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.

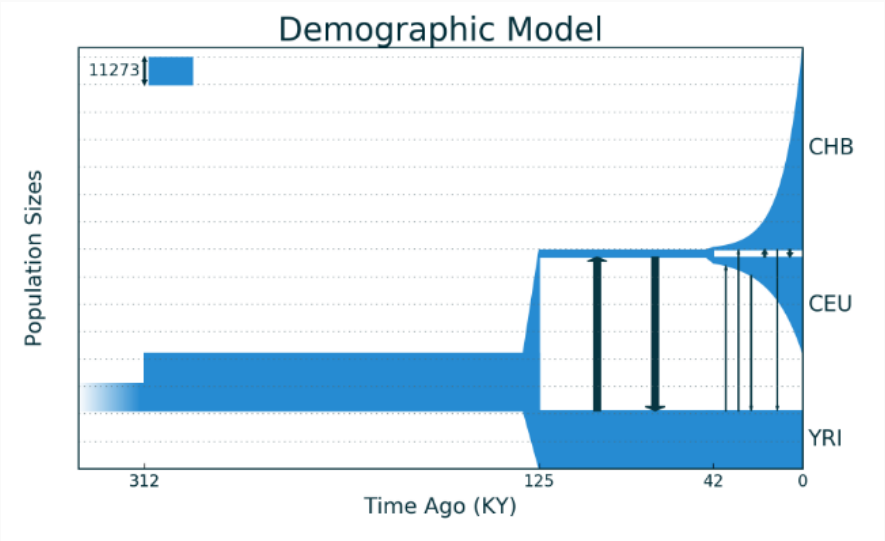


Рис. 1: Демографическая модель африканского происхождени человека

Аллель-частотный спектр

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

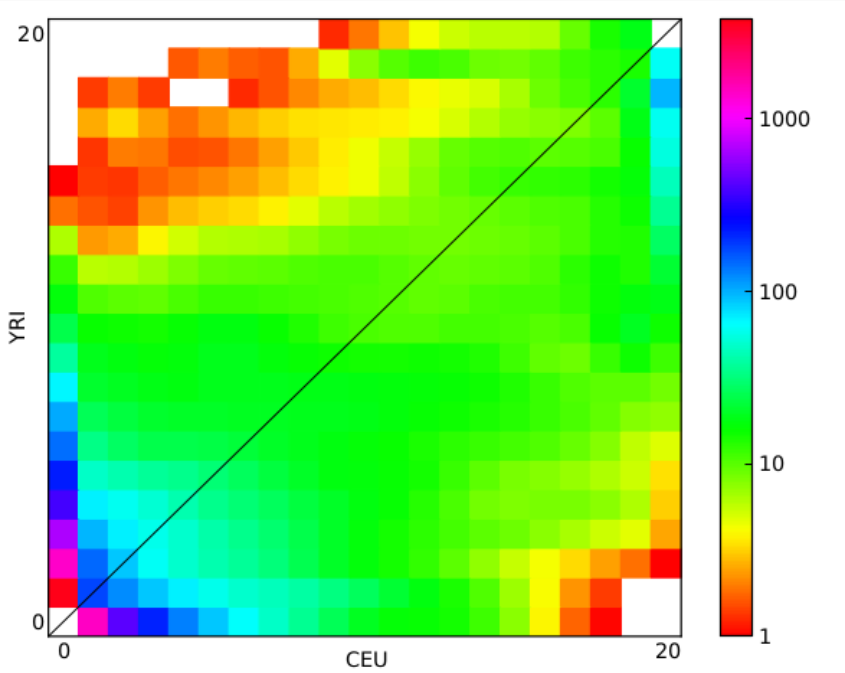


Рис. 2: График АЧС

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

moments

https://bitbucket.org/simongravel/moments

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

GADMA

https://github.com/ctlab/GADMA

- Основана на ∂a∂i и moments
- Использует генетический алгоритм для поиска значения параметров демографической модели
- Не требует человеческого вмешательства

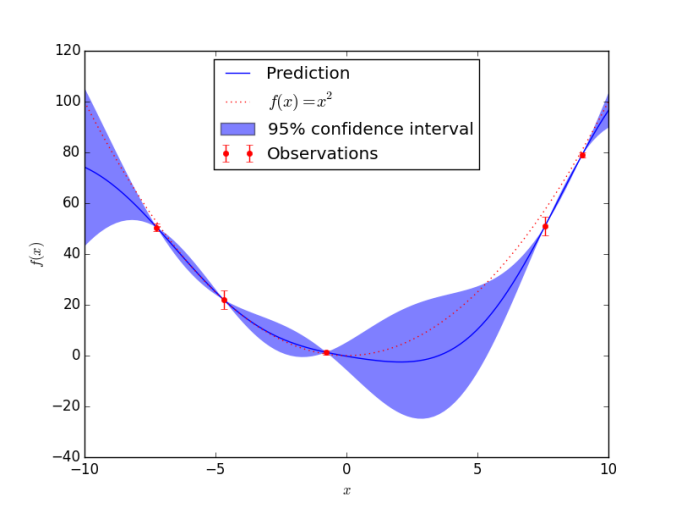
Что можно сделать

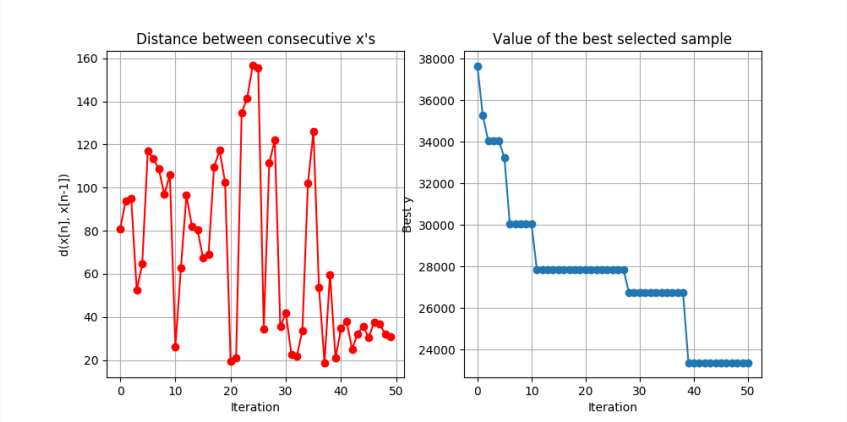
Заменим генетический алгоритм байесовской оптимизацей.

Байесовская оптимизация

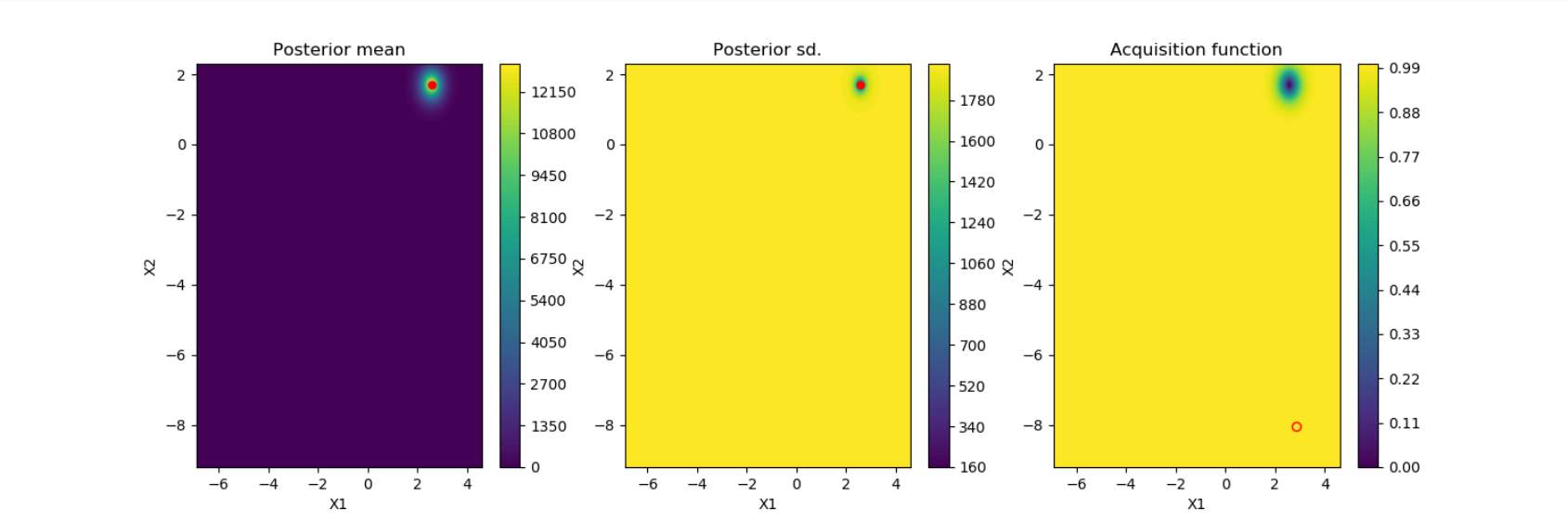
- Алгоритм глобальной оптимизации
- Хорошо работает для сложновычислимых функций (например, если нужно решать уравнение в частных производных), т.е. хорошо подходит для задачи
- Можно параллелить
- Менее эвристична, чем генетический алгоритм

Красивые графики работы GPyOpt





Красивые графики работы GPyOpt



Результаты

Планы (промежуточная презентация)

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

Реальность

- ☒ Заменить в ~~adai~~ moments алгоритм градиентного спуска на байесовскую оптимизацию.
- ☒ Посмотреть станет ли лучше
- ☐ Интегрировать в GADMA

Что мы делали

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

Сравнение

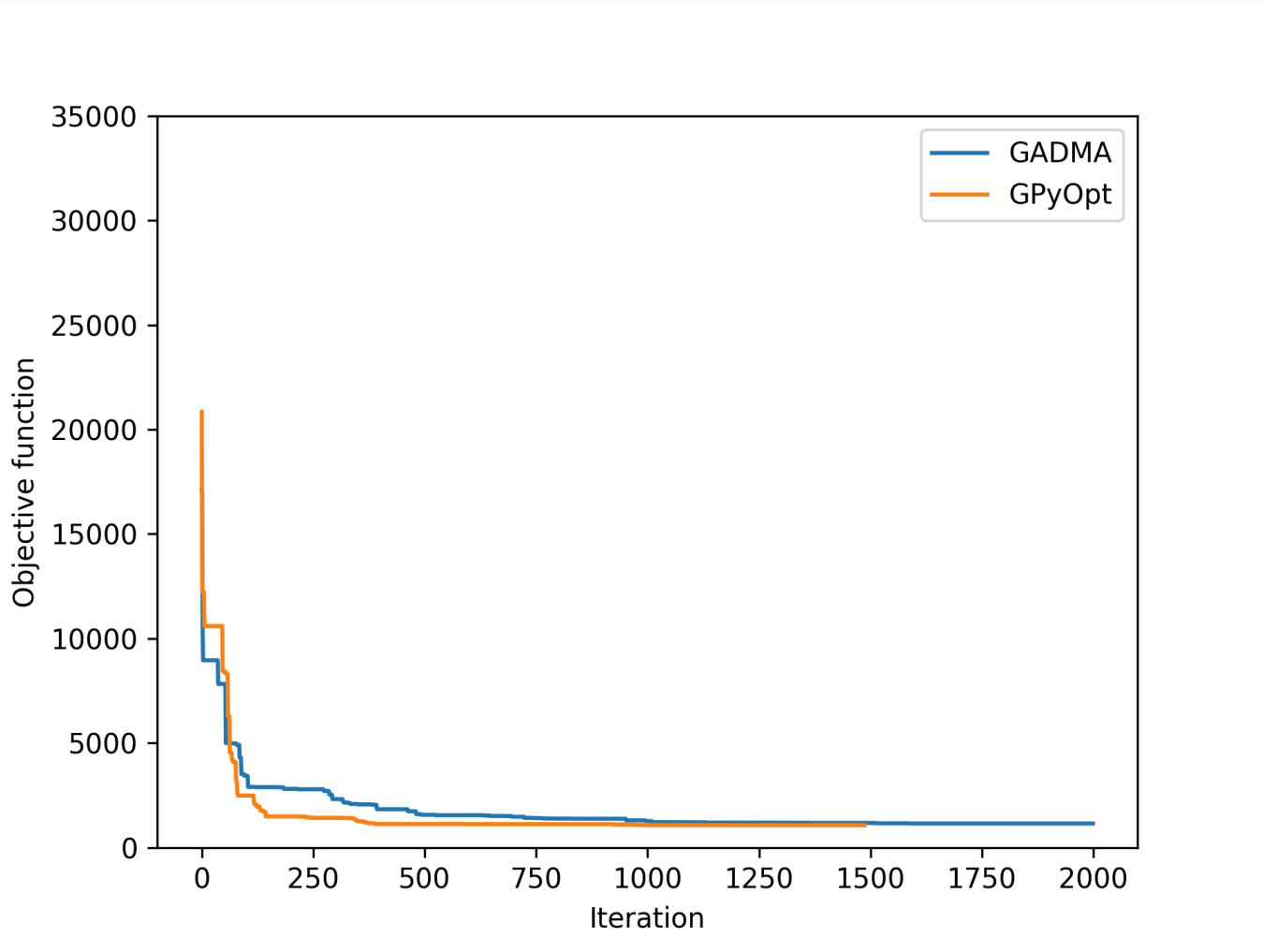


Рис. 3: 2 популяции, 6 переменных

Сравнение

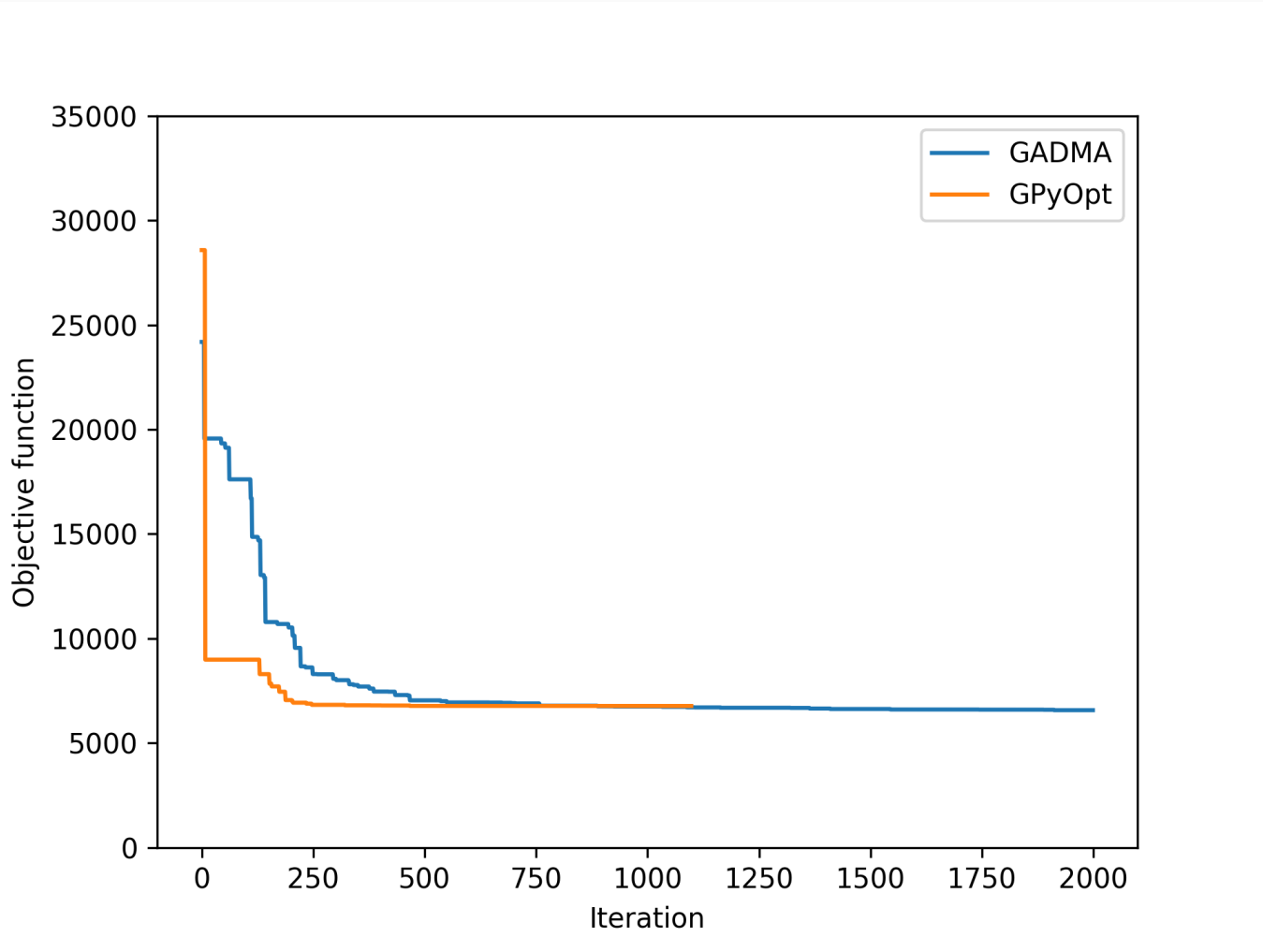


Рис. 4: 3 популяции, 13 переменных

Итого

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

Что бы хотелось ещё сделать

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

Конец

Спасибо за внимание

Github

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