**Иерархический подход к определению числа кластеров и параметров GMM для функции распределения ионов.**

Парамоник И.П. 1, Дивин А.В. 1, Зайцев И.В.2, Чжун Д.1, Семенов В.С. 1

*1 СПбГУ, г. Санкт-Петербург, Россия,* [*igorparamonik@gmail.com*](mailto:igorparamonik@gmail.com)

*2 Университет Хельсинки, г. Хельсинки, Финляндия*

Функция распределение ионов по скоростям может иметь сложную структуру, содержащую ускоренные холодные пучки, сильно тепловые популяция (гало), тяжелые хвосты, особенности вида "полумесяц" и другие. Для задачи автоматического определения особенностей функции распределения, полученных в результате кинетического PIC-моделирования [1], себя хорошо показал метод Gaussian Mixture Model (GMM) [2], являющийся алгоритмом кластеризации. Однако, как и для многих других алгоритмов этого класса методов машинного обучение без учителя, отдельной проблемой остается определение оптимального числа кластеров. Слишком малое число компонент смеси может скрыть локальные особенности функции распределения, а слишком большое – найдет избыточное число структур, чтобы лучшим образом описать начальное распределение.

Для определения оптимального числа кластеров принято проводить несколько расчетов модели, варьируя количество компонент, а наилучшая модель из них выбирается с помощью метода локтя или информационного критерия Акаике (AIC) или Байеса (BIC). Однако не всегда результаты выбора модели этими методами дают приемлемое для нас разбиение, поскольку они не учитывают параметры найденных кластеров, а опираются лишь на значение функции логарифма правдоподобия, которая может сходиться к локальному экстремуму из-за наличия элемента случайности при инициализации итерационного алгоритма.

Поскольку мы имеем ряд моделей GMM с разным количеством кластеров для одного распределения, то можем обобщить информацию для всех компонент смеси из такой серии моделей, используя иерархический метод. Сравнивая отдельные компоненты друг с другом, объединяя их по мере уменьшения количества заданных кластеров, возможно визуализировать дерево связей и сказать, где разделение становится избыточным, тем самым упростить анализ функции распределения и подобрать оптимальное разложение. Результаты совмещения иерархического подхода и GMM, мы исследуем на примере синтетического тестового распределения и функции распределения ионов из PIC-моделирования.

1. Markidis S. and Lapenta G. "Multi-scale simulations of plasma with iPIC3D." Mathematics and Computers in Simulation 80.7 (2010): 1509-1519.
2. Dupuis R. et al. "Characterizing magnetic reconnection regions using Gaussian mixture models on particle velocity distributions." *The Astrophysical Journal* 889.1 (2020): 22.