



Введение в гауссовские процессы на Python

Максим Николаев
ПОМИ, 5 октября 2025

В прошлый раз

- Оптимальное предсказание $\hat{Y}$ относительно MSE называется проекцией. Если предсказываем с помощью элемента линейного пространства, то $Y - \hat{Y}$ ортогонально этому пространству
- Оптимальное линейное предсказание: $\mathbb{E}Y + \text{cov}(Y, X)\text{cov}(X)^{-1}(X - \mathbb{E}X)$
- Многомерное нормальное распределение $\mathcal{N}(\mu, \Sigma)$ это $\Sigma^{1/2}Z + \mu$, где $Z \sim \mathcal{N}(0, I)$
- Если $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$, то $Y|(X = x) \sim \mathcal{N}(\mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X), \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})$



Случайные процессы

Как нетрудно догадаться, в анализе данных редко интересуются отдельными случайными величинами. Существенно более мощным инструментом являются **случайные процессы** — семейства случайных величин $\{X_t\}_{t \in T}$, индексируемые множеством индексов T , и имеющие совместное распределение.

Со случаем **конечного** T мы уже сталкивались — это просто случайные вектора.

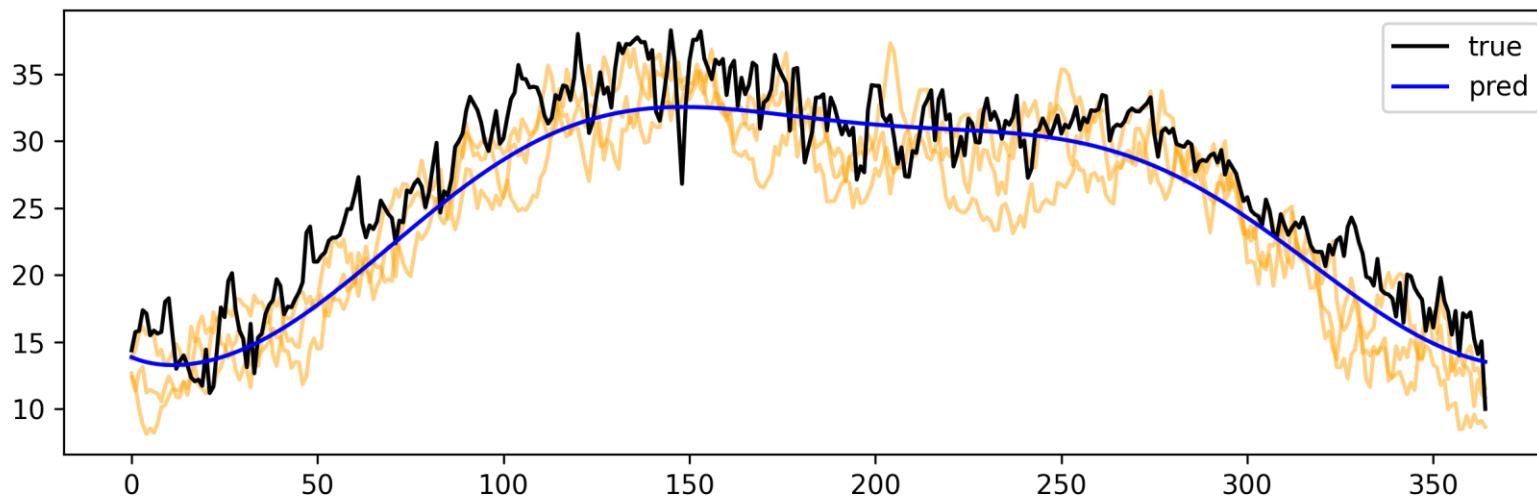
Принципиально более сложный случай — $T \subset \mathbb{R}$ или $T \subset \mathbb{R}^k$. Обычно под случайными процессами понимают процессы именно на таких T . Процессы при $T \subset \mathbb{R}^k$, $k \geq 2$, часто называют **случайными полями**.



Случайные процессы как траектории

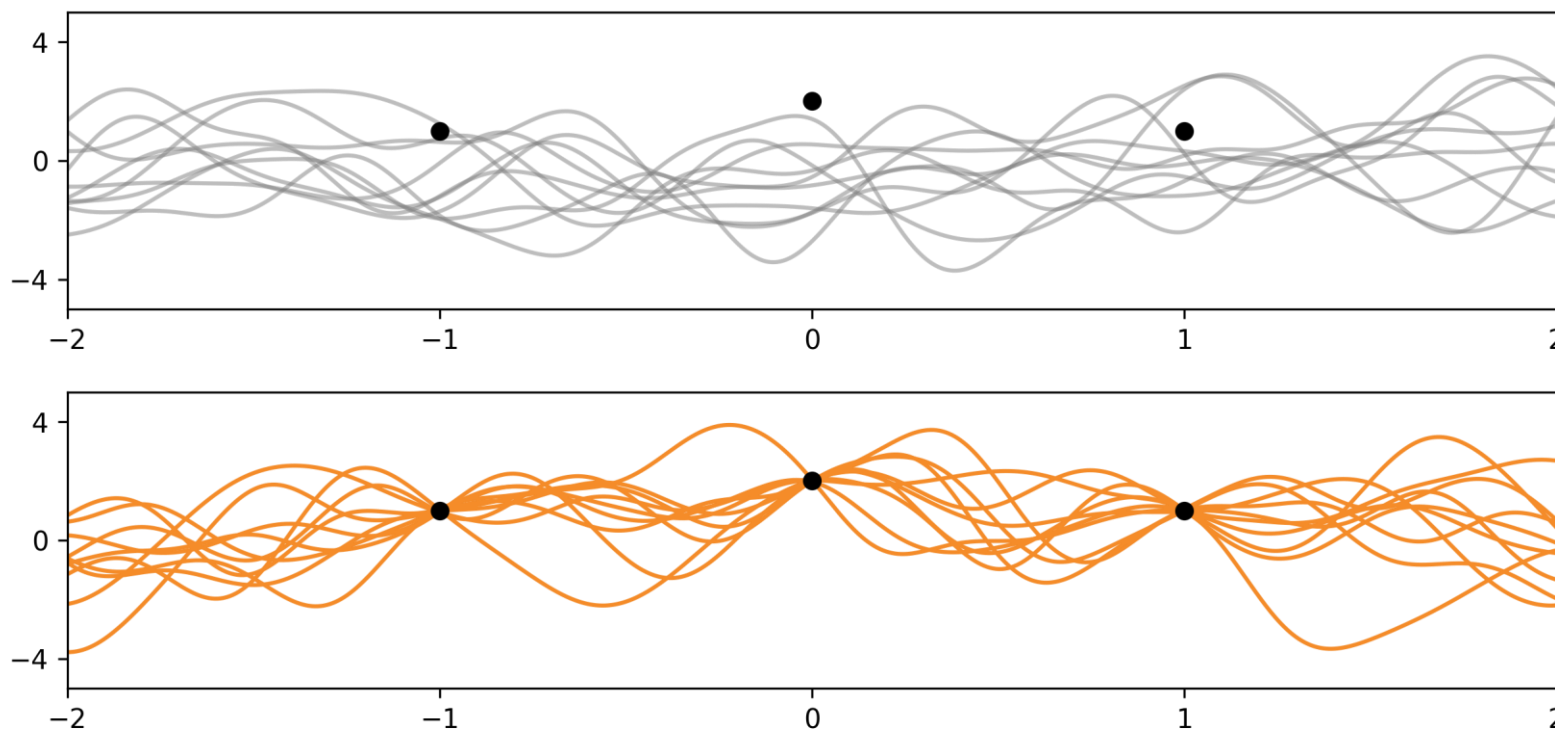
На случайные процессы можно смотреть двумя способами:

1. **Как на семейство случайных величин**, для которых мы знаем все конечномерные распределения $(X_{t_1}, \dots, X_{t_k})$. Это не очень наглядно
2. **Как на случайную функцию**: для каждого исхода ω все случайные величины $X_t(\omega)$ чему-то равны и мы получаем функцию $t \mapsto X_t(\omega)$, которую называют **траекторией процесса**. При разных исходах получаются разные траектории



Случайные процессы как траектории

Если мы пронаблюдали какие-то значения переменных, то мы можем рассмотреть случайный процесс при условии наблюдений. Если исходный процесс задает распределение на всех на свете траекториях, то новый процесс будет задавать **распределение только на тех траекториях, что проходят через наблюдения!**



Гауссовские процессы

Случайный процесс называется **гауссовским**, если для любых $t_1, \dots, t_n$ вектор $(X_{t_1}, \dots, X_{t_n})$ имеет многомерное нормальное распределение.

Гауссовский процесс полностью описывается своим **матожиданием** $m(t) := \mathbb{E}X_t$ и **ковариационной функцией** $K(t_1, t_2) = \text{cov}(X_{t_1}, X_{t_2})$. Для любых $t_1, \dots, t_k \in T$ вектор $(X_{t_1}, \dots, X_{t_k})$ имеет распределение $\mathcal{N}(\mu, \Sigma)$, где $\mu = (m(t_1), \dots, m(t_k))^T$, а $\Sigma_{i,j} = K(t_i, t_j)$.

Так как Σ должна быть неотрицательно определена, K должна быть **неотрицательно определенной функцией**: для любых $t_1, \dots, t_k \in T$ и любых $c_1, \dots, c_k \in \mathbb{R}$ должно быть выполнено

$$\sum_i \sum_j c_i c_j K(t_i, t_j) \geq 0.$$



Гауссовские процессы

Гауссовский процесс с матожиданием m и ковариационной функцией K обозначают с помощью $GP(m, K)$.

Пара свойств:

- если $f \sim GP(m, K)$, то $f(t) = m(t) + g(t)$, где $g \sim GP(0, K)$
- если $f \sim GP(m_1, K_1)$ и $g \sim GP(m_2, K_2)$ независимы, то $f + g \sim GP(m_1 + m_2, K_1 + K_2)$

Последнее свойство удобно для моделирования: можно подбирать свой процесс для каждого эффекта, а потом все сложить. Скажем, один процесс для тренда, другой для сезонности, третий для ошибок измерений и т.д.



Свойства ковариационных функций

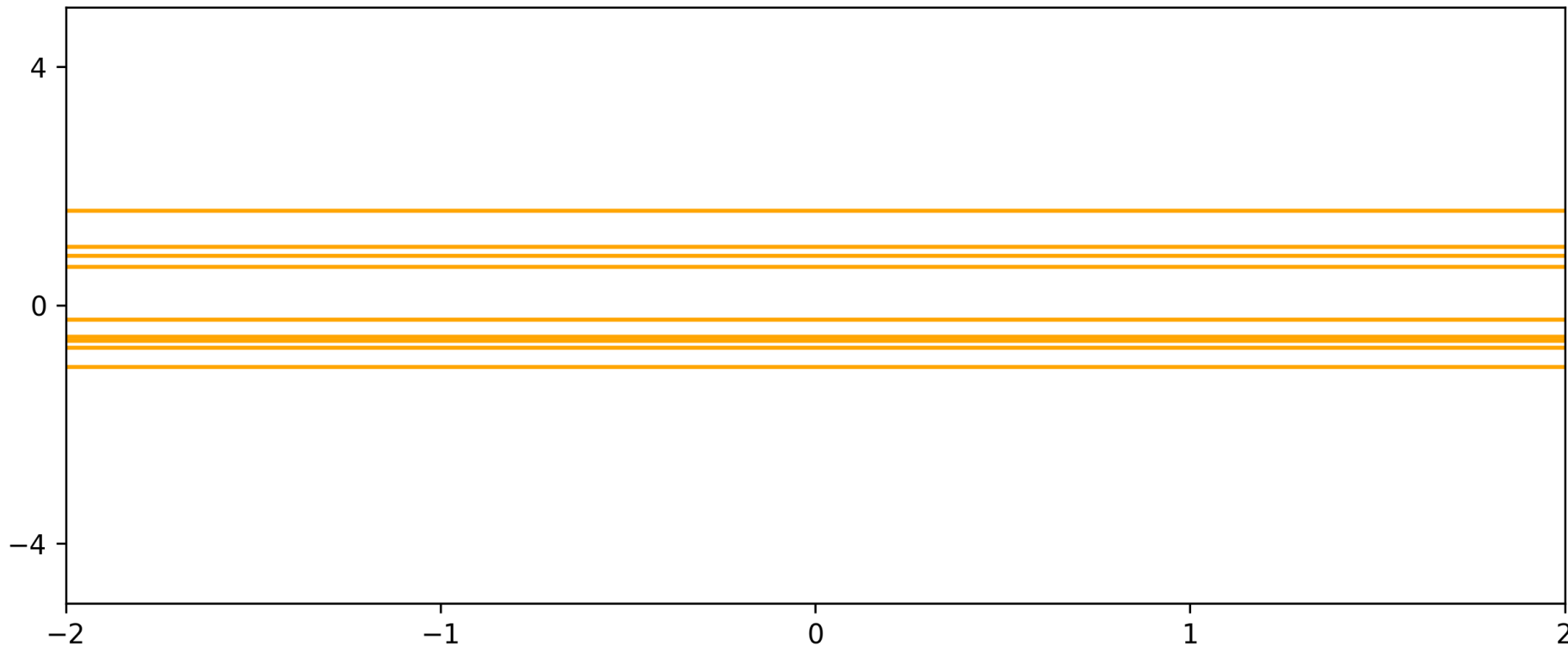
Если K_1 и K_2 это ковариационные функции, а $\alpha_1, \alpha_2 \geq 0$, то

- $\alpha_1 K_1 + \alpha_2 K_2$ это ковариационная функция
- $K_1 K_2$ это ковариационная функция



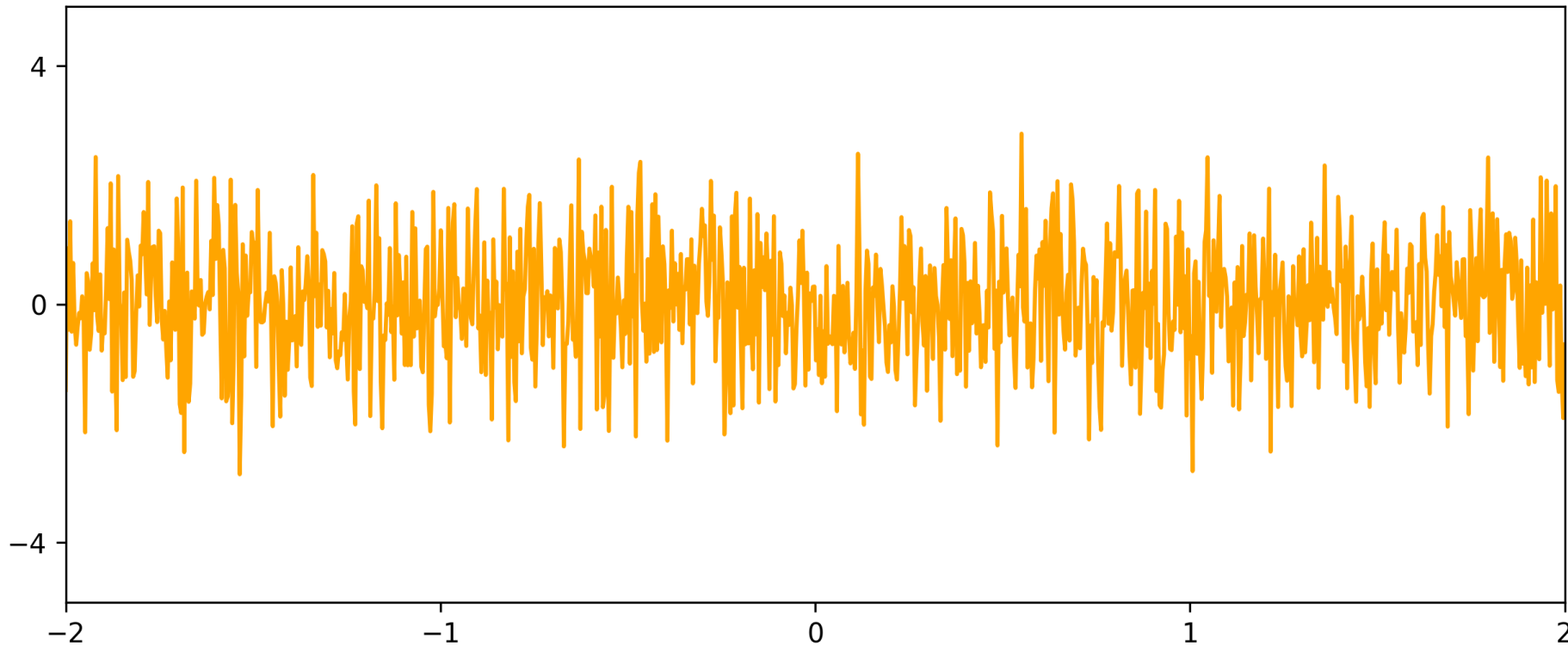
Примеры ковариационных функций

Постоянная ковариационная функция $K(s, t) \equiv \text{const} \geq 0$ порождает процесс с постоянными траекториями:



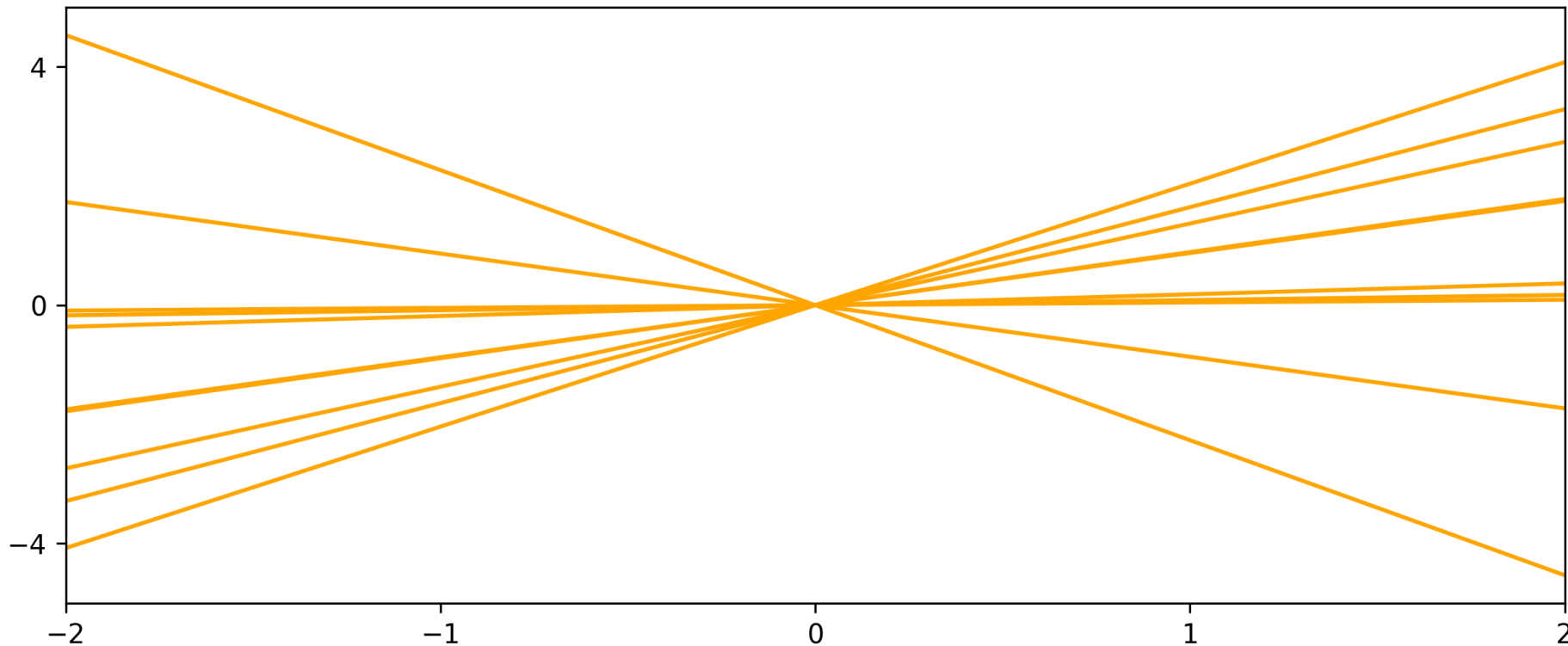
Примеры ковариационных функций

Ковариационная функция **Дирака** $K(s, t) = \delta_{s, t}$ порождает так называемый **белый шум**. С помощью этой ковариационной функции удобно моделировать **погрешность измерений**.



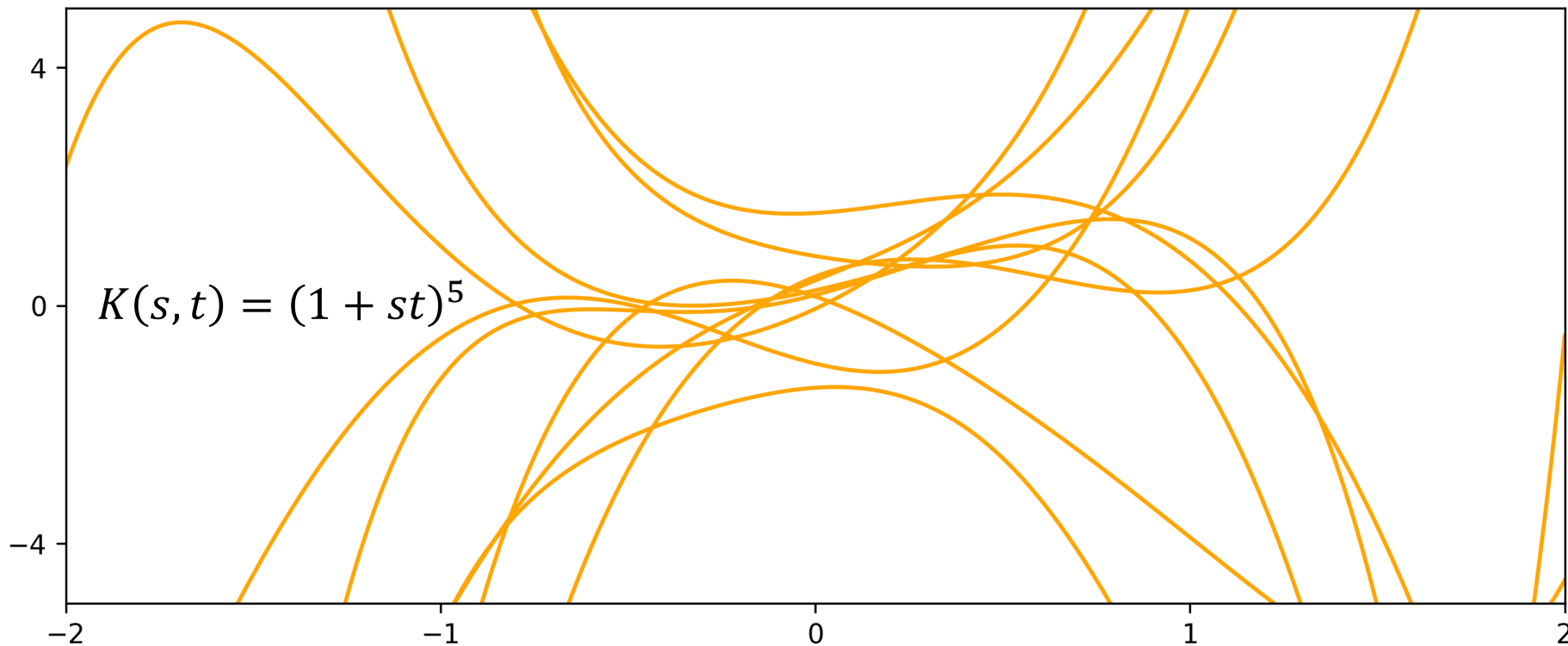
Примеры ковариационных функций

Линейная ковариационная функция $K(s, t) = \langle s, t \rangle$ порождает прямые проходящие через 0. Если хочется породить просто прямые, то можно добавить константу.



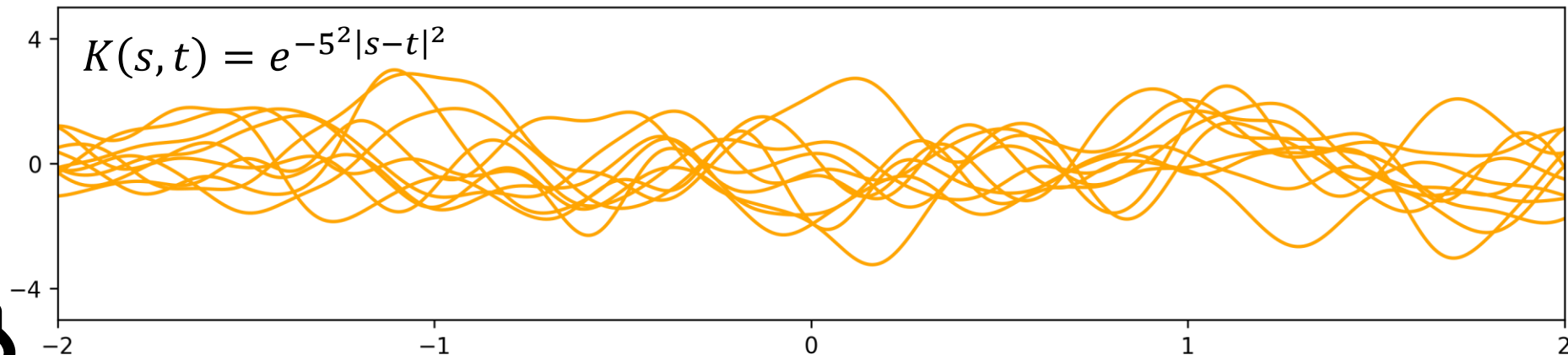
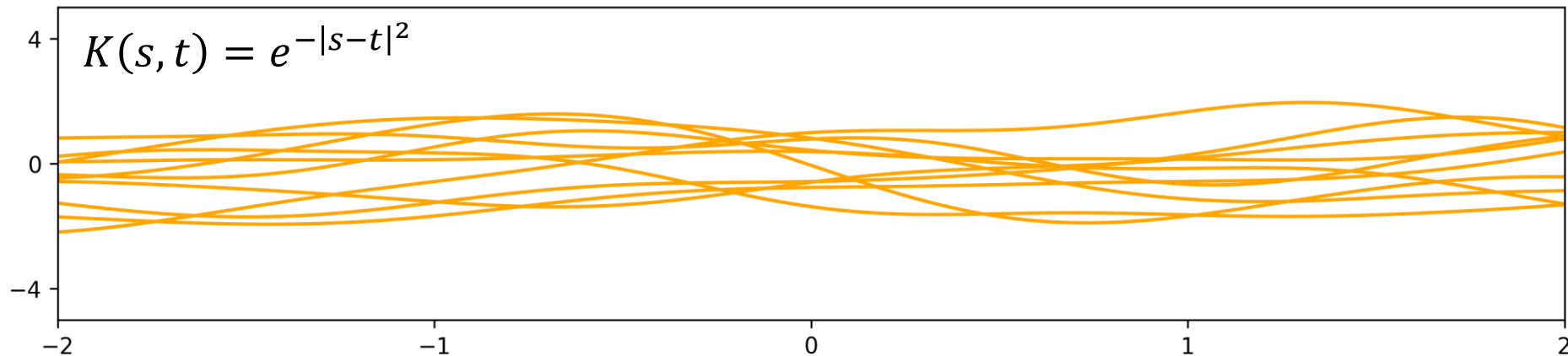
Примеры ковариационных функций

Полиномиальная ковариационная функция $K(s, t) = \langle s, t \rangle^k$ порождает однородные полиномы степени k . Если хочется породить и неоднородные, берем $K(s, t) = (1 + \alpha \langle s, t \rangle)^k$, $\alpha > 0$.



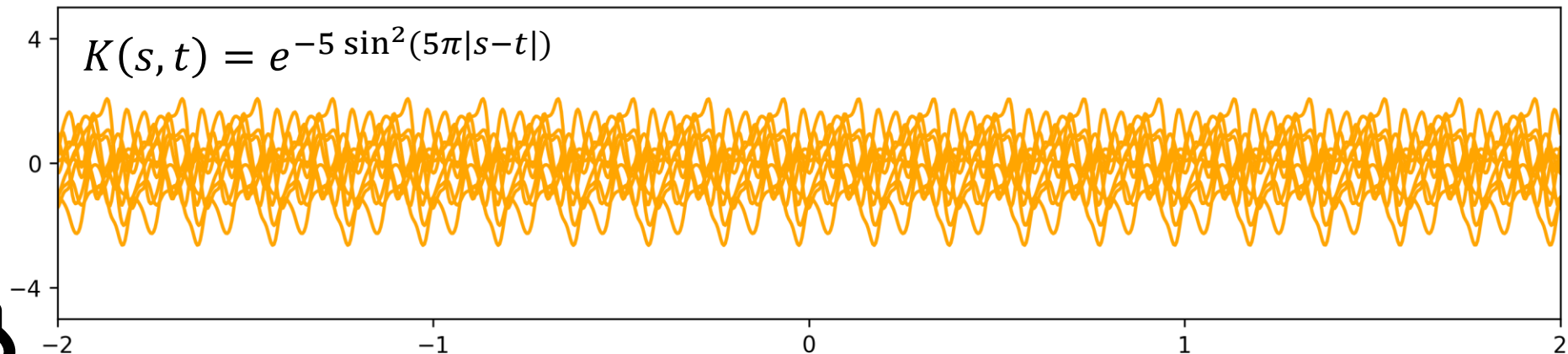
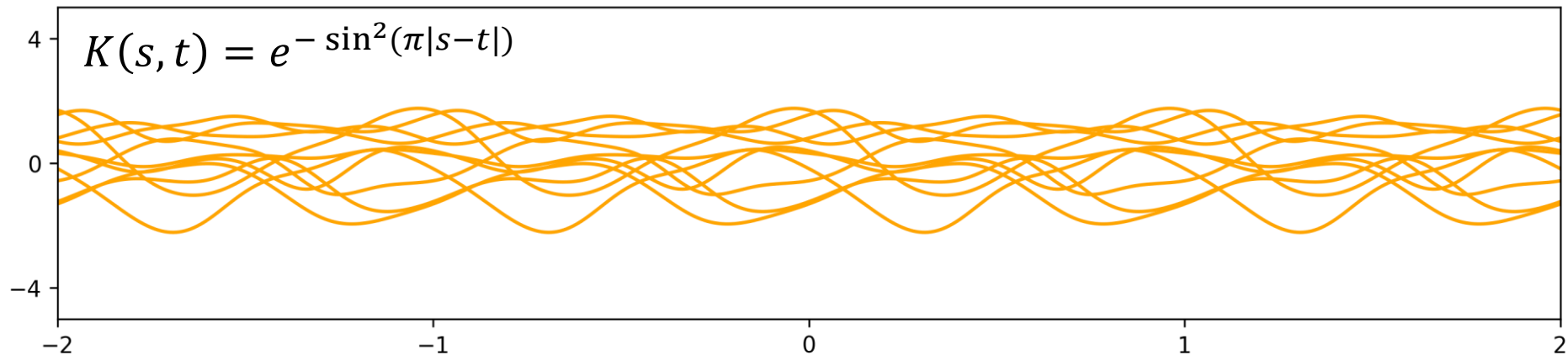
Примеры ковариационных функций

Squared exponential ковариационная функция $K(s, t) = e^{-|s-t|^2/2\ell^2}$. Параметр ℓ отвечает за то, насколько быстро уменьшается корреляция с ростом расстояния между точками.



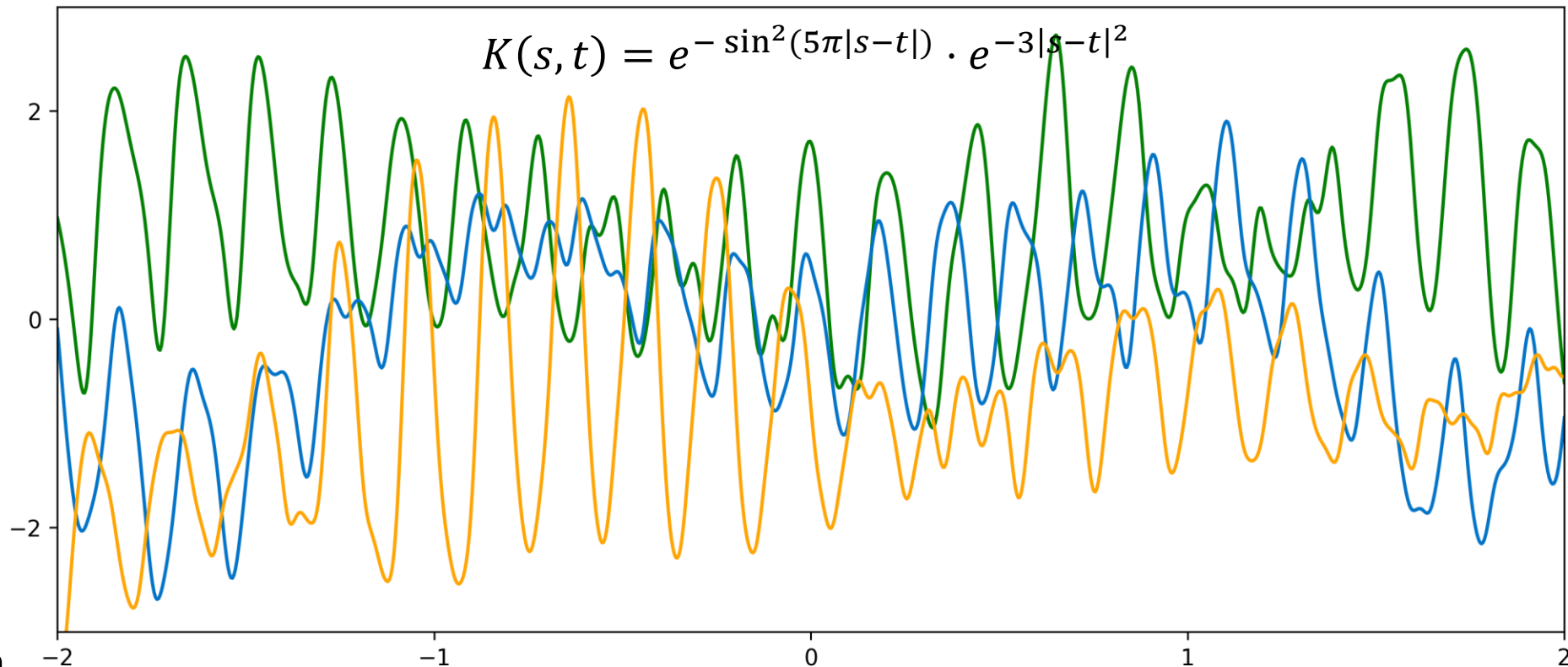
Примеры ковариационных функций

Периодическая ковариационная функция $K(s, t) = \exp\left(-\frac{2 \sin^2\left(\frac{\pi|s-t|}{\omega}\right)}{\ell^2}\right)$. Параметр ℓ выполняет ту же функцию, что и раньше, а ω отвечает за частоту.



Примеры ковариационных функций

Локально периодическая ковариационная функция $K(s, t)$ получается как произведение squared exponential и периодической. Получается что-то периодическое, но с нюансами.



Спектральный взгляд на вещи

Предположим, что T конечно. Это не супер сильное умаление общности применительно к реальной жизни, потому что любой процесс мы можем вычислить не более чем в конечном числе точек.

В этом случае ковариационная функция задается матрицей K такой, что $K(s, t) = K_{s,t}$, а все на свете матрицы ковариации для наборов индексов являются ее подматрицами.



Спектральный взгляд на вещи

Предположим, что T конечно. Это не супер сильное умаление общности применительно к реальной жизни, потому что любой процесс мы можем вычислить не более чем в конечном числе точек.

В этом случае ковариационная функция задается матрицей K такой, что $K(s, t) = K_{s,t}$, а все на свете матрицы ковариации для наборов индексов являются ее подматрицами.

Ясно, что K симметрична и неотрицательно определена. Тогда K диагонализуема: существует ортогональная матрица C и диагональная матрица Λ такие, что

$$K = C\Lambda C^T.$$

Диагональ матрицы Λ состоит из собственных чисел $\lambda_1, \lambda_2, \dots$. Все λ_i неотрицательны и можно считать, что $\lambda_1 \geq \lambda_2 \geq \dots$.

Спектральный взгляд на вещи

$$K = C\Lambda C^T.$$

Диагональ матрицы Λ состоит из собственных чисел $\lambda_1, \lambda_2, \dots$. Все λ_i неотрицательны и можно считать, что $\lambda_1 \geq \lambda_2 \geq \dots$.

Столбцы $v_1, \dots, v_n$ матрицы C образуют **ортонормированный базис**, при этом $K = \sum_i \lambda_i v_i v_i^T$.



Спектральный взгляд на вещи

$$K = C\Lambda C^T.$$

Диагональ матрицы Λ состоит из собственных чисел $\lambda_1, \lambda_2, \dots$. Все λ_i неотрицательны и можно считать, что $\lambda_1 \geq \lambda_2 \geq \dots$.

Столбцы $v_1, \dots, v_n$ матрицы C образуют **ортонормированный базис**, при этом $K = \sum_i \lambda_i v_i v_i^T$.

Пусть $Z \sim \mathcal{N}(0, \Lambda)$. Заметим, что координаты Z_i некоррелированы, а значит независимы.

Рассмотрим случайный вектор $V = \sum_i Z_i v_i = CZ$. Ясно, что $\mathbb{E}V = 0$, а

$$\text{cov}(V) = C \text{cov}(Z) C^T = C\Lambda C^T = K.$$



Спектральный взгляд на вещи

$$K = C\Lambda C^T.$$

Диагональ матрицы Λ состоит из собственных чисел $\lambda_1, \lambda_2, \dots$. Все λ_i неотрицательны и можно считать, что $\lambda_1 \geq \lambda_2 \geq \dots$.

Столбцы $v_1, \dots, v_n$ матрицы C образуют **ортонормированный базис**, при этом $K = \sum_i \lambda_i v_i v_i^T$.

Пусть $Z \sim \mathcal{N}(0, \Lambda)$. Заметим, что координаты Z_i некоррелированы, а значит независимы.

Рассмотрим случайный вектор $V = \sum_i Z_i v_i = CZ$. Ясно, что $\mathbb{E}V = 0$, а

$$\text{cov}(V) = C \text{cov}(Z) C^T = C\Lambda C^T = K.$$

Таким образом, $V \sim \text{GP}(0, K)$ и мы представили гауссовский процесс как **линейную комбинацию конкретных «функций» v_i со случайными и независимыми коэффициентами Z_i** .

Спектральный взгляд на вещи

Таким образом, $V \sim \text{GP}(0, K)$ и мы представили гауссовский процесс как линейную комбинацию конкретных «функций» v_i со случайными и независимыми коэффициентами Z_i .

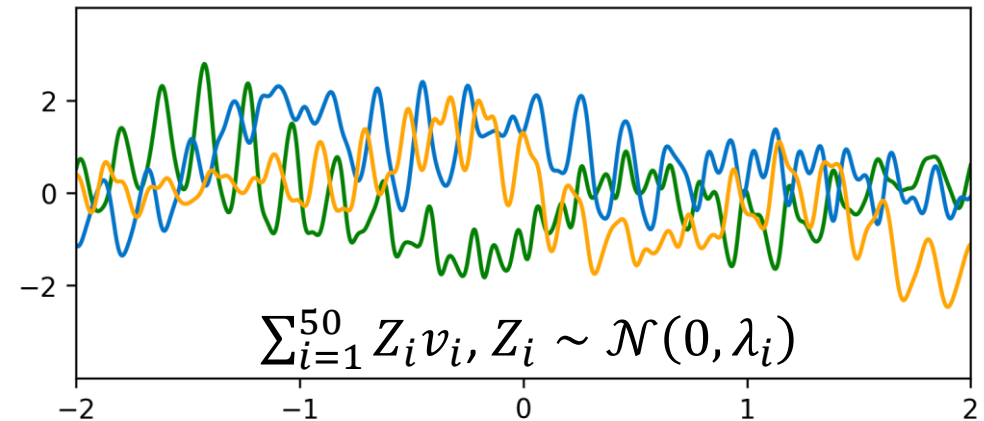
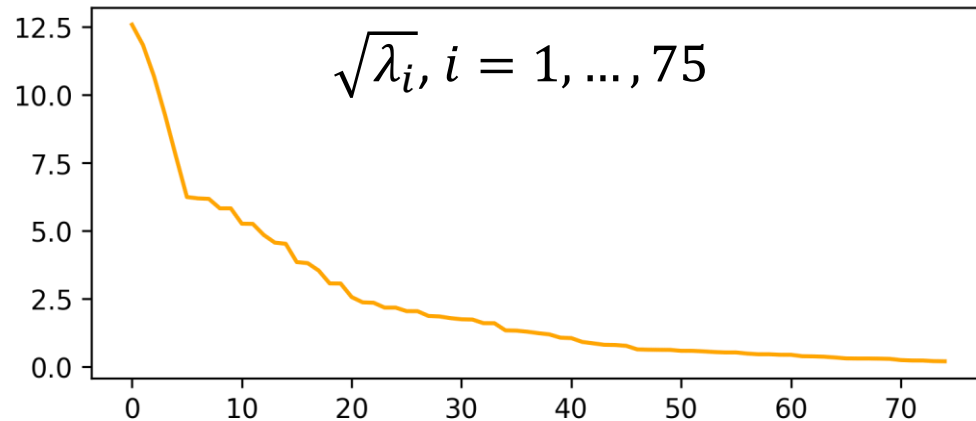
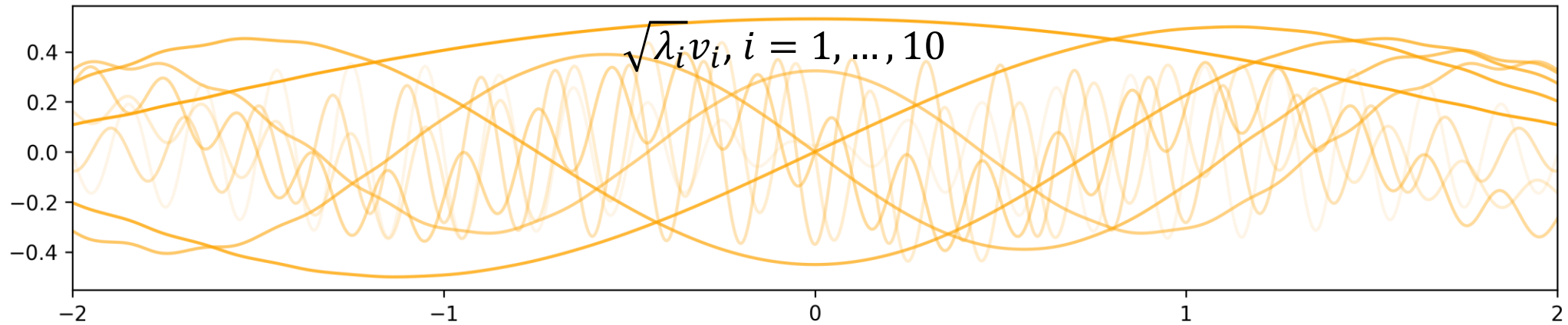
Этот результат верен и в общем случае:

- Утверждение, что неотрицательно определенная функция $K(s, t)$ представима в виде $K(s, t) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(s) \varphi_i(t)$, где $\{\varphi_i\}_{i=1}^{\infty}$ это ортонормированный базис в подходящем функциональном пространстве, носит название **теоремы Мерсера**
- Утверждение, что $f \sim \text{GP}(0, K)$ представимо в виде $f(t) = \sum_{i=1}^{\infty} Z_i \varphi_i(t)$, где $Z_i \sim \mathcal{N}(0, \lambda_i)$ и независимы, носит название **теоремы Карунена—Лоэва**

Эти и подобные теоремы являются классическими результатами функционального анализа и составляют фундамент так называемых **ядерных методов**, которые мы еще обсудим.

Пример для локально периодической функции

Рассмотрим $K(s, t) = e^{-\sin^2(5\pi|s-t|)} \cdot e^{-3|s-t|^2}$ на множестве `np.linspace(-2, 2, 1000)`.



Выводы

1. На гауссовские процессы можно смотреть не только как на случайные траектории, но и как на случайные линейные комбинации детерминированных функций
2. Если λ_i быстро убывают, то гауссовский процесс можно аппроксимировать с помощью линейной комбинации первых нескольких базисных векторов/функций. Это может существенно ускорить вычисления



Добавляем ошибку измерений

Очень часто данные, которые мы получаем, являются зашумленными. Обычно это моделируют так: $f = g + \varepsilon$, где $g \sim \text{GP}(m, K)$ — сигнал, который нам интересен, а $\varepsilon \sim \text{GP}(0, \sigma^2 \delta)$ — белый шум с дисперсией σ^2 . Таким образом, $f \sim \text{GP}(m, K + \sigma^2 \delta)$.



Добавляем ошибку измерений

Очень часто данные, которые мы получаем, являются зашумленными. Обычно это моделируют так: $f = g + \varepsilon$, где $g \sim \text{GP}(m, K)$ — сигнал, который нам интересен, а $\varepsilon \sim \text{GP}(0, \sigma^2 \delta)$ — белый шум с дисперсией σ^2 . Таким образом, $f \sim \text{GP}(m, K + \sigma^2 \delta)$.

Ясно, что при этом нас интересует информация именно о g в остальных точках. Иначе говоря,

мы рассматриваем вектор (X, Y) с ковариацией $\begin{pmatrix} \Sigma_X + \sigma^2 I & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}$ и формула для $Y|(X = x)$

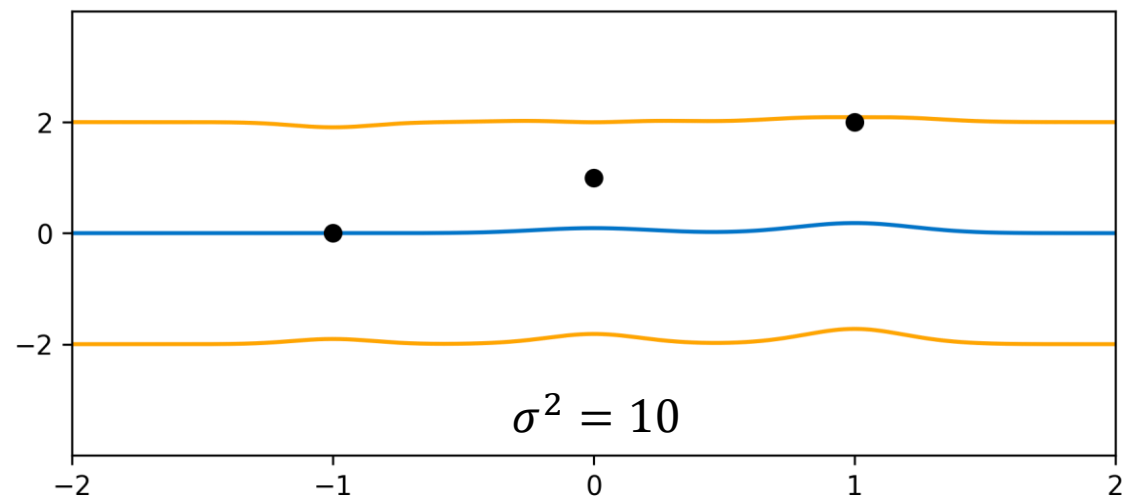
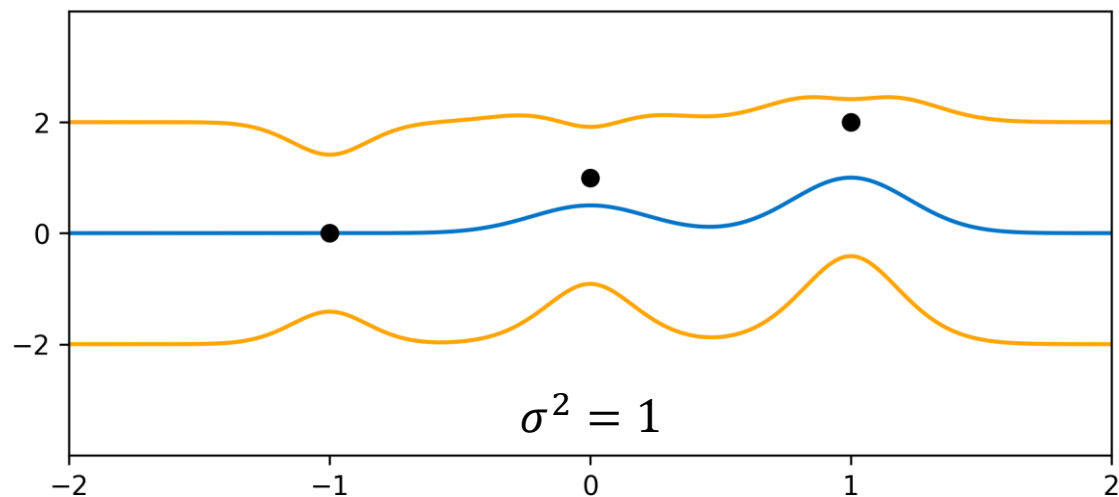
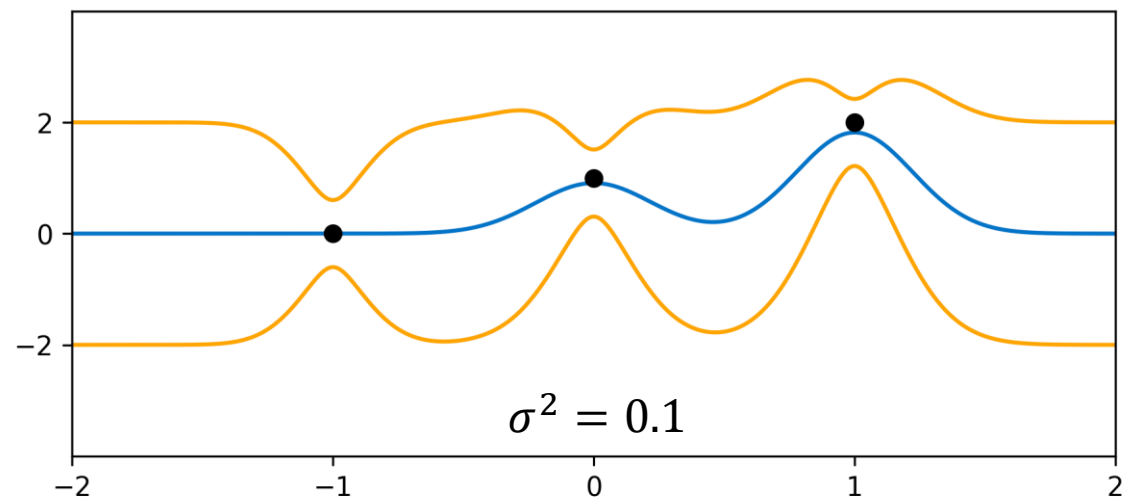
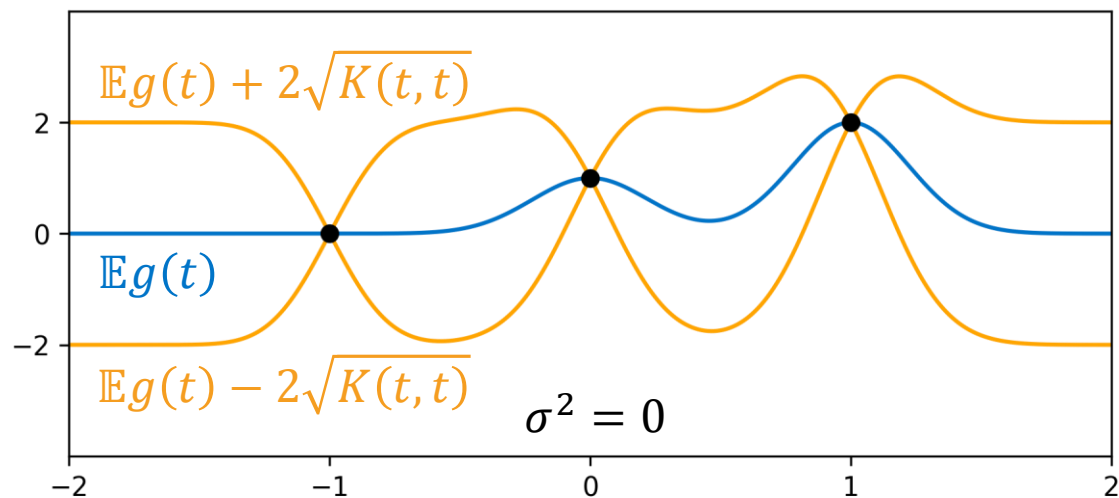
получается такая:

$$Y|(X = x) \sim \mathcal{N}(\mu_Y + \Sigma_{YX}(\Sigma_X + \sigma^2 I)^{-1}(x - \mu_X), \Sigma_Y - \Sigma_{YX}(\Sigma_X + \sigma^2 I)^{-1}\Sigma_{XY}).$$

Из интересного: матрица $\Sigma_X + \sigma^2 I$ всегда обратима.

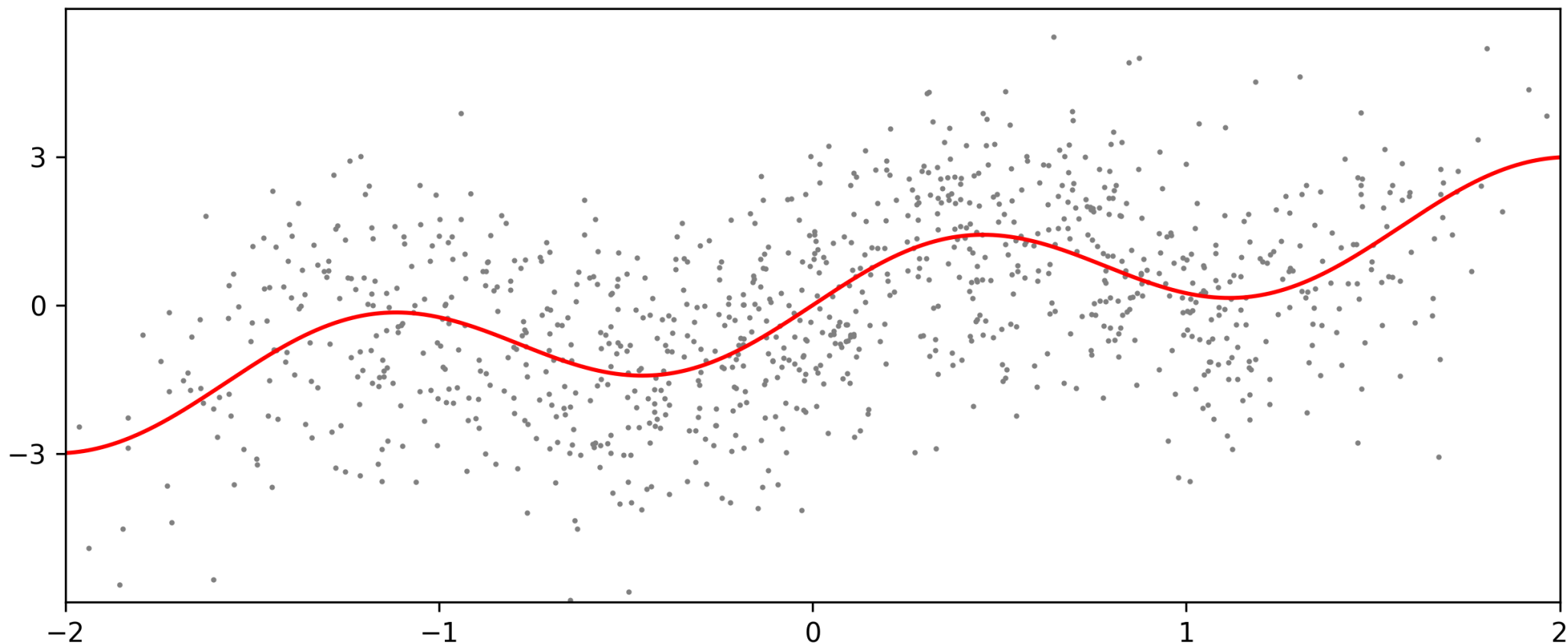
Добавляем ошибку измерений

Рассмотрим пример $g \sim \text{GP}(0, K)$ с $K(s, t) = e^{-10|s-t|^2}$. Наблюдения: $(-1, 0)$, $(0, 1)$, $(1, 2)$.



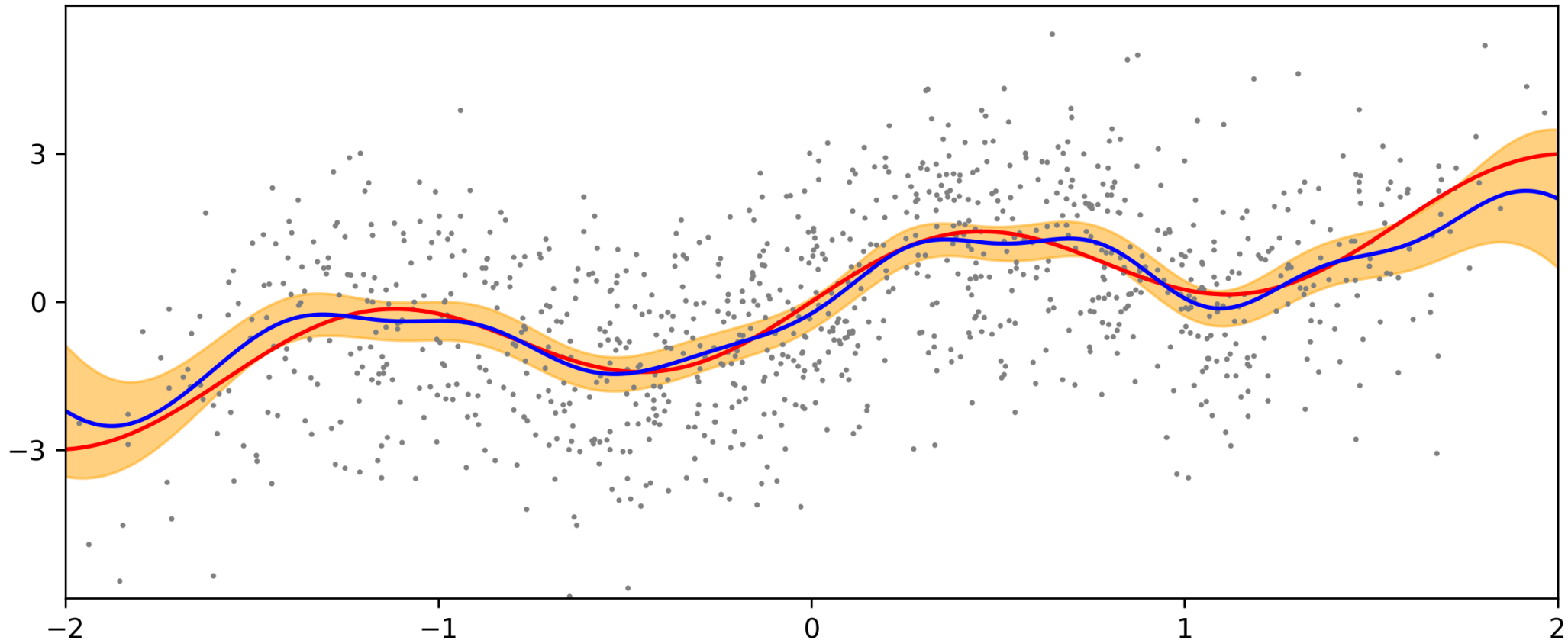
Пример, когда данных побольше

Серое это измерения, красное это сигнал.



Пример, когда данных побольше

Серое это измерения, красное это сигнал. Обучим $GP(0, K)$ с $K(s, t) = e^{-10|s-t|^2}$, $\sigma^2 = 2.5$.



3,5 ложки дегтя

1. Случайные процессы кажутся прямо созданными для однородно устроенных данных: временные ряды, геоданные и т.п. Что делать с **табличными данными**, которые обычно устроены очень неоднородно? Например, для фильмов это название фильма, год выпуска, жанр, главный режиссер, бюджет, сборы, и т.п. Как подбирать для этого всего ковариационные функции?
2. Как подбирать параметры ковариационных функций?
3. Что делать, когда данных реально много?



Предчувствие байесовского подхода

Мы выяснили, что на гауссовские процессы можно смотреть как на такую линейную регрессию, в которой коэффициенты являются случайными величинами:

$$Y(x) = \beta_1 \varphi_1(x) + \dots + \beta_k \varphi_k(x) + \varepsilon,$$

где $\beta \sim \mathcal{N}(0, \Lambda_k)$, а $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Имея набор наблюдений $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, мы можем пересчитать распределение коэффициентов: $\beta \rightarrow \beta|\mathcal{D}$.

Внешне это все очень похоже на **байесовский подход**, и это ж-ж-ж неспроста!

