Complex data clustering: from neural network architecture to theory and applications of nonlinear dynamics of pattern recognition

Delay Adaptation for Projective Clustering

Jianhong Wu

Laboratory for Industrial and Applied Mathematics,
Centre for Disease Modelling,
York University, Canada

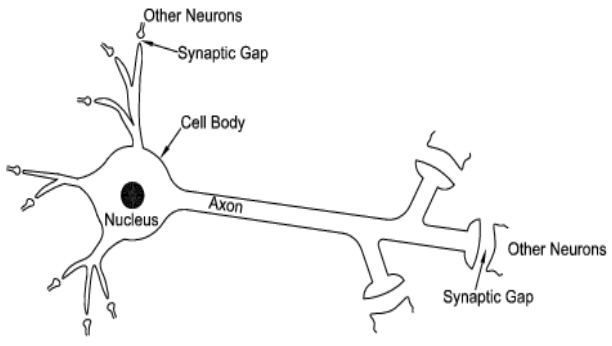
November 5, 2013

Outline

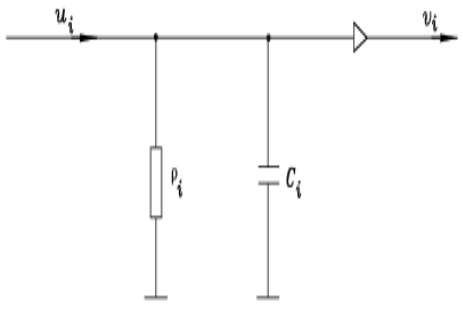
- Multi-scale dynamics in neural networks (transmission delay as a netwrok plasticity);
- Clustering: Projective clustering in high dimensional data;
- A dynamics system approach towards clustering–inverse process of pattern formation;
- Minimal network structure of a dynamical system for clustering (ART);
- Dimension relevant projective clustering (PART);
- Neurophysiological basis for delay adaptation (PART-D); Network performance/computational dynamics of PART-D;
- Minor component analysis and skewed subspace clustering; clustering in sub-manifolds;
- Some applications.

A Single Neuron

Biology

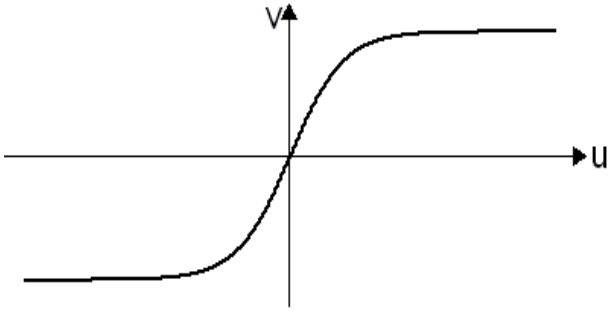


Implementation



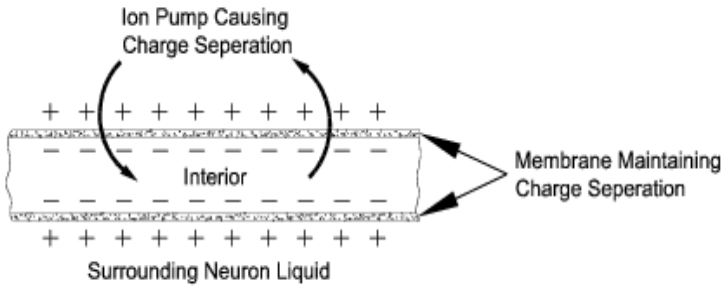
Mathematics

$$v = f(u)$$

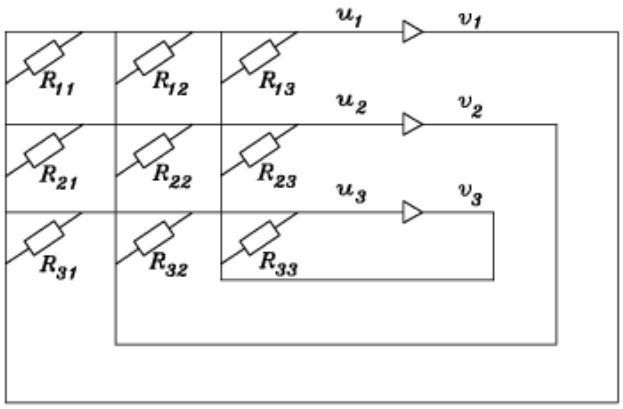


Network and Synaptic Connection: Excitation & Inhibition

Biology

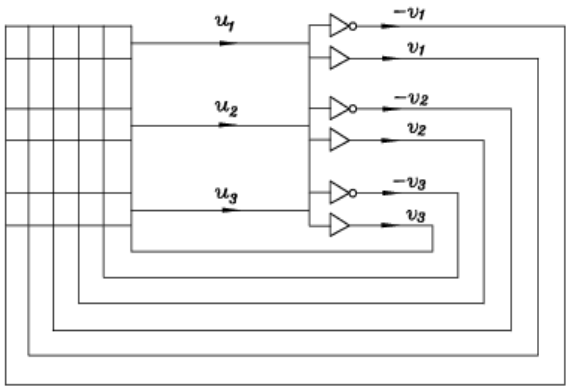


Implementation



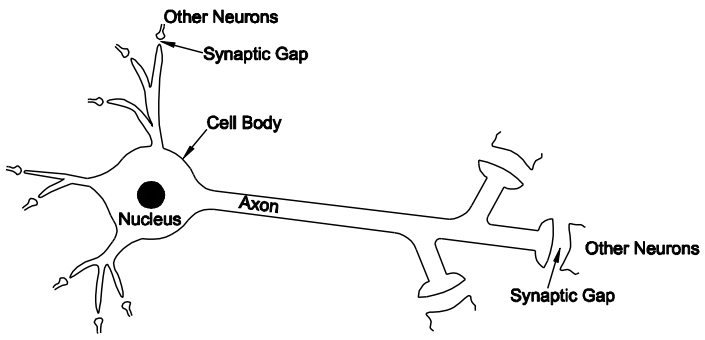
Mathematics

$$\dot{u}_i(t) = -\beta_i u_i(t) + \sum_{j=1}^{n} w_{ij} f_j(u_j(t))$$



Signal transmission delay: omnipresent in nets

- the finite rise time of postsynaptic potentials;
- signal integration time in dendrtic trees;
- conduction time for action potentials running down an axon



Multi-scale dynamics in neural signal processing

$$\dot{u}_i(t) = -\beta_i u_i(t) + \sum w_{ij} f_j(u(t - \tau_{ji}))$$

- Short-term memory trace (u_i);
- Long term memory trace (w_{ij})–the network's plasticity;
- Delay adaption in signal transmission (τ_{ji}): Glial-neuron interactions play important roles in determining axonal myelination and hence conduction velocity (Stevens et al., 1998; Fields et al., 2000);
- Signal loss due to delay: the signal losses that necessarily arises in the presence of transmission delay (Williams & Stuart, 2002; Bale & Petersen, 2009; Sincich et al., 2009).

Clustering: concept and challenge

- **Clustering:** Determine the intrinsic grouping in a set of unlabeled data to group similar objects together.
- **Abstract Formulation**: Given a data set $X \subset R^d$, find a partition $X = \cup_{i=1}^{k} P_i$ so that points within the same P_i are "close" to each other, but points from two different P_i and P_j are "far away".
- **Challenge:** Number of clusters and similarity criterion unknown, and more.

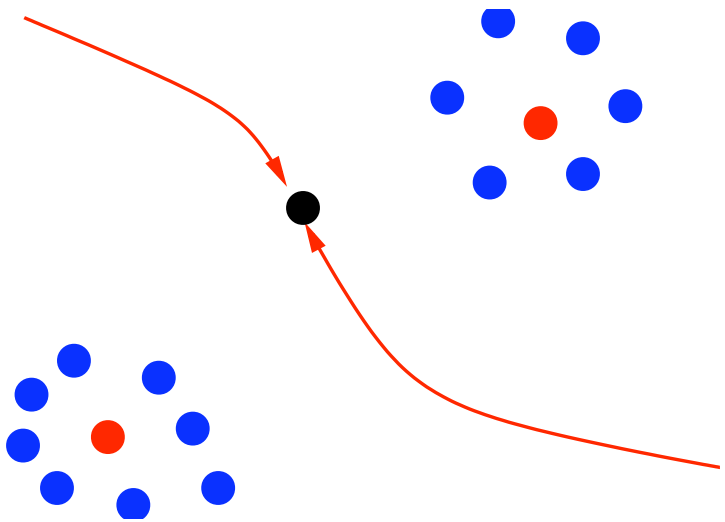
An example from taxonomy: Data, Clustering and Curse of Dimensionality

Animals	Bearing Progeny	Lune	Living Environment	...	...
dog	1	1	1 (outside water)	...	...
cat	1	1	1	...	...
sheep	1	1	1	...	...
sparrow	0	1	1	...	...
seagull	0	1	1	...	...
viper	0	1	1	...	...
lizard	0	1	1	...	...
goldfish	0	0	0 (in water)	...	...
red mullet	0	0	0	...	...
blue shark	1	0	0	...	...
frog	0	1	2 (both)	...	...

- projective subspaces are an essential part of clustering criteria, different choices of subspaces lead to different results;
- the higher the dimension of a projective space, the larger the number of clusters generated;
- If the dimensions of projective spaces are too high, the clustering results are useless (from Theodoridis and Koutroumbas (1999))

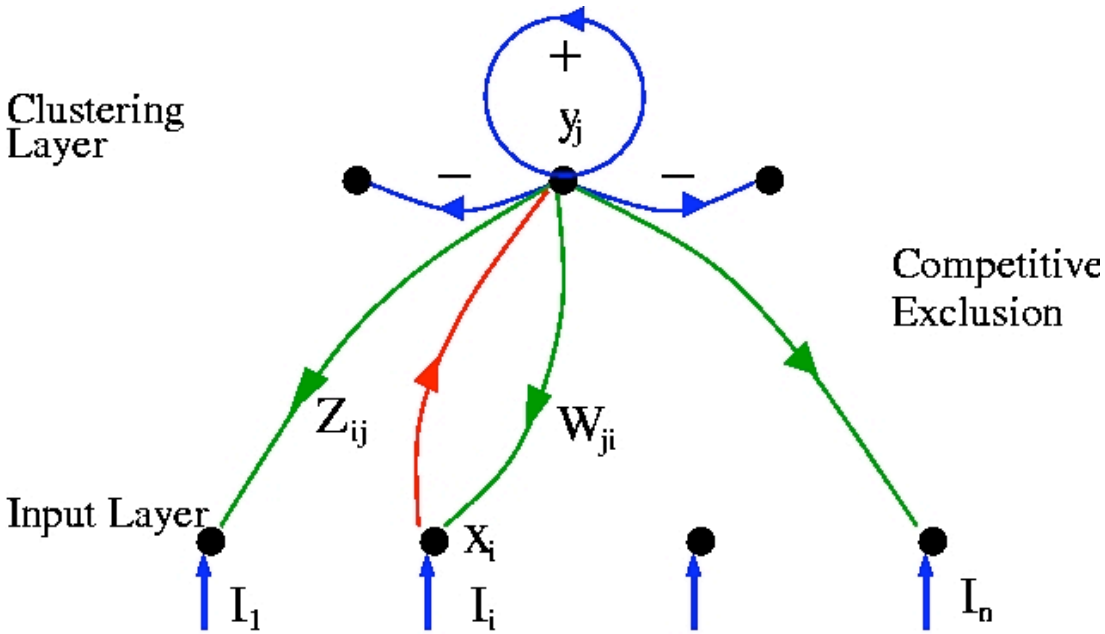
Clustering=an inverse problem of pattern formation

Our Goal: to develop some theoretical foundation from the view point of dynamical systems for data clustering: given a set of unlabeled data, construct a dynamical system so that each P_i (cluster) is contained in the domain of attraction of a stable solution (cluster center), and the clusters are separated by the boundaries of the domains of attraction (similarity measure).



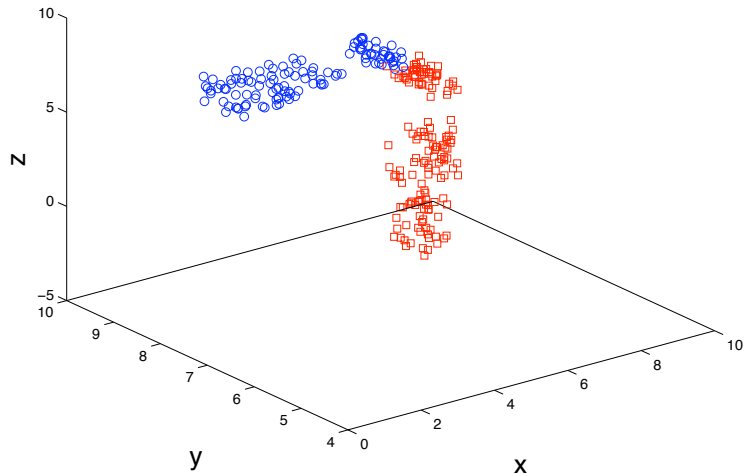
How to construct the dynamical system? what is the minimal network structure?

- Design Principle: Neural networks and central nervous systems



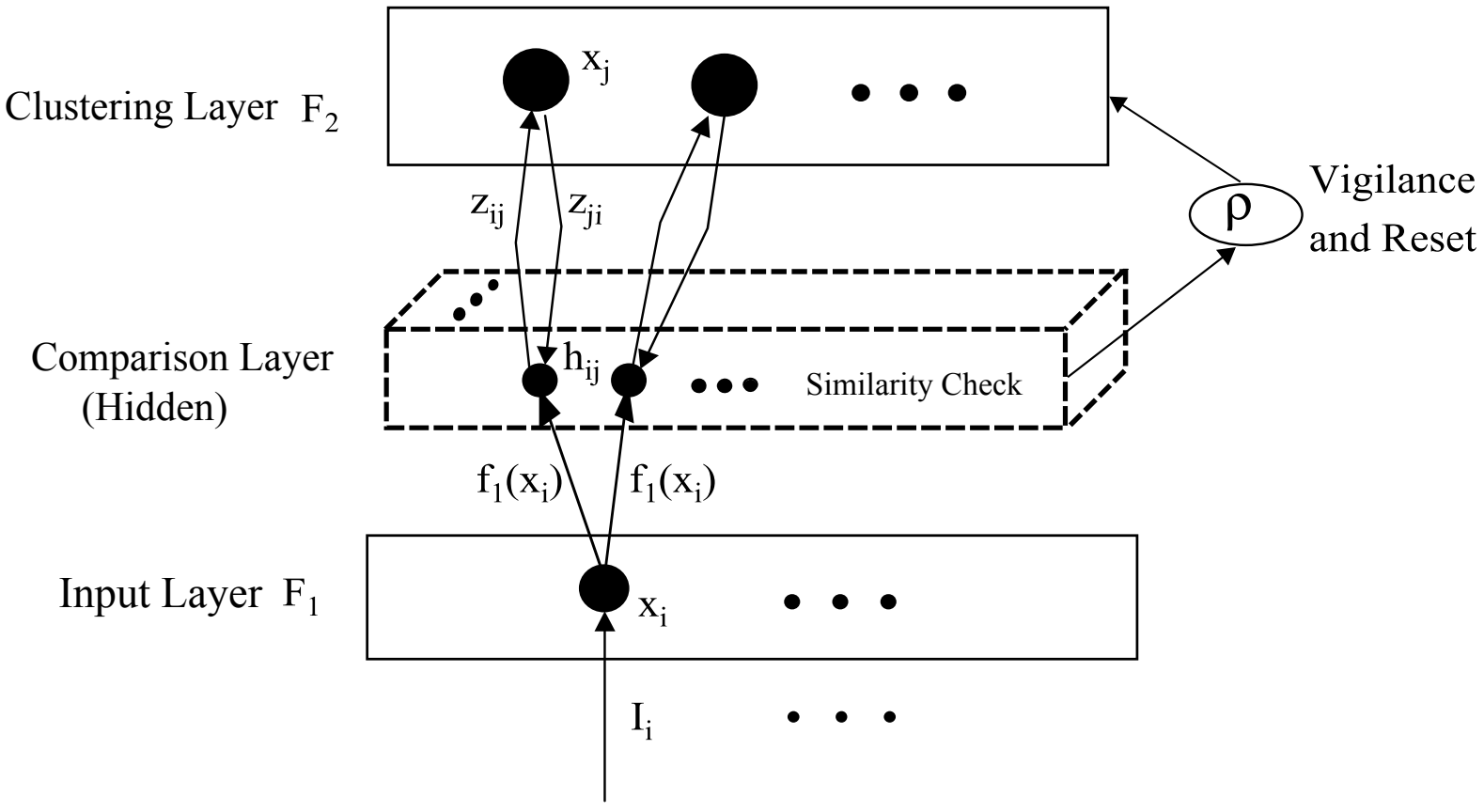
- Adaptive Resonance Theory (ART) Network (Carpenter, Grossberg and their collaborators, 1980's, 1990's);
- Features: clustering layer is a competitive net; top-down templates give the feature of the cluster.

The need for dimension-specific signal processing



- Data points form clusters only in low dimension subspaces: subspace (projective) clustering (IBM T. J. Watson group, 1999).
- Curse of dimensionality (feasibility) and difficulty of feature selection (reliability) combined yields the Feasibility-Reliability Dilemma

SOS for dimension-specific signal processing

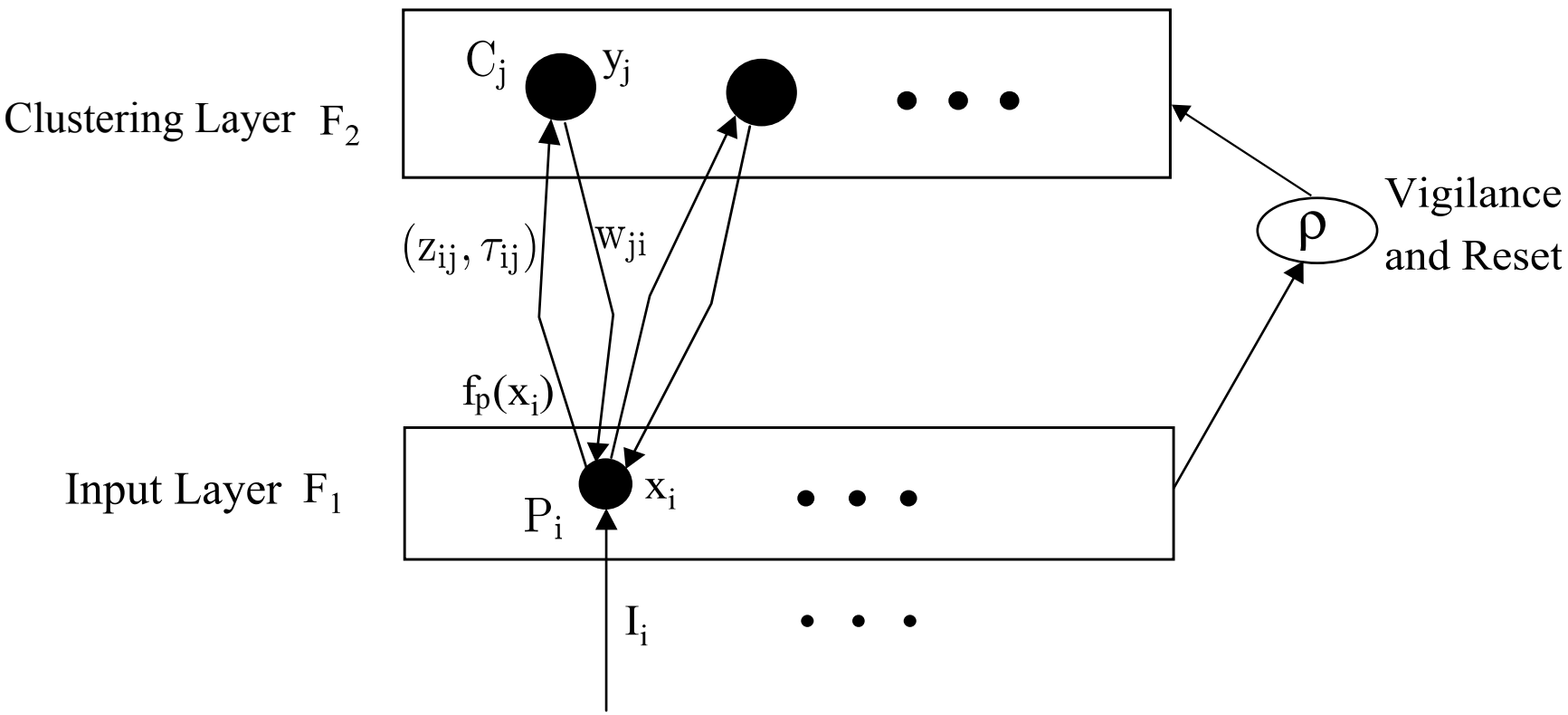


- PART Neural Network, Cao and Wu, Neural Networks (2002); IEEE NNs (2004).

Neurophysiological basis: PART-D Network

Wu, ZivariPiran, Hunter & Milton, Neural Computation, 2011.

- Dissimilarity driven transmission delay
- Delay induced signal loss



PART-D: differential equations with adaptive delays

$$\epsilon_p \frac{dx_i(t)}{dt} = -x_i(t) + I_i(t),$$
$$\epsilon_c \frac{dy_j(t)}{dt} = -y_j(t) + [1 - Ay_j(t)][f_c(y_j(t)) + T_j(t)] - [B + Cy_j(t)] \sum_{k \neq j, k \in \Lambda_2} f_c(y_k(t)) \text{(competitive net)}$$

$$T_j(t) = D \sum_{v_i \in F_1} z_{ij}(t) f_p(x_i(t - \tau_{ij}(t)) e^{-\alpha \tau_{ij}(t)}$$

$$\beta \frac{d\tau_{ij}(t)}{dt} = -\tau_{ij}(t) + E[1 - h_{ij}(t)] \textbf{(Delay Adaptation/Fitt's Law)}$$

$$h_{ij}(t) = h_\sigma(d(f_p(x_i(t)), w_{ji}(t))) l_\theta(z_{ij}(t)) \text{(similarity measure)}$$
$$\delta \frac{dz_{ij}(t)}{dt} = f_c(y_j(t))[(1 - z_{ij}(t)) L f_p(x_i(t - \tau_{ij}(t)) e^{-\alpha \tau_{ij}(t)} - z_{ij}(t) \sum_{k \neq i, k \in \Lambda_1} f_p(x_k(t - \tau_{kj}(t) e^{-\alpha \tau_{kj}(t)}]$$
$$\gamma \frac{dw_{ji}(t)}{dt} = f_c(y_j(t))[-w_{ji}(t) + f_p(x_i(t - \tau_{ij}(t)) e^{-\alpha \tau_{ij}(t)})] \text{(learning)}$$

Part-D Dynamics: Performance

Theorem We can choose small ϵ_p, ϵ_c and δ so that:

(i) *Inhibition of Non-Candidate Neurons*: *For* $j \neq J$ *and* $t \geq 0$, $y_j(t) < \eta_c$ *and* $f_c(y_j(t)) = 0$;

(ii) *Sustained Excitation of the Candidate Neuron*: *There exists* $\Gamma > 0$ *such that* $y_J(t) < \eta_c$ *and* $f_c(y_J(t)) = 0$ *when* $t < \Gamma$, *and* $y_J(t) \geq \eta_c$ *and* $f_c(y_J(t)) = 1$ *when* $t \geq \Gamma$;

(iii) *Invariance of Similarity*: *For any* $i \in \Lambda_p$, $j \in \Lambda_c$ *and* $t \geq 0$, $h_{ij}(t) = h_{ij}(0)$;

(iv) *Learning at Infinity*: *For any* $i \in \Lambda_p$ *and* $j \in \Lambda_c$ *with* $j \neq J$, $z_{ij}(t)$ *and* $w_{ji}(t)$ *remain unchanged for all* $t \geq 0$. *But* $\lim_{t\to\infty} w_{Ji}(t) = f_p(I_i)e^{-\alpha\tau^*_{ij}}$ *and*

$$\lim_{t\to\infty} z_{iJ}(t) = \begin{cases} 0 & \text{if } h_{iJ}(0) = 0, \\ \frac{L}{L+I_i} & \text{if } h_{iJ}(0) = 1, \end{cases}$$

where $I_i = \#\{k \in \Lambda_p \setminus \{i\}; h_{kJ}(0) = 1\}$.

Fast Learning & Thermal Dynamics in PC

Theorem We can choose small ϵ_p, ϵ_c and δ

(v) Fast Excitation: $\Gamma \in (0,1)$;

(vi) Fast Learning: Write $z_{ij}^{\epsilon_p,\epsilon_c,\delta}$ and $w_{ji}^{\epsilon_p,\epsilon_c,\delta}$ to indicate explicitly the dependence on $(\epsilon_p, \epsilon_c, \delta)$. Then we have (with $q = 1 - e^{-1/\gamma}$)

$$\lim_{\delta\to 0} z_{iJ}^{\epsilon_p,\epsilon_c,\delta}(1) = \begin{cases} 0 & \text{if } h_{iJ}(0) = 0, \\ \frac{L}{L+l_i} & \text{if } h_{iJ}(0) = 1, \end{cases}$$

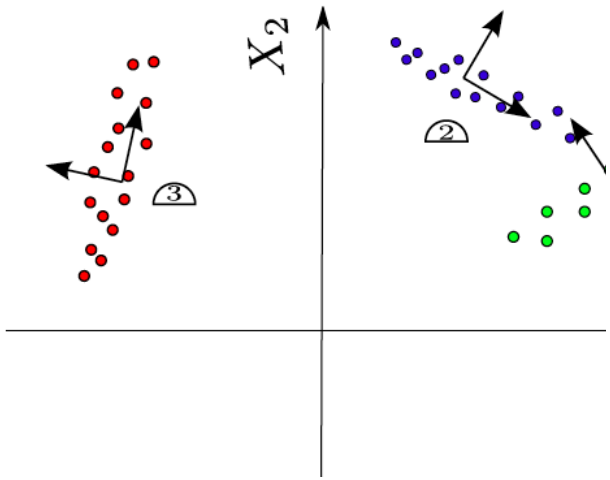
$$\lim_{\epsilon_p\to 0,\beta\to 0} w_{Ji}^{\epsilon_p,\epsilon_c,\delta}(1) = (1-q)w_{Ji}(0) + qf_p(I_i)e^{-\alpha\tau_{iJ}^*};$$

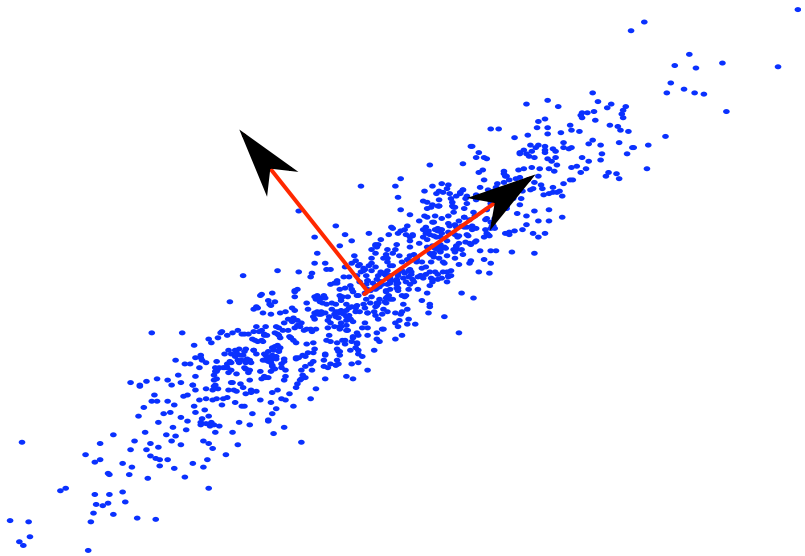
(vii) Convergence of Projective Subspace For any $j \in \Lambda_c$, define $D_j(t) = \{i \in \Lambda_p; I_\theta(z_{ij}(t)) = 1\}$. Then, as $\epsilon_p, \epsilon_c, \delta \to 0$, we have

$$\begin{aligned} &D_j(t) = D_j(0) \text{ for any } j \neq J; \\ &D_J(t_2) \subseteq D_J(t_1) \text{ if } t_2 \geq t_1 \geq 0; \\ &D_J(t) = D_J(1) \text{ for all } t \geq 1. \end{aligned}$$

Skewed Subspace Clustering & Minor Component Analysis

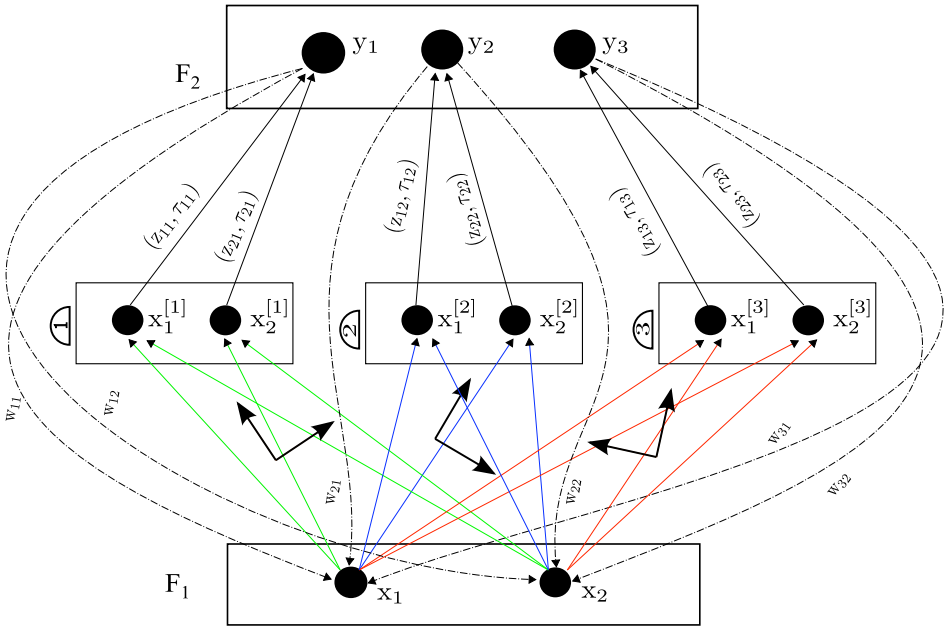
When input variables are not independent, PART-D fails





PART-D-MCA architecture

Wu and ZivariPiran, preprint.



$$\beta \frac{d\tau_{ij}(t)}{dt} = -\tau_{ij}(t) + E[1 - h_{ij}(t)], \quad t \geq 0, \ i \in \Lambda_p, j \in \Lambda_c,$$

$$\text{(PART-D)} \quad h_{ij}(t) = h_\sigma\big(d(f_p(x_i(t)), w_{ji}(t))\big) I_\theta(z_{ij}(t)).$$

$$\text{(PART-D-PCA)} \quad h_{ij}(t) = h_\sigma\big(d(x_i^{[j]}(t), 0)\big) I_\theta(z_{ij}(t)).$$

Part-D Dynamics: technical assumptions

Let

$$\tau_{ij}^* = E[1 - h_{ij}(0)],$$
$$T_j^* = D \sum_{i \in \Lambda_p} z_{ij}(0) f_p(I_i) e^{-\alpha \tau_{ij}^*}.$$

Assume that

$$\frac{L}{L+m-1} > \theta$$

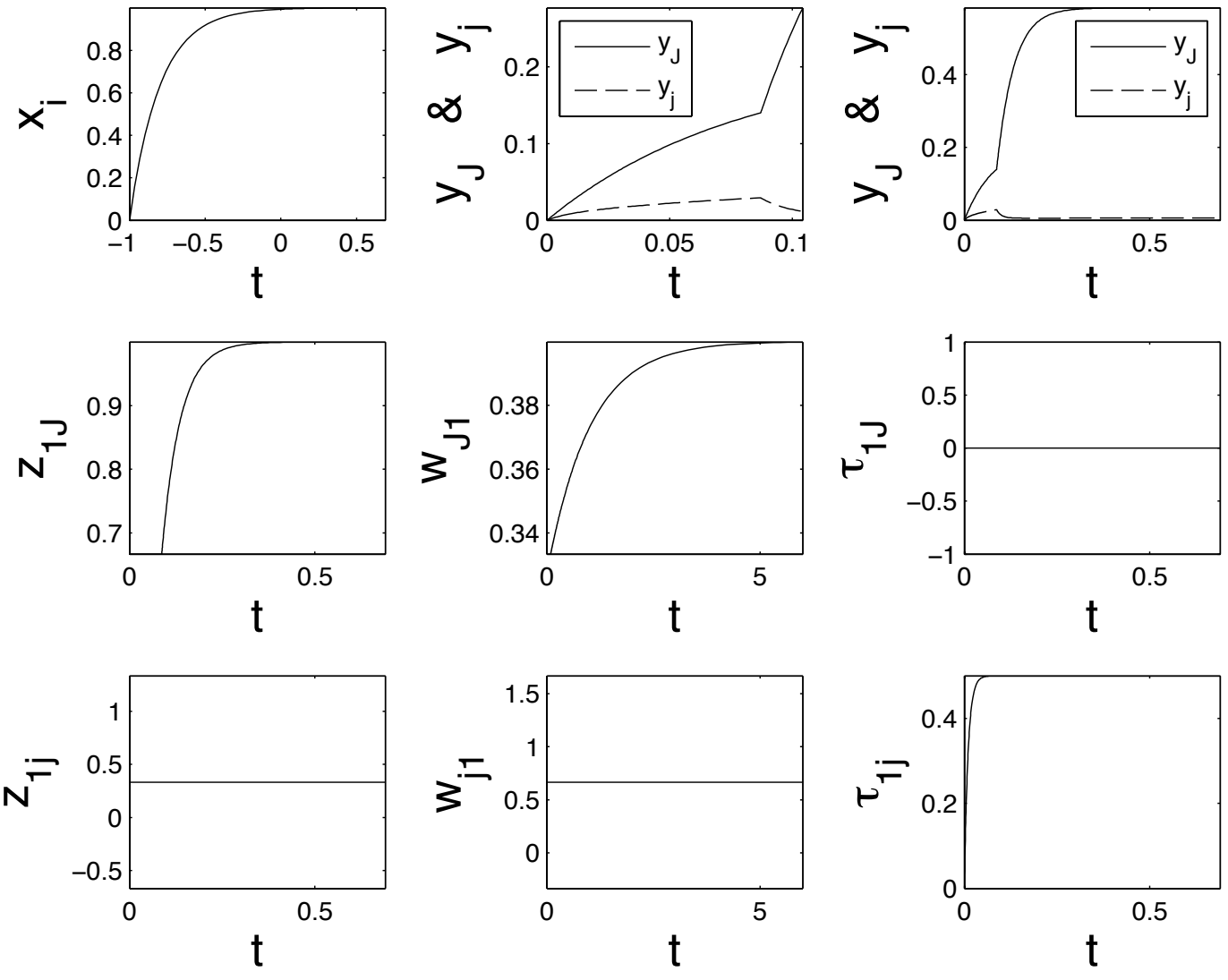
and that there exists $J \in \Lambda_c$ such that $T_j^* < T_J^*$ for all $j \in \Lambda_c \setminus \{J\}$. Let

$$M = \sum_{i \in \Lambda_p} z_{iJ}(0) f_p(I_i).$$

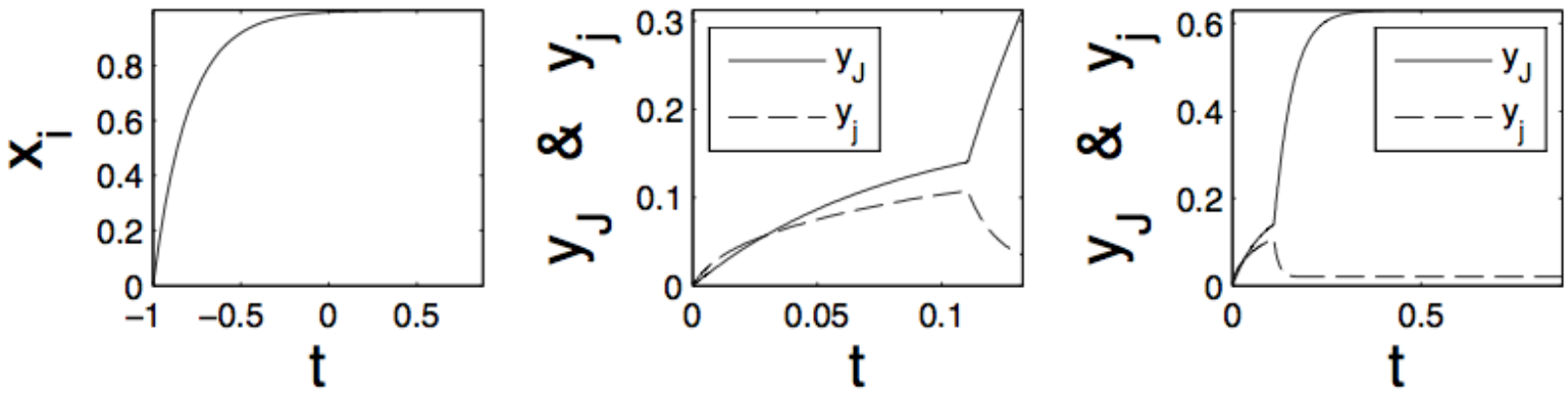
Assume further that there exists $T_{min} > 0$ so that

$$\frac{DL}{L+m-1} f_p(I_i) > T_{min}, \ i \in \Lambda_p,$$
$$T_J^* > T_{min} + DMe^{-\alpha E},$$
$$\frac{1+T_j^*}{2+T_j^*+C} < \eta_c < \min\{\frac{T_J^*}{1+T_J^*}, \frac{1+T_{min}}{2+T_{min}}\}, \ j \in \Lambda_c \setminus \{J\}.$$

PART-D Dynamics: Illustration



PART-D Dynamics: transient oscillation



- M. Eichmann, PhD Dissertation, Justus-Liebig University in Giessen, 2006); Jan Sieber, preprint, 2010;
- Q.Hu & J.Wu, JDE(2010); JDDE(2010); Q.Hu, J.Wu & X.Zou, SIAM J. Math. Anal., 2012.

Remarks on Adaptive Delay for Clustering

- Data clustering from dynamical systems point of view;
- Projective clustering based on adaptive delay: self-organization of transmission delays is an important, but under-recognized, mechanism for learning;
- Mathematical theory for delay adaption is still non-existant, except:
  - D. Beamish, S.A. Bhatti, C.S. Chubbs, I.S. MacKenzie, J. Wu and Z. Jing, Biological Cybernetics (2009, 2008);
  - J. Royal Soc. London Interface (2006); Neural Networks (2006);
  - Beamish D, Peskun C, Wu J , J. Math. Biology (2005).

Algorithms and applications

An effective algorithm based on the above results, specially the fast learning rules, has been developed. These algorithms consist of the following major steps:

- Input Processing and Select Output Signals from Input Layer;
- Activation, Inhibition, and Identification of a Potential Cluster;
- Confirmation, Vigilance and Reset;
- Fast Learning;
- Identification of Subspaces;
- Outliers collection.

The time cost of these algorithms is $O(mnNM)$, where m is the number of dimensions of data space, n is the number of clustering neurons, m is the number of all data points and M is the number of iterations.

Experiments on synthetic data

Example on a high dimensional synthetic data: The input data has 20,000 data points in a 100-dimensional space, which has 6 clusters generated in 20, 24, 17, 13, 16 and 28-dimensional subspaces respectively. The data points are presented in random order, and the clustering results can be reported as number of clusters found, dimensions found, centers of clusters found, and the contingency table of input clusters and output clusters.

Output\Input	1	2	3	4	5	6	Sums
1	5144	0	0	0	0	0	5144
2	0	1878	0	0	0	0	1878
3	0	0	4412	0	0	0	4412
4	0	0	0	2716	0	0	2716
5	0	0	0	0	2608	0	2608
6	0	0	1	0	0	1185	1186
Outliers	106	66	239	290	68	287	2056
Sums	5250	1944	4652	3006	2676	1472	20000

Experiments on online social network news aggregation

Digg.com: a content discovery and sharing application launched in 2004.

- Digg allowed people to vote web content up or down, called digging and burying, respectively. Users in Digg can share the content with other users who are connected to them by voting for or against the news.
- Data: from K. German (http://www.isi.edu/lerman/downloads/digg2009.html), who collected the information of stories on the Digg's front page over a period of a month in 2009. 3553 popular stories are voted for 3,018,197 times by 139,409 distinct users and on average, each story received about 850 votes.
- 2013 Fields-Mitacs Internship project by M. Freeman (Harvard), J. McVittie (Toronto), I. Sivak (Taras Shevchenko, Ukraine) and J. Yin (National University of Defense Technology, China).

Experiments on Digg.com

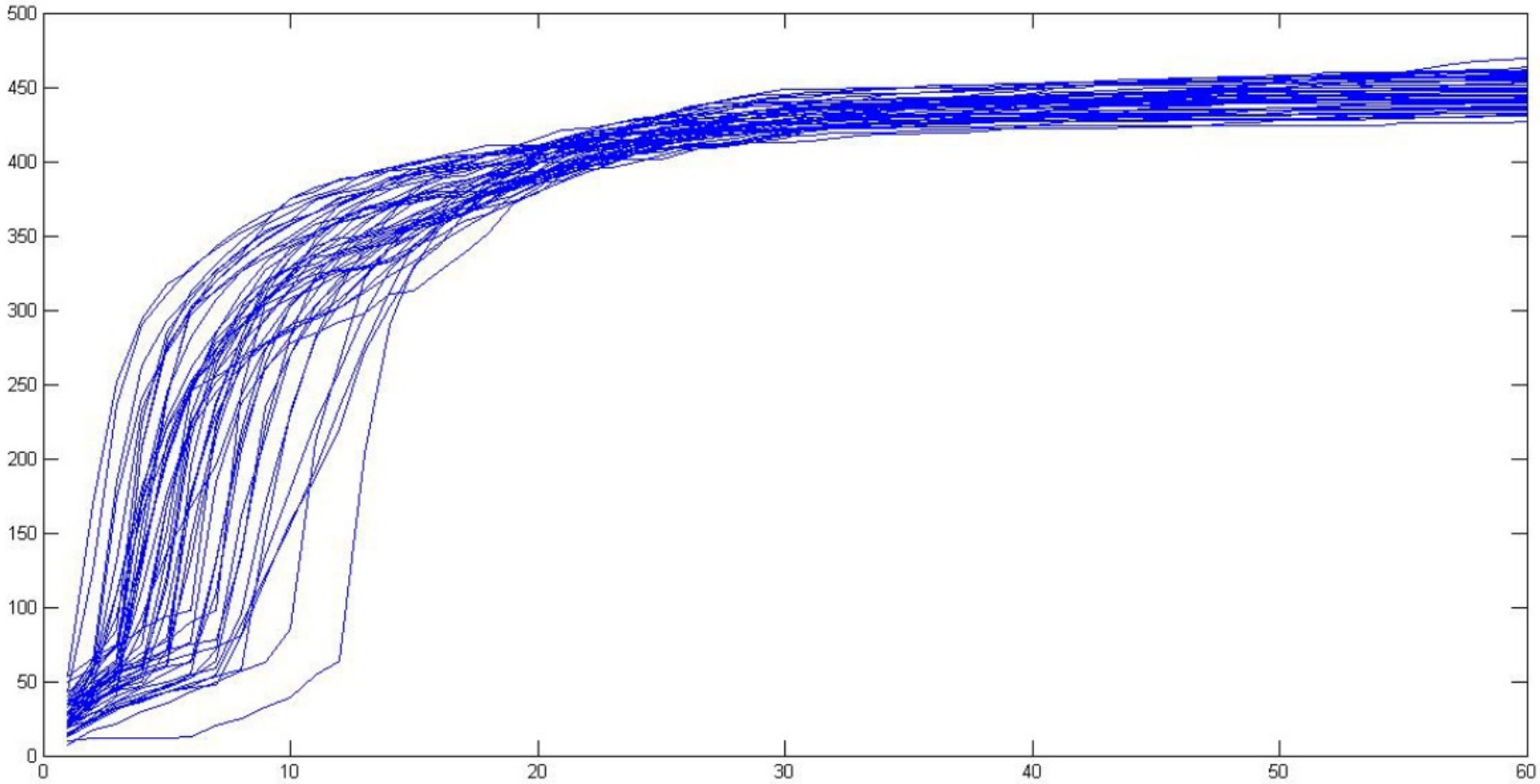


Figure: An example of projective clustering of the time series for the accumulated votes in the Digg networks. The cluster is formed based on the final size, and for those news reaching the equilibrium state within 20 ours since their release from the sources.

Experiments on Digg.com

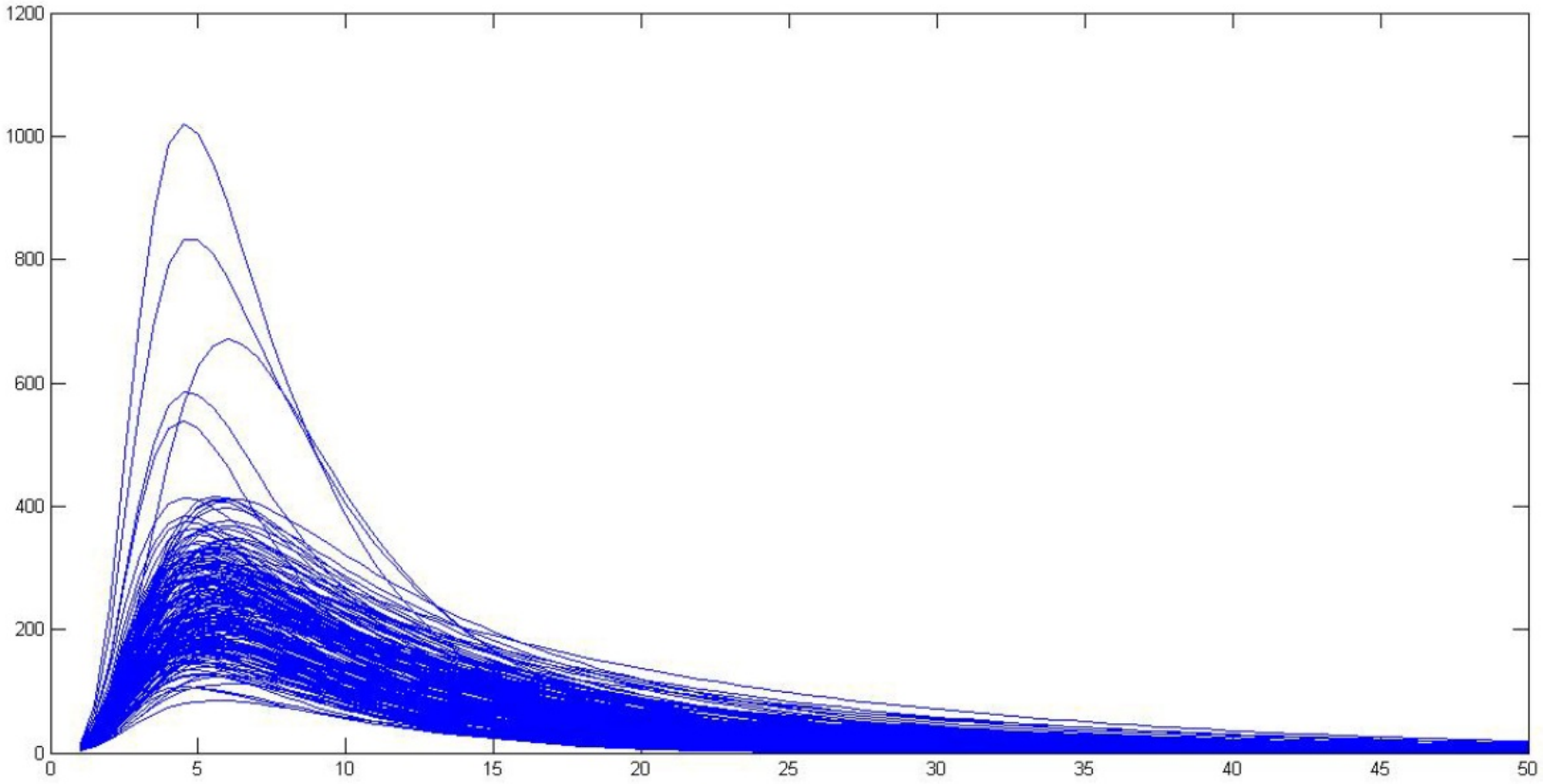


Figure: An example of projective clustering of the news in the Digg network, using features that are selected to reflect some "epidemic" nature of the number of new "influence votes."

Applications of PART

- Classification of neural spike trains (Hunter, Wu & Milton, IEEE on Decision and Control, 2008);
- Construction of robust prognostic predictors (Takahashi, Kobayashi & Honda, Bioinformatics, 2005);
- Diagnosis marker extraction, microarray data analysis for the extraction of subtype-specific genes (Takahashi, Nemoto & Yoshida, BMC Bioinformatics, 2006);
- Text mining (Chen & Chuang, Expert Systems, 2008);
- Clustering high-dimensional categorical data (Gan, Wu & Yang, IEEE Cong. Comp. Intelligence, 2006);
- Finding stock concurrence association rules (Liu, Huang, Lai & Ma, Neural Computing, 2009);
- Gene expression (Kawamura, Takahashi & Honda, J. Bioscience & Bioengineering, 2008).

Take Home Message

- Delay: omnipresent
- Delay: adaptive
- Delay transmission: not lossless
- Delay: not just destabilizing and *harmful* but critically important for effective signal processing.
- A dynamical system approach towards understanding the mechanism behind the remarkable clustering performance of our CNS is promising, and this approach also leads to quite powerful high dimensional data clustering algorithms and generates questions for neuroscience and mathematics.