МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ им. М.В. ЛОМОНОСОВА
ФИЗИЧЕСКИЙ ФАКУЛЬТЕТ

П.В. Голубцов

Теоретические основы аналитики больших данных

Учебно-методическое пособие

Москва, 2024 г.

П.В. Голубцов

Теоретические основы аналитики больших данных

В пособии рассматриваются концептуальная основа и математические инструменты, применимые к аналитике больших данных и вычислений в режиме реального времени. Изучаются основные фазы работы с большими данными, таких как извлечение, унификация, обновление и объединение информации. Особое внимание уделяется специфическим особенностей обработки больших данных, которая должны быть в высшей степени параллельной и распределённой.

Предназначено для слушателей специального курса «Теоретические основы аналитики больших данных», читающегося магистрам первого года обучения физического факультета (Осень: Отделение прикладной математики: каф. математики, каф. математического моделирования и информатики, каф. физико-математических методов управления; Весна: Астрономическое отделение, каф. астрофизики и звездной астрономии, каф. небесной механики, астрометрии и гравиметрии, каф. экспериментальной астрономии) и для всех интересующихся математическими аспектами анализа больших данных.

Рассчитано на студентов старших курсов физико-математических специальностей.

Автор — сотрудник кафедры математики физического факультета МГУ.

Объем: 82 стр. Тираж: 50 экз.

Оглавление

Предисловие

В настоящее время в различных вузах читается большое количество курсов, с различных точек зрения и с различной степенью полноты излагающих круг вопросов, которые можно объединить под общим названием наука о данных. Резкий всплеск интереса к этой тематике связан с проблематикой больших данных. При этом, в существующих курсах, посвященных анализу данных, основное внимание уделяется либо изложению классических методов, применимых, в основном, к небольшим объемам данных, либо отдельные, нередко эвристические подходы к адаптации этих методов для систем распределенного анализа больших объёмов данных. Представленный курс направлен на устранение этого очевидного несоответствия. В его основе лежит систематическое развитие общих методов работы с информацией в распределенных системах сбора и обработки данных. Методы, рассматриваемые в этом курсе, могут быть полезны для многих прикладных задач, например, при сборе и обработке информации в крупномасштабных экспериментах, где данные собираются в многочисленных исследовательских центрах, разбросанных по всему Земному шару. Данный курс рассчитан на широкий круг студентов-физиков, которые специализируются в самых разных областях физики, как теоретической, так и экспериментальной. Многие проблемы, решаемые в этих областях, являются не только фундаментальными, но и имеющими прямое практическое применение. Знание об общем состоянии исследований и проблем в этой области знаний и владение соответствующими практическими навыками может являться отличительной характеристикой высококлассного специалиста в любой области науки.

Цель курса – предоставить слушателям возможность приобрести концептуальную основу и математические инструменты, применимые к аналитике больших данных и вычислений в режиме реального времени.

Тесная связь между теоретической базой и практическими результатами определяет как подбор материала, так и характер его изложения. Данный курс предполагает выполнение домашних заданий, которые, по сути, представляют собой теоретические (исследовательские) или практические (программные) проекты, условия и основные этапы выполнения которых приведены в опубликованных учебных материалах.

1 Понятие информации в контексте задач обработки больших данных

В системах больших данных возникает необходимость в трансформации существующих алгоритмов, так, чтобы отдельные фрагменты данных обрабатывались независимо и параллельно. Соответствующий алгоритм должен, работая параллельно на многих компьютерах, извлекать из каждого набора исходных данных некоторую компактную промежуточную информацию, объединять ее и, наконец, использовать накопленную информацию для получения результата. Рассматриваются особенности такой хорошо организованной промежуточной формы информации, ее естественные алгебраические свойства и приводится иллюстративный пример.

1.1 О понятии информации

В последнее время наблюдается резкий всплеск исследований, связанных с большими данными (Big Data). Действительно, было обнаружено, что большие объемы данных могут содержать ценную информацию, возможность извлечения которой из такого рода данных ранее даже и не предполагалась. Можно сказать, что в задачах больших данных, как правило, речь идет об извлечении спрятанной информации и представлении ее в форме, пригодной для интерпретации или принятия решений. Такого рода процессы обычно проходят через несколько стадий, в которых информация извлекается из исходных данных, преобразуется, передается, накапливается и, в конце концов, трансформируется к удобному для интерпретации виду.

Отметим, использование термина «информация», в последнее время заметно возросло, особенно, в контексте анализа данных. Обычно он понимается слишком широко и неформально. Однако, возможно, такая

возросшая частота употребления этого термина свидетельствует о возрастающей потребности в более точном и формальном понимании феномена информации. Может ли проблематика больших данных приблизить нас к такому пониманию?

Исследования, связанные с системы больших данных, нацелены на проблемы обработки больших объемов распределенных данных и имеют, как правило, ярко выраженную практическую и техническую направленность. В то же время, основная масса исследований по теории информации проводится в контексте теории вероятностной и математической статистики и представляет преимущественно теоретический интерес.

Пожалуй, наиболее прикладная часть теории информации, берущая начало в работах Шеннона, связана с передачей сообщений при наличии помех. При этом речь идет не столько о «смысле» информации, сколько о ее количестве. Особое место в математической статистике занимает информация Фишера, описываемая матрицами [1], [7]. Она обеспечивает более детальное отражение понятия информации и, в частности обладает важной аддитивной структурой, в рамках которой объединению независимых статистик отвечает сумма их информационных матриц. Несмотря на многочисленные исследования по теории информации, проблема формализации понятия информации, отражающей именно смысл информации, содержащейся в данных, представляется еще далекой от удовлетворительного решения.

На данный момент сферы интересов больших данных и различных подходов к понятию информации слабо пересекаются. Однако, как уже было отмечено выше, проблематика больших данных требует более четкого, формального описания самого понятия информации и информационных процессов. Это необходимо для построения эффективных инструментов преобразования информации, опирающихся на математические (например, алгебраические) свойства информации. В связи с этим, по мнению автора,

большие данные станут в ближайшее время основным двигателем и потребителем (бенефициаром) общей теории информации. В этой работе мы попытаемся показать, как некоторая формализация понятия информации и ее алгебраические свойства могут следовать просто из рассмотрения задачи в контексте больших данных.

Чем же выделяются задачи «больших данных» на фоне задач анализа данных? Данные в таких задачах, как правило, имеют огромный объем, распределены между многочисленными сайтами и постоянно пополняются. В результате даже самый простой анализ больших данных сталкивается с серьезными трудностями. Действительно, традиционные подходы к обработке информации предполагают, что данные, предназначенные для обработки, собираются в одном месте, организуются в виде удобных структур (например, матриц), и только тогда соответствующий алгоритм обрабатывает эти структуры и выдает результат анализа. В случае больших данных невозможно собрать все данные, необходимые для исследовательского проекта на одном компьютере. Более того, это было бы непрактично, поскольку один компьютер не сможет обработать их в разумные сроки. В результате возникает необходимость в трансформации существующих алгоритмов, приводящих к их «распараллеливанию», или даже разработке новых подходов к обработке данных, которые по самой формулировке проблемы смогли бы обрабатывать отдельные фрагменты данных независимо и параллельно. Соответствующий алгоритм анализа данных должен, параллельно работая на многих компьютерах, извлекать из каждого набора исходных данных некоторую промежуточную компактную «информацию», постепенно объединять и обновлять ее и, наконец, использовать накопленную информацию для получения результата. По прибытии новых фрагментов данных он должен иметь возможность добавлять их к накопленной информации и, в конечном итоге, обновлять результат.

Мы обсудим особенности такой хорошо организованной промежуточной формы информации, выявим ее естественные алгебраические свойства и представим несколько примеров. Мы также увидим, что такая промежуточная форма представления информации в некотором смысле отражает саму суть информации, содержащейся в данных. Это приводит нас к совершенно новому подходу к самому понятию информации.

1.2 Особенности обработки в системах больших данных

Выделим следующие особенности задач обработки информации в системах больших данных:

a) Как правило, речь идет об огромных объемах данных.
b) Такие данные обычно не собраны воедино, а распределены по многочисленным, возможно, достаточно удаленным компьютерам.
c) Постоянно могут возникать новые данные, которые необходимо оперативно включать в обработку.

Традиционные методы обработки обычно не учитывают такую специфику и требуют серьезного пересмотра при необходимости их применения в задачах больших данных.

Рассмотрим бегло (и, конечно же, предельно упрощенно) стандартный подход к обработке данных. К задачам такого рода относятся задачи оценивания, принятия решений, обучения, классификации… Обычно в задачах с малым фиксированным набором данных обработка состоит в применении некоторого отображения (алгоритма, метода), определяющего обработку, к набору данных и получению результата обработки (например, оценки некоторой величины), Рис. 1.

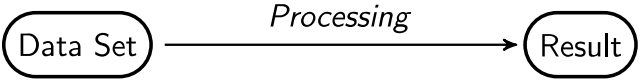

Рис. 1. Стандартный подход к обработке данных

Важным условием здесь является то, что все данные находятся в одном месте и готовы к применению к ним отображения обработки, например, представлены в виде подходящих структур, скажем, матриц. Если же данные распределены по многим различным локациям, для применения обработки их требуется сначала собрать в одном месте, организовать комбинированные данные в виде подходящих структур, и применить к ним алгоритм обработки (Рис. 2). Ключевым моментом здесь является необходимость собрать все данные в одном месте. Пунктирными стрелками здесь и далее обозначается передача данных в исходном или частично обработанном виде.

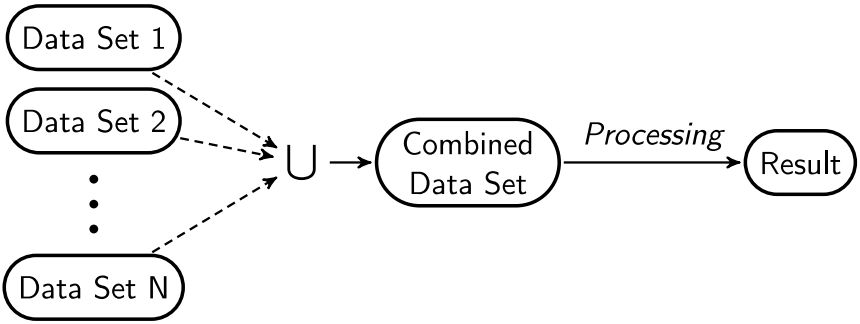

Рис. 2. Стандартный подход к обработке распределенных данных

Недостатки такого подхода к обработке распределенных данных достаточно очевидны:

a) Передача больших объемов исходных данных создаст чрезмерный траффик.
b) Хранение полного набора данных в одном месте потребует огромных объемов памяти.

c) Обработка всех данных на одном компьютере потребует чрезмерных вычислительных и временных ресурсов.

d) По мере поступления новых данных, комбинированный набор данных будет расти и, как следствие требовать постоянно возрастающих (потенциально бесконечных) ресурсов для хранения.

e) При этом, при поступлении новых данных, алгоритм обработки будет необходимо по новой применять к постоянно увеличивающемуся объему данных.

1.3 Выделение промежуточной информации в процессе обработки

Рассмотрим следующую модификацию процесса обработки, которая позволит преодолеть обозначенные выше недостатки. Предположим, что полный алгоритм обработки P допускает разбиение на две фазы $P = P_2 \circ P_1$ (Рис. 3):

a) P_1 – выделение из исходных данных некоторой промежуточной информации.

b) P_2 – вычисление результата на основании выделенной промежуточной информации.

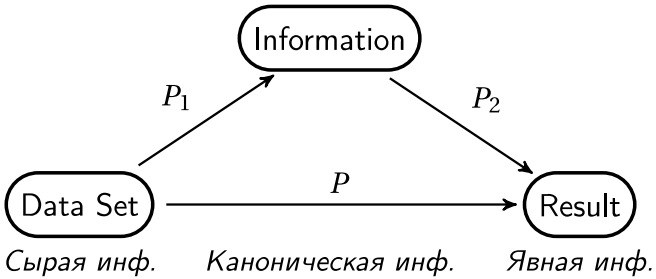

Рис. 3. Разбиение процесса обработки данных на две фазы

Выбор подходящей промежуточной формы представления информации определяется рассматриваемой задачей обработки данных. Будем называть

некоторую выбранную форму представления промежуточной информации канонической формой информации или короче, **канонической информацией**.

В определенном смысле узлы диаграммы на Рис. 3 отражают представления информации в разных формах:

c) Data Set – информация в сырой (исходной) форме.
d) Result – информация в явной (удобной для интерпретации) форме.
e) Information – информация в промежуточной (удобной для обработки) канонической форме.

Ниже мы более подробно обсудим желательные свойства канонической информации. Но сейчас отметим, что такая форма представления информации должна быть полна, то есть содержать всю необходимую для вычисления результата информацию (в этом и состоит коммутативность диаграммы на Рис. 3) и компактна, то есть иметь минимально возможный размер, в идеале, не зависящий от объема представленных данных.

Рассмотрим, как может быть трансформирована схема обработки информации, если полная обработка может быть разбита на две, указанные выше, фазы, Рис. 4.

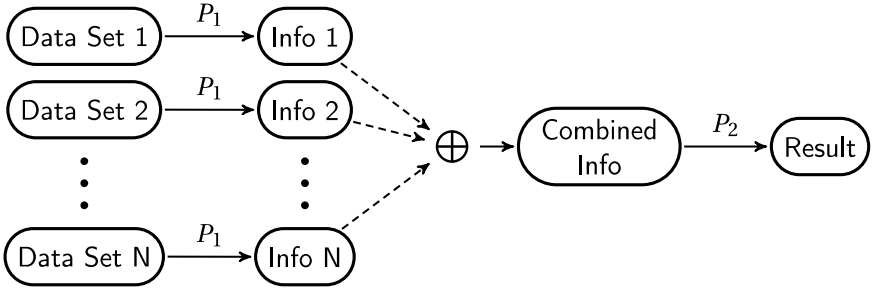

Рис. 4. Модифицированная схема обработки распределенных данных

Такая схема позволяет преодолеть все недостатки стандартной схемы обработки распределенных данных, отмеченные выше:

a) Передаются лишь компактные фрагменты выделенной промежуточной информации.
b) Хранение комбинированной информации потребует небольших объемов памяти, возможно, таких же, как и объемы, требуемые для хранения отдельных частей промежуточной информации.
c) Промежуточная информация выделяется параллельно из каждого отдельного набора данных (фаза P_1). Если основная часть обработки сосредоточена в первой фазе, то вторая фаза P_2, состоящая в построении результата по компактной накопленной информации, не потребует серьезных вычислительных и временных ресурсов.
d) По мере поступления новых данных, потребуется лишь выделить из них промежуточную информацию и «добавить» ее к накопленной.
e) При этом, алгоритм обработки будет необходимо снова применять к компактной информации фиксированного объема.

Отметим, что в приведённых выше рассуждениях мы предполагаем существование операции композиции (сложения) отдельных фрагментов канонической информации. Фактически, мы предполагаем, что на множестве всех фрагментов канонической информации определена операция композиции. При этом, объединению двух наборов данных отвечает композиция соответствующих фрагментов канонической информации, Рис. 5.

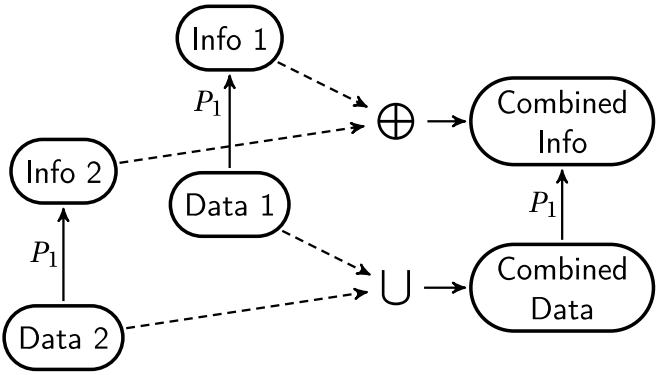

Рис. 5. Соответствие композиции фрагментов канонической информации и объединения наборов исходных данных

Это можно записать как $P_1(D_1) \oplus P_1(D_2) = P_1(D_1 \cup D_2)$, где под $D_1 \cup D_2$ понимается объединение двух наборов данных в один.

Заметим, наконец, что схема обработки распределенных данных, представленная на Рис. 4. идеально «вписывается» в архитектуру систем распределенного хранения и анализа данных. Здесь операция Map извлекает фрагменты информации из нескольких наборов данных, а операция Reduce объединяет все эти частичные фрагменты информации в один элемент, который представляет все исходные наборы данных.

1.4 Пример факторизации алгоритма путем выделения промежуточной информации

Рассмотрим следующую задачу, которая часто встречается в статистических приложениях. Подчеркнем, что мы используем для иллюстрации довольно простую задачу. При этом будем считать, что объемы наборов данных в этой задаче и количество таких наборов крайне велики.

Пусть $(x_1, x_2, \dots, x_n)$ - последовательность вещественных чисел и задача обработки состоит в вычислении выборочного среднего:

$$X = \frac{1}{n}\sum_{i=1}^{n} x_i. \quad (1)$$

Таким образом, исходным набором данных является последовательность векторов $(x_1, x_2, \dots, x_n)$, а требуемым результатом обработки P является среднее X, определяемое выражением (1), т.е., $P(x_1, \dots, x_n) = (X, V)$.

$$(x_1, \dots, x_n) \xrightarrow{\boldsymbol{P}} X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Рис. 6. Стандартный алгоритм P

1.4.1 Стандартный подход для распределенных данных

Если же исходные данные содержатся в N наборах $(x_1, \dots, x_{n_1})$,…,$(z_1, \dots, z_{n_N})$, размещенных на различных компьютерах, то для их обработки с помощью этого алгоритма придется собрать их в одном месте и применить преобразование P, Рис. 7.

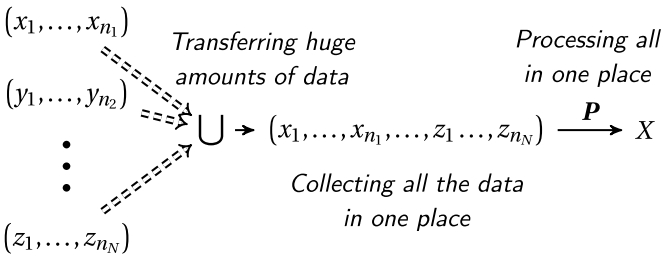

Рис. 7. Объединение исходных данных для обработки

При использовании такой схемы потребуется передавать большие объемы исходных данных, хранить и обрабатывать полный набор $(x_1, \dots, x_{n_1}, \dots, z_1, \dots, z_{n_N})$ на одном компьютере. При поступлении нового набора данных придется добавить его к уже имеющемуся полному набору и пересчитать результат X.

1.4.2 Каноническая информация

Заметим, однако, что вычисление X в (1) можно разбить на два этапа. Пусть

$$S = \sum_{i=1}^{n} x_i. \tag{2}$$

Тогда $X = \frac{S}{n}$.

Таким образом, вся информация, достаточная для вычисления X может быть представлена тройкой (n, S) и процесс обработки P может быть разбит на две стадии $P = P_2 \circ P_1$ (Рис. 8).

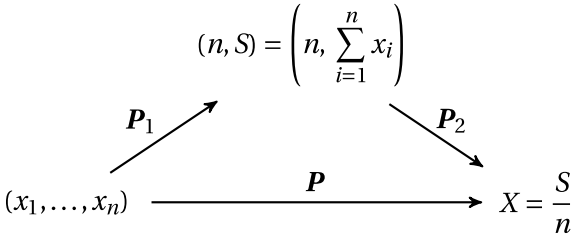

Рис. 8. Разбиение процесса обработки на две стадии путем выделения канонической информации

В данной задаче пара (n, S) представляет собой удобную промежуточную форму представления информации об исходных данных в рассматриваемой задаче - *каноническую информацию*. Заметим, что она определяется двумя числами, независимо от объема данных, которые он представляет. На самом деле, первое число n говорит, сколько данных представлено (n, S).

1.4.3 Пересмотренная схема обработки

В результате разбиения алгоритма P на две фазы и введения канонической информации, схема обработки распределенных данных, представленная на Рис. 7, может быть трансформирована следующим образом

(Рис. 9). Из каждого отдельного фрагмента данных выделяется каноническая информация (n_j, S_j), которая впоследствии объединяется и используется для вычисления результата.

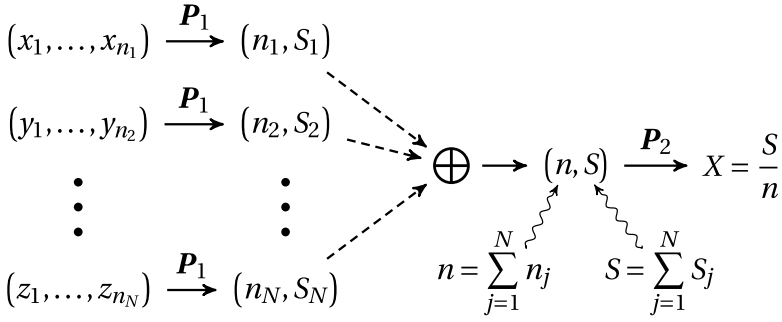

Рис. 9. Модифицированная схема обработки распределенных данных

Отметим основные особенности такой модифицированной схемы:

a) Выделение канонической информации (n_j, S_j) из j-того набора данных (преобразование P_1) может проводиться «на местах» параллельно и независимо. В результате, распределенность исходных данных способствует повышению эффективности обработки за счет распараллеливания.

b) Передаются лишь компактные фрагменты выделенной канонической информации одинакового объема (2 числа), не зависящего от объема исходного набора данных.

c) Сложение частей канонической информации максимально упрощено и определяется покомпонентным сложением пар (n_j, S_j):

$$(n_1, S_1) \oplus (n_2, S_2) = (n_1 + n_2, S_1 + S_2). \quad (3)$$

d) Хранение всей комбинированной канонической информации также требует такого же небольшого объема памяти (2 числа).

e) Поскольку основная часть обработки сосредоточена в первой фазе, то вторая фаза P_2, состоящая в построении результата по компактной

накопленной информации не зависит от объема исходных данных и не требует серьезных вычислительных и временных ресурсов.

f) По мере поступления новых данных, потребуется лишь выделить из них канонической информацию и «добавить» ее к накопленной.

g) При этом, алгоритм обработки будет необходимо снова применять к компактной информации фиксированного объема.

Заметим, что пары вида (n, S) можно рассматривать как элементы некоторого множества, наделенного дополнительной структурой – канонического **информационного пространства** $\mathfrak{I}$. В данном примере $\mathfrak{I} = \mathfrak{I}_1 = \mathbb{N} \times \mathbb{R}$, где $\mathbb{N} = \{0,1,\dots\}$ - множество натуральных чисел, а $\mathbb{R}$ – множество вещественных чисел. При этом согласно (3), на пространстве $\mathfrak{I}_1$ задана операция композиции $\oplus$, определяемая покомпонентно.

1.5 Пример 2. Добавление выборочной дисперсии к цели обработки

Теперь давайте немного изменим наш предыдущий пример, изменив цель обработки. Предположим, что набор данных тот же, что и раньше, т. е. $D = (x_1, x_2, \dots, x_n)$, но в дополнение к выборочному среднему $X = \frac{1}{n}\sum_{i=1}^{n} x_i$ нам нужно вычислить также выборочную дисперсию (Рис. 10)

$$V = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - X)^2. \tag{4}$$

$$(x_1, \dots, x_n) \xrightarrow{P} (X, V) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i,\ \ \frac{1}{n-1}\sum_{i=1}^{n}(x_i - X)^2\right)$$

Рис. 10 . Стандартный подход к обработке для новой задачи

Поскольку вычисления включают X и все исходные x_i, может показаться, что нам придется вычислять X и V за два прохода и, таким

образом, сохранять все исходные данные для выполнения таких вычислений. Однако, поскольку

$$\sum_{i=1}^{n}(x_i - X)^2 = \sum_{i=1}^{n} {x_i}^2 - X\sum_{i=1}^{n} x_i - \left(\sum_{i=1}^{n} x_i\right)X + nX^2,$$

Это выражение можно записать как

$$\sum_{i=1}^{n}(x_i - X)^2 = T - \frac{1}{n}S^2,$$

где

$$T = \sum_{i=1}^{n} {x_i}^2.$$

Это не только позволяет вычислять выборочное среднее и дисперсию за один проход, но и предлагает естественный способ распараллеливания вычислений для распределенных данных. Действительно, мы можем просто изменить нашу предыдущую форму канонической информации, добавив к ней T. В результате

$$X = \frac{\mathrm{S}}{n}, \qquad V = \frac{1}{n(n-1)}(nT - S^2) \tag{5}$$

и соответствующая факторизация обработки $\boldsymbol{P}$ примет вид, показанный на Рис. 11.

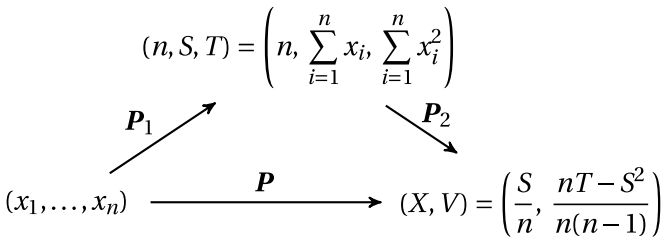

Рис. 11. Факторизация обработки на два этапа для измененной цели.

В результате мы приходим к новой форме канонической информации, подходящей для модифицированной проблемы обработки. Теперь каноническая информация представлена тройкой (n, S, T) и новым информационным пространством $\mathfrak{I}_2 = \mathbb{N} \times \mathbb{R} \times \mathbb{R}_+$, где $\mathbb{R}_+$ — множество неотрицательных вещественных чисел. Как и раньше, пространство $\mathfrak{I}_2$ снабжено покомпонентной операцией композиции $\oplus$:

$$(n_1, S_1, T_1) \oplus (n_2, S_2, T_2) = (n_1 + n_2, S_1 + S_2, T_1 + T_2), \tag{6}$$

что отвечает объединению двух наборов данных, т.е.

$$P_1(x_1, \ldots, x_{n_1}) \oplus P_1(y_1, \ldots, y_{n_2}) = P_1(x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}).$$

Любое отдельное наблюдение x может быть добавлено к собранной канонической информации «на лету» с помощью следующей операции обновления:

$$(n, S, T) \oplus x = (n + 1, S + x, T + x^2).$$

Легко видеть, что все упомянутые выше полезные особенности пересмотренной двухэтапной схемы обработки справедливы и для этой модифицированной задачи.

Более того, эти два примера показывают, что более сложная цель может потребовать более сложного информационного пространства. Точнее, $\mathfrak{I}_1$ можно рассматривать как подпространство $\mathfrak{I}_2$. Это иллюстрирует, что определенная иерархия целей обработки (для тех же типов данных) должна приводить к соответствующей иерархии информационных пространств.

1.6 Основные свойства хорошо организованной промежуточной информации

В приведенных выше примерах мы видели, что выбор удобной формы представления промежуточной информации позволяет существенно повысить эффективность обработки распределенных данных. Эти примеры наталкивают на следующие желательные свойства канонической информации.

Существование для любого исходного набора данных. Любой исходный набор данных должен допускать представление информации в каноническом виде. В частности, минимальный «атомарный» набор данных или даже «пустой» набор должны быть представимы в канонической форме. В нашем примере это условие выполнено. Так атомарному набору (x), состоящему из единственного столбца x, отвечает каноническая информация $P_1((x)) = (1, x, xx)$, а пустому набору (), отвечает «нулевая» каноническая информация $\mathbf{0} = P_1(()) = (0,0,0)$.

Заметим, что вычисление окончательного результата может оказаться невозможным для некоторых наборов исходных данных. В частности, согласно (4), для вычисления выборочной дисперсии необходимо, чтобы исходные данные содержали, как минимум, два элемента. Строго говоря, отображение P является не всюду определенным. В то же время, мы требуем, чтобы P_1 было всюду определено.

Полнота (или **достаточность**). Каноническая форма должна содержать всю информацию, содержащуюся в исходных данных, а именно, она должна приводить к тому же результату, что и исходные данные, из которых она получена. Формально это означает что $P(D) = P_2(P_1(D))$ для всех данных D из области определения преобразования P.

Единственность представления данных в каноническом виде. Фактически, это свойство означает отсутствие избыточности в канонической

информации. Отсюда, в частности, следует, что каноническая информация не должна зависеть от порядка данных в исходном наборе.

Операция композиции $\oplus$. Для обеспечения возможности «объединять» информацию, отвечающую отдельным наборам данных, на каноническом информационном пространстве должна быть определена операция композиции $\oplus$, обладающая следующими свойствами:

a) $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{b} \oplus \boldsymbol{a}$ – коммутативность. Комбинированная каноническая информация не должна зависеть от порядка поступления данных.

b) $(\boldsymbol{a} \oplus \boldsymbol{b}) \oplus \boldsymbol{c} = \boldsymbol{a} \oplus (\boldsymbol{b} \oplus \boldsymbol{c})$ – ассоциативность. Каноническая информация не должна зависеть от порядка комбинирования данных.

c) $\boldsymbol{a} \oplus \boldsymbol{0} = \boldsymbol{a}$ – свойство нейтрального элемента. Добавление к некоторой информация пустой информации не меняет эту информацию.

Таким образом, каноническое информационное пространство является коммутативным моноидом.

В приведенных выше примерах справедливость этих свойств сразу же следует из покомпонентного определения операции композиции, а именно, композиции двух элементов отвечает сумма их компонент (6).

Компактность. Информация, представленная в канонической форме, должна занимать небольшой (желательно минимальный) объем, по возможности, не зависящий от объема представленных данных.

Эффективность. Представление промежуточной информации в канонической форме должно обеспечивать эффективное выполнение всех стадий обработки данных:

a) Извлечение канонической информации из исходных данных;

b) Комбинирование и накопление канонической информации;

c) Вычисление результата из накопленной канонической информации.

Отметим, что свойства компактности и эффективности носят скорее технический характер, связанный с особенностями реализации соответствующих алгоритмов.

1.7 Выводы

Отметим, что чисто техническая попытка «распараллелить алгоритм», фактически привела нас к необходимости нахождения специального вида представления информации, обладающему удобными алгебраическими свойствами. В некотором смысле, такое представление отражает саму суть информации, содержащейся в данных. Можно сказать, что сама потребность эффективно манипулировать огромными распределенными массивами данных выдвигает новые требования к осмыслению и формализации понятия информации.

В рассмотренном выше примере выбор канонической формы информации довольно очевиден. В общем случае выбор компактной промежуточной информации может быть не очевиден или даже невозможен. В связи с этим, представляется важным выявление класса задач, в которых возможно выделение достаточно компактной промежуточной информации и нахождение эффективных методов построения подходящих информационных пространств.

В данной главе мы старались минимизировать формализм, чтобы акцентировать внимание на содержательной стороне проблемы. Мы наметили основные требования к хорошо организованной промежуточной информации. Это, в свою очередь, поднимает вопрос о выборе в некотором смысле оптимального (или идеального) вида промежуточной информации. Подобная проблематика потребует дальнейшей формализации и исследований.

2 Задача линейного оценивания и информация в системах больших данных

В данной главе рассмотрена проблема трансформации процедуры оптимального линейного оценивания так, чтобы отдельные фрагменты исходных данных могли обрабатываться независимо и параллельно. Предложена форма представления промежуточной информации, позволяющая алгоритму извлекать такую информацию параллельно из каждого набора исходных данных, объединять ее и использовать для получения результата. Показано, что на построенном информационном пространстве индуцируется упорядочение, отражающее понятие качества информации.

Выше мы рассмотрели в общих чертах специфику обработки информации в системах больших данных. В таких системах данные нередко собираются и хранятся распределённо и могут постоянно пополняться. Было показано, что для эффективной обработки таких распределенных данных ключевую роль играет возможность введения промежуточной формы представления информации, обладающей определенными алгебраическими свойствами. В данной главе мы исследуем задачу линейного оценивания с точки зрения распределенных систем сбора и обработки информации.

Подходы, развиваемые в этой статье, могут быть полезны для многих прикладных задач, например, при сборе и обработке информации в крупномасштабных экспериментах, где данные собираются в многочисленных исследовательских центрах, разбросанных по всему Земному шару. Однако для нас основной интерес представляют особенности информационных пространств, возникающих при необходимости распределенной обработки данных. Как мы увидим, на построенном

информационном пространстве естественным образом порождается алгебраическая структура, описывающая композицию отдельных фрагментов информации, и согласованное с ней отношение предпорядка, отражающее феномен качества информации.

2.1 Линейный эксперимент и задача линейного оценивания

Рассмотрим схему линейного измерения вида [2], [3]

$$y = Ax + \nu, \tag{7}$$

где $x \in \mathcal{D}$ – неизвестный вектор эвклидова пространства – объект измерения, $y \in \mathcal{R}$ - результат измерения, $A: \mathcal{D} \to \mathcal{R}$ - линейное отображение, описывающее искажения измерительной системы, и $\nu \in \mathcal{R}$ - случайный вектор шума с нулевым средним $\mathsf{E}\nu = 0$ и заданным ковариационным оператором $\mathsf{D}\nu = S$.

Ковариационный оператор случайного вектора $\mu \in \mathcal{R}$ является многомерным обобщением понятия дисперсии и определяется как $(\mathsf{D}\mu)(z) = \mathsf{E}\langle \mu - \mathsf{E}\mu, z\rangle(\mu - \mathsf{E}\mu)$[1]. Несложно проверить, что ковариационный оператор случайного вектора μ является самосопряженным неотрицательно определенным оператором и его матрица в ортонормированном базисе представляет собой ковариационную матрицу координат вектора μ в этом базисе.

Таким образом, вся информация об измерении — это модель измерения, описываемая парой (A, S) и результат измерения y. Будем рассматривать здесь лишь измерения, в которых оператор S положительно определен, $S > 0$ и, следовательно, обратим. По сути, это означает, что шум ν возможен во всех направлениях, т.е., не существует собственного подпространства $\tilde{\mathcal{R}} \subset \mathcal{R}$ такого, что $\nu \in \tilde{\mathcal{R}}$ с вероятностью единица.

[1] Здесь и далее $\langle \cdot, \cdot \rangle$ обозначает скалярное произведение.

Приведем здесь кратко постановку и решение задачи линейного оценивания. Более детальное и общее рассмотрение можно найти в [2], [3].

Задача линейного оценивания неизвестного вектора x состоит в построении такого линейного отображения $R\colon \mathcal{R} \to \mathcal{D}$, что оценка $\hat{x} = Ry$ максимально близка к x. Формально, рассмотрим погрешность оценки $\mathsf{E}\|Ry - x\|^2$

$$\begin{aligned}\mathsf{E}\|Ry - x\|^2 &= \mathsf{E}\|R(Ax + v) - x\|^2 \\ &= \|(RA - I)x\|^2 + 2\mathsf{E}\langle (RA - I)x, Rv\rangle + \mathsf{E}\|Rv\|^2 \\ &= \|(RA - I)x\|^2 + \operatorname{tr} RSR^*.\end{aligned}$$

В последнем равенстве мы воспользовались тем, что $\mathsf{E}v = 0$ и следующими свойствами: $\mathsf{D}(Rv) = RSR^*$ и $\mathsf{E}\|\mu\|^2 = \operatorname{tr}\mathsf{D}\mu$ для случайного вектора μ с нулевым средним.

Поскольку в выражении для $\mathsf{E}\|Ry - x\|^2$ присутствует неизвестный вектор x, определим погрешность оценивания, обеспечиваемую оператором R, как

$$H(R) = \sup_{x \in \mathcal{D}} \mathsf{E}\|Ry - x\|^2.$$

Легко видеть, что если $RA \neq I$ то $\|(RA - I)x\|^2$ может принимать сколь угодно большие значения и, следовательно,

$$H(R) = \begin{cases} +\infty, & \text{если } RA \neq I, \\ \operatorname{tr} RSR^*, & \text{если } RA = I. \end{cases}$$

Таким образом, линейное отображение R обеспечивает конечную погрешность оценивания $H(R)$ тогда и только тогда, когда $RA = I$, что, как легко убедиться, равносильно несмещённости оценки $\hat{x} = Ry$, т.е. $\mathsf{E}Ry = x$. С другой стороны, существование R, для которого $RA = I$, равносильно тому, что ядро линейного отображения A тривиально: $\mathcal{N}(A) = \{0\}$[2].

Итак, задача линейного оценивания — это задача условной минимизации:

[2] Будем обозначать ядро и образ линейного отображения $A\colon \mathcal{D} \to \mathcal{R}$, соответственно $\mathcal{N}(A) \subseteq \mathcal{D}$ и $\mathcal{R}(A) \subseteq \mathcal{R}$.

$$\min_R\{\mathrm{tr}RSR^* \mid RA = I\},$$

разрешимая лишь если $\mathcal{N}(A) = \{0\}$.

Пусть $\mathbb{S}_{\mathcal{D}}$ – пространство всех самосопряженных операторов на $\mathcal{D}$. Определим частичный порядок на $\mathbb{S}_{\mathcal{D}}$ следующим образом:

$$P \geq Q \;\Leftrightarrow\; P - Q \geq 0.$$

Заметим, что tr является строго монотонным отображением из пространства самосопряженных операторов в вещественную прямую, а именно, если $P \geq Q$ то $\mathrm{tr}P \geq \mathrm{tr}Q$, а если, кроме того, $P \neq Q$ то $\mathrm{tr}P > \mathrm{tr}Q$. Поэтому, рассмотрим задачу минимизации самого оператора RSR^* при условии $RA = I$. Докажем, что если $\mathcal{N}(A) = \{0\}$, то существует единственное линейное отображение R, доставляющее минимум оператору RSR^*.

Условие $\mathcal{N}(A) = \{0\}$ равносильно тому, что оператор $A^*S^{-1}A$ обратим. Очевидно, что $R = (A^*S^{-1}A)^{-1}A^*S^{-1}$ удовлетворяет условию $RA = I$. Докажем, что среди всех линейных отображений $\tilde{R}$ таких, что $\tilde{R}A = I$, оператор $\tilde{R}S\tilde{R}^*$ достигает минимума при $\tilde{R} = R$. Пусть $C = \tilde{R} - R$, тогда $CA = \big(\tilde{R} - R\big)A = 0$. Отсюда имеем:

$$\begin{aligned}\tilde{R}S\tilde{R}^* &= [(A^*S^{-1}A)^{-1}A^*S^{-1} + C]S[S^{-1}A(A^*S^{-1}A)^{-1} + C^*] \\ &= (A^*S^{-1}A)^{-1} + (A^*S^{-1}A)^{-1}(AC)^* + AC(A^*S^{-1}A)^{-1} + CSC^* \\ &= (A^*S^{-1}A)^{-1} + CSC^*.\end{aligned}$$

Оператор $CSC^* \geq 0$, а из невырожденности S сразу следует, что если $C \neq 0$, то и $CSC^* \neq 0$. Следовательно, $\tilde{R}S\tilde{R}^*$ и $\mathsf{E}\big\|\tilde{R}y - x\big\|^2 = \mathrm{tr}\,\tilde{R}S\tilde{R}^*$ достигают минимальных значений в единственной точке $\tilde{R} = R$.

Таким образам, задача оптимального оценивания вида

$$\min_{R:\mathcal{R}\to\mathcal{D}} \sup_{x\in\mathcal{D}} \mathsf{E}\|Ry - x\|^2$$

имеет решение тогда и только тогда, когда $\mathcal{N}(A) = \{0\}$. При этом оценка

$$\hat{x} = Ry = (A^*S^{-1}A)^{-1}A^*S^{-1}y$$

является несмещенной оценкой вектора x, обладающей наименьшим ковариационным оператором

$$\mathsf{D}\hat{x} = RSR^* = (A^*S^{-1}A)^{-1}.$$

Отсюда, в частности, следует, что оценка $\hat{x} = Ry$ обладает минимальными дисперсиями координат $\hat{x}_j$ в некотором ортонормированном базисе:

$$\mathsf{D}\hat{x}_j = ((A^*S^{-1}A)^{-1})_{jj}$$

и $\mathsf{E}\|\hat{x} - x\|^2$ достигает минимального значения

$$\mathsf{E}\|\hat{x} - x\|^2 = \mathrm{tr}(A^*S^{-1}A)^{-1}.$$

Итак, пусть заданы результат измерения y и модель измерения (A, S). Тогда исходные данные для линейного оценивания представляются тройкой (y, A, S) и процедура обработки P состоит в преобразовании исходных данных в результат оценивания: $\hat{x}$ - оптимальную оценку вектора x:

$$(y, A, S) \xrightarrow{P} \hat{x} = (A^*S^{-1}A)^{-1}A^*S^{-1}y.$$

При этом отображение P определено не всюду, а лишь тогда, когда оператор $A^*S^{-1}A$ обратим.

2.2 Линейное оценивание в случае многих независимых измерений

Пусть теперь имеется много независимых измерений одного и того же неизвестного вектора $x \in \mathcal{D}$:

$$y_i = A_i x + \nu_i, \qquad i = 1, \dots, n, \tag{8}$$

где $y_i \in \mathcal{R}_i$ – результаты измерений, $A_i : \mathcal{D} \to \mathcal{R}_i$ – линейные отображения, и $\nu_i \in \mathcal{R}_i$ – независимые случайные векторы с нулевыми средними $\mathsf{E}\nu_i = 0$ и ковариационными операторами $\mathsf{D}\nu_i = S_i : \mathcal{R}_i \to \mathcal{R}_i$. В общем случае пространства измерений $\mathcal{R}_i$ могут быть различными.

Чтобы воспользоваться результатом предыдущего раздела в случае серии измерений (8), представим эту серию в виде одного измерения вида (7). Рассмотрим сначала случай двух независимых измерений (8), т.е., $n = 2$.

Пару (y_1, y_2) векторов $y_1 \in \mathcal{R}_1$ и $y_2 \in \mathcal{R}_2$ можно рассматривать как элемент произведения эвклидовых пространств $\mathcal{R}_1 \times \mathcal{R}_2$. Удобно записывать такую пару в виде столбца $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \mathcal{R}_1 \times \mathcal{R}_2$. Линейные операции на $\mathcal{R}_1 \times \mathcal{R}_2$ определяются через операции на $\mathcal{R}_1$ и $\mathcal{R}_2$ покомпонентно, а скалярное произведение как $\langle \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \rangle = \langle y_1, z_1 \rangle + \langle y_2, z_2 \rangle$ для любых $y_1, z_1 \in \mathcal{R}_1$ и $y_2, z_2 \in \mathcal{R}_2$.

Пара линейных отображений $A_1 : \mathcal{D} \to \mathcal{R}_1$ и $A_2 : \mathcal{D} \to \mathcal{R}_2$, действующих из одного и того же пространства $\mathcal{D}$, взаимно однозначно определяются линейным отображением $\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} : \mathcal{D} \to \mathcal{R}_1 \times \mathcal{R}_2$, действующим по правилу $\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} x = \begin{pmatrix} A_1 x \\ A_2 x \end{pmatrix}$ для любого $x \in \mathcal{D}$. Аналогично, пара линейных отображений $B_1 : \mathcal{R}_1 \to Q$ и $B_2 : \mathcal{R}_2 \to Q$, действующих в одно и то же пространство Q взаимно однозначно определяются линейным отображением $(B_1 \quad B_2) : \mathcal{R}_1 \times \mathcal{R}_2 \to Q$, действующим по правилу $(B_1 \quad B_2) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = B_1 y_1 + B_2 y_2$ для любых $y_1 \in \mathcal{R}_1$ и $y_2 \in \mathcal{R}_2$. Манипулирование такими матрицами подчиняется обычным правилам матричной алгебры. Заметим, что сопряженной к матрице из линейных преобразований будет транспонированная матрица с сопряженными компонентами. Например, если $\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} : \mathcal{D} \to \mathcal{R}_1 \times \mathcal{R}_2$ то $\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}^* = (A_1^* \quad A_2^*) : \mathcal{R}_1 \times \mathcal{R}_2 \to \mathcal{D}$.

Наконец, для ковариационного оператора вектора $\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \in \mathcal{R}_1 \times \mathcal{R}_2$ с независимыми компонентами ν_1 и ν_2 имеем

$$\left(\mathsf{D} \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \right) \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \mathsf{E} \langle \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \rangle \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} = \mathsf{E}(\langle \nu_1, z_1 \rangle + \langle \nu_2, z_2 \rangle) \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}$$

$$= \mathsf{E} \begin{pmatrix} \langle \nu_1, z_1 \rangle \nu_1 + \langle \nu_2, z_2 \rangle \nu_1 \\ \langle \nu_1, z_1 \rangle \nu_2 + \langle \nu_2, z_2 \rangle \nu_2 \end{pmatrix} = \begin{pmatrix} (\mathsf{D}\nu_1) z_1 + 0 \\ 0 + (\mathsf{D}\nu_2) z_2 \end{pmatrix}$$

$$= \begin{pmatrix} \mathsf{D}\nu_1 & 0 \\ 0 & \mathsf{D}\nu_2 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

для любого $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \in \mathcal{R}_1 \times \mathcal{R}_2$. Следовательно, $\mathsf{D}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} = \begin{pmatrix} \mathsf{D}\nu_1 & 0 \\ 0 & \mathsf{D}\nu_2 \end{pmatrix} = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}$. Здесь мы воспользовались тем, что, в силу независимости ν_1 и ν_2 $\mathsf{E}\langle \nu_1, z_1 \rangle \nu_2 = \langle \mathsf{E}\nu_1, z_1 \rangle \mathsf{E}\nu_2 = 0$ и $\mathsf{E}\langle \nu_2, z_2 \rangle \nu_1 = 0$.

Таким образом, объединение исходных данных, отвечающих двум независимым измерениям, описывается операцией

$$(y_1, A_1, S_1) \cup (y_2, A_2, S_2) = \left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \right). \tag{9}$$

При наличии n независимых измерений (8) потребуется собрать соответствующие данные в одном месте, реорганизовать их в виде блочных матриц, возможно, очень больших размерностей и применить к объединенным данным отображение P (Рис. 12).

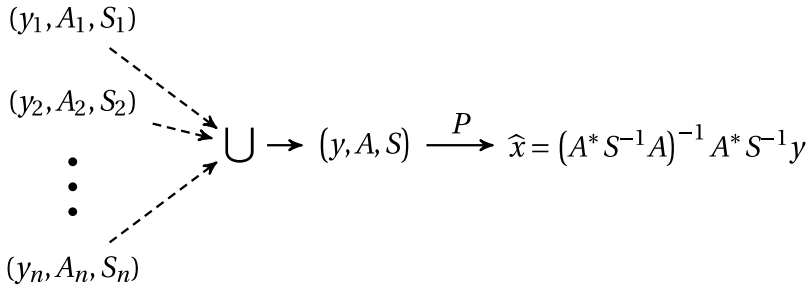

Рис. 12. Стандартная схема линейного оценивания для большого числа измерений.

Здесь

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathcal{R}, \qquad A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{pmatrix} : \mathcal{D} \to \mathcal{R}, \qquad S = \begin{pmatrix} S_1 & 0 & \cdots & 0 \\ 0 & S_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_n \end{pmatrix} : \mathcal{R} \to \mathcal{R},$$

$$\mathcal{R} = \mathcal{R}_1 \times \mathcal{R}_2 \times \cdots \times \mathcal{R}_n, \qquad \dim \mathcal{R} = \sum_{i=1}^{n} \dim \mathcal{R}_i.$$

При большом числе измерений размерность объединенных данных может стать крайне большой, в результате чего данный подход может оказаться практически нереализуемым. Кроме того, добавление новых данных будет приводить к увеличению размерностей объединенных данных, что, в свою очередь, будет требовать все больше ресурсов для их хранения и обработки (применения преобразования P).

2.3 Распараллеливание обработки за счет выделения промежуточной информации

Покажем, что обработку данных в задаче линейного оценивания можно разбить на две фазы $P = P_2 \circ P_1$, где первая фаза P_1 состоит в выделении некоторой компактной промежуточной информации из исходных данных, а вторая P_2 вычисляет результат оценивания на основании этой промежуточной информации. При этом, нашей целью будет найти такую факторизацию, что применение преобразования P_1 к объединенному набору данных может быть заменено параллельным применением P_1 к отдельным данным и последующему «сложению» полученных фрагментов информации.

Пусть (y_1, A_1, S_1) и (y_2, A_2, S_2) два набора данных. Рассмотрим результат оценивания, отвечающий объединённому набору данных (y, A, S), где $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$, $A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$, и $S = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}$. Заметим, что для вычисления оптимальной оценки вектора x $\hat{x} = (A^* S^{-1} A)^{-1} A^* S^{-1} y$ требуется вычислять выражения вида $A^* S^{-1} A$ и $A^* S^{-1} y$. Рассмотрим их для случая двух объединенных измерений.

Используя очевидное равенство $\begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}^{-1} = \begin{pmatrix} S_1^{-1} & 0 \\ 0 & S_2^{-1} \end{pmatrix}$, получаем

$$A^* S^{-1} A = \begin{pmatrix} A_1^* & A_2^* \end{pmatrix} \begin{pmatrix} S_1^{-1} & 0 \\ 0 & S_2^{-1} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} A_1^* S_1^{-1} & A_2^* S_2^{-1} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$$
$$= A_1^* S_1^{-1} A_1 + A_2^* S_2^{-1} A_2$$

и аналогично,

$$A^*S^{-1}y = A_1^*S_1^{-1}y_1 + A_2^*S_2^{-1}y_2.$$

Это означает, что вся необходимая для дальнейшей обработки информация, относящаяся к i-му измерению может быть представлена парой (v_i, T_i), где

$$v_i = A_i^*S_i^{-1}y_i \in \mathcal{D}, \quad T_i = A_i^*S_i^{-1}A_i : \mathcal{D} \to \mathcal{D},$$

и T_i – неотрицательно определенный оператор. При этом, объединенным данным будет отвечать пара (v, T), в которой $v = v_1 + v_2$ и $T = T_1 + T_2$.

Будем называть пару $(v, T) = (A^*S^{-1}y,\ A^*S^{-1}A)$ **канонической информацией** для данных (y, A, S), а множество $\mathfrak{I}$ всех таких пар каноническим **информационным пространством** для задачи линейного оценивания вектора из пространства $\mathcal{D}$. Заметим, что $\mathcal{R}(A^*S^{-1}) = \mathcal{R}(A^*S^{-1}A) = \mathcal{N}^{\perp}(A)$ [2], [3]. Следовательно, измерениям вида (y, A, S) могу отвечать лишь такие пары (v, T), в которых $v \in \mathcal{R}(T)$. Таким образом,

$$\mathfrak{I} = \{(v, T) \mid T \in \mathbb{S}_{\mathcal{D}}^+, v \in \mathcal{R}(T)\}$$

где $\mathbb{S}_{\mathcal{D}}^+$ - множество неотрицательно определенных операторов на $\mathcal{D}$ – выпуклый конус в линейном пространстве $\mathbb{S}_{\mathcal{D}}$ самосопряженных операторов на пространстве $\mathcal{D}$. Если $\dim \mathcal{D} = m$ то $\dim \mathbb{S}_{\mathcal{D}} = \frac{m(m+1)}{2}$. Тогда $\mathfrak{I} \subset \mathcal{D} \times \mathbb{S}_{\mathcal{D}}^+$ представляет собой выпуклый конус в $\frac{m(m+3)}{2}$-мерном линейном пространстве $\mathcal{D} \times \mathbb{S}_{\mathcal{D}}$. Отсюда, в частности, следует, что любой элемент информационного пространства $\mathfrak{I}$ может быть задан $\frac{m(m+3)}{2}$ числами.

Очевидно, процесс линейного оценивания можно разбить на две фазы $P = P_2 \circ P_1$, где первая фаза P_1 состоит в построении канонической информации:

$$(v, T) = P_1(y, A, S) = (A^*S^{-1}y, A^*S^{-1}A),$$

а вторая P_2 вычисляет результат оценивания на основании этой информации (Рис. 13):

$$\hat{x} = P_2(v, T) = T^{-1}v.$$

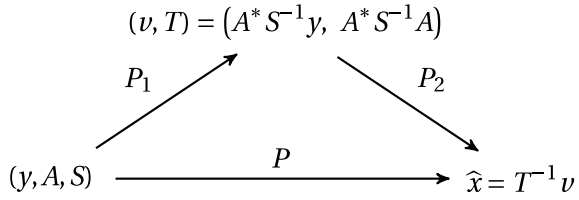

Рис. 13. Разбиение процесса обработки данных на две фазы.

Как было показано выше, объединению исходных данных (y_1, A_1, S_1) и (y_2, A_2, S_2) отвечает композиция соответствующих элементов канонической информации (v_1, T_1) и (v_2, T_2), определенная как

$$(v_1, T_1) \oplus (v_2, T_2) = (v_1 + v_2, T_1 + T_2).$$

Это можно записать как $P_1(y_1, A_1, S_1) \otimes P_1(y_2, A_2, S_2) = P_1((y_1, A_1, S_1) \cup (y_2, A_2, S_2))$, где под $(y_1, A_1, S_1) \cup (y_2, A_2, S_2)$ понимается определяемое выражением (2), объединение двух наборов данных в один, Рис. 14.

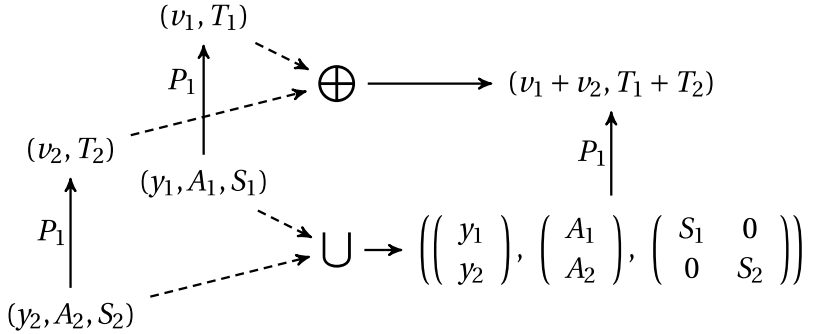

Рис. 14. Соответствие композиции фрагментов канонической информации и объединения наборов исходных данных.

В результате факторизации алгоритма P на две фазы и введения канонической информации, схема обработки распределенных данных, представленная на Рис. 12 может быть трансформирована следующим образом (Рис. 15). Из каждого отдельного фрагмента (y_i, A_i, S_i) данных

выделяется каноническая информация (v_i, T_i), которая впоследствии объединяется и используется для вычисления результата оценивания.

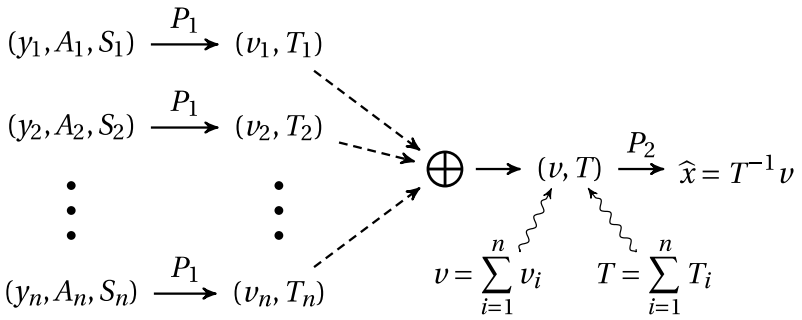

Рис. 15. Модифицированная схема обработки распределенных данных.

Отметим основные особенности такой модифицированной схемы:

a) Объем памяти, требуемый для хранения информации в каноническом виде не зависит от объема представляемых исходных данных и составляет $\frac{m(m+3)}{2}$ чисел (m-мерный вектор и самосопряженный оператор в -мерном пространстве).

b) Выделение канонической информации (v_i, T_i) из -того набора данных (преобразование P_1) может проводиться на тех компьютерах, где эти данные находятся, причем, параллельно и независимо.

c) Передаются лишь компактные фрагменты выделенной канонической информации одинакового объема.

d) Сложение частей канонической информации максимально упрощено и определяется покомпонентным сложением пар (v_i, T_i).

e) Ресурсоемкость второй фазы P_2, состоящей в построении результата по компактной накопленной информации (v, T), определяется только размерностью m пространства неизвестных и не зависит от объема исходных данных.

f) По мере поступления новых данных, потребуется лишь выделять из них каноническую информацию и «добавлять» ее к накопленной. При этом окончательную обработку P_2 будет необходимо снова применять к компактной информации фиксированного объема.

В результате, распределенность исходных данных способствует повышению эффективности обработки за счет естественного распараллеливания алгоритма.

В рассмотренной выше задаче линейного оценивания нашей целью было построение оценки $\hat{x}$, т.е., $P(y, A, S) = \hat{x}$. При построении оценки $\hat{x}$ важно также охарактеризовать ее точность, исчерпывающую информацию о которой содержит ковариационный оператор

$$Q = \mathsf{D}\hat{x} = (A^* S^{-1} A)^{-1} = T^{-1}.$$

В частности, он позволяет определить погрешности оценивания отдельных компонент вектора x (в произвольном ортонормированном базисе)

$$\mathsf{E}\left(\hat{x}_j - x_j\right)^2 = \mathsf{D}\hat{x}_j = Q_{jj}$$

и полную погрешность оценивания

$$\mathsf{E}\|\hat{x} - x\|^2 = \mathrm{tr} Q = \sum_{j=1}^{m} Q_{jj}.$$

Таким образом, каноническая информация вида (v, T) подходит также для построения результатов оценивания $P(y, A, S)$ в виде

$$P(y, A, S) = (\hat{x}, \mathsf{D}\hat{x}),$$

$$P'(y, A, S) = \left(\hat{x}, \left\{\mathsf{D}\hat{x}_j\right\}_{j=1,\dots,m}\right)$$

или

$$P''(y, A, S) = (\hat{x}, \mathsf{E}\|\hat{x} - x\|^2).$$

В этих случаях отображение P_1 остается прежним, а отображение P_2 заменяется, соответственно, на

$$P_2(v, T) = (T^{-1} v, T^{-1}),$$

$$P'_2(v, T) = \left(T^{-1} v, \left\{(T^{-1})_j\right\}_{j=1,\dots,m}\right)$$

или

$$P''_2(v,T) = (T^{-1}v, \mathrm{tr}T^{-1})$$

соответственно.

2.4 Качество информации и информативность линейного эксперимента

Как было отмечено выше, ковариационный оператор $Q = \mathsf{D}\hat{x} = (A^*S^{-1}A)^{-1} = T^{-1}$ полностью характеризует точность оценивания. При этом, чем меньше ковариационный оператор оценки $\hat{x}$ тем меньше погрешность оценивания: т.е., если $Q \le \tilde{Q}$ то $Q_{jj} \le \tilde{Q}_{jj}$ и $\mathrm{tr}Q \le \mathrm{tr}\tilde{Q}$.

Будем говорить, что информация (v,T) не хуже (не менее точна), чем $(\tilde{v},\tilde{T})$ если $T \ge \tilde{T}$ и писать $(v,T) \succcurlyeq (\tilde{v},\tilde{T})$. Если $(v,T) \succcurlyeq (\tilde{v},\tilde{T})$ и $(\tilde{v},\tilde{T}) \succcurlyeq (v,T)$, то будем говорить, что (v,T) и $(\tilde{v},\tilde{T})$ имеют одинаковую точность и обозначать это $(v,T) \approx (\tilde{v},\tilde{T})$. Очевидно, это равносильно условию $T = \tilde{T}$. Нетрудно видеть, что более точная информация обеспечивает более точное оценивание. Действительно, пусть $T \ge \tilde{T}$ и (v,T) и $(\tilde{v},\tilde{T})$ позволяют построить соответствующие оценки, т.е., T и $\tilde{T}$ обратимы. Согласно [2] отсюда следует, что $T^{-1} \le \tilde{T}^{-1}$ и значит $Q \le \tilde{Q}$, где Q и $\tilde{Q}$ – ковариационные операторы соответствующих оценок.

Как мы уже отмечали, построение оценки вектора x возможно лишь если $\mathcal{N}(A) = \{0\}$. Если $\mathcal{N}(A) \ne \{0\}$ то часть вектора x, принадлежащая $\mathcal{N}(A)$, обнуляется и не может быть оценена. Однако, как показано в [2], [3] можно оценить проекцию Px вектора x на подпространство $\mathcal{N}^{\perp}(A)$. Здесь $P = A^{-}A: \mathcal{D} \to \mathcal{D}$ – ортогональный проектор на $\mathcal{N}^{\perp}(A)$, а $A^{-}: \mathcal{R} \to \mathcal{D}$ – линейное отображение, псевдообратное к $A: \mathcal{D} \to \mathcal{R}$ [2]. При этом оптимальная оценка вектора Px и ее ковариационный оператор определяются выражениями

$$\widehat{Px} = (A^*S^{-1}A)^{-}A^*S^{-1}y, \quad \mathsf{D}(\widehat{Px}) = (A^*S^{-1}A)^{-},$$

или, в терминах канонической информации, $\widehat{Px} = T^{-}v, \quad \mathsf{D}(\widehat{Px}) = T^{-}$. Более того, поскольку $\mathcal{N}(T) = \mathcal{N}(A^{*}S^{-1}A) = \mathcal{N}(A)$, проектор P также может быть выражен через T, $P = T^{-}T$.

Таким образом, рассматриваемая нами каноническая информация содержит всю информацию, необходимую для построения оптимальной оценки и в этом более широком контексте.

Пусть информация (v, T) не хуже, чем $(\tilde{v}, \tilde{T})$, то есть $T \geq \tilde{T} \geq 0$. Тогда $\mathcal{N}(T) \subseteq \mathcal{N}(\tilde{T})$ и $\tilde{T}^{-} \geq \tilde{P}T^{-}\tilde{P}$ [2], где $\tilde{P} = \tilde{T}^{-}\tilde{T}$ - ортогональный проектор на $\mathcal{N}^{\perp}(\tilde{T})$. Включение $\mathcal{N}(T) \subseteq \mathcal{N}(\tilde{T})$ влечет $\mathcal{N}^{\perp}(\tilde{T}) \subseteq \mathcal{N}^{\perp}(T)$, т.е., информация (v, T) позволяет оценить большую часть вектора x, чем $(\tilde{v}, \tilde{T})$, а неравенство $\tilde{T}^{-} \geq \tilde{P}T^{-}\tilde{P}$ означает, что информация (v, T) обеспечивает более точное оценивание вектора $\tilde{P}x$. Иными словами, более точная информация обеспечивает более широкие возможности оценивания и при прочих равных условиях менее интенсивный шум оценки.

Отметим, что рассмотренное выше понятия точности информации приводит к такому же упорядочению на множестве моделей линейного измерения, как и понятие качества моделей измерений в [3], [4] или информативности преобразователей информации в [5].

2.5 Свойства канонической информации в задаче линейного оценивания

Рассмотрим свойства канонического информационного пространства $\mathfrak{I}$, определенного выше. Эти свойства не только представляют самостоятельный интерес, но и могут выступать в качестве примера общих свойств информационных пространств, возникающих в задачах обработки больших объемов распределенных данных.

Существование для любого исходного набора данных. Как мы видели, информации, содержащейся в данных (y, A, S) может быть недостаточно для

построения результата оценивания. А именно, если ядро отображения A нетривиально, т.е., $\mathcal{N}(A) \neq \{0\}$, то оценка неизвестного вектора не может быть построена. Тем не менее, каноническая информация (v, T) может быть построена для любых исходных данных. Отметим, что даже полное отсутствие измерений (несущее нулевую информацию) может быть представлено в каноническом виде. Формально, любое измерение (y, A, S), в котором $A = 0: \mathcal{D} \to \mathcal{R}$ - нулевое отображение, не несет никакой информации об измеряемом векторе. Любому такому измерению отвечает каноническая информация $\mathbf{0} = (0,0)$, т.е. $v = 0 \in \mathcal{D}$ и $T = 0: \mathcal{D} \to \mathcal{D}$.

Полнота (или **достаточность**). Каноническая форма содержит всю информацию, содержащуюся в исходных данных, а именно, она приводит к тому же результату, что и исходные данные, из которых она получена. Формально это означает что $P(y, A, S) = P_2(P_1(y, A, S))$ для всех данных (y, A, S) из области определения преобразования P. Это свойство напоминает понятие достаточности в математической статистике.

Операция композиции $\oplus$. На каноническом информационном пространстве $\mathfrak{I}$ определена операция композиции $\oplus$, описывающая сложение фрагментов информации, отвечающих данным. При этом $(\mathfrak{I}, \oplus, \mathbf{0})$ является коммутативным моноидом, т.е., выполнены следующие свойства для любых $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathfrak{I}$:

a) $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{b} \oplus \boldsymbol{a}$.

b) $(\boldsymbol{a} \oplus \boldsymbol{b}) \oplus \boldsymbol{c} = \boldsymbol{a} \oplus (\boldsymbol{b} \oplus \boldsymbol{c})$.

c) $\boldsymbol{a} \oplus \mathbf{0} = \boldsymbol{a}$.

Отметим, что моноид $(\mathfrak{I}, \oplus, \mathbf{0})$ обладает также свойством сокращения:

d) $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{a} \oplus \boldsymbol{c} \Rightarrow \boldsymbol{b} = \boldsymbol{c}$,

но не имеет обратимых элементов отличных от $\mathbf{0}$, т.е. не существует «отрицательной» информации.

Предпорядок $\succcurlyeq$, отражающий понятие точности информации. На каноническом информационном пространстве $\mathfrak{I}$ определено отношение $\succcurlyeq$, обладающее следующими свойствами:

e) $\boldsymbol{a} \succcurlyeq \boldsymbol{a}$ (рефлексивность).

f) $\boldsymbol{a} \succcurlyeq \boldsymbol{b}$ & $\boldsymbol{b} \succcurlyeq \boldsymbol{c} \Rightarrow \boldsymbol{a} \succcurlyeq \boldsymbol{c}$ (транзитивность).

Заметим, что отношение $\succcurlyeq$ не обладает свойством антисимметричности, т.е. не является частичным порядком. Действительно, из $(v, T) \succcurlyeq (\tilde{v}, \tilde{T})$ и $(\tilde{v}, \tilde{T}) \succcurlyeq (v, T)$ следует лишь, что $T = \tilde{T}$, но не обязательно $v = \tilde{v}$. Однако, на классах эквивалентной точности это отношение антисимметрично, $\boldsymbol{a} \succcurlyeq \boldsymbol{b}$ & $\boldsymbol{b} \succcurlyeq \boldsymbol{a} \Rightarrow \boldsymbol{a} \approx \boldsymbol{b}$ и, следовательно, является отношением частичного порядка.

Кроме того, алгебраическая структура информационного пространства согласована со структурой порядка, а именно:

g) $\boldsymbol{a} \succcurlyeq \boldsymbol{0}$. Любая информация точнее, чем отсутствие информации.

h) $\boldsymbol{a} \oplus \boldsymbol{b} \succcurlyeq \boldsymbol{a}, \boldsymbol{b}$. Композиция двух фрагментов информации точнее, чем каждый из них по отдельности.

i) $\boldsymbol{a} \succcurlyeq \boldsymbol{b}$ & $\boldsymbol{c} \succcurlyeq \boldsymbol{e} \Rightarrow \boldsymbol{a} \oplus \boldsymbol{c} \succcurlyeq \boldsymbol{b} \oplus \boldsymbol{e}$. Чем точнее фрагменты информации, тем точнее результат композиции.

Единственность представления данных в каноническом виде. Несложно убедиться, что различные данные (y, A, S) могут описываться одной и той же канонической информацией (v, T) и, как следствие, приводить к одному и тому же результату оценивания. Может показаться, что данные (y, A, S) могут быть представлены разными парами вида (v, T). Действительно, например, $(2v, 2T)$ приведет к тому же самому результату $\hat{x} = T^{-1}v$, что и пара (v, T). Однако, пары (v, T) и $(2v, 2T)$ будут вести себя по-разному при композиции с другими элементами информационного пространства и, в конечном итоге, будут приводить к разным результатам. Таким образом, для каждого элемента исходных данных (y, A, S) существует единственное представление элементом пространства $\mathfrak{I}$, согласованное с

операцией композиции и обеспечивающего соответствующий результат оценивания.

В частности, поскольку результат оценивания не зависит от порядка данных в исходном наборе, каноническая информация не должна зависеть от порядка данных. Это имеет место для информационного пространства $\mathfrak{I}$.

Отметим, наконец, два «практических» свойства данного способа представления промежуточной информации. Они носят скорее технический характер, связанный с особенностями реализации соответствующих алгоритмов.

Компактность. Информация, представленная в канонической форме, занимает фиксированный объем $\frac{m(m+3)}{2}$ чисел, не зависящий от объема представленных данных.

Эффективность. Представление промежуточной информации в канонической форме обеспечивает эффективное выполнение всех стадий обработки данных:

a) Извлечение канонической информации из исходных данных требует несколько матричных умножений для матриц, определяемых отдельными фрагментами данных. При этом извлечение канонической информации из отдельных фрагментов может производиться параллельно.
b) Комбинирование и накопление канонической информации сводится к сложению векторов и матриц фиксированной размерности и требует незначительных вычислительных ресурсов.
c) Вычисление результата на основании накопленной канонической информации требует решения системы линейных уравнений фиксированного размера $m \times m$ (или обращения соответствующей матрицы. Даже при постоянном поступлении новых данных обновление оценки может осуществляться лишь время от времени.

2.6 Заключение

Несмотря на то, что задача линейного оценивания имеет самостоятельную ценность и часто встречается в приложениях, мы использовали ее, в первую очередь, в качестве иллюстрации. Мы показали, что проблема адаптации алгоритма к работе в системах больших данных приводит к построению специального вида представления информации, обладающему естественными алгебраическими свойствами.

Во многих практических задачах преобразование обработки P, трансформирующее исходные данные в окончательный результат обработки, имеет специфическое «происхождение», а именно, является решением некоторой оптимизационной задачи. В нашем случае рассматривалась задача построения оценки минимальной погрешности. Оптимизационная постановка исходной задачи, фактически, привела к тому, что понятие качества решения (точности оценки) индуцировало на информационном пространстве упорядочение, отражающее «качество» информации.

3 Переход от априорной информации к апостериорной в распределенных системах обработки данных

Рассматривается процедура перехода от априорной к апостериорной информации для линейного эксперимента в контексте систем больших данных. Такой процесс носит, на первый взгляд, принципиально последовательный характер. А именно, в результате наблюдения, априорная информация трансформируется в апостериорную, которая впоследствии трактуется как априорная для следующего наблюдения, и т.д. Как показано ниже, данная процедура может быть распараллелена и унифицирована за счет преобразования как результатов измерений, так и исходной априорной информации к некоторому специальному виду. Исследуются и сравниваются свойства различных форм представления информации. Рассматриваемый подход позволяет эффективно масштабировать процедуру байесовского оценивания и, таким образом, адаптировать ее к проблемам обработки больших объемов распределённых данных.

Проблема линейного оценивания с априорной информацией, тесно связана с байесовским переходом от априорного распределения к апостериорному в математической статистике [6], [1], [7] и совпадает с байесовским переходом для нормальных распределений. Данная задача представляет интересную возможность исследовать и сравнить различные формы представления информации, такие как: исходная «сырая» информация; максимально удобная для интерпретации «явная» информация; и, наконец, специальная «каноническая» информация, максимально удобная для промежуточных манипуляций с информацией (таких как, слияние, обновление, передача, хранение и т.п.). Будет показано, что все эти три способа представления информации приводят к информационным

пространствам, обладающих определенными алгебраическими свойствами, исследованы свойства этих пространств и связи между ними. Основной интерес для нас представляют особенности этих пространств, в контексте распределенной обработки больших объемов данных.

Заметим, что байесовская процедура последовательного обновления информации, считается одним из важнейших инструментов в экспертных системах. Особый интерес к различным вариантам этой процедуры наблюдается в контексте Больших Данных, поскольку она позволяет обновлять информацию об объекте исследования по мере поступления данных, в результате чего отпадает необходимость накопления и хранения самих исходных данных. Как будет показано ниже, выбор адекватного канонического информационного пространства позволяет существенно повысить эффективность процесса обработки данных за счет унификации и минимизации вычислений. Более того, в последнее время, в проблематике Больших Данных, особое внимание уделяется методам анализа данных, допускающим параллельную и распределенную обработку. Ниже мы увидим, что введение подходящей промежуточной формы представления информации открывает возможность гибкого распараллеливания и масштабирования процедуры обновления информации в распределенных системах обработки данных.

3.1 Линейное оценивание с априорной информацией

Приведем здесь кратко постановку и решение задачи линейного оценивания с априорной информацией. Более детальное и общее рассмотрение можно найти в [2], [3], [8].

3.1.1 Линейное измерение

Рассмотрим схему линейного измерения вида

$$y = Ax + \nu, \tag{10}$$

где $x \in \mathcal{D}$ – объект измерения – вектор евклидова пространства, $y \in \mathcal{R}$ – результат измерения, $\mathcal{R}$ – пространство результатов измерения, $A: \mathcal{D} \to \mathcal{R}$ – линейное отображение, описывающее искажения измерительной системы, и $\nu \in \mathcal{R}$ – случайный вектор шума с нулевым средним $\mathsf{E}\nu = 0$ и заданным ковариационным оператором $\mathsf{D}\nu = S: \mathcal{R} \to \mathcal{R}$.

Ковариационный оператор случайного вектора $\mu \in \mathcal{R}$ является многомерным обобщением понятия дисперсии и определяется как

$$(\mathsf{D}\mu)(z) = \mathsf{E}\langle \mu - \mathsf{E}\mu, z\rangle(\mu - \mathsf{E}\mu)^3$$

для любого $z \in \mathcal{R}$. Ковариационный оператор случайного вектора μ является самосопряженным положительно полуопределенным оператором, и его матрица в ортонормированном базисе представляет собой ковариационную матрицу координат вектора μ в этом базисе.

Вся информация об измерении — это модель измерения, описываемая парой (A, S) и результат измерения y. Будем рассматривать здесь лишь измерения, в которых оператор S положительно определен, $S > 0$ и, следовательно, обратим. По сути, это означает, что шум ν возможен во всех направлениях, т.е., не существует собственного подпространства $\tilde{\mathcal{R}} \subset \mathcal{R}$ такого, что $\nu \in \tilde{\mathcal{R}}$ с вероятностью единица. Таким образом, исходные (или сырые) данные об измерении элемента $x \in \mathcal{D}$, как и в предыдущей главе, представляются тройками вида (y, A, S), где $y \in \mathcal{R}$, $A: \mathcal{D} \to \mathcal{R}$, $S: \mathcal{R} \to \mathcal{R}$, $S > 0$, а $\mathcal{R}$ – некоторое линейное пространство. Множество всех таких троек вида (y, A, S) будем обозначать $\mathfrak{R}$. Ниже мы рассмотрим структуру этого пространства более подробно.

Кроме того, в отличие от предыдущей главы, будем считать, что имеется априорная информация относительно объекта измерения x. В математической статистике априорная информация об x задается некоторым

[3] Все рассматриваемые линейные пространства являются конечномерными евклидовыми со скалярным произведением $\langle \cdot, \cdot \rangle$.

вероятностным распределением на пространстве $\mathcal{D}$ [6], [1], т.е. x рассматривается как случайный вектор, независимый с ν. Мы будем считать, что известны лишь некоторые свойства априорного распределения, а именно, его априорное среднее $\mathsf{E}x = x_0$ и априорный ковариационный оператор $\mathsf{D}x = F\colon \mathcal{D} \to \mathcal{D}$. Также будем считать, что оператор F положительно определен, $F > 0$, т.е. не существует собственного подпространства $\widetilde{\mathcal{D}} \subset \mathcal{D}$ такого, что $x - x_0 \in \widetilde{\mathcal{D}}$ с вероятностью единица. Множество всех таких пар (x_0, F) обозначим $\mathfrak{E}$, т.е.,

$$\mathfrak{E} = \{(x_0, F) \mid x_0 \in \mathcal{D}, F\colon \mathcal{D} \to \mathcal{D}, F > 0\}.$$

3.1.2 Оптимальное линейное оценивание

Задача линейного оценивания с априорной информацией о векторе x состоит в построении оценки $\hat{x}$ вида

$$\hat{x} = Ry + r, \tag{11}$$

определяемого линейным отображением $R\colon \mathcal{R} \to \mathcal{D}$ и вектором сдвига $r \in \mathcal{D}$. При этом оценка $\hat{x} = Ry + r$ должна быть в среднем максимально близка к x. Формально, рассмотрим среднюю погрешность оценки $H(R, r) = \mathsf{E}\|\hat{x} - x\|^2$. Из (10) и (11) сразу следует, что

$$\begin{aligned}\|\hat{x} - x\|^2 &= \|(RA - I)x + r + R\nu\|^2 \\ &= \|(RA - I)x + r\|^2 + 2\langle (RA - I)x + r, R\nu\rangle + \|R\nu\|^2.\end{aligned}$$

Усредняя это выражение по ν, и учитывая, что $\mathsf{E}_\nu \nu = 0$, получаем

$$\mathsf{E}_\nu\|\hat{x} - x\|^2 = \|(RA - I)x + r\|^2 + \mathsf{E}_\nu\|R\nu\|^2 = \|(RA - I)x + r\|^2 + \mathrm{tr} RSR^*.$$

В последнем равенстве мы воспользовались тем, что $\mathsf{D}(R\nu) = RSR^*$ и $\mathsf{E}\|\mu\|^2 = \mathrm{tr}\,\mathsf{D}\mu$ для случайного вектора μ с нулевым средним.

Поскольку в выражении для $\mathsf{E}_\nu\|\hat{x} - x\|^2$ присутствует неизвестный вектор x, определим погрешность оценивания, обеспечиваемую парой (R, r), как усредненную по априорному распределению вектора x, погрешность $\mathsf{E}_\nu\|\hat{x} - x\|^2$, т.е.,

$$H(R, r) = \mathsf{E}_x\mathsf{E}_\nu\|\hat{x} - x\|^2.$$

Обозначая $\tilde{x} = x - x_0$ и учитывая, что $\mathsf{E}_x\tilde{x} = 0$, получаем

$$\begin{aligned} H(R,r) &= \mathsf{E}_x\|(RA-I)(\tilde{x}+x_0)+r\|^2 + \mathrm{tr}RSR^* \\ &= \mathsf{E}_x\|(RA-I)\tilde{x}\|^2 + 2\mathsf{E}_x\langle(RA-I)\tilde{x},(RA-I)x_0+r\rangle \\ &+ \|(RA-I)x_0+r\|^2 + \mathrm{tr}RSR^* \\ &= \mathrm{tr}(RA-I)F(RA-I)^* + \|(RA-I)x_0+r\|^2 + \mathrm{tr}RSR^*. \end{aligned}$$

Итак, задача линейного оценивания с априорной информацией состоит в построении таких R и r, при которых средняя погрешность оценивания $H(R,r)$ минимальна:

$$\min_{R,r} H(R,r) = \min_{R,r}(\mathrm{tr}(RA-I)F(RA-I)^* + \|(RA-I)x_0+r\|^2 + \mathrm{tr}RSR^*). \tag{12}$$

Легко видеть, что r входит только во второе слагаемое $\|(RA-I)x_0 + r\|^2$, которое всегда неотрицательно и обращается в 0 только при

$$r = (I-RA)x_0. \tag{13}$$

Поэтому можно исключить r из задачи минимизации (12) и свести ее к проблеме минимизации только относительно R:

$$\min_R H(R) = \min_R \mathrm{tr}((RA-I)F(RA-I)^* + RSR^*).$$

Пусть $\mathbb{S}_{\mathcal{D}}$ – пространство всех самосопряженных операторов на $\mathcal{D}$. Определим частичный порядок на $\mathbb{S}_{\mathcal{D}}$ следующим образом:

$$P \geq Q \;\Leftrightarrow\; P - Q \geq 0.$$

Заметим, что tr является строго монотонным отображением из пространства самосопряженных операторов в вещественную прямую, а именно, если $P \geq Q$ то $\mathrm{tr}P \geq \mathrm{tr}Q$, а если, кроме того $P \neq Q$ то $\mathrm{tr}P > \mathrm{tr}Q$. Поэтому, рассмотрим задачу минимизации оператора $Q = (RA-I)F(RA-I)^* + RSR^*$. Фактически, оператор Q, который можно определить как $Q(z) = \mathsf{E}\langle\hat{x}-x,z\rangle(\hat{x}-x)$ для $z \in \mathcal{D}$, описывает корреляционные свойства погрешности оценивания $\hat{x}-x$, поэтому будем называть его оператором погрешности оценивания. Докажем, что существует единственное линейное отображение R, доставляющее минимум оператору Q.

Сначала преобразуем Q к виду, в котором явно выделены «квадратичные» и «линейные», относительно R, члены:

$$Q = R(AFA^* + S)R^* + RAF + FA^*R^* + F = RCR^* + RD^* + DR^* + F,$$

где $C = AFA^* + S > 0$ и $D = FA^*$.

Используя обратимость оператора C и, проводя процедуру, аналогичную «выделению полного квадрата» в выражении $RCR^* + RD^* + DR^*$, получаем

$$Q = (R - DC^{-1})C(R - DC^{-1})^* + F - DC^{-1}D^*.$$

В этом выражении только первое слагаемое включает R и для любого R оператор $(R - DC^{-1})C(R - DC^{-1})^* \geq 0$. Из невырожденности C сразу следует, что если $R - DC^{-1} \neq 0$ то и $(R - DC^{-1})C(R - DC^{-1})^* \neq 0$. Следовательно, Q и $H(R) = \operatorname{tr} Q$ достигают минимальных значений в единственной точке $R = DC^{-1}$. При этом оператор Q достигает минимального значения

$$Q = F - DC^{-1}D^* = F - RAF = (I - RA)F. \qquad (14)$$

Отсюда следует, что оптимальное R определяется выражением

$$R = FA^*(AFA^* + S)^{-1} = (F^{-1} + A^*S^{-1}A)^{-1}A^*S^{-1},$$

Чтобы убедиться в справедливости последнего равенства, достаточно домножить обе части легко проверяемого тождества

$$(F^{-1} + A^*S^{-1}A)^{-1}FA^* = A^*S^{-1}(AFA^* + S)$$

на $(F^{-1} + A^*S^{-1}A)^{-1}$ слева и на $(AFA^* + S)^{-1}$ справа.

Несложно убедиться, что

$$I - RA = [I - (F^{-1} + A^*S^{-1}A)^{-1}A^*S^{-1}A] = (F^{-1} + A^*S^{-1}A)^{-1}F^{-1}.$$

Используя это выражение в (13) и (14), получим явные выражения для оптимального вектора сдвига:

$$r = (F^{-1} + A^*S^{-1}A)^{-1}F^{-1}x_0$$

и для минимального значения оператора погрешности:

$$Q = (F^{-1} + A^*S^{-1}A)^{-1}.$$

Таким образам, задача оптимального оценивания с априорной информацией о сигнале вида

$$\min_{R:\mathcal{R}\to\mathcal{D},r\in\mathcal{D}} \mathsf{E}\|Ry+r-x\|^2$$

имеет единственное решение

$$R=(F^{-1}+A^*S^{-1}A)^{-1}A^*S^{-1},\quad r=(F^{-1}+A^*S^{-1}A)^{-1}F^{-1}x_0.$$

При этом оптимальная оценка вектора x

$$\hat{x}=Ry+r=(F^{-1}+A^*S^{-1}A)^{-1}(F^{-1}x_0+A^*S^{-1}y) \tag{15}$$

обладает наименьшим оператором погрешности оценивания

$$Q=RSR^*=(F^{-1}+A^*S^{-1}A)^{-1}. \tag{16}$$

Отсюда, в частности, следует, что оценка $\hat{x}=Ry+r$ обладает минимальными погрешностями оценивания для каждой из координат x_j в некотором ортонормированном базисе:

$$\mathsf{E}(\hat{x}_j-x_j)^2=Q_{jj}=((F^{-1}+A^*S^{-1}A)^{-1})_{jj}$$

и $\mathsf{E}\|\hat{x}-x\|^2$ достигает минимального значения

$$\mathsf{E}\|\hat{x}-x\|^2=\mathrm{tr}Q=\mathrm{tr}(F^{-1}+A^*S^{-1}A)^{-1}.$$

3.1.3 Байесовское оценивание в случае нормальных распределений

Отметим, что в случае нормальных распределений, т.е., если распределение погрешности ν, $P_\nu=N(0,S)$ и априорная информация $P_x=N(x_0,F)$, то условное распределение x при условии наблюдения y также нормально и $P_{x|y}=N(\hat{x},Q)$, где $\hat{x}$ и Q определяются формулами (15) и (16), см., напр., [6], [8], [9]. Таким образом, рассматриваемая процедура перехода от априорной информации к апостериорной полностью соответствует стандартному байесовскому переходу для нормальных распределений.

3.1.4 Исчезающая априорная информация

Рассмотрим предельный случай исчезающей априорной информации. Для этого будем считать, что ковариационный оператор $F=F_\alpha$ стремится к ∞ «во всех направлениях» при $\alpha\to\infty$. А именно, пусть $F_\alpha\geq\alpha I$. Отсюда сразу

следует, что $F_\alpha^{-1} \le \frac{1}{\alpha} I$ и $F_\alpha^{-1} \to 0$. При этом для оценки вектора x и оператора погрешности оценивания Q получаем

$$\lim_{\alpha\to\infty} \hat{x}_\alpha = \lim_{\alpha\to\infty} (F_\alpha^{-1} + A^* S^{-1} A)^{-1} (F_\alpha^{-1} x_0 + A^* S^{-1} y) = (A^* S^{-1} A)^{-1} A^* S^{-1} y,$$

$$\lim_{\alpha\to\infty} Q_\alpha = \lim_{\alpha\to\infty} (F_\alpha^{-1} + A^* S^{-1} A)^{-1} = (A^* S^{-1} A)^{-1},$$

что отвечает оптимальному линейному оцениванию неизвестного вектора x, рассмотренному в предыдущей главе.

3.1.5 Априорная информация как дополнительное измерение

Отметим, что априорная информация $(x_0, F) \in \Re$, описываемая средним x_0 и ковариационным оператором F может формально рассматриваться как дополнительное измерение. Действительно, пусть в дополнение к измерению вида (10) производится независимое измерение, описываемое моделью (I, F), т.е.,

$$x_0 = x + \mu,$$

сопровождающееся шумом $\mu \in \mathcal{D}$ с нулевым средним $\mathsf{E}\mu = 0$ и заданным ковариационным оператором $\mathsf{D}\mu = F : \mathcal{D} \to \mathcal{D}$. Как отмечалось ранее, такая пара измерений может рассматриваться как одно измерение неизвестного вектора x:

$$\begin{pmatrix} x_0 \\ y \end{pmatrix} = \begin{pmatrix} I \\ A \end{pmatrix} x + \begin{pmatrix} \mu \\ \nu \end{pmatrix},$$

где $\begin{pmatrix} x_0 \\ y \end{pmatrix} \in \mathcal{D} \times \mathcal{R}$ – результат такого двойного измерения, $\begin{pmatrix} I \\ A \end{pmatrix} : \mathcal{D} \to \mathcal{D} \times \mathcal{R}$ – линейное отображение, описывающее пару измерений, и $\begin{pmatrix} \mu \\ \nu \end{pmatrix} \in \mathcal{D} \times \mathcal{R}$ – случайный вектор шума с ковариационным оператором $\mathsf{D}\begin{pmatrix} \mu \\ \nu \end{pmatrix} = \begin{pmatrix} F & 0 \\ 0 & \mathrm{S} \end{pmatrix} : \mathcal{D} \times \mathcal{R} \to \mathcal{D} \times \mathcal{R}$.

Согласно предыдущей главе, оптимальная линейная оценка неизвестного и ее ковариационный оператор определяются выражениями

$$\hat{x} = (F^{-1} + A^* S^{-1} A)^{-1} (F^{-1} x_0 + A^* S^{-1} y),$$

$$\mathsf{D}\hat{x} = (F^{-1} + A^* S^{-1} A)^{-1}.$$

Эти формулы в точности совпадают с выражениями (15) и (16) для оптимальной оценки отвечающего ей оператора погрешности в задаче оценивания с априорной информацией. Таким образам, априорная информация о векторе x вида $(x_0, F) \in \mathfrak{E}$ может быть формально «заменена» дополнительным измерением, описываемым тройкой $(x_0, I, F) \in \mathfrak{R}$.

3.2 Переход от априорнрой к апостериорной информации

Эксперимент с априорной информацией нередко трактуется (см. напр. [1], [7]) как процедура перехода от априорной информации к апостериорной. Более того, полученная апостериорная информация рассматривается как априорная по отношению следующему измерению [6]. Строго говоря, понятия априорной и апостериорной информации концептуально различны. Поэтому, чтобы такой переход был обоснован, необходимо убедиться, что использование такой апостериорной информации в качестве априорной для следующего измерения приведет тому же результату, что и использование исходной априорной информации для пары измерений.

Итак, пусть априорная информация о векторе x имеет вид $(x_0, F_0) \in \mathfrak{E}$. Согласно (15) и (16), измерение вида

$$y_1 = A_1 x + v_1, \quad \mathsf{D}v_1 = S_1, \tag{17}$$

описываемое данными $(y_1, A_1, S_1) \in \mathfrak{R}$, обеспечит оценку

$$x_1 = (F_0^{-1} + A_1^* S_1^{-1} A_1)^{-1} (F_0^{-1} x_0 + A_1^* S_1^{-1} y_1) \tag{18}$$

и оператор погрешности оценивания

$$F_1 = (F_0^{-1} + A_1^* S_1^{-1} A_1)^{-1}. \tag{19}$$

Мы намеренно использовали обозначения x_1 и F_1 для обозначения оценки и ее оператора погрешности (вместо $\hat{x}$ и Q) чтобы подчеркнуть наше намерение использовать апостериорную информацию $(x_1, F_1) \in \mathfrak{E}$ в качестве априорной для следующего измерения

$$y_2 = A_2 x + \nu_2, \quad \mathsf{D}\nu_2 = S_2. \tag{20}$$

Аналогично предыдущему шагу, получаем оценку x_2 и оператор погрешности F_2:

$$\begin{aligned} x_2 &= (F_1^{-1} + A_2^* S_2^{-1} A_2)^{-1} (F_1^{-1} x_1 + A_2^* S_2^{-1} y_2) \\ &= (F_0^{-1} + A_1^* S_1^{-1} A_1 + A_2^* S_2^{-1} A_2)^{-1} (F_0^{-1} x_0 + A_1^* S_1^{-1} y_1 \\ &\quad + A_2^* S_2^{-1} y_2) \end{aligned} \tag{21}$$

и оператор погрешности оценивания

$$F_2 = (F_1^{-1} + A_2^* S_2^{-1} A_2)^{-1} = (F_0^{-1} + A_1^* S_1^{-1} A_1 + A_2^* S_2^{-1} A_2)^{-1}. \tag{22}$$

Нетрудно убедиться, что задача оценивания с априорной информацией $(x_0, F_0) \in \mathfrak{E}$ и комбинированным измерением

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} x + \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \quad \mathsf{D}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix},$$

описываемым данными $\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \right) \in \mathfrak{R}$ и представляющим пару независимых измерений (17) и (20), также приводят к результату оценивания, представленному выражениями (21) и (22). Это формально подтверждает правомерность использования апостериорной информации в качестве априорной для последующих измерений.

3.3 Последовательное обновление информации для серии измерений

Рассмотрим теперь серию независимых измерений

$$y_i = A_i x + \nu_i, \quad \mathsf{D}\nu_i = S_i, \quad i = 1, \ldots, n \tag{23}$$

Исходные данные, представляющие отдельное измерение, описываются тройкой $(y_i, A_i, S_i) \in \mathfrak{R}$. Рассмотрим схему последовательного «обновления» информации о векторе x, состоящую в переходе от априорной к апостериорной информации при поступлении очередного фрагмента данных (y_i, A_i, S_i).

Итак, пусть имеется исходная априорная информация $(x_0, F_0) \in \mathfrak{E}$. При поступлении первого измерения (y_1, A_1, S_1) оно «преобразует» априорную информацию (x_0, F_0), в апостериорную (x_1, F_1), согласно (18) и (19), которая теперь выступает в качестве априорной для второго измерения (y_2, A_2, S_2) и т.д., см. Рис. 16.

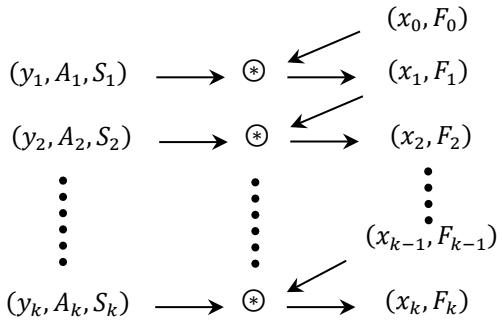

Рис. 16. Последовательное обновление информации.

На k-том шаге добавление данных (y_k, A_k, S_k) к текущей, априорной для данного шага, информации (x_{k-1}, F_{k-1}) и преобразовании ее в апостериорную (x_k, F_k) описывается как

$$(x_k, F_k) = (y_k, A_k, S_k) \circledast (x_{k-1}, F_{k-1}),$$

где

$$F_k = (F_{k-1}^{-1} + A_k^* S_k^{-1} A_k)^{-1}, \qquad x_k = F_k(F_{k-1}^{-1} x_{k-1} + A_k^* S_k^{-1} y_k).$$

Такая процедура последовательного обновления информации считается особенно важной в задачах обработки потоков больших данных («big data streams») поскольку позволяет избежать накопления и хранения больших наборов данных.

Отметим, однако, что такое обновление «явной» информации о векторе x выглядит неоправданно трудоемко, поскольку на каждом шаге требует обращения линейных операторов. Кроме того, в этом процессе комбинируется информация, представленная двумя различными формами: явной, $(x_k, F_k) \in \mathfrak{E}$

и сырой $(y_k, A_k, S_k) \in \mathfrak{R}$, формально, операция обновления $\circledast$ определена на $\mathfrak{R} \times \mathfrak{E}$, т.е., $\circledast: \mathfrak{R} \times \mathfrak{E} \longrightarrow \mathfrak{E}$.

3.4 Последовательное обновление информации в явной форме

Процесс обновления информации, описанный выше, можно сделать более однородным путем преобразования исходной информации в явную форму, перед добавлением ее к накопленной явной информации.

Выше мы видели, что если оператор $A_k^* S_k^{-1} A_k$ обратим, то сырую информацию $(y_k, A_k, S_k) \in \mathfrak{R}$, можно представить в явном виде $(x_k, F_k) \in \mathfrak{E}$, где

$$F_k = (A_k^* S_k^{-1} A_k)^{-1}, \quad x_k = F_k A_k^* S_k^{-1} y_k.$$

В данном случае x_k и F_k есть не что иное как оптимальная линейная оценка вектора x и ее ковариационный оператор, построенные на основании измерения (y_k, A_k, S_k).[4]

Если все исходные данные полученные в результате измерений (23) допускают подобное представление, то, перед добавлением к накопленной информации $(\bar{x}_{k-1}, \bar{F}_{k-1})$ сырой информации (y_k, A_k, S_k), преобразуем последнюю в явную форму (x_k, F_k), Модифицированная схема обновления информации представлена на Рис. 17

[4] Заметим, что здесь (x_k, F_k) обозначает «частичную» информацию, определяемую только k-тым измерением, в отличие от предыдущего раздела, где так обозначалась «полная» накопленная после k-того измерения информация.

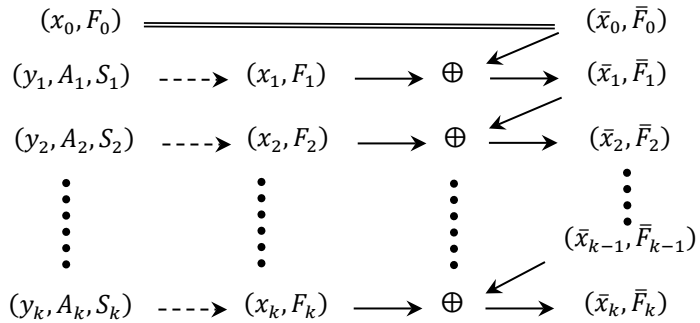

Рис. 17. Последовательное обновление информации с предварительным преобразованием сырой информации в явную форму. Пунктирные стрелки означают, что соответствующее преобразование не всюду определено.

Здесь добавление информации в явной форме (x_k, F_k) к накопленной $(\bar{x}_{k-1}, \bar{F}_{k-1})$, представлено выражением

$$(\bar{x}_k, \bar{F}_k) = (\bar{x}_{k-1}, \bar{F}_{k-1}) \oplus (x_k, F_k),$$

где

$$\bar{F}_k = (\bar{F}_{k-1}^{-1} + F_k^{-1})^{-1}, \qquad \bar{x}_k = \bar{F}_k(\bar{F}_{k-1}^{-1}\bar{x}_{k-1} + F_k^{-1}x_k).$$

Отметим особенности такого подхода. Обновление базируется на операции композиции $\oplus$ двух элементов одного вида – информации в явной форме, т.е., $\oplus$ – бинарная операция на $\mathfrak{E}$. Использование явной формы в качестве основной представляется привлекательным, поскольку явная форма информации представляет собой оценку и ее погрешность, определяемые соответствующим набором данных. Однако, как и для предыдущей схемы, накопление информации сопровождается многократными обращениями линейных операторов. Кроме того, представление сырой информации (y_k, A_k, S_k) в явной форме возможно не всегда, а лишь тогда, когда $A_k^* S_k^{-1} A_k$ обратим, что существенно ограничивает применимость такого подхода. Фактически, это означает, что явная форма не может быть использована как универсальная и эффективная форма представления информации.

3.5 Последовательное обновление информации в канонической форме

Как мы видели в первой главе, удобной промежуточной формой представления результатов измерения $(y, A, S) \in \Re$ является каноническая форма информации $(u, T) = (A^* S^{-1} y,\ A^* S^{-1} A) \in \Im$. При выборе ее в качестве основной формы представления информации, схема последовательного обновления информации принимает вид Рис. 18.

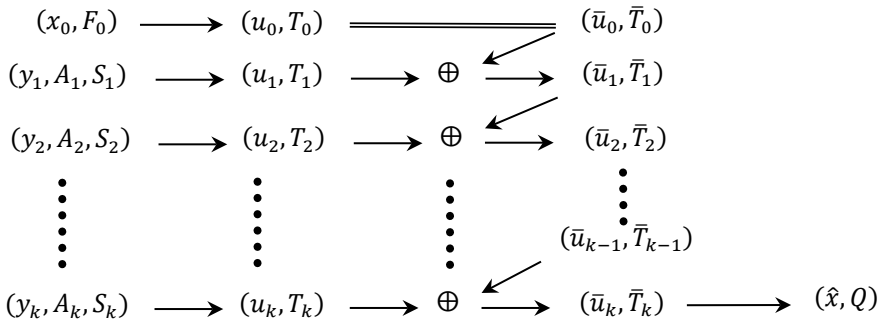

Рис. 18. Последовательное обновление канонической информации с предварительным преобразованием сырой информации в каноническую форму.

Отметим, что в канонической форме, все шаги обработки данных максимально упрощаются. Действительно, преобразования, представленные на Рис. 18, описываются следующими формулами:

Преобразование исходной априорной информации к каноническому виду:

$$(x_0, F_0) \longrightarrow (u_0, T_0) = (F_0^{-1} x_0, F_0^{-1});$$

Преобразование сырой информации к каноническому виду:

$$(y_k, A_k, S_k) \longrightarrow (u_k, T_k) = (A_k^* S_k^{-1} y_k,\ A_k^* S_k^{-1} A_k);$$

Композиция фрагментов канонической информации:

$$(\bar{u}_k, \bar{T}_k) = (\bar{u}_{k-1}, \bar{T}_{k-1}) \oplus (u_k, T_k) = (\bar{u}_{k-1} + u_k,\ \bar{T}_{k-1} + T_k).$$

Построение результата оценивания из канонической информации:

$$(\bar{u}_k, \bar{T}_k) \longrightarrow (\hat{x}, Q) = (\bar{T}_k^{-1},\ \bar{T}_k^{-1}\bar{u}_k).$$

Полученная схема обновления информации, использующая в качестве основной формы информации каноническую, значительно превосходит предыдущие:

- Все формы представления информации легко преобразуются в каноническую;
- Каноническая информация (в отличие от явной) определена для любой сырой информации;
- Композиция информации в канонической форме наиболее эффективна, поскольку описывается покомпонентной суммой пар вида (u, T), где $u \in \mathcal{D}$, $T: \mathcal{D} \to \mathcal{D}$.
- Наиболее трудоемкая часть – построение оценки $\hat{x}$ и оператора погрешности Q может производиться лишь после того, накоплена каноническая информация из большое количества данных. При поступлении новых данных эта процедура может проводиться время от времени по мере необходимости получения обновленной оценки.

3.6 Информационные пространства

Выше мы видели, что в задаче линейного оценивания можно использовать разные формы представления информации:

Исходная *сырая* форма (y, A, S), которая описывает исходные данные;

Окончательная *явная* форма $(\hat{x}, Q)$, в которой представляется результат оценивания или априорная информация (x_0, F);

Промежуточная *каноническая* форма (u, T), максимально удобная для накопления информации.

Определим формально соответствующие информационные пространства $\mathfrak{R}$, $\mathfrak{E}$, и $\mathfrak{I}$. Для некоторого линейного пространства $\mathcal{R}$ будем

обозначать $\mathbb{S}_{\mathcal{R}}$ – пространство самосопряженных операторов на пространстве $\mathcal{R}$, $\mathbb{S}_{\mathcal{R}}^{+} = \{S \in \mathbb{S}_{\mathcal{R}} | \, S > 0\}$ – выпуклый конус положительно определенных операторов на $\mathcal{R}$, $\overline{\mathbb{S}}_{\mathcal{R}}^{+} = \{S \in \mathbb{S}_{\mathcal{R}} \mid S \geq 0\}$ – замкнутый выпуклый конус положительно полуопределенных операторов на $\mathcal{R}$,

3.6.1 Каноническое информационное пространство

Пусть $\mathcal{D}$ – фиксированное пространство объекта измерения x. Мы видели, что элементы канонической информации образуют хорошо организованную алгебраическую структуру – **каноническое информационное пространство**

$$\mathfrak{I} = \{(u, T) \mid T \in \overline{\mathbb{S}}_{\mathcal{D}}^{+}, u \in \mathcal{R}(T)\}$$

с операцией композиции

$$(u_1, T_1) \oplus (u_2, T_2) = (u_1 + u_2, T_1 + T_2).$$

А именно, пусть $\mathbf{0} = (0,0) \in \mathfrak{I}$ (т.е. $u = 0 \in \mathcal{D}$ и $T = 0: \mathcal{D} \to \mathcal{D}$) – элемент, определяющий «отсутствие информации». Тогда $(\mathfrak{I}, \oplus, \mathbf{0})$ является *коммутативным моноидом с сокращением*, т.е., для всех $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathfrak{I}$ выполняются соотношения:

- $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{b} \oplus \boldsymbol{a}$ – коммутативность,
- $(\boldsymbol{a} \oplus \boldsymbol{b}) \oplus \boldsymbol{c} = \boldsymbol{a} \oplus (\boldsymbol{b} \oplus \boldsymbol{c})$ – ассоциативность,
- $\boldsymbol{a} \oplus \mathbf{0} = \boldsymbol{a}$ – свойство нейтрального элемента,
- $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{a} \oplus \boldsymbol{c} \Rightarrow \boldsymbol{b} = \boldsymbol{c}$ – свойство сокращения.

Пусть $\mathfrak{I}^{+}$ – множество всех элементов (u, T) пространства $\mathfrak{I}$, для которых оператор $T > 0$ и, следовательно, обратим, т.е.,

$$\mathfrak{I}^{+} = \{(u, T) \in \mathfrak{I} \mid T > 0\} = \{(u, T) \mid u \in \mathcal{D}, T \in \mathbb{S}_{\mathcal{D}}^{+}\} \subset \mathfrak{I}.$$

Очевидно, $\mathfrak{I}^{+}$ является *коммутативной подполугруппой* моноида $\mathfrak{I}$, но не подмоноидом, поскольку $\mathfrak{I}^{+}$ не содержит нейтральный элемент. Более того, $\mathfrak{I}^{+}$ является *идеалом* в $\mathfrak{I}$, т.е., если $\boldsymbol{a} \in \mathfrak{I}^{+}$ и $\boldsymbol{b} \in \mathfrak{I}$, то $\boldsymbol{a} \oplus \boldsymbol{b} \in \mathfrak{I}^{+}$.

3.6.2 Исходное информационное пространство

Аналогично, все элементы сырой информации, т.е., тройки вида (y, A, S), можно рассматривать как элементы другой алгебраической структуры – исходного (сырого) информационного пространства $\mathfrak{R}$. Определим его формально. Пусть

$$\mathfrak{R}_{\mathcal{R}} = \{(y, A, S) \mid y \in \mathcal{R}, A: \mathcal{D} \to \mathcal{R}, S \in \mathbb{S}^{+}_{\mathcal{R}}\}$$

обозначает множество всех возможных измерений, в пространстве $\mathcal{R}$. Определим **исходное информационное пространство** $\mathfrak{R}$ как объединение всех пространств $\mathfrak{R}_{\mathcal{R}}$, а именно,

$$\mathfrak{R} = \bigcup_{n=0}^{\infty} \mathfrak{R}_{\mathbb{R}^n}.$$

Теперь мы можем формально определить операцию композиции на $\mathfrak{R}$ как

$$(y_1, A_1, S_1) \oplus (y_2, A_2, S_2) = \left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}\right).$$

Нетрудно убедиться, что исходное информационное пространство ($\mathfrak{R}$,$\oplus$, $\mathbf{0}$) является (некоммутативным) *моноидом с сокращением*.

- $(\boldsymbol{a} \oplus \boldsymbol{b}) \oplus \boldsymbol{c} = \boldsymbol{a} \oplus (\boldsymbol{b} \oplus \boldsymbol{c})$,
- $\boldsymbol{a} \oplus \mathbf{0} = \boldsymbol{a} = \mathbf{0} \oplus \boldsymbol{a}$,
- $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{a} \oplus \boldsymbol{c} \;\Rightarrow\; \boldsymbol{b} = \boldsymbol{c} \quad \& \quad \boldsymbol{b} \oplus \boldsymbol{a} = \boldsymbol{c} \oplus \boldsymbol{a} \;\Rightarrow\; \boldsymbol{b} = \boldsymbol{c}$.

Нейтральным элементом $\mathbf{0} \in \mathfrak{R}$ является тройка $\mathbf{0} = (0,0,0) \in \mathfrak{R}_{\mathbb{R}^0}$, т.е., $y = 0 \in \mathbb{R}^0$ – единственный элемент 0-мерного пространства, $A = 0: \mathcal{D} \to \mathbb{R}^0$ – единственное линейное отображение из $\mathcal{D}$ в 0-мерное пространство и $S = I: \mathbb{R}^0 \to \mathbb{R}^0$ – единственный линейный оператор в 0-мерном пространстве.

Пусть $\mathfrak{R}^+$ – множество всех элементов (y, A, S) пространства $\mathfrak{R}$, для которых оператор $A^* S^{-1} A > 0$, т.е.,

$$\mathfrak{R}^+ = \{(y, A, S) \in \mathfrak{R} \mid A^* S^{-1} A > 0\} \subset \mathfrak{R}.$$

Очевидно, $\mathfrak{R}^+$ является (некоммутативной) *подполугруппой* моноида $\mathfrak{R}$, но не подмоноидом, т.к. не содержит нейтральный элемент. Более того, $\mathfrak{R}^+$ является двусторонним *идеалом* в $\mathfrak{R}$, т.е., если $\boldsymbol{a} \in \mathfrak{R}^+$ и $\boldsymbol{b} \in \mathfrak{R}$, то $\boldsymbol{a} \oplus \boldsymbol{b} \in \mathfrak{R}^+$ и $\boldsymbol{b} \oplus \boldsymbol{a} \in \mathfrak{R}^+$.

3.6.3 Явное информационное пространство

Наконец, как мы видели выше, все элементы явной информации, т.е., пары вида (x_0, F) также можно рассматривать как элементы алгебраической структуры – **явного информационного пространства**

$$\mathfrak{E} = \{(x, F) \mid x \in \mathcal{D}, F \in \mathbb{S}_{\mathcal{D}}^+\}$$

с операцией композиции

$$(x_1, F_1) \oplus (x_2, F_2) = ((F_1^{-1} + F_2^{-1})^{-1}(F_1^{-1}x_1 + F_2^{-1}x_2),\ (F_1^{-1} + F_2^{-1})^{-1}).$$

Несложно видеть, что пространство $\mathfrak{E}$ не имеет нейтрального элемента и, образует *коммутативную полугруппу с сокращением*.

- $(\boldsymbol{a} \oplus \boldsymbol{b}) \oplus \boldsymbol{c} = \boldsymbol{a} \oplus (\boldsymbol{b} \oplus \boldsymbol{c})$,
- $\boldsymbol{a} \oplus \boldsymbol{0} = \boldsymbol{a}$,
- $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{a} \oplus \boldsymbol{c} \ \Rightarrow\ \boldsymbol{b} = \boldsymbol{c}$.

3.6.4 Сравнение информационных пространств

Вкратце обсудим достоинства и недостатки работы с информацией в этих формах.

Сырая информация:

1. Тривиальным образом представляет всю информацию, содержащуюся в исходных данных.
2. По мере поступления новых данных, размер памяти, необходимый для их хранения будет расти и потенциально неограничен.
3. Комбинирование информации в этой форме не требует специальных вычислений, однако, с ростом размера, временные

затраты на организацию соответствующих массивов данных будут неограниченно расти.

4. Вычисление окончательного результата оценивания потребует значительных вычислительных ресурсов, в связи с необходимостью производить вычисления с матрицами огромных размеров.

Явная информация:

1. Не всегда позволяет представить информацию, содержащуюся в исходных данных. Для возможности представления в явном виде необходима обратимость оператора $A^*S^{-1}A$.
2. Размер памяти не зависит от объема представленных данных ($\frac{m(m+3)}{2}$ чисел).
3. Комбинирование информации в этой форме требует многократного обращения матриц фиксированного размера $m \times m$ и умножений таких матриц на столбцы.
4. Вычисление окончательного результата не требует никаких вычислений, поскольку, по самому определению, такое представление информации содержит результат оценивания в явной форме.

Каноническая информация:

1. Всегда может представить информацию, содержащуюся в исходных данных.
2. Размер памяти не зависит от объема представленных данных ($\frac{m(m+3)}{2}$ чисел).
3. Комбинирование информации в этой форме максимально упрощено и требует только сложения матриц фиксированного размера $m \times m$ и сложения m-мерных столбцов.
4. Вычисление окончательного результата требует умеренных вычислений (решения системы m уравнений с m неизвестными) и

может выполняться лишь тогда, когда требуется получить результат оценивания по накопленной информации.

Таким образом, каноническая форма представления информации является самой универсальной и эффективной, среди рассмотренных, и использование ее в качестве основной для проведения манипуляций с информацией, позволяет повысить эффективность обработки данных.

3.6.5 Связь с достаточными статистиками и информационными матрицами

Заметим, что в случае нормальных распределений компоненты u и T канонической информации (u, T) имеют интересный теоретико-статистический смысл. Вектор u является *минимальной достаточной статистикой*, а оператор T представляет собой *информационную матрицу Фишера* [1], [7] для измерения y (и достаточной статистики u), см. напр., [8]. Как известно, матрица Фишера описывает количество (возможно, правильнее сказать, качество) информации, содержащейся в измерении. Таким образом, каноническая информация (u, T) в данном контексте представляет собой минимальную достаточную статистику плюс детальную характеризацию ее информативности.

3.7 Работа с информацией в различных формах

На разных стадиях обработки требуется осуществлять преобразования между различными формами представления информации. Рассмотрим эти преобразования более подробно, Рис. 19.

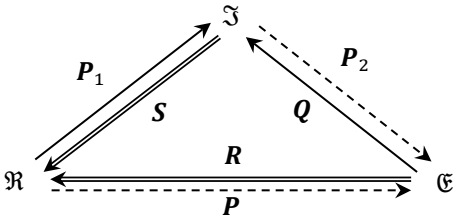

Рис. 19. Преобразования между разными формами информации. Отображения, обозначенные пунктиром, определены не всюду, двойными линиями – определены не однозначно.

Преобразования одних форм в другие описывается следующими формулами:

Преобразование $\boldsymbol{P}: \mathfrak{R} \longrightarrow \mathfrak{E}$ *сырой информации* (y, A, S) *в явную.* Фактически, реализует полное решение задачи оптимального оценивания, предоставляющее в качестве результата оптимальную линейную оценку $x_0 = \hat{x}$ и оператор погрешности оценивания $F = \mathsf{D}\hat{x}$. Отображение $\boldsymbol{P}$ является частично определенным, точнее, оно определено лишь если оператор $A^*S^{-1}A$ обратим, т.е., на подполугруппе $\mathfrak{R}^+ \subset \mathfrak{R}$:

$$\boldsymbol{P}: \mathfrak{R}^+ \longrightarrow \mathfrak{E}, \qquad \boldsymbol{P}: (y, A, S) \mapsto (x_0, F) = ((A^*S^{-1}A)^{-1}A^*S^{-1}y,\ (A^*S^{-1}A)^{-1}).$$

Нетрудно убедиться, что на своей области определения $\mathfrak{R}^+$ отображение $\boldsymbol{P}$ сохраняет алгебраическую структуры пространств, а именно, является частично определенным *гомоморфизмом полугрупп*, т.е.,

$$\forall \boldsymbol{d}_1, \boldsymbol{d}_2 \in \mathfrak{R}^+ \quad \boldsymbol{P}(\boldsymbol{d}_1 \oplus \boldsymbol{d}_2) = \boldsymbol{P}(\boldsymbol{d}_1) \oplus \boldsymbol{P}(\boldsymbol{d}_2).$$

Преобразование $\boldsymbol{P}_1: \mathfrak{R} \longrightarrow \mathfrak{I}$ *сырой информации* (y, A, S) *в каноническую.* Определено всюду (мы считаем, что оператор S всегда обратим):

$$\boldsymbol{P}_1: \mathfrak{R} \longrightarrow \mathfrak{I}, \qquad \boldsymbol{P}_1: (y, A, S) \mapsto (u, T) = (A^*S^{-1}y,\ A^*S^{-1}A)$$

и является (всюду определенным) *гомоморфизмом моноидов*, т.е.,

$$\boldsymbol{P}_1(\boldsymbol{0}_{\mathfrak{R}}) = \boldsymbol{0}_{\mathfrak{I}} \quad \text{и} \quad \forall \boldsymbol{d}_1, \boldsymbol{d}_2 \in \mathfrak{R} \quad \boldsymbol{P}_1(\boldsymbol{d}_1 \oplus \boldsymbol{d}_2) = \boldsymbol{P}_1(\boldsymbol{d}_1) \oplus \boldsymbol{P}_1(\boldsymbol{d}_2).$$

Преобразование $\boldsymbol{P}_2: \mathfrak{I} \longrightarrow \mathfrak{E}$ *канонической информации* (u, T) *в явную.* Определено лишь если оператор T обратим, т.е., на подполугруппе $\mathfrak{I}^+ \subset \mathfrak{I}$:

$$\boldsymbol{P}_2: \mathfrak{I}^+ \longrightarrow \mathfrak{E}, \qquad \boldsymbol{P}_2: (u, T) \mapsto (x_0, F) = (T^{-1}u, T^{-1})$$

и является *гомоморфизмом коммутативных полугрупп*, т.е.,

$$\forall \boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathfrak{I}^+ \quad \boldsymbol{P}_2(\boldsymbol{c}_1 \oplus \boldsymbol{c}_2) = \boldsymbol{P}_2(\boldsymbol{c}_1) \oplus \boldsymbol{P}_2(\boldsymbol{c}_2).$$

Отображения $\boldsymbol{P}_1$ и $\boldsymbol{P}_2$ обеспечивают факторизацию отображения $\boldsymbol{P}$, т.е., $\boldsymbol{P} = \boldsymbol{P}_2 \circ \boldsymbol{P}_1$, позволяющую разбивать обработку данных (отображение $\boldsymbol{P}$) на две фазы: выделение канонической информации (отображение $\boldsymbol{P}_1$) и построение результата обработки на основе канонической информации (отображение $\boldsymbol{P}_2$).

Преобразование $\boldsymbol{Q}$ *явной информации* (x_0, F) *в каноническую.* Определено всюду (мы считаем, что оператор F всегда обратим):

$$\boldsymbol{Q}: \mathfrak{E} \longrightarrow \mathfrak{I}, \qquad \boldsymbol{Q}: (x_0, F) \mapsto (u, T) = (F^{-1}x_0, F^{-1}).$$

Отображение $\boldsymbol{Q}$ позволяет представить любую явную (и в частности, априорную) информацию в каноническом виде. Нетрудно видеть, что отображения $\boldsymbol{Q}$ и $\boldsymbol{P}_2$ являются, в некотором смысле, взаимно обратными, точнее, они являются *изоморфизмами коммутативных полугрупп* $\mathfrak{E}$ и $\mathfrak{I}^+$, а именно, $\boldsymbol{P}_2 \circ \boldsymbol{Q} = \boldsymbol{I}_{\mathfrak{E}}$ и $\boldsymbol{Q} \circ \boldsymbol{P}_2 = \boldsymbol{I}_{\mathfrak{I}^+}$.

Преобразование $\boldsymbol{R}$ *явной информации* (x_0, F) *в сырую.* Заметим, что существует бесконечное множество элементов сырой информации (троек вида (y, A, S)), обеспечивающих результат оценивания (x_0, F). Наиболее очевидный из них $(x_0, I, F) \in \mathfrak{R}$. Определим отображение $\boldsymbol{R}$ как

$$\boldsymbol{R}: \mathfrak{E} \longrightarrow \mathfrak{R}, \qquad \boldsymbol{R}: (x_0, F) \mapsto (y, A, S) = (x_0, I, F).$$

Отображение $\boldsymbol{R}$ позволяет представить любую явную (и в частности, априорную) информацию в виде результата некоторого гипотетического измерения. Нетрудно проверить, что $\boldsymbol{R}$ является правым обратным к, $\boldsymbol{P}$ т.е., $\boldsymbol{P} \circ \boldsymbol{R} = \boldsymbol{I}_{\mathfrak{E}}$, но не является гомоморфизмом полугрупп, т.к. формально не сохраняет операцию $\oplus$.

Преобразование $\boldsymbol{S}$ *канонической информации* (u,T) *в сырую.* Как и в предыдущем случае существует бесконечное множество элементов сырой информации, приводящих к канонической информации (u,T). Один из вариантов можно определить следующим образом:

$$\boldsymbol{S}\colon \Im \longrightarrow \Re, \qquad \boldsymbol{S}\colon (u,T) \mapsto (y,A,S) = \big(u,T,T+(I-P)\big).$$

Здесь $P = TT^{-}$ – ортогональный проектор на $\mathcal{R}(T)$ – пространство значений оператора T [2], [10], а $I-P$ ортогональный проектор на $\mathcal{R}^{\perp}(T)$. Отображение $\boldsymbol{S}$ позволяет представить каноническую информацию в виде результата некоторого гипотетического измерения, но, по-видимому, не имеет практической ценности. Подобно $\boldsymbol{R}$, отображение $\boldsymbol{S}$ является правым обратным к, $\boldsymbol{P}_1$ т.е., $\boldsymbol{P}_1 \circ \boldsymbol{R} = \boldsymbol{I}_{\mathfrak{E}}$, но не является гомоморфизмом моноидов.

3.8 Параллельная распределенная обработка данных в задаче линейного оценивания с априорной информацией

Отображения информационных пространств, упомянутые выше, обеспечивают широкие возможности преобразования информации в процессе обработки. Наиболее универсальным и эффективным представляется преобразование как априорной информации, так и всех имеющихся данных в каноническую форму, комбинирование информации в этой форме и вычисление результата оценивания на основании $(\hat{x},Q)$ накопленной канонической информации (Рис. 20).

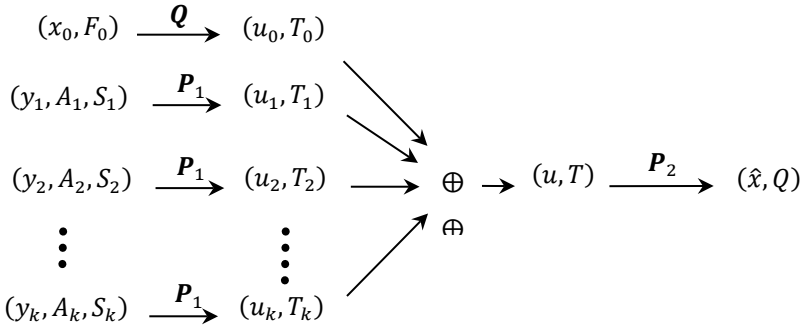

Рис. 20. Параллельная распределенная обработка данных и априорной информации.

Поскольку преобразование различных фрагментов данных и априорной информации может производиться независимо и параллельно на различных компьютерах. Сложение фрагментов канонической информации требует минимальных вычислительных ресурсов. При появлении новых фрагментов данных, будет необходимо преобразовать их в каноническую форму, добавить к накопленной канонической информации и пересчитать результат, используя обновленную каноническую информацию.

Одним из существенных недостатков Байесовского подхода считается сильная зависимость результата оценивания от априорного распределения, в результате чего, ошибочная априорная информация может приводить к ошибочным результатам оценивания. Однако, наличие сократимости и коммутативности в пространствах $\mathfrak{I}$ и $\mathfrak{E}$ позволяет в любой момент «вычесть» исходную априорную информацию из накопленной и «замените» ее на другую. Аналогично, из накопленной информации можно «вычесть» любую предварительно включенную информацию, если впоследствии выяснится, что по тем или иным причинам соответствующее измерение было недостоверно.

Отметим особо, что благодаря введению специальной промежуточной формы представления информации, появляется возможность преобразовать

последовательную, по своей природе, процедуру байесовского обновления информации в форму, допускающую высокую степень параллелизации и масштабирования. В результате этого, процедура накопления информации органично «вписывается» в модель распределенного хранения и анализа данных MapReduce [11].

3.9 Заключение

Рассмотренная задача линейного оценивания с априорной информацией, предоставляет целый спектр моделей информационных пространств с интересными соотношениями между ними. Два типа информационных пространств: исходное и явное, являются, в некотором смысле «естественными» и, фактически, определяются самой постановкой задачи. Первое формализует пространство исходных данных, а второе – априорные и апостериорные виды информации. Однако, достаточно бедные свойства этих пространств ограничивают возможности оптимизации, алгоритмов обработки данных, опирающихся только на такие формы представления информации. В связи с этим (особенно в контексте параллельной распределенной обработки данных) возникает потребность в построении некоторого «искусственного» информационного пространства, обладающим максимально богатой и универсальной структурой. В определенном смысле, такая специальная форма представления информации отражает саму суть информации, содержащейся в данных.

Выше было показано, что использование канонической формы представления информации в качестве основной для манипуляций с информацией, позволяет не только унифицировать процессы обработки данных, но и повысить их эффективность. Более того, благодаря богатым алгебраическим свойствам канонического информационного пространства, традиционно последовательная процедура байесовского уточнения информации (перехода от априорной к апостериорной информации)

допускает различные варианты распараллеливания. Это открывает возможности гибкого и эффективного масштабирования процедуры накопления информации в распределенных системах обработки данных.

4 Накопление информации в задачах калибровки

В этой главе определяются и исследуются информационные пространства, возникающие в задаче калибровки измерительной системы в случае, когда изначально модель измерения неизвестна. Информацию, извлекаемую из калибровочных измерений, предлагается представлять элементом соответствующего информационного пространства, наделенного определенной алгебраической структурой. Также рассматривается возможность дополнительного повышения точности оценивания за счет многократного измерения неизвестного объекта исследования, что приводит к информационному пространству другого типа. В результате строится алгоритм обработки, содержащий накопление информации двух типов и взаимодействие информационных потоков при одновременном накоплении калибровочных и повторных измерений.

4.1 Необходимость калибровки

Как мы видели ранее, для адекватной обработки экспериментальных данных требуется знание модели измерительной системы, описывающая связь между входом и выходом. Такое знание позволяет построить оптимальный алгоритм обработки для задач линейного оценивания.

Однако нередко модель измерения бывает известна неточно либо совсем неизвестна. В таком случае обычно проводится серия измерений эталонных сигналов, на основании этих результатов строить приближение модели и далее это приближение используется для решения задачи интерпретации. Нередко, например, модель выбирается из некоторого класса таким образом, чтобы обеспечить наилучшее в определенном смысле совпадение предсказаний и результатов измерений на обучающей выборке.

Известно, однако, даже малые отклонения модели, используемой в обработке, от истинной могут привести к большим погрешностям интерпретации. При этом, если калибровочных измерений немного, погрешность в оценке модели может быть довольно большой. В связи с этим если модель измерения известна неточно необходим аккуратный учет этой неточности в задаче интерпретации результатов измерений.

Еще одна проблема, связанная с калибровкой, состоит в необходимости сбора большого объема калибровочных данных. Однако, как будет показано, имеется возможность эффективного и компактного накопления калибровочной информации, избавляющей от необходимости хранить исходные калибровочные данные. Это приводит к калибровочному информационному пространству и к алгоритму обработки, идеально вписывающемуся в подход MapReduce [11], который является ключевым в параллельной распределенной обработке больших данных.

Мы также рассмотрим возможность дополнительного повышения точности оценивания за счет многократного измерения неизвестного объекта исследования, что приведет к потоку информации другого типа. В результате мы получим алгоритм обработки, содержащий накопление информации двух типов и взаимодействие информационных потоков для построения оптимальной оценки объекта исследования при одновременном накоплении калибровочных и повторных измерений.

4.2 Задача калибровки

В этом разделе мы рассмотрим проблему обработки результатов линейного эксперимента, в случае если модель измерения неизвестна, а вся информация о ней извлекается из специальной серии измерений известных объектов – калибровочных измерений. Поскольку информация, извлекаемая из калибровочных измерений с неизбежностью, является неточной, нам

потребуется сведения об оптимальном линейном оценивании при неточной информации о модели измерения [4], [8].

4.2.1 Линейное оценивание при неточной информации о модели измерения

Рассмотрим схему линейного измерения вектора $x \in \mathcal{D}$ вида

$$y = Ax + v,$$

где $y \in \mathcal{R}$ – результат измерения, $A: \mathcal{D} \to \mathcal{R}$ – линейный оператор и $v \in \mathcal{R}$ – случайный вектор шума с нулевым средним $\mathsf{E}v = 0$ и ковариационным оператором $S > 0$. Также предположим, что имеется априорная информация о векторе x, определяемая его априорным средним $\mathsf{E}x = x_0$ и ковариационным оператором $F > 0$.

Если линейное отображение A известно неточно, то согласно (Pyt'ev 1984, 2012), оптимальная линейная оценка $\hat{x}$ вектора x может быть представлена в виде

$$\hat{x} = (A_0^*(S+J)^{-1}A_0 + F^{-1})^{-1}(A_0^*(S+J)^{-1}y + F^{-1}x_0),$$

где операторы

$$A_0 = \mathsf{E}A: \mathcal{D} \to \mathcal{R}$$

и

$$J = \mathsf{E}(A - A_0)\,\overline{F}(A - A_0)^*: \mathcal{R} \to \mathcal{R}$$

описывают информацию об операторе A. А именно, A_0 является оценкой оператора A, а J описывает эффект неточности этой оценки на решение окончательной задачи – оценивание вектора x. Здесь $\overline{F}: \mathcal{D} \to \mathcal{D}$ – нецентральный оператор второго момента случайного вектора x, $\overline{F} = F + x_0 x_0^*$. Для простоты рассуждений мы будем отождествлять операторы с их матрицами в фиксированных ортонормированных базисах.

Точность оценки $\hat{x}$ характеризуется ковариационным оператором вектора $\hat{x} - x$:

$$Q = (A_0^*(S+J)^{-1}A_0 + F^{-1})^{-1}.$$

В частности, полная погрешность оценивания

$$\mathsf{E}\|\hat{x} - x\|^2 = \mathrm{tr} Q.$$

4.2.2 Калибровочные измерения

В случае, когда информация об операторе A изначально отсутствует, она может быть извлечена из результатов калибровочных измерений [12] известных сигналов φ_i:

$$\psi_i = A\varphi_i + \mu_i, \qquad i = 1, \dots, k.$$

Здесь $\varphi_i \in \mathcal{D}$ – известные калибровочные сигналы, $\psi_i \in \mathcal{R}$ – наблюдаемые результаты калибровочных измерений, $\mu_i \in \mathcal{R}$ – независимые случайные векторы погрешности, имеющие то же распределение, что и вектор ν, т.е. нулевое среднее $\mathsf{E}\nu = 0$ и ковариационный оператор $S > 0$.

Последовательность пар векторов $(\varphi_1, \psi_1), \dots, (\varphi_k, \psi_k)$ образует набор калибровочных данных. Пусть $\dim\mathcal{D} = m$ и $\dim\mathcal{R} = n$. Объем калибровочных данных составляет $(m + n)k$ и неограниченно растет при росте числа калибровочных измерений k.

4.2.3 Каноническая калибровочная информация

Однако, оказывается, вся информация, содержащаяся в калибровочных данных, может быть представлена парой линейных операторов вида

$$G = \sum_{i=1}^{k} \psi_i \, \varphi_i^* : \mathcal{D} \to \mathcal{R}, \qquad H = \sum_{i=1}^{k} \varphi_i \, \varphi_i^* : \mathcal{D} \to \mathcal{D},$$

где H – положительно полуопределенный, $H \geq 0$.

Будем говорить, что пара (G, H) представляет *каноническую калибровочную информацию* для набора калибровочных данных $(\varphi_1, \psi_1), \dots, (\varphi_k, \psi_k)$. Множество всех таких пар $\mathfrak{C}$ назовем *калибровочным информационным пространством*.

Поскольку матрицы G и H имеют фиксированные размеры $n \times m$ и $m \times m$ соответственно, то каноническая калибровочная информация занимает

фиксированный объем, не зависящий от количества калибровочных измерений. Более того, если две такие пары (G_1, H_1) и (G_2, H_2) получены из двух наборов калибровочных данных, то объединенный набор будет представляться парой

$$(G_1, H_1) \oplus (G_2, H_2) = (G_1 + G_2, H_1 + H_2).$$

Очевидно, любой набор калибровочных данных может быть представлен такой парой, при этом отсутствие данных представляется парой $\mathbf{0} = (0,0)$.

Несложно убедится, что $(\mathfrak{C}, \oplus, \mathbf{0})$ является коммутативным моноидом со свойством сокращения, т.е., для любых $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathfrak{I}$:

$$\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{b} \oplus \boldsymbol{a},$$

$$(\boldsymbol{a} \oplus \boldsymbol{b}) \oplus \boldsymbol{c} = \boldsymbol{a} \oplus (\boldsymbol{b} \oplus \boldsymbol{c}),$$

$$\boldsymbol{a} \oplus \mathbf{0} = \boldsymbol{a},$$

$$\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{a} \oplus \boldsymbol{c} \;\Rightarrow\; \boldsymbol{b} = \boldsymbol{c},$$

но не имеет обратимых элементов отличных от $\mathbf{0}$, т.е. не существует «отрицательной» информации. Наличие сократимости позволяет «вычесть» информацию если обнаружится ее недостоверность.

Как отмечалось ранее, в терминах алгебраической структуры информационного пространства оказывается возможным единообразно описывать последовательное «накопление» информации и «объединение» информации, полученной из разных источников. При этом многие интуитивно ожидаемые свойства самого понятия «информация» получают адекватное математическое отражение в терминах свойств информационного пространства.

4.2.4 Информация о модели измерения

Каноническая калибровочная информация позволяет получить явную информацию об операторе A, а именно, его оценку A_0 и характеризацию точности этой оценки J [12]:

$$A_0 = GH^{-1}, \quad J = \alpha S,$$

где

$$\alpha = \mathrm{tr}\left(H^{-1}\overline{F}\right).$$

Отметим, что неточность информации об операторе A проявляется не только в использовании приближенного значения A_0 вместо точного, но неизвестного A, но и в эффективном увеличении «шума» измерения: $\overline{S} = J + S = (\alpha + 1)S$ вместо S. Нередко при реализации обработки данных на основании приближенной модели, информация о точности приближенной модели не учитывается. Это может привести к большой неконтролируемой погрешности оценивания. Адекватный учет такой неточности, выражающийся оператором J позволяет не только получить верную погрешность оценивания, но и имеет регуляризирующий эффект и способствует снижению погрешности оценивания, особенно при небольшом объеме калибровочных данных.

При достаточно общих условиях, при неограниченном накоплении калибровочной информации (т.е. при $k \to \infty$) $A_0 \to A$, $J \to 0$ и погрешность оценки $Q \to (A^* S^{-1} A + F^{-1})^{-1}$, а именно, к погрешности, отвечающей точно заданной модели измерения.

4.3 Повышение точности оценивания посредством многократных измерений

4.3.1 Многократные измерения объекта исследования

Для дальнейшего повышения точности оценки рассмотрим возможность многократных измерений неизвестного вектора x:

$$y_j = Ax + v_j, \qquad j = 1, \dots, r.$$

Оказывается, такое r-кратное измерение эквивалентно однократному измерению вида

$$\overline{y} = Ax + \overline{v},$$

где

$$\overline{y} = \frac{1}{r}\sum_{j=1}^{r} y_j$$

и $\overline{\nu} \in \mathcal{R}$ – случайный вектор шума с ковариационным оператором $\frac{1}{r}S = \beta S$. Здесь коэффициент $\beta = \frac{1}{r}$ отражает эффект повышения точности оценивания за счет повторения измерений.

Таким образом, при наличии k калибровочных измерений и r повторных измерений неизвестного вектора оценка и ее погрешность определяются формулами

$$Q = \left(A_0^* \overline{S}^{-1} A_0 + F^{-1}\right)^{-1}, \quad \hat{x} = Q\left(A_0^* \overline{S}^{-1} \overline{y} + F^{-1} x_0\right),$$

где

$$\overline{S} = J + S = (\alpha + \beta)S.$$

4.3.2 Асимптотическое поведение точности оценивания и баланс вкладов в погрешность между калибровочными и повторными измерениями

Предположим, что калибровочные сигналы выбираются случайным образом из некоторого распределения со вторым моментом $\tilde{F}$. Тогда

$$\frac{1}{k}H = \frac{1}{k}\sum_{i=1}^{k} \varphi_i \varphi_i^* \to \tilde{F}$$

при $k \to \infty$ и при достаточно большом числе калибровочных измерений $H \approx k\tilde{F}$ и $\alpha \approx \frac{\mu}{k}$, где $\mu = \mathrm{tr}\left(\tilde{F}^{-1}\overline{F}\right)$. Заметим, что если калибровочные сигналы выбираются из того же «ансамбля», что и неизвестный вектор x, то $\tilde{F} = \overline{F}$ и $\mu = m$ – размерность оцениваемого вектора x. Таким образом,

$$\overline{S} \approx \left(\frac{\mu}{k} + \frac{1}{r}\right) S \to 0$$

при одновременном увеличении как калибровочных, так и повторных измерений, $k, r \to \infty$. При этом, если оператор A невырожден, т.е. $\ker A = \{0\}$,

то и $Q \to 0$. Иными словами, путем увеличения как калибровочных, так и повторных измерений, погрешность оценивания может быть сделана сколь угодно малой. Вклады в погрешность оценивания, определяемые неточностью калибровочной информации об A и погрешностью многократных измерений становятся сравнимыми если $\frac{\mu}{k} \approx \frac{1}{r}$ или $k \approx \mu r$.

4.3.3 Каноническая информация для повторяющихся измерений

Заметим, что в рассмотренной схеме калибровки и повторения измерений производится накопление информации из данных двух сортов: калибровочных данных $(\varphi_1, \psi_1), \dots, (\varphi_k, \psi_k)$ и данных многократных измерений $y_1, \dots, y_r$. Как было показано выше, все калибровочные данные могут быть эффективно представлены канонической калибровочной информацией (G, H).

Аналогично, поскольку данные $y_1, \dots, y_r$ требуются лишь для построения их среднего, как показано ранее, удобно накапливать *каноническую измерительную информацию* в виде (u, r), где $u = \sum_{j=1}^{r} y_j \in \mathcal{D}$ – сумма всех векторов y_j, а $r \in \mathbb{N}$ – их количество. Тогда $\overline{y} = \frac{u}{r}$. Пары вида (u, r) образуют информационное пространство со свойствами аналогичными свойствам калибровочного информационного пространства, а именно, коммутативный моноид со свойством сокращения $(\Re, \oplus, \mathbf{0})$. Как и раньше, $\oplus$ является покомпонентным сложением пар вида (u, r), а нейтральный элемент $\mathbf{0}$ (отсутствие измерений) представляется парой $(0,0)$ – нулевой вектор из $\mathcal{D}$ и натуральное число 0.

В свою очередь, *полная каноническая информация*, получаемая из двух потоков, представляется наборами $(G, H; u, r)$, т.е. элементами произведения моноидов $(\mathfrak{C}, \oplus, \mathbf{0}) \times (\Re, \oplus, \mathbf{0})$. Нетрудно убедиться, что полное информационное пространство, полученное как произведение коммутативных моноидов со свойством сокращения также является коммутативным моноидом с свойством сокращения.

4.4 Накопление канонической информации двух сортов в задаче калибровки с повторяющимися измерениями

4.4.1 Обновление информации для потоков данных

При одновременном проведении калибровочных измерений и многократных измерений объекта исследования, происходит накопление и взаимодействие канонической информации двух сортов, проиллюстрированное Рис. 21.

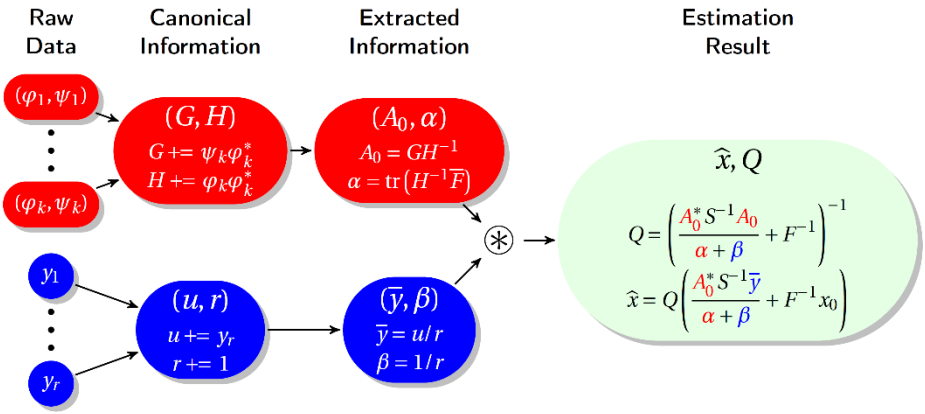

Рис. 21. Функциональная схема накопления и взаимодействия канонической информации двух сортов в задаче калибровки.

Как видно из рисунка, два типа канонической информации, калибровочная (G,H) и многократная измерительная (u,r), могут накапливаться совершенно независимо из соответствующих потоков данных. При этом отпадает необходимость хранить сами данные, они могут выбрасываться сразу же после добавления содержащейся в них информации к канонической информации соответствующего типа.

Отметим, процедуры накопления канонической информации довольно просты, могут быть реализованы на маломощных котроллерах и накапливать каноническую информацию по мере ее поступления. Более трудоемкие

процедуры выделения явной информации из калибровочных измерений об измерительной системе: (A_0, α) и из многократных измерений $(\overline{y}, \beta)$, а также построение окончательного результата оценивания $\hat{x}, Q$ могут производиться время от времени лишь по мере необходимости.

4.4.2 Распределенное накопление двух типов информации в модели MapReduce

В предыдущих главах было показано, что в задачах больших данных естественным образом возникают собственные информационные пространства. Мы видели, что структура адекватных информационных пространств позволяет эффективно распараллеливать процесс накопления информации с использованием модели распределенного анализа данных MapReduce [11] и организовывать эффективную обработку без необходимости накопления и хранения самих исходных данных. В результате процедура накопления информации органично «вписывается» в архитектуру распределенных систем хранения и анализа данных, таких как, например, Hadoop MapReduce или Spark.

На рис. 21 проиллюстрировано накопление информации для случая двух потоков данных: калибровочных и измерительных. В случае распределенных наборов данных схема обработки примет вид, представленный на Рис. 22. Здесь (Φ_i, Ψ_i), $i = 1, \dots, K$ — наборы данных калибровки, а $Y_j = \left(y_{j1} \cdots y_{jr_j}\right)$, $j = 1, \dots, R$ — наборы данных повторных измерений.

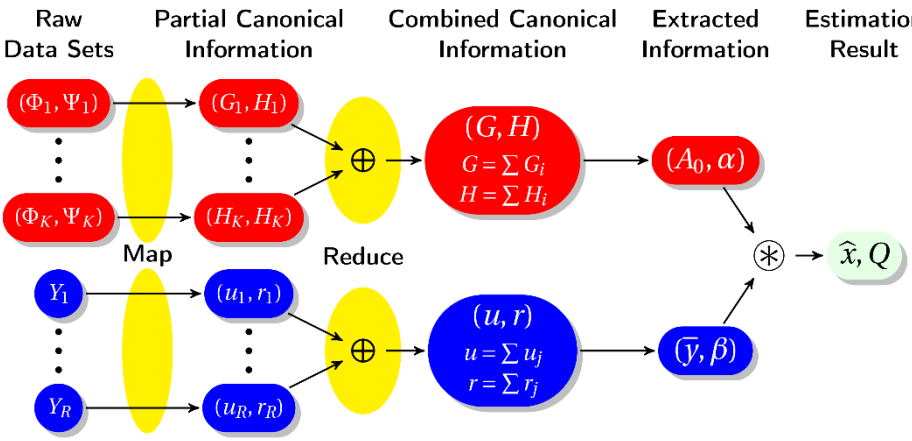

Рис. 22. Функциональная схема распределенного накопления и взаимодействия канонической информации двух видов в задаче калибровки в рамках модели MapReduce

Здесь операция Map извлекает фрагменты информации из нескольких наборов данных, а операция Reduce объединяет все эти частичные фрагменты информации в один элемент, который представляет все исходные наборы данных. Фактически, можно сказать, что любой алгоритм MapReduce основан на определенном информационном пространстве.

В нашей задаче калибровки мы имеем дело с двумя типами данных. В результате весь алгоритм обработки начинается с двух ветвей MapReduce, которые независимо и параллельно накапливают соответствующие типы информации наиболее эффективным способом. Затем определенная явная информация формы (A_0, α) или $(\overline{y}, \beta)$ извлекается из соответствующей накопленной канонической информации и, наконец, эти фрагменты информации взаимодействуют и производят окончательный результат оценки $\hat{x}, Q$.

4.5 Заключение

Рассмотренная в данной статье задача представляет элегантный пример процедуры обработки данных в реальном времени, в котором происходит накопление информации двух разных типов и взаимодействие двух типов информационных потоков. Фактически это пример сложной проблемы, в которой, для получения результата принципиально взаимодействие потоков данных различных сортов. Как было показано, для каждого из входных потоков данных удобно построить специальное информационное пространство, позволяющее производить накопление информации максимально эффективно. Каждое из этих пространств обладает алгебраической структурой отражающей свойства накапливаемой информации. Более того, полная накопленная информация описывается произведением этих пространств и наследует их алгебраические свойства. Таким образом, рассмотренная задача демонстрирует возможности разложения сложного алгоритма со многими потоками входных данных на максимально простые независимые составляющие, а процессы накопления информации их этих потоков и их взаимодействие представляются изящными алгебраическими конструкциями.

Список литературы

1. Барра Ж.-Р. Основные понятия математической статистики. – М.: Мир, 1974.
2. Пытьев Ю. П. Псевдообратный оператор. Свойства и применения // Мат. сб. – 1982. – Т. 118, № 5, – с. 19–49.
3. Пытьев Ю. П. Математические методы интерпретации эксперимента. – М.: Высшая школа, 1989.
4. Пытьев Ю. П. Задачи редукции в экспериментальных исследованиях // Мат. сб. – 1983. – Т. 120, № 2, – с. 240–272.
5. Голубцов П. В. Информативность в категории линейных измерительных систем // Пробл. передачи информ. – 1992. – Т. 28, № 2, – с. 30–46.
6. Lindley, D. Bayesian statistics: A review. – SIAM, Philadelphia, PA, 1972. – 89 p.
7. Боровков А. А. Математическая статистика. – Новосибирск: Наука, 1997. – 772 с.
8. Пытьев Ю. П. Методы математического моделирования измерительно-вычислительных систем. – М.: Физматлит, 2012. – 428 с.
9. Lindley D. V., Smith A. F. M., Bayes Estimates for the Linear Model // Journal of the Royal Statistical Society. Series B – 1972. V. 34, № 1, – P. 1-41.
10. Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание – М.: Наука, 1977. – 224 с.
11. Dean, J., Ghemawat, S. Mapreduce: simplified data processing on large clusters // Communications of the ACM – 2008. 51, № 1, – P. 107-113.
12. Голубцов П.В., Пытьев Ю.П., Чуличков А.И. Построение оператора редукции по тестовым измерениям. Дискретные системы обработки информации. Устинов: изд. Удмуртского государственного университета, 1986. – с. 68-71.