Застосування середовища статистичних обчислень R до спектроскопічних даних

Апуневич С.Є., Апуневич С.В., Дендебера М.П., Марковський А.А.

Львівський національний університет імені Івана Франка,
н.сп. астрономічної обсерваторії, доцент кафедри експериментальної фізики,
студенти 4-го та 3-го курсу фізичного факультету,
вул. Кирила і Мефодія, 8, Львів 79005

In this report we attempt to clarify how R statistical software can be applied to process and analyze the spectroscopical data. It appears that R Comprehensive Archive directly advertise quite a number of packages and tools for such application. We have tested one of them (prospectr) on Raman spectroscopy data obtained at Experimental Physics department. We find the usage of R very perspective for applying to real experiment data, as well as for learning students statistical methods.

Ускладнення експериментів, покращення точності та збільшення об'ємів даних, отриманих в експериментах, непрямий характер багатьох вимірювань кидають виклик фізикам-експериментаторам. Дедалі голосніше перед сучасним фізиком стає потреба в адекватному розумінні сучасних статистичних методів, в умінні їх застосувати до опрацювання даних, візуалізації даних та видобутку із них важливих фізичних результатів, обгрунтування їх статистичної значущості. Логічним у такому випадку виглядає не перевинайдення велосипеда, а використання наробків прикладної статистики у вигляді бібліотек до мов програмування або пакетів (середовищ).

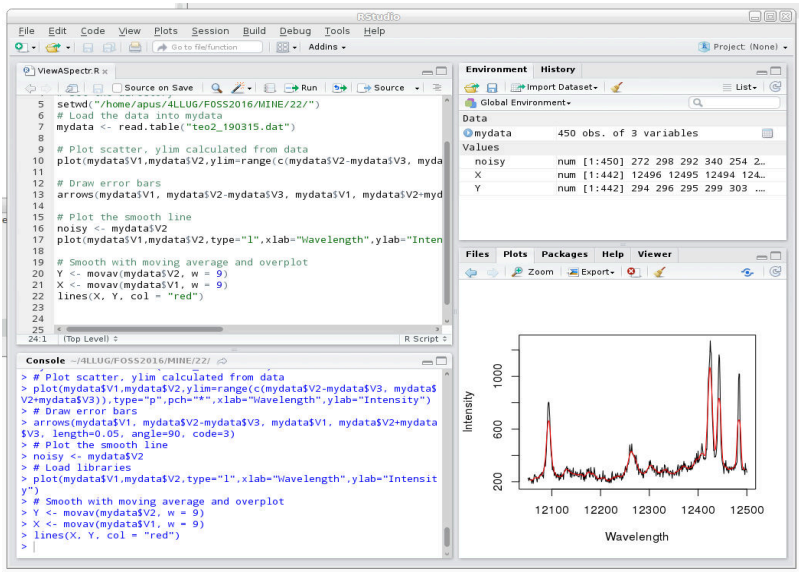
Однозначним лідером серед статистичних пакетів виглядає система (середовище) для статистичних обчислень R (R-project) [1]. Проект має тривалу історію понад 30 років, і, завдяки відкритості свого вихідного коду та потужній спільноті розробників, останнім часом переживає ренесанс не лише у середовищі професійних статистиків, але й серед спеціалістів інших галузей, від психологів до інженерів, у міру того як статистика дедалі більше входить у всі галузі життя.

До позитивних особливостей пакету R слід зарахувати: це вільне, відкрите ПЗ, один із проектів GNU [2]; це гнучка мова інтерактивних обчислень та програмування; це багатофункціональне середовище, не залишаючи котре можна виконати повний цикл роботи із даними; це архів-бібліотека CRAN (Comprehensive R Archive Network) [3] із понад 7000 пакетів; це сучасні засоби для високоякісної візуалізації даних, як інтерактивна так і для підготовки до друку; це векторизовані обчислення; це платформо-незалежність (є для Linux, Windows, MacOS); активна спільнота із професійних програмістів, академічних статистиків, активно просувається в Data Science; величезний обшир довідників/книжок/курсів/блогів.

Серед недоліків варто зауважити такі як: доволі складна і специфічна семантика функціонального програмування; деяка технологічна “несучасність”, мова мало змінилася за 30 років; доволі повільний інтерпретатор;

всі дані розміщуються в оперативній пам'яті. Здебільшого, ці недоліки або активно усуваються, або стають несуттєвими із зростанням потужності обчислювальної техніки.

Насправді, важко знайти статистичний метод чи алгоритм, який не втілено у пакеті R чи не опубліковано у вигляді рецептури або зразку. Навіть більше, R вже є дечим значно більшим ніж просто статистичним пакетом чи засобом візуалізації, він вже може потіснити MATLAB (Octave) як система наукових обчислень та моделювання.



Знімок екрану середовища розробки R-Studio із кодом, виконанням та графіком даних.

Отже, ми маємо такі завдання: є дані вимірювань; знайти що вже є в CRAN; адаптувати за мінімальної кількості кроків і застосувати для попередньої обробки даних.

Джерело [4] подає нам в одному місці аналіз наявних пакетів. Стисло подамо опис найбільш активних проектів: “hyperSpec” дає можливість візуалізувати/аналізувати гіперспектральні дані, тобто спектри із додатковою інформацією про розподіл у просторі, часі, концентрацію тощо; “ChemoSpec” є збіркою функцій для побудови графіків та загального аналізу спектральних даних; “prospectr” містить функції для попередньої обробки та формування вибірки для дифузних спектрів відбивання; “resemble” включає функції аналізу відмінностей, нелінійне моделювання спектральних даних; “TIMP” надає середовище розв'язування задач для апроксимації, розділення моделей, для часово-впорядкованих спектрів; “speaq” втілює засоби

калібровки спектрів Cluster-based Peak Alignment (CluPA); “Peaks” надає функції для маніпуляцій із спектром, перенесений із ROOT/TSpectrum.

Ми використали дані комбінаційного розсіяння сполуки TeO_2, отримані за допомогою спектрографа ДФС-52, джерело – твердотільний лазер працював на довжині хвилі 532 нм. Ми скористались пакетом prospectr для згладжування даних та усунення шуму методом біжучого середнього, а також для побудови графіків.

Висновок: Пакет статистичних обчислень R має все необхідне для застосування до опрацювання даних спектроскопічних вимірювань, і ми вважаємо доцільним внести R разом із сучасними методами статистики до програми навчання фізиків-експериментаторів.

Джерела:

1. http://www.r-project.org/
2. GNU Project: http://www.gnu.org
3. Бібліотека CRAN: http://cran.r-project.org/
4. Journal of Statistical Software http://www.jstatsoft.org/ January 2007, Volume 18, Issue 1. An Introduction to the Special Volume “Spectroscopy and Chemometrics in R”

Бестіарій Великих Даних, або про екологію проектів навколо Apache Hadoop

Апуневич С.Є.

Астрономічна обсерваторія Львівського національного університету імені Івана Франка,
EPAM Systems,
вул. Кирила і Мефодія, 8, Львів 79005

This short report is an attempt of holistic analysis of Hadoop framework for Big Data computations in terms of ecology and complexity, to expose the beauty and ugliness of this system, imagine its future.

Бестіарій [1] – це особливий жанр старовинної літератури, щедро ілюстровані збірки описів та оповідей, певною мірою предтечі енциклопедій. В них описували тварин, комах, потвор, і навіть камені, часто безоглядно до реальності описуваного, бо автори здебільшого самі не бачили цих сутностей і покладалися на непевні джерела. Бестіарії були популярні у Середньовіччі, коли різниця між реальністю та вигадкою була нечіткою, і ніхто не ставив під сумнів правдивість відомостей викладених у книжці, а істинність потверджувалася моральним висновком, що супроводжував кожен допис. До чого тут Apache Hadoop? Я висуваю тезу, що така ж суб'єктивність часто панує у сфері сучасних високих інформаційних технологій.