Локализация сигнала в задаче интерпретации множества статистических тестов[1]

Анастасия Процветкина

14 мая, 2021

Санкт-Петербургский государственный электротехнический университет "ЛЭТИ"
Санкт-Петербург, Россия.

[1] Руководитель: Сергей Васильевич Малов

Локализация в категориальных данных

Постановка задачи

Имея таблицу сопряженности размера $m \times n$, можно проверить гипотезу независимости признаков с помощью хи-квадрат статистики.

Известно, что статистика критерия при нулевой гипотезе имеет асимптотическое распределение

$$\chi^2_{(m-1)(n-1)}$$

По определению, хи-квадрат величина X^2 с k степенями свободы есть сумма квадратов k независимых одинаково распределенных нормальных случайных величин X_i, т.е.

$$X^2 = X_1^2 + X_2^2 + \cdots + X_k^2$$

Постановка задачи

С каждой таблицей $m \times n$ мы *теоретически* имеем ассоциированный с ней вектор $\mathbf{X} = (X_1, X_2, \ldots, X_{(m-1)(n-1)})^T$, где X_i асимптотически НОРСВ из $\mathcal{N}(0,1)$.

Пусть имеется w таких таблиц, имеющих общий признак. Сформируем длинный вектор из всех асимптотически нормальных компонент:

$$\boldsymbol{\xi} = (\underbrace{X_{11}, X_{12}, \ldots, X_{1,(m-1)(n-1)}}_{\text{норм. компоненты из 1-й таб.}}, \ldots, \underbrace{X_{w1}, X_{w2}, \ldots, X_{w,(m-1)(n-1)}}_{\text{норм. компоненты из w-й таб.}})^T$$

Очевидно, что $\boldsymbol{\xi} \Rightarrow \mathcal{N}(0, \Sigma_q),\ q = w(m-1)(n-1)$.

Постановка задачи

Теорема (Rao, Mitra)

$$\boldsymbol{\psi} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma),\ r = \mathrm{rk}(\Sigma),$$

$$\Sigma^{-} \text{ – обобщенная обратная матрица для } \Sigma$$

1. Квадратичная форма $\boldsymbol{\psi}^T \Sigma^{-} \boldsymbol{\psi}$ не зависит от выбора Σ^{-}.
2. $\boldsymbol{\psi}^T \Sigma^{-} \boldsymbol{\psi} \sim \chi^2_{\delta, r}$ ($\delta = \boldsymbol{\mu}^T \Sigma^{-} \boldsymbol{\mu}$ – параметр нецентральности).

Комбинированная статистика

Для вектора $\boldsymbol{\xi} \Rightarrow \mathcal{N}(0, \Sigma)$ из всех асимптотически нормальных величин, соответствующих компонентам хи-квадрат статистики из w таблиц сопряженности, построим квадратичную форму

$$\boldsymbol{\xi}^T \Sigma^{+} \boldsymbol{\xi} \overset{H_0}{\Rightarrow} \chi^2_r,$$

H_0: “Ни в одной из $W = \{1, \ldots, w\}$ таблиц нет зависимости”,
$r = \mathrm{rk}(\Sigma), \Sigma^{+}$ – обобщенная обратная матрица Мура-Пенроуза.

Задача 1

Хи-квадрат статистика для проверки гипотезы независимости обычно записывается как

$$X^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

где O_{ij} – наблюденные частоты в ячейке (i, j), E_{ij} – соответствующие ожидаемые частоты при справедливости гипотезы независимости.

X^2 имеет $(m-1)(n-1)$ степеней свободы, но *алгебраически* является суммой $m \cdot n$ зависимых слагаемых, а нам необходимо представление вида:

$$X^2 = X_1^2 + \dots X_{(m-1)(n-1)}^2$$

$\Longrightarrow$ необходимо преобразование.

Задача 1

Решение

Ирвином и Ланкастером было предложено ортогональное преобразование, приводящее хи-квадрат статистику с $m \cdot n$ слагаемыми в сумму квадратов $(m-1)(n-1)$ асимптотических независимых нормальных величин.

Задача 2

Вектор из всех асимп. нормальных компонент сформирован,т.е.

$$\boldsymbol{\xi}=(\underbrace{X_{11}, X_{12}, \ldots, X_{1,(m-1)(n-1)}}_{\text{норм. компоненты из 1-й таб.}}, \ldots, \underbrace{X_{w1}, X_{w2}, \ldots, X_{w,(m-1)(n-1)}}_{\text{норм. компоненты из w-й таб.}})^T$$

получен. Построение квадратичной формы $\boldsymbol{\xi}^T \Sigma^+ \boldsymbol{\xi}$ требует оценивания ковариационной матрицы Σ, т.е. оценивания ковариации между нормальными величинами, соответствующими компонентам хи-квадрат статистики из разных таблиц.

Замечание

Каждая величина X_{ij} представляет одну степень свободы статистики хи-квадрат, поэтому $\mathbf{var}\,(X_{ij}) = 1$. В дальнейшем матрицу Σ будем называть корреляционной, нежели просто ковариационной.

Задача 2

Решение

Оценки корреляционной матрицы Σ были получены в частных случаях 2×2 и 3×2. Формулы громоздки, поэтому не приводятся.

Задача 3

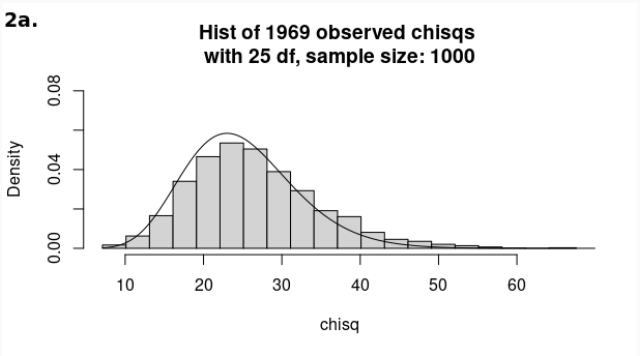
Теоретическая и неизвестная матрица Σ обладает неустойчивым свойством положительной полу-определенности. Оценивание матрицы может привести к его потере, поэтому необходимо учитывать такую потенциальную проблему и иметь решение в случае возникновения.

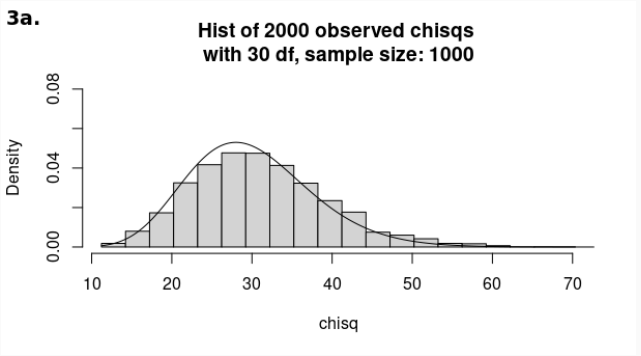
Решение

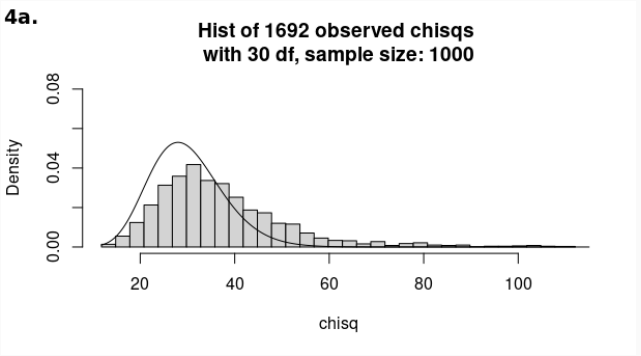
Согласно Nigham, ближайшую положительно-определенную матрицу для Σ можно построить, обнулив отрицательные собственные числа Σ.

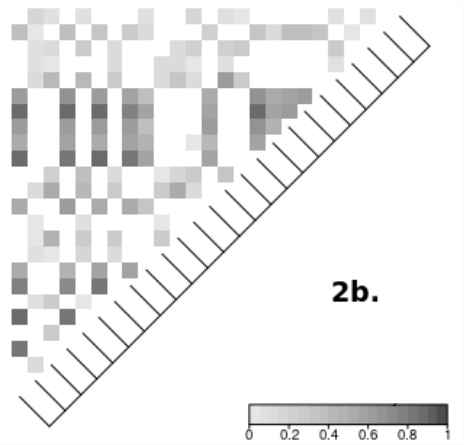
Результаты

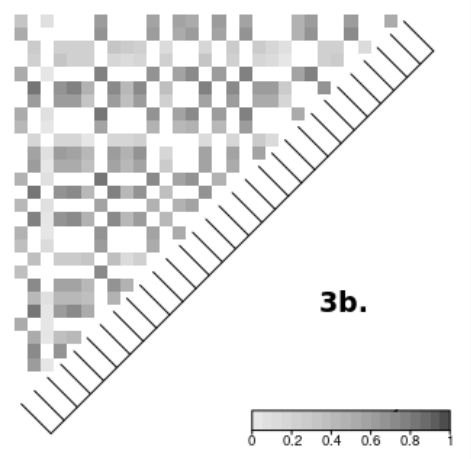
Сравнение распределений: 2×2 случай

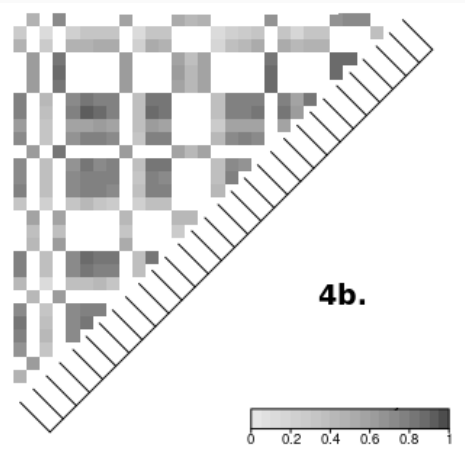








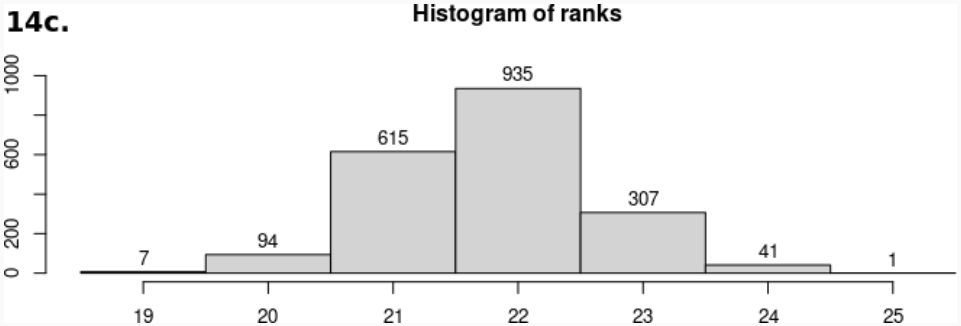


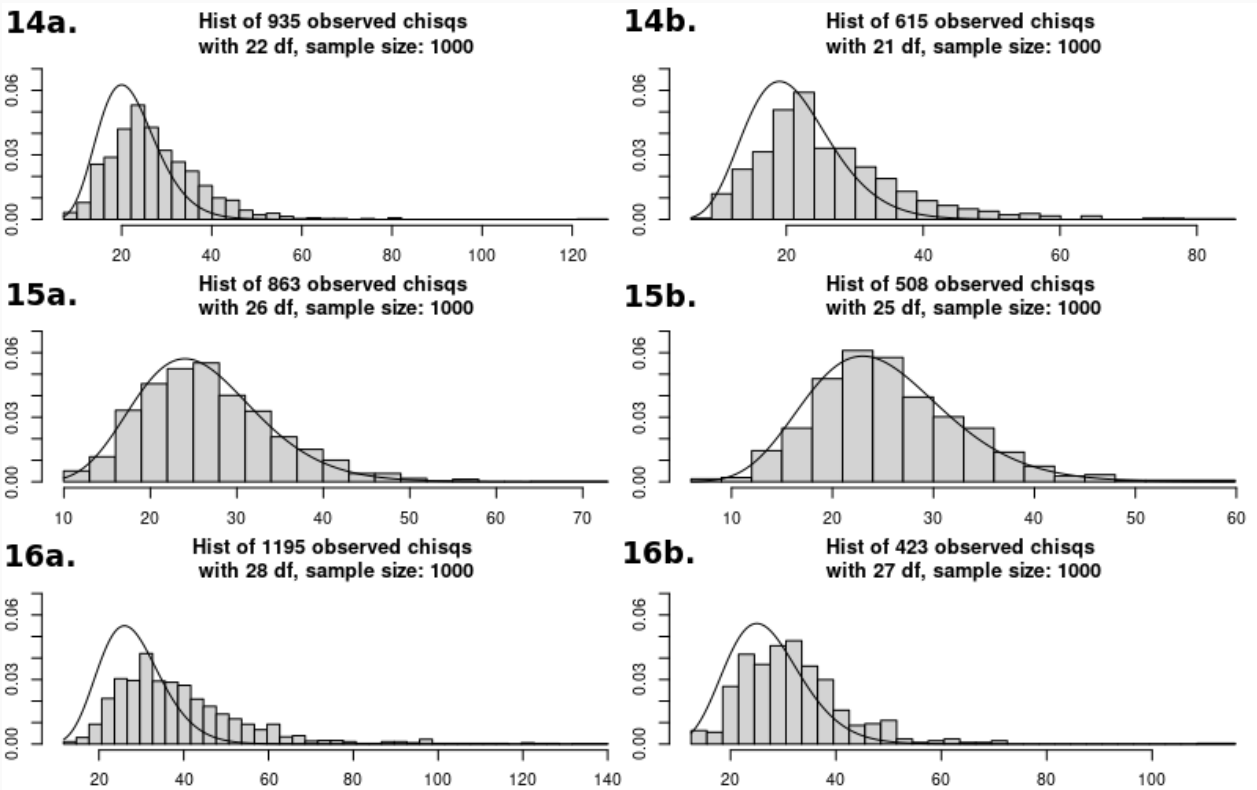


2a,b: $w = 25$, **3a,b**: $w = 30$, **4a,b**: $w = 30$

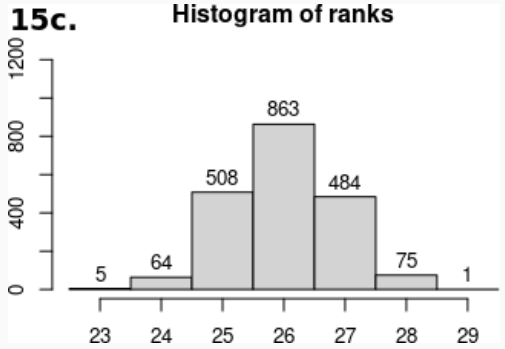
$N = 1000$, маргинальные вероятности по 0.5.

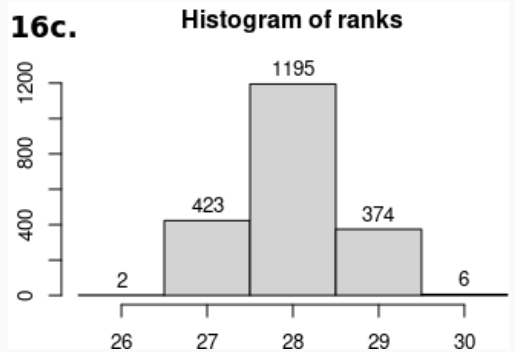
Сравнение распределений: 3×2 случай





Распределение для 2 наиболее частых рангов. $w = 15$, $N = 1000$





Заключение

- Разработан новый метод, позволяющий объединять отдельные слабые сигналы в единый более мощный сигнал
- Получены оценки корреляционной матрицы для асиптотически нормального базисного вектора
- Потенциальные проблемы вырождаемости матрицы Σ и потеря положительной определенности были предвидены и разрешены
- Разработанный метод протестирован на симулированных зависимых данных
- Отмечена сходимость распределений, но для окон малого размера
- Матрица Σ вырождается с увеличением размера окна и зависимости между маркерами