Применение Байесовской статистики для анализа наблюдательных данных

Сергей Анфиногентов[1,2]

[1]Институт Солнечно-земной физики СО РАН, Иркутск, Россия
[2]University of Warwick, UK

Онлайн семинар
5 октября 2016

Содержание

Постановка задачи

обозначение

- D — наблюдательные данные
  - Интенсивность излучения
  - Спектральные измерения
  - ...
- $\theta = [\theta_1, \theta_2, \cdots, \theta_N]$ — ненаблюдаемые параметры
  - Магнитное поле
  - Температура
  - ...
- $D_m = M(\theta)$ — модель, позволяющая построить синтетические наблюдения D_m по известным параметрам θ
  - Математическая формула
  - Сложное моделирование

Постановка задачи

Обратная задача

- Получение информации о параметрах модели θ по известным наблюдательным данным D
  - Средние значения
  - Доверительные интервалы
  - Функция распределения вероятности для каждого параметра
- Какая из конкурирующих моделей M_1, M_2, ... M_n лучше описывает наблюдательные данные?
  - Определение предпочтительной модели
  - Количественные критерий для сравнения моделей

Основные понятия

Обозначения

1. $P(\theta)$ — **Плотность распределения априорной вероятности** для параметров модели.
2. $P(D|\theta)$ — Условная **плотность вероятности получения наблюдательных данных** при заданных параметрах модели θ.
3. $P(D|\theta)$ как функция θ — **Функция правдоподобия** определяет степень правдоподобности θ при фиксированных наблюдательных данных D
4. $P(\theta|D)$ — **Апостериорная плотность распределения вероятности** для параметров модели.

Априорная вероятность

Равномерное распределение — наиболее простой вариант

Параметры равномерно распределены внутри заданного диапазона:

$$\theta_i^{\min} \le \theta_i \le \theta_i^{\max}$$

Функция априорного распределения вероятности:

$$P(\theta) = \prod_{i=1}^{N_p} H(\theta_i, \theta_i^{\min}, \theta_i^{\max}),$$

$$H(x, a, b) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

Модель

- Наблюдательные данные $D = [D_1, D_2, ..., D_{N_d}]$
- Ненаблюдаемые параметры $\theta = [\theta_1, \theta_2, \cdots, \theta_{N_p}]$
- Модель $D_i^m(\theta) = f_i(\theta) + N_i(0, \sigma^2), 1 < i < N_d,$
  - $f_i(\theta)$ — теоретическая зависимость
  - $N(0, \sigma)$ — белый шум с нулевым средним и дисперсией σ^2

Функция правдоподобия:

$$P(D|\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N_d}{2}}} \prod_{i=1}^{N_d} \exp\left\{-\frac{[D_i - D_i^m(\theta)]^2}{2\sigma^2}\right\}$$

Теорема Байеса

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

Уточнённое знание $[P(\theta|D)] \Leftarrow$ Знание априори $[P(\theta)]$

$$P(D) = \int P(D|\theta)P(\theta)d\theta$$

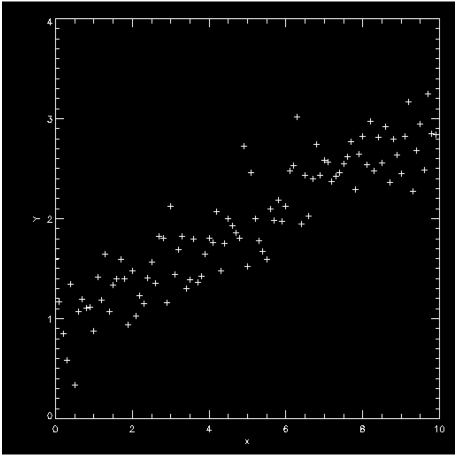
$P(D)$ — нормировочные коэффициент или свидетельство (evidence).

Свидетельство — мера того, насколько хорошо модель M описывает наблюдения D.

Линейная зависимость

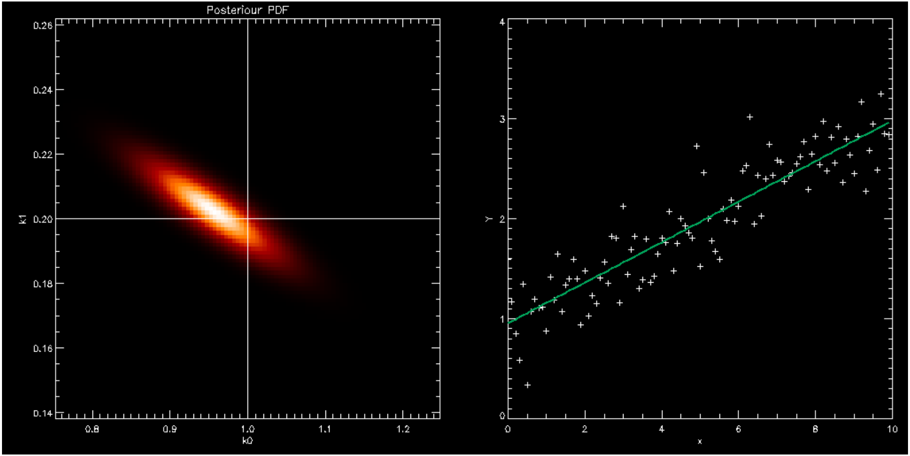
Данные

Модель: $Y(x) = k_0 + k_1 x + N(0, 0.3^2)$



Пример: линейная зависимость

Байесовский анализ



Проклятие размерностей

Для определение доверительных интервалов требуется знать маргинальное распределение параметров:

$$P(\theta_i|D) = \int P(\theta_1, \theta_2, \cdots, \theta_N|D) d\theta_{k \neq i}$$

Проклятие размерности:

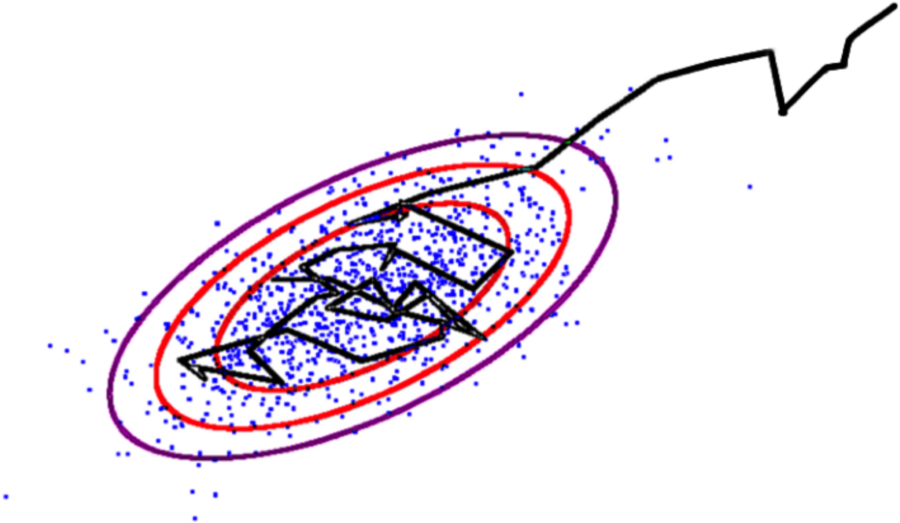
- Аналитическое интегрирование для большинства сложных моделей невозможно.
- При численном интегрировании каждый дополнительный параметр увеличивает время расчета на несколько порядков

Решение:

- Генерация выборки из апостериорного распределения
- Аппроксимация маргинальных распределений гистограммами

Метод Монте-Карло по схеме марковской цепи

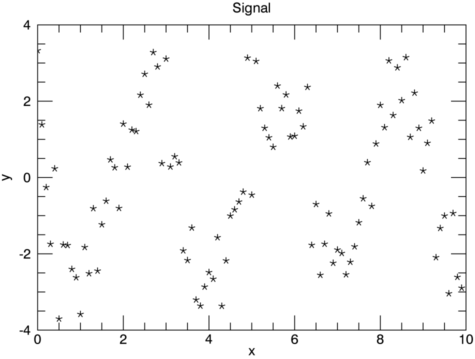
Markov Chain Monte-Carlo



Определение параметров синусоиды

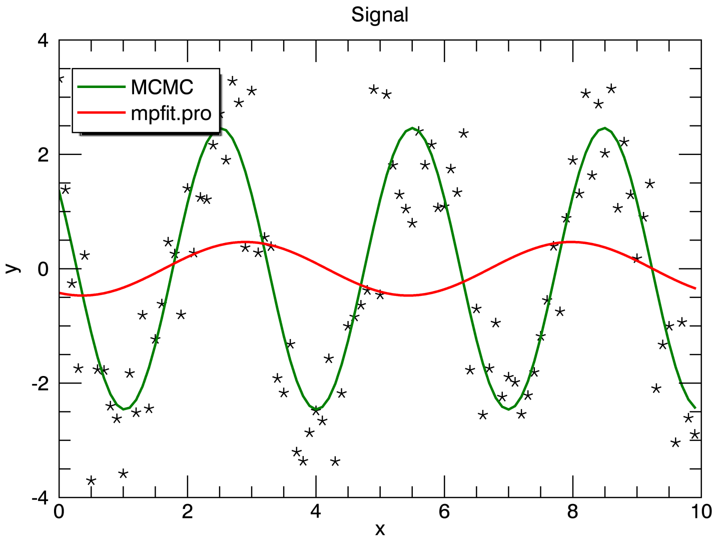
Модельные данные

$$y(x) = A\cos(\frac{2\pi x}{P} + \phi) + N(0, \sigma^2)$$



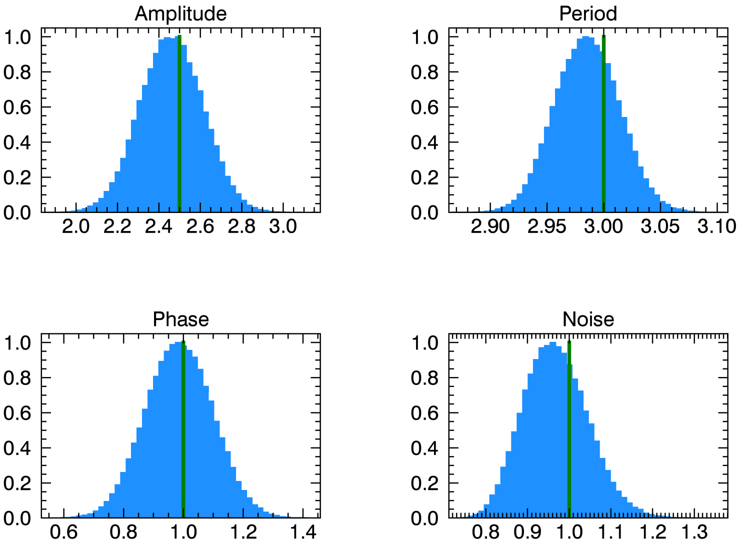
Определение параметров синусоиды

Результаты



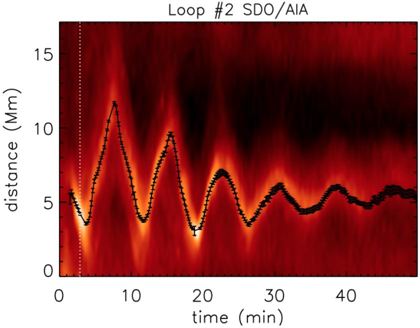
Определение параметров синусоиды

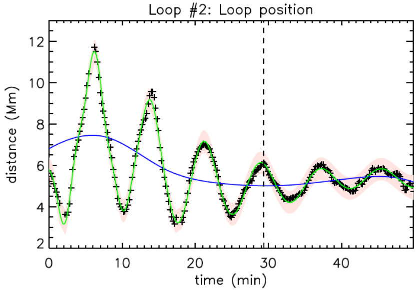
Гистограммы



Определение физических параметров корональной петли

Модель имеет более 20 параметров





Определение физических параметров корональной петли

Результаты