Исследование конформационных изменений белков с использованием коллективных движений в пространстве торсионных углов и регуляризации L1

Moscow Institute of Physics and Technology

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

21 апреля 2019 г.

Цель работы

Исследуются

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

Проблемы

Нет научных публикаций исследующих приложение L_1 регуляризации к данной проблеме. L_1 регрессия работает быстрее чем методы L_2 за счет того пространство торсионных углов разреженно. Она также позволяет выбрать произволное количество углов, что снижает размерность.

Методы

Канонический: Ridge regression
Исследуемые: LASSO, Elastic-net, LARS

Литература

Результаты в области

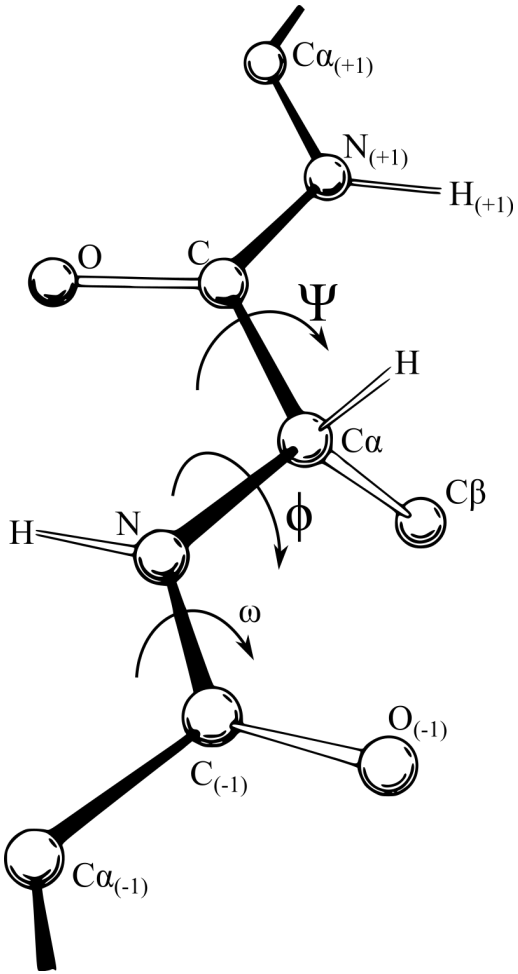
R. Mendez and U. Bastolla, Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins.
A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model
F. Tama 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.

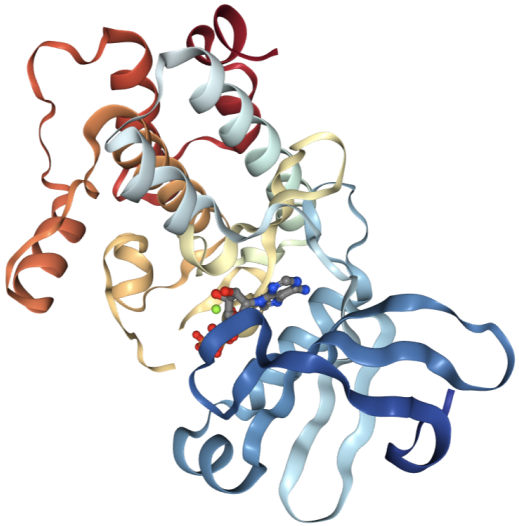
Исследуемые методы

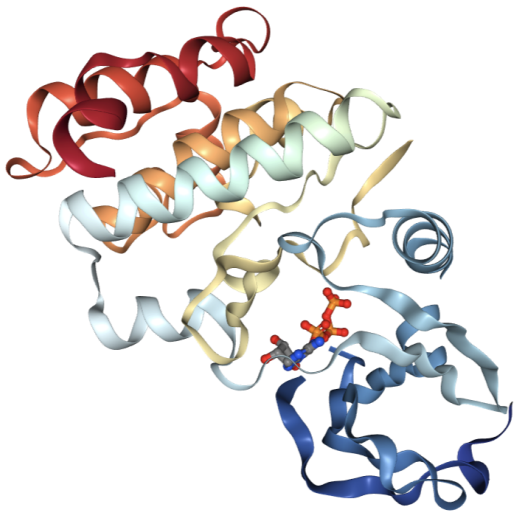
R. Tibshirani, Regression shrinkage and selection via the lasso
H. Zou and T. Hastie, Addendum: Regularization and variable selection via the elastic net
B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression

Конформационные изменения и нормальные моды в пространстве торсионных углов

Рис.: Структура белка и 1MQ4-10L6 конформационное изменение







Постановка задачи

Формулировка

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

Δr - изменения декартовых координат

$\Delta\phi$ - изменения углов скручивания(торсионные углы)

Регрессия с LASSO

$$min\ \frac{1}{2n}\|\Delta r - J\Delta\phi\|_2^2 + \alpha\|\Delta\phi\|_1$$

Входные данные

Датасет

Для этой мы используем 30 пар белков из RCSB Protein Data Bank и для каждой пары опираемся на полученные RMSD(initial,final) и RMSD(initial, predicted).

Обзор методов регуляризации для приближения к Ridge regression

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

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

Обзор методов регуляризации использованных для получения фиксированного количества компонент

LARS

$$\|\Delta r - J\Delta\phi\|_2^2 + \alpha\|\Delta\phi\|_1$$

$$\|\Delta r - J\Delta\phi\|_2^2 + \alpha s^\top \Delta\phi$$

$$s_j = 0, \phi_j = 0$$

$$s_j = 1, \phi_j > 0$$

$$s_j = -1, \phi_j < 0$$

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

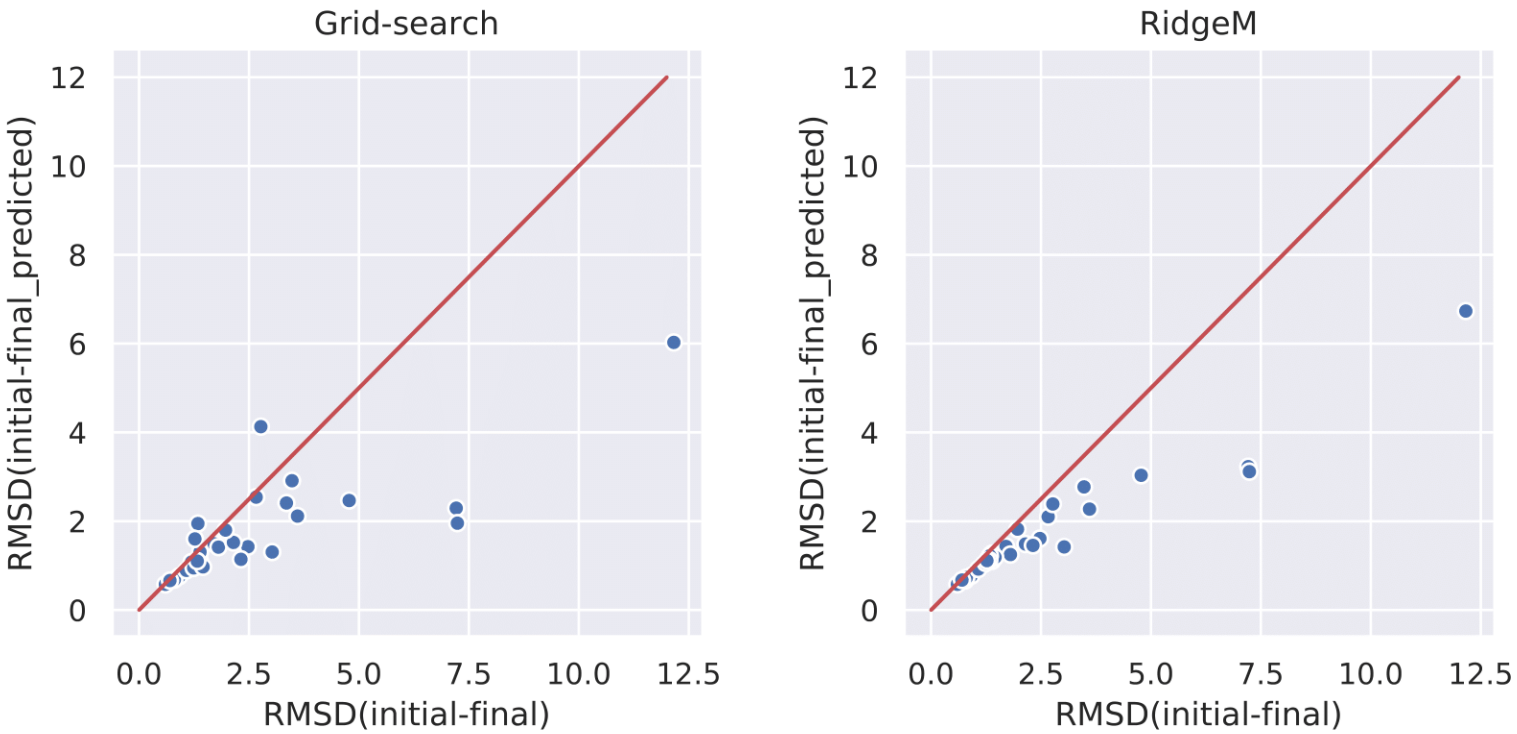
Исследованные методы

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.

Результаты

Рис.: Корреляция RMSD в Grid search Lasso и Ridge regression



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

Результаты

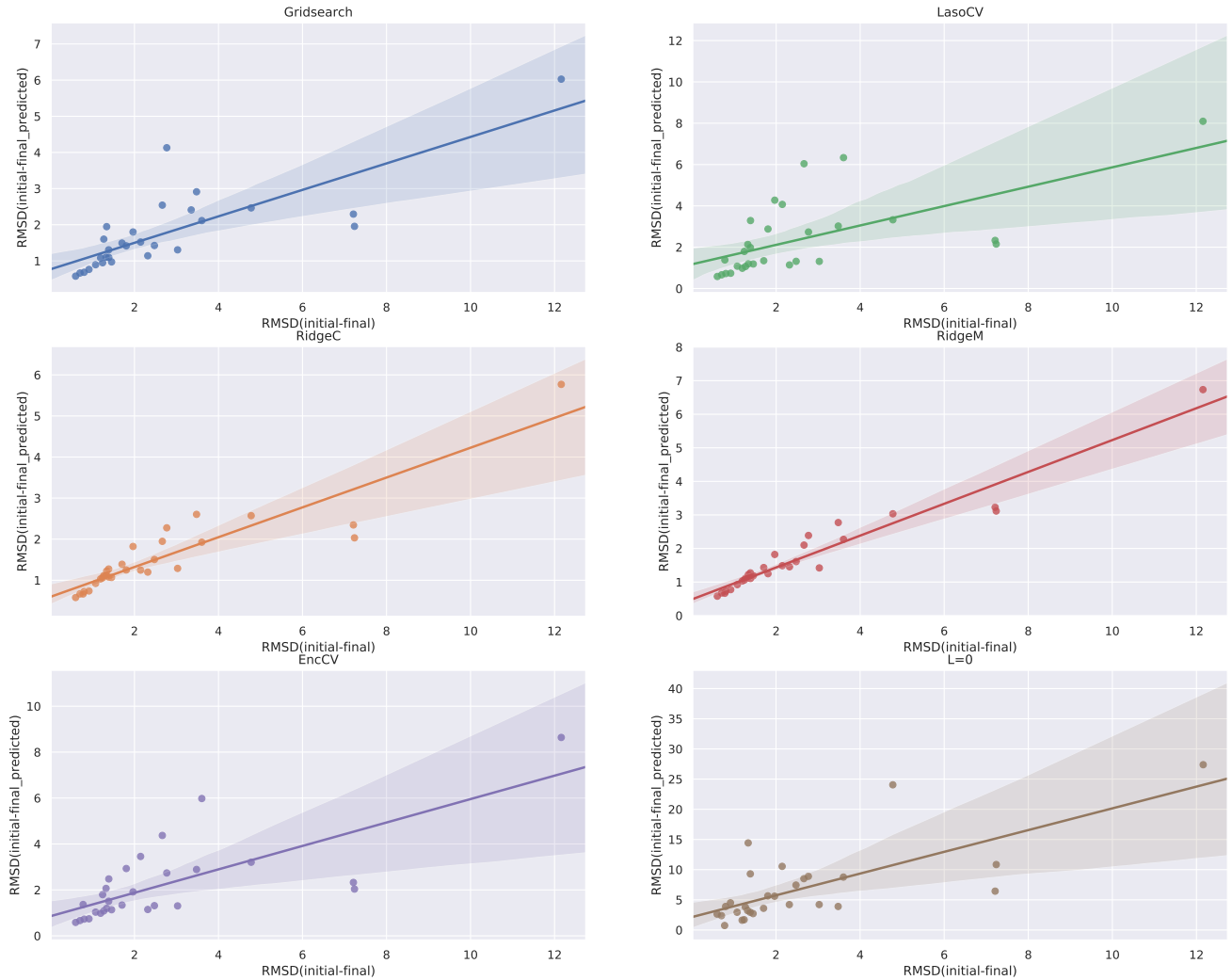


Рис.: Сравнение корреляции RMSD для всех методов(линейная апроксимация)

Результаты

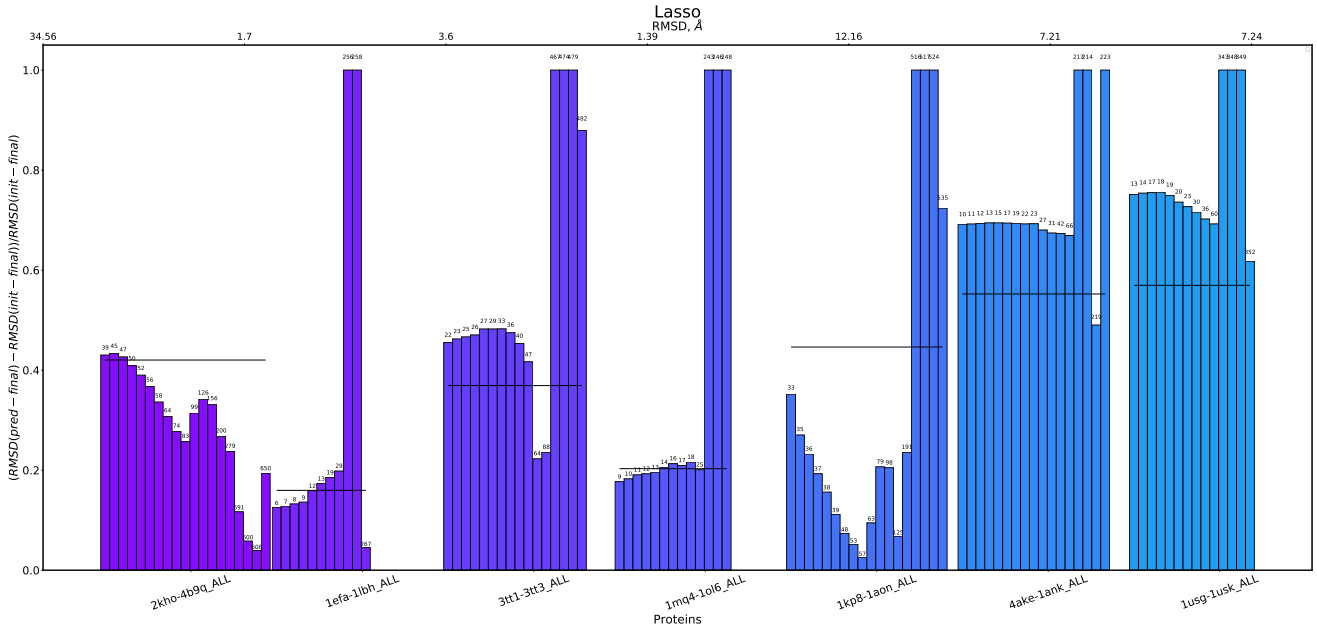
По сравнению с современной моделью (LASOO с поиском по сетке и kfold = 10 показал неплохие результаты).

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

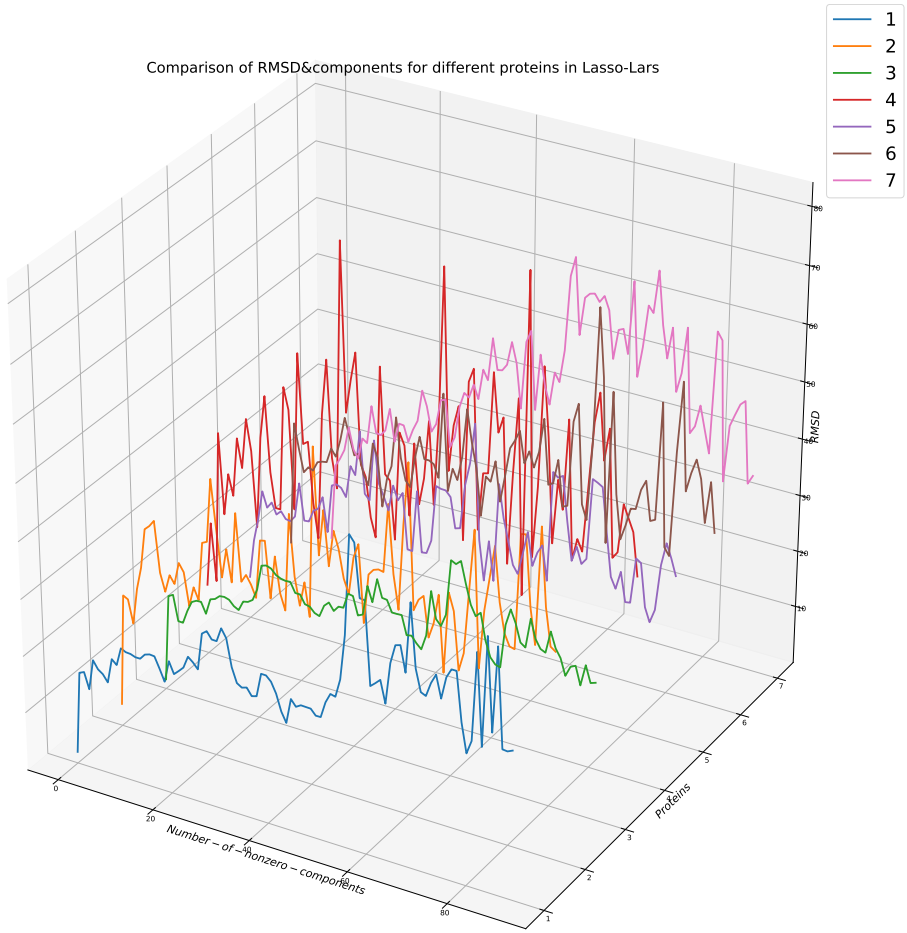
$$LassoCVGS - RMSD = 1,74A$$

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

- Показано, что возможно 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)