

Локальная корреляция как мера зависимости между случайными величинами

Григорий Стерлинг^{1,2}, Павел Приходько^{2,3}

¹ Московский Физико-Технический Институт, Москва,
`sterling@phystech.edu`

² Институт проблем передачи информации им. А.А. Харкевича Российской
академии наук (ИППИ РАН),
Москва, Большой Каретный переулок, д.19 стр. 1.

³ Datadvance, Москва, Покровский бульвар, д.3, стр. 1Б, 4 этаж

Аннотация В данной работе рассмотрен метод расчета локальной корреляции между произвольными случайными величинами. Показана сходимость и асимптотическая несмещенность. Исследованы свойства локальной корреляции в одномерной и многомерной постановках.

Ключевые слова: локальная корреляция, анализ чувствительности, аксиомы Реньи, частичная корреляция

1 Введение

Задача определения меры зависимости одной случайной величины от другой есть одна из актуальных задач математической статистики и анализа данных. В наиболее общем случае требуется построить функцию (или алгоритм), зависящую от случайных величин или выборок, которая будет обладать некоторыми свойствами. Например, равняться единице в случае функциональной зависимости, и нулю - для независимых случайных величин. Также существуют частичная и локальная постановки данной задачи. В первой нужно исследовать случайные величины при известном наборе других случайных величин, от которых, возможно, зависят исходные. Локальная постановка отличается от глобальной тем, что нас интересует зависимость в некоторой окрестности интересующей нас точки.

На данный момент существует большое количество мер, позволяющих в каком-либо смысле оценить зависимость между переменными. Наиболее простая из них - корреляция Пирсона - нормированная на стандартные отклонения величин ковариация между ними, однако, она может детектировать только линейную зависимость, что делает ее непригодной в более общем случае. Данную проблему позволили частично решить другие работы. Например, анализ копул [4], взаимная информация [3], [2] или Distance Correlation [5]. В русскоязычной литературе этот термин обычно не переводится. У всех этих методов есть свои плюсы и недостатки. Например, при

расчете локальной взаимной информации нужно восстанавливать исходное распределение, но не существует достаточно точного метода для этого, особенно в многомерном случае. Distance Correlation есть корреляция Пирсона между всеми попарными расстояниями в выборке, и такой алгоритм будет работать не быстрее, чем за $O(N^2)$ операций, что делает данный метод непригодным в случае больших выборок.

Помимо описанных выше мер существуют, очевидно, и другие. О некоторых из них можно прочесть в [6], [9].

Рассмотрим постановку задачи анализа чувствительности, предложенную А. Реньи [1]. Требуется построить функцию от двух произвольных случайных величин, удовлетворяющую свойствам:

- (A) $|R(X, Y)| \leq 1$.
- (B) $R(X, Y) = 0$ в случае независимых X и Y .
- (C) $R(X, Y) = \pm 1$ когда $Y = f(X)$.
- (D) $R(X, Y) = R(f(X), g(Y))$ для всех монотонных и непрерывных f и g .
- (E) $X_n \rightarrow X, Y_n \rightarrow Y$ тогда $R(X_n, Y_n) \rightarrow R(X, Y)$.
- (F) Если X и Y - нормально распределенные с.в., то $R(X, Y) = q(r)$ где r - корреляция Пирсона между X и Y , а $q(\cdot)$ - непрерывная функция $[-1, 1] \rightarrow [-1, 1]$.

Эти аксиомы являются разумными требованиями к некоторой функции, которая измеряет наличие и силу глобальной зависимости между случайными величинами. В данной работе предлагается рассмотреть локальную постановку данной задачи. Характерной ее особенностью является то, что нас интересует зависимость между случайными величинами на некотором подмножестве их области определения.

2 Локальная корреляция

В литературе локальная корреляция исследована не так широко, как глобальная. Однако, есть несколько работ, описывающих данную проблему. Например, локальный метод на основе взаимной информации описан в [8], а про корреляцию Пирсона на сужении исходного распределения на некоторое множество - в [10].

Рассмотрим случайные величины $X, Y \sim P_{XY}$, $X \in R^p$, $Y \in R$, и iid выборку размера N из них $\{X_i, Y_i\}_{i=1}^N$. Зафиксируем точку $x \in R^p$ и построим функцию $R_j(X, Y|x)$ и ее выборочную оценку $R_j^N(\{X_i, Y_i\}_{i=1}^N|x)$, где индекс j означает номер компоненты. Далее будем обозначать их как R_j и R_j^N соответственно.

Предложенный в настоящей работе метод состоит в замене классических моментов на их взвешенные аналоги, причем, как будет показано, выполнение аксиом не нарушится. Взвешивание производится путем домножения плотности распределения на некоторую ядерную функцию $w(x, x')$, определенную для всех $x, x' \in R^p$, $w(x, x') \geq 0$, $\forall x \in R^p : \int_{-\infty}^{\infty} w(x, x') dx' < \infty$.

Определение 1 Пусть величина $\int_{-\infty}^{\infty} w(x, x') p_{\xi}(x') dx'$ отлична от нуля. Взвешенным с весом $w(x, \cdot)$ k -м моментом случайной величины ξ назовем

$$E_w \xi^k = \frac{\int_{-\infty}^{\infty} w(x, x') p_{\xi}(x') x'^k dx'}{\int_{-\infty}^{\infty} w(x, x') p_{\xi}(x') dx'}. \quad (1)$$

Определение 2 Для случайных величин X и Y , таких, что их взвешенные дисперсии существуют и отличны от нуля, взвешенной с весом $w(x, \cdot)$ корреляцией j -го признака и Y назовем величину

$$R_j = \frac{E_w(X_j - E_w X_j)(Y - E_w Y)}{\sqrt{(E_w X_j^2 - (E_w X_j)^2)(E_w Y^2 - (E_w Y)^2)}}. \quad (2)$$

Например, можно показать, что для гауссовского ядра и невырожденных распределений локальная корреляция R_j существует.

Утверждение 1 Локальная корреляция R_j удовлетворяет аксиомам (A), (B), (E), (F), то есть, тем же, которым удовлетворяет корреляция Пирсона.

Рассмотрим новую случайную величины X'_j , которая задается плотностью:

$$P_{X'_j}(t) = \frac{w(x, t) P_{X_j}(t)}{\int w(x, s) P_{X_j}(s) ds}. \quad (3)$$

Тогда корреляция Пирсона между X'_j и Y в точности совпадает с R_j . В работе Реньи [1] показано, что для произвольных случайных величин указанные в утверждении аксиомы выполняются для обычной корреляции, и, значит, они выполняются и для локальной.

В многомерном случае без шума виден физический смысл предложенной величины. Разумно предположить, что если Y зависит от X_i сильнее в каком-либо смысле, чем от X_j , то и частная производная по X_i должна быть больше. Следующее утверждение показывает связь частных производных и локальной корреляции.

Утверждение 2 $x, X \in R^p, Y = f(X) \in R, \forall i, j \leq p : P_{X_i} = P_{X_j}, \theta_i = \theta_j, w(x, x') = w((\sum_{s=1}^p \frac{|x_s - x'_s|^2}{\theta_s^2})^{0.5})$. Тогда R_j есть отношение частной производной по j -й переменной к величине модуля градиента f .

Доказательство. Используя разложение Тейлора многомерной функции, $f(x') = f(x) + \nabla f(x)(x' - x) + \frac{1}{2}(x' - x)^T J(f(x))(x' - x) + \dots$, где $J(f(x))$ - матрица Якоби, нетрудно показать, что:

$$E_w g(x) = g(x) + \sum_{s=1}^p \theta_s \frac{\partial g}{\partial x_s} A_i(x) + \frac{1}{2} \sum_{s,l=1}^p \theta_s \theta_l A_{sl}(x) + \dots, \quad (4)$$

где t - вектор $\{t_s\}_{s=1}^p$, $A_s(x) = \frac{\int w(t)p_x(x+t\theta)t_s dt}{\int w(t)p_x(x+t\theta)dt}$, $A_{sl}(x) = \frac{\int w(t)p_x(x+t\theta)t_s t_l dt}{\int w(t)p_x(x+t\theta)dt}$. Далее мы воспользуемся тем, что $A_s(x) \sim \theta_s^2$. Это можно увидеть, разложив плотность $p_x(x+t\theta)$ в ряд Тейлора в окрестности x и собрав подобные члены.

$R_j^* = \lim_{\|\theta\| \rightarrow 0} R_j = \frac{\alpha_j \frac{\partial f}{\partial x_j}}{\sqrt{\sum_{s=1}^p \alpha_s^2 (\frac{\partial f}{\partial x_s})^2}}$, где $\alpha_i = \lim_{\|\theta\| \rightarrow 0} \frac{A_{ii}(x) - (A_i(x))^2}{\theta_i^2} = \frac{\int w(t)t_s^2 dt}{\int w(t)dt} - (\frac{\int w(t)t_s dt}{\int w(t)dt})^2$ - не зависящие от функции f коэффициенты. Если выбрать симметричное ядро $w(t) = w(-t)$, то выражение упростится до $\alpha_i = \frac{\int w(t)t_s^2 dt}{\int w(t)dt}$. По условию утверждения, исходное распределение P_X изотропно, а ширины ядра по каждой компоненте совпадают, тогда $\forall i, j, 1 \leq i, j \leq p : \alpha_i = \alpha_j$. Таким образом, $R_j^* = \frac{\partial f / \partial x_j}{\sqrt{\sum_{s=1}^p (\partial f / \partial x_s)^2}}$, что есть отношение производной по выбранной компоненте к величине градиента функции в точке x .

Следствие: $\sum_{j=1}^p R_j^{*2} = 1$, причем, это будет верно для произвольных распределений и весовой функции.

3 Выборочная оценка локальной корреляции

Мы выяснили, что локальная корреляция R_j , посчитанная для случайных величин X и Y - удовлетворяет аксиомам Реньи и имеет разумный физический смысл. Однако, на практике, мы наблюдаем не случайные величины, а их реализации. Поэтому необходимо построить выборочную оценку R_j и исследовать ее свойства.

Определение 3 Аналогично определению (1), выборочная оценка k -го момента произвольной случайной величины:

$$E_w^N \xi^k = \frac{\sum_{i=0}^N w(x, \xi_i) \xi_i^k}{\sum_{i=0}^N w(x, \xi_i)}; \quad (5)$$

а выборочной оценкой R_j назовем величину

$$R_j^N = \frac{E_w^N (X_j - E_w^N X_j)(Y - E_w^N Y)}{\sqrt{(E_w^N X_j^2 - (E_w^N X_j)^2)(E_w^N Y^2 - (E_w^N Y)^2)}}. \quad (6)$$

Данная величина есть корреляция Пирсона для выборки, полученной из случайных величин с плотностью $P'_{X,Y} \propto P_{X,Y} w(x, \cdot)$, поэтому удовлетворяет интересующим нас требованиям. Однако, требуется выяснить, насколько хорошо R_j^N оценивает R_j .

3.1 Предельные соотношения

Проверим, что оценка является асимптотически несмещенной, и сходится по распределению к истинному значению.

Утверждение 3 X, Y - случайные величины, у которых существуют конечные четвертые взвешенные моменты, то есть, $\forall j E_w X_j^4 < \infty, E_w Y^4 < \infty$. x - точка из того же пространства, что и любая реализация X . Тогда R_j^N есть состоятельная, асимптотически несмещенная оценка R_j .

Доказательство.

Для произвольной случайной величины ξ с конечным четвертым взвешенным моментом, оценка взвешенного матожидания $E_w^N \xi$ - состоятельная и асимптотически несмещенная, то есть $E_w^N \xi \xrightarrow{P} E_w \xi$. Это видно из того, что суммы в числителе и знаменателе сходятся к некоторым конечным значениям, а отношение сходящихся по вероятности последовательностей случайных величин сходится к отношению их предельных значений.

В монографии [11] показано, что если для случайной величины β существует конечный четвертый момент, то сумма N реализаций β , отнесенная к N , сходится по вероятности к матожиданию β . Рассмотрим случайные величины $w(x, X), w(x, X)X_j, w(x, X)Y$. По условию утверждения, у них существуют четвертые моменты, поэтому будет выполнено:

$$\frac{1}{N} \sum_{i=1}^N w(x, X_i) \xi_i \xrightarrow{P} \int w(x, x') p_X(x') \xi(x') dx'. \quad (7)$$

Если числитель и знаменатель сходятся по распределению к некоторым конечным величинам, причем предел числителя отличен от нуля, то и отношение сумм будет сходиться к отношению пределов.

$$E_w^N \xi = \frac{\sum_{i=1}^N w(x, X_i) \xi_i}{\sum_{i=1}^N w(x, X_i)} \xrightarrow{P} E_w \xi. \quad (8)$$

Подставив вместо ξ различные моменты, получим, что каждая из выборочных оценок, участвующих в R_j^N сходится по распределению к их истинным значениям. Тогда:

$$R_j^N \xrightarrow{P} \frac{E_w(X_j - E_w X_j)(Y - E_w Y)}{\sqrt{E_w(X_j - E_w X_j)^2 E_w(Y - E_w Y)^2}} = R_j. \quad (9)$$

4 Численные эксперименты

В данном разделе будут продемонстрированы примеры расчета локальной корреляции для различных задач. На самом деле не существует эталонного метода, меряющего локальную зависимость, поэтому предложенную меру не с чем сравнивать. Однако, можно придумать задачи, в которых можно оценить наличие или отсутствие зависимости качественно, но не количественно. Мы ожидаем, что, например, для сильно зашумленных данных мы

получим близкую к нулю оценку зависимости, а для бесшумных - близкую к единице. Чтобы продемонстрировать данный эффект, разумно выбрать задачи с нестационарным шумом и нелинейной зависимостью между случайными величинами.

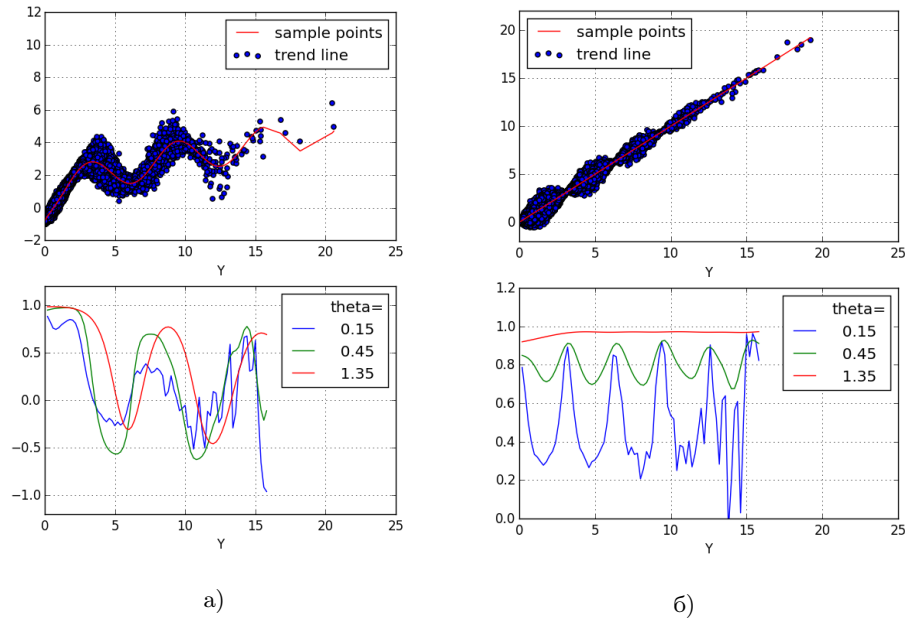


Рис. 1. Поведение R_j^N при различных θ .

На Рис.1а представлен пример расчета локальной корреляции на функции довольно сложного вида, для различных ширин ядра. Выборка получена из экспоненциального распределения с нестационарным шумом. Как видно из рисунка, слишком узкое и слишком широкое ядро в данной задаче дают неправдоподобный результат.

На Рис.1б показана линейная функция с шумом, дисперсия которого периодическая. Видно, что при большом числе точек около интересующей нас точки, R_j разумно меряет зависимость между случайными величинами, но если точек мало, корреляция с узким ядром вырождается и осциллирует.

В общем случае локальная корреляция зависит от весовой функции. Ее разумно выбирать из класса ядерных функций, параметризованных шириной ядра θ , то есть, $w(x, x') = w(\|x - x'\|_\theta)$. Интересно посмотреть, как ведет себя оценка R_j^N при различных θ .

Ниже, на Рис.2, показано, как смещение и дисперсия оценки меняются с ростом ширины окна, для различных размеров выборки, для функции $Y =$

$e^{X_1} + 2e^{X_2}$. Минимум bias-variance tradeoff позволяет найти оптимальную ширину ядра.

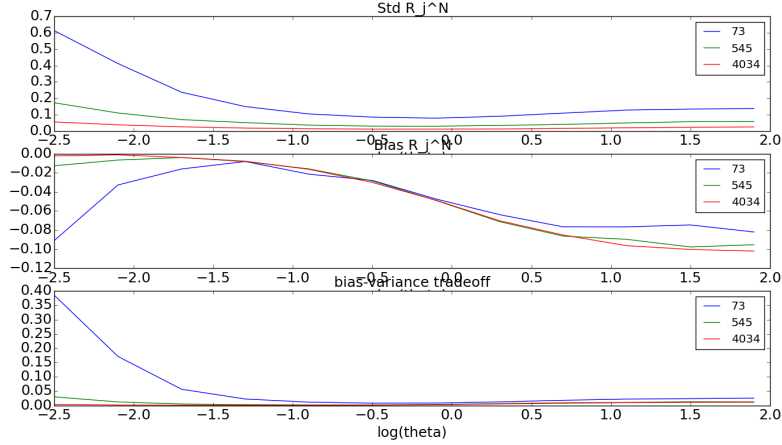


Рис. 2. Сравнение bias и variance для различных ширин ядра и размеров выборки.

5 Заключение

На данный момент открыт вопрос об аналитическом выражении для оптимальной ширины ядра весовой функции для выборочной оценки R_j^N . Если принять за истинное значение R_j^* , то, очевидно, с уменьшением ширины ядра смещение оценки относительно R_j^* будет уменьшаться, а дисперсия оценки - увеличиваться. Аналогично, в ростом ширины ядра $R_j \rightarrow \text{Pearson}(X_j, Y)$, а дисперсия уменьшится до нуля. Так как это происходит непрерывно, можно рассчитывать на то, что существует оптимальное значение θ , минимизирующее bias-variance tradeoff, однако, аналитическое выражение для θ_{opt} остается неизвестным.

Ввиду высокой скорости работы алгоритма расчета локальной корреляции, асимптотически равной $O(N)$, для достаточно больших выборок оптимальное значение ширины ядра можно найти из кросс-валидации, или других методов ресемплинга.

На данный момент среди линейных по сложности не существовало алгоритма, позволяющего детектировать зависимость общего вида. Эта проблема была решена путем добавления весовой функции специального вида во все интегралы, участвующие в определении корреляции Пирсона.

Предложенный алгоритм, на самом деле, зависит от многомерного параметра θ , имеющего смысл ширины ядра для весовой функции. Один из

недостатков состоит в том, что в предельном случае бесконечно малого θ выборочная оценка локальной корреляции для зашумленных данных вырождается в ноль. Это показывает, что чтобы получить информативную оценку, следует выбирать θ в каком-либо смысле оптимальным. Аналитического выражения, позволяющего оценивать оптимальную ширину ядра на данный момент не предложено, и это является одним из направлений будущей работы над локальной корреляцией.

Список литературы

1. A. Renyi, "On measures of dependence Acta Mathematica Academiae Scientiarum Hungarica 1959, Volume 10, Issue 3-4, pp 441-451
2. Mahito Sugiyama and Karsten M. Borgwardt, "Measuring Statistical Dependence via Mutual Information Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Paper 251
3. Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M.; Sabeti, P. C. (2011). "Detecting Novel Associations in Large Data Sets". Science 334 (6062): 1518–1524].
4. Zoubin Ghahramani, Barnabás Póczos and Jeff G. Schneider, "Copula-based Kernel Dependency Measures Proceedings of the 29th International Conference on Machine Learning (ICML-12), 2012, 775–782
5. Székely, Gábor J.; Rizzo, Maria L.; Bakirov, Nail K. "Measuring and testing dependence by correlation of distances". Ann. Statist. 35 (2007), no. 6, 2769-2794
6. O.V. Sarmanov, "The maximum correlation coefficient (symmetric case)"Dokl. Akad. Nauk SSSR , 120 : 4 (1958) pp. 715–718
7. Leo Breiman, Jerome H. Friedman, "Estimating Optimal Transformations for Multiple Regression and Correlation Journal of the American Statistical Association, Volume 80, Issue 391, 1985
8. Dag Tjøstheim, Karl Ove Hufthammer, "Local Gaussian correlation: A new measure of dependence Journal of Econometrics, Volume 172, Issue 1, January 2013, Pages 33-48
9. Pedro Delicado, Marcelo Smrekar, "Measuring non-linear dependence for two random variables distributed along a curve Statistics and Computing, September 2009, Volume 19, Issue 3, pp 255-269
10. Bouma, Gerloff (2009). "Normalized (Pointwise) Mutual Information in Collocation Extraction". Proceedings of the Biennial GSCL Conference.
11. Stirzaker D., Elementary probability, Cambridge University Press, 2003.