Гауссовские процессы, часть 1

Максим Кочуров

МГУ им. М.В. Ломоносова

Лекция 4

Содержание

Непараметрические модели

- Предположения обычно весьма условны условны
- Структура (функции) задаётся априорно
- Это характерно не только для Гауссовских процессов

Непараметрические модели

- Предположения обычно весьма условны условны
  - Априорные распределения на функции
  - Априорные распределения на временные или пространственные эффекты
- Структура (функции) задаётся априорно

- Это характерно не только для Гауссовских процессов

Непараметрические модели

- Предположения обычно весьма условны условны
  - Априорные распределения на функции
  - Априорные распределения на временные или пространственные эффекты
- Структура (функции) задаётся априорно
  - Почти не изменяется
  - Волатильна
  - Принимает значения из диапазона между y_0 и y_1
  - Экстраполирует периодически
  - Может иметь другие структурные допущения
- Это характерно не только для Гауссовских процессов

Непараметрические модели

- Предположения обычно весьма условны условны
  - Априорные распределения на функции
  - Априорные распределения на временные или пространственные эффекты
- Структура (функции) задаётся априорно
  - Почти не изменяется
  - Волатильна
  - Принимает значения из диапазона между y_0 и y_1
  - Экстраполирует периодически
  - Может иметь другие структурные допущения
- Это характерно не только для Гауссовских процессов
  - Процесс Дирихле
  - Байесовские аддитивные регессионые деревья (BART)
  - И еще много для чего

Обозначения

$x \in \mathbb{R}^n$, $y \in \mathbb{R}$

$$Y \sim \mathcal{GP}(m(x), k(x, x'))$$

Обозначения

$x \in \mathbb{R}^n$, $y \in \mathbb{R}$

$$Y \sim \mathcal{GP}(m(x), k(x, x'))$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_N) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \dots & k(x_N, x_N) \end{bmatrix} \right)$$

1. $\mathcal{GP}$ Гауссовский процесс - обычно, с $m(x)$ в качестве среднего и ковариационной матрицей $k(x, x')$

Обозначения

$x \in \mathbb{R}^n$, $y \in \mathbb{R}$

$$Y \sim \mathcal{GP}(m(x), k(x, x'))$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_N) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \dots & k(x_N, x_N) \end{bmatrix} \right)$$

1. $\mathcal{GP}$ Гауссовский процесс - обычно, с $m(x)$ в качестве среднего и ковариационной матрицей $k(x, x')$
2. $m(x)$ - функция среднего, например
   - Линейная регрессия $m(x) = x^\top \beta$
   - Константа или ноль $m(x) = c$
   - Другие функции на выбор, например $m(x) = \sin(x)$

Обозначения

$x \in \mathbb{R}^n$, $y \in \mathbb{R}$

$$Y \sim \mathcal{GP}(m(x), k(x, x'))$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_N) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \dots & k(x_N, x_N) \end{bmatrix} \right)$$

1. $\mathcal{GP}$ Гауссовский процесс - обычно, с $m(x)$ в качестве среднего и ковариационной матрицей $k(x, x')$
2. $m(x)$ - функция среднего, например
   - Линейная регрессия $m(x) = x^\top \beta$
   - Константа или ноль $m(x) = c$
   - Другие функции на выбор, например $m(x) = \sin(x)$
3. $k(x, x')$ - ядровая функция, мера схожести между x и x'
   - $[K]_{ij} = k(x_i, x_j)$ симметричная и положительно определенная матрица

Ядровая функция

Напомним, что $\mathcal{GP}(M(x), K(x, x'))$ является, по сути, нормальным распределением. Тогда ядро может быть задано следующим образом:

$$k(x, x') = RBF(x, x') = \exp(-||x - x'||/2L)$$

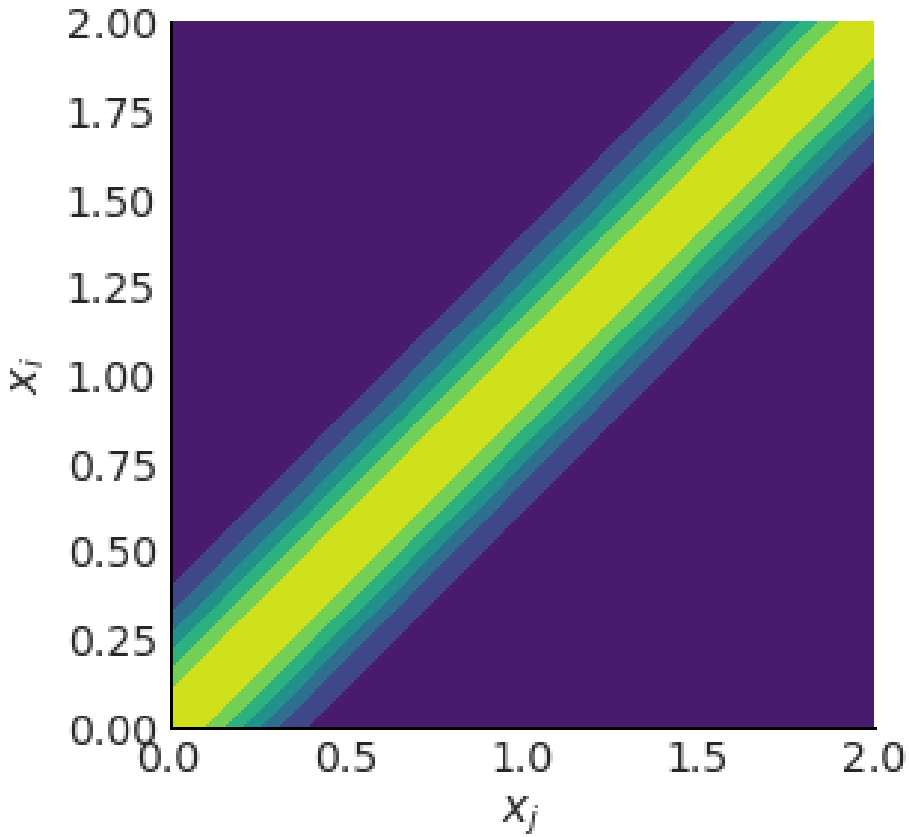


Рис.: Ядро RBF (пространство данных)

Ядровая функция

Напомним, что $\mathcal{GP}(M(x), K(x, x'))$ является, по сути, нормальным распределением. Тогда ядро может быть задано следующим образом:

$$k(x, x') = RBF(x, x') = \exp(-||x - x'||/2L)$$

Интерпретация параметра

L - **масштаб длины** x, что малые изменения не меняют y сильно

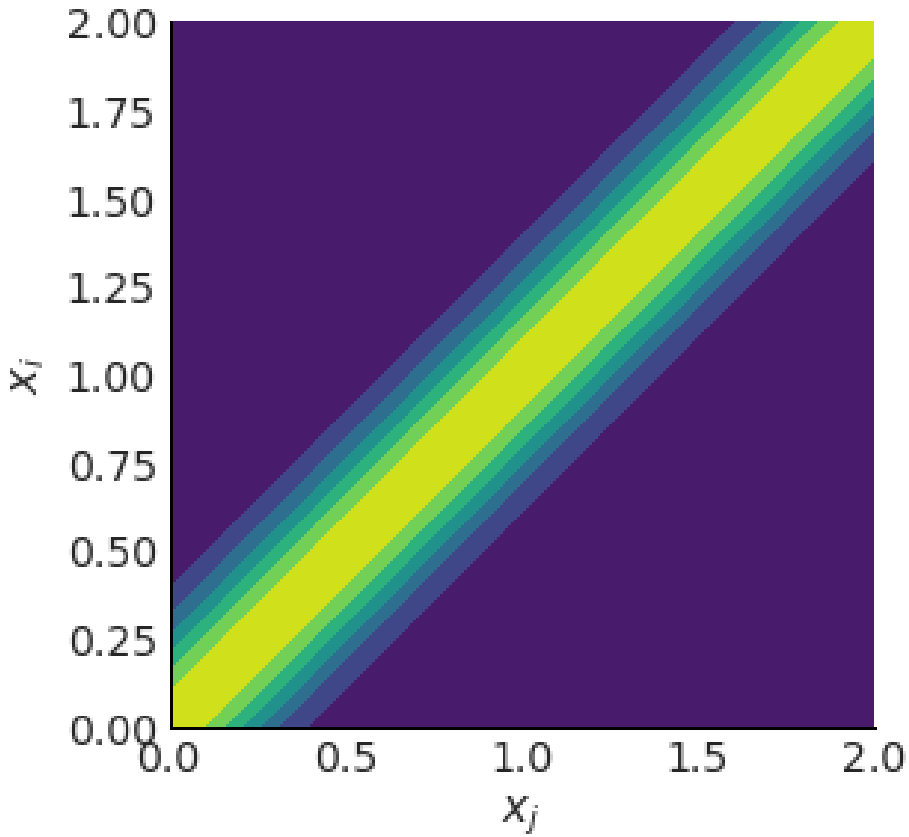


Рис.: Ядро RBF (пространство данных)

Ядровая функция

Напомним, что $\mathcal{GP}(M(x), K(x, x'))$ является, по сути, нормальным распределением. Тогда ядро может быть задано следующим образом:

$$k(x, x') = RBF(x, x') = \exp(-||x - x'||/2L)$$

Интерпретация параметра

L - **масштаб длины** x, что малые изменения не меняют y сильно

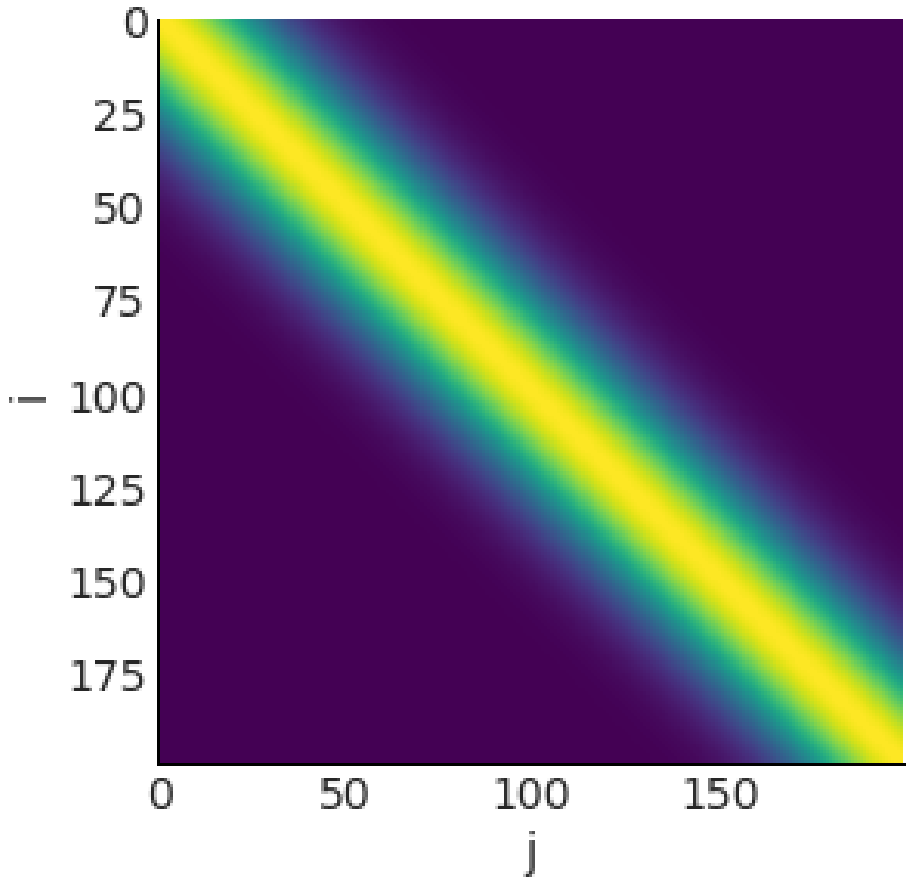


Рис.: Ядро RBF (ковариационная матрица)

Математика ядерных методов

Ядра можно комбинировать (тут подробнее). Если $k_1(x, x')$ и $k_2(x, x')$ удовлетворяют свойствам ядра, то

1. $k_*(x, x') = a \cdot k_1(x, x') + b \cdot k_2(x, x')$ удовлетворяет свойствам ядра
   - Правило сложения
   - $a, b > 0$
2. $k_*(x, x') = k_1(x, x')^a \cdot k_2(x, x')^b$ удовлетворяет свойствам ядра
   - Правило умножения
   - $a, b > 0$

Математика ядерных методов

Ядра можно комбинировать (тут подробнее). Если $k_1(x, x')$ и $k_2(x, x')$ удовлетворяют свойствам ядра, то

1. $k_*(x, x') = a \cdot k_1(x, x') + b \cdot k_2(x, x')$ удовлетворяет свойствам ядра
   - Правило сложения
   - $a, b > 0$
2. $k_*(x, x') = k_1(x, x')^a \cdot k_2(x, x')^b$ удовлетворяет свойствам ядра
   - Правило умножения
   - $a, b > 0$

При параметризации обычно настраиваются следующее

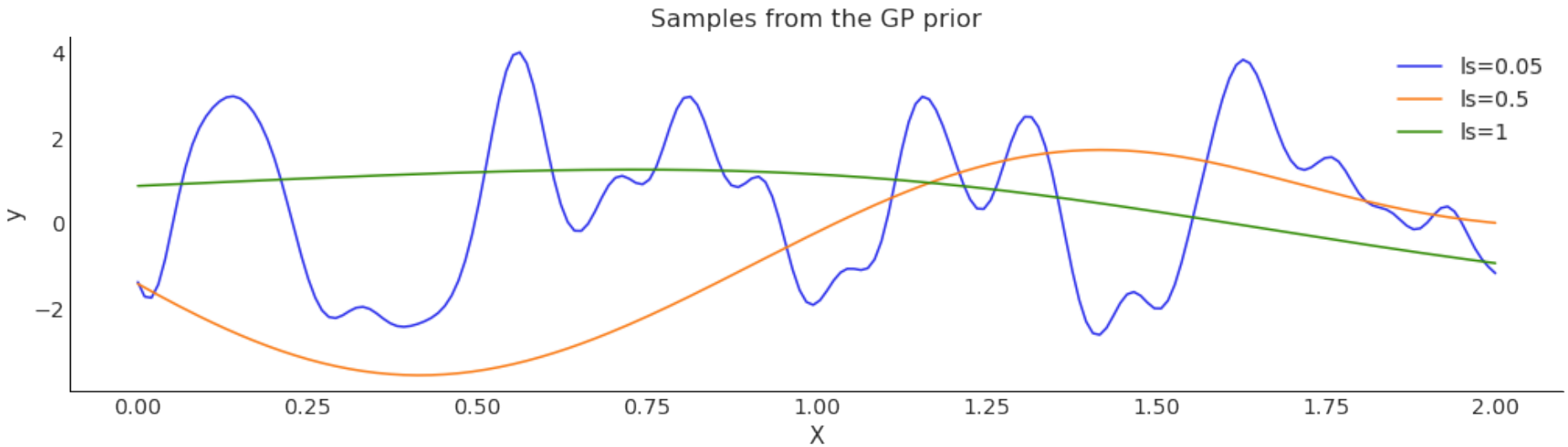
- Белый шум ε
- Амплитуда σ
- Масштаб длины L

$$k(x, x') \cdot \sigma^2 + \varepsilon^2$$

Масштаб длины

- Как **сильно** меняется y
- Не оценивает величину изменений!
- Хорошо подбирается итеративно
- Трудно оценить эмпирически

$$k(x, x') \cdot \sigma^2 + \varepsilon^2$$



Соображению по подбору масштаба длины

- **Периодичность** данных
  - Для годовых данных подойдет L в 1 год
  - Интерполяция на пропущенные наблюдения
  - Интерполяция на более частотные данные (помесячные)

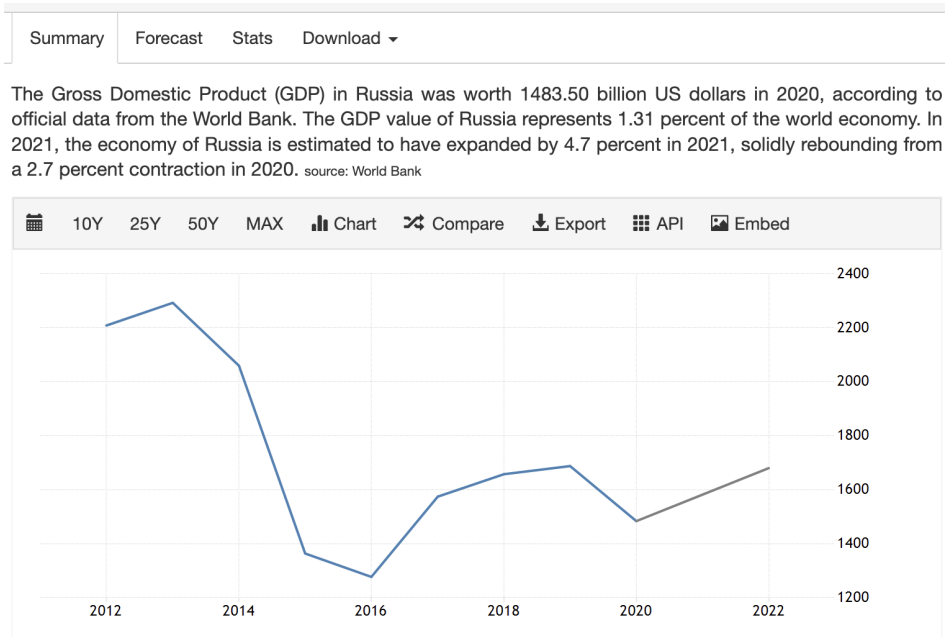


Рис.: ВВП России (tradingeconomics.com)

Соображению по подбору масштаба длины

- **Периодичность** данных
  - Для годовых данных подойдет L в 1 год
  - Интерполяция на пропущенные наблюдения
  - Интерполяция на более частотные данные (помесячные)
- **Другие примеры**
  - Дистанция (км, м, см)
  - Возраст
  - Длительность обучения

Рис.: ВВП России (tradingeconomics.com)

Амплитуда

$$k(x, x') \cdot \sigma^2 + \varepsilon^2$$

- Насколько изменчива зависимая переменная
- Не стандартное отклонение (т.е. не белый шум)
- Априорное распределение подбирается на основе предиктивного

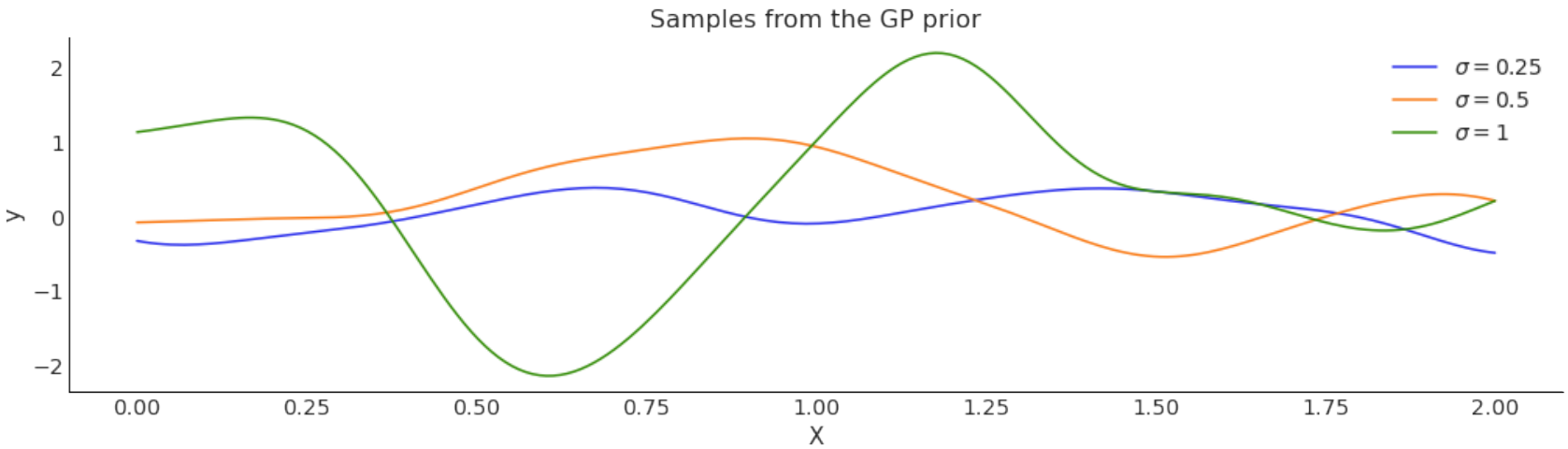


Рис.: Сравнение амплитуд (σ)

Амплитуда и Белый шум

$$k(x, x') \cdot \sigma^2 + \varepsilon^2$$

- Белый шум не стоит путатать с амплитудой

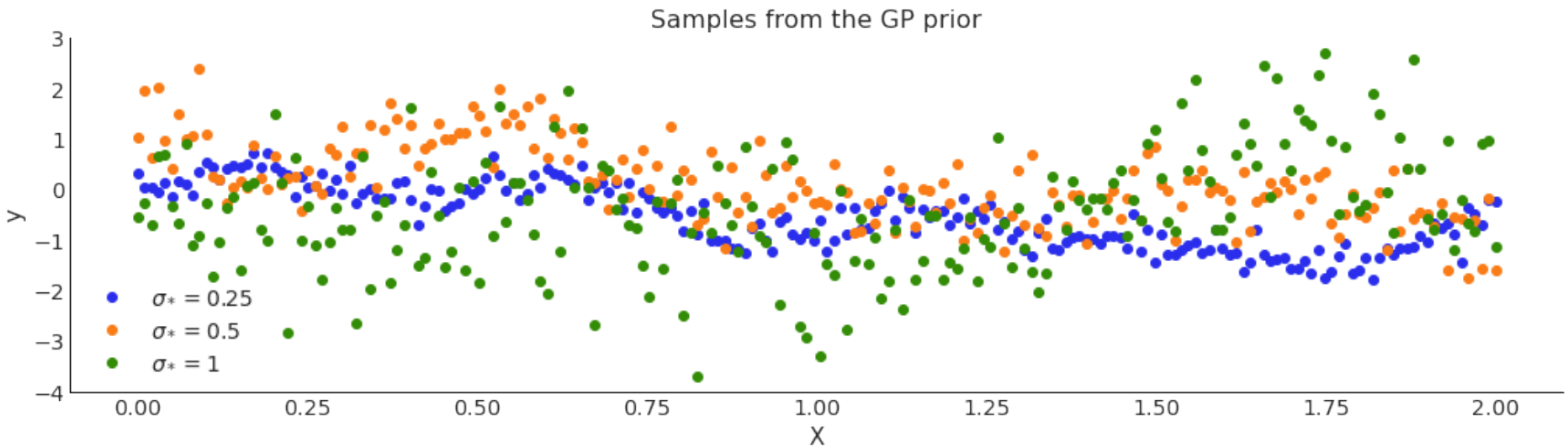


Рис.: Сравнение белого шума (ε)

Подводим итоги

$$k(x, x') = RBF(x, x') \cdot \sigma^2 + \varepsilon^2$$
$$= \exp(-||x - x'||/2L) \cdot \sigma^2 + \varepsilon^2$$

Подводим итоги

- масштаб длины L – размерность исходных данных

$$k(x, x') = RBF(x, x') \cdot \sigma^2 + \varepsilon^2$$
$$= \exp(-||x - x'||/2L) \cdot \sigma^2 + \varepsilon^2$$

Подводим итоги

- масштаб длины L – размерность исходных данных
- амплитуда σ – изменчивость результата

$$k(x, x') = RBF(x, x') \cdot \sigma^2 + \varepsilon^2$$
$$= \exp(-||x - x'||/2L) \cdot \sigma^2 + \varepsilon^2$$

Подводим итоги

- масштаб длины L – размерность исходных данных
- амплитуда σ – изменчивость результата
- ε – белый шум результата

$$k(x, x') = RBF(x, x') \cdot \sigma^2 + \varepsilon^2$$
$$= \exp(-||x - x'||/2L) \cdot \sigma^2 + \varepsilon^2$$

Подводим итоги

- масштаб длины L – размерность исходных данных
- амплитуда σ – изменчивость результата
- ε – белый шум результата

$$k(x, x') = RBF(x, x') \cdot \sigma^2 + \varepsilon^2$$
$$= \exp(-||x - x'||/2L) \cdot \sigma^2 + \varepsilon^2$$

Заметка

Масштаб длины можно вынести из ядровой функции, так как для большинства из них это не внутреннее свойство

$$\exp(-||x - x'||/2L) = \exp(-||x/L - x'/L||/2)$$

Виды ядер

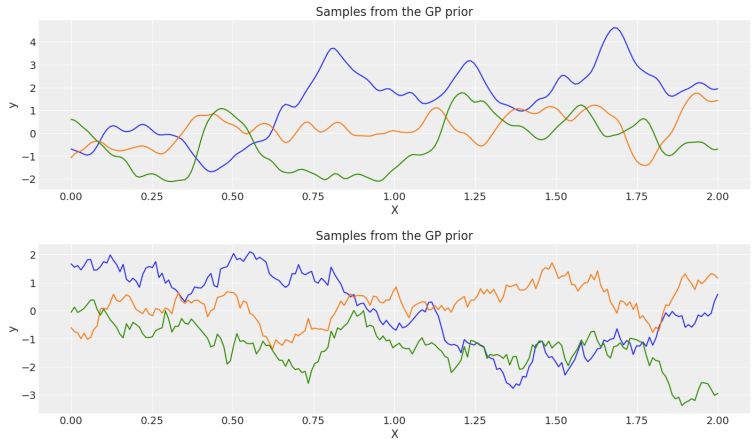
Каждое ядро – это структурное ограничение

- Стационарные
- Периодическое/цикличное
- Линейное/полиномиальное (нестационарные)

Виды ядер

Каждое ядро – это структурное ограничение

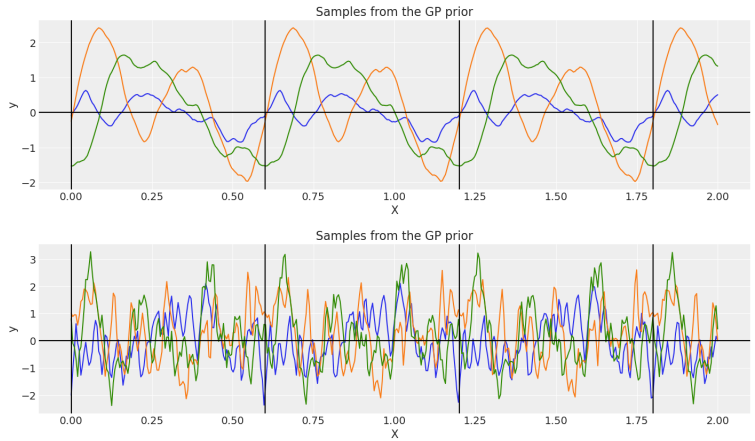
- Стационарные
  - "При экстраполяции распределение вернется к априорному"
- Периодическое/цикличное
- Линейное/полиномиальное (нестационарные)



Виды ядер

Каждое ядро – это структурное ограничение

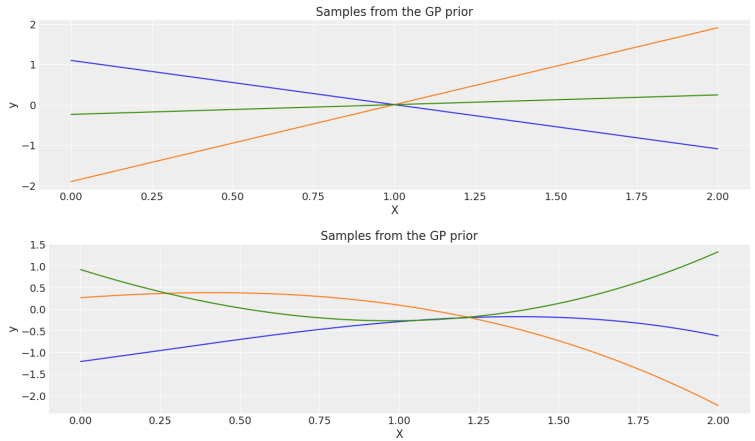
- Стационарные
  - "При экстраполяции распределение вернется к априорному"
- Периодическое/цикличное
  - "Наблюдения неизменны во времени"
- Линейное/полиномиальное (нестационарные)



Виды ядер

Каждое ядро – это структурное ограничение

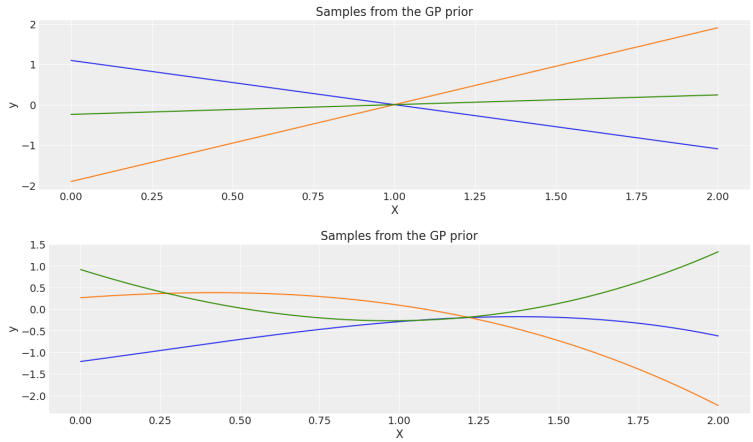
- Стационарные
  - "При экстраполяции распределение вернется к априорному"
- Периодическое/цикличное
  - "Наблюдения неизменны во времени"
- Линейное/полиномиальное (нестационарные)
  - Линейная регрессия
  - Полиномиальная регрессия



Виды ядер

Каждое ядро – это структурное ограничение

- Стационарные
  - "При экстраполяции распределение вернется к априорному"
- Периодическое/цикличное
  - "Наблюдения неизменны во времени"
- Линейное/полиномиальное (нестационарные)
  - Линейная регрессия
  - Полиномиальная регрессия



Математическое преимущество ядер

Можно комбинировать разные семейства ядер. Пример здесь.
Комбинирование ядер – это искусство (разберем на семинаре)

Комбинации ядер

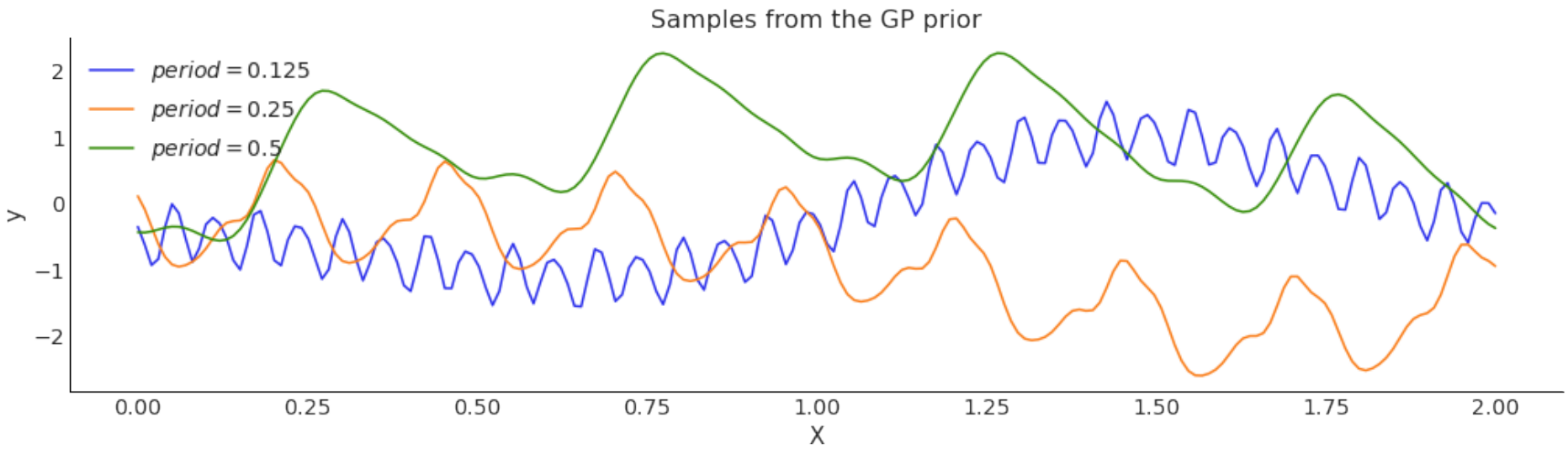


Рис.: Экспоненциальные и периодические ядра

Комбинации ядер

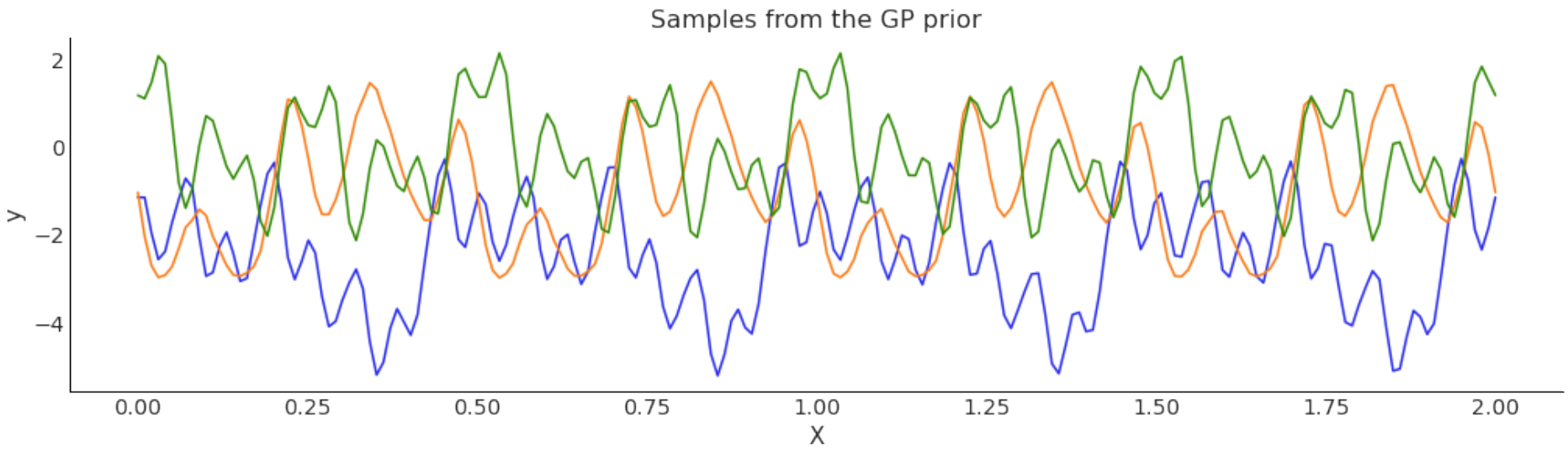


Рис.: Множественные периодические ядра

Комбинации ядер

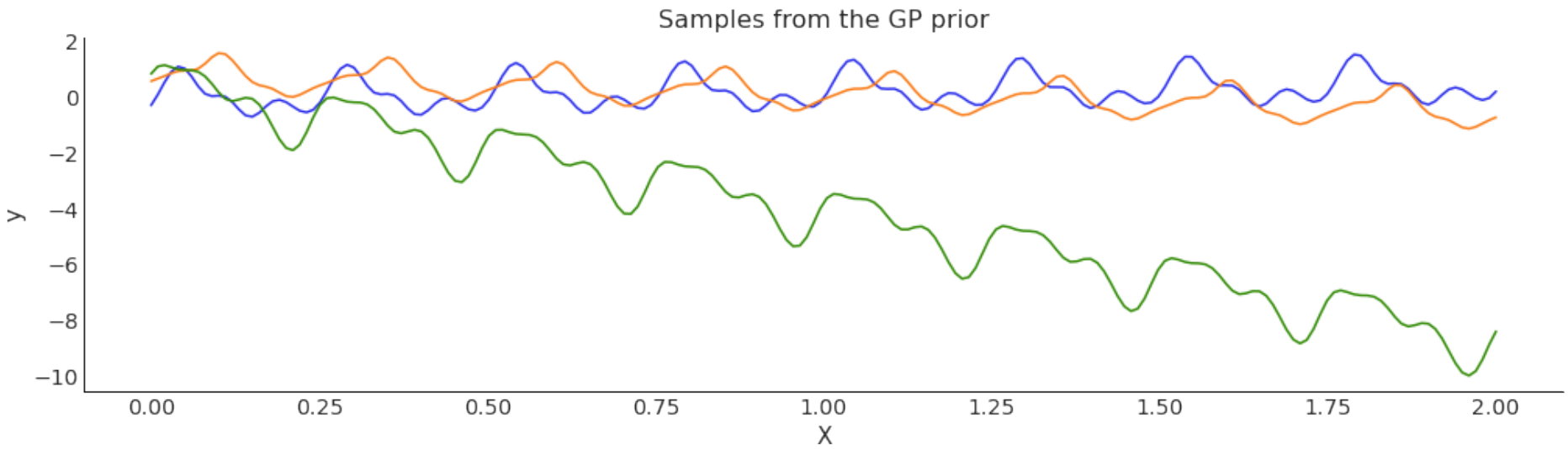


Рис.: Линейные и периодические ядра

Выводы

- Ядра отображают структурные особенности
- Особенности можно вычленить из данных
- Комбинируя ядра можно учитывать несколько особенностей данных

Мотивация

Есть несколько примеров, когда гауссовские процессы – решение проблемы. Например:

- Параметры изменяются во времени
- Данные – временные ряды
- Данные – пространственные
- Данные – пространственные, с панельной структурой

Разбираем пример

Смотрим на данные по 8 школам

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\theta_i \sim \text{Normal}(\mu, \tau)$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

Данные – пары рядов $\{(y_i, \sigma_i)\}$

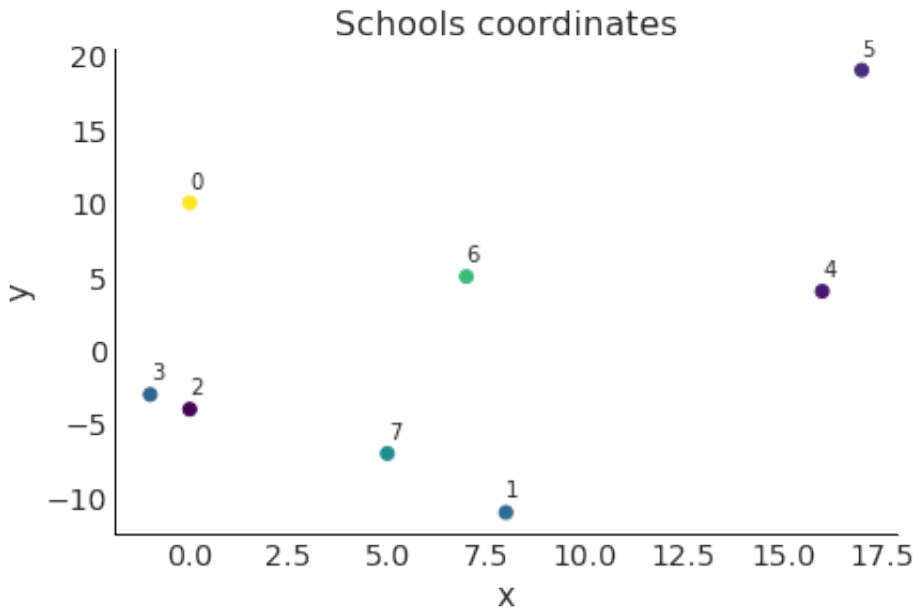
Разбираем пример

Смотрим на данные по 8 школам

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\theta_i \sim \text{Normal}(\mu, \tau)$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

Данные – пары рядов $\{(y_i, \sigma_i)\}$

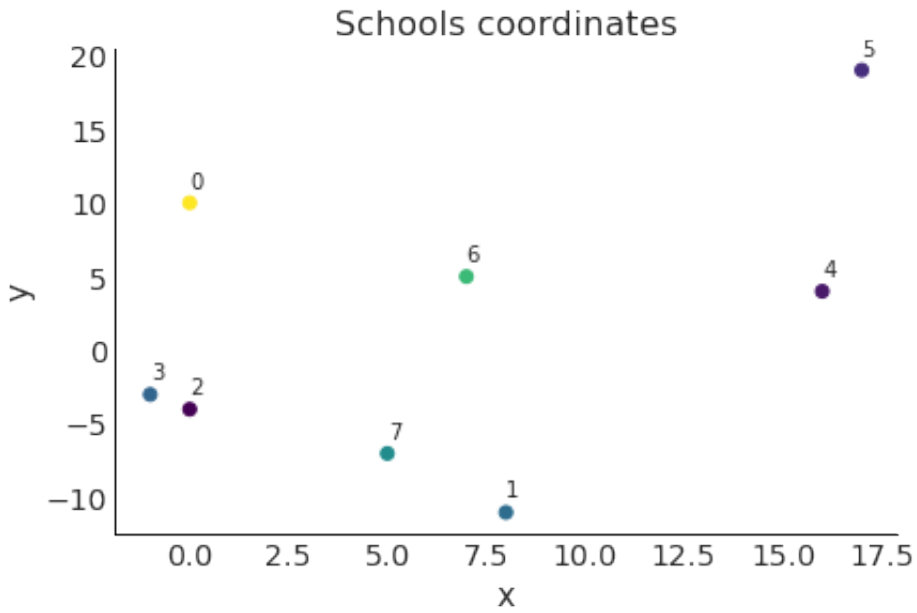
- Что если у нас появится дополнительная информация?
- Учтем расположение школ?



Разбираем пример

Смотрим на данные по 8 школам

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\theta_i \sim \text{Normal}(\mu, \tau)$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$



Данные – пары рядов $\{(y_i, \sigma_i)\}$

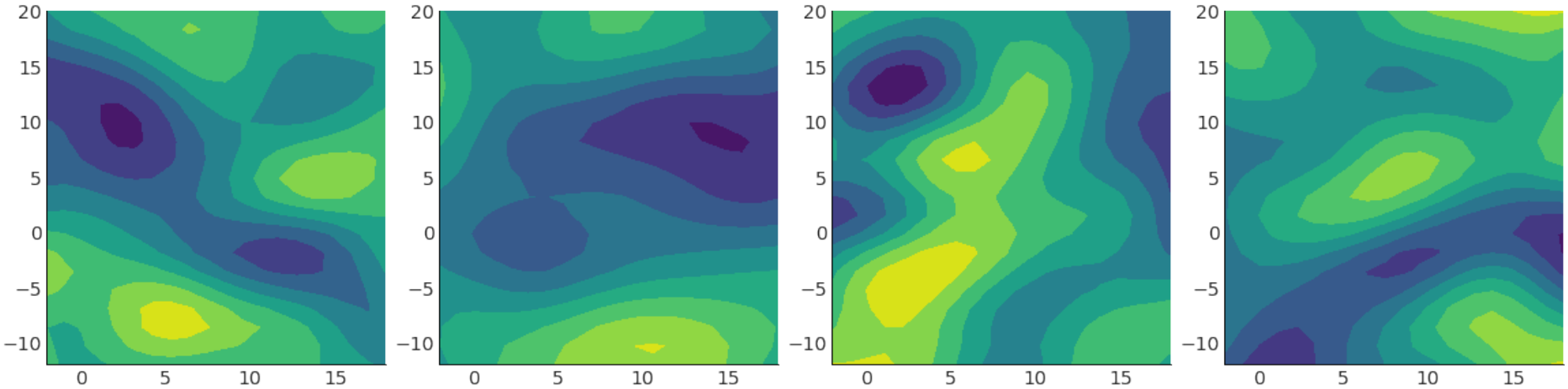
- Что если у нас появится дополнительная информация?
- Учтем расположение школ?

Предположение

Соседние школы похожи

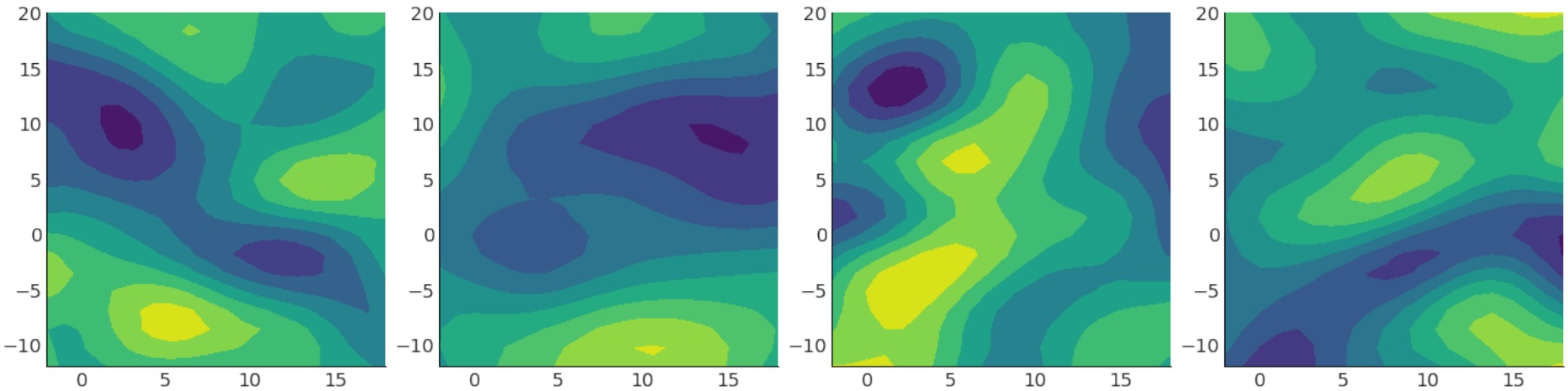
Гауссовский процесс в пространстве

- Гладкая функция в двумерном пространстве
- Масштаб длины важен и его можно вычислить



Гауссовский процесс в пространстве

- Гладкая функция в двумерном пространстве
- Масштаб длины важен и его можно вычислить



Идея

Вместо независимой иерархии применим GP-иерархию!

GP-иерархии

До: (центрированное)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\theta_i \sim \text{Normal}(\mu, \tau)$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

(нецентрированное)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\bar{\theta}_i \sim \text{Normal}(0, 1)$$
$$\theta_i = \mu + \tau \cdot \bar{\theta}_i$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

После: (нецентрированное + GP)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\bar{\theta}_i \sim \mathcal{GP}(x_i)$$
$$\theta_i = \mu + \tau \cdot \bar{\theta}_i$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

GP-иерархии

До: (центрированное)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\theta_i \sim \text{Normal}(\mu, \tau)$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

(нецентрированное)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\bar{\theta}_i \sim \text{Normal}(0, 1)$$
$$\theta_i = \mu + \tau \cdot \bar{\theta}_i$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

После: (нецентрированное + GP)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\bar{\theta}_i \sim \mathcal{GP}(x_i)$$
$$\theta_i = \mu + \tau \cdot \bar{\theta}_i$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

Комментарий

У центрированного распределения есть сложности с геометрией (см. лек. 2)

GP-иерархии

До: (центрированное)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\theta_i \sim \text{Normal}(\mu, \tau)$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

(нецентрированное)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\bar{\theta}_i \sim \text{Normal}(0, 1)$$
$$\theta_i = \mu + \tau \cdot \bar{\theta}_i$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

После: (нецентрированное + GP)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\bar{\theta}_i \sim \mathcal{GP}(x_i)$$
$$\theta_i = \mu + \tau \cdot \bar{\theta}_i$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

Комментарий

Проблемы можно решить неориентированной параметризацией

GP-иерархии

До: (центрированное)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\theta_i \sim \text{Normal}(\mu, \tau)$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

(нецентрированное)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\bar{\theta}_i \sim \text{Normal}(0, 1)$$
$$\theta_i = \mu + \tau \cdot \bar{\theta}_i$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

После: (нецентрированное + GP)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\bar{\theta}_i \sim \mathcal{GP}(x_i)$$
$$\theta_i = \mu + \tau \cdot \bar{\theta}_i$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

Комментарий

В изначальной модели θ_i (или $\bar{\theta}_i$) независимы для разных школ

GP-иерархии

До: (центрированное)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\theta_i \sim \text{Normal}(\mu, \tau)$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

(нецентрированное)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\bar{\theta}_i \sim \text{Normal}(0, 1)$$
$$\theta_i = \mu + \tau \cdot \bar{\theta}_i$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

После: (нецентрированное + GP)

$$\mu \sim \text{Normal}(0, 5)$$
$$\tau \sim \text{HalfCauchy}(5)$$
$$\bar{\theta}_i \sim \mathcal{GP}(x_i)$$
$$\theta_i = \mu + \tau \cdot \bar{\theta}_i$$
$$y_i \sim \text{Normal}(\theta_i, \sigma_i)$$

Комментарий

Гауссовский процесс добавляет зависимость, что делает соседские школы схожими. $\sigma_{\mathcal{GP}} = 1$

Результаты и выводы

Когда гауссовские процессы хороши

1. Гибкая структура
2. Построение непростых иерархий
3. Предсказания для новых наблюдений