Вывод демографических историй при помощи байесовской оптимизации

Илья Шешуков

Введение

Демографическая модель популяции

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

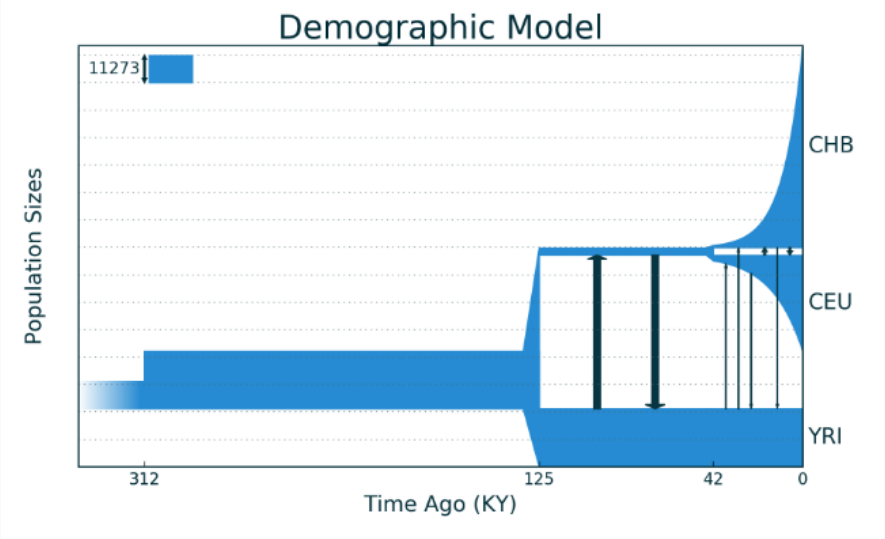


Рис. 1: Демографическая модель африканского происхождени человека

Аллель-частотный спектр

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

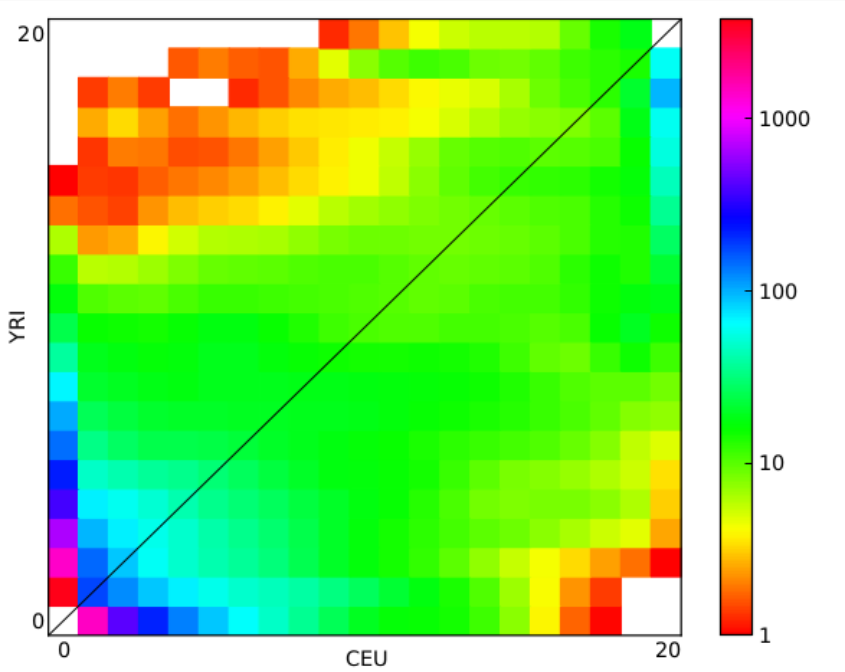


Рис. 2: График АЧС

Пример

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7	SNP 8
	0	1	0	0	0	0	1	0
	1	0	1	0	0	0	1	0
	0	1	1	0	0	1	0	0
	0	0	0	0	1	0	1	1
	0	0	1	0	0	0	1	0
	0	0	0	1	0	1	1	0
Сумма	1	2	3	1	1	2	5	1

Спектр: $\begin{pmatrix} 4 & 2 & 1 & 0 & 1 \end{pmatrix}$

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 ∂а∂і , 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

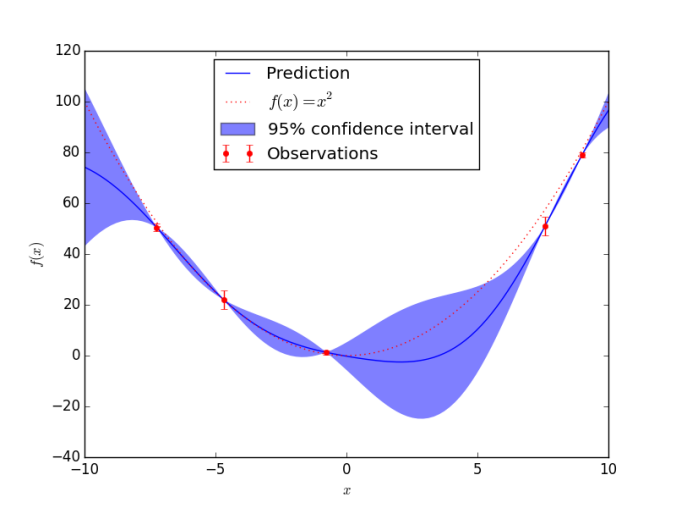
Что можно сделать

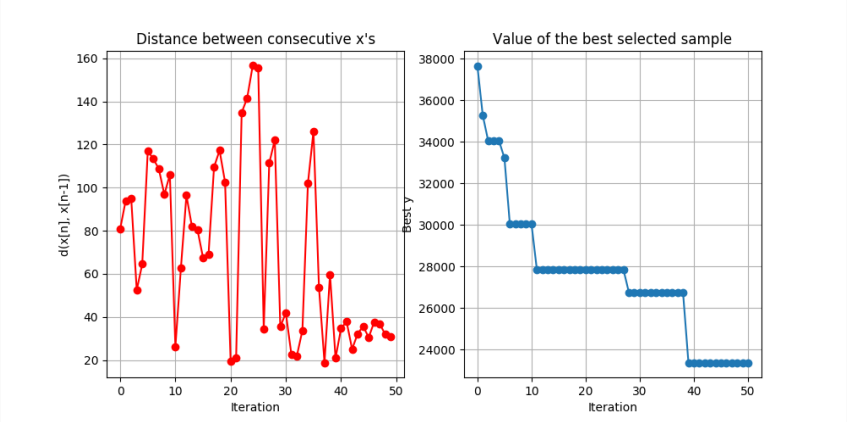
Заменим генетический алгоритм байесовской оптимизацей.

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





Результаты

Планы (промежуточная презентация)

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

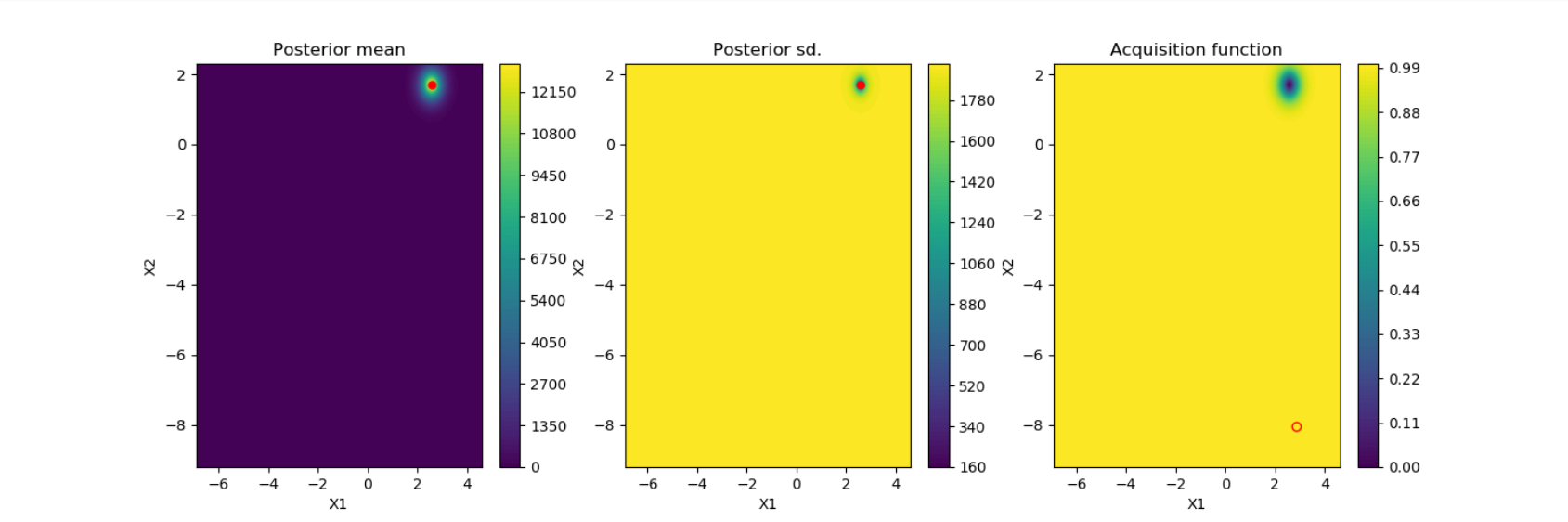
Реальность

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

Сравнительная таблица

Данные	Оптимум	dadi	moments	GPyOpt
2 популяции	1066.823	-	-	56 часов
6 переменных				f(x) = 1066.954
2 популяции	1070.048	-	-	24 часа
8 переменных				f(x) = 1160.432
3 популяции	6316.578	-	-	73 часа
13 переменных				f(x) = 7377.065

Torque Generating Mechanism



Спасибо за внимание

TODO

- время -> итерации
- анимированные графики в презентации
- убрать дади, получить данные по моментс
- графики сходимости по
- добавить лирики (что происходилов работе !!!!)
- сравнить на других данных