# Critical Temperature in Superconductors

Richard Kim
Southern Methodist University

Hien Lam
Southern Methodist University

Joaquin Dominguez
Southern Methodist University

September 8, 2022

**Abstract**

*Using statistical and machine learning methods, we analyzed a publicly-available dataset regarding superconductivity. Our variable of interest was* critical_temperature *and built a linear regression model based on the prediction thereof, considering more than 160 variables. After processing, we compared the results of Lasso (L1) and Ridge (L2) regularization methods using RMSE and $R^2$ as evaluation metrics. The model using Lasso regularization produced a lower RMSE as well as a larger $R^2$ value. Based on the predictive capacity of our model, we encourage the further application of these methods on different datasets regarding the topic of superconductivity.*

## I. Introduction

Superconductivity was discovered by Dutch physicist Heike Kamerlingh Onnes and his team in 1911.[1] Despite more than one hundred years passing since the discovery, it remains one of the most engaging and enigmatic topics in physics, chemistry, and materials science. As currently understood in simple terms, superconductivity is a collection of physical properties found to exist in certain material classes. Under specific conditions, these materials lose their electrical resistance (infinite conductivity) and, consequently, have the ability to generate very large electric currents and thus, equally large magnetic fields—at least in theory.[2] Modern study of the phenomena in those fields has not proven to be as fruitful, however, given the complex nature of the relationship between the state of superconductivity and the chemical structure of the materials being used.[3]

Engaging with alternative approaches to understanding superconductivity may provide insights not found by and through methods with physical limitations, such as those heretofore employed. Statistical and Machine Learning (ML) methods have advanced sufficiently in recent years to engage with important applications such as this. Specifically, in this case, ML models may predict highly useful information without needing to engage with the chemical structures of materials.

We employed such an approach by exploring the function of critical temperature on more than 160 variables. By using regression methods, such as L1 and L2 Regularization, we were able to create a reliable model that serves to predict critical temperature. As more data-sets related to superconductivity are made publicly available, similar models produced, in the interest of inter-disciplinary collaboration, may

[1] Superconductivity. CERN. (n.d.).
[2] Combescot, 2022
[3] Chu, Deng, and Lv, 2015

provide much needed clarity to this phenomena.

# II. Methods

## i. Data Preprocessing

The superconductivity data-set consisted of two files: 'train' and 'material.' The former file contained 82 relevant feature information from 21,263 superconductors including the response feature, 'critical_temp'. The material file contained 88 features which represented the chemical formula for material that were one hot encoded for all 21,263 superconductors. A simplistic example for water ($H_2O$) would have a two under the hydrogen column, one under oxygen, and zero for all the other elements. The two files were concatenated and material subsequently dropped since this feature denoted the full chemical formula and deemed unnecessary after hot encoding. Next, we identified the nine features that had a single value and consequently dropped them due to their lack of usefulness in predictive performance. Lastly, we confirmed that the data was free of missing values and duplicated records. The final unprocessed data frame resulted in 21263 rows and 159 columns. We will note the datatypes were all numerical and proceeded with exploratory data analysis.

## ii. Exploratory Data Analysis

### ii.1 Correlation

Correlations allow us to form a picture of the patterns and dynamics between the variables. We first looked at correlation of features in the following forms:

- Correlation to target variable (*critical_temp*)
- Mutual correlation of non-target variables
- Lack of correlation to target variable (*critical_temp*)

Identifying those variables that had a high correlation with the target variable provides a safeguard during feature selection. Next, we ran a "for" loop to determine those pairs that show a mutual correlation higher than 0.9 (0 to 1 scale), which gives us an indication of those variables that affect variance. In this case, 74 different pairs showed a correlation of 0.9 or higher. From those pairs, we concluded that 39 columns show high mutual correlation, which warrant exclusion prior to modeling.

Lastly, we addressed variables with very low correlation with respect to the target variable (< .01), which resulted in 69 additional variables being excluded. After these exclusions, our data-set contained 51 variables.

### ii.2 Outliers

Upon initial visual inspection of histograms, a few variables displayed skewness with the possibility of outliers. In order to identify and address outliers, and thus reduce MSE, we ran an anomaly detection algorithm ('*IsolationForest*') on the data-set in scaled form, with optimal hyper-parameter tuning. The algorithm only identified two rows as outliers, and said rows were removed.

### ii.3 Assumptions

**Linearity**
In all variables, there exists a linear relationship between the independent variable, x, and dependent variable, y.

**Independence**
We may assume independence for all instances.

**Multicollinearity**
Multicollinearity was addressed in the correlation stage of the EDA; those variables with high multicollinearity were removed.

**Homoscedasticity**
Residuals were shown to have constant variance.

**Normality**
Residuals were shown to be normally distributed.

### iii. Feature Scaling

Given that the features followed a normal distribution, we felt confident Scikit-Learn's standard scaler class was the most appropriate feature scaling for the dataset. This estimator bounded each feature to maintain a mean of zero and standard deviation of one. We conducted linear regression with scaled and unscaled data and compared their performance by way of five-fold internal cross validation. Looking at $R^2$ and root mean squared error (RMSE), both models produced equivalent metrics at 0.45 and 23.8, respectively. We split the data into X and y (training features and target feature, respectively) and proceeded with standard scaler going forward.

### iv. Modeling

As stated above, we would like to predict the critical temperature of a superconductor using linear regression with regularization then investigate the important features. Specifically, Lasso and Ridge regularization were explored. These methods allow us to reduce model complexity and prevent over-fitting. The function of Lasso is to employ a cost function (*Alpha* or *Lamda*), whereby the account of magnitudes can lead to zero coefficients—a highly useful tool for the purpose of feature selection. Ridge, on the other hand, employs the same cost function, but taking into account the square of the coefficients instead, shrinking coefficients and thus aiding in the prevention of overfitting.

Given our goal to investigate feature importance, an L1 model does make more conceptual sense. Nonetheless, we proceed by creating both models. We utilized Scikit-Learn's *pipeline* class to set up the gridsearch workflow, *Lasso* and *Ridge* classes to train the models, *gridsearchcv* class to execute the hyper-parameter search then refit the best performing model, and *cross_validate* class to ascertain their mean RMSE from five-fold, shuffled internal cross validation. The feature importance was extracted from the best model of each algorithm and its values transformed to the absolute root because we care about the magnitude of the coefficients, not the sign of the coefficient.

## III. Results

### i. Regularization

We conducted a Lasso and Ridge gridsearch with hyper-parameter tuning. Lasso's alpha was set to be between 0.1-10 with 200 samples and Ridge's alpha to be 0.1-100 with 300 samples. Displayed below is a table of the optimal alpha from each model and their respective five-fold cross validation metrics.

| Optimal Alpha | | |
|---|---|---|
| | Lasso | Ridge |
| Alpha | 2.03 | 100 |
| $R^2$ | 0.66 | 0.53 |
| RMSE | 20.09 | 22.43 |

**Table 1:** *Optimal Alpha*

### ii. Feature Importance

Given that the data were standardized prior to modeling, the coefficients across our variables are on the same scale. This allows for direct comparisons, where we may interpret larger coefficients as indicative of greater contribution or importance. Although the positive and negative signs of the coefficients provide information on directionality, we focus on the magnitude of the coefficients and summarize our findings on feature importance below. The most important variables identified by our L1 model were, in order:

- *wtd_std_ThermalConductivity*
- *Ba*
- *wtd_entropy_atomic_mass*
- *Ca*
- *wtd_gmean_ElectronAffinity*

By contrast, our L2 model identified the following as the most important variables:

- *wtd_std_ThermalConductivity*

- *std_ElectronAffinity*
- *range_ElectronAffinity*
- *Ba*
- *range_atomic_mass*

We found that *wtd_std_ThermalConductivity* came up with the greatest contribution in both cases, and *Ba* fell into the top 5 important variables in both of our models.

## IV. Conclusion

The linear regression model using Lasso regularization produced a lower RMSE at 20.09, as well as a larger $R^2$ value at 0.66. By comparison, the model using Ridge regularization had an RMSE of 22.43 and an $R^2$ value of 0.53. These results indicate that our Lasso model is a better fit than the Ridge model. This, in addition to the fact that Lasso regularization is generally more applicable for identifying important features, lends support to its preferred use over the ridge model for this case study.

In line with these results, we highlight our findings from the Lasso model, concluding that the most important feature in predicting critical temperatures is *wtd_std_ThermalConductivity*. *Ba* was identified as the second most important feature, which was also the only other feature that the Ridge model shared in the top 5 list. Above all other features, these two variables have the highest predictive power for critical temperatures after accounting for possible overfitting. It may also be worth noting that measures surrounding electron affinity consistently showed up as relatively important features in both of our models. Despite our findings here, this additional insight may warrant a deeper look at the relation between electron affinity and critical temperature.

### i. Code

Please refer to the attached Python notebook.

## References

[Chu, Deng, and Lv, 2015] Chu, C. W., Deng, L. Z., and Lv, B. (2015). Hole-doped cuprate high temperature superconductors. *Physica* C: Superconductivity and its Applications, 514, 290-313.

[Superconductivity. CERN, 2009] Superconductivity. CERN. (n.d.). Retrieved September 7, 2022, from https://home.cern/science/engineering/superconductivity

[Combescot, 2022] Combescot, R. (2022). Superconductivity: An Introduction. Singapore: *Cambridge University Press.*