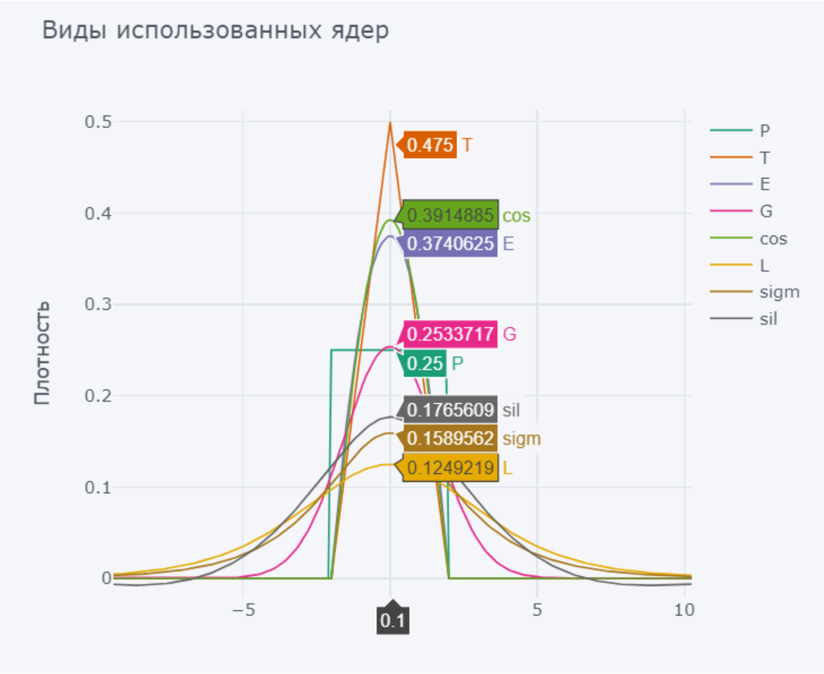
Анализ непараметрического подхода к восстановлению плотности Парзеновским окном.

Целью данного исследования было выявление функции и соответствующего параметра, которые с большей скоростью и точностью могли бы восстановить плотность распределения.

Было исследовано восемь видов ядер:

- прямоугольное/равномерное (P)
- треугольное (T)
- Епанечникова
- Гауссово (G)
- Косинусоидное (cos)
- Логистическое (L)
- Сигмоидальное (sigm)
- Сильвермана (sil)



Во всех ядрах варьировался параметр ширины окна кроме Гауссова, в котором дополнительно использовался параметр alfa = 1 / sqrt(D).

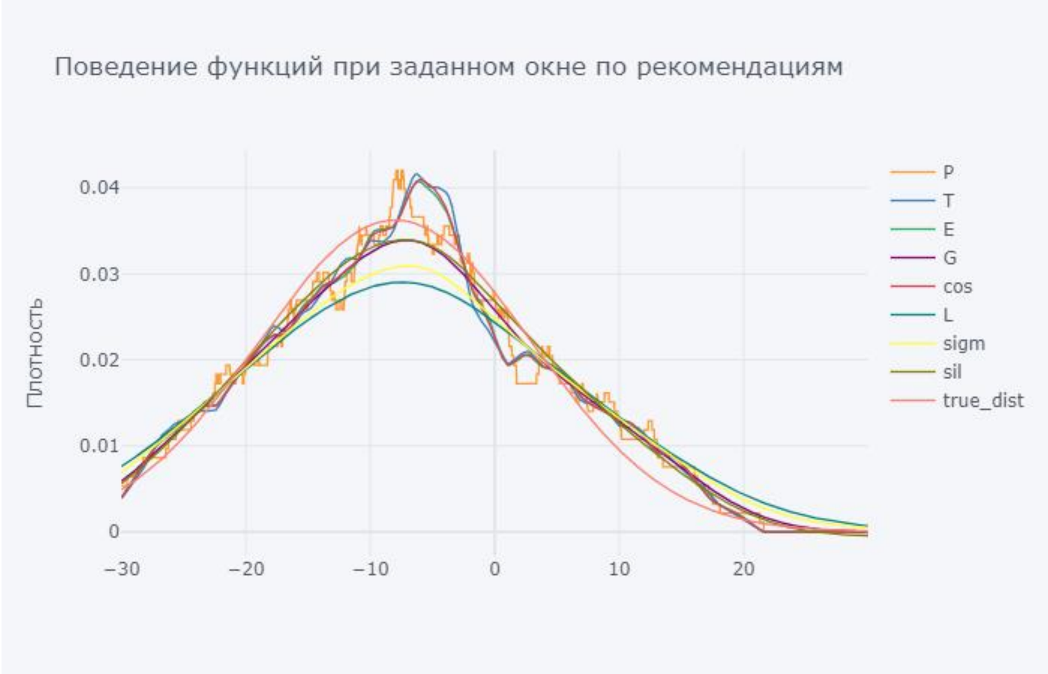
Для визуализации был использована библиотека plotly, которая позволяет рисовать интересные интерактивные графики. Если обратиться к файлу ipynb данной работы, там можно опробовать интерактивность графиков.

Для анализа сгенерирован набор данных из нормального распределения с произвольными параметрами (μ = -8, σ = 11).

В первую очередь было исследовано эмпирическое правило, согласно которому, оптимальный размер окна должен составлять:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1,06\hat{\sigma}n^{-1/5},$$

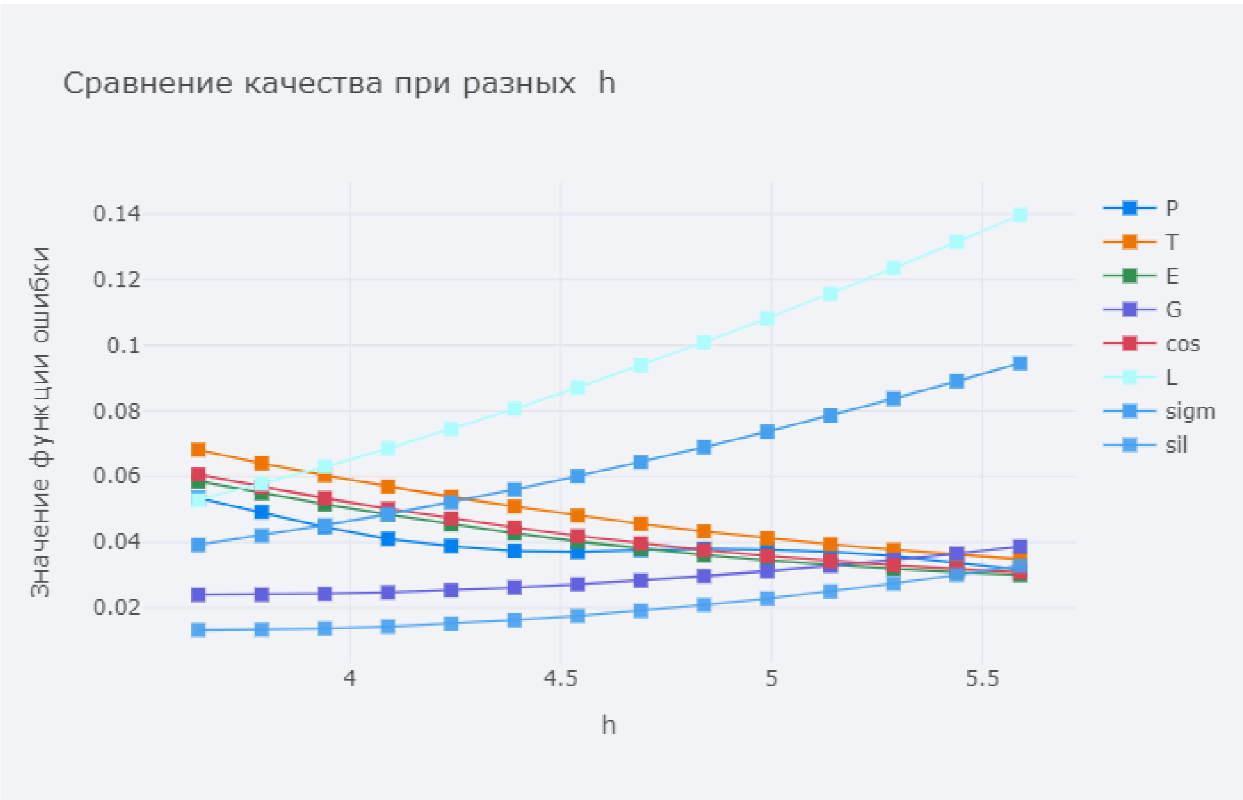
или в нашем случае эта величина составила ~4,64.



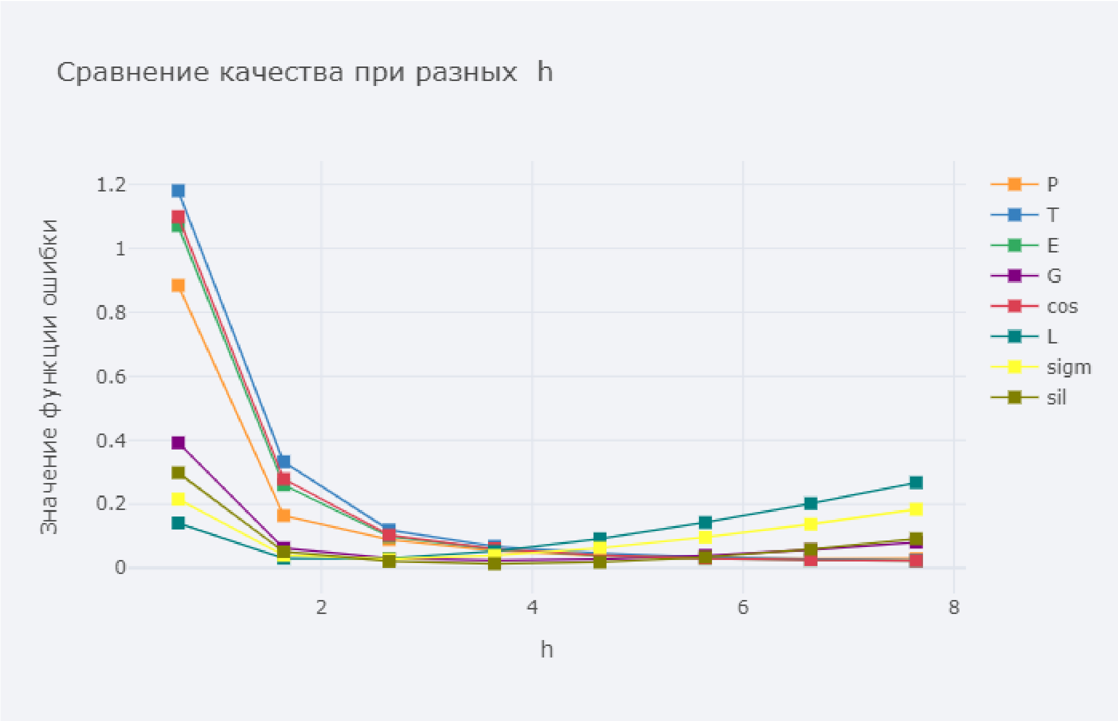
Качество модели решено было измерять стандартной суммой квадратов отклонений, посчитаны так называемые истинные значения при исходных параметрах, с которыми и проводилось сравнение.

Максимальное значение MSE получилось у логистической функции ядра, минимум дало ядро Сильвермана. Интересно, что функция равномерного ядра оказалась не самой худшей в плане MSE.

Ниже приведен график зависимости от величины окна и размера ошибки для разных ядер. И при столь малых изменениях h качество нельзя сказать, что стабильное.

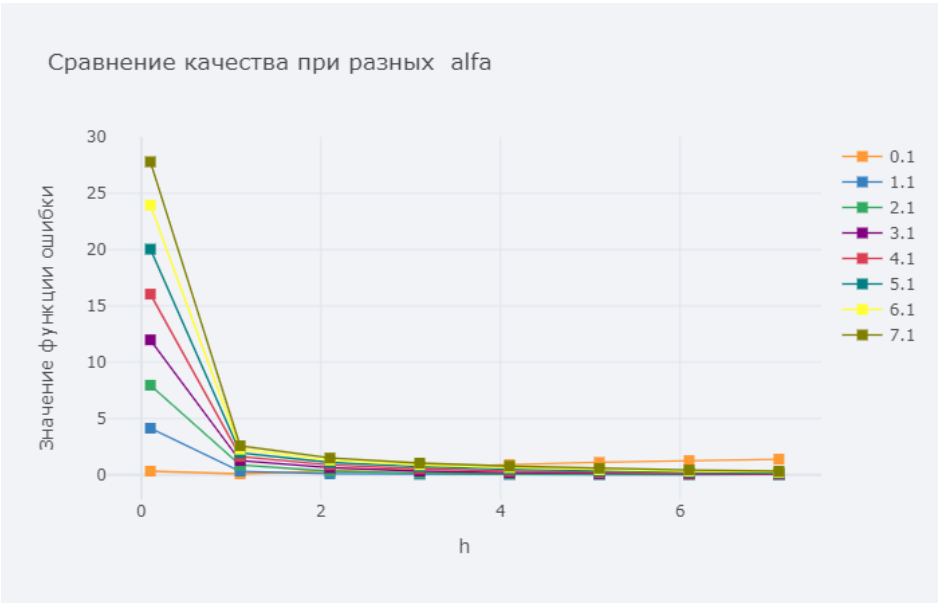


Для определения стабильности MSE был взят больший разброс h, результаты представленны ниже.



Тут сразу видим, что ядерные функции при снижении h ведут себя хуже. При этом стоит отметить, что наименьший разброс качества у логистической функции, «Треугольное» ядро дало самый высокий скачок, но при этом минимума еще не достигнуто. В целом можно разделить функции по данному графику на две группы. Для первой группы можно уже указать на диапазон, в котором достигается минимум ошибки (последние три функции в списке и Гауссово ядро) – примерно от 1,64 до 3,64. Вторая половина относится к тем, для которых можно попробовать еще улучшить результат.

Теперь рассмотрим функцию с использованием гауссова ядра как функцию двух переменных: h – окно, alfa – параметр самого ядра.



Минимум достигается при h = 6.1 и alfa = 1.1. Уменьшение окна также приводит к ухудшению результатов.