Parameters of Molecular Models beyond their initial Field of Application

Data Analysis for Machine Learning
Maximilian Fleck

November 14, 2023

Brief description of the data set and a summary of its attributes

The dataset consists information about chemical species like parameters of a physical model to describe static properties (pressure, density, phase equilibria,...), parameters of an empirical approach to predict transport properties based on the physical model for static properties (viscosities described as a polynomial in a 2D subspace), chemical information (molecular weight, polarities, associating) and information about the area of validity and quality of the empirical approach to predict viscosities.

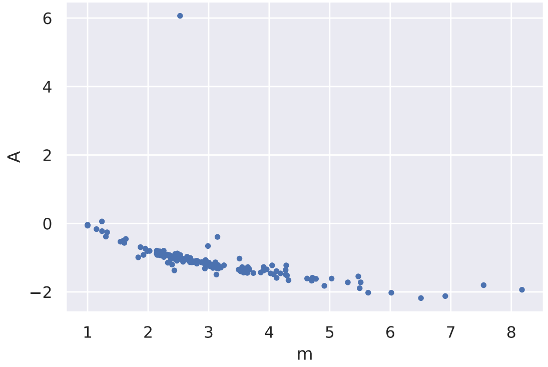

Figure 1: *The outlier is clearly visible.*

Initial plan for data exploration

The plan is to determine the extent to which the parameters of the physical model m, σ, ϵ_k, κ_{AB} and ϵ_{kAB} are related to those of the polynomial in the empirical approach. Although there is no clear physical connection between the two, correlations are conceivable and could form the basis for models to predict viscosities.

Such models could be powerful, as there is a strong thermodynamic intuition that a physical relationship exists and very good empirical models exist for individual substances. Only parameter A of the empirical approach will be analysed in this work. Parameters B, C and D can be investigated in follow-up studies. We want to point out that a correlation is most likely to be expected for A.

Actions taken for data cleaning and feature engineering

First, all columns not needed for our study are removed. If no parameter is assigned in the models, the parameter needs to be set to zero. There is also a clear outlier visible in Figure 1 which is removed.

Then we find and select those attributes which best explain the relationship of the independent variables with respect to the target variable. The target variable is A whereas the attributes are parameters of the physical model to describe static properties m, σ, ϵ_k, κ_{AB}, ϵ_{kAB} and mu.

Exploratory Data Analysis

During EDA correlation between all independent variables and A is calculated. Skew variables are log transformed. Only A and log(m) are correlating linearly. There are also physical reasons for the log transformation of m that will not be discussed here. Log transformed results are shown in Figure 4.

A and log(m) correlate linearly and are distributed with low skew in a normal-distributed manner. Consequently, the product of A and log(m) is also distributed with low skew in a normal-distributed manner. The shape of the distributions A and log(m) and their low skew is surprising. Therefore the analysis was repeated with a subset of this dataset where polar and associating molecules are excluded. In this subset, A and log(m) also correlate linearly and are distributed with low skew in a normal-distributed manner.

Hypothesis and Testing

This leads to the following 3 hypothesis:

(1) There is no difference between A of long (here: many segments m) molecules (m $\geq$ 3) and A of

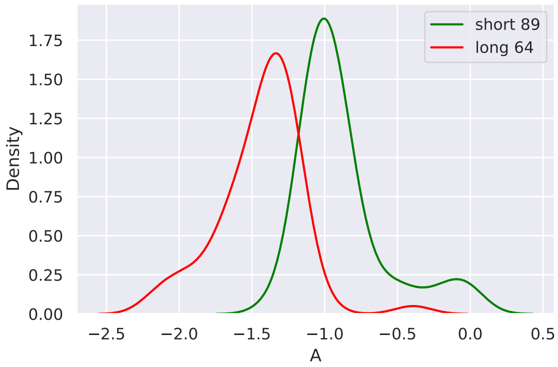

Figure 2: *A Distributions for long and short molecules.*

short molecules (m < 3).

(2) There is no difference between σ of long molecules and A of short molecules.

(3) There is no difference between ϵ_k of long molecules and A of short molecules.

Now we will test the hypothesis (1) if there is no difference between A of long molecules (m $\geq$ 3) and A of short molecules (m < 3). The sample size is 89 for short molecules and 64 for long molecules. The p-value is lower than $\alpha = 0.05$.

Therefore we reject the hypothesis and conclude that there is a difference between A of long molecules and A of short molecules. This outcome seems reasonable as differences between distributions of A for short and long molecules are clearly visible in Figure 2.

Next steps

We can approximate A as a linear function of log(m) as shown in Figure 3.

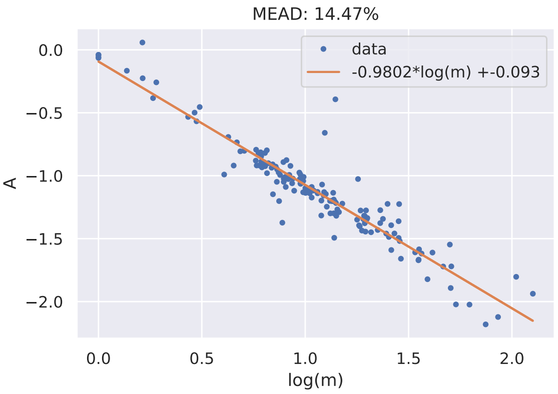

Figure 3: *A can be described as a linear function of log(m). The parameters of the linear function shown in the legend indicate the relationship A = -log(m).*

This is a good basis for less trivial descriptions of A depending on more than m and for analysing parameters B, C and D. Of course, a parameter-free approach would be best, as the parameters A, B, C and D themselves and their relation can be subject to errors and biases.

Quality and additional data

The quality of this data set is sufficient. Uncertainties and scatter are expected as some of the molecules are very complex but described using relatively simple models. Additional data for long-chained molecules is needed to deeply investigate A and log(m) distributions and the linear relation between A and log(m). More data on polar and associating species should be helpful for analysing C and D.

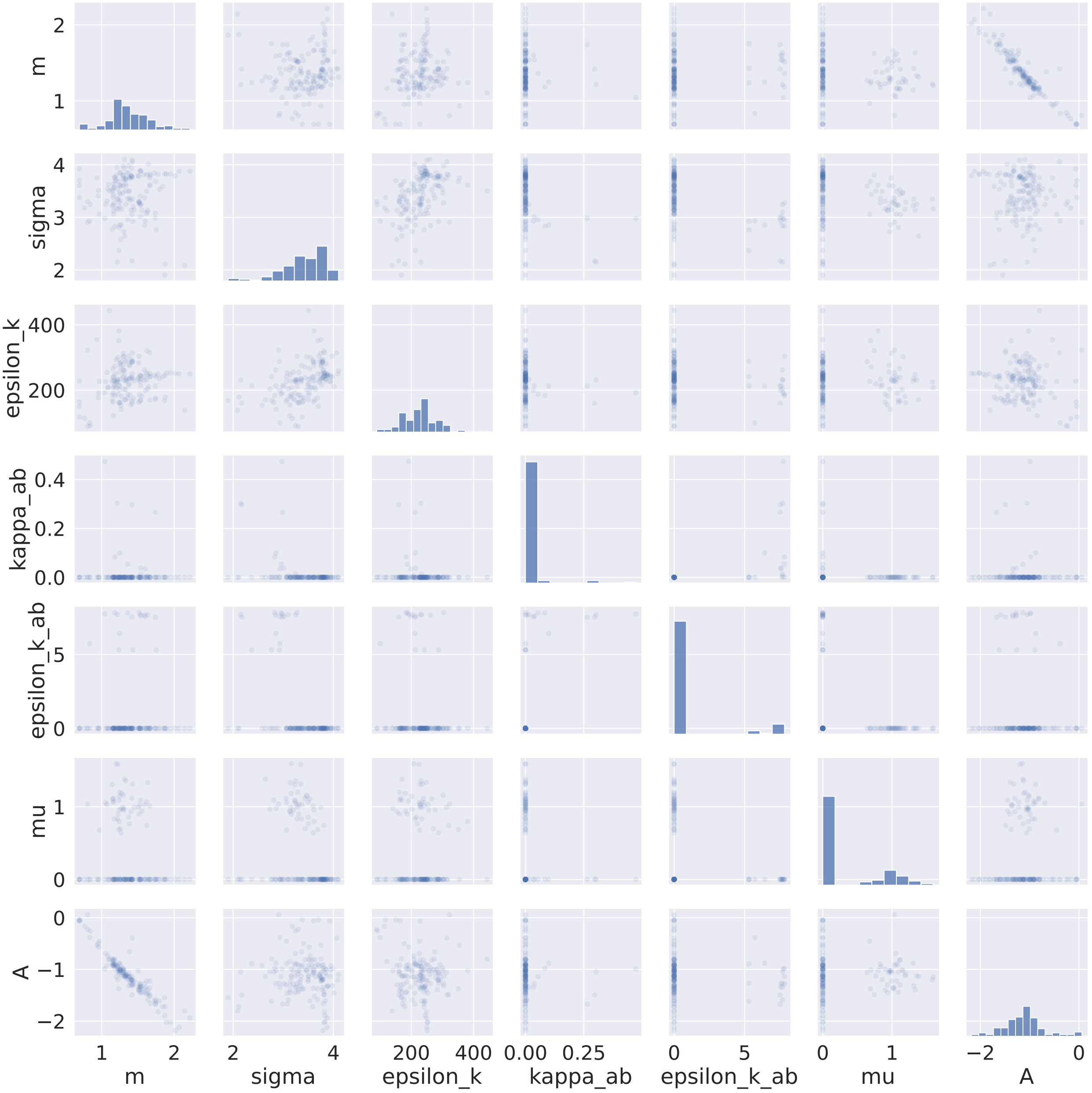

Figure 4: *If necessary log transformed data. A and log(m) are pearson correlate with 0.94. All other parameters correlate with less than 0.15.*