Исследование конформационных изменений белков при помощи L_1 регуляризации

Емцев Даниил, Рябинина Раиса

Московский физико-технический институт

Курс: Автоматизация научных исследований в машинном обучении (практика, В. В. Стрижов)/ весна 2019

Цель работы

Конформационные изменения белков

Исследовать методы L_1 регуляризации функции разницы наблюдаемых изменений, 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 на переход в новое положение атомов.

Методы регуляризации

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

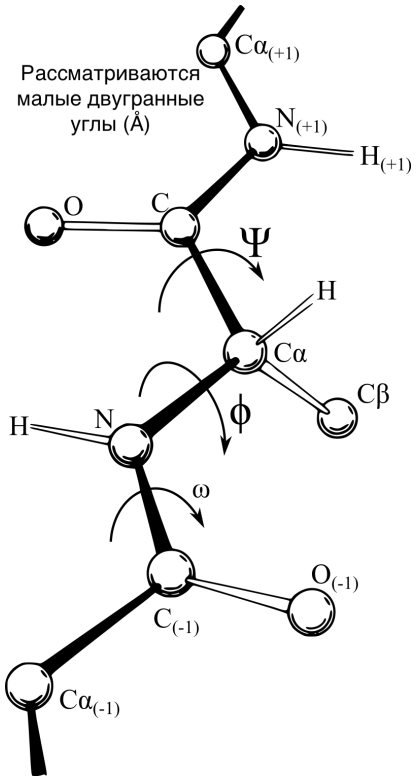
Критерий качества

Для оценки расстояния между белками используется RMSD.

Предмет исследования: конформационные изменения

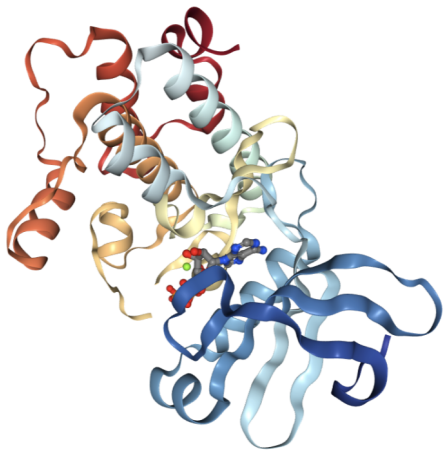
Пространство торсионных углов

Изменения в пространстве торсионных углов предсказанные при помощи методов L_1 и L_2 регрессии наиболее точно совпадают с заданными конформациями.

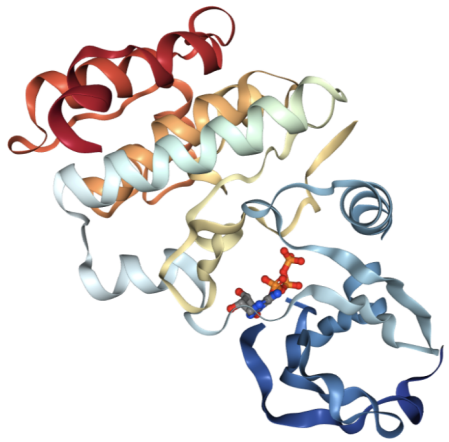


'Скелет' белка

Белок 1MQ4



Белок 10L6



Изменение белков с одинаковой цепочкой атомов

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

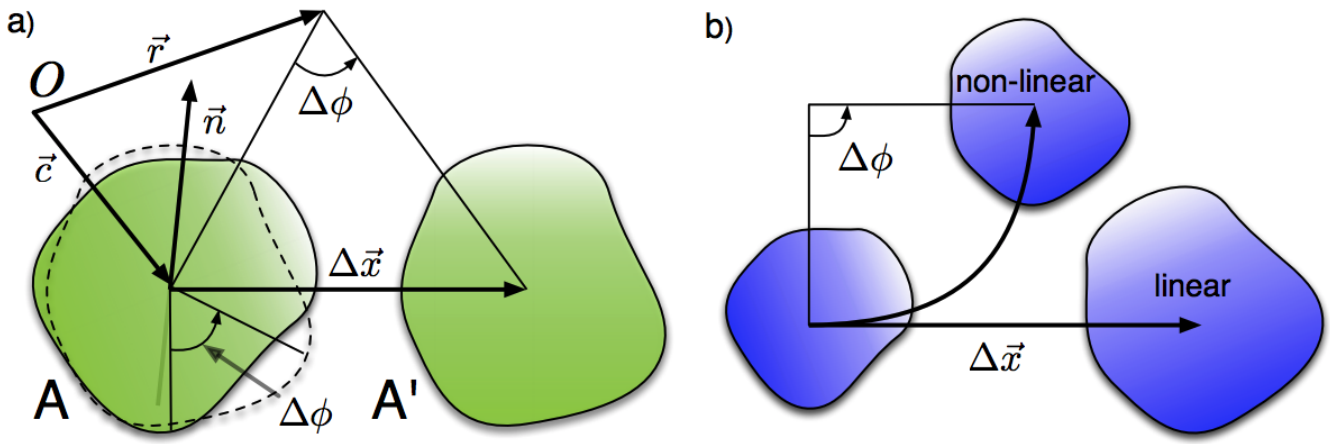
Экстраполяция мгновенного движения

Поступательное приращение в жесткого блока $\Delta\mathbf{x}$ и угловое приращение $\Delta\phi$

$$\Delta\vec{x} = a\vec{v}$$

$$\vec{n} = \frac{\vec{\omega}}{\|\|\vec{\omega}\|_2}$$

$$\Delta\phi = a\|\vec{\omega}\|_2$$



https://hal.inria.fr/hal-01505843/file/NOLB-Author-Version.pdf

Литература

Модели белков

- R. Mendez and U. Bastolla, Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins, 2010.
- A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model, 2001.

Конформационные изменения

- F. Tama and Y. H. Sanejouand, Conformational change of proteins arising from normal mode calculations, 2001.
- H. G. Dos Santos, J. Klett, R. Mendez, and U. Bastolla, Characterizing conformation changes in proteins through the torsional elastic response, 2013.
- I. Bahar and A. Rader, Coarse-grained normal mode analysis in structural biology, 2005.
- L. Meireles, M. Gur, A. Bakan, and I. Bahar, Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins, 2016.

О преймуществах LASSO

- S. Katayama and S. Imori, Lasso penalized model selection criteria for high-dimensional multivariate linear regression analysis, 2014.

Постановка задачи

Рассматриваются две конформации одного белка

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

A,B - различные конформации, $\Delta r = r^A - r^B$, $r_i^{A,B} \in \mathbb{R}^3$, i - номер атома, $\Delta r \in \mathbb{R}^{3n}$ - декартовы координаты, $\Delta\phi \in \mathbb{R}^n$ - торсионные углы, $J \in \mathbb{R}^{3n}x\mathbb{R}^n$ - матрица Якоби $\frac{\Delta r_{int_i}}{\Delta\phi_a}$, $M \in \mathbb{R}^{3n}x\mathbb{R}^{3n}$ - диагональная матрица весов.

Метод наименьших квадратов

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

L_1 регуляризация

- Матрица Якоби содержит коррелированные переменные, что приводит к переобучению.
- Методы L_1 регуляризации выбирают наболее важные признаки.

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

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

При расчетах использовались cross validation и grid search методы

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

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

Вычислительный эксперимент

Датасет

- 26 белков, каждый из которых представлен в двух конформациях
- Набор включает широкий спектр макромолекулярных движений
- Все структуры не имеют разорванных цепей и пропущенных атомов
- Данные загружены из RCSB Protein Data Bank

Среднеквадратичное отклонение позиций атомов

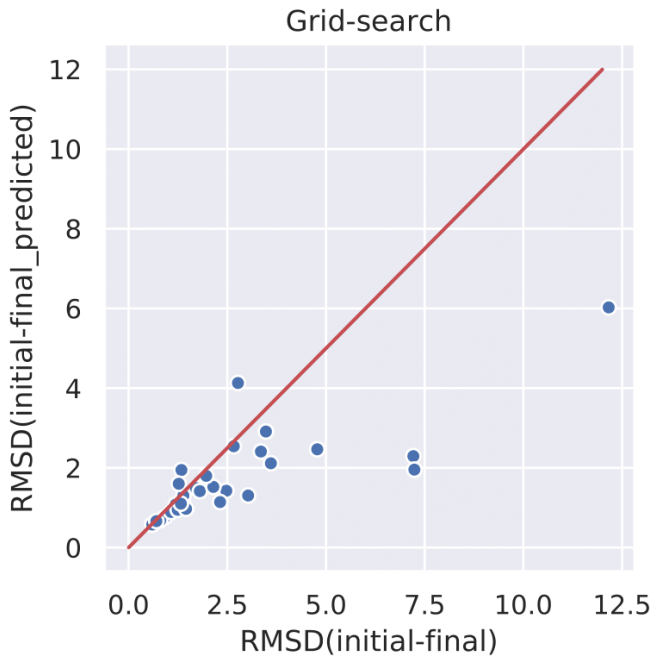
В работе сравниваются $RMSD_{(initial,final)}$ и $RMSD_{(initial,predicted)}$

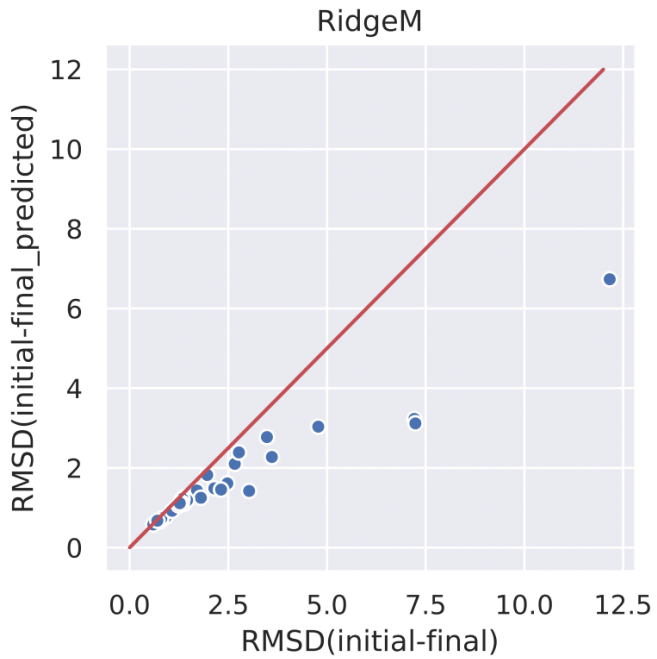
$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

Корреляция RMSD в Grid search Lasso и Ridge regression

Точность метода Grid search Lasso близка к Ridge regression.

На графиках можно увидеть насколько хорошо коррелируют предсказанные и заданные конформации. Прямая является биссектрисой.



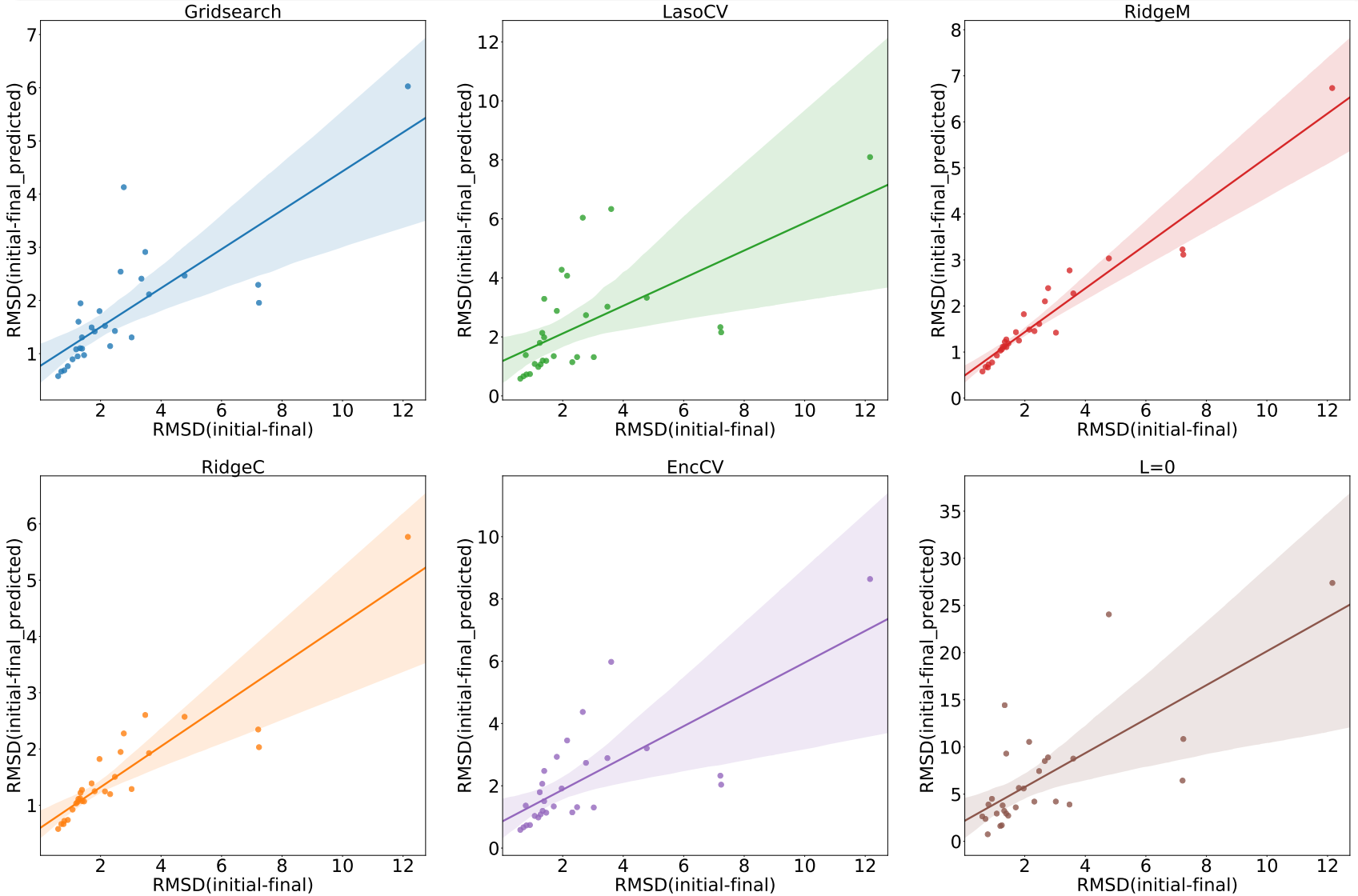


	LassoCV	EnCV	L=0	Ridge M	Ridge C	LassoCVGS
Å	2.39	2.17	6.79	1.71	1.53	1.74

Сравнение корреляции RMSD для всех методов

Линейное приближение сравниваемых RMSD

Область вокруг соответствует наиболее возможному значению коэффициента наклона.



Сравнения результатов для LASSO и Elastic-net с Ridge

Метод LASSO с grid search в среднем догоняет Ridge regression

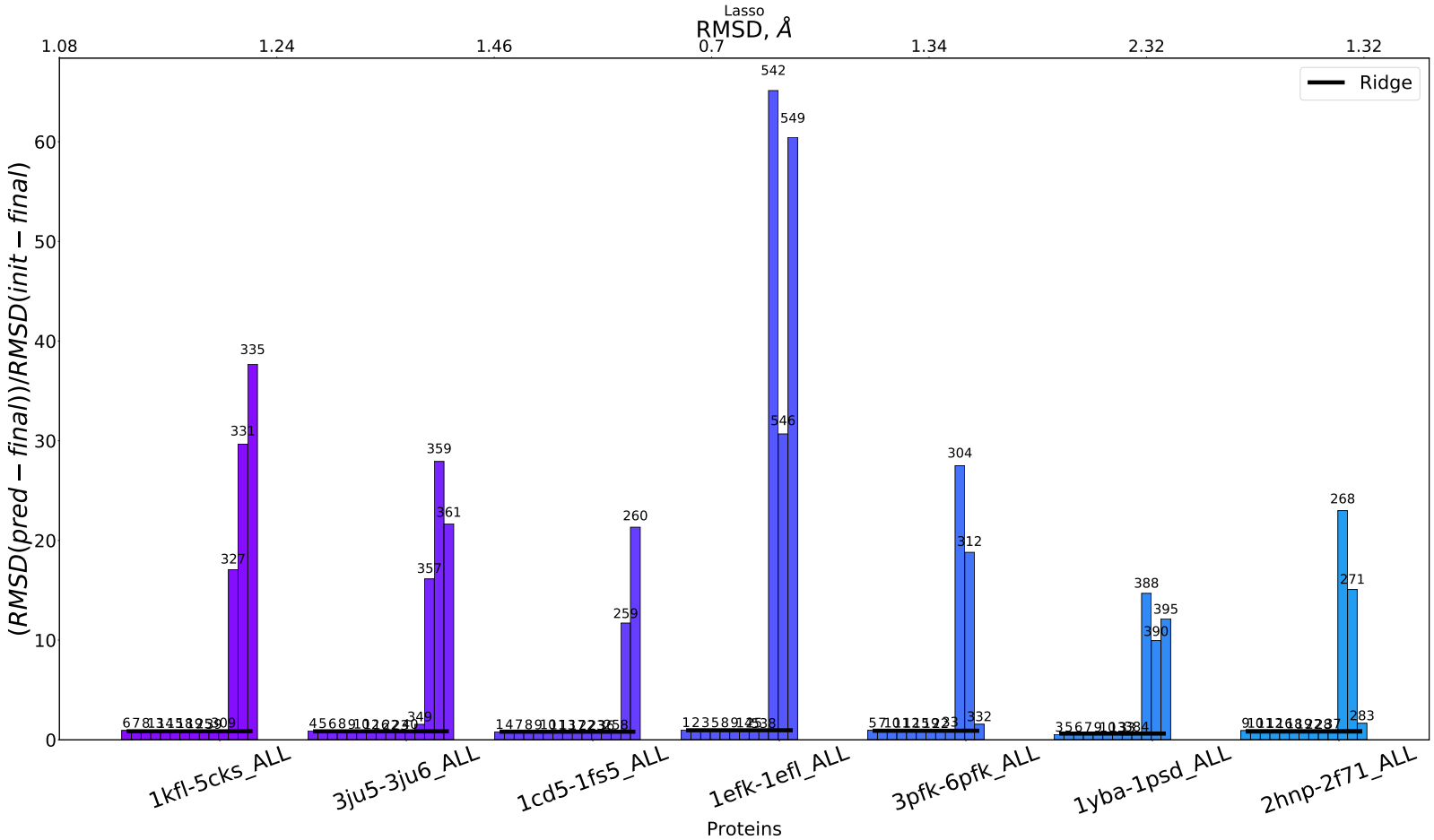
$$RMSD_{RidgeM} = 1,71\mathring{A} \quad RMSD_{LassoCVGS} = 1,74\mathring{A}$$

Elastic-net regression показал менее хороший результат

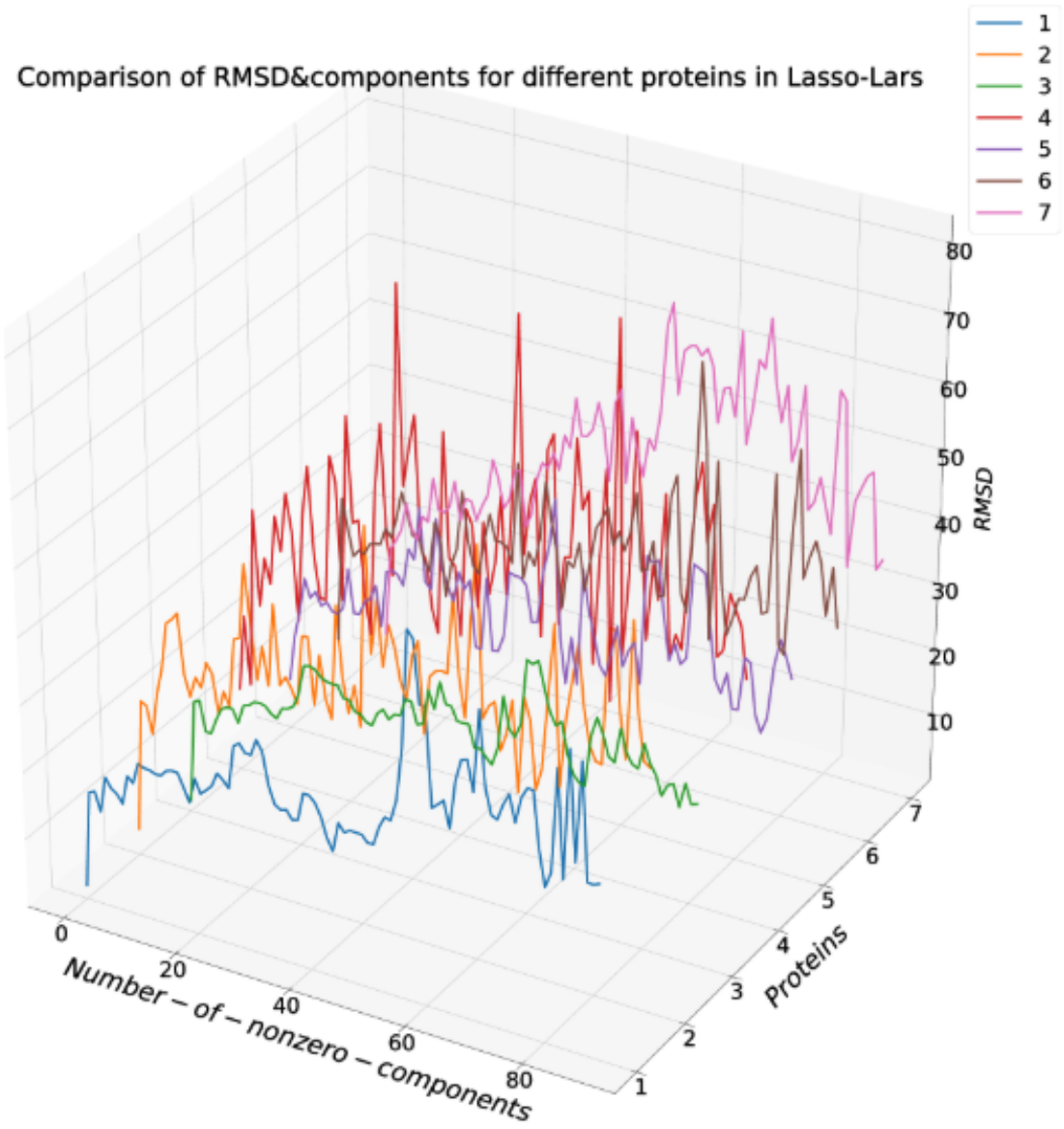
$$RMSD_{EnCV} = 2,17\mathring{A}$$

RMSD получено для фиксированного числа компонент

На графике представлены результаты для 7-ми предсказанных белков в сравнении с изначальным RMSD. Можно видеть как количество ненулевых компонент влияет на качество модели.



LARS и модели с небольшим количеством компонент



Выводы

L_1 методы показали близкие к L_2 результаты

Среднее *RMSD* по всем белкам из тестового набора для Lasso с использованием grid search близко к Rigde regression.

LASSO и LARS более оптимальны

За счет того, что пространство торсионных углов разрежено, отбор признаков происходит быстрее.

Метод LARS производит отбор небольшого колличества признаков

Из результатов видно, что наилучший вклад в предсказание другой конформации дают модели с не более чем 20-ю ненулевыми компонентами.