Asymptotic and Bayesian Approaches to Uncertainty Quantification of Interatomic Models

Kinamo Williams

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Bachelor of Science

Mark Transtrum, Advisor

Department of Physics and Astronomy

Brigham Young University

ABSTRACT

Asymptotic and Bayesian Approaches to Uncertainty Quantification
of Interatomic Models

Kinamo Williams
Department of Physics and Astronomy, BYU
Bachelor of Science

Interatomic models (IMs) are used in molecular modeling to predict material properties of interest. The development of an IM can take several months to years and relies on expert intuition, and yet these potentials are usually only valid for a particular application of interest. Extending existing IMs to new applications is an active area of research. Quantifying the uncertainty of an IM can tell us how much we can trust the predictions it makes. I discuss here two methods for analyzing uncertainty: Fisher Information Matrix (FIM) and Markov Chain Monte Carlo (MCMC). Using MCMC methods, I sample from the posterior distribution of the parameters when trained on data. I demonstrate this method on Lennard-Jones and Morse potentials fit to triclinic crystal configurations from the OpenKIM database. In particular, IMs are often sloppy, i.e., have likelihood surfaces with long, narrow canyons and broad, flat plateaus. I will be comparing the benefits and drawbacks of the two methods.

Keywords: Interatomic Models, Fisher Information Matrix, Uncertainty Quantification, Markov Chain Monte Carlo, Bayesian Analysis

ACKNOWLEDGMENTS

I would like to thank Dr. Mark Transtrum for advising me. The weekly meeting helped give perspective on this project and clear direction while working on the research. Your review of my final thesis was especially helpful.

I would also like to thank Yonatan Kurniawan, a BYU grad student who really helped me understand this research and answered my questions patiently. He also helped greatly with my various presentations over the years. He is also is the chief architect of the codebase behind this research. I would like to thank Dr. Cody Petrie who also helped me with my presentations and also encouraged me.

I would especially like to thank OpenKIM for the use of their API, IMs, and tests. Their documentation was also very helpful. I would also like to thank the National Science Foundation for funding this research under grant CMMT-1834332.

Contents

Chapter 1

Introduction

1.1 Interatomic Models

Interatomic models (IMs) are used widely in material modeling and simulations to predict material properties. These properties may include total energy, forces, lattice constants, and/or elastic constants. Many IMs were designed for specific conditions and applications. The development of a single IM can take anywhere from a few months to a couple of years and relies on expert intuition, yet these potentials are usually only valid for a particular application of interest. Extending existing IMs to new applications is an active area of research. An important question comes up as we consider extending these IMs to new applications: how much can we trust the predictions of the IMs for applications that they were not originally designed for? This is where the field of uncertainty quantification comes into play, which is the subject of my thesis.

An interatomic model is a function that approximates the potential energy of the nuclei of a collection of atoms based on their positions. Usually an IM is a function of r_{ij} where r_{ij} represents the distance between two atoms. However, as we will see with Stillinger-Weber in Eq. 2.13, more complicated IMs, including those used on tests with more than one element, can be a function of more than just r_{ij}. For example, Eq. 2.13 is the 3-body term that is a function of r_{ij}, r_{ik} and the bond angle between the $i-j$ and $i-k$ bonds.

The way these IMs and tests work together is that the atomic configuration and the interatomic model are brought together to give energy and forces which then undergo simulation to give predictions. The IM evaluates the energy and forces for a given r; r is based on the atomic configuration. The test contains the atomic configuration and simulation. For these research project, we take IMs and tests from the OpenKIM database, which is an interatomic potential repository. OpenKIM provides a framework for IMs and tests to work together. Together, the test and IM is the model.

1.2 Uncertainty Quantification

There are two ways to represent models geometrically: the model manifold in data space and the cost contour in parameter space. We represent models by taking a point in parameter space (this point is a specific combination of all parameters) and map it onto a point in prediction space using the model.

If we take the set of all possible parameter values, we can use the model to map them onto the model manifold in prediction space, i.e. the set of allowed model predictions. The uncertainty in the parameters transfers onto the model predictions. We use cost and sloppiness to determine the uncertainty of the parameters.

1.2.1 Cost

We define a cost function which tells us how far from the data is from the predictions. In our research, it is normally given by,

$$C(\theta) = \frac{1}{2}\sum_{m}\left(\frac{y_m - y_m(\theta)}{\sigma_m}\right)^2. \tag{1.1}$$

In this equation, θ is the parameter that is being evaluated (IMs have multiple parameters), m is the index of the data, y_m is the actually data, $y_m(\theta)$ is the predictions from the model, and σ_m is the inverse weights. The weights, which we took to be 10% of the data, tells us how much each data point contributes to the cost. The cost function can also be written as

$$C(\theta) = \frac{1}{2}\sum_m r_m^2(\theta), \tag{1.2}$$

where r_m is the residuals defined by as

$$r_m(\theta) = \frac{y_m - y(\theta)}{\sigma_m}. \tag{1.3}$$

The cost quantifies uncertainty in parameters which derives from the uncertainty in the original data. We then propagate that uncertainty to new predictions. The cost function has a probabilistic interpretation as the negative log-likelihood,

$$P(y|\theta) \sim \exp\left(-C(\theta)\right). \tag{1.4}$$

For a model with two parameters, we can plot the cost contour directly where the different colors correspond to different costs, see Figure 1.1. Low cost translates to lower uncertainty. We can propagate uncertainty in the original data to the cost contour which can be used to propagate the uncertainty to a new model and its predictions. (For our research, this new model has the same IM but a different test, meaning a different set of data points y_m.) Propagating the uncertainty from the cost contour to the new model is straightforward, but mapping the uncertainty to the cost contour is more complicated. For models with more than two parameters, it is not possible to plot the cost contour directly in respect to all the parameters simultaneously. Thus, we need other methods to get information about the cost contour.

1.2.2 Sloppiness

Sloppiness is a common property of IMs; this complicates uncertainty quantification. When a model is sloppy, a large change in certain parameter combinations, known as sloppy parameters, leads to

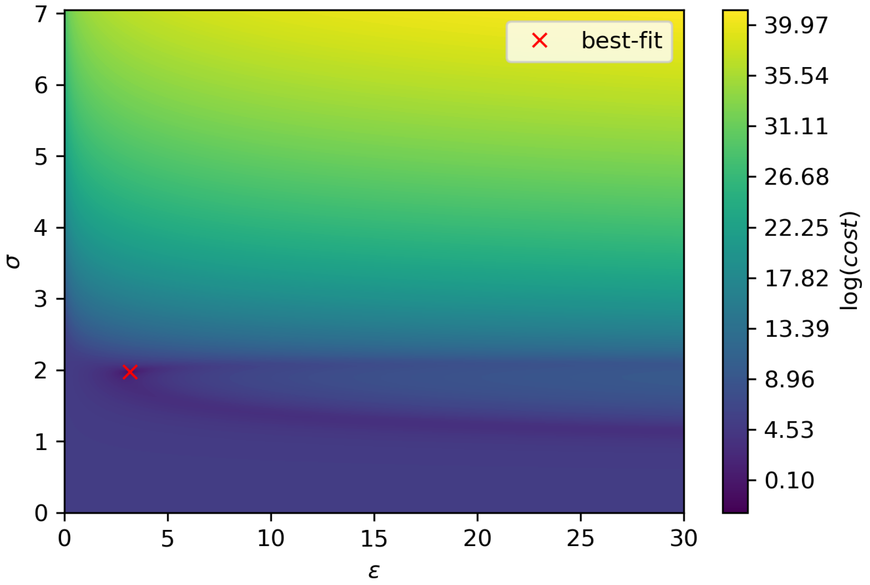

Figure 1.1 Cost contour plot of the Lennard-Jones potential.

only small changes in predictions. For a sloppy model, some parameters may be sloppy while other parameters may be stiff. A stiff parameter is the opposite of a sloppy parameter; for a stiff parameter, a small change in the parameter leads to large changes in the predictions. Additionally, with many models, its not just one or two parameters that are sloppy; there may be many combinations of parameters that are sloppy.

My thesis will focus on two different approaches to uncertainty quantification: the asymptotic approach and the Bayesian approach. The asymptotic approach uses the Fisher Information Matrix (FIM) and the Bayesian approach samples from the Bayesian posterior using Markov Chain Monte Carlo (MCMC) techinques. These methods will give us information about the cost contour which is impossible to plot for IMs with more than 2 parameters. These techniques, along with the IMs and tests that we used, will be discussed in the next section.

Chapter 2

Methods

In this chapter, I will be discussing the methods we used for uncertainty quantification and the IMs and tests that we applied these methods on. The methods we used were the Fisher Information Matrix and Bayesian analysis. The three IMs are Lennard-Jones, Morse, and Stillinger-Weber; the tests corresponding to each IM will also be explained. Lennard-Jones and Morse use a Triclinic PBC energy and forces test while Stillinger-Weber uses a test for monolayer MoS_2.

2.1 Fisher Information Matrix

The first method we use to analyze the uncertainty of an IM is the Fisher Information Matrix (FIM). FIM is calculated by using the Jacobian around the point of interest, usually the best fit. For our FIM analysis, we calculated the Jacobian at the default parameter values. For the local neighborhood around the best fit, we linearize the residuals, from Eq. 1.3:

$$r_m(\theta) \approx r_m(\theta^*) + \frac{\partial r_m}{\partial \theta}(\theta - \theta^*). \tag{2.1}$$

The derivative of the residuals with respect to the parameters is given by the Jacobian matrix,

$$J_{m\mu} = \partial_\mu r_m = \frac{\partial r_m}{\partial \theta_\mu}, \tag{2.2}$$

where μ and m are indices of the Jacobian matrix. The Jacobian matrix contains information about the sensitivity of each prediction to the change in each parameter.

FIM is especially useful for determining the sloppiness of a model. The sloppiness of a model can be quantified by analyzing the quadratic approximation of the cost around the best-fit parameters. The coefficient of the quadratic term of the approximation is given by the matrix of the second derivative of the cost, called the Hessian matrix $H_{\mu\nu}$ [1], where μ and ν are indices of the matrix. In terms of the residuals, the Hessian matrix is given by

$$\begin{aligned} H_{\mu\nu} &= \partial_\mu \partial_\nu C \\ &= \sum_m \partial_\mu r_m \partial_\nu r_m + \sum_m r_m \partial_\mu \partial_\nu r_m \\ &\approx \sum_m \partial_\mu r_m \partial_\nu r_m, \end{aligned} \tag{2.3}$$

assuming that the residuals around the best-fit parameters are small.

Thus, the approximation of the Hessian matrix can be written in terms of the Jacobian matrix as

$$H_{\mu\nu} = \mathcal{I}_{\mu\nu} = \left(J^T J\right)_{\mu\nu}, \tag{2.4}$$

where $\mathcal{I}$ is the Fisher Information Matrix. To understand how the FIM is related to the model manifold locally, consider the singular value decomposition of the Jacobian,

$$J = U \Sigma V^T, \tag{2.5}$$

where U is an $M \times N$ matrix, Σ is an $N \times N$ postive diagonal matrix, and V is an $N \times N$ unitary matrix, satisfying $V^T V = V V^T = 1$. After substituting this singular value decomposition into Eq.

2.4, the FIM can then be written as

$$\begin{aligned}
&\mathcal{I} = V\Sigma^2 V^T, \\
&\mathcal{I} = V\Sigma' U U' \Sigma V' = V\Sigma^2 V', \\
&\text{with } V = \begin{bmatrix} \overrightarrow{v_1} & \overrightarrow{v_2} & \dots & \overrightarrow{v_n} \end{bmatrix} \\
&\text{and } \Sigma^2 = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}.
\end{aligned} \tag{2.6}$$

In this equation, $\{\lambda_i\}_{i=1}^n$ and $\{\sigma_i\}_{i=1}^n$ are the eigenvalues of the FIM and the singular values of the Jacobian, respectively. The eigenvalues of the FIM and the singular values of the Jacobian are related by $\sigma_i = \sqrt{\lambda_i}$. V is a unitary matrix constructed from the orthonormal eigenvectors of $\mathcal{I}$. Eq. (2.6) is a very useful form to express the FIM. This equation shows how a unit ball around a point in the parameter space is mapped onto an ellipsoid in the prediction space, see Figure 2.1. The orientation of the ellipsoid is described by matrix V, which consists of the eigenvectors of the FIM, with the semi-major axis of the ellipsoid is aligned with the first eigenvector, $\vec{v}_1$, corresponding to the largest eigenvalue, λ_1. The ratio of the length of the semi-major axis to the length of the semi-minor axis is proportional to the ratio of the largest eigenvalue to the smallest eigenvalue of $\mathcal{I}$ (see Figure 2.1).

We can get even more information about the local geometry of the cost contour around the best fit point as shown by Figure 2.2. The eigenvectors and eigenvalues tell us the shape and direction of the cost contour around the point of interest in parameter space.

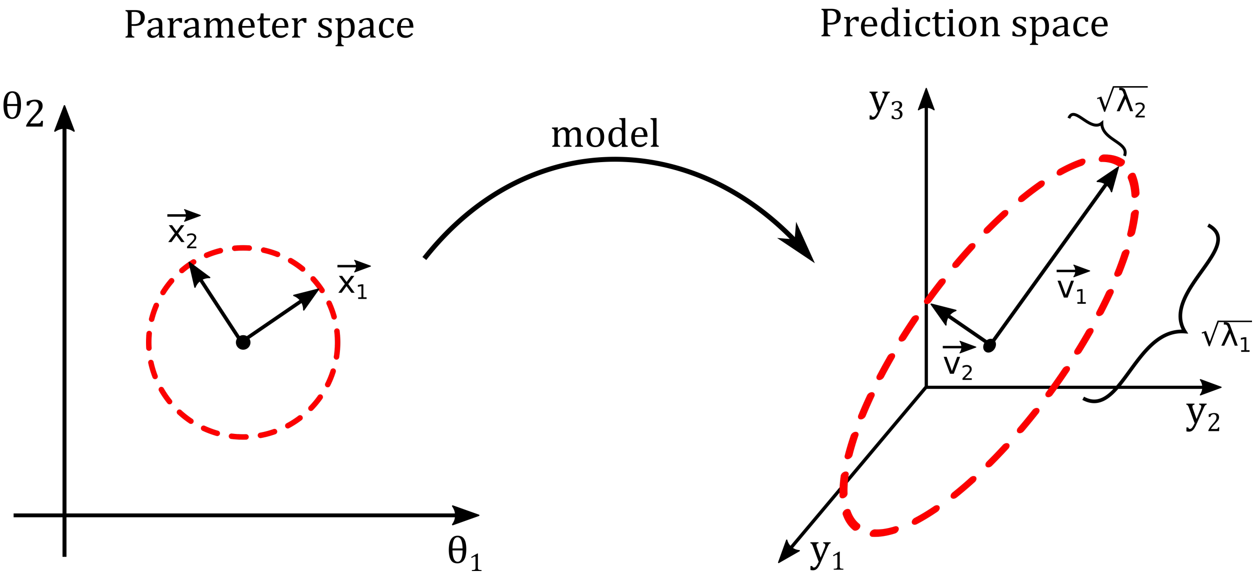

Figure 2.1 Unit ball in parameter space mapped onto an ellipsoid in prediction space. FIM tells us how the unit ball in parameter space is mapped onto an ellipsoid in the prediction space. The unit ball is stretched and compressed in the direction along the eigenvectors of FIM with magnitude of compression proportional to the eigenvalues of the FIM.

2.2 Bayesian analysis

Another common approach to uncertainty quantification in parameter space is using a Bayesian framework. In Bayesian statistics, the parameter uncertainty is described by a posterior distribution given by Bayes' theorem.

$$P(\theta|\mathbf{y}) \propto L(\theta|\mathbf{y}) \cdot \pi(\theta), \tag{2.7}$$

where $L(\theta|\mathbf{y})$ and $\pi(\theta)$ are the likelihood and the prior distribution of the model's parameters, respectively [2,3].

Additionally,

$$L(\theta|\mathbf{y}) \propto \exp(-C(\theta)/T), \tag{2.8}$$

where T is a "sampling temperature" as in [4]. The cost is analogous to the internal energy of a system, so low temperature distributions are concentrated near the low-energy (i.e., low cost) region of parameter space. Formally, the temperature uniformly scales the tolerances σ_m in Eq. (1.3).

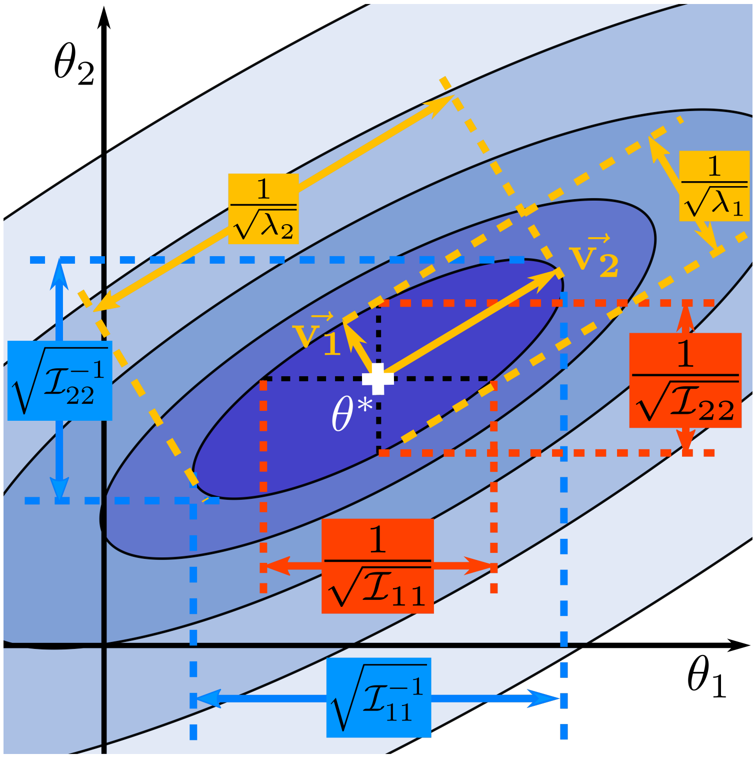
θ_2
θ_1
$\frac{1}{\sqrt{\lambda_2}}$
$\frac{1}{\sqrt{\lambda_1}}$
$\vec{v}_2$
$\vec{v}_1$
θ^*
$\sqrt{\mathcal{I}^{-1}_{22}}$
$\frac{1}{\sqrt{\mathcal{I}_{22}}}$
$\frac{1}{\sqrt{\mathcal{I}_{11}}}$
$\sqrt{\mathcal{I}^{-1}_{11}}$

Figure 2.2 Geometry of the Fisher Information Matrix (FIM). $\mathcal{I}$ is the FIM and λ is the eigenvalue of $\mathcal{I}$ with $\vec{v}$ being the corresponding eigenvector. $\mathcal{I}^{-1}$ is the inverse FIM. These quantities tell us the shape and direction of the cost contour around the point of interest in parameter space. $\vec{v}$ tells us the direction of the cost contour while λ tells us the corresponding length. The conditional variance is 1 over the square root of the diagonal of $\mathcal{I}$ which tells us how the cost contour changes by varying one parameter and holding the others fixed. The first diagonal element corresponds to θ_1, parameter 1, and so on. The variance of θ_1 is given by the square root of the first diagonal of the inverse FIM.

By sampling at multiple temperatures, we can easily assess how the choice of σ_m affects any conclusions we draw from the distribution.

The likelihood is functionally the same as the probability distribution, $L(\theta|\mathbf{y}) = P(\mathbf{y}|\theta)$, so Eq. 2.7 could be rewritten as,

$$P(\theta|\mathbf{y}) \propto P(\mathbf{y}|\theta) \cdot \pi(\theta) . \tag{2.9}$$

Here, θ represents the parameter, and $\mathbf{y}$ represents the data. The posterior distribution, $P(\theta|\mathbf{y})$, represents the probability of getting a certain parameter given the data. The likelihood, $L(\theta|\mathbf{y})$ or $P(\mathbf{y}|\theta)$, is the probability of getting that data given that the certain parameter is chosen. The prior, $\pi(\theta)$, is the probability of getting a certain parameter out of all possible parameters; as discussed later on in this paper, the choice of prior has important ramifications on the MCMC calculation.

Markov Chain Monte Carlo (MCMC) is an algorithm that is used to sample from the posterior. As one may expect, running MCMC depends on defining the posterior that the MCMC calculation will be sampling. Defining the posterior distribution requires making some nontrivial decisions that can affect the outcome of the calculations. These decisions include using bare parameters vs log parameters, the choice of prior, and the effect of temperature, described by Eq. 2.8.

The first decision, deciding whether to use bare parameters (parameters in normal scale) or log parameters (parameters in log scale), is significant because it affects the prior probability density of the parameters. For example, a parameter that is close to one would have a higher chance of being selected in log scale compared to a parameter that has a number that is much bigger than one. Additionally, parameters that are negative cannot be represented at all in log scale—though a solution could be to first take the absolute value of each parameter before converting to log scale.

The choice of prior, $\pi(\theta)$, is another important decision that affects the probability of parameters. The problem is that there is no simple way to choose a good prior. We choose to use a flat prior for all of our calculations; a flat prior means that all parameters within specified boundaries are equally likely to be chosen while parameters outside of those boundaries are rejected. Using a flat prior simplifies the process of choosing a prior, but it still requires making a good decision on the boundaries. Choosing too small of a range can lead to inaccurate results, but choosing too wide a range can affect speed of convergence.

2.3 Interatomic potentials and tests

We choose three different IMs to illustrate these methods of uncertainty quantification. The first IM we choose was the Lennard-Jones potential for silicon. The Lennard-Jones potential is a potential between pairs of atoms; this IM only has two parameters, ε and σ, as seen below:

$$\begin{aligned} V(r_{ij}) &= 4\varepsilon\left(\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right) + \Delta, \\ \text{with } \Delta &= -4\varepsilon\left(\left(\frac{\sigma}{r_{\text{cut}}}\right)^{12} - \left(\frac{\sigma}{r_{\text{cut}}}\right)^{6}\right). \end{aligned} \tag{2.10}$$

The potential is a function of r_{ij} where r_{ij} is distance between atoms i and j in the configuration. The parameter ε is an energy scaling factor in the potential, and σ is associated with the equilibrium distance of the pair interaction. Δ is chosen so that the potential goes to zero at the cutoff radius, r_{cut}.

For our chosen test, the default parameter values are $\varepsilon = 3.17431$ eV and $\sigma = 1.9778$ Å [5–9]. The test we chose for this potential was Triclinic PBC energy and forces [10]. A triclinic PBC is a crystal lattice structure similar to a body centered cubic lattice; however, in the triclinic PBC, the 3 side lengths corresponding to the height, width, and depth of the shape are of different lengths so that the shape is no longer a perfect cube and none of the angles between the sides are a perfect 90°. We use the pair potentials to predict the unrelaxed energy and forces of the atoms of a silicon system in a body-centered triclinic configuration.

The Morse potential, the second IM we used, is also a pair-wise function but with three parameters, ε, C, and r_0,

$$\begin{aligned} V(r_{ij}) &= \varepsilon\left(-e^{-2C(r_{ij}-r_0)} + 2e^{-C(r_{ij}-r_0)}\right) + \Delta, \\ \text{with } \Delta &= -\varepsilon\left(-e^{-2C(r_{\text{cut}}-r_0)} + 2e^{-C(r_{\text{cut}}-r_0)}\right). \end{aligned} \tag{2.11}$$

Similar to the Lennard-Jones potential, ε is also an energy scaling factor in the potential. The r_0 parameter shows the equilibrium distance of the pair interaction; C controls the width of the potential well. Again, Δ is chosen so that the potential is zero at the cut-off radius, r_{cut}; we only change the ε, r_0, and C parameters. The default parameters that we use are $\varepsilon = -0.4205$ eV, $C = 1.4199$ Å^{-1}, and $r_0 = 2.78$ Å [11–13]. We also used a Triclinic PBC energy and forces test for Morse but used nickel instead of silicon. We can see a graph of the Lennard-Jones and Morse potentials in Figure 2.3.

The third IM we studied was Stillinger-Weber for monolayer MoS_2. This potential contains both pair-wise interactions and 3-body interactions. The 2-body interaction is

$$\phi_2\left(r_{ij}\right) = A_{IJ}\left(B_{IJ}\left(\frac{r_{ij}}{\sigma_{IJ}}\right)^{-p_{IJ}} - \left(\frac{r_{ij}}{\sigma_{IJ}}\right)^{-q_{IJ}}\right)\exp\left(\frac{\sigma_{IJ}}{r_{ij}-r_{IJ}^{\text{cut}}}\right), \tag{2.12}$$

where the uppercase subscripts denote the types of atom, i.e. A_{IJ} is the parameter A corresponding to interaction between atoms type I and J. The 3-body term is given by

$$\phi_3\left(r_{ij}, r_{ik}, \beta_{jik}\right) = \lambda_{JIK}\left(\cos\beta_{jik} - \cos\beta_{JIK}^0\right)^2 \exp\left(\frac{\gamma_{IJ}}{r_{ij}-r_{IJ}^{\text{cut}}} + \frac{\gamma_{IK}}{r_{ik}-r_{IK}^{\text{cut}}}\right), \tag{2.13}$$

with β_{jik} be the bond angle between the $i-j$ and $i-k$ bonds.

We use this potential to predict the forces of the atoms in configurations around the equilibrium state. More detailed information of the fitting process of this potential are explained in [14]. Following the original paper, we set $q_{IJ} = 0$ and let γ be the same for all types of interaction.

We apply non-unity weighting factors in our cost function, Eq. 1.1. Since the fitting data are the forces around the equilibrium state, then we want the magnitudes of the force to be close to zero. Thus, we put larger weights on the force data that are closer to zero. We achieve this by setting

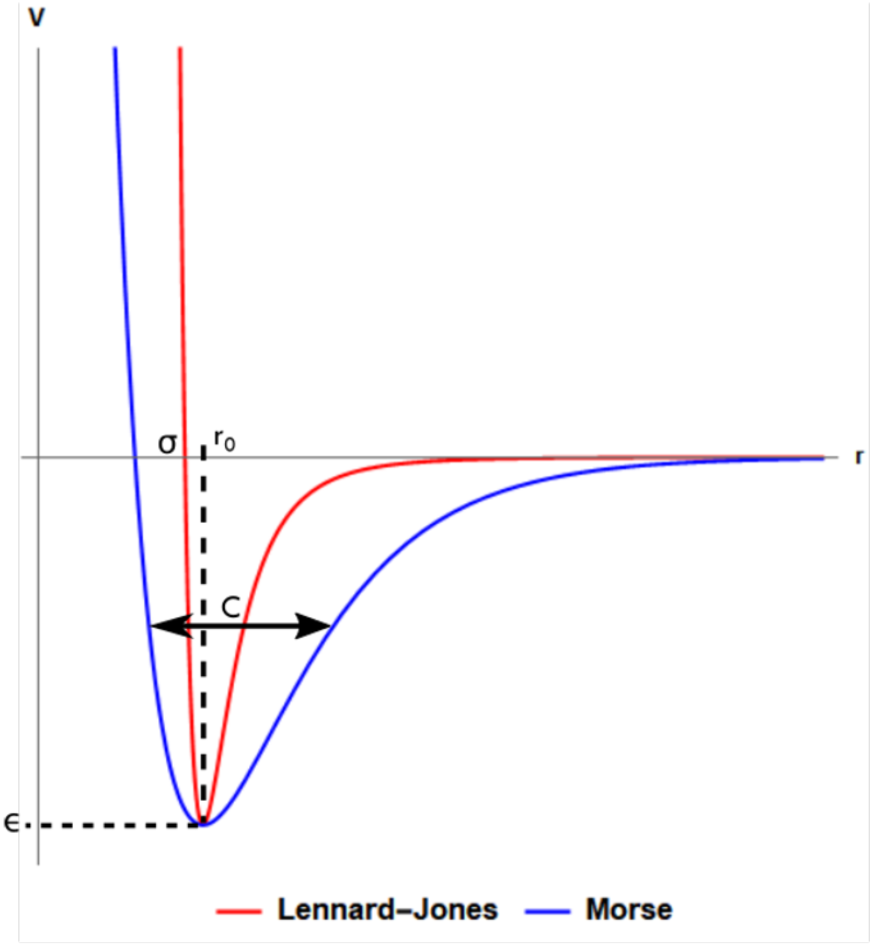

Figure 2.3 Comparative graph of the Lennard-Jones and Morse potentials.

the error tolerance of each prediction to be 10% of the magnitude of the force data acting on the corresponding atom. Additionally, we allow parameters p's to take any non-integer, positive value. We also remove the relation between σ's and the equilibrium lattice constants of the system. Thus, we don't require the force of each type of interaction to be zero, and we remove the constraint on parameters B's. Using these conditions, we consider parameters A, B, p, and σ for each type of pair-wise interaction ($Mo-Mo$, $Mo-S$ and $S-S$ interactions). Then for the 3-body interaction, we consider parameters λ, for $S-Mo-S$ and $Mo-S-Mo$ interactions, and γ. Applying these conditions and setup gives us a new set of default parameters. The values of these parameters are listed in Table 2.1 for the two-body interaction term and Table 2.2 for the three-body interaction term.

	Interaction		
Parameter	Mo-Mo	Mo-S	S-S
A (eV)	18.4310060	8.83861305	0.37463396
B	0.00641786	1.04793603	561.429270
p	4.73717813	8.26621744	2.66196913
σ(Å)	6.16940454	1.92967991	0.41904814

Table 2.1 Parameters of the 2-body term in the Stillinger-Weber fitted to the MoS_2 test.

Parameter	Value
λ_{S-Mo-S}	4.28784076 eV
$\lambda_{Mo-S-Mo}$	14.4285026 eV
γ	1.53800500 Å

Table 2.2 Parameters of the 3-body term in the Stillinger-Weber fitted to the MoS_2 test.

Chapter 3

Results and Conclusions

3.1 Fisher Information Matrix

From the plot of the eigenvectors for Lennard-Jones in Figure 3.1, we can see how the eigenvectors map onto the cost contour of Lennard-Jones. We plotted the eigenvectors of Lennard-Jones with the length of the arrow proportional to $1/\sqrt{\lambda}$ where λ is the corresponding eigenvalue in Figure 3.1. The longer vector represents the sloppy parameter direction while the shorter arrow represents the stiff direction. These eigenvectors were plotted against the cost contour zoomed in around the best fit point (see Figure 3.1). As we can see from Figure 3.1, the eigenvectors for Lennard-Jones show the direction of the cost contour.

One primary use of FIM is to evaluate the sloppiness of a model. The more parameters a model has, the more useful this analysis can be. The high eigenvalues represent stiff parameter directions while low eigenvalues represent sloppy combinations; these parameter combinations are in the direction of their respective eigenvectors. The eigenvalues are dimensionless.

In Figure 3.2, we can see the eigenvalues plotted for six different IMs and tests. The first three IMs are the three that we focused our analysis on: Lennard-Jones, Morse, and Stillinger-Weber

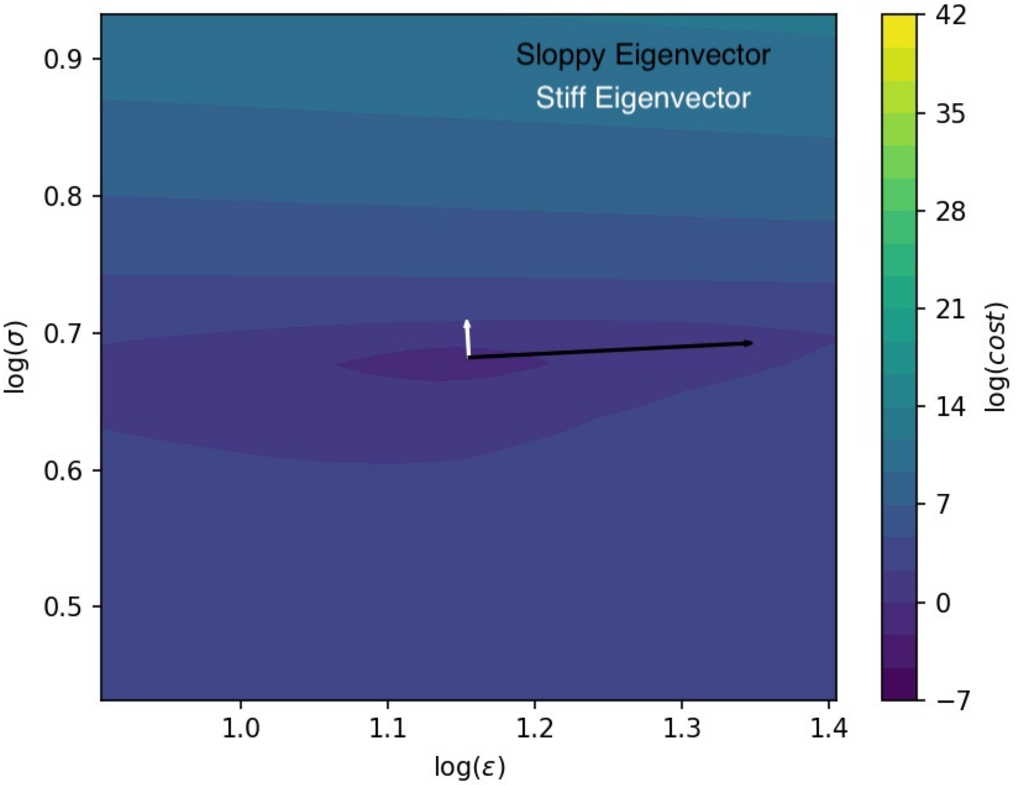

Figure 3.1 FIM results for Lennard-Jones plotted along the cost contour. The longer vector represents the sloppy direction while the short vector represents the stiff direction.

for MoS_2. The last three IMs are included for additional comparison and analysis of this method. These IMs are three-body bond order, environment dependent interatomic potential (EDIP), and Stillinger-Weber for a silicon crystal. (Stillinger-Weber for a silicon crystal doesn't have as many parameters because it doesn't have the three-body interaction that the Stillinger-Weber model we normally use has.) As we can see from Figure 3.2, the more parameters an IM has, the easier it is to tell which eigenvalues are sloppy and which are stiff. Just looking at Lennard-Jones, the difference between the highest and the lowest eigenvalues may seem like a lot, but comparing it to the other IMs with more parameters, it doesn't seem like a large range. The graph is in log scale to better compare eigenvalues.

As we can see in Figure 3.2, the eigenvalues can vary greatly in their magnitude. The high eigenvalues represent stiff parameter directions of the IM in the direction of their respective eigenvector while the low eigenvalues represent sloppy parameter combinations in the direction of their respective eigenvectors. However, it is not the actual value of the eigenvalue that matters as much as the value of the eigenvalue in respect to the other eigenvalues. For example, the highest eigenvalue for Three-Body Bond Order is lower than the highest eigenvalue for Lennard-Jones or Morse, but we know that the eigenvalue is in the stiff parameter direction because it is much larger than any of the other eigenvalues. We included these additional IMs in our dicussion to show how sloppiness is inherit in IMs; it's not just the IMs that we choose that are sloppy.

As mentioned above, the eigenvectors along with their respective eigenvalues can give information about how a unit sphere or hyper-sphere can be mapped from parameter space to an ellipsoid or hyper-ellipse in prediction space. The resulting hyper-ellipse is compressed a lot in stiff directions and stretched a lot in sloppy directions. This leads to large uncertainty, especially in the sloppy directions. Thus, the low eigenvalues correspond with higher uncertainties.

One major advantage of FIM is that it is fast computationally, especially compared to running a MCMC calculation. Accordingly, it is recommended that FIM analysis should be the first step in

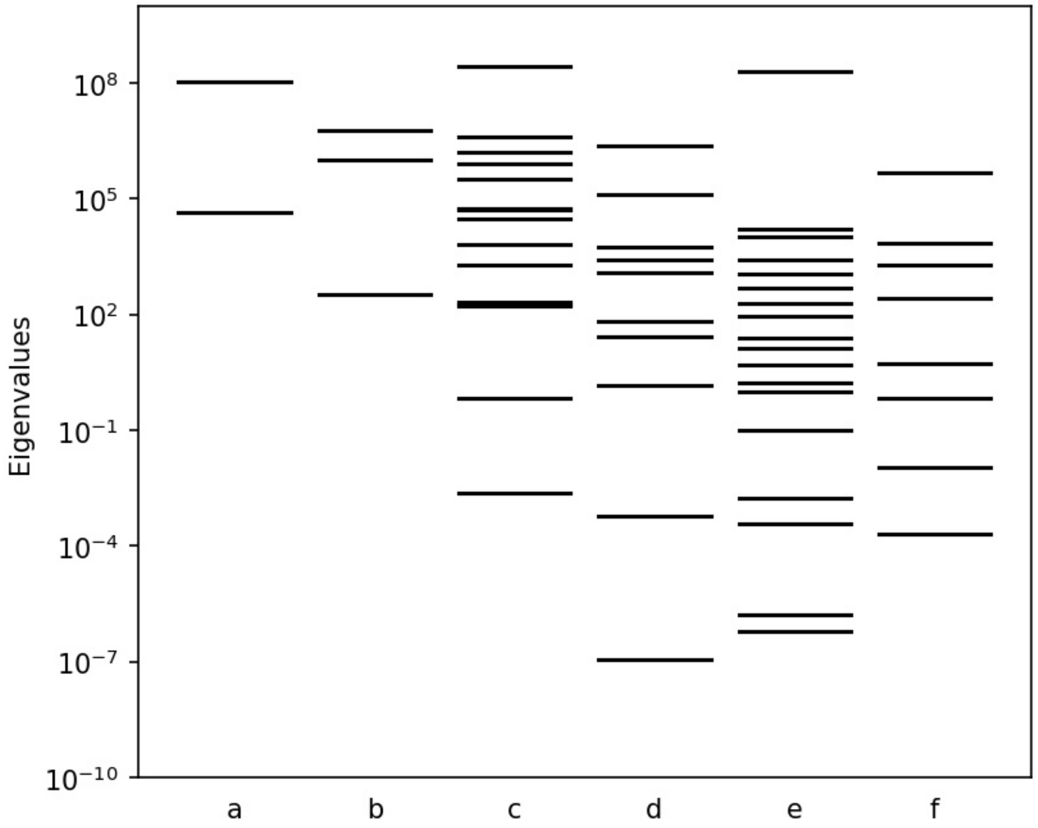

Figure 3.2 Eigenvalues from the FIM of a few IMs. The high eigenvalues represent stiff parameters of the IM in the direction of their respective eigenvectors. The low eigenvalues represent sloppy parameters in the direction of their respective eigenvectors. (a) Lennard-Jones, (b) Morse, (c) Stillinger-Weber with MoS_2, (d) Three-Body Bond Order, (e) Environment Dependent Interatomic Potential, (f) Stillinger-Weber for Si crystal.

analyzing the uncertainty of a model. However, the primary limitation of FIM analysis is that it only gives information about the cost contour locally, around the point of interest; it does not give any information about the cost contour globally. This is a result of FIM being a linear approximation.

3.2 Bayesian Posterior

3.2.1 Lennard-Jones

Since Lennard-Jones has only two parameters, we can see the results of the MCMC calculation plotted directly on top of the cost contour, see Figure 3.3. We can see how the MCMC points, plotted in black, follow the canyon of lowest cost on the cost contour. (The best fit point is the default parameters that this model was fitted with.) We can also see how the cost contour stretches until the end of the graph which is something that we cannot get from FIM alone. Additionally, we see that Lennard-Jones gets very sloppy as we move along the ε axis. We can see that the global approach of MCMC gives a better idea of the shape of the cost contour away from the best fit point.

One important choice discussed earlier is the decision to use log parameters vs. bare parameters. Comparing Figure 3.3 with Figure 3.4, we can see that for bare parameters, there is a greater concentration of regions with a high cost, i.e. regions that are yellow in the figure, in the cost contour compared to the cost contour with log parameters. On the other hand, for the cost contour with log parameters has a greater concentration of regions with lower cost—but not the lowest cost; these are the blue regions of the figure. Additionally, the bare parameters has a larger region with the minimum cost. This illustrates how the region that the MCMC calculation is sampling from depends on the choice of parameterization. As we can see from Figure 3.3 and Figure 3.4, the resulting MCMC points are different.

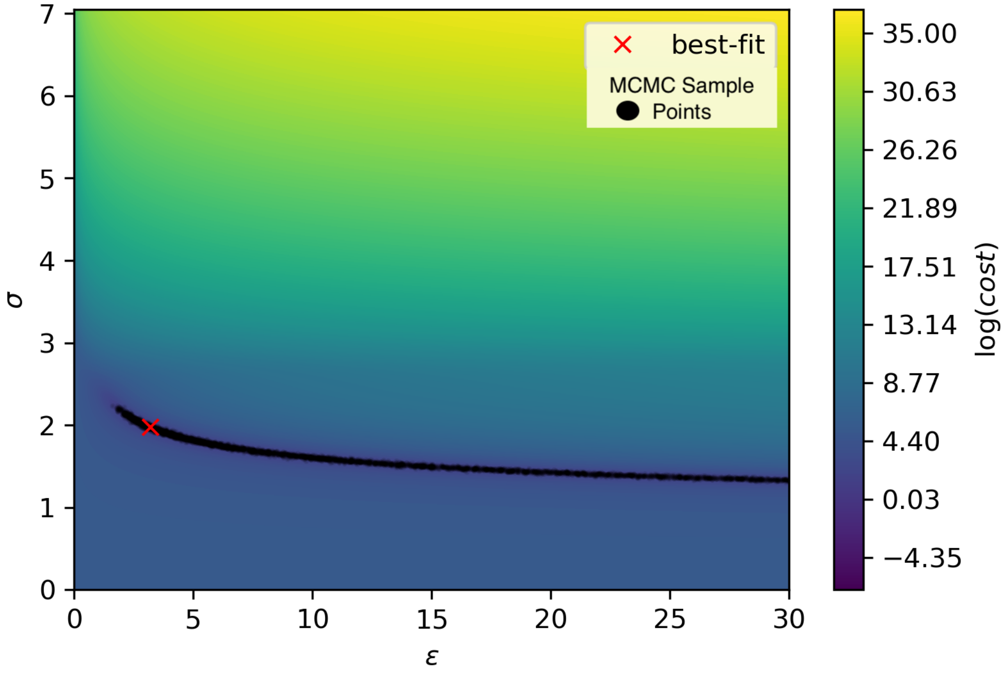

Figure 3.3 MCMC results for Lennard-Jones plotted against the respective cost contour. The best fit point is the parameter values with the lowest cost.

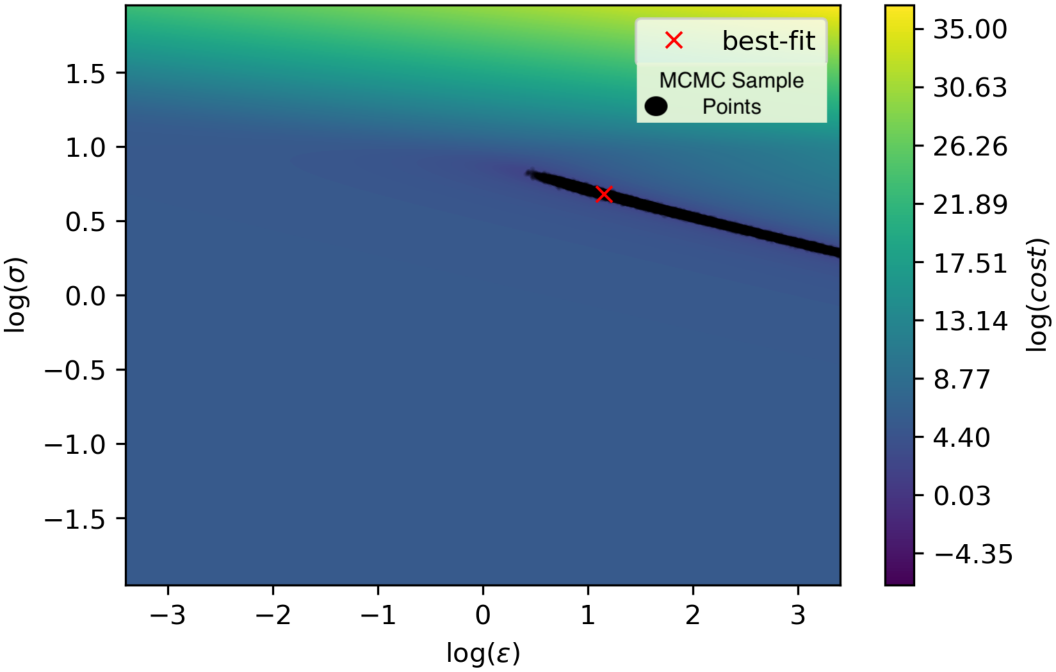

Figure 3.4 MCMC results for Lennard-Jones plotted against the respective cost contour in log scale. The best fit point is the parameter values with the lowest cost.

3.2.2 Morse

With the Morse potential, we are unable to plot the MCMC results directly against the cost contour because it has more than two parameters. However, we can see the different parameters plotted against each other in Figure 3.5. In this figure, we plotted the MCMC sample points onto the two-dimensional projections of the parameters. We also plotted the histograms for the parameters. By performing statistical analysis on the histograms, we can analyze the uncertainty of the parameters. Additionally, since Morse only has three parameters, we can see the MCMC points plotted in three dimensions in Figure 3.6 along with the projection onto each of the three planes of the parameters.

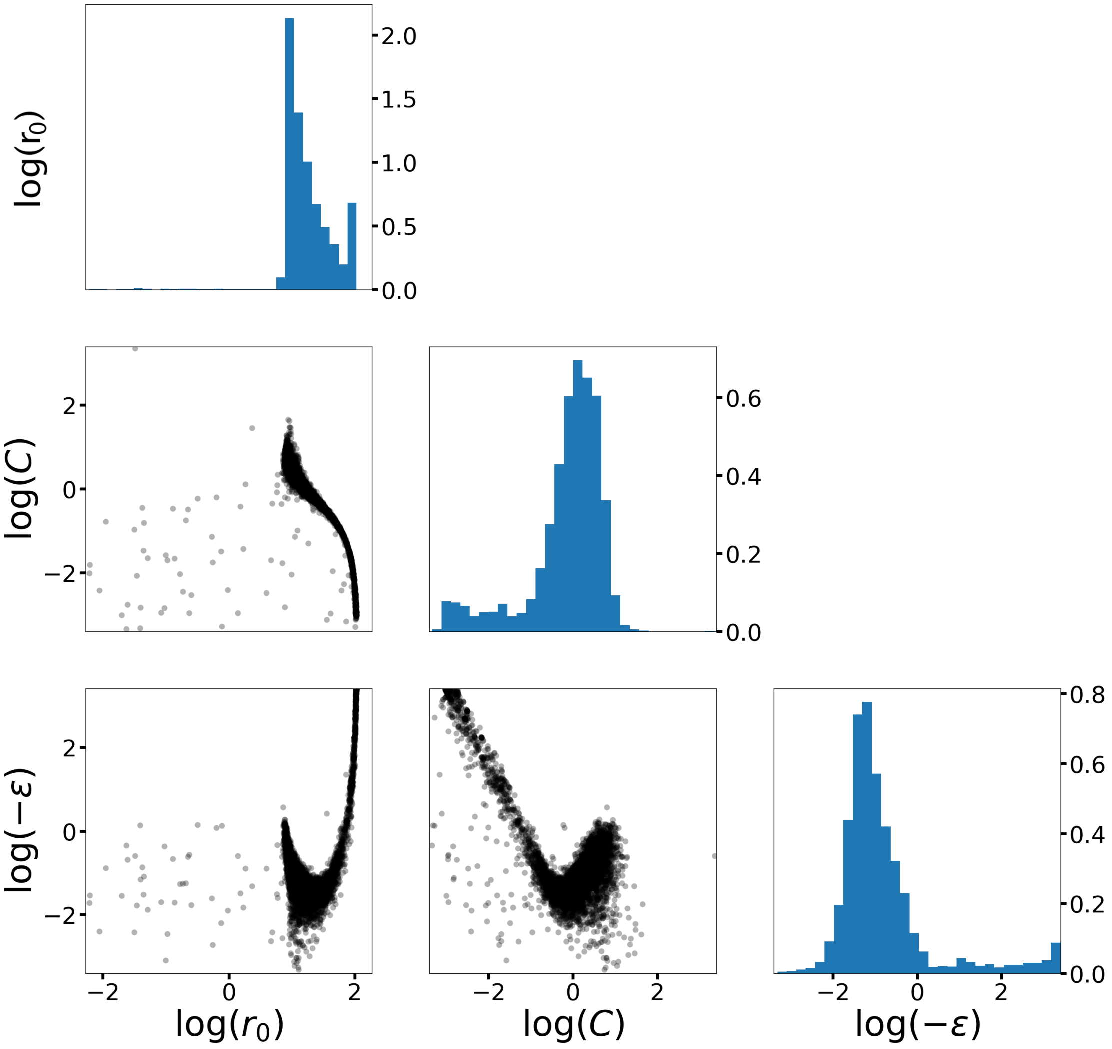

Figure 3.5 MCMC results for Morse with histograms. The scatter plot represents the MCMC sample points projected onto the two dimensional projection of parameters. The histogram shows the distribution of the sample points for a given parameter.

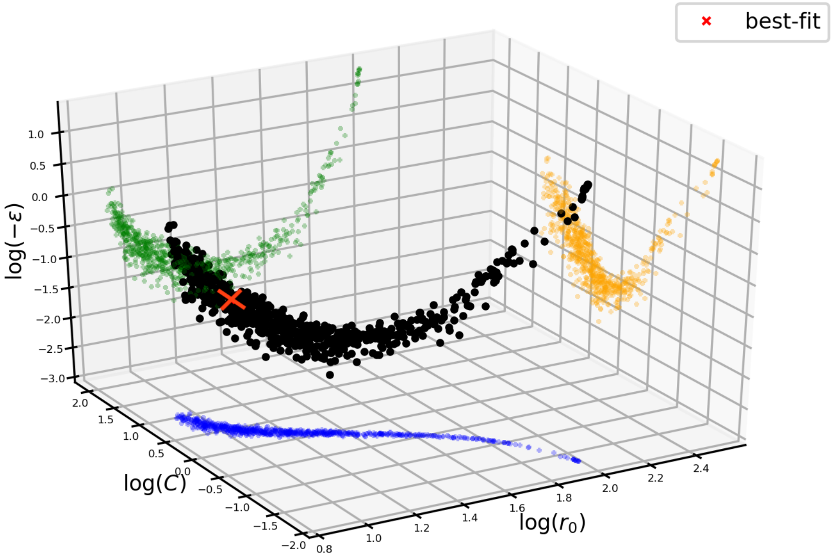

Figure 3.6 MCMC results for Morse in three dimensions. The different colors represent projections onto the planes: green is the projection onto the $\log(-\varepsilon) - \log(r_0)$ plane; blue is the projection onto the $\log(C) - \log(r_0)$ plane; yellow is the projection onto the $\log(-\varepsilon) - \log(C)$ plane. The black dots represent the MCMC samples in three dimensional space.

3.2.3 Stillinger-Weber

We plotted the MCMC results for Stillinger-Weber in Figure 3.7; this plot is a subsection of the complete figure with all 15 parameters (the complete figure has 120 subgraphs compared to the 10 subgraphs in Figure 3.7). This subsection was chosen because it illustrates parameter evaporation. Parameter evaporation is when the MCMC sample points run off to infinity (or reach the boundaries of the prior, $\pi(\theta)$). From Figure 3.7, we can see that some parameters evaporate while others do not. For example, parameters $\log(B_{Mo-Mo})$ and $\log(p_{Mo-Mo})$ both evaporate in these graphs; this is shown by how the sample points run off to either one or both boundaries of the graph. The boundaries of the graph are chosen so that they reflect $\pi(\theta)$. Conversely, parameters $\log(\sigma_{Mo-Mo})$ and $\log(A_{Mo-Mo})$ don't evaporate. The histograms of the parameters also show parameter evaporation. Histograms that spread out across the range of possible parameters (the prior, $\pi(\theta)$) show parameter evaporation, such as the histograms for parameters $\log(B_{Mo-Mo})$ and $\log(p_{Mo-Mo})$. Histograms that are more narrow, on the other hand, show parameters that don't evaporate, like the histograms for parameters $\log(\sigma_{Mo-Mo})$ and $\log(A_{Mo-Mo})$.

Parameter evaporation is the result of sloppiness in a model. Parameter evaporation makes uncertainty quantification difficult because a parameter that evaporates has infinite uncertainty. We can also see examples of parameter evaporation in the plots for Lennard-Jones and Morse, see Figures 3.3, 3.4, and 3.5. However, those plots have parameters that only evaporate in one direction.

3.2.4 Advantages and disadvantages

The primary benefit of using a Bayesian analysis to uncertainty quantification of IMs is that it is a global method. It can be used to see the shape of the cost contour far away from the best fit point. However, running a MCMC calculation takes a long time due to the difficulty of convergence. This issue of convergence is partially due to the fact that a Bayesian analysis depends heavily on the choice of prior, $\pi(\theta)$, and the choice of using bare parameters vs. log parameters. A

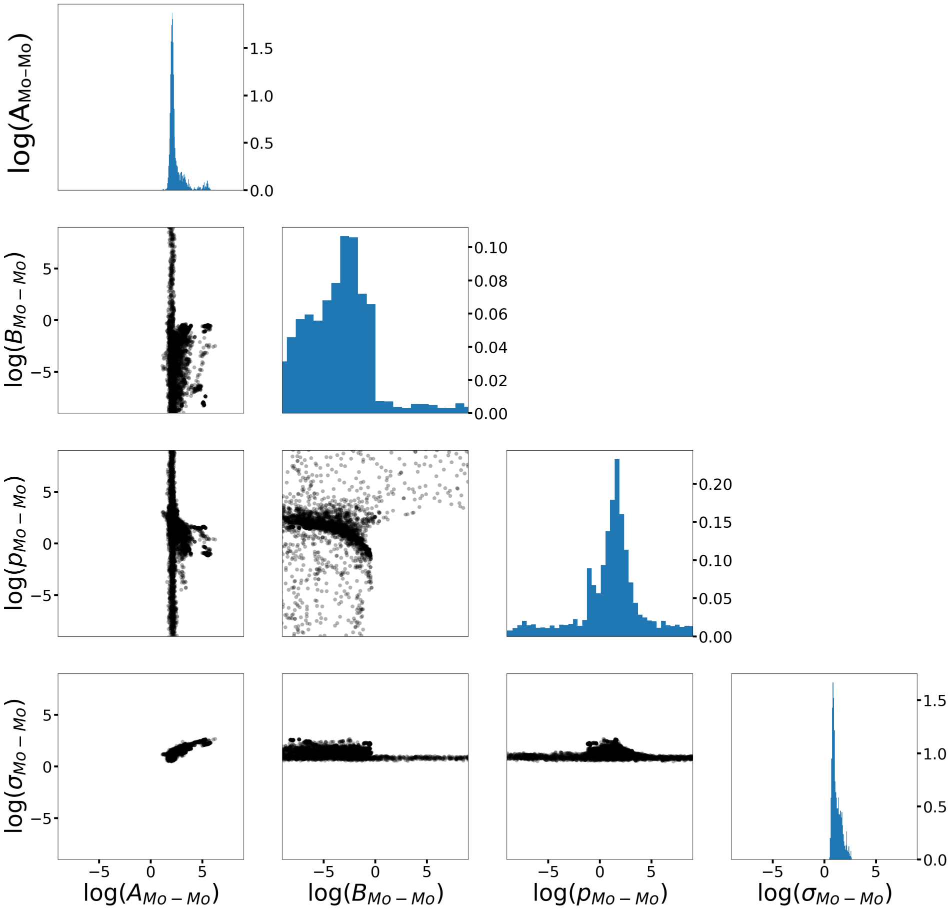

Figure 3.7 MCMC results for Stillinger-Weber with histograms. The scatter plot represents the MCMC sample points projected onto the two dimensional projection of parameters. The histogram shows the distribution of the sample points for a given parameter.

MCMC calculation can show parameter evaporation which can give additional information about the sloppiness of a model. This issues are the result of the inherent sloppy nature of IMs.

3.3 Conclusion

3.3.1 Summary

The cost contour, given by Eq. 1.1, can give useful information about how the parameters, θ, of an IM propagate uncertainty. However, the cost contour can't be plotted directly in respect to all the parameters simultaneously for IMs with more than two parameters. FIM and Bayesian anlysis are methods that are useful for extracting information about the cost contour for IMs with a higher number of parameters.

3.3.2 Future Work and Applications

Sloppiness is a common property of IMs. When a model is sloppy, a large change in certain parameter combinations, known as sloppy parameters, leads to only small changes in predictions. Sloppiness makes uncertainty quantification difficult. This is especially true when using a Bayesian approach to uncertainty quantification. However, there are other methods that can be used for uncertainty quantification. These methods include likelihood profile and information geometry. Our future work includes applying these other methods to the same IMs and tests that we have already looked at and applying FIM and MCMC to new IMs and tests.

List of Figures

Bibliography

[1] B. K. Mannakee, A. P. Ragsdale, M. K. Transtrum, and R. N. Gutenkunst, in *Uncertainty in Biology*, Vol. 17 of *Studies in Mechanobiology, Tissue Engineering and Biomaterials*, L. Geris and D. Gomez-Cabrero, eds., (Springer International Publishing, 2016), p. 271–299.

[2] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, in *Numerical recipes*, 3rd ed. ed., W. H. Press, ed., (Cambridge University Press, Cambridge, UK; New York, 2007).

[3] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice* (Chapman and Hall/CRC, 1995).

[4] S. L. Frederiksen, K. W. Jacobsen, K. S. Brown, and J. P. Sethna, “Bayesian Ensemble Approach to Error Estimation of Interatomic Potentials,” Physical Review Letters **93,** 165501 (2004).

[5] R. S. Elliott, “Efficient ’universal’ shifted Lennard-Jones model for all KIM API supported species developed by Elliott and Akerson (2015) v003,”, OpenKIM, https://doi.org/10.25950/962b4967, 2018.

[6] R. S. Elliott, “Efficient multi-species Lennard-Jones model with truncated or shifted cutoff v003,”, OpenKIM, https://doi.org/10.25950/962b4967, 2018.

[7] J. E. Jones, “On the Determination of Molecular Fields. I. From the Variation of the Viscosity of a Gas with Temperature,” Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **106,** 441–462 (1924).

[8] J. E. Jones, “On the Determination of Molecular Fields. II. From the Equation of State of a Gas,” Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **106,** 463–477 (1924).

[9] J. E. Lennard-Jones, “On the Forces between Atoms and Ions,” Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **109,** 584–497 (1925).

[10] D. S. Karls, “Potential energy and atomic forces of periodic, non-orthogonal cell of Si atoms in a perturbed bcc structure v003,”, OpenKIM, https://openkim.org/cite/TE_006970922000_003, 2019.

[11] R. S. Elliott, “Morse potential (shifted) for Ni by Girifalco and Weizer (1959) using a high-accuracy cutoff distance v004,”, OpenKIM, https://doi.org/10.25950/fc55a3bb, 2020.

[12] R. S. Elliott and Y. Afshar, “Morse pair potential shifted to zero energy at cutoff separation v004,”, OpenKIM, https://doi.org/10.25950/fc55a3bb, 2020.

[13] L. A. Girifalco and V. G. Weizer, “Application of the Morse Potential Function to Cubic Metals,” Physical Review **114,** 687–690 (1959).

[14] M. Wen, S. N. Shirodkar, P. Plecháč, E. Kaxiras, R. S. Elliott, and E. B. Tadmor, “A force-matching Stillinger-Weber potential for MoS_2: Parameterization and Fisher information theory based sensitivity analysis,” Journal of Applied Physics **122,** 244301 (2017).

Index