Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ПРИМЕНЕНИЕ БАЙЕСОВСКОЙ ОПТИМИЗАЦИИ ДЛЯ ВЫВОДА ДЕМОГРАФИЧЕСКОЙ ИСТОРИИ ПОПУЛЯЦИЙ ПО ГЕНОМНЫМ ДАННЫМ

Автор: Хужин Павел Андреевич _____

Направление подготовки: 01.03.02 Прикладная математика и информатика

Квалификация: Бакалавр

Руководитель ВКР: Ульянцев В.И., канд. техн. наук _____

Санкт-Петербург, 2020 г.

Обучающийся Хужин Павел Андреевич
Группа М3435 Факультет/институт/кластер ИТиП

Направленность (профиль), специализация
Математические модели и алгоритмы в разработке программного обеспечения

ВКР принята «____» _____ 20___ г.

Оригинальность ВКР _____%

ВКР выполнена с оценкой _____

Дата защиты «____» _____ 20___ г.

Секретарь ГЭК Павлова О.Н. _____

Листов хранения _____

Демонстрационных материалов/Чертежей хранения _____

Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

УТВЕРЖДАЮ

Руководитель ОП

проф., д.т.н. Парфенов В.Г. _____

«____» _____ 20___ г.

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

Обучающийся: Хужин Павел Андреевич

Группа: М3435 **Факультет/институт/кластер:** ИТиП

Квалификация: Бакалавр

Направление подготовки: 01.03.02 Прикладная математика и информатика

Направленность (профиль) образовательной программы: Математические модели и алгоритмы в разработке программного обеспечения

Тема ВКР: Применение байесовской оптимизации для вывода демографической истории популяций по геномным данным

Руководитель Ульянцев В.И., канд. техн. наук, доцент факультета информационных технологий и программирования университета ИТМО

2 Срок сдачи студентом законченной работы до: «____» _____ 20___ г.

3 Техническое задание и исходные данные к работе

Требуется разработать и проанализировать метод, основанный на байесовской оптимизации, для поиска параметров демографической истории четырёх и пяти популяций по геномным данным. Требуется провести сравнение сходимости полученного метода с существующими алгоритмами поиска параметров демографической истории, а именно с генетическим алгоритмом и случайным поиском. Сравнение необходимо провести как на симулированных данных, так и на реальных данных из работы Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation / J. Jouganous [et al.].

4 Содержание выпускной квалификационной работы (перечень подлежащих разработке вопросов)

а) Описание предметной области и существующих решений поиска параметров демографических историй по геномным данным;

б) Разработка и реализация метода, основанного на байесовской оптимизации, для поиска параметров демографических историй четырех и пяти популяций;

в) Проведение экспериментальных исследований для выявления эффективности метода на симулированных и реальных данных.

5 Перечень графического материала (с указанием обязательного материала)

Графические материалы и чертежи работой не предусмотрены

6 Исходные материалы и пособия

а) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data / R. N. Gutenkunst [et al.] // PLoSgenetics. — 2009. — Vol. 5, no. 10.

б) Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation / J. Jouganous [et al.] // Genetics. — 2017. — Vol.206, no. 3.

в) Frazier P. I.A tutorial on bayesian optimization // arXiv preprint arXiv:1807.02811. — 2018

7 Дата выдачи задания «____» _____ 20___ г.

Руководитель ВКР _____

Задание принял к исполнению _____ «____» _____ 20___ г.

Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Обучающийся: Хужин Павел Андреевич

Наименование темы ВКР: Применение байесовской оптимизации для вывода демографической истории популяций по геномным данным

Наименование организации, где выполнена ВКР: Университет ИТМО

ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

1 Цель исследования: Разработка и анализ метода, основанного на байесовской оптимизации, для поиска параметров демографической истории четырёх и пяти популяций по геномным данным.

2 Задачи, решаемые в ВКР:

а) Изучение существующих подходов;

б) Разработка и реализация метода, основанного на байесовской оптимизации;

в) Экспериментальные исследования и выявление эффективности метода.

3 Число источников, использованных при составлении обзора: 19

4 Полное число источников, использованных в работе: 26

5 В том числе источников по годам:

Отечественных			**Иностранных**		
Последние 5 лет	От 5 до 10 лет	Более 10 лет	Последние 5 лет	От 5 до 10 лет	Более 10 лет
0	0	0	6	3	17

6 Использование информационных ресурсов Internet: да, число ресурсов: 1

7 Использование современных пакетов компьютерных программ и технологий:

Пакеты компьютерных программ и технологий	**Раздел работы**
Среда разработки PyCharm и язык программирования Python 3	2
Библиотека GPyOpt	2
Библиотека *moments*	2.3, 2.4, 2.5, 2.6
Программное обеспечение GADMA	2.3, 2.4, 2.5, 2.6
LaTeX	1, 2

8 Краткая характеристика полученных результатов

В данной работе был рассмотрен и реализован метод, основанный на байесовской оптимизации, для поиска параметров эволюционных историй для четырех и пяти популяций. Метод был протестирован не только на симулированных данных, но и на реальных. Удалось показать превосходство предложенного алгоритма на начальных этапах оптимизации для демографических историй с нулевыми миграциями. Такой результат делает целесообразным

применение байесовской оптимизации в условиях ограниченного временного бюджета или в составе комбинированного подхода.

9 Полученные гранты, при выполнении работы

При выполнении работы грантов получено не было

10 Наличие публикаций и выступлений на конференциях по теме выпускной работы

Хужин П. А. Применение байесовской оптимизации для вывода демографической истории популяций по геномным данным // Сборник тезисов докладов конгресса молодых ученых. Электронное издание. – СПб: Университет ИТМО, 2020.

Обучающийся Хужин П.А. _____

Руководитель ВКР Ульянцев В.И. _____

«____» _____ 20___ г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 5
1. Обзор предметной области 7
1.1. Основные положения популяционной генетики 7
1.1.1. Аллель-частотный спектр 8
1.2. Сравнение на основе правдоподобия 10
1.3. Байесовская оптимизация 12
1.3.1. Гауссовский процесс 13
1.3.2. Ковариационная функция 13
1.3.3. Гиперпараметры ковариационной функции 14
1.3.4. Регрессия на основе Гауссовского процесса 15
1.3.5. Функция выбора 16
1.3.6. Основной алгоритм байесовской оптимизации 17
1.4. Постановка задачи 20
1.5. Параметры демографической истории 20
Выводы по главе 1 21
2. Экспериментальные исследования 22
2.1. Существующие реализации байесовской оптимизации 22
2.2. Реализация 22
2.3. Симулированные данные 24
2.3.1. Четыре популяции 25
2.3.2. Пять популяций 27
2.4. Реальные данные 29
2.4.1. Четыре популяции 29
2.4.2. Три популяции 32
2.4.3. Четыре популяции без миграций 33
2.5. Комбинированный подход 34
2.6. Локальные дооптимизации 36
Выводы по главе 2 37
ЗАКЛЮЧЕНИЕ 39
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 41

ВВЕДЕНИЕ

Построение демографической модели популяций является одной из важных проблем популяционной генетики. Настоящая работа посвящена разработке алгоритма, основанного на байесовской оптимизации и его применению для вывода демографических историй для четырёх и пяти популяций по геномным данным.

Демографическая история — это математическая модель истории развития и эволюции популяций. Эта модель обычно включает в себя такие параметры, как численность популяций, время разделения и темпы миграций. Перед исследователями ставится задача поиска оптимальных значений этих параметров по геномным данным.

В идеале требуется найти демографическую модель, которая бы полностью описывала геномные данные, однако существующие алгоритмы позволяют найти только приближенную модель. Следует отметить, что количество вариантов параметров может быть очень большим, что дополнительно усложняет поиск.

Существует несколько решений для симуляции генетических данных по заданной демографической истории. Одним из наиболее перспективных методов является аппроксимация моментов случайного процесса. Этот метод был реализован в программном обеспечении *moments* [13] и поддерживает симуляцию для демографических историй до пяти популяций. *moments* позволяет исследователю задать сценарий и затем сравнить симулированные и реальные данные с помощью значения правдоподобия.

Задача поиска параметров, дающих максимальное правдоподобие, была решена различными методами, например, в *moments* были использованы алгоритмы локального поиска: алгоритм Пауэлла [21], отмеченный авторами за свою эффективность, алгоритм Бройдена-Флетчера-Гольдфарба-Шанно (BFGS) [3, 7, 10, 24] и его модификация (L-BFGS-B) [1]. Несмотря на то, что алгоритмы локальной оптимизации обладают малой эффективностью на практике, *moments* остается действенным программным обеспечением для симулирования данных и вычисления правдоподобия. Недавно был представлен первый алгоритм глобальной оптимизации, основанный на генетическом алгоритме — GADMA (Genetic Algorithm for Demographic Model

Analysis) [9]. Он использует *moments* как один из симуляторов для вывода параметров демографических историй.

Однако, представленный метод, основанный на генетическом алгоритме, рекомендуется использовать до трех популяций. Это связано с тем, что вычисление значения правдоподобия в *moments* имеет экспоненциальную сложность от числа популяций, и задача эффективного и быстрого поиска параметров демографической истории для четырёх и особенно для пяти популяций до сих пор является вычислительно-сложной [13].

Первая глава данной работы посвящена основам популяционной генетики и теории байесовской оптимизации. В ней описаны основные определения и перечислены популярные методы для поиска параметров демографической истории по аллель-частотному спектру. Поставлена задача поиска параметров демографической истории для четырёх и пяти популяций.

Во второй главе представлена реализация предлагаемого метода и результаты экспериментальных исследований на симулированных и реальных данных.

ГЛАВА 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Основные положения популяционной генетики

Популяционная генетика — один из разделов генетики, который изучает распределение частот аллелей, а также их изменение под влиянием движущих сил эволюции, она пытается объяснить разнообразие внутри и между популяциями. Одним из базовых понятий генетики является понятие аллели — различной вариации гена в одинаковых местах (локусах) гомологичных хромосом. На основе информации об аллелях популяционная генетика позволяет построить историю развития популяций, называемую демографической историей.

На рисунке 1 приведён пример схематической структуры демографической истории, нарисованный с помощью программного обеспечения *moments* [13]. По оси ординат стоят отметки времени, когда популяции разделяются на две новые, начиная с предковой популяции в прошлом и до настоящего времени. Для численности популяций указывается масштаб (в левом верхнем углу).

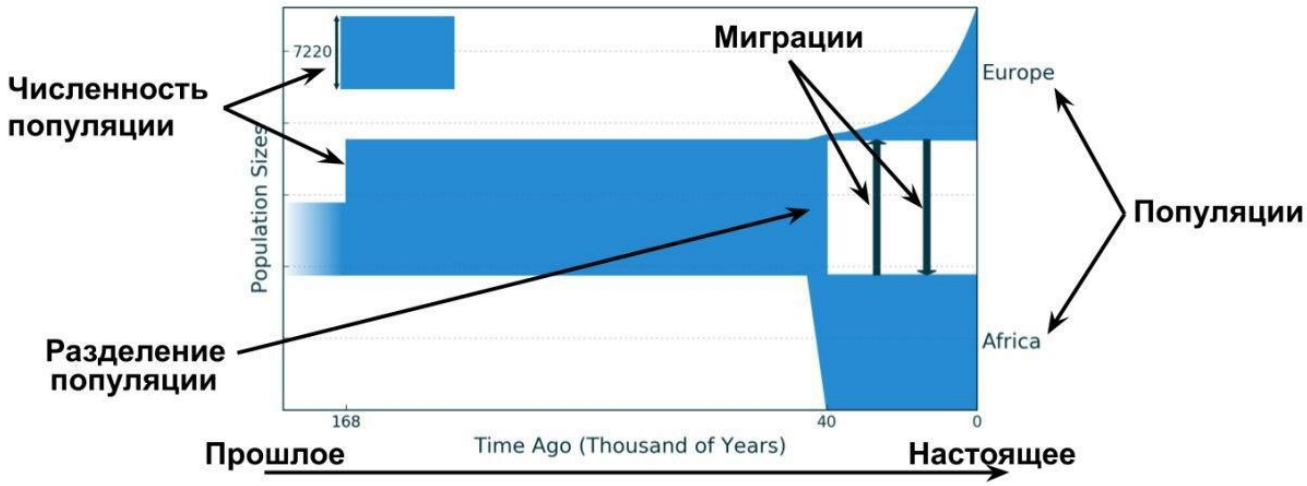


Рисунок 1 – Пример схематической структуры демографической истории [12]. Эта демографическая история показывает, что 168 тысяч лет назад существовала общая предковая популяция размера 7220 особей, которая затем внезапно выросла в 2 раза. Потом, 40 тысяч лет назад эта популяция разделилась на Африканскую и Европейскую. Последняя имела экспоненциальный рост. Также, после разделения между ними происходили миграции.

1.1.1. Аллель-частотный спектр

Обычно для анализа демографической истории не используют полные геномы. Это связано с отсутствием эффективных методов для большого объема данных. Вместо этого обычно используют различные сжатые представления данных [23].

Одним из наиболее популярных представлений генетических данных является аллель-частотный спектр (Allele Frequency Spectrum, AFS) — это совместное распределение частот полученных аллелей у P популяций [6]. Полученные аллели (derived allele) — это те аллели, которые в ходе эволюции произошли в результате мутаций от единого аллеля-предшественника, чаще всего они отличаются друг от друга заменой одного нуклеотида. Аллель-частотный спектр P популяций является P-мерным тензором $A \in \mathbb{N}^{(n_1+1)\times(n_2+1)\times\cdots\times(n_P+1)}$, где n_i — это количество хромосом в i-й популяции [6]. Каждый элемент спектра равен числу локусов, в которых полученная аллель встретилась в каждой из популяций у определенного числа особей. Более формально, элемент $A[d_1,\ldots,d_P] \in \mathbb{N}$, $d_i \in [0, n_i]$ — это число полученных аллелей, которые встретились у d_1 хромосом первой популяции, d_2 хромосом второй популяции, и так далее. Примеры AFS для моделей двух популяций представлены на рисунке 2.

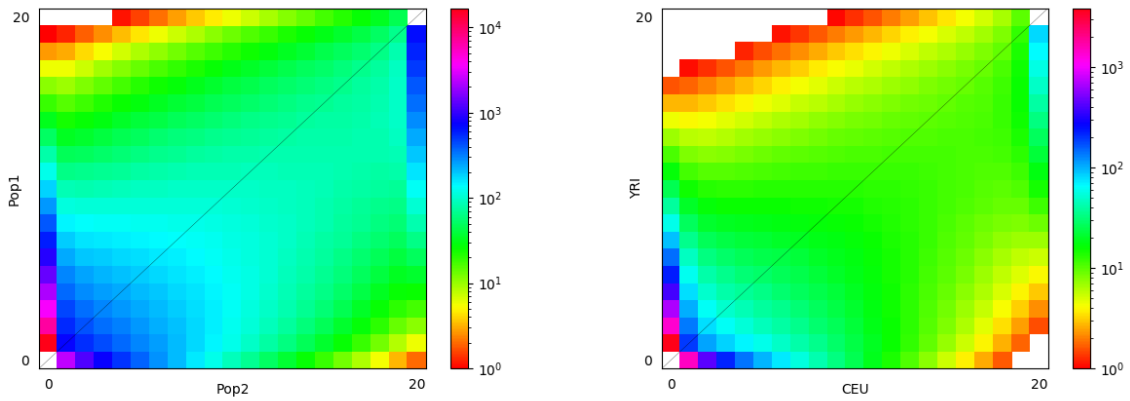


Рисунок 2 – Примеры аллель-частотного спектра двух популяций.

По построению, аллель-частотный спектр не учитывает информацию о связанности и позициях аллелей. Существуют и другие представления генетических данных, но AFS — это один из наиболее популярных данных

для получения информации об истории развития и эволюции популяций [2, 23].

В данный момент, подходы для вывода демографической истории нескольких популяций из наблюдаемого аллель-частотного спектра основаны на максимизации правдоподобия [12, 13]: по заданной демографической модели симулируется ожидаемый аллель-частотный спектр M и вычисляется логарифм правдоподобия $log(\mathcal{L}(M|S))$ с наблюдаемым аллель-частотным спектром S.

Правдоподобие — это мера, равная вероятности наблюдать спектр, построенный по геномным данным, при условии ожидаемого аллель-частотного спектра для известной демографической истории с заданными параметрами.

Для симулирования аллель-частотного спектра из демографической модели с некоторыми заданными параметрами используются разные методы. Среди них одними из самых популярных являются библиотеки для языка программирования Python:

— *∂a∂i* [12] численно решает дифференциальное уравнение диффузии в частных производных (Partial Differential Equation, PDE). При работе с демографическими моделями для большого числа популяций численное решение PDE занимает существенное время, поэтому данный метод поддерживает только до трех популяций;

— *moments* [13] основан на аппроксимации моментов случайного процесса. Этот метод был реализован в 2017 году, он способен поддерживать до пяти популяций. Но при этом, *moments* имеет бóльшую погрешность вычислений, чем *∂a∂i*.

Что примечательно, данные методы обладают одинаковым интерфейсом и имеют возможность вычислить логарифм правдоподобия $log(\mathcal{L})$ и максимизировать его различными методами локальной оптимизации. В данной работе выбран именно *moments*, потому что обладает возможностью поддерживать четыре и пять популяций из-за своей меньшей временной сложности по сравнению с *∂a∂i*. При сравнительном анализе в оригинальной работе [13] было показано, что *moments* в среднем работает в два раза быстрее, в работе по GADMA на экспериментальных исследованиях было достигнуто превосходство в 7.5 раз [9].

1.2. Сравнение на основе правдоподобия

Для определения того, насколько хорошо подобраны параметры для демографической модели, необходимо по этим параметрам построить ожидаемый аллель-частотный спектр с помощью *moments* и вычислить значение правдоподобия с наблюдаемыми данными. На рисунке 4 приведена общая схема оценки качества полученных параметров для метода аппроксимации моментов случайного процесса, реализованного в программном обеспечении *moments*.

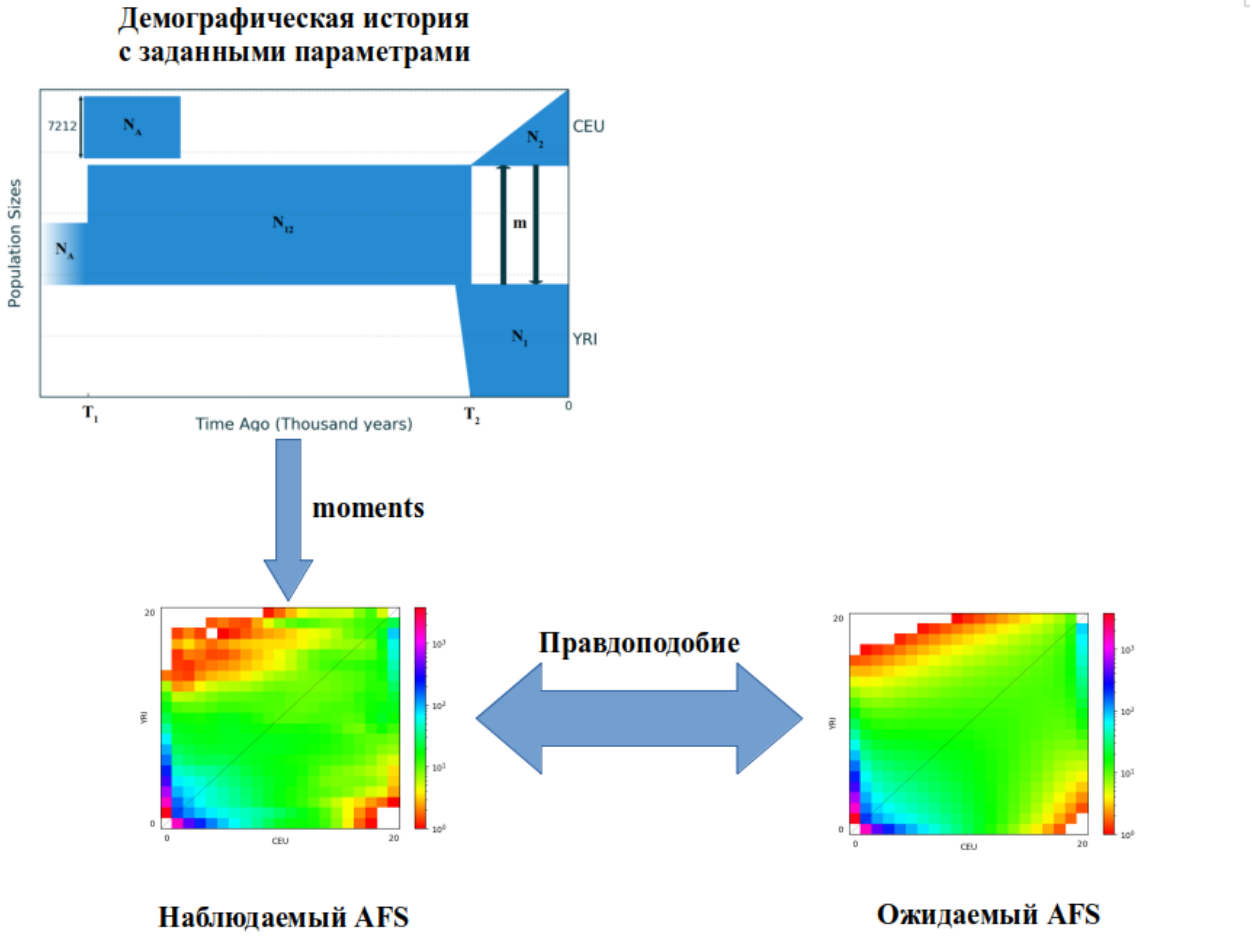


Рисунок 4 – Общая схема оценки параметров для *moments*. Чтобы сравнить заданные параметры демографической модели и реальный наблюдаемый AFS, *moments* симулирует ожидаемый AFS из демографической модели, и вычисляет правдоподобие между ожидаемым и наблюдаемым аллель-частотным спектром.

Сравнение демографических моделей происходит по значению правдоподобия — чем оно больше, тем лучше демографическая история описывает генетические данные. Пусть M — это ожидаемый AFS для

демографической истории P популяций с некоторыми параметрами. Так как аллель-частотный спектр предполагает независимость локусов, то каждый элемент AFS $S[d_1,..,d_P]$ — это независимая Пуассоновская величина [22] со средним $M[d_1,..,d_P]$. Тогда можно вычислить правдоподобие — вероятность получить наблюдаемый аллель-частотный спектр S, если ожидаемый равен M, как произведение $(n_1 + 1)(n_2 + 1)...(n_P + 1)$ Пуассоновских правдоподобий [12]:

$$\mathcal{L}(M|S) = \prod_{i=1,\cdots P} \prod_{d_i=1,\cdots n_i} \frac{e^{-M[d_1,\cdots,d_P]} M[d_1,\cdots,d_P]^{S[d_1,\cdots,d_P]}}{S[d_1,\cdots,d_P]!} \tag{1}$$

В данной работе логарифм правдоподобия $log(\mathcal{L}) \in [-\infty, 0]$ был использован в качестве целевой функции алгоритмов оптимизации.

Сложность вычисления логарифма правдоподобия $log(\mathcal{L})$ растёт экспоненциально от числа популяций [12, 13], что экспериментально подтверждается на рисунке 5, где представлено время вычисления правдоподобия для разного количества популяций и разного размера спектра с использованием *moments*. Число особей для построения каждого спектра было выбрано одинаковым для каждой популяции.

В качестве алгоритмов поиска параметров демографической истории, дающие максимальное значение логарифма правдоподобия, приведённые программные решения (∂a∂i, *moments*) предлагают только методы локального поиска, которые не эффективны для поиска глобального оптимума. Недавно, в работе [9] был представлен и реализован первый алгоритм глобальной оптимизации, основанный на генетическом алгоритме — GADMA (Genetic Algorithm for Demographic Model Analysis). Но задача эффективного и быстрого поиска параметров демографической истории для четырёх и особенно для пяти популяций до сих пор является вычислительно-сложной. К сожалению, генетический алгоритм не рекомендуется применять для четырёх и пяти популяций из-за длительности работы. Но среди алгоритмов глобального поиска существуют и другие подходы, которые потенциально смогут работать с большим числом популяций, например, байесовская оптимизация [17].

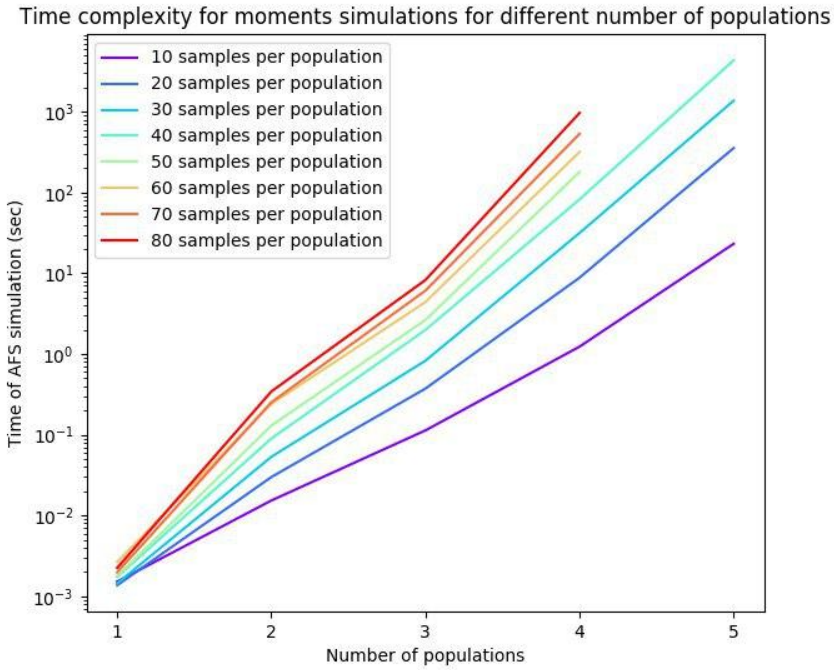


Рисунок 5 – Временные затраты на вычисление логарифма правдоподобия $log(\mathcal{L})$. По оси абсцисс — число популяций, по оси ординат — время для симуляции аллель-частотного спектра в логарифмической шкале. Цвета обозначают различные размеры спектров. В случае большого количества особей в 5 популяциях значения отсутствуют по причине слишком большой длительности вычисления.

1.3. Байесовская оптимизация

Байесовская оптимизация — это процедура глобальной оптимизации сложновычислимых функций, которые не имеют определённой структуры и требуют больших вычислительных затрат. Впервые данный подход был разработан Гарольдом Дж. Кушнером в 1964 [15], но само определение дал Джонас Моккус в своих работах по глобальной оптимизации [18, 20].

Байесовская оптимизация обычно применяется для целевых функций со следующими свойствами [8]:

— Непрерывность, для использования регрессии на основе Гауссовского процесса (подробнее в разделе 1.3.4);

— Размерность входных данных, равная числу параметров оптимизируемой функции, не очень большая;

— Пространство поиска оптимума имеет сложную структуру с наличием нескольких локальных минимумов и колебаниями абсолютного значения градиента, что мешает найти глобальный оптимум;

— Вычисление функции ограничено небольшим числом итераций (обычно несколько сотен), потому что каждая итерация занимает существенное время;
— Функция не дифференцируема в формальном виде;
— Требуется найти глобальный, а не локальный оптимум.

Процедура байесовской оптимизации состоит из двух частей: регрессии на основе Гауссовского процесса и функции выбора для определения точки для вычисления целевой функции на следующей итерации.

1.3.1. Гауссовский процесс

Случайная величина (random variable) — это измеримая функция $y = \xi(\omega)$, значения y которой численно выражают исходы ω случайного эксперимента.

Непрерывный случайный (стохастический) процесс — это семейство случайных величин $\{\mathbf{X}_t\}$, индексированных переменной $t \in T$, где $T \subset \mathbb{R}$ — произвольное непрерывное множество.

Гауссовский процесс (Gaussian process, $\mathcal{GP}$) — это непрерывный случайный процесс, такой что любой конечный набор случайных величин $\mathbf{X}_{t_1,\ldots,t_k} = (\mathbf{X}_{t_1}, \ldots, \mathbf{X}_{t_k})$, $k \in \mathbb{N}$, $t_i \in T$ имеет многомерное нормальное распределение $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ [16] с вектором средних значений $\boldsymbol{\mu}$ и ковариационной матрицей $\boldsymbol{\Sigma}$. Таким образом, любая конечная линейная комбинация случайных величин $Y = a_1 X_{t_1} + a_2 X_{t_2} + \cdots + a_k X_{t_k}$ нормально распределена. Гауссовский процесс полностью определяется функцией среднего $m(x)$ и ковариационной функцией $K(x, x')$:

$$f(x) \sim \mathcal{GP}(m(x), K(x, x')) \tag{2}$$

1.3.2. Ковариационная функция

Ковариационная функция (ядро) $K(x, x')$ — это функция, которая определяет ковариацию между двумя точками случайного процесса. Ковариационная функция используется при построении априорного распределения на основе Гауссовского процесса.

Функция ковариации позволяет описать некоторые свойства процесса, например, периодичность, тренд или гладкость [26]. Приведём некоторые из ковариационных функций:

— Константа: $K_C(x, x') = C$;

— Линейная функция: $K_{Lin}(x, x') = \sum_{i=1}^{d}(\sigma_i^2 x_i x_i')$;

— Периодическая функция: $K_{Per}(x, x') = \exp\left(-\frac{2\sin^2\left(\frac{x-x'}{2}\right)}{\ell^2}\right)$;

— Стационарные функции:

– Квадратичная экспоненциальная функция: $K_{SE}(x, x') = \sigma^2 \exp(-\frac{1}{2}r^2)$;

– Matérn 3/2: $K_{Matern3/2}(x, x') = \sigma^2(1 + \sqrt{3}r)exp(-\sqrt{3}r)$;

– Matérn 5/2: $K_{Matern5/2}(x, x') = \sigma^2(1 + \sqrt{5}r + \frac{5}{3}r)exp(-\sqrt{5}r)$,

где $r(x, x') = \sqrt{\sum_{i=1}^{d}\frac{(x_i - x_i')^2}{\ell_i^2}}$ — взвешенное Евклидово расстояние между двумя точками (x, x').

Ядра играют важную роль для описания процесса. В случае, если данные имеют различные типы параметров, то можно определить любую ковариационную функцию как линейную комбинацию других, более простых ковариационных функций, чтобы объединить информацию об имеющихся наборах данных [26].

1.3.3. Гиперпараметры ковариационной функции

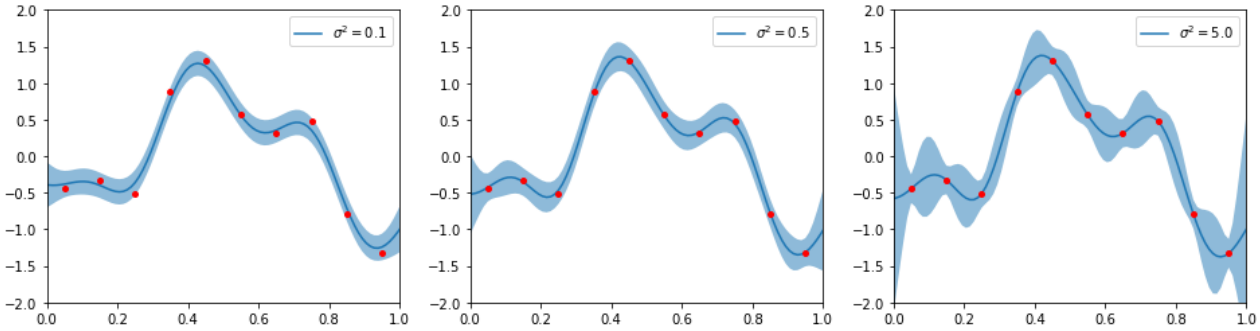
Ковариационные функции имеют параметры, которые обычно называются априорными гиперпараметрами. Например, для стационарных ядер, которые получили широкое практическое применение, определены следующие гиперпараметры:

— σ^2 — дисперсия (variance), отвечает за вертикальное изменение траектории функции;

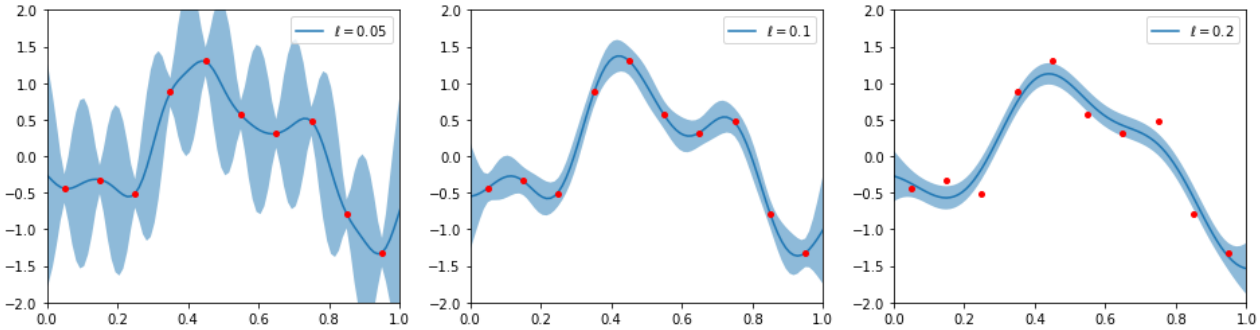
— ℓ^2 — масштаб (lengthscale), оценивает влияние расстояния между точками на корреляцию между ними.

На рисунке 6 показано, как изменяется среднее и доверительный интервал при изменении гиперпараметров ковариационной функции.

Заметим, что для стационарных ковариационных функций можно предположить независимость координат и использовать отдельный параметр ℓ_i для каждой из них. Такой подход называется автоматическим определением релевантности (automatic relevance determination, ARD) и имеет широкое применение [26]. Ковариационные функции с ARD более эффективны, благодаря большому числу гиперпараметров, но для подбора их значений требуется дополнительное время.



(а) Различные значения дисперсии $\sigma^2 = \{0.1, 0.2, 5.0\}$ при фиксированном масштабе $\ell = 0.1$. При увеличении дисперсии увеличивается вертикальное изменение траектории функции.



(б) Различные значения масштаба $\ell = \{0.05, 0.1, 0.2\}$ при фиксированной дисперсии $\sigma^2 = 1$. При увеличении масштаба уменьшается максимальное отклонение от среднего.

Рисунок 6 – Пример изменения регрессии на основе Гауссовского процесса при изменении гиперпараметров квадратичной экспоненциальной ковариационной функции K_{SE}. Красными точками изображена обучающая выборка. Синей линией на графиках изображено среднее, а голубая область отображает 95%-ный доверительный интервал.

Определение свойств данных и выбор ковариационной функции на их основе — это отдельная проблема, которая не рассматривается в данной работе. Для нашего исследования в качестве ковариационной функции используется наиболее часто используемая и рекомендуемая на практике [4, 25, 26] функция $Matern_{5/2}$ с использованием автоматического определения релевантности (ARD), так как в нашей задаче представлены демографические истории с параметрами различных типов.

1.3.4. Регрессия на основе Гауссовского процесса

Прогнозирующая часть байесовской оптимизации использует регрессию на основе Гауссовского процесса для моделирования функций. Пусть имеется конечный набор точек $X = x_1, \ldots, x_k$, $x_i \in \mathbb{R}^{N_D}$ и вектор значений целевой функции $f_{\overline{1,k}}(X) = f(x_1), \ldots, f(x_k)$. Предполагается, что выборка значений функции берётся случайным образом из некоторого

априорного распределения. Исходя из предположения об использовании Гауссовского процесса с функцией среднего $m(x)$ и ковариационной функцией $K(x, x')$, априорное распределение считается многомерным нормальным распределение с вектором средних значений $\mu_k = [m(x_i)]_{i=1}^k$ и ковариационной матрицей $\Sigma_k = \|K(x_i, x_j)\|_{i,j=1}^k$. Таким образом, априорное распределение будет равно:

$$f_{\overline{1,k}}(X) \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{3}$$

Полагая, что данные не зашумлены, то есть $y = f(x)$, рассмотрим уже известные значения целевой функции в n_0 точках. Требуется определить значение функции $f^* = f(x^*)$ в новой точке x^*. Пусть $k = n_0+1, x_k = x^*$. Тогда априорное распределение $[\mathbf{f}_{\overline{1,n_0}}, f^*]$ задаётся формулой 3. Используя теорему Байеса можно вычислить условное распределение f^* (жирным шрифтом выделены известные значения) [26]:

$$f^*|\mathbf{f}_{\overline{1,n_0}} \sim \mathcal{N}(\mu_{n_0}, \sigma_{n_0}^2),$$
$$\mu_{n_0}(x) = {k_*}^T \mathbf{\Sigma}_0^{-1}(\mathbf{f}_{\overline{1,n_0}} - \boldsymbol{\mu}_0) + m(x^*),$$
$$\sigma_{n_0}^2 = K(x^*, x^*) - {k_*}^T \mathbf{\Sigma}_0^{-1} k_*,$$

где

$$\boldsymbol{\mu}_0 = [m(x_i)]_{i=1}^{n_0},$$
$$\mathbf{\Sigma}_0 = \|K(x_i, x_j)\|_{i,j=1}^{n_0},$$
$$k_* = \left(K(x^*, x_1), \ldots, K(x^*, x_n)\right)^T.$$

Такое условное распределение называется апостериорным. Здесь апостеориорный вектор средних значений $\mu_{n_0}(x)$ представляет собой средневзвешенное значение между априорным вектором $\mu(x)$ значением и оценкой $\mathbf{f}_{\overline{1,n_0}}$, где веса задаются ковариационной функцией $K(x, x')$. Апостериорная дисперсия $\sigma_{n_0}^2$ равна априорной ковариации $K(x^*, x^*)$ без значения, соответствующего априорной дисперсии.

1.3.5. Функция выбора

Функция выбора (acquisition function) $\alpha(x)$ — это функция, основанная на вероятностной модели, используя которую можно выбрать точку для вычисления целевой функции на следующей итерации в процедуре

байесовской оптимизации. Функция выбора определяется так, чтобы точка, в которой функция достигает максимума, соответствовала потенциально большому значению оптимизируемой функции или области, про которую ещё ничего не известно.

Пусть имеется n точек $x_{\overline{1,n}} = x_1, \ldots, x_n$ с текущим наилучшим значением целевой функции $y_{best} = \max\limits_{i=\overline{1,n}} f(x_i)$. Требуется определить в какой следующей точке x_{new} производить вычисление. Тогда улучшением (improvement) будет функция $I(x) = \max(0, f(x) - y_{best})$. Приведём некоторые функции выбора:

— Максимальная вероятность улучшения (Maximum Probabity of Improvement, MPI): $\alpha_{MPI}(x) = \mathbb{P}(f(x) > y_{best}) = \Phi(\gamma(x))$, где $\Phi(\cdot)$ — функция стандартного нормального распределения $\mathcal{N}(0, 1)$, $\gamma(x) = \frac{\mu(x) - y_{best}}{\sigma(x)}$. Ожидается, что новое значение функции будет существенно меньше значений, которые наблюдались ранее.

— Ожидаемое улучшение (Expected Improvement, EI) [14, 19]: $\alpha_{EI}(x) = (y_{best} - \mu)\Phi\left(\frac{y_{best} - \mu}{\sigma}\right) + \sigma\phi\left(\frac{y_{best} - \mu}{\sigma}\right)$. Максимум достигается при одновременном увеличении как апостериорного среднего μ, так и среднеквадратичного отклонения σ.

Примеры различных функций выбора для произвольного апостериорного распределения в одномерном случае представлены на рисунке 8. Чтобы определить, в какой точке делать следующее вычисление, необходимо максимизировать функцию выбора. Для нахождения максимума функции выбора используются различные алгоритмы локальной оптимизации, например, запуск из различных точек алгоритма градиентного спуска L-BFGS [1] и выбор значения, соответствующего наилучшему результату.

1.3.6. Основной алгоритм байесовской оптимизации

Цель байесовской оптимизации — сделать как можно меньше вычислений целевой функции для поиска оптимума. Этот алгоритм решает задачу глобальной оптимизации:

$$x_{max} = \operatorname*{argmax}_{x \in X} f(x) \tag{4}$$

Так как вычисление целевой функции требует значительного времени, то для оптимизации используется следующий подход: после небольшого числа

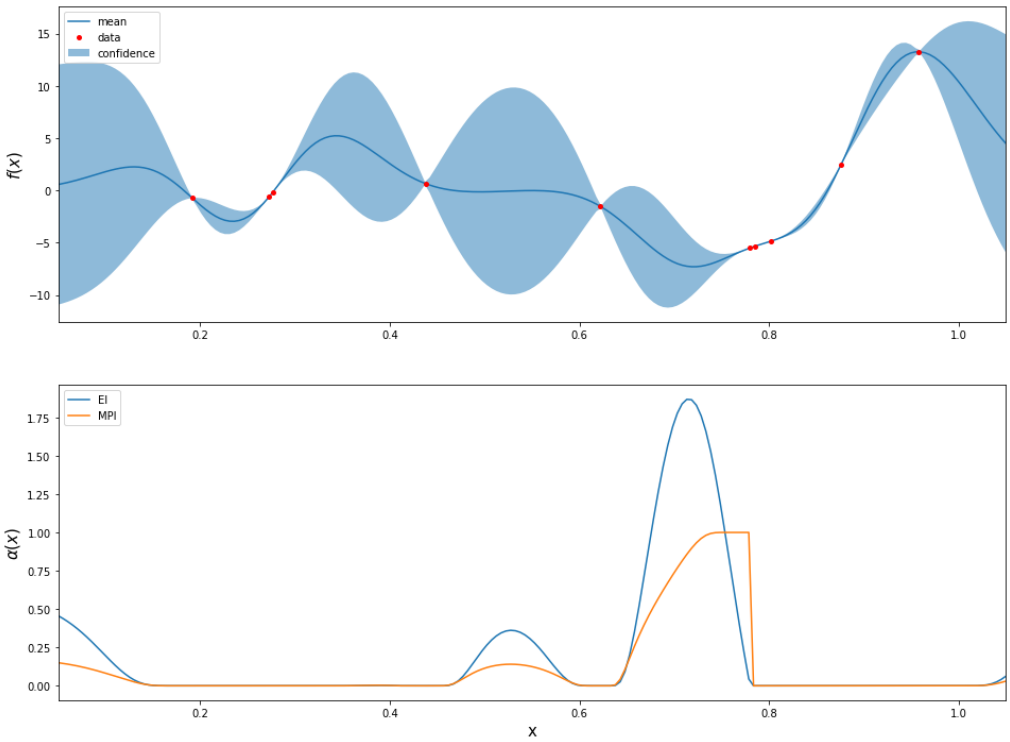


Рисунок 8 – Примеры различных функций выбора в одномерном случае.

итераций оценить возможную форму функции f и вычислять последующие точки только в тех местах, в которых достигается компромисс между возможностью получить лучшее значение и узнать больше о том, как устроена целевая функция. Общая схема данного подхода описана в листинге 1.

Листинг 1 – Псевдокод процедуры байесовской оптимизации.

Рассчитать априорное распределение Гауссовского процесса над f
Вычислить значение функции f в n_0 точках начального приближения
Задать максимальное число итераций n_{max}
for $i \leftarrow n_0, \ldots, n_{max}$ **do**
    Обновить апостериорное распределение над f, используя текущие точки
    Обновить функцию выбора, используя апостериорное распределение
    Найти значение x_{new}, которое соответствует максимуму функции выбора
    Добавить точку $(x_{new}, f(x_{new}))$ к существующим точкам
end for
return x_{max}, которая имеет максимальное значение целевой функции среди вычисленных в ходе оптимизации.
end

Использование функции выбора даёт возможность свести поиск решения задачи оптимизации целевой функции (4) к последовательному

решению серии следующих, более простых задач:

$$x_{t+1} = \underset{x \in X}{\operatorname{argmax}}\, \alpha(x|GP_t) \tag{5}$$

При этом, используя функцию выбора $\alpha(x)$ можно:

а) Обновить функцию выбора на основе нового апостериорного распределения;

б) Быстро вычислить точку максимума, используя алгоритм градиентного спуска L-BFGS;

в) И посчитать значение в новой точке.

Одна итерация процедуры байесовской оптимизации в случае одномерной функцией f из алгоритма 1 с использованием функции выбора EI представлена на рисунке 9.

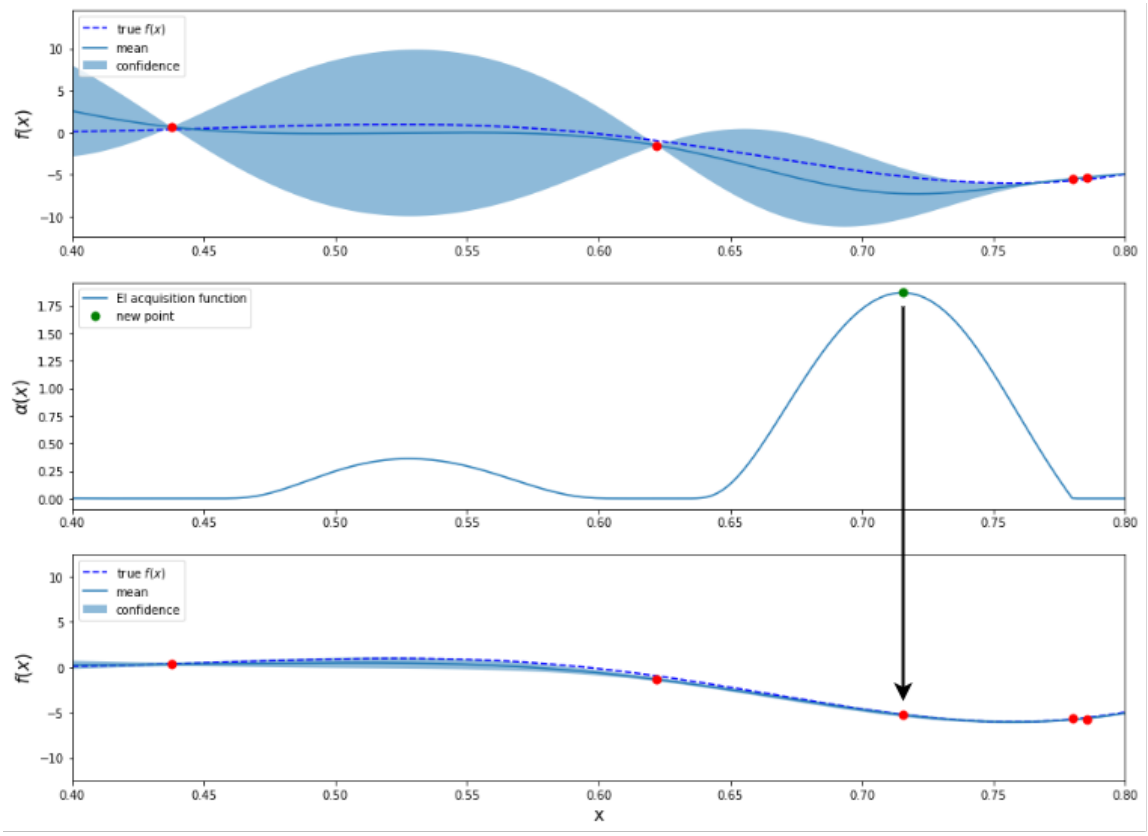


Рисунок 9 – На верхней панели изображены: пунктирной тёмно-синей линией истинное значение функции, красные точки — обучающая выборка, сплошная голубая линия — среднее, область вокруг — 95%-ный доверительный интервал. На второй панели показана функция выбора EI и зелёная точка, которая максимизирует функцию выбора. В ней снова вычисляется функция, что отображено на нижней панели.

1.4. Постановка задачи

Пусть имеется P популяций размера n_1 особей (хромосом) первой популяции, n_2 особей второй популяции, и так далее. Рассмотрим функцию $f_D(\Theta, A)$, которая построена для заданной демографической истории D. Она зависит от аллель-частотного спектра A, построенного по геномным данным и параметров демографической истории $\Theta = \{\theta_i\}_{i=1}^{N_D}, N_D \leqslant 20$ и возвращает меру $log(\mathcal{L})$ соответствия параметров данному аллель-частотному спектру.

Целью данной работы является разработка и анализ метода, основанного на байесовской оптимизации, для поиска параметров демографической истории четырёх и пяти популяций по геномным данным. Более формально:

Вход

— D — демографическая история для P популяций;

— $A \in \mathbb{N}^{(n_1+1)\times(n_2+1)\times\cdots\times(n_P+1)}$ — P-мерная матрица, $P \in \{4,5\}$.

Выход

— Набор $\Theta_{best} \in \mathbb{R}^{N_D}$ значений, который максимизирует значение функции f:

$$\Theta_{best}: \quad f_D(\Theta_{best}, A) \to \max$$

Так как данная функция обладает свойствами, перечисленными выше, а именно непрерывна, имеет небольшое число параметров, не представима в “closed-form expression” и вычислительно-сложна, то алгоритм байесовской оптимизации применим для решения поставленной задачи.

1.5. Параметры демографической истории

Определим набор параметров демографической истории $\Theta = \{\theta_i\}_{i=1}^{N_D}, N_D \leqslant 20$, которые входят в нашу оптимизируемую функцию.

— **N** — численность популяций;

— **T** — время разделения популяций;

— **m** — миграции между популяциями.

Пример изображения параметров на схематической структуре демографической истории приведен на рисунке 10.

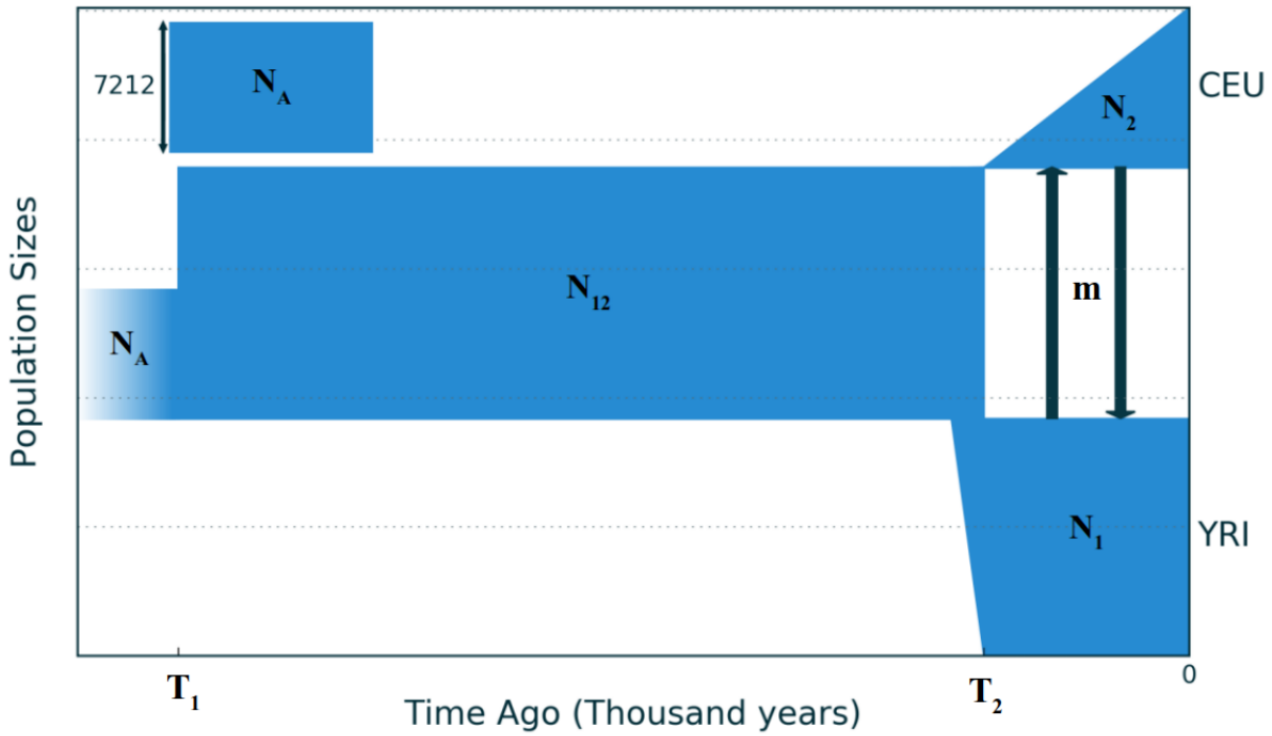


Рисунок 10 – Изображение параметров демографической истории двух популяций.

Выводы по главе 1

В данной главе приведены основные определения из предметной области, такие как аллель и демографическая история, описано популярное представление генетических данных — аллель-частотный спектр. Также, были перечислены основные методы для поиска параметров демографической истории по аллель-частотному спектру. Существующие методы имеют экспоненциальную сложность от числа популяций и неэффективны для поиска параметров демографической истории для четырёх, а особенно для пяти популяций. Была поставлена задача поиска параметров демографической истории для четырёх и пяти популяций. Для решения данной задачи в текущей работе предложена байесовская оптимизация. В главе описана схема алгоритма байесовской оптимизации и её основные элементы: регрессия на основе Гауссовского процесса и функция выбора.

ГЛАВА 2. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

Метод, основанный на байесовской оптимизации, для вывода параметров демографической истории из аллель-частотного спектра был реализован на основе существующего решения. Эффективность метода была проверена на симулированных и реальных данных. Было проведено сравнение предложенного метода с генетическим алгоритмом, представленным в GADMA, и со случайным поиском. В качестве функций выбора для процедуры байесовской оптимизации были выбраны: максимальная вероятность улучшения (Maximum Probabity of Improvement, MPI) и ожидаемое улучшение (Expected Improvement, EI).

2.1. Существующие реализации байесовской оптимизации

Байесовская оптимизация используется в различных областях и реализована на многих языках программирования, таких как Python, C++, Matlab, Java, R. В данной работе был выбран Python, так как основные инструменты для данной области (*∂a∂i*, *moments*, GADMA) реализованы именно на нём. Для этого языка программирования существует несколько реализаций байесовской оптимизации:

— BoTorch;
— GPyOpt;
— Spearmint.

BoTorch и Spearmint предназначен для параллельного вычисления на кластерах. Для решения нашей задачи был выбран именно GPyOpt [11], разработанный группой машинного обучения Университета Шеффилда, которая считается ведущей в этой области.

2.2. Реализация

Для нахождения максимума функции выбора в реализации GPyOpt используется алгоритм градиентного спуска L-BFGS [1]. Он запускается из 5 точек из равномерного распределения, и среди них выбирается та точка, в которой достигается максимум функции выбора. При этом, алгоритм сильно зависит от начальных точек и на текущей итерации не учитывает предыдущие вычисления целевой функции. Предполагая, что глобальный оптимум будет расположен в окрестностях текущих наилучших точек, было добавлено такое же количество (5) дополнительных точек, из которых

также запускается алгоритм L-BFGS. Эти дополнительные точки равномерно распределены вокруг точек, соответствующих текущим лучшим значениям целевой функции. Такой подход позволяет уточнить лучшее решение на поздних итерациях байесовской оптимизации. В последующих сравнениях представлено именно это решение, так как оно показало себя лучше на различных данных. Пример сравнения модифицированного (BayesOpt) алгоритма с классическим (GPyOpt) подходом представлен на рисунке 11.

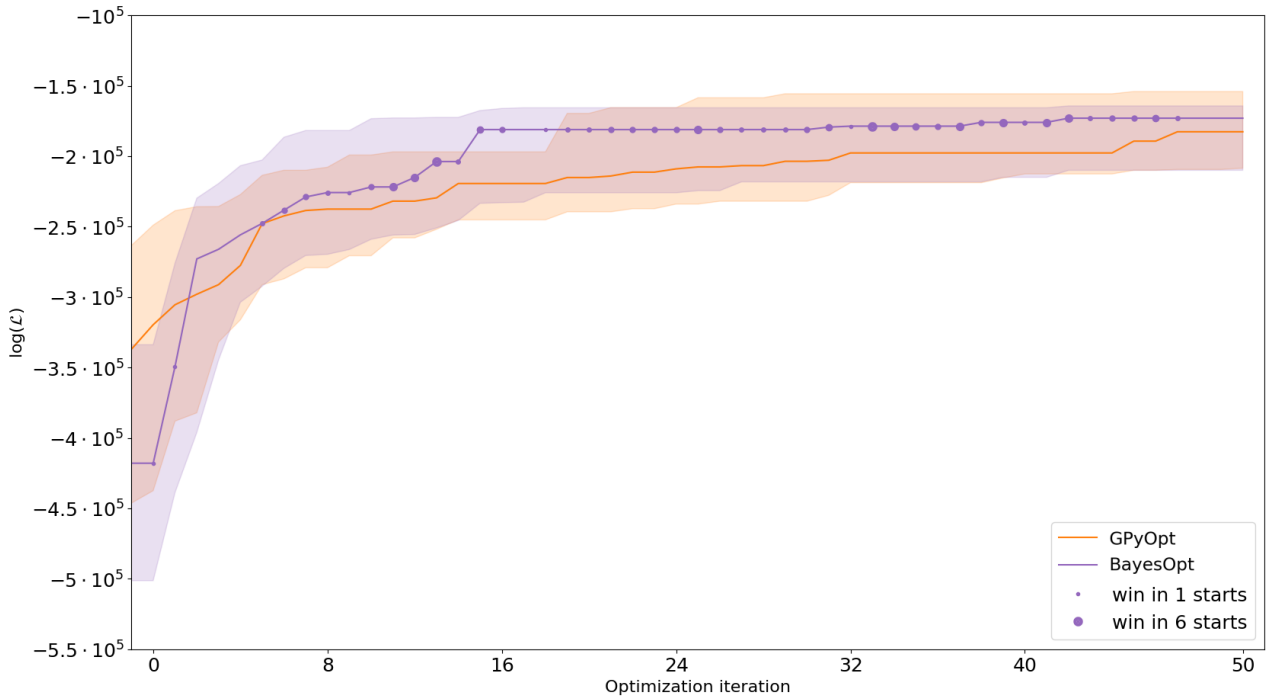


Рисунок 11 – График сходимости классического и модифицированного алгоритма байесовской оптимизации (32 запуска). Для наглядности представлены только итерации оптимизации, без начального приближения. Выделенные точки отражают выбор предложенной гипотезы (win в легенде). Размер точки равен количеству запусков, в которых подтвердилась гипотеза.

Перед запуском всех методов, для которых проводилось сравнение, выполняется случайный поиск по небольшому числу точек для поиска начального приближения (initial design). Для построения выборки этих точек было использовано распределение, которое зависит от типа параметра. Оно показало себя наилучшим для случайного поиска в работе про генетический алгоритм [9]. Для времени разделения или темпов миграции берётся выборка из равномерного распределения U по области значений. А для численности популяций точки выбираются из логнормального распределения со средним μ, равным размеру предковой популяции N_A, и среднеквадратическим

отклонением, покрывающим область значений по правилу 3σ. Количество точек для начального приближения требовалось выбрать небольшим, так как функция сложновычислима, но при этом сравниваемые методы генетического алгоритма и байесовской оптимизации должны были иметь близкие значения после начального приближения, чтобы обеспечить корректное сравнение.

Также, для байесовской оптимизации входные значения параметров были логарифмированы, как предлагается в локальных оптимизациях *moments* и ∂a∂i.

2.3. Симулированные данные

Для проверки эффективности предложенного метода байесовской оптимизации были проведены запуски на симулированных аллель-частотных спектрах для демографической истории четырёх и пяти популяций. Симуляции были проведены с помощью *moments*, поэтому для них были заданы оптимальные параметры и значение максимально возможного правдоподобия. Для симуляций размер предковой популяции составляет 10000 особей.

Для всех экспериментов сравнение проводилось на небольшом количестве итераций, потому что на последующих итерациях методы имеют схожее поведение и теряется наглядность.

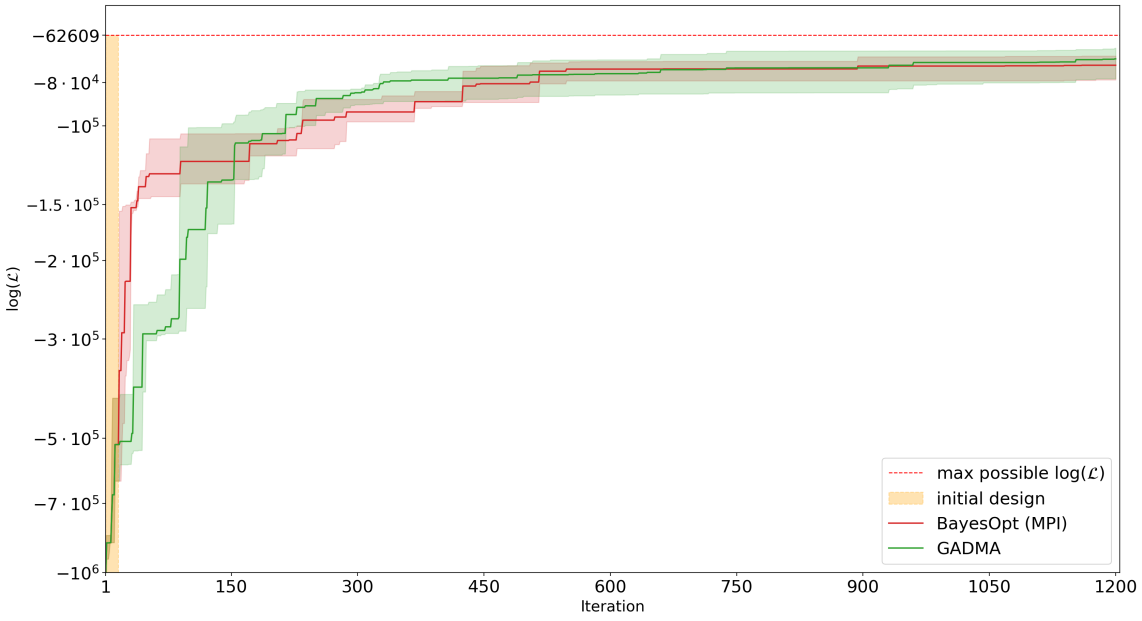


Рисунок 12 – График сходимости методов на большом количестве итераций. Модель — симулированные данные, 4 популяции (подробнее — далее).

Например, на рисунке 12 представлен график сходимости на 1200 итерациях в случае симулированных данных для четырёх популяций. По оси абсцисс — итерации алгоритмов. Для всех алгоритмов итерация — это вычисление целевой функции, то есть правдоподобия. По оси ординат — текущие лучшие значения правдоподобия в логарифмической шкале.

2.3.1. Четыре популяции

Аллель-частотный спектр размера $21 \times 21 \times 21 \times 21$ был симулирован с помощью *moments* для демографической модели четырёх популяций с девятью параметрами, заданные значения которых представлены в таблице 1.

Таблица 1 – Заданные значения параметров симулированной демографической модели четырёх популяций. Размеры популяций представлены в особях, время в поколениях.

Название параметра	Значение	Описание
N_A	10,000	Размер предковой популяции в прошлом.
N_1	15,000	Размер популяции 1 после отделения от предковой.
N_{234}	8,000	Размер второй образовавшейся популяции (общей для популяций 2, 3, и 4) при разделении предковой.
N_2	10,000	Размер популяции 2 после ее отделения.
N_{34}	5,000	Размер общей популяции популяций 3 и 4 после отделения популяции 2.
N_3	2,000	Размер популяции 3 после ее отделения.
N_4	3,000	Размер популяции 4 после ее отделения.
T_1	6,000	Время отделения популяции 1.
T_2	4,000	Время отделения популяции 2.
T_3	1,000	Время разделения популяции 3 и популяции 4.

Время одного вычисления целевой функции представлено на рисунке 13 и в среднем составляет 30 секунд.

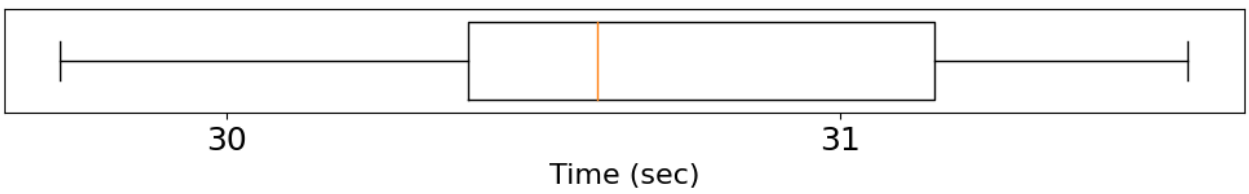


Рисунок 13 – Диаграмма размаха времени одной итерации на симулированных данных, 4 популяции. Медианное значение — 30.608 сек.

Для построения диаграммы размаха временных затрат значение логарифма правдоподобия было вычислено на 1000 точках из равномерного распределения.

На рисунке 14 представлена схематическая структура рассматриваемой демографической модели с оптимальными параметрами.

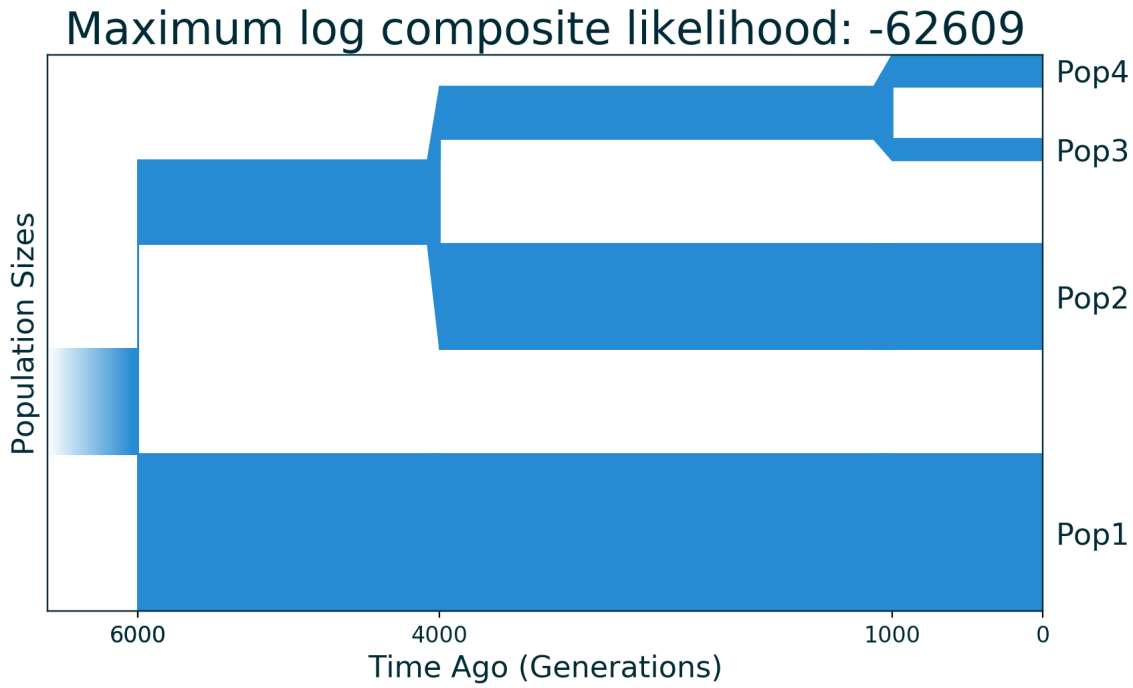


Рисунок 14 – Симулированная демографическая модель четырёх популяций, которая показывает, что 6000 поколений назад существовала общая предковая популяция размера 10 тысяч особей, которая разделилась на 2 популяции размера 15 и 8 тысяч особей. Последняя через 2000 поколений тоже разделилась на 2 популяции размера 10 и 5 тысяч особей. И ещё через 3000 поколений от последней популяции отделились другие две популяции размером в 2 и 3 тысячи особей, соответственно.

Байесовская оптимизация с разными функциями выбора (MPI, EI), генетический алгоритм, а также случайный поиск были запущены по 32 раза для поиска параметров демографической истории четырех популяций. По результатам запусков был построен график сходимости методов (рисунок 15), на котором представлены 256 итераций с 16 точками начального приближения. В среднем, байесовская оптимизация с функцией выбора MPI показала себя лучше, чем с функцией выбора EI, и оба метода показывают лучшую сходимость на первых (MPI — 70, EI — 30) итерациях, чем генетический алгоритм. Как и ожидалось, случайный поиск проигрывает всем рассмотренным методам в сходимости.

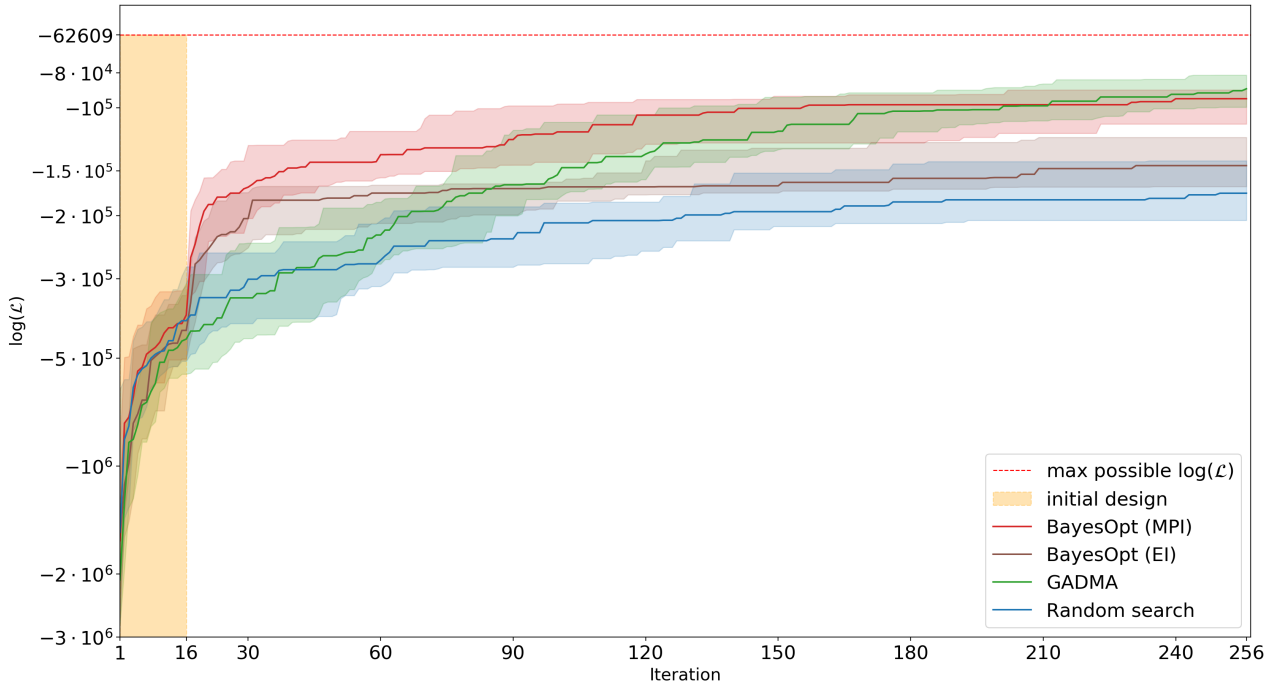


Рисунок 15 – График сходимости методов на **симулированных** данных для **четырёх** популяций. По оси абсцисс — итерации алгоритмов, ординат — текущие лучшие значения правдоподобия в логарифмической шкале. Цветные линии — вторая квартиль (медиана) множества из 32 значений правдоподобия на каждой итерации, область вокруг — диапазон между первой (0.25) и третьей (0.75) квартилями. В среднем, байесовская оптимизация с функцией выбора MPI показала лучшую сходимость на первых 70 итерациях.

2.3.2. Пять популяций

Далее была рассмотрена демографическая модель для пяти популяций. Модель также включает в себя девять параметров, заданные значения которых представлены в таблице 2. Аллель-частотный спектр размера $11 \times 11 \times 11 \times 11 \times 11$ был симулирован с помощью *moments*. Время одного вычисления целевой функции представлено на рисунке 16 и в среднем составляет 59 секунд. На рисунке 17 представлена схематическая структура используемой демографической модели с выбранными оптимальными параметрами.

Как и в случае четырех популяций, был построен график сходимости методов (рисунок 18) для первых 200 итераций. Каждый алгоритм был запущен 16 раз и первые 10 точек были использованы для начального приближения. По результатам сравнения сходимости методов байесовская оптимизация вновь продемонстрировала превосходство на первых итерациях.

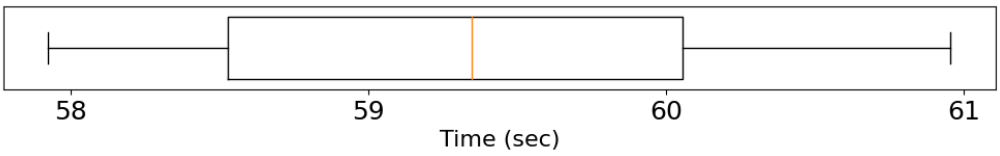


Рисунок 16 – Диаграмма размаха времени одной итерации на симулированных данных, 5 популяций. Медианное значение — 59.346 сек.

Таблица 2 – Заданные значения параметров симулированной модели для пяти популяций. Размеры популяций представлены в особях, время в поколениях.

Название параметра	Значение	Описание
N_A	10,000	Размер предковой популяции в прошлом.
N_1	10,000	Размер популяции 1 после отделения от предковой.
N_2	20,000	Размер популяции 2 после ее отделения.
N_3	15,000	Размер популяции 3 после ее отделения.
N_4	10,000	Размер популяции 4 после ее отделения.
N_5	5,000	Размер популяции 5 после ее отделения.
T_1	7,000	Время отделения популяции 1.
T_2	6,000	Время отделения популяции 2.
T_3	4,000	Время отделения популяции 3.
T_4	1,000	Время разделения популяции 4 и популяции 5.

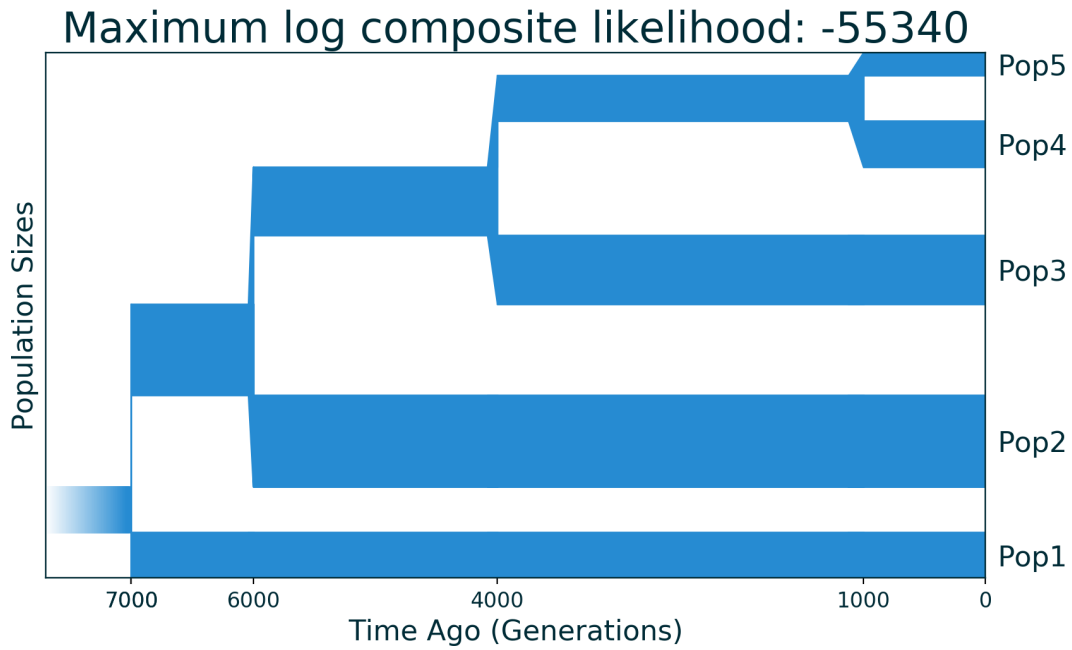


Рисунок 17 – Симулированная демографическая модель пяти популяций, которая показывает, что 7000 поколений назад существовала общая предковая популяция размера 10 тысяч особей, от которой отделилась популяция 1 размера 10 тысяч особей и так далее.

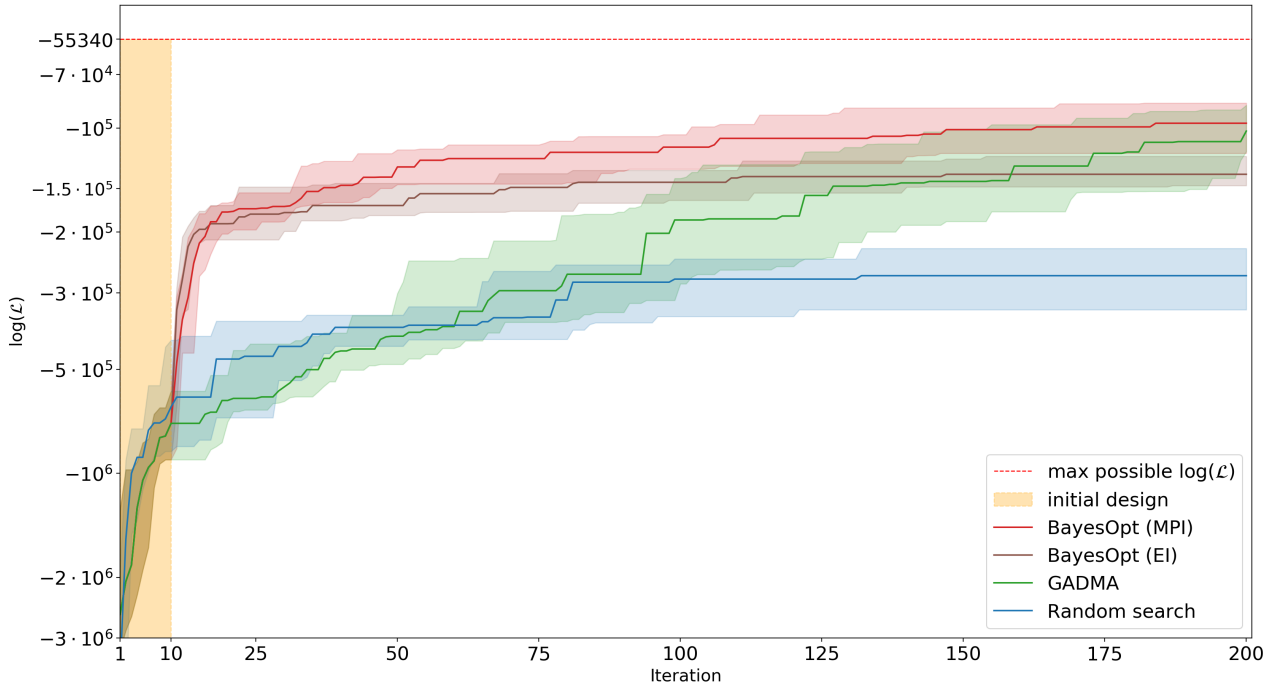


Рисунок 18 – График сходимости рассматриваемых методов на **симулированных** данных для **пяти** популяций. По оси абсцисс — итерации алгоритмов. По оси ординат — текущие лучшие значения правдоподобия в логарифмической шкале. Для всех алгоритмов представлены квартили множества из 16 значений правдоподобия на каждой итерации: цветные линии — вторая квартиль, область вокруг — диапазон между первой и третьей квартилями. Версия с функцией выбора MPI показала лучшую сходимость на показанных 200 итерациях.

2.4. Реальные данные

Реальные данные для экспериментальных исследований были взяты из статьи про *moments*, а именно данные, для которых была построена демографическая история “Out-of-Africa” [5, 12] четырёх популяций:

— люди народа Йоруба из города Ибадан, Нигерия (**YRI**);

— жители штата Юта с предками из северной и западной Европы (**CEU**);

— китайцы народа Хань из Пекина (**CHB**);

— и жители Японии из Токио (**JPT**) (добавлено в [13]).

2.4.1. Четыре популяции

Аллель-частотный спектр размера $21 \times 21 \times 21 \times 21$ был предоставлен авторами статьи [13] для демографической модели четырёх популяций, которая содержит семнадцать параметров. Их значения, найденные авторами статьи, представлены в таблице 3.

Время одного вычисления целевой функции представлено на рисунке 19 и в среднем составляет 713 секунд.

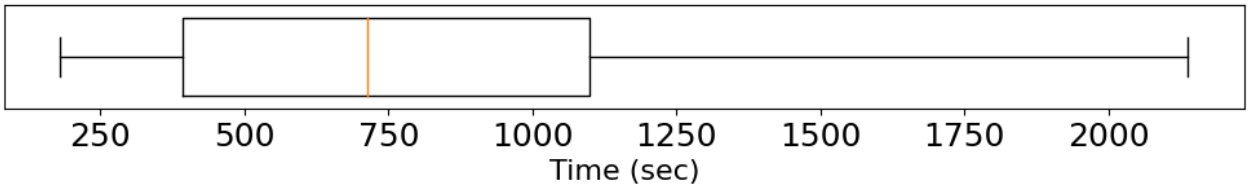


Рисунок 19 – Диаграмма размаха времени одной итерации на реальных данных, 4 популяции. Медианное значение — 713.757 сек.

На рисунке 20 представлена схематическая структура рассматриваемой демографической модели с параметрами, найденными авторами статьи.

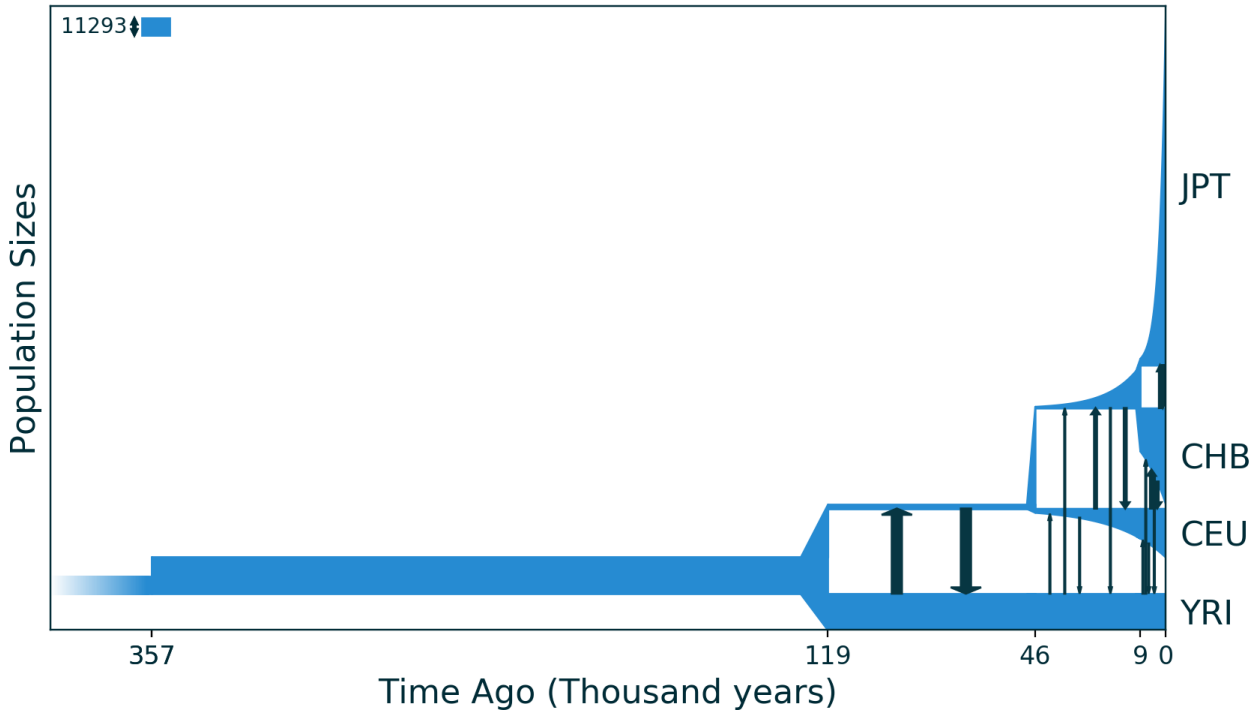


Рисунок 20 – Предполагаемая реальная модель популяций африканского (YRI), азиатского (CHB и JPT) и европейского (CEU) происхождения, полученная в статье [13].

На рисунке 21 по результатам 20 запусков изображен график сходимости различных методов на 100 итерациях с 10 точками начального приближения. Значение максимально возможного правдоподобия соответствует значению правдоподобия, полученного для параметров, найденных в работе по *moments*.

Таблица 3 – Параметры для демографической модели для четырёх популяций, представленной в статье про *moments*. Размеры популяций представлены в особях, темпы миграций в числе особей на поколение, время в тысячах лет (kilo years ago, kya).

Название параметра	Значение	Описание
N_A	11,293	Размер предковой популяции в прошлом.
N_{Af}	23,721	Размер предковой популяции после внезапного роста.
N_B	2,831	Размер Евроазиатской популяции после выхода из Африки (отделения от Африканской популяции).
N_{Eu0}	2,512	Размер Европейской популяции сразу после ее отделения
N_{Eu}	31,721	Размер Европейской популяции после экспоненциального роста.
N_{As0}	1,019	Размер популяции Китая после разделения с Европейской популяцией.
N_{As}	62,653	Размер популяции Китая после экспоненциального роста в настоящий момент времени.
N_{Jp0}	4,384	Размер популяции Японии в момент отделения от популяции Китая.
N_{Jp}	234,114	Размер популяции Японии после экспоненциального роста.
m_{Af-B}	16.8×10^{-5}	Темпы миграций между Африканской популяцией и Евроазиатской.
m_{Af-Eu}	1.14×10^{-5}	Темпы миграций между Африканской популяцией и Европейской.
m_{Af-As}	0.56×10^{-5}	Темпы миграций между Африканской популяцией и популяцией Китая.
m_{Eu-As}	4.75×10^{-5}	Темпы миграций между Европейской популяцией и популяцией Китая.
m_{Ch-Jp}	3.3×10^{-5}	Темпы миграций между популяцией Китая и Японии.
T_{Af}	357	Время внезапного роста численности предковой популяции.
T_B	119	Время выхода людей из Африки — разделения Африканской и Евроазиатской популяций.
T_{Eu-As}	46	Время разделения Европейской популяции и популяции Китая.
T_{Ch-Jp}	9	Время отделения популяции Японии от популяции Китая.

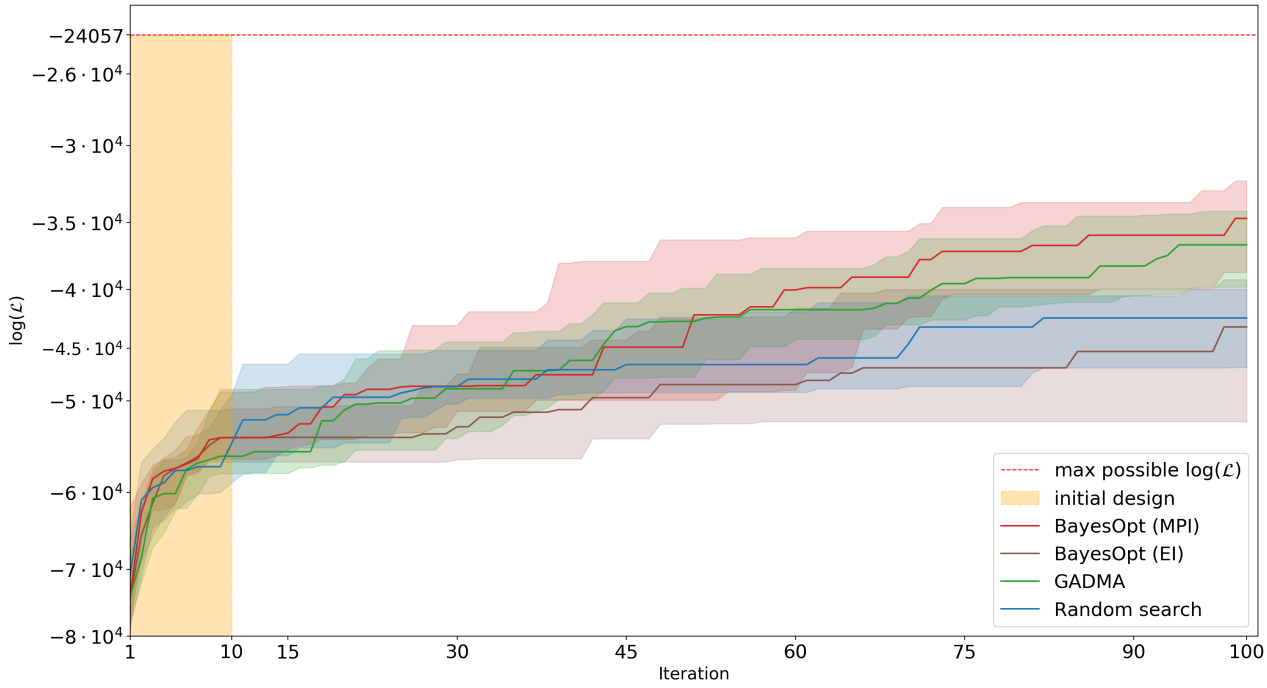


Рисунок 21 – График сходимости рассматриваемых методов на **реальных** данных для четырёх популяций. По оси абсцисс — итерации алгоритмов. По оси ординат — текущие лучшие значения правдоподобия в логарифмической шкале. Для всех алгоритмов представлены квартили множества из 20 значений правдоподобия на каждой итерации: цветные линии — вторая квартиль (медиана), область вокруг — диапазон между первой и третьей квартилями.

В результате данного исследования превосходство байесовской оптимизации не было выявлено. Учитывая, что для предлагаемого алгоритма рекомендуемое количество параметров не больше 20, то была выдвинута гипотеза, что низкая эффективность связана с большим количеством (17) параметров в исследуемой модели.

2.4.2. Три популяции

Чтобы выявить, является ли большое число параметров причиной плохой сходимости, было проведено исследование на той же модели, но без японской популяции из Токио **(JPT)**. Такая модель включает в себя тринадцать параметров, в том числе четыре миграции. Был взят аллель-частотный спектр размера $41 \times 41 \times 41$, полученный в работе [13]. На рисунке 22 по результатам 16 запусков представлен график сходимости рассматриваемых методов на 200 итерациях с 10 точками начального приближения.

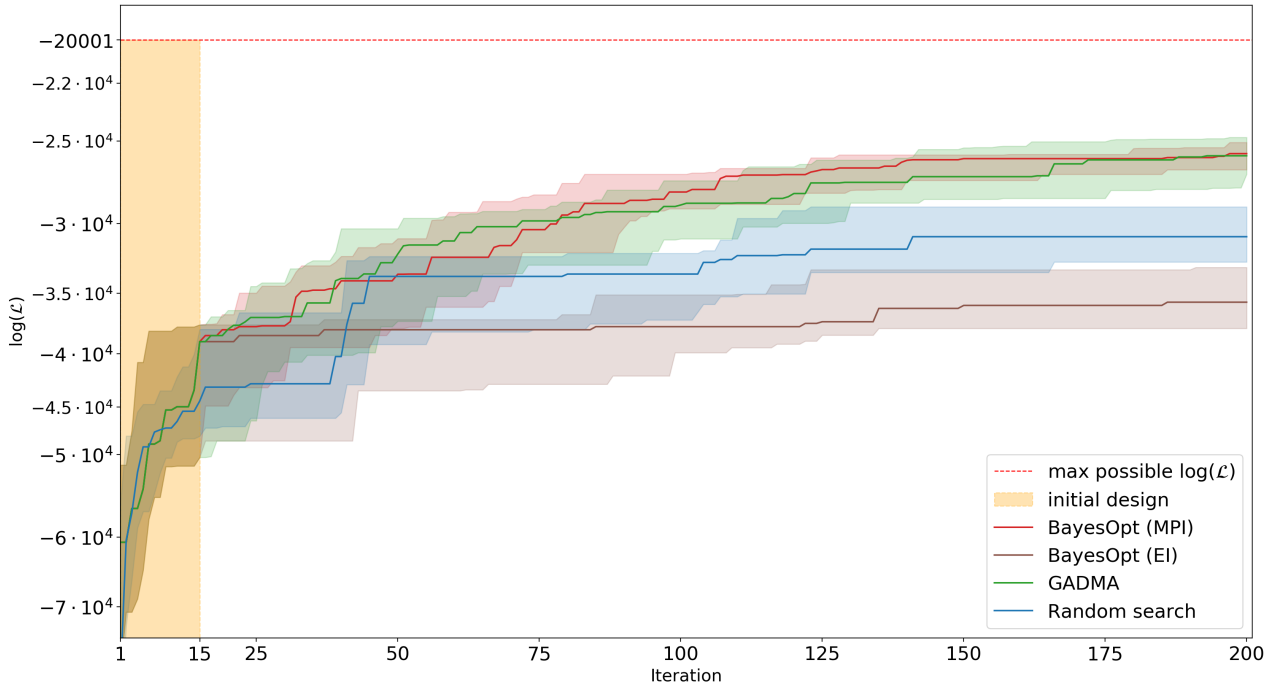


Рисунок 22 – График сходимости рассматриваемых методов на реальных данных для **трёх** популяций. Стоит отметить схожее с предыдущей моделью поведение.

Исходя из проведённого эксперимента можно предположить, что уменьшение количества параметров для данной модели не решило проблемы с плохой сходимостью.

Симулированные данные, на которых байесовская оптимизация показала себя успешно, не содержат среди параметров темпов миграций. В реальных моделях трёх и четырёх популяций присутствуют миграции. Поэтому была выдвинута новая гипотеза о том, что наличие миграций замедляет сходимость исследуемых методов.

2.4.3. Четыре популяции без миграций

Была исследована та же демографическая модель для четырёх популяции, но без учёта миграций. То есть значения параметров, отвечающих за темпы миграций, приведённые в таблице 3 не учитываются, и тогда можно считать, что данная модель содержит только двенадцать параметров. Время одного вычисления целевой функции представлено на рисунке 23 и в среднем составляет 31 секунду. Стоит отметить значительное уменьшение времени вычисления относительно модели с миграциями.

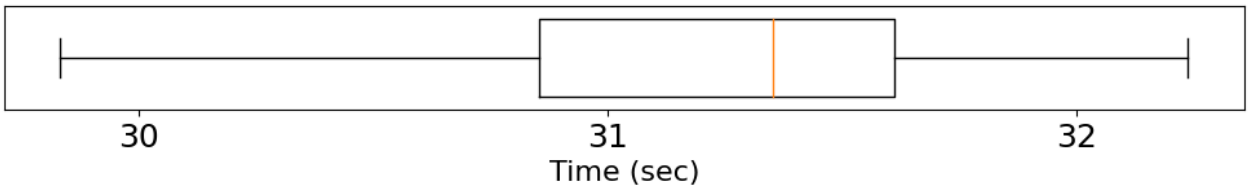


Рисунок 23 – Диаграмма размаха времени одной итерации на реальных данных, 4 популяции, **без миграций**. Медианное значение — 31.328 сек.

На рисунке 24 по результатам 32 запусков представлен график сходимости различных методов на 200 итерациях с 10 точками начального приближения.

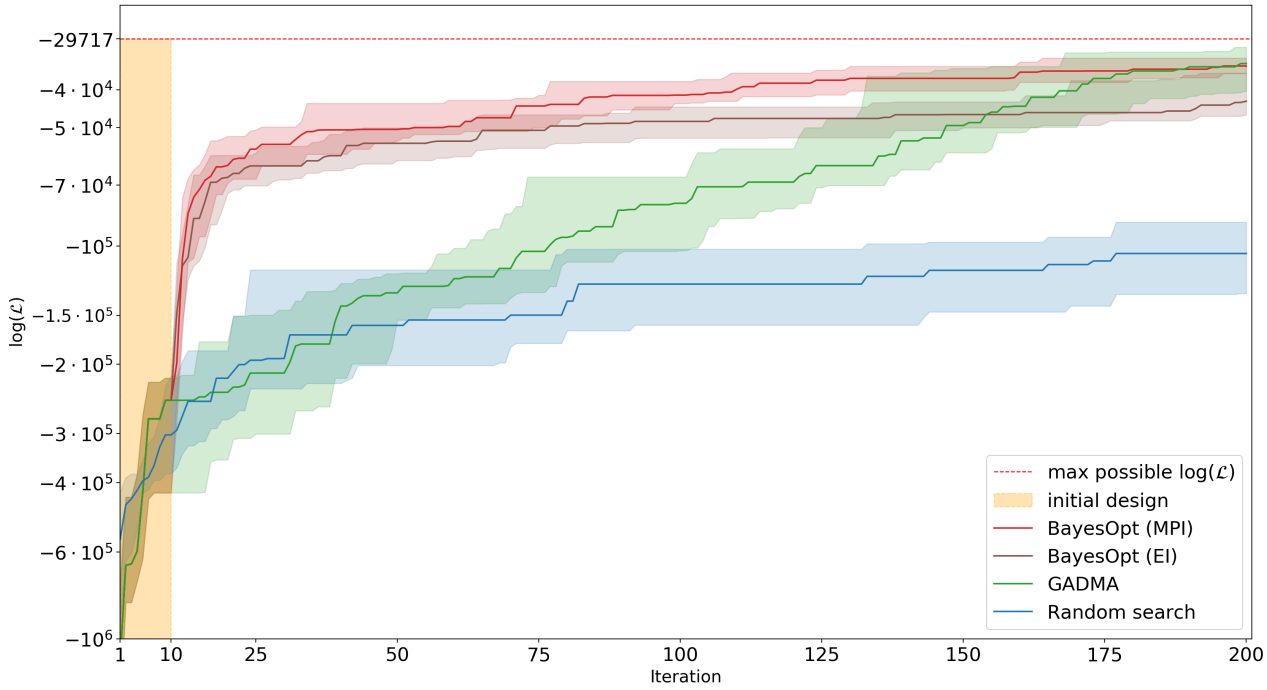


Рисунок 24 – График сходимости рассматриваемых методов на реальных данных для четырёх популяций **без учёта миграций**.

Байесовская оптимизация показала превосходство в сходимости на первых итерациях по сравнению с другими алгоритмами. Поэтому можно сделать предположение, что наличие темпов миграций среди параметров демографических историй отрицательно влияет на сходимость исследуемых методов.

2.5. Комбинированный подход

В качестве последнего экспериментального исследования был реализован и протестирован комбинированный подход: на первых 40 итерациях использовалась байесовская оптимизация с функцией выбора MPI,

после чего полученные точки были переданы для генетического алгоритма в качестве начального приближения. Эффективность разработанного подхода была проверена на симулированных данных четырех популяций (подробнее в подразделе 2.3.1). На рисунке 25 по результатам 32 запусков представлен график сходимости различных методов на 200 итерациях с 16 точками начального приближения (результаты запусков генетического алгоритма и байесовской оптимизации с функцией выбора MPI взяты из подраздела 2.3.1).

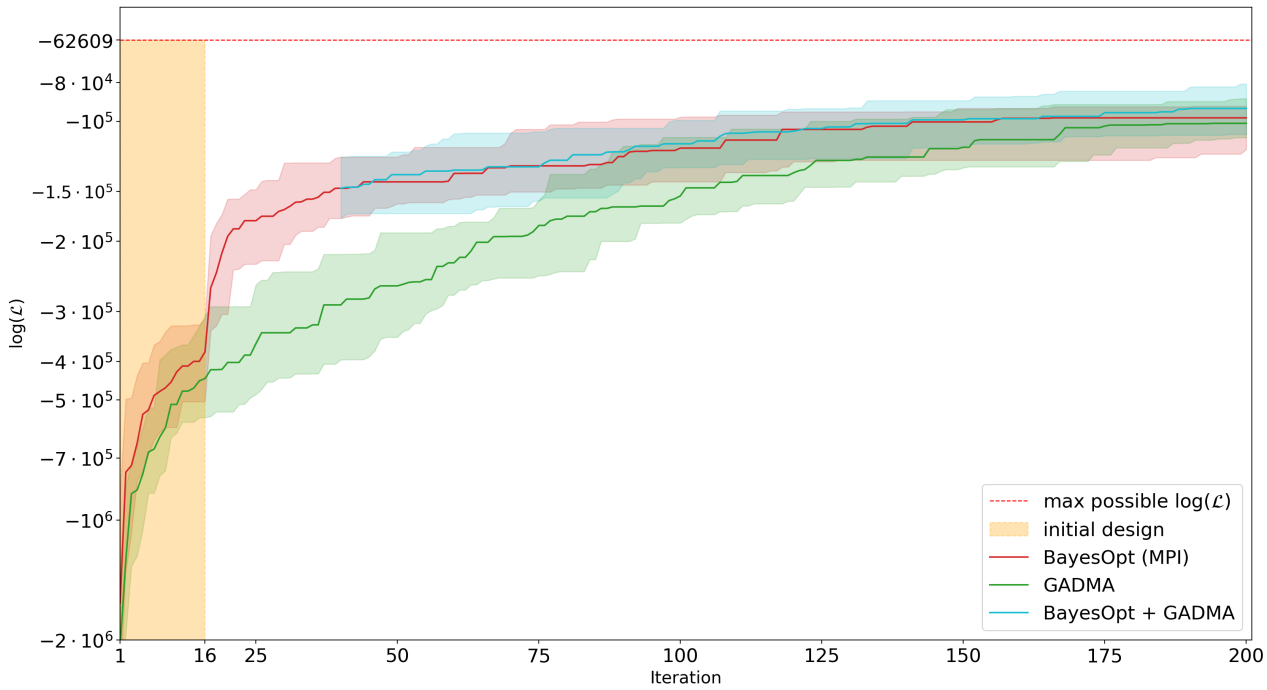


Рисунок 25 – График сходимости рассматриваемых методов и **комбинированного подхода** на симулированных данных для четырёх популяций. Можно отметить близкую с байесовской оптимизацией сходимость.

Такой подход имеет преимущество перед байесовской оптимизацией за счёт того, что время одной итерации исходной процедуры замедляется из-за увеличения нахождения обратной ковариационной матрицы в процессе обновления модели, на каждой итерации используются все предыдущие вычисления. На рисунке 26 показано общее время, затраченное на оптимизацию. Стоит отметить, что у комбинированного подхода достигается небольшое улучшение как в сходимости, так и во времени работы, по сравнению с байесовской опимизацией.

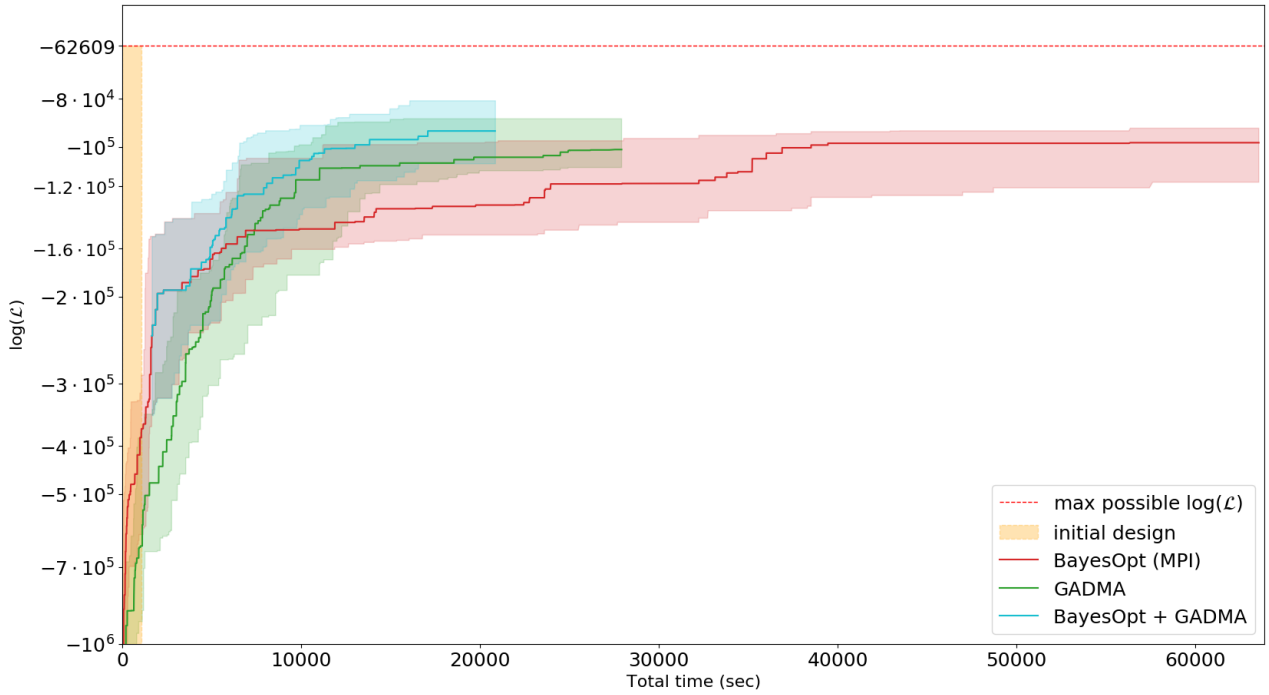


Рисунок 26 – График сходимости с комбинированным подходом. **По оси абсцисс — общее время оптимизации (в секундах).** По оси ординат — правдоподобие в логарифмической шкале. В среднем, комбинированный подход затратил наименьшее время для 200 итераций.

2.6. Локальные дооптимизации

Алгоритмы локального поиска обладают большей эффективностью, когда начальное приближение близко к оптимуму. Аккуратный подбор параметров может существенно улучшить результат, полученный во время глобальной оптимизации.

Для эксперимента с симулированными данными для пяти популяций (подробнее в подразделе 2.3.2) были дополнительно запущены локальные дооптимизации с помощью программного обеспечения *moments*, а именно, алгоритм локального поиска BFGS.

На рисунке 27 по результатам 32 запусков представлен график сходимости байесовской оптимизации и генетического алгоритма на 200 основных итерациях с 10 точками начального приближения и на 100 дополнительных итераций локальной оптимизации. Следует отметить, что одна итерация локального поиска на графиках сходимости соответствует одному вычислению целевой функции, то есть правдоподобия.

Из 32 запусков байесовской оптимизации и генетического алгоритма были выбраны запуски с наилучшими значениями правдоподобия после 200

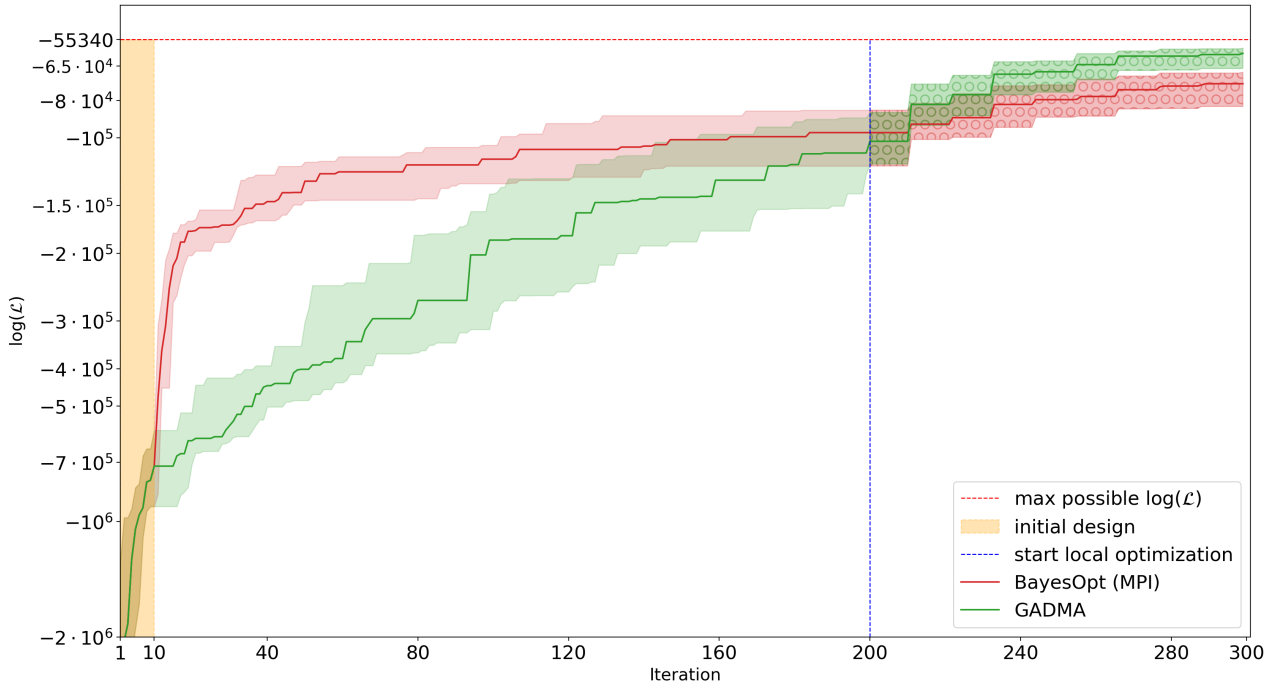
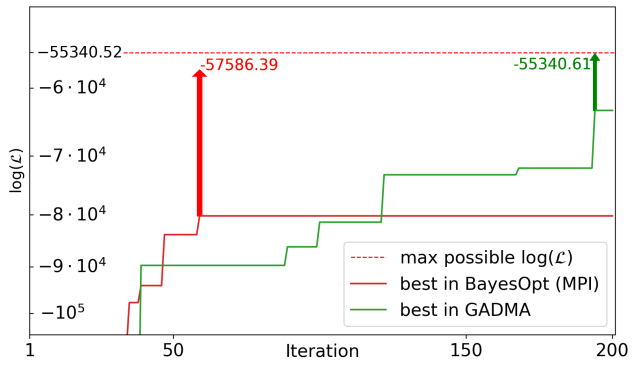


Рисунок 27 – График сходимости на симулированных данных для пяти популяций. После 200 итераций (синяя линия) глобальной оптимизации представлены 100 итераций локального поиска для *всех* запусков. В среднем, дооптимизации после генетического алгоритма показывают лучшую сходимость.

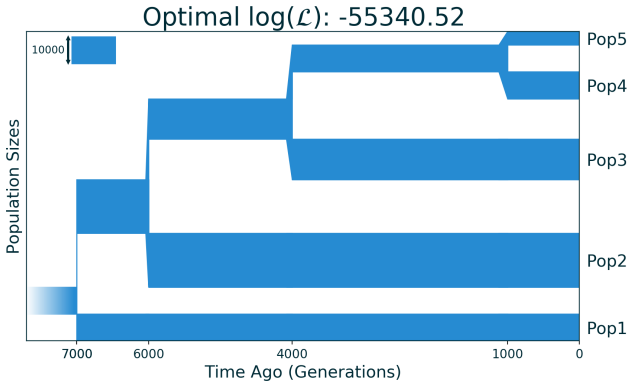
итераций. Найденные точки были переданы в *moments* для запуска локального поиска. На рисунке 28 приведены результаты локальной дооптимизации. Для значения, полученного после генетического алгоритма, поиск закончил работу после 980 итераций, лучшее значение (с точностью в 5 знаков) достигнуто на 265 итерации. В свою очередь, после байесовской оптимизации поиск сошелся за 1452 итераций, а лучшее значение (с точностью в 5 знаков) достигнуто на 628 итерации.

Выводы по главе 2

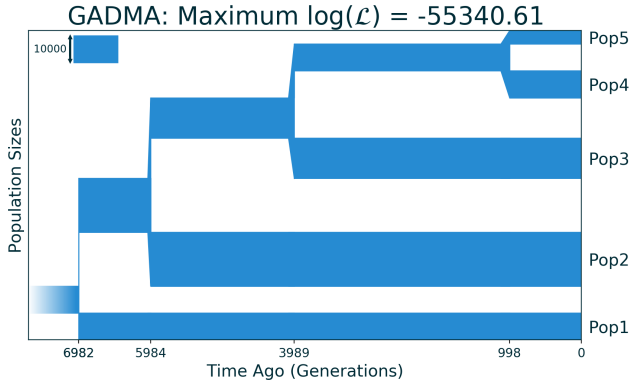
В данной главе перечислены существующие реализации байесовской оптимизации, приведены модификации выбранной реализации и описаны результаты экспериментальных исследований на симулированных и реальных данных для четырёх и пяти популяций. Были построены графики сходимости байесовской оптимизации с двумя функциями выбора: MPI и EI, генетического алгоритма и случайного поиска. По оси абсцисс отмечены итерации алгоритмов, по оси ординат отмечены текущие лучшие значения логарифма правдоподобия и сделаны выводы из этих графиков. По результатам



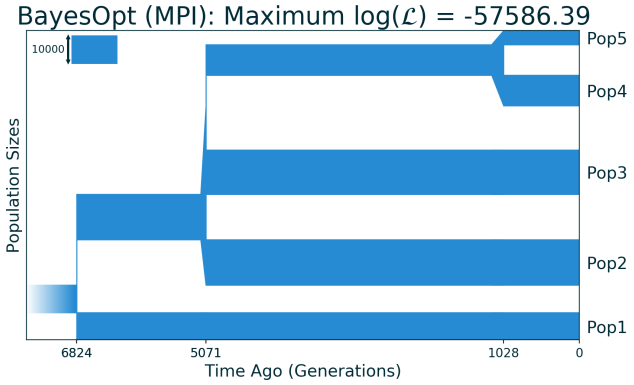
(а) График сходимости лучших запусков с локальной дооптимизацией.



(б) Демографическая модель с заданными параметрами.



(в) Демографическая модель, полученная с помощью локальной дооптимизации лучшего запуска GADMA.



(г) Демографическая модель, полученная с помощью локальной дооптимизации лучшего запуска BayesOpt (MPI).

Рисунок 28 – Результаты работы локальной дооптимизации для лучших запусков. Ширина стрелки на **(а)** отображает количество итераций локального поиска; демографические модели симулированных данных для пяти популяций с заданными параметрами **(б)** и найденными GADMA **(в)** и BayesOpt (MPI) **(г)**.

исследования на реальных данных превосходство не выявлено. Возможно, это связано с большим числом параметров (17 на реальных против 9 на симулированных) или с присутствием миграций среди параметров демографических историй. Для уточнения причин требуются дальнейшие исследования. При этом, на всех представленных моделях, в которых миграции отсутствуют, предлагаемый метод показал себя лучшим образом, как минимум на первых 50 итерациях оптимизации. Исходя из этого, был предложен и исследован алгоритм комбинирования байесовской оптимизации и генетического алгоритма. Также, была запущена локальная дооптимизация, в результате которой, было показано, что приведённые алгоритмы глобального поиска позволяют приблизиться к оптимуму.

ЗАКЛЮЧЕНИЕ

Настоящая работа посвящена выводу параметров демографической истории по генетическим данным для четырёх и пяти популяций. Был разработан и реализован метод, основанный на байесовской оптимизации для поиска параметров, который использует программное обеспечение *moments* для симулирования данных по заданной модели. Эффективность метода была проверена на симулированных и реальных данных для четырёх и пяти популяций.

В работе была использована существующая реализация байесовской оптимизации (GPyOpt) с дополнительными модификациями. Эффективность модификаций была проверена в ходе экспериментальных исследований. Также было проведено сравнение двух версий байесовской оптимизации с разными функциями выбора: максимальная вероятность улучшения (MPI) и ожидаемое улучшение (EI).

В ходе проведённых исследований байесовская оптимизация показала превосходство по сравнению с генетическим алгоритмом и случайным поиском на первых итерациях как на симулированных, так и на реальных данных (без учёта миграций). Однако, на последующих итерациях генетический алгоритм имеет лучшую сходимость к оптимуму. При сравнении двух байесовских оптимизаций с разными функциями выбора, версия с функцией MPI показала более высокую эффективность во всех экспериментальных исследованиях.

В данной работе был реализован простейший комбинированный подход: запуск байесовской оптимизации на первых итерациях и использование полученных точек в качестве начального приближения для генетического алгоритма. Этот подход был протестирован на симулированных данных для четырёх популяций и показал небольшое превосходство как в сходимости, так и во времени работы, однако требует дальнейшей модификации для повышения эффективности.

На реальных данных для демографической модели четырех популяций современных людей, включающей в себя миграции, байесовская оптимизация не продемонстрировала эффективности по сравнению с существующим решением, основанном на генетическом алгоритме. Была выдвинута и экспериментально проверена гипотеза о том, что проблема заключается в

наличии темпов миграций среди параметров. При рассмотрении моделей, в которых отсутствовали миграции, предлагаемый подход всегда оказывался эффективнее.

В качестве возможного дальнейшего развития работы стоит отметить модификацию существующего решения для моделей с миграциями, например, применение различных преобразований к области поиска (логарифмирование), подбор гиперпараметров байесовской оптимизации, разработка алгоритма поиска оптимальной итерации гибридизации для повышения эффективности представленного комбинированного подхода, а также дальнейшие совершенствования реализации для более быстрой работы.

По результатам работы было сделано выступление с докладом на IX Конгрессе молодых ученых в Университете ИТМО, в рамках секции «Технологии программирования, искусственный интеллект, биоинформатика».

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 A limited memory algorithm for bound constrained optimization / R. H. Byrd [et al.] // SIAM Journal on scientific computing. — 1995. — Vol. 16, no. 5. — P. 1190–1208.

2 *Beichman A. C.*, *Huerta-Sanchez E.*, *Lohmueller K. E.* Using genomic data to infer historic population dynamics of nonmodel organisms // Annual Review of Ecology, Evolution, and Systematics. — 2018.

3 *Broyden C. G.* The convergence of a class of double-rank minimization algorithms: 2. The new algorithm // IMA journal of applied mathematics. — 1970. — Vol. 6, no. 3. — P. 222–231.

4 *Cornford D.*, *Nabney I. T.*, *Williams C. K.* Modelling frontal discontinuities in wind fields // Journal of nonparametric statistics. — 2002. — Vol. 14, no. 1/2. — P. 43–58.

5 Demographic history and rare allele sharing among human populations / S. Gravel [et al.] // Proceedings of the National Academy of Sciences. — 2011. — Vol. 108, no. 29. — P. 11983–11988.

6 *Fisher R. A.* XVII.—The Distribution of Gene Ratios for Rare Mutations // Proceedings of the Royal Society of Edinburgh. — 1931. — Vol. 50. — P. 204–219.

7 *Fletcher R.* A new approach to variable metric algorithms // The computer journal. — 1970. — Vol. 13, no. 3. — P. 317–322.

8 *Frazier P. I.* A tutorial on bayesian optimization // arXiv preprint arXiv:1807.02811. — 2018.

9 GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data / E. Noskova [et al.] // GigaScience. — 2020. — Vol. 9, no. 3. — giaa005.

10 *Goldfarb D.* A family of variable-metric methods derived by variational means // Mathematics of computation. — 1970. — Vol. 24, no. 109. — P. 23–26.

11 GPyOpt: A Bayesian Optimization framework in python [Электронный ресурс]. — 2016.

12 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data / R. N. Gutenkunst [et al.] // PLoS genetics. — 2009. — Vol. 5, no. 10.

13 Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation / J. Jouganous [et al.] // Genetics. — 2017. — Vol. 206, no. 3. — P. 1549–1567. — ISSN 0016-6731. — DOI: 10.1534/genetics.117.200493. — eprint: http://www.genetics.org/content/206/3/1549.full.pdf. — URL: http://www.genetics.org/content/206/3/1549.

14 *Jones D. R.*, *Schonlau M.*, *Welch W. J.* Efficient global optimization of expensive black-box functions // Journal of Global optimization. — 1998. — Vol. 13, no. 4. — P. 455–492.

15 *Kushner H. J.* A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. — 1964.

16 *MacKay D. J.*, *Mac Kay D. J.* Information theory, inference and learning algorithms. — Cambridge university press, 2003.

17 *Mockus J.* On Bayesian Methods for Seeking the Extremum // Proceedings of the IFIP Technical Conference. — Berlin, Heidelberg : Springer-Verlag, 1974. — P. 400–404. — ISBN 3540071652.

18 *Mockus J.* The Bayesian approach to local optimization // Bayesian Approach to Global Optimization. — Springer, 1989. — P. 125–156.

19 *Mockus J.* Bayesian approach to global optimization: theory and applications. Vol. 37. — Springer Science & Business Media, 2012.

20 *Močkus J.* On Bayesian methods for seeking the extremum // Optimization techniques IFIP technical conference. — Springer. 1975. — P. 400–404.

21 *Powell M. J.* An efficient method for finding the minimum of a function of several variables without calculating derivatives // The computer journal. — 1964. — Vol. 7, no. 2. — P. 155–162.

22 *Sawyer S. A.*, *Hartl D. L.* Population genetics of polymorphism and divergence. // Genetics. — 1992. — Vol. 132, no. 4. — P. 1161–1176.

23 *Schraiber J. G.*, *Akey J. M.* Methods and models for unravelling human evolutionary history // Nature Reviews Genetics. — 2015. — Vol. 16, no. 12. — P. 727–740.

24 *Shanno D. F.* Conditioning of quasi-Newton methods for function minimization // Mathematics of computation. — 1970. — Vol. 24, no. 111. — P. 647–656.

25 *Snoek J.*, *Larochelle H.*, *Adams R. P.* Practical bayesian optimization of machine learning algorithms // Advances in neural information processing systems. — 2012. — P. 2951–2959.

26 *Williams C. K.*, *Rasmussen C. E.* Gaussian processes for machine learning. Vol. 2. — MIT press Cambridge, MA, 2006.