Министерство науки и высшего образования Российской Федерации  
Новосибирский государственный технический университет  
Кафедра теоретической и прикладной математики

Отчет по проектной деятельности  
по теме:  
**Компьютерный анализ связи между конформационными свойствами мутантных белков и особенностями заболеваний с использованием методов статистического анализа и молекулярного моделирования. Глава 1.**

|  |  |
| --- | --- |
| Факультет:  Группа:  Студенты:  Преподаватель: | ПМИ  ПМИ-81  Воронкина Д.К  Бортников Н.Е  Тимофеев В.С |

Новосибирск

2021

Оглавление

[Задачи проекта 3](#_Toc73981407)

[Введение 3](#_Toc73981408)

[Почему был использован PLS алгоритм? 3](#_Toc73981409)

[Модель и Методы 4](#_Toc73981410)

[Алгоритм стандартногоPLS: 4](#_Toc73981411)

[Алгоритм робастного PLS: 4](#_Toc73981412)

[Результаты 5](#_Toc73981413)

[Стандартный алгоритм PLS 5](#_Toc73981414)

[Робастный PLS алгоритм. 6](#_Toc73981415)

[Вывод 11](#_Toc73981416)

# Задачи проекта

* Изучение алгоритма для построения модели оценки времени дожития пациентов с боковым амиотрофическим склерозом.
* Реализация алгоритма.
* Исследования алгоритма.

# Введение

Для построения времени дожития пациентов с боковым амиотрофическим склерозом был использован алгоритм partial least square (PLS) или частичная регрессия наименьших квадратов.

## Почему был использован PLS алгоритм?

Так как в задачах подобного типа исследуемые признаки в разы превышают количество испытуемых, множественную регрессию невозможно построить. Проведенный анализ данных показал существование тесных связей между определенными признаками, а иногда даже группами признаков.

Данный факт может усугубить проблему, поэтому учитывая все особенности данных было принято решение использовать регрессию частичных наименьших квадратов или partial least square. Метод решает эту задачу проецируя предсказанные переменные и наблюдаемые переменные в новое пространство.

PLS регрессия выделяет небольшое количество латентных переменных, в пространстве которых связь между зависимой переменной и предикторами достигает максимального значения.

Коэффициенты регрессии находятся с помощью алгоритма нелинейной итеративной оценки (NIPALS). В процессе этого алгоритма мы инициализируем переменные, заполняем матрицы весов W, матрицу проекции Т, матрицу нагрузки Р. Эти данные необходимы для получения коэффициентов B и B0.

# Модель и Методы

## Алгоритм стандартногоPLS:

1. X(0) = X – инициализация Х на нулевой итерации
2. w(0) = XTy/||XTy|| - поиск вектора весов на нулевой итерации
3. Для k=0 до N – цикл, где N – количество латентных компонент
4. t(k) = Xw – поиск вектора проекции на k итерации
5. p(k) = XTt/tTt – поиск вектора нагрузки на k итерации
6. q = yTt/tTt – поиск значения оценочного вектора коэфф. регрессии
7. Если q = 0, то заканчиваем цикл
8. Иначе X(k) = X – tpT – вычисление матрицы X на k итерации
9. Составить матрицу W из векторов w(k), матрицу T из векторов t(k), матрицу P из векторов p(k)
10. B = W(PTW)-1q – вычисление коэффициентов регрессии
11. B0 = q - PTB
12. Y = XB + B0

Так как практически все данные содержат ошибки и, в некоторых случаях, выбросы. Для улучшения качества оценки также стояла задача реализации робастного метода частичных наименьших квадратов.

При вычислении функции потерь была выбрана функция Хьюбера.



## Алгоритм робастного PLS:

1. X(0) = X – инициализация Х на нулевой итерации
2. w(0) = XTy/||XTy|| - поиск вектора весов на нулевой итерации
3. Для k=0 до N – цикл, где N – количество латентных компонент
4. t(k) = Xw – поиск вектора проекции на k итерации
5. p(k) = XTt/tTt – поиск вектора нагрузки на k итерации
6. q = yTt/tTt – поиск значения оценочного вектора коэфф. регрессии
7. Если q = 0, то заканчиваем цикл
8. Иначе X(k) = X – tpT – вычисление матрицы X на k итерации
9. Составить матрицу W из векторов w(k), матрицу T из векторов t(k), матрицу P из векторов p(k)
10. B=W(PTW)-1q – вычисление начального приближения коэффициентов регрессии.
11. B= argmin(), где  - функция Хьюбера, *b* – элемент вектора B.
12. B0 = q - PTB
13. Y = XB + B0

# Результаты

## Стандартный алгоритм PLS

Для оценки точности модели была использована стандартная среднеквадратичная ошибка. Ошибка находилась по формуле:

, где .

Для каждого фиксированного значения количества компонент, от 1 до 20 строилась модель PLS и находилась ошибка.

Так как PLS регрессия выделяет небольшое количество латентных переменных, в пространстве которых связь между зависимой переменной и предикторами достигает максимального значения. Можно заметить то, что увеличение количества латентных переменных или компонент, ведет к увеличению точности оценки.

Также для понимания достоверности модели была построена кросс-валидация.

1.Процедура повторялась N раз, где N – размерность вектора Y:

1. Обучающая выборка разбивается на i непересекающихся одинаковых по объему частей;

2. Производится i итераций. На каждой итерации происходит следующее:

1. Модель обучается на i-1 части обучающей выборки;

2. Модель тестируется на части обучающей выборки, которая не участвовала в обучении.

Кросс-валидация в данном случае представлена формулой:, где - полученные значения после оценки с учетом того, что отсутствует i-ая строка.

Можно заметить, что ошибка кросс-валидации в данном случае растет.

## Робастный PLS алгоритм.

При реализации робастного PLS алгоритма были проведены следующие исследования:

**1. Зависимость результатов работы алгоритма от различных методов минимизации.**

Результаты работы алгоритма показали, что метод минимизации напрямую влияет на ошибку робастного метода. Рассматривался алгоритм: Хука и Дживса, Нелдера Мида и Пауэлла. Параметр функции Хьюбера брался равным 1,345.

Для выяснения корректности работы робастного алгоритма, было поставлено аномальное значение в точку Y[5] = 100.

Ошибка кросс валидации имела следующие значения:

Так как робастный алгоритм с минимизацией Хука и Дживса не давал сильных отличий в ошибке от классического PLS алгоритма, а минимизация с помощью метода Пауэлла давала слишком маленькую ошибку на первых компонентах, что тоже не верно так как она описывала выброс, было принято решение рассматривать дальнейшие исследования основываясь на методе минимизации Нелдера Мида.

**2. Зависимость результатов работы алгоритма от различных значений параметра функции Хьюбера.**

Для функции Хьюбера



рассматривалось несколько значений параметра δ.

δ={0.18442; 0.3; 0.5; 0.75; 1; 1.345; 2; 2.35; 3}

Было замечено что ошибка робастного метода с увеличением значения параметра Хьюбера уменьшается, вне зависимости от количества компонент. При этом чем больше значение количества латентных переменных (компонент), тем меньшее значение принимает ошибка робастного метода.

На диаграмме представлены результаты с выбросом 100 в одной из точек.

Ошибка кросс валидации ведет себя следующим образом.

На диаграммах представлены результаты с выбросом 100 в одной из точек.

-Уменьшалась, если количество компонент принимало значение до 6, включительно:

-Увеличивалась, если количество компонент было больше, либо равно 7.

Если выброс был равен 50, то ошибка кросс валидации увеличивалась, вне зависимости от количества компонент.

Если выбросов не наблюдалось, то ошибка кросс валидации росла более плавно.

# Вывод

Перед проектом стоит цель интерпретации полученных результатов. Определение оптимального числа компонент для оценки времени дожития пациентов с боковым амиотрофическим склерозом.