ECEN-662 Final Project

Title: Estimator Design of the Stacking-Fault Energy

Group members: Guanda Li, Shaowen Zhu

Abstract

Stacking-Fault Energy (SFE) is a materials property of alloys. Its value is dependent on the content of several metals in the alloy. In this project, we have got a SFE material data set as the training data set. It consists of 473 samples, and for every sample, we have its SFE and the content of 17 kinds of metals. However, we were not sure which metals can influence the SFE, as some of them may be independent with SFE or have little influence on it. In some documents, we learnt that the SFE has a linear relationship with related metals, that is, we can use a linear formula to calculate it according to the content of several metals. As a result, we can use the methods of linear regression or designing a best linear unbiased estimator (BLUE) to select the features and construct the formula. In this report, we firstly compared the results of choosing different features and then selected the features. We both used the filter approaches (independent of the classification rule, such as deleting the features with too many zeros, F-regression and Mutual-info-regression) and wrapper approaches (not independent of the classification rule, such as exhaustive search). Secondly, as we had got the features, we could use linear regression, ridge regression or the formula of BLUE to construct the estimator. Finally, we used R^2 statistic and plotted the graphs to test the estimator.

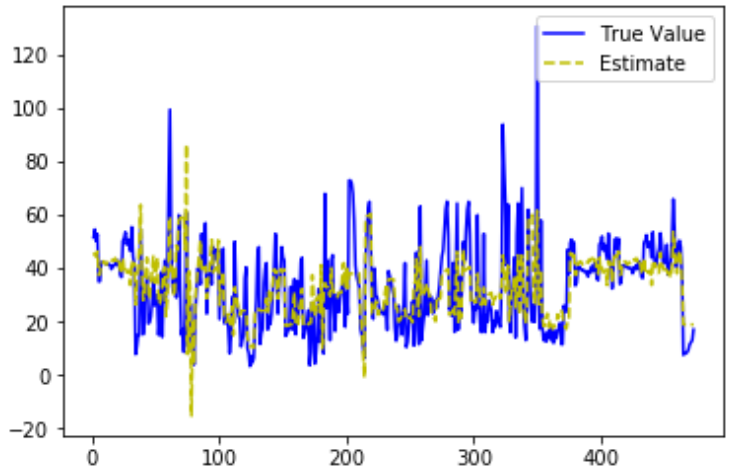
Feature Selection

In classification and estimation, the peaking phenomenon implies that the true error is likely to increase with more features. Also, reduce the number of features can decrease the computational load. Therefore, it's important to do feature selection before analysis of the data. In this project, both filter and wrapper method were used for feature selection. The criterion of feature selection is to minimize the loss of information. Thus, features that contain most information about SFE were selected. The idea and results are shown below.

Firstly, we used all the features and linear regression to construct the formula. We got the coefficient of the content of every kind of metals and estimated of mean-square error of the estimator we got. Table 1-a shows the coefficients and Graph 1-a shows the True SFE and our estimates of all the samples. The mean-square error is 142.6777.

Table 1-a Coefficients of the content of different metals (using all features)

C	N	P	S	V	Ni	Nb	Al	Ti
-3.1614	2.8377	308.87	-65.642	-7.0194	0.9906	-13.107	4.2971	-76.224
Fe	Hf	Mo	Mn	Co	Si	Cr	Cu	Intercept
-0.1111	-19.117	1.3741	0.1999	-0.7688	-2.8179	0.0021	3.7783	142.68



Graph 1-a True SFE and estimate of all the samples (using all features)

From the table and the graph we can see, the coefficients had a great span and the estimates seemed unbiased and tracked the true values. However the mean-square-error was large and the true value seemed like the estimate combined with an obvious noise. Therefore, there must be some features independent of SFE and playing a role as the noise in the estimator. As a result, we needed to delete some features.

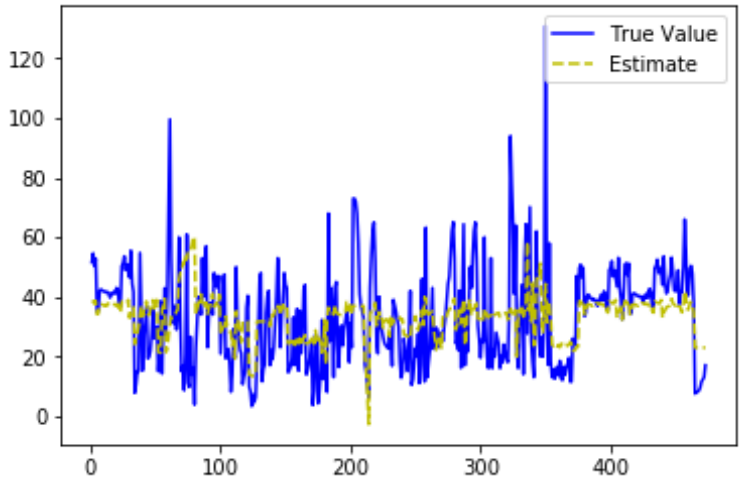
Filtering Approaches

Deleting the features and samples (Pre-processing)

From the training data set, we could see that there were some features of which the values are zero of the majority of samples. These features had less influence on the value of SFE and could be noises in regression, So we could delete the features of which the values are zero of more than 55% samples. After that, we retained these features: C, N, Ni, Fe, Mn, Si, Cr. Still. We used these features for linear regression. Table 1-b shows the coefficients of retained features and Graph 1-b shows the True SFE and the new estimates of all the samples.

Table 1-b Coefficients of the content of different metals (deleting features with too many zeros)

C	N	Ni	Fe
-6.8348	6.9851	-0.8004	-1.1482
Mn	Si	Cr	Intercept
-1.2425	-3.9415	-1.0633	143.58

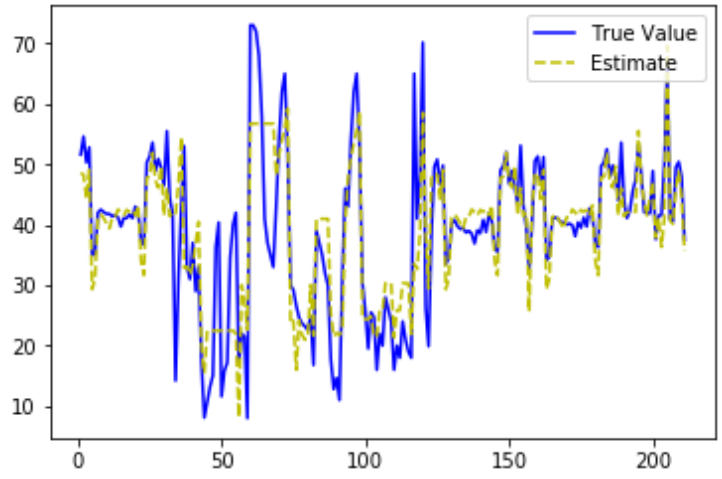


Graph 1-b True SFE and estimate of all the samples (deleting features with too many zeros)

The mean-square error was 215.9770, which was even larger than that of using all the features. And from the graph we could see that the estimates still tracked the true values, but it seemed smoother, so the mean-square-error could be larger. However, we could also see that there were still some features valued zero of many samples, and then these features would not contribute to the estimate of SFE, so these samples could also be noisy. As a result, we could also delete these samples. Table 1-c shows the coefficients of each feature and Graph 1-c shows the True SFE and the new estimates of retained samples.

Table 1-c Coefficients of the features (deleting samples with features valued zero)

C	N	Ni	Fe
-25.896	19.050	-1.8353	-4.0630
Mn	Si	Cr	Intercept
-3.0461	-7.4480	-4.0763	410.13



Graph 1-c True SFE and estimate (deleting samples with features valued zero)

The sample size reduced to 211, but the mean-square error also reduced to 53.2571. Besides, from Graph 1-c we could also see that the estimates tracked the true values much better.

F-regression

F-regression is based on F-value, which measures the linear relationship between each feature and SFE. This is done by calculating the correlations between each regressor and the target, that is:

$$\frac{(x(i)-\overline{x})*(y(i)-\overline{y})}{\sqrt{Var(x)*Var(y)}}$$

then, the result is converted to an F value then to a p-value. According to F value, top five features (Ni, Fe, N, Si, Mn) were selected.

Table 1-d F-value of the features

Element	F-value
C	1.4619
N	7.5712
Ni	207.3065
Fe	142.4360
Mn	1.7767
Si	1.9039
Cr	1.5271

Mutual-info-regression

Mutual information measures the dependency between the variables. This value is equal to zero if and only if the feature and SFE are independent, and higher values mean higher dependency between the feature and target. More specifically, it quantifies the amount of information obtained about SFE, through one of the feature. According to mutual information, top five features (Fe, Ni, Cr, Si, Mn) were selected.

Table 1-e Mutual information of the features

Element	Mutual information
C	0.245242
N	0.204064
Ni	0.784569
Fe	0.806204

Table 1-e (continued)

Element	Mutual information
Mn	0.250247
Si	0.331770
Cr	0.552505

Wrapper Approaches

From pre-processing we had already selected 7 features through filter approaches. Besides we could use wrapper approaches to find whether we could delete more features without loss of much information. We used R^2 statistic as the criteria, where

$$R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS},$$

$$\text{TSS} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2,$$

$$\text{RSS} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2,$$

and the estimator worked better if the R^2 statistic is larger. We used exhaustive searching method, that was, when we were about to select n features, we compared the criteria of all the combinations of n features in the retained features, and then choose the combination whose criteria was the largest. Table 1-f shows the features we selected when we selected 1~7 features from the retained 7 features and their related R^2 statistic and mean-square-error.

Table 1-f Exhaustive searching result

	Select 1 feature	Select 2 features	Select 3 features	Select 4 features
Selected features	Ni	N, Ni	Fe, Mn, Cr	C, Fe, Mn, Cr
R^2 statistic	0.4980	0.5535	0.5945	0.6322
MSE	84.6691	75.3009	68.3904	62.0328

	Select 5 features	Select 6 features	Select 7 features
Selected features	C, Fe, Mn, Si, Cr	C, Ni, Fe, Mn, Si, Cr	C, N, Ni, Fe, Mn, Si, Cr
R^2 statistic	0.6621	0.6751	0.6842
MSE	56.9792	54.7986	53.2571

From the table we could see that the R2 statistic increased as the number of features we selected increased, and the MSE decreased as the same time. However, when we selected 5 features, the MSE did not change much and neither did the R2 statistic. As a result, we could select the 5 features, C, Fe, Mn, Si and Cr to make the problem simpler without losing much information.

From above, we had selected three sets of 5 features by three different ways. The three sets are { Ni, Fe, N, Si, Mn }, { Fe, Ni, Cr, Si, Mn } and {C, Fe, Mn, Si, Cr}.

Formula Construction

Linear Regression

The model used in linear regression is:

$$Y = \omega_0 + \omega_1 X_1 + \cdots + \omega_d X_d + \varepsilon$$

where $X_1, X_2, \cdots, X_d$ are the predictors, ε is noise term, ω_0 and $\boldsymbol{\omega} = (\omega_1, \omega_2 \cdots, \omega_p)$ are the parameters. The equation can also be written as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \boldsymbol{\omega_0} + \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \boldsymbol{\omega} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{\omega}_0 + \boldsymbol{H\omega} + \boldsymbol{\varepsilon}$$

The key point is that the model be linear in the parameters. Linear regression fits a linear model with coefficients $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_p)$ to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation. Mathematically it solves a least squares problem of the form:

$$\min_{\boldsymbol{\omega}} \left(\hat{Y} - Y\right)^2$$

Without intercept($\omega_0 = 0$), the solution is found to be:

$$\hat{\theta} = (H^T C^{-1} H)^{-1} H^T C^{-1} Y$$

With intercept($\omega_0 \neq 0$), we can define $X_0 = (1,1,...,1)^T$ as a predictor, and then treat it like the condition above.

However, coefficient estimates for Ordinary Least Squares rely on the independence of the model terms. When terms are correlated and the columns of the design matrix X have an approximate linear dependence, the design matrix becomes close to singular and as a result, the least-squares estimate becomes highly sensitive to random errors in the observed response, producing a large variance. And ridge regression is used to solve the problem.

From the feature selection part we had ever selected three sets of 5 features, then we respectively used the three sets for linear regression and got three formulas of SFE. Table 2-a ~ Table 2-c show the coefficients we got for the three linear models.

Table 2-a Coefficients of linear model 1 (based on Ni, Fe, N, Si and Mn)

Ni	Fe	N	Si	Mn	Intercept
1.9542	-0.3921	27.7414	-5.9433	-0.3599	41.0295

Table 2-b Coefficients of linear model 2 (based on Fe, Ni, Cr, Si, Mn)

Fe	Ni	Cr	Si	Mn	Intercept
-4.3941	-2.2315	-4.2104	-7.6867	-3.0384	439.832

Table 2-c Coefficients of linear model 3 (based on C, Fe, Mn, Si, Cr)

C	Fe	Mn	Si	Cr	Intercept
-39.9924	-2.4441	-0.9586	-6.7142	-2.6265	252.0665

As a result, we got three linear estimator of SFE:

$\mathrm{SFE} = 27.742 * \mathrm{N} + 1.954 * \mathrm{Ni} - 0.392 * \mathrm{Fe} - 0.360 * \mathrm{Mn} - 5.943 * \mathrm{Si} + 41.030$;

$\mathrm{SFE} = -2.231 * \mathrm{Ni} - 4.394 * \mathrm{Fe} - 3.038 * \mathrm{Mn} - 7.687 * \mathrm{Si} - 4.210 * \mathrm{Cr} + 439.832$;

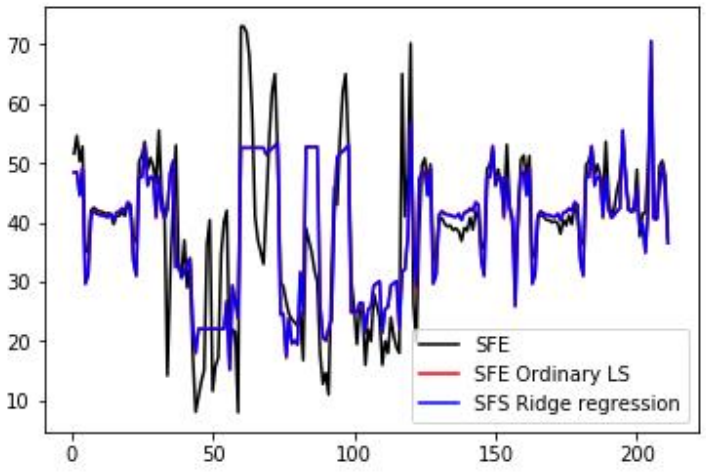
$\mathrm{SFE} = -39.992 * \mathrm{C} - 2.444 * \mathrm{Fe} - 0.959 * \mathrm{Mn} - 6.714 * \mathrm{Si} - 2.627 * \mathrm{Cr} + 252.066.$

Ridge Regression

Ridge regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of coefficients. The ridge coefficients minimize a penalized residual sum of squares：

$$\min_{\boldsymbol{\omega}} \left(\hat{Y} - Y\right)^2 + \alpha \left|\left|\omega\right|\right|_2^2$$

Here, $\alpha \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity.



Graph 2-a Comparison of Ridge Regression and Ordinary LS

Blue

Best linear unbiased estimator (BLUE) can be interpreted as linear regression in this project, as it is to minimize the variance of the estimator, and it has the formula

$$\hat{\theta} = (H^T C^{-1} H)^{-1} H^T C^{-1} Y.$$

And in this project, we assumed all the features are independent, so the covariance matrix C = σ^2I, so that

$$\hat{\theta} = (H^T H)^{-1} H^T Y.$$

However, as for the former part, we used the linear regression part in sklearn package, so we also used the BLUE theory to directly calculate the parameters, and we we got the parameters showed in Table 2-d ~ Table 2-f.

Table 2-d Coefficients of linear model 1 (BLUE based on Ni, Fe, N, Si and Mn)

Ni	Fe	N	Si	Mn	Intercept
1.9542	-0.3921	27.7414	-5.9433	-0.3599	41.0295

Table 2-e Coefficients of linear model 2 (BLUE based on Fe, Ni, Cr, Si and Mn)

Fe	Ni	Cr	Si	Mn	Intercept
-4.3941	-2.2315	-4.2104	-7.6867	-3.0384	439.832

Table 2-f Coefficients of linear model 3 (BLUE based on C, Fe, Mn, Si and Cr)

C	Fe	Mn	Si	Cr	Intercept
-39.9924	-2.4441	-0.9586	-6.7142	-2.6265	252.0665

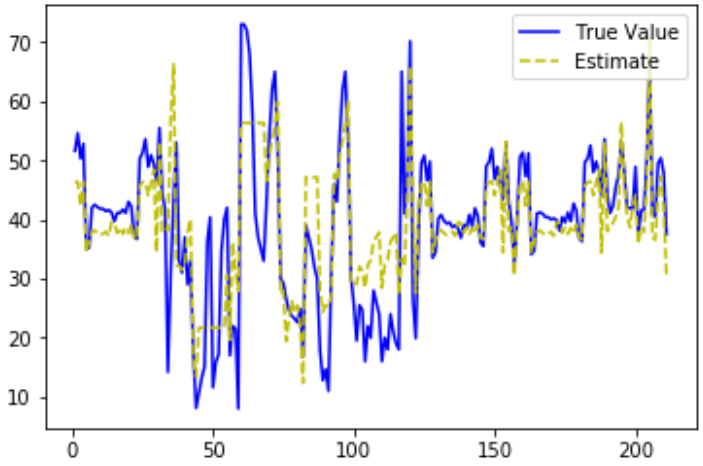
From the tables we can see that we got the same results as using linear regression methods.

Formula Testing

Linear regression was the main method we used in the project. As we have got three linear estimators to estimate the SFE, we used R2 statistic, mean-square-error and graphs (based on the training set from which we had deleted some features and samples) to test their performance.

Linear estimator 1

SFE = $27.742 * \mathrm{N} + 1.954 * \mathrm{Ni} - 0.392 * \mathrm{Fe} - 0.360 * \mathrm{Mn} - 5.943 * \mathrm{Si} + 41.030$;
R2 statistic: 0.5818;
MSE: 70.5368.



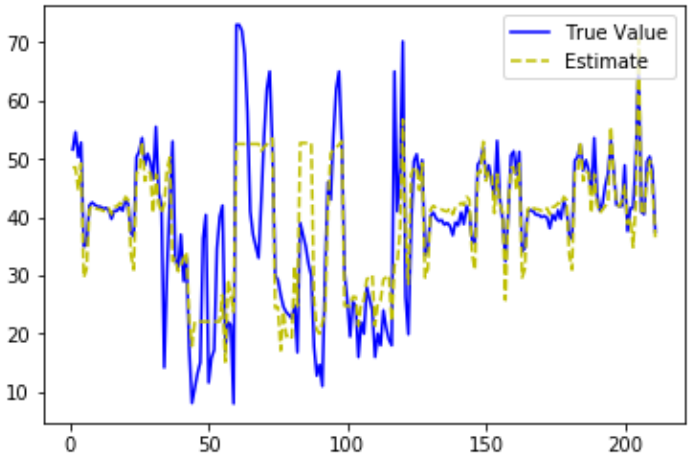
Graph 3-a True SFE and estimate (Linear estimator 1)

Linear estimator 2

SFE = $-2.231 * \mathrm{Ni} - 4.394 * \mathrm{Fe} - 3.038 * \mathrm{Mn} - 7.687 * \mathrm{Si} - 4.210 * \mathrm{Cr} + 439.832$;

R2 statistic: 0.6476;

MSE: 59.4376.



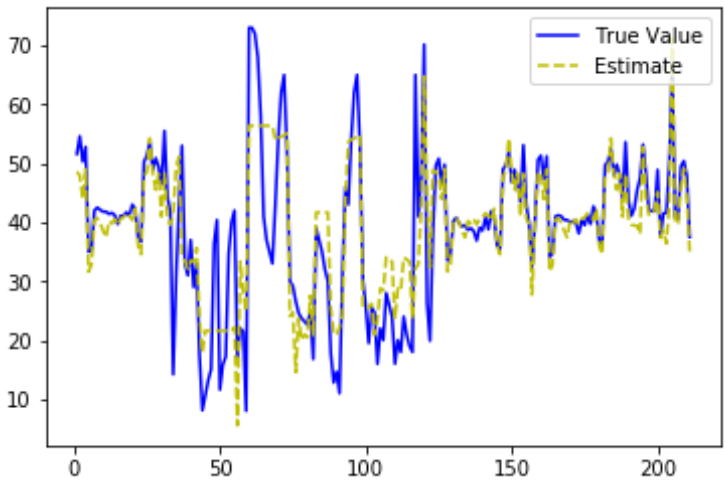
Graph 3-b True SFE and estimate (Linear estimator 2)

Linear estimator 3

SFE = $-39.992 * \mathrm{C} - 2.444 * \mathrm{Fe} - 0.959 * \mathrm{Mn} - 6.714 * \mathrm{Si} - 2.627 * \mathrm{Cr} + 252.066$.

R2 statistic: 0.6621;

MSE: 56.9792.



Graph 3-c True SFE and estimate (Linear estimator 3)

From the testing result we can see that all of the three estimators tracked the true value well. However, based on R2 statistic and mean-square-error, the third estimator based on feature set {C, Fe, Mn, Si, Cr} performed best. Therefore, the best linear estimator

we got for SFE was

$$\mathrm{SFE} = -39.992 * \mathrm{C} - 2.444 * \mathrm{Fe} - 0.959 * \mathrm{Mn} - 6.714 * \mathrm{Si} - 2.627 * \mathrm{Cr} + 252.066,$$

Where the element symbols represent the content of the related element in the alloy.

Work Arrangement

Guanda Li

Selecting features with F-regression and Multi-info-regression methods, processing ridge regression, and related part in the report. Analysis of the Linear regression. Test of linear regression part.

Shaowen Zhu

Preprocessing (Deleting the features and samples), selecting features with exhaustive search, processing linear regression, BLUE, formula testing and related part in the report. Complete the rest part of the report.