САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
математико-механический факультет
кафедра статистического моделирования

Андреев Роман Валерьевич
студент 322 группы

ОТЧЕТ ПО КУРСОВОЙ РАБОТЕ

на тему

"Дискриминация функциональных моделей методами перестановок"

Руководитель _____
проф. Мелас В.Б.

Санкт-Петербург
25 декабря 2012 г.

Введение

В данной работе рассматриваются различные методы сравнения групп кривых с целью выяснить, являются ли они различными. Одним из основных методов сравнения является метод перестановок.

Метод перестановок

Пусть нам даны две группы функций $\{f_i\}_{i=1}^N$ и $\{g_j\}_{j=1}^M$, полученных тем или иным способом. Для такого разбиения введем специальным образом функцию оценки разбиения функций на группы $G(\{f_i\}, \{g_j\})$. Чем меньше ее значение, тем «более плотно» сгруппированы кривые внутри групп.

Мы перебираем все разбиения множества из $N + M$ элементов на два множества размерами N и M и считаем для них значения нашей функции разбиения. В итоге мы получаем следущее $P-$значение:

$$P = \frac{1}{\binom{N+M}{N}} \sum_{A,B:|A|=N,|B|=M,A\cup B=\{f_i\}\cup\{g_j\}} I(G(A,B) > G(\{f_i\},\{g_j\})).$$

Обычно результат считается положительным, если $P < 5\%$, иногда 10%. В нашем случае положительный результат означает, что гипотеза о том, что это две разные группы, подтверждается.

На самом деле можно перебирать меньше значений, если $N = M$, ведь мы каждое разбиение переберем два раза. По-сути мы зафиксировали первый элемент в первой группе и перебираем разбиения оставшихся $2N - 1$ на $N - 1$ и N.

Функции разбиения

Для начала выберем функцию расстояния между кривыми. Есть три стандартных расстояния:

- В L_1 : $\rho_1(f,g) = \int |f(x) - g(x)|dx$
- В L_2 : $\rho_1(f,g) = \int (f(x) - g(x))^2 dx$
- В C : $\rho_\infty(f,g) = \sup |f(x) - g(x)|dx$

Но на практике считать расстояния в этих метриках слишком трудная вычислительная задача, а также нам данные не всегда даны в виде функций. Точнее наоборот, обычно сначала делаются измерения в некоторых точках, а потом результаты приближаются кривой. Так что мы заменим интегралы суммами по наблюдениям. Если у нас изначально все же была функция, то мы можем взять выборку из точек на нужном отрезке рассмотрения, например взять некоторое число точек с одинаковым шагом.

Можно выбирать функцию разбиения различными способами:

1. $G(A,B) = -\sum\limits_{f\in A}\sum\limits_{g\in B}\rho(f,g)$,

   что эквивалентно

   $G(A,B) = \sum\limits_{f_1\neq f_2\in A}\rho(f_1,f_2) + \sum\limits_{g_1\neq g_2\in B}\rho(g_1,g_2)$

2. Сначала для каждой группы подсчитаем «среднюю» кривую. Например среднее арифметическое $M_A(x) = \frac{1}{|A|} \sum\limits_{f \in A} f(x)$, или медиану.

   $G(A, B) = \sum\limits_{f \in A} \rho(f, M_A) + \sum\limits_{g \in B} \rho(g, M_B)$

3. Если еще упростить, то получим просто

   $G(A, B) = -\rho(M_A, M_B)$

Как мы видим, в данном методе есть огромное множество вариаций. Можно придумать еще много различных функций сравнения. В каждой задаче возможно будет более естественно использовать ту или иную комбинацию.

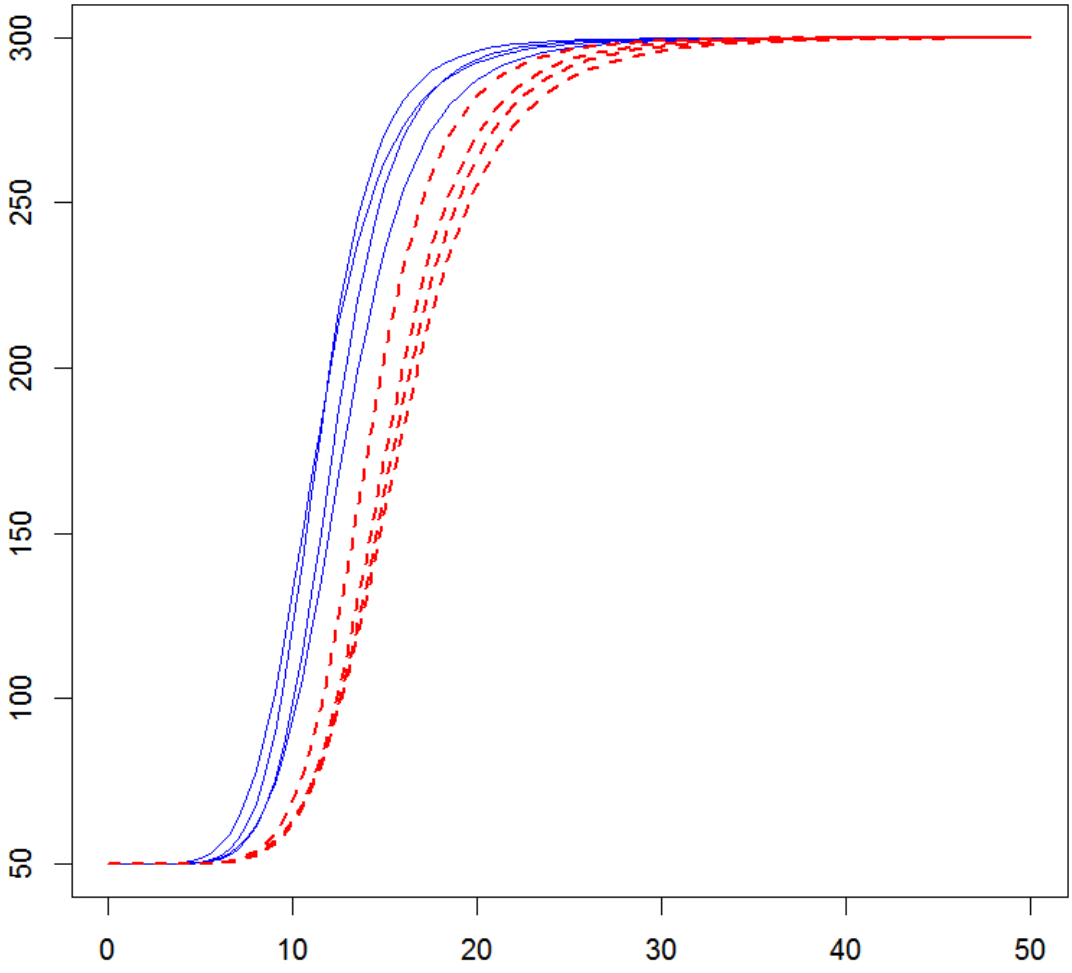
Рассматриваемые кривые

Кривые могут быть получены как из реальных задач, так и искусственно сгенерированные. Для начала были рассмотрены четырехпараметрические логистические функции следующего вида:

$$f(x, a, b, c, d) = a + \frac{b - a}{1 + \left(\frac{x}{c}\right)^{-2-d}}$$

$a_1 = a_2 = 50, b_1 = b_2 = 300, c_1 \in U(11, 13), c_2 \in U(11, 13) + \Delta, d_1, d_2 \in U(4, 6),$

где Δ - параметр сдига.



Проделанная работа

Была написана программа на языке R, реализующая метод перестановок для данных функций.

В следующей таблице приводятся результаты сравнения трех функций сравнения в метрике L_1 для описанных выше функций. В каждой из групп было по 4 функции. Данные в таблице усреднены по 10 экспериментам.

Δ	1	2	3
0.000000	0.494286	0.520000	0.380000
0.500000	0.265714	0.205714	0.257143
1.000000	0.111429	0.080000	0.148571
1.500000	0.020000	0.014286	0.031429
2.000000	0.000000	0.000000	0.000000

Из данной таблицы можно сделать вывод о том, что третий метод работает хуже, чем второй. Это логично, так как третий является в каком-то смысле упрощенной версией второго.

Дальнейшая работа

В дальнейшем планируется добавить к функциям одинаково нормально распределенные независимые ошибки, как бы имитируя эксперимент. Затем применить к этим функциям методы устранения ошибок и сглаживания и посмотреть, как эти действия влияют на результаты.

Список литературы

[1] Monica Sirski. *On the Statistical Analysis of Functional Data Arising from Designed Experiments.* Department of Statistics University of Manitoba, 2012.

[2] Joseph Sturino, Ivan Zorych, Bani Mallick, Karina Pokusaeva, Ying-Ying Chang, Raymond J. Carroll, and Nikolay Bliznuyk. *Statistical Methods for Comparative Phenomics Using High-Throughput Phenotype Microarrays.* The International Journal of Biostatistics, 6(1):Article 29, 2010.