

# Minimum Description Length, Graphs and Clustering with Exemplars

Po-Hsiang Lai<sup>1\*</sup>, Joseph A. O'Sullivan<sup>1</sup>, and Robert Pless<sup>2</sup>

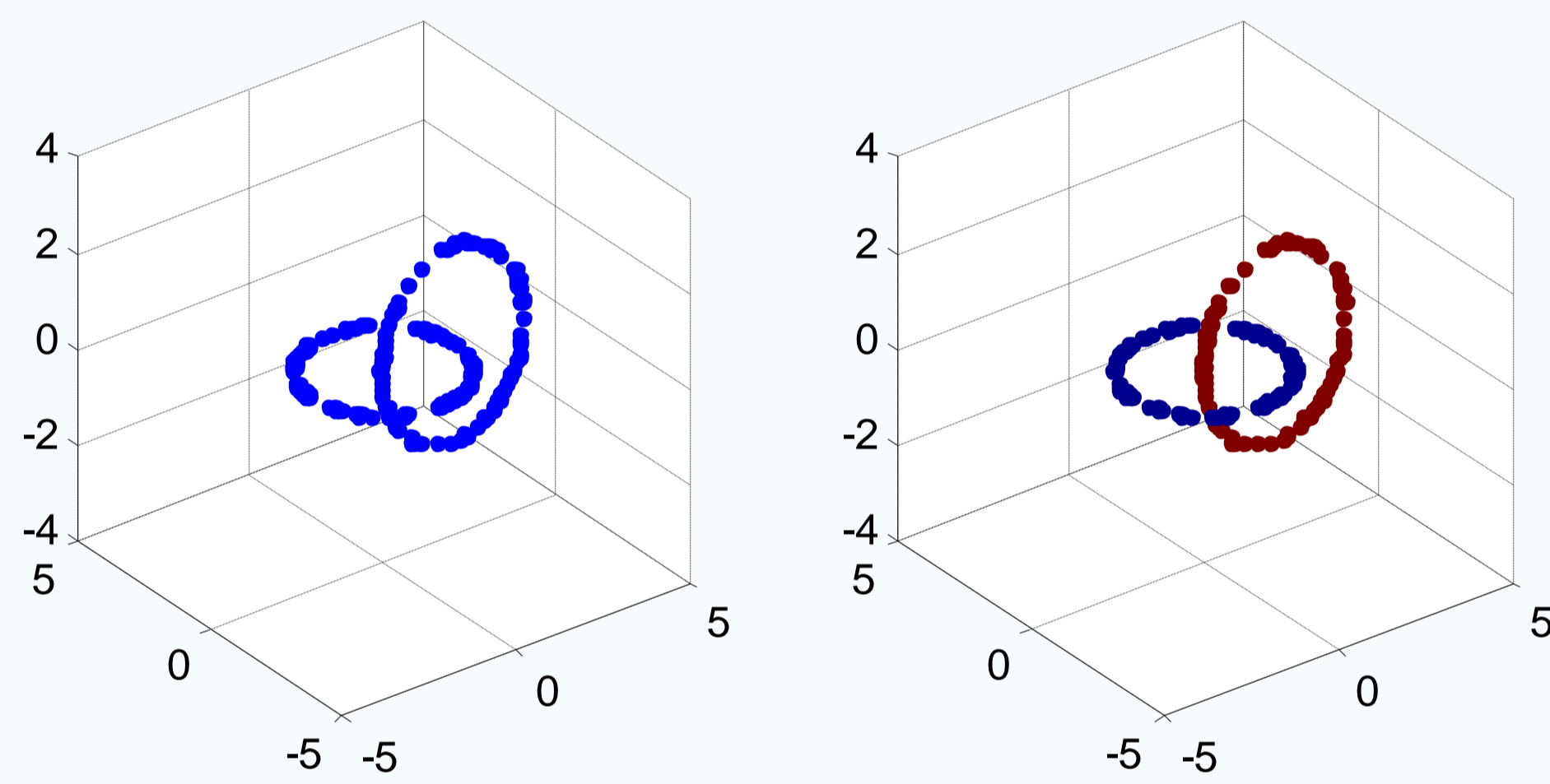
Washington University in St. Louis  
SCHOOL OF ENGINEERING & APPLIED SCIENCE

<sup>1</sup>Department of Electrical and Systems Engineering, <sup>2</sup>Department of Computer Science and Engineering, Washington University in Saint Louis

\*pl1@wustl.edu

## Clustering

- Clustering is a type of unsupervised learning that one seeks to partition data into reasonable groups.
  - Distribution based clustering
    - Fit data with a mixture model to partition data.
    - The objective is to maximize goodness of fit of a model.
    - The measure of relationship between pairs of data points changes with different choice of mixture parameters.
  - Distance (similarity) based clustering
    - Distance measure between data points is fixed.
    - View data points as vertices of a graph and distances as edge weights.
    - The objective is to remove a fixed number of edges or to optimize an objective function defined on graphs.



## Clustering and Model Selection

- Can the number of clusters and other parameters be determined in a principled way?
- Can a clustering algorithm balance the number of parameters used and the modeling error?
- Distance based clustering and MDL
  - Distances as an approximate/estimate for description length
  - Usually distances are defined between pairs of data points
  - Encoding a data point itself:  $L(x_i)$ .
  - Encoding given another data point:  $L(x_i|x_j)$ .
  - Use available compression algorithms:  $L(x_i|x_j) = L(x_i, x_j) - L(x_j)$ .
  - Quantize data and use universal code for integers.

- Distribution based clustering and MDL

$$L(x, \Theta, \gamma) = \sum_{i=1, \dots, N} -\log p(x_i | c_i, \theta_{c_i, \eta}) + \log \sum_{j=0, \dots, K-1} (-1)^j \binom{K}{K-j} (K-j)^N + \frac{d_\theta}{2} \log \frac{n}{2\pi} + \log \int |I(\theta)|^2 d\theta + o(1) + L(K, \eta)$$

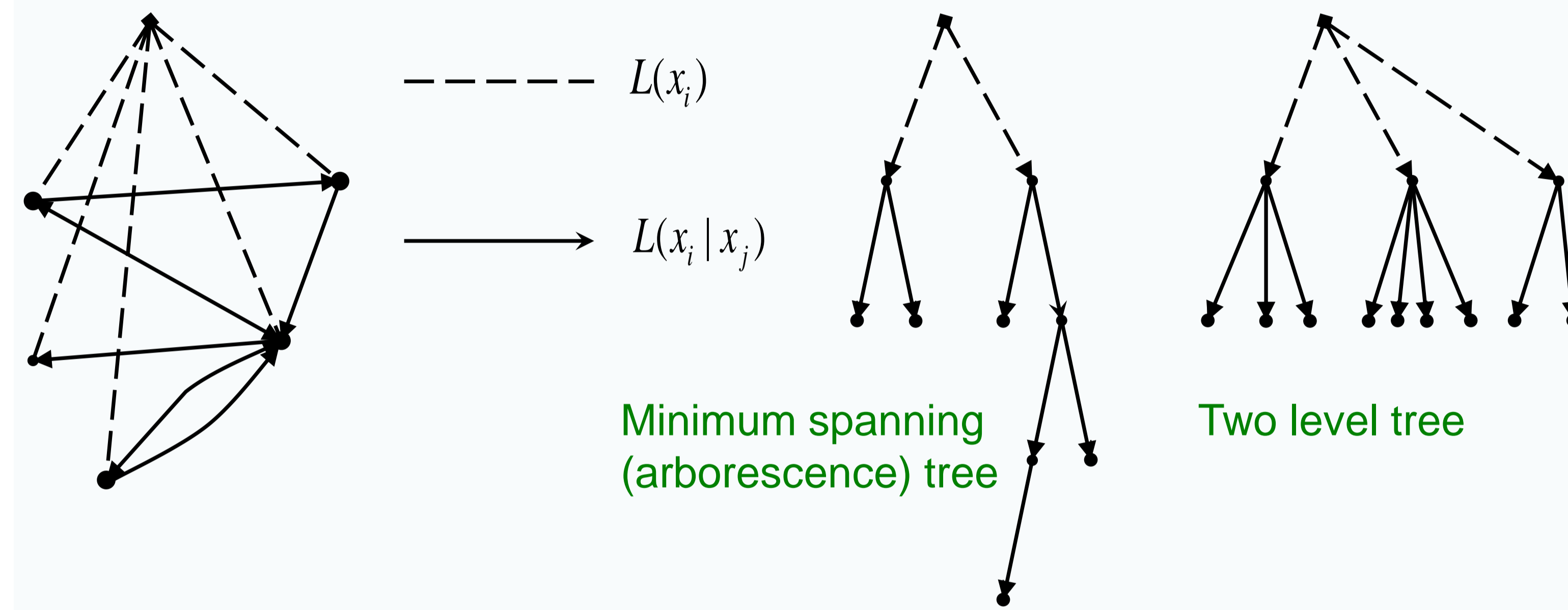
$L(x, \Theta, \gamma) = L(x | \Theta, \gamma) + L(\Theta | \gamma) + L(\gamma)$   
log of number of valid vectors  $c$   
Code length of continuous parameters (Rissanen 96)

## Objective Functions and Graphical Optimization

$$L(x | t) + L(t) = \sum_{x_i: i \neq t_i} L(x_i | x_{t_i}) + \sum_{x_i: i = t_i} L(x_i) + L(t)$$

- Constraints:

- Let  $t_i^q = t_{t_i^{q-1}}, t_i^1 = t_i$ , then  $t_i^N = t_{t_i^{N-1}}$  must hold (weak exemplar)
- $t_i^2 = t_i$  must hold (strong exemplar)



## Weak Exemplar Case

- Observe that
  - Must start with the root node
  - For every data point, there is one and only one edge pointing to it Tree
  - No cycles

- Minimum spanning tree: undirected graph, symmetric distance
- Minimum arborescence tree: directed graph, asymmetric distance

## Strong Exemplar Case

Objective function:

$$L(x | t) + L(t) = \sum_{x_i: i \neq t_i} L(x_i | x_{t_i}) + \sum_{x_i: i = t_i} L(x_i) + L(t)$$

subject to

$$t_i^2 = t_i$$

- Relax the search over  $t$  by assigning probabilities to cluster/exemplar membership:

$$\min_t L(x | t) \rightarrow \min_{P, Q} L(x | P, Q) =$$

$$\min_{P, Q} - \sum_{k=1}^N \log \left( P_k p(x_k) + (1 - P_k) \sum_{m \neq k} P_m Q(k, m) p(x_k | x_m) \right)$$

$$p(x_k) = \exp -L(x_k), \quad p(x_k | x_m) = \exp -L(x_k | x_m)$$

- Need one more minimization

## Alternating Minimization

- Convex decomposition lemma:

$$\log \left( \sum_k q_k \right) = - \min_{\pi \in P} \sum_k \pi_k \log \frac{\pi_k}{q_k}$$

- Use convex decomposition lemma to decouple the optimization problem:

$$\min_{P, Q} - \sum_{k=1}^N \log \left( P_k p(x_k) + (1 - P_k) \sum_{m \neq k} P_m Q(k, m) p(x_k | x_m) \right) = \min_{P, Q} \min_{\pi_k, q_{m\bar{k}}} \sum_{k=1}^N \pi_k \log \frac{\pi_k}{P_k p(x_k)} + (1 - \pi_k) \sum_{m \neq k} q_{m\bar{k}} \log \frac{(1 - \pi_k) q_{m\bar{k}}}{(1 - P_k) P_m Q(k, m) p(x_k | x_m)}$$

Objective function:

$$L(x | t) + L(t) = \sum_{x_i: i \neq t_i} L(x_i | x_{t_i}) + \sum_{x_i: i = t_i} L(x_i) + L(t)$$

where

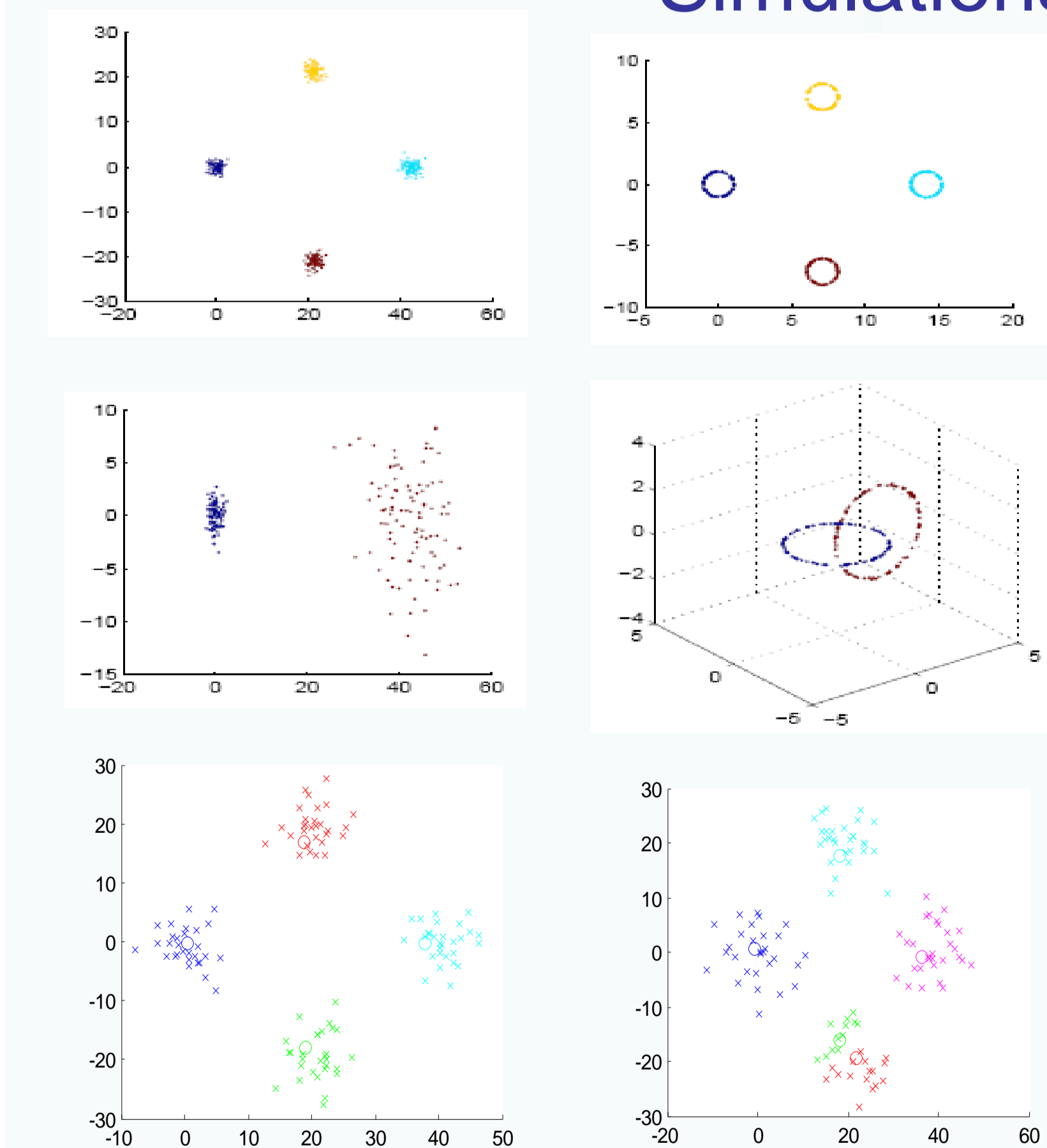
$$L(t) = (N - K) \log K + \log \binom{N}{K} + \log N, \quad \sum p_k \cong K$$

Assignments of non-exemplar points

Which ones are exemplars

How many exemplars

## Simulations



Weak exemplar clustering using MST, uniform quantization and Rissanen's universal code for integers

Strong exemplar clustering using AM algorithm under different signal to noise ratios

## References

- R. Cilibrasi and P. M. B. Vitányi, "Clustering by compression," *IEEE Trans. Info. Theory*, vol. 51, pp. 1523–1524, 2005.
- I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions, Supplement Issue*, vol. 1, pp. 205–237, 1984.
- B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.
- P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri, "An MDL framework for data clustering," in *Advances in Minimum Description Length*, P. D. Grunwald, I. J. Myung, and M. A. Pitt, Eds. MIT press, Cambridge, Massachusetts, 2005, pp. 323–353.
- J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Info. Theory*, vol. 42, pp. 40–47, 1996.
- J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Annals of Statistics*, vol. 11, pp. 417–431, 1983.
- J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Trans. Info. Theory*, vol. 30, pp. 629–636, 1984.
- A. Schrijver, *Combinatorial Optimization*. Springer, Berlin, 2003.