Вероятностные модели и статистика случайных процессов

Конспекты лекций

Лектор: А.В. Артёмов

НИУ ВШЭ, 2018

Оглавление

Глава 1

Лекции

1.1 Введение в курс

Как понятно из названия, этот пункт посвящён простой цели: вспомнить основные понятия, необходимые для дальнейшего чтения. Начнём с самого основного: *вероятностного* пространства.

Определение 1. Вероятностное пространство, или *тройка Колмогорова* — это тройка $(\Omega, \mathcal{F}, \mathsf{P})$, где

- Ω — простанство элементарных исходов.
- $\mathcal{F}$ — σ-алгебра над Ω, пространство событий. Неформально говоря, $\mathcal{F}$ определяет объекты, относительно которых делаются утверждения.
- $\mathsf{P} : \mathcal{F} \mapsto [0, 1]$ — вероятностная мера.

На всякий случай напомню определения сигма-алгебры и вероятностной меры.

Определение 2. *σ-алгебра* над множеством A — это множество $\mathcal{F} \subseteq 2^A$, обладающее следующими свойствами:

1. $\varnothing \in \mathcal{F}$.
2. Если $X \in \mathcal{F}$, то и дополнение $A \setminus X \in \mathcal{F}$.
3. $\mathcal{F}$ замкнуто относительно счётного объединения, то есть объединение счётного подсемейства из $\mathcal{F}$ лежит в $\mathcal{F}$.[1]

Определение 3. Пусть Ω — пространство элементарных исходов, а $\mathcal{F}$ — σ-алгебра его подмножеств (пространство событий). *Вероятностной мерой* или же вероятностью называется функция $\mathsf{P} : \mathcal{F} \mapsto [0, 1]$, удовлетворяющая двум свойствам:

1. (счётная аддитивность) Пусть $\{A_n\}_{n=1}^{\infty}$ — последовательность попарно не пересекающихся событий. Тогда
$$\mathsf{P}\left(\bigsqcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathsf{P}(A_n)$$
2. (нормированность) $\mathsf{P}(\Omega) = 1$.

Теперь посмотрим на несколько примеров вероятностных пространств.

[1] То есть это алгебра, к которой конечное объединение заменено на счётное.

Пример 1. Допустим, что человек бросает монетку. В данном случае элементарными исходами являются "орёл" и "решка". Обозначим их за 0 и 1 соответственно. Тогда вероятностное пространство будет устроено следующим образом: пространство элементарных исходов $\Omega = \{0, 1\}$, пространство событий $\mathcal{F} = \{\varnothing, \{0\}, \{1\}, \Omega\}$, а вероятность P определяется следующим образом:

$$\mathsf{P}(\varnothing) = 0, \quad \mathsf{P}(\{1\}) = p, \qquad \mathsf{P}(\{0\}) = 1 - p, \qquad \mathsf{P}(\Omega) = 1, \quad p \in [0, 1].$$

Пример 2. Теперь подбросим монетку n раз. В таком случае $\Omega = \{0, 1\}^n$, то есть если $\omega \in \Omega$, то $\omega = (\omega_1, \ldots, \omega_n)$, где $\omega_i \in \{0, 1\}$. Пространство событий же введём просто как множество всех подмножеств: $\mathcal{F} = 2^\Omega$. Введение же вероятностной меры оставим читателю.

В дискретном случае алгебра событий, как известно, очень проста: обычно это просто множество всех подмножеств. Теперь перейдём от дискретного случая к более общему. Пусть $\Omega = \mathbb{R}$. Как тогда ввести алгебру событий? Ведь, как известно, в $\mathbb{R}$ есть неизмеримые подмножества. Для этого вводят *борелевскую σ-алгебру* $\mathcal{B}(\mathbb{R})$ и говорят, что $\mathcal{F} = \mathcal{B}(\mathbb{R})$. Напомню определение:

Определение 4. Борелевская σ-алгебра над $\mathbb{R}$ $\mathcal{B}(\mathbb{R})$ — это минимальная (по включению) сигма-алгебра над $\mathbb{R}$, содержащая все полуинтервалы вида $(a, b]$.[2]

Всё это вводилось ради того, с чем мы работаем на постоянной основе: ради *случайных величин*.

Определение 5. Пусть $(\Omega, \mathcal{F}, \mathsf{P})$ — вероятностное пространство. Случайной величиной будем называть отображение $\xi : \Omega \mapsto \mathbb{R}$ такое, что для любого $x \in \mathbb{R}$ событие $\{\xi \leqslant x\} = \{\omega \in \Omega : \xi(\omega) \leqslant x\}$ лежит в $\mathcal{F}$.[3]

Характеризовать поведение случайной величины помогает *функция распределения*:

Определение 6. Пусть ξ — случайная величина на вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$. Функцией распределения случайной величины ξ называют функцию $F_\xi : \mathbb{R} \mapsto [0, 1]$, определяемую следующим образом: $F_\xi(x) = \mathsf{P}(\xi \leqslant x)$.

Дальше по плану сходимости. Как известно, их четыре типа, и они не равноценны. Чтобы не повторяться, сразу же введём обозначения. Пусть $(\Omega, \mathcal{F}, \mathsf{P})$ — вероятностное пространство, а $\{\xi_n\}_{n=1}^\infty$ и ξ — случайные величины на нём.

Определение 7. ξ_n сходится к ξ почти наверное, если

$$\mathsf{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)\right\}\right) = 1.$$

Обозначение: $\xi_n \xrightarrow{\text{п.н.}} \xi$ или же просто $\xi_n \to \xi$ (если не указан вид сходимости и он не ясен из контекста, то сходимость идёт почти наверное).

Определение 8. ξ_n сходится к ξ по вероятности, если для любого $\varepsilon > 0$

$$\lim_{n\to\infty} \mathsf{P}(\{\omega \in \Omega : |\xi_n(\omega) - \xi(\omega)| > \varepsilon\}) = 0.$$

Обозначение: $\xi_n \xrightarrow{\mathsf{P}} \xi$.

[2] Хоть она и минимальна, но всё равно она огромна — найти неборелевское множество не так уж и просто.

[3] Это свойство принято называть измеримостью. Такое требование нужно для того, чтобы были осмысленны вопросы типа "Чему равна вероятность того, что значение случайной величины будет лежать в таком-то отрезке?".

Определение 9. ξ_n сходится к ξ в среднем порядка p, если

$$\lim_{n\to\infty} \mathsf{E}[|\xi_n - \xi|^p] = 0.$$

Обозначение: $\xi_n \xrightarrow{L^p} \xi$.

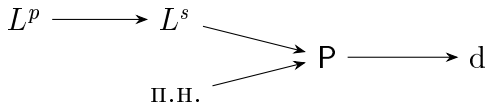
На практике обычно берётся $p = 2$. В таком случае говорят, что имеет место *сходимость в среднеквадратичном смысле*. У такой сходимости есть аж три разных обозначения: $\xi_n \xrightarrow{L^2} \xi$, $\xi_n \xrightarrow{\text{с.к.}} \xi$ или же вообще $\xi = \text{l.i.m.}\, \xi_n$.

Определение 10. ξ_n сходятся к ξ по распределению,[4] если для любой ограниченной непрерывной функции $f : \mathbb{R} \mapsto \mathbb{R}$

$$\lim_{n\to\infty} \mathsf{E}[f(\xi_n)] = \mathsf{E}[f(\xi)].$$

Обозначение: $\xi_n \xrightarrow{\text{d}} \xi$.

Цепочка взаимосвязей сходимостей устроена следующим образом. На данном рисунке $p > s \geqslant 1$ и стрелка из A в B означает “из A следует B”:



Поехали дальше. Для случайных величин вводились такие вещи, как матожидание и дисперсия. Напомню:

Определение 11. *Математическое ожидание* случайной величины ξ с функцией распределения F_ξ и плотностью p_ξ — это интеграл

$$\mathsf{E}[\xi] \equiv \int\limits_{-\infty}^{+\infty} x\, dF_\xi(x) = \int\limits_{-\infty}^{+\infty} x p_\xi(x)\, dx$$

Рядом с матожиданием вводятся *дисперсия*, *ковариация* и *корреляция*.

Определение 12. Дисперсией случайной величины ξ называется $\mathsf{D}[\xi] = \mathsf{E}[(\xi - \mathsf{E}[\xi])^2] = \mathsf{E}[\xi^2] - (\mathsf{E}[\xi])^2$. Корень из дисперсии σ называется *среднеквадратиным отклонением*.

Определение 13. Ковариацией случайных величин ξ и η называется

$$\mathsf{cov}(\xi, \eta) = \mathsf{E}[(\xi - \mathsf{E}[\xi])(\eta - \mathsf{E}[\eta])] = \mathsf{E}[\xi\eta] - \mathsf{E}[\xi]\,\mathsf{E}[\eta].$$

Определение 14. Корреляцией случайных величин ξ и η называется

$$\rho(\xi, \eta) = \frac{\mathsf{cov}(\xi, \eta)}{\sqrt{\mathsf{D}[\xi]\,\mathsf{D}[\eta]}}.$$

В многомерном случае ситуация немного изменяется. Матожидание случайного вектора определяется, как вектор из матожиданий компонент. Ковариация (да и дисперсия тоже) случайных векторов $\boldsymbol{\xi} \in \mathbb{R}^m$ и $\boldsymbol{\eta} \in \mathbb{R}^n$ равна

$$\mathsf{cov}(\boldsymbol{\xi}, \boldsymbol{\eta}) = \mathsf{E}[(\boldsymbol{\xi} - \mathsf{E}[\boldsymbol{\xi}])(\boldsymbol{\eta} - \mathsf{E}[\boldsymbol{\eta}])^\intercal] = (\mathsf{cov}(\xi_i, \eta_j))_{m\times n}.$$

При анализе некоторых вещей могут понадобиться условные матожидания. Введём их.

[4] Она так называется по той причине, что в одномерном случае её условие расносильно сходиости функций распределения почти наверное.

Определение 15. Пусть ξ — случайная величина, а $\boldsymbol{\eta}$ — случайный вектор из $\mathbb{R}^n$. Тогда условным математическим ожиданием ξ относительно $\boldsymbol{\eta}$ называется случайная величина $\mathsf{E}[\xi \mid \boldsymbol{\eta}]$, удовлетворяющая двум условиям:

1. $\mathsf{E}[\xi \mid \boldsymbol{\eta}] = \varphi(\boldsymbol{\eta})$, где $\varphi : \mathbb{R}^n \mapsto \mathbb{R}$ — некоторая борелевская функция (свойство измеримости).

2. Для любого $B \in \mathcal{B}(\mathbb{R}^n)$ $\mathsf{E}[\xi\,\mathrm{I}\{\boldsymbol{\eta} \in B\}] = \mathsf{E}[\mathsf{E}[\xi \mid \boldsymbol{\eta}]\,\mathrm{I}\{\boldsymbol{\eta} \in B\}]$ (интегральное свойство).

Но считать условное матожидание по определению весьма грустно. Поэтому введём две теоремы, которые упрощают жизнь. Они опираются на понятие *условного распределения* и условной плотности.

Определение 16. *Условным распределением* случайной величины ξ при условии, что $\boldsymbol{\eta} = \mathbf{y}$ назовём функцию $\mathsf{P}(\xi \in B \mid \boldsymbol{\eta} = \mathbf{y}) \equiv \mathsf{E}[\mathrm{I}\{\xi \in B\} \mid \boldsymbol{\eta} = \mathbf{y}]$, рассматриваемую, как функцию от $B \in \mathcal{B}(\mathbb{R}^n)$ при фиксированном $\mathbf{y} \in \mathbb{R}^k$.

Определение 17. Если условное распределение имеет плотность $p_{\xi|\boldsymbol{\eta}}(x \mid \mathbf{y})$, то назовём его *условной плотностью ξ относительно $\boldsymbol{\eta}$*. То есть для любого $B \in \mathcal{B}(\mathbb{R})$

$$\mathsf{P}(\xi \in B \mid \boldsymbol{\eta} = \mathbf{y}) = \int\limits_B p_{\xi|\boldsymbol{\eta}}(x \mid \mathbf{y})\,dx.$$

Теорема 1 (о вычислении условного математического ожидания)**.** *Пусть ξ — случайная величина, а $f(x)$ — некоторая борелевская функция. Если $\mathsf{E}[|f(\xi)|] < +\infty$ и существует плотность $p_{\xi|\boldsymbol{\eta}}(x \mid \mathbf{y})$, то*

$$\mathsf{E}[f(\xi) \mid \boldsymbol{\eta} = \mathbf{y}] = \int\limits_{\mathbb{R}} f(x) p_{\xi|\boldsymbol{\eta}}(x \mid \mathbf{y})\,dx.$$

Теорема 2. *Пусть ξ и η таковы, что есть совместная плотность $p_{\xi,\eta}(x,y)$. Тогда существует условная плотность $p_{\xi|\eta}(x \mid y)$ и она равна*

$$p_{\xi|\eta}(x \mid y) = \frac{p_{\xi,\eta}(x,y)}{p_\eta(y)}\,\mathrm{I}\{p_\eta(y) > 0\}$$

Обобщение на многомерный случай ровно такое же, как и для обычных матожиданий.

1.2 Основы теории случайных процессов

Теперь можно приступать к делу. Для начала введём понятие случайной функции.

Определение 18. Пусть $(\Omega, \mathcal{F}, \mathsf{P})$ — вероятностное пространство, а T — произвольное неслучайное множество (множество индексов). Тогда любое множество случайных величин $X = \{X_t \mid t \in T\}$ называется *случайной функцией.*

Случайные функции можно классифицировать по строению множества индексов.

- Если $T = \{1\}$ (или любое другое одноэлементное множество), то $X = X_1 : \Omega \mapsto \mathbb{R}$ — случайная величина.

- Если $T = \{1, 2, \ldots, N\}$, то $X = (X_1, \ldots, X_N)$ — случайный вектор.

- Если T дискретно (то есть не более, чем счётно), то X называют *случайной последовательностью*. Например, возьмём $T = \mathbb{N}$ и $\Omega = \{0, 1\}$. Тогда случайная последовательность $X = (X_t)_{t\in T}$ соотвествует броскам монетки.

- Если $T \subseteq \mathbb{R}$, то параметр $t \in T$ можно характеризовать, как время. В этом случае X принято называть *случайным процессом*.

- Аналогично, если $T \subseteq \mathbb{R}^d$, где $d \geqslant 1$, то параметр можно характеризовать, как точку в d-мерном пространстве. Тогда X называют *случайным полем*. Например, если $T = \mathbb{R}^3$, то $X_t(\omega)$ может соответствовать давлению в точке $t = (x, y, z)$ в момент времени ω.

Впрочем, данная классификация не является строгой. Например, понятие "случайный процесс" обычно считается безусловным синонимом термина "случайная функция".

Случайные процессы ещё разбивают по мощности множества индексов. Если оно не более, чем счётно, то говорят, что случайный процесс *дискретен во времени*. Если же T континуально, то у случайного процесса *непрерывное время*.

Рассмотрим несколько примеров.

Пример 3. Пусть $T = \mathbb{N}$, а $\{X_n\}_{n=1}^{\infty}$ — последовательность независимых и одинаково распределённых случайных величин с нулевым матожиданием.[5] Тогда $X = (X_t)_{t\in T}$ — случайный процесс с дискретным временем, называемый *белым шумом*.

Пример 4. Пусть $Y = (Y_t)_{t\in T}$ — последовательность iid случайных величин. Тогда построим новую последовательность случайных величин следующим образом: $X_t = Y_1 + \ldots + Y_t$. Тогда $X = (X_t)_{t\in T}$ — случайный процесс с дискретным временем, называемый *случайным блужданием*.

Пример 5. Пусть $\xi_1, \ldots, \xi_d$ — случайные величины. Тогда "случайный полином"

$$X = (X_t)_{t\in\mathbb{R}}, \quad X_t = \sum_{n=1}^{d} \xi_n t^n$$

образует случайный процесс.

Пример 6. Пусть некоторое устройство (например, жёсткий диск) начинает работу в момент времени 0 и ломается в момент времени T_1. В этот же момент его меняют на новое (временем замены пренебрегаем). Оно, в свою очередь, ломается спустя время T_2, после чего его снова заменяют и этот процесс продолжается. Введём случайные величины $S_n = T_1 + \cdots + T_n$. Случайный процесс $S = (S_t)_{t\in\mathbb{N}}$ называется процессом моментов восстановления. Сечения будут соответствовать моментам "восстановления". Далее, введём следующий случайный процесс:

$$N = (N_t)_{t\in\mathbb{R}}, \quad N_t = \sum_{n=1}^{\infty} \mathrm{I}\{S_n \leqslant t\} = \#\{n \in \mathbb{N} \mid S_n \leqslant t\}.$$

Этот процесс принято называть *процессом восстановления*.

В дальнейшем нам понадобятся понятия *сечения* и *траектории* случайной функции. Введём их.

[5] Для удобства в дальнейшем будем иногда сокращать "независимые и одинаково распределённые" до iid (independent and identically distributed).

Определение 19. Пусть $X = (X_t)_{t\in T}$ — случайная функция. Тогда X_t при фиксированном t называется *сечением* случайной функции.

Определение 20. Пусть $X = (X_t)_{t\in T}$ — случайная функция и $\omega \in \Omega$ — фиксированный исход. *Траекторией* случайной функции называется функция $\varphi_\omega : T \mapsto \mathbb{R}$ такая, что $\varphi_\omega(t) = X_t(\omega)$.

Теперь вопрос: есть случайный процесс. Как описать вероятность того, что реализация удовлетворяет какому-то условию? Да и вообще, какие вопросы можно задавать относительно реализаций?

Посмотрим на пару частных случаев случайного процесса $X = (X_t)_{t\in T}$

1. Пусть $T = \{1\}$. Тогда случайный процесс превращается в случайную величину ξ. Но её значения описываются функцией распределения $F_\xi(x) = \mathsf{P}(\xi \leqslant x)$.

2. Аналогично, пусть $T = \{1, 2, \ldots, n\}$. Тогда $X = (X_1, \ldots, X_n)$ — это случайный вектор, а его значения описываются совместной функцией распределения

$$F_X(x_1, \ldots, x_n) = \mathsf{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n)$$

В конечном случае всё хорошо и можно обойтись совместной функцией распределения. А как перейти в бесконечный случай? Сделать прямое обобщение вряд ли получится: учитывать бесконечное число условий сложно и не факт, что возможно. Поэтому будем смотреть на конечные поднаборы. Тем самым мы пришли к следующему определению.

Определение 21. Пусть $(\Omega, \mathcal{F}, \mathsf{P})$ — вероятностное пространство, T — множество индексов, а $X = (X_t)_{t\in T}$ — случайный процесс. Тогда *семейством конечномерных распределений* случайного процесса X называется набор F его конечномерных функций распределения

$$F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) = \mathsf{P}(X_{t_1} \leqslant x_1, \ldots, X_{t_n} \leqslant x_n),$$

то есть

$$F = \{F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) \mid n \in \mathbb{N}, t_1, \ldots, t_n \in T\}.$$

Пример 7. Пусть $X = (X_t)_{t\in\mathbb{N}}$ — случайный процесс, состоящий из iid случайных величин из распределения $\mathcal{N}(\mu, \sigma^2)$. Тогда

$$F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) = \prod_{k=1}^{n} F_{X_{t_k}}(x_k) = \prod_{k=1}^{n} \int_{-\infty}^{x_k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} dy.$$

Вообще говоря, конечномерное распределение $F_{t_1,\ldots,t_n}(x_1, \ldots, x_n)$ — это распределение случайного вектора $(X_{t_1}, \ldots, X_{t_n})$.

Как мы видим, по любому случайному процессу можно построить семейство конечномерных функций распределения. Теперь вопрос: а можно ли по семейству построить случайный процесс? Оказывается, что можно.

Теорема 3 (Колмогорова о существовании случайного процесса)**.** *Пусть F — заданное семейство конечномерных функций распределения над множеством индексов T, которое удовлетворяет свойству согласованности: для $k < n$*

$$F_{t_1,\ldots,t_k}(x_1, \ldots, x_k) = F_{t_1,\ldots,t_n}(x_1, \ldots, x_k, +\infty, \ldots, +\infty).$$

Тогда существует вероятностное пространство $(\Omega, \mathcal{F}, \mathsf{P})$ и случайный процесс $X = (X_t)_{t\in T}$ такой, что $\mathsf{P}(X_{t_1} \leqslant x_1, \ldots, X_{t_n} \leqslant x_n) = F_{t_1,\ldots,t_n}(x_1, \ldots, x_n)$.

> Вообще, доказательство этой теоремы не особо сложное, но оно требует введения не самых простых вещей наподобие борелевских сигма-аглебр над пространством функций. Поэтому я лишь напишу идею.

Теперь вспомним, что независимость бесконечного набора — это независимость в совокоуности любого конечного поднабора. Тогда из теоремы Колмогорова сразу же получаем следующее

Следствие. *Пусть для каждого $t \in T \subseteq \mathbb{R}$ определена одномерная функция распределения $F_t(x)$. Тогда существует вероятностное пространство $(\Omega, \mathcal{F}, \mathsf{P})$ и случайный процесс $X = (X_t)_{t\in T}$ с независимыми сечениями такой, что $\mathsf{P}(X_t \leqslant x) = F_t(x)$.*

Теперь рассмотрим пример применения этой теоремы. Пусть $T = [0, 1]$, а

$$F_{t_1,\dots,t_n}(x_1, \dots, x_n) = \prod_{k=1}^{n} \int_{-\infty}^{x_k} g(y)\,dy,$$

где g — симметричная относительно нуля плотность. Возьмём случайный процесс $X = (X_t)_{t\in T}$ из теоремы Колмогорова. Несложно понять, что в таком случае

$$\mathsf{P}(X_t > \varepsilon, X_s < -\varepsilon) = \left(\int_{\varepsilon}^{\infty} g(y)\,dy \right)^2$$

Теперь возьмём последовательность событий $A_n = \{X_t > \varepsilon, X_{t+1/n} < -\varepsilon\}$. Если процесс является непрерывным (поточечно), то $A_n \to \varnothing$ при $n \to \infty$. Но тогда нарушается непрерывность вероятностной меры в нуле, то есть $\mathsf{P}(A_n) \not\to 0$. Следовательно, непрерывность, независимость и одинаковая распределённость не совместимы.

Для описания случайных величин вводились такие вещи, как матожидание, дисперсия и так далее. Так вот: их можно ввести и для случайных процессов, хотя это уже будут не константы, а функции (неслучайные).

Определение 22. Математическое ожидание случайного процесса $X = (X_t)_{t\in T}$ — это функция $m : T \mapsto \mathbb{R}$, устроенная следующим образом: $m(t) = \mathsf{E}[X_t]$.

Определение 23. Дисперсия случайного процесса $X = (X_t)_{t\in T}$ — это функция $D : T \mapsto \mathbb{R}$, устроенная следующим образом: $D(t) = \mathsf{D}[X_t] = \mathsf{E}[(X_t - \mathsf{E}[X_t])^2]$.

Определение 24. Среднеквадратичное отклонение случайного процесса $X = (X_t)_{t\in T}$ — это функция $\sigma : T \mapsto \mathbb{R}$, устроенная следующим образом: $\sigma(t) = \sqrt{\mathsf{D}[X_t]}$.

Определение 25. Ковариационная функция случайного процесса $X = (X_t)_{t\in T}$ — это функция $R : T^2 \mapsto \mathbb{R}$, устроенная следующим образом:

$$R(t_1, t_2) = \mathsf{cov}(X_{t_1}, X_{t_2}) = \mathsf{E}[(X_{t_1} - \mathsf{E}[X_{t_1}])(X_{t_2} - \mathsf{E}[X_{t_2}])]$$

Наряду с ковариационной функцией вводят корреляционную функцию

$$r(t_1, t_2) = \frac{R(t_1, t_2)}{\sigma(t_1)\sigma(t_2)}$$

Про неё нужно сказать, что по неравенству Коши-Буняковского-Шварца $|r(t_1, t_2)| \leqslant 1$.

Осталось ввести ещё одну функцию $K : T^2 \mapsto \mathbb{R}$, определяемую следующим образом: $K(t_1, t_2) = \mathsf{E}[X_{t_1} X_{t_2}]$.

Теперь выпишем несколько очевидных свойств:

- $D(t) = R(t,t) \geqslant 0$, так как дисперсия неотрицательна.
- $K(t_1,t_2) = \mathsf{E}[(X_{t_1} - m(t_1) + m(t_1))(X_{t_2} - m(t_2) + m(t_2))] = R(t_1,t_2) + m(t_1)m(t_2)$ (проверьте!).
- $R(t_1,t_2) = R(t_2,t_1)$.
- Для любого $n \in \mathbb{N}$ и наборов $(z_1,\ldots,z_n)$, $(t_1,\ldots,t_n)$ матрица $(R(t_i,t_j))_{n\times n}$ неотрицательно определена:
$$\sum_{i,j=1}^{n} R(t_i,t_i) z_i z_j \geqslant 0$$
Это следует из того, что $(X_{t_1},\ldots,X_{t_n})$ — случайный вектор, а $(R(t_i,t_j))_{n\times n}$ есть его матрица ковариаций.

Допустим, что у нас есть какой-то случайный процесс. Что произойдёт с его параметрами при линейном преобразовании?

Пусть $X = (X_t)_{t\in T}$ — случайный процесс, а $a, b : T \mapsto \mathbb{R}$ — ограниченные неслучайные функции. Построим новый случайный процесс $Y = (Y_t)_{t\in T}$ следующим образом: $Y_t = a(t)X_t + b(t)$. Посмотрим на матожидание, дисперсию и ковариационную функцию данного процесса. Для того, чтобы различать их для разных процессов, будем указывать название процесса в качестве индекса: например, $R_Y(t_1,t_2)$.

$$m_Y(t) = \mathsf{E}[Y_t] = \mathsf{E}[a(t)X_t + b(t)] = a(t)m_X(t) + b(t)$$
$$R_Y(t_1,t_2) = \mathsf{cov}(Y_{t_1},Y_{t_2}) = \mathsf{cov}(a(t_1)X_{t_1} + b(t_1), a(t_2)X_{t_2} + b(t_2)) = a(t_1)a(t_2)R_X(t_1,t_2)$$
$$D_Y(t) = R_Y(t,t) = a^2(t)R_X(t,t) = a^2(t)D_X(t)$$

Теперь сделаем небольшое обобщение и возьмём линейную комбинацию n процессов. Пусть $X_i = (X_t^i)_{t\in T}$, $i = 1,\ldots,n$, а $a_1,\ldots,a_n$ — набор ограниченных функций из T в $\mathbb{R}$. Далее, строим новый случайный процесс $Y = (Y_t)_{t\in T}$ по следующему правилу:

$$Y_t = \sum_{i=1}^{n} a_i(t)X_t^i + b(t)$$

Для упрощения жизни введём *взаимную корреляционную функцию* $R_{X_i,X_j}(t_1,t_2) = \mathsf{cov}(X_{t_1}^i, X_{t_2}^j)$. Тогда несложно показать, что

$$m_Y(t) = \sum_{i=1}^{n} a_i(t)m_{X_i}(t) + b(t)$$
$$R_Y(t_1,t_2) = \sum_{i=1}^{n} a_i(t_1)a_i(t_2)R_{X_i}(t_1,t_2) + \sum_{\substack{i,j=1\\ i\neq j}}^{n} a_i(t_1)a_j(t_2)R_{X_i,X_j}(t_1,t_2)$$
$$D_Y(t) = \sum_{i=1}^{n} a_i^2(t)D_{X_i}(t) + \sum_{\substack{i,j=1\\ i\neq j}}^{n} a_i(t)a_j(t)R_{X_i,X_j}(t,t)$$

1.3 Непрерывность случайных процессов

Вообще, случайные функции — тоже вполне себе функции, поэтому вполне осмысленно посмотреть на её непрерывность. Только в данном случае можно вводить разные виды непрерывности, так как последовательности сходятся не только поточечно.

Определение 26. Случайный процесс $X = (X_t)_{t\in T}$ непрерывыен в среднеквадратичном смысле в точке $t \in T$, если

$$X_{t+\varepsilon} \xrightarrow{\text{с.к.}} X_t \text{ при } \varepsilon \to 0.$$

Если это выполнено для любого $t \in T$, то процесс называют непрерывным в среднеквадратичном смысле.

Вот есть у нас процесс. Можем ли мы сказать, что он будет непрерывным в среднеквадратичном смысле, не вспоминая определение каждый раз? Можем.

Теорема 4. *Случайный процесс $X = (X_t)_{t\in T}$ непрерывен в среднеквадратичном случае тогда и только тогда, когда непрерывны $R(t_1, t_2)$ и $m(t)$.*

Доказательство. Для начала проверим, что этого условия достаточно. Действительно, из непрерывности $R(t_1, t_2)$ и $m(t)$ следует непрерывность $K(t, t)$ и

$$\mathsf{E}\big[(X_{t+\varepsilon} - X_t)^2\big] = \mathsf{E}\big[X_{t+\varepsilon}^2\big] + \mathsf{E}\big[X_t^2\big] - 2\,\mathsf{E}[X_t X_{t+\varepsilon}] = K(t+\varepsilon, t+\varepsilon) + K(t,t) - 2K(t, t+\varepsilon).$$

Устремляя ε к нулю, получаем, что $\mathsf{E}[(X_{t+\varepsilon} - X_t)^2] \to 0$, что и даёт непрерывность в среднеквадратичном смысле. Теперь покажем, что это условие необходимо. Для этого воспользуемся следующей леммой:

Лемма. *Пусть $\{X_n\}_{n=1}^{\infty}$, $\{Y_n\}_{n=1}^{\infty}$, X и Y — случайные величины такие, что $X_n \xrightarrow{\text{с.к.}} X$, $Y_n \xrightarrow{\text{с.к.}} Y$, $\mathsf{E}[X^2] < \infty$ и $\mathsf{E}[Y^2] < \infty$. Тогда*

$$\lim_{m,n\to\infty} \mathsf{E}[X_n Y_m] = \mathsf{E}[XY].$$

Доказательство. Для начала покажем, что $(\mathsf{E}[XY])^2 \leqslant \mathsf{E}\big[X^2\big]\,\mathsf{E}\big[Y^2\big]$. Для этого рассмотрим $f(t) = \mathsf{E}\big[(tX+Y)^2\big]$. Понятно, что $f(t)$ есть квадратный трёхчлен от t и он неотрицателен. Тогда его дискриминант отрицателен и мы получаем желаемое. По сути, это просто неравенство Коши-Буняковского-Шварца.

Заметим, что по нему

$$|\,\mathsf{E}[(X_n - X)(Y_m - Y)]| \leqslant \sqrt{\mathsf{E}[(X_n - X)^2]\,\mathsf{E}[(Y_n - Y)^2]} \to 0.$$

Теперь раскроем скобки:

$$\mathsf{E}[(X_n - X)(Y_m - Y)] = \mathsf{E}[X_n Y_m] - \mathsf{E}[X_n Y] - \mathsf{E}[X Y_m] + \mathsf{E}[XY]$$

Покажем, что $\mathsf{E}[X_n Y] \to \mathsf{E}[XY]$. Действительно,

$$\mathsf{E}[(X_n - X)Y] \leqslant \sqrt{\mathsf{E}[Y^2]\,\mathsf{E}[(X_n - X)^2]} \to 0.$$

Аналогично, $\mathsf{E}[XY_m] \to \mathsf{E}[XY]$. Тем самым мы получаем желаемое. □

Из неё сразу же следует, что при $\varepsilon_1, \varepsilon_2 \to 0$

$$\begin{cases} X_{t_1+\varepsilon_1} \xrightarrow{\text{с.к.}} X_t \\ X_{t_2+\varepsilon_2} \xrightarrow{\text{с.к.}} X_t \end{cases} \implies \mathsf{E}[X_{t_1+\varepsilon_1} X_{t_2+\varepsilon_2}] \to \mathsf{E}[X_{t_1} X_{t_2}] \implies K(t_1+\varepsilon_1, t_2+\varepsilon_2) \to K(t_1, t_2)$$

Это означает непрерывность $K(t_1, t_2)$. Теперь вспомним, что

$$\mathsf{D}[|X_{t_1} - X_{t_2}|] = \mathsf{E}\big[(X_{t_1} - X_{t_2})^2\big] - (\mathsf{E}[|X_{t_1} - X_{t_2}|])^2 \geqslant 0$$

Отсюда сразу же следует, что при $t_1 \to t_2$ $\mathsf{E}[|X_{t_1} - X_{t_2}|] \to 0$. Следовательно, $\mathsf{E}[X_{t_1}] \to \mathsf{E}[X_{t_2}]$ и $m(t)$ непрерывна. Отсюда получаем, что и $R(t_1, t_2)$ тоже непрерывна. □

Но не среднеквадратичным случаем единым! У него весьма жёсткие требования, так что введём ещё пару видов.

Определение 27. Будем говорить, что $X = (X_t)_{t\in T}$ — случайный процесс с непрерывными траекториями, если для любого $\omega \in \Omega$ траектория φ_ω непрерывна, как функция от t.

Пример 8. Пусть $T = [0,1]$, $f : [0,1] \mapsto \mathbb{R}$ — какая-то непрерывная функция, а ξ — случайная величина. Тогда случайный процесс $X = (X_t)_{t\in T}$, построенный по правилу $X_t = f(t)\xi$, будет иметь непрерывные траектории.

Определение 28. Будем говорить, что процесс $X = (X_t)_{t\in T}$ стохастически непрерывен, если выполняется сходимость по вероятности: для любого $t \in X_s \xrightarrow{\mathsf{P}} X_t$ при $s \to t$.

Пример 9. Пусть для любого натурального n T_n — это iid случайные величины с распределением $\mathrm{Exp}(\lambda)$, то есть их плотность равна $p(z) = \lambda e^{-\lambda z}\,\mathrm{I}\{z \geqslant 0\}$. Дальше, $S_n = T_1 + \ldots + T_n$, а для любого $t \geqslant 0$

$$N_t = \sum_{n=1}^{\infty} \mathrm{I}\{S_n \leqslant t\} = \#\{n \in \mathbb{N} : S_n < t\}.$$

По сути, это процесс восстановления для экспоненциального распределения. Далее мы докажем, что для него выполнено следующее свойство: для всех $0 \leqslant s < t < +\infty$ $N_t - N_s \sim \mathrm{Pois}(\lambda(t-s))$. Но из него сразу же получается стохастическая непрерывность:

$$\mathsf{P}(|N_t - N_s| > \varepsilon) \leqslant \mathsf{P}(|N_t - N_s| > 0) = \sum_{k=1}^{\infty} \frac{(\lambda(t-s))^k}{k!} e^{-\lambda(t-s)} = 1 - e^{-\lambda(t-s)} \xrightarrow[s\to t]{} 0.$$

Вообще говоря, между непрерывностями случайных процессов и видами сходимостей есть прямая связь. Например, если процесс непрерывен в среднеквадратичном смысле, то он будет стохастически непрерывен.

Ранее мы показали, что непрерывность исключает независимость. Можем ли мы сказать, что стохастическая непрерывность тоже исключает независимость? Можем. Пусть $X = (X_t)_{t\in T}$ — случайный процесс такой, что для какой-то окрестности t_0 X_t независимо с X_{t_0} и X_t имеет плотность $g(x)$. Посмотрим на вероятность отклониться на ε:

$$\mathsf{P}(|X_t - X_{t_0}| > \varepsilon) = \iint\limits_{|x-y|>\varepsilon} g(x)g(y)\,dx\,dy.$$

Но этот интеграл не стремится к нулю при $t_0 \to t$. Тогда стохастической непрерывности нет. Впрочем, как и непрерывных траекторий.

1.4 Лирическое отступление: гауссовские векторы

Сделаю небольшое лирическое отступление и вспомним гауссовские векторы. Для этого вспомним, что такое *характеристическая функция* случайной величины и вектора.

Определение 29. Пусть ξ — случайная величина с плотностью p_ξ. Тогда её характеристической функцией называется функция $\varphi_\xi : \mathbb{R} \mapsto \mathbb{C}$, определяемая следующим образом:

$$\varphi_\xi(t) = \mathsf{E}\big[e^{it\xi}\big] = \int\limits_{-\infty}^{+\infty} e^{itx} p_\xi(x)\,dx.$$

Определение 30. Пусть $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ — случайный вектор с совместной плотностью $p_{\boldsymbol{\xi}}$. Тогда её характеристической функцией называется функция $\varphi_{\boldsymbol{\xi}} : \mathbb{R}^n \mapsto \mathbb{C}$, определяемая следующим образом: $\varphi_{\boldsymbol{\xi}}(\mathbf{t}) = \mathsf{E}\left[e^{i\langle \boldsymbol{\xi}, \mathbf{t}\rangle}\right]$, где $\langle \cdot, \cdot \rangle$ — скалярное произведение в $\mathbb{R}^n$.

По сути, характеристическая функция — это преобразование Фурье функции распределения.

Далее, из курса теории вероятности известно, что если $\xi \sim \mathcal{N}(\mu, \sigma^2)$, то

$$\varphi_\xi(t) = \exp\left\{ i\mu t - \frac{1}{2}\sigma^2 t^2 \right\}.$$

Поэтому гауссовский вектор вводят следующим образом:

Определение 31. Случайный вектор $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ подчиняется *многомерному нормальному распределению*, если его характеристическая функция равна

$$\varphi_{\boldsymbol{\xi}}(\mathbf{t}) = \exp\left\{ i\langle \boldsymbol{\mu}, \mathbf{t}\rangle - \frac{1}{2}\langle \boldsymbol{\Sigma}\mathbf{t}, \mathbf{t}\rangle \right\},$$

где $\boldsymbol{\mu} \in \mathbb{R}^n$ — некоторый фиксированный вектор, а $\boldsymbol{\Sigma}$ — некоторая симметрическая и неотрицательно определённая матрица. В таком случае пишут, что $\xi \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Это определение не очень удобно. Докажем одну теорему, которая даст несколько более удобное определение.

Теорема 5. *Пусть $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ — случайный вектор. Он будет гауссовским тогда и только тогда, когда для любого неслучайного вектора $\boldsymbol{\lambda} \in \mathbb{R}^n$ $\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle$ имеет нормальное распределение.*

Доказательство. Пусть $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Тогда посмотрим на характеристическую функцию $\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle$. Заметим, что она равна

$$\varphi_{\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle}(t) = \mathsf{E}\left[e^{it\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle}\right] = \mathsf{E}\left[e^{i\langle t\boldsymbol{\lambda}, \boldsymbol{\xi}\rangle}\right] = \varphi_{\boldsymbol{\xi}}(t\boldsymbol{\lambda}) = \exp\left\{ it\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle - \frac{t^2}{2}\langle \boldsymbol{\Sigma}\boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle \right\}.$$

Это означает, что $\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle \sim \mathcal{N}(\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle, \langle \boldsymbol{\Sigma}\boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle)$.

Теперь предположим, что $\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle$ имеет нормальное распределение для любого $\boldsymbol{\lambda}$. Тогда посмотрим на характеристическую функцию $\boldsymbol{\xi}$:

$$\varphi_{\boldsymbol{\xi}}(\boldsymbol{\lambda}) = \mathsf{E}\left[e^{i\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle}\right] = \varphi_{\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle}(1) = \exp\left\{ i\,\mathsf{E}[\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle] - \frac{1}{2}\,\mathsf{D}[\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle] \right\}.$$

Теперь заметим, что

$$\mathsf{E}[\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle] = \mathsf{E}\left[\sum_{k=1}^n \lambda_k \xi_k\right] = \sum_{k=1}^n \lambda_k\, \mathsf{E}[\xi_k] = \langle \boldsymbol{\lambda}, \mathsf{E}[\boldsymbol{\xi}]\rangle,$$

$$\mathsf{D}[\langle \boldsymbol{\lambda}, \boldsymbol{\xi}\rangle] = \mathsf{cov}\left(\sum_{k=1}^n \lambda_k \xi_k, \sum_{k=1}^n \lambda_k \xi_k\right) = \sum_{i,j=1}^n \mathsf{cov}(\xi_i, \xi_j)\lambda_i \lambda_j = \langle \mathsf{D}[\boldsymbol{\xi}]\boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle.$$

Отсюда получаем, что $\boldsymbol{\xi} \sim \mathcal{N}(\mathsf{E}[\boldsymbol{\xi}], \mathsf{D}[\boldsymbol{\xi}])$. □

Теперь выпишем следствия из этой теоремы.

Следствие (Смысл параметров)**.** *Если случайный вектор $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, то $\boldsymbol{\mu} = \mathsf{E}[\xi]$, $\boldsymbol{\Sigma} = \mathsf{D}[\xi]$.*

Следствие (Линейные преобразования)**.** *Любое линейное преобразование гауссовского вектора тоже является гауссовским вектором.*

Теперь вспомним про плотности. У гауссовских векторов она есть не всегда.

Теорема 6 (о плотности гауссовских векторов)**.** *Пусть $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ — n-мерный гауссовский вектор. Тогда, если $\boldsymbol{\Sigma}$ положительно определена, то существует плотность $p_{\boldsymbol{\xi}}(\mathbf{t})$ и она равна*

$$p_{\boldsymbol{\xi}}(\mathbf{t}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left\{-\frac{1}{2}\left\langle \boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu}), (\mathbf{t}-\boldsymbol{\mu})\right\rangle\right\}.$$

Задача 1. *Пусть $\mathbf{X} = (\xi, \eta)$ — гауссовский вектор, для которого:*[6]

$$\mathsf{E}[\mathbf{X}] = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad \mathsf{D}[\mathbf{X}] = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}, \quad |\rho| < 1.$$

Докажите, что плотность случайного вектора $\mathbf{X}$ равна

$$p_{\mathbf{X}}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \right.\right.$$
$$\left.\left. + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2}\right)\right\}$$

Порой хочется сказать, что любой вектор, состоящий из нормальных случайных величин, является гауссовским. Но это неверно.

Пример 10. Пусть ξ_1 и ξ_2 — это независимые стандартные нормальные случайные величины. Построим случайный вектор (X_1, X_2) следующим образом:

$$(X_1, X_2) = \begin{cases}(\xi_1, |\xi_2|), & \xi_1 \geqslant 0\\ (\xi_1, -|\xi_2|), & \xi_1 < 0\end{cases}$$

Данный случайный вектор не будет гауссовским (почему?).

У гауссовских векторов есть одно уникальное свойство, связанное с некоррелированностью.

Теорема 7. *Пусть $\boldsymbol{\xi}$ — гауссовский вектор. Тогда его компоненты независимы тогда и только тогда, когда они некоррелированны.*

Вопрос: допустим, что у нас есть вектор, состоящий из некоррелированных нормальных случайных величин. Можно ли сказать, что он гауссовский? Оказывается, что нет.

Пример 11. Пусть $X \sim \mathcal{N}(0,1)$ и $c \geqslant 0$. Построим по ней новую случайную величину Y следующим образом:

$$Y = \begin{cases}X, & |X| \leqslant c\\ -X, & |X| > c\end{cases}$$

Оказывается, что $Y \sim \mathcal{N}(0,1)$. Тогда $\mathsf{cov}(X,Y) = \mathsf{E}[XY]$, что, в свою очередь, равно $\mathsf{E}[X^2\,\mathrm{I}\{|X| \leqslant c\}] - \mathsf{E}[X^2\,\mathrm{I}\{|X| > c\}]$. Что мы можем сказать про ковариацию?

[6] Несложно понять, что ρ есть коэффициент корреляции между ξ и η.

- Она является непрерывной функцией от c.
- Если $c = 0$, то $\mathsf{cov}(X, Y) = -1$, а если $c = +\infty$, то $\mathsf{cov}(X, Y) = 1$.

В таком случае можно сказать, что есть c такая, что $\mathsf{cov}(X, Y) = 0$. Зафиксируем её. Можно ли сказать, что (X, Y) — это гауссовский вектор? Увы, но нет. Если бы это было так, то X и Y были бы независимы. Но $\mathsf{P}(X > c, Y > c) = 0 \neq \mathsf{P}(X > c)\,\mathsf{P}(Y > c)$.

Теперь расскажем два факта, которые могут понадобиться в дальнейшем и связаны с нормальным распределением.

Пример 12. Пусть (ξ, η) — случайный вектор с совместной плотностью

$$p_{\xi,\eta}(x, y) = C \exp\{-(1 + x^2)(1 + y^2)\}, \text{ где } C = \iint\limits_{\mathbb{R}^2} \exp\{-(1 + x^2)(1 + y^2)\}\, dx\, dy$$

Попробуем найти условную плотность $p_{\xi|\eta}(x \mid y)$. Сразу же заметим, что процесс аналогичен для $p_{\eta|\xi}(y \mid x)$. Для этого найдём плотность η:

$$p_\eta(y) = \int\limits_{-\infty}^{+\infty} Ce^{-(1+x^2)(1+y^2)}\, dx = Ce^{-(1+y^2)} \int\limits_{-\infty}^{+\infty} e^{-x^2(1+y^2)}\, dx =$$

$$= \frac{Ce^{-(1+y^2)}}{\sqrt{y^2+1}} \int\limits_{-\infty}^{+\infty} e^{-u^2}\, du = \frac{C\sqrt{\pi}e^{-(1+y^2)}}{\sqrt{y^2+1}}.$$

Теперь несложно посчитать условную плотность:

$$p_{\xi|\eta}(x \mid y) = \frac{g(x, y)}{p_\eta(y)} = \frac{\sqrt{y^2+1}}{\sqrt{\pi}} e^{-x^2(1+y^2)} = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{x^2}{2\sigma_y^2}}, \text{ где } \sigma_y^2 = \frac{1}{2 + 2y^2}.$$

Из математической статистики вам известны такие понятия, как оценки. Следующая теорема в некоторой степени связана с ними, но это станет ясней позднее, когда дело дойдёт до фильтров Калмана.

Теорема 8 (о нормальной корреляции, одномерный случай)**.** *Пусть (ξ, η) — двумерный гауссовский вектор. Тогда*

$$\mathsf{E}[\eta \mid \xi] = \mathsf{E}[\eta] + \frac{\mathsf{cov}(\xi, \eta)}{\mathsf{D}[\xi]}(\xi - \mathsf{E}[\xi]),$$

$$\Delta = \mathsf{E}\big[(\eta - \mathsf{E}[\eta \mid \xi])^2\big] = \mathsf{D}[\eta] - \frac{\mathsf{cov}^2(\xi, \eta)}{\mathsf{D}[\xi]}.$$

Доказательство. Докажем эту теорему в лоб. Для этого посчитаем условную плотность $p_{\eta|\xi}(y \mid x)$. Формула для совместной плотности была выведена в задаче 1. Пользуясь теми же обозначениями, получим (проверьте!), что

$$p_{\eta|\xi}(y \mid x) = \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}} \exp\left\{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(y - \mu_2 - \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)\right)^2\right\}.$$

Из этого можно сделать следующий вывод:

$$(\eta \mid \xi = x) \sim \mathcal{N}\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1-\rho^2)\right) \implies \mathsf{E}[\eta \mid \xi] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\xi - \mu_1).$$

Заменяя, получим первую часть утверждения. Для доказательства второго утверждения подставим полученный результат и раскроем скобки как квадрат разности:

$$\Delta = \mathsf{E}\left[\left(\eta - \mathsf{E}[\eta] - \frac{\mathsf{cov}(\xi,\eta)}{\mathsf{D}[\xi]}(\xi - \mathsf{E}[\xi])\right)^2\right] = \mathsf{E}\big[(\eta - \mathsf{E}[\eta])^2\big] - \\ - 2\frac{\mathsf{cov}(\xi,\eta)}{\mathsf{D}[\xi]}\,\mathsf{E}[(\eta - \mathsf{E}[\eta])(\xi - \mathsf{E}[\xi])] + \frac{\mathsf{cov}^2(\xi,\eta)}{\mathsf{D}^2[\xi]}\,\mathsf{E}\big[(\xi - \mathsf{E}[\xi])^2\big] = \mathsf{D}[\eta] - \frac{\mathsf{cov}^2(\xi,\eta)}{\mathsf{D}[\xi]}. \quad \square$$

У данной теоремы есть многомерное обобщение.

Теорема 9 (о нормальной корреляции, общий случай)**.** *Пусть* $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *— n-мерный гауссовский вектор. Сделаем следующее разбиение:*

$$\mathbf{X} = \begin{pmatrix}\boldsymbol{\xi}\\ \boldsymbol{\eta}\end{pmatrix} \text{ с размерами } \begin{pmatrix} q \times 1 \\ (n-q)\times 1\end{pmatrix}$$

Соответственным образом вводятся разбиения матожидания и матрицы ковариаций:

$$\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix} \text{ с размерами } \begin{pmatrix} q \times 1 \\ (n-q)\times 1\end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix} \text{ с размерами } \begin{pmatrix} q\times q & q\times(n-q) \\ (n-q)\times q & (n-q)\times(n-q)\end{pmatrix}$$

Тогда верны следующие формулы:

$$\mathsf{E}[\boldsymbol{\xi} \mid \boldsymbol{\eta}] = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\eta} - \boldsymbol{\mu}_2),$$
$$\Delta = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

Доказательство. Для начала найдём линейную комбинацию $\boldsymbol{\delta} = \boldsymbol{\xi} + \mathbf{A}\boldsymbol{\eta}$ такую, что она некоррелирована (а следовательно, и независима) с $\boldsymbol{\eta}$. Для этого распишем корреляцию:

$$\mathsf{cov}(\boldsymbol{\delta}, \boldsymbol{\eta}) = \mathsf{cov}(\boldsymbol{\xi}, \boldsymbol{\eta}) + \mathsf{cov}(\mathbf{A}\boldsymbol{\eta}, \boldsymbol{\eta}) = \mathsf{cov}(\boldsymbol{\xi}, \boldsymbol{\eta}) + \mathbf{A}\,\mathsf{D}[\boldsymbol{\eta}] = \boldsymbol{\Sigma}_{12} + \mathbf{A}\boldsymbol{\Sigma}_{22} \implies \mathbf{A} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}.$$

Отсюда получаем, что

$$\mathsf{E}[\boldsymbol{\xi} \mid \boldsymbol{\eta}] = \mathsf{E}[\boldsymbol{\delta} - \mathbf{A}\boldsymbol{\eta} \mid \boldsymbol{\eta}] = \mathsf{E}[\boldsymbol{\delta} \mid \boldsymbol{\eta}] - \mathsf{E}[\mathbf{A}\boldsymbol{\eta} \mid \boldsymbol{\eta}] = \mathsf{E}[\boldsymbol{\delta}] - \mathbf{A}\boldsymbol{\eta} = \boldsymbol{\mu}_1 + \mathbf{A}(\boldsymbol{\mu}_2 - \boldsymbol{\eta}) = \\ = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\eta} - \boldsymbol{\mu}_2).$$

Теперь посмотрим на то, как себя ведёт Δ. Как известно, по формуле полной вероятности:

$$\Delta = \mathsf{E}\big[(\boldsymbol{\xi} - \mathsf{E}[\boldsymbol{\xi} \mid \boldsymbol{\eta}])^2\big] = \mathsf{E}\big[\mathsf{E}\big[(\boldsymbol{\xi} - \mathsf{E}[\boldsymbol{\xi} \mid \boldsymbol{\eta}])^2 \bigm| \boldsymbol{\eta}\big]\big] = \mathsf{E}[\mathsf{D}[\boldsymbol{\xi} \mid \boldsymbol{\eta}]].$$

Теперь посмотрим на условную дисперсию:

$$\mathsf{D}[\boldsymbol{\xi} \mid \boldsymbol{\eta}] = \mathsf{D}[\boldsymbol{\delta} - \mathbf{A}\boldsymbol{\eta} \mid \boldsymbol{\eta}] = \mathsf{D}[\boldsymbol{\delta} \mid \boldsymbol{\eta}] + \mathsf{D}[\mathbf{A}\boldsymbol{\eta} \mid \boldsymbol{\eta}] - \mathsf{cov}(\boldsymbol{\delta}, \mathbf{A}\boldsymbol{\eta} \mid \boldsymbol{\eta}) - \mathsf{cov}(\mathbf{A}\boldsymbol{\eta}, \boldsymbol{\delta} \mid \boldsymbol{\eta}).$$

Как известно, матожидание вектора, умноженного на матрицу слева/справа, равно матожиданию вектора, умноженного на эту же матрицу слева/справа, а условная дисперсия случайной величины, не зависящей от условия, равна обычной дисперсии. Тогда это равно

$$\mathsf{D}[\boldsymbol{\delta} \mid \boldsymbol{\eta}] + \mathbf{A}\,\mathsf{D}[\boldsymbol{\eta} \mid \boldsymbol{\eta}]\mathbf{A}^\intercal - \mathsf{cov}(\boldsymbol{\delta}, \boldsymbol{\eta} \mid \boldsymbol{\eta})\mathbf{A}^\intercal - \mathbf{A}\,\mathsf{cov}(\boldsymbol{\eta}, \boldsymbol{\delta} \mid \boldsymbol{\eta}) = \mathsf{D}[\boldsymbol{\delta} \mid \boldsymbol{\eta}] = \mathsf{D}[\boldsymbol{\delta}].$$

Осталось посчитать эту дисперсию. Для этого распишем дисперсию суммы, пользуясь свойствами матриц $\mathbf{\Sigma}_{ij}$:

$$\begin{aligned}
\mathsf{D}[\boldsymbol{\delta}] &= \mathsf{D}[\boldsymbol{\xi} + \mathbf{A}\boldsymbol{\eta}] = \mathsf{D}[\boldsymbol{\xi}] + \mathsf{D}[\mathbf{A}\boldsymbol{\eta}] + \mathsf{cov}(\boldsymbol{\xi}, \mathbf{A}\boldsymbol{\eta}) + \mathsf{cov}(\mathbf{A}\boldsymbol{\eta}, \boldsymbol{\xi}) = \\
&= \mathsf{D}[\boldsymbol{\xi}] + \mathbf{A}\,\mathsf{D}[\boldsymbol{\eta}]\mathbf{A}^{\intercal} + \mathsf{cov}(\boldsymbol{\xi}, \boldsymbol{\eta})\mathbf{A}^{\intercal} + \mathbf{A}\,\mathsf{cov}(\boldsymbol{\eta}, \boldsymbol{\xi}) = \\
&= \mathbf{\Sigma}_{11} + (-\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1})\mathbf{\Sigma}_{22}(-\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1})^{\intercal} + \mathbf{\Sigma}_{12}(-\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1})^{\intercal} + (-\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1})\mathbf{\Sigma}_{21} = \\
&= \mathbf{\Sigma}_{11} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{22}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} = \\
&= \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}
\end{aligned}$$

Так как матожидание константы есть сама константа, то $\Delta = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$. □

1.5 Примеры случайных процессов

Наше небольшое введение закончилось. Теперь можно посмотреть на несколько основных примеров случайных процессов.

1.5.1 Гауссовский и винеровский процессы

Многие процессы, которые попадаются на практике, обладают так называемыми независимыми приращениями. Что это значит?

Определение 32. Пусть $X = (X_t)_{t \geqslant 0}$ — некоторый случайный процесс. Будем говорить, что X есть *процесс с независимыми приращениями*, если для любых $t_0, t_1, \ldots, t_n \in T$ таких, что $0 = t_0 < t_1 < \ldots < t_n$ случайные величины X_{t_0}, $X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}}$ независимы в совокупности.

В 1827 году Роберт Броун открыл движение пыльцевых зёрен в жидкости. Исследуя пыльцу под микроскопом, он установил, что в растительном соке плавающие пыльцевые зёрна двигаются совершенно хаотически зигзагообразно во все стороны. В дальнейшем это хаотическое движение назвали *броуновским*. Для его математического описания используется так называемый *винеровский процесс*. Как он вводится?

Определение 33. Случайный процесс $B = (B_t)_{t \geqslant 0}$ называется винеровским, если для него выполнены следующие условия:

1. $B_0 = 0$ почти наверное.
2. B — процесс с независимыми приращениями.
3. $B_t - B_s \sim \mathcal{N}(0, t - s)$ для любых $0 \leqslant s < t < +\infty$.
4. B имеет непрерывные почти наверное траектории, то есть с вероятностью 1 B_t непрерывна, как функция от t.

Обычно наряду с винеровскими процессами вводят *гауссовские процессы*.

Определение 34. Случайный процесс $X = (X_t)_{t \in T}$ называется гауссовским, если все его конечномерные функции распределения являются гауссовскими, то есть задают гауссовский вектор.

Свойство 1. *Гауссовский процесс однозначно определяется своим математическим ожиданием и ковариационной функцией.*

Пример 13. Оказывается, что винеровский процесс является гауссовским. Действительно, возьмём произвольные $t_0, t_1, \ldots, t_n$ таким образом, что $0 = t_0 < t_1 < \cdots < t_n$. Как известно, у винеровского процесса независимые приращения. Следовательно, B_{t_0}, $B_{t_1} - B_{t_0}, \ldots, B_{t_n} - B_{t_{n-1}}$ независимы в совокупности и образуют гауссовский вектор. Теперь поймём, какие распределения имеют $B_{t_0}, B_{t_1}, \ldots, B_{t_n}$. Для этого заметим, что

$$\begin{pmatrix} B_{t_0} \\ B_{t_1} \\ B_{t_2} \\ \vdots \\ B_{t_n} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} B_{t_0} \\ B_{t_1} - B_{t_0} \\ B_{t_2} - B_{t_1} \\ \vdots \\ B_{t_n} - B_{t_{n-1}} \end{pmatrix}$$

Следовательно, $B_{t_i} \sim \mathcal{N}(0, t_i)$ и они образуют гауссовский вектор.

Из этого сразу же получаем, что $\mathsf{E}[B_t] = 0$. Теперь покажем, что $R_B(t_1, t_2) = \min(t_1, t_2)$. Без ограничения общности будем считать, что $t_1 < t_2$. Тогда

$$R_B(t_1, t_2) = \mathsf{E}[B_{t_1} B_{t_2}] = \mathsf{E}[B_{t_1}((B_{t_2} - B_{t_1}) + B_{t_1})] = \mathsf{E}[B_{t_1}(B_{t_2} - B_{t_1})] + \mathsf{E}\big[B_{t_1}^2\big].$$

Так как B_{t_1} и $B_{t_2} - B_{t_1}$ независимы, то $\mathsf{E}[B_{t_1}(B_{t_2} - B_{t_1})] = \mathsf{E}[B_{t_1}]\,\mathsf{E}[B_{t_2} - B_{t_1}] = 0$. Тем самым мы получаем, что $R_B(t_1, t_2) = t_1$.

1.5.2 Процесс Орнштейна-Уленбека

Этот процесс пошёл из теории стохастических дифференциальных уравнений. Пусть $B = (B_t)_{t \geqslant 0}$ — винеровский процесс. Построим по нему новый процесс $X = (X_t)_{t \geqslant 0}$ по следующему правилу: $X_t = e^{-t} B_{e^{2t}}$. Полученный процесс называется *процессом Орнштейна-Уленбека*. Каковы его свойства?

Свойство 1. $\mathsf{E}[X_t] = 0$, $R_X(t, s) = e^{-|t-s|}$.

Доказательство. Первая часть очевидна: $\mathsf{E}[X_t] = \mathsf{E}[e^{-t} B_{e^{2t}}] = 0$. Теперь рассмотрим ковариационную функцию:

$$R_X(t, s) = \mathsf{E}[X_t X_s] = e^{-(s+t)}\,\mathsf{E}[B_{e^{2t}} B_{e^{2s}}] = e^{2\min(t,s)-(s+t)} = e^{-|t-s|}. \qquad \square$$

Свойство 2. *Процесс Орнштейна-Уленбека является гауссовским.*

Доказательство. Без ограничения общности зафиксируем числа $t_1 < t_2 < \ldots < t_n$. Рассмотрим случайный вектор $(X_{t_1}, \ldots, X_{t_n})$. Как он устроен? Распишем последний член:

$$X_{t_n} = e^{-t_n} B_{e^{2t_n}} = e^{-t_n}(B_{e^{2t_n}} - B_{e^{2t_n-1}}) + e^{-t_n} B_{e^{2t_n-1}} = \ldots =$$
$$= e^{-t_n} \sum_{k=1}^{n-1} (B_{e^{2t_{k+1}}} - B_{e^{2t_k}}) + e^{-t_n} B_{e^{2t_1}}$$

Далее, нам известно, что $B_{e^{2t_1}}$, $B_{e^{2t_2}} - B_{e^{2t_1}}, \ldots, B_{e^{2t_n}} - B_{e^{2t_n-1}}$ независимы в совокупности и имеют следующие распределения:

$$B_{e^{2t_1}} \sim \mathcal{N}(0, e^{2t_1}), \quad B_{e^{2t_k}} - B_{e^{2t_k-1}} \sim \mathcal{N}(0, e^{2t_k} - e^{2t_{k-1}}).$$

Осталось заметить, что

$$\begin{pmatrix} X_{t_1} \\ X_{t_2} \\ X_{t_3} \\ \vdots \\ X_{t_n} \end{pmatrix} = \begin{pmatrix} e^{-t_1} & 0 & 0 & \ldots & 0 \\ e^{-t_2} & e^{-t_2} & 0 & \ldots & 0 \\ e^{-t_3} & e^{-t_3} & e^{-t_3} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e^{-t_n} & e^{-t_n} & e^{-t_n} & \ldots & e^{-t_n} \end{pmatrix} \begin{pmatrix} B_{e^{2t_1}} \\ B_{e^{2t_2}} - B_{e^{2t_1}} \\ B_{e^{2t_3}} - B_{e^{2t_2}} \\ \vdots \\ B_{e^{2t_n}} - B_{e^{2t_n-1}} \end{pmatrix}$$

Отсюда получаем, что $(X_{t_1}, \ldots, X_{t_n})$ — действительно гауссовский вектор. $\square$

Теперь посчитаем совместную плотность такого вектора. Для этого выпишем совместную плотность $\boldsymbol{\xi} = (B_{e^{2t_1}}, B_{e^{2t_2}} - B_{e^{2t_1}}, \ldots, B_{e^{2t_n}} - B_{e^{2t_{n-1}}})$. Так как компоненты независимы, то она равна произведению плотностей каждой компоненты:

$$p_{\boldsymbol{\xi}}(x_1, \ldots, x_n) = \frac{1}{\sqrt{2\pi}e^{t_1}} \exp\left\{-\frac{x_1^2}{2e^{2t_1}}\right\} \prod_{k=2}^{n} \frac{1}{\sqrt{2\pi(e^{2t_k} - e^{2t_{k-1}})}} \exp\left\{-\frac{x_k^2}{2(e^{2t_k} - e^{2t_{k-1}})}\right\}$$

Теперь выразим компоненты вектора $\boldsymbol{\xi}$ через компоненты вектора $\boldsymbol{\eta} = (X_1, X_2, \ldots, X_n)$:

$$\begin{cases} B_{e^{2t_1}} = e^{t_1} X_1 \\ B_{e^{2t_2}} = e^{t_2} X_2 \\ \ldots \\ B_{e^{2t_n}} = e^{t_n} X_n \end{cases} \quad \Longrightarrow \quad \begin{cases} B_{e^{2t_1}} = e^{t_1} X_1 \\ B_{e^{2t_2}} - B_{e^{2t_1}} = e^{t_2} X_2 - e^{t_1} X_1 \\ \ldots \\ B_{e^{2t_n}} - B_{e^{2t_{n-1}}} = e^{t_n} X_n - e^{t_{n-1}} X_{n-1} \end{cases}$$

В матричном виде это можно записать следующим образом:

$$\begin{pmatrix} B_{e^{2t_1}} \\ B_{e^{2t_2}} - B_{e^{2t_1}} \\ B_{e^{2t_3}} - B_{e^{2t_2}} \\ \vdots \\ B_{e^{2t_n}} - B_{e^{2t_{n-1}}} \end{pmatrix} = \begin{pmatrix} e^{t_1} & 0 & 0 & \ldots & 0 \\ -e^{t_1} & e^{t_2} & 0 & \ldots & 0 \\ 0 & -e^{t_2} & e^{t_3} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & e^{t_n} \end{pmatrix} \begin{pmatrix} X_{t_1} \\ X_{t_2} \\ X_{t_3} \\ \vdots \\ X_{t_n} \end{pmatrix}$$

Несложно понять, что матрица перехода служит матрицей Якоби и её определитель равен $e^{t_1 + \cdots + t_n}$. Следовательно,

$$p_{\boldsymbol{\eta}}(x_1, \ldots, x_n) = e^{t_1 + \cdots + t_n} p_{\boldsymbol{\xi}}(e^{t_1} x_1, e^{t_2} x_2 - e^{t_1} x_1, \ldots, e^{t_n} x_n - e^{t_{n-1}} x_{n-1}).$$

Подставляя, получаем, что

$$p_{\boldsymbol{\eta}}(x_1, \ldots, x_n) = \frac{e^{t_1}}{\sqrt{2\pi}e^{t_1}} \exp\left\{-\frac{x_1^2 e^{2t_1}}{2e^{2t_1}}\right\} \prod_{k=2}^{n} \frac{e^{t_k}}{\sqrt{2\pi(e^{2t_k} - e^{2t_{k-1}})}} \exp\left\{-\frac{(e^{t_k} x_k - e^{t_{k-1}} x_{k-1})^2}{2(e^{2t_k} - e^{2t_{k-1}})}\right\} =$$

$$= \left(2\pi \prod_{k=2}^{n} \left(1 - e^{2(t_{k-1} - t_k)}\right)\right)^{-\frac{1}{2}} \exp\left\{-\frac{x_1^2}{2} - \frac{1}{2} \sum_{k=2}^{n} \frac{(x_k - x_{k-1} e^{t_{k-1} - t_k})^2}{1 - e^{2(t_{k-1} - t_k)}}\right\}.$$

1.5.3 Пуассновский процесс

Перейдём к одному из самых простых для исследования процессов — к *пуассоновскому потоку.*

Определение 35. Случайный процесс $X = (X_t)_{t \geqslant 0}$ называется пуассоновским потоком с *интенсивностью* λ, если он удовлетворяет трём условиям:

1. $X_0 = 0$ почти наверное.
2. X — процесс с независимыми приращениями.
3. Для всех $0 \leqslant s < t < +\infty$ $X_t - X_s \sim \mathrm{Pois}(\lambda(t - s))$, то есть для любого $n \in \mathbb{Z}_+$

$$\mathsf{P}(X_t - X_s = n) = \frac{(\lambda(t - s))^n}{n!} e^{\lambda(t-s)}.$$

Теперь построим пример такого процесса. Для этого вспомним процесс восстановления, описанный в примере 6.

Пусть для любого натурального n T_n — это iid случайные величины с распределением $\mathrm{Exp}(\lambda)$, то есть их плотность равна $p(z) = \lambda e^{-\lambda z}\,\mathrm{I}\{z \geqslant 0\}$. Дальше, $S_n = T_1 + \ldots + T_n$, а для любого $t \geqslant 0$

$$N_t = \sum_{n=1}^{\infty} \mathrm{I}\{S_n \leqslant t\} = \#\{n \in \mathbb{N} : S_n < t\}.$$

Полученный случайный процесс обладает некоторыми интересными свойствами.

Свойство 1. *Для любого $t > 0$ $N_t \sim \mathrm{Pois}(\lambda t)$.*

Доказательство. Понятно, что N_t принимает значения в $\mathbb{Z}_+$. Следовательно, достаточно показать, что вероятности принять нужное значение будут именно такими, какими они должны быть.

Пусть $n = 0$. Тогда

$$\mathsf{P}(N_t = 0) = \mathsf{P}(T_1 > t) = \int\limits_t^{\infty} \lambda e^{-\lambda x}\,dx = e^{-\lambda t} \int\limits_t^{\infty} \lambda e^{-\lambda(x-t)}\,dx = e^{-\lambda t} = \frac{(\lambda t)^0}{0!} e^{-\lambda t}.$$

Теперь посмотрим на вероятность события $N_t = n$. Она равна

$$\mathsf{P}(N_t = n) = \mathsf{P}(N_t \geqslant n) - \mathsf{P}(N_t \geqslant n+1) = \mathsf{P}(S_n \leqslant t) - \mathsf{P}(S_{n+1} \leqslant t).$$

Для того, чтобы посчитать полученные вероятности, вспомним один факт: сумма n iid случайных величин с распределением $\mathrm{Exp}(\lambda)$ имеет гамма-распределение $\Gamma(1/\lambda, n)$ с плотностью

$$p_n(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{\Gamma(n)} = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}.$$

Тогда эта вероятность равна

$$\mathsf{P}(N_t = n) = \int\limits_0^t \left(\frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} - \frac{\lambda^{n+1} x^n e^{-\lambda x}}{n!} \right) dx = \int\limits_0^t \left(\frac{\lambda^n x^n e^{-\lambda x}}{n!} \right)' dx = \frac{(\lambda t)^n}{n!} e^{-\lambda t}. \qquad \square$$

Следствие. $\mathsf{E}[N_t] = \lambda t$, $\mathsf{D}[N_t] = \lambda t$.

Свойство 2. *$N = (N_t)_{t \geqslant 0}$ — это пуассоновский поток с интенсивностью λ.*

Доказательство. Для начала покажем, что $N_0 = 0$ почти наверное. Действительно, если $N_0 \neq 0$, то $T_1 = 0$, что происходит с нулевой вероятностью. Тем самым $N_0 = 0$ почти наверное.

Теперь докажем, что $(N_t)_{t \geqslant 0}$ удовлетворяет двум последним свойствам. Для этого заметим, что

$$\begin{pmatrix} S_1 \\ S_2 \\ S_3 \\ \vdots \\ S_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ \vdots \\ T_n \end{pmatrix}$$

Так как у матрицы единичный определитель и замена линейна, то

$$p_{S_1,\ldots,S_n}(x_1, x_2, \ldots, x_n) = p_{T_1,\ldots,T_n}(x_1, x_2 - x_1, \ldots, x_n - x_{n-1})$$

Пользуясь независимостью T_n, получаем, что

$$p_{S_1,\dots,S_n}(x_1,x_2,\dots,x_n) = \lambda e^{-\lambda x_1}\prod_{k=2}^{n}\lambda e^{-\lambda(x_k-x_{k-1})}\,\mathrm{I}\{x_k - x_{k-1}\geqslant 0\}$$

После преобразования получаем, что

$$p_{S_1,\dots,S_n}(x_1,x_2,\dots,x_n) = \lambda^n e^{-\lambda x_n}\,\mathrm{I}\{x_n\geqslant x_{n-1}\geqslant\ldots\geqslant x_1\geqslant 0\}$$

Дальше, зафиксируем какие-либо числа $0\leqslant t_1<t_2<\dots<t_n$, $k_1\leqslant k_2\leqslant\ldots\leqslant k_n$ и посмотрим на следующую вероятность:

$$\mathsf{P}\big(N_{t_1}=k_1, N_{t_2}-N_{t_1}=k_2-k_1,\dots,N_{t_n}-N_{t_{n-1}}=k_n-k_{n-1}\big)$$

Поймём, как связать это с S_n. Для этого поймём, как устроено первое условие. Оно означает, что $S_1,\dots,S_{k_1}\leqslant t_1$, а $S_{k_1+1}>t_1$. Аналогично, получаем, что эта вероятность равна

$$\mathsf{P}\big(S_1,\dots,S_{k_1}\in(0,t_1], S_{k_1+1},\dots,S_{k_2}\in(t_1,t_2],\dots,S_{k_{n-1}+1},\dots,S_{k_n}\in(t_{n-1},t_n], S_{k_n+1}>t_n\big)$$

Пользуясь плотностью случайного вектора из S_k, запишем это в виде интеграла:

$$\int\limits_{\substack{0<x_1,\dots,x_{k_1}\leqslant t_1\\ t_1<x_{k_1+1},\dots,x_{k_2}\leqslant t_2\\ \dots\\ t_{n-1}<x_{k_{n-1}+1},\dots,x_{k_n}\leqslant t_n\\ x_{k_n+1}>t_n}}\cdots\int \lambda^{k_n+1}e^{-\lambda x_{k_n+1}}\,\mathrm{I}\{x_{k_n+1}\geqslant x_{k_n}\geqslant\ldots\geqslant x_1\geqslant 0\}\,dx_1\ldots dx_{k_n}\,dx_{k_n+1}$$

Разобъём его в произведение интегралов, положив $k_0=t_0=0$:

$$\lambda^{k_n}\int\limits_{t_n}^{\infty}\lambda e^{-\lambda x_{k_n+1}}\,dx_{k_n+1}\prod_{j=1}^{n}\int\limits_{t_{j-1}<x_{k_{j-1}+1}\leqslant\ldots\leqslant x_{k_j}\leqslant t_j} dx_{k_{j-1}+1}\ldots dx_{k_j}$$

Интегралы в произведении берутся достаточно просто: это объём симплекса. Рассуждая по аналогии с трёхмерным случаем, получаем, что интеграл равен

$$\lambda^{k_n}e^{-\lambda t_n}\prod_{j=1}^{n}\frac{(t_j-t_{j-1})^{k_j-k_{j-1}}}{(k_j-k_{j-1})!}$$

Теперь сделаем подгон: заметим, что

$$k_n = k_n-k_0=\sum_{j=1}^{n}(k_j-k_{j-1}),\quad t_n=t_n-t_0=\sum_{j=1}^{n}(t_j-t_{j-1})$$

Тогда интеграл равен

$$\prod_{j=1}^{n}\frac{\big(\lambda(t_j-t_{j-1})\big)^{k_j-k_{j-1}}}{(k_j-k_{j-1})!}e^{-\lambda(t_j-t_{j-1})}.$$

Какая красота. Отсюда мы сразу получаем оба свойства. Тем самым процесс восстановления для экспоненциального распределения является пуассоновским потоком с интенсивностью λ. □

Следующее свойство связано с понятием стационарных приращений.

Определение 36. Случайный процесс $X = (X_t)_{t\geqslant 0}$ имеет *стационарные приращения*, если для любых $0 \leqslant t_0 < t_1 < t_2 < \ldots < t_n < +\infty$ и $\forall h \geqslant 0$

$$(X_{t_1+h} - X_{t_0+h}, \ldots, X_{t_n+h} - X_{t_{n-1}+h}) \stackrel{d}{=} (X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}}).$$

Свойство 3. *Пуассоновский поток имеет стационарные приращения.*

Доказательство. Следует из того, что $X_{a+h} - X_{b+h} \sim \mathrm{Pois}(a-b) \sim X_a - X_b$ и приращения независимы. □

Свойство 4. *Пусть $N^1, \ldots, N^k$ — независимые пуассоновские потоки с интенсивностями $\lambda_1, \ldots, \lambda_k$. Тогда случайный процесс $N_t = N_t^1 + \ldots + N_t^k$ — это пуассоновский процесс с интенсивностью $\lambda = \lambda_1 + \ldots + \lambda_k$.*

1.6 Стационарность случайных процессов

На практике часто попадаются процессы, которые неизменны во времени. Их принято называть *стационарными*. Это понятие было введено и для случайных процессов, хоть и не в одной ипостаси.

Буквальный перевод вышесказанного на математический язык даёт *стационарность в узком смысле.*

Определение 37. Случайный процесс $X = (X_t)_{t\in T}$ называется сильно стационарным (strong sense stationary, SSS, стационарным в узком смысле), если для любого натурального n, любых индексов $t_1, t_2, \ldots, t_n \in T$ и любого сдвига $h \geqslant 0$

$$F_{t_1+h,\ldots,t_n+h}(x_1, \ldots, x_n) = F_{t_1,\ldots,t_n}(x_1, \ldots, x_n).$$

Пример 14. Последовательность iid случайных величин — это стационарная в узком смысле случайная последовательность.

Пример 15. Винеровский процесс не является стационарным в узком смысле. Действительно, пусть $n = 1$ и $h > 0$. Тогда $B_t \sim \mathcal{N}(0, t)$, а $B_{t+h} \sim \mathcal{N}(0, t+h)$ и $F_{t+h}(x) \neq F_t(x)$.

У стационарных в узком смысле процессов есть одно полезное свойство. Но у него есть требование — процесс должен быть второго порядка. Что это значит?

Определение 38. Случайный процесс $(X_t)_{t\in T}$ называется *процессом второго порядка*, если $\mathsf{E}[X_t^2]$ конечно для всех t (то есть это ограниченная функция от t).

Свойство 1. *Если $(X_t)_{t\in T}$ — стационарный в узком смысле процесс второго порядка, то*

1. *$m(t) = \mu = const$, $D(t) = \sigma^2 = const$.*
2. *Для любых $t_1, t_2, h \in T$ $R_X(t_1, t_2) = R_X(t_1 + h, t_2 + h)$.*

Данное свойство очевидным образом следует из определения стационарного в узком смысле процесса и того, что матожидание, дисперсия и ковариация конечны. Последнее свойство позволяет свести ковариационную функцию к одному аргументу: $R_X(t_1, t_2) = R_X(0, t_2 - t_1) \equiv R_X(t_2 - t_1)$.

Вообще говоря, выполнение этих трёх свойств — это тоже в некоторой степени стационарность. Только в широком смысле.

Определение 39. Случайный процесс $X = (X_t)_{t\in T}$ называется слабо стационарным (стационарным в широком смысле, wide sense stationary, WSS, ковариационно стационарным, стационарным второго порядка),[7] если выполнены следующие условия:

1. $m(t) = \mu = const$, $D(t) = \sigma^2 = const$.
2. Для любых $t_1, t_2, h \in T$ $R_X(t_1, t_2) = R_X(t_1 + h, t_2 + h)$.

Обычно сильная и слабая стационарности идут вместе (если это не так, то что-то пошло не туда). Например, если семейство конечномерных функций распределения полностью задаётся матожиданием и ковариационной функцией, то сильная стационарность равносильна слабой стационарности.

У ковариационной функции стационарного в широком смысле случайного процесса есть несколько свойств:

1. Она неотрицательна в нуле: $R_X(0) = R_X(t, t) = D(t) = \sigma^2 \geqslant 0$.
2. Она чётна: $R_X(-\tau) = R_X(0, -\tau) = R_X(-\tau, 0) = R_X(0, \tau) = R_X(\tau)$.
3. Она ограничена по модулю дисперсией. Действительно, по неравенству Коши-Буняковского-Шварца $|R_X(\tau)| = |R_X(t, t + \tau)| = |\operatorname{cov}(X_t, X_{t+\tau})| \leqslant \sqrt{\mathsf{D}[X_t]\,\mathsf{D}[X_{t+\tau}]} = \sigma^2$.
4. Аналог неотрицательной определённости: для любого натурального n, любого неслучайного вектора $(z_1, \ldots, z_n)$ и любого набора индексов $t_1, \ldots, t_n$
$$\sum_{i,j=1}^{n} R_X(t_i - t_j) z_i z_j \geqslant 0.$$
5. Если $R_X(\tau)$ непрерывна в нуле, то она непрерывна для любого τ.

Доказательство. Пусть $\xi_t = X_t - \mathsf{E}[X_t]$. Пользуясь этим обозначением, распишем разность ковариационных функций.

$$R_X(t_1 + h_1, t_2 + h_2) - R_X(t_1, t_2) = \mathsf{E}[\xi_{t_1+h_1}\xi_{t_2+h_2} - \xi_{t_1}\xi_{t_2}] = \\ = \mathsf{E}[\xi_{t_1+h_1}(\xi_{t_2+h_2} - \xi_{t_2})] + \mathsf{E}[(\xi_{t_1+h_1} - \xi_{t_1})\xi_{t_2}]$$

Далее, по неравенству треугольника:

$$|R_X(t_1 + h_1, t_2 + h_2) - R_X(t_1, t_2)| \leqslant |\,\mathsf{E}[\xi_{t_1+h_1}(\xi_{t_2+h_2} - \xi_{t_2})]| + |\,\mathsf{E}[(\xi_{t_1+h_1} - \xi_{t_1})\xi_{t_2}]|$$

Теперь воспользуемся неравенством Коши-Буняковского-Шварца:

$$|R_X(t_1 + h_1, t_2 + h_2) - R_X(t_1, t_2)| \leqslant \sqrt{\mathsf{E}\left[\xi^2_{t_1+h_1}\right]\mathsf{E}[(\xi_{t_2+h_2} - \xi_{t_2})^2]} + \sqrt{\mathsf{E}[(\xi_{t_1+h_1} - \xi_{t_1})^2]\,\mathsf{E}\left[\xi^2_{t_2}\right]}$$

Осталось показать, что эта сумма стремится к нулю. Покажем, что первый член уходит в ноль (второй рассматривается аналогично). Действительно, по непрерывности $R_X(t, t)$

$$\mathsf{E}\left[(\xi_{t_2+h_2} - \xi_{t_2})^2\right] = \mathsf{E}\left[\xi^2_{t_2+h_2}\right] - 2\,\mathsf{E}[\xi_{t_2+h_2}\xi_{t_2}] + \mathsf{E}\left[\xi^2_{t_2}\right] = R_X(t_2 + h_2, t_2 + h_2) + \\ + R_X(t_2, t_2) - 2R_X(t_2 + h_2, t_2) \xrightarrow[h_2\to 0]{} 0.$$

Следовательно, $R_X(t_1 + h_1, t_2 + h_2) \xrightarrow[h_1,h_2\to 0]{} R_X(t_1, t_2)$ и $R_X(t_1, t_2)$ непрерывна везде. □

[7] Название ковариационной стационарности появилось из-за того, что ковариационная функция зависит только от разности индексов. Стационарность второго порядка же означает постоянство второго момента.

Теперь посмотрим на три класса случайных процессов: IID, SSS и WSS. Есть ли между ними какая-либо связь? Есть. Начнём с очевидной цепочки вложений: IID $\subseteq$ SSS $\subseteq$ WSS. Хотя второе вложение не совсем корректно — не все сильно стационарные процессы являются процессами второго порядка. Если добавить это требование, то вложение станет корректным. Теперь покажем, что все вложения строгие.

- Начнём с SSS $\setminus$ IID. Возьмём случайную последовательность $X = (X_t)_{t\in T}$, устроенную следующим образом: для всех t $X_t = \xi$, где ξ — это какая-то фиксированная случайная величина. Она очевидно является стационарной в сильном смысле и все её сечения одинаково распределены, но она не задаёт последовательность независимых случайных величин.

- Теперь посмотрим на WSS $\setminus$ SSS. Суть примера в том, что мы будем брать разные распределения, у которых совпадают матожидание и дисперсия. Например, пусть $X = (X_t)_{t\geqslant 0}$ — случайная последовательность независимых случайных величин такая, что $X_{2n} \sim \mathrm{Bern}(p)$, а $X_{2n+1} \sim \mathcal{N}(p, p(1-p))$. Понятно, что ни о каком равенстве распределений и речи быть не может, а вот слабая стационарность выполнена (почему?).

- Приведём ещё один пример процесса из SSS $\setminus$ IID. Пусть $\{X_n\}_{n=1}^{\infty}$ — последовательность iid случайных величин. Построим по ней новую последовательность $\{Y_n\}_{n=1}^{\infty}$ по правилу $Y_n = X_n + X_{n+1}$. Понятно, что полученная последовательность не будет состоять из независимых случайных величин, но она будет сильно стационарной.

1.7 Эргодичность случайных процессов

Пусть $X = (X_t)_{t\geqslant 0}$ — какой-то случайный процесс. Как оценить $\mathsf{E}[X_t]$? Достаточно просто: берём N реализаций $X_t(\omega_1), \ldots, X_t(\omega_N)$ и берём их среднее арифметическое. Получается *среднее по ансамблю траекторий*:

$$\widehat{\mu}_\omega = \frac{1}{N}\sum_{k=1}^{N} X_t(\omega_k).$$

Теперь предположим, что у случайного процесса X постоянное матожидание μ. Можно ли тогда заменить усреднение по ансамблю траекторий на *усреднение по времени*

$$\widehat{\mu}_t = \frac{1}{n}\sum_{k=1}^{n} X_{t_k}(\omega)?$$

Если окажется так, что $\widehat{\mu}_t \xrightarrow{\text{с.к.}} \mu$ при $n \to \infty$, то процесс X называют *эргодическим в среднеквадратичном смысле.*

Пример 16. Возьмём последовательность iid случайных величин. Она эргодична в среднеквадратичном смысле. Действительно,

$$\mathsf{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right)^2\right] = \mathsf{D}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{\mathsf{D}[X_1]}{n} \xrightarrow[n\to\infty]{} 0.$$

Пример 17. Возьмём случайную последовательность $X = (X_n)_{n\in\mathbb{N}}$, устроенную следующим образом: для всех n $X_n = \xi$, где ξ — это какая-то фиксированная случайная величина. Она сильно стационарна, но не эргодична, так как все траектории — константы.

Проверять эргодичность по определению — это явно занятие на любителя. Поэтому сформулируем без доказательства одну теорему, которая даёт критерий эргодичности (но для сходимости в среднем, а не в среднеквадратичном):

Теорема 10 (условие Слуцкого). *Слабо стационарный случайный процесс $X = (X_t)_{t\geqslant 0}$ является эргодическим в среднем тогда и только тогда, когда*

$$\lim_{T\to\infty} \frac{1}{T^2} \int_0^T \int_0^T R_X(t_1, t_2)\, dt_1\, dt_2 = 0.$$

Если же X является процессом второго порядка, то он будет эргодическим тогда и только тогда, когда $R_X(t) \to 0$ при $t \to \infty$.

1.8 Генерирование реализаций случайных процессов

Теперь посмотрим, как симулировать различные процессы. Начнём с самого простого — с пуассоновского потока.

1.8.1 Генерирование пуассоновских случайных процессов

Однородный пуассоновский поток событий

Для тех, кто забыл — определение пуассоновского потока дано в примере 6. Единственная сложность в генерации реализации состоит в том, что нужно уметь генерирвать случайные величины из экспоненциального распределения. Но мы можем свободно генерировать случайные величины из $\mathrm{U}(0,1)$. Как получить из него экспоненциальное распределение? Для этого докажем одно утверждение:

Лемма (Метод обратного преобразования). *Пусть X — случайная величина с неубывающей функцией распределения $F : \mathbb{R} \mapsto [0,1]$. Введём обратную функцию $F^{-1} : [0,1] \mapsto \mathbb{R}$ следующим образом: $F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geqslant u\}$. Тогда, если $U \sim \mathrm{U}(0,1)$, то $F^{-1}(U)$ имеет функцию распределения F.*

Доказательство. Действительно, $\mathsf{P}(F^{-1}(U) \leqslant x) = \mathsf{P}(U \leqslant F(x)) = F(x)$. □

Теперь покажем, как генерировать случайные величины из распределения $\mathrm{Exp}(\lambda)$. Рассмотрим функцию распределения:

$$F(x) = 1 - e^{-\lambda x} \implies x = -\frac{1}{\lambda}\ln(1 - F(x)) \implies F^{-1}(x) = -\frac{1}{\lambda}\ln(1-x).$$

Тогда по методу обратного преобразования $-\ln(1-U)/\lambda$ будет иметь распределение $\mathrm{Exp}(\lambda)$. Теперь заметим, что $1 - U \overset{d}{=} U$. Тогда получаем, что $-\frac{1}{\lambda}\ln U$ будет иметь нужное распределение.

Теперь несложно написать алгоритм генерации реализации однородного пуассоновского потока.

Алгоритм 1 Алгоритм генерации реализации однородного пуассоновского потока

Вход: Интенсивность λ, максимальное время T.

1: $t \leftarrow 0, I \leftarrow 0, S \leftarrow \varnothing$
2: сгенерировать $U \sim \mathrm{U}(0,1)$
3: $t \leftarrow t - \ln(U)/\lambda$

4: **while** $t \leqslant T$ **do**
5: $I \leftarrow I + 1, S(I) \leftarrow t$
6: сгенерировать $U \sim \mathrm{U}(0, 1)$
7: $t \leftarrow t - \ln(U)/\lambda$

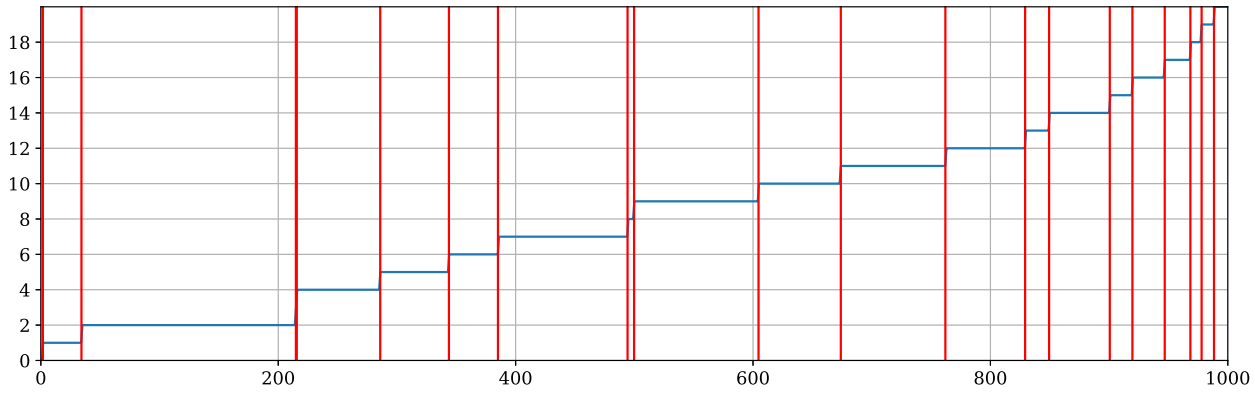


Рис. 1.1: Пример реализации пуассоновского потока с параметрами $T = 1000$, $\lambda = 0.02$

Неоднородный пуассоновский поток событий

Ранее мы смотрели на однородный пуассоновский процесс. Однородный он по той простой причине, что его интенсивность постоянна. Теперь скажем, что λ — это какая-то функция от t. В таком случае получим *неоднородный пуассоновский поток*. Определяется он почти так же, как и однородный, только немного изменяется третье свойство:

$$X_t - X_s \sim \mathrm{Pois}\left(\int_s^t \lambda(x)\,dx\right).$$

Но считать интегралы не очень приятно. Можно ли обойтись без них? Можно. Рассмотрим однородный пуассоновский поток N_t с интенсивностью λ. Пусть событие, появляющееся в момент времени t "засчитывается" с некоторой вероятностью $p(t)$, то есть

$$\mathsf{P}(N_t = N_{t-\varepsilon} + 1) = p(t), \quad \mathsf{P}(N_t = N_{t-\varepsilon}) = 1 - p(t)$$

Оказывается, что N_t — неоднородный пуассоновский поток событий с интенсивностью $\lambda(t) = \lambda p(t)$. Этот результат называется *теоремой Льюиса-Шедлера*.

Пусть $\tilde{\lambda} = \max_{t\in[0,T]} \lambda(t)$. Тогда алгоритм будет выглядеть так:

Алгоритм 2 Алгоритм генерации реализации неоднородного пуассоновского потока

Вход: Интенсивность $\lambda(t)$, максимальное время T.
1: $t \leftarrow 0, I \leftarrow 0, S \leftarrow \varnothing$,
2: сгенерировать $U_1 \sim \mathrm{U}(0, 1)$
3: $t \leftarrow t - \ln(U_1)/\tilde{\lambda}$
4: **while** $t \leqslant T$ **do**
5: сгенерировать $U_2 \sim \mathrm{U}(0, 1)$
6: **if** $U_2 \leqslant \lambda(t)/\tilde{\lambda}$ **then**
7: $I \leftarrow I + 1, S(I) \leftarrow t$
8: сгенерировать $U_1 \sim \mathrm{U}(0, 1)$
9: $t \leftarrow t - \ln(U_1)/\tilde{\lambda}$

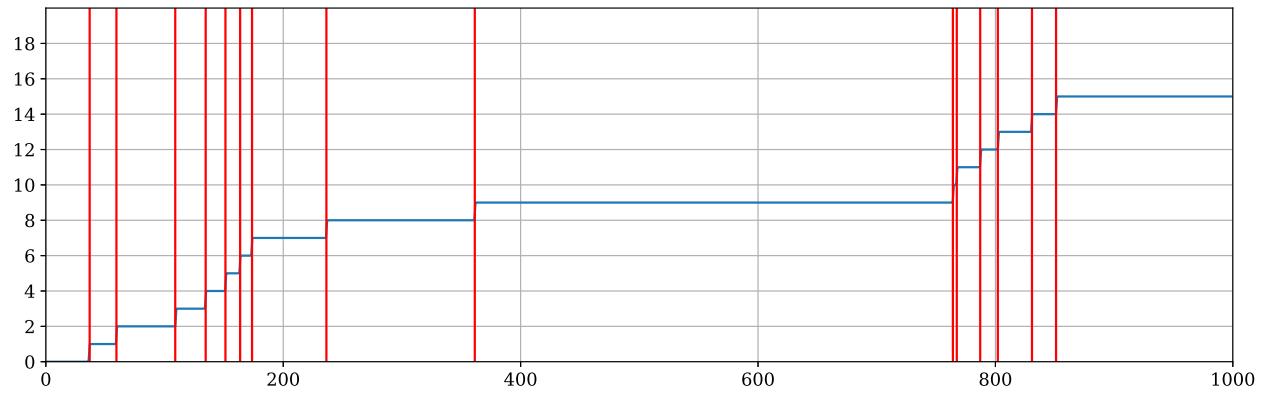


Рис. 1.2: Пример реализации неоднородного пуассоновского потока с параметрами $T = 1000$, $\lambda(t) = (\sin(t/100) + 1)/100$.

1.8.2 Метод стохастического интегрирования

Как известно, дифференциальные уравнения описывают очень многое. Но, оказывается, их можно приспособить и для описания случайных процессов. Основное отличие состоит в том, что в данном случае функция, относительно которой решается уравнение, является случайной величиной. Такие дифференциальные уравнения называют *стохастическими*.

Оказывается, что многие процессы, которые изучаются на практике, "управляются" броуновским движением. Однако есть проблема: траектории винеровского процесса нигде не дифференцируемы почти наверное. Поэтому манипулирование с процессами такого типа потребовало создания собственного исчисления, называемого теорией *стохастических интегралов*. Дадим определение:

Определение 40. Пусть $T = \{t_k\}_{k=0}^n$ — некоторое разбиение отрезка $[0, t]$. Далее, выберем точки $\tau = \{\tau_k\}_{k=1}^n$ по правилу $\tau_k \in [t_{k-1}, t_k]$. Составим по этому разбиению интегральную сумму:

$$S_n = \sum_{k=1}^{n} b(\tau_k)(B_{t_k} - B_{t_{k-1}})$$

Стохастическим интегралом от неслучайной функции $b(t)$ по броуновскому движению $B = (B_t)_{t \geqslant 0}$ называют предел интегральных сумм при диаметре разбиения, стремящемся к нулю:[8]

$$\int_0^t b(x)\, dB_x = \lim_{\Delta T \to 0} S_n$$

Примечание. Не стоит забывать, что $B_{t_i} - B_{t_{i-1}} \sim \mathcal{N}(0, t_i - t_{i-1})$. Это поможет при симуляции процесса.

Многие стохастические процессы могут быть записаны в виде

$$X_t = \int_0^t a(x)\, dx + \int_0^t b(x)\, dB_x,$$

[8] Вопрос о том, почему этот предел вообще существует и каким образом последовательность частичных сумм сходится к нему, оставим за кадром.

где $a(x)$ и $b(x)$ — некоторые неслучайные функции. Это же выражение можно записать в дифференциалах: $dX_t = a(t)\,dt + b(t)\,dB_t$.

Как использовать этот метод? Примерно так же, как и в численном интегрировании: заменить дифференциал на малое изменение и суммировать.

$$X_{t+\varepsilon} - X_t \approx a(t)\varepsilon + b(t)(B_{t+\varepsilon} - B_t).$$

1.8.3 Метод гауссовских векторов

Если нужно сгенерировать не слишком большую реализацию гауссовского процесса, то ситуация становится несколько проще. Как известно, у них все конечномерные функции распределения являются гауссовскими, поэтому реализация будет являться гауссовским вектором. Далее, нам известны математическое ожидание и ковариационная функция процесса. Из этого можно вытащить математическое ожидание и матрицу ковариаций нужного вектора.

Но есть проблема: как генерировать случайный гауссовский вектор с заданным распределением? Сходу это сделать не получится. Для этого проведём одно рассуждение.

Допустим, что вектор одномерный, то есть это просто случайная величина $\xi \sim \mathcal{N}(\mu, \sigma^2)$. Как из стандартного нормального распределения получить нужное распределение? Легко: $\xi \stackrel{d}{=} \mu + \sigma\eta$, где $\eta \sim \mathcal{N}(0, 1)$.

Оказывается, что для общего случая верно нечто похожее. Пусть $\boldsymbol{\mu}$ — некоторый фиксированный вектор, $\boldsymbol{\Sigma}$ — квадратная матрица, а $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{E})$. Какое распределение будет у случайного вектора $\boldsymbol{\eta} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{\xi}$? Так как преобразование линейно, то это гауссовский вектор с параметрами

$$\mathsf{E}[\boldsymbol{\eta}] = \mathsf{E}[\boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{\xi}] = \mathsf{E}[\boldsymbol{\mu}] + \boldsymbol{\Sigma}\,\mathsf{E}[\boldsymbol{\xi}] = \boldsymbol{\mu}$$
$$\mathsf{D}[\boldsymbol{\eta}] = \mathsf{E}[\boldsymbol{\Sigma}\boldsymbol{\xi}(\boldsymbol{\Sigma}\boldsymbol{\xi})^\intercal] = \boldsymbol{\Sigma}\,\mathsf{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\intercal]\boldsymbol{\Sigma}^\intercal = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\intercal.$$

Теперь вспомним один факт из линейной алгебры.

Теорема 11 (Разложение Холецкого)**.** *Пусть $\boldsymbol{\Sigma}$ — неотрицательно определённая симметричная матрица. Тогда существует нижнетреугольная матрица $\mathbf{C}$ с неотрицательными членами на диагонали такая, что $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^\intercal$. Если же $\boldsymbol{\Sigma}$ положительно определена, то все члены $\mathbf{C}$ на диагонали строго положительны.*

Выпишем формулы для вычисления матрицы $\mathbf{C}$:

$$\mathbf{C}_{kk} = \sqrt{\boldsymbol{\Sigma}_{kk} - \sum_{j=1}^{k-1} \mathbf{C}_{ki}^2}, \quad \mathbf{C}_{ij} = \frac{1}{\mathbf{C}_{jj}} \left(\boldsymbol{\Sigma}_{ij} - \sum_{k=1}^{j-1} \mathbf{C}_{ik}\mathbf{C}_{jk} \right)$$

Отсюда понятно, как генерировать гауссовский вектор с заданным распределением $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Для этого мы независимо генерируем n случайных величин с распределением $\mathcal{N}(0, 1)$ (это можно сделать с помощью того же преобразования Бокса-Мюллера) и получаем гауссовский вектор $\boldsymbol{\xi}$. Далее, берём матрицу $\mathbf{C}$ из разложения Холецкого и строим новый вектор $\boldsymbol{\eta} = \boldsymbol{\mu} + \mathbf{C}\boldsymbol{\xi}$. Полученный вектор будет иметь нужное распределение.

1.8.4 Генерирование гауссовских случайных процессов

Винеровский процесс

Сначала разберёмся, как генерировать его с помощью гауссовских векторов. Для этого достаточно сгенерировать матрицу ковариаций и вектор матожиданий. Как известно, $m(t) = 0$, а $R(s, t) = \min(s, t)$.

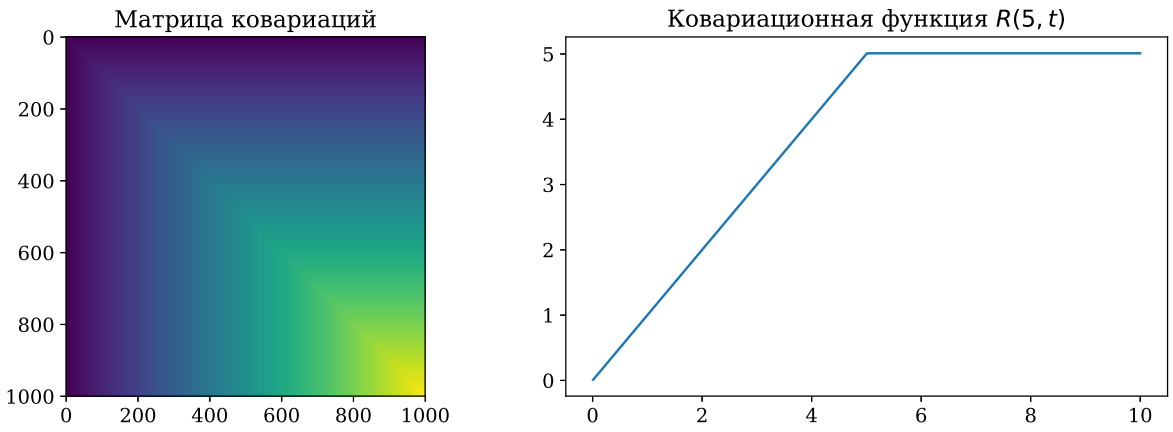


Рис. 1.3: Визуализация матрицы ковариаций и $R(5, t)$ для $t \in (0, 10)$.

Интереснее генерация с помощью стохастического интегрирования. Хотя и данном случае всё достаточно очевидно: B_t можно приблизить суммой достаточно большого числа номальных случайных величин:

$$B_t = \int_0^t dB_x \approx \sum_{k=1}^{N} (B_{t_k} - B_{t_{k-1}}).$$

Для примера посмотрим на первые десять секунд. Для этого разобьём отрезок $[0, 10]$ на 1000 равных кусков и посчитаем эту сумму. Это даст приемлемую точность.

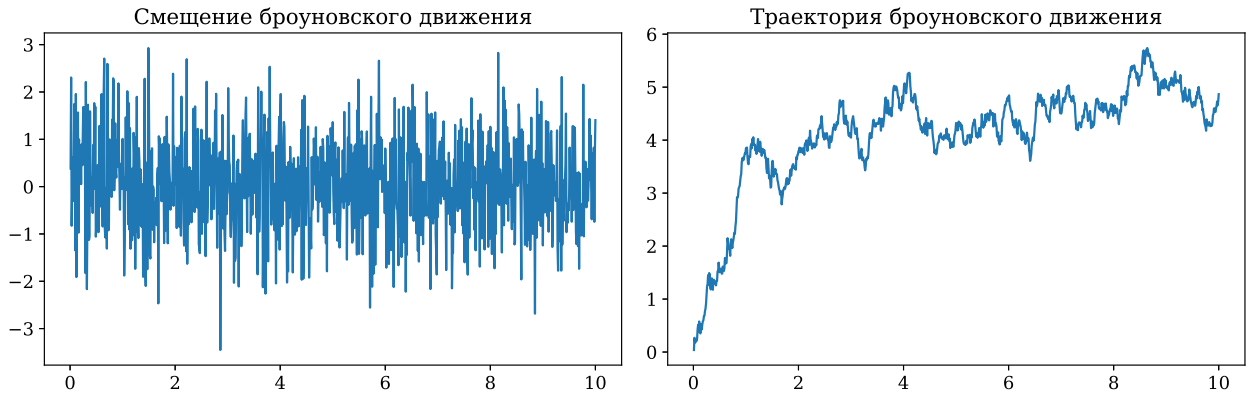


Рис. 1.4: Пример реализации первых десяти секунд броуновского движения.

Процесс Орнштейна-Уленбека

Ранее мы обсуждали процесс Орнштейна-Уленбека. Однако на самом деле он определяется немного по-другому:

Определение 41. Случайный процесс $X = (X_t)_{t \geqslant 0}$, удовлетворяющий стохастическому дифференциальному уравнению

$$dX_t = \theta(\mu - X_t)\, dt + \sigma\, dB_t, \quad X_0 = x_0,$$

называется процессом Орнштейна-Уленбека.

Для того, чтобы получить ранее описанный процесс, нужно подставить $\sigma = \sqrt{2}$, $\theta = 1$, $\mu = 0$ и $x_0 = 0$. Для генерации его реализации с помощью гауссовского вектора достаточно вспомнить, что $\mathsf{E}[X_t] = 0$ и $\mathsf{cov}(X_t, X_s) = e^{-|t-s|}$.

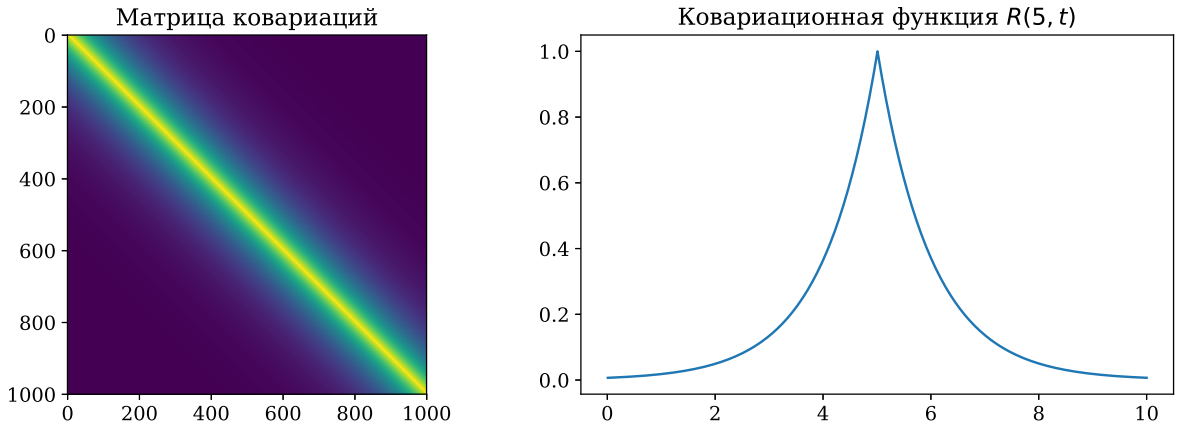


Рис. 1.5: Визуализация матрицы ковариаций и $R(5, t)$ для $t \in (0, 10)$.

Разностная схема устроена следующим образом:

$$X_{t_{i+1}} - X_{t_i} = \theta(\mu - X_{t_i})(t_{i+1} - t_i) + \sigma(B_{t_{i+1}} - B_{t_i}).$$

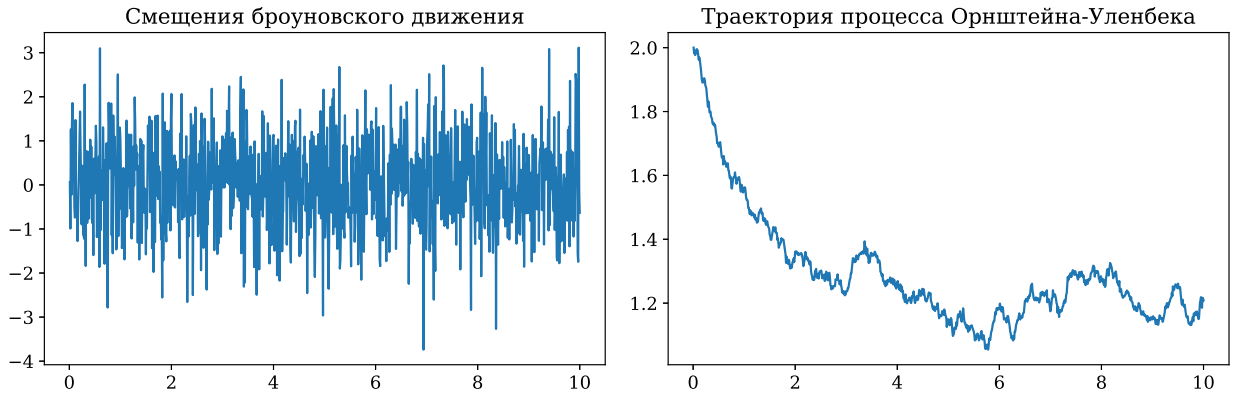


Рис. 1.6: Пример реализации первых десяти секунд процесса Орнштейна-Уленбека с параметрами $x_0 = 2$, $\sigma = 0.1$, $\mu = 1.2$, $\theta = 1$.

Фрактальное броуновское движение

Напоследок рассмотрим ещё один случайный процесс, называемый *фрактальным броуновским движением*.

Определение 42. Фрактальное броуновское движение с *параметром Хёрста* $H \in (0, 1)$ — это гауссовский случайный процесс с непрерывным временем $B^H = (B_t^H)_{t \in [0,T]}$, удовлетворяющий следующим условиям:

- $B_0^H = 0$ почти наверное,
- $\mathsf{E}\left[B_t^H\right] = 0$ для всех $t \in [0, T]$,
- $\mathsf{cov}(B_t^H, B_s^H) = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t - s|^{2H})$.

Оказывается, что если подставить $H = 1/2$, то получится обычное броуновское движение.[9] В остальных случаях получается некоторый гауссовский процесс.

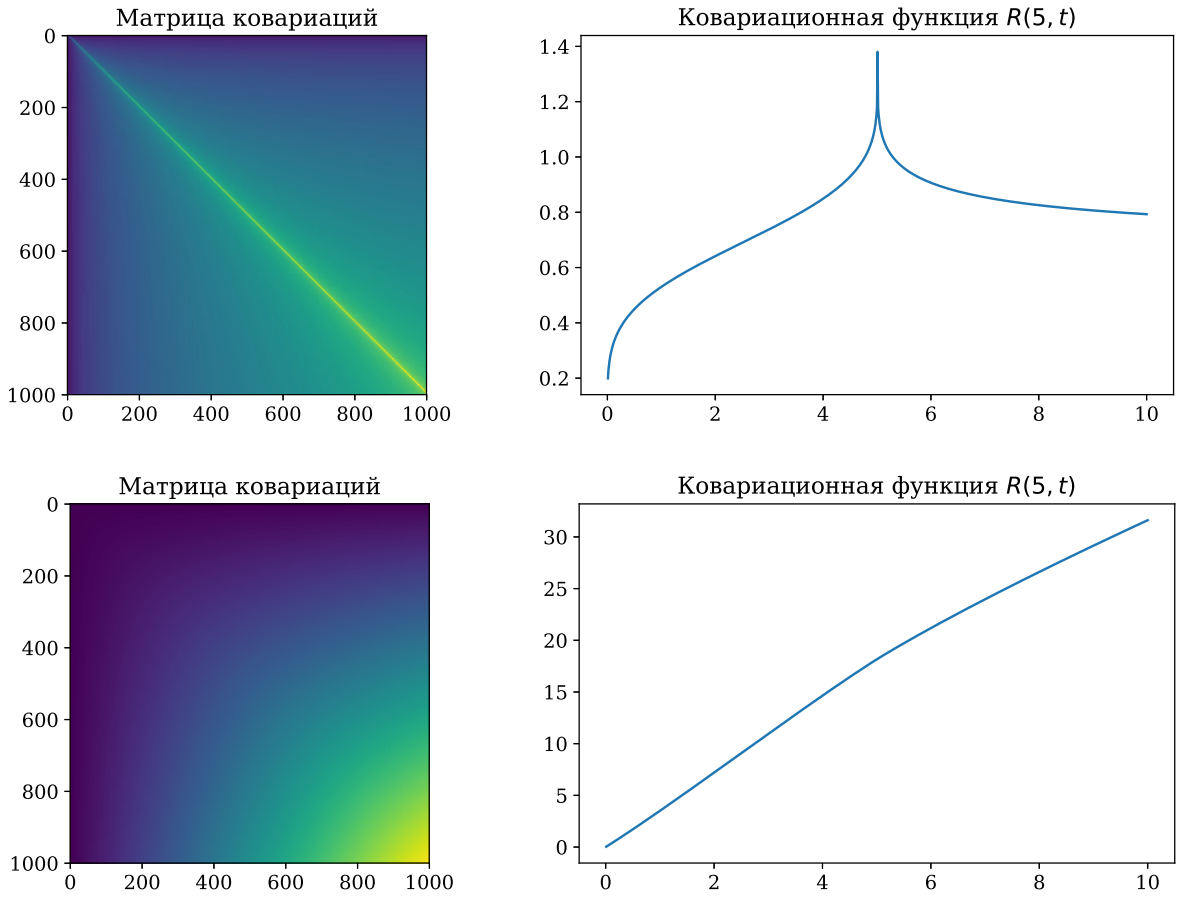


Рис. 1.7: Визуализация матрицы ковариаций и $R(5, t)$ для $t \in (0, 10)$ при $H = 0.1$ и $H = 0.9$.

После того, как была получена матрица ковариаций, дело остаётся за малым: получить нужнул реализацию с помощью разложения Холецкого. Я не буду описывать технические детали, ибо они и так очевидны. Теперь посмотрим, как себя ведут траектории в зависимости от параметра Хёрста.

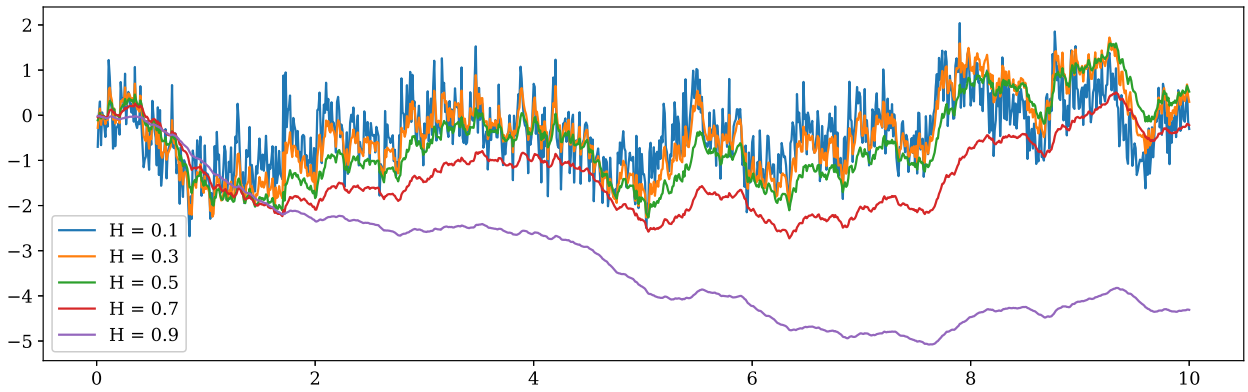


Рис. 1.8: Пример реализаций первых десяти секунд фрактального броуновского движения при разных параметрах Хёрста.

[9] Возникает вопрос о том, что делать с независимостью приращений. Но есть теорема, которая гласит, что фрактальное броуновское движение имеет независимые приращения только при $H = 1/2$.

1.9 Марковские цепи

1.9.1 Основные понятия

Наверное, многие слышали про такое понятие, как марковские цепи. Что это такое? Перед этим дадим несколько необходимых понятий.

Пусть E — это некоторое дискретное (конечное или счётное) множество, которое называют *пространством состояний*. Если система находится в состоянии $i \in E$ в момент времени n, то в момент времени $n+1$ она может перейти в состояние $j \in E$ с *переходной вероятностью* p_{ij}. Это сразу даёт два свойства переходной вероятности:

$$\forall i, j \in E \; p_{ij} \geqslant 0 \text{ и } \forall i \in E \; \sum_{j \in E} p_{ij} = 1.$$

Переходные вероятности образуют *матрицу переходных вероятностей* $P = (p_{ij})_{i,j \in E}$. Теперь можно дать определение марковской цепи.

Определение 43. *Марковская цепь* с пространством состояний E и матрицей переходных вероятностей P — это случайный процесс с дискретным временем $X = (X_n)_{n \in \mathbb{N}}$, $X_n \in E$, для которого

- известны начальные распределения $\alpha_i \equiv \mathsf{P}(X_0 = i)$,
- верно *марковское свойство*: для любого натурального n и любых $i_0, i_1, \ldots, i_{n-1}, i, j$

$$\mathsf{P}(X_{n+1} = j \mid X_n = i) = \mathsf{P}(X_{n+1} = j \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i) = p_{ij},$$

  если условные вероятности хорошо определены, то есть $\mathsf{P}(X_0 = i_0, \ldots, X_n = i) > 0$.

Неформально говоря, марковское свойство означает, что то, как система будет развиваться в текущий момент, не зависит от того, что было в прошлом и зависит только от настоящего.

Теперь вопрос: допустим, что у нас есть какая-то траектория (последовательность состояний). Какова её вероятность? Ответ на этот вопрос даст одна простая теорема.

Теорема 12 (о состояниях марковской цепи)**.** *Для любого натурального n и любых $i_0, i_1, \ldots, i_{n-1}, i, j$*

$$\mathsf{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \alpha_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n}.$$

Доказательство. Индукция по количеству состояний. Пусть $n = 0$. Тогда по определению марковской цепи $\mathsf{P}(X_0 = i_0) = \alpha_{i_0}$.

Теперь предположим, что утверждение верно для n состояний. Покажем, что оно верно и для $n+1$ состояния. Действительно, по определению условной вероятности и марковскому свойству

$$\mathsf{P}(X_0 = i_0, \ldots, X_{n+1} = i_{n+1}) = \mathsf{P}(X_0 = i_0, \ldots, X_n = i_n)\, \mathsf{P}(X_{i+1} = i_{n+1} \mid X_n = i_n) = \\ = \alpha_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n} p_{i_n i_{n+1}}. \quad \square$$

Следствие. *Для любого натурального n и любого $i_n \in E$*

$$\mathsf{P}(X_n = i_n) = \sum_{i_0, \ldots, i_{n-1} \in E} \alpha_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n}.$$

Доказательство. Прямое следствие из формулы полной вероятности и теоремы о состояниях марковской цепи. □

Но обычно нас не интересует полный путь, а лишь начало и конец. Поэтому вводят вероятность перейти из состояния i в состояние j за n шагов:

$$p_{ij}^{(n)} = \mathsf{P}(X_n = j \mid X_0 = i)$$

Чему равна эта вероятность? Воспользуемся теоремой о состояниях:

$$\mathsf{P}(X_n = j \mid X_0 = i) = \frac{\mathsf{P}(X_n = j, X_0 = i)}{\mathsf{P}(X_0 = i)} =$$
$$= \sum_{i_1,\ldots,i_{n-1}\in E} \frac{\mathsf{P}(X_0 = i, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n - j)}{\mathsf{P}(X_0 = i)} = \sum_{i_1,\ldots,i_{n-1}\in E} p_{ii_1}\ldots p_{i_{n-1}j}.$$

Если мы посмотрим на случай $n = 2$, то полученное выражение очень похоже на скалярное произведение строк матрицы переходной вероятности. Оказывается, что это не так уж и далеко от истины.

Теорема 13. *Пусть* $P^{(n)} = (p_{ij}^{(n)})_{i,j\in E}$. *Тогда* $P^{(n)} = P \cdot P \cdot \ldots \cdot P = P^n$.

Доказательство. Индукция по количеству шагов. База ($n = 1$) очевидна, так как $p_{ij}^{(1)} \equiv p_{ij}$. Теперь предположим, что утверждение выполнено для n шагов. Тогда $P^{(n)} = P^n$. Посмотрим на $P^{(n+1)}$:

$$p_{ij}^{(n+1)} = \mathsf{P}(X_{n+1} = j \mid X_0 = i) = \sum_{i_n\in E} \mathsf{P}(X_{n+1} = j \mid X_n = i_n, X_0 = i)\,\mathsf{P}(X_{n+1} = i_n \mid X_0 = i) =$$
$$= \sum_{i_n\in E} \mathsf{P}(X_{n+1} = j \mid X_n = i_n)\,\mathsf{P}(X_{n+1} = i_n \mid X_0 = i) = \sum_{i_n\in E} p_{ii_n}^{(n)} p_{i_nj}.$$

Отсюда получаем, что $P^{(n+1)} = P^{(n)}P = P^{n+1}$. □

Однако это доказательство работает не всегда. Почему же? Потому что никто не обещал, что переходная вероятность не зависит от шага. Если она действительно не зависит, то говорят, что марковская цепь *однородна*.

Теперь поговорим про состояния марковских цепей. В зависимости от переходных вероятностей поведение цепи в этом состоянии может кардинально различаться. Поэтому их классифицируют.

Первая классификация связана с важностью состояния. Может оказаться так, что из состояния можно выйти за конечное число шагов, но вернуться назад уже невозможно. Такие состояния не слишком влияют на долговременное поведение марковской цепи, поэтому их считают несущественными. Формализуем это:

Определение 44. Пусть $X = (X_n)_{n\in\mathbb{N}}$ — марковская цепь с матрицей переходных вероятностей P и дискретным множеством состояний E. Будем называть состояние $i \in E$ *несущественным*, если существует состояние j и натуральное n такое, что $p_{ij}^{(n)} > 0$, но $\forall m \in \mathbb{N}\ p_{ij}^{(m)} = 0$. В противном случае состояние называют *существенным*.

Вторая классификация связана с возвращением.

Определение 45. Состояние i марковской цепи называется *возвратным* (reccurent), если

$$\mathsf{P}(X_n = i \text{ для бесконечно многих } n) = 1.$$

Если же эта вероятность равна нулю, то состояние называют *невозвратным* (transient).

Так как на данный момент мы рассматриваем только дискретный случай, то у нас есть роскошь: можно смотреть на марковскую цепь, как на ориентированный граф, где вершины — это события, а вес ребра — переходная вероятность. Теперь вспомним, что в теории графов вводилось такое понятие, как связность. Для марковских цепей вводится похожее понятие.

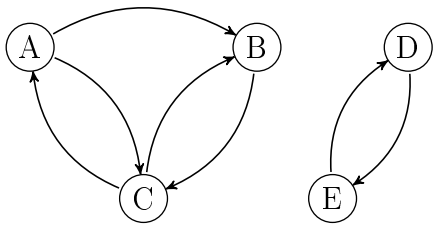


Рис. 1.9: Пример изображения марковской цепи в виде графа. В данном случае $E = \{A, B, C, D, E\}$. Отсутствие ребра означает, что переходная вероятность равна нулю.

Определение 46. Состояние j марковской цепи называются *достижимым* из состояния i, если существует такое натуральное n, что $p_{ij}^{(n)} > 0$. Обозначение: $i \to j$.

Определение 47. Существенные состояния i и j марковской цепи называются *сообщающимися*, если $i \to j$ и $j \to i$. Обозначение: $i \leftrightarrow j$.

У этого понятия есть одно полезное свойство:

Свойство 1. *Сообщаемость задаёт отношение эквивалентности.*

Доказательство. Для начала нужно показать, что отношение сообщаемости рефлексивно. Пусть $i \in E$ существенно. Это означает, что существуют такие натуральные m и n, что $p_{ij}^{(m)} > 0$ и $p_{ji}^{(n)} > 0$. А это и означает, что $i \leftrightarrow i$.

Коммутативность очевидна. Теперь покажем, что выполнена транзитивность. Для этого достаточно показать, что если $i \to j$ и $j \to k$, то и $i \to k$. Действительно, если $p_{ij}^{(m)} > 0$ и $p_{jk}^{(n)} > 0$, то $p_{ik}^{(m+n)} \geqslant p_{ij}^{(m)} p_{jk}^{(n)} > 0$. □

В итоге мы получаем, что марковскую цепь можно разбить на классы сообщающихся вершин, коих будет не более, чем счётное число. Если такой класс один, то марковскую цепь называют *неприводимой* (или *неразложимой*).

Допустим, что мы разбили множество состояний марковской цепи на классы сообщаемости $S_1, \ldots, S_n, \ldots$ и класс несущественных состояний S_0. Как будет выглядеть матрица переходных вероятностей $P^{(n)}$? Оказывается, что она будет иметь блочный вид:

$$P^{(n)} = \begin{pmatrix} S_0 \to S_0 & S_0 \to S_1 & S_0 \to S_2 & \ldots & S_0 \to S_n & \ldots \\ 0 & S_1 \to S_1 & 0 & \ldots & 0 & \ldots \\ 0 & 0 & S_2 \to S_2 & \ldots & 0 & \ldots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \\ 0 & 0 & 0 & \ldots & S_n \to S_n & \ldots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \end{pmatrix},$$

где $S_i \to S_j$ — блочная матрица переходных вероятностей из состояний класса S_i в состояния класса S_j.

Последняя классификация вершин связана с возвращением в состояние. Введём понятие периода состояния.

Определение 48. Пусть $X = (X_n)_{n \in \mathbb{N}}$ — однородная марковская цепь с матрицей переходных вероятностей P и дискретным множеством состояний E. Число $d_i = \gcd\{n \in \mathbb{N} \mid p_{ii}^{(n)} > 0\}$ будем называть *периодом* состояния j. Если $d_j > 1$, то состояние j *периодическое*, иначе же j *апериодическое*.

Осталось ввести одно понятие, которое может понадобиться в дальнейшем.

Определение 49. Состояние i называется *нулевым*, если $p_{ii}^{(n)} \to 0$ при $n \to \infty$.

Теперь посмотрим на какой-нибудь жизненный пример, который можно описать марковской цепью — например, простейшее случайное блуждание. Строится оно так же, как в примере 4, только берётся распределение, принимающее значения 1, 0 и -1 с какими-то вероятностями. Марковская цепь, симулирующая такой процесс, устроена просто. Множеством состояний является $\mathbb{Z}$, а матрица переходных вероятностей задаётся так:

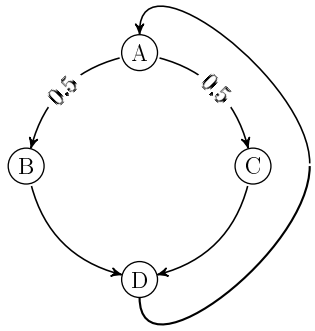


Рис. 1.10: В данной марковской цепи периодическими будут состояния B и C.

$$p_{ij} = \begin{cases} 0, & |i-j| > 1 \\ p_0, & j = i \\ p_1, & j = i+1 \\ p_2, & j = i-1 \end{cases}$$

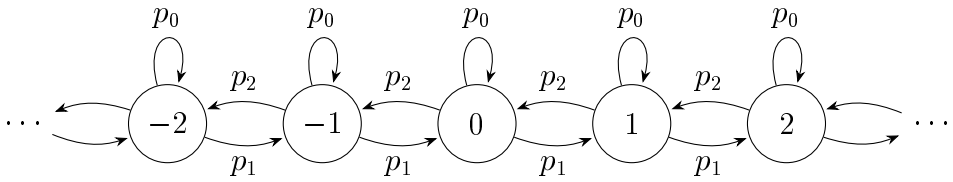


Рис. 1.11: Графическое изображение марковской цепи, соответствующей простому случайному блужданию.

Пример 18. Пусть блуждание начинается в нуле и $p_0 = p_1 = 1/2$. Отсюда сразу же получаем, что $p_2 = 0$. Что мы можем сказать про такое блуждание? А ничего хорошего. Все состояния несущественны, невозвратны и нулевые.

Пример 19. Теперь изменим вероятности: $p_1 = p_2 = 1/2$. Получится так называемое *симметричное случайное блуждание*. В таком случае мы можем сказать, что нулевое состояние периодично с $d_0 = 2$, так как есть ненулевая вероятность вернуться в начало за $2k$ шагов (и понятно, что вернуться за нечётное число шагов невозможно). Вообще, если и p_1, и p_2 больше нуля, то все состояния существенны и сообщаются.

Пример 20. Сузим множество состояний до $\{-k, \ldots, -1, 0, 1, \ldots, k\}$ и изменим вероятности перехода на границах: $p_{kk} = p_{-k,-k} = 1$. Получится *случайное блуждание с поглощающими экранами*. В таком случае существенными состояниями будут только $\pm k$.

Вернёмся к возвратности. Как можно определить, является ли состояние возвратным? По определению это сдедать весьма непросто. Но есть теорема, которая упрощает жизнь.

Теорема 14. *Пусть $f_i = \mathsf{P}(\exists\, n \in \mathbb{N} : X_n = i)$ — вероятность того, что мы хотя бы раз попали в состояние i. Тогда состояние i будет возвратным тогда и только тогда, когда $f_i = 1$.*

Как посчитать f_i? Пусть $A_i^{(n)} = \{$вернуться в i ровно за n шагов$\}$. Вероятность этого события равна:

$$f_i(n) \equiv \mathsf{P}\left(A_i^{(n)}\right) = \mathsf{P}(X_n = i, X_k \neq i, k \in \{1, \ldots, n-1\} \mid X_0 = i)$$

Отсюда несложно получить, что f_i равна сумме $f_i(n)$ по всем натуральным n.

Доказательство. Зафиксируем состояние i. Пусть $B_k = \{X_n = i$ хотя бы k раз$\}$. Это можно записать по-другому:

$$B_k = \{\exists\, n_1, \ldots, n_k \in \mathbb{N} : X_{n_1} = i, \ldots, X_{n_k} = i\}.$$

Что мы можем сказать про такие события? Во-первых, $B_{k+1} \subseteq B_k$, так как если мы посетили состояние i $k+1$ раз, то мы его посетили и k раз. Далее, марковское свойство даёт нам "отсутствие памяти". Тогда $\mathsf{P}(B_k) = f_i^k$.

Теперь заметим, что вероятность возвратности можно записать следующим образом:

$$\mathsf{P}(X_n = i \text{ для бесконечно многих } n) = \mathsf{P}\left(\bigcap_{n=1}^{\infty} B_n\right).$$

По непрерывности вероятностной меры

$$\mathsf{P}\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mathsf{P}(B_n) = \lim_{n\to\infty} f_i^n = \begin{cases} 1, & f_i = 1 \\ 0, & f_i < 1 \end{cases}$$ □

Теорема 15. *Пусть i — состояние марковской цепи. Оно будет возвратным тогда и только тогда, когда расходится ряд $\sum p_{ii}^{(n)}$.*

Доказательство. Зафиксируем состояние i и рассмотрим случайную величину

$$V_i = \sum_{k=1}^{\infty} \mathrm{I}\{X_k = i\} = \#\{n \in \mathbb{N} : X_n = i\}.$$

Что мы можем сказать про неё? Во-первых, $\mathsf{P}(V_i \geqslant k) = \mathsf{P}(B_k) = f_i^k$. Далее, мы можем посчитать матожидание двумя способами:

$$\mathsf{E}[V_i] = \mathsf{E}\left[\sum_{k=1}^{\infty} \mathrm{I}\{X_k = i\}\right] = \sum_{k=1}^{\infty} \mathsf{P}(X_k = i) = \alpha_i \sum_{k=1}^{\infty} p_{ii}^{(n)}.$$

$$\mathsf{E}[V_i] = \sum_{k=1}^{\infty} k\,\mathsf{P}(V_i = k) = \sum_{k=1}^{\infty} \left(\sum_{j=k}^{\infty} \mathsf{P}(V_i = j)\right) = \sum_{k=1}^{\infty} \mathsf{P}(V_i \geqslant k) = \sum_{k=1}^{\infty} f_i^k.$$

Теперь заметим одну вещь: если матожидание конечно, то вероятность того, что $V_i = +\infty$, нулевая. Тогда мы сразу получаем желаемое: если i возвратно, то $f_i = 1$ и ряд $\sum p_{ii}^{(n)}$ расходится. Иначе же ряд $\sum f_i^n$ сходится и матожидание конечно. Следовательно, ряд $\sum p_{ii}^{(n)}$ сходится. □

Теперь применим эту теорему к простейшим случайным блужданиям.

Теорема 16. *Пусть простейшее случайное блуждание задаётся случайными величинами, принимающими значения 1 и -1 с вероятностями p и $q \equiv 1 - p$ соответственно. Если $p = 1/2$, то все состояния возвратны. Иначе же все состояния невозвратны.*

Доказательство. Без ограничения общности рассмотрим состояние $i = 0$. Как известно, соответствующее случайное блуждание задаётся так:

$$X_n = \xi_1 + \ldots + \xi_n, \quad \{\xi_n\}_{n=1}^{\infty} \text{ — iid}, \quad \mathsf{P}(\xi_k = 1) = p, \quad \mathsf{P}(\xi_k = -1) = 1 - p.$$

Заметим, что $\mathsf{E}[\xi_k] = \mathsf{P}(\xi_k = 1) - \mathsf{P}(\xi_k = -1) = 2p - 1$. Тогда оно не ноль при $p \neq 1/2$. Далее, по усиленному закону больших чисел

$$\frac{X_n}{n} \xrightarrow{\text{п.н.}} \mathsf{E}[\xi_1] \implies \begin{cases} X_n \xrightarrow{\text{п.н.}} +\infty, & p > 1/2 \\ X_n \xrightarrow{\text{п.н.}} -\infty, & p < 1/2 \end{cases}$$

Из этого следует, что если $p \neq 1/2$, то все состояния невозвратны. Теперь покажем, что если $p = 1/2$, то состояние возвратно. Для этого посчитаем вероятность вернуться в ноль:

$$p_{ii}^{(n)} = \begin{cases} 0, & n \neq 2k, k \in \mathbb{Z} \\ 2^{-2k} C_{2k}^{k}, & n = 2k, k \in \mathbb{Z} \end{cases}$$

Теперь оценим поведение $p_{ii}^{(n)}$. Для этого можно вспомнить формулу Стирлинга:

$$p_{ii}^{(2k)} \sim \frac{1}{2^{2k}} \frac{\sqrt{4\pi k}(2k)^{2k} e^{-2k}}{2\pi k \cdot k^{2k} e^{-2k}} = \frac{1}{\sqrt{\pi k}} \implies \sum_{n=1}^{\infty} p_{ii}^{(n)} = +\infty.$$

Следовательно, состояние $i = 0$ возвратно. □

Пусть у нас есть какая-то марковская цепь и i — её состояние. Можно ли как-то оценить, сколько времени цепь проводит в этом состоянии? Пусть $T_1, T_2, \ldots$ — это упорядоченные моменты попадания в состояние i. Далее, введём "времена вне состояния i" $T^{(1)} = T_1, T^{(2)} = T_2 - T_1, \ldots$. Если мы рассматриваем однородную марковскую цепь, то эти случайные величины независимы в совокупности и одинаково распределены.

Нередко возникает вопрос об относительном времени пребывания в состоянии i, то есть рассматривается отношение n/N, где n — количество шагов, на которых цепь находилась в состоянии i, а N — общее количество шагов.

Теперь рассмотрим матожидание времени вне состояния i для однородной цепи:

$$\mathsf{E}\big[T^{(n)}\big] = \mathsf{E}\big[T^{(1)}\big] = \sum_{k=1}^{\infty} k\, \mathsf{P}\big(T^{(1)} = k\big).$$

Если состояние i возвратно, то по определению $\mathsf{P}\big(T^{(1)} < \infty\big) = 1$ и мы ничего не можем сказать про сходимость полученного ряда. Однако мы можем сказать две вещи:

- Если матожидание конечно, то $\mathsf{P}\big(T^{(1)} < \infty\big) = 1$ и состояние i возвратно.
- Если состояние невозвратно, то ряд точно расходится. Действительно, в таком случае $\mathsf{P}\big(T^{(1)} = \infty\big) = 1 - \mathsf{P}\big(T^{(1)} < \infty\big) > 0$.

Далее, введём обозначение $m_T \equiv \mathsf{E}\big[T^{(1)}\big]$. Что мы можем сказать про предел отношения n/N при $N \to \infty$?

- Пусть $m_T < \infty$. В таком случае состояние i возвратно и n неограниченно возрастает вместе с ростом $T_n \equiv N$. Далее, по усиленному закону больших чисел:

$$\frac{N}{n} = \frac{T_n}{n} = \frac{1}{n} \sum_{k=1}^{n} T^{(k)} \xrightarrow{\text{п.н.}} \mathsf{E}\big[T^{(1)}\big] = m_T \implies \frac{n}{N} \xrightarrow{\text{п.н.}} \frac{1}{m_T}.$$

  Как известно, если $Z_n \xrightarrow{\mathsf{P}} c$, где c — неслучайная величина, то $\mathsf{E}[Z_n] \to c$. Тогда $\mathsf{E}[n/N] \to 1/m_T$.

- Теперь предположим, что $m_T = \infty$. В таком случае мы уже не можем сказать, что состояние возвратно или не возвратно. Поэтому рассмотрим оба случая. Пусть i невозвратно. Тогда всё просто: мы посещаем состояние конечное число раз с вероятностью 1 и $n/N \xrightarrow{\mathsf{P}} 0$. Оказывается, что и для возвратного состояния выполнена та же сходимость.

Теперь докажем одну теорему, связанную с классами сообщаемости. Оказывается, у состояний в одном классе весьма немало общего.

Теорема 17. *Пусть X — это неразложимая однородная марковская цепь с множеством состояний E и матрицей переходных вероятностей P. Тогда*

- *Если одно из состояний цепи нулевое, то все состояния нулевые.*
- *Если одно из состояний цепи возвратное, то все состояния возвратные.*
- *Если одно из состояний цепи имеет период d, то все состояния имеют тот же период.*

Доказательство. Пусть $i \leftrightarrow j$, то есть существуют натуральные M и N такие, что $\alpha \equiv p_{ij}(M) > 0$ и $\beta \equiv p_{ji}(N) > 0$. Теперь возьмём произвольное натуральное n и распишем $p_{ii}(M+n+N)$:

$$p_{ii}(M+n+N) = \sum_{k,l \in E} p_{ik}(M)p_{kl}(n)p_{li}(N) = \sum_{\substack{k,l \in E \\ k \neq j, l \neq j}} p_{ik}(M)p_{kl}(n)p_{li}(N) + \\ + p_{ij}(M)p_{jj}(n)p_{ji}(N) \geqslant \alpha\beta p_{jj}(n).$$

Теперь можно приступить к доказательству:

1. Пусть i нулевое. Тогда и j нулевое:

$$0 = \lim_{n\to\infty} p_{ii}(M+n+N) = \lim_{n\to\infty} \alpha\beta p_{jj}(n) \implies \lim_{n\to\infty} p_{jj}(n) = 0.$$

2. Пусть i возвратное. По теореме 15:

$$\infty = \sum_{n=1}^{\infty} p_{ii}(M+n+N) = \alpha\beta \sum_{n=1}^{\infty} p_{jj}(n) \implies \sum_{n=1}^{\infty} p_{jj}(n) = \infty.$$

   Отсюда получаем, что и j является возвратным состоянием.

3. Теперь предположим, что i и j имеют периоды d_i и d_j соответсвенно. Докажем, что $d_i = d_j$. Введём два множества:

$$W_i = \{n \in \mathbb{N} : p_{ii}(n) > 0\}, \quad d_i = \gcd W_i,$$
$$W_j = \{n \in \mathbb{N} : p_{jj}(n) > 0\}, \quad d_j = \gcd W_j.$$

   Так как $p_{ii}(M+N) \geqslant \alpha\beta > 0$, то $M+N \in W_i$. Аналогично, $M+N \in W_j$. Теперь построим новое множество:

$$W = \{M+N+n \mid n \in W_i\} \subset W_i.$$

   По построению W имеет общий делитель d_i. Однако для любого $n \in W_i$ $p_{jj}(M+n+M) \geqslant \alpha\beta p_{ii}(n) > 0$. Тогда $W \subseteq W_j$ и любой элемент из W делится на d_j. Тогда можно сказаать, что $d_j \leqslant d_i$. Рассуждая аналогично, получим желаемое.

Рассуждая таким образом для всех состояний, получаем желаемое. □

Эргодичность вводится и для марковских цепей, хоть и немного по-другому.

Определение 50. Пусть $X = (X_n)_{n\in\mathbb{N}}$ — марковская цепь с множеством состояний E и матрицей переходных вероятностей $\mathbf{P}$. Будем называть её *эргодической*, если существует независимое от начального распределения ненулевое предельное распределение вероятностей и состояния $\{p_j^*\}_{j\in E}$ такие, что

$$\forall i,j \in E \lim_{n\to\infty} p_{ij}(n) = p_j^* > 0.$$

Когда марковская цепь является эргодической и как определить предельное распределение? Для этого введём понятие стационарного распределения:

Определение 51. Набор состояний $\{\overline{p}_j\}_{j\in E}$ называется *стационарным распределением вероятностей* (или *инвариантной мерой*) дискретной марковской цепи, если он не изменяется со временем.

Найти стационарное распределение $\mathbf{p}$ не так уж и сложно. Заметим, что

$$p_j(n+1) = \sum_{i\in E} p_i(n)p_{ij} \implies \overline{p}_j = \sum_{i\in E} \overline{p}_i p_{ij} \implies \mathbf{p} = \mathbf{pP}.$$

Теперь можно сформулировать теорему (к сожалению, без доказательства):

Теорема 18 (первая эргодическая)**.** *Марковская цепь является эргодической тогда и только тогда, когда она неразложима и периодична. При этом её предельное распределение равно стационарному распределению, которое единственно.*

1.9.2 Пример применения марковских сетей: модель системы массового обслуживания

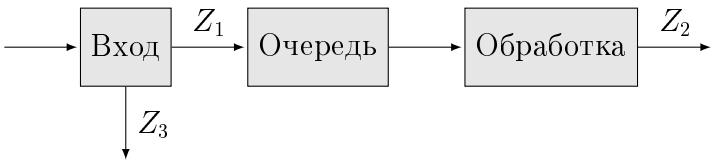


Рис. 1.12: Графическое изображение модели системы массового обслуживания. Z_1 — поток объектов, принятых на обработку, Z_2 — поток обработанных объектов, Z_3 — поток объектов, отклонённых от обработки из-за переполненности очереди.

Рассмотрим модель некоторой организационной системы, предназначенной для массовой обработки однотипных объектов и состоящей из *устройства обработки* и *очереди*, способной хранить N объектов, ожидающих обработки.

- Будем считать, что время дискретно: что-либо происходит только в моменты времени $t + k\tau$, где $k \in \mathbb{Z}$, а t и τ — какие-то фиксированные отрезки времени.
- В каждый момент времени с вероятностью p появляется новый вызов.
- Если в момент времени t ещё не была закончена обработка, то она заканчивается в момент времени $t + \tau$ с вероятностью q.

- Процессы поступления и обработки независимы.
- Объект проходит три стадии: приём (в случае, если в очереди есть место), ожидание и обработка, после чего направляется на выходной поток.
- Очередь имеет объём $N < \infty$. Из этого следует, что система всего вмещает в себя $N+1$ объект и множество состояний можно представить в виде $E = \{0, 1, \ldots, N, N+1\}$. $N+1$ возникает из-за того, что на обработке может находиться 0 или 1 объект.

Теперь введём три события:

$$A_k(t) = \{\text{в момент времени } t \text{ в системе ровно } k \text{ объектов}\}$$
$$B = \{\text{в момент времени } t \text{ в системе появился новый объект}\}$$
$$C = \{\text{в момент времени } t+\tau \text{ система закончила обрабатывать объект}\}$$

Теперь, как выразить $A_k(t+\tau)$ через них? Начнём с нуля. Заметим, что в системе может не оказаться объектов в двух случаях:

1. В системе в предыдущий момент времени не было объектов. Тогда либо в систему не подали новый объект, либо подали, но система закончила обрабатывать другой обьект.
2. В системе был один объект, обработка которого закончилась. При этом нового объекта не поступило.

Формально это можно записать так:

$$A_0(t+\tau) = (A_0(t) \cap (\overline{B} \cup (B \cap C))) \cup (A_1(t) \cap \overline{B} \cap C).$$

Теперь посмотрим на крайний случай — $A_N(t+\tau)$. Рассуждая аналогичным образом, получаем, что

$$A_N(t+\tau) = (A_{N-1}(t) \cap B \cap \overline{C}) \cup (A_N(t) \cap ((B \cap C) \cup (\overline{B} \cap \overline{C}))) \cup (A_{N+1}(t) \cap C).$$

Отсутсвие $\overline{B}$ в последней скобке объясняется тем, что $A_{N+1}(t)$ уже влечёт это событие. В остальных случаях же

$$A_k(t+\tau) = (A_{k-1}(t) \cap B \cap \overline{C}) \cup (A_k(t) \cap ((B \cap C) \cup (\overline{B} \cap \overline{C}))) \cup (A_{k+1}(t) \cap \overline{B} \cap C).$$

Осталось заметить, что $A_0(t) \cup A_1(t) \cup \ldots A_{N+1}(t) = V$, где V — достоверное событие. Теперь можно записать вероятности этих событий, пользуясь независимостью:

$$\begin{cases} \mathsf{P}_0(t+\tau) = \mathsf{P}_0(t)(1-p+pq) + \mathsf{P}_1(t)(1-p)q \\ \mathsf{P}_k(t+\tau) = \mathsf{P}_{k-1}(t)p(1-q) + \mathsf{P}_k(t)(pq + (1-p)(1-q)) + \mathsf{P}_{k+1}(t)(1-p)q \\ \mathsf{P}_N(t+\tau) = \mathsf{P}_{N-1}(t)p(1-q) + \mathsf{P}_N(t)(pq + (1-p)(1-q)) + \mathsf{P}_{N+1}(t)q \\ \mathsf{P}_0(t+\tau) + \ldots + \mathsf{P}_{N+1}(t+\tau) = \mathsf{P}_0(t) + \ldots + \mathsf{P}_{N+1}(t) = 1. \end{cases}$$

Теперь предположим, что мы хотим найти стационарное состояние. Тогда система превращается в не очень элегантную систему линейных уравнений:

$$\begin{cases} \mathsf{P}_0 = \mathsf{P}_0(1-p+pq) + \mathsf{P}_1(1-p)q \\ \mathsf{P}_k = \mathsf{P}_{k-1}\,p(1-q) + \mathsf{P}_k(pq + (1-p)(1-q)) + \mathsf{P}_{k+1}(1-p)q \\ \mathsf{P}_N = \mathsf{P}_{N-1}\,p(1-q) + \mathsf{P}_N(pq + (1-p)(1-q)) + \mathsf{P}_{N+1}\,q \\ \mathsf{P}_0 + \ldots + \mathsf{P}_{N+1} = 1. \end{cases}$$

В качестве упражнения оставим то, что решение этой системы выглядит так:

$$\forall k \in \{1, 2, \ldots, N\}\ \mathsf{P}_k = \left[\frac{p(1-q)}{q(1-p)}\right]^k \mathsf{P}_0, \quad \mathsf{P}_{N+1} = (1-p)\left[\frac{p(1-q)}{q(1-p)}\right]^{N+1} \mathsf{P}_0$$

$$\mathsf{P}_0 = \left(\sum_{k=1}^{N}\left[\frac{p(1-q)}{q(1-p)}\right]^k + (1-p)\left[\frac{p(1-q)}{q(1-p)}\right]^{N+1}\right)^{-1}.$$

Оказывается, что данная модель задаёт эргодичечкую марковскую цепь, поэтому стационарные вероятности являются предельными и неплохо описывают поведение системы, устаканивающееся после достаточно большого отрезка времени. Эти соотношения позволяют решать такие практические задачи, как вычисление вероятности отказа от обслуживания поступающего объекта (вероятность $P_{N+1}(t)$), среднего времени ожидания обслуживания, среднего числа объектов, находящихся в бункере и так далее.

1.10 Марковские процессы

Постепенно будем переходить от дискретного случая к общему. Пусть $X = (X_t)_{t\in T}$ — какой-то случайный процесс с непрерывным временем $T \subseteq \mathbb{R}$. Обычно полагается, что $T = \mathbb{R}_+$ или же $T = \mathbb{R}$. Как обобщить марковское свойство? Для этого нужно ввести понятие фильтрации:

Определение 52. Пусть $(\Omega, \mathcal{F}, \mathsf{P})$ — вероятностное пространство и $T \subseteq \mathbb{R}$. Семейство σ-алгебр $\{\mathcal{F}_t\}_{t\in T}$ такое, что

$$\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}, \quad \forall\, t > s, t, s \in T,$$

называется *фильтрацией вероятностного пространства.*

Данное определение можно прояснить следующей интерпретацией: $\mathcal{F}_t$ — совокупность событий, наблюдаемых до момента t (включительно).

Определение 53. Пусть дан случайный процесс $X = (X_t)_{t\in T}$ с непрерывным временем, определённый на некотором вероятностном пространстве. Определим

$$\mathcal{F}_t^X = \sigma\{X_s \mid s \leqslant t, s, t \in T\}.$$

Тогда $\mathbb{F}^X = \{\mathcal{F}_t^X\}_{t\in T}$ называется *естественной фильтрацией* случайного процесса X.

Теперь можно дать понятие марковского свойства.

Определение 54. Пусть $(\Omega, \mathcal{F}, \mathsf{P})$ — вероятностное пространство с фильтрацией $\{\mathcal{F}_t\}_{t\in T}$ по какому-то множеству $T \subseteq \mathbb{R}$. Далее, пусть $(E, \mathcal{E})$ — измеримое пространство (множество состояний и сигма-алгебра на нём). Будем говорить, что случайный процесс $X = (X_t)_{t\in T}$, определённый на этом фильтрованном вероятностном пространстве, удовлетворяет *марковскому свойству*, если $\forall A \in \mathcal{E}$ и любых $s, t \in T$ таких, что $s < t$

$$\mathsf{P}(X_t \in A \mid \mathcal{F}_s) = \mathsf{P}(X_t \in A \mid X_s).$$

Суть этого определения та же, что и в дискретном случае: вероятность не зависит от предыстории.

Определение 55. Случайный процесс $X = (X_t)_{t\in T}$ называется *марковским*, если он удовлетворяет марковскому свойству с естественной фильтрацией $\mathbb{F}^X$.

1.10.1 Дискретное множество событий. Ступенчатые процессы

Теперь посмотрим на случай, когда E не более, чем счётно. В таком случае случайный процесс называют *марковской цепью с непрерывным аргументом*. Для краткости такой процесс ещё называют *ступенчатым* (из-за вида траекторий).

Пример 21. Оказывается, что однородный пуассоновский процесс является марковской цепью с непрерывным аргументом (почему?).

Состояния удобно обозначать целыми числами, поэтому обычно полагается, что $E = \mathbb{Z}_+$ или $E = \mathbb{Z}$.

Далее, для них тоже вводятся матрицы переходных вероятностей $\mathbf{P}(t_1, t_2) = \|p_{ij}(t_1, t_2)\|$ по правилу:

$$p_{ij}(t_1, t_2) = \mathsf{P}(X_{t_2} = j \mid X_{t_1} = i)$$

Теперь зафиксируем $0 < t_1 < t < t_2$. Тогда распишем переходную вероятность, пользуясь марковским свойством.

$$p_{ij}(t_1, t_2) = \mathsf{P}(X_{t_2} = j \mid X_{t_1} = i) = \sum_{k \in E} \mathsf{P}(X_{t_2} = j \mid X_t = k, X_{t_1} = i)\, \mathsf{P}(X_t = k \mid X_{t_1} = i) =$$

$$= \sum_{k \in E} \mathsf{P}(X_{t_2} = j \mid X_t = k)\, \mathsf{P}(X_t = k \mid X_{t_1} = i) = \sum_{k \in E} p_{ik}(t_1, t) p_{kj}(t, t_2).$$

Отсюда получается так называемое *уравнение Колмогорова-Чепмена*, аналог которого мы получали ранее:

$$\mathbf{P}(t_1, t_2) = \mathbf{P}(t_1, t)\mathbf{P}(t, t_2), \quad t_2 > t > t_1 > 0.$$

Теперь приступим к конструктивному описанию ступенчатого процесса. При этом наша цель будет состоять в получении соотношений, описывающих изменение распределения вероятностей состояния процесса во времени.

Введём специальный процесс $Z = (Z_t)_{t \geqslant 0}$ по правилу $Z_t = \mathrm{I}\{\xi < t\}$, где ξ — неотрицательная случайная величина с функцией распределения F_ξ и плотностью p_ξ. Оказывается, что Z является ступенчатым процессом.

Действительно, для любого натурального n и индексов $t_1 < t_2 < \ldots < t_n$ верны следующие импликации: $X_{t_k} = 1 \implies X_{t_{k+1}} = 1$ и $X_{t_{k+1}} = 0 \implies X_{t_k} = 0$. Тогда

$$\mathsf{P}\big(Z_{t_n} = 1 \,\big|\, Z_{t_1} = z_1, \ldots, Z_{t_{n-1}} = z_{n-1}\big) = \begin{cases} \mathsf{P}\big(Z_{t_n} = 1 \,\big|\, Z_{t_{n-1}} = z_{n-1}\big), & z_{n-1} = 0 \\ 1, & z_{n-1} = 1 \end{cases}$$

Аналогично рассматривается случай с $Z_n = 0$. Тем самым мы показали, что выполнено марковское свойство и Z действительно является ступенчатым.

Теперь предположим, что выполнено следующее равенство, где $\lambda(t)$ — неотрицательная функция, называемая *интенсивностью*:

$$\mathsf{P}(X_{t+h} = 1 \mid X_t = 0) = \lambda(t)h + o(h).$$

Что из него следует? Оказывается, что это равенство задаёт распределение ξ:

$$\mathsf{P}(X_{t+h} = 1 \mid X_t = 0) = \mathsf{P}(\xi < t + h \mid \xi \geqslant t) = \frac{\mathsf{P}(t \leqslant \xi < t + h)}{\mathsf{P}(\xi \geqslant t)} = \frac{F_\xi(t + h) - F_\xi(t)}{1 - F_\xi(t)} =$$

$$= \frac{p_\xi(t)h + o(h)}{1 - F_\xi(t)} = \lambda(t)h + o(h) \implies \lambda(t) = \frac{p_\xi(t)}{1 - F_\xi(t)}$$

Из полученного равенства несложно вытащить F_ξ:

$$\frac{(1-F_\xi(t))'}{1-F_\xi(t)} = -\lambda(t) \implies F_\xi(t) = 1 - \exp\left\{-\int\limits_0^t \lambda(\tau)\,d\tau\right\}.$$

Если же $\lambda(t) = \lambda = const$, то $\xi \sim \mathrm{Exp}(\lambda)$. А, как известно, экспоненциальное распределение обладает эффектом отсутствия памяти, то есть

$$\mathsf{P}(\xi > s+t \mid \xi \geqslant s) = \mathsf{P}(\xi > t).$$

Воспользуемся процессом Z для описания марковской цепи X с непрерывным временем и конечным множеством состояний E. Сопоставим каждому состоянию $i \in E$ случайный процесс Z_i с конечной интенсивностью $\lambda_i(t)$. Каждый из процессов Z_i управляет моментом перехода процесса $X(t)$ из i-го состояния в какое-либо другое. Введем для каждого i также относительные переходные вероятности $\|q_{ij}(t)\|$, где $q_{ij}(t)$ есть вероятность перейти в состояние j в момент времени t при условии, что в этот момент произошёл скачок в процессе Z_i (он управляет переходами) и до этого процесс находился в состоянии i. Понятно, что

$$\sum_{j\in E, j\neq i} q_{ij}(t) = 1.$$

Теперь поймём, как устроена реализация такого процесса. Пусть i_0 — это начальное состояние процесса X. В момент времени $t_0 = 0$ запускается процесс Z_{i_0}. В момент его скачка t_1 в соответствии с вероятностями $\{q_{i_0 j}(t_1)\}_{j\in E}$ происходит переход его в иное (например, k-ое) состояние. В этот же момент запускается процесс Z_k с интенсивностью перехода $\lambda_k(t)$. Далее, в момент t_2 его скачка разыгрывается с вероятностями $\{q_{kj}(t_2)\}_{j\in E}$ переход процесса X в следующее состояние и так далее.

Теперь предположим, что все $\lambda_k(t)$ и $q_{ij}(t)$ есть константы. В таком случае процесс X однороден. Пока что ограничимся такими процессами.

Что мы можем сказать про вероятности для однородного ступенчатого процесса? Для достаточно малых h мы можем заменить интеграл его линейным приближением. Тогда

$$p_{ij}(h) \equiv p_{ij}(t, t+h) = \mathsf{P}(X_{t+h} = j \mid X_t = i) = \begin{cases} \lambda_i q_{ij} h + o_{ij}(h), & i \neq j \\ 1 - \lambda_i h + o_{ii}(h), & i = j \end{cases}$$

где $o_{ij}(h)/h \to 0$ при $h \to 0$.

Далее, пусть $p_i(t) = \mathsf{P}(X_t = i)$. Применим формулу полной вероятности и выразим $p_j(t+h)$:

$$p_j(t+h) = \sum_{i\in E} p_{ij}(h) p_i(t) = p_j(t)(1-\lambda_j h) + h\left(\sum_{i\in E, i\neq j} p_i(t)\lambda_i q_{ij} + \sum_{i\in E} \frac{o_{ij}(h)}{h} p_i(t)\right).$$

Преобразуем выражение:

$$\frac{p_j(t+h) - p_j(t)}{h} = -\lambda_j p_j(t) + \sum_{i\in E, i\neq j} p_i(t)\lambda_i q_{ij} + \sum_{i\in E} \frac{o_{ij}(h)}{h} p_i(t).$$

Теперь устремим h к нулю (это можно сделать, так как предел выражения справа действительно существует):

$$\frac{dp_j(t)}{dt} = -\lambda_j p_j(t) + \sum_{i\in E, i\neq j} p_i(t)\lambda_i q_{ij}.$$

Можно ввести коэффициенты

$$a_{ij} = \begin{cases} -\lambda_j, & i = j \\ \lambda_i q_{ij}, & i \neq j \end{cases} \implies \frac{dp_j(t)}{dt} = \sum_{i \in E} a_{ij} p_i(t).$$

Получилось уравнение, которое называют *уравнением Колмогорова*. В принципе, эти рассуждения работают и для счётного E, только нужно добавить равномерную по i сходимость $o_{ij}(h)/h$ к нулю и ограниченность λ_i.

Мы получили уравнения, описывающие эволюцию распределения вероятностей состояний ступенчатого процесса во времени. Теперь хотелось бы получить аналогичные уравнения для переходных вероятностей. В данном случае с одним состоянием i связывается не один процесс, а целое семейство процессов $\{Z_{ij}\}_{j \neq i}$ с интенсивностями $\{\lambda_{ij}(t)\}_{j \neq i}$.

Первый скачок в этой совокупности потоков, одновременно "запускаемых" каждый раз в момент перехода процесса X в какое-либо очередное i-ое состояние, определяет момент перехода его в новое состояние и номер этого состояния (равный значению второго индекса того случайного процесса $Z_{ij}(t)$, в котором раньше произошел скачок).

Опять же, несложно понять, что при малых h верны следующие формулы :

$$p_{ij}(t, t+h) = \begin{cases} \lambda_{ij}(t)h + o_{ij}(h), & i \neq j \\ 1 - \sum_{k \in E, k \neq j} \lambda_{ik}(t)h + o_{ii}(h), & i = j \end{cases}$$

Теперь введём три условия:

- Пусть все $\lambda_{ij}(t)$ есть константы λ_{ij}.
- Далее, пусть $o_{ij}(h)/h$ равномерно сходится к нулю одновременно по i и по j.
- Напоследок, затребуем сходимость ряда $\sum \lambda_{ij}$.

Теперь снова воспользуемся формулой полной вероятности для моментов времени $t_0 < t < t + h$:

$$\begin{aligned} p_{ij}(t_0, t+h) &= \sum_{k \in E} p_{ik}(t_0, t) p_{kj}(t, t+h) = \\ &= p_{ij}(t_0, t) p_{jj}(t, t+h) + \sum_{k \in E, k \neq j} p_{ik}(t_0, t) p_{kj}(t, t+h) = \\ &= p_{ij}(t_0, t) \left(1 - \sum_{k \in E, k \neq j} \lambda_{jk} h\right) + \sum_{k \in E, k \neq j} \lambda_{kj} h p_{ik}(t_0, t) + o(h) = \\ &= p_{ii}(t_0, t) + h \sum_{k \in E} b_{kj} p_{ik}(t_0, t) + o(h), \end{aligned}$$

где

$$b_{kj} = \begin{cases} \lambda_{kj}, & k \neq i \\ -\sum_{i \in E, i \neq j} \lambda_{ji}, & k = i \end{cases}$$

После преобразований и устремления h к нулю, получаем, что

$$\frac{\partial p_{ij}(t_0, t)}{\partial t} = \sum_{k \in E} p_{ik}(t_0, t) b_{kj}$$

Это так называемое *прямое уравнение Колмогорова*. Несложно показать, аналоничным образом рассматривая формулу полной вероятности для $t_0 < t_0 + h < t$, что верно и *обратное уравнение Колмогорова*:

$$\frac{\partial p_{ij}(t_0, t)}{\partial t_0} = -\sum_{k \in E} b_{ik} p_{kj}(t_0, t)$$

1.10.2 Общий случай

Теперь будем смотреть на случай, когда континуально не только время, но и множество состояний. В таком случае обычно полагается, что $E = \mathbb{R}$. Однако в данном случае рассматриваются не переходные вероятности, а совместные функции распределения или плотности (если они есть) для пары сечений $X_1 = X_{t_1}$ и $X_2 = X_{t_2}$:

$$F(x, t_2 \mid y, t_1) \equiv \mathsf{P}(X_{t_2} \leqslant x \mid X_{t_1} = y), \quad p(x, t_2 \mid y, t_1) \equiv \frac{\partial F(x, t_2 \mid y, t_1)}{\partial x}.$$

Можно ли найти связь для функций распределения в разные моменты времени, аналогичную дискретному случаю? Можно. Для начала сделаем неформальное обоснование. Пусть $t_0 < t < t_1$. Тогда для любого разбиения множества состояний $\{v_n\}_{n\in\mathbb{Z}}$ верно следующее:

$$\mathsf{P}(X_{t_1} \leqslant x \mid X_{t_0} = y) = \sum_{n=-\infty}^{+\infty} \mathsf{P}(X_{t_1} \leqslant x \mid v_n \leqslant X_t < v_{n+1}, X_{t_1} = y)\, \mathsf{P}(v_n \leqslant X_t < v_{n+1} \mid X_{t_0} = y)$$

Введём обозначение:

$$\Delta F(v_i, t \mid y, t_0) = F(v_{i+1}, t \mid y, t_0) - F(v_i, t \mid y, t_0) = \mathsf{P}(v_n \leqslant X_t < v_{n+1} \mid X_{t_0} = y)$$

Тогда

$$\mathsf{P}(X_{t_1} \leqslant x \mid X_{t_0} = y) = \sum_{n=-\infty}^{+\infty} \mathsf{P}(X_{t_1} \leqslant x \mid v_n \leqslant X_t < v_{n+1}, X_{t_1} = y) \Delta F(v_n, t \mid y, t_0)$$

Как мы сказали ранее, это верно для любого разбиения прямой. Далее, это будет верно и при предельном переходе при диаметре разбиения, стремящемся к нулю. В пределе получится интеграл Стильтьеса:

$$F(x, t_1 \mid y, t_0) = \int_E F(x, t_0 \mid z, t, y, t_0)\, dF(z, t \mid y, t_0),$$

где $F(x, t_0 \mid z, t, y, t_0) = \mathsf{P}(X_{t_1} \leqslant x \mid X_t = z, X_{t_0} = y)$. Согласно марковскому свойству второе условие можно убрать:

$$F(x, t_1 \mid y, t_0) = \int_E F(x, t_0 \mid z, t)\, dF(z, t \mid y, t_0)$$

Если же есть плотность, то это выражение можно записать немного по-другому, продифференцировав по x:

$$p(x, t_1 \mid y, t_0) = \int_E p(x, t_0 \mid z, t) p(z, t \mid y, t_0)\, dz.$$

Полученные уравнения называются *обобщёнными уравнениями Маркова.*

Основной вопрос, возникающий при изучении непрерывного марковского случайного процесса, как и прежде, состоит в анализе эволюции распределений (функции или плотности распределения) его состояний с течением времени. Эти эволюции описываются первым и вторым уравнениями Колмогорова. Мы докажем первое уравнение Колмогорова для скалярного марковского процесса при достаточно жестких ограничениях.

Теорема 19 (первое уравнение Колмогорова). *Пусть $X = (X_t)_{t\in T}$ — марковский процесс, удовлетворяющий трём условиям:*

- *Условная плотность $p(x, t_1 \mid y, t_0)$ существует, непрерывно дифференцируема по t_0, трижды непрерывно дифференцируема по x_0 и имеет ограниченные производные.*
- *Процесс "непрерывен в L^3", то есть для любого x*

$$\lim_{\Delta t\to 0} \frac{1}{\Delta t}\,\mathsf{E}\big[(X_{t+\Delta t} - X_t)^3 \;\big|\; X_t = x\big] = 0.$$

- *Существуют конечные пределы:*

$$a(x,t) \equiv \lim_{\Delta t\to 0} \frac{1}{\Delta t}\,\mathsf{E}[X_{t+\Delta t} - X_t \mid X_t = x] < \infty,$$
$$b(x,t) \equiv \lim_{\Delta t\to 0} \frac{1}{\Delta t}\,\mathsf{E}\big[(X_{t+\Delta t} - X_t)^2 \;\big|\; X_t = x\big] < \infty.$$

Тогда для любых $t_0 < t$

$$\frac{\partial p(x,t \mid x_0,t_0)}{\partial t_0} + a(x_0,t_0)\frac{\partial p(x,t \mid x_0,t_0)}{\partial x_0} + \frac{b(x_0,t_0)}{2}\frac{\partial^2 p(x,t \mid x_0,t_0)}{\partial x_0^2} = 0.$$

Примечание. Коэффициент $a(x,t)$ описывает начальную скорость изменения условного математического ожидания процесса, поэтому его называют *коэффициентом сноса*. Коэффициент $b(x,t)$ равным образом указывает на существование конечной начальной скорости изменения условной дисперсии процесса. Отсюда появилось его название — *коэффицент диффузии*.

Полученное дифференциальное уравнение в частных производных позволяет находить функцию условной плотности распределения по ее частным производным в начальный момент времени. В связи с этим это уравнение часто называют *обратным уравнением Колмогорова*.

Доказательство. Возьмём три момента времени $t_0 < t_0 + \Delta t < t$ и рассмотрим соответсвующие сечения. Для них обобщённое уравнение Маркова имеет вид

$$p(x,t \mid x_0,t_0) = \int\limits_E p(x,t \mid z,t_0+\Delta t)p(z,t_0+\Delta t \mid x_0,t_0)\,dz.$$

Разложим $p(x,t \mid z,t_0+\Delta t)$ в ряд Тейлора по z в окрестности x_0 до второго порядка с остаточным членом в форме Лагранжа ($\tilde{x} \in [x_0, x]$):

$$p(x,t \mid z,t_0+\Delta t) = p(x,t \mid x_0,t_0+\Delta t) + \frac{\partial p(x,t \mid x_0,t_0+\Delta t)}{\partial x_0}(z-x_0) + \\ + \frac{1}{2}\frac{\partial^2 p(x,t_0 \mid x_0,t_0+\Delta t)}{\partial x_0^2}(z-x_0)^2 + \frac{1}{6}\frac{\partial^3 p(x,t_0 \mid \tilde{x},t_0+\Delta t)}{\partial \tilde{x}^3}(z-x_0)^3,$$

Осталось подставить это разложение в интеграл:

$$p(x,t \mid x_0,t_0) = p(x,t \mid x_0,t_0+\Delta t)\int\limits_E p(z,t_0+\Delta t \mid x_0,t_0)\,dz + \\ + \frac{\partial p(x,t \mid x_0,t_0+\Delta t)}{\partial x_0}\int\limits_E (z-x_0)p(z,t_0+\Delta t \mid x_0,t_0)\,dz + \\ + \frac{1}{2}\frac{\partial^2 p(x,t_0 \mid x_0,t_0+\Delta t)}{\partial x_0^2}\int\limits_E (z-x_0)^2 p(z,t_0+\Delta t \mid x_0,t_0)\,dz + \\ + \frac{1}{6}\frac{\partial^3 p(x,t_0 \mid \tilde{x},t_0+\Delta t)}{\partial \tilde{x}^3}\int\limits_E (z-x_0)^3 p(z,t_0+\Delta t \mid x_0,t_0)\,dz$$

Немного преобразуем выражение и поделим на Δt:

$$-\frac{p(x,t \mid x_0,t_0+\Delta t)-p(x,t \mid x_0,t_0)}{\Delta t} = \\ = \frac{\partial p(x,t \mid x_0,t_0+\Delta t)}{\partial x_0}\left[\frac{1}{\Delta t}\int\limits_E (z-x_0)p(z,t_0+\Delta t \mid x_0,t_0)\,dz\right] + \\ + \frac{1}{2}\frac{\partial^2 p(x,t_0 \mid x_0,t_0+\Delta t)}{\partial x_0^2}\left[\frac{1}{\Delta t}\int\limits_E (z-x_0)^2 p(z,t_0+\Delta t \mid x_0,t_0)\,dz\right] + \\ + \frac{1}{6}\frac{\partial^3 p(x,t_0 \mid \tilde{x},t_0+\Delta t)}{\partial \tilde{x}^3}\left[\frac{1}{\Delta t}\int\limits_E (z-x_0)^3 p(z,t_0+\Delta t \mid x_0,t_0)\,dz\right].$$

Теперь устремим Δt к нулю. Тогда, пользуясь условиями теоремы, получаем, что

$$-\frac{\partial p(x,t \mid x_0,t_0)}{\partial t_0} = a(x_0,t_0)\frac{\partial p(x,t \mid x_0,t_0)}{\partial x_0} + \frac{b(x_0,t_0)}{2}\frac{\partial^2 p(x,t \mid x_0,t_0)}{\partial x_0^2}. \qquad \square$$

В целях упрощения вывода первого уравнения Колмогорова и получения её в форме, удобной для практики, мы исходили из существования условной плотности распределения процесса. Можно, однако получить это уравнение и в терминах условных функций распределения, что повышает её общность. Заметим также, что при несколько иных предположениях условие непрерывности процесса при выводе уравнения может быть ослаблено, имея вид:

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}\,\mathsf{P}(|X_{t+\Delta t}-X_t|>\delta \mid X_t = x) = 0.$$

При определенным образом измененных условиях справедливо второе уравнение Колмогорова (называемое также уравнением Колмогорова-Фоккера-Планка), которое приводим без доказательства:

Теорема 20 (второе уравнение Колмогорова)**.**

$$\frac{\partial p(x,t \mid x_0,t_0)}{\partial t} = -a(x,t)\frac{\partial p(x,t \mid x_0,t_0)}{\partial x} + \frac{b(x,t)}{2}\frac{\partial^2 p(x,t \mid x_0,t_0)}{\partial x^2}.$$

Заметим, что случайные процессы рассмотренного типа носят название *диффузионных*, поскольку они описывают, в частности, диффузию частиц.

1.10.3 Модель системы массового обслуживания с непрерывным временем

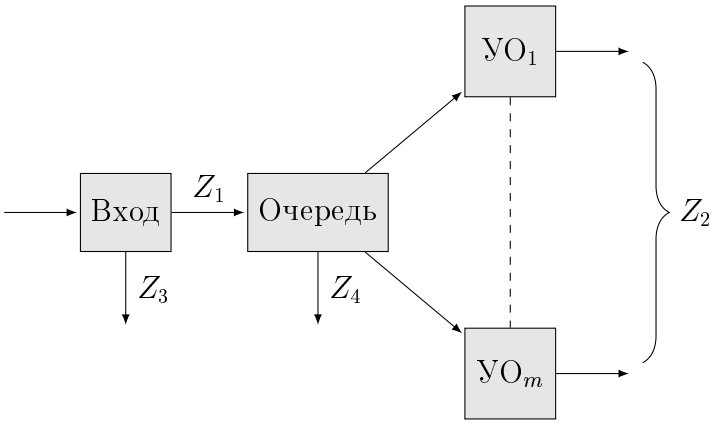


Рис. 1.13: Графическое изображение модели системы массового обслуживания с m устройствами обработки. Z_1 — поток объектов, принятых на обработку, Z_2 — поток обработанных объектов, Z_3 — поток объектов, отклонённых от обработки из-за переполненности очереди, Z_4 — поток объектов, не дождавшихся обработки.

Ранее мы изучали модель системы массового обслуживания в дискретном времени. Теперь обобщим это понятие на непрерывное время, но с некоторыми изменениями. Теперь предположим, что у системы не один поток обработки, а $m < \infty$. Далее, добавим новое свойство: пусть объект, ожидающий обслуживания в очереди (размер которой $n < \infty$), может в случайный момент времени покинуть её без обслуживания, если время пребывания его в очереди окажется больше значения случайной величины – времени ожидания $T_{\text{ож}}$. Такая "нетерпеливость" объектов определяется общим для них распределением этой случайной величины; каждому объекту приписывается её реализация.

Будем считать, что прибытие объектов, требующих обработки, покидание очереди без обработки, поступление на обработку, сама обработка и окончание из обслуживания независимы и могут происходить в любой момент времени.

Далее, состояние системы в момент времени t, как и в дискретном случае, описывается числом находящихся в системе объектов. Понятно, что у системы будет $n+m+1$ состояние. По аналогии с дискретным случаем, введём события:

$$A_k(t) = \{\text{на момент времени } t \text{ в системе ровно } k \text{ объектов}\}.$$

Количество объектов сразу задаёт то, где находятся объекты. Если $k \leqslant m$, то очередь пуста и заняты k линий обработки из m. Если же $m < k \leqslant m + n$, то все линии обработки заняты и в очереди $k - m$ объектов.

Далее, сделаем пару предположений о распределении параметров системы.

- Пусть входной поток описывается однородным пуассоновским процессом с интенсивностью λ.
- Далее, пусть в каждом из устройств обработки время обработки $T_{\text{обр}}$ имеет одно и то же экспоненциальное распределение с параметром $\mu = 1/\,\mathsf{E}[T_{\text{обр}}]$.
- Аналогично, время ожидания обработки для каждого объекта $T_{\text{ож}}$ имеет экспоненциальное распределение с параметром $\nu = 1/\,\mathsf{E}[T_{\text{ож}}]$.

Параметры μ и ν можно называть интенсивностями обслуживания объектов и покидания очереди соответственно. Это три предположения и свойства пуассоновского и экспоненциального распределений позволяют нам сказать, что полученная модель является марковским процессом.

Пусть t и $t+\Delta t$ — два близких момента времени. Введём три обозначения:

- Событие B_k будет обозначать "за промежуток времени $(t, t+\Delta t]$ в систему прибыло k новых объектов".

  Вероятность такого события описать несложно. По сути, она равна вероятности $\mathsf{P}(N_{t+\Delta t} - N_t = k)$:

$$\mathsf{P}(B_k) = \frac{(\lambda\Delta t)^k}{k!}e^{-\lambda\Delta t} \implies \begin{cases} \mathsf{P}(B_0) = 1 - \lambda\Delta t + o(\Delta t) \\ \mathsf{P}(B_1) = \lambda\Delta t + o(\Delta t) \\ \mathsf{P}(B_k) = o(\Delta t), & k > 1 \end{cases}$$

- Событие $D_k^{(l)}$ будет обозначать следующее: к моменту $t+\Delta t$ закончилась обработка l из k объектов, находящихся в момент t в устройствах обработки и (или) поступивших в них за указанный интервал.

  Как посчитать вероятность такого события? Воспользуемся независимостью и тем, что объект будет обрабатываться всё время с вероятностью $e^{-\mu\Delta t}$:

$$\mathsf{P}\left(D_k^{(l)}\right) = \binom{k}{l}(1-e^{-\mu\Delta t})^l(e^{-\mu\Delta t})^{k-l} \implies \begin{cases} \mathsf{P}\left(D_k^{(0)}\right) = 1 - k\mu\Delta t + o(\Delta t) \\ \mathsf{P}\left(D_k^{(1)}\right) = k\mu\Delta t + o(\Delta t) \\ \mathsf{P}\left(D_k^{(l)}\right) = o(\Delta t), & l > 1 \end{cases}$$

- Событие $F_k^{(l)}$ будет обозначать "за этот отрезок очередь покинуло l из k объектов, которые были в очереди до этого". Аналогично предыдущему пункту, вероятность считается так:

$$\mathsf{P}\left(F_k^{(l)}\right) = \binom{k}{l}(1-e^{-\nu\Delta t})^l(e^{-\nu\Delta t})^{k-l} \implies \begin{cases} \mathsf{P}\left(F_k^{(0)}\right) = 1 - k\nu\Delta t + o(\Delta t) \\ \mathsf{P}\left(F_k^{(1)}\right) = k\nu\Delta t + o(\Delta t) \\ \mathsf{P}\left(F_k^{(l)}\right) = o(\Delta t), & l > 1 \end{cases}$$

Если "загнать" события, вероятность которых есть $o(\Delta t)$, в события вида O_k, то можно записать следующие логические тождества:

$$A_0(t+\Delta t) = (A_0(t) \cap B_0) \cup (A_1 \cap B_0 \cap D_1^{(1)}) \cup O_0$$
$$A_1(t+\Delta t) = (A_0(t) \cap B_1 \cap D_1^{(0)}) \cup (A_1(t) \cap B_0 \cap D_1^{(0)}) \cup (A_2(t) \cap B_0 \cap D_2^{(1)}) \cup O_1$$
$$\dots$$

Но при индексах $m \leqslant k \leqslant m+n$ ситуация немного другая:

$$A_m(t+\Delta t) = (A_{m-1}(t) \cap B_1 \cap D_m^{(0)}) \cup (A_m(t) \cap B_0 \cap D_m^{(0)}) \cup$$
$$\cup (A_{m+1}(t) \cap B_0 \cap F_1^{(0)} \cap D_m^{(1)}) \cup (A_{m+1}(t) \cap B_0 \cap F_1^{(1)} \cap D_m^{(0)}) \cup O_m$$

$$A_{m+j}(t+\Delta t) = (A_{m+j-1}(t) \cap B_1 \cap D_m^{(0)} \cap F_{j-1}^{(0)}) \cup (A_m(t) \cap B_0 \cap D_m^{(0)} \cap F_j^{(0)}) \cup \\ \cup (A_{m+1}(t) \cap B_0 \cap ((F_{j+1}^{(0)} \cap D_m^{(1)}) \cup (F_{j+1}^{(1)} \cap D_m^{(0)}))) \cup O_{m+j}$$

$$A_{m+n}(t+\Delta t) = (A_{m+n-1}(t) \cap B_1 \cap D_m^{(0)}) \cup (A_{m+n} \cap F_n^{(0)} \cap D_m^{(0)}) \cup O_{m+n}$$

Напоследок, добавим очевидное тождество: $A_0(t) \cup \ldots \cup A_{n+m}(t) = \Omega$.

Осталось перейти к вероятностям, введя обозначение $p_i(t) = \mathsf{P}(A_i(t))$:

$$\begin{aligned}
p_0(t+\Delta t) &= p_0(t)(1-\lambda\Delta t + o(\Delta t)) + p_1(t)(1-\lambda\Delta t + o(\Delta t))(\mu\Delta t + o(\Delta t)) + o(\Delta t) = \\
&= p_0(t) - \lambda\Delta t p_0(t) + \mu\Delta t p_1(t) + o(\Delta t) \\
p_1(t+\Delta t) &= p_0(t)(\lambda\Delta t + o(\Delta t))(1-\mu\Delta t + o(\Delta t)) + p_1(t)(1-\lambda\Delta t + o(\Delta t))\times \\
&\times (1-\mu\Delta t + o(\Delta t)) + p_2(t)(1-\lambda\Delta t + o(\Delta t))(2\mu\Delta t + o(\Delta t)) + o(\Delta t) = \\
&= \lambda\Delta t p_0(t) + p_1(t) - (\lambda+\mu)\Delta t p_1(t) + 2\mu\Delta t p_2(t) + o(\Delta t) \\
&\ldots \\
p_m(t+\Delta t) &= p_{m-1}(t)(\lambda\Delta t + o(\Delta t))(1-m\mu\Delta t + o(\Delta t)) + p_m(t)(1-\lambda\Delta t + o(\Delta t))\times \\
&\times (1-m\mu\Delta t + o(\Delta t)) + p_{m+1}(t)(1-\lambda\Delta t + o(\Delta t))(m\mu\Delta t + o(\Delta t))\times \\
&\times (1-\nu\Delta t + o(\Delta t)) + p_{m+1}(t)(1-\lambda\Delta t + o(\Delta t))(1-m\mu\Delta t + o(\Delta t))\times \\
&\times (\nu\Delta t + o(\Delta t)) = p_{m-1}(t)\lambda\Delta t + p_m(t) - (\lambda+\mu)\Delta t p_m(t) + \\
&+ (m\mu+\nu)\Delta t p_{m+1}(t) + o(\Delta t) \\
p_{m+j}(t+\Delta t) &= p_{m+j-1}(t)(\lambda\Delta t + o(\Delta t))(1-m\mu\Delta t + o(\Delta t))(1-(j-1)\mu\Delta t + o(\Delta t)) + \\
&+ p_{m+j}(t)(1-\lambda\Delta t + o(\Delta t))(1-m\mu\Delta t + o(\Delta t))(1-j\mu\Delta t + o(\Delta t)) + \\
&+ p_{m+j+1}(t)(1-\lambda\Delta t + o(\Delta t))(m\mu\Delta t + o(\Delta t))(1-(j+1)\mu\Delta t + o(\Delta t)) + \\
&+ p_{m+j+1}(t)(1-\lambda\Delta t + o(\Delta t))(1-m\mu\Delta t + o(\Delta t))((j+1)\mu\Delta t + o(\Delta t)) = \\
&= \lambda\Delta t p_{m+j-1}(t) + p_m(t) - (\lambda + m\mu + j\nu)\Delta t p_{m+j}(t) + \\
&+ (m\mu + (j+1)\nu)\Delta t p_{m+j+1}(t) + o(\Delta t) \\
p_{m+n}(t+\Delta t) &= p_{m+n-1}(t)(\lambda\Delta t + o(\Delta t))(1-m\mu\Delta t + o(\Delta t)) + p_{m+n}(t)\times \\
&\times (1-m\mu\Delta t + o(\Delta t))(1-n\nu\Delta t + o(\Delta t)) = p_{m+n-1}(t)\lambda\Delta t + p_{m+n}(t) - \\
&- (m\mu + n\nu)p_{m+n}(t) + o(\Delta t) \\
1 &= p_0(t) + p_1(t) + \ldots + p_{n+m}(t)
\end{aligned}$$

Попробуем найти стационарное решение. Для этого скажем, что p_i есть константы, после чего устремим Δt к нулю:

$$\begin{cases}
0 = -\lambda p_0 + \mu p_1 & \\
0 = \lambda p_{k-1} - (\lambda + k\mu)p_k + (k+1)\mu p_{k+1}, & 1 \leqslant k < m \\
0 = \lambda p_{m-1} - (\lambda+\mu)p_m + (m\mu+\nu)p_{m+1} & \\
0 = \lambda p_{m+j-1} - (\lambda + m\mu + j\nu)p_{m+j} + (m\mu + (j+1)\nu)p_{m+j+1}, & 1 \leqslant j < n \\
0 = \lambda p_{m+n-1} - (m\mu + n\nu)p_{m+n} & \\
1 = p_0 + p_1 + \ldots + p_m + \ldots + p_{m+n} &
\end{cases}$$

Опять же, в качестве упражнения оставим вывод решения этой системы:

$$\forall k \in \{1,2,\ldots,m\}\ p_k = \frac{(\lambda/\mu)^k}{k!}p_0,$$
$$\forall j \in \{1,2,\ldots,n\}\ p_{m+j} = \frac{(\lambda/\mu)^m}{m!}\frac{\lambda^j}{\prod_{k=1}^{j}(m\mu+k\nu)}p_0,$$
$$p_0 = \left[\sum_{k=0}^{m}\frac{(\lambda/\mu)^k}{k!} + \frac{(\lambda/\mu)^m}{m!}\sum_{j=1}^{n}\frac{\lambda^j}{\prod_{k=1}^{j}(m\mu+k\nu)}\right]^{-1}$$

Осталось показать два частных случая:

- Если устремить ν к бесконечности, то получится *система без очереди*, в которой объекты, которые попали в очередь, моментально выходят из неё. В этом случае $p_{m+j} = 0$ для всех $j \in \{1,2,\ldots,n\}$.
- Если же устремить ν у нулю, то получится рассмотренная ранее система массового обслуживания без покидания очереди.

1.11 Стохастические модели с дискретным временем

Приступим к изучению *временных рядов*, то есть случайных процессов с множеством индекосв $T = \mathbb{Z}_+$. Пусть единица времени измерения есть сутки, а $S = (S_n)_{n\geqslant 0}$ — это какой-то финансовый индекс (например, раночная цена акции или же обменный курс валют). Практика показывает, что S_n ведёт себя весьма нерегулярно. Это привело Луи Башелье к идее использования аппарата теории вероятностей для изучения эмпирических феноменов, характеризующихся статистической неопределённостью, но при этом обладающих статистической устойчивостью частот.

Как обычно, будем считать, что все наблюдения проводятся на некотором вероятностном прострнастве $(\Omega, \mathcal{F}, \mathsf{P})$. Однако время и динамика являются неотъемлимыми частями финансовой теории, в связи с чем целесообразно специфицировать вероятностное пространство, добавив фильтрацию $(\mathcal{F}_n)_{n\geqslant 0}$. Интуитивно можно понимать, что $\mathcal{F}_n$ есть доступная наблюдателю "информация" о рынке вплоть до момента времени n. Получаемая четвёрка $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geqslant 0}, \mathsf{P})$ называется *фильтрованным вероятностным пространством* или же *стохастическим базисом*. Впрочем, во многих случаях целесообразно вводить не одну вероятностную меру, а целое семейство $\mathcal{P} = \{\mathsf{P}\}$ (это связано с тем, что бывает трудно выбрать какую-то конкретную меру P). Полученный набор объектов $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geqslant 0}, \mathcal{P})$ можно назвать *фильтрованным статистическим экспериментом*. Разумно предположить, что S_n полностью задаётся тем, что произошло до момента времени n включительно. Формально говоря, S_n является $\mathcal{F}_n$-измеримой случайной величиной.[10]

Теперь предположим, что для всех $n \geqslant 0$ $S_n > 0$. Как можно охарактеризовать S_n? Есть два способа.

Первый метод похож на метод сложных процентов (то есть проценты выплачиваются непрерывно). В нём $S_n = S_0 e^{H_n}$, где $H_n = h_0 + h_1 + \ldots + h_n$, $h_0 = 0$, а h_n есть $\mathcal{F}_n$-измеримая случайная величина. Несложно понять, что

$$H_n = \ln\frac{S_n}{S_0}, \quad h_n = \ln\frac{S_n}{S_{n-1}} = \ln\left(1 + \frac{\Delta S_n}{S_{n-1}}\right), \text{ где } \Delta S_n = S_n - S_{n-1}.$$

[10] То есть для любого $B \in \mathcal{B}(\mathbb{R})$ $S_n^{-1}(B) \in \mathcal{F}_n$.

Теперь введём следующие обозначения:

$$\widehat{h}_n = \frac{\Delta S_n}{S_{n-1}}, \quad \widehat{H}_n = \sum_{k=1}^{n} \widehat{h}_k.$$

Тогда

$$S_n = S_0 \prod_{k=1}^{n} e^{h_k} = S_0 \prod_{k=1}^{n} (1 + \widehat{h}_k) = S_0 \prod_{k=1}^{n} (1 + \Delta \widehat{H}_k) = S_0 e^{\widehat{H}_n} \prod_{k=1}^{n} (1 + \Delta \widehat{H}_k) e^{-\Delta \widehat{H}_k}.$$

Полученное разложение задаёт второй метод, который похож на метод простых процентов.

Теперь введём обозначение

$$\mathcal{E}(\widehat{H})_n \equiv e^{\widehat{H}_n} \prod_{k=1}^{n} (1 + \Delta \widehat{H}_k) e^{-\Delta \widehat{H}_k}.$$

Получаемая случайная последовательность $\mathcal{E}(\widehat{H}) \equiv (\mathcal{E}(\widehat{H})_n)_{n \geqslant 0}$ называется *экспонентой Долеан* или *стохастической экспонентой*, порождённой случайной последовательностью $\widehat{H} = (\widehat{H}_n)_{n \geqslant 0}$. Несложно понять, что в первом методе используется обычная экспонента: $S_n = S_0 e^{H_n}$, а во втором — стохастическая: $S_n = S_0 \mathcal{E}(\widehat{H})_n$.

Теперь покажем взаимосвязь между H_n и $\widehat{H}_n$. Заметим, что

$$\widehat{H}_n = \sum_{k=1}^{n} \widehat{h}_k = \sum_{k=1}^{n} (e^{h_k} - 1) = \sum_{k=1}^{n} (e^{\Delta H_k} - 1).$$

Тогда

$$\widehat{H}_n = H_n + (\widehat{H}_n - H_n) = H_n + \sum_{k=1}^{n} (e^{\Delta H_k} - 1 - \Delta H_k).$$

Также понятно, что

$$H_n = \sum_{k=1}^{n} h_k = \sum_{k=1}^{n} \ln(1 + \widehat{h}_k) = \sum_{k=1}^{n} \ln(1 + \Delta \widehat{H}_k).$$

Теперь покажем, что для $\mathcal{E}(\widehat{H})$ верно следующее разностное уравнение, от которого и пошло название:

$$\Delta \mathcal{E}(\widehat{H})_n = \mathcal{E}(\widehat{H})_{n-1} \Delta \widehat{H}_n, \quad \mathcal{E}(\widehat{H})_0 = 1.$$

Действительно,

$$\begin{aligned}
\Delta \mathcal{E}(\widehat{H})_n &= \mathcal{E}(\widehat{H})_n - \mathcal{E}(\widehat{H})_{n-1} = e^{\widehat{H}_n} \prod_{k=1}^{n} (1 + \Delta \widehat{H}_k) e^{-\Delta \widehat{H}_k} - e^{\widehat{H}_{n-1}} \prod_{k=1}^{n-1} (1 + \Delta \widehat{H}_k) e^{-\Delta \widehat{H}_k} = \\
&= (e^{\Delta \widehat{H}_n} (1 + \Delta \widehat{H}_n) e^{-\Delta \widehat{H}_n} - 1) e^{\widehat{H}_{n-1}} \prod_{k=1}^{n-1} (1 + \Delta \widehat{H}_k) e^{-\Delta \widehat{H}_k} = \\
&= \Delta \widehat{H}_n e^{\widehat{H}_{n-1}} \prod_{k=1}^{n-1} (1 + \Delta \widehat{H}_k) e^{-\Delta \widehat{H}_k} = \mathcal{E}(\widehat{H})_{n-1} \Delta \widehat{H}_n.
\end{aligned}$$

Примечание. Несложно заметить, что при малых h_k $\widehat{h}_k \approx h_k$, причём

$$\widehat{h}_k - h_k = \frac{1}{2} h_k^2 + \frac{1}{6} h_k^3 + \ldots$$

Впрочем, на данный момент ограничимся описанием распределения $S = (S_n)_{n\geqslant 0}$ и $H = (H_n)_{n\geqslant 0}$. С точки зрения классической теории вероятностей и продвинутой "статистики нормального распределения" было бы хорошо, если бы последовательность $H = (H_n)_{n\geqslant 0}$ была гауссовской. Если $H_n = h_1 + \ldots + h_n$, $n \geqslant 1$, то распределение H_n полностью бы задавалось распределением последовательности $h = (h_n)_{n\geqslant 1}$. Однако она полностью задаётся двумя параметрами: матожиданием $\mu_n = \mathsf{E}[h_n]$ и ковариациями $\mathsf{cov}(h_m, h_n)$.

Предположение о нормальности существенно упрощает решение многих задач, связанных с свойствами распределений. Например, теорема о нормальной корреляции даёт формулу для вычисления условного математического ожидания $\tilde{h}_{n+1} = \mathsf{E}[h_{n+1} \mid h_1, \ldots, h_n]$:

$$\tilde{h}_{n+1} = \mu_{n+1} + \sum_{k=1}^{n} a_k (h_k - \mu_k),$$

где a_k — коэффициенты, задаваемые матрицей ковариаций. Оказывается, что $\tilde{h}_{n+1}$ является *оптимальной* в среднеквадратичном смысле оценкой h_{n+1} по $h_1, \ldots, h_n$, то есть матожидание квадрата отклонения минимально.

Что будет в случае, если все h_k независимы? В таком случае матрица ковариаций диагональна и

$$\tilde{h}_{n+1} = \mathsf{E}[h_{n+1}] + \sum_{k=1}^{n} \frac{\mathsf{cov}(h_{n+1}, h_k)}{\mathsf{D}[h_k]} (h_k - \mathsf{E}[h_k]).$$

Формула для ошибки оценивания будет иметь вид (проверьте!):

$$\Delta_{n+1} = \mathsf{E}\Big[(\tilde{h}_{n+1} - h_{n+1})^2\Big] = \mathsf{D}[h_{n+1}] - \sum_{k=1}^{n} \frac{\mathsf{cov}^2(h_{n+1}, h_i)}{\mathsf{D}[h_i]}$$

Теперь вспомним один факт, связанный с нормальным распределением: с вероятностью около 90% значение случайной величины $\xi \sim \mathcal{N}(\mu, \sigma^2)$ будет лежать в интервале $[\mu - 1{,}65\sigma, \mu + 1{,}65\sigma]$. Тогда, пользуясь тем, что $h_{n+1} - \tilde{h}_{n+1} \sim \mathcal{N}(0, \Delta_{n+1})$, получаем, что

$$\mathsf{P}\Big(|h_{n+1} - \tilde{h}_{n+1}| \leqslant 1{,}65\sqrt{\Delta_{n+1}}\Big) \approx 0{,}90.$$

Отсюда получаем, что в 90% случаев прогнозируемое значение $\tilde{S}_{n+1}$ величины рыночной цены (по наблюдениям $h_1, \ldots, h_n$) лежит в интервале

$$[S_n e^{\tilde{h}_{n+1} - 1{,}65\sqrt{\Delta_{n+1}}}, S_n e^{\tilde{h}_{n+1} + 1{,}65\sqrt{\Delta_{n+1}}}].$$

Впрочем, к гипотезе нормальности нужно относиться с осторожностью. Практика показывает, что

- Число выборочных значений, не попадающих в интервал $[\overline{h}_n - k\widehat{\sigma}_n, \overline{h}_n + k\widehat{\sigma}_n]$, $k = 1, 2, 3$, где

$$\overline{h}_n = \frac{1}{n}\sum_{k=1}^{n} h_k \text{ — выборочное среднее,}$$

$$\widehat{\sigma}_n = \frac{1}{n-1}\sum_{k=1}^{n} (h_k - \overline{h}_n)^2 \text{ — выборочное стандартное отклонение,}$$

значительно больше, чем это должно быть при гипотезе нормальности. Это означает, что "хвосты" эмпирических распределений убывают значительно медленнее, чем у гауссовского распределения (тяжёлые хвосты).

- Может оказаться так, что *эксцесс*, или коэффициент вытянутости:

$$\widehat{k}_n = \frac{\widehat{m}_4}{\widehat{m}_2^2} - 3,$$

где $\widehat{m}_k$ есть выборочный k-й момент, получается положительным (хотя для нормального распределения он должен быть нулевым). Это означает сильную вытянутость пика плотности распределения в окрестности центральных значений.

Пожалуй, самым сильным предположением (относительно структуры распределения величин $h = (h_n)$) является, помимо нормальности, предположение *независимости и одинаковой распределённости* этих величин. В таком случае анализ цен легко проводится с помощью обычных методов теории вероятностей. Однако при таком предположении сразу же рушится надежда на то, что прошлые данные хоть как-то влияют на будущие.

Предположим, что в модели

$$S_n = S_0 e^{H_n}, \quad H_n = h_1 + \ldots + h_n,$$

случайные величины h_n имеют конечные абсолютные первые моменты: $\mathsf{E}[|h_n|] < +\infty$.

Разложение Дуба, о котором пойдёт речь дальше, предполагает изучение последовательности $H = (H_n)$ в зависимости от свойств фильтрации $(F_n)_{n\geqslant 0}$, то есть потока информаций, доступных наблюдателю. Положим $\mathcal{F}_0 = \{\varnothing, \Omega\}$.

Так как $\mathsf{E}[|h_n|] < +\infty$, $n \geqslant 1$, то определены условные математические ожидания $\mathsf{E}[h_n \mid \mathcal{F}_{n-1}]$. Тогда

$$H_n = \sum_{k=1}^{n} h_k = \sum_{k=1}^{n} \mathsf{E}[h_k \mid \mathcal{F}_{k-1}] + \sum_{k=1}^{n} (h_k - \mathsf{E}[h_k \mid \mathcal{F}_{k-1}]).$$

Если ввести обозначения

$$A_n = \sum_{k=1}^{n} \mathsf{E}[h_k \mid \mathcal{F}_{k-1}],$$
$$M_n = \sum_{k=1}^{n} (h_k - \mathsf{E}[h_k \mid \mathcal{F}_{k-1}]),$$

то для $H = (H_n)_{n\geqslant 0}, H_0 = 0$ справедливо разложение Дуба

$$H_n = A_n + M_n, \quad n \geqslant 0,$$

где

- $A = (A_n)_{n\geqslant 0}, A_0 = 0$ является *предсказуемой* случайной последовательностью. Другими словами, для любого $n \geqslant 1$ A_n есть $\mathcal{F}_{n-1}$-измеримой случайная величина.
- $M = (M_n)_{n\geqslant 0}, M_0 = 0$ является *мартингалом*, то есть для любого $n \geqslant 1$ $\mathsf{E}[M_n \mid \mathcal{F}_{n-1}] = M_{n-1}$, причём M_n есть $\mathcal{F}_n$-измеримые величины и $\mathsf{E}[|M_n|] < \infty$.

Примечание. Предположим, что наряду с фильтрацией $(\mathcal{F}_n)$ задана *подфильтрация* $(\mathcal{G}_n)$, где $\mathcal{G}_n \subseteq \mathcal{F}_n$ и $\mathcal{G}_n \subseteq \mathcal{G}_{n+1}$. Аналогичным образом можно написать разложение $H = (H_n)$ относительно потока $(\mathcal{G}_n)$:

$$H_n = \sum_{k=1}^{n} \mathsf{E}[h_k \mid \mathcal{G}_{k-1}] + \sum_{k=1}^{n} (h_k - \mathsf{E}[h_k \mid \mathcal{G}_{k-1}]).$$

Последовательность $A = (A_n)$ с элементами

$$A_n = \sum_{k=1}^{n} \mathsf{E}[h_k \mid \mathcal{G}_{k-1}]$$

будет $(\mathcal{G}_n)$-предсказуемой (то есть A_n являются $\mathcal{G}_{n-1}$-измеримыми). Однако $M = (M_n)$, задаваемая по правилу

$$M_n = \sum_{k=1}^{n} (h_k - \mathsf{E}[h_k \mid \mathcal{G}_{k-1}]).$$

не является мартингалом относительно фильтрации $(\mathcal{G}_n)$, так как h_k являются измеримыми относительно сигма-алгебры $\mathcal{F}_k$, что не означает $\mathcal{G}_k$-измеримость.

У этого разложения есть хорошее свойство: оно единственно. Действительно, пусть $H_n = A'_n + M'_n$ — другое разложение с $(\mathcal{F}_n)$-предсказуемой последовательностью $A' = (A'_n), A'_0 = 0$ и мартингалом $M' = (M'_n, \mathcal{F}_n)$. Заметим, что

$$A'_{n+1} - A'_n = H_{n+1} - H_n - (M'_{n+1} - M'_n) = (A_{n+1} - A_n) + (M_{n+1} - M_n) - (M'_{n+1} - M'_n).$$

Теперь возьмём от обеих частей условное матожидание $\mathsf{E}[\cdot \mid \mathcal{F}_n]$. Тогда мы получим, что

$$\mathsf{E}\big[A'_{n+1} - A'_n \;\big|\; \mathcal{F}_n\big] = \mathsf{E}[A_{n+1} - A_n \mid \mathcal{F}_n].$$

Так как A_{n+1} и A_n являются $\mathcal{F}_n$-измеримыми, то $A'_{n+1} - A'_n = A_{n+1} - A_n$. Пользуясь тем, что $A_0 = A'_0 = 0$, получаем желаемое.

Стоит заметить, что если в рассматриваемой модели $\mathsf{E}[h_k \mid \mathcal{F}_{k-1}] = 0$ для всех $k \geqslant 1$, то сама последовательность $H = (H_n)$ будет являться мартингалом.

Разложение Дуба не такое тривиальное, каким оно может показаться на первый взгляд.

Пример 22. Рассмотрим последовательность iid случайных величин $\{\xi_n\}_{n\in\mathbb{N}}$ таких, что

$$\mathsf{P}(\xi_n = 1) = \mathsf{P}(\xi_n = -1) = \frac{1}{2}.$$

Далее, пусть $X_n = \xi_1 + \ldots + \xi_n$. Другими словами, пусть есть простейшее случайное блуждание. Разложение Дуба для $H_n = |X_n|$, $n \geqslant 0$, $|X_0| = 0$ будет устроено следующим образом:

$$h_n = \Delta H_n = |X_n| - |X_{n-1}| = |X_{n-1} + \xi_n| - |X_{n-1}|.$$

Далее, пользуясь свойствами условного математического ожидания и независимостью ξ_n и X_{n-1}, получаем, что

$$\begin{aligned}\Delta M_n = h_n - \mathsf{E}[h_n \mid \mathcal{F}_{n-1}] = |X_{n-1} + \xi_n| - |X_{n-1}| - \mathsf{E}[|X_{n-1} + \xi_n| - |X_{n-1}| \mid \mathcal{F}_{n-1}] = \\ = |X_{n-1} + \xi_n| - \mathsf{E}[|X_{n-1} + \xi_n| \mid X_{n-1}] = (\operatorname{sgn} X_{n-1})\xi_n.\end{aligned}$$

Отсюда получаем, что мартингал в разложении Дуба имеет вид

$$M_n = \sum_{k=1}^{n} (\operatorname{sgn} X_{k-1})\xi_k = \sum_{k=1}^{n} (\operatorname{sgn} X_{k-1})\Delta X_k.$$

Далее,

$$\mathsf{E}[h_n \mid \mathcal{F}_{n-1}] = \mathsf{E}[|X_{n-1} + \xi_n| - |X_{n-1}| \mid \mathcal{F}_{n-1}] = \mathsf{E}[|X_{n-1} + \xi_n| \mid \mathcal{F}_{n-1}] - |X_{n-1}|.$$

Несложно заметить, что если $X_{n-1} \neq 0$, то это условное матожидание обращается в ноль. Если же $X_{n-1} = 0$, то оно равно единице. Тогда

$$\sum_{k=1}^{n} \mathsf{E}[h_k \mid \mathcal{F}_{k-1}] = \#\{1 \leqslant k \leqslant n : X_{k-1} = 0\}.$$

Пусть $L_n(0) = \#\{0 \leqslant k \leqslant n-1 : X_k = 0\}$ — число нулей последовательности $(X_k)_{0 \leqslant k \leqslant n-1}$. Тогда по разложению Дуба

$$|X_n| = \sum_{k=1}^{n} (\operatorname{sgn} X_{k-1}) \Delta X_k + L_n(0).$$

Теперь воспользуемся тем, что у мартингала матожидание постоянно и равно нулю. Тогда

$$\mathsf{E}[L_n(0)] = \mathsf{E}[|X_n|].$$

Согласно ЦПТ $X_n/\sqrt{n} \sim \mathcal{N}(0,1)$. Следовательно,

$$\mathsf{E}[|X_n|] \sim \sqrt{\frac{2n}{\pi}} \implies \mathsf{E}[L_n(0)] \sim \sqrt{\frac{2n}{\pi}}.$$

Полученная формула — это известный результат о среднем числе нулей в симметричном случайном блуждании Бернулли.

1.12 Гауссовские и условно-гауссовские модели

На абсолютно *эффективных рынках* наилучшим прогнозом будущего уровня цены финансового актива является текущая цена этого актива. Поэтому понятие мартингала стало одним из основных при исследовании динамики эволюции цен как стохастических последовательностей или процессов с определёнными свойствами их распределений. Однако при проведении конкретных расчетов одного лишь знания "мартингальности распределений" слишком мало — нужна более "тонкая" структура этих распределений, что приводит к необходимости детального рассмотрения самых разнообразных вероятностно-статистических моделей с целью выявления тех из них, свойства распределений которых лучше всего согласуются со свойствами эмпирических распределений, построенных по статистическим данным.

Предположение гауссовости распределений величин $h_1, \ldots, h_n$, конечно, выглядит привлекательным и с точки зрения теоретического анализа, и с точки зрения "статистики нормального распределения". Но это предположение не всегда соответсвтует истинной картине поведения цен. Но какую альтернативу можно привести? Для этого вспомним разложение Дуба. Как известно, оно определяется с привлечением условных матожиданий вида $\mathsf{E}[h_n \mid \mathcal{F}_{n-1}]$. Тогда было бы разумно предположить, что не безусловные, а *условные* распределения являются гауссовскими:

$$\operatorname{Law}(h_n \mid \mathcal{F}_{n-1}) = \mathcal{N}(\mu_n, \sigma_n^2)$$

с некоторыми $\mathcal{F}_{n-1}$-измеримыми величинами $\mu_n = \mu_n(\omega)$ и $\sigma_n^2 = \sigma_n^2(\omega)$.

Оказывается, что $\mathsf{E}[h_n \mid \mathcal{F}_{n-1}] = \mu_n$ и $\mathsf{D}[h_n \mid \mathcal{F}_{n-1}] = \sigma_n^2$ (это следует из регулярности условного распределения — за доказательством обращайтесь к первому тому Ширяева). Тем самым видем смысл этих параметров — условное среднее и условная дисперсия распределения $\operatorname{Law}(h_n \mid \mathcal{F}_{n-1})$.

Само же распределение $\mathrm{Law}(h_n)$ является *взвесью* условных гауссовских распределений $\mathrm{Law}(h_n \mid \mathcal{F}_{n-1})$ с усреднением по распределению величин μ_n и σ_n^2.

Обычно наряду с $h = (h_n)$ вводится "стандартная" условно-гауссовская последовательность $\varepsilon = (\varepsilon_n)_{n\geqslant 1}$ $\mathcal{F}_n$-измеримых случайных величин таких, что

$$\mathrm{Law}(\varepsilon_n \mid \mathcal{F}_{n-1}) = \mathcal{N}(0,1), \text{ где } \mathcal{F}_0 = \{\varnothing, \Omega\}.$$

Оказывается, что это будет последовательность *независимых* случайных величин с стандартным нормальным распределением, так как

$$\mathrm{Law}(\varepsilon_n \mid \varepsilon_1, \ldots, \varepsilon_{n-1}) = \mathcal{N}(0,1).$$

Если $\omega_n^2 \neq 0$ поточечно для всех $n \geqslant 1$, то величины ε_n, задаваемые по правилу $\varepsilon_n \equiv (h_n - \mu_n)/\omega_n$, будут задавать стандартную гауссовскую последовательность. Тогда можно считать, что рассматриваемые условно-гауссовские последовательности представимы в виде

$$h_n = \mu_n + \sigma_n \varepsilon_n,$$

где $\varepsilon = (\varepsilon_n)$ — последовательность независимых $\mathcal{F}_n$-измеримых случайных величин с распределением $\mathcal{N}(0,1)$.

Понятно, что более подробное изучение свойств последовательности $h = (h_n)$ зависит от структуры μ_n и σ_n^2. Именно это и делается в представляемых ниже моделях.

В теории временных рядов есть целый арсенал разнообразных *линейных* моделей, среди которых в первую очередь нужно назвать следующие:

- Модель *скользящего среднего* порядка q $\mathrm{MA}(q)$,
- Модель *авторегрессии* порядка p $\mathrm{AR}(p)$,
- Модель *авторегрессии и скользящего среднего* порядка (p,q) $\mathrm{ARMA}(p,q)$.

Эти модели широко исследуются в теории временных рядов, особенно в предположении *стационарности*. Вообще, для чего вводятся линейные модели? Они весьма просты, но при этом ими можно неплохо приближать весьма широкий класс стационарных последовательностей.

Вот только не все временные "эконометрические" ряды являются стационарными. Анализ показывает, что часто в данных вырисовываются три составляющие:

- Медленно меняющийся (например, "инфляционный") тренд x,
- Периодические или же апериодические циклы y,
- Нерегулярная, флуктуирующая ("стохастическая" или "хаотическая") компонента z.

В наблюдаемые данные h они могут входить весьма разнообразными способами. Образно это можно представить так: $h = x * y * z$, где вместо $*$ могут выступать сложение $+$, умножение $\times$ и так далее.

Ниже мы рассмотрим некоторые *линейные* (а затем и нелинейные) модели, преследуя цель дать представление об их структуре, особенностях, свойствах, применяемых в анализе данных.

Не стоит забывать, что конечной целью анализа статистических данных является *прогнозирование* дальнейшнего поведения. Качество этого прогнозирования зависит от удачного выбора модели, точности оценивания определяющих её параметров и качества экстраполяционного оценивания.

Во всех рассматриваемых далее моделях будем считать, что задана некоторая "базисная" последовательность $\varepsilon = (\varepsilon_n)$, которую в теории временных рядов обычно считают *белым шумом* и идентифицируют с источником случайности, определяющим стохастический характер исследуемых вероятностно-статистических объектов. При этом (в "L^2-теории") говорят, что последовательность является *белым шумом в широком смысле.*

Определение 56. Последовательность $\varepsilon = (\varepsilon_n)$ называется белым шумом в широком смысле, если

$$\forall m, n \in \mathbb{Z}, m \neq n : \mathsf{E}[\varepsilon_n] = 0, \mathsf{E}\left[\varepsilon_n^2\right] < \infty, \mathsf{E}[\varepsilon_m \varepsilon_n] = 0.$$

Другими словами, белый шум в широком смысле — это квадратично интегрируемая последовательность некоррелированных случайных величин с нулевыми средними.

Ещё вводят белый шум *в узком смысле*, который обычно называют просто *белым (гауссовским) шумом.*

Определение 57. Белый шум — это гауссовская последовательность, являющаяся белым шумом в широком смысле.

Другими словами, это последовательность независимых случайных величин с нормальными распределениями $\mathcal{N}(0, \sigma_n^2)$. Далее будем считать, что $\sigma_n^2 \equiv 1$. В таком случае обычно говорят, что ε есть стандартная гауссовская последовательность.

1.12.1 Модель скользящего среднего MA(q)

В модели скользящего среднего порядка q, описывающей эволюцию последовательности $h = (h_n)$, предполагается следующие способ формирования значений h_n по белому шуму в широком смысле $\varepsilon = (\varepsilon_n)$:

$$h_n = (\mu + b_1\varepsilon_{n-1} + \ldots + b_q\varepsilon_{n-q}) + b_0\varepsilon_n,$$

где параметр q определяет порядок зависимости от "прошлого", а ε_n играет роль величин, "обновляющих" информацию, содержащуюся в $\mathcal{F}_{n-1} = \sigma(\varepsilon_{n-1}, \varepsilon_{n-2}, \ldots)$.

Далее, для компактности вводят *лаговый оператор* L, действующий по правилу $Lx_n = x_{n-1}$. Так как $L(Lx_n) = x_{n-2}$, то разумно ввести обозначение

$$L^2 x_n \equiv L(Lx_n) = x_{n-2},$$

и, в общем случае, $L^k x_n = x_{n-k}$.

Отметим следующие свойства лагового оператора:

$$\begin{aligned} L(cx_n) &= cLx_n, \\ L(x_n + y_n) &= Lx_n + Ly_n, \\ (c_1L + c_2L^2)x_n &= c_1Lx_n + c_2L^2x_n = c_1x_{n-1} + c_2x_{n-2}, \\ (1 - \lambda_1 L)(1 - \lambda_2 L)x_n &= x_n - (\lambda_1 + \lambda_2)x_{n-1} + (\lambda_1\lambda_2)x_{n-2}. \end{aligned}$$

Пользуясь этими свойствами, модели MA(q) можно придать следующую форму: $h_n = \mu + \beta(L)\varepsilon_n$, где $\beta(L) = b_0 + b_1L + \ldots + b_qL^q$.

Теперь положим $q = 1$. В таком случае

$$h_n = \mu + b_0\varepsilon_n + b_1\varepsilon_{n-1}.$$

Несложно проверить, что

$$\mathsf{E}[h_n] = \mu, \mathsf{D}[h_n] = b_0^2 + b_1^2$$
$$\mathsf{cov}(h_{n+1}, h_n) = b_0 b_1, \mathsf{cov}(h_{n+k}, h_n) = 0, k > 1.$$

Это означает, что $h = (h_n)$ — это последовательность с коррелированными соседними значениями (h_n и h_{n+1}), причём корреляция значений h_{n+k} и h_n при $k > 1$ равна нулю.

Из соотношений сверху следует, что у элементов последовательности $h = (h_n)$ матожидание, дисперсия и ковариация не зависят от n (впрочем, это определяется предположением стандартности последовательности ε и тем, что b_k не зависят от n). Отсюда следует, что последовательность $h = (h_n)$ является стационарной в широком смысле (просто по определению). Если же добавить то, что ε является гауссовской, то и h тоже будет гауссовской. Это означает, что все её параметры полностью задаются средним, дисперсией и ковариацией. Но тогда h будет стационарной и в узком смысле, так как для произвольных n, k и $i_1, \ldots, i_n$

$$\mathrm{Law}(h_{i_1}, \ldots, h_{i_n}) = \mathrm{Law}(h_{i_1+k}, \ldots, h_{i_n+k})$$

Теперь покажем одно интересное свойство модели MA(1). Пусть $(h_1, \ldots, h_n)$ — некоторая реализация, полученная в результате наблюдений величин h_k в моменты времени $k = 1, \ldots, n$. Далее, пусть $\overline{h}_n = (\sum_{k=1}^n h_k)/n$ — это временное среднее. Со статистической точки зрения обращение к "статистике" $\overline{h}_n$ представляет тот интерес, что $\overline{h}_n$ является естественным кандидатом для оценивания среднего μ.

Оказывается, что для стационарной в широком смысле последовательности h_n есть хороший критерий эргодичности, который похож на условие Слуцкого (на самом деле это оно и есть):

Теорема 21. *Пусть $h = (h_n)$ — стационарная в широком смысле последовательность, а $R(k) = \mathsf{cov}(h_{n+k}, h_n)$. Тогда*

$$\lim_{n\to\infty} \mathsf{E}\big[(\overline{h}_n - \mu)^2\big] = 0 \iff \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n R(k) = 0.$$

Доказательство. Без ограничения общности скажем, что $\mu = \mathsf{E}[h_n] = 0$.

Пусть $\mathsf{E}\big[(\overline{h}_n - \mu)^2\big] \to 0$. Тогда по неравенству Коши-Буняковского:

$$\left|\frac{1}{n}\sum_{k=1}^n R(k)\right|^2 = \left|\mathsf{E}\left[\frac{h_0}{n}\sum_{k=1}^n h_k\right]\right|^2 \leqslant \mathsf{E}\big[h_0^2\big]\,\mathsf{E}\left[\left|\frac{1}{n}\sum_{k=1}^n h_k\right|^2\right] \xrightarrow[n\to\infty]{} 0.$$

Теперь докажем в другую сторону. Заметим, что

$$\mathsf{E}\left[\left|\frac{1}{n}\sum_{k=1}^n h_k\right|^2\right] = \frac{1}{n^2}\,\mathsf{E}\left[\sum_{k=1}^n h_k^2 + 2\sum_{i=1}^n\sum_{j=1}^{i-1} h_i h_j\right] = \frac{2}{n^2}\sum_{i=1}^n\sum_{j=0}^{i-1} R(j) - \frac{1}{n}R(0)$$

Зафиксируем произвольное $\delta > 0$ и найдём $n(\delta)$ такое, что для любого $l \geqslant n(\delta)$ выполнено, что

$$\left|\frac{1}{l}\sum_{k=0}^l h_k\right| \leqslant \delta.$$

Тогда для $n \geqslant n(\delta)$ имеем

$$\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=0}^{i-1}R(j)\right| = \left|\frac{1}{n^2}\sum_{i=1}^{n(\delta)}\sum_{j=0}^{i-1}R(j) + \frac{1}{n^2}\sum_{i=n(\delta)+1}^{n}\sum_{j=0}^{i-1}R(j)\right| \leqslant$$
$$\leqslant \left|\frac{1}{n^2}\sum_{i=1}^{n(\delta)}\sum_{j=0}^{i-1}R(j)\right| + \left|\frac{1}{n^2}\sum_{i=n(\delta)+1}^{n} i\cdot\frac{1}{i}\sum_{j=0}^{i-1}R(j)\right| \leqslant$$
$$\leqslant \frac{1}{n^2}\left|\sum_{i=1}^{n(\delta)}\sum_{j=0}^{i-1}R(j)\right| + \delta$$

Теперь вспомним, что $R(0) = const < \infty$. Тогда

$$\overline{\lim_{n\to\infty}}\,\mathsf{E}\left[\left|\frac{1}{n}\sum_{k=1}^{n}h_k\right|^2\right] \leqslant \delta.$$

Устремляя δ к нулю, получаем желаемое. □

Тем самым мы получили весьма полезное свойство: MA(1) эргодична в среднеквадратичном, то есть среднее по времени стремится в среднеквадратичном смысле к среднему по ансамблю μ.

Теперь вспомним про корреляционную функцию. Для MA(1) она будет иметь вид

$$r(k) = \frac{\mathsf{cov}(h_{n+k}, h_n)}{\sqrt{\mathsf{D}[h_{n+k}]\,\mathsf{D}[h_n]}} = \frac{R(k)}{R(0)} = \begin{cases} 1, & k = 0 \\ \frac{b_0 b_1}{b_0^2 + b_1^2}, & k = 1 \\ 0, & k > 1 \end{cases}$$

Вернёмся к общему случаю MA(q). В качестве упражнения оставим вывод следующих тождеств:

$$\mathsf{E}[h_n] = \mu, \quad \mathsf{D}[h_n] = \sum_{k=0}^{q} b_k^2,$$
$$R(k) = \begin{cases} \sum_{i=0}^{q-k} b_i b_{k+i}, & k \leqslant q \\ 0, & k > q \end{cases}$$

Пользуясь этими формулами, моделью MA(q) можно пытаться моделировать поведение последовательностей $h = (h_n)$, у которых корреляция величин h_n и h_{n+k}, где $k > q$, нулевая. Но как это делать? Общий принцип подгонки следующий:

- Для начала, по выборке $(h_1, \ldots, h_n)$ строятся некоторые эмпирические характеристики: например,

$$\overline{h}_n = \frac{1}{n}\sum_{k=1}^{n} h_k \text{ — выборочное среднее}$$
$$\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(h_k - \overline{h}_n) \text{ — выборочная дисперсия}$$
$$r_n(k) = \frac{1}{n\widehat{\sigma}_n^2}\sum_{i=k+1}^{n}(h_i - \overline{h}_n)(h_{i-k} - \overline{h}_n) \text{ — выборочное среднее}$$

- Далее, используя выражения для теоретических характеристик, производится варьирование параметров с целью подгонки теоретических значений под эмпирические.

- В конце провдодится *оценка кажества* подгонки, основываясь на знании эмпирических характеристик и их отклонений от теоретических распределений.

Ранее мы смотрели на модели с конечным q. Было бы разумно ввести обобщение, которое позволяет и бесконечное q: MA(∞). Устроено оно так:

$$h_n = \mu + \sum_{j=0}^{\infty} b_j \varepsilon_{n-j}.$$

Но, конечно, должны быть какие-то условия (как минимум, на сходимость). Если потребовать сходимость ряда $\sum b_j^2$, то ряд в формуле для h_n будет сходиться в среднеквадратичном смысле.

Для этой модели

$$\mathsf{E}[h_n] = \mu, \quad \mathsf{D}[h_n] = \sum_{k=0}^{\infty} b_k^2, \quad R(k) = \sum_{i=0}^{\infty} b_{k+i} b_i.$$

В теории стационарных случайных процессов принято говорить, что h_n есть "результат реакции физически осуществимого фильтра с импульсной переходной функцией $b = (b_n)$, когда на вход подается последовательность $\varepsilon = (\varepsilon_n)$".

Оказывается, что в определённом смысле "регулярная" стационарная (в широком смысле) последовательность может быть представлена моделью MA(∞). Результат, связанный с этим фактом, называется *разложением Вольда* стационарных последовательностей на «сингулярную» и «регулярную» части. За подробностями обращайтесь ко второму тому Ширяева.

1.12.2 Авторегрессионная модель AR(p)

Поехали дальше. Авторегрессионая модель порядка p определяется следующим образом:

$$h_n = \mu_n + \sigma\varepsilon_n, \quad \mu_n = a_0 + a_1 h_{n-1} + \ldots + a_p h_{n-p}.$$

Можно сказать, что модель AR(p) подчиняется *разностному уравнению* порядка p:

$$h_n = a_0 + a_1 h_{n-1} + \ldots + a_p h_{n-p} + \sigma\varepsilon_n.$$

Если совпользоваться лаговым оператором, то это уравнение можно записать в виде:

$$(1 - a_1 L - \ldots - a_p L^p) h_n = a_0 + \sigma\varepsilon_n.$$

Тогда, введя пару обознаений, уравнение приобретает вид

$$\alpha(L) h_n = \omega_n, \quad \omega_n = a_0 + \sigma\varepsilon_n, \quad \alpha(L) = 1 - a_1 L - \ldots - a_p L^p$$

В отличие от модели со скользящий средним, в этой модели нужно задавать *начальные* условия $(h_{1-p}, h_{2-p}, \ldots, h_0)$. Обычно их обнуляют, хотя можно считать их случайными и не зависящими от ε. В эргодических случаях асимптотическое поведение последовательности при $n \to \infty$ не зависит от начальных условий, и в этом смысле их конкретизация не столь существенна.

Рассмотрим модель AR(1) поподробнее. Выглядит она следующим образом:

$$h_n = a_0 + a_1 h_{n-1} + \sigma\varepsilon_n.$$

Этоа модель выделяется из общего класса AR(p) моделей тем, что из «прошлых» величин $h_{n-1}, \ldots, h_{n-p}$, h_n зависит только от ближайшего (по времени) значения h_{n-1}. Если добавить к этой модели независимость последовательности $\varepsilon = (\varepsilon_n)$ и независимость h_0 от неё, то получится *конструктивный* пример марковской цепи.

Несложно получить, что

$$h_n = a_0(1 + a_1 + a_1^2 + \ldots + a_1^{n-1}) + a_1^n h_0 + \sigma(\varepsilon_n + a_1\varepsilon_{n-1} + \ldots + a_1^{n-1}\varepsilon_1).$$

Отсюда видно, что свойства последовательности сильно зависят от a_1. Учитывая формулу, есть смысл различать три случая: $|a_1| < 1$, $|a_1| = 1$ и $|a_1| > 1$, причём второй случай является «пограничным». Смысл этого станет понятен позднее.

Из разложения понятно, что

$$\begin{aligned}
\mathsf{E}[h_n] &= a_0(1 + a_1 + \ldots + a_1^{n-1}) + a_1^n\,\mathsf{E}[h_0] = \frac{a_0(1-a_1^n)}{1-a_1} + a_1^n\,\mathsf{E}[h_0] \\
\mathsf{D}[h_n] &= a_1^{2n}\,\mathsf{D}[h_0] + \sigma^2(1 + a_1^2 + \ldots + a_1^{2(n-1)}) = a_1^{2n}\,\mathsf{D}[h_0] + \frac{\sigma^2(1-a_1^{2n})}{1-a_1^2} \\
\mathsf{cov}(h_n, h_{n-k}) &= a_1^{2n-k}\,\mathsf{D}[h_0] + \sigma^2 a_1^k(1 + a_2 + \ldots + a_1^{2(n-k-1)}) \\
&= a_1^{2n-k}\,\mathsf{D}[h_0] + \frac{\sigma^2 a_1^k(1-a_1^{2(n-k)})}{1-a_1^2}
\end{aligned}$$

Если $|a_1| < 1$ и $\mathsf{E}[h_0] < \infty]$, $\mathsf{D}[h_0] < \infty$, то при $n \to \infty$ последовательность «стационаризируется»:

$$\mathsf{E}[h_n] \to \frac{a_0}{1-a_1}, \quad \mathsf{D}[h_n] \to \frac{\sigma^2}{1-a_1^2}, \quad \mathsf{cov}(h_n, h_{n-k}) \to \frac{\sigma^2 a_1^k}{1-a_1^2}$$

Теперь заметим, что если начальное распределение для h_0 является нормальным:

$$h_0 \sim \mathcal{N}\left(\frac{a_0}{1-a_1}, \frac{\sigma^2}{1-a_1^2}\right),$$

то последовательность $h = (h_n)$ является стационарной в узком смысле гауссовской последовательностью. Заметим, что для такой последовательности корреляция равна

$$r(k) = \frac{\mathsf{cov}(h_{n-k}, h_n)}{\sqrt{\mathsf{D}[h_{n-k}]\,\mathsf{D}[h_n]}} = a_1^k.$$

Зафиксируем n. В таком случае

$$h_n = a_0(1 + a_1 + a_1^2 + \ldots + a_1^{n-1}) + a_1^n h_0 + \sigma(\varepsilon_n + a_1\varepsilon_{n-1} + \ldots + a_1^{n-1}\varepsilon_1).$$

Если h_0 есть константа, то, введя обозначение $\mu = a_0(1 + a_1 + a_1^2 + \ldots + a_1^{n-1}) + a_1^n h_0$, получится модель MA($n-1$). В этом смысле иногда несколько вольно говорят, что «модель AR(1) может рассматриваться как модель MA(∞)».

Теперь скажем, что $|a_1| = 1$. Тогда

$$h_n = a_0 n + h_0 + \sigma(\varepsilon_1 + \ldots + \varepsilon_n).$$

Если ввести обозначение $\omega_n = a_0 + \sigma\varepsilon_n$, то модель будет иметь вид

$$h_n = h_0 + \omega_1 + \omega_2 + \ldots + \omega_n.$$

Это есть ни что иное, как случайное блуждание. Заметим, что

$$\mathsf{E}[h_n] = a_0 n + \mathsf{E}[h_0], \quad \mathsf{D}[h_n] = \sigma^2 n \xrightarrow[n\to\infty]{} \infty.$$

Случай с $|a_1| > 1$ называют *взрывающимся*, так как и среднее значение, и дисперсия растут с ростом n, причём экспоненциально быстро.

Теперь посмотрим на модель AR(2):

$$h_n = a_0 + a_1 h_{n-1} + a_2 h_{n-2} + \sigma\varepsilon_n \iff (1 - a_1 L - a_2 L^2)h_n = a_0 + \sigma\varepsilon_n.$$

Если $a_2 = 0$, то мы возвращаемся к модели AR(1). Введя обозначение $\omega_n = a_0 + \sigma\varepsilon_n$, получаем, что

$$(1 - a_1 L)h_n = \omega_n.$$

Вопрос: можно ли как-то «обратить» это равенство и сразу считать h_n только по $\omega = (\omega_n)$, не обращаясь к предыдущим значениям? Воспользуемся свойствами оператора L и заметим, что

$$(1 + a_1 L + a_1^2 L^2 + \ldots + a_1^k L^k)(1 - a_1 L) = 1 - a_1^{k+1} L^{k+1}.$$

Тогда

$$h_n = (1 + a_1 L + a_1^2 L^2 + \ldots + a_1^k L^k)\omega_n + a_1^{k+1} L^{k+1} h_n.$$

Если положить $k = n - 1$, то получим разложение, которое получали ранее:

$$\begin{aligned} h_n &= (1 + a_1 L + a_1^2 L^2 + \ldots + a_1^{n-1} L^{n-1})\omega_n + a_1^n h_0 = \\ &= (a_0 + \sigma\varepsilon_n) + a_1(a_0 + \sigma\varepsilon_{n-1}) + \ldots + a_1^{n-1}(a_0 + \sigma\varepsilon_1) + a_1^n h_0 = \\ &= a_0(1 + a_1 + \ldots + a_1^{n-1}) + a_1^n h_0 + \sigma(\varepsilon_n + a_1\varepsilon_{n-1} + \ldots + a_1^{n-1}\varepsilon_1). \end{aligned}$$

Если $|a_1| < 1$ и n достаточно велико, то неформально можно сказать, что

$$h_n \approx (1 + a_1 L + a_1^2 L^2 + \ldots + a_1^{n-1} L^{n-1})\omega_n = (1 + a_1 L + a_1^2 L^2 + \ldots + a_1^{n-1} L^{n-1})(1 - a_1 L)h_n.$$

Тем самым мы получаем, что «обратный» оператор $(1 - a_1 L)^{-1}$ разумно определить следующим образом:

$$(1 - a_1 L)^{-1} \equiv \sum_{k=0}^{\infty} a_1^k L^k = 1 + a_1 L + a_1^2 L^2 + \ldots + a_1^n L^n + \ldots$$

Оказывается, что если получаемый ряд для h_n сходится в среднеквадратичном смысле, то это разложение единственно.

Пользуясь этим рассуждением, можно найти похожее представление для AR(2). Так как

$$(1 - \lambda_1 L)(1 - \lambda_2 L) = 1 - (\lambda_1 + \lambda_2)L + \lambda_1\lambda_2 L^2,$$

то, определяя λ_1 и λ_2 из системы

$$\begin{cases} \lambda_1 + \lambda_2 = a_1 \\ \lambda_1\lambda_2 = -a_2 \end{cases}$$

получим, что

$$1 - a_1 L - a_2 L^2 = (1 - \lambda_1 L)(1 - \lambda_2 L).$$

Тогда

$$(1 - \lambda_1 L)(1 - \lambda_2) h_n = \omega_n \implies h_n = (1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1}\omega_n$$

Воспользуемся методом неопределённых коэффициентов:

$$\frac{1}{(1 - \lambda_1 L)(1 - \lambda_2 L)} = \frac{A}{1 - \lambda_1 L} + \frac{B}{1 - \lambda_2 L} = \frac{A - A\lambda_2 L + B - B\lambda_1 L}{(1 - \lambda_1 L)(1 - \lambda_2 L)}$$

Тогда

$$\begin{cases} A + B = 1 \\ A\lambda_2 + B\lambda_1 = 0 \end{cases} \implies A = \frac{\lambda_1}{\lambda_1 - \lambda_2}, B = -\frac{\lambda_2}{\lambda_1 - \lambda_2}$$

Отсюда получаем, что

$$h_n = \frac{\lambda_1}{\lambda_1 - \lambda_2}(1 - \lambda_1 L)^{-1}\omega_n - \frac{\lambda_2}{\lambda_1 - \lambda_2}(1 - \lambda_2 L)^{-1}\omega_n.$$

Предположим, что все λ_i такие, что $|\lambda_i| < 1$. Тогда

$$h_n = \sum_{k=0}^{\infty}(c_1 \lambda_1^k + c_2 \lambda_2^k)\omega_k, \text{ где } c_1 = \frac{\lambda_1}{\lambda_1 - \lambda_2}, \quad c_2 = -\frac{\lambda_2}{\lambda_1 - \lambda_2}.$$

Это рассуждение обобщается и на модель AR(p). Для неё

$$(1 - a_1 L - a_2 L^2 - \ldots - a_p L^p) h_n = \omega_n.$$

Опять же, разложим его на множители:

$$1 - a_1 L - a_2 L^2 - \ldots - a_p L^p = (1 - \lambda_1 L) \ldots (1 - \lambda_p L)$$

Если все $|\lambda_i| < 1$, то получится стационарное решение, которое будет единственным среди решений с конечным вторым моментом:

$$h_n = (1 - \lambda_1 L)^{-1} \ldots (1 - \lambda_p L)^{-1}\omega_n.$$

Теперь сведём это в виду ряда. Для этого снова воспользуемся методом неопределённых коэффициентов:

$$\frac{1}{(1 - \lambda_1 z) \ldots (1 - \lambda_p z)} = \frac{c_1}{1 - \lambda_1 z} + \ldots + \frac{c_p}{1 - \lambda_p z}.$$

Умножая на $(1 - \lambda_1 z) \ldots (1 - \lambda_p z)$, получаем уравнение, которое должно выполняться для всех z:

$$1 = \sum_{k=1}^{p} c_k \prod_{i \neq k}(1 - \lambda_i z).$$

Пподставляя значения $z = 0, \lambda_k^{-1}$, $k \in \{1, 2, \ldots, p\}$, получаем, что $c_1 + \ldots + c_p = 1$ и

$$c_k = \frac{\lambda_k^{p-1}}{\prod_{i \neq k}(\lambda_k - \lambda_i)}.$$

Следовательно, решение имеет вид

$$h_n = \sum_{k=0}^{\infty} (c_1 \lambda_1^k + \ldots + c_p \lambda_p^k) \omega_{n-k}.$$

Это разложение помогает считать различные характеристики последовательности $h = (h_n)$: например, моменты $\mathsf{E}\left[h_n^k\right]$, ковариации, условные математические ожидания и так далее.

Что мы можем сказать про некоторые характеристики последовательности $h = (h_n)$, если она стационарна (в широком смысле)? Для этого воспользуемся определением модели. Тогда, если $\mu \equiv \mathsf{E}[h_k]$, то

$$\mu = a_0 + a_1\mu + \ldots + a_p\mu \implies \mu = \frac{a_0}{1 - (a_1 + \ldots + a_p)}.$$

Ковариация $R(k) = \mathsf{cov}(h_{n+k}, h_n)$ при $k > 0$ же удовлетворяет следующему уравнению:

$$\begin{aligned} R(k) = \mathsf{cov}(h_{n+k}, h_n) &= \mathsf{cov}(a_0 + a_1 h_{n+k-1} + \ldots + a_p h_{n+k-p} + \sigma\varepsilon_{n+k}, h_n) = \\ &= a_1 R(k-1) + a_2 R(k-2) + \ldots + a_p R(k-p). \end{aligned}$$

Если $k = 0$, то уравнение имеет вид

$$R(0) = a_1 R(1) + \ldots + a_p R(p) + \sigma^2.$$

Оказывается, что аналогичные самые уравнения верны и для корреляций $r(k)$. Их принято называть *уравнениями Юла-Уолкера*.

1.12.3 Модель авторегрессии и скользящего среднего ARMA(p, q) и интегральная модель ARIMA(p, d, q)

Модель ARMA(p, q) совмещает в себе возможности и модели скользящего среднего, и авторегрессионной модели. Это отображено и в названии: ARMA = AR + MA. Перейдём к определению.

Определение 58. Будем называть последовательность $h = (h_n)$ ARMA(p, q)-*моделью*, если

$$h_n = \mu_n + \sigma\varepsilon_n, \text{ где } \mu_n = (a_0 + a_1 h_{n-1} + \ldots + a_p h_{n-p}) + (b_1\varepsilon_{n-1} + \ldots + b_q\varepsilon_{n-q}).$$

Без ограничения общности можно полагать, что $\sigma = 1$. Тогда

$$h_n - a_1 h_{n-1} - \ldots - a_p h_{n-p} = a_0 + \varepsilon_n + b_1\varepsilon_{n-1} + \ldots + b_q\varepsilon_{n-q}$$

Введём два оператора:

$$\alpha(L) = 1 - a_1 L - \ldots - a_p L^p, \quad \beta(L) = 1 + b_1 L + \ldots + b_q L^q.$$

Тогда модель можно записать следующим образом:

$$\alpha(L) h_n = a_0 + \beta(L)\varepsilon_n \iff h_n = \frac{a_0}{1 - (a_1 + \ldots + a_p)} + \frac{\beta(L)}{\alpha(L)}\varepsilon_n.$$

Опять же, поднимем вопрос о существовании стационарного решения этого уравнения. Из предыдущих рассуждений и вида уравнения понятно, что стационарность задаётся авторегрессионной компонентой модели ARMA(p, q). Следовательно, если все корни

уравнения $1 - a_1 z - \ldots - a_p z^p = 0$ меньше единицы по модулю, то стационарное решение существует и единственно (в классе L^2). Для него

$$\mathsf{E}[h_n] = \frac{a_0}{1 - (a_1 + \ldots + a_p)}.$$

Со ковариацией же ситуация немного другая. Если $k > q$, то выполнено уравнение Юла-Уолкера:

$$\begin{aligned} R(k) &= \mathsf{cov}(a_0 + a_1 h_{n+k-1} + \ldots + a_p h_{n+k-p} + \varepsilon_{n+k} + b_1\varepsilon_{n+k-1} + \ldots + b_q\varepsilon_{n+k-q}, h_n) = \\ &= a_1 R(k-1) + \ldots + a_p R(k-p). \end{aligned}$$

Если же $k \leqslant q$, то уже нужно учитывать корреляционную зависимость между h_n и ε_{n-l}, $l \geqslant 0$.

Рассмотрим модель ARMA$(1,1)$:

$$h_n - a_1 h_{n-1} = a_0 + \varepsilon_n + b_1\varepsilon_{n-1}.$$

Предположим, что $|a_1| < 1$. Тогда

$$\begin{aligned} h_n &= \frac{a_0}{1-a_1} + \frac{1 + b_1 L}{1 - a_1 L}\varepsilon_n = \frac{a_0}{1-a_1} + \left(\sum_{k=0}^{\infty} a_1^k L^k\right)(1 + b_1 L)\varepsilon_n = \\ &= \frac{a_0}{1-a_1} + \sum_{k=0}^{\infty} a_1^k L^k \varepsilon_n + b_1 \sum_{k=0}^{\infty} a_1^k L^{k+1}\varepsilon_n = \frac{a_0}{1-a_1} + \varepsilon_n + (a_1 + b_1)\sum_{k=1}^{\infty} a_1^{k-1}\varepsilon_{n-k} \end{aligned}$$

Что мы можем сказать про ковариацию? На самом деле много. Пользуясь этим разложением, получаем, что при $k > 1$ $R(k) = a_1 R(k-1)$ и

$$R(1) = \mathsf{cov}(h_{n+1}, h_n) = \mathsf{cov}(a_0 + a_1 h_n + \varepsilon_{n+1} + b_1\varepsilon_n, h_n) = a_1 R(0) + b_1$$

Осталось решить:

$$\begin{aligned} R(0) &= \mathsf{D}[h_n] = (a_1 + b_1)^2 \sum_{k=0}^{\infty} a_1^{2k} + 1 = \frac{a_1^2 + 2a_1 b_1 + b_1^2}{1 - a_1^2} + 1 = \frac{1 + 2a_1 b_1 + b_1^2}{1 - a_1^2}, \\ R(1) &= \frac{a_1 + 2a_1^2 b_1 + a_1 b_1^2}{1 - a_1^2} + b_1 = \frac{a_1 + b_1 + a_1^2 b_1 + a_1 b_1^2}{1 - a_1^2} = \frac{(a_1 + b_1)(1 + a_1 b_1)}{1 - a_1^2} \\ R(k) &= \frac{(a_1 + b_1)(1 + a_1 b_1)}{1 - a_1^2} a_1^{k-1}, \quad k > 0. \end{aligned}$$

Следовательно, корреляция равна

$$r(k) = \frac{(a_1 + b_1)(1 + a_1 b_1)}{1 + 2a_1 b_1 + b_1^2} a_1^{k-1}, \quad k > 0.$$

Модели ARMA(p,q) достаточно хорошо изучены и успешно применяются при описании стационарных временных рядов. Однако стационарность есть не всегда. Но, переходя от временного ряда $x = (x_n)$ к ряду разностей $h = (h_n)$, где $h_n = \Delta x_n \equiv x_n - x_{n-1}$, или же разностей более высокого порядка: $h_n = \Delta^d x_n = (1 - L)^d x_n$, получается получить стационарность. Именно из этих соображений и появилась модель ARIMA(p,d,q).

Определение 59. Будем говорить, что последовательность $x = (x_n)$ образует ARIMA(p,d,q)-модель, если последовательность $\Delta^d x = (\Delta^d x)$ образует ARMA(p,q)-модель.

Неформально это можно записать так:

$$\Delta^d \mathrm{ARIMA}(p,d,q) = \mathrm{ARMA}(p,q).$$

Проясним смысл модели на примере $\mathrm{ARIMA}(0,1,1)$. Она устроена следующим образом: $h_n = \Delta x_n$, где $h = (h_n)$ является моделью $\mathrm{ARMA}(0,1) = \mathrm{MA}(1)$:

$$\Delta x_n = \mu + (b_0 + b_1 L)\varepsilon.$$

Если ввести оператор «интегрирования» S по правилу $S \equiv \Delta^{-1}$, или, что равносильно,

$$S = (1-L)^{-1} = 1 + L + L^2 + \ldots$$

то формально можно записать, что $x_n = (Sh)_n$, где $h_n = \mu + b_0\varepsilon_n + b_1\varepsilon_{n-1}$.

Следовательно, $x = (x_n)$ можно рассматривать, как результат «интегрирования» последовательности $h = (h_n)$, подчиняющейся модели $\mathrm{MA}(1)$. Это объясняет происхождение названия: $\mathrm{ARIMA} = \mathrm{AR} + \mathrm{I} + \mathrm{MA}$, где I происходит от слова «Integrated».

1.12.4 Нелинейные модели: ARCH и GARCH

Казалось бы, линейные модели всем хороши: имеют широкое применение на практике и устроены очень просто. Однако всё не так радужно. На практике в данных могут возникать самые разные феномены, которые линейная модель описать не может. Например, если мы смотрим на цены, то могут возникать: кластеризация, катастрофическое изменение, тяжёлые хвосты распределений величин $h = (h_n)$ (см. выше), наличие «долгой памяти» у цен и её свойств и так далее. Для того, чтобы как-то описать их, обращаются к *нелинейным* моделям. Таких моделей много, а особенностей в данных ещё больше, поэтому перед исследователями возникает далеко не самая тривиальная задача подбора «подходящей» модели.

Как и раньше, пусть $(\Omega, \mathcal{F}, \mathsf{P})$ — исходное вероятностное пространство, а $\varepsilon = (\varepsilon_n)$ — последовательность iid случайных величин с стандартным нормальным распределением (они будут моделировать «случайность» в рассматриваемых далее моделях). Далее, введём фильтрацию $(\mathcal{F}_n)_{n\geqslant 0}$ по правилу: $\mathcal{F}_0 = \{\varnothing, \Omega\}$, $\mathcal{F}_n = \sigma(\varepsilon_1, \ldots, \varepsilon_n)$.

Модель $\mathrm{ARCH}(p)$ была введена Робертом Энглем следующим образом:

Определение 60. Будем называть последовательность случайных величин $h = (h_n)$ $\mathrm{ARCH}(p)$-моделью, если

$$h_n = \sigma_n \varepsilon_n, \text{ где } \sigma_n^2 = \alpha_0 + \sum_{k=1}^{p} \alpha_i h_{n-k}^2,$$

$\alpha_0 > 0$, $\alpha_i \geqslant 0$, $h_0 = h_0(\omega)$ — случайная величина, не зависящая от $\varepsilon = (\varepsilon_n)_{n\geqslant 1}$.

Обычно h_0 полагают либо константой, либо случайной величиной, для которой второй момент выбирается из соображений «стационарности» значений $\mathsf{E}[h_n^2]$.

Из формулы для σ_n^2 видна явная зависимость от $h_{n-1}^2, \ldots, h_{n-p}^2$. При этом ясно, что большие (малые) значения h_{n-k}^2 приводят к большим (малым) значениям σ_n^2. Возникновение же большого значения h_n^2 при условии, что $h_{n-1}^2, \ldots, h_{n-p}^2$ были мылыми, происходит из-за возникновения большого значения ε_n. Это объясняет то, почему нелинейные модели могут помочь в описании событий наподобие кластерности, то есть группирования значений в пачки «больших» и пачки «маленьких» значений.

ARCHрасшифровывается, как *авторегрессионная модель условной неоднородности*, или AutoRegressive Conditional Heteroskedastic. Смысл каждого слова весьма понятен:

авторегрессивная = прямая зависимость от своих предыдущих значений, условная = задаётся условное распределение $\mathrm{Law}(h_n \mid \mathcal{F}_{n-1})$, неоднородность = σ_n^2 ведёт себя весьма неоднородно.

Теперь рассмотрим модель ARCH(1). Для неё

$$h_n = \sigma_n \varepsilon_n, \text{ где } \sigma_n^2 = a_0 + \alpha_1 h_{n-1}^2.$$

Несложно понять, что выполнены следующие условия:

$$\mathsf{E}[h_n] = 0, \quad \mathsf{E}\big[h^2\big] = \alpha_0 + \alpha_1 \mathsf{E}\big[h_{n-1}^2\big], \quad \mathsf{E}\big[h^2 \bigm| \mathcal{F}_{n-1}\big] = \sigma_n^2 = a_0 + \alpha_1 h_{n-1}^2.$$

Если предположить, что $\alpha_1 \in (0,1)$, то рекуррентное соотношение на матожидание квадрата будет иметь единственное «стационарное» решение: $\mathsf{E}\big[h_{n-1}^2\big] = \alpha_0/(1-\alpha_1)$. Если взять $h_0^2 = \alpha_0/(1-\alpha_1)$, то матожидание квадрата будет постоянно.

Далее, посчитаем четвёртый момент, пользуясь независимостью σ_n и ε_n и тем, что $\mathsf{E}[\varepsilon_n^4] = 3$ (проверьте!):

$$\begin{aligned}\mathsf{E}\big[h_n^4\big] &= \mathsf{E}\big[\sigma_n^4\big]\,\mathsf{E}\big[\varepsilon_n^4\big] = 3\,\mathsf{E}\big[(\alpha_0 + \alpha_1 h_{n-1}^2)^2\big] = \\ &= 3(\alpha_0^2 + 2\alpha_0\alpha_1\,\mathsf{E}\big[h_{n-1}^2\big] + \alpha_1^2\,\mathsf{E}\big[h_{n-1}^4\big]) = \\ &= \frac{3\alpha_0^2(1+\alpha_1)}{1-\alpha_1} + 3\alpha_1^2\,\mathsf{E}\big[h_{n-1}^4\big].\end{aligned}$$

Если предположить, что $\alpha_1 \in (0,1)$ и $3\alpha_1^2 < 1$, то можно найти «стационарное решение»:

$$\mathsf{E}\big[h_n^4\big] = \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}.$$

Из полученных формул несложно получить, что стационарное значение коэффициента эксцесса равно

$$K \equiv \frac{\mathsf{E}[h_n^4]}{(\mathsf{E}[h_n^2])^2} - 3 = \frac{3(1-\alpha_1^2)}{1-3\alpha_1^2} - 3 = \frac{6\alpha_1^2}{1-3\alpha_1^2}.$$

Его положительность говорит о том, что плотность «установившегося» распределения $h = (h_n)$ в окрестности среднего значения «вытянута» вверх. Напомним, что для нормального распределения эксцесс равен нулю.

Теперь заметим, что наша модель задаёт *мартингал-разность*, то есть $\mathsf{E}[h_n \mid \mathcal{F}_{n-1}] = 0$. Отсюда следует, что для любого $k < n$

$$\mathsf{E}[h_n h_k] = \mathsf{E}[\mathsf{E}[h_n h_k \mid \mathcal{F}_{n-1}]] = \mathsf{E}[h_k\,\mathsf{E}[h_n \mid \mathcal{F}_{n-1}]] = 0.$$

Это означает *ортогональность* значений: $\mathsf{cov}(h_n, h_m) = 0$ при $n \neq m$. Но из ортогональности этих значений не следует независимость, так как совместное распределение $\mathrm{Law}(h_m, h_n)$ не является гауссовским при $\alpha_1 > 0$ (почему?). Но если они не независимы, то между ними есть какая-то зависимость. Как её исследовать? Для этого посмотрим на корреляционную зависимость *квадратов* h_n^2 и h_m^2 в «стационарном» случае. Посчитаем

дисперсию и ковариацию для соседних значений:

$$\begin{aligned}
\mathsf{D}\big[h_n^2\big] &= \mathsf{E}\big[h_n^4\big] - (\mathsf{E}\big[h_n^2\big])^2 = \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)} - \frac{\alpha_0^2}{(1-\alpha_1)^2} = \\
&= \left(\frac{\alpha_0}{1-\alpha_1}\right)^2 \left(\frac{3-3\alpha_1^2}{1-3\alpha_1^2} - 1\right) = \frac{2}{1-3\alpha_1^2}\left(\frac{\alpha_0}{1-\alpha_1}\right)^2. \\
\mathsf{E}\big[h_n^2 h_{n-1}^2\big] &= \mathsf{E}\big[(\alpha_0 + \alpha_1 h_{n-1}^2)\varepsilon_n^2 h_{n-1}^2\big] = \mathsf{E}\big[\alpha_0 h_{n-1}^2\big] + \mathsf{E}\big[\alpha_1 h_{n-1}^4\big] = \\
&= \frac{\alpha_0^2}{1-\alpha_1} + \frac{2\alpha_1}{1-3\alpha_1^2}\left(\frac{\alpha_0}{1-\alpha_1}\right)^2 = \frac{\alpha_0^2}{1+\alpha_1}\left(1 - \frac{2\alpha_1}{(1-\alpha_1)(1-3\alpha_1^2)}\right) = \\
&= \frac{\alpha_0^2}{1-\alpha_1}\left(\frac{1-\alpha_1-3\alpha_1^2+3\alpha_1^2+2\alpha_1}{(1-3\alpha_1^2)(1-\alpha_1)}\right) = \frac{\alpha_0^2}{1-\alpha_1}\frac{1+3\alpha_1}{1-3\alpha_1^2}, \\
\mathsf{cov}(h_n^2, h_{n-1}^2) &= \mathsf{E}\big[h_n^2 h_{n-1}^2\big] - \mathsf{E}\big[h_n^2\big]\,\mathsf{E}\big[h_{n-1}^2\big] = \frac{1+3\alpha_1}{1-3\alpha_1^2}\frac{\alpha_0^2}{1-\alpha_1} - \frac{\alpha_0^2}{(1-\alpha_1)^2} = \\
&= \left(\frac{\alpha_0}{1-\alpha_1}\right)^2 \left(\frac{(1-\alpha_1)(1+3\alpha_1)}{1-3\alpha_1^2} - 1\right) = \frac{2\alpha_1}{1-3\alpha_1^2}\left(\frac{\alpha_0}{1-\alpha_1}\right)^2, \\
r(1) &\equiv \frac{\mathsf{cov}(h_n^2, h_{n-1}^2)}{\sqrt{\mathsf{D}[h_n^2]\,\mathsf{D}\big[h_{n-1}^2\big]}} = \alpha_1.
\end{aligned}$$

Далее посчитаем корреляцию в общем случае. Заметим, что для $k < n$

$$\begin{aligned}
\mathsf{E}\big[h_n^2 h_{n-k}^2\big] &= \mathsf{E}\big[\mathsf{E}\big[h_n^2 h_{n-k}^2 \bigm| \mathcal{F}_{n-1}\big]\big] = \mathsf{E}\big[h_{n-k}^2\,\mathsf{E}\big[h_n^2 \bigm| \mathcal{F}_{n-1}\big]\big] = \\
&= \mathsf{E}\big[h_{n-k}^2\,\mathsf{E}\big[\sigma_n^2\varepsilon_n^2 \bigm| \mathcal{F}_{n-1}\big]\big] = \mathsf{E}\big[h_{n-k}^2\sigma_n^2\,\mathsf{E}\big[\varepsilon_n^2 \bigm| \mathcal{F}_{n-1}\big]\big] = \\
&= \mathsf{E}\big[h_{n-k}^2(\alpha_0 + \alpha_1 h_{n-1}^2)\big] = \alpha_0\,\mathsf{E}\big[h_{n-k}^2\big] + \alpha_1\,\mathsf{E}\big[h_{n-1}^2 h_{n-k}^2\big].
\end{aligned}$$

Тогда в стационарном случае это равенство можно преобразовать следующим образом:

$$\mathsf{E}\big[h_n^2 h_{n-k}^2\big] - \left(\frac{\alpha_0}{1-\alpha_1}\right)^2 = \frac{\alpha_0^2}{1-\alpha_1} - (1-\alpha_1)\left(\frac{\alpha_0}{1-\alpha_1}\right)^2 + \alpha_1\left(\mathsf{E}\big[h_{n-1}^2 h_{n-k}^2\big] - \left(\frac{\alpha_0}{1-\alpha_1}\right)^2\right).$$

Отсюда следует, что

$$r(k) = \alpha_1 r(k-1) \implies r(k) = \alpha_1^k.$$

В названии модели ARCH(p) фигурирует слово «авторегрессионая». Оказывается, что модель ARCH(p) сводится к AR(p)-модели. Действительно, пусть $\nu_n = h_n^2 - \sigma_n^2$. Если $\mathsf{E}[h_n^2] < \infty$, то $\mathsf{E}[\nu_n \mid \mathcal{F}_{n-1}] = 0$ и $\nu = (\nu_n)$ образует мартингал-разность относительно $(\mathcal{F}_n)_{n\geqslant 0}$. Далее, введём обозначение $x_n = h_n^2$. Тогда

$$x_n = \alpha_0 + \alpha_1 x_{n-1} + \ldots + \alpha_p x_{n-p} + \nu_n.$$

Успех условно-гауссовской модели ARCH(p), давшей объяснение многим феноменам в поведении финансовых индексов, породил целую кучу различных её обобщений, преследующих цель «ухватить», дать описание ряда других эффектов. Исторически первое обобщение было введено Тимом Боллерслевом в 1986-м году: так называемая *обобщённая* ARCH-*модель*, характеризуемая двумя параметрами (p, q). Её принято обозначать GARCH(p, q). Определяется она следующим образом:

$$h_n = \sigma_n\varepsilon_n, \text{ где } \sigma_n^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i h_{n-i}^2 + \sum_{j=1}^{q}\beta_j\sigma_{n-j}^2.$$

Основное преимущество GARCH(p,q)-моделей перед их прародительницей, ARCH(p)-моделью, состоит в подборе параметров модели. На практике периодически может оказаться так, что при подгонке статистических данных моделями ARCH(p) параметр p становится слишком большим (что усложняет анализ модели), в то время как при подгонке GARCH(p,q)-моделями можно ограничиваться небольшими значениями p и q (экспериментальный факт!).

Как и всегда, подробнее рассмотрим частный случай: GARCH$(1,1)$. Для него

$$h_n = \sigma_n\varepsilon_n, \text{ где } \sigma_n^2 = \alpha_0 + \alpha_1 h_{n-1}^2 + \beta_1\sigma_{n-1}^2.$$

Отсюда ясно, что

$$\mathsf{E}[h_n] = 0, \quad \mathsf{E}\big[h_n^2\big] = \mathsf{E}\big[\sigma_n^2\big] = \alpha_0 + \alpha_1\,\mathsf{E}\big[h_{n-1}^2\big] + \beta_1\,\mathsf{E}\big[\sigma_{n-1}^2\big] = \alpha_0 + (\alpha_1+\beta_1)\,\mathsf{E}\big[h_{n-1}^2\big].$$

Если $\alpha_1+\beta_1<1$, то существует «стационарное» значение $\mathsf{E}[h_n^2]$, равное

$$\mathsf{E}\big[h_n^2\big] = \frac{\alpha_0}{1-\alpha_1-\beta_1}.$$

Далее, посчитаем четвёртый момент.

$$\begin{aligned}\mathsf{E}\big[h_n^4\big] &= \mathsf{E}\big[\sigma_n^4\big]\,\mathsf{E}\big[\varepsilon_n^4\big] = 3\,\mathsf{E}\big[(\alpha_0+\alpha_1 h_{n-1}^2+\beta_1\sigma_{n-1}^2)^2\big] = \\ &= 3(\alpha_0^2+\alpha_1^2\,\mathsf{E}\big[h_{n-1}^4\big]+\beta_1^2\,\mathsf{E}\big[\sigma_1^4\big]+2\alpha_0\alpha_1\,\mathsf{E}\big[h_{n-1}^2\big]+2\alpha_0\beta_1\,\mathsf{E}\big[\sigma_{n-1}^2\big]+ \\ &+ 2\alpha_1\beta_1\,\mathsf{E}\big[h_{n-1}^2\sigma_{n-1}^2\big]) = 3\alpha_0^2+3\alpha_1^2\,\mathsf{E}\big[h_{n-1}^4\big]+\beta_1^2\,\mathsf{E}\big[h_{n-1}^4\big]+ \\ &+ 6\alpha_0\alpha_1\,\mathsf{E}\big[h_{n-1}^2\big]+6\alpha_0\beta_1\,\mathsf{E}\big[h_{n-1}^2\big]+2\alpha_1\beta_1\,\mathsf{E}\big[h_{n-1}^4\big].\end{aligned}$$

Если предположить, что $3\alpha_1^2+\beta_1^2+2\alpha_1\beta_1<1$, то можно найти «стационарное» решение:

$$(1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1)\,\mathsf{E}\big[h_n^4\big] = 3\alpha_0^2\left(1+\frac{2(\alpha_1+\beta_1)}{1-\alpha_1-\beta_1}\right),$$

$$\mathsf{E}\big[h_n^4\big] = \frac{3\alpha_0^2(1+\alpha_1+\beta_1)}{(1-\alpha_1-\beta_1)(1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1)}.$$

Отсюа можно получить коэффициент эксцесса:

$$K = \frac{3(1-(\alpha_1+\beta_1)^2)}{1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1} - 3 = \frac{6\alpha_1^2}{1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1}.$$

Теперь посчитаем корреляционную функцию для этой модели. Заметим, что

$$\begin{aligned}\mathsf{E}\big[h_n^2h_{n-1}^2\big] &= \mathsf{E}\big[(\alpha_0+\alpha_1h_{n-1}^2+\beta_1\sigma_{n-1}^2)\varepsilon_n^2h_{n-1}^2\big] = \\ &= \alpha_0\,\mathsf{E}\big[h_{n-1}^2\big]+(\alpha_1+\beta_1/3)\,\mathsf{E}\big[h_{n-1}^4\big] = \\ &= \frac{\alpha_0^2}{1-\alpha_1-\beta_1}+\frac{\alpha_0^2(3\alpha_1+\beta_1)(1+\alpha_1+\beta_1)}{(1-\alpha_1-\beta_1)(1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1)} = \\ &= \frac{\alpha_0^2}{1-\alpha_1-\beta_1}\left(1+\frac{(3\alpha_1+\beta_1)(1+\alpha_1+\beta_1)}{1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1}\right) \\ \mathsf{cov}(h_n^2,h_{n-1}^2) &= \mathsf{E}\big[h_n^2h_{n-1}^2\big]-\mathsf{E}\big[h_n^2\big]\,\mathsf{E}\big[h_{n-1}^2\big] = \\ &= \frac{\alpha_0^2}{1-\alpha_1-\beta_1}\left(1+\frac{(3\alpha_1+\beta_1)(1+\alpha_1+\beta_1)}{1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1}\right)-\frac{\alpha_0^2}{(1-\alpha_1-\beta_1)^2} = \\ &= \frac{\alpha_0^2}{(1-\alpha_1-\beta_1)^2}\left(1-\alpha_1-\beta_1+\frac{(3\alpha_1+\beta_1)(1-(\alpha_1+\beta_1)^2)}{1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1}-1\right) = \\ &= \frac{\alpha_0^2}{(1-\alpha_1-\beta_1)^2}\left(\frac{(3\alpha_1+\beta_1)(1-\alpha_1^2-\beta_1^2-2\alpha_1\beta_1)}{1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1}-(\alpha_1+\beta_1)\right) = \\ &= \left(\frac{\alpha_0}{1-\alpha_1-\beta_1}\right)^2\frac{2\alpha_1(1-\alpha_1\beta_1-\beta_1^2)}{1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1}\end{aligned}$$

$$\mathsf{D}\big[h_n^2\big] = \mathsf{E}\big[h_n^4\big] - (\mathsf{E}\big[h_n^2\big])^2 = \frac{3\alpha_0^2(1+\alpha_1+\beta_1)}{(1-\alpha_1-\beta_1)(1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1)} - \frac{\alpha_0^2}{(1-\alpha_1-\beta_1)^2} =$$
$$= \frac{\alpha_0^2}{(1-\alpha_1-\beta_1)^2}\left(\frac{3(1-\alpha_1^2-\beta_1^2-2\alpha_1\beta_1)}{1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1} - 1\right) =$$
$$= \left(\frac{\alpha_0}{1-\alpha_1-\beta_1}\right)^2 \frac{2(1-2\alpha_1\beta_1-\beta_1^2)}{1-3\alpha_1^2-\beta_1^2-2\alpha_1\beta_1}$$

Отсюда сразу же получаем, что

$$\rho(1) \equiv \frac{\mathsf{cov}(h_n^2, h_{n-1}^2)}{\sqrt{\mathsf{D}[h_n^2]\,\mathsf{D}\big[h_{n-1}^2\big]}} = \frac{\alpha_1(1-\alpha_1\beta_1-\beta_1^2)}{1-2\alpha_1\beta_1-\beta_1^2}.$$

Далее заметим, что

$$\begin{aligned}\mathsf{E}\big[h_n^2 h_{n-k}^2\big] &= \mathsf{E}\big[\mathsf{E}\big[h_n^2 h_{n-k}^2 \mid \mathcal{F}_{n-1}\big]\big] = \mathsf{E}\big[h_{n-k}^2\,\mathsf{E}\big[h_n^2 \mid \mathcal{F}_{n-1}\big]\big] = \\ &= \mathsf{E}\big[h_{n-k}^2\,\mathsf{E}\big[\sigma_n^2\varepsilon_n^2 \mid \mathcal{F}_{n-1}\big]\big] = \mathsf{E}\big[h_{n-k}^2\sigma_n^2\,\mathsf{E}\big[\varepsilon_n^2 \mid \mathcal{F}_{n-1}\big]\big] = \\ &= \mathsf{E}\big[h_{n-k}^2(\alpha_0 + \alpha_1 h_{n-1}^2 + \beta_1\sigma_{n-1}^2)\big] = \\ &= \alpha_0\,\mathsf{E}\big[h_{n-k}^2\big] + (\alpha_1+\beta_1)\,\mathsf{E}\big[h_{n-1}^2 h_{n-k}^2\big]\end{aligned}$$

Следовательно, в стационарном случае это можно перезаписать следующим образом:

$$\mathsf{E}\big[h_n^2 h_{n-k}^2\big] - \left(\frac{\alpha_0}{1-\alpha_1-\beta_1}\right)^2 = \frac{\alpha_0^2}{1-\alpha_1-\beta_1} + (\alpha_1+\beta_1-1)\left(\frac{\alpha_0}{1-\alpha_1-\beta_1}\right)^2 +$$
$$+ (\alpha_1+\beta_1)\left(\mathsf{E}\big[h_{n-1}^2 h_{n-k}^2\big] - \left(\frac{\alpha_0}{1-\alpha_1-\beta_1}\right)^2\right)$$

Тогда $\mathsf{cov}(h_n^2, h_{n-k}^2) = (\alpha_1+\beta_1)\,\mathsf{cov}(h_{n-1}^2, h_{n-k}^2)$ и

$$\rho(k) \equiv \frac{\mathsf{cov}(h_n^2, h_{n-k}^2)}{\sqrt{\mathsf{D}[h_n^2]\,\mathsf{D}\big[h_{n-k}^2\big]}} = \frac{\alpha_1(1-\alpha_1\beta_1-\beta_1^2)}{1-2\alpha_1\beta_1-\beta_1^2}(\alpha_1+\beta_1)^{k-1}.$$

1.13 Назад в прошлое: динамическое программирование

Для описания некоторых алгоритмов, которые появятся дальше, будет нужен метод динамического программирования. Рассмотрим его на примере *задачи объезда стран*:

Задача 1. *Путешественник хочет устроить турне: начиная с города x_s и заканчивая городм x_f, он хочет посетить k стран в определённой последовательности. Далее, он ночует в каждой стране, которую посещает. Однако у него мало денег, поэтому он хочет устроить турне с минимальными затратами. Какие города он должен посетить?*

Понятно, что решать эту задачу перебором слишком долго: если в каждой стране n городов, то придётся рассмотреть n^k вариантов поездок. Попробуем немного улучшить эту оценку.

Пусть города задаются следующим образом: x_{ij} означает «j-й город в i-й стране» (для этого занумеруем все страны и все города в них). Далее, пусть $f(x_{ij}, x_{i+1,k})$ — стоимость

проезда из города x_{ij} в город $x_{i+1,k}$, а $h(x_{ij})$ — стоимость ночлега в городе x_{ij}. Пользуясь этими обозначениями, введём *функцию Беллмана* $V(x)$ по следующему правилу:

$$V(x_s) = 0, \quad V(x_{i+1,k}) = \min_j \left[V(x_{ij}) + f(x_{ij}, x_{i+1,k}) + h(x_{i+1,k})\right].$$

Оказывается, что смысл функции Беллмана $V(x)$ — это наименьшая сумма, которую нужно затратить для того, чтобы добраться из города x_s в город x. Это утверждение несложно доказать, пользуясь индукцией.

Пользуясь этой функцией, несложно получить решение. Введём функцию $S(x)$, действующую по правилу

$$S(x_{i+1,k}) = \arg\min_j \left[V(x_{ij}) + f(x_{ij}, x_{i+1,k}) + h(x_{i+1,k})\right].$$

Тогда оптимальный путь $(x_1^*, \ldots, x_k^*)$ может быть построен рекурсивным вызовом функции S: $x_k^* = S(x_f)$, $x_m^* = S(x_{m+1}^*)$.

Решение задачи объезда стран показывает пример применения метода динамического программирования. Но, понятное дело, этот метод далеко не универсален. Когда его применение оправдано?

- Когда для решения задачи нужно решить перекрывающиеся подзадачи и их решения можно запомнить.
- Когда задача обладает оптимальной подструктурой, или же кусочной оптимальностью. Это означает следующее: например, если известно, что оптимальный путь из A в B содержит C и D и известен оптимальный путь из C в D, то этот путь станет частью оптимального пути из A в B.

1.14 Скрытые марковские модели

1.14.1 Мотивация и основные понятия

Допустим, что мы хотим определить среднюю годовую температуру в каком-то месте на Земле за определённый период времени. Далее, усложним задачу тем, что возьмём очень давние времена — когда ещё не существовало термометров. К сожалению, отправиться назад в прошлое и замерить температуру не представляется возможным, поэтому придётся ориентироваться на косвенные признаки определённых температур.

Для простоты скажем, что температуру мы измеряем очень просто — различаются лишь состояния «горячо» и «холодно». Предположим, что современные измерения показывают, что тёплый год следует за тёплым годом с вероятностью 0,7, а холодный за холодным — с вероятностью 0,6. Далее предположим (на не самом понятном основании, но не суть), что это предположение имеет место и для давних времён. Тогда переходная матрица будет устроена следующим образом (Г — горячо, Х – холодно):

$$\begin{array}{c} \\ \Г \\ \text{Х} \end{array} \begin{array}{c} \begin{array}{cc} \Г & \text{Х} \end{array} \\ \begin{pmatrix} 0{,}7 & 0{,}3 \\ 0{,}4 & 0{,}6 \end{pmatrix} \end{array}$$

Преположим, что исследование показало зависимость диаметра древесных колец от средней годовой температуры. Опять же, для простоты будем различать только три раз-

мера — маленькие (М), средние (С) и большие (Б). Напоследок предположим, что исследование показало следующую вероятностную зависимость между ними:

$$\begin{array}{c} \\ \text{Г} \\ \text{Х} \end{array} \begin{array}{c} \begin{array}{ccc} \text{М} & \text{С} & \text{Б} \end{array} \\ \begin{pmatrix} 0{,}1 & 0{,}4 & 0{,}5 \\ 0{,}7 & 0{,}2 & 0{,}1 \end{pmatrix} \end{array}$$

Для данной системы *состоянием* служит среднегодовая температура — либо «горячо», либо «холодно». При этом переход из текущего состояния в следующее является марковским (первого порядка[11]), так как мы предположили, что он зависит только от текущего состояния и полностью задаётся переходной матрицей. Однако настоящее состояние «спрятано», так как мы не можем его определить напрямую.

Хоть мы и не можем наблюдать состояние (температуру) в прошлом, но мы можем замерить размер древесных колец. Вероятностная зависимость между диаметром колец и температурой даёт нам информацию о состоянии. Так как состояния «скрыты», то такие системы называют *скрытыми марковскими моделями.*

Теперь можно дать формальное определение.

Определение 61. Скрытая марковская модель (первого порядка) — это вероятностная модель последовательности, которая состоит из набора *наблюдаемых* переменных $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, где $\mathbf{x}_k \in \mathbb{R}^d$, и набора *латентных* (или *скрытых*) переменных

$$\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_n\}, \quad \mathbf{t}_k \in \{0, 1\}^K, \quad \sum_{i=1}^{K} t_{ki} = 1.$$

В данной модели латентные переменные являются *бинарными* и кодируют K состояний, поэтому их иногда называют переменными состояния. Значения наблюдаемого вектора $\mathbf{x}_k$, взятого в момент времени k, зависит только от скрытого состояния $\mathbf{t}_k$, которое, в свою очередь, зависит только от скрытого состояния в предыдущий момент времени $\mathbf{t}_{k-1}$.

Скрытые марковские модели имеют множество применений: распознание речи, образов на видео, поведения, анализирование фондовых рынков, естественных языков, ДНК и так далее.

Теперь опишем то, как можно задать некоторые параметры вероятносной модели. Начнём с распределения $\mathbf{t}_n$. Пусть в скрытой марковской модели K состояний. Закодируем их состояния в момент времени n бинарным вектором $\mathbf{t}_n = (t_{n1}, \ldots, t_{nK})$ по правилу:

$$t_{ij} = \begin{cases} 1, & \text{система находится в состоянии } j \text{ в момент времени } i, \\ 0, & \text{иначе.} \end{cases}$$

Так как в векторе $\mathbf{t}_n$ может быть только один ненулевой элемент (предполагается, что система не может находиться в двух разных состояниях одновременно), то распределение $\mathbf{t}_n$ относительно $\mathbf{t}_{n-1}$ $p(\mathbf{t}_n \mid \mathbf{t}_{n-1})$ можно задать матрицей $\mathbf{A}$, где $\mathbf{A}_{ij} = p(t_{nj} = 1 \mid t_{n-1,i} = 1)$. Стоит заметить, что $\sum_{j=1}^{K} \mathbf{A}_{ij} = 1$. Следовательно, распределение можно записать следующим образом:

$$p(\mathbf{t}_n \mid \mathbf{t}_{n-1}) = \prod_{i=1}^{K} \prod_{j=1}^{K} \mathbf{A}_{ij}^{t_{n-1,i} t_{nj}}.$$

[11] Марковская цепь называется цепью k-го порядка, если состояние зависит от k предыдущих состояний. По умолчанию будем считать, что все рассматриаемые нами модели имеют первый порядок.

Далее, нужно задать начальное распределение $p(\mathbf{t}_1)$. Пусть $\pi_i = p(t_{1i} = 1)$. Тогда

$$p(\mathbf{t}_1) = \prod_{i=1}^{K} \pi_i^{t_{1i}}.$$

Хоть матрица $\mathbf{A}$ может быть почти любой (нужна только неотрицательность элементов и равенство единице сумме элементов в строке), но с точки зрения скрытых марковских моделей более интересны матрицы $\mathbf{A}$ с преобладающими элементами на диагонали (то есть более вероятно то, что система не изменит своего состояния). В этом случае можно сказать, что процесс находится в одном и том же состоянии в течение какого-то отрезка времени. Отсюда получаем простую физическую интерпретацию скрытой марковской модели: это процесс, который иногда меняет свои характеристики.

Теперь вспомним, что наблюдаемая переменная $\mathbf{x}_n$ зависит только от переменной состояния $\mathbf{t}_n$. Следовательно, разумно рассматривать условное распределение $p(\mathbf{x}_n \mid \mathbf{t}_n)$. Обычно предполагается, что оно известно с точностью до параметров $\varphi_k, k \in \{1, \ldots, K\}$: то есть, если $t_{ni} = 1$, то $p(\mathbf{x}_n \mid \mathbf{t}_n) = p(\mathbf{x}_n \mid \varphi_i)$. Следовательно,

$$p(\mathbf{x}_n \mid \mathbf{t}_n) = \prod_{k=1}^{K} p(\mathbf{x}_n \mid \varphi_k)^{t_{nk}}.$$

Введённых выше параметров достаточно для полного описания скрытой марковской модели. Их собирают в *набор параметров*

$$\Theta = (\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\varphi}), \text{ где } \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K), \quad \boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_K).$$

Теперь можно сформулировать основные задачи теории скрытых марковских процессов:

- *Обучение с учителем.* Пусть есть некоторая последовательность $\mathbf{X}$, для которой известны латентные переменные $\mathbf{T}$. По обучающей выборке нужно оценить набор параметров Θ.
- *Сегментация.* Пусть известна последовательность наблюдаемых переменных $\mathbf{X}$ и набор параметров Θ. По ним нужно построить максимально правдоподобный набор латентных переменных $\mathbf{T}$, то есть найти $\arg\max_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta)$.
- *Обучение без учителя.* Пусть известна последовательность наблюдаемых переменных $\mathbf{X}$ и число состояний K. Нужно оценить набор параметров Θ. Подзадача — *нахождение маргинального распределения*: найти $p(t_n \mid \mathbf{X}, \Theta)$.
- *Прогнозирование.* Пусть известна некоторая последовательность $\mathbf{X}$ длины N. Нужно оценить наблюдаемый вектор в момент времени $N + 1$, то есть найти $p(\mathbf{x}_{N+1} \mid \mathbf{X})$.

1.14.2 Пример 1: обучение с учителем

Пусть дана обучающая выборка $(\mathbf{X}, \mathbf{T})$, представляющая собой одну или несколько последовательностей, в которых известны значения скрытых компонент. По ней нужно оценить набор параметров Θ. Как это сделать?

Для начала посмотрим на вероятность того, что обучающая выборка «соответствует» набору, то есть на $p(\mathbf{X}, \mathbf{T} \mid \Theta)$. По построению скрытой марковской модели это распределение задаётся следующим образом:

$$p(\mathbf{X}, \mathbf{T} \mid \Theta) = p_{\boldsymbol{\pi}}(\mathbf{t}_1) \prod_{k=1}^{N} p_{\boldsymbol{\varphi}}(\mathbf{x}_k \mid \mathbf{t}_k) \prod_{k=2}^{N} p_{\mathbf{A}}(\mathbf{t}_k \mid \mathbf{t}_{k-1}).$$

Подставляем ранее выписанные формулы для $p_{\boldsymbol{\pi}}(\mathbf{t}_1)$, $p_{\boldsymbol{\varphi}}(\mathbf{x}_k \mid \mathbf{t}_k)$ и $p_{\mathbf{A}}(\mathbf{t}_k \mid \mathbf{t}_{k-1})$:

$$p(\mathbf{X}, \mathbf{T} \mid \Theta) = \prod_{i=1}^{K} \pi_i^{t_{1i}} \left(\prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_n \mid \varphi_k)^{t_{nk}} \right) \left(\prod_{n=2}^{K} \prod_{i=1}^{K} \prod_{j=1}^{K} \mathbf{A}_{ij}^{t_{n-1,i} t_{nj}} \right)$$

Оценивать Θ будем с помощью метода максимального правдоподобия, то есть

$$\Theta_{ML} = \arg\max_{\Theta} p(\mathbf{X}, \mathbf{T} \mid \Theta) = \arg\max_{\Theta} \log p(\mathbf{X}, \mathbf{T} \mid \Theta)$$

Второе равенство корректно, так как $\log x$ — монотонно возрастающая функция. Теперь выпишем $\log p(\mathbf{X}, \mathbf{T} \mid \Theta)$:

$$\log p(\mathbf{X}, \mathbf{T} \mid \Theta) = \sum_{i=1}^{K} t_{1i} \log \pi_i + \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log p(\mathbf{x}_n \mid \varphi_k) + \sum_{n=2}^{N} \sum_{i=1}^{K} \sum_{j=1}^{K} t_{n-1,i} t_{nj} \log \mathbf{A}_{ij}.$$

Для нахождения оценок воспользуемся методом Лагранжа. Вспомним, что на параметры наложены следующие ограничения:

$$\sum_{i=1}^{K} \pi_i = 1, \quad \forall i \in \{1, 2, \ldots, K\} \ \sum_{j=1}^{K} A_{ij} = 1.$$

Выпишем лагранжиан:

$$\mathcal{L}(\Theta, \lambda, \boldsymbol{\mu}) = \log p(\mathbf{X}, \mathbf{T} \mid \Theta) + \lambda \left(\sum_{i=1}^{K} \pi_i - 1 \right) + \sum_{i=1}^{K} \mu_i \left(\sum_{j=1}^{K} A_{ij} - 1 \right) \to \text{extr}.$$

Теперь начнём искать оценки.

- Начнём с оценки $\boldsymbol{\pi}$. Для неё

$$\frac{\partial \mathcal{L}(\Theta, \lambda, \boldsymbol{\mu})}{\partial \pi_i} = \frac{t_{1i}}{\pi_i} + \lambda = 0 \implies \pi_i = -\frac{t_{1i}}{\lambda}.$$

  Теперь вспомним ограничения на $\boldsymbol{\pi}$ и на $\mathbf{t}_1$:

$$\sum_{i=1}^{K} \pi_i = -\frac{1}{\lambda} \sum_{i=1}^{K} t_{1i} = -\frac{1}{\lambda} = 1 \implies \pi_i = t_{1i}.$$

- Теперь оценим матрицу $\mathbf{A}$:

$$\frac{\partial \mathcal{L}(\Theta, \lambda, \boldsymbol{\mu})}{\partial \mathbf{A}_{ij}} = \sum_{n=2}^{N} \frac{t_{n-1,i} t_{nj}}{A_{ij}} + \mu_i = 0 \implies A_{ij} = \sum_{n=2}^{N} \frac{t_{n-1,i} t_{nj}}{\mu_i}.$$

  Осталось избавиться от μ_i.

$$\sum_{j=1}^{K} A_{ij} = 1 \implies \sum_{j=1}^{K} \sum_{n=2}^{N} \frac{t_{n-1,i} t_{nj}}{\mu_i} = \sum_{n=2}^{N} \frac{t_{n-1,i}}{\mu_i} = 1 \implies \mu_i = \sum_{n=2}^{N} t_{n-1,i}$$

  Отсюда получаем оценку:

$$\mathbf{A}_{ij} = \frac{\sum_{n=2}^{N} t_{n-1,i} t_{nj}}{\sum_{n=2}^{N} t_{n-1,i}}.$$

- Осталось оценить $\boldsymbol{\varphi}$. Снова возьмём частную производную:

$$\frac{\partial \mathcal{L}(\Theta, \lambda, \boldsymbol{\mu})}{\partial \varphi_i} = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{\partial \log p(\mathbf{x}_n \mid \varphi_k)}{\partial \varphi_i} = \sum_{n\,:\,t_{ni}=1} \frac{\partial \log p(\mathbf{x}_n \mid \varphi_i)}{\partial \varphi_i} = 0.$$

Получилась классическая задача на максимизацию правдоподобия по выборке iid объектов. Получаем, что

$$\varphi_i = \arg\max \sum_{n\,:\,t_{ni}=1} \log p(\mathbf{x}_n \mid \varphi_i).$$

Для нахождения этого решения можно воспользоваться методами восстановления плотностей. Об одном из них, EM-алгоритме, будет рассказано позднее.

1.14.3 Пример 2: сегментация. Алгоритм Витерби

Пусть нам известна последовательность наблюдаемых переменных $\mathbf{X}$ и набор параметров скрытой марковской модели Θ. Как по ним построить максимально правдоподобный набор латентных переменных $\mathbf{T}$?

Как мы сказали ранее, нужный набор равен

$$\arg\max_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta).$$

Теперь надо заметить, что $p(\mathbf{X} \mid \Theta)$ не зависит от $\mathbf{T}$. Тогда можно провести следующую цепочку равенств:

$$\arg\max_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta) = \arg\max_{\mathbf{T}} \frac{p(\mathbf{X}, \mathbf{T} \mid \Theta)}{p(\mathbf{X} \mid \Theta)} = \arg\max_{\mathbf{T}} p(\mathbf{X}, \mathbf{T} \mid \Theta) = \arg\max_{\mathbf{T}} \log p(\mathbf{X}, \mathbf{T} \mid \Theta).$$

Оказывается, что это есть ни что иное, как задача динамического программирования! Действительно, вспомним задачу об объезде стран. Посмотрим на логарифм:

$$\log p(\mathbf{X}, \mathbf{T} \mid \Theta) = \sum_{i=1}^{K} t_{1i} \log \pi_i + \sum_{n=2}^{N} \sum_{i=1}^{K} \sum_{j=1}^{K} t_{n-1,i} t_{nj} \log \mathbf{A}_{ij} + \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log p(\mathbf{x}_n \mid \varphi_k).$$

По аналогии с задачей, первое слагаемое определяет «пункт отбытия», слагаемые второй суммы — стоимость переезда из одной страны в другую, а слагаемые третьей — «стоимость ночлега» в выбранном городе. Составим функцию Беллмана:

$$V(t_{1j}) = \log \pi_j, \quad V(t_{nj}) = \max_i \left[V(t_{n-1,i}) + t_{n-1,i} t_{nj} \log \mathbf{A}_{ij} + t_{nj} \log p(\mathbf{x}_n \mid \varphi_j) \right]$$

Теперь, как и раньше, определяем функцию S по правилу

$$S(t_{1j}) = \varnothing, \quad S(t_{nj}) = \arg\max_i \left[V(t_{n-1,i}) + t_{n-1,i} t_{nj} \log \mathbf{A}_{ij} + t_{nj} \log p(\mathbf{x}_n \mid \varphi_j) \right]$$

Выполнив прямой обход по сигналу, мы определим значения $V(t_{ij})$ и $S(t_{ij})$. Далее, выполнив обратный проход, мы получим номера оптимальных состояний $(i^*(1), \ldots, i^*(N))$ по следующему правилу:

$$i^*(N) = \arg\max_i V(t_{Ni}), \quad i^*(k) = S(t_{k+1,i^*(k+1)})$$

Легко понять, что $\mathbf{t}_k$ определяются так: $t_{k,i^*(k)} = 1$, $t_{kl} = 0$ при $l \neq i^*(k)$.

Чем так хорош этот алгоритм? Он позволяет достаточно быстро производить сегментацию очень длинных сигналов. Более того, при некоторых модификациях он может делать это в реальном времени (с небольшой задержкой, конечно).

1.14.4 Пример 3: обучение без учителя. EM-алгоритм

Преположим, что есть некоторая графическая модель, в которой известна лишь часть значений переменных, при этом атомарные распределения известны с точностью до набора параметров Θ. Нужно оценить Θ по наблюдаемым величинам методом максимального правдоподобия, то есть найти

$$\Theta_{ML} = \arg\max_{\Theta} p(\mathbf{X} \mid \Theta).$$

Это называют *методом неполного правдоподобия*. По правилу суммирования неполное правдоподобие может быть получено суммированием по скрытым переменных полного правдоподобия, то есть

$$p(\mathbf{X} \mid \Theta) = \sum_{\mathbf{T}} p(\mathbf{X}, \mathbf{T} \mid \Theta).$$

Для многих моделей (в частности, для байесовских сетей) полное правдоподобие считается достаточно просто.

Далее, для удобства часто переходят к логарифму, что возможно из-за строгой монотонности $\log x$. В частности, раньше мы получили явные формулы для $\arg\max_\Theta p(\mathbf{X}, \mathbf{T} \mid \Theta) = \arg\max_\Theta \log p(\mathbf{X}, \mathbf{T} \mid \Theta)$. Вот здесь и возникает первый подводный камень. Из-за возникающего «логарифма суммы» оптимизация становится крайне сложной.

Теперь построим итеративный алгоритм для нахождения Θ_{ML}, называемый *EM-алгоритмом* (expectation-maximization algorithm). Смысл названия будет понятен немного позднее. Для начала введём обозначение для логарифмической функции правдоподобия: $L(\Theta) = \log p(\mathbf{X} \mid \Theta)$.

Пусть после n-й итерации алгоритма мы получили значение Θ_n. Так как мы хотим максимизировать $L(\Theta)$, то мы хотим получить значение Θ такое, что разность

$$L(\Theta) - L(\Theta_n) = \log p(\mathbf{X} \mid \Theta) - \log p(\mathbf{X} \mid \Theta_n).$$

была максимальной. Заметьте, что мы пока что никак не трогали латентные или утерянные переменные. Если они есть, то EM-алгоритм предлагает естественный метод их включения. Периодически латентные переменные можно ввести просто как «трюк» для упрощения нахождения оценки максимального правдоподобия. В этом случае предполагается, что знание скрытых переменных успрощает анализ функции правдоподобия.

Пусть $\mathbf{T}$ — вектор скрытых переменных. Тогда по формуле полной вероятности

$$p(\mathbf{X} \mid \Theta) = \sum_{\mathbf{T}} p(\mathbf{X}, \mathbf{T} \mid \Theta) = \sum_{\mathbf{T}} p(\mathbf{X} \mid \mathbf{T}, \Theta) p(\mathbf{T} \mid \Theta).$$

Тогда разность лог-функций правдоподобия можно записать в следующем виде:

$$L(\Theta) - L(\Theta_n) = \log \left(\sum_{\mathbf{T}} p(\mathbf{X} \mid \mathbf{T}, \Theta) p(\mathbf{T} \mid \Theta) \right) - \log p(\mathbf{X} \mid \Theta_n)$$

Мы вернулись к логарифму суммы. Для того, чтобы избавиться от него, воспользуемся *неравенством Йенсена*, которое доказывалось в курсе математического анализа.

Теорема 22. *Пусть f — выпуклая (вниз) функция, определённая на интервале (a, b). Тогда для любых $x_1, \ldots, x_n \in (a, b)$ и $\lambda_1, \ldots, \lambda_n > 0$ таких, что $\lambda_1 + \ldots + \lambda_n = 1$, выполнено следующее неравенство:*

$$f\left(\sum_{k=1}^{n} \lambda_k x_k \right) \leqslant \sum_{k=1}^{n} \lambda_k f(x_k).$$

Если функция f выпукла вверх, то знак неравенства меняется на противоположный.

Применим это неравенство следующим образом. Пусть $\lambda_i = p(\mathbf{T} \mid \mathbf{X}, \Theta_n)$. Несложно понять, что такие константы подходят. Далее, логарифм — выпуклая (вверх) функция, поэтому использование неравенства легально. Тогда «подгоним» выражение внутри логарифма под эти константы:

$$\begin{aligned}
L(\Theta) - L(\Theta_n) &= \log\left(\sum_{\mathbf{T}} p(\mathbf{X} \mid \mathbf{T}, \Theta) p(\mathbf{T} \mid \Theta)\right) - \log p(\mathbf{X} \mid \Theta_n) = \\
&= \log\left(\sum_{\mathbf{T}} p(\mathbf{X} \mid \mathbf{T}, \Theta) p(\mathbf{T} \mid \Theta) \frac{p(\mathbf{T} \mid \mathbf{X}, \Theta_n)}{p(\mathbf{T} \mid \mathbf{X}, \Theta_n)}\right) - \log p(\mathbf{X} \mid \Theta_n) = \\
&= \log\left(\sum_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta_n) \frac{p(\mathbf{X} \mid \mathbf{T}, \Theta) p(\mathbf{T} \mid \Theta)}{p(\mathbf{T} \mid \mathbf{X}, \Theta_n)}\right) - \log p(\mathbf{X} \mid \Theta_n) \geqslant \\
&\geqslant \sum_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta_n) \log\left(\frac{p(\mathbf{X} \mid \mathbf{T}, \Theta) p(\mathbf{T} \mid \Theta)}{p(\mathbf{T} \mid \mathbf{X}, \Theta_n)}\right) - \log p(\mathbf{X} \mid \Theta_n) = \\
&= \sum_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta_n) \left(\log\left(\frac{p(\mathbf{X} \mid \mathbf{T}, \Theta) p(\mathbf{T} \mid \Theta)}{p(\mathbf{T} \mid \mathbf{X}, \Theta_n)}\right) - \log p(\mathbf{X} \mid \Theta_n)\right) = \\
&= \sum_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta_n) \log\left(\frac{p(\mathbf{X} \mid \mathbf{T}, \Theta) p(\mathbf{T} \mid \Theta)}{p(\mathbf{X} \mid \Theta_n) p(\mathbf{T} \mid \mathbf{X}, \Theta_n)}\right) \equiv \Delta(\Theta \mid \Theta_n).
\end{aligned}$$

В результате получаем, что $L(\Theta) \geqslant L(\Theta_n) + \Delta(\Theta \mid \Theta_n)$. Для удобства введём функцию $l(\Theta \mid \Theta_n) \equiv L(\Theta_n) + \Delta(\Theta \mid \Theta_n)$. Тогда наше неравенство имеет очень простой вид: $L(\Theta) \geqslant l(\Theta \mid \Theta_n)$.

Теперь у нас есть функция $l(\Theta \mid \Theta_n)$, ограниченная сверху лог-функцией правдоподобия $L(\Theta)$. Заметим, что $l(\Theta_n \mid \Theta_n) = L(\Theta_n)$:

$$\begin{aligned}
l(\Theta_n) &= L(\Theta_n) + \sum_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta_n) \log\left(\frac{p(\mathbf{X} \mid \mathbf{T}, \Theta_n) p(\mathbf{T} \mid \Theta_n)}{p(\mathbf{T} \mid \mathbf{X}, \Theta_n) p(\mathbf{X} \mid \Theta_n)}\right) = \\
&= L(\Theta_n) + \sum_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta_n) \log\left(\frac{p(\mathbf{X}, \mathbf{T} \mid \Theta_n)}{p(\mathbf{X}, \mathbf{T} \mid \Theta_n)}\right) = L(\Theta_n).
\end{aligned}$$

Наша цель состоит в выборе значения Θ такого, что $L(\Theta)$ максимально. Мы показали, что функция $l(\Theta \mid \Theta_n)$ ограничена сверху $L(\Theta)$ и равенство достигается при текущей оценке, то есть при $\Theta = \Theta_n$. Следовательно, если Θ увеличивает $l(\Theta \mid \Theta_n)$, то она увеличивает и $L(\Theta)$. Для получения максимального прироста в значении $L(\Theta)$, EM-алгоритм ищет Θ такое, что $l(\Theta \mid \Theta_n)$ максимально. Это значение обозначают через Θ_{n+1}. Теперь посчитаем значение Θ_{n+1}:

$$\begin{aligned}
\Theta_{n+1} &\equiv \arg\max_{\Theta} l(\Theta \mid \Theta_n) = \\
&= \arg\max_{\Theta} \left[L(\Theta_n) + \sum_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta_n) \log\left(\frac{p(\mathbf{X} \mid \mathbf{T}, \Theta) p(\mathbf{T} \mid \Theta)}{p(\mathbf{X} \mid \Theta_n) p(\mathbf{T} \mid \mathbf{X}, \Theta_n)}\right)\right]
\end{aligned}$$

Выбросим все члены, не зависящие от Θ — они не повлияют на значение Θ_{n+1}:

$$\begin{aligned}
\Theta_n &= \arg\max_{\Theta} \left[\sum_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta_n) \log\left(p(\mathbf{X} \mid \mathbf{T}, \Theta) p(\mathbf{T} \mid \Theta)\right)\right] = \\
&= \arg\max_{\Theta} \left[\sum_{\mathbf{T}} p(\mathbf{T} \mid \mathbf{X}, \Theta_n) \log p(\mathbf{X}, \mathbf{T} \mid \Theta)\right] = \\
&= \arg\max_{\Theta} \left[\mathsf{E}_{\mathbf{T} \mid \mathbf{X}, \Theta_n} [\log p(\mathbf{X}, \mathbf{T} \mid \Theta)]\right].
\end{aligned}$$

В формуле появляется $p(\mathbf{T} \mid \mathbf{X}, \Theta_n)$. Возникает вопрос: а как это посчитать? Для этого воспользуемся формулой Байеса:

$$p(\mathbf{T} \mid \mathbf{X}, \Theta_n) = \frac{p(\mathbf{X}, \mathbf{T} \mid \Theta_n)}{p(\mathbf{X} \mid \Theta_n)} = \frac{p(\mathbf{X}, \mathbf{T} \mid \Theta_n)}{\sum_{\mathbf{T}} p(\mathbf{X}, \mathbf{T} \mid \Theta_n)}.$$

Отсюда и получается название: мы *максимизируем* (M) условное математическое *ожидание* (E). Теперь можно записать алгоритм формально:

Алгоритм 3 EM-алгоритм

Вход: Набор наблюдаемых переменных $\mathbf{X}$
Выход: Набор параметров Θ_{ML}.
1: Инициализируем Θ_1 каким-либо образом.
2: $n \leftarrow 1$.
3: **while** алгоритм не сошёлся **do**
4: E-шаг: считаем $\mathsf{E}_{\mathbf{T}|\mathbf{X},\Theta_n}[\log p(\mathbf{X}, \mathbf{T} \mid \Theta)]$
5: M-шаг: $\Theta_{n+1} \leftarrow \arg\max_{\Theta} \left[\mathsf{E}_{\mathbf{T}|\mathbf{X},\Theta_n}[\log p(\mathbf{X}, \mathbf{T} \mid \Theta)]\right]$
6: $n \leftarrow n + 1$

По сути, этот алгоритм есть ни что иное, как покоординатный спуск: на каждой итерации последовательно уточняются возможные значения $\mathbf{T}$ (E-шаг), после чего пересчитывается значение Θ (M-шаг).

После всех логических изысканий можно задать вопрос: а зачем всё это? Почему бы не максимизировать $L(\Theta)$ сразу, а не ворочаться с $l(\Theta \mid \Theta_n)$? Ответ достаточно прост: $l(\Theta \mid \Theta_n)$ учитывает скрытые переменные T. В случае, когда мы хотим их определить, EM-алгоритмы предоставляют необходимые для этого инструменты. Более того, как мы говорили ранее, введение скрытых переменных может быть весьма удобным в том плане, что оптимизация $l(\Theta \mid \Theta_n)$ будет гораздо проще прямой оптимизации $L(\Theta)$. Например, во многих случаях на M-шаге можно получить явные формулы, так как происходит оптимизация выпуклой комбинации логарифмов полных правдоподобий.

Рассмотрим пример применения EM-алгоритма на задаче *разделения гауссовской смеси*. Она появляется при попытке приблизить плохо параметризуемые распределения смесью гауссиан. Пусть

$$\mathbf{X} \text{ — выборка размера } n \text{ из смеси } \sum_{k=1}^{l} w_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in \mathbb{R}^d, \quad \sum_{k=1}^{l} w_k = 1.$$

Задача следующая: нужно восстановить плотность генеральной совокупности, то есть определить $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ и w_k. Попробуем применить EM-алгоритм. Для начала каким-либо образом инициализируем эти параметры, соблюдая ограничения, наложенные на них.

E-шаг будет устроен следующим образом: для начала определим значение $p(\mathbf{T} \mid \mathbf{X}, \Theta)$. Для этого введём скрытые переменные $\mathbf{z}_k \in \{0, 1\}^l$, $z_{k1} + \ldots + z_{kl} = 1$. Они будут определять, к какой компоненте принадлежит $\mathbf{x}_k$. Тогда

$$p(\mathbf{T} \mid \mathbf{X}, \Theta) = \gamma(z_{ij}) = \frac{w_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{k=1}^{l} w_k \mathcal{N}(\mathbf{x}_k \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}.$$

Далее, выпишем $\mathsf{E}_{\mathbf{T}|\mathbf{X},\Theta_n}[\log p(\mathbf{X},\mathbf{T}\mid\Theta)]$:

$$\begin{aligned}\mathsf{E}_{\mathbf{T}|\mathbf{X},\Theta_n}[\log p(\mathbf{X},\mathbf{T}\mid\Theta)] &= \sum_{\mathbf{T}} p(\mathbf{T}\mid\mathbf{X},\Theta_n)\log p(\mathbf{X},\mathbf{T}\mid\Theta) = \\ &= \sum_{i=1}^{n}\sum_{j=1}^{l}\gamma(z_{ij})\log(w_j\mathcal{N}(\mathbf{x}_i\mid\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)) = \\ &= \sum_{i=1}^{n}\sum_{j=1}^{l}\gamma(z_{ij})\left(\log w_j + \log\mathcal{N}(\mathbf{x}_i\mid\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\right)\end{aligned}$$

Теперь начинаем оптимизировать. Покажем вывод формулы для w_i. Воспользуемся методом Лагранжа и выпишем лагранжиан:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n}\sum_{j=1}^{l}\gamma(z_{ij})\log w_j + \lambda\left(\sum_{k=1}^{l} w_k - 1\right).$$

Далее, дифференцируем его по w_j и получаем оптимальное значение:

$$\sum_{i=1}^{n}\frac{\gamma(z_{ij})}{w_j} + \lambda = 0 \implies w_j = -\frac{1}{\lambda}\sum_{i=1}^{n}\gamma(z_{ij})$$

Теперь воспользуемся свойствами w_j и $\gamma(z_{ij})$:

$$\sum_{j=1}^{l} w_j = -\frac{1}{\lambda}\sum_{i=1}^{n}\sum_{j=1}^{l}\gamma(z_{ij}) = -\frac{n}{\lambda} = 1 \implies \lambda = -n \implies w_j = \frac{1}{n}\sum_{i=1}^{n}\gamma(z_{ij}).$$

Без доказательства выпишем формулы для $\boldsymbol{\mu}_k$ и $\boldsymbol{\Sigma}_k$:

$$\begin{aligned}N_j &= \sum_{i=1}^{n}\gamma(z_{ij}),\\ \boldsymbol{\mu}_j &= \frac{1}{N_j}\sum_{i=1}^{n}\gamma(z_{ij})\boldsymbol{x}_i,\\ \boldsymbol{\Sigma}_j &= \frac{1}{N_j}\sum_{i=1}^{n}\gamma(z_{ij})(\boldsymbol{x}_i-\boldsymbol{\mu}_j)(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^{\mathsf{T}}.\end{aligned}$$

Теперь возвращаемся к E-шагу и повторяем процедуру до тех пор, пока алгоритм не сойдётся.

У EM-алгоритма много плюсов, но есть и недостатки. Например, он чувствителен к начальным приближениям — в зависимости от их значений он может сходиться к разным точкам. Далее, он находит локальный экстремум. Другими словами, алгоритм может сойтись к точке, крайне далёкой от глобального максимума — к седловой точке или, более того, к локальному минимуму. Впрочем, такое происходит не слишком часто. Стоит заметить, что в задаче разделения смеси EM-алгоритм не может определить количество компонентов смеси l — оно является *структурным параметром*.

1.14.5 Сегментация временных рядов

Задачи о сегментации временных рядов вполне распространены на практике и с ними связаны многие методы решения: например, расчёт агрегированных показателей, разбиение на равные сегменты и принцип скользящего окна. Однако они не универсальны.

Иногда на практике попадаются временные ряды, устроенные следующим образом: они весьма естественно разбиваются на достаточно однородные сегменты. Внутри них происходят колебания вокруг какого-то среднего, а после этого происходит резкий скачок и сигнал стабилизируется на другом уровне. Сами же скачки происходят неравномерно. Такая структура не позволяет эффективно использовать классические методы решения. Но при исследовании таких рядов возникла задача применимого на практике алгоритма, обеспечивающего выделение достаточно однородных фрагментов, удобных для последующей обработки.

Вообще, алгоритмы сегментации можно разделить на две большие группы: последовательные (on-line) и апостериорные (off-line). Но большинство апостериорных алгоритмов либо разбивает на два сегмента, либо использует итерационные вычислительные схемы, где одна итерация имеет сложность порядка $O(N^2)$, где N — длина известной реализации временного ряда. Было необходимо разработать апостериорный алгоритм, который, с одной стороны, мог разбивать временной ряд на произвольное число сегментов, а, с другой стороны, имел адекватную временную сложность.

Бурнаев и Меньшиков разработали такой апостериорный алгоритм, основанный на скрытых марковских моделях. В нём разбиение получается с помощью метода максимального правдоподобия. Такое разбиение является «оптимальным» в том смысле, что с точки зрения теории вероятностей ему соответствует локальный максимум функции правдоподобия, а с чисто вычислительной стороны оно минимизирует среднеквадратичное отклонение ряда от его средних значений, посчитанным по соответствующим сегментам однородности.

Ближе к делу. Переформулируем задачу сегментации временного ряда в терминах задачи максимизации функции правдоподобия некоторой скрытой марковской модели. Пусть $\mathbf{X}^N = (x_1, \ldots, x_N)$ — известная реализация временного ряда. Будем говорить, что $\mathbf{X}^N$ является реализацией наблюдаемой части $\mathbf{X} = (X_n)_{n\geqslant 1}$ скрытой марковской модели $(\mathbf{S}, \mathbf{X}) = (S_n, X_n)_{n\geqslant 1}$, где $\mathbf{S} = (S_n)_{n\geqslant 1}$ — это марковская цепь с M состояниями, то есть $S_i \in \{1, 2, \ldots, M\}$ для всех i. По умолчанию будем считать, что начальное состояние марковской цепи $S_0 = 1$ почти наверное. Далее, введём переходную матрицу $\mathbf{P}^M = (p_{ij}) \in \mathrm{M}_M(\mathbb{R})$, для которой $p_{ij} = 0$ при $|i - j| > 1$ и $p_{MM} = 1$. Тогда по теореме 12

$$\mathsf{P}(S_1 = s_1, \ldots, S_N = s_N) = \prod_{k=1}^{N} p_{s_{k-1}s_k}, \quad s_0 = 1.$$

Параметрами марковской цепи $\mathbf{S}$ являются количество состояний M и переходная матрица $\mathbf{P}^M$.

Теперь посмотрим на $\mathbf{X}$. Будем считать, что это последовательность условно независимых (при фиксированных значениях процесса $\mathbf{S}$) случайных величин с распределением $\mathcal{N}(\mu_{\mathbf{S}}, \sigma^2)$, то есть

$$\mathsf{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n \mid S_1 = s_1, \ldots, S_n = s_n) = \\ = \frac{1}{(2\pi\sigma^2)^{n/2}} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} \exp\left\{ -\sum_{k=1}^{n} \frac{(y_k - \mu_{s_k})^2}{2\sigma^2} \right\} dy_1 \ldots dy_n.$$

Таким образом, параметрами процесса $(\mathbf{S}, \mathbf{X})$ являются вектор средних значений $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$, дисперсия σ^2 и матрица переходных вероятностей $\mathbf{P}^M$.

Введём функцию $J(\mathbf{z})$, равную количеству компонент с разными значениями в произвольном векторе $\mathbf{z} \in \mathbb{R}^N$. Далее, пусть $\mathbf{S}_N = (s_1, \ldots, s_N)$ — наблюдаемая реализация значений марковской цепи, соответсвующая реализации наблюдаемой части $\mathbf{X}^N =$

$(x_1, \ldots, x_N)$ скрытой марковской модели. Теперь введём обозначение $M_S = J(\mathbf{S}^N)$. Вектором координат сегментов однородности временного ряда будем называть вектор $\mathbf{n}_S = (n_0, n_1, \ldots, n_{M_S})$, $n_0 = 0$, $n_{M_s} = N$, для координат которого выполнено неравенство $s_{n_i} \neq s_{n_i+1}$.

Достаточно очевидно, что $M_S \leqslant M$ и каждому сегменту однородности $[n_{i-1}, n_i]$ соответствует состояние i марковской цепи $\mathbf{S}$ и параметр μ_i распределений процесса $\mathbf{X}$. Таким образом, оценив по реализации $\mathbf{X}^n$ процесса $\mathbf{X}$ реализацию $\mathbf{S}_N$ процесса $\mathbf{S}$, можно получить оценку вектора $\mathbf{n}_S$ и, тем самым, сегментировать сигнал на участки однородности (в смысле постоянства среднего значения).

Будем оценивать $\mathbf{S}_N$ с помощью максимизирования правдоподобия при фиксированном $\mathbf{X}^N$. Обозначим условную функцию правдоподобия $\mathbf{S}_N$ через $\mathcal{L}(\mathbf{S}_N \mid \mathbf{X}^N; M, \mathbf{P}^M, \boldsymbol{\mu}, \sigma)$. Тогда

$$\mathbf{S}_N^{\text{opt}} = \arg \max_{\mathbf{S}_N \in \{1,\ldots,M\}^N} \mathcal{L}(\mathbf{S}_N \mid \mathbf{X}^N; M, \mathbf{P}^M, \boldsymbol{\mu}, \sigma).$$

С помощью формулы Байеса можно получить связь между $\mathcal{L}(\mathbf{S}_N \mid \mathbf{X}^N; M, \mathbf{P}^M, \boldsymbol{\mu}, \sigma)$ и $\mathcal{L}(\mathbf{S}_N, \mathbf{X}^N; M, \mathbf{P}^M, \boldsymbol{\mu}, \sigma)$, то есть между условной и совместной функциями правдоподобия:

$$\mathcal{L}(\mathbf{S}_N \mid \mathbf{X}^N; M, \mathbf{P}^M, \boldsymbol{\mu}, \sigma) = \frac{\mathcal{L}(\mathbf{S}_N, \mathbf{X}^N; M, \mathbf{P}^M, \boldsymbol{\mu}, \sigma)}{\mathcal{L}(\mathbf{X}^N; M, \mathbf{P}^M, \boldsymbol{\mu}, \sigma)}.$$

Как известно, $\mathcal{L}(\mathbf{X}^N; M, \mathbf{P}^M, \boldsymbol{\mu}, \sigma)$ есть ни что иное, как безусловная плотность распределения случайного вектора $\mathbf{X}^N$. Следовательно, она не зависит от $\mathbf{S}^N$ и задача равносильна максимизации $\mathcal{L}(\mathbf{S}_N, \mathbf{X}^N; M, \mathbf{P}^M, \boldsymbol{\mu}, \sigma)$ по $\mathbf{S}^N$, где

$$\mathcal{L}(\mathbf{S}_N, \mathbf{X}^N; M, \mathbf{P}^M, \boldsymbol{\mu}, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{k=1}^{N} p_{s_{k-1}s_k} \exp\left\{-\frac{(x_k - \mu_{s_k})^2}{2\sigma^2}\right\}.$$

Приступим к описанию алгоритма. Но для начала сделаем одно уточнение. Будем считать, что на матрицу $\mathbf{P}^M$ наложены следующие условия: $p_{ii} = p$, $p_{i,i+1} = 1 - p$ для какого-то $p \in (0, 1)$ при $i \in \{1, 2, \ldots, M-1\}$ и $p_{MM} = 1$. В дальнейшем в аргументах функции правдоподобия вместо матрицы $\mathbf{P}^M$ будем писать параметр p.

Теперь приступаем к описанию. Начнём с входных и выходных данных.

- На вход алгоритму подаются $\mathbf{X}^N = (x_1, \ldots, x_N)$ — известная реализация ряда, $M \leqslant N/2$ — оценка сверху количества сегментов и ε — пороговое значение.
- Результатом работы алгоритма являются оценки параметров p^{opt},$\boldsymbol{\mu}^{\text{opt}}$, σ^{opt}, $\mathbf{S}_N^{\text{opt}}$, $\mathbf{n}_S^{\text{opt}}$ и $M_S^{\text{opt}} = J(\mathbf{S}_N^{\text{opt}})$. При этом вполне себе возможна ситуация, когда $M_S^{\text{opt}} < M$.

Инициализация алгоритма состоит в следующем:

- Начнём с инициализации параметра p: $p^{(0)} = (N - M)/N$.
- Далее, $\mathbf{S}_N^{(0)} = (s_1^{(0)}, \ldots, s_N^{(0)})$ генерируется случайным образом из чисел множества $\{1, 2, \ldots, M\}$ с соблюдением следующих правил:

$$J(\mathbf{S}_N^{(0)}) = M, \quad s_1^{(0)} \leqslant s_2^{(0)} \leqslant \ldots s_N^{(0)}.$$

- Сразу же инициализируем оптимальную оценку стандартного отклонения:

$$\sigma^{\text{opt}} = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(x_k - \overline{x})^2}, \text{ где } \overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_k.$$

Теперь оговорим условие остановки алгоритма. Пусть m — номер текущей итерации. Алгоритм остановит работу в том случае, если изменение функции правдоподобия стало достаточно малым:

$$|\mathcal{L}(\mathbf{S}_N^{(m)}, \mathbf{X}^N; M_S^{(m)}, p^{(m)}, \boldsymbol{\mu}^{(m)}, \sigma^{\text{opt}}) - \mathcal{L}(\mathbf{S}_N^{(m-1)}, \mathbf{X}^N; M_S^{(m-1)}, p^{(m-1)}, \boldsymbol{\mu}^{(m-1)}, \sigma^{\text{opt}})| < \varepsilon.$$

Теперь опишем, что происходит на m-й итерации алгоритма:

1. По значению $\mathbf{S}_N^{(m)}$ строится вектор $\mathbf{n}_S^{(m)}$ и определяется $M_S^{(m)} = J(\mathbf{S}_N^{(m)})$.

2. По вектору $\mathbf{n}_S^{(m)}$ строится вектор матожиданий $\boldsymbol{\mu}^{(m)} \in \mathbb{R}^{M_S^{(m)}}$ по следующему правилу:
$$\mu_i^{(m)} = \frac{1}{n_i - n_{i-1}} \sum_{k=n_{i-1}+1}^{n_i} x_k, \quad i \in \{1, 2, \ldots, M_S^{(m)}\}.$$

3. Ищется вероятность $p^{(m)} = (N - M_S^{(m)})/N$.

4. Оценивается значение функции правдоподобия $\mathcal{L}(\mathbf{S}_N^{(m)}, \mathbf{X}^N; M^{(m)}, p^{(m)}, \boldsymbol{\mu}^{(m)}, \sigma^{\text{opt}})$ и проверяется, выполнено ли условие остановки алгоритма. Если оно подтверждено, то мы получили ответ: $p^{\text{opt}} = p^{(m)}$,$\boldsymbol{\mu}^{\text{opt}} = \boldsymbol{\mu}^{(m)}$, $\mathbf{S}_N^{\text{opt}} = \mathbf{S}_N^{(m)}$, $\mathbf{n}_S^{\text{opt}} = \mathbf{n}_S^{(m)}$ и $M_S^{\text{opt}} = M_S^{(m)} = J(\mathbf{S}_N^{(m)})$. Иначе же переходим к следующему шагу.

5. С помощью алгоритма Витерби инициализируем $\mathbf{S}_N^{(m+1)}$:
$$\mathbf{S}_N^{(m+1)} = \arg \max_{\mathbf{S}_N \in \{1,\ldots,M_S^{(m)}\}^N} \mathcal{L}(\mathbf{S}_N, \mathbf{X}^N; M_S^{(m)}, p^{(m)}, \boldsymbol{\mu}^{(m)}, \sigma^{\text{opt}}).$$

После этого m увеличивается на единицу и запускается следующая итерация.

Описанный выше алгоритм, по сути, можно отнести к классу EM-алгоритмов, так как сначала алгоритм сводится к подсчёту функции правдоподобия и её максимизации. На практике обычно смотрят не на саму функцию правдоподобия, а на её логарифм.

Теперь докажем одно простое утверждение, которое доказывает корректность алгоритма.

Теорема 23. *Описанный выше алгоритм сходится.*

Доказательство. Введём следующую функцию:

$$F(\mathbf{S}_N, \mathbf{X}^N; M, p, \boldsymbol{\mu}, \sigma) \equiv -\ln \mathcal{L}(\mathbf{S}_N, \mathbf{X}^N; M, p, \boldsymbol{\mu}, \sigma).$$

Несложно получить, что

$$F(\mathbf{S}_N, \mathbf{X}^N; M, p, \boldsymbol{\mu}, \sigma) = \sum_{k=1}^{N} \frac{(x_k - \mu_{s_k})^2}{2\sigma^2} - \sum_{k=1}^{N} \ln p_{s_{k-1}s_k} + N \ln(\sigma\sqrt{2\pi}).$$

Рассмотрим сумму логарифмов отдельно. Мы можем сказать (по построению $\mathbf{S}_N$), что $p_{s_{k-1}s_k}$ равно либо p, либо $1-p$. Если $p_{s_{k-1}s_k} = 1-p$, то $s_k = s_{k-1} + 1$. Если же $p_{s_{k-1}s_k} = p$, то $s_k = s_{k-1}$. Следовательно,

$$\sum_{k=1}^{N} \ln p_{s_{k-1}s_k} = J(\mathbf{S}_N) \ln(1-p) + (N - J(\mathbf{S}_N)) \ln p = J(\mathbf{S}_N) \ln \frac{1-p}{p} + N \ln p.$$

Отсюда мы получаем, что верно следующее представление:

$$F(\mathbf{S}_N, \mathbf{X}^N; M, p, \boldsymbol{\mu}, \sigma) = \frac{1}{2\sigma^2} F_1(\mathbf{S}_N, \mathbf{X}^N; M, \boldsymbol{\mu}) + F_2(\mathbf{S}_N; M, p) + N \ln(\sigma\sqrt{2\pi}),$$

где функции F_1 и F_2 опредеяются следующим образом:

$$F_1(\mathbf{S}_N, \mathbf{X}^N; M, \boldsymbol{\mu}) = \sum_{k=1}^{N} (x_k - \mu_{s_k})^2, \quad F_2(\mathbf{S}_N; M, p) = J(\mathbf{S}_N) \ln \frac{p}{1-p} - N \ln p.$$

Теперь заметим, что выполнены три неравенства:

- Для всех $\boldsymbol{\mu} \in \mathbb{R}^{M_S^{(m)}}$, где $M_S^{(m)} = J(\mathbf{S}_N^{(m)})$, выполнено следующее неравенство:

$$F_1(\mathbf{S}_N^{(m)}, \mathbf{X}^N; M_S^{(m)}, \boldsymbol{\mu}) \geqslant F_1(\mathbf{S}_N^{(m)}, \mathbf{X}^N; M_S^{(m)}, \boldsymbol{\mu}^{(m)}).$$

- Для всех $p \in (0, 1)$ выполнено, что

$$F_2(\mathbf{S}_N^{(m)}; M_S^{(m)}, p) \geqslant F_2(\mathbf{S}_N^{(m)}; M_S^{(m)}, p^{(m)}).$$

  Действительно, возьмём производную по p и приравняем её к нулю:

$$J(\mathbf{S}_N^{(m)}) \frac{1-p}{p} \frac{1-p+p}{(1-p)^2} - \frac{N}{p} = 0.$$

  Отсюда получаем, что

$$\frac{J(\mathbf{S}_N^{(m)})}{1-p} = N \implies p = 1 - \frac{J(\mathbf{S}_N^{(m)})}{N} = \frac{N - J(\mathbf{S}_N^{(m)})}{N} \equiv p^{(m)}.$$

  Так как при увеличении p знак производной меняется с отрицательного на положительный, то это действительно минимум.

- Так как алгоритм Витерби позволяет получить глобальный максимум функции правдоподобия, то для любого $\mathbf{S}_N \in \{1, 2, \ldots, M_S^{(m)}\}^N$

$$F(\mathbf{S}_N, \mathbf{X}^N; M_S^{(m)}, p, \boldsymbol{\mu}^{(m)}, \sigma) \geqslant F(\mathbf{S}_N^{(m)}, \mathbf{X}^N; M_S^{(m)}, p, \boldsymbol{\mu}^{(m)}, \sigma).$$

Отсюда получаем следующую цепочку равенств:

$$F(\mathbf{S}_N^{(m-1)}, \mathbf{X}^N; M_S^{(m-1)}, p^{(m-1)}, \boldsymbol{\mu}^{(m-1)}, \sigma) \geqslant F(\mathbf{S}_N^{(m)}, \mathbf{X}^N; M_S^{(m-1)}, p^{(m-1)}, \boldsymbol{\mu}^{(m-1)}, \sigma) \geqslant$$
$$\geqslant F(\mathbf{S}_N^{(m)}, \mathbf{X}^N; M_S^{(m)}, p^{(m)}, \boldsymbol{\mu}^{(m)}, \sigma) \geqslant F(\mathbf{S}_N^{(m+1)}, \mathbf{X}^N; M_S^{(m)}, p^{(m)}, \boldsymbol{\mu}^{(m)}, \sigma).$$

Отсюда получаем, что каждая итерация уменьшает $F(\mathbf{S}_N^{(m)}, \mathbf{X}^N; M_S^{(m)}, p^{(m)}, \boldsymbol{\mu}^{(m)}, \sigma)$. Так как это значение ограничено снизу нулём, то существует предел и алгоритм сходится. □

Теперь выпишем несколько свойств этого алгоритма.

- Каждая итерация имеет сложность порядка $O(NM^2)$ (такая оценка возникает из-за алгоритма Витерби). Эксперименты показывают, что он сходится за 10-15 итераций, что характерно для алгоритмов типа EM.

- Как мы показали выше, минимум $F_2(\mathbf{S}_N; M, p)$ достигается при $p^{\text{opt}} = (N - J(\mathbf{S}_N))/N$ и равен

$$F_2(\mathbf{S}_N; M, p^{(m)}) = J(\mathbf{S}_N) \ln\left(\frac{N - J(\mathbf{S}_N)}{J(\mathbf{S}_N)}\right) - N \ln\left(\frac{N - J(\mathbf{S}_N)}{N}\right) = \\ = J(\mathbf{S}_N) \ln\left(\frac{N}{J(\mathbf{S}_N)} - 1\right) - N \ln\left(1 - \frac{J(\mathbf{S}_N)}{N}\right).$$

  Оказывается, что $F_2(\mathbf{S}_N; M, p^{\text{opt}})$ отвечает за регуляризацию, поскольку при увеличении $J(\mathbf{S}_N)$ это слагаемое возрастает, в то время как $F_1(\mathbf{S}_N, \mathbf{X}^N; M, \boldsymbol{\mu})$ убывает (и наоборот). Следовательно, при использовании алгоритма Витерби при минимизации $F(\mathbf{S}_N, \mathbf{X}^N; M, p, \boldsymbol{\mu}, \sigma)$ по $\mathbf{S}_N \in \{1, 2, \ldots, M\}^N$ при фиксированных значениях M и $\boldsymbol{\mu}$ будет так же выбрано оптимальное число сегментов разбиения, равное $M_S = J(\mathbf{S}_N) \leqslant M$. Вообще говоря, $M \geqslant M_S^{(0)} \geqslant M_S^{(1)} \geqslant \ldots$, причём процесс понижения значения $J(\mathbf{S}_N)$ со временем стабилизируется, так как иначе значение F начнёт возрастать.

- Далее, после стабилизации процесса понижения значения $J(\mathbf{S}_N)$, значение F понижается дальше за счёт «подстройки» значений $\mathbf{n}_S$ (и, следовательно, значений, $\boldsymbol{\mu}$). Описанный нами алгоритм, по сути, ищет такое разбиение, при котором приближение временного ряда его средними значениями по соответствующим сегментам будет оптимальным в смысле среднеквадратичного отклонения.

- Практика показала следующее: результат работы алгоритма слабо зависит от $p^{(0)}$ и $\mathbf{S}_N^{(0)}$, но сильно зависит от M. Если взять $M \ll N$, то с большой вероятностью $M_S^{\text{opt}} = M$. В противном случае $M_S^{\text{opt}} < M$. Более того, если запустить алгоритм с двумя разными параметрами M_1 и M_2, то полученные значения $M_S^{\text{opt},1}$ и $M_S^{\text{opt},2}$ не обязательно равны. Это вытекает из того факта, что оптимизируемая функция многоэкстремальна.

1.15 Фильтр Калмана

Перейдём к статистике. Для начала введём понятие *оптимальной в среднеквадратическом смысле* (MMSE — Minimum Mean Square Error) оценки.

Определение 62. *Среднеквадратическая ошибка* оценки $\widehat{\boldsymbol{\xi}}$ случайного вектора $\boldsymbol{\xi}$ равна

$$\text{MSE} = \operatorname{tr} \mathsf{E}\left[(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}})(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}})^\top\right] = \mathsf{E}\left[(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}})^\top(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}})\right].$$

Определение 63. Будем называть $\widehat{\boldsymbol{\xi}}$ *MMSE-оценкой* случайного вектора $\boldsymbol{\xi}$, если для неё среднеквадратическая ошибка минимальна.

Теперь приступим к так называемым фильтрам Калмана для линейных систем с дискретным временем. Его вывод будет основан на двух предположениях:

- В гауссовском случае фильтр Калмана является оптимальной оценкой состояния в среднеквадратичном смысле (MMSE state estimator).

- В остальных случаях фильтр Калмана является оптимальной *линейной* оценкой состояния в среднеквадратничном смысле (LMMSE state estimator).

Также далее мы опишем детерменистическую модель (метод наименьших квадратов). Но начнём с описания базового пространства состояний.

1.15.1 Стохастическое пространство состояний

И сразу же дадим определение.

Определение 64. Линейное (изменяющееся во времени) пространство состояний с дискретным временем задаётся парой уравнений:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{F}_k\mathbf{x}_k + \mathbf{G}_k\mathbf{w}_k && \text{(уравнение эволюции системы)} \\ \mathbf{z}_{k+1} &= \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k && \text{(уравнение измерения)} \end{aligned}$$

где

- $\mathbf{F}_k \in \mathrm{Mat}_{n\times n}(\mathbb{R})$, $\mathbf{G}_k \in \mathrm{Mat}_{n\times n_w}(\mathbb{R})$, $\mathbf{H}_k \in \mathrm{Mat}_{m\times n}(\mathbb{R})$ — известные матрицы,
- $\mathbf{x}_k \in \mathbb{R}^n$ — *вектор состояния*,
- $\mathbf{w}_k \in \mathbb{R}^{n_w}$ — *шум состояния*,
- $\mathbf{z}_k \in \mathbb{R}^m$ — *вектор наблюдений*,
- $\mathbf{v}_k \in \mathbb{R}^m$ — *шум наблюдений*,

Начальным условием для такой системы является $\mathbf{x}_0$, которая обычно считается за случайную величину.

Теперь докажем одно свойство:

Свойство 1. *Если $\{\mathbf{w}_n\}_{n\in\mathbb{N}}$ является последовательностью независимых случайных величин и она не зависит от $\mathbf{x}_0$, то процесс $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ является марковским.*

Доказательство. Раскрыв формулу для $\mathbf{x}_{k+1}$, получим, что это линейная функция от независимых случайных величин $\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_k$. Из этого следует, что $\mathbf{w}_k$ не зависит от $\mathbf{x}_i$ при $i \leqslant k$. Отсюда следует, что $\mathbf{x}_{k+1}$ разбивается на две части: на $\mathbf{F}_k\mathbf{x}_k$, которая однозначно задаётся значением $\mathbf{x}_k$, и на $\mathbf{G}_k\mathbf{w}_k$, которая не зависит от $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k$. Следовательно,

$$\mathsf{P}(\mathbf{x}_{k+1} = \mathbf{a}_{k+1} \mid \mathbf{x}_0 = \mathbf{a}_0, \ldots, \mathbf{x}_k = \mathbf{a}_k) = \mathsf{P}(\mathbf{x}_{k+1} = \mathbf{a}_{k+1} \mid \mathbf{x}_k = \mathbf{a}_k). \qquad \square$$

Примечание. На самом деле тут не обязательна линейность — достаточно, что $\mathbf{x}_{k+1}$ есть функция от $\mathbf{x}_k$ и $\mathbf{w}_k$.

Примечание. Процесс $(\mathbf{z}_n)_{n\in\mathbb{N}}$ обычно не является марковским.

Следствие. *В принципе, плотность распределения вектора $\mathbf{x}_{k+1}$ можно считать с помощью аналога обобщённого уравнения Маркова:*

$$p_{\mathbf{x}_{k+1}}(\mathbf{y}_{k+1}) = \int_{\mathbb{R}^n} \cdots \int p_{\mathbf{x}_{k+1}|\mathbf{x}_k}(\mathbf{y}_{k+1} \mid \mathbf{y}_k) p_{\mathbf{x}_k}(\mathbf{y}_k)\, d\mathbf{y}_k,$$

где $p_{\mathbf{x}_{k+1}|\mathbf{x}_k}(\mathbf{y}_{k+1} \mid \mathbf{y}_k)$ задаётся с помощью $p_{\mathbf{w}_k}(\mathbf{y}_{k+1})$.

Теперь рассмотрим так называемые *гауссовские пространства событий*. Они характеризуются тем, что последовательности шумовые последовательности $\{\mathbf{w}_n\}$, $\{\mathbf{v}_n\}$ и начальное условие $\mathbf{x}_0$ образуют гауссовскую последовательность, то есть имеют совместное нормальное распределение. Отсюда сразу же получаем, что тогда процессы $\{\mathbf{x}_n\}$ и $\{\mathbf{z}_n\}$ тоже являются гауссовскими (как линейное преобразование). Если же выполнено свойство марковости, то такие пространства называют *моделями Гаусса-Маркова*.

Требование независимости весьма сильное, поэтому иногда его ослабляют следующим образом:

- Считают, что $\{\mathbf{w}_n\}$ — это белый шум в широком смысле, то есть $\mathsf{E}[\mathbf{w}_k] = \mathbf{0}$ и $\mathsf{cov}(\mathbf{w}_k, \mathbf{w}_l) = \mathbf{Q}_k\delta_{kl}$, где δ_{kl} — дельта Кронекера.

- Аналогичное предположение делается относительно $\{\mathbf{v}_n\}$: $\mathsf{E}[\mathbf{v}_k] = \mathbf{0}$ и $\mathsf{cov}(\mathbf{v}_k, \mathbf{v}_l) = \mathbf{R}_k\delta_{kl}$.

- Шумы некоррелированы: $\mathsf{cov}(\mathbf{w}_k, \mathbf{v}_l) = \mathbf{0}$.

- $\mathbf{x}_0$ некоррелирован с шумовыми последовательностями. При этом про $\mathbf{x}_0$ известно, что $\mathsf{D}[\mathbf{x}_0] = \mathbf{P}_0$.

Если выполнены эти условия, то пространство событий называют *стандартным пространством второго порядка*. Иногда бывает полезно позволять корреляцию между шумами:

$$\mathsf{cov}(\mathbf{w}_k, \mathbf{v}_l) = \mathbf{S}_k\delta_{kl}.$$

В таком случае пространство событий называют *пространством второго порядка с коррелированным шумом*.

Всё вышесказанное можно записать одним уравнением:

$$\mathsf{cov}\left(\begin{pmatrix}\mathbf{w}_k\\ \mathbf{v}_k\\ \mathbf{x}_0\end{pmatrix}, \begin{pmatrix}\mathbf{w}_l\\ \mathbf{v}_l\\ \mathbf{x}_0\end{pmatrix}\right) = \begin{pmatrix}\mathbf{Q}_k\delta_{kl} & \mathbf{S}_k\delta_{kl} & \mathbf{0}\\ \mathbf{S}_k^\top\delta_{kl} & \mathbf{R}_k\delta_{kl} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{P}_0\end{pmatrix}$$

Стоит заметить, что модели Гаусса-Маркова являются частным случаем моделей второго порядка с коррелированным шумом.

Для стандартной нормальной модели несложно рекурсивно задать матожидание и дисперсию $\mathbf{x}_k$. Действительно, по линейности матожидания

$$\mathsf{E}[\mathbf{x}_{k+1}] = \mathbf{F}_k\,\mathsf{E}[\mathbf{x}_k] + \mathbf{G}_k\,\mathsf{E}[\mathbf{w}_k] = \mathbf{F}_k\,\mathsf{E}[\mathbf{x}_k].$$

Далее, введём обозначение $\mathbf{P}_k = \mathsf{D}[\mathbf{x}_k] = \mathsf{E}\big[(\mathbf{x}_k - \mathsf{E}[\mathbf{x}_k])(\mathbf{x}_k - \mathsf{E}[\mathbf{x}_k])^\top\big]$. Заметим, что $\mathbf{x}_{k+1} - \mathsf{E}[\mathbf{x}_{k+1}] = \mathbf{F}_k(\mathbf{x}_k - \mathsf{E}[\mathbf{x}_k]) + \mathbf{G}_k\mathbf{w}_k$. Далее, вспомним, что $\mathbf{x}_k$ есть линейная функция от $\mathbf{x}_0, \mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{k-1}$. Отсюда следует, что $\mathbf{w}_k$ некоррелированно с $\mathbf{x}_k$, а, следовательно, и с $\mathbf{x}_{k+1} - \mathsf{E}[\mathbf{x}_{k+1}]$. Следовательно,

$$\begin{aligned}\mathbf{P}_{k+1} &= \mathsf{E}\big[(\mathbf{F}_k(\mathbf{x}_k - \mathsf{E}[\mathbf{x}_k]) + \mathbf{G}_k\mathbf{w}_k)(\mathbf{F}_k(\mathbf{x}_k - \mathsf{E}[\mathbf{x}_k]) + \mathbf{G}_k\mathbf{w}_k)^\top\big] =\\ &= \mathsf{E}\big[\mathbf{F}_k(\mathbf{x}_k - \mathsf{E}[\mathbf{x}_k])(\mathbf{x}_k - \mathsf{E}[\mathbf{x}_k])^\top\mathbf{F}_k^\top\big] + \mathsf{E}\big[\mathbf{G}_k\mathbf{w}_k\mathbf{w}_k^\top\mathbf{G}_k^\top\big] =\\ &= \mathbf{F}_k\mathbf{P}_k\mathbf{F}_k^\top + \mathbf{G}_k\mathbf{Q}_k\mathbf{G}_k^\top.\end{aligned}$$

Полученное уравнение относится к так называемым *разностным уравнениям Ляпунова*.

Теперь посмотрим на $\mathbf{z}_k$. Так как по определению $\mathbf{z}_{k+1} = \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k$, где $\mathbf{x}_k$ и $\mathbf{v}_k$ некоррелированны, то несложно посчитать и матожидание, и дисперсию:

$$\begin{aligned}\mathsf{E}[\mathbf{z}_{k+1}] &= \mathbf{H}_k\,\mathsf{E}[\mathbf{x}_k] + \mathsf{E}[\mathbf{v}_k] = \mathbf{H}_k\,\mathsf{E}[\mathbf{x}_k]\\ \mathsf{D}[\mathbf{z}_k] &= \mathsf{E}\big[(\mathbf{H}_k(\mathbf{x}_k - \mathsf{E}[\mathbf{x}_k]) + \mathbf{v}_k)(\mathbf{H}_k(\mathbf{x}_k - \mathsf{E}[\mathbf{x}_k]) + \mathbf{v}_k)^\top\big] =\\ &= \mathbf{H}_k\mathbf{P}_k\mathbf{H}_k^\top + \mathbf{R}_k.\end{aligned}$$

1.15.2 Фильтр Калмана в гауссовском случае

Рассмотрим следующую модель Гаусса-Маркова:

$$\mathbf{x}_{k+1} = \mathbf{F}_k\mathbf{x}_k + \mathbf{G}_k\mathbf{w}_k$$
$$\mathbf{z}_{k+1} = \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k$$

для которой

- $\{\mathbf{w}_n\}$ и $\{\mathbf{v}_n\}$ — независимые белые гауссовские шумы, ковариации которых равны $\mathsf{cov}(\mathbf{w}_k, \mathbf{w}_l) = \mathbf{Q}_k\delta_{kl}$ и $\mathsf{cov}(\mathbf{v}_k, \mathbf{v}_l) = \mathbf{R}_k\delta_{kl}$.
- Начальное состояние системы $\mathbf{x}_0$ не зависит от шумов и имеет нормальное распределение с дисперсией $\mathbf{P}_0$.

Далее, пусть $\mathbf{Z}_k = (\mathbf{z}_0, \ldots, \mathbf{z}_k)$. Наша цель — найти рекурсифную формулу для следующей *оптимальной* в среднеквадратичном смысле оценки $\mathbf{x}_k$:

$$\widehat{\mathbf{x}}_k^+ \equiv \widehat{\mathbf{x}}_{k|k} = \mathsf{E}[\mathbf{x}_k \mid \mathbf{Z}_k].$$

Далее, введём *одношаговую оценку* (one-step predictor):

$$\widehat{\mathbf{x}}_k^- \equiv \widehat{\mathbf{x}}_{k|k-1} = \mathsf{E}[\mathbf{x}_k \mid \mathbf{Z}_{k-1}].$$

Для них вводятся соответствующие матрицы условных ковариаций:

$$\mathbf{P}_k^+ \equiv \mathbf{P}_{k|k} = \mathsf{D}[\mathbf{x}_k \mid \mathbf{Z}_k], \quad \mathbf{P}_k^- \equiv \mathbf{P}_{k|k-1} = \mathsf{D}[\mathbf{x}_k \mid \mathbf{Z}_{k-1}]$$

Примечание. На матрицу $\mathbf{P}_k^+$ (и, аналогично, на матрицу $\mathbf{P}_k^-$) можно смотреть, как на

1. матрицу ковариации *постериорной ошибки измерения* $\mathbf{e}_k = \mathbf{x}_k - \widehat{\mathbf{x}}_k^+$. Стоит заметить, что
$$\mathrm{MMSE} = \arg\min_{\widehat{\mathbf{x}}_k^+} \operatorname{tr} \mathsf{E}\big[(\mathbf{x} - \widehat{\mathbf{x}}_k^+)^\top(\mathbf{x} - \widehat{\mathbf{x}}_k^+)\big] = \operatorname{tr}\mathbf{P}_k^+.$$
2. матрицу ковариаций *условной случайной величины* $(\mathbf{x}_k \mid \mathbf{Z}_k)$.

Напоследок введём два обозначения: $\mathbf{P}_0^- = \mathbf{P}_0$ и $\widehat{\mathbf{x}}_0^- = \mathsf{E}[\mathbf{x}_0]$.
Теперь вспомним теорему о нормальной корреляции. Она гласит следующее:

- Если ξ и η имеют совместное нормальное распределение, то $(\xi \mid \eta)$ "имеет нормальное распределение" $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, где
$$\boldsymbol{\mu} = \boldsymbol{\mu}_\xi + \boldsymbol{\Sigma}_{\xi\eta}\boldsymbol{\Sigma}_{\eta\eta}^{-1}(\eta - \boldsymbol{\mu}_\eta),$$
$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\xi\xi} - \boldsymbol{\Sigma}_{\xi\eta}\boldsymbol{\Sigma}_{\eta\eta}^{-1}\boldsymbol{\Sigma}_{\eta\xi}.$$
- Более того, эта теорема корректна и в том случае, когда на всё навешивается условие на какой-то другой случайный вектор.

Теперь запишем одно простое свойство описанной выше модели:

Свойство 1. *Все случайные процессы описанной выше модели, то есть шумы,* $\mathbf{x}_k$ *и* $\mathbf{z}_k$*, являются гауссовскими.*

Доказательство. Следует из того, что начальные условия и шумы имеют совместное гауссовское распределение (так как независимость и каждый элемент имеет нормальное распределение), и того, что линейное преобразование гауссовского вектора тоже является гауссовским вектором. □

Отсюда, недолго думая, получаем гауссовость условных случайных величин $(\mathbf{x}_k \mid \mathbf{Z}_m)$ при всех $k, m \in \mathbb{N}$. Заметим, что

$$(\mathbf{x}_k \mid \mathbf{Z}_k) \sim \mathcal{N}(\widehat{\mathbf{x}}_k^+, \mathbf{P}_k^+), \quad (\mathbf{x}_k \mid \mathbf{Z}_{k-1}) \sim \mathcal{N}(\widehat{\mathbf{x}}_k^-, \mathbf{P}_k^-).$$

Теперь приступим к выводу. Допустим, что нам известна $(\widehat{\mathbf{x}}_k^-, \mathbf{P}_k^-)$. Как нам оценить $(\widehat{\mathbf{x}}_k^+, \mathbf{P}_k^+)$ и $(\widehat{\mathbf{x}}_{k+1}^-, \mathbf{P}_{k+1}^-)$? Это делается следующим образом.

Сначала проходит *шаг обновления измерений*. Вспомним, что $\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k$. Следовательно (почему?),

$$\left(\begin{pmatrix} \mathbf{x}_k \\ \mathbf{z}_k \end{pmatrix} \,\middle|\, \mathbf{Z}_{k-1} \right) \sim \mathcal{N}\left(\begin{pmatrix} \widehat{\mathbf{x}}_k^- \\ \mathbf{H}_k \widehat{\mathbf{x}}_k^- \end{pmatrix}, \begin{pmatrix} \mathbf{P}_k^- & \mathbf{P}_k^- \mathbf{H}_k^\top \\ \mathbf{H}_k \mathbf{P}_k^- & \mathbf{M}_k \end{pmatrix} \right),$$

где $\mathbf{M}_k \equiv \mathbf{H}_k \mathbf{P}_k^- \mathbf{H}_k^\top + \mathbf{R}_k$.

Теперь надо определить распределение $(\mathbf{x}_k \mid \mathbf{Z}_k) = (\mathbf{x}_k \mid \mathbf{z}_k, \mathbf{Z}_{k-1})$. Для этого воспользуемся теоремой о нормальной корреляции. Тогда $(\mathbf{x}_k \mid \mathbf{Z}_k)$ имеет нормальное распределение с параметрами

$$\begin{aligned} \widehat{\mathbf{x}}_k^+ &= \mathsf{E}[\mathbf{x}_k \mid \mathbf{Z}_k] = \widehat{\mathbf{x}}_k^- + \mathbf{P}_k^- (\mathbf{M}_k)^{-1} (\mathbf{z}_k - \mathbf{H}_k \widehat{\mathbf{x}}_k^-) \\ \mathbf{P}_k^+ &= \mathsf{D}[\mathbf{x}_k \mid \mathbf{Z}_k] = \mathbf{P}_k^- - \mathbf{P}_k^- \mathbf{H}_k^\top (\mathbf{M}_k)^{-1} \mathbf{H}_k \mathbf{P}_k^-. \end{aligned}$$

Дальше идёт *шаг обновления времени*. Теперь вспомним, что $\mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{G}_k \mathbf{w}_k$. Далее, заметим, что $\mathbf{x}_k$ и $\mathbf{w}_k$ независимы при условии $\mathbf{Z}_k$ (почему?). Тогда

$$\begin{aligned} \widehat{\mathbf{x}}_{k+1}^- &= \mathbf{F}_k \widehat{\mathbf{x}}_k^+, \\ \mathbf{P}_{k+1}^- &= \mathbf{F}_k \mathbf{P}_k^+ \mathbf{F}_k^\top + \mathbf{G}_k \mathbf{Q}_k \mathbf{G}_k^\top. \end{aligned}$$

Теперь выпишем несколько примечаний к этому:

- Фильтр Калмана считает и оценку $\widehat{\mathbf{x}}_k^+$, и её матрицу ковариаций $\mathbf{P}_k^+$ (аналогично для $\widehat{\mathbf{x}}_k^-$). Заметьте, что вычисление ковариации необходимо, так как оно является частью вычисления оценки. Впрочем, оно важно само по себе, так как оно задаёт меру неопределённости (или достоверности) оценки.

- Стоит заметить, что матрицы $\mathbf{P}_k^\pm$ не зависят от измерений $\{\mathbf{z}_k\}$. Следовательно, их можно посчитать заранее, если известны матрицы ковариаций для шума и *системные матрицы*: $\mathbf{F}_k$, $\mathbf{G}_k$, $\mathbf{H}_k$, $\mathbf{Q}_k$, $\mathbf{R}_k$ и $\mathbf{P}_0$.

- В гауссовском случае матрица $\mathbf{P}_k^+$ является безусловной ковариационной матрицей: $\mathbf{P}_k^+ = \mathsf{D}\left[\mathbf{x}_k - \widehat{\mathbf{x}}_k^+\right]$.

  В общем случае безусловная ковариация будет играть ключевую роль в выведении линейной оптимальной в среднеквадратичном смысле оценки.

- Допустим, что мы хотим оценить $\mathbf{s}_k = \mathbf{C}\mathbf{x}_k$. В таком случае оптимальной оценкой будет $\widehat{\mathbf{s}}_k = \mathsf{E}[\mathbf{C}\mathbf{x}_k \mid \mathbf{Z}_k] = \mathbf{C}\widehat{\mathbf{x}}_k^+$.

- Отклонение наблюдения от ожидаемого

  $$\tilde{\mathbf{z}}_k \equiv \mathbf{z}_k - \mathbf{H}_k \widehat{\mathbf{x}}_k^- = \mathbf{z}_k - \mathsf{E}[\mathbf{z}_k \mid \mathbf{Z}_{k-1}]$$

  называется *инновацией*, а процесс $\{\tilde{\mathbf{z}}_k\}$ называется *процессом инноваций*. Этот процесс сыграет немалую роль в дальнейшем. К слову: $\mathbf{M}_k = \mathsf{D}[\tilde{\mathbf{z}}_k]$.

1.15.3 Линейные оценки. Инновационный подход

Теперь перейдём к общему случаю. Но для начала немного пробежимся по линейным оценкам. Докажем одно утверждение:

Теорема 24. *Пусть* $\mathbf{x}$, $\mathbf{y}$ *— случайные векторы. Тогда оптимальная в среднеквадратичном смысле линейная (LMMSE) оценка* $\mathbf{x}$ *по* $\mathbf{y}$ *равна*

$$\widehat{\mathbf{x}} = \mathbf{m}_{\mathbf{x}} + \mathbf{\Sigma}_{\mathbf{xy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{y}}),$$

где $\mathbf{m}_{\mathbf{x}} = \mathsf{E}[\mathbf{x}]$, $\mathbf{\Sigma}_{\mathbf{xy}} = \mathsf{cov}(\mathbf{x}, \mathbf{y})$.

Доказательство. Так как мы ищем линейную оценку, то скажем, что $\widehat{\mathbf{x}} = \mathbf{A}^\top\mathbf{y} + \mathbf{b}$ (смысл транспонирования станет ясен немного позднее). Покажем, что оценка должна быть несмещённой. Допустим, что это не так, и $\mathbf{x} - \widehat{\mathbf{x}} = \mathbf{a} = (a_1, \ldots, a_n)$. Распишем MSE:

$$\mathrm{MSE} = \mathsf{E}\big[(\mathbf{x} - \widehat{\mathbf{x}})^\top(\mathbf{x} - \widehat{\mathbf{x}})\big] = \sum_{k=1}^{n}\mathsf{E}\big[(x_k - \widehat{x}_k)^2\big].$$

Отсюда получаем, что достаточно рассмотреть одну компоненту — например, k-ю. Заметим, что $\mathsf{E}[x_k] - \mathsf{E}[\widehat{x}_k] = a_k$. Пусть $\tilde{x}_k = (x_k - \mathsf{E}[x_k]) - (\widehat{x}_k - \mathsf{E}[\widehat{x}_k])$, то есть разность между оцениваемым параметром и несмещённой оценкой. Тогда $x_k - \widehat{x}_k = \tilde{x}_k + a_k$. Следовательно,

$$\mathsf{E}\big[(x_k - \widehat{x}_k)^2\big] = \mathsf{E}\big[(\tilde{x}_k + a_k)^2\big] = \mathsf{E}\big[\tilde{x}_k^2\big] + a_k^2 + 2a\,\mathsf{E}[\tilde{x}_k] = \mathsf{E}\big[\tilde{x}_k^2\big] + a_k^2.$$

Следовательно, $a_k = 0$ и оценка должна быть несмещённой. Тогда можно сказать, что $\mathsf{E}[\mathbf{x}] = \mathsf{E}[\widehat{\mathbf{x}}] = \mathbf{0}$.

Распишем среднеквадратическую ошибку, как функцию от матрицы $\mathbf{A}$:

$$\begin{aligned}\mathrm{MSE}(\mathbf{A}) &= \operatorname{tr}\mathsf{E}\big[(\mathbf{x} - \widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}})^\top\big] = \\ &= \operatorname{tr}\left(\mathsf{E}\big[\mathbf{x}\mathbf{x}^\top\big] - \mathsf{E}\big[\widehat{\mathbf{x}}\mathbf{x}^\top\big] - \mathsf{E}\big[\mathbf{x}\widehat{\mathbf{x}}^\top\big] + \mathsf{E}\big[\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\big]\right) = \\ &= \operatorname{tr}\left(\mathbf{\Sigma}_{\mathbf{xx}} - \mathbf{A}^\top\mathbf{\Sigma}_{\mathbf{yx}} - \mathbf{\Sigma}_{\mathbf{xy}}\mathbf{A} + \mathbf{A}^\top\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{A}\right)\end{aligned}$$

Теперь продифференцируем по $\mathbf{A}$. Для этого вспомним три правила матричного дифференцирования (если не знали их ранее, то их несложно вывести покомпонентно):

$$\frac{\partial \operatorname{tr}(\mathbf{X}^\top\mathbf{A})}{\partial \mathbf{X}} = \mathbf{A}, \quad \frac{\partial \operatorname{tr}(\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}^\top, \quad \frac{\partial \operatorname{tr}\mathbf{X}^\top\mathbf{A}\mathbf{X}}{\partial \mathbf{X}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{X}.$$

Следовательно,

$$\frac{\partial}{\partial \mathbf{A}}\mathrm{MSE}(\mathbf{A}) = -2\mathbf{\Sigma}_{\mathbf{yx}} + 2\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{A} = 0 \implies \mathbf{A} = \mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}.$$

Отсюда получаем, что $\widehat{\mathbf{x}} = \mathbf{\Sigma}_{\mathbf{xy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{y}$. Переходя к произвольным значениям матожиданий, получаем желаемое. □

Рассуждая аналогично доказательству теоремы о нормальной компоненте, можно получить, что

$$\mathsf{D}[\mathbf{x} - \widehat{\mathbf{x}}] = \mathbf{\Sigma}_{\mathbf{xx}} - \mathbf{\Sigma}_{\mathbf{xy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}.$$

Для удобства введём обозначение $\mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}] \equiv \widehat{\mathbf{x}}$. Стоит заметить, что это *не* стандартное обозначение.

Теперь покажем одно свойство этой оценки:

Свойство 1 (принцип ортогональности)**.** *Для любой линейной функции* $L(\mathbf{y})$ *выполнено*

$$\mathsf{E}\big[(\mathbf{x} - \mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}])L(\mathbf{y})^\top\big] = 0.$$

Доказательство. Так как мы рассматриваем линейные функции от $\mathbf{y}$, то скажем, что $L(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{b}$. Далее, заметим, что $\mathsf{E}\big[\mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}]\big] = \mathsf{E}[\mathbf{x}]$. Тогда от значения $\mathbf{b}$ ничего не зависит и можно считать, что

$$L(\mathbf{y}) = \mathbf{A}(\mathbf{y} - \mathsf{E}[\mathbf{y}]).$$

Теперь можно сказать, что достаточно смотреть на величины $\mathbf{x}$ и $\mathbf{y}$ с нулевым матожиданием. Теперь же можно просто аккуратно расписать:

$$\begin{aligned}\mathsf{E}\big[(\mathbf{x} - \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{y})(\mathbf{A}\mathbf{y})^\top\big] &= \mathsf{E}\big[\mathbf{x}\mathbf{y}^\top\big]\mathbf{A}^\top - \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\,\mathsf{E}\big[\mathbf{y}\mathbf{y}^\top\big]\mathbf{A}^\top = \\ &= \boldsymbol{\Sigma}_{\mathbf{xy}}\mathbf{A}^\top - \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yy}}\mathbf{A}^\top = \mathbf{0}.\end{aligned}$$

Тем самым мы получили желаемое. □

Следующее свойство будет крайне полезно в дальнейшем описании. Оно получается, если положить $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$.

Свойство 2. *Пусть* $\mathsf{cov}(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{0}$*. Тогда*

$$\mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2] = \mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}_1] + \mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}_2] - \mathsf{E}[\mathbf{x}].$$

Более того,

$$\mathsf{D}\big[\mathbf{x} - \mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2]\big] = \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_1}\boldsymbol{\Sigma}_{\mathbf{y}_1\mathbf{y}_1}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}_1\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_2}\boldsymbol{\Sigma}_{\mathbf{y}_2\mathbf{y}_2}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}_2\mathbf{x}}.$$

Доказательство. Докажем только первое утверждение, оставив второе в качестве упражнения. Заметим, что матрица ковариаций имеет следующий вид:

$$\left(\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \hline \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}} \end{array}\right) = \left(\begin{array}{c|cc} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_1} & \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_2} \\ \hline \boldsymbol{\Sigma}_{\mathbf{y}_1\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{y}_1\mathbf{y}_1} & \mathbf{0} \\ \boldsymbol{\Sigma}_{\mathbf{y}_2\mathbf{x}} & \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{y}_2\mathbf{y}_2} \end{array}\right).$$

Теперь посмотрим на $\mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2]$. Оно равно

$$\mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2] = \mathsf{E}[\mathbf{x}] + \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}(\mathbf{y} - \mathsf{E}[\mathbf{y}]).$$

Подставим полученные выше блочные матрицы:

$$\mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2] = \mathsf{E}[\mathbf{x}] + \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_1} & \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_2}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{y}_1\mathbf{y}_1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{y}_2\mathbf{y}_2}\end{pmatrix}^{-1}(\mathbf{y} - \mathsf{E}[\mathbf{y}]).$$

Теперь вспомним, что обратная к диагональной блочной матрице равна матрице с обратными блоками. Следовательно, это равно

$$\mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2] = \mathsf{E}[\mathbf{x}] + \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_1} & \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_2}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{y}_1\mathbf{y}_1}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{y}_2\mathbf{y}_2}^{-1}\end{pmatrix}(\mathbf{y} - \mathsf{E}[\mathbf{y}]).$$

Осталось заметить, что произведение матриц ведёт себя, как квадратичная форма (и такое рассмотрение корректно из-за совпадений размеров матриц). Тогда

$$\begin{aligned}\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_1} & \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_2}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{y}_1\mathbf{y}_1}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{y}_2\mathbf{y}_2}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{y}_1 - \mathsf{E}[\mathbf{y}_1] \\ \mathbf{y}_2 - \mathsf{E}[\mathbf{y}_2]\end{pmatrix} &= \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_1}\boldsymbol{\Sigma}_{\mathbf{y}_1\mathbf{y}_1}^{-1}(\mathbf{y}_1 - \mathsf{E}[\mathbf{y}_1]) + \\ &+ \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}_2}\boldsymbol{\Sigma}_{\mathbf{y}_2\mathbf{y}_2}^{-1}(\mathbf{y}_2 - \mathsf{E}[\mathbf{y}_2])\end{aligned}$$

Отсюда получаем желаемое. □

Теперь рассмотрим случайную последовательность $\{\mathbf{z}_k\}_{k\geqslant 0}$. По ней можно построить так называемую *последовательность инноваций* (в широком смысле):

$$\tilde{\mathbf{z}}_k = \mathbf{z}_k - \mathsf{E}^{\mathbf{L}}[\mathbf{z}_k \mid \mathbf{Z}_{k-1}].$$

На условную случайную величину $\tilde{\mathbf{z}}_k$ можно смотреть, как на случайную величину, содержащую только новую статистическую информацию, которой нет в $\mathbf{Z}_{k-1}$.

Пользуясь свойствами LMMSE-оценки, можно сразу же выписать три свойства:

- $\mathsf{E}[\tilde{\mathbf{z}}] = \mathbf{0}$.
- $\tilde{\mathbf{z}}_k$ есть линейная функция от $\mathbf{Z}_k$.
- Из второго пункта следует, что $\mathsf{cov}(\tilde{\mathbf{z}}_k, \tilde{\mathbf{z}}_l) = \mathbf{0}$ при $k \neq l$.

Тогда можно сказать, что $\{\mathbf{z}_k\}_{k\geqslant 0}$ является *белым шумом в широком смысле*.

Теперь введём обозначение $\tilde{\mathbf{Z}}_k = (\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_k)$. Сразу же заметим, что $\tilde{\mathbf{Z}}_k$ есть линейная функция от $\mathbf{Z}_k$ (так как при построении берутся только линейные преобразования). Следовательно, $\mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \mathbf{Z}_{k-1}] = \mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \tilde{\mathbf{Z}}_{k-1}]$ для любой случайной величины $\mathbf{x}$. Действительно, если $\tilde{\mathbf{Z}}_k = \mathbf{A}\mathbf{Z}_k + \mathbf{b}$, то

$$\begin{aligned}
\mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \tilde{\mathbf{Z}}_k] &= \mathsf{E}[\mathbf{x}] + \mathsf{cov}(\mathbf{x}, \mathbf{A}\mathbf{Z}_k + \mathbf{b})\,\mathsf{D}[\mathbf{A}\mathbf{Z}_k + \mathbf{b}]^{-1}(\mathbf{A}\mathbf{Z}_k + \mathbf{b} - \mathsf{E}[\mathbf{A}\mathbf{Z}_k + \mathbf{b}]) = \\
&= \mathsf{E}[\mathbf{x}] + \mathsf{cov}(\mathbf{x}, \mathbf{Z}_k)\mathbf{A}^\top(\mathbf{A}\,\mathsf{D}[\mathbf{Z}_k]\mathbf{A}^\top)^{-1}\mathbf{A}(\mathbf{Z}_k - \mathsf{E}[\mathbf{Z}_k]) = \\
&= \mathsf{E}[\mathbf{x}] + \mathbf{\Sigma}_{\mathbf{x}\mathbf{Z}_k}\mathbf{A}^\top(\mathbf{A}^\top)^{-1}\mathbf{\Sigma}_{\mathbf{Z}_k\mathbf{Z}_k}^{-1}\mathbf{A}^{-1}\mathbf{A}(\mathbf{Z}_k - \mathsf{E}[\mathbf{Z}_k]) = \\
&= \mathsf{E}[\mathbf{x}] + \mathbf{\Sigma}_{\mathbf{x}\mathbf{Z}_k}\mathbf{\Sigma}_{\mathbf{Z}_k\mathbf{Z}_k}^{-1}(\mathbf{Z}_k - \mathsf{E}[\mathbf{Z}_k]) = \mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \mathbf{Z}_k]
\end{aligned}$$

Из этого следует, что (если $\mathsf{E}[\mathbf{x}] = 0$)

$$\mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \mathbf{Z}_k] = \mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \tilde{\mathbf{Z}}_k] = \mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \tilde{\mathbf{Z}}_{k-1}] + \mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \tilde{\mathbf{z}}_k] = \sum_{i=0}^{k} \mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \tilde{\mathbf{z}}_i]$$

В принципе, вся необходимая теория была введена, так что можно начинать вводить фильтр Калмана для линейных негауссовских моделей. Мы немного обобщим задачу, разрешив корреляцию между шумами. Рассмотрим следующую модель:

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{F}_k\mathbf{x}_k + \mathbf{G}_k\mathbf{w}_k \\
\mathbf{z}_{k+1} &= \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k
\end{aligned}$$

Скажем, что $\{\mathbf{w}_k\}$ и $\{\mathbf{v}_k\}$ — белые шумы в широком смысле с нулевым матожиданием и ковариацией

$$\mathsf{cov}\left(\begin{pmatrix}\mathbf{w}_k\\ \mathbf{v}_k\end{pmatrix}, \begin{pmatrix}\mathbf{w}_l\\ \mathbf{v}_l\end{pmatrix}\right) = \begin{pmatrix}\mathbf{Q}_k & \mathbf{S}_k\\ \mathbf{S}_k^\top & \mathbf{R}_k\end{pmatrix}\delta_{kl}$$

Далее, начальное условие $\mathbf{x}_0$ имеет матрицу ковариаций $\mathbf{P}_0$ и не коррелирует с шумами.

Теперь введём несколько обозначений:

$$\begin{aligned}
\mathbf{Z}_k &= (\mathbf{z}_1, \ldots, \mathbf{z}_k), \\
\widehat{\mathbf{x}}_{k|k} &= \mathsf{E}^{\mathbf{L}}[\mathbf{x}_k \mid \mathbf{Z}_k], & \widehat{\mathbf{x}}_{k|k-1} &= \mathsf{E}^{\mathbf{L}}[\mathbf{x}_k \mid \mathbf{Z}_{k-1}], \\
\tilde{\mathbf{x}}_{k|k} &= \mathbf{x}_k - \widehat{\mathbf{x}}_{k|k}, & \tilde{\mathbf{x}}_{k|k-1} &= \mathbf{x}_k - \widehat{\mathbf{x}}_{k|k-1}, \\
\mathbf{P}_{k|k} &= \mathsf{D}\big[\widehat{\mathbf{x}}_{k|k}\big], & \mathbf{P}_{k|k-1} &= \mathsf{D}\big[\widehat{\mathbf{x}}_{k|k-1}\big].
\end{aligned}$$

Инновации будем вводить, как и раньше:

$$\tilde{\mathbf{z}}_k = \mathbf{z}_k - \mathsf{E}^{\mathbf{L}}[\mathbf{z}_k \mid \mathbf{Z}_{k-1}].$$

Заметим, что второй член можно преобразовать:

$$\begin{aligned}\mathsf{E}^{\mathbf{L}}[\mathbf{z}_k \mid \mathbf{Z}_{k-1}] &= \mathsf{E}[\mathbf{z}] + \boldsymbol{\Sigma}_{\mathbf{z}_k\mathbf{Z}_{k-1}}\boldsymbol{\Sigma}^{-1}_{\mathbf{Z}_{k-1}\mathbf{Z}_{k-1}}(\mathbf{Z}_{k-1} - \mathsf{E}[\mathbf{Z}_{k-1}]) = \\ &= \mathbf{H}_k\,\mathsf{E}[\mathbf{x}_k] + \mathbf{H}_k\boldsymbol{\Sigma}_{\mathbf{x}_k\mathbf{Z}_{k-1}}\boldsymbol{\Sigma}^{-1}_{\mathbf{Z}_{k-1}\mathbf{Z}_{k-1}}(\mathbf{Z}_{k-1} - \mathsf{E}[\mathbf{Z}_{k-1}]) = \\ &= \mathbf{H}_k\,\mathsf{E}^{\mathbf{L}}[\mathbf{x}_k \mid \mathbf{Z}_{k-1}] = \mathbf{H}_k\widehat{\mathbf{x}}_{k|k-1}.\end{aligned}$$

Пользуясь этим, получаем, что

$$\tilde{\mathbf{z}}_k = \mathbf{z}_k - \mathbf{H}_k\widehat{\mathbf{x}}_{k|k-1} = \mathbf{H}_k\tilde{\mathbf{x}}_{k|k-1} + \mathbf{v}_k.$$

Приступим к описанию самого фильтра. Он состоит из шага обновления измерения, шага обновления времени. Иногда вводят шаг обновления ковариации.

1. Начнём с шага обновления измерения, то есть с перехода от $\widehat{\mathbf{x}}_{k|k-1}$ к $\widehat{\mathbf{x}}_{k|k}$. Пользуясь свойством 2 и тем, что $\mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \mathbf{Z}_k] = \mathsf{E}^{\mathbf{L}}[\mathbf{x} \mid \tilde{\mathbf{Z}}_k]$, получаем, что

$$\widehat{\mathbf{x}}_{k|k} = \mathsf{E}^{\mathbf{L}}[\mathbf{x}_k \mid \mathbf{Z}_k] = \mathsf{E}^{\mathbf{L}}[\mathbf{x}_k \mid \tilde{\mathbf{Z}}_k] = \mathsf{E}^{\mathbf{L}}[\mathbf{x}_k \mid \tilde{\mathbf{Z}}_{k-1}] + \mathsf{E}^{\mathbf{L}}[\mathbf{x}_k \mid \tilde{\mathbf{z}}_k] - \mathsf{E}[\mathbf{x}_k].$$

   Это выражение и лежит в основе инновационного подхода. Остальное следует из прямых вычислений и принципа ортогональности. Для начала заметим, что

$$\mathsf{E}^{\mathbf{L}}[\mathbf{x}_k \mid \tilde{\mathbf{z}}_k] - \mathsf{E}[\mathbf{x}_k] = \mathsf{cov}(\mathbf{x}_k, \tilde{\mathbf{z}}_k)\,\mathsf{D}[\tilde{\mathbf{z}}_k]^{-1}\tilde{\mathbf{z}}_k.$$

   Теперь считаем эти ковариации. Начнём с первой:

$$\mathsf{cov}(\mathbf{x}_k, \tilde{\mathbf{z}}_k) = \mathsf{cov}(\mathbf{x}_k, \mathbf{H}_k\tilde{\mathbf{x}}_{k|k-1} + \mathbf{v}_k) = \mathsf{cov}(\mathbf{x}_k, \tilde{\mathbf{x}}_{k|k-1})\mathbf{H}_k^\top + \mathsf{cov}(\mathbf{x}_k, \mathbf{v}_k).$$

   Распишем первый член подробнее. Заметим, что $\widehat{\mathbf{x}}_k$ есть несмещённая оценка $\mathbf{x}_k$. Тогда, добавив принцип ортогональности, получаем, что

$$\begin{aligned}\mathsf{cov}(\mathbf{x}_k, \tilde{\mathbf{x}}_{k|k-1}) &= \mathsf{E}\big[(\mathbf{x}_k - \mathsf{E}[\mathbf{x}_k])(\mathbf{x}_k - \widehat{\mathbf{x}}_{k|k-1})^\top\big] = \mathsf{E}\big[\mathbf{x}_k(\mathbf{x}_k - \widehat{\mathbf{x}}_{k|k-1})^\top\big] = \\ &= \mathsf{E}\big[\mathbf{x}_k(\mathbf{x}_k - \widehat{\mathbf{x}}_{k|k-1})^\top\big] + \big(\mathsf{E}\big[(\mathbf{x}_k - \widehat{\mathbf{x}}_k)\widehat{\mathbf{x}}_{k|k-1}^\top\big]\big)^\top = \\ &= \mathsf{D}\big[\mathbf{x}_k - \widehat{\mathbf{x}}_{k|k-1}\big] = \mathbf{P}_{k|k-1}.\end{aligned}$$

   Следовательно, так как $\mathbf{x}_k$ не коррелирует с $\mathbf{v}_k$, то $\mathsf{cov}(\mathbf{x}_k, \tilde{\mathbf{z}}_k) = \mathbf{P}_{k|k-1}\mathbf{H}_k^\top$. Далее заметим, что $\widehat{\mathbf{x}}_{k-1}$ тоже не коррелирует с $\mathbf{v}_k$, так как зависит только от $\mathbf{Z}_{k-1}$ и $\mathbf{x}_k$. Следовательно,

$$\mathsf{D}[\tilde{\mathbf{z}}_k] = \mathsf{D}\big[\mathbf{H}_k\tilde{\mathbf{x}}_{k|k-1} + \mathbf{v}_k\big] = \mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^\top + \mathbf{R}_k \equiv \mathbf{M}_k.$$

   Собирая результаты выше в один большой результат, получаем, что

$$\widehat{\mathbf{x}}_{k|k} = \widehat{\mathbf{x}}_{k|k-1} + \mathbf{P}_{k|k-1}\mathbf{H}_k^\top\mathbf{M}_k^{-1}\tilde{\mathbf{z}}_k.$$

2. Теперь приступим к шагу обновления времени, то есть к переходу от $\widehat{\mathbf{x}}_{k|k}$ к $\widehat{\mathbf{x}}_{k+1|k}$. Распишем:

$$\begin{aligned}\widehat{\mathbf{x}}_{k+1|k} &= \mathsf{E}^{\mathbf{L}}[\mathbf{x}_{k+1} \mid \mathbf{Z}_k] = \mathsf{E}^{\mathbf{L}}[\mathbf{F}_k\mathbf{x}_k + \mathbf{G}_k\mathbf{w}_k \mid \mathbf{Z}_k] = \\ &= \mathsf{E}[\mathbf{F}_k\mathbf{x}_k + \mathbf{G}_k\mathbf{w}_k] + \mathsf{cov}(\mathbf{F}_k\mathbf{x}_k + \mathbf{G}_k\mathbf{w}_k, \mathbf{Z}_k)\,\mathsf{D}[\mathbf{Z}_k]^{-1}(\mathbf{Z}_k - \mathsf{E}[\mathbf{Z}_k]) = \\ &= \mathbf{F}_k\,\mathsf{E}^{\mathbf{L}}[\mathbf{x}_k \mid \mathbf{Z}_k] + \mathbf{G}_k\,\mathsf{E}^{\mathbf{L}}[\mathbf{w}_k \mid \mathbf{Z}_k].\end{aligned}$$

Дальше воспользуемся тем, что $\mathbf{w}_k$ не коррелирует с $\mathbf{Z}_{k-1}$, а, следовательно, и с $\tilde{\mathbf{Z}}_{k-1}$. Тогда это равно

$$\begin{aligned}\widehat{\mathbf{x}}_{k+1|k} &= \mathbf{F}_k\,\mathsf{E}^{\mathbf{L}}[\mathbf{x}_k \mid \mathbf{Z}_k] + \mathbf{G}_k\,\mathsf{E}^{\mathbf{L}}[\mathbf{w}_k \mid \tilde{\mathbf{Z}}_k] = \\ &= \mathbf{F}_k\widehat{\mathbf{x}}_{k|k} + \mathbf{G}_k(\mathsf{E}^{\mathbf{L}}[\mathbf{w}_k \mid \tilde{\mathbf{z}}_k] + \mathsf{E}^{\mathbf{L}}[\mathbf{w}_k \mid \tilde{\mathbf{Z}}_{k-1}] = \\ &= \mathbf{F}_k\widehat{\mathbf{x}}_{k|k} + \mathbf{G}_k\,\mathsf{E}^{\mathbf{L}}[\mathbf{w}_k \mid \tilde{\mathbf{z}}_k].\end{aligned}$$

Осталось посчитать $\mathsf{E}^{\mathbf{L}}[\mathbf{w}_k \mid \tilde{\mathbf{z}}_k] = \mathsf{E}[\mathbf{w}_k] + \mathsf{cov}(\mathbf{w}_k, \tilde{\mathbf{z}}_k)\,\mathsf{D}[\tilde{\mathbf{z}}_k]^{-1}\tilde{\mathbf{z}}_k$. Распишем ковариацию подробнее, пользуясь тем, что и $\mathbf{x}_k$, и $\widehat{\mathbf{x}}_{k|k-1}$ не коррелируют с $\mathbf{w}_k$:

$$\begin{aligned}\mathsf{cov}(\mathbf{w}_k, \tilde{\mathbf{z}}_k) &= \mathsf{cov}(\mathbf{w}_k, \mathbf{H}_k\tilde{\mathbf{x}}_{k|k-1} + \mathbf{v}_k) = \mathsf{cov}(\mathbf{w}_k, \tilde{\mathbf{x}}_{k|k-1})\mathbf{H}^\top + \mathbf{S}_k = \\ &= \mathsf{cov}(\mathbf{w}_k, \mathbf{x}_k) - \mathsf{cov}(\mathbf{w}_k, \widehat{\mathbf{x}}_k) + \mathbf{S}_k = \mathbf{S}_k.\end{aligned}$$

Следовательно, шаг обновления времени записывается так:

$$\widehat{\mathbf{x}}_{k+1|k} = \mathbf{F}_k\widehat{\mathbf{x}}_{k|k} + \mathbf{G}_k\mathbf{S}_k\mathbf{M}_k^{-1}\tilde{\mathbf{z}}_k.$$

3. Предыдущие шаги можно скомпоновать в один:

$$\widehat{\mathbf{x}}_{k+1|k} = \mathbf{F}_k\widehat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k\tilde{\mathbf{z}}_k,$$

где $\mathbf{K}_k = (\mathbf{F}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^\top + \mathbf{G}_k\mathbf{S}_k)\mathbf{M}_k^{-1}$.

4. Теперь рассмотрим шаг обновления ковариации, то есть переход от $\mathbf{P}_{k|k-1}$ к $\mathbf{P}_{k+1|k}$. Стоит сказать, что переход от $\mathbf{P}_{k|k-1}$ к $\mathbf{P}_{k|k}$ полностью соответствует гауссовскому случаю, поэтому мы сразу смотрим комбинацию переходов. Заметим, что

$$\begin{aligned}\tilde{\mathbf{x}}_{k+1|k} &= \mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k} = \mathbf{x}_{k+1} - \mathbf{F}_k\widehat{\mathbf{x}}_{k|k-1} - \mathbf{K}_k\tilde{\mathbf{z}}_k = \\ &= \mathbf{F}_k\mathbf{x}_k + \mathbf{G}_k\mathbf{w}_k - \mathbf{F}_k\widehat{\mathbf{x}}_{k|k-1} - \mathbf{K}_k(\mathbf{H}_k\tilde{\mathbf{x}}_{k|k-1} + \mathbf{v}_k) = \\ &= (\mathbf{F}_k - \mathbf{K}_k\mathbf{H}_k)\tilde{\mathbf{x}}_{k|k-1} + \mathbf{G}_k\mathbf{w}_k - \mathbf{K}_k\mathbf{v}_k.\end{aligned}$$

Теперь, пользуясь некоррелированностью $\tilde{\mathbf{x}}_{k|k-1}$ с $\mathbf{w}_k$ и $\mathbf{v}_k$, получаем, что

$$\begin{aligned}\mathbf{P}_{k+1|k} &= \mathsf{D}\big[(\mathbf{F}_k - \mathbf{K}_k\mathbf{H}_k)\tilde{\mathbf{x}}_{k|k-1}\big] + \mathsf{D}[\mathbf{G}_k\mathbf{w}_k - \mathbf{K}_k\mathbf{v}_k] = \\ &= (\mathbf{F}_k - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k|k-1}(\mathbf{F}_k - \mathbf{K}_k\mathbf{H}_k)^\top + \mathsf{D}[\mathbf{G}_k\mathbf{w}_k - \mathbf{K}_k\mathbf{v}_k].\end{aligned}$$

Значение второго члена получить несложно, но формула для него достаточно длинная. Запишем её без доказательства:

$$\mathsf{D}[\mathbf{G}_k\mathbf{w}_k - \mathbf{K}_k\mathbf{v}_k] = \mathbf{G}_k\mathbf{Q}_k\mathbf{G}_k^\top + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^\top - \mathbf{G}_k\mathbf{S}_k\mathbf{K}_k^\top - \mathbf{K}_k\mathbf{S}_k^\top\mathbf{G}_k^\top.$$

Определения и результаты, связанные с линейными оценками, можно хорошо интерпретировать с помощью гильбертовых пространств.

Пусть, для простоты, все случайные величины (то есть, $\mathbf{w}_k$, $\mathbf{v}_k$ и $\mathbf{x}_0$) имеют нулевое матожидание.

Напомню, что гильбертово пространство — это полное пространство со скалярным произведением. Другими словами, это линейное пространство V (для простоты скажем, что над $\mathbb{R}$) с определённой на нём операцией скалярного произведения $\langle\cdot,\cdot\rangle : V \times V \mapsto \mathbb{R}$, которая обладает следующими свойствами:

1. Оно линейно по первому аргументу: для любых $\alpha, \beta \in \mathbb{R}$, $x, y, z \in V$

$$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle.$$

2. Оно симметрично: для любых $x, y \in V$ $\langle x, y \rangle = \langle y, x \rangle$.

3. Оно положительно определено: для любого $x \in V$ $\langle x, x \rangle \geqslant 0$, причём $\langle x, x \rangle = 0 \iff x = 0$.

Полнота означает, что любая фундаментальная (т.е. для которой выполнено условие Коши) последовательность имеет предел. Далее, это скалярное произведение порождает норму: $\|x\| = \sqrt{\langle x, x \rangle}$. Теперь выпишем несколько стандартных утверждений из линейной алгебры:

- *Подпространство* S — это замкнутое относительно линейных преобразований подмножество V. Другими словами, оно является линейной оболочкой каких-то векторов $\{v_\alpha\}$.

- *Ортогональная проекция* $\Pi_S v$ вектора v на подпространство S — это ближайший к v элемент из S, то есть это вектор $w \in S$, который минимизирует $\|v - w\|$. Такой вектор действительно существует и для него верно, что $v - \Pi_S v \perp S$, то есть $\langle v - \Pi_S v, w \rangle = 0$ для любого $w \in S$.

- Если $S = \operatorname{span}(s_1, \ldots, s_n)$, то

$$\Pi_S v = \sum_{k=1}^{n} \alpha_k s_k, \text{ где } \begin{pmatrix} \alpha_1 & \ldots & \alpha_n \end{pmatrix} = \begin{pmatrix} \langle v, s_1 \rangle & \ldots & \langle v, s_1 \rangle \end{pmatrix} \begin{pmatrix} \langle s_1, s_1 \rangle & \ldots & \langle s_1, s_n \rangle \\ \vdots & \ddots & \vdots \\ \langle s_n, s_1 \rangle & \ldots & \langle s_n, s_n \rangle \end{pmatrix}^{-1}$$

  Если $(s_1, \ldots, s_n)$ — это ортогональный базис S, то

$$\Pi_S v = \sum_{k=1}^{n} \langle v, s_k \rangle \langle s_k, s_k \rangle^{-1} s_k.$$

- Если $S = S_1 \oplus S_2$ (то есть, S есть прямая сумма двух ортогональных подпространств S_1 и S_2), то $\Pi_S v = \Pi_{S_1} v + \Pi_{S_2} v$.

- Если есть набор линейно независимых векторов $\{v_1, v_2, \ldots\}$, то его можно превратить в ортогональный базис, используя *процесс Грама-Шмидта*:

$$\tilde{v}_k = v_k - \Pi_{\operatorname{span}(v_1, \ldots, v_{k-1})} v_k = v_k - \sum_{i=1}^{k-1} \langle v_k, \tilde{v}_i \rangle \langle \tilde{v}_i, \tilde{v}_i \rangle^{-1} \tilde{v}_i.$$

Теперь можно провести аналогию между полученными ранее результатами, связанными с линейными оценками, с этими фактами, заметив следующее:

- В качестве гильбертова пространства возьмём пространство случайных векторов, для которых $\mathsf{E}[\mathbf{x}] = \mathbf{0}$ и $\mathsf{E}\big[\mathbf{x}^\top \mathbf{x}\big] < +\infty$. В таком случае скалярное произведение задаётся следующим образом: $\langle \mathbf{x}, \mathbf{y} \rangle = \mathsf{E}\big[\mathbf{x}^\top \mathbf{y}\big] = \operatorname{tr} \mathsf{E}\big[\mathbf{x}\mathbf{y}^\top\big]$.

- Оптимальная линейная оценка $\mathsf{E}^{\mathrm{L}}[\mathbf{x} \mid \mathbf{Z}_k]$ является ничем иным, как ортогональная проекция $\mathbf{x}$ на линейную оболочку векторов $(\mathbf{z}_1, \ldots, \mathbf{z}_k)$ (почему?).

- Процесс инноваций $\{\tilde{\mathbf{z}}_k\}$ есть ортогонализированная версия процесса $\{\mathbf{z}_k\}$.

Формулировка через гильбертовы пространства может дать несколько хороших идей и результатов (особенно в случае с непрерывным временем, так называемым фильтром Калмана-Бьюси). Но мы остановимся на этом.

1.15.4 Фильтр Калмана и метод наименьших квадратов

Рассмотрим следующую задачу оптимизации: на переменные $\mathbf{x}_0, \dots, \mathbf{x}_k$ и $\mathbf{w}_0, \dots, \mathbf{w}_{k-1}$ наложены следующие ограничения:

$$\mathbf{x}_{i+1} = \mathbf{F}_i \mathbf{x}_i + \mathbf{G}_i \mathbf{w}_i, \quad i \in \{0, 1, \dots, k-1\}.$$

Далее, платёжная функция, которую нужно минимизировать, равна

$$J_k = \frac{1}{2}(\mathbf{x}_0 - \overline{\mathbf{x}}_0)^\top \mathbf{P}_0^{-1}(\mathbf{x}_0 - \overline{\mathbf{x}}_0) + \frac{1}{2}\sum_{i=0}^{k}(\mathbf{z}_i - \mathbf{H}_i\mathbf{x}_i)^\top \mathbf{R}_i^{-1}(\mathbf{z}_i - \mathbf{H}_i\mathbf{x}_i) + \frac{1}{2}\sum_{i=0}^{k-1}\mathbf{w}_i^\top \mathbf{Q}_i^{-1}\mathbf{w}_i.$$

В этой задаче $\overline{\mathbf{x}}_0$, $\{\mathbf{z}_k\}$ — это известные векторы, а $\mathbf{P}_0$, $\mathbf{R}_k$ и $\mathbf{Q}_k$ — положительно определённые симметричные матрицы.

Пусть $\mathbf{x}_0^{(k)}, \dots, \mathbf{x}_k^{(k)}$ — это оптимальное решение задачи. Утверждается, что $\mathbf{x}_k^{(k)}$ можно вычислить точно так же, как вычисляется $\widehat{\mathbf{x}}_{k|k}$ в фильтре Калмана.

Этот результат можно получить в лоб, расписывая решение методом наименьших квдаратов для $k-1$ и k и преобразованиями матриц. Но мы пойдём проще, используя гауссовский подход.

Теорема 25. *Оптимальное решение $\mathbf{x}_0^{(k)}, \dots, \mathbf{x}_k^{(k)}$ описанной выше задачи оптимизации является максимизатором условной вероятности (другими словами, это оценка апостериорного максимума, или MAP-оценка)*

$$(\mathbf{x}_0^{(k)}, \dots, \mathbf{x}_k^{(k)}) = \arg\max_{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_k} p(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_k \mid \mathbf{Z}_k),$$

которая связана с следующей гауссовской моделью:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{F}_k\mathbf{x}_k + \mathbf{G}_k\mathbf{w}_k, & \mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_0), \mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k) \\ \mathbf{z}_{k+1} &= \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k & \mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k) \end{aligned}$$

Для этой модели $\{\mathbf{w}_k\}$ и $\{\mathbf{v}_k\}$ — белые шумы, не зависящие от начального условия $\mathbf{x}_0$, матожидание которого равно $\overline{\mathbf{x}}_0$.

Указание. Распишите $p(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{Z}_k)$.

Теперь нужно указать ещё один факт (без доказательства):

Теорема 26. *Для таких гауссовских моделей MAP = MMSE.*

Из указанных выше фактов следует то, что $\mathbf{x}_k^{(k)} = \mathbf{x}_k^+$, что и требовалось доказать.