L_1 Norm Based Shrinkage and Selection Regression of Cluster Expansion

Funded Project: Modeling and Simulation of High Dimensional Stochastic Multiscale PDE Systems at the Exascale

Guang Lin
Fundamental & Computational Sciences Directorate
Pacific Northwest National Laboratory
902 Battelle Boulevard, P.O. Box 999, MSIN K7-90
Richland, WA 99352
and
Nicholas Zabaras
Materials Process Design and Control Laboratory
Sibley School of Mechanical and Aerospace Eng.
101 Rhodes Hall, Cornell University, Ithaca, NY 14853-3801
http://mpdc.mae.cornell.edu/

Abstract

The cluster expansion has been widely used in recent years to identify ground states, calculate phase diagrams, and study ordering of systems of substitutional disorder. In addition, it has been employed to model kinetic activation energies, tensor properties, band gaps, orientations of complex ions, amino acid sequences in proteins, and configurational electronic entropy.

Two versions of the cluster expansion dominate the majority of the practical applications of the cluster expansion: the discrete Fourier transform (DFT) equivalent conventional cluster expansion and the variable basis cluster expansion whose rate of convergence is superior when applied to properties that are nonlinear functions of the subcomponent concentration.

Practical applications of the cluster expansion involves truncating the full cluster expansion and regressing the coefficients of the correlation functions (also known as Effective Cluster Interactions) using data from DFT calculations. The predictive power of the truncated and regressed cluster expansion depends heavily on the choices of the correlation functions to retain (which to date has exclusively been determined manually relying on experience) and the regression strategy.

Here we propose to apply an L_1 norm based shrinkage and selection regression scheme to automate the truncation and regression process in constructing the cluster expansion. Our algorithm automatically determines the optimal set of correlation functions to retain and their corresponding coefficients in order to achieve the best possible predictive power with the computation cost in the same order of magnitude as that of an ordinary least square regression. Another advantageous feature of this regression algorithm is that it is a path algorithm able to generate the entire shrinkage path within one single computation process.

From a Bayesian point of view, the L_1 norm based constraint is equivalent to assigning double exponential prior distributions to the regressors. The equivalence is established by taking the posterior modes, i.e., maximizers of the posterior, derived from such prior to be the estimates of the coefficients. The equivalence is established by taking the posterior modes, i.e., maximizers of the posterior, derived from such prior to be the estimates of the coefficients.