My primary research interests lie in theoretical machine learning and practical climate forecasting and energy system optimization. Particularly, the main contributions of my dissertation research emphasize the creation of a new measure of clusterability by defining a min-max based matrix multiplication and an unsupervised model selection method with the help of generalized partitional entropy. I also attributed to several projects in precipitation estimation, water price prediction, and power system stability operations. Due to the increasing need for large data processing, my research agenda focuses on scaling the existence of potential data structures and identifying ways to leverage such scales to improve the algorithm performance in practice.

In this statement, I will briefly introduce my dissertation research about clusterability and unsupervised model selection. I will also discuss some of my works in the field of energy system stability and climate analysis. My future research direction will also be detailed in the end.

## Dissertation Research: Clusterability and Model Selection

Clustering is a prototypical, unsupervised machine learning activity which consists in identifying cohesive and well-differentiated groups of records in data. However, due to the variation in data distributions and the inadequate formalization of certain basic notions of clustering, it is still difficult to determine when two objects are similar and to what degree. This issue is related to two typical problems of clustering: clustering tendency and clustering validity. The first problem focuses on determining the data clusterability before applying specific clustering algorithms, while the second one looks at how to select the best parameter (usually the number of clusters) of the clustering model.

Evaluating data clusterability before applying clustering algorithms can be very helpful because clustering algorithms are expensive. According to Shai Ben-David's work, a good clusterability measure should satisfy three criteria: efficiency, algorithm independence, and effectiveness [1]. Generally, users prefer to use data reduction methods to see the data structure directly. However, many traditional dimension reduction algorithms like Principle Component Analysis (PCA), which may not be well suited for non-linear structures, could fail to realize the real structures of some datasets. Statistical tools like multimodality tests or spatial randomness are also popular for measuring clusterability. However, multimodality tests, such as the Silverman-test or Dip-test, still have unclear behaviors on multi-dimensional datasets, while measures of spatial randomness like Hopkins' Statistic could be sensitive to the outliers [2].

In my dissertation, I proposed a novel matrix multiplication method via min-max to quantify the degree of clusterability. We show that by applying our technique to a dissimilarity matrix, a so-called ultrametric dissimilarity matrix is generated [3]. We then prove that such a matrix can form a hierarchical clustering tree of the corresponding data set in $O(nlogn)$ time. The value in this final matrix can be regarded as the distance between two nodes in a weighted graph; this is found by identifying the set of largest edges along all possible connecting paths between these two nodes and defining the pairwise distance as the minimum hop (edge).

Our measure successfully solved the previously mentioned problems of other current methods. Due to the property of ultrametric space, our measure can handle non-spherical cluster structures, fitting to the high dimensional data space. It also does not require any prior assumption of data distribution. The experiments also show its robustness to outliers.

In addition, since our measure represents the degree of how well a dataset can be grouped into meaningful clusters, we can improve the performance of clustering algorithms on some challenging datasets, for example, the dataset formed as two entangled spirals. Experiments show that if we apply the same clustering algorithm on the final ultrametric distance matrix instead of the original one, the final clustering result is much better [4].

The second part of my dissertation is related to model selection. Even if we can determine there is a clustering structure in a data set, we still face new problems after running the clustering algorithms. It is always necessary to justify whether the produced clustering is the best clustering representation of the data set. Such issues will be even worse if the inherent clusters are imbalanced. Currently, most of the works on unsupervised model selection have not paid attention to the imbalance-distributed clusters.

We invented a technique grounded in information theory for determining the "natural" number of clusters that exist in a data set [5]. It involves a bi-criteria optimization based on generalized partitonal entropy [6] and the cohesion of a partition. The results are promising, and it can also be a competitive evaluation index for judging clustering qualities. The most significant contribution of this model selection index is its ability to deal with clusterings of imbalanced data in an unsupervised manner.

## Climate Analysis and Power System Optimization

Besides the experience of theoretical machine learning, climate analysis with spatio-temporal data (NCEP-NCAR Re-analysis) for long-term (5-14 days) precipitation forecasting is also one of my research topics. In a project of this topic, I conducted a Hierarchical Clustering-based Bayesian Structural Vector Autoregression (HC-BSVAR) to predict the precipitation amounts at particular locations [7]. The single-linkage hierarchical clustering was first applied to identify the "similar" locations and generate a new multivariate time-series data. Then, the new dataset was processed with the Bayesian SVAR to get the final prediction. Instead of grouping data based on the temporal information, our method clusters location for each time spot and generates a time-series dataset based on only hydro-meteorological variables. This model also gives the relationship between each hydro-meteorological variable on specific target locations and provides a set of "similar" locations that help predict the precipitation of the target location.

I also cooperated with researchers from the Ministry of Water Resources of China to create a model that predicts water rights price, which is a significant factor in the cost of hydro-power generation [8]. Currently, the existence of water markets establishes water rights prices, promoting water to be traded from low to high-valued uses. However, market participants can face uncertainty in asking and offering prices as the water rights are heterogeneous, resulting in the inefficiency of the water markets. We proposed a random forest model with feature selection to achieve the right prediction of future water right price in a specific place. Transactions of 12 semiarid States since 1987 and part of the NCEP-NCAR Re-analysis dataset containing various hydro-meteorological predictor variables were assembled. The importance of influencing factors associated with water prices was then analyzed based on the plausible variable importance rankings generated by our model. Results showed that the models had good predictive power for water prices and revealed active assistance for potential water market participants to make more efficient decisions.

I also have research experience under the supervision of Professor Hsiao-Dong Chiang at Cornell, where we looked at the minimax approximation optimization on state estimation of the power system. Due to inevitable abnormal data reading from the power system, most companies need state estimation to calibrate the power system information into an optimal estimate. However, the traditional weighted least square method leads to some unexpected residuals, which may confuse the users. Our algorithm embedded the minimax approximation to minimize the maximum residuals so that all of the residuals are within a reasonable or customized range. Because of its excellent performance, the product that uses our algorithm as its cornerstone was adopted by a power system company in South China.

## Future Research Direction

Although our clusterability measure has several benefits in dealing with arbitrary datasets, the method still needs users to set up a threshold to empirically classify whether a dataset is clusterable or un-clusterable, according to its continuous output. From the experiments, the method reveals a clear threshold for discriminating the existence of latent clustering structure, which is good for comparing the clusterability between

different datasets. However, a more convincing way to instruct users of judging clusterability on a specific dataset is necessary. One of my interests is to leverage the multimodality test to statistically provide a threshold or a confidence interval for judging whether a dataset is clusterable or not.

We proved that the final ultrametric matrix (as mentioned in the fifth paragraph) we generated from the original dissimilarity matrix is identical to the cophenetic distance matrix of the result of the single-linkage hierarchical clustering on the same dataset (this is also a possible explanation for why our measure is robust to outliers) [3]. The family of linkage-based hierarchical clustering algorithms can be generalized by the nearest-neighbor chain algorithm, with the only difference being the cluster merging scheme, and the merging process can also be generalized by the Lance-Williams formula. Since our method can achieve the same result of single-link hierarchical clustering, the process of matrix multiplication could be further enhanced via GPU, and the speed of hierarchical clustering will be sharply improved. However, single-link has its own merits and demerits. If we can find a generalized matrix multiplication that can achieve other linkage-based hierarchical clustering algorithms by only changing some parameters, the speed of different hierarchical clustering algorithms will also be largely increased. Many applications that prefer to use the hierarchical clustering will also speed up their process of analysis.

## Selected References

[1] S. Ben-David, "Computational feasibility of clustering under clusterability assumptions," *arXiv preprint arXiv:1501.00437*, 2015.

[2] A. Adolfsson, M. Ackerman, and N. C. Brownstein, "To cluster, or not to cluster: An analysis of clusterability methods," *Pattern Recognition*, vol. 88, pp. 13–26, 2019.

[3] D. Simovici and K. Hua, "Data ultrametricity and clusterability," in *Journal of Physics: Conference Series*, vol. 1334, p. 012002, IOP Publishing, 2019.

[4] K. Hua, *Clusterability, Model Selection and Evaluation*. PhD thesis, University of Massachusetts Boston, 2019.

[5] K. Hua and D. A. Simovici, "Dual criteria determination of the number of clusters in data," in *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pp. 201–208, IEEE, 2018.

[6] D. Simovici, "On generalized entropy and entropic metrics," *Multiple-Valued Logic and Soft Computing*, vol. 13, no. 4-6, pp. 295–320, 2007.

[7] K. Hua and D. A. Simovici, "Long-lead term precipitation forecasting by hierarchical clustering-based bayesian structural vector autoregression," in *2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC)*, pp. 1–6, IEEE, 2016.

[8] Z. Xu, J. Lian, L. Bin, K. Hua, K. Xu, and H. Y. Chan, "Water price prediction for increasing market efficiency using random forest regression: A case study in the western united states," *Water*, vol. 11, no. 2, p. 228, 2019.