

Тема X3. Обработка и анализ данных с интервальной неопределённостью.

А.Н. Баженов

ФТИ им. А.Ф.Иоффе

a_bazhenov@inbox.ru

31.03.2022

Обработка и анализ данных с интервальной неопределённостью.

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 неопределённостью. РХД.
Серия «Интервальный анализ и его приложения». Ижевск. 2021. с.200.

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 неопределённость*, когда нам известны лишь некоторые интервалы, дающие двусторонние границы измеренных значений.

Таким образом, результатом i -го измерения являются такие интервалы $\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \dots, \mathbf{x}_m^{(i)}, \mathbf{y}^{(i)}$, относительно которых мы предполагаем, что истинное значение x_1 лежит в пределах $\mathbf{x}_1^{(i)}$, истинное значение x_2 лежит в $\mathbf{x}_2^{(i)}$ и т.д. вплоть до y , истинное значение которого находится в интервале $\mathbf{y}^{(i)}$.

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

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

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, ограничиваемому в многомерном пространстве \mathbb{R}^{m+1} этими интервалами по отдельным координатным осям.

Definition

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

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.

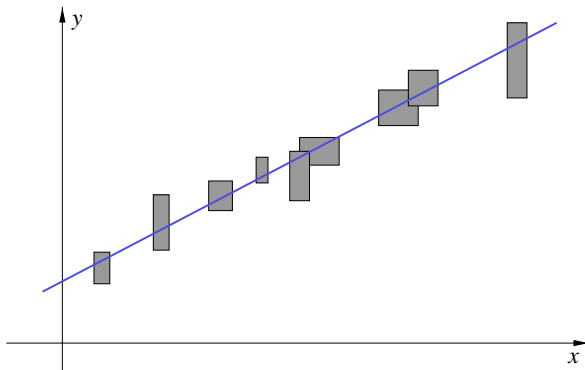


Рис.: Наглядная иллюстрация задачи восстановления линейной зависимости по данным с интервальной неопределённостью.

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 (бруса неопределённости). При отсутствии неопределённости 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.

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 подобранной 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 интервальному 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.

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.

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).

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

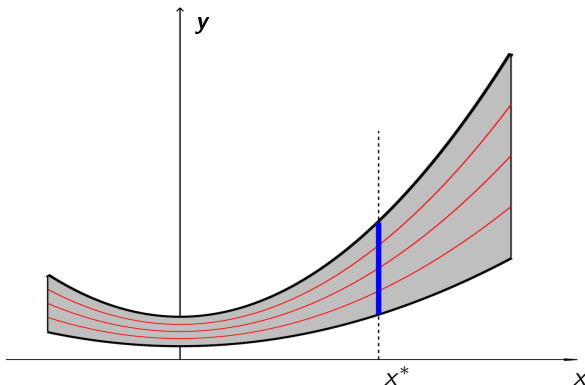


Рис.: Коридор совместных зависимостей и его сечение для какого-то значения аргумента x^* .

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 \mathcal{Y} , сопоставляющее каждому значению аргумента x множество

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

Сечение коридора совместных зависимостей

Значение $\mathcal{Y}(\tilde{x})$ 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 брусков 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

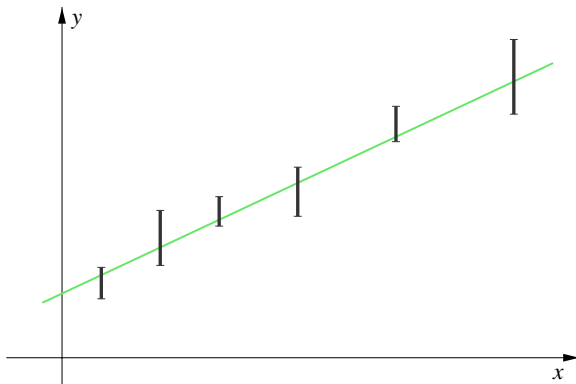


Рис.: Частный случай задачи восстановления линейной зависимости по неточным данным, когда входные переменные измеряются точно.

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.

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.

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

$$\begin{cases} \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1m}\beta_m = y_1, \\ \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \dots + x_{2m}\beta_m = y_2, \\ \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \ddots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ \beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \dots + x_{nm}\beta_m = y_n, \end{cases}$$

у которой интервальность присутствует только в правой части.

С другой стороны, (4) равносильно системе

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

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

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

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

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

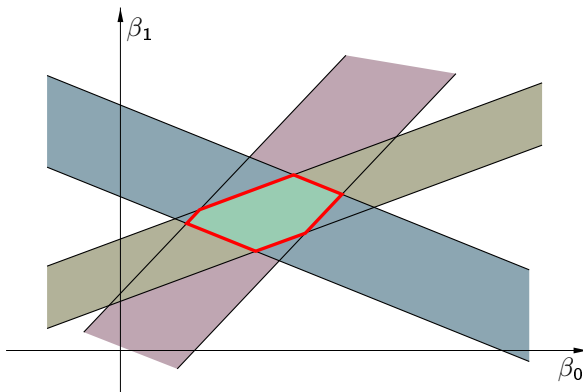


Рис.: Образование информационного множества параметров
линейной зависимости (ограничено красной линией)
для случая точных входных переменных.

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 [3].

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.

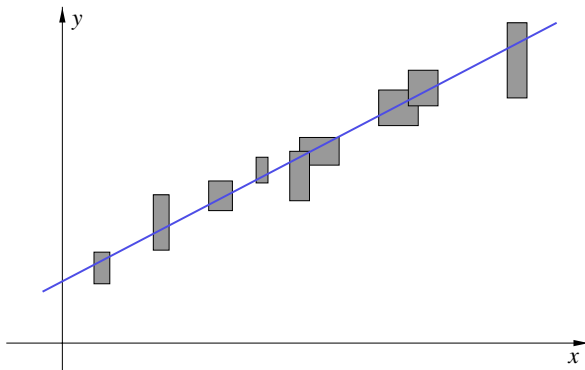


Рис.: Наглядная иллюстрация задачи восстановления линейной зависимости по данным с интервальной неопределённостью.

`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

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

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. [3]).

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).

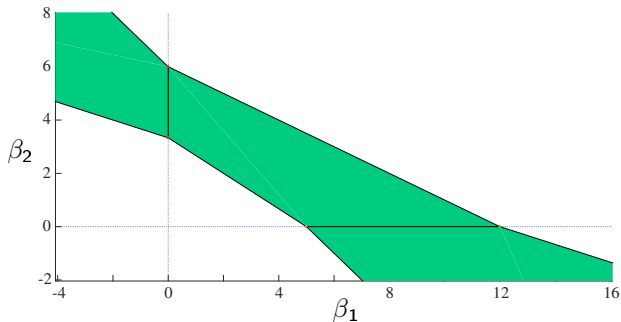


Рис.: Объединённое множество решений интервального линейного уравнения $[1, 2]\beta_1 + [2, 3]\beta_2 = [10, 12]$.

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 брусам 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_i 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

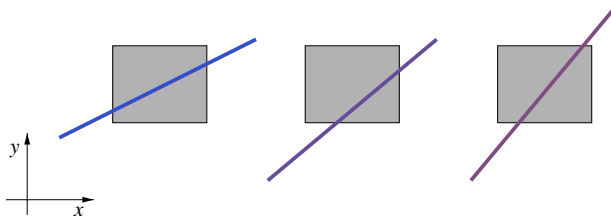


Рис.: Различные способы пересечения линии с бруском
неопределённости измерения зависимости.

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.

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).

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

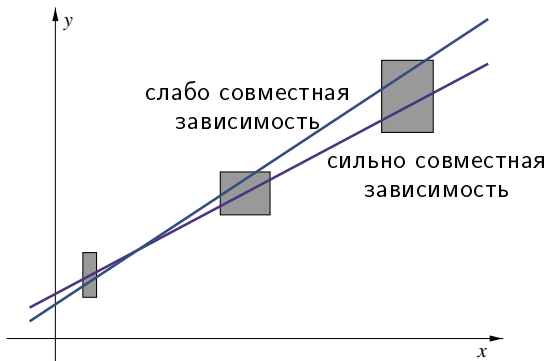


Рис.: Линейные зависимости с разными типами согласования с данными.

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.

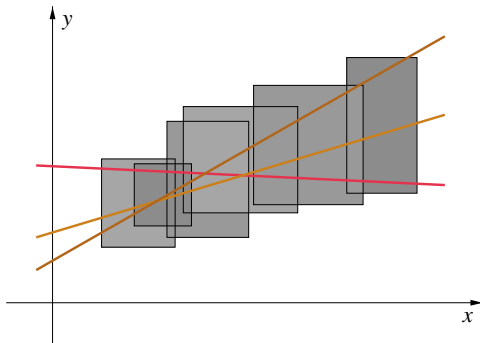


Рис.: Сложный случай восстановления зависимости по широким перекрывающимся интервальным данным.

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.

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

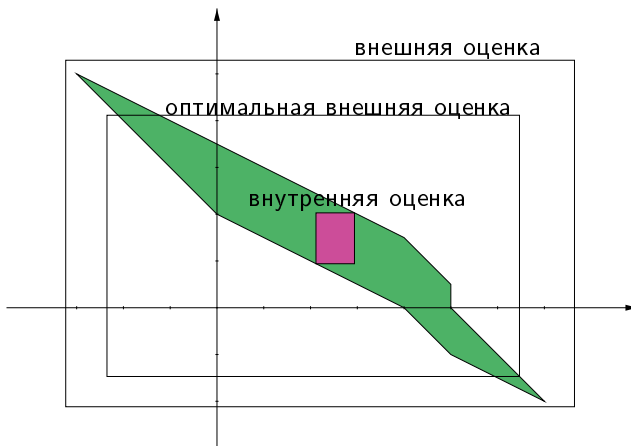


Рис.: Различные способы оценивания
информационного множества.

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.

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

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 брусков 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 выбрать 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.

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.

Разведочный анализ на основе здравого смысла.

- Раздельная обработка данных для каждой ветви
- Совместная обработка всех данных
-
- Намётки общей теории

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.

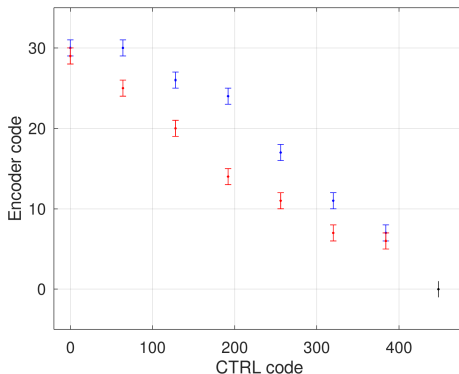


Рис.: Диаграмма рассеяния движения «вперёд-назад».

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.

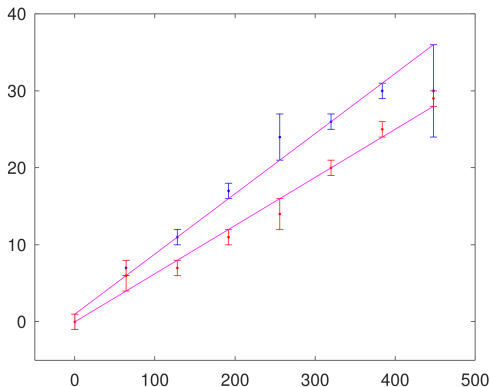


Рис.: Регрессии на разные ветви данных для движения «вперёд-назад» с оценкой по норме L_1 .

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 .

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

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

Разница между ветвями проявляется в величинах коэффициентов регрессии:

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

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

Определим теперь интервальные параметры регрессии [4].

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.

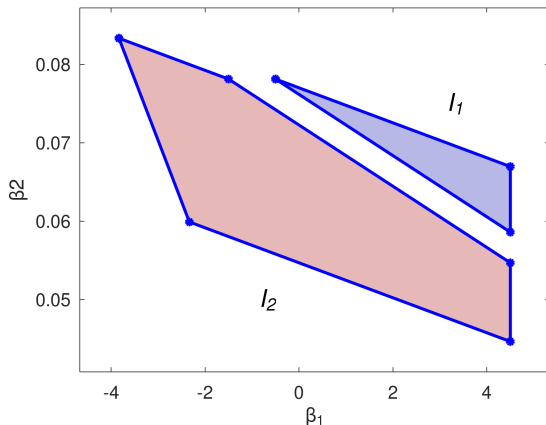


Рис.: Интервальные оценки для разных ветвей зависимости. Назначенное значение погрешности данных $\varepsilon=4.5$.

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

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

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.

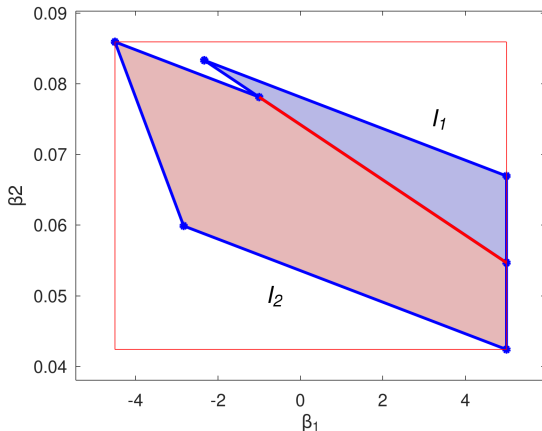
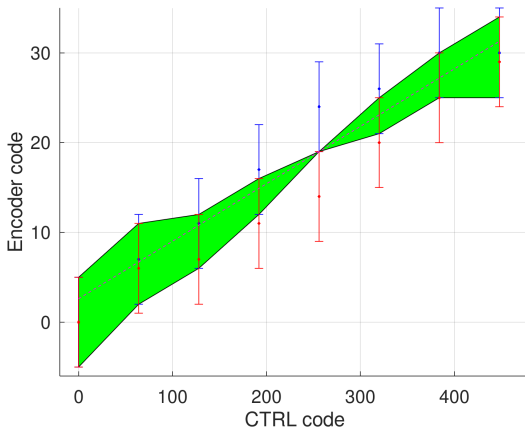


Рис.: Интервальные оценки для разных ветвей зависимости и множество $I(\beta_1, \beta_2)$. Назначенное значение погрешности данных (11) $\varepsilon=5$.

Коридор совместности γ .

На Рис. 15 приведены диаграмма рассеяния данных и коридор совместности параметров модели регрессии γ 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 $\mathcal{I}(x^*)$ состоит из одной точки.

Линейные регрессии.

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

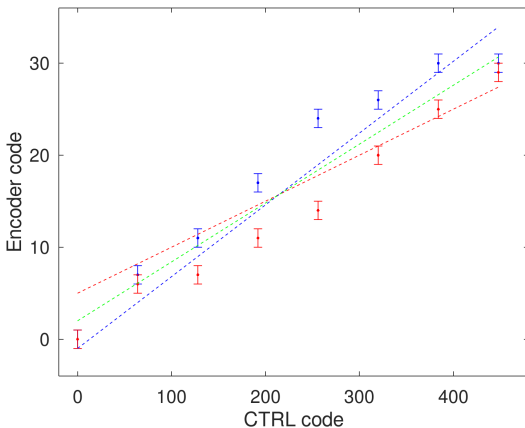


Рис.: Набор линейных регрессий.

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.

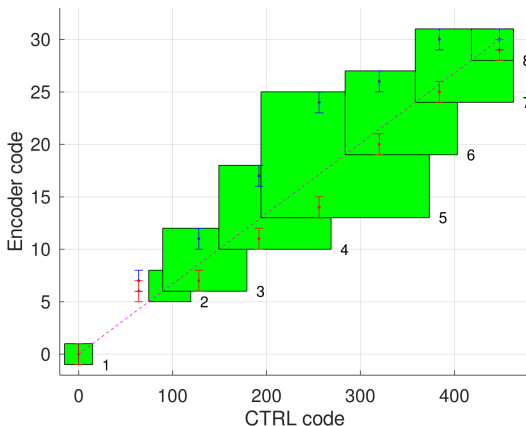


Рис.: Брусы совместности данных. Номер замера дан правее его правого нижнего угла бруса.

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

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

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.

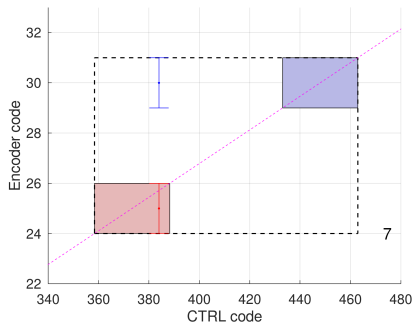
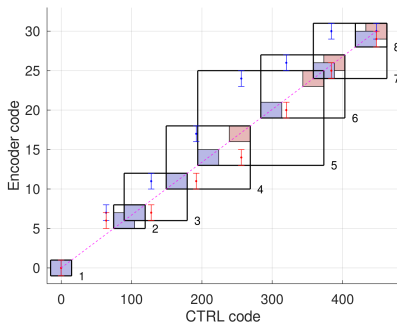


Рис.: Брусы совместности данных по отдельности для каждого замера и для пары «вперёд-назад» совместно. Номер замера дан правее его правого нижнего угла бруса.. Справа — один замер.

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

$$x_i \rightarrow \mathbf{x}_i, \quad i = 1, 2, \dots, 15.$$

так чтобы регрессионная прямая прошла через все брусы (\mathbf{x}_i, y_i) .

Совместимость за счет коррекции входных данных.

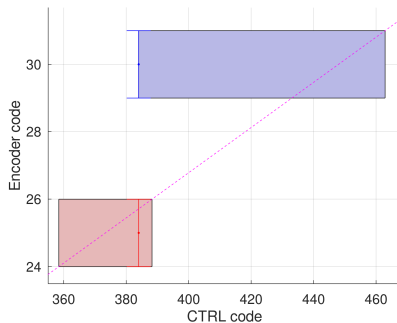
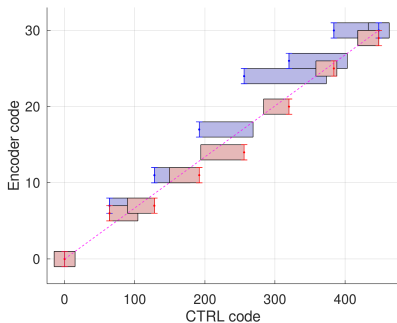


Рис.: Совместимость за счет входных данных. Справа — один замер.

Совместимость за счет коррекции входных данных.

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

При постановке задачи линейного программирования

$$\sum_i \text{rad } \mathbf{x}_i \rightarrow \min,$$

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.

Диаграмма рассеяния данных.

Вернёмся к исходным данным.

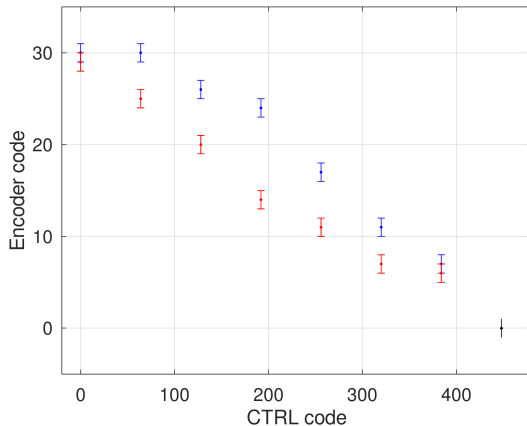


Рис.: Диаграмма рассеяния движения «вперёд-назад».

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 интервальной регрессии.

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

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 в виду 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

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

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.

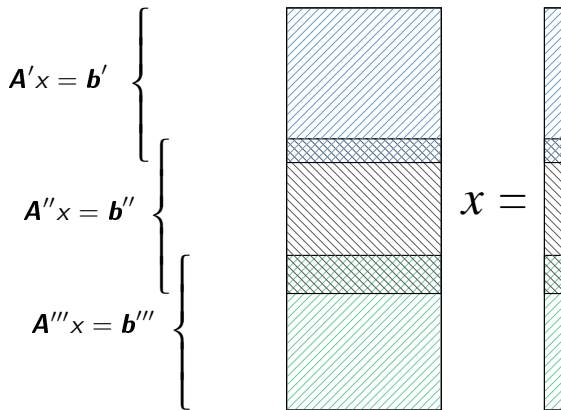


Рис.: Иллюстрация разбиения интервальной линейной системы уравнений на квадратные подсистемы, которые выделены штриховками с разными наклонами.

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 рассения данных и регрессия методом квадратных подсистем.

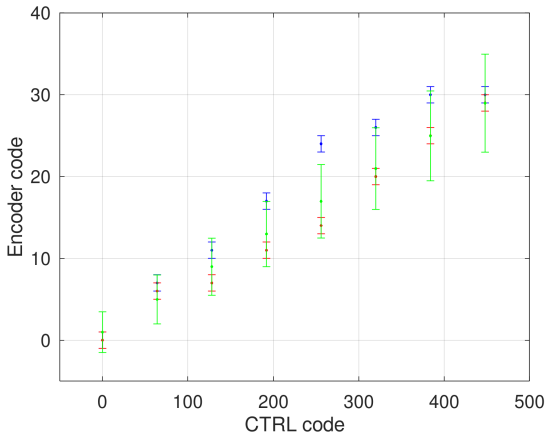


Рис.: Диаграмма рассения данных и регрессия методом квадратных подсистем.

Метод квадратных подсистем.

На Рис. 22 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* [3] с помощью библиотеки

`kinterval`

<https://github.com/szhilin/kinterval>.

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. 22 оценки выходных данных даны зеленым цветом.

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 \mathcal{Y} .

Коридор совместности \mathcal{Y} .

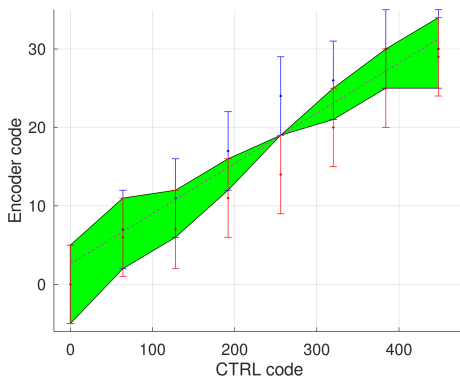


Рис.: Коридор совместности \mathcal{Y} , погрешность данных (11) $\varepsilon = 4.5$.

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


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.


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. »

-  А.Н. БАЖЕНОВ, С.И. Жилин, С.И. Кумков, С.П. ШАРЫЙ. Обработка и анализ данных с интервальной неопределённостью. (готовится к изданию). с.312.

-  А.Н. БАЖЕНОВ. Введение в анализ данных с интервальной неопределённостью. 2023.
<https://elib.spbstu.ru/dl/2/id22-247.pdf/info>

-  С.П. Шарый. Конечномерный интервальный анализ. — Новосибирск: XYZ, 2022. — 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>