LIME: Estimating the interaction depth z

Best efforts with Linear Regression

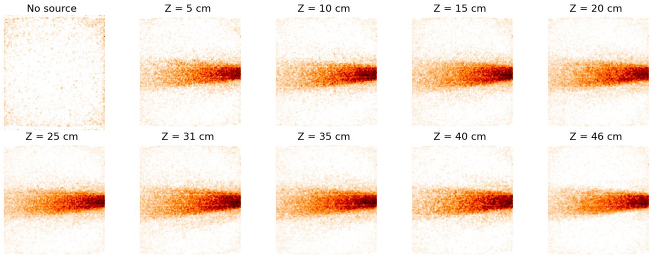
CYGNO Experiment

R. Roque| CYGNO Reconstruction & Analysis Meeting | 20/10/2022

An Update of the Last Efforts

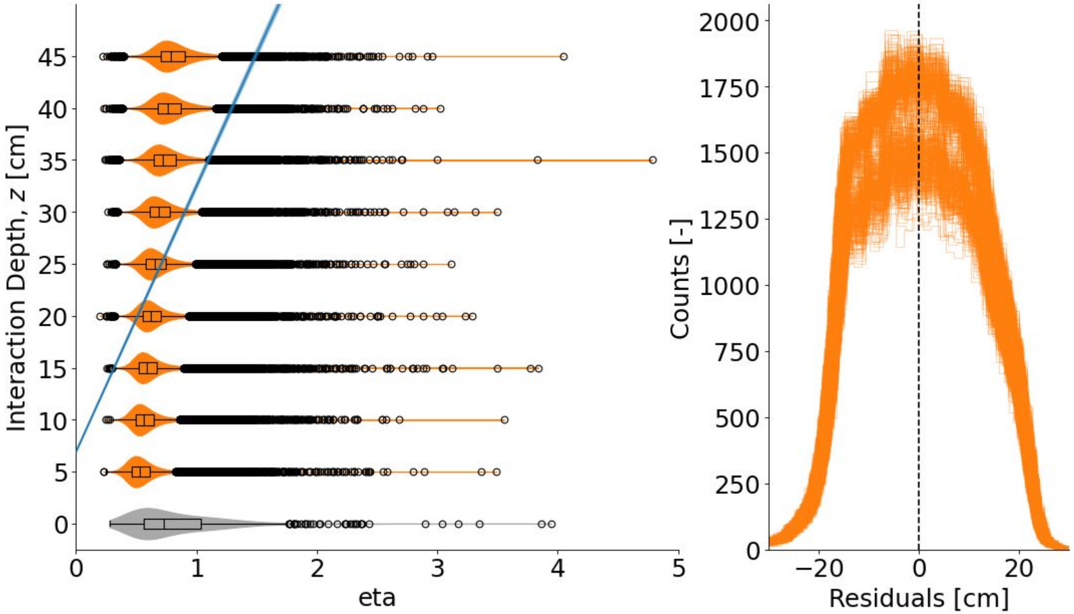
Data Information

- Runs 5861 -> 5911 taken on 04/11
- Water cooled, dark lab, He-40%CF_4
- Scan in z with ^{55}Fe source



I am working with 19.6% of the original dataset (background clusters were discarded).

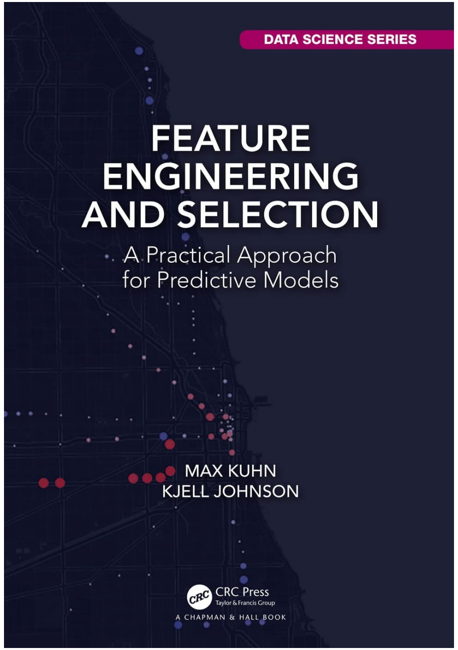
Linear Regression with the Transverse Profile, η



	1st order	2nd order	3rd order	4th order
r^2	0.0170(16)	0.246(32)	0.271(33)	0.278(15)
RMSE [cm]	11.25(14)	10.73(24)	10.54(25)	10.49(13)

A New Strategy based on Feature Engineering

Feature Removal → Feature Correlation → Feature Interactions → Feature Selection → Linear Models



scikit learn

Pandas

https://github.com/RitaROK/Analysis/blob/main/Estimation_of_z.ipynb

A New Strategy based on Feature Engineering

Feature Removal | Feature Correlation | Feature Interactions | Feature Selection | Linear Models

To guarantee a model valid for other energies, **the energy-dependent features were discarded**:

- sc_integral
- sc_corrintegral
- sc_tgaussamp
- sc_size
- sc_nhits
- sc_length
- sc_width

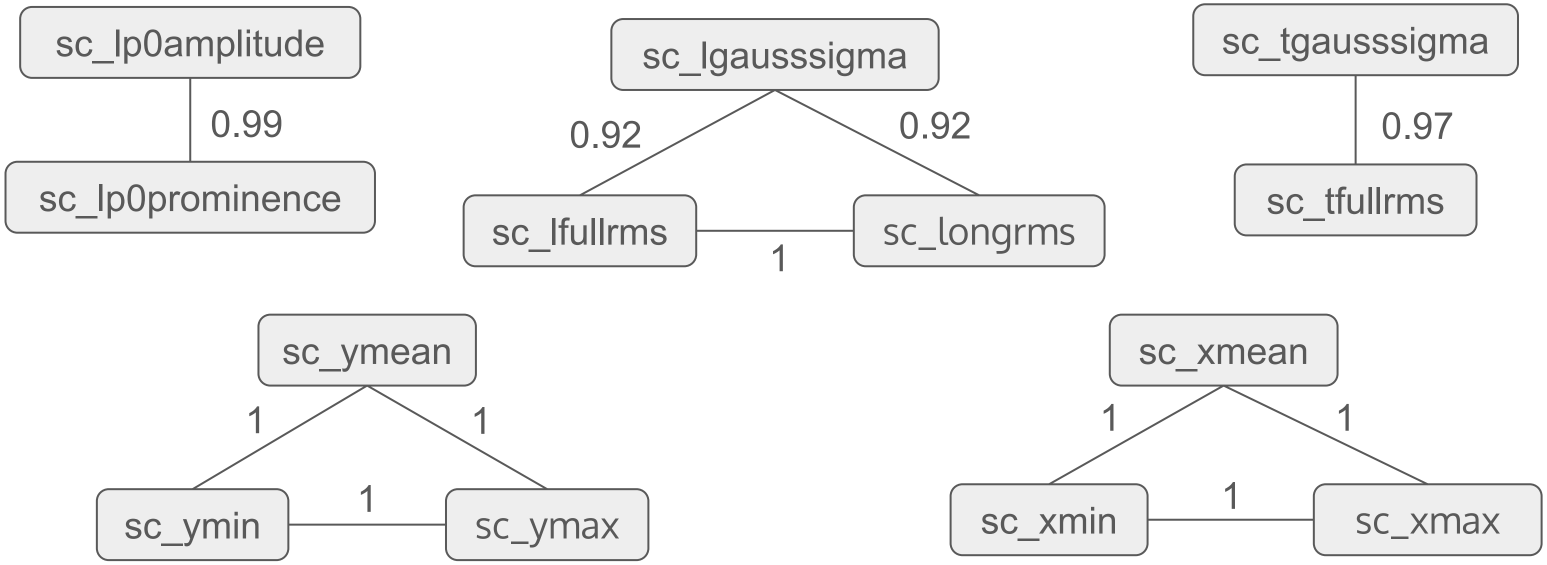
I also **discarded quasi-constant features** from the dataset:

- sc_energy
- sc_pathlength
- sc_lstatus
- slimness
- sc_pearson
- sc_tstatus

A New Strategy based on Feature Engineering

Feature Removal → **Feature Correlation** → Feature Interactions → Feature Selection → Linear Models

Groups of features with redundant information ($r^2 > 0.9$).



A New Strategy based on Feature Engineering

Feature Removal → Feature Correlation → **Feature Interactions** → Feature Selection → Linear Models

New features were created by multiplying and dividing all the original features. The following interactions show a promising correlation with z:

- sc_lgausssigma*sc_tfullrms
- sc_lfullrms*sc_tfullrms
- sc_longrms*sc_tfullrms
- sc_lgausssigma*sc_tgausssigma
- sc_tgausssigma*sc_lfullrms
- sc_tgausssigma*sc_longrms
- sc_tfullrms/sc_rms

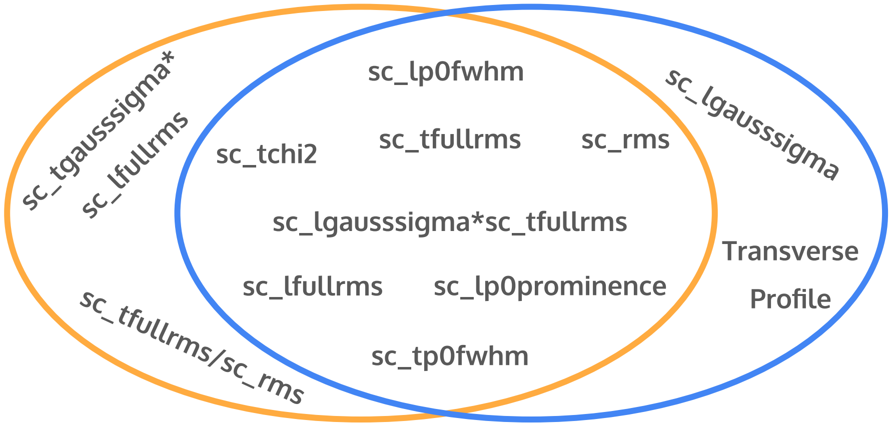
Almost all of them are a combination of the transverse and longitudinal profile of the clusters.

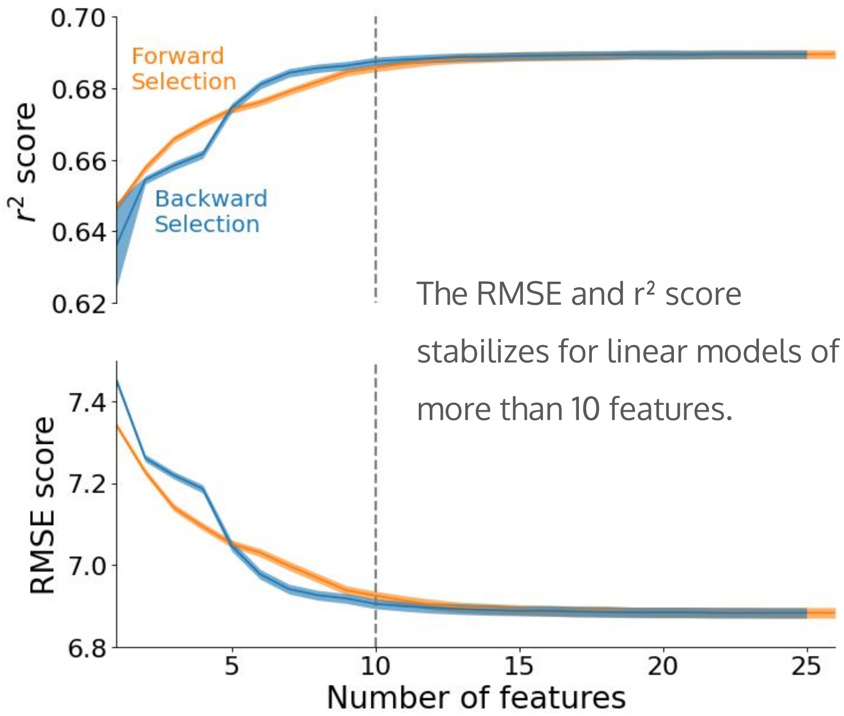
A New Strategy based on Feature Engineering

Feature Removal | Feature Correlation | Feature Interactions | **Feature Selection** | Linear Models

Using the original features and relevant interactions, **forward** and **backward** feature selection were applied.

The best 10 features for multilinear regression are:





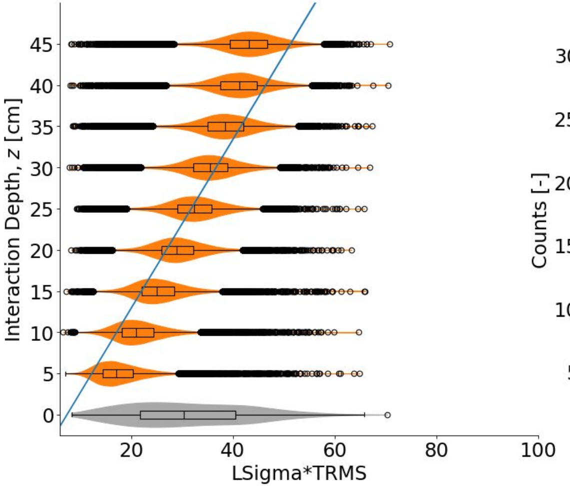
The RMSE and r^2 score stabilizes for linear models of more than 10 features.

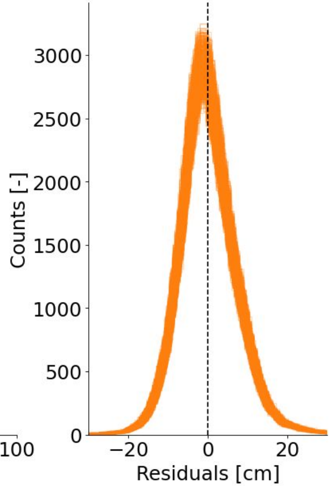
A New Strategy based on Feature Engineering

Feature Removal → Feature Correlation → Feature Interactions → Feature Selection → **Linear Models**

The Best Linear Regression

z = -7.52(5) + 1.0233(15) * sc_lgausssigma*sc_tfullrms



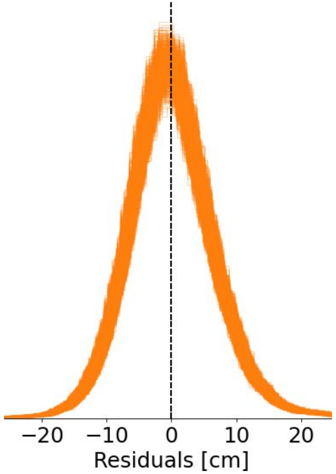


r^2=0.647(12), RMSE = 7.34(13) cm

The Best Multilinear Regression

Regression coefficients:

- Intercept: 18.4(8)
- sc_lgausssigma: -2.26(10)
- Transverse profile: 1.27(23)
- sc_lp0fwhm: 1.557(21)
- sc_tfullrms: -2.00(12)
- sc_tchi2: 0.1130(24)
- sc_lgausssigma*sc_tfullrms: 0.896(17)
- sc_tp0fwhm: 0.752(25)
- sc_rms: -1.826(29)
- sc_lfullrms: 0.52(5)
- sc_lp0prominence: 0.00639(13)



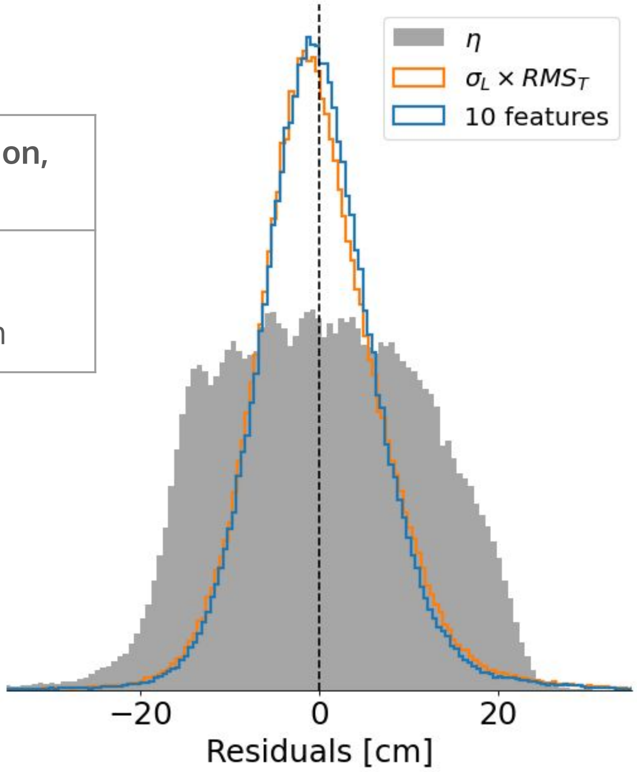
r^2=0.686(12), RMSE = 6.92(13) cm

Conclusions

- With a well developed linear model, the model accuracy is significantly improved

  The RMSE can be improved from 11.25(14) cm to 6.92(13) cm.

Linear Regression, Transverse profile	Linear Regression, TSigma*LRMS	Multi-Linear Regression, 10 features
r^2=0.0170(16) RMSE = 11.25(14) cm	r^2=0.647(12) RMSE = 7.34(13) cm	r^2=0.686(12) RMSE = 6.92(13) cm



Next steps

- Explore other Non-Linear Regression Models.

  BDTs, NN and Random Forests

- Consider cluster variables that are not in the trees

  Flaminia suggested using kurtosis/integral

The speaker acknowledges the FCT PhD studentship (ref. SFRH/BD/143355/2019).