Министерство образования и науки Российской Федерации

Санкт-Петербургский политехнический университет Петра Великого

Институт компьютерных наук и кибербезопасности

**Высшая школа компьютерных технологий и информационных систем**

**ЛАБОРАТОРНАЯ РАБОТА №1**

Классификация данных на плоскости с помощью байесовского классификатора

Выполнил

студент гр. 5140901/31502 <*подпись*> A.A. Алешковский

Проверил

доцент, к.т.н. <*подпись*> К.В. Никитин

Санкт-Петербург

2023

**Часть 3 Параметрическое оценивание распределений вероятностей**

1. Построим истинные плотности и плотности оценок

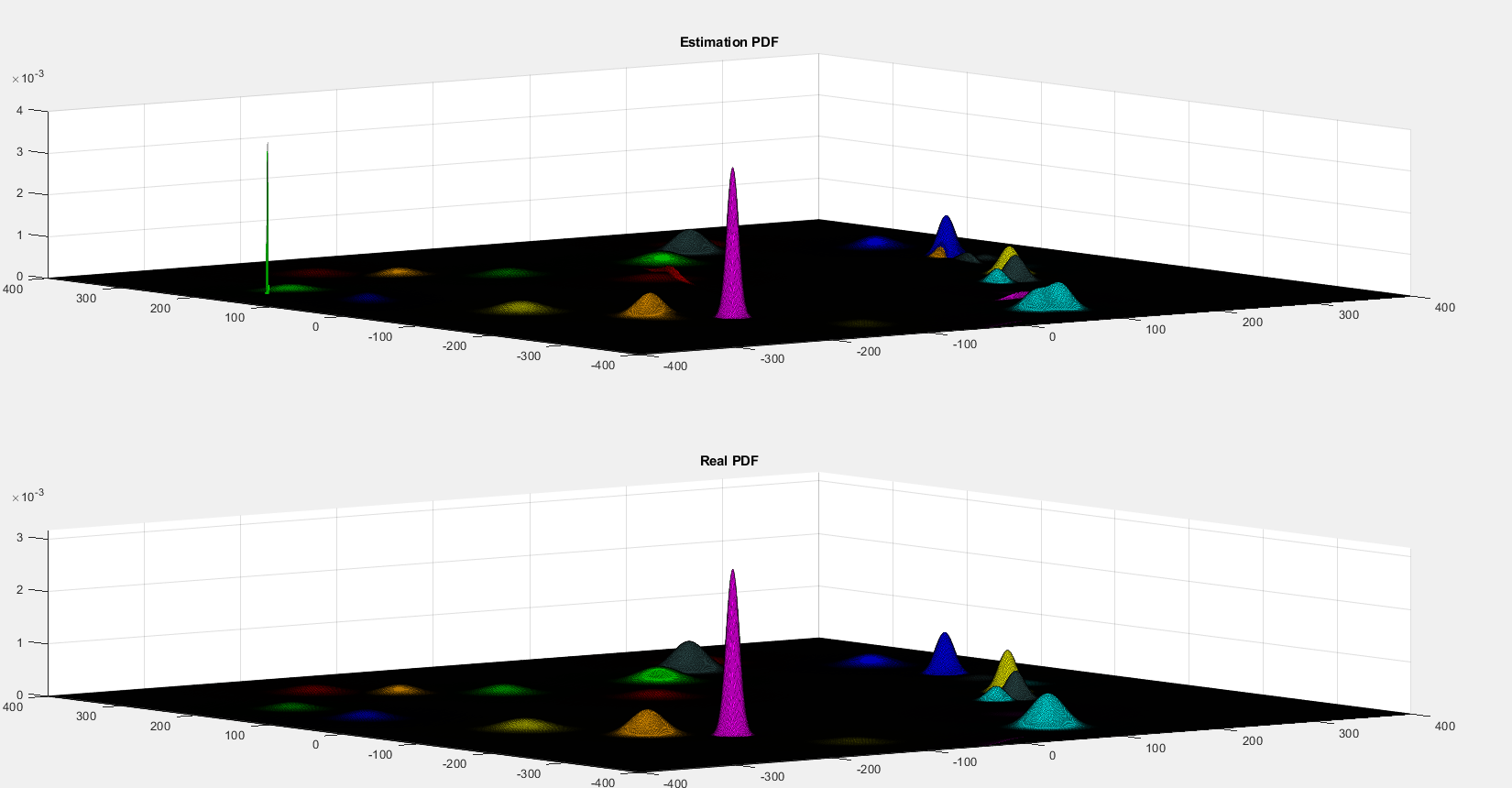


Рисунок 1 – Графики истинной плотности и плотности оценок

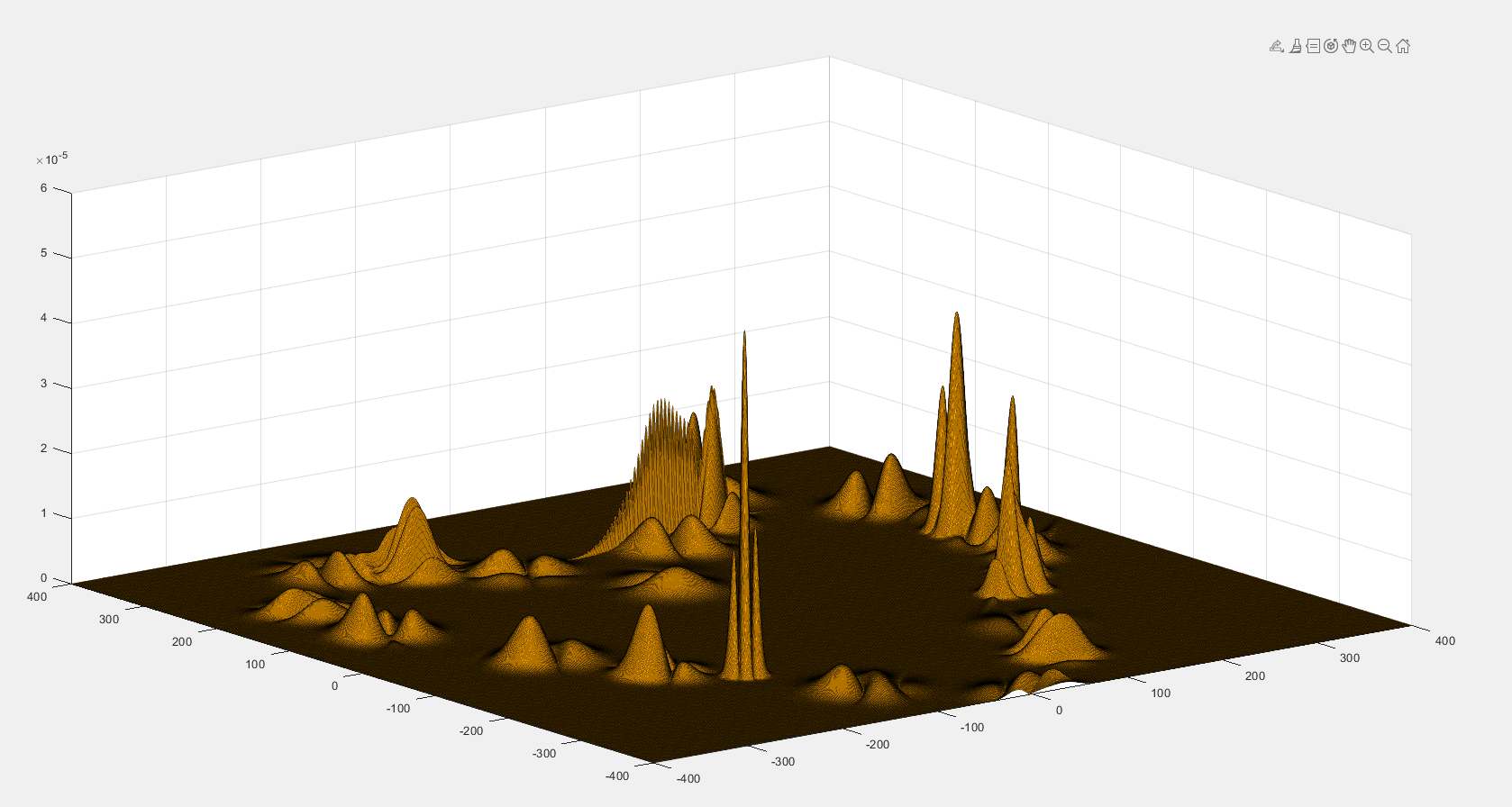


Рисунок 2 – График ошибок

Построим одномерные плотности:

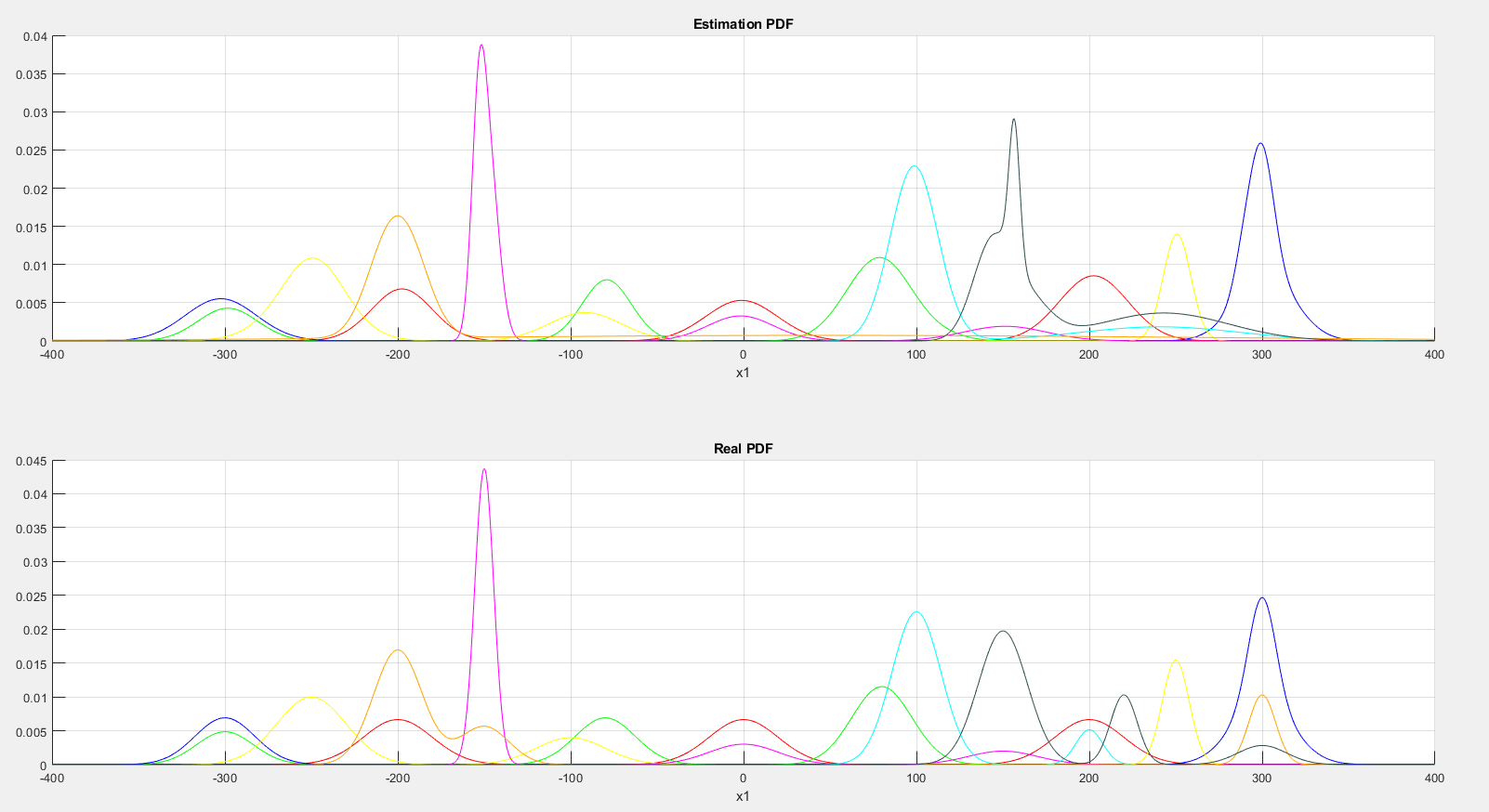


Рисунок 3 – Графики истиной плотности и плотности оценок X1

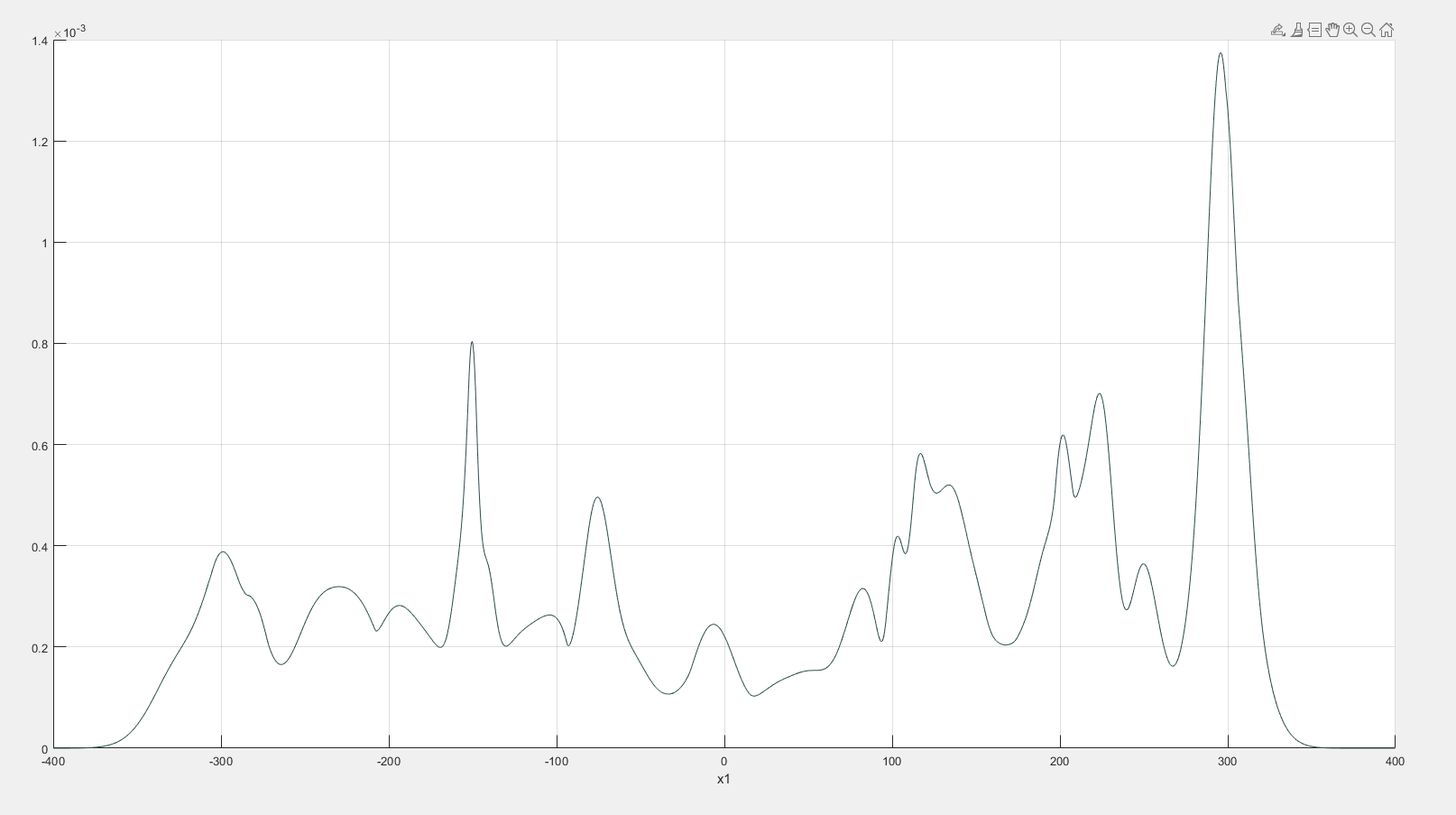


Рисунок 4 – График ошибок X1

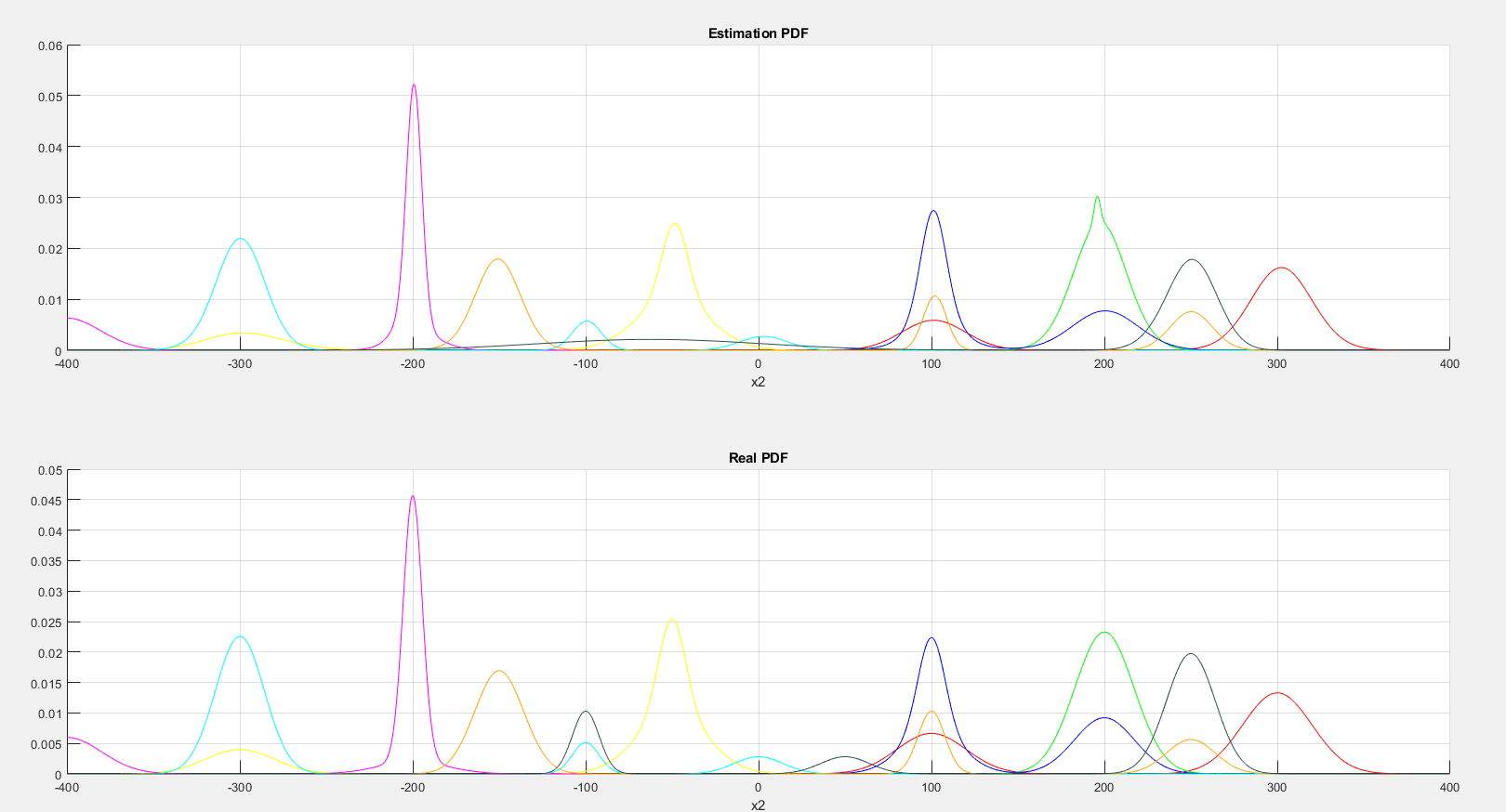


Рисунок 5 – Графики истиной плотности и плотности оценок X2

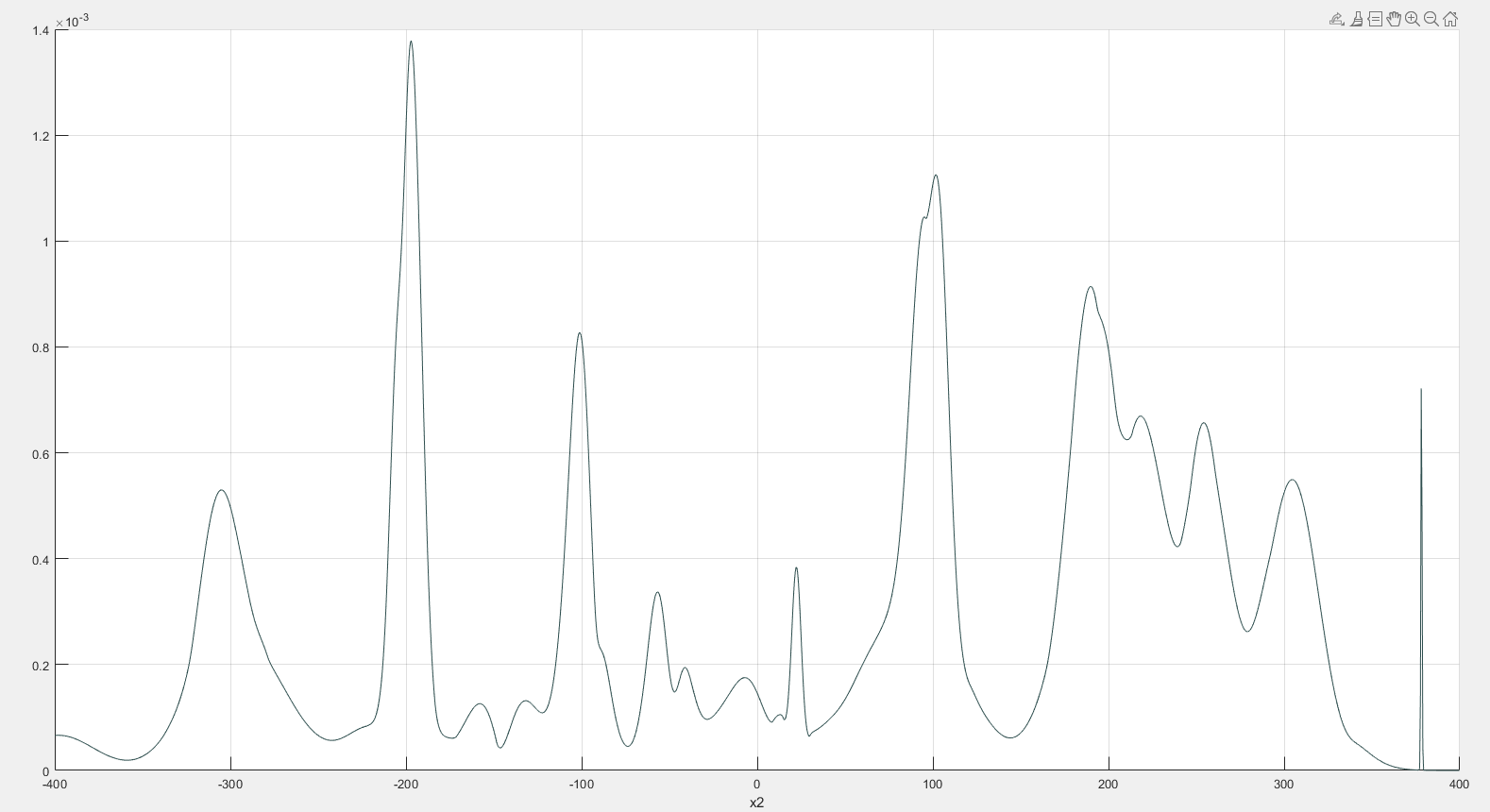


Рисунок 6 – График ошибок X2

Решим задачу классификации и построим график типа Т1:

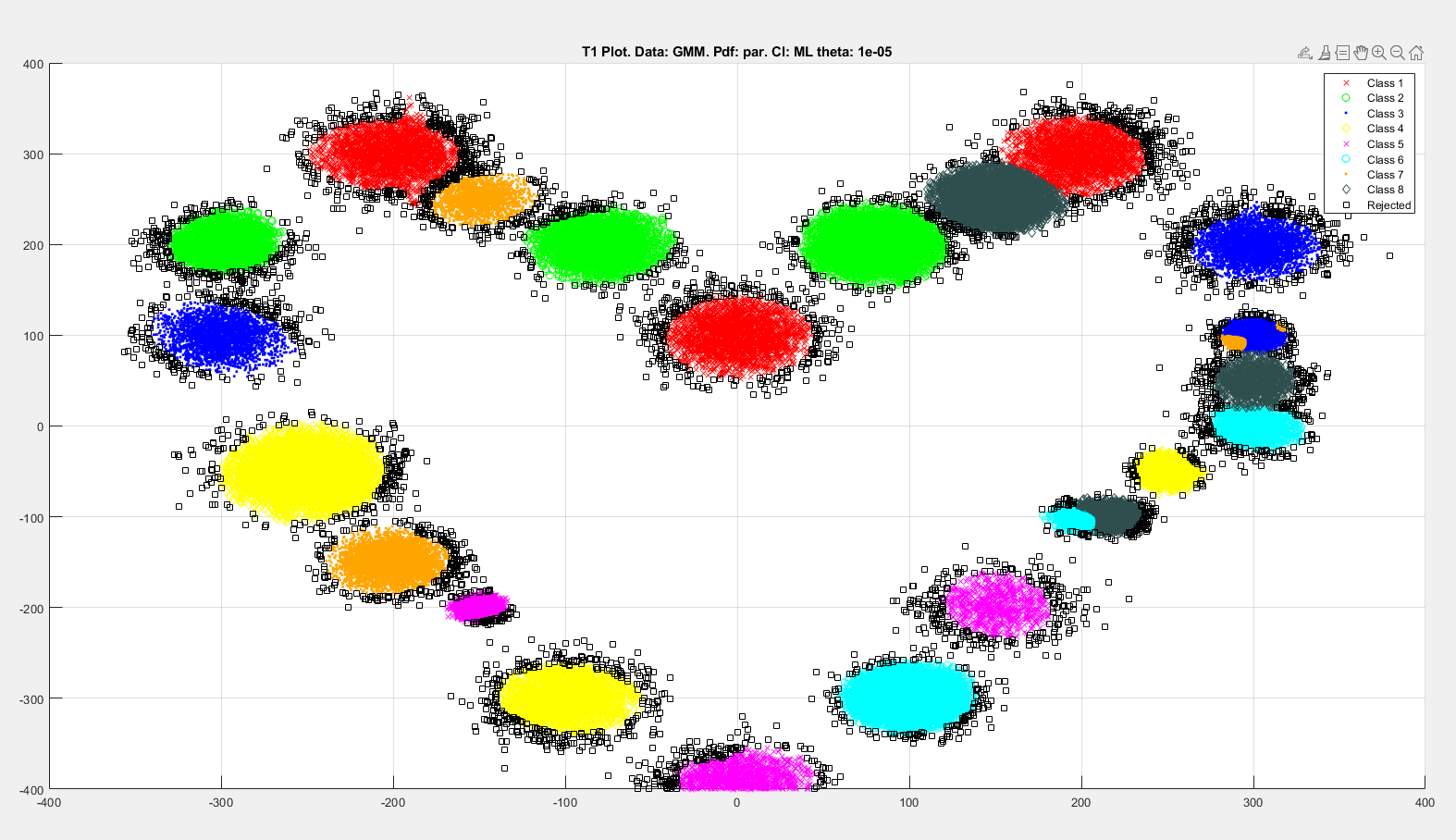
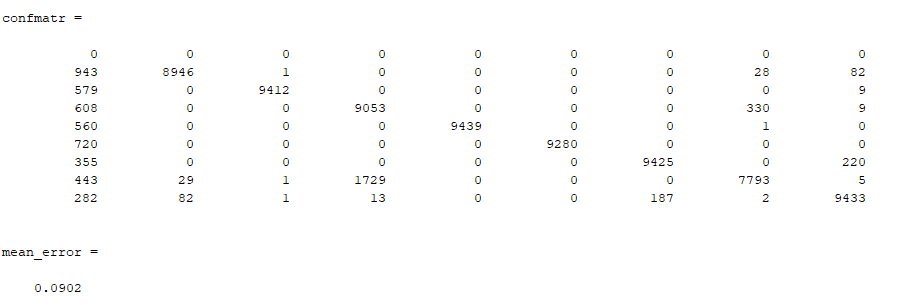


Рисунок 7 – График типа Т1



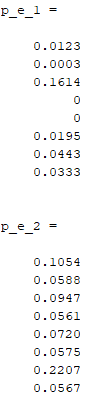


Рисунок 8 – Вывод в консоли при построении графика T1

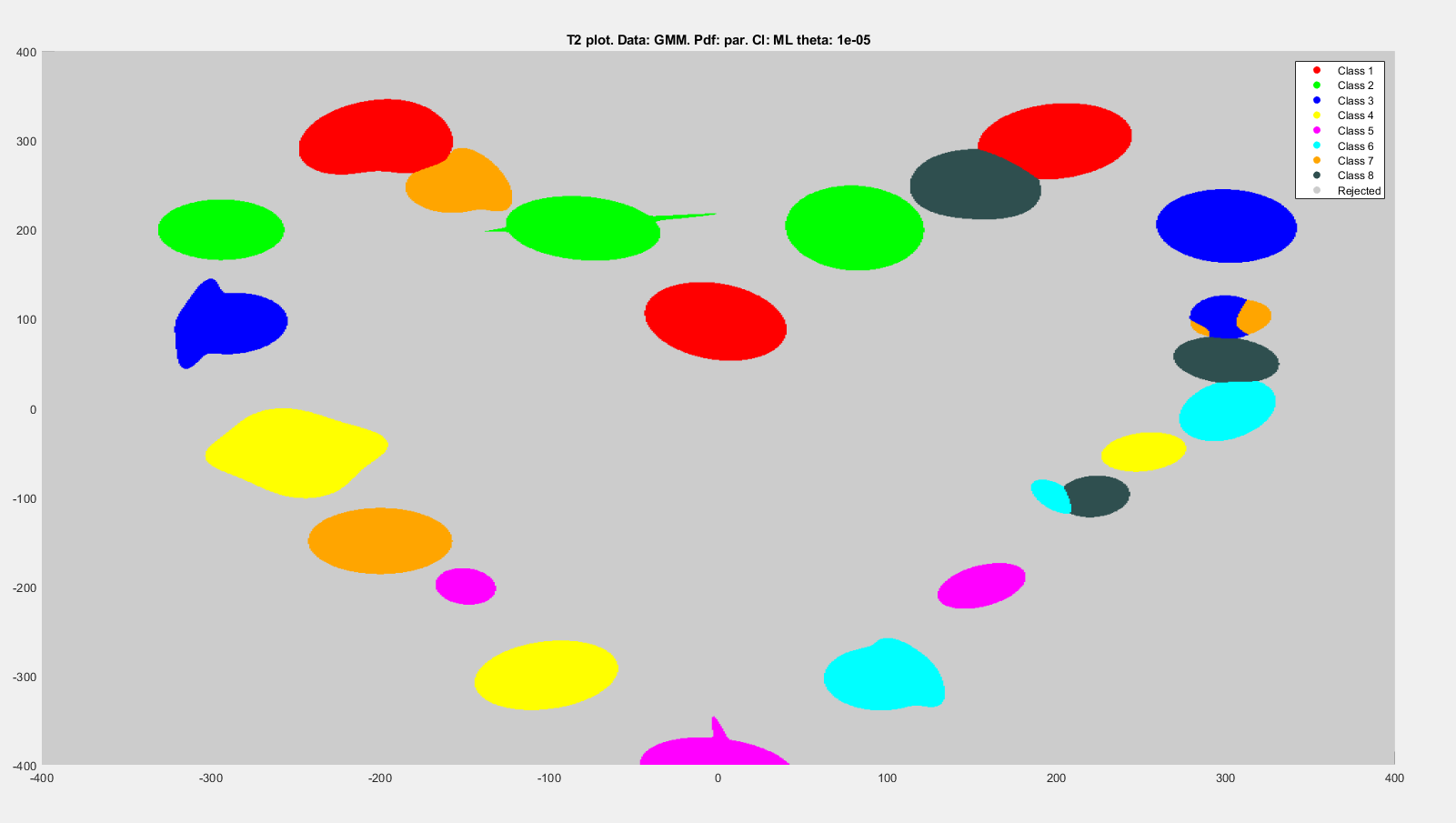


Рисунок 9 – График типа Т2

1. Определим, насколько адекватна гипотеза о независимости признаков

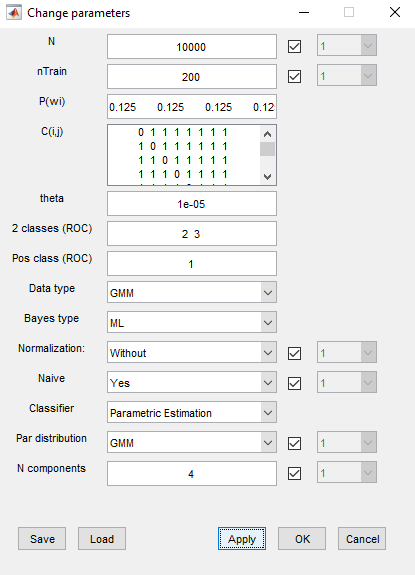


Рисунок 10– Native = yes

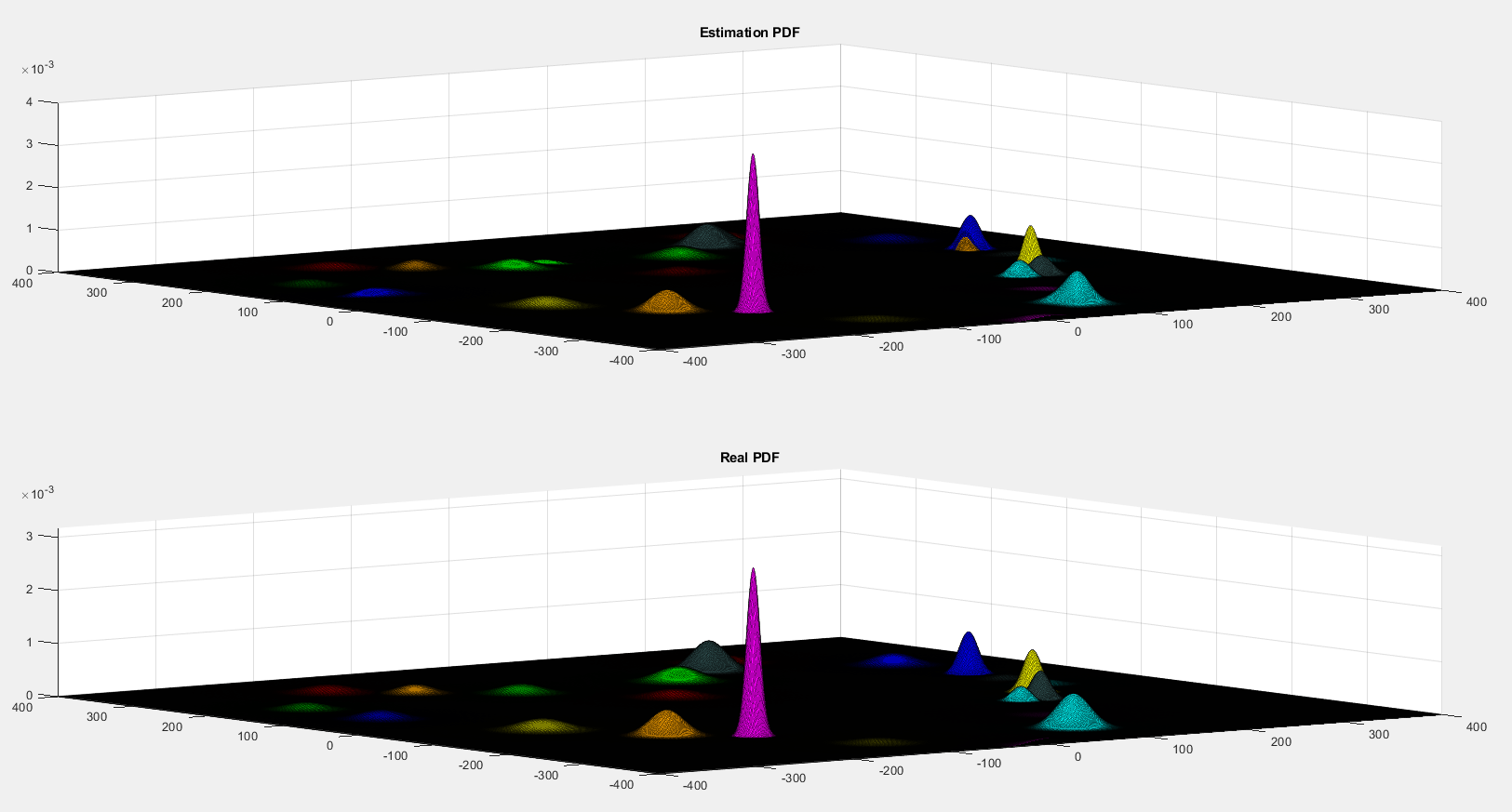


Рисунок 11 – Графики истинной плотности и плотности оценок

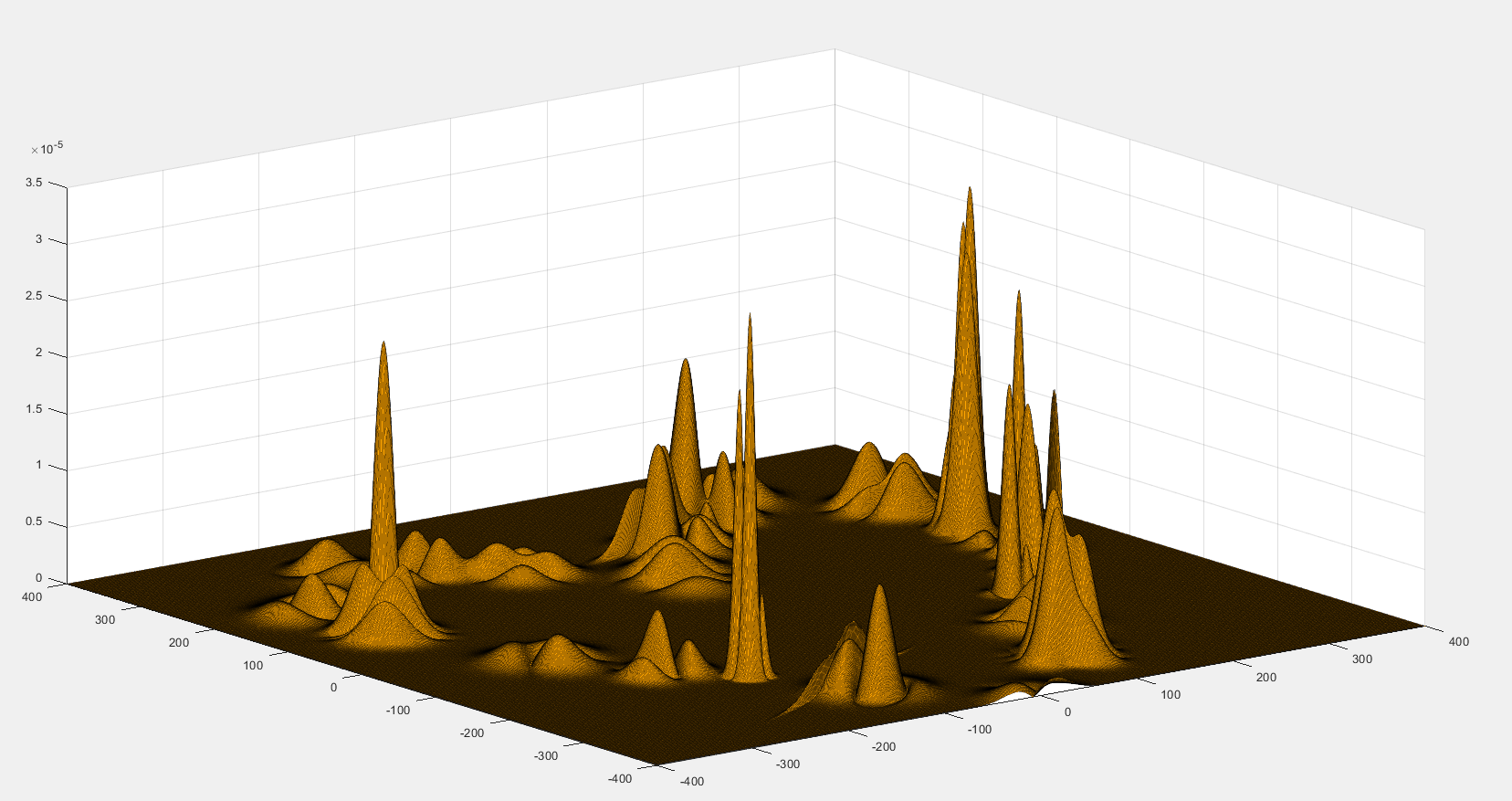


Рисунок 12 – График ошибок

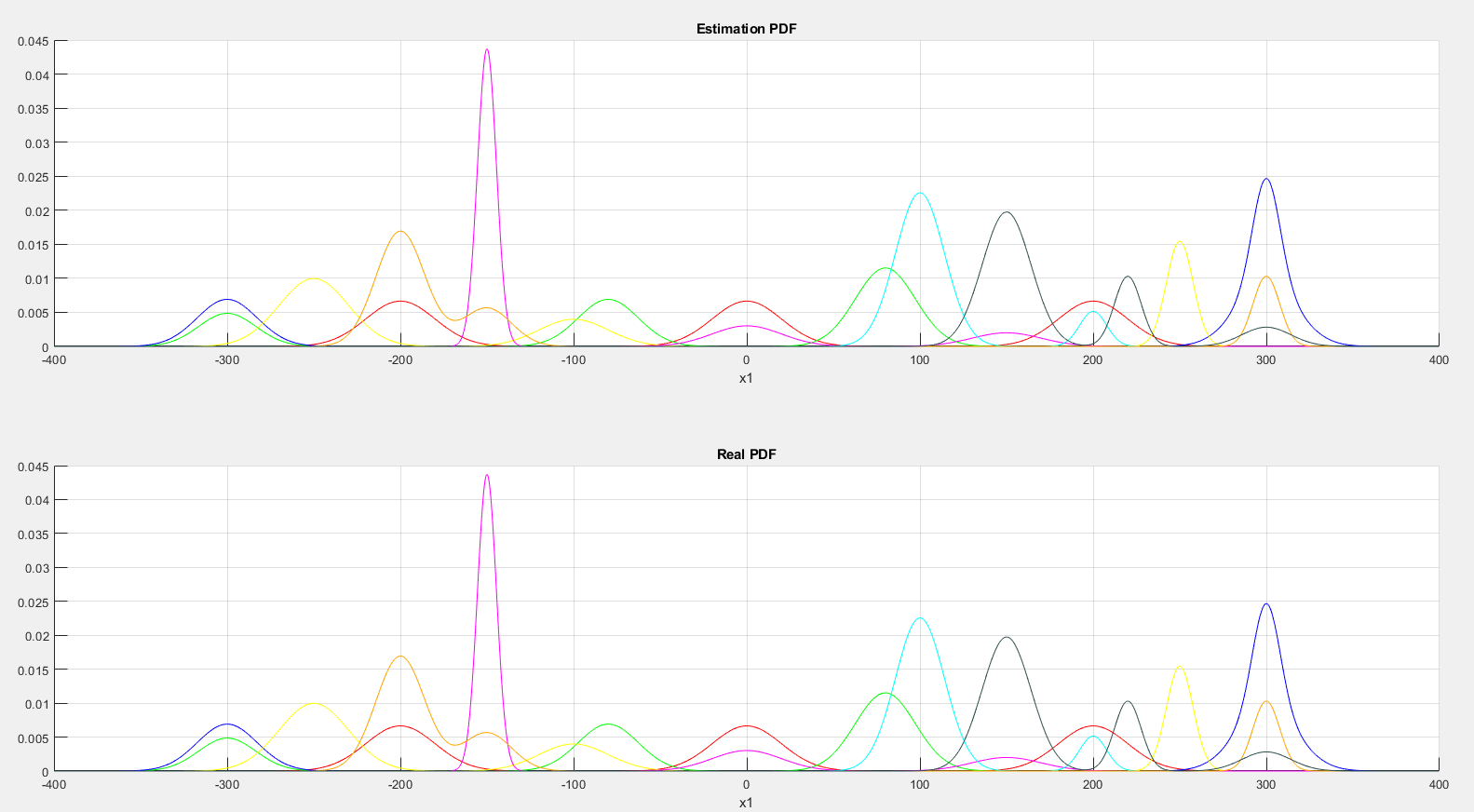


Рисунок 13 – Графики истинной плотности и плотности оценок X1

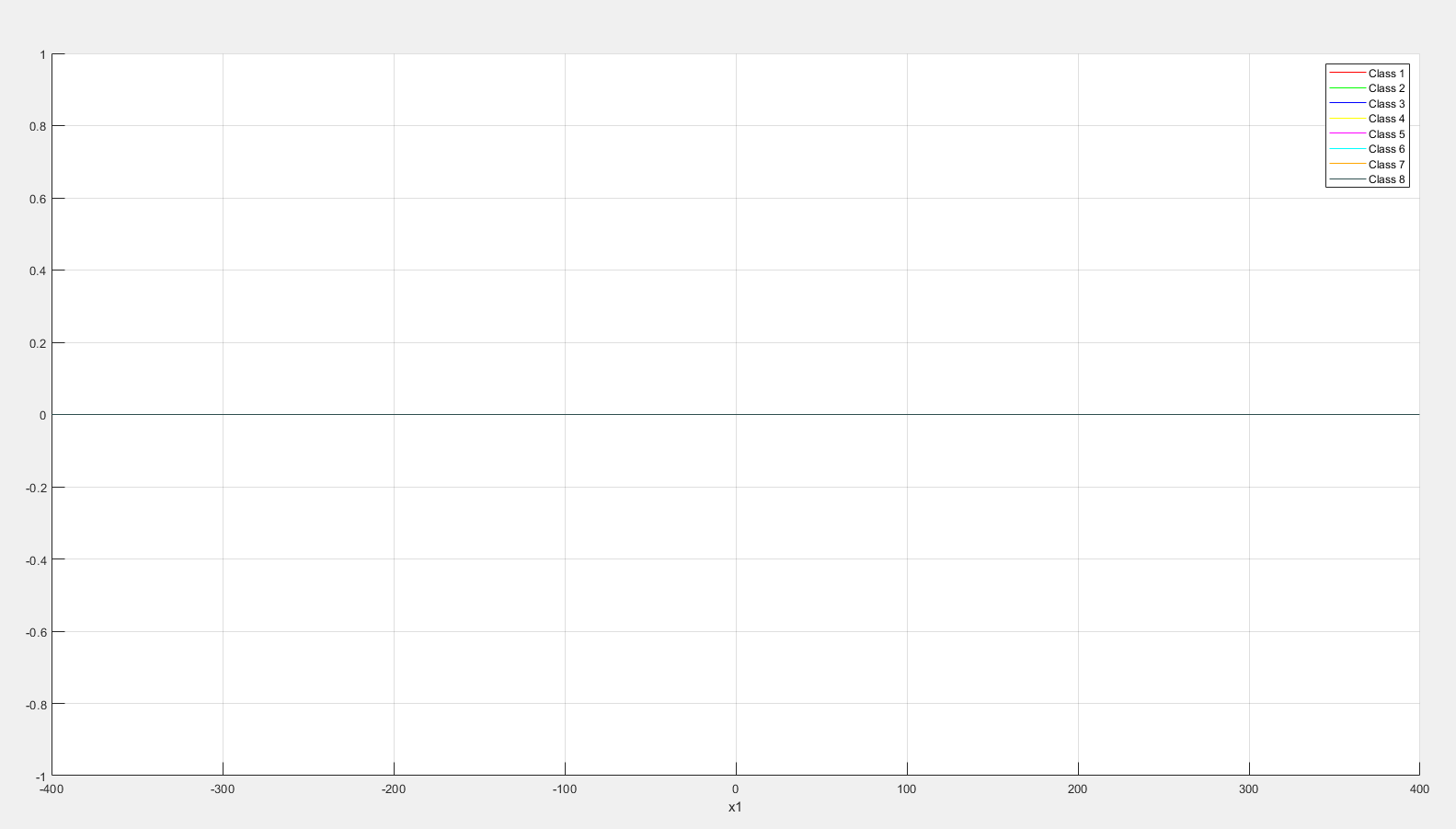


Рисунок 14 – График ошибок X1

По непонятной и неочевидной мне причине график имеет приведенный выше вид.

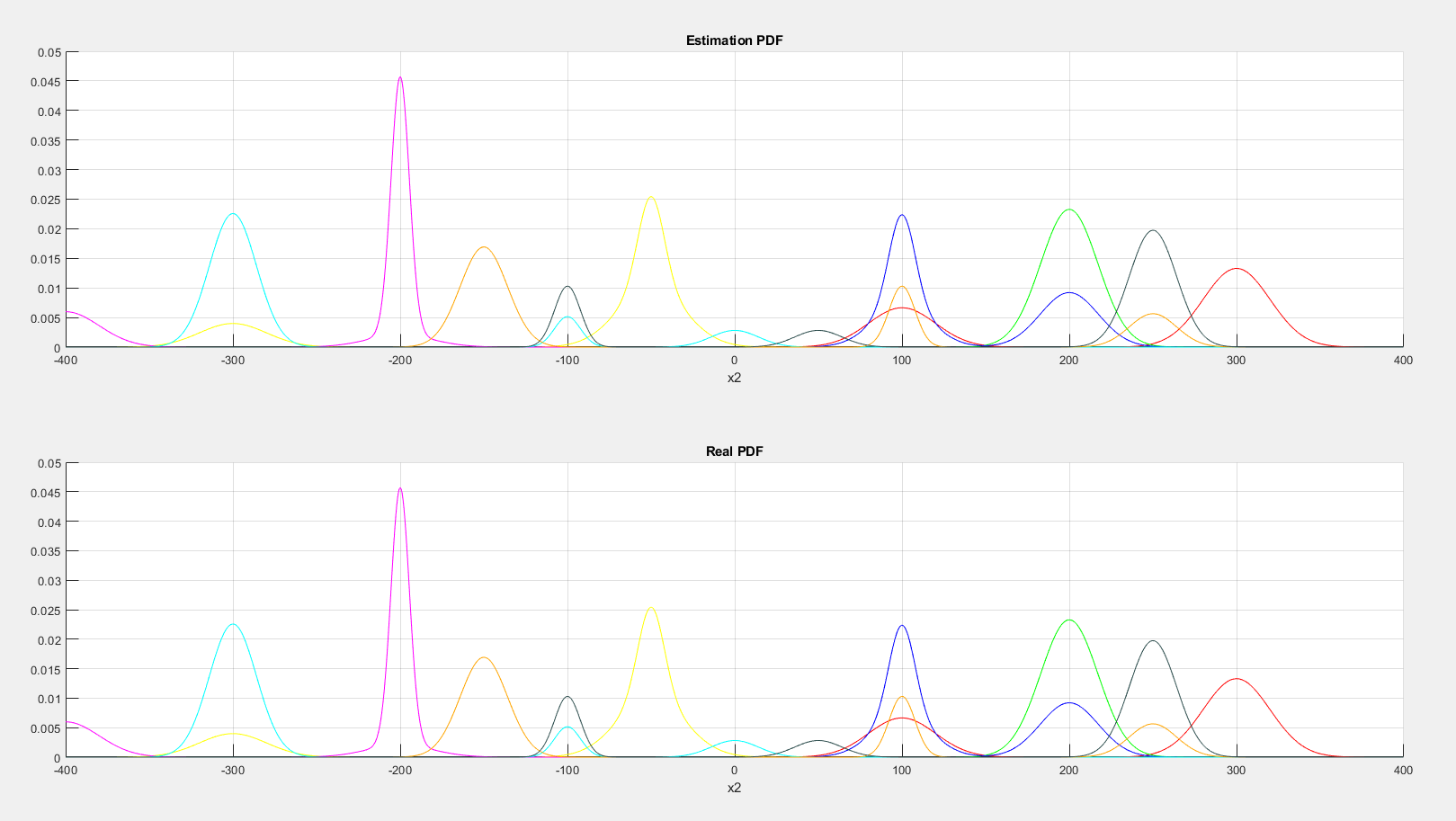


Рисунок 15 – Графики истинной плотности и плотности оценок X2

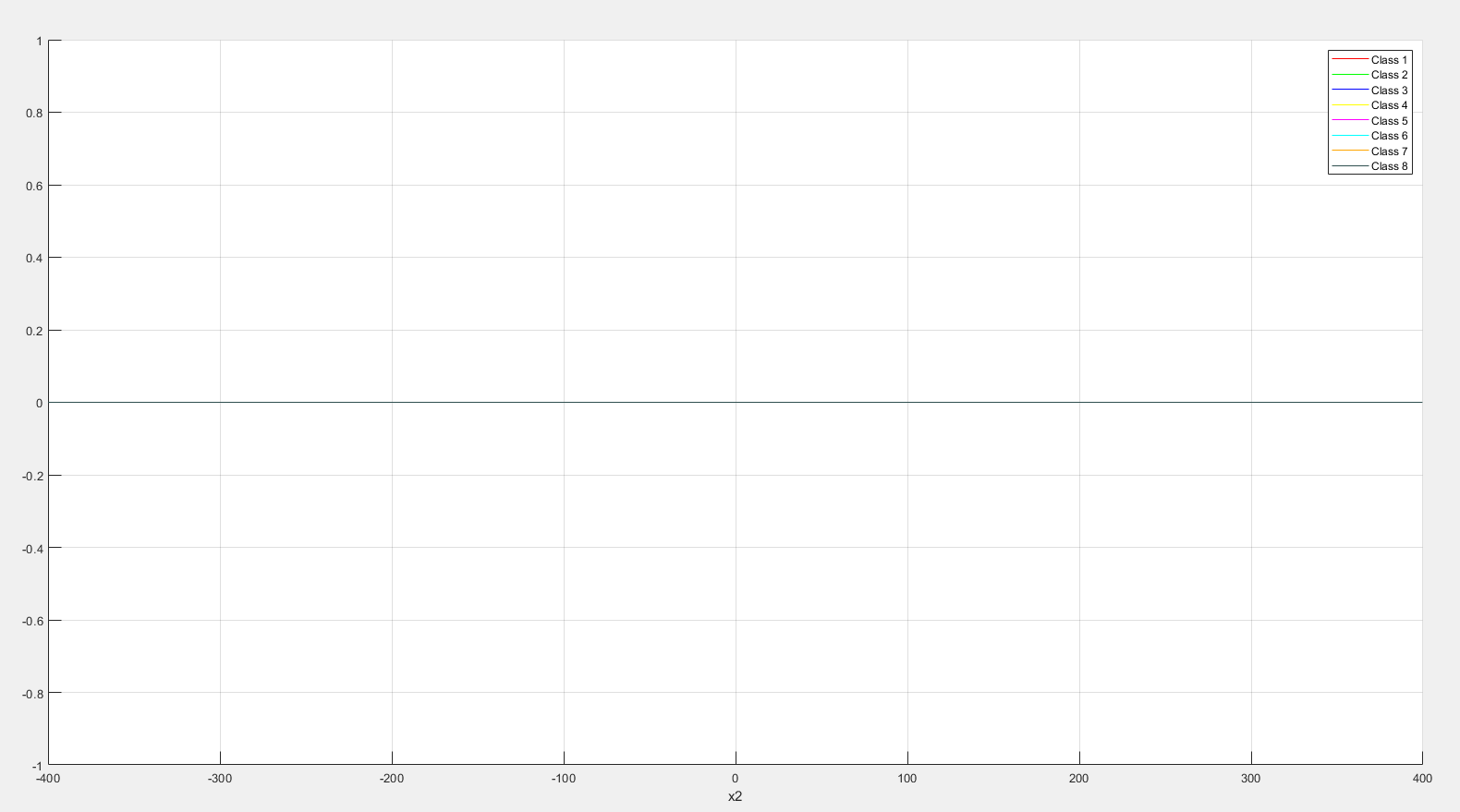


Рисунок 16 – График ошибок X2

Аналогичная ситуация и с X2.

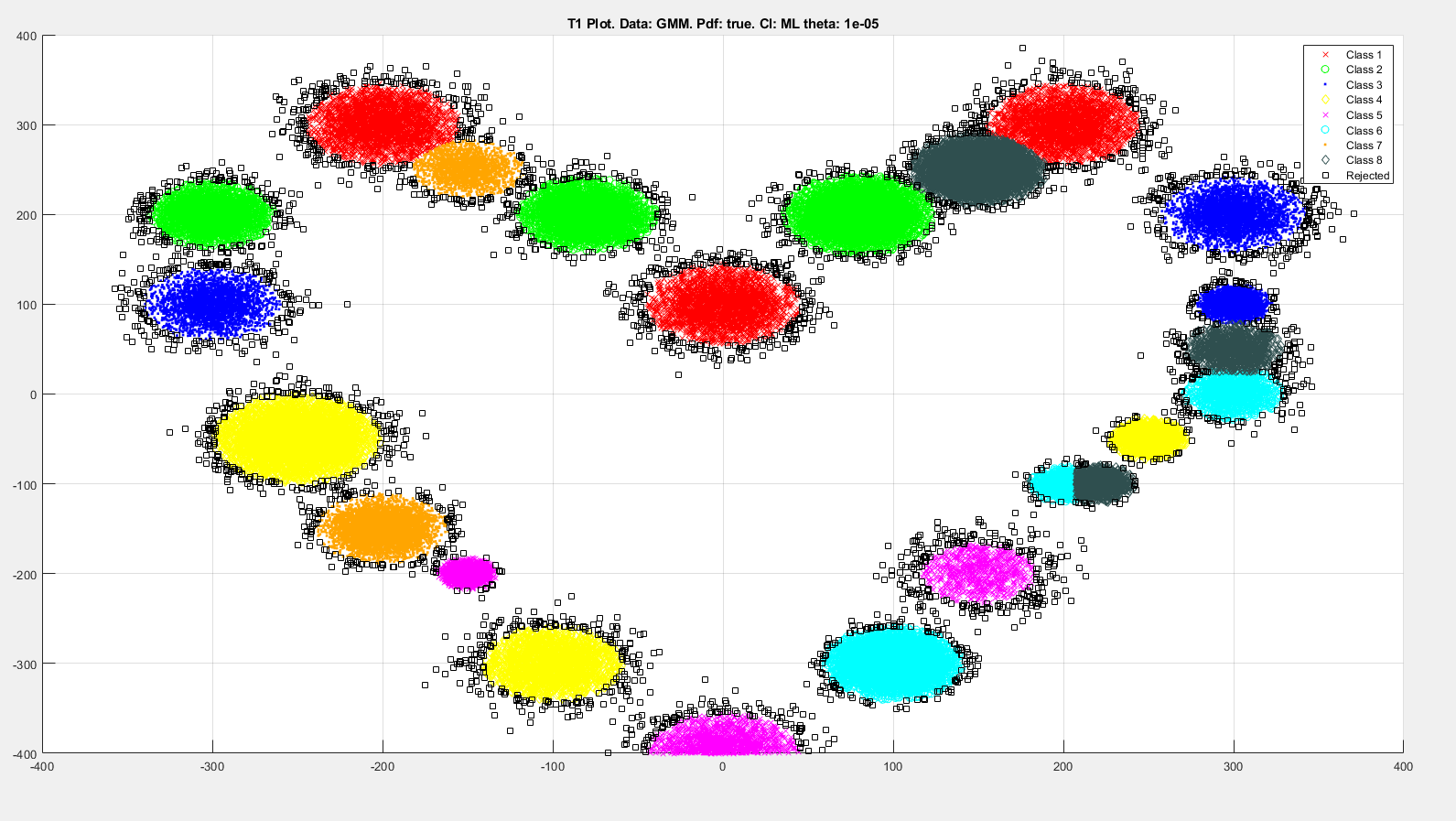
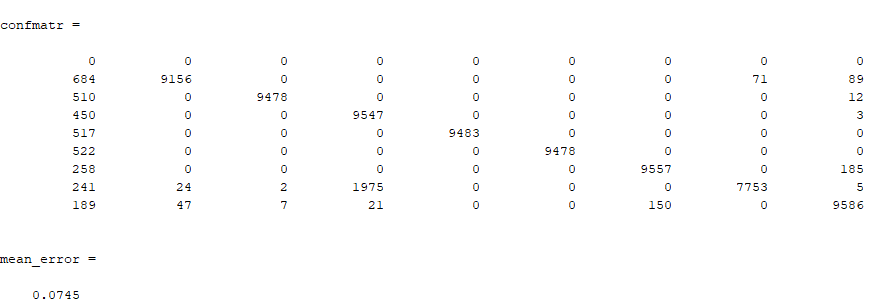


Рисунок 17 – График типа Т1



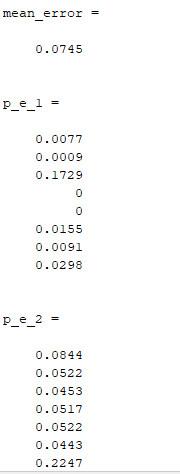


Рисунок 18 – Вывод в консоли при построении графика T1

Исходя из полученных результатов, сделаем промежуточные выводы.

При выставлении значение **Naïve=Yes** в настройках мы добились уменьшения

**Mean\_error (В первом случае – 0.902, во втором – 0.745)**

Оценим при различных способах нормализации входных признаков. Отключим опцию с гипотезой о независимости признаков (Naïve=No). Включим последовательно опции нормализации с выравниваем дисперсий (Normalization=Norm Variances) и отбеливанием (Normaliztion=Whitening), когда данные приводятся к единичной корреляционной матрице.

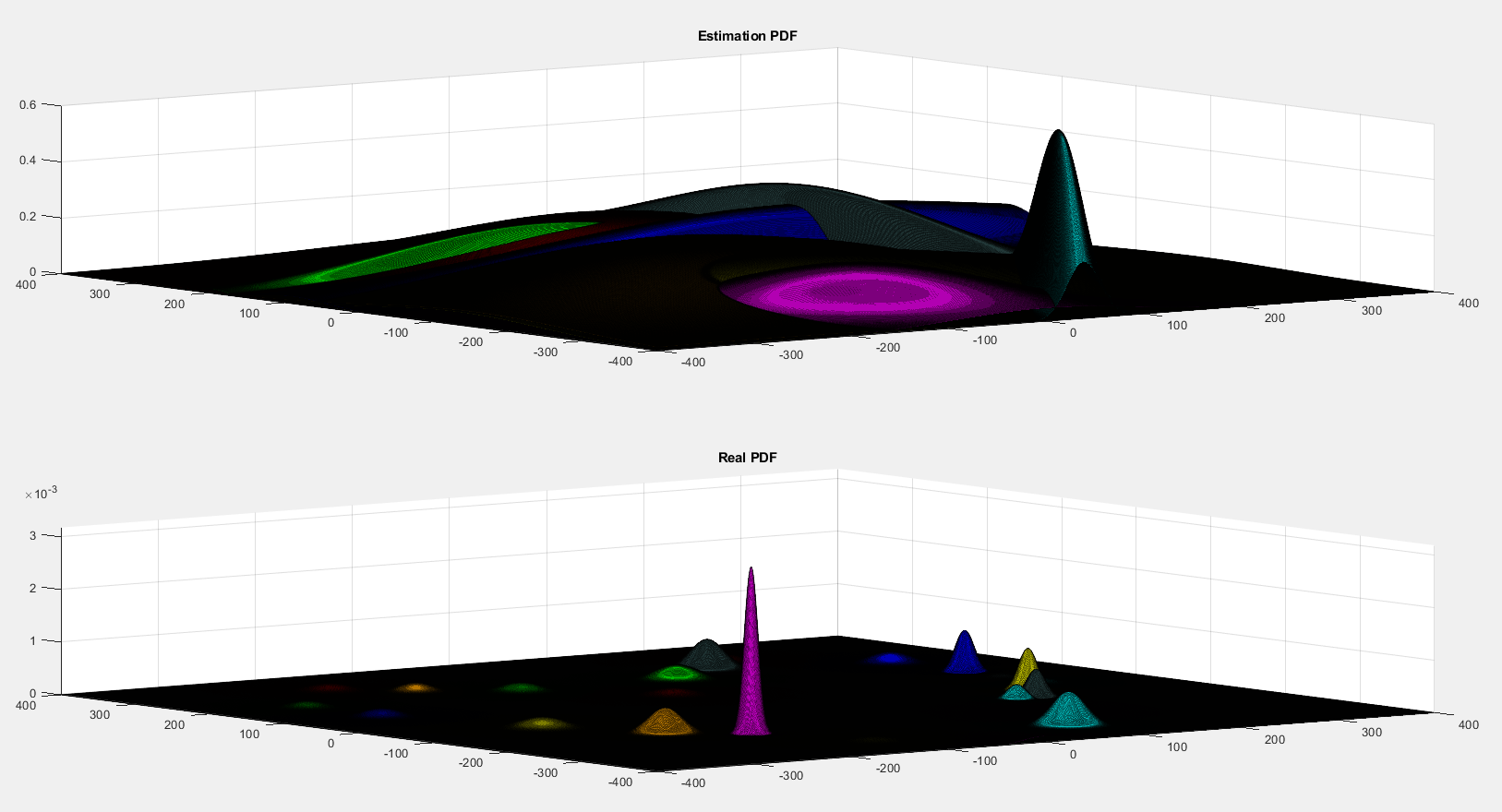
****

Рисунок 19 – Графики истинной плотности и плотности оценок

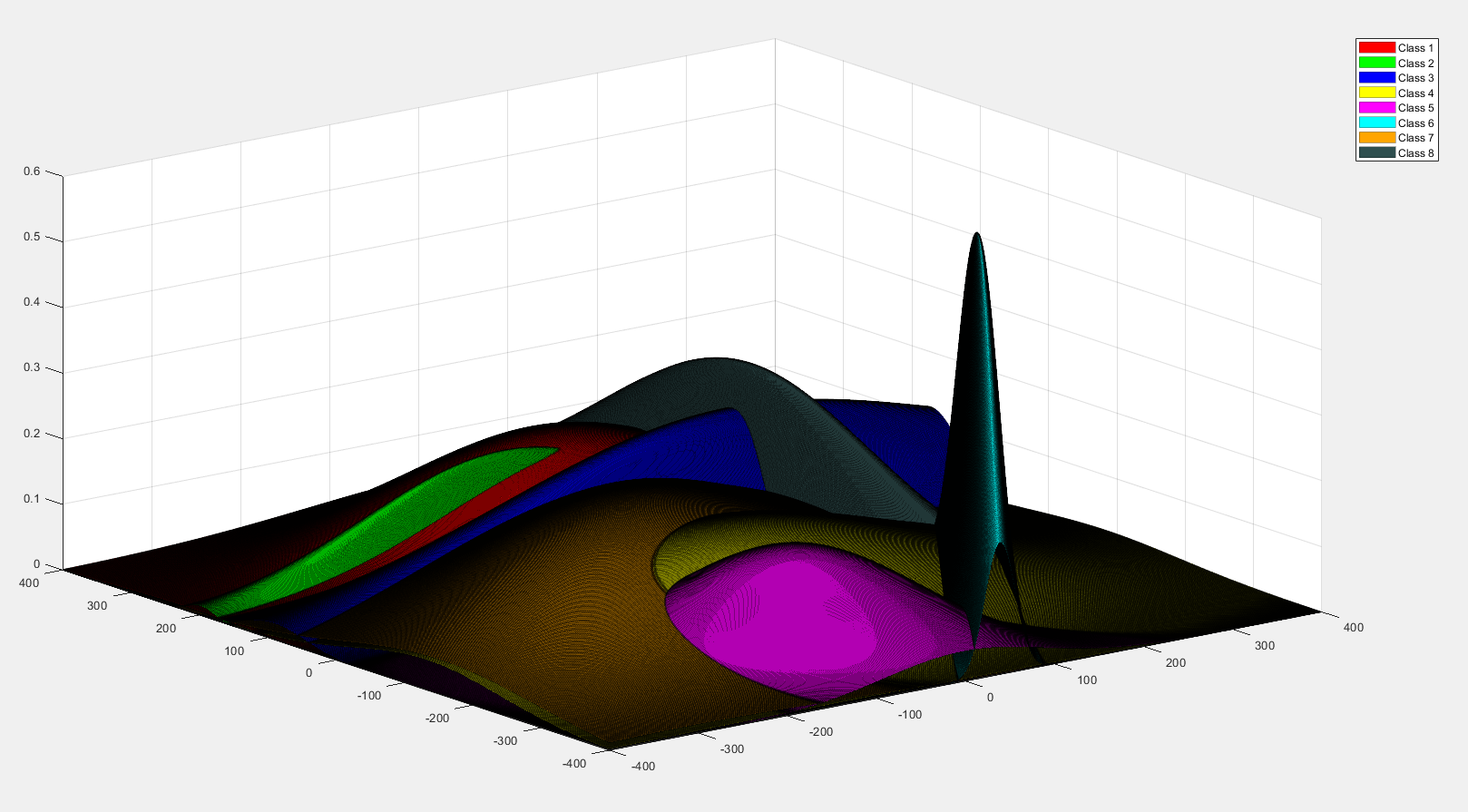
****

Рисунок 20 – График ошибок

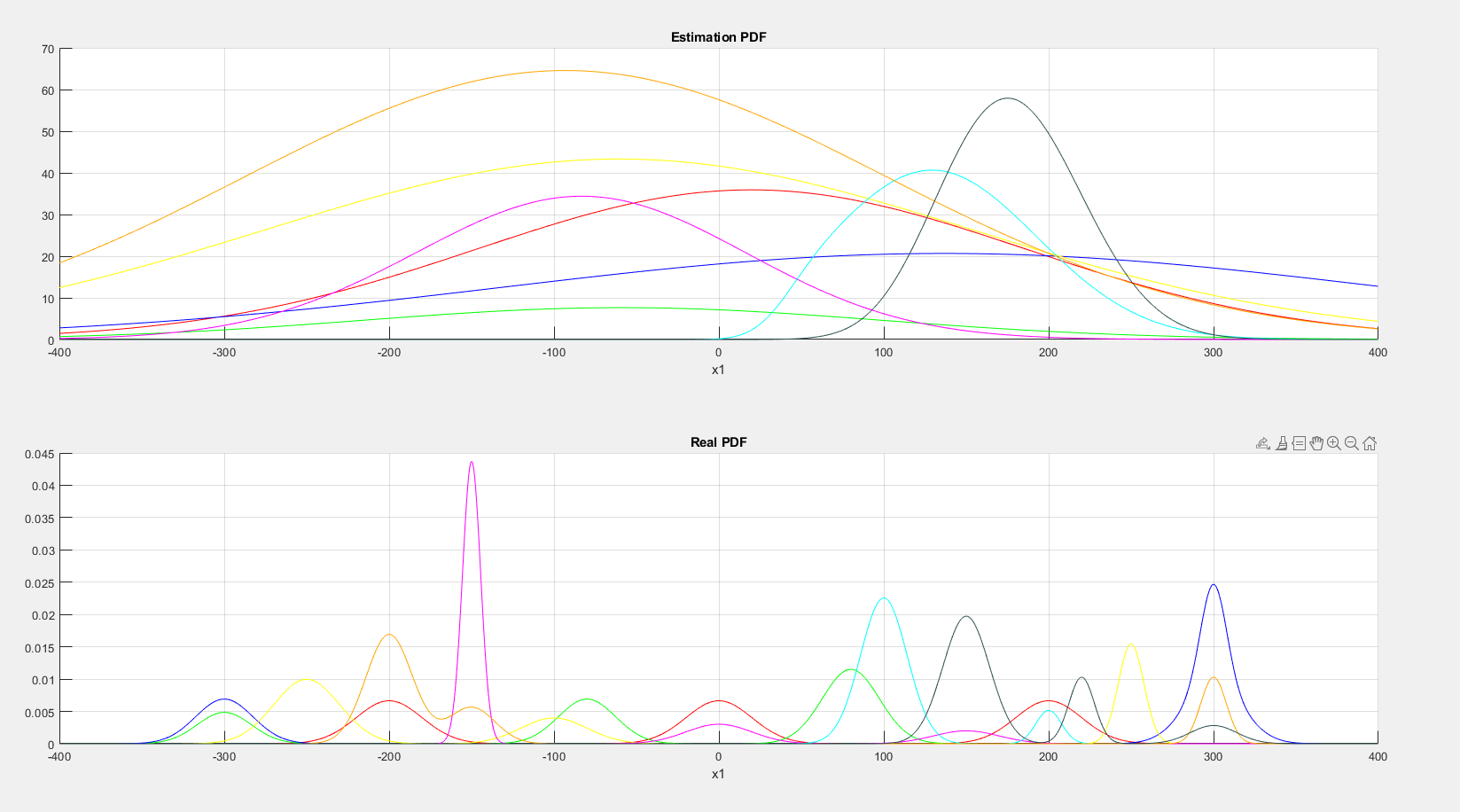


Рисунок 21 – Графики истинной плотности и плотности оценок X1

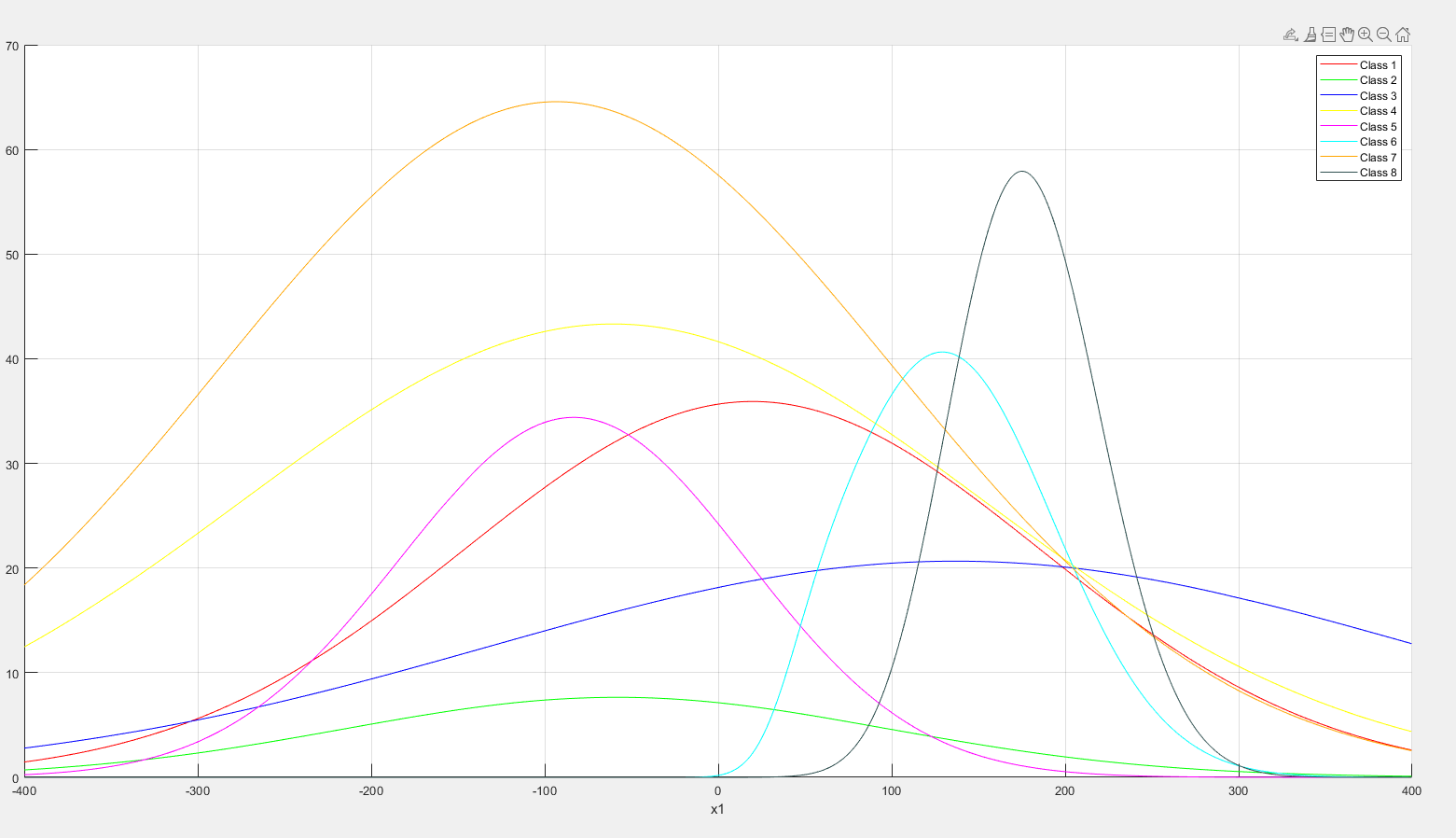


Рисунок 22 – График ошибок X1

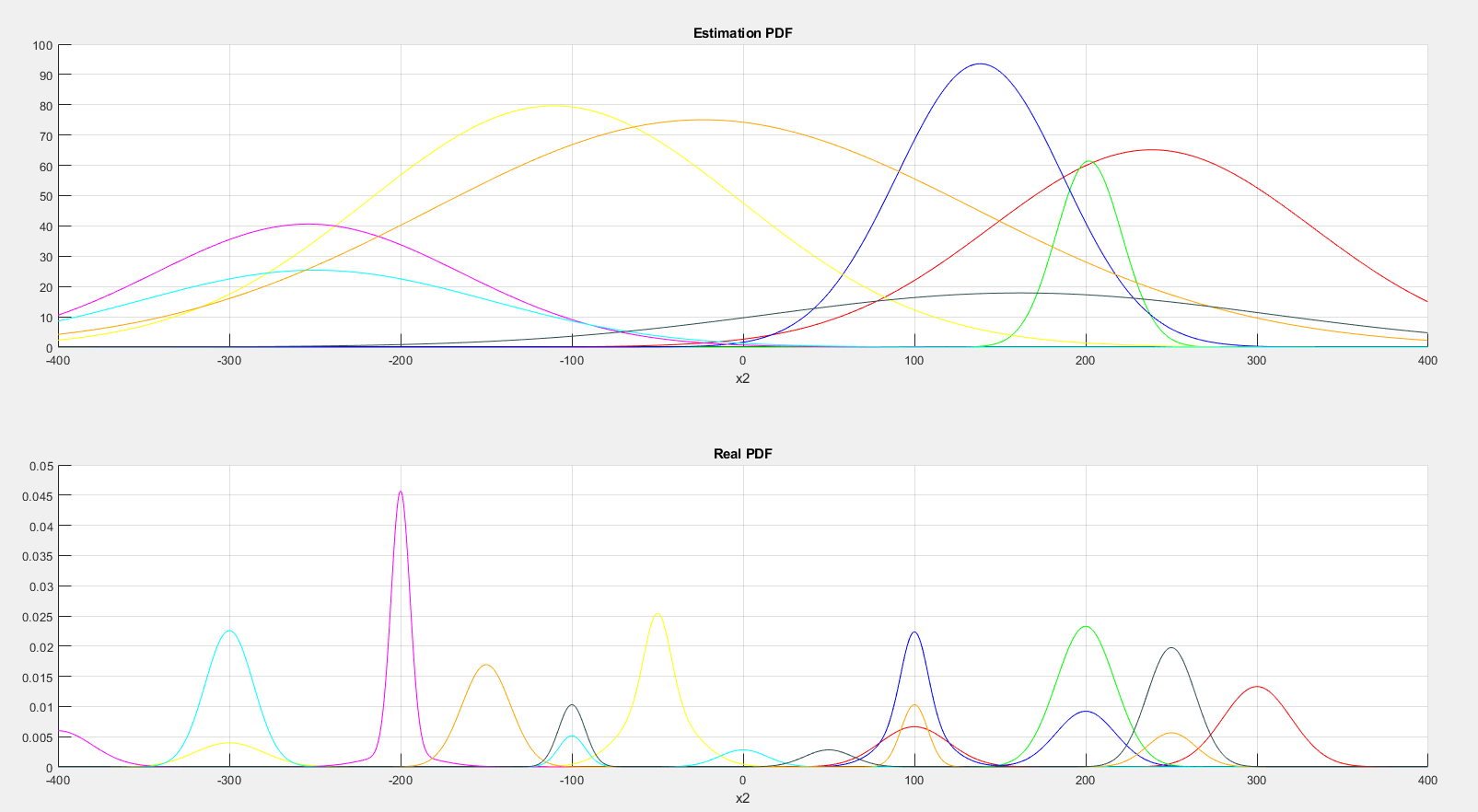


Рисунок 23 – Графики истинной плотности и плотности оценок X2

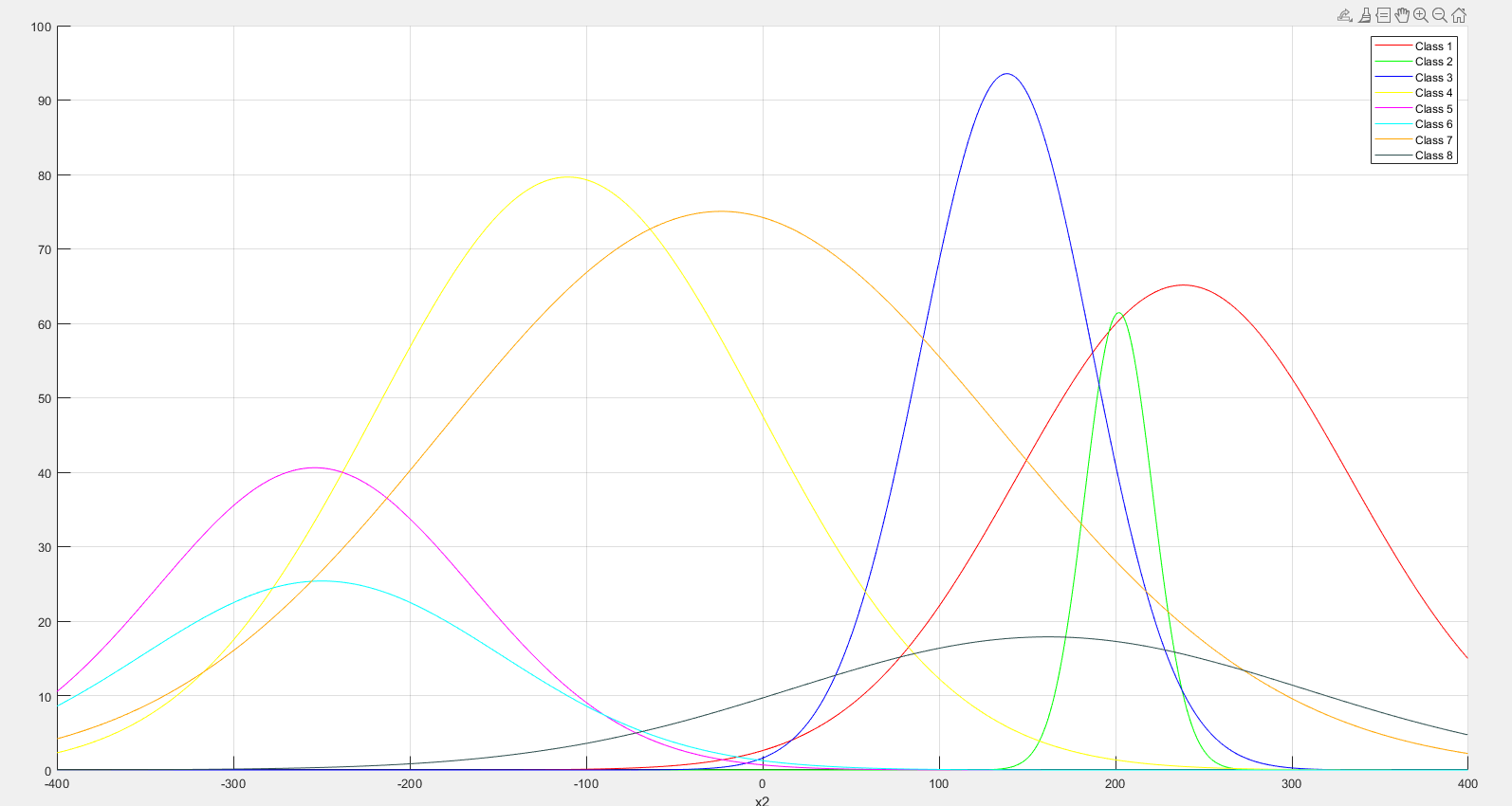


Рисунок 24 – График ошибок X2

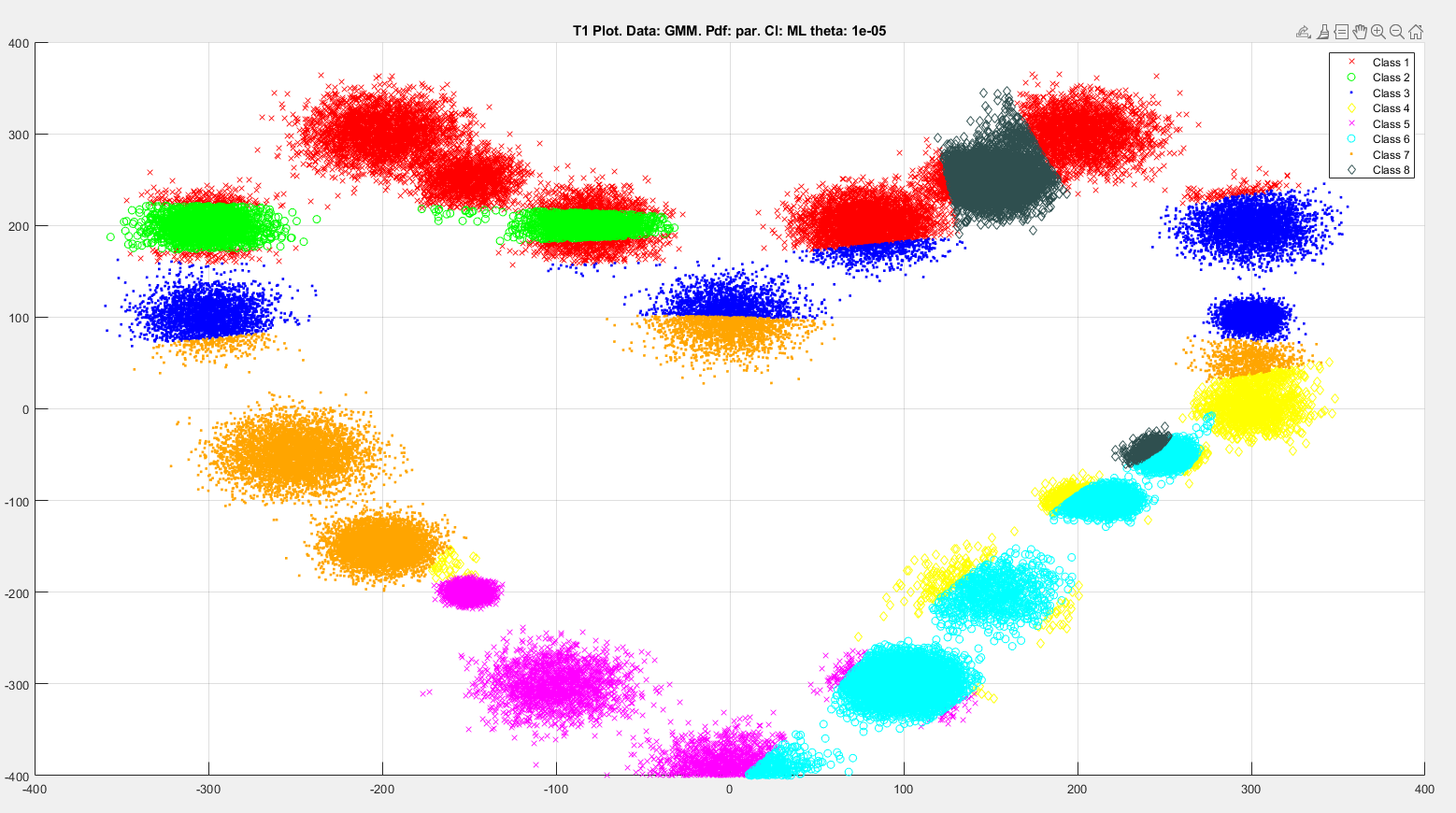


Рисунок 25 – График типа Т1

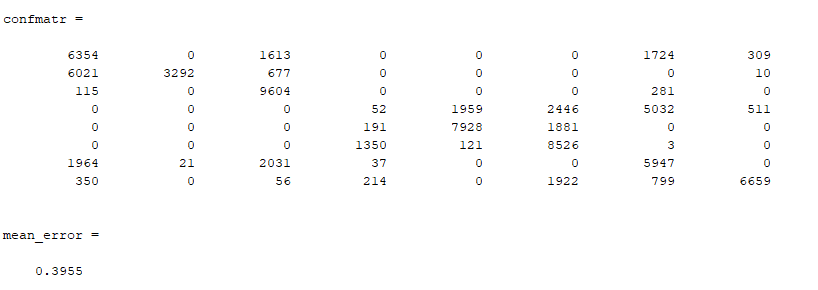


Рисунок 26 – Вывод в консоли при построении графика T1

Выводы по эксперименту:

Мы отключили предположение о том, что признаки независимы.

Применили вид нормализации для выравнивания дисперсий признаков. Признаки стали более сравнимыми. Форма распределений значительно изменилась по сравнению с результатами предыдущих экспериментов.

**Показатель Mean\_error значительно возрос (0.3955)**

Производительность модели, как и качество, сильно уменьшилась, следовательно, **нормализацию в данном случае использовать нецелесообразно.**

Включим опцию с отбеливанием, когда данные приводятся к единичной корреляционной матрице

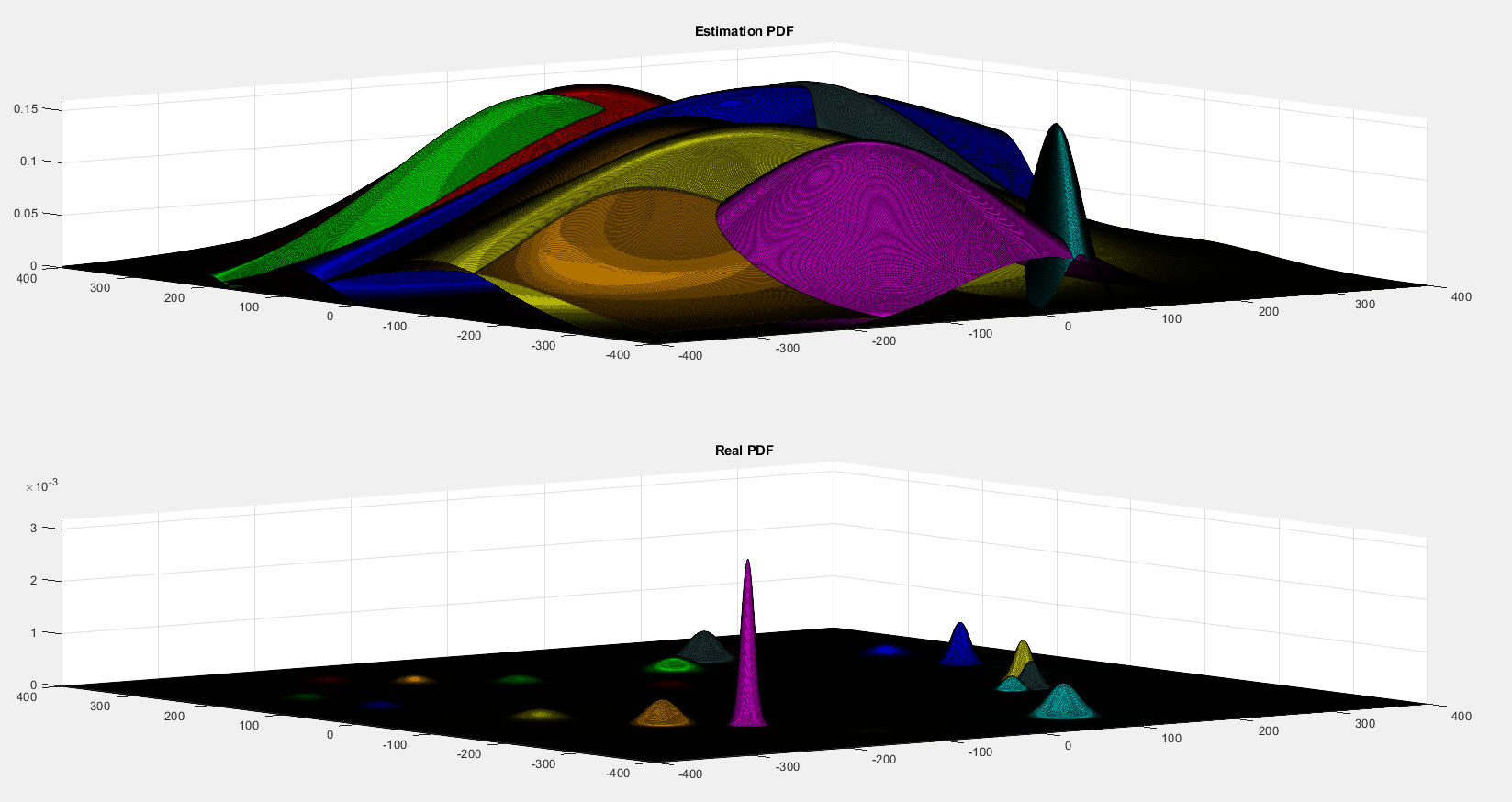


Рисунок 27– Графики истинной плотности и плотности оценок

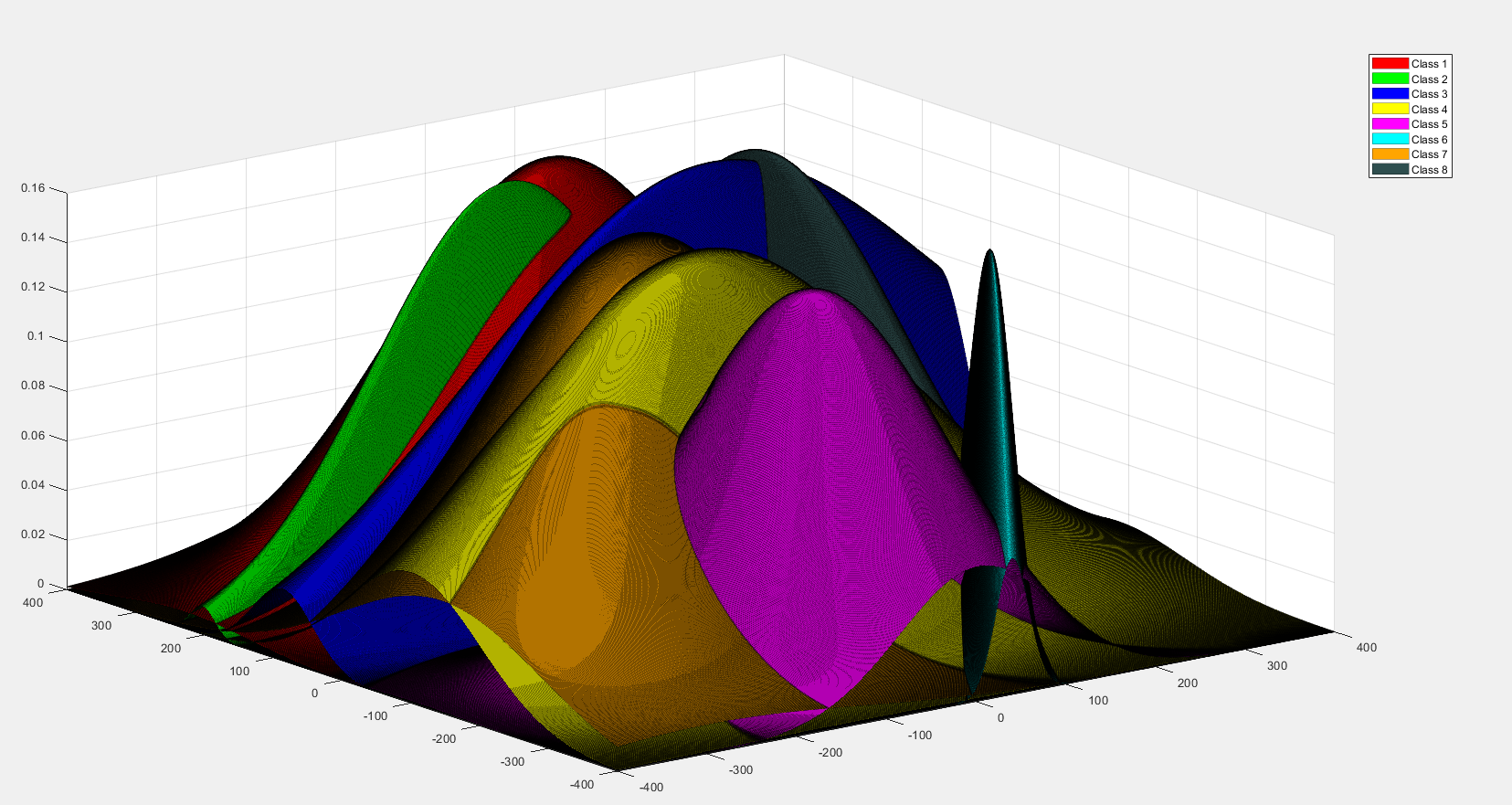


Рисунок 28 – График ошибок

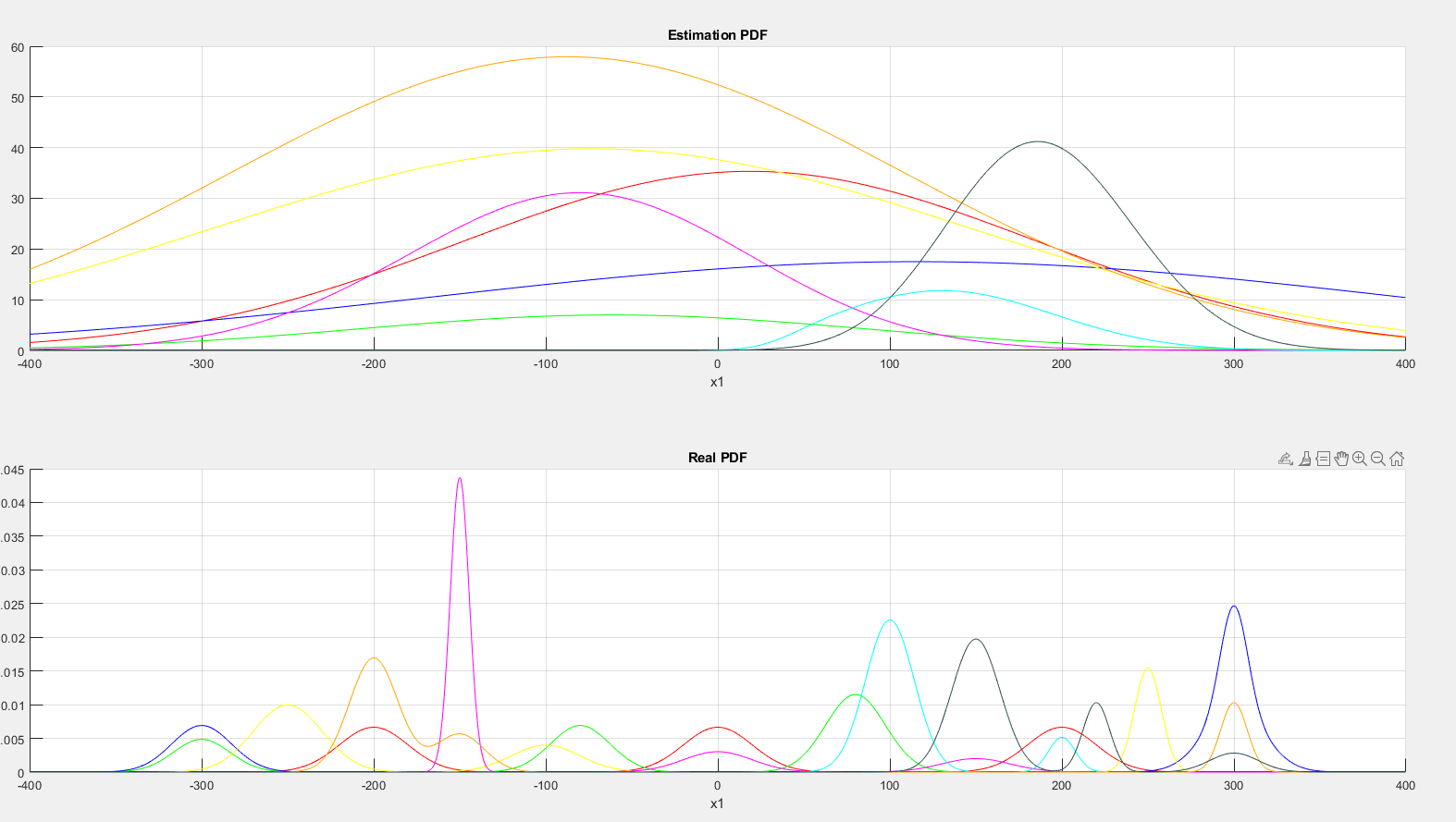


Рисунок 29 – Графики истинной плотности и плотности оценок X1

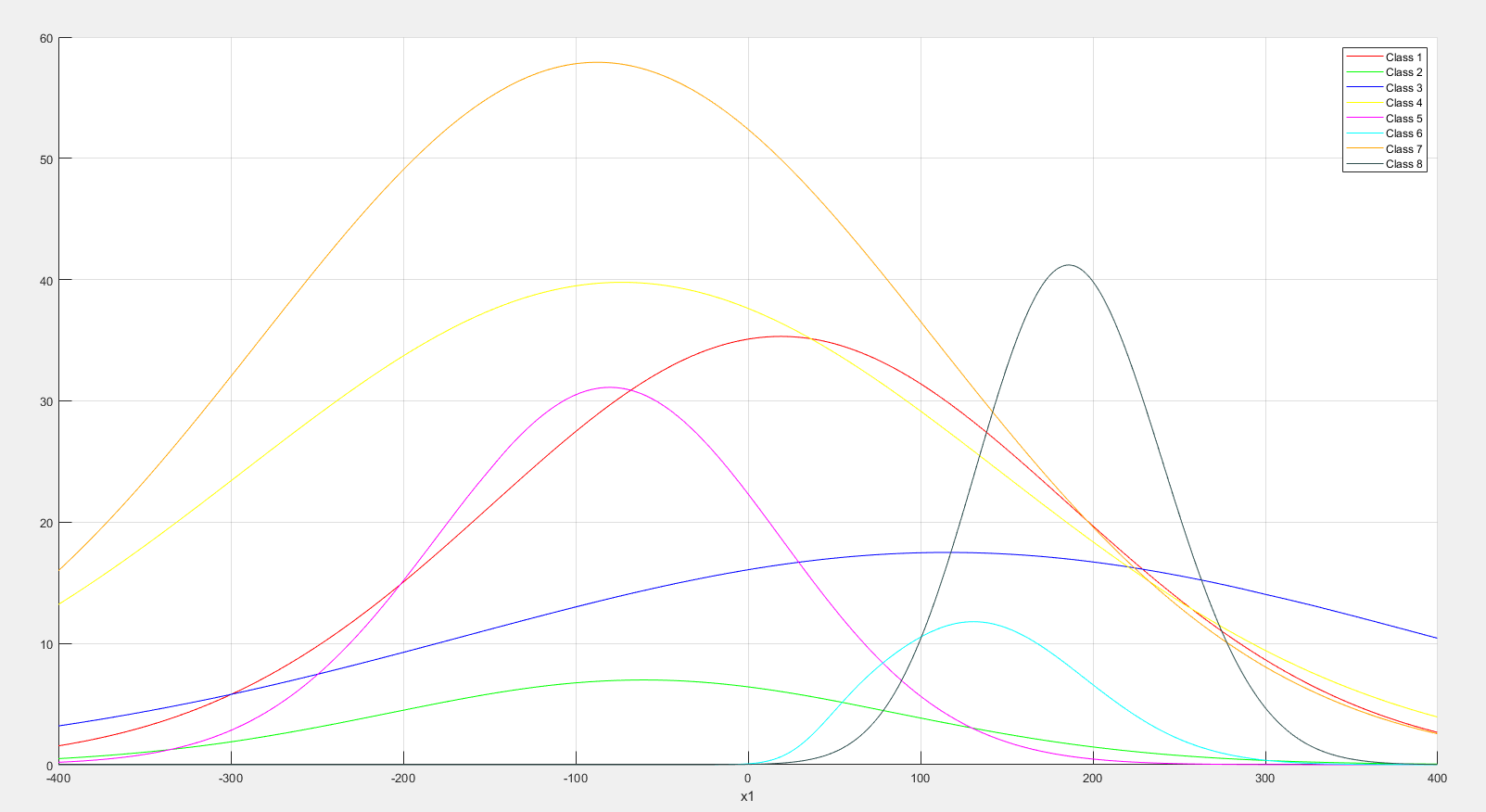


Рисунок 30 – График ошибок X1

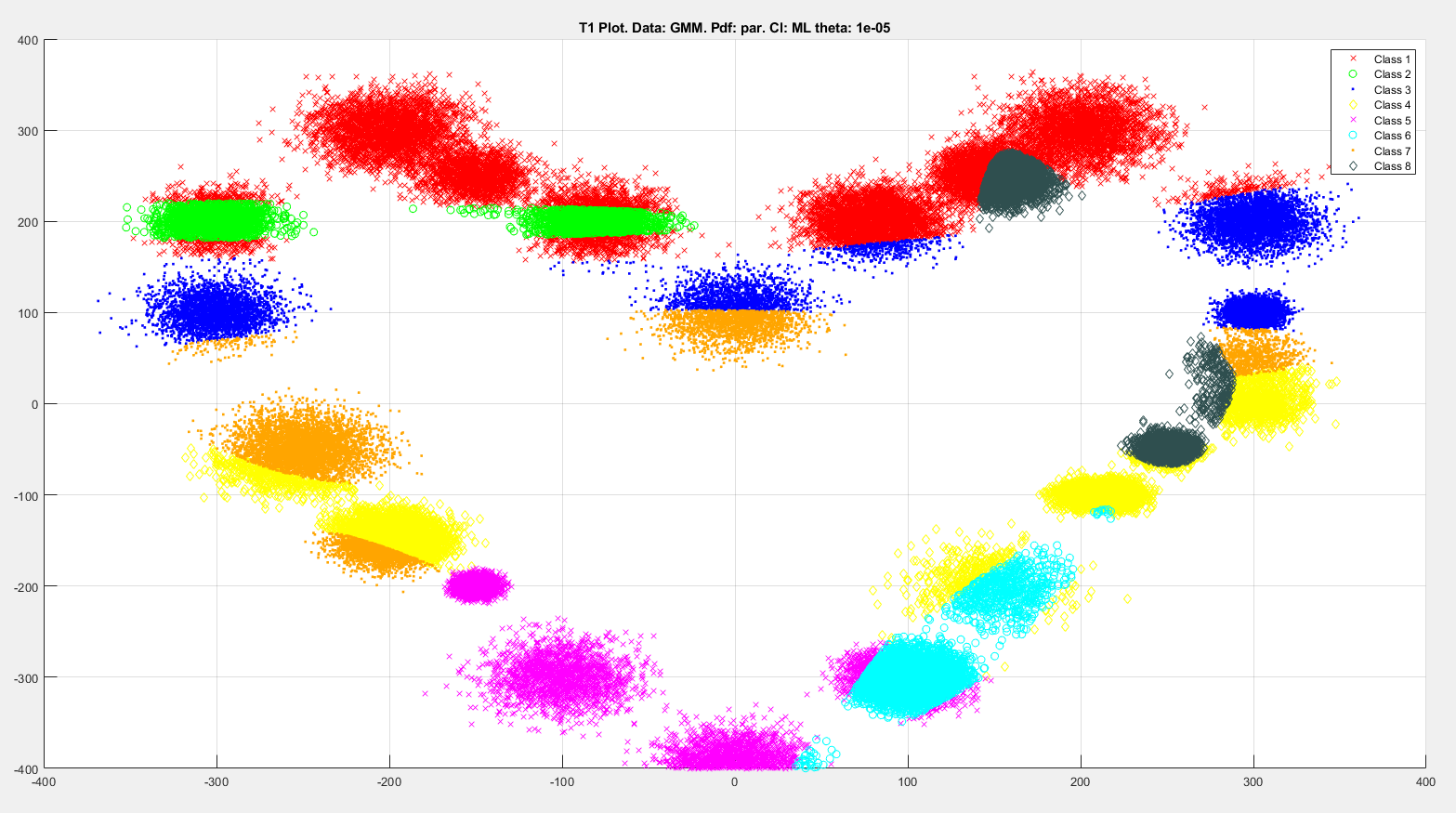


Рисунок 31 – График типа Т1

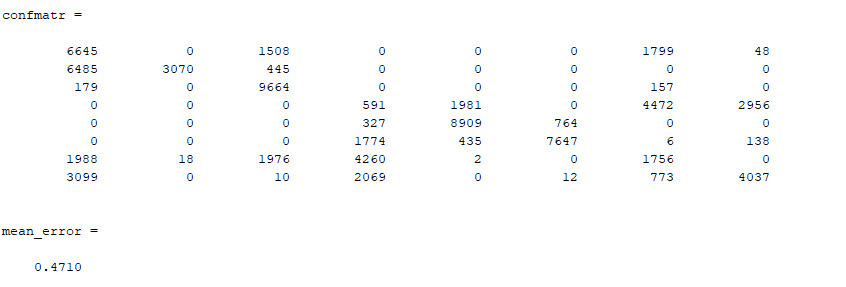


Рисунок 32 – Вывод в консоли при построении графика T1

Выводы по эксперименту:

Мы отключили предположение о том, что признаки независимы.

Применили вид нормализации для выравнивания дисперсий признаков. Признаки стали более сравнимыми. Форма распределений значительно изменилась по сравнению с результатами предыдущих экспериментов.

**Показатель Mean\_error значительно возрос (0.4710)**

Производительность модели, как и качество, сильно уменьшилась, следовательно, **нормализацию в данном случае использовать нецелесообразно.**

Используем гипотезу о нормальной смеси (в настройках выберите Par distribution = GMM).

****

Рисунок 33 – Установка параметров с двумя компонентами

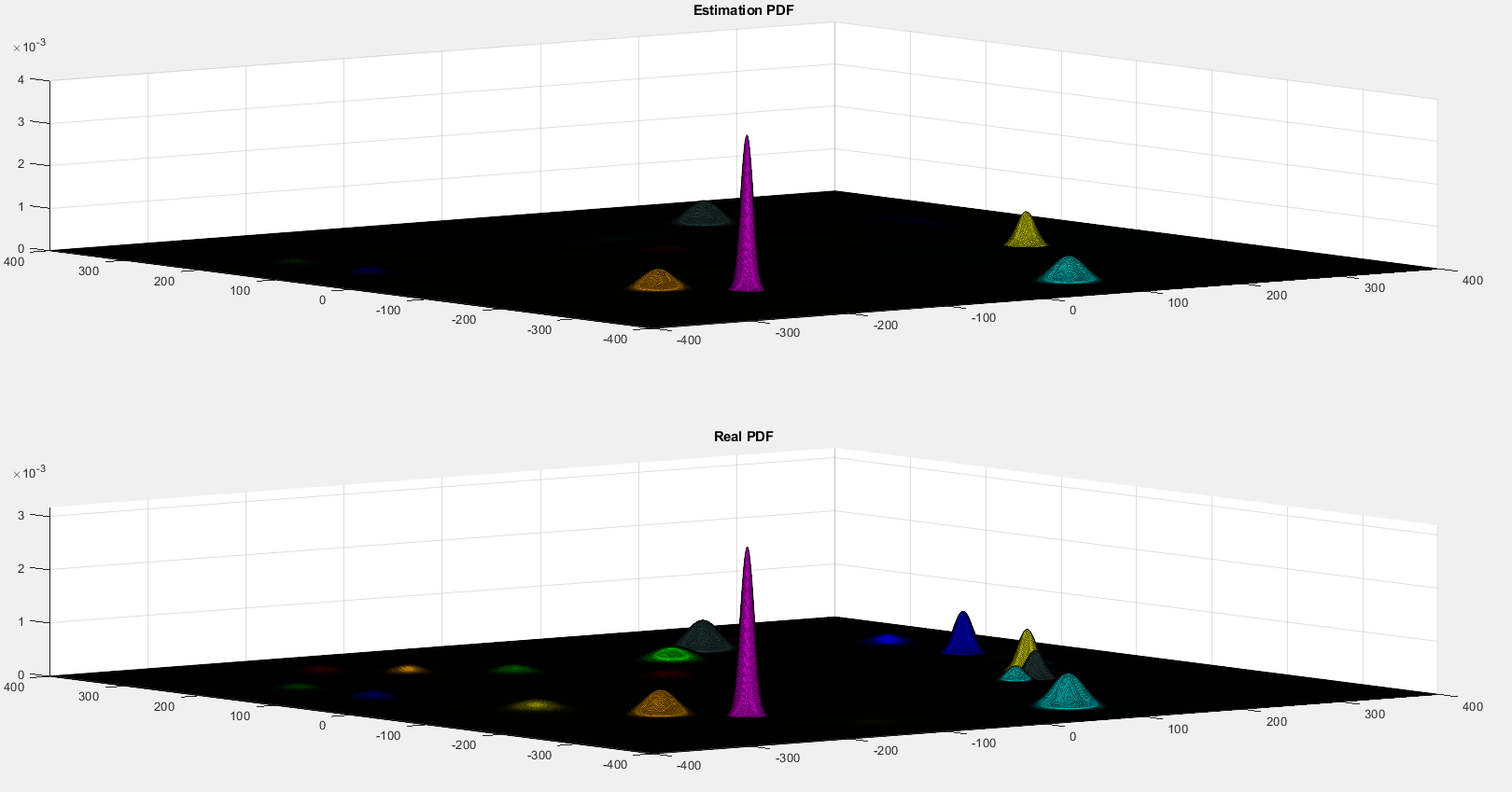


Рисунок 34– Графики истинной плотности и плотности оценок

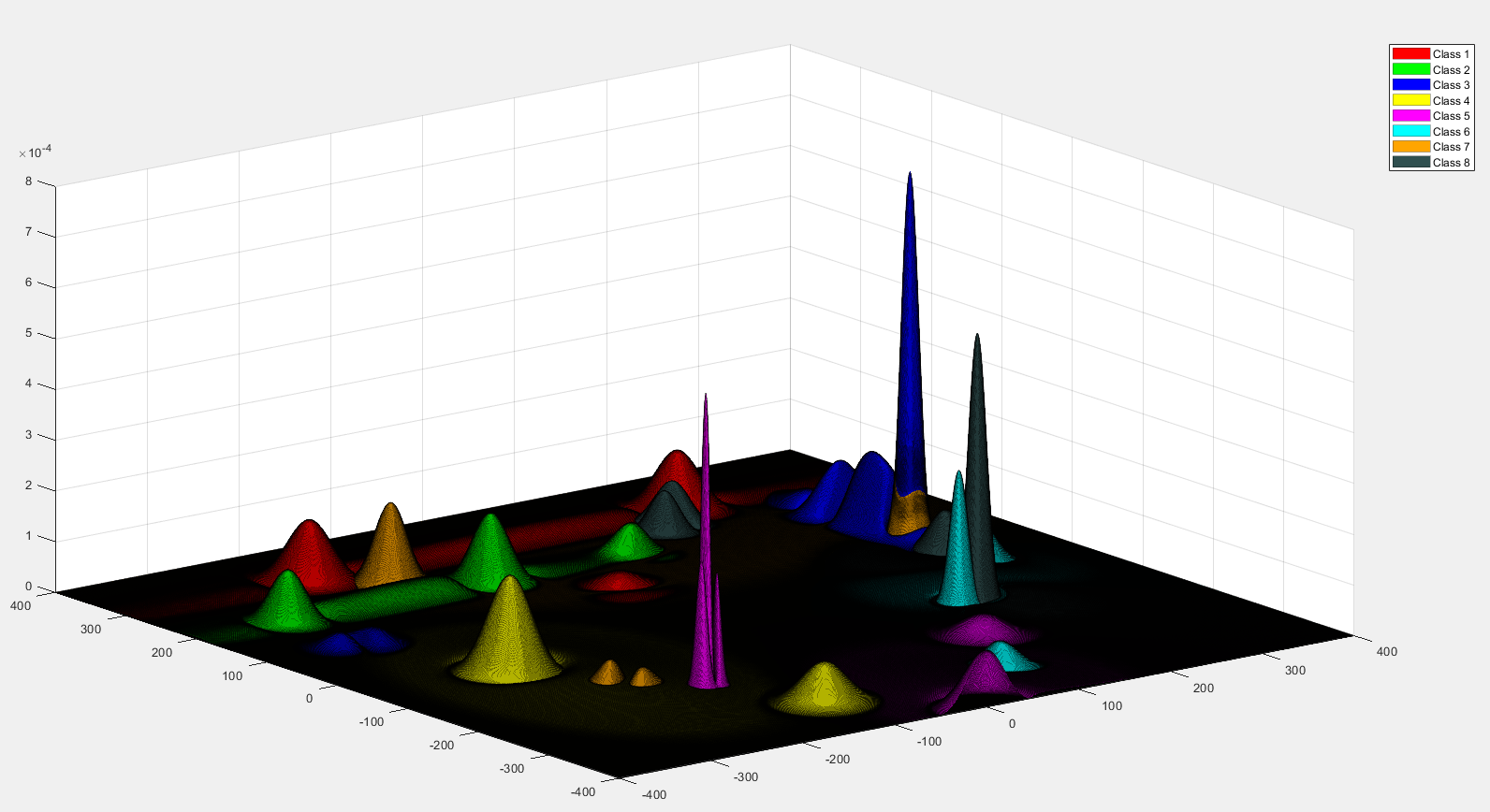


Рисунок 35 – График ошибок

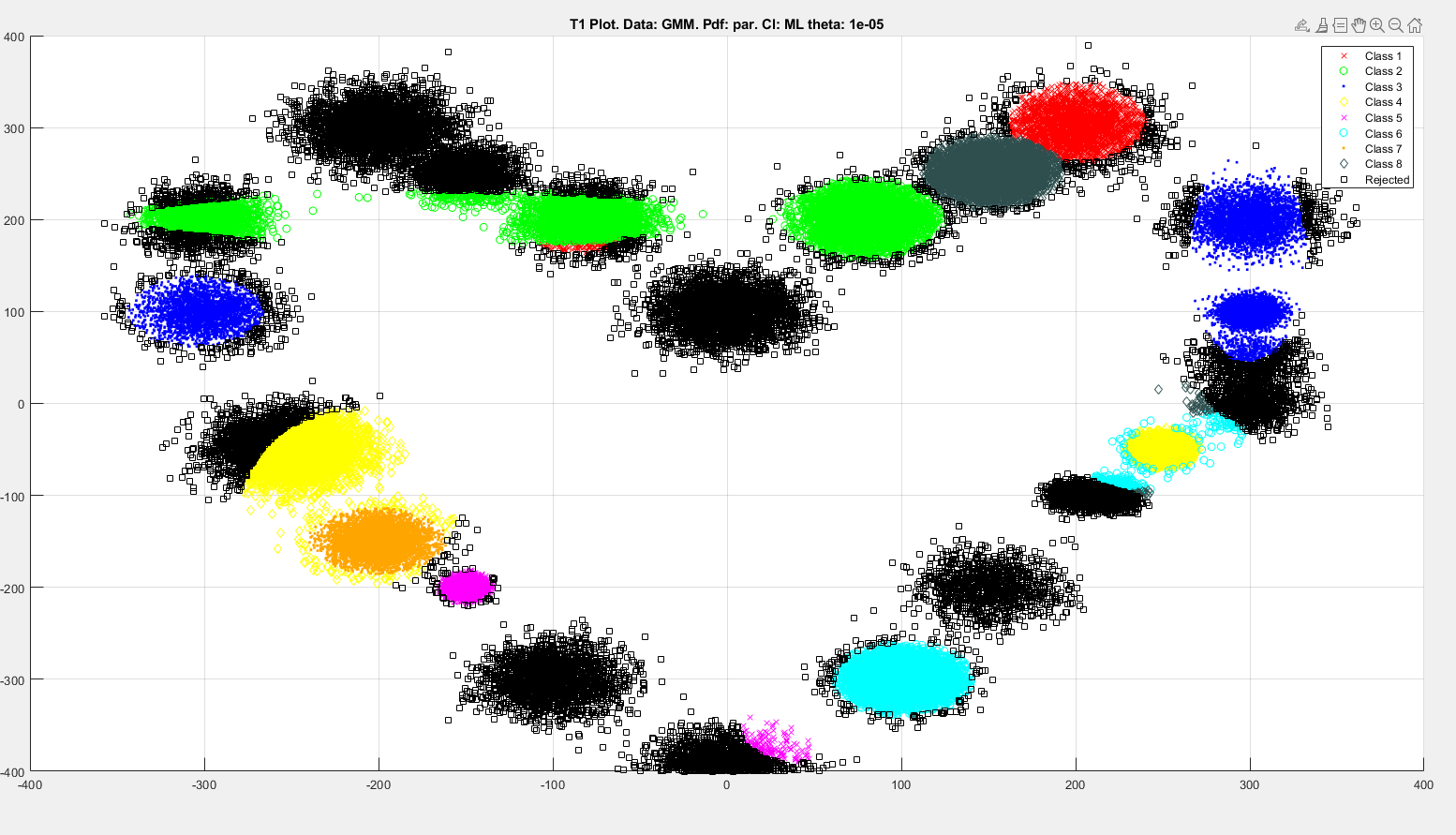


Рисунок 36 – График типа Т1

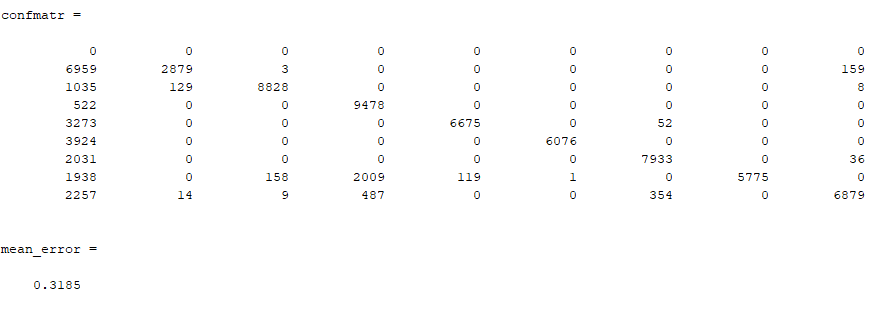


Рисунок 37 - Вывод в консоли при построении графика T1

Ошибка довольно велика.

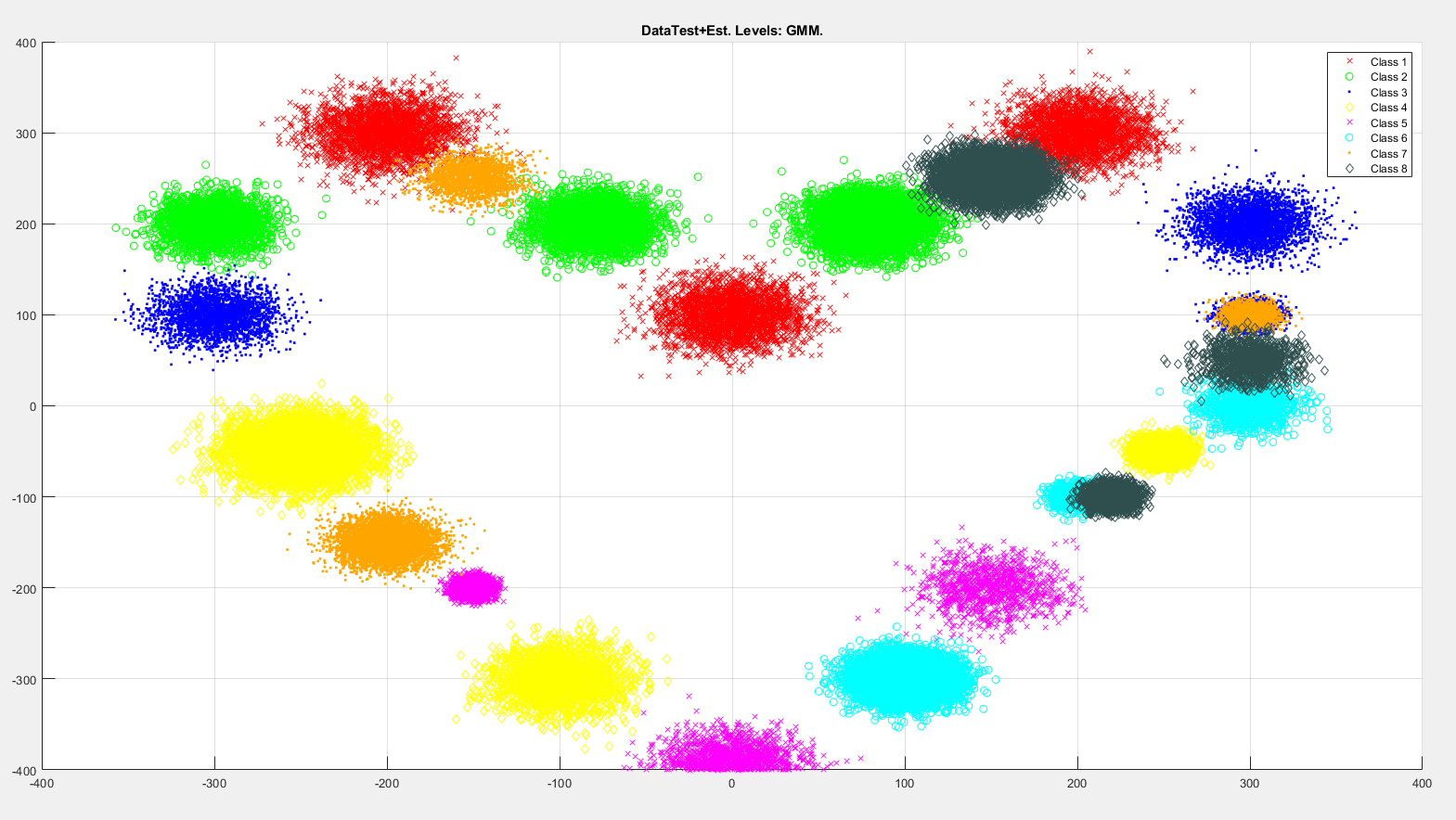


Рисунок 38 – Гауссовы смеси в виде их центров

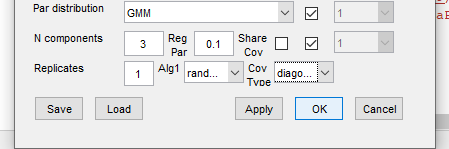


Рисунок 39– Установка параметров с двумя компонентами

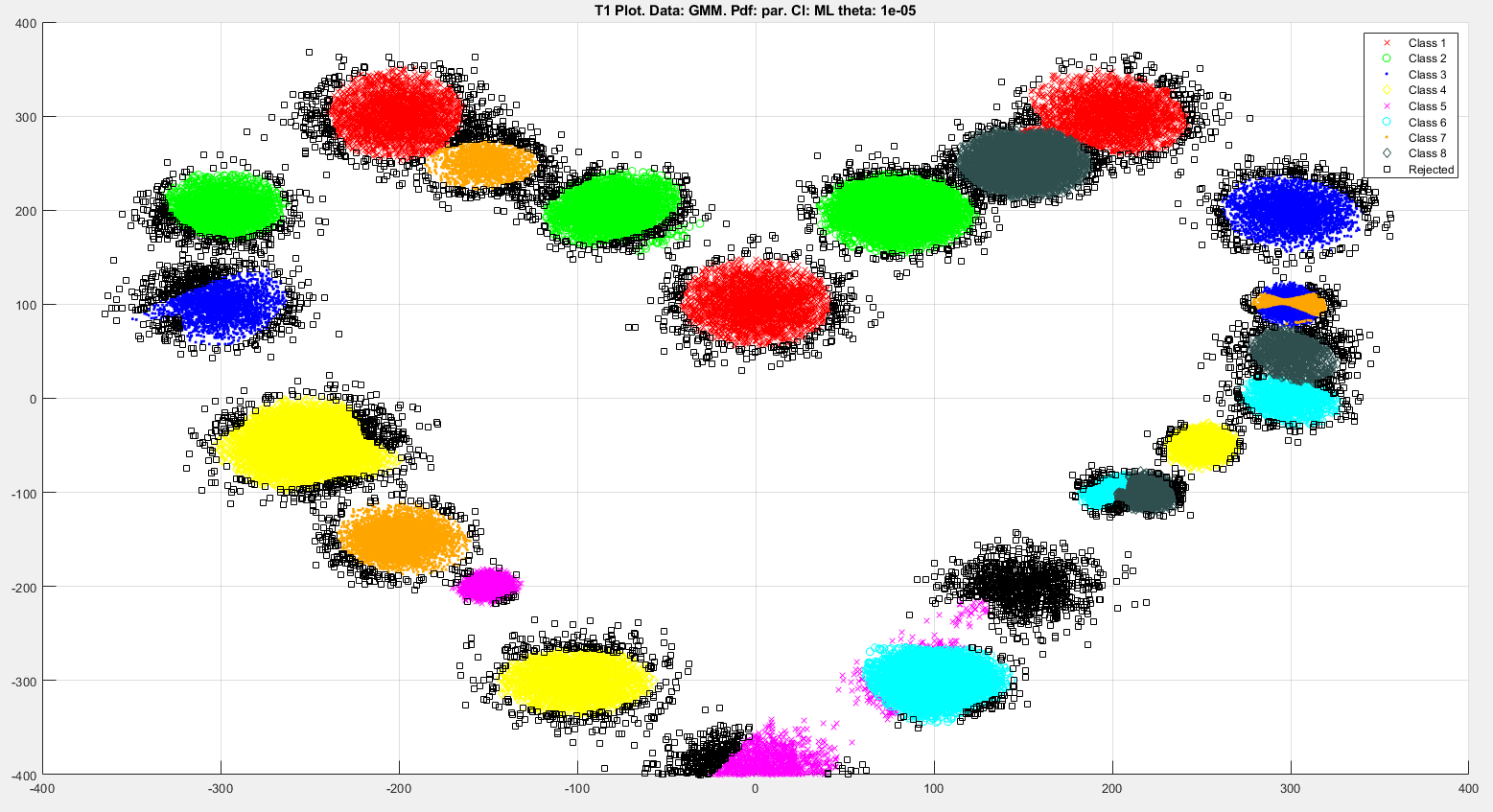


Рисунок 40 – График типа Т1

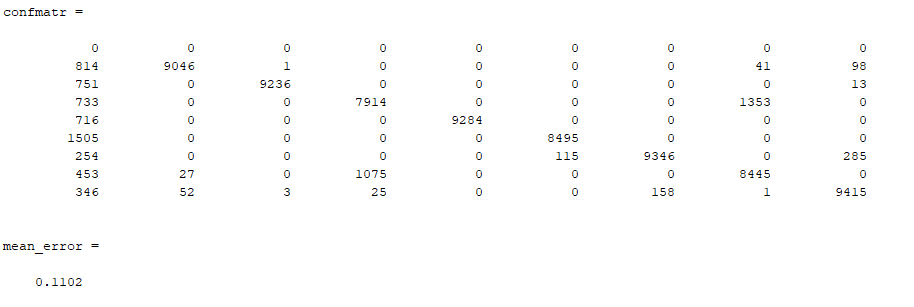


Рисунок 41 - Вывод в консоли при построении графика T1

Ошибка стала в три раза меньше, но все равно очень велика.

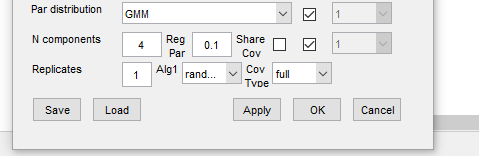


Рисунок 42– Установка параметров с двумя компонентами

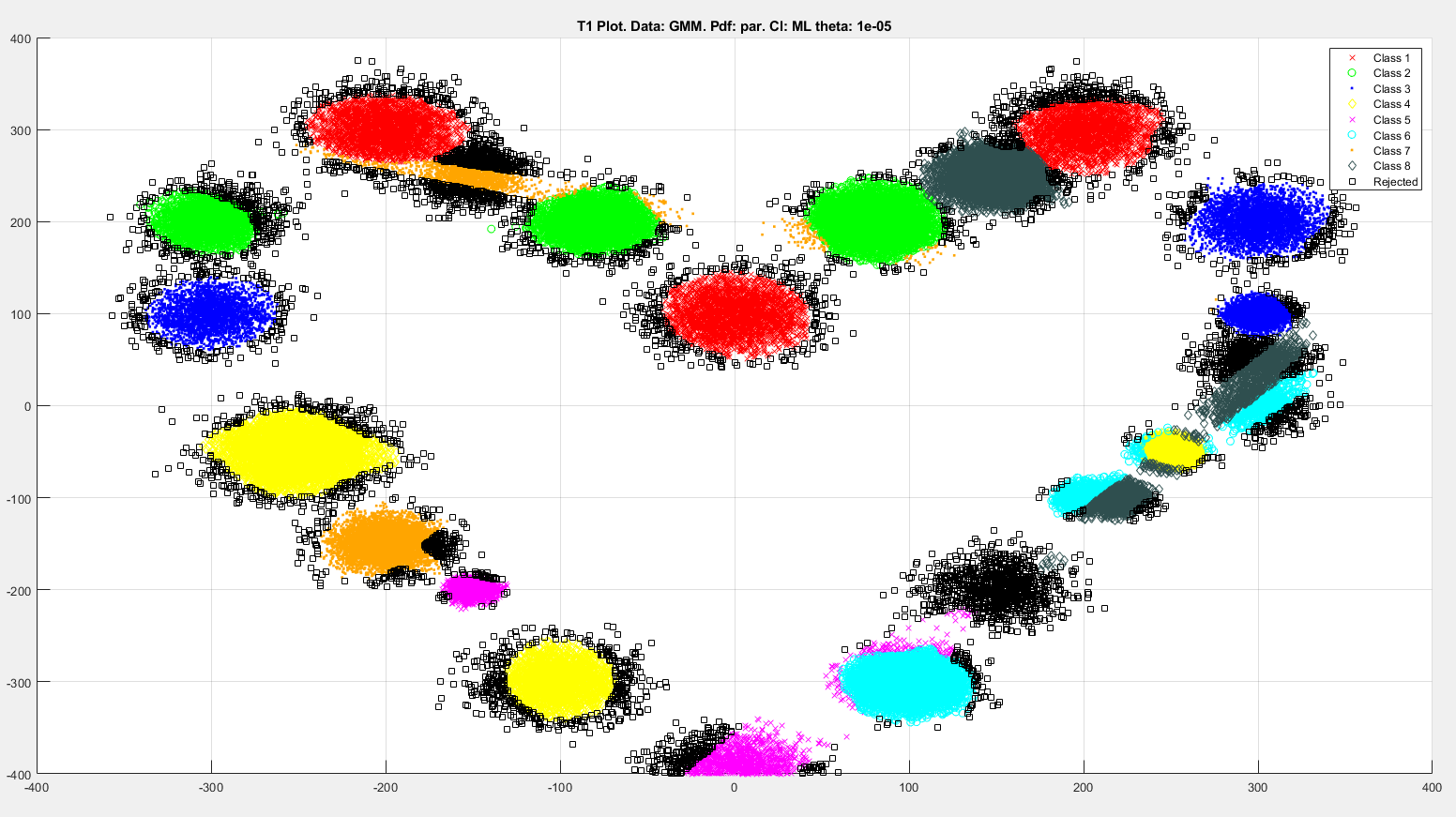


Рисунок 43– График типа Т1

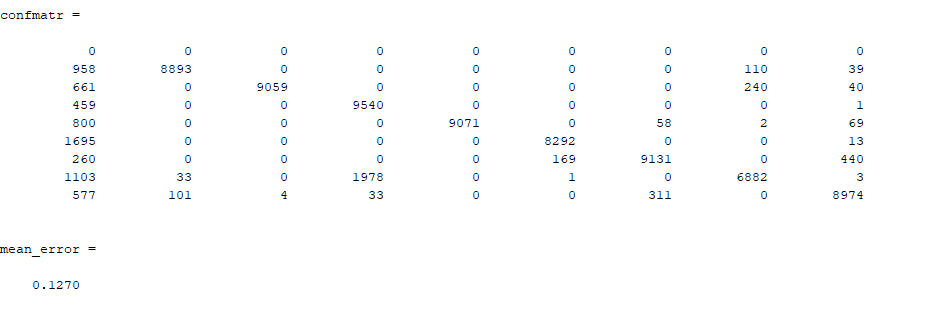


Рисунок 44 - Вывод в консоли при построении графика T1

Ошибка немного возросла.

**Выводы:**

**5.1** Методы параметрического оценивания подходят для моделей, которые предполагают определенное распределение данных. Например:

Нормальное (Гауссово) распределение: Параметрические методы, такие как GMM, предполагают, что данные имеют нормальное распределение. Это распределение широко используется в статистике из-за своей математической удобства и предположения о нормальности данных.

Биномиальное распределение: Используется для бинарных данных или счетных данных с фиксированным числом испытаний.

Пуассоновское распределение: Применяется для моделирования событий, происходящих с фиксированными интервалами времени или пространства.

Экспоненциальное распределение

**5.2**

**Преимущества:**

Интерпретируемость: Параметры модели имеют ясный смысл, что облегчает интерпретацию результатов.

Эффективность: При правильном выборе распределения параметрические методы могут быть более эффективными и точными.

Обобщение: Параметрические модели обычно легко обобщаются на новые данные, если они соответствуют тому же распределению.

**Недостатки:**

Ограничения в выборе распределения: Выбор неправильного распределения может привести к искаженным результатам.

Чувствительность к выбросам: Некорректные или аномальные данные могут исказить параметры модели, особенно в случае нормального распределения.

Ограничения в сложности модели: Параметрические модели ограничены сложностью выбранного распределения, что может быть недостаточно для точного моделирования сложных данных.

**5.3**

Сложность оценивания параметров модели GMM увеличивается с увеличением числа компонент. При большом количестве компонент оценка параметров может стать неустойчивой и требовать больших вычислительных ресурсов.

Способы борьбы с этим включают:

Уменьшение числа компонент: Попробуйте начать с небольшого числа компонент и постепенно увеличивать его, оценивая, когда добавление новых компонент перестает улучшать качество модели.

Использование начальных приближений: Используйте хорошие начальные приближения для параметров GMM, чтобы ускорить процесс обучения.

Регуляризация: Применяйте регуляризацию для уменьшения переобучения, особенно при большом количестве компонент.

Кластеризация данных: Попробуйте предварительно кластеризовать данные с использованием более простых методов кластеризации, а затем использовать GMM для каждого кластера.

**5.4**

Выбор оптимального числа компонент для модели GMM может быть сложной задачей. Некоторые методы включают:

**Критерии информационных критериев (AIC, BIC):** Выбирайте количество компонент, которое минимизирует AIC или BIC.

**Elbow method:** Постройте график зависимости критерия, такого как AIC или BIC, от количества компонент, и выберите точку, где изменение критерия замедляется (похоже на локоть).

**Кросс-валидация:** Используйте кросс-валидацию для оценки качества модели с разным числом компонент и выберите количество, которое дает наилучшее качество на тестовых данных.

**5.5**

Гипотеза о независимости признаков: Используется, когда предполагается, что признаки независимы друг от друга. Это может быть полезно в случае большого количества признаков, когда их взаимосвязи сложно моделировать.

Нормализация данных: Предварительная нормализация данных может улучшить производительность моделей, особенно в случае, если различные признаки имеют разные шкалы. Нормализация может ускорить сходимость методов оптимизации и улучшить обобщающую способность модели.

Оба эти подхода зависят от конкретного набора данных и задачи классификации. Важно экспериментировать с различными методами и анализировать их влияние на производительность модели.