Data-driven estimation of atomistic support for continuum stress using the Gaussian mixture model

Sean J. Moran[1,*] and **Manfred H. Ulz**[2]

[1] Institute for Language, Cognition and Computation, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom

[2] Institute for Strength of Materials, Graz University of Technology, Kopernikusgasse 24/I, 8010 Graz, Austria

The notion of stress being an inherent continuum concept has been a matter of discussion at the atomistic level. The atomistic stress measure at a given spatial position contains a space averaging volume over nearby atoms to provide an averaged macroscopic stress measure. Previous work on atomistic stress measures introduce the characteristic length as an a priori given parameter. In this contribution we learn the characteristic length directly from the atomistic data itself. Central to our proposed approach is the grouping of atoms with highly similar values of position and stress into the same atomistic sub-population. We hypothesise that atoms with similar values for position and stress are those atoms which harbour the greatest influence over each other and therefore should be contained within the same space averaging volume. Consequently the characteristic length can be computed directly from the discovered sub-populations by averaging over the maximum extent of each sub-population. We motivate the Gaussian mixture model (GMM) as a principled probabilistic method of estimating the similarity between atoms within position-stress space. The GMM parameters are learnt from the atomistic data using the Expectation Maximization (EM) algorithm. To form a parsimonious representation of the dataset we regularise our model using the Bayesian Information Criterion (BIC) which maintains a balance between too few and too many atomistic sub-populations. We use the GMM to segment the atoms into homogeneous sub-populations based on the probability of each atom belonging to a particular sub-population. Thorough evaluation is conducted on a numerical example of an edge dislocation in a single crystal. We derive estimates of the space averaging volume which are in very close agreement to the corresponding analytical solution.

1 Atomistic stress

The state of stress at a material point is an important continuum concept. There is a wide body of existing work that defines an atomistic stress based upon kinematics and kinetics at the microscale, see [1] for a comprehensive review. The seminal work of Irving and Kirkwood gave pointwise definitions of continuum quantities. The limitations of this approach were resolved by using a kernel function with wider support, resulting in the Hardy stress at a continuum point $\mathbf{y}$:

$$\boldsymbol{\sigma}(\mathbf{y}) \quad = \quad -\sum_{i=1}^{N} m^i \mathbf{u}^i \otimes \mathbf{u}^i \psi(\mathbf{y}-\mathbf{y}^i) - \frac{1}{2!} \sum_{i,j=1,j\neq i}^{N} \mathbf{f}^{ij} \otimes \mathbf{y}^{ij} B^{ij}(\mathbf{y}) - \dots \tag{1}$$

Here, m^i represents the mass of atom i, $\mathbf{u}^i$ the velocity of atom i relative to the mean velocity of the system, $\mathbf{y}^{ij} = \mathbf{y}^i - \mathbf{y}^j$, $\mathbf{y}^i$ is the position vector of atom i, $\mathbf{f}^{ij}$ gives the force on atom i due to its pair interaction with atom j, the kernel function $\psi(\mathbf{y}-\mathbf{y}^i)$ defines a space averaging volume within which the stress of the contained atoms are aggregated. The bond function $B^{ij}(\mathbf{y})$ between atoms i and j is defined in Eq. 1 as $B^{ij}(\mathbf{y}) = \int_0^1 \psi(\mathbf{y} - \mathbf{y}^i + \lambda \mathbf{y}^{ij})\, d\lambda$. The kernel bandwidth depends on the physics about the continuum point and is crucial for the correct estimation of continuum stress.

2 Data-driven estimation of the kernel function bandwidth

The data-driven estimation of the kernel bandwidth in Eq. 1 forms the focus of our work. We frame this problem as one of probabilistic clustering: atoms with similar values of position and stress are grouped together into homogeneous sub-populations. If we assume that atoms with similar stress and position are indicative of those atoms that should belong within the support of the kernel function $\psi(\mathbf{y}-\mathbf{y}^i)$, then it follows that the discovered sub-populations can be used to estimate the size of the characteristic length. To do so, we advocate the measurement of the maximum spatial extent of each sub-population: the maximum extent of a sub-population equals the furthest distance of a contained data point to the mean of all data points in the sub-population. We take the median of the maximum extents as an estimate of the size of the average characteristic length.

In our contribution we address the question of how to effectively estimate the similarity between atoms as represented by their position and stress attributes, and therefore how to form a representative grouping of the atomistic data with no requirement for manual tuning of model parameters. We demonstrate that the Gaussian mixture model (GMM) [2] performs

[*] Corresponding author: Email s.j.moran@sms.ed.ac.uk, phone +44 131 650 4449

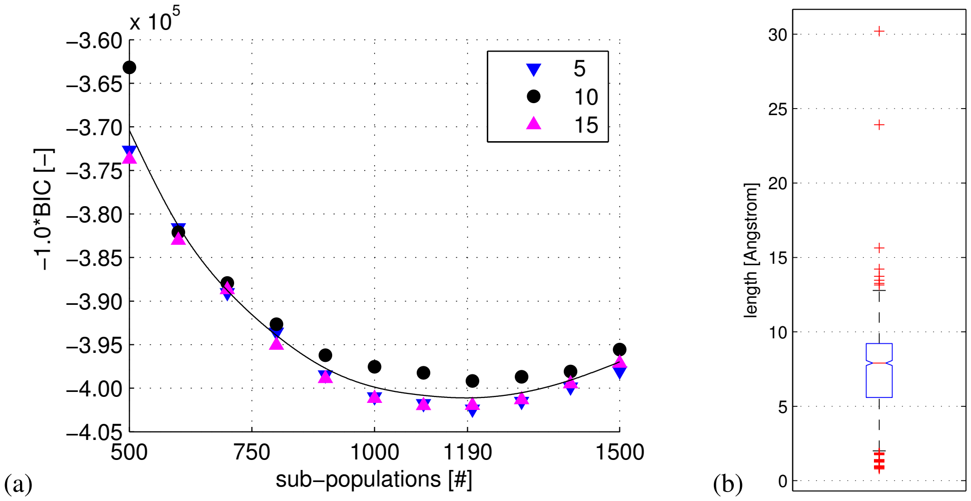

Fig. 1 (a) BIC for varying sub-population counts and random initializations of EM (as the EM algorithm is sensitive to the random initialization of the sub-populations, this experimental procedure is repeated for three different random seeds of 5, 10 and 15). The solid line is a non-parametric smoothing spline fitted to the data. (b) Boxplot of the maximum extent of GMM computed sub-populations.

well at this task. Mixtures of Gaussian functions are ideally suited to modelling sub-populations of data-points. In this case, each sub-population is modelled by a multivariate Gaussian distribution, with its mean indicating the center of the sub-population and the covariance measuring the spread. We apply the GMM to probabilistically cluster the atomistic dataset: each atom is assigned to a sub-population which has the highest probability of containing that atom.

The parameters of the Gaussian mixture model are learnt directly from the atomistic data using the Expectation Maximization (EM) algorithm [3]. EM is an efficient iterative algorithm which finds a set of maximum likelihood parameters for the GMM. We assume the EM algorithm has converged to a local maximum after either 500 iterations, or when the relative increase in log-likelihood was below 10^{-4}. The EM parameter learning step is repeated for different values of the sub-population count. We select for our experimentation the GMM with a sub-population count that leads to a minimum of the negative Bayesian Information Criterion (BIC) [4]. The BIC is an additional parameter dependent factor that is added to the likelihood function, penalizing overly complex (many sub-population) models. By conducting a systematic search for the model that minimizes the negative BIC we are able to discover a representative set of sub-populations, and therefore compute the characteristic length, in an entirely data-driven manner.

3 Numerical example

We consider the distribution of residual stress about the core of an edge dislocation in an elastic solid. In this case, the stress arises from the defect itself and not from external loading or inhomogeneities. LAMMPS is used to generate an atomistic model of a single crystal of copper in a face-centred cubic lattice. The edges are aligned to the crystallographic axes and have $1000 \times 1000 \times 3$ unit cells with periodic boundaries on the two square surfaces and free boundaries on the remaining four surfaces. A $\langle 100 \rangle$ edge dislocation is created at the center of the simulation box. An embedded-atom method (EAM) potential is used to determine the atomic interactions with cut-off radius of 2.5 times the unit cell length a of 3.615 Å. Energy minimization at zero temperature with relative tolerance 10^{-15} brings the system to equilibrium.

As the stress distribution about the edge dislocation may be described in a plane, we only consider a five-dimensional data space for this problem set: each atom presents a data point with two spatial dimensions (x and y) and three stress dimensions ($\boldsymbol{\sigma}_{xx}$, $\boldsymbol{\sigma}_{yy}$ and $\boldsymbol{\sigma}_{xy}$). A dataset takes the atoms in a $100a \times 100a$ slice surrounding the edge dislocation in a single-crystal. The optimal number of sub-populations is shown in Fig. 1(a) that occurs at 1190 sub-populations for this copper data set. Furthermore, the maximum extent of all sub-populations are plotted Fig. 1(b). This box plot excludes those sub-populations within close proximity to the edge dislocation which contain only a single atom. We take the median as our estimate of the size of the space averaging volume that evaluates to approximately 7.8 Å.

Finally, we evaluate the Hardy stress for different sizes of the space averaging volume through a parameter study conducted with LAMMPS. The lower bound for the size of the space averaging volume is the smallest size which results in a stress distribution about an edge dislocation that converges to the local analytical solution. This is found to be 16.3 Å and agrees well with the GMM estimated value of 15.6 Å (two times the median of the maximum extents). Further details of the outlined procedure and a complete presentation of the results are given in [5].

References

[1] N. C. Admal and E. B. Tadmor, J. Elasticity **100**, 63 (2010).
[2] B. Everitt and D. J. Hand, Finite mixture distributions (London: Chapman and Hall, London, 1981).
[3] A. P. Dempster, N. M. Laird and D. B. Rubin, J. R. Stat. Soc. Ser. B-Stat. Methodol. **39**, 1 (1977).
[4] G. Schwarz, Ann. Stat. **6**, 461 (1978).
[5] M. H. Ulz and S. J. Moran, Modelling Simulation Mater. Sci. Eng. (submitted).