Эффективное оценивание параметров смеси распределений

Казанцев А.А., СПбГУ, Санкт-Петербург anton.a.kazancev@gmail.com,
Гориховский В.И., доцент кафедры системного программирования СПбГУ, к.ф.-м.н.,
Санкт-Петербург.

Аннотация

Оценка параметров смесей распределений встречается в большом спектре прикладных задач. Наиболее часто с решением данной проблемы сталкиваются при анализе смесей нормальных распределений. Однако бывают задачи, в которых необходимо разделять и оценивать смеси в более общем виде. Например, задержки пакетов в маршрутизаторе могут рассматриваться как распределения Вейбулла, если исключить единичные выбросы, связанные с работой сетевого оборудования. Таким образом, определение компонент смеси распределений Вейбулла позволяет строить качественные модели задержек передачи пакетов в сетях.

Среди существующих статей и библиотек присутствует большое количество узконаправленных алгоритмов для анализа распределений и их смесей. Однако в ходе анализа предметной области, не удалось найти универсальных инструментов для нахождения параметров произвольных смесей распределений.

Основная цель работы — разработка и реализация универсального метода для оценки параметров произвольных смесей непрерывных распределений.

Введение и постановка задачи

Во многих прикладных задачах, связанных с математической статистикой, возникает необходимость оценки параметров распределения случайной величины. Для решения такой задачи могут быть использованы известные для конкретного распределения оценки параметров, такие как математическое ожидание и дисперсия для нормального распределения. В том случае, если хороших известных оценок нет, можно воспользоваться математическими оптимизаторами, позволяющими находить локальные экстремумы функции от вектора параметров.

Помимо задачи об оценке параметров для одного распределения случайной величины, существует также задача оценки параметров для смеси

распределений случайной величины — комбинации нескольких распределений [1]. Такая задача возникает при наблюдении более сложных процессов, в которых участвуют случайные величины более чем из одного распределения. При оценке смеси решается ряд задач: узнать, сколько различных распределений находится в смеси, определить вид и параметры каждого распределения, узнать соотношение распределений в смеси друг к другу (априорная вероятность).

Наиболее часто с оценкой параметров смеси распределений сталкиваются при анализе смесей нормальных распределений [3]. Однако бывают задачи, в которых необходимо разделять и оценивать смеси в более общем виде. Например, смеси произвольных распределений, или смеси распределений из разных семейств.

Компания Huawei в октябре 2023 года опубликовала интернет-проект, посвящённый предсказанию задержек отправки пакетов через сеть, по наблюдаемым отправителем задержкам пакетов [7]. Задержки пакетов в маршрутизаторе могут рассматриваться с помощью распределения Вейбулла [2, 5], если исключить единичные выбросы, связанные с работой сетевого оборудования. Таким образом, определение компонент смеси распределений Вейбулла позволяет строить качественные модели задержек передачи пакетов в сетях.

Среди существующих статей и библиотек присутствует большое количество узконаправленных алгоритмов для анализа распределений и их смесей. Однако в ходе анализа предметной области не удалось найти универсальных инструментов для нахождения параметров произвольных смесей распределений.

Цель работы — создание библиотеки для оценки параметров произвольных смесей непрерывных распределений случайных величин, что позволит решать большой спектр прикладных задач.

Для осуществления этой цели были сформулированы следующие **задачи**.

- Создать алгоритм для оценки параметров смесей распределений на основе EM-алгоритма с поддержкой использования различных методов математической оптимизации.
- Спроектировать архитектуру библиотеки, реализующую данный алгоритм.
- Выполнить и опубликовать реализацию спроектированной библиотеки.
- Выполнить эксперименты для анализа работы алгоритма при оценке параметров смесей распределений с различными параметрами алгоритма.

Обзор

Смесь распределений

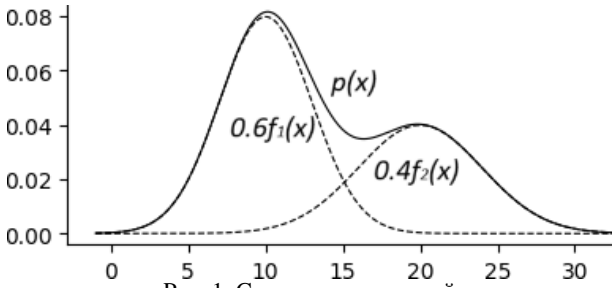

Рис. 1: Смесь распределений

Плотность смеси распределений задается следующим образом:

$$p(x|\Omega, F, \Theta) = \sum_{j=1}^{k} \omega_j f_j(x|\theta_j),$$

где k — это количество компонент смеси, $\omega_j \geq 0, \sum_{j=1}^{k} \omega_j = 1$ — априорная вероятность j компоненты смеси, $f_j(x, \theta_j), \theta_j \in \Theta_{f_j}$ — функция правдоподобия j компоненты смеси; $\Omega = [\omega_1, .., \omega_k]$, $F = [f_1, .., f_k]$, $\Theta = [\theta_1, .., \theta_k]$ — параметры смеси.

На рис. 1 показана смесь, состоящая из двух распределений Гаусса, с априорными вероятностями 0.4 и 0.6.

Математическая постановка задачи

По заданной выборке X^m случайных независимых наблюдений из смеси распределений $p(x|\Omega, F, \Theta)$ оценить параметры Ω, F, Θ и количество компонент смеси k.

В данной работе рассматривается упрощенная версия задачи, в которой параметры F, k смеси известны. Полученный алгоритм в дальнейшем может использоваться внутри алгоритма, решающего полную задачу об оценке параметров смеси распределений.

EM-алгоритм

EM-алгоритм [6, 4] является итеративным алгоритмом для нахождения локального максимума логарифма функции максимального правдоподобия статистической модели. Он применяется в случае, когда модель зависит от некоторых скрытых параметров.

Алгоритм состоит из двух шагов:

- expectation (E-шаг): вычисление оценки функции максимального правдоподобия модели, при этом скрытые параметры рассматриваются как наблюдаемые.
- maximization (M-шаг): подбор скрытых параметров для того что бы максимизировать оценку, найденную на E-шаге.

Для начала работы EM-алгоритма необходимо стартовое предположение о скрытых параметрах, от которого напрямую зависит работа алгоритма.

```
function EM_step(X, Ω, F, Θ)
    \\E-step
    for j ∈ 1..k do
        for i ∈ 1..m do
            H_ij ← ω_j f_j(X_i|θ_j) / Σ_{s=1}^{k} ω_s f_s(X_i|θ_s)
    \\M-step
    for j ∈ 1..k do
        ω'_j ← (1/m) Σ_{i=1}^{m} H_ij
        θ'_j ← argmax_{γ∈Θ_{f_j}} (Σ_{i=1}^{m} H_ij ln F_j(X_i, γ))
    return Ω', Θ'
```

Алгоритм для оценки параметров смесей распределений

В основе используемого в библиотеке алгоритма лежит описанная ранее версия EM-алгоритма в общем виде с добавлением функции останова и проверки рассматриваемых распределений на их корректность.

```
function EM_solve(X, Ω, F, Θ)
    Ω′, F′, Θ′ ← Ω, F, Θ
    repeat
        Ω, F, Θ ← Ω′, F′, Θ′
        Ω′, Θ′ ← EM_step(X, Ω, F, Θ)
        F′ ← [F_j ∈ F : valid(Ω′_j, F_j, Θ′_j)]
    until breakpoint(Ω, F, Θ, Ω′, F′, Θ′)
    return F′, Ω′, Θ′
```

- $valid(\omega, f, \theta)$ — проверка, является ли распределение с соответствующими параметрами корректным
- $breakpoint(\Omega, F, \Theta, \Omega', F', \Theta')$ — функция точки останова

Критерии останова

Так как EM-алгоритм итеративный, можно выделить большое количество различных критериев останова. Примеры возможных критериев останова:

- время исполнения:
  - максимальное количества шагов;
  - максимальное времени исполнения;
- относительная точность — изменение параметров;
- абсолютная точность:
  - логарифм максимального правдоподобия;
  - разделение выборки на рабочую и тестовую.

«Плохие» распределения

В связи с использованием оптимизатора в алгоритме, параметры одного или нескольких распределений в выборке могут стать экстремально большими. Данная ситуация может возникнуть по ряду причин, среди которых слишком большой параметр k и плохое стартовое предположение. Так как распределение в выборке имеет параметр ω (априорная вероятность), с точки зрения алгоритма ошибок нет, так как при экстремально больших ложных параметрах ω стремится к нулю. Наличие в выборке «плохого» распределения может привести к значительному снижению скорости работы алгоритма, или к его остановке.

После очередного М-шага алгоритма следует проводить проверку корректности полученных распределений. В случае нахождения «плохого распределения» следует отбросить его и не использовать в дальнейшей работе алгоритма, сообщив что стартовое предположение, приведшее к возникновению этого распределения, ложно.

Таким образом, к возможным критериям «плохого» распределения можно отнести малую априорную вероятность или экстремально большие значения параметров.

Проектирование и реализация библиотеки

Была спроектирована библиотека, позволяющая решать задачу оценки параметров произвольных смесей непрерывных или дискретных распределений [10].

Библиотека была спроектирована с требованиями к универсальности и простоте модификаций.

Архитектура библиотеки

Основной класс библиотеки позволяет применять описанный выше EM-алгоритм к произвольным распределениям, описанным с помощью содержащихся в библиотеке классов обёрток. При его инициализации необходимо в качестве параметров передать оптимизатор, класс реализующий условие останова и класс определяющий, является ли распределение из смеси корректным.

В соответствии с требованиями EM-алгоритма и используемого в нём оптимизатора, сформулируем следующие требования к описанию модели распределения:

- плотность;
- логарифм плотности;
- логарифм частных производных плотности по каждому параметру (опционально для работы с оптимизаторами второго порядка).

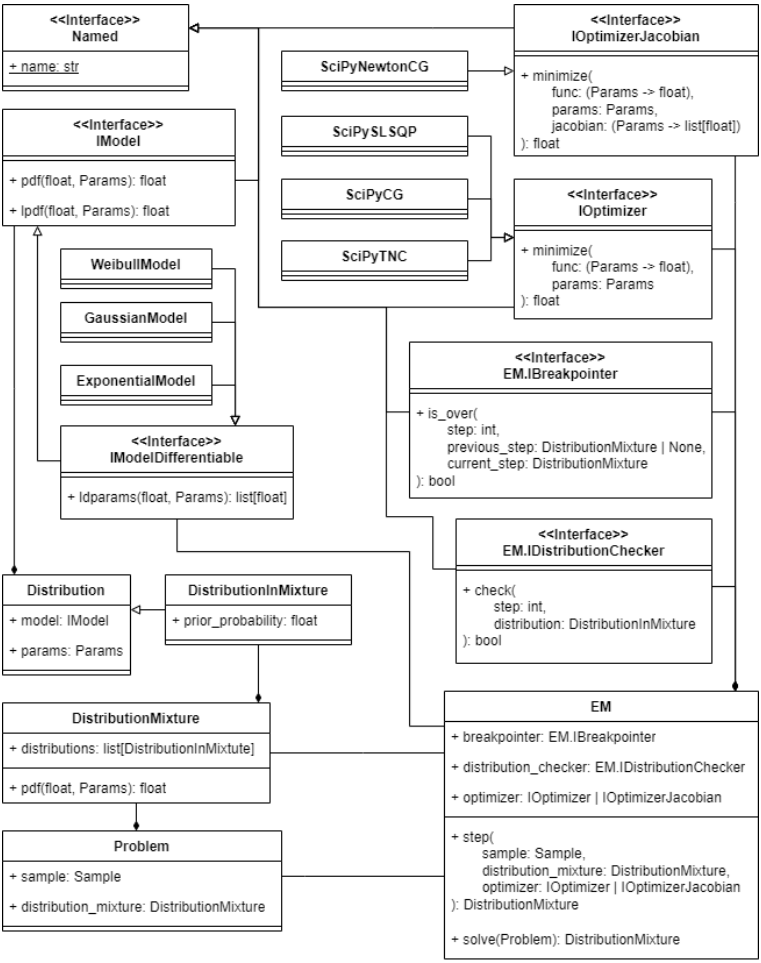

Рис. 2: UML диаграмма классов библиотеки

Эксперименты

Дизайн эксперимента

Характеристики оборудования

- Ноутбук: HP Probook 440 G5, подключенный к сети питания
- Процессор: Intel Core i5-8250U
- Оперативная память: DDR4 16Гб
- Версия ОС: Ubuntu

Рассматриваемые распределения

Так как в ходе работы используемого алгоритма подразумевается использование функций оптимизации (максимизации), плотность рассматриваемых распределений должна быть всюду дифференцируемой по её параметрам. Для того что бы обеспечить такое свойство распределению, параметры которого не могут быть отрицательными, применяется замена параметра на его экспоненту.

Распределение Вейбулла Наиболее интересное для изучения в данной работе распределение, так как его параметры k и λ трудно выразить через выборку.

Плотность распределения Вейбулла:

$$f(x|k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, k > 0, \lambda > 0, x \geq 0$$

Распределение Гаусса

$$f(x|m,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(n-x)^2}{2\sigma^2}}, m \in \mathbb{R}, \sigma >= 0, x \in \mathbb{R}$$

Экспоненциальное распределение

$$f(x|\lambda) = \lambda e^{(-\lambda x)}, \lambda > 0, x >= 0$$

Эксперимент со смесями одного семейства распределений

Условия эксперимента

- Смеси распределений Вейбулла, Гаусса, экспоненциальные
- Оптимизаторы из пакета SciPy [9]: CG, NewtonCG, SLSQP, TNC
- Количество компонент от 1 до 3
- Генерируются 32 выборки из 2048 случайных независимых наблюдений для каждого количества компонент и типов распределений, откуда в дальнейшем случайно выбираются выборки меньших размеров.
- Размеры компонент выборки: 50, 100, 200, 500, 1000, одинаковые для одного теста
- По 2 выборки меньшего размера из каждой большой выборки
- По 8 независимых тестов со случайными стартовыми параметрами
- Максимальное количество шагов алгоритма 16
- Минимальная априорная вероятность распределения в выборке 0.1%

Обоснование Данный эксперимент направлен на то, что бы понять, как работает алгоритм и какие ситуации могут возникнуть во время работы. А так же сравнить работу наиболее интересных оптимизаторов.

Были взяты одинаковые размеры компонент выборки для упрощения анализа результатов.

На практике хочется получить не очень точный, но быстрый алгоритм, поэтому размеры компонент начинаются с 50. Выборки с компонентами размера 1000 нужны для понимания того, как сильно увеличится время работы алгоритма и насколько точнее он будет работать.

Метрики

- скорость схождения;
- усреднённый логарифм максимального правдоподобия;

$$NLL = \frac{1}{m} \sum_{x \in X^m} \ln p(x|\Omega, F, \Theta)$$

- шанс ошибки алгоритма: алгоритм не смог найти ни одного распределения или остановил работу из-за экстремально больших параметров смеси.

Результаты

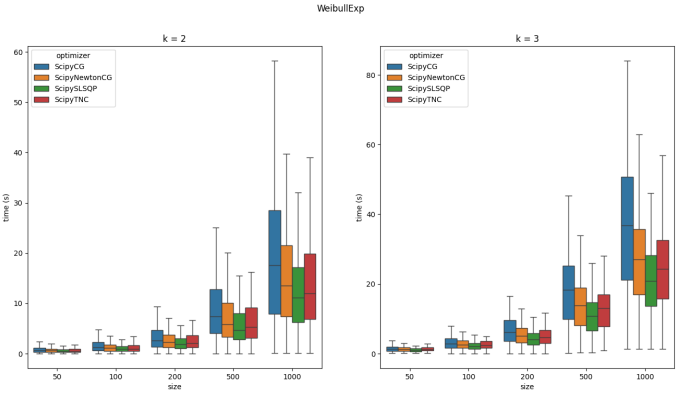

Рис. 3: Время работы алгоритма при оценке смесей Вейбулла

На рис. 3 изображён график зависимости времени исполнения алгоритма от размера выборки, количества распределений в смеси и оптимизатора для смесей Вейбулла. Чем больше размер выборки, тем больше времени необходимо алгоритму, что бы сойтись. Так же видно, что оптимизаторы работают с разной скоростью. Самым медленным оказался оптимизатор CG (метод сопряженных градиентов). Самым же быстрым — SLSQP (последовательное квадратичное программирование).

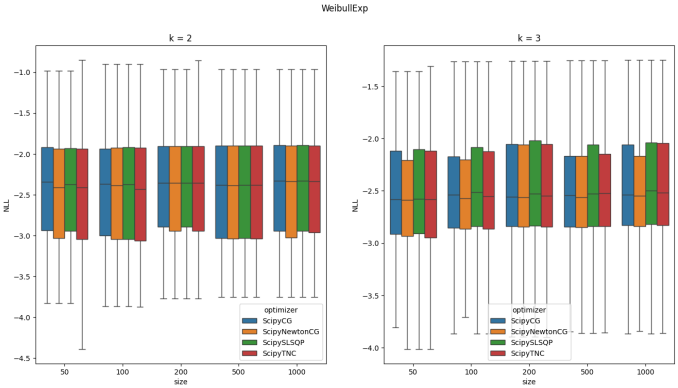

Рис. 4: NLL алгоритма при оценке смесей Вейбулла

На рис. 4 видно, что различные оптимизаторы практически не отличаются в точности для задачи оценки параметров смесей распределений Вейбулла.

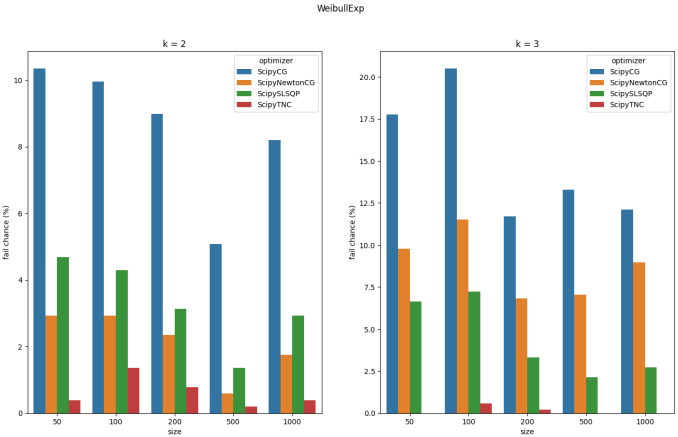

Рис. 5: Шанс ошибки алгоритма при оценке смесей Вейбулла

На рис. 5 видно, что шанс ошибки алгоритма зависит от оптимизатора. Для смесей Вейбулла самым стойким оказался оптимизатор TNC (Truncated Newton method).

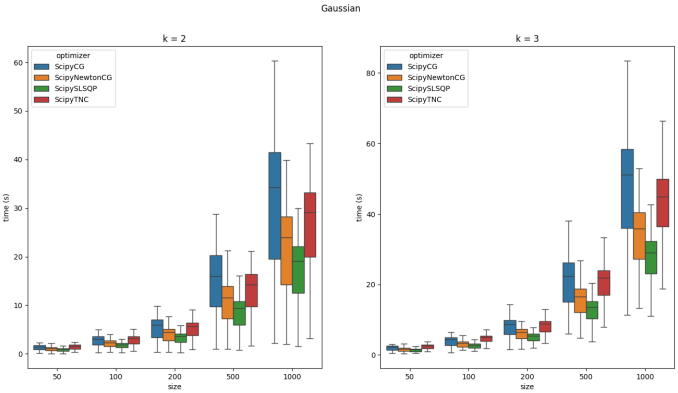

Рис. 6: Время работы алгоритма при оценке смесей Гаусса

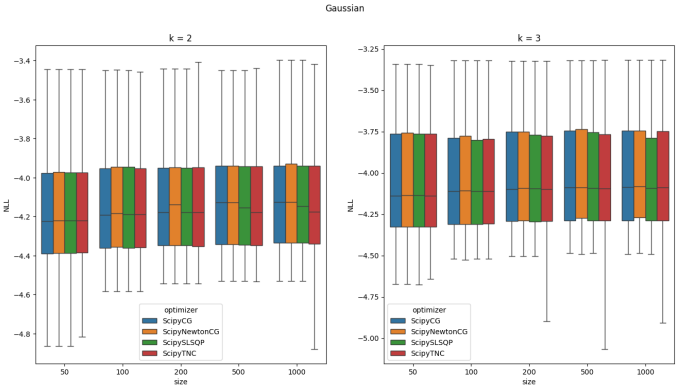

Рис. 7: NLL алгоритма при оценке смесей Гаусса

Для смесей распределений Гаусса время работы и точность (Рис. 6, Рис. 7) зависят от оптимизатора практически так же как и для смесей Вейбулла.

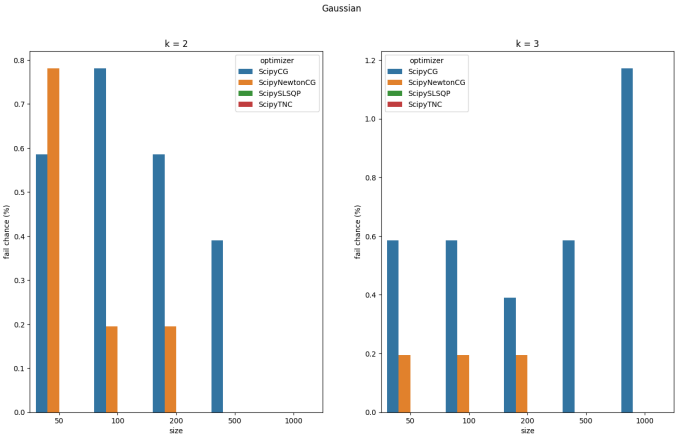

Рис. 8: Шанс ошибки алгоритма при оценке смесей Гаусса

Шанс ошибки алгоритма при оценке параметров смесей Гаусса (Рис. 8), в отличие от Вейбулла, минимален не только у оптимизатора TNC, но и у SLSQP. Таким образом для смесей Гаусса оптимизатор SLSQP является лучшим по времени работы и по шансу ошибки.

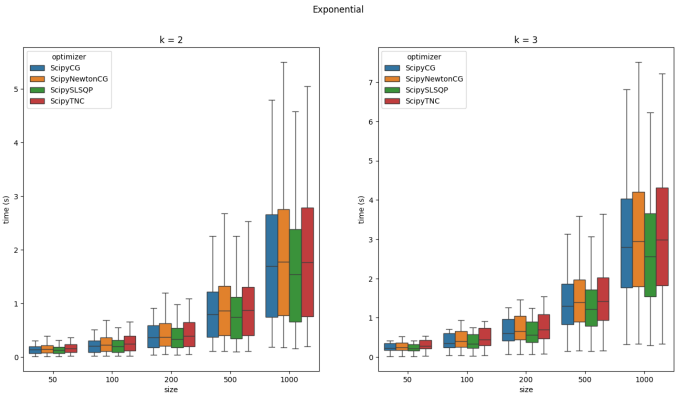

Рис. 9: Время работы алгоритма при оценке смесей экспоненциальных распределений

Как видно из рис. 9, при оценке параметров смеси экспоненциальных распределений разница между рассмотренными оптимизаторами не столь велика, как для смесей Вейбулла и Гаусса.

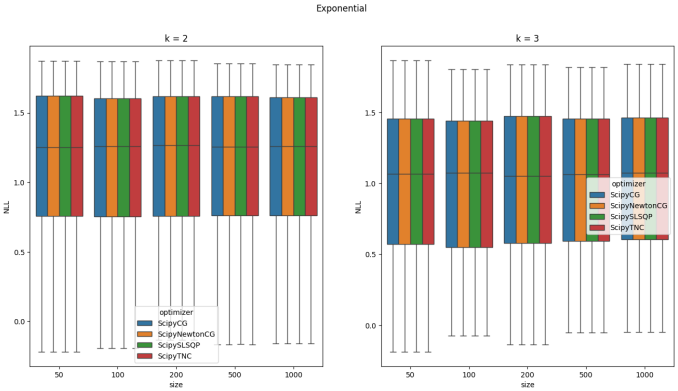

Рис. 10: NLL алгоритма при оценке смесей экспоненциальных распределений

Так же как и для смесей Вейбулла и Гаусса, точность алгоритма при оценке параметров смеси экспоненциальных распределений (Рис. 10) не сильно зависит от оптимизатора.

При оценке параметров смесей экспоненциальных распределений ни один из оптимизаторов не ошибся.

По графикам можно сделать следующие наблюдения:

- Оценка параметров смеси экспонент быстрее и точнее, чем Гаусса и Вейбулла, что вероятно связано с тем, что экспоненциальное распределение зависит от одного параметра.
- Различные оптимизаторы работают с разной скоростью. Самый быстрый оптимизатор среди рассмотренных: SLSQP.
- У разных оптимизаторов разная устойчивость к ошибкам. Самыми устойчивыми оказались оптимизаторы SLSQP и TNC.
- Точность оценки параметров практически не зависит от оптимизатора.

Заключение

- Был создан алгоритм для решения поставленной задачи (на основе EM-алгоритма в общем виде). Работа полученного алгоритма зависит от следующих параметров:
  - метод математической оптимизации;
  - условие остановки;
  - условие корректности рассматриваемых распределений в смеси.
- Спроектирована архитектура библиотеки, реализующей предложенный алгоритм, которая отвечает требованиям по универсальности и расширяемости, позволяет работать со смесями распределений из различных семейств и дополнять библиотеку произвольными моделями распределений и различными реализациями ключевых параметров алгоритма.
- Библиотека реализована на языке Python.
- Выполнено экспериментальное исследование, получены следующие результаты/выводы:
  - доказана корректность разработанного алгоритма;
  - выявлено, что с увеличением количества распределений в смеси растёт время работы и падает точность алгоритма, а с увеличением размера выборки растёт время работы и точность алгоритма;
  - показано, что методы оптимизации обладают разными достоинствами/недостатками и для эффективного решения задачи они могут быть использованы в комбинации друг с другом.

Список литературы

[1] Bruce G. Lindsay: Mixture Models: Theory, Geometry and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics. 1995. Hayward, CA, USA. https://www.jstor.org/stable/4153184

[2] Elmahdy Emad E., Aboutahoun Abdallah W.: A new approach for parameter estimation of finite Weibull mixture distributions for reliability modeling. Applied Mathematical Modelling. 2013. с. 1800–1810. https://www.sciencedirect.com/science/article/pii/S0307904X12002545

[3] Wardatul Jannah, Dewi R.S. Saputro: Parameter estimation of Gaussian mixture models (GMM) with expectation maximization (EM) algorithm. Conference Proceedings 2566, 040002. 2022. https://web.archive.org/web/20221201095213id_/https://aip.scitation.org/doi/pdf/10.1063/5.0117119

[4] Jason Brownlee: A Gentle Introduction to Expectation-Maximization (EM Algorithm). 2020. https://machinelearningmastery.com/expectation-maximization-em-algorithm

[5] И.М. Макуха: Оценка параметров смеси распределений Вейбулла, выпускная квалификационная работа Санкт-Петербургского государственного университета. 2023. https://se.math.spbu.ru/thesis/texts/Makuha_Il'ja_Mihajlovich_Bachelor_Thesis_2023_text.pdf

[6] J.W. Davenport, J.C. Bezdek, R.J. Hataway: Parameter estimation for finite mixture distributions. Mathematies and Computer Science Department. 1988. Southern College, Georgia. https://core.ac.uk/download/pdf/82096497.pdf

[7] Dmitriy Moskvitin, Evgeny Onegin, Rachel Huang, Hanlin Luo, Qichang Chen: Forward Erasure Correction for Short-Message Delay-Sensitive QUIC Connections. 2023. https://datatracker.ietf.org/doc/draft-dmoskvitin-quic-short-message-fec

[8] Matthew Reid: Reliability — a Python library for reliability engineering. 2022. https://doi.org/10.5281/ZENODO.3938000

[9] SciPy — an open-source software for mathematics, science, and engineering. https://docs.scipy.org/doc/scipy/

[10] Реализация библиотеки. https://github.com/toxakaz/EM-algo