О равномерной состоятельности непараметрического критерия Неймана

Ермаков М. С., д. ф.-м. н., профессор кафедры статистического моделирования СПбГУ, erm2512@mail.ru[1]

Капаца Д. Ю., студент-магистр кафедры статистического моделирования СПбГУ, david@kapatsa.com

Аннотация

Для задачи проверки гипотезы согласия в работе исследуется равномерная состоятельность критериев против непараметрических множеств альтернатив, когда тестовой статистикой является бесконечная линейная комбинация квадратов оценок коэффициентов Фурье разложения плотности распределения в ряд Фурье.

Получены условия равномерной состоятельности, близкие к необходимым. Доказана асимптотическая нормальность тестовой статистики при гипотезе или альтернативах.

Введение

Состоятельность непараметрических критериев при параметрическом задании альтернатив довольно хорошо изучена ([2], [15], [16], [18]). С конца восьмидесятых годов прошлого века началось интенсивное исследование задач непараметрической проверки гипотез при непараметрических множествах альтернатив ([12], [9], [10], [11], [4]). До этих публикаций данная проблематика исследовалась только в отдельных работах ([17], [13], [14]).

Изучению скорости сходимости классических непараметрических критериев при этом уделялось немного внимания. В то же время, на тот момент уже были получены определённые результаты о состоятельности критериев против непараметрических множеств альтернатив ([17], [16], [15]).

Следует отметить, что исследования в конце восьмидесятых годов прошлого века начались под влиянием крайне интенсивного в этот период развития теории непараметрического оценивания. В свою очередь, развитие теории непараметрического оценивания базировалась на уже созданные к тому времени теорию апроксимации и в какой-то мере на теорию информации. В результате исследований теория непараметрического оценивания в какомто смысле стала аналогом теории аппроксимации, но с введённой случайной

[1] Исследование поддержано грантом РФФИ 20-01-00273

ошибкой. Как следствие, теория непараметрической проверки гипотез — использующая схожую технику — есть не что иное, как теория проверки гипотез в функциональных пространствах.

Принятие гипотезы носит рекомендательный характер и говорит лишь о том, что данные не сильно противоречат модели. Понятно, что взяв большое количество данных, мы всегда сможем опровергнуть простую статистическую модель. Поэтому важным фактором в выборе критерия является различимость гипотезы и альтернативы в асимптотическом сценарии. Описание наибольших множеств альтернатив —– способных быть «различимыми» непараметрическими критериями — получило развитие в работах ([3], [6], [8], [1]). В них авторы, по сути, показывали состоятельность критериев против альтернатив, изучаемых в рамках метода расстояний (distance method), на основе которого построены базовые статистические непараметрические критерии (в том числе построенный в данной работе).

Метод расстояний Опишем метод расстояний в контексте проверки гипотезы согласия. Пусть дана выборка независимых одинаково распределённых случайных наблюдений $X^{(n)} = X_1, X_2, \dots, X_n$, имеющих функцию распределения $F(x)$, $x \in (0, 1)$. Мы проверяем простую гипотезу

$$\mathbb{H}_0 : F(x) = F_0(x)$$

против альтернатив

$$\mathbb{H}_n : F(x) \neq F_0(x).$$

Для проверки гипотезы мы задаём тестовые статистики $T_n(\hat{F}_n)$, где

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i < x\}}$$

— эмпирическая функция распределения. Здесь $T_n(F) = \bar{T}_n(F - F_0)$ — функционал, заданный на множестве всех функций распределения $\mathcal{F}$. В качестве $\bar{T}_n(F - F_0)$ обычно берётся некоторая псевдометрика, заданная на множестве $\mathcal{F} - \mathcal{F}$ разностей функций распределения.

Например, для критерия Колмогорова

$$T_n(X^{(n)}) = \sup_x |F_0(x) - \hat{F}_n(x)|,$$

критерия фон Мизеса

$$T_n(X^{(n)}) = \int_{-\infty}^{+\infty} (\hat{F}_n(x) - F_0(x))^2 \, dF_0(x).$$

Равномерная состоятельность Ясно, что мы не можем состоятельно проверить гипотезу $\mathbb{H}_0$ для всех возможных функциональных альтернатив. В работах [5], [7], [8] показывалось, что возможно состоятельно проверить гипотезу для таких альтернатив F, для которых $T(F) > \rho_n$, где $\rho_n \to 0$ специальным образом при $n \to \infty$.

Равномерная состоятельность в теории проверки гипотез при этом рассматривается обычно в двух наиболее распространённых смыслах. Пусть стоит задача проверки гипотезы

$$\mathbb{H}_0 : \; F = F_0 \tag{1}$$

против альтернатив

$$\mathbb{H}_n : \; F_n \in \mathcal{F}_n \subset \mathcal{F}. \tag{2}$$

Для критерия K_n проверки гипотезы обозначим $\alpha(K_n) := \mathbf{E}_{F_0} K_n$ — вероятность ошибки первого рода и $\beta(K_n, F_n) := \mathbf{E}_{F_0}(1 - K_n)$ — вероятность ошибки второго рода при альтернативе $F_n \in \mathcal{F}_n$.

Скажем, что последовательность критериев K_n, где $0 < \alpha(K_n) = \alpha < 1$, *слабо равномерно состоятельна*, если

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}_n} \beta(K_n, F) = 0,$$

и *равномерно состоятельна*, если

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}_n} \beta(K_n, F) < 1 - \alpha. \tag{3}$$

Слабая равномерная состоятельность непараметрических критериев довольно хорошо изучена ([16], [10], [11]). Равномерная состоятельность (3) изучается по настоящее время; её изучению для случая непараметрического критерия Неймана посвещена настоящая работа.

Приведём одно следствие из равномерной состоятельности. В работе Ле Кама [14] утверждается, что если критерии равномерно состоятельны против множества альтернатив $\mathcal{F}_0 \subset \mathcal{F}$ в следующем смысле: если существует n_0 и критерий K_{n_0}, $0 < \alpha(K_{n_0}) = \alpha < 1$ такой, что

$$\sup_{F\in\mathcal{F}_0} \beta(K_{n_0}, F) < 1 - \alpha,$$

то существуют $c, c_1 > 0$ и последовательность критериев K_n такая, что

$$\sup_{F\in\mathcal{F}_n} \beta(K_n, F) < c_1 \exp\{-Cn\}.$$

Также, если непараметрические критерии равномерно состоятельны для $r_n \to 0$ при $n \to \infty$, то они слабо равномерно состоятельны против множеств альтернатив

$$\mathcal{F}_{1n} = \{F : \bar{T}(F - F_0) > r_n > 0,\, F \in \mathcal{F}\},$$

где $\rho_n / r_n \to \infty$ при $n \to \infty$.

Постановка задачи

Далее предполагается, что рассматриваемые в задаче функции распределения абсолютно непрерывны, а соответствующие им плотности являются ограниченными в пространстве L_2. Поэтому можем рассматривать задачу в терминах плотностей распределения.

Пусть $X_1, \ldots, X_n$ — независимые одинаково распределённые с.в., имеющие неизвестную плотность распределения $p(x)$ в $L_2(\nu)$. Предполагается, что плотность представима в виде

$$p(x) = 1 + f(x),\, f(x) = \sum_{j=1}^{\infty} \theta_j \varphi_j(x) \tag{4}$$

где $p(x)$ — плотность распределения, $\{\varphi_j(x)\}_0^{\infty}$ — ортонормированная система функций в $L_2(\nu)$, ν — вероятностная мера, $\varphi_0(x) = 1$ для всех $x \in \Omega$. Функция $f(x)$ предполагается принадлежащей некоторому компакту $\mathcal{U}$ из L_2.

Требуется проверить гипотезу о равномерности распределения наблюдаемых случайных величин

$$\mathbb{H}_0 :\, p(x) = 1,\, x \in [0, 1], \tag{5}$$

т.е. $\theta_j = 0$ для всех j против **альтернатив**

$$\mathbb{H}_n :\, f \in \Psi_n(R_n, c), \Psi_n = \{f : R_n(f) > c,\, f \in \mathcal{U}\}\,, \tag{6}$$

где

$$R_n(f) = A_n(\boldsymbol{\theta}) = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^2 \theta_j^2, \tag{7}$$

где $\varkappa_{nj}^2$ — некоторая убывающая последовательность весов, которая удовлетворяет условиям B1–B3 (см. далее). Соответствующие функциям $f \in \Psi_n(R_n, c)$ бесконечномерные векторы $\boldsymbol{\theta}$ условимся относить к множеству Q_n.

Критерии типа Неймана

Последовательность весов $\varkappa_{nj}$ выбирается из условий B1–B3. Введём

$$R_n(f) = A_n(\boldsymbol{\theta}) = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^2 \theta_j^2, \quad A_n = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^4.$$

$f(x) = \sum_{j=1}^{\infty} \theta_j \varphi_j(x)$ — разложение функции $f(x)$ по ортонормированному базису. Определим тестовые статистики

$$T_n(X^{(n)}) = \sum_{j=1}^{\infty} \varkappa_{nj}^2 \left(\sum_{s=1}^{\infty} \varphi_j(X_s) \right)^2. \tag{8}$$

Данная статистика, помимо условий на веса $\varkappa_{nj}^2$, отличается от предложенной в [5] тем, что соответствующие суммы по $\varkappa_{nj}^2$ имеют бесконечное число слагаемых.

Условия на последовательности $\{\varkappa_j^2\}_{j=1}^{\infty}$

B1. Для любого n последовательность $\varkappa_{nj}^2$ является убывающей.

B2. Существуют константы C_1 и C_2 такие, что для любого n

$$C_1 < A_n = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^4 < C_2. \tag{9}$$

B3. Существуют C_1 и $\lambda > 1$, что для любого $\delta > 0$ и n имеет место

$$\varkappa_{[n,(1+\delta)k_n]}^2 < C_1 (1+\delta)^{-\lambda} \varkappa_n^2,$$

где $k_n = \sup \left\{ k : \sum_{j<k} \varkappa_{nj}^2 \leqslant \frac{1}{2} \sum_{j=1}^{\infty} \varkappa_{nj}^2 < \infty \right\}$. Также имеет место конечность сумм $\sum_{j=1}^{\infty} \varkappa_{nj} < C < \infty$.

Для любого события D обозначим за $\chi(D)$ индикатор этого события. Зададим критерии

$$K_n(X^{(n)}) = \chi \left(T_n(X^{(n)}) - n \sum_{j=1}^{k} \varkappa_j^2 > (2A_n)^{1/2} x_{\alpha_n} \right).$$

Задачей исследования является обобщение результатов работы [5] на большее множество критериев[2] и на более широкий класс альтернатив. Таким образом, главным итогом работы является доказательство следующей теоремы.

Теорема 1 (О равномерной состоятельности критериев типа Неймана) *Пусть выполнены условия B1–B3, а также условие ограниченности системы ортонормированных функций $\{\varphi_j\}_{j=1}^{\infty}$. Тогда для последовательности критериев K_n имеют место $\alpha(K_n) = \alpha + o(1)$ и*

$$\beta(K_n, f_n) = \Phi(x_\alpha - R_n(\boldsymbol{\theta}_n)(2A_n)^{-1/2})(1 + o(1))$$

равномерно для всех последовательностей $\boldsymbol{\theta}_n$, для которых выполнено условие ограниченности f.

Из данного результата сразу следует результат о равномерной состоятельности последовательности заданных критериев.

Результаты

Для доказательства теоремы 1 достаточно доказать две следующих леммы.

Лемма 1 (Об асимптотической нормальности) *Пусть $\boldsymbol{\theta} = \boldsymbol{\theta}^{(n)} \in \Psi_n$ и $\mathbf{D}_\theta T_n = O(A_n)$, $A_n(\boldsymbol{\theta}) \asymp A_n$. $A_n(\boldsymbol{\theta}) := n^2 \sum_{j=1}^{\infty} \theta_j^2 \varkappa_{nj}^2$. Тогда P_θ-распределения тестовых статистик*

$$\left(T_n(X^{(n)}) - n \sum_{j=1}^{k} \varkappa_{nj}^2 - A_n(\boldsymbol{\theta}) \right) (2A_n)^{-1/2}$$

асимпотически нормальны.

Доказательство леммы 1 основано на проверке условий теоремы из работы [19] и на оценках, схожих с построенными в следующей лемме.

Лемма 2 *Пусть $\boldsymbol{\theta} = \theta^{(n)} \in \Psi_n$. Тогда*

$$\mathbf{E}_{\boldsymbol{\theta}} T_n = A_n(\boldsymbol{\theta}) + o(A_n(\boldsymbol{\theta})),$$
$$\mathbf{D}_{\boldsymbol{\theta}} T_n = 2A_n + o(A_n^2(\boldsymbol{\theta})).$$

[2]за счёт неявного задания последовательностей весов $\varkappa_{nj}^2$

Заключение

По результатам работы для задачи проверки гипотезы о распределении построены асимптотические оценки для математического ожидания и дисперсии статистики типа Неймана в случае гипотезы и альтернативы, а также проведено доказательство асимптотической нормальности данной статистики при заданных условиях на рассматриваемое пространство альтернатив.

Таким образом, построено новое семейство непараметрических критериев о распределении, обладающее свойством равномерной состоятельности против широких множеств непараметрических альтернатив.

Список литературы

[1] A. R. Barron, Uniformly powerful goodness of fit tests.— Ann. Statist., v.17 (1989), 107-124.

[2] J. Durbin, Distribution Theory for Tests Based on the Sample Distribution function.— Regional Conference Series in Applied Mathematics, v.9 (1973), SIAM, Philadelphia.

[3] J. Horowitz, V. Spokoiny, An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative.— Econometrica, v.69(3) (2001), 599-631.

[4] М. С. Ермаков, Минимаксное обнаружение сигнала в гауссовском белом шуме.— Теория вероятн. и ее примен., в.35(4) (1990), 704-715.

[5] М. С. Ермаков, Минимаксная проверка гипотез о плотности распределения.— Теория вероятн. и ее примен., в.39(5) (1994), 488-512.

[6] М. С. Ермаков, Асимптотическая минимаксность критериев хи-квадрат.— Теория вероятн. и ее примен., в.42(4) (1997), 668-695.

[7] M. S. Ermakov, On asymptotic minimaxity of kernel-based tests.— ESAIM Probab. Stat., v.7 (2003), 279-312.

[8] М. С. Ермаков, Минимаксное обнаружение сигнала в весовом гауссовском белом шуме.— Зап. научн. сем. ПОМИ, в.320 (2004), 54-68.

[9] И. А. Ибрагимов, Р. З. Хасьминский, Об оценке бесконечномерного параметра в гауссовом белом шуме.— Докл. АН СССР, в.236(5) (1977), 1053-1055.

[10] Ю. И. Ингстер, О сравнении минимаксных свойств тестов Колмогорова, ω^2 и χ^2.— Теория вероятн. и ее примен., в.32(2) (1987), 374-378.

[11] Yu. I. Ingster, I. A. Suslina, Nonparametric Goodness-of-fit Testing under Gaussian Models.— Lecture Notes in Statistics, 169, Springer: N.Y., 2002.

[12] E. Gine, R. Nickl, Mathematical Foundation of Infinite–Dimensional Statistical Models.— Cambridge University Press, Cambridge, 2015.

[13] L. Le Cam, L. Schwartz, A necessary and sufficient conditions for the existence of consistent estimates.— Ann. Math. Statist., v.31 (1960), 140-150.

[14] L. Le Cam, Convergence of estimates under dimensionality restrictions.— Ann. Statist., v.1 (1973), 38-53.

[15] М. Кендалл, Ф. Стьюарт, Статистические выводы и связи.— М. : Наука, 1973, 810с.

[16] E. L. Lehmann, Testing Statistical Hypotheses.— Wiley: N.Y., 1986, 604p.

[17] H. Mann, A Wald, On the Choice of the Number of Class Intervals in the Application of Chi-Square Test.— Annals of Mathematical Statistics, v.13 (1942), 306-317.

[18] G. R. Shorack, J. A. Wellner, Empirical Processes with Application to Statistics.— Wiley: N.Y, 1986.

[19] P. Hall, Central limit theorem for integrated square error of multivariate nonparametric density estimators.— J. Multivar. Anal., v.14(1) (1984), 1-16.