Исследование конформационных изменений белков при помощи L_1 регуляризации

Moscow Institute of Physics and Technology

daniil.emcev.ru@yandex.ru, *ryabinina.rb@phystech.edu*

22 апреля 2019 г.

Цель работы

Конформационные изменения белков

Исследовать методы L_1 регуляризации, способные приближать конформационные изменения белков в пространстве торсионных углов

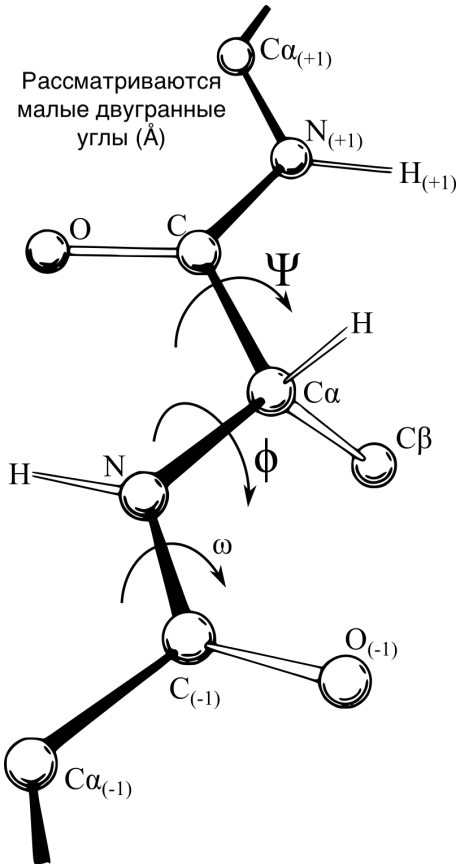
Проблемы

Методы L_1 регресии работают быстрее чем методы L_2 за счет того пространство торсионных углов разреженно. Они также позволяет выбрать произволное число углов, что снижает размерность. Методы L_1 позволяют получить интерпретируемые модели - отбираются признаки, оказывающие наибольшее влияние.

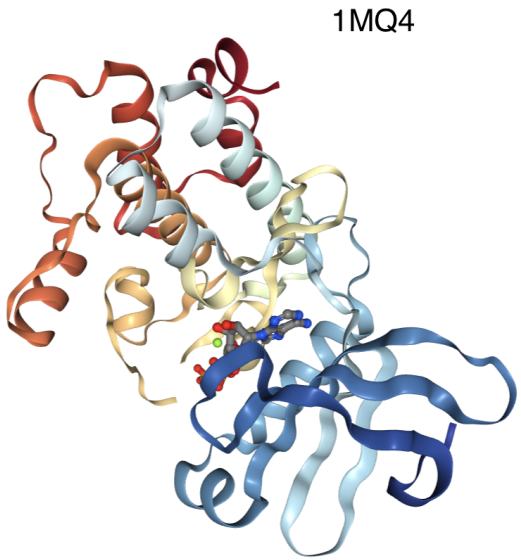
Модели

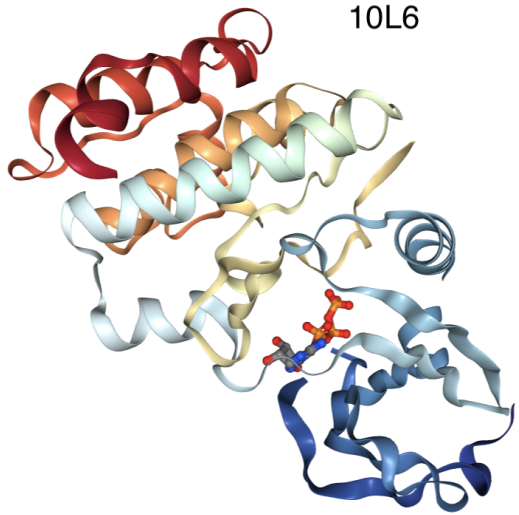
L_2: Ridge regression; L_1: LASSO, Elastic-net, LARS

Предмет исследования



'Скелет' белка





Конформационное изменение белков с одинаковой цепочкой атомов

https://en.wikipedia.org/wiki/Dihedral_angle

Литература

- R. Mendez and U. Bastolla, Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins.
- A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model
- F. Tama and Y. H. Sanejouand, Conformational change of proteins arising from normal mode calculations.
- H. G. Dos Santos, J. Klett, R. Mendez, and U. Bastolla, Characterizing conformation changes in proteins through the torsional elastic response.
- R. Tibshirani, Regression shrinkage and selection via the lasso
- H. Zou and T. Hastie, Addendum: Regularization and variable selection via the elastic net
- B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression

Постановка задачи

Рассматриваются две структуры одного белка, представляющимие из себя различные конформации.

$$\Delta r = J\Delta\phi$$

$\Delta r = r^A - r^B$, A,B - конформации белка, $r_i^{A,B} \in R^3$, i - номер атома,
$\Delta r \in R^{3n}$ - декартовых координаты, $\Delta\phi$ - торсионные углы,
J - матрица Якоби $\frac{\Delta r_{int_i}}{\Delta\phi_a}$, M - диагональная матрица весов
Для формулировки проблемы используется метод наименьших квадратов и предположение о том, что углы малы

$$\min_{\Delta\phi} (\Delta\phi, J^\top MJ\Delta\phi) - 2(\Delta\phi, J^\top M\Delta r).$$

Этот метод неприменим к коррелированным переменным, которые являются компонентами матрицы Якоби, потому что он приводит к переобучению.
Таким образом, мы рассматриваем подход регуляризации L_1, позволяющий выбирать наиболее важные признаки.

Работа с RCSB Protein Data Bank

Датасет

Тестовый набор сформирован из 26 белков, каждый из которых представлен в двух конформациях, включающих широкий спектр макромолекулярных движений. Размеры варьировались от 100 до 1000 аминокислот. Все структуры не имеют разорванных цепей и пропущенных атомов. Набор состоит из белков, которые демонстрируют крупномасштабные коллективные тепловые движения. Данные загружены из RCSB Protein Data Bank.

RMSD

Для оценки расстояния между белками используется RMSD - среднеквадратичное отклонение позиций атомов. В работе сравниваются $RMSD_{(initial,final)}$ и $RMSD_{(initial,predicted)}$

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

Сравниваются

- Ridge regression

$$\min_{\Delta\phi} (\Delta\phi, J^\top MJ\Delta\phi) - 2(\Delta\phi, J^\top M\Delta r) + \lambda(\Delta\phi, \Delta\phi)$$

- Least absolute shrinkage and selection operator

$$\min_{\Delta\phi} (\Delta\phi, J^\top MJ\Delta\phi) - 2(\Delta\phi, J^\top M\Delta r) + \lambda \sum_{j=1}^{p} |\Delta\phi_j|$$

- Elastic-net regularization

$$\min_{\Delta\phi} (\Delta\phi, J^\top MJ\Delta\phi) - 2(\Delta\phi, J^\top M\Delta r) + \alpha(\Delta\phi, \Delta\phi) + (1-\alpha) \sum_{j=1}^{p} |\Delta\phi_j|$$

- Least-angle regression

$$\min_{\Delta\phi} \|\Delta r - J\Delta\phi\|_2^2 + \alpha s^\top \Delta\phi \text{ , при } s_j = 0, \phi_j = 0, s_j = 1, \phi_j > 0, s_j = -1, \phi_j < 0$$

Исследованные методы

Lasso with cross validation

Разделить набор данных на 10 частей, используя координаты спуска из библиотеки sklearn

Lasso with grid search and cross validation

Автоматическая настройка гиперпарамера α по сетке

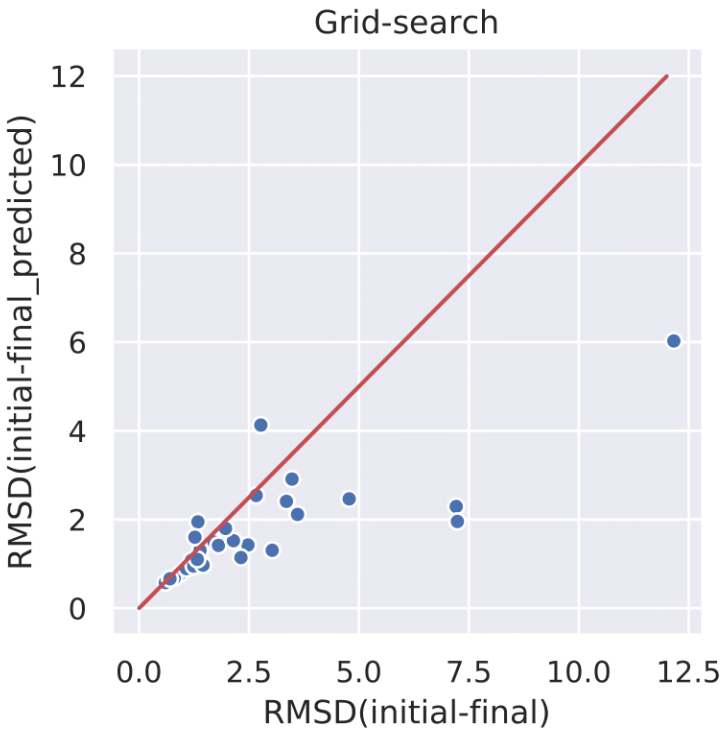
Elastic net regularization

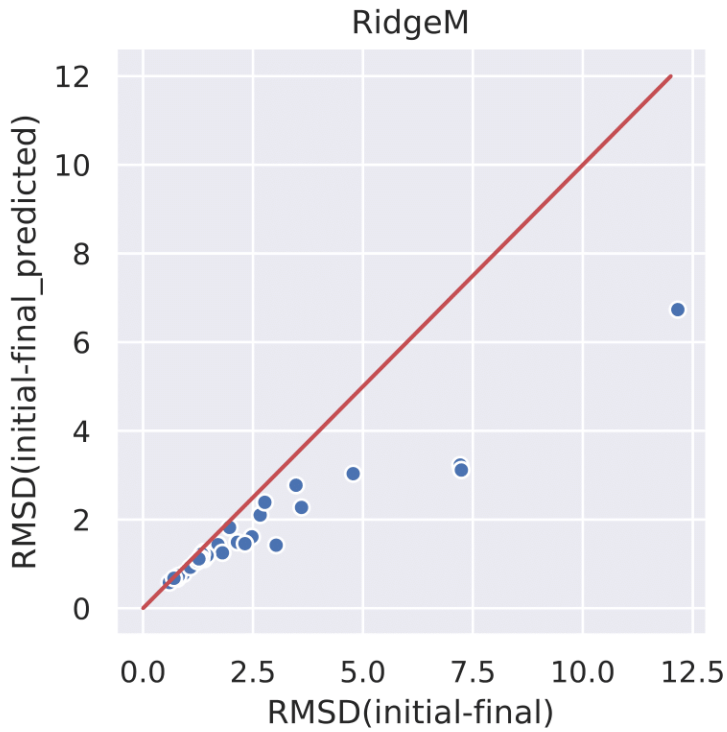
Использование Ridge regression и LASSO одновременно

LARS

Для каждой пары белков рассмотрен путь компонент (500 итераций) и число ненулевых компонент для набора весов на каждой итерации. В случае нахождения ряда из числа ненулевых компонет, производилась сортировка с использованием loss function и выбиралось RMSD соответвующее наименьшему значению.

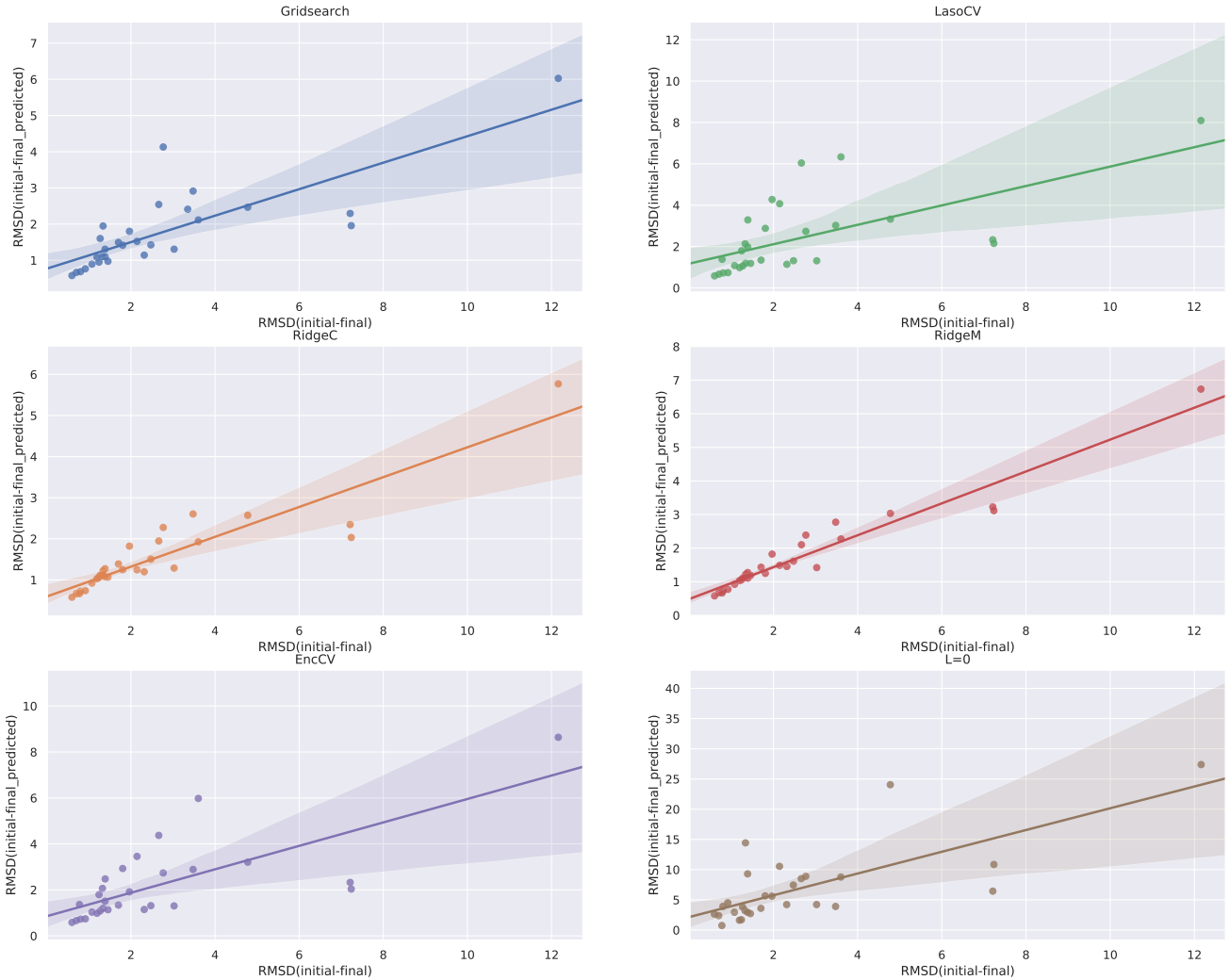
Корреляция RMSD в Grid search Lasso и Ridge regression





	LassoCV	EnCV	L=0	Ridge M	Ridge C	LassoCVGS
$\bar{\lambda}$	2.39	2.17	6.79	1.71	1.53	1.74

Сравнение корреляции RMSD для всех методов)



Результаты

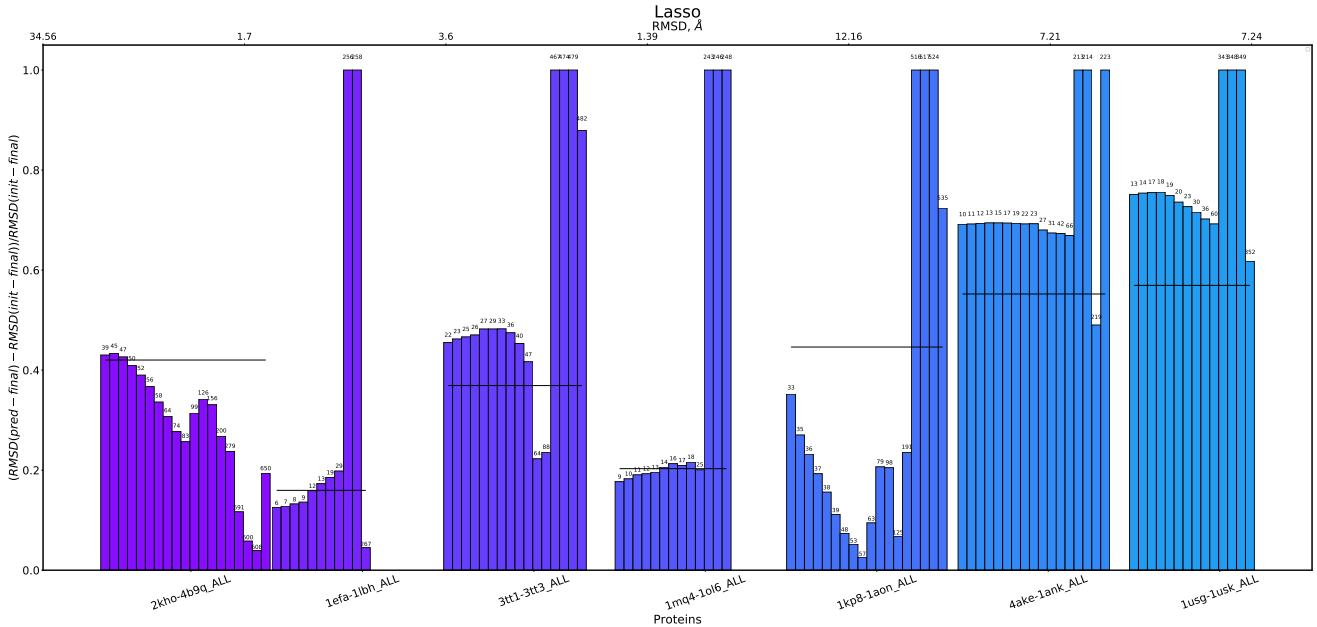
По сравнению с современной моделью (LASOO с поиском по сетке и kfold = 10 показал неплохие результаты).

$$RidgeM - RMSD = 1,71A$$

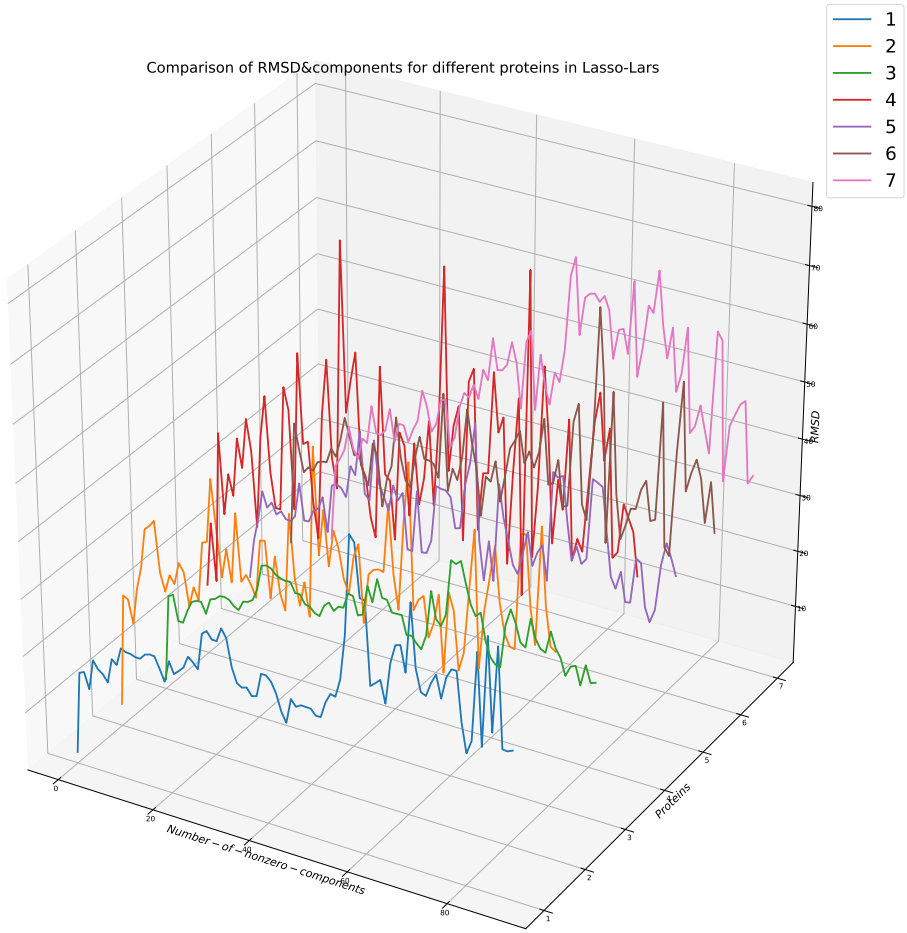
$$LassoCVGS - RMSD = 1,74A$$

EnCV показал менее хороший результат - RMSD=2.17 A

Результаты


Lasso
RMSD, Å
34.56
1.7
3.6
1.39
12.16
7.21
7.24
1.0
0.8
0.6
0.4
0.2
0.0
(RMSD(pred − final) − RMSD(init − final))/RMSD(init − final)
2kho-4b9q_ALL
1efa-1lbh_ALL
3tt1-3tt3_ALL
1mq4-1ol6_ALL
1kp8-1aon_ALL
4ake-1ank_ALL
1usg-1usk_ALL
Proteins

Результаты


Comparison of RMSD&components for different proteins in Lasso-Lars
1
2
3
4
5
6
7
RMSD
Proteins
Number − of − nonzero − components

Выводы

- Получилось приблизиться к результатам L_2 регрессии при помощи L_1 методов
- При этом LASSO и LARS более оптимальны в пространстве разреженных торсионных углов
- Полученные результаты для LARS говорят о том, что наилучший вклад в предсказание дает модель с небольшим количеством компонент(1-20)