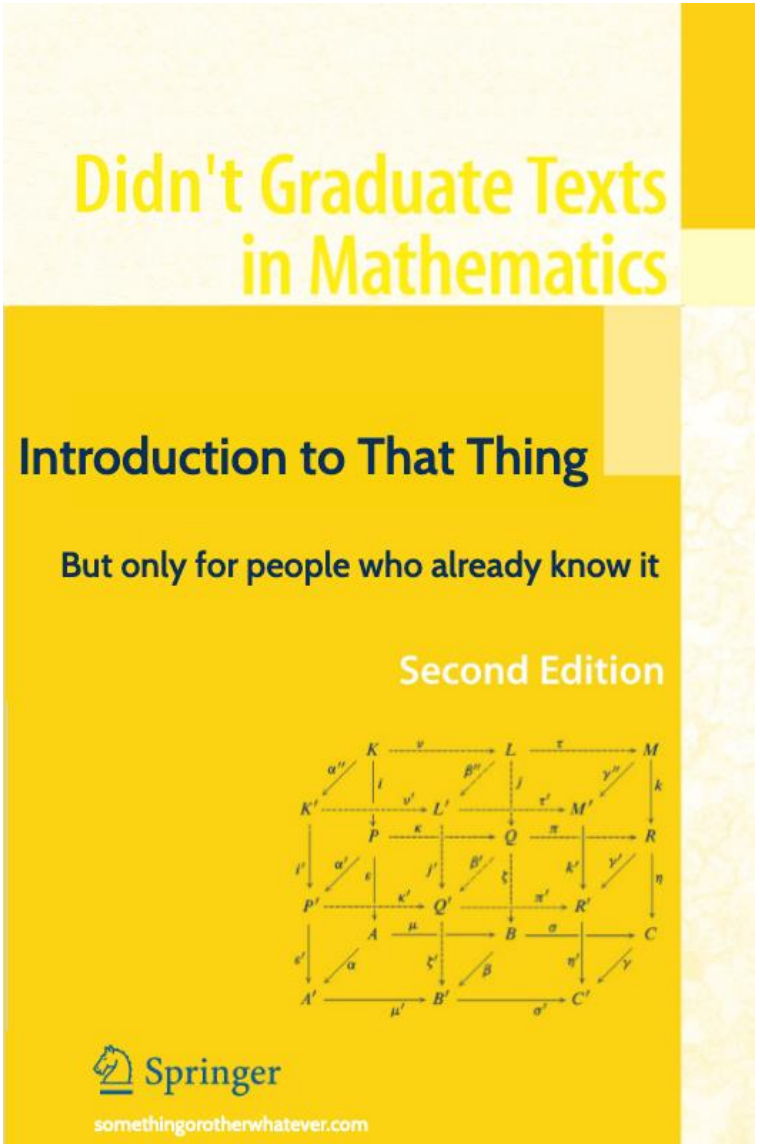
Введение в гауссовские процессы на Python

Максим Николаев
ПОМИ, 4 октября 2025

Дисклеймер

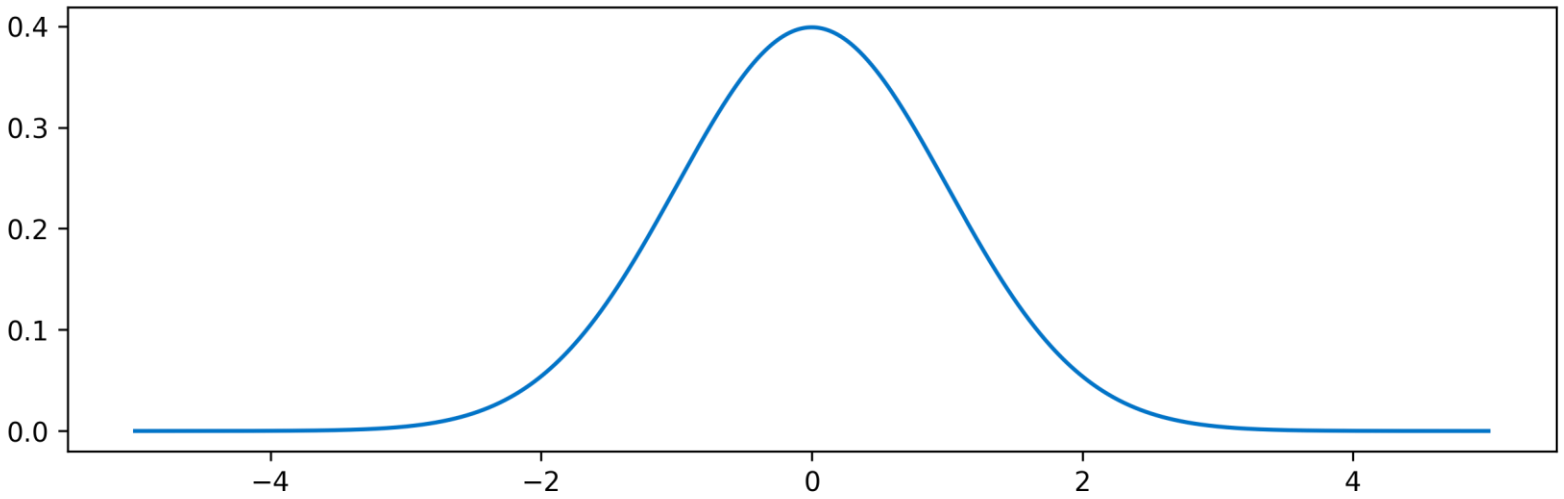

Didn't Graduate Texts in Mathematics
Introduction to That Thing
But only for people who already know it
Second Edition
Springer
somethingorotherwhatever.com

Одномерное нормальное распределение

Одномерное нормальное распределение

Говорят, что случайная величина X имеет **стандартное нормальное распределение** $\mathcal{N}(0,1)$, если ее плотность равна

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$



Шаг в сторону

Почему интеграл от e^{-x^2} равен $\sqrt{\pi}$?

Шаг в сторону

Почему интеграл от e^{-x^2} равен $\sqrt{\pi}$?

Пусть $I \coloneqq \int_{\mathbb{R}} e^{-x^2} dx$. Тогда

$$I^2 = \left(\int_{\mathbb{R}} e^{-x^2} dx\right)\left(\int_{\mathbb{R}} e^{-y^2} dy\right) = \int_{\mathbb{R}^2} e^{-(x^2+y^2)} dxdy\,.$$

Шаг в сторону

Почему интеграл от e^{-x^2} равен $\sqrt{\pi}$?

Пусть $I \coloneqq \int_{\mathbb{R}} e^{-x^2} dx$. Тогда

$$I^2 = \left(\int_{\mathbb{R}} e^{-x^2} dx\right)\left(\int_{\mathbb{R}} e^{-y^2} dy\right) = \int_{\mathbb{R}^2} e^{-(x^2+y^2)} dx dy\,.$$

По теореме Пифагора $x^2 + y^2 = \rho^2$, где (ρ, ϕ) — полярные координаты (x, y). Переходя в полярные координаты, получаем

$$I^2 = \int_0^{2\pi} \int_0^{+\infty} e^{-\rho^2} \cdot \rho \cdot d\rho d\phi = \int_0^{2\pi} d\phi \cdot \int_0^{+\infty} \frac{1}{2} e^{-\rho^2} d\rho^2 = 2\pi \cdot \frac{1}{2} = \pi.$$

Нормальное распределение и геометрия

Заметим, что если случайные величины X и Y независимы, то их совместная плотность равна произведению плотностей:

$$p(x, y) = p(x)p(y).$$

В случае нормального распределения этот факт приобретает удивительную геометрическую интерпретацию — совместная плотность X и Y зависит только от нормы вектора $z = (x, y)$:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} = \frac{1}{2\pi} e^{-\frac{1}{2}|z|^2}.$$

Отсюда, например, следует, что если X и Y имеют распределение $\mathcal{N}(0,1)$ и независимы, то $X\cos\alpha - Y\sin\alpha$ и $X\sin\alpha + Y\cos\alpha$ тоже имеют распределение $\mathcal{N}(0,1)$ и независимы.

Многомерное нормальное распределение

Многомерное нормальное распределение

Случайный вектор $Z = (Z_1, \dots, Z_m)^{\mathrm{T}}$, у которого все Z_i независимы и имеют распределение $\mathcal{N}(0,1)$, называется **стандартным гауссовским вектором**.

Случайный вектор $X \in \mathbb{R}^n$ называется **гауссовским вектором** или **нормальным вектором**, если существуют такие $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$ и стандартный гауссовский $Z \in \mathbb{R}^m$, что $X = AZ + b$.

Распределение нормального вектора обозначается через $\mathcal{N}(\mu, \Sigma)$ или $\mathcal{N}_n(\mu, \Sigma)$, если хочется подчеркнуть размерность, где $\mu = \mathbb{E}X$, $\Sigma = \mathrm{cov}(X) \coloneqq \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^{\mathrm{T}}$.

Ясно, что $Z \sim \mathcal{N}(0, I)$, а $X \sim \mathcal{N}\left(b, AA^{\mathrm{T}}\right)$.

значок «имеет распределение»

единичная матрица

столбец × строка

Многомерное нормальное распределение

Мы выяснили, что если $X = AZ + b$, то $X \sim \mathcal{N}\left(b, AA^{\mathrm{T}}\right)$, и это дает нам понятный способ генерации X, если мы умеем генерировать $\mathcal{N}(0,1)$: сначала генерируем Z, потом применяем к нему преобразование $Z \to AZ + b$.

Что делать в случае, если мы хотим сгенерировать $X \sim \mathcal{N}(\mu, \Sigma)$?

Многомерное нормальное распределение

Мы выяснили, что если $X = AZ + b$, то $X \sim \mathcal{N}\left(b, AA^{\mathrm{T}}\right)$, и это дает нам понятный способ генерации X, если мы умеем генерировать $\mathcal{N}(0,1)$: сначала генерируем Z, потом применяем к нему преобразование $Z \to AZ + b$.

Что делать в случае, если мы хотим сгенерировать $X \sim \mathcal{N}(\mu, \Sigma)$? Можно взять «корень из Σ», получить равенство $X = \Sigma^{1/2}Z + \mu$ и воспользоваться предыдущим способом!
Формально, существует много корней из матрицы, но для определенности можно брать матрицу

$$\Sigma^{1/2} \coloneqq U\Lambda^{1/2}U^{\mathrm{T}},$$

где $\Sigma = U\Lambda U^{\mathrm{T}}$ это диагонализация матрицы Σ, а Λ это матрица собственных чисел Σ.
Стоит отметить, что это не самый эффективный способ генерации, и на практике его никто не использует.

Плотность нормального распределения

Пусть $Z \in \mathbb{R}^m$ это стандартный гауссовский вектор. Тогда

$$p_Z(z) = \prod_{i=1}^{m} p(z_i) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} = \frac{1}{(2\pi)^{m/2}} e^{-\frac{1}{2}\sum_i z_i^2} = \frac{1}{(2\pi)^{m/2}} e^{-\frac{1}{2}|z|^2}.$$

Если $X = AZ + b$ и A обратима (в частности, $m = n$), то $Z = A^{-1}(X - b)$ и

$$p_X(x) = \det(A^{-1})\, p_Z\big(A^{-1}(X-b)\big) = \frac{1}{(2\pi)^{m/2}\det(A)} e^{-\frac{1}{2}|A^{-1}(x-b)|^2} = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}.$$

Если Σ необратима, то ситуация чуть сложнее — надо разбираться, что такое псевдообратная матрица.

Независимость и некоррелированность

Пусть (X, Y) — гауссовский вектор, состоящий из двух векторов $X \sim \mathcal{N}_n(\mu_X, \Sigma_X)$ и $Y \sim \mathcal{N}_m(\mu_Y, \Sigma_Y)$.

Пусть также X и Y некоррелированны, то есть $\operatorname{cov}(X, Y) = 0$. Тогда для $V = (X, Y)$ выполнено

$$\mathbb{E}V = (\mathbb{E}X, \mathbb{E}Y)^{\mathrm{T}}, \operatorname{cov}(V) = \Sigma_V = \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix},$$

$$p(x, y) = p(v) = \frac{1}{\sqrt{2\pi}^{n+m}\sqrt{\det(\Sigma_V)}} e^{-\frac{1}{2}(v-\mathbb{E}V)^{\mathrm{T}}\Sigma_V^{-1}(v-\mathbb{E}V)}$$

$$= \frac{1}{\sqrt{2\pi}^{n}\sqrt{\det(\Sigma_X)}} e^{-\frac{1}{2}(x-\mu_X)^{\mathrm{T}}\Sigma_X^{-1}(x-\mu_X)} \frac{1}{\sqrt{2\pi}^{m}\sqrt{\det(\Sigma_Y)}} e^{-\frac{1}{2}(y-\mu_Y)^{\mathrm{T}}\Sigma_Y^{-1}(y-\mu_Y)} = p(x)p(y).$$

Получили, что из некоррелированности X и Y следует их независимость.

Условное нормальное распределение

Пусть (X, Y) это гауссовский вектор, состоящий из двух векторов X и Y:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}\right).$$

Тогда распределение Y при условии, что $X = x$, тоже является нормальным

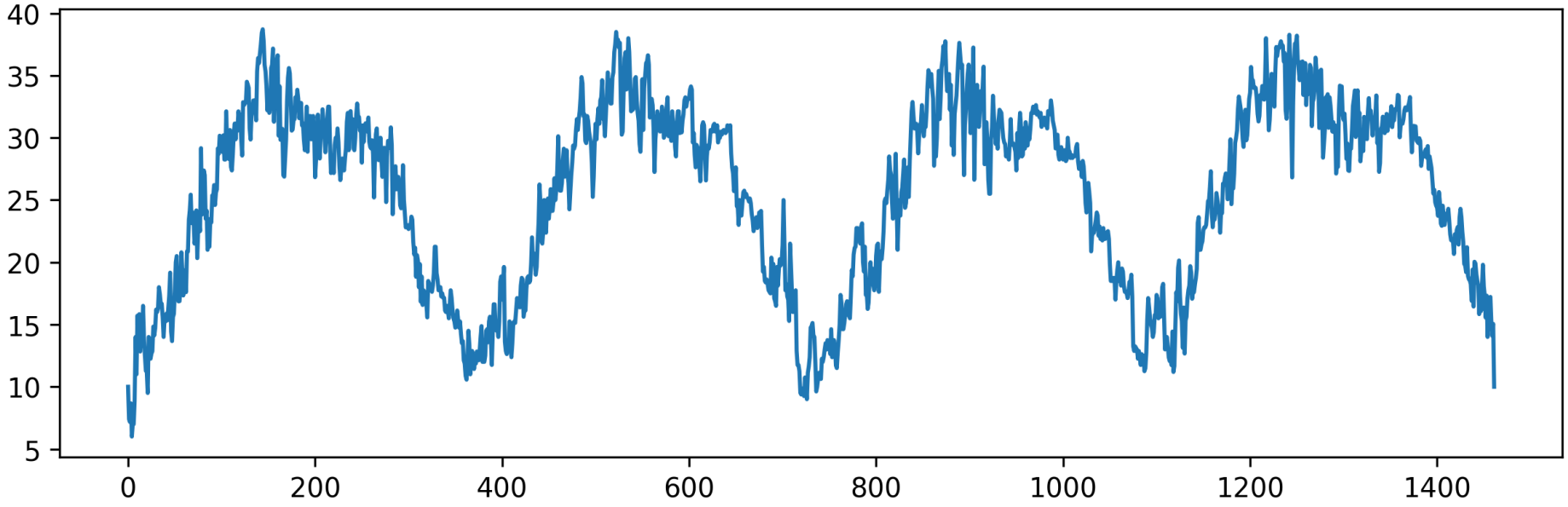
$$Y|(X = x) \sim \mathcal{N}(\mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X), \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}).$$

Докажем мы этот факт немного позже.

Анализ временных рядов

Пусть (Y_p, Y_f) это нормальный вектор, где Y_p это «прошлое», а Y_f это «будущее». Мы знаем Y_p и хотим что-то понять про Y_f. Если мы понимаем, как устроена ковариация между разными моментами времени, мы можем посчитать условное распределение.

В качестве примера рассмотрим среднедневную температуру в Дели:



Анализ временных рядов

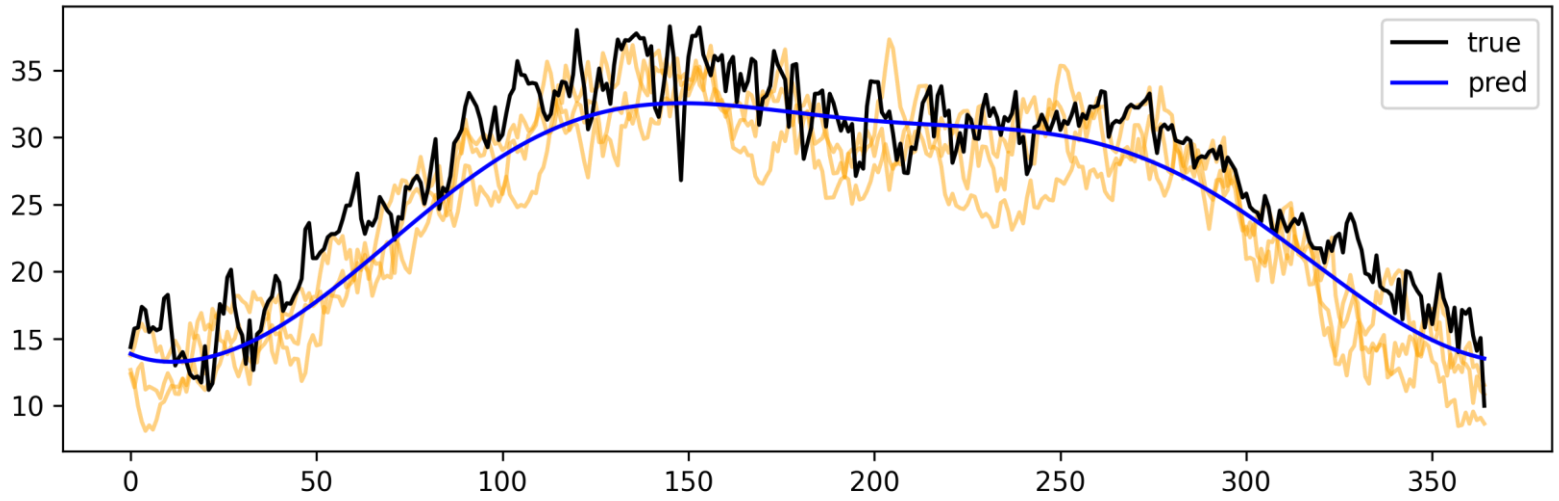
Попробуем предсказать температуру в четвертом году по третьему. Для этого зададим матрицу ковариации с помощью **ковариационной функции**

$$K(t_1, t_2) = 10 + 44.5\cos(2\pi|t_1 - t_2|/365) + 4\cos(2\pi|t_1 - t_2|/182.5) + 4.5e^{-|t_1-t_2|/6},$$

где t_1 и t_2 это номера дней. После этого предскажем Y_f с помощью

$$\Sigma_{Y_f Y_p}\Sigma_{Y_p}^{-1}Y_p,$$

где Y_p это вектор температур за третий год. Получается так:



Очень много линала

Минимизация MSE: проекции

Пусть Y это случайный вектор, а $\mathfrak{X}$ это семейство случайных векторов той же размерности с конечными вторыми моментами: $\mathbb{E}XX^{\mathrm{T}} < \infty$, $X \in \mathfrak{X}$.

Случайная величина $\hat{Y}$ называется **проекцией** или **L_2-проекцией** Y на $\mathfrak{X}$, если $\hat{Y} \in \mathfrak{X}$ и

$$\mathbb{E}|Y - \hat{Y}|^2 \leq \mathbb{E}|Y - X|^2$$

для всех $X \in \mathfrak{X}$. Например, $\mathbb{E}X$ это проекция Y на семейство констант.

Минимизация MSE: проекции

Часто $\mathfrak{X}$ является **линейным пространством**, то есть $\alpha_1 X_1 + \alpha_2 X_2 \in \mathfrak{X}$ для $\alpha_1, \alpha_2 \in \mathbb{R}$ и $X_1, X_2 \in \mathfrak{X}$.

В этом случае $\hat{Y}$ является проекцией Y тогда и только тогда, когда $Y - \hat{Y}$ **ортогонально** $\mathfrak{X}$: в смысле скалярного произведения $\langle Y, X \rangle \coloneqq \mathbb{E} Y^{\mathrm{T}} X$. Более формально:

Теорема

Пусть $\mathfrak{X}$ это линейное пространство случайных величин с конечными вторыми моментами.

- $\hat{Y}$ является проекцией Y на $\mathfrak{X} \Leftrightarrow \hat{Y} \in \mathfrak{X}$ и $\langle Y - \hat{Y}, X \rangle \coloneqq \mathbb{E}\left(Y - \hat{Y}\right)^{\mathrm{T}} X = 0$, для всех $X \in \mathfrak{X}$.
- Любые две проекции совпадают.
- Если $\mathfrak{X}$ содержит константы, то $\mathbb{E} Y = \mathbb{E} \hat{Y}$ и $\operatorname{cov}\left(Y - \hat{Y}, X\right) = 0$ для всех $X \in \mathfrak{X}$.

Минимизация MSE: проекции

- $\hat{Y}$ является проекцией Y на $\mathfrak{X} \Leftrightarrow \hat{Y} \in \mathfrak{X}$ и $\langle Y - \hat{Y}, X \rangle \coloneqq \mathbb{E}\left(Y - \hat{Y}\right)^{\mathrm{T}} X = 0$, для всех $X \in \mathfrak{X}$.
- Любые две проекции совпадают.

$$|Y - X|^2 = \left|Y - \hat{Y} + \hat{Y} - X\right|^2 = \left|Y - \hat{Y}\right|^2 + 2\langle Y - \hat{Y}, \hat{Y} - X \rangle + \left|\hat{Y} - X\right|^2.$$

Если $Y - \hat{Y}$ ортогонально $\mathfrak{X}$, то центральное слагаемое равно нулю, а равенство достигается только если $\left|\hat{Y} - X\right|^2 = 0$, то есть если $\hat{Y} = X$.

Таким образом «ортогональность» $\Longrightarrow$ «единственная проекция».

Минимизация MSE: проекции

- $\hat{Y}$ является проекцией Y на $\mathfrak{X} \Leftrightarrow \hat{Y} \in \mathfrak{X}$ и $\langle Y - \hat{Y}, X \rangle := \mathbb{E}\left(Y - \hat{Y}\right)^{\mathrm{T}} X = 0$, для всех $X \in \mathfrak{X}$.
- Любые две проекции совпадают.

В обратную сторону:

Если $\hat{Y}$ это проекция, то для всех $\alpha \in \mathbb{R}$ и $X \in \mathfrak{X}$ выполнено:

$$\left|Y - \hat{Y}\right|^2 \leq \left|Y - \hat{Y} - \alpha X\right|^2 = \left|Y - \hat{Y}\right|^2 - 2\alpha\langle Y - \hat{Y}, X\rangle + \alpha^2 |X|^2$$

Иначе говоря,

$$-2\alpha\langle Y - \hat{Y}, X\rangle + \alpha^2 |X|^2 \geq 0.$$

Минимизация MSE: проекции

- $\hat{Y}$ является проекцией Y на $\mathfrak{X} \Leftrightarrow \hat{Y} \in \mathfrak{X}$ и $\langle Y - \hat{Y}, X \rangle := \mathbb{E}\left(Y - \hat{Y}\right)^{\mathrm{T}} X = 0$, для всех $X \in \mathfrak{X}$.
- Любые две проекции совпадают.

В обратную сторону:

Если $\hat{Y}$ это проекция, то для всех $\alpha \in \mathbb{R}$ и $X \in \mathfrak{X}$ выполнено:

$$\left|Y - \hat{Y}\right|^2 \leq \left|Y - \hat{Y} - \alpha X\right|^2 = \left|Y - \hat{Y}\right|^2 - 2\alpha\langle Y - \hat{Y}, X\rangle + \alpha^2 |X|^2$$

Иначе говоря,

$$-2\alpha\langle Y - \hat{Y}, X\rangle + \alpha^2 |X|^2 \geq 0.$$

В то же время парабола $\alpha \mapsto |X|^2\alpha^2 - 2\langle Y - \hat{Y}, X\rangle\alpha$ неотрицательна на $\mathbb{R}$ только если

$$\langle Y - \hat{Y}, X\rangle = 0.$$

Минимизация MSE: проекции

– Если $\mathfrak{X}$ содержит константы, то $\mathbb{E}Y = \mathbb{E}\hat{Y}$ и $\operatorname{cov}\left(Y - \hat{Y}, X\right) = 0$ для всех $X \in \mathfrak{X}$.

В этом случае $\mathbb{E}\left(Y - \hat{Y}\right)^{\mathrm{T}} c = 0$, откуда $\mathbb{E}Y = \mathbb{E}\hat{Y}$.

$$\operatorname{cov}\left(Y - \hat{Y}, X\right) = \mathbb{E}\left(Y - \hat{Y}\right)X^{\mathrm{T}} - \mathbb{E}\left(Y - \hat{Y}\right)\mathbb{E}X^{\mathrm{T}} = 0.$$

Линейное предсказание в одномерном случае

Пусть X и Y случайные величины. Рассмотрим пространство $\mathfrak{X} = \{\alpha + \beta X \mid \alpha, \beta \in \mathbb{R}\}$. В этом случае наилучшее предсказание Y с помощью $\alpha + \beta X$ имеет вид

$$\mathbb{E}Y + \rho(X, Y)\frac{\mathrm{std}(Y)}{\mathrm{std}(X)}(X - \mathbb{E}X).$$

Линейное предсказание в многомерном случае

Пусть (X, Y) это случайный вектор состоящий из двух векторов $X \in \mathbb{R}^n$ и $Y \in \mathbb{R}^m$. Чему равно наилучшее линейное предсказание Y по X? Иначе говоря, при каких $A \in \mathbb{R}^{m \times n}$ и $b \in \mathbb{R}^m$ выражение $|Y - AX - b|^2$ минимально?

Как мы уже знаем, оптимальные A^* и b^* должны удовлетворять условию ортогональности:

$$\mathbb{E}(Y - A^*X - b^*)^{\mathrm{T}}(AX + b) = 0$$

для всех A и b.

Линейное предсказание в многомерном случае

Пусть (X, Y) это случайный вектор состоящий из двух векторов $X \in \mathbb{R}^n$ и $Y \in \mathbb{R}^m$. Чему равно наилучшее линейное предсказание Y по X? Иначе говоря, при каких $A \in \mathbb{R}^{m \times n}$ и $b \in \mathbb{R}^m$ выражение $|Y - AX - b|^2$ минимально?

Как мы уже знаем, оптимальные A^* и b^* должны удовлетворять условию ортогональности:

$$\mathbb{E}(Y - A^*X - b^*)^{\mathrm{T}}(AX + b) = 0$$

для всех A и b.

Ортогональность со всеми $b \in \mathbb{R}^m$ означает, что $\mathbb{E}(Y - A^*X - b^*) = 0$, то есть $b^* = \mathbb{E}Y - A^*\mathbb{E}X$.

Линейное предсказание в многомерном случае

Пусть (X, Y) это случайный вектор состоящий из двух векторов $X \in \mathbb{R}^n$ и $Y \in \mathbb{R}^m$. Чему равно наилучшее линейное предсказание Y по X? Иначе говоря, при каких $A \in \mathbb{R}^{m \times n}$ и $b \in \mathbb{R}^m$ выражение $|Y - AX - b|^2$ минимально?

Как мы уже знаем, оптимальные A^* и b^* должны удовлетворять условию ортогональности:

$$\mathbb{E}(Y - A^*X - b^*)^{\mathrm{T}}(AX + b) = 0$$

для всех A и b.

Ортогональность со всеми $b \in \mathbb{R}^m$ означает, что $\mathbb{E}(Y - A^*X - b^*) = 0$, то есть $b^* = \mathbb{E}Y - A^*\mathbb{E}X$.

Ортогональность со всеми AX означает, что $\mathbb{E}(Y - A^*X - b^*)X^{\mathrm{T}} = 0$. Получаем

$$\mathbb{E}\left((Y - \mathbb{E}Y) - A^*(X - \mathbb{E}X)X^{\mathrm{T}}\right) = 0 \Longrightarrow A^*\mathbb{E}(X - \mathbb{E}X)X^{\mathrm{T}} = \mathbb{E}(Y - \mathbb{E}Y)X^{\mathrm{T}} \Longrightarrow A^* = \mathrm{cov}(Y, X)\mathrm{cov}(X)^{-1}.$$

Таким образом, оптимальное линейное предсказание это $\mathbb{E}Y + \mathrm{cov}(Y, X)\mathrm{cov}(X)^{-1}(X - \mathbb{E}X)$.

Когда константа это часть вектора

Ради развлечения посмотрим, что будет, если $X' = (1, X) = (1, X_1, \ldots, X_n)^{\mathrm{T}}$. В этом случае приближение достаточно искать в виде AX'. Как и раньше, оптимальная A^* удовлетворяет условию

$$\mathbb{E}(Y - A^*X')X'^{\mathrm{T}} = 0 \Longrightarrow A^* = \mathbb{E}YX'^{\mathrm{T}}\left(\mathbb{E}X'X'^{\mathrm{T}}\right)^{-1}.$$

По идее, A^*X' должно быть равно $\mathbb{E}Y + \mathrm{cov}(Y, X)\mathrm{cov}(X)^{-1}(X - \mathbb{E}X)$. Как это увидеть?

Что такое $\mathbb{E}YX'^{\mathrm{T}}$ более менее понятно — это $(\mathbb{E}Y \quad \mathbb{E}YX^{\mathrm{T}})$. Но что делать с $\left(\mathbb{E}X'X'^{\mathrm{T}}\right)^{-1}$.

Формула обращения блочной матрицы

Пусть $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$, где $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{m \times n}$, $D \in \mathbb{R}^{m \times m}$. Можно ли выразить M^{-1} через A, B, C, D?

Пусть A обратима. Тогда методом Гаусса можно избавиться от C:

$$\begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & B \\ 0 & D - CA^{-1}B \end{pmatrix}.$$

А потом от B:

$$\begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix}.$$

Величина $H = D - CA^{-1}B$ называется **дополнением Шура** блока A и обозначается как M/A.

Формула обращения блочной матрицы

$$\begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix}.$$

Заметим, что $\begin{pmatrix} I & H \\ 0 & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -H \\ 0 & I \end{pmatrix}$ и аналогично с нижнетреугольной матрицей. Стало быть,

$$\begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix}\begin{pmatrix} A & 0 \\ 0 & M/A \end{pmatrix}\begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Тогда $\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1}$ равна

$$\begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix}\begin{pmatrix} A^{-1} & 0 \\ 0 & (M/A)^{-1} \end{pmatrix}\begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix} = \begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix}.$$

Когда константа это часть вектора

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix}.$$

Пусть $M = \mathbb{E}X'X'^{\mathrm{T}}$, где $X' = (1, X)$. Тогда $A = 1$, $C = B^{\mathrm{T}} = \mathbb{E}X$, $D = \mathbb{E}XX^{\mathrm{T}}$. Получаем

$$M/A = D - CA^{-1}B = \mathbb{E}XX^{\mathrm{T}} - \mathbb{E}X\mathbb{E}X^{\mathrm{T}} = \mathrm{cov}(X).$$

Откуда

$$\left(\mathbb{E}X'X'^{\mathrm{T}}\right)^{-1} = \begin{pmatrix} 1 + \mathbb{E}X^{\mathrm{T}}\mathrm{cov}(X)^{-1}\mathbb{E}X & -\mathbb{E}X^{\mathrm{T}} \cdot \mathrm{cov}(X)^{-1} \\ -\mathrm{cov}(X)^{-1}\mathbb{E}X & \mathrm{cov}(X)^{-1} \end{pmatrix}.$$

Когда константа это часть вектора

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix}.$$

Пусть $M = \mathbb{E}X'X'^{\mathrm{T}}$, где $X' = (1, X)$. Тогда $A = 1$, $C = B^{\mathrm{T}} = \mathbb{E}X$, $D = \mathbb{E}XX^{\mathrm{T}}$. Получаем

$$M/A = D - CA^{-1}B = \mathbb{E}XX^{\mathrm{T}} - \mathbb{E}X\mathbb{E}X^{\mathrm{T}} = \mathrm{cov}(X).$$

Откуда

$$\left(\mathbb{E}X'X'^{\mathrm{T}}\right)^{-1} = \begin{pmatrix} 1 + \mathbb{E}X^{\mathrm{T}}\mathrm{cov}(X)^{-1}\mathbb{E}X & -\mathbb{E}X^{\mathrm{T}} \cdot \mathrm{cov}(X)^{-1} \\ -\mathrm{cov}(X)^{-1}\mathbb{E}X & \mathrm{cov}(X)^{-1} \end{pmatrix}.$$

Тогда $\mathbb{E}YX'^{\mathrm{T}}\left(\mathbb{E}X'X'^{\mathrm{T}}\right)^{-1}$ это $\begin{pmatrix} \mathbb{E}Y & \mathbb{E}YX^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} 1 + \mathbb{E}X^{\mathrm{T}}\mathrm{cov}(X)^{-1}\mathbb{E}X & -\mathbb{E}X^{\mathrm{T}} \cdot \mathrm{cov}(X)^{-1} \\ -\mathrm{cov}(X)^{-1}\mathbb{E}X & \mathrm{cov}(X)^{-1} \end{pmatrix} =$

Когда константа это часть вектора

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix}.$$

Пусть $M = \mathbb{E}X'X'^{\mathrm{T}}$, где $X' = (1, X)$. Тогда $A = 1$, $C = B^{\mathrm{T}} = \mathbb{E}X$, $D = \mathbb{E}XX^{\mathrm{T}}$. Получаем

$$M/A = D - CA^{-1}B = \mathbb{E}XX^{\mathrm{T}} - \mathbb{E}X\mathbb{E}X^{\mathrm{T}} = \mathrm{cov}(X).$$

Откуда

$$\left(\mathbb{E}X'X'^{\mathrm{T}}\right)^{-1} = \begin{pmatrix} 1 + \mathbb{E}X^{\mathrm{T}}\mathrm{cov}(X)^{-1}\mathbb{E}X & -\mathbb{E}X^{\mathrm{T}} \cdot \mathrm{cov}(X)^{-1} \\ -\mathrm{cov}(X)^{-1}\mathbb{E}X & \mathrm{cov}(X)^{-1} \end{pmatrix}.$$

Тогда $\mathbb{E}YX'^{\mathrm{T}}\left(\mathbb{E}X'X'^{\mathrm{T}}\right)^{-1}$ это $\begin{pmatrix} \mathbb{E}Y & \mathbb{E}YX^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} 1 + \mathbb{E}X^{\mathrm{T}}\mathrm{cov}(X)^{-1}\mathbb{E}X & -\mathbb{E}X^{\mathrm{T}} \cdot \mathrm{cov}(X)^{-1} \\ -\mathrm{cov}(X)^{-1}\mathbb{E}X & \mathrm{cov}(X)^{-1} \end{pmatrix} =$

$$= \left(\mathbb{E}Y + \mathbb{E}Y\mathbb{E}X^{\mathrm{T}}\mathrm{cov}(X)^{-1}\mathbb{E}X - \mathbb{E}YX^{\mathrm{T}}\mathrm{cov}(X)^{-1}\mathbb{E}X, \quad -\mathbb{E}Y\mathbb{E}X^{\mathrm{T}} \cdot \mathrm{cov}(X)^{-1} + \mathbb{E}YX^{\mathrm{T}}\mathrm{cov}(X)^{-1}\right) =$$

$$(\mathbb{E}Y - \mathrm{cov}(Y, X)\mathrm{cov}(X)^{-1}\mathbb{E}X, \quad \mathrm{cov}(Y, X)\mathrm{cov}(X)^{-1}).$$

Условное матожидание

Минимизация MSE: Условное матожидание

Пусть у нас есть пара (X, Y) и мы хотим предсказывать Y с помощью X.

Предположим, что $\mathfrak{X}$ это множество всех X-измеримых функций $g(X)$ с конечными вторыми моментами: $\mathbb{E}g^2(X) < \infty$. Это наиболее широкий класс функций, который мы можем рассмотреть в задаче регрессии с функцией качества MSE.

В этом случае наша задача — найти проекцию Y на $\mathfrak{X}$. Если проекция существует, то она называется **условным матожиданием Y относительно X** и обозначается как $\mathbb{E}(Y|X)$.

Еще раз: $\mathbb{E}(Y|X)$ это X-измеримая функция $g_0(X)$, которая доставляет минимум $\mathbb{E}\big(Y - g(X)\big)^2$ на множестве всех X-измеримых функций с конечными вторыми моментами.

Минимизация MSE: Условное матожидание

Так как $\mathfrak{X}$ является линейным пространством (и содержит константы, кстати), то $\mathbb{E}(Y|X)$ это единственная функция такая, что

$$\mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)g(X) = 0 \text{ для всех } g.$$

Если $\mathbb{E}(Y|X) = g_0(X)$, то для $g_0(x)$ обычно пишут $\mathbb{E}(Y|X = x)$. Такая запись интерпретируется как среднее значение Y при условии, что мы пронаблюдали реализацию x величины X.

Свойства УМО

$$\mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)g(X) = 0 \text{ для всех } g.$$

Пусть $g(X) \equiv 1$. Тогда $\mathbb{E}\big(Y - \mathbb{E}(Y|X)\big) = 0$ и значит

$$\mathbb{E}\mathbb{E}(Y|X) = \mathbb{E}Y.$$

Говорят, что матожидание снимает условие.

Свойства УМО

$$\mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)g(X) = 0 \text{ для всех } g.$$

Пусть $Y = f(X)$. Тогда $\mathbb{E}(Y|X) = f(X)$, то есть Y полностью предсказуема по X.

Свойства УМО

Предположим, что совместное распределение (X, Y) имеет плотность $p(x, y)$ и пусть

(Это такая формула полной вероятности) $$p(y|x) = \frac{p(x,y)}{p_X(x)}$$

$$p_X(x) = \int_y p(x,y)dy$$

это условная плотность Y при условии $X = x$. Тогда

$$\mathbb{E}(Y|X = x) = \int yp(y|x)dy\,,$$

то есть условное матожидание это матожидание условного распределения.

Свойства УМО

$$\mathbb{E}(Y|X=x)=\int yp(y|x)dy$$

В самом деле,

$$\mathbb{E}\big(Y-g(X)\big)^2=\int_{X\times Y}\big(y-g(x)\big)^2p(x,y)dxdy=\int_X\left[\int_Y\big(y-g(x)\big)^2p(y|x)dy\right]p_X(x)dx.$$

Таким образом, минимизировать $\mathbb{E}\big(Y-g(X)\big)^2$ можно для каждого x в отдельности.

Свойства УМО

$$\mathbb{E}(Y|X=x)=\int yp(y|x)dy$$

В самом деле,

$$\mathbb{E}\big(Y-g(X)\big)^2=\int_{X\times Y}\big(y-g(x)\big)^2p(x,y)dxdy=\int_X\left[\int_Y\big(y-g(x)\big)^2p(y|x)dy\right]p_X(x)dx.$$

Таким образом, минимизировать $\mathbb{E}\big(Y-g(X)\big)^2$ можно для каждого x в отдельности.

Для данного x величина $g(x)$ это просто константа, а значит минимум $\int_Y\big(y-g(x)\big)^2p(y|x)dx$

достигается при $g(x)=\int_Y yp(y|x)dx$.

Свойства УМО

$$\mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)g(X) = 0 \text{ для всех } g.$$

Если X и Y независимы, то $\mathbb{E}(Y|X) = \mathbb{E}Y$.

В самом деле, в этом случае Y и $g(X)$ независимы и

$$\mathbb{E}(Y - \mathbb{E}Y)g(X) = \mathbb{E}(Y - \mathbb{E}Y)\mathbb{E}g(X) = 0.$$

Иначе говоря, $Y - \mathbb{E}Y$ ортогонально $\mathfrak{X}$, то есть $\mathbb{E}Y$ это проекция.

Свойства УМО

$$\mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)g(X) = 0 \text{ для всех } g.$$

Если $f(X)$ это X-измеримая функция, то

$$\mathbb{E}(f(X)Y|X) = f(X)\mathbb{E}(Y|X).$$

Иначе говоря, $f(X)$ при условии X ведет себя как константа.

В самом деле, для всякой $g(X)$

$$\mathbb{E}\big(f(X)Y - f(X)\mathbb{E}(Y|X)\big)g(X) = \mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)f(X)g(X) = 0,$$

так как $Y - \mathbb{E}(Y|X)$ ортогонально $f(X)g(X)$. Стало быть, $f(X)Y - f(X)\mathbb{E}(Y|X)$ ортогонально $g(X)$.

Причем тут условное нормальное распределение?

Причем тут условное нормальное распределение?

Если $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$, то $Y|(X = x) \sim \mathcal{N}(\mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X), \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{YX})$. Заметим, что

- $\mu_Y + \Sigma_{YX}\Sigma_X^{-1}(X - \mu_X)$ это УМО и при этом оптимальное линейное предсказание Y по X,
- $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{YX}$ это дополнение Шура Σ_X для $\Sigma_{(X,Y)}$.