САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

МАТЕМАТИКО-МЕХАНИЧЕСКИЙ ФАКУЛЬТЕТ

Кафедра теории вероятностей и математической статистики

Проектная работа

студента 311 группы:

Подлужного Ивана Андреевича

Руководитель:
к. ф-м. н., О.В. Русаков

8 июня 2018 года

Содержание

1 Постановка задачи

Рассмотрим следующий процесс: путь на плоскость $\mathbb{R}^2$ случайным образом ставят точки – независимые реализации случайного вектора, имеющего нормальное распределение с параметрами σ_X, σ_Y, ρ. На каждом шаге проводится прямая МНК и прямые регрессии X на Y и Y на X. К чему и как быстро будут сходится предельные $n \longrightarrow \infty$ параметры прямых? В работе также рассмотрены смеси нормального распределения со случайным параметром ρ, где нам удалось ответить на часть поставленных вопросов. Результаты подкреплены математической моделью данного процесса в среде `R`.

Работа разбита на 3 части:

1. Исследование сходимости при постоянном параметре ρ
2. Исследование смесей нормального распределения
3. Практическая реализация

Все необходимые предварительные выкладки находятся в Приложении.

Часть I
Теоретическая модель

2 Сходимость параметров выборки гауссова вектора

2.1 Базовые сходимости

Пусть $\mathbf{z_k} = (x_k, y_k)$ н.о.р с плотностью
$p_{\mathbf{z}}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2} - 2\frac{\rho xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}\right)\right)$
То есть $\mathbb{E}x_k = 0$, $\mathbb{E}y_k = 0$, $\mathbb{D}x_k = \sigma_X^2$, $\mathbb{D}y_k = \sigma_Y^2$.

Рассмотрим поведение прямых выборки $(x_k, y_k)_{k=1}^n$ при $n \longrightarrow \infty$. Рассмотрим сходимости почти наверное и в среднем порядка 2.

2.1.1 Сходимость почти наверное

Рассмотрим сходимости $\frac{\sum\limits_{k=1}^{n} x_k}{n} = \bar{x}_n$ и $\frac{\sum\limits_{k=1}^{n} y_k}{n} = \bar{y}_n$:
По закону повторного логарифма почти наверное

$$\begin{aligned} \limsup \frac{\sum\limits_{k=1}^{n} x_k}{n} \frac{1}{\sigma_X}\sqrt{\frac{n}{\ln\ln n}} = \sqrt{2} \\ \liminf \frac{\sum\limits_{k=1}^{n} x_k}{n} \frac{1}{\sigma_X}\sqrt{\frac{n}{\ln\ln n}} = -\sqrt{2} \end{aligned} \tag{1}$$

то есть $|\bar{x}_n| = O(\sigma_X\sqrt{\frac{2\ln\ln n}{n}})$ и $\exists\, n_k\ |\bar{x}_{n_k}| > (\sigma_X - \varepsilon)\sqrt{\frac{2\ln\ln n}{n}})$. Аналогичное верно и для $\bar{y}_k$.

Перейдём к выборочным дисперсиям. Заметим, что $n\overline{x^2}_n$ тоже сумма случайных величин, только теперь уже имеющая распределение χ^2 с числом степеней свободы равном 1, и матожидание σ_X^2, дисперсию $2\sigma_X^4$. Тогда по закону повторного логарифа

$$\overline{Dx}_n = \frac{\sum\limits_{k=1}^{n} x_i^2}{n} - \left(\frac{\sum\limits_{k=1}^{n} x_i}{n}\right)^2$$

По закону повторного логарифма, $\frac{\sum\limits_{k=1}^{n} x_i^2 - \sigma_X^2}{\sqrt{2\sigma_X^2\sqrt{n\ln\ln n}}}$, строго "лежит"между $-\sqrt{2}$ и $\sqrt{2}$. Но последнее выражение равно:

$$\begin{aligned} &\left(\overline{Dx}_n + \left(\frac{\sum\limits_{k=1}^{n} x_i}{n}\right)^2 - \sigma_X^2\right) \frac{\sqrt{n}}{\sqrt{2\sigma_X^2\sqrt{\ln\ln n}}} = \\ &= \left(\overline{Dx}_n - \sigma_X^2\right) \frac{\sqrt{n}}{\sqrt{2\sigma_X^2\sqrt{\ln\ln n}}} + \left(\frac{\sum\limits_{k=1}^{n} x_i}{n} \frac{\sqrt[4]{n}}{\sqrt[4]{4\sigma_X^4\ln\ln n}}\right)^2 \end{aligned} \tag{2}$$

, где последнее слагаемое сходится к 0. Таким образом, $\overline{Dx}_n$ сходится к σ_X^2 со "скоростью"$\sigma_X\sqrt{\frac{\ln\ln n}{n}}$, (т.е. почти наверное разность есть $O\left(2\sigma_X^2\sqrt{\frac{\ln\ln n}{n}})\right)$)

Аналогично $\overline{Dy}_n$ сходится к σ_Y^2 .

Рассмотрим $\overline{Cov(x,y)}_n$ По закону повторного логарифма, $\frac{\sum\limits_{i=1}^{n} X_iY_i - n\mathbb{E}XY}{\sqrt{n(\mathbb{D}XY)n\ln\ln n}}$ "строго"лежит между $[-\sqrt{2},\sqrt{2}]$

Но $DXY = \mathbb{E}X^2Y^2 - (\mathbb{E}XY)^2 \leqslant \frac{\mathbb{E}X^4+\mathbb{E}Y^4}{2} = 3\frac{\sigma_X^4+\sigma_Y^4}{2}$ и как в случае с дисперсиями, слагаемое с $\sum\limits_{i=1}^{n} X_iY_i$ будет сходится медненее всего и в итоге получим оценку $\frac{\sqrt{3(\sigma_X^4+\sigma_Y^4)}}{2}\sqrt{\frac{\ln\ln n}{n}}$.

2.1.2 Сходимость в L^2

Рассмотрим аналогичную ситуацию в L^2. Ограничимся здесь цепочками переходов к нужным неравенствам:

$$\sqrt{\mathbb{E}\left(\frac{\sum\limits_{k=1}^{n} X_i}{n}\right)^2} = \sqrt{\frac{\mathbb{E}X^2}{n}} = \frac{\sigma_X}{\sqrt{n}} \tag{3}$$

Теперь для $\overline{Dx}$:

$$\begin{gathered}\sqrt{\mathbb{E}\left(\overline{\mathbb{D}X} - \sigma_X^2\right)^2} = \sqrt{\mathbb{E}\left(\frac{\sum\limits_{k=1}^{n} X_i^2}{n} - \left(\frac{\sum\limits_{k=1}^{n} X_i}{n}\right)^2 - \sigma_X^2\right)^2} \leqslant \\ \leqslant \sqrt{\mathbb{E}\left(\frac{\sum\limits_{k=1}^{n} X_i^2}{n} - \sigma_X^2\right)^2} + \sqrt{\mathbb{E}\left(\frac{\sum\limits_{k=1}^{n} X_i}{n}\right)^4} = \sqrt{\mathbb{E}\left(\frac{\sum\limits_{k=1}^{n} X_i^2-\sigma_X^2}{n}\right)^2} + \sqrt{\mathbb{E}\left(\frac{\sum\limits_{k=1}^{n} X_i}{n}\right)^4} = \\ = \sqrt{\frac{3\sigma_X^4}{n}} + \sqrt{\frac{3\sigma_X^4}{n^3} + \frac{6\sigma_X^4(n^2-n)}{n^4}} \leqslant \frac{2\sigma_X^2}{\sqrt{n}}\end{gathered} \tag{4}$$

И $\overline{Cov(X,Y)}$

$$\begin{gathered}\sqrt{\mathbb{E}\left(\overline{Cov(X,Y)} - \rho\sigma_X\sigma_Y\right)^2} \leqslant \sqrt{\mathbb{E}\left(\frac{\sum\limits_{k=1}^{n} X_iY_i}{n} - \rho\sigma_X\sigma_Y\right)^2} + \sqrt{\mathbb{E}\left(\frac{\sum\limits_{k=1}^{n} X_i}{n}\frac{\sum\limits_{k=1}^{n} Y_i}{n}\right)^2} = \\ = \sqrt{\frac{\mathbb{E}(XY-\rho\sigma_X\sigma_Y)^2}{n}} + \sqrt{\mathbb{E}\left(\frac{(\sum\limits_{k=1}^{n} X_i^2+\sum X_iX_j)(\sum\limits_{k=1}^{n} Y_i^2+\sum Y_iY_j)}{n^4}\right)} = \\ = \frac{C}{\sqrt{n}} + \sqrt{\mathbb{E}\left(\frac{\sum X_i^2Y_j^2+\sum X_i^2Y_jY_k+\sum Y_i^2X_jX_k+\sum X_iX_jY_kY_l}{n^4}\right)} = \\ = \frac{C}{\sqrt{n}} + \sqrt{\mathbb{E}\left(\frac{\sum X_i^2Y_i^2+\sum\limits_{i\neq j} X_i^2Y_j^2+2\sum X_iX_jY_iY_j}{n^4}\right)} = \\ \frac{C}{\sqrt{n}} + \sqrt{\left(\frac{\mathbb{E}X^2Y^2+n(n-1)\sigma_X^2\sigma_Y^2+2n(n-1)\rho\sigma_X\sigma_Y}{n^4}\right)} \leqslant \frac{C+\varepsilon}{\sqrt{n}}\end{gathered} \tag{5}$$

2.2 Сходимость прямых регрессии

Ещё раз выпишем выборки уравнений регресии:

$$\begin{array}{l} \widehat{a_{x,n}} = \frac{\overline{Cov(x,y)}_n}{\overline{Dx}_n} \\ \widehat{b_{x,n}} = \overline{y}_n - \widehat{a_{x,n}}\overline{x}_n \\ \widehat{a_{y,n}} = \frac{\overline{Cov(x,y)}_n}{\overline{Dy}_n} \\ \widehat{b_{y,n}} = \overline{x}_n - \widehat{a_{y,n}}\overline{y}_n \end{array} \tag{6}$$

Заметим, что как мы упомянули ранее, $\overline{Dx}_n \longrightarrow \sigma_X^2 \neq 0$, $\overline{Dy}_n \longrightarrow \sigma_Y^2 \neq 0$, $\overline{Cov(x,y)}_n \longrightarrow \rho\sigma_X\sigma_Y$. Таким образом, почти наверное:

$$\begin{array}{l} \widehat{a_{x,n}} \longrightarrow \frac{\rho\sigma_Y}{\sigma_X} \\ \widehat{a_{y,n}} \longrightarrow \frac{\rho\sigma_X}{\sigma_Y} \end{array} \tag{7}$$

и, в силу того, что $\widehat{a_{x,n}}$ и $\widehat{a_{y,n}}$ почти наверное:

$$\begin{array}{l} \widehat{b_{x,n}} = \overline{y}_n - \widehat{a_x}\overline{x}_n \longrightarrow 0 \\ \widehat{b_{y,n}} = \overline{x}_n - \widehat{a_y}\overline{y}_n \longrightarrow 0 \end{array} \tag{8}$$

Так как $\overline{Dx}_n \longrightarrow \sigma_X$, то скорость сходимости $\widehat{a_{x,n}}$ будет равна хотя бы $\frac{\sqrt{3(\sigma_X^4+\sigma_Y^4)}}{2\sigma_X^2}\sqrt{\frac{\ln\ln n}{n}}$. Аналогично, скорость сходимости $\widehat{a_{y,n}}$ будет $\frac{\sqrt{3(\sigma_X^4+\sigma_Y^4)}}{2\sigma_Y^2}\sqrt{\frac{\ln\ln n}{n}}$.

2.3 Сходимость наилучших прямых выборки гауссова вектора к главным осям гауссова вектора

2.3.1 Сходимость прямых при $\sigma_X \neq \sigma_Y$

Из сходимостей $\overline{\mathbb{D}x}_n$, $\overline{\mathbb{D}y}_n$ и $\overline{Cov(x,y)}_n$ следует сходимость $A_n = \begin{pmatrix} \overline{\mathbb{D}x}_n & \overline{Cov(x,y)}_n \\ \overline{Cov(x,y)}_n & \overline{Dy}_n \end{pmatrix}$ к $A = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$ по L_2, так как $||A||_2 \leqslant \sqrt{|a_{11}|^2+|a_{12}|^2+|a_{21}|^2+|a_{22}|^2}$. Тогда имеем

$$||A - A_n||_2 \to 0$$

со скоростью $C\sqrt{\frac{\ln\ln n}{n}}$. Разложим A_n в сумму A и $A - A_n$. Заметим, что так как $A - A_n$ симметрична, все собственные числа также вещественны и $\max\limits_{i=1,2}|\mu_{i,n}| = ||A - A_n||_2 \to 0$. Тогда, разложим A_n в сумму A и $A_n - A$ и так как обе матрицы симметричны, выполнены следующие тождества:

Пусть $\alpha_{1,n} \leqslant \alpha_{2,n}$ – собственные числа A_n, $\mu_{1,n} \leqslant \mu_{2,n}$ – собственные числа $A_n - A$ и $\lambda_1 \leqslant \lambda_2$ собственные числа A. Используем такой факт: если A_n, A симметричные матрицы и α, μ, λ такие как выше, то верны следующие неравенства:

$$\begin{cases} \alpha_{1,n} \geqslant \lambda_1 + \mu_{1,n} \\ \alpha_{1,n} \leqslant \lambda_1 + \mu_{2,n} \\ \alpha_{2,n} \geqslant \lambda_2 + \mu_{1,n} \\ \alpha_{2,n} \geqslant \lambda_2 + \mu_{2,n} \end{cases} \tag{9}$$

Откуда $|\lambda_1 - \alpha_{1,n}| \leqslant \max\limits_{i=1,2}|\mu_{i,n}| = ||A - A_n||_2 \to 0$. Аналогично $|\lambda_2 - \alpha_{2,n}| \leqslant \max\limits_{i=1,2}|\mu_{i,n}| = ||A - A_n||_2 \to 0$. При этом:

$$||A - A_n||_2 \leqslant \sqrt{\sum |a_{ij}^0 - a_{ij}^n|^2} \leqslant \sqrt{\frac{\ln\ln n}{n}}\sqrt{4(\sigma_X^4 + \sigma_Y^4) + \frac{3}{4}(\sigma_X^4 + \sigma_Y^4)} \leqslant \sqrt{5(\sigma_X^4 + \sigma_Y^4)} \tag{10}$$

Аналогично для второго собственного числа.

Но раз так, то пусть $u_{1,n}$ собственный вектор A_n при с.ч. $\lambda_{1,n}$ и $||u_{1,n}|| = 1$, т.е. $A_n u_{1,n} = \lambda_{1,n} u_{1,n}$, $\lambda_{1,n} \to \lambda_1$. Тогда

$$|(A - \lambda_1 E)u_{1,n}| \leqslant |(A_n - \lambda_{1,n}E)u_n| + |u_{1,n}||\lambda_1 - \lambda_{1,n}| + ||A - A_n|||u_{1,n}| \to 0 \tag{11}$$

причём порядок сходимости $\sqrt{\frac{\ln\ln n}{n}}$ не меняется –обе части убывают одинаково быстро.

Пусть же теперь ещё и $\dim\ker(A - \lambda_1 E) = 1$, то есть $\lambda_1 \neq \lambda_2$. Тогда $||A - \lambda E||_2 = |\lambda_2 - \lambda_1|$, тогда имеем

$$d_2\left(u_{1,n}, \ker(A - \lambda_1 E)\right) = \frac{|(A-\lambda_1 E)u_{1,n}|}{||A-\lambda E||_2} = \frac{|(A-\lambda_1 E)u_{1,n}|}{|\lambda_2 - \lambda_1|} \to 0 \tag{12}$$

Но по предположению, $\ker(A - \lambda_1 E)$ однмерно, значит $\ker(A - \lambda_1 E) = \langle u_1 \rangle$. То есть точка $u_{1,n}$ приближается к прямой, заданной вектором u_1. Аналогичное верно и для u_2.

Напоследок $|\lambda_2 - \lambda_1| = \left|\frac{\sigma_X^2 - \sigma_Y^2}{\cos 2\varphi}\right| = \left|\frac{\sigma_X^2 - \sigma_Y^2}{\cos 2\varphi}\right| = \sqrt{(\sigma_X^2 - \sigma_Y^2)^2 + 4\sigma_X^2\sigma_Y^2\rho^2} = \sqrt{(\sigma_X^2 + \sigma_Y^2 - 2\sqrt{1-\rho^2}\sigma_X\sigma_Y)(\sigma_X^2 + \sigma_Y^2 + 2\sqrt{1-\rho^2}\sigma_X\sigma_Y)}$

Итог: Пусть есть выборка (x_i, y_i) гауссова вектора $\mathbf{z}$ с нулевым средним, дисперсиями компонент σ_X, σ_Y и ковариацией $\rho\sigma_X\sigma_Y$. Пусть $\begin{pmatrix} \overline{\mathbb{D}x}_n & \overline{Cov(x,y)}_n \\ \overline{Cov(x,y)}_n & \overline{Dy}_n \end{pmatrix}$ и $A = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$ Тогда

1. $\bar{x} \to 0$ и $\bar{y} \to 0$ почти наверное со "скоростями"$\sigma_X\sqrt{\frac{2\ln\ln n}{n}}$ и $\sigma_Y\sqrt{\frac{2\ln\ln n}{n}}$.
2. $||A_n - A||_2 \to 0$ хотя бы как $\sqrt{5(\sigma_X^4 + \sigma_Y^4)}\sqrt{\frac{\ln\ln n}{n}}$ почти наверное
3. Собственные числа A_n сходятся к собственным числам A с той же скоростью.

Пусть дополнительно собственные числа матрицы $\begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$ не равны друг другу, т.е. $\sigma_X \neq \sigma_Y$ и $\rho \neq 0$ при $\sigma_X = \sigma_Y$, то тогда:

1. Собственные вектора A_n сближаются с прямыми, порождёнными соответствующими собственными веторами A.
2. Наилучшая прямая для выборки (x_i, y_i), порождена собственным вектором u матрицы A_n при её наибольшем собственном числе и сходится к главному направлению $\mathbf{z}$

Обе этих сходимости порядка $\frac{1}{\sqrt{(\sigma_X^2 - \sigma_Y^2)^2 + 4\sigma_X^2\sigma_Y^2\rho^2}}\sqrt{\frac{\ln\ln n}{n}}$

2.3.2 Случай $\sigma_X = \sigma_Y$, $\rho = 0$

Пусть теперь $A = \sigma_X E$ и соответственно $\dim\ker(A - \sigma_X E) = 2$. Посмотрим на ситуацию иначе:

Наши $\{(x_i, y_i)\}_{i=1}^n$ это выборка гауссова вектора $\mathbf{z}$, $\sigma_X = \sigma_Y$ и $\rho = 0$, а значит некоторая случайный вектор в $\mathbb{R}^{2n}$. Угол наилучшей прямой это почти наверное определённая функция $\psi \in [0, 2\pi)/\pi\mathbb{Z}$ от этого вектора, причём если каждую пару $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ заменить на $U_\varphi \begin{pmatrix} x_i \\ y_i \end{pmatrix}$, то функция ψ_n перейдёт в $\psi_n + \varphi$. Тогда рассмотрим такое событие: $\psi_n \in (\alpha, \beta)$ $\alpha, \beta \in [0, 2\pi)/\pi\mathbb{Z}$. Но если заменить $\{(x_i, y_i)\}_{i=1}^n$ на $\{(x_i, y_i)U_\varphi\}_{i=1}^n$, то ничего не изменится, так как распределения (x_i, y_i) и $(x_i, y_i)U_\varphi$ равны. Но, как мы говорили, при этой замене событие $\psi_n \in (\alpha, \beta)$ перейдёт в $\psi_n + \varphi \in (\alpha, \beta)$. Значит:

$$\mathbb{P}(\psi_n \in (\alpha, \beta)) = \mathbb{P}(\psi \in (\alpha - \varphi, \beta - \varphi)) \quad \forall \alpha, \beta, \varphi \in \varphi \in [0, 2\pi)/\pi\mathbb{Z} \tag{13}$$

И ясно, что $\mathbb{P}(\psi_n \in (0, \pi)) = 1$. Тогда отсюда следует, что ψ_n равномерно распределена на $[0, 2\pi)/\pi\mathbb{Z}$. Заметим, что этот результат не зависит от n.

2.4 Сходимость сумм расстояний до наилучшей прямой

Как мы отметитли ранее

$$\begin{aligned} R^2 &= \tfrac{n}{2}\left(\left(\overline{Dx} + \overline{Dy}\right) + \cos 2\varphi\left(\overline{Dx} - \overline{Dy}\right) + 2\sin 2\varphi\overline{Cov(x,y)}\right) = \\ &= \tfrac{n}{2}\left(\overline{Dx} + \overline{Dy} - \sqrt{\left(\overline{Dx} - \overline{Dy}\right)^2 + 4\overline{Cov(x,y)}^2}\right) \end{aligned} \tag{14}$$

То есть при $n \longrightarrow \infty$ почти наверное

$$\left(\tfrac{R}{\sqrt{n}}\right)^2 \longrightarrow \tfrac{1}{2}\left(\sigma_X^2 + \sigma_Y^2 - \sqrt{\left(\sigma_X^2 - \sigma_Y^2\right)^2 + 4\rho\sigma_X^2\sigma_Y^2}\right) \tag{15}$$

То есть, R имеет асимптотику порядка

$$\sqrt{n}\sqrt{\sigma_X^2 + \sigma_Y^2 - \sqrt{\left(\sigma_X^2 - \sigma_Y^2\right)^2 + 4\rho\sigma_X^2\sigma_Y^2}} \tag{16}$$

2.5 Сходимость при случайной величине ρ

Пусть ρ теперь случайная величина, расперделённая на $[-1, 1]$, имеющая плотность $p_\rho(t)$. Тогда заметим, что у теперь уже не гауссова вектора $\mathbf{z}$ можно вычислить $\mathbb{E}\mathbf{z}$, σ_X^2, σ_Y^2 и $Cov(x, y)$ достаточно просто, так как его плотность будет равна $p_{\mathbf{z}}(x, y) = \int\limits_{-1}^{1} p_\rho(t) p_{XY}(x, t)\, dt$ по формуле условной вероятности. Тогда, применяя теорему Фубини:

$$p_{\mathbf{z}}(x, y) = \int\limits_{-1}^{1} p_\rho(t) \tfrac{1}{2\pi\sigma_X\sqrt{1-t^2}} \exp\left(-\tfrac{1}{2\sigma_X^2(1-t^2)}\left(\mathbf{x}, B(t)\mathbf{x}\right)\right) dt \tag{17}$$

где $B(t) = \begin{pmatrix} 1 & -t \\ -t & 1 \end{pmatrix}$.

$$\mathbb{E}x = \int\limits_{\mathbb{R}^2} x \int\limits_{-1}^{1} p_\rho(t) p_{XY}(x, t)\, dt\, dx\, dy = \int\limits_{-1}^{1} p_\rho(t) \int\limits_{\mathbb{R}^2} x p_{XY}(x, t)\, dx\, dy\, dt = 0 \tag{18}$$

Аналогично с $\mathbb{E}y$

$$\begin{aligned}\mathbb{D}x = \int\limits_{\mathbb{R}^2} x^2 \int\limits_{-1}^{1} p_\rho(t) p_{XY}(x,t)\,dt\,dx\,dy = \int\limits_{-1}^{1} p_\rho(t) \int\limits_{\mathbb{R}^2} x^2 p_{XY}(x,t)\,dx\,dy\,dt = \\ = \sigma_X^2 \int\limits_{-1}^{1} p_\rho(t)\,dt = \sigma_X^2\end{aligned} \tag{19}$$

Аналогично с $\mathbb{D}y$

$$\begin{aligned}Cov(x,y) = \int\limits_{\mathbb{R}^2} xy \int\limits_{-1}^{1} p_\rho(t) p_{XY}(x,t)\,dt\,dx\,dy = \\ = \int\limits_{-1}^{1} p_\rho(t) \int\limits_{\mathbb{R}^2} xy p_{XY}(x,t)\,dx\,dy\,dt = \sigma_X\sigma_Y \int\limits_{-1}^{1} p_\rho(t)\rho\,dt = \\ = \sigma_X\sigma_Y\mathbb{E}\rho\end{aligned} \tag{20}$$

Таким образом, ситуация со сходимостью $\bar{x}_n$ и $\bar{y}_n$ будет похожа: они будут сходиться к 0 со "скоростью"$\sigma_X\sqrt{\frac{2\ln\ln n}{n}}$, аналогичное верно и для $\bar{y}_n$.

Заметим, что в предыдущих выкладках мы почти ничего не требовали от распределения $\mathbf{z}$, кроме существования дисперсии обеих компонент. Значит, выкладки для линий регрессии останутся полностью без изменений. Для наилучших прямых можно также рапространить имеющиеся результаты, лишь бы $\sigma_X \neq \sigma_Y$ и $\mathbb{E}\rho \neq 0$ при $\sigma_X = \sigma_Y$.

В отличие от нормального вектора, пределельное распределение в случае $\mathbb{E}\rho = 0\sigma_X = \sigma_Y$ описывается по-другому.

Для начала отметим, что если ρ имеет плотность, то:

$$\frac{2\overline{Cov(x,y)}}{\overline{Dx}-\overline{Dy}} - \frac{2\overline{XY}_n}{\overline{X^2}_n - \overline{Y^2}_n} \longrightarrow 0 \tag{21}$$

по вероятности. Это можно доказать напрямую, а можно заметить, что при вычитании, числитель интговой дроби будет иметь больший порядок малости, чем нижний.

Таким образом, мы можем перейти к изучению дроби $\frac{2\overline{XY}_n}{\overline{X^2}_n - \overline{Y^2}_n}$.

Рассмотрим $U_i = 2X_iY_i$, $V_i = X_i^2 - Y_i^2$, $\overline{2U} = \overline{2XY}$, $\overline{V} = \overline{X^2 - Y^2}$. Тогда вектор $\frac{\sum_i (U_i, V_i)}{\sqrt{n}} \to N(0,K)$ по многомерной ЦПТ, где $K = \begin{pmatrix} \mathbb{D}U_i & \rho Cov(U_i,V_i) \\ Cov(U_i,V_i) & \mathbb{D}V_i \end{pmatrix}$

Вычислим элементы этой матрицы. Для этого воспользуемся представлением:

$$\begin{aligned}U_i = 2X_iY_i = 2\sigma_X^2\alpha\left(\rho\alpha + \sqrt{1-\rho^2}\beta\right) \\ V_i = X_i^2 - Y_i^2 = \sigma_X^2\left((\alpha^2-\beta^2)(1-\rho^2) - 2\rho\sqrt{1-\rho^2}\alpha\beta\right)\end{aligned}, \tag{22}$$

где $\alpha, \beta \in N(0,1)$ и вычислим элементы K:

$$\begin{aligned}\mathbb{D}U_i = 4\mathbb{E}X_iY_i = 4\sigma_X^4\left(\mathbb{E}\rho^2 + 1\right) \\ \mathbb{D}V_i = \mathbb{E}X_i^2 - Y_i^2 = 4\sigma_X^4\left(1 - \mathbb{E}\rho^2\right) \\ Cov(U_i,V_i) = 0\end{aligned}. \tag{23}$$

Теперь заметим, что угол наклона прямой $\varphi \in S_1/\pi\mathbb{Z}$ непрерывно зависит от $\frac{\sum U_i}{\sqrt{n}}$ и $\frac{\sum V_i}{\sqrt{n}}$ в $\mathbb{R}^2 \setminus \{0\}$. Отсюда следует, что $\varphi(\sqrt{n}\overline{U}_n, \sqrt{n}\overline{V}_n) \to \varphi(U_0, V_0)$, где (U_0, V_0) распределён как $N(0,K)$. Вычислим это распределение.

Мы знаем, что

$$\begin{array}{l} \sin 2\varphi = \frac{-U_0}{\sqrt{U_0^2+V_0^2}} \\ \cos 2\varphi = \frac{-V_0}{\sqrt{U_0^2+V_0^2}} \end{array} . \tag{24}$$

Тогда

$$\tan\varphi = \frac{\sin 2\varphi}{1+\cos 2\varphi} = -\frac{U_0}{\sqrt{U_0^2+V_0^2}-V_0} . \tag{25}$$

Пусть $U_0 = r\sin\xi$, $V_0 = r\cos\xi$. Тогда

$$\tan\varphi = \frac{\sin 2\varphi}{1+\cos 2\varphi} = -\frac{U_0}{\sqrt{U_0^2+V_0^2}-V_0} = -\tan\frac{\xi}{2} . \tag{26}$$

То есть $\varphi = -\frac{\xi}{2}$, где $\xi \in (-\pi, \pi)$.
Совместная плотность (U_0, V_0) выглядит так:

$$p_{(U_0,V_0)}(r,\xi) = \frac{r}{2\mu_X\mu_Y\pi}\exp\left(-\frac{r^2}{2}\left(\frac{\sin^2\xi}{\mu_X^2}+\frac{\cos^2\xi}{\mu_Y^2}\right)\right) , \tag{27}$$

где

$$\begin{array}{l} \mu_X = \sqrt{\mathbb{D}U_0} = 2\sigma_X^2\sqrt{\mathbb{E}\rho^2+1} \\ \mu_Y = \sqrt{\mathbb{D}V_0} = 2\sigma_X^2\sqrt{1-\mathbb{E}\rho^2} \end{array} . \tag{28}$$

Найдём плотность ξ:

$$\begin{array}{l} p_\xi(t) = \int\limits_0^\infty p_{(U_0,V_0)}(r,t)\,dr = \int\limits_0^\infty \frac{r}{2\pi\mu_X\mu_Y\sqrt{1-\eta^2}}\exp\left(-\frac{r^2}{2}\left(\frac{\sin^2 t}{\mu_X^2}+\frac{\cos^2 t}{\mu_Y^2}\right)\right)\,dr = \\ = \frac{1}{2\pi\left(\frac{\sin^2 t}{\mu_X^2}+\frac{\cos^2 t}{\mu_Y^2}\right)} \end{array} . \tag{29}$$

При этом, $\varphi = -\frac{\xi}{2}$. То есть плотность $p_\varphi(t)$ равна

$$p_\varphi(t) = \frac{1}{\pi\mu_X\mu_Y\left(\frac{\sin^2 2t}{\mu_X^2}+\frac{\cos^2 2t}{\mu_Y^2}\right)} \quad t \in (-\tfrac{\pi}{2}, \tfrac{\pi}{2}) . \tag{30}$$

Рассмотрим случай ρ распределённую как $2B(k,k)-1$, где $B(k,k)$ бета распределение.

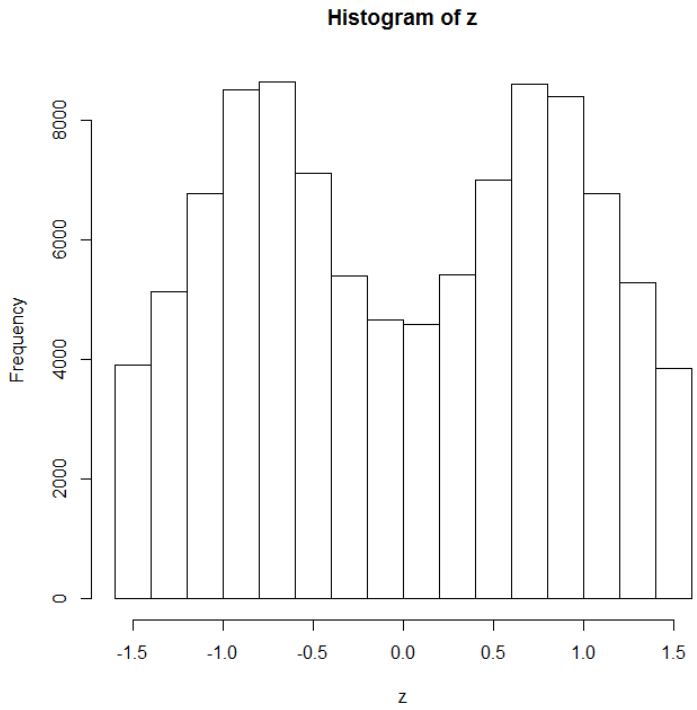


Рис. 1: Гистограмма частот угла наилучшей прямой для $\rho \in \mathrm{U}[-1,1]$, $n = 1000$, число проб 10^5

Как мы видим, при $k \to 0$, предельное распределение всё больше походит на дискретное распределение с параметрами

$$\begin{cases} -\frac{\pi}{4} & \frac{1}{2} \\ \frac{\pi}{4} & \frac{1}{2} \end{cases} \tag{31}$$

А при $k \longrightarrow \infty$ распределение переходит в $\mathrm{U}[-\frac{\pi}{2}, \frac{pi}{2}]$.

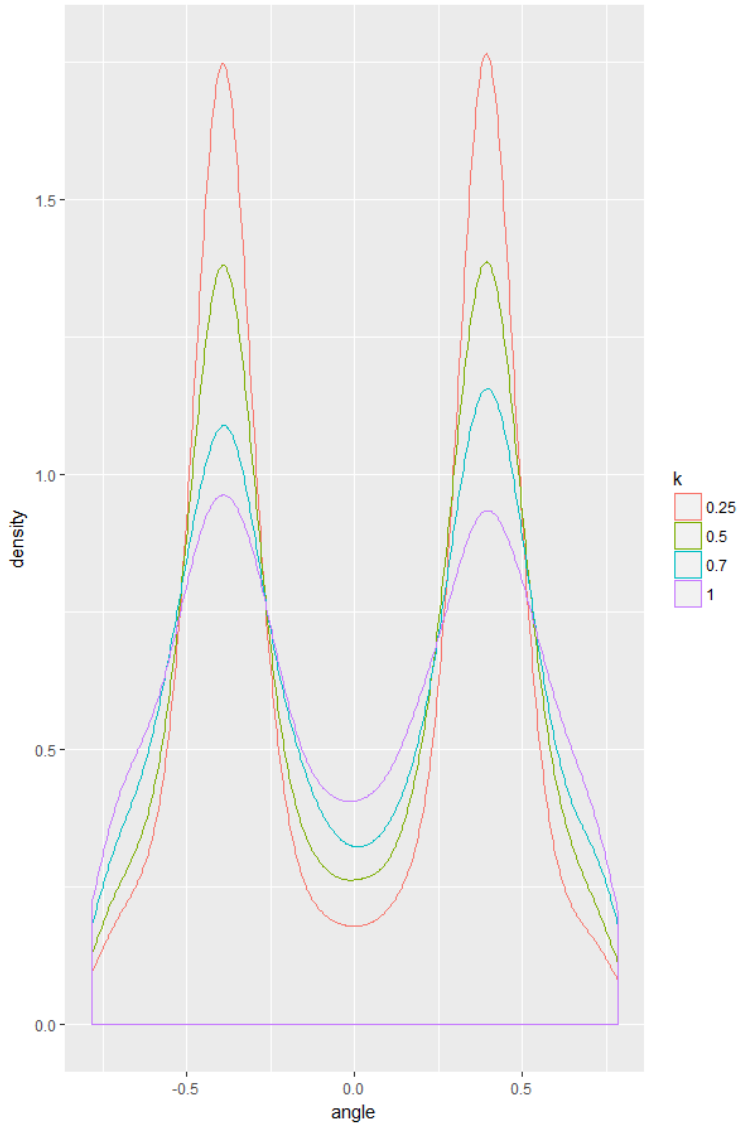


Рис. 2: Эмпирические плотности угла наилучшей прямой для $\rho \in 2\mathrm{B}(k,k) - 1$, для различных $n = 1000$, число проб 10^4

Часть II

Практическая реализация

3 Моделирование распределений

3.1 Нормальное распределение

Рассмотрим $\mathbf{z} = (\xi, \eta)$ стандартный гауссовский вектор, т.е. $p_{\mathbf{z}}(x,y) = \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right)$. Если мы перейдём к полярным координатам $x = r\sin\varphi$ $y = r\cos\varphi$, то:

$$p_{\mathbf{z}}(r,\varphi) = \frac{r}{2\pi}\exp\left(-\frac{r^2}{2}\right)$$

Таким образом, $\operatorname{Arg}\xi + i\eta$ равномерно распределён на $[0, 2\pi)$, а $r = |\xi + i\eta|$ имеет плотность $p_{|\mathbf{z}|}(r) = r\exp\left(-\frac{r^2}{2}\right)$ и $F_{|\mathbf{z}|}(t) = 1 - \exp\left(-\frac{t^2}{2}\right)$. . Но по формуле обращения $\xi = \sqrt{-2\ln\alpha}$, где α имеет равномерное распределение на

$[0,1]$. Таким образом, будем моделировать нормальную величину следующим образом:

$\xi = \sqrt{-2\ln\alpha_1}\cos 2\pi\alpha_2$

$\eta = \sqrt{-2\ln\alpha_1}\sin 2\pi\alpha_2$

где α_1, α_2 незаависимы и равномерно распределены на $[0,1]$.

Теперь смоделируем произвольный гауссовский вектор с параметрами σ_X, σ_Y, ρ. Для этого достаточно подобрать матрицу $A = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}$, такую, что умножив её на вектор $\begin{pmatrix} \xi \\ \eta \end{pmatrix}$ из независимых стандартных нормальных с.в. мы бы получили требуемое.

Распишем $\begin{pmatrix} a & 0 \\ b & c \end{pmatrix}\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} a\xi \\ b\xi + c\eta \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix}$.

И заметим, что $\sigma_X = a$, $\mathbb{E}XY = ab = \rho\sigma_X\sigma_Y$ и значит $b = \rho\sigma_Y$ и $\mathbb{D}Y = b^2 + c^2 = \sigma_Y^2$

$c^2 = \sigma_Y^2 - b^2 = \sigma_Y^2(1-\rho^2)$

$c = \sigma_Y\sqrt{1-\rho^2}$

Соответственно, получаем всё коэффициенты матрицы:

$a = \sigma_X$

$b = \rho\sigma_Y$

$c = \sigma_Y\sqrt{1-\rho^2}$

Алгоритм ниже иллюстрирует принцип работы $\texttt{Gauss}(\sigma_X, \sigma_Y, \rho)$

Data: Числа σ_X, σ_Y, ρ, генератор равномерного на $[0,1]$ распределения `RANDOM()` с независимыми вызовами

Result: Вектор (y_1, y_2) имеющий нормальное распределение с заданными параметрами

1 α_1 = `RANDOM()`
2 α_2 = `RANDOM()`
3 $x_1 = \sqrt{-2\ln\alpha_1}\cos 2\pi\alpha_2$
4 $x_2 = \sqrt{-2\ln\alpha_1}\sin 2\pi\alpha_2$
5 $y_1 = \sigma_X x_1$
6 $y_2 = \rho\sigma_Y x_1 + \sigma_Y\sqrt{1-\rho^2}x_2$

Algorithm 1: Схема работы $\texttt{Gauss}(\sigma_X, \sigma_Y, \rho)$

3.2 Бета-распределение

Также нам потребуется смоделировать с.в. ξ, имеющую бета-распределение. Для этого сформулируем такой факт:

Теорема 1. *Пусть* $\{\alpha_i\}_{i=0}^{\infty}$ *совместно независимые,* $\alpha_i \in \mathrm{U}[0,1]$. *Пусть* $k = \left\{\min N > 0 : \alpha_{2N-1}^{\frac{1}{\nu}} + \alpha_{2N}^{\frac{1}{\mu}} \leqslant 1\right\}$. *Тогда* $\xi = \frac{\alpha_{2k-1}^{\frac{1}{\nu}}}{\alpha_{2k-1}^{\frac{1}{\nu}} + \alpha_{2k}^{\frac{1}{\mu}}} \in \mathrm{B}(\nu,\mu)$

Доказательство. Пусть $B_N = \left\{\omega : \alpha_{2N-1}^{\frac{1}{\nu}} + \alpha_{2N}^{\frac{1}{\mu}} \leqslant 1\right\}$ и $\zeta_N = \alpha_{2N-1}^{\frac{1}{\nu}} + \alpha_{2N}^{\frac{1}{\mu}}$. Попробуем упростить формулу распределения (ξ, ζ). Сделаем невырожденную замену:

$y_1 = \frac{x_1^{\frac{1}{\nu}}}{x_1^{\frac{1}{\nu}} + x_2^{\frac{1}{\mu}}}$

$y_2 = x_1^{\frac{1}{\nu}} + x_2^{\frac{1}{\mu}}$

и
$x_1 = (y_1 y_2)^\nu$
$x_2 = y_2^\mu (1 - y_1)^\mu$

Якобиан равен $J = \mu\nu y_1^{\nu-1}(1-y_1)^{\mu-1} y_2^{\mu+\nu-1}$. И $p_{\xi\eta}(y_1, y_2) = J = \mu\nu y_1^{\nu-1}(1-y_1)^{\mu-1} y_2^{\mu+\nu-1}$ на $\left\{ y_1 \in [0,1]; y_2 \in \left[0, \min\{\frac{1}{y_1}, \frac{1}{1-y_1}\}\right]\right\}$.

Таким образом, $\frac{\alpha_{2N-1}^{\frac{1}{\nu}}}{\alpha_{2N-1}^{\frac{1}{\nu}} + \alpha_{2N}^{\frac{1}{\mu}}}$ при B_N

Тогда $p_\xi(y_1) = p_{\xi|_{B_N}}(y_1) = \frac{\int_0^1 p_{\xi\eta}(y_1,y_2)\,dy_2}{\int_0^1\int_0^1 p_{\xi\eta}(y_1,y_2) p_{\xi\eta}(y_1,y_2)\,dy_1\,dy_2} = \frac{\int_0^1 y_1^{\nu-1}(1-y_1)^{\mu-1} y_2^{\mu+\nu-1}\,dy_2}{\int_0^1\int_0^1 y_1^{\nu-1}(1-y_1)^{\mu-1} y_2^{\mu+\nu-1}\,dy_1\,dy_2} = \frac{y_1^{\nu-1}(1-y_1)^{\mu-1}}{\int_0^1\int_0^1 y_1^{\nu-1}(1-y_1)^{\mu-1}\,dy_1} = \frac{y_1^{\nu-1}(1-y_1)^{\mu-1}}{B(\nu,\mu)}$

что и требовалось доказать. □

Таким образом, применим следующий алгоритм:

Data: Числа ν и μ, генератор равномерного на $[0,1]$ распределения `RANDOM()` с независимыми вызовами
Result: ξ имеющая распределение $\mathrm{B}(\nu,\mu)$ или -1

```
α[1] = RANDOM()
α[2] = RANDOM()
k=1
while α[1]^(1/ν) + α[2]^(1/μ) > 1 OR (k<10) do
    α[1] = RANDOM()
    α[2] = RANDOM()
    k=k+1
end
if k<100 then
    ξ = α[1]^(1/ν) / (α[1]^(1/ν) + α[2]^(1/μ))
else
    ξ = -1
end
```

Algorithm 2: Схема работы `Beta`(ν,μ)

Таким образом, мы получили генератор `Beta`(ν,μ) бета-распределения.

4 Программное обеспечение

Данная модель представлена несколькими файлами в R. Приведём здесь краткое описание основных фукций:

- `Gauss`$(\sigma_X, \sigma_Y, \rho, N)$ – выдаёт массив N пар векторов (x_N^1, x_N^2), представляющих выборку независимых одинаково распределённых гауссовых векторов с параметрами σ_X, σ_Y, ρ.
- `Beta`(ν,μ,N) – выдаёт выборку из N независимых одинаково распределённых величин, имеющих $\mathrm{B}(\nu,\mu)$
- `GaussB`$(\sigma_X, \sigma_Y, \rho, \nu, \mu, N)$ – выдаёт массив N пар векторов (x_N^1, x_N^2), представляющих выборку независимых одинаково распределённых векторов

при независимо выбираемом $\rho \in 2\mathrm{B}(\nu,\mu) - 1$ и гаусовом распределении с параметрами σ_X, σ_Y, ρ.

- `SumInf`$(selections_{array})$ – по выборке из двумерного распределения рассчитывает: выборочные средние по X и Y, выборочные дисперсии, ковариацию, коэффициенты a и b наилучшей прямой $y = ax+b$ для выборки, и ортогональной к ней, и собственные вектора матрицы $\begin{pmatrix} \overline{\mathbb{D}x}_n & \overline{Cov(x,y)}_n \\ \overline{Cov(x,y)}_n & \overline{Dy}_n \end{pmatrix}$

Все функции прописаны в файле `func.R`, перед началом работы нужно запустить этот файл. Для построения следующих графиков нам понадобится подключить пакет `ggplot2`, хотя для работы самих функций он не обязателен.

Примеры работы программы:

Выборка нормального распределения Построим выборку нормального распределения на плоскости и найдём для неё наилучшую прямую (красным цветом) и прямую, отвечающую второму собственноу числу "ковариационной матрице"выборки:

```
dat <- Gauss(1,1,0.7,1000)
S <- SumInf(dat)

library(ggplot2)
ggplot(dat, aes(x=xvar, y=yvar)) +
  geom_point(shape=1)+
  geom_abline(intercept = (S$cx)[1], slope = (S$cy)[1] ,color="red")+
  geom_abline(intercept = (S$cx)[2], slope = (S$cy)[2] ,color="green")
```

Выборка `GaussB`$(\sigma_X, \sigma_Y, \rho, \nu, \mu, N)$

```
dat <- GaussB(1,1,0.1,0.1,1000)
S <- SumInf(dat)

library(ggplot2)
ggplot(dat, aes(x=xvar, y=yvar)) +
  geom_point(shape=1)+
  geom_abline(intercept = (S$cx)[1], slope = (S$cy)[1] ,color="red")+
  geom_abline(intercept = (S$cx)[2], slope = (S$cy)[2] ,color="green")
```

Оценка сходимости вектора наилучшей прямой

```
N=10000
u=rep(0,N)
sx=1
sy=1
r=0.5
if(sx == sy){
  tg=c(1,-1)
```

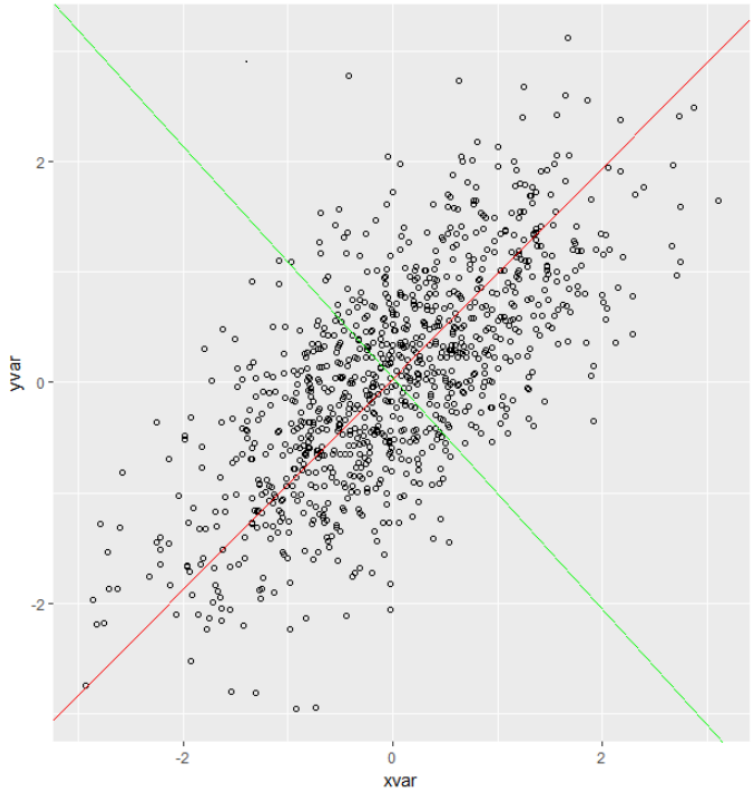


Рис. 3: Выборка из 1000 точек с помощью `Gauss`(1,1,0.7 1000)

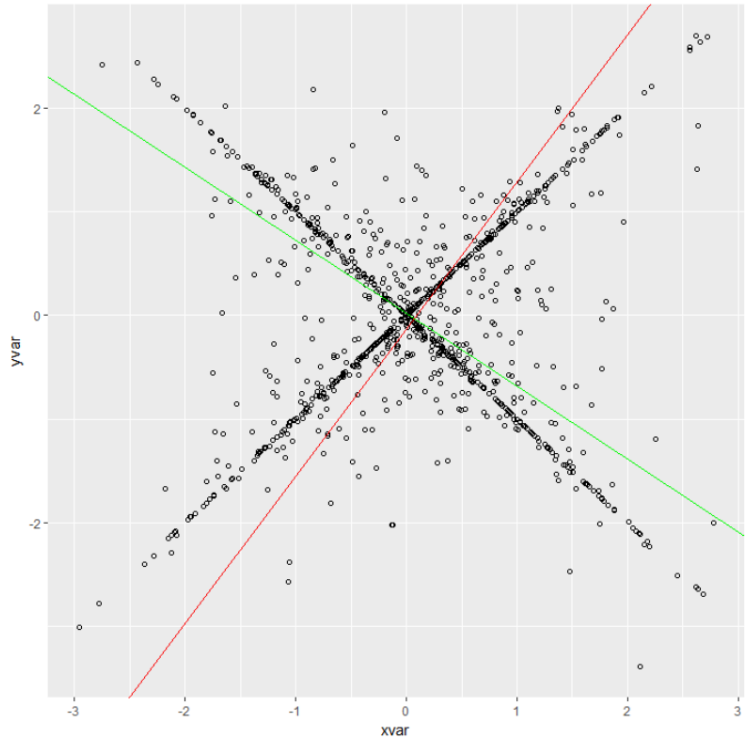
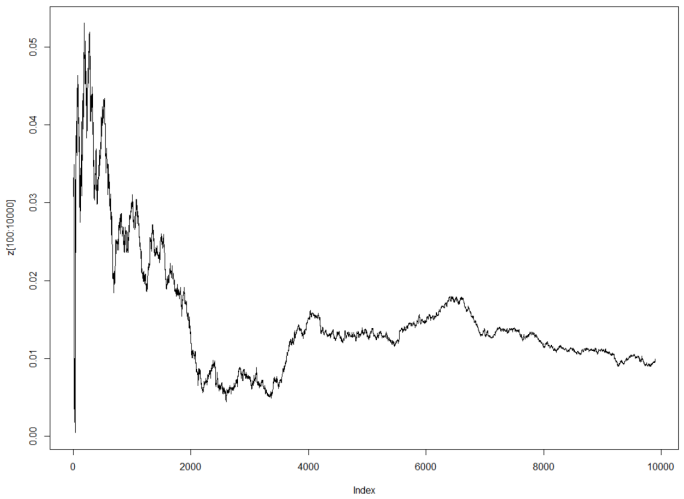
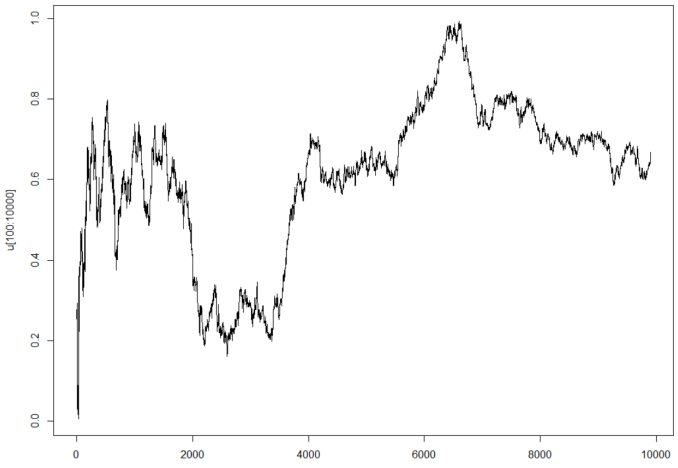


Рис. 4: Выборка из 1000 точек с помощью `GaussB`(1,1,0.1,0.1,1000)



$||u_n - u_0||_2$



$||u_n - u_0||_2\sqrt{\frac{n}{\ln\ln n}}$

Рис. 5: Расстояние между единичным вектором предельной прямой и единичным вектором наилучшей прямой

```
}else{
tg2=(2*sx*sy*r)/(sx^2-sy^2)
if(r>0){
  tg=c(-1/tg2+sqrt(1+1/tg2^2),-1/tg2-sqrt(1+1/tg2^2))
}else{
  tg=c(-1/tg2-sqrt(1+1/tg2^2),-1/tg2+sqrt(1+1/tg2^2))
}
}
costg=1/sqrt(1+tg^2)
sintg=tg/sqrt(1+tg^2)

d <- Gauss(sx,sy,r,N+1)
z=rep(0,N)
for(n in 2:(N+1))
  {
  dat <- SumInf(head(d,n))
  z[n-1]=sqrt( (((dat)$evp)[1] - costg[1])^2+(((dat)$evp)[2]-sintg[1])^2)
  u[n-1]=z[n-1]*sqrt((n)/(log(log(max(exp(1),n)))))
}
plot(z[100:10000],type="l")
```

Часть III
Приложение

4.1 Регрессии

Пусть на плоскости есть n точек: $\{(x_i, y_i)\}_{i=1}^n$. Будем искать прямую $\widehat{a_x}x+\widehat{b_x}$, то есть регрессию y на x, такую, что $R(\widehat{a_x}, \widehat{b_x}) = \sum\limits_{i=1}^{n}(y_i - \widehat{a_x}x_i - \widehat{b_x})^2$.

Чтобы найти минимум R, вычислим частные производные по $\widehat{a_x}$ и $\widehat{b}$:

$$\begin{cases} \frac{\partial R}{\partial \widehat{a_x}} = -2\sum\limits_{k=1}^{n} 2x_k(y_k - \widehat{a_x}x_k - \widehat{b_x}) = 0 \\ \frac{\partial R}{\partial \widehat{b_x}} = -2\sum\limits_{k=1}^{n} 2(y_k - \widehat{a_x}x_k - \widehat{b_x}) = 0 \end{cases} \tag{32}$$

$$\begin{cases} \widehat{a_x}\overline{x^2}_n + \widehat{b_x}\overline{x}_n = \overline{xy}_n \\ \widehat{a_x}\overline{x}_n + \widehat{b_x} = \overline{y}_n \end{cases} \tag{33}$$

$$\begin{cases} \widehat{a_x} = \frac{\overline{xy}_n - \overline{x}_n\overline{y}_n}{\overline{x^2}_n - \overline{x}_n^2} \\ \widehat{b_x} = \overline{y}_n - \widehat{a_x}\overline{x}_n \end{cases} \tag{34}$$

При $\overline{x^2}_n = \overline{x}_n^2$, коэффициент $\widehat{a_x}$ не определён. Но в этом случае $x_i = Const$ по неравенству Коши-Буняковского и соответственно все точки лежат на одной вертикальной прямой. Ясно, что в таком случае прямой такого вида не сущетвует.

Аналогично, если мы ищем прямую вида $\widehat{a_y}y + \widehat{b_y}$, или x на y, то уравнения будут:

$$\begin{cases} \widehat{a_y} = \frac{\overline{xy}_n - \overline{x}_n\overline{y}_n}{\overline{y^2}_n - \overline{y}_n^2} \\ \widehat{b_y} = \overline{x}_n - \widehat{a_y}\overline{y}_n \end{cases} \tag{35}$$

4.2 Наименее уклоняющаяся прямая

Пусть на плоскости есть n точек: $\{(x_i, y_i)\}_{i=1}^n$. Будем искать наименее уклоняющуюся от них прямую. Для этого, повернём всю плоскость на угол φ, сдвинем на a по оси Ox и вычислим сумму квадратов расстояний до оси x, то есть:

$$x'_k = \cos\varphi\, x_k - \sin\varphi\, y_k \tag{36}$$

$$\sum_{k=1}^{n}(x'_k - a)^2 = R^2 \tag{37}$$

Будем минимизировать (37) по a и φ.

$$\begin{cases} \frac{\partial R}{\partial a} = -\sum\limits_{k=1}^{n} 2(\cos\varphi\, x_k - \sin\varphi\, y_k - a) = 0 \\ \frac{\partial R}{\partial \varphi} = \sum\limits_{k=1}^{n} -2(\sin\varphi\, x_k + \cos\varphi\, y_k)(\cos\varphi\, x_k - \sin\varphi\, y_k - a) = 0 \end{cases} \tag{38}$$

Введём обозначения:

$$\begin{aligned}\bar{x} &= \frac{\sum\limits_{k=1}^{n} x_k}{n} \\ \overline{Dx} &= \overline{x^2} - \bar{x}^2 \\ \overline{Cov(x,y)} &= \overline{xy} - \bar{x}\bar{y}\end{aligned} \tag{39}$$

Раскроем (37)

$$\sum_{k=1}^{n}(\cos\varphi\, x_k - \sin\varphi\, y_k - a)^2 = n\left(\cos^2\varphi\overline{Dx} + \sin^2\varphi\overline{Dy} + 2\cos\varphi\sin\varphi\overline{Cov(x,y)}\right) = R^2 \tag{40}$$

или

$$R^2 = \frac{n}{2}\left(\left(\overline{Dx} + \overline{Dy}\right) + \cos 2\varphi\left(\overline{Dx} - \overline{Dy}\right) + 2\sin 2\varphi\overline{Cov(x,y)}\right) \tag{41}$$

Минимизируем последние два слагаемых в 41:

$$\cos 2\varphi\left(\overline{Dx} - \overline{Dy}\right) + 2\sin 2\varphi\overline{Cov(x,y)} \tag{42}$$

Минимум будет равен $\sqrt{\left(\overline{Dx} - \overline{Dy}\right)^2 + 4\overline{Cov(x,y)}^2}$ и достигаться при

$$\sin 2\varphi = -2\frac{2\overline{Cov(x,y)}}{\sqrt{\left(\overline{Dx} - \overline{Dy}\right)^2 + 4\overline{Cov(x,y)}^2}} \tag{43}$$

Или

$$\tan 2\varphi = -2\frac{\overline{xy} - \bar{x}\bar{y}}{(\overline{x^2} - \bar{x}^2) - (\overline{y^2} - \bar{y}^2)} \tag{44}$$

Соответственно $a = \cos\varphi\,\bar{x} - \sin\varphi\,\bar{y}$.

Чтобы получить искомую прямую, нужно снова прямую $x' = a$ повернуть обратно на φ то есть:

$$\cos\varphi\, x - \sin\varphi\, y = \cos\varphi\,\bar{x} - \sin\varphi\,\bar{y} \tag{45}$$

Или

$$(x - \bar{x}) = \tan\varphi(y - \bar{y}) \tag{46}$$

где φ угол наклона прямой относительно Oy

В более привычном виде,

$$\begin{aligned}&\tan\psi(x - \bar{x}) = (y - \bar{y}) \\ &\tan 2\psi = \frac{2\overline{Cov(x,y)}}{\overline{Dx} - \overline{Dy}} \\ &\sin 2\psi = -2\frac{2\overline{Cov(x,y)}}{\sqrt{\left(\overline{Dx} - \overline{Dy}\right)^2 + 4\overline{Cov(x,y)}^2}}\end{aligned} \tag{47}$$

Также, как будет доказано позже вектора $\begin{pmatrix}\cos\varphi \\ \sin\varphi\end{pmatrix}$ и $\begin{pmatrix}-\sin\varphi \\ \cos\varphi\end{pmatrix}$ будут собственными векторами $\begin{pmatrix}\overline{\mathbb{D}x} & \overline{Cov(x,y)} \\ \overline{Cov(x,y)} & \overline{Dy}\end{pmatrix}$, $\begin{pmatrix}\cos\varphi \\ \sin\varphi\end{pmatrix}$, то есть вектор направления кратчайшей прямой, будет собственным при наибольшем собственном значении. Случай, когда такого вектора нет: $\overline{Cov(x,y)} = 0$, $\overline{\mathbb{D}x} = \overline{Dy}$.

В этом случае, $\sqrt{\left(\overline{Dx} - \overline{Dy}\right)^2 + 4\overline{Cov(x,y)}^2} = 0$ и R^2 не меняется при выборе ψ

4.3 О двумерном гауссовом векторе

Пусть (X, Y) невырожденный гауссов вектор с нулевым средним, т.е.

$$p_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2} - 2\frac{\rho xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}\right)\right) \tag{48}$$

$\sigma_X, \sigma_Y \in \mathbb{R}_+$, $\rho \in (-1,1)$. Или, если $R = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$, $|R| = \sigma_X^2\sigma_Y^2(1-\rho^2)$

$$p_{XY}(\mathbf{x}) = \frac{1}{2\pi\sqrt{|R|}} \exp\left(-\frac{1}{2}\left(\mathbf{x}, R^{-1}\mathbf{x}\right)\right) \tag{49}$$

Пусть $\mathbf{z}$ -стандартный гауссовский вектор, т.е.
$p_{\mathbf{z}}(\mathbf{x}) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}(\mathbf{x},\mathbf{x})\right)$
И пусть A невыродженная матрица и $|A| > 0$. Тогда плотность вектора $A\mathbf{z}$ равна $p_{A\mathbf{z}} = \frac{1}{2\pi|A|}\exp\left(-\frac{1}{2}(\mathbf{x}, R^{-1}\mathbf{x})\right)$, где $R = AA^\top$. Легко видеть, что $\sqrt{\det R} = |A|$ и тогда она превращается в формулу (49)
Буде искать A в виде $A = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}\begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$, $(\alpha \geqslant \beta > 0)$
Тогда $AA^\top = R$ распиывется как:

$$\begin{cases} \alpha^2\cos^2\varphi + \beta^2\sin^2\varphi = \sigma_X^2 \\ \alpha^2\sin^2\varphi + \beta^2\cos^2\varphi = \sigma_Y^2 \\ \cos\varphi\sin\varphi(\alpha^2 - \beta^2) = \rho\sigma_X\sigma_Y \end{cases} \tag{50}$$

Или

$$\begin{cases} \alpha^2 + \beta^2 = \sigma_X^2 + \sigma_Y^2 \\ (\alpha^2 - \beta^2)\cos 2\varphi = \sigma_X^2 - \sigma_Y^2 \\ \sin 2\varphi(\alpha^2 - \beta^2) = 2\rho\sigma_X\sigma_Y \end{cases} \tag{51}$$

Откуда:

$$\begin{cases} \alpha^2 + \beta^2 = \sigma_X^2 + \sigma_Y^2 \\ (\alpha^2 - \beta^2)\cos 2\varphi = \sigma_X^2 - \sigma_Y^2 \\ \tan 2\varphi = \frac{2\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2} \end{cases} \tag{52}$$

Причём: $tg(2\varphi)$ даёт 2 ортогональные прямые с углами φ и $\varphi + \frac{\pi}{2}$, соответственно. Но $\cos 2\varphi = -\cos(2\varphi + \frac{\pi}{2})$ и, по выбору $\alpha \geqslant \beta$
$\varphi = \frac{1}{2}\arctan\frac{2\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}$ при $\frac{2\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2} > 0$ и $\varphi = \frac{1}{2}\arctan\frac{2\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2} + \frac{pi}{2}$ при $\frac{2\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2} < 0$
При $\rho = 0$ $\varphi = 0$ при $\sigma_X > \sigma_Y$, и $\rho = \frac{\pi}{2}$ при $\sigma_X < \sigma_Y$. При $\sigma_X = \sigma_Y$ и $\rho > 0$ $\varphi = \frac{\pi}{4}$, при $\rho < 0$ $\varphi = -\frac{\pi}{4}$. Если $\sigma_X = \sigma_Y$ и $\rho = 0$, то возможны оба поворота.
α^2 и β^2 выводятся из первых двух уравнений:

$$\begin{cases} \alpha^2\cos^2\varphi + \beta^2\sin^2\varphi = \sigma_X^2 \\ \alpha^2\sin^2\varphi + \beta^2\cos^2\varphi = \sigma_Y^2 \end{cases} \tag{53}$$

$$\begin{aligned} \alpha^2 &= \frac{\sigma_X^2\cos^2\varphi - \sigma_Y^2\sin^2\varphi}{\cos^2\varphi - \sin^2\varphi} \\ \beta^2 &= \frac{\sigma_Y^2\cos^2\varphi - \sigma_X^2\sin^2\varphi}{\cos^2\varphi - \sin^2\varphi} \end{aligned} \tag{54}$$

при $\sigma_X = \sigma_Y$ и $\rho \neq 0$, собственные числа матрицы можно вычислить явно, так как матрица имеет простой вид:

$$\sigma_X^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{55}$$

и её собственные числа равны $\lambda_{1,2} = \sigma_X^2(1 \pm \rho)$.

Таким образом, оси Ox и Oy под углом в φ. Будем их называть главными направлениями $\mathbf{z}$

Заметим, что если $R = AA^\top = U_\varphi \begin{pmatrix} \alpha^2 & 0 \\ 0 & \beta^2 \end{pmatrix} U_{-\varphi}$, то собственные числа R совпадают с такими у $\begin{pmatrix} \alpha^2 & 0 \\ 0 & \beta^2 \end{pmatrix}$, а собственные вектора равны $U_\varphi \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix}$ и $U_\varphi \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sin\varphi \\ \cos\varphi \end{pmatrix}$

Это в точности главные направления нашего гауссова вектора $\mathbf{z}$, причём α, при $\alpha \neq \beta$ соответствует именно $\begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix}$

5 Литература

Список литературы

[1] Ю.С.Харин, В.И.Малюгин, В.Л. Кирлица, В.И.Лобач, Г.А.Хацкевич *Основы иммитационного и статистического моделирования*, Дизайн-ПРО, 1997 г.

[2] Прасолов В. В. *Задачи и теоремы линейной алгебры.*, М.: Наука, 1996. — 304 с. — ISBN 5-02-014727-3.

[3] А.В.Булинский, А.Н. Ширяев *Теория случайных процессов*, Физматлит, 2005 г.