Pattern Decomposition with Complex Combinatorial Constraints: Application to Materials Discovery

Stefano Ermon
Computer Science Department
Stanford University
ermon@cs.stanford.edu

Ronan Le Bras
Department of Computer Science
Cornell University
lebras@cs.cornell.edu

Santosh K. Suram, John M. Gregoire
Joint Center for Artificial Photosynthesis
California Institute of Technology
{sksuram, gregoire}@caltech.edu

Carla P. Gomes, Bart Selman
Department of Computer Science
Cornell University
{gomes,selman}@cs.cornell.edu

Robert B. van Dover
Department of Materials Science and Engineering
Cornell University
rbv2@cornell.edu

Abstract

Identifying important components or factors in large amounts of noisy data is a key problem in machine learning and data mining. Motivated by a pattern decomposition problem in materials discovery, aimed at discovering new materials for renewable energy, e.g. for fuel and solar cells, we introduce CombiFD, a framework for factor based pattern decomposition that allows the incorporation of a-priori knowledge as constraints, including complex combinatorial constraints. In addition, we propose a new pattern decomposition algorithm, called AMIQO, based on solving a sequence of (mixed-integer) quadratic programs. Our approach considerably outperforms the state of the art on the materials discovery problem, scaling to larger datasets and recovering more precise and physically meaningful decompositions. We also show the effectiveness of our approach for enforcing background knowledge on other application domains.

Introduction

In recent years, we have seen an enormous growth in data generation rates in many fields of science (Halevy, Norvig, and Pereira 2009). For instance, in combinatorial materials discovery, scientists search for new materials with desirable properties by obtaining measurements on hundreds of samples in a single batch experiment using a composition spread technique (Ginley et al. 2005; Narasimhan, Mallapragada, and Porter 2007). Providing computational tools for automatically analyzing and for determining the structure of the materials formed in a composition spread is an important and exciting direction in the emerging field of Computational Sustainability (Gomes 2009). For example, this approach has been successfully applied to speed up the discovery of new materials with improved catalytic activity for fuel cell applications (Gregoire et al. 2010; van Dover, Schneemeyer, and Fleming 1998) and of oxygen evolution catalysts with applications to solar fuel generation (Haber et al. 2014). Long-term solutions to several sustainability issues in energy and transportation are likely to come from break-through innovations in materials (White 2012), and computer science can play a role and provide support to accelerate new materials discovery (Le Bras et al. 2011).

To accelerate the discovery of new materials, materials scientists have developed high-throughput experimental procedures to create composition-spread libraries of new materials, a process that can be intuitively understood as generating an enormous number of compounds in a single experiment by mixing different amounts of a small number of basic elements (Takeuchi, van Dover, and Koinuma 2002). The promising libraries are then characterized using X-ray diffraction to determine the underlying crystal structure and composition. Specifically, a set of X-ray diffraction signals are sampled at n different material compositions, each one corresponding to a different mixture of some basic elements. A key problem in materials discovery is called the *phase map identification problem*, defined as finding $k \ll n$ basic phases (basis functions that change gradually with composition, in terms of structure and intensity), such that all the n X-ray measurements can be explained as a mixture of the k basic phases. The decomposition is subject to physical constraints that govern the underlying crystallographic process.

The *phase map identification* is effectively a source separation or spectral unmixing problem (Berry et al. 2007) where the sources are the k non-negative, basic x-ray diffraction signals and each observation is a non-negative combination of these k sources. Therefore, a standard approach from the literature is non-negative matrix factorization (NMF) (Long et al. 2009). Nevertheless, this approach overlooks the physical constraints from the crystal formation. For example, it does not guarantee connectivity of the "phase regions" in the phase map, nor can it handle basis patterns that are slightly changing ("shifting") as the crystal lattice constants change. Recent development (Kusne et al. 2014), while not enforcing these constraints *per se*, have shown to be resilient to peak shifting for example. To obtain physically meaningful decompositions, researchers (Le Bras et al. 2011; Ermon et al. 2012) have looked at constraint programming formulations that can incorporate all the necessary constraints. The down side of these approaches is that they are fully discrete (they require a discretization of the data through peak detection) and they cannot directly deal with continuous measurement data. On the other side of the spectrum with respect to NMF, the unsupervised nature is lost, and scalability becomes an issue as, for example, in a fully discrete problem there is no notion of gradient anymore. In addition, these approaches are not robust to the presence of noise in the data, as noise considerably impacts the efficiency of their filtering and propagation mechanisms.

In this work, we aim to achieve the best of both worlds by bridging the previous approaches and providing a new hybrid formulation, where we integrate additional domain knowledge as additional constraints into the basic NMF approach. We introduce CombiFD, a novel pattern decomposition framework that allows the specification of very general constraints. These include constraints used with some matrix factorization and clustering approaches (such as non-negativity or partial labeling information) as well as more general ones that require a richer representation language, powerful enough to capture more complex, combinatorial dependencies. For example, we show how to encode complex a-priori scientific domain knowledge by specifying combinatorial dependencies on the variables imposed by physical laws. We also propose a novel solution technique called AMIQO (**A**lternating **M**ixed **I**nteger **Q**uadratic **O**ptimization), that involves the solution of a sequence of (mixed-integer) quadratic programs. We solve these optimization problems using state-of-the-art combinatorial optimization techniques. Overall, our constrained factorization algorithm clearly outperforms previous approaches: it scales to large real world datasets, and recovers physically meaningful and significantly more accurate interpretations of the data when prior knowledge is taken into account.

Framework

Given n data points $a_i \in \mathbb{R}^m$, each one represented by an m-dimensional vector of real-valued features, we represent the input data compactly as a matrix $A \in \mathbb{R}^{m \times n}$, each column corresponding to a data point and each row to a feature. In the context of the *phase map identification*, A corresponds to the n observed X-ray diffraction patterns, each of them as a vector of m scattering intensity values.

We are interested in low-dimensional representations which can approximate the input data A by identifying its essential components or factors. Namely, for a given number k of basic phases, we want to approximate A as $A \approx WH$, where $W \in \mathbb{R}^{m \times k}$ represents k basic phases (or phase patterns) and $H \in \mathbb{R}^{k \times n}$ the combination coefficients at each data point. This problem belongs to the family of low-rank approximation problems, an important research theme in machine learning and data mining with numerous applications, including source separation, denoising, compression, and dimensionality reduction (Berry et al. 2007).

Low-Rank Approximation

Given a non-negative matrix $A \in \mathbb{R}^{m \times n}$ and a desired rank $k < \min(n, m)$, we seek matrices $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$, $H, W \geq 0$ that give the best low rank approximation of A, i.e. $A \approx WH$. Typically this is formulated as the following optimization problem:

$$\min_{W,H} ||A - WH||_2 \tag{1}$$

where $W \in \mathbb{R}^{m \times k}$ is a matrix of *basis vectors* or *patterns* and $H \in \mathbb{R}^{k \times n}$ the *coefficient* matrix. The symbol $|| \cdot ||_2$ indicates (entry-wise) Frobenius norm.

This basic problem can be solved by the so-called truncated Singular Value Decomposition (SVD) approach, which produces the best approximation in terms of squared distance. It can be computed efficiently and robustly, obtaining a representation where data points can be interpreted as linear combinations of a small number of basis vectors.

In many applications input data are non-negative, for instance representing color intensities, word counts (Tjioe et al. 2008) or x-ray scattering intensities in our motivating application. The basis vectors computed with an SVD, however, are generally not guaranteed to be non-negative, and this leads to an undesirable *lack of interpretability*. For example, it is not possible to interpret an image as the superposition of simpler patches, or an X-ray diffraction pattern as the composition of several basic compounds when obtaining negative values for some of the basis vectors or the coefficients. To overcome this limitation, researchers have introduced the NMF approach, which explicitly enforces non-negativity constraints on the basis vectors and coefficients.

While non-negativity is a very common constraint in many domains, in some applications we have additional valuable *a priori* information on the features (each feature corresponds to a row of A and W). For instance, we might also know *a priori* an upper bound on the value of some features, or that two Boolean features are incompatible, or that non-negativity only holds for a subset of the features. In particular, in our materials science domain basis, vectors (patterns) correspond to chemical compounds and their underlying crystal structures. For chemical systems in thermodynamic equilibrium, the compositional variation of the concentration and lattice parameters of each compound follow well-defined rules, from which constraints on the coexistence and variation of basis patterns can be defined (Le Bras et al. 2011; Ermon et al. 2012). While some constraints such as non-negativity can be individually enforced in some approaches, others such as connectivity and the complex rules defining shifting basis patterns (see below for details) have not been considered before. To the best of our knowledge, there is no general factor analysis framework that can handle the *combination* of all these constraints. This motivates the definition of the following general pattern decomposition subject to combinatorial constraints problem.

CombiFD: Pattern Decomposition with Combinatorial Constraints

Given a (general) matrix $A \in \mathbb{R}^{m \times n}$ and a desired rank $k < \min(n, m)$ we seek matrices $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$ that minimize $||A - WH||_p$ where $p \in \{1, 2\}$ and $|| \cdot ||_p$ is an entry-wise norm (e.g. $p = 2$ corresponds to the Frobenius norm). Moreover, the factors need to satisfy an additional set of J linear inequality constraints, possibly requiring binary or integer variables. This is formalized as the following optimization problem:

$$\begin{aligned} \underset{W,H,x,b}{\text{minimize}} \quad & ||A - WH||_p = f(W,H) \\ \text{subject to} \quad & C[\text{vec}(W), \text{vec}(H), x, b]^T \leq d \\ & b_i \in \{0,1\},\ i \in [1,N]. \end{aligned} \tag{2}$$

where $x \in \mathbb{R}^M$ is a vector of additional real-valued variables, b is a vector of N binary variables, $d \in \mathbb{R}^J$, $C \in \mathbb{R}^{J\times(mk+nk+M+N)}$ and $\text{vec}(\cdot)$ denotes vectorization (stacking a matrix into a vector). That is, we have J linear inequalities involving the entries of W, H, x and b, with coefficients given by C and right-hand side given by d. As Integer Linear Programming is well known to be NP-complete, very general constraints can be encoded by appropriately choosing C and d. These additional (combinatorial) constrains are extremely useful to encode prior knowledge we might have about the domain. These include non-negativity ($W, H \geq 0$), upper bounds ($W_{i,j} \leq u$), sparsity (see Example 1), semi-supervised clustering (see Example 2) as well as many others. Other examples of intricate constraints can be found in the experimental section below.

Example 1: L_0 sparsity Suppose we want to explicitly formulate a sparsity constraint on the coefficient matrix. That is, we want to find W, H such that $A \approx WH$ and each column of H has at most S non-zero entries, i.e. $||H_i||_0 \leq S$ where H_i is the i-th column of H. For instance, in a topic modeling application where the basis vectors correspond to topics, this constraint ensures that each document can have at most S topics. This can be encoded as follows:

$$\begin{aligned} \min_{W,H,b} \quad & ||A - WH||_2 \\ \text{s.t.} \quad & b_{i,j} \geq h_{i,j}, \sum_j h_{i,j} = 1, \sum_i b_{i,j} \leq S \\ & b_{i,j} \in \{0,1\}\ i \in [1,k], j \in [1,n] \end{aligned} \tag{3}$$

which can be easily rewritten more compactly in the form (2) by selecting appropriate C and d (in this case, $M = 0$ and $N = kn$).

Example 2: Semi-Supervised Clustering As an example, in a semi-supervised clustering problem, we can easily include partial labeling information as constraints on H, e.g. enforcing Must-Link or Cannot-Link constraints (Liu and Wu 2010; Basu, Davidson, and Wagstaff 2008; Choo et al. 2013). Suppose we have prior information on P_{ML} pairs of data points $ML = \{(i_1, j_1), \cdots, (i_{P_{ML}}, j_{P_{ML}})\}$ that are known to belong to the same cluster (Must-Link) , and p_{CL} pairs of data points $CL = \{(i_1, j_1), \cdots, (i_{p_{CL}}, j_{p_{CL}})\}$ that are known not to belong to the same cluster (Cannot-Link).We encode this prior knowledge into our CombiFD framework as follows:

$$\begin{aligned} \min_{W,H,b} \quad & ||A - WH||_2 \\ \text{s.t.} \quad & W, H \geq 0, \textstyle\sum_j h_{i,j} = 1, \sum_i b_{i,j} \leq S \\ & b_{i,j} \geq h_{i,j},\ b_{i,j} \in \{0,1\}\ i \in [1,k], j \in [1,n] \\ & b_{i,i_s} = b_{i,j_s}\ i \in [1,k], (i_s, j_s) \in ML \\ & b_{i,i_s} + b_{i,j_s} \leq 1\ i \in [1,k], (i_s, j_s) \in CL \end{aligned} \tag{4}$$

Related work

Many low rank approximation schemes are available, including QR decomposition, Independent Component Analysis, truncated Singular Value Decomposition, and Nonnegative Matrix Factorization (Berry et al. 2007).

While these basic methods are unsupervised, there is a growing interest in incorporating prior knowledge or user guidance into these frameworks (Zhi et al. 2013). For example, in semi-supervised clustering applications, user guidance is often given by partial labeling information, which can be incorporated using hard constraints (Liu and Wu 2010; Basu, Davidson, and Wagstaff 2008; Choo et al. 2013). Typical constraints used in this case are Must-Link and Cannot-Link, enforcing that two data points must or cannot be in the same cluster, respectively. For example, (Hossain et al. 2010) present an integrated framework for clustering non-homogenous data, and show how to turn Must-Link and Cannot-Link constraints into dependent and disparate clustering problems. Recently, researchers have also considered interactive matrix factorization schemes for topic modeling that can take into account user feedback on the quality of the decomposition (Choo et al. 2013) (topic refinement, merging or splitting) and semi-supervised NMF with label information as constraints (Liu and Wu 2010). Constraint clustering (Basu, Davidson, and Wagstaff 2008) is another example of this approach. Alternatively, regularizations or penalty terms are also used to obtain solutions with certain desired properties such as sparsity (Hoyer 2004; Cai et al. 2011), convexity (Ding, Li, and Jordan 2010), temporal structural properties and shift invariances (Smaragdis 2004; Smaragdis, Raj, and Shashanka 2008).

Most of the work in the literature is however confined to a single type of constraints or simple conjunctions, which limits their usability. With CombiFD, we propose a new approach for finding a low dimensional approximation of some data (decomposition into basic patterns) that is able to incorporate not only existing types of constraints but also more complex logical combinations.

Constrained Factorization Algorithm

Solving the general CombiFD optimization problem (2) is challenging for two reasons. First, the objective function is not convex, hence minimization is difficult even in the presence of simple non-negativity constraints (Berry et al. 2007). Second, we are allowing a very expressive class of constraints, which can potentially specify very complex, intricate dependencies among the variables. Unfortunately, general nonconvex mixed-integer non-linear programming has not seen as much progress as their linear (MILP) and quadratic (MIQP) counterparts, and most approaches are either application specific or do not scale well. Even in the presence of simple constraints (as it is the case for NMF), the problem is rarely solved to optimality and in practice heuristic approaches are used. Yet, simple heuristic approaches such as multiplicative update rules (Lee and Seung 1999), which is one of the most widely used algorithms, do not apply to our case due to the integer variables. Similarly, projected gradient techniques cannot be directly applied here (Lin 2007). We thereofore introduce a new approximate technique called AMIQO which exploits the special structure of the problem and takes full advantage of advanced MIQP optimization techniques from the OR literature.

To solve the general CombiFD optimization problem (2), we introduce AMIQO (**A**lternating **M**ixed **I**nteger **Q**uadratic **O**ptimization) with pseudocode reported as Algorithm 1. AMIQO is an iterative two-block coordinate descent procedure enhanced with sophisticated combinatorial optimization techniques beyond the standard convex optimization methods. In fact, the key advantage of our CombiFD framework is that for either H or W fixed, problem (2) is a mixed-integer quadratic program.[1] Mixed-integer quadratic programs have been widely studied in the operations research literature, and we can leverage a wide range of techniques that have been developed and are implemented in state-of-the-art mixed-integer quadratic programming (MIQP) solvers such as IBM CPLEX. These integer programs do not have to be solved to optimality, and it is sufficient to improve the objective function with respect to the factorization found at the previous step (which is used to warm-start the search). Notice that when there are no binary variables ($N = 0$), the optimization problems in the inner loop of AMIQO correspond to standard quadratic programs that can be solved efficiently, even in the presence of (linear) constraints in the form (2), which are more general than non-negativity. AMIQO is inspired by the seminal work of Paatero and Tapper who initially proposed the use of a block coordinate descent procedure for NMF (Paatero and Tapper 1994), and was later followed upon by a number of researchers, including an unconstrained least squares version (Berry et al. 2007), and solution techniques based on projected gradient descent (Lin 2007), Quasi-Newton (Kim, Sra, and Dhillon 2007), and Active-set (Kim and Park 2008). However, our approach is novel in that it uses a mixed integer solver in each coordinate descent step, and is the only one that can take into account combinatorial constraints, guaranteeing feasibility of the solution at every iteration even in the presence of intricate combinatorial constraints.

Algorithm 1 AMIQO

```
  Find feasible $W^0, H^0, x^0, b^0$ for (2)    ▷ Use MIP solver
  for $j = 0, \cdots, t-1$ do
      $W^{j+1}, \tilde{x}^{j+1}, \tilde{b}^{j+1} \leftarrow \arg\min_{W,x,b} f(W, H^j)$ s.t. (2) and $H = H^j$    ▷ Use MIQP solver
      $H^{j+1}, x^{j+1}, b^{j+1} \leftarrow \arg\min_{H,x,b} f(W^{j+1}, H)$ s.t. (2) and $W = W^{j+1}$    ▷ Use MIQP solver
  end for
  return $W^t, H^t$
```

We summarize the properties of AMIQO with the following proposition:

Proposition 1. *Let W^j, H^j be as in Algorithm 1. If (2) is feasible, the following two properties hold: 1) For all j, $0 \leq j \leq t$, the optimization problems in the inner loop of the algorithm are feasible and (W^j, H^j, x^j, b^j) is feasible for (2). 2) The objective function $||A - W^j H^j||_p$ is monotonically non-increasing, i.e. $||A - W^j H^j||_p \geq ||A - W^{j+1} H^{j+1}||_p$*

Theorem 1. *AMIQO run on CombiFD problem (4) with $S = 1, ML = CL = \emptyset$ is equivalent to k-means.*

Proof. See Appendix for both proofs. □

Although the objective function is monotonically non-increasing, AMIQO is not guaranteed to converge to a global minimum. This is consistent with the hardness of problem (2). More specifically, the quality of the final solution found might depend on the initialization of W^0 and H^0. This issue is common to other standard matrix factorization algorithms, and several heuristic initialization schemes have been proposed to mitigate the issue (Albright et al. 2006). Nevertheless, in our experimental evaluation, the initialization did not play a major role, and we typically converged to the same solution, regardless of the initial conditions.

Experiments – Encoding domain knowledge as additional constraints

In order to show the generality of our approach, we provide experimental results on three application domains, with increasingly more complex constraints capturing a-priori domain knowledge. We start with semi-supervised clustering with partial labeling information, i.e. simple Must-Link and Cannot-Link constraints. We then consider another clustering problem, where we include more complex logical constraints on the features, describing higher-level biological knowledge. Finally we consider our motivating application: the phase map identification problem.

Clustering with partial labeling information

NMF has become a popular approach for clustering, with application domains ranging from document clustering (Shahnaz et al. 2006) and graph clustering (Kuang, Park, and Ding 2012) to gene expression data clustering (Tjioe et al. 2008). Cluster membership is determined by the coefficient matrix H, which reflects how each data point decomposes into the basis vectors.

There are several ways to obtain hard clustering assignments (binary indicators) from the coefficient matrix H (real valued). We follow (Ding, Li, and Peng 2008) and normalize the matrix H so that the entries can be interpreted as the posterior probability $p(c_s|d_j)$ that a data point j belongs to cluster s. Specifically, we let $D_W = diag(\mathbf{1}^T W)$ and estimate the cluster membership probability as $p(c_s|d_j) \propto [D_W H]_{sj}$. As a result, a data point j is assigned a cluster $s^*(j)$ such that $s^*(j) = argmax_s [D_W H]_{sj}$.

We consider a semi-supervised clustering task where we assume to have some prior information on the labels (equivalently, on the cluster assignment) of a subset of datapoints. Specifically, we assume to have information about pairs of data points, which should either belong to the same cluster (Must-Link) or not (Cannot-Link). This information is obtained using standard labeled datasets from the UCI repository (Bache and Lichman 2013) for which a ground truth clustering is known. To generate various amounts of prior

[1] For $p = 1$, it simplifies to a mixed-integer linear program.

knowledge, we randomly select P pairs of data points, using their labels to specify a MustLink or CannotLink constraint.

We compare our CombiFD formulation of the problem (4) with two previous approaches from the literature: CNMF (Constrained NMF) (Liu and Wu 2010) and NMFS (NMF-based semi supervised clustering) (Li, Ding, and Jordan 2007). The first approach is based on enforcing non combinatorial constraints on H, while the second approach captures the ML and CL constraints using penalty terms in a modified objective function which is then approximately minimized using multiplicative updates.

We report in Figure 1 the accuracy obtained using these methods as a function of the amount of supervision, i.e. the number of Must-Link or Cannot-Link constraints used. Accuracy is defined as in (Liu and Wu 2010) and corresponds to $AC = 1/n \cdot max_{\sigma} \sum_{i=1}^{k} |r_i \cap c_{\sigma_i}|$, where $\sigma : 1..k \to 1..k$ is a bijection mapping clusters r_i to ground-truth classes c_j. Namely, each cluster is assigned a label such that the labeling best matches the ground truth labels. Note that the accuracy can be computed efficiently using, for example, the Hungarian algorithm.

Results are averaged over 100 runs. We see that CombiFD significantly outperforms the competing techniques across all levels of prior knowledge. Intuitively, this is because by properly taking into account the combinatorial nature of the problem, CombiFD can automatically make logically sound inferences about the data: for example, we can take into account transitive closure (i.e., if a and b must link, and b and c must link, then a and c must link as well) and other logical implications. The deeper reasoning power and greater accuracy provided by AMIQO however involves a small computational overhead, with a typical runtime in the order of a couple of minutes for AMIQO versus a few tens of seconds for the competing techniques.

Clustering with more complex prior knowledge

In this experiment, we consider the Zoo dataset from UCI (Bache and Lichman 2013). This is a dataset of 101 animals, each one represented as a vector of 17 non-negative features (e.g. whether it has hair or not, whether it has feathers or not, or its number of legs). There are 7 class labels.

We solve the clustering problem using our CombiFD approach (2), where we enforce non-negativity as well as additional constraints capturing some well known biological facts. Specifically, we enforce the following constraints on the basis vectors using our CombiFD framework: $\neg(HasMilk \wedge HasEggs)$, $HasFeathers \to HasEggs$, and $\neg(HasFeathers \wedge HasHair)$. Since we are not aware of any other matrix factorization technique that can take into account this kind of complex, logically structured prior knowledge, we compare with standard NMF (problem (1)), which is a totally unsupervised technique.

Table 1 reports the accuracy (defined as before) and runtime of the two approaches averaged over 100 runs. The results show that the CombiFD approach, which incorporates a limited amount of logically structured prior knowledge, significantly improves the accuracy of standard NMF, while still running within seconds using AMIQO.

Approach	**Accuracy**		**Time (*seconds*)**	
	Avg.	Std. Dev.	Avg.	Std. Dev.
NMF	0.72	0.08	**1.84**	0.1
CombiFD	**0.82**	0.05	4.06	0.8

Table 1: Accuracy and runtime of NMF vs. CombiFD on the UCI Zoo dataset for 100 runs.

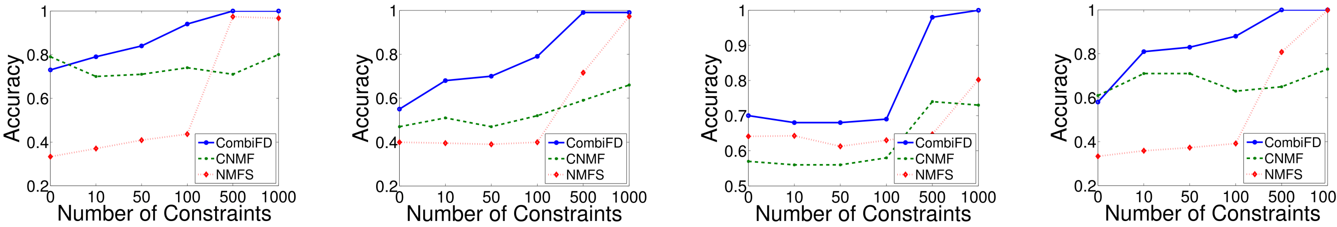
Spectral Unmixing for Materials Discovery

We first present how to incorporate complex constraints capturing some of the key physical laws that govern the data-generating process for the phase map identification problem.

Sparsity: According to the so-called Gibbs phase rule, under equilibrium conditions at fixed temperature and pressure, there can be mixtures of at most M phases occurring at each data point (chemical composition) in a library involving M basic elements. This is encoded as $||H_s||_0 \leq M$, for $s \in [1, k]$ as in (3).

Shifting: Each basis pattern may be slightly stretched by a different amount for each composition sample. Indeed, chemical alloying within a given compound may alter the crystal lattice constants, leading to a systematic shift (as a function of composition) of peak positions in the measured X-ray patterns. For isotropic lattice expansion and signals measured versus the scattering vector magnitude, the peak shifts are proportional to the peak positions, corresponding to a linear stretch of the signal. Therefore, we use Qk basis vectors, where k are free and $(Q-1)k$ are constrained to be shifted versions of the free basis vectors. This is encoded as follows: $A_{i,j} = interpolate(A_{\lfloor i/(1+\ell\gamma)\rfloor, zQ}, A_{\lceil i/(1+\ell\gamma)\rceil, zQ})$ for $j = zQ + \ell$, $z \in [0, k-1]$, $\ell \in [1, Q-1]$, where γ is a constant for the shifting granularity and $interpolate(x, y)$ denotes linear interpolation between x and y. By choosing Q to be no smaller than the ratio of the maximum peak shift and the minimum peak width (a known value for a given experiment), a linear sum of the Q mutually-constrained basis patterns can accurately model each instance of the shifting phase pattern.

Connectivity: The compositions at which a phase (or basis) is observed should form a connected region in composition space, and its lattice parameters should vary smoothly across the region. Hence we build a graph $G = (V, E)$ with n vertices (one vertex per data point) and edges between sample points that have similar compositions. Let $G(H_\ell)$ be the subgraph induced by the vertex set $V(H_\ell) = \{i : H_{\ell,i} > 0\} \subseteq V$ (the set of vertices where phase ℓ is used). We want to enforce $G(H_\ell)$ is connected for $\ell = 1, \cdots, k$. The first formulation we consider is flow based. Intuitively, this flow-based encoding defines a flow for each phase ℓ that can only pass through vertices where the phase is present. There is a source node that injects positive flow in the network, and there is some outgoing flow at every vertex where a phase is used. In order to satisfy flow conservation, there has to be a path from the source to any other node belonging to the same phase in the network. This constraint enforces connectivity but it can be expensive to include in the MIQP formulation when there is a large number of points.



(a) Iris (b) Wine (c) Ionosphere (d) Seeds

Figure 1: Accuracy as a function of the amount of prior knowledge. CombiFD (solid blue line) clearly outperforms competing techniques across all levels of supervision (number of constraints).

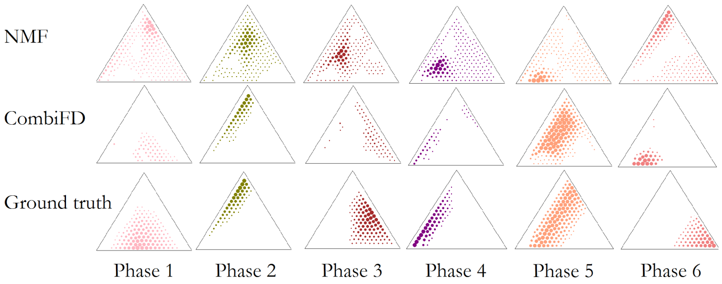


Figure 2: Results of NMF (top row), CombiFD (middle row), and ground truth (bottom row) for the Al-Li-Fe oxide system. In each row, each one of the 6 plots corresponds to the map of the concentration of a basis pattern (crystal phase). The size of each dot is proportional to the phase concentration estimated at that point from the coefficient matrix H. It can be seen that standard NMF overlooks the physical constraints, e.g. there are often mixtures of more than 3 basis patterns and the phase regions are disconnected. CombiFD recovers accurately all the phases except for the last one.

We therefore also consider the following variation. For all triples of points v_1, v_2, v_3 that lie on a straight line (in this order) in the composition space, and for every phase ℓ, we enforce the constraint that $h_{\ell,v_2} \geq \min\{h_{\ell,v_1}, h_{\ell,v_3}\}$. Although these constraints do not strictly enforce connectivity, they are simpler and often strong enough that the solution obtained is actually connected.

We consider synthetic data from (Le Bras et al. 2014), generated from the Aluminium(Al)-Lithium(Li)-Iron(Fe) oxide system and for which the ground truth is known. This dataset has 219 points and 6 underlying phases, where each phase has up to 42 diffraction peaks. We report in Fig. 2 the data interpretation obtained using regular NMF, CombiFD and the ground truth. On average, the number of iterations of AMIQO is about 8, with runtimes ranging from minutes to a few hours. We can see that although we cannot recover exactly the ground truth solution, our interpretation obtained considering prior domain knowledge is much more accurate. For example, using NMF the sparsity constraint is violated for many sample points. In terms of evaluation metric we adapt the previous definition of accuracy to the case of soft cluster assignment, to reflect the fact that a sample point might be assigned to multiple clusters. Namely, we define $AC = 1/n \cdot max_\sigma \sum_{i=1}^{k} |r_i \cap c_{\sigma_i}|/|r_i \cup c_{\sigma_i}|$. Note that this metric does not explicitly reflect the amount of violated combinatorial constraints, which would be interesting to evaluate in future work. We report the results in Table 2. Overall, CombiFD outperforms NMF, confirming the visual that incorporating prior knowledge indeed improves the accuracy of the interpretation. Compared with the previous constraint programming formulation (SMT) (Ermon et al. 2012), the CombiFD algorithm exhibits much better scalability, allowing us to quickly analyze more realistic sized datasets with hundreds of sample points in a few hours. In fact, SMT could not find solutions after 20 hours on all instances but one.

Finally, we show results obtained with CombiFD on a real dataset (Fe-Bi-V oxide system, Fig. 3). While the phase map is unknown for this system, Fig. 3 shows the excellent match between the phase 1 basis pattern from the CombiFD solution and the pattern of the known $Bi_4V_2O_{11}$ phase (Bergerhoff and Brown 1987). Chemical alloying of Fe into this compound, of the form $Bi_4V_{2-x}Fe_xO_{11-x}$, has also been observed, which may be related to the identification of this phase 1 basis pattern over a wide composition range in the library (Vannier et al. 2003). The results of Fig. 3 demonstrate the ability of CombiFD to identify well-connected phase regions from experimental data, and further investigations are underway to evaluate the presence of minority phases with weak signals in this composition library.

Conclusions

The ability to integrate complex prior knowledge into unsupervised data analysis approaches is a rich and challenging research problem, pervasive in a variety of domains, such as scientific discovery. In particular, the motivating appli-

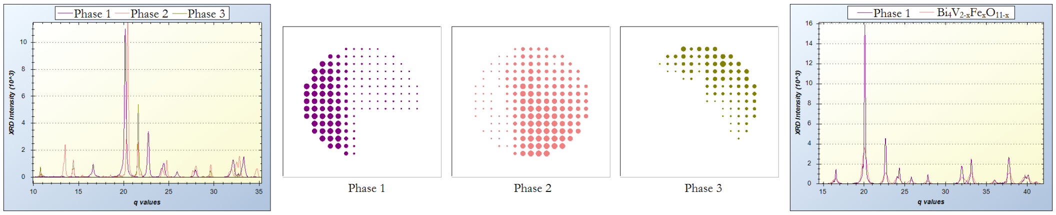


Figure 3: Results obtained with CombiFD and 3 phases (k=3) on the Fe-Bi-V oxide system. The X-ray diffraction data was collected at the SLAC National Accelerator Laboratory. The plot on the left shows the patterns of the 3 phases for low values of q (10-30 nm^{-1}), while the center 3 diagrams indicate where each phase appears. The right plot shows how one (recovered) phase matches a phase from the literature.

Parameters			Accuracy	
n	m	**NMF**	**CombiFD**	**SMT**
28	650	0.65	0.86	**0.96**
60	300	0.68	**0.77**	*t.-o.*
60	650	0.66	**0.75**	*t.-o.*
219	300	0.64	**0.73**	*t.-o.*
219	650	0.64	**0.73**	*t.-o.*
219	100	0.64	**0.75**	*t.-o.*
219	200	0.63	**0.73**	*t.-o.*

Table 2: Accuracy of NMF vs. CombiFD vs. SMT for different instances of the Al-Li-Fe oxide system. *t.-o.* indicates a time-out after 20 hours of CPU time.

cation of this work is the discovery of new fuel cell and solar fuel materials, thereby addressing pressing issues in sustainability such as the need for renewable and clean energy. We introduced CombiFD, a novel factor-based pattern decomposition framework that significantly generalizes and extends prior approaches. Our framework allows the specification of general constraints (including combinatorial ones), which are used to specify a-priori domain knowledge on the factorization that is being sought. These include traditional constraints such as non-negativity as well as more intricate dependencies, such as the ones coming from known phase behavior of chemical systems. We introduced a general factorization algorithm called AMIQO, based on solving a sequence of (mixed-integer) quadratic programs. We showed that AMIQO outperforms state-of-the-art approaches on a key problem in materials discovery: it scales to large real world datasets, it can handle complex, logically structured prior knowledge and by including prior knowledge into the model, we obtain significantly more accurate interpretations of the data. There are many directions to further extend our work, namely concerning representation formalism to capture other combinatorial constraints, with good performance/runtime trade-offs, as well as other algorithms to solve the combinatorial optimization problem.

Acknowledgments

Work supported by the National Science Foundation (NSF Expeditions in Computing award for Computational Sustainability, grant 0832782, NSF Inspire grant 1344201 and NSF Eager grant 1258330). Experiments were run on an infrastructure supported by the NSF Computing research infrastructure for Computational Sustainability grant 1059284. In collaboration with the Energy Materials Center at Cornell (emc2), an Energy Frontier Research Center funded by the U.S. Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences (Award No. DE-SC0001086). Research conducted at the Cornell High Energy Synchrotron Source (CHESS) is supported by the NSF and the National Institutes of Health/National Institute of General Medical Sciences under NSF award DMR-0936384. Based upon work performed by the Joint Center for Artificial Photosynthesis (JCAP), a DOE Energy Innovation Hub, supported through the Office of Science of the U.S. Department of Energy (Award No. DE-SC0004993). Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. DOE, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.

Appendix: Proofs

Proof of Proposition 1 The proof is by induction. Suppose (W^j, H^j, x^j, b^j) is feasible for (2) (the base case $j = 0$ holds by construction). It follows that (2) augmented with the additional constraint $H = H^j$ is still feasible, and therefore $\min_{W,x,b} f(W, H^0)$ subject to (2) and $H = H^j$ admits an optimal solution $W^{j+1}, \tilde{x}^{j+1}, \tilde{b}^{j+1}$. Since $W^{j+1}, H^j, \tilde{x}^{j+1}, \tilde{b}^{j+1}$ is feasible for (2), it follows that (2) augmented with the additional constraint $W = W^{j+1}$ is also feasible. Therefore, $\min_{H,x,b} f(W^{j+1}, H)$ subject to (2) and $W = W^{j+1}$ admits an optimal solution $H^{j+1}, x^{j+1}, b^{j+1}$. It also follows that $W^{j+1}, H^{j+1}, x^{j+1}, b^{j+1}$ is feasible for (2). Finally, since we are optimizing at every step, it follows that $||A - W^j H^j||_p \geq ||A - W^{j+1} H^j||_p \geq ||A - W^{j+1} H^{j+1}||_p$.

Proof of Theorem 1 The initial (feasible) values of H^0, b^0 can be seen as an initial (hard) assignment of data points to clusters, where data point i belongs to cluster s if $b^0_{s,i} = 1$. The optimal solution for $\min_{W,b} f(W, H^j)$ subject to (2) and $H = H^j$ is to choose each column W to be the centroid of the data points assigned to the corresponding cluster by b^j at iteration j, i.e. for each s set the s-th column of W to be $w_s = 1/(\sum_i b_{s,i}) \sum a_i b_{s,i}$ (non-negative because the data points are assumed to be non-negative). An optimal solution for $\min_{H,b} f(W^{j+1}, H)$ subject to (2) and $W = W^{j+1}$ can be found by assigning each data point i to the cluster whose centroid w_s is closest to data point a_i, i.e. setting $h_{s^*(i),i} = b_{s^*(i),i} = 1$ where $s^*(i) = \arg\min_s ||w_s - a_i||_2$. These operations exactly correspond to k-means clustering initialized with the hard cluster assignment given by b^0.

References

Albright, R.; Cox, J.; Duling, D.; Langville, A.; and Meyer, C. 2006. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Technical report, NCSU Tech Report.

Bache, K., and Lichman, M. 2013. UCI ML repository.

Basu, S.; Davidson, I.; and Wagstaff, K. 2008. *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press.

Bergerhoff, G., and Brown, I. 1987. Crystallographic databases. *International Union of Crystallography, Chester* 77–95.

Berry, M. W.; Browne, M.; Langville, A. N.; Pauca, V. P.; and Plemmons, R. J. 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis* 52(1):155–173.

Cai, D.; He, X.; Han, J.; and Huang, T. S. 2011. Graph regularized nonnegative matrix factorization for data representation. *PAMI, IEEE Transactions on* 33(8):1548–1560.

Choo, J.; Lee, C.; Reddy, C. K.; and Park, H. 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *Visualization and Computer Graphics, IEEE Transactions on* 19(12):1992–2001.

Ding, C. H.; Li, T.; and Jordan, M. I. 2010. Convex and semi-nonnegative matrix factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions.* 32(1):45–55.

Ding, C.; Li, T.; and Peng, W. 2008. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Stat. & Data Analysis* 52(8):3913–3927.

Ermon, S.; Le Bras, R.; Gomes, C. P.; Selman, B.; and van Dover, R. B. 2012. Smt-aided combinatorial materials discovery. In *SAT 2012*, 172–185. Springer.

Ginley, D.; Teplin, C.; Taylor, M.; van Hest, M.; and Perkins, J. 2005. Combinatorial materials science. In *AccessScience*. McGraw-Hill Companies.

Gomes, C. P. 2009. Computational sustainability: Computational methods for a sustainable environment, economy, and society. *The Bridge* 39(4):5–13.

Gregoire, J. M.; Tague, M. E.; Cahen, S.; Khan, S.; Abruna, H. D.; DiSalvo, F. J.; and van Dover, R. B. 2010. Improved fuel cell oxidation catalysis in pt1-xtax. *Chem. Mater.* 22(3):1080.

Haber, J. A.; Cai, Y.; Jung, S.; Xiang, C.; Mitrovic, S.; Jin, J.; Bell, A. T.; and Gregoire, J. M. 2014. Discovering ce-rich oxygen evolution catalysts, from high throughput screening to water electrolysis. *Energy Environ. Sci.* 7:682–688.

Halevy, A.; Norvig, P.; and Pereira, F. 2009. The unreasonable effectiveness of data. *Intelligent Systems, IEEE* 24(2):8–12.

Hossain, M. S.; Tadepalli, S.; Watson, L. T.; Davidson, I.; Helm, R. F.; and Ramakrishnan, N. 2010. Unifying dependent clustering and disparate clustering for non-homogeneous data. In *SIGKDD*, 593–602. ACM.

Hoyer, P. O. 2004. Non-negative matrix factorization with sparseness constraints. *Journal of ML Research* 1457–1469.

Kim, H., and Park, H. 2008. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM* 30(2):713–730.

Kim, D.; Sra, S.; and Dhillon, I. S. 2007. Fast newton-type methods for the least squares nonneg. matrix approx. problem. In *SDM*.

Kuang, D.; Park, H.; and Ding, C. H. 2012. Symmetric nonnegative matrix factorization for graph clustering. In *SDM*, 106–117.

Kusne, A. G.; Gao, T.; Mehta, A.; Ke, L.; Nguyen, M. C.; Ho, K.-M.; Antropov, V.; Wang, C.-Z.; Kramer, M. J.; Long, C.; et al. 2014. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Scientific reports* 4.

Le Bras, R.; Damoulas, T.; Gregoire, J. M.; Sabharwal, A.; Gomes, C. P.; and Van Dover, R. B. 2011. Constraint reasoning and kernel clustering for pattern decomposition with scaling. In *CP*. 508–522.

Le Bras, R.; Bernstein, R.; Gregoire, J. M.; Suram, S. K.; Gomes, C. P.; Selman, B.; and van Dover, R. B. 2014. A computational challenge problem in materials discovery: Synthetic problem generator and real-world datasets. In *AAAI*.

Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.

Li, T.; Ding, C.; and Jordan, M. I. 2007. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, 577–582. IEEE.

Lin, C.-J. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural comp.* 19(10):2756–2779.

Liu, H., and Wu, Z. 2010. Non-negative matrix factorization with constraints. In *Twenty-Fourth AAAI Conference*.

Long, C.; Bunker, D.; Karen, V.; Li, X.; and Takeuchi, I. 2009. Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization. *Rev. Sci. Instruments* 80.

Narasimhan, B.; Mallapragada, S.; and Porter, M. 2007. *Combinatorial materials science*. John Wiley and Sons.

Paatero, P., and Tapper, U. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5(2):111–126.

Shahnaz, F.; Berry, M. W.; Pauca, V. P.; and Plemmons, R. J. 2006. Document clustering using nonnegative matrix factorization. *Information Processing & Management* 373–386.

Smaragdis, P.; Raj, B.; and Shashanka, M. V. 2008. Sparse and shift-invariant feature extraction from non-negative data. In *ICASSP*, 2069–2072.

Smaragdis, P. 2004. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Indep. Component Analysis and Blind Signal Separation*. 494–499.

Takeuchi, I.; van Dover, R. B.; and Koinuma, H. 2002. Combinatorial synthesis and evaluation of functional inorganic materials using thin-film techniques. *MRS bulletin* 27(04):301–308.

Tjioe, E.; Berry, M.; Homayouni, R.; and Heinrich, K. 2008. Using a literature-based nmf model for discovering gene functional relationships. *BMC Bioinformatics* 9.

van Dover, R. B.; Schneemeyer, L.; and Fleming, R. 1998. Discovery of a useful thin-film dielectric using a composition-spread approach. *Nature* 392(6672):162–164.

Vannier, R.; Pernot, E.; Anne, M.; Isnard, O.; Nowogrocki, G.; and Mairesse, G. 2003. $Bi_4V_2O_{11}$ polymorph crystal structures related to their electrical properties. *Solid State Ionics* 157(1):147–153.

White, A. 2012. The materials genome initiative: One year on. *MRS Bulletin* 37(08):715–716.

Zhi, W.; Wang, X.; Qian, B.; Butler, P.; Ramakrishnan, N.; and Davidson, I. 2013. Clustering with complex constraints-algorithms and applications. In *AAAI*.