

РОБАСТНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ МОДЕЛИ ГРУППИРОВАННЫХ БИНАРНЫХ ДАННЫХ

М. А. Пашкевич

Введение

Для моделирования совокупности наблюдаемых последовательностей бинарных исходов традиционно используется бета-биномиальное распределение вероятностей (ББР) [1; 2; 3; 4]. Классические методы оценивания параметров ББР достаточно хорошо исследованы для гипотетической модели наблюдений [5; 6; 7; 8; 9]. Однако на практике гипотетическая вероятностная модель наблюдений, как правило, оказывается неадекватной в силу искажений различных типов [10; 11; 12], в результате чего классические оценки теряют свойство состоятельности. Поэтому возникает задача построения новых статистических оценок, устойчивых к этим искажениям. В данной работе предлагается новый метод оценивания параметров ББР, позволяющий получить состоятельные оценки в случае стохастических искажений бинарных данных.

1. Постановка задачи

Пусть имеются некоторая совокупность из K объектов и некоторое случайное событие A . Над каждым объектом этой совокупности производится серия из n испытаний. Результаты испытаний описываются бинарной матрицей $B = (b_{ij})_{K \times n}$, при этом выполняются предположения:

- П1. Вероятностные свойства объектов в процессе испытаний не изменяются.
- П2. Вероятностные свойства объектов внутри группы неоднородны, и для i -того объекта вероятность p_i наступления случайного события A является случайной величиной, которая имеет бета распределение с параметрами α, β , причем $p_1, p_2, \dots, p_K$ независимы в совокупности.

Пусть на бинарную матрицу B воздействуют случайные ошибки $\{\eta_{ij}\}$ и наблюдается искаженная матрица $\tilde{B} = (\tilde{b}_{ij})$:

$$\tilde{b}_{ij} = b_{ij} \oplus \eta_{ij}, \quad P\{\eta_{ij} = 1 | b_{ij} = 0\} = \varepsilon_0, \quad P\{\eta_{ij} = 1 | b_{ij} = 1\} = \varepsilon_1, \quad (1)$$

где $\oplus$ – операция сложения по модулю два, $\{\eta_{ij}\}$ – независимые случайные величины Бернулли, $\varepsilon_0, \varepsilon_1 \in [0, 1]$ – уровни искажений. Рассматривается задача робастного статистического оценивания пара-

метров α, β по косвенным данным – искаженной выборке объема K : $X = (x_1, x_2, \dots, x_K)$, где $x_i = \sum_{j=0}^n \tilde{b}_{ij}$.

2. Алгоритм оценивания

В предыдущих работах доказано [10, 11], что при наличии искажений (1) классические оценки, построенные по методу моментов (ММ-оценки) и методу максимального правдоподобия (ММП-оценки), теряют свойство состоятельности. Предложен также метод построения состоятельных оценок (МСО-оценок) параметров ББР для случая, когда $\varepsilon_0, \varepsilon_1$ априорно известны:

$$\alpha(\varepsilon_0, \varepsilon_1) = \frac{(m_1^* - \varepsilon_0 n)(m_1^* n - m_2^* + \varepsilon_0(n-1)(m_1^* - (1-\varepsilon_1)n) - m_1^* \varepsilon_1(n-1))}{(1-\varepsilon_0-\varepsilon_1)(m_2^* n - m_1^* n - m_1^{*2}(n-1))}, \quad (2)$$

$$\beta(\varepsilon_0, \varepsilon_1) = \frac{(m_1^* - (1-\varepsilon_1)n)(m_2^* + m_1^* \varepsilon_1(n-1) - m_1^* n - \varepsilon_0(n-1)(m_1^* - n(1-\varepsilon_1)))}{(1-\varepsilon_0-\varepsilon_1)(m_2^* n - m_1^* n - m_1^{*2}(n-1))}, \quad (3)$$

где m_1^*, m_2^* – соответственно первый и второй выборочные моменты.

Рассмотрим теперь наиболее общий случай, когда неизвестными являются параметры α, β и уровни искажений $\varepsilon_0, \varepsilon_1$. Пусть по выборке X вычислены пять первых выборочных начальных моментов $m_1^*, m_2^*, m_3^*, m_4^*, m_5^*$. Можно показать, что при условии, когда первые два начальных момента фиксированы, задача оценивания методом моментов сводится к решению системы двух нелинейных уравнений вида:

$$m_3^* = m_3(\alpha(\varepsilon_0, \varepsilon_1), \beta(\varepsilon_0, \varepsilon_1), \varepsilon_0, \varepsilon_1), \quad m_4^* = m_4(\alpha(\varepsilon_0, \varepsilon_1), \beta(\varepsilon_0, \varepsilon_1), \varepsilon_0, \varepsilon_1), \quad (4)$$

где $m_3(\cdot), m_4(\cdot)$ – это теоретические моменты для искаженного ББР, а $\alpha(\varepsilon_0, \varepsilon_1), \beta(\varepsilon_0, \varepsilon_1)$ – это оценки, построенные при заданных уровнях искажений по формулам (2), (3). Для решения этой системы применим модифицированный метод Ньютона. Обозначим матрицу якобиана системы (4) при условии, что первые два начальных момента фиксированы через J_0^{cond} . Тогда итерационная процедура для решения системы (4) имеет следующий вид [13]:

$$\begin{pmatrix} \varepsilon_0^{k+1} \\ \varepsilon_1^{k+1} \end{pmatrix} = \begin{pmatrix} \varepsilon_0^k \\ \varepsilon_1^k \end{pmatrix} + \lambda \cdot (J_0^{cond})^{-1} \cdot \begin{pmatrix} m_3^* - m_3(\alpha(\varepsilon_0^k, \varepsilon_1^k), \beta(\varepsilon_0^k, \varepsilon_1^k), \varepsilon_0^k, \varepsilon_1^k) \\ m_4^* - m_4(\alpha(\varepsilon_0^k, \varepsilon_1^k), \beta(\varepsilon_0^k, \varepsilon_1^k), \varepsilon_0^k, \varepsilon_1^k) \end{pmatrix}, \quad (5)$$

где λ – параметр из промежутка $[0, 1]$, обеспечивающий сходимость при больших уровнях искажений $\varepsilon_0, \varepsilon_1$. Получен явный вид матрицы

J_0^{cond} , который в силу громоздкости выражений не приводится. В результате получаем следующий алгоритм оценивания.

Алгоритм. Вычисление МСО-оценок α , β , ϵ_0 , ϵ_1 методом моментов.

Вход: Начальные выборочные моменты m_1^* , m_2^* , m_3^* , m_4^* , m_5^* .

Выход: МСО-оценки параметров α , β и уровней искажений ϵ_0 , ϵ_1 .

Шаг 1. Инициализация: полагаем $\epsilon_0 = 0$, $\epsilon_1 = 0$.

Шаг 2. Для текущих ϵ_0 , ϵ_1 по формулам (2), (3) вычисляем оценки α , β .

Шаг 3. Для найденных значений по полученным формулам вычисляем J_0^{cond} .

Шаг 4. Методом равномерного поиска находим наилучшее приближение для оценок ϵ_0 , ϵ_1 по принципу минимума нормы вектора невязки для третьего и четвертого моментов m_3^* , m_4^* .

Шаг 5. В случае нескольких вариантов для начального приближения оценок ϵ_0 , ϵ_1 разрешаем неоднозначность по принципу наилучшей аппроксимации пятого начального момента m_5^* .

Шаг 6. Используя итерационную процедуру (5), находим новые оценки ϵ_0 , ϵ_1 .

Шаг 7. Проверяем stop-критерий (для нормы вектора невязок); если он не выполняется, то переходим на шаг 2, в противном случае заканчиваем работу алгоритма.

3. Компьютерное моделирование

Для иллюстрации полученных теоретических результатов была проведена серия компьютерных экспериментов. При этом в качестве истинных значений параметров бета-биномиального распределения были выбраны следующие значения: $\alpha_0 = 0,5$, $\beta_0 = 9,5$, $n = 10$, а в качестве уровней искажений следующие: $\epsilon_0 = \epsilon_1 = 0,01$. В процессе эксперимента генерировались $L = 100$ случайных выборок объема $K = 1000$ реализаций ББР, затем производилось искажение каждой выборки и сравнивались смещение и среднеквадратичное отклонение для ММ-оценок, ММП-оценок и МСО-оценок. Как следует из таблицы, предложенный алгоритм позволяет существенно уменьшить смещение оценок параметров ББР.

Таблица

**Сравнение метода совместного оценивания параметров распределения
и уровней искажений с классическими методами**

Параметр	α (истинное значение 0.5)			β (истинное значение 9.5)		
Метод	ММ	ММП	МСО	ММ	ММП	МСО
Среднее	0.73	0.76	0.49	11.81	12.28	9.20
Среднеквадратичное отклонение	0.12	0.11	0.21	2.01	1.97	2.63

Заключение

Данные исследования были частично поддержаны грантом БГУ для аспирантов и студентов 680/30.

Литература

1. *Green J.D.* Personal Media Probabilities // *Journal of Advertising Research*. 1970. Vol. 10. P. 12–18.
2. *Jonson N.L., Kotz S., Kemp A.W.* Univariate Discrete Distributions. New York, 1996.
3. *Pearson E.S.* Bayes' Theorem in the Light of Experimental Sampling // *Biometrika*. 1925. Vol. 17. P. 338–442.
4. *Skellam J.G.* A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as a Variable between the Sets of Trials // *Journal of the Royal Statistical Society*. 1948. Vol. 10. P. 257–261.
5. *Lee J.C., Sabavala D.J.* Bayesian Estimation for the Beta-Binomial Model // *Journal of Business & Economic Statistics*. 1987. Vol. 5. P. 357–367.
6. *Smith D.M.* Maximum Likelihood Estimation of the Parameters of the Beta-Binomial Distribution // *Journal of the Royal Statistical Society*. 1983. Vol. 24. P. 217–229.
7. *Tripathi R.C., Gupta R.C., Gurland J.* Estimation of parameters in the beta binomial model // *Ann. Inst. Statist. Math.* 1994. Vol. 46. P. 317–331.
8. *Yamamoto E, Yanagimoto T.* Moment estimators for the beta-binomial distribution // *Journal of Applied Statistics*. 1992. Vol. 19. P. 273–283.
9. *Griffiths D.A.* Maximum Likelihood Estimation for the Beta Binomial Distribution and an Application to the Household Distribution of the Total number of Cases of a Disease // *Biometrics*. 1973. Vol. 29. P. 673–648.
10. *Хьюбер П.* . Робастность в статистике. М.: Мир, 1984.
11. *Хампель Ф.* . Робастность в статистике. Подход на основе функции влияния. М.: Мир, 1984.
12. *Kharin Yu.* Robustness in Statistical Pattern Recognition. Dordrecht, 1996.
13. *Демидович Б.П., Марон. И.А.* Основы вычислительной математики. М.: Мир, 1970.