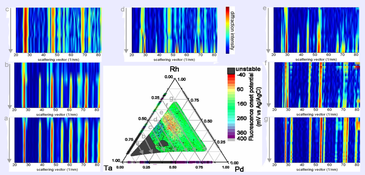
Constraint Reasoning and Kernel Clustering for Pattern Decomposition With Scaling

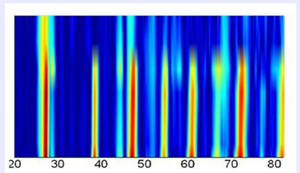
Ronan LeBras, Theodoros Damoulas, John M. Gregoire,
Ashish Sabharwal, Carla P. Gomes, R. Bruce van Dover

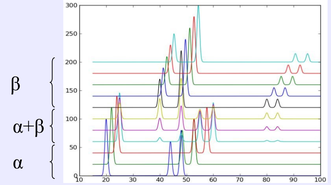
Motivation

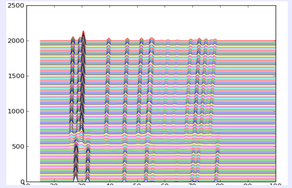
Material Discovery through Combinatorial Method: sputtering 3 metals (or oxides) onto a silicon wafer (which produces a *thin-film*) and using x-ray diffraction to obtain structural information about crystal lattice.

Input: Diffraction patterns $Y_1, \ldots, Y_n$ of n points on the thin-film.









Output: Set of k basis patterns (or *phases*) $X_1, \ldots, X_k$.
Weights $A_1, \ldots, A_n$ and shifts $B_1, \ldots, B_n$ of these basis patterns in the n points.

- Finding new products
- Finding product substitutes
- Understanding material properties (such as catalysts for fuel cell technologies)

Problem Definition

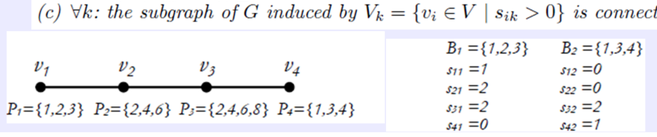
Input

- $G = (V, E)$ *be an undirected graph with* $V = \{v_1, \ldots, v_N\}$,
- $\mathcal{P} = \{P_1, \ldots, P_N\}$ *be a collection of* N *patterns over a finite set* $S \subseteq \mathbb{Q}^+$,
- $M \leq K \leq N$ *be positive integers, and* $\delta \geq 1$ *be a rational.*

Output

Determine whether there exists a collection $\mathcal{B}$ *of* K basis patterns *over* S *and* scaling factors $s_{ik} \in \{0\} \cup [1/\delta, \delta]$ *such that:*

(a) $\forall i$: $P_i = \bigcup_{k=1}^{K} s_{ik} B_k$
(b) $\forall i$: $|\{k \mid s_{ik} > 0\}| \leq M$
(c) $\forall k$: *the subgraph of* G *induced by* $V_k = \{v_i \in V \mid s_{ik} > 0\}$ *is connected*



	B_1 ={1,2,3}	B_2 ={1,3,4}
s_{11}	=1	s_{12} =0
s_{21}	=2	s_{22} =0
s_{31}	=2	s_{32} =2
s_{41}	=0	s_{42} =1

Problem Complexity

Assumptions: ***No experimental noise / Each B_k appears by itself in some v_i***

The problem can be solved* in **polynomial time**.

Assumptions: ***No experimental noise***

The problem becomes ***NP-hard*** (reduction from the *"Set Basis"* problem)

Assumptions used in this work: ***Experimental noise in the form of missing elements in P_i, and each B_k need not be sampled***

Constraint Programming Model

Advantage: Captures physical properties and relies on peak location rather than height.
Drawback: Does not scale to realistic instances; poor propagation if experimental noise.

Variables	Description	Type
p_{ki}	Normalizing peak for phase k in pattern c_i	Decision
a_{ki}	Whether phase k is present in pattern c_i	Auxiliary
q_k	Set of normalized peak locations of phase k	Auxiliary

$$a_{ki} = 0 \iff p_{ki} = 0 \qquad \forall\, 1 \leq k \leq K, 1 \leq i \leq n \tag{1}$$

$$1 \leq \sum_{s=1}^{K} a_{si} \leq 3 \qquad \forall\, 1 \leq i \leq n \tag{2}$$

$$p_{ki} = j \wedge \sum_{s=1}^{K} a_{si} = 1 \rightarrow q_k \subseteq r_{ij} \qquad \forall\, 1 \leq k \leq K, 1 \leq i \leq n, 1 \leq j \leq |c_i| \tag{3}$$

$$p_{ki} = j \wedge \sum_{s=1}^{K} a_{si} = 1 \rightarrow r_{ij} \subseteq q_k \qquad \forall\, 1 \leq k \leq K, 1 \leq i \leq n, 1 \leq j \leq |c_i| \tag{4}$$

$$P(k,k',i,j,j') \rightarrow \begin{cases} member(r_{ij}[j''], q_k) \\ \vee \\ member(r_{ij'}[j''], q_{k'}) \end{cases} \qquad \forall\, 1 \leq k < k' \leq K, 1 \leq i \leq n, 1 \leq j, j', j'' \leq |c_i| \tag{5}$$

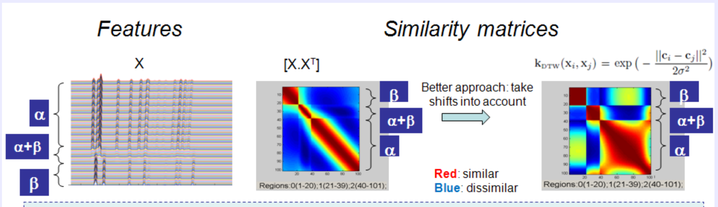
where $P(k,k',i,j,j')$ is the proposition: $p_{ki} = j \wedge p_{k'i} = j' \wedge \sum_{s=1}^{K} a_{si} = 2$.

$$p_{ki} = j \rightarrow p_{ki'} \neq j' \qquad \forall\, 1 \leq k \leq K, (i,j,i',j') \in \Phi \tag{6}$$

$$phaseConnectivity(\{a_{ki} | 1 \leq i \leq n\}) \qquad \forall\, 1 \leq k \leq K \tag{7}$$

Unsupervised Machine Learning

Advantage: Provides a data driven global picture, incorporates complex dependencies.
Drawback: Misses critical details (e.g. connectivity, linear phase shifts).



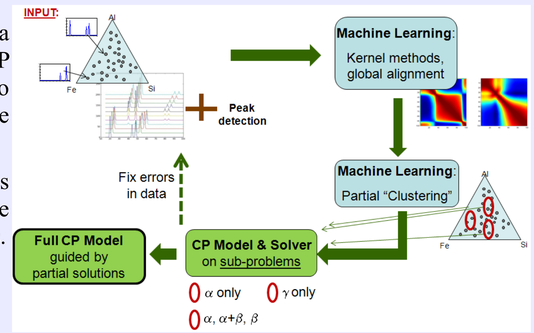
Method: K-means on Dynamic Time-Warping kernel

Goal: Select groups of samples that belong to the same **phase region** to feed the CP model, in order to extract the underlying phases of these sub-problems.

Integrating both approaches

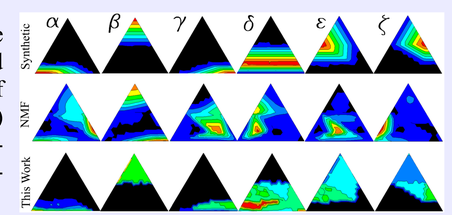
Each cluster represents a sub-problem that the CP procedure attempts to solve by extracting the underlying phases.

The full CP model is then solved using the pool of extracted phases.



Experimental Results: Example

As illustrated on the Al-Li-Fe diagram (right), our method outperforms the current state of the art (NMF, Long et al. '09) as it better captures the underlying phases and ensures connectivity in composition space.



Acknowledgments

NSF
IISI Intelligent Information Systems Institute Cornell
emc2 Energy Materials Center at Cornell
NSERC CRSNG