DOI 10.15826/B978-5-7996-3229-8.18

АЛГОРИТМ ВЕЙВЛЕТ-АНАЛИЗА ДВУ- И ТРЕХМЕРНЫХ СТАТИСТИЧЕСКИХ РАСПРЕДЕЛЕНИЙ

Р. В. Балуев
Санкт-Петербургский государственный университет

Рассматривается задача поиска статистически значимых структур в распределении каких-либо астрономических объектов. Для этого используется метод вейвлет-анализа в дву- или трехмерном пространстве параметров. Алгоритм двумерного анализа полностью завершен и опубликован в виде открытого C++ кода стабильной версии, тогда как трехмерный алгоритм находится на экспериментальной стадии.

WAVELET ANALYSIS ALGORITHM FOR BI- AND TRIVARIATE STATISTICAL DISTRIBUTIONS

R. V. Baluev
Saint Petersburg State University

We consider the task of detecting statistically significant patterns in a distribution of some astronomical objects. For this goal we use the wavelet analysis method in bi- or trivariate parametric space. The bivariate analysis algorithm is finished and released as an open-source C++ code of a stable version, while the trivariate algorithm is at an experimental stage.

Введение

Вейвлет-анализ восходит к Гроссману и Морле [1] и представляет собой математический инструмент, способный представить какую-либо функцию через совокупность множества уровней разрешения. В настоящее время этот метод часто используется в различных областях науки, наиболее известным приложением являются анализ временных рядов (одномерный вейвлет-анализ) и обработка изображений (двумерный вейвлет-анализ).

Классический вейвлет-анализ нацелен на исследование функции $f(x)$ детерминированного аргумента x (возможно, многомерного). Сама функция может содержать случайный шум, но она определена на детерминированной области. В задачах такого типа измеряются значения $f_i = f(x_i)$, а затем к ним применяется дискретная версия вейвлет-преобразования. Шум в этой классической формулировке принимается обычно аддитивным, т. е. $f(x) = s(x) + n(x)$, где $s(x)$ — детерминированный сигнал; $n(x)$ — шум.

Здесь, однако, мы рассматриваем другую практическую задачу: анализ распределения случайной величины. В такой формулировке мы не можем измерять значения $f(x_i)$ напрямую. Вместо этого мы должны оценить плотность вероятности $f(x)$ или получить какие-либо научные знания об $f(x)$, основываясь на случайной выборке из N независимых случайных величин x_i. Тогда $f(x)$ определяется неявно как локальная плотность x_i в окрестности заданного x. Важно, что шум в такой оценке плотности не является аддитивным, поскольку возникает по причине случайных флуктуаций выборки. В этой задаче он представляет собой шум дробового типа. Таким образом, наша задача оказывается существенно отличной от того, что мы имеем в классическом вейвлет-анализе.

Хотя имеется обширная математическая литература по классическому вейвлет-анализу, эти результаты относятся в основном к анализу временных рядов или обработке изображений и предполагают аддитивный шум. Для задачи анализа распределений имеющаяся

литература все еще довольно скудна и до сих пор не позволяет построить самодостаточного алгоритма анализа, хотя первые попытки решения этой задачи (в применении к конкретным астрономическим приложениям) восходят к 1990-м [2, 3].

При ближайшем рассмотрении используемые методы включают существенные дефекты либо не учитывают важные аспекты задачи:

1. Вейвлет-преобразование часто применяется к бинированной выборке, что позволяет использовать далее теорию классического вейвлет-анализа (с небольшими модификациями). Однако такой подход приводит к потере мелких структур из-за бинирования и к дополнительным ошибкам интерполяционного типа.

2. Большой проблемой является фильтрация шума в вейвлет-преобразовании. Многие методы такой фильтрации были разработаны еще в 1990-х гг. [4], однако они не подходят для наших целей, поскольку не позволяют выразить статистическую значимость как таковую (через доверительную вероятность или подобную ей меру). В других работах, где пытались оценить значимость в нужном нам определении, не учитывался эффект множественного тестирования (одновременная проверка большого числа независимых вейвлет-коэффициентов). Пренебрежение этим эффектом [3] приводит к необоснованному завышению значимости и числа выявленных структур.

3. Другая проблема с фильтрацией шума возникает по причине того, что необходимые для этого уровни значимости оцениваются с помощью численного моделирования Монте-Карло. Это очень медленный и вычислительно затратный подход, потому очевидна необходимость более быстрых аналитических оценок.

4. До сих пор почти не уделялось внимание вопросу оптимальности применяемых вейвлетов, т. е. улучшению отношения сигнал/шум и соответственно улучшению чувствительности анализа к малоамплитудным структурам.

Метод

Основы разработанной нами методики представлены в работе [5], где описан самосогласованный алгоритм вейвлет-анализа одномерных статистических распределений. В дальнейшем он был обобщен на двумерный случай, т. е. для анализа распределения объектов на плоскости двух параметров [6], а сейчас ведется работа по расширению методики на три измерения. Повышение размерности ведет к усложнению некоторых формул, а также к резкому росту сложности вычислений, однако общая схема анализа остается примерно одинаковой.

Используется изотропная версия многомерного вейвлет-преобразования:

$$Y(a, \mathbf{b}) = \int\limits_{\mathbb{R}^n} f(\mathbf{x}) \psi \left(\frac{|\mathbf{x} - \mathbf{b}|}{a} \right) d\mathbf{x}, \tag{1}$$

где $f(\mathbf{x})$ — плотность вероятности вектора $\mathbf{x}$, а ψ — анализирующий вейвлет (см. ниже). Изотропная она потому, что вейвлет радиально симметричен (зависит только от модуля аргумента).

По общим требованиям вейвлет-анализа вейвлет ψ должен быть функцией, хорошо локализованной вместе со своим фурье-образом $\hat{\psi}$. Конкретная форма ψ определяется потребностями задачи. В нашем алгоритме ψ определяется как лапласиан

$$\psi = \Delta \varphi, \tag{2}$$

где φ — некоторая нормируемая функция колоколообразной формы. При таком выборе вейвлет-преобразование Y представляет собой сглаженный ядром φ лапласиан функции f, что позволяет придать функции Y интерпретацию, связанную со скоростью ухода f от локальной касательной плоскости. Большой лапласиан потенциально указывает на наличие локального максимума или минимума или как минимум на высокую кривизну графика $f(x)$, что является признаком сильной неоднородности распределения в окрестности заданной точки.

Однако мы не можем вычислить Y напрямую, поскольку оно определено через неизвестную нам f. Можно заметить, что (1) является математическим ожиданием случайной величины

$$y = \psi\left(\frac{|\mathbf{x}-\mathbf{b}|}{a}\right). \tag{3}$$

Следовательно, мы можем построить статистическую *оценку* функции Y через выборочное среднее y:

$$\widetilde{Y}(a,\mathbf{b}) = \langle y\rangle = \left\langle \psi\left(\frac{|\mathbf{x}-\mathbf{b}|}{a}\right)\right\rangle, \tag{4}$$

где угловые скобки означают усреднение по выборке из N объектов, для каждого из которых задан параметрический вектор x_i. Аналогично можно построить оценку дисперсии $\widetilde{Y}$ как функцию $\widetilde{D}(a,\mathbf{b})$ (по классическим формулам выборочной дисперсии для y_i).

Также необходимо задать нулевую гипотезу, т. е. модель сравнения $Y_0(a,\mathbf{b})$, соответствующую какой-то простой плотности распределения (можно выбрать $Y_0 \equiv 0$ или взять Y_0 на основе гауссовой аппроксимации $f(\mathbf{x})$). В итоге можно построить нормализованную величину, основную статистику, служащую для фильтрации шума:

$$z(a,\mathbf{b}) = \frac{Y(a,\mathbf{b}) - Y_0(a,\mathbf{b})}{\sqrt{\widetilde{D}(a,\mathbf{b})}}. \tag{5}$$

Если модель Y_0 верна, то $z(a,\mathbf{b})$ должна оставаться небольшой по модулю, а большие значения указывают на наличие в данной точке $(a,\mathbf{b})$ статистически значимой структуры. Заметим, что функция $z(a,\mathbf{b})$ определена через случайную выборку, а значит, каждое ее значение является случайной величиной. Таким образом, сама $z(a,\mathbf{b})$ есть случайное поле. Критерий разделения значимых структур от незначимых требует вычисления функции распределения максимального отсчета этого случайного поля, что сделано в указанных выше работах.

Также в работе проводился поиск оптимальных вейвлетов, точнее — оптимальных производящих функций φ, которые позволили бы минимизировать возникающий в $z(a,\mathbf{b})$ шум (хотя z нормирована так, что дисперсия каждого ее значения всегда равна единице, дисперсия максимального отсчета зависит от корреляционных свойств поля, которые определяются формой вейвлета). Оптимальные вейвлеты получены для всех размерностей задачи от 1 до 3, при этом они существенно отличаются от широко распространенного классического MHAT-вейвлета (лапласиана гауссианы).

Результаты

Алгоритм одно- и двумерного вейвлет-анализа распределений реализован в виде открытого C++ кода, который доступен по адресу: `https://sourceforge.net/projects/waveletstat/`.

Одномерный алгоритм применялся к исследованию распределений экзопланет [7]. Была найдена новая статистически значимая группа планет-гигантов с большой полуосью орбит

около 1 а. е. Она, вероятно, связана с эффектом ледяной аккумуляции в протопланетном диске, существование этого семейства планет подтверждает важность данного физического явления в планетообразовании.

Двумерный алгоритм анализа применялся в двух независимых задачах: поиск астероидных семейств Главного пояса [8] и поиск движущихся групп звезд солнечной окрестности по каталогу GAIA DR2 [9]. Основной итог по первой задаче: вейвлет-анализ позволил обнаружить лишь часть известных астероидных семейств, т. е. многие семейства, обнаруженные обычным здесь методом HCM (иерархический кластерный анализ), в рамках вейвлет-анализа оказываются статистически незначимы. Тем не менее это может быть эффектом проекции и наложения семейств, так как мы изучали фактически трехмерное распределение по его двумерным проекциям. Во второй задаче, наоборот, удалось найти около 20 ранее не известных движущихся групп звезд. Все они оказались высокоскоростными (в низкоскоростной зоне весьма силен эффект наложения и перекрытия известных групп).

Таким образом, применение двумерного алгоритма анализа к практическим задачам показало, что полноценное исследование требует обобщения методики на три измерения. Астероиды требуют полного исследования как минимум пространства собственных элементов $(a, e, \sin i)$, а звездное население — пространства скоростей (U, V, W). Использование лишь двумерных проекций не позволяет сделать однозначных выводов о соответствующем трехмерном распределении, хотя иногда позволяет получить интересные результаты.

Обобщение нашего алгоритма на случай трех измерений в целом выполнено, первых результатов его применения к этим задачам можно ожидать в ближайшем будущем.

Работа выполнена при поддержке Министерства науки и высшего образования РФ, проект 075-15-2020-780 (N13.1902.21.0039).

Библиографические ссылки

[1] *Grossman A., Morlet J.* Decomposition of Hardy functions into square integrable wavelets of constant shape // SIAM J. Math. Anal. — 1984. — Vol. 15. — P. 723–736.

[2] *Chereul E., Crézé M., Bienaymé O.* The distribution of nearby stars in phase space mapped by Hipparcos. II. Inhomogeneities among A-F type stars // Astron. Astrophys. — 1998. — Vol. 340. — P. 384–396.

[3] *Skuljan J., Hearnshaw J. B., Cottrell P. L.* Velocity distribution of stars in the solar neighbourhood // Mon. Not. R. Astron. Soc. — 1999. — Vol. 308. — P. 731–740.

[4] *Abramovich Felix, Bailey Trevor C., Sapatinas Theofanis.* Wavelet Analysis and its Statistical Applications // J. R. Stat. Soc. D (The Statistician). — 2000. — Vol. 49. — P. 1–29.

[5] *Baluev Roman V.* Statistical detection of patterns in unidimensional distributions by continuous wavelet transforms // Astron. Comput. — 2018. — Vol. 23. — P. 151–165.

[6] *Baluev R. V., Rodionov E. I., Shaidulin V. Sh.* Isotropic wavelet denoising algorithm for bivariate density analysis and estimation // preprint (submitted to Appl. & Comput. Harmonic Anal.). — 2020. — Vol. arXiv.org:1903.10167.

[7] *Baluev Roman V., Shaidulin Vakhit Sh.* Fine-resolution wavelet analysis of exoplanetary distributions: hints of an overshooting iceline accumulation // Astrophys. Space. Sci. — 2018. — Vol. 363. — P. 192.

[8] *Baluev R. V., Rodionov E. I.* Analysing the Main Belt asteroid distributions by wavelets // Celest. Mech. Dyn. Astron. — 2020. — Vol. 132. — P. 34.

[9] *Baluev R. V., Shaidulin V. Sh., Veselova A. V.* High-velocity moving groups in the Solar neighborhood in GAIA DR2 // Acta Astron. — 2020. — Vol. 70. — P. 141–168.