Hierarchically Clustered PCA and CCA via a Convex Clustering Penalty

Amanda M. Buch
Department of Psychiatry
Weill Cornell Medicine
Cornell University
New York, NY 10065
amb2022@med.cornell.edu

Conor Liston
Department of Psychiatry
Weill Cornell Medicine
Cornell University
New York, NY 10065
col2004@med.cornell.edu

Logan Grosenick
Department of Psychiatry
Weill Cornell Medicine
Cornell University
New York, NY 10065
log4002@med.cornell.edu

Abstract

We introduce an unsupervised learning approach that combines the truncated singular value decomposition with convex clustering to estimate within-cluster directions of maximum variance/covariance (in the variables) while simultaneously hierarchically clustering (on observations). In contrast to previous work on joint clustering and embedding, our approach has a straightforward formulation, is readily scalable via distributed optimization, and admits a direct interpretation as hierarchically clustered principal component analysis (PCA) or hierarchically clustered canonical correlation analysis (CCA). Through numerical experiments and real-world examples relevant to precision medicine, we show that our approach outperforms traditional and contemporary clustering methods on underdetermined problems ($p \gg N$ with tens of observations) and scales to large datasets (e.g., $N = 100,000$; $p = 1,000$) while yielding interpretable dendrograms of hierarchical per-cluster principal components or canonical variates.

1 Introduction

Explainable clustering across samples or individuals is important in countless real-world applications. In medical diagnoses, interpretable clustering of patients into distinct subtypes is increasingly important for personalized treatment [50, 45, 17, 49, 42, 2]. Due to the curse of dimensionality, similarity metrics (and thus clustering algorithm outcomes) degrade in high dimensions (i.e. the “$p > N$” setting common in medical imaging, genomics, and multomics, where we have $p$ correlated variables and a limited number of observations $N$). As a result, it is popular to use a two-stage procedure where high dimensional data are first embedded into a low-rank representation, and then clustered in this reduced dimension space. The mapping to the low-rank space (e.g., principal component loadings) are then often used to explain which variables are important (e.g., which differences in brain regions or genes relate to cluster differences [17, 14, 16, 23]).

Unfortunately, such two-stage procedures can lead to suboptimal and hard-to-explain results [6], as the embedding ignores important clustered structure in the data, which may harm the embedding. Further, as the embedding is agnostic to the underlying clusters, it may not provide a good space in which to separate the clusters (see Fig. 1). These issues motivate a need for joint clustering and embedding for such data. Further, such concerns extend to multiple data sets (multiview learning problems), where clustering and embedding has also frequently been approached as a two-step process where a low-rank, multiview embedding is obtained first and then input into a clustering algorithm (e.g., canonical correlation analysis followed by clustering, a common approach: [7, 8, 17, 18, 37]).

In the area of joint clustering and embedding, a number of pioneering approaches have emerged, including cluster-aware feature selection [56], CCA mixture models [20, 31], non-negative matrix factorization approaches (NMF) [21, 60, 63], and a variety of neural networks (e.g., [27, 57, 61, 35, 48, 3, 30]). However, these

Preprint. Under review.
approaches lack explainability as they rely on neural networks or complex formulations, prioritize clustering over interpretability, and perform poorly given limited data.

To develop an explainable and scalable formulation for joint clustering and embedding, we aimed to combine two widely used and interpretable methods: principal components analysis (PCA) and hierarchical clustering. To do so in a tractable way, we turned to convex clustering. For the past decade, this convex approach to clustering has reinvigorated clustering research, providing both firm theoretical footing and yielding strong empirical results within the framework of convex optimization. Since initial work using sparse, structured convex penalties to encourage fusion of variables into clusters [41, 25, 33], there has been a flurry of theoretical and algorithmic improvements [38, 10, 54, 51, 58, 55, 53]. A desirable characteristic of the convex clustering approach is the simple convex penalty it uses to encourage clustering of the optimization variable, as this penalty can be readily incorporated into more complex problems. This has recently led to new approaches to biclustering [1, 10], multi-view clustering [56], and supervised convex clustering [55].

To jointly fit PCA and cluster hierarchically, we exploit this flexibility and add a convex clustering penalty to the classical Frobenius loss formulation of the truncated singular value decomposition (tSVD) [19]. This results in a simple formulation for joint hierarchical clustering and principal component embedding. Although nonconvex, the objective function consists of only the tSVD loss term and a convex penalty, allowing for rapid
fitting to a path of hierarchical solutions. Further it is amenable to scalable optimization for fitting larger \( N \) than previously possible for convex clustering methods (while estimating within-cluster principal components).

As this method learns a smooth, continuous path of hierarchical clustering solutions as a penalty parameter \( (\lambda) \) is varied, we call it Pathwise Clustered Matrix Factorization (PCMF). The approach’s ability to cluster and fit principal components simultaneously allows it to handle the inherently uncentered structure of clustered components with different means, and to yield explainable loadings by correctly fitting components within clusters (Fig. 1g). It is highly interpretable as it returns both a dendrogram describing the cluster structure and the principal component scores and loadings per-cluster. Importantly, it can also be extended to joint multiview (CCA) embedding and clustering.

### 2 Main Contributions and Significance

1. **We introduce a method for joint hierarchical clustering and embedding (PCMF)** that adds a convex clustering penalty to the tSVD to yield joint hierarchically clustering and principal components within each cluster (Fig. 1). This is significant, as it is a new and uniquely explainable (e.g., yielding per-cluster principal component loadings), sensitive (effective for small \( N \)/small clusters), and scalable (able to fit large \( N \), e.g., \( N = 100,000 \)) approach to joint embedding and clustering.

2. **We present two algorithms to solve the PCMF problem along a path of solutions**, accelerate them using Algorithmic Regularization [58], and introduce a consensus approach that scales to much larger \( N \) than previously possible with convex clustering.

3. **We introduce a method to recover a dendrogram from the clustered matrix approximations**.

4. **We generalize our method to multiview learning**, introducing a novel approach and algorithm: Pathwise Clustered Canonical Correlation Analysis (P3CA).

5. **We demonstrate the efficacy of our approach relative to state-of-the-art methods on real data**.

### 3 Related Work: Convex Clustering

Traditional clustering algorithms such as K-means or hierarchical clustering aim to partition observations into more homogeneous subsets and are traditionally formalized as discrete optimization problems that are NP-hard. By relaxing the hard clustering constraint to a convex penalty [41], clustering can be reformulated as a convex optimization problem. In such “convex clustering” (also referred to as “clusterpath” or “sum-of-norms” clustering), the fitting procedure trades off approximating the data well with minimizing the sum of between-observation distances via a tuning penalty parameter, \( \lambda \). The number of clusters is indirectly controlled by \( \lambda \), and when solved along a path of \( \lambda \)s, convex clustering can exactly recover true data partitions among a mixture of Gaussians [33, 25, 29]. Further, this path of clustering solutions can be used to visualize hierarchical structure among the clusters as a dendrogram [58].

More explicitly, for data matrix \( X \in \mathbb{R}^{N \times p} \) with \( N \) observations in the rows and \( p \) variables in the columns, convex clustering solves the problem:

\[
\min_{X \in \mathbb{R}^{N \times p}} \frac{1}{2} \|X - \hat{X}\|_F^2 + \lambda \sum_{i<j} w_{ij} \|\hat{X}_i - \hat{X}_j\|_q.
\]  

(1)

The problem takes a “Lagrangian” form, with the penalty parameter \( \lambda > 0 \) continuously trading off a relaxed clustering inequality constraint (the sum of convex \( q \)-norms of differences between observations—typically \( q \in \{1, 2, \infty\} \)) with model fit to the data. Without this penalty, the problem is highly overparameterized and every row (sample/observation) has its own centroid. Weights \( w_{ij} > 0 \) can be added to increase the flexibility in modeling local structure in the row differences such as with a radial basis function (\( w_{ij} = \exp(-\gamma \|X_i - X_j\|_2^2) \) [25, 10]). However, multiplicative weights may provide better theoretical guarantees [29]. To speed up optimization and (sometimes significantly) improve results, weights may be constrained to only be nonzero for nearest neighbors [10, 56].

Multiple algorithms have been proposed to solve agglomerative convex clustering including subgradient descent [25], the stochastic incremental algorithm [39], the semismooth Newton-CG augmented Lagrangian method [52], and primal-dual operator splitting methods using the Alternating Direction Method of Multipliers (ADMM) or the Alternating Minimization Algorithm (AMA) [10]. Agglomerative convex clustering has been shown to converge and yield hierarchical dendrograms under certain satisfiable regularity conditions [43, 11, 58]. Recently, a warm-started ADMM approach—Algorithmic Regularization, which converges along a path of values of regularization penalty (\( \lambda \))—has been shown to strongly converge to ADMM solutions;
this enables feasible computation of dense convex clustering $\lambda$ paths by increasing the speed in which the algorithm converges by >100 times [58]. Multiple studies have evaluated the statistical properties of convex optimization and its extension to clustering [22, 4, 39, 13, 52, 28, 29, 32], but none have explored using the convex clustering penalty for hierarchical clustering of PCA and CCA, as we do here. Convex biclustering [12] is perhaps the most similar in spirit, but significantly, it does not allow the same variables to contribute differentially to more than one cluster of observations, as our pathwise clustered PCA and CCA approaches do.

4 Pathwise Clustered Matrix Factorization (PCMF)

4.1 Problem Formulation

Consider a data matrix $X \in \mathbb{R}^{N \times P}$ with $N$ observation in the rows and $P$ variables in the columns, rank $R \leq \min(N, P)$, and without loss of generality assume $X$ has been centered (see Appendix §1.2 for further discussion of uncentered effects). Among the most well-known and widely-employed algorithms is the truncated singular value decomposition (tSVD), where the rank-$R$ reconstruction to $X$ is given by $X = USV^T$, where $U^T U = I_N, V^T V = I_P$, $S = \text{diag}(s_1, \ldots, s_r)$ and $s_1 \geq s_2 \geq \cdots \geq s_R > 0$. It is a classic result [19] that for any $r \leq R$ the “truncated” SVD satisfies:

$$\hat{X} = U_r S_r V_r^T = \arg\min_{\hat{X} \in M(r)} \frac{1}{2} \|X - \hat{X}\|_F^2,$$

where $M(r)$ is the set of rank-$r \leq R$ matrices of $N$ rows and $P$ columns, $U_r$ and $V_r$ contain the first $r$ left and right singular vectors in their respective columns, and $S_r$ is square with the first $r$ ordered singular values on its diagonal.

This decomposition is ubiquitously employed throughout statistics, engineering, and life sciences for dimensionality reduction, denoising, and data modeling. It is frequently used for capturing directions of maximum variance or covariance in data, as it is the algorithmic basis of principal components analysis (PCA), total least squares regression (TLS), and canonical correlation analysis (CCA).

We aim to generalize the tSVD to allow clustering among the observations, such that different clusters contain different variance-maximizing tSVD approximations. To do this we follow the convex clustering literature and augment (2) with the following sum-of-norms regularization (with $q \in \{1, 2, \infty\}$):

$$\min_{\hat{X} \in M(r)} \frac{1}{2} \|X - \hat{X}\|_F^2 + \lambda \sum_{i<j} w_{ij} \|\hat{X}_i - \hat{X}_j\|_q.$$

Note that when $r = R$ (that is, if $\text{rank}(\hat{X}) = \text{rank}(X)$), this problem reduces to standard convex clustering (1). Expressing the matrix rank constraint $\hat{X} \in M(r)$ explicitly in terms of the tSVD, this optimization problem can be written:

$$\min_{\hat{X}, U_r, S_r, V_r} \frac{1}{2} \|X - \hat{X}\|_F^2 + \lambda \sum_{i<j} w_{ij} \|\hat{X}_i - \hat{X}_j\|_q$$

subject to $\hat{X} - U_r S_r V_r^T = 0, U_r^T U_r = V_r^T V_r = I_r, S_r = \text{diag}(s_1, \ldots, s_r)$, for $s_1 \geq s_2 \geq \cdots \geq s_r > 0$, which can be interpreted as a convex clustering problem with truncated SVD constraints. We next present two algorithmic approaches to solving this problem.

4.2 Solving PCMF with ADMM

First, we present an approach solving PCMF with ADMM. Because in most cases it is desirable for many weights $w_{ij}$ in the convex clustering penalty to be exactly zero [10], we can represent the relevant nonzero distances as a sparse graph $G$. We then introduce an auxiliary variable $G = D\hat{X} \in \mathbb{R}^{E \times N}$, where $D \in \mathbb{R}^{E \times n}$ is a sparse matrix containing the weighted pairwise distances defined by edges $E$. This allows us to rewrite the PCMF problem as:

$$\min_{\hat{X}, G, U_r, S_r, V_r} \frac{1}{2} \|X - \hat{X}\|_F^2 + \lambda \sum_{\ell \in E} w_\ell \|G_\ell\|_q$$

subject to $\hat{X} - U_r S_r V_r^T = 0, G - D\hat{X} = 0, U_r^T U_r = V_r^T V_r = I_r, S_r = \text{diag}(s_1, \ldots, s_r)$,
for \( s_1 \geq \cdots \geq s_r > 0 \), which yields a problem separable in its objective and penalty subject to (nonconvex) constraints—a common application for ADMM. Algorithm 1 shows the ADMM updates, using Algorithmic Regularization [58] along the \( \lambda \) path. ADMM solutions fit along a path of \( \lambda s \) benefit from “warm-starting” by initializing the next problem along the path at the previous solution. Algorithmic Regularization takes this to the extreme, shortening steps along the path and decreasing the number of ADMM iterations at each point to a small number (achieved by making \( K \) small in Algorithm 1). For an appropriately chosen step size, this has been proven to converge to the true path solutions and speed up the computation of path estimation by \( >100 \) times [58]. Critically, this improves computational feasibility since the algorithm requires solving over many path parameter (\( \lambda \)) values (see Appendix for derivation, convergence details, and computational complexity).

Algorithm 1 PCMF-ADMM

\begin{algorithm}
\begin{algorithmic}
\State \textbf{Input:} data \( X \), decreasing path \( \{ \lambda \} \), weights \( w \), pairwise distance matrix \( D \)
\State \textbf{Notation:} data mean, \( \hat{X} \), rank \( r \), iteration \( k \), norm \( q \in \{ 1, 2, \infty \} \), \( \rho \geq 1 \), proximal operator \( \text{prox} \hat{P}_{w,q}(G) \)
\State
\State 1: \( G^0 \leftarrow Z^0 \leftarrow DX; \hat{X} \leftarrow \hat{X}; (U^0_r, S^0_r, V^0_r) \leftarrow \text{SVD}(\hat{X}); L = \text{chol}(I + \rho I + \rho D^T D) \)
\State 2: for \( \lambda \in \{ \lambda \} \) do
\State 3: for \( k = 1, \ldots, K \) do
\State 4: \( \hat{X}^{k+1} \leftarrow L^{-T} L^{-1}(X + \rho D^T (G^k - Z^k) + \rho(U^k_r S^k_r V^k_r^T - Z^k_2)) \)
\State 5: \( G^{k+1} \leftarrow \text{prox} \hat{P}_{w,q}(G)(D \hat{X}^{k+1} + Z^k_2) \)
\State 6: \( (U^k_r + 1, S^k_r + 1, V^k_r + 1) \leftarrow \text{SVD}(\hat{X}^{k+1} + Z^k_2) \)
\State 7: \( Z^{k+1}_1 \leftarrow Z^k_1 + D^T \hat{X}^{k+1} - G^{k+1} \)
\State 8: \( Z^{k+1}_2 \leftarrow Z^k_2 + \hat{X}^{k+1} - U^{k+1}_r S^{k+1}_r V^{k+1} \)
\State 9: end for
\State 10: Save current path solutions: \( \hat{X}_\lambda \leftarrow \hat{X}^K \), \( G_\lambda \leftarrow G^K \), \( (U_r, S_r, V_r) \leftarrow (U^K_r, S^K_r, V^K_r) \)
\State 11: Initialize for next path solution: \( \hat{X}^0 \leftarrow \hat{X}^K \), \( G^0 \leftarrow G^K \), \( (U^0_r, S^0_r, V^0_r) \leftarrow (U^K_r, S^K_r, V^K_r) \)
\State 12: end for
\State 13: return pathwise solutions \( \{ \hat{X}_\lambda \}, \{ G_\lambda \}, \{ U_r, S_r, V_r \} \)
\end{algorithmic}
\end{algorithm}

### 4.3 Solving PCMF with a Penalized Alternating Least Squares Relaxation

Second, we present a Penalized Alternating Least Squares (PALS) approach that, when relaxed, provides more flexible fitting of per-cluster factors than the ADMM solution and will subsequently motivate our multiview approach. Without loss of generality, we consider the rank-1 version of this problem (which can be generalized to rank-\( r \)) using an appropriate deflation approach; see Mackey [34], Witten et al. [59]). Denoting the \( i \)th column vector of \( X \) as \( x_i = (X)_{ij} \) and defining penalty \( \hat{P}_{w,q}(u, v) = \sum_{(i,j) \in E} w_{ij} \| u_i v_i - u_j v_j \|_q \), we can write the rank-1 truncated SVD with a convex clustering penalty (see Appendix §4) as:

\[
\min_{s,u,v} \frac{1}{2} \sum_{i=1}^N \| x_i - su_i v_i \|_2^2 + \lambda \hat{P}_{w,q}(u, v) \quad \text{subject to} \quad \| u \|_2^2 = 1, \| v \|_2^2 = 1, s > 0. \quad (6)
\]

To introduce the overparameterization necessary for convex clustering (without loss of generality) we center and scale \( X \), set \( s = 1 \), and replace the single vector \( v \) with a matrix \( V \in \mathbb{R}^{p \times N} \) with column vectors \( v_i = V_i \) (denoting the the set of these column vectors as \( \{ v \} \), \( i = 1, \ldots, N \))—this allows each observation to potentially be its own cluster in the limit \( \lambda \rightarrow 0 \). We define \( P_{w,q}(u, V) = \sum_{(i,j) \in E} w_{ij} \| u_i v_i - u_j v_j \|_q \) and arrive at the overparameterized problem:

\[
\min_{u,V} \sum_{i=1}^N \| x_i - u_i v_i \|_2^2 + \lambda P_{w,q}(u, V) \quad \text{subject to} \quad \| u \|_2^2 = 1, \| v \|_2^2 = 1, i = 1, \ldots, N. \quad (7)
\]

To decrease computation time and improve the flexibility of fitting per-cluster factors, we remove the cross terms in the penalty; we relax the problem formulation by replacing \( P_{w,q}(u, V) \) with \( Q_u^w(u) = \sum_{(i,j) \in E} w_{ij} | u_i - u_j | \) and \( Q_V^w(V) = \sum_{(i,j) \in E} w_{ij} \| v_i - v_j \|_q \). Next, using fixed values from iterate \( k \),
We next extend PCMF to multiview learning, where we aim to learn the covariance structure between clustered observations measured using multiple data views (i.e., fitting canonical correlation analysis or CCA within clusters). We will subsequently refer to a convex clustering solver as CONVEXCLUSTER. Here we implement our CONVEXCLUSTER as an ADMM algorithm that can be scaled using consensus ADMM (Appendix §8); however, other convex clustering solvers could be used instead. The CONVEXCLUSTER algorithm and derivation of the PCMF-PALS problem is given in the Appendix.

Algorithm 2 Pathwise Clustered Canonical Correlation Analysis (P3CA)

| Input: data $(X, Y)$, decreasing path $\{\lambda\}$, weights $w$, norm $q \in \{1, 2, \infty\}$ |
| Notation: iteration $k$, data means $(X, Y)$, $v_i = V_i$, $x_i = (X_i)^T$, $\bar{y}_i = (\bar{Y}_i)^T$, $\rho \geq 1$ |
| 1: $U \leftarrow X, V \leftarrow Y$ |
| 2: for $\lambda \in \{\lambda\}$ do |
| 3: for $k = 1, \ldots, K$ do |
| 4: $\bar{x}_i^{k+1} \leftarrow \Sigma_i v_i^k, (\Sigma_i = X_i Y_i^T \in \mathbb{R}^{P \times P})$ for $i = 1, \ldots, N$ |
| 5: $u_i^{k+1} \leftarrow \text{CONVEXCLUSTER}(\bar{x}_i^{k+1}, U^k, \lambda, w, q)$ |
| 6: $v_i^{k+1} \leftarrow \text{prox}_{\|_{2}^{2}}(u_i^{k+1})$ for $i = 1, \ldots, N$ |
| 7: $\bar{y}_i^{k+1} \leftarrow \Sigma_i u_i^{k+1}, (\Sigma_i = Y_i X_i^T \in \mathbb{R}^{P \times P})$ for $i = 1, \ldots, N$ |
| 8: $v_i^{k+1} \leftarrow \text{prox}_{\|_{2}^{2}}(v_i^{k+1})$ for $i = 1, \ldots, N$ |
| 9: end for |
| 10: Save path solutions: $U_i^K \leftarrow u_i^{KT}; V_i^K \leftarrow v_i^{KT}$ for $i = 1, \ldots, N$; $(U_{\lambda}, V_{\lambda}) \leftarrow (U^K, V^K)$ |
| 11: Initialize for next path solution: $(U^0, V^0) \leftarrow (U^K, V^K)$ |
| 12: end for |
| 13: return pathwise solutions $\{U_{\lambda}\}, \{V_{\lambda}\}$ |

5 Pathwise Clustered CCA (P3CA): Multiview Learning PCMF

We next extend PCMF to multiview learning, where we aim to learn the covariance structure between clustered observations measured using multiple data views (i.e., fitting canonical correlation analysis or CCA within clusters). To do so we introduce the overparameterized pathwise clustered canonical correlation analysis (P3CA) optimization problem (recall $v_i = V_i$ are column vectors of $V \in \mathbb{R}^{P \times N}$). We have data matrices $X \in \mathbb{R}^{N \times P}, Y \in \mathbb{R}^{N \times P}$ and variables $u_i \in \mathbb{R}^{P \times 1}, v_i \in \mathbb{R}^{P \times 1}$, and we define $\Sigma_i = X_i^T Y_i \in \mathbb{R}^{P \times P}$ and $Q_{w,q}(V) = \sum_{(i,j) \in \mathbb{E}} w_{ij} \|v_i - v_j\|_q$. This yields the problem:

$$\begin{align*}
\max_{\{u_i\}, \{v_i\}} & \sum_{i=1}^{N} u_i^T \Sigma_i v_i - \lambda Q_{w,q}(U) - \lambda Q_{w,q}(V) \quad \text{subject to} \quad \|u_i\|_2^2 = 1, \ \|v_i\|_2^2 = 1,
\end{align*}$$

for $i = 1, \ldots, N$. Without inequality constraints, this is a biconvex problem in the $\{u_i\}$ and $\{v_i\}$ when the subproblems are relaxed by fixing $\bar{x}_i = \Sigma_i v_i$ and $\bar{y}_i = \Sigma_i^T u_i$ at each subiterate:

$$\begin{align*}
\{u_i\}^{k+1} & \leftarrow \argmin_{\{u_i\}} \sum_{i=1}^{N} \frac{1}{2} \|\bar{x}_i - u_i\|_2^2 + \lambda Q_{w,q}(U) \quad \text{subject to} \quad \|u_i\|_2^2 = 1, \ i = 1, \ldots, N,
\end{align*}$$

where $y_{x,i}^k = x_i^T v_i^k$ and $y_{x,i}^k = u_i^k x_i$, we can write iterative updates:

$$\begin{align*}
\{u_i\}^{k+1} & \leftarrow \argmin_{u} \sum_{i=1}^{N} \|y_{x,i}^k - u_i\|_2^2 + \lambda Q_{w,q}(u) \quad \text{subject to} \quad \|u\|_2^2 = 1, \quad (9a) \\
\{v_i\}^{k+1} & \leftarrow \argmin_{\{v_i\}} \sum_{i=1}^{N} \|y_{x,i}^k - v_i\|_2^2 + \lambda Q_{w,q}(V) \quad \text{subject to} \quad \|v_i\|_2^2 = 1, \ i = 1, \ldots, N. \quad (9b)
\end{align*}$$

This results in the iterative updates each being standard convex clustering problems with added constraints. We will subsequently refer to a convex clustering solver as CONVEXCLUSTER. Here we implement our CONVEXCLUSTER as an ADMM algorithm that can be scaled using consensus ADMM (Appendix §8); however, other convex clustering solvers could be used instead. The CONVEXCLUSTER algorithm and derivation of the PCMF-PALS problem is given in the Appendix.
Figure 2: P3CA recovers 3 penguin species and identifies a hierarchy of shared embeddings (U on left and V on right). a. X variables colored by P3CA clustering solution indicated by the arrow. b. Ground truth clustering solution for X and Y variable sets. c. Y variables colored by best P3CA clustering solution indicated by the arrow. d-e. P3CA path diagrams for X and variable sets (colors indicate variable; pink is intercept term).

\[
\{v_i\}_k^{k+1} \leftarrow \arg\min_{\{v_i\}_k} \sum_{i=1}^N \frac{1}{2} \|\tilde{y}_i - v_i\|^2_2 + \lambda Q_{w,q}(V) \text{ subject to } \|v_i\|_2^2 = 1, \ i = 1, \ldots, N. \tag{11b}
\]

Each of these updates is a convex clustering problem in a particular representation of the data, leading to Algorithm 2. Here and for the algorithms above, we find empirically that for sufficiently small steps sizes the Algorithmic Regularization algorithm closely approaches the ADMM solutions with a significant speed up (see Appendix §5−7 for more details).

6 Dendrograms and Model Selection

PCMF fits a path of solutions along a sequence of values of \(\lambda\), and when using the \(\ell_2\)-norm (\(q = 2\)) (as we do for the remainder of the paper given its desirable rotational symmetry), not all members of a cluster are shrunk to exactly the same value [25]. Previous work has forced hard clustering at each agglomerative stage along the \(\lambda\) path [25, 58, 29]. This may artificially force observations into one cluster that may then later switch to another, resulting in nonsmooth paths in practice. We choose to instead let the paths be unconstrained and smooth during solving, and then generate a dendrogram using a wrapper function that estimates sequential split points from the fully fit paths by sequentially testing whether increasing the number of clusters at each step would improve overall model fit in terms of the penalized log-likelihood. Clustering at each \(\lambda\) is performed on the weighted affinity matrix generated from differences matrix defined by the dual variables as recommended in [10]. Thus this procedure estimates the connected components of the affinity graph defined by the dual variables at each value of \(\lambda\). Further details on model selection are described in Appendix §9.

7 Experiments

7.1 PCMF on Numerical Experiments and Real-World Omic Datasets.

Here, we evaluate PCMF in numerical experiments using small and large synthetic data (Table 1 and Appendix Tables 1-2) and 4 real-world datasets (Tables 2 and 3). We measure clustering performance of PCMF against 3 classical (PCA+K-means, Ward, spectral, and DP-GMM) and 10 state-of-the-art (Elastic Subspace Clustering [62], gMADD [46, 40], HDCC, Leiden, Louvain, DP-GMM, CARP) clustering methods including convex clustering (CARP) and 3 deep learning methods (DEC, IDEC, CarDEC) to accurately recover cluster labels in 18 synthetic datasets across 10 runs (Appendix Table 2) and in 4 complex real-world datasets (Table 2-3).
We further show our consensus ADMM runs on larger \( N > 4 \) complex real-world omic datasets as PCMF, but split the variables into Consensus PCMF and PCMF. Dataset [47], we find P3CA identifies a metabolome-proteome embedding that predicts both COVID-19 outperforms or ties all 13 clustering methods (Table 3). Subtype and Cell/Tissue Classification using Real-World Omic Datasets. To evaluate the PCA interpretation of PCMF embeddings, we compare and show high similarity to tSVD estimates on the ground-truth clusters (Fig. 1g and Appendix §1). Consensus ADMM and PCMF-PALS with nearest neighbors \( p > N \) outperforms in 18 synthetic datasets, especially for alternative clustering (DEC, IDEC; Table 3). Finally, we evaluate our method on 4 complex real-world omic datasets and find PCMF outperforms all 13 clustering methods (including CARP and deep learning methods) in nearly all cases (except on the SRBCT data against DEC/IDEC; Table 3).

### Numerical Experiments

We find that PCMF-ADMM and PCMF-PALS with nearest neighbors \( N > N \) performs competitively in 18 synthetic datasets, especially for \( p > N \) (\( p = 200, p = 2,000 \); Appendix Table 1). We further show our consensus ADMM runs on larger datasets and compare time of completion (TOC) to PCMF-ADMM (which like convex clustering cannot run on \( N > 1,000 \)) (Table 1; n.b. consensus ADMM code is unoptimized). To evaluate the PCA interpretation of PCMF embeddings, we compare and show high similarity to tSVD estimates on the ground-truth clusters (Fig. 1g and Appendix §1.2). Cancer Subtype and Cell/Tissue Classification using Real-World Omic Datasets. Finally, we evaluate our method on 4 complex real-world omic datasets and find PCMF outperforms all 13 clustering methods (including CARP and deep learning methods) in nearly all cases (except on the SRBCT data against DEC/IDEC; Table 3).

#### 7.2 P3CA on Real-world Physiological, Omic, Behavioral, and Neuroimaging Datasets

We demonstrate P3CA in six real-world datasets (Tables 2-3). Species Classification. First we apply P3CA to phenotypic data and classify penguin species with 98.25% accuracy [26] (Table 3 and Fig. 2). Cancer Subtype and Cell/Tissue Classification using Real-World Omic Datasets. We evaluate P3CA on the same 4 complex real-world omic datasets as PCMF, but split the variables into \( X \) and \( Y \) (Table 2). We find P3CA outperforms or ties all 13 clustering methods (Table 3). Predicting COVID-19 Severity. Using a COVID-19 dataset [47], we find P3CA identifies a metabolome-proteome embedding that predicts both COVID-19

### Table 1: Clustering accuracy of consensus PCMF on 4-class data.

| \( N = 100 \) | \( N = 1,000 \) | \( N = 10,000 \) | \( N = 100,000 \) |
|---|---|---|---|
| PCMF | 1.0 | 0.1 | 169.65 |
| Consensus PCMF | 1.0 | 1.0 | 71.6s | 1.0 | 1.0 | 467.2s | 1.0 | 1.0 | 4,897.8s | 1.0 | 1.0 | 137,488.9s |

### Table 2: Real-world datasets. See Appendix §11 for more details.

| Dataset | Variables (\( p \)) | Samples (\( N \)) | Classes |
|---|---|---|---|
| NCI (Multiview) | 6,830 genes (expression) (\( X \)) | 64 | 13 cell types |
| SRBCT (Multiview) | 2,318 genes (expression) (\( X \)) | 88 | 4 cancer diagnoses |
| Mouse Organ (Multiview) | 16,941 genes (scRNA-seq) (\( Y \)) | 125 | 7 mouse organs |
| Tumors (Multiview) | 11,931 expression/methylation (\( X \)) | 142 | 3 cancer diagnoses |
| COVID-19 (Multiview) | 403 metabolites (\( X \)); 382 proteins (\( Y \)) | 45 | 3 severities |
| NCI (Multiview) | 1,000 genes (expression) (\( X \)); 100 genes (expression) (\( Y \)) | 64 | 13 cell types |
| SRBCT (Multiview) | 1,000 genes (expression) (\( X \)); 100 genes (expression) (\( Y \)) | 88 | 4 cancer diagnoses |
| Mouse Organ (Multiview) | 100,000 genes (scRNA-seq) (\( X \)); 100 genes (scRNA-seq) (\( Y \)) | 125 | 7 mouse organs |
| Tumors (Multiview) | 1,000 genes (expression) (\( X \)); 100 genes (expression) (\( Y \)) | 142 | 3 cancer diagnoses |
| Autism (ASD) (Multiview) | 3 behaviors (\( X \)); 20 RSFC features (\( Y \)) | 299 | unknown |

### Table 3: Clustering accuracy on real-world datasets ("MV" abbreviates "Multiview").

| Dataset | NCI | SRBCT | Mouse Organ | Tumors | COVID-19 | Penguins | NCI-MV | SRBCT-MV | Mouse-MV | Tumors-MV |
|---|---|---|---|---|---|---|---|---|---|---|
| PCMF-PALS | 64.06% | 55.42% | 80.00% | 97.89% | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| PCMF-ADMM | 43.79% | 51.8% | 73.6% | 92.25% | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| P3CA | 91.11% | 98.25% | 56.25% | 65.06% | 63.20% | 98.59% | 100 | 100 | 100 | 100 |
| PCA + K-means | 39.06% | 40.96% | 45.60% | 50.00% | 51.11% | 79.82% | 31.25% | 37.35% | 27.20% | 50.70% |
| CCA + K-means | 56.25% | 40.96% | 46.40% | 94.37% | 68.89% | 96.78% | 51.56% | 40.96% | 30.40% | 94.36% |
| Spectral | 43.75% | 43.37% | 45.60% | 93.66% | 82.22% | 96.78% | 50.00% | 43.37% | 40.00% | 93.66% |
| Elastic Subspace | 59.38% | 49.40% | 73.60% | 94.37% | 51.11% | 97.37% | 48.43% | 40.96% | 52.00% | 94.37% |
| gMADD | 42.19% | 46.99% | 42.40% | 72.34% | 51.11% | 67.25% | 39.06% | 44.58% | 35.20% | 58.45% |
| HDCC | 59.38% | 34.94% | 29.60% | 50.00% | 40.00% | 88.01% | 51.50% | 38.55% | 29.60% | 50.00% |
| Leiden | 50.00% | 46.99% | 68.00% | 71.12% | 82.22% | 40.00% | 48.43% | 46.99% | 49.60% | 71.13% |
| Louvain | 42.19% | 48.19% | 76.00% | 94.34% | 82.22% | 65.20% | 45.31% | 48.43% | 60.80% | 93.66% |
| DP-GMM | 46.88% | 43.37% | 54.40% | 85.92% | 73.33% | 68.42% | 45.31% | 45.31% | 39.20% | 92.96% |
| hCARP | 43.75% | 46.99% | 36.00% | 75.25% | 71.11% | 79.82% | 34.37% | 43.37% | 30.40% | 93.66% |
| DEC | 45.31% | **71.08%** | 46.40% | 94.37% | 88.89% | 54.69% | **65.06%** | 33.60% | 94.37% |
| IDEC | 48.44% | 67.47% | 61.60% | 92.96% | 73.33% | 73.33% | 50.00% | 43.37% | 30.40% | 93.66% |
| CarDEC | 51.56% | 40.96% | 75.20% | 90.14% | 84.44% | 90.14% | 51.56% | 40.96% | 30.40% | 93.66% |
1. GBM (N = 68)
2. Lung (N = 36), GBM (N = 3)
3. Breast (N = 35)

Variable: MIRLET7I microRNA
Samples in Leaf
Accuracy: 97.89%

Figure 3: PCMF dendrogram on Tumors data shows cancer taxonomy with dendrogram splits consistent with observed confusion rate, highlighting interpretability.

Figure 4: Associations between P3CA variates and protein / metabolite abundance predict COVID-19 severity. a. Proteome (X) P3CA variate is associated with SERPINA3 abundance. Right: Distribution of SERPINA3 abundance by patient group. b. Left: Metabolome (Y) P3CA variate is associated with stachydrine abundance, and strength/direction of association differs between groups. Right: Distribution of stachydrine abundance by patient group. r_all indicates correlation between all patients.

Severity and potential protein and metabolite biomarkers. P3CA recovers true patient severity subgroups (healthy, non-severe, severe) with the highest accuracy (91.11%) compared to other methods (Table 3). In this COVID-19 example P3CA is successful on a very small sample of N = 45 subjects with p ≈ 400 variables for each data view. By correlating the metabolome/proteome P3CA variate with metabolite/protein abundance, we identify severity-associated proteins and metabolites (Fig. 4). Importantly, the direction of association can be shared across subgroups (SERPINA3 in Fig. 4a) or can differ across subgroups (see r-values for stachydrine in Fig. 4b)— uniquely distinguishable in P3CA compared to standard canonical correlation analysis that only captures shared associations across all subjects/samples. The associated proteins/metabolites to COVID-19 are supported by prior work: SERPINA3 (alpha-1 antichimotrypsin) is an immune marker in the coagulation/fibrinolytic cascade [44] that has been associated with COVID-19 [47, 15]; stachydrine (alias proline betaine) is an osmotic protective agent/compound [5] associated with a number of medical disorders [9], and osmotic crisis is a risk factor for severe COVID-19 [36]. Discovery of Subtypes of Autism Spectrum Disorder (ASD). In Appendix §10.2 (Appendix Fig. 6 and Appendix Tables 2-3), we demonstrate deploying P3CA for subtype discovery, in which true biological subtypes are unknown but could inform diagnosis and the development of personalized treatments [17, 24]. P3CA identifies distinct biological subtypes of N = 299 ASD patients by finding a clustered CCA embedding of multivariate clinical behavior (X = 3) features and functional brain connectivity (Y = 20 top features) with stable subject-level U and V coefficients (cosine similarity: 0.93 ± 0.05 for U estimates and 0.97 ± 0.03 for V estimates; 10 subsamples 70% patients).

8 Conclusion

Here we have introduced Pathwise Clustered Matrix Factorization (PCMF) and its generalization to multiview data—Pathwise Clustered Canonical Correlation Analysis (P3CA)—both approaches that jointly hierarchically cluster and embed high dimensional data using a convex clustering penalty. These methods are significant, as they are uniquely interpretable and effective for both small N and small cluster size as well as scalable to very large N. We present novel, accelerated, and scalable algorithms to solve these problems along a path of solutions and introduce a method to recover a dendrogram from the pathwise solutions. We then validate our approaches on synthetic and real-world data. Our results showed our methods to be highly competitive in terms of clustering accuracy and in terms of the enhanced explainability allowed by the hierarchical discovery of cluster-wise principal components or canonical variates.
References

[1] G. I. Allen, L. Grosenick, and J. Taylor. A generalized Least-Square matrix decomposition. *J. Am. Stat. Assoc.*, 109(505):145–159, Jan. 2014.

[2] R. Bonacchi, A. Meani, C. Bassi, E. Pagani, M. Filippi, and M. A. Rocca. MRI-Based clustering of multiple sclerosis patients in the perspective of personalized medicine (3930). *Neurology*, 94(15 Supplement), Apr. 2020.

[3] A. Boubekki, M. Kampffmeyer, U. Brefeld, and R. Jenssen. Joint optimization of an autoencoder for clustering and embedding. *Mach. Learn.*, 110(7):1901–1937, July 2021.

[4] S. Boyd, N. Parikh, and E. Chu. *Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers*. Now Publishers Inc, 2011.

[5] S. T. Chambers and C. M. Kunin. Isolation of glycine betaine and proline betaine from human urine, assessment of their role as osmoprotective agents for bacteria and the kidney. *J. Clin. Invest.*, 79(3):731–737, Mar. 1987.

[6] W.-C. Chang. On using principal components before separating a mixture of two multivariate normal distributions. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 32(3):267–275, 1983.

[7] C.-L. Chen, Y.-C. Gong, and Y.-J. Tian. KCK-Means: A clustering method based on kernel canonical correlation analysis. In *Computational Science – ICCS 2008*, pages 995–1004. Springer Berlin Heidelberg, 2008.

[8] J. Chen and I. D. Schizas. Distributed sparse canonical correlation analysis in clustering sensor data. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 639–643, Nov. 2013.

[9] F. Cheng, Y. Zhou, M. Wang, C. Guo, Z. Cao, R. Zhang, and C. Peng. A review of pharmacological and pharmacokinetic properties of stachydrine. *Pharmacol. Res.*, 155:104755, May 2020.

[10] E. C. Chi and K. Lange. Splitting methods for convex clustering. *J. Comput. Graph. Stat.*, 24(4):994–1013, Dec. 2015.

[11] E. C. Chi and S. Steinerberger. Recovering trees with convex clustering. *SIAM JOURNAL ON MATHEMATICS OF DATA SCIENCE*, 1(3):383–407, 2019.

[12] E. C. Chi, G. I. Allen, and R. G. Baraniuk. Convex biclustering. *Biometrics*, 73(1):10–19, Mar. 2017.

[13] J. Chiquet, P. Gutierrez, and G. Rigail. Fast tree inference with weighted fusion penalties. *J. Comput. Graph. Stat.*, 26(1):205–216, Jan. 2017.

[14] M. Ciortan and M. Defrance. GNN-based embedding for clustering scRNA-seq data. *Bioinformatics*.

[15] A. D’Alessandro, T. Thomas, M. Dzieciatkowska, R. C. Hill, R. O. Francis, K. E. Hudson, J. C. Zimring, E. A. Hod, S. L. Spitalnik, and K. C. Hansen. Serum proteomics in COVID-19 patients: Altered coagulation and complement status as a function of IL-6 level. *J. Proteome Res.*, 19(11):4417–4427, Nov. 2020.

[16] S. Danda. *Identification of Cell Types in scRNA-seq Data via Enhanced Local Embedding and Clustering*. PhD thesis, 2021.

[17] A. T. Drysdale, L. Grosenick, J. Downar, K. Dunlop, F. Mansouri, Y. Meng, R. N. Fetcho, B. Zebley, D. J. Oathes, A. Etkin, A. F. Schatzberg, K. Sudheimer, J. Keller, H. S. Mayberg, F. M. Gunning, G. S. Alexopoulos, M. D. Fox, A. Pascual-Leone, H. U. Voss, B. J. Casey, M. J. Dubin, and C. Liston. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.*, 23 (1):28–38, Jan. 2017.

[18] L. Du, K. Liu, T. Zhang, X. Yao, J. Yan, S. L. Risacher, J. Han, L. Guo, A. J. Saykin, L. Shen, and A. Initiative. A Novel SCCA Approach via Truncated $\ell_1$-norm and Truncated Group Lasso for Brain Imaging Genetics. *Bioinform Oxf Engl*, 2017.

[19] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1 (3):211–218, Sept. 1936.
[20] X. Z. Fern, C. E. Brodley, and M. A. Friedl. Correlation clustering for learning mixtures of canonical correlation models. Apr. 2005.

[21] P. Fogel, Y. Gaston-Mathé, D. Hawkins, F. Fogel, G. Luta, and S. S. Young. Applications of a novel clustering approach using Non-Negative matrix factorization to environmental research in public health. *Int. J. Environ. Res. Public Health*, 13(5), May 2016.

[22] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *AOAS*, 1(2):302–332, Dec. 2007.

[23] E. Gharavi, A. Gu, G. Zheng, J. P. Smith, H. J. Cho, A. Zhang, D. E. Brown, and N. C. Sheffield. Embeddings of genomic region sets capture rich biological associations in lower dimensions. *Bioinformatics*, 37(23):4299–4306, Dec. 2021.

[24] L. Grosenick, T. C. Shi, F. M. Gunning, M. J. Dubin, J. Downar, and C. Liston. Functional and optogenetic approaches to discovering stable Subtype-Specific circuit mechanisms in depression. *Biol Psychiatry Cogn Neurosci Neuroimaging*, 4(6):554–566, June 2019.

[25] T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath an algorithm for clustering using convex fusion penalties. *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[26] A. M. Horst, A. P. Hill, and K. B. Gorman. palmerpenguins: Palmer archipelago (antarctica) penguin data. *R package version 0. 1. 0*, 2020.

[27] P. Huang, Y. Huang, W. Wang, and L. Wang. Deep embedding network for clustering. *2014 22nd International Conference on Pattern Recognition*, pages 1532–1537, 2014.

[28] IEEE and M. Weylandt. Splitting methods for convex bi-clustering and co-clustering. In *2019 IEEE DATA SCIENCE WORKSHOP (DSW)*, pages 237–242, 2019.

[29] T. Jiang, S. Vavasis, and C. W. Zhai. Recovery of a mixture of gaussians by sum-of-norms clustering. *J. Mach. Learn. Res.*, 21(225):1–16, 2020.

[30] J. Lakkis, D. Wang, Y. Zhang, G. Hu, K. Wang, H. Pan, L. Ungar, M. P. Reilly, X. Li, and M. Li. A joint deep learning model enables simultaneous batch effect correction, denoising, and clustering in single-cell transcriptomics. *Genome Res.*, 31(10):1753–1766, Oct. 2021.

[31] E. Lei, K. Miller, and A. Dubrawski. Learning mixtures of Multi-Output regression models by correlation clustering for Multi-View data. Sept. 2017.

[32] Y. X. Lin and S. C. Chen. A centroid Auto-Fused hierarchical fuzzy c-means clustering. *IEEE Trans. Fuzzy Syst.*, 29(7):2006–2017, 2021.

[33] F. Lindsten, H. Ohlsson, and L. Ljung. Clustering using sum-of-norms regularization: With application to particle filter output computation. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 201–204, June 2011.

[34] L. W. Mackey. Deflation methods for sparse pca. In *NIPS*, volume 21, pages 1017–1024, 2008.

[35] D. Mautz, C. Plant, and C. Böhm. DeepECT: The deep embedded cluster tree. *Data Science and Engineering*, 5(4):419–432, Dec. 2020.

[36] D. Muhanna, S. R. Arnipalli, S. B. Kumar, and O. Ziouzenkova. Osmotic adaptation by Na+-Dependent transporters and ACE2: Correlation with hemostatic crisis in COVID-19. *Biomedicines*, 8(11), Oct. 2020.

[37] Q. Ouyang. *Canonical Correlation and Clustering for High Dimensional Data*. PhD thesis, 2019.

[38] W. Pan, X. Shen, and B. Liu. Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *J. Mach. Learn. Res.*, 14(7):1865, July 2013.

[39] A. Panahi, D. Dubhashi, F. D. Johansson, and C. Bhattacharyya. Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2769–2777. PMLR, 2017.
[40] B. Paul, S. K. De, and A. K. Ghosh. Some clustering-based exact distribution-free k-sample tests applicable to high dimension, low sample size data. *J. Multivar. Anal.*, page 104897, Nov. 2021.

[41] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor. Convex clustering shrinkage. In *PASCAL Workshop on Statistics and Optimization of Clustering Workshop*, 2005.

[42] T. Qian, S. Zhu, and Y. Hoshida. Use of big data in drug development for precision medicine: an update. *Expert Review of Precision Medicine and Drug Development*, 4(3):189–200, May 2019.

[43] P. Radchenko and G. Mukherjee. Convex clustering via 1-fusion penalization. *J. R. Stat. Soc. Series B Stat. Methodol.*, 79(5):1527–1546, Nov. 2017.

[44] H. Rubin, Z. M. Wang, E. B. Nickbarg, S. McLaren, N. Naïdo, O. L. Schoenberger, J. L. Johnson, and B. S. Cooperman. Cloning, expression, purification, and biological activity of recombinant native and variant human alpha 1-antichymotrypsins. *J. Biol. Chem.*, 265(2):1199–1207, Jan. 1990.

[45] C. Santos, R. Sanz-Pampulona, E. Nadal, J. Grasselli, S. Pernas, R. Dienstmann, V. Moreno, J. Tabernero, and R. Salazar. Intrinsic cancer subtypes—next steps into personalized medicine. *Cell. Oncol.*, 38(1):3–16, Feb. 2015.

[46] S. Sarkar and A. K. Ghosh. On perfect clustering of high dimension, low sample size data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(9):2257–2272, Sept. 2020.

[47] B. Shen, X. Yi, Y. Sun, X. Bi, J. Du, C. Zhang, S. Quan, F. Zhang, R. Sun, L. Qian, W. Ge, W. Liu, S. Liang, H. Chen, Y. Zhang, J. Li, J. Xu, Z. He, B. Chen, J. Wang, H. Yan, Y. Zheng, D. Wang, J. Zhu, Z. Kong, Z. Kang, X. Liang, X. Ding, G. Ruan, N. Xiang, X. Cai, H. Gao, L. Li, S. Li, Q. Xiao, T. Lu, Y. Zhu, H. Liu, H. Chen, and T. Guo. Proteomic and metabolic characterization of COVID-19 patient sera. *Cell*, 182(1):59–72.e15, July 2020.

[48] S.-J. Shin, K. Song, and I.-C. Moon. Hierarchically clustered representation learning. *AAAI*, 34(04):5776–5783, Apr. 2020.

[49] A. Singh and B. Pandey. A new intelligent medical decision support system based on enhanced hierarchical clustering and random decision forest for the classification of alcoholic liver damage, primary hepatoma, liver cirrhosis, and cholelithiasis. *J. Healthc. Eng.*, 2018:1469043, Feb. 2018.

[50] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorisen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønnning, and A.-L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001. doi: 10.1073/pnas.191367098. URL https://www.pnas.org/content/98/19/10869.

[51] X. L. Sui, L. Xu, X. Qian, and T. Liu. Convex clustering with metric learning. *Pattern Recognit.*, 81:575–584, Sept. 2018.

[52] D. Sun, K.-C. Toh, and Y. Yuan. Convex clustering: Model, theoretical guarantee and efficient algorithm preprint. Oct. 2018.

[53] D. Sun, K.-C. Toh, and Y. Yuan. Convex clustering: Model, theoretical guarantee and efficient algorithm. *J. Mach. Learn. Res.*, 22(9):1–32, 2021.

[54] K. M. Tan and D. Witten. Statistical properties of convex clustering. *Electron J Stat*, 9(2):2324–2347, Oct. 2015.

[55] M. Wang, T. Yao, and G. I. Allen. Supervised convex clustering. May 2020.

[56] M. J. Wang and G. I. Allen. Integrative generalized convex clustering optimization and feature selection for mixed Multi-View data. *J. Mach. Learn. Res.*, 22, 2021.

[57] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep Multi-View representation learning: Objectives and optimization. Feb. 2016.

[58] M. Weylandt, J. Nagorski, and G. I. Allen. Dynamic visualization and fast computation for convex clustering via algorithmic regularization. *J. Comput. Graph. Stat.*, 29(1):87–96, 2020.
[59] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, July 2009.

[60] W. Wu and X. Ma. Joint learning dimension reduction and clustering of single-cell RNA-sequencing data. *Bioinformatics*, 36(12):3825–3832, June 2020.

[61] J. Yang, D. Parikh, and D. Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5147–5156, 2016.

[62] C. You, C.-G. Li, D. P. Robinson, and R. Vidal. Oracle based active set algorithm for scalable elastic net subspace clustering. May 2016.

[63] L. Zhou, G. Du, K. Lü, and L. Wang. A network-based sparse and multi-manifold regularized multiple non-negative matrix factorization for multi-view clustering. *Expert Syst. Appl.*, 174:114783, July 2021.