Міністерство освіти і науки України

Київський національний університет імені Тараса Шевченка

Кафедра алгебри і комп'ютерної математики

Курсовий проект

На тему: «EM-алгоритм для аналізу регресійної суміші гамма-розподілів»

Виконав:

Тунік Вадим Романович,

студент 3 курсу, Комп'ютерна математика.

Науковий керівник:

Мірошниченко Віталій Олегович,

Асистент кафедри Теорії ймовірностей,

статистики та актуарної математики

Київ-2024

Зміст

1 Вступ.

В цій роботі розглянута параметрична модель регресійної суміші, в якій функції регресії - лінійні. В моделі набір даних розглядяється як суміш об'єктів, кожен з яких належить до певної компоненти з певною ймовірністю (концентрацією). Кожен набір даних зі спільною компонентою відповідно належить до певного ймовірнісного розподілу. В даній роботі розглянутий випадок суміші гамма-розподілів. Тобто модель, за припущенням, використовує лінійну залежність з відповідними параметрами до розподілу компоненти регресорів.

Подібна модель може бути застосована до статистичних моделювань випадкових процесів в економіці, біології, медицині, соціології.

Також дослідженням цих моделей займалася низка вчених, наприклад: Bilmes, Jeff. (2000) [1], Faria, S. & Soromenho, Gilda. (2010) [2], Miroshnichenko V., Maiboroda R. (2018) [3], Quandt, R. E., & Ramsey, J. B. (1978) [4], Liubashenko D., Maiboroda R. (2015) [5].

2 Модель даних.

Маємо набір даних, що являють собою суміш гамма-розподілів з невідомими концентраціями. Розподіл суміші можна визначити, записавши її щільність:

$$p(x) = \sum_{l=1}^{M} w_l p_l(x)$$

Тобто:

$$X_i = \left(1, X_i^1, \ldots, X_i^d\right), i = \overline{1, N};$$
$$X_i \sim p_l(x) = \frac{1}{\Gamma(k_l)\theta_l^k} x^{k_l - 1} e^{-x/\theta_l};$$
$$w_l = P\left(\mathrm{lnd}\,(x) = l\right).$$

Також маємо залежну від X_i характеристику $Y_i, i = \overline{1, N}$. Тобто:

$$Y_i = \langle X_i, b_l \rangle + \varepsilon_i,$$

де $\langle \cdot, \cdot \rangle$ - скалярний добуток, а ε_i - випадкова похибка вимірювання, причому $\varepsilon_i \sim p_{\varepsilon,l}(x) = \frac{1}{\sqrt{2\pi\sigma^2_{\varepsilon,l}}} e^{-\frac{x^2}{2\sigma^2_{\varepsilon,l}}}$.

$$y_i = \begin{cases} \langle x_i, b_1 \rangle + \varepsilon_{i1} & \text{з ймовірністю } p_1, \\ \langle x_i, b_2 \rangle + \varepsilon_{i2} & \text{з ймовірністю } p_2, \\ & \vdots \\ \langle x_i, b_M \rangle + \varepsilon_{iM} & \text{з ймовірністю } p_M \end{cases},$$

де p_l - апріорні ймовірності змішування.

Виведемо модель, що одночасно буде вирішувати задачу кластеризації даних та задачу регресії. Для оцінки невідомих параметрів використаємо EM-алгоритм.

3 Теоретичні відомості.

3.1 Лінійна регресія.

Метою регресії є визначення залежності кількісного впливу одних змінних (регресорів) на числові значення інших змінних (відгуків). Лінійна регресія розглядає лінійну залежність відгуків від регресорів.

$$y_i = \langle x_i, b \rangle + \varepsilon_i, \quad i = \overline{1, N},$$

де y_i - і-те спостереження залежної змінної (відгук), x_i - і-те спостереження незалежної змінної (регресор), b - невідомий вектор параметрів, що підлягає оцінюванню, ε - вектор випадкових похибок вимірювання, N - кількість спостережень (розмір вибірки). Природньо припускати, що випадковий вектор похибок ε містить незалежні, однаково розподілені, з нульовим середнім (не систематичні) та дисперсією σ^2_ε похибки.

3.2 Оцінка максимальної вірогідності.

Оцінка максимальної вірогідності - це метод оцінювання невідомого параметра $\theta \in \Theta$ за вибіркою $X = (X_1, \cdots, X_n)$, який повертає статистику, що максимізує вибіркову функцію вірогідності $L(X, \theta)$:

$$\hat{\theta} = \hat{\theta}(X) = \arg\max_{\theta \in \Theta} L(X, \theta)$$

Для незалежних та однаково розподілених випадкових величин, $L(X, \theta)$ буде добутком одновимірних функцій щільності:

$$L(X, \theta) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

Іноді простішим та еквівалетним способ обчислення ОМВ є використання логарифмічної функції вірогідності:

$$\hat{\theta} = \hat{\theta}(X) = \arg\max_{\theta \in \Theta} \ln L(X, \theta)$$

Еквівалентність випливає з властивості монотонності логарифму: максимум будь-якої функції $f(x)$ є максимумом функції $\ln f(x)$, і навпаки.

Для знаходження безпосередньо ОМВ використовують рівняння максимальної вірогідності:

$$\frac{\partial}{\partial \theta} \ln L(X, \theta) \mid_{\theta = \hat{\theta}} = 0,$$

за умови того, що функція вірогідності диференційована, а параметр є векторним. Для деяких моделей це рівняння можна розв'язати в явному вигляді, але в загальному випадку аналітичний розв'язок задачі максимізації не є відомим, і ОМВ може бути знайдене лише шляхом чисельної оптимізації, наприклад, методом Ньютона-Рафсона.

Оцінку знаходять таким чином, щоб максимізувати ймовірність того, що процес, описаний моделлю, призвів до отримання даних, які дійсно

спостерігалися. Словом, використовується принцип максимальної вірогнідності - "те, що спостерігається, є найбільш імовірним серед усіх можливих альтернатив". Принцип інтуїтивно зрозумілий та дає стабільні оцінки, тож широко використовується для статистичного моделювання.

3.3 EM-алгоритм.

EM-алгоритм - це ітеративний алгоритм без вчителя, назва якого походить від Expectation – Maximization (максимізація очікування). Алгоритм використовується для знаходження оцінок максимальної вірогідності параметрів ймовірнісних моделей, у випадку, коли модель залежить від деяких латентних змінних, що дозволяє наближати складні ймовірнісні розподіли із суміші розподілів. Тому алгоритм часто розглядається як метод кластеризації випадкової величини, що має ймовірнісний розподіл. Однак слід зазначити, що алгоритм має ширше застосування.

Кожна ітерація алгоритму складається з двох кроків: кроку очікування (E-крок), який створює функцію для очікування логарифмічної вірогідності, оціненої з використанням поточної оцінки параметрів, і кроком максимізації (M-крок), який обчислює параметри, що максимізують очікувану логарифмічну вірогідність, знайдену на E-кроці.

Припустимо, що набір даних X згенеровано деяким розподілом. Надалі називатимемо X неповними даними. Припустимо, що існує повний набір даних $Z = (X, Y)$, а також припускаємо спільну функцію щільності розподілу:

$$p(z \mid \Theta) = p(x, y \mid \Theta) = p(y \mid x, \Theta)p(x \mid \Theta)$$

За допомогою цієї нової функції щільності ми можемо визначити нову функцію вірогідності повних даних:

$$L(\Theta \mid Z) = L(\Theta \mid X, Y) = p(X, Y \mid \Theta)$$

Е-крок:

Знайдемо очікуване значення логарифмічної функції вірогідності повних даних $\log p(X, Y \mid \Theta)$ для невідомих даних Y, маючи спостережувані дані X та поточні оцінки параметрів:

$$Q\left(\Theta, \Theta^{i-1}\right) = E\left[\log p(X, Y \mid \Theta) \mid X, \Theta^{i-1}\right],$$

де Θ^{i-1} - поточні оцінки параметрів, які ми використовуємо для оцінки очікування, а Θ - нові параметри, які ми оптимізуємо для максимізації вірогідності Q.

М-крок:

Максимізуємо очікування, яке ми обчислили на Е-кроці:

$$\Theta^i = \arg\max_{\Theta} Q\left(\Theta, \Theta^{i-1}\right)$$

Ці два кроки повторюються до збіжності алгоритму. Кожна ітерація гарантовано збільшує логарифмічну функцію вірогідності та алгоритм гарантовано збігається до локального максимуму вірогідності.

4 Знаходження параметрів суміші за допомогою EM - алгоритму.

За припущенням, маємо суміш з гамма-розподілів. Стандартний вигляд щільності гамма-розподілу:

$$p(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$$

Як ми побачимо далі, для отримання оцінки $\hat{k}$ нам не вдасться явно розв'язати рівняння. Тому для знаходження $\hat{k}$ необхідно застосувати приблизний метод Ньютона — Рафсона. Проте, у зв'язку з цим може виникнути проблема, що метод повертає недодатнє значення оцінки параметру, що суперечить простору параметрів гамма-розподілу. Отже, задля уникнення появи від'ємних значень параметру k, експоненціюємо цей параметр, тобто: e^k. Тепер щільність нашого модифікованого гамма-розподілу виглядає наступним чином:

$$p(x) = \frac{1}{\Gamma(e^k)\theta^{e^k}} x^{e^k-1} e^{-x/\theta}$$

Побудуємо оцінки за допомогою EM алгоритму.

$$p(x|\Theta) = \sum_{l=1}^{M} w_l p_l\left(x|\Theta_l\right),$$

де $\Theta = (w_1, \cdots, w_M, \Theta_1, \cdots, \Theta_M)$; $w_l = P(Ind(x) = l)$ - коефіцієнт (вага) змішування в суміші, причому $\sum_{l=1}^{M} w_l = 1$; Θ_l - параметри щільності компоненти; $p_l\left(x|\Theta_l\right) = \frac{1}{\Gamma(e^{k_l})\theta_l^{e^{k_l}}} x^{e^{k_l}-1} e^{-x/\theta_l}$; M - кількість компонент суміші.

E-крок:

Спочатку виведемо вираз для розподілу неспостережуваних даних. Для цього вгадаємо параметри густини суміші, тобто припустимо, що

$$\Theta^g = (w_1^g, \cdots, w_M^g, \Theta_1^g, \cdots, \Theta_M^g)$$

є відповідними параметрами для ймовірності $L\left(\Theta^g \mid X, Y\right)$ [1] с.3:

$$p\left(y_i \mid x_i, \Theta^g\right) = \frac{w_{y_i}^g p_{y_i}\left(x_i \mid \Theta_{y_i}^g\right)}{\sum_{l=1}^{M} w_l^g p_l\left(x_i \mid \Theta_l^g\right)}$$

та

$$p\left(y \mid X, \Theta^g\right) = \prod_{i=1}^{N} p\left(y_i \mid x_i, \Theta^g\right)$$

Тепер можемо вивести зручну для нас форму функції очікування логарифмічної вірогідності [1] с.4:

$$Q\left(\Theta,\Theta^{g}\right)=\sum_{l=1}^{M}\sum_{i=1}^{N}\log\left(w_{l}\right)p\left(l\mid x_{i},\Theta^{g}\right)+\sum_{l=1}^{M}\sum_{i=1}^{N}\log\left(p_{l}\left(x_{i}\mid\Theta_{l}\right)\right)p\left(l\mid x_{i},\Theta^{g}\right)$$

де w_l - концентранція змішування l-тої компоненти суміші.

М-крок:

Максимізуємо $Q\left(\Theta,\Theta^{g}\right)$, вивівши оцінки нових параметрів в термінах старих параметрів і позначимо нову концентрацію змішування l-тої компоненти як w_l^{new} , а нові параметри щільності компоненти з гамма-розподілом як k_l^{new} і θ_l^{new} відповідно:

4.1 Оцінка для концентрації l-компоненти w_l^{new}.

Використаємо відомий результат [1] с.5:

$$w_{l}^{\text{new}}=\frac{1}{N}\sum_{i=1}^{N}p\left(l|x_{i},\theta^{g}\right)$$

4.2 Оцінка для параметру масштабу θ_l^{new} .

$$\begin{aligned}\frac{\partial Q}{\partial\theta_{l}}&=\left(\sum_{l=1}^{M}\sum_{i=1}^{N}\log\left(\frac{x_{i}^{\left(e^{k_{l}}-1\right)}e^{\left(-\frac{x}{\theta_{l}}\right)}}{\Gamma\left(e^{k_{l}}\right)\theta_{l}^{\left(e^{k_{l}}\right)}}\right)p\left(l\mid x_{i},\Theta^{g}\right)\right)'_{\theta_{l}}=\\&=\sum_{i=1}^{N}p\left(l\mid x_{i},\Theta^{g}\right)\left(\log\left(\frac{x_{i}^{\left(e^{k_{l}}-1\right)}e^{\left(-\frac{x}{\theta_{l}}\right)}}{\Gamma\left(e^{k_{l}}\right)\theta_{l}^{\left(e^{k_{l}}\right)}}\right)\right)'_{\theta_{l}}=\\&=\sum_{i=1}^{N}p\left(l\mid x_{i},\Theta^{g}\right)\left(\log x_{i}^{e^{k_{l}}-1}-\frac{x_{i}}{\theta_{l}}\log e-\log\Gamma\left(e^{k_{l}}\right)-e^{k_{l}}\log\theta_{l}\right)'_{\theta_{l}}\\&=\sum_{i=1}^{N}p\left(l\mid x_{i},\Theta^{g}\right)\frac{x_{i}-e^{k_{l}}\theta_{l}}{\theta_{l}^{2}}\end{aligned}$$

Прирівняємо останній вираз до нуля.

$$\sum_{i=1}^{N} p\left(l \mid x_i, \Theta^g\right) \frac{x_i - e^{k_l}\theta_l}{\theta_l^2} = 0 \quad \Big| \times \theta_l^2$$

$$\sum_{i=1}^{N} x_i p\left(l \mid x_i, \Theta^g\right) = \theta_l \sum_{i=1}^{N} e^{k_l} p\left(l \mid x_i, \Theta^g\right)$$

$$\Rightarrow \theta_l^{\text{new}} = \frac{\sum_{i=1}^{N} x_i p\left(l \mid x_i, \Theta^g\right)}{\sum_{i=1}^{N} e^{k_l} p\left(l \mid x_i, \Theta^g\right)}$$

4.3 Оцінка для параметру форми k_l^{new} .

$$\frac{\partial Q}{\partial k_l} = \left(\sum_{l=1}^{M} \sum_{i=1}^{N} \log \left(\frac{x_i^{\left(e^{k_l}-1\right)} e^{\left(-\frac{x}{\theta_l}\right)}}{\Gamma\left(e^{k_l}\right) \theta_l^{\left(e^{k_l}\right)}} \right) p\left(l \mid x_i, \Theta^g\right) \right)'_{k_l} =$$

$$= \sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) \left(\log \frac{x_i^{\left(e^{k_l}-1\right)} e^{\left(-\frac{x_i}{\theta_l}\right)}}{\Gamma\left(e^{k_l}\right) \theta_l^{\left(e^{k_l}\right)}} \right)'_{k_l} =$$

$$= \sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) \left(\left(e^{k_l} - 1\right) \log_i + \left(-\frac{i}{\theta_l}\right) \log e - \log \Gamma\left(e^{k_l}\right) - e^{k_l} \log \theta_l \right)'_{k_l} =$$

$$= \sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) \left(\left(e^{k_l} \log_i\right)'_{k_l} - \left(\log \Gamma\left(e^{k_l}\right)\right)'_{k_l} - \log \theta \left(e^{k_l}\right)'_{k_l} \right) =$$

$$= \sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) \left(e^{k_l} \left(\log x_i - \log \theta_l\right) - e^{k_l} \psi\left(e^{k_l}\right) \right) =$$

$$= e^{k_l} \sum_{i=1}^{N} p\left(l|x_i, \Theta^g\right) \left(\log x_i - \log \theta_l - \psi\left(e^{k_l}\right) \right)$$

Прирівняємо останній вираз до нуля і підставимо оцінку для θ_l у вираз.

$$\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\left(\log x_i-\log\hat{\theta}_l-\psi\left(e^{k_l}\right)\right)=0;$$

$$\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\log x_i=\left(\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\right)\left(\log\hat{\theta}_l-\psi\left(e^{k_l}\right)\right);$$

$$\frac{\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\log x_i}{\left(\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\right)}=\log\frac{\sum_{i=1}^{N} x_i p\left(l\mid x_i,\Theta^g\right)}{e^{k_l}\sum_{i=1}^{N} p\left(l\mid x_i,\Theta^g\right)}+\psi\left(e^{k_l}\right);$$

$$\frac{\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\log x_i}{\left(\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\right)}-\log\frac{\sum_{i=1}^{N} x_i p\left(l\mid x_i,\Theta^g\right)}{\sum_{i=1}^{N} p\left(l\mid x_i,\Theta^g\right)}-\psi\left(e^{k_l}\right)+k_l=0.$$

Отримане рівняння неможливо розв'язати явно, тому застосуємо приблизний метод Ньютона-Рафсона. В результаті знайдемо наближений параметр k_l^{new} .

5 Лінійна регресійна модель.

Застосуємо EM-алгоритм до нашої регресійної моделі. Для зручності позначимо апостеріорні ймовірності l-компоненти наступним чином:

$$w_{i,l}=\frac{w_l^* p_{\varepsilon,l}\left(\varepsilon_{i,l}\right)p_{x,\hat{\Theta}_{n,l}}\left(x_i\right)}{\sum_{k=1}^{M} w_k^* p_{\varepsilon,k}\left(\varepsilon_{i,l}\right)p_{x,\hat{\Theta}_{n,k}}\left(x_i\right)}$$

де $\varepsilon_{i,l}:=y_i-\left\langle x_i,\hat{b}_{n,l}\right\rangle$ - нормально розподілені залишки; w_l^* - апріорні ймовірності (4.1).

5.1 Оцінка дисперсії залишків

Давайте виведемо оцінку для дисперсії залишків:

$$Q\left(\hat{\theta}_n,\theta\right)=\sum_{l=1}^{M}\sum_{i=1}^{N} w_{i,l}\ln\left(p_l\right)+\sum_{l=1}^{M}\sum_{i=1}^{N} w_{i,l}\ln\left(p_{\theta,x}\left(x_i\right)\right)+$$
$$+\sum_{l=1}^{M}\sum_{i=1}^{N} w_{i,l}\ln\left(p_{\theta,\varepsilon}\left(\varepsilon_{i,l}\right)\right),$$

Нехай $\varepsilon_{i,l}$ розподілені нормально, тобто $p_{\varepsilon,l}(x) = \frac{1}{\sqrt{2\pi\sigma^2_{\varepsilon,l}}} e^{-\frac{x^2}{2\sigma^2_{\varepsilon,l}}}$. Тоді виведемо оцінку для $\sigma^2_{\varepsilon,l}$:

$$\frac{\partial Q}{\partial \sigma^2_{\varepsilon,l}} = \frac{\partial}{\partial \sigma^2_{\varepsilon,l}} \sum_{i=1}^{N} w_{i,l} \ln\left(p_{\varepsilon,l}\left(\varepsilon_{i,l}\right)\right) = \sum_{i=1}^{N} w_{i,l} \frac{\partial}{\partial \sigma^2_{\varepsilon,l}} \ln\left(p_{\varepsilon,l}\left(\varepsilon_{i,l}\right)\right) =$$
$$= \sum_{i=1}^{N} w_{i,l} \frac{\partial}{\partial \sigma^2_{\varepsilon,l}} \left(-\frac{1}{2}\frac{\varepsilon^2_{i,l}}{\sigma^2_{i,l}} - \frac{1}{2}\ln\left(2\pi\sigma^2_{\varepsilon,l}\right)\right) = \sum_{i=1}^{N} w_{i,l} \left(\frac{\varepsilon^2_{i,l}}{2}\frac{1}{\sigma^4_{\varepsilon,l}} - \frac{1}{2\sigma^2_{\varepsilon,l}}\right)$$

Прирівняємо крайній вираз до нуля і домножимо на $2\sigma^2_{\varepsilon,k}$:

$$\sum_{i=1}^{N} w_{i,l} \left(\frac{\varepsilon^2_{i,l}}{2}\frac{1}{\sigma^4_{\varepsilon,l}} - \frac{1}{2\sigma^2_{\varepsilon,l}}\right) = 0 \quad \Big| \times 2\sigma^2_{\varepsilon,k}$$
$$\frac{1}{\sigma^2_{\varepsilon,l}} \sum_{i=1}^{N} w_{i,l}\varepsilon^2_{i,l} = \sum_{i=1}^{N} w_{i,l}$$
$$\hat{\sigma}^2_{\varepsilon,l} = \frac{\sum_{i=1}^{N} w_{i,l}\varepsilon^2_{i,l}}{\sum_{i=1}^{N} w_{i,l}}.$$

5.2 Параметри регресії

Давайте виведемо оцінку для параметрів регресії:

$$\mathcal{L}_{b_l}(X,Y) = \sum_{i=1}^{N} w_{i,l}\left(y_i - \langle x_i, b_l\rangle\right)^2, W = \begin{bmatrix} w_{1,l} & 0 & 0 & \cdots \\ 0 & w_{2,l} & 0 & \cdots \\ 0 & 0 & w_{3,l} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

де W - матриця апостеріорних ймовірностей.

$$\nabla \mathcal{L}_{b_l} = \sum_{i=1}^{N} -2w_{i,l}x_i\left(y_i - \langle x_i, b_l\rangle\right)$$

Прирівняємо крайній вираз до нуля та поділимо на -2

$$\sum_{i=1}^{N} -2w_{i,l}x_i\left(y_i - \langle x_i, b_l\rangle\right) = 0 \quad \Big| \times -2$$

Перейдемо до матричного запису задля спрощення операцій.

$$\nabla \mathcal{L}_{B_l} = (W_l X)^\top Y - (W_l X)^\top X B_l = 0$$

Так як $W_l^\top = W_l \Rightarrow (W_l X)^\top = X^\top W_l^\top = X^\top W_l$. Тоді:

$$\begin{aligned} &(W_l X)^\top Y = (W_l X)^\top X B_l \quad \Big| \times (X^\top W_l X)^{-1} \\ &\hat{B}_l = \left(X^\top W_l X\right)^{-1} X^\top W_l Y. \end{aligned}$$

6 Моделювання.

Для перевірки якості отриманих оцінок, була проведена серія експериментів на модельованих даних. Дані для експериментів генерувалися на основі моделі двокомпонентної суміші ($M = 2$). Для кожної вибірки розміру n було проведено 1000 експериментів.

6.1 Експеримент 1.

Даний експеримент перевіряє якість побудованих оцінок для параметрів суміші гамма-розподілів за допомогою EM-алгоритму. В якості метрики якості використаємо MSE (Середньоквадратична похибка). Значення параметрів для цього експерименту зазначено у таблиці 6.1.

	$l = 1$	$l = 2$
w_l	0.5	0.5
k_l	$\log 2$	$\log 6$
θ_l	2	2

Таблиця 6.1. Параметри розподілів для моделювання експерименту 1.

Для кращої наочності зобразимо графік теоретичної функції щільності гамма-розподілів для заданих параметрів жовтим кольором на Рис. 1. Також на цьому рисунку синім кольором зобразимо гістограму вибірки даних.

Як видно з результатів експериментів у таблиці 6.2, на заданих параметрах суміші, що не сильно різняться між собою, маємо відносно невеликі значення середньоквадритичної похибки.

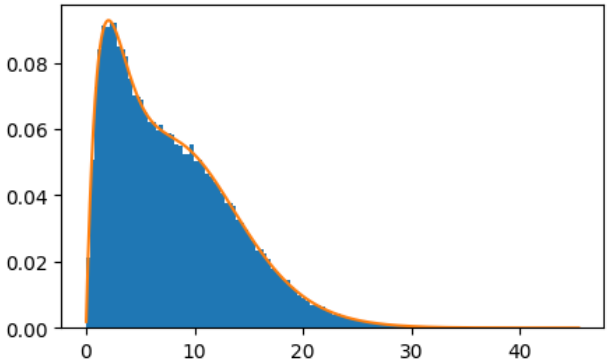

Рис. 1: Теоретична функція щільності суміші та гістограма вибірки.

	w_l		k_l		θ_l	
n	$l=1$	$l=2$	$l=1$	$l=2$	$l=1$	$l=2$
1×10^2	1.3e-05	1.3e-05	2.7e-02	3.0e-01	9.7e-01	1.2e+00
1×10^4	1.9e-06	1.9e-06	7.6e-03	2.1e-01	1.0e+00	1.0e+00
1×10^5	2.7e-06	2.7e-06	1.8e-01	6.9e-01	1.2e+00	1.4e+00

Таблиця 6.2. MSE (Середньоквадратична похибка).

6.2 Експеримент 2.

Слід зауважити, що деколи EM-алгоритм може повертати оцінки відмінні від очікуваних, оскільки заходить у локальний максимум правдоподібності (вірогідності). Для явної демонстрації цього випадку розглянемо суміш, котра містить параметри k, що значно відрізняються один від одного. Значення параметрів для цього експерименту зазначено у табл. 6.3.

Як бачимо з таблиці результатів 6.4, аномальні випадки (Рис. 3) призводять до значного зросту MSE параметру θ_l.

	$l=1$	$l=2$
w_l	0.5	0.5
k_l	$\log 10$	$\log 100$
θ_l	2	2

Таблиця 6.3. Параметри розподілів для моделювання експерименту 2.

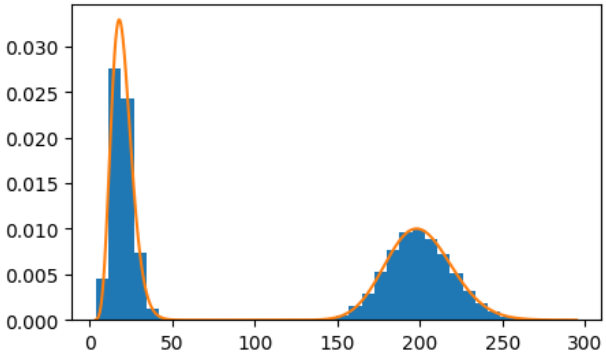

Рис. 2: Теоретична функція щільності суміші та гістограма вибірки.

	w_l		k_l		θ_l	
n	$l=1$	$l=2$	$l=1$	$l=2$	$l=1$	$l=2$
1×10^2	0.0e+00	4.3e-34	7.8e-01	3.0e+00	1.7e+03	1.7e+03
1×10^4	0.0e+00	0.0e+00	8.3e-01	3.3e+00	1.8e+03	1.8e+03
1×10^5	1.1e-32	2.6e-33	7.6e-01	3.0e+00	1.7e+03	1.7e+03

Таблиця 6.4. MSE (Середньоквадратична похибка).

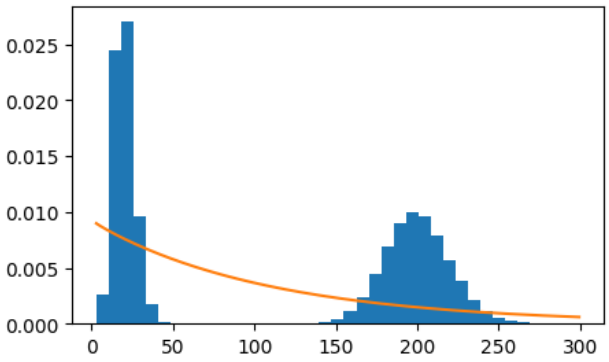

Рис. 3: Приклад аномального випадку.

6.3 Експеримент 3.

Розглянемо суміш, в котрій параметри розподілу θ кількісно значно відрізняються. Значення параметрів для цього експерименту зазначено у табл. 6.5.

Як бачимо з таблиці результатів 6.6, аномальні випадки (Рис. 5) знову призводять до значного зросту MSE параметру θ_l, як і в попередньому експерименті.

	$l = 1$	$l = 2$
w_l	0.5	0.5
k_l	$\log 5$	$\log 5$
θ_l	45	8

Таблиця 6.5. Параметри розподілів для моделювання експерименту 3.

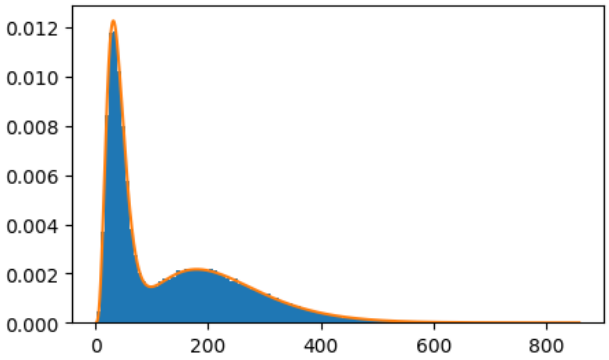

Рис. 4: Теоретична функція щільності суміші та гістограма вибірки.

	w_l		k_l		θ_l	
n	$l = 1$	$l = 2$	$l = 1$	$l = 2$	$l = 1$	$l = 2$
1×10^2	5.0e-06	5.0e-06	2.7e-01	2.7e-01	5.0e+02	7.3e+02
1×10^4	6.9e-06	6.9e-06	2.0e-01	2.0e-01	3.6e+02	9.8e+02
1×10^5	4.0e-08	4.0e-08	2.4e-01	2.4e-01	4.4e+02	1.2e+03

Таблиця 6.6. MSE (Середньоквадратична похибка).

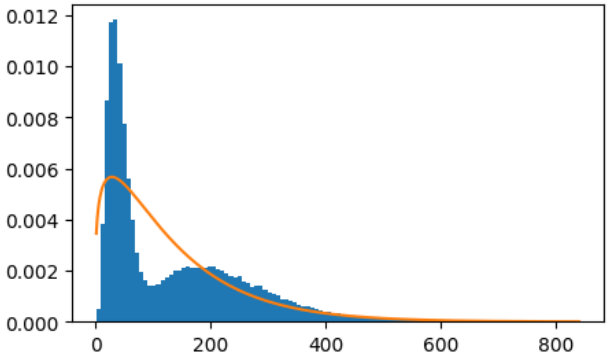

Рис. 5: Приклад аномального випадку.

6.4 Експеримент 4.

Повторимо Експеримент 3, але цього разу будемо відкидати аномальні випадки, проводячи один і той же експеримент двічі, та обиратимемо найкращий. Значення параметрів для цього експерименту зазначено у табл. 6.5.

В результаті отримали значне покращення результатів: середньоквадратична похибка набула адекватних значень для параметру θ та зменшується для всіх параметрів зі збільшенням обсягу вибірки.

	w_l		k_l		θ_l	
n	$l=1$	$l=2$	$l=1$	$l=2$	$l=1$	$l=2$
1×10^2	3.5e-04	3.5e-04	8.2e-03	8.3e-02	8.2e+00	9.6e-01
1×10^4	1.1e-06	1.1e-06	5.0e-04	2.6e-04	1.8e+00	2.2e-02
1×10^5	1.4e-09	1.4e-09	2.2e-05	5.6e-05	5.3e-02	3.3e-03

Таблиця 6.7. MSE (Середньоквадратична похибка).

6.5 Експеримент 5.

Тепер проведемо експеримент для перевірки якості оцінок лінійної регресійної моделі. Для моделювань використаємо просту регресію:

$$Y_l = b_{l,0} + b_{l,1} * X_l + \varepsilon_l,$$

де $X_l \sim Gamma\,(k_l, \theta_l)$ - регресори, Y_l - відгуки, l - латентний індекс компоненти, ε_l - незміщена похибка регресії.

Значення параметрів для цього експерименту зазначено у табл. 6.8. Як бачимо, результати перевірки нашої моделі в табл. 6.9 для кожної вибірки мають достатньо низьку середньоквадратичну похибку.

	$l=1$	$l=2$
w_l	0.5	0.5
k_l	$\log 10$	$\log 100$
θ_l	2	2
b_0	1	1
b_1	2	3

Таблиця 6.8. Параметри регресійної суміші експерименту 5.

	b_0		b_1	
n	$l=1$	$l=2$	$l=1$	$l=2$
1×10^2	6.7e-02	4.4e-02	8.2e-02	6.4e-02
1×10^4	3.3e-03	5.9e-03	6.3e-03	9.3e-03
1×10^5	4.1e-04	9.0e-04	3.0e-04	8.4e-04

Таблиця 6.9. MSE (Середньоквадратична похибка).

7 Висновки.

Математичні моделі розглянуті в цій роботі не нові та вивчаються багатьма вченими. Проте, від цього робота не втрачає цінність:

- Було розглянуто специфічний випадок суміші для гамма-розподілів та виведено відповідні теоретичні оцінки. Також, було виявлено цікаву проблему пов'язану з наближеними обчисленнями оцінок параметрів та наведено метод її розв'язання.

- Під час моделювань в експериментах виявили систематичну проблему, що впливала на якість алгоритму в середньому. В наступних експериментах наводиться приклад як вирішити цю проблему, що значно покращує результати.

В майбутньому в цій роботі варто глибше дослідити якість моделі на більш складних функціях регресії, більшій кількості компонент суміші та застосувати на справжніх даних.

Також хочу висловити подяку моєму керівнику Віталію Мірошниченко за менторство впродовж виконання цієї роботи.

8 Список літератури.

[1] Bilmes, Jeff. (2000). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report ICSI-TR-97-021, University of Berkeley.

[2] Faria, S. & Soromenho, Gilda. (2010). Fitting mixtures of linear regressions. Journal of Statistical Computation and Simulation.

[3] Miroshnichenko V., Maiboroda R., Confidence ellipsoids for regression coefficients by observations from a mixture, Modern Stoch. Theory Appl. 5(2018), no. 2, 225-245, DOI 10.15559/18-VMSTA105

[4] Quandt, R. E., & Ramsey, J. B. (1978). Estimating Mixtures of Normal Distributions and Switching Regressions. Journal of the American Statistical Association, 73(364), 730–738.

[5] Liubashenko D., Maiboroda R. "Linear regression by observations from mixture with varying concentrations". Modern Stochastics: Theory and Applications, Vol.2, Iss.4 pp. 343 - 353, - 2015