Отчёт о проделанной работе

Лазар В. И.

12.11.2024

1 Проведённые исследования

1.1 Модель

Программно реализована однопиковая модель **PBFTPK** с возможностью обучения на данных и генерации сэмплов с параметрами, подобранными при обучении.

1.2 Метрика

В качестве метрики для оценки моделей большинство уже реализованных алгоритмов используют метрику

$$\frac{1}{n}\sum_{i=1}^{n}(f(t_i) - X(t_i))^2$$

, которая мало подходит в задачах биоэквивалентности в силу слабой интерпретируемости. Был реализован аналог алгоритма подбора параметров модели для метрики

$$\frac{1}{n}\sum_{i=1}^{n}|f(t_i) - X(t_i)|$$

, что дало большее качество при обучени модели.

1.3 Исследование моментных характеристик остатков

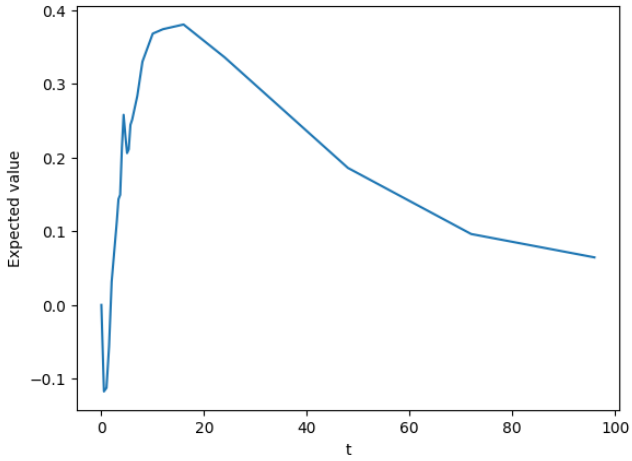
Здесь и далее будем действовать в предположении о том, что величина

$$X(t) - f(t)$$

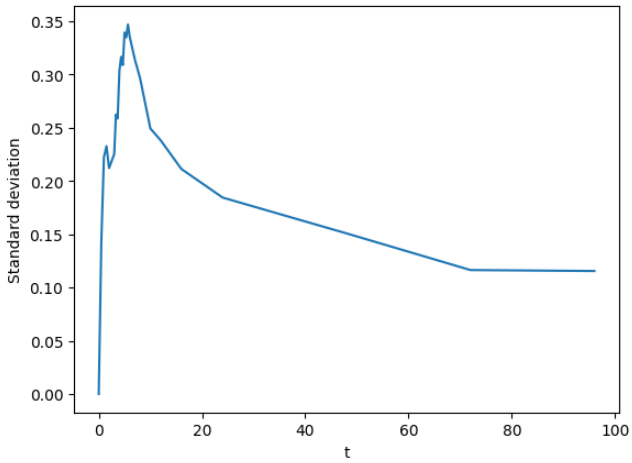
, где X(t) - исходный случайный процесс, а f(t) - тректория предсказанная моделью, является процессом Леви

Для получения более точного вида процесса было решено исследовать матожидание и стандартное отклонение проекций процесса остатков

Получен следующий график поведения для матожидания



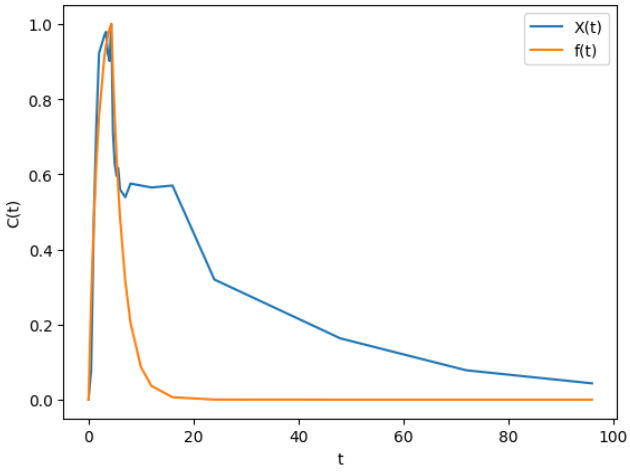
И для стандартного отклонения

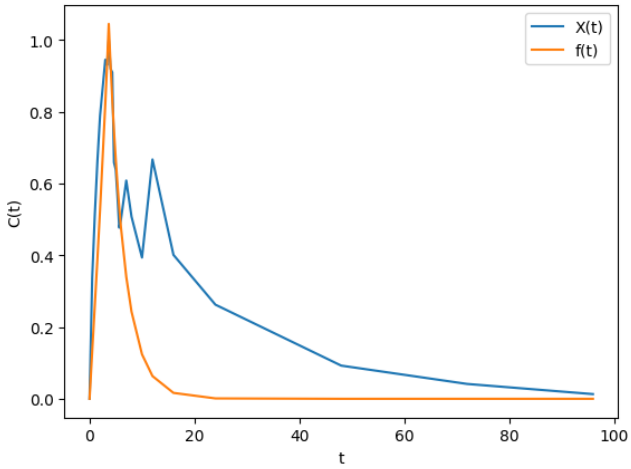


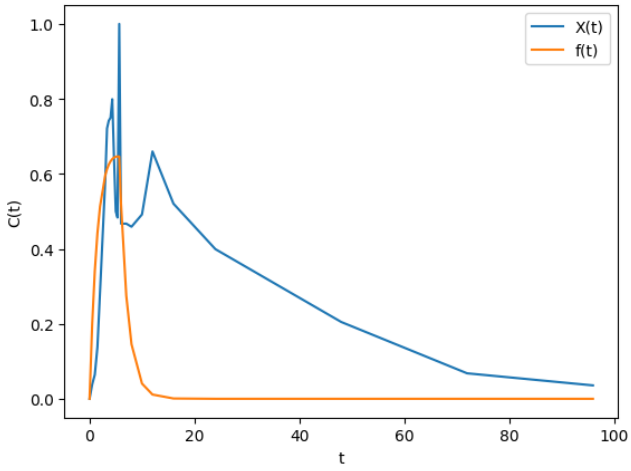
Заметно, что в среднем наша модель сильнее всего ошибается в окрестности 20-и. Запомним это и рассмотрим несколько графиков процессов

1.4 Исследование поведения исходного процесса

Здесь показаны одни из типовых случаев процессов, на которых моделm значительно ошибается ошибается


X(t)
f(t)
C(t)
t
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100


X(t)
f(t)
C(t)
t
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100


X(t)
f(t)
C(t)
t
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100

Заметим, что второй пик здесь находится в окрестности 20, что согласуется с полученными выше результатами. Поскольку подобных процессов в датасете достаточно много, логично предположить, что однопиковой модели **PBFTPK** будет недостаточно для предсказания поведения процесса. Более того, учитывая, что однопиковых процессов в датасете также много, возможно логично будет предположение о небиоэквивалентности референтного и тестового препаратов

2 Гипотезы и планы

2.1 Разметка датасета

Поскольку была обнаружена особенность сильно влияющая на работу модели, появилась идея вручную разметить датасет и для каждого процесса указать возможное количество максимумов (от 1 до 3). Возможно, имеет смысл создать алгоритм (или обучить отдельную модель) для поиска количества пиков процесса и оценки вероятности того, что количество пиков найдено верно

2.2 Модель

По этой же причине было принято решение усовершенствовать модель для учёта возможных многопиковых случаев. Возможно также имеет смысл изменить функцию потерь на следующую:

$$L(f, \alpha) = \frac{1}{n} \sum_{i=1}^{i_0-1} |f(t_i) - X(t_i)| + \frac{1}{n} \sum_{i=i_0}^{n} \alpha |f(t_i) - X(t_i)|$$

$$\alpha \geq 0$$

$$X(t_i) = \tau$$

Коэффициент α отвечает за то, насколько сильно на общую ошибку влияет отклонение после времени абсорбции τ. Это может быть полезным, так как сильнее всего модель ошибается после достижения пика.