Метод главных компонент

5) (Д.А.Борзых, Б.Б.Демешев, задача 7.4)
Пионеры, Крокодил Гена и Чебурашка собирали металлолом несколько дней подряд. В распоряжение иностранной шпионки, гражданки Шапокляк, попали ежедневные данные по количеству собранного металлолома: вектор g – для Крокодила Гены, вектор h – для Чебурашки и вектор x – для пионеров.Гена и Чебурашка собирали вместе, поэтому выборочная корреляция $\hat{cor}(g,h) = -0.9$. Гена и Чебурашка собирали независимо от пионеров, поэтому $\hat{cor}(g,x) = 0$, $\hat{cor}(h,x) = 0$. Если регрессоры g, h, x центрировать и нормировать, то получится матрица $\tilde{X}$.
1) Найдите параметр обусловленности матрицы $\tilde{X}'\tilde{X}$.
2) Вычислите одну или две главные компоненты (выразите их через вектор-столбцы матрицы $\tilde{X}$), объясняющие не менее 70% общей выборочной дисперсии регрессоров.

4. (5 баллов) Исследователь Д'Артаньян стандартизировал (центрировал и нормировал) все имеющиеся регрессоры и поместил их в столбцы матрицы $\tilde{X}$. Выборочная корреляционная матрица регрессоров равна:

$$\begin{pmatrix} 1 & 0.85 & 0 \\ 0.85 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

   а) Найдите параметр обусловленности (condition number) матрицы $\tilde{X}^T\tilde{X}$.
   б) Вычислите одну или две главные компоненты, объясняющие не менее 70% суммарной дисперсии стандартизированных регрессоров. Выпишите найденные компоненты как линейные комбинации столбцов матрицы $\tilde{X}$.

LASSO и Ridge регрессии

1. (5 баллов) Рассмотрим алгоритм LASSO с параметром регуляризации λ для модели $Y = X\beta + \varepsilon$, где все переменные центрированы.
   а) Выпишите целевую функцию алгоритма.
   б) Что произойдет с оценками $\hat{\beta}_{LASSO}$ при $\lambda \to \infty$?
   в) Что произойдет с оценками $\hat{\beta}_{LASSO}$ при $\lambda \to 0$?

LASSO и Ridge оценки

1) Используя данные файла Dougherty, оцените LASSO регрессию с зависимой переменной EARNING.
2) Найдите оптимальное значение параметра регуляризации с помощью кросс-валидации.
3) Постройте графики оценок коэффициентов при выбранных факторов при различных значениях параметра регуляризации.
4) Какие из оценок коэффициентов отличны от нуля при оптимальном значении параметра регуляризации?