Cluster-Aware Algorithms
for AI-Enabled Precision Medicine

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

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

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

Abstract

AI-enabled precision medicine promises a transformative improvement in healthcare outcomes. However, training on biomedical data presents a challenge: such data are often high dimensional, clustered, and of limited sample size. To overcome this, we propose a simple and scalable approach for cluster-aware embedding that augments embedding methods with a convex clustering penalty. This approach outperforms fourteen widely-used clustering methods on both highly underdetermined problems and on large sample datasets, yielding interpretable dendrograms of the embedded clusters. Our approach improves on existing methods and introduces a modular framework for interpretable biomarker discovery in precision medicine.

1 Introduction and Related Work

Interpretable clustering of patients into distinct subtypes is important for personalized biomarker discovery, diagnosis, prognosis, and treatment selection [4, 6, 9, 19, 53, 58, 57, 49]. However, due to the "curse of dimensionality", similarity metrics (and thus clustering algorithm outcomes) degrade in high dimensions (the "$p > N$" setting common in medicine and genomics, where we have p correlated variables and N observations where N is fewer than p). It is popular to use a two-stage procedure, first embedding high dimensional data into a low-rank representation, and then clustering in this latent space [9, 11, 12, 15, 16, 19, 27, 20, 45]. Unfortunately, such two-stage procedures can lead to suboptimal and hard-to-explain results [10], as the embedding ignores important clustered structure in the data, thereby harming the embedding (see Fig. 1).

These issues motivate a need for joint clustering and embedding methods for such data. Here, we develop an explainable and scalable formulation for joint clustering and embedding ("cluster-aware embedding") relevant to precision medicine applications; we show that an addition of a convex clustering penalty (λ) to standard embedding methods yields a simple and modular approach to cluster-aware embedding that is highly competitive in practice.

Exciting methods have emerged for jointly clustering and embedding data, including cluster-aware feature selection [64], CCA mixture models [23, 40], non-negative matrix factorization (NMF)-based models [25, 70, 75], and a number of neural networks (e.g., [7, 35, 39, 43, 56, 65, 73]). Although pioneering, these approaches involve complicated many-objective or deep neural network formulations that prioritize clustering over interpretability and underperform on restricted data cases.

While clustering algorithms are classically formalized as discrete optimization problems that are NP-hard, by relaxing the hard clustering constraint to a convex penalty [48], clustering can be reformulated as a convex optimization problem (referred to as "convex clustering," "clusterpath" or "sum-of-norms"). A range of theoretical/algorithmic developments and approaches for convex clustering [14, 24, 36, 41, 46, 59, 60, 62, 67], have been developed to solve the problem [13, 31, 46, 60, 68]. Crucially, a recent warm-started ADMM approach—Algorithmic Regularization—enables feasible computation of dense convex clustering λ paths, speeding convergence more than 100-fold [68].
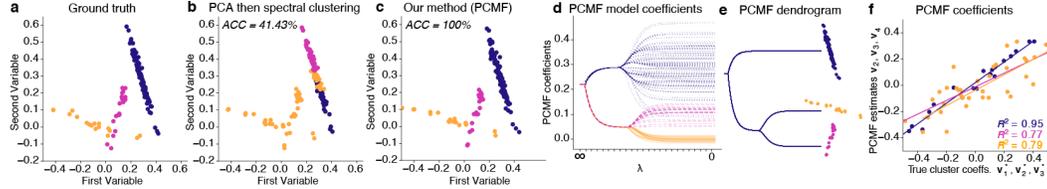
Figure 1: PCMF for explainable joint PCA and hierarchical clustering. **a.** Scatterplot of reconstructed ground truth data (PCA rank $r = 4$) for 3-class problem; $p = 20$; $N_1 = 100$ (blue), $N_2 = 25$ (pink), $N_3 = 25$ (orange), colored by true cluster membership. **b.** PCA ($r = 4$) sequentially followed by spectral clustering on PCA components. **c.** Joint PCA and clustering with PCMF ($r = 4$; $\lambda = 3.0$). Two-step procedures in **b-c** fail to find correct clusters while PCMF succeeds. Color indicates predicted clusters. **d.** PCMF paths for variable 1 fit along decreasing penalty path ($\lambda = \infty$ to $\lambda = 0$). **e.** Interpretable PCMF dendrogram estimated from paths. **f.** PCMF coefficients accurately fit ground truth cluster-specific coefficients used to generate data. PCMF coefficients $\mathbf{v_2}$, $\mathbf{v_3}$, and $\mathbf{v_4}$ approximate true cluster coefficients ("slopes") $\mathbf{v_1^*}$ (blue), $\mathbf{v_2^*}$ (pink), and $\mathbf{v_3^*}$ (orange).

2 Our Approach: Pathwise Clustered Matrix Factorization (PCMF)

We use the convex clustering penalty (λ) as a modular addition to common embedding methods, making them cluster-aware (i.e., enabling them to jointly cluster and embed). Given data matrix $X \in \mathbb{R}^{N \times p}$ (with N observations in the rows, p variables in the columns, and rank $R \leq \min(N, p)$), we can express the embedding constraint, $\widehat{X}$, in terms of the widely-used truncated singular value decomposition (tSVD) [21]. The rank-$r \leq R$ tSVD embedding is given by $\widehat{X} = U_r S_r V_r^T$, subject to orthogonality constraints on the first r left and right singular vectors (collected in U_r and V_r, respectively) and the first r singular values on the diagonal of S_r [21]. This yields the PCMF problem:

$$\underset{\widehat{X}, U_r, S_r, V_r}{\text{minimize}} \quad \frac{1}{2}\|X - \widehat{X}\|_F^2 + \lambda \sum_{i<j} w_{ij}\|\widehat{X}_{i\cdot} - \widehat{X}_{j\cdot}\|_q$$
$$\text{subject to} \quad \widehat{X} - U_r S_r V_r^T = 0, \ U_r^T U_r = V_r^T V_r = I_r, \ S_r = \text{diag}(s_1, \ldots, s_r), \tag{1}$$

for $s_1 \geq s_2 \geq \cdots \geq s_r > 0$. We use the ℓ_2-norm ($q = 2$). If X is centered, the tSVD is also principal components analysis (PCA). Next, we present algorithms for solving this nonconvex problem.

PCMF dendrograms for explainability and model selection. PCMF does not require choosing the number of clusters prior to fitting, and can generate a dendrogram. First, it fits a path of solutions along a sequence of λs (Fig. 1e–f). Unlike previous convex clustering approaches [31, 68, 36], we solve divisively and do not constrain the paths. Second, we sequentially (at each λ) estimate split points based on whether or not increasing the number of clusters would improve model fit (based on minimizing the penalized log-likelihood). Clustering at each λ is performed on the weighted affinity matrix that estimates the connected components (the differences matrix defined by the dual variables [13]). To output the dendrogram, we average $\widehat{X}$ within clusters using the split points along the paths.
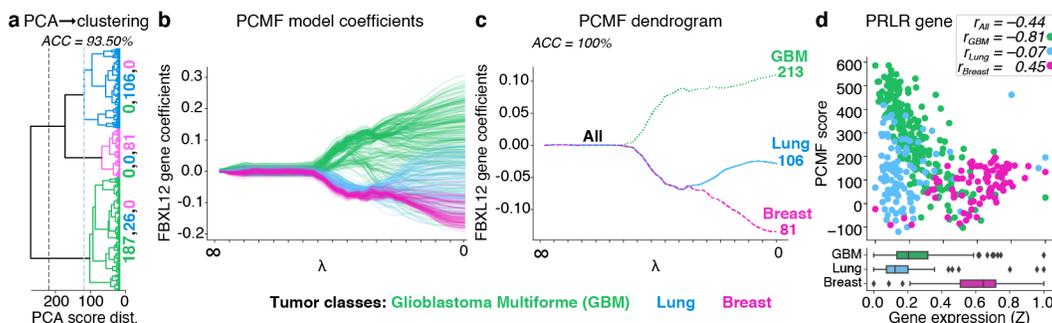
Figure 2: PCMF identifies tumor clusters and embeddings using gene expression ($p = 11,931$) from $N = 400$ samples. **a.** Dendrogram shows hierarchical clustering on PCA embedding. **b.** PCMF path and **c.** dendrogram show PCMF perfectly recovers clusters. **d.** Scatter/boxplots show PRLR gene expression versus PCMF expression scores for each sample colored by PCMF-predicted clusters.

Hierarchically-clustered PCA via Algorithmic Regularization. We first re-represent the relevant nonzero distances as a sparse graph, G [13], and then introduce auxiliary variable $G = D\widehat{X} \in \mathcal{R}^{|\mathcal{E}| \times p}$, where $D \in \mathbb{R}^{|\mathcal{E}| \times N}$ is a sparse matrix with the weighted pairwise distances defined by edges, $\mathcal{E}$. Then adding weights, w_ℓ, we rewrite the PCMF problem as:

$$\underset{\widehat{X}, G, U_r, S_r, V_r}{\text{minimize}} \quad \frac{1}{2}\|X - \widehat{X}\|_F^2 + \lambda \sum_{\ell \in \mathcal{E}} w_\ell \|G_{\ell\cdot}\|_q \tag{2}$$

$$\text{subject to } \widehat{X} - U_r S_r V_r^T = 0, \ G - D\widehat{X} = 0, U^T U = V^T V = I_r, \ S_r = \text{diag}(s_1, \ldots, s_r),$$

for $s_1 \geq \cdots \geq s_r > 0$, yielding a problem separable in its objective and penalty subject to (nonconvex) constraints. We solve along a λ path, and use Algorithmic Regularization [68] (making K small in Alg. 1) and mini-batches to dramatically speed up computation time and feasibility.

Algorithm 1 PCMF (**Input:** data X, decreasing path $\{\lambda\}$, weights $\mathbf{w}$, pairwise distance matrix D)

Notation: data mean $\overline{X}$, rank r, iteration k, norm $q \in \{1, 2, \infty\}$, $\rho \geq 1$, operator $\text{prox}_{\frac{\lambda}{\rho} P_{\mathbf{w},q}(G)}$

1: $G^0 \leftarrow Z_1^0 \leftarrow DX; \widehat{X} \leftarrow Z_2^0 \leftarrow \overline{X}, (U_r^0, S_r^0, V_r^0) \leftarrow \text{SVD}_r(\widehat{X}), L = \text{chol}(I + \rho I + \rho D^T D)$
2: **for** $\lambda \in \{\lambda\}$ **do**
3:     **for** $k = 1, \ldots, K$ **do**
4:         $\widehat{X}^{k+1} \leftarrow L^{-T} L^{-1} \big(X + \rho D^T (G^k - Z_1^k) + \rho(U_r^k S_r^k V_r^{kT} - Z_2^k) \big)$
5:         $G^{k+1} \leftarrow \text{prox}_{\frac{\lambda}{\rho} P_{\mathbf{w},q}(G)}(D\widehat{X}^{k+1} + Z_1^k); \ (U_r^{k+1}, S_r^{k+1}, V_r^{k+1}) \leftarrow \text{SVD}_r(\widehat{X}^{k+1} + Z_2^k)$
6:         $Z_1^{k+1} \leftarrow Z_1^k + D^T \widehat{X}^{k+1} - G^{k+1}; \quad Z_2^{k+1} \leftarrow Z_2^k + \widehat{X}^{k+1} - U_r^{k+1}, S_r^{k+1}, V_r^{k+1}$
7:     **end for**
8:     Save current path solutions: $\widehat{X}_\lambda \leftarrow \widehat{X}^K, G_\lambda \leftarrow G^K, (U_{r,\lambda}, S_{r,\lambda}, V_{r,\lambda}) \leftarrow (U_r^K, S_r^K, V_r^K)$
9:     Initialize for next path solution: $\widehat{X}^0 \leftarrow \widehat{X}^K, G^0 \leftarrow G^K, (U_r^0, S_r^0, V_r^0) \leftarrow (U_r^K, S_r^K, V_r^K)$
10: **end for**; **return pathwise solutions** $\{\widehat{X}_\lambda\}, \{G_\lambda\}, \{U_{r,\lambda}\}, \{S_{r,\lambda}\}, \{V_{r,\lambda}\}$

A nonlinear extension: Locally Linear PCMF (LL-PCMF). We introduce LL-PCMF using Penalized Alternating Least Squares (PALS) to solve it [51]. Without loss of generality, we center and scale X, set $s_1 = 1$, and consider the rank-1 version (generalizable to rank-r via deflation [42, 69]). Denoting the ith column vector of X^T as $\mathbf{x}_i = (X^T)_{\cdot i}$ and defining penalty, $\tilde{P}_{\mathbf{w},q}(\mathbf{u}, \mathbf{v}) = \sum_{(i,j) \in \mathcal{E}} w_{ij} \|u_i \mathbf{v} - u_j \mathbf{v}\|_q$ (weights $w_{ij} > 0$), we re-write the rank-1 tSVD with a convex clustering penalty, λ. We introduce overparameterization, replacing the single vector $\mathbf{v}$ with matrix $V \in \mathbb{R}^{p \times N}$ ($\mathbf{v}_i = V_{\cdot i}$; column vector set $\{\mathbf{v}\}_i, \ i = 1, \ldots, N$)—this allows each observation to potentially be its own cluster in the limit $\lambda \to 0$. Defining $P_{\mathbf{w},q}(\mathbf{u}, V) = \sum_{(i,j) \in \mathcal{E}} w_{ij} \|u_i \mathbf{v}_i - u_j \mathbf{v}_j\|_q$, we arrive at:

$$\underset{\mathbf{u},V}{\text{minimize}} \sum_{i=1}^N \frac{1}{2}\|\mathbf{x}_i - u_i \mathbf{v}_i\|_2^2 + \lambda P_{\mathbf{w},q}(\mathbf{u}, V) \text{ subject to } \|\mathbf{u}\|_2^2 = 1, \ \|\mathbf{v}_i\|_2^2 = 1, \text{ for } i = 1, \ldots, N. \tag{3}$$

We remove the penalty cross-terms, allowing $\mathbf{u}$ and $\mathbf{v}$ to independently vary with now locally-linear weights, w_{ij}. We replace $P_{\mathbf{w},q}(\mathbf{u}, \mathbf{v})$ with $Q_{\mathbf{w},q}^{\mathbf{u}}(\mathbf{u}) = \sum_{(i,j) \in \mathcal{E}} w_{ij} |u_i - u_j|$ and $Q_{\mathbf{w},q}^V(V) = \sum_{(i,j) \in \mathcal{E}} w_{ij} \|\mathbf{v}_i - \mathbf{v}_j\|_q$, and use fixed iterate k values, $y_{\mathbf{u},i}^k = \mathbf{x}_i^T \mathbf{v}_i^k$ and $\mathbf{y}_{\mathbf{v},i}^k = u_i^k \mathbf{x}_i$, for updates:

$$\mathbf{u}^{k+1} \leftarrow \underset{\mathbf{u}}{\text{argmin}} \sum_{i=1}^N \|y_{\mathbf{u},i}^k - u_i\|_2^2 + \lambda Q_{\mathbf{w},q}^{\mathbf{u}}(\mathbf{u}) \text{ subject to } \|\mathbf{u}\|_2^2 = 1, \text{for } i = 1, \ldots, N, \tag{4a}$$

$$\{\mathbf{v}_i\}^{k+1} \leftarrow \underset{\{\mathbf{v}_i\}}{\text{argmin}} \sum_{i=1}^N \|\mathbf{y}_{\mathbf{v},i}^k - \mathbf{v}_i\|_2^2 + \lambda Q_{\mathbf{w},q}^V(V) \text{ subject to } \|\mathbf{v}_i\|_2^2 = 1, \text{ for } i = 1, \ldots, N. \tag{4b}$$

A multi-view extension: Pathwise Clustered CCA (P3CA). We extend our approach to jointly learn low-rank correlation structure while clustering samples across multiple data views (i.e., fitting canonical correlation analysis (CCA) within clusters). As in LL-PCMF, we overparameterized ($\mathbf{v}_i = V_{\cdot i}$ are column vectors of $V \in \mathbb{R}^{p \times N}$), yielding P3CA problem:

$$\underset{\{\mathbf{u}_i\}, \{\mathbf{v}_i\}}{\text{minimize}} - \sum_{i=1}^N \mathbf{u}_i^T \Sigma_i \mathbf{v}_i + \lambda Q_{\mathbf{w},q}(U) + \lambda Q_{\mathbf{w},q}(V) \text{ subject to } \|\mathbf{u}_i\|_2^2 = 1, \ \|\mathbf{v}_i\|_2^2 = 1, \tag{5}$$

for $i = 1, \ldots, N$. Without inequality constraints, this is biconvex in the $\{\mathbf{u}_i\}$ and $\{\mathbf{v}_i\}$, with the subproblems relaxed by fixing $\tilde{\mathbf{x}}_i = \Sigma_i \mathbf{v}_i$ and $\tilde{\mathbf{y}}_i = \Sigma_i^T \mathbf{u}_i$ at each subiterate, leading to Alg. 2.

3

Algorithm 2 P3CA (**Input:** data (X, Y), decreasing path $\{\lambda\}$, weights $\mathbf{w}$, norm $q \in \{1, 2, \infty\}$)

Notation: iteration k, data means $(\bar{X}, \bar{Y})$, $\mathbf{v}_i = V_{\cdot i}$, $\tilde{\mathbf{x}}_i = (\tilde{X}_{i\cdot})^T$, $\tilde{\mathbf{y}}_i = (\tilde{Y}_{i\cdot})^T$, $\rho \geq 1$

1: $U \leftarrow \bar{X}, V \leftarrow \bar{Y}$
2: **for** $\lambda \in \{\lambda\}$ **do**
3:      **for** $k = 1, \ldots, K$ **do**
4:          $\tilde{\mathbf{x}}_i^{k+1} \leftarrow \Sigma_i \mathbf{v}_i^k$ $(\Sigma_i = X_{i\cdot} Y_{i\cdot}^T \in \mathbb{R}^{p_X \times p_Y})$ for $i = 1, \ldots, N$
5:          $\mathbf{u}_i^{k+\frac{1}{2}} \leftarrow \text{CONVEXCLUSTER}(\tilde{X}^{k+1}, U^k, \lambda, \mathbf{w}, \mathbf{q}); \mathbf{u}_i^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{u}_i^{k+\frac{1}{2}})$, $i = 1, \ldots, N$
6:          $\tilde{\mathbf{y}}_i^{k+1} \leftarrow \Sigma_i^T \mathbf{u}_i^{k+1}$ $(\Sigma_i^T = Y_{i\cdot} X_{i\cdot}^T \in \mathbb{R}^{p_Y \times p_X})$ for $i = 1, \ldots, N$
7:          $\mathbf{v}_i^{k+\frac{1}{2}} \leftarrow \text{CONVEXCLUSTER}(\tilde{Y}^{k+1}, V^k, \lambda, \mathbf{w}, \mathbf{q}); \mathbf{v}_i^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{v}_i^{k+\frac{1}{2}})$, $i = 1, \ldots, N$
8:      **end for**
9:      Save path solutions: $U_{i\cdot}^K \leftarrow \mathbf{u}_i^{KT}$; $V_{i\cdot}^K \leftarrow \mathbf{v}_i^{KT}$ for $i = 1, \ldots, N$; $(U_\lambda, V_\lambda) \leftarrow (U^K, V^K)$
10:     Initialize: $(U^0, V^0) \leftarrow (U^K, V^K)$
11: **end for**; **return pathwise solutions** $\{U_\lambda\}, \{V_\lambda\}$

---

3 Results and Discussion

First, we evaluate our unsupervised cluster-aware approach (PCMF, LL-PCMF, and P3CA) using synthetic data and 13 real-world biomedical datasets (7 single-view; 6 multi-view; Figs. 1–2; Table 1). Our approach outperforms 14 other methods on both underdetermined datasets ($p > N$) and large sample datasets (up to $N = 100,000$), except versus DEC/IDEC on SRBCT. This includes a small, multi-view COVID-19 dataset ($N = 45$) where P3CA identified hierarchically clustered metabolome-proteome embeddings that predict severity (accuracy = 91.11%) and biomarkers.

Next, in the Tumors-Large dataset ($N = 400$), we found PCMF model coefficients for the FBXL2 gene reveal a cluster hierarchy between GBM, lung, and breast cancer while a two-step approach does not (Fig. 2a-c). The branching structure reflects the suspected role of FBXL2 as a metastatic biomarker of breast-to-lung metastasis [66] and suggests a druggable target [18]. In Fig. 2d, correlations between the PCMF score and PRLR gene expression reveal strong slope differences between the three cancer tumor types. PRLR is a mammary proto-oncogene [28, 52], and a suggested prognostic biomarker of GBM progression (higher expression with shorter survival in males) [2] and therapeutic target [1, 52]. Interestingly, PRLR is strongly but oppositely associated with the GBM ($R = -0.81$) and breast tumor clusters ($R = 0.45$), as suggested in literature on triple-negative breast cancer (higher expression associated with lower recurrence and longer survival [44]).

Table 1: Clustering accuracy on real-world datasets ("MV" abbreviates "Multi-view").
("X" indicates computationally infeasible to run. "T" indicates infeasible due to run time out.)

	NCI [30, 50]	SRBCT [30, 37]	Mouse [61, 38]	Tumors [26]	Tumors-Large [26]	MNIST [17]	Fashion [71]	Synthetic	Penguins-MV [32]	COVID-19-MV [55]	NCI-MV [30, 50]	SRBCT-MV [30, 37]	Mouse-MV [61, 38]	Tumors-MV [26]
Variables (p)	6,830	2,318	16,944	11,931	11,931	784	784	1,000	2; 2	403; 382	1,000; 100	1,000; 100	1,000; 100	1,000; 100
Samples (N)	64	88	125	142	400	36,000	36,000	100,000	342	45	64	88	125	142
Classes	13	4	7	3	3	6	6	4	3	3	13	4	7	3
PCMF	43.79%	51.8%	73.6%	92.25%	**100.00%**	99.93%	99.94%	**100.00%**	—	—	—	—	—	—
LL-PCMF	64.06%	55.42%	**80.00%**	97.89%	—	—	—	—	—	—	—	—	—	—
P3CA	—	—	—	—	—	—	—	—	98.25%	91.11%	56.25%	65.06%	63.20%	98.59%
PCA + K-means [33]	39.06%	40.96%	45.60%	50.00%	89.75%	29.64%	45.00%	50.09%	—	—	—	—	—	—
CCA + K-means [34]	—	—	—	—	—	—	—	—	79.82%	51.11%	31.25%	37.35%	27.20%	50.70%
Ward [30]	56.25%	40.96%	46.40%	94.37%	90.50%	–X–	–X–	–X–	96.78%	68.89%	51.56%	40.96%	30.40%	94.36%
Spectral [30]	43.75%	43.37%	45.60%	93.66%	92.00%	–X–	–X–	–X–	96.78%	82.22%	50.00%	43.37%	40.00%	93.66%
Elastic Subspace [74]	59.38%	49.40%	73.60%	94.37%	—	—	—	—	97.37%	51.11%	48.43%	40.96%	52.00%	94.37%
gMADD [47, 54]	42.19%	46.99%	42.40%	72.54%	61.50%	–X–	–X–	–X–	67.25%	51.11%	39.06%	44.58%	35.20%	58.45%
HDCC [3, 8]	59.38%	34.94%	29.60%	50.00%	—	—	—	—	88.01%	40.00%	51.50%	38.55%	29.60%	50.00%
Leiden [63]	50.00%	46.99%	68.00%	71.12%	66.25%	60.62%	38.31%	10.88%	40.06%	82.22%	48.43%	46.99%	49.60%	71.13%
Louvain [5]	42.19%	48.19%	76.00%	94.34%	72.25%	69.88%	42.26%	10.85%	65.20%	82.22%	45.31%	48.19%	60.80%	93.66%
DP-GMM [22]	46.88%	43.37%	54.40%	85.92%	–X–	–X–	–X–	–X–	68.42%	73.33%	45.31%	44.58%	39.20%	92.96%
hCARP [68]	43.75%	46.99%	36.00%	75.25%	–X–	–X–	–X–	–X–	79.82%	71.11%	34.37%	43.37%	30.40%	93.66%
DEC [72]	45.31%	**71.08%**	46.40%	99.25%	–T–	–T–	–T–	86.67%	94.37%	88.89%	54.69%	**65.06%**	33.60%	94.37%
IDEC [29]	48.44%	67.47%	61.60%	92.96%	86.50%	55.25%	48.98%	–T–	—	73.33%	—	—	—	—
CarDEC [39]	51.56%	40.96%	75.20%	90.14%	–X–	–X–	–X–	–X–	—	84.44%	—	–X–	—	—

4 Conclusion

Facilitating adoption of AI-enabled precision medicine by professionals will require explainable, sensitive, and scalable methods appropriate for biomedical data. To meet this need, we have introduced an interpretable joint clustering and embedding strategy using a modular convex clustering penalty, and instantiated it in three scalable algorithms that solve linear (PCMF), nonlinear (LL-PCMF), and multi-view (P3CA) problems. We show that our method performs competitively across biomedical datasets against 14 commonly-used clustering approaches (including three deep learning methods).

References

[1] A. S. Asad, A. J. Nicola Candia, N. Gonzalez, C. F. Zuccato, A. Abt, S. J. Orrillo, Y. Lastra, E. De Simone, F. Boutillon, V. Goffin, A. Seilicovich, D. A. Pisera, M. J. Ferraris, and M. Candolfi. Prolactin and its receptor as therapeutic targets in glioblastoma multiforme. *Sci. Rep.*, 9 (1):19578, Dec. 2019.

[2] A. S. Asad, A. J. Nicola Candia, N. Gonzalez, C. F. Zuccato, A. Seilicovich, and M. Candolfi. The role of the prolactin receptor pathway in the pathogenesis of glioblastoma: what do we know so far? *Expert Opin. Ther. Targets*, 24(11):1121–1133, Nov. 2020.

[3] L. Bergé, C. Bouveyron, and S. Girard. HDclassif: An R package for Model-Based clustering and discriminant analysis of High-Dimensional data. *J. Stat. Softw.*, 46:1–29, Jan. 2012.

[4] J. R. Bishop, L. Zhang, and P. Lizano. Inflammation subtypes and translating inflammation-related genetic findings in schizophrenia and related psychoses: A perspective on pathways for treatment stratification and novel therapies. *Harv. Rev. Psychiatry*, 30(1):59–70, Feb. 2022.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008(10):P10008, Oct. 2008.

[6] 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.

[7] 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.

[8] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Comput. Stat. Data Anal.*, 52(1):502–519, Sept. 2007.

[9] A. M. Buch, P. E. Vértes, J. Seidlitz, S. H. Kim, L. Grosenick, and C. Liston. Molecular and network-level mechanisms explaining individual differences in autism spectrum disorder. *Nat. Neurosci.*, Mar. 2023.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] M. Ciortan and M. Defrance. GNN-based embedding for clustering scRNA-seq data. *Bioinformatics*, 2022.

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

[17] L. Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.*, 29(6):141–142, Nov. 2012.

[18] L. Deng, T. Meng, L. Chen, W. Wei, and P. Wang. The role of ubiquitination in tumorigenesis and targeted drug discovery. *Signal Transduct Target Ther*, 5(1):11, Feb. 2020.

[19] 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.

[20] 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 l1-norm and Truncated Group Lasso for Brain Imaging Genetics. *Bioinform Oxf Engl*, 2017.

[21] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, Sept. 1936.

[22] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, 1995.

[23] X. Z. Fern, C. E. Brodley, and M. A. Friedl. Correlation clustering for learning mixtures of canonical correlation models. In *Proceedings of the 2005 SIAM International Conference on Data Mining (SDM)*, Proceedings, pages 439–448. Society for Industrial and Applied Mathematics, Apr. 2005.

[24] L. Fodor, D. Jakovetić, D. Boberić Krstićev, and S. Škrbić. A parallel ADMM-based convex clustering method. *EURASIP J. Adv. Signal Process.*, 2022(1):1–33, Nov. 2022.

[25] 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.

[26] E. F. Franco, P. Rana, A. Cruz, V. V. Calderón, V. Azevedo, R. T. J. Ramos, and P. Ghosh. Performance comparison of deep learning autoencoders for cancer subtype detection using Multi-Omics data. *Cancers*, 13(9), Apr. 2021.

[27] 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.

[28] J. M. Grible, P. Zot, A. L. Olex, S. E. Hedrick, J. C. Harrell, A. E. Woock, M. O. Idowu, and C. V. Clevenger. The human intermediate prolactin receptor is a mammary proto-oncogene. *NPJ Breast Cancer*, 7(1):37, Mar. 2021.

[29] X. Guo, L. Gao, X. Liu, and J. Yin. Improved deep embedded clustering with local structure preservation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, California, Aug. 2017. International Joint Conferences on Artificial Intelligence Organization.

[30] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer-Verlag New York, 2 edition, 2009.

[31] 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.

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

[33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24(6):417–441, Sept. 1933.

[34] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, Dec. 1936.

[35] 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.

[36] 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.

[37] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, 7(6):673–679, June 2001.

[38] A. Kopf, V. Fortuin, V. R. Somnath, and M. Claassen. Mixture-of-Experts variational autoencoder for clustering and generating from similarity-based representations on single cell data. *PLoS Comput. Biol.*, 17(6):e1009086, June 2021.

[39] 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.

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

[41] 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.

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

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

[44] B. Motamedi, H.-A. Rafiee-Pour, M.-R. Khosravi, A. Kefayat, A. Baradaran, E. Amjadi, and P. Goli. Prolactin receptor expression as a novel prognostic biomarker for triple negative breast cancer patients. *Ann. Diagn. Pathol.*, 46:151507, June 2020.

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

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, 24(3):227–235, Mar. 2000.

[51] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[52] D. Sa-Nguanraksa, C. Thasripoo, N. Samarnthai, T. Kummalue, T. Thumrongtaradol, and P. O-Charoenrat. The role of Prolactin/Prolactin receptor polymorphisms and expression in breast cancer susceptibility and outcome. *Transl. Cancer Res.*, 9(10):6344–6353, Oct. 2020.

[53] C. Santos, R. Sanz-Pamplona, 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.

[54] 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.

[55] 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 metabolomic characterization of COVID-19 patient sera. *Cell*, 182(1):59–72.e15, July 2020.

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

[57] 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.

[58] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, 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.

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

[60] 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.

[61] Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, Oct. 2018.

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

[63] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.*, 9(1):1–12, Mar. 2019.

[64] 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.

[65] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning: Objectives and optimization. *arXiv*, Feb. 2016.

[66] X. Wang, T. Zhang, S. Zhang, and J. Shan. Prognostic values of f-box members in breast cancer: an online database analysis and literature review. *Biosci. Rep.*, 39(1), Jan. 2019.

[67] M. Weylandt. Splitting methods for convex bi-clustering and co-clustering. In *2019 IEEE Data Science Workshop (DSW)*, pages 237–242, June 2019.

[68] 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.

[69] 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.

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

[71] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. Aug. 2017.

[72] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. *arXiv*, Nov. 2015.

[73] 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.

[74] C. You, C.-G. Li, D. P. Robinson, and R. Vidal. Oracle based active set algorithm for scalable elastic net subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3928–3937, 2016.

[75] 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.