

«Настоящая книга восполняет важную потребность: она знакомит студентов с элегантными реализациями классических алгоритмов в области обработки сигналов и изображений, теории сетей и биоинформатики».  
— Лэв Варини, Университет шт. Иллинойс, США

Добро пожаловать в научное программирование на Python и его сообщество. Если вы — ученый, который программирует на Python, то это практическое руководство для вас! Оно не только познакомит вас с основополагающими компонентами библиотеки SciPy и другими связанными с ней библиотеками, но и даст вам ощущение красоты и удобочитаемости программного кода, который вы сможете применять на практике. Вы научитесь писать элегантный программный код, который ясен, краток и эффективен при исполнении решаемой задачи.

На протяжении всей книги вы будете работать с примерами из обширной научной экосистемы Python, используя программный код, который иллюстрирует кратко очерченные принципы. Используя реальные научные данные, вы будете работать с практическими задачами вместе с SciPy, NumPy, Pandas, scikit-image и другими библиотеками Python.

#### Вы будете:

- исследовать массивы NumPy, то есть структуры данных, которые лежат в основе численных научных вычислений;
- применять квантильную нормализацию, чтобы данные измерений гарантированно укладывались в заданное распределение;
- представлять отдельные области изображения при помощи графа смежности областей;
- преобразовывать временные и пространственные данные в данные частотной области при помощи быстрого преобразования Фурье;
- решать задачи с разреженными матрицами, включая сегментацию изображений при помощи модуля sparse библиотеки SciPy;
- выполнять линейно-алгебраические задачи при помощи пакетов SciPy;
- исследовать выравнивание (регистрацию) изображений при помощи модуля SciPy optimize;
- обрабатывать крупные наборы данных при помощи потоковых примитивов и библиотеки Toolz.

**Хуан Нуньес-Иглесиас (Juan Nunez-Iglesias)** — свободный консультант и научный сотрудник Университета Австралии. Его главные интересы лежат в области нейробиологии и анализа изображений. Он также интересуется графовыми методами в биоинформатике и биостатистике.

**Штефан ван дер Уолт (Stefan van der Walt)** — младший научный сотрудник Института науки о данных в Калифорнийском университете и старший лектор прикладной математики в Университете Штелленбоша, Южная Африка. Он активно участвует в разработке научного программного обеспечения с открытым исходным кодом и преподает Python на семинарах и конференциях, а также является создателем библиотеки scikit-image и одним из разработчиков библиотек NumPy, SciPy и cesium-ml.

**Харриет Дэшноу (Harriet Dashnow)** — биоинформатик, работала в Детском научно-исследовательском институте Мердока, в отделе биохимии в Мельбурнском университете. В настоящее время одновременно с написанием докторской диссертацией она занимается преподаванием на профессиональных семинарах в области геномики, принципов разработки программного обеспечения, Python, R, Unix и системы управления версиями Git.

Интернет-магазин:  
www.dmkpress.com  
Книга — почтой:  
orders@alians-kniga.ru  
Оптовая продажа:  
“Альянс-книга”  
тел. (499) 782-38-89  
books@alians-kniga.ru

**ДМК**  
ИЗДАТЕЛЬСТВО  
www.dmk.pf

ISBN 978-5-97060-600-1



9 785970 606001 >



# Элегантный SciPy

Научное программирование на Python

Хуан Нуньес-Иглесиас  
Штефан ван дер Уолт  
Харриет Дэшноу

**ДМК**  
ИЗДАТЕЛЬСТВО

---

Хуан Нуньес-Иглесиас, Штефан ван дер Уолт  
и Харриет Дэшноу

# Элегантный SciPy



---

Juan Nunez-Iglesias, Stéfan van der Walt  
and Harriet Dashnow



# Elegant SciPy

*The Art of Scientific Python*



Beijing • Boston • Farnham • Sebastopol • Tokyo

**O'REILLY®**

---

Хуан Нуньес-Иглесиас, Штефан ван дер Уолт  
и Харриет Дэшноу



# Элегантный SciPy


*Искусство научного программирования на Python*



Москва, 2018



УДК 373.167.1:004.42+004.42(075.3)  
ББК 32.973.721  
Н87

**Нуньес-Иглесиас Х., Уолт ван дер Ш., Дэшноу Х.**   
Н87 Элегантный SciPy / пер. с англ. А. В. Логунова. – М.: ДМК-Пресс, 2018. –  
266 с.: ил.

**ISBN 978-5-97060-600-1**

Книга познакомит вас с основополагающими компонентами библиотеки SciPy языка Python. Вы научитесь писать элегантный, ясный, краткий и эффективный программный код благодаря примерам из обширной научной экосистемы Python. Кроме SciPy, вы узнаете много нового про сопутствующие библиотеки, такие как NumPy, Pandas, scikit-image.

Издание будет полезно всем программистам на Python, желающим использовать научные библиотеки в своей работе.

УДК 373.167.1:004.42+004.42(075.3)  
ББК 32.973.721

Authorized Russian translation of the English edition of Elegant SciPy ISBN 9781491922873  
© 2017 Juan Nunez-Iglesias, Stéfan van der Walt, and Harriet Dashnow.

This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN 978-1-491-92287-3 (анг.)

Copyright © 2017 Juan Nunez-Iglesias, Stéfan van der Walt, and Harriet Dashnow

ISBN 978-5-97060-600-1 (рус.)

© Оформление, издание, перевод, ДМК Пресс, 2018

# Содержание



<b>Предисловие</b> .....	9
<b>Глава 1. Элегантный NumPy: фундамент научного программирования на Python</b> .....	32
Введение в данные: что такое экспрессия гена? .....	34
N-мерные массивы NumPy .....	38
Зачем использовать массивы ndarray вместо списков Python? .....	39
Векторизация .....	41
Транслирование .....	41
Исследование набора данных экспрессии генов .....	43
Чтение данных при помощи библиотеки pandas .....	43
Нормализация .....	46
Нормализация между образцами .....	46
Нормализация между генами .....	52
Нормализация по образцам и генам: RPKM .....	54
Подведение итогов .....	61
<b>Глава 2. Квантильная нормализация с NumPy и SciPy</b> .....	62
Получение данных .....	64
Разница в распределении экспрессии генов между индивидуумами .....	65
Бикластеризация количественных данных .....	68
Визуализация кластеров .....	70
Предсказание выживаемости .....	72
Дальнейшая работа: использование кластеров пациентов TCGA .....	77
Дальнейшая работа: воспроизведение кластеров TCGA .....	77
<b>Глава 3. Создание сетей из областей изображений при помощи ndimage</b> .....	78
Изображения – это просто массивы NumPy .....	79
Задача: добавление сеточного наложения .....	84
Фильтры в обработке сигналов .....	84
Фильтрация изображений (двумерные фильтры) .....	90
Универсальные фильтры: произвольные функции от соседних значений .....	92
Задача: игра «Жизнь» Конуэя .....	93
Задача: магнитуда градиента Собела .....	94
Графы и библиотека NetworkX .....	94
Задача: подбор кривой при помощи SciPy .....	98

Графы смежности областей.....	98
Элегантный пакет ndimage: как строить графы из областей изображений ...	102
Собираем все вместе: сегментация по среднему цвету .....	105
<b>Глава 4. Частота и быстрое преобразование Фурье .....</b>	<b>107</b>
Введение в частоту .....	107
Иллюстрация: спектрограмма пения птиц .....	110
История .....	115
Реализация .....	115
Выбор длины ДПФ .....	116
Дополнительные понятия ДПФ .....	118
Частоты и их упорядочивание .....	118
Оконное преобразование.....	124
Практическое применение: анализ радарных данных.....	128
Свойства сигнала в частотной области .....	133
Оконное преобразование на практике .....	136
Радарные изображения.....	138
Дополнительные применения БПФ .....	142
Дополнительные материалы для чтения.....	143
Задача: свертывание изображения .....	143
<b>Глава 5. Таблицы сопряженности на основе разреженных координатных матриц .....</b>	<b>144</b>
Таблицы сопряженности.....	146
Задача: вычислительная сложность матриц ошибок.....	147
Задача: альтернативный алгоритм вычисления матрицы ошибок.....	147
Задача: мультиклассовая матрица ошибок .....	148
Форматы данных модуля scipy.sparse .....	148
Формат COO .....	148
Задача: представление в формате COO .....	149
Формат сжатой разреженной строки .....	150
Применения разреженных матриц: преобразования изображений .....	152
Задача: поворот изображения .....	156
Назад к таблицам сопряженности .....	157
Задача: сокращение объема потребляемой оперативной памяти .....	158
Таблицы сопряженности в сегментации изображений .....	159
Теория информации вкратце .....	160
Задача: вычисление условной энтропии .....	163
Теория информации применительно к сегментации: изменчивость информации.....	163
Конвертирование программного кода массивов NumPy под использование разреженных матриц .....	166

Применение изменчивости информации .....	167
Дальнейшая работа: сегментация на практике.....	173
<b>Глава 6. Линейная алгебра в SciPy .....</b>	<b>174</b>
Основы линейной алгебры .....	174
Лапласова матрица графа .....	175
Задача: матрица поворота .....	176
Лапласовы матрицы с данными о мозге.....	181
Задача: изображение аффинного подобия.....	186
Задача: линейная алгебра с разреженными матрицами .....	186
PageRank: линейная алгебра для репутации и важности .....	187
Задача: обработка висячих узлов .....	192
Задача: эквивалентность разных методов получения собственного вектора .....	192
Заключительные замечания .....	192
<b>Глава 7. Оптимизация функций в SciPy .....</b>	<b>193</b>
Оптимизация в SciPy: <code>scipy.optimize</code> .....	195
Пример: вычисление оптимального сдвига изображения.....	195
Регистрация изображения при помощи <code>optimize</code> .....	201
Предотвращение локальных минимумов на основе алгоритма <code>basin hopping</code> .....	204
Задача: модификация функции <code>align</code> .....	205
«Что лучше?»: выбор правильной целевой функции.....	205
<b>Глава 8. Большие данные с Toolz в маленьком ноутбуке .....</b>	<b>212</b>
Потоковая передача при помощи <code>yield</code> .....	214
Введение в потоковую библиотеку Toolz .....	217
Подсчет k-мер и исправление ошибок.....	219
Каррирование: изюминка потоковой обработки.....	223
Возвращаясь к подсчету k-мер .....	226
Задача: анализ главных компонент потоковых данных.....	227
Марковская модель на основе полного генома .....	228
Задача: онлайн-овая распаковка архива .....	231
<b>Эпилог .....</b>	<b>233</b>
Что дальше?.....	233
Списки рассылок .....	233
GitHub .....	234
Конференции .....	235
За пределами SciPy .....	235
Содействие этой книге .....	236

---

До следующей встречи.....	237
<b>Приложение. Решения задач.....</b>	<b>238</b>
Решение: добавление сеточного наложения.....	238
Решение: игра «Жизнь» Конуэя.....	239
Решение: магнитуда градиента Собела.....	240
Решение: подбор кривой при помощи SciPy.....	241
Решение: свертывание изображения.....	243
Решение: вычислительная сложность матриц ошибок.....	243
Решение: альтернативный алгоритм вычисления матрицы ошибок.....	243
Решение: вычисление матрицы ошибок.....	244
Решение: представление в формате COO.....	244
Решение: поворот изображения.....	245
Решение: сокращение объема потребляемой оперативной памяти.....	246
Решение: вычисление условной энтропии.....	247
Решение: матрица поворота.....	247
Решение: изображение аффинного подобия.....	248
Решение: линейная алгебра с разреженными матрицами.....	249
Решение: обработка висячих узлов.....	252
Решение: методы проверки.....	253
Решение: модификация функции align.....	253
Решение: анализ главных компонент потоковых данных при помощи библиотеки scikit-learn.....	255
Решение: добавление шага в начало конвейера.....	257
<b>Предметный указатель.....</b>	<b>259</b>

# Предисловие



В отличие от стереотипного подвенечного стиля, оно было – если использовать технический термин – элегантным, как компьютерный алгоритм, который всего несколькими строками исходного кода достигает впечатляющего результата.

– Грэм Симсион, «Проект “Рози”»

Добро пожаловать в книгу «*Элегантный SciPy*». Мы собираемся провести довольно много времени, сосредоточившись на той части заголовка книги, которая относится к «SciPy», поэтому давайте воспользуемся моментом, чтобы поразмышлять над словом «элегантный». Существует масса руководств, учебных пособий и веб-сайтов документации, которые дают всестороннее описание пакета SciPy. Книга «*Элегантный SciPy*» идет дальше. Она представляет собой нечто большее, чем просто обучение приемам написания по-настоящему рабочего программного кода. Мы вдохновим вас на написание программного кода, который будет по-настоящему потрясающим!

В романе «Проект “Рози”» (между прочим, уморительная книга; когда закончите читать «*Элегантный SciPy*», обязательно прочтите в Википедии<sup>1</sup> материал, описывающий предысторию ее написания) Грэм Симсион переиначивает общепринятые нормы, связанные со словом «элегантный». Большинство людей использует это слово для описания визуальной простоты, стиля и изящества, например, первого мобильного iPhone. Вместо этого герой Грэма Симсона, Дон Тиллман, дает *определение* элегантности, используя термин «компьютерный алгоритм». Мы надеемся, что после прочтения этой книги вы получите ясное понимание того, что он имеет в виду, и что впредь, занимаясь чтением или написанием куска элегантного программного кода, вы будете ощущать умиротворение в лучах его красоты и изящества. (Возьмите на заметку: авторы могут быть подвержены гиперболе.)

Хороший фрагмент программного кода вызывает ощущение удовлетворенности. Когда вы на него смотрите, его замысел *ясен*, он нередко *краток* (но не настолько краток, чтобы быть туманным), и он *эффективен* при выполнении практической задачи. Для авторов удовольствие от анализа элегантного программного кода лежит в скрытых внутри него уроках и в том, как он вдохновляет нас быть *творческими* в подходах к новым алгоритмическим задачам.

Как ни странно, креативность, помимо всего прочего, может также заставлять нас умничать за счет читателя и писать маловразумительный про-

<sup>1</sup> См. [https://en.wikipedia.org/wiki/The\\_Rosie\\_Project](https://en.wikipedia.org/wiki/The_Rosie_Project) и [https://ru.wikipedia.org/wiki/Проект\\_«Рози»](https://ru.wikipedia.org/wiki/Проект_«Рози»).

граммный код, который очень трудно понять. PEP8<sup>1</sup> (руководство по стилю программирования на Python) и PEP20<sup>2</sup> (дзэн языка Python) нам напоминают, что «программный код читается намного больше раз, чем пишется» и поэтому «читаемость имеет значение».

Краткость элегантного кода обеспечивается не за счет упаковки кучи вложенных вызовов функций, а за счет абстракции и рационального использования функций. Она требует одной-двух минут на то, чтобы вникнуть, но в конечном счете обеспечивает вам четкий момент понимания. Как только вы начнете разбираться в различных компонентах программного кода, его правильность должна стать очевидной. Этому способствуют ясные имена переменных и функций и тщательно продуманные комментарии, которые программный код *объясняют*, а не просто его *описывают*.

Инженер-программист Дж. Брэдфорд Хиппс (J. Bradford Hipps) недавно в газете «*Нью-Йорк таймс*»<sup>3</sup> заявил: «для того чтобы писать хороший программный код, следует почитать Вирджинию Вульф»:

*На практике разработка программного обеспечения имеет гораздо более творческий характер, чем алгоритмический.*

*Разработчик обращается к своему редактору исходного кода таким же образом, как писатель к чистому листку бумаги. [...] Разработчика и писателя также могут отличать общее здоровое нетерпение по отношению к тому, как все «делалось всегда», и воспроизводящееся стремление нарушать общепринятые правила. Когда модуль закончен или страницы завершены, их качество оценивается в сопоставлении со многими из тех же самых стандартов: элегантностью, краткостью, целостностью; обнаружением симметрий там, где их существование ни разу не было замечено. И даже красотой.*

Это именно та позиция, которую мы примем в этой книге.

Теперь, когда мы рассмотрели «элегантную» часть заголовка, давайте возвратимся к «SciPy».

В зависимости от контекста «SciPy» может означать пакет программного обеспечения, экосистему или сообщество разработчиков. Отчасти величие SciPy обусловлено тем, что он имеет превосходную онлайн-документацию<sup>4</sup> и учебные пособия<sup>5</sup>, поэтому предлагать читателю очередной справочник было бы бессмысленно. Вместо этого в книге «*Элегантный SciPy*» будет представлен самый лучший программный код, который был создан при помощи SciPy.

<sup>1</sup> См. <https://www.python.org/dev/peps/pep-0008/>.

<sup>2</sup> См. <https://www.python.org/dev/peps/pep-0020/>.

<sup>3</sup> См. <https://www.nytimes.com/2016/05/22/opinion/sunday/to-write-software-read-novels.html>.

<sup>4</sup> См. <https://docs.scipy.org/>.

<sup>5</sup> См. <http://www.scipy-lectures.org/>.

Отобранный нами программный код подчеркивает умное, элегантное использование расширенного функционала NumPy, SciPy и других связанных с ними библиотек. Начинающий читатель научится применять эти библиотеки к реальным задачам, используя красивый программный код. При этом в обоснование наших примеров мы используем реальные научные данные.

Как и сам SciPy, мы хотели построить книгу «*Элегантный SciPy*» на основе сообщества разработчиков. Многие приводимые в книге примеры были взяты из рабочего программного кода, обнаруженного в обширной экосистеме научного программирования на Python, и отобраны для демонстрации кратко очерченных выше принципов элегантного программного кода.

## Для кого эта книга предназначена?

Цель книги «*Элегантный SciPy*» – побудить вас поднять свои навыки программирования на языке Python на более высокий уровень. Вы изучите пакет SciPy на примере самого лучшего программного кода.

Прежде чем приступить к работе, вы должны иметь представление о языке Python и знать, что такое переменные, функции, циклы, и, возможно, немного разбираться в библиотеке NumPy. Хотя вполне может быть, что вы даже отточили свои навыки программирования на Python на основе такого продвинутого материала, как, например, «Python. К вершинам мастерства» (Fluent Python)<sup>1</sup>. Если это к вам не относится, то, прежде чем продолжить чтение этой книги, вам следует начать с каких-нибудь учебных пособий по Python для начинающих, таких как Software Carpentry<sup>2</sup>.

Но, возможно, для вас «стек SciPy» – все равно, что пункт меню сети экспресс-блинных ИНОР, и вы чувствуете себя не в своей тарелке, когда речь идет о его применении на практике. Или, возможно, вы являетесь ученым, который прочитал в Сети несколько учебных пособий по Python и скачал несколько аналитических сценариев из другой лаборатории или у другого сотрудника вашей собственной лаборатории и повозился с ними некоторое время. И, возможно, думаете, что в той или иной степени одиноки в своей попытке научиться программировать SciPy. Вы ошибаетесь.

По ходу изложения мы научим вас использовать Интернет в качестве своего справочника. И мы будем направлять вас к спискам рассылок, хранилищам и конференциям, где вы встретите аналогично мыслящих ученых, которые в своем опыте работы находятся немного дальше, чем вы.

Это такая книга, которую вы прочтете один раз, но будете к ней возвращаться в дальнейшем в поисках вдохновения (и, возможно, чтобы еще раз восхищаться некоторыми элегантными фрагментами программного кода!).

<sup>1</sup> См. <http://shop.oreilly.com/product/0636920032519.do>.

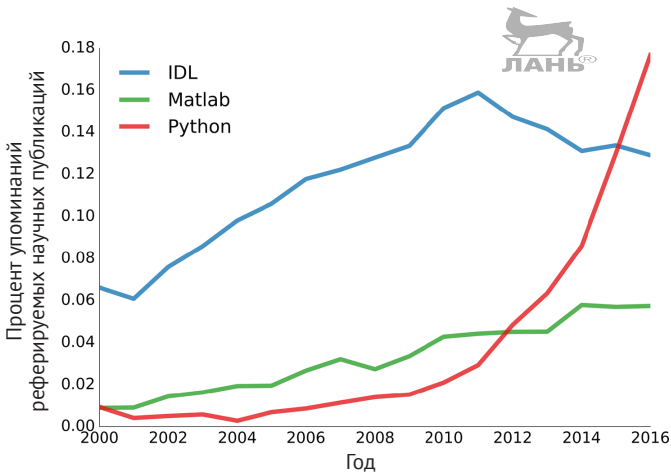
<sup>2</sup> См. <http://software-carpentry.org/>.



## ПОЧЕМУ ИМЕННО SciPy?

Библиотеки NumPy и SciPy составляют ядро научной экосистемы языка Python. В программной библиотеке SciPy реализован набор функций для обработки научных данных из таких областей, как статистика, обработка сигналов, обработка изображений и математическая оптимизация. Библиотека SciPy надстроена поверх библиотеки NumP, которая предназначена для вычислительной обработки числовых массивов. За последние несколько лет вся экосистема приложений и библиотек продемонстрировала существенный рост, опираясь как раз на NumPy и SciPy, с охватом широкого спектра дисциплин, который среди прочих включает астрономию, биологию, метеорологию, климатологию и материаловедение.

И этот рост не проявляет никаких признаков к ослаблению. В 2014 г. Томас Робитэйл и Крис Бомон (Thomas Robitaille и Chris Beaumont) задокументировали<sup>1</sup> рост применения языка Python в астрономии. Вот то, что мы обнаружили, когда обновили<sup>2</sup> их график во второй половине 2016 г.:



Совершенно очевидно, что в течение многих последующих лет SciPy и связанные с ней библиотеки будут основными в подавляющей части аналитической обработки научных данных.

Еще одним убедительным примером служит тот факт, что организация Software Carpentry, которая обучает ученых вычислительным навыкам, используя для этого чаще всего язык Python, сегодня едва справляется со спросом на свои учебные курсы.

<sup>1</sup> См. [https://nbviewer.jupyter.org/github/ChrisBeaumont/adass\\_proceedings/blob/master/Mining%20acknowledgments%20in%20ADS.ipynb](https://nbviewer.jupyter.org/github/ChrisBeaumont/adass_proceedings/blob/master/Mining%20acknowledgments%20in%20ADS.ipynb).

<sup>2</sup> См. <https://gist.github.com/jni/3339985a016572f178d3c2f18e27ec0d>.

## Что такое экосистема SciPy?

SciPy (произносится как «Сай Пи») – это экосистема программного обеспечения с открытым исходным кодом на основе Python для математики, науки и техники.

– <http://www.scipy.org>

Экосистема SciPy представляет собой нестрого определенную коллекцию пакетов Python. В книге «Элегантный SciPy» мы встретимся с большинством ее главных компонентов:

- **NumPy** – фундамент научных вычислений на Python. Эта библиотека обеспечивает эффективные числовые массивы и широкую поддержку численных вычислений, включая линейную алгебру, случайные числа и разложение в ряды Фурье. Уникальная особенность NumPy заключена в ее «N-мерных массивах», или массивах ndarray. Эти структуры данных эффективным образом хранят числовые величины и задают решетку в любом количестве размерностей (подробнее об этом чуть позже)<sup>1</sup>;
- **SciPy**, как библиотека, является коллекцией эффективных численных алгоритмов для таких областей, как обработка сигналов, интеграция, оптимизация и статистика. Эти алгоритмы обернуты в легкие для использования интерфейсы<sup>2</sup>;
- **Matplotlib** – мощный пакет для построения графиков в двух измерениях (и элементарных 3D-графиков). Он берет свое название от навевшего его разработку синтаксиса Matlab<sup>3</sup>;
- **IPython** – интерактивный интерфейс для языка Python, который позволяет оперативно взаимодействовать с вашими данными и тестировать свои идеи<sup>4</sup>;
- блокнот **Jupyter** работает в вашем браузере и позволяет составлять документы с широкими функциональными возможностями, которые объединяют в себе программный код, текст, математические выражения и интерактивные элементы интерфейса<sup>5</sup>. На самом деле при подготовке настоящей книги текст программ был преобразован в блокноты Jupyter и выполнялся там (благодаря этому мы знаем, что все примеры выполняются правильно). Проект Jupyter, начавшийся как расширение IPython,

<sup>1</sup> См. <http://www.numpy.org/>.

<sup>2</sup> См. <http://www.scipy.org/scipylib/index.html>.

<sup>3</sup> См. <http://matplotlib.org/>.

<sup>4</sup> См. <https://ipython.org/>.

<sup>5</sup> См.: Перес Ф. «Грамотные вычисления» и вычислительная воспроизводимость: IPython в эпоху журналистики, ориентированной на данные (публикация в блоге). 19 апреля 2013 г. (Fernando Perez. 'Literate computing' and computational reproducibility: IPython in the age of data-driven journalism. <http://blog.fperez.org/2013/04/literate-computing-and-computational.html>).



теперь поддерживает многочисленные языки, включая Cython, Julia, R, Octave, Bash, Perl и Ruby<sup>1</sup>;

- **pandas** обеспечивает быстрые столбчатые структуры данных в простом для применения пакете. Он в особенности подходит для работы с помеченными наборами данных, такими как таблицы или реляционные базы данных, и для управления данными временных рядов и скользящими окнами. Кроме того, пакет **pandas** располагает несколькими удобными инструментами, предназначенными для анализа данных, в т. ч. разбором, очисткой и агрегированием данных, а также построением графиков<sup>2</sup>;
- **scikit-learn** предоставляет унифицированный интерфейс к алгоритмам машинного обучения<sup>3</sup>;
- **scikit-image** обеспечивает инструменты анализа изображений, которые напрямую интегрируются в остальную часть экосистемы SciPy<sup>4</sup>.

Помимо перечисленных выше библиотек, существует целый ряд других пакетов Python, которые являются частью экосистемы SciPy, и некоторые из них мы также увидим. Хотя в центре внимания настоящей книги будут библиотеки NumPy и SciPy, именно большое количество окружающих их пакетов делает Python движущей силой научных вычислений.

## ВЕЛИКИЙ КАТАКЛИЗМ: ПУТНОН 2 ПРОТИВ ПУТНОН 3



В ваших путешествиях по Python вы, вероятно, уже слышали пересуды о том, какая версия Python лучше. Возможно, вы задавались вопросом, а разве нельзя просто взять последнюю версию. (Сразу отвечаем: можно.)

В конце 2008 г. разработчики ядра Python выпустили Python 3, который среди прочих улучшений стал основным обновлением языка с более оптимальной обработкой текста (на многочисленных естественных языках) в кодировке Юникод, согласованностью типов и обработкой потоковых данных. Как язвительно заметил Дуглас Адамс<sup>5</sup> по поводу создания Вселенной, «это рассердило многих людей и повсеместно признавалось плохим ходом». И все потому, что программный код на Python 2.6 или 2.7 обычно не может исполняться интерпретатором Python 3, по крайней мере без небольшой модификации (хотя изменения, как правило, не слишком обременительны).

Всегда существует трение между неумолимой силой прогресса и обратной совместимостью. В этом случае команда разработчиков ядра языка Python решила, чтобы устранить некоторые несоответствия, в особенности в лежащем в основе программном интерфейсе C, требуется полный разрыв, и продвинула

<sup>1</sup> См. <http://jupyter.org/>.

<sup>2</sup> См. <http://pandas.pydata.org/>.

<sup>3</sup> См. <http://scikit-learn.org/>.

<sup>4</sup> См. <http://scikit-image.org/>.

<sup>5</sup> См.: Адамс Д. Путеводитель автостопщика по Галактике (*Douglas Adams. The Hitchhiker's Guide to the Galaxy*. London: Pan Books, 1979).

язык в XXI век (Python 1.0 появился в 1994 г., более 20 лет назад, что в технологическом мире представляет собой целую жизнь).

Вот один из тех приемов, благодаря которым при переходе к версии 3 язык Python стал лучше:

```
print "Привет, Мир!" # инструкция print в Python 2
print("Привет, Мир!") # функция print в Python 3
```

Зачем создавать столько шума, чтобы добавить несколько круглых скобок! Все верно. Но если вместо этого вы хотите направить печать в другой *поток*, такой, например, как *стандартная ошибка*, т. е. в обычное место для отладочной информации?

```
print >>sys.stderr, "фатальная ошибка" # Python 2
print("фатальная ошибка", file=sys.stderr) # Python 3
```

Такое изменение, несомненно, выглядит целесообразнее. А что же происходит в Python версии 2? Авторы не знают, с полным на то правом.

Еще одно изменение состоит в том, как в Python 3 рассматривается деление целых чисел. Это делается точно так же, как большинство людей выполняет данную операцию. (Обратите внимание, что серия символов >>> говорит о том, что мы набираем программный код в интерактивной оболочке Python.)

```
# Python 2
>>> 5 / 2 2
# Python 3
>>> 5 / 2 2.5
```

Кроме того, нас также сильно порадовал новый оператор *умножения матриц* @, введенный в Python 3.5 в 2015 г. Обратитесь к главам 5 и 6, чтобы увидеть несколько практических примеров применения этого оператора!

Возможно, самым большим улучшением в Python 3 является поддержка Юникода, т. е. приема кодирования текста, который позволяет использовать не только английский алфавит, но и любой другой существующий в мире алфавит. Python 2 позволял вам определять строковое значение в кодировке Юникода следующим образом:

```
beta = u"β"
```

Но в Python 3 абсолютно все является Юникодом:

```
β = 0.5
print(2 * β)
```

```
1.0
```

Команда разработчиков ядра языка Python приняла абсолютно правильное решение, что в программном коде на Python имеет смысл в качестве объектов первого класса поддерживать символы всех естественных языков. Это в особенности актуально теперь, когда большинство новых программистов проживают не в англоязычных странах. Ради функциональной совместимости мы рекомендуем в большей части исходного кода по-прежнему использовать ан-



глийские символы. Эта возможность может пригодиться, например, в блокнотах Jupyter, утяжеленных математическими формулами.

- ☑ Наберите в терминале IPython или в блокноте Jupyter LaTeX-е имя символа, после чего нажмите клавишу Tab. В результате это имя будет расширено в символ Юникода. Например, `\beta<ТАВ>` станет  $\beta$ .

Обновление до Python 3 также разрушает большую часть существующего программного кода в версии 2.x и в некоторых случаях Python 3 выполняется медленнее, чем прежде. Несмотря на эти недостатки, мы рекомендуем всем пользователям как можно скорее обновиться до третьей версии (до 2020 г. Python 2.x теперь находится только в режиме обслуживания), поскольку большинство существовавших проблем будет решено вместе с совершенствованием версий 3.x. И действительно, в этой книге мы используем многие новые возможности Python 3.

В данной книге мы используем **Python 3.6**.

Дополнительные материалы для чтения, касающиеся перехода на версию 3, вы сможете найти на ресурсе Эда Шофилда Python-Future<sup>1</sup> и в справочнике книжного формата<sup>2</sup> Ника Коглана.

## ЭКОСИСТЕМА И СООБЩЕСТВО SciPy

SciPy – это главная библиотека с довольно-таки большой функциональностью. Вместе с NumPy она является одним из уникальных приложений Python. Она положила начало огромному количеству связанных с ней библиотек, которые опираются на ее функционал. Со многими этими библиотеками вы будете работать на протяжении всей книги.

Создатели данных библиотек и многие их пользователи встречаются на многочисленных мероприятиях и конференциях по всему миру. Это такие мероприятия, как ежегодная конференция SciPy в Остин (США), EuroSciPy, SciPy Индия, PyData, и другие. Мы настоятельно вам рекомендуем посетить одну из них и встретиться с авторами лучшего научного программного обеспечения в мире Python. Если же у вас нет возможности попасть на эти конференции или же вы просто хотите прочувствовать их характер, воспользуйтесь Сетью<sup>3</sup>, где участники этих конференций публикуют свои дискуссии.

## Бесплатное программное обеспечение и программное обеспечение с открытым исходным кодом (FOSS)

Сообщество SciPy приветствует разработку программного обеспечения с открытым исходным кодом. Исходный код почти всех библиотек SciPy находится

<sup>1</sup> См. <http://python-future.org/>.

<sup>2</sup> См. [http://python-notes.curiousefficiency.org/en/latest/python3/questions\\_and\\_answers.html](http://python-notes.curiousefficiency.org/en/latest/python3/questions_and_answers.html).

<sup>3</sup> См. <https://www.youtube.com/user/EnthoughtMedia/playlists>.

в свободном доступе для чтения, правки и повторного использования компонентов любым, кто в этом заинтересован.

Если вы хотите, чтобы другие разработчики использовали ваш программный код, то один из лучших способов этого добиться состоит в том, чтобы сделать его свободным и открытым. Если вы используете программное обеспечение с закрытым исходным кодом и получаете не тот результат, который вы хотели достигнуть, то вам не повезло. Вы можете послать разработчику электронное сообщение и попросить, чтобы он добавил новый функционал (что часто не срабатывает!), либо написать новое программное обеспечение самостоятельно. Если исходный код находится в открытом доступе, то вы с легкостью можете добавить или изменить его функциональность, используя приемы, которые вы узнаете из этой книги.

Таким же образом, если вы находите в компоненте программного обеспечения системную ошибку, то наличие доступа к исходному коду может в значительной степени облегчить работу как пользователя, так и разработчика. Даже если вы не вполне разбираетесь в исходном коде, вы, как правило, продвинетесь в диагностике возникшей проблемы гораздо дальше и поможете разработчику с ее исправлением. Обычно это полезный познавательный опыт для всех!

### ***Открытый исходный код, открытая наука***

В научном программировании все вышеупомянутые сценарии чрезвычайно распространены и важны: научное программное обеспечение часто строится на предыдущей работе либо ее видоизменяет самым интересным образом. А вследствие высокого темпа научных публикаций и прогресса большое количество программного кода остается без тщательного тестирования перед его выпуском, что приводит к незначительным или системным ошибкам.

Еще одна веская причина, чтобы сделать исходный код открытым, состоит в том, чтобы способствовать развитию производимых исследований. Многие из нас по своему опыту знают, когда, читая действительно актуальную исследовательскую работу и затем скачивая исходный код, чтобы проверить его на своих собственных данных, мы обнаруживаем: исполняемый файл не скомпилирован для вашей операционной системы, невозможно разобраться, как его запустить, код имеет ошибки, отсутствуют важные компоненты. Или мы вообще получаем неожиданные результаты. Делая научное программное обеспечение открытым, мы не только улучшаем качество этого программного обеспечения, но и позволяем ясно увидеть, каким образом был написан код, какие допущения были приняты и жестко запрограммированы. Открытый исходный код помогает решать многие из этих проблем. Он также позволяет другим ученым опираться на исходный код своих коллег, способствуя новому сотрудничеству и ускоряя научный прогресс.

### ***Лицензии на программное обеспечение с открытым исходным кодом***

Если вы хотите, чтобы ваш программный код использовали другие, то вы *должны* его лицензировать. Если вы его не лицензируете, то по умолчанию он будет

закрытым. Даже если вы свой код опубликуете (например, разместив его в публичном хранилище GitHub), то без лицензии на программное обеспечение ваш программный код никто не имеет права использовать, править или распространять.

Выбирая среди многих вариантов лицензирования, сначала необходимо решить, что именно вы хотите позволить людям делать с вашим исходным кодом. Предоставить людям право продавать ваш исходный код для получения прибыли? Или право продавать программное обеспечение, в котором используется ваш исходный код? Или же вы хотите ограничить использование своего исходного кода только бесплатным программным обеспечением?

Есть две широкие категории FOSS-лицензий (лицензий на свободное и открытое программное обеспечение, Free and Open Source Software):

- разрешительная лицензия;
- свободная лицензия (Copy-left).

Разрешительная лицензия означает, что вы предоставляете любому пользователю право использовать, править и распространять ваш исходный код любым способом, который ему нравится, включая применение вашего исходного кода в качестве коммерческого программного обеспечения. В этой категории популярными вариантами являются лицензии MIT (программная лицензия Массачусетского технологического института) и BSD (программная лицензия университета Беркли). Сообщество SciPy приняло новую лицензию BSD (так называемую «Модифицированную BSD», или «3-пунктовую лицензию BSD»). Использование такой лицензии подразумевает получение помощи относительно исходного кода от огромного количества людей, включая тех, кто трудится в информационной индустрии и стартапах.

Свободные лицензии также позволяют другим разработчикам использовать, править и распространять ваш исходный код. Вместе с тем эти лицензии еще предписывают, что производный исходный код должен распространяться в соответствии со свободной лицензией. Таким образом свободные лицензии ограничивают то, что именно пользователи могут делать с этим исходным кодом.

Самой популярной свободной лицензией является Публичная лицензия GNU или GPL. Главный ее недостаток связан с использованием свободной лицензии и состоит в том, что часто ваш исходный код становится недоступным для любых потенциальных пользователей или участников из частного сектора. И среди них в будущем можете оказаться вы сами! Как результат этот факт может существенно уменьшить вашу пользовательскую базу и, следовательно, успех вашего программного обеспечения. В науке это может означать меньшее количество цитирования.

Для получения более подробной справки относительно выбора лицензии обратитесь на веб-сайт, посвященный выбору лицензии Choose a License<sup>1</sup>. От-

<sup>1</sup> См. <http://choosealicense.com/>.



носителю лицензирования в научном контексте мы рекомендуем публикацию в блоге «The Whys and Hows of Licensing Scientific Code» (Вопросы «почему» и «как» относительно лицензирования научного программного кода) под авторством Джейка Вандерпласа (Jake VanderPlas), директора по исследованиям в области естествознания Вашингтонского университета и разносторонней суперзвезды SciPy. Собственно, здесь мы процитируем Джейка, чтобы убедительно довести до вас ключевые моменты лицензирования программного обеспечения:

...если вы извлечете из статьи всего три порции информации, то пусть они будут следующими:

1. Всегда лицензируйте свой код. Нелицензированный код является закрытым, поэтому любая открытая лицензия лучше, чем ничего (но см. п. 2).
2. Всегда используйте лицензию, совместимую с общей публичной лицензией (GPL). GPL-совместимые лицензии гарантируют вашему исходному коду широкую совместимость и включают GPL, новую BSD, MIT и другие (но см. п. 3).
3. Всегда используйте разрешительную лицензию в стиле BSD. Разрешительная лицензия, такая как новая лицензия BSD или лицензия MIT, более предпочтительна, по сравнению со свободной лицензией, такой как GPL или LGPL.

Весь исходный код в этой книге доступен в соответствии с 3-пунктовой лицензией BSD (3-clause BSD license). Там, где мы разместили фрагменты исходного кода других авторов. Этот исходный код обычно действует в соответствии с разрешительной открытой лицензией в той или иной форме (хотя не обязательно в соответствии с лицензией BSD).

Что касается вашего собственного исходного кода, то мы рекомендуем вам применять практику своего сообщества. В научном Python это означает 3-пунктовую лицензию BSD, в то время как в сообществе разработчиков на языке R, например, принята лицензия GPL.

## GitHub: исходный код в социальном пространстве

Мы немного поговорили о распространении своего исходного кода в соответствии с лицензией на открытое программное обеспечение. И будем надеяться, что огромное число людей будут скачивать ваш исходный код, использовать его, исправлять ошибки и добавлять новые свойства. В связи с этим возникает вопрос: где разместить свой исходный код, чтобы люди смогли его найти? Как эти исправления ошибок и новые свойства вернуться в ваш исходный код? Каким образом отслеживать все вопросы и изменения? Можно представить, как все это совсем скоро выйдет из-под контроля.

Присоединяйтесь к GitHub.

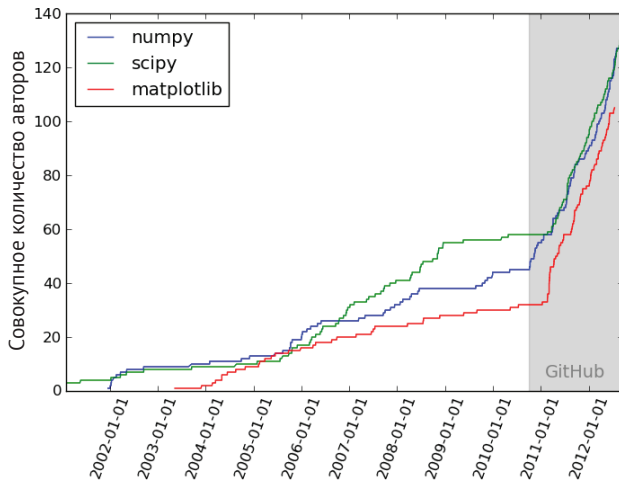
GitHub<sup>1</sup> – это веб-сайт для размещения, совместного использования и разработки исходного кода. В его основе лежит программная система управления

<sup>1</sup> См. <https://github.com/>.



версиями Git<sup>1</sup>. Обучению работе с GitHub посвящено несколько замечательных ресурсов, таких как «Введение в GitHub»<sup>2</sup> Питера Белла и Brenta Бира. Подавляющее большинство проектов в экосистеме SciPy размещено на GitHub, поэтому, безусловно, стоит научиться его использовать!

Веб-сайт GitHub оказал значительное влияние на участие разработчиков и внесение ими открытого исходного кода. Это было достигнуто за счет предоставления пользователям возможности публиковать исходный код и свободно сотрудничать. Любой участник может зайти и создать копию (так называемое ответвление *fork*) исходного кода и отредактировать, как его душе будет угодно. Он в конечном счете может внести эти изменения назад в оригинальный исходный код, создав *запрос на включение внесенных изменений*. Существует целый ряд хороших возможностей, таких как управление текущими вопросами и запросами на изменение, а также возможность определять, кто непосредственно может редактировать ваш исходный код. Вы даже можете отслеживать правки участников и получать другую детальную статистику. Помимо вышперечисленного, существует целый ряд других замечательных особенностей GitHub. Однако мы предоставим вам самим возможность обнаружить многие из них самостоятельно. С некоторыми из них вы познакомитесь, когда будете читать последующие главы. В сущности, GitHub демократизировал разработку программного обеспечения (рис. П.1) и существенно снизил барьер доступа к нему.



**Рис. П.1** ❖ Влияние GitHub  
(используется с разрешения автора, Джейка Вандерпласа)

<sup>1</sup> См. <http://git-scm.com/>.

<sup>2</sup> См. <http://shop.oreilly.com/product/0636920033059.do>.

## Оставьте свой след в экосистеме SciPy

По мере накопления опыта работы с SciPy и после того, как вы начнете использовать эту библиотеку в вашей исследовательской работе, может оказаться, что тому или иному пакету не хватает функционала, в котором вы нуждаетесь. Либо вы посчитаете, что можете сделать что-то эффективнее. А возможно, найдете ошибку. Когда вы дойдете до этого, значит, пришла пора начать принимать участие в экосистеме SciPy.

Мы настоятельно рекомендуем вам попробовать. Сообщество существует, потому что люди готовы делиться своим исходным кодом и совершенствовать существующий исходный код. И если каждый из нас внесет свой небольшой вклад, то вместе мы добьемся многого. Но помимо любых альтруистических причин содействия сообществу, существует несколько вполне практических личных выгод. Сотрудничая с сообществом, вы станете более профессиональным программистом. Любой вносимый вами исходный код будет рассмотрен другими участниками, и вы получите отзыв в виде замечаний и комментариев. В качестве побочного эффекта вы научитесь использовать Git и GitHub, которые сами по себе являются очень полезными инструментами для технической поддержки и совместного использования вашего собственного исходного кода. Вы даже можете обнаружить, что взаимодействие с сообществом SciPy предоставляет вам более широкую научную сеть и удивительные возможности карьерного роста.

Мы хотим, чтобы вы задумались над тем, чтобы стать больше, чем просто пользователем SciPy. Присоединившись к сообществу, ваша работа сделает его лучше для всех научных программистов.

## Капля эксцентричности Python

Если вы обеспокоены тем, что сообщество SciPy может стать для вновь прибывшего участника вызывающим благоговение местом, то следует напомнить, что это сообщество состоит из таких же людей, как и вы, ученых, которые, как правило, имеют прекрасное чувство юмора.

В стране Python вы неизбежно найдете ссылки на шоу «Монти Пайтон»<sup>1</sup>. Пакет Airspeed Velocity<sup>2</sup> измеряет быстродействие вашего программного обеспечения (подробнее о нем позднее) и ссылается на строчку из «Монти Пайтона и Священного Грааля» «какова воздушная скорость полета необремененной ласточки?».

А вот еще один забавно названный пакет – «Six». Он позволяет использовать пакеты Python 2 из Python 3. В имени пакета обыгрывается слово «six» (шесть),

<sup>1</sup> Монти Пайтон (англ. Monty Python) – комик-группа из Великобритании, состоявшая из шести человек. Благодаря своему новаторскому, абсурдистскому юмору участники «Монти Пайтон» находятся в числе самых влиятельных комиков всех времен. Группа известна во многом благодаря юмористическому телешоу «Летающий цирк Монти Пайтона». – *Прим. перев.*

<sup>2</sup> См. <http://spacetelescope.github.io/asv/using.html>.



а сам пакет позволяет использовать синтаксис Python 3 в Python 2 с новозеландским акцентом. Синтаксис `Sux` уменьшает разочарование от использования пакетов, предназначенных только для Python 2, после того как вы перешли на Python 3:

```
import sux
p = sux.to_use('my_py2_package')
```

В целом имена библиотек Python могут вызывать бурное веселье, и мы надеемся, что вы хорошо проведете время, когда будете придумывать свое собственное!

## ПОЛУЧЕНИЕ ПОМОЩИ

Когда мы попадаем в тупик, наш первый шаг – поискать в Интернете подсказку, которая поможет найти решение текущей задачи, либо информацию об ошибке, сообщение о которой мы получили. Эти поиски обычно приводят нас на [Stack Overflow](http://StackOverflow.com)<sup>1</sup>, превосходный вопросно-ответный сайт для программистов. Если вы с первого раза не нашли того, что ищете, то попробуйте обобщить критерии поиска, чтобы найти кого-то, кто имеет схожие проблемы.

Иногда вы можете оказаться фактически первым человеком, который задаст этот конкретный вопрос (это наиболее вероятно в тех случаях, когда вы используете совершенно новый пакет). Но не отчаивайтесь! Не все потеряно! Как было отмечено выше, сообщество SciPy дружелюбно и разбросано в различных частях межсетей. Ваш следующий шаг состоит в том, чтобы обратиться в поисковик с запросом «<имя библиотеки> список рассылки» и найти список рассылки с адресами, куда можно обратиться за помощью. Авторы библиотек и компетентные пользователи регулярно их читают и очень радушны ко вновь прибывшим участникам. Обратите внимание, что общепринятое правило поведения в сообществе требует, чтобы, перед тем как отправлять свои вопросы, вы *подписались* на список. Если вы не подпишетесь, это будет означать, что перед регистрацией вашего электронного адреса в списке кто-то должен будет вручную проверить, что данный адрес не является спамным. Присоединение к очередному списку рассылки может показаться ненужным, но мы настоятельно рекомендуем поступать именно так: это как раз то место, где надо учиться!

## ИНСТАЛЛЯЦИЯ ЯЗЫКА PYTHON

В настоящей книге мы исходим из того, что у вас уже установлен Python 3.6 (либо его более поздняя версия) и все необходимые пакеты SciPy. Мы перечислим все использованные нами необходимые библиотеки и их версии в файле `environment.yml`, прилагаемом вместе с файлами данных и исходными кодами

<sup>1</sup> См. <http://stackoverflow.com/>.

к этой книге. Самый легкий способ получить все необходимые компоненты – установить `conda`<sup>1</sup>, инструмент для управления средой разработки на Python. Затем вы можете передать файл `environment.yml` в `conda`, чтобы за один шаг установить правильные версии всех необходимых пакетов.

```
conda env create --name elegant-scipy -f путь/к/environment.yml
source activate elegant-scipy
```

За дополнительной информацией обратитесь к хранилищу книги на GitHub<sup>2</sup>.

## Доступ к книжным материалам

Весь исходный код и данные, приводимые в этой книге, доступны в нашем хранилище на GitHub. В файле README хранилища вы найдете инструкции по созданию блокнотов Jupyter из специальных исходных файлов с облегченной разметкой markdown. После этого блокноты Jupyter можно запускать в интерактивном режиме, используя данные, которые включены в хранилище.

## НАЧИНАЕМ

Мы выбрали самый элегантный исходный код, который предлагает сообщество SciPy. Мы также займемся исследованием нескольких реальных научных задач, решаемых при помощи SciPy. Эта книга еще дает возможность мимолетно взглянуть на радушное и готовое к сотрудничеству научное программистское сообщество, которое надеется, что вы присоединитесь.

Добро пожаловать в элегантный SciPy.



## УСЛОВНЫЕ ОБОЗНАЧЕНИЯ, ПРИНЯТЫЕ В КНИГЕ

В книге используются следующие типографские условные обозначения.

### *Курсивный шрифт*

Указывает новые термины, URL-адреса, адреса электронной почты, имена и расширения файлов.

### Моноширинный шрифт

Используется для листингов программ, а также внутри абзацев для отсылки на элементы программ, такие как переменные или имена функций, базы данных, типы данных, переменные окружающей среды, операторы и ключевые слова.

### **Жирный моноширинный шрифт**

Показывает команды либо другой текст, который должен быть напечатан самим пользователем.

<sup>1</sup> См. <http://conda.pydata.org/miniconda.html>.

<sup>2</sup> См. <https://github.com/elegant-scipy/elegant-scipy>.

### Курсивный моноширинный шрифт

Показывает текст, который должен быть заменен на предоставленные пользователем значения либо на значения, определяемые контекстом.

- ➔ Данный элемент обозначает общее замечание.
- ✓ Данный элемент обозначает подсказку или совет.
- ! Данный элемент обозначает предупреждение или предостережение.

## ИСПОЛЬЗОВАНИЕ ЦВЕТА

В некоторых приводимых в книге примерах используются различные цвета, которые не видны в печатной версии этой книги. Читателям печатной версии книги рекомендуется обратиться к исходным блокнотам на <http://elegant-scipy.org/>.



## ИСПОЛЬЗОВАНИЕ ПРИМЕРОВ ПРОГРАММ

Дополнительный материал (примеры программного кода, упражнения и т. д.) доступен для скачивания с <https://github.com/elegant-scipy/elegant-scipy>.

Эта книга предназначена, чтобы помочь вам в решении своих задач. В целом, если код примеров предлагается вместе с книгой, вы можете использовать его в своих программах и документации. Вам не нужно связываться с нами с просьбой о разрешении, если вы не воспроизводите значительную часть кода. Например, написание программы, которая использует несколько фрагментов кода из данной книги, официального разрешения не требует. Продажа либо распространение компакт-диска с примерами из книг издательства O'Reilly требует официального разрешения. Ответ на вопрос цитированием данной книги и приведение выдержек кода примеров в качестве цитат разрешения не требует. Включение значительного количества кода примеров из данной книги в документацию на ваш продукт требует официального разрешения.

Мы ценим, но не требуем атрибуции. Атрибуция обычно включает в себя титул, автора, издателя и ISBN. Например: «Нуньес-Иглесиас Х., Уолт ван дер Ш., Дэшноту Х. Элегантный SciPy. O'Reilly. ISBN 978-1-491-92287-3».

Если вы считаете, что применяете примеры кода, выходя за рамки их справедливого использования или разрешения, выданного выше, то обращайтесь к нам по адресу [permissions@oreilly.com](mailto:permissions@oreilly.com).

## БЛАГОДАРНОСТИ

Выражаем признательность многим и многим людям, которые внесли существенный вклад в эту книгу. Она бы не появилась без вашей помощи.

Прежде всего мы хотим поблагодарить многих участников, внесших свой вклад в разработку NumPy, SciPy и связанных с ними библиотек. Мы надеемся, что в этой книге мы отдали должное вашей удивительной работе.

Также благодарим многих участников, внесших свой вклад в развитие более широкой научной экосистемы Python, включая тех, кто обеспечил фундамент для нескольких наших глав: Винеша Бирокдара, Мэтта Роклин и Уоррена Векессер. Мы также должны поблагодарить тех, чей вклад мы не смогли включить к моменту публикации. Ваша работа нас вдохновила, и мы надеемся включить ее в будущие версии книги. Мы также благодарим Николаса Ругира за многие из его предположений, которые мы включили в качестве примеров и упражнений.

Другие предоставили нам данные и исходный код, который сэкономил нам часы поиска и изысканий. Мы благодарим Лэв Варшни за исходный программный код на MATLAB для макета спектрального графа мозга червя (главы 3 и 6) и Стефано Аллезина за данные о пищевой сети заповедника Сент-Марка (глава 6).

Мы обязаны всем, кто внес исправления и предложения во время предварительного показа книги, включая Билла Каца, Мэттиаса Бассоннира и Марка Хюн-ки Кима.

Мы благодарим наших технических рецензентов, Томаса Кэзуэлла, Нелли Вароко, Лэва Варшни и Грега Уилсон, которые великодушно нашли время в своих плотных графиках, чтобы причесать наши заключительные черновики и поделиться с нами своими экспертными оценками.

Хотя мы продолжим совершенствовать эту книгу на основе полученных от вас, наши читатели, комментариев, должны отдать должное нашим друзьям и семьям, которые корректировали самые ранние версии и предоставили ценные отзывы, предложения и поддержку. Малкольм Горман, Алисия Ошэк, ПВ Ван дер Уолт, Саймон Кокбек, Нелли Вароко и Ариэль Рокем: спасибо вам.

И конечно, мы благодарим наших редакторов издательства О'Рейли, Мэг Бланшетт, Брайана Макдональда и Нэн Барбер. Мы особенно благодарны Мэг, которая первая сделала нам предложение по поводу книги и предоставила бесценные указания на раннем этапе работы, когда у нас не было ни малейшего понятия, что мы должны делать.

## КОММЕНТАРИИ ПЕРЕВОДЧИКА

Весь материал настоящей книги протестирован в среде Windows 10. При тестировании исходного кода за основу взят Python версии 3.6.4 (время перевода книги – март 2018 г.).

Прилагаемый к книге адаптированный и скорректированный исходный код примеров лучше всего разместить в подпапке домашней папки пользователя (/home/python\_projects или C:\Users\[ИМЯ\_ПОЛЬЗОВАТЕЛЯ]\python\_projects). Ниже приведена структура папки с прилагаемыми примерами:

data	Файлы данных, используемые во всех главах и упражнениях
figures	Иллюстрации из книги
images	Фотографии, используемые в главах и упражнениях
markdown	Файлы с упрощенной разметкой глав книги на английском языке для генерирования блокнотов Jupyter
notes	Дополнительные примечания авторов с примерами исходного кода
pics	Изображения и графики, полученные в результате применения примеров программного кода
script	Дополнительные сценарии Python по темам книги
style	Стилевые файлы, используемые в примерах программного кода
tools	Дополнительные вспомогательные сценарии для работы с LaTeX, HTML и Jupyter
Глава 1 – глава 8, решения задач	Исходный код примеров в виде сценариев Python, блокнотов Jupyter и файлов HTML

## БАЗОВЫЙ НАБОР БИБЛИОТЕК ДЛЯ РАЗРАБОТЧИКА

В обычных условиях библиотеки Python можно скачать и установить из каталога библиотек Python PyPi (<https://pypi.python.org/>). Однако следует учесть, что для работы некоторых библиотек, в частности SciPy и Scikit-learn, в ОС Windows требуется, чтобы в системе была установлена библиотека Numpy+MKL. Библиотека **Numpy+MKL** привязана к библиотеке Intel® Math Kernel Library и включает в свой состав необходимые динамические библиотеки (DLL), расположенные в каталоге [numpy.org](http://numpy.org). Библиотеку Numpy+MKL следует скачать с хранилища whl-файлов на веб-странице Кристофа Голька из Лаборатории динамики флуоресценции Калифорнийского университета в г. Ирвайн (<http://www.lfd.uci.edu/~gohlke/pythonlibs/>) и установить при помощи менеджера пакетов pip как whl (соответствующая процедура установки пакетов в формате whl описана ниже). Например, для 64-разрядной операционной системы Windows и среды Python 3.6 команда будет такой:

```
pip install numpy-1.14.2+mkl-cp36-cp36m-win_amd64.whl
```

Подобный режим установки также касается библиотек scipy, scikit-image и scikit-learn. Стоит также отметить, что эти особенности установки не относятся к ОС Linux и Mac. Далее приводятся сведения об основополагающих библиотеках.

- **NumPy**, основополагающая библиотека, необходимая для научных вычислений на Python.
- **SciPy**, библиотека, используемая в математике, естественных науках и инженерном деле. Требует наличия numpy+mkl.
- **Matplotlib**, библиотека для работы с двумерными графиками.
- **Pandas**, инструмент для анализа структурных данных и временных рядов. Требует наличия numpy и некоторых других. Для чтения файлов Excel требует установки библиотеки xlrd.
- **Scikit-learn**, интегратор классических алгоритмов машинного обучения. Требует наличия numpy+mkl.

- **Jupyter**, интерактивная онлайн-вычислительная среда.
- **PyQt5**, библиотека инструментов для программирования графического интерфейса пользователя, требуется для работы инструментальной среды программирования Spyder.
- **Spyder**, инструментальная среда программирования на Python.

## НАБОР БИБЛИОТЕК, ИСПОЛЬЗУЕМЫХ В КНИГЕ

Ниже перечислены так называемые зависимости, т. е. библиотеки и их версии, от которых зависит приводимый в книге исходный код. Все они устанавливаются при помощи менеджера пакетов `pip` стандартным образом, за исключением первых четырех библиотек, особенности установки которых в ОС Windows были упомянуты выше:

- `numpy=1.12*`
- `scipy=0.19*`
- `scikit-learn=0.18*`
- `scikit-image=0.13*`
- `pandas=0.19*`
- `matplotlib=2.0*`
- `xlrd=1.0*`
- `jupyter=1.0*`
- `sympy=1.*`
- `toolz=0.8*`
- `six=1.10*`
- `beautifulsoup4=4.5*`
- `pillow`
- `lxml`
- `networkx`
- `notedown>=1.5`
- `jupytercontrib>=0.0.5`
- `bs4`



## ПРОТОКОЛ УСТАНОВКИ БИБЛИОТЕК

Если вы хотите устанавливать более свежие версии библиотек и держать весь процесс установки под своим контролем, то ниже предлагается список команд установки библиотек. В случае нескольких библиотек может потребоваться установка из файлов `whl`, которые можно взять из хранилища `whl`-файлов (<http://www.lfd.uci.edu/~gohlke/pythonlibs/>).

```
pip install --upgrade pip
pip install numpy
```

либо как `whl`: `pip install numpy-1.14.2+mkl-cp36-cp36m-win_amd64.whl`



```
pip install scipy
```

либо как whl: `pip install scipy-1.0.1-cp36-cp36m-win_amd64.whl`



```
pip install Scikit-learn
```

либо как whl: `pip install scikit_learn-0.19.1-cp36-cp36m-win_amd64.whl`

```
pip install scikit-image
```

либо как whl: `pip install scikit_image-0.13.1-cp36-cp36m-win_amd64.whl`

Все остальные библиотеки устанавливаются стандартным образом:

```
pip install pandas
```

```
pip install matplotlib
```

```
pip install jupyter
```

```
...
```

```
pip install pyqt5
```

```
pip install spyder
```



В зависимости от базовой ОС, версий языка Python и версий программных библиотек устанавливаемые вами версии whl-файлов могут отличаться от приведенных выше, где показаны последние на апрель 2018 г. версии для 64-разрядной ОС Windows и Python 3.6.4.

## УСТАНОВКА БИБЛИОТЕК PYTHON ИЗ WHL-ФАЙЛОВ

Библиотеки для Python можно разрабатывать не только на чистом Python. Довольно часто библиотеки пишутся на C (динамические библиотеки), и для них создается обертка Python, или же библиотека пишется на Python, но для оптимизации узких мест часть кода пишется на C. Такие библиотеки получаются очень быстрыми, однако библиотеки с вкраплениями кода на C программисту на Python тяжелее установить ввиду банального отсутствия соответствующих знаний либо необходимых компонентов и настроек в рабочей среде (в особенности в Windows). Для решения описанных проблем разработан специальный формат (файлы с расширением .whl) для распространения библиотек, который содержит заранее скомпилированную версию библиотеки со всеми ее зависимостями. Формат whl поддерживается всеми основными платформами (Mac OS X, Linux, Windows).

Установка производится с помощью менеджера библиотек pip. В отличие от обычной установки командой `pip install <имя_библиотеки>`, вместо имени библиотеки указывается путь к whl-файлу `pip install <путь/к/whl_файлу>`. Например:

```
pip install C:\temp\scipy-1.0.1-cp36-cp36m-win_amd64.whl
```

Откройте окно командой строки и при помощи команды `cd` перейдите в каталог, где размещен ваш whl-файл. Скопируйте в этот каталог имя вашего whl-файла. В этом случае полный путь указывать не понадобится. Например:

```
pip install scipy-1.0.1-cp36-cp36m-win_amd64.whl
```

При выборе библиотеки важно, чтобы разрядность устанавливаемой библиотеки и разрядность интерпретатора совпадали. Пользователи Windows могут брать whl-файлы с веб-сайта Кристофа Голька, где библиотеки постоянно обновляются. В архиве содержатся все библиотеки, какие только будут нужны.

## УСТАНОВКА И НАСТРОЙКА ИНСТРУМЕНТАЛЬНОЙ СРЕДЫ SPYDER

Spyder – это инструментальная среда для научных вычислений для языка Python (Scientific Python Development Environment) для Windows, Mac OS X и Linux. Это простая, легковесная и бесплатная интерактивная среда разработки на Python, которая предлагает функционал, аналогичный среде разработки на MATLAB, включая готовые к использованию элементы интерфейса PyQt5 и PySide: редактор исходного кода, редактор массивов данных NumPy, редактор словарей, а также консоли Python и IPython и многое другое.

Чтобы установить среду Spyder в Ubuntu Linux, используя официальный менеджер библиотек, нужна всего одна команда:

```
sudo apt-get install spyder
```



Чтобы установить с использованием менеджера библиотек pip:

```
sudo apt-get install python-qt5 python-sphinx
sudo pip install spyder
```

Для обновления:

```
sudo pip install -U spyder
```

Установка среды Spyder в Fedora 27:

```
dnf install python-spyder
```

Установка среды Spyder в Windows:

```
pip install spyder
```

➔ Среда Spyder требует обязательной установки библиотеки PyQt5 (`pip install pyqt5`).

## БЛОКНОТЫ JUPYTER

Среди файлов с исходным кодом Python можно найти файлы с расширением .html и .ipynb. Последние – это файлы блокнотов интерактивной среды программирования Jupyter (<http://jupyter.org/>). Блокноты Jupyter позволяют иметь в одном месте исходный код, результаты выполнения исходного кода, графики данных и документацию, которая поддерживает синтаксис упрощенной разметки Markdown и мощный синтаксис LaTeX.

Интерактивная среда программирования Jupyter – это зонтичный проект, который наряду с Python предназначен для выполнения в обычном веб-браузере

небольших программ и фрагментов программного кода на других языках программирования, в том числе Julia, R и многих других (уже более 40 языков).

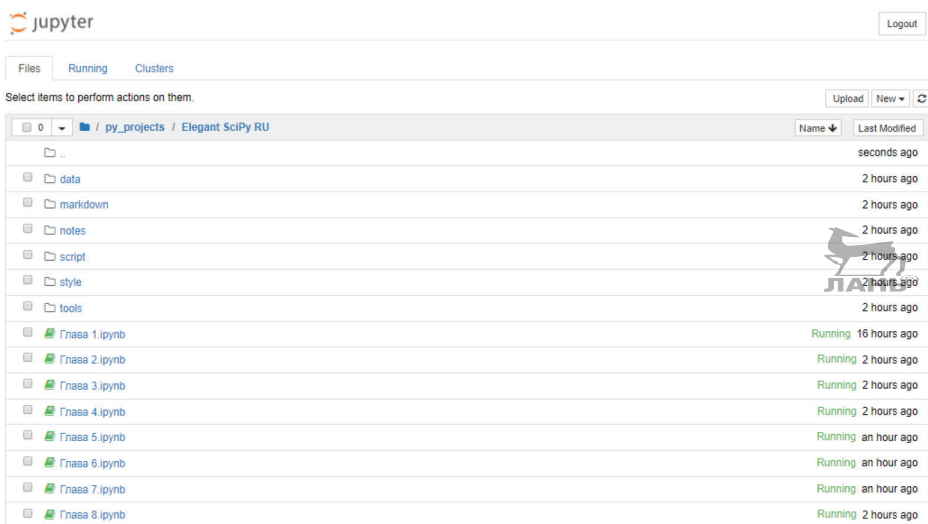
Интерактивная среда программирования Jupyter устанавливается, как обычно, при помощи менеджера пакетов `pip`.

```
pip install jupyter
```

Чтобы запустить интерактивную среду Jupyter, следует в командной оболочке или окне терминала набрать и исполнить приведенную ниже команду:

```
jupyter notebook
```

Локальный сервер интерактивной среды Jupyter запустится в браузере, заданном по умолчанию (как правило, по адресу <http://localhost:8888/>). После этого нужный блокнот Jupyter можно просто выбрать из меню Jupyter.



Файлы HTML представляют собой простые копии блокнотов Jupyter для просмотра в веб-браузере.

## ПРЕДСТАВЛЕНИЕ ЧИСЕЛ

Используемая в языке Python и в настоящей книге форма записи вещественных чисел и больших чисел с несколькими группами разрядов отличается от принятой в России. Ниже в таблице показаны эти отличия:

Разделитель групп разрядов		Десятичный разделитель	
США	Россия	США	Россия
запятая (,)	пробел	точка (.)	запятая (,)
65,535	65 535	3.14159	3,14159



Во время работы с Python вы часто будете встречать это разночтение – программируя на Python, вам придется иметь дело с отличающимся форматом записи чисел и сталкиваться с задачей форматирования чисел в локальное для России представление во время вывода результатов вычислений.

Для решения этой задачи, в качестве одного из вариантов, в Python имеется встроенная библиотека `locale`, которая позволяет настраивать форму представления чисел и соответствующим образом их форматировать. В общем случае, когда в программе используется форматирование чисел, следует начинать программу с приведенных ниже строк кода:

```
import locale # импортировать библиотеку по работе с локалями
loc = locale.getlocale() # запомнить текущую локаль
locale.setlocale(locale.LC_ALL, "Russian_Russia.1251") # изменить локаль
```

В последней строке задается форма написания чисел, принятая у нас в стране. Далее вы размещаете свой собственный программный код, к примеру:

```
cost = 50000.0 # стоимость в рублях
print('Стоимость изделия составляет ₺',
      locale.format('%.2f', cost, grouping=True),
      sep='')
```

И в конце программы следует вернуть локаль, которая была вначале:

```
locale.setlocale(locale.LC_ALL, loc) # вернуть локаль назад
```

Если выполнить этот фрагмент кода, то вы получите результат, в котором представление чисел будет соответствовать принятому у нас стандарту, т. е. с разделением групп разрядов пробелом и использованием в качестве десятичного разделителя запятой:

```
Стоимость изделия составляет ₺50 000,00
```

# Глава 1

---

## Элегантный NumPy: фундамент научного программирования на Python

[Библиотека NumPy] повсюду. Она окружает нас. Даже сейчас она с нами рядом. Ты видишь ее, когда смотришь в окно или включаешь телевизор. Ты ощущаешь ее, когда работаешь, идешь в церковь, когда платишь налоги.

– Морфейс, к/ф «Матрица»

Эта глава затрагивает некоторые статистические функции SciPy, однако особое внимание в ней уделено исследованию массива NumPy, структуры данных, которая лежит в основе почти всех численных научных вычислений в Python. Мы увидим, как операции с массивами NumPy позволяют создавать краткий и эффективный исходный код для управления числовыми данными.

В нашем случае мы будем использовать данные экспрессии генов из проекта «Атлас ракового генома» (The Cancer Genome Atlas, TCGA) для предсказания смертности среди больных раком кожи. Мы будем работать в направлении этой цели на протяжении всей этой главы и главы 2, по ходу знакомясь с некоторыми ключевыми понятиями SciPy. Прежде чем мы сможем предсказать смертность, мы должны нормализовать данные экспрессии, используя метод под названием «нормализация RPKM». Он позволяет сравнивать результаты измерений между различными образцами и генами. (Мы раскроем смысл понятия «экспрессия гена» буквально через мгновение.)

Чтобы вас заинтриговать и познакомить с идеями этой главы, давайте начнем с фрагмента кода. Так же, как и в других главах, мы начинаем с примера исходного кода, который, по нашему мнению, воплощает элегантность и мощь той или иной функции экосистемы SciPy. В данном случае мы хотим подчеркнуть правила векторизации и транслирования библиотеки NumPy, ко-



торые позволяют нам очень эффективно управлять массивами данных и делать о них выводы.

```
def rpkм(counts, lengths):
    """Вычислить прочтения на тысячу оснований экзона на миллион
    картированных прочтений (reads per kilobase transcript per million reads).

    RPKM = (10^9 * C) / (N * L)

    где:
    C = количество прочтений, картированных на ген
    N = суммы количеств картированных (выровненных) прочтений в эксперименте
    L = длина экзона в парах оснований для гена

    Параметры
    -----
    counts: массив, форма (N_genes, N_samples)
             РНК-сек (или подобные) количественные данные, где столбцы являются
             отдельными образцами и строки - генами.
    lengths: массив, форма (N_genes,)
             Длины генов в парах оснований в том же порядке, что и
             строки в counts.

    Возвращает
    -----
    normed: массив, форма (N_genes, N_samples)
            Матрица количеств counts, нормализованная согласно RPKM.

    «»»
    N = np.sum(counts, axis=0) # просуммировать каждый столбец, чтобы
                              # получить суммы количеств прочтений на образец

    L = lengths
    C = counts

    normed = 1e9 * C / (N[np.newaxis, :] * L[:, np.newaxis])

    return(normed)
```



Этот пример иллюстрирует несколько приемов, благодаря которым массивы NumPy могут сделать ваш код элегантнее:

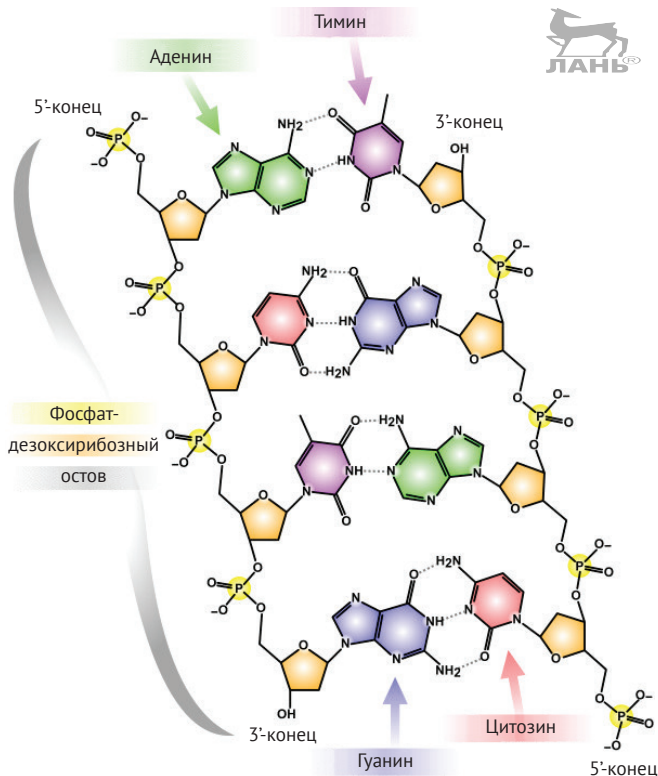
- массивы могут быть одномерными, как списки, но могут быть и двумерными, как матрицы, и даже более высокой размерности. Это позволяет представлять множество различных видов числовых данных. В нашем случае мы манипулируем двумерной матрицей;
- с массивами можно выполнять операции вдоль осей. В первой строке мы вычисляем сумму каждого столбца, задав ось `axis=0`;
- массивы позволяют выражать сразу несколько числовых операций. Например, ближе к концу функции мы делим двумерный массив количеств (`C`) на одномерный массив постолбцовых сумм (`N`). Такая операция называется транслированием. Дополнительная информация о том, как она работает, буквально через мгновение!

Прежде чем мы начнем вникать в мощные возможности NumPy, давайте потратим немного времени, чтобы разобраться в биологических данных, с которыми мы будем работать.

## ВВЕДЕНИЕ В ДАННЫЕ: ЧТО ТАКОЕ ЭКСПРЕССИЯ ГЕНА?

Мы построим свою работу, опираясь на *анализ экспрессии генов*, который позволит продемонстрировать силу библиотек NumPy и SciPy в решении реальной биологической задачи. Мы воспользуемся библиотекой pandas, которая опирается на NumPy, чтобы прочитать и преобразовать наши файлы данных, и затем мы будем эффективно манипулировать нашими данными в массивах NumPy.

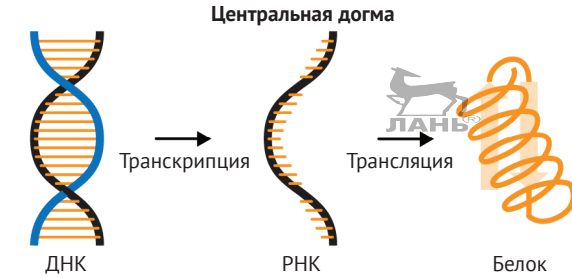
Так называемая *центральная догма молекулярной биологии*<sup>1</sup> постулирует, что вся информация, необходимая для анализа клетки (или в данном случае организма), хранится в молекуле, называемой дезоксирибонуклеиновой кислотой, или ДНК. Эта молекула имеет периодически повторяющийся остов, на котором лежат химические группы, именуемые *основаниями*, в последовательности (рис. 1.1). Имеется четыре вида оснований, сокращенно А, С, G и Т, составляющих алфавит, посредством которого сохраняется информация.



**Рис. 1.1** ❖ Химическая структура ДНК (автор Маделин Прайс Болл, изображение используется в соответствии с лицензией общего пользования ССО)

<sup>1</sup> См. [https://en.wikipedia.org/wiki/Central\\_dogma\\_of\\_molecular\\_biology](https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology).

Чтобы получить доступ к этой информации, ДНК *транскрибируется* в родственную молекулу, которая называется матричной рибонуклеиновой кислотой, или мРНК. Наконец, мРНК *транслируется* в белки, «рабочие лошадки» клетки (рис. 1.2). Участок ДНК, в котором закодирована информация, дающая белок (через мРНК), называется геном.



**Рис. 1.2** ❖ Центральная догма молекулярной биологии

Количество мРНК, получаемых из конкретного гена, называется *экспрессией* этого гена. В идеальном случае мы хотели бы измерить уровни белка. Однако это намного более трудная задача, чем измерение мРНК. К счастью, уровни экспрессии мРНК и уровни соответствующего ей белка обычно коррелируются<sup>1</sup>. Поэтому мы обычно измеряем уровни мРНК и на этом основании проводим наши исследования. Как вы увидите ниже, зачастую это не имеет значимой разницы, потому что уровни мРНК используются из-за их способности предсказывать биологические исходы, избавляя от необходимости делать определенно сформулированные утверждения о белках.

Важно отметить, что ДНК в каждой клетке вашего тела идентична. Поэтому различия между клетками являются результатом дифференциальной экспрессии этой ДНК на РНК: в разных клетках разные участки ДНК обрабатываются в нисходящие молекулы (рис. 1.3). Аналогичным образом, как мы увидим в этой и следующей главах, дифференциальная экспрессия может идентифицировать различные виды рака.

Современная технология измерения количества мРНК основывается на методе секвенирования (расшифровки) РНК, который носит сокращенное название РНК-сек (RNA-seq). РНК извлекается из образца ткани (например, в результате взятой у пациента биопсии), обратно транскрибируется в ДНК (чья стабильность выше) и затем прочитывается с использованием химически модифицированных оснований, которые светятся<sup>2</sup> при включении их в последовательность ДНК.

<sup>1</sup> См.: Майер Т., Гвелл М., Серрано Л. Корреляция мРНК и белка в составных биологических образцах (Tobias Maier, Marc Güell, and Luis Serrano. Correlation of mRNA and protein in complex biological samples. FEBS Letters 583, no. 24 (2009). <https://www.sciencedirect.com/science/article/pii/S0014579309008126>).

<sup>2</sup> См.: И все-таки ДНК светится // <https://22century.ru/biology-and-biotechnology/31442>. – Прим. перев.



В настоящее время высокопроизводительные секвенирующие машины способны прочитать лишь короткие фрагменты (как правило, приблизительно 100 оснований). Эти короткие последовательности называются «прочтениями»<sup>1</sup>. Мы измеряем миллионы прочтений и затем на основе их последовательностей подсчитываем, сколько прочтений пришло из каждого гена (рис. 1.4). Мы начнем наш анализ непосредственно с этих количественных данных.

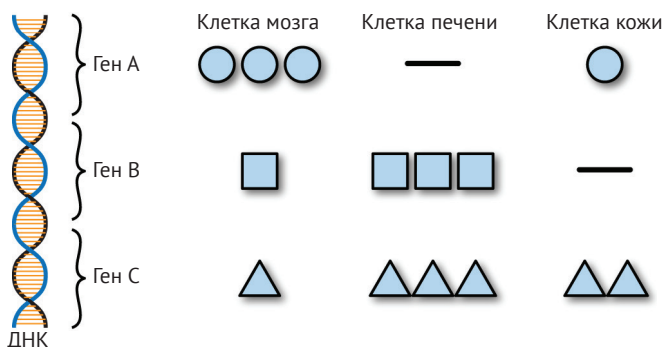


Рис. 1.3 ❖ Экспрессия гена

Таблица 1.1 показывает минимальный пример количественных данных об экспрессии генов.

Таблица 1.1. Количественные данные об экспрессии генов

	Тип клетки А	Тип клетки В
Ген 0	100	200
Ген 1	50	0
Ген 2	350	100

Эти данные представляют собой таблицу количеств прочтений в виде целых чисел, показывающих, сколько прочтений наблюдалось относительно каждого гена в каждом типе клетки. Вы можете заметить, насколько эти количества по каждому гену разнятся в зависимости от типов клетки. Эта информация может использоваться для того, чтобы узнать различия между этими двумя типами клеток.

Одним из способов представить эти данные в Python является список списков:

```
gene0 = [100, 200]
gene1 = [50, 0]
gene2 = [350, 100]
expression_data = [gene0, gene1, gene2]
```

<sup>1</sup> Прочтение, или рид (read) – это отсеквенированная последовательность коротких отрезков, полученных при разбиении молекулы, в данном случае молекулы мРНК. – Прим. перев.

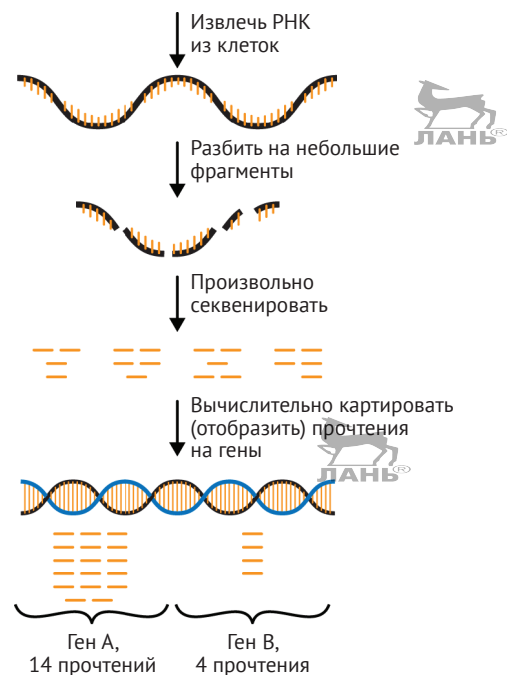


Рис. 1.4 ❖ Секвенирование РНК (РНК-сек)

Выше показано, что экспрессия каждого гена по разным типам клеток хранится в списке целых чисел Python. Затем мы сохраняем все эти списки в другом списке (метасписке, если быть точнее). При этом отдельные точки данных можно извлекать, используя два уровня индексации списка:

```
expression_data[2][0]
```

350

Этот способ хранения точек данных является весьма неэффективным вследствие характера работы интерпретатора Python. Прежде всего списки Python всегда являются списками *объектов*. Поэтому приведенный выше список `gene2` является не списком целых чисел, а списком *указателей* на целые числа, т. е. ненужными дополнительными издержками. Кроме того, такое представление означает, что каждый из этих списков и каждое из этих целых чисел в конечном итоге занимают совершенно разные случайные участки оперативной памяти вашего компьютера. Вместе с тем современные процессоры на практике предпочитают извлекать данные из памяти *блоками*, поэтому такое распределение данных по всей оперативной памяти является неэффективным.

Как раз эта проблема эффективно решается благодаря *массивам NumPy*.

## N-МЕРНЫЕ МАССИВЫ NUMPY

Одним из ключевых типов данных NumPy является N-мерный массив (`ndarray`, или просто массив). Массивы `ndarray` лежат в основе многих потрясающих методов управления данными в SciPy. В частности, мы займемся исследованием методов, которые позволяют писать мощный и элегантный программный код для управления данными, конкретно методы векторизации и транслирования.

Прежде всего давайте разберемся с массивом `ndarray`. Эти массивы должны быть гомогенными: все значения в массиве должны иметь один и тот же тип. В нашем случае мы должны хранить целые числа. Массивы `ndarray` называются N-мерными, потому что они могут иметь любое количество размерностей. Одномерный массив примерно эквивалентен списку Python:

```
import numpy as np

array1d = np.array([1, 2, 3, 4])
print(array1d)
print(type(array1d))

[1 2 3 4]
<class 'numpy.ndarray'>
```

Массивы имеют особые атрибуты и методы, к которым можно получить доступ, поставив точку после имени массива. Например, *форму* массива можно получить следующим образом:

```
print(array1d.shape)

(4,)
```

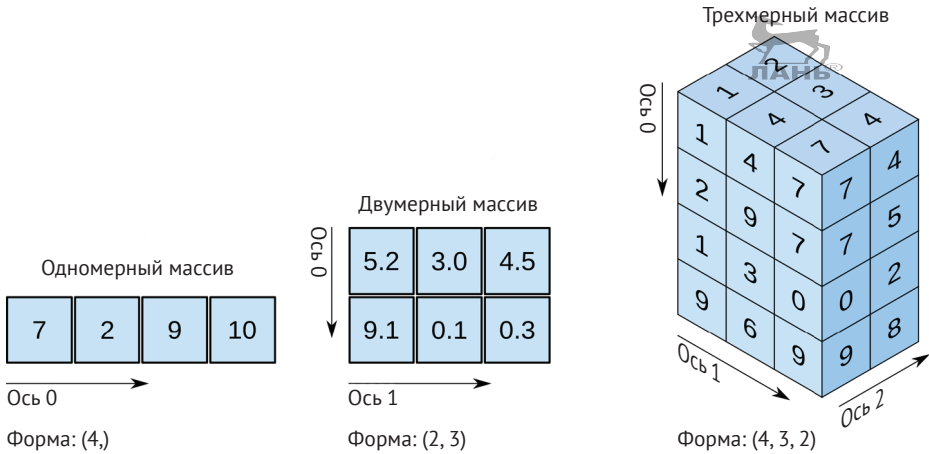
Здесь это просто кортеж с единственным числом. Вы можете спросить, почему просто не применить функцию `len`, как это было бы сделано в случае списка. И вы будете правы, это сработает. Однако на двумерные массивы этот прием распространяться не будет.

Вот что мы используем для представления данных в табл. 1.1:

```
array2d = np.array(expression_data)
print(array2d)
print(array2d.shape)
print(type(array2d))

[[100 200]
 [ 50  0]
 [350 100]]
(3, 2)
<class 'numpy.ndarray'>
```

Теперь вы видите, что атрибут `shape` обобщает функцию `len`, создавая отчет о величине многочисленных размерностей массива данных.



**Рис. 1.5** ❖ Визуализация массивов `ndarrays` NumPy в одной, двух и трех размерностях

Массивы имеют и другие атрибуты, такие как `ndim`, количество размерностей:

```
print(array2d.ndim)
```

2

Вы познакомитесь со всеми этими атрибутами, когда начнете широко использовать NumPy при выполнении своего собственного анализа данных.

Массивы NumPy могут представлять данные с еще большим количеством размерностей в таких случаях, как, например, данные магнитно-резонансной томографии (МРТ), которые включают результаты измерений внутри трехмерного объема. Если хранить значения МРТ во времени, то нам, возможно, понадобится четырехмерный массив NumPy.

Пока же мы будем придерживаться двумерных данных. В последующих главах будут введены многомерные данные с размерностями числом более двух, и вы научитесь писать программный код, который работает для данных любого количества размерностей.

## Зачем использовать массивы `ndarray` вместо списков Python?

Массивы имеют высокое быстродействие, потому что они задействуют векторизованные операции, написанные на низкоуровневом языке C. Эти операции работают на всем массиве в целом. Положим, у вас есть список, и вы хотите умножить каждый элемент в списке на пять. Стандартный подход Python состоит в написании цикла, который перебирает элементы списка и умножает каждый элемент на пять. Однако если вместо этого ваши данные представлены в виде массива, то вы можете одновременно умножить каждый элемент массива на пять. Высокооптимизированная библиотека NumPy максимально быстро выполнит эту итеративную обработку за кадром.

```
import numpy as np
```

```
# Создать массив ndarray целочисленных в диапазоне
# от 0 и до (но не включая) 1 000 000
array = np.arange(1e6)
```

```
# Конвертировать его в список
list_array = array.tolist()
```

Давайте сравним, сколько потребуется времени, чтобы умножить все значения в массиве на пять, воспользовавшись для этого волшебной функцией IPython `timeit`. Сначала возьмем данные, которые находятся в списке:

```
%timeit -n10 y = [val * 5 for val in list_array]
```

10 loops, average of 7: 102 ms +- 8.77 ms per loop (using standard deviation)

Теперь выполним ту же операцию с использованием встроенных *векторизованных* операций NumPy:

```
%timeit -n10 x = array * 5
```

10 loops, average of 7: 1.28 ms +- 206 µs per loop (using standard deviation)

Быстродействие более чем в 50 раз быстрее, а команда к тому же короче!

Массивы также эффективно экономят объем используемой оперативной памяти. В языке Python каждый элемент в списке является объектом, для которого выделяется порядочный участок оперативной памяти (что выглядит как расточительность). В массивах же каждый элемент занимает лишь необходимый объем оперативной памяти. Например, массив 64-разрядных целых чисел займет ровно 64 бита в расчете на элемент, плюс очень маленький расход на метаданные массива, такие как атрибут `shape`, который мы упоминали выше. Это, как правило, намного меньше, чем было бы выделено на объекты в списке Python. (Если вам интересно узнать, каким образом в Python работает процедура выделения оперативной памяти, обратитесь к публикации в блоге Джейка Вандерпласа «Почему Python медленный: взгляд изнутри»<sup>1</sup>.)

Кроме этого, при вычислениях с использованием массивов вы также можете использовать *срезы*, которые извлекают подмножество массива, *не копируя основные данные*.

```
# Создать массив ndarray x
x = np.array([1, 2, 3], np.int32)
print(x)
```

```
[1 2 3]
```

```
# Создать «срез» массива x
y = x[:2]
print(y)
```

```
[1 2]
```

<sup>1</sup> См. <https://jakevdp.github.io/blog/2014/05/09/why-python-is-slow/>.

```
# Назначить первому элементу среза у значение 6
```

```
y[0] = 6
print(y)
```

```
[6 2]
```

Обратите внимание: хотя мы редактировали только срез `y`, массив `x` тоже изменился, так как срез `y` ссылается на те же самые данные!

```
# Теперь первый элемент в массиве x поменялся на 6!
```

```
print(x)
```

```
[6 2 3]
```



Это означает, что во время работы с указателями на массив вам следует проявлять осторожность. Если вы хотите управлять данными, не касаясь оригинала, сделайте копию. Это очень легко сделать:

```
y = np.copy(x[:2])
```

## Векторизация

Ранее мы говорили о быстродействии операций с массивами. Одной из хитростей, которую NumPy использует для ускорения вычислений, является *векторизация*. Благодаря векторизации можно применять вычисление к каждому элементу в массиве без необходимости использования цикла `for`. В дополнение к ускорению вычислений векторизация может приводить к более естественному, удобочитаемому программному коду. Давайте рассмотрим несколько примеров.

```
x = np.array([1, 2, 3, 4])
print(x * 2)
```

```
[2 4 6 8]
```

В приведенном выше примере у нас массив `x` из 4 значений, и мы неявным образом умножили каждый элемент в `x` на одиночное значение, равное 2.

```
y = np.array([0, 1, 2, 1])
print(x + y)
```

```
[1 3 5 5]
```

Теперь мы сложили каждый элемент в `x` с соответствующим ему элементом в массиве `y` той же самой формы.

Обе эти операции просты и, надеемся, представляют собой интуитивно понятные примеры векторизации. Более того, NumPy выполняет их очень быстро, намного быстрее, чем итеративный обход массива в ручном режиме. (Вы можете смело поэкспериментировать с векторизацией самостоятельно, применив волшебную функцию IPython `%timeit`, которую мы упоминали ранее.)

## Транслирование

Одной из самых мощных и часто недооцененных особенностей массивов `ndarray` является операция трансляции. Трансляция – это прием вы-

полнения неявных операций между двумя массивами. Этот метод позволяет выполнять операции с массивами *совместимых* форм, создавать более крупные массивы, чем оба исходных массива. Например, мы можем вычислить внешнее векторное произведение<sup>1</sup> двух векторов, изменив их форму соответствующим образом:

```
x = np.array([1, 2, 3, 4])
x = np.reshape(x, (len(x), 1))
print(x)
```

```
[[1]
 [2]
 [3]
 [4]]
```

```
y = np.array([0, 1, 2, 1])
y = np.reshape(y, (1, len(y)))
print(y)
```

```
[[0 1 2 1]]
```

Две формы совместимы, когда по каждой размерности обе равны единице либо они совпадают друг с другом<sup>2</sup>.

Давайте проверим формы двух этих массивов.

```
print(x.shape)
print(y.shape)
```

```
(4, 1)
(1, 4)
```

Оба массива имеют две размерности, и внутренние размерности обоих массивов равняются 1, следовательно, размерности совместимы!

```
outer = x * y
print(outer)
```

```
[[0 1 2 1]
 [0 2 4 2]
 [0 3 6 3]
 [0 4 8 4]]
```

Внешние размерности говорят о размере результирующего массива. В нашем случае мы ожидаем получить массив (4, 4):

```
print(outer.shape)
```

```
(4, 4)
```

Вы можете сами убедиться, что  $outer[i, j] = x[i] * y[j]$  для всех  $(i, j)$ .

<sup>1</sup> См. [https://en.wikipedia.org/wiki/Outer\\_product](https://en.wikipedia.org/wiki/Outer_product).

<sup>2</sup> Мы всегда начинаем со сравнения последних размерностей и продвигаемся вперед, игнорируя избыточные размерности, в случае если размерность одного из массивов больше, чем размерность другого массива (например, (3, 5, 1) и (5, 8) совпадают).

Это было достигнуто за счет *правил трансляции* NumPy<sup>1</sup>, которые не-равно расширяют размерности величиной 1 в одном массиве, чтобы данная размерность совпадала с соответствующей размерностью другого массива. Не волнуйтесь, позже в этой главе мы поговорим об этих правилах подробнее.

Как мы убедимся в оставшейся части настоящей главы во время исследования реальных данных, операция трансляции имеет чрезвычайную ценность для реальных вычислений, связанных с массивами данных. Она позволяет выражать сложные операции сжато и эффективно.



## ИССЛЕДОВАНИЕ НАБОРА ДАННЫХ ЭКСПРЕССИИ ГЕНОВ

Используемый нами набор данных является экспериментом по РНК-секвенированию (РНК-сек) образцов рака кожи из проекта «Атлас ракового генома» (TCGA)<sup>2</sup>. Предварительно мы уже очистили и отсортировали данные, поэтому вы можете использовать файл *data/counts.txt* в хранилище настоящей книги.

В главе 2 мы будем использовать эти данные экспрессии генов, чтобы предсказать смертность среди больных раком кожи, и воспроизведем упрощенную версию рис. 5А и 5В<sup>3</sup> исследовательской работы<sup>4</sup> консорциума TCGA. Но сначала нам нужно разобраться в смещениях в наших данных и подумать о том, как их уменьшить.

### Чтение данных при помощи библиотеки pandas

Сначала мы воспользуемся библиотекой pandas, чтобы прочитать таблицу количеств. Заметим, pandas – это библиотека Python, предназначенная для управления данными и их анализа. При этом в этой библиотеке особый акцент делается на табличных данных и данных временных рядов. В нашем случае мы воспользуемся этой библиотекой, чтобы прочитать табличные данные смешанного типа. В этой библиотеке имеется тип DataFrame, представляющий собой гибкий табличный формат, основанный на объекте языка R, в котором данный формат носит название «фрейм данных». Например, данные, которые мы будем читать, включают в себя столбец имен генов (строковых значений) и многочисленные столбцы количеств (целых чисел). Поэтому было бы неправильным заносить их в гомогенный массив чисел. Хотя NumPy располагает некоторой поддержкой смешанных типов данных (реализованной за счет так называемых «структурированных массивов»), эта библиотека в основном не рассчитана на применение, усложняющее последующие операции.

При чтении данных в виде фрейма данных pandas мы поручаем библиотеке pandas выполнить их разбор, затем извлечь релевантную информацию и со-

<sup>1</sup> См. <https://docs.scipy.org/doc/numpy/user/basics.broadcasting.html>.

<sup>2</sup> См. <http://cancergenome.nih.gov/>.

<sup>3</sup> См. [http://www.cell.com/action/showImagesData?pii=S0092-8674\(15\)00634-0](http://www.cell.com/action/showImagesData?pii=S0092-8674(15)00634-0).

<sup>4</sup> См. [http://www.cell.com/cell/fulltext/S0092-8674\(15\)00634-0?\\_returnURL=http://linkinghub.elsevier.com/retrieve/pii/S0092867415006340%3Fshowall%3Dtrue](http://www.cell.com/cell/fulltext/S0092-8674(15)00634-0?_returnURL=http://linkinghub.elsevier.com/retrieve/pii/S0092867415006340%3Fshowall%3Dtrue).



хранить ее в более эффективном типе данных. Здесь мы используем pandas для быстрого импорта данных. В последующих главах мы познакомимся с pandas поближе, однако более подробную информацию вы можете получить, прочитав книгу «Python для анализа данных» (О’Рейли)<sup>1</sup>, написанную создателем библиотеки pandas Уэса Маккинни (Wes McKinney).

```
import numpy as np
import pandas as pd
```

```
# Импортировать данные TCGA по меланоме
```

```
filename = 'data/counts.txt'
```

```
with open(filename, 'rt') as f:
```

```
    data_table = pd.read_csv(f, index_col=0) # pandas выполняет разбор данных
```

```
print(data_table.iloc[:5, :5])
```

	00624286-41dd-476f-a63b-d2a5f484bb45	TCGA-FS-A1Z0	TCGA-D9-A3Z1 \	
A1BG	1272.36	452.96	288.06	
A1CF	0.00	0.00	0.00	
A2BP1	0.00	0.00	0.00	
A2LD1	164.38	552.43	201.83	
A2ML1	27.00	0.00	0.00	
	02c76d24-f1d2-4029-95b4-8be3bda8fdbe	TCGA-EB-A51B		
A1BG	400.11	420.46		
A1CF	1.00	0.00		
A2BP1	0.00	1.00		
A2LD1	165.12	95.75		
A2ML1	0.00	8.00		

Мы видим, что библиотека pandas любезно извлекла строку заголовка и использовала ее, чтобы назвать столбцы. Первый столбец дает имя каждому гену, а остальные столбцы представляют отдельные образцы.

Нам также будут нужны соответствующие метаданные, включая информацию об образце и длине генов.

```
# Имена образцов
```

```
samples = list(data_table.columns)
```

Информация о длине генов нам потребуется для нормализации. Поэтому, чтобы воспользоваться причудливой индексацией, принятой в pandas, мы зададим индексирование таблицы pandas по именам генов в первом столбце.

```
# Импортировать длины генов
```

```
filename = 'data/genes.csv'
```

```
with open(filename, 'rt') as f:
```

```
    # Разобрать файл при помощи pandas, индексировать по GeneSymbol
```

```
    gene_info = pd.read_csv(f, index_col=0)
```

```
print(gene_info.iloc[:5, :])
```

GeneSymbol	GeneID	GeneLength
CPA1	1357	1724

<sup>1</sup> См. <http://shop.oreilly.com/product/0636920050896.do>.

GUCY2D	3000	3623
UBC	7316	2687
C11orf95	65998	5581
ANKMY2	57037	2611

Давайте проверим, насколько хорошо наши данные о длине генов совпадают с данными о количествах экспрессии.

```
print("Гены в data_table: ", data_table.shape[0])
print("Гены в gene_info: ", gene_info.shape[0])
```

```
Гены в data_table: 20500
Гены в gene_info: 20503
```



В данных о длине генов имеется больше генов, чем фактически было изменено в эксперименте. Давайте выполним фильтрацию, чтобы получить только релевантные гены. При этом мы хотим удостовериться, что они находятся в том же самом порядке, что и в наших количественных данных. Как раз здесь пригодится индексированный функционал библиотеки pandas! Из наших двух источников данных можно получить пересечение имен генов и их использовать для индексации обоих наборов данных, тем самым гарантируя, что они будут иметь одинаковые гены, расположенные в том же самом порядке.

```
# Взять подмножество генной информации, которая
# совпадает с количественными данными
matched_index = pd.Index.intersection(data_table.index, gene_info.index)
```

Теперь давайте применим пересечение имен генов, чтобы проиндексировать количественные данные.

```
# Двумерный массив ndarray, содержащий количества экспрессии
# для каждого гена в каждом индивидууме
counts = np.asarray(data_table.loc[matched_index], dtype=int)
gene_names = np.array(matched_index)

# Проверить, сколько генов и индивидуумов измерено
print(f'{{counts.shape[0]}} генов измерено в {{counts.shape[1]}} индивидуумах.')
```

20500 генов измерено в 375 индивидуумах.

И данные с длинами генов:

```
# Одномерный массив ndarray, содержащий длины каждого гена
gene_lengths = np.asarray(gene_info.loc[matched_index]['GeneLength'],
                           dtype=int)
```

И теперь проверим размерности объектов:

```
print(counts.shape)
print(gene_lengths.shape)

(20500, 375)
(20500,)
```

Как и ожидалось, они теперь полностью совпадают!

## НОРМАЛИЗАЦИЯ

Реальные данные содержат самые разные виды артефактов измерений. Прежде чем выполнять какой-либо вид анализа данных, очень важно их рассмотреть и определить, является ли обоснованной какая-либо нормализация. Например, результаты измерения температуры цифровыми термометрами могут систематически отличаться от показаний ртутных термометров, читаемых человеком. Следовательно, сравнение образцов часто требует выполнения своего рода преобразования данных, благодаря которому каждый результат измерений будет приведен к общей шкале.

В нашем случае мы хотим удостовериться, что любые обнаруженные нами различия соответствуют реальным биологическим различиям и не относятся к техническому артефакту. Мы рассмотрим два уровня нормализации, которые часто совместно применяются к набору данных экспрессии генов: нормализацию между образцами (столбцами) и нормализацию между генами (строками).



### Нормализация между образцами

Например, количества для каждого индивидуума в экспериментах по РНК-сек могут существенно варьироваться. Давайте взглянем на распределение количеств экспрессии по всем генам. Сначала мы просуммируем столбцы и в результате получим суммы количеств экспрессии всех генов по каждому индивидууму, чтобы мы смогли просто рассмотреть вариацию между индивидуумами. Для визуализации распределения сумм количеств мы будем использовать метод *ядерной оценки плотности* (KDE), широко применяемый для сглаживания гистограмм, потому что этот метод дает более четкую картину лежащего в основе распределения.

Прежде чем начать, выполним небольшую настройку графика (которую мы будем делать в каждой главе). Смотрите ниже заметку «Короткое примечание по построению графиков» для получения подробной информации относительно каждой строки приведенного далее фрагмента кода.

```
# Заставить все графики в блокноте Jupyter
# в дальнейшем появляться локально
%matplotlib inline
# Применить к графикам собственный стилевой файл
import matplotlib.pyplot as plt
plt.style.use('style/elegant.mplstyle')
```

#### Краткое замечание по поводу построения графиков

Приведенный выше фрагмент кода делает несколько изящных трюков, которые придают нашим графикам более симпатичный вид.

Во-первых, строка `%matplotlib inline` представляет собой волшебную команду блокнота Jupyter<sup>1</sup>. Эта команда отображает все графики не во всплывающем окне, а в блокноте. Если блокнот Jupyter работает в интерактивном режиме, то вместо команды `%matplotlib inline` вы можете применить команду `%matplotlib notebook`. В результате вместо статического изображения вы получите интерактивное изображение каждого графика.

Во-вторых, мы импортируем модуль `matplotlib.pyplot` и затем сообщаем ему, чтобы он использовал наш собственный стиль создаваемых графиков `plt.style.use('style/elegant.mplstyle')`.

Аналогичный блок программного кода вы будете встречать в каждой главе перед первым графиком.

Вы, возможно, обратили внимание, что иногда разработчики импортируют существующие стили типа `plt.style.use('ggplot')`. Однако нам бы хотелось применить несколько конкретных параметров настройки. Желательно, чтобы все графики в этой книге соблюдали один и тот же стиль. Поэтому нами был собран наш собственный стиль Matplotlib. Чтобы увидеть, как мы это сделали, взгляните на файл таблицы стилей в хранилище книги «Элегантный SciPy»: `style/elegant.mplstyle`. Для получения дополнительной информации о стилях обратитесь к документации Matplotlib по таблицам стилей<sup>2</sup>.

Теперь вернемся к построению графика распределения количеств!

```
total_counts = np.sum(counts, axis=0) # просуммировать столбцы
                                     # (axis=1 будет суммировать строки)
```

```
from scipy import stats
```

```
# Применить гауссово сглаживание для оценки плотности
density = stats.kde.gaussian_kde(total_counts)
```

```
# Создать значения, для которых оценить плотность, с целью построения графика
x = np.arange(min(total_counts), max(total_counts), 10000)
```

```
# Создать график плотности
fig, ax = plt.subplots()
ax.plot(x, density(x))
ax.set_xlabel("Суммы количеств на индивидуум")
ax.set_ylabel("Плотность")
plt.show()
```

```
print(f'Количественная статистика:\n минимум: {np.min(total_counts)}'\n
      f'\n среднее: {np.mean(total_counts)}'\n
      f'\n максимум: {np.max(total_counts)}')
```

```
Количественная статистика:
минимум: 6231205
среднее: 52995255.33866667
максимум: 103219262
```

Мы видим, что суммы количеств экспрессии между самым низким и самым высоким индивидуумами разнятся на порядок (рис. 1.6). Это означает, что для

<sup>1</sup> См. <http://ipython.org/ipython-doc/dev/interactive/tutorial.html#magics-explained>.

<sup>2</sup> См. [https://matplotlib.org/users/style\\_sheets.html](https://matplotlib.org/users/style_sheets.html).

каждого индивидуума было сгенерировано разное число прочтений РНК-сек. Мы говорим, что эти индивидуумы имеют разные размеры библиотеки прочтений.

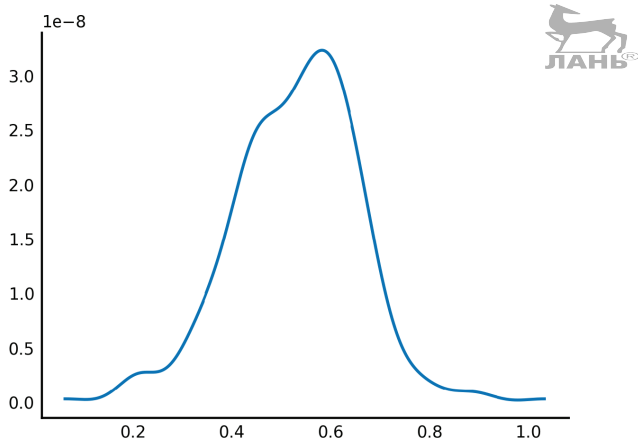


Рис. 1.6 ❖ График плотности количества экспрессии генов в расчете на индивидуум на основе сглаживания по методу ядерной оценки плотности (KDE)

### Нормализация размера библиотеки между образцами

Давайте взглянем поближе на диапазоны экспрессии генов для каждого индивидуума. При применении нормализации мы сможем увидеть ее в действии. Чтобы не слишком загрязнять результирующий график, извлечем случайную выборку, состоящую всего из 70 столбцов.

```
# Извлечь выборку для построения графика
np.random.seed(seed=7) # Задать начальное значение случайного числа,
                        # чтобы получить устойчивые результаты
# Случайно отобразить 70 образцов
samples_index = np.random.choice(range(counts.shape[1]), size=70, replace=False)
counts_subset = counts[:, samples_index]

# Индивидуальная настройка меток оси X, чтобы легче было читать графики
def reduce_xaxis_labels(ax, factor):
    «»Показать только каждую i-ю метку для предотвращения скапливания на
    оси X, например factor = 2 будет наносить каждую вторую метку оси X,
    начиная с первой.

    Параметры
    -----
    ax : ось графика matplotlib, подлежащая корректировке
    factor : int, коэффициент уменьшения числа меток оси X
    «»»
    plt.setp(ax.xaxis.get_ticklabels(), visible=False)
    for label in ax.xaxis.get_ticklabels()[factor-1::factor]:
        label.set_visible(True)

# Коробчатая диаграмма количества экспрессии на индивидуум
```



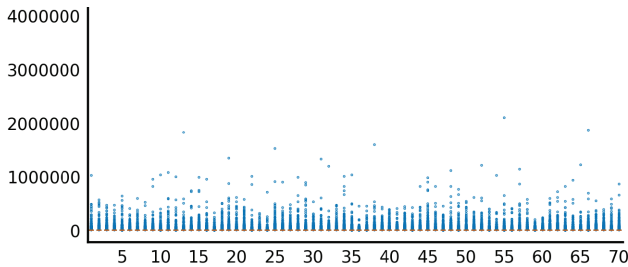
```
fig, ax = plt.subplots(figsize=(4.8, 2.4))
with plt.style.context('style/thinner.mplstyle'):
    ax.boxplot(counts_subset)
    ax.set_xlabel("Индивидуумы")
    ax.set_ylabel("Количества экспрессии генов")
    reduce_xaxis_labels(ax, 5)
```

Совершенно очевидно, что в верхнем конце шкалы экспрессии довольно много выбросов. Имеется большая вариация между индивидуумами, но их трудно заметить, так как все данные кластеризованы вокруг нуля (рис. 1.7). Поэтому давайте переведем наши данные в логарифмическую шкалу  $\log(n + 1)$ , чтобы их было легче рассмотреть (рис. 1.8). Чтобы выполнить функцию  $\log$  с шагом  $n + 1$ , следует использовать операцию транслирования, упрощающую программный код и ускоряющую обработку.

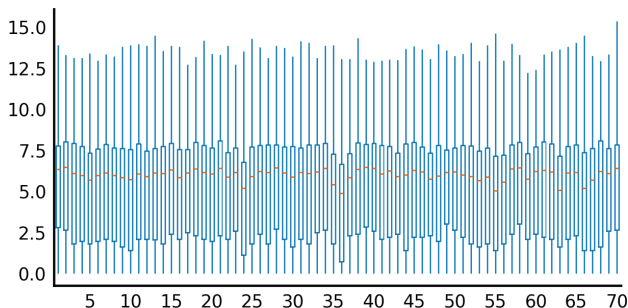
*# Коробчатая диаграмма количеств экспрессии генов на индивидуум*

```
fig, ax = plt.subplots(figsize=(4.8, 2.4))
```

```
with plt.style.context('style/thinner.mplstyle'):
    ax.boxplot(np.log(counts_subset + 1))
    ax.set_xlabel("Индивидуумы")
    ax.set_ylabel("Лог-количества экспрессии генов") # логарифмические кол-ва
    reduce_xaxis_labels(ax, 5)
```



**Рис. 1.7** ❖ Коробчатая диаграмма количества экспрессии генов на индивидуум



**Рис. 1.8** ❖ Коробчатая диаграмма количеств экспрессии генов на индивидуум (логарифмическая шкала)

Теперь давайте посмотрим, что происходит, когда мы выполняем нормализацию по размеру библиотеки (рис. 1.9).

```
# Нормализовать по размеру библиотеки
# Разделить количества экспрессии на суммы количеств
# для конкретного индивидуума
# Умножить на 1 миллион, чтобы вернуться к аналогичной шкале
counts_lib_norm = counts / total_counts * 1000000
# Обратите внимание, как здесь мы применили трансляцию дважды!
counts_subset_lib_norm = counts_lib_norm[:, samples_index]

# Коробчатая диаграмма количеств экспрессии на индивидуум
fig, ax = plt.subplots(figsize=(4.8, 2.4))

with plt.style.context('style/thinner.mplstyle'):
    ax.boxplot(np.log(counts_subset_lib_norm + 1))
    ax.set_xlabel("Индивидуумы")
    ax.set_ylabel("Лог-количества экспрессии генов")
    reduce_xaxis_labels(ax, 5)
```

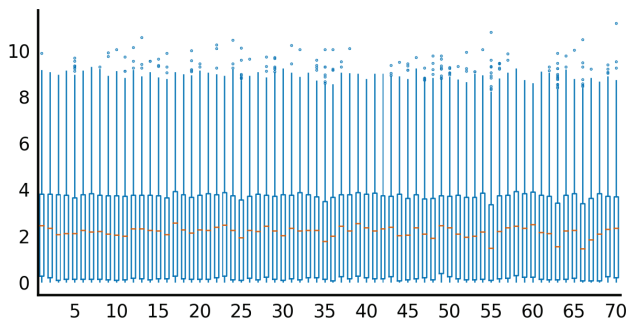


Рис. 1.9 ❖ Коробчатая диаграмма количеств экспрессии генов на индивидуум, нормализованных по библиотеке (логарифмическая шкала)

Теперь график выглядит намного лучше! Также обратите внимание: мы дважды применили трансляцию. Один раз, чтобы разделить все количества экспрессии генов на сумму для этого столбца, и потом еще раз, чтобы умножить все значения на 1 миллион.

Теперь давайте сравним наши нормализованные данные с необработанными данными.

```
import itertools as it
from collections import defaultdict

def class_boxplot(data, classes, colors=None, **kwargs):
    """Создать коробчатую диаграмму, в которой коробки расцвечены,
    согласно классу, к которому они принадлежат.
```

*Параметры*

-----

*data : массивоподобный список вещественных значений*

*Входные данные. Один коробчатый график будет сгенерирован для каждого элемента в `data`.*  
*classes : список строковых значений той же длины, что и `data`*  
*Класс, к которому принадлежит каждое распределение в `data`.*

*Другие параметры*

*kwargs : словарь*

*Именованные аргументы для передачи в `plt.boxplot`.*  
 """

```

all_classes = sorted(set(classes))
colors = plt.rcParams['axes.prop_cycle'].by_key()['color']
class2color = dict(zip(all_classes, it.cycle(colors)))

# Отобразить классы на векторы данных
# другие классы получают пустой список в этой позиции для смещения
class2data = defaultdict(list)
for distrib, cls in zip(data, classes):
    for c in all_classes:
        class2data[c].append([])
    class2data[cls][-1] = distrib

# Затем по очереди построить каждый коробчатый график
# с соответствующим цветом
fig, ax = plt.subplots()
lines = []
for cls in all_classes:
    # задать цвет для всех элементов коробчатого графика
    for key in ['boxprops', 'whiskerprops', 'flierprops']:
        kwargs.setdefault(key, {}).update(color=class2color[cls])
    # нарисовать коробчатый график
    box = ax.boxplot(class2data[cls], **kwargs)
    lines.append(box['whiskers'][0])
ax.legend(lines, all_classes)
return ax
    
```



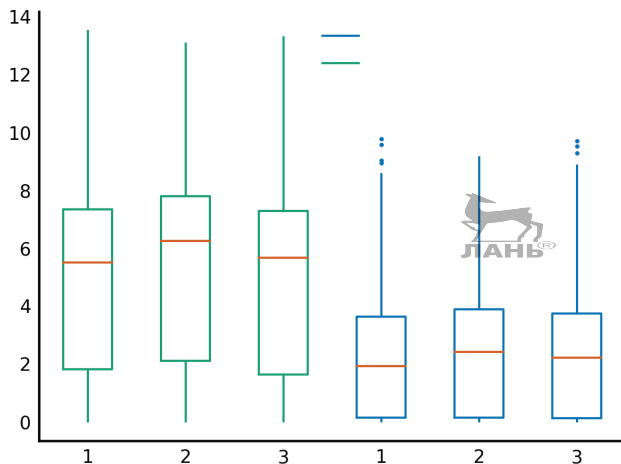
Теперь можно построить цветную коробчатую диаграмму, противопоставив нормализованные и ненормализованные образцы. Для иллюстрации мы показываем только три образца из каждого класса:

```

log_counts_3 = list(np.log(counts.T[:3] + 1))
log_ncounts_3 = list(np.log(counts_lib_norm.T[:3] + 1))
ax = class_boxplot(log_counts_3 + log_ncounts_3,
    ['сырые количества'] * 3 + ['нормализовано по размеру библиотеки'] * 3,
    labels=[1, 2, 3, 1, 2, 3])
ax.set_xlabel('номер образца')
ax.set_ylabel('лог-количества экспрессии генов');
    
```

Заметьте, когда мы учитываем размер библиотеки (сумму этих распределений), нормализованные распределения становятся чуть более похожими (рис. 1.10). Теперь между образцами мы сопоставляем подобное с подобным! Но как быть с различиями между генами?

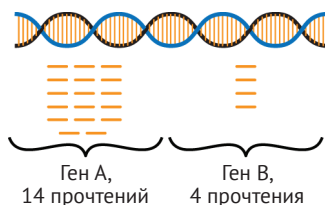




**Рис. 1.10** ❖ Сравнение необработанных и нормализованных по размеру библиотеки количество экспрессии генов в трех образцах (логарифмическая шкала)

## Нормализация между генами

Мы также можем получить неприятности, пытаясь сравнивать разные гены. Количества по конкретному гену связаны с длиной гена. Предположим, что ген В в два раза длиннее гена А. Оба экспрессируются в образце на аналогичных уровнях (т. е. оба производят аналогичное число молекул мРНК). Напомним, что в эксперименте РНК-сек мы подразделяем транскрипты на фрагменты и выборочно отбираем прочтения из пула фрагментов. Поэтому, если ген будет вдвое длиннее, он произведет вдвое больше фрагментов, и мы с удвоенной вероятностью отберем именно его. Следовательно, мы ожидаем, что ген В будет иметь вдвое больше количеств, чем ген А (рис. 1.11). Если мы хотим сравнивать уровни экспрессии разных генов, то нам следует выполнить еще одну нормализацию.



**Рис. 1.11** ❖ Связь между количествами и длиной генов

Давайте посмотрим, работает ли в нашем наборе данных связь между количествами и длиной генов. Сначала мы определим служебную функцию для построения графика:

```
def binned_boxplot(x, y, *, # относится только к Python 3! (*см. совет ниже)
                  xlabel='длина гена (логарифмическая шкала)',
                  ylabel='средние логарифмические количества'):
    «»»Построить график распределения `y` независимо от `x`, используя
        большое число коробчатых графиков.
        Примечание: ожидается, что все входные данные приведены в
        логарифмическую шкалу.

    Параметры
    -----
    x: Одномерный массив вещественных значений
        Значения независимых переменных.
    y: Одномерный массив вещественных значений
        Значения зависимых переменных.
    «»»
    # Определить интервалы для `x` в зависимости от плотности
    # результатов наблюдений
    x_hist, x_bins = np.histogram(x, bins='auto')

    # Применить `np.digitize` для нумерации интервалов
    # Отбросить последний край интервала, так как он нарушает допущение
    # метода `digitize` об открытости справа. Максимальный результат наблюдения
    # правильно попадает в последний интервал.
    x_bin_idx = np.digitize(x, x_bins[:-1])

    # Применить эти индексы для создания списка массивов, где каждый содержит
    # значения `y`, соответствующие значениям `x` в последнем интервале.
    # Этот формат входных данных ожидается на входе в `plt.boxplot`
    binned_y = [y[x_bin_idx == i]
                 for i in range(np.max(x_bin_idx))]
    fig, ax = plt.subplots(figsize=(4.8,1))

    # Создать метки оси X, используя центры интервалов
    x_bin_centers = (x_bins[1:] + x_bins[:-1]) / 2
    x_ticklabels = np.round(np.exp(x_bin_centers)).astype(int)

    # Создать коробчатую диаграмму
    ax.boxplot(binned_y, labels=x_ticklabels)

    # Показать только каждую 10-ю метку, чтобы
    # предотвратить скапливание на оси X
    reduce_xaxis_labels(ax, 10)

    # Скорректировать имена осей
    ax.set_xlabel(xlabel)
    ax.set_ylabel(ylabel);
```



Теперь мы вычислим длины генов и количества экспрессии:

```
log_counts = np.log(counts_lib_norm + 1)
mean_log_counts = np.mean(log_counts, axis=1) # по всем образцам
log_gene_lengths = np.log(gene_lengths)

with plt.style.context('style/thinner.mplstyle'):
    binned_boxplot(x=log_gene_lengths, y=mean_log_counts)
```



### Совет по Python 3: использование символа \* для создания только именованных аргументов

Начиная с версии 3.0 Python допускает использование «только именованных аргументов»<sup>1</sup>. Это такие аргументы, которые необходимо вызывать лишь с использованием ключевого слова, не полагаясь только на позицию. Например, только что написанную функцию `binned_boxplot` можно вызвать следующим образом:

```
>>> binned_boxplot(x, y, xlabel='my x label', ylabel='my y label')
```

но не так, как показано ниже. Это было бы допустимо в Python 2, но вызовет ошибку в Python 3:

```
>>> binned_boxplot(x, y, 'my x label', 'my y label')
```

```
-----
TypeError                                 Traceback (most recent call last)
<ipython-input-58-7a118d2d5750in <module>()
      1 x_vals = [1, 2, 3, 4, 5]
      2 y_vals = [1, 2, 3, 4, 5]
----> 3 binned_boxplot(x, y, 'my x label', 'my y label')
```

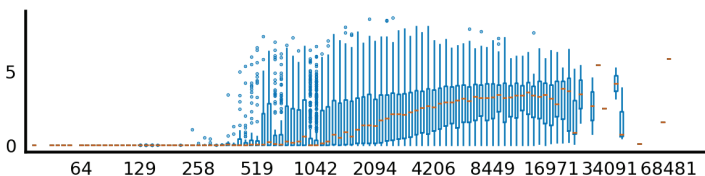
TypeError: binned\_boxplot() takes 2 positional arguments but 4 were given

Идея состоит в том, чтобы оградить вас от чего-то, похожего на это:  
`binned_boxplot(x, y, 'моя метка')`

Данная команда приведет к тому, что ваша метка `y` будет на оси `X`, т.е. вызовет распространенную ошибку для сигнатур со многими необязательными параметрами, которые не имеют очевидного порядка следования.



На приведенном ниже изображении мы видим, чем длиннее ген, тем выше его измеренные количества! Как отмечалось ранее, это артефакт метода, а не биологический сигнал! Как это объяснить?



## Нормализация по образцам и генам: RPKM

Одним из самых простых методов нормализации, применяемых для данных РНК-сек, является метод определения показателя RPKM, т.е. определение количества прочтений на тысячу оснований экзона на миллион картированных

<sup>1</sup> См. <https://www.python.org/dev/peps/pep-3102/>.

прочтений (reads per kilobase transcript per million reads)<sup>1</sup>. Показатель RPKM соединяет идеи нормализации по образцу и по гену. Когда мы вычисляем показатель RPKM, мы нормализуем размер библиотеки (сумму каждого столбца) и длину гена.

Чтобы разобраться в том, каким образом рассчитывается показатель RPKM, давайте определим следующие ниже величины:

- C = количества прочтений, картированных на ген;
- L = длина экзона в парах оснований для гена;
- N = суммы количеств картированных прочтений в эксперименте.

Давайте сначала вычислим прочтения на килобазу, или на тысячу оснований экзона.

Прочтения в расчете на основание выполняются следующим образом:

$$\frac{C}{L}.$$

Эта формула запрашивает прочтения в расчете на тысячу оснований вместо прочтений в расчете на основание. Одна килобаза = 1000 оснований, поэтому нам нужно разделить длину (L) на 1000.

Прочтения в расчете на тысячу оснований экзона вычисляются следующим образом:

$$\frac{C}{L} \frac{10^3 C}{L}.$$

Далее нам нужно выполнить нормализацию по размеру библиотеки. Если просто поделить на количество картированных прочтений, мы получим:

$$\frac{10^3 C}{LN}.$$

Но биологи предпочитают размышлять в миллионах прочтений, с тем чтобы числа не становились слишком большими. Рассчитав на миллион прочтений, мы получим:

$$\frac{10^3 C}{L \left( \frac{N}{10^6} \right)} = \frac{10^9 C}{LN}.$$

Таким образом, формула вычисления прочтений на тысячу оснований экзона на миллион картированных прочтений будет такой:

$$RPKM = \frac{10^9 C}{LN}.$$

<sup>1</sup> Показатель RPKM получают при помощи нормирования в пределах образца, который удаляет эффекты длины гена и размера библиотеки. – Прим. перев.

Теперь давайте реализуем вычисление показателя RPKM по всему массиву количеств.

```
# Создать переменные в соответствии с формулой RPKM, чтобы легче было сравнивать
C = counts
N = counts.sum(axis=0) # просуммировать каждый столбец, чтобы получить суммы
                        # прочтений на образце
L = gene_lengths      # длины для каждого гена, совпадающего со строками в `C`
```



Сначала мы умножаем на  $10^9$ . Поскольку переменная *C* является массивом `ndarray`, мы можем применить операцию транслирования. Если умножить массив `ndarray` на одиночное значение, то это значение будет транслировано по всему массиву.

```
# Умножить все количества на 10^9
C_tmp = 10^9 * C
```



Далее нам нужно разделить на длину гена. Процесс транслирования одиночного значения по всему двумерному массиву был довольно понятным. Мы просто умножали каждый элемент в массиве на заданное значение. Но что происходит, когда нам нужно разделить двумерный массив на одномерный?

### Правила транслирования

Операция транслирования позволяет выполнять вычисления между массивами `ndarray`. Как говорилось ранее, у этих массивов разные формы. Для того чтобы сделать эти манипуляции немного легче, в библиотеке NumPy используются правила транслирования. Когда два массива имеют одинаковые размерности, операция транслирования может выполняться, если размеры каждой размерности совпадают либо одна из них равна 1. Если массивы имеют разные размерности, то в начало более короткого массива добавляется (1,) до тех пор, пока размерности не будут совпадать, и затем применяются стандартные правила транслирования.

Например, предположим, что у нас два массива `ndarray`, *A* и *B*, с формами (5,2) и (2,). Мы определяем произведение  $A * B$ , используя транслирование. Размерность массива *B* меньше, чем массива *A*. Поэтому во время вычисления новая размерность добавляется в начало *B* со значением 1, в результате чего новая форма массива *B* становится (1,2). Наконец, там, где форма массива *B* не совпадает с формой массива *A*, она *умножается* путем накапливания достаточного количества версий массива *B*, давая в итоге форму (5,2). Это действие выполняется «виртуально», не расходуя дополнительную оперативную память. В этой точке произведение представляет собой поэлементное умножение, давая выходной массив той же самой формы, что и у массива *A*.

Теперь предположим, у нас есть еще один массив, *C*, с формой (2,5). Чтобы умножить *C* на *B* (или их сложить), мы можем попытаться добавить (1, ) в начало формы массива *B*, но в этом случае мы по-прежнему в итоге получим несовместимые формы: (2,5) и (1,2). Чтобы выполнить транслирование массивов, мы должны вручную добавить размерность *в конец* массива *B*. Тогда мы в итоге

получим формы (2,5) и (2,1), в результате чего может быть осуществлена операция транслирования.

В NumPy можно явным образом добавить новую размерность в массив B, применив `np.newaxis`. Давайте посмотрим, как это делается в нашей нормализации на основе показателя RPKM.

Сначала взглянем на размеры наших массивов.

```
print('C_tmp.shape', C_tmp.shape)
print('L.shape', L.shape)
```

```
C_tmp.shape (20500, 375)
L.shape (20500,)
```

Мы видим, что массив `C_tmp` имеет две размерности, тогда как `L` имеет одну. Поэтому во время транслирования в начало массива `L` будет добавлена еще одна размерность. В результате мы получим:

```
C_tmp.shape (20500, 375)
L.shape (1, 20500)
```

Размеры не будут совпадать! Мы же хотим транслировать массив `L` на первую размерность массива `C_tmp`, поэтому нам следует скорректировать размерности массива `L` самостоятельно.

```
L = L[:, np.newaxis] # добавить размерность в L со значением 1
print('C_tmp.shape', C_tmp.shape)
print('L.shape', L.shape)
```

```
C_tmp.shape (20500, 375)
L.shape (20500, 1)
```

Теперь, когда наши размерности совпадают или равны единице, мы можем выполнить транслирование.

```
# Разделить каждую строку на длину гена для этого гена (L)
C_tmp = C_tmp / L
```

Наконец, мы должны выполнить нормализацию по размеру библиотеки, т. е. по сумме количеств для этого столбца. Напомним, что мы уже вычислили `N` при помощи:

```
N = counts.sum(axis=0) # просуммировать каждый столбец, чтобы получить суммы
# количеств прочтений на образец
```

```
# Проверить формы массивов C_tmp и N
print('C_tmp.shape', C_tmp.shape)
print('N.shape', N.shape)
```

```
C_tmp.shape (20500, 375)
N.shape (375,)
```

При запуске трансляции в начало массива `N` будет добавлена дополнительная размерность:

```
N.shape (1, 375)
```

Размерности будут совпадать, и поэтому нам ничего делать не нужно. Тем не менее для удобочитаемости полезно добавить в `N` дополнительную размерность.

*# Добавить в `N` дополнительную размерность*

```
N = N[np.newaxis, :]
print('C_tmp.shape', C_tmp.shape)
print('N.shape', N.shape)
```

```
C_tmp.shape (20500, 375)
N.shape (1, 375)
```

*# Разделить каждый столбец на суммы количеств прочтений для этого столбца (`N`)*

```
rpkm_counts = C_tmp / N
```

Давайте поместим этот программный код в функцию, чтобы его можно было использовать повторно.

```
def rpkm(counts, lengths):
```

*«»»Вычислить прочтения на тысячу оснований экзона на миллион картированных прочтений.*

$$RPKM = (10^9 * C) / (N * L)$$

*где:*

*`C` = количества прочтений, картированных на ген*

*`N` = суммы количеств картированных (выровненных) прочтений в эксперименте*

*`L` = длина экзона в парах оснований для гена*

*Параметры*

-----

*counts: массив, форма (`N_genes`, `N_samples`)*

*РНК-сек (или подобные) количественные данные, где столбцы являются отдельными образцами и строки являются генами.*

*lengths: массив, форма (`N_genes`,)*

*Длины генов в парах оснований в том же порядке, что и строки в `counts`.*

*Возвращает*

-----

*normed: массив, форма (`N_genes`, `N_samples`)*

*Матрица `counts`, нормализованная согласно RPKM.*

*«»»*

```
N = np.sum(counts, axis=0) # просуммировать каждый столбец, чтобы
                          # получить суммы количеств прочтений на образец
```

```
L = lengths
```

```
C = counts
```

```
normed = 1e9 * C / (N[np.newaxis, :] * L[:, np.newaxis])
```

```
return(normed)
```

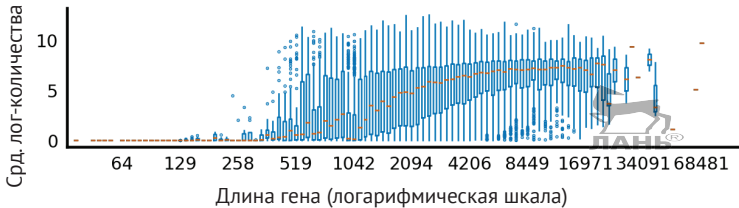
```
counts_rpkm = rpkm(counts, gene_lengths)
```

### **RPKM между нормализацией генов**

Давайте посмотрим на влияние нормализации RPKM в действии. Сначала, в качестве напоминания, посмотрите распределение средних логарифмических количеств как функции длины генов (см. рис. 1.12):

```
log_counts = np.log(counts + 1)
mean_log_counts = np.mean(log_counts, axis=1)
log_gene_lengths = np.log(gene_lengths)

with plt.style.context('style/thinner.mplstyle'):
    binned_boxplot(x=log_gene_lengths, y=mean_log_counts)
```

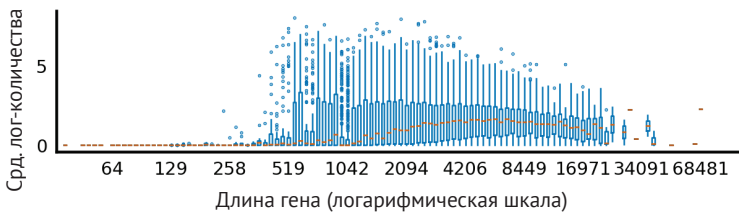


**Рис. 1.12** ❖ Связь между длиной гена и средней экспрессией до нормализации RPKM (логарифмическая шкала)

Теперь тот же самый график с нормализованными по RPKM значениями:

```
log_counts = np.log(counts_rpkм + 1)
mean_log_counts = np.mean(log_counts, axis=1)
log_gene_lengths = np.log(gene_lengths)

with plt.style.context('style/thinner.mplstyle'):
    binned_boxplot(x=log_gene_lengths, y=mean_log_counts)
```



**Рис. 1.13** ❖ Связь между длиной гена и средней экспрессией после нормализации RPKM (логарифмическая шкала)

Вы видите, что средние количества экспрессии значительно сгладились, в особенности для генов длиной более 3000 пар оснований. (Гены меньшей длины, похоже, по-прежнему имеют низкую экспрессию – она может быть слишком малой для статистической мощности метода RPKM.)

Нормализация RPKM может быть полезной для сравнения профиля экспрессии разных генов. Мы уже видели, что более длинные гены имеют на графике более высокие количества, но это не означает, что уровень их экспрессии на самом деле выше. Давайте выберем короткий ген и длинный ген, сравним их количества до и после нормализации RPKM и покажем, что мы имеем в виду.

```
gene_idxs = np.array([80, 186])
gene1, gene2 = gene_names[gene_idxs]
```



```

len1, len2 = gene_lengths[gene_idxs]
gene_labels = [f'{gene1}, {len1}bp', f'{gene2}, {len2}bp']

log_counts = list(np.log(counts[gene_idxs] + 1))
log_ncounts = list(np.log(counts_rpkм[gene_idxs] + 1))

ax = class_boxplot(log_counts,
                   ['сырые количества'] * 3,
                   labels=gene_labels)
ax.set_xlabel('Гены')
ax.set_ylabel('лог-количества экспрессии генов по всем образцам');

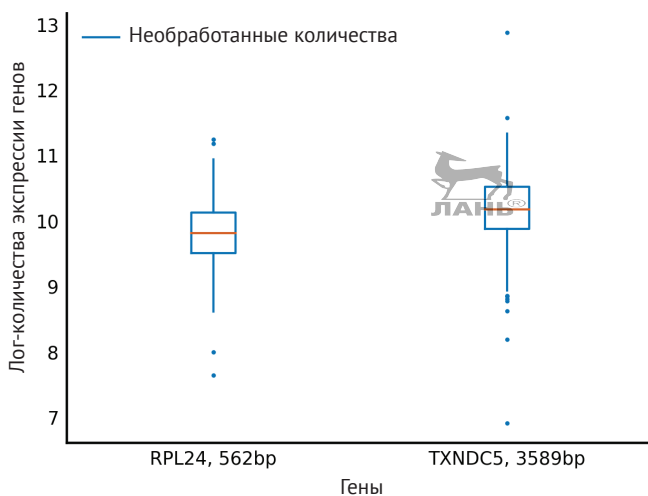
```

Если мы посмотрим только на необработанные количества, то, по-видимому, более длинный ген, TXNDC5, экспрессируется немного больше, чем более короткий, RPL24 (рис. 1.14). Однако после нормализации RPKM проявляется другая картина:

```

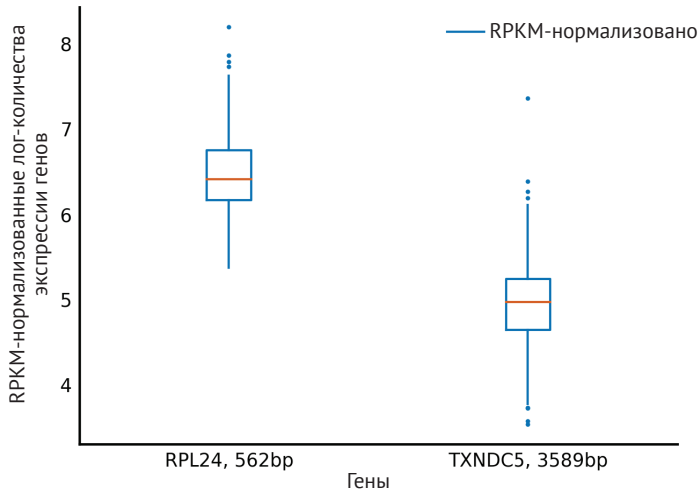
ax = class_boxplot(log_ncounts,
                   ['RPKM-нормализовано'] * 3,
                   labels=gene_labels)
ax.set_xlabel('Гены')
ax.set_ylabel('лог-количества экспрессии генов после RPKM');

```



**Рис. 1.14** ❖ Сравнение экспрессии двух генов до нормализации RPKM

Теперь похоже, что RPL24 фактически экспрессируется на гораздо более высоком уровне, чем TXNDC5 (см. рис. 1.15). Это вызвано тем, что RPKM содержит нормализацию длины гена. Теперь мы можем непосредственно выполнить сравнение между генами разной длины.



**Рис. 1.15** ❖ Сравнение экспрессии двух генов после нормализации RPKM

## ПОДВЕДЕНИЕ ИТОГОВ

На данный момент мы сделали следующее:

- импортировали данные, используя библиотеку `pandas`;
- познакомились с ключевым классом объектов NumPy – массивом `ndarray`;
- применили силу операции трансляции, которая сделала наши вычисления элегантнее.

В главе 2 мы продолжим работать с тем же самым набором данных, выполнив реализацию более сложного метода нормализации, затем применим кластеризацию, чтобы сделать предсказания относительно смертности среди больных раком кожи.

---

# Глава 2



## Квантильная нормализация с NumPy и SciPy

Не печальтесь, если вы не сможете сразу постичь более глубокие тайны Трехмерия. Постепенно они откроются перед вами.

– Эдвин Э. Эбботт.

«Флатландия: роман о четвертом измерении»

В этой главе мы продолжим анализировать данные экспрессии генов из главы 1, но немного с другой целью: мы хотим использовать *профиль экспрессии генов* каждого пациента (полный вектор замеров экспрессии его генов) для предсказания ожидаемой выживаемости. Чтобы использовать полные профили, нам нужна более глубокая нормализация, чем та, которую обеспечивает показатель RPKM главы 1. Вместо применения показателя RPKM мы выполним *квантильную нормализацию*<sup>1</sup>, т. е. прием, обеспечивающий укладку замеров в определенное распределение. В этом методе принимается большое допущение: если данные не распределены согласно желаемой форме, то мы просто заставляем их укладываться в метод! Это немного похоже на обман, но на деле такой подход оказывается простым и полезным в случаях, где конкретное распределение не имеет значения, и важность представляют относительные изменения значений внутри популяции. Например, Болстад (Bolstad) и его коллеги продемонстрировали<sup>2</sup>, что данный метод показывает превосходные результаты в восстановлении известных уровней экспрессии в данных ДНК-микрочипов.

По ходу главы мы воспроизведем упрощенную версию рис. 5А и 5В<sup>3</sup> из работы<sup>4</sup> «Геномная классификация кожной меланомы» проекта «Атлас ракового генома» (TCGA).

---

<sup>1</sup> См. [https://en.wikipedia.org/wiki/Quantile\\_normalization](https://en.wikipedia.org/wiki/Quantile_normalization).

<sup>2</sup> См. <https://academic.oup.com/bioinformatics/article/19/2/185/372664>.

<sup>3</sup> См. [http://www.cell.com/action/showImagesData?pii=S0092-8674\(15\)00634-0](http://www.cell.com/action/showImagesData?pii=S0092-8674(15)00634-0).

<sup>4</sup> См. [http://www.cell.com/cell/fulltext/S0092-8674\(15\)00634-0?\\_returnURL=http://linkinghub.elsevier.com/retrieve/pii/S0092867415006340%3Fshowall%3Dtrue](http://www.cell.com/cell/fulltext/S0092-8674(15)00634-0?_returnURL=http://linkinghub.elsevier.com/retrieve/pii/S0092867415006340%3Fshowall%3Dtrue).

Наша реализация квантильной нормализации эффективно использует возможности библиотек NumPy и SciPy, производя быструю, эффективную и элегантную функцию. Квантильная нормализация включает в себя три шага:

- 1) отсортировать значения по каждому столбцу;
- 2) найти среднее каждой результирующей строки;
- 3) заменить квантиль каждого столбца на квантиль среднего столбца.

```
import numpy as np
from scipy import stats
```



```
def quantile_norm(X):
```

*«»»Нормализовать столбцы X, чтобы каждый имел одинаковое распределение.*

*При заданной матрице экспрессии (данных микрочипов, количеств прочтений и пр.), состоящей из M генов на N образцов, квантильная нормализация обеспечивает, что все образцы будут иметь одинаковый разброс данных (в силу своей конструкции).*

*Данные по каждой строке усредняются, чтобы получить среднее значение столбца. Каждый квантиль столбца заменяется соответствующим квантилем среднего столбца.*

*Параметры*

-----

*X : двумерный массив вещественных, форма (M, N)*

*Входные данные с M строками (гены/признаки) и N столбцами (образцы).*

*Возвращает*

-----

*Xn : двумерный массив вещественных, форма (M, N)*

*Нормализованные данные.*

*«»»*

*# Вычислить квантили*

```
quantiles = np.mean(np.sort(X, axis=0), axis=1)
```

*# Вычислить ранги по столбцам. Каждое наблюдение заменяется на его ранг*

*# в этом столбце: наименьшее наблюдение заменяется на 1, следующее*

*# наименьшее на 2, ..., и наибольшее на M, т. е. на количество строк.*

```
ranks = np.apply_along_axis(stats.rankdata, 0, X)
```

*# Преобразовать ранги в целочисленные индексы от 0 до M-1*

```
rank_indices = ranks.astype(int) - 1
```

*# Проиндексировать квантили для каждого ранга ранговой матрицей*

```
Xn = quantiles[rank_indices]
```

```
return(Xn)
```



По причине характера вариабельности, присутствующей в количественных данных экспрессии генов, общепринято перед квантильной нормализацией логарифмически преобразовывать данные. Поэтому мы напишем дополнительную вспомогательную функцию, которая будет выполнять это преобразование:

```
def quantile_norm_log(X):
    logX = np.log(X + 1)
```

```
logXn = quantile_norm(logX)
return logXn
```

Обе эти функции иллюстрируют многое из того, что превращает библиотеку NumPy в мощный инструмент (первые три шага вам известны из главы 1):

- массивы могут быть одномерными, подобно спискам, но они также могут быть двумерными, как матрицы, и даже более многомерными. Этот прием дает возможность представлять многообразные виды числовых данных. В нашем случае мы представляем двумерную матрицу;
- массивы позволяют одновременно выражать многие числовые операции. В первой строке функции `quantile_norm_log` мы одним вызовом добавляем единицу и берем логарифм по каждому значению в  $X$ . Такая операция называется векторизацией;
- операции с массивами выполняются вдоль осей. В первой строке функции `quantile_norm` мы сортируем данные вдоль каждого столбца, задав в методе сортировки `np.sort` параметр оси `axis`. Затем мы берем среднее значение вдоль каждой строки, указав другую ось `axis`;
- в основе научной экосистемы Python лежат массивы. Функция `scipy.stats.rankdata` оперирует не списками Python, а массивами NumPy. Это же относится ко многим научным библиотекам в Python;
- даже те функции, которые не имеют ключевого слова `axis=`, можно заставить оперировать вдоль осей путем применения функции NumPy `apply_along_axis`;
- массивы поддерживают многочисленные виды манипуляций с данными посредством *чудесной индексации*: `Xn = quantiles[ranks]`. Она, возможно, представляет собой самую хитроумную часть библиотеки NumPy, которая к тому же является одной из самых полезных. Мы займемся ее подробным исследованием в последующих разделах.

## Получение данных

Как и в главе 1, мы будем работать с набором данных РНК-сек рака кожи из проекта TCGA. Наша цель состоит в том, чтобы предсказать смертность среди больных раком кожи, используя данные этого проекта об экспрессии РНК. Как отмечалось ранее, к концу этой главы мы воспроизведем упрощенную версию рис. 5A и 5B работы «Геномная классификация кожной меланомы» консорциума TCGA.

Как и в главе 1, мы сначала воспользуемся библиотекой `pandas`, чтобы намного упростить работу по считыванию данных в оперативную память. Прежде всего мы прочитаем количественные данные как таблицу `pandas`.

```
import numpy as np
import pandas as pd
```

```
# Импортировать данные TCGA о меланоме
filename = 'data/counts.txt'
```

```
data_table = pd.read_csv(filename, index_col=0) # Выполнить разбор файла
print(data_table.iloc[:, :5])
```

	00624286-41dd-476f-a63b-d2a5f484bb45	TCGA-FS-A1Z0	TCGA-D9-A3Z1 \
A1BG	1272.36	452.96	288.06
A1CF	0.00	0.00	0.00
A2BP1	0.00	0.00	0.00
A2LD1	164.38	552.43	201.83
A2ML1	27.00	0.00	0.00
	02c76d24-f1d2-4029-95b4-8be3bda8fdbe	TCGA-EB-A51B	
A1BG	400.11	420.46	

Глядя на строки и столбцы таблицы `data_table`, мы видим, что столбцы являются образцами, а строки – генами. Теперь давайте поместим наши количества в массив NumPy.

```
# Двумерный массив, содержащий количества экспрессии по каждому гену
# в каждом индивидууме
counts = data_table.values
```

## РАЗНИЦА В РАСПРЕДЕЛЕНИИ ЭКСПРЕССИИ ГЕНОВ МЕЖДУ ИНДИВИДУУМАМИ

Теперь давайте получим представление о наших количественных данных, построив график распределения количеств по каждому индивидууму. Чтобы сгладить неровности в наших данных, мы будем использовать гауссово ядро. Тем самым мы получим более оптимальное представление об общей форме данных.

Сначала, как обычно, мы настроим стиль построения графиков:

```
# Заставить графики появляться локально, задать стиль графиков
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('style/elegant.mplstyle')
```

Затем напишем функцию построения графика, применяющую функцию библиотеки SciPy `gaussian_kde` для построения сглаженных распределений:

```
from scipy import stats

def plot_col_density(data):
    «»»По каждому столбцу произвести график плотности по всем строкам.«»»

    # Применить гауссово сглаживание с целью получения оценки плотности
    density_per_col = [stats.gaussian_kde(col) for col in data.T]
    x = np.linspace(np.min(data), np.max(data), 100)

    fig, ax = plt.subplots()
    for density in density_per_col:
        ax.plot(x, density(x))
    ax.set_xlabel('Значения данных (в расчете на столбец)')
    ax.set_ylabel('Плотность')
```

Теперь мы можем применить эту функцию для построения графика распределения необработанных данных, т. е. до выполнения какой-либо нормализации:

```
# До нормализации
log_counts = np.log(counts + 1)
plot_col_density(log_counts)
```

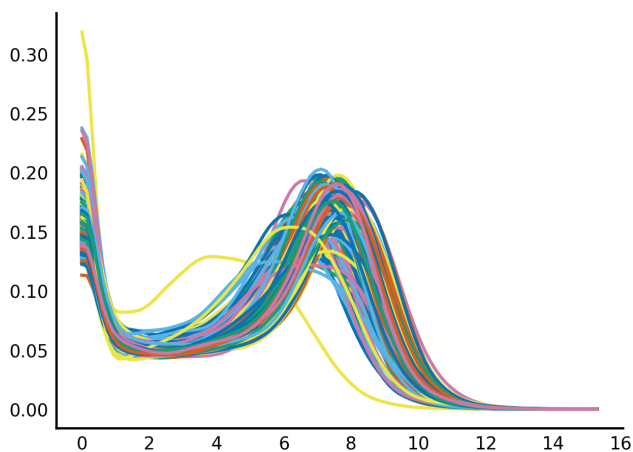


Рис. 2.1 ❖ График распределения до квантильной нормализации

Мы видим, что, в то время как распределения количеств в основном похожи, некоторые индивидуумы имеют более плоские распределения, а некоторые смещены влево. На самом деле если учесть, что мы имеем логарифмическую шкалу, то положение пика распределений фактически варьируется более чем на порядок! Далее в этой главе, во время выполнения анализа количественных данных, мы будем исходить из того, что изменения в экспрессии генов происходят из-за биологических различий между образцами. Но главный сдвиг в распределении, который мы наблюдаем, говорит о том, что эти различия технические. Иными словами, наличие изменений, скорее всего, вызвано различиями в обработке каждого образца, а не биологической вариацией. Поэтому мы попытаемся нормализовать эти глобальные различия между индивидуумами.

Чтобы выполнить эту нормализацию, мы задействуем квантильную нормализацию, как было описано в начале главы. Идея состоит в том, что все наши образцы должны иметь похожее распределение, поэтому любые различия в форме, скорее всего, происходят из-за некой технической вариации. Если выразаться более формально, то при заданной матрице экспрессии (данных микрочипов, количеств прочтений и пр.) с формой  $(n\_genes, n\_samples)$  квантильная нормализация гарантирует одинаковый разброс данных образцов (столбцов) в силу своей конструкции.

В NumPy и SciPy квантильная нормализация делается легко и эффективно. Ниже приведена наша реализация квантильной нормализации, которую мы ввели в начале данной главы.

Давайте предположим, что мы прочитали входную матрицу как  $X$ :

```
import numpy as np
from scipy import stats
```

```
def quantile_norm(X):
```

*«»Нормализовать столбцы  $X$ , чтобы каждый имел одинаковое распределение.*

*При заданной матрице экспрессии (данных микрочипов, количеств прочтений и пр.), состоящей из  $M$  генов на  $N$  образцов, квантильная нормализация обеспечивает, что все образцы будут иметь одинаковый разброс данных (в силу своей конструкции).*

*Данные по каждой строке усредняются, чтобы получить среднее значение столбца. Каждый квантиль столбца заменяется соответствующим квантилем среднего столбца.*

*Параметры*

-----

*$X$  : двумерный массив вещественных, форма ( $M$ ,  $N$ )*

*Входные данные с  $M$  строками (гены/признаки) и  $N$  столбцами (образцы).*

*Возвращает*

-----

*$X_n$  : двумерный массив вещественных, форма ( $M$ ,  $N$ )*

*Нормализованные данные.*

*«»*

*# Вычислить квантили*

```
quantiles = np.mean(np.sort(X, axis=0), axis=1)
```

*# Вычислить ранги по столбцам. Каждое наблюдение заменяется на его ранг*

*# в этом столбце: наименьшее наблюдение заменяется на 1, следующее*

*# наименьшее на 2, ..., и наибольшее на  $M$ , т. е. количество строк.*

```
ranks = np.apply_along_axis(stats.rankdata, 0, X)
```

*# Преобразовать ранги в целочисленные индексы от 0 до  $M-1$*

```
rank_indices = ranks.astype(int) - 1
```

*# Проиндексировать квантили для каждого ранга ранговой матрицей*

```
Xn = quantiles[rank_indices]
```

```
return(Xn)
```

```
def quantile_norm_log(X):
```

```
logX = np.log(X + 1)
```

```
logXn = quantile_norm(logX)
```

```
return logXn
```

Посмотрим, как будут выглядеть наши распределения после квантильной нормализации:

```
# После нормализации
```

```
log_counts_normalized = quantile_norm_log(counts)
```

```
plot_col_density(log_counts_normalized)
```



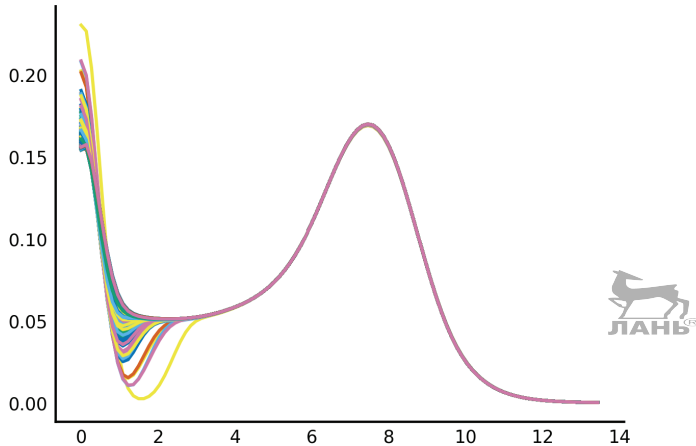


Рис. 2.2 ❖ График распределения после квантильной нормализации

Как и следовало ожидать, теперь распределения выглядят фактически идентичными! (Различающиеся левые хвосты распределения связаны с разным числом совпадений для значений с низкими количествами – 0, 1, 2, ... – в разных столбцах данных.)

Далее, когда мы нормализовали количества, начнем использовать данные экспрессии генов для предсказания дальнейшего течения болезни у пациентов.

## БИКЛАСТЕРИЗАЦИЯ КОЛИЧЕСТВЕННЫХ ДАННЫХ

Кластеризация образцов сообщает, какие образцы имеют похожие профили экспрессии генов. Это может свидетельствовать о похожих характеристиках образцов в других шкалах. После нормализации данных мы можем сгруппировать гены (строки) и образцы (столбцы) матрицы экспрессии. Кластеризация строк говорит о том, какие значения экспрессии генов связаны между собой. Это является признаком, что в изучаемом процессе они работают вместе. Под *бикластеризацией* подразумевается одновременное группирование строк и столбцов данных. В результате кластеризации вдоль строк мы обнаруживаем, какие гены взаимодействуют, и в итоге кластеризации вдоль столбцов будет показана схожесть образцов.

Поскольку операция кластеризации может быть дорогостоящей, мы ограничим наш анализ наиболее изменчивыми генами в количестве 1500, поскольку они будут объяснять большую часть сигнала корреляции в любой размерности.

```
def most_variable_rows(data, *, n=1500):
```

*«»Извлечь подмножество n наиболее изменчивых строк*

*В данном случае нам нужны n наиболее изменчивых генов.*

*Параметры*

-----

*data* : двумерный массив вещественных  
Данные, из которых будет извлечено подмножество  
*n* : целое, необязательный  
Количество возвращаемых строк.



Возвращает

```
-----
variable_data : двумерный массив вещественных
                `n` строк данных `data`, которые проявляют наибольшую дисперсию.
«»»
# Вычислить дисперсию вдоль оси столбцов
rowvar = np.var(data, axis=1)
# Получить отсортированные индексы (в порядке возрастания), взять последние n
sort_indices = np.argsort(rowvar)[-n:]
# Использовать в качестве индекса для данных
variable_data = data[sort_indices, :]
return variable_data
```

Далее нам нужна функция для выполнения бикластеризации данных. Обычно для этого используется сложный алгоритм кластеризации из библиотеки `scikit-learn`<sup>1</sup>. В нашем случае для простоты и удобства отображения результатов мы применим *иерархическую кластеризацию*. Как оказалось, библиотека `SciPy` располагает вполне приличным модулем иерархической кластеризации, хотя потребуется немного времени, чтобы разобраться в его интерфейсе.

Напомним, иерархическая кластеризация – это метод группирования наблюдений с использованием последовательного слияния кластеров. Первоначально каждое наблюдение принадлежит своему собственному кластеру. Затем два ближайших кластера неоднократно объединяются в один, далее приходит очередь следующих двух кластеров, и т. д. до тех пор, пока каждое наблюдение не будет находиться в единственном кластере. Эта последовательность операций слияния формирует *дерево слияния*. Подрезая дерево на определенной высоте, можно получить более мелкозернистую или более крупнозернистую кластеризацию наблюдений.

Функция связи `linkage` в модуле `scipy.cluster.hierarchy` выполняет иерархическую кластеризацию строк матрицы, используя конкретный метрический показатель (например, евклидово расстояние, манхэттенское расстояние или другие) и конкретный метод связи, т. е. расстояние между двумя кластерами (например, среднее расстояние между всеми наблюдениями в паре кластеров).

Функция возвращает дерево слияния в виде «матрицы связей», содержащей каждую операцию слияния вместе с вычисленным для слияния расстоянием и количеством наблюдений в результирующем кластере. Из документации по функции `linkage`:

Кластер с индексом меньше *n* соответствует одному из *n* исходных наблюдений. Расстояние между кластерами  $Z[i, 0]$  и  $Z[i, 1]$  дает  $Z[i, 2]$ . Четвертое значение  $Z[i, 3]$  представляет количество исходных наблюдений во вновь сформированном кластере.

<sup>1</sup> См. <http://scikit-learn.org/>.

Целый кладезь информации! Однако приступим к делу. Надеемся, что вы быстро освоитесь. Сначала мы определим функцию `bicluster`, кластеризирующую строки и столбцы матрицы:

```
from scipy.cluster.hierarchy import linkage

def bicluster(data, linkage_method='average', distance_metric='correlation'):
    """Кластеризовать строки и столбцы матрицы.

    Параметры
    -----
    data : двумерный массив ndarray
        Входные данные, подлежащие бикластеризации.
    linkage_method : строковый, необязательный
        Метод связи, передаваемый в `linkage`.
    distance_metric : строковый, необязательный
        Метрический показатель расстояния, применяемый для кластеризации.
        См. документацию по ``scipy.spatial.distance.pdist`` относительно
        допустимых метрических показателей расстояния

    Возвращает
    -----
    y_rows : матрица связей
        Кластеризация строк входных данных.
    y_cols : матрица связей
        Кластеризация столбцов входных данных.
    «»»
    y_rows = linkage(data, method=linkage_method, metric=distance_metric)
    y_cols = linkage(data.T, method=linkage_method, metric=distance_metric)
    return y_rows, y_cols
```

Все просто: мы вызываем функцию `linkage` для входной матрицы и для транспонированной входной матрицы, в которой столбцы становятся строками, а строки – столбцами.

## Визуализация кластеров

Далее определим функцию, визуализирующую результат кластеризации. Мы перестроим строки и столбцы входных данных так, чтобы похожие строки были вместе. Таким же способом выполним перестроение столбцов, чтобы похожие столбцы были вместе. Дополнительно к этому мы собираемся показать дерево слияния для строк и для столбцов, демонстрирующее, какие наблюдения подходят друг другу. Деревья слияния представляются в виде дендограмм, в которых длины ответвлений говорят, насколько наблюдения похожи друг на друга (чем короче дендограмма, тем больше сходства).

Заметьте, здесь используется много жестко запрограммированных параметров. Этого при построении графиков трудно избежать, так как дизайн графиков нередко требует визуального осмотра с целью поиска правильных пропорций.

```

from scipy.cluster.hierarchy import dendrogram, leaves_list

def clear_spines(axes):
    for loc in ['left', 'right', 'top', 'bottom']:
        axes.spines[loc].set_visible(False)
    axes.set_xticks([])
    axes.set_yticks([])

def plot_bicluster(data, row_linkage, col_linkage,
                  row_nclusters=10, col_nclusters=3):
    """Выполнить бикластеризацию, построить тепловую карту с дендограммами
    на каждой оси.

    Параметры
    -----
    data : массив вещественных, форма (M, N)
           Входные данные, подлежащие бикластеризации.
    row_linkage : массив, форма (M-1, 4)
                 Матрица связей для строк данных `data`.
    col_linkage : массив, форма (N-1, 4)
                 Матрица связей для столбцов данных `data`.
    n_clusters_r, n_clusters_c : целое, необязательный
                                 Количество кластеров для строк и столбцов.
    «»»
    fig = plt.figure(figsize=(4.8, 4.8))

    # Вычислить и построить дендограмму по строкам
    # `add_axes` принимает «прямоугольный» вход и добавляет подграфик
    # в изображение.
    # Принимается, что изображение имеет длину стороны 1 на каждой стороне,
    # и его левый нижний угол находится в (0, 0).
    # В `add_axes` передаются измерения со значениями: левая сторона,
    # нижняя сторона, ширина и высота подграфика. Таким образом, чтобы
    # построить левую дендограмму (для строк), мы создаем прямоугольник, чей
    # левый нижний угол находится в (0.09, 0.1) с шириной 0.2 и высотой 0.6.
    ax1 = fig.add_axes([0.09, 0.1, 0.2, 0.6])

    # Для заданного количества кластеров мы можем получить отрезок дерева
    # связей, обратившись к соответствующей аннотации расстояния в
    # матрице связи.
    threshold_r = (row_linkage[-row_nclusters, 2] +
                   row_linkage[-row_nclusters+1, 2]) / 2
    with plt.rc_context({'lines.linewidth': 0.75}):
        dendrogram(row_linkage, orientation='left',
                  color_threshold=threshold_r, ax=ax1)
    clear_spines(ax1)

    # Вычислить и построить дендограмму по столбцам
    # См. примечания выше, где дается объяснение параметров для `add_axes`
    ax2 = fig.add_axes([0.3, 0.71, 0.6, 0.2])
    threshold_c = (col_linkage[-col_nclusters, 2] +
                   col_linkage[-col_nclusters+1, 2]) / 2
    with plt.rc_context({'lines.linewidth': 0.75}):
        dendrogram(col_linkage, color_threshold=threshold_c, ax=ax2)

```





```

clear_spines(ax2)

# Построить тепловую карту данных
ax = fig.add_axes([0.3, 0.1, 0.6, 0.6])

# Отсортировать данные в соответствии с листьями дендограммы
idx_rows = leaves_list(row_linkage)
data = data[idx_rows, :]
idx_cols = leaves_list(col_linkage)
data = data[:, idx_cols]

im = ax.imshow(data, aspect='auto', origin='lower', cmap='YlGnBu_r')
clear_spines(ax)

# Метки осей
ax.set_xlabel('Образцы')
ax.set_ylabel('Гены', labelpad=125)

# Нанести пояснительную надпись
axcolor = fig.add_axes([0.91, 0.1, 0.02, 0.6])
plt.colorbar(im, cax=axcolor)

# Показать график
plt.show()

```

Теперь мы применим эти функции к нормализованной матрице количеств, чтобы показать кластеризации по строкам и по столбцам (рис. 2.3).

```

counts_log = np.log(counts + 1)
counts_var = most_variable_rows(counts_log, n=1500)
yr, yc = bicluster(counts_var, linkage_method='ward',
                   distance_metric='euclidean')
with plt.style.context('style/thinner.mplstyle'):
    plot_bicluster(counts_var, yr, yc)

```



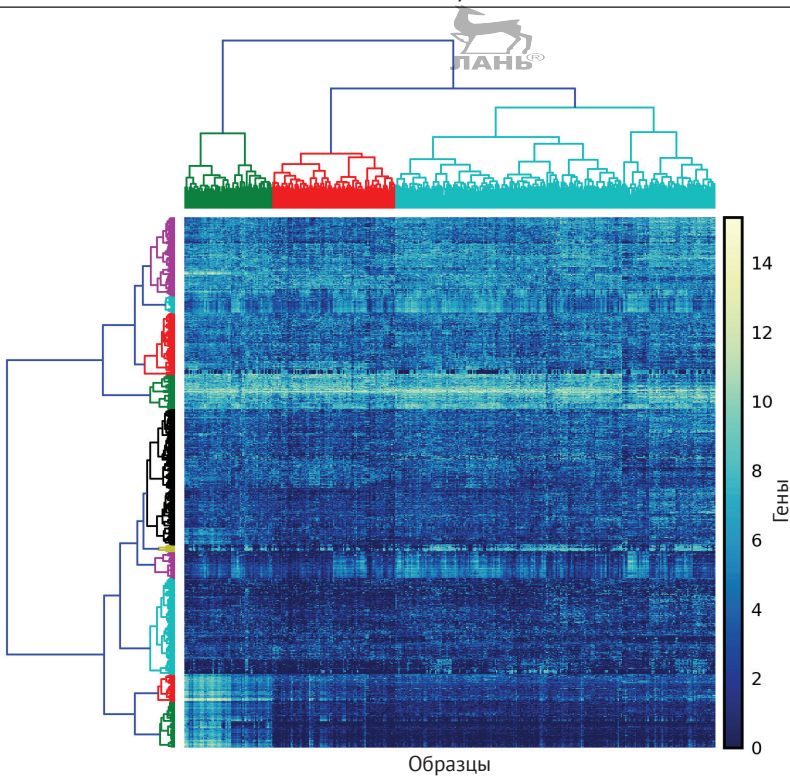
## ПРЕДСКАЗАНИЕ ВЫЖИВАЕМОСТИ

Мы видим, что данные образцов естественным образом попадают как минимум в два или три кластера. В чем смысл этих кластеров? Чтобы ответить на этот вопрос, мы можем обратиться к данным пациентов, доступным в хранилище данных, прилагаемых к исследовательской работе. После небольшой предварительной обработки мы получим таблицу пациентов, содержащую информацию о жизни каждого пациента. Затем ее можно сопоставить с кластерами количеств и понять, сможет ли экспрессия генов пациентов предсказать различия в их патологии.

```

patients = pd.read_csv('data/patients.csv', index_col=0)
patients.head()

```



**Рис. 2.3** ❖ Эта тепловая карта показывает уровень экспрессии генов по всем образцам и генам. Цвет говорит об уровне экспрессии. Строки и столбцы сгруппированы в соответствии с кластерами. Мы видим кластеры генов вдоль оси Y и кластеры образцов поперек оси X

**Таблица 2.1. Таблица пациентов**

	Сигнатура в УФ-области спектра	Исходные кластеры	Время жизни с меланомой	Смерть от меланомы
TCGA-BF-A1PU	УФ-сигнатура	кератин	NaN	NaN
TCGA-BF-A1PV	УФ-сигнатура	кератин	13.0	0.0
TCGA-BF-A1PX	УФ-сигнатура	кератин	NaN	NaN
TCGA-BF-A1PZ	УФ-сигнатура	кератин	NaN	NaN
TCGA-BF-A1Q0	не УФ	резистентный	17.0	0.0

По каждому пациенту (строкам) мы имеем:

*Сигнатура в ультрафиолетовой области спектра*

Ультрафиолетовый свет имеет тенденцию вызывать определенные мутации ДНК. Отыскивая эту сигнатуру мутации, исследователи могут заключить, смог ли ультрафиолетовый свет вызвать мутацию (мутации), приведшую к появлению рака у этих пациентов.

### Исходный кластер

В исследовательской работе пациенты были кластеризованы на основе данных экспрессии генов. Эти кластеры были классифицированы в соответствии с типами генов, типичными для этого кластера. Главными кластерами были «резистентный» (immune) ( $n = 168$ ; 51%), «кератин» ( $n = 102$ ; 31%) и «MITF-low» ( $n = 59$ ; 18%)<sup>1</sup>.

### Время жизни с меланомой

Количество дней, которые пациент прожил.

### Смерть от меланомы

Один (1), если пациент умер от меланомы, ноль (0), если они жив или умер от чего-то другого.

Теперь для каждой группы пациентов, определенной в результате кластеризации, нам нужно начертить *кривые выживания*. Этот график показывает доли популяции, остающейся живой в течение некоторого времени. Обратите внимание, некоторые данные *цензурованы справа*. Это означает, что в некоторых случаях мы фактически не знаем дату смерти пациента, либо причины смерти пациента не связаны с меланомой. Мы рассматриваем этих пациентов как «живых» на всем протяжении кривой выживаемости, однако более тщательные исследования могут попытаться оценить их вероятное время смерти.

Чтобы получить кривую выживания из времен жизни, мы создадим ступенчатую функцию, уменьшающуюся на  $1/n$ , где  $n$  – это количество пациентов в группе. Затем мы сопоставим эту функцию относительно нецензуированных времен жизни.

```
def survival_distribution_function(lifetimes, right_censored=None):
```

*«»Вернуть функцию распределения выживаемости из набора времен жизни.*

*Параметры*

-----

*lifetimes : массив вещественных либо целых*

*Наблюдавшиеся времена жизни популяции. Они должны быть неотрицательными.*

*right\_censored : массив булевых, такой же формы, что и `lifetimes`*

*Здесь значение `True` говорит о том, что это время жизни не наблюдалось.*

*Значения `np.nan` в `lifetimes` также рассматриваются как цензурированные справа.*

*Возвращает*

-----

*sorted\_lifetimes : массив вещественных*

*Отсортированные времена жизни*

*sdf : array of float*

*Значения, начинающиеся с 1 и прогрессивно уменьшающиеся, на один уровень для каждого наблюдения в `lifetimes`.*



<sup>1</sup> MITF – транскрипционный фактор, ассоциированный с микрофтальмией. – Прим. перев.

## Примеры

-----

В этом примере в популяции из четырех человек двое умирают во время 1, третий умирает во время 2, и последний умирает в неизвестное время. (Отсюда ``np.nan``.)



```
>>> lifetimes = np.array([2, 1, 1, np.nan])
>>> survival_distribution_function(lifetimes)
(array([ 0., 1., 1., 2.]), array([ 1. , 0.75, 0.5 , 0.25]))
"""
```

```
n_obs = len(lifetimes)
rc = np.isnan(lifetimes)
if right_censored is not None:
    rc |= right_censored
observed = lifetimes[~rc]
xs = np.concatenate( ([0], np.sort(observed)) )
ys = np.linspace(1, 0, n_obs + 1)
ys = ys[:len(xs)]
return xs, ys
```

Теперь, когда можно легко получить кривые выживания из данных о жизни пациентов, мы сможем вывести их на график. Мы напишем функцию, которая группирует времена жизни по идентификатору (identity) кластера и выводит каждую группу на графике как отдельную прямую:

```
def plot_cluster_survival_curves(clusters, sample_names, patients,
                                censor=True):
```

```
    """Вывести на график данные о жизни из набора кластеров образцов.
```

## Параметры

-----

*clusters* : массив целых либо серия *pd.Series* категориальных значений  
Идентификатор кластера каждого образца, кодированный как простое целое либо как категориальная переменная *pandas*.

*sample\_names* : list of string

Имя, соответствующее каждому образцу. Должно иметь такую же длину, что и массив *clusters*.

*patients* : *pandas.DataFrame*

Фрейм данных *DataFrame*, содержащий информацию о жизни каждого пациента. Индексы этого объекта *DataFrame* должны соответствовать именам *sample\_names*. Образцы, не представленные в этом списке, будут проигнорированы.

*censor* : булев, необязательный

Если *True*, то использовать *patients['melanoma-dead']*, чтобы цензурировать данные о жизни справа.

```
    """
```

```
fig, ax = plt.subplots()
if type(clusters) == np.ndarray:
    cluster_ids = np.unique(clusters)
    cluster_names = ['cluster {}'.format(i) for i in cluster_ids]
elif type(clusters) == pd.Series:
    cluster_ids = clusters.cat.categories
    cluster_names = list(cluster_ids)
```



```

n_clusters = len(cluster_ids)
for c in cluster_ids:
    clust_samples = np.flatnonzero(clusters == c)
    # Отбросить пациентов, которые не представлены в данных о жизни
    clust_samples = [sample_names[i] for i in clust_samples
                    if sample_names[i] in patients.index]
    patient_cluster = patients.loc[clust_samples]
    survival_times = patient_cluster['melanoma-survival-time'].values
    if censored:
        censored = ~patient_cluster['melanoma-dead'].values.astype(bool)
    else:
        censored = None
    stimes, sfracs = survival_distribution_function(survival_times,
                                                censored)

    ax.plot(stimes / 365, sfracs)

ax.set_xlabel('время жизни (годы)')
ax.set_ylabel('доля живых')
ax.legend(cluster_names)

```



Теперь можно применить функцию `fcluster`, чтобы получить идентификаторы кластеров для образцов (столбцы данных о количествах) и вывести на график каждую кривую выживания по отдельности. Функция `fcluster` принимает матрицу связей, возвращаемую функцией `linkage`, и порог и возвращает идентификаторы кластеров. Очень трудно узнать *априори*, каким должен быть порог, но мы можем получить соответствующий порог для фиксированного количества кластеров в результате проверки расстояний в матрице связей.

```

from scipy.cluster.hierarchy import fcluster
n_clusters = 3
threshold_distance = (yc[-n_clusters, 2] + yc[-n_clusters+1, 2]) / 2
clusters = fcluster(yc, threshold_distance, 'distance')
plot_cluster_survival_curves(clusters, data_table.columns, patients)

```

Как показано на рис. 2.4, кластеризация профилей экспрессии генов, по всей видимости, идентифицировала высокорисковый подтип меланомы (группа 2). Исследовательская работа TCGA поддерживает это утверждение более робастной кластеризацией и проверкой статистических гипотез. Это совсем недавнее исследование показывает полученный здесь результат, а также другие результаты, в которых идентифицируются такие подтипы рака, как лейкемия (рак крови), рак пищеварительного тракта и другие. Вышеупомянутый метод кластеризации не надежен. Вместе с тем имеются другие, более надежные способы исследования этого и подобных ему наборов данных<sup>1</sup>.

<sup>1</sup> См. сеть «Атласа генома рака» // Геномная классификация кожной меланомы. Cell 161, no. 7 (2015): 1681–1696 (<http://dx.doi.org/10.1016/j.cell.2015.05.044>).

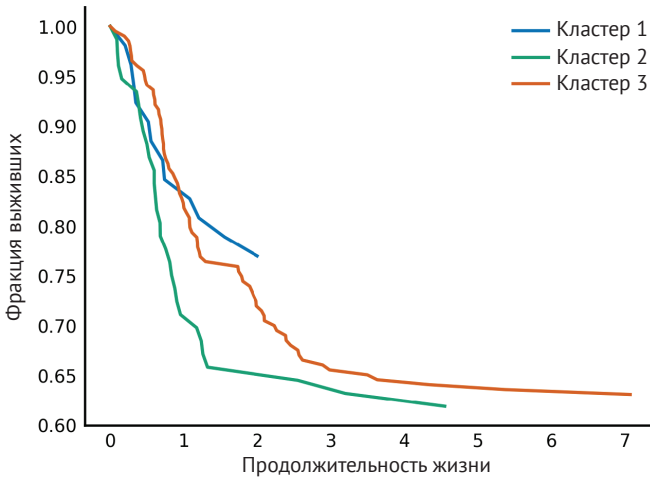


Рис. 2.4 ❖ Кривые выживания для пациентов, кластеризованных с использованием данных экспрессии генов

## Дальнейшая работа: использование кластеров пациентов TCGA

Предсказывают ли наши кластеры выживание более эффективно, чем исходные кластеры, приведенные в исследовательской работе? А что относительно УФ-сигнатуры? Постройте график кривых выживания с использованием исходных кластеров и столбцов с УФ-сигнатурами данных пациентов. Как они соотносятся с нашими кластерами?

## Дальнейшая работа: воспроизведение кластеров TCGA

В качестве упражнения мы оставляем вам реализацию подхода, описанного в исследовательской работе<sup>1</sup>.

1. Возьмите бутстраповские выборки генов (случайный отбор с возвратом), которые используются для кластеризации образцов.
2. По каждому образцу произведите иерархическую кластеризацию.
3. При бутстраповской кластеризации в матрице формы  $(n\_samples, n\_samples)$  сохраните частотность (количество) одновременного появления пары образцов.
4. На результирующей матрице выполните иерархическую кластеризацию.

Это позволит идентифицировать группы образцов, которые часто появляются в кластерах вместе, независимо от выбранных генов. В результате можно будет считать, что эти образцы надежно кластеризуются вместе.

✓ **Подсказка.** Для создания бутстраповских выборок индексов строк следует применить функцию `np.random.choice` с аргументом `replacement=True`.

<sup>1</sup> См. сеть «Атласа генома рака» // Геномная классификация кожной меланомы. Cell 161, no. 7 (2015): 1681–1696 (<http://dx.doi.org/10.1016/j.cell.2015.05.044>).

# Глава 3



## Создание сетей из областей изображений при помощи ndimage



Тигр, о тигр, светло горящий  
В глубине полночной чаши,  
Кем задуман огневой  
Симметричный образ твой?

– Уильям Блэйк. «Тигр»

Скорее всего, вам известно, что цифровые изображения состоят из пикселей. В целом их следует представлять не как крошечные квадраты, а как *точечные образцы* светового сигнала, *вычисленные на регулярной сетке*<sup>1</sup>.

Более того, во время обработки изображений мы часто имеем дело с объектами, которые намного крупнее отдельных пикселей. Пейзаж, небо, земля, деревья и скалы – все эти объекты состоят из большого количества пикселей. Для их представления используется универсальная структура под названием «граф смежности областей», или RAG (region adjacency graph, лоскутный граф). Его *узлы* содержат свойства каждой области в изображении, а *связи* – пространственные отношения между областями. Во входном изображении два узла связаны между собой, когда их соответствующие области соприкасаются друг с другом.

Создание такой структуры может представлять сложную задачу, и эта задача становится еще сложнее, когда изображения не двумерные, а трехмерные и даже четырехмерные, что, помимо прочего, широко распространено в микроскопии, материаловедении и климатологии. Тем не менее здесь мы вам покажем метод генерирования RAG-графа всего в нескольких строках программ-

<sup>1</sup> См. *Смит А. Р.* Пиксель – это не крошечный квадрат: техническая записка. 17 июля 1995 г. (Alvy Ray Smith. A Pixel Is Not A Little Square // [http://alvyray.com/Memos/CG/Microsoft/6\\_pixel.pdf](http://alvyray.com/Memos/CG/Microsoft/6_pixel.pdf)).

ного кода с использованием библиотеки NetworkX (библиотеки Python для анализа графов и сетей) и фильтра из подмодуля SciPy ndimage для обработки N-мерных изображений.

```
import networkx as nx
import numpy as np
from scipy import ndimage as ndi

def add_edge_filter(values, graph):
    center = values[len(values) // 2]
    for neighbor in values:
        if neighbor != center and not graph.has_edge(center, neighbor):
            graph.add_edge(center, neighbor)
    return 0.0

def build_rag(labels, image):
    g = nx.Graph()
    footprint = ndi.generate_binary_structure(labels.ndim, connectivity=1)
    _ = ndi.generic_filter(labels, add_edge_filter, footprint=footprint,
                          mode='nearest', extra_arguments=(g,))
    return g
```

### Истоки книги «Элегантный SciPy»

*(Примечание от Хуана Нуньес-Иглесиаса)*

Эта глава требует особого упоминания, так как она вдохновила авторов на написание всей книги. Винеш Биркодар (Vighnesh Birodkar) написал данный фрагмент кода в 2014 г., когда, будучи студентом Университета, участвовал в инициативной программе компании Google, известной как Google Summer of Code (GSoC). Когда я увидел этот программный код, он поразил мое воображение. Если говорить о целях настоящей книги, этот код затрагивает многие аспекты научного программирования на Python. Когда вы закончите читать данную главу, вы научитесь обрабатывать массивы *любой* размерности и прекратите о них думать только как об одномерных списках или двумерных таблицах. Более того, вы поймете основы фильтрации изображений и обработки сетей.

Здесь будет происходить следующее: мы представим изображения как массивы NumPy, *фильтрацию* этих изображений выполним с использованием модуля `scipy.ndimage`. А для превращения областей изображений в граф (сеть) используем библиотеку NetworkX.

## ИЗОБРАЖЕНИЯ – ЭТО ПРОСТО МАССИВЫ NumPy

В предыдущей главе было показано, что массивы NumPy могут эффективно представлять табличные данные и являются удобным инструментом выполнения вычислений. Как оказалось, массивы одинаково эффективно способны представлять изображения.

Ниже показано, как можно создать изображение белого шума, используя только библиотеку NumPy, и с помощью библиотеки Matplotlib вывести его на

экран. Мы сначала импортируем необходимые пакеты и применим волшебную команду IPython `matplotlib inline`, чтобы наши изображения появлялись внизу программного кода:

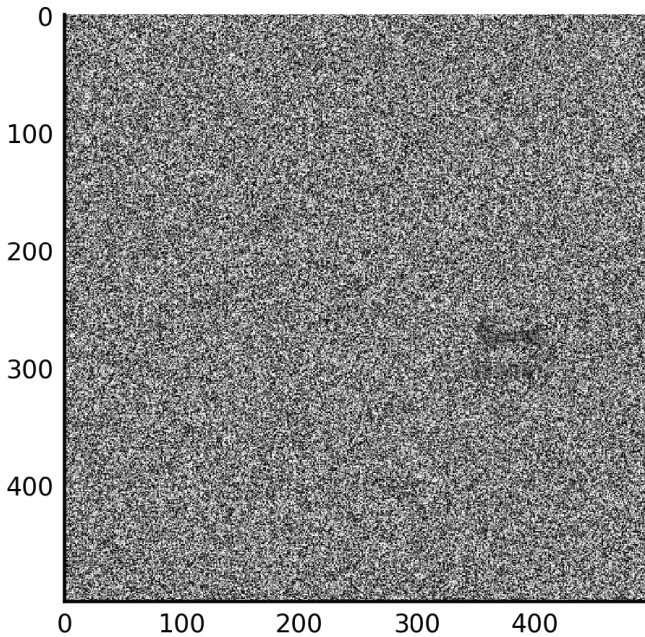
```
# Заставить графики появляться локально, задать индивидуальный стиль графиков  
%matplotlib inline  
import matplotlib.pyplot as plt  
plt.style.use('style/elegant.mplstyle')
```

Затем мы «добавим немного шума» и выведем его как изображение:

```
import numpy as np  
random_image = np.random.rand(500, 500)  
plt.imshow(random_image);
```



Функция `imshow` показывает массив NumPy как изображение:



Также справедливо и обратное: изображение может рассматриваться как массив NumPy. Для этого примера мы используем библиотеку `scikit-image`, представляющую собой коллекцию инструментов обработки изображений, настроенных поверх библиотек NumPy и SciPy.

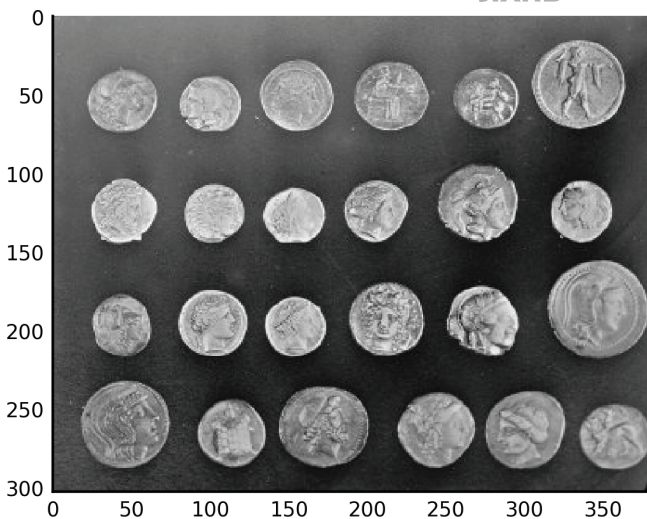
Вот изображение в формате PNG из хранилища `scikit-image`. Это черно-белая (иногда именуемая «полутоновой») фотография нескольких древних римских монет из Помпеи, снятая в Бруклинском музее:



Ниже показано, как загрузить изображение монет при помощи библиотеки scikit-image:

```
from skimage import io
url_coins = ('https://raw.githubusercontent.com/scikit-image/scikit-image/'
            'v0.10.1/skimage/data/coins.png')
coins = io.imread(url_coins)
print("Тип:", type(coins), "Форма:", coins.shape, "Тип данных:", coins.dtype)
plt.imshow(coins);
```

Тип: <class 'numpy.ndarray'> Форма: (303, 384) Тип данных: uint8



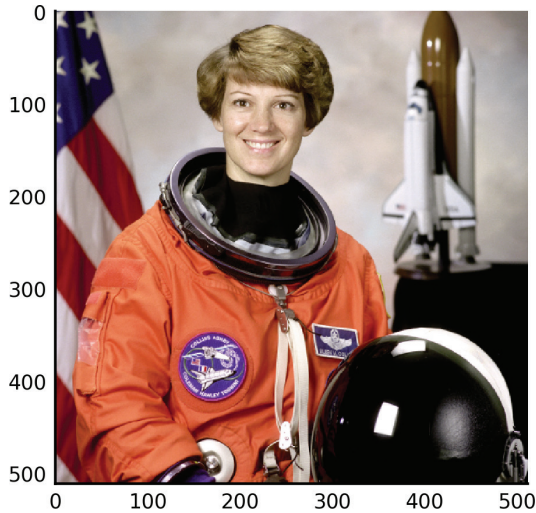
Полутоновое изображение может быть представлено как *двумерный* массив, в котором каждый элемент массива в конкретной позиции содержит полутоновую интенсивность. Одним словом, *изображение – это всего лишь массив NumPy*.



Цветные изображения являются *трехмерными* массивами, где первые две размерности представляют пространственные позиции изображения, в то время как заключительная размерность представляет цветовые каналы, как правило, три основных аддитивных цвета: красный, зеленый и синий. Чтобы показать, что можно сделать с этими размерностями, давайте поэкспериментируем с приведенной ниже фотографией астронавта Айлин Коллинз (Eileen Collins):

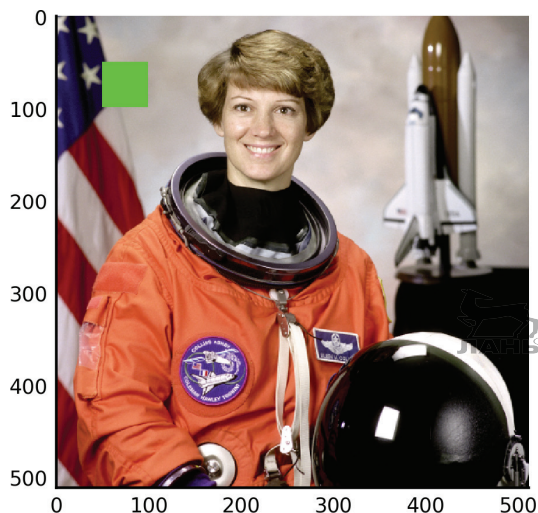
```
url_astronaut = ('https://raw.githubusercontent.com/scikit-image/scikit-image/'
                'master/skimage/data/astronaut.png')
astro = io.imread(url_astronaut)
print("Тип:", type(astro), "Форма:", astro.shape, "Тип данных:", astro.dtype)
plt.imshow(astro);
```

Тип: <class 'numpy.ndarray'> Форма: (512, 512, 3) Тип данных: uint8



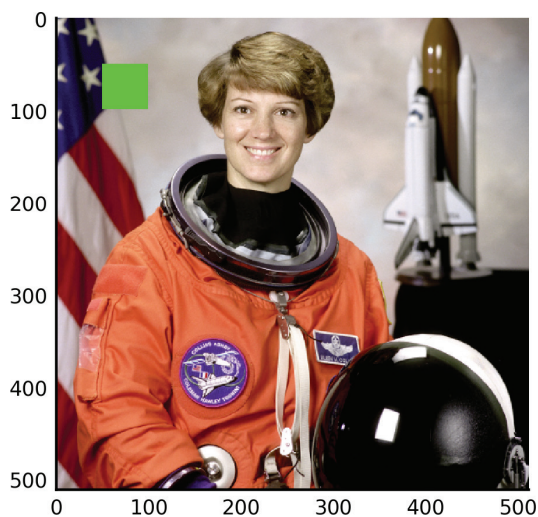
Это изображение является *простым массивом NumPy*. Как только вы это поймете, вы очень легко сможете в изображение добавить зеленый квадрат, используя простой срез массива NumPy:

```
astro_sq = np.copy(astro)
astro_sq[50:100, 50:100] = [0, 255, 0] # красный, зеленый, синий
plt.imshow(astro_sq);
```



Кроме того, вы можете применить булеву *маску*, т. е. массив, состоящий из значений True или False. Мы эти значения встречали в главе 2 как способ отбора строк таблицы. В данном случае для отбора пикселей можно применить массив такой же формы, что и изображение:

```
astro_sq = np.copy(astro)
sq_mask = np.zeros(astro.shape[:2], bool)
sq_mask[50:100, 50:100] = True
astro_sq[sq_mask] = [0, 255, 0]
plt.imshow(astro_sq);
```





## Задача: добавление сеточного наложения

Только что было показано, каким образом можно выбрать квадрат и окрасить его в зеленый цвет. Есть ли возможность ли расширить этот способ на другие формы и цвета? Давайте создадим функцию, рисующую синюю сетку на цветном изображении, и применим ее к приведенной выше фотографии. Ваша функция должна содержать два параметра: входное изображение и интервал сетки. Мы предлагаем использовать шаблон, приведенный ниже. Этот шаблон поможет вам начать:

```
def overlay_grid(image, spacing=128):
    «»Вернуть изображение с сеточным наложением, используя
    предоставленный интервал.

    Параметры
    -----
    image : массив, форма (M, N, 3)
            Входное изображение.
    spacing : целое
            Интервал между линиями сетки.

    Возвращает
    -----
    image_gridded : массив, форма (M, N, 3)
            Исходное изображение с наложенной синей сеткой.
    «»
    image_gridded = image.copy()
    pass # замените эту строку своим программным кодом...
    return image_gridded
# plt.imshow(overlay_grid(astro, 128)); # раскомментировать эту строку, чтобы
# протестировать вашу функцию
```



Обратите внимание: решение задачи «Добавление сеточного наложения» вы найдете в конце книги.

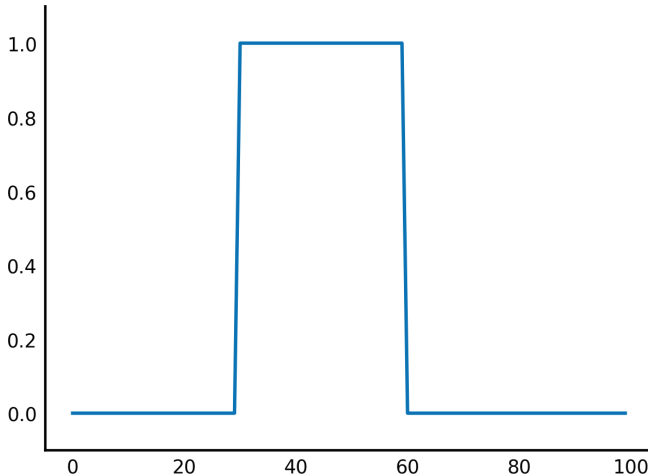
## ФИЛЬТРЫ В ОБРАБОТКЕ СИГНАЛОВ

Фильтрация – это одна из самых фундаментальных и распространенных операций в обработке изображений. Вы можете отфильтровать изображение, чтобы удалить шум, усилить его свойства или обнаружить края между объектами в изображении.

Чтобы понять, как работают фильтры, проще всего начать не с изображения, а с одномерного сигнала. Например, вы можете измерить свет, поступающий в ваш конец оптоволоконного кабеля. Если вы будете *отбирать* (сэмплировать) сигнал каждую миллисекунду (мс) в течение 100 мс, то в итоге вы получите массив длиной 100 элементов. Предположим, что на 30-й миллисекунде световой сигнал был включен, и через 30 миллисекунд снова выключен. В результате вы получите сигнал, как показано ниже:

```

sig = np.zeros(100, np.float) #
sig[30:60] = 1 # сигнал = 1 в течение периода 30-60 мс,
                # так как свет наблюдается
fig, ax = plt.subplots()
ax.plot(sig);
ax.set_ylim(-0.1, 1.1);
    
```



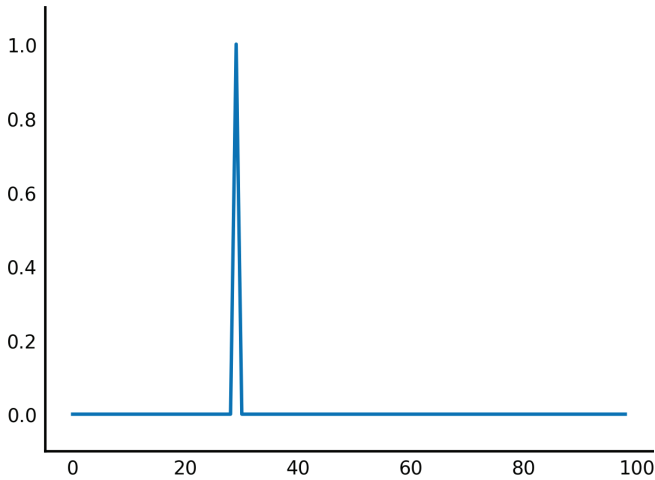
Чтобы найти точку, когда свет был включен, можно задержать сигнал на 1 мс, затем вычесть исходный сигнал из задержанного сигнала. Тем самым, когда между текущей миллисекундой и следующей сигнал остается без изменения, вычитание будет давать 0, но, когда сигнал увеличится, вы получите положительный сигнал.

Когда сигнал уменьшится, мы получим отрицательный сигнал. Если мы заинтересованы только в точном определении времени включения света, мы можем обрезать разностный сигнал, чтобы любые отрицательные значения были преобразованы в 0:

```

sigdelta = sig[1:] # sigdelta[0] равняется sig[1] и т.д.
sigdiff = sigdelta - sig[:-1]
sigon = np.clip(sigdiff, 0, np.inf) # сигнал включен
fig, ax = plt.subplots()
ax.plot(sigon)
ax.set_ylim(-0.1, 1.1)
print('Сигнал включен на:', 1 + np.flatnonzero(sigon)[0], 'мс')
    
```

Сигнал включен на: 30 мс



(Здесь мы применили функцию NumPy `flatnonzero`, чтобы получить первый индекс, где массив `sigon` не равен 0.)

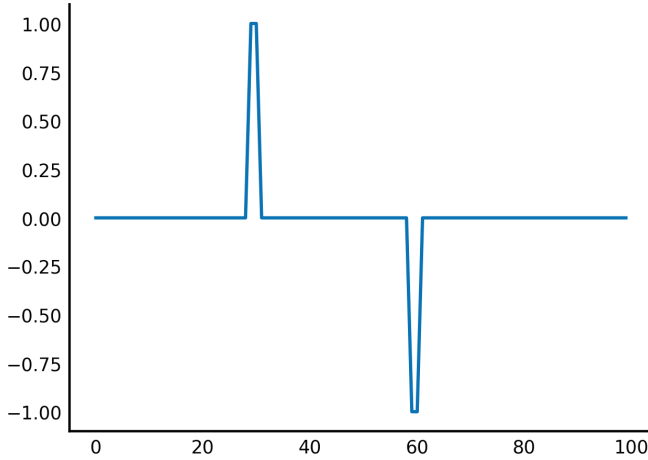
Оказывается, это может быть реализовано операцией обработки сигналов, именуемой *конволюцией*, или сверткой. В каждой точке сигнала мы вычисляем скалярное произведение между окружающими эту точку значениями и *ядром*, или *фильтром*, который представляет собой предопределенный вектор значений. Затем в зависимости от ядра конволюция выявляет различные свойства сигнала.

Теперь подумайте, что произойдет, когда для сигнала  $s$  ядро равняется  $(1, 0, -1)$ , т. е. представляет собой разностный фильтр. В любой позиции  $i$  результат конволюции равняется  $1*s[i+1] + 0*s[i] - 1*s[i-1]$ , что сводится к  $s[i+1] - s[i-1]$ . Таким образом, когда смежные с  $s[i]$  значения идентичны, конволюция дает 0, но когда  $s[i+1] > s[i-1]$  (сигнал увеличивается), конволюция дает положительное значение, и наоборот, когда  $s[i+1] < s[i-1]$ , конволюция дает отрицательное значение. Это можно представить как вычисление производной из входной функции.

В целом формула конволюции имеет вид  $s'(t) = \sum_{j=i-\tau}^i s(j)f(t-j)$ , где  $s$  – это сигнал,  $s'$  – отфильтрованный сигнал,  $f$  – фильтр и  $\tau$  – длина фильтра.

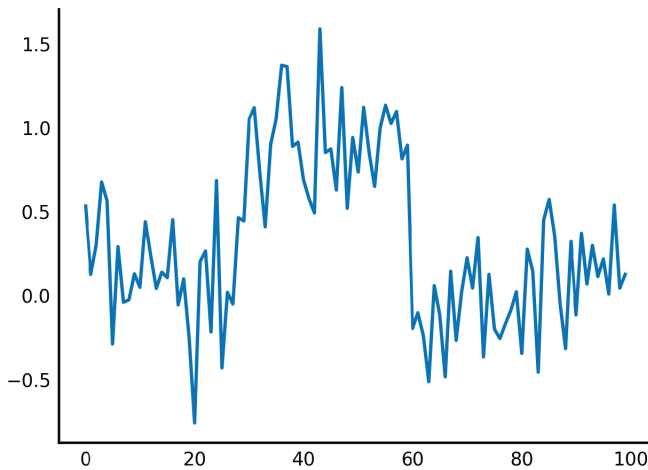
В библиотеке SciPy для работы с конволюцией можно использовать функцию `scipy.ndimage.convolve`:

```
diff = np.array([1, 0, -1])
from scipy import ndimage as ndi
dsig = ndi.convolve(sig, diff)
plt.plot(dsig);
```



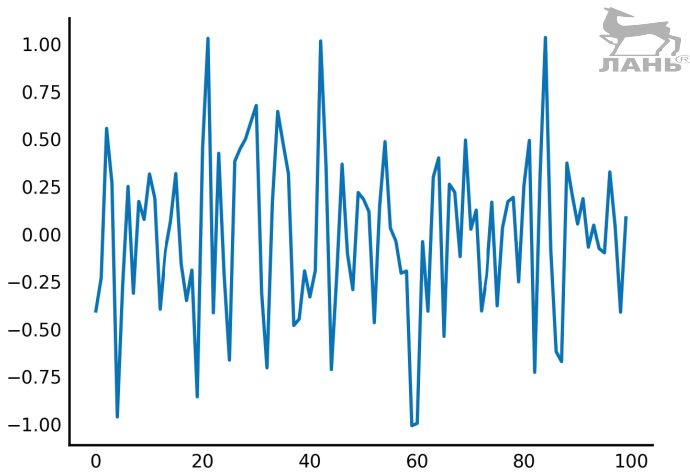
Правда, приведенные выше сигналы обычно зашумлены и не идеальны:

```
np.random.seed(0)
sig = sig + np.random.normal(0, 0.3, size=sig.shape)
plt.plot(sig);
```



Простой разностный фильтр может этот шум усилить:

```
plt.plot(ndi.convolve(sig, diff));
```



В таких случаях в фильтр можно добавить сглаживание. Наиболее распространенной формой сглаживания является *гауссово* сглаживание, которое берет взвешенное среднее соседних точек в сигнале, используя гауссову функцию<sup>1</sup>. Мы можем написать функцию, создающую гауссово ядро сглаживания, которая будет выглядеть следующим образом:

```
def gaussian_kernel(size, sigma):
    «»Создать одномерное гауссово ядро заданного размера и стандартного отклонения.

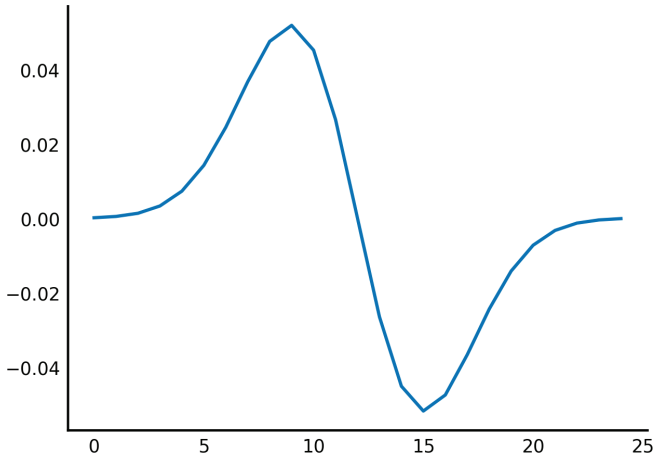
    Размер должен иметь нечетное число и стандартное отклонение (сигма) как минимум в ~6 раз больше, чтобы обеспечить достаточное покрытие.
    «»»

    positions = np.arange(size) - size // 2
    kernel_raw = np.exp(-positions**2 / (2 * sigma**2))
    kernel_normalized = kernel_raw / np.sum(kernel_raw)
    return kernel_normalized
```

Поистине замечательная особенность операции конволюции состоит в том, что она *ассоциативна*. То есть если вы хотите найти производную сглаженного сигнала, то в равной степени можете свернуть сигнал сглаженным разностным фильтром! Это может сэкономить уйму вычислительного времени, так как будет сглаживаться только фильтр, размер которого обычно намного меньше данных.

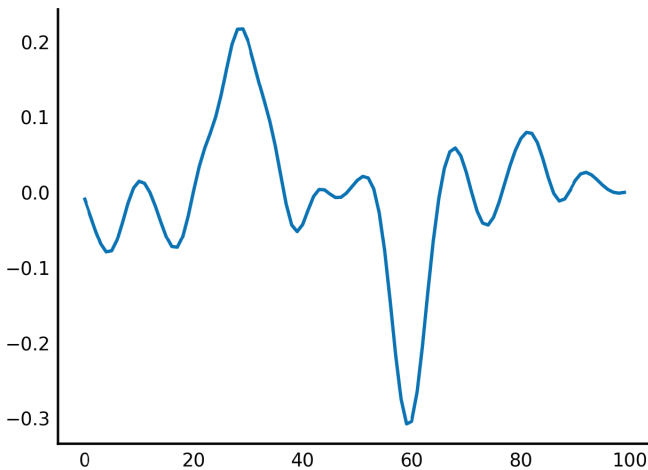
```
smooth_diff = ndi.convolve(gaussian_kernel(25, 3), diff)
plt.plot(smooth_diff);
```

<sup>1</sup> См. [https://ru.wikipedia.org/wiki/Гауссова\\_функция](https://ru.wikipedia.org/wiki/Гауссова_функция).



Этот сглаженный разностный фильтр отыскивает край в центральной позиции и выполняет проверку на продолжение разницы. Это продолжение происходит не на «паразитных» краях, вызванных шумом, а в истинном крае. Обратите внимание на результат (рис. 3.1):

```
sdsig = ndi.convolve(sig, smooth_diff)
plt.plot(sdsig);
```



**Рис. 3.1** ❖ Сглаженный разностный фильтр, примененный к зашумленному сигналу

Хотя он все еще выглядит неустойчивым, *соотношение сигнал-шум* (SNR) в этой версии намного больше, чем тогда, когда мы используем простой разностный фильтр.

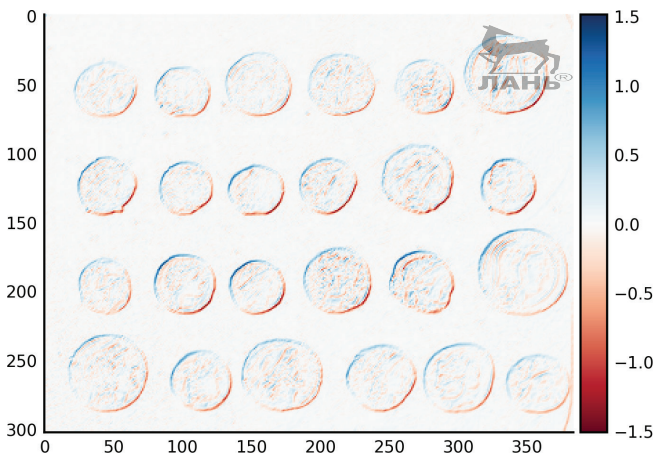
- ☑ **Фильтрация.** Эта операция называется фильтрацией, потому что в физических электрических цепях большое число таких операций реализуются аппаратными средствами, позволяющими пропускать определенные виды тока, блокируя другие. Такие аппаратные компоненты называются фильтрами. Например, распространенный фильтр, удаляющий из тока высокочастотные колебания напряжения, называется *фильтром нижних частот*.



## ФИЛЬТРАЦИЯ ИЗОБРАЖЕНИЙ (ДВУМЕРНЫЕ ФИЛЬТРЫ)

Теперь, когда вы увидели фильтрацию в одном измерении, мы надеемся, что вам удастся легко расширить эти понятия на двумерные сигналы, такие как изображения. Вот двумерный разностный фильтр для нахождения краев на изображении монет:

```
coins = coins.astype(float) / 255 # предотвращает ошибки переполнения
diff2d = np.array([[0, 1, 0], [1, 0, -1], [0, -1, 0]])
coins_edges = ndi.convolve(coins, diff2d)
io.imshow(coins_edges);
```



Принцип тот же, что с одномерным фильтром: в каждой точке в изображении поместить фильтр, вычислить скалярное произведение значений фильтра со значениями изображения и поместить результат в выходном изображении в той же самой позиции. Подобно одномерному разностному фильтру, когда фильтр помещается в позицию с небольшой вариацией, а скалярное произведение сводится к нулю, при помещении фильтра в позицию с изменяемой яркостью изображения значения, умноженные на 1, будут отличаться от значений, умноженных на -1. При этом отфильтрованный результат станет положительной или отрицательной величиной (в зависимости от того, является ли изображение, находящееся справа внизу или слева сверху от этой точки, ярче).

Точно так, как с одномерным фильтром, вы можете усложнить задачу и сгладить шум прямо в фильтре. Фильтр *Собела* предназначен как раз для этой цели. Данный фильтр существует как в горизонтальном, так и в вертикальном варианте и отыскивает края в одной из указанных в данных ориентаций. Давайте начнем с горизонтального фильтра. Чтобы на фотографии найти горизонтальный край, вы можете попробовать следующий ниже фильтр:

```
# столбцовый (вертикальный) вектор для нахождения горизонтальных краев
hdiff = np.array([[1], [0], [-1]])
```

Однако, как мы убедились на примере одномерных фильтров, это приводит к зашумленности оценки краев в изображении. Но вместо применения гауссового сглаживания, вызывающего во многих случаях расплывчатые края, фильтр Собела использует свойство, демонстрирующее тенденцию непрерывности краев. Например, фотография океана будет содержать горизонтальный край вдоль всей линии, а не только в отдельных точках изображения. Поэтому фильтр Собела сглаживает вертикальный фильтр горизонтально: он ищет сильный край в центральной позиции, подкрепляемый смежными позициями:

```
hsobel = np.array([[ 1, 2, 1],
                  [ 0, 0, 0],
                  [-1, -2, -1]])
```

Вертикальный фильтр Собела попросту является результатом транспонирования горизонтального:

```
vsobel = hsobel.T
```



В результате в изображении монет мы можем найти горизонтальные и вертикальные края:

```
# Нанесение индивидуальных меток на оси X, чтобы графики легче читались
def reduce_xaxis_labels(ax, factor):
    «»Показывать только каждую i-ю метку, чтобы предотвратить скапливание на оси X, например factor = 2 будет выводить каждую вторую метку оси X, начиная с первой.
```

*Параметры*

-----

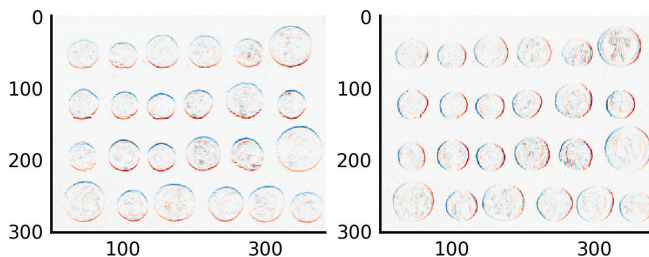
*ax* : ось графика matplotlib, подлежащая корректировке  
*factor* : целое, коэффициент сокращения количества меток оси X  
 «»

```
plt.setp(ax.xaxis.get_ticklabels(), visible=False)
for label in ax.xaxis.get_ticklabels()[::factor]:
    label.set_visible(True)
```

```
coins_h = ndi.convolve(coins, hsobel)
coins_v = ndi.convolve(coins, vsobel)
```

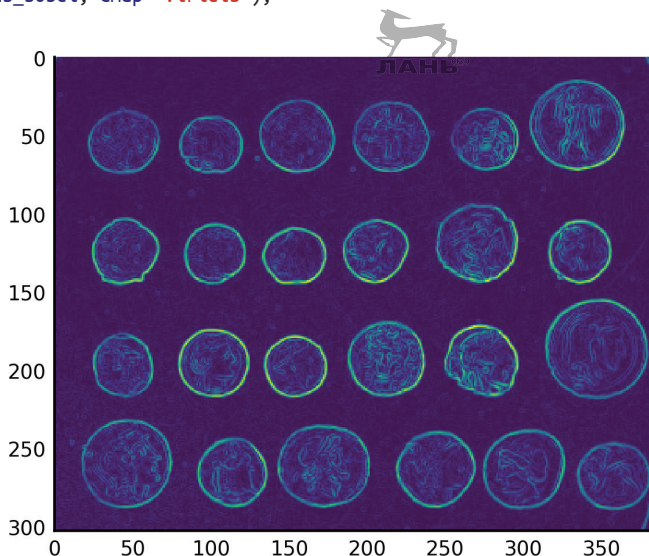
```
fig, axes = plt.subplots(nrows=1, ncols=2)
axes[0].imshow(coins_h, cmap=plt.cm.RdBu)
axes[1].imshow(coins_v, cmap=plt.cm.RdBu)
for ax in axes:
    reduce_xaxis_labels(ax, 2)
```





И наконец, вы можете констатировать, что, как в теореме Пифагора, величина края в *любом* направлении равна квадратному корню суммы квадратов горизонтальных и вертикальных составляющих:

```
coins_sobel = np.sqrt(coins_h**2 + coins_v**2)
plt.imshow(coins_sobel, cmap='viridis');
```



## УНИВЕРСАЛЬНЫЕ ФИЛЬТРЫ: ПРОИЗВОЛЬНЫЕ ФУНКЦИИ ОТ СОСЕДНИХ ЗНАЧЕНИЙ

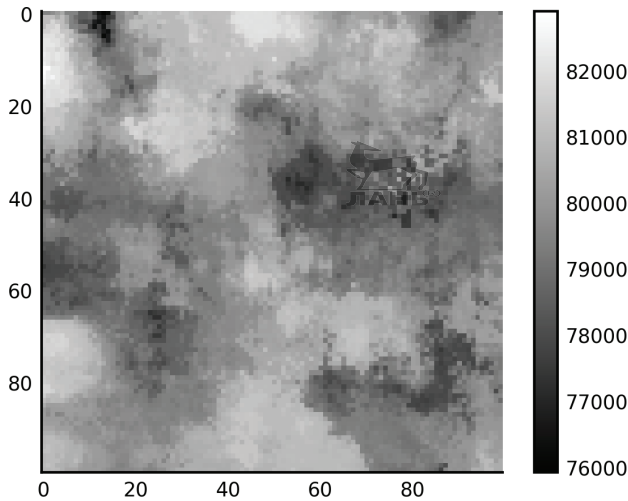
В дополнение к скалярным произведениям, реализованным в функции `ndi.convolve`, библиотека SciPy позволяет определять фильтр, являющийся *произвольной функцией* точек в окрестности. Этот фильтр реализован в функции `ndi.generic_filter`, позволяющей выражать произвольно сложные фильтры.

Предположим, что изображение представляет собой медианную стоимость домов в административном округе с пространственной разрешающей способностью 100 м в продольном и поперечном направлениях (100×100 м). Местный



совет принимает решение облагать налогом продажи домов на уровне \$10 000 плюс 5% от 90-й перцентили цен на дома в радиусе 1 километр. (Поэтому продажа дома в дорогом районе обходится дороже.) При помощи универсального (генерического) фильтра `generic_filter` мы можем создать карту налоговой ставки по всей географической карте округа:

```
from skimage import morphology
def tax(prices):
    return 10000 + 0.05 * np.percentile(prices, 90)
house_price_map = (0.5 + np.random.rand(100, 100)) * 1e6
footprint = morphology.disk(radius=10)
tax_rate_map = ndi.generic_filter(house_price_map, tax, footprint=footprint)
plt.imshow(tax_rate_map)
plt.colorbar();
```



## Задача: игра «Жизнь» Конуэя»

Предложено Николасом Ругиром

Игра «Жизнь» (англ. Conway's Game of Life) – это на вид простая конструкция, в которой «клетки» на регулярной квадратной сетке живут или умирают в зависимости от непосредственно окружающих их клеток. В каждый такт времени мы определяем состояние позиции  $(i, j)$  согласно его предыдущему состоянию и состоянию его восьми соседей (выше, ниже, слева, справа и по диагоналям):

- живая клетка умирает, если она имеет всего одного живого соседа или ни одного;
- живая клетка продолжает жить в течение еще одного поколения, если она имеет двух или трех живых соседей;
- живая клетка умирает от перенаселенности, если она имеет четырех или более живых соседей;

- мертвая клетка оживает вследствие воспроизводства, если она имеет ровно трех живых соседей.

Хотя эти правила смотрятся как надуманная математическая задача, они на самом деле дают начало невероятным шаблонам, начиная с планеров (небольших скоплений живых клеток, которые медленно перемещаются в каждое поколение) и орудий планеров (постоянных скоплений, которые вырастают из планеров), вплоть до генераторов простых чисел (например, почитайте статью «Генерирование последовательностей простых чисел в игре «Жизнь» Конуэя» Натаниеля Джонстона<sup>1</sup>) и даже симулирование самой игры «Жизнь»<sup>2</sup>!

Сможете ли вы реализовать игру «Жизнь» при помощи `ndi.generic_filter?`

Обратите внимание: решение задачи «Игра «Жизнь» Конуэя» находится в конце книги.

## Задача: магнитуда градиента Собела

Мы недавно увидели, каким образом можно объединять результаты двух разных фильтров: горизонтального фильтра Собела с вертикальным. Сможете ли вы написать функцию, которая делает это за один проход при помощи `ndi.generic_filter?`

Обратите внимание: решение задачи «Магнитуда градиента Собела» находится в конце книги.

## Графы и БИБЛИОТЕКА NETWORKX

Графы – это естественная форма описания, подходящая для удивительного разнообразия данных. Например, страницы в сети могут содержать узлы, тогда как связи между этими страницами могут быть, ну, в общем, связями. Или же в биологии в так называемых *транскрипционных сетях* гены представлены вершинами. При этом гены, имеющие непосредственное влияние на экспрессию друг друга, соединены ребрами.

✔ **Графы и сети.** В данном контексте термин «граф» синонимичен с термином «сеть» и не имеет никакого отношения к слову «график». Математики и специалисты в области информатики используют слегка отличающиеся слова, которые относятся к данной теме: граф = сеть, вершина = узел, ребро = связь = дуга. Как и большинство людей, мы будем использовать эти термины взаимозаменяемо.

Возможно, вам чуть больше знакома терминология сетей: сеть состоит из *узлов* и *связей* между узлами. Эквивалентно, граф состоит из *вершин* и *ребер* между вершинами. В библиотеке NetworkX имеются графовые объекты `Graph`, состоящие из узлов `node` и ребер `edge` между узлами. Такое использование терминов, вероятно, является наиболее распространенным.

<sup>1</sup> См. <http://www.njohnston.ca/2009/08/generating-sequences-of-primess-in-conways-game-of-life/>.

<sup>2</sup> См. <https://youtu.be/xP5-ileKXE8>.

Чтобы познакомить вас с графами, мы воспроизведем несколько результатов из исследовательской работы Лэва Варшни (Lav Varshney) и др. «Структурные свойства нейронной сети свободно живущей нематоды»<sup>1</sup>.

В этом примере мы представим нейроны в нервной системе червя нематоды в виде узлов и поместим ребро между двумя узлами, когда нейрон образует синапс с другим узлом. (*Синапсы* – это химические контакты, посредством которых нейроны передают информацию.) Червь предоставляет потрясающий пример анализа структуры контактов между нейронами, потому что каждый червь (этого вида) имеет одинаковое количество нейронов (302), и все контакты между ними известны. Данный факт привел к созданию фантастического проекта Openworm<sup>2</sup>, о котором мы рекомендуем почитать подробнее.

Вы можете скачать нейронный набор данных в формате Excel из базы данных WormAtlas<sup>3</sup>. Библиотека pandas позволяет читать таблицу Excel по сети, поэтому для прочтения данных мы воспользуемся этой библиотекой. Затем передадим эти данные в NetworkX.

```
import pandas as pd
connectome_url = 'http://www.wormatlas.org/images/NeuronConnect.xls'
conn = pd.read_excel(connectome_url)
```

Переменная conn теперь содержит объект библиотеки pandas DataFrame со строками в следующем формате:

```
[Нейрон1, Нейрон2, тип контакта, сила]
```

Мы собираемся исследовать только коннектом<sup>4</sup> химических синапсов, поэтому мы отфильтруем другие типы синапсов следующим образом:

```
conn_edges = [(n1, n2, {'weight': s})
               for n1, n2, t, s in conn.itertuples(index=False, name=None)
               if t.startswith('S')]
```

(Обратитесь к странице WormAtlas относительно описания различных типов контактов.) В приведенном выше словаре мы используем слово weight, так как это специальное ключевое слово для свойств ребер в библиотеке NetworkX. Затем мы строим граф, используя класс библиотеки NetworkX DiGraph:

```
import networkx as nx
wormbrain = nx.DiGraph()
wormbrain.add_edges_from(conn_edges)
```

Теперь можно приступить к исследованию некоторых свойств этой сети. Один из главных вопросов, который задают исследователи в отношении на-

<sup>1</sup> См. <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1001066>.

<sup>2</sup> См. <http://www.openworm.org/>.

<sup>3</sup> См. <http://www.wormatlas.org/neuronalwiring.html#Connectivitydata>.

<sup>4</sup> Коннектом (connectome) – это полное описание структуры связей в нервной системе организма. Они определяют поведение, с их помощью преобразуется вся информация из внутренней и внешней среды. – *Прим. перев.*



правленных сетей, – это какие узлы имеют внутри нее критическое значение для информационного потока. Узлы с высокой *центральностью по посредничеству* характеризуются тем, что находятся на кратчайшем пути между многочисленными отличающимися парами узлов. Представьте железнодорожную сеть: определенные станции будут связывать множество железнодорожных веток в таком порядке, что для различных поездок вам придется менять линии. Эти станции имеют высокую центральность по посредничеству.

При помощи библиотеки NetworkX мы можем легко находить не менее важные нейроны. В документации по API библиотеки NetworkX<sup>1</sup> раздел «centrality», строка документации `betweenness_centrality`<sup>2</sup>, определена функция, принимающая на входе граф и возвращающая словарь, в котором идентификаторам узлов поставлены в соответствие значения центральности по посредничеству (значения с плавающей запятой).

```
centrality = nx.betweenness_centrality(wormbrain)
```

Теперь мы можем найти нейроны с самой высокой центральностью, используя встроенную функцию Python `sorted`:

```
central = sorted(centrality, key=centrality.get, reverse=True)
print(central[:5])
```



```
['AVAR', 'AVAL', 'PVCR', 'PVT', 'PVCL']
```

В результате будут возвращены нейроны AVAR, AVAL, PVCR, PVT и PVCL, причастные к тому, как червь отвечает на нажатие: нейроны AVA (среди прочего) связывают передние рецепторы прикосновения червя с нейронами, которые отвечают за обратное движение, в то время как нейроны PVC связывают задние рецепторы прикосновения с движением вперед.

Варшни и др. провели исследование свойств *компонент сильной связности* 237 нейронов из общего числа 279 нейронов. В графах *компонента связности* представляет собой множество узлов, достижимых благодаря некому пути, проходящему через все связи. Коннектом представляется *направленный граф*, характеризующийся тем, что его ребра не просто соединяют узлы между собой, а *направлены* из одного узла в другой узел. В этом случае компонента сильной связности отличается достижимостью всех узлов друг из друга путем обхода связей *в правильном направлении*. Таким образом, путь  $A \rightarrow B \rightarrow C$  не характеризуется сильной связностью, так как отсутствует путь в A из B или C. И напротив, путь  $A \rightarrow B \rightarrow C \rightarrow A$  обладает сильной связностью.

В нейронной цепи компонента сильной связности может рассматриваться как «мозг» цепи, в котором происходит обработка, в то время как вышестоящие узлы являются входами, а нижестоящие узлы – выходами.

<sup>1</sup> См. <https://networkx.readthedocs.io/en/stable/>.

<sup>2</sup> См. <https://networkx.github.io/documentation/stable/reference/algorithms/centrality.html?highlight=centrality>.



**Циклы в нейронных сетях.** Идея циклических нейронных цепей восходит к 1950-м годам. Вот прекрасный абзац из статьи Аманды Джефтер в научном журнале Nautilus «Человек, который попытался искупить мир логикой»<sup>1</sup> с описанием этой идеи:

Если человек увидит вспышку молнии на небе, то его глаза пошлют сигнал в головной мозг, пронеся его через цепь нейронов. Начиная с любого заданного нейрона в цепи вы можете проследить шаги сигнала и выяснить, сколько времени прошло от вспышки молнии. С одним исключением – если только эта цепь не является циклом. В случае цикла информация, кодирующая вспышку молнии, будет до бесконечности вращаться в цикле. И она не будет иметь никакой связи со временем вспышки самой молнии. Как выразился Маккалок, она становится «идеей, вырванной из временного пространства». Другими словами, памятью.



Библиотека NetworkX облегчает работу по получению из сети wormbrain самой большой компоненты сильной связности:

```
sccs = nx.strongly_connected_component_subgraphs(wormbrain)
giantscc = max(sccs, key=len)
```

```
print(f'Самая большая компонента сильной связности имеет '
      f'{giantscc.number_of_nodes()} узлов из общего числа '
      f'{wormbrain.number_of_nodes()} узлов.')
```



Самая большая компонента сильной связности имеет 237 узлов из общего числа 279 узлов.

Как отмечается в исследовательской работе, размер этой компоненты *меньше*, чем ожидалось вследствие случайности. Это демонстрирует следующее: сеть подразделяется на три слоя: входной, центральный и выходной.

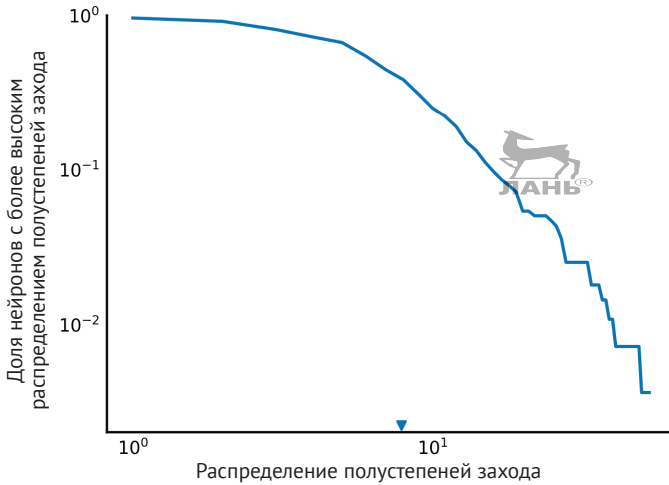
Теперь мы воспроизведем рис. 6В из исследовательской работы: функцию выживания распределения полустепеней захода. Сначала вычислим соответствующие количества:

```
in_degrees = list(dict(wormbrain.in_degree()).values())
in_deg_distrib = np.bincount(in_degrees)
avg_in_degree = np.mean(in_degrees)
cumfreq = np.cumsum(in_deg_distrib) / np.sum(in_deg_distrib)
survival = 1 - cumfreq
```

Затем, используя библиотеку Matplotlib, построим график:

```
fig, ax = plt.subplots()
ax.loglog(np.arange(1, len(survival) + 1), survival)
ax.set_xlabel('распределение полустепеней захода')
ax.set_ylabel('доля нейронов с более высоким распределением полустепеней захода')
ax.scatter(avg_in_degree, 0.0022, marker='v')
ax.text(avg_in_degree - 0.5, 0.003, 'среднее=%2f' % avg_in_degree)
ax.set_ylim(0.002, 1.0);
```

<sup>1</sup> См. <http://nautilus.us/issue/21/information/the-man-who-tried-to-redeem-the-world-with-logic>.



А вот и результат использования SciPy: воспроизведение научного анализа. Мы пропускаем подбор кривой ..., именно для этого и существуют упражнения.

### Задача: подбор кривой при помощи SciPy

Это упражнение является чем-то вроде предварительного ознакомления с материалом главы 7 (посвященной оптимизации): используйте функцию `scipy.optimize.curve_fit`, чтобы подогнать хвост функции выживания полустепеней захода под степенной закон,  $f(d) \sim d^{-\gamma}$ ,  $d > d_0$ , для  $d = 10$ , для  $d_0 = 10$  (красная прямая на рис. 6В указанной работы), и видоизмените график, чтобы включить эту прямую.

Обратите внимание: решение задачи «Подбор кривой при помощи SciPy» находится в конце книги.

Теперь у вас должно быть фундаментальное понимание не только графов как научной абстракции, но и как с помощью языка Python и библиотеки NetworkX можно легко этими графами манипулировать и анализировать их. Теперь мы пойдем дальше и обратимся к конкретному виду графов, используемых в обработке изображений и компьютерном зрении.

## ГРАФЫ СМЕЖНОСТИ ОБЛАСТЕЙ

Граф смежности областей (RAG) представляет изображение в виде, широко применяемом для *сегментации*, т. е. для подразделения изображений на содержательные области (или сегменты). Если вы видели к/ф «Терминатор-2», значит, вы видели и пример сегментации (рис. 3.2).







8868865868888888888888666655556686688665  
 688888866668888888888886666656686665  
 66888888456888899988886666556866655  
 666888886245666886666654431268686655  
 68688888668969666655655313668688655  
 6888888668998998885356888986655  
 6868889888866899999986666668986655  
 688888888666668886666666688866655  
 56888888868898688655566688886555  
 3666888888888888888688666688866655  
 26686888888888888888866688886654  
 28688888888888888886666686866655  
 286666888888888888866868888666548



Наша визуальная система оптимизирована определять лица так, что вы сможете увидеть лицо даже в этом скоплении чисел! Но мы надеемся, что донесли свою точку зрения. К тому же вы можете взглянуть на примеры лиц в Twitter<sup>1</sup>, которые с юмором демонстрируют оптимизацию обнаружения лиц нашими визуальными системами.

Во всяком случае, проблема состоит в придании этим числам смысла и в нахождении расположения границ, делящих различные части изображения. Популярный подход состоит в отыскании небольших областей (так называемых суперпикселей), принадлежащих тому же самому сегменту, и их объединения в соответствии с неким более сложным правилом.

В качестве простого примера предположим, что вам необходимо вычлнить в приведенном ниже изображении из набора данных эталонных сегментаций Университета Беркли (Berkeley Segmentation Dataset, BSDS) тигра.



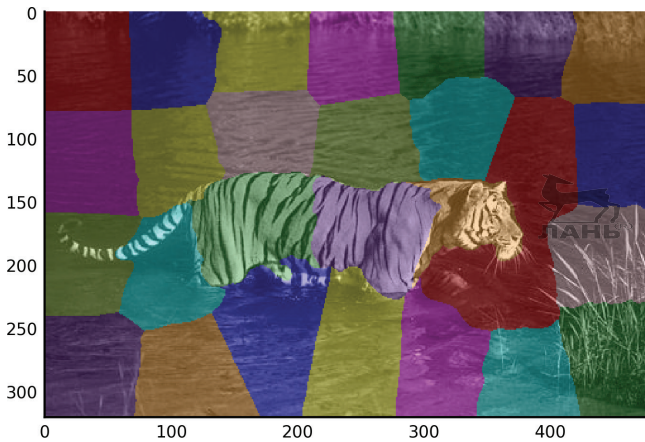
Алгоритм кластеризации под названием «Простая линейная итеративная кластеризация» (simple linear iterative clustering, SLIC) может предоставить нам хорошую отправную точку. Этот алгоритм находится в библиотеке scikit-image.

<sup>1</sup> См. <https://twitter.com/facespics>.

```
url = ('http://www.eecs.berkeley.edu/Research/Projects/CS/vision/'
      'bsds/BSDS300/html/images/plain/normal/color/108073.jpg')
tiger = io.imread(url)
from skimage import segmentation
seg = segmentation.slic(tiger, n_segments=30, compactness=40.0,
                       enforce_connectivity=True, sigma=3)
```

Библиотека `scikit-image` также имеет функцию для вывода сегментации на экран. Мы применим ее для визуализации результата работы алгоритма SLIC:

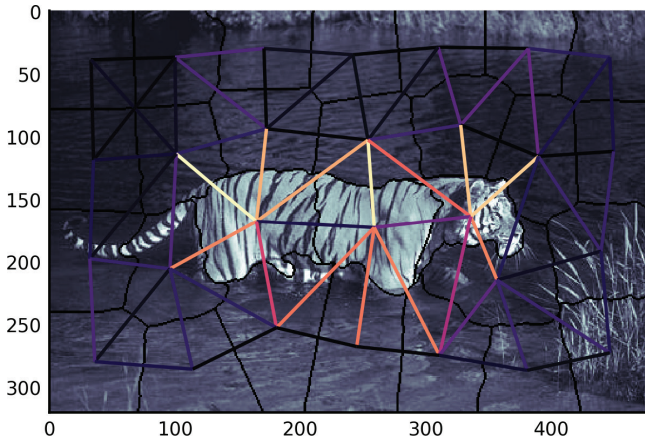
```
from skimage import color
io.imshow(color.label2rgb(seg, tiger));
```



Результат показывает, что тело тигра будет разбито на три части, а остальная часть изображения находится в других сегментах.

Граф смежности областей (RAG) – это граф, в котором каждый узел представляет одну из вышеупомянутых областей, а ребро соединяет два узла при их соприкосновении. Чтобы до построения собственного графа посмотреть, как это работает, мы используем функцию `show_rag` из библиотеки `scikit-image`. На самом деле эта библиотека содержит фрагмент кода данной главы!

```
from skimage.future import graph
g = graph.rag_mean_color(tiger, seg)
graph.show_rag(seg, g, tiger);
```



Здесь вы видите узлы, соответствующие каждому сегменту, и ребра между смежными сегментами. Они, в соответствии с разницей в цвете между двумя узлами, окрашены в палитру YlGnBu (желтый-зеленый-синий) из библиотеки Matplotlib.

Приведенный выше рисунок также показывает чудодейственность взгляда на сегментацию как на графы: вы видите, что ребра между узлами внутри тигра и ребра за его пределами ярче (т. е. с более высокой величиной), чем ребра внутри самого объекта. Таким образом, если мы сможем вырезать граф вдоль этих ребер, получим нашу сегментацию. Мы выбрали простой пример сегментации на основе цвета, однако те же самые принципы сохраняются и для графов с более сложными попарными связями.

## ЭЛЕГАНТНЫЙ ПАКЕТ NDIMAGE: КАК СТРОИТЬ ГРАФЫ ИЗ ОБЛАСТЕЙ ИЗОБРАЖЕНИЙ

Теперь все части на месте: вы знакомы с массивами NumPy, фильтрацией изображений, универсальными фильтрами и графами, в частности с графами RAG. Давайте построим граф RAG, чтобы выдернуть тигра из этой фотографии!

Очевидный подход состоит в применении двух вложенных циклов `for` для перебора всех пикселей изображения, просмотра соседних пикселей и проверки на разницу меток:

```
import networkx as nx

def build_rag(labels, image):
    g = nx.Graph()
    nrows, ncols = labels.shape
    for row in range(nrows):
        for col in range(ncols):
            current_label = labels[row, col]
```

```

if not current_label in g:
    g.add_node(current_label)
    g.node[current_label]['total color'] = np.zeros(3,
                                                    dtype=np.float)
    g.node[current_label]['pixel count'] = 0
if row < nrows - 1 and labels[row + 1, col] != current_label:
    g.add_edge(current_label, labels[row + 1, col])
if col < ncols - 1 and labels[row, col + 1] != current_label:
    g.add_edge(current_label, labels[row, col + 1])
g.node[current_label]['total color'] += image[row, col]
g.node[current_label]['pixel count'] += 1
return g

```

Отлично! Все работает. Однако если вы захотите сегментировать трехмерное изображение, вам придется написать другую версию:

```
import networkx as nx
```

```

def build_rag_3d(labels, image):
    g = nx.Graph()
    nplns, nrows, ncols = labels.shape
    for pln in range(nplns):
        for row in range(nrows):
            for col in range(ncols):
                current_label = labels[pln, row, col]
                if not current_label in g:
                    g.add_node(current_label)
                    g.node[current_label]['total color'] = np.zeros(3,
                                                                    dtype=np.float)
                    g.node[current_label]['pixel count'] = 0
                if pln < nplns - 1 and labels[pln + 1, row, col] != \
                    current_label:
                    g.add_edge(current_label, labels[pln + 1, row, col])
                if row < nrows - 1 and labels[pln, row + 1, col] != \
                    current_label:
                    g.add_edge(current_label, labels[pln, row + 1, col])
                if col < ncols - 1 and labels[pln, row, col + 1] != \
                    current_label:
                    g.add_edge(current_label, labels[pln, row, col + 1])
                g.node[current_label]['total color'] += image[pln, row, col]
                g.node[current_label]['pixel count'] += 1
    return g

```

Обе версии довольно страшные и громоздкие. И более того, их трудно расширить. Если мы захотим подсчитать диагонально соседние пиксели как смежные (т. е. [row, col] «смежен с» [row + 1, col + 1]), то программный код станет еще запутаннее. И если мы захотим проанализировать трехмерное видео, то потребуется еще одна размерность и еще один уровень вложенности. Полный бардак!

Взгляните на эту проблему с точки зрения Винеша: функция универсальной фильтрации SciPy `generic_filter` уже делает эту итерацию за нас! Мы исполь-

зовали ее выше, чтобы вычислить произвольно сложную функцию на окрестности каждого элемента массива NumPy. Только теперь от этой функции нам требуется не отфильтрованное изображение, а граф. Оказывается, функция `generic_filter` позволяет передавать в функцию фильтрации дополнительные аргументы. И мы можем это задействовать для построения графа:

```
import networkx as nx
import numpy as np
from scipy import ndimage as nd

def add_edge_filter(values, graph):
    center = values[len(values) // 2]
    for neighbor in values:
        if neighbor != center and not graph.has_edge(center, neighbor):
            graph.add_edge(center, neighbor)
    # Возвращаемое вещественное значение не используется, но
    # требуется функцией `generic_filter`
    return 0.0

def build_rag(labels, image):
    g = nx.Graph()
    footprint = ndi.generate_binary_structure(labels.ndim, connectivity=1)
    _ = ndi.generic_filter(labels, add_edge_filter, footprint=footprint,
                           mode='nearest', extra_arguments=(g,))

    for n in g:
        g.node[n]['total color'] = np.zeros(3, np.double)
        g.node[n]['pixel count'] = 0
    for index in np.ndindex(labels.shape):
        n = labels[index]
        g.node[n]['total color'] += image[index]
        g.node[n]['pixel count'] += 1
    return g
```



Вот несколько причин, почему это действительно прекрасный фрагмент кода:

- функция `ndi.generic_filter` перебирает все элементы массива вместе с их соседями (используйте `numpy.ndindex`, чтобы перебирать только индексы массива);
- из функции фильтрации мы возвращаем «0.0», так как `generic_filter` требует, чтобы функция фильтрации возвращала вещественное значение. Однако мы игнорируем результат фильтрации (который повсюду равеняется нулю) и используем его только из-за его «побочного эффекта», заключающегося в добавлении в граф ребер;
- циклы не вложены на несколько уровней в глубину. Это делает программный код компактнее и проще для восприятия;
- этот программный код работает не только для одно-, двух-, трех-, но даже и для восьмимерных изображений!
- чтобы добавить поддержку диагональной связности, достаточно поменять параметр `connectivity` на `ndi.generate_binary_structure`.

## СОБИРАЕМ ВСЕ ВМЕСТЕ: СЕГМЕНТАЦИЯ ПО СРЕДНЕМУ ЦВЕТУ

Теперь, чтобы сегментировать тигра на приведенном выше изображении, мы можем применить все, чему научились:



```
g = build_rag(seg, tiger)
for n in g:
    node = g.node[n]
    node['mean'] = node['total color'] / node['pixel count']
for u, v in g.edges:
    d = g.node[u]['mean'] - g.node[v]['mean']
    g[u][v]['weight'] = np.linalg.norm(d)
```

Каждое ребро содержит разность между средним цветом каждого сегмента. Сейчас мы зададим порог для графа:

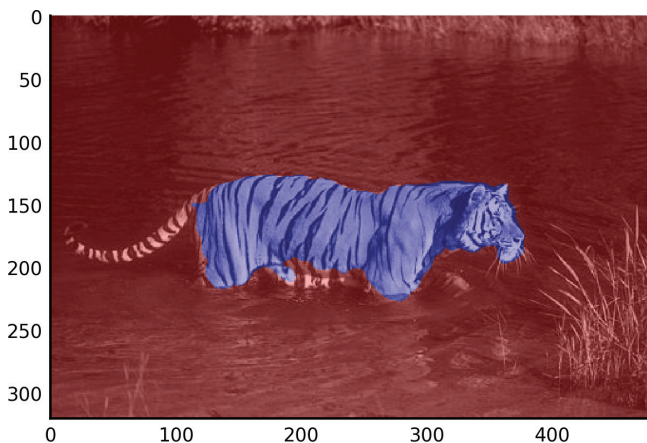
```
def threshold_graph(g, t):
    to_remove = [(u, v) for (u, v, d) in g.edges(data=True)
                 if d['weight'] > t]
    g.remove_edges_from(to_remove)
threshold_graph(g, 80)
```

И применим трюк NumPy с индексацией в массиве, с которым мы познакомились в главе 2:

```
map_array = np.zeros(np.max(seg) + 1, int)
for i, segment in enumerate(nx.connected_components(g)):
    for initial in segment:
        map_array[int(initial)] = i
segmented = map_array[seg]
plt.imshow(color.label2rgb(segmented, tiger));
```



Ой! Похоже, кошка потерял хвост!





И тем не менее мы считаем, что это была прекрасная демонстрация возможностей графов RAG и красоты, с которой библиотеки SciPy и NetworkX их делают выполнимыми. Многие из этих функций имеются в библиотеке scikit-image. Если вы интересуетесь анализом изображений, то с этой библиотекой следует познакомиться поближе!



---

# Глава 4

.....

## Частота и быстрое преобразование Фурье

Если вы хотите раскрыть тайны вселенной, думайте с точки зрения энергии, частоты и вибрации.



— Никола Тесла

*Эта глава была написана в сотрудничестве с отцом Штефана, ПВ ван дер Уолтом (PW van der Walt).*

Данная глава немного отойдет от формата остальной части книги. Здесь, в частности, вы обнаружите не очень большой программный код. Цель этой главы – показать элегантный и очень полезный алгоритм под названием «быстрое преобразование Фурье» (БПФ, Fast Fourier Transform, FFT), реализованный в SciPy и, разумеется, работающий с массивами NumPy.

### ВВЕДЕНИЕ В ЧАСТОТУ

Как всегда, мы начнем с настройки стиля графиков и импорта обычных составляющих:

*# Заставить графики появляться локально, задать индивидуальный стиль графиков*

```
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('style/elegant.mplstyle')
import numpy as np
```

Дискретное<sup>1</sup> преобразование Фурье (ДПФ, Decrete Fourier Transform, DFT) – это математический метод, используемый для преобразования временных или пространственных данных в данные *частотной области*. Понятие *частоты* широко известно. Благодаря звуковым колебаниям разной частоты

---

<sup>1</sup> ДПФ оперирует с выборочными данными, в отличие от стандартного преобразования Фурье, которое определено для непрерывных данных.



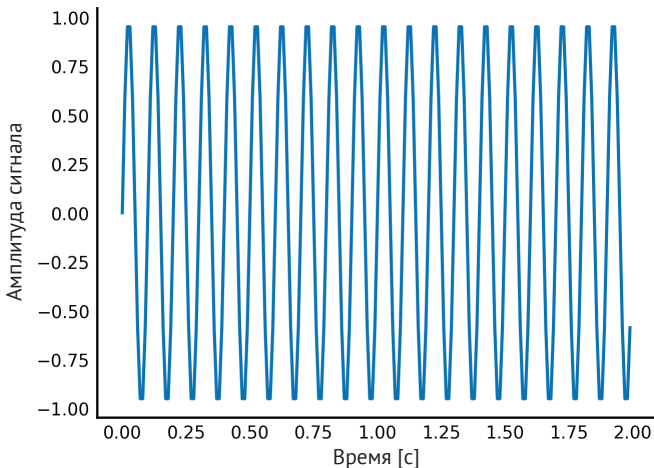


мы слышим друг друга. Частоты, которые слышит человек, лежат в пределах от 20 и до 16 000 Гц. Но способность слышать крайние частоты у каждого человека строго индивидуальна. Самые низкие ноты, воспроизводимые головными телефонами, имеют частоту порядка 20 Гц, а средняя нота «до» на фортепьяно имеет частоту примерно 261.6 Гц. Герцы, или колебания в секунду, в данном случае буквально означают количество движений мембраны в наушнике из стороны в сторону в секунду. Эти движения, в свою очередь, создают сжатые импульсы воздуха, которые, достигнув вашей барабанной перепонки, вызывают вибрацию с той же самой частотой. Так, если взять простую периодическую функцию,  $\sin(10 \times 2\pi t)$ , то ее можно представить как волну:

```
f = 10 # Частота колебаний в циклах в секунду, или в герцах
f_s = 100 # Частота дискретизации (количество замеров в секунду)

t = np.linspace(0, 2, 2 * f_s, endpoint=False)
x = np.sin(f * 2 * np.pi * t)

fig, ax = plt.subplots()
ax.plot(t, x)
ax.set_xlabel('Время [с]')
ax.set_ylabel('Амплитуда сигнала');
```



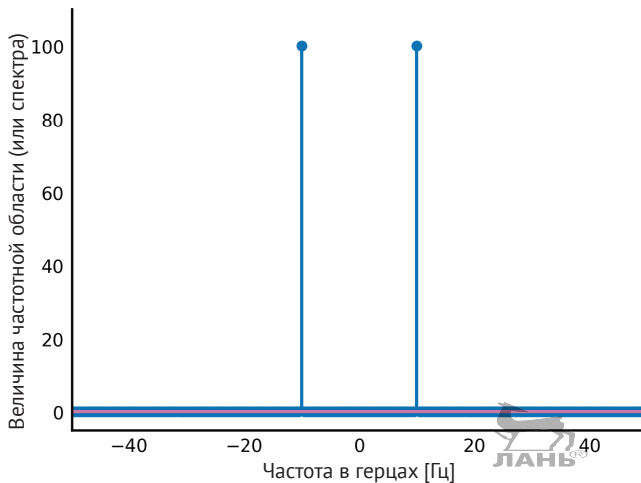
Или как повторяющийся сигнал частотой 10 Гц (он повторяется один раз в 1/10 секунды – этот отрезок времени мы называем периодом). Хотя мы, вполне естественно, связываем частоту со временем, она может одинаково хорошо применяться и к пространству. Например, фотография текстильных узоров показывает высокую *пространственную частоту*, тогда как небо или другие гладкие объекты имеют низкую пространственную частоту.

Теперь давайте исследуем нашу синусоиду, применив БПФ:

```

from scipy import fftpack
X = fftpack.fft(x)
freqs = fftpack.fftfreq(len(x)) * f_s
fig, ax = plt.subplots()
ax.stem(freqs, np.abs(X))
ax.set_xlabel('Частота в герцах [Гц]')
ax.set_ylabel('Величина частотной области (или спектра)')
ax.set_xlim(-f_s / 2, f_s / 2)
ax.set_ylim(-5, 110)
(-5, 110)

```



Мы видим, что на выходе из БПФ мы получим одномерный массив той же самой формы, содержащий комплексные величины, что и на входе. Все величины, за исключением двух записей, равны нулю. По традиции мы визуализируем магнитуду результата в виде графика стебель–листья, в котором высота каждого стебля соответствует базовой величине.

(Позже в разделе «Дискретные преобразования Фурье» мы объясним, почему вы видите положительные и отрицательные колебания частоты. Вы также можете обратиться к упомянутому разделу для получения более глубокого обзора лежащих в основе математических выкладок.)

Преобразование Фурье уводит нас из *временной* области в *частотную*, и это влечет за собой огромное количество применений. Быстрое преобразование Фурье (БПФ) представляет собой алгоритм вычисления ДПФ. Его высокое быстродействие достигается за счет сохранения и повторного использования результатов вычислений по мере продвижения.

В этой главе мы исследуем несколько применений ДПФ, которые продемонстрируют, что БПФ может применяться к многомерным (а не только к одномерным) данным, достигая при этом самых разнообразных целей.

## Иллюстрация: СПЕКТРОГРАММА ПЕНИЯ ПТИЦ

Давайте начнем с одного из наиболее распространенных применений – преобразования аудиосигнала (состоящего из вариаций давления воздуха во времени) в спектрограмму. Вы, возможно, уже встречали спектрограммы в окне эквалайзера своего аудиоплеера или даже в старомодном стереопроеигрывателе (рис. 4.1).

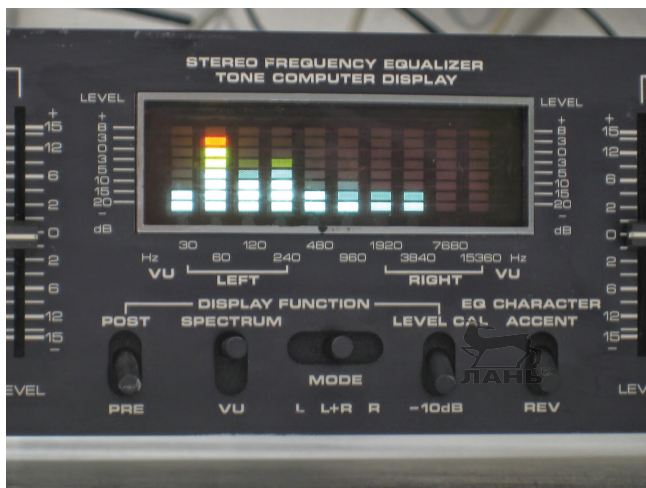


Рис. 4.1 ❖ Стереоэквалайзер Numark EQ2600<sup>1</sup>  
(изображение используется с разрешения автора Сергея Герасимука)

Прослушайте этот отрывок пения соловья<sup>2</sup> (выпущенный в соответствии с лицензией CC 4.0):

```
from IPython.display import Audio
Audio('data/nightingale.wav')
```



Если вы читаете бумажную версию этой книги, то вам придется применить свое воображение! Он звучит примерно так: чи-чи-вурррр-хи-хи чиит-виит-хуррр-чирр-ви-вео-вео-вео-вео-вео-вео.

<sup>1</sup> См. <https://sgerasimuk.blogspot.ru/2014/06/numark-eq-2600-10-band-stereo-graphic.html>.

<sup>2</sup> См. <http://www.orange-freesounds.com/nightingale-sound/>.

Поскольку мы понимаем, что не все бегло говорят на птичьем языке, будет лучше выполнить визуализацию всех замеров – именуемых как «сигнал».

Мы загрузим аудиофайл, который даст нам частоту дискретизации (количество замеров в секунду), а также аудиоданные в виде массива формы (N, 2) – два столбца, потому что это стереозапись.

```
from scipy.io import wavfile
rate, audio = wavfile.read('data/nightingale.wav')
```

Далее преобразуем этот файл в моно, усреднив левый и правый каналы.

```
audio = np.mean(audio, axis=1)
```

Затем мы вычислим длину отрывка и выведем аудио на график (рис. 4.2).

```
N = audio.shape[0]
L = N / rate
print(f'Длина аудио: {L:.2f} секунд')
f, ax = plt.subplots()
ax.plot(np.arange(N) / rate, audio)
ax.set_xlabel('Время [с]')
ax.set_ylabel('Амплитуда [неизвестно]');
```

Длина аудио: 7.67 секунды

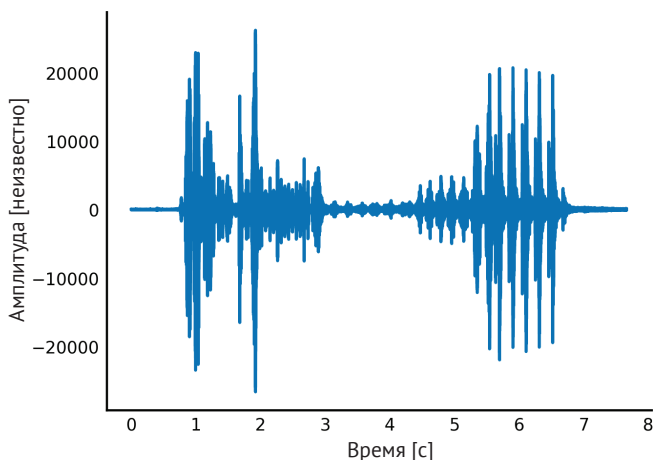


Рис. 4.2 ❖ График формы аудиосигнала пения соловья

Надо сказать, его вид не очень-то вдохновляет, не правда ли? Если бы я направил этот сигнал в громкоговоритель, то смог бы услышать щебет птицы. Но я не смогу достаточно хорошо представить, как оно будет звучать в моей голове. Есть ли более подходящий способ *увидеть*, что происходит?

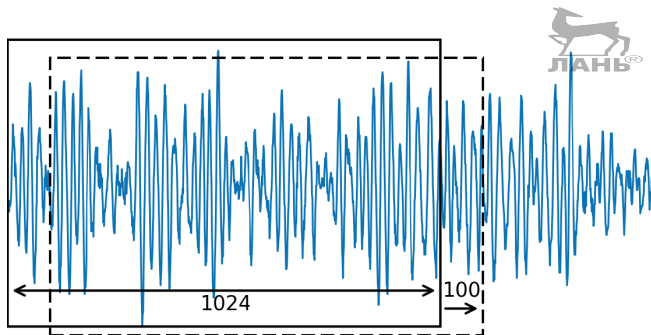
Да, есть. И этот метод называется дискретным преобразованием Фурье, или ДПФ. Слово «дискретный» относится к записи, состоящей, в отличие от непре-

рывной записи (как, например, на магнитной ленте катушечного магнитофона или аудиокассете), из разделенных во времени замеров звука. ДПФ часто вычисляется с использованием алгоритма БПФ, название которого неофициально используется для обозначения ДПФ как такового. ДПФ сообщает нам о том, какие частоты или «ноты» можно ожидать в нашем сигнале.

Конечно же, в пении птицы содержится большое количество нот, поэтому мы также хотели бы знать, когда появляется каждая нота. Преобразование Фурье берет сигнал во временном интервале (т. е. серию замеров во времени) и превращает его в спектр – серию частот с соответствующими (комплексными<sup>1</sup>) величинами. Спектр не содержит какой-либо информации о времени<sup>2</sup>!

Так вот, чтобы найти частоты и временную точку, когда они были спеты, нам придется применить неформальный подход. Наша стратегия будет следующей: взять аудиосигнал, разделить его на небольшие накладывающиеся фрагменты и применить преобразование Фурье к каждому из них (такой прием называется кратковременным преобразованием Фурье).

Мы разделим сигнал на фрагменты, состоящие из 1024 выборок (сэмплов), – продолжительность каждого составит порядка 0.02 секунды аудио. Почему мы выбрали именно 1024, а не 1000, мы объясним через секунду, когда займемся исследованием производительности. Фрагменты будут накладываться на 100 выборок, как показано ниже:



Начнем с нарезки сигнала на фрагменты по 1024 выборки, причем каждый фрагмент будет накладываться на предыдущие 100 выборок. Результирующий объект `slices` содержит один фрагмент в расчете на строку.

<sup>1</sup> Преобразование Фурье, по существу, говорит о том, как объединить серию синусоид переменной частоты, чтобы сформировать входной сигнал. Спектр состоит из комплексных чисел – одно число для каждой синусоиды. В комплексном числе закодированы две вещи: магнитуда и угол. Магнитуда – это сила синусоиды в сигнале, и угол – насколько она смещена во времени. На данном этапе нас интересует только магнитуда, которую мы вычисляем с использованием `np.abs`.

<sup>2</sup> Для получения дополнительной информации о методах вычисления (приблизительных) частот и времени появления почитайте материалы по анализу формы сигнала, или по вейвлет-анализу.



```
from skinage import util
```

```
M = 1024
```

```
slices = util.view_as_windows(audio, window_shape=(M,), step=100)
print(f'Форма аудио: {audio.shape}, форма нарезанного аудио: {slices.shape}')
```

```
Форма аудио: (338081,), форма нарезанного аудио: (3371, 1024)
```

Сгенерируем оконную функцию (см. ниже раздел «Оконное преобразование», в котором обсуждаются лежащие в основе допущения и интерпретации) и умножим ее на сигнал:

```
win = np.hanning(M + 1)[: -1]
slices = slices * win
```

Удобнее иметь один фрагмент в расчете на столбец. Поэтому мы транспонируем массив:

```
slices = slices.T
print('Форма объекта `slices`:', slices.shape)
```



```
Форма объекта `slices`: (1024, 3371)
```

Для каждого фрагмента вычислим ДПФ-преобразование, возвращающее положительные и отрицательные частоты (подробнее об этом ниже в разделе «Частоты и их упорядочивание»). На данный момент мы нарежем положительные M2-частоты.

```
spectrum = np.fft.fft(slices, axis=0)[:M // 2 + 1: -1]
spectrum = np.abs(spectrum)
```

(В примечании заметим, что используются взаимозаменяемые функции `scipy.fftpack.fft` и `np.fft`. Библиотека NumPy обеспечивает базовый функционал БПФ, расширяемый библиотекой SciPy. При этом обе библиотеки содержат функцию `fft`, основанную на динамической библиотеке FFTPACK языка Fortran.)

Спектр может содержать очень большие и очень малые значения. Если взять логарифм этого спектра, то этот диапазон значений будет существенно сжат.

Ниже мы строим логарифмический график соотношения сигнала, деленного на максимальный сигнал (график показан на рис. 4.3). Специфической единицей измерения, используемой для данного соотношения, является децибел,  $20 \log_{10}$  (амплитудный коэффициент).

```
f, ax = plt.subplots(figsize=(4.8, 2.4))
S = np.abs(spectrum)
S = 20 * np.log10(S / np.max(S))

ax.imshow(S, origin='lower', cmap='viridis',
           extent=(0, L, 0, rate / 2 / 1000))
ax.axis('tight')
ax.set_ylabel('Частота [кГц]')
ax.set_xlabel('Время [с]');
```

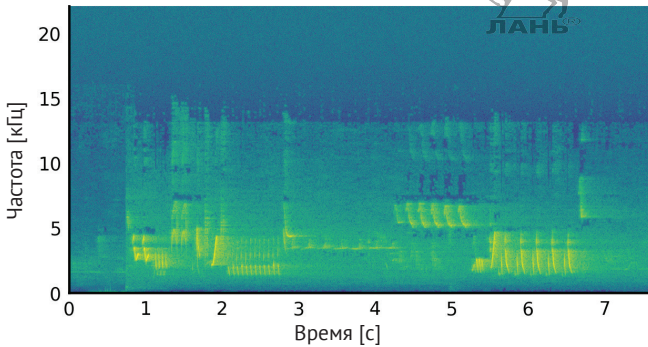


Рис. 4.3 ❖ Спектрограмма пения птицы

Теперь намного лучше! Мы видим, что частоты варьируются во времени, и спектрограмма соответствует тому, как звучит аудио. Проверьте, можно ли сопоставить наше предыдущее описание: чи-чи-вуррррр-хи-хи чиит-виит-хурррр-чирр-ви-вео-вео-вео-вео-вео-вео-вео. (Я не расшифровал вторую отметку с 3 по 5 – это другая птица.)

Библиотека SciPy уже содержит реализацию этой процедуры как `scipy.signal.spectrogram` (рис. 4.4), которую можно вызвать следующим образом:

```
from scipy import signal
freqs, times, Sx = signal.spectrogram(audio, fs=rate, window='hanning',
                                     nperseg=1024, noverlap=M - 100,
                                     detrend=False, scaling='spectrum')
f, ax = plt.subplots(figsize=(4.8, 2.4))
ax.pcolormesh(times, freqs / 1000, 10 * np.log10(Sx), cmap='viridis')
ax.set_ylabel('Частота [кГц]')
ax.set_xlabel('Время [с]');
```

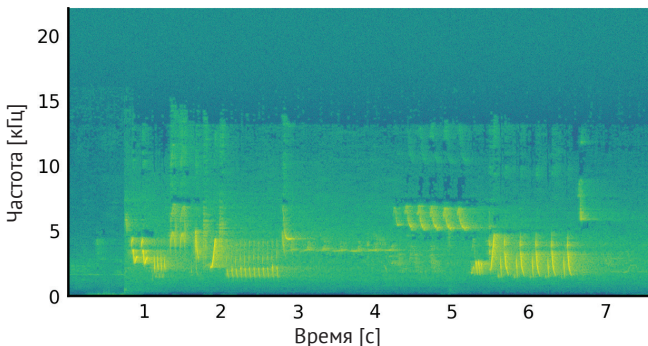


Рис. 4.4 ❖ Встроенное в SciPy исполнение спектрограммы пения птицы



Единственная разница между спектрограммой, созданной нами вручную, и встроенной в SciPy функцией состоит в том, что SciPy возвращает квадрат магнитуды спектра (превращающий измеренное напряжение в измеренную энергию) и умножает его на коэффициенты нормализации<sup>1</sup>.

## ИСТОРИЯ

Отследить точное происхождение преобразования Фурье довольно сложно. Некоторые связанные с ним процедуры восходят еще к временам Вавилона. Вместе с тем это преобразование было актуальным для орбит астероидов и решения уравнения теплопроводности, что в итоге привело к нескольким прорывам в начале 1800-х годов. Кого мы должны благодарить: Клеро, Лагранжа, Эйлера, Гаусса и Д'Аламбера, – не совсем понятно. Но Гаусс был первым, кто описал быстрое преобразование Фурье (т. е. алгоритм вычисления ДПФ, популяризированный Кули и Тьюки в 1965 г.). Жан Батист Жозеф Фурье, в честь которого это преобразование названо, был первым, кто утверждал, что *произвольные* периодические<sup>2</sup> функции могут быть выражены как сумма тригонометрических функций.

## РЕАЛИЗАЦИЯ

Функционал ДПФ библиотеки SciPy расположен в модуле `scipy.fftpack`. Кроме того, он обеспечивает следующую связанную с ДПФ функциональность:

`fft`, `fft2`, `fftn`

Вычисляют ДПФ, используя алгоритм БПФ в 1, 2 или  $n$  размерностях.

`ifft`, `ifft2`, `ifftn`

Вычисляют инверсию ДПФ.

`dct`, `idct`, `dst`, `idst`

Вычисляют косинус- и синус-преобразования и их инверсии.

`fftshift`, `ifftshift`

Смещают постоянную компоненту (компоненту с нулевой частотой) соответственно в центр спектра и назад (подробнее об этом ниже).

<sup>1</sup> Библиотека SciPy прилагает некоторые усилия по сохранению энергии в спектре. Поэтому, беря только половину компонент (для  $N$  четного), она умножает оставшиеся компоненты, кроме первой и последней, на два (эти две компоненты используются «совместно» двумя половинами спектра). Она также нормализует окно, деля его на их сумму.

<sup>2</sup> На самом деле период тоже может быть бесконечным! Обобщенное непрерывное преобразование Фурье предусматривает эту возможность. ДПФ-преобразования, как правило, определяются на конечном интервале, и этот интервал неявно является периодом функции преобразуемой временной области. Другими словами, если вы выполняете обратное дискретное преобразование Фурье, то на выходе вы *всегда* получите периодический сигнал.



`fftfreq`

Возвращает выборочные частоты ДПФ.

`rfft`

Для увеличения производительности вычисляет ДПФ вещественной последовательности, эксплуатируя симметрию результирующего спектра. В зависимости от ситуации также используется функцией `fft` для внутренних целей.

Этот список дополняется следующими ниже функциями библиотеки NumPy:

`np.hanning`, `np.hamming`, `np.bartlett`, `np.blackman`, `np.kaiser`

Функции суженного оконного преобразования.



ДПФ также применяется для выполнения быстрой свертки больших входных данных функцией `scipy.signal.fftconvolve`.

Библиотека SciPy служит оберткой для динамической библиотеки FFTPACK языка Fortran. Это не самая быстрая библиотека, но, в отличие от таких пакетов, как FFTW, она имеет разрешительную лицензию бесплатного программного обеспечения.

## Выбор длины ДПФ

Для простого вычисления ДПФ требуется  $\mathcal{O}(N^2)$  операций<sup>1</sup>. Почему? Дело в том, что у вас  $N$  (комплексных) синусоид разных частот ( $2\pi f \times 0$ ,  $2\pi f \times 1$ ;  $2\pi f \times 3$ , ...,  $2\pi f \times (N - 1)$ ), и вы хотите увидеть, насколько ваш сигнал соответствует каждой из них. Начиная с первой вы берете скалярное произведение с сигналом (который сам по себе влечет за собой  $N$  операций умножения). Тогда повторение этой операции  $N$  раз, один раз для каждой синусоиды, дает  $N^2$  операций.

Теперь сопоставьте это с алгоритмом БПФ, имеющим в идеальном случае вычислительную сложность  $\mathcal{O}(N \log N)$ . Это достигнуто благодаря грамотному повторному использованию результатов вычислений. Громадное улучшение! Однако классический алгоритм Кули-Тьюки, реализованный в динамической библиотеке FFTPACK (и используемый библиотекой SciPy), рекурсивно раз-

<sup>1</sup> В информатике вычислительная сложность алгоритма часто выражается в форме математического обозначения « $\mathcal{O}$ » большое. Это обозначение указывает, как масштабируется время выполнения алгоритма с ростом количества элементов. Если алгоритм имеет сложность  $\mathcal{O}(N)$ , то время его выполнения увеличивается линейно вместе с количеством входных элементов (например, поиск заданного значения в неотсортированном списке имеет сложность  $\mathcal{O}(N)$ ). Сортировка пузырьком является примером алгоритма со сложностью  $\mathcal{O}(N^2)$ . Точное количество выполненных операций теоретически может равняться  $N + \frac{1}{2}N^2$ . При этом считается, что вычислительная сложность растет квадратически вместе с количеством входных элементов.

бывает преобразование на меньшие фрагменты (имеющие размер, равный простому числу) и показывает это улучшение только для «гладких» входных длин (входная длина считается гладкой, когда ее самый большой простой множитель является малым, как показано на рис. 4.5). Для фрагментов с размером, равным большому простому числу, вместе с алгоритмом Кули-Тьюки могут применяться алгоритмы Блуштайна или Рейдера. Но такая оптимизация в FFTPACK не реализована<sup>1</sup>.

Давайте посмотрим на примере:

```
import time
from scipy import fftpack
from sympy import factorint

K = 1000
lengths = range(250, 260)

# Вычислить гладкость для всех входных длин
smoothness = [max(factorint(i).keys()) for i in lengths]

exec_times = []
for i in lengths:
    z = np.random.random(i)

    # Для каждой входной длины i исполнить БПФ K раз и
    # сохранить время исполнения

    times = []
    for k in range(K):
        tic = time.monotonic()
        fftpack.fft(z)
        toc = time.monotonic()
        times.append(toc - tic)

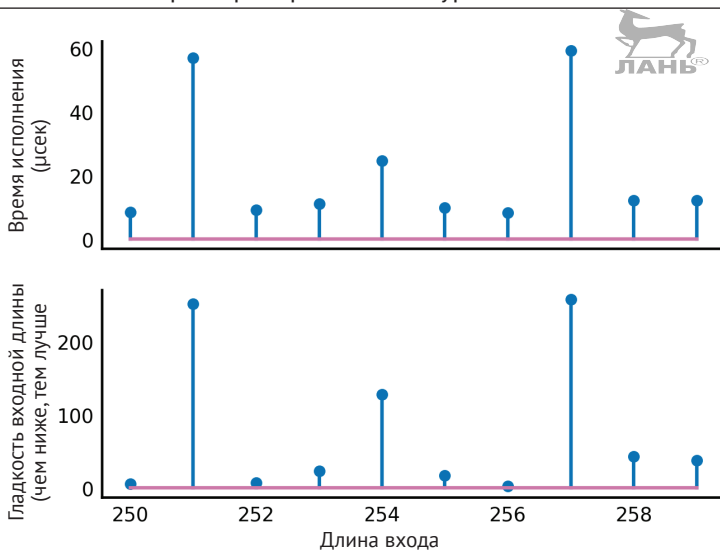
    # Для каждой входной длины запомнить *минимальное* время исполнения
    exec_times.append(min(times))

f, (ax0, ax1) = plt.subplots(2, 1, sharex=True)
ax0.stem(lengths, np.array(exec_times) * 10**6)
ax0.set_ylabel('Время исполнения (мсек)')

ax1.stem(lengths, smoothness)
ax1.set_ylabel('Гладкость входной длины\n(чем ниже, тем лучше)')
ax1.set_xlabel('Длина входа');
```



<sup>1</sup> Хотя в идеальном случае нам бы не хотелось повторно реализовывать существующие алгоритмы, иногда возникает необходимость получить наилучшие возможные скорости исполнения, и такие инструменты, как Cython, компилирующий Python на C, и Numba, производящий JIT-компиляцию программного кода Python, намного облегчают жизнь (и ускоряют работу алгоритмов!). Если у вас есть возможность использовать программное обеспечение с общедоступной GPL-лицензией, то вы могли бы рассмотреть использование библиотеки PyFFTW, в которой применяются более быстрые алгоритмы БПФ.



**Рис. 4.5** ❖ Время исполнения БПФ относительно гладкости разных входных длин

Интуитивный вывод состоит в том, что для гладких количеств преобразование БПФ может быть разбито на множество небольших фрагментов. После выполнения БПФ на первом фрагменте мы можем повторно воспользоваться этими результатами в последующих вычислениях. Это объясняет, почему ранее для наших кусков аудио мы выбрали длину 1024, – она имеет гладкость, равную всего 2, что приводит к оптимальному алгоритму Кули-Тьюки для «корня степени 2», вычисляемому БПФ, используя всего  $(N/2)\log_2 N = 5120$  комплексных умножений вместо  $N^2 = 1\,048\,576$ . Выбор  $N = 2^m$  всегда гарантирует максимально гладкое  $N$  (и, значит, самое быстрое исполнение алгоритма БПФ).

## Дополнительные понятия ДПФ

Далее мы представим несколько общих понятий, которые стоит узнать, прежде чем приступать к работе с тяжелыми механизмами преобразования Фурье. Мы займемся решением еще одной практической задачи: анализом обнаружения цели в радиолокационных данных.

## Частоты и их упорядочивание

По историческим причинам большинство реализаций возвращает массив, в котором частоты варьируются от низких до высоких и снова низких (см. раздел «Дискретные преобразования Фурье» относительно более подробного объяснения частот). Например, когда мы выполняем вещественное преобразование сигнала, состоящего только из единиц, вход не имеет вариации. Поэтому на

входе в качестве первой записи появляется лишь самая медленная постоянная компонента Фурье (так называемая «DC-компонента», или постоянно токовая компонента – жаргон из радиоэлектроники, обозначающий просто «среднее значение сигнала»):

```
from scipy import fftpack
N = 10
fftpack.fft(np.ones(N)) # Первая компонента равняется np.mean(x) * N
array([ 10.+0.j,   0.+0.j,   0.+0.j,   0.+0.j,   0.+0.j,   0.+0.j,
        0.-0.j,   0.-0.j,   0.-0.j,   0.-0.j])
```

Когда мы проверяем БПФ на быстро изменяющемся сигнале, видим появление высокочастотной компоненты:

```
z = np.ones(10)
z[::2] = -1
print(f'Применение БПФ к {z}')
fftpack.fft(z)
Применение БПФ к [-1.  1. -1.  1. -1.  1. -1.  1. -1.  1.]
array([ 0.+0.j,   0.+0.j,   0.+0.j,   0.+0.j,   0.+0.j,  -10.+0.j,
        0.-0.j,   0.-0.j,   0.-0.j,   0.-0.j])
```

Обратите внимание, БПФ возвращает комплексный спектр, являющийся в случае вещественных входных данных сопряженно-симметричным (т. е. симметричным в вещественной части и антисимметричным в мнимой части):

```
x = np.array([1, 5, 12, 7, 3, 0, 4, 3, 2, 8])
X = fftpack.fft(x)
np.set_printoptions(precision=2)
print("Вещественная часть: ", X.real)
print("Мнимая часть:", X.imag)
np.set_printoptions()
Вещественная часть: [ 45.   7.09 -12.24 -4.09 -7.76 -1. -7.76 -4.09 -12.24  7.09]
Мнимая часть:      [ 0.  -10.96 -1.62 12.03  6.88  0. -6.88 -12.03  1.62 10.96]
```

(Напоминаем: первая компонента равняется  $\text{np.mean}(x) * N$ .)

Функция `fftfreq` показывает, на какие конкретно частоты мы смотрим:

```
fftpack.fftfreq(10)
array([ 0. ,  0.1,  0.2,  0.3,  0.4, -0.5, -0.4, -0.3, -0.2, -0.1])
```

Этот результат показывает, что наша максимальная компонента произошла на частоте 0.5 цикла в расчете на выборку. Результат согласуется со входом, где цикл «минус один плюс один» повторялся каждую вторую выборку.

Иногда удобно рассматривать спектр, организованный немного по-другому: от высокоотрицательного до «от низко- до высокоположительного» (на дан-

ном этапе мы не будем детально разбирать понятие отрицательной частоты и просто скажем, что реальная синусоидальная волна порождается комбинацией положительных и отрицательных частот). Мы перетасовываем спектр, используя функцию `fftshift`.

### Дискретные преобразования Фурье



ДПФ преобразовывает последовательность из  $N$  равномерно расположенных вещественных или комплексных выборок  $x_0, x_1, \dots, x_{n-1}$  функции  $x(t)$  времени (либо другой переменной, в зависимости от приложения) в последовательность из  $N$  комплексных чисел  $X_k$  путем приведенного ниже суммирования:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{j2\pi kn}{N}}, \quad k = 0, 1, \dots, N-1.$$

Если числа  $X_k$  известны, то обратное ДПФ-преобразование с помощью приведенного ниже суммирования в точности восстанавливает выборочные значения  $x_n$ :

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{\frac{j2\pi kn}{N}}.$$

Учитывая, что  $e^{j\theta} = \cos\theta + j\sin\theta$ , последнее уравнение показывает, что ДПФ-преобразование разложило последовательность  $x_n$  в комплексный дискретный ряд Фурье с коэффициентами  $X_k$ . Сравним ДПФ с непрерывным комплексным рядом Фурье:

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{jn\omega_0 t}.$$



ДПФ представляет собой конечный ряд с  $N$  членами, определенными в равномерно расположенных дискретных экземплярах угла  $(\omega_0 t_n) = 2\pi(k/N)$  в интервале  $[0, 2\pi)$  – т. е. включая 0 и исключая  $2\pi$ . Это автоматически нормализует ДПФ так, что в прямом или обратном преобразовании время явным образом не появляется.

Если исходная функция  $x(t)$  будет ограничиваться по частоте менее половиной частоты дискретизации (так называемой *частотой Найквиста*), то интерполяция между выборочными значениями, производимая обратным ДПФ-преобразованием, обычно будет давать верную реконструкцию  $x(t)$ . Если  $x(t)$  как таковая не ограничивается, то обратное ДПФ-преобразование не может в целом путем интерполяции использоваться для реконструкции  $x(t)$ . Обратите внимание, данное ограничение не подразумевает отсутствия методов, позволяющих выполнять такую реконструкцию. Возьмем, например, методы восстановления сигнала с использованием знаний о его предыдущих разреженных или сжатых значениях (*compressed sensing*) или методы выборки сигналов с конечной интенсивностью обновления (*FRI-сигналов*)<sup>1</sup>.

Функция  $e^{(j2\pi k/N)n} = (e^{(j2\pi k/N)})^k = w^k$  принимает дискретные значения между 0 и на единичном круге в комплексной плоскости. Функция  $e^{(j2\pi kn/N)} = w^{kn}$  окружает источник  $n[(N-1)/N]$  раз, в результате генерируя гармонику фундаментальной синусоиды, для которой  $n = 1$ .

<sup>1</sup> Интенсивность обновления сигнала (*rate of innovation*, FRI) – это количество степеней свободы на единицу времени. См. <https://infoscience.epfl.ch/record/34271/files/VetterliMB01.pdf>. – Прим. перев.

То, чем мы определили ДПФ, приводит к нескольким тонкостям при  $n > (N/2)$  для  $N$  четных<sup>1</sup>. На рис. 4.6 график функции  $e^{j2\pi kn/N}$  построен для увеличивающихся значений  $k$  для случаев от  $n = 1$  до  $n = N - 1$  для  $N = 16$ . При увеличении  $k$  от  $k$  до  $k + 1$  угол увеличивается на  $2\pi n/N$ . При  $n = 1$  шаг равняется  $2\pi/N$ . При  $n = N - 1$  угол увеличивается на  $2\pi[(N - 1)/N] = 2\pi - 2\pi/N$ . Поскольку  $2\pi$  равно одному обороту вокруг круга, шаг будет равен  $-(2\pi/N)$ , т. е. в направлении отрицательной частоты. Компоненты до  $N/2$  представляют *положительные* компоненты частоты. Компоненты частоты, что выше  $N/2$ , и до  $N - 1$  представляют *отрицательные* частоты. Угловое приращение для компоненты  $N/2$  для  $N$  четного занимает половину круга для каждого приращения в  $k$ . Поэтому может интерпретироваться как положительная или же как отрицательная частота. Эта компонента ДПФ представляет частоту Найквиста (т. е. половину частоты дискретизации) и служит ориентиром при рассмотрении графика ДПФ.

БПФ, в свою очередь, просто является специальным и очень эффективным алгоритмом вычисления ДПФ. В отличие от прямого вычисления ДПФ, занимающего порядка  $N^2$  вычислений, алгоритм БПФ занимает порядка  $N \log N$  вычислений. БПФ стал ключевым в широком распространении ДПФ в приложениях, работающих в режиме реального времени, и в 2000 г. журналом IEEE Computing Science & Engineering он был включен в список лучших 10 алгоритмов XX века.

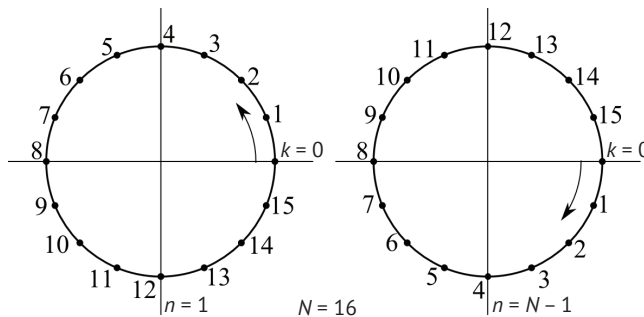


Рис. 4.6 ❖ Образцы единичного круга

Давайте исследуем частотные компоненты в зашумленном изображении (рис. 4.7). Обратите внимание: когда статическое изображение не имеет меняющейся во времени компоненты, его значения варьируются в пространстве. ДПФ применяется одинаково к любому случаю.

Сначала загрузим и покажем изображение:

```
from skimage import io
image = io.imread('images/moonLanding.png')
M, N = image.shape
f, ax = plt.subplots(figsize=(4.8, 4.8))
```

<sup>1</sup> Как упражнение мы оставляем читателям задачу изобразить ситуацию для  $N$ . Обратите внимание: в этой главе все примеры используют ДПФ-преобразования с четным порядком.

```
ax.imshow(image)
print((M, N), image.dtype)
(474, 630) uint8
```

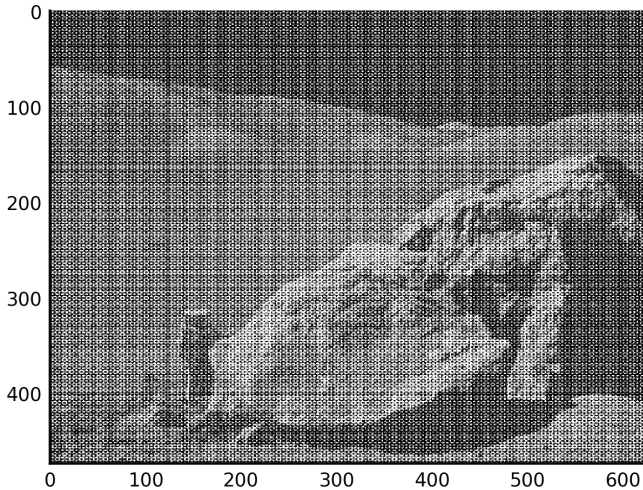


Рис. 4.7 ❖ Зашумленное изображение посадки на Луну

Не пытайтесь регулировать свой монитор! Показанное изображение настоящее, хотя искажено передающим или приемным оборудованием.

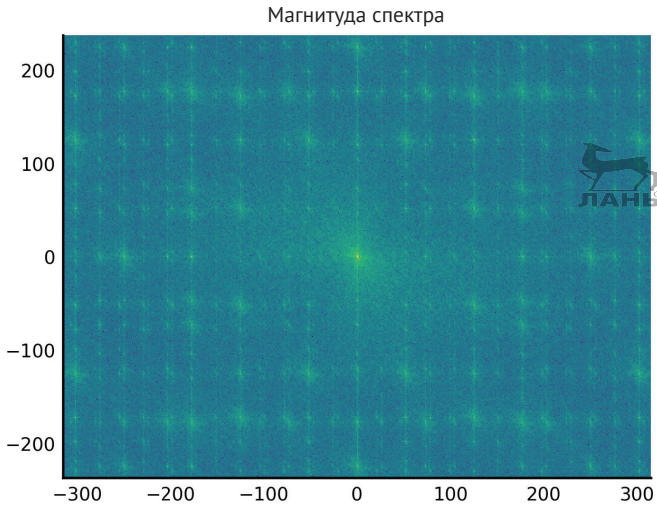
Для исследования спектра, поскольку изображение имеет более одной размерности, чтобы вычислить ДПФ, применим вместо функции `fft` функцию `fftn`. Двумерное БПФ-преобразование эквивалентно взятию одномерного БПФ в строках и затем в столбцах, или наоборот.

```
F = fftpack.fftn(image)
F_magnitude = np.abs(F)
F_magnitude = fftpack.fftshift(F_magnitude)
```

Снова, чтобы сжать диапазон значений перед выводом на экран, возьмем логарифм спектра:

```
f, ax = plt.subplots(figsize=(4.8, 4.8))
ax.imshow(np.log(1 + F_magnitude), cmap='viridis',
          extent=(-N // 2, N // 2, -M // 2, M // 2))
ax.set_title('Магниту́да спектра');
```





Обратите внимание на высокие значения вокруг источника (середины) спектра. Эти коэффициенты описывают низкие частоты или сглаживают части изображения, размывшие полотно фотографии. Более высокочастотные компоненты, распространенные по всему спектру, заполняют края и детализацию. Пики вокруг более высоких частот соответствуют периодическому шуму.

Из фотографии мы видим, что шум (артефакты измерения) имеет высокопериодический характер. Поэтому попробуем удалить его, обнулив соответствующие части спектра (рис. 4.8).

Изображение с этими подавленными пиками действительно **выглядит лучше!**

```
# Назначить блоку вокруг центра спектра значение ноль
K = 40
F_magnitude[M // 2 - K: M // 2 + K, N // 2 - K: N // 2 + K] = 0

# Найти все пики выше 98-го перцентиля
peaks = F_magnitude < np.percentile(F_magnitude, 98)

# Сдвинуть пики назад, чтобы выровнять с исходным спектром
peaks = fftpack.ifftshift(peaks)

# Сделать копию исходного (комплексного) спектра
F_dim = F.copy()

# Установить эти пиковые коэффициенты в ноль
F_dim = F_dim * peaks.astype(int)

# Выполнить обратную трансформацию Фурье, чтобы вернуться к изображению.
# Поскольку мы начали с вещественного изображения, то обратимся только
# к вещественной части результата.
image_filtered = np.real(fftpack.ifft2(F_dim))

f, (ax0, ax1) = plt.subplots(2, 1, figsize=(4.8, 7))
```



```
ax0.imshow(np.log10(1 + np.abs(F_dim)), cmap='viridis')
ax0.set_title('Спектр после подавления')

ax1.imshow(image_filtered)
ax1.set_title('Реконструированное изображение');
```

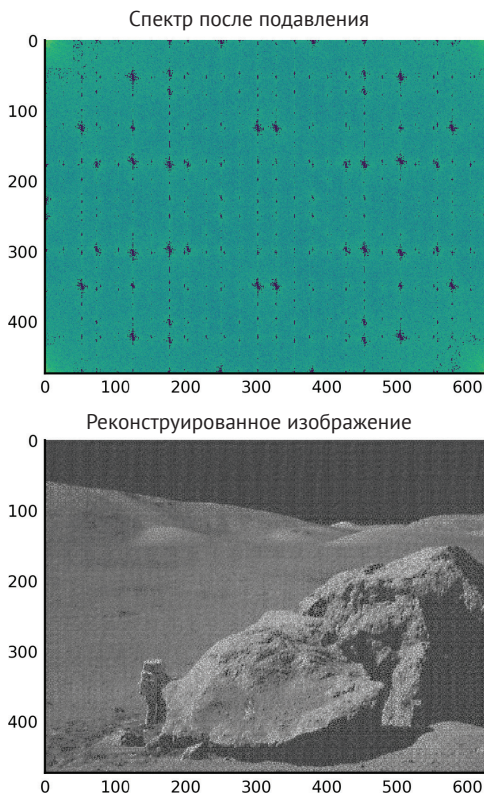


Рис. 4.8 ❖ Отфильтрованное изображение посадки на Луну и его спектр

## Оконное преобразование

Если исследовать преобразование Фурье прямоугольного импульса, то мы увидим значительные боковые лепестки в спектре:

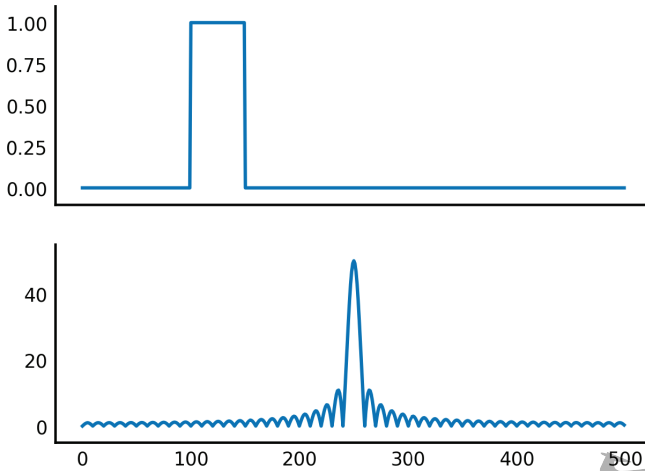
```
x = np.zeros(500)
x[100:150] = 1
X = fftpack.fft(x)

f, (ax0, ax1) = plt.subplots(2, 1, sharex=True)

ax0.plot(x)
ax0.set_ylim(-0.1, 1.1)
```

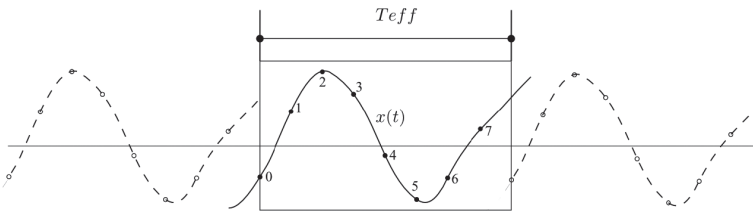


```
ax1.plot(fftpack.fftfshift(np.abs(X)))
ax1.set_ylim(-5, 55);
```



В теории для представления любого резкого перехода вам потребуется комбинация бесконечно многочисленных синусоид (частот). Коэффициенты, как правило, имеют ту же самую структуру боковых лепестков, как показано в случае с пульсом.

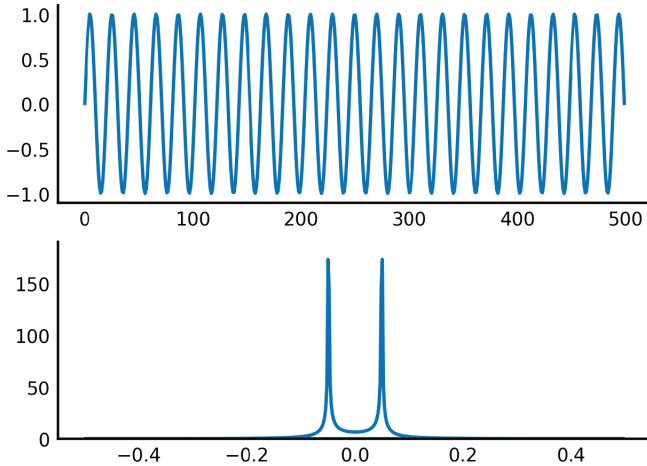
Немаловажно, что ДПФ исходит из периодического характера входного сигнала. Если сигнал не периодический, принимается допущение, что в конце сигнала он отскакивает к своему начальному значению. Рассмотрим функцию  $x(t)$ , показанную ниже:



Мы измеряем сигнал в течение короткого времени, помеченного как  $T_{eff}$ . Преобразование Фурье строится на основании, что  $x(8) = x(0)$ , и сигнал продолжается как прерывистая, а не сплошная линия. Это вносит большой скачок на краю с ожидаемой осцилляцией в спектре:

```
t = np.linspace(0, 1, 500)
x = np.sin(49 * np.pi * t)
X = fftpack.fft(x)
```

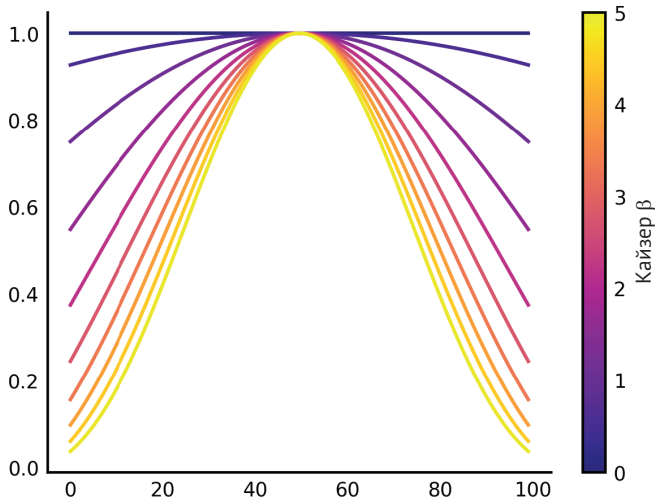
```
f, (ax0, ax1) = plt.subplots(2, 1)
ax0.plot(x)
ax0.set_ylim(-1.1, 1.1)
ax1.plot(fftpack.fftfreq(len(t)), np.abs(X))
ax1.set_ylim(0, 190);
```



Вместо ожидаемых двух линий пики распределены по спектру.

Этому эффекту можно противопоставить процесс, называемый оконным преобразованием. Исходная функция умножается на функцию окна, такую как окно Кайзера  $K(N, \beta)$ . Ниже мы его визуализируем для  $\beta$  в пределах от 0 до 100:

```
f, ax = plt.subplots()
N = 10
beta_max = 5
colormap = plt.cm.plasma
norm = plt.Normalize(vmin=0, vmax=beta_max)
lines = [
    ax.plot(np.kaiser(100, beta), color=colormap(norm(beta)))
    for beta in np.linspace(0, beta_max, N)
]
sm = plt.cm.ScalarMappable(cmap=colormap, norm=norm)
sm._A = []
plt.colorbar(sm).set_label(r'Кайзер  $\beta$ ');
```



Изменяя параметр  $\beta$ , мы можем изменять форму окна из прямоугольного ( $\beta = 0$ , т. е. оконное преобразование отсутствует) в окно, производящее сигналы, плавно увеличивающиеся от нуля и уменьшающиеся до нуля в конечных точках выборочного интервала. При этом будут производиться очень низкие боковые лепестки ( $\beta$ , как правило, между 5 и 10)<sup>1</sup>.

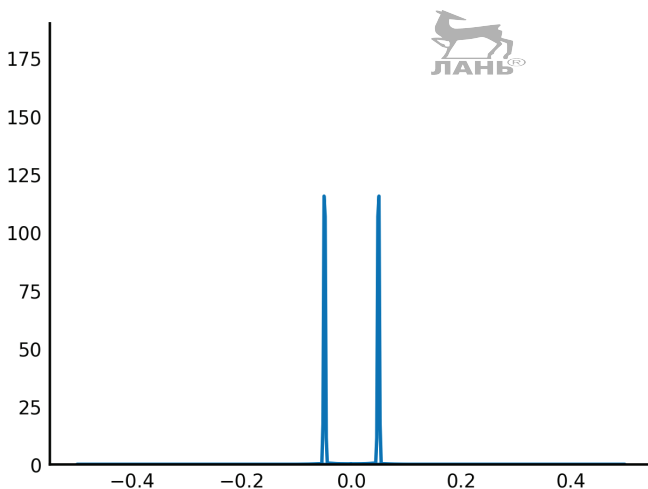
Применяя окно Кайзера, мы видим, что боковые лепестки были сильно сокращены за счет небольшого расширения в главном лепестке.

Эффект оконного преобразования нашего предыдущего примера примечателен:

```
win = np.kaiser(len(t), 5)
X_win = fftpack.fft(x * win)

plt.plot(fftpack.fftfreq(len(t)), np.abs(X_win))
plt.ylim(0, 190);
```

<sup>1</sup> Классические оконные функции включают функции Ханна, Хемминга и Блэкмана. Они различаются по их уровням боковых лепестков и расширением главного лепестка (в пространстве Фурье). Современной и гибкой оконной функцией, близкой к оптимальной для большинства приложений, является оконная функция Кайзера — хорошая аппроксимация оптимального вытянутого сферического окна, концентрирующего большинство энергии в главном лепестке. Путем корректировки параметра  $\beta$  мы можем выполнить тонкую настройку окна Кайзера, чтобы приспособить под конкретное приложение, как проиллюстрировано в основном тексте.



## ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ: АНАЛИЗ РАДАРНЫХ ДАННЫХ

В линейно модулированных РЛС непрерывного излучения с частотной модуляцией (Frequency-Modulated Continuous-Wave Radars), или FMCW-радары, алгоритм БПФ широко применяется для обработки сигналов. FMCW-радары обеспечивают примеры самых разнообразных применений БПФ. Мы воспользуемся фактическими данными FMCW-радара, чтобы продемонстрировать обнаружение цели (целеуказание).

В общем и целом FMCW-радар работает следующим образом (для получения более подробной информации см. раздел «Простая радарная система FMCW» и рис. 4.9):

1. Генерируется сигнал с изменяющейся частотой. Этот сигнал передается антенной, посылающей его от радара вовне. Когда сигнал попадает на объект, часть сигнала отражается обратно на радар. Радар этот сигнал и умножает на копию переданного сигнала. Далее полученный результат сэмплируется (превращается в упакованные в массив числа). Наша задача состоит в том, чтобы проинтерпретировать эти числа и сформировать содержательные результаты.
2. Предыдущий шаг умножения имеет большое значение. Вспомните тригонометрическое тождество из школьной программы:

$$\sin(xt)\sin(yt) = \frac{1}{2} \left[ \sin \left( (x - y)t + \frac{\pi}{2} \right) - \sin \left( (x + y)t + \frac{\pi}{2} \right) \right].$$

3. Так, если умножить принятый на переданный сигнал, получим в спектре две частотные компоненты: разностную между частотами переданного и принятого сигналов и суммарную, состоящую из переданного и принятого сигналов.

4. Нас в особенности интересует первая, разностная, показывающая время, которое потребуется сигналу для отражения от цели на радар. Мы с помощью фильтра нижних частот (фильтр, отсекающий высокочастотную компоненту) отбрасываем другую, суммирующую компоненту.

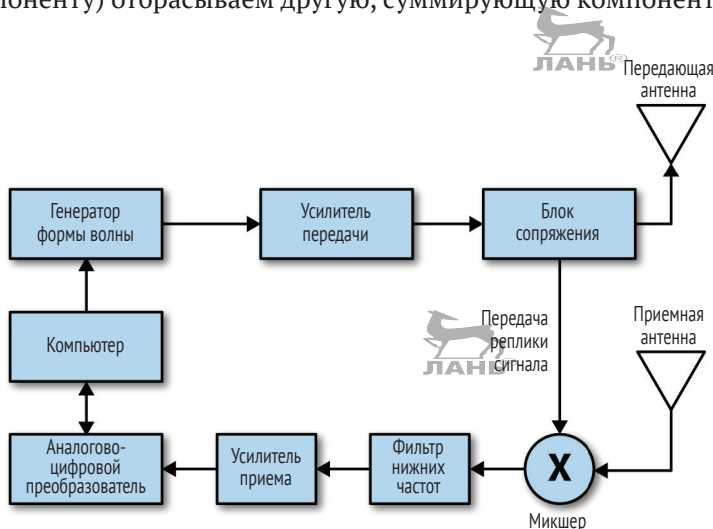


Рис. 4.9 ❖ Блок-схема простой радарной системы FMCW

### Простая радарная система FMCW

Блок-схема простой радарной системы FMCW, использующей отдельно передающую и приемную антенны, показана выше. Радар состоит из генератора волны, производящей синусоидальный сигнал. Частота этого сигнала линейно варьируется вокруг требуемой частоты передачи. Сгенерированный сигнал усиливается до нужного уровня услителем передачи и через цепь блока сопряжения направляется в передающую антенну. Обратите внимание, в блоке сопряжения мы отделяем копию сигнала, идущего в передающую антенну. Передающая антенна излучает полученный сигнал в направлении цели, подлежащей обнаружению. Сигнал излучается в виде узконаправленной электромагнитной волны. Заметьте, любая цель в той или иной мере частично отражает электромагнитную волну, которой он облучается. После встречи электромагнитной волны с объектом часть облучающей цель энергии отразится и вернется назад в виде вторичной электромагнитной волны. После встречи отраженной от цели электромагнитной волны с приемной антенной эта волна преобразуется в электрический сигнал и направится в микшер. Микшер после умножения отраженного сигнала на копию переданного сигнала произведет синусоидальный сигнал с частотой, равной разнице частот между переданным и приемным сигналами. Этот разностный сигнал поступит на фильтр нижних частот. Фильтр нижних частот отсекает частоты, в которых мы не заинтересованы. Приемный усилитель усилит сигнал до амплитуды, необходимой для аналого-цифрового преобразователя (ADC). ADC, в свою очередь, передаст данные в компьютер.

Подводя итоги, мы должны отметить, что:

- данные, достигающие компьютера, состоят из  $N$  выборок (из умноженного и отфильтрованного сигнала) на частоте дискретизации  $f_s$ ;
- амплитуда возвращенного сигнала варьируется в зависимости от силы отражения (т. е. является свойством целевого объекта и расстояния между целью и радаром);
- измеренная частота является признаком расстояния целевого объекта от радара.

Чтобы начать анализ радарных (радиолокационных) данных, сгенерируем несколько синтетических сигналов, после чего сосредоточим наше внимание на сигнале, исходящем из фактического радара.

Напомним, радар увеличивает свою частоту по мере передачи в размере  $S$  Гц/с. По прошествии определенного временного промежутка  $t$  частота будет на  $tS$  выше (рис. 4.10). В тот же отрезок времени радарный сигнал прошел расстояние  $d = t/v$  метров, где  $v$  – это скорость переданной по воздуху электромагнитной волны (примерно такая же, что и скорость света,  $3 \times 10^8$  м/с).

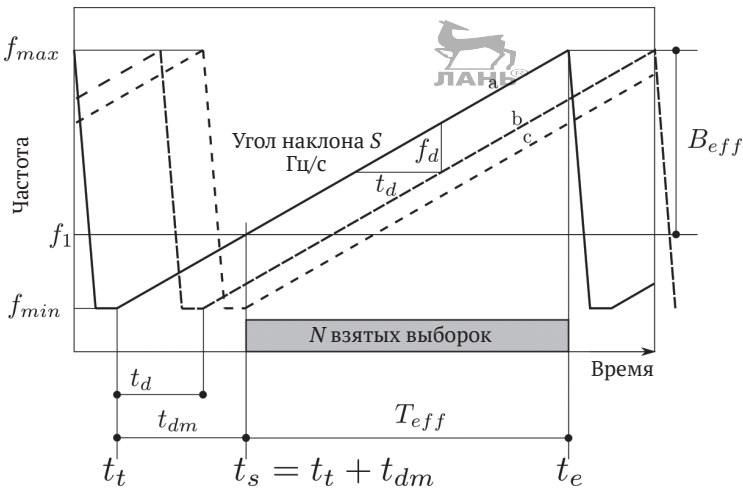


Рис. 4.10 ❖ Частотные связи в радаре FMCW с линейной частотной модуляцией

Объединяя приведенные выше наблюдения, мы можем вычислить временной промежуток, требующийся сигналу, чтобы пройти, отразиться и вернуться от цели, находящейся на расстоянии  $R$ :

$$t_R = \frac{2R}{v}.$$

$p_i = \text{nr.p}_i$

# Параметры радара

```

fs = 78125          # Частота дискретизации в герцах, т. е. мы извлекаем
                  # 78125 выборок в секунду

ts = 1 / fs        # Период дискретизации, т. е. одна выборка
                  # извлекается ts секунд

Teff = 2048.0 * ts # Общее время дискретизации для 2048 выборок
                  # (т. н. эффективная длительность развертки) в секундах.

Beff = 100e6       # Диапазон частоты проходящего сигнала во время взятия радаром выборок,
                  # т. н. «эффективная полоса частот» (задан в герцах)

S = Beff / Teff    # Скорость качания частоты в Гц/с

# Описания целей. Это выдуманные цели, которые представляют собой
# видимые радаром объекты с заданной дальностью и размером.

R = np.array([100, 137, 154, 159, 180]) # Дальности (в метрах)
M = np.array([0.33, 0.2, 0.9, 0.02, 0.1]) # Размер цели
P = np.array([0, pi / 2, pi / 3, pi / 5, pi / 6]) # Случайно отобранные фазовые сдвиги

t = np.arange(2048) * ts # Периоды дискретизации

fd = 2 * S * R / 3E8      # Разности частот для этих целей

# Сгенерировать пять целей
signals = np.cos(2 * pi * fd * t[:, np.newaxis] + P)

# Сохранить сигнал, связанный с первой целью, в качестве примера
# для обследования в дальнейшем
v_single = signals[:, 0]

# Взвесить сигналы в соответствии с размером и суммы целей, чтобы
# сгенерировать комбинированный сигнал, видимый радаром.
v_sim = np.sum(M * signals, axis=1)

## Приведенный выше программный код эквивалентен следующему:
#
# v0 = np.cos(2 * pi * fd[0] * t)
# v1 = np.cos(2 * pi * fd[1] * t + pi / 2)
# v2 = np.cos(2 * pi * fd[2] * t + pi / 3)
# v3 = np.cos(2 * pi * fd[3] * t + pi / 5)
# v4 = np.cos(2 * pi * fd[4] * t + pi / 6)
#
## Смешать их вместе
# v_single = v0
# v_sim = (0.33 * v0) + (0.2 * v1) + (0.9 * v2) + (0.02 * v3) + (0.1 * v4)

```

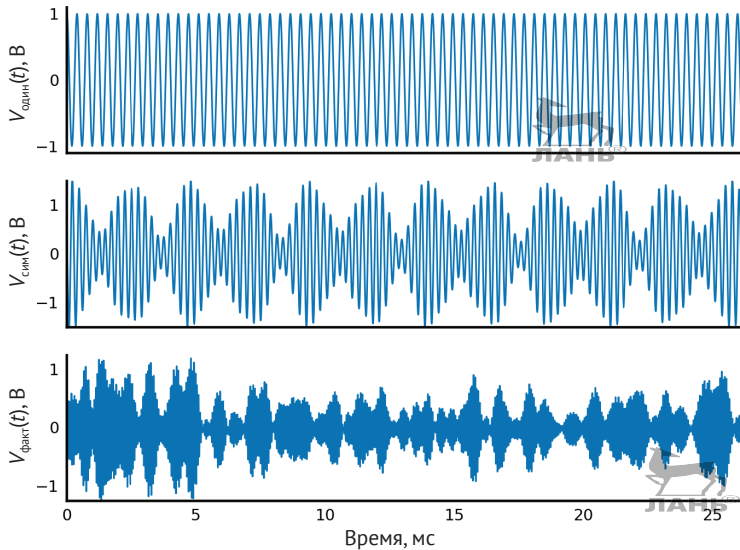


Здесь мы сгенерировали синтетический сигнал, полученный при рассмотрении одиночной цели (см. рис. 4.11). Подсчитав количество циклов, обнаруженных в заданном периоде времени, мы можем вычислить частоту сигнала и, следовательно, расстояние до цели.

Вместе с тем реальный радар редко будет получать только одно отраженное эхо. Симулируемый сигнал показывает, как выглядит радарный сигнал с пятью целями с различными дальностями (включая две, находящиеся рядом друг с другом на расстоянии 154 и 159 м); показывает выходной сигнал, полученный фактическим радаром. Когда мы складываем многочисленные эхо



вместе, до тех пор, пока мы не проанализируем его более тщательно через линзу ДПФ, результат не будет иметь какого-то интуитивно понятного смысла (рис. 4.11).



**Рис. 4.11** ❖ Выходные сигналы приемника:  
(а) одиночная симулированная цель, (б) пять симулированных целей  
и (с) фактические радарные данные

Реальные радарные данные читаются в формате NumPy из файла `.npz`. (Формат NumPy представляет собой легковесный, кросс-платформенный и версионно-совместимый формат хранения.) Эти файлы могут быть сохранены функциями `np.savez_compressed` или `np.savez`. Обратите внимание, что подмодуль SciPy `io` также легко может читать другие форматы, такие как файлы NetCDF и MATLAB.

```
data = np.load('data/radar_scan_0.npz')
# Загрузить переменную 'scan' из 'radar_scan_0.npz'
scan = data['scan']

# Набор данных содержит многочисленные замеры, при этом каждый взят,
# когда радар был направлен в разные стороны. Здесь мы берем один
# такой замер с заданным азимутом (позицией лево-право) и высотой
# (позицией верх-низ). Замер имеет форму (2048,).
v_actual = scan['samples'][5, 14, :]

# Амплитуда сигнала варьируется от -2.5V до +2.5V. 14-разрядный
# аналогово-цифровой преобразователь в радаре выдает целые числа
# между -8192 и 8192. Мы преобразовываем назад в напряжение путем
# умножения $(2.5 / 8192)$.
v_actual = v_actual * (2.5 / 8192)
```

Поскольку файлы `.prz` могут хранить многочисленные переменные, нам необходимо отобразить одну, которая нам нужна: `data['scan']`. В результате будет возвращен *структурированный массив NumPy* со следующими полями:

`time`

Беззнаковое 64-разрядное (8-байтовое) целое число (np.uint64).



`size`

Беззнаковое 32-разрядное (4-байтовое) целое число (np.uint32).

`position`

`az`

32-разрядное вещественное (np.float32).

`el`

32-разрядное вещественное (np.float32).

`region_type`

Беззнаковое 8-разрядное (1-байтовое) целое число (np.uint8).

`region_ID`

Беззнаковое 16-разрядное (2-байтовое) целое число (np.uint16).

`gain`

Беззнаковое 8-разрядное (1-байтовое) целое число (np.uint8).



`samples`

2048 беззнаковых 16-разрядных (2-байтовых) целых чисел (np.uint16).

Хотя утверждение, что массивы NumPy гомогенные, верно (т. е. тип всех элементов внутри него одинаков), это не означает, что это не составные элементы, как было здесь.

Доступ к отдельному полю осуществляется на основе словарного синтаксиса:

```
azimuths = scan['position']['az'] # Получить все замеры азимута
```

Резюмируем: все, что мы видели до сих пор, – показанные замеры  $v_{\text{сим}}$  (и  $v_{\text{факт}}$ ) являются суммой синусоидальных сигналов, отраженных каждым из нескольких объектов. Нам нужно определить каждую составляющую компоненту этих составных радарных сигналов. БПФ – это как раз тот инструмент, который делает это за нас.

## Свойства сигнала в частотной области

Прежде всего мы выполняем БПФ-преобразование наших трех сигналов (синтетического с единственной целью, синтетического с мультицелью и реального) и затем показываем компоненты с положительной частотой (т. е. компоненты от 0 до  $N/2$ ; см. рис. 4.12). В терминологии радаров они называются *трассировками дальности* (range traces, или сопровождениями по дальности).

```
fig, axes = plt.subplots(3, 1, sharex=True, figsize=(4.8, 2.4))
```

```
# Взять БПФ наших сигналов. Обратите внимание на правило именования
# БПФ-преобразований с заглавной буквы.
```

```
V_single = np.fft.fft(v_single)
```

```
V_sim = np.fft.fft(v_sim)
```

```
V_actual = np.fft.fft(v_actual)
```

```
N = len(V_single)
```

```
with plt.style.context('style/thinner.mplstyle'):
```

```
axes[0].plot(np.abs(V_single[:N // 2]))
```

```
axes[0].set_ylabel("$|V_{\mathrm{один}}|$")
```

```
axes[0].set_xlim(0, N // 2)
```

```
axes[0].set_ylim(0, 1100)
```

```
axes[1].plot(np.abs(V_sim[:N // 2]))
```

```
axes[1].set_ylabel("$|V_{\mathrm{сим}}|$")
```

```
axes[1].set_ylim(0, 1000)
```

```
axes[2].plot(np.abs(V_actual[:N // 2]))
```

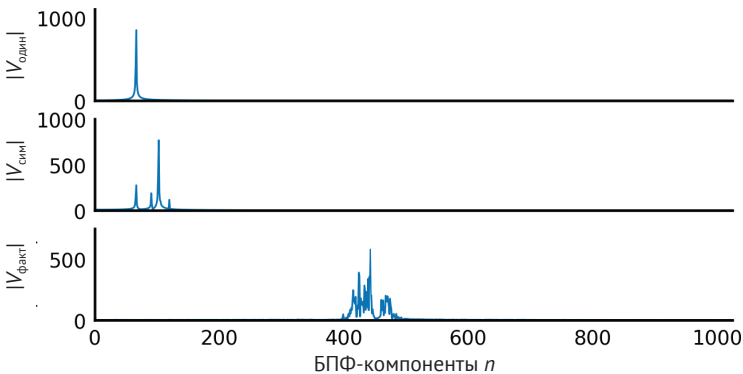
```
axes[2].set_ylim(0, 750)
```

```
axes[2].set_ylabel("$|V_{\mathrm{факт}}|$")
```

```
axes[2].set_xlabel("БПФ-компоненты $n$")
```

```
for ax in axes:
```

```
ax.grid(False)
```



**Рис. 4.12** ❖ Трассировки дальности для: (а) одиночной симулируемой цели, (б) многочисленных симулируемых целей и (с) реальных целей

И внезапно информация приобретает смысл!

График для  $|V_0|$  ясно показывает цель в компоненте 67 и для  $|V_{\text{сим}}|$  показывает цели, созданные сигналом, которые не поддавались интерпретации во временной области. Реальный радарный сигнал  $|V_{\text{факт}}|$  показывает большое количество целей между компонентами 400 и 500 с большим пиком в компоненте 443. Это оказывается эхо-сигналом из радара, просвечивающего откос открытой горнорудной разработки.

Чтобы получить полезную информацию из графика, мы должны определить дальность! И снова мы используем формулу:

$$R_n = \frac{nv}{2B_{\text{eff}}}.$$



В терминологии радаров каждая компонента ДПФ называется *интервалом дальности* (range bin, или элементом разрешения по дальности).

Это уравнение также определяет разрешающую способность радара по дальности: цели будут различаться, только если они отстоят более чем на два интервала дальности. Например:

$$\Delta R = \frac{1}{B_{\text{eff}}}.$$

Это фундаментальное свойство всех типов радаров.

Мы получили удовлетворительный результат – но динамический диапазон столь большой, что мы очень легко могли пропустить некоторые пики. Давайте, как и ранее, возьмем логарифм спектрограммы:

```
c = 3e8 # Приблизительно скорость света и
        # электромагнитных волн в воздухе
```

```
fig, (ax0, ax1, ax2) = plt.subplots(3, 1)
```



```
def dB(y):
```

```
    "Вычислить логарифмическое соотношение y / max(y) в децибелах."
```

```
    y = np.abs(y)
```

```
    y /= y.max()
```

```
    return 20 * np.log10(y)
```

```
def log_plot_normalized(x, y, ylabel, ax):
```

```
    ax.plot(x, dB(y))
```

```
    ax.set_ylabel(ylabel)
```

```
    ax.grid(False)
```

```
rng = np.arange(N // 2) * c / 2 / Beff
```

```
with plt.style.context('style/thinner.mplstyle'):
```

```
    log_plot_normalized(rng, V_single[:N // 2], "$|V_0|$ [дБ]", ax0)
```

```
    log_plot_normalized(rng, V_sim[:N // 2], "$|V_5|$ [дБ]", ax1)
```

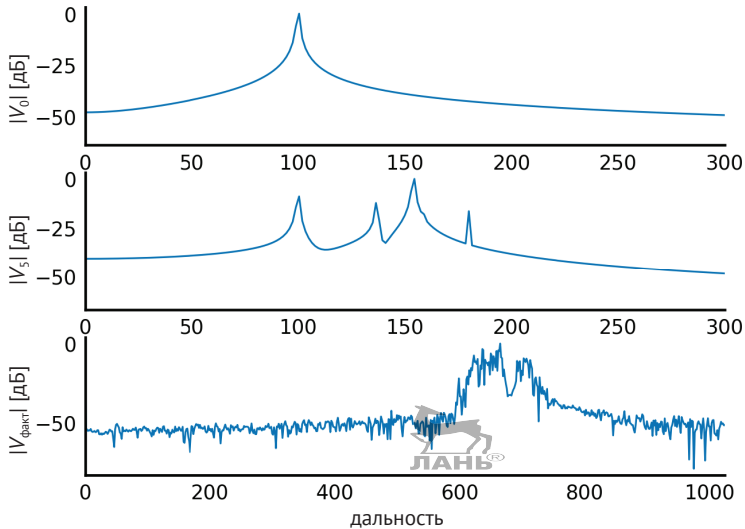
```
    log_plot_normalized(rng, V_actual[:N // 2], "$|V_{\mathbf{факт}}|$ [дБ]", ax2)
```

```
ax0.set_xlim(0, 300) # Для этих графиков изменить границы x, чтобы
```

```
ax1.set_xlim(0, 300) # можно было лучше увидеть форму пиков.
```

```
ax2.set_xlim(0, len(V_actual) // 2)
```

```
ax2.set_xlabel('дальность')
```



На этих графиках наблюдаемый динамический диапазон стал намного лучше. Например, в реальном радарном сигнале *уровень собственных шумов* радара стал видимым (т. е. уровень, где электронный шум в системе начинает ограничивать способность радара обнаруживать цель).

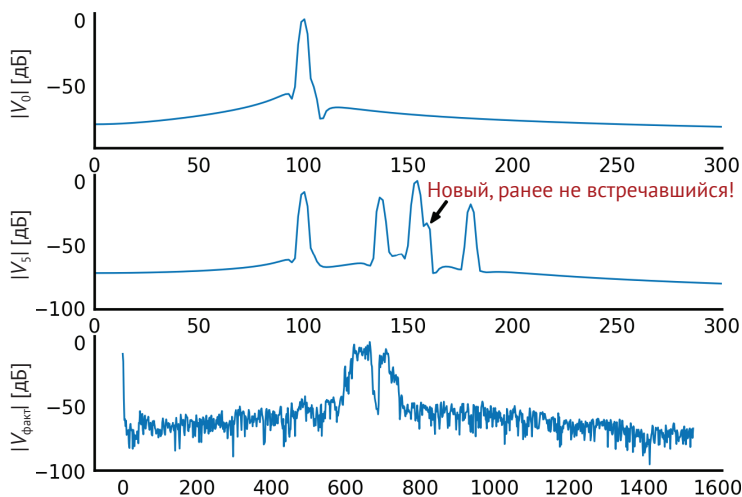
## Оконное преобразование на практике

Мы почти у цели. Однако в спектре симулируемого сигнала мы по-прежнему не можем различить пики на 154 и 159 метрах. Кто знает, что бы мы пропустили в реальном сигнале! Чтобы заострить пики, мы вернемся к нашему комплекту инструментов и применим *оконное преобразование*.

Вот сигналы, которые до настоящего времени использовались в этом примере, преобразованные окном Кайзера при  $\beta = 6.1$ :

```
f, axes = plt.subplots(3, 1, sharex=True, figsize=(4.8, 2.8))
t_ms = t * 1000 # Периоды дискретизации в миллисекундах
w = np.kaiser(N, 6.1) # Окно Кайзера при beta = 6.1
for n, (signal, label) in enumerate([(v_single, r'$v_0 [V]$',
                                     (v_sim, r'$v_5 [V]$',
                                      (v_actual, r'$v_{\mathrm{факт}} [V]$',
                                     ])]):
    with plt.style.context('style/thinner.mplstyle'):
        axes[n].plot(t_ms, w * signal)
```





Сравните их с более ранними трассировками дальности. Имеется сильнейшее понижение в уровне боковых лепестков. Но за счет изменения пиков по форме пики расширились и стали более тупыми. Следовательно, снизилась разрешающая способность радара, т. е. способность радара различать между двумя близко расположенными целями. Выбор окна является компромиссом между уровнем боковых лепестков и разрешающей способностью. Несмотря на это, обращаясь к трассировке для  $V_{\text{сим}}$  оконное преобразование существенно увеличило нашу способность отличать небольшую цель от ее крупного соседа.

В трассировке дальности реальных радарных данных оконное преобразование также уменьшило боковые лепестки. Это особенно хорошо видно по глубине зазубрины между двумя группами целей.

## Радарные изображения

Сведения о том, как проводить анализ одиночной трассировки, позволяют расширить эту методологию до рассмотрения радарных изображений.

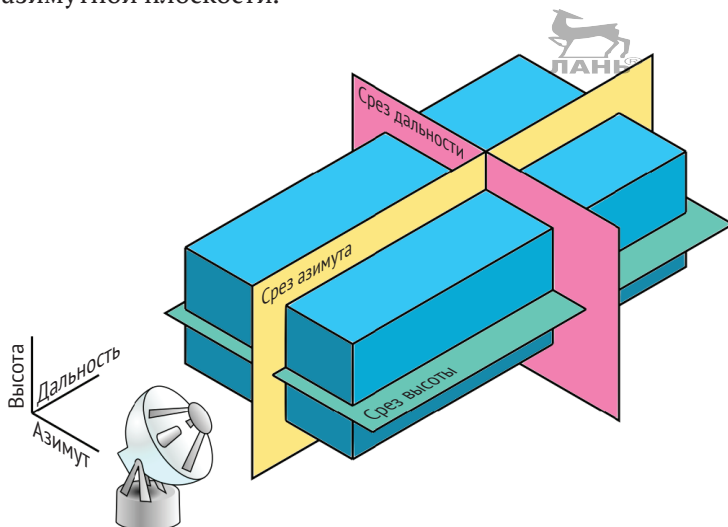
Данные производятся радаром с параболической отражающей антенной. Она производит остронаправленный круговой карандашный луч с двухградусным углом расходимости между точками половинной мощности. При направленности с нормальным падением на поверхность радар будет просвечивать пятно диаметром порядка 2 м на расстоянии 60 м. Вне этого пятна мощность довольно быстро понижается, однако сильные эхо-сигналы снаружи пятна тем не менее будут по-прежнему видимы.

Изменяя азимут карандашного луча (позицию лево-право) и высоту (позицию верх-низ), мы можем распространить его на интересующую целевую

область. Поймав отражения, получим возможность вычислить расстояние до отражателя (объекта, в который попал радарный сигнал). Текущий азимут и высота карандашного луча определяют положение отражателя в трехмерном измерении.

Склон скалы состоит из тысяч отражателей. Интервал дальности можно представить в виде большой сферы с радаром в центре, пересекающей склон вдоль рваной линии. Рассеиватели на данной линии будут производить отражения в этом интервале дальности. Длина волны радара (расстояние, которое передаваемая волна проходит за одну секунду осцилляции) составляет приблизительно 30 мм. Отражения от рассеивателей, отделенных нечетными кратными четверти длины волны, порядка 7.5 мм будут создавать деструктивную интерференцию в радаре, в то время как от рассеивателей, отделенных четными кратными половины длины волны, будут создавать конструктивную интерференцию. Чтобы произвести различимые пятна сильных отражений, отражения объединяются. Этот конкретный радар перемещает свою антенну так, чтобы сканировать малые области, состоящие из интервалов с 20-градусным азимутом и 30-градусной высотой, сканируемых с шагом по 0.5 градуса.

Теперь мы создадим несколько контурных графиков, результирующих радарные данные. Обратитесь к рис. 4.13, чтобы увидеть, каким образом берутся различные срезы. Первый срез с фиксированной дальностью показывает мощность эхо-сигналов относительно высоты и азимута. Еще два среза при фиксированной высоте и азимуте, соответственно, показывают склон (см. рис. 4.13 и 4.14). Ступенчатая конструкция откоса открытой горнорудной разработки видима в азимутной плоскости.



**Рис. 4.13** ❖ Диаграмма, показывающая срезы азимута, высоты и дальности через объем данных





```

data = np.load('data/radar_scan_1.npz')
scan = data['scan']

# Амплитуда сигнала варьируется от -2.5V до +2.5V. 14-разрядный
# аналогово-цифровой преобразователь в радаре выдает целые числа
# между -8192 и 8192. Мы преобразуем назад в напряжение путем
# умножения $(2.5 / 8192)$.

v = scan['samples'] * 2.5 / 8192
win = np.hanning(N + 1)[: -1]

# Взять БПФ для каждого замера
V = np.fft.fft(v * win, axis=2)[: -1, :, :N // 2]

contours = np.arange(-40, 1, 2)

# игнорировать предупреждения matplotlib
import warnings
warnings.filterwarnings('ignore', '.*Axes.*compatible.*tight_layout.*')

f, axes = plt.subplots(2, 2, figsize=(4.8, 4.8), tight_layout=True)
labels = ('Дальность', 'Азимут', 'Высота')

def plot_slice(ax, radar_slice, title, xlabel, ylabel):
    ax.contourf(dB(radar_slice), contours, cmap='magma_r')
    ax.set_title(title)
    ax.set_xlabel(xlabel)
    ax.set_ylabel(ylabel)
    ax.set_facecolor(plt.cm.magma_r(-40))

with plt.style.context('style/thinner.mplstyle'):
    plot_slice(axes[0, 0], V[:, :, 250], 'Дальность=250', 'Азимут', 'Высота')
    plot_slice(axes[0, 1], V[:, 3, :], 'Азимут=3', 'Дальность', 'Высота')
    plot_slice(axes[1, 0], V[6, :, :].T, 'Высота=6', 'Азимут', 'Дальность')
    axes[1, 1].axis('off')

```



### Трехмерная визуализация

Мы также можем визуализировать объем в трех измерениях (рис. 4.15).

Сначала вычисляем `argmax` (индекс максимального значения) в направлении дальности. Это должно дать представление о дальности, с которой радарный луч попал на скальный склон. Каждый индекс `argmax` преобразуется в трехмерную координату (дальности-азимута-высоты):

```

r = np.argmax(V, axis=2)

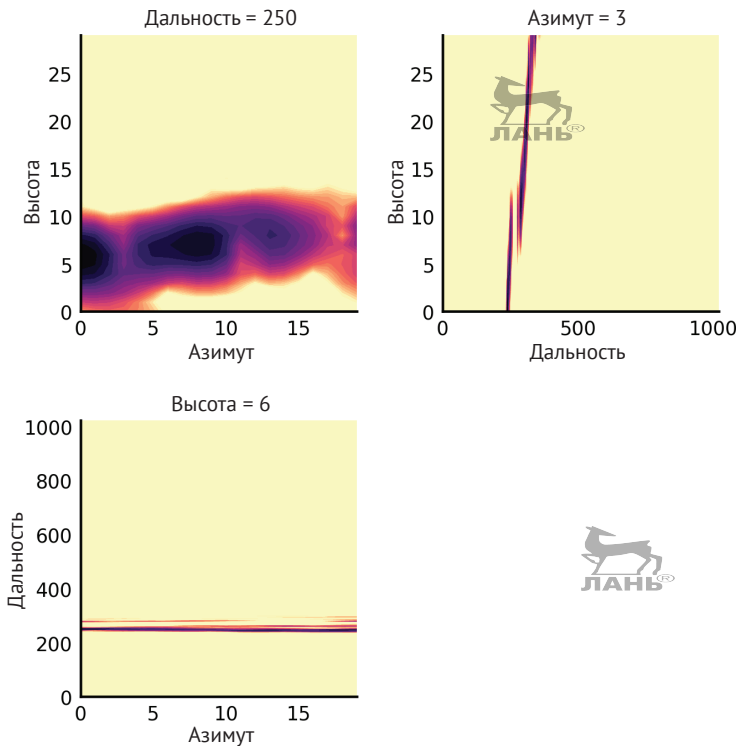
el, az = np.meshgrid(*[np.arange(s) for s in r.shape], indexing='ij')
axis_labels = ['Высота', 'Азимут', 'Дальность']
coords = np.column_stack((el.flat, az.flat, r.flat))

```

Беря эти координаты, мы проецируем их на плоскость азимута-высоты (пропуская координату дальности) и выполняем тесселяцию Делоне. Тесселяция возвращает набор индексов на наши координаты, которые определяют треугольники (или симплексы). Хотя треугольники, строго говоря, определены

на спроецированных координатах, мы используем наши исходные координаты для реконструкции, для чего добавляем назад компонент дальности:

```
from scipy import spatial
d = spatial.Delaunay(coords[:, :2])
simplexes = coords[d.vertices]
```



**Рис. 4.14** ❖ Контурные графики трассировок дальностей вдоль различных осей (см. рис. 4.13)

Для демонстрационных целей мы меняем местами ось диапазона, делая его первым:

```
coords = np.roll(coords, shift=-1, axis=1)
axis_labels = np.roll(axis_labels, shift=-1)
```

Теперь функция Matplotlib `trisurf` может использоваться для визуализации результата:

```
# Эта строка импорта инициализирует трехмерный механизм Matplotlib
from mpl_toolkits.mplot3d import Axes3D
```



```
# Задать трехмерную ось
f, ax = plt.subplots(1, 1, figsize=(4.8, 4.8),
                    subplot_kw=dict(projection='3d'))

with plt.style.context('style/thinner.mplstyle'):
    ax.plot_trisurf(*coords.T, triangles=d.vertices, cmap='magma_r')

    ax.set_xlabel(axis_labels[0])
    ax.set_ylabel(axis_labels[1])
    ax.set_zlabel(axis_labels[2], labelpad=-3)
    ax.set_xticks([0, 5, 10, 15])

# Скорректировать позицию камеры, чтобы она совпала с нашей диаграммой выше
ax.view_init(azim=-50);
```

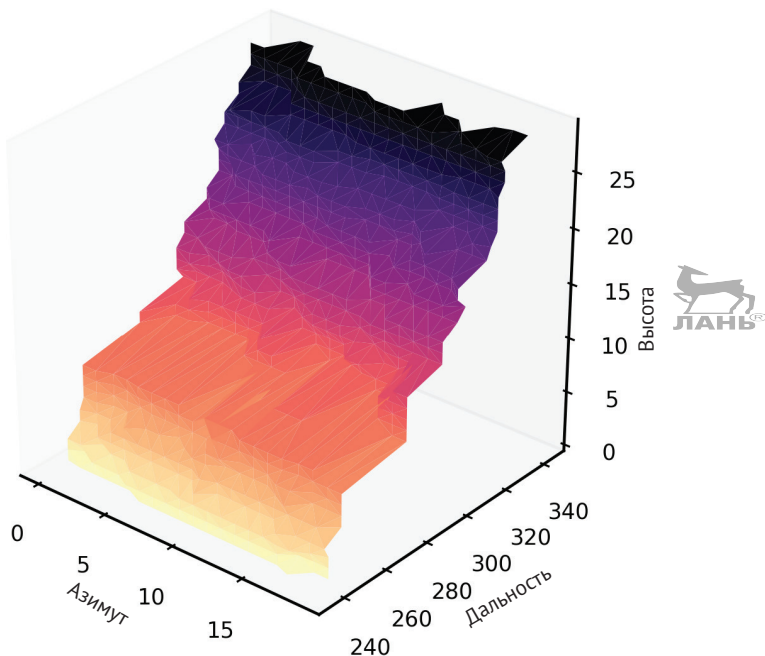


Рис. 4.15 ❖ Трехмерная визуализация положения предполагаемого скального склона

## Дополнительные применения БПФ

Предыдущие примеры показывают только одно из применений БПФ в радаре. Кроме него, существует целый ряд других, таких как (доплеровское) измерение движения и распознавание цели. БПФ вездесущ и встречается везде, от МРТ до статистики. Имея на руках базовые методы, которые кратко были обрисованы в этой главе, вы теперь в достаточной мере оснащены, чтобы начать его использовать!

## Дополнительные материалы для чтения

По преобразованию Фурье:

- *Популис А.* Интеграл Фурье и его применения (*Athanasios Papoulis. The Fourier Integral and Its Applications.* New York: McGraw-Hill, 1960);
- *Брейсуэлл Р. А.* Преобразование Фурье и его применения (*Ronald A. Bracewell. The Fourier Transform and Its Applications.* New York: McGraw-Hill, 1986).

По обработке радарных сигналов:

- *Ричардс М. А., Шир Дж. А., Холм У. А.* Принципы современного радара: Основные принципы (*Mark A. Richards, James A. Scheer, and William A. Holm, eds. Principles of Modern Radar: Basic Principles,* Raleigh. NC: SciTech, 2010);
- *Ричардс М. А.* Основные принципы обработки радарных сигналов (*Mark A. Richards. Fundamentals of Radar Signal Processing.* New York: McGrawHill, 2014).

## Задача: свертывание изображения

БПФ часто используется для ускорения свертывания изображения (свертывание, или конволюция, связано с применением скользящего фильтра). Сверните изображение при помощи `np.ones((5, 5))`, используя а) функцию `NumPy np.convolve` и б) функцию `np.fft.fft2`. Подтвердите, что результаты идентичны.

Подсказки:

- свертывание  $x$  и  $y$  эквивалентно `ifft2(x * y)`, где  $x$  и  $y$  – это БПФ-преобразования соответственно  $x$  и  $y$ ;
- для того чтобы умножить  $X$  и  $Y$ , они должны иметь одинаковый размер. Примените функцию `np.pad`, чтобы расширить  $x$  и  $y$  нулями (вправо и вниз) *перед* взятием их БПФ-преобразования;
- вы можете увидеть некоторые краевые эффекты. Их можно удалить, увеличив размер дополнения так, чтобы  $x$  и  $y$  имели размерности `shape(x) + shape(y) - 1`.

Обратите внимание: решение задачи «Свертывание изображений» вы найдете в конце книги.

---

## Таблицы сопряженности на основе разреженных координатных матриц



Мне нравится разреженность. Что-то есть в этой минималистской атмосфере, которая заставляет нечто оказывать непосредственное влияние и делает его уникальным. Я бы, наверное, всегда работал с этой формулой. Я просто не знаю, как.

– Бритт Дэниел, солист группы *Spoon*

Многие матрицы, используемые в реальных условиях, являются разреженными. Имеется в виду, что большинство их значений равняется нулю.

При использовании массивов NumPy для управления разреженными матрицами много времени и энергии тратится впустую на умножение огромного количества значений на 0. Вместо этого мы можем воспользоваться модулем SciPy `sparse`, позволяющим, исследуя только ненулевые значения, эффективно вычислять такие матрицы. В дополнение, что этот модуль способствует разрешению таких «канонических» разреженно-матричных проблем, модуль `sparse` может использоваться для задач, связанных очевидным образом с разреженными матрицами.

Сравнение сегментаций изображения является одной из таких задач. (Обратитесь к главе 3 относительно определения понятия «сегментация».)

Пример программного кода, мотивирующий эту главу, дважды использует разреженные матрицы. Сначала для вычисления *матрицы сопряженности* мы используем программный код, предложенный Андреасом Мюллером (Andreas Mueller). Эта матрица подсчитывает соответствие меток между двумя сегментациями. Затем, по предложению Хайме Фернандеса дель Рио и Уоррена Векеккера (Jaime Fernández del Río и Warren Weckesser), мы для вычисления *изменчивости информации* используем матрицу сопряженности, измеряющей разницу между сегментациями.

```
def variation_of_information(x, y):
    # Вычислить матрицу сопряженности, т. н. матрицу совместной вероятности
    n = x.size
    Pxy = sparse.coo_matrix((np.full(n, 1/n), (x.ravel(), y.ravel())),
                            dtype=float).tocsr()

    # Вычислить маргинальные вероятности, преобразовав в одномерный массив
    px = np.ravel(Pxy.sum(axis=1))
    py = np.ravel(Pxy.sum(axis=0))

    # Использовать линейную алгебру разреженных матриц, чтобы сначала
    # вычислить изменчивость информации (VI),
    # вычислить обратные диагональные матрицы
    Px_inv = sparse.diags(invert_nonzero(px))
    Py_inv = sparse.diags(invert_nonzero(py))

    # затем вычислить энтропии
    hygx = px @ xlog1x(Px_inv @ Pxy).sum(axis=1)
    hxgy = xlog1x(Pxy @ Py_inv).sum(axis=0) @ py

    # Вернуть их сумму
    return float(hygx + hxgy)
```

### Профессиональный совет по Python 3.5!

Символы @ в приведенном выше абзаце представляют собой оператор *умножения матриц* и были введены в Python 3.5 в 2015 г. Для научных программистов, чтобы использовать Python 3, это один из самых востребованных аргументов: он позволяет программировать линейно-алгебраические алгоритмы, используя программный код, который остается очень близким к исходному математическому аппарату. Сравните строку из приведенного выше программного кода:

```
hygx = px @ xlog1x(Px_inv @ Pxy).sum(axis=1)
```

с эквивалентом на Python 2:

```
hygx = px.dot(xlog1x(Px_inv.dot(Pxy))).sum(axis=1)
```

При помощи оператора @, позволяющего оставаться максимально близко к форме математической записи, мы избегаем ошибок реализации и производим программный код, который намного легче читается.

На самом деле авторы библиотеки SciPy учили это задолго до того, как оператор @ был введен, и фактически изменили значение оператора \*, когда входными данными являются матрицы SciPy. Этот оператор существует в Python 2.7 и позволяет создавать хороший, удобочитаемый программный код, как и показанный выше:

```
hygx = -px * xlog(Px_inv * Pxy).sum(axis=1)
```

Однако есть одна большая сложность: этот фрагмент кода будет вести себя по-другому, когда px или Px\_inv станут матрицами SciPy, когда они таковыми не являются! Если Px\_inv и Pxy являются массивами NumPy, то оператор \* производит поэлементное умножение, а если они являются матрицами SciPy, то он производит матричное умножение! Как вы можете справедливо предположить, эта разница является источником очень многих ошибок, и большая часть сообщества SciPy отказалась от его использования в пользу более ужасного, но однозначного метода .dot.

Оператор @ языка Python версии 3.5 предлагает нам лучший из обоих методов!

## ТАБЛИЦЫ СОПРЯЖЕННОСТИ

Но давайте начнем с простого и займемся сегментациями.

Предположим, что вы только что начали работать аналитиком данных в стартапе Spam-o-matic, занимающимся электронными сообщениями. Вам поручена разработка детектора спама. Вы кодируете выход из детектора в виде числового значения, назначая 0 для неспамных сообщений и 1 для спамных.

Если вы хотите разбить набор из 10 электронных сообщений на классы, то в итоге вы получите вектор *предсказаний*:

```
import numpy as np
pred = np.array([0, 1, 0, 0, 1, 1, 1, 0, 1, 1])
```

Вы можете проверить результативность работы детектора, сравнив этот вектор с *контрольным* вектором, т. е. результатом классификации, полученным в результате обследования каждого сообщения вручную

```
gt = np.array([0, 0, 0, 0, 0, 1, 1, 1, 1, 1])
```

Вообще, для компьютеров задача классификации представляет трудность. Так, значения в массивах `pred` и `gt` не полностью совпадают. В тех позициях, где `pred` равняется 0, и `gt` тоже равняется 0, предиктор правильно идентифицировал сообщение как неспамное. Такой исход называется *истинно отрицательным*. С другой стороны, в позициях, где оба значения равняются 1, предиктор правильно идентифицировал спамное сообщение и получил *истинно положительный* исход.

Далее. Есть два вида ошибок. Если мы допустили попадание спамного сообщения (где `gt` равняется 1) в почтовый ящик входящих сообщений пользователя (`pred` равняется 0), то мы совершили *ложноотрицательную* ошибку. При предсказании, что допустимое сообщение (`gt` 0) является спамным (`pred` равняется 1), будет сделано *ложноположительное* предсказание. (Электронное сообщение, пришедшее от директора моего научно-исследовательского института, однажды попало в мою папку для спама. И причина состояла в том, что его объявление о конкурсе работ, участвовавших в защите кандидатской диссертации, начиналось со слов «Вы можете получить приз \$500!».)

Чтобы измерить результативность работы детектора, следует подсчитать приведенные выше виды ошибок, используя *матрицу сопряженности*. (Иногда ее также называют *матрицей ошибок*, что вполне соответствует ее сути.) Для этого мы помещаем метки предсказания вдоль строк и контрольные метки вдоль столбцов. Затем подсчитываем число совпадений. Например, если имеется 4 истинно положительных исхода (где `pred` и `gt` оба равняются 1), матрица будет иметь значение 3 в позиции (1, 1).

<sup>1</sup> В англоязычной терминологии *vector of ground truth*, т. е. вектор данных полевых исследований. – Прим. перев.



В общем случае:

$$C_{i,j} = \sum_k \mathbb{I}(p_k = i) \mathbb{I}(g_k = j).$$

Вот интуитивно понятный, но неэффективный способ реализовать приведенное выше уравнение:

```
def confusion_matrix(pred, gt):
    cont = np.zeros((2, 2))
    for i in [0, 1]:
        for j in [0, 1]:
            cont[i, j] = np.sum((pred == i) & (gt == j))
    return cont
```

Можно проверить, что эта функция дает правильные количества:

```
confusion_matrix(pred, gt)
array([[ 3.,  1.],
       [ 2.,  4.]])
```



## Задача: вычислительная сложность матриц ошибок

Почему мы назвали этот программный код неэффективным? Смотрите решение задачи «Вычислительная сложность матриц ошибок» в конце книги.

## Задача: альтернативный алгоритм вычисления матрицы ошибок

Напишите альтернативный способ вычисления матрицы ошибок, который выполняет всего один обход векторов `pred` и `gt`.

```
def confusion_matrix1(pred, gt):
    cont = np.zeros((2, 2))
    # здесь идет ваш программный код
    return cont
```

Обратите внимание на решение задачи «Альтернативный алгоритм вычисления матрицы ошибок» в конце книги.

Этот пример можно слегка обобщить. Вместо классификации на спам и не спам мы можем классифицировать спам, информационные бюллетени, акции по распродаже и стимулированию продаж, списки рассылок и личные сообщения. Получим 5 категорий, которые пометим от 0 до 4. Матрица ошибок теперь будет иметь размер 5 на 5, в которой совпадения будут подсчитываться в диагональных ячейках, а ошибки – во внедиагональных ячейках.

Определение приведенной выше функции `confusion_matrix` плохо масштабируется на более крупную матрицу, так как теперь мы должны выполнить 25 обходов массивов результирующих и контрольных данных. Эта проблема будет только нарастать по мере добавления новых категорий почтовых сообщений, таких как уведомления социальных сетей.



## Задача: мультиклассовая матрица ошибок

Как и в задании выше, напишите функцию вычисления матрицы ошибок за один проход. Но теперь вместо принятия двух категорий она должна логически выводить количество категорий на основе входных данных.

```
def general_confusion_matrix(pred, gt):
    n_classes = None # заменить `None` на что-то полезное
    # здесь идет ваш программный код
    return cont
```



Ваше однопроходное решение хорошо масштабируется в зависимости от количества классов. Но, так как цикл `for` выполняется в интерпретаторе Python, при большом количестве документов это решение станет медленным. Учитывая, что некоторые классы легко спутать с другими, матрица будет *разреженной* и иметь много нулевых записей. И действительно, по мере увеличения количества классов выделение растущего пространства оперативной памяти под нулевые записи матрицы сопряженности станет наиболее расточительным. Вместо этого мы можем воспользоваться модулем SciPy `sparse`, который содержит объекты для эффективного представления разреженных матриц.

## ФОРМАТЫ ДАННЫХ МОДУЛЯ SCIPY.SPARSE

Мы в главе 1 затронули внутренний формат данных массивов NumPy и надеемся, что он интуитивно понятен и в некотором смысле представляет собой неизбежный формат для хранения данных  $n$ -мерных массивов. На самом деле для разреженных матриц имеется огромное количество возможных форматов, и «правильный» формат зависит от решаемой задачи. Мы рассмотрим два широко применяемых формата. Чтобы получить полный список, обратитесь к сравнительной таблице, приводимой далее в этой главе, или к онлайн-документации по модулю `scipy.sparse`.

### Формат COO

Возможно, наиболее интуитивно понятным является координатный формат, или формат COO. Для представления двумерной матрицы  $A$  в нем используется три одномерных массива.  $A$  именно каждый такой массив имеет длину, равную количеству ненулевых значений в  $A$ , и вместе они формируют перечень координат в формате  $(i, j, \text{значение})$  каждой записи, которая не равна 0.

- Массивы `row` и `col`, вместе задающие позицию каждой ненулевой записи (индексы соответствуют индексам строки и столбца).
- Массив `data`, задающий значение в каждой из этих позиций.

Каждая часть матрицы, которая не представлена парами  $(row, col)$ , считается равной 0. Это гораздо эффективнее! Поэтому, чтобы представить матрицу

```
s = np.array([[ 4,  0,  3],
              [ 0, 32,  0]], dtype=float)
```

мы можем сделать следующее:

```
from scipy import sparse
data = np.array([4, 3, 32], dtype=float)
row = np.array([0, 0, 1])
col = np.array([0, 2, 1])
s_coo = sparse.coo_matrix((data, (row, col)))
```

Метод `.toarray()` каждого разреженного формата в модуле `scipy.sparse` возвращает представление разреженных данных в виде массива NumPy. Его можно применять для проверки правильности создания разреженного массива `s_coo`:

```
s_coo.toarray()
array([[ 4.,  0.,  3.],
       [ 0., 32.,  0.]])
```

Таким же образом мы можем воспользоваться *свойством* `.A`, которое очень похоже на атрибут, но на самом деле исполняет функцию. Свойство `.A` представляет собой исключительно опасное свойство, потому что за ним может скрываться потенциально очень емкая операция: плотная версия разреженной матрицы может быть на порядок больше разреженной матрицы как таковой, ставя компьютер на колени всего тремя нажатиями клавиш!

```
s_coo.A
array([[ 4.,  0.,  3.],
       [ 0., 32.,  0.]])
```

В этой главе, как и в других случаях, если он не ухудшает удобочитаемость, мы рекомендуем использовать метод `toarray()`, так как он яснее сигнализирует о потенциально дорогостоящей операции. Однако мы будем использовать свойство `.A` там, где оно упрощает восприятие программного кода благодаря своей краткости (например, при реализации последовательности математических уравнений).

## Задача: представление в формате COO

Напишите следующую ниже матрицу в формате COO:

```
s2 = np.array([[0, 0, 6, 0, 0],
               [1, 2, 0, 4, 5],
               [0, 1, 0, 0, 0],
               [9, 0, 0, 0, 0],
               [0, 0, 0, 6, 7]])
```

К сожалению, несмотря на то что формат COO интуитивно понятен, он не очень оптимизирован для минимизации потребляемого объема оперативной памяти или для оптимизации скорости обхода массива во время вычисления. (Из главы 1 вы помните, что для эффективных вычислений очень важное значение имеет *сосредоточенность данных*!) Однако вы можете взглянуть на ваше

представление в формате COO и идентифицировать избыточную информацию. Обратите внимание на повторяющиеся единицы.

## Формат сжатой разреженной строки

Если мы используем COO не для того, чтобы перечислить ненулевые записи в произвольном порядке (которое этот формат допускает), а для того, чтобы их перечислить построчно, мы в итоге получим множество последовательных повторяющихся значений в массиве `row`. Вместо того чтобы неоднократно записывать индекс строки, эти значения можно сжать, указывая в `col` индексы, где начинается следующая строка. На этом основывается *формат сжатой разреженной строки*, или формат CSR (Compressed Sparse Row).

Давайте проанализируем приведенный выше пример. В формате CSR массивы `col` и `data` неизменны (но `col` переименовывается в `indices`). Вместе с тем массив `row`, вместо чтобы указывать на строки, указывает место, где в `col` начинается каждая строка, и переименовывается в `indptr`, т. е. в «указатель индекса».

Рассмотрим `row` и `col` в формате COO без учета `data`:

```
row = [0, 1, 1, 1, 1, 2, 3, 4, 4]
col = [2, 0, 1, 3, 4, 1, 0, 3, 4]
```

Каждая новая строка начинается в индексе, где изменяется значение `row`. Нулевая строка начинается в индексе 0, а первая строка начинается в индексе 1, но вторая строка начинается там, где в строке `row` в первый раз появляется «2», т. е. в индексе 5. Затем индексы увеличиваются на 1 для строк 3 и 4 до 6 и 7. Заключительный индекс, указывающий на конец матрицы, представляет собой общее количество ненулевых значений (9). Поэтому:

```
indptr = [0, 1, 5, 6, 7, 9]
```

Давайте применим эти вычисленные вручную массивы, чтобы построить матрицу CSR в SciPy. Мы можем проверить нашу работу, сравнив свойства `.A` наших представлений в формате COO и CSR с массивом NumPy `s2`, который мы определили ранее.

```
data = np.array([6, 1, 2, 4, 5, 1, 9, 6, 7])
coo = sparse.coo_matrix((data, (row, col)))
csr = sparse.csr_matrix((data, col, indptr))

print('Массивы COO и CSR эквивалентны: ',
      np.all(coo.A == csr.A))
print('Массивы CSR и NumPy эквивалентны: ',
      np.all(s2 == csr.A))
```

```
Массивы COO и CSR эквивалентны: True
Массивы CSR и NumPy эквивалентны: True
```

Способность хранить большие разреженные матрицы и выполнять с ними вычисления имеет невероятную силу и может применяться во многих предметных областях.

	<b>bsr_matrix</b>	<b>coo_matrix</b>	<b>csc_matrix</b>	<b>csr_matrix</b>	<b>dia_matrix</b>	<b>dok_matrix</b>	<b>lil_matrix</b>
Полное имя	Блочная разреженная строка (BSR)	Координатная матрица (COO)	Сжатый разреженный столбец (CSC)	Сжатая разреженная строка (CSR)	Диагональная матрица (DIA)	Словарь ключей (DOK)	Построенный связанный список (LIL)
Случай применения	Хранение плотных подматриц. Часто применяется в численном анализе дискретизированных задач, таких как конечные элементы, дифференциальные уравнения	Быстрый и прямой линейный способ создания разреженных матриц. Во время создания дублируются координаты суммируются – например, полезно для анализа конечных элементов	Арифметические операции (поддержка сложения, вычитания, умножения, деления и степени матрицы). Гибкая нарезка столбцов. Быстрые матрично-векторные произведения (CSR, BSR могут быть быстрее, в зависимости от задачи)	Арифметические операции. Эффективная нарезка строк. Быстрые матрично-векторные произведения	Арифметические операции	Недорогое внесение изменений в структуру разреженности. Арифметические операции. Быстрый доступ к индивидуальным элементам. Эффективное преобразование в COO (но дубликаты недопустимы)	Недорогое внесение изменений в структуру разреженности. Гибкая нарезка
Недостатки	Нет арифметических операций. Нет срезов	Медленная нарезка строк (см. CSR). Дорогостоящее внесение изменений в структуру разреженности (см. LIL, DOK)	Медленная нарезка столбцов (см. CSC). Дорогостоящее внесение изменений в структуру разреженности (see LIL, DOK)	Структурная разреженности ограничена значениями по диагонали	Дорогостоящие арифметические операции. Медленное матрично-векторное произведение	Дорогостоящие арифметические операции. Медленная нарезка столбцов. Медленное матрично-векторное произведение	Дорогостоящие арифметические операции. Медленная нарезка столбцов. Медленное матрично-векторное произведение

Например, всю Мировую паутину можно представить как большую разреженную матрицу размера  $N \times N$ . Каждая запись  $X_{ij}$  указывает на то, связана ли веб-страница  $i$  со страницей  $j$ . Нормализовав эту матрицу и найдя ее доминирующий собственный вектор, можно получить так называемую меру PageRank – одно из чисел, которое поисковик Google использует для упорядочивания результатов вашего поискового запроса. (В следующей главе вы узнаете о ней подробнее.)

В качестве еще одного примера мы можем представить человеческий мозг как большой граф размера  $m \times m$ , где есть  $m$  узлов (позиций), в которых вы измеряете активность, используя МРТ-сканер. Через какое-то время после снятия измерений могут быть вычислены корреляции и введены в матрицу  $C_{ij}$ . Пороговая обработка этой матрицы порождает разреженную матрицу, состоящую из единиц и нулей. Собственный вектор, соответствующий второму самому маленькому собственному значению этой матрицы, подразделяет  $m$  мозговых областей в подгруппы, которые, как оказывается, часто связаны с функциональными областями мозга<sup>1</sup>!

## ПРИМЕНЕНИЯ РАЗРЕЖЕННЫХ МАТРИЦ: ПРЕОБРАЗОВАНИЯ ИЗОБРАЖЕНИЙ

Такие библиотеки, как scikit-image и SciPy, уже содержат алгоритмы эффективного преобразования (поворота и деформирования) изображений, но что, если вы являетесь главой Агентства NumPy по космическим делам и должны вращать миллионы изображений, поступающих потоком от недавно запущенного орбитального аппарата Jupyter?

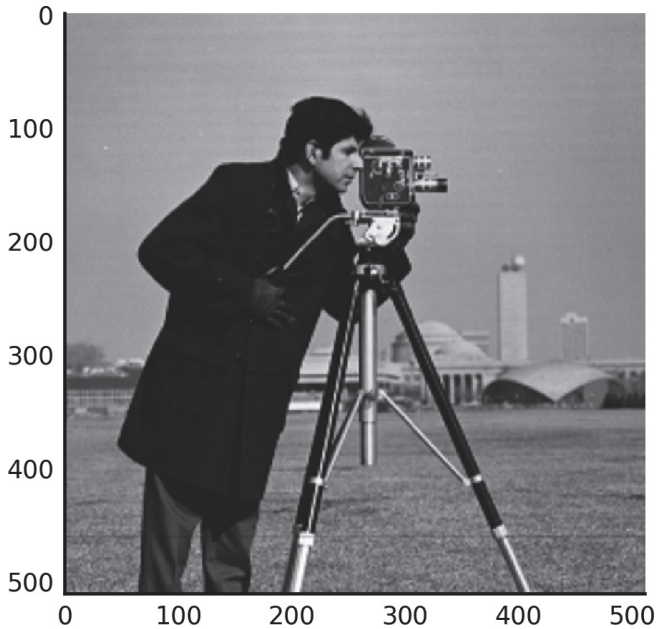
В таких случаях вы захотите выжать из своего компьютера каждую каплю производительности. Оказывается, что мы можем получить намного лучшую результативность, чем даже оптимизированный код на C в модуле SciPy ndimage, если будем многократно применять *одинаковое* преобразование.

В качестве примера данных мы используем приведенное ниже тестовое изображение кионооператора из scikit-image:

```
# Заставить графики появляться локально, задать индивидуальный стиль графиков
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('style/elegant.mplstyle')

from skimage import data
image = data.camera()
plt.imshow(image);
```

<sup>1</sup> Ньюман М. Э. Дж. Модульность и структура сообществ в сетях // PNAS 103. № 23 (2006):8577-8582 (<http://dx.doi.org/DOI:10.1073/pnas.0601602103>).



В качестве тестовой операции будем поворачивать изображение на 30 градусов. Мы начнем с определения матрицы преобразования,  $H$ , которая при умножении с координатой из входного изображения,  $[r, c, 1]$ , даст нам соответствующую координату на выходе,  $[r', c', 1]$ . (Обратите внимание: мы используем *однородные координаты*<sup>1</sup>, где к координатам прибавляется 1 и которые при определении линейных преобразований предоставляют нам большую гибкость.)

```
angle = 30
c = np.cos(np.deg2rad(angle))
s = np.sin(np.deg2rad(angle))
H = np.array([[c, -s, 0],
              [s, c, 0],
              [0, 0, 1]])
```

Вы можете проверить, все работает. Для этого умножьте  $H$  на точку  $(1, 0)$ . 30-градусный поворот против часовой стрелки вокруг начала системы координат  $(0, 0)$  должен нас привести к точке  $(\sqrt{3}/2, 1/2)$ :

```
point = np.array([1, 0, 1])
print(np.sqrt(3) / 2)
print(H @ point)
```

<sup>1</sup> См. [https://en.wikipedia.org/wiki/Homogeneous\\_coordinates](https://en.wikipedia.org/wiki/Homogeneous_coordinates) и [https://ru.wikipedia.org/wiki/Однородная\\_система\\_координат](https://ru.wikipedia.org/wiki/Однородная_система_координат).

0.866025403784

[ 0.8660254 0.5 1. ]



И точно так же трехкратное применение 30-градусного поворота должно нас привести к оси столбцов в точке (0, 1). Мы видим, что за минусом погрешности аппроксимации чисел с плавающей точкой все работает:

```
print(H @ H @ H @ point)
```

```
[ 2.7755756e-16 1.0000000e+00 1.0000000e+00]
```

Теперь мы построим функцию, которая определяет «разреженный оператор» (sparse operator). Задача разреженного оператора состоит в том, чтобы взять все пиксели выходного изображения, выяснить, где они находятся во входном изображении, и, чтобы вычислить их значения, выполнить соответствующую (билинейную) интерполяцию (см. рис. 5.1). Этот оператор работает, применяя к значениям изображения только умножение матриц. Поэтому его быстрое действие чрезвычайно большое.

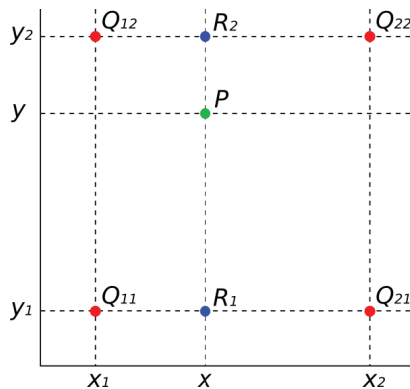


Рис. 5.1 ❖ Диаграмма, объясняющая билинейную интерполяцию, – значение в точке  $P$  оценивается как взвешенная сумма значений в  $Q_{11}$ ,  $Q_{12}$ ,  $Q_{21}$ ,  $Q_{22}$

Давайте рассмотрим функцию, которая строит наш разреженный оператор:

```
from itertools import product
```

```
def homography(tf, image_shape):
```

*«»Проективное (гомографическое) преобразование и интерполяция в качестве линейного оператора.*

*Параметры*

*-----*

*tf : (3, 3) массив ndarray*

*Матрица преобразования.*

*image\_shape : (M, N)*

*форма входного полутонового изображения.*

Возвращает

-----  
**A** : ( $M * N$ ,  $M * N$ ) разреженная матрица

Линейный оператор, представляющий преобразование + билинейная интерполяция.

«»»

# Инvertировать матрицу. Сообщает по каждому выходному пикселу,  
 # где искать соответствующий ему входной пиксел.

`H = np.linalg.inv(tf)`

`m, n = image_shape`

# Мы построим COO-матрицу, так называемую IJK-матрицу, для  
 # которой нам потребуются координаты строк (I),  
 # координаты столбцов (J) и значения (K).

`row, col, values = [], [], []`

# Для каждого пиксела в выходном изображении...

**for** `sparse_or_row, (out_row, out_col) in \`  
`enumerate(product(range(m), range(n))):`

# Вычислить, где он находится во входном изображении

`in_row, in_col, in_abs = H @ [out_row, out_col, 1]`

`in_row /= in_abs`

`in_col /= in_abs`

# Если координаты лежат за пределами исходного изображения, то  
 # проигнорировать эту координату; в этой позиции у нас будет 0

**if** (`not 0 <= in_row < m - 1 or`  
`not 0 <= in_col < n - 1`):

`continue`

# Мы хотим найти четыре окружающих пиксела и интерполировать их  
 # значения, чтобы рассчитать точное значение выходного пиксела.

# Мы начинаем с левого верхнего угла, отмечая, что остальные

# точки отстоят на 1 в каждом направлении.

`top = int(np.floor(in_row))`

`left = int(np.floor(in_col))`

# Вычислить позицию выходного пиксела, отображенного на

# входное изображение, в пределах четырех отобранных пикселей.

# <https://commons.wikimedia.org/wiki/File:BilinearInterpolation.svg>

`t = in_row - top`

`u = in_col - left`

# Текущая строка разреженно-операторной матрицы задается

# развернутыми (`m. e. обработанными функцией gavel`) координатами

# выходных пикселей, содержащимися в `sparse_or_row`.

# Мы возьмем взвешенное среднее четырех окружающих входных

# пикселей, соответствующих четырем столбцам. Поэтому нам нужно

# повторить индекс строки четыре раза.

`row.extend([sparse_or_row] * 4)`

# Фактические веса вычисляются в соответствии с алгоритмом

# билинейной интерполяции, как показано в Википедии

# [https://en.wikipedia.org/wiki/Bilinear\\_interpolation](https://en.wikipedia.org/wiki/Bilinear_interpolation)





```

sparse_op_col = np.ravel_multi_index(
    ([top, top, top + 1, top + 1 ],
     [left, left + 1, left, left + 1]), dims=(m, n))
col.extend(sparse_op_col)
values.extend([(1-t) * (1-u), (1-t) * u, t * (1-u), t * u])

operator = sparse.coo_matrix((values, (row, col)),
                             shape=(m*n, m*n)).tocsr()

return operator

```

Напомним, что мы применяем разреженный оператор следующим образом:

```

def apply_transform(image, tf):
    return (tf @ image.flat).reshape(image.shape)

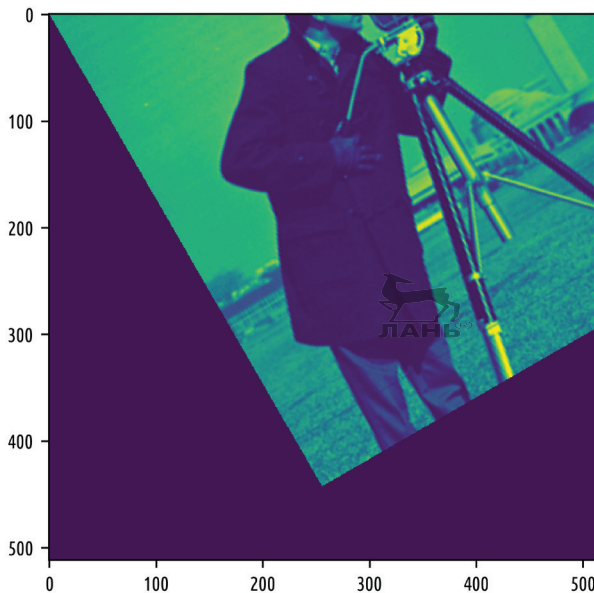
```

Давайте испытаем!

```

tf = homography(H, image.shape)
out = apply_transform(image, tf)
plt.imshow(out);

```



А вот и поворот, о котором идет речь!

### Задача: поворот изображения

Поворот происходит вокруг начала системы координат, т. е. координаты (0, 0). Сможете ли вы выполнить поворот изображения вокруг его центра?

**Совет:** матрица преобразования для *трансляции* (т. е. сдвига изображения вверх/вниз или налево/направо) задается следующим образом:

$$H_{tr} = \begin{bmatrix} 1 & 0 & t_r \\ 0 & 1 & t_c \\ 0 & 0 & 1 \end{bmatrix},$$



когда вы хотите переместить изображение на  $t_r$  пикселей вниз и  $t_c$  пикселей вправо.

Как было отмечено ранее, подход к преобразованию изображения на основе разреженного линейного оператора имеет очень высокое быстродействие. Давайте измерим его производительность по сравнению с `ndimage`. Чтобы сопоставление было справедливым, мы должны сообщить `ndimage`, что нам требуется выполнить линейную интерполяцию с `order=1`, а пиксели за пределами исходной формы с `reshape=False` проигнорируем.

```
%timeit apply_transform(image, tf)
100 loops, average of 7: 3.35 ms +- 270 µs per loop (using standard deviation)

from scipy import ndimage as ndi
%timeit ndi.rotate(image, 30, reshape=False, order=1)
100 loops, average of 7: 19.7 ms +- 988 µs per loop (using standard deviation)
```

На наших машинах мы видим приблизительно 10-кратное ускорение. Хотя этот пример выполняет только поворот, мы можем выполнять и более сложные операции по деформированию, такие как корректировка искаженных линз во время формирования изображения или для придания людям забавных лиц. После вычисления преобразования его многократное применение будет быстрым благодаря разреженно-матричной алгебре.

Итак, мы увидели «стандартное» применение разреженных матриц SciPy. Теперь рассмотрим инновационное применение, вдохновившее нас на написание этой главы.

## НАЗАД К ТАБЛИЦАМ СОПРЯЖЕННОСТИ

Следует напомнить, что мы пытаемся быстро построить разреженную матрицу совместных вероятностей с использованием разреженных форматов SciPy. Мы знаем, что формат COO хранит разреженные данные в виде трех массивов, содержащих координаты строк и столбцов с ненулевыми записями, а также их значения. Однако мы можем воспользоваться малоизвестным свойством формата COO, которое позволяет чрезвычайно быстро получить нашу матрицу.

Взгляните на эти данные:

```
row = [0, 0, 2]
col = [1, 1, 2]
dat = [5, 7, 1]
S = sparse.coo_matrix((dat, (row, col)))
```

Обратите внимание, запись в формате (строка, столбец) в позиции (0, 1) появляется дважды: сначала как 5, далее как 7. Каким должно быть значение



матрицы в позиции (0, 1)? Как вариант может быть выбрана самая первая встретившаяся запись или самая последняя, однако в действительности была выбрана сумма первой и последней записей:

```
print(S.toarray())
```

```
[[ 0 12 0]
 [ 0 0 0]
 [ 0 0 1]]
```

Иными словами, формат COO суммирует повторяющиеся записи. Это именно то, что нам нужно для создания матрицы сопряженности! И действительно, наша задача в значительной степени решена: массиву `pred` можно назначить строки, массиву `gt` назначить столбцы и в качестве значения использовать 1. Затем мы просуммируем единицы и подсчитаем количество раз, когда метка  $i$  из `pred` встречается вместе с меткой  $j$  из `gt` в позиции  $i, j$  матрицы! Давайте это проверим:

```
from scipy import sparse
```

```
def confusion_matrix(pred, gt):
    cont = sparse.coo_matrix((np.ones(pred.size), (pred, gt)))
    return cont
```

Чтобы взглянуть на небольшую матрицу, применим метод `.toarray`, как было показано выше:

```
cont = confusion_matrix(pred, gt)
print(cont)
```

```
(0, 0) 1.0
(1, 0) 1.0
(0, 0) 1.0
(0, 0) 1.0
(1, 0) 1.0
(1, 1) 1.0
(1, 1) 1.0
(0, 1) 1.0
(1, 1) 1.0
(1, 1) 1.0
```



```
print(cont.toarray())
```

```
[[ 3.  1.]
 [ 2.  4.]]
```

Работает!

## Задача: сокращение объема потребляемой оперативной памяти

В главе 1 было сказано, что NumPy имеет встроенные инструменты для повторения массивов на основе операции *транслирования*. Каким образом можно

уменьшить объем потребляемой оперативной памяти, требуемой для вычисления матрицы сопряженности?

**Совет:** обратитесь к документации по функции `np.broadcast_to`.



## ТАБЛИЦЫ СОПРЯЖЕННОСТИ В СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ

Сегментацию изображения можно представить в том же виде, что и выше задачу классификации: сегментная метка в каждом *пикселе* представляет собой *предсказание*, к какому классу принадлежит пиксел. И массивы NumPy позволят нам делать это прозрачно, так как их метод `.ravel()` возвращает одномерное представление лежащих в основе данных.

В качестве примера – сегментация крошечного изображения размером  $3 \times 3$ :

```
seg = np.array([[1, 1, 2],
               [1, 2, 2],
               [3, 3, 3]], dtype=int)
```

Вот контрольные данные со слов некоего человека о том, как правильно сегментировать это изображение:

```
gt = np.array([[1, 1, 1],
              [1, 1, 1],
              [2, 2, 2]], dtype=int)
```

Эти две классификации можно представить точно так же, как и прежде. Каждый пиксел является другим предсказанием.

```
print(seg.ravel())
print(gt.ravel())

[1 1 2 1 2 2 3 3 3]
[1 1 1 1 1 1 2 2 2]
```

Затем, как и прежде, получаем матрицу сопряженности:

```
cont = sparse.coo_matrix((np.ones(seg.size),
                        (seg.ravel(), gt.ravel()))))
print(cont)

(1, 1) 1.0
(1, 1) 1.0
(2, 1) 1.0
(1, 1) 1.0
(2, 1) 1.0
(2, 1) 1.0
(3, 2) 1.0
(3, 2) 1.0
(3, 2) 1.0
```

Некоторые индексы появляются несколько раз, но мы можем воспользоваться свойством суммирования формата COO и подтвердить, что этот результат представляет нужную нам матрицу:

```
print(cont.toarray())
```

```
[[ 0. 0. 0.]
 [ 0. 3. 0.]
 [ 0. 3. 0.]
 [ 0. 0. 3.]]
```

Как преобразовать эту таблицу в меру того, насколько хорошо массив `seg` представляет массив `gt`? Сегментация представляет собой сложную задачу, поэтому важно измерить, как хорошо алгоритм сегментации с ней справляется. Для этого сравним его результат с «контрольной» сегментацией, выполненной человеком вручную.

Но даже такое сравнение не является легкой задачей. Тогда как определить, насколько автоматическая сегментация будет «близкой» по сравнению с контрольной? Мы проиллюстрируем один замечательный метод под названием «*изменчивость информации*» (*variation of information*, VI), или ВИ (Meila, 2005). Он определяется как ответ на следующий вопрос: в среднем, если для случайного пиксела дан его сегментный идентификатор в одной сегментации, сколько еще потребуется информации для определения его идентификатора в другой сегментации?

На интуитивном уровне, если эти две сегментации полностью одинаковы, сведения о сегментном идентификаторе в одной сегментации говорят о сегментном идентификаторе в другой без дополнительной информации. Но по мере того, как сегментации все больше отличаются, сведения об идентификаторе в одной сегментации не говорят об идентификаторе в другой без дополнительной информации.

## ТЕОРИЯ ИНФОРМАЦИИ ВКРАТЦЕ



Чтобы ответить на этот вопрос, нам потребуется оперативная сводка о теории информации. Эта сводка будет краткой. Однако если вам требуется более подробная информация, то (увы) вам следует обратиться к звездной публикации в блоге Кристофера Олаха (Christopher Olah) «Визуальная теория информации»<sup>1</sup>.

Базовой единицей информации является *бит*. Бит представляется как 0 или 1, равновероятностный выбор между двумя вариантами. Это прямолинейно просто: если я хочу вам сказать, повернулась ли монета в результате броска орлом либо решкой, мне потребуется один бит, и такой вариант может принимать много разных форм: длинный или короткий сигнал по телеграфному проводу (как в азбуке Морзе), луч света, сверкающий одним цветом из двух, или же одиночное число, принимающее значения 0 либо 1. Важным является то, что мне *всегда* нужен один бит, так как исход броска монеты случаен.

Оказывается, что это понятие можно расширить до *дробных битов* для менее случайных событий. Предположим, вам требуется передать, шел сегодня в Лос-Анджелесе дождь или нет. На первый взгляд кажется, что это тоже потребует 1 бит: 0 для того, что дождя не было, 1 для того, что лил дождь. Однако дождь

<sup>1</sup> См. <https://colah.github.io/posts/2015-09-Visual-Information/>.

в Лос-Анджелесе является редким событием. Поэтому со временем мы можем обойтись передачей намного меньшей информации: изредка будем передавать 0, чтобы удостовериться, что наш канал связи по-прежнему работает. В противном случае примем как допущение, что сигнал равняется 0, и отправлять 1 только в редких случаях, когда идет дождь.

Следовательно, когда два события *не* равновероятны, для их представления потребуется меньше одного бита. Обычно мы это измеряем для любой случайной величины  $X$  (которая может иметь больше двух возможных значений) при помощи функции энтропии  $H$ :

$$\begin{aligned} H(X) &= \sum_x p_x \log_2 \left( \frac{1}{p_x} \right) \\ &= -\sum_x p_x \log_2(p_x), \end{aligned}$$

где значения  $x$  – это возможные значения  $X$  и  $p_x$  – вероятность, что  $X$  примет значение  $x$ . Поэтому энтропия броска монеты  $T$ , могущего принимать значения «орел» ( $h$ ) и «решка» ( $t$ ), равняется:

$$\begin{aligned} H(T) &= p_h \log_2(1/p_h) + p_t \log_2(1/p_t) \\ &= 1/2 \log_2(2) + 1/2 \log_2(2) \\ &= 1/2 \cdot 1 + 1/2 \cdot 1 \\ &= 1. \end{aligned}$$



Долгосрочная вероятность дождя в любой конкретный день в Лос-Анджелесе равняется приблизительно 1 из 6. Поэтому энтропия дождя в Лос-Анджелесе,  $R$ , принимающая значения дождливо ( $r$ ) либо солнечно ( $s$ ), равняется:

$$\begin{aligned} H(R) &= p_r \log_2(1/p_r) + p_s \log_2(1/p_s) \\ &= 1/6 \log_2(6) + 5/6 \log_2(6/5) \\ &\approx 0.65 \text{ бита.} \end{aligned}$$

Особым видом энтропии является *условная энтропия*. Это энтропия величины *при условии*, что об этой величине вы также знаете что-то еще. Например, какова энтропия дождя, *если* известен месяц? Она записывается так:

$$H(R|M) = \sum_{m=1}^{12} p(m) H(R|M=m)$$

и:

$$\begin{aligned} H(R|M=m) &= p_{r|m} \log_2 \left( \frac{1}{p_{r|m}} \right) + p_{s|m} \log_2 \left( \frac{1}{p_{s|m}} \right) \\ &= \frac{p_{rm}}{p_m} \log_2 \left( \frac{p_m}{p_{rm}} \right) + \frac{p_{sm}}{p_m} \log_2 \left( \frac{p_m}{p_{sm}} \right) \\ &= -\frac{p_{rm}}{p_m} \log_2 \left( \frac{p_{rm}}{p_m} \right) - \frac{p_{sm}}{p_m} \log_2 \left( \frac{p_{sm}}{p_m} \right). \end{aligned}$$

Теперь вы располагаете достаточными сведениями обо всей теории информации, которая вам потребуется для понятия изменчивости информации. В предыдущем примере событиями являются дни, и они имеют два свойства:

- дождливо/солнечно;
- месяц.

Проведя наблюдения в течение многих дней, мы можем построить *матрицу сопряженности*, которая будет точно такой, что и в примерах классификации, с показаниями за определенный день месяца, была ли в этот день погода дождливой. Чтобы это сделать, мы не собираемся ехать в Лос-Анджелес (как бы весело это не было) и вместо этого будем использовать приведенную ниже историческую таблицу, составленную на глаз на основе данных метеорологического веб-сайта WeatherSpark<sup>1</sup>:

Месяц	P(дождливо)	P(солнечно)
1	0.25	0.75
2	0.27	0.73
3	0.24	0.76
4	0.18	0.82
5	0.14	0.86
6	0.11	0.89
7	0.07	0.93
8	0.08	0.92
9	0.10	0.90
10	0.15	0.85
11	0.18	0.82
12	0.23	0.77



Тогда условная энтропия дождя при заданном месяце будет следующей:

$$\begin{aligned}
 H(R|M) &= -\frac{1}{12}(0.25 \log_2(0.25) + 0.75 \log_2(0.75)) - \frac{1}{12}(0.27 \log_2(0.27) + 0.73 \log_2(0.73)) \\
 &\quad - \dots - \frac{1}{12}(0.23 \log_2(0.23) + 0.77 \log_2(0.77)) \\
 &\sim 0.626 \text{ бита.}
 \end{aligned}$$

Итак, используя месяц, мы уменьшили хаотичность сигнала, но ненамного!

Мы также можем вычислить условную энтропию месяца при наличии дождя, которая измеряет количество информации, требующейся для определения месяца, если мы знаем, что шел дождь. На интуитивном уровне мы знаем, это лучше, чем идти вслепую, так как вероятность дождливой погоды гораздо выше в зимние месяцы.

<sup>1</sup> См. <https://weatherspark.com/y/1705/Average-Weather-in-Los-Angeles-California-United-States-Year-Round>.



## Задача: вычисление условной энтропии

Вычислите условную энтропию месяца при наличии дождя. Какова энтропия переменной месяца? (Не учитывайте разное количество дней в месяце.) Какая из них больше?

➔ Представленные в таблице вероятности являются условными вероятностями дождя при наличии месяца.

```
prains = np.array([25, 27, 24, 18, 14, 11, 7, 8, 10, 15, 18, 23]) / 100
pshine = 1 - prains
p_rain_g_month = np.column_stack([prains, pshine])
# Замените 'None' на выражение для таблицы безусловной сопряженности
# Совет: сумма значений в таблице должна равняться 1.
p_rain_month = None
# Добавьте свой собственный исходный код ниже, чтобы вычислить
# H(M|R) и H(M)
```

Эти два значения вместе определяют изменчивость информации (VI):

$$VI(A, B) = H(A|B) + H(B|A).$$



## ТЕОРИЯ ИНФОРМАЦИИ ПРИМЕНительно К СЕГМЕНТАЦИИ: ИЗМЕНЧИВОСТЬ ИНФОРМАЦИИ

Если вернуться назад в контекст сегментации изображений, то «дни» становятся «пикселями», а «дождь» и «месяц» становятся «меткой в автоматической сегментации ( $S$ )» и «контрольной меткой ( $T$ )». Затем условная энтропия автоматической сегментации при наличии контрольных данных дает меру количества дополнительной информации, которая нам потребуется, для определения идентичности пикселя в  $S$ , если нам известна его идентичность в  $T$ . Например, если каждый сегмент  $g$  в  $T$  разбивается на два равноразмерных сегмента  $a_1$  и  $a_2$  в  $S$ , то  $H(S|T) = 1$ . Если пиксел находится в  $g$ , чтобы узнать, принадлежит ли он  $a_1$  или  $a_2$ , вам по-прежнему нужен 1 дополнительный бит. Однако  $H(T|S) = 0$ , потому что, независимо от того, находится ли пиксел в  $a_1$  или  $a_2$ , он гарантированно находится в  $g$ , и поэтому, в отличие от сегмента в  $S$ , дополнительная информация не нужна.

Поэтому в данном случае все вместе:

$$VI(S, T) = H(S|T) + H(T|S) = 1 + 0 = 1 \text{ бит.}$$

Вот простой пример:

```
S = np.array([[0, 1],
              [2, 3]], int)
T = np.array([[0, 1],
              [0, 1]], int)
```



Здесь мы имеем две сегментации четырехпиксельных изображений:  $S$  и  $T$ . Сегментация  $S$  помещает каждый пиксел в свой собственный сегмент, в то время как сегментация  $T$  помещает левые два пиксела в сегмент 0 и правые два пиксела – в сегмент 1. Теперь точно так же, как мы сделали с метками предсказания спама, мы создадим таблицу сопряженности пиксельных меток. Единственная разница – в том, что массивы меток, в отличие от одномерных массивов предсказаний, являются двумерными. По сути, это не имеет значения: вспомните, что на самом деле массивы NumPy представляют собой линейные (одномерные) блоки данных, за которыми закреплена информация об их форме и другие метаданные. Как мы отмечали ранее, мы можем проигнорировать форму при помощи метода `ravel()` для массивов:

```
S.ravel()
array([0, 1, 2, 3])
```

Теперь можно легко создать таблицу сопряженности, точно так, как мы делали, когда предсказывали спам:

```
cont = sparse.coo_matrix((np.broadcast_to(1., S.size),
                          (S.ravel(), T.ravel())))
cont = cont.toarray()
cont
array([[ 1., 0.],
       [ 0., 1.],
       [ 1., 0.],
       [ 0., 1.]])
```



Чтобы вместо количеств эта таблица содержала вероятности, мы просто выполним деление на общее количество пикселей:

```
cont /= np.sum(cont)
```

Наконец, эту таблицу можно применить для вычисления вероятности меток в  $S$  либо в  $T$ , используя суммы вдоль осей:

```
p_S = np.sum(cont, axis=1)
p_T = np.sum(cont, axis=0)
```

В написании исходного кода Python для вычисления энтропии имеется небольшое отклонение: хотя  $0 \log(0)$  по определению равняется 0, в Python это выражение не определено и возвращает значение `nan` (not a number, не цифра):

```
print('Логарифм 0 равняется: ', np.log2(0))
print('Произведение 0 на логарифм 0 равняется: ', 0 * np.log2(0))
```

```
Логарифм 0 равняется: -inf
Произведение 0 на логарифм 0 равняется: nan
```

Поэтому, чтобы наложить маску на нулевые значения, мы должны воспользоваться индексацией NumPy. Кроме того, в зависимости, является ли

вход массивом NumPy или разреженной матрицей SciPy, нам потребуется не-много другая стратегия. Мы напишем приведенную ниже вспомогательную функцию:

```
def xlog1x(arr_or_mat):
    «»Вычислить поэлементную функцию энтропии для массива или матрицы.

    Параметры
    -----
    arr_or_mat : массив numpy или разреженная матрица scipy
    Входной массив вероятностей. Поддерживаются только форматы
    разреженных матриц с атрибутом `data`.

    Возвращает
    -----
    out : массив или разреженная матрица, тот же тип, что и на входе
    Результирующий массив. Нулевые записи во входных данных
    остаются нулевыми; все другие записи умножаются на логарифм
    (по основанию 2) их инверсии.
    «»»»

    out = arr_or_mat.copy()
    if isinstance(out, sparse.spmatrix):
        arr = out.data
    else:
        arr = out
    nz = np.nonzero(arr)
    arr[nz] *= -np.log2(arr[nz])
    return out
```

Давайте проверим, работает ли эта функция:

```
a = np.array([0.25, 0.25, 0, 0.25, 0.25])
xlog1x(a)
array([ 0.5, 0.5, 0. , 0.5, 0.5])

mat = sparse.csr_matrix([[0.125, 0.125, 0.25, 0],
                        [0.125, 0.125, 0, 0.25]])
xlog1x(mat).A
array([[ 0.375, 0.375, 0.5, 0. ],
       [ 0.375, 0.375, 0. , 0.5 ]])
```

Поэтому условная энтропия  $S$  при наличии  $T$ :

```
H_ST = np.sum(np.sum(xlog1x(cont / p_T), axis=0) * p_T)
H_ST
1.0
```

И обратное:

```
H_TS = np.sum(np.sum(xlog1x(cont / p_S[:, np.newaxis]), axis=1) * p_S)
H_TS
0.0
```



## КОНВЕРТИРОВАНИЕ ПРОГРАММНОГО КОДА МАССИВОВ NumPy ПОД ИСПОЛЬЗОВАНИЕ РАЗРЕЖЕННЫХ МАТРИЦ

В приведенных выше примерах мы использовали массивы NumPy и операцию транслирования. Мы уже не раз убеждались, это мощные инструменты анализа данных в Python. Однако для сегментации сложных изображений, содержащих тысячи сегментов, этот способ не эффективен. Лучше во время вычислений воспользоваться модулем `sparse` и придать части чудесных особенностей NumPy форму линейно-алгебраических операций. Это было предложено<sup>1</sup> Уорреном Векессером (Warren Weckesser) на StackOverflow.

Линейно-алгебраическая версия элегантно короткая и эффективно вычисляет матрицу сопряженности для очень больших объемов данных, порядка миллиардов точек.

```
import numpy as np
from scipy import sparse

def invert_nonzero(arr):
    arr_inv = arr.copy()
    nz = np.nonzero(arr)
    arr_inv[nz] = 1 / arr[nz]
    return arr_inv
```



```
def variation_of_information(x, y):
    # Вычислить матрицу сопряженности, т. н. матрицу совместной вероятности
    n = x.size
    Pxy = sparse.coo_matrix((np.full(n, 1/n), (x.ravel(), y.ravel()))),
                           dtype=float).tocsr()

    # Вычислить маргинальные вероятности, преобразовав в одномерный массив
    px = np.ravel(Pxy.sum(axis=1))
    py = np.ravel(Pxy.sum(axis=0))

    # Использовать разреженно-матричную линейную алгебру, чтобы сначала
    # вычислить изменчивость информации (VI),
    # вычислить обратные диагональные матрицы
    Px_inv = sparse.diags(invert_nonzero(px))
    Py_inv = sparse.diags(invert_nonzero(py))

    # затем вычислить энтропии
    hygx = px @ xlog1x(Px_inv @ Pxy).sum(axis=1)
    hxgy = xlog1x(Pxy @ Py_inv).sum(axis=0) @ py

    # вернуть их сумму
    return float(hygx + hxgy)
```

Мы можем проверить, что эта функция дает правильное значение (1) для изменчивости информации наших игрушечных S и T:

<sup>1</sup> См. <https://stackoverflow.com/questions/16043299/substitute-for-numpy-broadcasting-using-scipy-sparse-csc-matrix>.

variation\_of\_information(S, T)

1.0

Вы видите, каким образом мы используем три типа разреженных матриц (COO, CSR и диагональную) для эффективного вычисления энтропии разреженных матриц сопряженности, в которых библиотека NumPy неэффективна. (На самом деле появление этого подхода стало следствием ошибки Python MemoryError!)

## ПРИМЕНЕНИЕ ИЗМЕНЧИВОСТИ ИНФОРМАЦИИ



В заключение продемонстрируем применение изменчивости информации для вычисления наилучшей автоматической сегментации изображения из возможных. Вы, вероятно, помните нашего дружелюбного тигра из главы 3 (см. рис. 5.2). (Если нет, то вам следует поработать над своими навыками оценивания угроз!) Используя навыки, полученные в главе 3, сегментируем ряд возможных вариантов сегментирования изображения тигра и затем выясним, какой из них самый лучший.

```
from skimage import io
```

```
url = ('http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds'  
      '/BSDS300/html/images/plain/normal/color/108073.jpg')
```

```
tiger = io.imread(url)
```



```
plt.imshow(tiger);
```



Рис. 5.2 ❖ Изображение тигра из набора данных BSDS, номер 108073

Чтобы проверить нашу сегментацию изображения, нам потребуется немного контрольных данных. Оказывается, люди обладают удивительной способностью идентифицировать тигров (естественный отбор, знаете ли!), поэтому от нас требуется только попросить человека найти тигра на фотографии. К сча-

стью, исследователи в Беркли уже попросили десятки людей посмотреть на это изображение и вручную его сегментировать<sup>1</sup>.

Давайте возьмем одно из отсегментированных изображений из набора данных эталонных сегментаций университета Беркли (Berkeley Segmentation Dataset and Benchmark, BSDS) (см. рис. 5.3)<sup>2</sup>. Следует отметить, между сегментациями, выполненными людьми, имеется довольно существенная вариация. Если вы просмотрите различные сегментации тигра<sup>3</sup>, то обнаружите, что некоторые люди педантичнее других в прорисовке тростников. Другие же считают, что водяные блики заслуживают выделения в виде сегментов из остальной части воды. Мы выбрали понравившуюся нам сегментацию (с педантичной прорисовкой тростника, потому что мы принадлежим к тому типу ученых, которых называют перфекционистами). Однако сразу же внесем ясность, у нас нет ни одного фрагмента контрольных данных!

```
from scipy import ndimage as ndi
from skimage import color
```

```
human_seg_url = ('http://www.eecs.berkeley.edu/Research/Projects/CS/'
                'vision/bsds/BSDS300/html/images/human/normal/'
                'outline/color/1122/108073.jpg')
boundaries = io.imread(human_seg_url)
plt.imshow(boundaries);
```

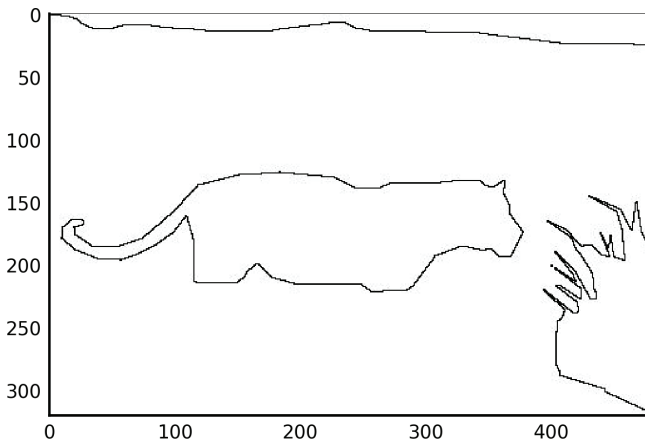


Рис. 5.3 ❖ Сегментация изображения тигра, выполненная человеком

<sup>1</sup> См.: Арбелаес П., Мэр М., Фолкс Ч., Малик Дж. Обнаружение контура и иерархическая сегментация изображения (*Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik*. Contour Detection and Hierarchical Image Segmentation // IEEE TPAMI 33. № 5 (2011): 898–916).

<sup>2</sup> См. <https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/>.

<sup>3</sup> См. <https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images/color/108073.html>.



```

for n in g:
    g.node[n]['total color'] = np.zeros(3, np.double)
    g.node[n]['pixel count'] = 0

for index in np.ndindex(labels.shape):
    n = labels[index]
    g.node[n]['total color'] += image[index]
    g.node[n]['pixel count'] += 1
return g

def threshold_graph(g, t):
    to_remove = [(u, v) for (u, v, d) in g.edges(data=True)
                  if d['weight'] > t]
    g.remove_edges_from(to_remove)

# Базовая сегментация
from skimage import segmentation
seg = segmentation.slic(tiger, n_segments=30, compactness=40.0,
                        enforce_connectivity=True, sigma=3)
plt.imshow(color.label2rgb(seg, tiger));

```

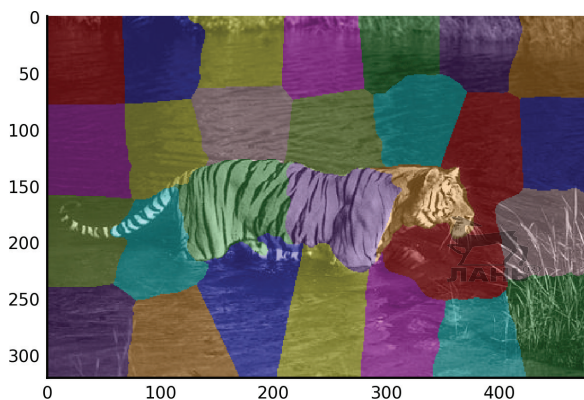


Рис. 5.5 ❖ Базовая сегментация изображения тигра на основе SLIC-алгоритма (простая линейная итеративная кластеризация)

В главе 3 мы назначили графу порог величиной 80 и для ясности опустили технические детали. Теперь мы собираемся рассмотреть внимательней, как этот порог влияет на точность нашей сегментации. Давайте вставим исходный код сегментации в функцию, чтобы с ним поэкспериментировать.

```

def rag_segmentation(base_seg, image, threshold=80):
    g = build_rag(base_seg, image)
    for n in g:
        node = g.node[n]
        node['mean'] = node['total color'] / node['pixel count']

    for u, v in g.edges():
        d = g.node[u]['mean'] - g.node[v]['mean']
        g[u][v]['weight'] = np.linalg.norm(d)

```



```

threshold_graph(g, threshold)

map_array = np.zeros(np.max(seg) + 1, int)
for i, segment in enumerate(nx.connected_components(g)):
    for initial in segment:
        map_array[int(initial)] = i
segmented = map_array[seg]
return(segmented)

```

Теперь попробуем несколько порогов и посмотрим, что произойдет (см. рис. 5.6 и 5.7):



```

auto_seg_10 = rag_segmentation(seg, tiger, threshold=10)
plt.imshow(color.label2rgb(auto_seg_10, tiger));

```

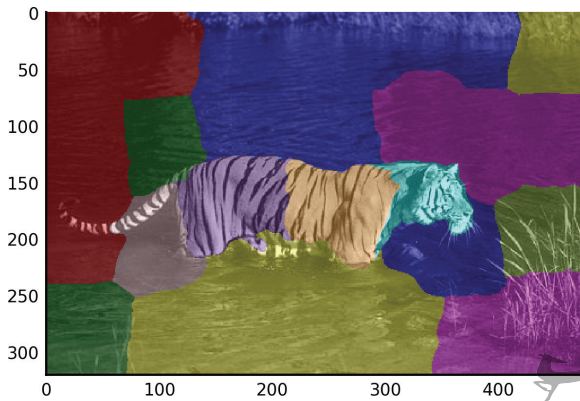


Рис. 5.6 ❖ Сегментация тигра на основе графа RAG с порогом 10

```

auto_seg_40 = rag_segmentation(seg, tiger, threshold=40)
plt.imshow(color.label2rgb(auto_seg_40, tiger));

```

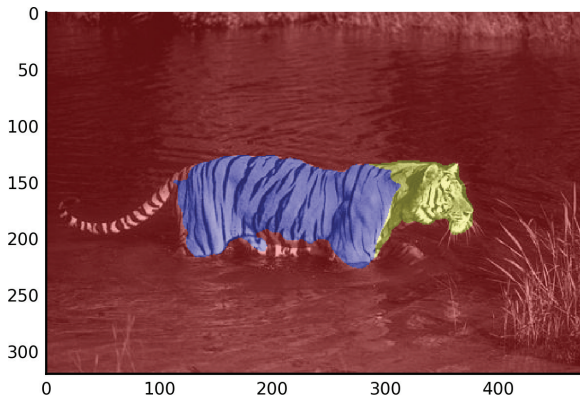


Рис. 5.7 ❖ Сегментация тигра на основе графа RAG с порогом 40



На самом деле в главе 3 мы выполнили сегментацию несколько раз с разными порогами и затем (потому что мы – люди, поэтому имеем право) выбрали ту, который произвел хорошую сегментацию. Это абсолютно неудовлетворительный подход к программированию сегментации изображения. Очевидно, нам нужно этот процесс как-то автоматизировать.

Мы видим, что более высокий порог порождает более оптимальную сегментацию. Но у нас есть контрольные данные, поэтому в действительности мы можем назначить сегментации число! Используя все наши навыки работы с разреженной матрицей, можем вычислить изменчивость информации для каждой сегментации.

```
variation_of_information(auto_seg_10, human_seg)
```

```
3.44884607874861
```

```
variation_of_information(auto_seg_40, human_seg)
```

```
1.0381218706889725
```

Высокий порог имеет более низкую изменчивость информации и, значит, более оптимальную сегментацию! Теперь можем вычислить изменчивость информации для диапазона возможных порогов и увидеть, какой из них дает сегментацию, ближе всего соответствующую контрольным данным человека (рис. 5.8).

```
# Проверить несколько порогов
```

```
def vi_at_threshold(seg, tiger, human_seg, threshold):
    auto_seg = rag_segmentation(seg, tiger, threshold)
    return variation_of_information(auto_seg, human_seg)
```

```
thresholds = range(0, 110, 10)
```

```
vi_per_threshold = [vi_at_threshold(seg, tiger, human_seg, threshold)
                    for threshold in thresholds]
```

```
plt.plot(thresholds, vi_per_threshold);
```

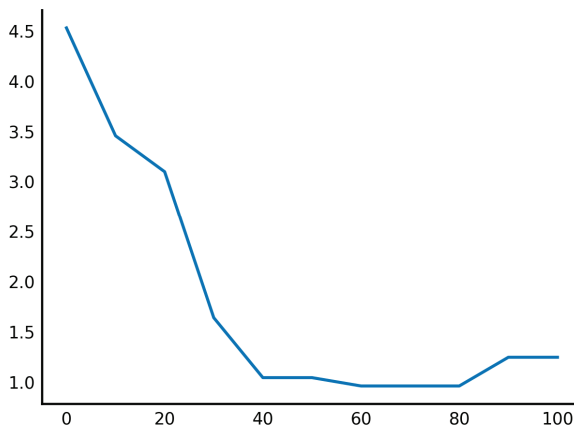


Рис. 5.8 ❖ Сегментация на основе изменчивости информации как функции порога

Как и следовало ожидать, оказывается, визуальный осмотр и выбор порога `threshold=80` действительно дали одну из лучших сегментаций (рис. 5.9). Однако теперь у нас есть способ автоматизации этого процесса для любого изображения!

```
auto_seg = rag_segmentation(seg, tiger, threshold=80)
plt.imshow(color.label2rgb(auto_seg, tiger));
```

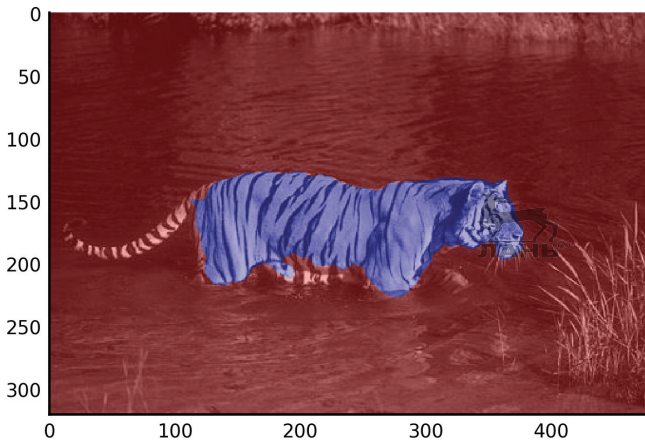


Рис. 5.9 ❖ Оптимальная сегментация тигра на основе кривой изменчивости информации

### Дальнейшая работа: сегментация на практике

Попытайтесь найти наилучший порог для подборки других изображений из набора данных эталонных сегментаций университета Беркли<sup>1</sup>. Используя среднее или медиану этих порогов, попробуйте сегментировать новое изображение. Насколько удачной будет полученная вами сегментация?

Разреженные матрицы являются эффективным способом представления данных со многими промежутками – такая ситуация происходит довольно часто. После прочтения этой главы вы, вероятно, начнете постоянно замечать возможности для их применения.., и вы будете знать, как это делать.

Один из конкретных случаев, где разреженные матрицы по-настоящему приходят на выручку, лежит в области линейной алгебры. Читайте дальше, и в следующей главе вы узнаете об этом больше!

<sup>1</sup> См.: Арбелаес П., Мэр М., Фолкс Ч., Малик Дж. Обнаружение контура и иерархическая сегментация изображения (Contour Detection and Hierarchical Image Segmentation // IEEE TPAMI 33. № 5 (2011): 898–916).

---

# Глава 6



.....

## Линейная алгебра в SciPy

Увы, невозможно объяснить, что такое Матрица... Ты должен увидеть это сам.

– Морфеус, к/ф «Матрица»

Точно так же, как и в главе 4, посвященной быстрому преобразованию Фурье (БПФ), в центре внимания настоящей главы будет элегантный *метод*. Мы хотим выделить имеющиеся в SciPy программные пакеты, позволяющие применять линейную алгебру, формирующую основу большинства научных вычислений.

### ОСНОВЫ ЛИНЕЙНОЙ АЛГЕБРЫ

Глава в книге по программированию является не совсем подходящим местом, где можно изучить линейную алгебру как таковую. Поэтому мы исходим из того, что читатель знаком с понятиями линейной алгебры. Как минимум, вы должны знать, что линейная алгебра связана с векторами (упорядоченной коллекцией чисел) и их преобразованием путем умножения на матрицы (коллекции векторов). Если все это для вас выглядит как тарабарщина, прежде чем приступить к чтению этой главы, вам, вероятно, следует подобрать учебник с вводным курсом линейной алгебры. Мы настоятельно рекомендуем учебник «Линейная алгебра и ее применения» Джила Стрэнга (Linear Algebra and Its Applications, Gil Strang, Pearson, 1994). Причем вам потребуются только введение – мы надеемся передать силу методов линейной алгебры, сохранив операции относительно простыми!

Попутно заметим: мы нарушим общепринятые правила написания программного кода Python в угоду соблюдению линейно алгебраической формы записи. В Python имена переменных обычно начинаются буквами в нижнем регистре. Однако в линейной алгебре матрицы обозначаются прописной буквой, а векторы и скалярные величины – буквами в нижнем регистре. Поскольку мы собираемся работать с большим количеством матриц и векторов, соблюдение линейно-алгебраической формы записи помогает сохранить прямолинейное соответствие. Поэтому переменные, представляемые матрицами, будут начинаться с прописной буквы, в то время как векторы и числа – с букв в нижнем регистре:

```
import numpy as np
m, n = (5, 6)           # скаляры
M = np.ones((m, n))    # матрица
v = np.random.random((n,)) # вектор
w = M @ v              # еще один вектор
```



В математической форме записи векторы, в отличие от скалярных величин, как правило, записываются полужирным шрифтом, как  $\mathbf{v}$  и  $\mathbf{w}$ . Скалярные величины записываются как  $m$  и  $n$ . Мы не сможем поддерживать это различие в программном коде Python. Поэтому, чтобы различать скаляры и векторы, мы будем опираться на контекст.

## ЛАПЛАСОВА МАТРИЦА ГРАФА

Графы мы рассмотрели в главе 3, где представляли области изображения как узлы, связанные между собой ребрами. Однако мы использовали довольно простой метод анализа: выполняли *пороговую обработку* графа, удаляя все ребра, которые были выше некоторого значения. Пороговая обработка хорошо работает лишь в простых случаях. Но достаточно одного значения, попавшего на неправильную сторону порога, чтобы этот подход не сработал.

Как пример предположим, что вы находитесь в состоянии войны и вражеские войска дислоцировались на другом берегу реки напротив ваших сил. Вы хотите отрезать все подходы и поэтому решаете взорвать все мосты между вами. Разведка предлагает, что  $t$  килограмм тротиловой взрывчатки будет достаточно для подрыва всех мостов через реку. Но мосты на вашей собственной территории могут выдержать  $t + 1$  кг. Прочитав главу 3, вы можете отдать своей диверсионно-разведывательной группе приказ взорвать  $t$  кг тротила под каждым мостом в этом районе. Но если сведения разведки оказались верными ко всем мостам, кроме одного, и он останется стоять, то вражеская армия сможет переправиться! И тогда беда!

Поэтому в этой главе мы займемся исследованием некоторых альтернативных подходов к анализу графов на основе линейной алгебры. Оказывается, можно представить граф  $G$  как *матрицу смежности*, в которой мы пронумеруем узлы графа от 0 до  $n - 1$ , и помещаем 1 в строке  $i$ , столбце  $j$  матрицы всякий раз, когда имеется ребро из узла  $i$  в узел  $j$ . Другими словами, если мы назовем матрицу смежности  $A$ , тогда  $A_{i,j} = 1$ , если и только если ребро  $(i, j)$  находится в  $G$ . И тогда мы можем применить линейно-алгебраические методы исследования этой матрицы, часто получая поразительные результаты.

*Степень*, или валентность, узла (вершины) графа определяется как количество ребер, инцидентных этому узлу. Например, если узел связан с пятью другими узлами в графе, то его степень равняется 5. (Позже мы будем различать между полустепенью захода и полустепенью исхода, когда ребра имеют «вход» и «выход».) В терминах матриц степень соответствует сумме значений в строке или столбце матрицы.

Лапласова матрица графа (для краткости иногда «лапласиан») определяется как степенная матрица,  $D$ , содержащая степень каждого узла вдоль диагонали и нуль в остальных ячейках минус матрица смежности  $A$ :

$$L = D - A.$$

Мы определенно не сможем втиснуть всю теорию линейной алгебры, чтобы разобраться в свойствах этой матрицы. Достаточно сказать, что она имеет несколько *замечательных* свойств. И мы задействуем пару таких свойств в следующих абзацах.

Прежде всего мы рассмотрим *собственные векторы*  $L$ . Собственный вектор  $v$  матрицы  $M$  является вектором, который удовлетворяет свойству  $Mv = \lambda v$  для некоторого числа  $\lambda$ , которое называется собственным значением. Другими словами,  $v$  – это особый вектор относительно  $M$ , потому что  $Mv$  просто изменяет размер вектора, не изменяя его направления. Как мы вскоре увидим, собственные векторы имеют много полезных и волшебных свойств!

Например,  $3 \times 3$  – матрица  $R$  поворота при умножении на любой трехмерный вектор  $p$  поворачивает его на 30 градусов вокруг оси  $z$ . Матрица  $R$  будет поворачивать все векторы, за исключением лежащих на оси  $z$ . Для них мы не увидим эффекта, или  $Rp = p$  (т. е.  $Rp = \lambda p$ ) с собственным значением  $\lambda = 1$ .

## Задача: матрица поворота

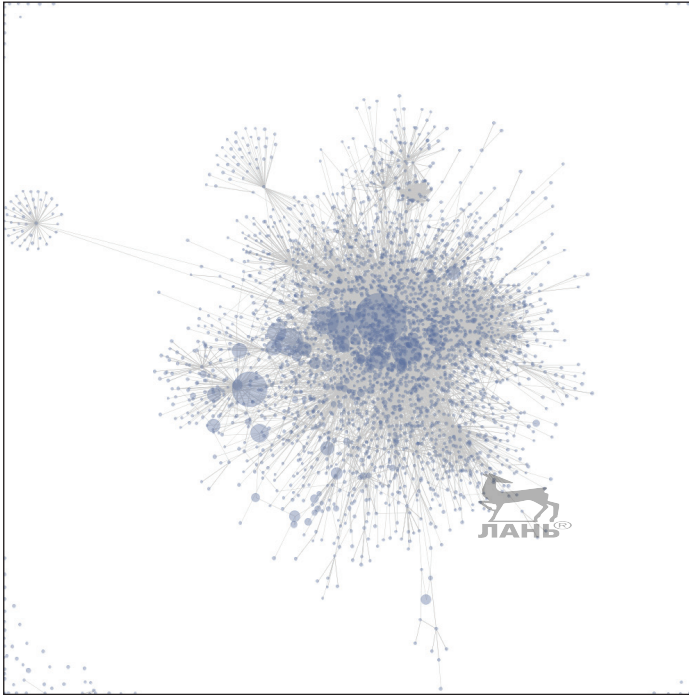
Рассмотрим матрицу поворота:

$$R = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

При умножении  $R$  на трехмерный столбцовый вектор  $p = [x \ y \ z]^T$  результирующий вектор  $Rp$  поворачивается на  $\theta$  градусов вокруг оси  $z$ .

1. Для  $\theta = 45^\circ$  следует подтвердить (путем проверки на нескольких произвольных векторах), что  $R$  поворачивает эти векторы вокруг оси  $z$ . Напомним, что матричное умножение в Python обозначается символом `@`.
2. Что делает матрица  $S = RR$ ? Проверить это в Python.
3. Подтвердить, что умножение на  $R$  оставляет вектор  $[0 \ 0 \ 1]^T$  неизменным. Другими словами,  $Rp = 1p$ , т. е.  $p$  – это собственный вектор матрицы  $R$  с собственным значением 1.
4. Применить функцию `np.linalg.eig` для нахождения собственных значений и собственных векторов матрицы  $R$  и подтвердить, что  $[0 \ 0 \ 1]^T$  действительно находится среди них и соответствует собственному значению 1.

Вернемся к лапласовой матрице. Типичной задачей в сетевом анализе является визуализация. Как нарисовать узлы и ребра так, чтобы не получить полный беспорядок, как, например, на рис. 6.1?



**Рис. 6.1** ❖ Визуализация структуры Википедии  
(созданной Крисом Дэвисом и выпущенной согласно лицензии CC SA 3.0)

Один из способов состоит в том, чтобы разместить узлы, делящие между собой много ребер, близких друг к другу. Оказывается, это можно сделать при помощи второго наименьшего собственного значения лапласовой матрицы и соответствующего ей собственного вектора, важность которого настолько высока, что он имеет свое собственное имя: *вектор Фидлера*<sup>1</sup>.

Для иллюстрации давайте воспользуемся минимальной сетью. Мы начнем с создания матрицы смежности:

```
import numpy as np
A = np.array([[0, 1, 1, 0, 0, 0],
              [1, 0, 1, 0, 0, 0],
              [1, 1, 0, 1, 0, 0],
              [0, 0, 1, 0, 1, 1],
              [0, 0, 0, 1, 0, 1],
              [0, 0, 0, 1, 1, 0]], dtype=float)
```

Для изображения сети можно воспользоваться библиотекой NetworkX. Сначала, как обычно, мы инициализируем Matplotlib:

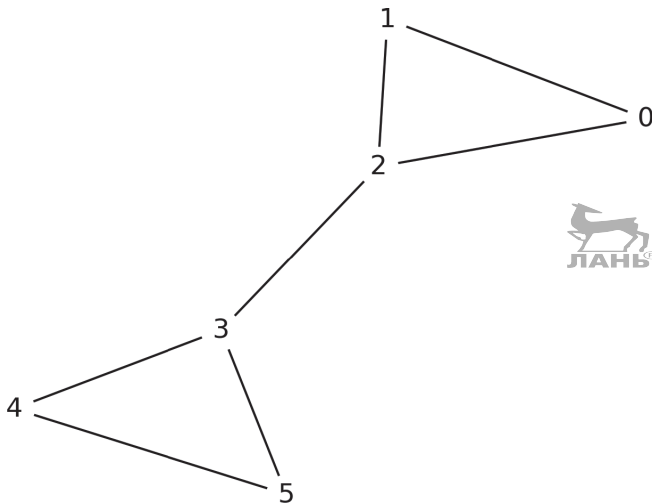
<sup>1</sup> См. [https://en.wikipedia.org/wiki/Algebraic\\_connectivity#The\\_Fiedler\\_vector](https://en.wikipedia.org/wiki/Algebraic_connectivity#The_Fiedler_vector) или [https://ru.wikipedia.org/wiki/Алгебраическая\\_связность#Вектор\\_Фидлера](https://ru.wikipedia.org/wiki/Алгебраическая_связность#Вектор_Фидлера).



```
# Заставить графики появляться локально, задать индивидуальный стиль графиков
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('style/elegant.mplstyle')
```

Теперь можно изобразить сеть:

```
import networkx as nx
g = nx.from_numpy_matrix(A)
layout = nx.spring_layout(g, pos=nx.circular_layout(g))
nx.draw(g, pos=layout,
        with_labels=True, node_color='white')
```



Вы видите, что узлы естественным образом попадают в две группы: 0, 1, 2 и 3, 4, 5. Сможет ли вектор Фидлера нам об этом сообщить? Прежде всего мы должны вычислить степенную матрицу и лапласову матрицу. Сначала мы получаем степени, выполняя суммирование вдоль любой оси  $A$ . (Подойдет любая из осей, потому что матрица  $A$  симметрична.)

```
d = np.sum(A, axis=0)
print(d)
[ 2.  2.  3.  3.  2.  2.]
```

Затем помещаем эти степени в диагональную матрицу такой же формы, что и  $A$ , в *степенную матрицу*. Для этого можно применить функцию `np.diag`:

```
D = np.diag(d)
print(D)
```

```
[[ 2. 0. 0. 0. 0. 0.]
 [ 0. 2. 0. 0. 0. 0.]
 [ 0. 0. 3. 0. 0. 0.]
 [ 0. 0. 0. 3. 0. 0.]
 [ 0. 0. 0. 0. 2. 0.]
 [ 0. 0. 0. 0. 0. 2.]]
```



Наконец, из этого определения мы получаем лапласову матрицу:

```
L = D - A
print(L)
```

```
[[ 2. -1. -1. 0. 0. 0.]
 [-1. 2. -1. 0. 0. 0.]
 [-1. -1. 3. -1. 0. 0.]
 [ 0. 0. -1. 3. -1. -1.]
 [ 0. 0. 0. -1. 2. -1.]
 [ 0. 0. 0. -1. -1. 2.]]
```

Поскольку матрица  $L$  симметрична, для вычисления собственных значений и собственных векторов мы можем применить функцию `np.linalg.eigh`:

```
val, Vec = np.linalg.eigh(L)
```

Вы можете проверить, что возвращенные значения удовлетворяют определению собственных значений и собственных векторов. Например, одно из собственных значений равняется 3:

```
np.any(np.isclose(val, 3))
```

```
True
```

И мы можем проверить, что умножение матрицы  $L$  на соответствующий собственный вектор действительно умножает вектор на 3:

```
idx_lambda3 = np.argmin(np.abs(val - 3))
v3 = Vec[:, idx_lambda3]
```

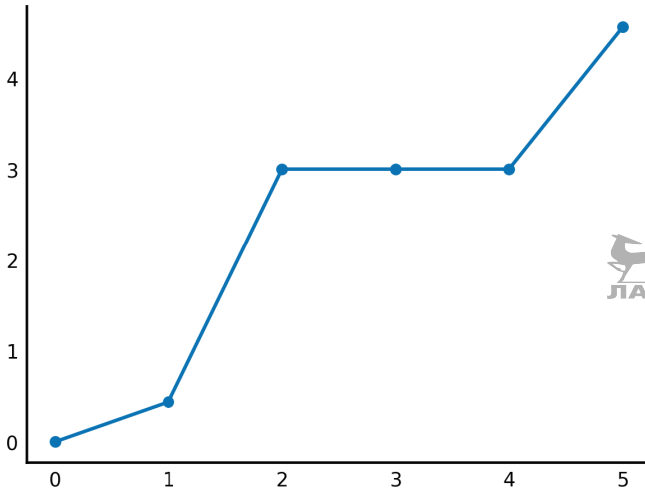
```
print(v3)
print(L @ v3)
```

```
[ 0.          0.37796447 -0.37796447 -0.37796447  0.68898224 -0.31101776]
[ 0.          1.13389342 -1.13389342 -1.13389342  2.06694671 -0.93305329]
```

Как было отмечено ранее, вектор Фидлера – это вектор, который соответствует второму наименьшему собственному значению матрицы  $L$ . Сортировка собственных значений сообщает, который из них является вторым наименьшим:



```
plt.plot(np.sort(val), linestyle='-', marker='o');
```



Это первое ненулевое собственное значение, находящееся рядом с 0.4. Вектор Фидлера является соответствующим собственным вектором (см. рис. 6.2):

```
f = Vec[:, np.argsort(val)[1]]
plt.plot(f, linestyle='-', marker='o');
```

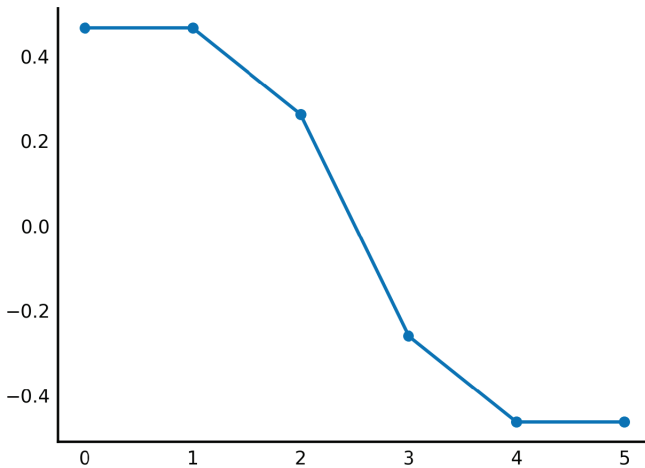


Рис. 6.2 ❖ Вектор Фидлера матрицы L

Поразительно! Глядя на *знак* элементов вектора Фидлера, мы можем разделить узлы на две группы, которые были определены на рисунке (см. рис. 6.3)!

```
colors = ['orange' if eigv > 0 else 'gray' for eigv in f]
nx.draw(g, pos=layout, with_labels=True, node_color=colors)
```

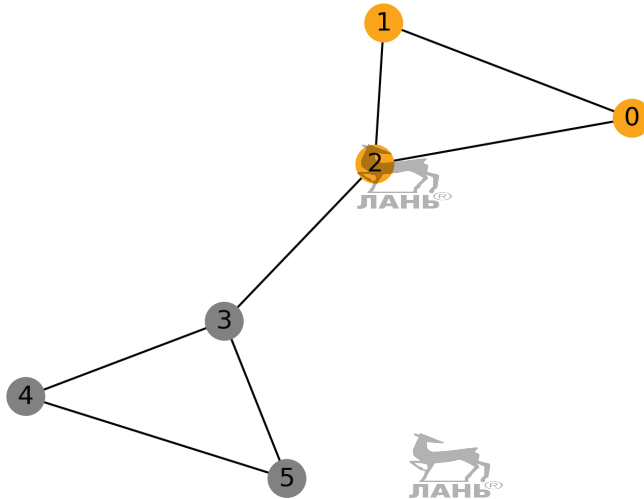


Рис. 6.3 ❖ Узлы, окрашенные по их знаку в векторе Фидлера матрицы  $L$

## ЛАПЛАСОВЫ МАТРИЦЫ С ДАННЫМИ О МОЗГЕ

Продемонстрируем этот процесс на реальном примере, расположив клетки головного мозга червя так, как показано на рис. 2 исследовательской работы Варшни и др.<sup>1</sup> С этой работой мы познакомили вас в главе 3. (Информация о том, как это сделать, находится в дополнительном материале, прилагаемом к указанной работе<sup>2</sup>.) Чтобы получить полученное исследователями расположение нейронов мозга червя, была применена сходная матрица, называемая *лапласовой матрицей, нормализованной по степени*.

Поскольку в этом анализе важен порядок следования нейронов, чтобы не загромождать эту главу очисткой данных, мы используем предварительно обработанный набор данных. Исходные данные были получены на веб-сайте Лэва Варшни<sup>3</sup>, а обработанные данные находятся в нашем каталоге *data/*.

Сначала загрузим данные. Есть четыре компонента:

- сеть химических синапсов, через которые *предсинаптический нейрон* посылает химический сигнал в *постсинаптический нейрон*;
- сеть щелевых контактов, содержащая прямые электрические контакты между нейронами;

<sup>1</sup> См. <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1001066>.

<sup>2</sup> См. <http://journals.plos.org/ploscompbiol/article/file?id=info:doi/10.1371/journal.pcbi.1001066.s001&type=supplementary>.

<sup>3</sup> См. <http://www.ifp.illinois.edu/~varshney/elegans>.

- идентификаторы нейронов (имена);
- три типа нейронов:
  - *сенсорные (рецепторные) нейроны* – это нейроны, которые обнаруживают сигналы, поступающие из внешней среды; закодированы как 0;
  - *моторные (эффекторные) нейроны* – это нейроны, активизирующие мышцы, позволяя червю перемещаться; закодированы как 2;
  - *промежуточные (вставочные) нейроны* – это нейроны, выполняющие сложную обработку сигналов между сенсорными и моторными нейронами; закодированы как 1.

```
import numpy as np

Chem = np.load('data/chem-network.npy')
Gap = np.load('data/gap-network.npy')
neuron_ids = np.load('data/neurons.npy')
neuron_types = np.load('data/neuron-types.npy')
```

Затем мы упрощаем сеть, т. е. складываем два вида связей и удаляем из сети направленность, беря среднее значение входящих и исходящих связей нейронов. Это слегка напоминает обман, но, поскольку мы стремимся получить только расположение нейронов в графе, нас интересует не направленность связи между нейронами, а факт наличия этой связи. Мы называем результирующую матрицу матрицей *связности*,  $C$ , являющейся другой разновидностью матрицы смежности.

```
A = Chem + Gap
C = (A + A.T) / 2
```

Чтобы получить лапласову матрицу  $L$ , нам потребуется матрица степени  $D$ , которая содержит степень узла  $i$  в позиции  $[i, i]$  и нули в остальных позициях.

```
degrees = np.sum(C, axis=0)
D = np.diag(degrees)
```

Теперь можно получить лапласову матрицу:

```
L = D - C
```

Вертикальные координаты на рис. 2 из исследовательской работы получены путем расположения узлов так, что в среднем нейроны находятся максимально близко и «чуть выше» своих нисходящих соседей. Варшни и др. называют эту меру «глубиной обработки» (processing depth, вычислительной глубиной). Она получена в результате решения линейного уравнения с участием лапласовой матрицы. Для его решения мы используем псевдоинверсию<sup>1</sup>, т. е. функцию `scipy.linalg.pinv`:

```
from scipy import linalg
b = np.sum(C * np.sign(A - A.T), axis=1)
z = linalg.pinv(L) @ b
```

<sup>1</sup> См. [https://en.wikipedia.org/wiki/Moore%E2%80%93Penrose\\_inverse](https://en.wikipedia.org/wiki/Moore%E2%80%93Penrose_inverse).

(Обратите внимание на использование символа @, который был введен в Python 3.5 для обозначения операции умножения матриц. Как мы отмечали в предисловии и в главе 5, в предыдущих версиях Python использовалась функция `np.dot()`.)

Чтобы получить нормализованную по степени лапласову матрицу,  $Q$ , нам потребуется обратный квадратный корень матрицы  $D$ :

```
Dinv2 = np.diag(1 / np.sqrt(degrees))
Q = Dinv2 @ L @ Dinv2
```

Наконец, мы можем извлечь координаты  $x$  нейронов, тем самым гарантируя, что сильно связанные нейроны остаются рядом: собственный вектор  $Q$ , соответствующий своему второму наименьшему собственному значению, нормализованному по степени:

```
val, Vec = linalg.eig(Q)
```

Обратите внимание: согласно документации по функции `numpy.linalg.eig`:

Собственные значения не обязательно упорядочены.

Хотя документация SciPy по функции `eig` такого предупреждения не содержит, в данном случае оно остается верным. Следовательно, мы должны самостоятельно отсортировать собственные значения и соответствующие им столбцы собственных векторов:

```
smallest_first = np.argsort(val) # наименьший первый
val = val[smallest_first]
Vec = Vec[:, smallest_first]
```

Теперь найдем собственный вектор, требующийся для вычисления аффинных координат (координат аффинного подобия):

```
x = Dinv2 @ Vec[:, 1]
```

(Объяснение причин использования этого вектора потребует слишком много времени и места. Эти объяснения можно найти в дополнительном материале, прилагаемом к исследовательской работе, ссылка на которую приведена выше. Если коротко, то выбор такого вектора позволяет минимизировать общую длину связей между нейронами.)

Есть одна небольшая сложность, которую мы должны решить, перед тем как пойти дальше: собственные векторы определены только до мультипликативной константы. Это следует из определения собственного вектора. Предположим, что  $v$  – это собственный вектор матрицы  $M$  с соответствующим собственным значением  $\lambda$ . Тогда  $\alpha v$  – это тоже собственный вектор матрицы  $M$  для любого скалярного числа  $\alpha$ , потому что  $Mv = \lambda v$  влечет  $M(\alpha v) = \lambda(\alpha v)$ . Поэтому, когда у программного пакета запрашиваются собственные векторы матрицы  $M$ , не важно, возвращает он  $v$  или  $-v$ . Чтобы убедиться, что мы воспроизводим то расположение узлов, которое было получено в работе Варшни и др., нам необходимо удостовериться, что вектор указывает в том же самом направлении,

что и у них, а не в противоположном. Это делается путем выбора произвольного нейрона из рис. 2 работы Варшни и др. и проверки знака переменной  $x$  в этой позиции. Далее, если знак не совпадает со своим знаком на рис. 2 исследовательской работы, мы меняем его на обратный.

```
vc2_index = np.argwhere(neuron_ids == 'VC02')
if x[vc2_index] < 0:
    x = -x
```

Теперь все дело сводится к отрисовке узлов и ребер. Мы окрашиваем их согласно типу, хранящемуся в `neuron_types`, используя симпатичную и функционально «нечувствительную к цвету» палитру `colorbrewer`<sup>1</sup>:

```
from matplotlib.colors import ListedColormap
from matplotlib.collections import LineCollection

def plot_connectome(x_coords, y_coords, conn_matrix, *,
                   labels=(), types=None, type_names=('',),
                   xlabel='', ylabel=''):
    """Вывести нейроны в виде точек, соединенных линиями.
    Нейроны могут иметь разные типы (до 6 разных цветов).

    Параметры
    -----
    x_coords, y_coords : массив вещественных, форма (N,)
        Координаты x и y нейронов.
    conn_matrix : массив или разреженная матрица вещественных, форма (N, N)
        Матрица связности с ненулевыми записями (i, j), если и только если
        узел i и узел j связаны.
    labels : массивоподобный с типом string, форма (N,), необязательный
        Имена узлов.
    types : массив целочисленных, форма (N,), необязательный
        Тип (т. е. сенсорный нейрон, промежуточный нейрон) каждого узла.
    type_names : массивоподобный с типом string, необязательный
        Имя каждого значения параметра `types`. Например, если a 0 в
        `types` означает «сенсорный нейрон», то `type_names[0]` должен
        быть «сенсорный нейрон».
    xlabel, ylabel : строковый, необязательный
        Метки для осей.

    «»»
    if types is None:
        types = np.zeros(x_coords.shape, dtype=int)
        ntypes = len(np.unique(types))
        colors = plt.rcParams['axes.prop_cycle'][:ntypes].by_key()['color']
        cmap = ListedColormap(colors)

    fig, ax = plt.subplots()

    # Вывести на график позиции нейронов:
    for neuron_type in range(ntypes):
        plotting = (types == neuron_type)
```

<sup>1</sup> См. [http://chrisalbon.com/python/seaborn\\_color\\_palettes.html](http://chrisalbon.com/python/seaborn_color_palettes.html).

```

pts = ax.scatter(x_coords[plotting], y_coords[plotting],
                c=cmap(neuron_type), s=4, zorder=1)
pts.set_label(type_names[neuron_type])

# Добавить текстовые метки:
for x, y, label in zip(x_coords, y_coords, labels):
    ax.text(x, y, ' ' + label,
            verticalalignment='center', fontsize=3, zorder=2)

# Вывести ребра
pre, post = np.nonzero(conn_matrix)
links = np.array([[x_coords[pre], x_coords[post]],
                  [y_coords[pre], y_coords[post]]]).T
ax.add_collection(LineCollection(links, color='lightgray',
                                lw=0.3, alpha=0.5, zorder=0))

ax.legend(scatterpoints=3, fontsize=6)

ax.set_xlabel(xlabel, fontsize=8)
ax.set_ylabel(ylabel, fontsize=8)

plt.show()

```

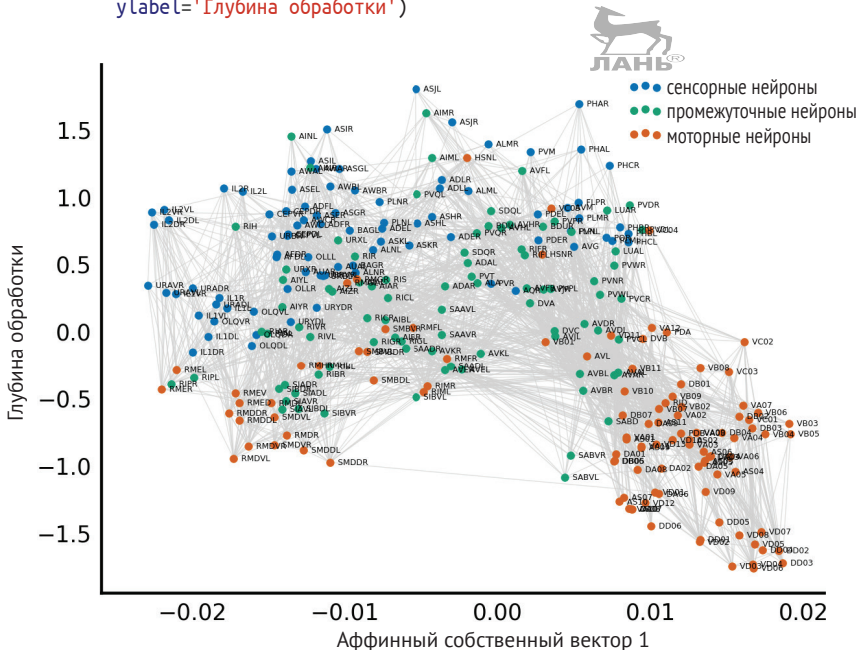


Теперь, чтобы вывести нейроны на график, применим эту функцию:

```

plot_connectome(x, z, C, labels=neuron_ids, types=neuron_types,
                type_names=['сенсорные нейроны',
                             'промежуточные нейроны',
                             'моторные нейроны'],
                xlabel='Аффинный собственный вектор 1',
                ylabel='Глубина обработки')

```



Извольте получить: мозг червя! Как отмечалось в оригинальной исследовательской работе, вы видите нисходящую обработку от сенсорных нейронов до моторных нейронов через сеть промежуточных нейронов. Здесь показаны две разные группы моторных нейронов: соответствующие шейному (слева) и телесному (справа) сегментам червя.

### Задача: изображение аффинного подобия

Как видоизменить приведенный выше программный код, чтобы показать аффинное подобие, показанное на рис. 2В исследовательской работы?

### Задача: линейная алгебра с разреженными матрицами

В приведенном выше программном коде массивы NumPy используются для хранения матриц и выполнения необходимых вычислений. Это оправдано, потому что мы используем небольшой граф, состоящий из менее чем 300 узлов. Однако в случае более крупных графов этот подход потерпит неудачу.

Например, можно было бы проанализировать связи между библиотеками, перечисленными в Каталоге пакетов Python, или PyPI, который содержит более 10 тысяч пакетов. Хранение лапласовой матрицы для этого графа заняло бы  $8(100 \times 10^3)^2 = 8 \times 10^{10}$  байт, или 80 Гб ОЗУ. Если к этому добавить матрицы смежности, симметрической смежности, псевдоинверсии и, скажем, две временные матрицы, используемые во время вычислений, то в результате вы получите 480 Гб, что будет вне досягаемости большинства настольных компьютеров.

Многие из вас могут воскликнуть: «Ха! Да, в моем настольном компьютере ОЗУ 512 Гб! Ему ничего не стоит вмиг справиться с этим так называемым “большим” графом!»

Может быть. Но вы, возможно, также захотите проанализировать цитатный граф Ассоциации вычислительной техники (Association for Computing Machinery, ACM), а это сеть, состоящая из более чем двух миллионов научных работ и ссылок. Такая лапласова матрица заняла бы ОЗУ объемом 32 терабайта.

Однако нам известно, что графы зависимостей и ссылок являются разреженными: программные пакеты обычно зависят лишь от нескольких других пакетов, а не всего каталога PyPI в целом. А научные работы и книги обычно ссылаются лишь на некоторые другие. Поэтому мы можем хранить вышеупомянутые матрицы, используя разреженные структуры данных из модуля `scipy.sparse` (см. главу 5), и для вычисления требующихся для нас значений применять линейно-алгебраические функции из подмодуля `scipy.sparse.linalg`.

Попробуйте исследовать документацию по `scipy.sparse.linalg`, чтобы разработать разреженную версию вышеупомянутого вычисления.



Псевдоинверсия разреженной матрицы в целом не является разреженной, поэтому она здесь не применима. Аналогичным образом вы не сможете получить все собственные векторы разреженной матрицы, потому что все вместе они образуют плотную матрицу.

Вы найдете части решения ниже (и, разумеется, в приложении в конце книги), однако мы настоятельно рекомендуем попробовать вам это сделать самостоятельно.

### Решатели

SciPy располагает несколькими разреженными итеративными решателями, и не всегда понятно, какой из них применять. К сожалению, на этот вопрос тоже нет простого ответа: у разных алгоритмов есть много сильных сторон с точки зрения скорости сходимости, стабильности, точности и использования оперативной памяти (среди прочих). Кроме того, глядя на входные данные, невозможно предсказать, какой алгоритм покажет наилучшую результативность.

Вот грубый ориентир для выбора итеративного решателя:

- если входная матрица  $A$  симметрична и положительно-определенная, лучше использовать решатель сопряженных градиентов `cg`. Если  $A$  симметричная, но почти сингулярная или неопределенная, попробуйте итеративный метод минимальных остатков `minres`;
- для несимметричных систем воспользуйтесь стабилизированным методом бисопряженных градиентов `bicgstab`. Квадратичный метод сопряженных градиентов, `cgs`, выполняется немного быстрее, но имеет более неустойчивую сходимость;
- если решается много аналогичных систем, используйте LGMRES-алгоритм `lgmres`;
- если  $A$  не является квадратной, воспользуйтесь алгоритмом наименьших квадратов `lsqr`.

А вот дополнительные материалы для чтения:

- *Noël M. Nachtigal, Satish C. Reddy, and Lloyd N. Trefethen*. How Fast Are Nonsymmetric Matrix Iterations? (Какова скорость несимметричных матричных итераций) *SIAM Journal on Matrix Analysis and Applications* 13, no. 3 (1992): 778–795;
- *Донгарра Дж.* Обзор новейших методов Крылова<sup>1</sup> (*Jack Dongarra*. Survey of Recent Krylov Methods. November 20, 1995).

## PAGERANK: ЛИНЕЙНАЯ АЛГЕБРА ДЛЯ РЕПУТАЦИИ И ВАЖНОСТИ

Еще одно применение линейной алгебры и собственных векторов представлено алгоритмом компании Google PageRank для обозначения веб-страниц, остроумно названным по имени одного из соучредителей компании, Ларри Пейджа.

Чтобы ранжировать веб-страницы по важности, вы можете подсчитать, сколько других веб-страниц на нее ссылаются. В конце концов, если каждая страница связана с конкретной страницей, это хорошо, разве не так? Но этот метрический показатель легко обойти: для повышения ранга вашей собственной веб-страницы просто создайте столько других веб-страниц, сколько сможете. И пусть они ссылаются на вашу исходную страницу.

Ключевой вывод, стимулировавший ранний успех Google, состоял в том, что на важные веб-страницы ссылаются не просто многие веб-страницы, а *важные*

<sup>1</sup> См. [http://www.netlib.org/linalg/html\\_templates/node50.html](http://www.netlib.org/linalg/html_templates/node50.html).



веб-страницы. Но как узнать, какие другие страницы важные? Ведь на них самих ссылаются другие важные страницы. И т. д.

Из этого рекурсивного определения следует, что важность страницы может быть измерена собственным вектором так называемой *матрицы переходов*, содержащей связи между веб-страницами. Предположим, у вас есть свой вектор важности  $r$  и своя матрица связей  $M$ . Вы еще не знаете  $r$ , но знаете, что важность страницы пропорциональна сумме важностей страниц, которые на нее ссылаются:  $r = \alpha Mr$ , или  $Mr = \lambda r$ , для  $\lambda = 1/\alpha$ . Но ведь это и есть определение собственного значения!

Добившись, что матрица переходов удовлетворяет некоторым особым свойствам, далее мы можем определить, что необходимое собственное значение равняется 1 и что оно является наибольшим собственным значением  $M$ .

Матрица переходов обозначает интернет-пользователя, нередко именуемого Вэбстером, который беспорядочно нажимает на ссылку на каждой посещаемой им веб-странице и затем задается вопросом, какую вероятность он в итоге получит на той или иной заданной странице. Эта вероятность называется *рангом* страницы, или PageRank.

С началом роста популярности Google исследователи стали применять меру PageRank ко всем видам сетей. Мы будем использовать пример Стефано Аллезина и Мерседес Паскуаль (Stefano Allesina и Mercedes Pascual), опубликованный<sup>1</sup> в журнале «*PLoS Вычислительная биология*». Они намеревались применить этот метод в экологических *пищевых сетях*, т. е. сетях, в которых организмы связаны с теми организмами, которыми они питаются.

В простейшем плане, если вам интересно, насколько важным для экосистемы является какой-либо организм, то вы посмотрите, сколько организмов им питаются. Если их много и этот организм исчез, то все «зависимые» от него организмы могут исчезнуть вместе с ним. На языке сетей можно сказать, что его *полустепень захода* определяет его экологическую важность.

Может ли PageRank быть оптимальной мерой важности для экосистемы?

Профессор Аллезина любезно предоставил нам несколько пищевых сетей для экспериментирования. Мы сохранили одну из них в формате языка разметки графов. Эта сеть получена в Природном заповеднике Сент-Марка во Флориде. Данная сеть была описана<sup>2</sup> в 1999 г. Робертом Р. Кристианом и Джозефом Й. Луцзовичем (Robert R. Christian и Joseph J. Luczovich). В наборе данных узел  $i$  имеет ребро в узел  $j$ , если организм  $i$  питается организмом  $j$ .

Начнем с загрузки данных, которые библиотека NetworkX читает тривиальным образом:

```
import networkx as nx
stmarks = nx.read_gml('data/stmarks.gml')
```

<sup>1</sup> См. <https://doi.org/10.1371/journal.pcbi.1000494>.

<sup>2</sup> См. <https://www.sciencedirect.com/science/article/pii/S0304380099000228>.



Затем получим разреженную матрицу, которая соответствует графу. Поскольку матрица содержит только числовую информацию, мы должны поддерживать отдельный список имен пакетов, соответствующих строкам/столбцам матрицы:

```
species = np.array(stmarks.nodes()) # массив для мультииндексации
Adj = nx.to_scipy_sparse_matrix(stmarks, dtype=np.float64)
```

Из матрицы смежности мы можем вывести матрицу *переходных вероятностей*, в которой каждое ребро заменяется *вероятностью* 1 на количестве ребер, исходящих из этого организма. В пищевой сети целесообразнее назвать эту матрицу матрицей обеденных вероятностей.

Общее количество организмов в нашей матрице будет использоваться неоднократно, поэтому давайте назовем его  $n$ :

```
n = len(species)
```

Затем нам понадобятся степени и, в частности, диагональная матрица, содержащая инверсию полустепеней исхода каждого узла на диагонали:

```
np.seterr(divide='ignore') # игнорировать ошибки деления на ноль
from scipy import sparse
degrees = np.ravel(Adj.sum(axis=1))
DegInv = sparse.diags(1 / degrees).tocsr()
Trans = (DegInv @ Adj).T
```

Как правило, мера PageRank представлена первым собственным вектором матрицы переходов. Если мы назовем матрицу переходов  $M$  и вектор PageRank-значений  $r$ , получим:

$$r = Mr.$$

Однако решение с вызовом функции `np.seterr` не такое простое. Метод на основе меры PageRank работает в случае, если матрица переходов является *по-столбцовой стохастической* матрицей, где каждый столбец в сумме составляет 1. Кроме того, каждая страница должна быть достижима из любой другой страницы, даже если путь до ее достижения очень длинный.

В нашей пищевой сети это вызывает проблемы, потому что основание пищевой цепи, т. е. то, что авторы называют органическим детритом (в основном это донные осадки), фактически ничем не питается (несмотря на круговорот жизни), и поэтому из него невозможно достигнуть других организмов.

*Молодой Симба:* Но, папа, разве мы не питаемся антилопой?

*Муфаса:* Да, Симба, только дай-ка я тебе объясню. Когда мы умираем, наши тела становятся травой, и антилопы едят траву. И поэтому мы все связаны между собой в большом круговороте жизни.

– Король-лев

Для решения этой проблемы в алгоритме PageRank используется так называемый «фактор затухания» (или демпфирования), обычно принимаемый

в размере 0.85. Этот фактор означает, что 85% времени алгоритм переходит по случайно отбираемой ссылке, но для остальных 15% он беспорядочно перескакивает на любую произвольную страницу. Так, если бы каждая страница имела низковоероятностную ссылку на любую страницу. Или если бы креветка изредка питалась акулами. Это может показаться бессмысленным, но послушайте! Ведь это вообще-то математическое представление круговорота жизни. Мы назначим фактору затухания величину 0.85, но для анализа его величина совсем не имеет значения: результаты будут аналогичными для большого диапазона возможных факторов затухания.

Если мы назовем фактор затухания  $d$ , то видоизмененное уравнение Page-Rank будет таким:

$$r = dMr + \frac{1-d}{n} \mathbf{1},$$

и:

$$(I - dM)r = \frac{1-d}{n} \mathbf{1}.$$

Мы можем решить это уравнение, используя прямой решатель `spsolve` подмодуля `scipy.sparse.linalg`. Правда, в зависимости от структуры и размера линейно-алгебраической задачи применение итеративного решателя может оказаться эффективнее. По этому поводу обратитесь к документации<sup>1</sup> по `scipy.sparse.linalg` для получения более подробной информации.

```
from scipy.sparse.linalg import spsolve
```

```
damping = 0.85
beta = 1 - damping
```

```
I = sparse.eye(n, format='csc') # Такой же разреженный формат, что и Trans
pagerank = spsolve(I - damping * Trans,
                   np.full(n, beta / n))
```

Теперь у нас есть «пищевой ранг» пищевой сети заповедника Сент-Марк!

Итак, каким образом пищевой ранг организма соотносится с рядом других организмов, которые им питаются?

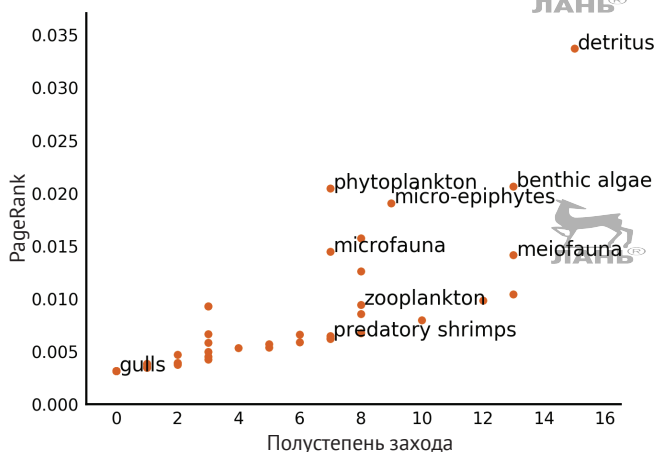
```
def pagerank_plot(in_degrees, pageranks, names, *,
                 annotations=[], **figkwargs):
    """Построит график рангов узлов в сопоставлении
    с полустепенью захода с отобранными вручную именами узлов.»»
    fig, ax = plt.subplots(**figkwargs)
    ax.scatter(in_degrees, pageranks, c=[0.835, 0.369, 0], lw=0)
    for name, indeg, pr in zip(names, in_degrees, pageranks):
        if name in annotations:
            text = ax.text(indeg + 0.1, pr, name)
    ax.set_ylim(0, np.max(pageranks) * 1.1)
```

<sup>1</sup> См. <https://docs.scipy.org/doc/scipy/reference/sparse.linalg.html#solving-linear-problems>.

```
ax.set_xlim(-1, np.max(in_degrees) * 1.1)
ax.set_ylabel('PageRank')
ax.set_xlabel('Полустепень захода')
```

Теперь построим график. Предварительно изучив набор данных, мы заранее поместили в графике некоторые интересные узлы:

```
interesting = ['detritus', 'phytoplankton', 'benthic algae', 'micro-epiphytes',
              'microfauna', 'zooplankton', 'predatory shrimps', 'meiofauna',
              'gulls']
in_degrees = np.ravel(Adj.sum(axis=0))
pagerank_plot(in_degrees, pagerank, species, annotations=interesting)
```



Донные осадки (органический детрит, detritus) являются самым важным элементом как по количеству организмов, которые им питаются (15), так и по величине меры PageRank ( $> 0.003$ ). Оданко вторым по важности элементом являются не донные водоросли (benthic algae), которые кормят 13 других организмов, а фитопланктон (phytoplankton), который кормит всего 7! Все потому, что им питаются другие *важные* организмы. Слева внизу у нас морские чайки (gulls), которые, как теперь можно подтвердить, являются для экосистемы настоящими бездельниками. Злобные *хищные креветки* (predatory shrimps) (мы это не выдумываем) поддерживают то же самое количество организмов, что и фитопланктон, но они являются менее существенными организмами, и поэтому они в итоге получают более низкий пищевой ранг.

Хотя мы это здесь не показываем, Аллесина и Паскуаль продолжили моделировать влияние исчезновения организмов на экологию и действительно обнаружили, что мера PageRank предсказывает экологическую важность лучше, чем полустепени захода.

Прежде чем закончить, хотелось бы отметить, что мера PageRank может вычисляться несколькими разными способами. Один способ, комплементарный тому, что мы сделали выше, называется *степенным методом*, и он действитель-

но мощный! Этот метод вытекает из теоремы Фробениуса–Перрона<sup>1</sup>, которая, среди всего прочего, гласит, что стохастическая матрица имеет 1 в качестве собственного значения и что единица является ее наибольшим собственным значением. (Соответствующий собственный вектор является вектором мер PageRank.) Это означает, что всякий раз, когда мы умножаем любой вектор на  $M$ , его компонента, указывающая на этот главный собственный вектор, остается прежней, а *все другие компоненты сужаются* мультипликативным фактором. Вследствие, если мы неоднократно умножаем некий случайный начальный вектор на  $M$ , в конечном счете мы должны получить вектор мер PageRank!

SciPy делает это очень эффективно при помощи своего модуля `sparse` для работы с разреженными матрицами:

```
def power(Trans, damping=0.85, max_iter=10**5):
    n = Trans.shape[0]
    r0 = np.full(n, 1/n)
    r = r0
    for _iter_num in range(max_iter):
        rnext = damping * Trans @ r + (1 - damping) / n
        if np.allclose(rnext, r):
            break
        r = rnext
    return r
```



## Задача: обработка висячих узлов

Следует отметить, что в приведенной выше итеративной обработке матрица `Trans` не является постолбцово-стохастической матрицей. Поэтому при каждой итерации вектор  $r$  сужается. Чтобы сделать матрицу стохастической, мы должны заменить каждый нулевой столбец на столбец, где все элементы равняются  $1/n$ . Это слишком дорого. С другой стороны, вычисление итераций обходится дешевле. Каким образом модифицировать приведенный выше программный код, чтобы вектор  $r$  гарантированно оставался вектором вероятностей на всем протяжении?

## Задача: эквивалентность разных методов получения собственного вектора

Удостоверьтесь, что все три метода производят одинаковое ранжирование узлов. Функция `numpy.coef.coef` может оказаться для этого полезной.

## ЗАКЛЮЧИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

Область линейной алгебры слишком широка, чтобы ее можно было полноценно рассмотреть в главе книги. Но эта глава позволила нам увидеть ее силу и как язык Python и библиотеки NumPy и SciPy делают эти элегантные алгоритмы доступными.

<sup>1</sup> См. [https://en.wikipedia.org/wiki/Perron%E2%80%93Frobenius\\_theorem](https://en.wikipedia.org/wiki/Perron%E2%80%93Frobenius_theorem) и [https://ru.wikipedia.org/wiki/Теорема\\_Фробениуса\\_–\\_Перрона](https://ru.wikipedia.org/wiki/Теорема_Фробениуса_–_Перрона).

# Глава 7



## Оптимизация функций в SciPy

«Что новенького?» – интересный и уходящий вширь вечный вопрос, но если пользоваться исключительно только им, то он приводит лишь к бесконечному параду тривиальностей, формальностей и пыли завтрашнего дня. Напротив, я хотел бы задуматься над вопросом «Что лучше?», над вопросом, который врезается вглубь, а не вширь, над вопросом, ответы на который смывают все наносное вниз по течению.

– Роберт М. Пирсиг,  
«Дзен и искусство обслуживания мотоцикла»

С первого раза вы не повесите картину на стену ровно. Вы ее поправляете, отходите назад, оцениваете горизонтальное положение и повторяете все заново. Этот процесс называется процессом *оптимизации*: мы изменяем ориентацию картины до тех пор, пока она не будет удовлетворять нашему требованию, т. е. пока ориентация картины не составит нулевой угол с горизонтом.

В математике наше требование называется «*функцией стоимости*», а ориентация портрета – «параметром». В типичной задаче оптимизации мы изменяем параметры до тех пор, пока функция стоимости не будет минимизирована.

Например, рассмотрим смещенную параболу,  $f(x) = (x - 3)^2$ . Мы хотели бы найти значение  $x$ , минимизирующее функцию стоимости. Нам известно, что эта функция с параметром  $x$  имеет минимум в значении 3, так как можем вычислить производную, уравнять ее нулем и увидеть, что  $2(x - 3) = 0$  (т. е.  $x = 3$ ).

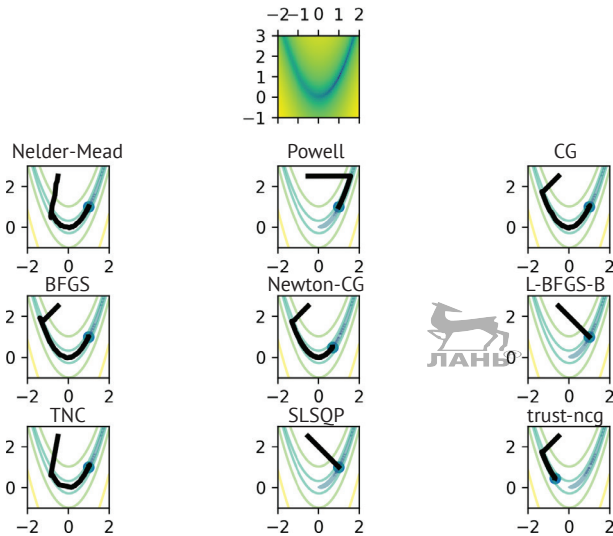
Но если бы эта функция была намного сложнее (например, если бы выражение имело много членов, многочисленные точки нулевой производной, содержало нелинейности или зависело от большого количества переменных), то ручное вычисление потребовало бы большого напряжения.

Функцию стоимости можно представить как рельеф местности, где мы пытаемся найти самую низкую точку. Такая аналогия непосредственно подчеркивает одну из сложных составляющих этой задачи: если вы находитесь в какой-то долине, окруженной горами, то как узнать, что вы находитесь в самой низкой долине, и не выглядит ли эта долина низкой, потому что она окружена очень

высокими горами? На языке оптимизации: как узнать, не застряли ли вы в *локальном минимуме*? Большинство алгоритмов оптимизации предпринимает ту или иную попытку решить эту проблему<sup>1</sup>.

➔ Рисунок 7.1 показывает все доступные в SciPy методы, некоторые из которых мы будем использовать, а другие оставим вам для исследования.

Существует целый ряд различных алгоритмов оптимизации (см. рис. 7.1). Вам предстоит решить, примет ли ваша функция стоимости на входе скаляр или вектор (т. е. имеется ли у вас один или несколько оптимизируемых параметров?). Существуют функции, которые требуют задание градиента функции стоимости, и функции, вычисляющие его автоматически. Некоторые функции отыскивают только параметры в заданной области (*оптимизация при заданных ограничениях*), а другие исследуют все параметрическое пространство.



**Рис. 7.1** ❖ Сравнение путей оптимизации, принимаемых разными алгоритмами оптимизации на функции Розенброка (вверху). Метод Пауэлла (Powell) выполняет линейный поиск вдоль первой размерности перед выполнением градиентного спуска. Метод сопряженных градиентов (CG), с другой стороны, выполняет градиентный спуск, начиная с отправной точки

<sup>1</sup> Алгоритмы оптимизации решают этот вопрос различными способами, но два общепринятых подхода представлены линейным (одномерным) поиском и доверительной областью. В случае с *линейным поиском* вы пытаетесь найти минимум функции стоимости вдоль определенной размерности и затем последовательно делаете ту же самую попытку вдоль остальных размерностей. В случае с *доверительными областями* мы перемещаем наше предположение относительно минимума в том направлении, в котором его ожидаем; если мы видим, что мы ожидаемо приближаемся к минимуму, то будем повторять эту процедуру с увеличенной уверенностью. В противном случае мы понижаем нашу уверенность и ищем более широкую область.

## ОПТИМИЗАЦИЯ В SciPy: SCIPY.OPTIMIZE

В оставшейся части настоящей главы мы для выравнивания двух изображений собираемся использовать модуль SciPy `optimize`. Применения выравнивания изображений, или *регистрации* изображений, включают сшивку панорамных снимков, совмещение разных сканов головного мозга, создание изображений со сверхвысокой разрешающей способностью и очистку объектов от зашумленности (шумоподавление) посредством комбинации многочисленных экспозиций в астрономии.

Как обычно, мы настраиваем нашу среду построения графиков:

```
# Заставить графики появляться локально, задать индивидуальный стиль графиков
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('style/elegant.mplstyle')
```

Начнем с самой простой версии задачи: у нас есть два изображения, одно смещено относительно другого. Требуется восстановить смещение, выравнивающее наилучшим образом наши изображения.

Наша функция оптимизации будет «сдвигать» одно из изображений и смотреть, уменьшает ли сдвиг в том или ином направлении их различие между собой. Выполнив это многократно, мы попробуем отыскать правильное выравнивание.

### Пример: вычисление оптимального сдвига изображения

Вы должны помнить нашего астронавта – Айлин Коллинз – из главы 3. Мы будем смещать это изображение на 50 пикселей вправо, затем сравнивать его с оригиналом, пока не найдем смещение, которое совпадает лучше всего. Очевидно, это глупость, так как мы знаем исходное положение. Но тем самым мы узнаем правду, и мы сможем проверить результативность нашего алгоритма. Вот оригинальное и смещенное изображения:

```
from skimage import data, color
from scipy import ndimage as ndi

astronaut = color.rgb2gray(data.astronaut())
shifted = ndi.shift(astronaut, (0, 50))

fig, axes = plt.subplots(nrows=1, ncols=2)
axes[0].imshow(astronaut)
axes[0].set_title('Оригинальное')
axes[1].imshow(shifted)
axes[1].set_title('Смещенное');
```





Чтобы алгоритм оптимизации выполнил свою работу, нам нужен некий способ определения «несхожести», т. е. функции стоимости. Самый легкий способ состоит в вычислении среднего квадрата разностей, который часто называется *среднеквадратической ошибкой*, или СКО (mean squared error, MSE).

```
import numpy as np
```

```
def mse(arr1, arr2):
```

```
    «»Вычислить среднеквадратическую ошибку между двумя массивами.«»
```

```
    return np.mean((arr1 - arr2)**2)
```



Эта функция вернет 0, когда оба изображения идеально выровнены, и более высокое значение в противном случае. При помощи функции стоимости мы можем проверить, выровнены два изображения или нет:

```
ncol = astronaut.shape[1]
```

```
# Покрывать расстояние в 90% от длины в столбцах,
```

```
# с одним значением в расчете на процентный пункт
```

```
shifts = np.linspace(-0.9 * ncol, 0.9 * ncol, 181)
```

```
mse_costs = []
```

```
for shift in shifts:
```

```
    shifted_back = ndi.shift(shifted, (0, shift))
```

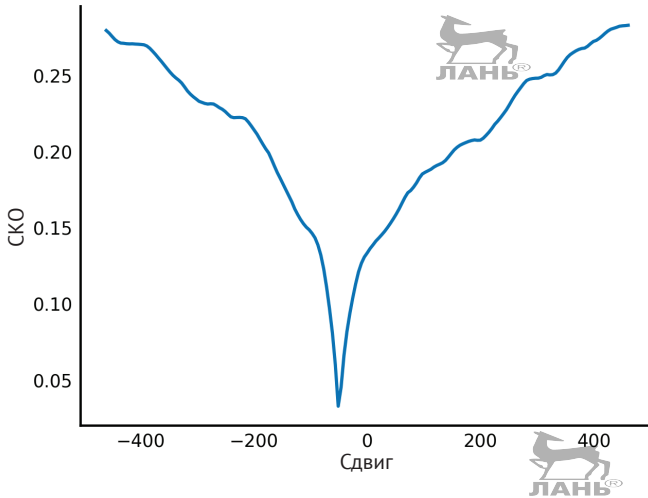
```
    mse_costs.append(mse(astronaut, shifted_back))
```

```
fig, ax = plt.subplots()
```

```
ax.plot(shifts, mse_costs)
```

```
ax.set_xlabel('Сдвиг')
```

```
ax.set_ylabel('СКО');
```



В случае если функция стоимости определена, мы можем запросить функцию `scipy.optimize.minimize`, чтобы отыскать оптимальные параметры:

```
from scipy import optimize

def astronaut_shift_error(shift, image):
    corrected = ndi.shift(image, (0, shift))
    return mse(astronaut, corrected)

res = optimize.minimize(astronaut_shift_error, 0, args=(shifted,),
                       method='Powell')

print(f'Оптимальный сдвиг для коррекции составляет: {res.x}')
```

Оптимальный сдвиг для коррекции составляет: -49.99997565757551

Сработало! Мы сместили его на +50 пикселей, и благодаря нашей мере СКО функция SciPy `optimize.minimize` выдала правильный размер сдвига (-50), при котором изображение возвращается в свое исходное состояние.

Правда, эта чрезвычайно легкая задача оптимизации приводит нас к главной трудности данного вида выравнивания: иногда мера СКО должна ухудшиться, чтобы потом улучшиться.

Давайте снова взглянем на смещение изображений, начав с неизмененного изображения:

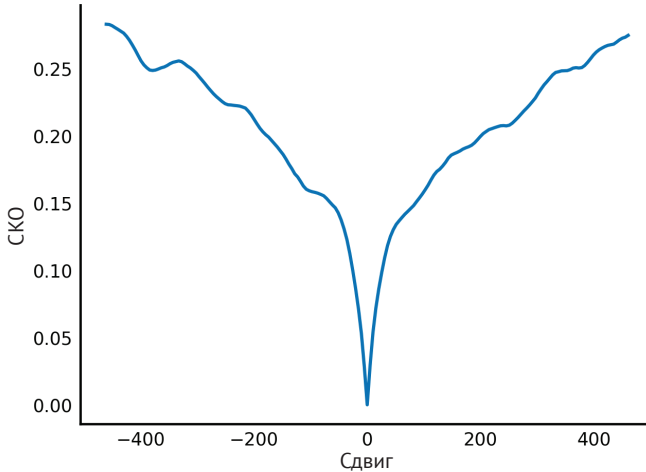
```
ncol = astronaut.shape[1]

# Покрыть расстояние в 90% от длины в столбцах,
# где одно значение приходится на один процентный пункт
shifts = np.linspace(-0.9 * ncol, 0.9 * ncol, 181)
mse_costs = []

for shift in shifts:
    shifted1 = ndi.shift(astronaut, (0, shift))
    mse_costs.append(mse(astronaut, shifted1))
```



```
fig, ax = plt.subplots()
ax.plot(shifts, mse_costs)
ax.set_xlabel('Сдвиг')
ax.set_ylabel('СКО');
```



Начиная с нулевого сдвига взгляните на значение СКО. Сначала сдвиг становится все более отрицательным. Он последовательно увеличивается приблизительно до  $-300$  пикселей сдвига, после чего снова начинает уменьшаться! Незначительно, но тем не менее он уменьшается. Мера СКО достигает нижнего предела приблизительно в  $-400$ , после чего снова увеличивается. Это называется *локальным минимумом*. Из-за того, что методы оптимизации имеют доступ только к «соседним» значениям функции стоимости, в ситуациях, когда функция улучшается, перемещаясь в «неправильном» направлении, процесс минимизации все равно будет двигаться в этом направлении. Поэтому если мы начинаем с изображения, сдвинутого на  $-340$  пикселей:

```
shifted2 = ndi.shift(astronaut, (0, -340))
```

функция `minimize` сдвинет его приблизительно на лишние 40 пикселей или около того, вместо того чтобы вернуть оригинальное изображение:

```
res = optimize.minimize(astronaut_shift_error, 0, args=(shifted2,),
                       method='Powell')
```

```
print(f'Оптимальный сдвиг для коррекции составляет {res.x}')
```

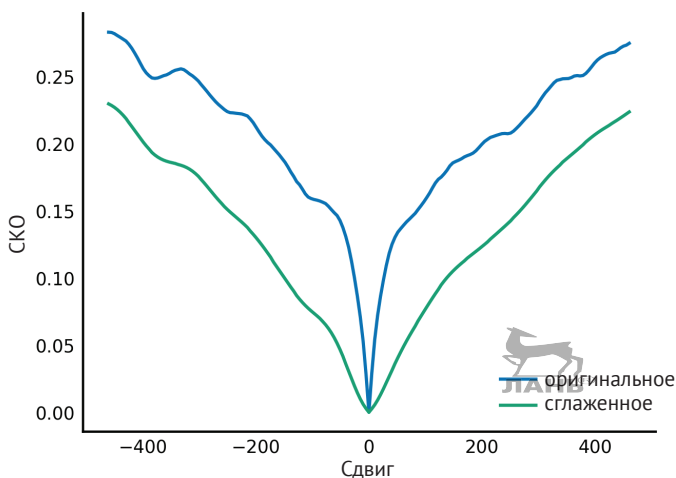
Оптимальный сдвиг для коррекции составляет  $-38.51778619397471$

Общепринятым решением этой задачи является сглаживание или понижение масштаба изображения. Это решение имеет побочный результат сглаживания целевой функции. Взгляните на тот же график после сглаживания изображения гауссовым фильтром:

```

from skimage import filters
astronaut_smooth = filters.gaussian(astronaut, sigma=20)
mse_costs_smooth = []
shifts = np.linspace(-0.9 * ncol, 0.9 * ncol, 181)
for shift in shifts:
    shifted3 = ndi.shift(astronaut_smooth, (0, shift))
    mse_costs_smooth.append(mse(astronaut_smooth, shifted3))

fig, ax = plt.subplots()
ax.plot(shifts, mse_costs, label='оригинальное')
ax.plot(shifts, mse_costs_smooth, label='сглаженное')
ax.legend(loc='lower right')
ax.set_xlabel('Сдвиг')
ax.set_ylabel('СКО');
    
```



Как вы видите, в случае экстремального сглаживания «полоса» функции ошибки становится шире и менее бугристой. Вместо того чтобы сглаживать саму функцию, мы можем получить аналогичный эффект путем размывки изображений перед их сравнением. Поэтому современное программное обеспечение выравнивания использует метод, называемый *гауссовой пирамидой*. Гауссова пирамида состоит из набора версий с пошагово уменьшающейся разрешающей способностью одного и того же изображения. Сначала мы выравниваем изображения с наиболее низкой разрешающей способностью (наиболее расплывчатые), чтобы получить приблизительное выравнивание. Затем последовательно уточняем выравнивание, основываясь на более четких изображениях.

```

def downsample2x(image):
    offsets = [(s + 1) % 2] / 2 for s in image.shape]
    slices = [slice(offset, end, 2)]
    
```

```

        for offset, end in zip(offsets, image.shape)]
    coords = np.mgrid[slices]
    return ndi.map_coordinates(image, coords, order=1)

```

```

def gaussian_pyramid(image, levels=6):
    «»Создать гауссову пирамиду изображения.

```

*Параметры*

-----

*image* : массив вещественных  
Входное изображение.

*max\_layer* : целочисленный, необязательный  
Количество уровней в пирамиде.

*Возвращает*

-----

*pyramid* : итератор массива вещественных  
Итератор уровней гауссовой пирамиды, начиная с верхнего  
уровня (наименьшей разрешающей способности).

.....

```
pyramid = [image]
```

```

for level in range(levels - 1):
    blurred = ndi.gaussian_filter(image, sigma=2/3)
    image = downsample2x(image)
    pyramid.append(image)

```

```
return reversed(pyramid)
```



Давайте проверим, как выглядит одномерное выравнивание вдоль этой пирамиды:

```

shifts = np.linspace(-0.9 * ncol, 0.9 * ncol, 181)
nlevels = 8
costs = np.empty((nlevels, len(shifts)), dtype=float)
astronaut_pyramid = list(gaussian_pyramid(astronaut, levels=nlevels))
for col, shift in enumerate(shifts):
    shifted = ndi.shift(astronaut, (0, shift))
    shifted_pyramid = gaussian_pyramid(shifted, levels=nlevels)
    for row, image in enumerate(shifted_pyramid):
        costs[row, col] = mse(astronaut_pyramid[row], image)

fig, ax = plt.subplots()
for level, cost in enumerate(costs):
    ax.plot(shifts, cost, label='Уровень %d' % (nlevels - level))
ax.legend(loc='lower right', frameon=True, framealpha=0.9)
ax.set_xlabel('Сдвиг')
ax.set_ylabel('СКО');

```

Как вы видите, на верхнем уровне пирамиды в сдвиге размером порядка –325 пикселей эта выбоина исчезает. Следовательно, на этом уровне мы можем получить приблизительное выравнивание, затем перейти к более низким уровням, чтобы уточнить полученное выравнивание (см. рис. 7.2).

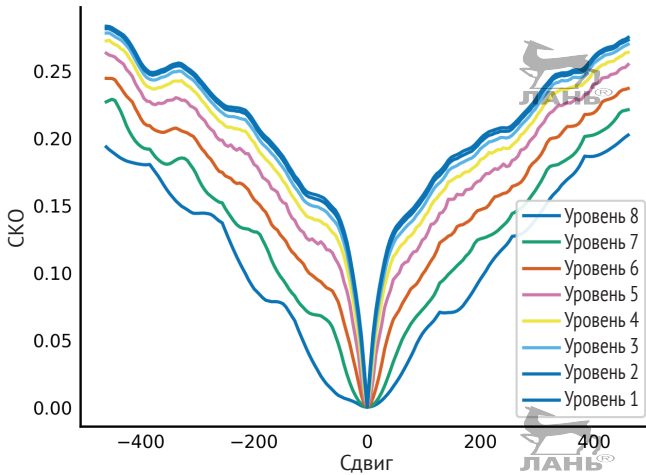


Рис. 7.2 ❖ Среднеквадратическая ошибка сдвига на разных уровнях гауссовой пирамиды

## РЕГИСТРАЦИЯ ИЗОБРАЖЕНИЯ ПРИ ПОМОЩИ OPTIMIZE

Давайте данный метод автоматизируем и попробуем выполнить «реальное» выравнивание с тремя параметрами: поворотом, трансляцией в размерность строк и трансляцией в размерность столбцов. Такое выравнивание называется «жесткой регистрацией», потому что отсутствуют какие-либо деформации (масштабирование, искажение или иное растягивание). Объект считается монолитным и перемещается в разные стороны (включая поворот) до тех пор, пока не будет найдено совпадение.

В целях упрощения исходного кода мы воспользуемся модулем библиотеки `scikit-image transform`, который будет вычислять сдвиг и поворот изображения. Модуль `SciPy optimize` предполагает на входе наличие вектора параметров. Мы сначала создадим функцию, принимающую такой вектор и производящую жесткое преобразование с правильными параметрами:

```
from skimage import transform

def make_rigid_transform(param):
    r, tc, tr = param
    return transform.SimilarityTransform(rotation=r,
                                         translation=(tc, tr))

rotated = transform.rotate(astronaut, 45)

fig, axes = plt.subplots(nrows=1, ncols=2)
axes[0].imshow(astronaut)
axes[0].set_title('Оригинальное')
axes[1].imshow(rotated)
axes[1].set_title('Повернутое');
```

Далее нам потребуется функция стоимости. Ею будет простая функция СКО, но SciPy требует определенного формата: первый аргумент должен быть *вектором параметров*, который она оптимизирует. Последующие аргументы могут быть переданы через ключевое слово `args` в виде кортежа, но должны оставаться фиксированными. Оптимизируется только вектор параметров. В нашем случае это просто угол поворота и два параметра трансляции:

```
def cost_mse(param, reference_image, target_image):
    transformation = make_rigid_transform(param)
    transformed = transform.warp(target_image, transformation, order=3)
    return mse(reference_image, transformed)
```



Наконец, мы напишем функцию выравнивания, которая, используя результат предыдущего уровня как отправную точку для следующего уровня, оптимизирует функцию стоимости *на каждом уровне гауссовой пирамиды*:

```
def align(reference, target, cost=cost_mse):
    nlevels = 7
    pyramid_ref = gaussian_pyramid(reference, levels=nlevels)
    pyramid_tgt = gaussian_pyramid(target, levels=nlevels)

    levels = range(nlevels, 0, -1)
    image_pairs = zip(pyramid_ref, pyramid_tgt)
    p = np.zeros(3)

    for n, (ref, tgt) in zip(levels, image_pairs):
        p[1:] *= 2

        res = optimize.minimize(cost, p, args=(ref, tgt), method='Powell')
        p = res.x

        # Печатаем текущий уровень, всякий раз перезаписывая информацию
        # (как индикатор выполнения)
        print(f'Уровень: {n}, Угол: {np.rad2deg(res.x[0]) :.3}, '
              f'Сдвиг: ({res.x[1] * 2**n :.3}, {res.x[2] * 2**n :.3}), '
              f'Стоимость: {res.fun :.3}', end='\r')

    print('\n') # новая строка, когда выравнивание завершено
    return make_rigid_transform(p)
```



Давайте проверим ее на нашем изображении астронавта. Будем его поворачивать на 60 градусов и добавлять немного шума. Сможет ли SciPy восстановить правильное преобразование? (См. рис. 7.3.)

```
from skimage import util

theta = 60
rotated = transform.rotate(astronaut, theta)
rotated = util.random_noise(rotated, mode='gaussian',
                             seed=0, mean=0, var=1e-3)

tf = align(astronaut, rotated)
corrected = transform.warp(rotated, tf, order=3)

f, (ax0, ax1, ax2) = plt.subplots(1, 3)
ax0.imshow(astronaut)
ax0.set_title('Оригинальное')
ax1.imshow(rotated)
ax1.set_title('Повернутое')
ax2.imshow(corrected)
ax2.set_title('Зарегистрированное')
for ax in (ax0, ax1, ax2):
    ax.axis('off')
```



Уровень: 1, Угол: -60.0, Сдвиг: (-1.87e+02, 6.98e+02), Cost: 0.0369



Рис. 7.3 ❖ Применение оптимизации для восстановления выравнивания изображения

Теперь мы чувствуем себя гораздо лучше. Однако наш выбор параметров фактически замаскировал трудность оптимизации. Давайте посмотрим, что произойдет после поворота на 50 градусов, который находится ближе к исходному изображению:

```
theta = 50
rotated = transform.rotate(astronaut, theta)
rotated = util.random_noise(rotated, mode='gaussian',
                             seed=0, mean=0, var=1e-3)

tf = align(astronaut, rotated)
corrected = transform.warp(rotated, tf, order=3)
```



```
f, (ax0, ax1, ax2) = plt.subplots(1, 3)
ax0.imshow(astronaut)
ax0.set_title('Оригинальное')
ax1.imshow(rotated)
ax1.set_title('Повернутое')
ax2.imshow(corrected)
ax2.set_title('Зарегистрированное')
for ax in (ax0, ax1, ax2):
    ax.axis('off')
```



Уровень: 1, Угол: 0.414, Сдвиг: (2.85, 38.4), Cost: 0.141

Несмотря на то что мы начали обработку ближе к исходному изображению, нам не удалось восстановить правильный поворот (рис. 7.4). Причина неудачи в том, что методы оптимизации могут застревать в локальных минимумах и небольших выбоинах. В этом мы уже убедились в случае с выравниванием только для сдвига. И как результат итог может быть довольно чувствительным по отношению к исходным параметрам.



Рис. 7.4 ❖ Неудавшаяся оптимизация

## ПРЕДОТВРАЩЕНИЕ ЛОКАЛЬНЫХ МИНИМУМОВ НА ОСНОВЕ АЛГОРИТМА BASIN HOPPING

Алгоритм 1997 года, разработанный Дэвидом Уэйлсом и Джонатаном Дойлом<sup>1</sup> под названием *basin hopping* (прыжки по котловине), пытается избежать локальных минимумов. Чтобы проскочить локальный минимум, следует попробовать оптимизацию с несколькими первоначальными параметрами. Затем уйти от найденного локального минимума в произвольном направлении

<sup>1</sup> Уэйлс Д. Дж., Дойл Дж. П. К. Глобальная оптимизация путем отскока от котловины и структуры кластеров с наименьшей энергией Леннарда–Джонса, содержащие до 110 атомов (David J. Wales, Jonathan P. K. Doyle. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms // Journal of Physical Chemistry 101, no. 28 (1997): 5111–5116).

и снова оптимизироваться. Выбрав соответствующий размер шага для случайных шагов, данный алгоритм может избежать повторного попадания в тот же самый локальный минимум и, по сравнению с простыми методами оптимизации на основе градиента, исследует гораздо большую площадь параметрического пространства.

В качестве упражнения мы предлагаем читателям включить реализацию алгоритма `basin hopping` средствами SciPy в нашу функцию выравнивания. Эта функция потребуется вам для последующих частей данной главы. Если же вы попадете в тупик, то смело можете подсмотреть решение в конце книги.

## Задача: модификация функции `align`

Попробуйте модифицировать функцию `align` для использования функции `scipy.optimize.basinhopping`, которая имеет четкие стратегии уклонения от локальных минимумов.

➔ Ограничьте использование алгоритма `basin hopping` только верхними уровнями пирамиды, т. к. этот метод оптимизации более медленный и может занять гораздо больше времени в случае работы с полной разрешающей способностью изображения.

## «Что лучше?»: ВЫБОР ПРАВИЛЬНОЙ ЦЕЛЕВОЙ ФУНКЦИИ

На данном этапе у нас есть рабочий подход к регистрации, дающий наилучшие результаты. Но оказывается, что мы решили лишь самую простую из регистрационных задач: выравнивание изображений той же самой *модальности*. Это значит, что мы ожидаем совпадения ярких пикселей в опорном изображении с яркими пикселями в тестовом изображении.

Теперь выровняем в этом изображении каналы цветности. Обратите внимание, здесь мы больше не можем опираться на каналы, имеющие ту же самую модальность. В данной задаче есть историческая ценность: между 1909 и 1915 г., до изобретения цветной фотографии, фотограф Сергей Михайлович Прокудин-Горский уже создавал цветные фотографии Российской империи. Фотографии создавались с помощью трех монохромных снимков, сделанных через помещенные перед линзой объектива три фильтра синего, красного и зеленого цвета.

В этом случае совместное выравнивание ярких пикселей, которое неявным образом выполняет СКО, работать не будет. Возьмем, к примеру, три снимка витража в церкви Святого Иоанна Богослова, взятые из коллекции Прокудина-Горского Библиотеки конгресса<sup>1</sup> (см. рис. 7.5):

```
from skimage import io
stained_glass = io.imread('data/00998v.jpg') / 255 # в [0, 1] использовать
# вещественное изображение
fig, ax = plt.subplots(figsize=(4.8, 7))
```

<sup>1</sup> См. <http://www.loc.gov/pictures/item/prk2000000263/>.

```
ax.imshow(stained_glass)  
ax.axis('off');
```



**Рис. 7.5** ❖ Пластина Прокудина-Горского:  
три фотографии одного и того же витража,  
снятые с использованием трех разных фильтров

Взгляните на одежды Святого Иоанна: на одном изображении они выглядят черными как смола, серыми на другом и ярко-белыми на третьем! Это привело бы к ужасной оценке СКО даже с прекрасным выравниванием.

Посмотрим, что можно с этим сделать. Начнем с разбиения пластины на составляющие ее каналы:

```
ngrows = stained_glass.shape[0]
step = ngrows // 3
channels = (stained_glass[:step],
            stained_glass[step:2*step],
            stained_glass[2*step:3*step])
channel_names = ['синий', 'зеленый', 'красный']
fig, axes = plt.subplots(1, 3)
for ax, image, name in zip(axes, channels, channel_names):
    ax.imshow(image)
    ax.axis('off')
    ax.set_title(name)
```

синий



зеленый



красный



Сначала наложим все три изображения друг на друга, чтобы убедиться в необходимости точной настройки выравнивания между этими тремя каналами:

```
blue, green, red = channels
original = np.dstack((red, green, blue))
fig, ax = plt.subplots(figsize=(4.8, 4.8), tight_layout=True)
ax.imshow(original)
ax.axis('off');
```



По наличию цветных «ореолов» вокруг объектов на изображении вы видите, что цвета близки к выравниванию, но не совсем. Давайте попробуем выровнять их так же, как мы выравнивали приведенное выше изображение астронавта, используя СКО. В качестве опорного изображения используем один канал зеленого цвета и по нему выравниваем синий и красный каналы.

```
print('*** Выравнивание синего с зеленым ***')
tf = align(green, blue)
cblue = transform.warp(blue, tf, order=3)

print('*** Выравнивание красного с зеленым ***')
tf = align(green, red)
cred = transform.warp(red, tf, order=3)

corrected = np.dstack((cred, green, cblue))
f, (ax0, ax1) = plt.subplots(1, 2)
ax0.imshow(original)
ax0.set_title('Оригинальный')
ax1.imshow(corrected)
ax1.set_title('Скорректированный')
for ax in (ax0, ax1):
    ax.axis('off')

*** Выравнивание синего с зеленым ***
Уровень: 1, Угол: -0.0474, Сдвиг: (-0.867, 15.4), Cost: 0.0499
** Выравнивание красного с зеленым ***
Уровень: 1, Угол: 0.0339, Сдвиг: (-0.269, -8.88), Cost: 0.0311
```

Выравнивание получилось немного лучше, чем с необработанными изображениями (рис. 7.6), потому что благодаря гигантскому желтому участку неба красный и зеленый каналы выровнены правильно. Однако синий канал по-прежнему смещен, и яркие пятна синего цвета не совпадают с зеленым каналом. Это означает, что СКО будет ниже, когда каналы *рассогласованы* из-за того, что синие участки частично накладываются на ярко-зеленые участки.

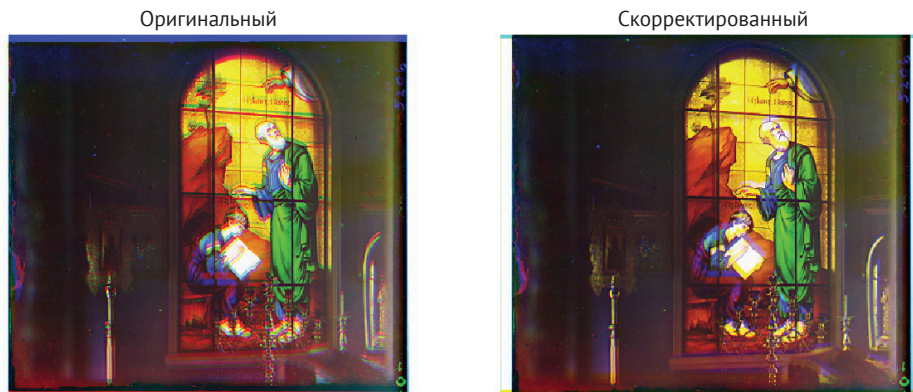


Рис. 7.6 ❖ Выравнивание на основе СКО сокращается, но не устраняет цветных ореолов

Теперь мы обратимся к мере под названием «*нормализованная взаимная информация*» (НВИ, *normalized mutual information*, NMI), которая измеряет корреляции между разными полосами яркости различных изображений. Когда изображения идеально выровнены, любой объект однородного цвета создаст большую корреляцию между оттенками различных компонентных каналов и соответствующим образом большую величину НВИ. В некотором смысле НВИ измеряет, насколько легко получается предсказать значение пиксела в одном изображении при наличии значения соответствующего пиксела в другом. Эта мера была определена в работе «Инвариантная к наложению мера энтропии выравнивания трехмерного медицинского изображения»<sup>1</sup>:

$$I(X, Y) = \frac{H(X) + H(Y)}{H(X, Y)},$$

где  $H(X)$  – это энтропия  $X$  и  $H(X, Y)$  – совместная энтропия  $X$  и  $Y$ . Числитель описывает энтропию двух изображений, рассматриваемых отдельно, а знаменатель – общую энтропию, если они наблюдаются вместе. Величины могут варьироваться между 1 (максимально выровнены) и  $\frac{1}{2}$  (минимально выровнены)<sup>2</sup>. См. главу 5.

В программном коде Python получится:

```
from scipy.stats import entropy
```

```
def normalized_mutual_information(A, B):
```

```
    «»»Вычислить нормализованную взаимную информацию.
```

```
    Нормализованная взаимная информация задается формулой:
```

$$I(A, B) = \frac{H(A) + H(B)}{H(A, B)}$$

```
    где H(X) – это энтропия ``- sum(x log x) для x в X``.
```

```
    Параметры
```

```
    -----
```

```
    A, B : массив ndarray
```

```
    Регистрируемые изображения.
```

```
    Возвращает
```

```
    -----
```

```
    nmI : вещественное
```

<sup>1</sup> См.: Стадхолм К., Хилл Д. Л. Г., Хоукс Д. Дж. Инвариантная к наложению мера энтропии выравнивания трехмерного медицинского изображения (С. Studholme, D. L. G. Hill, and D. J. Hawkes. An Overlap Invariant Entropy Measure of 3D Medical Image Alignment // Pattern Recognition 32, no. 1 (1999): 71–86). (<https://www.sciencedirect.com/science/article/pii/S0031320398000910>).

<sup>2</sup> Объяснение на скорую руку состоит в следующем: энтропия вычисляется из рассматриваемой гистограммы количества. Если  $X = Y$ , то совместная гистограмма  $(X, Y)$  – это диагональ, и эта диагональ совпадает с диагональю  $X$  или  $Y$ . Отсюда  $H(X) = H(Y) = H(X, Y)$  и  $I(X, Y) = 2$ .



*Нормализованная взаимная информация между двумя массивами, вычисленная с гранулярностью 100 интервалов на ось (всего 10 тыс. интервалов).*

```
«»»
hist, bin_edges = np.histogramdd([np.ravel(A), np.ravel(B)], bins=100)
hist /= np.sum(hist)

H_A = entropy(np.sum(hist, axis=0))
H_B = entropy(np.sum(hist, axis=1))
H_AB = entropy(np.ravel(hist))

return (H_A + H_B) / H_AB
```



Теперь мы определяем оптимизируемую *функцию стоимости* в соответствии с определением `cost_mse`, приведенным выше:

```
def cost_nmi(param, reference_image, target_image):
    transformation = make_rigid_transform(param)
    transformed = transform.warp(target_image, transformation, order=3)
    return -normalized_mutual_information(reference_image, transformed)
```

Наконец, мы применим ее с нашим оптимизирующим выравнивателем, работающим на основе алгоритма *basin hopping* (рис. 7.7):

```
print('*** Выравнивание синего с зеленым ***')
tf = align(green, blue, cost=cost_nmi)
cblue = transform.warp(blue, tf, order=3)

print('*** Выравнивание красного с зеленым ***')
tf = align(green, red, cost=cost_nmi)
cred = transform.warp(red, tf, order=3)

corrected = np.dstack((cred, green, cblue))
fig, ax = plt.subplots(figsize=(4.8, 4.8), tight_layout=True)
ax.imshow(corrected)
ax.axis('off')
```

\*\*\* Выравнивание синего с зеленым \*\*\*  
 Уровень: 1, Угол: 0.444, Сдвиг: (6.07, 0.354), Cost: -1.08  
 \*\* Выравнивание красного с зеленым \*\*\*  
 Уровень: 1, Угол: 0.000657, Сдвиг: (-0.635, -7.67), Cost: -1.11  
 (-0.5, 393.5, 340.5, -0.5)



Какое великолепное изображение! Только представьте, что этот экспонат был создан еще до появления цветной фотографии! Обратите внимание на жемчужно-белые одежды Господа, белую бороду Иоанна и белые страницы книги в руках Прохора, его писца – все они отсутствовали на выравнивании, основанном на СКО, но выглядят резко и четко после применения НВИ. Также обратите внимание на реалистичное золото подсвечников, расположенных на переднем плане.

В этой главе мы проиллюстрировали два ключевых понятия функциональной оптимизации: понимание локальных минимумов, как их избежать, и выбор правильной оптимизируемой функции для достижения конкретной цели. Решение этих функций позволяет применить оптимизацию к огромному спектру научных задач!



**Рис. 7.7** ❖ Каналы Prokudin-Gorskii  
выровнены с нормализованной взаимной информацией



---

# Глава 8

.....

## Большие данные с Toolz в маленьком ноутбуке

ГРЕЙСИ: Нож? Парень двенадцать футов ростом!  
ДЖЕК: Семь. Да не волнуйся ты, думаю, я с ним справлюсь.

– Джек Бертон,  
к/ф «Большой переполох в маленьком Китае»

Потоковая обработка не является изюминкой SciPy. Это скорее подход, позволяющий эффективно обрабатывать большие наборы данных, подобные тем, с которыми часто сталкиваются в науке. Язык Python содержит несколько полезных примитивов для обработки потоковых данных, и они могут быть объединены с библиотекой Toolz Мэтта Роклина (Matt Rocklin), предназначенной для генерирования элегантного и сжатого программного кода. Этот код эффективно экономит потребление оперативной памяти. В данной главе мы покажем, как применять принципы потоковой обработки, чтобы позволить вам работать с более крупными наборами данных, чем те, которые смогут поместиться в ОЗУ вашего компьютера.

Скорее всего, с потоковой обработкой вы уже работали. Самая простая потоковая форма состоит в итеративном переборе строк файла и обработке каждой строки без считывания всего файла в оперативную память. Например, подобный цикл будет состоять в вычислении среднего значения каждой строки и их суммировании:

```
import numpy as np

with open('data/expr.tsv') as f:
    sum_of_means = 0
    for line in f:
        sum_of_means += np.mean(np.fromstring(line, dtype=int, sep='\t'))
print(sum_of_means)
```

1463.0

Эта стратегия прекрасно работает, когда ваша задача легко решается построчной обработкой. Но при усложнении программного кода она может быстро выйти из-под контроля.

В потоковых программах некая функция обрабатывает фрагмент входных данных, возвращает обработанный блок, затем, пока нисходящие функции работают с этим блоком, сама функция получает следующий фрагмент и т. д. И все эти вещи происходят одновременно! Как обеспечить их надежную работу?

Для нас это тоже представляло трудность, пока мы не обнаружили библиотеку Toolz. Ее конструкции делают потоковые программы столь элегантными, что написание этой книги просто невозможно представить без включения главы, посвященной данной теме.

Давайте уточним, что мы имеем в виду под «потоковой обработкой» и зачем она вам может понадобиться. Допустим, что у вас в текстовом файле есть некие данные и вы хотите вычислить постолбцовое среднее логарифма значений по формуле  $\log(x + 1)$ . В таких случаях принято использовать библиотеку NumPy, при помощи которой загружают значения, вычисляют функцию логарифма для всех значений в полной матрице и затем берут среднее на первой оси:

```
import numpy as np

expr = np.loadtxt('data/expr.tsv')
logexpr = np.log(expr + 1)
np.mean(logexpr, axis=0)

array([ 3.11797294,  2.48682887,  2.19580049,  2.36001866,  2.70124539,
        2.64721531,  2.43704834,  3.28539133,  2.05363724,  2.37151577,
        3.85450782,  3.9488385 ,  2.46680157,  2.36334423,  3.18381635,
        2.64438124,  2.62966516,  2.84790568,  2.61691451,  4.12513405])
```



Все хорошо работает, следуя успокаивающе знакомой вычислительной модели ввода-вывода. Однако такой способ работы совершенно неэффективен! Мы загружаем полную матрицу в оперативную память (1), затем делаем ее копию, добавляя 1 в каждое значение (2), потом делаем еще одну копию для вычисления логарифма (3) и только после этого передаем ее в `np.mean`. Целых три экземпляра массива данных для выполнения операции, не требующей хранения в памяти даже одного экземпляра. Для любого вида операции с «большими данными» такой подход просто не будет работать.

Создатели языка Python это хорошо понимали и создали ключевое слово `yield`, наделяющее функции возможностью обрабатывать всего один «глоток» данных. Потом передавать результат дальше в следующий процесс и завершать цепь обработки одного фрагмента данных, перед тем как перейти к следующему. Слово «yield» (производить, передавать) очень хорошо это характеризует: функция передает управление в следующую функцию, ожидая продолжения обработки данных до тех пор, пока все нисходящие шаги не обработают эту точку данных.

## ПОТОКОВАЯ ПЕРЕДАЧА ПРИ ПОМОЩИ YIELD

Описанный выше поток управления довольно сложно отследить. Потрясающая особенность языка Python заключается в том, что он абстрагируется от этой сложности, позволяя вам сосредоточиваться на аналитическом функционале. Вот как это можно представить: для каждой обрабатываемой функции, которая обычно принимала список (коллекцию данных) и выполняла преобразование этого списка, вы можете переписать эту функцию как принимающую *поток* и *производящую* результат из каждого элемента этого потока.

Ниже приводится пример, где мы берем логарифм каждого элемента списка, используя два метода: стандартный метод копирования данных и потоковый метод:

```
def log_all_standard(input):
    output = []
    for elem in input:
        output.append(np.log(elem))
    return output

def log_all_streaming(input_stream):
    for elem in input_stream:
        yield np.log(elem)
```

Давайте проверим, дают ли оба метода одинаковый результат:

```
# Установить начальное значение случайного числа,
# чтобы получать воспроизводимые результаты
np.random.seed(seed=7)
```

```
# Задать настройки печати для отображения только 3 значащих разрядов
np.set_printoptions(precision=3, suppress=True)
```

```
arr = np.random.rand(1000) + 0.5
result_batch = sum(log_all_standard(arr))
print('Пакетный результат: ', result_batch)
result_stream = sum(log_all_streaming(arr))
print('Потоковый результат: ', result_stream)
```

```
Пакетный результат: -48.2409194561
Потоковый результат: -48.2409194561
```

Преимущество потокового метода состоит в том, что элементы потока не обрабатываются, пока в них не возникнет необходимость, будь то вычисление нарастающего итога или запись данных на жесткий диск либо что-то еще. Это позволяет сэкономить большой объем оперативной памяти при большом объеме входных значений или когда каждое значение очень большое. (Или и то, и другой вместе!) Приведенная ниже цитата из публикации в блоге<sup>1</sup> Мэтта очень кратко резюмирует полезность потокового анализа данных:

<sup>1</sup> См. <http://matthewrocklin.com/blog/work/2015/02/17/Towards-OOC-Bag>.

Исходя из моего небольшого опыта, люди редко идут по такому [потоковому] пути. Они используют однопоточный Python с хранением данных в оперативной памяти, пока все не зависнет, и затем ищут инфраструктуру обработки больших данных типа Hadoop/Spark с относительно высокими издержками по производительности.

И действительно, эта цитата идеально описывает нашу вычислительную карьеру. Однако промежуточный подход способен продвинуть вас *гораздо* дальше, чем вы думаете. В некоторых случаях он способен сделать это быстрее, чем подход на основе сверхвысокопроизводительных вычислений, устраняя издержки, связанные со взаимодействием многочисленных ядер и произвольным доступом к базам данных. (Обратитесь, например, к публикации в блоге «Большие данные; тот самый ноутбук»<sup>1</sup> Фрэнка Макшерри (Frank McSherry), где он обрабатывает граф с 128 миллиардами ребер на своем ноутбуке *быстрее*, чем используя графовую базу данных на суперкомпьютере.)

Чтобы внести ясность, каким образом при использовании потоковых функций разворачивается поток управления, полезно написать детализированные версии функции, печатающей сообщение вместе с каждой операцией.

```
import numpy as np

def tsv_line_to_array(line):
    lst = [float(elem) for elem in line.rstrip().split('\t')]
    return np.array(lst)

def readtsv(filename):
    print('вход в readtsv')
    with open(filename) as fin:
        for i, line in enumerate(fin):
            print(f'чтение строки {i}')
            yield tsv_line_to_array(line)
    print('выход из readtsv')

def add1(arrays_iter):
    print('вход в add1')
    for i, arr in enumerate(arrays_iter):
        print(f'добавление 1 к строке {i}')
        yield arr + 1
    print('выход из add1')

def log(arrays_iter):
    print('вход в log')
    for i, arr in enumerate(arrays_iter):
        print(f'взятие логарифма массива {i}')
        yield np.log(arr)
    print('выход из log')

def running_mean(arrays_iter):
    print('вход в running_mean')
    for i, arr in enumerate(arrays_iter):
```

<sup>1</sup> См. <http://www.frankmcsherry.org/graph/scalability/cost/2015/02/04/COST2.html>.

```

if i == 0:
    mean = arg
    mean += (arg - mean) / (i + 1)
    print(f'добавление строки {i} к скользящему среднему')
print('возвращение среднего значения')
return mean

```

Давайте посмотрим на эти функции в действии на примере небольшого демонстрационного файла:

```

fin = 'data/expr.tsv'
print('Создание итератора строк')
lines = readtsv(fin)
print('Создание итератора логарифма строк')
loglines = log(add1(lines))
print('Вычисление среднего')
mean = running_mean(loglines)
print(f'Среднее логарифма строк: {mean}')

```



Создание итератора строк

Создание итератора логарифма строк

Вычисление среднего

вход в running\_mean

вход в log

вход в add1

вход в readtsv

чтение строки 0

добавление 1 к строке 0

взятие логарифма массива 0

добавление строки 0 к скользящему среднему

чтение строки 1

добавление 1 к строке 1

взятие логарифма массива 1

добавление строки 1 к скользящему среднему

чтение строки 2

добавление 1 к строке 2

взятие логарифма массива 2

добавление строки 2 к скользящему среднему

чтение строки 3

добавление 1 к строке 3

взятие логарифма массива 3

добавление строки 3 к скользящему среднему

чтение строки 4

добавление 1 к строке 4

взятие логарифма массива 4

добавление строки 4 к скользящему среднему

выход из readtsv

выход из adding 1

выход из log

возвращение среднего значения

Среднее логарифма строк: [ 3.118 2.487 2.196 2.36 2.701 2.647 2.437 3.285  
2.054 2.372

3.855 3.949 2.467 2.363 3.184 2.644 2.63 2.848 2.617 4.125]

Примечание:

- при создании итератора строк и итератора логарифма строк никаких вычислений не происходит. Это связано с характером итераторов – они ленивые, имея в виду, что они не вычисляются (не *потребляются*) до тех пор, пока результат не станет необходим;
- когда в результате вызова функции `gunning_mean` вычисление наконец запускается, перед тем как перейти дальше к следующей строке, поток управления прыгает туда-сюда между всеми функциями, по мере того как выполняются различные вычисления, связанные с каждой строкой данных.

## ВВЕДЕНИЕ В ПОТОВОКУЮ БИБЛИОТЕКУ TOOLZ

В примере программного кода этой главы, предоставленном Мэттом Роклином, мы создаем марковскую модель на основе полного генома плодовой мушки, дрозофилы. Эта модель создается на ноутбуке менее чем за пять минут, используя всего несколько строк программного кода. (Мы его немного отредактировали, чтобы упростить его нисходящую обработку.) В примере Мэтта используется геном человека, но по вполне понятным причинам быстрдействие наших ноутбуков не было достаточным, поэтому вместо него мы собираемся использовать геном плодовой мушки (его размер составляет 1/20 генома человека). В течение этой главы данный пример будет немного расширен, чтобы можно было начинать обработку со сжатых данных (неужели же кто-то желает хранить несжатый набор данных на своем жестком диске?). Данная модификация почти *тривиальна*, что говорит в пользу элегантности этого примера.

```
import toolz as tz
from toolz import curried as c
from glob import glob
import itertools as it

LDICT = dict(zip('ACGTacgt', range(8)))
PDICT = {(a, b): (LDICT[a], LDICT[b])
         for a, b in it.product(LDICT, LDICT)}

def is_sequence(line):
    return not line.startswith('>')

def is_nucleotide(letter):
    return letter in LDICT # ignore 'N'

@tz.curry
def increment_model(model, index):
    model[index] += 1

def genome(file_pattern):
    «»Передать геном потоком, буква за буквой, из списка имен файлов FASTA.«»
    return tz.pipe(file_pattern, glob, sorted, # Имена файлов
                  c.map(open), # строки
                  # конкатенировать строки из всех файлов:
```

```
tz.concat,
# отбросить заголовок из каждой последовательности
c.filter(is_sequence),
# конкатенировать символы из всех строк
tz.concat,
# отбросить символы новой строки и 'N'
c.filter(is_nucleotide))
```



```
def markov(seq):
    «»Получить марковскую модель первого порядка
    из последовательности нуклеотидов.«»
    model = np.zeros((8, 8))
    tz.last(tz.pipe(seq,
                    c.sliding_window(2), # каждый последующий кортеж
                    c.map(PDICT.__getitem__), # местоположение кортежа
                                         # в матрице
                    c.map(increment_model(model)))) # прирастить матрицу
    # преобразовать количества в матрицу вероятностей переходов
    model /= np.sum(model, axis=1)[: , np.newaxis]
    return model
```



Затем, чтобы получить марковскую модель последовательностей, повторяющихся в геноме дрозофилы, мы можем сделать следующее:

```
%timeit -r 1 -n 1
dm = 'data/dm6.fa'
model = tz.pipe(dm, genome, c.take(10**7), markov)
# Мы используем `take`, чтобы построить модель на первых 10 млн оснований
# с целью ускорения обработки. Шаг, связанный с функцией take,
# можно пропустить, если вы готовы подождать ~5-10 мин.

1 loop, average of 1: 24.3 s +- 0 ns per loop (using standard deviation)
```

В этом примере происходит целый ряд событий, и мы собираемся его разложить по полочкам шаг за шагом. В конце главы мы выполним данный пример.

Первое, на что следует обратить внимание, – это ряд функций из библиотеки Toolz<sup>1</sup>. Например, из библиотеки Toolz мы взяли функции `pipe`, `sliding_window`, `frequencies` и каррированную версию функции `map` (подробнее об этом позднее). Это вызвано тем, что библиотека Toolz специально написана для мобилизации преимуществ итераторов Python и может легко управлять потоками.

Начнем с функционального конвейера `pipe`. Эта функция представляет собой синтаксический сахар, облегчающий чтение вызовов вложенных функций. Данная возможность приобретает особое значение, так как такой функциональный шаблон получает все большее распространение, когда приходится иметь дело с итераторами.

В качестве простого примера перепишем наше скользящее среднее с использованием функции `pipe`:

```
import toolz as tz
filename = 'data/expr.tsv'
```

<sup>1</sup> См. <http://toolz.readthedocs.org/en/latest/>.

```
mean = tz.pipe(filename, readtsv, add1, log, running_mean)
```

```
# Это эквивалентно следующему вложению функций:
# running_mean(log(add1(readtsv(filename))))
```

```
вход в running_mean
вход в log
вход в add1
вход в readtsv
чтение строки 0
добавление 1 к строке 0
взятие логарифма массива 0
добавление строки 0 к скользящему среднему
чтение строки 1
добавление 1 к строке 1
взятие логарифма массива 1
добавление строки 1 к скользящему среднему
чтение строки 2
добавление 1 к строке 2
взятие логарифма массива 2
добавление строки 2 к скользящему среднему
чтение строки 3
добавление 1 к строке 3
взятие логарифма массива 3
добавление строки 3 к скользящему среднему
чтение строки 4
добавление 1 к строке 4
взятие логарифма массива 4
добавление строки 4 к скользящему среднему
выход из readtsv
выход из add1
выход из log
возвращение среднего значения
```



То, что первоначально представляло собой многочисленные строки или кучу беспорядочных круглых скобок, теперь является четким описанием последовательных преобразований входных данных. И его гораздо легче понять!

Данная стратегия также имеет преимущество над оригинальной реализацией NumPy: если мы масштабируем наши данные до миллионов или миллиардов строк, нашему компьютеру пришлось бы из последних сил держать все данные в оперативной памяти. Напротив, здесь мы загружаем по одной строке из жесткого диска и обеспечиваем поддержание данных в объеме всего одной строки.

## ПОДСЧЕТ К-МЕР И ИСПРАВЛЕНИЕ ОШИБОК

Чтобы освежить свою память по информации о ДНК и геномике, вы, возможно, захотите просмотреть главы 1 и 2. Если коротко, то ваша генетическая информация, т. е. рецепт, по которому *вы* были созданы, закодирована в вашем *геноме* в виде последовательности химических *оснований*. Они чрезвычайно малы,



поэтому вы не сможете просто взять микроскоп и их прочитать. Вы также не сможете прочитать их длинную цепочку: ошибки накапливаются, и считывание становится непригодным. (Новые технологии это исправляют, но здесь мы сосредоточимся на наиболее распространенных сегодня коротко прочитанных данных секвенирования.) К счастью, все до единой клетки имеют идентичную копию вашего генома. Поэтому мы можем расщепить эти копии на крошечные сегменты (приблизительно 100 оснований в длину) и затем собрать их как огромный пазл из 30 миллионов фрагментов.

Перед выполнением сборки чрезвычайно важно выполнить коррекцию прочтений. За время секвенирования ДНК некоторые основания прочитываются неправильно и должны быть исправлены, иначе они испортят сборку. (Представьте фрагменты пазла неправильной формы.)

Одна из стратегий исправления состоит в том, чтобы в вашем наборе данных найти схожие прочтения и исправить ошибку, выхватив из этих прочтений правильную информацию. Или, в качестве альтернативы, вы можете полностью отказаться от прочтений с ошибками.

Однако этот способ очень неэффективный, потому что отыскание схожих прочтений означает необходимость сравнивать каждое прочтение со всеми другими прочтениями. На это потребуется  $N^2$  операций, или  $9 \times 10^{14}$  для 30-миллионного набора прочтений! (К тому же эти операции далеко не дешевые.)

Существует еще один способ. Павел Певзнер (Pavel Pevzner) и другие<sup>1</sup> поняли, что прочтения могут быть разбиты на меньшие, накладывающиеся  $k$ -меры<sup>2</sup>, подстроки длины  $k$ , которые затем сохраняются в хеш-таблице (в Python это словарь). У этого способа есть масса преимуществ. Главное преимущество – в том, что вместо вычислений на общем количестве прочтений, которое может быть произвольно большим, мы можем вычислять на общем количестве  $k$ -мер, равном величине самого генома. Это обычно на один-два порядка меньше количества прочтений.

Если мы выбираем довольно большую величину  $k$ , чтобы гарантировать появление в геноме любого  $k$ -мер только один раз, то количество появлений  $k$ -мер в точности равно количеству прочтений, приходящих из этого участка генома. Такое количество называется *покрытием* данного участка.

Если в прочтении есть ошибка, то существует высокая вероятность, что накладывающиеся на ошибку  $k$ -меры будут в геноме уникальными или близкими к уникальным. Ближайшую аналогию можно найти в английской литературе: если бы вам пришлось брать прочтения из Шекспира и одним из прочтений было «быть или ну быть», то встречаемость 7-мер «ну быть» будет редкой либо вовсе нулевой, тогда как встречаемость «не быть» будет очень частой.

<sup>1</sup> См. <http://www.pnas.org/content/98/17/9748.full>.

<sup>2</sup> Термин  $k$ -мер ( $k$ -mer) обычно относится ко всем возможным подстрокам длины  $k$ , содержащимся в строке. В вычислительной геномике  $k$ -мер относится ко всем возможным подпоследовательностям (длины  $k$ ) из прочтения, полученного при помощи ДНК-секвенирования. – *Прим. перев.*

В этом основа для исправления ошибок k-мер: разбить прочтения на k-меры, подсчитать количество появлений каждого k-мер и применить логику, чтобы заменить в прочтениях редкие k-меры на схожие общие. (Или же в качестве альтернативы отбросить прочтения с ошибочными k-мерами. Последнее возможно благодаря тому, что прочтения присутствуют в таком изобилии, что мы можем позволить себе отбраковать ошибочные данные.)

Этот пример ярко демонстрирует *чрезвычайную* важность потоковой обработки. Как было отмечено ранее, количество прочтений может быть огромным, поэтому мы не хотим хранить их в оперативной памяти.

Данные последовательности ДНК общепринято представлять в формате FASTA. Это текстовый формат, состоящий из одной или нескольких последовательностей ДНК в расчете на файл, при этом каждая имеет имя и саму последовательность.

Типичный файл FASTA:

```
> sequence_name1
TCAATCTCTTTATATAGATCTCGTTAAAGTAAAATTTGGTTTGTGTTAAAGTACAAG
GGGTACCTATGACCACGGAACCAACAAAGTCCSTAAATAGGACATCAAGTAACTACCGGT
ACGT
```

```
> sequence_name2
ATGTCCCGGGCTTCTTTTGCATTTGCTTCGCATTAACAGAATATCCAGCGTACTTAGG
ATTGTGACCTGTCTTGTGCTACGTGGCCGCAACACCAGGTATAGTGCCAATACAAGTCA
GACTAAAACCTGGTTC
```

Теперь у нас есть требуемая информация для преобразования потока строк из файла FASTA и подсчета k-мер:

- отфильтровать строки так, чтобы использовались только строки последовательностей;
- для каждой строки последовательности произвести поток k-мер;
- добавить каждый k-мер в счетчик количеств в словаре.

Вот как это делается на чистом Python, не используя ничего, кроме встроенных функций:

```
def is_sequence(line):
    line = line.rstrip() # удалить '\n' в конце строки
    return len(line) > 0 and not line.startswith('>')

def reads_to_kmers(reads_iter, k=7):
    for read in reads_iter:
        for start in range(0, len(read) - k):
            yield read[start : start + k] # обратите внимание на yield,
            # это генератор

def kmer_counter(kmer_iter):
    counts = {}
    for kmer in kmer_iter:
        if kmer not in counts:
            counts[kmer] = 0
        counts[kmer] += 1
    return counts
```

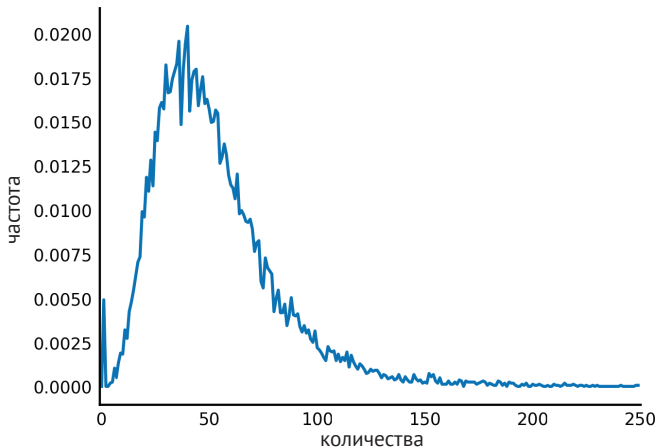
```
with open('data/sample.fasta') as fin:
    reads = filter(is_sequence, fin)
    kmers = reads_to_kmers(reads)
    counts = kmer_counter(kmers)
```

Этот код прекрасно работает, и данные поступают в поток – прочтения загружаются по одному из жесткого диска и передаются по конвейеру через конвертор k-мер в счетчик k-мер. В результате мы можем построить гистограмму количеств и подтвердить, что действительно имеется две хорошо разделенные популяции правильных и ошибочных k-мер:

```
# Заставить графики появляться локально, задать стиль графиков
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('style/elegant.mplstyle')

def integer_histogram(counts, normed=True, xlim=[], ylim=[],
                    *args, **kwargs):
    hist = np.bincount(counts)
    if normed:
        hist = hist / np.sum(hist)
    fig, ax = plt.subplots()
    ax.plot(np.arange(hist.size), hist, *args, **kwargs)
    ax.set_xlabel('количества')
    ax.set_ylabel('частота')
    ax.set_xlim(*xlim)
    ax.set_ylim(*ylim)

counts_arr = np.fromiter(counts.values(), dtype=int, count=len(counts))
integer_histogram(counts_arr, xlim=(-1, 250))
```



Обратите внимание на замечательное распределение частот k-мер и большую вогнутость k-мер (в левой части графика), которая появляется всего один раз. Такие низкочастотные k-меры, скорее всего, являются ошибками.

Но в приведенном выше программном коде мы на самом деле выполняем слишком много работы. Значительная часть функциональности, которую мы поместили в циклы `for` и инструкцию `yield`, на самом деле манипулирует потоком: преобразование потока данных в другой вид данных и его накопление в конце. Библиотека `Toolz` имеет целый ряд примитивов для управления потоком, которые облегчают написание вышеупомянутого программного кода всего в одном вызове функции; причем по именам преобразующих функций также становится легче визуализировать то, что происходит в каждой точке с вашим потоком данных.

Например, функция скользящего окна `sliding_window` представляет собой именно то, что нам нужно для создания  $k$ -мер:

```
print(tz.sliding_window.__doc__)
```

Последовательность накладываемых последовательностей

```
>>> list(sliding_window(2, [1, 2, 3, 4]))
[(1, 2), (2, 3), (3, 4)]
```

Эта функция создает скользящее окно, подходящее для таких преобразований, как скользящие средние / сглаживание

```
>>> mean = lambda seq: float(sum(seq)) / len(seq)
>>> list(map(mean, sliding_window(2, [1, 2, 3, 4])))
[1.5, 2.5, 3.5]
```

Кроме того, функция частот `frequencies` подсчитывает количество появлений отдельных значений в потоке данных. Вместе с конвейером `pipe` теперь мы можем подсчитать  $k$ -меры в одном вызове функции:

```
from toolz import curried as c
```

```
k = 7
counts = tz.pipe('data/sample.fasta', open,
                 c.filter(is_sequence),
                 c.map(str.rstrip),
                 c.map(c.sliding_window(k)),
                 tz.concat, c.map(''.join),
                 tz.frequencies)
```

Но постойте: что это за вызовы функций с префиксом `c` из пространства имен `toolz.curried`?

## КАРРИРОВАНИЕ: ИЗЮМИНКА ПОТОКОВОЙ ОБРАБОТКИ

Ранее мы запросто использовали каррированную версию функции `map`, которая применяет заданную функцию к каждому элементу в последовательности. Теперь, когда мы примешали туда еще несколько каррированных вызовов, пора поделиться с вами, что это означает! Операция каррирования получила такое название не по имени пряной приправы «карри» (хотя, надо признать, она действительно приправляет ваш программный код). Она названа в честь Кар-

ри Хаскелла, математика, который изобрел это понятие. Карри Хаскелл также является тезкой языка программирования Haskell, в котором *все* функции каррированы!

«Каррирование» означает *частичное* вычисление функции и возвращение еще одной «меньшей» функции. В обычной ситуации, если в Python не предоставить функции все необходимые аргументы, она закати́т истерику. Напротив, каррированная функция может принимать лишь *некоторые* из этих аргументов. Если каррированная функция не получает достаточно аргументов, то возвращает новую функцию, принимающую остальные аргументы. Когда эта вторая функция вызывается с остальными аргументами, она может выполнить исходную задачу. Еще одним термином для каррирования является *частичное применение*. В функциональном программировании каррирование является средством порождения функции, ожидающей остальных аргументов, которые появятся позже.

Поэтому, в отличие от вызова функции `map(np.log, numbers_list)`, применяющей функцию `np.log` ко всем числам в списке `numbers_list` (возвращая последовательность логарифмированных чисел), вызов каррированной функции `toolz.curried.map(np.log)` возвращает *функцию*, принимающую последовательность чисел и возвращающую последовательность логарифмированных чисел.

Как оказалось, наличие функции, которая уже знает некоторые аргументы, идеально подходит для потоковой обработки! В приведенном выше фрагменте кода мы уже увидели наметки, показывающие, насколько мощным может быть каррирование совместно с конвейерами.

Однако когда вы начинаете впервые, каррирование может заставить вас поднапрячься. Поэтому мы попробуем эту операцию на нескольких простых примерах, чтобы продемонстрировать, как она работает. Давайте начнем с того, что напишем простую некаррированную функцию:

```
def add(a, b):  
    return a + b
```

```
add(2, 5)
```

7

Теперь мы пишем похожую функцию, которую прокарируем вручную:

```
def add_curried(a, b=None):  
    if b is None:  
        # Второй аргумент не задан, поэтому создать функцию и ее вернуть  
        def add_partial(b):  
            return add(a, b)  
        return add_partial  
    else:  
        # Оба аргумента заданы, поэтому просто вернуть значение  
        return add(a, b)
```

Теперь давайте испытаем каррированную функцию, чтобы удостовериться в том, что она делает именно то, что мы ожидаем:

```
add_curried(2, 5)
```

```
7
```

Замечательно. Когда заданы обе переменные, она действует как нормальная функция. Теперь давайте исключим вторую переменную:

```
add_curried(2)
```

```
<function __main__.add_curried.<locals>.add_partial>
```

Как мы и ожидали, она вернула функцию. Теперь давайте применим эту возвращенную функцию:

```
add2 = add_curried(2)
```

```
add2(5)
```

```
7
```

Сработало! Правда, читабельность функции `add_curried` вне всякой критики. При ее использовании в будущем, вероятно, могут возникнуть затруднения, и придется поднапрячься, чтобы вспомнить, зачем мы написали этот фрагмент кода. К счастью, библиотека `Toolz` имеет инструменты, которые способны нас выручить.

```
import toolz as tz
```

```
@tz.curry # Применить curry в качестве декоратора
```

```
def add(x, y):
    return x + y
```

```
add_partial = add(2)
```

```
add_partial(5)
```

```
7
```

Подводя итог сделанному, функция `add` теперь является каррированной функцией. Поэтому она может принимать один из аргументов и возвращать еще одну функцию, `add_partial`, которая «запоминает» этот аргумент.

На самом деле все функции библиотеки `Toolz` также доступны в каррированном виде и находятся в пространстве имен `toolz.curried`. Библиотека `Toolz` также включает каррированные версии некоторых вспомогательных функций Python высшего порядка, таких как `map`, `filter` и `reduce`. Мы импортируем пространство имен `curried` под псевдонимом `c`, чтобы не загромождать наш программный код. Так, к примеру, каррированной версией функции `map` будет `c.map`. Стоит отметить, что каррированные функции (например, `c.map`) отличаются от декораторов `@curry`, которые используются для создания каррированных функций.

```
from toolz import curried as c
```

```
c.map
```

```
<class 'map'>
```



В качестве напоминания заметим, функция `map` встроена в Python. Из документации<sup>1</sup>:

`map(функция, итерируемый_объект, ...)` Возвращает итератор, применяющий функцию к каждому значению итерируемого объекта, производя результат ее применения.

Каррированная версия функции `map` особенно удобна во время работы в конвейере библиотеки Toolz. Используя конвейер `tz.pipe`, вы можете передать свою функцию в каррированную функцию `c.map` и позже направить потоком итератор. Взгляните еще раз на нашу функцию чтения генома, чтобы увидеть, как все это работает на практике.

```
def genome(file_pattern):
    «»»Передать геном потоком, буква за буквой, из списка имен файлов FASTA.»»»
    return tz.pipe(file_pattern, glob, sorted, # Имена файлов
                  c.map(open), # строки
                  # конкатенировать строки из всех файлов:
                  tz.concat,
                  # отбросить заголовок из каждой последовательности
                  c.filter(is_sequence),
                  # конкатенировать символы из всех строк
                  tz.concat,
                  # отбросить символы новой строки и 'N'
                  c.filter(is_nucleotide))
```



## ВОЗВРАЩАЯСЬ К ПОДСЧЕТУ К-МЕР

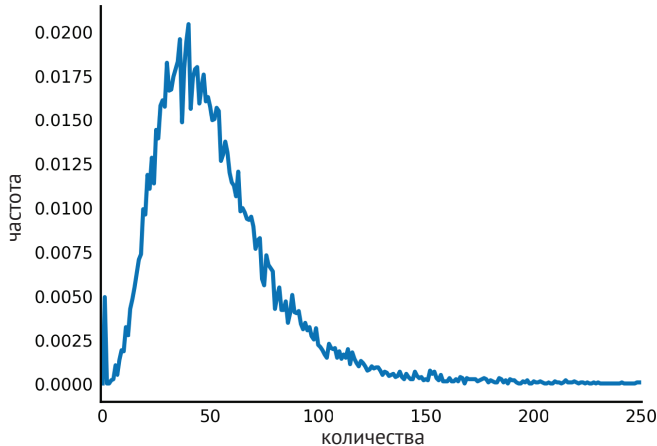
Итак, надо полагать, теперь мы ориентируемся в каррировании. Давайте вернемся к нашему программному коду подсчета k-мер. Ниже еще раз приведем фрагмент кода с использованием этих каррированных функций:

```
from toolz import curried as c
k = 7
counts = tz.pipe('data/sample.fasta', open,
                 c.filter(is_sequence),
                 c.map(str.rstrip),
                 c.map(c.sliding_window(k)),
                 tz.concat, c.map(''.join),
                 tz.frequencies)
```

Теперь мы можем увидеть частоту различных k-мер:

```
counts = np.fromiter(counts.values(), dtype=int, count=len(counts))
integer_histogram(counts, xlim=(-1, 250), lw=2)
```

<sup>1</sup> См. <https://docs.python.org/3.4/library/functions.html%23map>.



### Советы для работы с потоками

- «Список списков» следует преобразовывать в «длинный список», используя для этого функцию `tz.concat`.
- Не попадитесь на следующем:
  - итераторы потребляются. Поэтому если вы создаете объект генератора и выполняете с ним некую обработку и потом на каком-то последующем шаге терпите неудачу, то вам следует создать генератор заново. Оригинал перестает существовать;
  - итераторы ленивы. Вам необходимо иногда вызывать вычисление.
- Когда в конвейере много функций, иногда трудно установить, где работа пошла не так, как надо. Возьмите небольшой поток и добавляйте в свой конвейер по одной функции, начиная с первой/крайней левой, до тех пор, пока вы не найдете ту, которая вызывает ошибку. Вы также можете вставить выражение `map(do(print))` (функции `map` и `do` взяты из пространства имен `toolz.curried`) в любую точку в потоке, чтобы распечатывать каждый элемент во время его потоковой передачи.

### Задача: анализ главных компонент потоковых данных

В библиотеке `scikit-learn` существует класс `IncrementalPCA`, позволяющий выполнять анализ главных компонент (`principal components analysis`, PCA) набора данных, не загружая полный набор данных в оперативную память. Однако при этом вам необходимо разбить свои данные на блоки самостоятельно, что делает использование вашего программного кода немного неудобным. Создайте функцию, которая способна принимать поток образцов данных и выполнять PCA. Затем примените эту функцию, чтобы вычислить PCA набора данных `iris` («Ирисы Фишера»), предназначенного для тестирования алгоритмов машинного обучения, который находится в файле `data/iris.csv`. (Доступ к этому набору данных можно также получить из модуля `datasets` библиотеки `scikit-learn`, используя `datasets.load_iris()`.) По желанию вы можете раскрасить точки с номерами видов цветков из файла `data/iris-target.csv`.



➔ Класс `IncrementalPCA` находится в модуле `sklearn.decomposition` и для пакетной тренировки модели требует размер пакета больше 1. Что касается того, как создавать поток пакетов из потока точек данных, следует обратить внимание на функцию `toolz.curried.partition`.

## МАРКОВСКАЯ МОДЕЛЬ НА ОСНОВЕ ПОЛНОГО ГЕНОМА

Вернемся к нашему первоначальному примеру программного кода. Что такое марковская модель, и в чем ее прелесть?

В целом марковская модель, или модель Маркова, исходит из того, что вероятность перехода системы в заданное состояние зависит только от состояния, в котором она находилась непосредственно перед этим. Например, если сейчас солнечно, то имеется высокая вероятность, что завтра тоже будет солнечно. И тот факт, что вчера шел дождь, не имеет никакого значения. В этой теории вся информация, необходимая для предсказания будущего, закодирована в текущем состоянии дел. Прошое не имеет значения. Это допущение полезно тем, что упрощает решение задач, которые не поддаются решению в других условиях, и часто дает хорошие результаты. Марковские модели, например, лежат в основе большинства методов обработки сигналов в мобильных телефонах и спутниковой связи.

В контексте геномики, как мы увидим, разные функциональные участки генома имеют разные *переходные вероятности* между аналогичными состояниями. Встречая их в новом геноме, мы можем делать некие предсказания о функции этих участков. Возвращаясь к погодной аналогии, вероятность перехода от солнечного дня к дождливому очень отличается в зависимости от местоположения: находитесь вы в Лос-Анджелесе или в Лондоне. Поэтому, если я дам вам последовательность дней (солнечно, солнечно, солнечно, дождливо, солнечно, ...), при условии что ваша модель предварительно была натренирована, вы можете предсказать, откуда она поступила: из Лос-Анджелеса или из Лондона.

В этой главе мы пока затронем только построение модели.

Для этого вам надо скачать файл с геномом дрозофилы фруктовой (плодовой мушки) `dm6.fa.gz`<sup>1</sup>. И вам придется его распаковать, применив команду `gzip -d dm6.fa.gz`.

В геномных данных генетическая последовательность, состоящая из букв А, С, G и T, содержит информацию о принадлежности к *повторяющимся элементам*, т. е. особому классу ДНК. Эта информация закодирована регистром букв, т. е. тем, находятся ли буквы последовательности в нижнем регистре (повторяющиеся элементы) или в верхнем регистре (неповторяющиеся элементы). Мы можем использовать эту информацию во время построения марковской модели.

Запрограммируем марковскую модель в виде массива NumPy. А чтобы индексировать по буквам в индексы в [0, 7] (имя `LDICT` обозначает «словарь букв»)

<sup>1</sup> См. <http://hgdownload.cse.ucsc.edu/goldenPath/dm6/bigZips/>.

и по парам букв в двумерные индексы в ([0, 7], [0, 7]) (имя PDICT обозначает «словарь пар»), создадим словари:

```
import itertools as it
LDICT = dict(zip('ACGTacgt', range(8)))
PDICT = {(a, b): (LDICT[a], LDICT[b])
         for a, b in it.product(LDICT, LDICT)}
```

Кроме того, следует отфильтровать данные, которые не принадлежат последовательностям: имена последовательностей, находящиеся в строках и начинающиеся с символа >, и неизвестную последовательность, которая помечена как N. Для этого мы создадим функции с критериями, на основе которых будет выполняться фильтрация:

```
def is_sequence(line):
    return not line.startswith('>')

def is_nucleotide(letter):
    return letter in LDICT # игнорировать 'N'
```

Наконец, всякий раз, когда мы будем получать новую нуклеотидную пару, скажем, (A, T), будем наращивать нашу марковскую модель (матрицу NumPy) в соответствующей позиции. Для выполнения этой работы мы создадим каррированную функцию:

```
import toolz as tz

@tz.curry
def increment_model(model, index):
    model[index] += 1
```



Теперь мы можем эти элементы объединить и направить геном потоком в нашу матрицу NumPy. Обратите внимание, если seq является потоком, то нам не придется втискивать в оперативную память ни полный геном, ни даже большой кусок генома!

```
from toolz import curried as c

def markov(seq):
    «»Получить марковскую модель первого порядка из
    последовательности нуклеотидов.»»
    model = np.zeros((8, 8))
    tz.last(tz.pipe(seq,
                    c.sliding_window(2), # каждый последующий кортеж
                    c.map(PDICT.__getitem__), # позиция в матрице кортежей
                    c.map(increment_model(model)))) # нарастить матрицу
    # Преобразовать матрицу переходных вероятностей
    model /= np.sum(model, axis=1)[:, np.newaxis]
    return model
```

Теперь от нас требуется только запустить этот геномный поток и создать нашу марковскую модель:

```

from glob import glob

def genome(file_pattern):
    «»»Передать геном потоком, буква за буквой, из списка имен файлов FASTA.»»»
    return tz.pipe(file_pattern, glob, sorted, # Имена файлов
                  c.map(open), # строки
                  # конкатенировать строки из всех файлов:
                  tz.concat,
                  # отбросить заголовок из каждой последовательности
                  c.filter(is_sequence),
                  # конкатенировать символы из всех строк
                  tz.concat,
                  # отбросить символы новой строки и 'N'
                  c.filter(is_nucleotide))

```



Давайте теперь проверим геном дрозофилы (плодовой мушки):

```

# Скачать dm6.fa.gz из ftp://hgdownload.cse.ucsc.edu/goldenPath/dm6/bigZips/
# Перед использованием распаковать: gzip -d dm6.fa.gz
dm = 'data/dm6.fa'
model = tz.pipe(dm, genome, c.take(10**7), markov)
# Мы используем `take`, чтобы построить модель на первых 10 млн оснований
# с целью ускорения обработки. Шаг, связанный с функцией take,
# можно пропустить, если вы готовы подождать ~5-10 мин.

```

А теперь давайте посмотрим на результирующую матрицу:

```

print(' ', ' '.join('ACGTacgt'), '\n')
print(model)

  A C G T a c g t
[[ 0.348 0.182 0.194 0.275 0. 0. 0. 0. ]
 [ 0.322 0.224 0.198 0.254 0. 0. 0. 0. ]
 [ 0.262 0.272 0.226 0.239 0. 0. 0. 0. ]
 [ 0.209 0.199 0.245 0.347 0. 0. 0. 0. ]
 [ 0.003 0.003 0.003 0.003 0.349 0.178 0.166 0.296]
 [ 0.002 0.002 0.003 0.003 0.376 0.195 0.152 0.267]
 [ 0.002 0.003 0.003 0.002 0.281 0.231 0.194 0.282]
 [ 0.002 0.002 0.003 0.003 0.242 0.169 0.227 0.351]]

```



Вероятно, полученный результат будет яснее, если продемонстрировать его на диаграмме (рис. 8.1):

```

def plot_model(model, labels, figure=None):
    fig = figure or plt.figure()
    ax = fig.add_axes([0.1, 0.1, 0.8, 0.8])
    im = ax.imshow(model, cmap='magma');
    axcolor = fig.add_axes([0.91, 0.1, 0.02, 0.8])
    plt.colorbar(im, cax=axcolor)
    for axis in [ax.xaxis, ax.yaxis]:
        axis.set_ticks(range(8))
        axis.set_ticks_position('none')
        axis.set_ticklabels(labels)
    return ax

plot_model(model, labels='ACGTacgt');

```

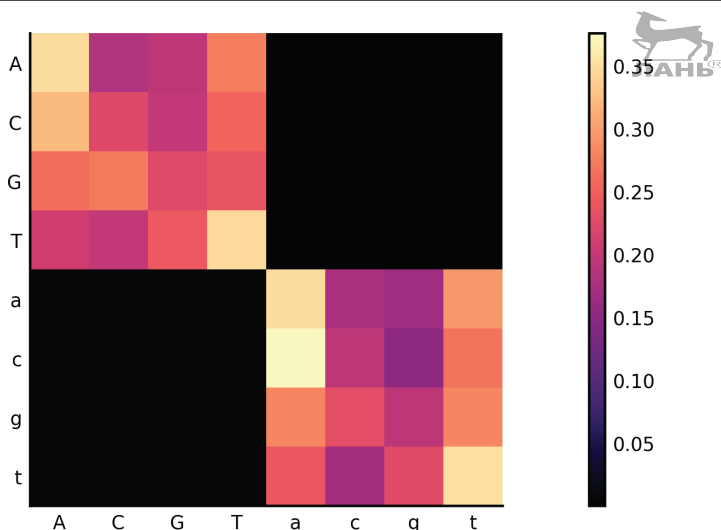


Рис. 8.1 ❖ Матрица переходных вероятностей для генетической последовательности в геноме дрозофилы фруктовой

Обратите внимание, насколько переходы C–A и G–C отличаются в повторяющихся и неповторяющихся участках генома. Эта информация может использоваться для классификации ранее не встречавшейся последовательности ДНК.

## Задача: онлайн-распаковка архива

Рекомендуем в начало конвейера добавить шаг, который будет выполнять распаковку данных из архива, избавляя вас от необходимости держать распакованную версию данных на своем жестком диске. Будучи в сжатом виде (модулем `gzip`), геном дрозофилы, например, занимает меньше одной трети пространства жесткого диска. И разумеется, распаковка архива тоже может выполняться в потоке!

➔ Пакет `gzip`, являющийся частью стандартной библиотеки Python, позволяет открывать файлы `.gz`, как если бы они были обычными файлами.

Мы надеемся, что нам удалось хотя бы дать вам намек, как легко можно выполнять потоковую обработку в Python, при использовании всего нескольких абстракций, наподобие тех, которые обеспечивает библиотека `Toolz`.

Потоковая обработка может сделать вашу работу продуктивнее, потому что работа с большим объемом данных занимает линейно более продолжительное время, чем с малым объемом данных. В пакетном анализе работа с большим объемом данных может продолжаться бесконечно, потому что операционной системе придется все время передавать данные из ОЗУ на жесткий диск и обратно. Или, что еще хуже, Python и вовсе может отказаться их обрабатывать,

просто показав ошибку MemoryError! Все это означает, что в случае проведения различных видов анализа больших наборов данных вам не потребуется более мощная машина. И если ваши тесты успешно проходят на малых объемах данных, то они успешно пройдут и на больших объемах данных!

Наша ключевая мысль заключается в следующем: при написании алгоритма или исследования подумайте, сможете ли вы применить потоковую обработку. Если да, то делайте это с самого начала. В будущем вы сами скажете себе спасибо. Позже сделать это будет сложнее, и в конечном счете приведет к ситуации, как на рис. 8.2.

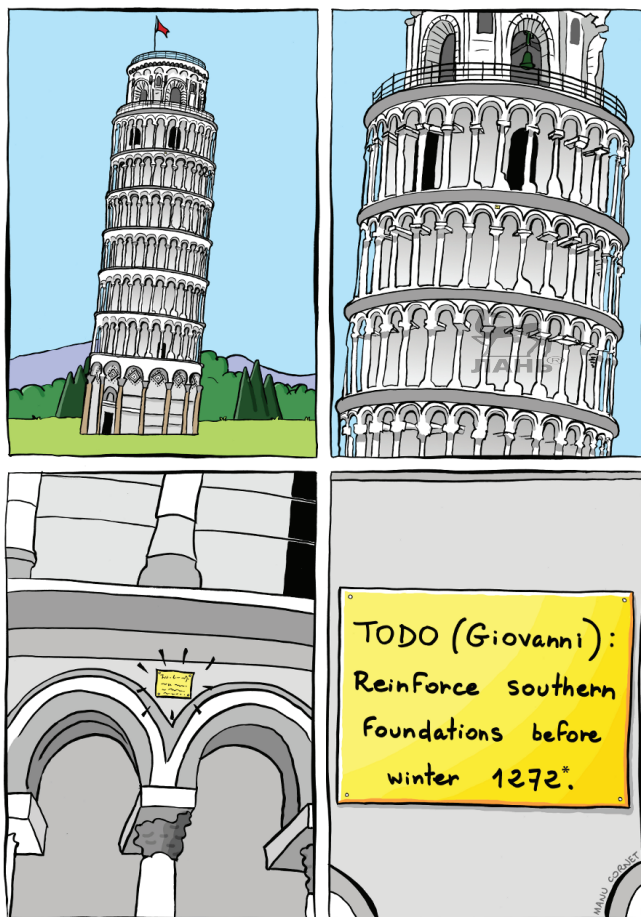


Рис. 8.2 ❖ Списки неотложных дел в истории  
(карикурист Ману Корне, использовано с разрешения)

\* TODO (Джованни): укрепить южный фундамент до наступления зимы 1273

# Эпилог



Качество означает делать все правильно, когда никто не смотрит.

– Генри Форд

В этой книге наша главная задача состояла в том, чтобы помочь в распространении элегантного применения библиотек NumPy и SciPy. Обучив вас тому, как при помощи библиотеки SciPy выполнять эффективный научный анализ, мы надеемся, что вселили в вас ощущение, что качественный программный код заслуживает необходимых для него усилий.

## Что дальше?

Теперь, когда вы знаете библиотеку SciPy в мере, чтобы проводить анализ любых данных, встающих на вашем пути, встает вопрос: что нужно сделать, чтобы пойти дальше? В самом начале книги мы сказали, что у нас не было никаких замыслов охватить все, что только можно узнать о библиотеке и всех ее ответвлениях. Прежде чем мы разойдемся, хотим направить вас к многочисленным ресурсам, которые существуют, чтобы вам помочь.

## Списки рассылок

В предисловии мы упомянули, что SciPy является сообществом. Отличный способ продолжить обучение – подписаться на главные списки рассылок на библиотеки NumPy, SciPy, pandas, Matplotlib, scikit-image и другие, в которых вы можете быть заинтересованы, и регулярно их читать.

И когда вы по-настоящему попадете в тупик в своей работе, не бойтесь обратиться туда за помощью! Мы – дружелюбная компания! При поиске помощи от вас требует только *одно*. Вы должны показать, что вы пытались решить задачу самостоятельно, и предоставить другим участникам минимальный сценарий и образец данных в достаточном объеме, чтобы продемонстрировать вашу проблему и то, как вы попытались ее исправить.

- **Нет:** «Мне нужно сгенерировать большой массив случайных гауссиан. Кто-нибудь сможет помочь?»
- **Нет:** «У меня вот тут огромная библиотека: [https://github.com/ron\\_obvious](https://github.com/ron_obvious). Если вы посмотрите в статистическую библиотеку, то там есть фрагмент, для которого реально требуются случайные гауссианы. Кто-нибудь сможет взглянуть???»
- **Да:** «Я попытался сгенерировать большой список случайных гауссиан, вот так: `gauss = [np.random.randn ()] * 10 ** 5`. Но когда я вычисляю `np.mean(gauss)`, у меня практически ни разу не получались значения,

близкие к 0, как хотел бы я. Что я делаю неправильно? Полный сценарий прилагается ниже».



## GitHub

В предисловии мы также поговорили о GitHub. Весь рассмотренный нами в этой книге программный код и другие материалы лежат на GitHub:

- NumPy<sup>1</sup>;
- SciPy<sup>2</sup>.

Когда что-то, вопреки вашим ожиданиям, не работает, возможно, причиной тому является ошибка. Если после небольшого расследования вы убеждены, что действительно обнаружили ошибку, вам следует пройти в закладку «issues» (вопросы) соответствующего хранилища GitHub и создать новый вопрос. Тем самым будет гарантировано, что ошибка попадет на радар разработчиков этой библиотеки и что она (надо надеяться) будет исправлена в следующей ее версии. Между прочим, этот совет также относится и к «ошибкам» в документации: если в документации библиотеки вам что-то не ясно, зарегистрируйте вопрос!

Помимо регистрации вопроса, будет еще лучше, если вы *отправите запрос на включение внесенных изменений*. Такой запрос с улучшением документации библиотеки является отличным способом попробовать свои силы в открытом исходном коде! Мы здесь не будем рассматривать весь этот процесс, однако предложим вам целый ряд книг и ресурсов, которые вам помогут:

- «Эффективное вычисление в физике» (*Anthony Scopatz, Katy Huff. Effective Computation in Physics. O'Reilly, 2015*)<sup>3</sup>. В этой книге среди многих других тем в области научных вычислений рассматриваются Git и GitHub;
- «Введение в GitHub» (*Peter Bell, Brent Beer. Introducing GitHub, O'Reilly, 2015*)<sup>4</sup>. В этой книге GitHub рассматривается более подробно;
- «Организация Software Carpentry»<sup>5</sup> содержит уроки по Git и круглый год предлагает бесплатные семинары по всему миру;
- один из авторов книги, частично опираясь на эти уроки, создал законченное учебное пособие по включению внесенных изменений с участием Git и GitHub «Открытая наука с Git и GitHub» (Open Source Science with Git and GitHub)<sup>6</sup>;
- наконец, многие проекты с открытым исходным кодом на GitHub имеют файл «CONTRIBUTING»<sup>7</sup>, который содержит ряд рекомендаций по участию своим программным кодом в проекте.

<sup>1</sup> См. <https://github.com/numpy/numpy>.

<sup>2</sup> См. <https://github.com/scipy/scipy>.

<sup>3</sup> См. <http://shop.oreilly.com/product/0636920033424.do>.

<sup>4</sup> См. <http://shop.oreilly.com/product/0636920033059.do>.

<sup>5</sup> См. <https://software-carpentry.org/>.

<sup>6</sup> См. <http://jni.github.io/git-tutorial/>.

<sup>7</sup> См. <https://github.com/scikit-image/scikit-image/blob/master/CONTRIBUTING.txt>.

Одним словом, интеллектуальный голод из-за отсутствия помощи по этой теме вам не грозит!

Мы приглашаем вас как можно чаще принимать участие в экосистеме SciPy не только потому, что вы сможете поспособствовать, чтобы эти библиотеки стали лучше для всех, но и потому, что это один из лучших способов развить свои навыки программирования. С каждым отправляемым вами запросом на включение внесенных изменений вы получите отзыв о своем программном коде, который поможет вам стать лучше. Вы сможете ближе познакомиться с процессом участия в GitHub и этикетом, которые являются весьма ценными навыками на сегодняшнем рынке труда.

## Конференции

Если вам понравилась эта книга, мы настоятельно рекомендуем посещать конференции по программированию в этой области. Фантастическая конференция SciPy каждый год проводится в Остине, шт. Техас. Вероятно, это самая лучшая конференция из возможных вариантов. Также существует европейская версия конференции, EuroSciPy, которая каждые два года проводится в другом городе-организаторе. Наконец, существует более общая конференция PyCon, проводимая в Соединенных Штатах. Но эта конференция имеет ответвления по всему миру, такие как PyCon-AU в Австралии, у которой имеется мини-конференция «Наука и данные», проводимая за день до главной конференции.

Какую бы конференцию вы ни выбрали, дождитесь спринтерских состязаний в конце конференции. Программистский спринт – это интенсивный сеанс командного программирования и фантастическая возможность изучить процесс внесения своего вклада в открытый исходный код, независимо от вашего уровня квалификации. Именно так один из ваших авторов (Хуан) начал свое путешествие по открытому исходному коду.

## ЗА ПРЕДЕЛАМИ SciPy

Библиотека SciPy написана не только на Python, но и на высокооптимизированных языках C и Fortran, которые взаимодействуют с Python. Вместе с NumPy и связанными с ней библиотеками SciPy пытается охватить значительную часть вариантов использования, которые возникают в научном анализе данных, и для этого обеспечивает очень быстрые функции. Тем не менее иногда научная задача совершенно не совпадает с тем, что уже есть в SciPy, и чистое Python'овское решение может оказаться бесполезным, так как не дает достаточной скорости. Что же делать?

*Высокоэффективный Python*<sup>1</sup> (High Performance Python. O'Reilly, 2014) под авторством Мичи Горелика и Иэна Озсвалда (Micha Gorelick, Ian Ozsvald) рассматривает все, что вам нужно знать в таких ситуациях: как найти то место, где вам

<sup>1</sup> См. <http://shop.oreilly.com/product/0636920028963.do>.





действительно нужна производительность, и какие существуют варианты получения этой производительности. Мы настоятельно рекомендуем эту книгу.

Далее мы кратко упомянем два варианта, которые особенно актуальны для мира SciPy.

В первую очередь это Cython. Cython является версией Python, допускающей компиляцию в C и затем обратное импортирование в Python. Добавление некоторых аннотаций типов в переменные Python приводит к тому, что скомпилированный в C программный код может быть в сотни, а то и тысячи раз быстрее сравнимого программного кода на Python. Cython сегодня является промышленным стандартом и используется в NumPy, SciPy и многих связанных с ними библиотеках (таких как scikit-image), обеспечивая быстрые алгоритмы в программном коде на основе массивов. Курт Смит (Kurt Smith) написал книгу с простым названием Cython<sup>1</sup> (O'Reilly, 2015), которая научит вас основам этого языка.

Нередко более легкой в использовании альтернативой Cython является Numba, JIT-компилятор для программного кода Python на основе массивов. JIT-компиляторы («just-in-time», или динамические, компилирующие «налету») ждут первого исполнения функции и в точке исполнения логически выводят тип всех аргументов функции и результатов и компилируют программный код в высокоэффективную форму для этих конкретных типов. В Numba аннотировать типы не нужно: Numba выведет их тип логически, когда впервые будет вызвана функция. Вместо этого вам просто нужно придерживаться правила: использовать только элементарные типы (целые, вещественные и т. д.) и массивы, исключив более сложные объекты Python. В этих случаях Numba сможет скомпилировать программу на Python в очень эффективный код и ускорить вычисления на порядки.

Numba по-прежнему остается очень молодым JIT-компилятором Python, но уже очень полезен. Немаловажно: Numba показывает, какие существуют возможности при реализации приложений на основе JIT-компиляторов. А они, судя по всему, собираются стать более распространенным явлением: в Python 3.6 добавлены возможности, облегчающие использование новых JIT-компиляторов (Pyjion JIT основан как раз на них). В блоге Хуана<sup>2</sup> вы сможете увидеть некоторые примеры использования Numba, включая те, которые показывают, как его объединять с SciPy. И разумеется, у Numba есть свой собственный, очень активный и дружелюбный список рассылки.

## СОДЕЙСТВИЕ ЭТОЙ КНИГЕ

Непосредственно сам источник книги размещен на GitHub<sup>3</sup> (а также на веб-сайте Elegant SciPy<sup>4</sup>). Вы точно так, как если бы вы участвовали в любом другом проекте с открытым исходным кодом, можете поднимать любые вопросы или,

<sup>1</sup> См. <http://shop.oreilly.com/product/0636920033431.do>.

<sup>2</sup> См. <https://ilovesymposia.com/tag/numba/>.

<sup>3</sup> См. <https://github.com/elegant-scipy/elegant-scipy>.

<sup>4</sup> См. <http://elegant-scipy.org/>.

если найдете ошибки или опечатки, отправлять любые запросы на включение внесенных изменений. Мы будем вам весьма признательны за это.

Мы, чтобы проиллюстрировать различные части библиотек SciPy и NumPy, использовали самые лучшие примеры программного кода, которые смогли найти. Если у вас есть более оптимальный пример, пожалуйста, поднимите вопрос в хранилище. Мы с удовольствием включим его в будущие издания книги.

Мы также находимся в Twitter по хештегу @elegantsciPy. Напишите нам, если захотите поговорить о книге! Каждого автора по отдельности вы найдете по хештегам @jnuneziglesias, @stefanvdwalt и @hdashnow.

Нам очень хотелось бы услышать, используете ли вы какую-либо из идей или какой-либо из фрагментов кода из этой книги для совершенствования своего научного исследования. Ведь в этом и заключается суть SciPy!

## До следующей встречи...

Тем временем мы надеемся, что эта книга вам понравилась и вы нашли ее полезной для себя. Если это так, то расскажите об этом всем своим друзьям и заходите, чтобы поздороваться в списках рассылки, на конференции, на GitHub и в Twitter. Спасибо за чтение, и пусть будет еще более Элегантный SciPy!




### РЕШЕНИЕ: ДОБАВЛЕНИЕ СЕТОЧНОГО НАЛОЖЕНИЯ

Здесь показано решение задачи «Добавление сеточного наложения» главы 3.

Мы можем применить сразу NumPy, чтобы отобрать строки сетки, назначить им синий цвет и затем отобрать столбцы и тоже назначить им синий цвет (рис. A1):

```
def overlay_grid(image, spacing=128):  
    «»»Вернуть изображение с сеточным наложением, используя  
    предоставленный интервал.  
  
    Параметры  
    -----  
    image : массив, форма (M, N, 3)  
           Входное изображение.  
    spacing : целое  
            Интервал между линиями сетки.  
  
    Возвращает  
    -----  
    image_gridded : массив, форма (M, N, 3)  
                  Исходное изображение с наложенной синей сеткой.  
    .....
```



```
    image_gridded = image.copy()  
    image_gridded[spacing:-1:spacing, :] = [0, 0, 255]  
    image_gridded[:, spacing:-1:spacing] = [0, 0, 255]  
    return image_gridded  
  
plt.imshow(overlay_grid(astro, 128));
```

Обратите внимание, согласно стандартной индексации в Python, мы использовали -1, имея в виду последнее значение оси. Вы можете это значение опустить, но тогда смысл будет немного отличаться. Без него (т. е. `spacing::spacing`) вы пройдете до самого конца массива, включая заключительную строку/столбец. Когда вы используете его в качестве индекса остановки, то препятствуете отбору заключительной строки. В случае наложения сетки такое поведение, вероятно, будет желательным.

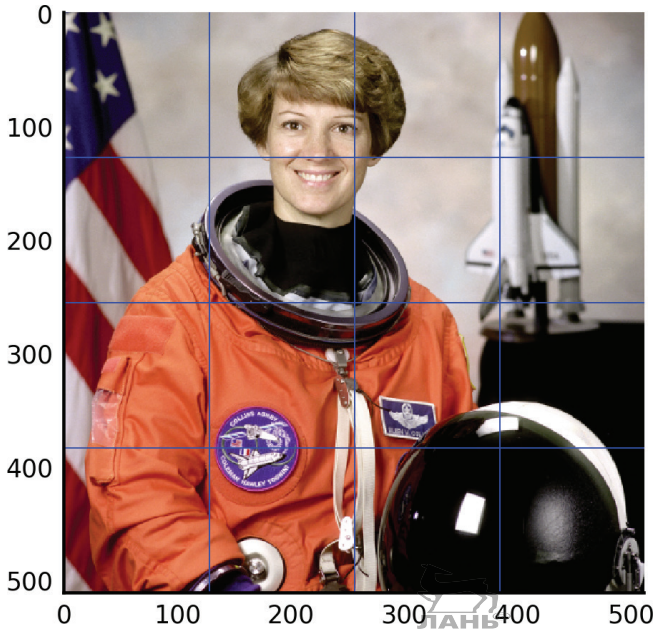


Рис. А1 ❖ Изображение астронавта с наложенной сеткой

## РЕШЕНИЕ: ИГРА «ЖИЗНЬ» КОНУЭЯ»

Здесь показано решение задачи «Игра «Жизнь» Конуэя» из главы 3.

На своей странице со 100 упражнениями NumPy<sup>1</sup> в упражнении под номером 79 Николас Ругир<sup>2</sup> (@NPRougier) предлагает решение с использованием только NumPy:

```
def next_generation(Z):
    N = (Z[0:-2,0:-2] + Z[0:-2,1:-1] + Z[0:-2,2:] +
         Z[1:-1,0:-2] + Z[1:-1,2:] +
         Z[2:,0:-2] + Z[2:,1:-1] + Z[2:,2:])

    # Применить правила
    birth = (N==3) & (Z[1:-1,1:-1]==0)
    survive = ((N==2) | (N==3)) & (Z[1:-1,1:-1]==1)
    Z[...] = 0
    Z[1:-1,1:-1][birth | survive] = 1
    return Z
```

Затем мы можем начать доску так:

```
random_board = np.random.randint(0, 2, size=(50, 50))
n_generations = 100
```

<sup>1</sup> См. <http://www.labri.fr/perso/nrougier/teaching/numpy.100/>.

<sup>2</sup> См. <https://github.com/rougier>.

```
for generation in range(n_generations):
    random_board = next_generation(random_board)
```

Использование универсального фильтра `generic_filter` делает это еще проще:

```
def nextgen_filter(values):
    center = values[len(values) // 2]
    neighbors_count = np.sum(values) - center
    if neighbors_count == 3 or (center and neighbors_count == 2):
        return 1.
    else:
        return 0.

def next_generation(board):
    return ndi.generic_filter(board, nextgen_filter,
                             size=3, mode='constant')
```

Замечательно то, что в некоторых формулировках игры «Жизнь» используются так называемая *тороидальная доска*, а именно левый и правый концы доски «циклически переносятся» и соединяются друг с другом. То же самое происходит с верхним и нижним концами. С участием `generic_filter` модификация нашего решения для ее включения реализуется обычным образом:

```
def next_generation_toroidal(board):
    return ndi.generic_filter(board, nextgen_filter,
                             size=3, mode='wrap')
```

Теперь можно просимулировать эту тороидальную доску для нескольких поколений:

```
random_board = np.random.randint(0, 2, size=(50, 50))
n_generations = 100
for generation in range(n_generations):
    random_board = next_generation_toroidal(random_board)
```



## РЕШЕНИЕ: МАГНИТУДА ГРАДИЕНТА СОБЕЛА

Здесь показано решение задачи «Магнитуда градиента Собела» из главы 3:

```
hsobel = np.array([[ 1, 2, 1],
                  [ 0, 0, 0],
                  [-1, -2, -1]])

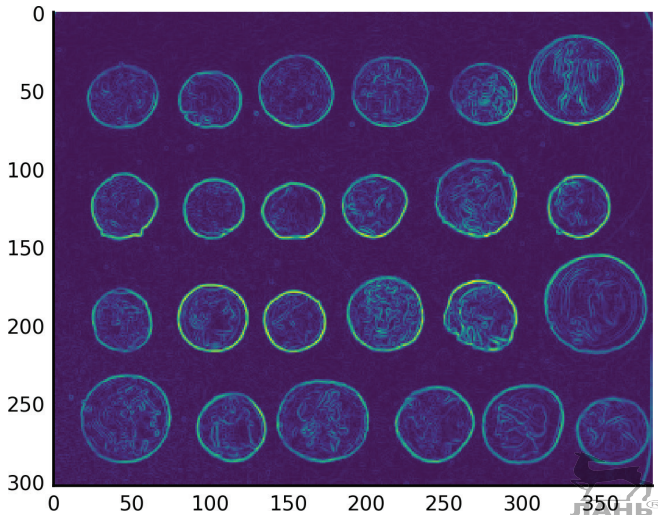
vsobel = hsobel.T

hsobel_r = np.ravel(hsobel)
vsobel_r = np.ravel(vsobel)

def sobel_magnitude_filter(values):
    h_edge = values @ hsobel_r
    v_edge = values @ vsobel_r
    return np.hypot(h_edge, v_edge)
```

Теперь мы можем проверить его на изображении монет:

```
sobel_mag = ndi.generic_filter(coins, sobel_magnitude_filter, size=3)
plt.imshow(sobel_mag, cmap='viridis');
```



## РЕШЕНИЕ: ПОДБОР КРИВОЙ ПРИ ПОМОЩИ SciPy

Здесь показано решение задачи «Подбор кривой с SciPy» из главы 3.

Давайте посмотрим на начало описания функции `curve_fit` в строке документации Python docstring:

Использует нелинейные наименьшие квадраты для подгонки функции,  $f$ , к данным.

Принимает `'ydata = f(xdata, *params) + eps'`

Параметры

-----

$f$ : вызываемая функция

Модельная функция,  $f(x, \dots)$ . Она должна принимать независимую переменную в качестве первого аргумента и подгоняемые параметры в качестве отдельных остальных аргументов.

$xdata$ : последовательность длины  $M$  или массив формы  $(k, M)$  для функций с  $k$  предикторами.

Независимая переменная, где данные подлежат измерению.

$ydata$ : последовательность длины  $M$

Зависимые данные --- номинально  $f(xdata, \dots)$

Похоже, что нам просто нужно предоставить функцию, принимающую точку данных и несколько параметров, и вернуть предсказанное значение. В нашем случае мы хотим, чтобы кумулятивная оставшаяся частота,  $f(d)$ , была пропорциональной  $d^{-\gamma}$ . Это означает, что нам нужно  $f(d) = \alpha d^{-\text{gamma}}$ .

```
def fraction_higher(degree, alpha, gamma):
    return alpha * degree ** (-gamma)
```



Затем нам нужны данные  $X$  и  $Y$ , под которые выполняется подгонка, для  $d > 10$ :

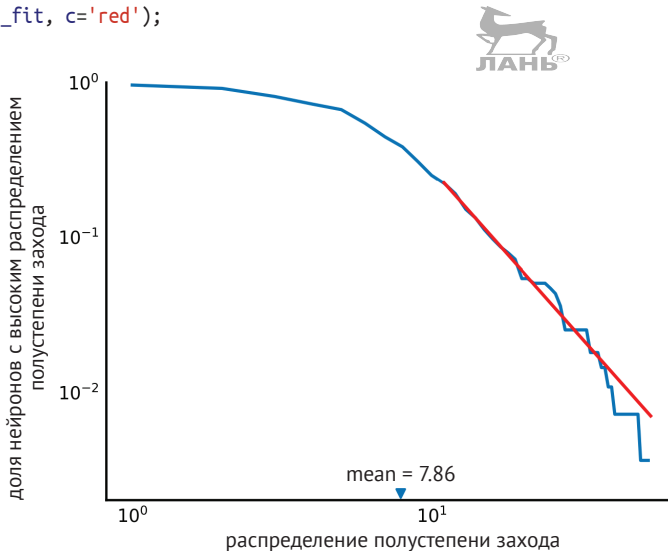
```
x = 1 + np.arange(len(survival))
valid = x > 10
x = x[valid]
y = survival[valid]
```

Теперь можно применить функцию `curve_fit`, чтобы получить подогнанные параметры:

```
from scipy.optimize import curve_fit
alpha_fit, gamma_fit = curve_fit(fraction_higher, x, y)[0]
```

Давайте выведем результаты на график, чтобы увидеть состояние дел:

```
y_fit = fraction_higher(x, alpha_fit, gamma_fit)
fig, ax = plt.subplots()
ax.loglog(np.arange(1, len(survival) + 1), survival)
ax.set_xlabel('распределение полустепени захода')
ax.set_ylabel('доля нейронов с высоким распределением полустепени захода')
ax.scatter(avg_in_degree, 0.0022, marker='v')
ax.text(avg_in_degree - 0.5, 0.003, 'среднее=%2f' % avg_in_degree)
ax.set_ylim(0.002, 1.0)
ax.loglog(x, y_fit, c='red');
```



Вуаля! Полный рисунок 6В и полное соответствие!

## РЕШЕНИЕ: СВЕРТЫВАНИЕ ИЗОБРАЖЕНИЯ

Здесь показано решение задачи «Свертывание изображения» из главы 4.

```
from scipy import signal

x = np.random.random((50, 50))
y = np.ones((5, 5))

L = x.shape[0] + y.shape[0] - 1
Px = L - x.shape[0]
Py = L - y.shape[0]

xx = np.pad(x, ((0, Px), (0, Px)), mode='constant')
yy = np.pad(y, ((0, Py), (0, Py)), mode='constant')

zz = np.fft.ifft2(np.fft.fft2(xx) * np.fft.fft2(yy)).real
print('Результирующая форма:', zz.shape, ' <-- Почему?')

z = signal.convolve2d(x, y)
print('Результаты равны?', np.allclose(zz, z))

Результирующая форма: (54, 54) <-- Почему?
```



Результаты равны? True

## РЕШЕНИЕ: ВЫЧИСЛИТЕЛЬНАЯ СЛОЖНОСТЬ МАТРИЦ ОШИБОК

Здесь показано решение задачи «Вычислительная сложность матриц ошибок» из главы 5.

Из главы 1 вы помните, что выражение `agg == k` создает массив булевых (True или False) значений того же самого размера, что и массив `agg`. Это, как можно ожидать, требует полного обхода массива `agg`. Поэтому в приведенном выше решении мы выполняем обход массивов `pred` и `gt` для каждой комбинации значений в `pred` и `gt`. По сути дела, мы можем вычислить `cont` всего за один обход обоих массивов. Поэтому многократные обходы неэффективны.

## РЕШЕНИЕ: АЛЬТЕРНАТИВНЫЙ АЛГОРИТМ ВЫЧИСЛЕНИЯ МАТРИЦЫ ОШИБОК

Здесь показано решение задачи «Альтернативный алгоритм вычисления матрицы ошибок» главы 5.

Мы предлагаем два решения, хотя решений может быть много.

В первом решении используется встроенная в Python функция `zip`, чтобы выдавать попарно метки из массивов `pred` и `gt`.

```
def confusion_matrix1(pred, gt):
    cont = np.zeros((2, 2))
    for i, j in zip(pred, gt):
        cont[i, j] += 1
    return cont
```



Второе решение состоит в том, чтобы перебрать все возможные индексы массивов `pred` и `gt` в цикле и вручную извлечь соответствующее значение из каждого массива:

```
def confusion_matrix2(pred, gt):
    cont = np.zeros((2, 2))
    for idx in range(len(pred)):
        i = pred[idx]
        j = gt[idx]
        cont[i, j] += 1
    return cont
```



Первый вариант может рассматриваться как более Python'овский из двух, но второй легче ускорить путем трансляции и компиляции на таких языках или в таких инструментах, как C, Cython и Numba (это тема для еще одной книги).

## РЕШЕНИЕ: ВЫЧИСЛЕНИЕ МАТРИЦЫ ОШИБОК

Здесь показано решение задачи «Мультиклассовая матрица ошибок» главы 5.

Чтобы определить максимальную метку, нужно выполнить только первоначальный обход обоих входных массивов. Затем, чтобы учесть нулевую метку и индексацию Python с отсчетом от 0, мы добавляем в нее 1. Далее создаем матрицу и заполняем ее так, как было выше:

```
def general_confusion_matrix(pred, gt):
    n_classes = max(np.max(pred), np.max(gt)) + 1
    cont = np.zeros((n_classes, n_classes))
    for i, j in zip(pred, gt):
        cont[i, j] += 1
    return cont
```



## РЕШЕНИЕ: ПРЕДСТАВЛЕНИЕ В ФОРМАТЕ COO

Здесь показано решение задачи «Представление в формате COO» из главы 5.

Сначала мы формируем перечень ненулевых элементов массива, слева направо и сверху вниз, как при чтении книги:

```
s2_data = np.array([6, 1, 2, 4, 5, 1, 9, 6, 7])
```

Затем формируем перечень строчных индексов этих значений в том же самом порядке:

```
s2_row = np.array([0, 1, 1, 1, 1, 2, 3, 4, 4])
```

И наконец, столбцовых индексов:

```
s2_col = np.array([2, 0, 1, 3, 4, 1, 0, 3, 4])
```

Проверив их равенство в обоих направлениях, можно легко убедиться, что они порождают правильную матрицу:

```
s2_coo0 = sparse.coo_matrix(s2)
print(s2_coo0.data)
print(s2_coo0.row)
print(s2_coo0.col)
```

```
[6 1 2 4 5 1 9 6 7]
[0 1 1 1 1 2 3 4 4]
[2 0 1 3 4 1 0 3 4]
```

и:

```
s2_coo1 = sparse.coo_matrix((s2_data, (s2_row, s2_col)))
print(s2_coo1.toarray())
```

```
[[0 0 6 0 0]
 [1 2 0 4 5]
 [0 1 0 0 0]
 [9 0 0 0 0]
 [0 0 0 6 7]]
```



## РЕШЕНИЕ: ПОВОРОТ ИЗОБРАЖЕНИЯ

Здесь показано решение задачи «Поворот изображения» главы 5.

Мы можем *скомпоновать* преобразования путем их умножения. Мы знаем, как поворачивать изображение вокруг начала координат, а также как двигать его в разные стороны. Поэтому сдвинем изображение так, чтобы его центр находился в начале координат, повернем его и затем сдвинем назад.

```
def transform_rotate_about_center(shape, degrees):
    «»Вернуть гомографическую матрицу для поворота
    вокруг центра изображения.»»
    c = np.cos(np.deg2rad(angle))
    s = np.sin(np.deg2rad(angle))

    H_rot = np.array([[c, -s, 0],
                     [s, c, 0],
                     [0, 0, 1]])

    # Вычислить координаты центра изображения
    center = np.array(image.shape) / 2
    # Матрица для центрирования изображения по началу координат
    H_tr0 = np.array([[1, 0, -center[0]],
                    [0, 1, -center[1]],
                    [0, 0, 1]])

    # Матрица для перемещения центра назад
    H_tr1 = np.array([[1, 0, center[0]],
                    [0, 1, center[1]],
                    [0, 0, 1]])

    # Полная матрица преобразования
    H_rot_cent = H_tr1 @ H_rot @ H_tr0

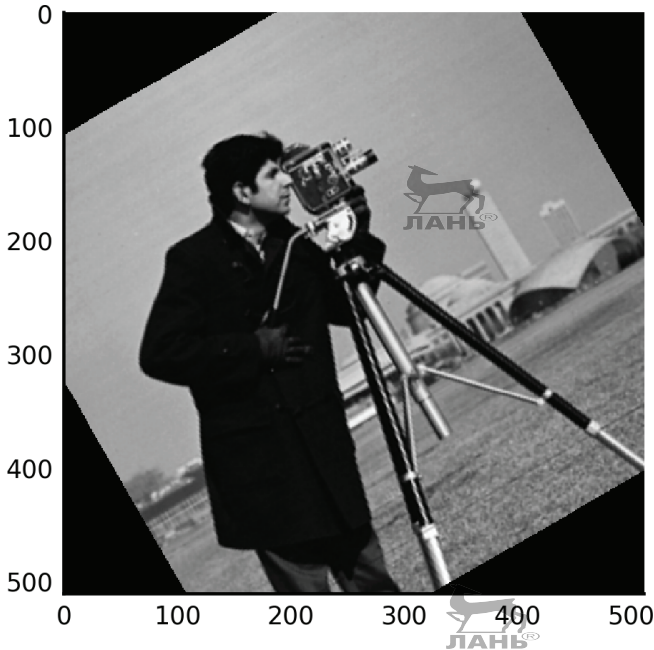
    sparse_op = homography(H_rot_cent, image.shape)

    return sparse_op
```



Можно проверить, что все работает:

```
tf = transform_rotate_about_center(image.shape, 30)
plt.imshow(apply_transform(image, tf));
```



## РЕШЕНИЕ: СОКРАЩЕНИЕ ОБЪЕМА ПОТРЕБЛЯЕМОЙ ОПЕРАТИВНОЙ ПАМЯТИ

Здесь показано решение задачи «Сокращение объема потребляемой оперативной памяти» из главы 5.

Создаваемый нами массив `np.ones` предназначен только для чтения: он будет использоваться лишь как значения, суммируемые функцией `coo_matrix`. Мы можем применить функцию `broadcast_to`, чтобы создать аналогичный массив всего с одним элементом, «виртуально» повторенным  $n$  раз:

```
def confusion_matrix(pred, gt):
    n = pred.size
    ones = np.broadcast_to(1., n) # виртуальный массив единиц размера n
    cont = sparse.coo_matrix((ones, (pred, gt)))
    return cont
```

Удостоверимся, что все по-прежнему работает ожидаемым образом:

```
cont = confusion_matrix(pred, gt)
print(cont.toarray())
```

```
[[ 3.  1.]
 [ 2.  4.]]
```

Ба-бах! Вместо того чтобы создавать такой же большой массив, что и исходные данные, мы просто создаем массив размера 1. По мере обработки все более крупных наборов данных такая оптимизация приобретает все более важное значение.

## РЕШЕНИЕ: ВЫЧИСЛЕНИЕ УСЛОВНОЙ ЭНТРОПИИ

Чтобы получить таблицу совместных вероятностей, мы просто делим таблицу на ее общее количество – в данном случае на 12:

Здесь показано решение задачи «Вычисление условной энтропии» из главы 5.

```
print('сумма таблицы:', np.sum(p_rain_g_month))
p_rain_month = p_rain_g_month / np.sum(p_rain_g_month)
```

```
table total: 12.0
```

Теперь мы можем вычислить условную энтропию месяца при условии дождя. (Это похоже на вопрос: если мы знаем, что идет дождь, сколько еще нам нужно узнать информации, чтобы выяснить, какой это месяц, в среднем?)

```
p_rain = np.sum(p_rain_month, axis=0)
p_month_g_rain = p_rain_month / p_rain
Hmr = np.sum(p_rain * p_month_g_rain * np.log2(1 / p_month_g_rain))
print(Hmr)
```

```
3.5613602411
```

Сравним этот результат с энтропией месяцев:

```
p_month = np.sum(p_rain_month, axis=1) # 1/12, но этот метод более общий
Hm = np.sum(p_month * np.log2(1 / p_month))
print(Hm)
```

```
3.58496250072
```

Таким образом, мы видим, что сведения о том, шел ли сегодня дождь, приблизили нас на две сотых бита к тому, чтобы угадать, какой это месяц! Не стоит ставить все на эту догадку.

## РЕШЕНИЕ: МАТРИЦА ПОВОРОТА

Здесь показано решение задачи «Матрица поворота» из главы 6.

### Часть 1

```
import numpy as np

theta = np.deg2rad(45)
R = np.array([[np.cos(theta), -np.sin(theta), 0],
              [np.sin(theta), np.cos(theta), 0],
              [0, 0, 1]])
```

```
print("Произведение R на ось X:", R @ [1, 0, 0])
print("Произведение R на ось Y:", R @ [0, 1, 0])
print("Произведение R на 45-градусный вектор:", R @ [1, 1, 0])
```

Произведение R на ось X: [ 0.70710678 0.70710678 0. ]

Произведение R на ось Y: [-0.70710678 0.70710678 0. ]

Произведение R на 45-градусный вектор: [ 1.11022302e-16 1.41421356e+00 0.00000000e+00]



### Часть 2

Поскольку умножение вектора на  $R$  поворачивает его на 45 градусов, повторное умножение результата на  $R$  должно привести к тому, что исходный вектор будет повернут на 90 градусов. Умножение матриц ассоциативно, имея в виду, что  $R(Rv) = (RR)v$ , поэтому  $S = RR$  должно поворачивать векторы на 90 градусов вокруг оси  $z$ .

$S = R @ R$

$S @ [1, 0, 0]$

array([ 2.22044605e-16, 1.00000000e+00, 0.00000000e+00])

### Часть 3

```
print("R @ ось Z:", R @ [0, 0, 1])
```

R @ ось Z: [ 0. 0. 1.]

$R$  поворачивает ось  $x$  и ось  $y$ , но не ось  $z$ .

### Часть 4

Если взглянуть на документацию по функции `eig`, то увидим, что она возвращает два значения: одномерный массив собственных значений и двумерный массив, в котором каждый столбец содержит собственный вектор, соответствующий каждому собственному значению.

```
np.linalg.eig(R)
```

```
(array([ 0.70710678+0.70710678j, 0.70710678-0.70710678j, 1.00000000+0.j ]),
 array([[ 0.70710678+0.j           , 0.70710678-0.j           , 0.00000000+0.j ],
        [ 0.00000000-0.70710678j, 0.00000000+0.70710678j, 0.00000000+0.j ],
        [ 0.00000000-0.j           , 0.00000000+0.j           , 1.00000000+0.j ]]))
```

В дополнение к нескольким комплексным собственным значениям и векторам мы видим значение, связанное с вектором  $[0, 0, 1]^T$ .

## РЕШЕНИЕ: ИЗОБРАЖЕНИЕ АФФИННОГО ПОДОБИЯ

Здесь показано решение задачи «Изображение аффинного подобия» из главы 6.

В аффинном представлении вместо использования обработки на оси  $y$  мы используем нормализованный третий собственный вектор  $Q$  точно так же, как поступили с  $x$ . (И при необходимости мы его инвертируем точно так же, как поступили с  $x$ !)

```

y = Dinv2 @ Vec[:, 2]
asjl_index = np.argwhere(neuron_ids == 'ASJL')
if y[asjl_index] < 0:
    y = -y

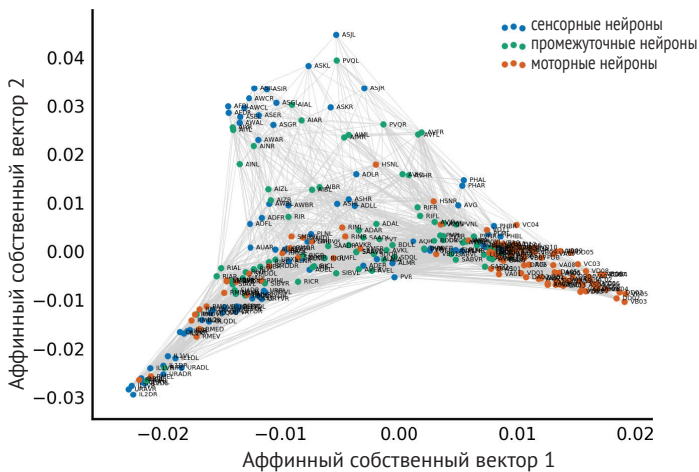
```



```

plot_connectome(x, y, C, labels=neuron_ids, types=neuron_types,
type_names=['сенсорные нейроны', 'промежуточные нейроны',
            'моторные нейроны'],
xlabel='Аффинный собственный вектор 1',
ylabel='Аффинный собственный вектор 2')

```



## РЕШЕНИЕ: ЛИНЕЙНАЯ АЛГЕБРА С РАЗРЕЖЕННЫМИ МАТРИЦАМИ

Здесь показано решение задачи «Линейная алгебра с разреженными матрицами» из главы 6.

Для целей этой задачи мы используем небольшой коннектом, потому что так легче визуализировать то, что происходит. В последующих частях упражнения мы будем использовать эти методы для анализа более крупных сетей.

Прежде всего начинаем с матрицы смежности,  $A$ , в формате разреженной матрицы, в данном случае в формате CSR, т. е. в формате, который наиболее распространен в линейно-алгебраических задачах. К именам всех матриц добавим букву «s» (от sparse), тем самым обозначив, что они являются разреженными.

```

from scipy import sparse
import scipy.sparse.linalg

```

```
As = sparse.csr_matrix(A)
```

Мы можем создать матрицу связности таким же образом:

```
Cs = (As + As.T) / 2
```



Чтобы получить степенную матрицу, можем применить разреженный диагональный формат «diags», в котором хранятся диагональные и внедиагональные матрицы.

```
degrees = np.ravel(Cs.sum(axis=0))
Ds = sparse.diags(degrees)
```

Матрица Лапласа выводится прямолинейно:

```
Ls = Ds - Cs
```

Теперь следует получить глубину обработки. Напомним, что о получении псевдоинверсии лапласовой матрицы не может быть и речи, так как это будет плотная матрица (инверсия разреженной матрицы в целом не является разреженной как таковой). Однако мы фактически использовали псевдоинверсию, чтобы вычислить вектор  $z$ , удовлетворяющий уравнению  $Lz = b$ , где  $b = c \odot \text{знак}(A - A^T)1$ . (Вы можете это увидеть в дополнительном материале, прилагаемом к работе Варшни и др.) В случае плотных матриц мы можем просто применить  $z = L^+b$ . В случае разреженных, тем не менее, можно применить один из решателей (см. в главе 6 раздел «Решатели») в модуле `sparse.linalg.isolve`. Чтобы получить вектор  $z$  после предоставления  $L$  и  $b$ , никакой инверсии не требуется!

```
b = Cs.multiply((As - As.T).sign()).sum(axis=1)
z, error = sparse.linalg.isolve.cg(Ls, b, maxiter=10000)
```



Наконец, следует найти собственные векторы матрицы  $Q$ , нормализованной по степени лапласовой матрицы, соответствующей ее второму и третьему наименьшим собственным значениям.

Из главы 5 вы, скорее всего, помните, что числовые данные в разреженных матрицах находятся в атрибуте `.data`. Мы используем его, чтобы инвертировать степенную матрицу:

```
Dsinv2 = Ds.copy()
Dsinv2.data = 1 / np.sqrt(Ds.data)
```

Наконец, используем линейно-алгебраические функции из модуля SciPy `sparse`, чтобы отыскать требующиеся собственные векторы. Матрица  $Q$  симметрична, поэтому можем применить функцию `eigsh`, специально предназначенную для вычисления симметрических матриц. Используем именованный аргумент `which`, чтобы указать, что нам необходимо получить собственные векторы, которые соответствуют наименьшим собственным значениям, и  $k$  и указать, что нам нужны три наименьших:

```
Qs = Dsinv2 @ Ls @ Dsinv2
vals, Vecs = sparse.linalg.eigsh(Qs, k=3, which='SM')
sorted_indices = np.argsort(vals)
Vecs = Vecs[:, sorted_indices]
```

Наконец, чтобы получить координаты  $x$  и  $y$  (и при необходимости меняем их местами), нормализуем собственные векторы:



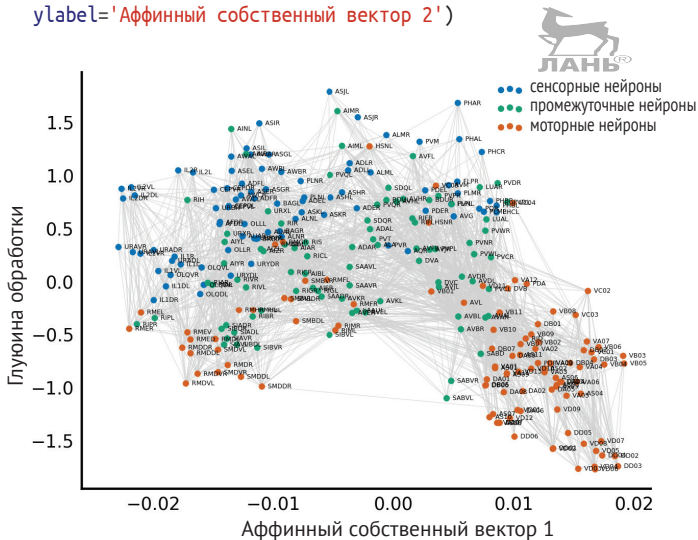
```
_dsinv, x, y = (Dsinv2 @ Vecs).T
vc2_index = np.argwhere(neuron_ids == 'VC02')
```

```
if x[vc2_index] < 0:
    x = -x
if y[asj_l_index] < 0:
    y = -y
```

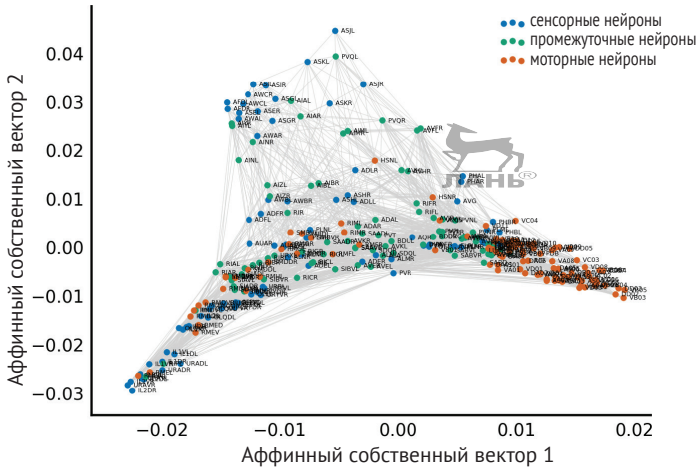
(Обратите внимание, собственный вектор, соответствующий наименьшему собственному значению, всегда является вектором единиц, который нас не интересует.) Теперь можно воспроизвести следующие ниже графики!

```
plot_connectome(x, z, C, labels=neuron_ids, types=neuron_types,
               type_names=['сенсорные нейроны', 'промежуточные нейроны',
                           'моторные нейроны'],
               xlabel='Аффинный собственный вектор 1',
               ylabel='Глубина обработки')
```

```
plot_connectome(x, y, C, labels=neuron_ids, types=neuron_types,
               type_names=['сенсорные нейроны', 'промежуточные нейроны',
                           'моторные нейроны'],
               xlabel='Аффинный собственный вектор 1',
               ylabel='Аффинный собственный вектор 2')
```







## РЕШЕНИЕ: ОБРАБОТКА ВИСЯЧИХ УЗЛОВ

Здесь показано решение задачи «Обработка висячих узлов» из главы 6.

Чтобы получить стохастическую матрицу, в сумме все столбцы матрицы переходов должны составлять 1. Это требование не удовлетворяется, когда организм не служит пищей для других: этот столбец будет состоять из одних нулей. Вместе с тем замена всех этих столбцов на  $1/n$  будет дорогостоящей.

Ключ к решению этой задачи состоит в том, чтобы осознать: *каждая строка* вносит *одинаковый вклад* в умножение матрицы переходов на вектор текущих вероятностей. Иными словами, сложение этих столбцов добавит единое значение к результату итеративного умножения. Каково это значение?  $1/n$ , умноженное на элементы  $r$ , которые соответствуют висячему узлу. Это может быть выражено как скалярное произведение вектора, содержащего  $1/n$  для позиций, соответствующих висячим узлам, и нуль во всех остальных случаях, на вектор  $r$  для текущей итерации.

```
def power2(Trans, damping=0.85, max_iter=10**5):
    n = Trans.shape[0]
    dangling = (1/n) * np.ravel(Trans.sum(axis=0) == 0)
    r0 = np.full(n, 1/n)
    r = r0
    for _ in range(max_iter):
        rnext = (damping * (Trans @ r + dangling @ r) +
                (1 - damping) / n)
        if np.allclose(rnext, r):
            return rnext
    else:
        r = rnext
    return r
```

Попробуйте это выполнить вручную для нескольких итераций. Обратите внимание, если вы начинаете со стохастического вектора (вектора, все элементы которого в сумме составляют 1), то следующий вектор по-прежнему будет стохастическим вектором. Таким образом, получаемая на выходе из этой функции мера PageRank будет вектором истинных вероятностей, и значения представят вероятность, что мы, следуя по ссылкам в пищевой цепи, придем к конкретному организму.



## РЕШЕНИЕ: МЕТОДЫ ПРОВЕРКИ

Здесь показано решение задачи «Эквивалентность различных методов получения собственного вектора» из главы 6.

Функция `np.corrcoef` дает коэффициент корреляции Пирсона между всеми парами списка векторов. Этот коэффициент будет равен 1, если и только если два вектора являются скалярными кратными друг друга. Поэтому коэффициент корреляции, равный 1, будет достаточен, чтобы показать, что вышеупомянутые методы порождают одинаковое ранжирование.

```
pagerank_power = power(Trans)
pagerank_power2 = power2(Trans)
np.corrcoef([pagerank, pagerank_power, pagerank_power2])
array([[ 1.,  1.,  1.],
       [ 1.,  1.,  1.],
       [ 1.,  1.,  1.]])
```

## РЕШЕНИЕ: МОДИФИКАЦИЯ ФУНКЦИИ ALIGN

Здесь показано решение задачи «Модификация функции align» главы 7.

Мы используем метод `basin hopping` на более высоких уровнях пирамиды. Однако для более низких уровней используем метод Пауэлла, потому что при полной разрешающей способности выполнение метода `basin hopping` в вычислительном плане обходится слишком дорого:

```
def align(reference, target, cost=cost_mse, nlevels=7, method='Powell'):
    pyramid_ref = gaussian_pyramid(reference, levels=nlevels)
    pyramid_tgt = gaussian_pyramid(target, levels=nlevels)

    levels = range(nlevels, 0, -1)
    image_pairs = zip(pyramid_ref, pyramid_tgt)

    p = np.zeros(3)

    for n, (ref, tgt) in zip(levels, image_pairs):
        p[1:] *= 2
        if method.upper() == 'BH':
            res = optimize.basinhopping(cost, p,
                                       minimizer_kwargs={'args': (ref, tgt)})
```

```

if n <= 4: # avoid basin hopping in lower levels
    method = 'Powell'
else:
    res = optimize.minimize(cost, p, args=(ref, tgt), method='Powell')
p = res.x
# print current level, overwriting each time (like a progress bar)
print(f'Уровень: {n}, Угол: {np.rad2deg(res.x[0]) :.3}, '
      f'Сдвиг: ({res.x[1] * 2**n :.3}, {res.x[2] * 2**n :.3}), '
      f'Стоимость: {res.fun :.3}', end='\r')

print(') # новая строка, когда выравнивание завершено
return make_rigid_transform(p)

```



Теперь попробуем это выравнивание:

```

from skimage import util

theta = 50
rotated = transform.rotate(astronaut, theta)
rotated = util.random_noise(rotated, mode='gaussian',
                           seed=0, mean=0, var=1e-3)

tf = align(astronaut, rotated, nlevels=8, method='BH')
corrected = transform.warp(rotated, tf, order=3)

f, (ax0, ax1, ax2) = plt.subplots(1, 3)
ax0.imshow(astronaut)
ax0.set_title('Оригинальное')
ax1.imshow(rotated)
ax1.set_title('Повернутое')
ax2.imshow(corrected)
ax2.set_title('Зарегистрированное')
for ax in (ax0, ax1, ax2):
    ax.axis('off')

```



Уровень: 1, Угол: -50.0, Сдвиг: (-2.09e+02, 5.74e+02), Стоимость: 0.0385

Оригинальное



Повернутое



Зарегистрированное



Успешно! Метод basin hopping смог восстановить правильное выравнивание даже в проблематичном случае, где функция minimize попала в тупик.

## РЕШЕНИЕ: АНАЛИЗ ГЛАВНЫХ КОМПОНЕНТ ПОТОКОВЫХ ДАННЫХ ПРИ ПОМОЩИ БИБЛИОТЕКИ SCIKIT-LEARN

Здесь показано решение задачи «Анализ главных компонент потоковых данных» из главы 8.

Сначала пишется функция для тренировки модели. Эта функция должна принимать поток образцов и производить модель PCA, которая способна преобразовывать новые образцы, проецируя их из исходного  $n$ -мерного пространства в пространство главной компоненты.

```
import toolz as tz
from toolz import curried as c
from sklearn import decomposition
from sklearn import datasets
import numpy as np

def streaming_pca(samples, n_components=2, batch_size=100):
    ipca = decomposition.IncrementalPCA(n_components=n_components,
                                       batch_size=batch_size)
    tz.pipe(samples,
            c.partition(batch_size), # итератор кортежей
            c.map(np.array),        # итератор двумерных массивов
            c.map(ipca.partial_fit), # частичная подгонка partial_fit по каждому
            tz.last)               # Пропустить поток данных через конвейер
    return ipca
```

Теперь можно применить эту функцию, чтобы *натренировать* (или выполнить *подгонку*) модели PCA:

```
reshape = tz.curry(np.reshape)

def array_from_txt(line, sep=',', dtype=np.float):
    return np.array(line.rstrip().split(sep), dtype=dtype)

with open('data/iris.csv') as fin:
    pca_obj = tz.pipe(fin, c.map(array_from_txt), streaming_pca)
```

Наконец, исходные образцы направляются потоком через модельную функцию `transform`. Мы укладываем их воедино, чтобы получить матрицу данных, состоящую из  $n$  образцов на  $n$  компонент ( $n\_samples \times n\_components$ ):

```
with open('data/iris.csv') as fin:
    components = tz.pipe(fin,
                        c.map(array_from_txt),
                        c.map(reshape(newshape=(1, -1))),
                        c.map(pca_obj.transform),
                        np.vstack)

print(components.shape)

(150, 2)
```

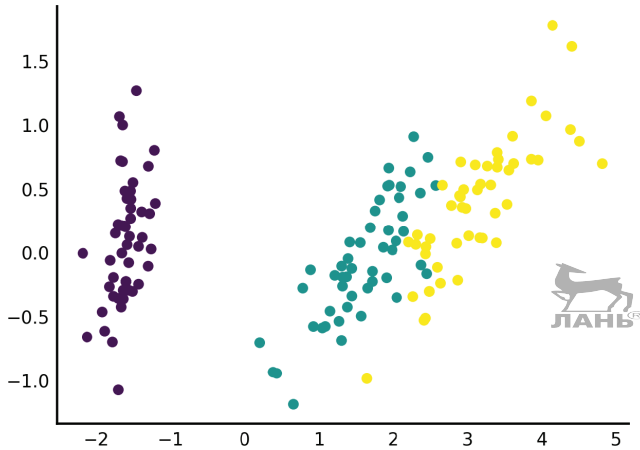


Теперь компоненты можно вывести на график:

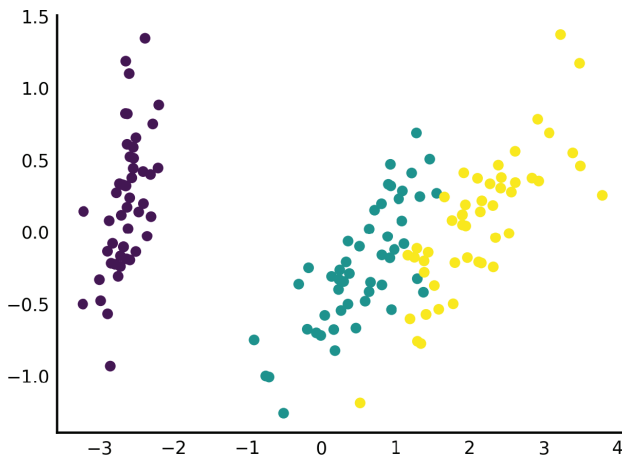
```
iris_types = np.loadtxt('data/iris-target.csv')
plt.scatter(*components.T, c=iris_types, cmap='viridis');
```

Вы можете проверить, что это дает (приблизительно) такой же результат, что и стандартный PCA (сравните рис. А2 и А3):

```
iris = np.loadtxt('data/iris.csv', delimiter=',')
components2 = decomposition.PCA(n_components=2).fit_transform(iris)
plt.scatter(*components2.T, c=iris_types, cmap='viridis');
```



**Рис. А2** ❖ Главные компоненты набора данных цветков ириса, вычисленных при помощи потокового PCA



**Рис. А3** ❖ Главные компоненты набора данных цветков ириса, вычисленные при помощи обычного PCA

Разница, конечно же, состоит в том, что потоковый анализ главных компонент масштабируется до чрезвычайно больших наборов данных.

## РЕШЕНИЕ: ДОБАВЛЕНИЕ ШАГА В НАЧАЛО КОНВЕЙЕРА

Здесь показано решение задачи «Онлайновая распаковка данных» из главы 8.

Функция `open` в первоначальном исходном коде `genome` может быть заменена на каррированную версию `gzip.open`. Функция `open` в библиотеке `gzip` по умолчанию использует режим `rb` (read bytes, т. е. читать байты) вместо встроенной в Python функции `open`, которая использует режим `rt` (read text, т. е. читать текст). Поэтому мы должны ее предоставить.

```
import gzip

gzopen = tz.curry(gzip.open)

def genome_gz(file_pattern):
    «»»Передать геном потоком, буква за буквой, из списка имен файлов FASTA.»»»
    return tz.pipe(file_pattern, glob, sorted, # Имена файлов
                   c.map(gzopen(mode='rt')), # строки
                   # конкатенировать строки из всех файлов:
                   tz.concat,
                   # отбросить заголовок из каждой последовательности
                   c.filter(is_sequence),
                   # конкатенировать символы из всех строк
                   tz.concat,
                   # отбросить символы новой строки и 'N'
                   c.filter(is_nucleotide))
```

Вы можете испытать эту функцию на сжатом файле генома дрозофилы:

```
dm = 'data/dm6.fa.gz'
model = tz.pipe(dm, genome_gz, c.take(10**7), markov)
plot_model(model, labels='ACGTacgt')
```



Если вы хотите иметь единую функцию `genome`, то вы могли бы написать собственную функцию `open`, которая по имени файла или путем проб и ошибок будет решать, является ли файл архивным файлом `gzip`.

Таким же образом, если у вас есть архив `.tar.gz` с файлами FASTA, вместо модуля `glob` вы можете использовать модуль Python `tarfile`, чтобы прочитать каждый файл по отдельности. Единственная оговорка состоит в том, что для декодирования каждой строки вам придется использовать функцию `bytes.decode`, поскольку `tarfile` возвращает строки как байты, а не как текст.



# Предметный указатель



## Символы

`%matplotlib inline`, команда, 47  
`@`, оператор умножения матриц, 145

## В

basin hopping, алгоритм, 204

## С

conda, 23  
Cython, 235

## Ф

FASTA, формат, 221  
fcluster, функция, 76  
FFTPACK, динамическая библиотека, 113  
frequencies, функция, 223

## Г

generic\_filter, функция, 92, 103  
GitHub, 19, 234

## И

imshow, функция, 80  
IncrementalPCA, класс, 227  
IPython, 13

## Ж

ИТ-компиляторы  
(динамические), 236  
Jupyter  
волшебные команды, 46  
команда `%matplotlib inline`, 47  
роль в научной экосистеме Python, 13

## М

Matplotlib  
`%matplotlib inline`, 46  
вывод изображений, 79

модуль `matplotlib.pyplot`, 47  
роль в научной экосистеме языка Python, 13  
создание таблицы стилей, 47  
стилизация, 46  
функция `trisurf`, 141

## Н

ndimage, библиотека, 79, 86, 102  
NetworkX, библиотека, 94  
Numba, 236  
NumPy  
N-мерные массивы, 38  
квантильная нормализация, 62  
метод `.ravel()`, 159  
нормализация данных, 46  
правила векторизации и транслирования, 32, 38, 56, 64, 158  
преобразование в разреженные матрицы, 166  
роль в научной экосистеме языка Python, 13, 32  
создание изображений, 79  
функциональность, связанная с ДПФ, 115  
N-мерные массивы (массивы `ndarray`)  
векторизация, 64  
в сопоставлении со списками Python, 39  
метод `.ravel()`, 159  
особенности, 38, 63  
представление изображений, 79  
преобразование в разреженные матрицы, 166  
транслирование, 41, 56, 158

## О

optimize, модуль, 195



**P**

- PageRank, мера, 187
- pandas, библиотека
  - роль в научной экосистеме Python, 14
  - чтение данных, 43, 64
- Python 3
  - анализ экспрессии генов, 36
  - в сопоставлении с Python 2, 14
  - в сопоставлении с массивами ndarray, 39
  - инсталляция версии 3.6, 22
  - ключевое слово yield, 214
  - оператор умножения матриц, 145
  - только именованные аргументы, 54

**R**

- RAG (графы смежности областей), 78.
- См. Области изображения лапласова матрица графа, 175
- лапласова матрица, нормализованная по степени, 181
- порождение, 78
- сегментация, 98
- узлы и связи, 78

- RPKM (количество прочтений на тысячу оснований экзона на миллион картированных прочтений), 32, 54

**S**

- scikit-image
  - алгоритм SLIC, 100
  - роль в научной экосистеме Python, 14
- scikit-learn
  - класс IncrementalPCA, 227
  - роль в научной экосистеме Python, 14
- SciPy, библиотека
  - альтернативы, 235
  - документация и обучающие пособия, 10
  - значение понятия «элегантный», 9

- квантильная нормализация, 62
- линейная алгебра, 174
- модуль scipy.fftpack, 113
- модуль sparse, 144, 148, 192
- модуль иерархической кластеризации, 69
- оптимизация функций, 193
- преимущества, 13
- разреженные итеративные решатели, 187
- ресурсы
  - GitHub, 19, 234
  - конференции, 16
  - списки рассылок, 22, 233
- роль в научной экосистеме Python, 13
- требуемые предварительные знания, 11
- функция scipy.signal.fftconvolve, 116

**T**

- Toolz, библиотека, 217, 223

**Y**

- yield, ключевое слово, 214

**A**

- Алгоритм Блуштайна, 117
- Алгоритм Кули-Тьюки, 116
- Алгоритм Рейдера, 117
- Алгоритмы линейного поиска, 193
- Анализ потоковых данных
  - выигрыш
    - в производительности, 231
  - каррирование данных, 223
  - ключевое слово yield, 214
  - основные понятия, 212
  - подсчет k-мер и исправление ошибок, 219
  - подсчет k-мер при помощи каррированных функций, 226
  - преимущества библиотеки Toolz, 213, 217
  - применения, 213
  - советы для работы, 227

управление потоком, 223  
 Анализ экспрессии генов  
 бикластеризация количественных данных, 68  
 визуализация кластеров, 70  
 квантильная нормализация, 62  
 марковская модель на основе полных геномов, 228  
 нормализация, 46  
 подсчет k-мер и исправление ошибок, 219  
 предсказание выживаемости, 72  
 пример, 32  
 разница в распределении между индивидуумами, 65  
 чтение данных при помощи библиотеки pandas, 43, 64  
 Аргументы, только именованные, 54  
 Атлас ракового генома (TCGA), 32, 43

**Б**

Бесплатное программное обеспечение и программное обеспечение с открытым исходным кодом (FOSS), 16  
 Бикластеризация, 68  
 Благодарности, 24  
 Большие данные, 212, 232. См.  
 Анализ потоковых данных  
 Быстрое преобразование Фурье (FFT)  
 в анализе радарных данных, 128  
 в вычислении дискретного преобразования Фурье, 110  
 выбор длины, 116  
 дополнительные применения, 142  
 история, 115  
 оконное преобразование, 124, 136  
 преимущества, 109  
 реализация, 115  
 ресурсы, 143  
 частоты и их упорядочивание, 118

**В**

Векторизация, 41  
 Вектор Фидлера, 177

Визуальная теория информации, 160.  
 См. Теория информации  
 Волшебные команды Jupyter, 46  
 Вопросы и комментарии, 236  
 Временные данные, преобразование, 107

**Г**

Гауссово ядро, 65, 88  
 Графики, стилизация, 46  
 Графики стебель-листья, 109  
 Графы  
 в сопоставлении с сетями, 94  
 использование библиотеки NetworkX, 94  
 компоненты связности, 96  
 лапласова матрица графа, 175  
 лапласова матрица, нормализованная по степени, 181  
 построение из областей изображения, 102

**Д**

Данные, цензурированные справа, 74  
 Деревья слияния, 69  
 Дискретное преобразование Фурье (DFT)  
 анализ радарных данных, 128  
 выбор длины, 116  
 дополнительные применения, 142  
 иллюстрация спектрограммы, 110  
 история, 115  
 математические преобразования, 120  
 оконное преобразование, 136  
 преобразование данных, 107  
 реализация, 115  
 ресурсы, 143  
 частоты и их упорядочивание, 118  
 ДНК (дезоксирибонуклеиновая кислота), 34  
 Доверительные области, 194  
**И**  
 Иерархическая кластеризация, 69

Изменчивость информации

вычисление, 144, 160

использование, 167

Интервал дальности, 135

Истинно отрицательные/истинно

положительные

исходы, 146

Итеративные решатели, 187

## К

Каррирование, 223

Квантильная нормализация

бикластеризация количественных  
данных, 68

визуализация кластеров, 70

логарифмическое преобразование  
данных, 63

преимущества, 62

разница в распределении между  
индивидуумами, 65

чтение данных при помощи  
библиотеки pandas, 64

шаги, 63

Комментарии и вопросы, 23, 236

Компоненты связности, 96

Конволюция, 143. См. Свертка

Контактная информация, 237

Контрольные данные (ground  
truth), 146

Кривые выживания, 74

## Л

Лапласова матрица графа

анализ графа, 175

лапласова матрица,

нормализованная по степени, 181

Линейная алгебра

лапласова матрица графа, 175

лапласова матрица,

нормализованная по степени, 181

основы, 174

реализация алгоритма

PageRank, 187

Лицензии, 17

Лицензии на программное  
обеспечение, 17

Локальные минимумы

предотвращение при помощи  
алгоритма basin hopping, 204

пример, 195

проблемы, 204

решение проблемы разными  
алгоритмами, 193



## М

Марковские модели, 228

Математическое обозначение O  
большое, 116

Матрица переходов, 188

Матрица сопряженности, 162

Матрицы ошибок, 146. См. Матрицы  
сопряженности

Матрицы смежности, 175

Молекулярная биология,  
ее центральная догма, 34

## Н

Набор данных эталонных  
сегментаций университета  
Беркли, 168

Научная экосистема языка Python

бесплатное программное  
обеспечение и программное  
обеспечение с открытым исходным  
кодом (FOSS), 16

компоненты, 13

отсылки на шоу «Монти

Пайтон», 21

поддержка со стороны сообществ,  
16, 21. См. Ресурсы

роль NumPy, 32

участие, 21

Нормализация

квантильная нормализация, 62

между генами, 52

между образцами, 46

по образцам и генам (RPKM), 54

потребность, 46

Нормализованная взаимная информация (NMI), 209

## О

Области изображения  
 графовые представления, 94  
 двумерные фильтры (изображения), 90  
 изображения в виде массивов NumPy, 79  
 преобразования изображений с использованием разреженных матриц, 152  
 сегментация по среднему цвету, 105  
 создание графов, 102  
 универсальные фильтры, 92  
 фильтры в обработке сигналов, 84

Обработка сигналов  
 гауссово сглаживание, 88  
 двумерные фильтры (изображения), 90  
 зашумленные сигналы, 87  
 свертка, 86  
 соотношение сигнал-шум (SNR), 89  
 универсальный фильтр, 92  
 фильтры, 84

Однородные координаты, 153

Окно Кайзера, 126

Оконное преобразование, 124, 136

Оператор умножения матриц, 145

Оптимизация, 193. См. Оптимизация функции

Оптимизация при заданных ограничениях, 194

Оптимизация функции  
 выбор алгоритма оптимизации функции, 194  
 выбор целевых функций, 205  
 вычисление оптимального смещения изображения, 195  
 определение, 193  
 предотвращение локальных минимумов при помощи алгоритма basin hopping, 204

регистрация изображения, 201  
 функции стоимости, 193

## П

Палитра colorbrewer, 184

Переходные вероятности, 228

Пиксел, 78. См. Области изображения

Подсчет k-мер  
 и исправление ошибок, 219  
 использование каррированных функций, 226

Получение помощи, 22. См. Ресурсы

Предсказания, 146

Примеры программного кода, получение и использование, 19, 22, 24, 234

Причудливая индексация, 64

Простая линейная итеративная кластеризация (SLIC), 100

Пространственная частота, 108

Пространственные данные, преобразование, 107

Публичная лицензия GNU, 18

## Р

Радарные данные, анализ, 128

Разработка программного обеспечения с открытым исходным кодом, 16

Разреженные итеративные решатели, 187

Разреженная матрица  
 преимущества модуля sparse, 144  
 преобразование изображений, 152  
 преобразование массива NumPy, 166  
 сравнение форматов, 152  
 формат COO (координатный), 148, 157  
 формат сжатых разреженных строк (CSR), 150

Разреженная оптимизация, сравнение методов, 194

Разрешительные лицензии, 18

## Ресурсы

- GitHub, 19, 234
- конференции, 16
- списки рассылок, 22, 233

## Решения задач, 238

## С

- Свертка, 86. См. Конволюция
- Свободная лицензия (copyleft), 18
- Система управления версиями, 19
- Собственные векторы, 176, 183
- Сосредоточенность данных, 149
- Спектрограммы, 110
- Среднеквадратическая ошибка (СКО, MSE), 196
- Степенная матрица (D), 176
- Степенной метод, 191

## Т

## Таблицы

## сопряженности

- измерение результативности, 147
- пример таблицы с исторической информацией, 160
- примеры разреженных матриц, 144
- сегментация, 146, 159
- формат COO (координатный), 157

## Теорема Фробениуса-Перрона, 192

## Теория информации

- введение, 160
- визуальная теория информации, 160
- в сегментации, 163

## Только именованные аргументы, 54

## Транскрипционные сети, 94

## Транслирование

## массивов, 33, 41, 56, 158

## Трассировки дальности, 133

## У

## Управление потоком, 223

## Упражнения

- альтернативный алгоритм вычисления матрицы

## ошибок, 147, 243

## анализ главных компонент (PCA)

## потоковых данных, 227, 255

## вычисление условной

## энтропии, 163, 247

## вычислительная сложность матриц

## ошибок, 147, 243

## добавление сеточного

## наложения, 84, 238

## игра «Жизнь», 93, 239

## изображение аффинного

## подобия, 186, 248

## линейная алгебра с разреженными

## матрицами, 186, 249

## магнитуда градиента

## Собела, 94, 240

## матрица поворота, 176, 247

## модификация функции align, 205, 253

## мультиклассовая матрица

## ошибок, 148, 244

## обработка висячих узлов, 192, 252

## онлайновая распаковка

## архива, 231, 257

## поворот изображения, 156, 245

## подбор кривой при помощи

## SciPy, 98, 241

## предсказание выживаемости, 77

## представление в формате

## COO, 149, 244

## решения, 238

## свертка изображения, 143, 243

## сегментация на практике, 173

## сокращение объема потребляемой

## оперативной памяти, 158, 246

## эквивалентность разных методов

## получения собственного

## вектора, 192, 253

## Уровень собственных шумов, 136

## Условная энтропия, 161

## Условные обозначения, 23

## Ф

## Фильтры

## в обработке сигналов, 84

двумерные фильтры  
(изображения), 90  
универсальные фильтры, 92  
фильтры Собела, 91, 94  
функция `generic_filter`, 102

Фильтры Собела, 91, 94  
Функция скользящего окна, 223  
Функция стоимости, 193.  
См. Оптимизация функции

## Ц

Центральная догма молекулярной  
биологии, 34  
Цифровые изображения, 78.  
См. Области изображения

## Ч

Частота  
анализ данных FMCW-радаров, 128  
данные частотной области, 107  
Найквиста, 120  
оконное преобразование, 136  
понятие, 107  
пространственная частота, 108  
упорядочивание частот, 118

## Э

Элегантный программный код,  
особенности, 9

## Я

Ядерная оценка плотности (KDE), 46





Книги издательства «ДМК Пресс» можно **заказать**<sup>®</sup>  
в торгово-издательском холдинге «Планета Альянс» наложенным платежом,  
выслав открытку или письмо по почтовому адресу:

115487, г. Москва, 2-й Нагатинский пр-д, д. 6А.

При оформлении заказа следует указать адрес (полностью),  
по которому должны быть высланы книги;  
фамилию, имя и отчество получателя.

Желательно также указать свой телефон и электронный адрес.

Эти книги вы можете заказать и в интернет-магазине: **www.aliants-kniga.ru**.

Оптовые закупки: тел. **(499) 782-38-89**.

Электронный адрес: **books@aliants-kniga.ru**.



Хуан Нуньес-Иглесиас, Штефан ван дер Уолт,  
Харриет Дэшноу

### **Элегантный SciPy**

Главный редактор *Мовчан Д. А.*  
dmkpress@gmail.com

Перевод *Логунов А. В.*

Корректор *Синяева Г. И.*

Верстка *Чаннова А. А.*

Дизайн обложки *Мовчан А. Г.*

Формат 70×100 1/16.

Гарнитура «PT Serif». Печать офсетная.

Усл. печ. л. 24,9375. Тираж 200 экз.

Веб-сайт издательства: **www.dmkpress.com**