Занятие седьмое. О мерах зависимости и корреляционных расстояних, о корреляции около кривой и главных компонентах.

В рамках этого занятия мы будем заниматься вопросами зависимости и независимости. Простейшей мерой выражения зависимости для нас является коэффициент корреляции

$$\rho_{X,Y} = \frac{E(X-EX)(Y-EY)}{\sqrt{DXDY}}.$$

Этот коэффициент $\rho \in [-1,1]$, причем для независимых величин $\rho = 0$, а для линейно зависимых величин (и только для них) $\rho = 1$ или $\rho = -1$. ρ хорошо отражает прямую (линейную зависимость), но не отслеживает более сложной зависимости, скажем, для $X \sim \mathcal{N}(0,1)$, $\rho_{X,X^2} = 0$, хотя величины существенно зависимы. Тем не менее, существенное отличие коэффициента корреляции от нуля, показывает нам наличие зависимости. В связи с этим используют оценку для ρ, имеющую вид

$$\hat{\rho} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{(\overline{X^2} - \overline{X}^2)(\overline{Y^2} - \overline{Y}^2)}}.$$

Мы обсудим эту меру зависимости и ее модификации.

Независимость дискретных выборок

Для проверки независимости двух дискретных выборок данные удобно представить в виде так называемых таблиц сопряженности (contigency table). Если X принимает значения $x_1, ..., x_k$, а Y — $y_1, ..., y_l$, то составим таблицу со столбцами $y_1, ..., y_l$ и строками $x_1, ..., x_k$, где в ячейке (i,j) стоит $p_{i,j} = P(X = x_i, Y = y_j)$. Если гипотеза верна, то $p_{i,j} = p_i q_j$, где $p_i = P(X = x_i)$, $q_j = P(Y = y_j)$.

1) Используем параметрический критерий хи-квадрат на основе частот $\nu_{i,j}$ появления пар (x_i, y_j). Тогда гипотеза заключается в том, что наше $kl - 1$ параметрическое семейство вероятностей $p_{i,j}$ выражается через $k + l - 2$ вероятности p_i, q_j, $i < k$, $j < l$. Функция правдоподобия при выполнении гипотезы имеет вид

$$\prod_{i=1}^{k}\prod_{j=1}^{l}(p_i q_j)^{\nu_{i,j}} = \prod_{i=1}^{k} e^{\ln p_i \sum_{j=1}^{l}\nu_{i,j}} \prod_{j=1}^{l} e^{\ln q_j \sum_{i=1}^{k}\nu_{i,j}}$$

Отсюда ОМП $\hat{p}_i = \nu_{i,\cdot}/N = \sum_{j=1}^{l}\nu_{i,j}/N$, $\hat{q}_j = \sum_{i=1}^{k}\nu_{i,j}/N = \nu_{\cdot,j}/N$.
Следовательно, при выполнении гипотезы

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(\nu_{i,j} - N\hat{p}_i\hat{q}_j)^2}{N\hat{p}_i\hat{q}_j} \sim \chi^2_{(k-1)(l-1)}$$

Отсюда возникает критерий хи-квадрат для проверки независимости, который в R задан все тем же chisq.test.

2) Что даст для нашего случая критерий отношения правдоподобий?

$$2\ln\left(\frac{\sup_{p_{i,j}} L(x,p)}{\sup_{p_{i,q_j}} L(x,p)}\right) = 2\sum_{i,j}\ln\left(\frac{\nu_{i,j}^{\nu_{i,j}}}{(\nu_{i,\cdot}\nu_{\cdot,j}/N)^{\nu_{i,j}}}\right) = 2\sum_{i,j}\nu_{i,j}\ln\left(\frac{\nu_{i,j}N}{\nu_{i,\cdot}\nu_{\cdot,j}}\right) \sim \chi^2_{(k-1)(l-1)}.$$

Упражнение 1. Исследовать критерии на данных survey пакета MASS. Взаимосвязаны ли курение и выполнение домашних задач? Пол и выполнение д.з.?

Те же методы работают для общих распределений за счет группировки данных.

С критериями 1) и 2) связаны коэффициенты, оценивающие взаимосвязь величин X, Y:

1) Коэффициент Крамера (V)

$$V = \frac{\chi^2}{N\min(k-1, l-1)}$$

Этот коэффициент лежит в диапазоне $[0,1]$, причем 0 получается только при $\frac{n_{i,j}}{N} = \frac{n_{i,\cdot}}{N}\frac{n_{\cdot,j}}{N}$, а 1 только в том случае, когда в каждой строке или столбце таблицы сопряженности не более 1 непустого значения.

2) Коэффициент неопределенности или U-коэффициент Тэйла

$$U_{Y|X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\nu_{i,j}\ln\left(\frac{\nu_{i,j}N}{\nu_{i,\cdot}\nu_{\cdot,j}}\right)}{\sum_{i=1}^{k}\nu_{i,\cdot}\ln\left(\frac{\nu_{i,\cdot}}{N}\right)}.$$

В силу его асимметричности, зачастую рассматривают коэффицент

$$U_{Y,X} = \frac{U_{Y|X}H_X + U_{X|Y}H_Y}{H_X + H_Y},$$

где $H_X = \sum_{i=1}^{k}\nu_{i,\cdot}\ln\left(\frac{\nu_{i,\cdot}}{N}\right)$, $H_Y = \sum_{i=1}^{k}\nu_{\cdot,j}\ln\left(\frac{\nu_{\cdot,j}}{N}\right)$.
Этот коэффициент также находится в диапазоне $[0,1]$, причем принимает значения 0 и 1 в тех же случаях, что и V.

Меры связи

Мерами связи (measure of association) называют коэффициент, аналогичные коэффициентам Крамера и Тэйла, показывающие взаимосвязь. Рассмотрим еще несколько таких мер в общем, недискретном случае.
1) Коэффициент корреляции Пирсона ρ.
Оценим коэффициент корреляции выборочной величиной

$$\rho = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{S_X^2 S_Y^2}},$$

где S_X^2, S_Y^2 — смещенные оценка дисперсии $\overline{(X-\overline{X})^2}$, $\overline{(Y-\overline{Y})^2}$. Этот коэффициент лежит в диапазоне $[-1,1]$, причем 1 будет достигаться только при $X = aY+b$ п.н. при некотором $a>0$, -1 при $X = aY+b$ п.н., $a<0$. Теоретический коэффициент корреляции измеряет наличие прямой линейной зависимости, поэтому коэффициент Пирсона также будет принимать большие абсолютные значения именно при линейной зависимости, но хуже будет отслеживать другие виды зависимостей.
Коэффициент Пирсона обычно применяют для двумерных нормальных выборок, для которых он является неплохой мерой зависимости.
2) Коэффициент Спирмена.
Этот коэффициент позволяет уйти от условия линейности зависимости, заменив наблюдения X_i на их ранги R_i в ряду X, а Y_i — на их ранги T_i в ряду Y. Тогда

$$\rho_S = \frac{\overline{RT} - \overline{R}\,\overline{T}}{\sqrt{S_R^2 S_T^2}}.$$

Тогда близость ρ_S по абсолютному значению к 1 означает, что R_i линейно зависят от T_i, т.е. зависимость Y от X монотонна.
3) Коэффициент Кенделла τ.
Назовем две пары значений x_i, y_i, x_j, y_j согласованными, если $x_i - x_j$, $y_i - y_j$ — одного знака. Пусть C — количество согласованных, D — несогласованных пар. Тогда

$$\tau = \frac{C-D}{C+D} = \frac{2(C-D)}{n(n-1)}$$

называют коэффициентом согласия Кенделла или τ-коэффициентом. Он также лежит в диапазоне $[-1,1]$. Этот коэффициент сильно коррелирован с коэффициентом ρ_S.
В R один из способов получения коэффициентов 1)-3) является cor(X,Y,method = c("pearson", "kendall", "spearman")). Методы CramerV, KendallTauB, SpearmanRho, Lambda, UncertCoef пакета DescTools строят коэффициенты вместе с доверительными интервалами.

Для двумерных нормальных данных удобно использовать так называемое преобразование Фишера $arctanh(z) = 0.5\ln((1+z)/(1-z))$, приводящее коэффициент ρ к асимптотической нормальности со среднем $arctanh(\rho)$ и дисперсией $1/(n-3)$, откуда можно построить доверительный интервал для ρ. Аналогичный результат верен и для коэффициентов Спирмена и Кенделла.

При выполнении гипотезы независимости справедливы сходимости

$$\frac{\rho_S}{\sqrt{D\rho_S}} = \rho_S\sqrt{n-1} \xrightarrow{d} Z \sim \mathcal{N}(0,1),\ \frac{\tau}{\sqrt{D\tau}} = \tau\sqrt{\frac{9n(n-1)}{2(2n+5)}} \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

откуда вытекают асимптотические критерии для проверки гипотезы независимости. Эти критерии, а также их точные версии для небольших n реализованы в cor.test пакета stats, где, опять же, можно выбрать конкретный метод. Более точную версию критерия Спирмена можно найти в spearman.test.

Упражнение 2. Исследовать коэффициенты на выборках $X, f(X)+U$, где X, U — нез-ы $\mathcal{N}$ или R (и а) $f(x)=2x+5$, б) $f(x)=x^2$, в) $f(x)=\ln x$, г) $f(x)=\sin x$.

Упражнение 3. Исследовать коэффициенты на (X,Y), распределенных на кольце $\sqrt{X^2+Y^2} \in [a,b]$, а) $a=0.5$, $b=1$, б) $a=0.75$, $b=1$.

Для многих выборок ($k \geq 3$) аналогичную роль может сыграть коэффициент конкордации, подсчитанный по многим выборкам. Одним из таких коэффициентов является коэффициент Кенделла

$$W = \frac{12}{k^2(n^3-n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{k} R_{i,j} - \frac{k(n+1)}{2} \right)^2,$$

где $R_{i,j}$ — ранг i-ого элемента j-ой выборки внутри своей выборки. Эта величина лежит в диапазоне $[0,1]$. При больших n $k(n-1)W$ близок по распределению к χ^2_{n-1}. В R этот коэффициент задается функцией KendallW из DescTools, позволяющей также найти соответствующее p-value гипотезы о совместной независимости.

Еще одним важным коэффициентом является частный или очищенный коэффициент корреляции, позволяющий исключить зависимость двух переменных через третью

$$\frac{\rho_{X,Y} - \rho_{X,Z}\rho_{Y,Z}}{\sqrt{(1-\rho^2_{X,Z})(1-\rho^2_{Y,Z})}},$$

который оценивается исходя из выборки с помощью статистики

$$\frac{\hat\rho_{X,Y} - \hat\rho_{X,Z}\hat\rho_{Y,Z}}{\sqrt{(1-\hat\rho^2_{X,Z})(1-\hat\rho^2_{Y,Z})}},$$

где $\hat\rho$ — выборочные коэффициенты корреляции Пирсона. Этот коэффициент эффективно работает для многомерных нормальных данных, для общих данных можно использовать выборочные коэффициенты корреляции Кенделла. В R такие величины могут быть рассчитаны с помощью функции pcor пакета ppcor. Полученная матрица estimate будет содержать исключенные корреляции по каждой паре переменных при условии всех остальных.

Упражнение 4. Исследовать выборки X,Y,Z с помощью частных корреляций: а) $U \sim \mathcal{N}(0,1)$, $V \sim \mathcal{N}(0,1)$, $Z \sim \mathcal{N}(1,1)$, где $X=U+Z$, $Y=V+Z$, б) $U \sim R[0,1]$, $V \sim R[0,1]$, $Z \sim \exp(\lambda)$, $X=U^2Z$, $Y=V^2Z$.

Ковариационные расстояния

Работа с ковариацией затрудняется тем, что она измеряет именно линейную зависимость. Для измерения общей зависимости можно использовать другие функционалы, ковариационные расстояния.

1) Метод Хёфдинга (Hoeffding). Этот метод основан на том, что величина

$$D = \int_{\mathbb{R}^2} (F_{X,Y}(x,y) - F_X(x)F_Y(y))^2 dF_{X,Y}(x,y)$$

неотрицательна и равна нулю в абсолютно-непрерывном случае тогда и только тогда, когда величины независимы.

Вопрос 1. Показать, что в дискретном случае это не так.

При этом $D \leq 1/30$. Для оценки D используется довольно сложная оценка, нахождение которое реали-

зовано в R в виде функции hoeffd пакета Hmisc, которая по матрице данных строит матрицу расстояний.
2) Более новый метод Шекели-Риззо (Szekely-Rizzo, 2009) предлагает рассматривать в качестве меры зависимости между случайными векторами $X \in \mathbb{R}^k$, $Y \in \mathbb{R}^l$, имеющими конечное математическое ожидание, величину

$$V^2_{X,Y} = \int_{t\in\mathbb{R}^k, s\in\mathbb{R}^l} |\psi_{X,Y}(t,s) - \psi_X(t)\psi_Y(s)|^2 w(t,s)dtds,$$

обнуляющуюся только при независимых векторах X, Y, где w — некоторая (интегрируемая) весовая функция. Тогда

$$\rho_V = \frac{V^2_{X,Y}}{\sqrt{V_{X,X}V_{Y,Y}}}$$

будет принимать значения в отрезке $[0,1]$. В роли $w(t,s)$ предлагается рассматривать

$$\frac{1}{c_k c_l ||t||_k^{1+k} ||s||_l^{1+l}},\ c_m = \frac{\pi^{1+m}}{\Gamma((1+m)/2)},$$

где норма рассматривается Евклидова. Для оценки ρ_V на выборке из n выборок (X,Y) введем

$$a_{i,j} = ||X_i - X_j||_k,\ a_{i,\cdot} = \frac{1}{n}\sum_{j=1}^n a_{i,j},\ a_{\cdot,\cdot} = \frac{1}{n^2}\sum_{i,j} a_{i,j},\ b_{i,j} = ||Y_i - Y_j||_l,\ b_{\cdot,j} = \frac{1}{n}\sum_{i=1}^n b_{i,j},\ b_{\cdot,\cdot} = \frac{1}{n^2}\sum_{i,j} b_{i,j},$$

положим $A_{i,j} = a_{i,j} - a_{i,\cdot} - a_{\cdot,j} + a_{\cdot,\cdot}$, $B_{i,j} = b_{i,j} - b_{i,\cdot} - b_{\cdot,j} + b_{\cdot,\cdot}$. Тогда

$$\hat{\rho}_V = \frac{\sum_{i,j} A_{i,j}B_{i,j}}{\sqrt{\sum_{i,j} A^2_{i,j} \sum_{i,j} B^2_{i,j}}}.$$

Это сильно состоятельная оценка ρ_V.
При этом можно построить асимптотический критерий, основанный на том, что при выполнении гипотезы независимости

$$\limsup_{n\to\infty} P\left(\frac{n\hat{\rho}_V}{S_2} > z^2_{1-\alpha/2}\right) \le \alpha,$$

при некотором S_2, где z — квантиль $\mathcal{N}(0,1)$.
В R этот метод реализован функцией dcor, а соответствующий критерий — функцией dcor.ttest пакета Energy.
Упражнение 5. Исследовать метод на тех же примерах, что и в предыдущей части.
В случае многих переменных строят матрицу попарных расстояний и таким образом исследуют, какие переменные сильно зависимы.
Метод главных компонент
Рассмотрим данные $X = (x_{i,j}, i \le n, j \le m)$. Будем считать, что у нас есть выборка из n объектов, где по строкам записаны признаки объектов. выборки расположены по столбцам и центрированы, $n^{-1}\sum_{i=1}^n x_{i,j} = 0$. Положим $\hat{\Sigma}^2 = (\hat{\sigma}^2_{k,l}) = X^tX$ – выборочная матрица ковариации признаков, характеризующая их зависимость.
Предположим, что $\lambda_1 > \lambda_2 .... > \lambda_m > 0$ — собственные значения $\hat{\Sigma}^2$, $e_1, ..., e_m$ — соответствующий ортогональный собственный базис. Рассмотрим сумму выборочных дисперсий признаков

$$\sum_{j=1}^m \sum_{i=1}^n X^2_{i,j} = tr(\hat{\Sigma}^2) = \lambda_1 + ... + \lambda_m.$$

Предположим, что для достаточно малого ε $\lambda_1 + \lambda_2 + ... + \lambda_i \ge (1-\varepsilon)(\lambda_1 + ... + \lambda_m)$. Тогда первые i признаков фактически определяют все остальные.
Переходя от исходных признаков к их линейным комбинациям, соответствующим e_1,...,e_i, мы снижаем размерность пространства признаков. Этот метод называют методом главных компонент.
В случае, если данные независимы, выборочная матрица ковариации близка к теоретической и при больших n их собственные значения также будут близки.

В R анализ главных компонент осуществляется функицей princomp (напрямую) и prcomp пакета stats. Полезно также использовать график biplot, аргументом которого является результат princomp. В случае princomp матрица loadings, а в случае prcomp матрица rotation по столбцам содержит собственные векторы в порядке убывания собственных значений, scores содержит координаты точек в новых координатах.

Метод главных компонент хорошо работает в случае данных, близких к нормальным. В связи с этим зачастую перед использованием к данным применяют преобразования, например, Box-Cox transformation. Функция preProcess пакета caret дает возможность использовать больше преобразований, например, BoxCox, center, scale, pca.

Упражнение 6. Применить метод главных компонент к данным базе iris и определить, можно ли по первым главным компонентам отделить цветы друг от друга.

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 $(x, y) = c(t)$, где t — натуральный параметр ($|c'(t)| = 1$), если $(X, Y) = c(S) + Vd(T)$ для некоторого вектора (S, T), где $d(t)$ — касательный вектор к кривой в точке $c(t)$. При этом будем требовать: а) $E(T|S) = 0$ п.н. б) $D(T|S) \leq DS$ п.н.

Вопрос 2. Найти распределения S, $T|S$ для выборки а) равномерно распределенной на квадрате с вершинами $(1, 1)$, $(1, -1)$, $(-1, -1)$, $(-1, 1)$ для $c(t) = (t, 0)$.

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

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

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

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

Вопрос 3. Показать, что если $c(t)$ — прямая, а $D(T|S = s)$ не зависит от s, то $DLoc_X = DX$, $DLoc_Y = DY$, $covLoc_{X,Y} = cov(X, Y)$.

Тогда ковариацией и корреляцией около кривой называют

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

Вопрос 4. Найти ковариацию и корреляцию (X, Y) для распределения на кольце, описанного выше.

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, вычисляющая covGC и corrGC, находится в приложенном файле CovrGC.

Упражнение 7. Исследовать метод на (X, Y), распределенных на кольце $\sqrt{X^2 + Y^2} \in [a, b]$, а) $a = 0.5$, $b = 1$, б) $a = 0.75$, $b = 1$.