Вычислительные и статистические аспекты модели IRT оценивания результатов тестов и вопросов

Понизова Вероника Сергеевна, гр. 14.Б02-ММ

Санкт-Петербургский Государственный Университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели.

Научный руководитель — к.ф.-м.н., **доцент А.И. Коробейников**
Рецензент — **м.н.с. А.Ю. Шлемов**

Санкт-Петербург
2018г.

Латентные признаки – скрытые качества личности, не поддаются непосредственному измерению.
Пример: способность человека к некоторому предмету учебной программы.

Item Response Theory (IRT): по ответам на тестовые вопросы можно оценивать способность людей и сложность вопросов в рамках некоторой параметрической модели.

Модель Раша, 1960: $\theta \in \mathbb{R}$ — способность человека, $\beta \in \mathbb{R}$ — сложность вопроса, бернуллиевская с.в. $\xi \in \{0, 1\}$ — правильность ответа человека на вопрос.

$$P(\xi = 1|\theta, \beta) = \frac{\exp(\theta - \beta)}{1 + \exp(\theta - \beta)}.$$

Постановка задачи

Для N человек и J вопросов: $\theta = (\theta_1 \dots \theta_N)$, $\beta = (\beta_1 \dots \beta_J) \Rightarrow$

$$P(\xi_{ij} = 1 | \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}.$$

Выборка: матрица ответов $\mathbf{X} = \{x_{ij}\}_{i,j=1}^{N,J}$

$$x_{ij} = \begin{cases} 0, & \text{если } i\text{-й человек ответил верно на } j\text{-й вопрос;} \\ 1, & \text{иначе.} \end{cases}$$

Проблема: при $N \to \infty$ растет размерность $\mathbf{X}$.

Задача: по матрице ответов $\mathbf{X}$ необходимо:

- оценить набор параметров способностей респондентов $\theta = (\theta_1 \dots \theta_N)$ при мешающих параметрах сложности вопросов $\beta = (\beta_1, \dots, \beta_J)$;
- исследовать возможность сравнения оценок параметров способности между разными группами респондентов.

Векторы $\theta = (\theta_1, \dots \theta_N), \beta = (\beta_1, \dots, \beta_J)$ — неизвестный, но фиксированный набор параметров.
Предположение о независимости: $\{x_{ij}\}_{i,j=1}^{N,J}$ независимы в совокупности.
Тогда:

$$P(\mathbf{X}|\theta, \beta) = \prod_{i=1}^{N} \prod_{j=1}^{J} \frac{\exp(x_{ij}(\theta_i - \beta_j))}{1 + \exp(\theta_i - \beta_j)}$$

Линейное ограничение:

$$\sum_{j=1}^{J} \beta_j = 0$$

Оценки с пом. метода полного правдоподоибия:

$$(\hat{\theta}_{JML}, \hat{\beta}_{JML}) = \arg\max_{\theta, \beta} P(\mathbf{X}|\theta, \beta)$$

Предложение [Ghosh, 1995]

При фиксированном количестве вопросов J и $N \to \infty$ оценки параметров модели, полученные с помощью метода полного максимального правдоподобия вообще говоря являются несостоятельными.

Свойство модели: $r_i = \sum_{j=1}^{J} x_{ij}$ — достаточная статистика для параметра θ_i.

Оценивание параметра сложности:

$$L_\beta(\beta|r_1,\ldots,r_N) = \prod_{i=1}^{N} \frac{\exp(\sum_{j=1}^{J} -\beta_j x_{ij})}{\sum_{\mathbf{Y}|r_i} \exp(\sum_{j=1}^{J} -\beta_j y_j)},$$

где $\mathbf{Y}|r_i$: $\mathbf{Y} = (y_1,\ldots,y_J) \in \{0,1\}^J$ такие, что $\sum_{j=1}^{J} y_j = r_i$. Тогда:

$$\hat{\beta}_{CML} = \arg\max_\beta L_\beta(\beta|r_1,\ldots,r_N).$$

Свойства $\hat{\beta}_{CML}$ [Andersen, 1970]

Оценки $\hat{\beta}_{CML}$ являются состоятельными при $J \to \infty$.

Свойство модели: $r_i = \sum_{j=1}^{J} x_{ij}$ — достаточная статистика для параметра θ_i.

Оценивание параметра сложности:

$$L_\beta(\beta|r_1, \dots, r_N) = \prod_{i=1}^{N} \frac{\exp(\sum_{j=1}^{J} -\beta_j x_{ij})}{\sum_{\mathbf{Y}|r_i} \exp(\sum_{j=1}^{J} -\beta_j y_j)},$$

где $\mathbf{Y}|r_i$: $\mathbf{Y} = (y_1, \dots, y_J) \in \{0,1\}^J$ такие, что $\sum_{j=1}^{J} y_j = r_i$. Тогда:

$$\hat{\beta}_{CML} = \arg\max_\beta L_\beta(\beta|r_1, \dots, r_N).$$

Свойства $\hat{\beta}_{CML}$ [Andersen, 1970]

Оценки $\hat{\beta}_{CML}$ являются состоятельными при $J \to \infty$.

Оценивание параметра способности: считаем, что уже получены $\hat{\beta}_{CML}$:

$$\hat{\theta}_{CML} = \arg\max_\theta P(\mathbf{X}, \hat{\beta}_{CML}|\theta) = \prod_{i=1}^{N} \prod_{j=1}^{J} \frac{\exp(x_{ij}(\theta_i - \hat{\beta}_j))}{1 + \exp(\theta_i - \hat{\beta}_j)}$$

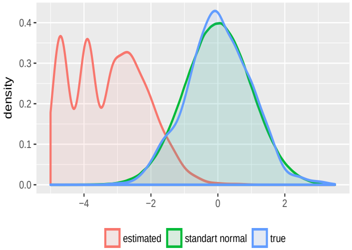
Задача оценивания параметров: Conditional Max. Likelihood

Задача: исследовать свойства оценок параметра способности θ, получаемых с помощью алгоритма условного максимального правдоподобия.

Эксперимент: $N = 1000$, $J = 40 \Rightarrow$ моделирование матрицы ответов с входными параметрами $\theta_i \sim \mathbf{N}(a_\theta, \sigma_\theta^2), \beta_j \sim \mathbf{N}(a_\beta, \sigma_\beta^2)$.

Два случая:

1. $\theta \sim \mathbf{N}(0, 1)$, параметры a_β и σ_β распределения β варьируются $\Rightarrow$ применяется алгоритм усл. макс. правдоподобия с учетом $\sum_{j=1}^{J} \beta_j = 0$.
   Вопрос: свойства $\hat{\theta}_{CML}$ в совокупности?
2. $\theta_1, \ldots, \theta_{N/2} \sim \mathbf{N}(-5, 1)$, $\theta_{N/2+1}, \ldots, \theta_N \sim \mathbf{N}(5, 1)$ — необходимо сравнить две группы респондентов между собой.
   Вопрос: сохранится ли разница в 10 между средними в выделенных группах?



(a) Сравнение исходного распределения $\theta \sim \mathbf{N}(0,1)$ и распределения $\hat{\theta}_{CML}$ в случае $\beta \sim \mathbf{N}(3,1)$

(b) Сравнение доверительного интервала для разницы средних и истинного значения разницы в случае $\beta \sim \mathbf{N}(0,1)$

Вывод: Cond. Max. Likelihood как двухшаговая процедура оценивания параметров не может быть применена для оценивания θ и корректного сравнения этих оценок между собой.
$\Rightarrow$ необходимо изменить или модель, или способ оценивания.

Смена модели: θ — с. в. с функцией распределения $F(\theta)$, β — фиксированный набор параметров.
$\Rightarrow (\theta_1 \dots \theta_N)$ — выборка из распределения $\mathcal{L}(\theta)$.

Дискретный случай: $\theta \sim \begin{pmatrix} q_1 & \dots & q_K \\ \pi_1 & \dots & \pi_K \end{pmatrix}$, где

- $(q_1 \dots q_K)$ — известные значения;
- $(\pi_1 \dots \pi_K)$ – неизвестные вероятности.

Преимущества: количество K неизвестных параметров модели не растет с увеличением N.
Обозначения:

- n_k — число респондентов, для которых значение параметра способности — q_k;
- r_{jk} — число респондентов из n_k, ответивших верно на j-ый вопрос;
- $P(q_k, \beta_j) = \frac{\exp(q_k - \beta_j)}{1 + \exp(q_k - \beta_j)}$

Оценки параметров можно получать с помощью EM–алгоритма.

- Если значения параметров $(\pi_1 \ldots \pi_K)$ неизвестны:

$$\log L(\mathbf{X}|\beta,\theta) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} P(q_k,\beta_j)^{x_{ij}} (1 - P(q_k,\beta_j))^{1-x_{ij}} \right)$$

Недостатки: вычислительная сложность при максимизации $\log L(X|\beta,\theta)$.

- Предположим, что значения $(\pi_1 \ldots \pi_K)$ известны:

$$\log L(\mathbf{X}, n_k, r_{jk}|\beta,\pi) = \sum_{j=1}^{J} \sum_{k=1}^{K} r_{jk} \log P(q_k,\beta_j) + (n_k - r_{jk}) \log(1 - P(q_k,\beta_j)) + n_k \pi_k$$

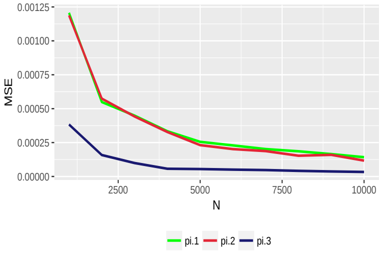
Преимущества: максимизация такого выражения проста в вычислительном плане.

Схема алгоритма:

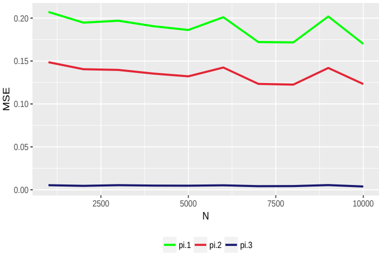
1. Е–шаг: вычисление $n_k^{(s)} = \mathbb{E}(n_k|\mathbf{X},\beta^{(s)},\pi^{(s)})$ и $r_{ij}^{(s)} = \mathbb{E}(r_{ij}|\mathbf{X},\beta^{(s)},\pi^{(s)})$.
2. М–шаг: $\pi_k^{(s+1)}$ и $\beta_j^{(s+1)}$ — т. максимума условного математического ожидания $\log L(\mathbf{X}, n_k, r_{jk}|\beta,\pi)$ относительно n_k и r_{jk}.

Задача: реализация ЕМ–алгоритма и проверка свойств получаемых оценок.

- $K = 3,\ \theta \sim \begin{pmatrix} 2 & 3 & 5 \\ 0.4 & 0.2 & 0.4 \end{pmatrix}$;
- $N = 1000,\ J = 10,$ $\beta_j \sim \mathbf{N}(3, 1)$.



(a) хороший случай



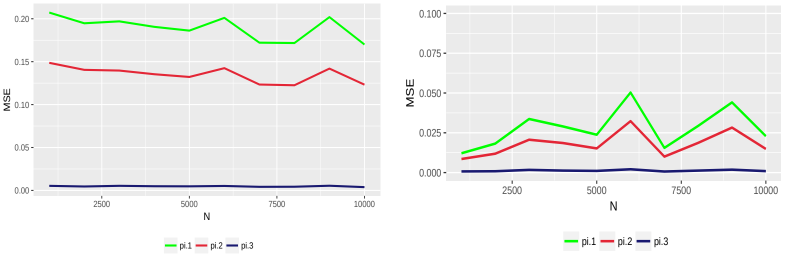
(b) плохой случай

Рис.: Зависимость $\hat{\mathbb{E}}(\pi_i - \hat{\pi}_i)^2$ от кол-ва человек N

Задача: подобрать более устойчивую к выбору начального приближения модификацию базового алгоритма.
Идея: модифицировать M–шаг исходного алгоритма: $\lambda^{(s+1)} = \kappa\lambda^{(s)}$, и вместо $\log L(\mathbf{X}, n_k, r_{jk}|\beta, \pi)$ рассматривать

$$\log L(\mathbf{X}, n_k, r_{jk}|\beta, \pi) - \lambda^{(s+1)}\mathcal{R}(\pi, \beta).$$

Результаты для $\mathcal{R}(\pi, \beta) = \|\beta\|_2^2$:



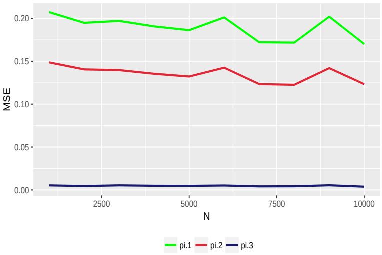
(a) Базовый алгоритм

(b) Регуляризация

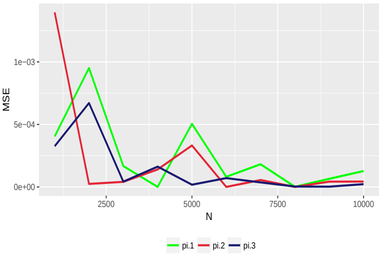
Рис.: Зависимость $\hat{\mathbb{E}}(\pi_i - \hat{\pi}_i)^2$ от кол-ва человек N

Задача: подобрать более устойчивую к выбору начального приближения модификацию базового алгоритма.

Идея: запускать алгоритм из Q случайных начальных приближений $\Rightarrow$ выбирать лучшую оценку. Результат для $Q = 200$:



(a) Базовый алгоритм



(b) Мультистарт

Рис.: Зависимость $\hat{\mathbb{E}}(\pi_i - \hat{\pi}_i)^2$ от кол-ва человек N

Вывод: в дискретном случае за счет структуры модели (фиксированное количество параметров K) удается оценивать $\mathcal{L}(\theta)$. С помощью модификаций можно повысить точность оценок.

Пусть $\theta = (\theta_1, \ldots, \theta_N)$ и $\beta = (\beta_1, \ldots, \beta_N)$ — случайные вектора с непрерывным распределением. Из т. Байеса:

$$p(\theta, \beta|\mathbf{X}) \propto p(\theta, \beta)p(\mathbf{X}|\theta, \beta)$$

Предложение [Tierney, 1994]

Можно построить марковскую цепь $M_0, M_1, \ldots, M_n, \ldots$ где $M_n = (\theta^{(n)}, \beta^{(n)})$ такую, что её стационарное распределение совпадает с апостериорным распределением параметров θ и β (при выполнении некоторых условий регулярности), то есть:

$$\mathcal{L}(M_n) \underset{n\to\infty}{\to} \mathcal{L}(\theta, \beta|\mathbf{X}).$$

В работе рассмотрен алгоритм Metropolis-Hastings within Gibbs для построения марковской цепи с вышеуказанным свойством.

Задача: исследовать возможность применения алгоритма для получения оценок распределения θ и сравнения распределений между собой.

Эксперимент: моделируется матрица $\mathbf{X}$ с входными параметрами $\theta = (\theta_1, \ldots, \theta_N)$ и $\beta = (\beta_1, \ldots, \beta_J)$, причем

- распределение $\theta_i \sim \mathbf{N}(5, 1)$ фиксировано;
- $\beta_j \sim \mathbf{N}(a_\beta, \sigma_\beta^2)$, параметры a_β, σ_β варьируются.

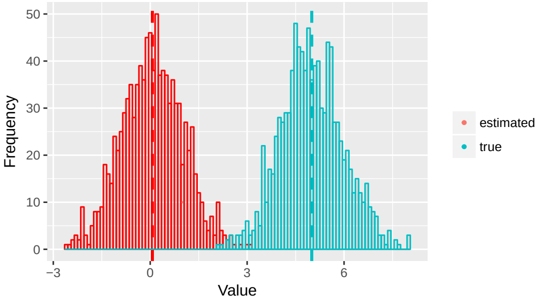


Рис.: Апостериорное распределение параметра θ

Вывод: не представляется возможным оценивать исходное распределение параметров. θ.

Пусть $\theta_1, \ldots, \theta_{N/2} \sim \mathbf{N}(-5, 1)$ и $\theta_{N/2+1}, \ldots, \theta_N \sim \mathbf{N}(5, 1)$.
Вопрос: сохранится ли разница между средними в двух выделенных группами?

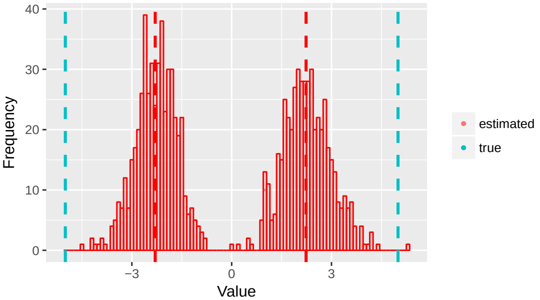


Рис.: Апостериорное распределение параметра θ

Вывод: разница в средних составляет ≈ 5, вместо ожидаемых 10 $\Rightarrow$
за счет отсутствия каких-либо ограничений на параметры алгоритм нельзя использовать для сравнения оценок между разными группами людей.

Заключение:

- Рассмотрено три подхода (**CML**, **MML**, **MCMC**) для получения оценок параметров модели Раша;
- Показано, что для поставленной задачи нельзя использовать **CML**;
- Реализован EM–алгоритм (**MML**) для оценивания априорного распределения параметра способности в дискретном случае. Показано, что параметры модели удается оценивать, и что базовый алгоритм неустойчив к выбору начального приближения. Предложено несколько модификаций, обладающих большей устойчивостью к выбору начального приближения;
- Реализован алгоритм Metropolis-Hastings within Gibbs для моделирования случайных величин с распределением, совпадающим с апостериорным распределением параметров модели. Показано, что он не может быть использован для оценивания исходного распределения параметра способности, а также для сравнения оценок между разными группами людей.
- Модель Раша непригодна для использования в оценивании параметра способности.