TIF345/FYM345 Project-2a: Alloy cluster expansions

Markus Utterström Salar Ghanbari

January 27, 2025

Introduction

Alloys, materials composed of multiple elements, are among the most common forms of metals found in nature, as pure metals are rare. The Au-Cu alloy, in particular, is significant due to its unique properties and applications. This project focuses on the cluster expansion of the Au-Cu alloy, a technique widely used in materials science to model and predict thermodynamic properties. Bayesian statistics are applied to fit the parameters of the cluster expansion, offering insights into behavior of the alloy. In this analysis, increasingly complex solvers and methods are explored and discussed, including Ordinary Least Squares (OLS), Ridge Regression, Covariance-Regularized Regression, Bayesian Cluster Expansion, and Automatic Relevance Detection Regression (ARDR)[1]. Finally, these methods are applied to predict the ground-state configuration of candidate clusters and assess their performance in identifying the structure with the lowest energy. The data analysis was conducted using Python, leveraging its libraries for statistical modeling and machine learning.

Task 1

For the first task, the database with the atomic structures was read using the ASE package. From there a cluster space object was created with the icet package. The relevant orbits were extracted with the cutoffs for pairs, triplets and quadruples at respectively 8,6 and 5 Å [2]. With this cluster space, a cluster vector and mixing energy was extracted for all orbits. The first task was then to plot the concentration of Cu-atoms against the mixing energy which is found in the figure 1.

The second part of this task was to standardize the cluster vectors and mixing energy. This was made using the functionality of the `sklearn` package with the functions `StandardScaler`. Standardization means setting the mean of all mixing energies to zero and variance to one. This negates the effect of outliers of the data so that every data point contributes equally. The data points with a larger scale will not contribute disproportionally. Standardization is often done in the context of machine learning when a gradient descent algorithm is applied. For this project, using standardized data is beneficial as regression techniques such as ridge regression will be used where using standardized data is better [3]. Ridge regression assumes that all data points contribute equally to the penalty term, by using standardized data the scale is the same for all points.

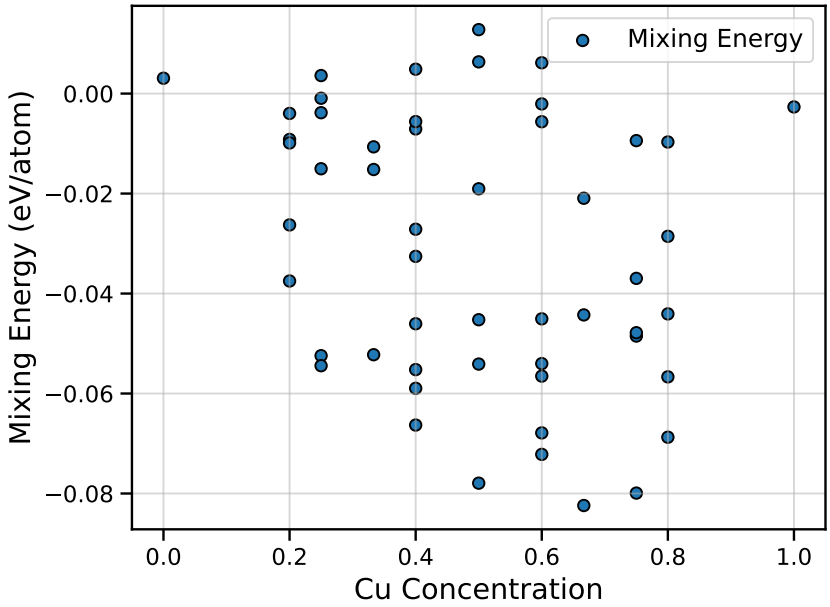

Figure 1: Concentration of Cu-atoms against the mixing energy for that orbit.

Task 2 & 3

In task 2, ordinary least squares (OLS) and Ridge Regression are used to fit the effective cluster index (ECI). The loss function is$L = ||\boldsymbol{E} - \boldsymbol{X}\boldsymbol{J}||^2 + \alpha||\boldsymbol{J}||^2$. $\boldsymbol{E}$ is the mixing energy, $\boldsymbol{X}$ is the cluster vector design matrix and $\boldsymbol{J}$ is the ECI. The equation for OLS looks like: $\boldsymbol{J}_{opt,ols} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{E}$. Ridge regression has the following form $\boldsymbol{J}_{opt,Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{E}$, where alpha is the hyperparameter than penalizes large values of ECI. The best value for α is calculated with the help of k-fold from the `sklearn` package. To compare the two models, Cross validation root-mean-square error (CV-RMSE) is applied, which gives a score.

Physical intuition can be a very strong tool when there is high dimensionality in a problem. By replacing $\alpha\boldsymbol{I}$ in the ridge model with a regularization matrix Λ, which is the same as applying a prior on the form $P(J|X) \propto e^{-J^T\Lambda J/2}$. We use a reduced version here and make Λ a diagonal matrix with elements

$$\lambda_\alpha = \frac{\sigma^2}{\sigma_\alpha^2} = \gamma_1 r + \gamma_2 n, \tag{1}$$

where r is the radius and n number of sites of the orbits. This imposes a restriction that $\lambda_\alpha > 0$. To find these values of λ_α, CV-RMSE is used in tandem with `scipy` minimize to find the best fits for the values of γ.

Figure 2 illustrates a comparison of the parameter values for the models, specifically the previously mentioned ECI. The Ridge model used an α value of 0.0764, while the blue bars represent OLS, the orange bars represent Ridge regression, and the green bars represent covariance-regularized regression. The CV-RMSE value was calculated to be 0.9928, with $\gamma_1 = 0.135$ and $\gamma_2 = 0.056$, indicating the contributions of radius and the number of sites, respectively.

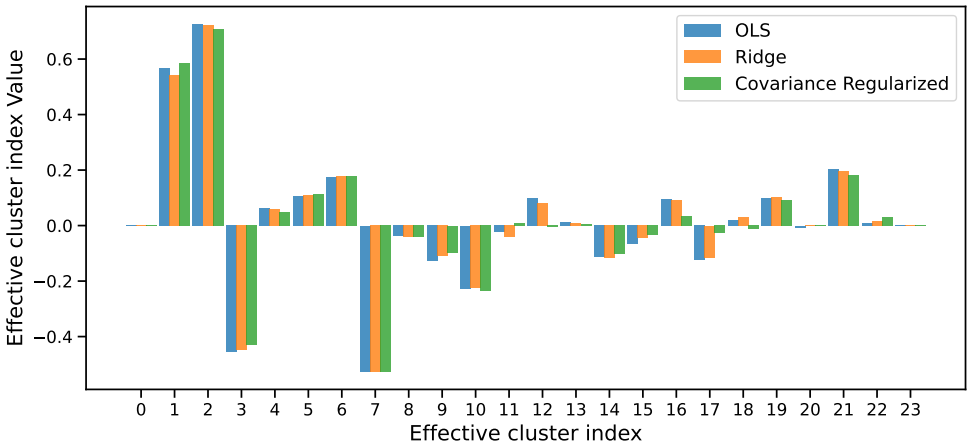

Figure 2: Comparison of ECI values obtained using OLS, Ridge, and Covariance Regularized methods for effective cluster indices.

The associated CV-RMSE scores for the models are, for OLS 0.1060, for Ridge 0.1057. It is a very small difference between them with a very slight favoring of the Ridge regression model. The purpose of using Ridge over OLS is to combat overfitting. Ridge introduces a penalty term α that favors smaller parameter coefficients as can be seen in figure 2. While OLS does not impose any restrictions on itself, Ridge lowers the overall variance of the model as well by favoring the smaller parameter values. For an even higher dimensional model, the threat of overfitting increases and Ridge should give a better fit. The small value of α indicates that Ridge almost behaves like OLS which means that the two models give roughly the same fits.

The CV-RMSE lower score shows favor for using covariance regularized over the two other models. Using this method over a uniform penalty for all values considers the varying sizes of the orbits and how important they are. From the values of γ we see that the orbit radius contributes more to it being regularized over the number of sites in it. Using only Ridge does not consider this and OLS ignores this completely.

Task 4

In the previous task, various methods have been applied to fit the parameters of data and for this task, a full Bayesian analysis using Markov chain Monte Carlo (MCMC) is done. For ECI, the following prior is used:

$$P(\boldsymbol{J}) = \frac{1}{(2\pi\alpha^2)^{N_p/2}} \exp\left(-||\boldsymbol{J}||^2/2\alpha^2\right). \tag{2}$$

Where α^2 is the variance and the mean is zero. Homoscedastic errors are assumed and thus σ is set to be the standard deviation. Both α and σ have a weak inverse gamma prior. The total prior function is the product of the three priors. The log-likelihood is a Gaussian distribution on the form

$$P(D|J,\sigma) = \frac{1}{(2\pi\sigma^2)^{N_d/2}} \exp\left(-\frac{(\boldsymbol{E} - \boldsymbol{X}\cdot\boldsymbol{J})^2}{2\sigma^2}\right) \tag{3}$$

Where $\boldsymbol{E} - \boldsymbol{X}\cdot\boldsymbol{J}$ is the residual that is used. To use MCMC effectively the log is taken of the prior and the likelihood to deal with sums rather than products. To gather

uncorrelated samples, the autocorrelation of the chain was calculated for each walker. The highest number was then used to thin out the samples which resulted in 4600 uncorrelated samples.

Figure 3 is a violin plot showing the various posterior distributions for each ECI. The colored areas represent the minimum and maximum values that were encountered during sampling. As can be seen, the errors start to increase for the higher indices.

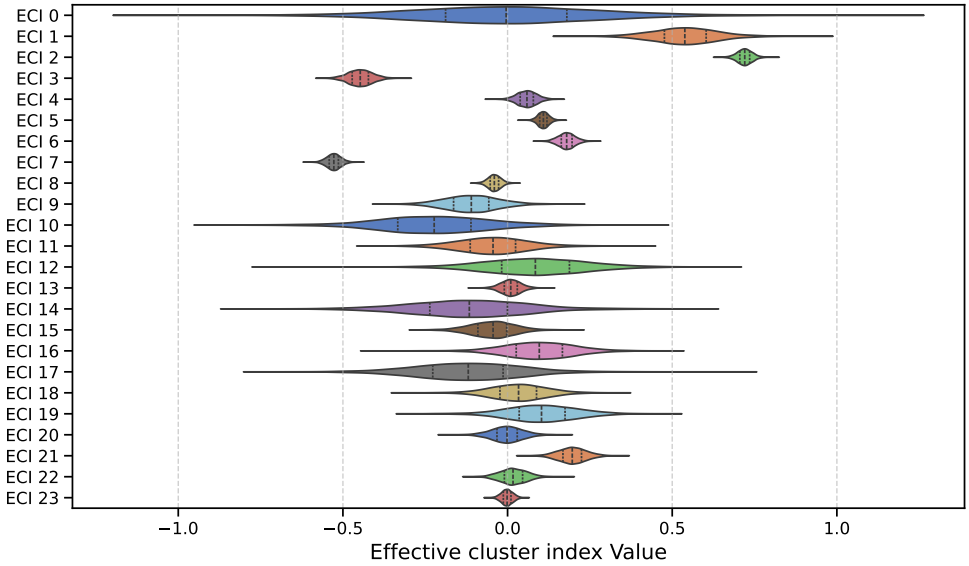

Figure 3: Violin plot of the posterior distributions of the ECIs from Bayesian MCMC sampling. The colored regions indicate the range of sampled values, with higher uncertainty observed for indices with larger values

When inspecting the figure above Figure 3, Eight parameters are strictly non-zero. By setting priors that favor higher order clusters, would increase the complexity of the model and probably increase the risk of overfitting. The higher order clusters are in general not as necessary as the others because they are smaller.

Task 5

For this task, a final method to fit the parameters is tested, Automatic Relevance Detection Regression (ARDR). This method uses a scaling parameter called threshold-λ that is varied to find the one that minimizes the Cross Validation (CV) error with k-folding. This technique is tested on other cutoffs for the orbit which results in more parameters that need to be determined. ARDR will be powerful enough that it will not cause any problems.

As shown in figure 4 for CV-error and threshold lambda from values 50 up to 1000. The red dashed line presents the lowest error. To compare how many non-zero coefficients the model needs, AIC and BIC score was calculated for each of the different iterations. The result is presented below.

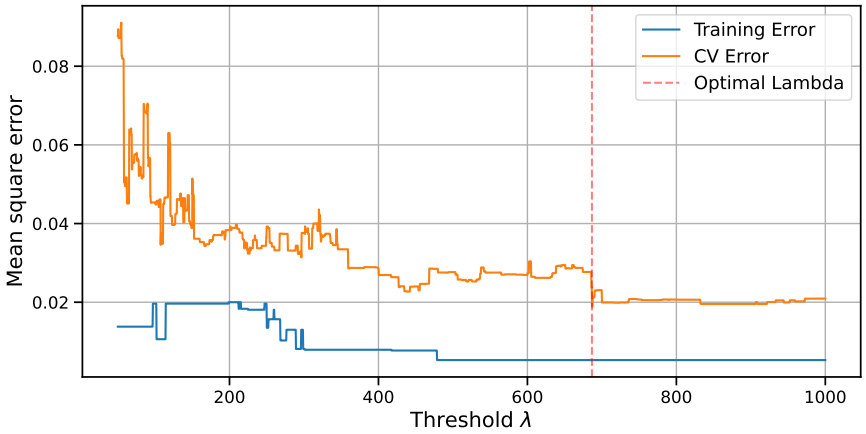

Figure 4: Training and CV errors vs. threshold λ, with the optimal λ minimizing CV error highlighted.

The non-zero ECI values identified by the ARDR model were (1, 2, 3, 5, 6, 7, 27, 76, 84, 85, 88). Most of the lower-order ECIs were retained as non-zero, with only a few higher-order ECIs selected. This indicates that lower-order clusters play a more significant role in the model, as they contribute more parameters compared to the less frequently selected higher-order clusters.

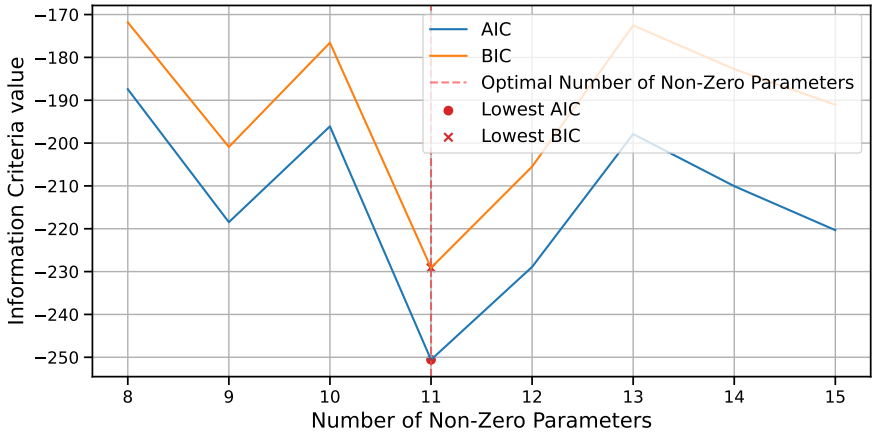

Figure 5: AIC and BIC values plotted against the number of non-zero parameters, showing the optimal number with the lowest AIC and BIC values highlighted.

With ARDR, we identified the most relevant ECIs, highlighting the significance of lower-order clusters. In Task 6, we test all the derived models on ground-state candidate data to evaluate their performance in predicting the structure with the lowest energy.

Task 6

For this final task, all the techniques presented in this report will be tested on provided ground state candidates. The models were tested on the candidates and the calculated mixing energy for each candidate can be seen in figure 6. As can be seen in the figure every model except the regularized covariance (RC) predicts the ground state to be at index 4, The (RC) one says it is at 29. From the Bayesian analysis, the frequency of each

candidate being the lowest yielded three possible ones: Index 4 30 % of the time, Index 28 0.25 % and Index 29 45%. This would indicate that index 29 is more probable which is not the case for the other models. For example, the ARDR model is far away from it.

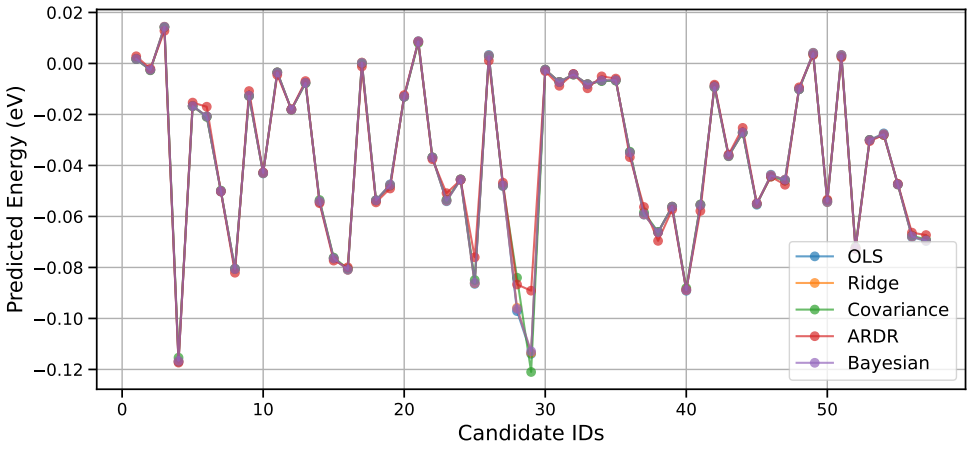

Figure 6: Comparison of predicted ground-state energies across OLS, Ridge, Covariance, and ARDR models. The overlap suggests similarity in model predictions for limited datasets.

The figure below shows the distribution of the ground state energies from the Bayesian analysis. The most frequent value matches the ones from the other models being just above -0.12 eV.

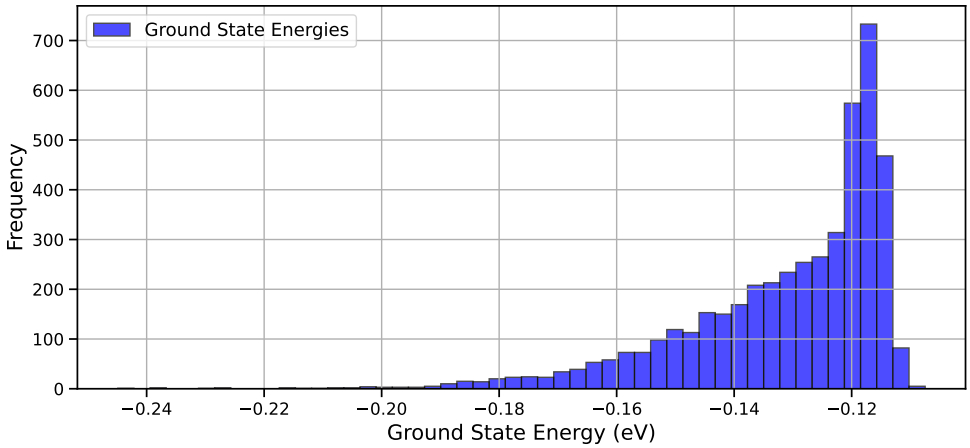

Figure 7: Distribution of the ground state energy for each uncorrelated sample from the Bayesian analysis.

To say which method is most suitable for this problem is difficult as they are all similar to each other. The results do not vary much even if there is a slight difference between the two ground states. There is also the consideration of how long it takes to run each model with the Bayesian analysis taking far more time than the rest. With more data points the use of these complex models might prove more worthwhile but for this smaller scale problem, OLS or Ridge gives a lot of insight despite being the simpler ones. Still using the physical intuition such as (RC) gives a lot of opportunities to make use of phenomena that the more static models do not have. In a physical problem using methods that maximize the utility of physics over statistics can be very powerful and should be used.

References

[1] Paul Erhart, Andreas Ekström, and Arkady Gonoskov. *Advanced Simulation and Machine Learning*. Lecture notes, 2024.

[2] Michael Widom and Marek Mihalkovič. Bulk properties of disordered metallic alloys, intermetallic compounds, and glasses. *Physical Review B*, 80:024103, 2009.

[3] Wessel N. van Wieringen. Lecture notes on ridge regression, 2023.