Alloy cluster expansions

Erik Andersson, Sebastian Holmin

January 2021

In this report we will investigate the issue of parameter selection and estimation in cluster expansions of alloys. To do this we will use the `icet` package which implements symmetry transformations to expand the mixing energy of alloy structures into

$$E_{\text{mix}} = J_0 + \sum_\alpha N_\alpha J_\alpha,$$

where N_α is the number of a α-clusters per atom and J_α is the effective cluster interaction (ECI), which are the parameters that we seek to estimate from energy data.

Assuming i.i.d. errors this can be written in matrix notation as $\boldsymbol{E} = \boldsymbol{X}\boldsymbol{J} + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and thus the likelihood function is given by

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

The Bayesian and Akaike information criteria are defined at the maximum likelihood, which can be shown to be equivalent to

$$\begin{aligned}\text{BIC} &= -N_d \log(\text{MSE}) - N_p \log\left(N_d\right) + \text{const} \\ \text{AIC} &= -N_d \log(\text{MSE}) - 2N_p + \text{const}\end{aligned}$$

where MSE is the mean squared error.

1 The data

The data used for fitting throughout this project is shown in fig. 1. Here, the energy of the given structures are plotted against the Pd concentration of the structure.

2 Parameter selection through cutoff

We first investigate a simple criteria for parameter selection which only includes pair clusters below a given cutoff. To investigate the optimal cutoff we fit the selected ECIs using OLS and measure the cross validation errors as well as information criteria.

Specifically we use 10-fold cross validation to extract both training and validation errors and then use the full data set to evaluate Akaike and Bayesian information criteria. The root mean

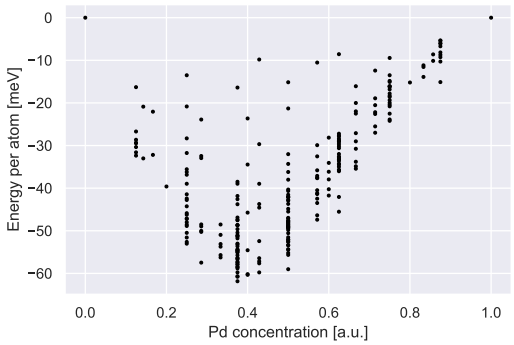
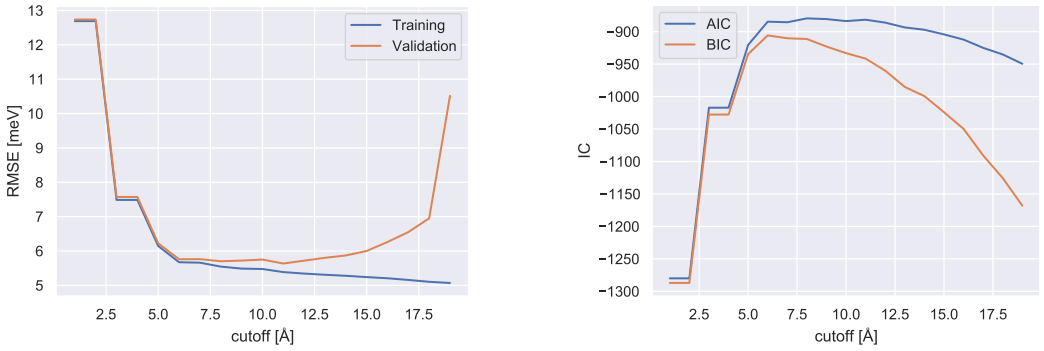


Figure 1: Data of mixing energies for different concentrations of Pd in a Ag/Pd alloy.



(a) Root mean squared error for cross validation.

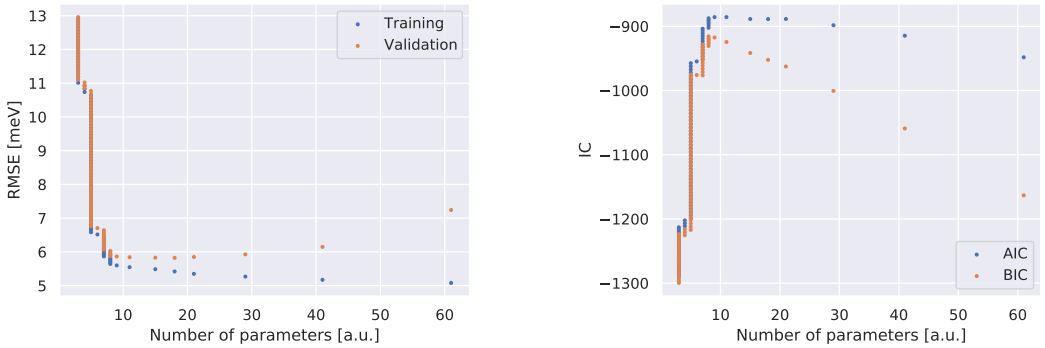
(b) Bayesian and Akaike information criteria.

Figure 2: The RMSE and information criteria scores of the OLS as a function of the pair cutoff used for the cluster space.

squared error (RMSE) for training and cross validation is shown in fig. 2a, we can see that there's no significant gain in validation performance after a roughly 6 Å cutoff (which corresponds to 6 parameters). Since more parameters will always yield smaller RMSE for the training set due to overfitting, we are only really concerned of the validation RMSE. The information criteria is shown in fig. 2b. We can see that BIC tends to prefer a simpler model and has its maximum at 6 Å, while AIC is mostly flat throughout 6-11 Å.

3 Feature selection

A more sophisticated method of choosing ECIs is utilizing feature selection algorithms. We will compare LASSO and ARDR. These are variations of linear regression algorithms that perform variable selection, which is controlled by parameters α and λ respectively (a large α induces fewer parameters in LASSO while a small λ does the same in ARDR).



(a) Root mean squared error for cross validation.

(b) Bayesian and Akaike information criteria.

Figure 3: The RMSE and information criteria scores of the LASSO regression as a function of the number of parameters. Note that several values of the hyperparameter α corresponds to the same number of non-zero parameters in the fit.

To fairly evaluate the importance of parameters J_α, normalization is required. For example a low energy cluster (small J_α) that occurs often (large N_α) still might contribute significantly to the total energy. Note here that the bias is a constant 1 and is left unmodified.

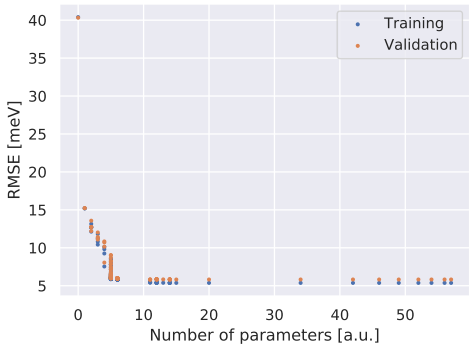
The RMSE and information criteria scores are shown for LASSO in fig. 3 and for ARDR in fig. 4. In the case of LASSO, the BIC score is clearly maximised at 9 parameters, while the AIC has a its maximum and is fairly stable in the region of ~ 10 to 21 parameters. This is also supported by the validation RMSE. Note however that multiple α/λ values can give the same number of parameters but still differing scores. In the case of ARDR, however, the measures agree. Both information criteria favour 11 parameters, at which point the validation RMSE plateaus. We can see that in our experiment that ARDR shows a more robust behaviour compared to LASSO. The training and validation RMSEs are closer and stable after ~ 10 parameters, while for Lasso they diverge, i.e. the algorithm shows some overfitting, which starts to grow significantly after ~ 20 parameters.

4 Bayesian cluster expansion

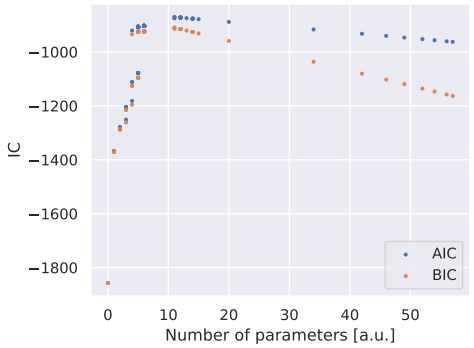
In the Bayesian approach we construct the posterior distribution as the product of the likelihood of eq. (1) and the prior $p(\boldsymbol{J}, \sigma^2, \alpha^2) = p(\boldsymbol{J})p(\sigma^2)p(\alpha^2)$. As a prior for $\boldsymbol{J}$ we use a Gaussian i.i.d prior

$$P(\boldsymbol{J}) = \frac{1}{\alpha^{N_p}} \exp\left[-\frac{||\boldsymbol{J}||^2}{2\alpha^2}\right]$$

and for σ^2 and α^2 we use inverse gamma priors, shown in fig. 5. The mean for α^2 was chosen close to the variance in the ECI's obtained from the OLS method and the mean for σ^2 to correspond to the RMSE observer earlier. The width was chosen to give quite lenient intervals of relatively high probability. Note that, since the model only uses the variances α^2 and σ^2, we can for convenience use priors for the variances instead of the stds. We then sampled the log-posterior using MCMC sampling through the `emcee` package. In the sampling we used 100 walkers taking 5000 steps each with a burn-in period of 3000 steps, i.e. effectively 2000 samples per walker. The walkers were

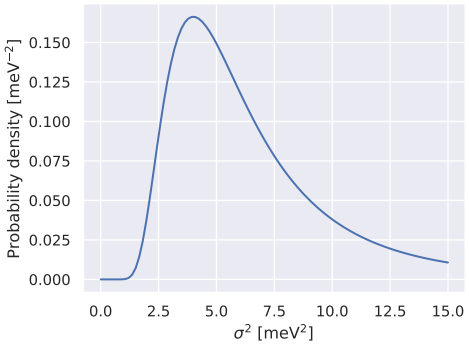


(a) Root mean squared error for cross validation.



(b) Bayesian and Akaike information criteria.

Figure 4: The RMSE and information criteria scores of the ARD regression as a function of the number of parameters. Note that several values of the hyperparameter λ corresponds to the same number of non-zero parameters in the fit.



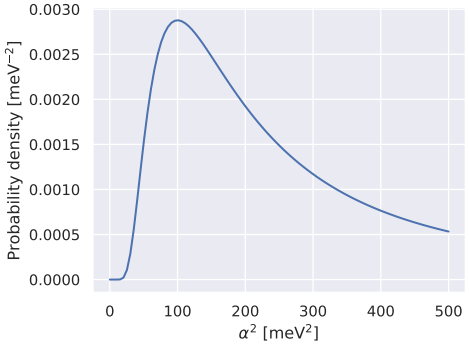


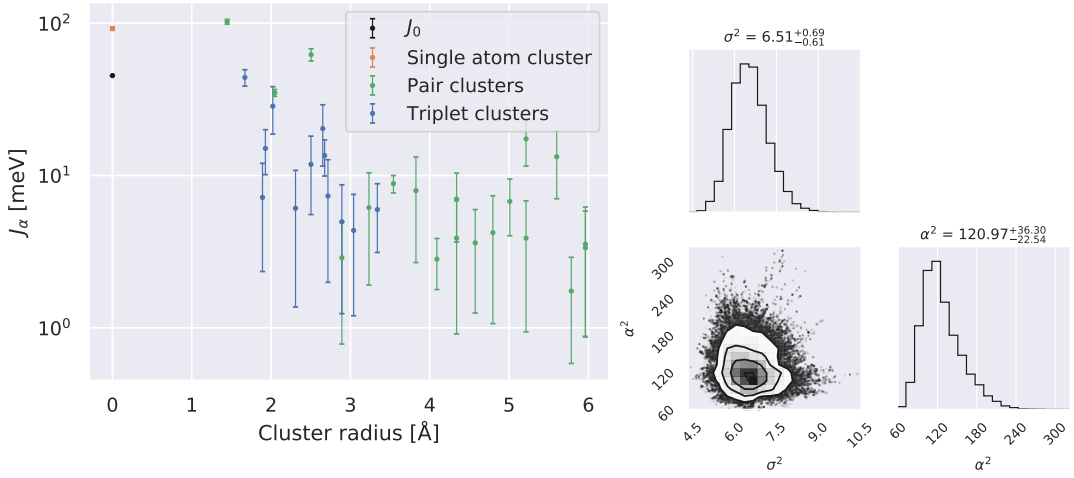
Figure 5: Prior probability functions for σ^2 and α^2.

independently and randomly initialized using normal distributions. For the J_n we used a normal distribution with mean zero and standard deviation 10 and for the α^2 and σ^2 we used means 7 and 30 respectively and standard deviation 10.

The MCMC sampled posterior distributions are shown in fig. 6. The mean and standard deviation of the ECI ordered by cluster radius is shown in fig. 6a. Note here that we used the standardized design matrix $\boldsymbol{X}$, meaning that the resulting values are scaled in proportion to their 'importance'. We can see that ECIs of large radii clusters are generally less important than smaller radii clusters. We can also note a few clusters stick out as significantly more important than the others.

5 Ground state analysis

We will now compare methods to predict the ground state from a list of candidate structures. First, we simply extract the energy of each ground state candidate using the ECIs of an OLS fit. We



(a) Mean and standard deviation of effective cluster interactions J_n (using the standardized design matrix $\boldsymbol{X}$) ordered by radius and colored by cluster type.

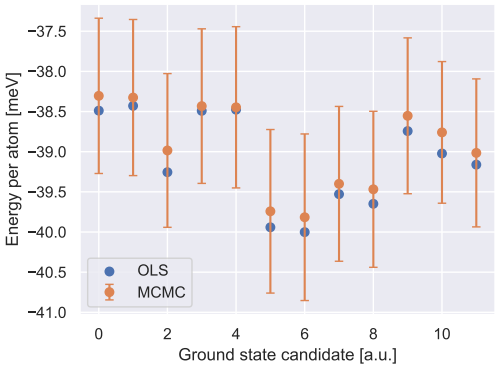
(b) Posterior plots for σ^2 and α^2.

Figure 6: MCMC sampled posterior distributions.

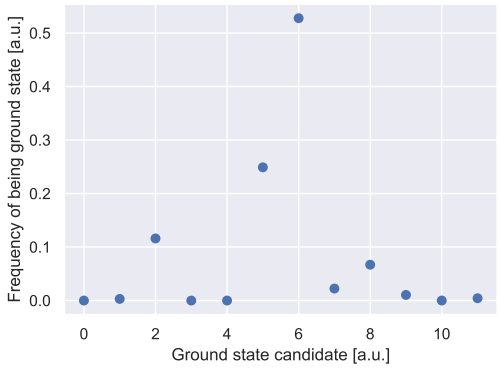
then use the ECIs from our MCMC samples to compute mean and standard deviations of energy for each candidate. The results can be seen in fig. 7a. We can see that the MCMC mean ground state energy corresponds quite well to the OLS values, but the standard deviations are fairly large. In both cases candidate six seems to be the most likely ground state, with five as a close second.

Another approach for determining the ground state comes from looking at the frequency of cluster candidate being the ground state in the MCMC samples of ECI parameters. This can be seen in fig. 7b. Using the lowest energy candidate for each sample we also plot the distribution of the ground state energy, which can be seen in fig. 8.

We can see that the two approaches agree very well. Both on which candidate is most probably the ground state as well as the ground state energy at about -40 meV/atom. Note however that the lead of candidate six in fig. 7b is much greater than in fig. 7a, where in particular the size of the error bars far exceeds the lead in mean value. This can be explained by the fact that candidate six consistently has a slightly lower energy than five for sampled parameters, as opposed to them alternating often. From this information we can fairly confidently say that candidate six is a more likely choice despite the lead being narrow.



(a) The predicted energy per atom in meV of the given ground state candidates. For the MCMC evaluation the error bars correspond to one standard deviation.



(b) The normalized frequency of each candidate being the ground state in the Bayesian analysis.

Figure 7: Predictions of ground state clusters from a set of candidates using two different methods.

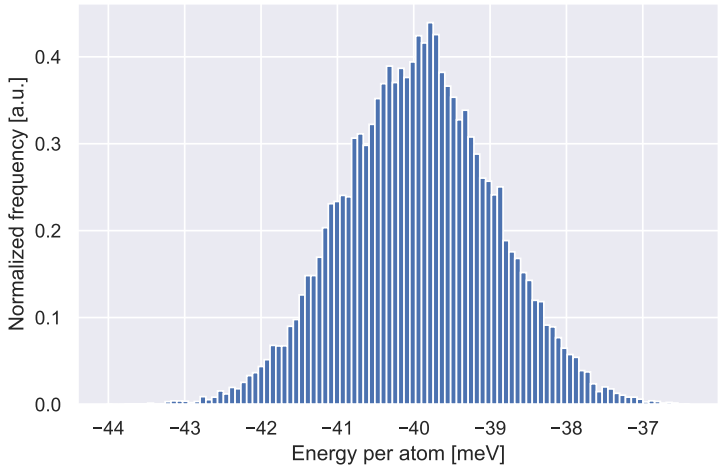


Figure 8: The normalized distribution of the ground state energy from the MCMC samples.