Численные исследования статистических методов сравнения двух наборов кривых

Шишкова Полина Олеговна
группа 622
Научный руководитель: доктор физико-математических наук,профессор В.Б. Мелас
Рецензент: доктор технических наук, профессор Ю.Д.Григорьев

Санкт-Петербургский Государственный Университет

12 июня 2019 г.

Постановка задачи

Пусть $Y_{ij}(t) = f_i(t) + \varepsilon_i, i = 1, 2, j = 1, \ldots, n$, где $f_i(t)$ — некоторые кривые, а ε_i — случайные величины. Обозначим через $F_1(t)$ и $F_2(t)$ функции распределения случайных величин $Y_1(t)$ и $Y_2(t)$ соответственно. Будем проверять гипотезу

$$H_0 : F_1(t) = F_2(t), t \in \{t_1, t_2, \ldots, t_k\} \tag{1}$$

против альтернативной гипотезы

$$H_1 : F_1(t) \neq F_2(t), t \in \{t_1, t_2, \ldots, t_k\}. \tag{2}$$

Краткий обзор литературы

- Перестановочные тесты были исследованы в работе [James and Sood 2006].
- В недавней работе [Monika Sirski 2012] исследование продолжено, предложены тесты, основанные на попарном сравнении элементов двух наборов.

Перестановочные критерии. Описание

Для произвольных кривых $g_1(t)$ и $g_2(t)$, заданных на интервале $t \in [a, b]$ введем расстояние δ_1 между кривыми как

$$\delta_1(g_1, g_2) = \int_a^b |g_1(t) - g_2(t)| dt. \quad (3)$$

Перестановочные критерии. Описание. Продолжение

Для наборов Y_1, Y_2 размера n введем расстояние

$$\overline{T}_1 = \frac{1}{n^2} \sum_{l=1}^{n} \sum_{m=1}^{n} \delta_1(Y_{1l}, Y_{2m}), \tag{4}$$

где Y_{1l} и Y_{2m} — кривые из первого и второго наборов соответственно.

Перестановочные критерии. Описание. Продолжение

Будем менять местами кривые из двух наборов, по $1, 2, \ldots, n-1$ кривых. Для каждой перестановки $b = 1, 2, \ldots, B$ рассчитываем $\overline{T}_1(b)$ согласно (4). Вычисляем величину γ

$$\gamma = \frac{1}{B} \sum_{b=1}^{B} \mathcal{I}(\overline{T}_1(b) \geq \overline{T}_1). \tag{5}$$

Рассматриваемые метрики

В качестве $\delta(g_1, g_2)$ будем выбирать следующие метрики:

$$\delta_1(g_1, g_2) = \int_a^b |g_1(t) - g_2(t)| dt,$$

$$\delta_2(g_1, g_2) = \int_a^b (g_1(t) - g_2(t))^2 \, dt,$$

$$\delta_3(g_1, g_2) = \int_a^b \ln(1 + |g_1(t) - g_2(t)|) \, dt.$$

Перестановочный T-test, описание

- В каждой точке вычислим

$$T(t) = \frac{|\overline{Y}_1(t) - \overline{Y}_2(t)|}{\sqrt{\frac{1}{n_1}\mathrm{Var}(Y_1(t)) + \frac{1}{n_2}\mathrm{Var}(Y_2(t))}}. \quad (6)$$

- Пусть $\overline{Y}_1(t), \overline{Y}_2(t)$ — средние значения для наборов в каждой точке

$$\overline{Y}_1(t) = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i}(t), \overline{Y}_2(t) = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{2i}(t),$$

где Y_{1i} — кривые из первого набора, $i = 1, 2, \ldots, n_1$, а Y_{2i} — кривые из второго набора, $i = 1, 2, \ldots, n_2$.

Перестановочный T-test, описание

- Получаем вектор $Z = (T(t_1), T(t_2), \ldots, T(t_k)$. Критерий — максимальный элемент вектора Z, T_{max}.
- Переставляем кривые из набора в набор. После каждой перестановки получаем вектор $Z_j, j = 1, 2, \ldots, N$, считаем $T_{max,j}$.

Эксперименты. Случай констант

Проведены следующие эксперименты (число повторений — 500):

- $f_1(t) = 2$ и $f_2(t) = 2 + i$, $t_1, t_2, \ldots, t_k \in [0, 1]$, $k = 100$, $k = 20$, ошибки $\varepsilon_j \sim N(0, \sigma_j^2), j = 1, 2$;
- $f_1(t) = 2$ и $f_2(t) = 2 + i$, $t_1, t_2, \ldots, t_k \in [0, 1]$, $k = 100$, ошибки из смеси нормального и Коши $\varepsilon_1 = \varepsilon_2 = 0.9N(0, 1) + 0.1K(0, 1)$;
- $f_1(t) = 2$ и $f_2(t) = 2 + i$, $t_1, t_2, \ldots, t_k \in [0, 1]$, $k = 100$, ошибки из смеси нормального и Коши распределений $\varepsilon_1 = 0.95N(0, 1) + 0.05K(0, (0.5)^2)$, $\varepsilon_2 = 0.95N(0, 1) + 0.05K(0, 1)$.

Пример численных результатов. Константы

Таблица: Мощности критериев. Базовые функции — константы, независимые ошибки из смеси распределений
$\varepsilon_1 \sim 0.95N(0,1) + 0.05K(0,(0.5)^2), \varepsilon_2 \sim 0.95N(0,1) + 0.05K(0,1)$

$2+i$	$\delta_1(f_1,f_2)$	$\delta_2(f_1,f_2)$	$\delta_3(f_1,f_2)$
2	0.116	0.117	0.12
2.01	0.171	0.16	0.178
2.015	0.21	0.25	0.29
2.02	0.33	0.31	0.39
2.025	0.45	0.43	0.51
2.03	0.69	0.67	0.73
2.035	0.84	0.72	0.88
2.04	0.98	0.89	0.99

Эксперименты. Случай полиномов

- Полином второй степени $f_1(t) = t^2 + 2t + 1$, $f_2(t) = c \cdot t^2 + 2t + 1$, $t \in [-4, 4]$, ошибки $\varepsilon_i \sim N(0, \sigma_i^2)$.
- Тригонометрический полином $f_1(t) = \cos(t) + \sin(t)$, $f_2(t) = c \cdot \cos(t) + \sin(t)$, $t \in [0, 2\pi]$, ошибки $\varepsilon_i \sim N(0, \sigma_i^2)$.

Пример численных результатов. Полиномы

Таблица: Мощности критериев. Базовые функции — полиномы второй степени, независимые нормально распределенные ошибки, $\varepsilon_1 \sim N(0, (0.5)^2), \varepsilon_2 \sim N(0, 1)$

c	$\delta_1(f_1, f_2)$	$\delta_2(f_1, f_2)$	$\delta_3(f_1, f_2)$
0.7	1	1	1
0.75	0.93	0.97	0.99
0.8	0.9	0.92	0.95
0.85	0.79	0.83	0.87
0.9	0.52	0.54	0.66
0.95	0.36	0.34	0.41
1	0.18	0.21	0.24

Эксперименты. Случай плотности Бета-распределений

- Пусть $f_1(t)$ — плотность Бета-распределения при $t \in [0, 1]$, ε — авторегрессионная ошибка.
- $f_2(t)$ — смесь двух Бета-распределений в соотношении $1 - p$ (первое распределение) и p(второе распределение).
- Повторим то же самое, однако ошибки будем считать независимыми.

Примеры численных результатов

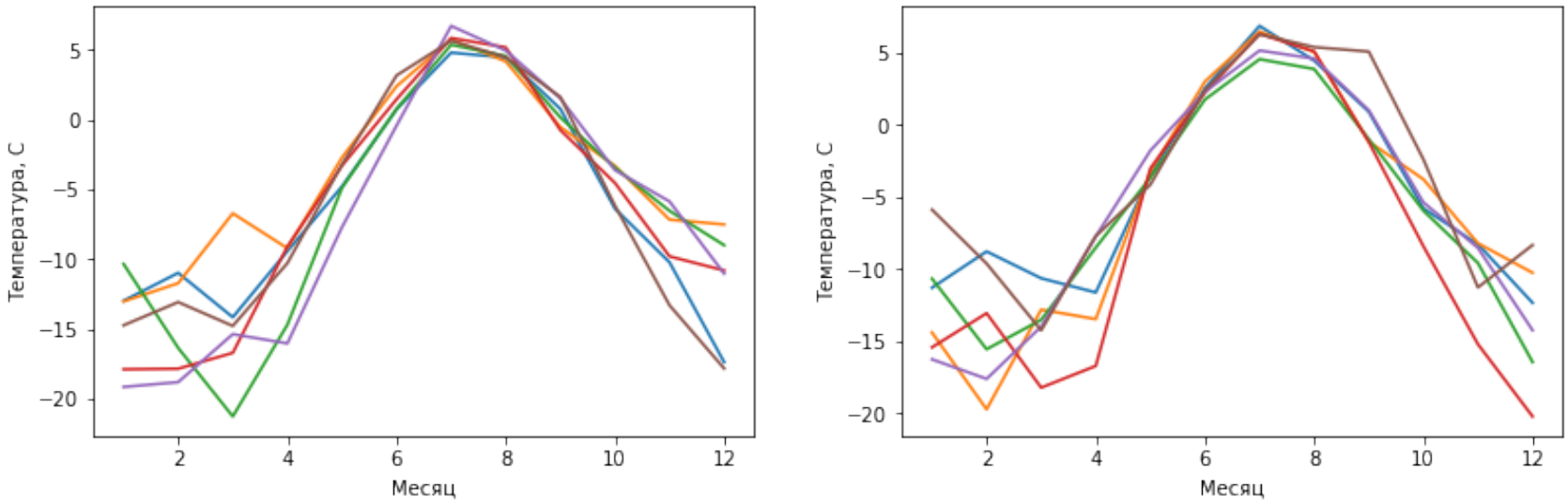
	$p = 0$	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
$\delta_1(f_1, f_2)$	0.05	0.16	0.17	0.484	0.778	0.918
$\delta_2(f_1, f_2)$	0.05	0.14	0.15	0.4	0.712	0.904
$\delta_3(f_1, f_2)$	0.048	0.12	0.19	0.504	0.784	0.912
$t - test$	0.044	0.049	0.108	0.329	0.629	0.801
	$p = 0.5$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$	$p = 1$
$\delta_1(f_1, f_2)$	0.918	0.941	0.986	0.989	0.994	1
$\delta_2(f_1, f_2)$	0.904	0.935	0.978	0.982	0.988	1
$\delta_3(f_1, f_2)$	0.912	0.952	0.976	0.976	0.99	1
$t - test$	0.801	0.817	0.821	0.853	0.872	0.881

Рис.: Мощности критериев при разных значениях p, авторегрессионные ошибки

Применение критериев к реальным данным

- Рассмотрены данные о среднемесячной температуре на острове Шпицберген с двух метеорологических станций на расстоянии 1 км друг от друга за 6 лет. Предоставлены НИИ Арктики и Антарктики.
- Проверяется гипотеза 1.

Реальные данные



а) б)

Рис.: Кривые среднемесячных температур на острове Шпицберген, измеренные на первой станции а) и на второй станции б).

Применение критериев к реальным данным, результаты

Таблица: % перестановок для каждого из критериев, для которых $\bar{T}_i(b) \geq \overline{T}$

	δ_1	δ_2	δ_3	$t-test$
%	$7,2$	$7,3$	$9,6$	$5,4$

Мы отвергаем нулевую гипотезу на уровне значимости 5%

Результаты

Почти для всех типов кривых, которые были рассмотрены, в случае, когда ошибки взяты из

- одного и того же нормального распределения, наиболее мощным оказывается критерий δ_2
- нормальных распределений с разными дисперсиями или из разных смесей распределений Коши и нормального, наиболее мощным оказывается критерий δ_3
- одной и той же смеси распределений Коши и нормального, лучшим оказывается критерий δ_1

Заключение

В рамках работы было сделано:

- сравнение уже исследованных критериев между собой на разных типах кривых
- исследование нового критерия
- критерии применены к реальным данным, сделан практически значимый вывод

Результаты эксперимента показали, что новый критерий в ряде случаев является наиболее мощным.