New Geometric Methods of Mixture Models for Interactive Visualization

Jia Li[1], Bruce Lindsay[1], Xiaolong (Luke) Zhang[2]

[1]Department of Statistics

[2]College of Information Sciences and Technology

The Penn State University

Goals

- Develop theories and algorithms for revealing prominent geometric features of mixture density.
- Develop approaches to clustering, dimension reduction, and variable selection based on the geometry of mixture density.
- Develop interactive visualization systems empowered by a suite of statistical learning tools.
- Apply the statistical methods and visualization paradigm to meteorology data for weather prediction and engineering design data

Our Work

- Theories and algorithms
  - Modal EM algorithm for solving modes of mixture density.
  - Clustering methods based on mode association.
  - Variable selection based on the geometry of mixture density.
  - Two-way mixture model for high dimensional data.
- Visualization system design
  - A work-centered visual analytics model
  - Explored applications to meteorology data and engineering design data.
    - Preliminary evaluation: engineering design case
- Parallelization of data clustering algorithms

Model EM (MEM)

- Let a mixture density be $f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$.
  - $x \in \mathcal{R}^d$
  - π_k is the prior probability of mixture component k.
  - $f_k(x)$ is the density of component k.

- Given any initial value $x^{(0)}$ MEM solves a local maximum of the mixture by alternating two steps.

Mode Association Clustering (MAC)

- The MAC Algorithm

1. Form kernel density $f(x \mid S, \sigma^2) = \sum_{i=1}^{n} \frac{1}{n}\phi(x \mid x_i, D(\sigma^2))$, where $S = \{x_1, x_2, ..., x_n\}$.
2. Use $f(x|S, \sigma^2)$ as the density function. Use each x_i, $i = 1, 2, ..., n$, as the initial value in the MEM algorithm to find a mode of $f(x|S, \sigma^2)$. Let the mode identified by starting from x_i be $\mathcal{M}_\sigma(x_i)$.
3. Extract distinctive values from the set $\{\mathcal{M}_\sigma(x_i), i = 1, 2, ..., n\}$ to form a set G. Label the elements in G from 1 to $|G|$.
4. If $\mathcal{M}_\sigma(x_i)$ equals the kth element in G, x_i is put in the kth cluster.

Hierarchical Mode Association Clustering (HMAC)

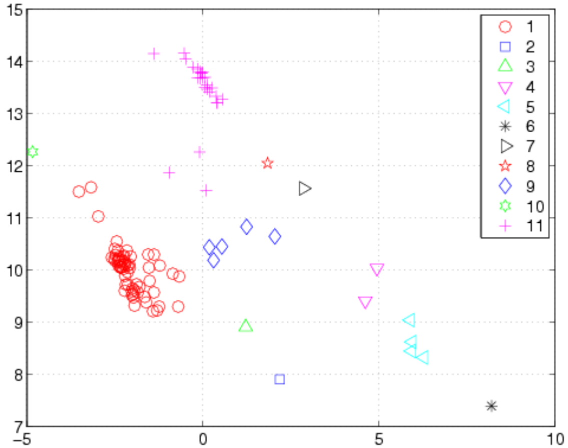
- Gradually increase kernel bandwidth:

$$\sigma_1 < \sigma_2 < \sigma_3 \cdots$$

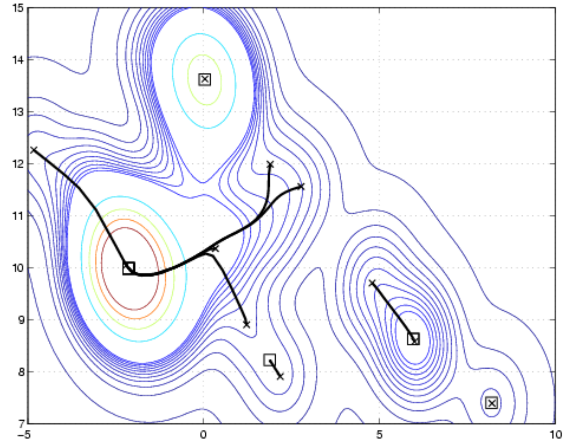
- Kernel density at level i: $f(x \mid S, \sigma_i^2)$
  - $\sigma_i \uparrow \longrightarrow$noother density, fewer modes
- Starting points at level *i* are the modes acquired at the previous level *i* - 1.
- The hierarchy by design:

$$x_i \to \mathcal{M}_{\sigma_1}(x_i) \to \mathcal{M}_{\sigma_2}(\mathcal{M}_{\sigma_1}(x_i)) \to \cdots$$

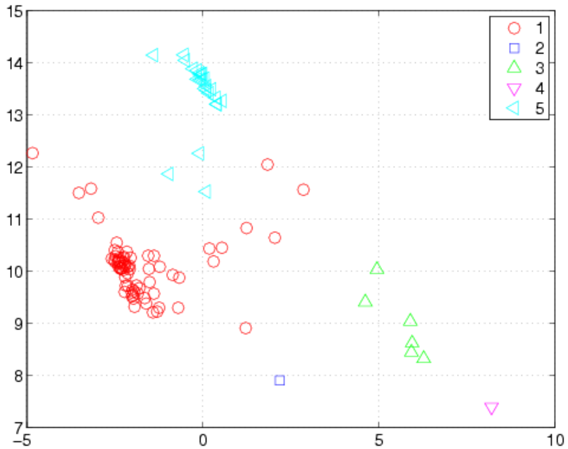
Geometry of Mixture Models



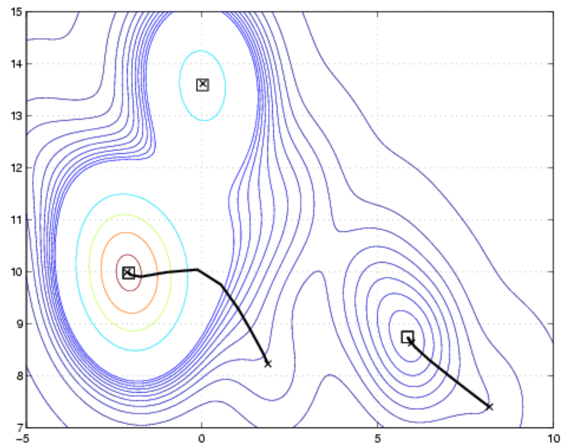
Clustering result at level 2



At level 3, merge the modes from level 2

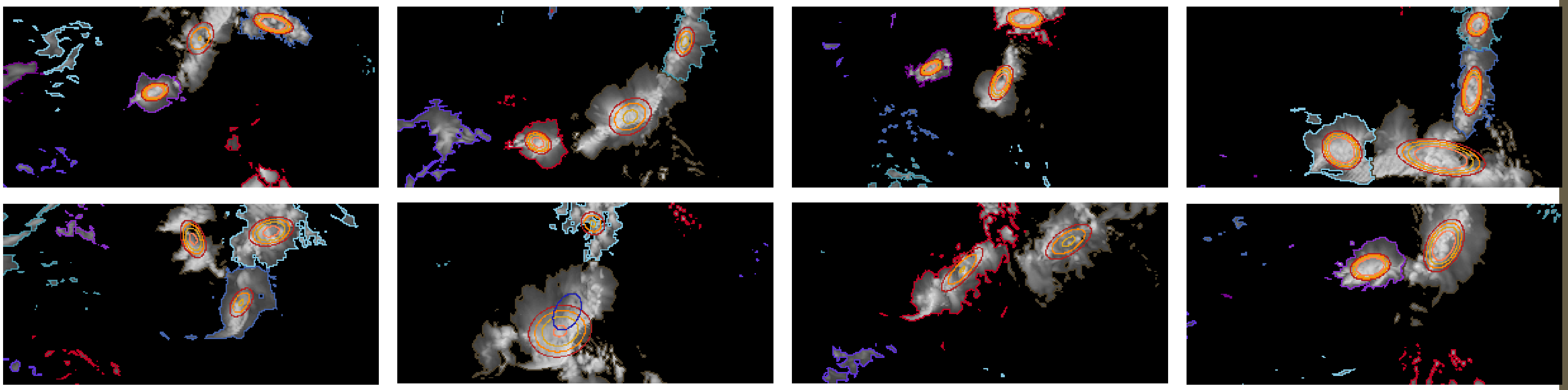


Clustering result at level 3

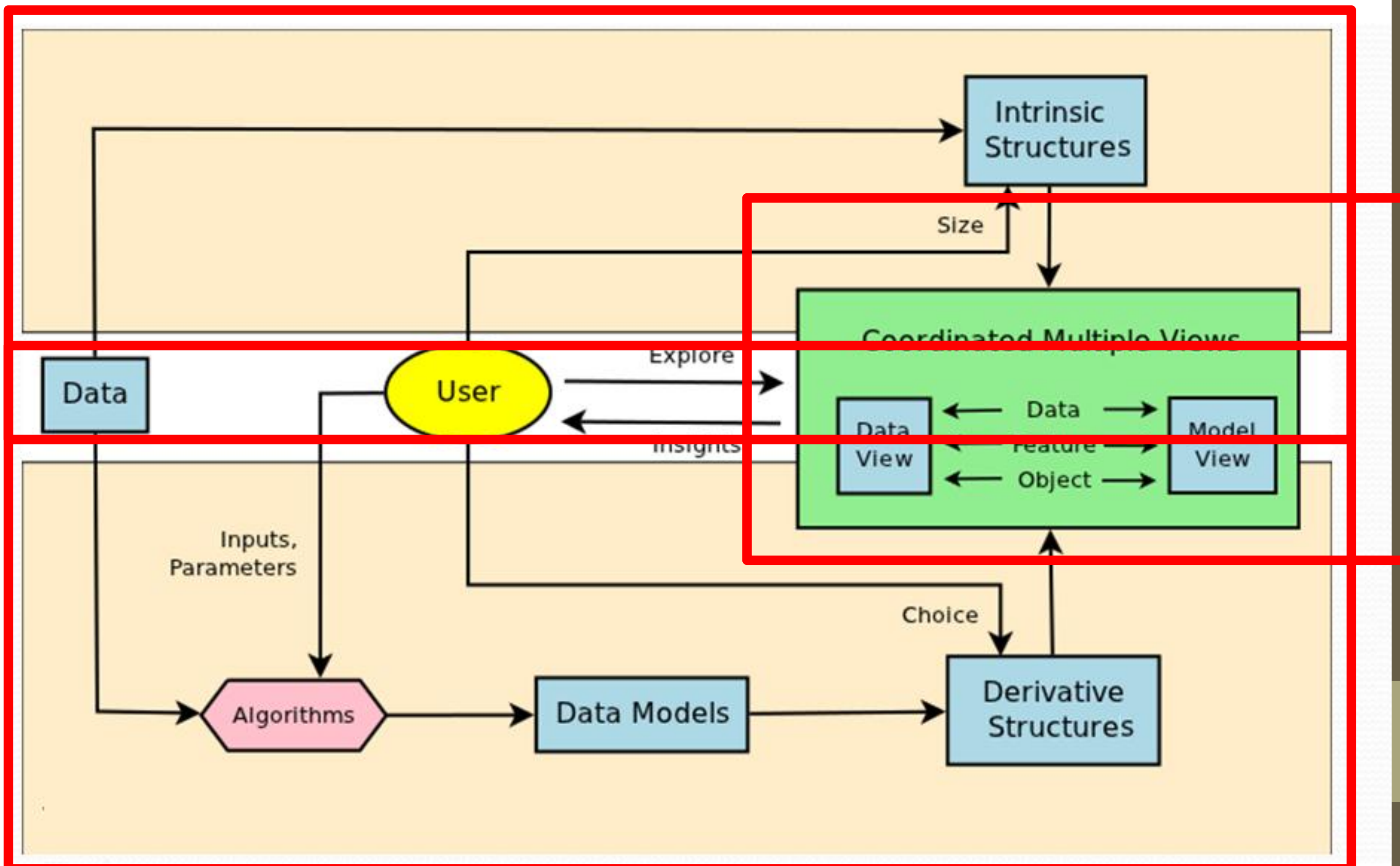


At level 4, merge the modes from level 3

Cloud Map Segmentation



A Work-Centered Model for Visual Analytics



Visual Analytics System: LIVE

Evaluation:
Conceptual Ship Design

Design input variables:

Length (L), Beam (B), Depth (D), Draft (T), Block Coeff (C_B), and Speed (V_k).

Design output variables :

Transportation Cost (TC), Light Ship Weight (LSM) and Annual Cargo (AC).

Goal

Minimize TC, minimize LSM, and maximize AC.

Constraints:

$L/B \geq 6$;

$L/D \leq 15$;

$L/T \leq 19$;

$F_n \leq 0.32$;

$25{,}000 \leq DWT \leq 50{,}000$;

$Const\_1 = T - 0.45DWT^{0.31} \leq 0$;

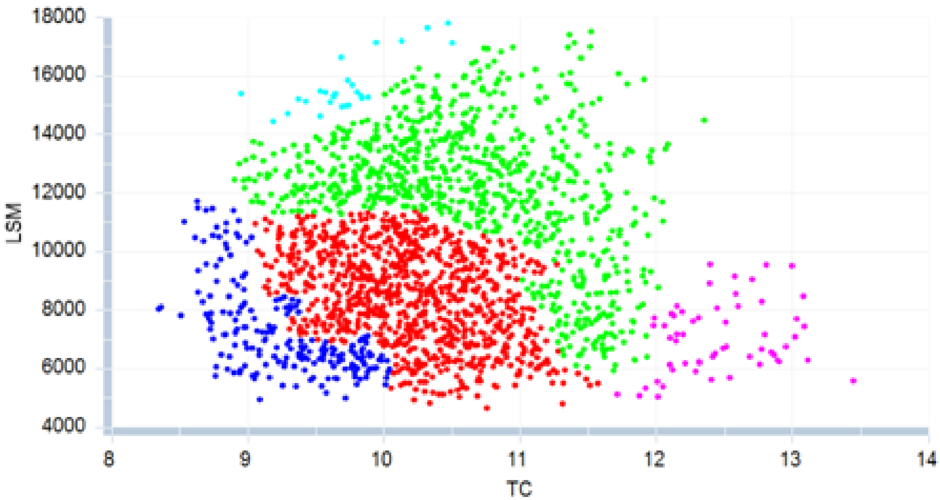
$Const\_2 = T - (0.7D + 0.7) \leq 0$;

$Const\_3 = 0.07B - GM_T \leq 0$;

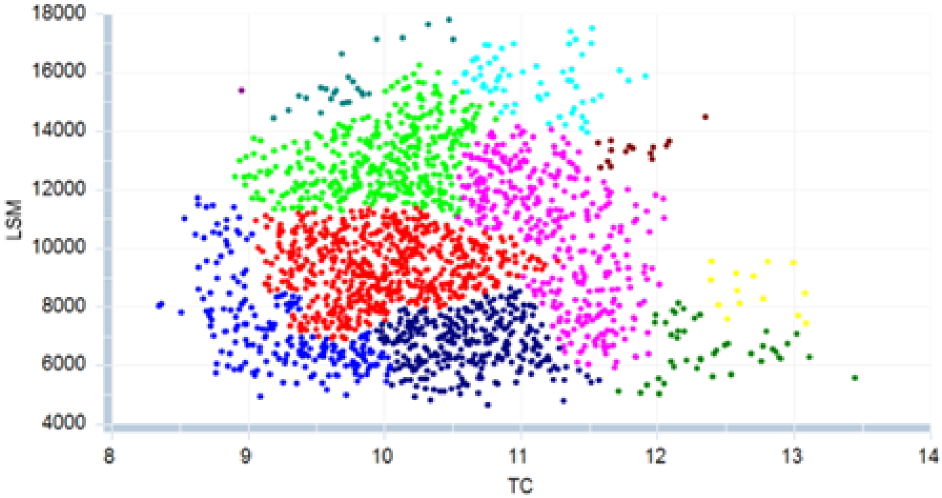
Multi-Objective Optimization (MOO)

Preliminary Result

- Our system can facilitate an iterative design optimization process.
  - Use our algorithm to indentify similar design alternatives
  - Use our algorithm to discover the values of design inputs based on desired outputs
  - Control the process of data clustering and classification
    - Step-by-step vs. batch



18000
16000
14000
12000
10000
8000
6000
4000
LSM
8
9
10
11
12
13
14
TC



18000
16000
14000
12000
10000
8000
6000
4000
LSM
8
9
10
11
12
13
14
TC

18000
16000
14000
12000
10000
8000
6000
4000
LSM
8
9
10
11
12
13
14
TC
T <= 11.52756
D > 17.437906
T <= 10.914416
B <= 28.314922
D > 18.416422
L <= 190.893982
D > 13.718345
D > 19.8305
D > 20.93
B > 20.44
B <= 22.39
B > 24.7
L <= 228.529
D > 17.4715
B > 25.840878
L <= 171.309

Preliminary Result

- Our system can facilitate an iterative design optimization process.
  - Use our algorithm to indentify similar design alternatives
  - Use our algorithm to discover the values of design inputs based on desired outputs
  - Control the process of clustering
    - Step-by-step vs. batch

- Challenges
  - Knowledge about clustering algorithms by domain experts
    - Validation
  - Speed of clustering algorithms
    - Real-time interaction

Parallelization of HMAC

- Hadoop
- MPI

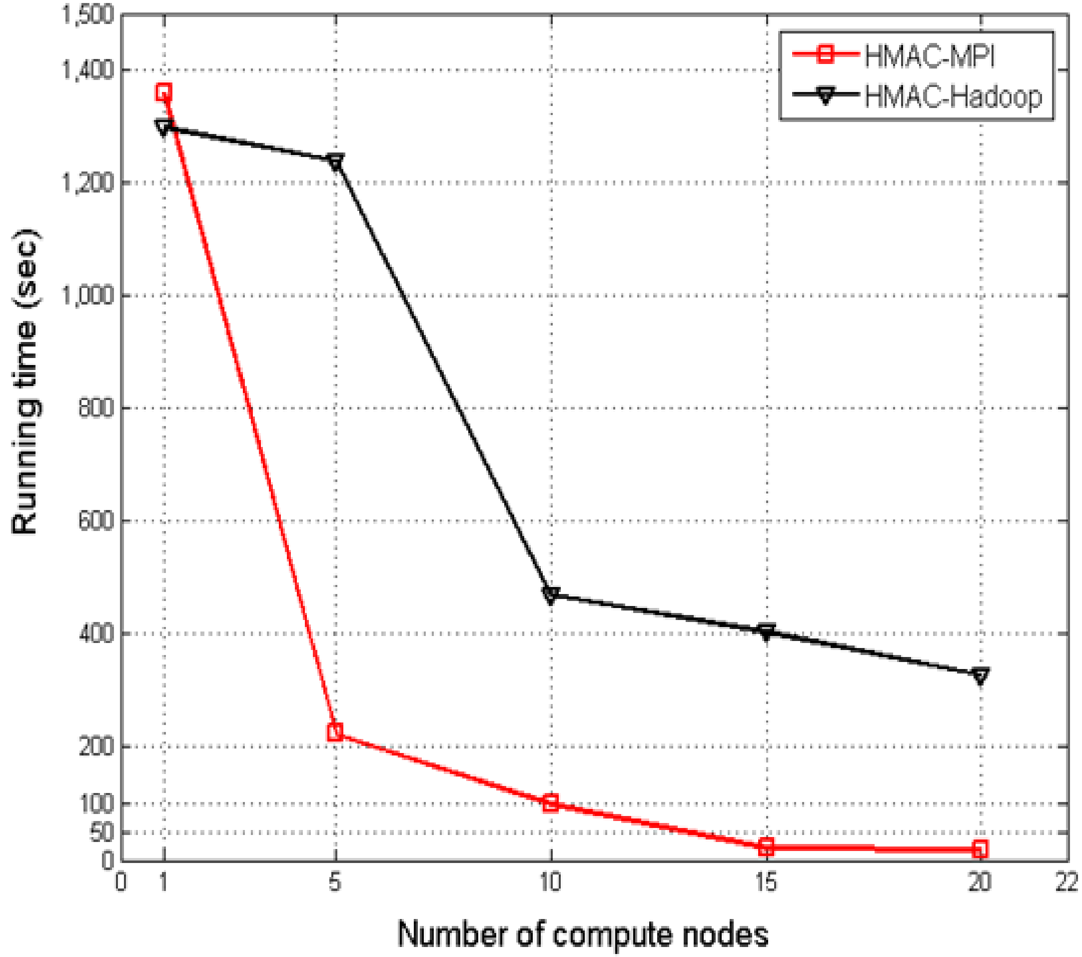
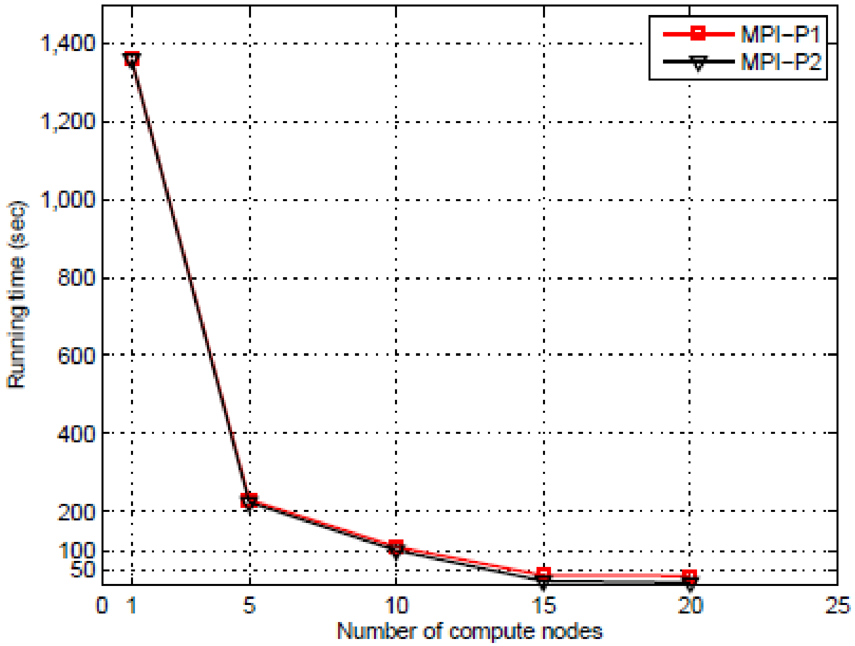


Image Data : 1,400 * 64

More Results



Ship design data: 2,000 * 17

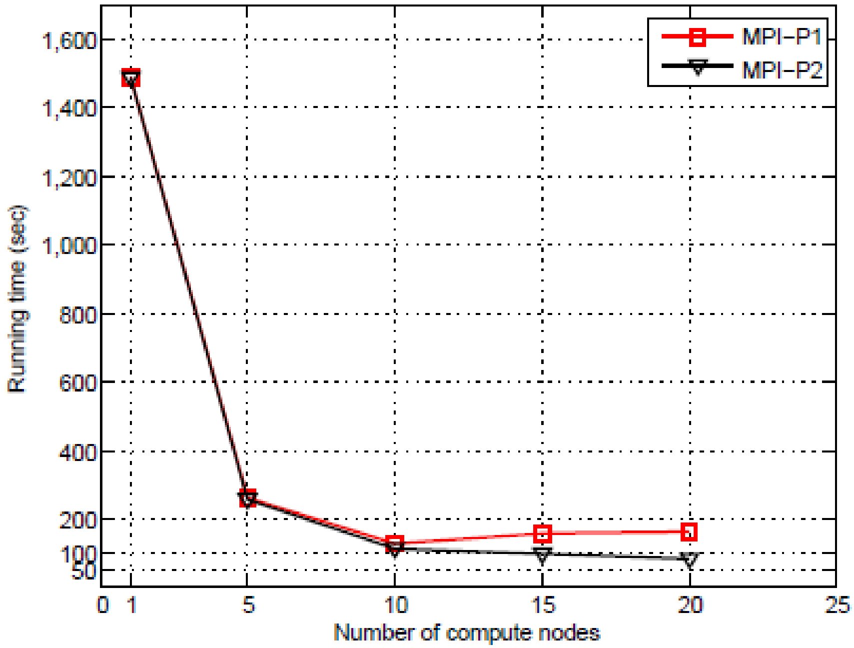


Image Data : 1,400 * 64

Project Accomplishments

- Algorithms
  - Downloadable from our project website
- Visualization design
  - A work-centered model for visual analytics
  - A system prototype to support engineering design
    - Plan to build a system for meteorology data analysis

Selected Publications

- H. M. Lee, J. Li, “Variable selection for clustering by separability based on ridgelines,” Journal of Computational and Graphical Statistics, 2012.
- M. Qiao, J. Li, "Gaussian Mixture Models with Component Means Constrained in Pre-selected Subspaces", Journal of Computational and Graphical Statistics, 2012.
- L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, J. Li, "OSCAR: On-Site Composition and Aesthetics Feedback through Exemplars for Photographers", International Journal of Computer Vision (IJCV).
- X. Yan, M. Qiao, J. Li, T. W. Simpson, G. M. Stump and X. Zhang, "A Work-Centered Visual Analytics Model to Support Engineering Design with Interactive Visualization and Data-Mining", HICSS 45.
- X. Yan, M. Qiao, T. W. Simpson, J. Li, and X. Zhang, "LIVE: A Work-centered Approach to Support Visual Analytics of Multi-dimensional Engineering Design Data with Interactive Visualization and Data-mining", ASME 2011 Design Engineering Technical Conferences - Design Automation Conference.
- M. Qiao, J. Li, "Two-way Gaussian Mixture Models for High Dimensional Classification", Statistical Analysis and Data Mining (SAM), 2010.
- M. Qiao, J. Li, “Two-way Gaussian mixture models for high dimensional classification”, *Journal of Statistical Analysis and Data Mining*, 2010.
- J. Li, S. Ray, B. G. Lindsay, “A nonparametric statistical approach to clustering via mode identification,” *Journal of Machine Learning Research*, 2007.

Impact

- Training Ph.D. students
  - Three Ph.D. dissertations
    - Statistics, CSE, Information Sciences and Technology
  - Two other Ph.D. students involved
- Led to new projects
  - Health informatics (NSF –SHB, NIH)
  - Spatial-temporal data analysis (Industrial collaboration)
- Outreach
  - Invited session in Joint Statistical Meetings (JSM), 2010 (**J. Li**)
  - Invited panelist on the Panel of Visualization in the Annual Workshop of Human-Computer Interaction Consortium, 2010 (**X. Zhang**)
  - Invited talks
    - Institute of Software at Chinese Academy of Sciences, 2011 (**X. Zhang**)
    - Xerox Research Center Europe, 2012 (**X. Zhang**)
    - NSF EarthCube Workshop, 2012 (**X. Zhang**)