7 Факультатив. Корреляция около кривой и MIC

7.1 Корреляция около кривой

Как мы уже говорили, коэффициент корреляции хорошо измеряет наблюдения, распределенные вдоль некоторой прямой. Введем аналогичные понятия, связанные с распределением вдоль кривой. Вектор (X, Y) назовем распределенным вдоль параметрической кривой $(x, y) = c(s)$, где s — натуральный параметр ($|c'(s)| = 1$), если $(X, Y) = c(S) + Td(S)$ для некоторого вектора (S, T), где $d(s)$ — единичный нормальный вектор к кривой в точке $c(s)$. При этом будем требовать: а) $\mathbf{E}(T|S) = 0$ п.н. б) $\mathbf{D}(T|S) \leq \mathbf{D}S$ п.н.

Вопрос 1. Найти распределения S, $T|S$ для выборки

- равномерно распределенной на квадрате с вершинами $(1, 1)$, $(1, -1)$, $(-1, -1)$, $(-1, 1)$ для $c(t) = (t, 0)$.
- равномерно распределенной на квадрате с вершинами $(1, 0), (0, 1), (-1, 0), (0, -1)$ для $c(t) = (t, 0)$ и для $c(t) = (t, t)/\sqrt{2}$.
- (S, T) для выборки, равномерно распределенной на кольце $\sqrt{X^2 + Y^2} \in [0.5, 1]$ для $c(t) = R(\cos at, \sin at)$. Какими должны быть R, a?

Положим $\alpha(t)$ — угол между касательной к $c(t)$ и осью абцисс и введем локальные дисперсии и ковариацию:

$$DLoc_X(s) = \mathbf{D}S \cos^2 \alpha(s) + \mathbf{D}(T|S = s) \sin^2 \alpha(s),\ DLoc_Y(s) = \mathbf{D}S \sin^2 \alpha(s) + \mathbf{D}(T|S = s) \cos^2 \alpha(s), \quad (1)$$

$$covLoc_{(X,Y)}(s) = (\mathbf{D}S - \mathbf{D}(T|S = s)) \cos \alpha(s) \sin \alpha(s),\ corrLoc_{(X,Y)}(s) = \frac{covLoc_{(X,Y)}}{\sqrt{DLoc_X DLoc_Y}}. \quad (2)$$

В частности, если $c(t)$ — прямая, а $D(T|S = s)$ не зависит от s, то $DLoc_X = DX$, $DLoc_Y = DY$, $covLoc_{X,Y} = cov(X, Y)$.
Ковариацией и корреляцией около кривой называют

$$covGC(X, Y) = \sqrt{\mathbf{E}(covLoc_{X,Y}(S))^2},\ corrGC(X, Y) = \sqrt{\mathbf{E}(corrLoc_{X,Y}(S))^2}.$$

Вопрос 2. Найти ковариацию (X, Y) для равномерного распределения на круге радиуса 3/2 с $c(t) = (\cos t, \sin t)$.

Для независимых величин эти показатели оказываются нулевыми, зато даже для тех видов зависимости, для которых ковариация нулевая (например, для примера в)), мы можем ее отследить. Алгоритм построения кривой c(t) называют principle curve fitting. В некотором смысле этот подход близок к методу главных компонент, но здесь выделяется не главная прямая, а главная кривая. Программа на R, вычисляющая covGC и corrGC, находится в приложенном файле CovrGC. Аналога на Python у меня нет, однако, вы можете запустить этот скрипт из под Python, используя пакет rpy2.

Упражнение 1. Исследовать метод на (X, Y), распределенных на кольце $\sqrt{X^2 + Y^2} \in [a, b]$, а) $a = 0.5$, $b = 1$, б) $a = 0.75$, $b = 1$.

7.2 Maximal Information Coefficient

Максимальный коэффициент информации (MIC, Reshef and co, 2011) был предложен рядом авторов в 2011 году и показывает достаточно хорошие результаты при тестировании.

Общая идея подхода Решефа заключается в следующем. Рассмотрим множество $\mathcal{G}_{k,l}$ решеток (то есть прямоугольных разбиений плоскости на x частей по вертикали и y по горизонтали). Каждая решетка G задает разбиение выборки, для которых мы можем подсчитать частоты попадания $\nu_{i,j}/N$ в каждый из промежутков, заданных решеткой. Для полученных частот ищется взаимная информация, по существу задающая все тот же критерий отношения правдоподобий для нашей дискретной модели

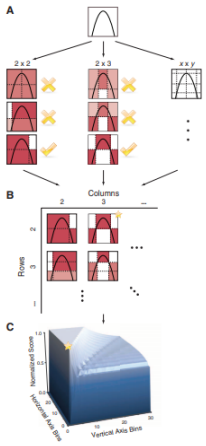
$$I_G = \sum_{i,j} \frac{\nu_{i,j}}{N} \ln \frac{\nu_{i,j} N}{\nu_{i,\cdot} \nu_{\cdot,j}}.$$

Пусть

$$I^*(k,l) = \max_{G \in \mathcal{G}_{k,l}} I_G(x,y), \quad MIC = \max_{kl<g(n)} \frac{I^*(kl)}{\log_2 \min(k,l)},$$

где $g(n) = o(n)$ — некоторая функция размера выборки. В оригинальной статье предлагалось рассматривать $g(n) = n^{0.6}$.

Рис. 1: Иллюстрация из работы Reshef and co, 2011: для каждых размеров выбираем решетку с самой большой статистикой LLR (информацией I), собираем все такие максимумы и строим функцию под аргументом MIC.



Полученная величина устойчива к выбросам (нетрудно понять, что мы смотрим только на ранги наблюдений) и устойчива к замене переменных.

Процедура подсчета MIC достаточно сложна, поскольку нам требуется эффективно осуществить максимизацию по всем решеткам. Подробное описание алгоритма и его реализация на Java есть в `https://science.sciencemag.org/content/sci/suppl/2011/12/14/334.6062.1518.DC1/Reshef.SOM.v2.pdf`

В R MIC реализован в cstats пакете minerva в функции, mictools среди прочего подсчитывает p-value. В Python mic реализован в пакете minepy.mine в функции mic.

7.3 Copule Dependence Coefficient

Jiang и Ding (2014) предложили еще один достаточно удачный коэффициент — CDC.

Его идея восходит к следующему. Копулой называют функцию распределения, все маргинальные распределения которого стандартные равномерные.

Замечательная теорема утверждает, что для любого вектора $X_1, \ldots, X_m$ можно найти такую копулу C, что

$$F_{\vec{X}}(\vec{x}) = C(F_{X_1}(x_1), \ldots, F_{X_m}(x_m)).$$

Чтобы оценить копулу C мы можем найти $(U_i, V_i) = (\widehat{F}_X(x_i), \widehat{F}_Y(y_i))$, где $\widehat{F}$ — оценки маргинальных распределений, откуда оценить

$$\widehat{C}(u,v) = \frac{1}{n} \sum_{i=1}^{n} I_{U_i \le u, \ldots, V_i \le v}.$$

В оригинальном алгоритме маргинальные распределения оцениваются как интеграл от ядерной оценки плотности.

При независимости функция $C(x,y)$ близка к xy. Мы можем оценить тем или иным способом расстояние от $C(x,y)$ до xy. Авторы предлагают следующий подход: рассмотрим

$$MCC(U,V) = \max(\rho(f(U), g(V))), \ \ f, g \in L^2(\mathbf{P}),$$

где ρ — коэффициент корреляции Пирсона. Критерия на основе CDC мне неизвестно, но соответствующий коэффициент корреляции реализован в коде, приложенном в файле CDC.R.

7.4 Ответы на вопросы

1. а) Кривая $c(t)$ идет вдоль оси OX, значит нормаль к ней — вдоль OY. Отсюда $S = X$, $T = Y$, S распределено $R[-1,1]$, $\mathbf{P}(T \in \cdot | S = s)$ будет также равномерным $R[-1,1]$ распределением независимо от s.
   б) В первом случае опять же $S = X$, $T = Y$. При этом $\mathbf{P}(S \le x) = x^2/2$, $x \in [0,1/2]$, $\mathbf{P}(S \le x) = 1 - (1-x)^2/2$, $x \in [1/2,1]$. Эти вычисления вытекают из простого подсчета площади множества $X \le x$, поскольку X, Y распределены равномерно. $\mathbf{P}(T \le x | S = s) = (x + 1 - s)/(2(1-s))$ при $x \in [s-1, 1-s]$, поскольку при фиксированном $S = s$ величина T распределена по $[s-1, 1-s]$ равномерно.
   Во втором случае мы попадаем задачу 4а).
   в) Поскольку параметр натуральный, $Ra = 1$. По свойствам математического ожидания $\mathbf{E}(T|S) = 0$. Но $T + R$ при условии S имеет то же распределение, что и расстояние $\tilde{R}$ от точек кольца до центра. Это распределение легко находится, поскольку

$$\mathbf{P}(\tilde{R} \le x) = \frac{x^2 - 1/4}{3/4},\ x \in [1/2, 1].$$

   Значит,

$$\mathbf{E}(T|S) = \frac{8}{3}\int_{1/2}^{1} x^2 dx - R = 7/9 - R,$$

   т.е. $R = 7/9$, $a = 9/7$. S при этом распределен равномерно по $[0, 14\pi/9]$, откуда

$$f_{T|S}(x|s) = \frac{8(x + 7/9)}{3},\ x \in [-5/18, 2/9].$$

2. Имеем

$$covLoc_{X,Y}(s) = -(\mathbf{D}S - \mathbf{D}T)\sin s \cos s,\ covGC(X,Y) = (\mathbf{D}S - \mathbf{D}T)/2\sqrt{\mathbf{E}\sin^2(2S)},$$

   где

$$\mathbf{E}\sin^2(2S) = \frac{1}{2\pi}\int_0^{2\pi} \sin^2(2s) ds = \frac{1}{2},\ \mathbf{D}S = \mathbf{E}S^2 - (\mathbf{E}S)^2 = \frac{\pi^2}{3},\ \mathbf{D}T = \frac{2}{(3/2)^2}\int_0^{3/2} t(t-1)^2 dt = \frac{1}{8}.$$

   Как мы видим, ковариация ненулевая.