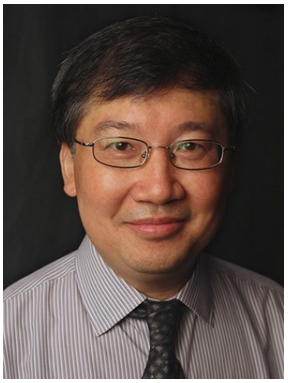
UNIVERSITY of WISCONSIN
UWMILWAUKEE

Special Interest Group
on Big Data

Biweekly Seminar
(lunch provided)

12pm - 1pm, Friday, March 15, 2019
UWM EMS 715

Principal component analysis combined with truncated-Newton minimization for dimensionality reduction of chemical databases



Dexuan Xie, Ph.D.
Professor
Department of Mathematical Sciences
University of Wisconsin-Milwaukee
Home page: www.uwm.edu/~dxie

The similarity and diversity sampling problems are two challenging optimization tasks that arise in the analysis of chemical databases. As a first step to their solution, we propose an efficient projection/refinement protocol based on the principal component analysis (PCA) and the truncated-Newton minimization method. We show that PCA can provide the same initial guess as the singular value decomposition (SVD) for the optimization task of solving the distance-geometry optimization problem if each column of a database matrix has a mean of zero. Using PCA/TNPACK and the Merck MDDR database (MDL Drug Data Report), we further investigate the projection/refinement procedure with regards to the preservation of the original clusters of chemical compounds, the accuracy of similarity and diversity sampling of chemical compounds, and the potential application in the study of structure activity relationships. We also explore by simple experiments accuracy and efficiency aspects of the PCA/TNPACK procedure compared to those of a global optimization algorithm (simulated annealing, as implemented by the program package SIMANN) in terms of producing the projection mapping of a database. Numerical results show that the 2D PCA/TNPACK mapping can preserve the distance relationships of the original database and is thus valuable as a first step in similarity and diversity applications. Since all numerical tests are performed on the Merck MDDR database, results are representative of realistic cases encountered in the field of drug design, and may help analyze properties of medicinal compounds