

Байесовская оптимизации для вывода демографических историй

Илья Шешуков (СПбГУ)

Руководители: Екатерина Носкова (Университет ИТМО)
Вячеслав Боровицкий (СПбГУ, ПОМИ РАН)

Введение

Демографическая модель популяции

Имея геномы людей, хотим понять как изменялись их популяции. Как менялась численность, когда популяции разделялись, как сильно они мигрировали.

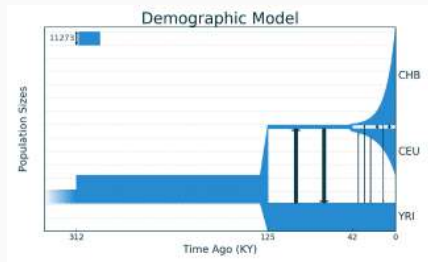


Рис. 1: Демографическая модель африканского происхождения человека

Аллель-частотный спектр

Аллель-частотный спектр это распределение частоты аллелей в данных локусах в популяции или выборке.

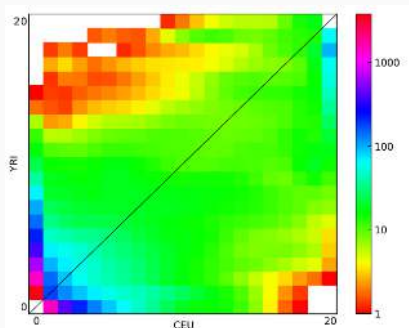


Рис. 2: График АЧС

Как это делается сейчас

<https://bitbucket.org/gutenkunstlab/dadi/>

- Плюсы
 - Она работает
 - Ей пользуются реальные люди
- Минусы
 - Решает дифференциальное уравнение в частных производных, что долго
 - Использует методы локальной оптимизации, что малоэффективно
 - Для работы необходимо руками писать Питон

<https://bitbucket.org/simongravel/moments>

- Плюсы
 - Эффективнее, чем dad1 , особенно на больших популяциях

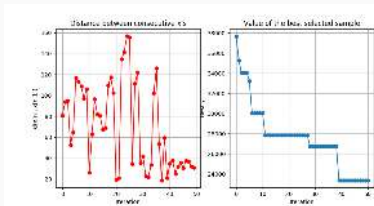
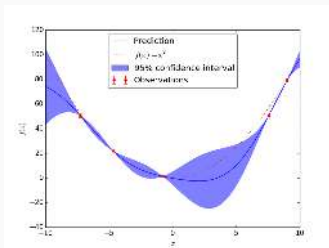
<https://github.com/ctlab/GADMA>

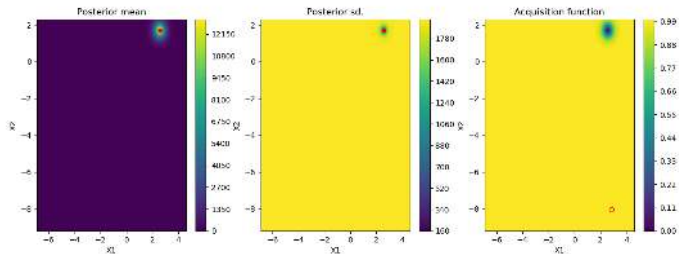
- Основана на `dad1` и `moments`
- Использует генетический алгоритм для поиска значения параметров демографической модели
- Не требует человеческого вмешательства

Заменяем генетический алгоритм байесовской оптимизацией.

- Алгоритм глобальной оптимизации
- Хорошо работает для сложновычислимых функций (например, если нужно решать уравнение в частных производных), т.е. хорошо подходит для задачи
- Можно параллелить
- Менее эвристична, чем генетический алгоритм

Красивые графики работы GPyOpt





Результаты

- Заменить в `dad1` алгоритм градиентного спуска на байесовскую оптимизацию.
- Посмотреть станет ли лучше
- Интегрировать в `GADMA`

- Заменить в `dad1` moments алгоритм градиентного спуска на байесовскую оптимизацию.
- Посмотреть станет ли лучше
- Интегрировать в GADMA

- Копались в библиотеках
- Нашли баги в GPyOpt
- Играли с гиперпараметрами
- Думали, почему всё работает плохо
- Очень долго ждали

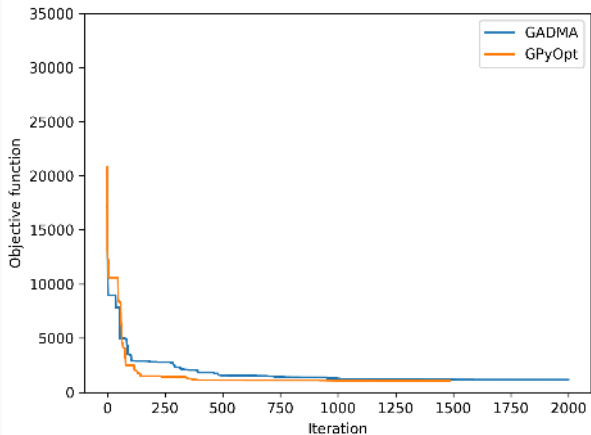


Рис. 3: 2 популяции, 6 переменных

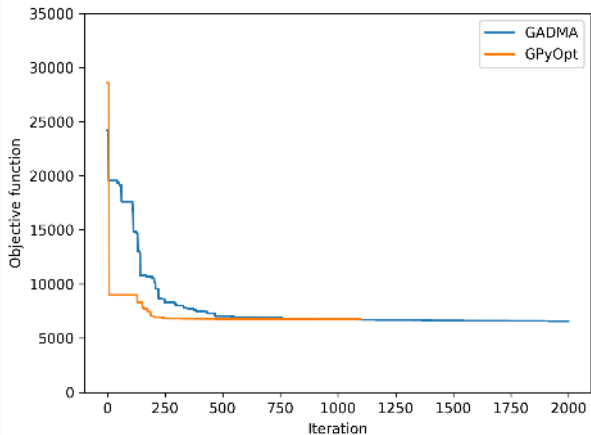


Рис. 4: 3 популяции, 13 переменных

- Байесовская оптимизация оправдывает себя, особенно если вычисления функции очень дорогие (как у `dad1`)
- Но всё равно, **пока** работает не так хорошо, как могла бы (т.е. лучше всех)

- Потестировать на других данных
- Потестировать на разных гиперпараметрах; найти такие, которые будут работать лучше всего
- Интегрировать в GADMA

Конец

Спасибо за внимание



<https://github.com/isheshukov/bioinf-sem-project>