, \max\{\bar{\mathbf{y}}_i^1, \bar{\mathbf{y}}_i^2\} \right].$$

Брусы совместности данных.

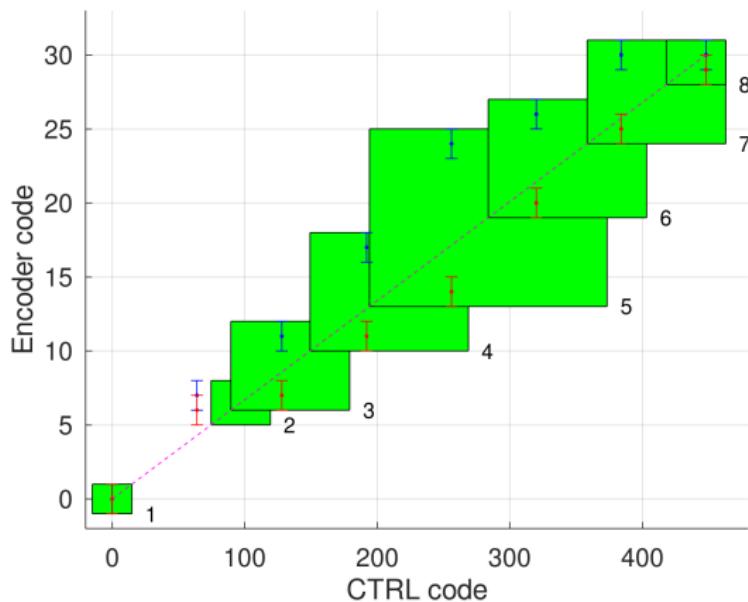


Рис.: Брусы совместности данных. Номер замера дан правее его правого нижнего угла бруса.

Брусы совместности данных.

Считая, что модель линейна, отнесем неопределённость на входные данные \mathbf{x}_i . В таком случае, модель неопределённости данных будет выглядеть как брусы $(\mathbf{x}_i, \mathbf{y}_i)$.

Внешнюю оценку входных данных примем как

$$\mathbf{x}_i = [\min\{\underline{\mathbf{x}}_i^1, \underline{\mathbf{x}}_i^2\}, \max\{\bar{\mathbf{x}}_i^1, \bar{\mathbf{x}}_i^2\}] .$$

При этом имеем в виду, что

$$\mathbf{y}_i = \beta_1 + \beta_2 \cdot \mathbf{x}_i, \quad i = 1, 2, \dots, m.$$

Рис. 17 даёт пример модели для данных Табл. 1. Регрессионная прямая проведена через «центры» первой и последней пар точек выборки. В такой постановке необходимо найти параметры линейной регрессии β_1 , β_2 и радиусы $\text{rad } \mathbf{x}_i$.

Брусы совместности данных.

В более подробном виде данные представлены на Рис. 18.

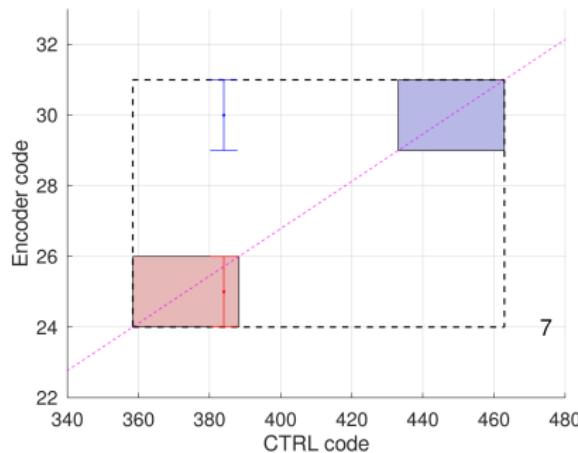
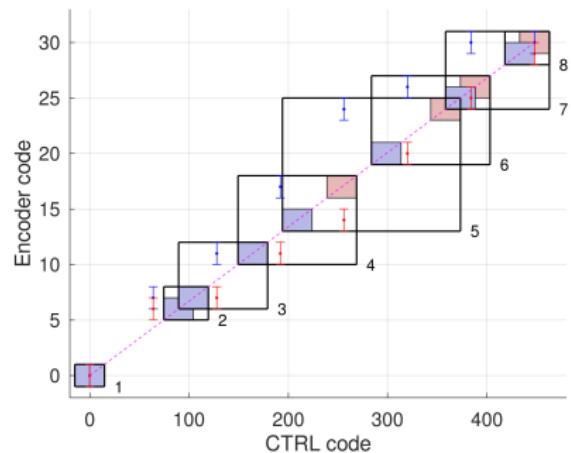


Рис.: Брусы совместности данных по отдельности для каждого замера и для пары «вперёд-назад» совместно. Номер замера дан правее его правого нижнего угла бруса.. Справа — один замер.

Брусы совместности данных.

Исходные данные для измерения 7 по данным Табл. 1

$$x_7 = 384, \quad y_7 = [24, 26] \cup [29, 31].$$

Брус совместности на Рис. 18

$$x_7 = [358, 462], \quad y_7 = [24, 31].$$

Совместимость за счет коррекции входных данных.

Пусть выходные данные y считаются абсолютно надёжными. В таком случае вся неопределённость содержится во входных данных.

Будем считать теперь данные Табл. 1 индивидуальными, не зависящими от ветви замеров, на которой они были получены.

Сделаем точечные значения x_i интервальными

$$x_i \rightarrow x_i, \quad i = 1, 2, \dots, 15.$$

так чтобы регрессионная прямая прошла через все брусы (x_i, y_i) .

Совместимость за счет коррекции входных данных.

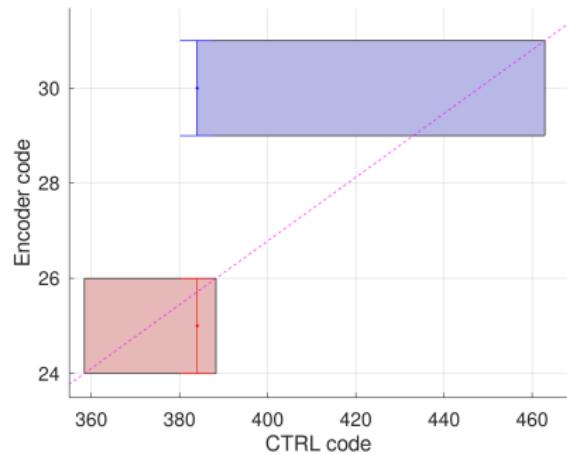
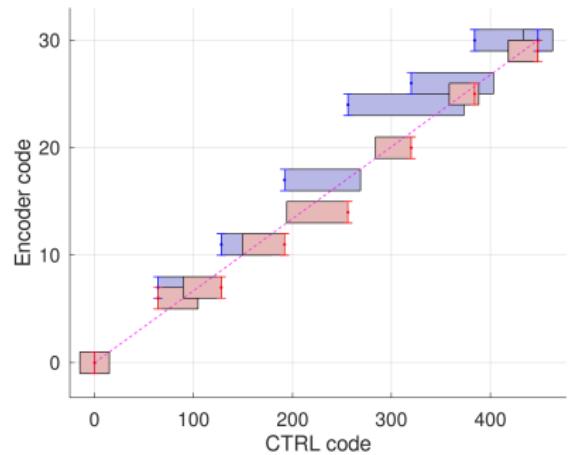


Рис.: Совместимость за счет входных данных. Справа — один замер.

Совместимость за счет коррекции входных данных.

Рис. 19 даёт представление о том, как выглядят совместные данные при таком подходе.

При постановке задачи линейного программирования

$$\sum_i \text{rad } x_i \rightarrow \min,$$

можно достигать получения совместной (в идеале, накрывающей) выборки при минимальном «расширении» входных данных.

Совместимость за счет коррекции входных данных.

В зависимости от конкретного характера данных, можно ставить и более общие постановки задач оптимизации, такие как

$$a \cdot \sum_i \text{rad } \mathbf{x}_i + b \cdot \sum_i \text{rad } \mathbf{y}_i \rightarrow \min, \quad (13)$$

где a, b — параметры, характеризующие предпочтения (веса) входным и выходным данным.

Сходный анализ данных можно найти в работах различных исследователей, начиная с диссертации Р.Мура 1962 г., и в самых современных публикациях С.И. Кумков конференция Scan2020, 2021.

Совместная обработка всех данных.

Совместная обработка всех данных.

Диаграмма рассеяния данных.

Вернёмся к исходным данным.

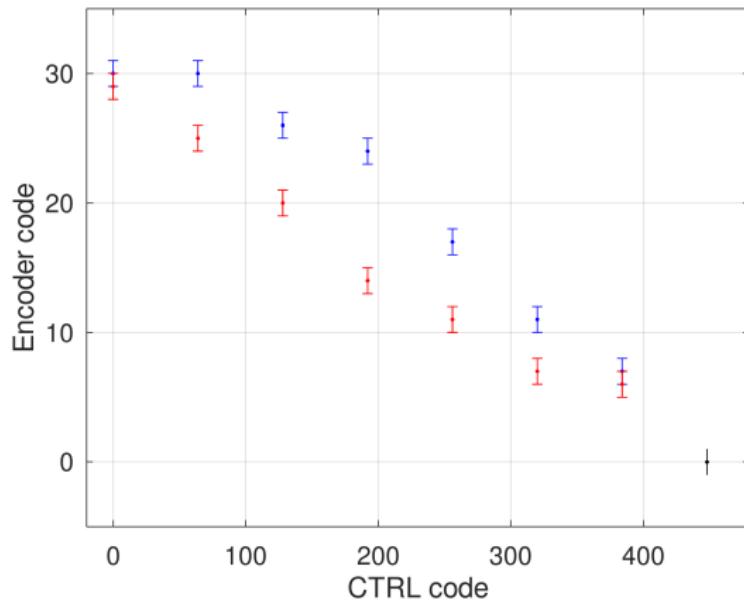


Рис.: Диаграмма рассеяния движения «вперёд-назад».

Совместная обработка всех данных в \mathbb{KR} .

Рассмотрим данные энкодера для двух ветвей зависимости. Как работать с данными, имеющими одинаковое значение независимой переменной?

Ненулевое пересечение имеют немногие из данных двух ветвей зависимости.

Поэтому рассматривать ситуацию следует в полной интервальной арифметике \mathbb{KR} и пользоваться конструкциями для объектов этой арифметики.

Вектор минимумов по включению.

Составим вектор *минимумов по включению* для 2-х ветвей, который планируем использовать как набор данных для проведения вычислений для построения интервальнойной регрессии.

$$\mathbf{y}_k = \mathbf{y}_k^{fw} \wedge \mathbf{y}_k^{bk} = \left[\max\{\underline{\mathbf{y}}_k^{fw}, \underline{\mathbf{y}}_k^{bk}\}, \min\{\bar{\mathbf{y}}_k^{fw}, \bar{\mathbf{y}}_k^{bk}\} \right]. \quad (14)$$

Выборка данных в КР.

k	y_k
.	единицы энкодера
1	[-1, 1]
2	[6, 7]
3	[10, 8]
4	[16, 12]
5	[23, 15]
6	[25, 21]
7	[29, 26]
8	[29, 30]

Таблица: Вектор минимумов по включению (14) для 2-х ветвей данных
Табл. 1.

Большая часть компонент y_k в Табл.2 — неправильные интервалы.

Задача нахождения максимума совместности.

Теперь можно поставить задачу нахождения *максимума совместности* для оценивания информационного множества.

$$X \cdot \beta \subseteq y. \quad (15)$$

Знак принадлежности в (15) вместо равенства использован ввиду того, что мы не можем требовать точного удовлетворения всех условий, наложенных данными, но ограничиваемся более слабым удовлетворением принадлежности.

Оценивание множеств решений переопределённых ИСЛАУ.

В книге [1] раздел «Численные методы для интервальных линейных систем» предлагается следующий практический рецепт решения задач внутреннего и внешнего оценивания множеств решений переопределённых интервальных систем уравнений.

Разобъём исходную систему уравнений на подсистемы

$$\mathbf{X}^{(1)}\boldsymbol{\beta} = \mathbf{y}^{(1)}, \quad \dots, \quad \mathbf{X}^{(k)}\boldsymbol{\beta} = \mathbf{b}^{(k)},$$

которые можно рассматривать и решать отдельно друг от друга.

Метод квадратных подсистем.

Решим задачи внутреннего или внешнего оценивания для полученных подсистем с помощью численных методов, предназначенных для квадратных интервальных линейных систем уравнений.

Затем пересечём полученные интервальные оценки, и полученный брус будет внутренней или внешней оценкой множества решений исходной системы.

Метод квадратных подсистем.

Пусть решениями подсистем будут множества

$$\Xi^{(1)}, \dots, \Xi^{(k)}.$$

Составим пересечение этих множеств

$$\Xi = \bigcap_i \Xi^{(i)},$$

которое будет оценкой решения системы включений (15).

Рассмотренный метод предложено в [1] называть *методом квадратных подсистем*.

Результаты очень сильно зависят от способа выбора квадратных матриц. В частности, в случае одинаковых строк в точечной матрице $X^{(i)}$, соответствующее множество $\Xi^{(i)}$ будет неограниченным. В случае соседних строк оценка также может быть весьма грубой.

Диаграмма рассеяния данных и регрессия методом квадратных подсистем.

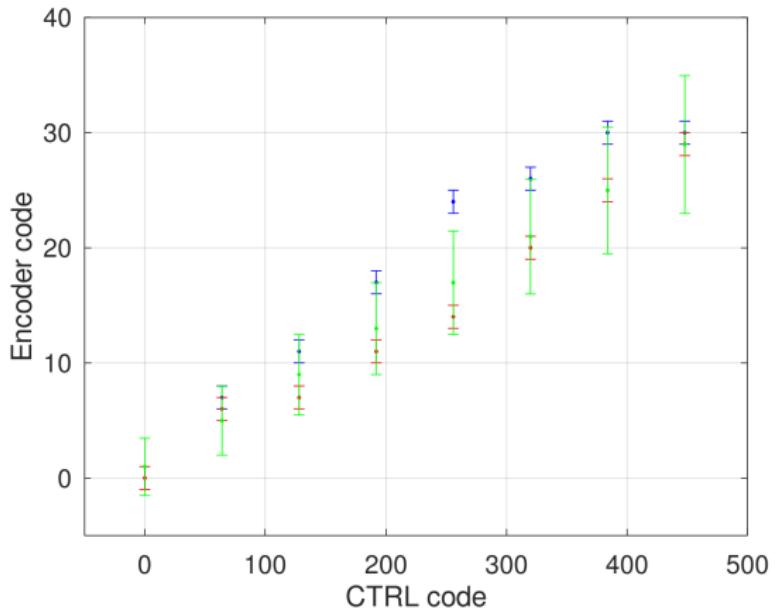


Рис.: Диаграмма рассеяния данных и регрессия методом квадратных подсистем.

Метод квадратных подсистем.

На Рис. 21 приведены оценки коридора совместности для решения системы включений (15) с перебором строк матрицы X размером 8×2 .

Для расчета были взяты 4 матрицы 2×2 .

Решение проводилось *субдифференциальным методом Ньютона* с помощью библиотеки `kinterval` [4].

Метод квадратных подсистем.

Пересечением значений β_1^i, β_2^i , $i = 1, 2, \dots, 4$ в частных решениях получены значения параметров регрессии

$$\beta_1 = \bigcap_i \beta_1^i = [-1.5159, 3.4648],$$

$$\beta_2 = \bigcap_i \beta_2^i = [0.054688, 0.070312].$$

На Рис. 21 оценки выходных данных даны зеленым цветом.

Метод квадратных подсистем.

В целом результат выглядит приемлемым, при этом для некоторых замеров интервальных границы оценок выходят за исходную диаграмму рассеяния, а для одного значения ($x_5 = 256$) не полностью покрывают «зазор» в неправильном интервале $y_5 = [23, 15]$.

Вспомним коридор совместности Υ .

Коридор совместности Υ .

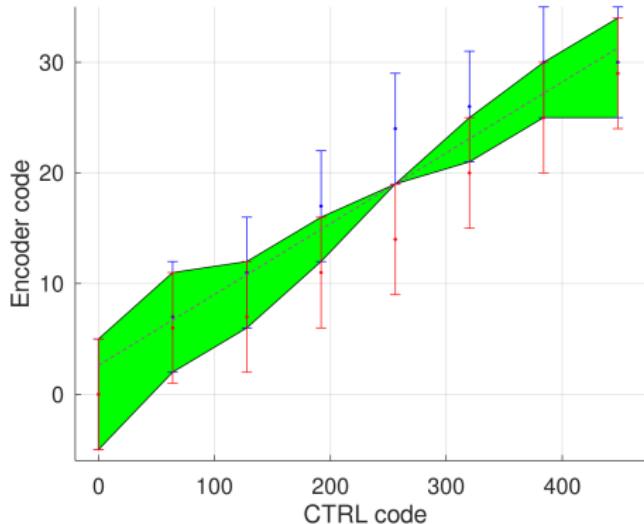


Рис.: Коридор совместности Υ , погрешность данных (11) $\varepsilon = 4.5$.

При значении $x^* = 256$, сечение коридора совместности $\Upsilon(x^*)$ состоит из одной точки.

Заключение.

Представленные вычисления дают различные оценки параметров регрессии ненакрывающей выборки. Вместе с тем очевидно, что исследование нельзя назвать исчерпывающим.

Этот факт отражает современное состояние теории оценок ненакрывающих выборок.

Заключение.

Приведём цитату из книги [1]:

«... некоторые из задач, возникших в анализе интервальных данных, на настоящий момент проработаны относительно слабо. Это относится, прежде всего, к решению интервальных линейных систем с общими прямоугольными матрицами, у которых число уравнений может не совпадать с числом неизвестных.

Кроме того, подавляющее большинство численных методов для интервальных систем уравнений, линейных и общих нелинейных, разработаны для задачи *внешнего интервального оценивания объединённого множества решений*, тогда как другие способы оценивания и другие множества решений получили гораздо меньшее внимание. »

Литература

-  А.Н. Баженов, С.И. Жилин, С.И. Кумков, С.П. Шарый.
Обработка и анализ данных с интервальной неопределенностью.
РХД. Серия «Интервальный анализ и его приложения». Ижевск.
2021. с.200.
-  С.П. Шарый. Конечномерный интервальный анализ. —
Новосибирск: XYZ, 2021. — Электронная книга, доступная на
<http://interval.ict.nsc.ru/Library/InteBooks/SharyBook.pdf>
-  С.И.Жилин. Примеры анализа интервальных данных в Octave
<https://github.com/szhilin/octave-interval-examples>
-  С.И.Жилин. Библиотека полной интервальной арифметики
kinterval в среде Octave. Частное сообщение.

Тема X2. Обработка и анализ данных с интервальной неопределенностью.

А.Н. Баженов

Санкт-Петербургский политехнический университет Петра Великого

a_bazhenov@inbox.ru

12.10.2021

Обработка и анализ данных с интервальной неопределённостью.

ПЛАН

ПЛАН

- Общие понятия
- Обработка константы
- Задача восстановления зависимостей
- Обработка выбросов

Теория:

А.Н. Баженов, С.И. Жилин, С.И. Кумков, С.П. Шарый.
Обработка и анализ данных с интервальной неопределённостью. РХД.
Серия «Интервальный анализ и его приложения». Ижевск. 2021. с.200.

ПЛАН

Обработка выбросов.

Обработка выбросов

Даются определения новых терминов и понятий, которые возникают в связи с восстановлением функциональных зависимостей по данным их измерений и наблюдений, имеющих интервальную неопределённость.

Мы рассмотрим основные идеи и типичные приёмы восстановления зависимостей по интервальным данным, а также возникающие при этом проблемы.

Подробно исследуется случай простейшей линейной зависимости, но большинство построений и рассуждений легко переносятся на общий нелинейный случай.

Постановка задачи

Предположим, что величина y является функцией некоторого заданного вида от независимых аргументов x_1, x_2, \dots, x_m , т. е.

$$y = f(x, \beta), \quad (1)$$

где $x = (x_1, \dots, x_m)$ — вектор независимых переменных,
 $\beta = (\beta_1, \dots, \beta_l)$ — вектор параметров функции. Имея набор значений переменных x и y , нам нужно найти β_1, \dots, β_l , которые соответствуют конкретной функции f из параметрического семейства (1).

Мы будем называть эту задачу *задачей восстановления зависимости*.

Постановка задачи

Важнейший частный случай поставленной задачи — определение параметров линейной функциональной зависимости вида

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m, \quad (2)$$

в которой x_1, x_2, \dots, x_m — независимые переменные (которые называются также *экзогенными*, *предикторными* или просто *входными* переменными), y — это зависимая переменная (которая называется также *эндогенной*, *критериальной* или *выходной* переменной), а $\beta_0, \beta_1, \dots, \beta_m$ — некоторые коэффициенты.

Эти неизвестные коэффициенты должны быть определены из ряда измерений значений x_1, x_2, \dots, x_m и y .

Постановка задачи

Результаты измерений неточны, и мы предполагаем что они имеют ограниченную неопределенность, когда нам известны лишь некоторые интервалы, дающие двусторонние границы измеренных значений.

Таким образом, результатом i -го измерения являются такие интервалы $x_1^{(i)}, x_2^{(i)}, \dots, x_m^{(i)}, y^{(i)}$, относительно которых мы предполагаем, что истинное значение x_1 лежит в пределах $x_1^{(i)}$, истинное значение x_2 лежит в $x_2^{(i)}$ и т.д. вплоть до y , истинное значение которого находится в интервале $y^{(i)}$.

В целом имеется n измерений, так что индекс i может принимать значения из множества натуральных чисел $\{1, 2, \dots, n\}$.

Постановка задачи

Далее для удобства построений и выкладок обозначим номер измерения i не верхним, а нижним индексом, который мы поставим первым при обозначении входов. Таким образом, полный набор данных будет иметь вид

$$\begin{aligned} & x_{11}, \quad x_{12}, \quad \dots \quad x_{1m}, \quad y_1, \\ & x_{21}, \quad x_{22}, \quad \dots \quad x_{2m}, \quad y_2, \\ & \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ & x_{n1}, \quad x_{n2}, \quad \dots \quad x_{nm}, \quad y_n. \end{aligned} \tag{3}$$

Нам необходимо найти или как-то оценить коэффициенты β_j , $j = 0, 1, \dots, m$, для которых линейная функция (2) «наилучшим образом» приближала бы интервальные данные измерений (3).

Постановка задачи

Для обозначения $n \times m$ -матрицы, составленной из данных (3) для независимых переменных часто используют термины **матрица плана эксперимента** или просто **матрица плана**, которые возникли в теории планирования эксперимента .

Интервалы $x_{i1}, x_{i2}, \dots, x_{im}, y_i$ мы называем, как и раньше, **интервалами неопределённости i -го измерения**.

Но кроме них нам также потребуется обращаться ко всему множеству, ограничеваемому в многомерном пространстве \mathbb{R}^{m+1} этими интервалами по отдельным координатным осям.

Общие идеи

Понятие «выброс» в статистике и анализе данных, как правило, определяется нечётко и неформально. Объясняется это тем, что основания для признания измерения выбросом лежат за пределами формальной математической постановки задачи анализа данных и требуют привлечения внешних по отношению к ней знаний из предметной области и истории происхождения данных, специфичных в каждом конкретном случае.

Тем не менее, главный объединяющий смысл различных определений — указание на нарушение измерением-выбросом некоторой однородности (согласованности, непротиворечивости), ожидаемой для большинства наблюдений выборки по отношению к заданной математической модели.

Общие идеи

Подчеркнём эту особую роль модели и неабсолютный характер понятия «выброс», вкупе означающие, что статус измерения в одной и той же выборке может меняться в зависимости от вида модели, рассматриваемой на конкретном этапе анализа данных.

Поэтому, строго говоря, утверждения вида «измерение x_i является выбросом в выборке X » всякий раз должны сопровождаться оговоркой — «относительно такой-то модели», если это явно не следует из контекста.

Общие идеи

Интервальный подход даёт естественный формальный индикатор согласованности данных, модели и априорной информации — непустоту информационного множества, соответствующего задаче. Пустота информационного множества свидетельствует о наличии тех или иных противоречий между данными и моделью.

Поиск причин появления противоречий, а также выбор путей их преодоления — процесс творческий и неформальный, большей частью опирающийся на прикладные соображения и экспертные знания о моделируемом явлении или процессе и условиях получения данных.

Общие идеи

Формальные приёмы и математические методы, задействованные в этом процессе, выполняют важную, но подчиненную роль. Они используются для получения информации о данных и модели, позволяющей выдвигать гипотезы о причинах противоречий, вырабатывать способы коррекции данных или модели и оценивать обеспечиваемые ею результаты.

Иными словами, математические методы отвечают на вопрос «как устроены данные?», в то время как ответы на вопросы «почему так устроены данные?» и «что делать?» может дать только содержательный анализ моделируемого явления.

Общие идеи

Причинами возникновения противоречий в задаче анализа данных могут служить как

некорректность измерений (вследствие нарушений условий их проведения, регистрации, сбоев при передаче, некорректной оценки уровня неопределённости, нештатного поведения моделируемой системы и т.п.),

так и некорректность модели (вид модели не соответствует моделируемому явлению и т.п.).

Общие идеи

При использовании формальных методов выявления выбросов следует иметь в виду, что выбросы могут оказаться наиболее существенной частью выборки, проливающей свет на то, как собирались данные или каково истинное поведение изучаемой системы или процесса, не укладывающееся в исходные предположения.

Учитывая, что предметом интервального анализа часто становятся малые выборки, обычная тактика удаления «подозрительных» измерений должна использоваться с особой осторожностью.

Обозначения.

$s_i = (x_i, y_i)$ — наблюдение, состоящее из значения входной переменной $x \in \mathbb{R}^m$ и интервального измерения y_i выходной переменной $y \in \mathbb{R}$.

$S_n = \{s_i\}_{i=1,\dots,n} = \{(x_i, y_i)\}_{i=1,\dots,n}$ — выборка из n наблюдений.

$y(x) = f(x, \beta)$ — модель с параметрами $\beta \in \mathbb{R}^{m+1}$, например, линейная $y(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$.

Обозначения.

$\Upsilon(x)$ — коридор совместных зависимостей.

$\Upsilon(x; S_n)$ — коридор совместных зависимостей, построенных по выборке S_n .

$\Omega_i = \Omega(s_i) = \{\beta \mid f(x_i, \beta) \subset y_i\}$ — информационное множество наблюдения $s_i = (x_i, y_i)$.

$\Omega = \Omega(S_n) = \cap_{i=1}^n \Omega_i$ — информационное множество задачи построения модели $y(x) = f(x, \beta)$ по выборке S_n .

Статус измерений.

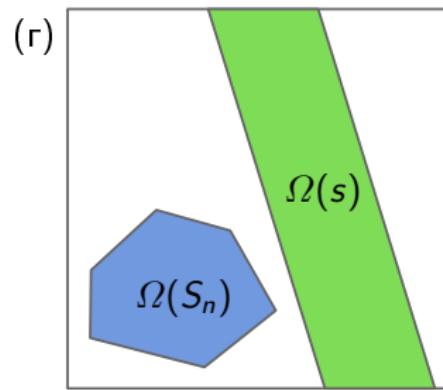
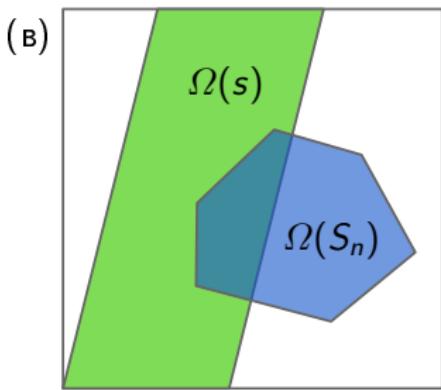
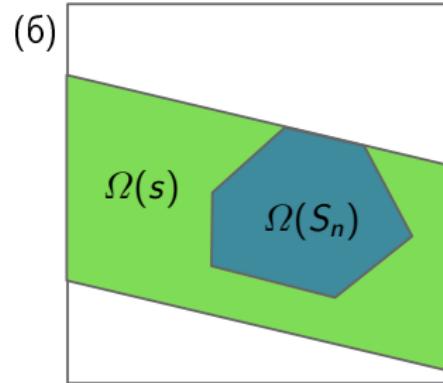
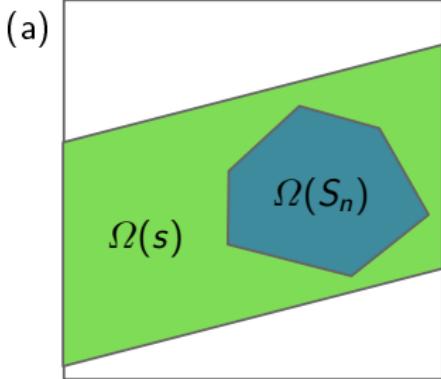
О влиянии некоторого интервального измерения $s = (x, y)$ на модель, построенную по выборке S_n , можно судить на основе того, в каком взаимоотношении находятся информационные множества $\Omega(s)$ и $\Omega(S_n)$.

Такая характеристика полезна как для «новых» измерений ($s \notin S_n$), так и для измерений, уже входящих в выборку ($s \in S_n$).

Измерения, добавление которых к выборке не приводит к модификации модели ($\Omega(S_n) = \Omega(S_n \cup s)$), именуются *внутренними*, изменяющие же модель ($\Omega(S_n) \supset \Omega(S_n \cup s)$) — *внешними*.

В каждом из этих классов измерений дополнительно выделяют специальные подклассы — *границные измерения* и *выбросы* соответственно (Рис. 20).

Статус измерений.



Статус измерений.

Информационные множества, построенные по выборке S_n и наблюдению s с различными статусами:

- (а) — внутреннее
- (б) — граничное
- (в) — внешнее
- (г) — выброс

Границные измерения.

Границными называют измерения, определяющие какой-либо фрагмент границы информационного множества. Очевидно, это свойство имеет смысл рассматривать для наблюдений, принадлежащих выборке S_n , по которой сконструирована модель и её информационное множество $\Omega(S_n)$.

Подмножество всех границных наблюдений в S_n играет особую роль, поскольку оно является

минимальной подвыборкой, полностью определяющей модель.

Удаление неграницных наблюдений из выборки не изменяет модель.

Коридор совместности — Лекция 2.

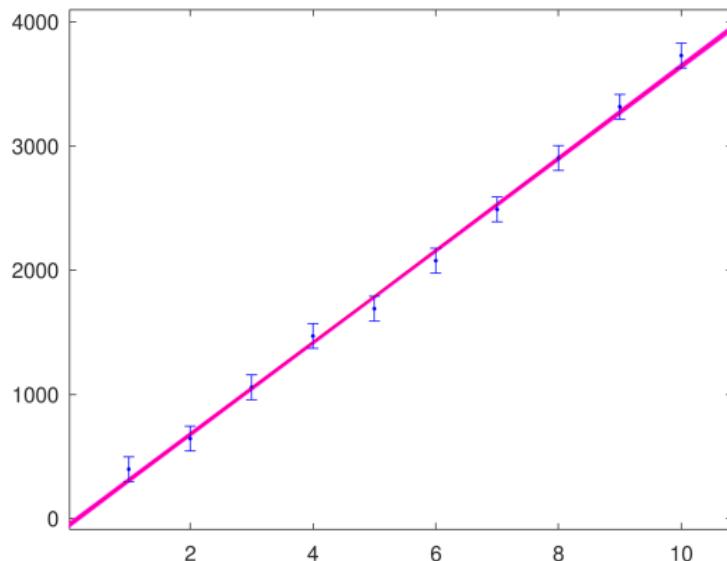


Рис.: Диаграмма рассеяния и коридор совместности \mathcal{Y} , погрешность данных $\varepsilon = 100$.

Коридор совместности — Лекция 2.

Коридор совместности Υ представляет собой узкую полосу, проходящую через крайние значения нескольких брусов.

Именно, коридор совместности касается вершин брусов

- \underline{y}_1 ,
- \bar{y}_5, \bar{y}_6 ,
- \underline{y}_{10} .

Как уже было замечено ранее, в середине графика имеется «излом».

Дальнейшее уменьшение ε приводит к пустоте множества параметров. Выборка становится *ненакрывающей*.

Таким образом,

$$\Omega_B = \{y_1, y_5, y_6, y_{10}\} \quad (4)$$

— подмножество всех граничных наблюдений, минимальная подвыборка, полностью определяющая модель.

Выбросы.

Среди внешних измерений особым образом выделяют *выбросы*.

Построение модели по выборке, пополненной таким наблюдением, приводит не просто к уменьшению информационного множества, а к его пустоте

$$\Omega(S_n \cup s) = \emptyset,$$

то есть к «разрушению» модели.

Статус измерений.

Существует экономичный способ определения статуса измерения, не требующий явного перестроения модели для выборки, расширенной анализируемым измерением.

Анализ взаимоотношений информационных множеств $\Omega(S_n)$ и $\Omega(S_n \cup s)$ или $\Omega(S_n)$ и $\Omega(s)$ можно заменить выяснением отношений интервала неопределённости y анализируемого измерения $s = (x, y)$ и интервального прогнозного значения рассматриваемой модели в той же точке $T(x; S_n)$.

На Рис. 2 анализируемые измерения показаны чёрными линиями, а соответствующие им интервалы прогнозов — широкими цветными линиями (в данном случае их ширина не имеет содержательного смысла, а лишь упрощает восприятие наложенных друг на друга интервалов).

Интервальные наблюдения с различными статусами.

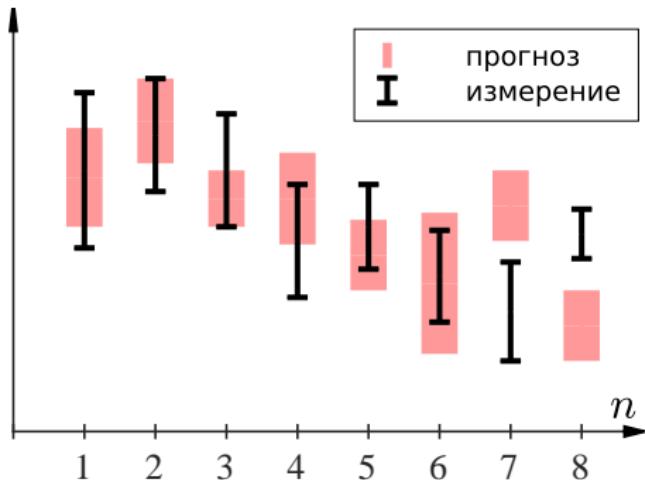


Рис.: Интервальные наблюдения с различными статусами: *внутреннее* ($n = 1, \dots, 3$), *границные* ($n = 2, 3$), *внешние* ($n = 4, \dots, 8$), *строго внешнее* ($n = 6$), *выбросы* ($n = 7, 8$).

Диаграмма статусов для интервальных наблюдений.

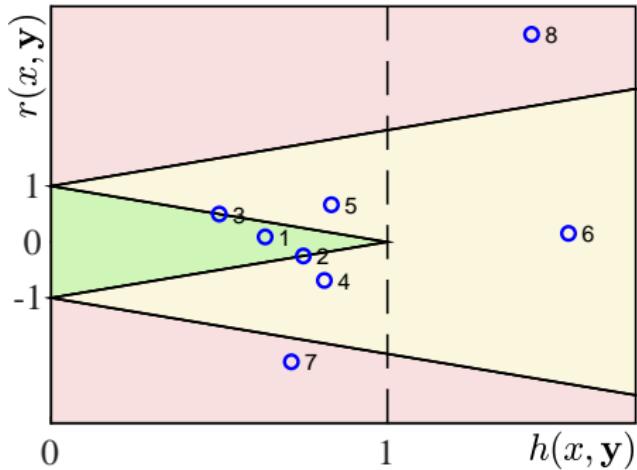


Диаграмма статусов для интервальных наблюдений, показанных на Рис. 2. Зоны наблюдений с различными статусами обозначены цветами: зелёный — внутренние наблюдения, жёлтый — внешние, красный — выбросы.

Внутреннее интервальное измерение.

Внутреннее интервальное измерение $s = (x, y)$ полностью содержит в себе прогнозный интервал, оцененный с помощью модели $\mathcal{Y}(x; S_n)$, или, иными словами, пересечение двух этих интервалов совпадает с прогнозным:

$$y \cap \mathcal{Y}(x; S_n) = \mathcal{Y}(x; S_n).$$

Будучи перестроенной по выборке, пополненной подобным измерением, модель не претерпит изменений, поскольку соответствующее ей информационное множество окажется внутри ограничения, порожденного добавленным внутренним измерением, а, следовательно, пересечение с ним не изменится. Коридор совместных зависимостей при этом также сохранит прежний вид.

Внешнее интервальное измерение.

Если внешнее интервальное измерение и соответствующий ему интервал прогноза имеют непустое пересечение, то результирующий интервал сужается по сравнению с прогнозным:

$$y \cap \Upsilon(x; S_n) \subset \Upsilon(x; S_n).$$

Это означает, что добавление внешнего измерения в модель уменьшит информационное множество задачи и коридор совместных зависимостей.

Получение пустого множества в пересечении свидетельствует о том, что измерение является выбросом по отношению к используемой модели.

В некоторых ситуациях, при более высоком уровне подозрительности, «быть тревогу» можно не при строгой пустоте информационного множества, а уже при некотором неестественно малом его размере.

Размах и остаток.

Взаимные отношения интервалов анализируемого наблюдения (x, y) и прогнозного интервала рассматриваемой модели $\Upsilon(x)$ удобно характеризовать в терминах

размаха (плечо, англ. – high leverage)

$$\ell(x, y) = \frac{\text{rad } \Upsilon(x)}{\text{rad } y} \quad (5)$$

и

остатка (остаточное отклонение, смещение, англ. – residual)

$$r(x, y) = \frac{\text{mid } y - \text{mid } \Upsilon(x)}{\text{rad } y}. \quad (6)$$

Диаграмма статусов для интервальных наблюдений.

Обе величины являются относительными, поскольку нормируются на величину неопределённости наблюдения u . Размах наблюдения косвенно характеризует положение наблюдения в пространстве независимых переменных x_i .

Наблюдения с размахом выше единицы $\ell > 1$ лежат за пределами «области определения» зависимости, образованной наблюдениями выборки, по которой построена зависимость.

Остаток характеризует смещение наблюдения по откликовой переменной u относительно коридора совместных зависимостей. Наблюдения с большими значениями размаха и остатка при их включении в выборку, по которой построен коридор совместных зависимостей, могут существенно повлиять на его вид.

Диаграмма статусов для интервальных наблюдений.

Размах и остаток позволяют установить статус наблюдения, проверив некоторые простые неравенства.

Так для внутренних наблюдений, содержащих в себе прогнозный интервал модели, выполняется нестрогое неравенство

$$|r(x, \mathbf{y})| \leq 1 - \ell(x, \mathbf{y}), \quad (7)$$

а точное равенство в нём является характеристическим условием для граничных наблюдений.

Выбросы — наблюдения, не пересекающиеся с коридором совместных зависимостей, а потому они удовлетворяют неравенству

$$|r(x, \mathbf{y})| > 1 + \ell(x, \mathbf{y}). \quad (8)$$

Диаграмма статусов для интервальных наблюдений.

Интервальные измерения, у которых величина неопределенности меньше, чем ширина прогнозного интервала, то есть

$$\ell(x, y) > 1, \quad (9)$$

могут оказывать очень сильное влияние на модель и потому называются *строго внешними*.

Иногда для обозначения строго внешнего наблюдения используется термин «абсолютно внешнее наблюдение», который по мнению авторов книги является менее удачным из-за невольной интерференции смыслов с общематематическими понятиями «абсолютная величина», «абсолютная погрешность», «абсолютно непрерывный» и т.п.

Диаграмма статусов для интервальных наблюдений.

Неравенства (7)–(9) на плоскости r, ℓ задают границы областей, соответствующих различным статусам наблюдений.

- Зона внутренних наблюдений выделена зелёным цветом.
Наблюдения, размещенные на границе зелёной зоны, являются граничными для информационного множества задачи.
- Зона внешних наблюдений — жёлтая. Правее вертикали $\ell(x, y) = 1$ лежат абсолютно внешние наблюдения.
- Выбросы локализуются в красной зоне.

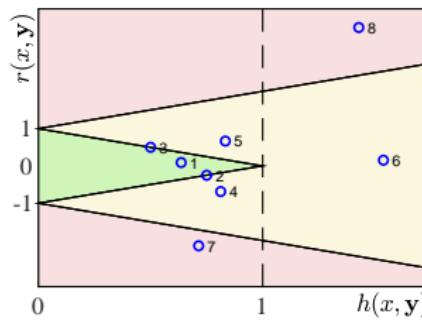


Диаграмма статусов для интервальных наблюдений.

Примечательно, что характеристизация наблюдений в терминах размахов и остатков *не зависит от размерности* входной переменной x и позволяет поддержать анализ статусов наблюдений визуальными инструментами даже в случаях, когда явное отображение информационного множества задачи и коридора совместных зависимостей затруднительно.

По своему назначению диаграмма статусов интервальных наблюдений является содержательным аналогом широко используемого в классическом регрессионном анализе *графика влияния* (англ. – influence plot), который также служит для оценки степени однородности (похожести) наблюдений и их потенциальной влиятельности на конструируемую зависимость.

Варьирование величины неопределённости измерений

Один из приёмов выявления выбросов в задаче построения зависимости по интервальным наблюдениям основан на интерпретации выбросов как наблюдений с недооценённой величиной неопределённости [5].

Закономерным шагом в этом случае становится поиск некоторой минимальной коррекции величин неопределённости интервальных наблюдений, необходимой для обеспечения совместности задачи построения зависимости.

Если величину коррекции каждого интервального наблюдения $y_i = [\hat{y}_i - \epsilon_i, \hat{y}_i + \epsilon_i]$ выборки S_n выражать коэффициентом его уширения $w_i \geq 1$, а общее изменение выборки характеризовать суммой этих коэффициентов, то минимальная коррекция выборки в виде вектора коэффициентов $w^* = (w_1^*, \dots, w_n^*)$, необходимая для совместности задачи построения зависимости $y = f(x, \beta)$ может быть найдена решением задачи условной оптимизации

$$\text{найти } \min_{w, \beta} \sum_{i=1}^n w_i \quad (10)$$

при ограничениях

$$\begin{cases} \hat{y}_i - w_i \epsilon_i \leq f(x_i, \beta) \leq \hat{y}_i + w_i \epsilon_i, \\ w_i \geq 1, \end{cases} \quad i = 1, \dots, n. \quad (11)$$

Варьирование величины неопределённости измерений

Результирующие значения коэффициентов w_i^* , строго превосходящие единицу, указывают на наблюдения, требующие уширения интервалов неопределённости для обеспечения совместности данных и модели.

Именно такие наблюдения заслуживают внимания при анализе данных на выбросы.

Значительное количество подобных наблюдений может говорить либо о неверно выбранной структуре зависимости, либо о том, что величины неопределённости измерений занижены во многих наблюдениях (например, в результате неверной оценки точности измерительного прибора).

Варьирование величины неопределённости измерений

Следует отметить значительную гибкость языка неравенств. Он даёт возможность переформулировать и расширять систему ограничений (11) для учёта специфики данных и задачи при поиске допустимой коррекции данных, приводящей к разрешению исходных противоречий.

Например, если имеются основания считать, что величина неопределённости некоторой группы наблюдений одинакова и при коррекции должна увеличиваться синхронно, то система ограничений (11) может быть пополнена равенствами вида

$$w_{i_1} = w_{i_2} = \dots = w_{i_K},$$

где i_1, \dots, i_K — номера наблюдений группы.

В случае, когда в надёжности каких-либо наблюдений исследователь уверен полностью, при решении задачи (10)–(11) соответствующие им величины w_i можно положить равными единице, т.е. запретить варьировать их неопределённость.

Варьирование величины неопределённости измерений

Задача поиска коэффициентов масштабирования величины неопределённости (10)–(11) сформулирована для распространённого случая уравновешенных интервалов погрешности и подразумевает синхронную подвижность верхней и нижней границ интервалов неопределённости измерений y_i при сохранении базовых значений интервалов \hat{y}_i неподвижными.

Варьирование величины неопределённости измерений

При необходимости постановка задачи легко обобщается.

Например, если интервалы наблюдений не уравновешены относительно базовых значений (то есть $y_i = [\hat{y}_i - \epsilon_i^-, \hat{y}_i + \epsilon_i^+]$ и $\epsilon^- \neq \epsilon^+$), то границы интервальных измерений можно варьировать независимо, масштабируя величины неопределённости ϵ_i^- и ϵ_i^+ с помощью отдельных коэффициентов w_i^- и w_i^+ :

$$\text{найти} \quad \min_{w^-, w^+, \beta} \sum_{i=1}^n (w_i^- + w_i^+) \quad (12)$$

при ограничениях

$$\left\{ \begin{array}{l} \hat{y}_i - w_i^- \epsilon_i^- \leq f(x_i, \beta) \leq \hat{y}_i + w_i^+ \epsilon_i^+, \\ w_i^- \geq 1, \\ w_i^+ \geq 1, \end{array} \right. \quad i = 1, \dots, n. \quad (13)$$

Варьирование величины неопределённости измерений

Для линейной по параметрам β зависимости $y = f(x, \beta)$ задача (10)–(11) представляет собою задачу линейного программирования, решатели которой широко доступны и в виде библиотек на различных языках программирования, и в виде стандартных процедур систем компьютерной математики, и в виде интерактивных подсистем электронных таблиц.

Пример

Наблюдения из таблицы 1 получены четырьмя различными способами A, B, C и D , обеспечивающими различную величину неопределённости ϵ ; измерений выходной переменной $y_i = [\hat{y} - \epsilon_i, \hat{y} + \epsilon_i]$ для точно задаваемых значений входной переменной x_i .

Диаграмма рассеяния интервальных данных приведена на рисунке 3.

По данным требуется построить линейную зависимость

$$y = \beta_0 + \beta_1 x.$$

Пример

Таблица: Данные с выбросами

Номер измерения <i>i</i>	Способ измерения	x_i	\hat{y}_i	ϵ_i
1	<i>A</i>	1	2.13	0.20
2	<i>A</i>	2	2.95	0.20
3	<i>A</i>	3	5.01	0.20
4	<i>A</i>	4	4.99	0.20
5	<i>A</i>	5	5.97	0.20
6	<i>B</i>	6	7.04	0.40
7	<i>B</i>	7	8.02	0.40
8	<i>C</i>	8	8.15	0.40
9	<i>C</i>	9	10.01	0.40
10	<i>D</i>	10	10.98	0.50

Пример

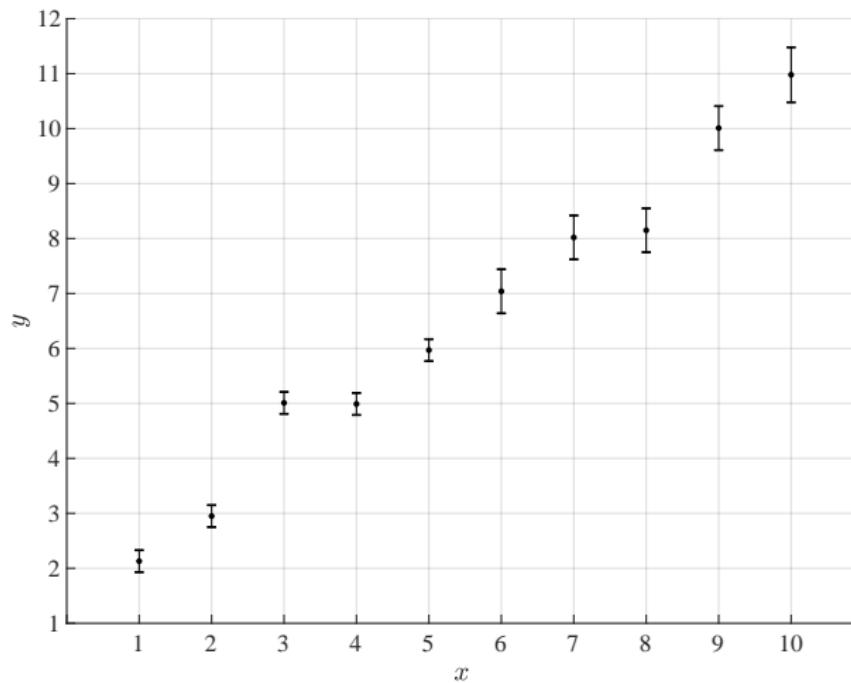


Рис.: Данные с выбросами.

Пример

Попытка построения зависимости по исходным данным приводит к пустому информационному множеству задачи и необходимости выяснения причин несовместности задачи.

Чтобы понять, имеются ли выбросы, для начала пытаемся решить в отношении данных задачу (10)–(11). Полученные в результате значения масштабирующих коэффициентов величины неопределённости w_i^* приведены в первом столбце таблицы 2.

Пример

Анализируя эти числа, можно прийти к выводу, о том, что третье и восьмое наблюдения несовместны с остальными.

Причём третье наблюдение особенно не вписывается в общую картину, поскольку сделать его совместным с прочими возможно только при расширении интервала измерения более, чем в четыре с половиной раза, и это при том, что способ A , которым получено третье наблюдение, является наиболее точным из всех четырёх.

Эти соображения позволяют нам склониться к заключению, что третье измерение вполне вероятно может оказаться результатом грубых промахов и потому стоит проанализировать данные с исключением этого измерения из всех дальнейших построений.

Пример

Что касается восьмого наблюдения, то его несовместность менее выражена. Поэтому здесь разумно отработать как гипотезу о грубых промахах во время проведения этого измерения, так и о возможной переоценке точности способа C .

С этой целью задачу (10)–(11) нужно решать, либо исключая из рассмотрения восьмое наблюдение, либо предполагая, что способ C менее точен, чем продекларировано в таблице, а значит величины неопределенности всех наблюдений, выполненных способом C , должны быть откорректированы синхронно, то есть

$$w_8 = w_9. \quad (14)$$

Пример

Таблица: Коэффициенты масштабирования величины неопределённости интервальных измерений для данных из таблицы 1

Номер измерения <i>i</i>	Решение задачи (10)–(11) w_i^*	Решение задачи (10)–(11), (14) w_i^*
1	1.000	1.000
2	1.000	1.000
3	4.686	–
4	1.000	1.000
5	1.000	1.000
6	1.000	1.000
7	1.000	1.000
8	1.343	1.143
9	1.000	1.143
10	1.000	1.000

Пример

Результат решения задачи (10) при ограничениях (11) и (14), приведенный во втором столбце таблицы 2, говорит о том, что для совместности задачи исходная величина неопределённости измерений, полученных способом C , не может иметь значение менее, чем $w_8^* \epsilon_8 = w_9^* \epsilon_9 = 1.143 \cdot 0.40 \approx 0.46$.

Такой вывод, конечно, не может служить основанием автоматического увеличения ширины интервалов неопределённости восьмого и девятого измерений до указанного уровня, а может означать лишь необходимость дополнительной проверки точности способа измерений C .

Пример

Таким образом, исследования, проведённые в отношении задачи построения линейной зависимости по данным из таблицы 1 позволяют сформулировать следующие гипотезы о причинах несовместности задачи, заслуживающие содержательной проверки:

- и третье, и восьмое наблюдение являются результатами грубых промахов и должны быть исключены из дальнейшего рассмотрения;
- третье наблюдение является результатом грубых промахов и должно быть удалено из набора данных, а способ измерений С менее точен и поэтому величина неопределенности всех выполненных им измерений должны быть увеличена.

Пример

Конечно же, наряду с гипотезами о некорректности данных не стоит забывать о всегда имеющейся альтернативной гипотезе о некорректном виде конструируемой модели, хотя для выбора иной структуры модели (скажем, квадратичной вместо линейной), как правило, нужны довольно весомые основания.

Отработка этих гипотез даёт шанс конструктивно преодолеть несовместность задачи построения зависимости и перейти к задаче построения зависимости с непустым информационным множеством, которое может подвергаться дальнейшему содержательному анализу.

Точки чебышёвского альтернанса

Следуя [?], изложим простой полуэвристический приём для выявления измерений, подозрительных на выбросы, в рамках общей вычислительной схемы метода максимума совместности (см. §??). Он основан на гипотезе о том, что «выбросы — это наиболее конфликтующие между собой измерения» \Rightarrow точки чебышёвского альтернанса.

Точки чебышёвского альтернанса

Исходным пунктом нашей методики является то простое наблюдение, что выражения для распознающих функционалов (??) имеют весьма специальный вид, в котором окончательное значение получается как минимум от значений ряда выражений одинаковой структуры (стоящих внутри фигурных скобок в (??)), которые вычисляются по строкам матрицы данных (3).

Мы будем называть их *образующими* распознающих функционалов. Фактически, их значения в точке $x = (x_1, x_2, \dots, x_n)^\top$ характеризуют отдельные измерения, давая для каждого из них меру совместности (согласования) данных в этом измерении с вектором параметров $x = (x_1, x_2, \dots, x_n)^\top$.

Точки чебышёвского альтернанса

С другой стороны, выбросы — это измерения, удаление которых резко увеличивает меру совместности оставшейся части выборки. Как следствие, приходим к следующей естественной идее. В точке максимума распознающего функционала нужно посмотреть на значения его образующих, соответствующих отдельным измерениям, и если какие-то из этих образующих существенно меньше остальных, то они и являются кандидатами на выбросы.

Высказанная идея верна по сути, но на пути её успешного применения стоят некоторые принципиальные ограничения, которые следует учитывать при интерпретации результатов расчётов.

Точки чебышёвского альтернанса

Напомним, что в пределе, когда интервалы неопределённости данных вырождаются в точки и мы должны восстанавливать зависимость по точным данным, метод максимума совместности (как слабая, так и сильная версии) переходит в чебышёвское сглаживание данных (см. обоснование в [?, ?, ?], т. е. в их приближение в равномерной метрике. Один из основных результатов теории равномерного приближения функций — это теорема Чебышёва.

Теорема Чебышёва

Теорема Чебышёва (см., к примеру, [?, ?])

Для того, чтобы многочлен n -ой степени $P(x)$ являлся многочленом наилучшего равномерного приближения непрерывной на интервале $[a, b]$ функции $f(x)$, необходимо и достаточно, чтобы на $[a, b]$ существовали по крайней мере $(n + 2)$ точки $x_0 < x_1 < \dots < x_n < x_{n+1}$, такие что разность $f(x_i) - P(x_i)$, $i = 0, 1, \dots, n + 1$, принимает в них равные по абсолютной величине значения, которые последовательно меняют знак от точки к точке.

Точки чебышёвского альтернанса

Точки $x_0 < x_1 < \dots < x_n < x_{n+1}$, о которых идёт речь в теореме Чебышёва, называются, как известно, точками чебышёвского альтернанса. Если ищется наилучшее равномерное приближение линейной функцией, т. е. полиномом первой степени $n = 1$, то $n + 2 = 3$, так что точек альтернанса должно быть не менее трёх штук. Но нередко их бывает гораздо больше. Нетрудно понять, что точки альтернанса соответствуют тем измерениям, значения образующих для которых — наименьшие, и из сделанного наблюдения следует, что таких точек не может одна или две. Их принципиально не меньше трёх, и, вообще говоря, может быть больше.

Точки чебышёвского альтернанса

Что происходит в случае интервальных данных? Вместо точек мы имеем брусы неопределённости измерений в пространстве \mathbb{R}^{n+1} , так что в общем случае теорема Чебышёва здесь, строго говоря, неприменима. Тем не менее, если интервалы данных «не слишком широки» (или «достаточно узки»), то теорема Чебышёва всё-таки остаётся верной, и мы можем считать, что количество точек альтернанса остаётся равным как минимум $n + 2$, т. е. 3 в линейном случае. Опять-таки, в реальных ситуациях их может быть довольно много, что хорошо демонстрируется при работе с практическими задачами.

Точки чебышёвского альтернанса

Таким образом, в методе максимума совместности выбросы, если они имеются, в силу принципиальных математических причин могут маскироваться обычными информативными измерениями.

Тем не менее, если количество обрабатываемых измерений велико, то любая дополнительная информация о выбросах, любая техника, позволяющая сузить «круг подозреваемых», может оказаться полезной и имеет смысл быть применённой. Особенно, когда затраты на её реализацию пренебрежимо малы, как это имеет место с предложенной выше методикой исследования образующих распознающего функционала в точке максимума.

Литература

-  А.Н. Баженов, С.И. Жилин, С.И. Кумков, С.П. Шарый. Обработка и анализ данных с интервальной неопределённостью. РХД. Серия «Интервальный анализ и его приложения». Ижевск. 2021. с.200.
-  С.П. Шарый. Конечномерный интервальный анализ. — Новосибирск: XYZ, 2021. — Электронная книга, доступная на <http://interval.ict.nsc.ru/Library/InteBooks/SharyBook.pdf>
-  С.И.Жилин. Примеры анализа интервальных данных в Octave <https://github.com/szhilin/octave-interval-examples>
-  С.И.Жилин. Библиотека полной интервальной арифметики `kinterval` в среде Octave. Частное сообщение.
-  Жилин С.И. Нестатистические методы и модели построения и анализа зависимостей. – Барнаул, 2004. – Диссертация на соискание учёной степени канд. физ.-мат. наук по специальности 09.00.02